MANG6513 Foundation of Business Analytics and Management Science

EXAMINATIONS 2021-22

Student number: 31302815

Question 1

Question 1 - a

The primary technique chosen for visual analysis is the ggplot plotting package. First, name a dataframe “data1” and count the total number of accidents in each week of the day by function “aggregate.” Then, change the name of the number presented in the column Day_of_Week in data1. For example, “1” represent “Sunday”; ”2” represent “Monday.” Finally, use the function “ggplot” to create a bar chart from data in a dataframe ”data1”. As can be seen in figure 1, accidents tend to happen on Friday with the most number of accidents at 950, which is followed by Tuesday and Thursday at around 870 accidents. The number of accidents that happened on Monday and Wednesday were 853 and 843, respectively. Also, the number of accidents that happened on the weekend represents the least the number of accidents with only 727 as well as 687.

```r
#Q1-a
data1 <- aggregate(data1$Day_of_Week, by=list(data1$Day_of_Week), length)
colnames(data1) <- c("Day_of_Week","Number_of_Accident")
data1
data1$Day_of_Week <- ifelse(data1$Day_of_Week == '1', 'Sunday', data1$Day_of_Week
)
data1$Day_of_Week <- ifelse(data1$Day_of_Week == '2', 'Monday', data1$Day_of_Week
)
data1$Day_of_Week <- ifelse(data1$Day_of_Week == '3', 'Tuesday',
data1$Day_of_Week)
data1$Day_of_Week <- ifelse(data1$Day_of_Week == '4', 'Wednesday',
data1$Day_of_Week)
data1$Day_of_Week <- ifelse(data1$Day_of_Week == '5', 'Thursday',
data1$Day_of_Week)
data1$Day_of_Week <- ifelse(data1$Day_of_Week == '6', 'Friday', data1$Day_of_Week
)
data1$Day_of_Week <- ifelse(data1$Day_of_Week == '7', 'Saturday',
data1$Day_of_Week)

ggplot(data=data1,mapping=aes(x=Day_of_Week,y=Number_of_Accident))+
  geom_bar(stat='identity')+
  geom_text(aes(label=Number_of_Accident))+
  ggtitle("Number of Accidents Across Each Day In The Week")
```

