Capstone project on BIKE SHARING DEMAND PREDICTION

By sneha owandkar

CONTENT

- BUSINESS UNDERSTANDING
- DATA SUMMARY
- FEATURE ANALYSIS
- EXPLORATORY DATA ANALYSIS
- DATA PREPROCESSING
- IMPLEMENTING
- ALGORITHMS
- CHALLENGES& CONCLUSIONS

Business Understanding

- Bike rentals have became a popular service in recent years and it seems people are using it more often. With relatively cheaper rates and ease of pick up and drop at own convenience is what making this business thrive. Mostly used by people having no personal vehicles and also to avoid congested public transport which that’s why they prefer rental bikes. Therefore, the business to strive and profit more, it has to be always ready and supply no. of bikes at different locations, to fulfil the demand. Our project goal is a pre planned set of bike count values that can be a handy solution to meet all demands.

Data Summary

Feature Name	Description	Range	Feature Type
instant	Record Index		Numerical - Discrete
dteday	Date	all dates in 2011 & 2012	Numerical - Discrete
season	Season/Climate	(1: Spring, 2: Summer, 3: Fall, 4: Winter)	Categorical- Nominal
Yr	Year	(0: 2011, 1: 2012)	Categorical- Nominal
Hr	Hour	(0,23)	Numerical- Discrete
holiday	Whether the day is a holiday or not	(0: No Holiday, 1: Holiday)	Categorical- Nominal
weekday	Day of the week	(0,6)	Categorical- Nominal
working day	Whether the day is a working day or not	(0: Non-working day, 1: Working Day)	Categorical- Nominal
weathersit	Weather situation	(1,4)	Categorical- Nominal
temp	Normalized temperature in Celsius		Numerical- Continuous
atemp	Normalized feeling-like temperature in Celsius.		Numerical- Continuous
hum	Normalized humidity		Numerical- Continuous
windspeed	Normalized windspeed		Numerical- Continuous
casual	Count of non-registered users		Numerical- Continuous
registered	Count of registered users		Numerical- Continuous
cnt	Count of total rental bikes including both casual and registered		Numerical- Continuous

contains 8760 lines and 14 columns. Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’. One Datetime features ‘Date’. We have some numerical type variables such as temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall which tells the environment conditions at that particular hour of the day.

