Yulu Case Study

YULU is a business that rents bikes for a variety of uses throughout the city, but recently the business has noticed a decline in renting bikes, so they have given us information to analyse it and provide suitable explanations for this decline in rental bikes as well as some suggestions for improvement.

We have included a variety of parameters in this case study, including the season, holiday, working day, date, temperature, weather, and the total number of rental bikes for each hour over the years.

Some **Problem Statements** we can draw from the above description:

1. Which months are facing the fall in Rental bikes?
2. What drives the trend?
3. If the trend is negative, what can we do to fix it?

Some of the **approaches** we can initially starts with:

1. Drawing **Month** wise trend for total Count of Rental bikes.

- (*To understand the over-all trend of the company*)

2. Do **Weather** affect the total Count of Rental bikes?

- (*Weather could be one factor affecting in Renting the bikes*)

3. Workingdays has some impact on Renting bike?

- (*Do any particular month has more holidays or working days which is affecting Renting bikes*)

4. No. of cycles rented similar or different in different seasons?

- (*Season could be one factor*)

5. If **Weather** & **Season** both are affecting then do they both are relatable to each other?

Note-**Inferences drawn from the Case Study will be presented at the end of the Study in Conclusion section.**

Several of these columns have int and object datatypes, but DataTime columns must be in DataTime format. Therefore we'll convert this column to DataTime format and delete any null entries from the data before proceeding with Analysis.

```
import numpy as np
import pandas as pd
from scipy.stats import ttest_1samp,ttest_ind,shapiro,levene,kruskal,chi2_contingency
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm


df=pd.read_csv("yulu.csv")

""" Converting Datatime column and creating Month and year columns so that we can build our intitutions on the over-al

df["datetime"]=pd.to_datetime(df["datetime"])
df["Month"]=df["datetime"].dt.month
df["year"]=df["datetime"].dt.year
```

Approach 1

With the aid of plots, we will first make conclusions about the data at the monthly level. For example, will these months follow the same pattern each year? which aid us in heading in the right direction.

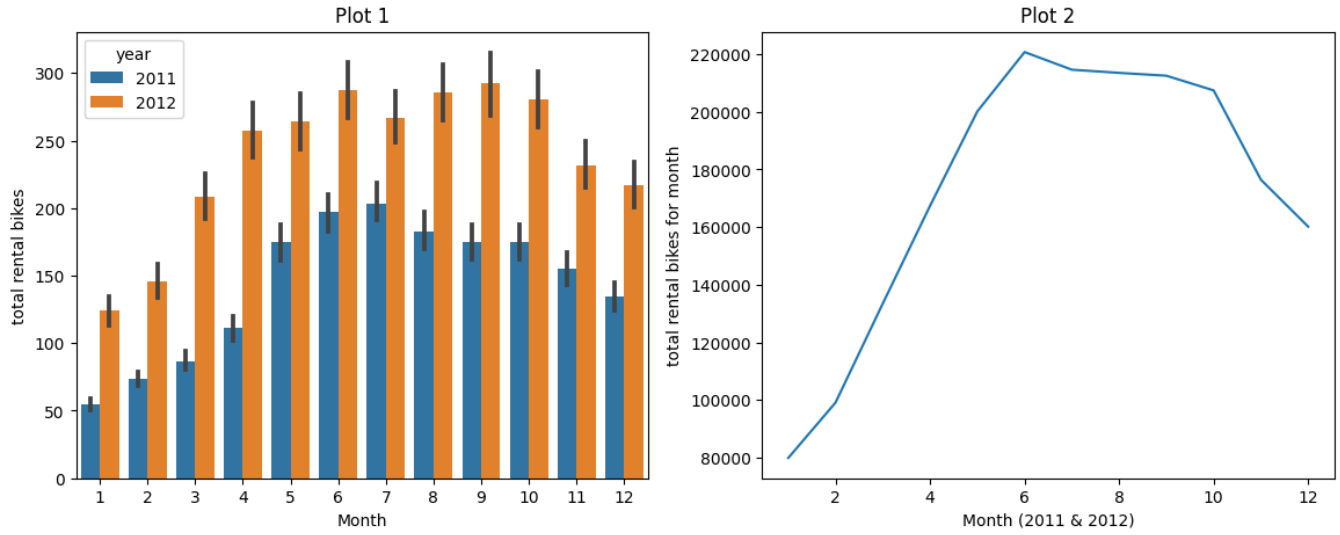
- First, we'll plot the total number of bikes rented over time using a month-by-month barplot.
- Next, we'll plot the combined data to help us understand the situation.

```
plt.figure(figsize=(14,5))

plt.subplot(1,2,1)
sns.barplot(df,x="Month",y="count",hue="year")
plt.title("Plot 1")
plt.xlabel("Month")
plt.ylabel("total rental bikes for month")

plt.subplot(1,2,2)
plt.plot(df.groupby("Month")["count"].sum())
plt.title("Plot 2")
plt.xlabel("Month (2011 & 2012)")
plt.ylabel("total rental bikes for month")

plt.show()
```



