Yulu - Hypothesis Testing case study

Business Problem -

Yulu suffered considerabble dips in its revenue market , so it wants to know the variables which are affecting the revenue of thier company, and which variable would help them to improve thier revenue and how significantly predicting the demand for shared electric cycles in the Indian market?

```
import pandas as pd #importing pandas for cleaning and manipulation
import numpy as np #importing numpy for statistical calculations
```

```
df= pd.read_csv('yulu.txt')# reading the csv file of YULU data
df
```

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
0	2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0000	3	13	16
1	2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0000	8	32	40
2	2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0000	5	27	32
3	2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0000	3	10	13
4	2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0000	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
10881	2012-12-19 19:00:00	4	0	1	1	15.58	19.695	50	26.0027	7	329	336
10882	2012-12-19 20:00:00	4	0	1	1	14.76	17.425	57	15.0013	10	231	241
10883	2012-12-19 21:00:00	4	0	1	1	13.94	15.910	61	15.0013	4	164	168
10884	2012-12-19 22:00:00	4	0	1	1	13.94	17.425	61	6.0032	12	117	129
10885	2012-12-19 23:00:00	4	0	1	1	13.12	16.665	66	8.9981	4	84	88

10886 rows × 12 columns

Next steps: Generate code with df View recommended plots New interactive sheet

EDA

- Checking the Data types of data set columns
- Checking the structure (Shape of Datasets)
- Summary Statistics Analysis of Data (describe)
- Analyzing Non null values

```
df.dtypes # Analyzing the data types of columns
```

	0
datetime	object
season	int64
holiday	int64
workingday	int64
weather	int64
temp	float64
atemp	float64
humidity	int64
windspeed	float64
casual	int64
registered	int64
count	int64

dtype: object

```
df.shape # Analyzing the Shape of data

df2= str(df.shape)
shaped= df2.split(',')
print(f'The shape of the data is Row = {shaped[0]} and columns{shaped[1]}')
```

```
The shape of the data is Row = (10886 and columns 12)
```

```
df.describe() # Analyzing the statistics data  of data
```

	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	re
count	10886.000000	10886.000000	10886.000000	10886.000000	10886.00000	10886.000000	10886.000000	10886.000000	10886.000000	1088
mean	2.506614	0.028569	0.680875	1.418427	20.23086	23.655084	61.886460	12.799395	36.021955	15
std	1.116174	0.166599	0.466159	0.633839	7.79159	8.474601	19.245033	8.164537	49.960477	15
min	1.000000	0.000000	0.000000	1.000000	0.82000	0.760000	0.000000	0.000000	0.000000	
25%	2.000000	0.000000	0.000000	1.000000	13.94000	16.665000	47.000000	7.001500	4.000000	3
50%	3.000000	0.000000	1.000000	1.000000	20.50000	24.240000	62.000000	12.998000	17.000000	11
75%	4.000000	0.000000	1.000000	2.000000	26.24000	31.060000	77.000000	16.997900	49.000000	22
max	4.000000	1.000000	1.000000	4.000000	41.00000	45.455000	100.000000	56.996900	367.000000	88

```
df.info()
# there are  12  columns and there is no null value at all in all the columns
# 12 columns comprises of 8 numerical columns and 1 categorial (date an time)column
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10886 entries, 0 to 10885
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   datetime    10886 non-null  object
 1   season      10886 non-null  int64
 2   holiday     10886 non-null  int64
 3   workingday  10886 non-null  int64
 4   weather     10886 non-null  int64
 5   temp        10886 non-null  float64
 6   atemp       10886 non-null  float64
 7   humidity    10886 non-null  int64
 8   windspeed   10886 non-null  float64
 9   casual      10886 non-null  int64
 10  registered  10886 non-null  int64
 11  count       10886 non-null  int64
dtypes: float64(3), int64(8), object(1)
memory usage: 1020.7+ KB
```

```
df.nunique()
#finding unique values of each column have
```

