Stat 305 Project Part A

Ibrahim Alwishah

11/10/2022

Teammates' Names: Ted Zybin, Rohit Nair

Research Question:

How do the provided variables associate with the prediction that a customer is high value?

External Requirements: Data Read-In and Package Loading

```
df = read.csv("Insurance_policy.csv")
```

```
library(ggplot2)
library(dplyr)
library(mosaic)
library(ggplot2)
```

```
insurance_policy = df


# There are 48842 rows and 8 columns.
dim(insurance_policy)
```

```
## [1] 48842     8
```

```
# The summary function has shown the min,
#max, median, and mean of the variables but #there does not seem to be any missing values.

summary(insurance_policy)
```

```
##       age         education_num   marital_status      occupation
##  Min.   :17.00   Min.   : 1.00   Length:48842       Length:48842
##  1st Qu.:28.00   1st Qu.: 9.00   Class :character   Class :character
##  Median :37.00   Median :10.00   Mode  :character   Mode  :character
##  Mean   :38.64   Mean   :10.08
##  3rd Qu.:48.00   3rd Qu.:12.00
##  Max.   :90.00   Max.   :16.00
##     cap_gain     hours_per_week      score        value_flag
##  Min.   :    0   Min.   : 1.00   Min.   :43.94   Length:48842
```

```
##  1st Qu.:    0   1st Qu.:40.00   1st Qu.:57.50   Class :character
##  Median :    0   Median :40.00   Median :60.24   Mode  :character
##  Mean   : 1079   Mean   :40.42   Mean   :60.23
##  3rd Qu.:    0   3rd Qu.:45.00   3rd Qu.:62.95
##  Max.   :99999   Max.   :99.00   Max.   :76.53
```

```
# This has confirmed that there are no missing variables.
colSums( is.na(insurance_policy)  )
```

```
##            age  education_num marital_status     occupation       cap_gain
##              0              0              0              0              0
## hours_per_week          score     value_flag
##              0              0              0
```

```
# All 48842 rows are complete there are no missing NAs.
sum( complete.cases(insurance_policy) )
```

```
## [1] 48842
```

```
# 0% of the rows are incomplete.
sum( !complete.cases(insurance_policy)  )
```

```
## [1] 0
```

Dataset Description

Describe the data set:

The data set has 48842 rows and 8 columns, above it also shows the mean and max of different categories using the 'summary' function. The data set also does not have NAs or incomplete rows.

Data Transformation

```
# check that variables you think should be factors are factors!
dg = insurance_policy
str(dg)
```

```
## 'data.frame':    48842 obs. of  8 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital_status: chr  "Never-married" "Married-civ-spouse" "Divorced" "Married-civ-spouse" ...
##  $ occupation    : chr  "Group 2" "Group 5" "Group 1" "Group 1" ...
##  $ cap_gain      : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ score         : num  59 55.8 62.8 60.1 53.3 ...
##  $ value_flag    : chr  "Low" "Low" "Low" "Low" ...
```

```
# It is giving me age and education number as integers

# Attempting to make age and education number proper factors below using 'as.factor'.

dg$age = as.factor(dg$age)
dg$education_num = as.factor(dg$education_num)
str(dg)
```

```
## 'data.frame':    48842 obs. of  8 variables:
##  $ age           : Factor w/ 74 levels "17","18","19",..: 23 34 22 37 12 21 33 36 15 26 ...
##  $ education_num : Factor w/ 16 levels "1","2","3","4",..: 13 13 9 7 13 14 5 9 14 13 ...
##  $ marital_status: chr  "Never-married" "Married-civ-spouse" "Divorced" "Married-civ-spouse" ...
##  $ occupation    : chr  "Group 2" "Group 5" "Group 1" "Group 1" ...
##  $ cap_gain      : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ score         : num  59 55.8 62.8 60.1 53.3 ...
##  $ value_flag    : chr  "Low" "Low" "Low" "Low" ...
```