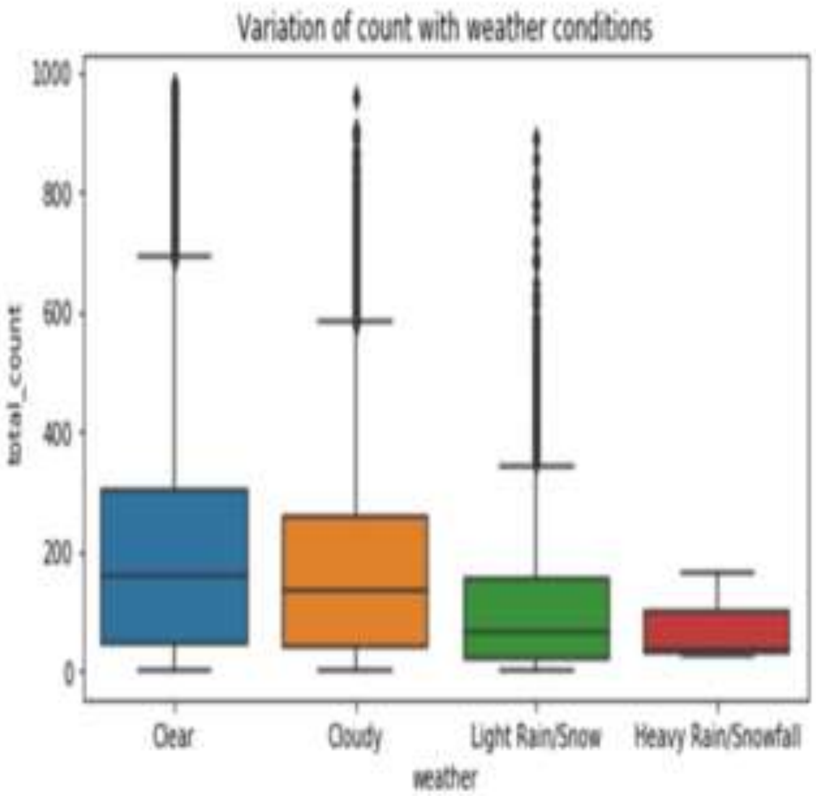
Feature summary

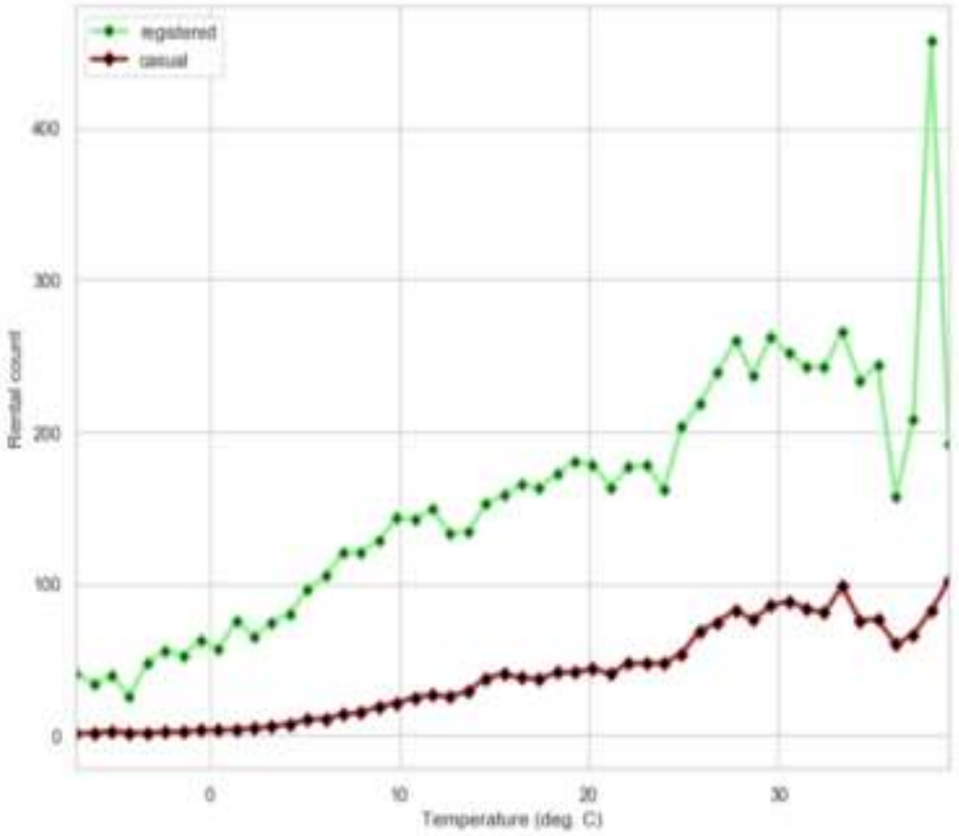
- Date : Year-Month-Day
- Rented Bike Count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature - Temperature in Celsius
- Humidity - % Wind Speed - m/s
- Visibility - 10m Dew point
- temperature -Celsius
- Solar radiation -MJ/m2
- Rainfall -mm
- Snowfall -cm
- Seasons -Winter, Spring, Summer, Autumn
- Holiday -Holiday/No Holiday
- Functional Day - NoFunc(Non Functional Hrs),Fun(Func

Insights from our dataset

- There are No Missing Values present There are No Duplicate values present There are No null values. And finally we have 'rented bike count' variable which we need to predict for new observations The dataset shows hourly rental data for one year (1 December 2017 to 31 November(2018)(365 days).
- we consider this as a single year data So we convert the "date" column into 3 different column i.e "year","month","day".
- We change the name of some features for our convenience , they are as below 'Rented_Bike_Count', 'Hour', 'Temperature', 'Humidity', 'Wind_speed', 'Visibility', ‘Dew_point_temperature', 'Solar_Radiation', 'Rainfall', 'Snowfall', 'Seasons', 'Holiday', 'Functioning_Day', 'month','weekdays_weeken

