FINAL PROJECT: CRISP DM

PREDICTING STUDENT GRADES

BY TEAM QUAD SQUAD

- ★ **BRIAN BUMPERS**
- ★ **AMER IMSIC**
- ★ **ASHRITH MADAN**
- ★ **SONIKA SHIVANI VIJAYKUMAR**

TABLE OF CONTENTS

1. Data Collection Report:

The data for this project was acquired from www.kaggle.com

The link to data is https://www.kaggle.com/daviddraper1518/predicting-student-grades , which is a public domain database.

No other data sources were used for this project.

Context:
These are the previous GCSE examination grades for the last four years at an English Underperforming School. The idea is to find a ML model that would predict future grades for students based on their profile (SEN,PP,HML,Reading, Writing and Maths Grades).

Idea is to predict the following grades (1-9 whole grades) Att8Act = Attainment 8 Actual (Best 8 grades for students) EngAct = English grade MatAct = Maths grade

PP = Pupil Premium EAL = English as Additional Language HML = Ability based on previous results ages 10/11. H= Higher, M = Middle, L = Lower Re = Reading grade at aged 10 Wr = Writing grade at aged 10 Ma = Maths grade at aged 10

2. Data Description Report:

Variable	N	N Miss	Mean	Std Dev	Std Error	Variance	Minimum	Maximum	Mode	Range
PPID	1037	0	500.3789778	294.7119365	9.151839	86855.13	1	1037	881	1036
PP	1034	3	0.3452611	0.4756834	0.014793	0.2262747	0	1	0	1
EAL	1037	0	0.1031823	0.3043436	0.009451	0.092625	0	1	0	1
SEN	1033	4	0.2400774	0.4273372	0.013296	0.182617	0	1	0	1
Att8Est	987	50	4.2616008	1.274717	0.040575	1.6249033	1.5	7.7	3.9	6.2
Att8Act	1036	1	3.6611004	1.6645965	0.051717	2.7708815	0	8.5	3.7	8.5
Att8Diff	987	50	-0.6316819	1.3230294	0.042113	1.7504069	-6.81	4.04	-1.02	10.85
EngEst	987	50	4.6200608	1.1601929	0.036929	1.3460475	1.9	7.6	4.3	5.7
EngAct	1036	1	4.0793436	1.8231305	0.056642	3.3238048	0	9	4	9
EngDiff	987	50	-0.5500608	1.5910034	0.050642	2.5312917	-6.94	4.67	-1.37	11.61
MathsEst	987	50	4.1037487	1.4074332	0.044799	1.9808683	1.1	7.9	4.3	6.8
MathsAct	1036	1	3.7007722	1.920834	0.059677	3.6896033	0	9	4	9
MathsDiff	987	50	-0.4459372	1.4652898	0.046641	2.1470742	-7.01	4.8	0.19	11.81
EbaccEst	987	50	3.7743668	1.482994	0.047204	2.1992713	0.7	7.9	3.1	7.2
EbaccAct	1036	1	2.669112	1.9409624	0.060303	3.7673349	0	9	0	9
EbaccDiff	987	50	-1.1203242	1.6119435	0.051309	2.598362	-6.65	4.68	-1.16	11.33

Variable	Sum	1st Pctl	5th Pctl	10th Pctl	Lower Quartile	Median	Upper Quartile	90th Pctl	95th Pctl	99th Pctl
PPID	518893	10	42	85	259	498	747	914	967	1029
PP	357	0	0	0	0	0	1	1	1	1
EAL	107	0	0	0	0	0	0	1	1	1
SEN	248	0	0	0	0	0	0	1	1	1
Att8Est	4206.2	1.6	2.2	2.7	3.4	4.1	5.1	6.1	6.4	7.2
Att8Act	3792.9	0.1	0.8	1.5	2.5	3.65	4.7	5.8	6.6	7.5
Att8Diff	-623.47	-3.86	-2.86	-2.32	-1.46	-0.62	0.21	1.03	1.51	2.62
EngEst	4560	2.1	2.7	3.1	3.9	4.6	5.4	6.2	6.6	7.2
EngAct	4226.2	0	1	2	3	4	5.5	7	7	8.5
EngDiff	-542.91	-4.63	-3.16	-2.5	-1.52	-0.57	0.47	1.51	2.14	3.14
MathsEst	4050.4	1.1	1.9	2.3	3.1	4.1	5	6.2	6.6	7.4
MathsAct	3834	0	0	1	3	4	5	6	7	8.5
MathsDiff	-440.14	-4.35	-2.77	-2.13	-1.28	-0.53	0.42	1.44	2.09	3.46
EbaccEst	3725.3	0.7	1.4	1.9	2.8	3.7	4.8	6	6.3	7.1
EbaccAct	2765.2	0	0	0.3	1	2.3	4	5.5	6.3	7.7
EbaccDiff	-1105.76	-4.88	-3.63	-3.03	-2.15	-1.16	-0.14	0.86	1.59	3.19

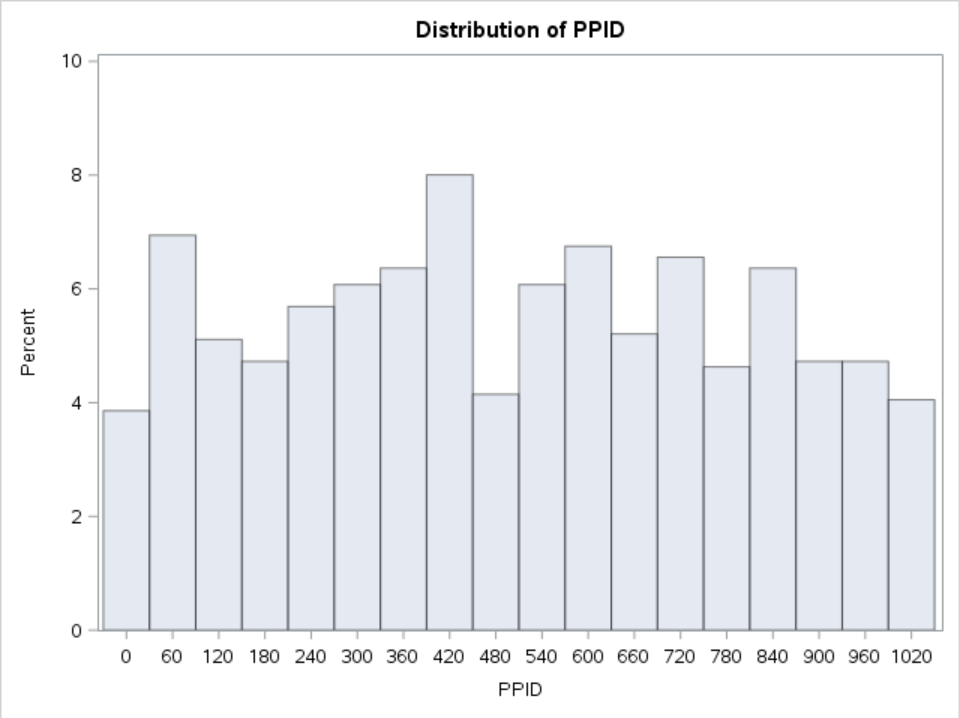
Correlation Matrix:

	PPID	PP	EAL	SEN	Att8Est	Att8Act	Att8Diff	EngEst	EngAct	EngDiff	MathsEst	MathsAct	MathsDiff	EbaccEst	EbaccAct	EbaccDiff
PPID	1															
PP	0.00053	1														
EAL	0.038241	0.000381	1													
SEN	0.025255	0.119679	-0.10181	1												
Att8Est	-0.0184	-0.15235	-0.03496	-0.22095	1											
Att8Act	0.010437	-0.24566	0.144429	-0.28526	0.60719	1										
Att8Diff	0.017333	-0.17431	0.200803	-0.13448	-0.1885	0.662576	1									
EngEst	-0.02022	-0.15367	-0.03756	-0.22925	0.997536	0.607114	-0.18694	1								
EngAct	-0.01675	-0.21595	0.06616	-0.25834	0.513751	0.861427	0.579396	0.512418	1							
EngDiff	-0.01563	-0.15046	0.100768	-0.12869	-0.13874	0.549	0.804817	-0.14162	0.777374	1						
MathsEst	-0.01903	-0.15063	-0.0353	-0.22691	0.996693	0.609872	-0.18277	0.99548	0.513804	-0.13702	1					
MathsAct	0.0069	-0.21333	0.167711	-0.25196	0.642458	0.87853	0.482602	0.643705	0.644985	0.277721	0.648243	1				
MathsDiff	0.009997	-0.15957	0.209874	-0.10289	-0.12234	0.562562	0.806046	-0.11968	0.3534	0.494606	-0.11732	0.67999	1			
EbaccEst	-0.01515	-0.14709	-0.02495	-0.20315	0.994005	0.601724	-0.189	0.989228	0.509837	-0.13727	0.985972	0.631457	-0.12641	1		
EbaccAct	0.027201	-0.19093	0.163172	-0.23374	0.572635	0.912265	0.594688	0.567961	0.716331	0.413253	0.56521	0.784819	0.492201	0.581889	1	
EbaccDiff	0.031494	-0.10343	0.203662	-0.08201	-0.22863	0.543086	0.88643	-0.22997	0.393938	0.622307	-0.23023	0.370425	0.705234	-0.22266	0.663078	1

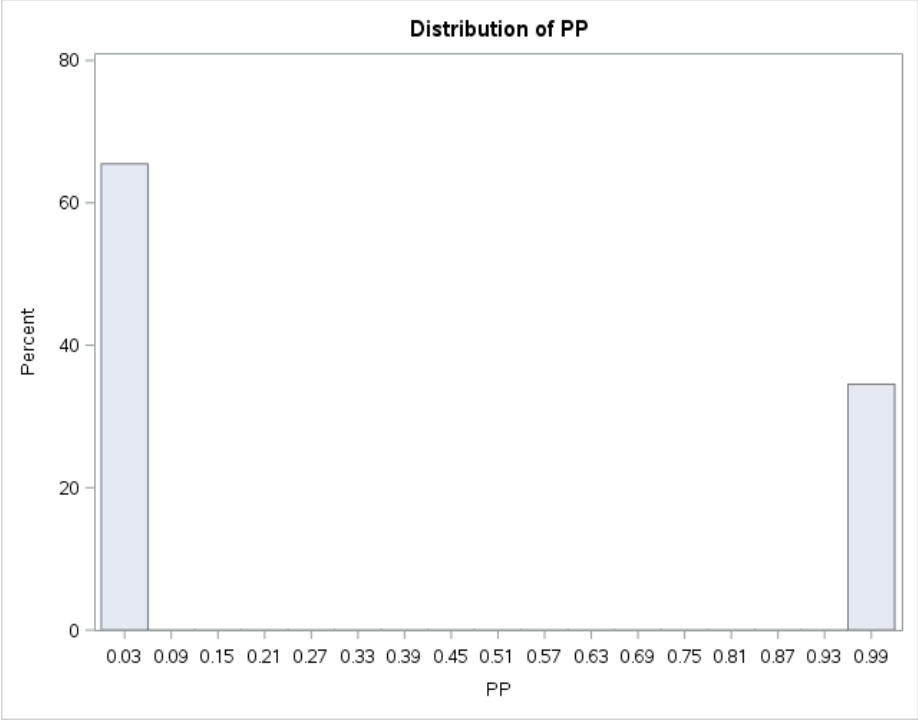
3. Data Exploration Report

During this process, we detailed the relationships between each variable/feature with the dependent variable/label. The histograms were found as below:

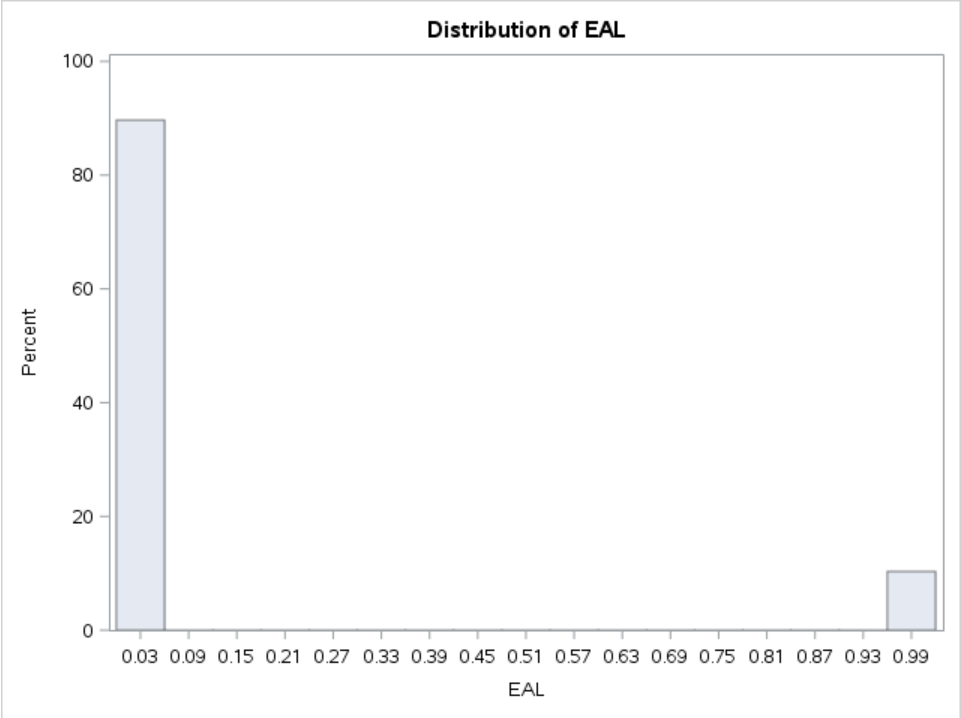
1. PPID


Distribution of PPID
Percent
10
8
6
4
2
0
0 60 120 180 240 300 360 420 480 540 600 660 720 780 840 900 960 1020
PPID

