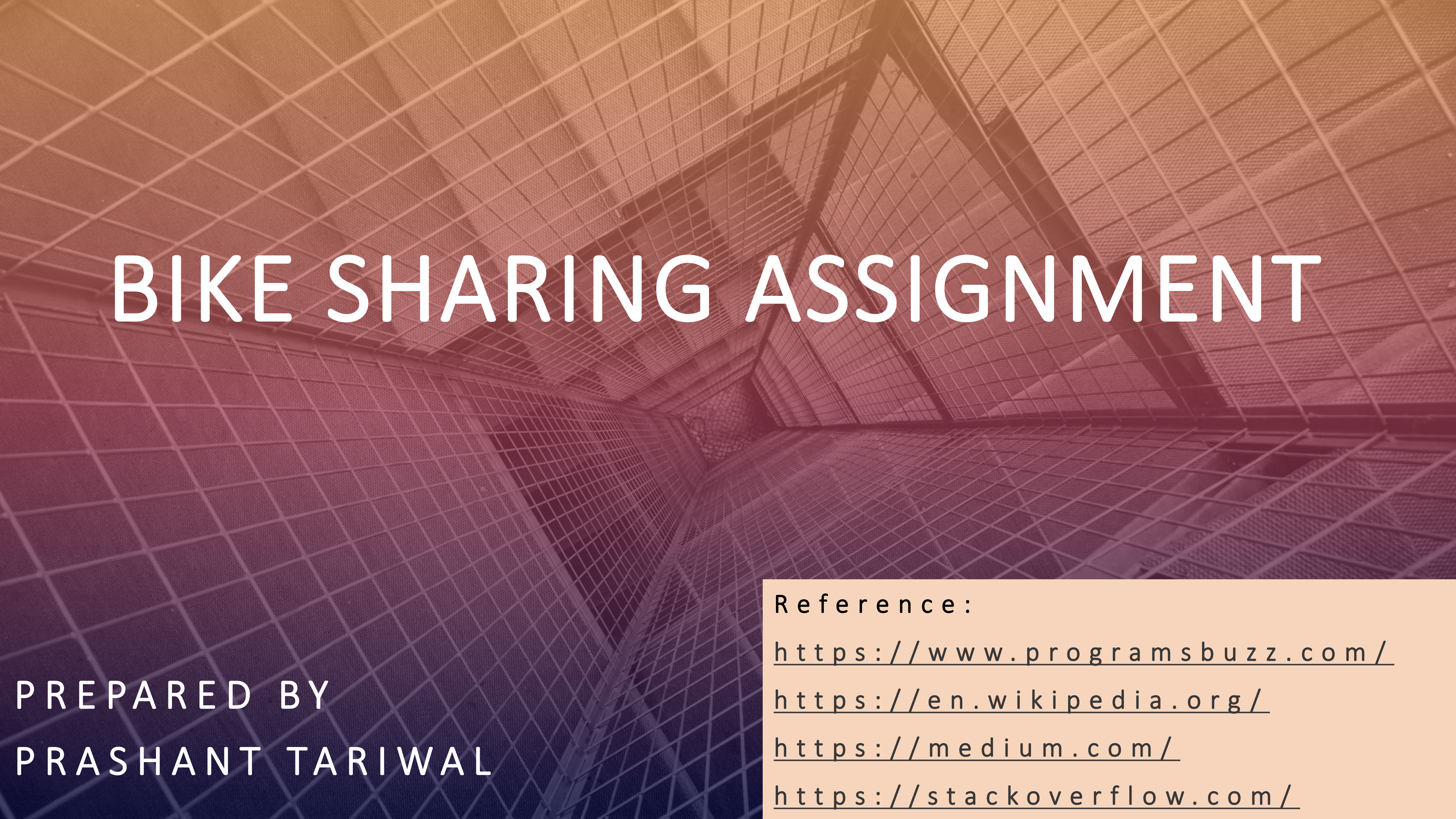


BIKE SHARING ASSIGNMENT

PREPARED BY

PRASHANT TARIWAL

Reference:

https://www.programsbuzz.com/

https://en.wikipedia.org/

https://medium.com/

https://stackoverflow.com/

OBJECTIVE

Understand the factors on which the demand for these shared bikes depends. Specifically, they want to understand the factors affecting the demand for these shared bikes in the American market. The company wants to know:

1.To understand which variables are significant in predicting the demand for shared bikes.
2.To understand how well those variables describe the bike demands.
3.To model the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features. They can accordingly manipulate the business strategy to meet the demand levels and meet the customer's expectations. Further, the model will be a good way for management to understand the demand dynamics of a new market.

CONCLUSION

As per our final Model, the below predictor variables influences bike booking :

Temp (Temperature), Windspeed, Season_Summer, Season_Winter, Year_2019, Month_Sept
Weekday_Sunday, Workingday_Yes, Weather_Good/Clear, Weather_Moderate/Misty

DATASET CHARACTERISTICS - DAY.CSV HAVE THE FOLLOWING FIELDS:

- instant: record index
- dteday : date
- season : season (1:spring, 2:summer, 3:fall, 4:winter)
- yr : year (0: 2018, 1:2019)
- mnth : month (1 to 12)
- holiday : weather day is a holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.

+ weathersit :

  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

- temp : temperature in Celsius
- atemp: feeling temperature in Celsius
- hum: humidity
- windspeed: wind speed
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

STEPS FOLLOWED BY CASE STUDY

Step 1: Data Understanding

Understanding the every column and its significant in the analysis.

Step 2: Data Cleaning

Removing the ‘NaN’ columns, unwanted rows and columns after understanding the importance of that, resolve issues of missing values and standardization of data for easy and effective analysis.

Step 3: Univariate Analysis

By using the every column analyzing the risk factors of the Charged off % and fully paid%.

Step 4: Segmented Univariate Analysis

By using single and multi column we are going to analyzing the risk factors of the Charged off % and fully paid% by clustering the data. Ex : Purpose of loan, Month, occupation, Company grade and state wise.

Step 5: Bivariate Analysis

By using multi column we are going to analyzing the risk factors of the Charged off % and fully paid% by clustering the data. Ex : Purpose of loan, Month, occupation, Company grade and state wise.

Step 6: Recommendations/Results

To reduce the risk factor and the importance of the business we are going to give recommendations and explain the result

DATA CLEANING STEPS

- Delete columns : Delete unnecessary columns.
- Remove outliers : Remove high and low values that would disproportionately affect the results of our analysis.
- Missing values : Treat missing values with appropriate approach.
- Duplicate data : Remove identical rows, remove rows where some columns are identical.
- Filter rows : Filter by segment to get only the rows relevant to the analysis.

ANALYSIS

Need to build model and residual analysis and have made predictions on the test set, just make sure you use the following two lines of code to calculate the R-squared score on the test set.

```
from sklearn.metrics import r2_score r2_score(y_test, y_pred)
```

where y_test is the test data set for the target variable, and y_pred is the variable containing the predicted values of the target variable on the test set

READING AND UNDERSTAND THE DATA

- Importing Numpy and Pandas libraries
- To suppress warnings
- Reading the day.csv file and assigning data frame as bike_sharing
- Shape of the dataframe.
- Information of the dataframe.
- Describe of the dataframe.

Insights:

1. 730 rows and 16 columns.
2. from data dictionary we understood that:
   - dteday is date column and we already have month and year columns in dataframe. thus we can drop dteday column as it not bringing any new information.
   - Addition of casual and registered columns gives cnt column, and we will not going to get much information from these columns, thus we will drop these two as well.
3. There are no missing values from non-null value count.
4. TARGET variable/column is cnt.

Checking all the column names and renaming

EXPLORATORY DATA ANALYSIS.

Data Visualization : **Categorical Variables**

Import Matplot and Seaborn libraries for visualization purpose.