Exploratory Data Analysis: Descriptive Statistics and Visualizations

1)

```
insuranceNum <- insurance_policy %>%
  select(age, education_num, cap_gain, hours_per_week, score)


favstats_vec = c()
columns = colnames(insuranceNum)

total_favstats = data.frame()

for (i in 1:ncol(insuranceNum)){
  total_favstats <- rbind(total_favstats, favstats(insuranceNum[,i]))

}

total_favstats <- cbind(names = columns, total_favstats)
rownames(total_favstats) <- NULL

total_favstats
```

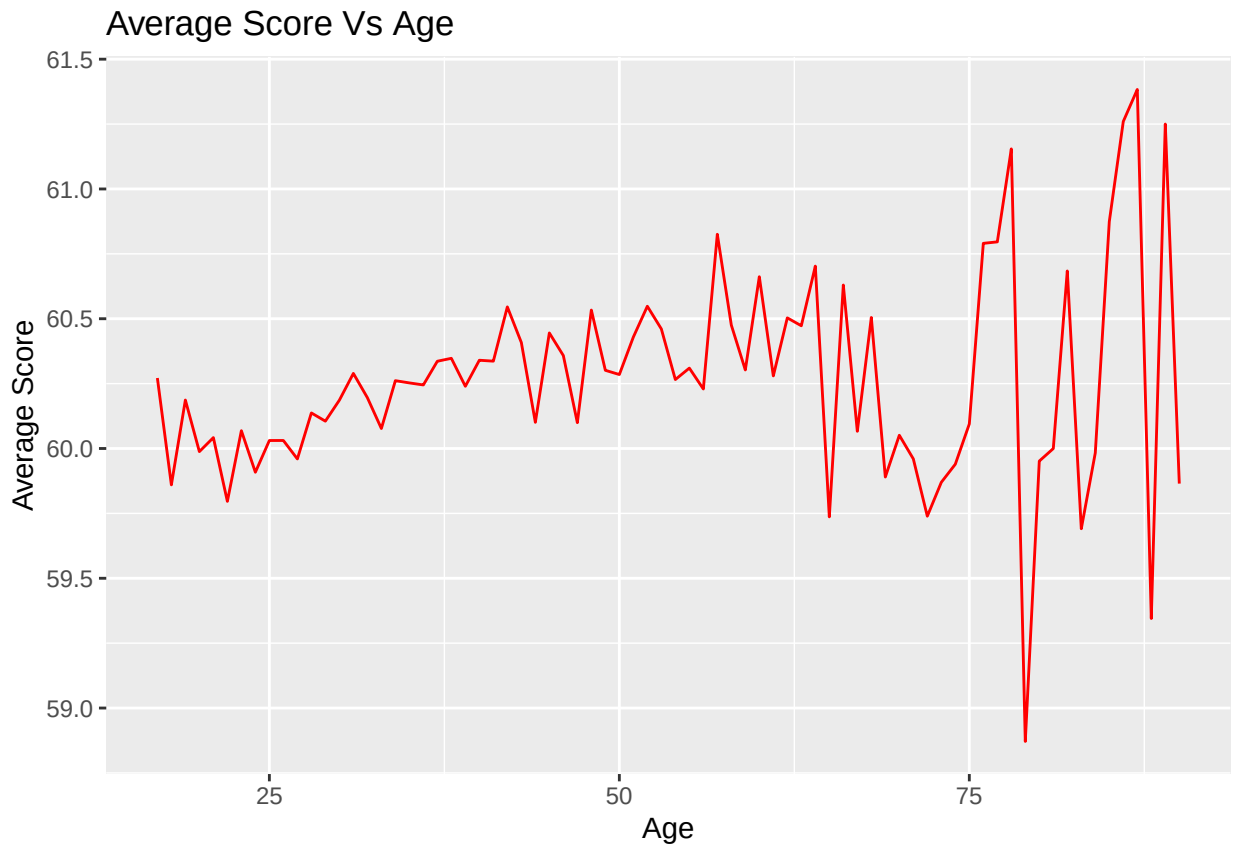
```
##            names   min   Q1 median    Q3      max       mean          sd     n
## 1            age 17.00 28.0  37.00 48.00    90.00   38.64359   13.710510 48842
## 2  education_num  1.00  9.0  10.00 12.00    16.00   10.07809    2.570973 48842
## 3       cap_gain  0.00  0.0   0.00  0.00 99999.00 1079.06763 7452.019058 48842
## 4 hours_per_week  1.00 40.0  40.00 45.00    99.00   40.42238   12.391444 48842
## 5          score 43.94 57.5  60.24 62.95    76.53   60.22825    4.025339 48842
##   missing
## 1       0
## 2       0
## 3       0
## 4       0
```

```
## 5       0
```

```
# This chart gives organized information for each row.
```

