Mining of Texas Standardized Exam Grades

Abstract

This research paper uses a data sets related to demographics information and student scores on two different standardized tests and by combining them together and analyzing them, tries to answer different questions and open up a discussion regarding academic performance and ways in which it can be predicted.

Introduction

Performance of students in schools has always been an important topic. On one hand, every single parent wants their children to have the best education possible and have the best grades as well. On the other, the government needs to have at least a minimum standard on how well educated the citizens are. But are there any ways in which we can determine beforehand how well a student will perform in school? Do variables outside the student's control responsible for a big part of how they perform?

For example, is gender a determinant factor in how well a student performs? If you read conventional sociological studies, many researchers' answer to this question is yes. The common belief is that girls perform better than boys in school. And how about demographic factors? Is the student's environment responsible in some way of how they perform?

On this research paper we explore those two factors and try to verify or contradict the conventional wisdom surrounding academic performance for students.

Sources of Data (Gathering and Cleaning)

Through Ruben's company we had access to the grades of over 100,000 students from Kindergarten to the 3rd Grade for the Texas Primary Reading Inventory (TPRI) assessment and its Spanish version, Tejas LEE. This data set (**Raw Student Info Data Set**) was contained in a CSV file with 22 fields related to a student's performance on standardized tests. These fields are:

- CountyName - county where the school student attends resides
- SchoolStreetAddress - postal address of the school student attends
- SchoolCity - city of the school student attends
- SchoolState - state of the school student attends
- SchoolZip - zip code of the school student attends
- DistrictId - unique identifier of the school district the student is in
- District - name of the school district the student is in
- SchoolId - unique numerical identifier for the school the student attends
- SchoolName - name for the school the student attends
- TeacherName - name of teacher for the student
- ClassName - class during which the standardized test was administered
- DateAdmin - date the standardized test was taken by the student

- StudentKey - unique GUID number that idenfities a student
- StudentId - unique 10 digit number that identifies each student
- StudentName - the student's name
- StudentGradeLevel - student's grade level
- DOB - student's date of birth
- Gender - student's gender
- Factor - An overall score specific to the kind of test taken
- CalculatedScore - A value derived from Factor in a scale of 0 to infinity where 0 is the highest score. This was necessary to do so we could measure each student on the same scale.
- ScoreAsPercentage - A value derived from CalculatedScore in a scale of negative infinity to 100 where 100 is the highest value
- GradeLetter - A value derived from ScoreAsPercentage used to simplify our outcomes in a standard scale normally used in a school setting. This is the standard A to F scale used in schools where A, B and C are assumed as passing grades.

We also looked up demographic data for all zip codes specified in the **Raw Student Info Data Set** from the following website (http://zipwho.com/). We wrote a program in python to record all data for each zip code and save it to a CSV (**Raw Zipcode Info Data Set**). The site provided many pieces of information for each zip code, out of which we used the following:

- Zipcode - the zip code for our demographic data
- Median_Household_Income - median household income for this zip code
- Married - percent of adults married in this zip code
- Divorced - percent of adults divorced in this zip code
- White - percent of adults that self identify as White in this zip code
- Black - percent of adults that self identify as Black in this zip code
- Asian - percent of adults that self identify as Asian in this zip code
- Hispanic - percent of adults that self identify as Hispanic in this zip code

Our sources where purged of bad data by various ways. For the **Raw Student Info Data**, we eliminated records for which the scores did not make sense. For example we eliminated 4 records for which the kids had an impossible score of over 100%. This was interesting because 4 records out of 338,000 is a miniscule amount of error but it showed that there is always human error involved when doing data analysis and that the analysts need to decide the best approach to take with the unusable data. There where also a small amount of zip codes for which we did not found demographic data and we decided to remove those zip codes from the **Raw Zipcode Info Data Set** as well as whichever records on the **Raw Student Info** matched them. This accounted for about 150 more records eliminated from the student information data set.

Goals

During the semester we learned data mining techniques that allows us to better analyze data. We have learned how to:

- perform Principal Component Analysis
- create linear regression models to try to understand and predict data
- use classification techniques to categorize data
- use computer tools to aid us in working with this data

Using these techniques among other concepts we have learned in our class, we applied current data mining tools and techniques to analyze all our gathered data. By doing this, we where able to better answer questions about gender and demographic factors' influence on a student's or school's performance.

Since our **Raw Student Info Data Set** consisted of 2 versions of a standardized test we analyzed them separately and then verified that our findings where the same for both. The approach we took for this research was to perform different kinds of analysis to answer the following questions. These helped clear up some of the general conceptions about gender and performance in an academic setting and also gave us a clearer view about the many things that in an indirect way may be influencing performance as well. The questions selected where:

- In general who performs better in Texas schools, girls or boys?
- Do girls in lower grades outperform boys? If so, do boys begin to outperform girls as they get older?
- Does the small difference in age on each grade have an impact on performance?
- Are there demographic characteristics that can influence the performance of schools? If so, which are they?
- Which is the school or district with the best overall performance in Texas? Do they follow any trend found in the analysis?

We will answer each of these questions on the subsequent sections and explain the approach taken for each of them.

Analysis and Findings

In general who performs better in Texas schools, girls or boys?

Do girls in lower grades outperform boys?

If so, do boys begin to outperform girls as they get older?

To answer this question we analyzed all of the data for each of the assessments at an overall level as well as at each of the grade levels.

At the overall level we noticed that the girls outperform the boys. We felt this information needed to be investigated at a deeper level (dimension) as the number of students varies greatly per grade hence the general idea could possibly not translate to the individual grade levels dimension. Additionally, this would give us the answer to our second part of the question. Does the trend still continue or do boys get better as tests as time passes.

After analyzing the data by performing various SQL queries on our data sets (APPENDIX A), we determined that indeed the trend found at an overall level keeps being true at each of the grade levels. Girls keep outperforming boys independently of which grade level they are at.

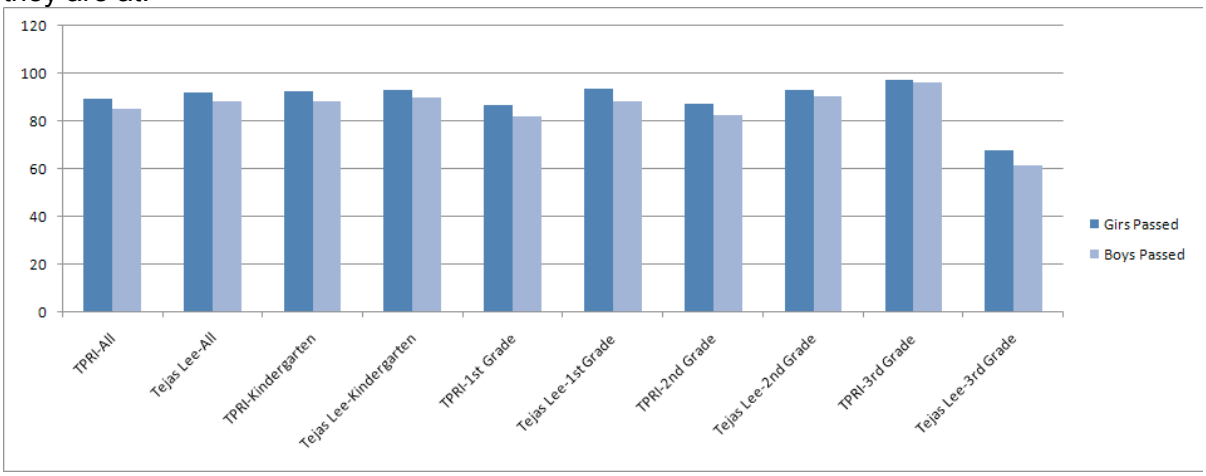

Figure 1

Does the small difference in age on each grade have an impact on student performance?

To answer this question, we charted the average scores of students from each grade, sliced by the month they were born in. All charts begin with children born in September for that school year and end with children born in August of that school year. These charts only show students born between this 12-month period, for example a student held back a year is not shown. Also, for this specific analysis no difference between assessments needed to be made as we where looking at students on the same grade level who took the same tests.