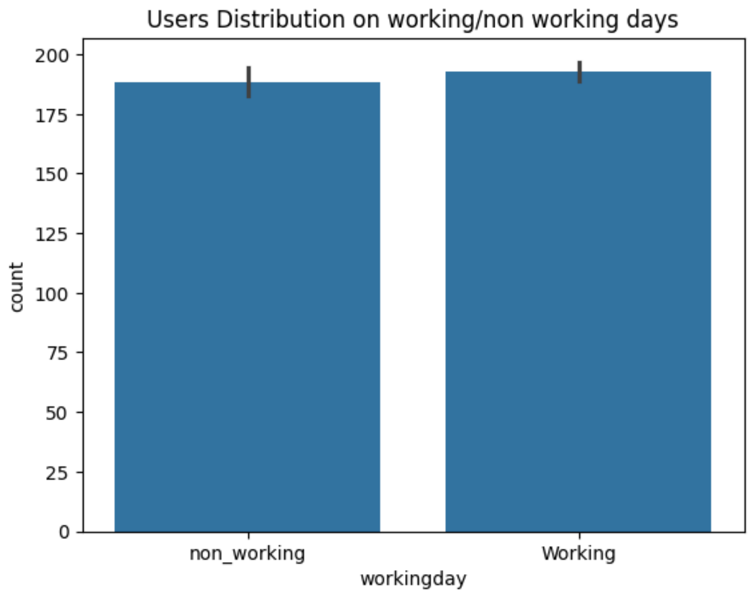
	0
season	4
holiday	2
workingday	2
weather	4
temp	49
atemp	60
humidity	89
windspeed	28
casual	309
registered	731
count	822
date	456
time	24

dtype: int64

Visual Analysis

Univariate Analysis

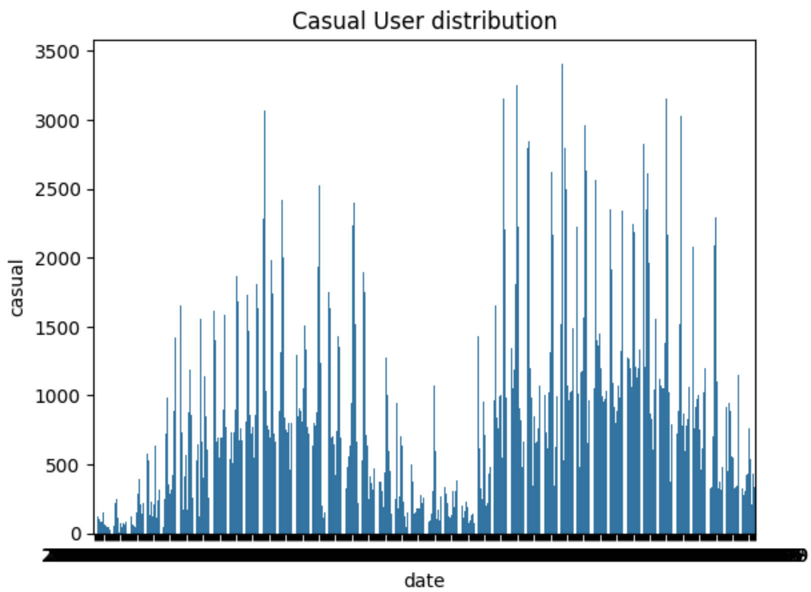
```
import matplotlib.pyplot as plt #import matplotlib library for visual analysis
import seaborn as sns #import seaborn library for visual analysis
df['workingday'] = df['workingday'].replace({0: 'non_working' , 1: 'Working'}) #replace 0 and 1 to actual meaniong which they are refferi
sns.barplot(x= 'workingday', y='count', data =df)

plt.title('Users Distribution on working/non working days')
plt.show()
```



