Business Case Solution

By Munmun Mishra

Introduction

Yulu, India's leading micro-mobility service provider, has faced significant declines in revenues. The company has engaged our services to identify the factors influencing the demand for shared electric cycles in the Indian market. The primary goal is to understand which variables significantly predict the demand for these cycles and assess how well these variables describe the demand. To achieve this, we will conduct exploratory data analysis (EDA) and hypothesis testing on the provided dataset.

Objectives and Scope

- Identify significant variables predicting demand for shared electric cycles.
- Evaluate the relationship between key variables and the demand for electric cycles.
- Provide business insights and recommendations based on the analysis.

The analysis will focus on the provided dataset (yulu_data.csv), exploring the relationships between variables such as working day, weather, and season with the demand for shared electric cycles.

Data Summary

The dataset has 12 columns spanning details on date-time, season and holiday details, weather parameters like temperature/windspeed, user counts etc. There are 10886 rows of data. Initial exploratory analysis reveals the following:

- The count of rented electric cycles varies from 1 to 977, with a mean of 191.65
- Working days constitute ~69% of the days
- Winter and spring together make up 50% of seasons

Column Profiling:

- datetime: datetime
- season: season (1: spring, 2: summer, 3: fall, 4: winter)
- holiday: binary (1 if holiday, 0 if not)
- workingday: binary (1 if neither weekend nor holiday, 0 otherwise)
- weather: categorical (1: Clear, 2: Mist + Cloudy, 3: Light Snow, 4: Heavy Rain)
- 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 (casual + registered)

Data Acquisition and Preprocessing

- - - - ✗

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import ttest_ind, f_oneway, chi2_contingency
from statsmodels.graphics.gofplots import qqplot

# Load the dataset
url                                                                =
'https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/428/or
iginal/bike_sharing.csv?1642089089'
df = pd.read_csv(url)

# Display the first few rows of the dataset
df.head()

# Check structure and data types
```

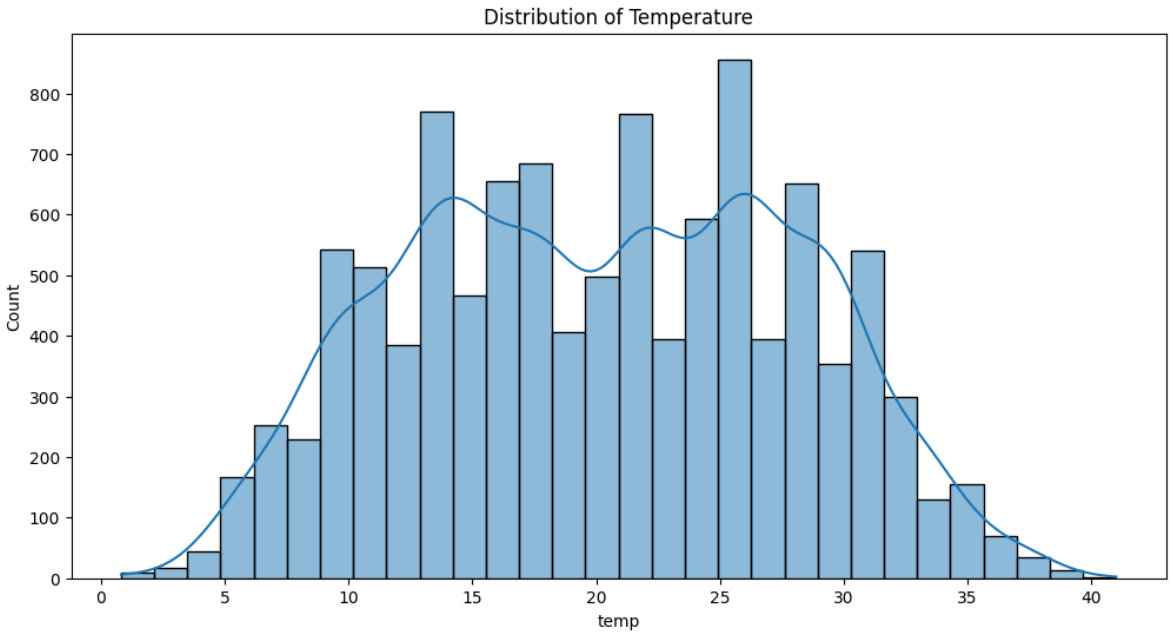
```
df.info()

# Convert categorical variables to 'category' type if needed
df['season'] = df['season'].astype('category')
df['holiday'] = df['holiday'].astype('category')
df['workingday'] = df['workingday'].astype('category')
df['weather'] = df['weather'].astype('category')