Kindergarten

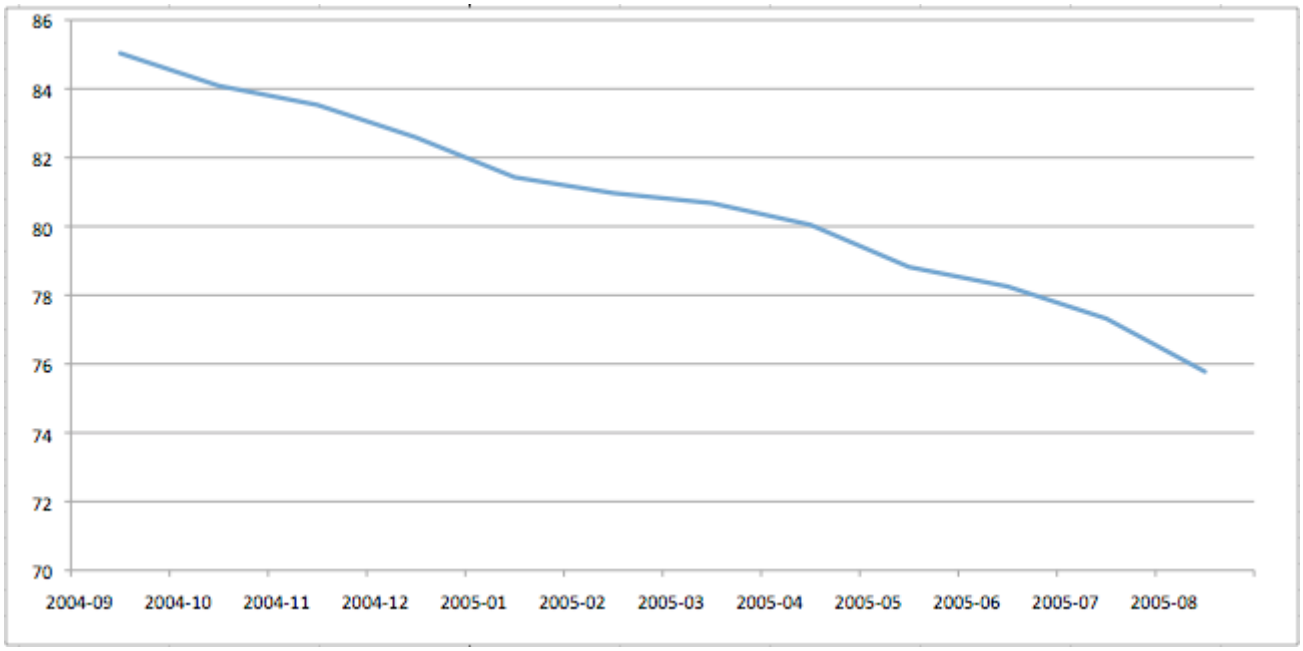

Figure 2 - Kindergarten Scores by Date of Birth

First Grade

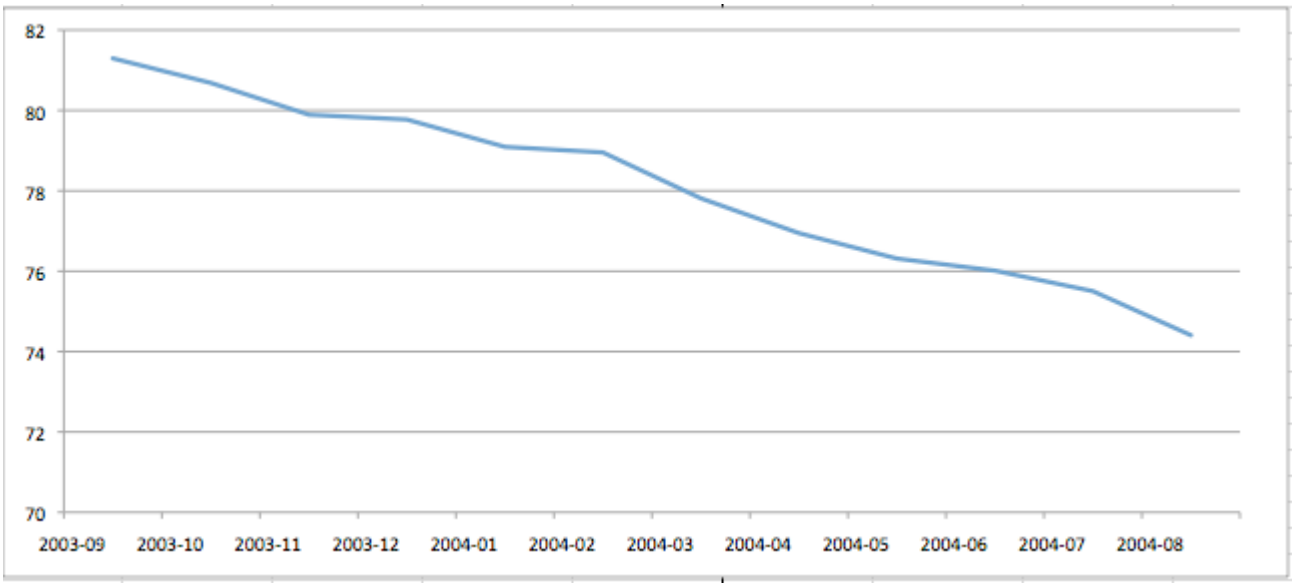

Figure 3 - 1st Grade Scores by Date of Birth

Second Grade

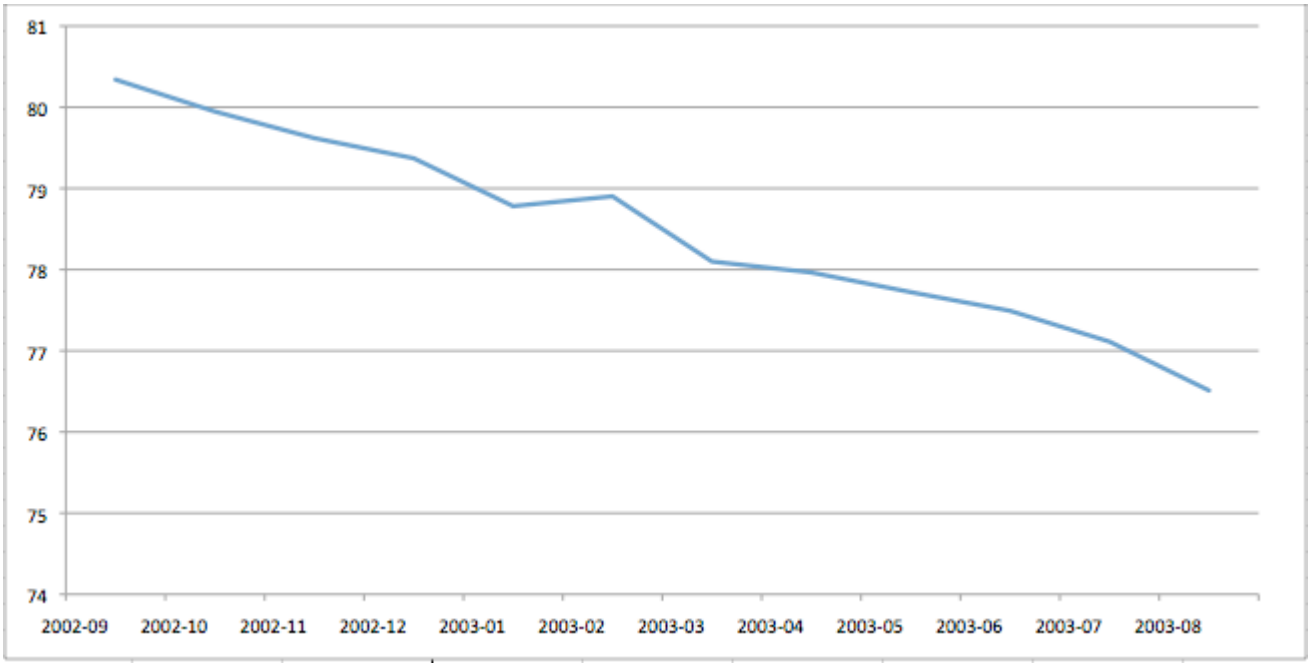

Figure 4 - 2nd Grade Scores by Date of Birth

Third Grade

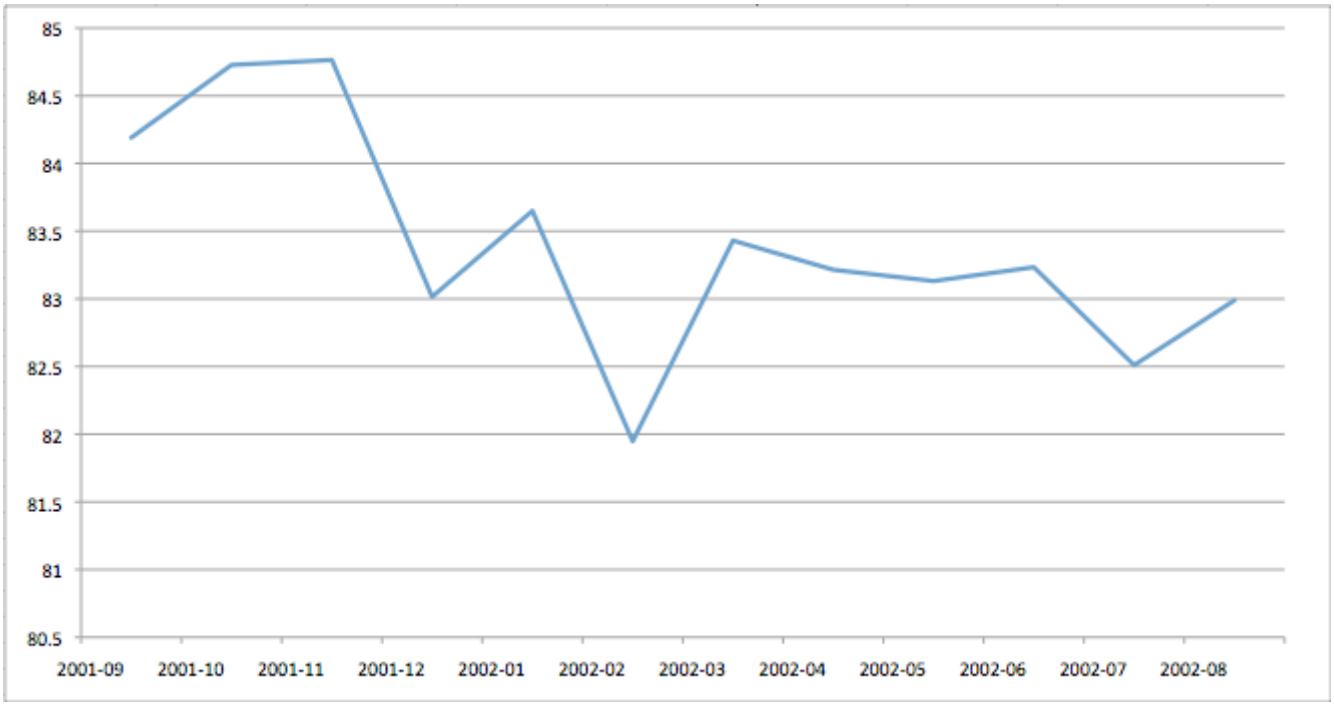

Figure 5 - 3rd Grade Score Scores by Date of Birth

As shown by the graphs, there is an obvious trend for grades Kindergarten through Second. The charts clearly show that children born earlier in the school year on average outperform children born later in the school year. However, it was surprising that this trend is disrupted in third grade. This analysis opened up a lot of questions. Does this trend keep occurring at other grade levels? Or is it something that only presents itself on those early grades? Unfortunately, we can't answer or investigate further due to the fact we do not have data for more grades and or data for other types of standardized tests.

We also charted the performance of students who by age should have been in Kindergarten and students held back a year. Below is the chart that represents this data.

Kindergarten

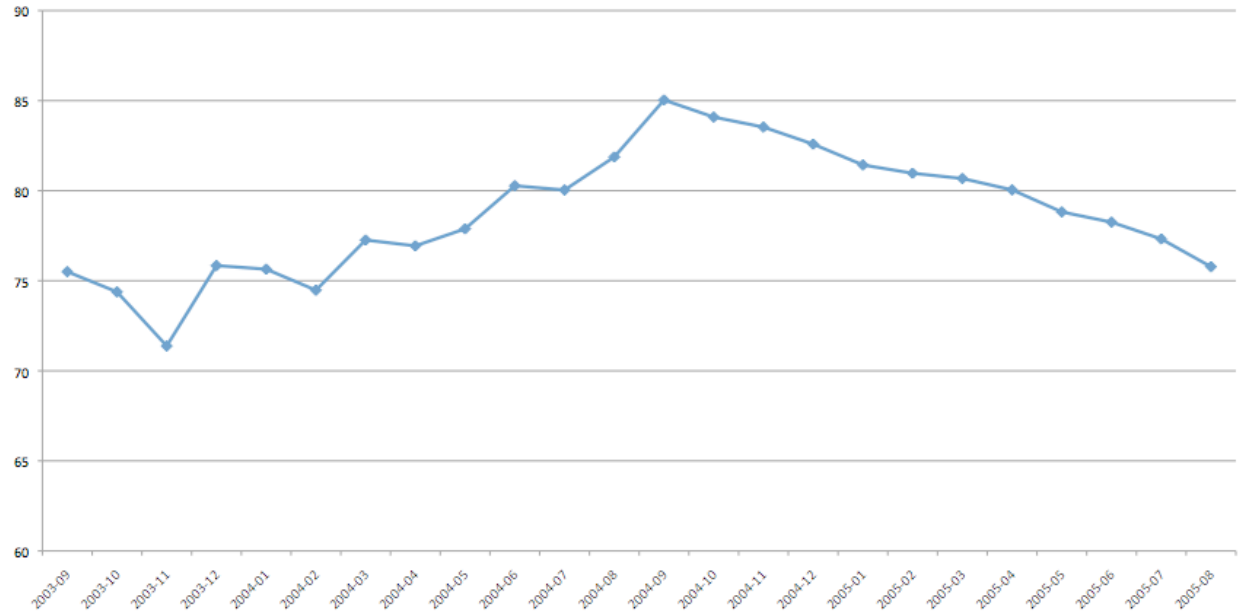

Figure 6 - 3rd Grade Score Scores by Date of Birth

First Grade

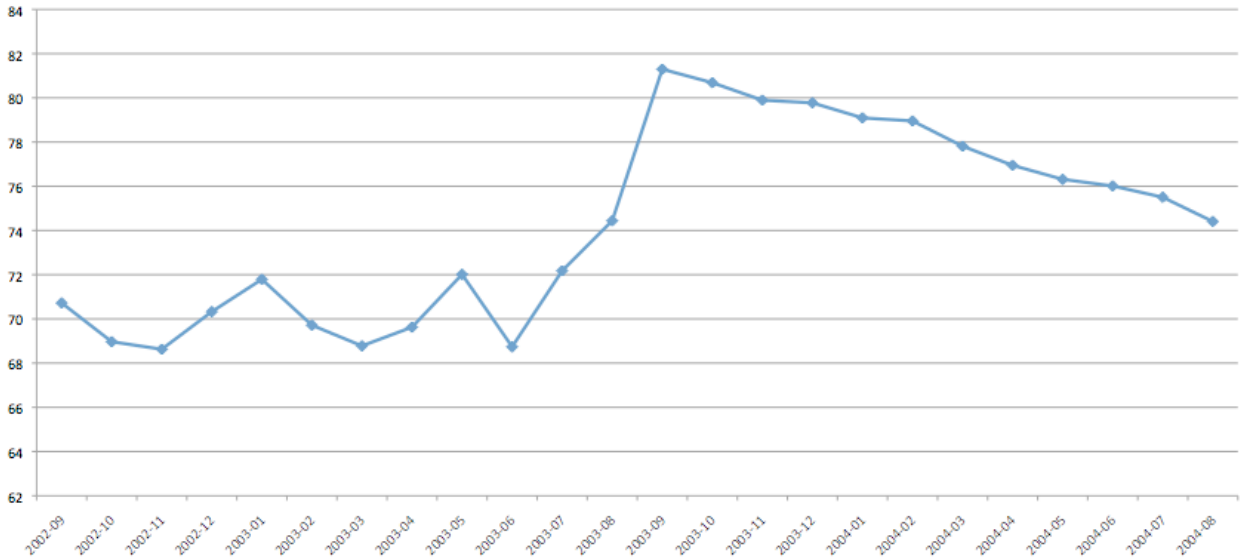

Figure 7 - 3rd Grade Score Scores by Date of Birth

Second Grade

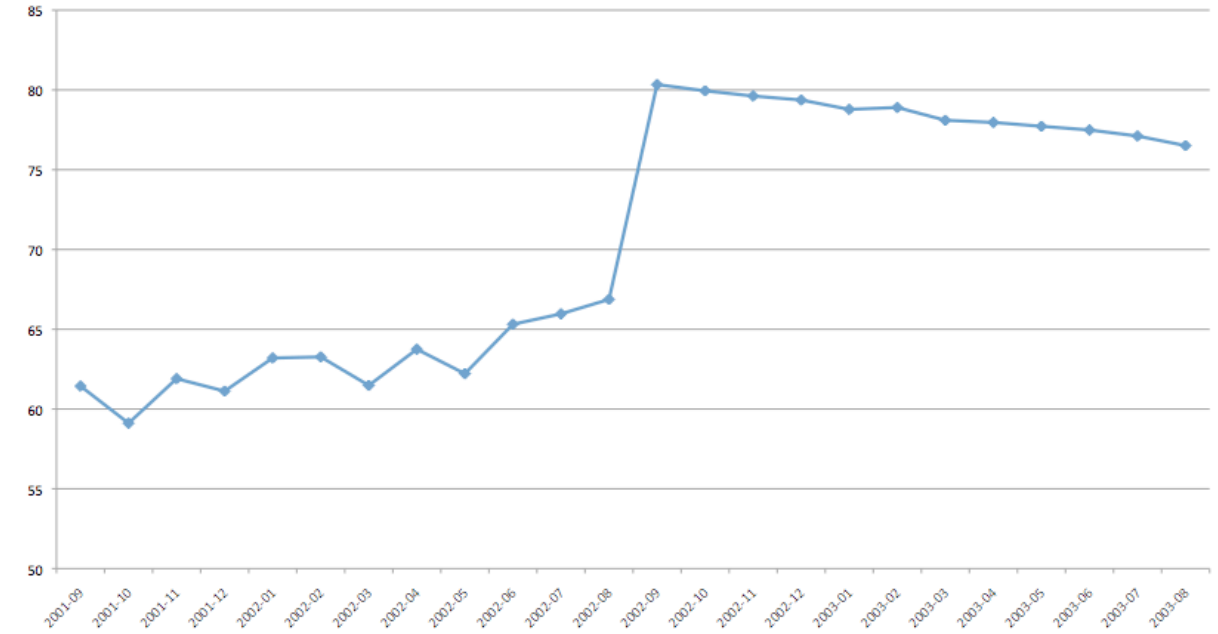

Figure 8 - 3rd Grade Score Scores by Date of Birth

Third Grade

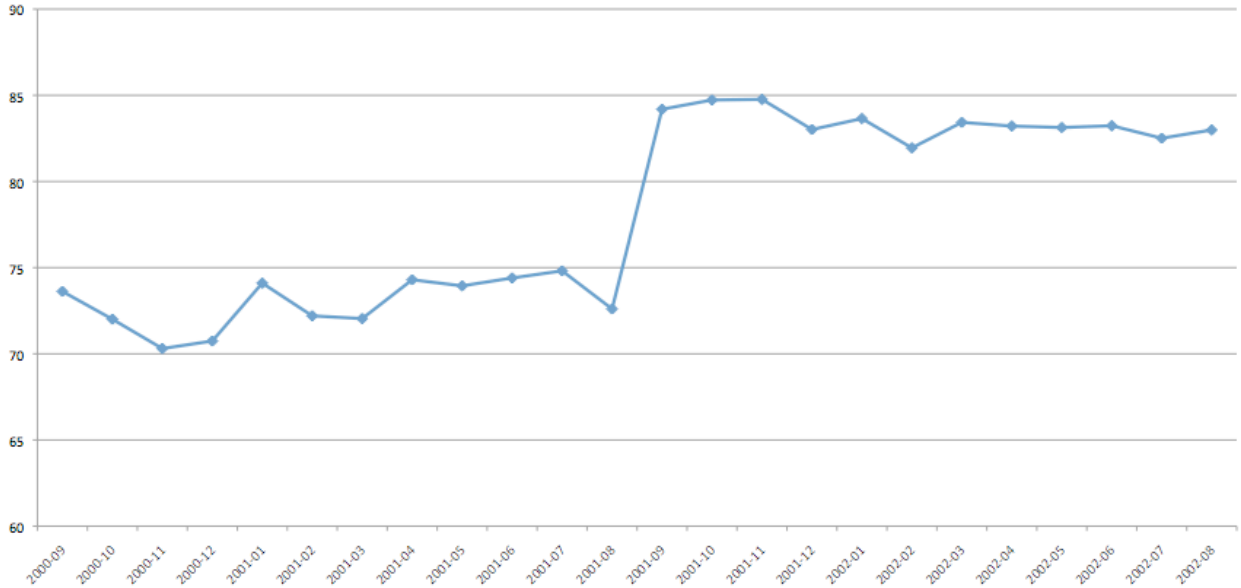

Figure 9 - 3rd Grade Score Scores by Date of Birth

We found these results to be surprising at first. It seems that children held back in Kindergarten often perform around the same level as the children who are not. However, for every grade after Kindergarten, if a child has been held back a year this child badly underperforms all other children in the grade. The drop off in performance cannot be explained by saying that children who were a year older than their peers in Kindergarten suddenly performed worse. A more likely explanation would be that the population of students a “year

older" in first through third grade changes as children being forced to repeat a grade are added to the population. We found this finding very intriguing, and it warrants further investigation.