Here by we can state that

- On the Working day there are more users in comparision of non working day

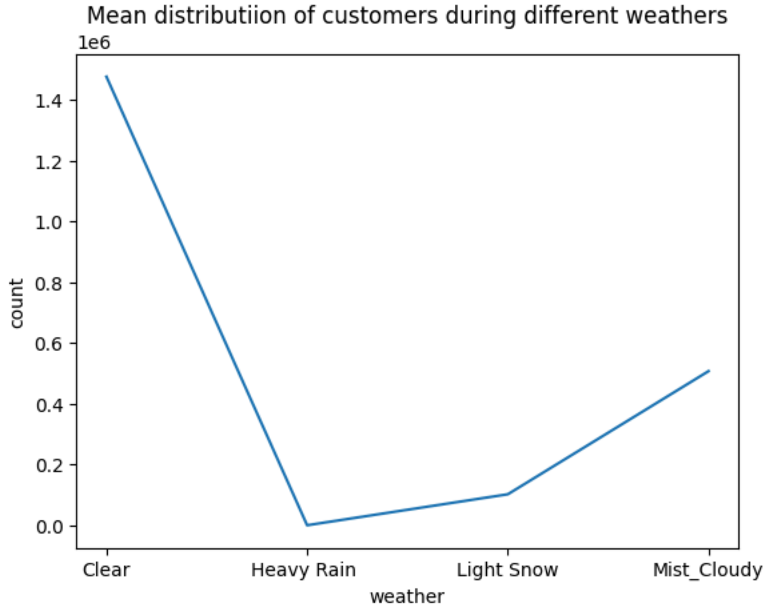
```
casul_sums =df.groupby(['date'])['casual'].sum() #summing up all the casual users on each day
df8= casul_sums.reset_index()
sns.barplot(data =df8 , x= 'date' , y='casual')
plt.title('Casual User distribution')
plt.show()
```



Bivariate Analysis

Double-click (or enter) to edit

```
df['weather'] = df['weather'].replace({1 : 'Clear' , 2 : 'Mist_Cloudy', 3 : 'Light Snow', 4 :'Heavy Rain '})
  ## Replace the weather codes with descriptive labels
customers_weather = df.groupby(['weather'])['count'].sum() ## Group by weather and sum the count
df7= customers_weather.reset_index()
sns.lineplot(data=df7 , x= 'weather', y='count')
plt.title('Mean distributiion of customers during different weathers')
plt.show()
```