Creating new datafrme by removing unnecessary columns (dteday, casual and registered)

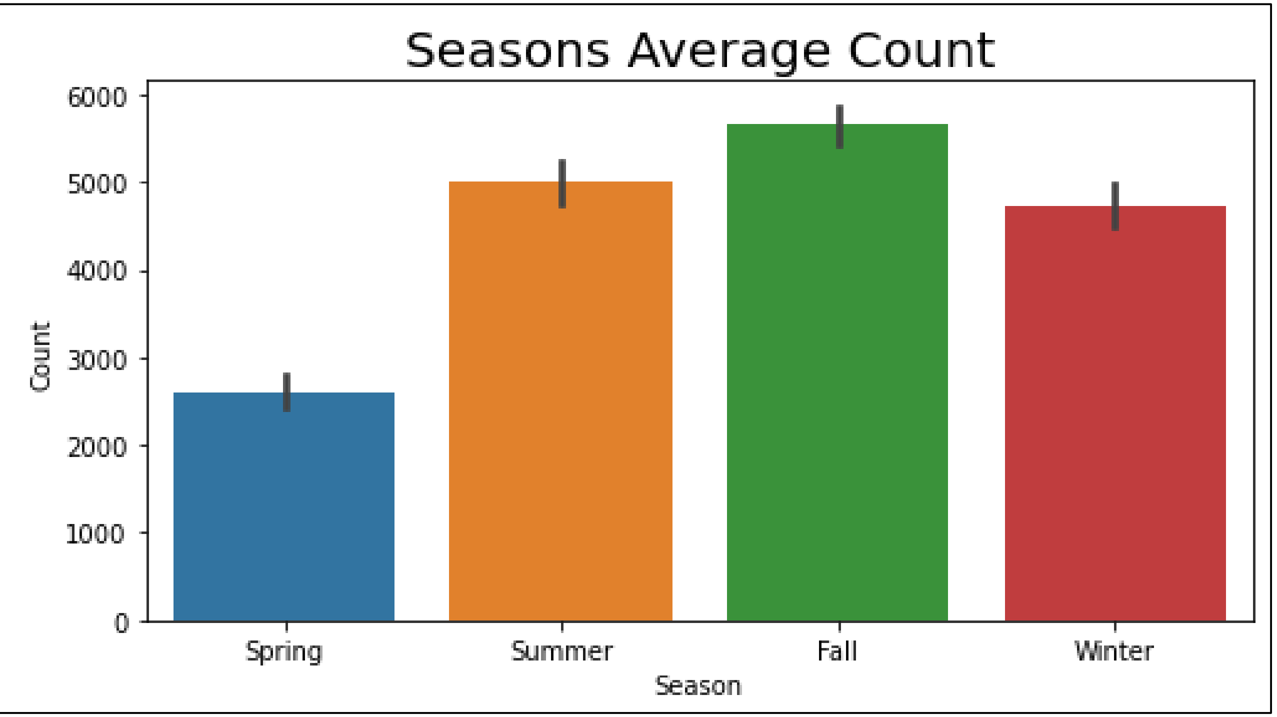
As observed, all the columns are of numerical type, but from the data dictionary we see that there are some columns which represents categorical data as well.

1. Binary Types : Year, Holiday, Workingday
2. Categorical Types: Season, Month, Weekday, Weather
3. Numeric types: Temp, Atemp, Humidity, Windspeed, Count

Data Visualization

EXPLORATORY DATA ANALYSIS.

Data Visualization : Count of rented bike correlation with different seasons.

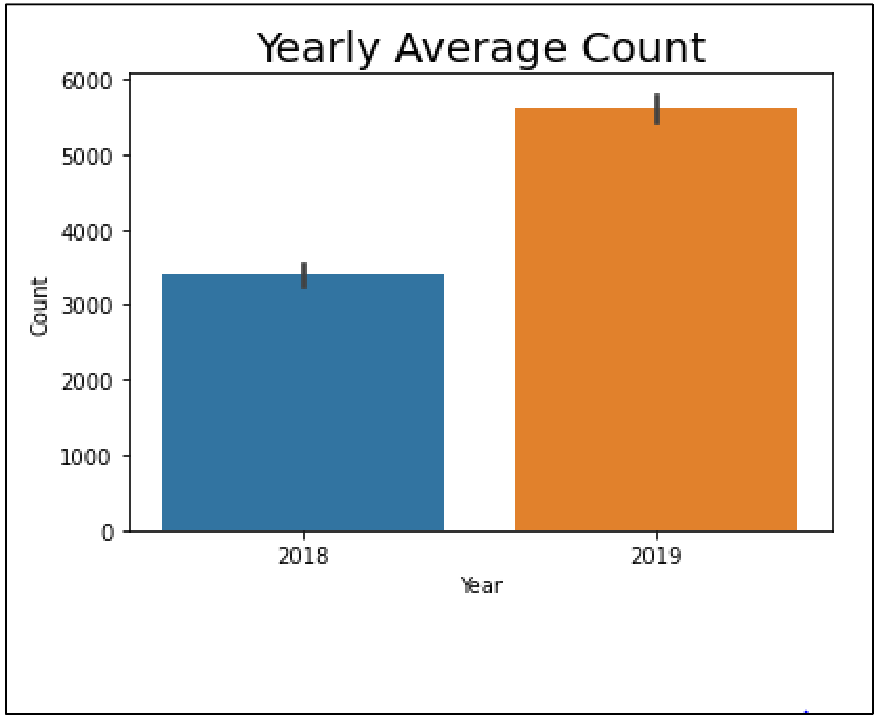


Insights:

1.In Fall highest demand of rented the bikes, followed by Summer and Winter

2.Spring least season where people uses of rent bikes

Data Visualization : Count of rented bike correlation with years.

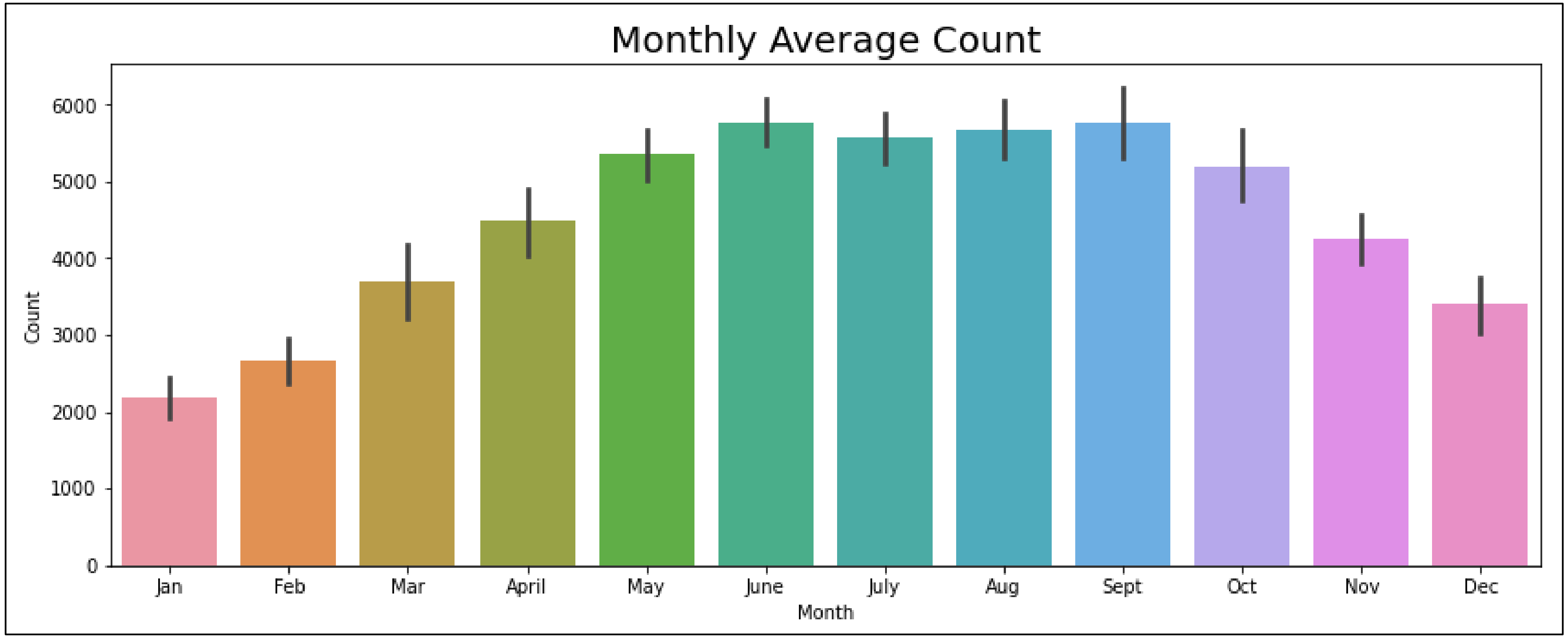


Insights:

We can observe here, average rented bikes has increased in 2019 compare to 2018

EXPLORATORY DATA ANALYSIS.

