Syracuse University

Project Portfolio Milestone

Detailed Description and Procedures

Mackenzie Houser

M.S. Applied Data Science

March 2024

Table of Contents

Learning Goals

1.) **Collect, store, and access data by identifying and leveraging applicable technologies.**

   a.) There are three parts to this goal. First, collection, this refers to obtaining data successfully. Whether it be from personal data accumulation or scrapping from external sources. Next, storing data, is alluding to keeping track of data that will potentially be used in one's data science process. This can look like building a database for larger schemes or simply downloading files to a pc or notebook for smaller data. Lastly, access is the main reason the first two steps are performed. Without completion of the previous part, the next section can't be successfully carried out. Access is the ability to read the data being collected and stored into a program that allows for further analysis. The data science program has taught all of the above.

2.) **Create actionable insight across a range of contexts using data and the full data science life cycle.**

   a.) The data science life cycle refers to: Problem Definition, Data Investigation & Cleaning, Data Exploration, Modeling Data, and Evaluation.

      i.) Problem Definition

         (1) This is where an idea of what the variables look like and possible need for cleaning present themselves.

         (2) Understanding data can mean more than just looking at the data points in the file. In order to solve the problem, it is important to have a deeper understanding.

      ii.) Data Investigation & Cleaning

         (1) Scrubbing data is important because if not performed, results could be inaccurate. In some cases, scrubbing isn't necessary because the data is already clean.

      iii.) Data Exploration

         (1) Exploring data is important to do because it leads into modeling. This is where the variables that have relationships are revealed.

      iv.) Modeling Data

         (1) Modeling relationships is where EDA is taken into account and different variables are used to test out which models are best at predicting.

      v.) Evaluation

         (1) Evaluation and Interpreting findings is where questions are answered.

3.) **Apply visualization and predictive models to help generate actionable insight.**

   a.) Visualizations make understanding data clear and simple. Visualizations can include things like: histograms, bar plots, count plots, heat maps, etc.

   b.) Predictive models are crucial to data scientists because they assist in arriving at findings and conclusions. These models consist of: decision trees, time series, random forest, regression, etc.

4.) **Use programming languages such as R and Python to support the generation of actionable insight.**

   a.) R and Python are the two main programming languages taught. R is ideal for calculations and visualizations. Python is most commonly used for handling big data and deep learning. Another programming language is SQL. SQL is optimal for storage, updating, removing, and retrieving data from a database.

   b.) Microsoft Excel isn't a programming language but is another very important tool used in analytics.

5.) **Communicate insights gained via visualization and analytics to a broad range of audiences.**

   a.) Visualizations are key in data science, especially when presenting data to those who might not have the necessary background to interpret data findings.Data analysis is oftentimes performed in order to arrive at a business decision. An example of a visualization communicating analytics is a time series model.

6.) **Apply ethics in the development, use and evaluation of data and predictive models.**

   a.) Ethics is the concern of human conduct and behavior. Ethics is a large part of data analysis and predictive modeling to make sense of findings. This could be something like using ethics to build models according to factors that would be expected. For example, ice cream sales most likely increase when the weather is warm.

Projects Overview

1.) IST 718 - Lab - Individual Assignment - Python

a.) Objective: Demonstrate ability to produce meaningful analysis. Specifically, provide a decision maker with more than just data.

b.) Question: How can we recommend the best salary for our next head football coach?

c.) Deliverables: Python code, written report

2.) IST 707 - Student Grade Prediction - Group Project - R-Studio

a.) Objective: Use skills taught in class to solve a real data mining problem.

b.) Problem: Schools can not seem to identify students who may need additional support before they are unsuccessful in the course.

c.) Goal: Determine what is contributing to the students failing the classes and what these schools can do to identify them early and offer additional assistance.

d.) Deliverables: R code, written report, presentation slides

3.) IST 659 - Whole-ER Foods - Group Project - SQL

a.) Objective: Demonstrate ability to work in a team to design and implement a functional system with a database, based on what you have learned in the course.Devise your own database to design and implement.

b.) Problem: Whole-ER Foods are able to get customers in with their greener, fresher, and healthier products, but are unable to supply enough products to keep up with demand. Currently, the store tracks inventory manually in Microsoft Excel.

c.) Goal: Upgrade store to a Relational Database in order to improve inventory tracking to keep up with customer demand.

d.) Deliverables: data files, presentation slides

4.) **MBC 638 - Process Improvement - Individual Project - Microsoft Exce**l

a.) Objective: Select an issue or opportunity that can be written as a problem statement. Use data that is accessible to you or can be collected in a reasonable amount of effort/time. Fixing this problem will provide value; develop a business case to support working this issue.

b.) Problem: Having 24/7 access to a cell phone due to remote work and school leads to excessive use of the phone. Leading to impacted time management and ability to focus.

c.) Goal: Decrease daily screen time on weekdays.

d.) Deliverables: Microsoft Excel file, presentation slide

Project One

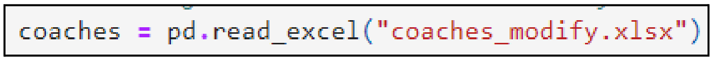
Learning Objectives: 1, 2, 3, 4, 6
IST 718 - Lab One

Coaches

The objective of this assignment was to demonstrate the ability to produce meaningful analysis using Python in order to answer the question: *How can we recommend the best salary for our next head football coach?*

Learning Objective One

Collect: The csv file was provided by Professor Lando.
Store: Stored on pc in JupyterNotebook.
Access: Data read into JupyterNotebook python file using the pandas package.

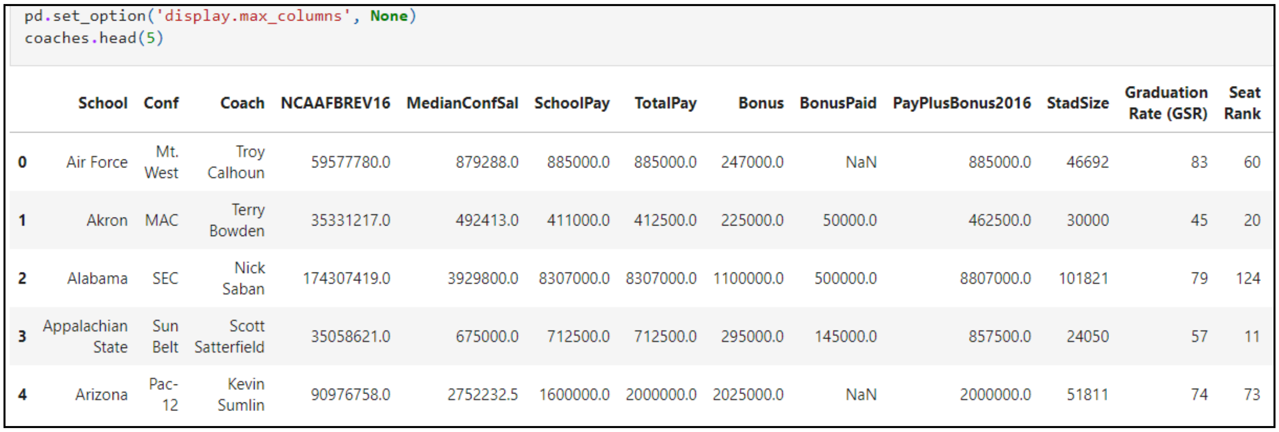


```
coaches = pd.read_excel("coaches_modify.xlsx")
```

Learning Objective Two

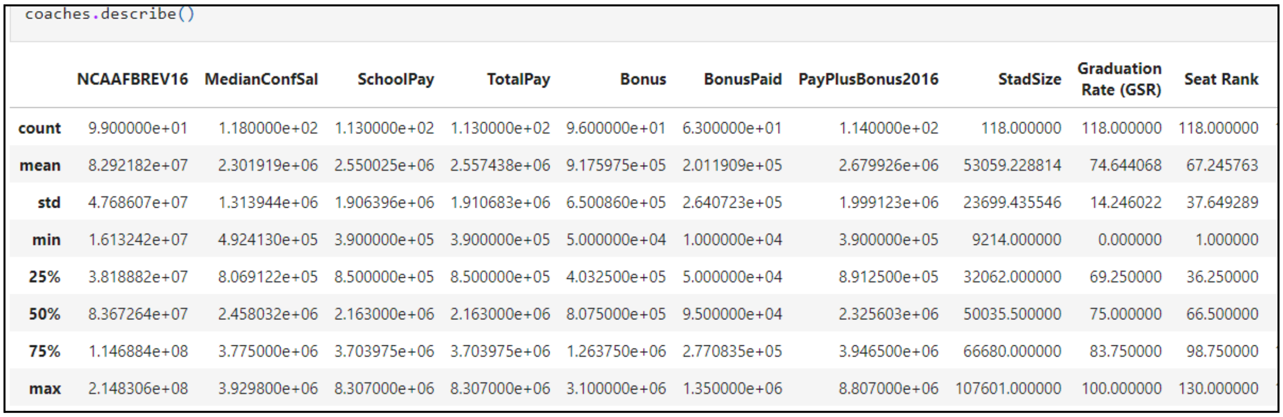
Understand

- Using functions like head() and describe() in Python can provide a glimpse of the data that will soon be analyzed in the future.



```
pd.set_option('display.max_columns', None)
coaches.head(5)
```

	School	Conf	Coach	NCAAFBREV16	MedianConfSal	SchoolPay	TotalPay	Bonus	BonusPaid	PayPlusBonus2016	StadSize	Graduation Rate (GSR)	Seat Rank
0	Air Force	Mt. West	Troy Calhoun	59577780.0	879288.0	885000.0	885000.0	247000.0	NaN	885000.0	46692	83	60
1	Akron	MAC	Terry Bowden	35331217.0	492413.0	411000.0	412500.0	225000.0	50000.0	462500.0	30000	45	20
2	Alabama	SEC	Nick Saban	174307419.0	3929800.0	8307000.0	8307000.0	1100000.0	500000.0	8807000.0	101821	79	124
3	Appalachian State	Sun Belt	Scott Satterfield	35058621.0	675000.0	712500.0	712500.0	295000.0	145000.0	857500.0	24050	57	11
4	Arizona	Pac-12	Kevin Sumlin	90976758.0	2752232.5	1600000.0	2000000.0	2025000.0	NaN	2000000.0	51811	74	73