Are there any demographic characteristics that can influence the performance? If so, which are they?

As we tried to determine what exactly described a having good performance we determined that it may depend on what level or dimension you wanted to find the performance at.

Because of this, we consulted with a district director from a school district in Texas and he stated that from the standpoint of the department of education the measure they would care about would be at a campus level (or dimension). That is, to measure the performance at that scope we need to calculate a score for each campus by performing a query on our database that did an average on all of their students together (APPENDIX B – Overall Query). With this derived data set (**Overall Analysis Data Set**) exported as a CSV file we used Microsoft Excel and the XLMiner plugin to perform various analysis.

However, there was another way that we could measure the performance of a particular school, which was more inclusive of what it meant for each student to perform well or not. That is, to do a measure of what percentage of students for the school have passed or not the assessment. By passing, we mean all students that had a grade of C or better on their scores (standard A to F scale). This kind of analysis would also help us determine the demographics' influence from a different perspective. It could help us determine if a student is most likely to be successful or not depending on the school they are in and in turn, should there be a trend on the demographics of the schools whose passage percentage themselves are above a "passing grade" then we could infer that those demographics can be good predictors of a student's performance. By performing a more complex query (APPENDIX B – Passed Percentages Queries) we aggregated the data to create a data set that tells us for each of the campuses the percentage of students that passed. We then converted each schools percentage into a grade letter (again following the standard A to F scale) This second derived data set (**Passing Percentage Data Set**) was also exported as a CSV file and analyzed using Microsoft Excel and the XLMiner plugin.

The first thing we did to get an overall idea or hints of any patterns on the data was to do simple 2 dimensional graphs of the attributes for both the **Overall Analysis Data Set** and the **Passing Percentage Data Set** against the outcome for each of the schools. This can be seen on Figures 10 to 21.