Figure1: R commands for Q1-a

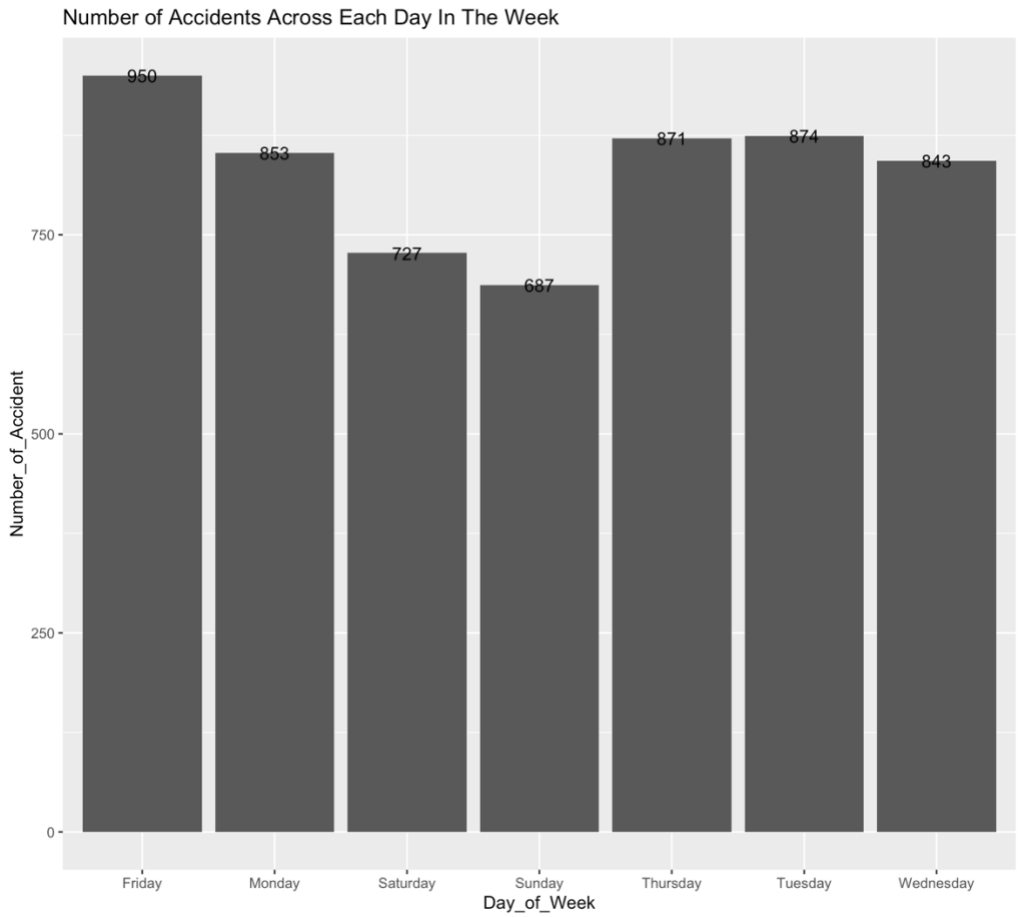

Figure2: Number of accidents across day of the week

Question 1 – b

There are four components of time series data which are the trend, seasonal, cyclical and random. Among those components, trend data refer to a pattern that describes whether the data for a time series is moving upward or downward (Trend – Statista Definition, 2022). Besides, according to Kenton (2022), seasonal data may be noted as a feature in which the data undergoes regular and predictable changes that repeat each calendar year.

To determine whether the number of accidents is seasonal or trending, first, count the number of accidents that occur every day (with format day/month/year) and assign it to “df4”(*Figure3*). Create a dataframe “df3” and compute the format of date from day/month/year to month/year as we would like to analyze data between 01/2013 and 12/2017(*Figure3*). Subsequently, build the time sires model and plot the result by “autoplot .”According to figure 4, data has a trend but is not obvious. Therefore, in order to see more clear pattern, take the first difference in the data to remove the trend(figure 5).

```r
#Q1-b
library(dplyr)
#count the number of accidents by date(year-month-day)
df4 <- aggregate(accident$Date, by=list(accident$Date), length)
colnames(dataframe2) <- c("Date","Number_of_Accident")

#count the number of accidents by date(year-month)
df3 <- ddf4 %>%
  mutate(date = as.Date(Date)) %>%
  mutate(Date = format(date, '%Y-%m')) %>%
  group_by(Date) %>%
  summarize(Number_of_Accident = sum(Number_of_Accident))
  as.data.frame(df3)

#build the time sires model
install.packages("fpp2", dependencies=TRUE)
library(fpp2)
library(ggplot2)
Y <- ts(dataframe3[,2],start=c(2013,1),frequency = 12)

#plot the chart to see the pattern
autoplot(Y) +
  ggtitle("Time Plot: Car Accidents over Time") +
  ylab("Number of Accidents")
```

Figure3: R commands for Q1-b(1)