Data Visualization : Count of rented bike correlation with monthly average count.

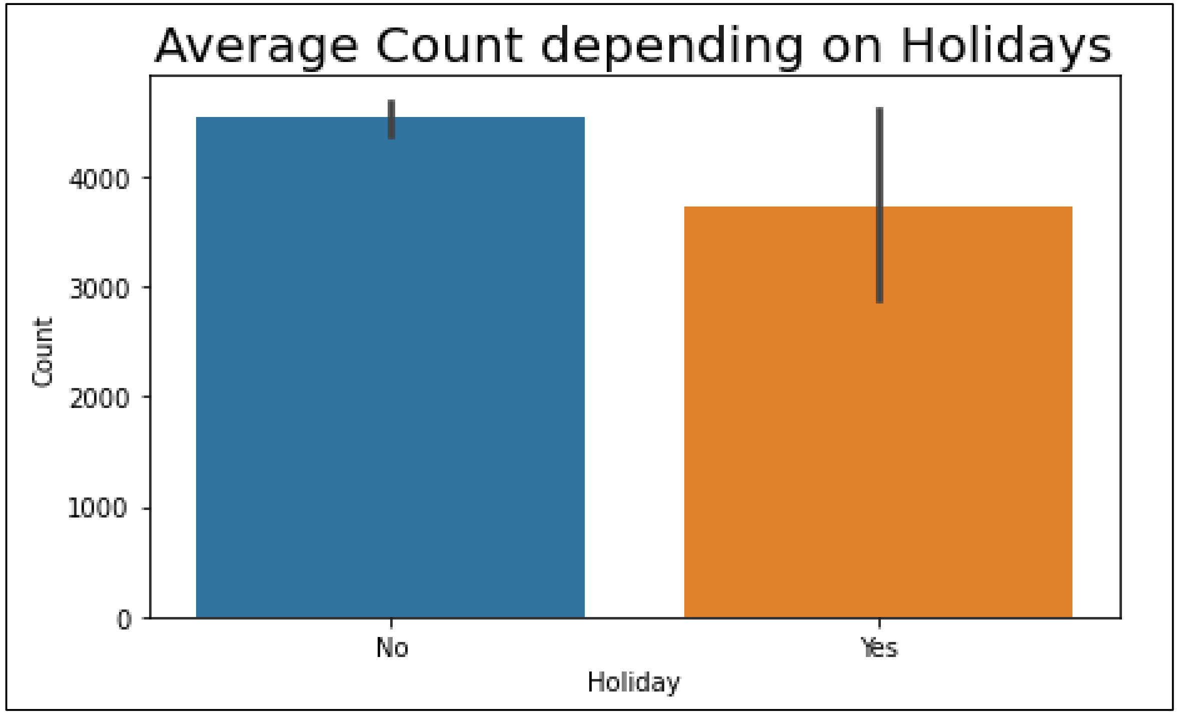


Insights:

1. Almost similar average count of rented bikes in June, September, August, July followed by May, October. Company should make sure they prepared for supply according to the demand of the bikes.
2. December, January, February have the least demand probably due to winter season

EXPLORATORY DATA ANALYSIS.

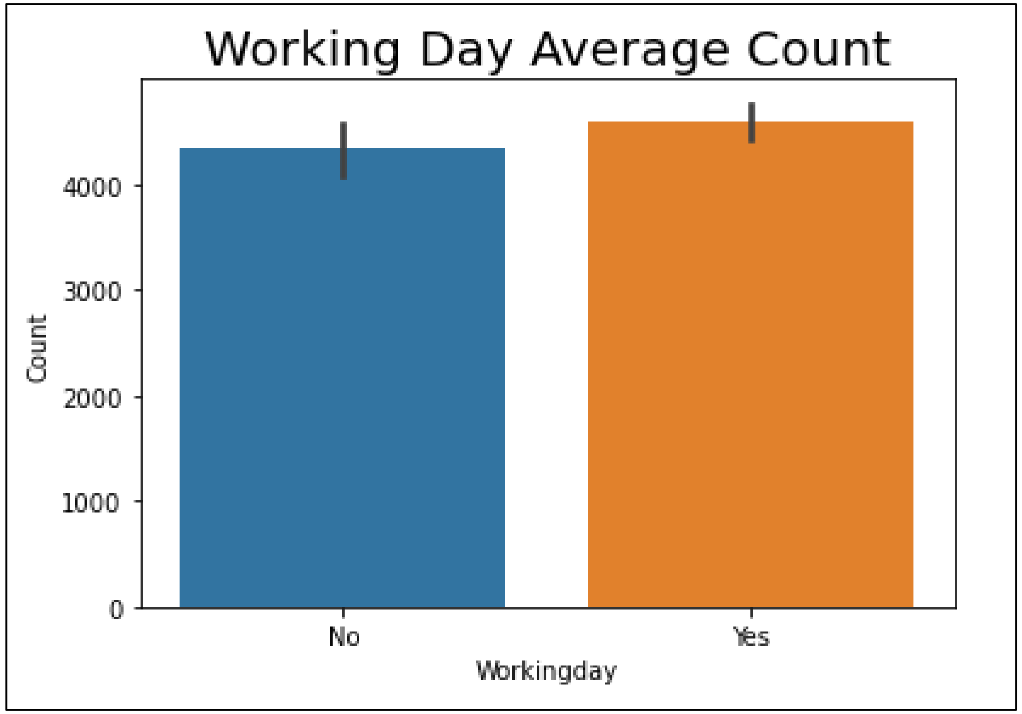
Data Visualization : Count of rented bike correlation with holidays



Insights:

There is high decrease of demand if it is a holiday may be traveling to other cities or prefer own vehicle and public transports.

Data Visualization : Count of rented bike correlation with workingday

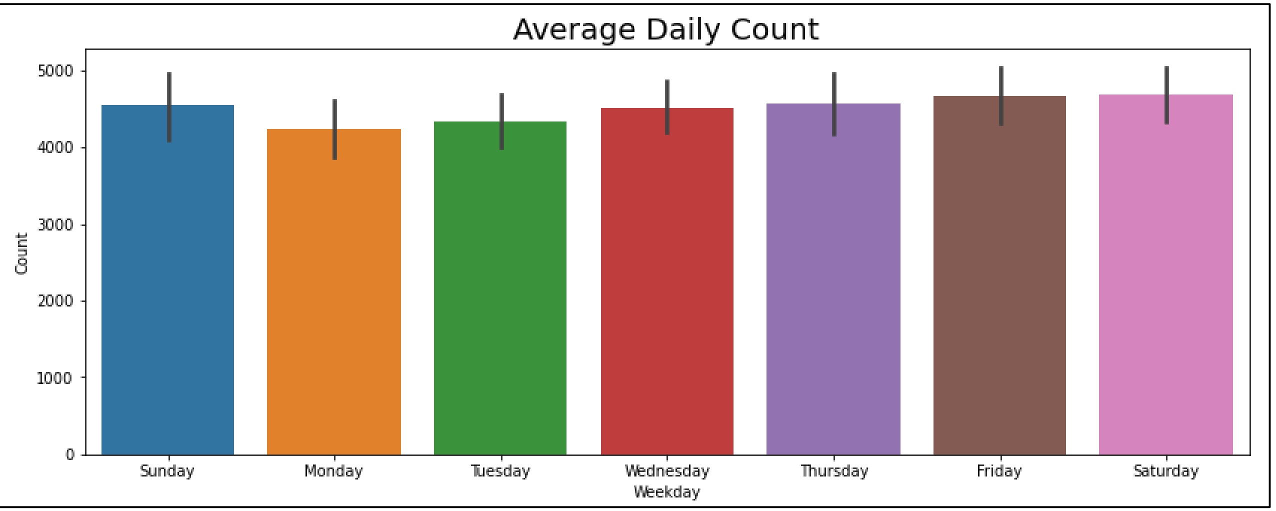


Insights:

There are similar demand during working day and non working day.