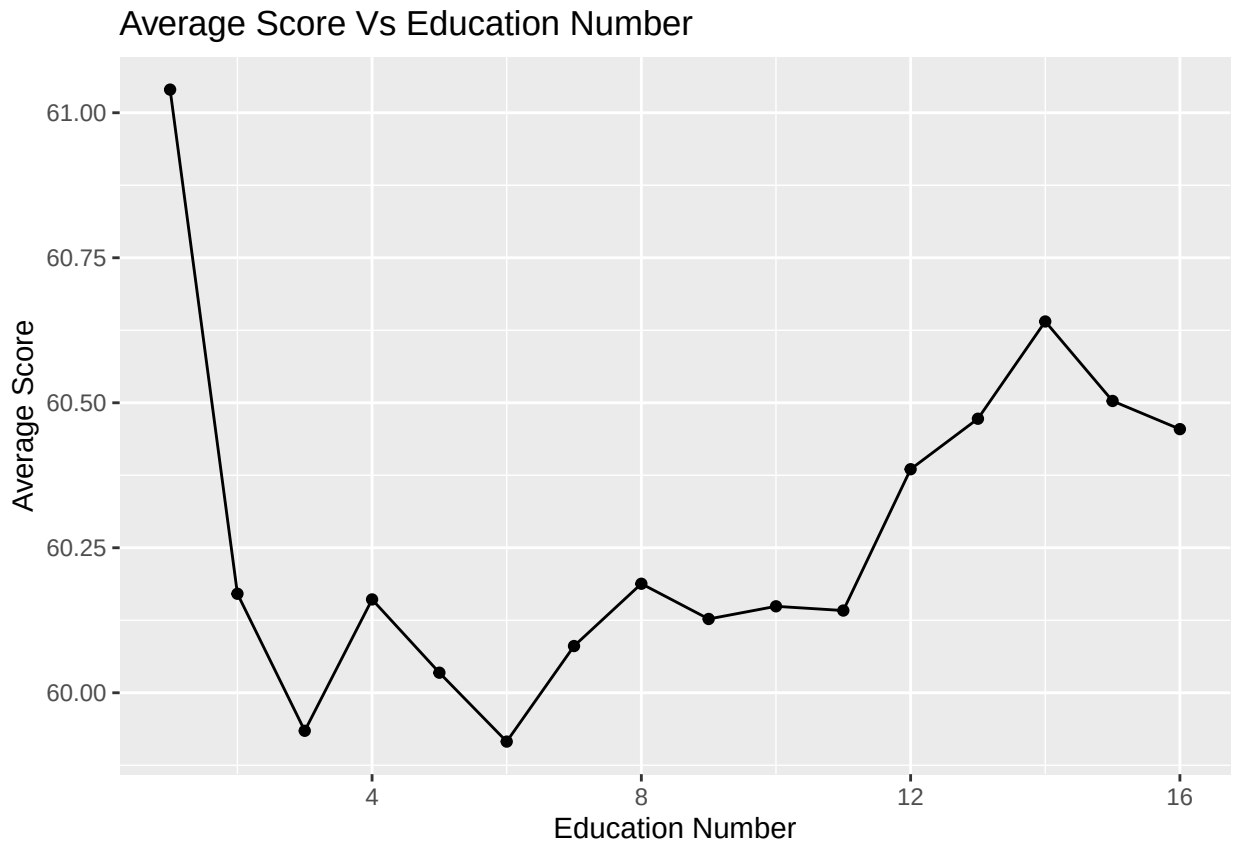
2)

```
# Graph used to compare age and average score.

# If you notice the variablity it may be due to the short range on the y-axis.

insurance_policy %>%
  group_by(age) %>%
  summarize(Average_score_by_age = mean(score, na.rm = T)) %>%
  ggplot(aes(x = age, y = Average_score_by_age)) +
  geom_line(color = "red") +
  ggtitle("Average Score Vs Age") +
  xlab("Age") +
  ylab("Average Score")
```



