R Notebook

INTRODUCTION

The purpose of this study is to examine Uber and other for-hire vehicle (FHV) pick-up data in New York City. Our client will be able to make decisions based on the results of the analysis. A descriptive and exploratory study was requested by the client to compare Uber pickups against other for-hire cars.

DATASET: https://github.com/fivethirtyeight/uber-tlc-foil-response

Given the project's scope and timeline, I concentrated on giving a high-level comparison of trends in Uber and other FHV providers. As a result, the majority of my research is focused on the aggregated trip dataset.

DATASET: https://github.com/fivethirtyeight/uber-tlc-foil-response/Aggregate%20FHV%20Data.xlsx

1. Data Understanding From the provided URL the data was fetched and stored into current working directory.

```
library(readxl)
url = "https://github.com/fivethirtyeight/uber-tlc-foil-response/raw/master/Aggregate%20FHV%20Data.xlsx
des_data = paste(getwd(),"Aggregate FHV Data.xlsx", sep = '/')
download.file(url, destfile = des_data ,mode="wb")
FHV_data = read_excel("Aggregate FHV Data.xlsx", sheet=1)
```

These are the fields and information provided in this dataset.

```
FHV_data
```

```
## # A tibble: 92 x 13
##    Date                American Carmel `Dial 7` Diplo Firstclass Highclass
##    <dttm>                 <dbl>  <dbl>    <dbl> <dbl>      <dbl>     <dbl>
##  1 2014-07-01 00:00:00      921   2871     2233  1046       1744      1368
##  2 2014-07-02 00:00:00     1028   2965     2409  1275       2228      1661
##  3 2014-07-03 00:00:00     1068   3361     2520  1200       2121      1599
##  4 2014-07-04 00:00:00     1008   2174     1955  1171       1459      1622
##  5 2014-07-05 00:00:00     1214   1846     1371  1371       1703      1898
##  6 2014-07-06 00:00:00     1048   2480     1872  1251       1501      1738
##  7 2014-07-07 00:00:00      893   3028     2213  1009       1768      1457
##  8 2014-07-08 00:00:00      916   2706     2073  1065       1815      1387
##  9 2014-07-09 00:00:00      841   2883     2209   987       1827      1342
## 10 2014-07-10 00:00:00      823   3222     2425   904       1746      1367
## # ... with 82 more rows, and 6 more variables: Prestige <dbl>, Skyline <dbl>,
## #   Lyft <dbl>, Uber <dbl>, `Yellow Taxis` <dbl>, `Green Taxis` <dbl>
```

```
FHV_colname = colnames(FHV_data)
FHV_colname[2:13]
```

```
##  [1] "American"     "Carmel"       "Dial 7"       "Diplo"        "Firstclass"
##  [6] "Highclass"    "Prestige"     "Skyline"      "Lyft"         "Uber"
## [11] "Yellow Taxis" "Green Taxis"
```

The information contains the company names.

2. Preparation of the Data.

If you go thgrough the data there are no missing values. An extra column was added to help with the analysis, and new column names were assigned.

```
library(reshape)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following objects are masked from 'package:reshape':
##
##     colsplit, melt, recast
```

```
FHV_colnames = colnames(FHV_data)

reshape = melt(FHV_data, id.vars = c("Date"),
             measure.vars = FHV_colnames[2:13])
```

```
FHV_data <- reshape
colnames(FHV_data) <-  c("Date","Company","TripPerDay")

FHV_data$Year <- as.numeric(format(reshape$Date,'%Y'))
FHV_data$Month <- as.numeric(format(reshape$Date,'%m'))
FHV_data$Day <- as.numeric(format(reshape$Date,'%d'))
FHV_data$Weekday <- weekdays(as.Date(reshape$Date))

head(reshape, 5)
```

```
##         Date variable value
## 1 2014-07-01 American   921
## 2 2014-07-02 American  1028
## 3 2014-07-03 American  1068
## 4 2014-07-04 American  1008
## 5 2014-07-05 American  1214
```

The weekdays were set for the data and was set starting from Monday. Created a copy of dataset including the non-zero values.

```
FHV_data$Weekday = factor(FHV_data$Weekday, levels = c("Monday","Tuesday","Wednesday","Thursday" ,"Frida
null_val_Lyft = FHV_data[!(FHV_data$Company=='Lyft'& FHV_data$TripPerDay==0),]
```

3. DATA ANALYSIS

I. ASSERTIVE COMPANIES

Here, we're looking to figure out how big these firms are in terms of market share, which we can accomplish by looking at their total number of trips. I start by looking at these firms' minimum and maximum number of trips each day. Based on the maximum number of trip the Company names are arranged.

```
library(moments)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:reshape':
##
##     rename

## The following objects are masked from 'package:stats':
```

