**UE19CS203 Statistics for Data Science**

**Final Project**

**Bike Rentals**

**Analysis of bike renting trends during the years 2011-2012 in the U.S**

Submitted by,

1. Pruthvi Karigiri - PES1UG19CS354
2. Raghav V Pandit - PES1UG19CS364
3. Rishab Kashyap - PES1UG19CS385
4. Royston E Tauro – PES1UG19CS396

Section F

PES University, RR Campus

Abstract

This is a project to analyse the bike renting trends during the years of 2011-12 in the US. The objective of this project is to observe the effect of seasonal and weather factors on number of bikes rented. This is specifically targeted to a client who is a management and operations head of a company. Visualisation of the sales and dependencies of the different factors on sales, view of the sales in a day, the growth or loss over the 2 years, who is majorly contributing to the sale the casual customers or the registered ones, Investigation of the change in number of bikes rented on calendar events or holidays, Demand analysis based on the sales of the day with the given factors of weather, holidays, etc. a 4 step process is carried out to address and help the client by presenting solutions. Initially the dataset given is cleaned up and pre-processed by filling up empty data entries by methods like imputation or deletion depending on the conditions in Python majorly using pandas and numpy libraries. This cleaned data is now visually represented by bar graphs, box plots and line graphs depending on the parameters selected and insights are drawn from them. Correlation being a crucial step is carried about next and finally Hypothesis testing is done using 2-sample z test, 1sample z test and ANOVA test. Relationship between the seasonal effects and sales were clearly seen visually and ideal weather conditions were extracted from these graphs. Adding to which the above mentioned hypothesis tests were performed on the data set which further made the sales dependence on seasonal and weather factors more concrete and also the general the view of the overall performance and demand of these bikes were seen clearly.

Keywords: Data set, correlation, EDA, Hypothesis testing, seasonal variations, Demand analysis, sales,

Introduction

Viewing of the performance and the different factors affecting the the business is a crucial part of an operations and management head of a company. The data being collected by the company has to be analysed and insights and solutions need to be drawn from the data collected. Data analysis is important in business to understand problems facing an organisation, and to explore data in meaningful ways. Data in itself is merely facts and figures. Data analysis organises, interprets, structures and presents the data into useful information that provides context for the data.

This project poses solutions and analysis and visual representations of the problems being posed by a client regarding the seasonal effects on the sales of the bike rentals. A 4 Step process of Data pre-Processing, EDA, Correlation and hypothesis testing are performed to go about solving the problems posed by the client.

The data set contains 16 features with over 17000 rows, 4 categorical variables, 5 ordinal variables and 7 numerical variables in the dataset. The features are date , season , hour, year, month, holiday, day of the week, working day, weather, temp, felling temperature, humidity , wind speed , casual riders, registered riders and total riders with 5% i.e. approx. 750  missing values.

The following are few of the issues posed by the client:

1. Seasonal and effect of different weather factors on sales.
2. View of sales throughout the day.
3. The growth/loss of sales over the 2 years and contributions of casual and registered users.
4. Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations.
5. Demand analysis

**Step-1-Data pre-Processing**

-Rishab Kashyap

1) First we install the dependencies that is pandas and numpy

/\*Pandas is an open-source library, that provides high-level data manipulation tool developed by Wes McKinney. It is built on the Numpy package and its key data structure is called the Data Frame. Data Frames allow you to store and manipulate tabular data in rows of observations and columns of variables.

\*/

2) Then we read the files using the command df = pd.read\_csv('./SDS-Project-Bike-rentals.csv')

3)We extract all the necessary information regarding our csv files by using type, shape, info, head and then start analyzing where the changes have to be made and where tidying of data is necessary.

4) The unique values assigned to the weekdays and seasons are extracted using np.unique...The names of the weekdays, seasons and months fit better in those areas, so we replace them accordingly.

Since in place is true, changes are made on the data and nothing is returned...In case of false, it performs operations on data and returns a new copy of data.

5) Now, using the 'df.isnull().sum()' function, we can obtain the number of missing values in the data set.