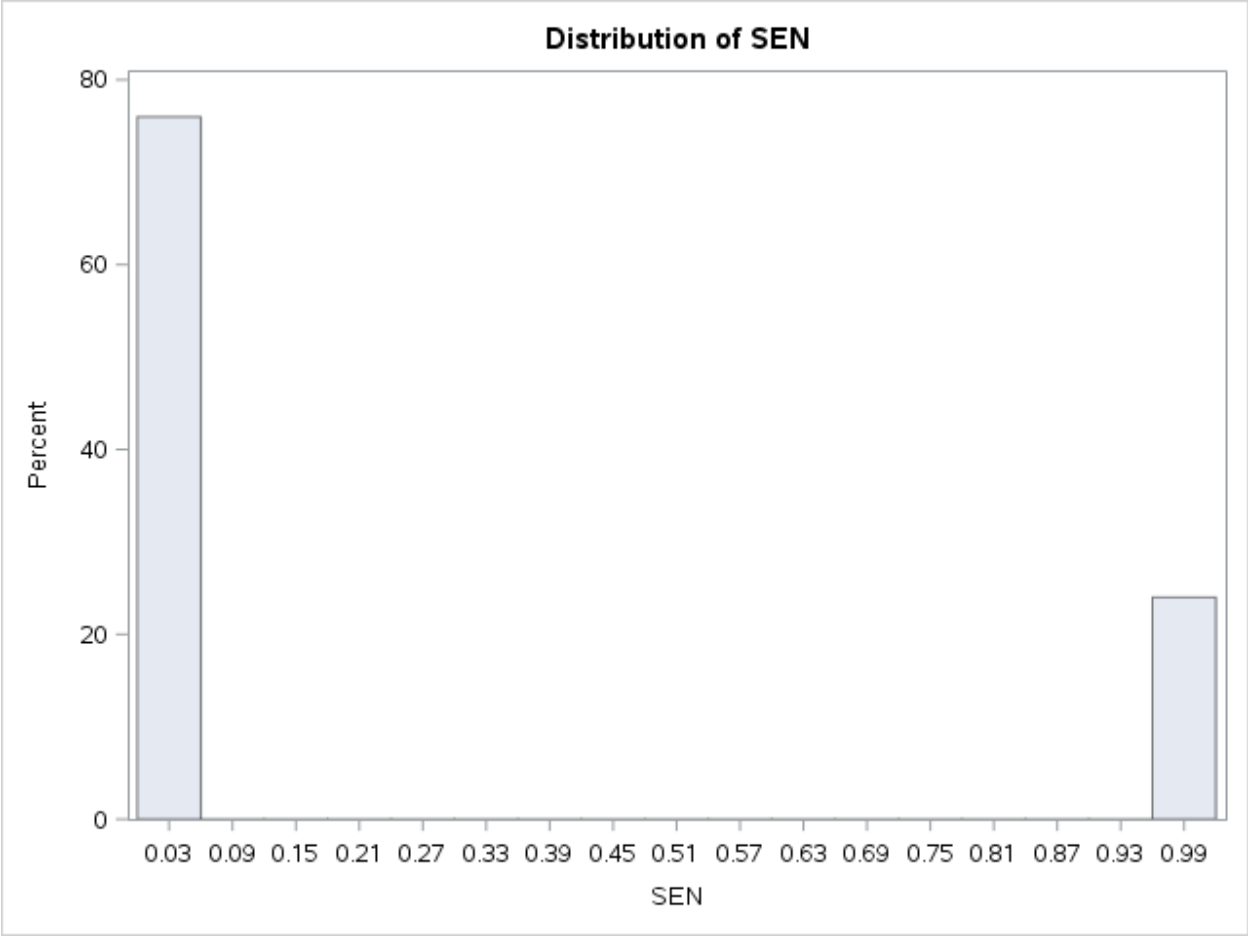
2. PP


Distribution of PP
Percent
80
60
40
20
0
0.03 0.09 0.15 0.21 0.27 0.33 0.39 0.45 0.51 0.57 0.63 0.69 0.75 0.81 0.87 0.93 0.99
PP

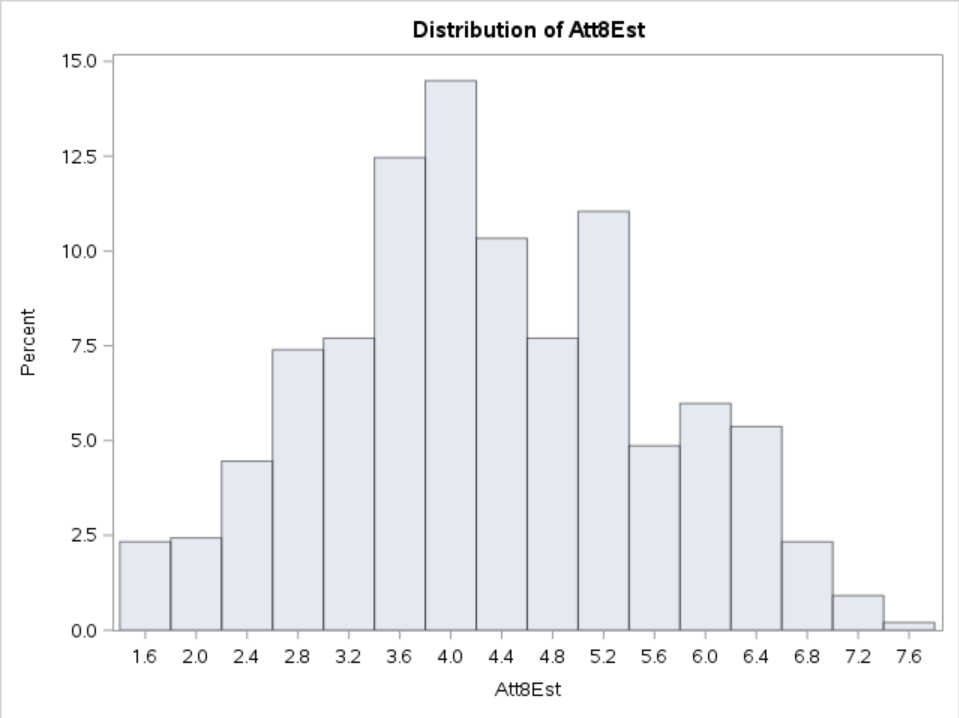
3. EAL


Distribution of EAL
Percent
100
80
60
40
20
0
0.03 0.09 0.15 0.21 0.27 0.33 0.39 0.45 0.51 0.57 0.63 0.69 0.75 0.81 0.87 0.93 0.99
EAL

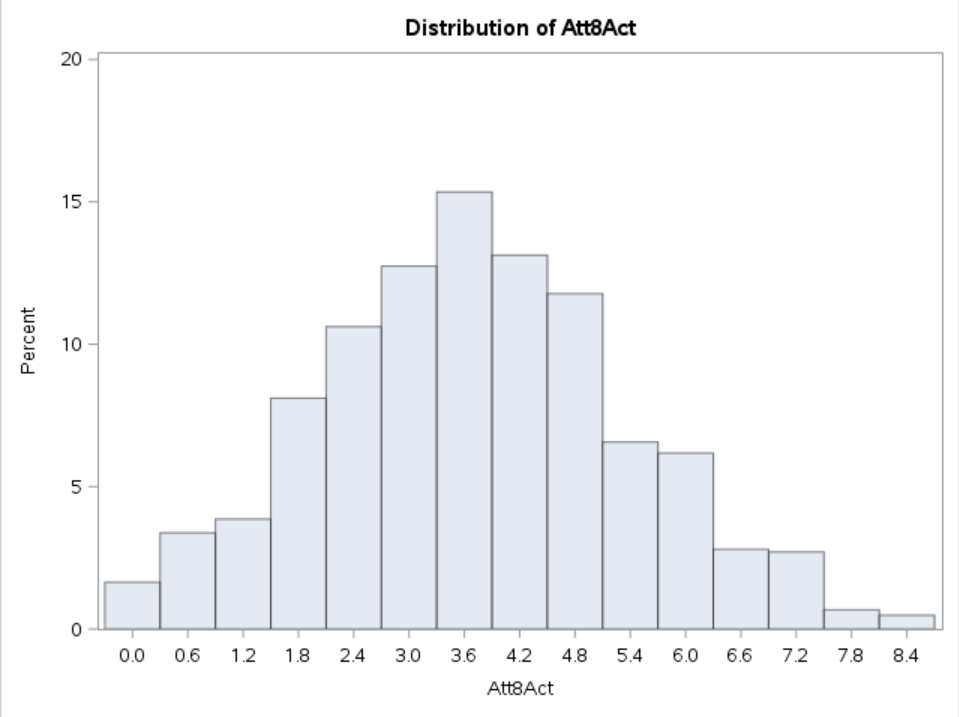
4. SEN


Distribution of SEN
Percent
80
60
40
20
0
0.03 0.09 0.15 0.21 0.27 0.33 0.39 0.45 0.51 0.57 0.63 0.69 0.75 0.81 0.87 0.93 0.99
SEN