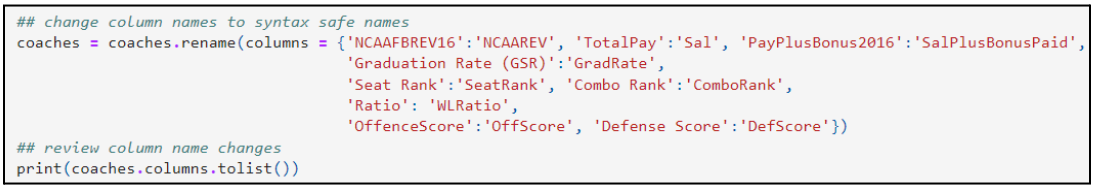


```
coaches.describe()
```

	NCAAFBREV16	MedianConfSal	SchoolPay	TotalPay	Bonus	BonusPaid	PayPlusBonus2016	StadSize	Graduation Rate (GSR)	Seat Rank
count	9.900000e+01	1.180000e+02	1.130000e+02	1.130000e+02	9.600000e+01	6.300000e+01	1.140000e+02	118.000000	118.000000	118.000000
mean	8.292182e+07	2.301919e+06	2.550025e+06	2.557438e+06	9.175975e+05	2.011909e+05	2.679926e+06	53059.228814	74.644068	67.245763
std	4.768607e+07	1.313944e+06	1.906396e+06	1.910683e+06	6.500860e+05	2.640723e+05	1.999123e+06	23699.435546	14.246022	37.649289
min	1.613242e+07	4.924130e+05	3.900000e+05	3.900000e+05	5.000000e+04	1.000000e+04	3.900000e+05	9214.000000	0.000000	1.000000
25%	3.818882e+07	8.069122e+05	8.500000e+05	8.500000e+05	4.032500e+05	5.000000e+04	8.912500e+05	32062.000000	69.250000	36.250000
50%	8.367264e+07	2.458032e+06	2.163000e+06	2.163000e+06	8.075000e+05	9.500000e+04	2.325603e+06	50035.500000	75.000000	66.500000
75%	1.146884e+08	3.775000e+06	3.703975e+06	3.703975e+06	1.263750e+06	2.770835e+05	3.946500e+06	66680.000000	83.750000	98.750000
max	2.148306e+08	3.929800e+06	8.307000e+06	8.307000e+06	3.100000e+06	1.350000e+06	8.807000e+06	107601.000000	100.000000	130.000000

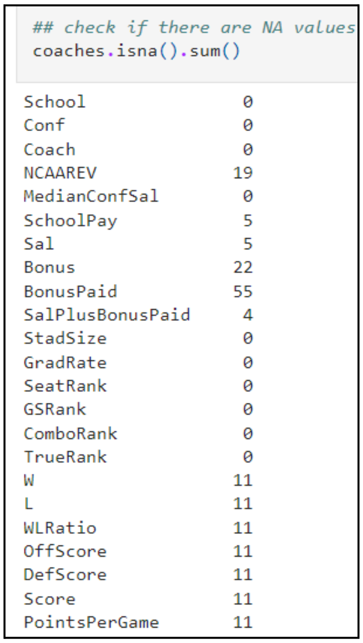
Scrub

- The data contains: non-syntax friendly column names and missing data.
- Adjusting variable names to make them syntax friendly- Python has a function called rename().



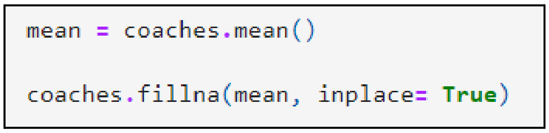
```
## change column names to syntax safe names
coaches = coaches.rename(columns = {'NCAAFBREV16':'NCAAREV', 'TotalPay':'Sal', 'PayPlusBonus2016':'SalPlusBonusPaid',
                                    'Graduation Rate (GSR)':'GradRate',
                                    'Seat Rank':'SeatRank', 'Combo Rank':'ComboRank',
                                    'Ratio': 'WLRatio',
                                    'OffenceScore':'OffScore', 'Defense Score':'DefScore'})
## review column name changes
print(coaches.columns.tolist())
```

- In order to pinpoint where missing values are located, the function isna() runs through each variable. To accommodate for the missing values, there are a few options like removing them or replacing them. In this case, missing values are replaced using fillna() and are replaced with the variable mean.



```
## check if there are NA values
coaches.isna().sum()
```

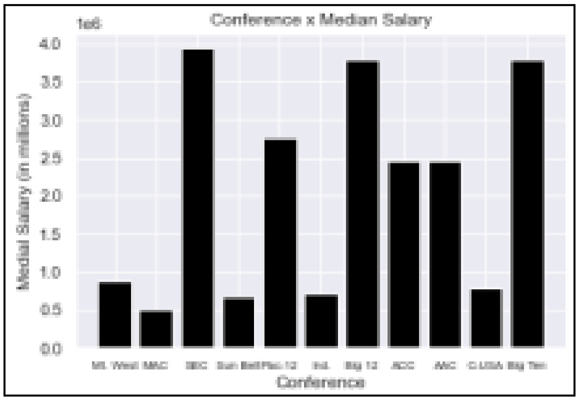
```
School               0
Conf                 0
Coach                0
NCAAREV             19
MedianConfSal        0
SchoolPay            5
Sal                  5
Bonus               22
BonusPaid           55
SalPlusBonusPaid     4
StadSize             0
GradRate             0
SeatRank             0
GSRank               0
ComboRank            0
TrueRank             0
W                   11
L                   11
WLRatio             11
OffScore            11
DefScore            11
Score               11
PointsPerGame       11
```



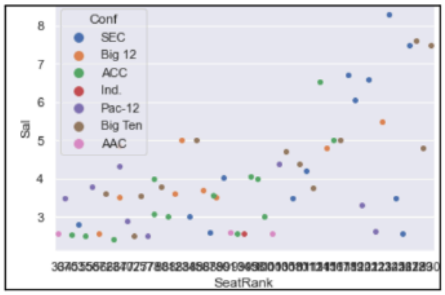
```
mean = coaches.mean()

coaches.fillna(mean, inplace= True)
```

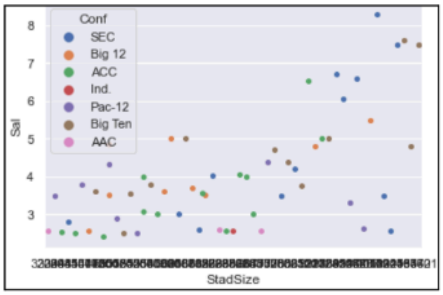
Exploration Findings

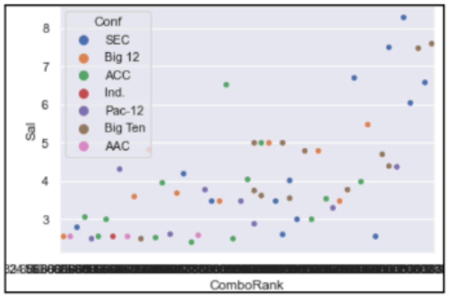
- The highest level of median salary seems to come from the SEC, Big 12, and Big Ten.

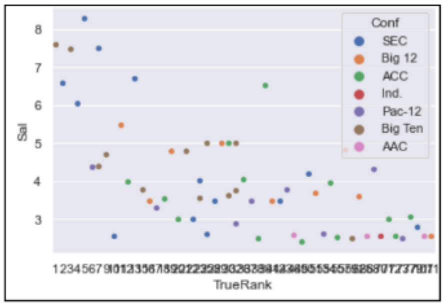


- Upper level salary range coaches have positive and fairly linear trends with SeatRank / StadSize, ComboRank, and TrueRank.



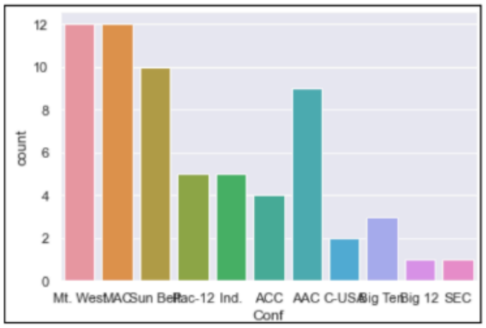




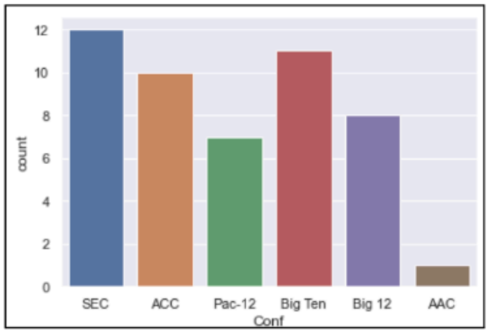


- All conferences contain at least one coach with a low salary, but not all conferences have a coach with and above average salary.

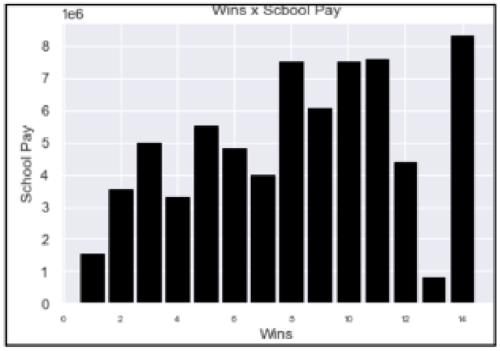
Lower Salary

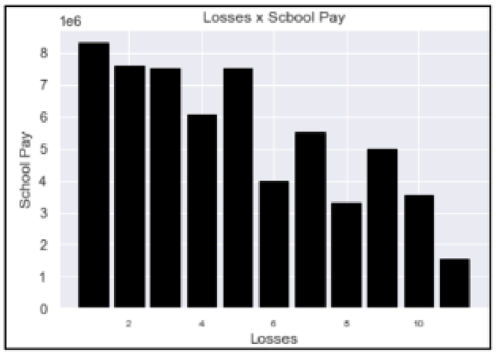


Higher Salary

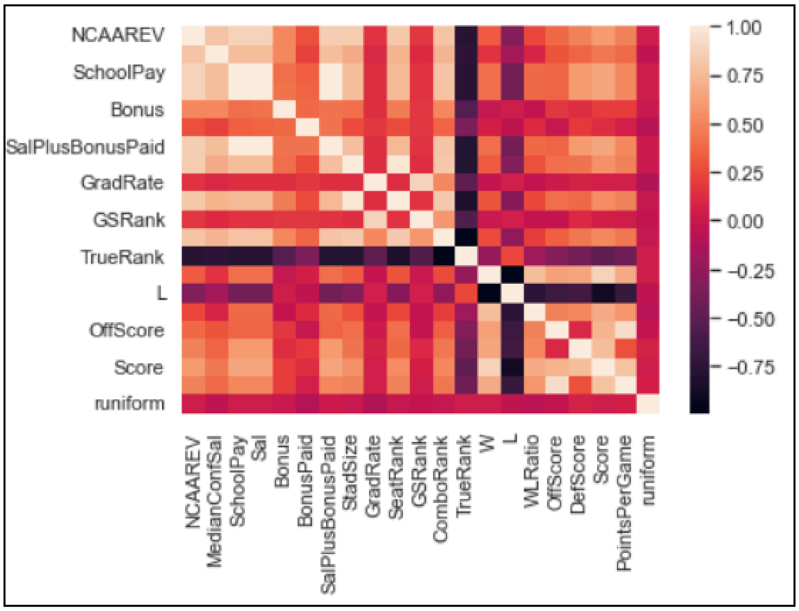


- More wins, higher salary, and vice versa.