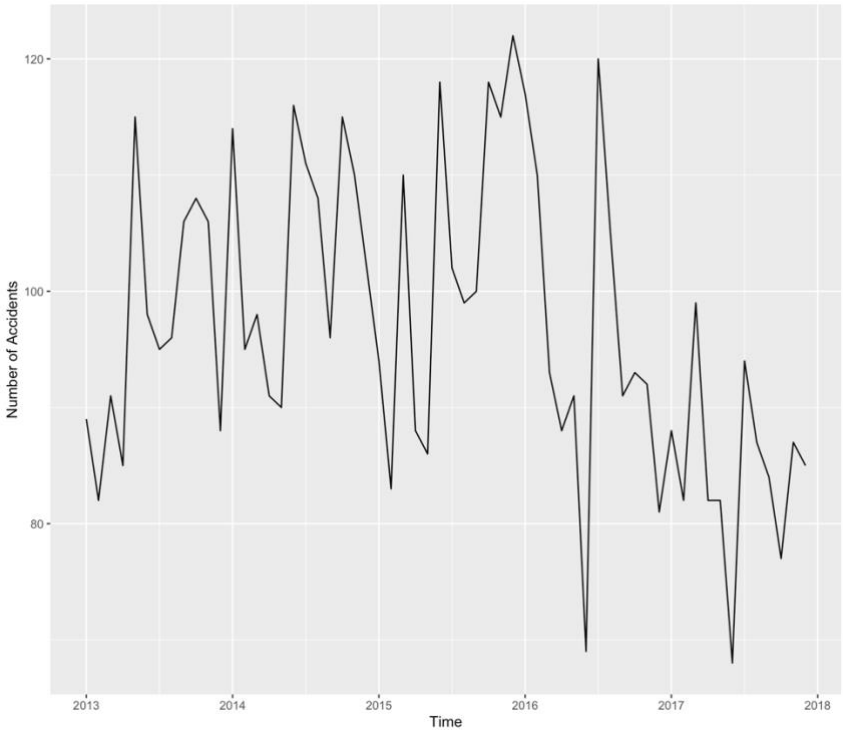

Figure4: Number of accidents across *time plot*

```
#take the first difference of the data
DY <- diff(Y)
autoplot(DY) +
  ggtitle("Time Plot: Change in Car Accidents month to month") +
  ylab("Number of Accidents")

#Seasonal Plot
ggseasonplot(DY) +
  ggtitle("Seansonal Plot") +
  ylab("Number of Accidents")
```

Figure5: R commands for Q1-b(2)

Figure 6 below shows the result after taking the first difference of the data. The trend of the data nearly cannot be seen, while the seasonality of data still appears uncertain. Therefore, investigate the seasonality by R command “ggseasonplot”(*Figure5*).