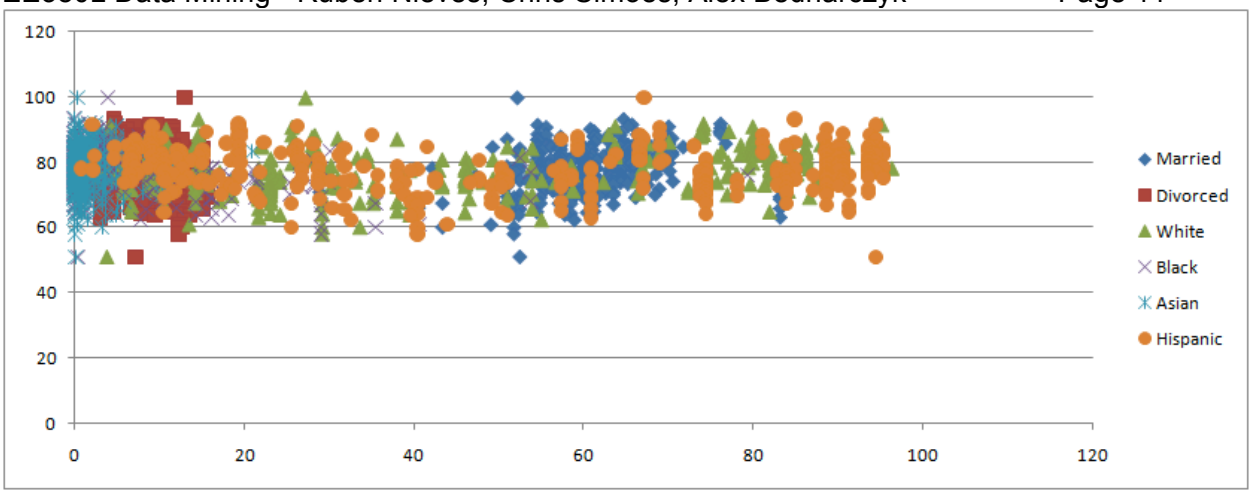

Figure 10 - Overall Analysis Natonialities % Plot - TPRI

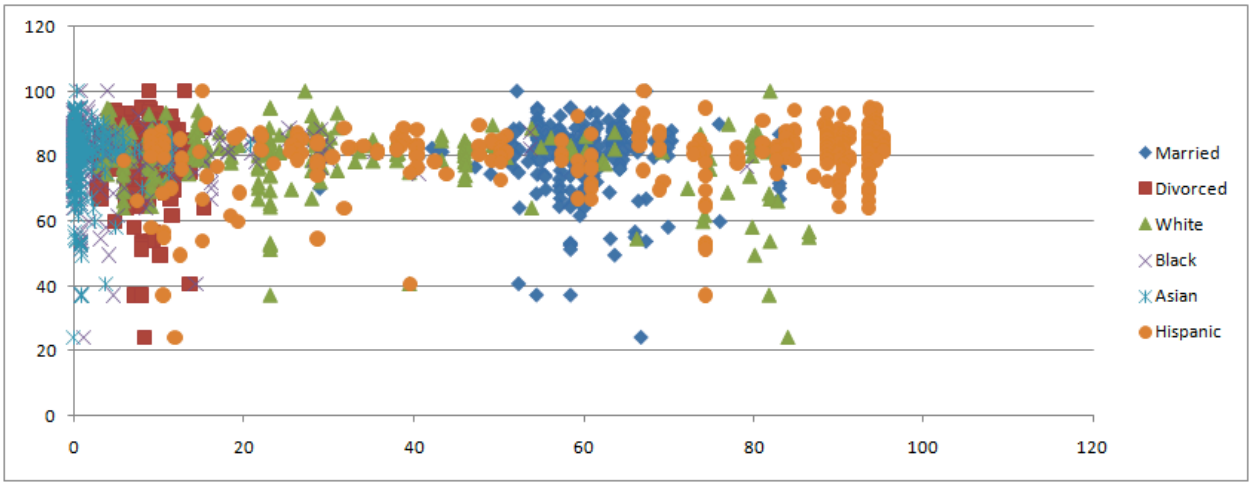

Figure 11 - Overall Analysis Nationalities % Plot - TejasLEE

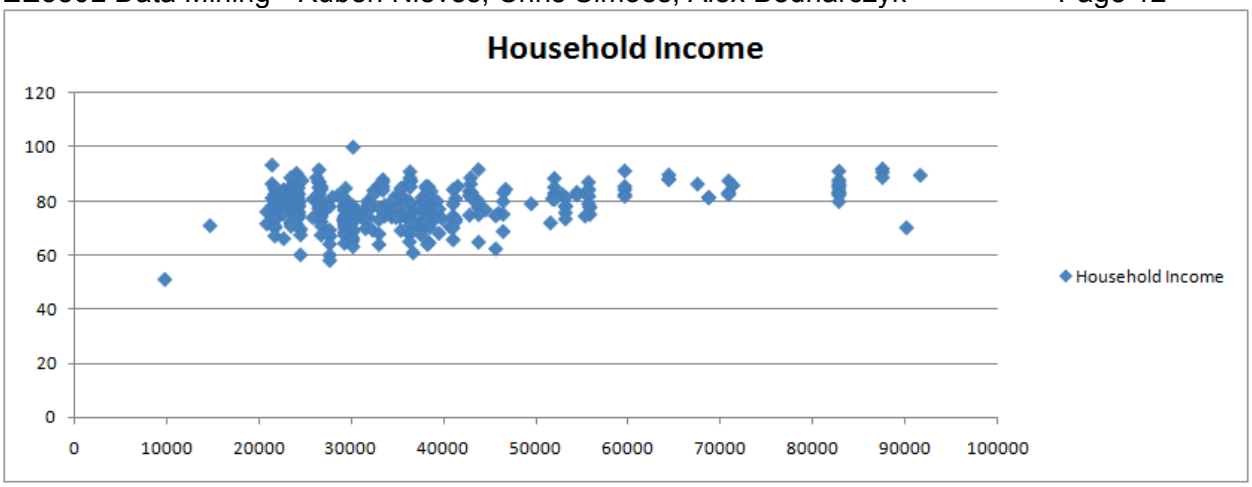

Figure 12 - Overall Analysis Household Income Plot - TPRI

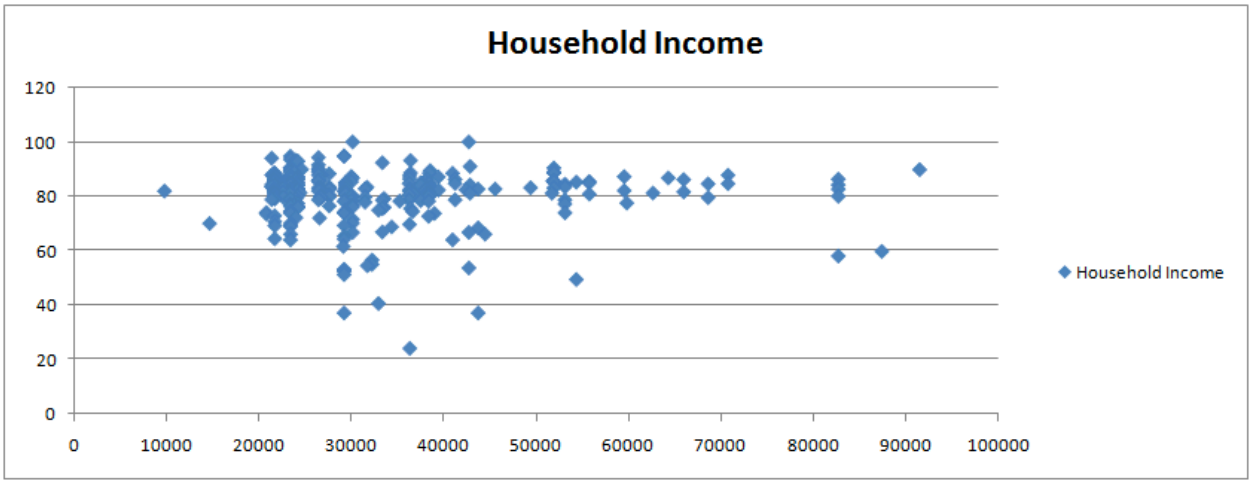

Figure 13 - Overall Analysis Household Income Plot - TejasLEE

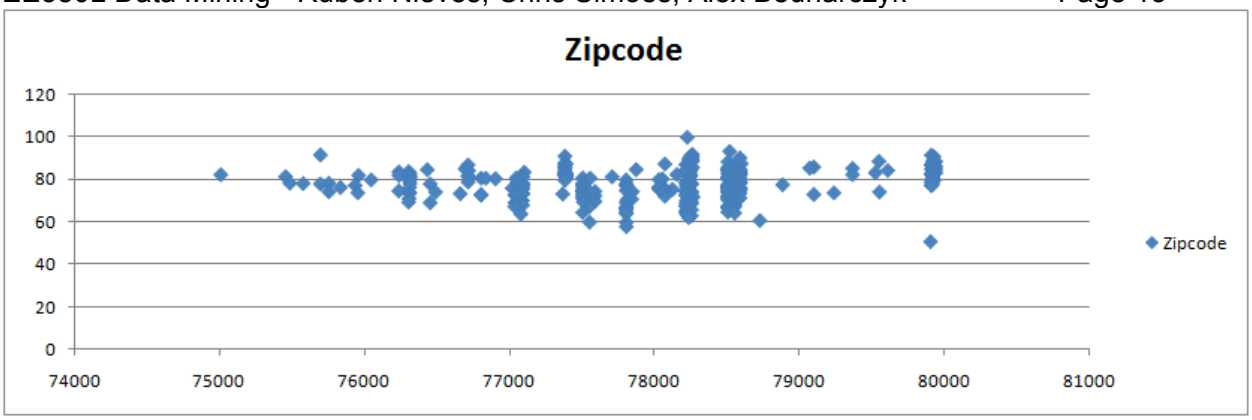

Figure 14 - Overall Analysis Zipcode Plot – TPRI

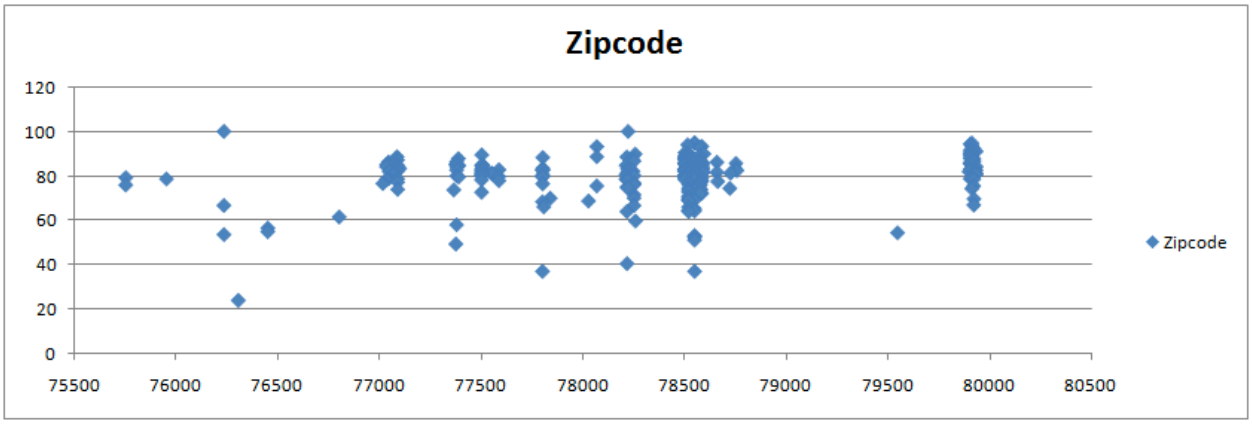

Figure 15 - Overall Analysis Zipcode Plot - TejasLEE

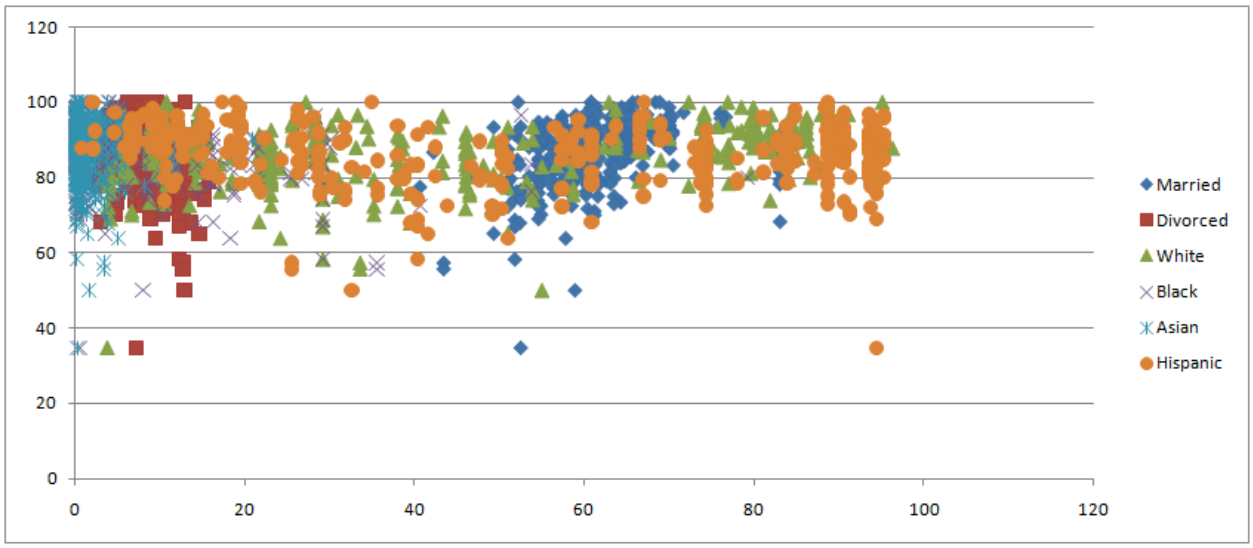

Figure 16 - Passing Percentage Analysis Nationality % Plot - TPRI

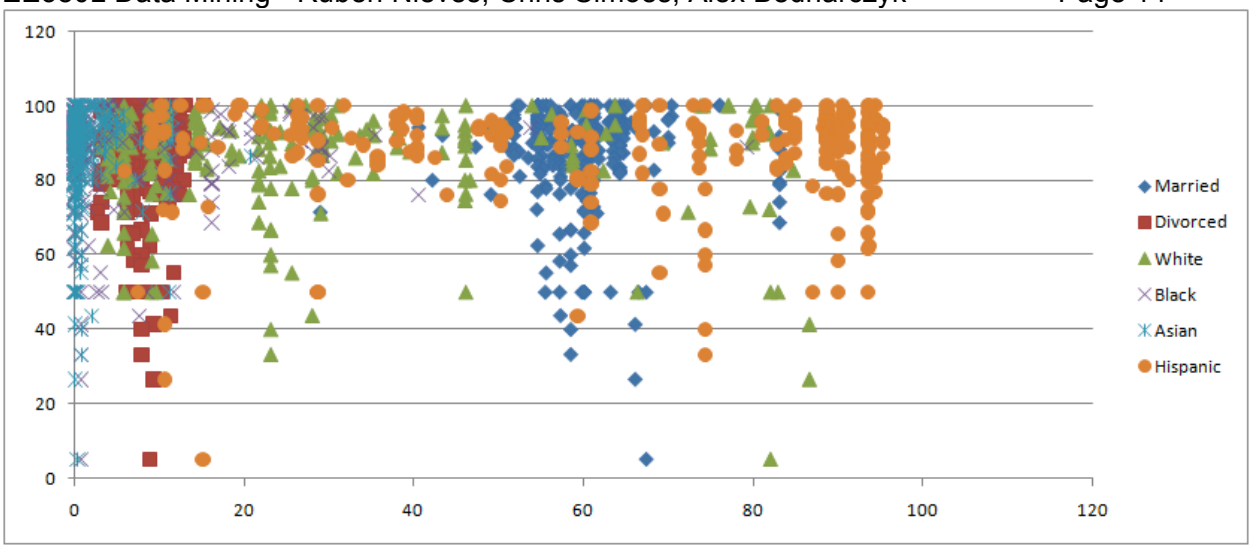

Figure 17 - Passing Percentage Analysis Nationality % Plot - Tejas LEE

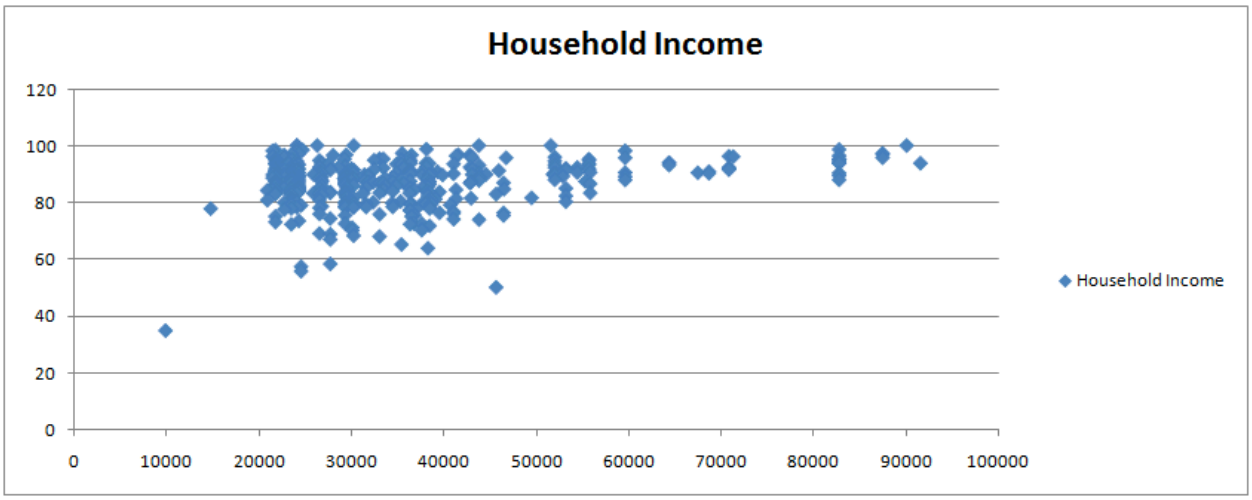

Figure 18 - Passing Percentage Analysis Household Income Plot - TPRI

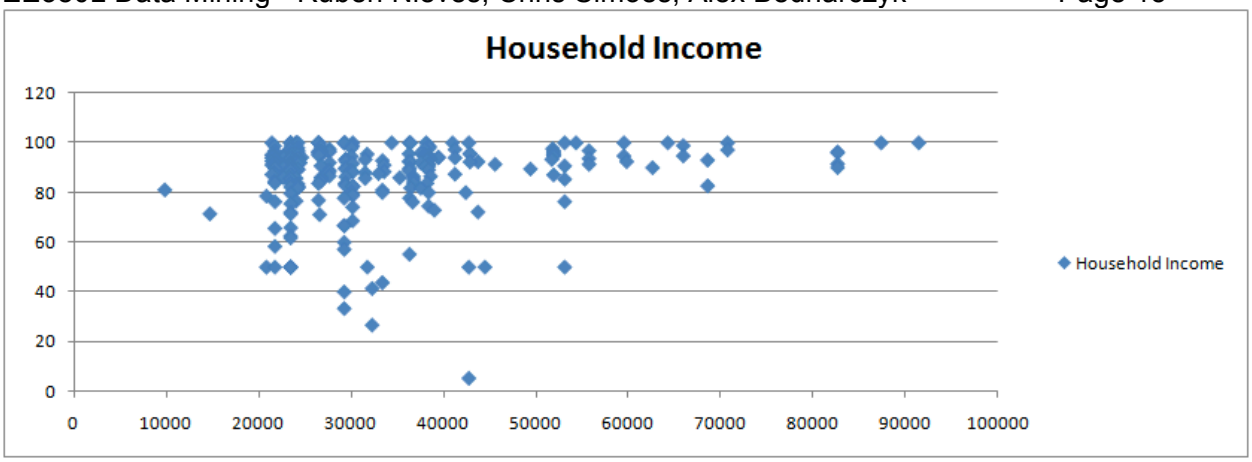

Figure 19 - Passing Percentage Analysis Household Income Plot - Tejas LEE

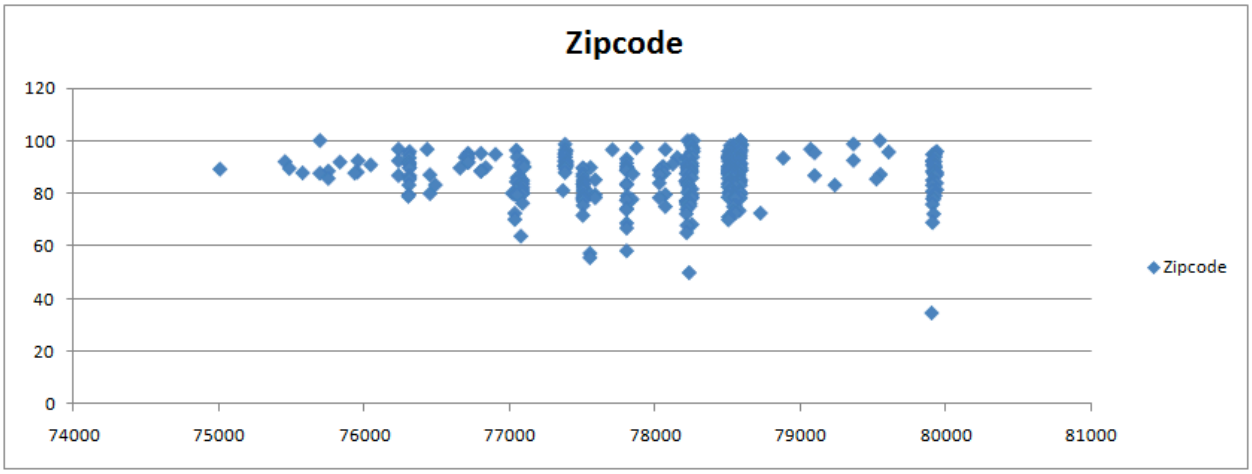

Figure 20 - Passing Percentage Analysis Zipcode Plot – TPRI

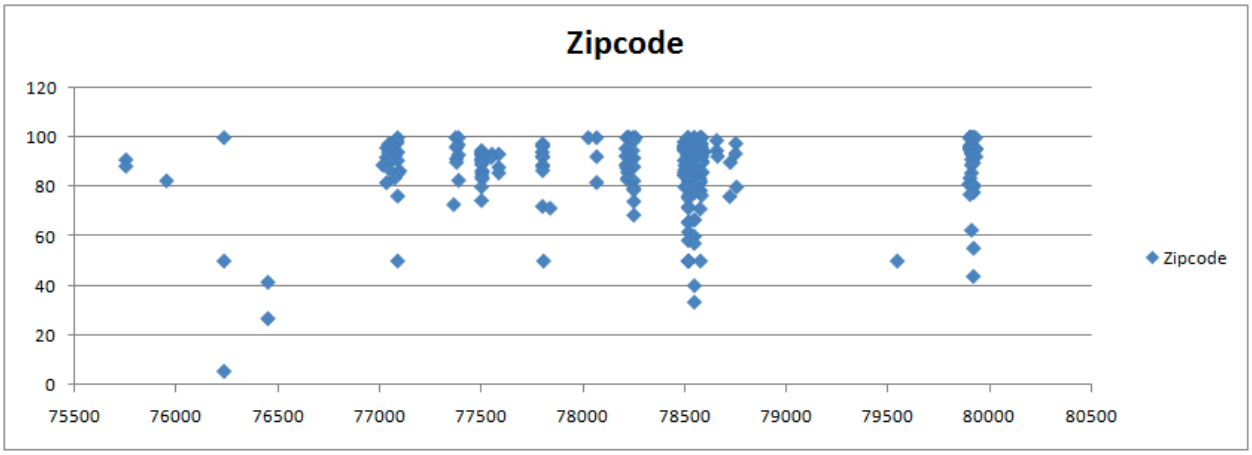

Figure 21 - Passing Percentage Analysis Zipcode Plot - Tejas LEE

Just by looking at these graphs, we saw that there was no clear pattern. For all of the different attributes plotted, there was either a straight-line trend, meaning that they all pretty much do the same in average or nothing following a path at all.

So the next step was then to try different kinds of analysis to see if something could be derived out of this data.

The first step was to perform a PCA analysis on the data sets per assessment. We first started with the **Overall Analysis Data Set**. For both assessments the principal component identified by XLMiner was the percentage of married people living on the school's zip code location as shown in Figure 22 and Figure 23. We also plotted the first two principal components against the grade letter the school would have gotten based on their score (using a standard A to F grade letter scale) but did not notice any clusters which could graphically show us that just by knowing those two components we could make a prediction as to what grade they would get. These plots can be seen in Figure 23 for TPRI and Figure 24 for TejasLEE.

Inputs

Data	
Input data	['MyAnalysisOverallLevel.xlsx']'OverallScoresPerSchool-
# Records in the input data	425

Variables							
# Selected Variables	7						
Selected variables	Married	Divorced	White	Black	Asian	Hispanic	MedianHouseholdIncome

Parameters/Options	
Fixed # of components	7
Method	Covariance matrix
Show data scores	Yes

Principal Components

	Components						
Variable	1	2	3	4	5	6	7
Married	0.00065847	0.22737008	0.57806563	0.36152014	-0.63164556	0.04687065	-0.28683227
Divorced	0.00009694	0.00672224	0.06887183	0.09528715	0.51046395	0.0644412	-0.84935606
White	0.00119626	-0.42577711	0.77630842	-0.14154965	0.35863563	0.00448809	0.25957978
Black	0.00007694	-0.04487091	-0.08533587	0.90747541	0.297571	-0.08630515	0.26682508
Asian	0.000064	-0.00635964	-0.04095352	0.0553569	0.02364598	0.99303699	0.09239282
Hispanic	-0.00047505	0.8745954	0.22244343	-0.11666708	0.35031882	-0.0077023	0.2218283
MedianHouseh	0.99999893	0.00077831	-0.00120112	-0.00020674	0.00007942	-0.00010306	0.00003961

Variance	307886200	1467.18457	295.6411438	91.26940918	20.01994133	3.12321949	2.36878538
Variance%	99.99938965	0.00047653	0.00009602	0.00002964	0.0000065	0.00000101	0.00000077
Cum%	99.99938965	99.99986267	99.99996185	99.99999237	99.99999237	99.99999237	100

Figure 22 - PCA Analysis of TPRI with Overall Score

Inputs

Data	
Input data	['MyAnalysisOverallLevel.xlsx']'OverallScoresPerSchool-
# Records in the input data	304

Variables							
# Selected Variables	7						
Selected variables	Married	Divorced	White	Black	Asian	Hispanic	MedianHouseholdIncome

Parameters/Options	
Fixed # of components	7
Method	Covariance matrix
Show data scores	Yes

Principal Components

	Components						
Variable	1	2	3	4	5	6	7
Married	0.00044978	-0.25566548	0.45530972	0.53823733	-0.59873629	0.02440313	-0.28026408
Divorced	0.0000954	0.0019095	0.02926715	0.07460696	0.49347991	-0.00975553	-0.86599994
White	0.00121221	0.37115338	0.81458479	0.02321703	0.37094116	0.04893343	0.24117343
Black	0.00012919	0.10820118	-0.30648687	0.83751905	0.34909862	-0.05143369	0.26154307
Asian	0.00008859	0.01505705	-0.07080103	0.03065783	0.01029675	0.99683946	-0.00508037
Hispanic	-0.00089682	-0.88596916	0.17128988	-0.04262737	0.37204728	0.02409626	0.21189816
MedianHouseh	0.99999875	-0.00114498	-0.00099555	-0.00042649	0.00006021	-0.00012942	0.00007302