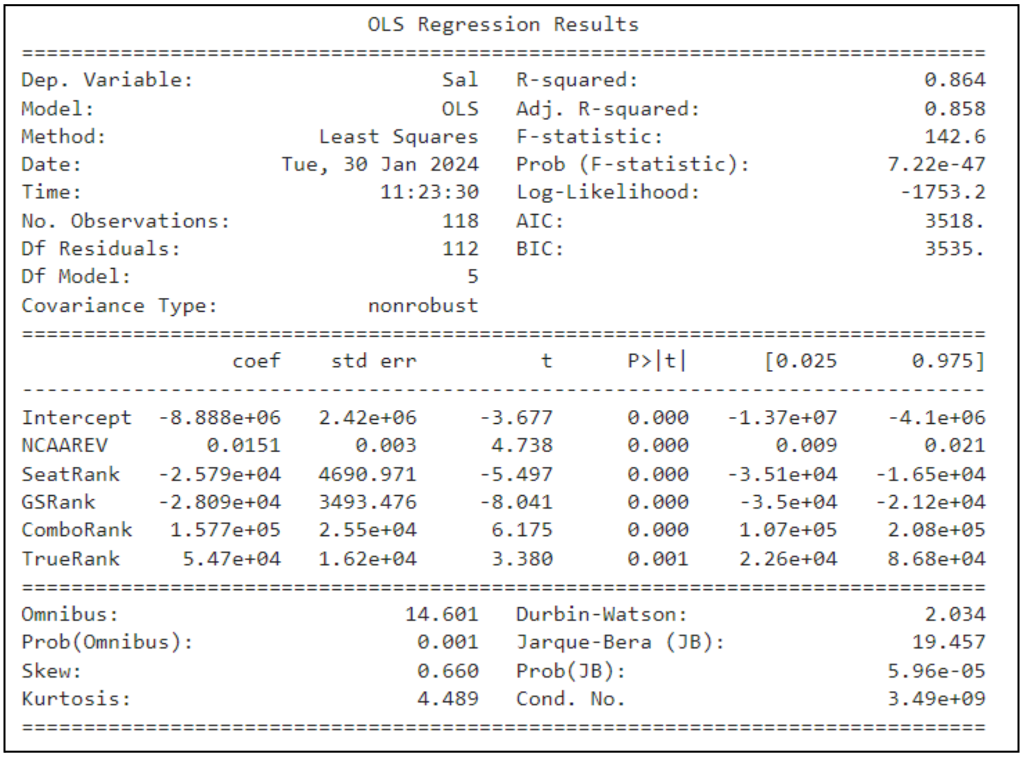


- The variables with a decently strong correlation to the dependent variable, Salary are: *NCAAREV, MedianConfSal, StadSize, SeatRank, ComboRank, TrueRank, and Score*.



Model

- Regression was used. This requires splitting the data into testing and training sets in order to validate the model. After running through a few models with different variables, the best model was found.



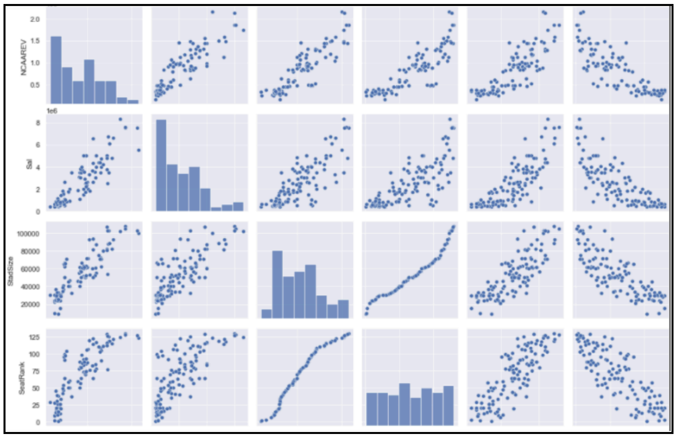
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    Sal   R-squared:                       0.864
Model:                            OLS   Adj. R-squared:                  0.858
Method:                 Least Squares   F-statistic:                     142.6
Date:                Tue, 30 Jan 2024   Prob (F-statistic):           7.22e-47
Time:                        11:23:30   Log-Likelihood:                -1753.2
No. Observations:                 118   AIC:                             3518.
Df Residuals:                     112   BIC:                             3535.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -8.888e+06   2.42e+06     -3.677      0.000   -1.37e+07    -4.1e+06
NCAAREV        0.0151      0.003      4.738      0.000       0.009       0.021
SeatRank   -2.579e+04   4690.971     -5.497      0.000   -3.51e+04   -1.65e+04
GSRank     -2.809e+04   3493.476     -8.041      0.000    -3.5e+04   -2.12e+04
ComboRank   1.577e+05   2.55e+04      6.175      0.000    1.07e+05    2.08e+05
TrueRank     5.47e+04   1.62e+04      3.380      0.001    2.26e+04    8.68e+04
==============================================================================
Omnibus:                       14.601   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               19.457
Skew:                           0.660   Prob(JB):                     5.96e-05
Kurtosis:                       4.489   Cond. No.                     3.49e+09
==============================================================================
```

Interpret

- The question: *What is the recommended salary for the Syracuse football coach?* Using the regression model that was built above, the formula looks like: *-8.888e+06 + 0.0151*(8.292182e+07) - 2.579e+04*(63) - 2.809e+04*(87) + 1.577e+05*(73.333333) + 5.47e+04*(49)* . To answer the question, the recommended salary for the Syracuse football coach is **$2,540,486.096**.

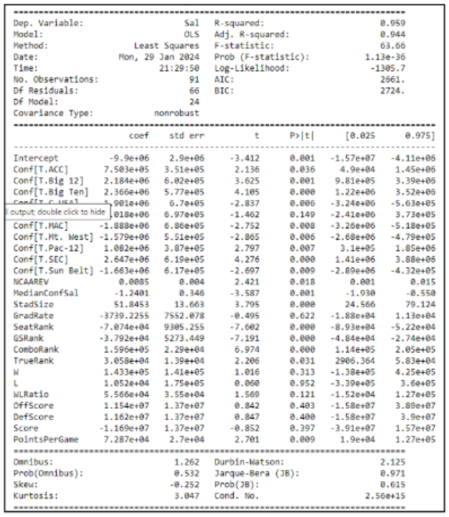
Learning Objective Three

Visualization: The variables used to create the pair plot visualization below were chosen by using strong correlation values. Visualizing correlation displays a better understanding than just numbers.



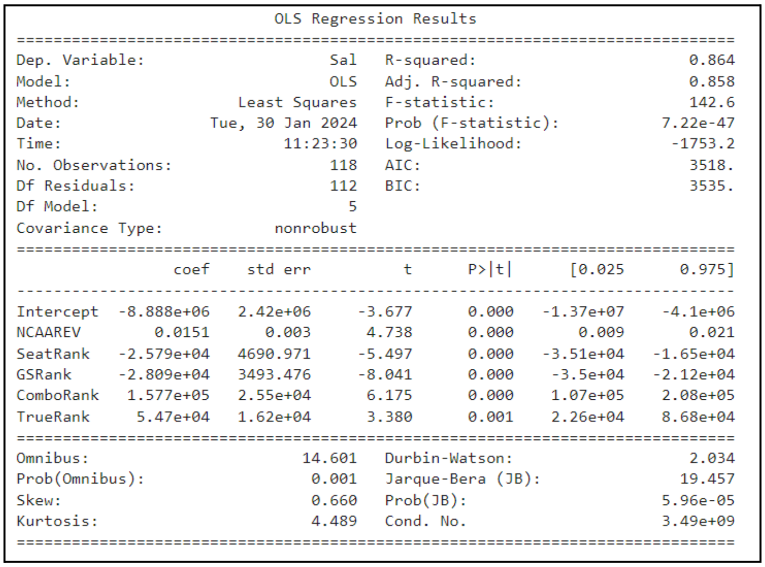
Predictive Model: The variables that were deemed correlated by visualization leads to building of models by starting with all the variables and narrowing them down to the most significant, ending with the best fit model.

- Beginning Model: more variables, less significant



```
==============================================================================
Dep. Variable:                    Sal   R-squared:                       0.959
Model:                            OLS   Adj. R-squared:                  0.944
Method:                 Least Squares   F-statistic:                     63.66
Date:                Mon, 29 Jan 2024   Prob (F-statistic):           1.13e-36
Time:                        21:29:50   Log-Likelihood:                -1305.7
No. Observations:                  91   AIC:                             2661.
Df Residuals:                      66   BIC:                             2724.
Df Model:                          24
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept        -9.9e+06    2.9e+06     -3.412      0.001   -1.57e+07   -4.11e+06
Conf[T.ACC]     7.503e+05   3.51e+05      2.136      0.036     4.9e+04    1.45e+06
Conf[T.Big 12]  2.184e+06   6.02e+05      3.625      0.001    9.81e+05    3.39e+06
Conf[T.Big Ten] 2.366e+06   5.77e+05      4.105      0.000    1.22e+06    3.52e+06
Conf[T.C-USA]  -1.901e+06    6.7e+05     -2.837      0.006   -3.24e+06   -5.63e+05
output double click to hide .018e+06   6.97e+05     -1.462      0.149   -2.41e+06    3.73e+05
Conf[T.MAC]    -1.888e+06   6.86e+05     -2.752      0.008   -3.26e+06   -5.18e+05
Conf[T.Mt. West] -1.579e+06 5.51e+05     -2.865      0.006   -2.68e+06   -4.79e+05
Conf[T.Pac-12]  1.082e+06   3.87e+05      2.797      0.007     3.1e+05    1.85e+06
Conf[T.SEC]     2.647e+06   6.19e+05      4.276      0.000    1.41e+06    3.88e+06
Conf[T.Sun Belt] -1.663e+06 6.17e+05     -2.697      0.009   -2.89e+06   -4.32e+05
NCAAREV            0.0085      0.004      2.421      0.018       0.001       0.015
MedianConfSal     -1.2401      0.346     -3.587      0.001      -1.930      -0.550
StadSize          51.8453     13.663      3.795      0.000      24.566      79.124
GradRate       -3739.2255   7552.078     -0.495      0.622   -1.88e+04    1.13e+04
SeatRank       -7.074e+04   9305.255     -7.602      0.000   -8.93e+04   -5.22e+04
GSRank         -3.792e+04   5273.449     -7.191      0.000   -4.84e+04   -2.74e+04
ComboRank       1.596e+05   2.29e+04      6.974      0.000    1.14e+05    2.05e+05
TrueRank        3.058e+04   1.39e+04      2.206      0.031    2906.364    5.83e+04
W               1.433e+05   1.41e+05      1.016      0.313   -1.38e+05    4.25e+05
L               1.052e+04   1.75e+05      0.060      0.952   -3.39e+05     3.6e+05
WLRatio         5.566e+04   3.55e+04      1.569      0.121   -1.52e+04    1.27e+05
OffScore        1.154e+07   1.37e+07      0.842      0.403   -1.58e+07    3.89e+07
DefScore        1.162e+07   1.37e+07      0.847      0.400   -1.58e+07     3.9e+07
Score          -1.169e+07   1.37e+07     -0.852      0.397   -3.91e+07    1.57e+07
PointsPerGame   7.287e+04    2.7e+04      2.701      0.009     1.9e+04    1.27e+05
==============================================================================
Omnibus:                        1.262   Durbin-Watson:                   2.125
Prob(Omnibus):                  0.532   Jarque-Bera (JB):                0.971
Skew:                          -0.252   Prob(JB):                        0.615
Kurtosis:                       3.047   Cond. No.                     2.56e+15
==============================================================================
```

- Ending Model: less variables, more significant



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    Sal   R-squared:                       0.864
Model:                            OLS   Adj. R-squared:                  0.858
Method:                 Least Squares   F-statistic:                     142.6
Date:                Tue, 30 Jan 2024   Prob (F-statistic):           7.22e-47
Time:                        11:23:30   Log-Likelihood:                -1753.2
No. Observations:                 118   AIC:                             3518.
Df Residuals:                     112   BIC:                             3535.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -8.888e+06   2.42e+06     -3.677      0.000   -1.37e+07    -4.1e+06
NCAAREV        0.0151      0.003      4.738      0.000       0.009       0.021
SeatRank   -2.579e+04   4690.971     -5.497      0.000   -3.51e+04   -1.65e+04
GSRank     -2.809e+04   3493.476     -8.041      0.000    -3.5e+04   -2.12e+04
ComboRank   1.577e+05   2.55e+04      6.175      0.000    1.07e+05    2.08e+05
TrueRank     5.47e+04   1.62e+04      3.380      0.001    2.26e+04    8.68e+04
==============================================================================
Omnibus:                       14.601   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               19.457
Skew:                           0.660   Prob(JB):                     5.96e-05
Kurtosis:                       4.489   Cond. No.                     3.49e+09
==============================================================================
```

