Data Acquisition & Data Wrangling

Presented by:- Meet patel

Objectives:

- This project are to do Data Acquisition and Data Wrangling on 3 different dataset.
- Check : skewness analysis, identify unique values, datatype of the dataset, missing values

Tools and Libraries used:

Tools:

- Jupyter Notebook (interactive coding environment)
- Vs code for batter understanding (application)
- Python (programming language)

Libraries:

- Pandas (data manipulation and analysis)
- NumPy (numerical computations
- SciPy (scientific computing, specifically for Spearman rank correlation)
- Seaborn (data visualization, specifically for heatmaps and pair plots)
- Matplotlib (data visualization)

- Imported and merged datasets
- Handled missing values and outliers
- Removed unnecessary columns and redundant data
- Visualized relationships using heatmaps and pair plots

Merged dataset:

	Index	Date	Season	Year	Month	Hour	Holiday	Weekday	Weathersit	Tempreature	Humdity	Windspeed	Casual	Registered	Count
0	1	01-01-2011	1	0	1	0	False	6	1	0.24	0.81	0.0000	3	13	16
1	2	01-01-2011	1	0	1	1	False	6	1	0.22	0.80	0.0000	8	32	40
2	3	01-01-2011	1	0	1	2	False	6	1	0.22	0.80	0.0000	5	27	32
3	4	01-01-2011	1	0	1	3	False	6	1	0.24	0.75	0.0000	3	10	13
4	5	01-01-2011	1	0	1	4	False	6	1	0.24	0.75	0.0000	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
385	615	28-01-2011	1	0	1	20	False	5	2	0.24	0.70	0.1940	1	61	62
386	616	28-01-2011	1	0	1	21	False	5	2	0.22	0.75	0.1343	1	57	58
387	617	28-01-2011	1	0	1	22	False	5	1	0.24	0.65	0.3582	0	26	26
388	618	28-01-2011	1	0	1	23	False	5	1	0.24	0.60	0.2239	1	22	23
389	619	29-01-2011	1	0	1	0	False	6	1	0.22	0.64	0.3582	2	26	28

Null Values Check:

```
Index          0
Date           0
Season         0
Year           0
Month          0
Hour           0
Holiday        0
Weekday        0
Weathersit     0
Tempreature    0
Humdity        0
Windspeed      0
Casual         0
Registered     0
Count          0
dtype: int64
```

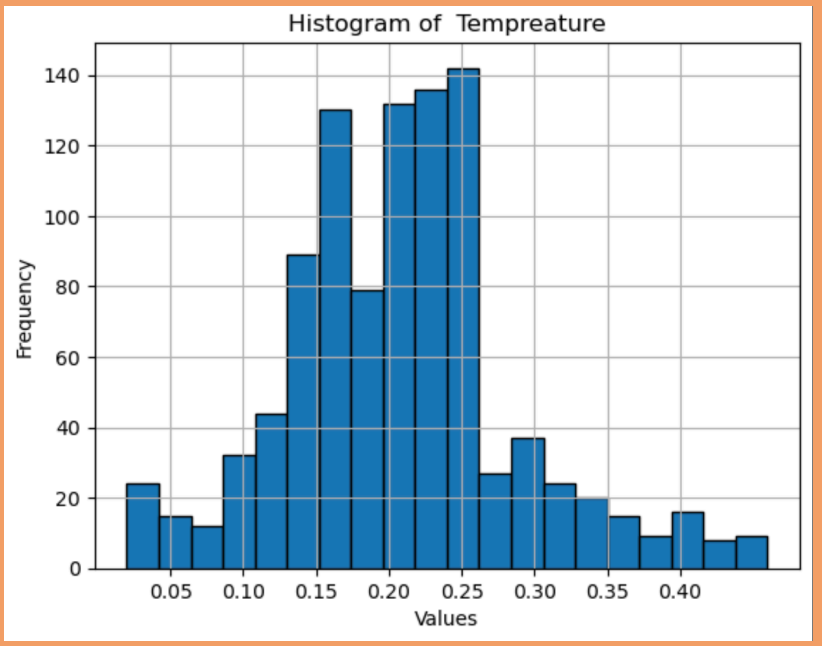
Skewness analysis:

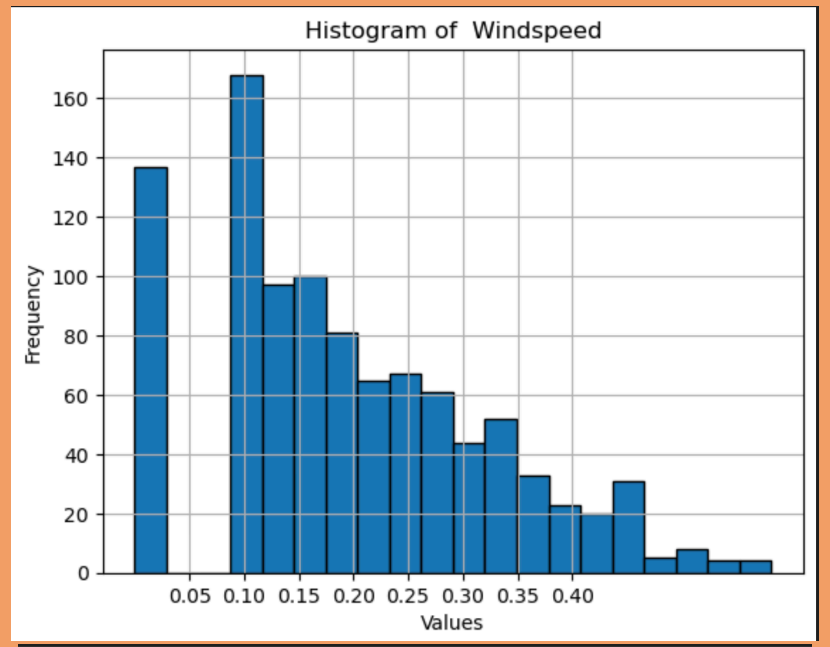
```
...   Tempreature    0.547997
      Hour          -0.063524
      Season         0.000000
      Year           0.000000
      Month          0.812772
      Weekday        0.021925
      Weathersit     1.044098
      Humdity        0.436893
      Windspeed      0.433675
      Casual         3.414105
      Registered     1.296723
      Count          1.137709
      dtype: float64
```

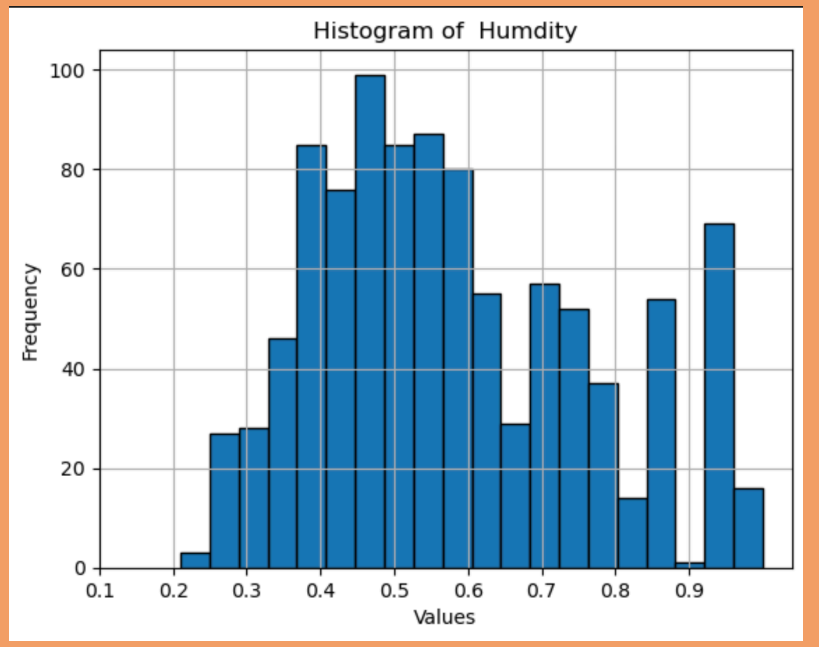
Data summary statistics:

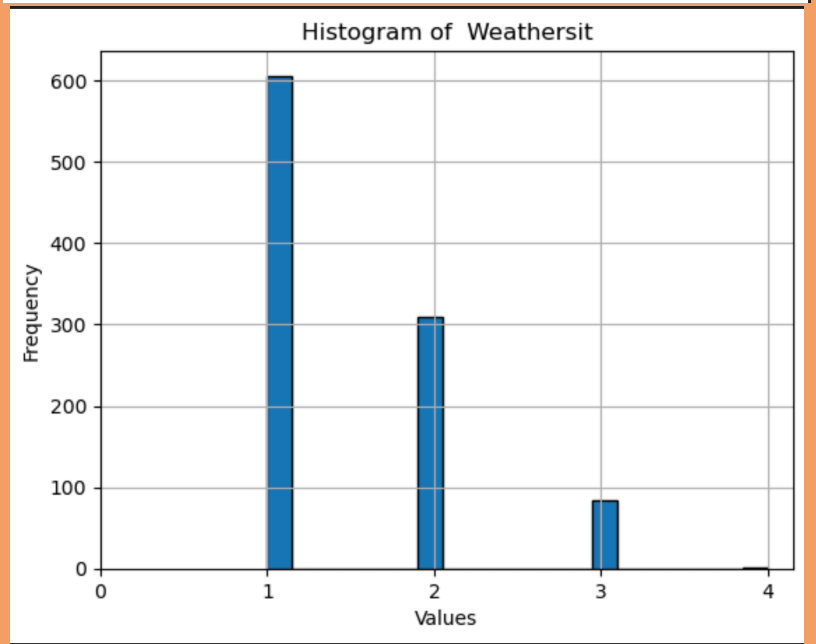
	Index	Season	Year	Month	Hour	Weekday	Weathersit	Tempreature	Humdity	Windspeed	Casual	Registered
count	1000.000000	1000.0	1000.0	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	500.500000	1.0	0.0	1.312000	11.753000	2.982000	1.480000	0.205900	0.582480	0.194931	4.921000	53.383000
std	288.819436	0.0	0.0	0.463542	6.899101	2.091423	0.651171	0.078977	0.187977	0.129126	7.643899	47.893968
min	1.000000	1.0	0.0	1.000000	0.000000	0.000000	1.000000	0.020000	0.210000	0.000000	0.000000	0.000000
25%	250.750000	1.0	0.0	1.000000	6.000000	1.000000	1.000000	0.160000	0.440000	0.104500	0.000000	15.000000
50%	500.500000	1.0	0.0	1.000000	12.000000	3.000000	1.000000	0.200000	0.550000	0.164200	3.000000	46.000000
75%	750.250000	1.0	0.0	2.000000	18.000000	5.000000	2.000000	0.240000	0.700000	0.283600	6.000000	74.000000
max	1000.000000	1.0	0.0	2.000000	23.000000	6.000000	4.000000	0.460000	1.000000	0.582100	62.000000	247.000000

Distribution plots:









the dimensions of the dataset:

```
ds = pd.DataFrame(merge_1)
dimensions = ds.shape
print(f"Dimensions: {dimensions}")
```

✓ 0.0s

```
Dimensions: (1000, 15)
```

The datatype of the dataset:

```
Index            int64
Date            object
Season           int64
Year             int64
Month            int64
Hour             int64
Holiday           bool
Weekday          int64
Weathersit       int64
Tempreature    float64
Humdity        float64
Windspeed      float64
Casual           int64
Registered       int64
Count            int64
dtype: object
```

Main_dataset:

https://docs.google.com/spreadsheets/d/1a9RAV96a8VBjLpXR23L-TdzhxOkrfHRcuY67vSTF5b4/edit?usp=sharing

Result:

- The data is all cleaned now
- The dataset has a total of 1000rows and 15 columns
- The data is now ready to use for inspection and work on it

Conclusion:

- This data wrangling project successfully cleaned, transformed, and prepared the Bike Rental dataset for analysis and modeling
- the data is easy to understand and visualisation

Thank you