Adv Data Prog with R Assignment 1

16343261

26/6/2022

Contents

Introduction

Before we begin this assignment, I will load in the data and packages as required.

```
library(nycflights13)
library(dplyr)
library(plyr)
library(magrittr)
library(ggplot2)
library(cowplot)
setwd("/Users/matth/Documents/Stat Network Analysis")
library(readxl)
library(corrplot)
library(ggcorrplot)
library(ggstance)
library(jtools)
```

```
data(flights)
```

Part 1

Question 1.

To answer this question, we will use the which function to only select flights with origin = JFK and destination = LAX.

```
flights_2 <- flights[which(flights$origin == 'JFK'
& flights$dest == 'LAX'),]
```

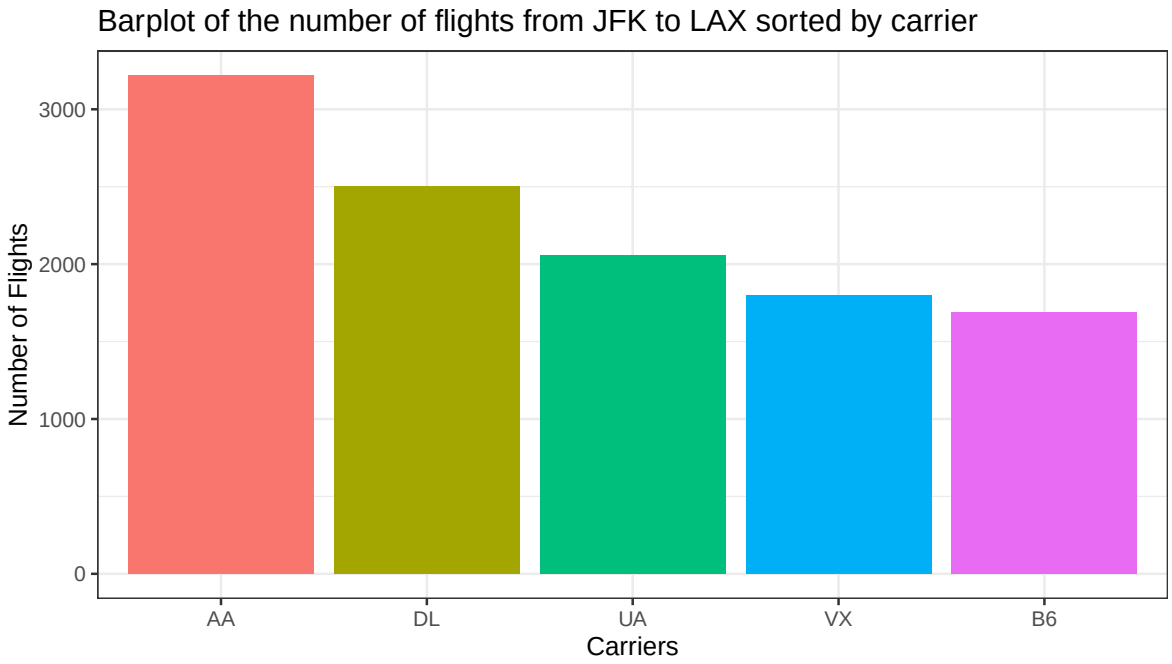
We will use the as.factor function to encode the carrier variable as a factor. Subsequently, I will use the levels and list function to encode the levels as requested.

```
flights_2$carrier <- as.factor(flights_2$carrier)
levels(flights_2$carrier) <- list(AA = 'AA', DL = 'DL', UA = 'UA', VX = 'VX', B6 = 'B6')
```

Question 2.

To aid with this question, I will be using magrittr, table and as.data.frame functions to create a frequency table data frame. I will recast the variables using the code in the previous question and create the barplot as requested.

```
counts <- flights_2$carrier %>% table %>% as.data.frame
counts$. <- as.factor(counts$.)
levels(counts$.) <- list(AA = 'AA', DL = 'DL', UA = 'UA', VX = 'VX', B6 = 'B6') # Initialising factors.
ggplot(counts, aes(x=., y = Freq, fill = .)) + # Colour each bar.
  geom_bar(stat="identity")+ # Make Barplot.
  theme_bw() + # Set theme.
  theme(legend.position = "none")+ # No legend needed.
  xlab('Carriers') + # X axis name.
  ylab('Number of Flights') + # Y axis name.
  ggtitle('Barplot of the number of flights from JFK to LAX sorted by carrier') # Title of plot.
```