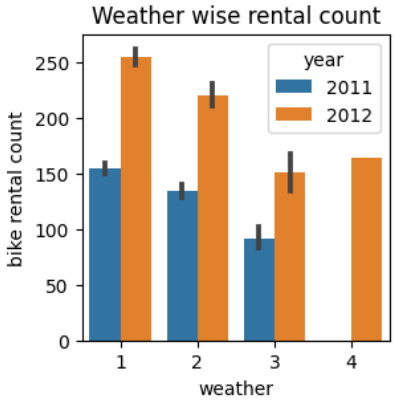
From the above plots we came to know that there's a peak from 4th month to 10th month in rental count in both the years. And there must be chances of some relation between months and rental count but there are other factors as well to analysis on. Factors helping these months to keep the trend same in both the years could be:-

- **Weather of months**
- **Temperature throuthout the month**
- **Number of holidays in a month**
- **Number of working days**

Approach 2

Let's check the weather-wise rental count ratio.

```
plt.figure(figsize=(3,3))
sns.barplot(df,x="weather",y='count',hue="year")
plt.title("Weather wise rental count")
plt.ylabel("bike rental count")
plt.show()
```



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

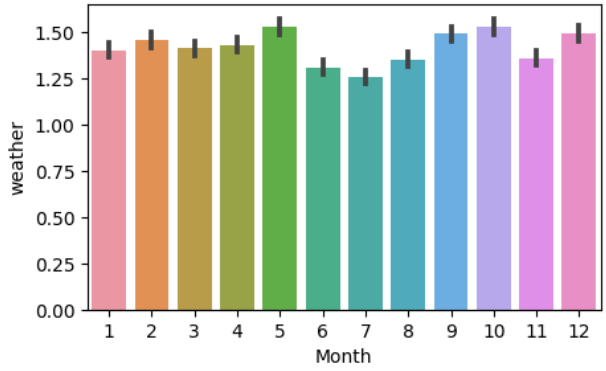
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

From the above observation we can plot that WEATHER Type 1 & 2 has more rental counts then of other weathers. I.e. Clear sky or cloudy sky has some relation with the hike in rental count.

But we need to check **month wise weather** of INDIA because INDIAN weather doesn't fluctuate frequently, it remains almost same throuth-out the year.

Plot for Month Wise Average Weather for combined both year

```
plt.figure(figsize=(5,3))
sns.barplot(x=df["Month"],y=df["weather"])
plt.show()
```



From the above plot we can estimate that through-out the year on an average weather remains between Type **1 to 2** i.e. **clear sky or partly cloudy condition**

So the clear sky or partly cloudy condition have more siginificant positive affect in total rental counts. But Weather almost remain between type 1 & 2

To give strength to our analysis we can perform T-Test or ANOVA over Weather and count column.

