**Yulu - Hypothesis Testing**

**Problem Statement :**

Yulu is India’s leading micro-mobility service provider, which offers unique vehicles for the daily commute. Yulu has recently suffered considerable dips in its revenues. The company wants to know:

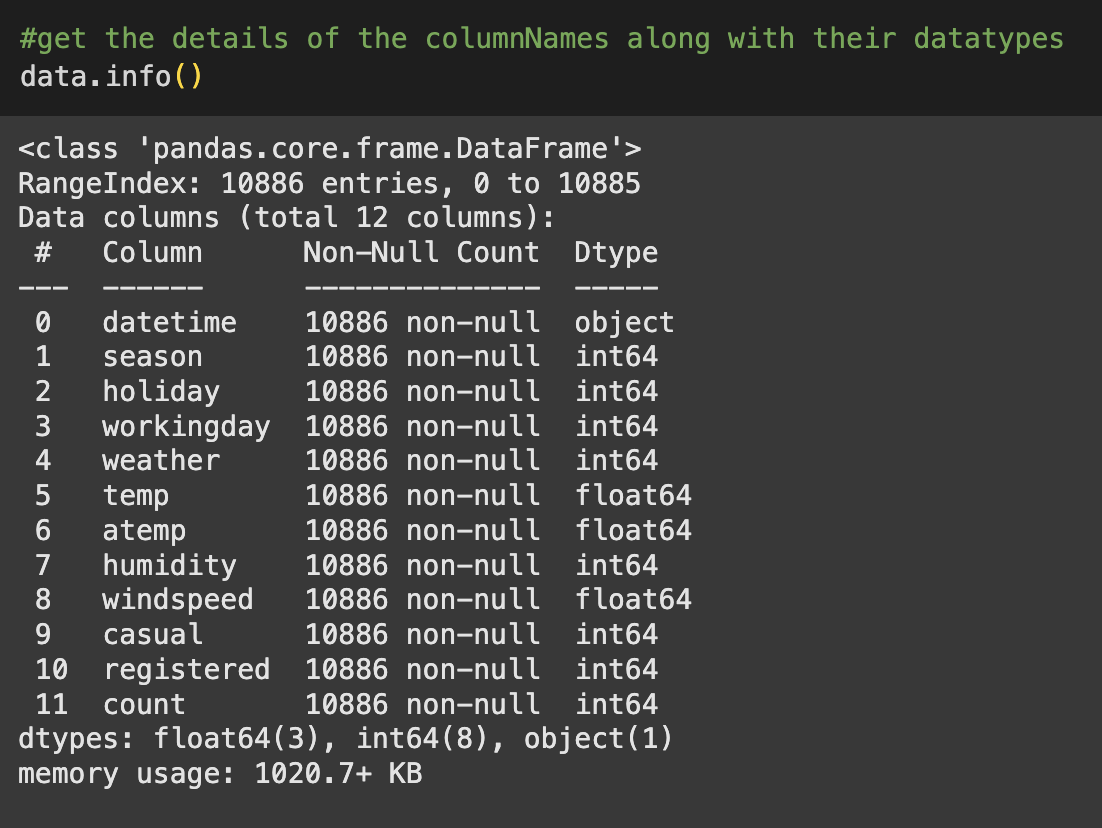
* Which variables are significant in predicting the demand for shared electric cycles in the Indian market?
* How well those variables describe the electric cycle demands

**Exploratory Data Analysis.**

Q) Examine dataset structure, characteristics, and statistical summary.

1. What are the attributes available for the dataset and what are their datatypes ?

Ans )



Insights : Of the 11 attributes, only datetime has the object datatypes. Rest all the columns are numerical among them temp, atemp(feel factor temperature) and windspeed being float values and remaining being integer values.

1. How does the data generally look like ?

Ans)