Learning Objective Four

Python is used to complete the data science life cycle and provide key insights using data analysis, visualizations, modeling, etc as shown in the first three learning objectives. Python is better equipped for handling large datasets and stronger visualizations, at least in this experience. The weaknesses of Python are: difficulties connecting to the database, slow speed, and a weak memory.

Learning Objective Six

Ethics were used in project one by understanding the bias, fairness, and politics around college football. Whenever thinking about the big teams in cfb, most humans assume the conferences are Big Ten or SEC. These ethics came into play when analyzing and suggesting the next Syracuse football coach salary because it is known that the conference is ACC and the salary won't be as high as Big Ten or SEC, but also won't be as low as MAC.

Project Two

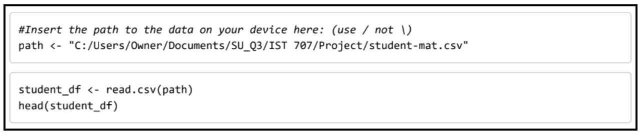
Learning Objectives: 1, 2, 3, 4, 5, 6
IST 707 - Group Project

<u>Student Grade Prediction</u>

The objective of this project was to use skills taught in the class to solve the problem: *Schools can not seem to identify students who may need additional support before they are unsuccessful in the course.*

<u>Learning Objective One</u>

Collect: Data retrieved from Kaggle.
Store: Stored on pc.
Access: Data read into R-Studio RMD file using read.csv() function.



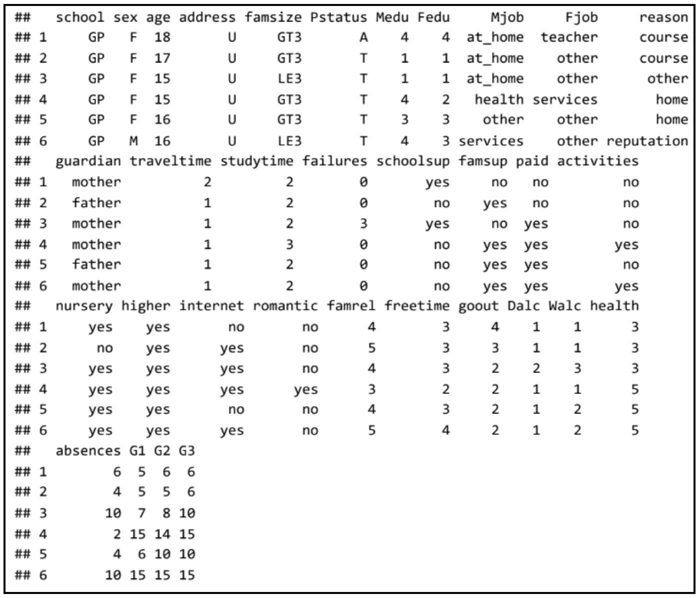
```
#Insert the path to the data on your device here: (use / not \)
path <- "C:/Users/Owner/Documents/SU_Q3/IST 707/Project/student-mat.csv"
```

```
student_df <- read.csv(path)
head(student_df)
```

<u>Learning Objective Two</u>

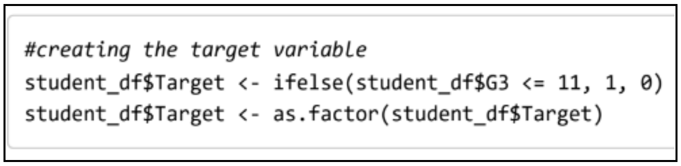
Understand

- Business Understanding:
    - The school that can identify who might struggle and how to best adapt for them will receive the best reputation.
    - Collecting historical data about students, the school can identify trends and adjust teaching strategies to meet particular needs.
    - A school is looked at by the grades it produces. With a successful study, the entire district can better prepare students for College while also benefiting their reputation which leads to increases in funding and more families desire to attend,
- Data Understanding:
    - head() function used in R to view the first few rows of the data set.



```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        3       yes     no  yes         no
## 4   mother          1         3        0        no    yes  yes        yes
## 5   father          1         2        0        no    yes  yes         no
## 6   mother          1         2        0        no    yes  yes        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3
## 1        6  5  6  6
## 2        4  5  5  6
## 3       10  7  8 10
## 4        2 15 14 15
## 5        4  6 10 10
## 6       10 15 15 15
```

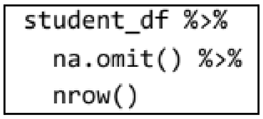
- Since the problem being solved is focused around students who are failing, a dummy variable is created to identify those students who are passing vs failing.



```
#creating the target variable
student_df$Target <- ifelse(student_df$G3 <= 11, 1, 0)
student_df$Target <- as.factor(student_df$Target)
```

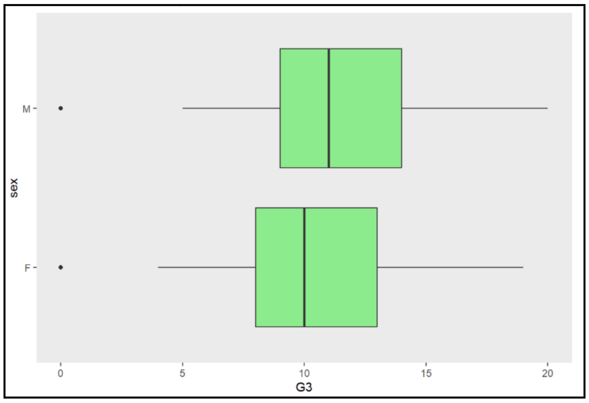
Scrub

- This data did not have to be cleaned, but it is important to check and make sure that the existing data is already clean. Null values were checked for by using na.omit() which removes rows with missing values and it ended up having the same amount of rows as started with.



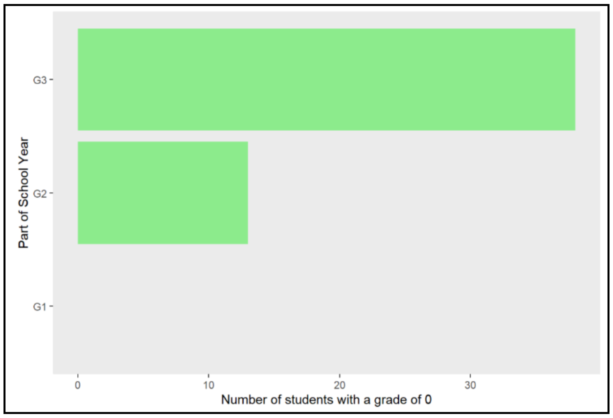
```
student_df %>%
  na.omit() %>%
  nrow()
```

- Some outliers were found, but in this case that is okay since the focus is on those students with failing grades.

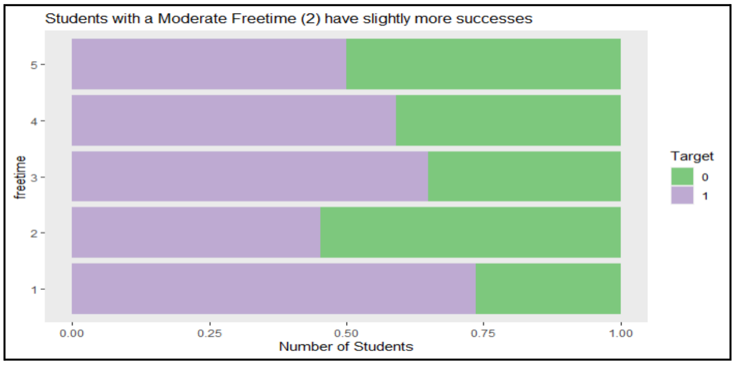


Exploration Findings

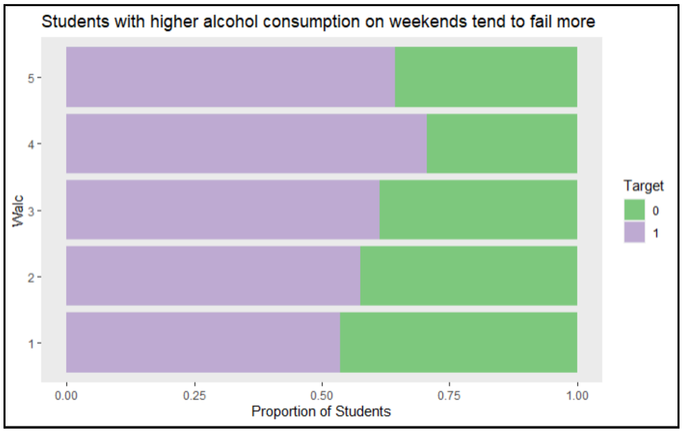
- Students who receive zeros increase significantly by period.