Distribution Of Users during different types of weather

Clear 205.236791

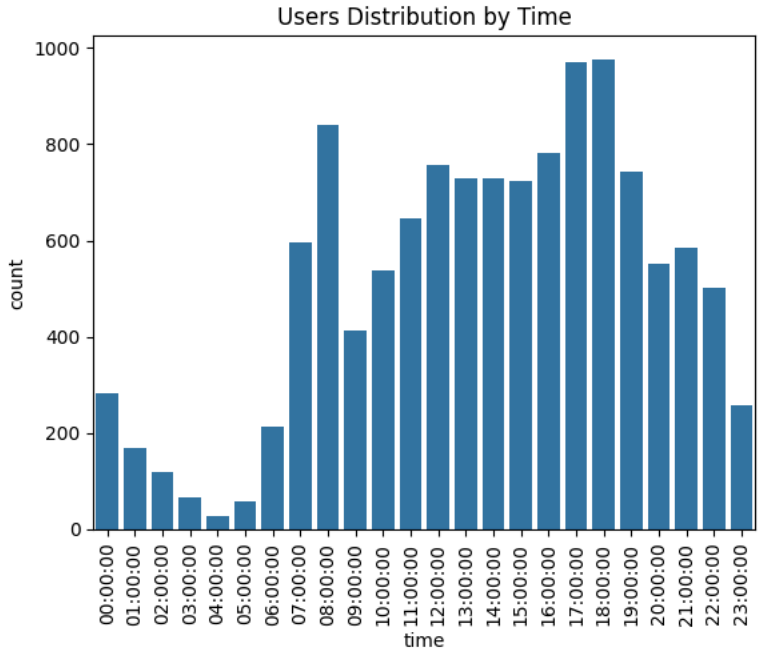
Heavy Rain 164.000000

Light Snow 118.846333

Mist_Cloudy 178.955540

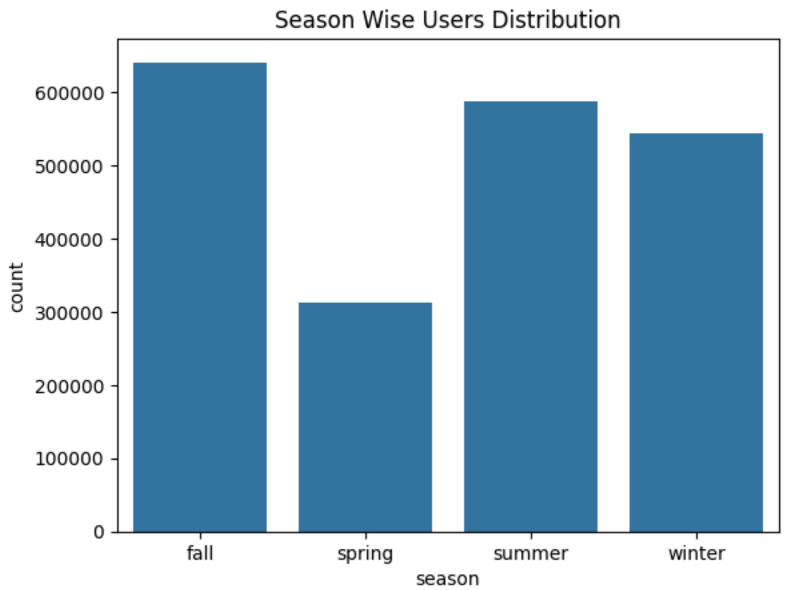
This States that there are very much users who enjoys riding YULU in clear weather the most , where in Heavy rain there are users who are very less or least

```
#at which time the hoghest number of users using YULU

time_max= df.groupby(['time'])['count'].max(5) # Group by time and max the count
timetime= time_max.reset_index() #reseting index
timetime
sns.barplot(data=timetime, x='time' , y='count')
plt.xticks(rotation=90) # rotating x- labels to 90 degree for clear visualisation
plt.title('Users Distribution by Time')
plt.show()
```



According to this here we state that **peak** time when most users enjoying riding YULU is **18:00:00**

```
#Replace the weather codes with descriptive labels
df['season']= df['season'].replace({1:'spring', 2:'summer', 3:'fall',4:'winter'})
df10= df.groupby(['season'])['count'].sum() # grouping season and sum the counts
df11 = df10.reset_index()
sns.barplot(data =df11 , x='season' ,y='count')
plt.title('Season Wise Users Distribution')
plt.show()
```



By this Analysis we can state that in the season of Fall or Rain Fall , the most number of users enjoying riding YULU

Insights Based on EDA

- The shape of the data is Row = (10886 and columns 12)
- On the Working day there are more users in comparision of non working day
- there are 12 columns and there is no null value at all in all the columns
- 12 columns comprises of 8 numerical columns and 1 categorial (date an time)column
- There are four season 1: spring, 2: summer, 3: fall, 4: winter

- There are four unnique Weather types : Clear , Mist , Light Snow , Heavy Rain
- Here by we can state that , On the Working day there are more users in comparision of non working day
- Clear 205.236791

  Heavy Rain 164.000000

  Light Snow 118.846333

  Mist_Cloudy 178.955540

  This States that there are very much users who enjoys riding YULU in clear weather the most , where in Heavy rain there are users who are very less or least
- According to this here we state that **peak** time when most users enjoying riding YULU is **18:00:00**
- By this Analysis we can state that in the season of Fall or Rain Fall , the most number of users enjoying riding YULU

HYPOTHESIS TESTING

Question 1- Working Day has effect on number of electric cycles rented

As we have to compare the two categories we will decide to do

Two-sample t-test (Independent t-test)

Here **NULL HYPOTHESIS** Is

H0 : The mean number of cycles rented on working days is the same as on non-working days.

Alternate Hypothesis is

Ha : The mean number of cycles rented on working days is different from non-working days.

AT **95% Confidence** we will do this Two sample t-test

here **alpha** = 5% or 0.05%

```
import scipy.stats as stats #importing scipy stats for hypothesis testing
working_days = df[df['workingday']== 'Working']['count'] # filter where working day is come and count that
non_working_days = df[df['workingday']== 'non_working']['count'] ## filter where working day is come and count that
t_test , p_val = stats.ttest_ind(working_days, non_working_days)
t_test,p_val
```

```
(np.float64(1.2096277376026694), np.float64(0.22644804226361348))
```

Here **p_val = 0.22** which is Greater than alpha value **p_val > alpha value** [this mean we fail to reject our null hypothesis]

Conclusion - The mean number of cycles rented on working days is same from non-working days.