A screenshot of a graph

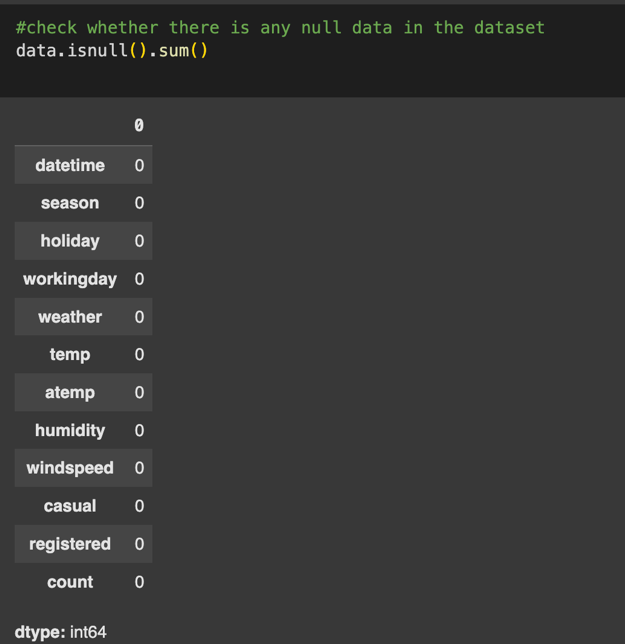
Description automatically generated

Insights : Various columns have numerical values though being categorical variables. This is done for the easiness of the data manipulation and analysis. PFB the details for the same.

* datetime: datetime
* season: season (1: spring, 2: summer, 3: fall, 4: winter)
* holiday: whether day is a holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
* workingday: if day is neither weekend nor holiday is 1, otherwise is 0.
* weather:
  + 1: Clear, Few clouds, partly cloudy, partly cloudy
  + 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  + 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  + 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
* temp: temperature in Celsius
* atemp: feeling temperature in Celsius
* humidity: humidity
* windspeed: wind speed
* casual: count of casual users
* registered: count of registered users
* count: count of total rental bikes including both casual and registered

1. Is there any missing values / null values present in the data set?

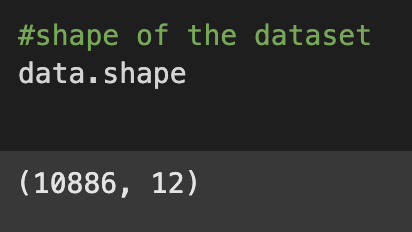
Ans)



Insights : there are no null values present in the dataset. This points that all the data are valid entries and with this data set we should be able to deduce a critical analysis on the requirement of the client.

1. What is the shape for the case study records?

Ans)



Insights : This shows that there are 10886 records or data set which has 12 different attributes or columns associated with them. Shape function help us to know the vastness of the data that we are dealing with.

1. Is there any duplicate records present in the data set ?

Ans)

