TRAN5340M

Module Name: TRANSPORT DATA SCIENCE

Assignment Title: Cycle Hire Scheme Analysis

Student Name: SOUNDAR JAMBU

Student ID: 201562661

Word Count: 1774 words

Lecturer: ROBIN LOVELACE

Submission Date: 12/08/2022

Semester: 2nd semester

Academic Year: 2021 - 2022

UNIVERSITY OF LEEDS

Coursework submission for Transport Data Science (TRAN5340M)

Cycle Hire Scheme Analysis

Soundar Jambu

Introduction

This project aims at understanding the usage of bicycles in the city of London in different time scale. The Santander Cycle Hire Scheme is popularly known as Boris Bikes, as it came into operational when he was Mayor of London. The reson behind the name is because of the sponsorship. It was ealrier called as 'Barclays Cycle hire' until april 2015, which then the sponsorship was took over by Santander. In an intresting study, it has been found that people using hire scheme cycles are tend to incur less injury than the normal cyclist. Also, customer research in 2013 showed that, 49 percent of the members of cycle hire scheme told that the sheme has prompted them to involve in more cycling than before. The scheme has been in use since June 30,2010 started with 5000 bicycles and 315 docking stations.

```
# required packages
library(sf)       # spatial vector data classes
library(stats19)  # get stats19 data
library(stplanr)  # transport planning tools
library(spData)   # example spatial data sets
library(tidyverse)# packages for data science
library(dplyr)    # pipelines for manipulating and plotting data
library(sp)       # dealing with spatial data
library(ggplot2)  # plotting graphs
library("corrplot") # plotting correlation
```

Dataset cleaning and manipulation

The data for this analysis has been obtained from two different resources. The day wise bicycle hire count for the city of London has been obtained from the official database of government of uk (data.london.gov.uk). And then, the daily weather data has been obatined from the site called visualcrossing.com . Both the obtained datasets are available in the excel format.

Then the two datasets are merged manually using the dates as a common point and added a categorical variable called 'Season' with four values as 1 denoting winter (December, January, February), 2 denoting spring (March, April, May), 3 denoting summer (June, July, August) and 4 denoting fall (September, October, November). Also, the weather conditions have been converted to numerical for convenience (Clear = 1, Cloudy = 2, Rain = 3, Snow = 4).

```
##          Day Hire_count temp feelslike humidity precip  dew windspeed
## 1 01-01-2017       6534  6.5       3.8     90.5   8.97  5.1      20.3
## 2 02-01-2017      11954  3.5       0.9     82.4   2.00  0.7      15.4
## 3 03-01-2017      19622  2.1      -1.0     85.5   0.00 -0.1      20.2
## 4 04-01-2017      22122  5.3       2.5     74.8   0.00  1.0      17.6
```

```
## 5 05-01-2017      23580  2.6       2.2     78.8   0.00 -0.8       7.4
## 6 06-01-2017      18973  3.3       1.8     90.8   1.00  1.8      16.4
##   conditions Season
## 1          3      1
## 2          3      1
## 3          3      1
## 4          1      1
## 5          1      1
## 6          3      1
```

```
# Checking the number of categorical values in applicable variable (condit
ions and Season)
sapply(bike_count,n_distinct)
```

The column weather condition has four categorical values such as clear sky, Cloudy, Rain and Snow.
The column Seasons has four categorical values such as Winter, Spring, Summer and Fall.

```
# Statistical summary
summary(bike_count)
```

```
##      Day              Hire_count         temp          feelslike
##  Length:1461        Min.   : 4872   Min.   :-3.20   Min.   :-8.00
##  Class :character   1st Qu.:22631   1st Qu.: 8.10   1st Qu.: 6.40
##  Mode  :character   Median :28707   Median :11.90   Median :11.70
##                     Mean   :28660   Mean   :12.37   Mean   :11.59
##                     3rd Qu.:35477   3rd Qu.:16.80   3rd Qu.:16.80
##                     Max.   :70170   Max.   :28.30   Max.   :28.90
##     humidity         precip            dew           windspeed
##  Min.   :39.60   Min.   : 0.000   Min.   :-6.900   Min.   : 0.10
##  1st Qu.:66.00   1st Qu.: 0.000   1st Qu.: 4.200   1st Qu.:15.40
##  Median :75.50   Median : 0.000   Median : 7.500   Median :20.10
##  Mean   :74.18   Mean   : 1.626   Mean   : 7.477   Mean   :21.17
##  3rd Qu.:82.90   3rd Qu.: 1.030   3rd Qu.:11.000   3rd Qu.:25.70
##  Max.   :97.20   Max.   :43.720   Max.   :17.900   Max.   :59.00
##    conditions        Season
##  Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.000
##  Median :1.000   Median :3.000
##  Mean   :1.949   Mean   :2.503
##  3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :4.000   Max.   :4.000
```