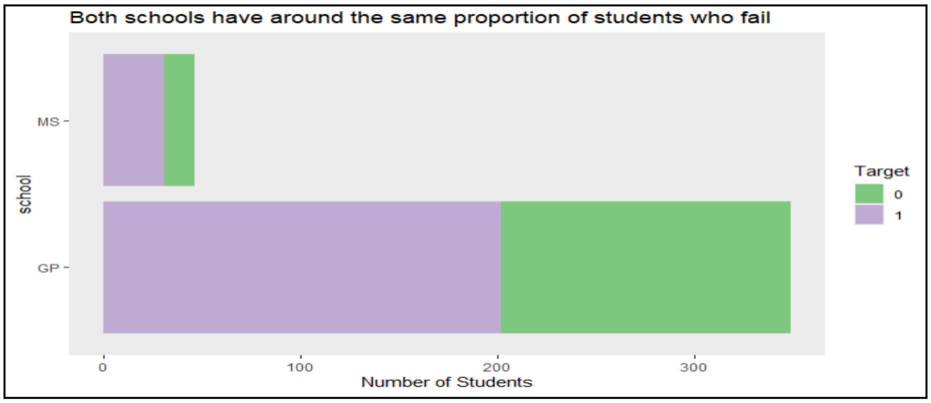
- Students with more freetime most likely are the ones not taking the class seriously.



- Direct increase in students who fail the class as their weekend alcohol consumption increases.



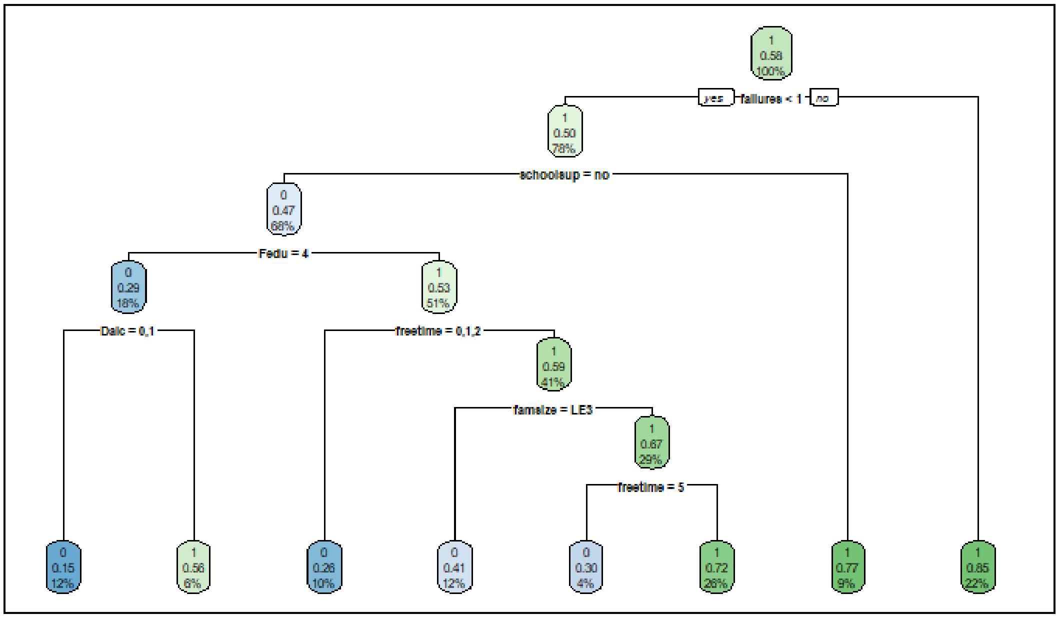
- Both schools have around the same proportion of students who fail, suggesting that the environment isn’t coming into play, the students are struggling with the course as a whole

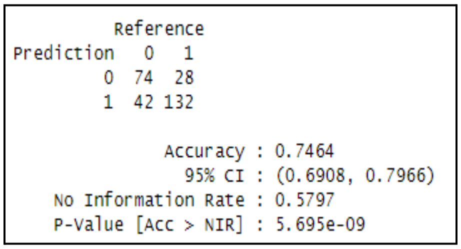


Model

Importance of variables checked before building in order to build the best models.

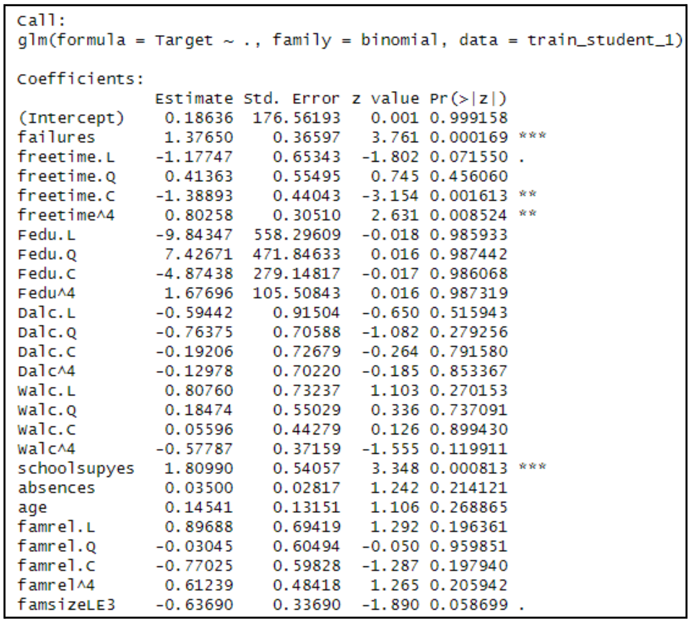
- Decision Tree





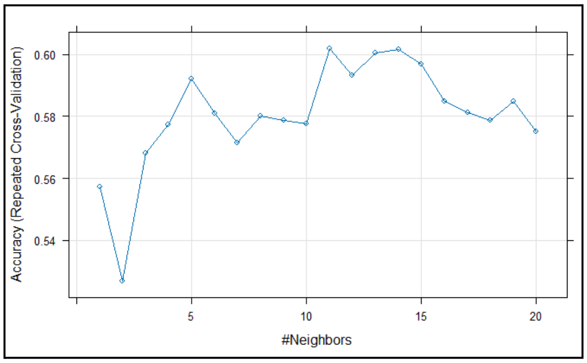
```
          Reference
Prediction   0   1
         0  74  28
         1  42 132

               Accuracy : 0.7464
                 95% CI : (0.6908, 0.7966)
    No Information Rate : 0.5797
    P-Value [Acc > NIR] : 5.695e-09
```

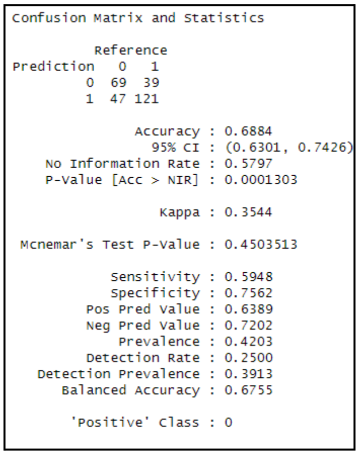
- Logistic Regression
    - Accuracy: 71.74%



```
Call:
glm(formula = Target ~ ., family = binomial, data = train_student_1)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.18636  176.56193   0.001 0.999158
failures       1.37650    0.36597   3.761 0.000169 ***
freetime.L    -1.17747    0.65343  -1.802 0.071550 .
freetime.Q     0.41363    0.55495   0.745 0.456060
freetime.C    -1.38893    0.44043  -3.154 0.001613 **
freetime^4     0.80258    0.30510   2.631 0.008524 **
Fedu.L        -9.84347  558.29609  -0.018 0.985933
Fedu.Q         7.42671  471.84633   0.016 0.987442
Fedu.C        -4.87438  279.14817  -0.017 0.986068
Fedu^4         1.67696  105.50843   0.016 0.987319
Dalc.L        -0.59442    0.91504  -0.650 0.515943
Dalc.Q        -0.76375    0.70588  -1.082 0.279256
Dalc.C        -0.19206    0.72679  -0.264 0.791580
Dalc^4        -0.12978    0.70220  -0.185 0.853367
Walc.L         0.80760    0.73237   1.103 0.270153
Walc.Q         0.18474    0.55029   0.336 0.737091
Walc.C         0.05596    0.44279   0.126 0.899430
Walc^4        -0.57787    0.37159  -1.555 0.119911
schoolsupyes   1.80990    0.54057   3.348 0.000813 ***
absences       0.03500    0.02817   1.242 0.214121
age            0.14541    0.13151   1.106 0.268865
famrel.L       0.89688    0.69419   1.292 0.196361
famrel.Q      -0.03045    0.60494  -0.050 0.959851
famrel.C      -0.77025    0.59828  -1.287 0.197940
famrel^4       0.61239    0.48418   1.265 0.205942
famsizeLE3    -0.63690    0.33690  -1.890 0.058699 .
```

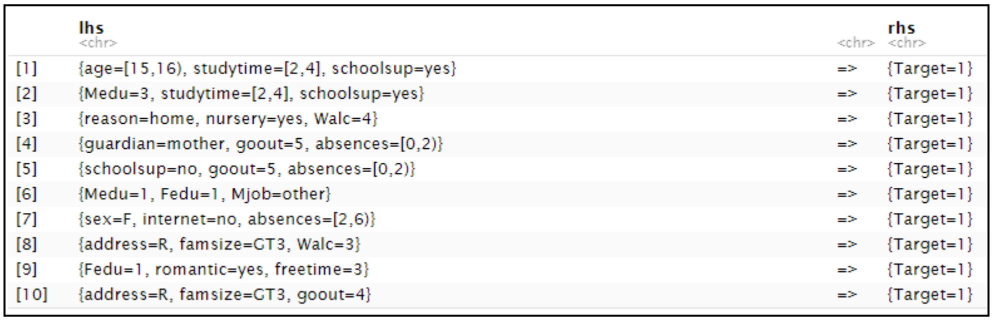
- kNN





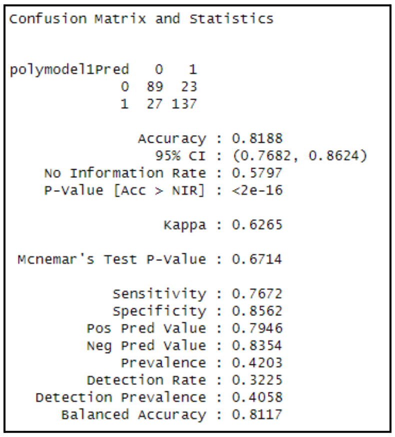
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  69  39
         1  47 121

               Accuracy : 0.6884
                 95% CI : (0.6301, 0.7426)
    No Information Rate : 0.5797
    P-Value [Acc > NIR] : 0.0001303

                  Kappa : 0.3544

 Mcnemar's Test P-Value : 0.4503513

            Sensitivity : 0.5948
            Specificity : 0.7562
         Pos Pred Value : 0.6389
         Neg Pred Value : 0.7202
             Prevalence : 0.4203
         Detection Rate : 0.2500
   Detection Prevalence : 0.3913
      Balanced Accuracy : 0.6755