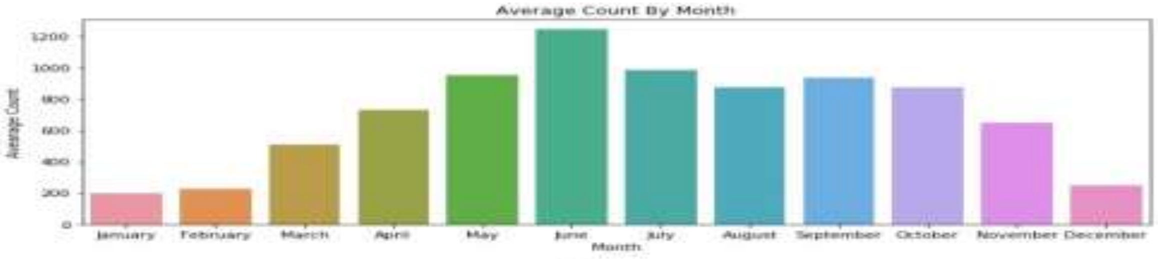
Rental bike count



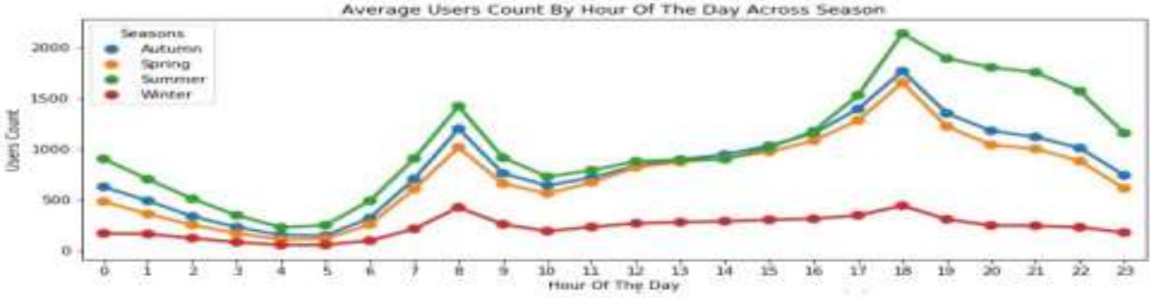


The following plots show the variations of ridership with weather conditions. As we know the ridership is high during the clear climate and decrease as the climate gets worse, this can also be inferred from the graph.

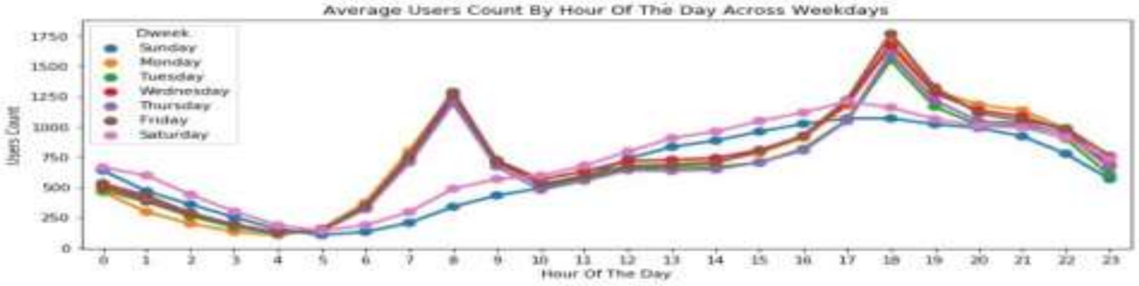
Analysis of month variable



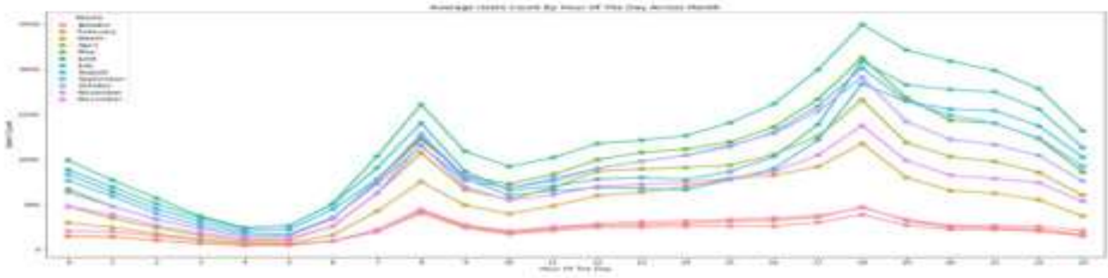
(A)