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

```
min_trips = null_val_Lyft %>%
  group_by(Company)%>%
  summarize(minTrip=min(TripPerDay))

max_trips = FHV_data %>%
  group_by(Company)%>%
  summarize(maxTrip=max(TripPerDay))

FHV_trip_stats = merge(min_trips, max_trips)
FHV_trip_stats = arrange(FHV_trip_stats,- maxTrip)
FHV_trip_stats
```

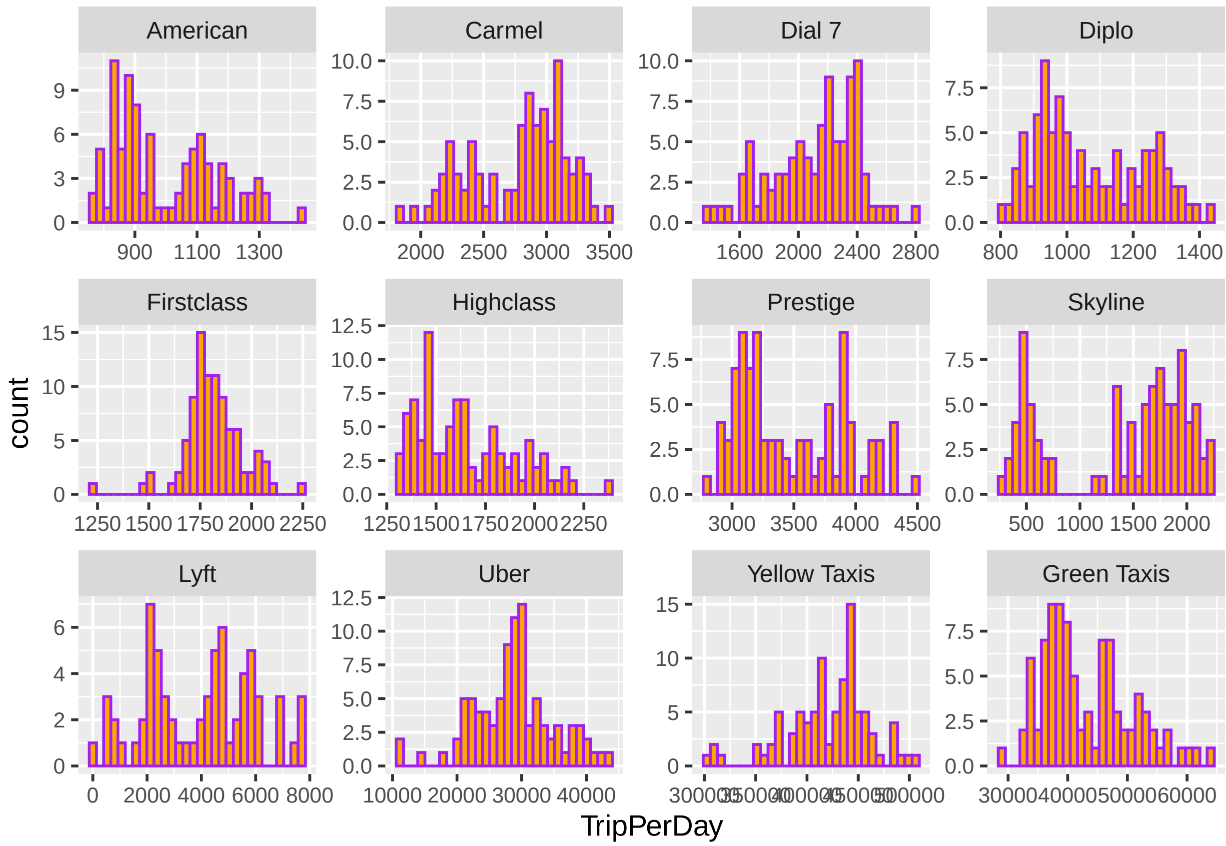
```
##         Company minTrip maxTrip
## 1  Yellow Taxis  305653  509480
## 2   Green Taxis   29186   64184
## 3          Uber   10890   43205
## 4          Lyft      40    7740
## 5      Prestige    2781    4470
## 6        Carmel    1846    3507
## 7        Dial 7    1371    2795
## 8     Highclass    1315    2375
## 9       Skyline     276    2230
## 10   Firstclass    1211    2228
## 11     American     768    1440
## 12        Diplo     810    1440
```