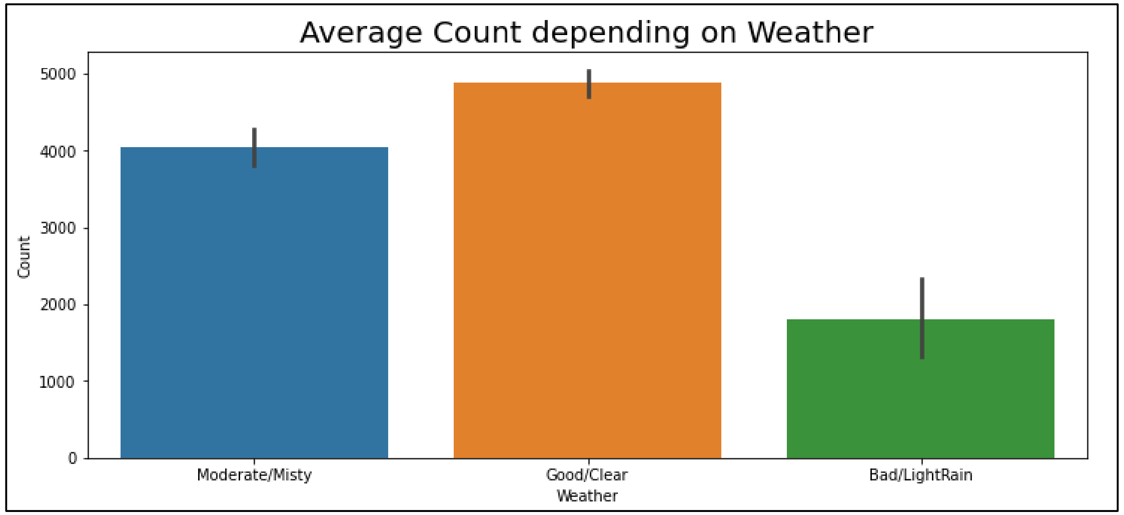
EXPLORATORY DATA ANALYSIS.

Data Visualization : Count of rented bike correlation with weekdays



Insights:
looks like all days have similar demands, but still Sunday, Thursday, Friday and Saturday has high demands than other days People less prefer rented bike on Monday, Tuesday and Wednesday.

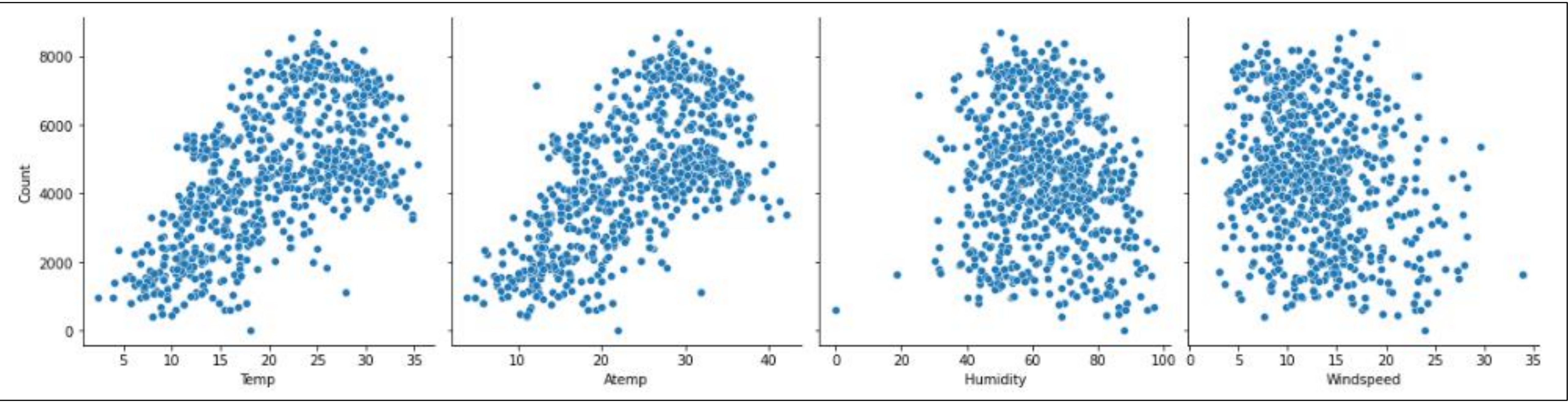
Data Visualization : Count of rented bike correlation with weather



Insights:
It clearly shows that if the weather is clear, the demand is more.
If the weather is bad, demand decreases.
Company has to look for forecast of weather to fullfill demands.

EXPLORATORY DATA ANALYSIS.

Data Visualization : Count of rented bike correlation with Numerical variables.



Insight:

As can be seen from above plots, there is some linear relation between temp, atemp with Count.

This shows that we can do linear regression for solving the problem.

Independent variables which could be a good predictor from EDA are:

Temp, Weather, Months, Seasons, Workingday

DATA PREPARATION

1. Creating a list and putting all category columns in to it and converting them to category data type
2. For Linear model creating dummies
3. Combining both the dataframe, bike_sharing_df and of dummy variables
4. Dropping columns from which dummy variables were created
5. Importing statsmodel and sklearn libraries for Linear regression model building
6. Splitting the date into two train and test dataframes
7. Verify the columns and rows

RESCALING THE FEATURES

1. importing MinMax scaler from preprocessing module of sklearn library.
2. Defining a variable scaler for minmax scaling.
3. Performing scaling on all the numerical variables of train dataset and leaving Count variable aside.
4. Checking all columns and all the variables after scaling.
5. Let's check the correlation coefficients to see which variables are highly correlated.

MODEL BUILDING

1. Dividing training set into X_train and y_train sets for the model building.
2. Importing RFE library for feature selection and after this will perform manual feature selection.
3. Using RFE for feature selection and limiting to selection to 15 features.
4. Creating a list of features selected by RFE.
5. Feature which are choose by RFE during feature selection.(so un-supported columns)
6. Creating new train dataframe with RFE selected features.

MODEL COMPARISON AND INSIGHTS

Model No (Removing)	R-squared	Adj. R-squared	P>|t| (Max)	VIF (Max)	Insights
Model_1	0.851	0.847	0.127	25.15	Month_Nov high P-valve and Humidity high VIF
Model_2 (Month_Nov)	0.850	0.846	< 0.05	25.11	Humidity due very high VIF values
Model_3 (Humidity)	0.845	0.841	0.005	14.86	Weather_Good/Clear due very high VIF values as all the p-values < 0.05 but R-squared is reducing more than 5% so Month_july is having more p-valve removing that.
Model_4 (Month_July)	0.842	0.839	0.004	14.82	removing feature - Weather_Good/Clear due very high VIF values as all the p-values < 0.05 but R-squared is reducing more than 5% so Season_Spring is having more p-valve removing that.
Model_5 (Season_Spring)	0.840	0.837	0.004	8.38	removing feature - Holiday_Yes is having more p-valve removing that.
Model_6	0.837	0.834	0.000	8.01	This model looks good, as there seems to be VERY LOW Multicollinearity between the predictors and the p-values for all the predictors seems to be significant. For now, we will consider this as our final model

FINAL MODEL INTERPRETATION

Hypothesis Testing:

Hypothesis testing states that:

H0:B1=B2=   =Bn=0

H1: at least one Bi!=0

Model 6 coefficient values

const = -0.2239, Temp = 0.5601, Windspeed = -0.1521, Season_Summer = 0.0912, Season_Winter = 0.1375
Year_2019 = 0.2306, Month_Sept = 0.0959, Weekday_Sunday = 0.0697, Workingday_Yes = 0.0591
Weather_Good/Clear = 0.2890, Weather_Moderate/Misty = 0.2082

