Report_D17129705

Question1

1.1 Recreation of plot “dublin employment trends.txt”.

Observations about the plot :-

Below graph shows trends of employment in 7 different sectors. Each and every sector has different trend but most dramatic sector is scientific which has extream high and extream low. I have explained my observation about each sector below.

1.Construction: Trend of construction is significantly moving up and down.As compared to other trends it does not have high pick low trends. Trend of construction is growing and decreasing at a moderate rate.

2.Finance: Finance sector has sudden high and low in trends but it does have extreme picks.If we will see on an average trend is neither growing nor declining.In first half finance trends is little high but in second half values are fluctuating at the median.

3.IT: IT is the only sector which is having growing trend as compare to other sectors.Although it is moderately growing at the start but in second part of graph it shows significant growth .

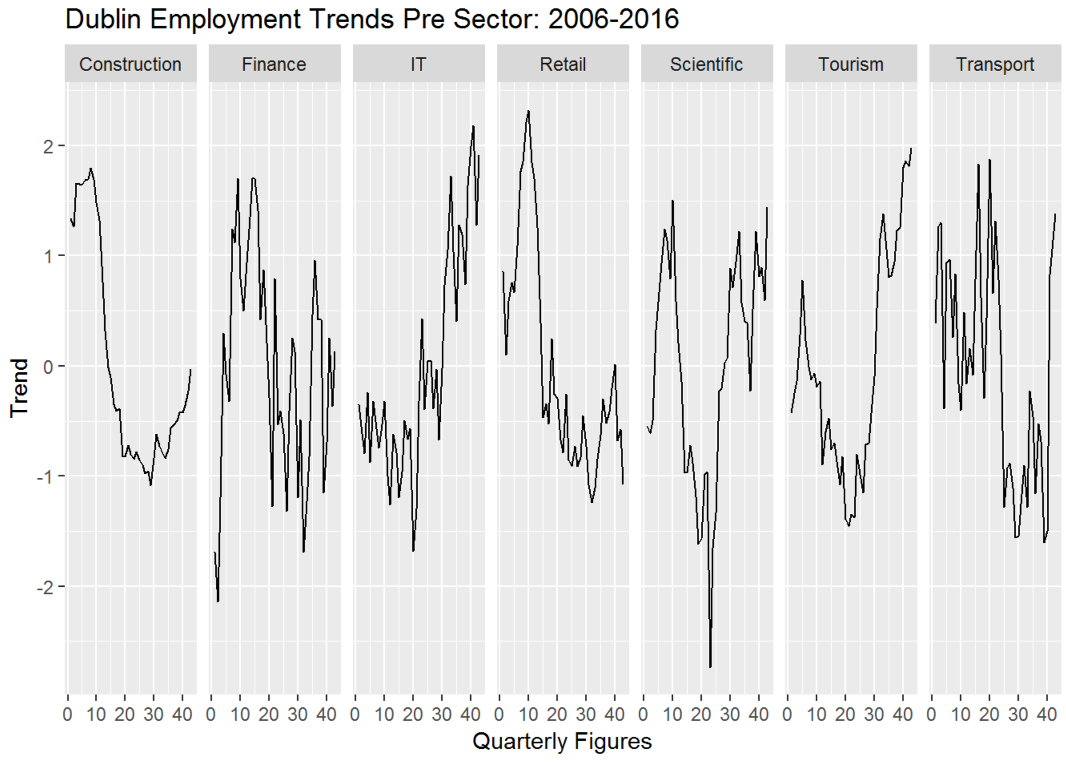
4.Retail: Unlike IT trend of retail sector decreased, but it sharply increased and then keep on decreasing.There is a fluctuation at the end of the graph but overall the trend is decreasing

5.Scientific: The trend of this trend is dramatically increasing and deceasing. At initial of the graph it has high peak and at the end of the graph it has low peak. There are two sharp up and down for this trend.

6.Tourism: Tourism has taken lot of time to grow, as first of the graph have most of the values below average. After certain low it started leveling up.

7.Transport: Transport sector has two different trends which is clearly visible in graph. In First half it has above average trend and in second half trend is below average trend.

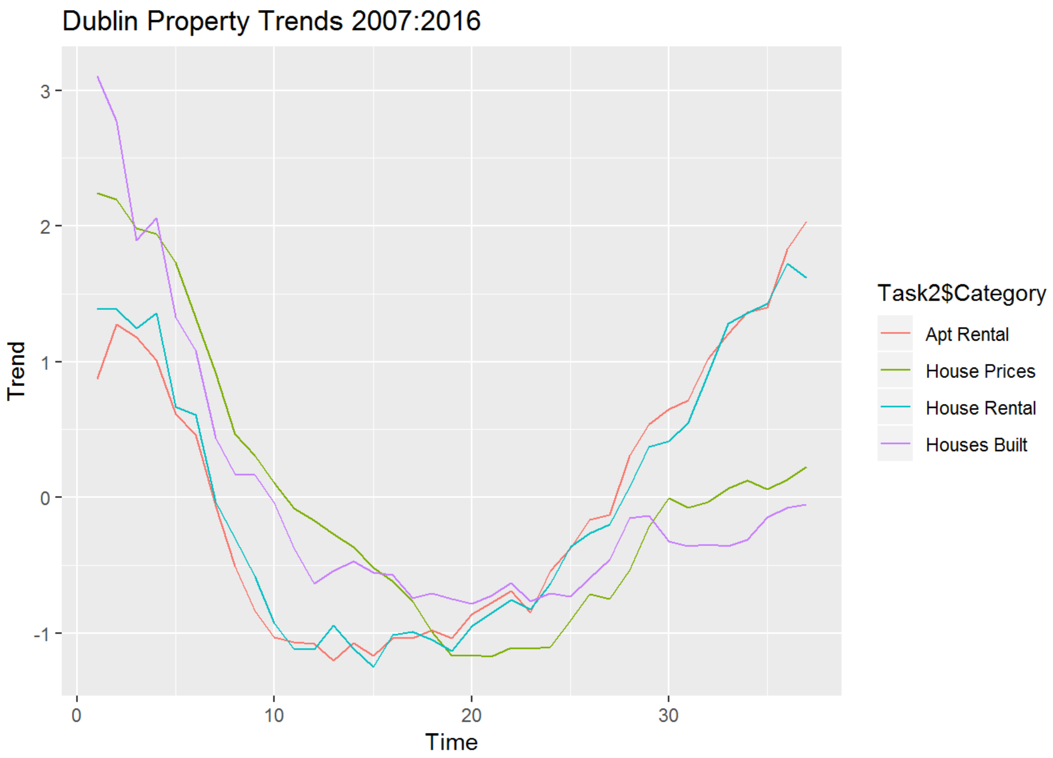
```r
library(dplyr)
library(ggplot2)
Task1<- read.table("dublin employment trends.txt", header = TRUE, dec = ".", sep = ":", quote = "", fill = FALSE)
qplot(data = Task1,x= Time, y= Employment, geom = 'line', main = "Dublin Employment Trends Pre Sector: 2006-2016") + facet_grid(~Sector) + xlab('Quarterly Figures') + ylab('Trend')
```