6) Then, we use 'df[df['column\_name'].isnull()]' to obtain a data frame depicting null values in a particular column for the columns [temp],[atemp] and [hum]

7) Now there are 2 ways to go about the missing values - a) deleting the rows having the missing values itself or b) imputation of the missing values.

a) This would be the simplest method in case of very large datasets which have negligible null values, but it isn't the most efficient method, when datasets are small and precious collected data is lost.

Hence, we are going to be using the second method, i.e. imputation

b) We are going to impute the missing values with the Mean value using: data["column\_name"] = data["column\_name"].replace(np.NaN, data["column\_name"].mean())

But in our case, we can't take the whole column instead we take them common for each season and take the mean...SO it is more organized and comparitively accurate.

So we use, seasons = df.groupby('season')['temp','atemp','hum'].mean()

and "mask = df['season'] == 'summer'

df.loc[mask , 'temp'] = df[df['season'] == 'summer']['temp'].fillna(seasons['temp']['summer']) " code for all seasons and atemp ,temp and hum for each season respectively

8)Finally, Checking whether the missing values/null values have been replaced using df.isnull().sum() and converting this new dataset into another csv file, so as to keep the original file as is using df.to\_csv(r'./Data\_Handling\_Final.csv', index = False, header = True)

**Step-2 - EDA – Visualization**

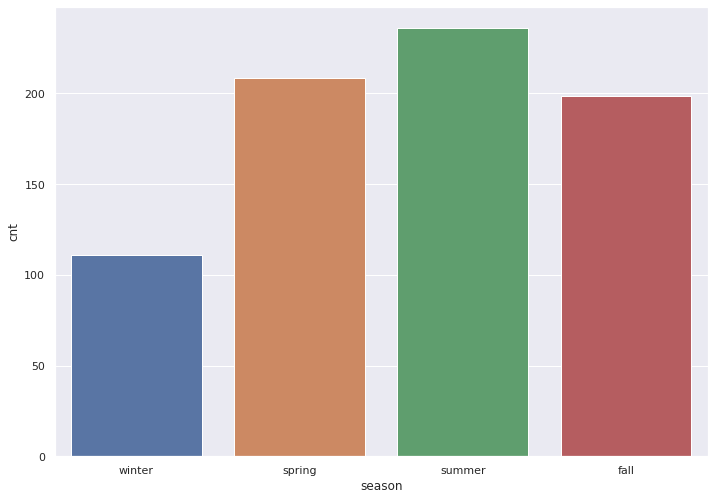
-Raghav V Pandit

3 issues posed by the client are addressed here:

1. Seasonal and effect of different weather factors on sales
2. View of sales throughout the day
3. The growth/loss of sales over the 2 years and contributions of casual and registered users

Shown below is the visual representation of the client’s problem and the insights:

**1. Seasonal effect of different weather factors on sales**



Metrics used:

1. Season
2. Total number of bikes rented
3. Temperature
4. Humidity
5. Wind speed

The bar graph represented here is a season Verses total bikes rented graph.

Insights:

Summer season has maximum demand and sale for bike rentals followed spring, fall and then winter.

The line graphs attached beside illustrate the total number of bikes rented Verses temperature, humidity and wind speed respectively.

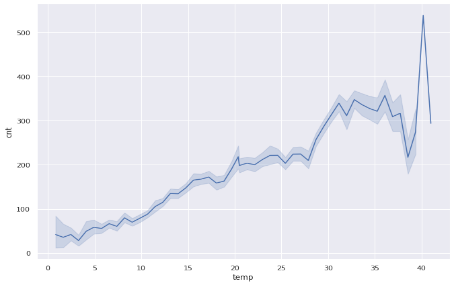
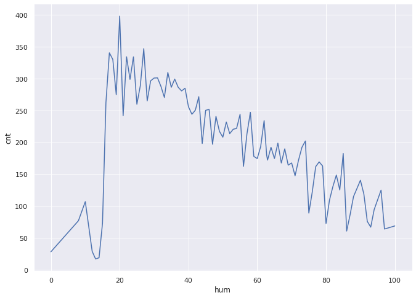
From these graphs optimal weather conditions where the maximum number of bikes rented are seen.