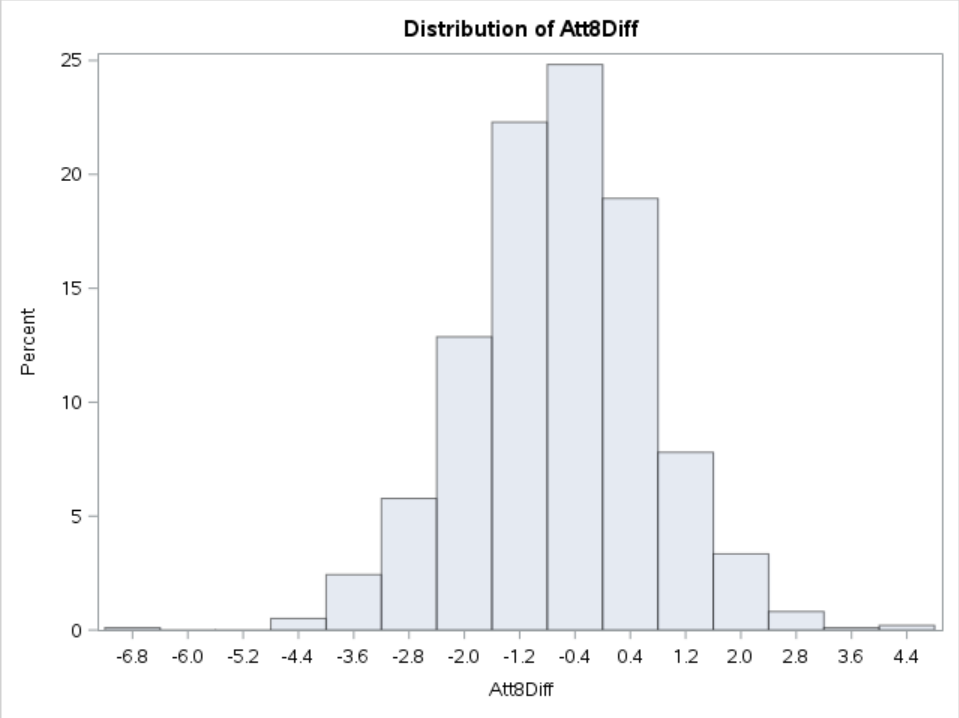
5. **Att8Est**


Distribution of Att8Est
Percent
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8 5.2 5.6 6.0 6.4 6.8 7.2 7.6
Att8Est

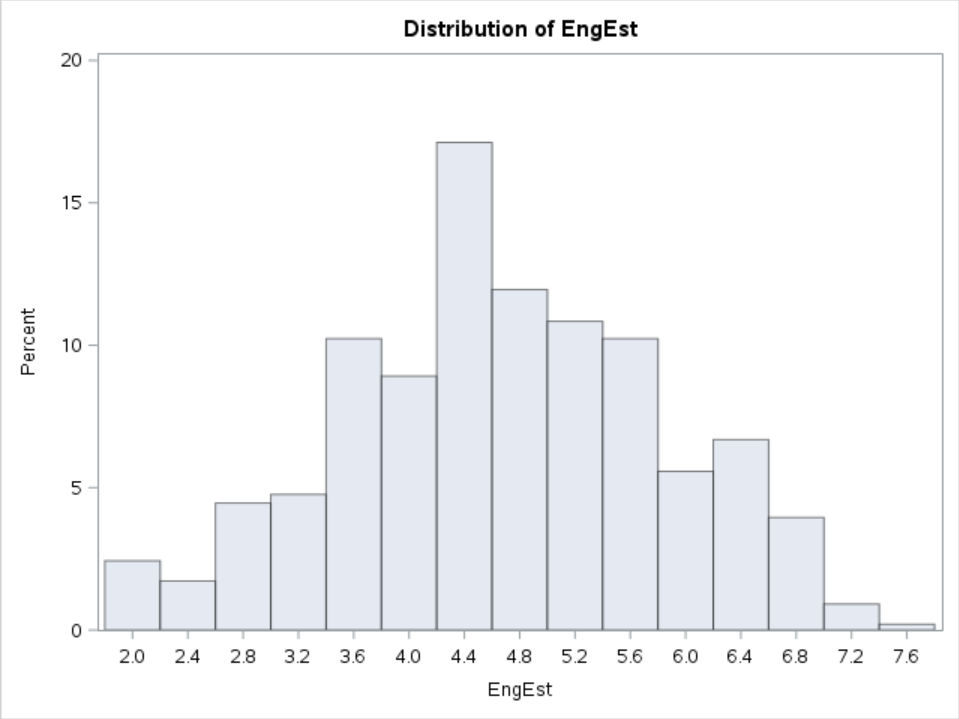
6. Att8Act


Distribution of Att8Act
Percent
20
15
10
5
0
0.0 0.6 1.2 1.8 2.4 3.0 3.6 4.2 4.8 5.4 6.0 6.6 7.2 7.8 8.4
Att8Act

