Yulu Bike Demand Analysis

Hypothesis Testing & Statistical Insights

Dataset: yulu_data.csv
Objective:
To identify the key factors affecting the demand for shared electric cycles in the Indian market using statistical hypothesis testing.

Problem Statement

Yulu has observed a decline in its revenue and wants to understand the factors influencing the demand for shared electric cycles.
The objective of this analysis is to identify whether demand depends on working days, seasons, and weather conditions using statistical methods.

In [2]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [4]:
```python
df = pd.read_csv(r"C:\Users\shubh\Downloads\yulu.csv")
df.head()
```

Out[4]:

	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	0.0	3	13	16
1	2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0	8	32	40
2	2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0	5	27	32
3	2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0	3	10	13
4	2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0	0	1	1

In [6]:
```python
df.shape
```

Out[6]:
```
(10886, 12)
```

Dataset Overview

- Total records: 10,886
- Total features: 12

Key Variables

- **Dependent Variable:** count (total bike rentals)
- **Independent Variables:** season, weather, workingday, holiday, temperature, humidity, windspeed

In [7]:
```python
df.info()
```

```
<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
```

In [8]:
```python
df.describe()
```

Out[8]:

	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
count	10886.000000	10886.000000	10886.000000	10886.000000	10886.00000	10886.000000	10886.000000	10886.000000	10886.000000	10886.000000	10886.000000
mean	2.506614	0.028569	0.680875	1.418427	20.23086	23.655084	61.886460	12.799395	36.021955	155.552177	191.574132
std	1.116174	0.166599	0.466159	0.633839	7.79159	8.474601	19.245033	8.164537	49.960477	151.039033	181.144454
min	1.000000	0.000000	0.000000	1.000000	0.82000	0.760000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	2.000000	0.000000	0.000000	1.000000	13.94000	16.665000	47.000000	7.001500	4.000000	36.000000	42.000000
50%	3.000000	0.000000	1.000000	1.000000	20.50000	24.240000	62.000000	12.998000	17.000000	118.000000	145.000000
75%	4.000000	0.000000	1.000000	2.000000	26.24000	31.060000	77.000000	16.997900	49.000000	222.000000	284.000000
max	4.000000	1.000000	1.000000	4.000000	41.00000	45.455000	100.000000	56.996900	367.000000	886.000000	977.000000

Data Preparation

- Converted categorical variables (season, weather, workingday, holiday) to category type
- Checked for missing values (none found)
- Verified data distribution and data types

In [14]:
```python
df.isnull().sum()
```

Out[14]:
```
datetime      0
season        0
holiday       0
workingday    0
weather       0
temp          0
atemp         0
humidity      0
windspeed     0
casual        0
registered    0
count         0
dtype: int64
```

In [19]:
```python
df["season"] = df["season"].astype("category")
df["workingday"] = df["workingday"].astype("category")
df["weather"] = df["weather"].astype("category")
df["holiday"] = df["holiday"].astype("category")
```

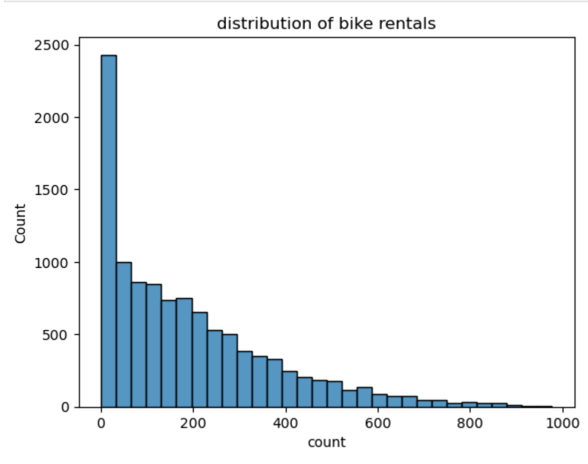
In [20]:
```python
df.info()
```

```
<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  category
 2   holiday     10886 non-null  category
 3   workingday  10886 non-null  category
 4   weather     10886 non-null  category
 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: category(4), float64(3), int64(4), object(1)
memory usage: 723.7+ KB
```