The ideal weather conditions:-

1. Temperature: 25-35 degree Celsius

2. Humidity: 15-25 g per cc

3. Wind speed: 30-50 kmph



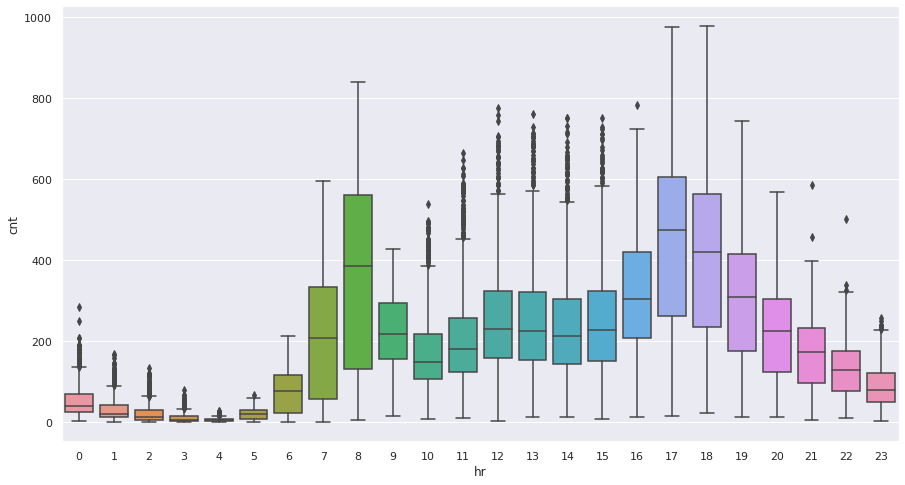
2. View of sales throughout the day

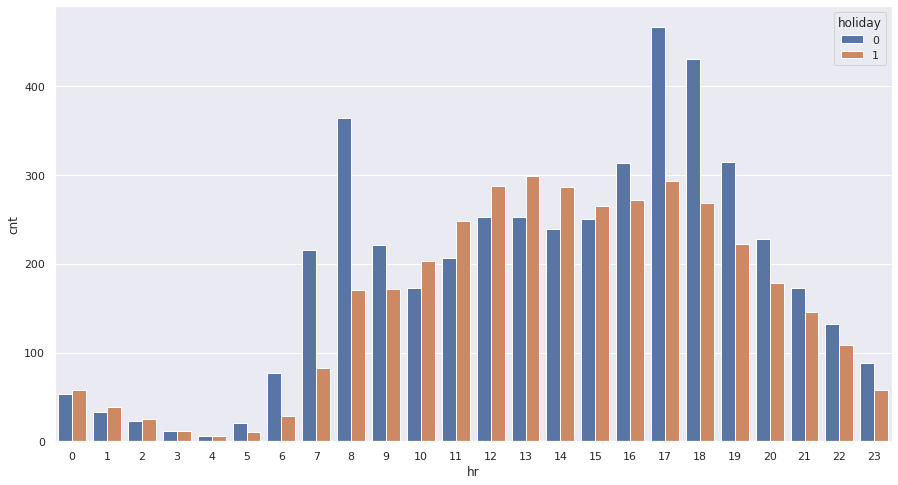
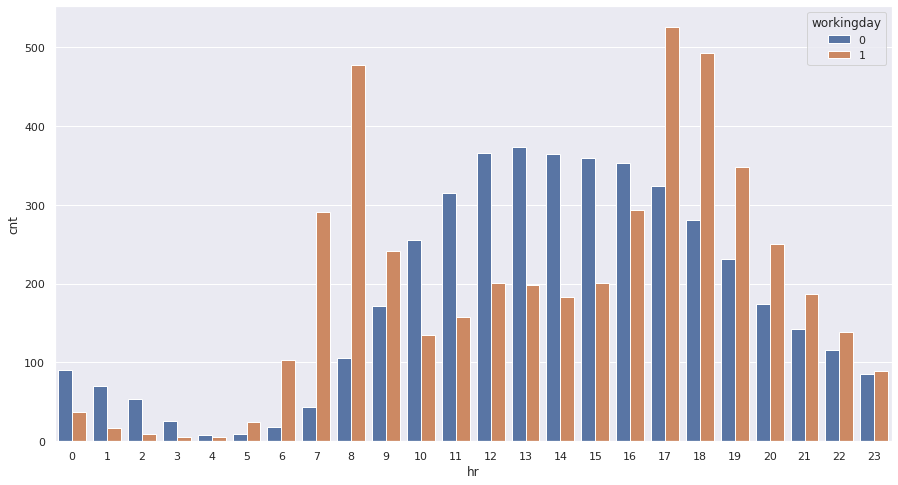
In the boxplot chart attached following insights were drawn.