To perform ANOVA we need to check whether data follows Normal distribution or not that is **Shapiro test**. And the Variance between different Weather type data are equal or not i.e. **Levene test**. And if either of the test fails we cannot perform ANOVA, instead we can go for KRUSKAL H Test.

For Normal distribution Shapiro test

We take 5000 random sample values of **Rental count** bold text from the data

```
ran_sam=np.random.choice(df["count"],size=5000)
```

This is a hypotheses test and the two hypotheses are as follows:

- Ho(Accepted): Sample is from the normal distributions.(Po>0.05)
- Ha(Rejected): Sample is not from the normal distributions.
- Alfa = 0.05

```
shapiro(ran_sam)
```

```
ShapiroResult(statistic=0.8928359746932983, pvalue=0.0)
```

Pvalue is 0 i.e. < 0.05 (Alfa), we'll reject the null Hypotheses. Hence It is not normal distribution.

For **Variance equal** Levene test

Hypotheses test and the two hypotheses are as follows:

- Ho(Accepted): Samples have equal variance.(Po>0.05)
- Ha(Rejected): Samples have not equal variance.
- Alfa = 0.05

```
weather_1=df[df["weather"]==1]["count"]
weather_2=df[df["weather"]==2]["count"]
weather_3=df[df["weather"]==3]["count"]
weather_4=df[df["weather"]==4]["count"]
```

```
levene(weather_1,weather_2,weather_3,weather_4)
```

```
LeveneResult(statistic=54.85106195954556, pvalue=3.504937946833238e-35)
```

Pvalue is almost 0 i.e. < 0.05 (Alfa), we'll reject the null Hypotheses. Hence Samples have no equal variance.

Thus, ANOVA test can't be applicaple. So we go with Kruskal TEST

Kruskal test

Hypothesis for Kruskal test are:

- Ho(Accepted):Weather doesn't affect count of rental bike.(Po>0.05)
- Ha(Rejected):Weather do affect the count of rental bike.
- Alfa = 0.05

```
kruskal(weather_1,weather_2,weather_3,weather_4)
```

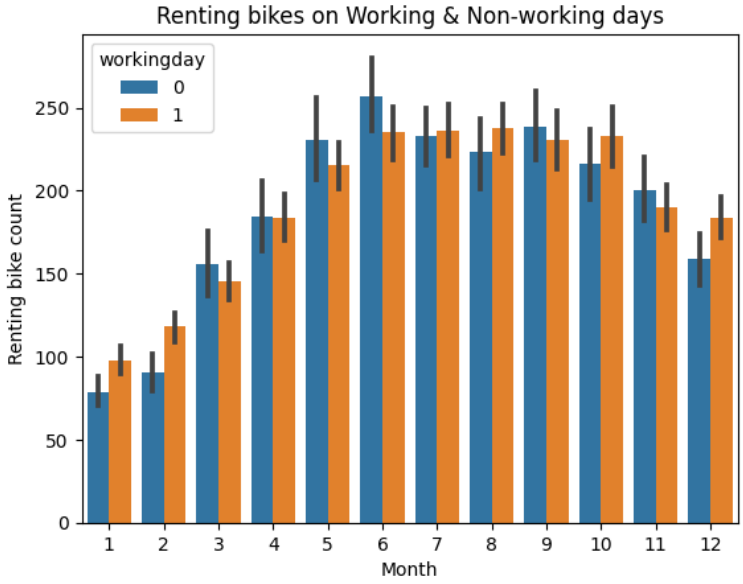
```
KruskalResult(statistic=205.00216514479087, pvalue=3.501611300708679e-44)
```

Pvalue is < 0.05 (Alfa value), We'll reject our Null hypothesis. Hence, **Weather do affect the count of rental bikes**

Approach 3

Let's check weather workingdays has some impact on Renting bike or not.

```
sns.barplot(df,x="Month",y="count",hue="workingday",)
plt.title("Renting bikes on Working & Non-working days")
plt.ylabel("Renting bike count")
plt.show()
```



Workingday:– if day is neither weekend nor holiday is 1 (orange), otherwise is 0 (blue).