1.2 Recreation of plot “dublin property trends”.

The most interesting observation about property trend is,over the period,highest trend went down and lowest trend went up.If we look at the overall graph between time 10-30 most of the trends are low and after 30 trends of property in each sector significantly increased. Over the course of period 20-25 house prices are at lowest but after 25 it notably increases. Moreover house built has significantly decreased and has created visible downfall in trend. whereas apartment rentals shows drastically improvement after time 20.Other than house rental and house built other two variables shows moderate increase over the time.

```
Task2<- read.table("dublin property trends.txt", header = TRUE, sep = "\t" , stringsAsFactors = F, fill = T
RUE)
qplot(x= Task2$Time, y=Task2$Trend, geom = 'line', colour= Task2$Category, xlab = 'Time', ylab = 'Trend', ma
in = 'Dublin Property Trends 2007:2016')
```



Question2

2.1 : Summarise the data available in the different variables.

Overview of the current situation with the bikes.

Main focus is on getting the available number of bike at each location. There are 114 stations available in Dublin city, out of them there are stations where bikes availablelity is 0 and there are stations where maximum bikes are available. The stations where bike availablelity is 0 are point of concern as it may require to increase number of bikes at that particular station. The station where number of bikes available in more there we can decrease the number of bikes as those stations are not used often. To analyze situation of bikes there are some variable available in data set, this will be helpful for overall understanding of bikes. For analysing the data I have used variable name “bikedata” which is containing all the record of bikes.

1. Total number of bike stand avaibale in Dublin region and names of bike stands available in Dublin.(Variable Used “name”) Ans : There are 114 bike stands available in Dublin region. To get the list of bike stands see variable “Name_of_bikestand”.
2. Location having Maximum number of available bike stands.(variable used “available_bike_stands”) Ans. Maximum number of available number of bike stand is also 40 but it different from total number of bike stands at each location.when we compare the list “List_max_bikestand” and “ava_max_bikestand”, it is observed that available bike stands are less then number of bike stands available at each location. Available bike stand is more useful because it gives idea about the relevent bike stand which are available for use. please see the variable “ava_max_bikestand” to see list of bike stands available bike stand for use.
3. Number of bikes available at each location. (Variable used “available_bikes”,“position.lat”, “position.lng”)
   Ans.To show number of bike available I have used Map representation og bikes which will show number of bikes available at each location.when we hover on the location it will show number of available at the location.

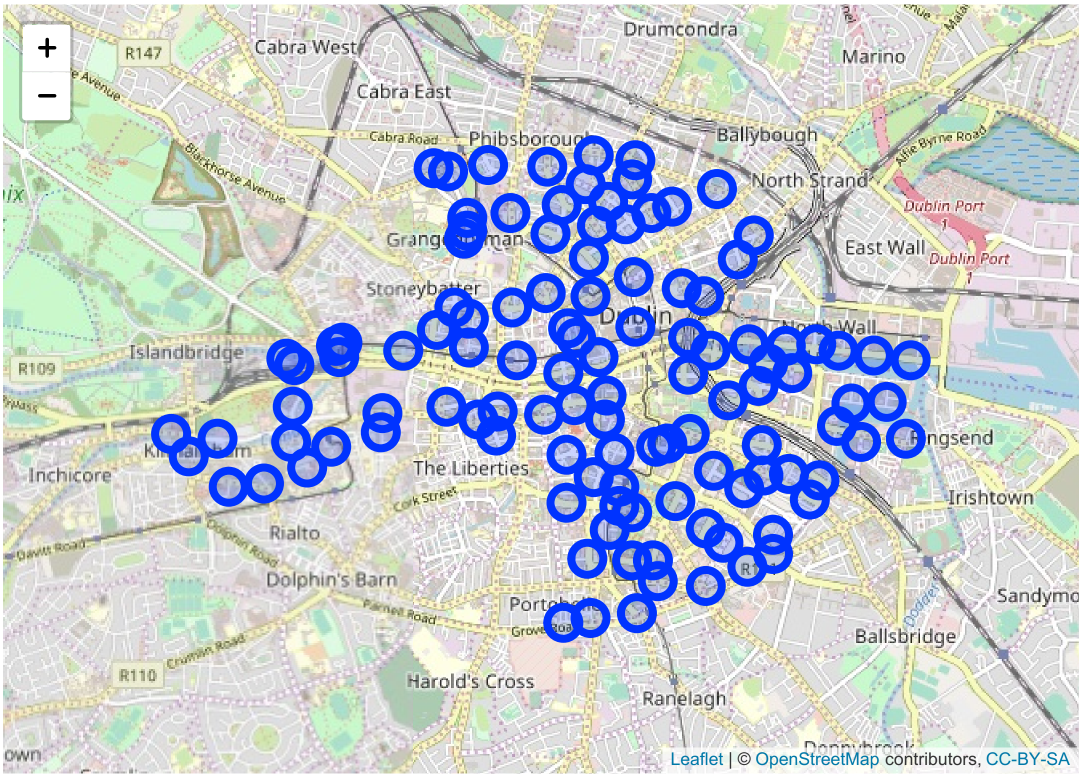
```
library(httr)
library(leaflet)
response <- GET('https://api.jcdecaux.com/vls/v1/stations?contract=dublin&apiKey=09edbcd3e39bc1b224db355e836
9d49edd9687cc')
result <- content(response)
bikedata <- data.frame(NULL)
limit <- length(result)
for (i in 1:limit) {
  df <- data.frame(result [i])
  bikedata <- rbind(bikedata, df)
}
View(bikedata)
count(data.frame(bikedata)) #Number of bike stands available
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   114
```

```
Name_of_bikestand <- data.frame(table(bikedata$name))
max(data.frame(bikedata$available_bike_stands))#Maximum number of available bike stands
```

```
## [1] 40
```

```
ava_max_bikestand <- subset(bikedata, available_bike_stands == 40)
bikedat2 <- data.frame(bikedata)#Number of bikes available
location_map <- leaflet(bikedat2) %>%
  addTiles() %>%
  addCircleMarkers(lng = ~ position.lng, lat = ~ position.lat, popup = ~available_bikes, label = ~as.charact
er(available_bikes))
location_map
```