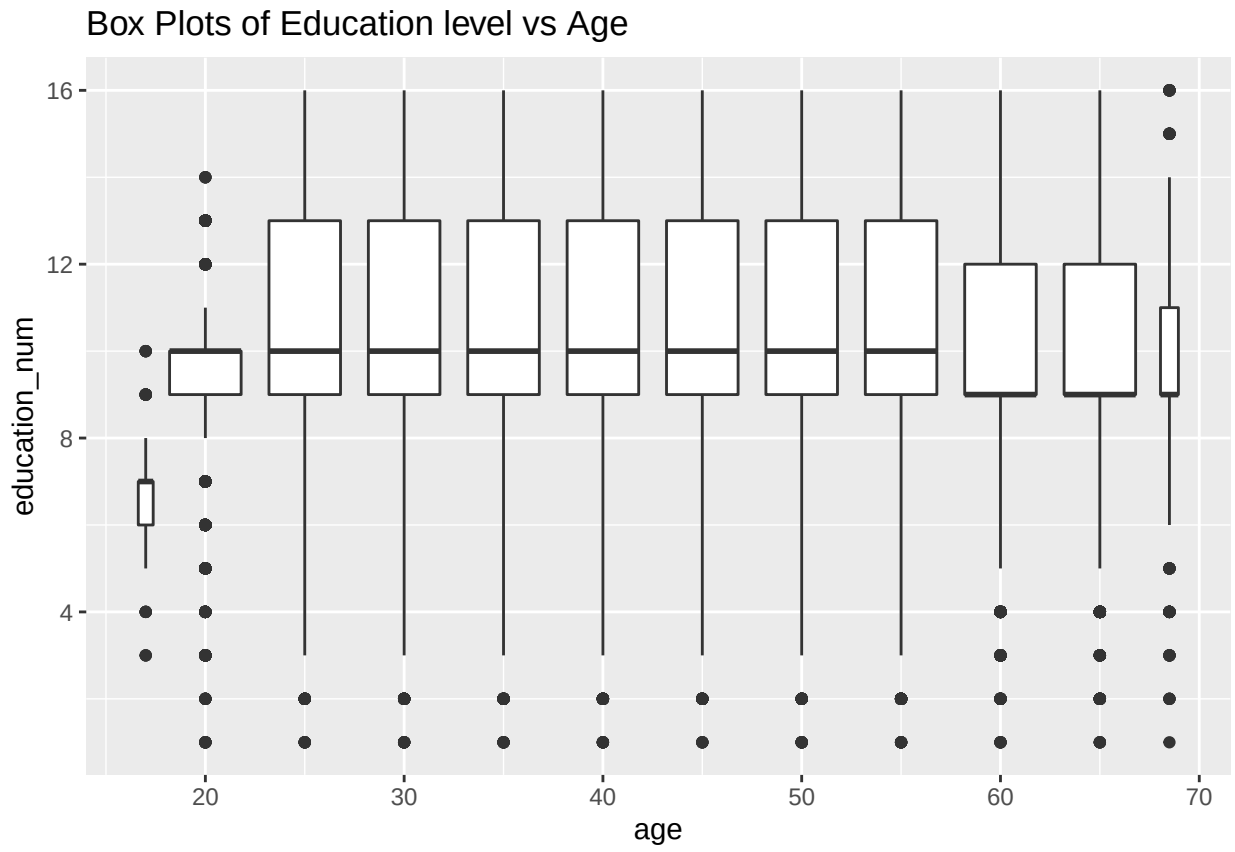
Graph 1: The idea behind this graph is to show how age may play a part in someone score. Since the insurance score is a proprietary value that shows the consumers trustworthiness. I expected an outcome where middle aged people would have a higher score. There is a lot of variability in this graph or it may seem that way bu these values are not far off from each other as it is scaled from 59-61. There is some fluctuation with the older groups but that is mainly because there weren't much scores to be reported for their age so the average would fluctuate. Relating to the research question, we can highlight that on average score isnt much different when predicting high value customers.

```
# This graph is focused on the average score compared to their education level.

# It may seem strange with the variability and outliers but the range is what sets it off.

insurance_policy %>%
  group_by(education_num) %>%
  summarise(average_score = mean(score)) %>%
  ggplot(aes(x= education_num, y = average_score )) +
  geom_line() +
  geom_point() +
  ggtitle("Average Score Vs Education Number") +
  xlab("Education Number") +
  ylab("Average Score")
```