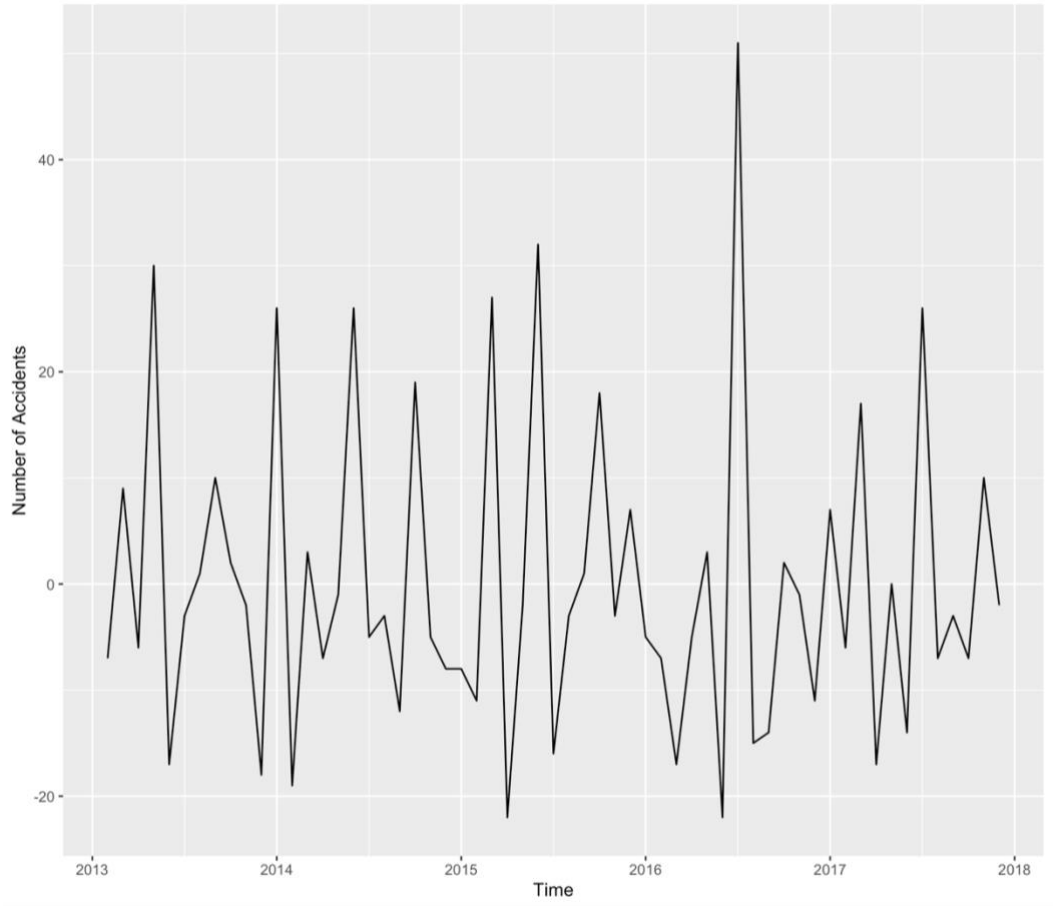

Figure6: Number of accidents across time plot (first difference)

Overall, in Figure7 below, the series appears non-stationary. The data appears to have some tendency but is not strong. There are no regular or repeat changes in the graph as well.

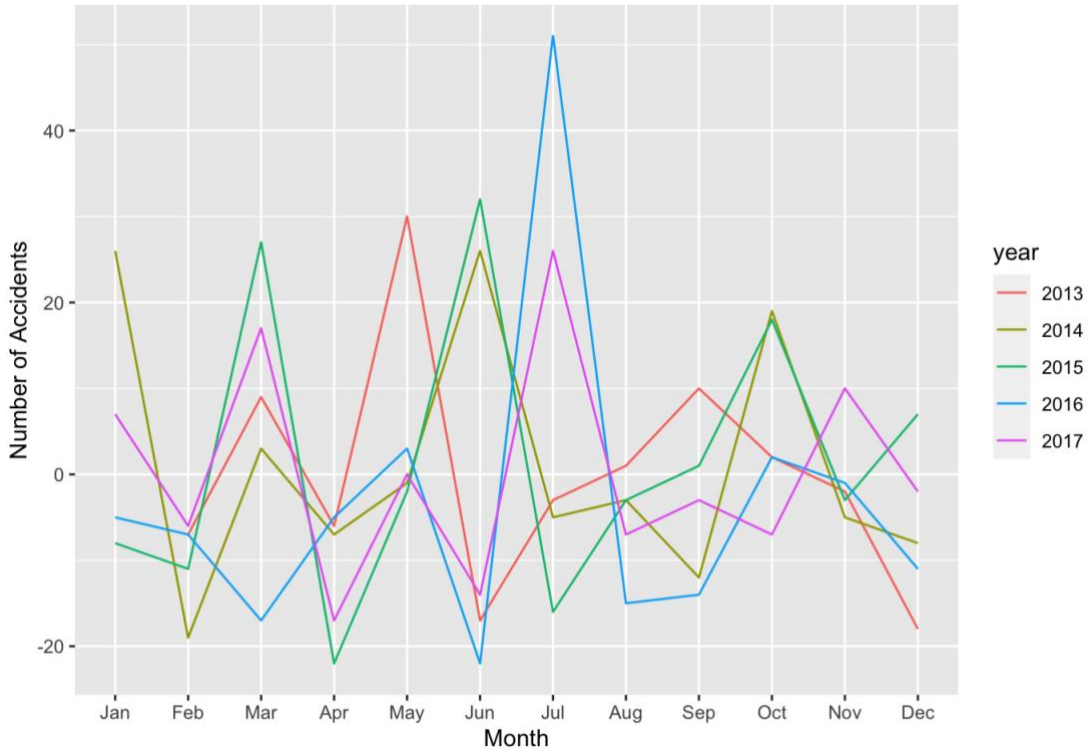

Figure7: Number of accidents across seasonal plot

Question 1 – c

Use R commands to create a data frame and count the number of accidents in different road types so that we can get the figure from the data frame (Figure10). As provided in the Data_Guide.xls., 1 represents roundabout and 6 represents Single carriageway. Figure8 can only provide the information that there are more car accidents happening on Single carriageway(4851) than accidents happening at roundabouts (224). Apply the function" pie" to the pie chart (Figure9). Figure 8 shows that the number of accidents that happened in roundabouts accounts for 3.86%. Single carriageway takes the largest part of the chart with 83.57%.