Variance	220671500	999.8872681	160.6746216	92.48401642	20.4437027	3.96244884	2.05554867
Variance%	99.99942017	0.00045311	0.00007281	0.00004191	0.00000926	0.0000018	0.00000093
Cum%	99.99942017	99.99987793	99.99994659	99.99999237	100	100	100

Figure 23 - PCA Analysis of TejasLEE with Overall Score

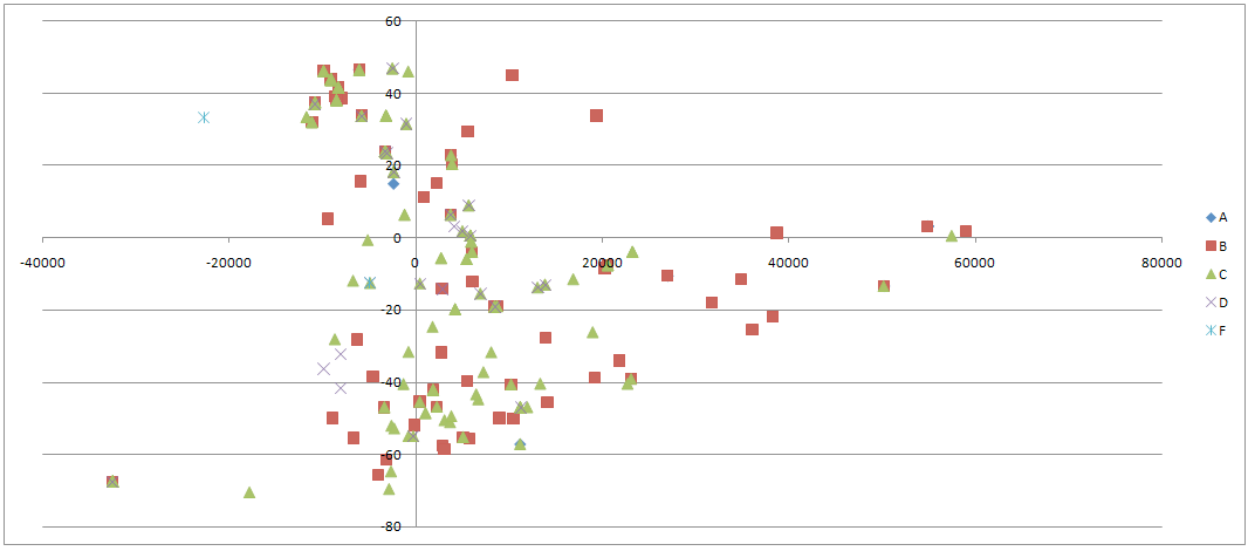

Figure 24 - TPRI Principal Components and Grades Plot for Overall Scores

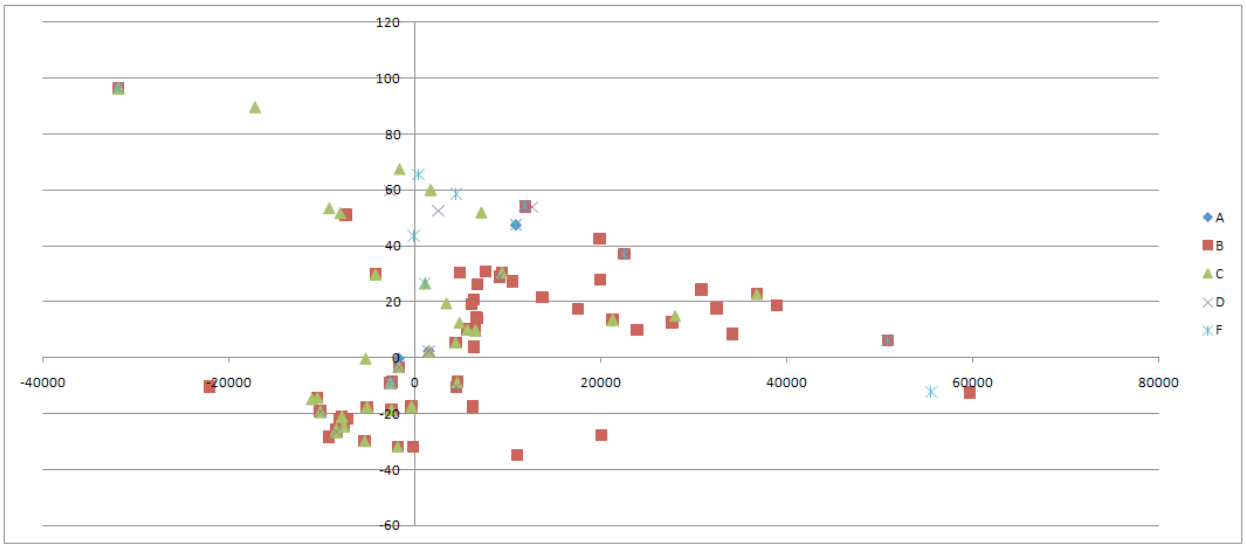

Figure 25 - TejasLEE Principal Components and Grades Plot for Overall Scores

We then proceeded to perform the same procedure with the **Passing Percentage Data Set**. For both assessments the principal component identified by XLMiner was again the percentage of married people living on the schools zip code's location as shown in Figure 26 and Figure 27. We also plotted the first two principal components against the grade letter the school would have gotten based on their passed percentage (again using the standard A to F grade letter scale) and again did not notice any clusters which could graphically show us that just by knowing those two components we could make a prediction as to what grade they would get. These plots can be seen in Figure 28 for TPRI and Figure 29 for TejasLEE

Inputs

Data	
Input data	['MyAnalysis.xlsx']'TPRI Correctly Grouped'!$B$5:$M$429
# Records in the input data	425

Variables							
# Selected Variables	7						
Selected variables	MARRIED	DIVORCED	WHITE	BLACK	ASIAN	HISPANIC	MEDIAN_HOUSEHOLD_INC

Parameters/Options	
Fixed # of components	7
Method	Covariance matrix
Show data scores	Yes

Principal Components

	Components						
Variable	1	2	3	4	5	6	7
MARRIED	0.00065847	0.22737008	0.57806563	0.36152014	-0.63164556	0.04687065	-0.28683227
DIVORCED	0.00009694	0.00672224	0.06887183	0.09528715	0.51046395	0.0644412	-0.84935606
WHITE	0.00119626	-0.42577711	0.77630842	-0.14154965	0.35863563	0.00448809	0.25957978
BLACK	0.00007694	-0.04487091	-0.08533587	0.90747541	0.297571	-0.08630515	0.26682508
ASIAN	0.000064	-0.00635964	-0.04095352	0.0553569	0.02364598	0.99303699	0.09239282
HISPANIC	-0.00047505	0.8745954	0.22244343	-0.11666708	0.35031882	-0.0077023	0.2218283
MEDIAN_HOUS	0.99999893	0.00077831	-0.00120112	-0.00020674	0.00007942	-0.00010306	0.00003961

Variance	307886200	1467.18457	295.6411438	91.26940918	20.01994133	3.12321949	2.36878538
Variance%	99.99938965	0.00047653	0.00009602	0.00002964	0.0000065	0.00000101	0.00000077
Cum%	99.99938965	99.99986267	99.99996185	99.99999237	99.99999237	99.99999237	100

Figure 26 - PCA Analysis of TPRI with Passing Percentages

Inputs

Data	
Input data	['MyAnalysis.xlsx']'Tejas LEE Correctly
# Records in the input data	291

Variables							
# Selected Variables	7						
Selected variables	MARRIED	DIVORCED	WHITE	BLACK	ASIAN	HISPANIC	MEDIAN_HOUSEHOLD_INC

Parameters/Options	
Fixed # of components	7
Method	Covariance matrix
Show data scores	Yes

Principal Components

	Components						
Variable	1	2	3	4	5	6	7
MARRIED	0.00044096	-0.26028347	0.43165895	0.56344128	-0.59229189	0.01784541	-0.2780804
DIVORCED	0.00009702	0.00205648	0.02281779	0.06927676	0.48890752	-0.02323558	-0.86896801
WHITE	0.0012026	0.35773832	0.81625533	0.05331216	0.37837437	0.0559865	0.23791841
BLACK	0.00013584	0.11896925	-0.34292948	0.82046765	0.35584107	-0.04320059	0.25804895
ASIAN	0.00009115	0.01685614	-0.07263949	0.02510508	0.00475518	0.99660331	-0.02383907
HISPANIC	-0.00091059	-0.88872534	0.1549159	-0.03308786	0.37462917	0.03037536	0.20929213
MEDIAN_HOUS	0.99999875	-0.00114261	-0.00097992	-0.00046316	0.00005108	-0.00013026	0.0000785

Variance	215395400	965.4563599	152.0139008	94.04866028	20.70024681	4.10601521	1.97747636
Variance%	99.9994278	0.00044822	0.00007057	0.00004366	0.00000961	0.00000191	0.00000092
Cum%	99.9994278	99.99987793	99.99994659	99.99999237	100	100	100

Figure 27 - PCA Analysis of TejasLEE with Passing Percentages

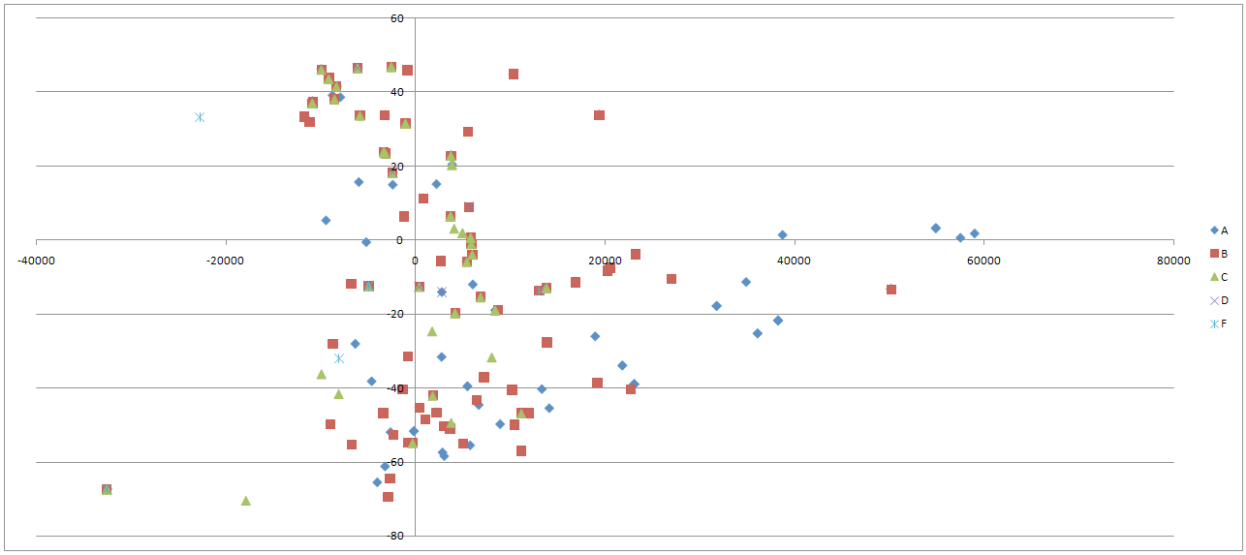

Figure 28 - TPRI Principal Components and Grades Plot for Passing Percentages

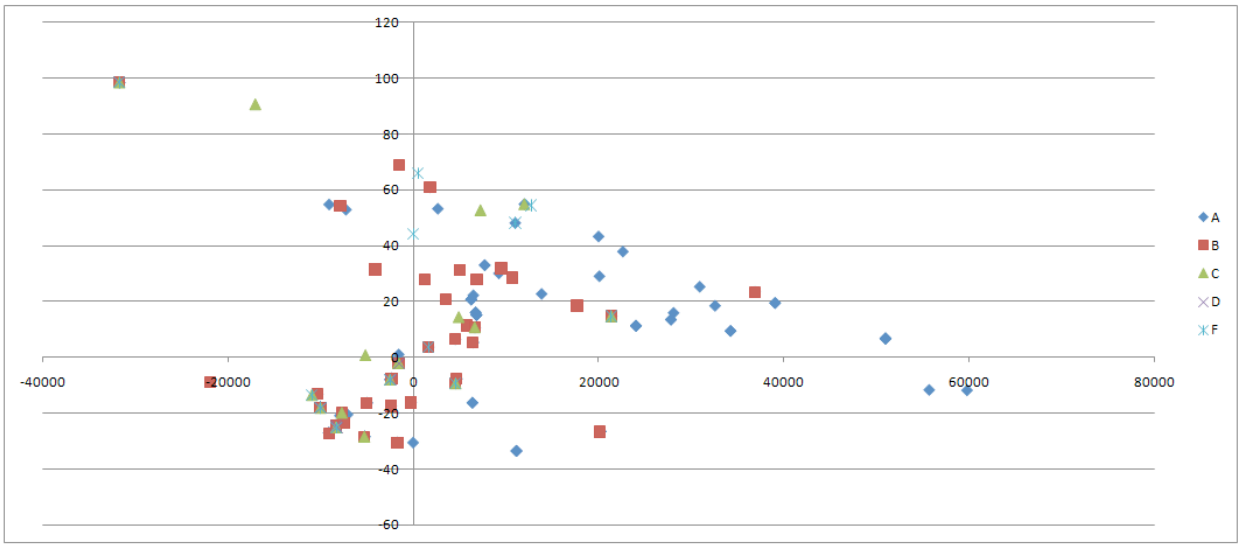

Figure 29 - TejasLEE Principal Components and Grades Plot for Passing Percentages

Since a PCA analysis did not give us a clear indicator of any attribute influencing the performance we moved on to another type of approach to see if we could get some answers. We performed a multiple linear regression on both the **Overall Analysis Data Set** and the **Passing Percentage Data Set** using XLMiner. By following XLMiner's best subset recommendations (Figure 30 through Figure 33) we did a regression with 6 attributes on the **Overall Analysis Data Set** for TPRI and with all 7 for Tejas LEE. We also did the regression for all 7 demographic attributes on the **Passing Percentage Data Set** for both TPRI and Tejas LEE.