(B)



(C)



(D)

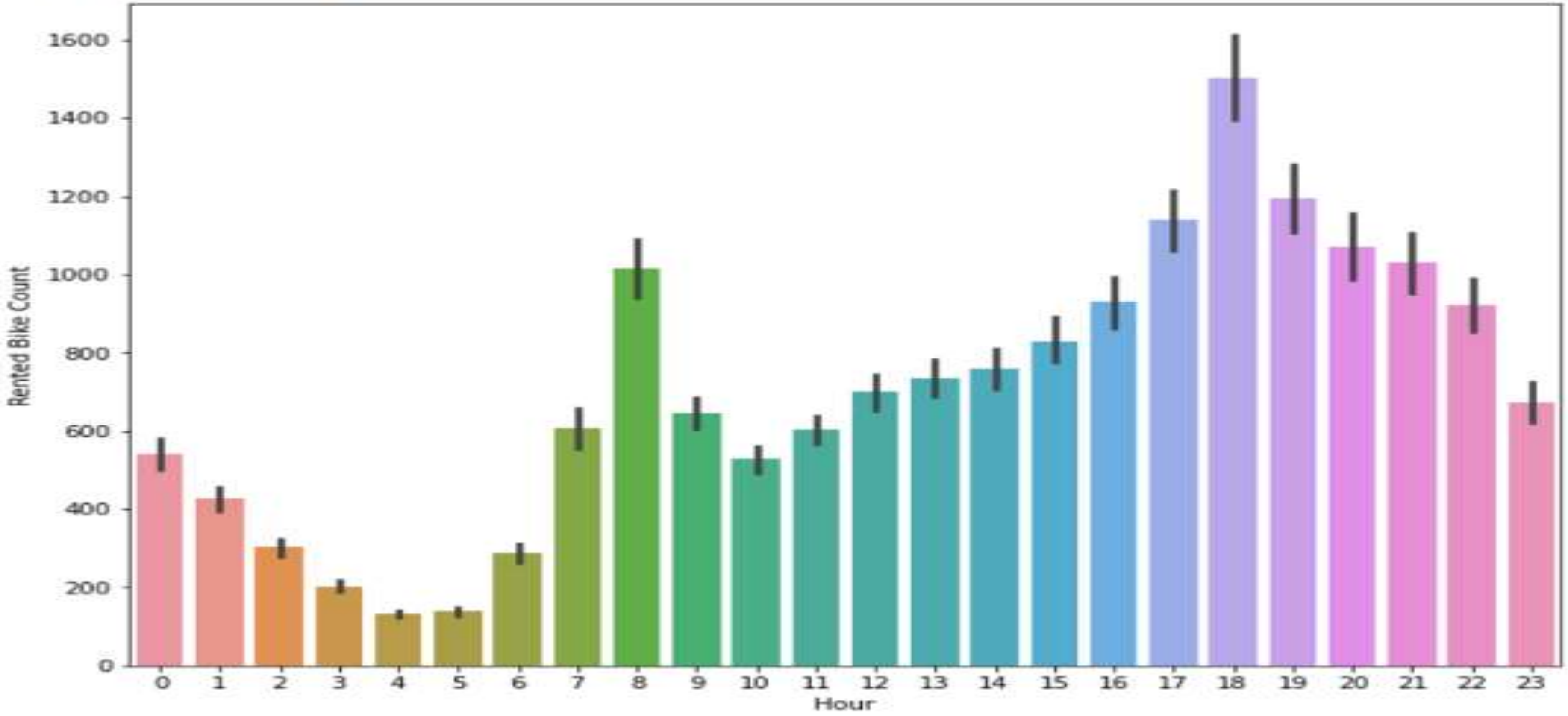
So the constant raise of users necessitates the prediction of the number of rental bikes that were needed to make the bike sharing system to consistently work. Therefore, this research aims to use machine learning and data mining based algorithms to predict required number of rental bikes required at eachour. In this method, data mining is used as it has the reliability to solve complicated issues.