       'Positive' Class : 0
```

- Association Rule Mining



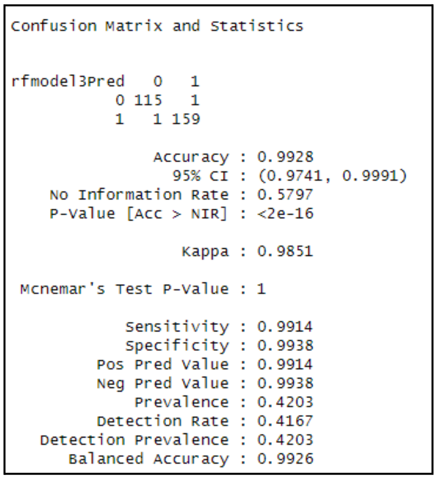
	lhs <chr>	<chr>	rhs <chr>
[1]	{age=[15,16), studytime=[2,4], schoolsup=yes}	=>	{Target=1}
[2]	{Medu=3, studytime=[2,4], schoolsup=yes}	=>	{Target=1}
[3]	{reason=home, nursery=yes, Walc=4}	=>	{Target=1}
[4]	{guardian=mother, goout=5, absences=[0,2)}	=>	{Target=1}
[5]	{schoolsup=no, goout=5, absences=[0,2)}	=>	{Target=1}
[6]	{Medu=1, Fedu=1, Mjob=other}	=>	{Target=1}
[7]	{sex=F, internet=no, absences=[2,6)}	=>	{Target=1}
[8]	{address=R, famsize=GT3, Walc=3}	=>	{Target=1}
[9]	{Fedu=1, romantic=yes, freetime=3}	=>	{Target=1}
[10]	{address=R, famsize=GT3, goout=4}	=>	{Target=1}

- SVM



```
Confusion Matrix and Statistics


polymodel1Pred   0   1
             0  89  23
             1  27 137

               Accuracy : 0.8188
                 95% CI : (0.7682, 0.8624)
    No Information Rate : 0.5797
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6265

 Mcnemar's Test P-Value : 0.6714

            Sensitivity : 0.7672
            Specificity : 0.8562
         Pos Pred Value : 0.7946
         Neg Pred Value : 0.8354
             Prevalence : 0.4203
         Detection Rate : 0.3225
   Detection Prevalence : 0.4058
      Balanced Accuracy : 0.8117
```

- Random Forest



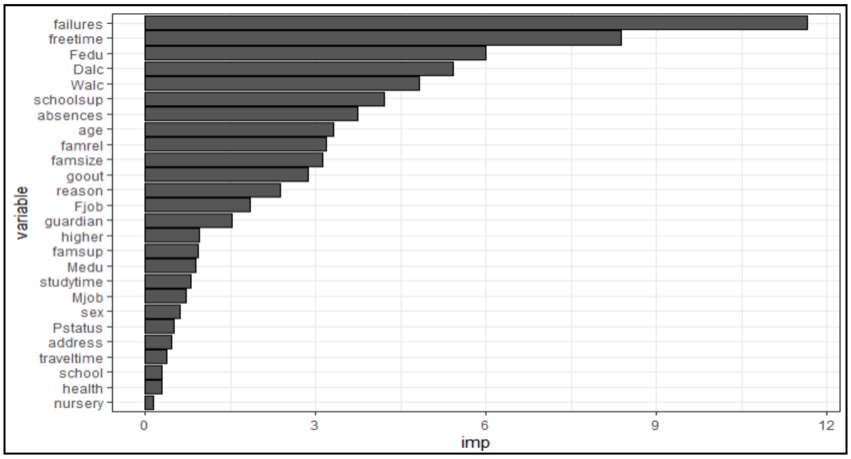
```
Confusion Matrix and Statistics


rfmodel3Pred   0   1
           0 115   1
           1   1 159

               Accuracy : 0.9928
                 95% CI : (0.9741, 0.9991)
    No Information Rate : 0.5797
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9851