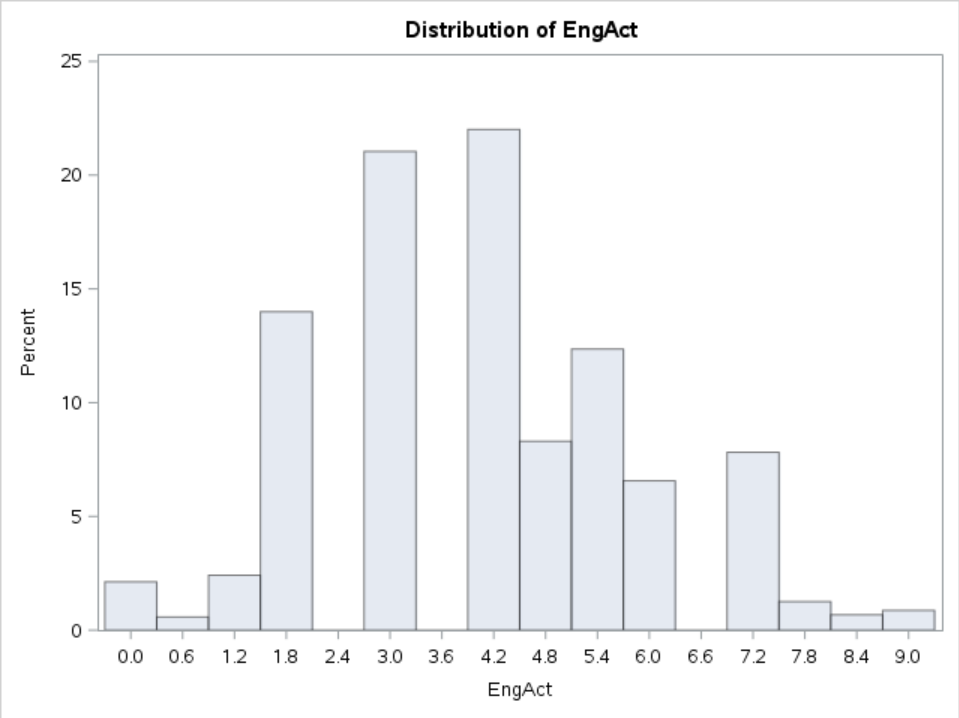
7. **Att8Diff**



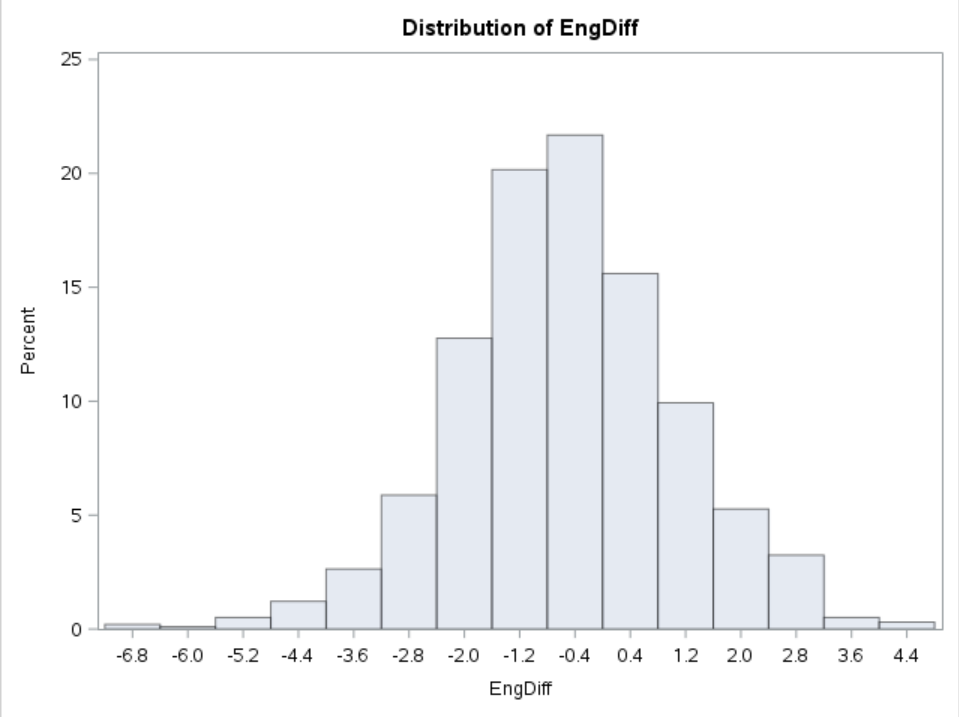
8. **EngEst**



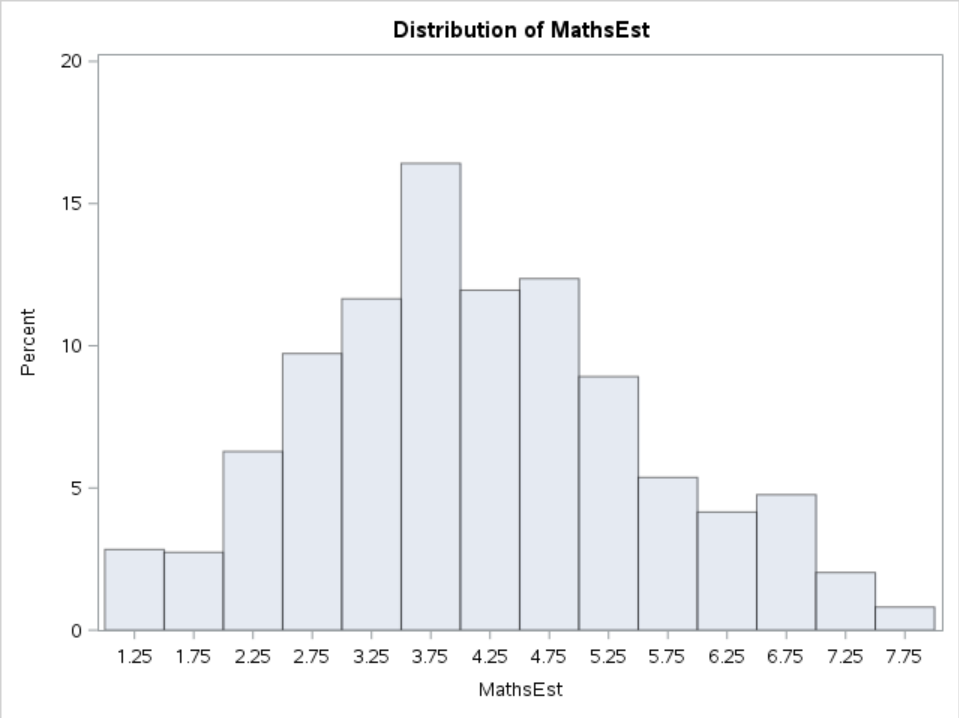
9. **EngAct**



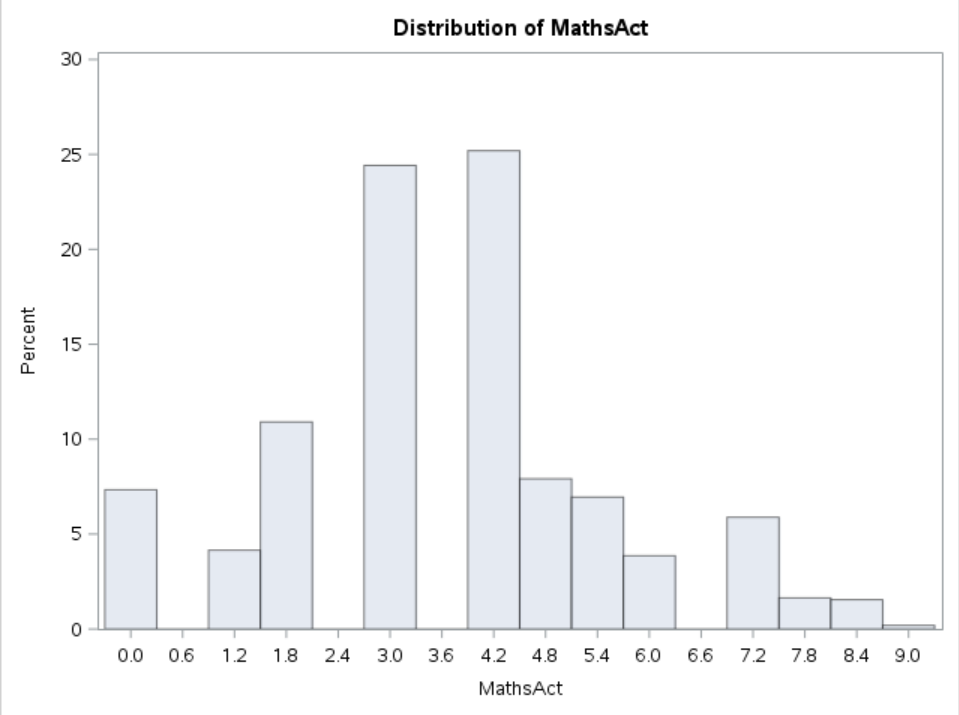
10. EngDiff



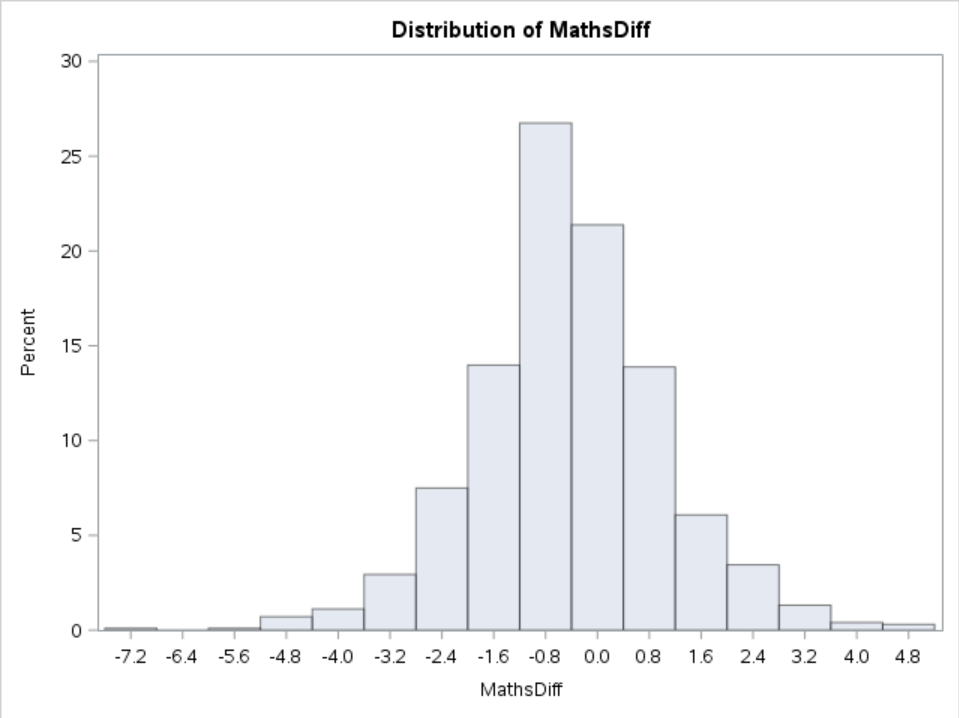
11. MathsEst



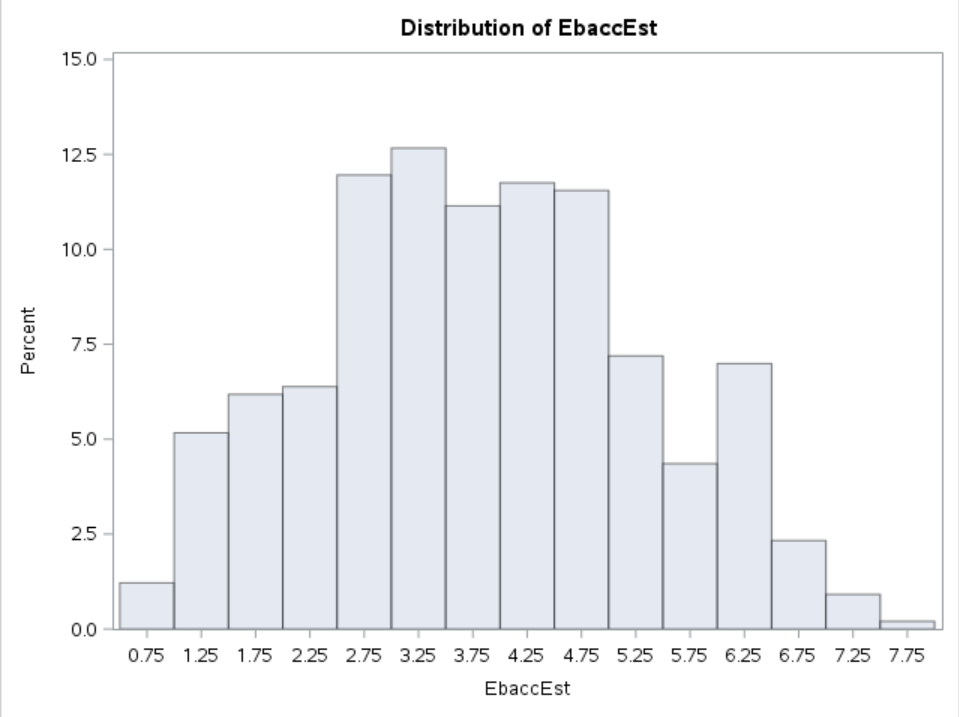
12. MathsAct



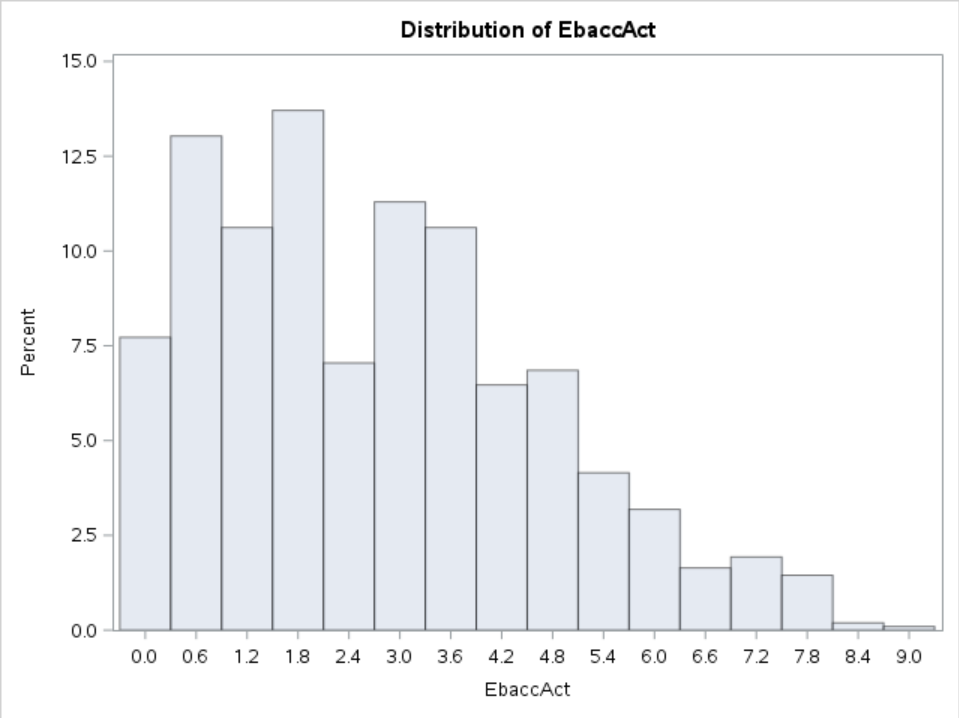
13. MathsDiff



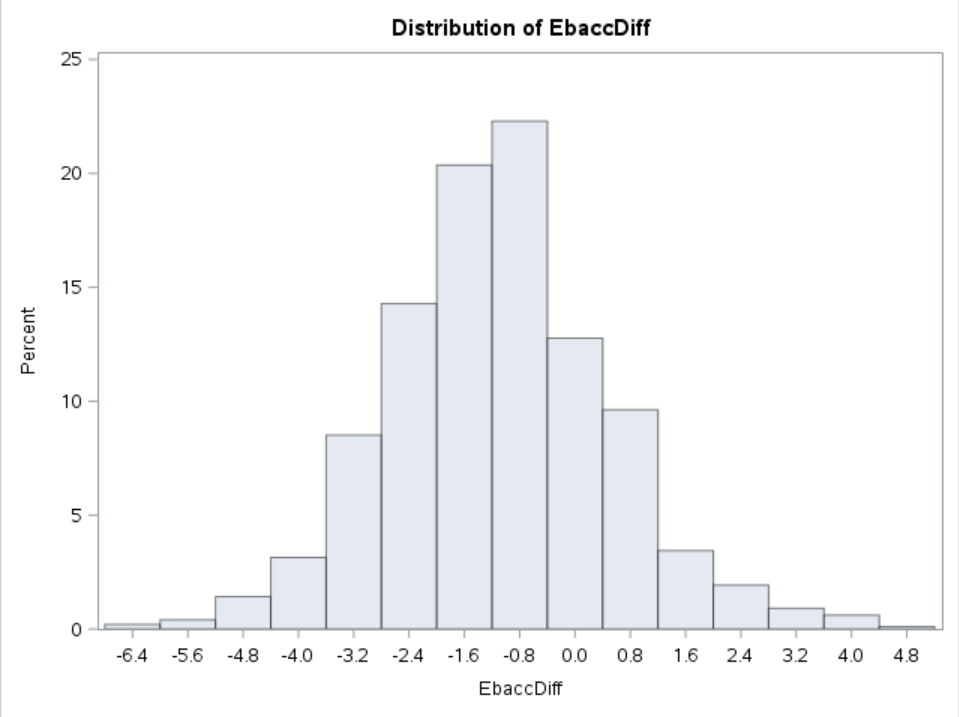
14. EbaccEst



15. EbaccAct



16. EbaccDiff



4. Data Quality Report:

The report documents the predictions from a dataset containing samples from 4 years ranging over a 1000 student grades. It contains the results in different subjects such as Math and English following each an estimate and an actual grade. The data also shows English Baccalaureate scores. Some of the data was blank like reading scores, writing scores, etc. We used MICE to help alleviate some of the drawbacks from that but having the complete data would be most beneficial.

5. Data Preparation Report:

- ✓ Feature Selection: Data Inclusion/Exclusion: For predicting the student's final grade, we included the following columns: PP, EAL, SEN, HML, RE, WR, MA, Att8Est, Att8Act, EngEst, EbaccEst, EbaccAct. It was not clear how the estimates were made for the data provided but we decided that that was out of our scope for this project. We did not include the primary key, the difference between the estimated grades and actual grades because we didn't have the final grades yet so having the difference didn't make sense. We also left out the actual grades for English and Math. We left out Ebacc difference since it didn't support the prediction in a good way.
- ✓ Data Cleaning/Missing Data: For data cleaning, we used the replace using MICE cleaning mode. This mode supplied a good prediction value as well as letting us leave some string values in our dataset that were important in determining student grades.
- ✓ Derived Attributes/Feature engineering (Transform &Simplify): We didn't find a way to incorporate feature engineering into our models. Given more details and time, we may have been able to create ratios of current grades to help improve the overall prediction.
- ✓ Data Integration: We didn't integrate any outside data. We did have to research Ebacc scores, what they meant, and how they were derived. We learned that the Ebacc scores were the sum of the scores from students, divided by the number of students taking the exam. We thought that this was important to include because the level of education may vary on the or even the professor. Using previous data from similar groups may help determine/predict future grades. We thought about using predicted scores to help predict other scores but decided it would get very messy and inaccurate very quick, so we decided to run the predictive analysis separately.

6. Data Modeling Report:

✓ Regressions/Classifications

✓ For Att8Act and EngAct we used the Bayesian Linear Regression Model. The regularization weights were kept at 1 for both variables. For MathsAct, we used a Linear Regression Model. The L2 regularization weight was kept at 0.001 and the random number seed was 12345.