INSIGHT: From the Model 6 model summary, it is evident that all our coefficients are not equal to zero. which means We REJECT the NULL HYPOTHESIS

F Statistics : F-Statistics is used for testing the overall significance of the Model: Higher the F-Statistics, more significant Model is

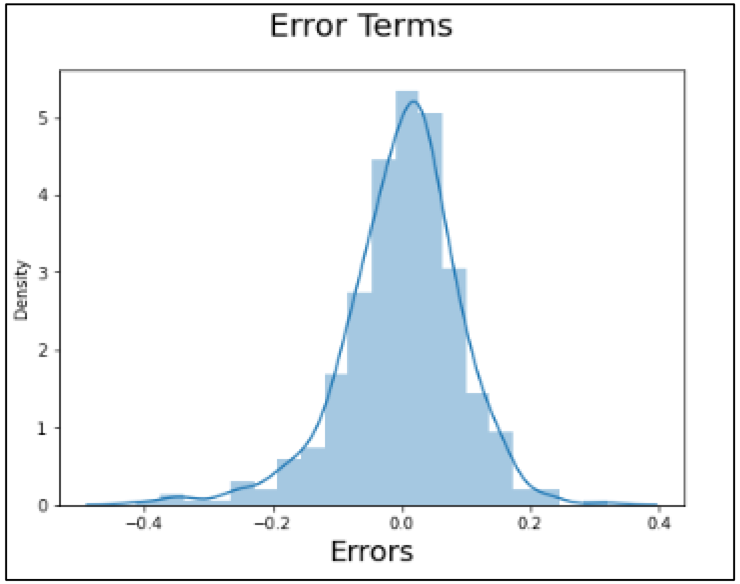
F-statistic: 276.0

Prob (F-statistic): 4.26e-204

The F-Statistics value of 276 (which is greater than 1) and the p-value of '~0.0000' states that the overall model is significant

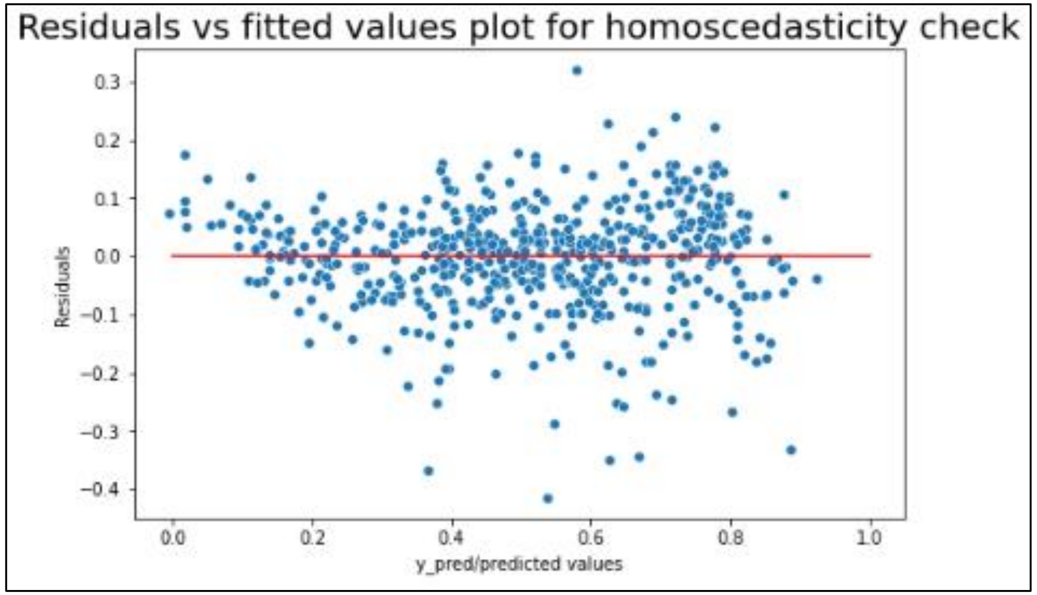
VALIDATE ASSUMPTIONS

Error terms are normally distributed with mean zero (not X, Y)
Residual Analysis Of Training Data



INSIGHT: - From the above histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.

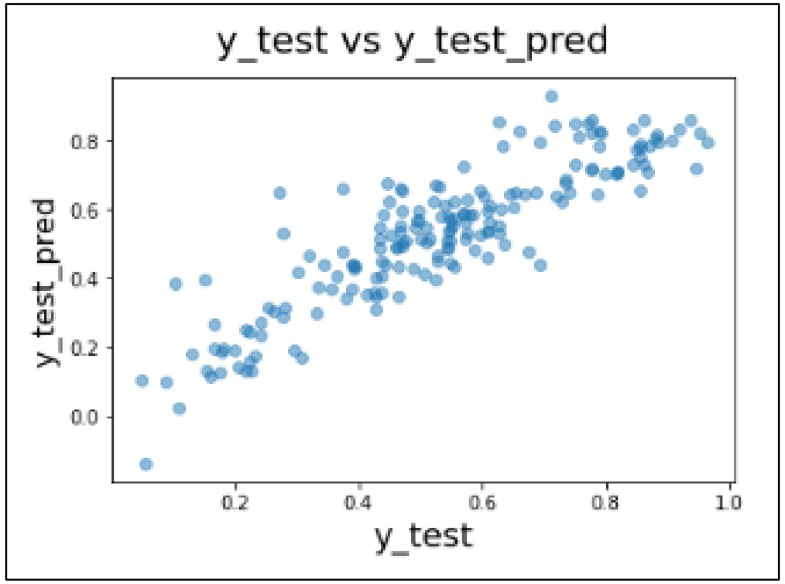
Check for Homoscedasticity



INSIGHT: - From the above plot, we can see that residuals have equal or almost equal variance across the regression line.

MODEL EVALUATION

y_test and y_test_pred

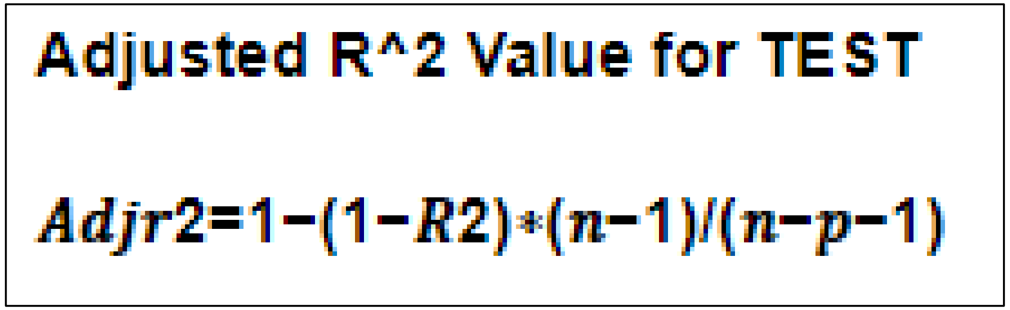


There is linear relationship between y_test and y_test_pred

Residual Analysis

Test data r^2 : 77.79
Train data r^2 : 83.74

Test data adjusted r^2 : 76.5
Train data adjusted r^2 : 83.44



Adjusted R^2 Value for TEST

$$Adjr2=1-(1-R2)*(n-1)/(n-p-1)$$

This seems to be a really good model that can very well 'Generalize' various datasets.

LINEAR REGRESSION SUBJECTIVE QUESTIONS

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Seasons insight:

1. In Fall highest demand of rented the bikes, followed by Summer and Winter
2. Spring least season where people uses of rent bikes

Years insight:

1. We can observe here, average rented bikes has increased in 2019 compare to 2018

Months insight:

1. Almost similar average count of rented bikes in June, September, August, July followed by May, October. Company should make sure they prepared for supply according to the demand of the bikes.
2. December, January, February have the least demand probably due to winter season

Holidays insight:

1. There is high decrease of demand if it is a holiday may be traveling to other cities or prefer own vehicle and public transports.