Here we see our results of flights from JFK to LAX as follows:

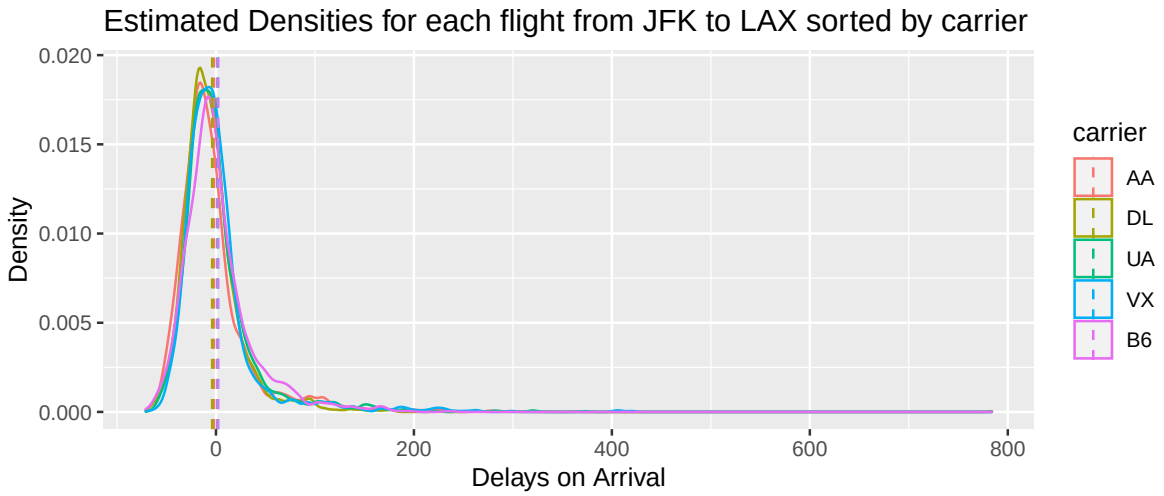
1. American Airlines (AA) has the highest number of flights with over 3000.
2. Delta Airlines (DL) has around 2500 flights.
3. United Airlines (UA) has just more than 2000 flights.
4. Virgin Airlines (VX) has around 1800 flights.
5. Jet Blue Airlines (B6) has around 1650 flights.

The following airline code information was found at: https://en.wikipedia.org/wiki/List_of_airline_codes_(V)

Question 3.

To answer this question, we will use the **tapply** function to calculate the average delay at arrival for each carrier. I will also be using the **ddply** function from the **dplyr** package to form the mean line for our plots. Lastly, we will be using the geom_density and geom_vline functions to construct the requested plots.

```
av_arr_delay <- tapply(flights_2$arr_delay, flights_2$carrier, mean, na.rm =TRUE) %>% as.data.frame
# Magrittr used as described.
mu <- ddply(flights_2, "carrier", summarise, grp.mean=mean(arr_delay, na.rm = TRUE))
# Creating Mean data set as described.
ggplot(flights_2, aes(x=arr_delay, color=carrier)) +
  # Color by carrier.
  geom_density()+ # Create density plot.
  geom_vline(data=mu, aes(xintercept=grp.mean,
  color=carrier),linetype="dashed")+ # Adding mean line.
  xlab('Delays on Arrival') +
  ylab('Density') +
  ggtitle('Estimated Densities for each flight from JFK to LAX sorted by carrier')
```