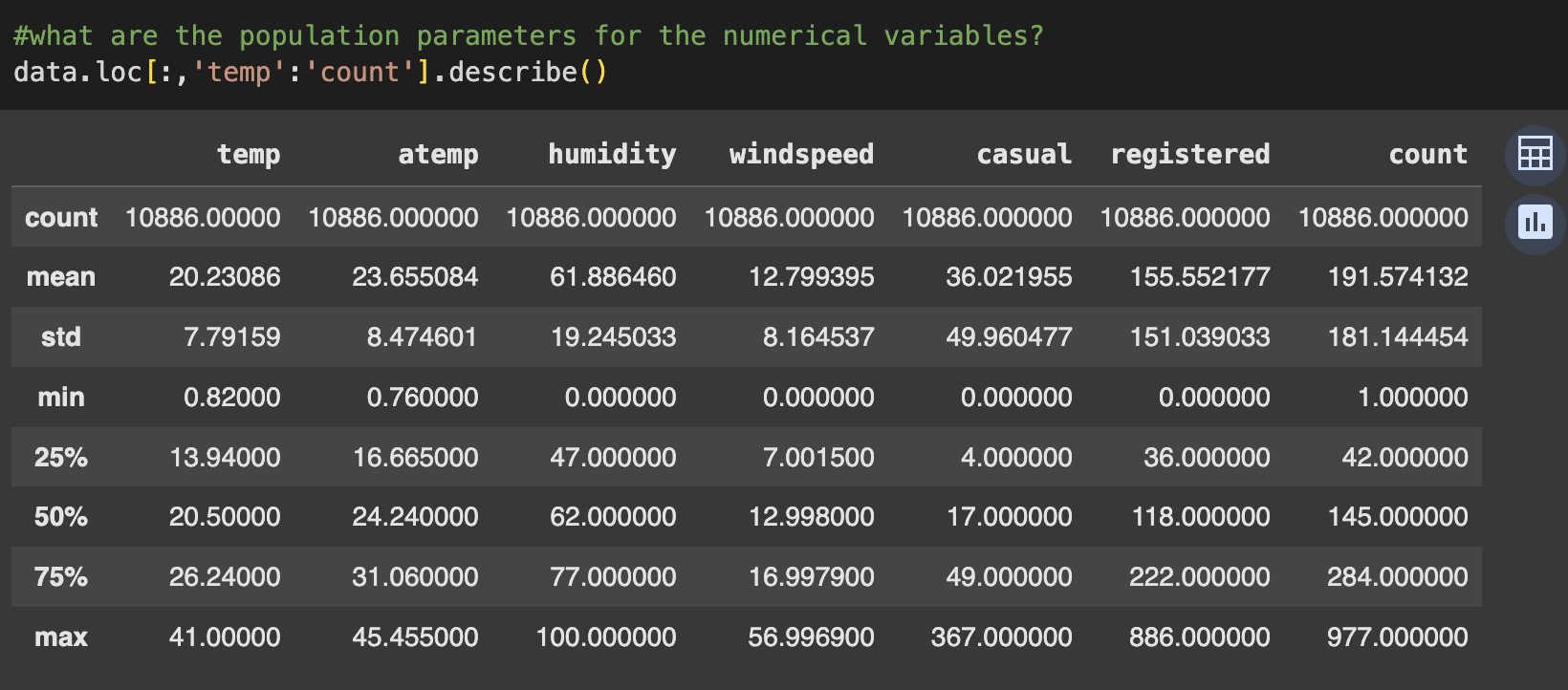
A screen shot of a computer

Description automatically generated

Insights : This says that there are no duplicates present in the given data set. The Flase value attributes to the condition of is duplicate record present. False with a value of 10886 which matches the total number of records as per the shape of the data set points to the fact that there are absolutely no duplicate records present in the data set.

1. Do the dataset has any numerical variables? If yes what are their population parameters ?

Ans)



Insights : The 25%, 50% and 75% refers to the percentile marks for the data set. The mean value refers to the average mean of the population. Std refers to the standard deviation in the population. The min and max records are the values of the minimum and maximum values of the particular column / attribute. The count variable gives the value of 10886 which corresponds to the total number of records.

Q) Analyse the relationship between the dependent and independent variable.

Ans)

* Temperature

A screenshot of a graph

Description automatically generated

Insights : Looking upon the distribution, the temp of 15 has the max count of above 800.The range of the temperature varies from a little above the zero to the max of little over 40 degree.

There are no big outliers for the graph. There distribution seems to moving towards being normal. The distribution of temperature values indicates that the majority of the data points cluster around the central value of approximately 20-25°C. This suggests that the climate represented by this data is relatively temperate, with temperatures not too extreme. The normal distribution shape implies that temperature variations occur symmetrically around the central value.

Recommendations : Most of users are getting the bikes during the temp: range of 20-25 degree. This corresponds to the timings of either morning and during the late evening. Yulu could provide offers or promotional discounts for improving the customer base during this time. They could also ensure that more products are available during these timings so that the customers gets the transportation every time and they don’t have to wait or look for any other alternate.

* Feel Factor Temperature :
* A screenshot of a graph

  Description automatically generated

Insights : The distribution seems to be appearing close to be normal. There are no significant outliers visible in the distribution. The data is spread out over a range of approximately 0°C to 40°C. The peak of the distribution seems to be around the 25-30°C range, suggesting that this is the most common "feel factor temperature. This suggests that people generally perceive the temperature to be in this range, which could be influenced by factors like humidity, wind speed, and solar radiation. The max: value gets peaked when the attribute is above 30

Recommendations : Since the feel factor gives an increased measurement for the temp: we can assume that the people might be responding to the higher values of temp. where it would be comfortable for them to ride. This can lead to a shortage of resources. So stocking of resources during these time would be helpful. Also we should be able to reduce the run around time for increased return of the vehicle than keeping it locked under the same individual. Loyalty points could be given to the customers

* Humidity

A screenshot of a graph

Description automatically generated

Insights :The distribution appears to be approximately normal. It has one peak and symmetric around the central value. The common humidity level is around 50% – 60% humidity though the max count was shown. The data is spread out over a range of approximately 0% to 100%.

There are no significant outliers visible in the distribution. The climate represented by this data has a moderate level of humidity. High humidity can affect the performance of electric vehicles, particularly batteries. This can lead to reduced battery life and range, limiting the usage time of the vehicles. High humidity can accelerate the degradation of infrastructure, such as charging stations and vehicle components.

Recommendations : Yulu could look into the development of vehicles that are resistant to high humidity and can operate efficiently in adverse weather conditions, if they are not operating currently.They could also Implement features like improved waterproof components and efficient cooling systems. Yulu could also have their pricing adjusted based on weather conditions. During periods of high humidity, offer discounts or incentives to encourage usage. Implementation of rigorous maintenance schedules to ensure the optimal performance of vehicles in humid conditions.