Workingday insight:

1. There are similar demand during working day and non working day.

Weather insight:

1. **It clearly shows that if the weather is clear, the demand is more.**
2. **If the weather is bad, demand decreases.**
3. **Company has to look for forecast of weather to full fill demands.**

2. Why is it important to use **drop_first=True** during dummy variable creation?

- Creating dummies which converts categorical data into a form which is understandable by ml model.
- It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- In pair-plots, there is some linear relation between Temp, Atemp with Count.
- This shows that we can do linear regression for solving the problem.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- Error terms are normally distributed with mean zero (not X, Y)
- Residual Analysis Of Training Data.
    - ➢ From histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.
- Check for Homoscedasticity
    - ➢ From homoscedasticity plot, we can see that residuals have equal or almost equal variance across the regression line.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

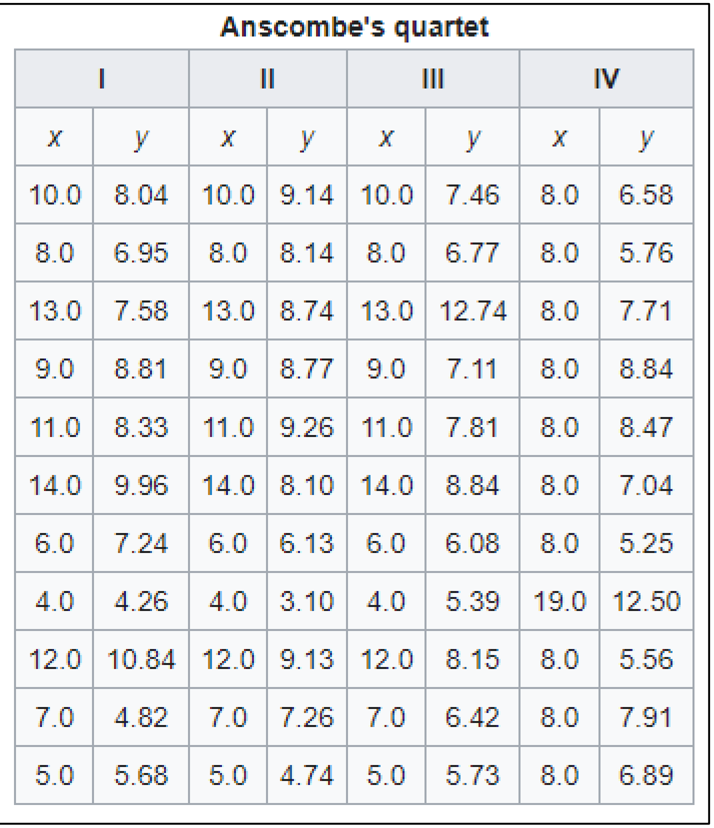
These are features which contributing significantly.
Temp = 0.5601, Weather_Good/Clear = 0.2890, Year_2019 = 0.2306, Weather_Moderate/Misty = 0.2082

1. Explain the linear regression algorithm in detail

- Data Understanding, Data Cleaning, Univariate Analysis, Segmented Univariate Analysis, Bivariate Analysis, Recommendations/Results
- Creating a list and putting all category columns in to it and converting them to category data type
- For Linear model creating dummies
- Combining both the dataframe, bike_sharing_df and of dummy variables
- Dropping columns from which dummy variables were created
- Importing statsmodel and sklearn libraries for Linear regression model building
- Splitting the date into two train and test dataframes
- Verify the columns and rows
- importing MinMax scaler from preprocessing module of sklearn library.
- Defining a variable scaler for minmax scaling.
- Performing scaling on all the numerical variables of train dataset and leaving Count variable aside.
- Checking all columns and all the variables after scaling.
- Let's check the correlation coefficients to see which variables are highly correlated.
- Dividing training set into X_train and y_train sets for the model building.
- Importing RFE library for feature selection and after this will perform manual feature selection.
- Using RFE for feature selection and limiting to selection to 15 features.
- Creating a list of features selected by RFE.
- Feature which are choose by RFE during feature selection.(so un-supported columns)
- Creating new train dataframe with RFE selected features.

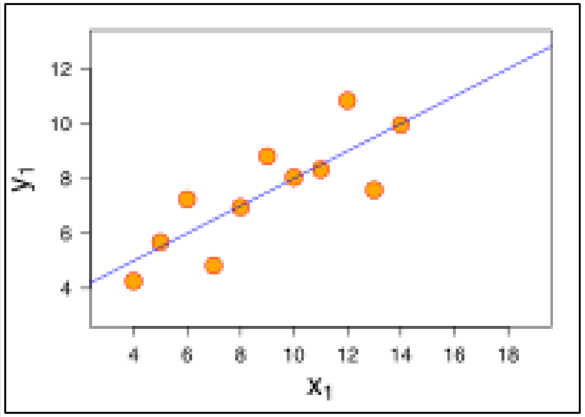
2. Explain the Anscombe's quartet in detail.

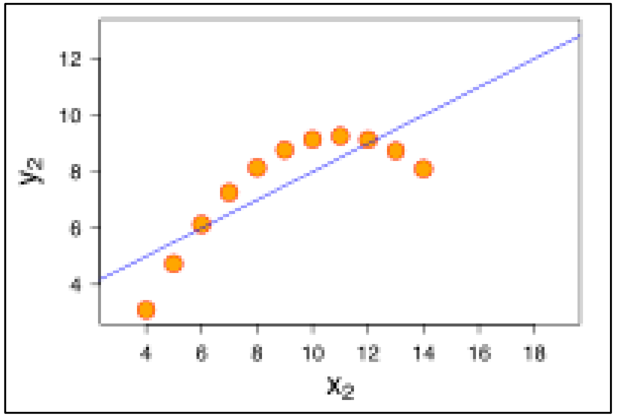
- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics,
- yet have very different distributions and appear very different when graphed.
- Each dataset consists of eleven *(x,y)* points. The effect of outliers and other influential observations on statistical properties.
- He describe intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.“



Anscombe's quartet

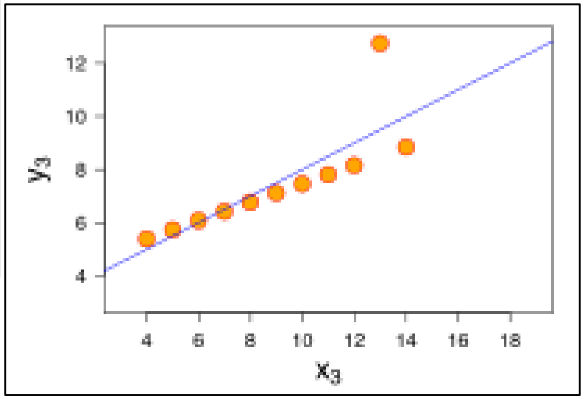
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

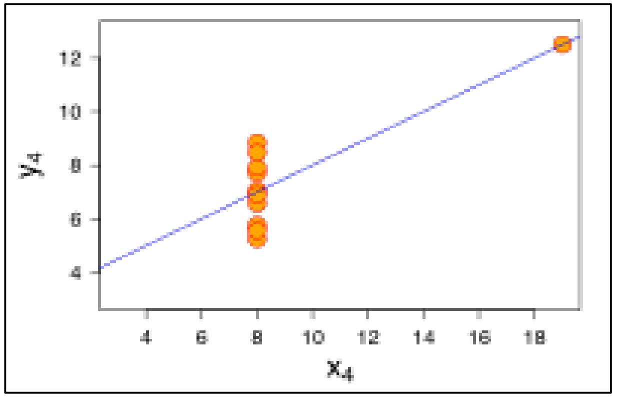




It is simple linear relationship

•It is not linear, and the Pearson correlation coefficient is not relevant.





•One high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

model relationship is linear, but it have a different regression line.

