Business Case: **Yulu**

Yulu is India’s leading mircro-mobility service provider, which offers unique vehicles for the daily commute. Starting off as a mission to eliminate traffic congestion in India, Yulu provides the safest commute solution through a user-friendly mobile app to enable shared,solo and sustainable commuting. Here we are going to analyze some factors that affect the demand for these shared electric cycles in the Indian market.

**Analysing basic metrics**

1. From google.colab import files
2. df= pd.read\_csv(‘bike\_sharing.txt’)
3. df.head(10)
4. df.info()
5. df.isnull().sum()
6. df[‘date’]= pd.to\_datetime(df[‘date’])

**Shape & Data Types of all the Attributes**

1. df.shape

2. df.dtypes

**Conversion of Categorical attributes**

df['season']= df['season'].astype('category')

df['holiday'] = df['holiday'].astype('category')

df['workingday'] = df['workingday'].astype('category')

df['weather'] = df['weather'].astype('category')

**Statistical Summary**

1. df.describe()

**Non- Graphical Analysis**

1. df.nunique()

**Visual Analysis- Univariate & Bivariate**

**For Univariate -Continous Variable**

**(Histogram)**

sns.histplot(x= 'humidity', data = df, bins = 20, kde = True, color = 'blue')

plt.xlabel('Humidity')

plt.ylabel('Frequency')

plt.title('Distribution of Humidity Levels')

**(Distplot)**

sns.displot(data=df,x= ‘temp’,kde = True, bins  = 20,color= ‘green’)

plt.title('Temperature Distribution')

plt.xlabel('Temperature', fontsize = 12)

plt.ylabel('Frequency',fontsize = 12)

plt.show()

**(Countplot)**

sns.countplot(x= 'season', data = df,palette= 'pastel')

plt.title('Season Distribution')

plt.xlabel('Season')

plt.ylabel('Count')

plt.show()

**For Categorical Variable**

sns.boxplot(x = 'season', y = 'count',data = df)

plt.title('Season Distribution')

plt.show()

sns.boxplot(x='weather', y='count', data=df, palette= 'coolwarm')

plt.xlabel('Weather Condition')

plt.ylabel('Bike Rentals')

plt.show()

**For Correlation- Heatmap & Pairplot**

**Heatmap**

corr = df.select\_dtypes(include=['number']).corr()

plt.figure(figsize=(10,6))

sns.heatmap(corr,annot=  True, cmap= 'coolwarm', fmt = '.2f')

plt.title('Correlation Heatmap')

plt.show()

**Pairplot**

sns.pairplot(df, vars=['temp', 'humidity', 'windspeed' ,'count'],hue= 'season')

plt.show()

**Missing Value & Outlier Detection**

**Missing Value**

missing\_values = df.isnull().sum()

missing\_percentage  = (df.isnull().mean())\*100

missing\_data = pd.DataFrame({'Missing Values': missing\_values, 'Percentage': missing\_percentage})

print(missing\_data)

**Outlier Detection**

numerical = df.select\_dtypes(include = ['float64','int64']).columns

Q1 = df[numerical].quantile(0.25)

Q3 = df[numerical].quantile(0.75)

IQR= Q3-Q1

lower\_bound = Q1-1.5\*IQR

upper\_bound = Q3+1.5\*IQR

outlier = df[(df[numerical]<lower\_bound)|(df[numerical]>upper\_bound)].any(axis = 1)

df['outlier']= outlier

**Business Insights (Non-Graphical & Visual Analysis)**

**Comments on Range of Attributes:**

1. **Season:** It represent 4 different seasons which are Winter, Summer, Fall & Spring.
2. **Temperature:** Most values are between 10 Degree c to 30 Degree C.
3. **Humidity:**  Most values are between 40% to 90%.
4. **Windspeed:**  Data is right skewed where the most values are below 30km/h.
5. **Count:** It represents bike rentals where cost varies from 1 to 1000 per hour.

**Comments on outliers of various attributes**

**Humidity:** There are no major outliers but have some extreme values at 0% and 100%.

**Windspeed:** It shows outliers as some of the values exceeds 40km/h which is very rare normally.

**Comments on Distributions & Relationships**

**Histogram:** It has a right skewed distribution where most humidity values range between 50%-90%.

**Displot:** It follows a normal distribution, peaking around 14.C to 32.C.

**Countplot:** Distribution across all seasons are balanced where season(3)-fall has the highest rentals.

**Boxplot (Season vs Count):** Out of 4 seasons, Season(2) and Season(3) shows higher median rentals while winter has the lowest.

**Boxplot (Weather vs Count):** In clear weather(category 1) rentals are the highest, while in rainy and snowy days leads to a drop.

**Comments for each univariate & bivariate plots**

1. Temperature and season have the strongest impact on rentals.
2. Demand for bikes occurs in Fall & Summer & low in winters specially when weather is not clear.
3. Holidays & working days impact rentals differently, on working days have higher rentals due to office commuters.

**Hypothesis Testing**

**(2\_Sample T-Test)**

**T\_Test:** This test determines if working days impact rentals, which can inform pricing strategies for weekdays vs weekends.