7. Results and Recommendations:

Below are the R squared and RMSE scores for Att8Act, MathsAct, and EngAct

Att8Act		
Random Seed: 12345	USING MICE	
Type of Regression	**R Squared**	**RMSE**
Boosted Decision Tree	0.83601	0.655149
Bayesian Linear Regression	0.87018	0.582911
Decision Forest Regression	0.851354	0.629174
Linear Regression	0.868711	0.586198
Neural Network Regression	0.849598	0.62742
Poisson Regression	0.803785	0.716635

MathsAct		
Random Seed: 12345	USING MICE	
Type of Regression	**R Squared**	**RMSE**
Boosted Decision Tree	0.630482	1.159188
Bayesian Linear Regression	0.718707	1.011382
Decision Forest Regression	0.649859	1.127998
Linear Regression	0.722665	1.004242
Neural Network Regression	0.718602	1.01157
Poisson Regression	0.650253	1.127749

EngAct		
Random Seed: 12345	USING MICE	
Type of Regression	**R Squared**	**RMSE**
Boosted Decision Tree	0.457972	1.325066
Bayesian Linear Regression	0.548831	1.208917
Decision Forest Regression	0.527045	1.255107
Linear Regression	0.545139	1.213853

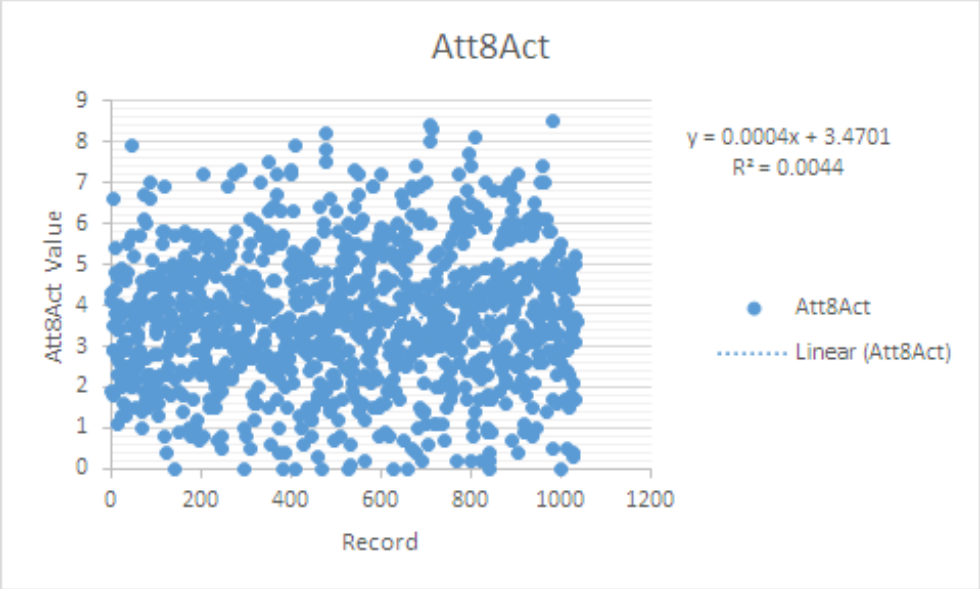
Neural Network Regression	0.544262	0.544262
Poisson Regression	0.511437	1.258019

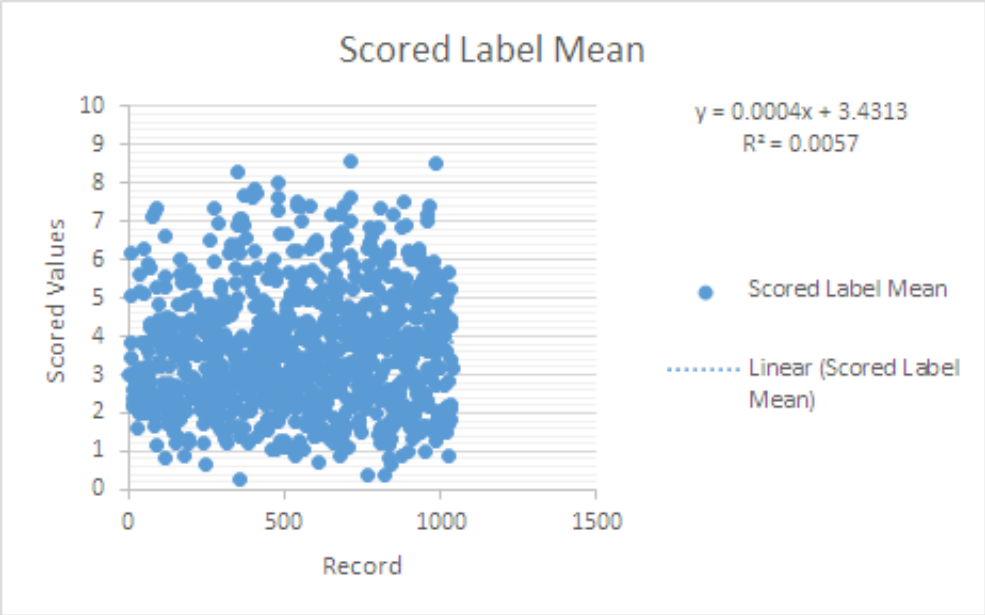
Below are charts with scatter plots of all the records.

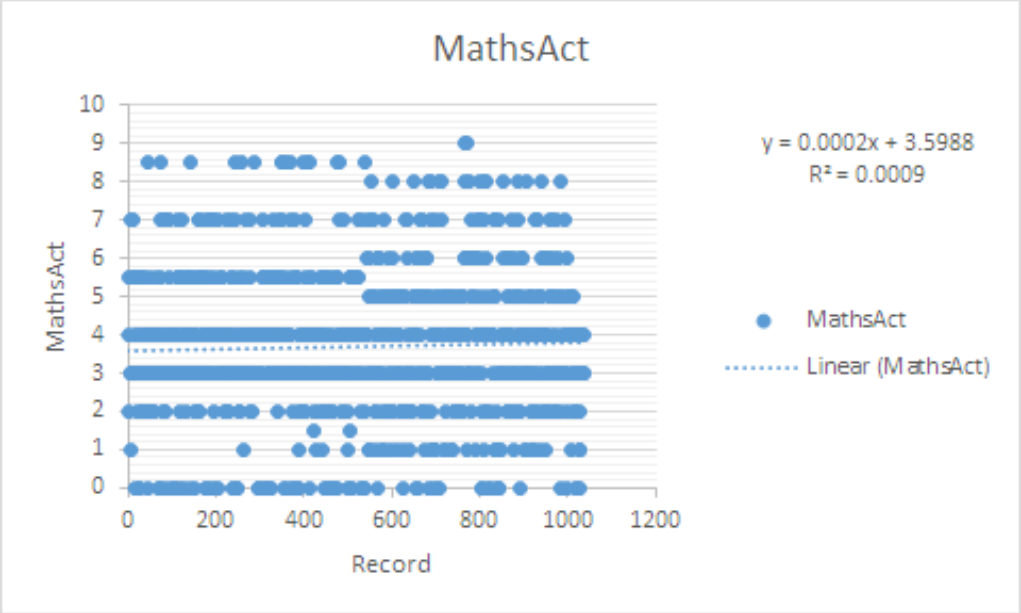
Our findings show that our Azure models had the Att8Act as the best predicted variable. According to the scatter plots below, EngAct has the most resemblance. Looking at the predicted values in the Excel file, there are some predicted values that are spot on and others that are completely off. For further research, we would investigate any similarities between data that has bad prediction values. From here, we could decide to what we would do with the data, like binning for example or clipping depending on the cause of bad prediction.