* Windspeed

A screenshot of a graph

Description automatically generated

Insights : Customers are using the bikes the most when there is no wind or less windspeed. For the windspeeds of higher values, people don’t preference to use bikes for commute. Though the max count is during no wind, the rest of the average lies in the area of the moderate wind, wherein it would ease the effort given by the customer while biking during the commute. There are a few data points with very high wind speeds, which might be considered outliers. High wind speeds can increase energy consumption for electric vehicles, leading to reduced battery life and range.

Recommendations : Stocking of resources for the days with lower or no windspeed would be helpful as the usage of the bikes during those days increases tremendously.

Q) Relationship between categorical variables (independent variables) and dependent variables (count )

* Holiday

A graph of a couple of blue squares

Description automatically generated

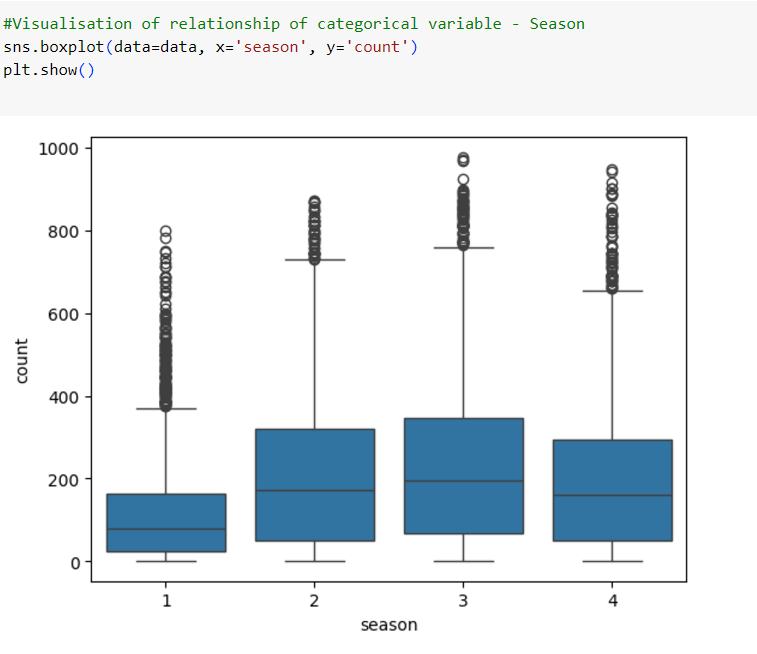
Insights : Holiday 0 and 1 means – false and true for the variable value.

The plot shows that the people uses more of the bikes for commute on a non holiday. Also it is to be noted that there are many outliers for the non holiday when compared to none on a holiday. Both distributions have a similar spread, with the interquartile range (IQR) being roughly the same for both categories.

The lower whishkers for both holiday and non holiday are set at or near to zero, while the upper is set at different values. For a non holiday, the value is near a little above 600 and for holiday it’s a bit more than holiday count. At the same time it should be noted that the average value for both holiday and non holiday are approximately similar values. A little near to 150 value.

Recommendations : On a non holiday the resources should be stocked up so that the higher demand is met accordingly. It should be noted that the increased outliers can also point to the extreme events.