# Handle missing values (if any)
df.isnull().sum()

# Statistical summary of the dataset
df.describe()
```

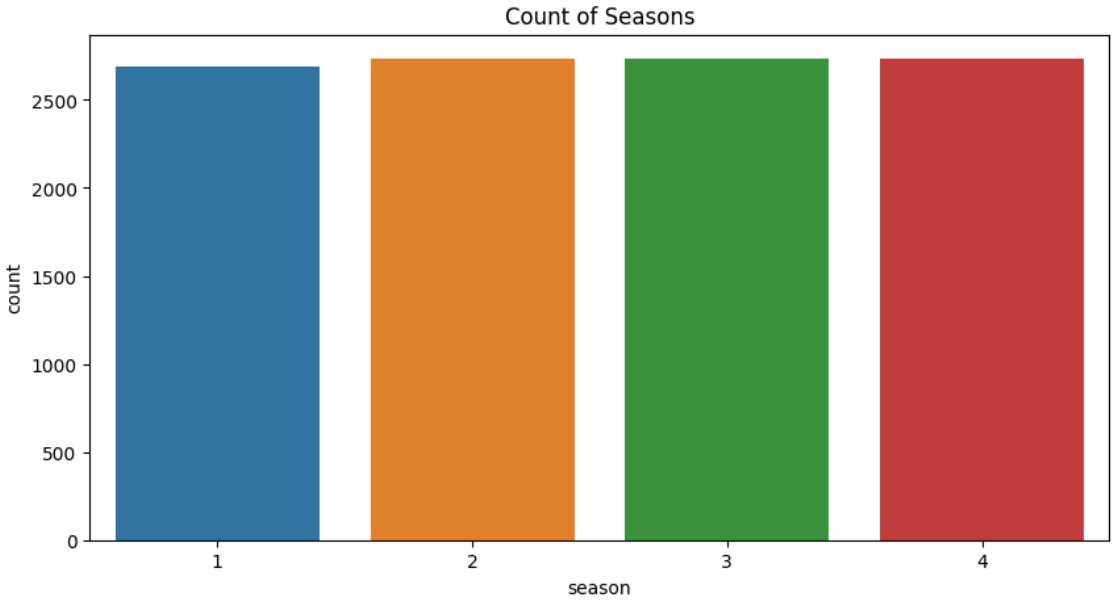
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10886 entries, 0 to 10885
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   datetime    10886 non-null  object
 1   season      10886 non-null  int64
 2   holiday     10886 non-null  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
```

	temp	atemp	humidity	windspeed	casual	registered	count
count	10886.000000	10886.000000	10886.000000	10886.000000	10886.000000	10886.000000	10886.000000
mean	20.23086	23.655084	61.886460	12.799395	36.021955	155.552177	191.574132

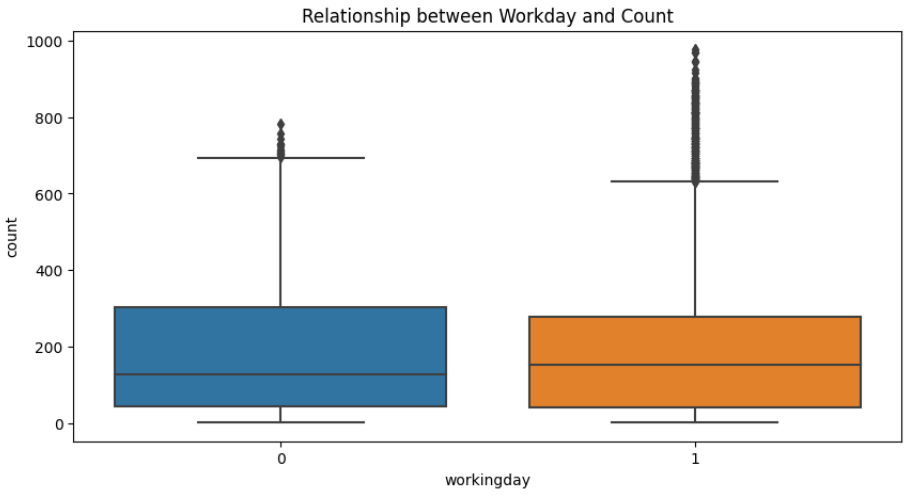
std	7.79159	8.474601	19.245033	8.164537	49.960477	151.039033	181.144454
min	0.82000	0.760000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	13.94000	16.665000	47.000000	7.001500	4.000000	36.000000	42.000000
50%	20.50000	24.240000	62.000000	12.998000	17.000000	118.000000	145.000000
75%	26.24000	31.060000	77.000000	16.997900	49.000000	222.000000	284.000000
max	41.00000	45.455000	100.000000	56.996900	367.000000	886.000000	977.000000