2.2:Relevant information from this dataset.

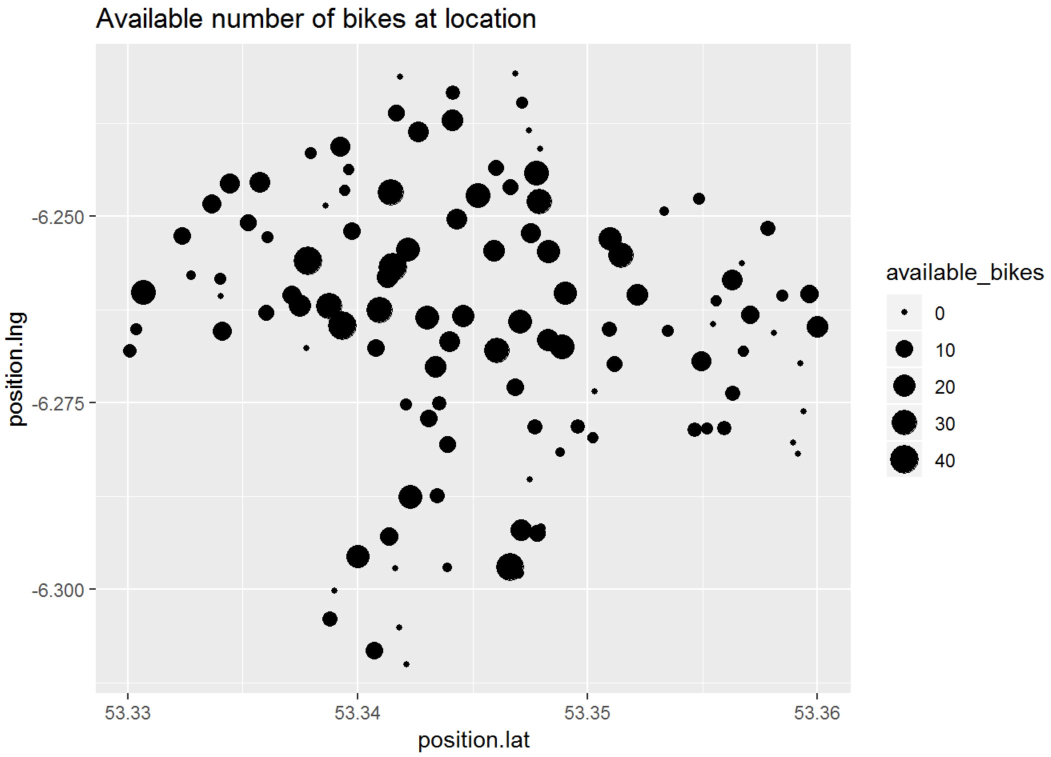
As a time pressured individual, I will look for available bike present near my location.

To analyze this situation I am filtering data with the bike stands which has bikes more than 5.I chose number 5, because if stand has only 1 or 2 it might be possible that someone might reach before and bike will not be available at that point. In this case probability of getting bike will be increased. I have created two graphs, One is simple Qplot and another one is geographical representation of data.

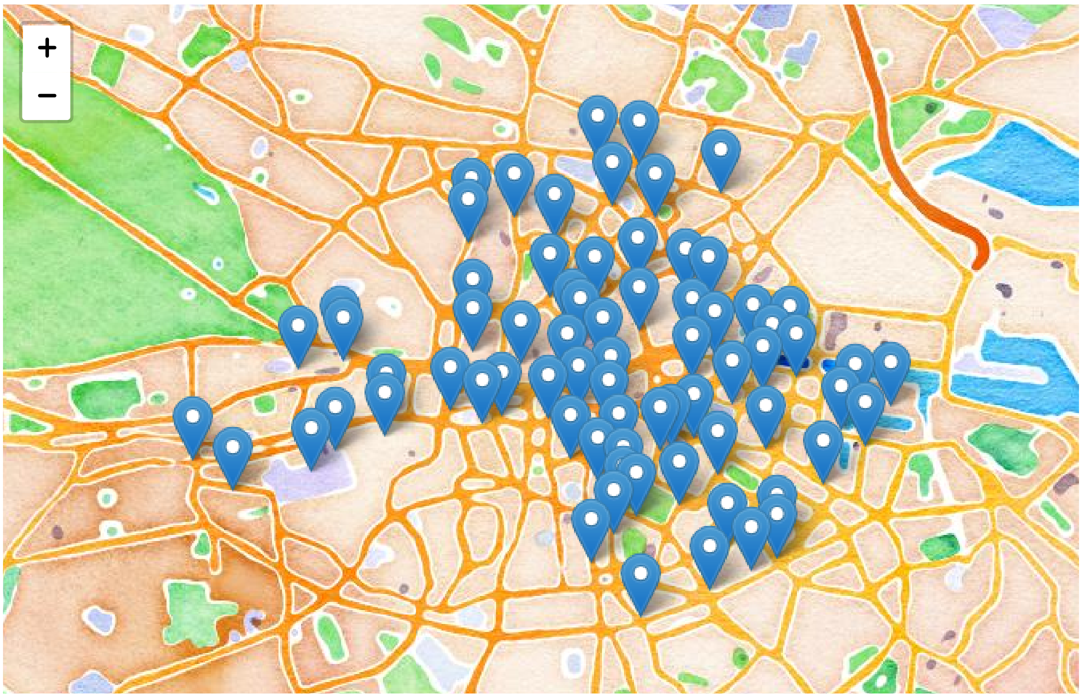
1. Plotted Qplot by using location and available bikes. This will give answer to questions at what location bikes are available and how many bikes are available . I have used bubble Qplot, in which greater number of bikes available at a location has bigger bubble.
2. This is another type representation of same data which I used in graph 1. I find it more interesting way of representing location and available number of bike. Hover on the location on map to see available number of bikes at a particular location.