* Season

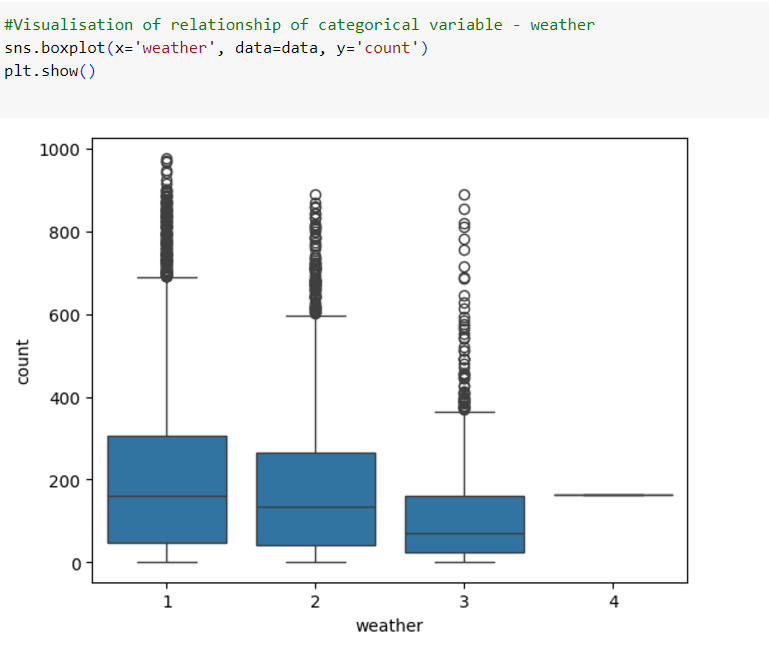


Insights: There are 4 seasons - (1: spring, 2: summer, 3: fall, 4: winter) If we look into the plot, we can understand that except for spring season the rest of the three season has almost similar spread values. All of them are less than 400 count. If we look at the outliers, most of them are for the Spring season followed by Winter, Fall and then Summer. Of the 4 seasons, spring seasons seems to be one with most outliers that signifies the use of the bikes at a large extend by the users. Since the climate is good for biking this could be another reason for the increased outliers for spring season.

The upper whiskers for fall summer and winter are near to 800 value.for Spring its near 400 mark value. Most of the outliers for spring season are between the count of slightly less than 400 and around 600 . For Summer, the outliers starts majorly from littles less than 800 to around 900 value count. Of the 4 seasons, spring seasons seems to be one with most outliers that signifies the use of the bikes at a large extend by the users. Since the climate is good for biking this could be another reason for the increased outliers for spring season.

Recommendations : Since the use of bikes during spring season is more, we could add possible surge value or even we could explore the concept of lower time period for the various slab of usages so that users would be less likely to use it more for per session.

* Weather



* Insights : There are 4 weather pattern as per the plot that is defined as : weather:
  + 1: Clear, Few clouds, partly cloudy, partly cloudy
  + 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  + 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  + 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

If we analyse the plot we can understand that biking is the minimal for the season 4 that is characterised by heavy rain. The spread for the other 3 types of weather are not very slim nor very large. They could be categorised as medium level spread. In case of outliers they are the maximum weather type 1 and 2. It can concluded that the biking becomes a really good experience for the mentioned two types of weather.In case of type 1 of weather the outliers are majorly concentrated around values 700-900 and for type 2 of weather the outliers are majorly concentrated around values 600- little below 800

Recommendations : The mean value for the Weather type 1 ( which is the highest too ) is just below 400 that says that resource to be stocked to ensure that the supply meets the 75 percentile of the demand. The rest of the major crowd being in the outlier section, it cannot be guaranteed for every time. So we may have to rotate the resources based on when it is getting free. Almost we should be able to tell the users the waiting time in case there is no resources available at the moment when it is searched. This would help the user to plan their ride.

* Working Day

A screenshot of a graph

Description automatically generated

Insights : The working day or not is plotted as 0 and 1 which shows false and true respectively. It is evident from the plot that the maximum usage is during the working day. but it should also be noted that the spread for the both the days are almost similar. While the outliers makes the difference for both the days. The higher number of outlier for working days shows the extreme nature of usage. The mean value for the non working day is lower when compared to the spread. But at the same time, it is a almost half value for the working days when compared to the spread. There is a steady usage for the bikes during the working days looking at the 50 percentile mark for the spread.

The outliers for working day = 1 goes all the value from little above 600 to around 900 while that for non working day goes only just a little above and below of 700

Recommendations : During the working days we should be able to allocate as much as possible in the count of around 200-400 so that the experience remains comfortable for the users.

Q) Check if there is any significant difference between the number of bike rides on weekdays and weekends?

Ans)

* setting up null and alternate hypothesis.

Ho – There is no significant difference between the number of bikes on weekdays and weekends.

H1 - There is significant difference between the number of bikes on weekdays and weekends.

* Choosing the distribution – as per histplot, this is a right skewed distribution.

A graph of a number of numbers

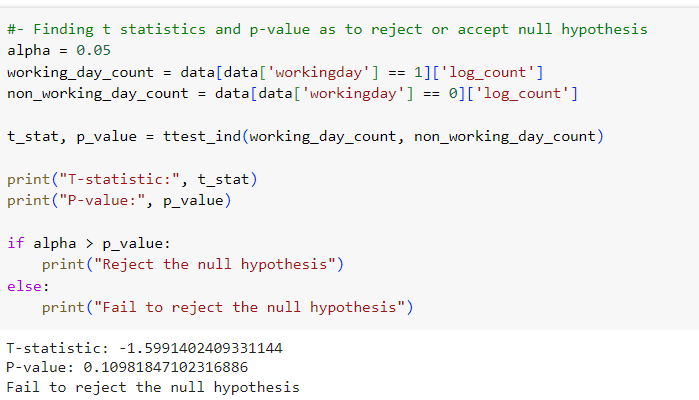
Description automatically generated with medium confidence

This needs to be converted to log normal distribution.