Insights:

On a regular weekday, the sales are high during the start of the day at around 8 am and in the evening at around 5 pm.

Whereas over the holidays or weekends it’s more uniformly spread during the day, minimal demand at night.





Metrics used:

1. Total number rented
2. Hour of the day
3. Holiday
4. weekend

The bar graph beside represents the total bikes rented verses the hour of the day and the hue being set to whether it’s a holiday or not.

Insights:

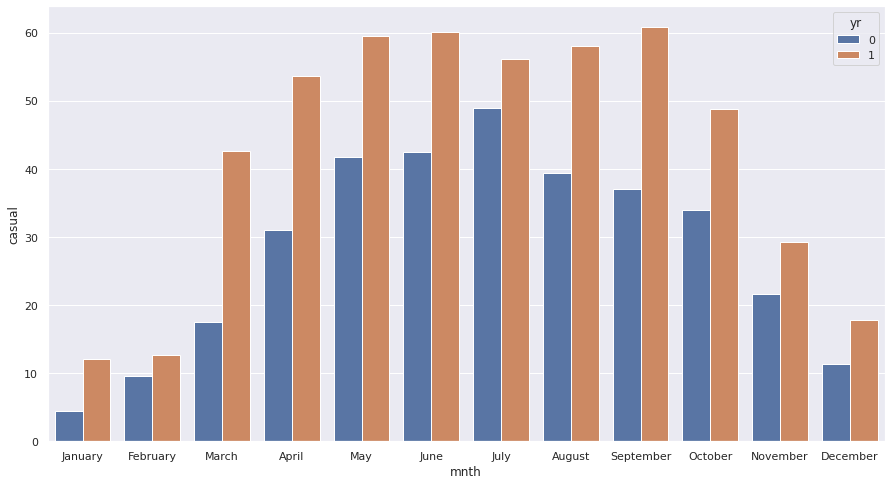
Sales are high mid-year and clearly year 2 has been more profitable in comparison.

The bar graph beside represents the total bikes rented verses the hour of the day and the hue being set to whether it’s a weekend or not.

Insights:

Sales are again high toward the mid and end of year and again there is a significant profit in year 2 than year 1.

3. The growth/loss of sales over the 2 years and contributions of casual and registered users



1. Casual users’ statistics

The attached bar graph shows the plot of the months and total number of bikes rented over the 2 years by casual users.

Insights:

Sales are high mid-year and clearly year 2 has been more profitable

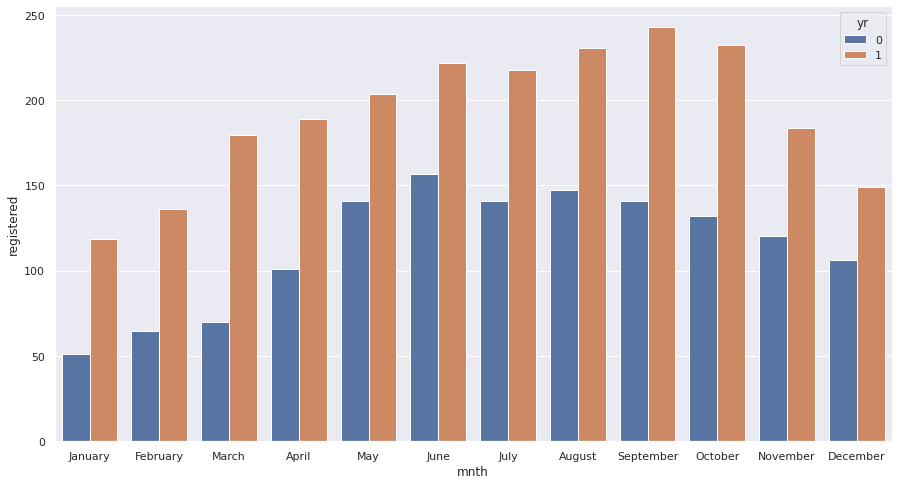
2. Registered users’ statistics

The attached bar graph shows the plot of the months and total number of bikes rented over the 2 years by registered users.