Across various cities, a growing body of research has investigated weather and climate impacts on cycling, usually across combination with several other factors that can affect cycling. The results differ in the degree to which climate influences use. Pucher et al. [3] shows that U.S. cities with relatively high levels of cycling have mild winters and often little rain compared to the extreme heat and moisture that disrupts cycling.

Rented bike count per hour

```
#Rented Bike Count per Hour using barplot
plt.figure(figsize=(10,8))
sns.barplot(x = data.Hour , y=data['Rented Bike Count'])
```

<matplotlib.axes._subplots.AxesSubplot at 0x7fb96f957c90>

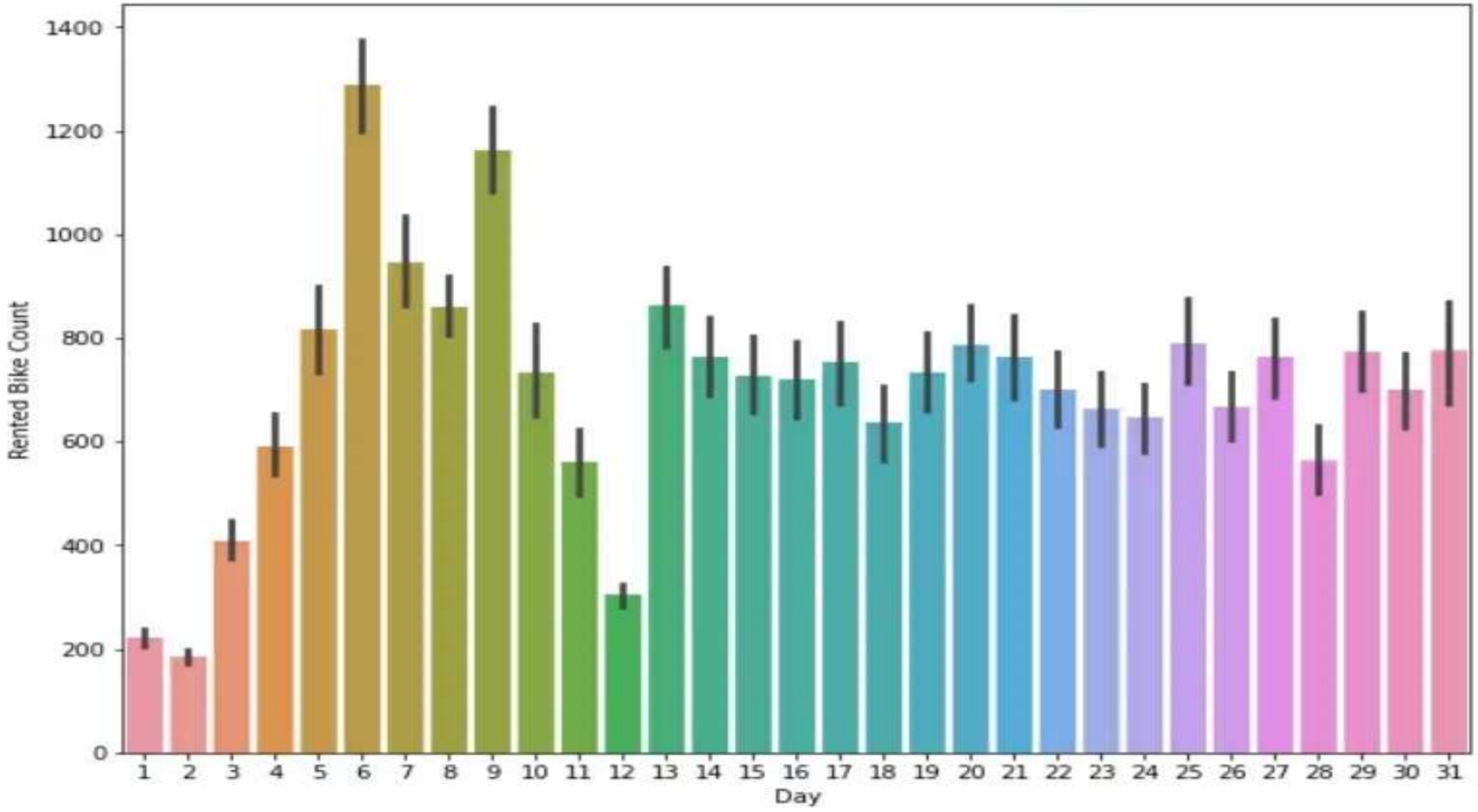


The above plot shows that in hour 17 , 18 & 19 have higher Rented Bike demand.

