NUMBER OF BIKE USERS IN THE CITY PREDICTION

Team 9:

Team Members:

B.Dharani

N.Shailaja

Madhuraju suresh

Borado laxmikanth

B.Harshavardhan Reddy

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0000	3	13	16
1	2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0000	8	32	40
2	3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0000	5	27	32
3	4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0000	3	10	13
4	5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0000	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
17374	17375	2012-12-31	1	1	12	19	0	1	1	2	0.26	0.2576	0.60	0.1642	11	108	119
17375	17376	2012-12-31	1	1	12	20	0	1	1	2	0.26	0.2576	0.60	0.1642	8	81	89
17376	17377	2012-12-31	1	1	12	21	0	1	1	1	0.26	0.2576	0.60	0.1642	7	83	90
17377	17378	2012-12-31	1	1	12	22	0	1	1	1	0.26	0.2727	0.56	0.1343	13	48	61
17378	17379	2012-12-31	1	1	12	23	0	1	1	1	0.26	0.2727	0.65	0.1343	12	37	49

17379 rows × 17 columns

Dataset :

Here, the Bikes sharing data consists of various information related to bike users in different weather conditions which includes season, temp, windspeed etc.it has 17 columns and 17379

Data Pre-processing :

- For machine learning algorithm it is necessary to convert raw data into clean data set which means converting the data set into numeric data

- Here we defined new columns which are(data,season,month,hour,holiday, hour,windspeed,etc)

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 17379 entries, 0 to 17378
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Date        17379 non-null  object
 1   season      17379 non-null  int64
 2   year        17379 non-null  int64
 3   month       17379 non-null  int64
 4   hour        17379 non-null  int64
 5   holiday     17379 non-null  int64
 6   Day         17379 non-null  int64
 7   workingday  17379 non-null  int64
 8   weather     17379 non-null  int64
 9   temp        17379 non-null  float64
 10  atemp       17379 non-null  float64
 11  hum         17379 non-null  float64
 12  windspeed   17379 non-null  float64
 13  casual      17379 non-null  int64
 14  registered  17379 non-null  int64
 15  count       17379 non-null  int64
dtypes: float64(4), int64(11), object(1)
memory usage: 2.3+ MB
```

	season	year	month	hour	holiday	Day	workingday	weather	temp	atemp	hum	windspeed	casual	registered	count
count	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000
mean	2.501640	0.502561	6.537775	11.546752	0.028770	3.003683	0.682721	1.425283	0.496987	0.475775	0.627229	0.190098	35.676218	153.786869	189.463088
std	1.106918	0.500008	3.438776	6.914405	0.167165	2.005771	0.465431	0.639357	0.192556	0.171850	0.192930	0.122340	49.305030	151.357286	181.387599
min	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.020000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	2.000000	0.000000	4.000000	6.000000	0.000000	1.000000	0.000000	1.000000	0.340000	0.333300	0.480000	0.104500	4.000000	34.000000	40.000000
50%	3.000000	1.000000	7.000000	12.000000	0.000000	3.000000	1.000000	1.000000	0.500000	0.484800	0.630000	0.194000	17.000000	115.000000	142.000000
75%	3.000000	1.000000	10.000000	18.000000	0.000000	5.000000	1.000000	2.000000	0.660000	0.621200	0.780000	0.253700	48.000000	220.000000	281.000000
max	4.000000	1.000000	12.000000	23.000000	1.000000	6.000000	1.000000	4.000000	1.000000	1.000000	1.000000	0.850700	367.000000	886.000000	977.000000

DATA CLEANING:

- In this process we are going to find out the null values (missing data)
- Then we find the unique vales which helps in data analysis and pre- processing