From the above plot we cannot estimate that workingdays do affect the Rental bike count, as both working and non-working are close to each other.

We can apply **Two sample T-Test** here.

Hypothesis for Two sample T-test are:

- Ho(Accepted):Working days doesn't affect count of rental bike.(Po>0.05)
- Ha(Rejected):Working days do affect the count of rental bike.
- Alfa = 0.05

```
working_1 = df[df["workingday"]==1]["count"]
non_work_2 = df[df["workingday"]==0]["count"]
```

```
work=[]
non_work=[]
for i in range (1000):
     tt=np.random.choice(working_1,size=1000).mean()
     yy=np.random.choice(non_work_2,size=1000).mean()
     work.append(tt)
     non_work.append(yy)
```

```
work=pd.Series(work)
non_work=pd.Series(non_work)
```

```
ttest_ind(work,non_work)
```

```
Ttest_indResult(statistic=18.624099958908957, pvalue=1.6366382549345115e-71)
```

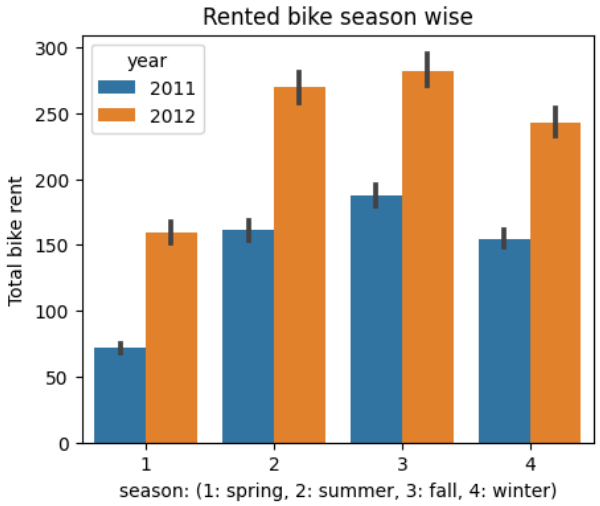
- Hence, the pvalue is significantly low i.e almost 0, we will reject out Null hypothesis. Pvalue shows there is difference between mean of working and non-working days
- That is, Working days might affects the Rental bikes

Approach 4

No. of cycles rented similar or different in different seasons

Lets check visually whether **Season** affects the bike rentak count.

```
plt.figure(figsize=(5,4))
plt.title("Rented bike season wise")
sns.barplot(x=df["season"],y=df["count"],hue=df["year"])
plt.ylabel("Total bike rent")
plt.xlabel("season: (1: spring, 2: summer, 3: fall, 4: winter)")
plt.show()
```



As we can observe from the graph that Season **1,2 & 3** have significantly hive Rented bikes as compare to Season **1**. And To give strong support to our observation we can use **ANOVA** for **4 different category** i.e. Rental count for all four Season.

Lets create 4 lists of mean of mean for Rental bike of all four seasons.

```
S1=df[df["season"]==1]["count"]
S2=df[df["season"]==2]["count"]
S3=df[df["season"]==3]["count"]
S4=df[df["season"]==4]["count"]
```

```
s1=[]
s2=[]
s3=[]
s4=[]
for i in range (1000):
   a=np.random.choice(S1,size=1000).mean()
   s1.append(a)
   b=np.random.choice(S2,size=1000).mean()
   s2.append(b)
```

```
c=np.random.choice(S3,size=1000).mean()
s3.append(c)
d=np.random.choice(S4,size=1000).mean()
s4.append(d)
```

Now we have 4 (s1,s2,s3,s4) lists of mean of mean for all four season.