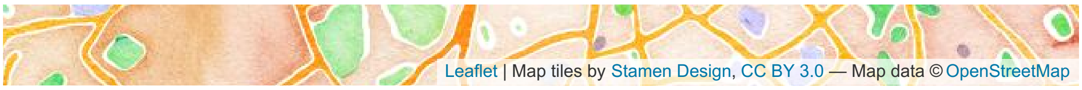
Variable used for the analysis is “bike_ave”.

```
bike_ave<- subset(bikedata, available_bikes >= 5, available_bike_stands >= 5)
bike_ave<- data.frame(bike_ave)
qplot(position.lat, position.lng, data = bikedata, size = available_bikes) +
  ggtitle("Available number of bikes at location")#graph1
```



```
#graph2
location_map <- leaflet(bike_ave) %>%
  addProviderTiles("Stamen.Watercolor") %>%
  addMarkers(lng = ~ position.lng, lat = ~ position.lat, popup = ~available_bikes, label = ~as.character(available_bikes))
location_map
```





Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap

Question3

3.1 : Initial Exploratory Analysis

Here I am using entire data set by joining different text file with the help of left_join. The variable name is “Final” which is combination of all the data.

For Exploratory Data Analysis, my main veriable of interests are Routes and Stops related variables.

For initial analysis need to used variables which provide number of stops at a particular location. There will analysis on routes available for Dublin bus and what are the numbers of stops available. Longest routes and area with maximum stops also needs to be analyzed, which will provide information on the routes covering maximum distance and maximum stops.

```
unzip("DublinBus.zip", exdir = 'googletransitdublinbusp20130315-1546')
stop_times <- read.table("stop_times.txt", sep = ",", header = TRUE)
colnames(stop_times)[1]<- "trip_id"
Trips <- read.table("trips.txt", sep = ",", header = TRUE)
Routes <- read.table("routes.txt", sep = ",", header = TRUE)
stops <- read.table("stops.txt", sep = ",", header = TRUE)
colnames(stops)[1] <- "stop_id"
Final <- left_join(Trips, Routes , by = 'ï..route_id') %>% left_join(stop_times, by = "trip_id")%>% left_joi
n(stops, by = "stop_id")
```

1.Below are the few observation of variables in the data set:

I have explained observation of data set in form of Question and Answers for better understanding of variables and data set.

a. Analysis of stops realted variables :

1. How many Stops are available.(Varibale used name Stop_id) Ans: 4275
2. How many different stops available with same name. (Variable used name stop_name) Ans: view “Name_of_Stops” variable.
3. Which road has maximum number of stops available. Ans : Blessington Rd with 79 stops

b. Analysis of routes related variables :

1. How many Routes are available in data set? (Varibale used route_short_name) Ans: 119
2. State the names of routes available in data set. (Variable used route_long_name) Ans: View “Name_of_routes” variable.

c. Analysis common in routes and stops:

1. longest route have how many stops? (Variable used stop_sequence) Ans : 109
2. name of route with maximum number od stops. Ans : St. Paul’s Crescent - Irish Life Mall

2. Other variable that are related to stops and routes.

1. Pick and Drop type of stop.(variable used pickup_type and drop_off_type) Ans: as value is 0 for all data set it is reregular pick up and drop off.(GTFS information available on site)
2. Number of routes with one direction and opposite direction. (Variable Used direction_id) Ans : one direction - 454618 opposite direction - 467675.

```
library(tidyverse)
#variable 1 (variables related to Stop)
No_of_Stops <- data.frame(table(Final$stop_id))
count(No_of_Stops) #there are 4724 stops available
```

```
## # A tibble: 1 x 1
##        n
##    <int>
## 1   4724
```

```
Name_of_Stops <- data.frame(table(stops$stop_name)) # Number of stops with same name
max(Name_of_Stops$Freq)#Road with maximum number of stops
```

```
## [1] 79
```

```
Name_of_RD_maxstops <- subset(Name_of_Stops, Freq == 79)# Name ofstop with maximum number of stops
No_of_Route_short_name <- data.frame(table(Final$route_short_name))#Number of routes avilable in data set
count(No_of_Route_short_name)# there are 119 routes available in data set
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   119
```

```
Name_of_routes <- table(Final$route_long_name)#Names of routes available
max(Final$stop_sequence)#longest route with stops
```

```
## [1] 109
```

```
Name_of_route1 <- subset(Final, stop_sequence == 109)
# pickup and drop off type of stop , these are other vaiable which are elated to Stops.
table(Final$pickup_type) #all are regular pickup scheduled for buses as value is 0
```

```
##
##      0
## 922293
```