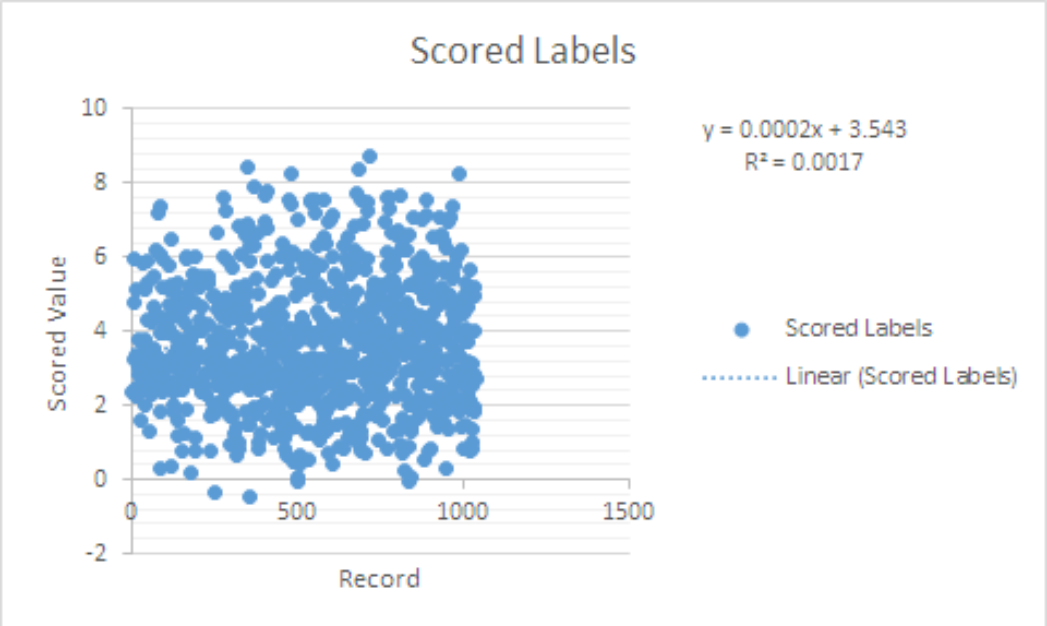
We started off this project thinking that we should do separate machine learning models. Then we moved towards thinking that we should predict Att8Act because that had the highest r squared value, then use that value to predict MathsAct followed by EngAct. However, the R squared and RMSE values would be very bad since we're already using predicted data. We scrapped this and went back to our original idea which is shown in this paper.

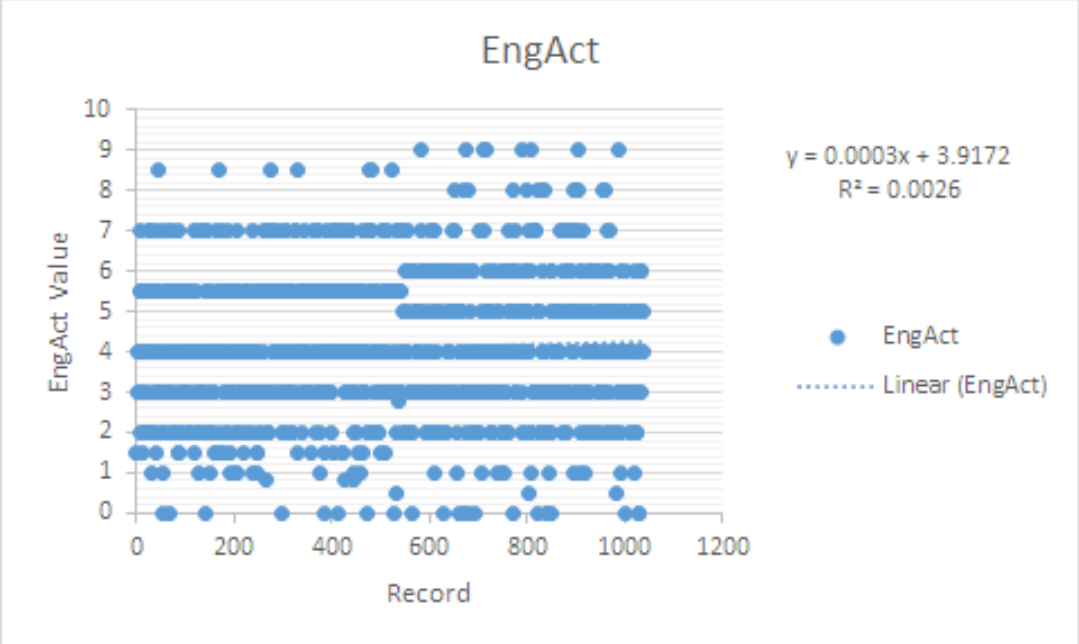
So, can this predictive model be used to predict the Att8Act? Not entirely. We have some predictions that are spot on, but we also have a difference that would be substantial enough in a real-world application where it shouldn't be depended on. There is potential for improvement with having more complete data and more time to figure out the details of what causes the poor predictions to be that wrong. I would recommend more research and clarification on the data provided, then approaching it with the same model as we did and some additional models depending on how the data is prepared.

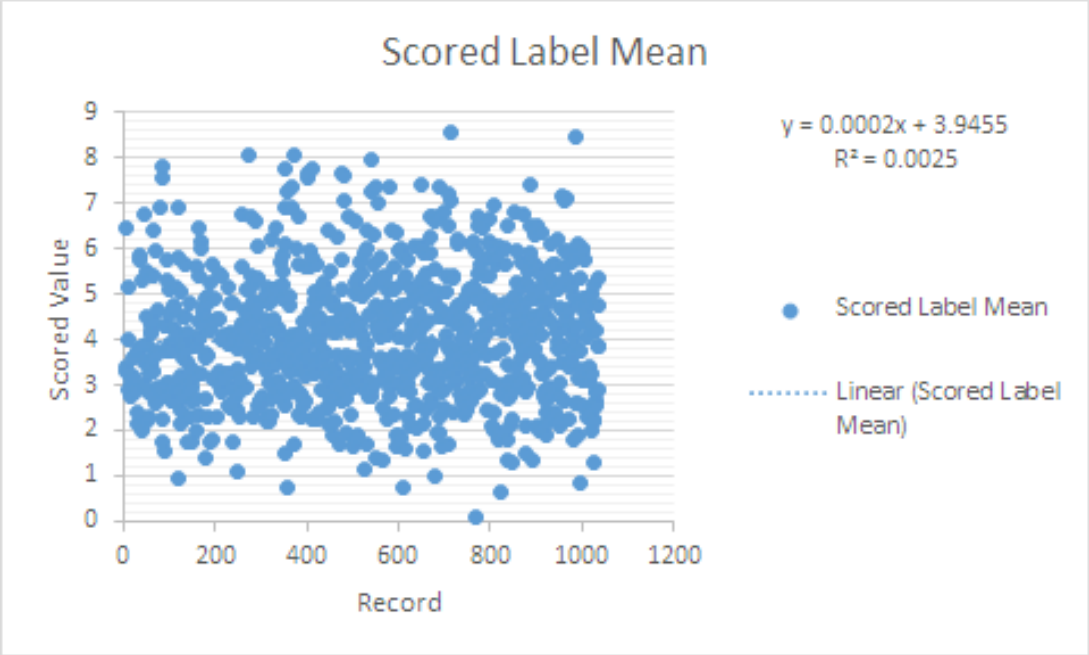

Att8Act
Att8Act Value
Record
y = 0.0004x + 3.4701
R² = 0.0044
Att8Act
Linear (Att8Act)


Scored Label Mean
Scored Values
Record
y = 0.0004x + 3.4313
R² = 0.0057
Scored Label Mean
Linear (Scored Label Mean)


MathsAct
MathsAct
Record
y = 0.0002x + 3.5988
R² = 0.0009
MathsAct
Linear (MathsAct)
0
1
2
3
4
5
6
7
8
9
10
0
200
400
600
800
1000
1200


Scored Labels
Scored Value
Record
y = 0.0002x + 3.543
R² = 0.0017
Scored Labels
Linear (Scored Labels)
-2
0
2
4
6
8
10
0
500
1000
1500


EngAct
EngAct Value
Record
y = 0.0003x + 3.9172
R² = 0.0026
EngAct
Linear (EngAct)
0
1
2
3
4
5
6
7
8
9
10
0
200
400
600
800
1000
1200



8. References:

→ https://www.kaggle.com/daviddraper1518/predicting-student-grades#AllResults.csv

→ https://app.myeducator.com/reader/web/1561b/