```
df.isnull().sum()
```

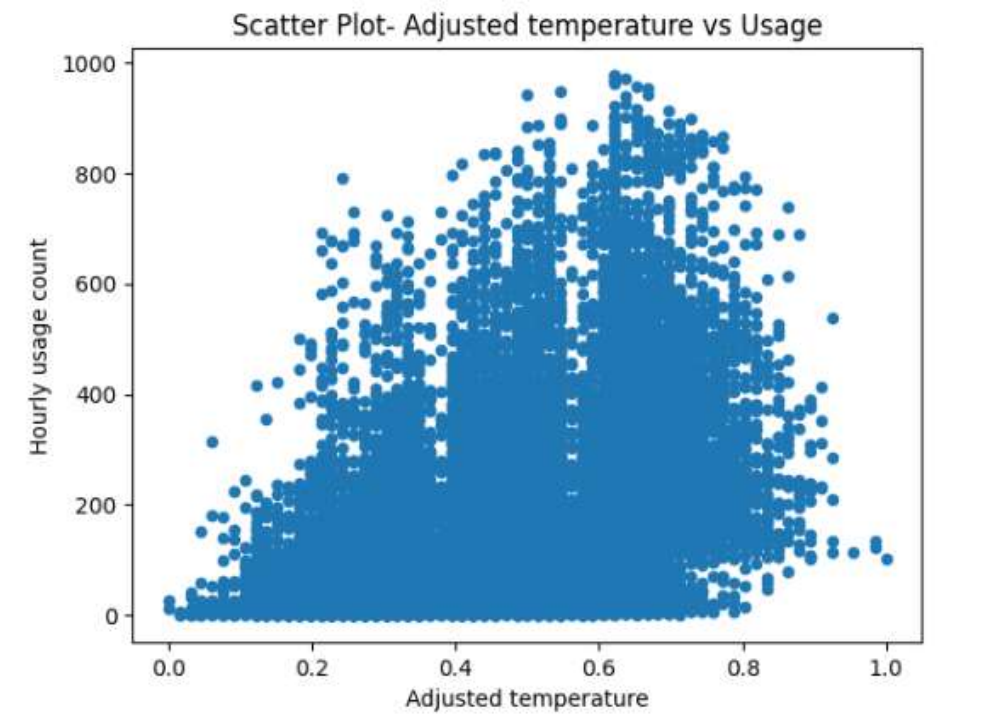
```
Date          0
season        0
year          0
month         0
hour          0
holiday       0
Day           0
workingday    0
weather       0
temp          0
atemp         0
hum           0
windspeed     0
casual        0
registered    0
count         0
dtype: int64
```