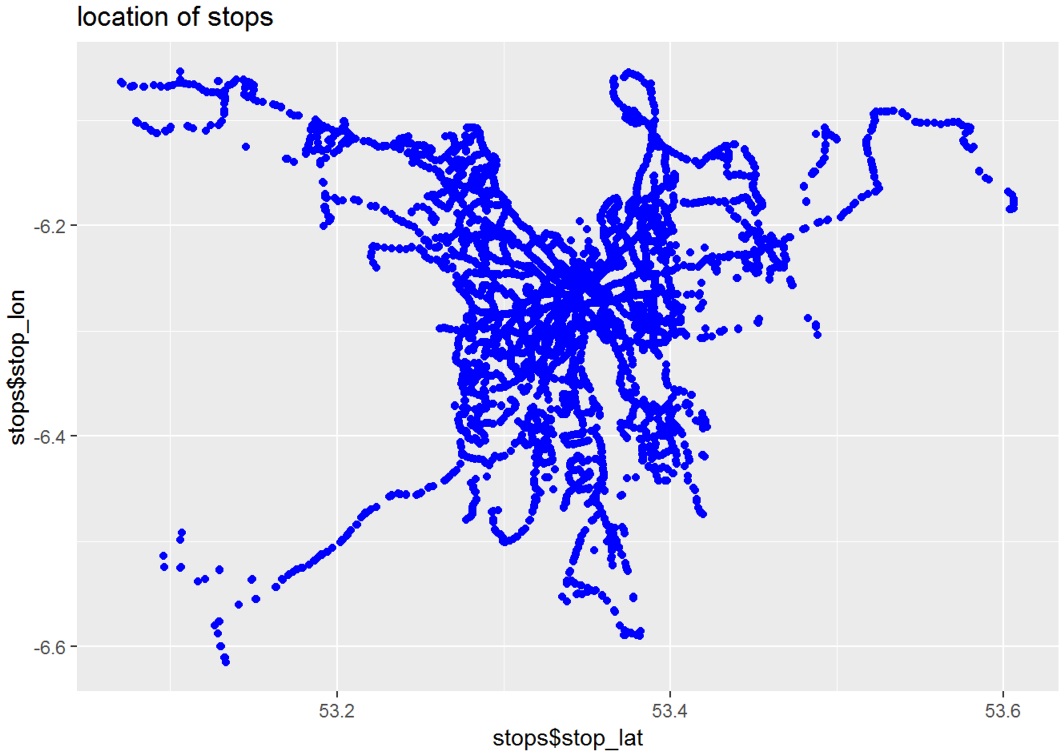
```
table(Final$drop_off_type) #all are regular drop off for buses as value is 0
```

```
##
##      0
## 922293
```

```
table(Final$direction_id)#Direction of route
```

```
##
##      0      1
## 454618 467675
```

```
Distance_travelled <- table(Final$shape_dist_traveled)#Distance travvelled on each route
#Below is the graph shows distibution of stops as per lattitude and longitude
ggplot(data= stops) +
  geom_point(mapping = aes(x = stops$stop_lat, y = stops$stop_lon), color = "blue") +
  ggtitle("location of stops")
```



3.2: Pick a bus route

Route from Swards Road to UCD.

For analysis of route we required data with number of stops at Origin and number of stops at destination. As per below analysis two routes “41X” and “32X” are there from which frequently travelled passenger can use to travel between Swards Road to UCD. The reason of using these routes are that direction of the route is available for forward and reverse direction both and frequency of these routes is also more. At time between 7:00 to 9:00 frequencies of buses on these routes is more which can be used for forward journey. Moreover for reverse journey also frequency of buses on these routes is good i.e. between 16:00 to 19:00. Below are the variable which has been used for analysis and approach towards the route. There are two geographical graphs which represent path followed by these two routes“41X” and “32X”.

a.Analysis for starting point:

1. What are the routes which are starting from “Swards road”? Ans: View data set “Home_stop”.
2. What are the frequency of these routes? Ans: View data “Routes_homey”.
3. Routes with highest frequency. Ans. View dataset “Valid_route”

b.Analysis for destination point:

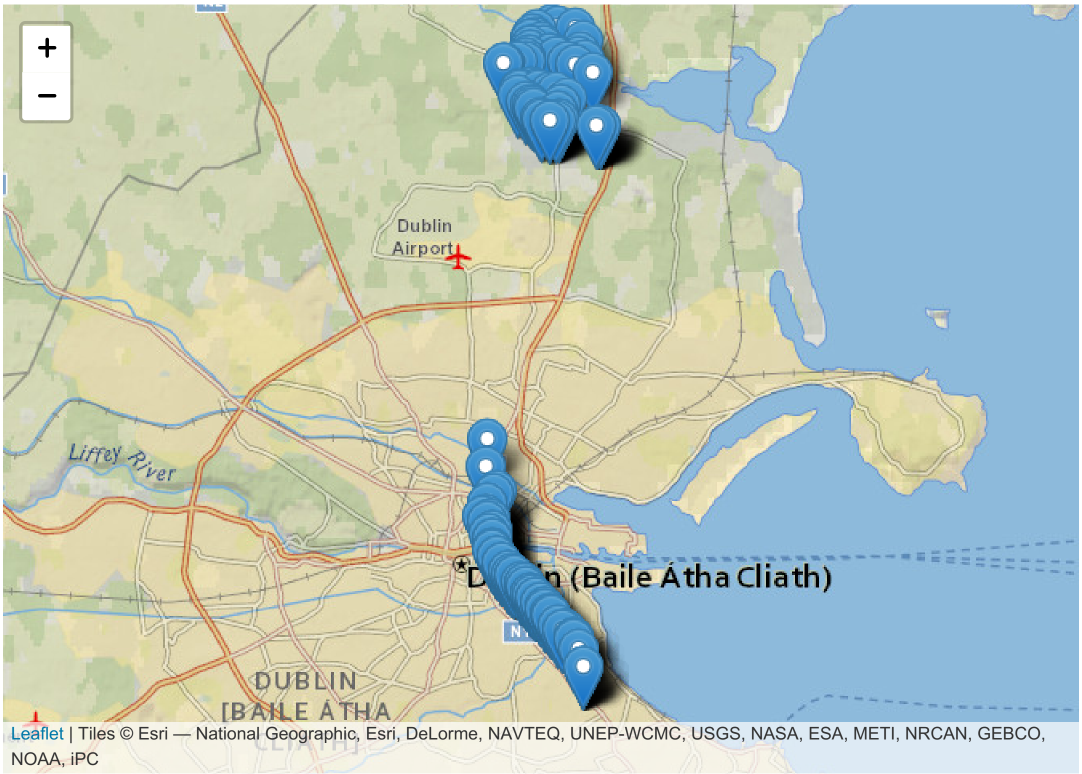
1. What are the routes which are going to “UCD”? Ans: View data set “Des_stop”.
2. What are the frequency of these routes? Ans: View data “Routes_Des”.
3. Routes with highest frequency. Ans. View dataset “valid_des_route”.

c. Analysis of routes:

