**UBER DATA ANALYSIS**

**Project Code:**

library(ggplot2)

library(ggthemes)

library(lubridate)

library(DT)

library(scales)

library(ggmap)

library(tidyverse)

library(lubridate)

library(knitr)

apr\_data <- read.csv("uber-raw-data-apr14.csv")

may\_data <- read.csv("uber-raw-data-may14.csv")

jun\_data <- read.csv("uber-raw-data-jun14.csv")

jul\_data <- read.csv("uber-raw-data-jul14.csv")

aug\_data <- read.csv("uber-raw-data-aug14.csv")

sep\_data <- read.csv("uber-raw-data-sep14.csv")

data\_2014 <- rbind(apr\_data,may\_data, jun\_data, jul\_data, aug\_data, sep\_data)

summary(data\_2014)

data\_2014

dim(data\_2014)

any(is.na(data\_2014))

sum(is.na(data\_2014))

data\_2014$Date.Time <- strptime(data\_2014$Date.Time, format = "%m/%d/%Y %H:%M:%S")

data\_2014$Day <- as.integer(format(data\_2014$Date.Time, "%d"))

data\_2014$Week\_day <- as.integer(format(data\_2014$Date.Time, "%w"))

data\_2014$Time\_hour <- as.integer(format(data\_2014$Date.Time, "%H"))

data\_2014$Time\_minute <- as.integer(format(data\_2014$Date.Time, "%M"))

data\_2014$Time\_second <- as.integer(format(data\_2014$Date.Time, "%S"))

data\_2014$month <- factor(month(data\_2014$Date.Time, label = TRUE))

data\_2014$Particular\_Day <- factor(wday(data\_2014$Date.Time, label = TRUE))

data\_2014

summary(data\_2014)

count(data\_2014, month)

count(data\_2014, Day)

library(dplyr)

Month\_analysis <- data\_2014 %>% group\_by(month) %>% dplyr::summarise(Total=n())

Month\_analysis

set.seed(20)

clusters <- kmeans(data\_2014[,2:3], 5)

data\_2014$cluster\_no <- as.factor(clusters$cluster)

str(clusters)

print(clusters)

library(factoextra)

hcluster <-hkmeans(data\_2014[1:10000,2:3], 5)

names(hcluster)

hcluster

hkmeans\_tree(hcluster, cex = 0.6)

fviz\_cluster(hcluster, frame.type = "norm", frame.level = 0.68)

fviz\_cluster(hcluster, palette = "jco", repel = TRUE,

ggtheme = theme\_classic())

month\_weekday <- data\_2014 %>% group\_by(month, Particular\_Day) %>% dplyr::summarize(Total = n())

month\_weekday

ggplot( Month\_analysis, aes(x = month, y = Total, color="blue", fill = month)) +

geom\_bar( stat = "identity") +

ggtitle("Trips by Month") +

theme(plot.title = element\_text(colour = "darkred", size = 25),

axis.title = element\_text(color = "darkred", size = 25))

ggplot(month\_weekday, aes(x = month, y = Total, fill = Particular\_Day)) +

geom\_bar( stat = "identity", color="blue", position = "dodge") +

ggtitle("Trips by Day and Month")

theme(plot.title = element\_text(colour = "darkred", size = 25),

axis.title = element\_text(color = "darkred", size = 25))

average\_weekday <- data\_2014 %>% group\_by(Week\_day) %>% summarise(average = n() / 24)

average\_weekday

average\_perday <- data\_2014 %>% group\_by(Day) %>% summarise(average = n() / 6)

average\_perday

ggplot(average\_weekday, aes(Week\_day, average)) +

geom\_line(aes(group = 1), size = 1.5, col = "orangered2") +

geom\_point(size = 3, col = "deepskyblue4") +

ggtitle("Avg count of trips per week day") +

scale\_x\_continuous(breaks = seq(0, 7, 1)) +

theme(plot.title = element\_text(colour = "darkred", size = 20),

axis.title = element\_text(color = "darkred", size = 12))

month\_weekday <- data\_2014 %>%

group\_by(Particular\_Day, month) %>%

dplyr::summarize(Total = n())

month\_weekday

month\_base <- data\_2014 %>%

group\_by(Base, month) %>%

dplyr::summarize(Total = n())

month\_base

ggplot(month\_weekday, aes(Particular\_Day, month, fill = Total)) +

geom\_tile(color = "Black") +

ggtitle("Heat Map by Month and Day of Week")

monthly\_growth <- month\_borough\_14 %>%

mutate(Date = paste("04", Month)) %>%

ggplot(aes(Month, n, colour = Borough)) + geom\_line() +

ggtitle("Uber Monthly Growth - 2014")

monthly\_growth

ggplot(hour\_data, aes(hour, Total)) +

geom\_bar( stat = "identity", fill = "steelblue", color = "red") +

ggtitle("Trips Every Hour") +

theme(legend.position = "none") +

scale\_y\_continuous(labels = comma)

ggplot(month\_hour, aes(hour, Total, fill = month)) +

geom\_bar( stat = "identity") +

ggtitle("Trips by Hour and Month") +

scale\_y\_continuous(labels = comma)

ggplot(day\_group, aes(day, Total)) +

geom\_bar( stat = "identity", fill = "steelblue") +

ggtitle("Trips Every Day") +

theme(legend.position = "none") +

scale\_y\_continuous(labels = comma)

min\_lat <- 40.5774

max\_lat <- 40.9176

min\_long <- -74.15

max\_long <- -73.7004

ggplot(data\_2014, aes(x=Lon, y=Lat)) +

geom\_point(size=1, color = "blue") +

scale\_x\_continuous(limits=c(min\_long, max\_long)) +

scale\_y\_continuous(limits=c(min\_lat, max\_lat)) +

theme\_map() +

ggtitle("NYC MAP BASED ON UBER RIDES DURING 2014 (APR-SEP)")

clusters$cluster

head(clusters$cluster,5)

clusters$size

clusters$centers

plot(clusters$centers)