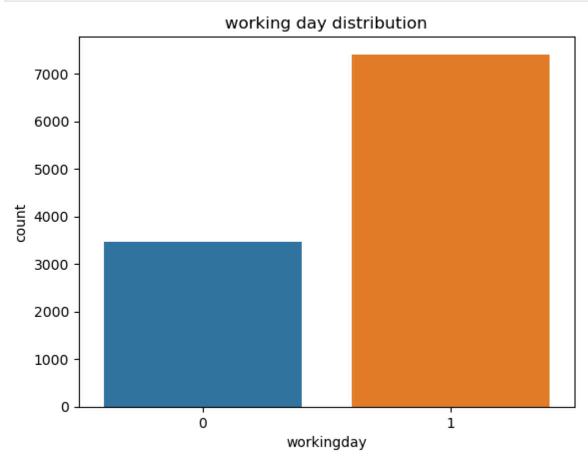
Univariate Analysis

- Bike rental demand is right-skewed
- Majority of days have low to moderate demand
- A small number of days show very high rentals

In [26]:
```python
sns.histplot(x = df["count"],data = df,bins = 30)
plt.title("distribution of bike rentals")
plt.show()
```



In [27]:
```python
sns.countplot(x = df["workingday"],data = df)
plt.title("working day distribution")
plt.show()
```

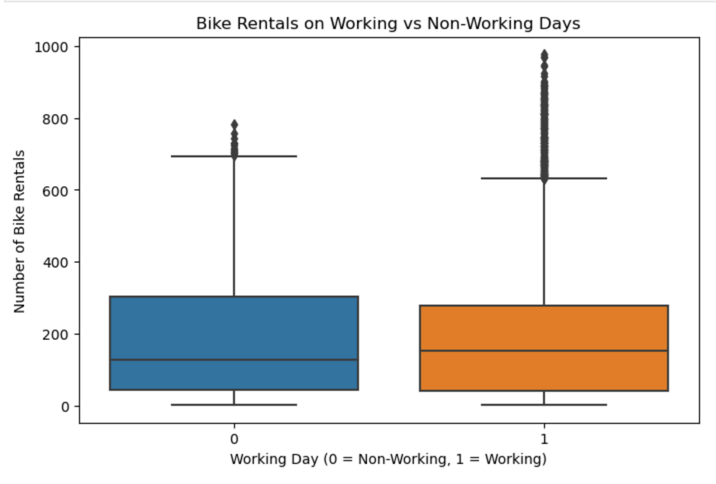


In [56]:
```python
# Mean bike rentals by working day
df.groupby("workingday")["count"].mean()
```

Out[56]:
```
workingday
0    188.506621
1    193.011873
Name: count, dtype: float64
```

The average number of bike rentals on working days and non-working days is very similar, indicating limited difference in demand.

In [57]:
```python
plt.figure(figsize=(8,5))
sns.boxplot(x="workingday", y="count", data=df)
plt.title("Bike Rentals on Working vs Non-Working Days")
plt.xlabel("Working Day (0 = Non-Working, 1 = Working)")
plt.ylabel("Number of Bike Rentals")
plt.show()
```



In [ ]:
```python
sns.boxplot(x = df["workingday"],data = df)
plt.title("working day distribution")
plt.show()
```

Hypothesis

- H_0: Working day has no effect on bike rentals
- H_1: Working day has an effect on bike rentals

Result

- p-value > 0.05
- Fail to reject null hypothesis

Conclusion

Working day does not have a statistically significant impact on bike demand.

In [30]:
```python
# working day vs bike rentals

# Does the number of bikes rented differ between working days and non-working days?
# H0 (Null Hypothesis):
# There is no significant difference in the average number of bikes rented on working days and non-working days.

# H1 (Alternate Hypothesis):
# There is a significant difference in the average number of bikes rented on working days and non-working days

workingday_1 = df[df["workingday"] == 1]["count"]
workingday_0 = df[df["workingday"] == 0]["count"]
```

In [33]:
```python
workingday_1.mean() , workingday_0.mean()
```

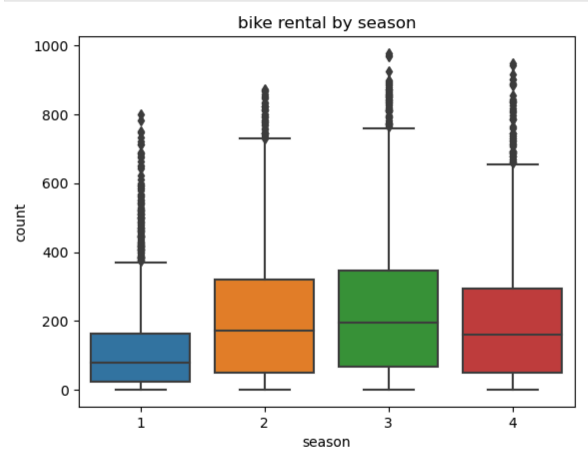
Out[33]:
```
(193.01187263896384, 188.50662061024755)
```