A graph of a logistic

Description automatically generated

as we are comparing the values of two different samples from same group and these two are independent of each other. To find the p-value and t-statistics we can use the approach of two sample independent test.



Insight : The alpha value is 0.05. as per the independent sample test, the p-value calculated is 0.1098. when this value is compared to the alpha value, we can deduce the result that – we fail to reject the null hypothesis. This means that null hypothesis is true. i.e. there is no significant difference between the number of bike rides on weekdays and weekends. If we look at the boxplot distribution for working day, we can understand that the spread for the variable is not so different for both weekend and weekday. This means that our findings here supports the boxplot projection.

Recommendations : since there is no significant difference between weekdays and weekends, we can provide promotional offers or increased incentives like loyalty points during the weekday usage. The weekend usages can be given a different pricing.

Q)Check if the demand of bicycles on rent is the same for different weather conditions?

Ans )

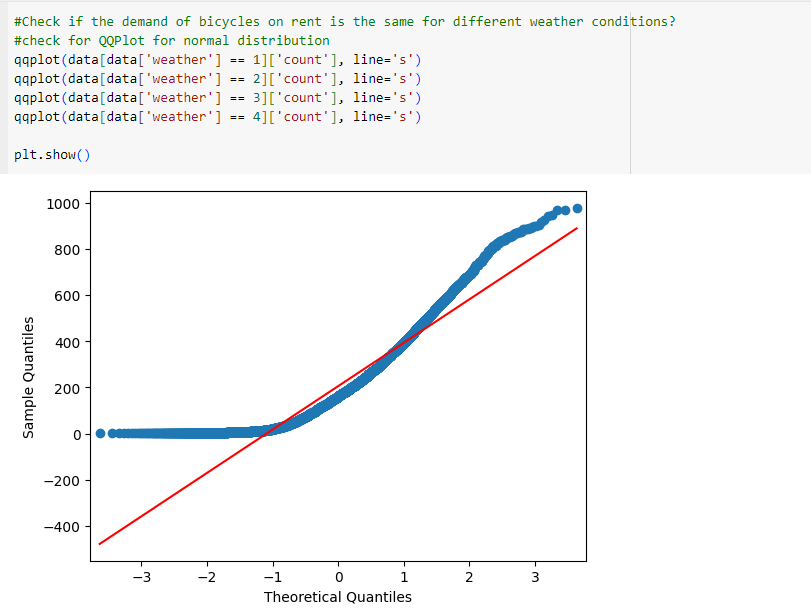
* setting up null and alternate hypothesis.

Ho – There is no difference in demand of bikes for different weather conditions.

H1 - There is significant difference in demand of bikes for different weather conditions.

We are selecting ANVOA test to confirm our hypothesis because this is a question that involves one categorical value and one numerical value.But we need to check whether the assumptions for ANOVA stands valid or not. For this we will have to check for normal distribution or not via qqplot and check for equal variance among groups using levenes test.

* as per qqplot this is not a normal distribution

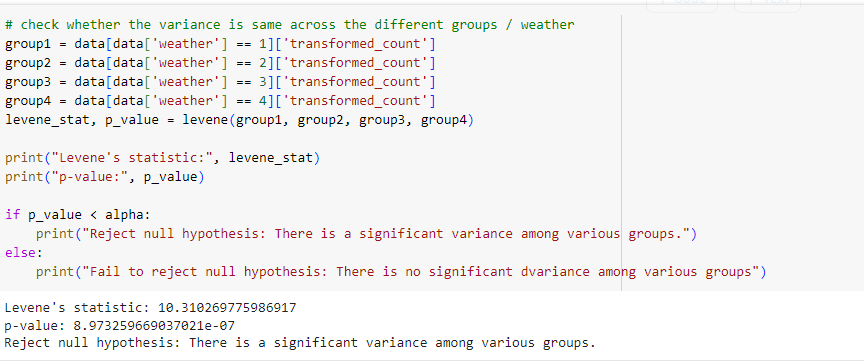


Hence we are doing box cox transformation to make it normal and then do the ANOVA test

and with help of f\_oneway function we do the ANOVA test. We are selecting ANVOA test because this is a question that involves one categorical value and one numerical value.

Since there are 4 different weather conditions, we are grouping them into 4.

We check for the variance of the different groups of weather conditions.