Insights:

Sales are again high toward the mid and end of year and again there is a significant profit in year 2 than year 1

And the majority of the sales has been from the registered users



Metrics:

* Month
* Casual users statistic
* Year

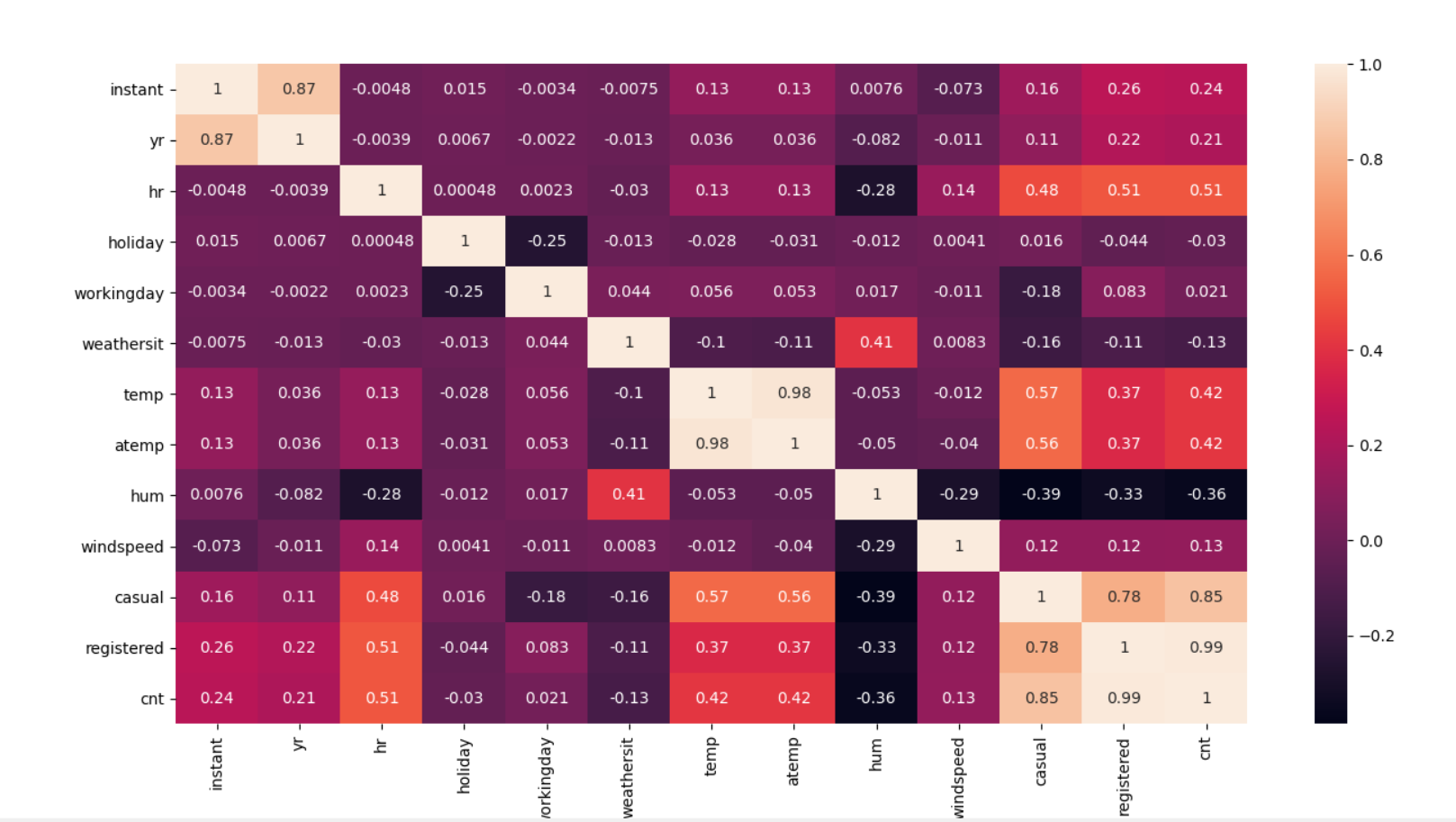
2011 -> 0

2012 -> 1

**Step-3-Correlation**

- Pruthvi Karigiri

Correlation is used to test relationships between quantitative variables or categorical variables. In other words, it’s a measure of how things are related. The study of how variables are correlated is called correlation analysis. A positive correlation indicates the extent to which those variables increase or decrease in parallel, a negative correlation indicates the extent to which one variable increases as the other decreases or vice versa.