Graph 2: Using the 'group_by' and 'summarise' functions I was able to group by the education number and then using the 'summarise' function I was able to take the average score of people and plot it against their education number. In the graph was see what looks like an out liar but it seems to be an optical illusion. We see that the y-axis is ranging only from 60-61 so that we know that there isn't actually much difference when it comes to average score ploted against age as there isnt much difference.

```
insurance_policy %>%
  group_by(education_num) %>%
  filter( age < 70) %>%
  ggplot( aes(x = age, y = education_num)) +
  geom_boxplot(aes(  group = cut_width(age, 5))) +
  ggtitle("Box Plots of Education level vs Age")
```




```
# Box plot used to show multiple results at different ages.

# The Box plot shows the variability in different ages.
```

Graph 3: Graph 3 was done as a box plot that illustrated the education number against their age. We don't see much data until we get older than 25 and younger than 60. It seems as tough the average ecuation number for most people is about the same and a little bit less for the older groups. This can be taken into account when finding out who is a high value customer.

3)

```
insurance_policy %>%
  group_by(education_num) %>%
  summarise(min_value_capital_gain = min(cap_gain),
            max_value_capital_gain = max(cap_gain),
            average_capital_gain = mean(cap_gain))
```

```
## # A tibble: 16 x 4
##    education_num min_value_capital_gain max_value_capital_gain average_capital~1
##            <int>                  <int>                  <int>             <dbl>
##  1             1                      0                  41310              732
##  2             2                      0                   7688              124.
##  3             3                      0                  99999              360.
##  4             4                      0                  10566              243.
```

```
##  5             5                     0                   99999             313.
##  6             6                     0                   99999             323.
##  7             7                     0                   15024             204.
##  8             8                     0                   18481             209.
##  9             9                     0                   99999             573.
## 10            10                     0                   99999             560.
## 11            11                     0                   99999             779.
## 12            12                     0                   99999             637.
## 13            13                     0                   99999            1763.
## 14            14                     0                   99999            2584.
## 15            15                     0                   99999           10586.
## 16            16                     0                   99999            5728.
## # ... with abbreviated variable name 1: average_capital_gain
```

Explanation of question 3: This chart summarizes the education number as well as their minimum, maximum and average capital gain. This data can be used to support our hypothesis. When we look deeper into the data we notice that the average capital gain of higher education numbers is higher than those with a lower education number. This can be used to look at their value as a lender.