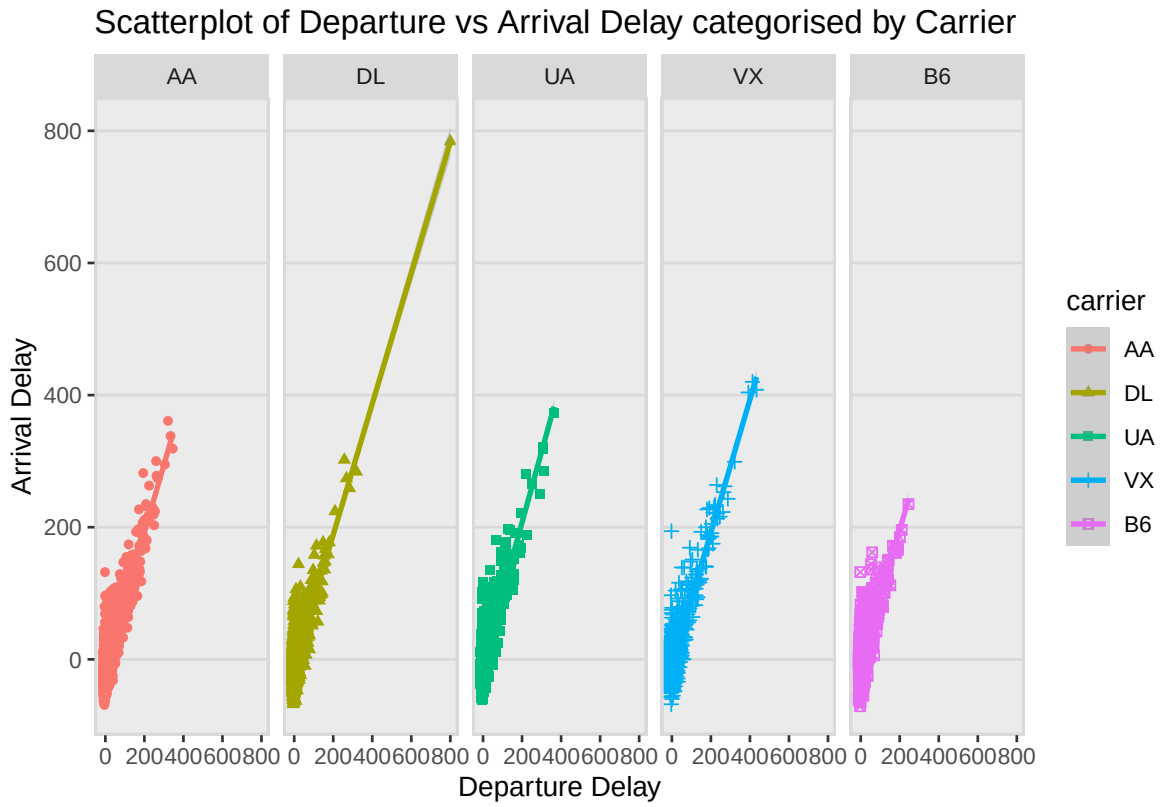


We find that most Flights from JFK to LAX have little to no Delays on Arrival. Here we see a right skewed distribution where each airline is relatively overlapping in their distributions with long tails exhibiting uncommon outliers. For example Jet Blue Airlines having a single delay on arrival of around 800 is extending it's tail considerably far out. It is somewhat difficult to interpret the end of our tail as our lines are overlapping.

Question 4.

To contruct this scatterplot we will use the *geom_point* function. To seperate our scatterplots by carrier, we will use the *facet_grid* function across the carrier variable.

```
ggplot(flights_2, aes(dep_delay, arr_delay, shape=carrier, color=carrier)) +
    geom_point() +
    facet_grid(. ~ carrier) + # Seperate plot by carrier.
    stat_smooth(method = "lm") +  # Include line of best fit.
    background_grid(major = 'y', minor = "none") +  # Include background grid.
    panel_border() + # Include panel border.
    xlab('Departure Delay') +
    ylab('Arrival Delay') +
    ggtitle('Scatterplot of Departure vs Arrival Delay categorised by Carrier')
```



Here we see a strong positive relationship between departure and arrival delays, meaning each positive increase in departure delay results on average in a positive increase in arrival delay, which is be expected. Our values including outlier values all seem relatively close to our lines of best fit. Intuitively we would expect a flight delayed at departure to be delayed at arriving at it's destination. This appears to be a very similar relationship for all airline carriers.

Question 5.