Rented bike count per day

```
#Rented Bike Count per Day using barplot
plt.figure(figsize=(10,8))
sns.barplot(x = data.Day , y=data['Rented Bike Count'])
```

<matplotlib.axes._subplots.AxesSubplot at 0x7fb96f883710>

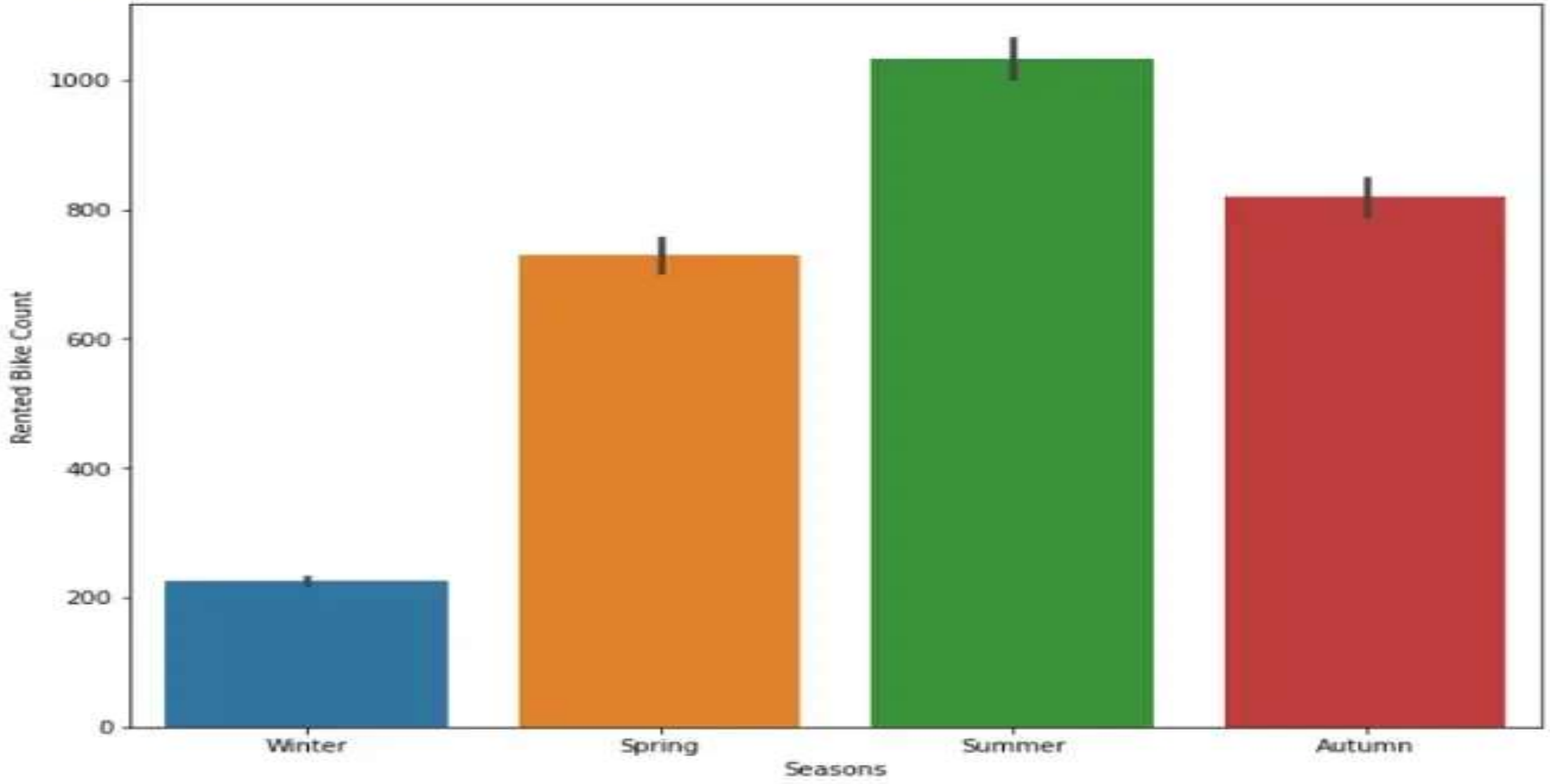


The above plot shows that in day 6 , 7 & 9 have higher Rented Bike demand .

Rented bike count per seasons

```
#Rented Bike Count per Seasons using barplot
plt.figure(figsize=(10,8))
sns.barplot(x = data.Seasons , y=data['Rented Bike Count'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb96f478450>
```



The above plot shows that in Summer, Autumn and, Spring seasons have Higher Rented bike demand.

Correlation matrix

We use the correlation matrix for numerical data.

We observe a highly positive correlation between 'temp' and 'atemp' and between 'casual' and 'registered'.

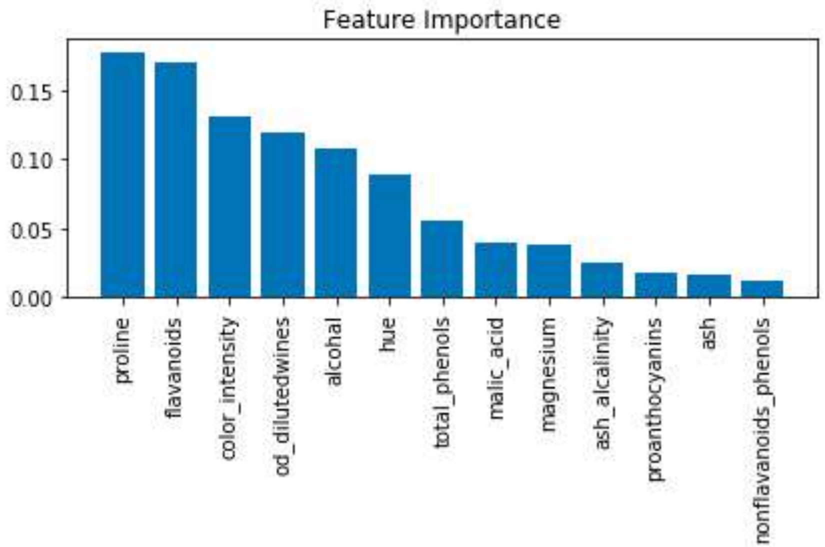
'Windspeed' displays an insignificant contribution to the count.

Hence, we will drop a few unnecessary column

Model building

- LINEAR REGRESSION
- LASSO REGRESSION
- RIDGE REGRESSION
- DECISION TREES REGRESSOR
- RANDOM FOREST REGRESSOR
- GRADIENT BOOSTED REGRESSOR
- GRADIENT BOOSTING REGRESSOR WITH GRIDSEARCHCV

```
In [148]: plt.title('Feature Importance')
          plt.bar(range(X_train.shape[1]), importances[sorted_indices], align='center')
          plt.xticks(range(X_train.shape[1]), X_train.columns[sorted_indices], rotation=90)
          plt.tight_layout()
          plt.show()
```



the selected feature is used to make decision how to divide the data set into two separate sets with similars responses within

CHALLENGES

- Large Dataset to handle.
- Needs to plot lot of Graphs to analyse.
- Feature engineering
- Feature selection
- Optimising the model Carefully
- tuned Hyperparameters as it affects the R2 score.

CONCLUSION

- ‘Hour’ of the day holds the most important feature.
- Bike rental count is mostly correlated with the time of the day as it is peak at 10 am morning and 8 pm at evening. We observed that bike rental count is high during working days than non working day.
- We see that people generally prefer to bike at moderate to high temperatures, and when little windy It is observed that highest number bike rentals counts in Autumn & Summer seasons & the lowest in winter season.
- We observed that the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day. We observed that with increasing humidity, the number of bike rental counts decreases.

THANKU