Statistical Analysis: Confidence Interval, Hypothesis Test, and Model or Machine Learning

1)

```
Married_overTwentyFive <- insurance_policy %>%
  filter(marital_status == "Married-civ-spouse" & age > 25 )
```

```
bootstrap_samplemeans =
  do(500)*mean(~age, data = sample_n(Married_overTwentyFive,
                                     size=dim(Married_overTwentyFive)[1],
                                     replace=TRUE))


bootstrap_samplemeans = bootstrap_samplemeans %>% rename(bootstrap_samplemean = mean)
head(bootstrap_samplemeans)
```

```
##   bootstrap_samplemean
## 1             44.35932
## 2             44.24988
## 3             44.33912
## 4             44.30343
## 5             44.31148
## 6             44.25830
```

```
confidence_interval = quantile(
  bootstrap_samplemeans$bootstrap_samplemean, probs = c(.025, .975))

confidence_interval
```

```
##     2.5%    97.5%
## 44.13669 44.45083
```

Explanation of Bootstrap confidence interval: The bootstrap sample is focused on couples married over the age of 25. The probability means are taken at 97.5% and 2.5% showing that they are very close to each other.

2)

My hypothesis is

```
Scoresyounger50 <- insurance_policy %>%
  group_by(score) %>%
  filter(age < 50)

Scoreolder50 <- insurance_policy %>%
  group_by(score) %>%
  filter(age > 50)



t.test(Scoresyounger50$score, Scoreolder50$score)
```

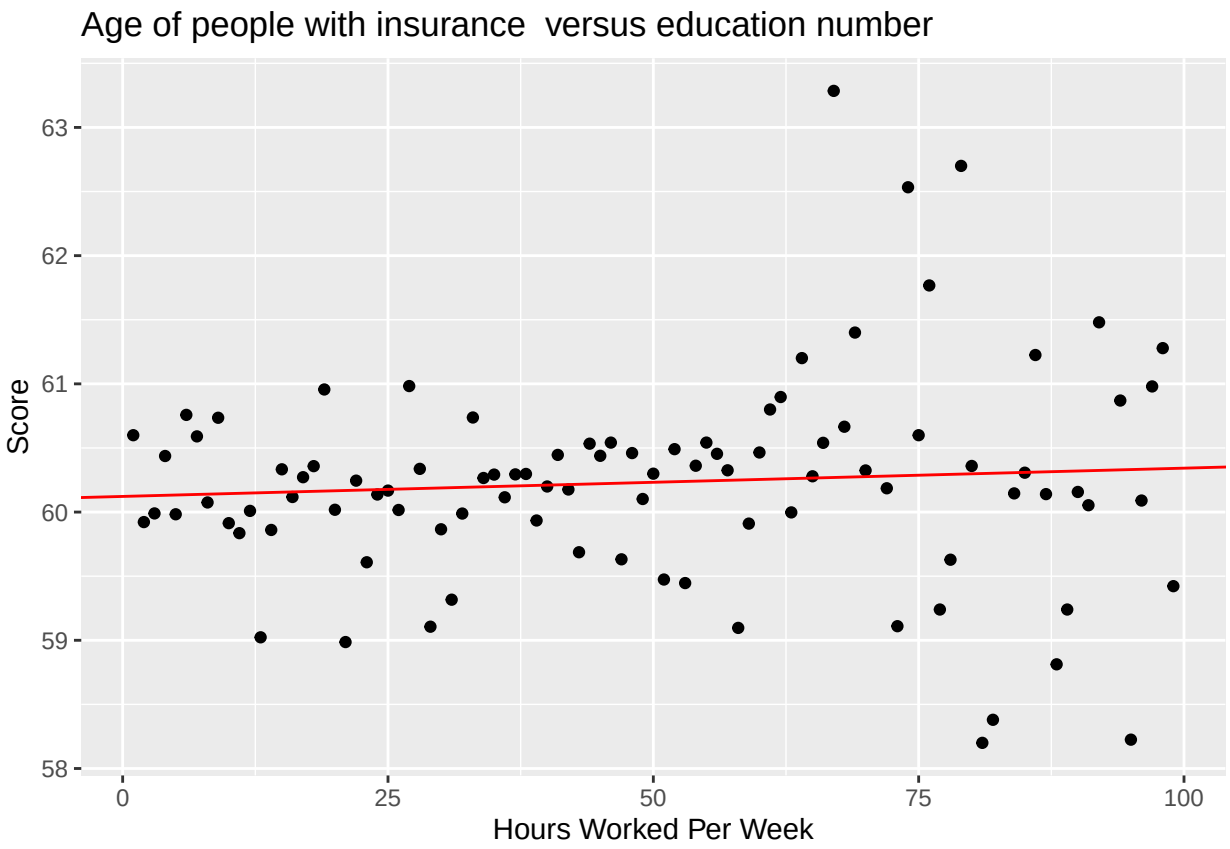
```
##
##  Welch Two Sample t-test
##
## data:  Scoresyounger50$score and Scoreolder50$score
## t = -4.1582, df = 15079, p-value = 3.225e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2811794 -0.1010179
## sample estimates:
## mean of x mean of y
##  60.18816  60.37926
```

