Assignment-based

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans. There are 6 categorical variables used in the dataset and Cnt is the dependent variable. According to my analysis, each value of the categorical variable becomes a predictor for the dependent variable. So, we can clearly understand the exact value which is more important to make final prediction about the dependent variable. Eg. Apart from 4 categories of Weather Situation variable, Light snow has a negative effect on the dependent variable, in other words, we can say that when there is less snow, people are less likely to take a rental bike.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans. When we use drop_first=True, the first category of the variable is dropped. Do, if we have k categories then instead of k dummy variables, we will have $k - 1$ variables. When we have less variables, we can interpret the data more clearly. That's why we drop the first category during dummy variable creation.

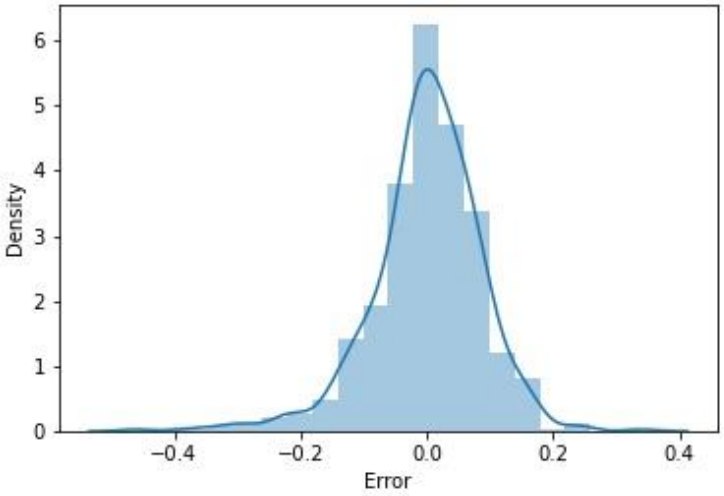
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. Temp and Atemp has the highest correlation with the target variable Cnt.

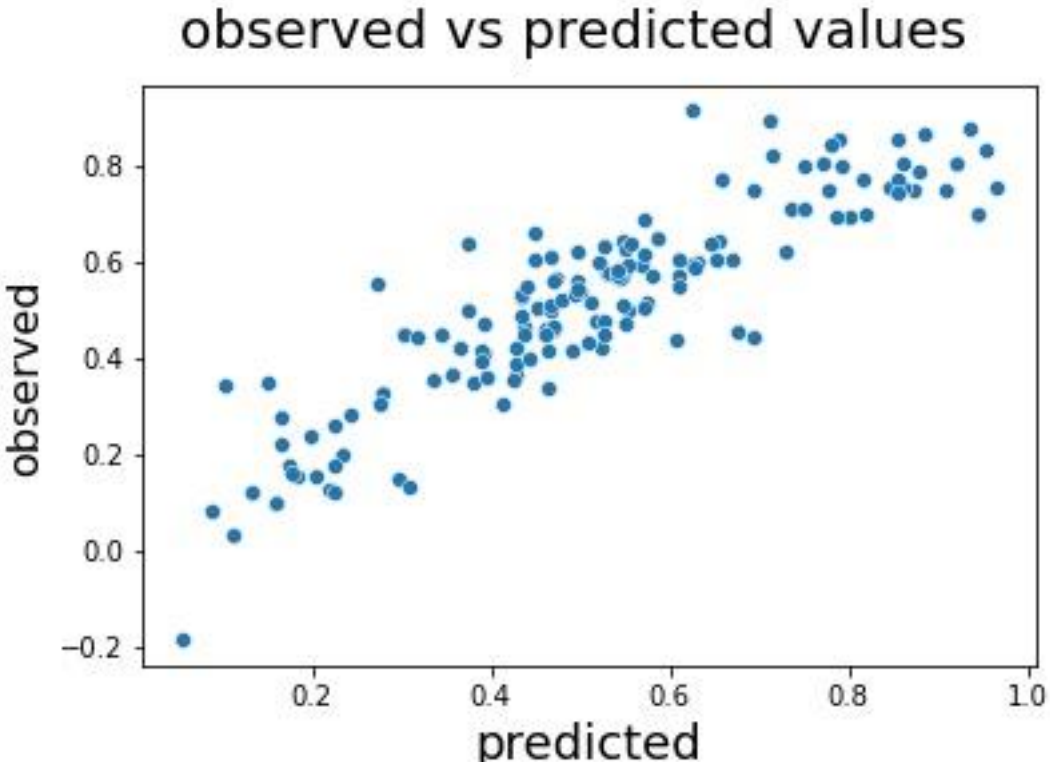
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans.