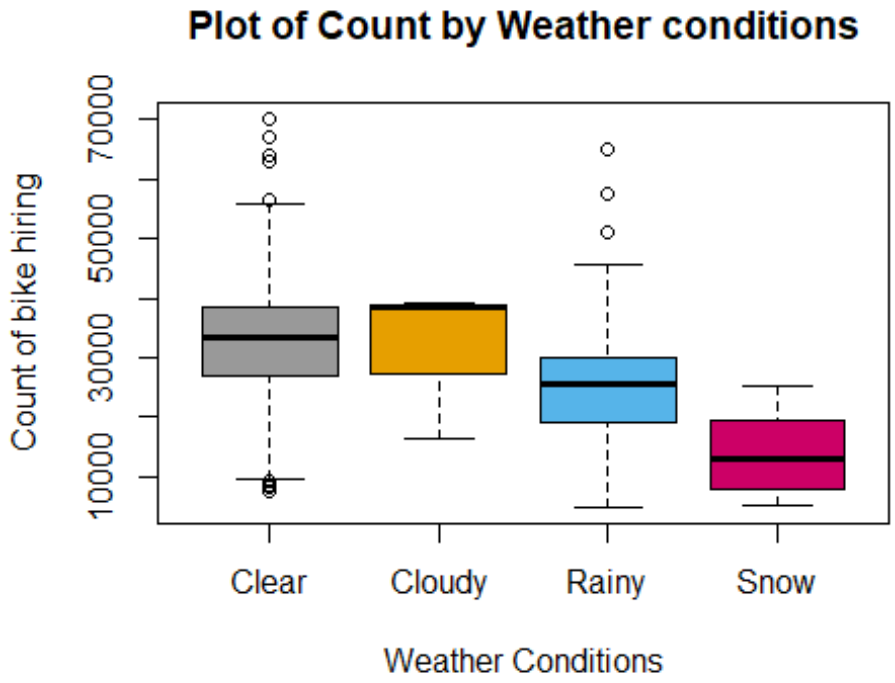
In the period of four years between 2017 and 2020, the lowest temperature recorded is around -3.3°C and the highest recorded temperature is around 28.30°C (which is better than this year as of now). The highest hire count per day tops at 70170 and the lowest per day hire count is at 4872. The highest wind range has been pretty worse recorded at the level of 59 kph speed.

```
# Subsetting dataset into individual years
year2017 <-bike_count[bike_count$Day < "2018-01-01",]
year2018 <-bike_count[bike_count$Day > "2017-12-31" & bike_count$Day < "20
19-01-01",]
year2019 <-bike_count[bike_count$Day > "2018-12-31" & bike_count$Day < "20
```

```
20-01-01",]
year2020 <-bike_count[bike_count$Day > "2019-12-31",]
# Subsetting date column into separate year, month and day
temp_df <- data.frame(date = bike_count$Day,
                      year = as.numeric(format(bike_count$Day, format = "%Y")),
                      month = as.numeric(format(bike_count$Day, format = "%m"))
,
                      day = as.numeric(format(bike_count$Day, format = "%d")))
```

Outlier detection and Analysis

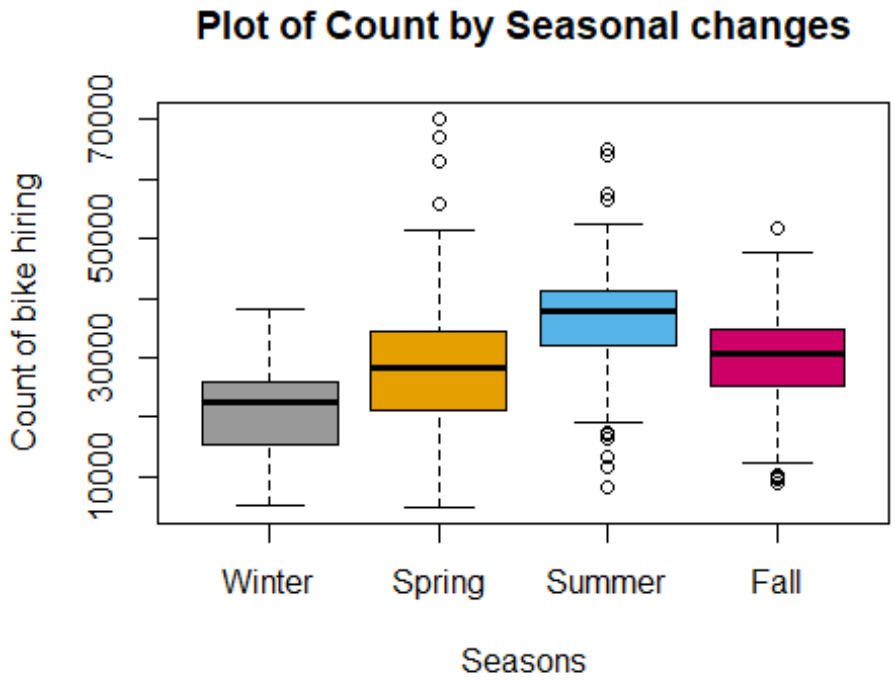
```
# Boxplot of count against daily weather condition
boxplot(Hire_count ~ conditions, data = bike_count,xlab = "Weather Conditi
ons",ylab = "Count of bike hiring",names = c("Clear", "Cloudy", "Rainy", "
Snow"), col = c("#999999", "#E69F00", "#56B4E9", "#CC0066"),main = "Plot o
f Count by Weather conditions")
```