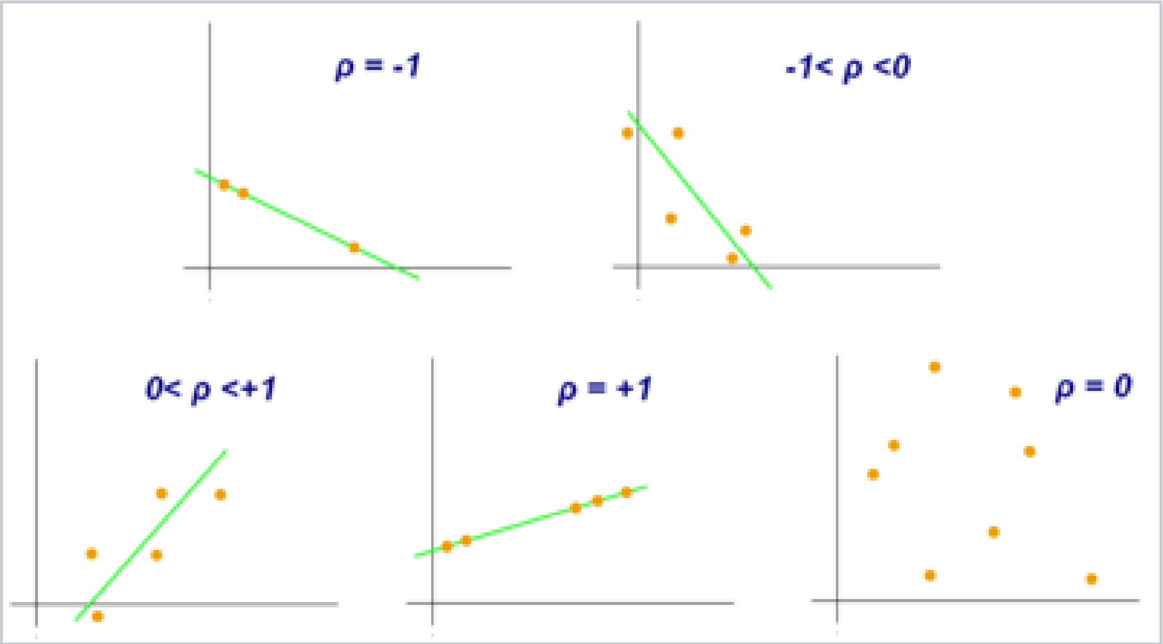
3. What is Pearson's R?

Pearson correlation coefficient (PCC)also known as Pearson's *r*, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient is a measure of linear correlation between two sets of data.

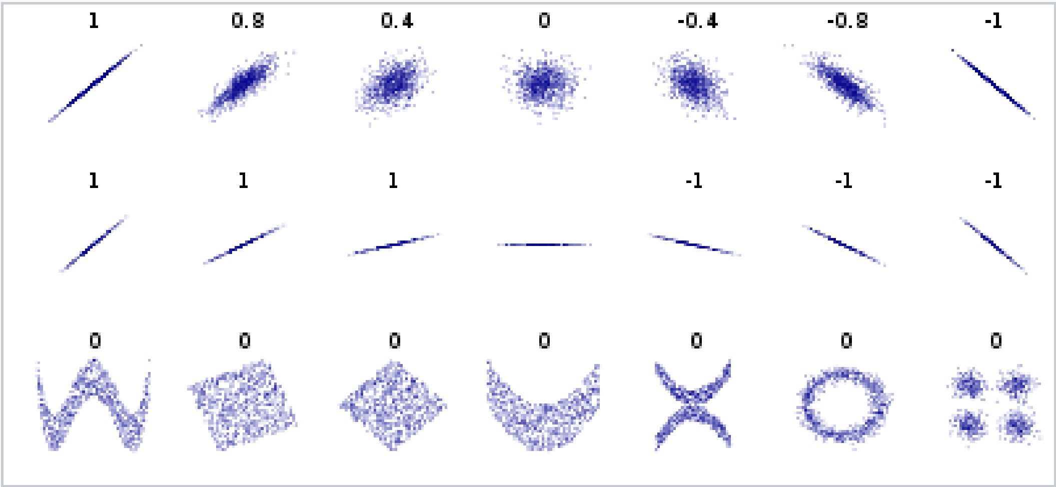
It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1.



Examples of scatter diagrams with different values of correlation coefficient (*ρ*)



Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

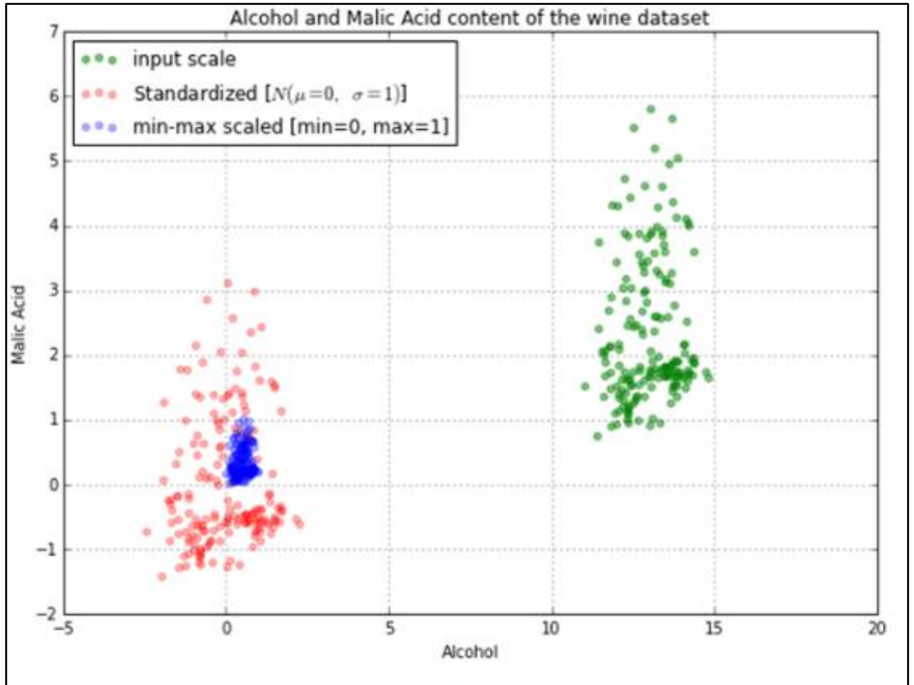
It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