```
df.nunique()
```

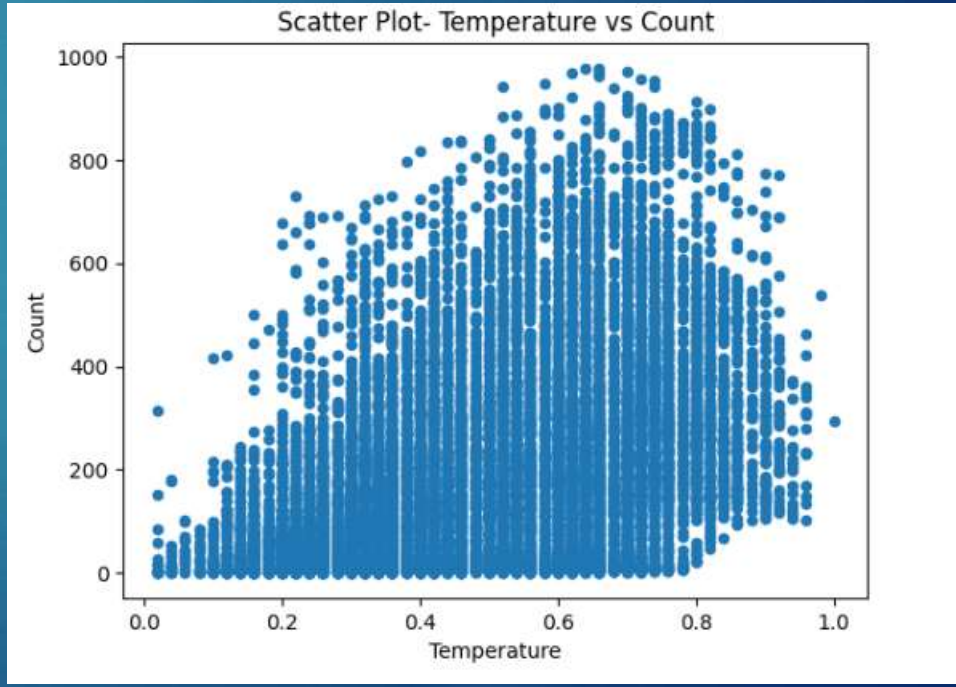
```
Date          731
season          4
year            2
month          12
hour           24
holiday         2
Day             7
workingday      2
weather         4
temp           50
atemp          65
hum            89
windspeed      30
casual        322
registered    776
count         869
dtype: int64
```

Exploratory Data Analysis(EDA):

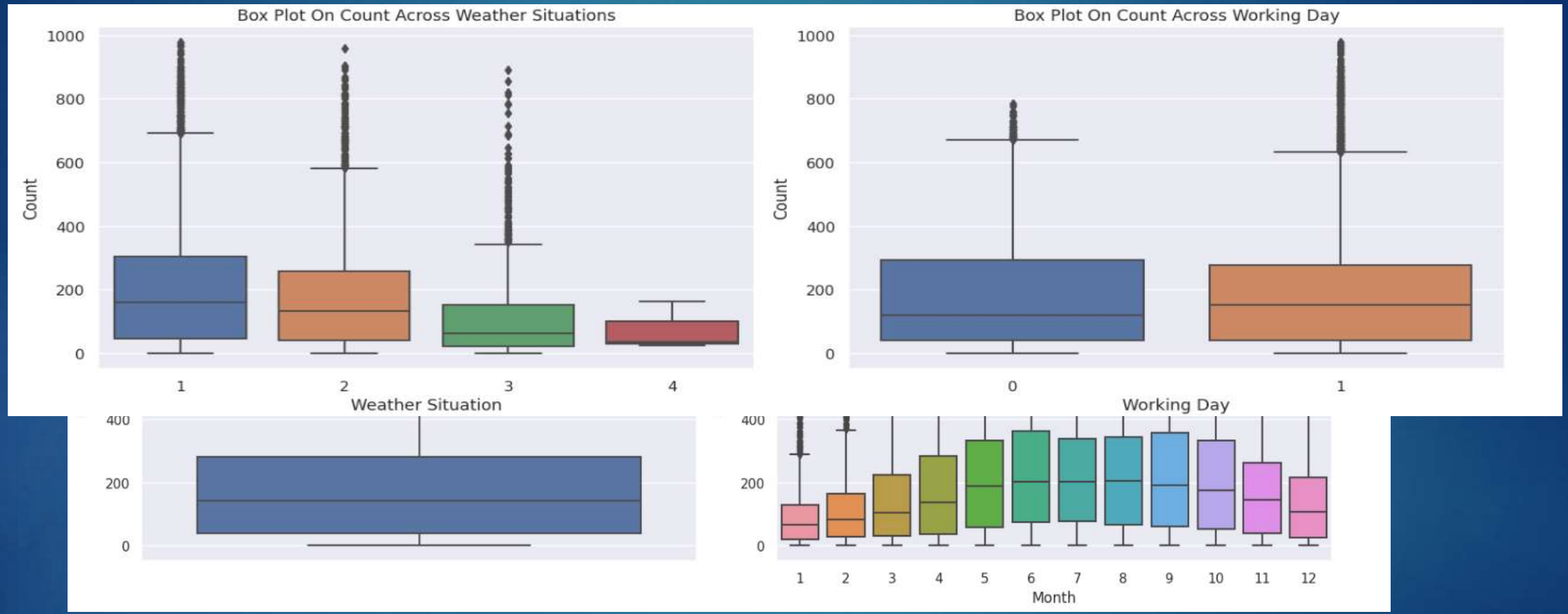
- *Scatter plots:* The below scatter plot shows the information Scatter Plot- Adjusted temperature vs Usage



- The below scatter plot shows the information Scatter Plot- Temperature vs Count



Bar plots:


Box Plot On Count Across Weather Situations
Count
1000
800
600
400
200
0
1
2
3
4
Weather Situation
Box Plot On Count Across Working Day
Count
1000
800
600
400
200
0
0
1
Working Day
400
200
0
1
2
3
4
5
6
7
8
9
10
11
12
Month

Regression:

- Evaluate the Model

Common metrics for regression include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). These metrics measure the goodness of fit and predictive accuracy.

- List Of Regression Models

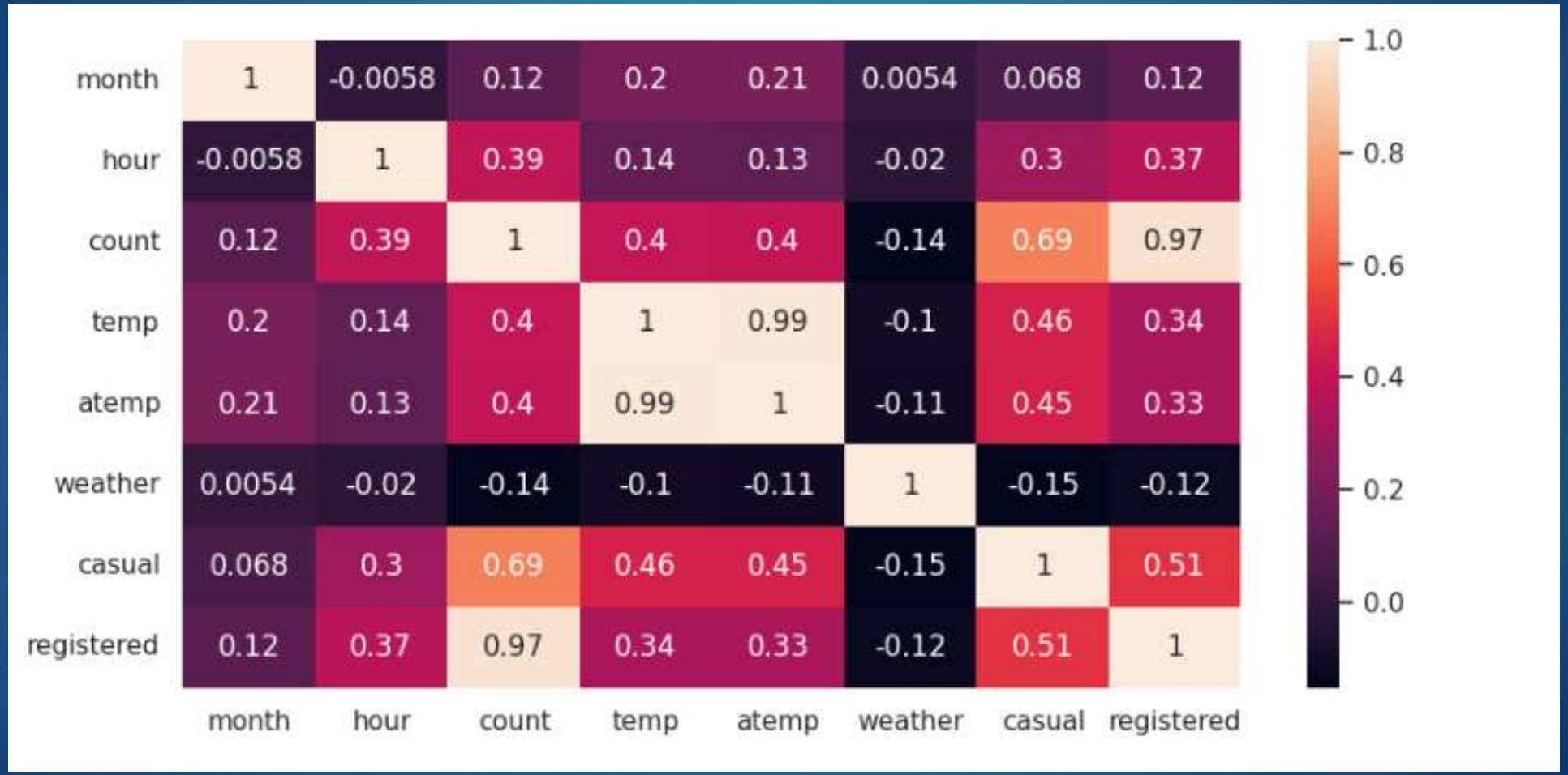