The Histogram plot below shows the companies per day trips.

```
library(ggplot2)

max_trips <- max_trips [order(max_trips$maxTrip),]
FHV_data$Company <- factor(FHV_data$Company, levels=max_trips$Company)

ggplot(null_val_Lyft, aes(x=TripPerDay))+
  geom_histogram(bins=30,color="purple", fill="orange")+
  theme(axis.text = element_text(size = 8))+
  facet_wrap(.~Company ,scales='free', ncol=4)
```



From this Histogram we can analyse the number of trips per day of each company. The top 3 operators can be easily identified: 1. YELLOW TAXIS

2. GREEN TAXIS

3. UBER

The top 3 operators, looking at the market share are the same with Yellow Taxis have the highest market share followed by Green Taxis and Uber.

```
FHV_market_share <- FHV_data%>%
  group_by(Company)%>%
  summarise(sum_trips=sum(TripPerDay))

FHV_market_share$percent <- round(FHV_market_share$sum_trips/sum(FHV_market_share$sum_trips)*100,2)
FHV_market_share <- arrange(FHV_market_share,-percent)
FHV_market_share
```

```
## # A tibble: 12 x 3
##    Company      sum_trips percent
##    <fct>            <dbl>   <dbl>
##  1 Yellow Taxis  38768702   82.4
##  2 Green Taxis    3975664    8.45
##  3 Uber           2653532    5.64
##  4 Prestige        320641    0.68
##  5 Lyft            267701    0.57
##  6 Carmel          256519    0.54
##  7 Dial 7          194992    0.41
##  8 Firstclass      166769    0.35
```

```
##   9 Highclass       151925    0.32
## 10 Skyline         127696    0.27
## 11 Diplo            98550    0.21
## 12 American         91712    0.19
```

‘

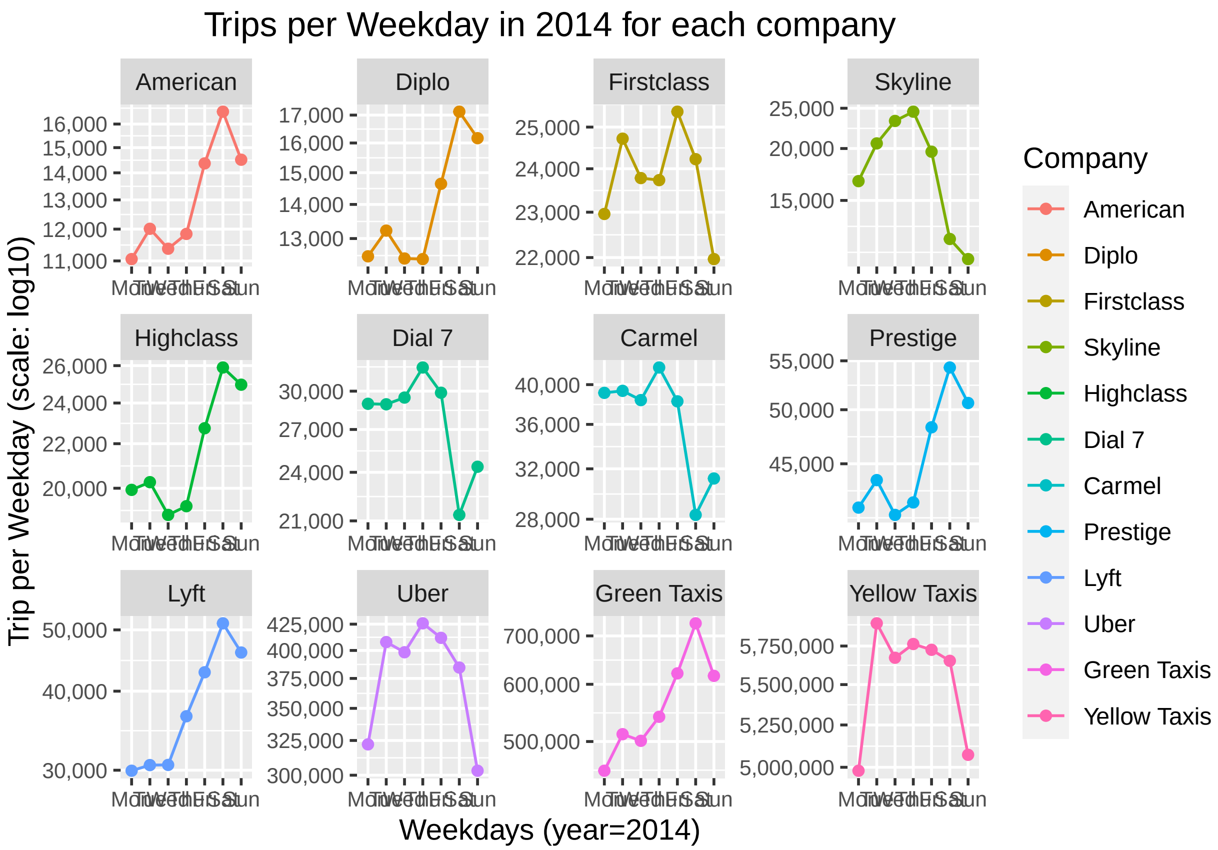
II. Finding Trends and Patterns in the Trip.