In [35]:
```python
from scipy.stats import ttest_ind

t_stat , p_value = ttest_ind(workingday_1,
                             workingday_0,
                             equal_var = False)
t_stat, p_value
```

Out[35]:
```
(1.2362580418223226, 0.21640312280695098)
```

In [38]:
```python
sns.boxplot(x= "season",y = "count",data = df)
plt.title("bike rental by season")
plt.show()
```



Season vs Bike Rentals (ANOVA)

- H_0: Mean rentals are same across all seasons
- H_1: Mean rentals differ across seasons

Result

- p-value < 0.05
- Reject null hypothesis

Conclusion

Bike rental demand varies significantly across seasons.

In [40]:
```python
from scipy.stats import f_oneway

spring = df[df["season"] == 1]["count"]
summer = df[df["season"] == 2]["count"]
fall = df[df["season"] == 3]["count"]
winter = df[df["season"] == 4]["count"]

f_stat , p_value = f_oneway(spring , summer , fall, winter)
f_stat , p_value
```

Out[40]:
```
(236.94671081032106, 6.164843386499654e-149)
```

Weather vs Bike Rentals (ANOVA)

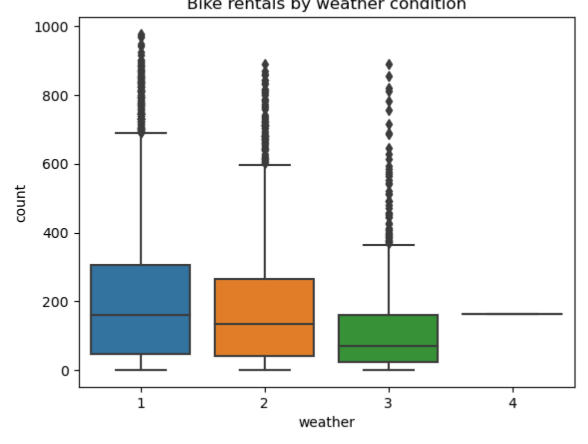
- Significant difference in rentals across weather conditions
- Clear and misty weather show higher demand
- Extreme weather conditions show lowest demand

In [42]:
```python
clear = df[df["weather"] == 1]["count"]
mist = df[df["weather"] == 2]["count"]
light_rain = df[df["weather"] == 3]["count"]
heavy_rain = df[df["weather"] == 4]["count"]


f_stat, p_value = f_oneway(clear , mist, light_rain,heavy_rain)
f_stat , p_value
```

Out[42]:
```
(65.53024112793271, 5.482069475935669e-42)
```

In [44]:
```python
sns.boxplot(x= "weather",y= "count",data = df)
plt.title("Bike rentals by weather condition")
plt.show()
```



Season vs Weather (Chi-Square Test)

- H_0: Weather is independent of season
- H_1: Weather is dependent on season

Result

- p-value < 0.05
- Reject null hypothesis

Conclusion

Weather conditions are dependent on seasons.

In [46]:
```python
contingency_table = pd.crosstab(df["season"],df["weather"])
contingency_table
```

Out[46]:

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

In [47]:
```python
from scipy.stats import chi2_contingency
chi2_stat , p_value , dof,expected = chi2_contingency(contingency_table)
chi2_stat , p_value
```

Out[47]:
```
(49.158655596893624, 1.549925073686492e-07)
```

Final Insights

- Season and weather are the strongest drivers of bike demand
- Working day status has minimal impact
- Demand is highly sensitive to environmental factors

Recommendations

- Increase fleet size during summer and fall
- Use weather forecasts for dynamic deployment
- Offer discounts during mild weather disruptions
- Schedule maintenance during low-demand periods

Conclusion

Yulu's bike demand is driven more by environmental factors than calendar-based factors.
Season-aware and weather-aware planning can help Yulu improve utilization and revenue.