Ridge Regression, Support Vector Regression, Ensemble Regressor, Random Forest Regressor

Model	Mean Squared Error	R² score
SGDRegressor	1448191638580559134976901 12.00	-4254497266532227219456.00
Lasso	0.00	1.00
ElasticNet	0.00	1.00
Ridge	0.00	1.00
SVR	0.01	1.00
SVR	32326.91	0.05
BaggingRegressor	9.54	1.00
BaggingRegressor	1765.56	0.95
NuSVR	31149.68	0.08
RandomForestRegressor	6.45	1.00

Random Forest Model :

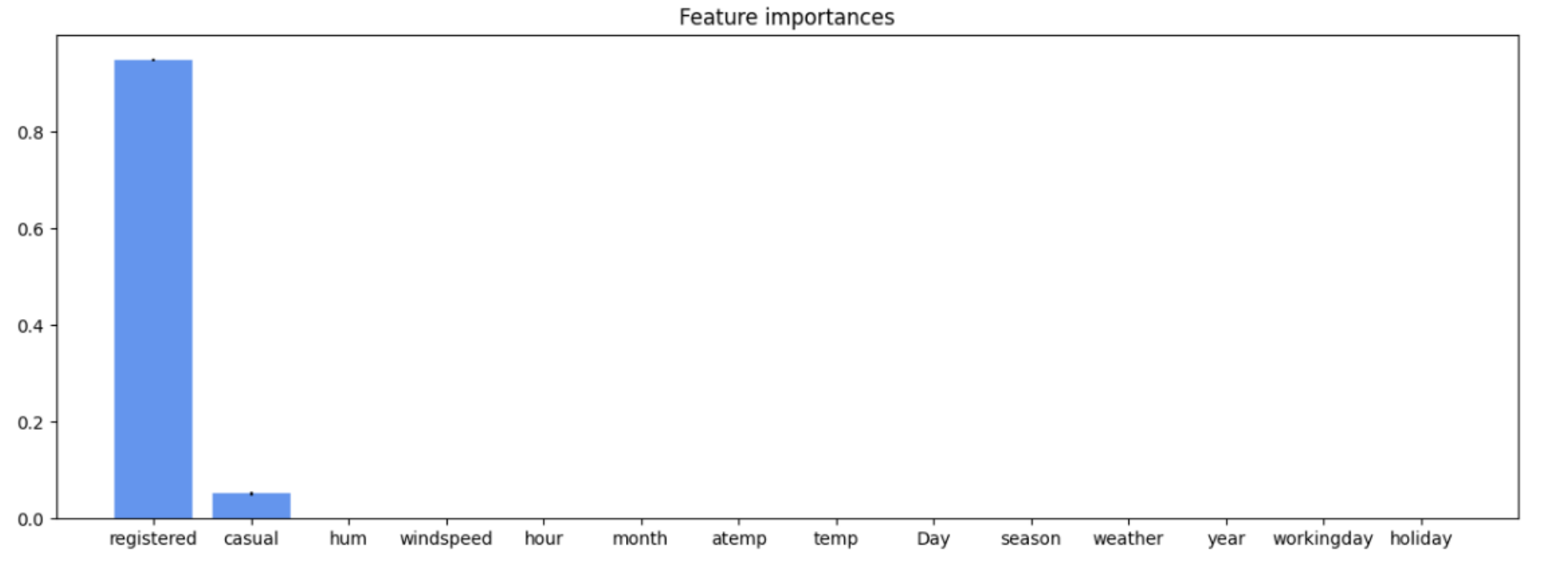
Model	Dataset	MSE	MAE	RMSLE	R^2 score
RandomForestRegressor	training	1.30	0.39	0.00	1.00
RandomForestRegressor	validation	6.61	0.97	0.01	1.00

Heat Map:


month 1 -0.0058 0.12 0.2 0.21 0.0054 0.068 0.12
hour -0.0058 1 0.39 0.14 0.13 -0.02 0.3 0.37
count 0.12 0.39 1 0.4 0.4 -0.14 0.69 0.97
temp 0.2 0.14 0.4 1 0.99 -0.1 0.46 0.34
atemp 0.21 0.13 0.4 0.99 1 -0.11 0.45 0.33
weather 0.0054 -0.02 -0.14 -0.1 -0.11 1 -0.15 -0.12
casual 0.068 0.3 0.69 0.46 0.45 -0.15 1 0.51
registered 0.12 0.37 0.97 0.34 0.33 -0.12 0.51 1
month hour count temp atemp weather casual registered
1.0
0.8
0.6
0.4
0.2
0.0

Feature Importance:

```
Feature ranking:
1. feature registered (0.947873)
2. feature casual (0.051893)
3. feature hum (0.000046)
4. feature windspeed (0.000038)
5. feature hour (0.000032)
6. feature month (0.000031)
7. feature atemp (0.000026)
8. feature temp (0.000025)
9. feature Day (0.000015)
10. feature season (0.000008)
11. feature weather (0.000005)
12. feature year (0.000004)
13. feature workingday (0.000003)
14. feature holiday (0.000001)
```



OBSERVATIONS:

The Value Of Mean Squared Error For Linear regression prediction vs actual is :
144819163858055913497690112.00

The value of Mean squared Error for Lasso regression predicted vs actual is :
144819163858055913497691 58.23589

The result corresponds to the high correlation of the registered and casual usage variable with the bike sharing count in the feature correlation matrix.

INFERNCES & CONCLUSION :

We can conclude that the temperature and other things like weather have the impact in the people riding the bikes

The given features in the dataset have the most influence on the number of bike users.

From the given raw data and the processed data we can understand that the bikers have different interests in riding the bikes

https://colab.research.google.com/drive/1xjd_mxaaPADKwR8rAk4yDe2JDTV1Qix2?usp=sharing