We reject the null hypothesis (H0). The data provide convincing evidence at the .05 significance level that the p-value is significantly lower than the pre-selected level of significance.05. Meaning that we would have to reject the null hypothesis. The hypothesis test was taken on the scores of people older than 50 and younger than 50. The hypothesis test was taken on a comparison of both to see if there is enough statistical evidence to reject this comparison.

3)

```
# This was created using a linear model

# The linear model is used to predict the score compared to the hours per week and the

data <- insurance_policy %>%
  group_by(hours_per_week) %>%
  summarise(scoreHoursperWeek = mean(score))
# Created a data frame taking the average of the score while grouping by hours per week

model <- lm(data$scoreHoursperWeek ~ data$hours_per_week)
```

```
data %>%
  ggplot() +
  geom_point(aes(x = hours_per_week,
                 y = scoreHoursperWeek)) +
  labs(title = "Age of people with insurance  versus education number",
       x = "Hours Worked Per Week", y = "Score") +
  geom_abline(intercept = model$coefficients[1],
              slope = model$coefficients[2],
              color="red")
```



Explanation of the model:

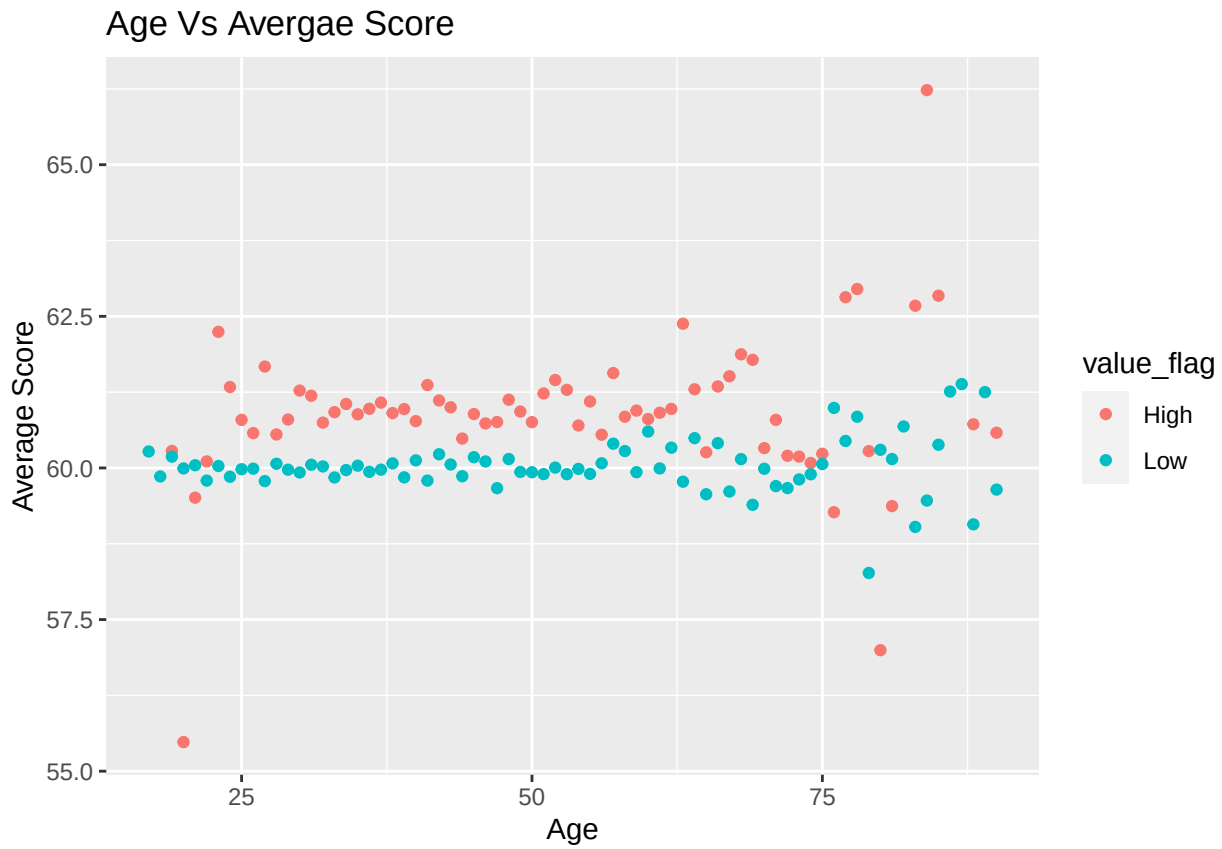
This model is showing the comparison of the hours per week people work to their scores. It seems there is interesting data before 50 hours but after that it does spread out more. We do however have to keep in mind that this graph is ranging from 58 to 63 on the y-axis so it would not show much major difference when drawing the line of best fit. This graph though does however show the line of best fit starting at around 60.2 score at the y-axis. This data can be used to support our research question by showing how the correlation between a hours worked and their score can make that person a better lender.