This means that the variance across the groups are different. Since the Distribution is not normal and the variance are not equal among the groups, we cannot use ANOVA test. Instead we will have to use Kruskal – Walli’s test to compare the medians of the independent groups.

A screen shot of a computer

Description automatically generated

On application of Kruskal walli’s test we get the value of p\_value as much less than alpha value. Hence we reject the null hypothesis. i.e. there is significant difference in bike rentals made across different weather conditions.

This hypothesis is also vetted by the box plot of the weather and count where we can see the outliers reach as high as 1000 and the lowest too.

Recommendations : Based upon the different weather conditions we can dynamically change the pricing. More the demand we can add the component of surge pricing but note that it should limited to a ceiling value as the surge pricing beyond threshold will make the users not interested in using our services. We can also rope in promotional offers and discounts for regular users and frequent users.

Q) Check if the demand of bicycles on rent is the same for different seasons ?

Ans)

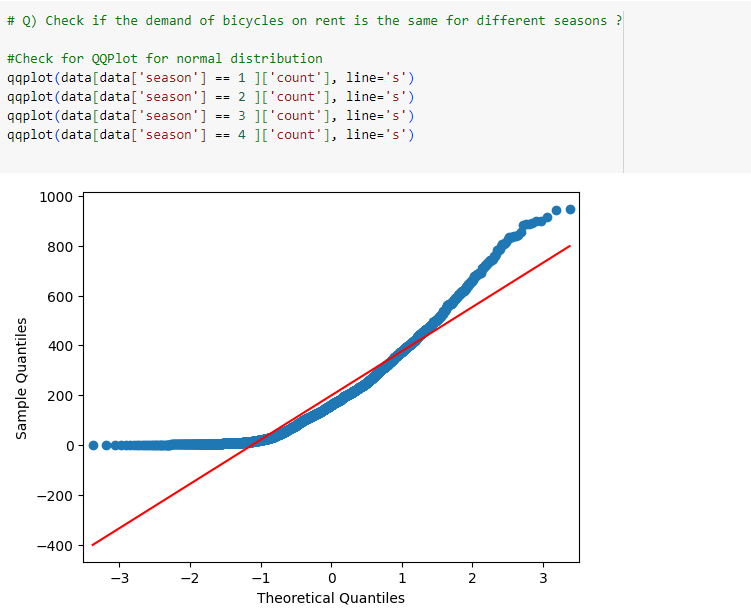
* setting up null and alternate hypothesis.

Ho – There is no difference in demand of bikes based on different seasons.

H1 - There is significant difference in demand of bikes based on different seasons.

We are selecting ANVOA test to confirm our hypothesis because this is a question that involves one categorical value and one numerical value.But we need to check whether the assumptions for ANOVA stands valid or not. For this we will have to check for normal distribution or not via qqplot and check for equal variance among groups using levenes test.

* as per qqplot this is not a normal distribution



We need to also check for variance among different seasons. Using boxcox transformation we change the count values to normal distribution.

#convert the distribution to normal using boxcox transformation

transformed\_data, lambda\_value = boxcox(data['count'])

data['transformed\_count'] = transformed\_data

A screenshot of a computer program

Description automatically generated

Since we are getting value of p\_value less than alpha we conclude that the variance among groups are different. Hence with this it is confirmed that we wont be able to use ANOVA.

Hence we will have to use Kruskal walli’s test to confirm our hypothesis.

A screenshot of a computer

Description automatically generated

It is evident that p\_value is much less than the alpha value. Hence we can reject the null hypothesis. i.e. there is difference in number of bikes that are rented during different seasons in the year.

This observation is in line with the box plot representation of seasons and count that we had tested earlier.

Recommendations : The pricing could be changed as per seasons. Different seasons have different usages, pricing can be also be altered. Looking at the box plot we can understand the spring seasons has max count for usage. Hence lucrative prices can be used here. Since the usage is more, we can also focus on quantum rather than on individual pricing. This would allow us to give a more competitive pricing to the customers. This would be a win-win situation for both customers and company.

Q) Check if weather conditions are significantly different during different seasons ?

Ans)

* setting up null and alternate hypothesis.

Ho – There is no difference in weather conditions based on different seasons.

H1 - There is significant difference in weather conditions based on different seasons.

We are selecting Chi square test to confirm our hypothesis because this is a question that involves two categorical values. We can use test for Independence to check the same.

First we need to create the matrix that maps the values of weather and seasons. For that we can use the crosstabs from pandas library. With this as input we can use contingency function of chi to deduce the p\_value so that we can confirm on the rejection or acceptance of the null hypothesis.