Best subset selection

	#Coeffs	RSS	Cp	R-Squared	Adj. R-Squared	Probability	Model (Constant present in all models)								
							1	2	3	4	5	6	7	8	
Choose Subset	2	9585.260742	55.95184326	0.081331553	0.077354633	0	Constant	seholdIncome	*	*	*	*	*	*	9641.212585
Choose Subset	2	9964.84375	67.23613739	0.044951642	0.040817234	0	Constant	Married	*	*	*	*	*	*	10032.07989
Choose Subset	2	10341.28027	78.42688751	0.008873296	0.004582704	0	Constant	White	*	*	*	*	*	*	10419.70716
Choose Subset	3	8638.916016	29.81878853	0.172030916	0.164831185	0.00001479	Constant	Hispanic	seholdIncome	*	*	*	*	*	8668.734804
Choose Subset	3	8801.149414	34.64168167	0.156482178	0.149147241	0.00000215	Constant	White	Black	*	*	*	*	*	8835.791096
Choose Subset	3	8916.50293	38.07092667	0.145426492	0.137995418	0.00000055	Constant	White	Hispanic	*	*	*	*	*	8954.573856
Choose Subset	4	7653.747559	2.53157592	0.266451214	0.256841404	0.63951385	Constant	White	Hispanic	seholdIncome	*	*	*	*	7656.279135
Choose Subset	4	8690.166016	33.34235382	0.167119025	0.15620792	0.00000271	Constant	Married	White	Black	*	*	*	*	8723.508369
Choose Subset	4	8787.319336	36.23054123	0.157807677	0.14677459	0.00000083	Constant	White	Black	Hispanic	*	*	*	*	8823.549877
Choose Subset	5	7584.843262	2.48318028	0.273055124	0.260301705	0.92245203	Constant	White	Black	Hispanic	seholdIncome	*	*	*	7587.326442
Choose Subset	5	8662.046875	34.50642395	0.169814013	0.155249347	0.00000111	Constant	Married	White	Black	Asian	*	*	*	8696.553299
Choose Subset	5	8664.148438	34.56890106	0.169612595	0.155044395	0.00000108	Constant	Married	White	Black	Hispanic	*	*	*	8698.717339
Choose Subset	6	7574.22998	4.16766739	0.27407232	0.258082723	0.91961271	Constant	Married	White	Black	Hispanic	seholdIncome	*	*	7578.397648
Choose Subset	6	8634.543945	35.68881607	0.172449943	0.154221968	0.00000036	Constant	Married	Divorced	White	Black	Asian	*	*	8670.232761
Choose Subset	6	8662.025391	36.5057869	0.169816072	0.151530082	0.00000025	Constant	Married	White	Black	Asian	Hispanic	*	*	8698.531178
Choose Subset	7	7569.337402	6.02222013	0.274541233	0.255281266	0.88163608	Constant	Married	White	Black	Asian	Hispanic	seholdIncome		7575.359622
Choose Subset	7	8630.575195	37.5708313	0.172830315	0.150870058	0.00000006	Constant	Married	Divorced	White	Black	Asian	Hispanic	*	8668.146027
Choose Subset	8	7568.589844	7.99999666	0.274612881	0.252045281	1	Constant	Married	Divorced	White	Black	Asian	Hispanic	seholdIncome	7576.58984

Figure 30 - XLMiner Overall Analysis Set Best Subset Selection Table – TPRI

Best subset selection

	#Coeffs	RSS	Cp	R-Squared	Adj. R-Squared	Probability	Model (Constant present in all models)								
							1	2	3	4	5	6	7	8	
Choose Subset	2	17031.83594	36.76807404	0.120647079	0.115564115	0.00000174	Constant	Hispanic	*	*	*	*	*	*	17068.60401
Choose Subset	2	19060.25195	61.5123291	0.015920052	0.010231729	0	Constant	Married	*	*	*	*	*	*	19121.76428
Choose Subset	2	19341.80859	64.94698334	0.001383295	-0.00438906	0	Constant	Divorced	*	*	*	*	*	*	19406.75558
Choose Subset	3	15667.375	22.12327957	0.191094136	0.181688254	0.00037369	Constant	Divorced	White	*	*	*	*	*	15689.49828
Choose Subset	3	16049.68359	26.7869873	0.171355561	0.16172016	0.00006114	Constant	Hispanic	seholdIncome	*	*	*	*	*	16076.47058
Choose Subset	3	16341.11426	30.34209442	0.156309009	0.146498648	0.0000155	Constant	Black	Hispanic	*	*	*	*	*	16371.45635
Choose Subset	4	14403.47363	8.70519066	0.256349307	0.243302803	0.07367925	Constant	Black	Hispanic	seholdIncome	*	*	*	*	14412.17882
Choose Subset	4	15557.97461	22.7887249	0.196742473	0.182650236	0.0002501	Constant	Married	Divorced	White	*	*	*	*	15580.76333
Choose Subset	4	15663.93652	24.08133316	0.191271664	0.177083448	0.00014807	Constant	Divorced	White	Black	*	*	*	*	15688.01786
Choose Subset	5	13851.99902	3.97785807	0.284821916	0.267994196	0.57819569	Constant	Divorced	Black	Hispanic	seholdIncome	*	*	*	13855.97688
Choose Subset	5	15044.09766	18.5200386	0.223273917	0.204998009	0.00123934	Constant	Divorced	White	Black	Asian	*	*	*	15062.61769
Choose Subset	5	15053.53027	18.63510513	0.222786912	0.204499545	0.00117936	Constant	Divorced	Black	Asian	Hispanic	*	*	*	15072.16538
Choose Subset	6	13705.39941	4.18951797	0.292390847	0.271455665	0.90964049	Constant	Divorced	Black	Asian	Hispanic	seholdIncome	*	*	13709.58893
Choose Subset	6	14839.05957	18.01881981	0.233860024	0.211193161	0.00118997	Constant	Married	Divorced	White	Black	Asian	*	*	14857.07839
Choose Subset	6	15031.05664	20.36095428	0.223947224	0.200987083	0.00040633	Constant	Divorced	White	Black	Asian	Hispanic	*	*	15051.41759
Choose Subset	7	13700.04395	6.12418747	0.29266735	0.267405469	0.72497809	Constant	Divorced	White	Black	Asian	Hispanic	seholdIncome	*	13706.16813
Choose Subset	7	14832.82324	19.9427433	0.234182006	0.206831363	0.00025777	Constant	Married	Divorced	White	Black	Asian	Hispanic	*	14852.76599
Choose Subset	8	13689.86328	7.99999571	0.293192977	0.263566335	1	Constant	Married	Divorced	White	Black	Asian	Hispanic	seholdIncome	13697.86328

Figure 31 - XLMiner Overall Analysis Set Best Subset Selection Table - Tejas LEE

Best subset selection

	#Coeffs	RSS	Cp	R-Squared	Adj. R-Squared	Probability	Model (Constant present in all models)								
							1	2	3	4	5	6	7	8	
Choose Subset	2	17837.9922	50.8250	0.038847965	0.035048945	0	Constant	DIVORCED	*	*	*	*	*	*	17888.8172
Choose Subset	2	18496.5859	61.9687	0.00336142	-0.00057786	0	Constant	MARRIED	*	*	*	*	*	*	18558.5546
Choose Subset	3	15943.8477	20.7755	0.140908828	0.134090644	0.0005392	Constant	DIVORCED	OLD_INCOME	*	*	*	*	*	15964.6231
Choose Subset	3	16942.3203	37.6700	0.087108825	0.079863657	0.00000058	Constant	DIVORCED	BLACK	*	*	*	*	*	16979.9903
Choose Subset	3	17310.1113	43.8931	0.0672914	0.059888951	0.00000005	Constant	MARRIED	DIVORCED	*	*	*	*	*	17354.0044
Choose Subset	4	15025.8535	7.2427	0.190372463	0.18069564	0.12724844	Constant	DIVORCED	BLACK	OLD_INCOME	*	*	*	*	15033.0962
Choose Subset	4	16031.3770	24.2565	0.136192548	0.125868156	0.00011306	Constant	MARRIED	DIVORCED	WHITE	*	*	*	*	16055.6334
Choose Subset	4	16198.1123	27.0777	0.127208464	0.116776693	0.00003447	Constant	DIVORCED	BLACK	HISPANIC	*	*	*	*	16225.1900
Choose Subset	5	14801.9902	5.4548	0.202434731	0.189673687	0.32888961	Constant	DIVORCED	BLACK	HISPANIC	OLD_INCOME	*	*	*	14807.4451
Choose Subset	5	15019.7373	9.1392	0.190702018	0.177753251	0.07022202	Constant	MARRIED	DIVORCED	BLACK	HISPANIC	*	*	*	15028.8765
Choose Subset	5	15266.2266	13.3099	0.17742061	0.16425934	0.01128274	Constant	MARRIED	DIVORCED	WHITE	BLACK	*	*	*	15279.5364
Choose Subset	6	14680.3613	5.3968	0.208988376	0.193104609	0.49834433	Constant	MARRIED	DIVORCED	BLACK	HISPANIC	OLD_INCOME	*	*	14685.7582
Choose Subset	6	14930.9619	9.6371	0.195485441	0.17933053	0.06159901	Constant	MARRIED	DIVORCED	WHITE	BLACK	HISPANIC	*	*	14940.5990
Choose Subset	6	15225.5479	14.6216	0.179612473	0.163138828	0.00551304	Constant	MARRIED	DIVORCED	WHITE	BLACK	ASIAN	*	*	15240.1694
Choose Subset	7	14636.4570	6.6540	0.211354041	0.192273897	0.41947716	Constant	MARRIED	DIVORCED	WHITE	BLACK	HISPANIC	OLD_INCOME	*	14643.1110
Choose Subset	7	14915.4209	11.3741	0.196322826	0.176879024	0.02125171	Constant	MARRIED	DIVORCED	WHITE	BLACK	ASIAN	HISPANIC	*	14926.7950
Choose Subset	8	14597.8076	8.0000	0.21343656	0.191145289	1	Constant	MARRIED	DIVORCED	WHITE	BLACK	ASIAN	HISPANIC	OLD_INCOME	14605.8076

Figure 32 - XLMiner Passing Percentage Set Best Subset Selection table TPRI

Best subset selection

	#Coeffs	RSS	Cp	R-Squared	Adj. R-Squared	Probability	Model (Constant present in all models)								
							1	2	3	4	5	6	7	8	
Choose Subset	2	38832.29297	21.92109489	0.015024795	0.009331297	0.00043545	Constant	OLD_INCOME	*	*	*	*	*	*	38854.21406
Choose Subset	2	39060.3125	23.0539093	0.009241114	0.003514184	0.00028516	Constant	WHITE	*	*	*	*	*	*	39083.36641
Choose Subset	2	39423.86328	24.86005211	1.97041E-05	-0.00576053	0.00014516	Constant	MARRIED	*	*	*	*	*	*	39448.72333
Choose Subset	3	35334.76563	6.54517603	0.103739044	0.093317405	0.13494965	Constant	WHITE	OLD_INCOME	*	*	*	*	*	35341.3108
Choose Subset	3	39018.65625	24.84696007	0.010297719	-0.00121045	0.00012969	Constant	MARRIED	WHITE	*	*	*	*	*	39043.50321
Choose Subset	3	39271.00781	26.10065651	0.00389686	-0.00768573	0.00007975	Constant	MARRIED	DIVORCED	*	*	*	*	*	39297.10847
Choose Subset	4	34766.75	5.72324133	0.118146674	0.102675563	0.22582848	Constant	MARRIED	WHITE	OLD_INCOME	*	*	*	*	34772.47324
Choose Subset	4	38422.23438	23.8839016	0.025425869	0.008328077	0.0001604	Constant	MARRIED	DIVORCED	WHITE	*	*	*	*	38446.11828
Choose Subset	4	38589.11328	24.71296501	0.021193011	0.004020958	0.00011467	Constant	MARRIED	WHITE	BLACK	*	*	*	*	38613.82625
Choose Subset	5	34366.64453	5.73549414	0.128295288	0.107784589	0.29502699	Constant	MARRIED	WHITE	ASIAN	OLD_INCOME	*	*	*	34372.38003
Choose Subset	5	38216.61328	24.86236382	0.030641417	0.007832979	0.00008238	Constant	MARRIED	DIVORCED	WHITE	BLACK	*	*	*	38241.47565
Choose Subset	5	38571.20703	26.62400627	0.0216472	-0.00137287	0.00003918	Constant	MARRIED	WHITE	BLACK	ASIAN	*	*	*	38597.83104
Choose Subset	6	33701.50391	4.43103647	0.145166479	0.119875547	0.80634475	Constant	MARRIED	WHITE	BLACK	ASIAN	OLD_INCOME	*	*	33705.93494
Choose Subset	6	38199.17578	26.77573204	0.031083716	0.002417554	0.00002293	Constant	MARRIED	DIVORCED	WHITE	BLACK	ASIAN	*	*	38225.95151
Choose Subset	6	38491.65625	28.22879219	0.023664994	-0.00522066	0.00001212	Constant	MARRIED	WHITE	BLACK	ASIAN	HISPANIC	*	*	38519.88504
Choose Subset	7	33619.30859	6.02268553	0.147251351	0.116796042	0.88045859	Constant	MARRIED	WHITE	BLACK	ASIAN	HISPANIC	OLD_INCOME	*	33625.33128
Choose Subset	7	38040.67578	27.9882946	0.035104045	0.000643475	0.00000565	Constant	MARRIED	DIVORCED	WHITE	BLACK	ASIAN	HISPANIC	*	38068.66408
Choose Subset	8	33614.74219	7.99999952	0.147367177	0.111628077	1	Constant	MARRIED	DIVORCED	WHITE	BLACK	ASIAN	HISPANIC	OLD_INCOME	33622.74219

Figure 33 - XLMiner Passing Percentage Set Best Subset Selection table Tejas LEE

By looking at the lift charts returned for both sets we see that for all of them both the baseline line and the lift curve are pretty much the same (Figures 34 through Figure 37). This means that the area between them (what would define the lift number) is pretty much 0 and hence the multiple line regression model is not very good for predicting them either.