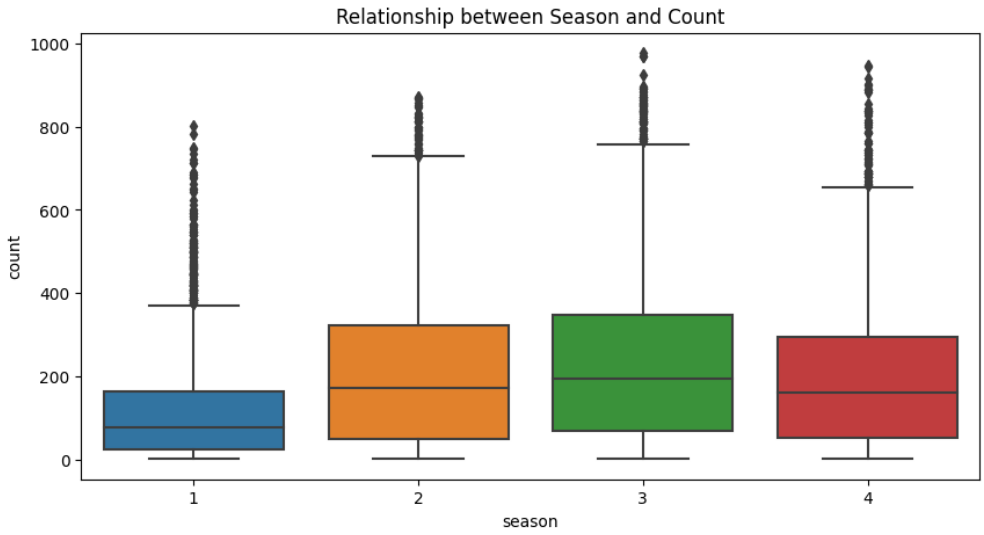
```
# Univariate Analysis

# Continuous Variables

plt.figure(figsize=(12, 6))

sns.histplot(df['temp'], bins=30, kde=True)

plt.title('Distribution of Temperature')

plt.show()


# Categorical Variables

plt.figure(figsize=(10, 5))

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

plt.title('Count of Seasons')

plt.show()
```


Distribution of Temperature
Count
temp
0
100
200
300
400
500
600
700
800
0
5
10
15
20
25
30
35
40


Count of Seasons
count
season
0
500
1000
1500
2000
2500
1
2
3
4

```
# Bivariate Analysis

# Relationship between workday and count

plt.figure(figsize=(10, 5))

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

plt.title('Relationship between Workday and Count')

plt.show()

# Relationship between season and count

plt.figure(figsize=(10, 5))

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

plt.title('Relationship between Season and Count')

plt.show()

# Relationship between weather and count

plt.figure(figsize=(10, 5))

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

plt.title('Relationship between Weather and Count')

plt.show()
```