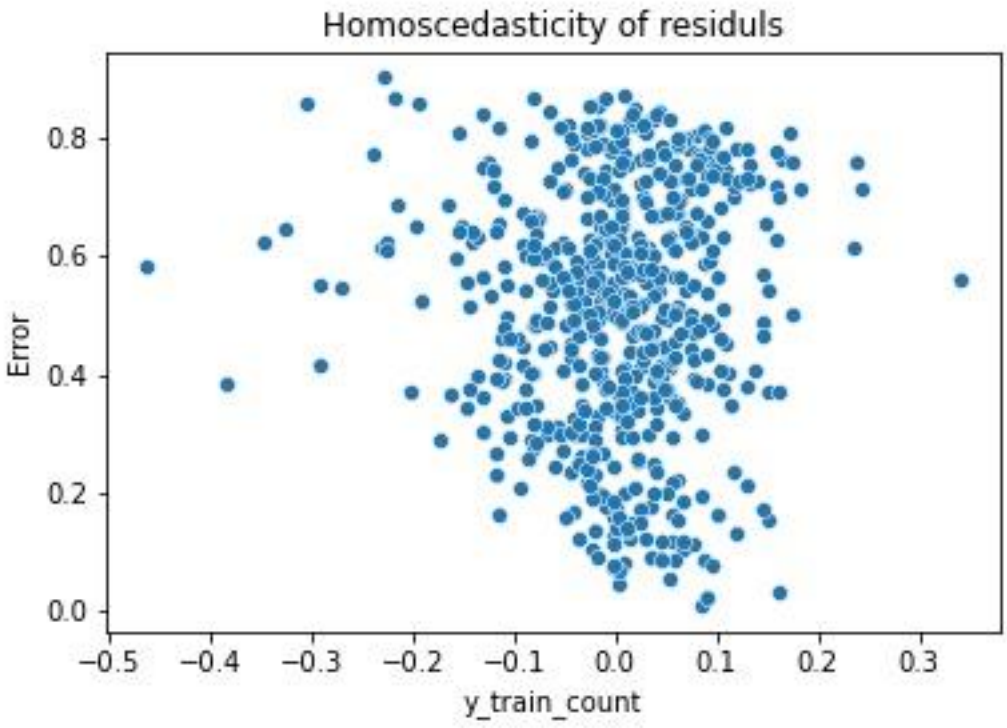
- To check that the error terms are normalized, I have plotted a distribution plot of the error terms, which shows a mean of 0.



- To check the linearity of X variables and Y, I have plotted a scatter plot between the observed Y and predicted Y.



- To check homoscedasticity, I have plotted errors and y_pred and observed that error terms are scattered



- To check that there is no multicollinearity, the final model has all features' VIF less than 5.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans. The top 3 features contributing towards explaining the demand of the shared bikes:

1. Temp (coeff= 0.443)
2. WeatherSituation_LightSnow (coeff= -0.263)
3. Year (coeff=0.228)
   As year is included in the model, but I find that it is not a right feature to be included.
4. Windspeed(coeff=0.184)

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a supervised learning technique which is used for predictive analysis and shows that there is a linear relationship between the independent variable(X-axis) and dependent variable(Y-axis). If we plot datapoints on the X and Y axis, we will observe that there is a line with goes through most of the data points. The equation of line is given by $Y = mX + c$, where m is the slope of the line and c is the intercept of this line on the Y axis.

The aim of the linear regression algorithm is to get the best values of m and c to find the best fit line with the least error. Error is calculated by taking the difference of predicted value with the actual value.

A cost function is used to estimate the values of m and c for the best fit line. Basically, it measures how well our model is performing. For linear regression, we use Mean Squared Error (MSE) cost function, which is the average of the squares of the error. Gradient descent is used to minimize the cost function.

Finally, to measure the performance of the linear regression model we use R^2 or R-squared. It is a statistical method to determine the goodness of fit. It measures the strength of relationship between the dependent and independent variable on the scale of 0-100. The formula of R-square is given by $R^2 = \frac{explained\ variance}{total\ variance}$

There are few assumptions made in linear regression:

- There is a linear relationship between the dependent and independent variables.
- There is little or no multicollinearity between features. It is because the presence of multicollinearity does not define which feature is affecting the target variable.
- Homoscedasticity. There is same error term for all the independent variables. For this, there should not be any clear pattern distribution of the data points.
- Error terms are normally distributed.

