ANA 515 Assignment 4 Data Analytics Project

Mohammed Sohail Ameer

2022-10-15

## Business Goal

Business problem or goal is a essential part of a business dataset analysis. Here we are utilized Uber’s data where we are finding the goals or problems in it. Here the goal is to find out the frequency of a customer travelled in Uber and what type of service they are willing to avail. Again to achieve a good service and customer reputation, the cancelled and completed trips are also analysed with their fare, distance, ride time etc.

## Dataset Retrieval

This Uber’s data is retrieved from the following link:

<https://drive.google.com/file/d/1emopjfEkTt59jJoBH9L9bSdmlDC4AR87/view>

## Intializing required libraries

library(ggplot2)  
library(ggthemes)

## Warning: package 'ggthemes' was built under R version 4.2.1

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
library(DT)  
library(scales)  
library(Hmisc)

## Warning: package 'Hmisc' was built under R version 4.2.1

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

## Loading Data

apr\_data <- read.csv("Uber/uber-raw-data-apr14.csv")  
may\_data <- read.csv("Uber/uber-raw-data-may14.csv")  
jun\_data <- read.csv("Uber/uber-raw-data-jun14.csv")  
jul\_data <- read.csv("Uber/uber-raw-data-jul14.csv")  
aug\_data <- read.csv("Uber/uber-raw-data-aug14.csv")  
sep\_data <- read.csv("Uber/uber-raw-data-sep14.csv")  
  
data\_2014 <- rbind(apr\_data,may\_data, jun\_data, jul\_data, aug\_data, sep\_data)

## Describing Data

describe((data\_2014))

## (data\_2014)   
##   
## 4 Variables 4534327 Observations  
## --------------------------------------------------------------------------------  
## Date.Time   
## n missing distinct   
## 4534327 0 260093   
##   
## lowest : 4/1/2014 0:00:00 4/1/2014 0:01:00 4/1/2014 0:02:00 4/1/2014 0:03:00 4/1/2014 0:04:00  
## highest: 9/9/2014 9:55:00 9/9/2014 9:56:00 9/9/2014 9:57:00 9/9/2014 9:58:00 9/9/2014 9:59:00  
## --------------------------------------------------------------------------------  
## Lat   
## n missing distinct Info Mean Gmd .05 .10   
## 4534327 0 7092 1 40.74 0.04024 40.67 40.69   
## .25 .50 .75 .90 .95   
## 40.72 40.74 40.76 40.77 40.79   
##   
## lowest : 39.6569 39.6686 39.7214 39.8416 39.9055  
## highest: 41.3730 41.3737 41.5016 41.5276 42.1166  
## --------------------------------------------------------------------------------  
## Lon   
## n missing distinct Info Mean Gmd .05 .10   
## 4534327 0 11453 1 -73.97 0.04633 -74.01 -74.01   
## .25 .50 .75 .90 .95   
## -74.00 -73.98 -73.97 -73.94 -73.87   
##   
## lowest : -74.9290 -74.8594 -74.8575 -74.8260 -74.8258  
## highest: -72.3304 -72.3097 -72.2999 -72.1801 -72.0666  
## --------------------------------------------------------------------------------  
## Base   
## n missing distinct   
## 4534327 0 5   
##   
## lowest : B02512 B02598 B02617 B02682 B02764, highest: B02512 B02598 B02617 B02682 B02764  
##   
## Value B02512 B02598 B02617 B02682 B02764  
## Frequency 205673 1393113 1458853 1212789 263899  
## Proportion 0.045 0.307 0.322 0.267 0.058  
## --------------------------------------------------------------------------------

## Data Preparation

data\_2014$Date.Time <- as.POSIXct(data\_2014$Date.Time, format = "%m/%d/%Y %H:%M:%S")  
data\_2014$Time <- format(as.POSIXct(data\_2014$Date.Time, format = "%m/%d/%Y %H:%M:%S"), format="%H:%M:%S")  
data\_2014$Date.Time <- ymd\_hms(data\_2014$Date.Time)  
data\_2014$day <- factor(day(data\_2014$Date.Time))  
data\_2014$month <- factor(month(data\_2014$Date.Time, label = TRUE))  
data\_2014$year <- factor(year(data\_2014$Date.Time))  
data\_2014$dayofweek <- factor(wday(data\_2014$Date.Time, label = TRUE))  
data\_2014$hour <- factor(hour(hms(data\_2014$Time)))  
data\_2014$minute <- factor(minute(hms(data\_2014$Time)))  
data\_2014$second <- factor(second(hms(data\_2014$Time)))

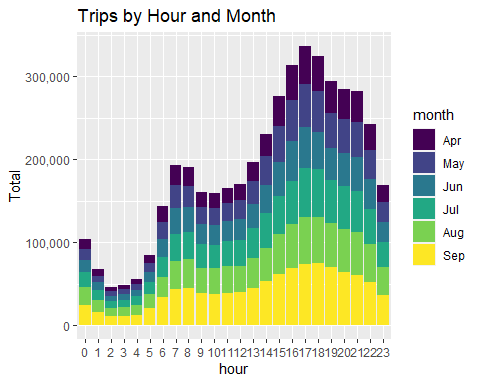
## Data Visualization

Here a bar plot has been developed with Trips by Hour and Month.

hour\_data <- data\_2014 %>%  
 group\_by(hour) %>%  
 dplyr::summarize(Total = n())   
month\_hour <- data\_2014 %>%  
 group\_by(month, hour) %>%  
 dplyr::summarize(Total = n())

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

ggplot(month\_hour, aes(hour, Total, fill = month)) +   
 geom\_bar( stat = "identity") +  
 ggtitle("Trips by Hour and Month") +  
 scale\_y\_continuous(labels = comma)

 ## Data Visualization

Here a heat map plot has been developed with Heat Map by Hour and Day.

day\_and\_hour <- data\_2014 %>%  
 group\_by(day, hour) %>%  
 dplyr::summarize(Total = n())

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

ggplot(day\_and\_hour, aes(day, hour, fill = Total)) +  
 geom\_tile(color = "black") +  
 ggtitle("Heat Map by Hour and Day")