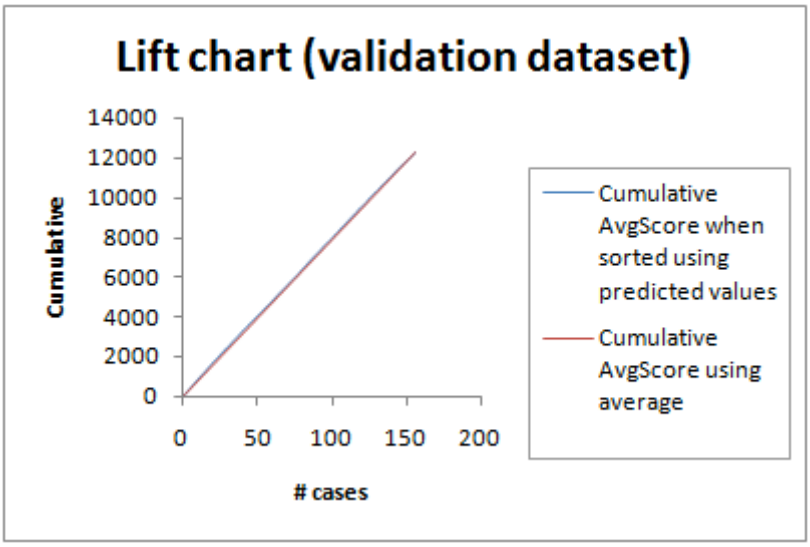

Figure 34 - Overall Analysis MLR Lift Chart – TPRI

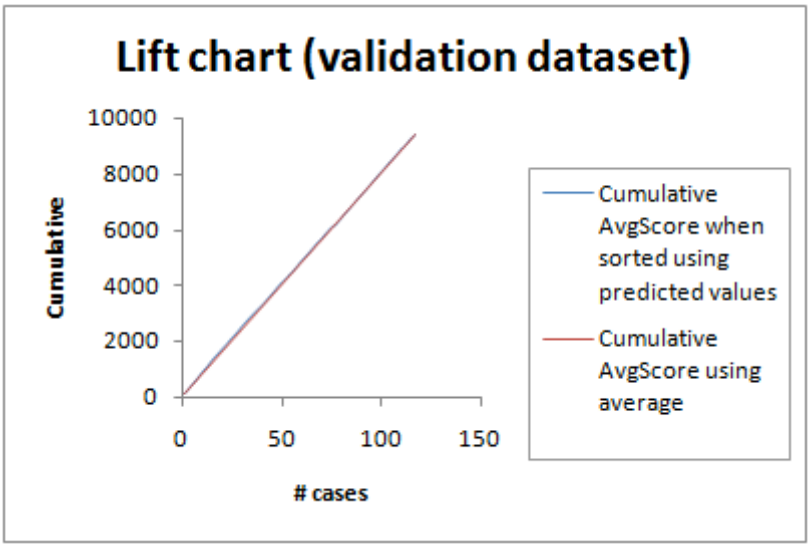

Figure 35 - Overall Analysis MLR Lift Chart - Tejas LEE

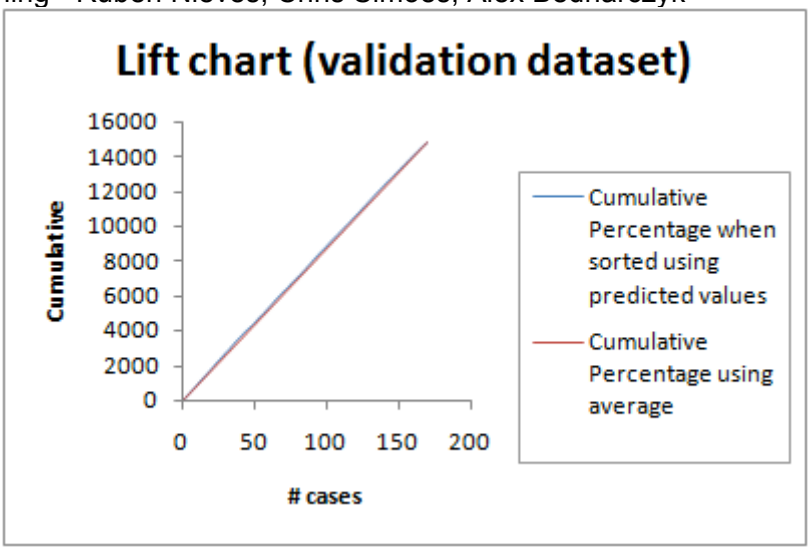

Figure 36 - Passing Percentage MLR Lift Chart – TPRI

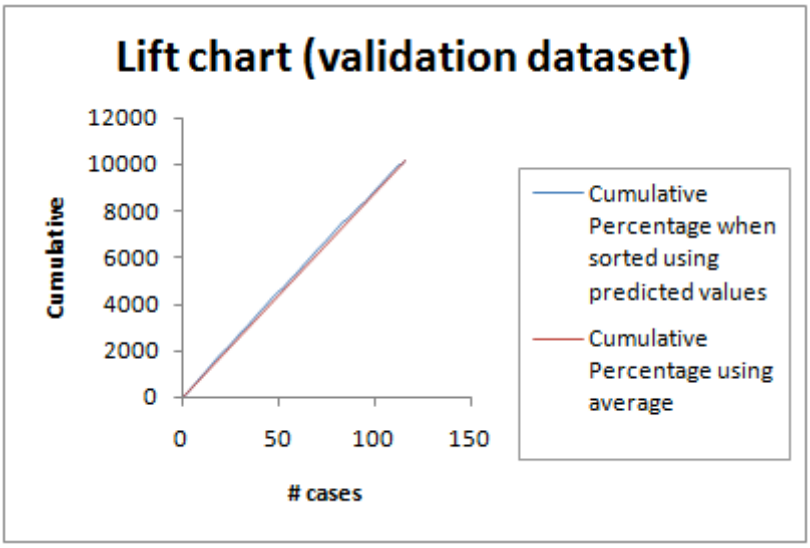

Figure 37 - Passing Percentage MLR Lift Chart - Tejas LEE

Neither of those techniques showed any indication of a trend so we decided to do a couple of multidimensional plots using a subset of the on the **Raw Student Info Data Set**. Because of the limits of XLMiner, the data was partitioned creating a subset of 10,000 records. PCA was performed on the partition set and is shown in Figure 34 below. This confirmed that most of the variance is on the zip code as scores change from one to the other.

Principal Components

	Components										
Variable	1	2	3	4	5	6	7	8	9	10	11
SchoolZip	-0.99936086	0.03517504	0.0063388	0.00052613	-0.00026825	0.00015535	-0.00001526	0.00002742	-0.00001328	-0.00001845	-0.00000292
TestLang	-0.00003243	-0.00227096	0.00395814	0.0019031	-0.00432891	0.00443124	0.02012816	0.06189891	0.00904359	-0.9967916	0.04500339
StudentGradeL	-0.00009735	-0.00156596	-0.00236683	-0.00763314	0.00466995	0.02396166	-0.03593626	-0.99706429	0.00022212	-0.06239085	0.00404901
Gender	0.00000892	-0.00005659	-0.00020856	-0.00036909	0.00101606	-0.00188014	0.02734818	-0.00070298	-0.99957716	-0.0083609	0.0047303
GradeLetter	0.00004924	-0.00060763	0.00536513	0.00614103	0.05142868	0.19590683	-0.97785407	0.04118061	-0.02694673	-0.01659019	0.00356177
married	0.00015541	0.04759983	-0.16607475	-0.95813692	-0.01209708	0.22438478	0.03784584	0.0116396	0.00101372	0.00030812	0.00795332
divorced	0.00125618	0.02053629	0.04852388	0.2203941	0.12685451	0.94342923	0.19779924	0.01380269	0.00382757	0.01158257	0.05508134
WHITE	0.02160883	0.69343966	-0.43507564	0.09185249	0.27269247	-0.07919506	-0.0023625	-0.00040121	0.0025524	0.01739939	0.48989671
BLACK	0.00468361	0.00018964	0.7683621	-0.15335499	0.37549707	-0.08446558	0.00736287	-0.00063992	0.00275298	0.02288155	0.48719773
ASIAN	0.00113684	0.01040439	0.08289823	0.01834187	-0.83731347	0.0779109	-0.02645302	-0.00102424	0.00054431	0.02779244	0.53295803
HISPANIC	-0.02803664	-0.71770495	-0.42828843	0.03176414	0.25420809	-0.03375793	0.00634537	0.00271038	0.00272335	0.02087591	0.48313707
Variance	644475.9375	976.6658936	148.0648346	42.46942139	8.06528091	2.55486965	2.28152084	0.88553602	0.24817285	0.15749992	0.04256021
Variance%	99.8170166	0.1512669	0.02293242	0.0065777	0.00124916	0.0003957	0.00035336	0.00013715	0.00003844	0.00002439	0.00000659
Cum%	99.8170166	99.96828461	99.99121857	99.9977951	99.99904633	99.99943542	99.99979401	99.99993134	99.99996948	99.99999237	100

Figure 38 - PCA for Student Info Data Set

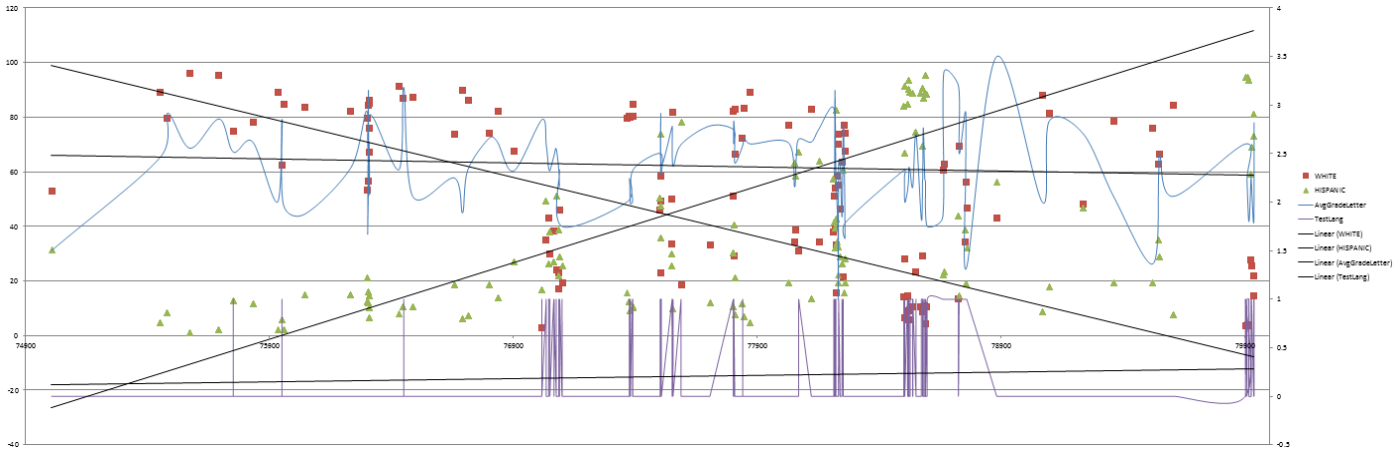

Figure 39 - Multidimensional Plot for Student Info Data Set

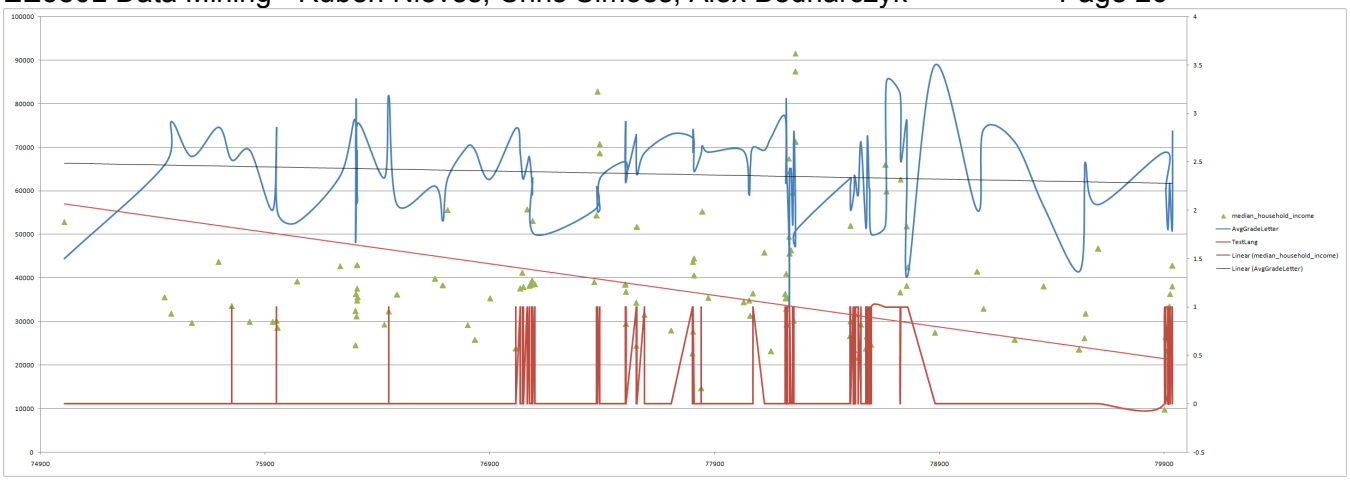

Figure 40 - Multidimensional Plot for Student Info Data Set 2

On Figure 35 we plotted zip code for the X-axis, percentage of population for the left Y-axis and grade for the right Y-axis. The legend includes components from the PCA analysis shown in Figure 34. The components include White and Hispanic population, as these are the most predominant in Texas, grade, and the language of the exam given. For the grade letter specifically we changed the scale and a lower number is better than a higher number. This means the lower peaks are the better scores. For the language of the exam, a value of 1 (peak) means it is the Tejas LEE version of the test and a 0 means it is TPRI. Each of these components is plotted against the zip code (X-axis). Looking at this figure, we did notice something really interesting that we did not notice before. It seemed that on most of the higher peaks in average grade letter (lower performance) the population percentages where mostly White (notice the left side of the graph). Based off of this we could infer that a higher Hispanic population would account for better scores. However, as we move along the zip codes it evens out and thus the results we had in all of our previous tests came to fruition. At an overall level it didn't seem to be true that a higher Hispanic population implied a better grade percentage.

We also created another plot where the zip code is used for the X-axis, the median household income on the left Y-axis, and average grade for each zip along the right Y-axis (Figure 36). At first look, based on the trend lines plotted against the information, it appeared that the average grade improved as median household income dropped. Once again, this didn't seemed to show up on our previous analysis of the data so we decided to take a closer look at it. The first thing we did notice was that most of the points on household income were on the lower end hence it may have been misleading to the naked eye at first look. To investigate this further we went back to the household income vs grade average plots done initially (Figure 8, 9, 14 and 15) and noticed that indeed we had a lot more points on the lower end of the household income spectrum, hence our numbers where skewed towards that side. Our detailed analysis techniques helped mitigate this situation and because of that it became clear that household income didn't really influence outcome either.

After all of that work, none of the processes we went through helped us derive any kind of significant conclusion. As a final result we decided to take a more abstract look at the **Overall Analysis Data Set** since the aggregated data was much smaller compared to the raw data sets

(425 and 304 records for TPRI and TejasLEE respectively). Doing an analysis on the TPRI data by looking at the top schools revealed that most zip codes just had a majority population of Hispanics or Whites and it was pretty evenly distributed. This was expected because the Hispanic and White communities are the biggest in Texas (Census 2010). When observing the same data set for TejasLEE the majority of the top score schools are on areas where the population is mostly Hispanic. This may be misleading at first view as you may think that the percentage of Hispanics living on the area has some influence but when confronted with all of the analysis done before we took a closer look and realized this just seemed like a trend because Tejas LEE in itself is mostly given on areas where the population is mostly Hispanic. That made sense since the TejasLEE assessment is for the Spanish version of the TPRI assessment.