One more thing

```
insurance_policy %>%
  group_by(age, value_flag) %>%
  summarise(average_score = mean(score)) %>%
```

```
  ggplot(aes(x= age, y = average_score )) +
  geom_point(aes(color = value_flag)) +
  ggtitle("Age Vs Avergae Score") +
  xlab("Age") +
  ylab("Average Score")
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```




```
# Graph colored to show the value of people comparing their age to their score.

# This can be used to support the research question.
```

Conclusion

This RMD file is used to answer my teams research question which is how you would provide variables that associate with the prediction that a customer is high value. I have created three graph to support my research question. The first graph uses 'group_by' to group by age then uses 'summarise' to summarise the average score. This gives us a better understanding of the correlation between the average score and the education number. The second graph groups by education number and summarises by the average score. The third graph is a box plot that offers a comparison to the education number and age. All these graphs can be used to support our research question by seeing their correlation to the value flag. As done in the final model above we see how we colored by the value flag and were given a clear comparison of how a higher

average score shows that they are usually a high value. The bootstrap confidence interval is on married couples over the age of 25. This is taken to show the confidence in the data for couples over the age of 25 and the benefits of the data. I then took a hypothesis test on difference age groups and compared them using the hypothesis test. The null hypothesis failed. The data collected should be used to stregthen my research in upcoming projects.