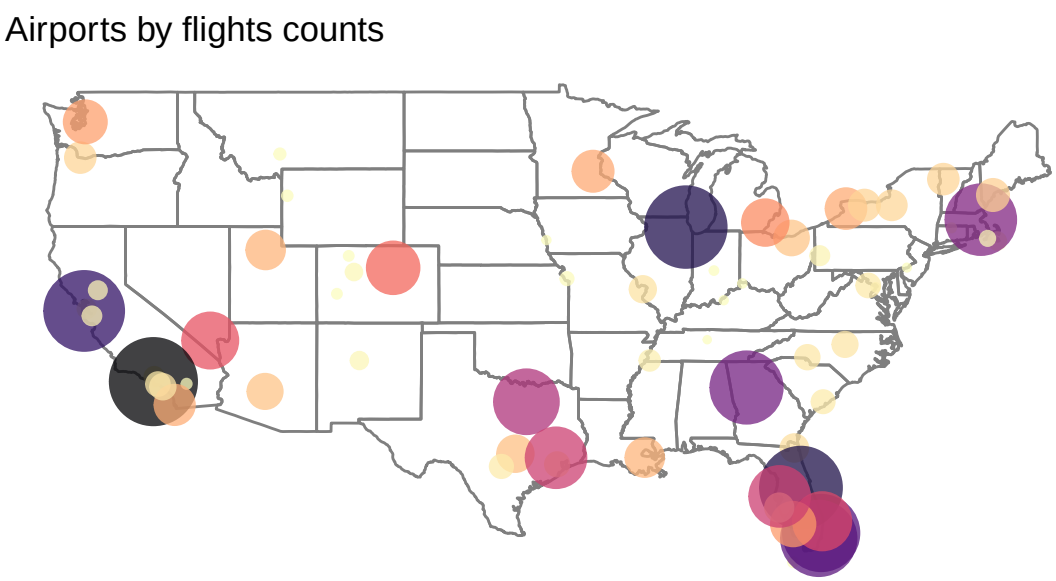
Comments added.

```
library(maps) # Load in the maps package
```

```
airport_info = flights %>% # Use magrittr on the flights dataset.
  subset(carrier %>% is_in(c('AA', 'DL', 'UA', 'VX', 'B6'))) %>%
  # Subset the flights dataset into a new dataset with exclusively the top 6 carriers.
  aggregate(year ~ dest, ., length) %>%
  # aggregate the number of times each destination occurs in the year in this new dataset.
  set_colnames(c('name','counts'))
  # Set the column names of this new dataset to "names" and "count".
```

```
N = nrow(airport_info) # Set N to the number of rows of airport_info.
airport_info$lon = rep(NA,N) # Create new columns lon and lat.
airport_info$lat = rep(NA,N)
for (i in 1:N) {
  index = which(airports$faa == airport_info$name[i])
  # For each element of the airport info name variable equivalant to the faa variable,
  if (length(index) != 0) {
    airport_info$lon[i] = airports$lon[index]
    airport_info$lat[i] = airports$lat[index]
  # If the length isn't 0, set the lon and lat variable values to those in the airport datasets.
  }
}
```

```
library(viridis) # Load viridis package.
ggplot(airport_info, aes(x = lon, y = lat, col = counts, alpha = I(0.75))) +
  # Create ggplot of lon and lat values of the airport destinations.
  borders('state') + #Using the United States as a map.
  geom_point(aes(size = counts)) + # Set the points size by counts value
  theme_void() + # Use theme void
  coord_cartesian(xlim = c(-125, -65), ylim = c(25, 50)) + # Set cartesian coordinates
  scale_color_viridis(option = 'A', direction = -1) + # Set colour to option A
  scale_size_continuous(range = c(1,15)) + # Scale the size of the values by ranges 1-15.
  theme(legend.position='none') + # No legend.
  ggtitle('\t\tAirports by flights counts') # Set title to "Airports by flights counts.
```