As P_value which we calculate is greater than alpha value i.e, **0.22 >0.05** ,

Means there is no significant diffeernce between the mean of cycles renrted on working days or on non working days

Question 2- No. of cycles rented similar or different in different seasons

In this test we have to compare different categories (more than 2) i.e We have to calculate 4 different seasons count of users

so **ANOVA** is most suitable test here

ANOVA Test

Here the **NULL HYPOTHESIS** Is

H0: The mean number of cycles rented is the same across all seasons

The **Alternate Hypothesis** is

Ha: The mean number of cycles rented acroos all the seasons is different

Here also we set the **Significance value** at **0.05%**

i.e, alpha value = 0.05

```
from scipy.stats import f_oneway
spring_season = df[df['season']== 'spring']['count'] # filtering and counting the numbers of seasons
summer_season = df[df['season']== 'summer']['count'] # filtering and counting the numbers of seasons
fall_season = df[df['season']=='fall']['count'] # filtering and counting the numbers of seasons
winter_season = df[df['season']== 'winter']['count'] # filtering and counting the numbers of seasons

f_stats , p_value = f_oneway(spring_season, summer_season,fall_season,winter_season )
f_stats, p_value
```

```
(np.float64(236.94671081032106), np.float64(6.164843386499654e-149))
```

Conclusion - Here Pvalue is much larger than our expected significance value or alpha value p_Value = 6.16 which is much larger than 0.05

as **6.16 > 0.05** (p_value > alpha_value)

so here **WE FAIL TO REJECT OUR NULL HYPOTHESIS**

Means - the conclusion we drawn from this test is

The mean number of cycles rented is the same across all seasons

Question 3- No. of cycles rented similar or different in different weather

Here again we have to compare different categorical columns or we can say we have to compare Different weather types in which user enjoy the ride

so again **ANOVA** is best suitable test here

ANOVA Test

Here **NULL HYPOTHESIS** is

H0: The mean of all the users is same across all weather types

and **ALTERNATE HYPOTHESIS** is

Ha: The Mean of all users is different across all the weather types

Here also we set the **Significance value** at **0.05%**

i.e, alpha value = 0.05

```
df['weather'].unique()
```

```
array(['Clear', 'Mist_Cloudy', 'Light Snow', 'Heavy Rain '], dtype=object)
```

```
from scipy.stats import f_oneway
Clear_weather = df[df['weather'] == 'Clear']['count'] # filtering and counting the numbers of weathers
Mist_weather = df[df['weather']== 'Mist_Cloudy']['count'] # filtering and counting the numbers of weathers
Light_weather = df[df['weather']=='Light Snow']['count'] # filtering and counting the numbers of weathers

f_statsist , p_valuess = f_oneway(Clear_weather, Mist_weather, Light_weather)
f_statsist, p_valuess
```

```
(np.float64(98.28356881946706), np.float64(4.976448509904196e-43))
```

```
print(Clear_weather.shape[0])
print(Mist_weather.shape[0])
print(Light_weather.shape[0])
print(Heavy_weather.shape[0]) # analyzing the shape of every weather types came across the data
```

```
7192
2834
859
0
```

Here there is no user who ride the bicycle at HEAVY WEATHER weather type , so its count came 0 across all columns , so we can not take that type of weather into consideration otherwise it will affect the f_ratio

So here p_value comes 4.9 which is much greater than 0.05

Conclusion :
here our p_value comes 4.97 which is much larger then alpha value 0.05

4.97 > 0.05 (p_value > alpha value) so here once again

WE FAIL TO REJECT OUR NULL HYPHOTHESIS

and the conclusion is

The mean of all the users is same across all weather types

Question 4- to check the weather dependency on seasons

For checking the dependancy of one categorical column with the other

CHI-SQUARE Independence is best suitable test for this .