In conclusion, after trying all of these techniques and analysis tools we can safely say that none of the demographic features based on the zip code used are good predictors on performance.

Conclusion

Our data mining research helped us clarify some questions and opened up many other questions for future investigation. We confirmed that in overall, girls do outperform boys in school. We also discovered that at least a lower grade levels the younger kids outperform the older ones.

We also confirmed that for the most part the type of test analyzed is not of importance to an analysis on schools performance for a certain subject. This opens up the question to investigate if the same is true across tests of different subjects.

Finally we discovered that a schools location and the demographic attributes on the location itself do not have an impact on the performance of the students. There can be many others attributes that may do but the scope of our research doesn't cover them but, ultimately this is great news for students because it means that at least these external attributes that most of the time a student doesn't have power over do not determine how well they will do.

APPENDIX A

Queries to analyze girls vs boys performance

OVERALL PERFORMANCE

**********TPRI**********

A: Total number of girls?

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'F' and Assessments_Description LIKE '%TPRI%'

44967 Girls

B: How many girls got a passing Grade (C or better) in the year 2011 FOR TPRI?

```
Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity,
St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'F' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND St.Assessments_Description LIKE '%TPRI%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName,
St.StudentGradeLevel
```

40206 Girls

89.4% of Girls Passed TPRI

C: Total number of boys

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'M' and Assessments_Description LIKE '%TPRI%'

47512 Boys

D: How many boys got a passing Grade (C or better) in the year 2011 FOR TPRI?

```
Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity,
St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'M' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND St.Assessments_Description LIKE '%TPRI%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName,
St.StudentGradeLevel
```

40499 Boys

85.2% of Boys Passed TPRI

****

**********Tejas LEE**********

A: Total number of girls?

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'F' and Assessments_Description LIKE '%Tejas%'

11741 Girls

B: How many girls got a passing Grade (C or better) in the year 2011 FOR TPRI?

```
Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity,
```

St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'F' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND St.Assessments_Description LIKE '%Tejas%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel

10817 Girls

92.1% of Girls Passed Tejas LEE

C: Total number of boys

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'M' and Assessments_Description LIKE '%Tejas%'

12380 Boys

D: How many boys got a passing Grade (C or better) in the year 2011 FOR TPRI?

Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'M' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND St.Assessments_Description LIKE '%Tejas%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel

10941 Boys

88.4% of Boys Passed Tejas LEE

****

ANALYSIS PER GRADE (Kindergarten to 3rd Grade)

----------KinderGarten----------

**********TPRI**********

A: Total number of girls?

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'F' and Assessments_Description LIKE '%TPRI%' and StudentGradeLevel = 'Kindergarten'

12178 Girls

B: How many girls got a passing Grade (C or better) in the year 2011 for TPRI?

Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'F' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C') AND StudentGradeLevel = 'Kindergarten' AND St.Assessments_Description LIKE '%TPRI%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel

11254 Girls

92.4% of Girls Passed TPRI

C: Total number of boys

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'M' and Assessments_Description LIKE '%TPRI%' and StudentGradeLevel = 'Kindergarten'

12986 Boys

D: How many boys got a passing Grade (C or better) in the year 2011 for TPRI?

Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'M' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C') AND StudentGradeLevel = 'Kindergarten' AND St.Assessments_Description LIKE '%TPRI%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel

11476 Boys

88.4% of Boys Passed TPRI

****

**********Tejas LEE**********

A: Total number of girls?

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'F' and Assessments_Description LIKE '%Tejas%' and StudentGradeLevel = 'Kindergarten'

3963 Girls

B: How many girls got a passing Grade (C or better) in the year 2011 for Tejas LEE?

```
Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity,
St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'F' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = 'Kindergarten' AND St.Assessments_Description LIKE
'%Tejas%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName,
St.StudentGradeLevel
```

3687 Girls

93.0% of Girls Passed Tejas LEE

C: Total number of boys

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'M' and Assessments_Description LIKE '%Tejas%' and StudentGradeLevel = 'Kindergarten'

4282 Boys

D: How many boys got a passing Grade (C or better) in the year 2011 for Tejas LEE?

```
Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity,
St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'M' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = 'Kindergarten' AND St.Assessments_Description LIKE
'%Tejas%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName,
St.StudentGradeLevel
```

3842 Boys

89.7% of Boys Passed Tejas LEE

****

----------1st Grade----------

**********TPRI**********

A: Total number of girls?

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'F' and Assessments_Description LIKE '%TPRI%' and StudentGradeLevel = '1st Grade'

12966 Girls

B: How many girls got a passing Grade (C or better) in the year 2011 for TPRI?

```
Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity,
St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'F' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = '1st Grade' AND St.Assessments_Description LIKE '%TPRI%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName,
St.StudentGradeLevel
```

11229 Girls

86.6% of Girls Passed TPRI

C: Total number of boys

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'M' and Assessments_Description LIKE '%TPRI%' and StudentGradeLevel = '1st Grade'

13829 Boys

D: How many boys got a passing Grade (C or better) in the year 2011 for TPRI?

```
Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity,
St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'M' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = '1st Grade' AND St.Assessments_Description LIKE '%TPRI%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName,
St.StudentGradeLevel
```

11337 Boys

82.0% of Boys Passed TPRI

****

**********Tejas LEE**********

A: Total number of girls?

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'F' and Assessments_Description LIKE '%Tejas%' and StudentGradeLevel = '1st Grade'

3637 Girls

B: How many girls got a passing Grade (C or better) in the year 2011 for Tejas LEE?

Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'F' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = '1st Grade' AND St.Assessments_Description LIKE '%Tejas%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel

3394 Girls

93.3% of Girls Passed Tejas LEE

C: Total number of boys

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'M' and Assessments_Description LIKE '%Tejas%' and StudentGradeLevel = '1st Grade'

3722 Boys

D: How many boys got a passing Grade (C or better) in the year 2011 for Tejas LEE?

Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'M' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = '1st Grade' AND St.Assessments_Description LIKE '%Tejas%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel

3292 Boys

88.4% of Boys Passed Tejas LEE

****

----------2nd Grade----------

**********TPRI**********

A: Total number of girls?

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'F' and Assessments_Description LIKE '%TPRI%' and StudentGradeLevel = '2nd Grade'

15509 Girls

B: How many girls got a passing Grade (C or better) in the year 2011 for TPRI?

```
Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity,
St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'F' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = '2nd Grade' AND St.Assessments_Description LIKE '%TPRI%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName,
St.StudentGradeLevel
```

13523 Girls

87.2% of Girls Passed TPRI

C: Total number of boys

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'M' and Assessments_Description LIKE '%TPRI%' and StudentGradeLevel = '2nd Grade'

16317 Boys

D: How many boys got a passing Grade (C or better) in the year 2011 for TPRI?

```
Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity,
St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'M' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = '2nd Grade' AND St.Assessments_Description LIKE '%TPRI%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName,
St.StudentGradeLevel
```

13484 Boys

82.6% of Boys Passed TPRI

****

**********Tejas LEE**********

A: Total number of girls?

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'F' and Assessments_Description LIKE '%Tejas%' and StudentGradeLevel = '2nd Grade'

3684 Girls

B: How many girls got a passing Grade (C or better) in the year 2011 for Tejas LEE?

```
Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity,
St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'F' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = '2nd Grade' AND St.Assessments_Description LIKE '%Tejas%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName,
St.StudentGradeLevel
```

3425 Girls

93% of Girls Passed Tejas LEE

C: Total number of boys

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'M' and Assessments_Description LIKE '%Tejas%' and StudentGradeLevel = '2nd Grade'

3896 Boys

D: How many boys got a passing Grade (C or better) in the year 2011 for Tejas LEE?

```
Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity,
St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'M' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = '2nd Grade' AND St.Assessments_Description LIKE '%Tejas%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName,
St.StudentGradeLevel
```

3513 Boys

90.2% of Boys Passed Tejas LEE

****

----------3rd Grade----------

**********TPRI**********

A: Total number of girls?

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'F' and Assessments_Description LIKE '%TPRI%' and StudentGradeLevel = '3rd Grade'

4315 Girls

B: How many girls got a passing Grade (C or better) in the year 2011 for TPRI?

Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'F' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = '3rd Grade' AND St.Assessments_Description LIKE '%TPRI%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel

4200 Girls

97.3% of Girls Passed TPRI

C: Total number of boys

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'M' and Assessments_Description LIKE '%TPRI%' and StudentGradeLevel = '3rd Grade'

4376 Boys

D: How many boys got a passing Grade (C or better) in the year 2011 for TPRI?

Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'M' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = '3rd Grade' AND St.Assessments_Description LIKE '%TPRI%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel

4200 Boys

96% of Boys Passed TPRI

****

**********Tejas LEE**********

A: Total number of girls?

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'F' and Assessments_Description LIKE '%Tejas%' and StudentGradeLevel = '3rd Grade'

458 Girls

B: How many girls got a passing Grade (C or better) in the year 2011 for Tejas LEE?

Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'F' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = '3rd Grade' AND St.Assessments_Description LIKE '%Tejas%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel

311 Girls

67.9% of Girls Passed Tejas LEE

C: Total number of boys

Select Distinct StudentKey From StudentsInfoDetailed Where Gender = 'M' and Assessments_Description LIKE '%Tejas%' and StudentGradeLevel = '3rd Grade'

477 Boys

D: How many boys got a passing Grade (C or better) in the year 2011 for Tejas LEE?

Select Distinct AVG(St.ScoreAsPercentage) AS finalScore, St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zip ON St.SchoolZip = Zip.ZipCode
Where Gender = 'M' and (GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND StudentGradeLevel = '3rd Grade' AND St.Assessments_Description LIKE '%Tejas%'
Group by St.CountyName, St.SchoolCity, St.SchoolState, St.StudentKey, St.StudentName, St.StudentGradeLevel

292 Boys

61.2% of Boys Passed Tejas LEE

****

APPENDIX B

Average Score per School Queries

**********TPRI**********

```
Select AVG(St.ScoreAsPercentage) AS AvgScore, St.SchoolName, St.SchoolCity,
St.SchoolZip,
Zp.MARRIED, Zp.DIVORCED, Zp.WHITE, Zp.BLACK, Zp.ASIAN, Zp.Hispanic,
Zp.MEDIAN_HOUSEHOLD_INCOME
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zp ON St.SchoolZip = Zp.ZipCode
WHERE St.Assessments_Description LIKE '%TPRI%'
GROUP BY
St.SchoolName, St.SchoolCity, St.SchoolZip,
Zp.MARRIED, Zp.DIVORCED, Zp.WHITE, Zp.BLACK, Zp.ASIAN, Zp.Hispanic,
Zp.MEDIAN_HOUSEHOLD_INCOME
ORDER BY AvgScore DESC
```

**********Tejas LEE**********

```
Select AVG(St.ScoreAsPercentage) AS AvgScore, St.SchoolName, St.SchoolCity,
St.SchoolZip,
Zp.MARRIED, Zp.DIVORCED, Zp.WHITE, Zp.BLACK, Zp.ASIAN, Zp.Hispanic,
Zp.MEDIAN_HOUSEHOLD_INCOME
From StudentsInfoDetailed AS St
INNER JOIN ZipCodeInfo as Zp ON St.SchoolZip = Zp.ZipCode
WHERE St.Assessments_Description LIKE '%Tejas%'
GROUP BY
St.SchoolName, St.SchoolCity, St.SchoolZip,
Zp.MARRIED, Zp.DIVORCED, Zp.WHITE, Zp.BLACK, Zp.ASIAN, Zp.Hispanic,
Zp.MEDIAN_HOUSEHOLD_INCOME
ORDER BY AvgScore DESC
```

Passed Percentages per School Queries

**********TPRI**********

```
Select Query1.NumberOfStudentsPassed, Query2.NumberOfStudents,
CAST((Query1.NumberOfStudentsPassed)AS FLOAT)/Query2.NumberOfStudents * 100
AS Percentage, Query1.SchoolName,
Query3.Zipcode, Query3.MARRIED, Query3.DIVORCED, Query3.WHITE, Query3.BLACK,
Query3.ASIAN, Query3.Hispanic, Query3.MEDIAN_HOUSEHOLD_INCOME
FROM
(Select COUNT(Distinct StudentKey) AS NumberOfStudentsPassed, SchoolName,
SchoolZip
FROM DataMiningClass.dbo.StudentsInfoDetailed
Where(GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND Assessments_Description LIKE '%TPRI%'
Group BY SchoolName, SchoolZip) AS Query1
INNER JOIN
(Select COUNT(Distinct StudentKey) As NumberOfStudents, SchoolName, SchoolZip
From DataMiningClass.dbo.StudentsInfoDetailed
Where Assessments_Description LIKE '%TPRI%'
Group BY SchoolName, SchoolZip) AS Query2
ON Query1.SchoolName = Query2.SchoolName
AND Query1.SchoolZip = Query2.SchoolZip
INNER JOIN
(Select Zipcode, MARRIED, DIVORCED, WHITE, BLACK, ASIAN, Hispanic,
MEDIAN_HOUSEHOLD_INCOME
FROM DataMiningClass.dbo.ZipCodeInfo) AS Query3
ON Query3.ZipCode = Query2.SchoolZip
ORDER BY Percentage DESC
```

**********Tejas LEE**********

```
Select Query1.NumberOfStudentsPassed, Query2.NumberOfStudents,
CAST((Query1.NumberOfStudentsPassed)AS FLOAT)/Query2.NumberOfStudents * 100
AS Percentage, Query1.SchoolName,
Query3.Zipcode, Query3.MARRIED, Query3.DIVORCED, Query3.WHITE, Query3.BLACK,
Query3.ASIAN, Query3.Hispanic, Query3.MEDIAN_HOUSEHOLD_INCOME
FROM
(Select COUNT(Distinct StudentKey) AS NumberOfStudentsPassed, SchoolName,
SchoolZip
FROM DataMiningClass.dbo.StudentsInfoDetailed
Where(GradeLetter = 'A' OR GradeLetter = 'B' OR GradeLetter = 'C')
AND Assessments_Description LIKE '%Tejas%'
Group BY SchoolName, SchoolZip) AS Query1
INNER JOIN
(Select COUNT(Distinct StudentKey) As NumberOfStudents, SchoolName, SchoolZip
From DataMiningClass.dbo.StudentsInfoDetailed
Where Assessments_Description LIKE '%Tejas%'
Group BY SchoolName, SchoolZip) AS Query2
ON Query1.SchoolName = Query2.SchoolName
AND Query1.SchoolZip = Query2.SchoolZip
INNER JOIN
(Select Zipcode, MARRIED, DIVORCED, WHITE, BLACK, ASIAN, Hispanic,
MEDIAN_HOUSEHOLD_INCOME
FROM DataMiningClass.dbo.ZipCodeInfo) AS Query3
ON Query3.ZipCode = Query2.SchoolZip
ORDER BY Percentage DESC
```