A screenshot of a computer

Description automatically generated

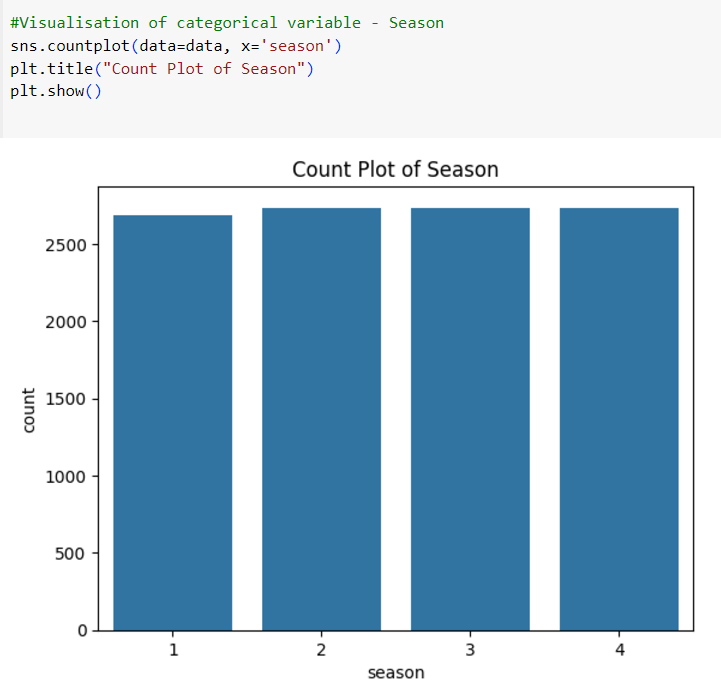
A screenshot of a computer

Description automatically generated

From the above code, we can understand that the p\_value is very much less than the alpha value. This says that we have to reject null hypothesis. i.e. there is significant impact on weather conditions and the seasons.

Q) What is the usage of bike rentals across the seasons ?

Ans )

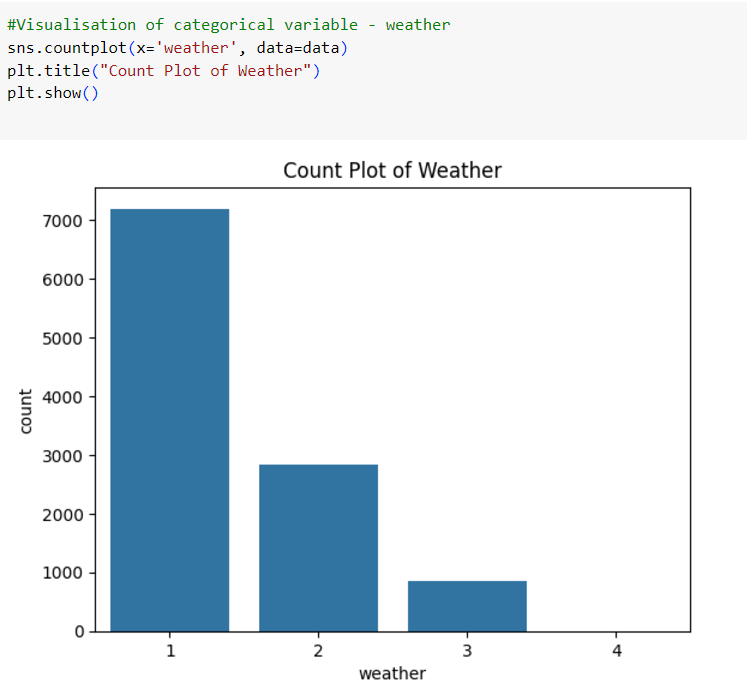


Insights : this says that the seasons other than spring have almost identical count for the bike rentals made. But even though spring is behind in the count, they are almost similar to the other three seasons values.

Recommendations : We can offer a stable pricing since through out the seasons the count remains almost the same. So daily users or regular users could be incentivized to keep them in lieu. We could also introduce new offers as to attract new customer s.

Q) What are the bike rentals count based on weather types ?

Ans )



Insights : From this, it is clear that the weather type 1 is preferred by most of the customers and weather type 4 is not preferred by the customers. The steep height of the weather type 1 implies that a good weather would encourage user to take bikes for their daily commute

Recommendations : When the weather is good, we could offer better pricing for the users as to incentivize them more to use bike for their commute. We can also introduce the loyalty points for increased frequency of usage of the rentals.