There are 3 methods of correlation. In this instance we have used Spearman Correlation as the variables are measured on a scale that is at least ordinal. We can infer that count (cnt) and humidity (hum) have a good negative association of 36% which means that the number of rentals decrease as humidity increases. There is also a very good positive association of 99% between count (cnt) and the number of registered users. This shows that as the number of registered users increases, so do the number of bicycles rented. All these data shows us the conditions when higher number of bicycles are rented.

Step-4-Hypothesis Testing

-Royston E Tauro

The Hypothesis statements for this project are taken from two sources as prior research is required to accept the research/hypothesis or disprove the research/hypothesis. The Hypothesis statements are based mainly on the count of bike rental users and factors affecting it. The significance level for all the tests is 0.05 or 5%. The Hypothesis statements are as follows

1. According to [**https://nacto.org/bike-share-statistics-2017/**](https://nacto.org/bike-share-statistics-2017/) , The demand for bikes is more during peak hours i.e., 7-9 am and 4-6 pm.
   * Null Hypothesis: - The mean count of bike rental users during 7-9 am and 4-6 pm is equal to the mean count of bike rental users during 6am –7 am, 9am-4pm and 6-9pm.
   * Alternate Hypothesis: - The mean count of bike rental users during 7-9 am and 4-6 pm is greater than the mean count of bike rental users during 6am –7 am, 9am-4pm and 6-9pm.
   * Test Performed: - Two sample z test
   * Result: Null Hypothesis Rejected as p value was lesser than 0.05
   * Possible Reason: A lot of people travel to work every day between these hours which makes the count increase.
2. According to <https://nacto.org/bike-share-statistics-2017/> , the demand for bikes is more on working days than non-working days.
   * Null Hypothesis: - The mean count of bike rental users is same on all days
   * Alternate Hypothesis: - The mean count of bike rental users on working days is greater than the mean count of bike rental users on non-working days
   * Test Performed: - Two sample z test
   * Result: Null Hypothesis Rejected as p value was lesser than 0.05
   * Possible reason: People who go to work rent bikes more frequently than those who go on vacations
3. According to <https://nacto.org/bike-share-statistics-2017/> , the demand for bikes doubled in 2017 than 2016. The hypothesis is that the same trend follows for 2011 and 2012
   * Null Hypothesis: - The mean count of bike rental users in 2012 is not double of that in 2011
   * Alternate Hypothesis: - The mean count of bike rental users in 2012 is double or greater of that in 2011
   * Test Performed: - One sample z test
   * Result: Null Hypothesis Accepted as p value was greater than 0.05
   * Possible reason: It's possible that the trend is not double every year but a geometric/arithmetic progression. On further digging, I found that the mean count in 2012 was 1.6 times that of 2011
4. According to the paper “Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations”, the number of bike rentals decrease at temperatures above 30°
   * Null Hypothesis: - The mean count of bike rental users does not depend on temperature between 6am and 10pm
   * Alternate Hypothesis: - The mean count of bike rental users decreases at temperatures above 30 between 6am and 10pm
   * Test Performed: -Two sample z test
   * Result: Null Hypothesis Accepted as p value was greater than 0.05
   * Possible reason: It's possible that because people travel to work and from work at times like 9am and 4-5pm, the temperatures at which might be greater than 30
5. According to the paper “Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations”, the bike rental count during light rain is lesser as compared to cloudy and clear sky
   * Null Hypothesis: - The mean count of bike rental users does not depend on the weather unless its heavy rain
   * Alternate Hypothesis: - The mean count of bike rental users decreases on light rain
   * Test Performed: -One-way ANOVA test
   * Result: Null Hypothesis Rejected as p value was lesser than 0.05
   * Possible reason: It's possible that people assume that the rain might get worse as it begins and wait for it to stop