1.Verify the common bus routes for Home and destination valid data. Ans: This will give two common routes i.e.“41X” and “32X”. 2.What are the timings for bus? Ans:View data set “Arr_Time1” and “Dep_time1” for route 1 i.e “41X”. View data set “Arr_Time1” and “Dep_time1” for route 1 i.e “32X”. 3.Direction ID for the routes. Ans:For Route1 Direction_IDs are : 0 = 128(buses at forward direction) 1 = 196(buses at reverse direcion) For Route2 Direction_IDs are : 0 = 94 (buses at forward direction) 1 = 88 (buses at reverse direcion)

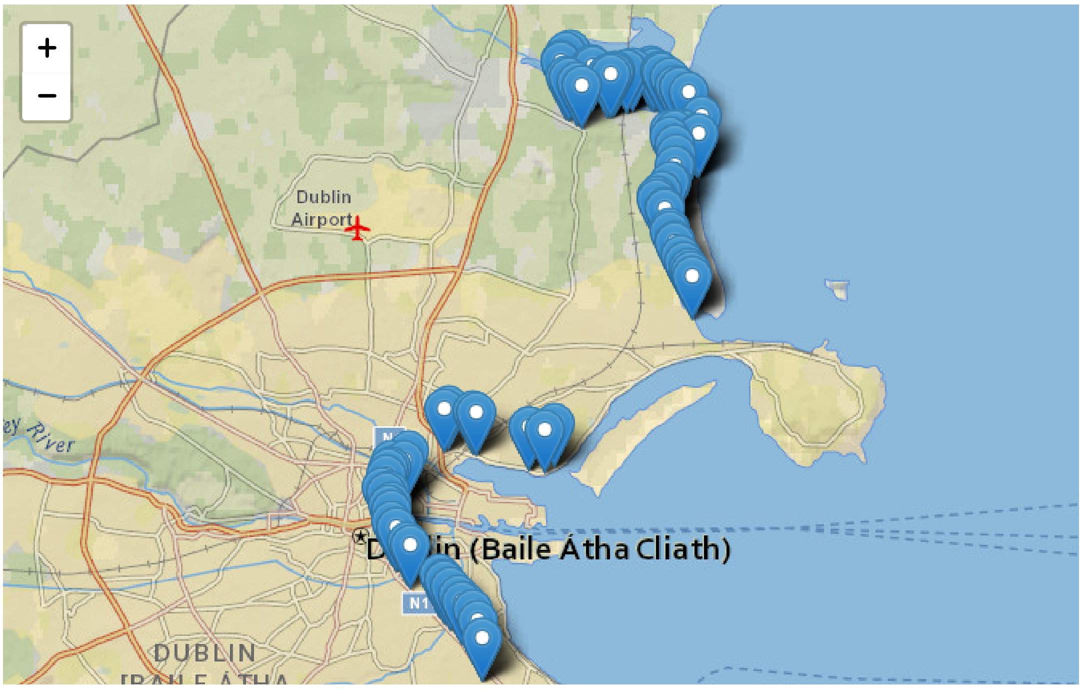
There are two plots, one will give route of “41X” and another give route of “32X”.

```
library(leaflet)
Home_stop <-subset(Final, stop_name == "Swords Road")#Home Point analysis
Routes_home <- data.frame(table(Home_stop$route_short_name))
Valid_route <- subset(Routes_home, Freq>=1)
Des_stop <- subset(Final, stop_name == "UCD")#destination point analysis
Routes_Des <- data.frame(table(Des_stop$route_short_name))
valid_des_route <- subset(Routes_Des, Freq >= 1)
My_route1 <-  filter(Final, route_short_name == "41X")#Route1 analysis
leaflet(My_route1) %>%
  addProviderTiles(providers$Esri.NatGeoWorldMap) %>%
  addMarkers(lng = ~ stop_lon, lat = ~ stop_lat)
```



Leaflet | Tiles © Esri — National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC

```
Arr_Time1 <- table(My_route1$arrival_time)#Time of the route1
Dep_time1 <- table(My_route1$departure_time)
Direction1 <- table(My_route1$direction_id)#Direction for route1
My_route2 <- filter(Final, route_short_name == "32X") #Route2 Analysis
leaflet(My_route2) %>%
  addProviderTiles(providers$Esri.NatGeoWorldMap) %>%
  addMarkers(lng = ~ stop_lon, lat = ~ stop_lat)
```



Leaflet | Tiles © Esri — National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC

```
Arr_Time2 <- table(My_route1$arrival_time)  #Time for the route2
Dep_time2 <- table(My_route1$departure_time)
Direction2 <- table(My_route2$direction_id)#Direction for route2
Direction2
```

```
##
##  0  1
## 94 88
```

Question4

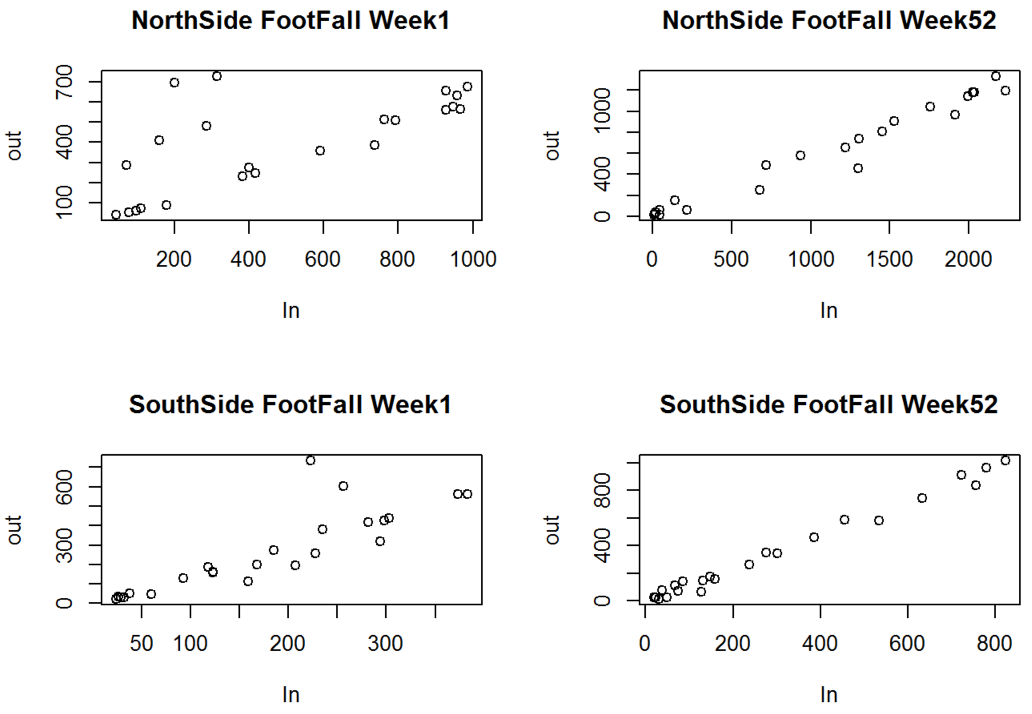
4.1 NorthSide and southSide FootFall of year 2013