**Null Hypothesis (Ho):** The mean number of electric cycle rented is the same on working & non working days.

**Alternate Hypothesis (Ha):**  The mean number of electric cycle rented differs between working & non working days.

**Checking Assumptions**

1. If the data is independent
2. If the data is in normality or not.
3. Homogeneity of Variance

from scipy.stats import levene ,ttest\_ind,Shapiro, mannwhitneyu

Non\_working\_day = df[df['workingday']==0]['count']

working\_day = df[df['workingday']==1]['count']

**#Check Test for normality**

shapiro\_non\_working = shapiro(non\_working\_day)

shapiro\_working = shapiro(working\_day)

P\_val for working: 2.25e-61(extremely small)

P\_val for Non\_working: 4.47e-45(extremely small)

Since both the p\_values are smaller than 0.05, we reject the null hypothesis of normality. This means the rental count data is not normally distributed for both working and non-working days.

**#Check homogeneity of variance**

stat, pval = levene(working\_day,Non\_working\_day)

stat,pval=(np.float64(0.004972848886504472),np.float64(0.9437823280916695))

**As per pval (0.943) which is much greater than 0.05, so we fail to reject the null hypothesis, meaning the variances are equal.**

As we can see, data is not normal, we can use a non-parametric alternative (Mann-Whitney U Test)

stat,pval = mannwhitneyu(working\_day,non\_working\_day, alternative= 'two-sided')

stat,pval= 12868495.5, 0.96

**Interpretation:**  Since the pval (0.96) is greater than 0.05, we fail to reject the null hypothesis (Ho.)so there is no statistically significant difference in the number of electric cycles rented between working days and non working days.

**(ANNOVA)**

**Annova:** Understanding seasonal impact can help optimize inventory and marketing efforts. **Weather** and **Season** are categorical variables with multiple groups.

**(For Weather)**

**Null Hypothesis (Ho):** The mean number of electric cycle rented is the same across all weather conditions.

**Alternate Hypothesis (Ha):**  Atleast one weather condition has a different mean rental count.

**(For Season)**

**Null Hypothesis (Ho):** The mean number of electric cycle rented is the same across all seasons.

**Alternate Hypothesis (Ha):**  Atleast one season has a different mean rental count.

**Checking Assumptions**

1. If the data is independent
2. If the data is in normality or not.
3. Homogeneity of Variance

# **Levene’s test for homogeneity of variance**

**(For Weather)**

stat,pval= levene(df[df['weather']==1]['count'],df[df['weather']==2]['count'],df[df['weather']==3]['count'])

**(For Season)**

stat,pval= levene(df[df['season']==1]['count'],df[df['season']==2]['count'],df[df['season']==3]['count'],df[df['season']==4]['count'])

**Test Intepretations of Leven’s Test**

1. **P\_val for Weather =** 6.18 (very high) which means variance are equal.
2. **P\_val for Season =** 1.0147e-118 (very low) which means variance are not equal.

**Performing Anova Test**

weather\_anova=f\_oneway(df[df['weather']==1]['count'],df[df['weather']==2]['count'],df[df['weather']==3]['count'],df[df['weather']==4]['count'])

season\_anova=f\_oneway(df[df['season']==1]['count'], df[df['season']==2]['count'],df[df['season']==3]['count'], df[df['season']==4]['count'])

**Test Intepretations of Anova Test**

1. **P\_val for Weather =** 5.48e-42 (very small,<0.05) , it mean we reject the Ho. It means weather significantly affects bike rentals, number of bikes rented varies depending on the weather condition. For ex: (Clear weather vs Rainy weather)
2. **P\_val for Season=** 6.16e-149 (extremely small,<0.5), it means we reject the Ho. It means season significantly affects bike rentals. The demand of bike changes across seasons, meaning people rent more bikes in some seasons(like summer) than in others(like winters).

**(Chi-square test)**

**Chi-square test:** Relate to Business. For example: Since weather depends on seasons, Yulu can plan bike availability accordingly. **Weather** and **Season** are categorical variables.

**Null Hypothesis (Ho):**  Weather is independent of season.

**Alternate Hypothesis (Ha):**  Weather depends on season.

**Creating a contingency table between weather and season:**

contigency\_table= pd.crosstab(df['weather'],df['season'])

contigency\_table

**Running a Chi-square test:**

chi2,pval,dof,expected = chi2\_contingency(contigency\_table)

**Intepretations of Chi-square test:** Since the pval is (1.55e-07) is smaller than 0.05 so we reject the null hypothesis . It means that weather is dependent on the season, implying that certain weather conditions are more likely to occur in specific reasons.

**Summary of Key Insights:**

1. Weather and Season are not independent (certain weather conditions are more likely in specific seasons).
2. Working days do not significantly impact rentals.
3. Clear weather has higher rentals; rentals drop significantly in rainy/snowy conditions.
4. Temperature and season have the strongest impact on rentals.
5. Season significantly affects bike demand(highest in Fall & Summer,lowest in Winter).