First we test for **Normal Distribution**(Shapiro test) then for **Equal variance**(Levene Test) and if any of these fails then we go with **Kruskal Test**

Hypothesis for Normal Distribution are:

- **Null Hypothesis**: The distribution is normal
- **Alternate Hypothesis**: Data doesn't follow normal distribution
- Alfa: 0.05

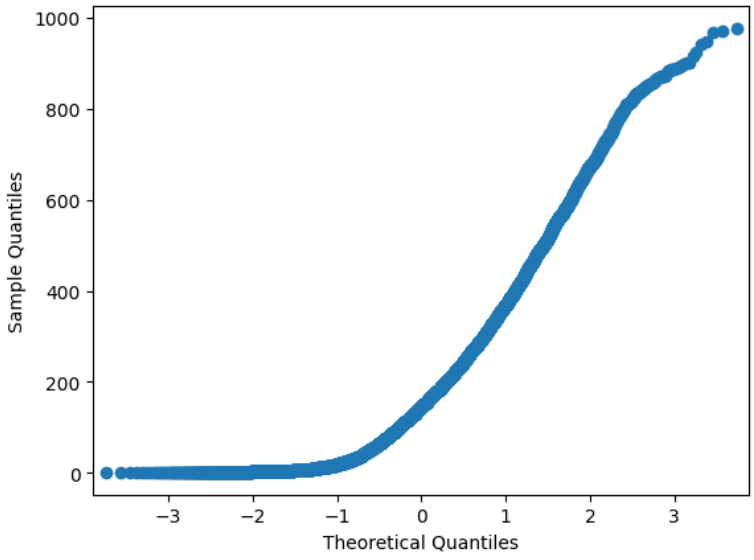
```
y=np.random.choice(df["count"],size=5000)
shapiro(y)
```

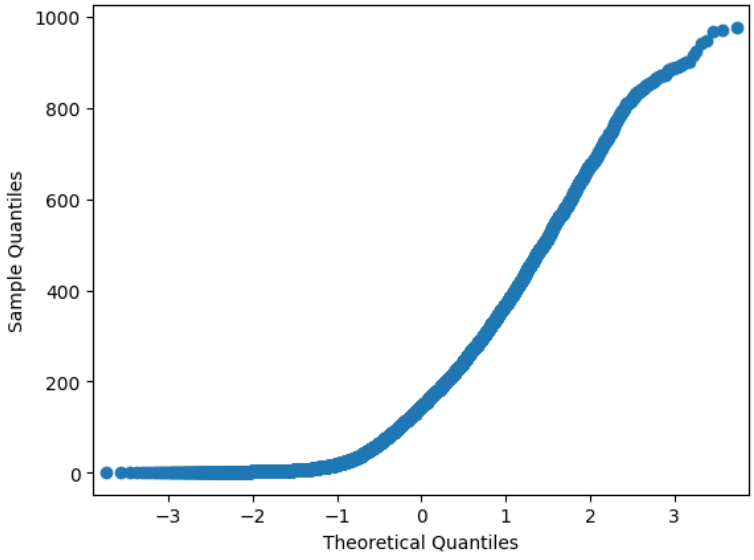
```
ShapiroResult(statistic=0.8817499876022339, pvalue=0.0)
```

Pvalue is less then our Alfa, Hence we will **reject** our Null Hypothesis. Thus, data is **not normally distributed**.

Let's Support out test with the help of QQ-Plot.

```
sm.qqplot(df["count"])
```





Hence, we can clearly see the down curve that shows the data is not **Normally Distributed**

Lets check for **Equal Variance**

Hypothesis for Equal Variance are:

- **Null Hypothesis**: Are Equally Variance
- **Alternate Hypothesis**: Data is not equally variance
- Alfa: 0.05

We have s1,s2,s3,s4 lists of mean of mean of four seasons.

```
levene(s1,s2,s3,s4)
```

```
LeveneResult(statistic=60.07403711803162, pvalue=5.556915712239204e-38)
```