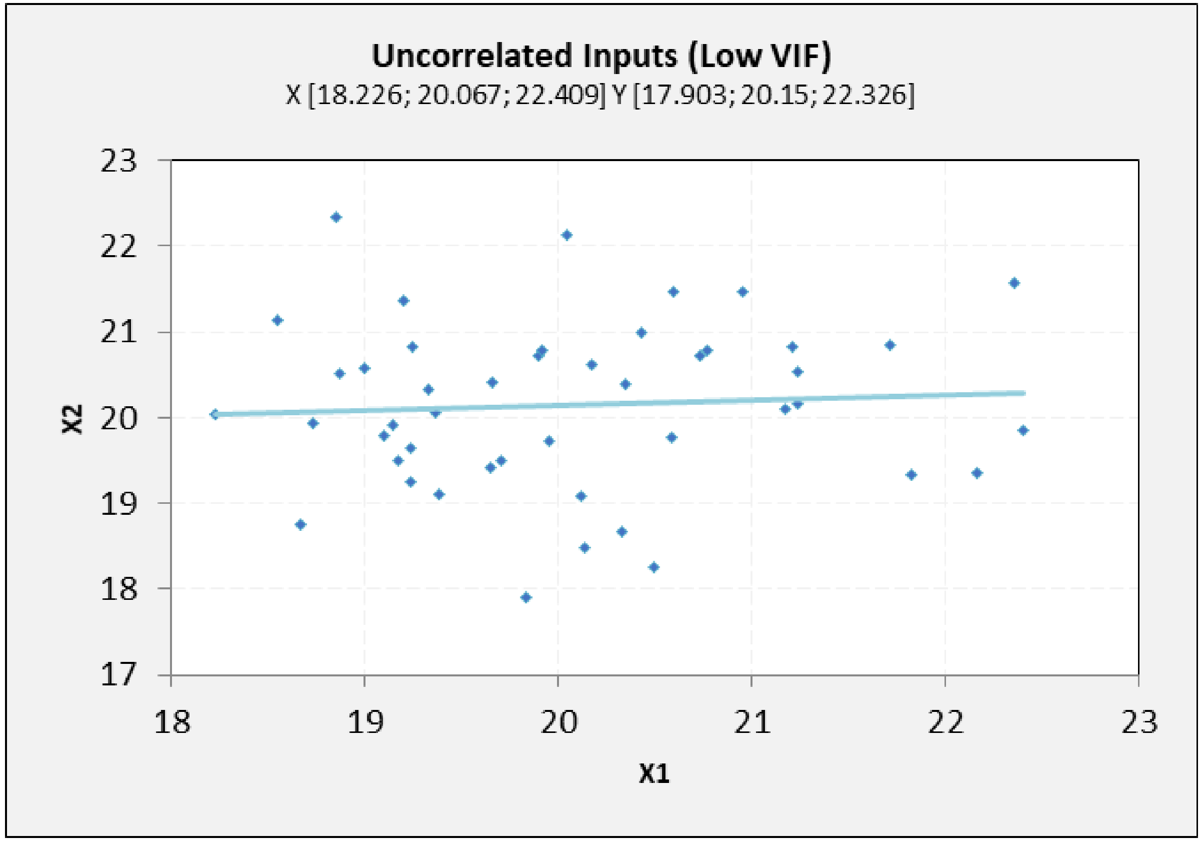


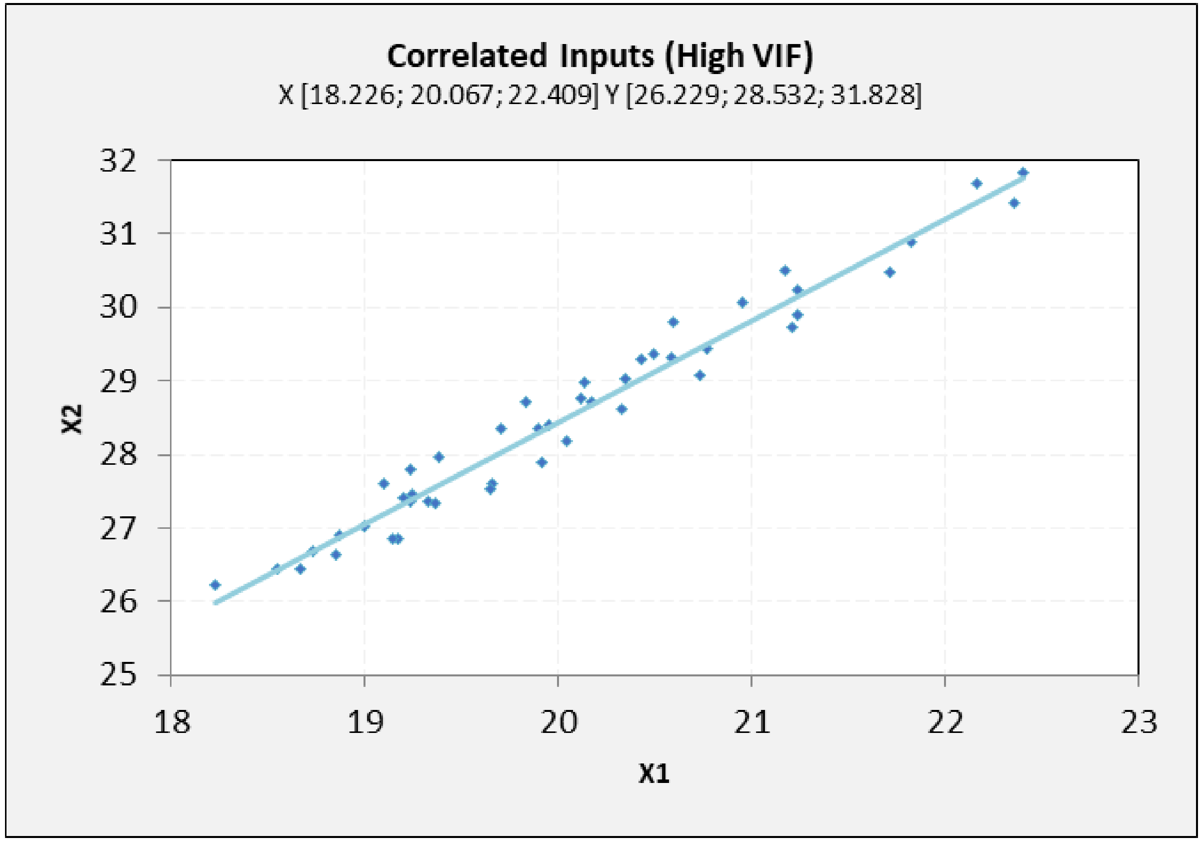
The impact of Standardization and Normalisation on the Wine dataset

- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.
In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

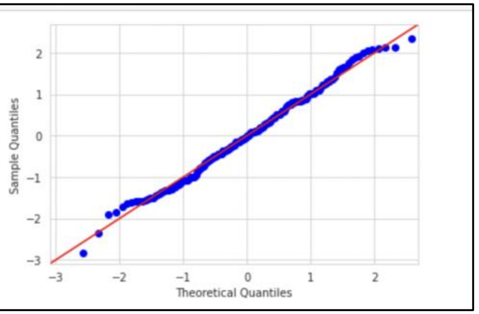


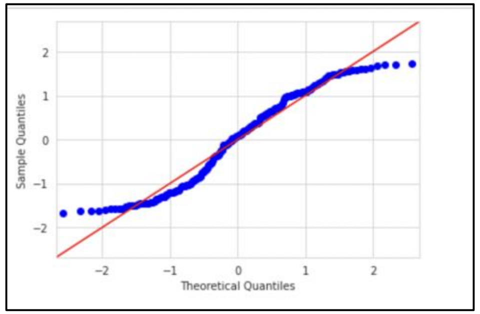


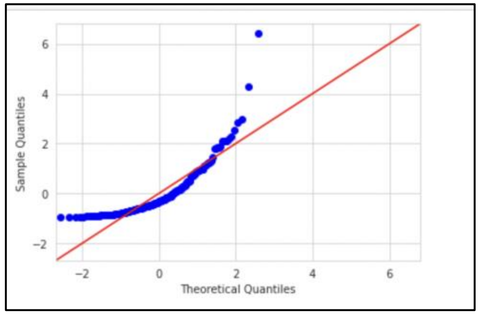
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

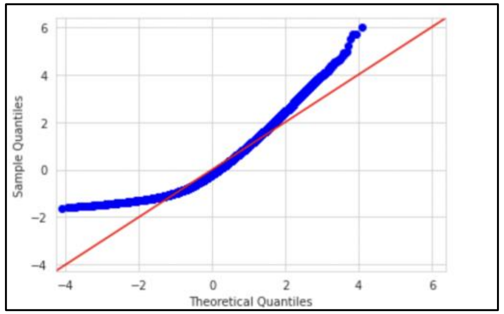
Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. Before we dive into the Q-Q plot, let’s discuss some of the probability distributions.

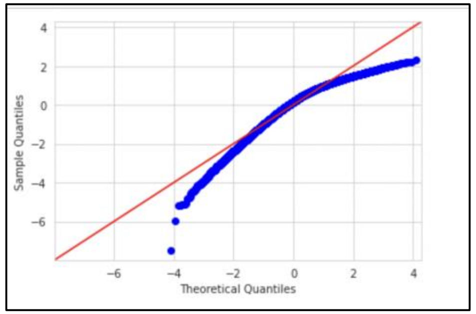
QQ plots is very useful to determine
If two populations are of the same distribution
If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
Skewness of distribution











Conclusion
As you build your machine learning model, ensure you check the distribution of the error terms or prediction error using a Q-Q plot. If there is a significant deviation from the mean, you might want to check the distribution of your feature variable and consider transforming them into a normal shape.

THANK YOU

👤 PRASHANT TARIWAL

☎ +91-9740248597

✉ prashanttariwal@gmail.com