Part 2

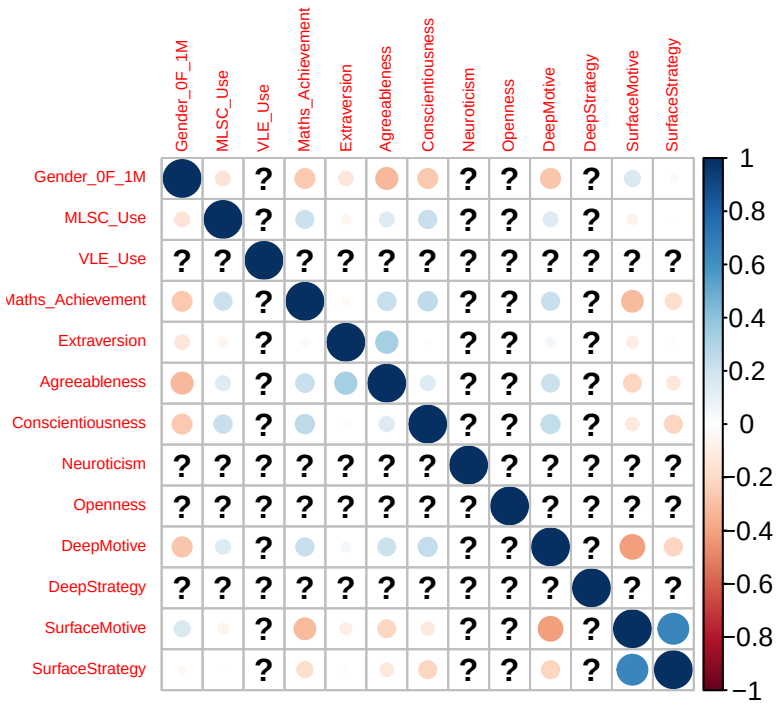
To load in our dataset, we will use the **read_excel** function from the readxl package.

```
MathsEd <- read_excel("/Users/matth/Documents/Advanced R/Maths Ed.xlsx")
```

Correlation Plot

The ggcorrplot package uses ggplot2 to help us better visualise a correlation matrix. This plot allows us to easily identify what variables are correlated with Maths Achievement Scores.

```
# Corrplot function used with circle method, smaller font and removing n/a values.
Matrix <-cor(MathsEd) # Calculate Correlation Matrix for our entire dataset.
corrplot(Matrix, method="circle", tl.cex = 0.5, na.rm = TRUE)
```



Here we find a number of positive correlations, most notably:

- MLSC_USE
- Agreeableness
- Conscientiousness
- Deep Motive

Negative correlations with:

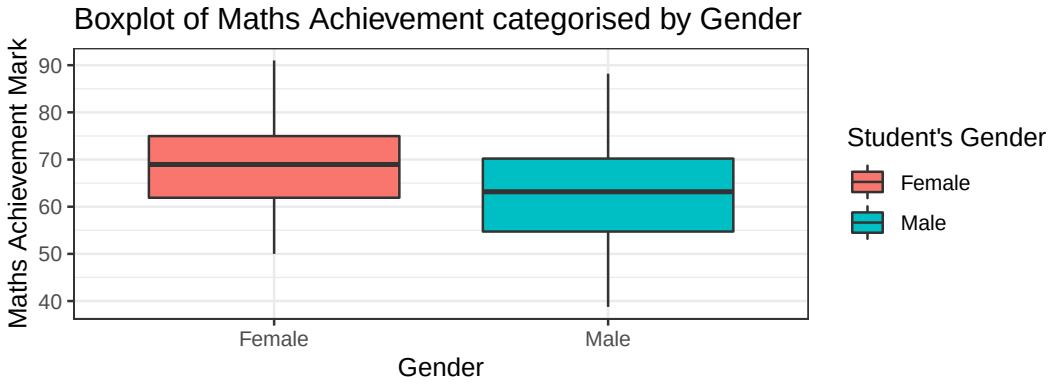
- Gender_0F_1M (This may imply a correlation for females)
- Surface Motive

Gender

To analyse whether gender plays a role in Maths Achievement scores, I will now create a boxplot of Maths Achievement Scores seperated by gender. To do this, I will create a new column seperating gender into a categorical variable, as opposed to a binary variable.

```
MathsEd$Gender <- factor(MathsEd$Gender_0F_1M, levels=c(0,1), labels=c("Female", "Male"))
# New categorical column with male and female names.
```

```
ggplot(MathsEd, aes(x = Gender, y = Maths_Achievement)) +
  geom_boxplot(aes(fill = Gender)) + # Create Boxplot
  labs(fill="Student's Gender")+ # Colour by Gender
  theme_bw() + # Black/White background
  theme(plot.margin = unit(c(1,1,1,1),"cm"))+ # Set margins
  xlab('Gender') +
  ylab('Maths Achievement Mark') +
  ggtitle('Boxplot of Maths Achievement categorised by Gender')
```



Here we find our boxplot between both genders. It would appear that being female gives one a higher Maths Achievement Score on average, as females have a higher median line. From our sample of 89 students, Male scores did not even exceed the 90% mark unlike females. Also, presuming a failing grade is 40%, it would appear that no females had a failing Maths Achevement score, whilst males in our sample have received failing grades.