We can see each company has its own pattern and trend depending upon their trips. To have a better knowledge of the potential weekday or weekly patterns, I plotted the number of trips fro each company versus the weekdays.

```
FHV_per_week_trip <- FHV_data %>%
  group_by(Company,Weekday) %>%
  summarise(sum_trips_weekday = sum(TripPerDay))
```

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

```
ggplot(data=FHV_per_week_trip, aes(x = Weekday , y = sum_trips_weekday, group=Company, colour=Company))
  geom_line() +
  geom_point()+
  scale_y_log10(labels=scales::comma)+
  scale_x_discrete(labels=c("Monday"="Mon","Tuesday"="Tue",
                            "Wednesday"="Wed","Thursday"="Thu" ,
                            "Friday"="Fri" ,"Saturday"="Sat","Sunday"="Sun"))+
  labs(title="Trips per Weekday in 2014 for each company",
       x = "Weekdays (year=2014)",
       y = "Trip per Weekday (scale: log10)")+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text = element_text(size=8))+
  facet_wrap(.~Company ,scales='free', ncol=4)
```



Here, every company has different trends followed by the changes in the number of trips over weekdays. For example we can see a quite start for Yellow Taxis on Monday and a hugh spike on Tuesday, During working days, Green Taxis see a significant rise in journeys, with Saturdays seeing the biggest number of trips. Uber appears to follow a pattern similar to Yellow Taxis, whereas Lyft appears to follow Green Taxis.

Prestige, Highclass, Diplo, and American has a noticeable tend which is calm Wednesdays and Thursdays are also visible, with a jump on Fridays and Saturdays. The other trend which seems unique is that the companies Dail7 and Carmel have skipe starting monday and looks quite on Saturday.

I re-plotted this chart to see the possible similarities in these trends more clearly, putting organisations with similar patterns and trends together.

Data Normalisation will help us identify more similarities in the pattern. normalize function will help us scale this values between 0 and 1. Based on the minimum and maximum trips per day, The data is normalized for each company and an extra field is added to represent normalized values for sum of trips for each company.

```
Func_normalisation <- function(a) {
  return ((a - min(a)) / (max(a) - min(a)))
}

null_val_Lyft%>%
  group_by(Company)%>%
  mutate(Normalized_TripPerDay = Func_normalisation(TripPerDay)) -> null_val_Lyft

FHV_per_week_trip_normalized <- null_val_Lyft %>%
  group_by(Company,Weekday) %>%
  summarise(sum_trips_weekday_normalized = sum(Normalized_TripPerDay))
```