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9914
            Specificity : 0.9938
         Pos Pred Value : 0.9914
         Neg Pred Value : 0.9938
             Prevalence : 0.4203
         Detection Rate : 0.4167
   Detection Prevalence : 0.4203
      Balanced Accuracy : 0.9926
```

Interpret

- Given the high accuracy levels, recommend the Random Forest model after assuring that no overfitting has taken place.
- The top 10 variables identified, the schools survey to students can be drastically reduced which will promote better completion rates and accuracy

Learning Objective Three

Visualization: Before beginning predictive models, instead of jumping in with all variables included, check the importance of each variable to see which ones have the most impact on the dependent variable. This can better be shown in a visualization, instead of looking at raw numbers.



Predictive Model(s): Images of Project Two's predictive models can be found under **Learning Objective Two**.

As and overview:

Model	Accuracy
No Information Rate	.5897
Decision Tree	.7464
Logistic Regression	.7174
kNN	.6884
SVM	.8188
Random Forest	.9928

Learning Objective Four

R-Studio, R, is used to complete the data science life cycle and provide key insights using data analysis, visualizations, modeling, etc as shown in the first three learning objectives. R-Studio was the beginning point for coding education. R is best suited for basic plots / graphics and various types of modeling including: decision trees, logistic regression, kNN, SVM, and random forest. The weaknesses of R include: slow speed, weak security, and poor data handling.

Note: I am currently in the process of learning different types of modeling in Python.

Learning Objective Five

Findings from analysis in R-Studio were communicated on a powerpoint presentation using visualizations produced from the R code to simplify findings. Exploratory data analysis was visualized in basic charts including box plots and bar charts. Models were visualized as kNN plot, Tree model, etc. Key numbers from modeling like accuracy value were pulled and compared so that the majority of people can understand percentages.

Learning Objective Six

Ethics were used in project two by applying what is personally known as a student and what factors in day to day life or academics impact student standings in school. For example, a finding was that students' grades decrease as the year progresses. Which makes sense because material tends to get more difficult. Another finding was that there is a direct increase in students who fail the class as their weekend alcohol consumption increases. More time partying equals less time focused on school.

Project Three

Learning Objectives: 1, 5, 6
IST 659- Group Project

Whole-ER Foods

The objective of this project was to demonstrate the ability to work in a team to design and implement a functional system with a database. Devise your own database to design and implement.

Learning Objective One

Collect: Data was hand collected and formatted into csv files manually using the assistance of https://www.mockaroo.com/ . Multiple csv files were created and formatted into tables.

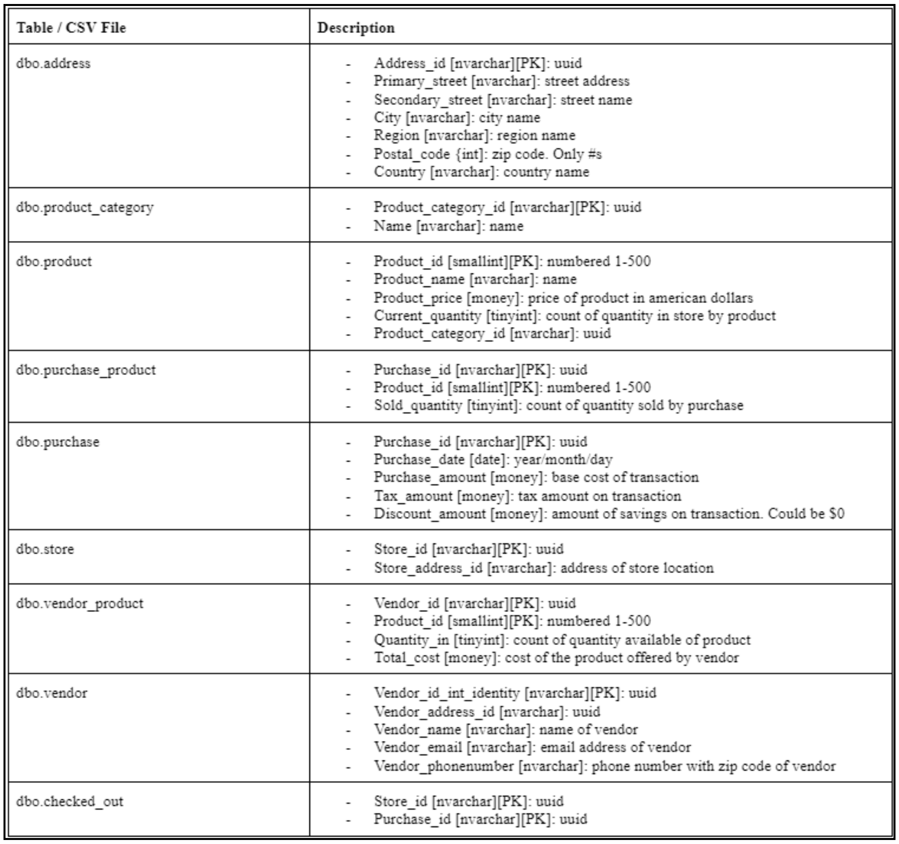
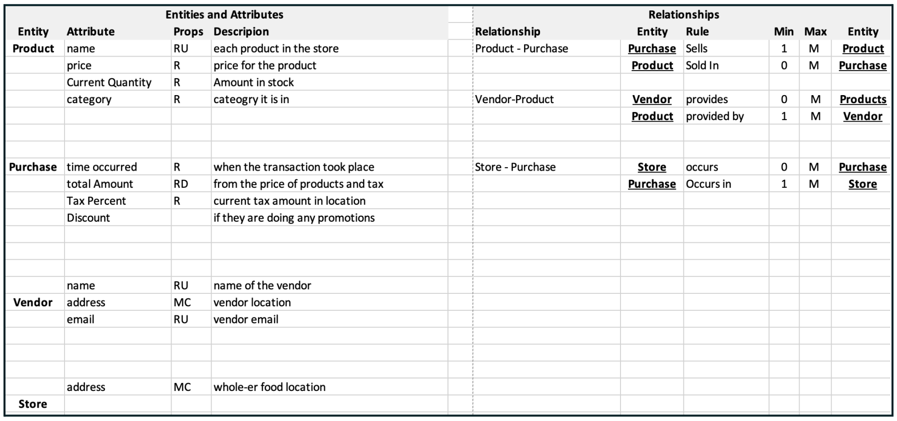


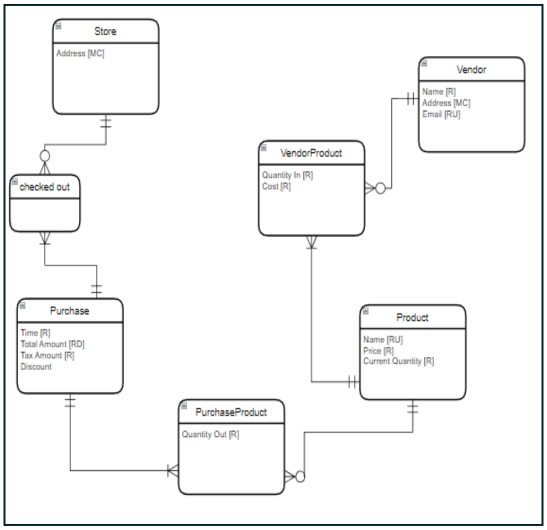
Table / CSV File	Description
dbo.address	- Address_id [nvarchar][PK]: uuid - Primary_street [nvarchar]: street address - Secondary_street [nvarchar]: street name - City [nvarchar]: city name - Region [nvarchar]: region name - Postal_code {int]: zip code. Only #s - Country [nvarchar]: country name
dbo.product_category	- Product_category_id [nvarchar][PK]: uuid - Name [nvarchar]: name
dbo.product	- Product_id [smallint][PK]: numbered 1-500 - Product_name [nvarchar]: name - Product_price [money]: price of product in american dollars - Current_quantity [tinyint]: count of quantity in store by product - Product_category_id [nvarchar]: uuid
dbo.purchase_product	- Purchase_id [nvarchar][PK]: uuid - Product_id [smallint][PK]: numbered 1-500 - Sold_quantity [tinyint]: count of quantity sold by purchase
dbo.purchase	- Purchase_id [nvarchar][PK]: uuid - Purchase_date [date]: year/month/day - Purchase_amount [money]: base cost of transaction - Tax_amount [money]: tax amount on transaction - Discount_amount [money]: amount of savings on transaction. Could be $0
dbo.store	- Store_id [nvarchar][PK]: uuid - Store_address_id [nvarchar]: address of store location
dbo.vendor_product	- Vendor_id [nvarchar][PK]: uuid - Product_id [smallint][PK]: numbered 1-500 - Quantity_in [tinyint]: count of quantity available of product - Total_cost [money]: cost of the product offered by vendor
dbo.vendor	- Vendor_id_int_identity [nvarchar][PK]: uuid - Vendor_address_id [nvarchar]: uuid - Vendor_name [nvarchar]: name of vendor - Vendor_email [nvarchar]: email address of vendor - Vendor_phonenumber [nvarchar]: phone number with zip code of vendor
dbo.checked_out	- Store_id [nvarchar][PK]: uuid - Purchase_id [nvarchar][PK]: uuid

Store: Relational Database made from start to finish to store data collected and establish relationships between each table.

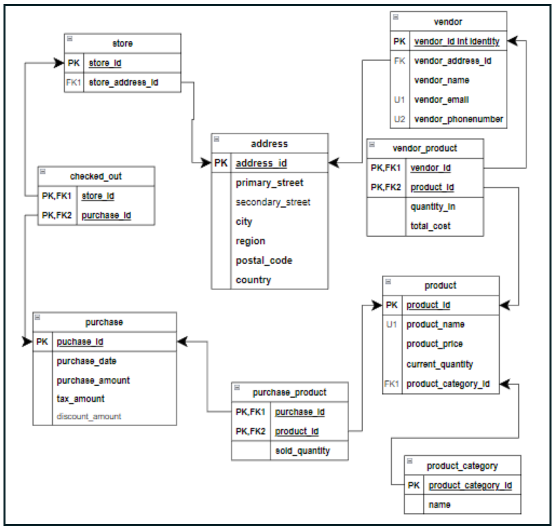


Entities and Attributes				Relationships					
Entity	**Attribute**	**Props**	**Descripion**	**Relationship**	**Entity**	**Rule**	**Min**	**Max**	**Entity**
Product	name	RU	each product in the store	Product - Purchase	**Purchase**	Sells	1	M	**Product**
	price	R	price for the product		**Product**	Sold In	0	M	**Purchase**
	Current Quantity	R	Amount in stock						
	category	R	cateogry it is in	Vendor-Product	**Vendor**	provides	0	M	**Products**
					Product	provided by	1	M	**Vendor**
Purchase	time occurred	R	when the transaction took place	Store - Purchase	**Store**	occurs	0	M	**Purchase**
	total Amount	RD	from the price of products and tax		**Purchase**	Occurs in	1	M	**Store**
	Tax Percent	R	current tax amount in location						
	Discount		if they are doing any promotions						
	name	RU	name of the vendor						
Vendor	address	MC	vendor location						
	email	RU	vendor email						
	address	MC	whole-er food location						
Store									

Conceptual Data Model



Logical Data Model


store
PK store_id
FK1 store_address_id

vendor
PK vendor_id int identity
FK vendor_address_id
vendor_name
U1 vendor_email
U2 vendor_phonenumber

address
PK address_id
primary_street
secondary_street
city
region
postal_code
country

checked_out
PK,FK1 store_id
PK,FK2 purchase_id

vendor_product
PK,FK1 vendor_id
PK,FK2 product_id
quantity_in
total_cost

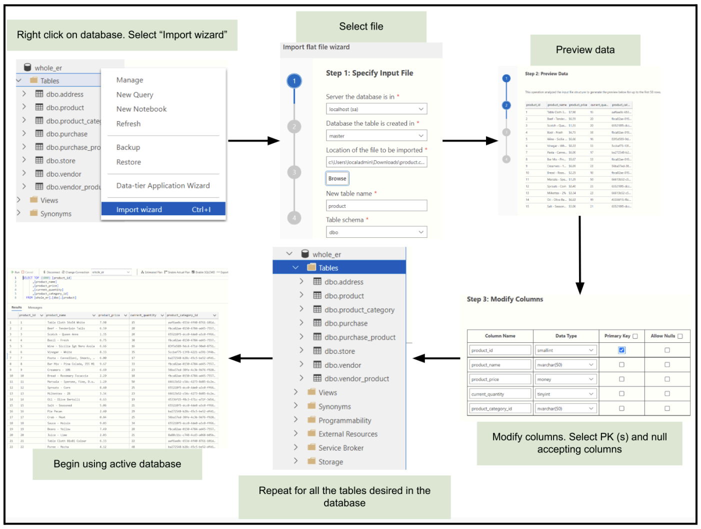
product
PK product_id
U1 product_name
product_price
current_quantity
FK1 product_category_id

purchase
PK puchase_id
purchase_date
purchase_amount
tax_amount
discount_amount

purchase_product
PK,FK1 purchase_id
PK,FK2 product_id
sold_quantity

product_category
PK product_category_id
name

Implementation


Right click on database. Select "Import wizard"

Select file

Import flat file wizard

Step 1: Specify Input File

Server the database is in * localhost (sa)

Database the table is created in * master

Location of the file to be imported * c:\Users\localadmin\Downloads\product.c...

Browse

New table name * product

Table schema * dbo

Preview data

Step 2: Preview Data

Step 3: Modify Columns

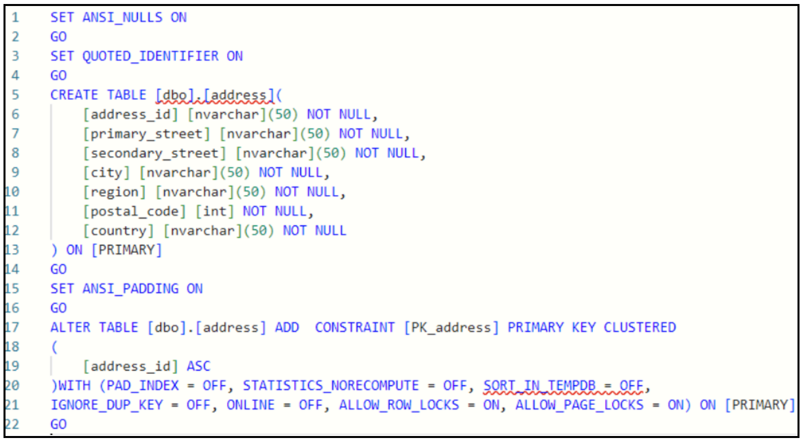
Column Name	Data Type	Primary Key ☐	Allow Nulls ☐
product_id	smallint	☑	☐
product_name	nvarchar(50)	☐	☐
product_price	money	☐	☐
current_quantity	tinyint	☐	☐
product_category_id	nvarchar(50)	☐	☐

Modify columns. Select PK (s) and null accepting columns

Repeat for all the tables desired in the database

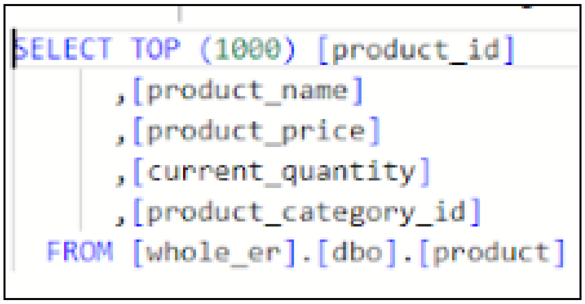
```
SELECT TOP (1000) [product_id]
      ,[product_name]
      ,[product_price]
      ,[current_quantity]
      ,[product_category_id]
  FROM [whole_er].[dbo].[product]
```

Begin using active database



```
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[address](
    [address_id] [nvarchar](50) NOT NULL,
    [primary_street] [nvarchar](50) NOT NULL,
    [secondary_street] [nvarchar](50) NOT NULL,
    [city] [nvarchar](50) NOT NULL,
    [region] [nvarchar](50) NOT NULL,
    [postal_code] [int] NOT NULL,
    [country] [nvarchar](50) NOT NULL
) ON [PRIMARY]
GO
SET ANSI_PADDING ON
GO
ALTER TABLE [dbo].[address] ADD  CONSTRAINT [PK_address] PRIMARY KEY CLUSTERED
(
    [address_id] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, SORT_IN_TEMPDB = OFF,
IGNORE_DUP_KEY = OFF, ONLINE = OFF, ALLOW_ROW_LOCKS = ON, ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
GO
```

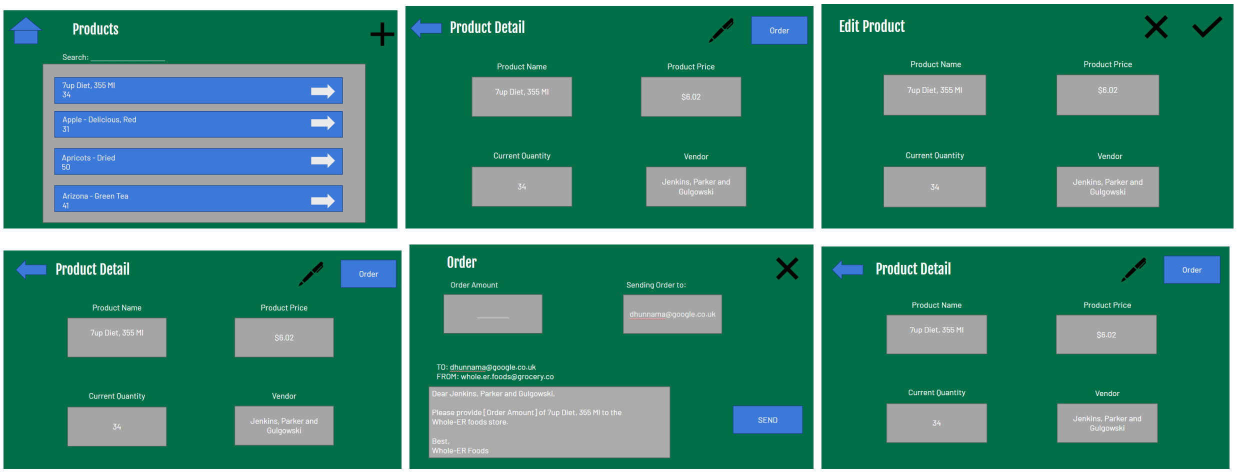
Access: All data implemented into the database can be accessed through SQL queries including but not limited to: select, update, delete, and insert.