Here **NULL HYPOTHESIS** is

H0: Weather and Seasons are independent (the distribution of weather does not depend on the season).

and **ALTERNATIVE HYPOTHESIS** is

Ha : Weather and Season are dependent (the distribution of weather varies across different seasons).

Here also we set the **Significance value** at **0.05%**

i.e, alpha value = 0.05

```
# Create a contingency table of counts for 'weather' and 'season'
contingency_table = pd.crosstab(df['weather'], df['season'])

print(contingency_table)
```

```
season       fall  spring  summer  winter
weather
Clear        1930    1759    1801    1702
Heavy Rain      0       1       0       0
Light Snow    199     211     224     225
Mist_Cloudy   604     715     708     807
```

```
from scipy.stats import chi2_contingency

# Perform the Chi-Square Test of Independence
chi2_stat, p_value, dof, expected = chi2_contingency(contingency_table)

chi2_stat, p_value, dof , expected
```

```
(np.float64(49.15865559689363),
 np.float64(1.5499250736864862e-07),
 9,
 array([[1.80559765e+03, 1.77454639e+03, 1.80559765e+03, 1.80625831e+03],
        [2.51056403e-01, 2.46738931e-01, 2.51056403e-01, 2.51148264e-01],
        [2.15657450e+02, 2.11948742e+02, 2.15657450e+02, 2.15736359e+02],
        [7.11493845e+02, 6.99258130e+02, 7.11493845e+02, 7.11754180e+02]]))
```

Conclusion - Here P_value is large than our expected significance value or alpha value p_Value = 1.54 which is larger than 0.05

as **1.54 > 0.05** (p_value > alpha_value)

so here **WE FAIL TO REJECT OUR NULL HYPOTHESIS**

Means - the conclusion we drawn from this test is

Weather and Seasons are independent (the distribution of weather does not depend on the season).

Business Insights -

- . According to this here we state that **peak** time when most users enjoying riding YULU is **18:00:00**
- . The mean number of cycles rented on working days is same from non-working days. As P_value which we calculate is greater than alpha value i.e, **0.22 >0.05 , Means there is no significant diffeernce between the mean of cycles renrted on working days or on non working days**
- . The mean number of cycles rented is the same across all seasons p_Value = 6.16 which is much larger than 0.05 as **6.16 > 0.05** (p_value > alpha_value) so here we conclude **WE FAIL TO REJECT OUR NULL HYPOTHESIS**

  The mean number of cycles rented is the same across all seasons
- . The mean of all the users is same across all weather types
  as p_value comes 4.97 which is much larger then alpha value 0.05

  4.97 > 0.05 (p_value > alpha value)

  and the conclusion is

  The mean of all the users is same across all weather types
- . Also there is no one who is riding the bicycle in **heavy rain** weather (as it is expected also)
- . the most number of users enjoying their rides in **CLEAR WEATHER**
- . the distribution of weather does not depend on the season

  as **1.54 > 0.05** (p_value > alpha_value)

  WE FAIL TO REJECT OUR NULL HYPOTHESIS

  Weather and Seasons are independent

- . According to this here we state that **peak** time when most users enjoying riding YULU is **18:00:00**
- . The mean number of cycles rented on working days is same from non-working days. As P_value which we calculate is greater than alpha value i.e, **0.22 >0.05** , **Means there is no significant diffeernce between the mean of cycles renrted on working days or on non working days**
- . The mean number of cycles rented is the same across all seasons p_Value = 6.16 which is much larger than 0.05 as **6.16 > 0.05** (p_value > alpha_value) so here we conclude **WE FAIL TO REJECT OUR NULL HYPOTHESIS**

  The mean number of cycles rented is the same across all seasons
- . The mean of all the users is same across all weather types
  as p_value comes 4.97 which is much larger then alpha value 0.05

  4.97 > 0.05 (p_value > alpha value)

  and the conclusion is

  The mean of all the users is same across all weather types
- . Also there is no one who is riding the bicycle in **heavy rain** weather (as it is expected also)
- . the most number of users enjoying their rides in **CLEAR WEATHER**
- . the distribution of weather does not depend on the season

  as **1.54 > 0.05** (p_value > alpha_value)