Picked camera of “O’Connell Street at Clerys” which is situated at north side of Dublin and “South Great Georges” create which is situated at the south side of Dublin. Just by viewing data it is not possible to come up with some observation as all are discrete data. I have created a comparative analysis plot for Week1 and Week52 of North side camera and South side camera. I plotted all the 4 graphs in one frame for better understanding of comparison. Southside footfall of week 52 provides better correlation then other three weeks. There are multiple outliers in week 1 for south and north side so there is something happing in first week which are causing these outliers.

```
library(readODS)
get_num_sheet_in_ods("pedestrianfootfall2013.ods")
```

```
## [1] 52
```

```
Foot_fall<- data.frame(read_ods("pedestrianfootfall2013.ods", sheet = 1, formula_as_formula = TRUE))
#O'Connell Street at Clerys analysis
Week1_north <- read_ods("pedestrianfootfall2013.ods", sheet = 1, formula_as_formula = TRUE, range = "R7C4:R29C5", col_names = FALSE)
Week52_north <- read_ods("pedestrianfootfall2013.ods", sheet = 52, formula_as_formula = TRUE, range = "R8C4:R29C5", col_names = FALSE)
#South Great Georges analysis
Week1_south <- read_ods("pedestrianfootfall2013.ods", sheet = 1, formula_as_formula = TRUE, range = "R63C4:R86C5", col_names = FALSE)
Week52_south <- read_ods("pedestrianfootfall2013.ods", sheet = 52, formula_as_formula = TRUE, range = "R64C4:R86C5", col_names = FALSE)
par(mfrow = c(2, 2))
plot(Week1_north$D, Week1_north$E , xlab = "In" , ylab = "out", main = "NorthSide FootFall Week1")
plot(Week52_north$D, Week52_north$E, xlab = "In" , ylab = "out", main = "NorthSide FootFall Week52")
plot(Week1_south$D, Week1_south$E, xlab = "In" , ylab = "out", main = "SouthSide FootFall Week1")
plot(Week52_south$D, Week52_south$E, xlab = "In" , ylab = "out", main = "SouthSide FootFall Week52")
```



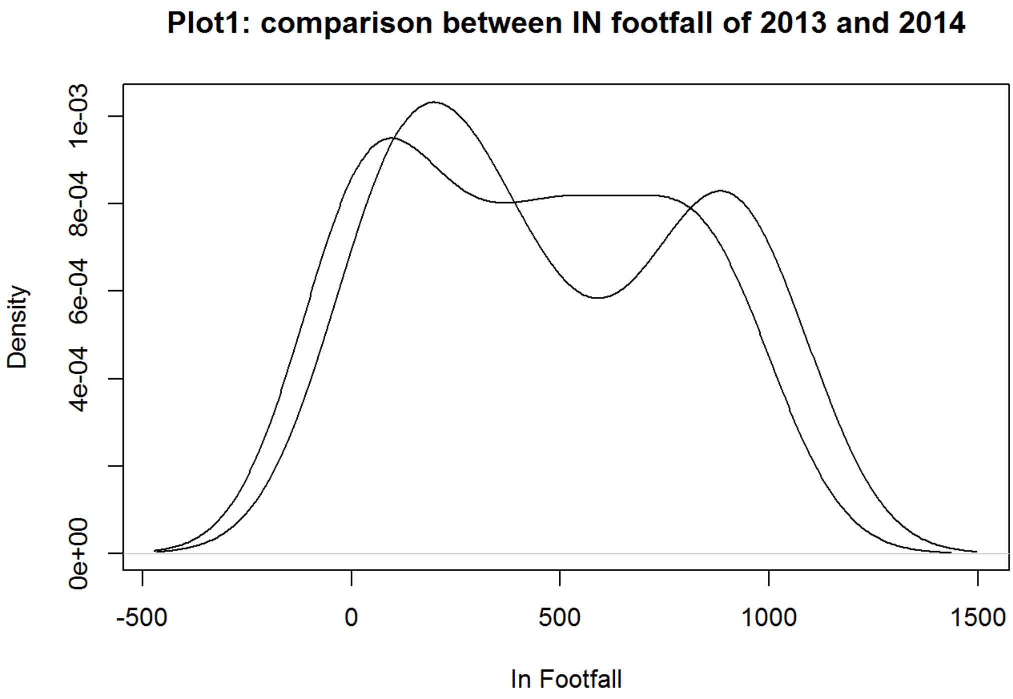
4.2 Comparative plot against the footfall data from one of the cameras.

I have taken camera of “O’Connell Street at Clerys” and compare IN and OUT Footfall for 2013 and 2014. Approach of analysis was to compare density of footfall in the same region in two different years. If we will compare between density of “In” Footfall in year 2013 and 2014 then in 2013 it is bit fluctuating but in 2014 density curve is flat. Little skewness is observed at the right side of the plot. However the comparison of “out” Footfall shows lot of changes in year 2013 and 2014. In year 2014 the density of “OUT” Footfall data is neither increasing nor decreasing. The “OUT” data of 2013 has fluctuation and most of the values can be considered moving at the left of the graph.