Therefore, further analysis is needed to support whether roundabouts tend to reduce the number of accidents. According to Crosby et al. (2016), the Highways Agency England and the Department for Transport indicated there were 116,615 instances of daily average traffic flow counts for a 16-year period. Take this number as an assumption to compute the ratio: the number of daily car accidents / average daily traffic flow counts(116,615). Then take the ratio as y and road types as variables. Firstly, define the hypothesis. H0: There is no correlation between ratio and road types. H1: There is a

significant correlation between ratio and road types. Then, build a regression model.

According to the output, the p-value of all the variables(Roundabout, One_way_street,Dual_carriageway,Single_carriageway,Slip_road+regression $Unknown) are larger than 0.05, we cannot reject H0, meaning there is no significant correlation between ratio and road types. However, as the coefficient of the roundabout is negative, the road types of roundabouts can reduce the number of accidents.

	Road_Type	Number_of_Accident
1	1	224
2	2	65
3	3	608
4	6	4851
5	7	18
6	9	39

Figure8: the table of road type and the number of accidents

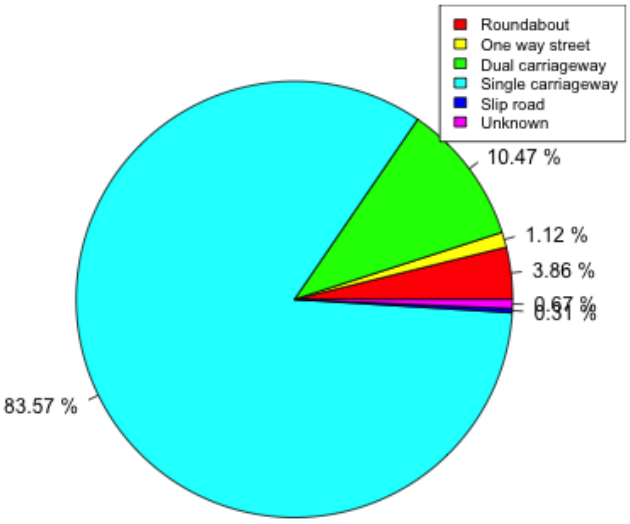

Figure9: The pie chart (road type)

```r
#Q1-c
dataframe <- aggregate(accident$Road_Type, by=list(accident$Road_Type), length)
colnames(dataframe) <- c("Road_Type","Number_of_Accident")

# Create data for the graph
x <- c(224, 65, 608, 4851, 18, 39)
labels <- c("Roundabout", "One way street", "Dual carriageway", "Single
carriageway","Slip road","Unknown")

piepercent<- paste(round(100*x/sum(x), 2), "%")

# Give the chart file a name.
png(file = "/Users/vivian/Desktop/研究所/Semester1/piechart.jpg")

# Plot the chart.
pie(x, labels = piepercent,radius = 0.7, main = "piechart",col = rainbow(length(x
)))
legend("topright", c("Roundabout", "One way street", "Dual carriageway", "Single
carriageway","Slip road","Unknown"), cex = 0.8,
       fill = rainbow(length(x)))

# Save the file.
dev.off()
```

Figure10: R commands for Q1-c(1)

```
> summary(mod)

Call:
lm(formula = regression$ratio ~ regression$Roundabout + regression$One_way_street +
    regression$Dual_carriageway + regression$Single_carriageway +
    regression$Slip_road + regression$Unknown, data = regression)

Residuals:
       Min         1Q     Median         3Q        Max
-3.034e-05 -1.114e-05 -2.569e-06  6.007e-06  5.746e-05

Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.892e-05  2.555e-06  15.232   <2e-16 ***
regression$Roundabout       -1.517e-06  2.768e-06  -0.548    0.584
regression$One_way_street   -3.958e-06  3.232e-06  -1.225    0.221
regression$Dual_carriageway -1.867e-06  2.636e-06  -0.708    0.479
regression$Single_carriageway -2.049e-06  2.565e-06  -0.799    0.424
regression$Slip_road         2.052e-06  4.547e-06   0.451    0.652
regression$Unknown                  NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.596e-05 on 5799 degrees of freedom
Multiple R-squared:  0.0005249, Adjusted R-squared:  -0.0003369
F-statistic: 0.6091 on 5 and 5799 DF,  p-value: 0.693
```