Note: 1= clear sky, 2 = Cloudy, 3 = Rainy, 4 = Snow The above plot clearly shows that the hire count is high during clear sky condition and somewhat cloudy. As the weather gets worse towards rainy and snow the hire count also decreases. During clear sky there are many outliers, its due to the tendency that more people every randomn day likes to go out during clear sky than a cloudy sky as it clouds people's decision whether to go out or not (unsure). In addition, under cloudy days there are less ouliers, its beacause quite a number of cloudy days gets classified under clear sky and rainy days in the process of strict classification into categorical values.

```
# Boxplot of count against Seasonal changes
boxplot(Hire_count ~ Season, data = bike_count,xlab = "Seasons",ylab = "Co
unt of bike hiring",names = c("Winter", "Spring", "Summer", "Fall"), col =
```

```
c("#999999", "#E69F00", "#56B4E9", "#CC0066"),main = "Plot of Count by Sea
sonal changes")
```



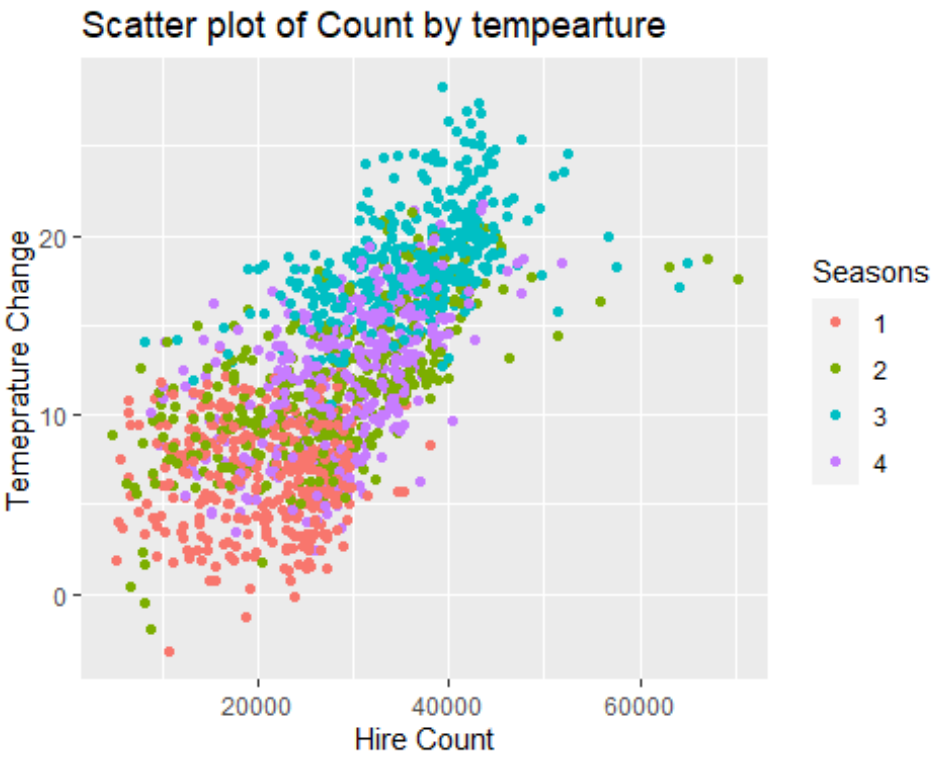
Note: 1= winter, 2 = spring, 3 = summer, 4 = fall. The above plot clearly shows that the hire count is highest during summer, least during winter and moderate during spring and fall. In addition, during winter there isn't much or in fact no outliers are there, which makes sense that no one really wants to go out suddenly in a bike during winter. This doesn't mean there isn't any hire count. It's just there isn't any sudden rise in the hire count during any such day in winter.

```
# Year wise hiring count
yhirecount<- c(sum(year2017$Hire_count),sum(year2018$Hire_count),sum(year2
019$Hire_count),sum(year2020$Hire_count))
year<- c('2017','2018','2019','2020')
year
```

```
## [1] "2017" "2018" "2019" "2020"
```