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

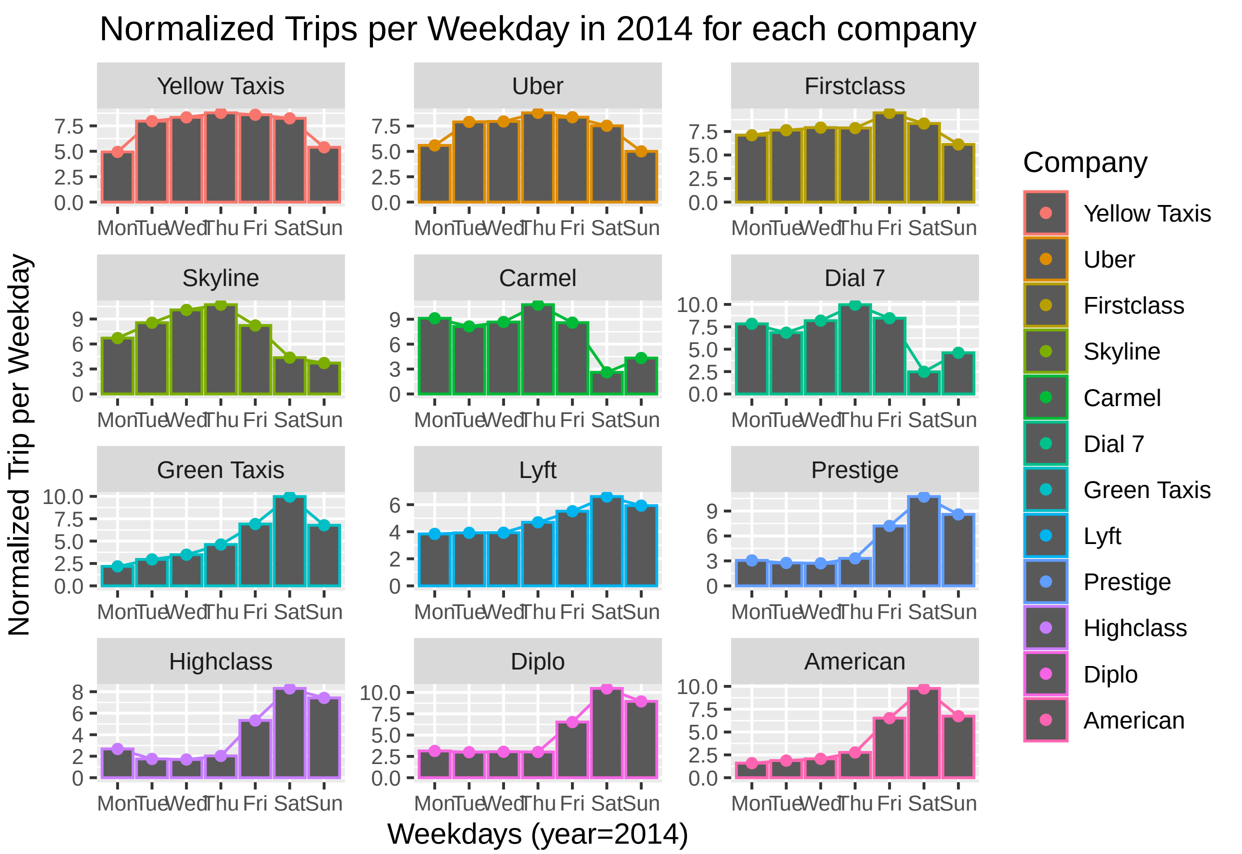
```
## `.groups` argument.
```

Now we have gained the normalized values and further can re-plot the previous diagram.

```
FHV_per_week_trip_normalized$Company <- factor(FHV_per_week_trip_normalized$Company,
                                         levels = c("Yellow Taxis","Uber", "Firstclass","Skyline",
                                                    "Carmel","Dial 7","Green Taxis","Lyft",
                                                    "Prestige", "Highclass", "Diplo","American"))

ggplot(data=FHV_per_week_trip_normalized, aes(x = Weekday , y = sum_trips_weekday_normalized, group=Comp
  geom_line() +
  geom_col()+
  geom_point()+

  scale_x_discrete(labels=c("Monday"="Mon","Tuesday"="Tue",
                            "Wednesday"="Wed","Thursday"="Thu" ,
                            "Friday"="Fri" ,"Saturday"="Sat","Sunday"="Sun"))+
  labs(title="Normalized Trips per Weekday in 2014 for each company",
       x = "Weekdays (year=2014)",
       y = "Normalized Trip per Weekday")+
  theme(plot.title = element_text(hjust = 0.5),
        axis.text = element_text(size=8))+
  facet_wrap(.~Company ,scales='free', ncol=3)
```



INTERESTING INSIGHTS

The normalisation enhances the visibility of defined groups in travel patterns. Depending on this research, I divided the companies into three categories based on their weekday trips pattern. I have formed groups to explain the trends and patterns.

Group 1. Yellow Taxis, Uber, and Firstclass: Quiet Sundays, Busy Saturdays.

Thursday or Friday are the busiest days. Except for Mondays, working days are pretty comparable in demand. Sundays and Mondays are calm days for this group.

Group 2. Skyline, Carmel, and Dial7: Quiet Saturday and Sunday(Weekends)

Working days have comparably similar demand. Thurday is a peak day for this group of companies and has quite weekends.

Group 3. Green Taxis, Lyft, Prestige, Highclass, Diplo, and American: Busy weekends.

This group has a very busy weekends and is quite during the weekdays.

Different pricing models are most likely to blame for these phenomena. For example, after analysis we can say that group 2 may provide weekend discounts to entice its clients. To have a better understanding on impact of pricing model, more information or data on trip pricing might help.

CONCLUSION

The analysis can be improved by gaining more recent infomation of data. Furthermore, the analysis and insights can be improved if we have more data. Event, weather, fuel pricing can be added with the dataset to gain better knowledge on the trips patterns and trends.

please find the code link attached below: http://rpubs.com/Opawar/882633