Explanation of code:

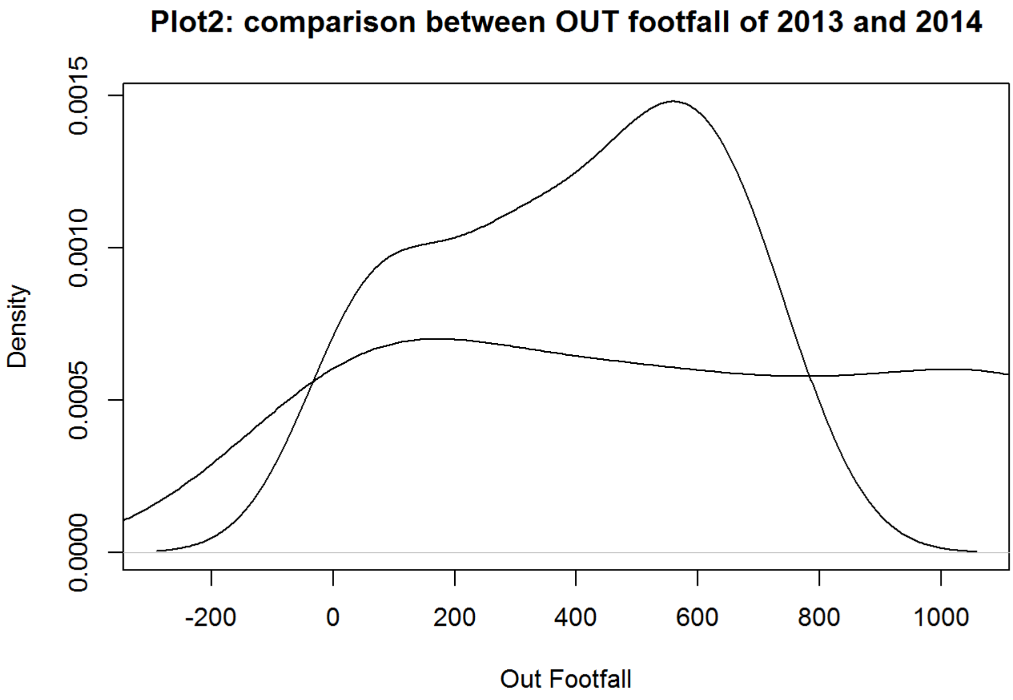
1. I have created four variables containing density of “IN” and “OUT” data of 2013 and 2014.
2. Week1_north is “In” Footfall data of 2013 and Week1 is data of 2014. I have created my first density plot by comparing these two data sets. The names of variable is “x” and “Z”.
3. Week52_north is “OUT” Footfall data of 2013 and Week52 is data of 2014. My another density plot is comparing these two data with variable name “Y” and “W”.

```
library(readxl)
Week1 <- read_excel("dublinfootfall2014.xls", sheet = 1 , range = "D7:E31")
Week52 <- read_excel("dublinfootfall2014.xls", sheet = 1 , range = "F7:G31")
x <- density(as.numeric(Week1_north$D))
y <- density(as.numeric(Week1_north$E))
Z <- density(as.numeric(Week52$In))
w <- density(as.numeric(Week1$Out))
plot(x, main = "Plot1: comparison between IN footfall of 2013 and 2014", xlab = "In Footfall") + lines(Z)
```



```
## integer(0)
```

```
plot(y, main = "Plot2: comparison between OUT footfall of 2013 and 2014", xlab = "Out Footfall")+ lines(w)
```



```
## integer(0)
```

Question 5

5.1 : Analytics

Approach of this data set is to convert it into text file and create two columns. One column “V1” will contain timestamp of command and column “V2”contains the name of activity which executed on R console. First of all, the cleaning of data required. For cleaning of data I have replaced blank,“{”, “)” and similar type of text which does not have any significance with “NA” this stored in data set named “History”. After cleaning of data need to convert the time field into readable format which has been done in variable named “History1”. As previous question have some keywords like Week , par function , plot function so I wanted to draw a graph that will show time taken by each command in the previous question. However, limitation of data is not allowing to draw graph therefore proceeded with basic validation. Explanation of code is below:

```
    1. Created a "History" data set with data wrangling.
    2. Find out number of rows in dataset "History".
    3. Counting the number of "na" in each column.
    4. created three variables first is "variable_with_week" which have history data that contain value rela
ted to Week. Second variable is  "Commad_par" which contain variable that start with command par. Third and
the last variable is to "Command_plot" which contain variable that contain command related to plot.
```

```
library(sqldf)
library(lubridate)
History <- read.table("R_history.txt",sep = ":", fill = TRUE, header = FALSE, na= c("{", "}", "", ")","(",
"```"))
nrow(History)
```

```
## [1] 5479
```

```
History1 <- as_datetime(as.numeric(History$V1), tz = 'GMT')
sum(is.na(History$V1))
```

```
## [1] 51
```

```
sum(is.na(History$V2))
```

```
## [1] 484
```

```
variables_with_Week <- sqldf("select * from History where V2 like'Week%'")
Command_par <- sqldf("select * from History where V2 like'par%'")
Command_plot <- sqldf("select * from History where V2 like 'plot%'")
```

5.2 : Personal Reflection of project.

As learning prospective project was good but it is time consuming. For each and every question lot of research required. Single line of code is demanding lot of background research. Sometime it is very difficult to understand data and huge volume of data is taking lot of time to read.

1. Challenges Overcome: First biggest challenge is to understand the flow of data and analyses the dependency of variables. To overcome this analysis of data in chunks will be beneficial to avoid confusions. Another challenge to plot the data because having the result does not make sense until and unless we are able to visualize it. The major part of work around taken in task 5 because data is unstructured and it very difficult to find useful information. By converting the time also not able to analyze the time taken by each command to run. So need to follow another approach to explore the data set by creating variables.
2. Efficient Work practice: The efficient work practice is to analyze each part of data first then create questions for the data.
3. Area of Frustration: There are multiple areas in the project where it seems very difficult to overcome. First is to get location on map of Dublin because as Google introduced API key for using map it become difficult to find another package to plot a graph. Reading with ODS file was difficult as it was difficult to put for look of that huge amount of data.
4. Time management: Each question was very time consuming but time can be saved by using same package in multiple questions. Although able to find answers of maximum question but plotting map and find routes was difficult to manage and get the result in short amount of time.