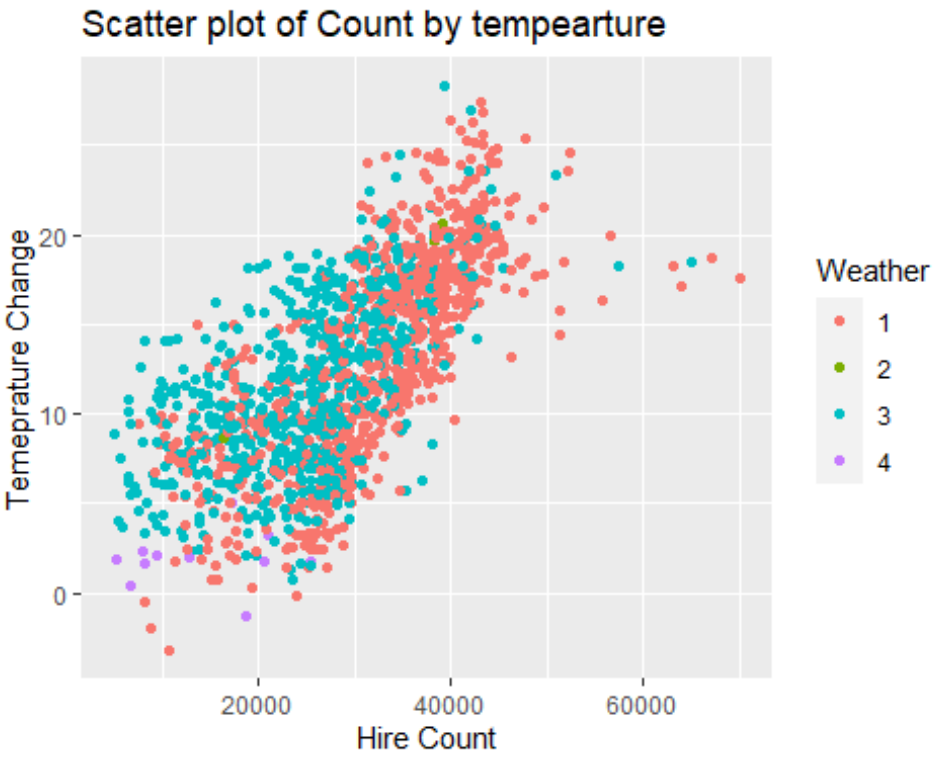
```
# Scatter group plot in accordance to Seasonal Changes

ggplot(bike_count, aes(Hire_count,temp,col=factor(Season))) +ggtitle("Scat
ter plot of Count by tempearture ") +xlab("Hire Count") + ylab("Temepratur
e Change") + labs(color='Seasons') +geom_point()
```



The above plot helps us to visualize the hire count and temperature relationship in accordance to the seasonal changes

```
# Scatter group plot in accordance to Daily weather Changes
ggplot(bike_count, aes(Hire_count,temp,col=factor(conditions))) +ggtitle("
Scatter plot of Count by tempearture ") +xlab("Hire Count") + ylab("Tempepr
ature Change") + labs(color='Weather') +geom_point()
```