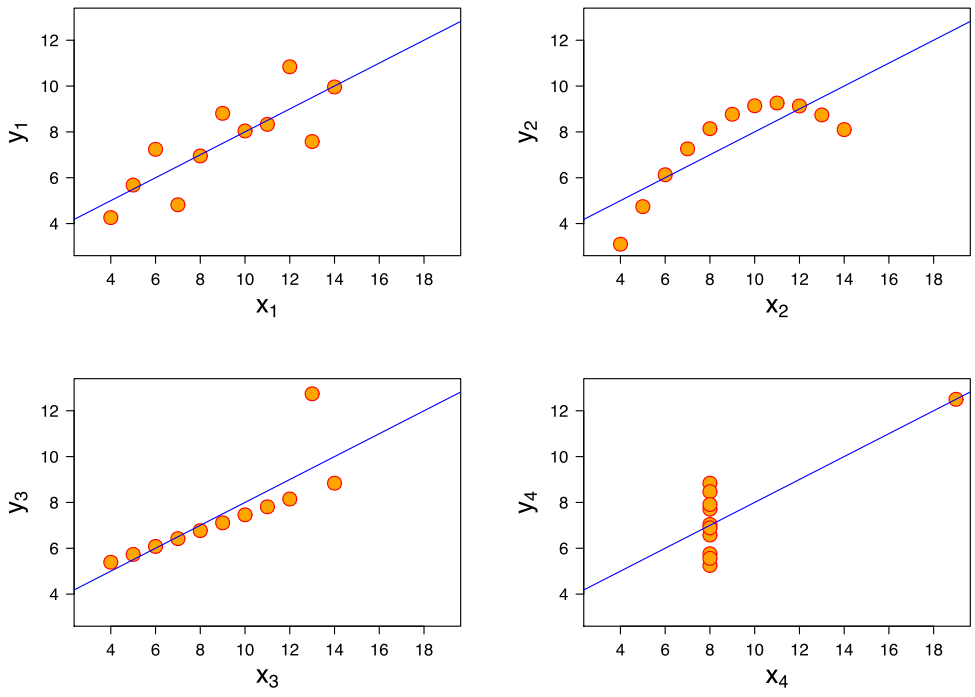
Q2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet comprises of four data sets having identical statistical features such as mean, variance, correlation, etc. yet they appear differently when they are visualized in a graph. Each data set has 11 data points (x, y). These data sets were first illustrated by a famous statistician Francis Anscombe to demonstrate the importance of plotting the data sets before analysing it.

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

Anscombe's quartet 4 data sets

(image reference: wikipedia)



The data sets when graphed

(image reference: wikipedia)

Explanation about each graph (from left to right row-wise)

1. There is a linear relationship between X_1 and Y_1
2. There is a non-linear relationship between X_2 and Y_2
3. There is a perfect linear relationship between X_3 and Y_3
4. This graph shows that an outlier is enough to produce a high correlation coefficient.

Q3. What is Pearson's R?

Ans. Pearson's correlation coefficient (R) is the most widely used correlation coefficient which describes the strength and direction of the linear relationship between two numerical variables. It can also be used to test the statistical hypothesis, that is we can measure how close the observations are to the best fit line.

We can choose to use Pearson's correlation coefficient in our model when all the below points are true:

- Both variables are numerical
- The variables are normally distributed.
- The data has no outliers.
- The relationship is linear.

The formula to calculate Pearson's correlation coefficient is:

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (y)^2]}}$$

$$where\ (x, y)\ are\ the\ data\ points.$$

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling is a pre-processing step in machine learning algorithm to handle highly varying values of a feature. Using scaling we bring the values of a feature to a fixed range.

Scaling is performed as a pre-processing step so that a machine learning algorithm weighs all the features equally. If few features have larger values,

the ML algorithm will produce lower coefficients and when features have smaller values, higher coefficients are produced regardless of the unit of values. For example, in the housing example, values of the feature “area” is higher than other features: bedrooms, bathrooms, guestroom, basement, etc. We must do the scaling here to bring them to a comparable scale. Otherwise, higher, or lower coefficients will be given to few features and the model will not predict well.

There are two common ways of scaling:

1. Normalization: This technique scales the values of a feature between o and 1. The formula to perform normalization is: $X_{new} = \frac{X_i - \min(X)}{\max(X) - \min(X)}$
2. Standardization: This technique scales the values of a feature so that the distribution has a mean of 0 and standard deviation of 1. The formula to perform standardization is: $X_{new} = \frac{X_i - X_{mean}}{Standard\ Deviation}$.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. Variance Inflation Factor is an index that measures how much correlated are the variables in the regression model. If there is a perfect correlation between two independent variables, then we get R^2 as 1. The formula for VIF calculation is $\frac{1}{1-R^2}$, this leads to VIF=infinity.

To solve the problem of perfect multicollinearity, one simple solution is to drop one of the variables which is causing this.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. Q-Q plots are also know as Quantile-Quantile plots. The purpose of this plot is to find put if the two data sets come from the same distribution or not. The quantiles of the first dataset are plotted again the quantiles of the second data set. For example, median is a quantile which divides the dataset into two, one half lies above median and other half lies below median.

When we use Q-Q plot, we get to know what kind of probability distribution is present in the dataset; normal, uniform or exponential. These plots are very useful to find:

- If the two populations are of the same distribution
- Residuals follow a normal distribution (which is an assumption of linear regression model)
- Skewness of the distribution

These plots are important because they can help us in choosing the type of regression model depending upon the type of the distribution.