Figure11: R commands for Q1-c(2)

Question 1 - d

It can be seen in figure 12 that the number of accidents that occur under speed limits of 30 accounts for 86.18%. The rest of the speed limits (20,40,50,60,70) took a small portion with around 1% to 7%. This can only prove that there are relatively more accidents that occur under 30-speed limits. In figure 13, despite the number of accidents that happened under lighting conditions in daylight accounts for the most with 67.73%, it still cannot be proved that improving lighting conditions can decrease the number of accidents.

Further analysis is needed. As mentioned, there are 116,615 cases of daily average traffic flow counts across the 16 years period (Crosby et al.,2016). Calculate the ratio and selected it as y, and lighting conditions and speed limit as variables x. Let H0: There is no correlation between the ratio and lighting conditions, and speed limit. H1: There is a significant correlation between the ratio and lighting conditions, and speed limit. Build a regression model(*Figure14*).

Based on the result *(Figure14)*, the p-value of the variable speed limit is more than 0.05, which means there is no correlation between ratio and speed limit. On the contrary, the p-value of the variable lighting conditions is less than 0.1, meaning there is a slight correlation between ratio and lighting conditions. Besides, the coefficient of the lighting conditions is negative, so the lighting conditions can reduce the number of accidents. This result is also supported by Yannis et al.(2013) that improved road lighting has a positive impact on reducing accident safety and frequency.

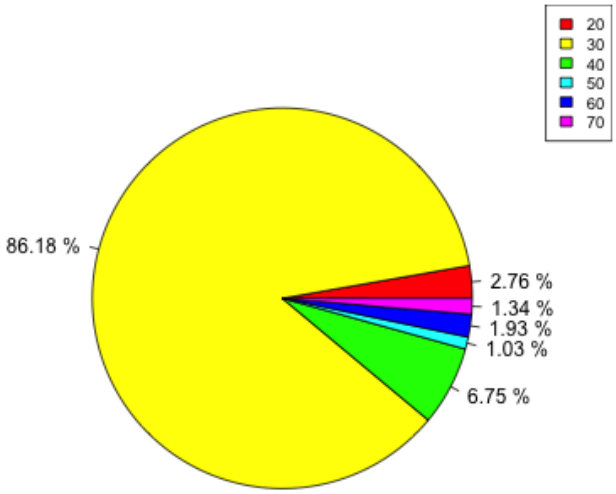

Figure12: The pie chart(speed limit)