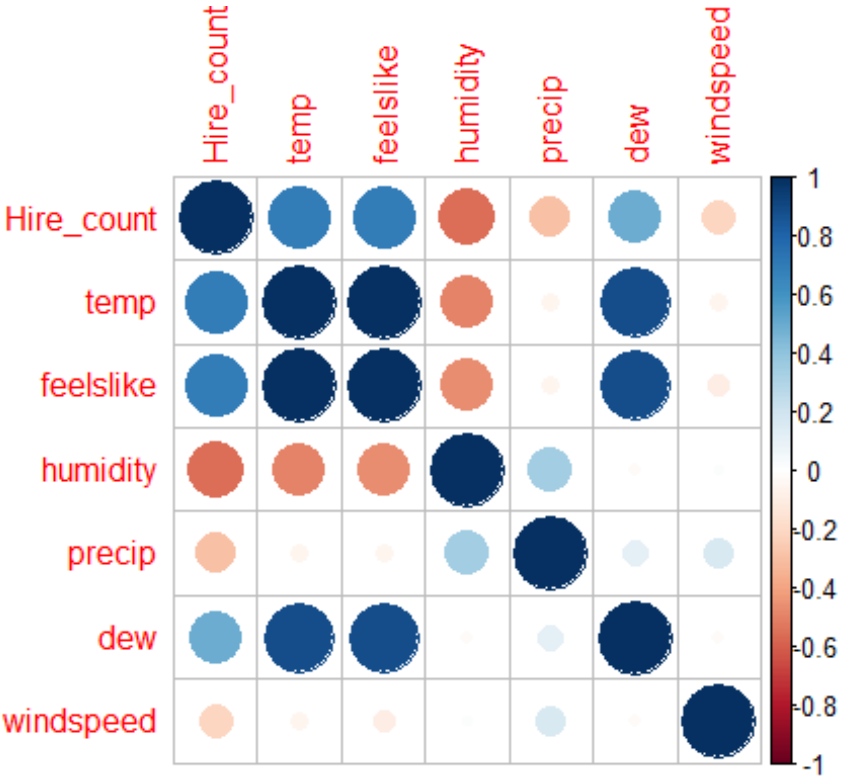
The above plot helps us to visualize the hire count and temperature relationship in accordance to the daily weather changes.

Evaluating Correlation

```
correlation <-cor(select(bike_count, Hire_count, temp, feelslike,humidity,
precip,dew,windspeed))
correlation
```

```
##             Hire_count        temp   feelslike    humidity      precip
## Hire_count   1.0000000  0.69074921  0.69279505 -0.55061052 -0.29350803
## temp         0.6907492  1.00000000  0.99360761 -0.49229510 -0.05279802
## feelslike    0.6927950  0.99360761  1.00000000 -0.46709121 -0.05724136
## humidity    -0.5506105 -0.49229510 -0.46709121  1.00000000  0.34027368
## precip      -0.2935080 -0.05279802 -0.05724136  0.34027368  1.00000000
## dew          0.4943964  0.88141545  0.88763089 -0.02792707  0.11583858
## windspeed   -0.2151913 -0.05194572 -0.09137838  0.01738121  0.16668759
##                    dew   windspeed
## Hire_count  0.49439644 -0.21519127
## temp        0.88141545 -0.05194572
## feelslike   0.88763089 -0.09137838
## humidity   -0.02792707  0.01738121
## precip      0.11583858  0.16668759
## dew         1.00000000 -0.02899988
## windspeed  -0.02899988  1.00000000
```

```
corrplot(correlation,method="circle")
```



Correlation Matrix Plot

As per the correlation matrix plot, we can observe that the hire count is in positive correlation to the temperature. So overall the main influence in people's decision to take go out depends on the heat range of the day. Even the variable dew has pretty high correlation to the hire count. It's interesting that the wind speed is in negative correlation to the hire count. It may be because of the factor that winds do not follow overall average pattern for a certain period of time unlike a season which almost covers a period four months.

Discussion

The introduction of bicycle hire scheme has been helpful to people in many ways. It could be for their daily commute, quick short distance travels to avoid traffic, as quick refreshing physical exercise in the day to day busy life and more over it just gives a chance to try new things and go out, especially after covid. Also, it is an fully green mode of transport. With all these things being said, it still involves a huge amount of money and efforts to pull this more effectively. Maintaining real time availability check, damage / theft of bikes, identifying the people responsible for any misconduct using the services, repairs and maintenance, policy and pricing updates for the scheme to make it reliable and affordable for short and long term users, expanding and encouraging more green routes inside the city and a lot more.

It is really good in this digital age, all these data are available, so that planned and required measures can be taken effectively based on the past data. As we all heard "History repeats itself". So based on the insights gained from the usage records and past data, time period to take measures like maintenance of the docking stations, repairing the damaged bikes, annual inspection, necessity of more or less bikes in certain region can be found.

Conclusion

With more people flocking towards cities for business opportunities and lifestyle changes and their financial needs, the congestion, crowd, traffic and pollution increases more, especially in big cities and developing nations. To tackle that more green routes and encouragement of green mode of transport should be in place. Time series data provides so much useful insights when manipulated and used in a right way to achieve what we need. Further work like based on the future weather forecast, we could predict the hire count in regional and seasonal wise. This will help in saving time to allocate the resources to meet the demands in right time and to provide promotional offers to encourage emission free transport.

References

1. Hiring data has been collected from the UK government official website. (data.london.gov.uk)
2. Weather data has been accumulated from an open source weather forecasting site. (https://www.visualcrossing.com)