Hence, Pvalue is very less then our **Alfa**, we will **reject our Null Hypothesis**

We will go with **Kruskal Test**

Hypothesis for Kruskal are:

- Null Hypothesis: Season has no relation in Rental bikes
- Alternate Hypothesis: Season has relation in Rental bikes
- Alfa: 0.05

```
kruskal(s1,s2,s3,s4)
```

```
KruskalResult(statistic=3690.854636977922, pvalue=0.0)
```

As Pvalue is less then **Alfa**, We will reject our **Null Hypothesis** Hence,

- **Season Has An Impact On Rental Bikes**

Approach 5

Lets Check if Weather is dependent on the season

We have two columns, first of "**Weather**" and second of "**Season**". To find out whether Weather depends of Season or not we need to do the test specifically for **Categorial variables** as both the columns are Categorical. And for Categorical vs Categorical test we have "**ChiSquare test**" i.e. **Chi2_contingency**.

- First we need to create a variabele using Crosstab of these columns.

```
X=pd.crosstab(df["season"],df["weather"])
```

X

weather	1	2	3	4
season				
1	1759	715	211	1
2	1801	708	224	0
3	1930	604	199	0
4	1702	807	225	0

As we got the cross table of **Weather vs Season**

- Now we can apply **Chi2_contingency** on it to check are these two relatable or not.

Setting Hypothesis for the test:

- Null Hypothesis: **Weather is not dependent on the season**
- Alternate Hypothesis: **Weather is dependent on the season**
- Alfa: 0.05

```
chi2_contingency(X)
```

```
Chi2ContingencyResult(statistic=49.158655596893624, pvalue=1.549925073686492e-07, dof=9, expected_freq=array([[1.77454639e+03,
6.99258130e+02, 2.11948742e+02, 2.46738931e-01],
       [1.80559765e+03, 7.11493845e+02, 2.15657450e+02, 2.51056403e-01],
       [1.80559765e+03, 7.11493845e+02, 2.15657450e+02, 2.51056403e-01],
       [1.80625831e+03, 7.11754180e+02, 2.15736359e+02, 2.51148264e-01]]))
```

As we observe the **Pvalue** is almost 0 which is less then our **Alfa** i.e. 0.05

- We will **Reject out Null Hypothesis**
- Hence **Weather is dependent on Season**

CONCLUSION

Inference from the analysis

From the Approach 1,

- By looking at the graph we can see that from month **4th to 10th** we have hike in Rental bikes but Year end is not doing great for **both years**
- There must be some factor which are related to months which are affecting in rental bikes like Weather,Season,Workingdays,etc

From Approach 2,

- From graph we have observed that different Weathers has drastic affects in rental bikes.
- But we also have seen that India doesn't have drasyic changes in Weather. It remains under the conditions which are good for rental bikes.
- Infact, Months having more bad weathers has no negative impact on rental bikes.

From Approach 3,

- From the grap it was not clear as Working days has positive or negative impact.
- We do the T-Test in which we reject the Hypothesis i.e. Working days or holidays has impact on rental bikes.
- If we go with Test then Working days and holidays do affect the rental bikes.

From Approach 4,

- From the graph we can see that Seasons are affecting the Rental bikes.
- Season 2 & 3 i.e Summer & Fall season has positive impact where as Winter has negative and Spring has low rental bikes.
- To support these claim we perform Kruskal test which hold the claim stongly that Season has impact on Rental bikes.

From Approach 5,

- From the ChiSquare Test Weather is dependent on Season, so indirectly Season & weather has link in positive and negative trend of rental bikes.

Suggestions

1. As Season 4 arrives or at the end of the Season 3 we can boost our **Marketing campaigns** to stop the fall in trend in year end.
2. We can have specific team only for Winter to drive the customer's attraction towards Renting bikes.
3. Need to make changes in our product to make it Winter friendy.
4. We need to bring offers specifically for winter season as Winter only shows Negative trend.

✓ 0s completed at 1:23 PM