```
SELECT TOP (1000) [product_id]
      ,[product_name]
      ,[product_price]
      ,[current_quantity]
      ,[product_category_id]
  FROM [whole_er].[dbo].[product]
```

Learning Objective Five

The building of a relational database was communicated in a powerpoint presentation. Visualizations were made manually in order to successfully communicate the process and organization of the database. SQL code was used in the assignment and no visualizations were made using the code. So in order to communicate with a range of audiences, a step by step beautified mockup was presented.



Products

Search: ___________

7up Diet, 355 Ml
34

Apple - Delicious, Red
31

Apricots - Dried
50

Arizona - Green Tea
41

Product Detail

Product Name: 7up Diet, 355 Ml
Product Price: $6.02
Current Quantity: 34
Vendor: Jenkins, Parker and Gulgowski

Edit Product

Product Name: 7up Diet, 355 Ml
Product Price: $6.02
Current Quantity: 34
Vendor: Jenkins, Parker and Gulgowski

Product Detail

Product Name: 7up Diet, 355 Ml
Product Price: $6.02
Current Quantity: 34
Vendor: Jenkins, Parker and Gulgowski

Order

Order Amount: ______
Sending Order to: dhunnama@google.co.uk

TO: dhunnama@google.co.uk
FROM: whole.er.foods@grocery.co

Dear Jenkins, Parker and Gulgowski,

Please provide [Order Amount] of 7up Diet, 355 Ml to the Whole-ER foods store.

Best,
Whole-ER Foods

Product Detail

Product Name: 7up Diet, 355 Ml
Product Price: $6.02
Current Quantity: 34
Vendor: Jenkins, Parker and Gulgowski

Learning Objective Six

Ethics were key in project three in order to build a relational database. The database kept track of sales and inventory. To be successful in building the database from the bottom up, ethics were used to analyze all variables and steps that could possibly go into getting an apple from a tree, into a store, and back to a kitchen. These business ethics include: accountability, responsibility, choice, relationship, society behavior, etc.

Project Four

Learning Objectives: 1, 2, 3, 5, 6
MBC 638- Individual Project

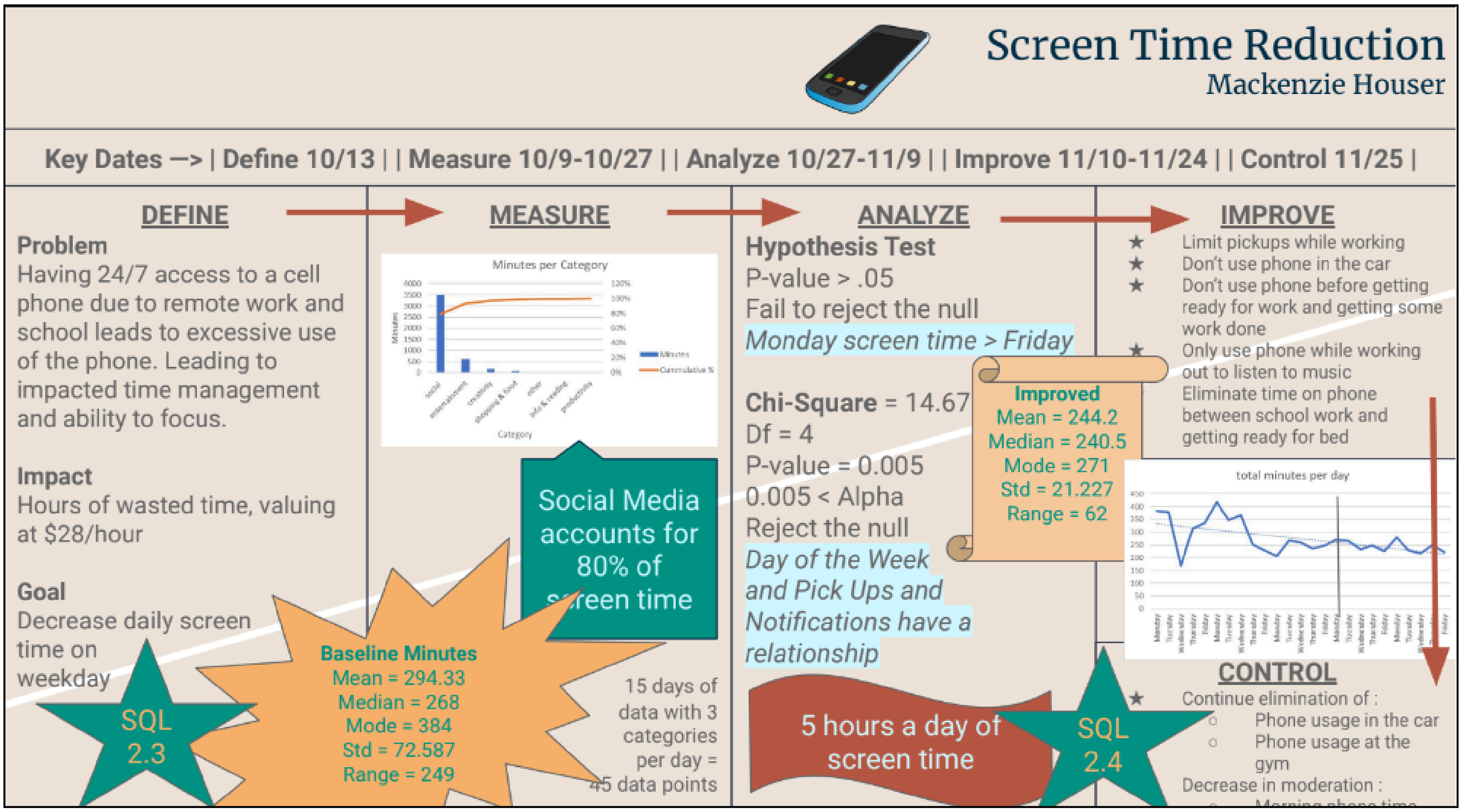
Process Improvement

The objective of this project was to select an issue or opportunity that can be written as a problem. Use data that is accessible to you or can be collected in a reasonable amount of effort/time. Fixing this problem will provide value; develop a business case to support working this issue.

Learning Objective One

Collect: Data manually collected from iPhone Phone Usage in settings.
Store: All data was stored in Microsoft Excel on local pc.
Access: Since data was collected personally, export access was not necessary. Access was automatic.

Learning Objective Two



Screen Time Reduction
Mackenzie Houser

Key Dates —> | Define 10/13 | | Measure 10/9-10/27 | | Analyze 10/27-11/9 | | Improve 11/10-11/24 | | Control 11/25 |

DEFINE

Problem
Having 24/7 access to a cell phone due to remote work and school leads to excessive use of the phone. Leading to impacted time management and ability to focus.

Impact
Hours of wasted time, valuing at $28/hour

Goal
Decrease daily screen time on weekday

SQL 2.3

MEASURE

Social Media accounts for 80% of screen time

Baseline Minutes
Mean = 294.33
Median = 268
Mode = 384
Std = 72.587
Range = 249

15 days of data with 3 categories per day = 45 data points

ANALYZE

Hypothesis Test
P-value > .05
Fail to reject the null
Monday screen time > Friday

Chi-Square = 14.67
Df = 4
P-value = 0.005
0.005 < Alpha
Reject the null
Day of the Week and Pick Ups and Notifications have a relationship

Improved
Mean = 244.2
Median = 240.5
Mode = 271
Std = 21.227
Range = 62

5 hours a day of screen time

SQL 2.4

IMPROVE

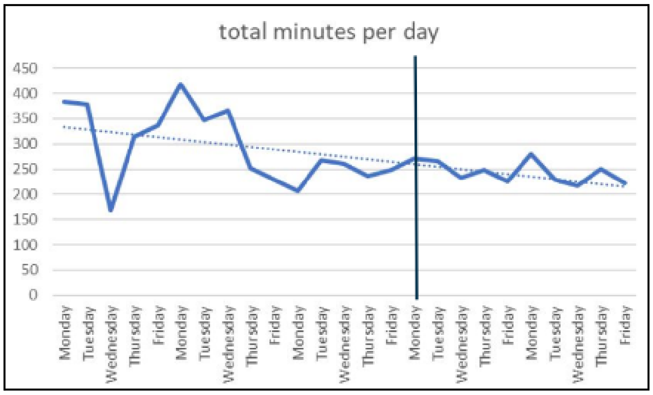
★ Limit pickups while working
★ Don't use phone in the car
★ Don't use phone before getting ready for work and getting some work done
★ Only use phone while working out to listen to music
Eliminate time on phone between school work and getting ready for bed

CONTROL

★ Continue elimination of :
○ Phone usage in the car
○ Phone usage at the gym
Decrease in moderation :
○ Morning phone time

Learning Objective Three

Visualization / Predictive Model: Sometimes visualizations and predictive models can be combined into one, like with a time series. In this project, it was analyzed over time, so the time series was a great predictive model to see progress of phone usage over three weeks.



Learning Objective Five

Excel was used for data, exploration, and modeling. Communicating insights to a wide range of audiences using visualizations was simplified to plots and time series. Basic data was communicated through the improvement of calculated statistics like mean, median, mode, etc. Lastly, findings were communicated in the form of null hypothesis.

Learning Objective Six

Project four was all about ethics because it tracked personal behavior. The entire project was a time series of trying to improve phone usage during weekdays. So using ethics was a way to pinpoint the inappropriate times to be using a cell phone like while driving a car or at work. Honesty and integrity were also key ethics used during this process because I had to hold myself accountable in order to collect accurate data.

Strengths

- Visualizations
    - In the Data Science program as a Business Analytics focusing student, visualizations are a strength because it is the best way to present findings to everyone with different educational backgrounds.
- Problem Solving
    - The entire data science life cycle is based upon solving problems. The first step in every analysis is to identify why the analysis is being performed and what the solution resolves.
- Identifying Issues
    - Identifying issues is the result of attention to detail. This is important because if there is an underlying problem, information and studies might not be accurate. An example of this could be dirty and missing data, coding errors, or misclassification.

Challenges

- Program to Program
    - Performing similar tasks in multiple programming languages such as Python and R-Studio can become complicated due to the different syntax. R-Studio was taught first and Python has been learned most recently. Functions and capabilities often get mixed up. A solution to this is to focus on mastery in one programming language or become fluent in writing both languages.
- Simplifying Code
    - The thing about writing code is that it can be done in a multitude of ways and arrive at the same conclusions. Every data scientist has their own writer's voice, just like in literature. The challenging part is sometimes the longer way of writing code is taught before the simplified version. The more extensive writing often becomes natural but the simplified is more efficient.

Future Goals

- Maintain data science skills by currently applying education in day to day work.
- Improve data science skills by taking on new challenges.
- Collaborate with professional teams to find significant data to solve problems and improve functionality.