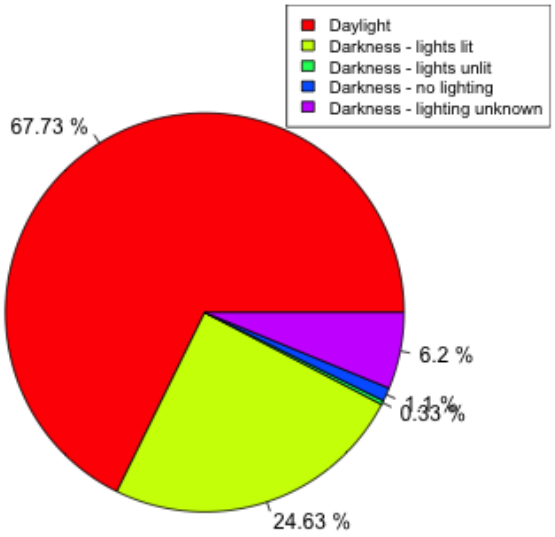

Figure13: The pie chart (lighting conditions)

```
> mod1<-lm(lightspeed_regression$ratio~lightspeed_regression$Light_Conditions+lightspeed
_regression$Speed_limit,data=lightspeed_regression)
> summary(mod1)

Call:
lm(formula = lightspeed_regression$ratio ~ lightspeed_regression$Light_Conditions +
    lightspeed_regression$Speed_limit, data = lightspeed_regression)

Residuals:
       Min         1Q     Median         3Q        Max
-2.862e-05 -1.143e-05 -2.853e-06  6.907e-06  5.777e-05

Coefficients:
                                     Estimate Std. Error t value
(Intercept)                          3.747e-05  9.889e-07  37.893
lightspeed_regression$Light_Conditions -1.975e-07  1.139e-07  -1.734
lightspeed_regression$Speed_limit      -4.043e-09  2.977e-08  -0.136
                                    Pr(>|t|)
(Intercept)                           <2e-16 ***
lightspeed_regression$Light_Conditions   0.0829 .
lightspeed_regression$Speed_limit        0.8920
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.595e-05 on 5802 degrees of freedom
Multiple R-squared:  0.0005256, Adjusted R-squared:  0.0001811
F-statistic: 1.526 on 2 and 5802 DF,  p-value: 0.2176
```

Figure14: R commands for Q1-d(2)

Question 2

Question 2- a

The problem is whether the local authority should take the option of constructing a bridge, a tunnel or a ferry. Moreover, how to make the decision will be based on how much cost the local authority will spend in 20 years according to different choices. This problem can be addressed by using a decision tree because there are various conditions with different probabilities. The decision tree can list every condition with its corresponding probability so that we can calculate the expected total cost. Then we can choose the greatest cost according to the decision tree and make a decision under

conditions of uncertainty. The following figure is the decision tree, the cost will be set with(-), and income will be set with (+):