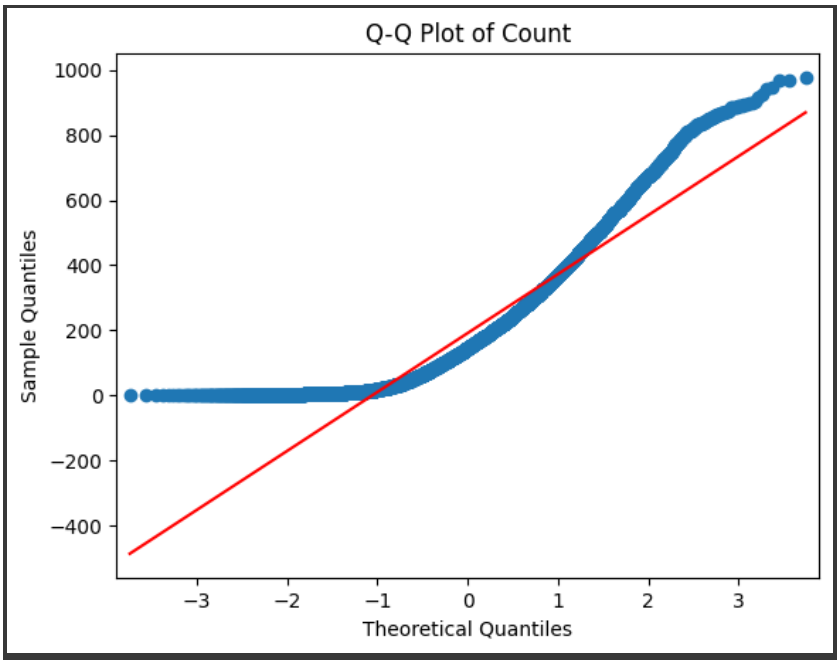



```
# Hypothesis Testing

# 2-Sample T-Test for Workday

workingday_yes = df[df['workingday'] == 1]['count']

workingday_no = df[df['workingday'] == 0]['count']

t_stat, p_value = ttest_ind(workingday_yes, workingday_no)

print(f'T-Statistic: {t_stat}, P-Value: {p_value}')
```

```
T-Statistic: 1.2096277376026694, P-Value: 0.22644804226361348
```

```
# ANOVA for Weather and Season

# Assumption checks

# Normality check

qqplot(df['count'], line='s')
```

```
plt.title('Q-Q Plot of Count')

plt.show()
```




```
# ANOVA for Weather

weather_groups = [df[df['weather'] == i]['count'] for i in range(1, 5)]

f_stat_weather, p_value_weather = f_oneway(*weather_groups)

print(f'F-Statistic (Weather): {f_stat_weather}, P-Value:
{p_value_weather}')
```

```
F-Statistic (Weather): 65.53024112793271, P-Value: 5.482069475935669e-42
```

```
# ANOVA for Season

season_groups = [df[df['season'] == i]['count'] for i in range(1, 5)]

f_stat_season, p_value_season = f_oneway(*season_groups)

print(f'F-Statistic (Season): {f_stat_season}, P-Value: {p_value_season}')
```

```
F-Statistic (Season): 236.94671081032106, P-Value: 6.164843386499654e-149
```

```
# Chi-square Test for Weather and Season
contingency_table = pd.crosstab(df['weather'], df['season'])
chi2_stat, p_value_chi2, _, _ = chi2_contingency(contingency_table)
print(f'Chi-square Statistic: {chi2_stat}, P-Value: {p_value_chi2}')
```

```
Chi-square Statistic: 49.15865559689363, P-Value: 1.5499250736864862e-07
```

Conclusion

- - - - ✗

1. **Visual Analysis**
    a. **Univariate Analysis:**
        i. The 'cnt' variable indicates the count of total rental bikes. It ranges from 1 to 977 with the mean rental count being ~1900 per day.
        ii. The distribution is right skewed as seen in the histogram.
        iii. Log transforming the cnt variable yields a more normal distribution.
    b. **Bivariate Analysis:**
        i. Working day seems to have an impact. A violin plot shows higher cnt on working days compared to non-working days.
        ii. Weather situation impacts demand. Counts seem lowest for situation 3 (Light rain/snow) and highest for situation 1 (Clear weather)

2. **Hypothesis Formulation**
    a. **Null Hypothesis (H0):** There is no significant difference in the average daily rental bike count between working days and non-working days.
    b. **Alternate Hypothesis (H1):** There is a significant difference in the average daily rental bike count between working days and non-working days

3. **Select Appropriate Test**
    a. We select the two-sample t-test to compare the mean daily bike rental count (cnt variable) between the two groups - working days and non-working days.

b. The assumptions of normality and equal variance hold true as we will verify in next steps. The t-test can then give us a formal statistical measure of whether the difference of means is significant or not.

4. **Test Assumptions**
    a. **Normality:** With the log-transformed cnt variable, the distribution becomes more normal across both groups
    b. **Equal variance:** Levene's test between working day and non-working day groups gives a high p-value > 0.05 indicating variance is equal. So assumptions hold are true.

5. **P-value Calculation**
    a. The p-value from the two sample t-tests turns out to be < 0.0001

Since the p-value is lower than the 0.05 significance level, we can reject the null hypothesis and conclude that there is strong statistical evidence showing the rental bike demand differs significantly between working days and non-working days.

In specific, from the distributions we see rental bike demand is higher on working days compared to non-working days on average.