Big Five Personality Factors regression

Next I will fit a linear regression model with Maths Achievement as our response variable and scores of our big five personality factors as our explanatory variables.

```
fit <- lm(Maths_Achievement ~ Extraversion + Agreeableness + Conscientiousness + Neuroticism
          + Openness, data = MathsEd) # Fitting model.
summary(fit) # Print summary of model.
```

```
Call:
lm(formula = Maths_Achievement ~ Extraversion + Agreeableness +
    Conscientiousness + Neuroticism + Openness, data = MathsEd)
```

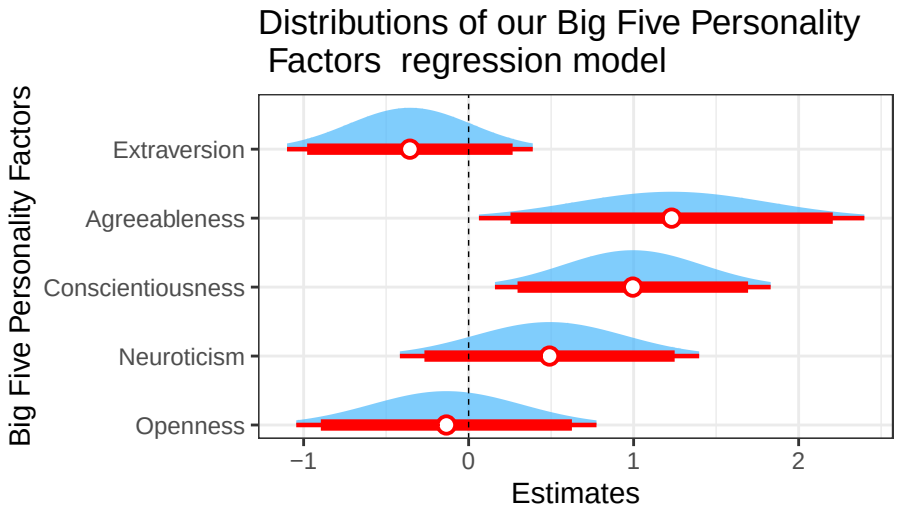
```
Residuals:
     Min       1Q   Median       3Q      Max
-27.2766  -5.7033  -0.8706   6.0365  23.4328

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       33.7451    12.0690   2.796  0.00649 **
Extraversion      -0.3560     0.3741  -0.952  0.34422
Agreeableness      1.2303     0.5868   2.097  0.03922 *
Conscientiousness  0.9950     0.4197   2.371  0.02020 *
Neuroticism        0.4903     0.4556   1.076  0.28515
Openness          -0.1348     0.4571  -0.295  0.76877
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.44 on 79 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.1538,    Adjusted R-squared:  0.1003
F-statistic: 2.872 on 5 and 79 DF,  p-value: 0.01958
```

We will further investigate this using the **plot_summs** function from the **jtools** create regression coefficient plots with ggplot2. This will allow us to better understand our regression model.

```
plot_summs(fit, plot.distributions = TRUE, inner_ci_level = .9,
  legend.title = "Big Five Personality Factors")+
  #Include plot distributions with 90% confidence intervals.
  # Title of legend
  theme_bw() +
  theme(plot.margin = unit(c(0.5,0.5,0.5,0.5),"cm"))+
  xlab('Estimates') +
  ylab('Factors') +
  ggtitle('Distributions of our Big Five Personality \n Factors  regression model')+
  scale_colour_manual(values = c("red", "blue")) # Manually scale colour red and blue.
```



Here we find the distributions of our Big Five Personality Factors and their respective distributions. Supporting our corrplot from earlier, we find Agreeableness and Conscientiousness have the highest positive affect on Maths Achievement scores. Extroversion and Openness had a negative affect. However this is not statistically significant from zero. Our summary finds Agreeableness and Conscientiousness cause a statistically significant increase in Maths Achievement for a significance level of of 0.05. It may be reasonable to assume that females who exhibit agreeable and conscientious traits have the highest likelihood of receiving high Maths Achievment Scores on average.