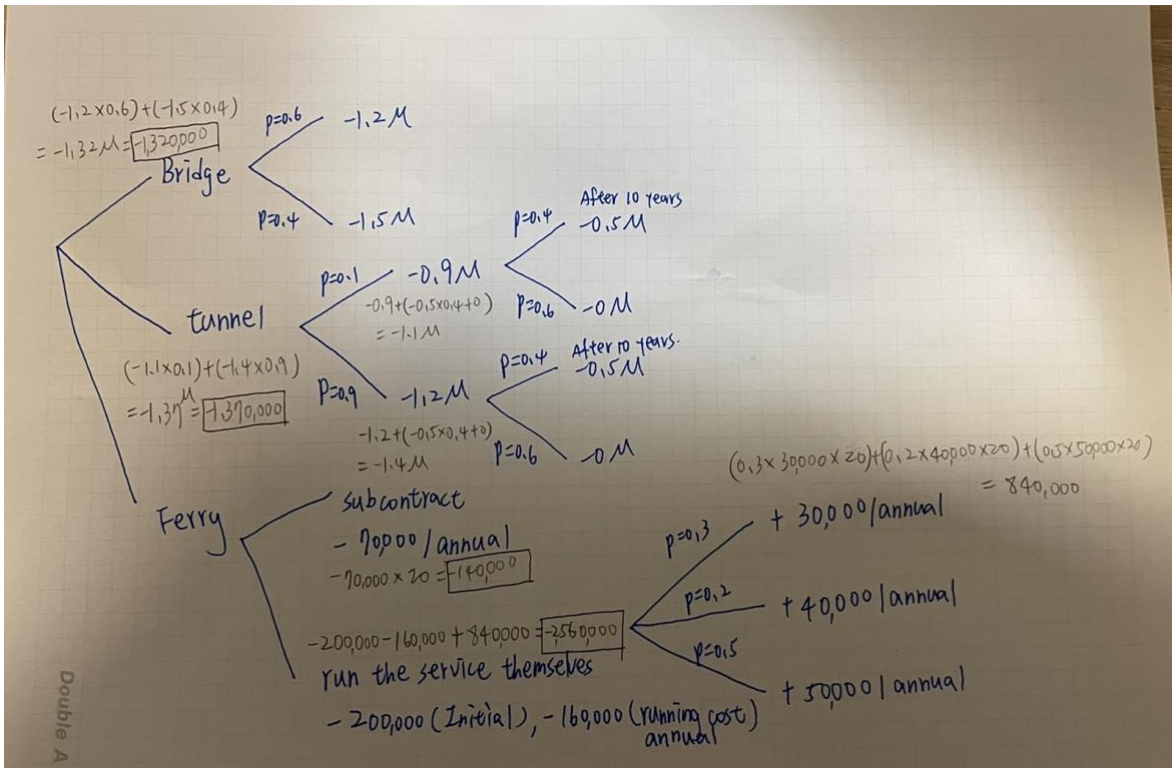

Figure15: Decision tree

The 20 years of total cost spent on the bridge and tunnel is calculated by adding each expected cost from each branch of the decision tree. For bridge construction, calculate the total cost by different costs with its given probability. As a result, the equation is : [(-1.2M * 0.6)+(-1.5M*0.4)]= -1.32M

As for tunnel construction, first, calculate the reinforcement cost by the number of times by the probability, respectively. For example:(-0.5M*0.4+0M*0)=-0.2M. Next, the equation in the second layer of branch will be: -0.9M+(-0.5M*0.4+0M*0)=-1.1M and -1.2M+(-0.5M*0.4+0M*0)=-1.4M. So the total cost can be output by adding two number(-1.1M, -1.4M) multiples with their corresponding probability(0.1, 0.9).

Besides, the total cost spent on the ferry will be calculated from 2 situations. First, if the local authority subcontracts a private company, the total cost will be (70,000 pounds * 20 years)= 140,000 pounds.

Second, if the local authority runs the service by themselves, the total cost will be [-2000,000(initial investment)+(-160,000 pounds(running cost) * 20

years)] minus annual income from ticket sales in 20 years by three difference probability.

As is shown in the decision tree, after 20 years, the total cost can be 1,320,000 pounds if the local authority constructs the bridge, while 1,370,000 pounds for building a tunnel. Regarding constructing the ferry, it costs the most either the local authority subcontract a private company or runs the service themselves with 140,000 pounds and 2,560,000 pounds. In conclusion, it is obvious that the local authority should choose to construct a bridge as this is the best choice of three options with the least cost.

Question 2 - b

In a nutshell, it is not a good investment for the local authority to increase the cost of £50,000 to know in advance whether the tunnel should be reinforced after ten years. If the local authority increase the cost of £50,000, the equation will be 0.1*[(0.9M+50000)+(0.4*0.5M+0)]+0.9*[(1.2M+50000)+ (0.4*0.5M+0)]=1,420,000 pounds. As a result, it's better for the local authority to maintain the same option for the construction of the bridge.

Although the local authority can save money (0.5M pounds) as there is a chance of no need to pay for the reinforcement fee after ten years, the total expected cost calculated by the mentioned equation is 1,420,000 pounds, and the amount is even higher than the tunnel option provided before. Also, if the local authority is willing to spend above 1,40,000 pounds, they can also consider the option of subcontracting a private company. Moreover, if the local authority chooses to build the tunnel, no matter whether they have to pay the fee of £50,000 or not, they cannot prevent the truth that the tunnel has to be reinforced or not in the future.

Reference:

Crosby, H., Davis, P. and Jarvis, S.A., 2016, September. Spatially-intensive decision tree prediction of traffic flow across the entire UK road network. In *2016 IEEE/ACM 20th International Symposium on Distributed Simulation and Real Time Applications (DS-RT)* (pp. 116-119). IEEE.

Kenton, W., 2022. *Seasonality*. [online] Investopedia. Available at: <https://www.investopedia.com/terms/s/seasonality.asp> [Accessed 24 August 2022].

Statista Encyclopedia. 2022. *Trend - Statista Definition*. [online] Available at: <https://www.statista.com/statistics-glossary/definition/425/trend/> [Accessed 24 August 2022].

Yannis, G., Kondyli, A. and Mitzalis, N., 2013, October. Effect of lighting on frequency and severity of road accidents. In *Proceedings of the Institution of Civil Engineers-transport* (Vol. 166, No. 5, pp. 271-281). Thomas Telford Ltd.