Assignment-based Subjective Questions

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

   Ans- Categorial variables analyzed in this dataset-

   1. Year- the demand of bike was higher in 2019 compared to previous year.

   2. Season – demand is reduced during spring as compared to other seasons.

   3. Weather Situation- Demand is higher during clear weather.

   4. Weekdays- When comparing demand across weekdays Fridays on an average has more demands where 50 percentiles of demand is between 300 to 6000. Whereas demands peaks during Monday and Weekend where it can go as high > 8000

   Further analysis points that there is when weather and season taken into account for weekday analysis there appears to be correlation between them. Which may be affecting the demand.

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

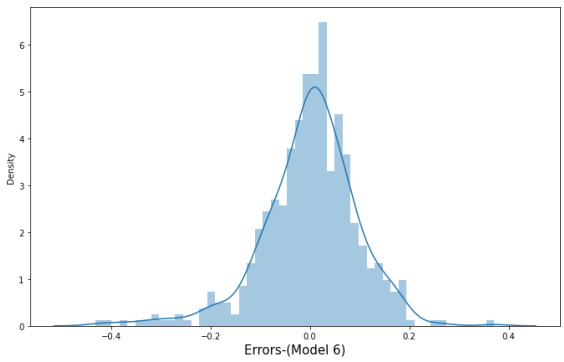
   Ans- The parameter drop_first=True is used while creating dummy variable. This prevents the creation of extra column, which may create unnecessary correlation. For example – house type is categorial attribute and has values- Furnished, Semi Furnished and Unfurnished. Here only 2 columns are required because to identify a value belonging to any of category it should not belong to 2 categories.

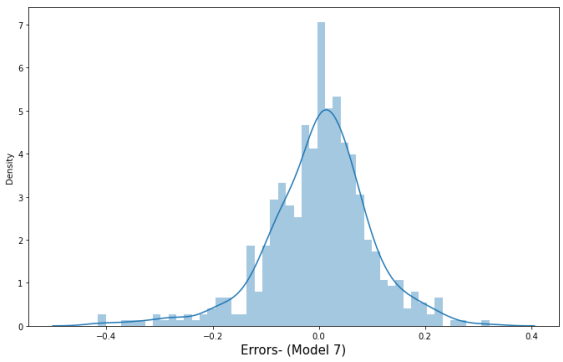
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   Ans- The highest correlation with the target variable (cnt) is with **temp** and **atemp** attributes. The correlation between them is **0.64** and **0.65** respectively

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

Ans – Model is validated by plotting the residual distribution. The distribution of the error term should follow the normal distribution. Another validation can be done by comparing R Squared and Adj R Squared. The difference should be minimum.





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

Ans - Top 3 features

a. Year- with coefficient- 0.2330
b. Weather Situation – Winter (Coefficients – 0.078)
c. Weather Situation- Snow (with -ve Coefficient of 0.297)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- This is a machine learning algorithm which is used for predictive analysis and based on supervised learning. This is the basic form of machine learning where the training data is used to derive a model based on some variables of the dataset. As the name suggests the two variables (dependent and target) which can be represented on x and y axis should be linearly related. In some case where the graph direction is linearly upward, this is called positive correlation and when the graph is linearly downward means the variables have -ve correlation.

  The method is used to predict a quantitative response Y from the predictor variable X.

  Mathematical equation for linear regression is –

$$y = mx + c$$

  Here, x and y are two variables on the regression line.

  m = Slope of the line

  c = y-intercept of the line

  x = Independent variable from dataset

  y = Dependent variable from dataset

  During regression task the cost function is used to derive the best possible value for m and c. Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the mapping function that maps the input variable to the output variable. This mapping function is also known as the Hypothesis function.

  In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

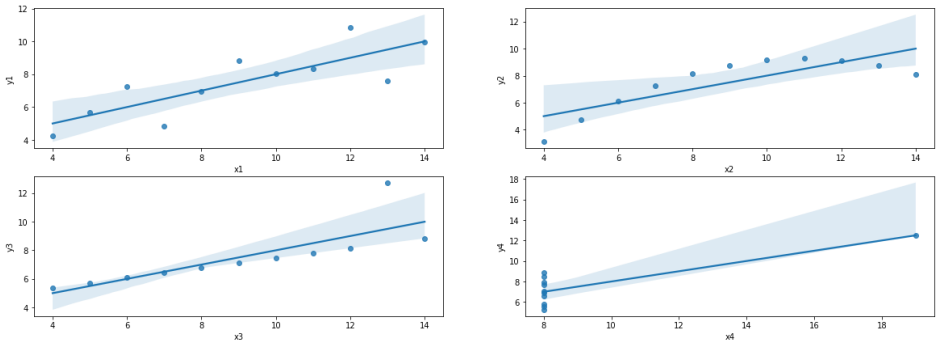
$$MSE = \frac{1}{N}\sum_{i=1}^{n}\left(y_i - (mx + c)\right)^2$$

Using the MSE function, the values of m and c can be changed such that the MSE value settles at the minima.

2. Explain the Anscombe's quartet in detail.

Ans- Anscombe's quartet was constructed by the statistician Francis Anscombe. This consists of 4 datasets that have nearly identical statistical properties but appear very different when plotted. Each dataset consists of eleven (x,y) points. The quartet was constructed to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties

When plotted these datasets looked as below



a. In the first (top left) there is linear relationship between x and y.
b. In the second (top right) the relationship between x and y is non linear.
c. In the third (bottom left) it can be concluded that there exists a perfect linear relationship between x and y but because of an outlier datapoint the line is skewed towards the outlier.
d. Finally, the fourth (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

3. What is Pearson's R?

Ans- The Pearson product-moment correlation coefficient or the Pearson's coefficient is the measure of strength of linear relationship between two variables. This is denoted by r.
Using this coefficient one can attempt to draw a best fit line through the dataset of two variables and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit.
Pearson's coefficient can take value ranging from -1 to 1,
Where if coefficient value is

< 0 denotes negative relationship i.e., value of both the variable will move in opposite direction

>0 denotes positive relationship i.e., value of both variables will move in same direction.

=0 denotes no relationship.

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

Ans- Scaling is crucial step in data preprocessing. A dataset which has multiple features and in different units will pose challenge to the performance of some machine learning algorithm like Linear regression. For example, a feature in Kg another in gram or mg will have values with different ranges. Linear regression uses Gradient Descent as optimization technique. So, the presence of feature with varied units will affect the step size for the features.

There are 2 ways in which scaling can be done-

a. Normalization- This is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
   Min Max scaling can be done using below formula-

$$X' = \frac{\text{X - Xmin}}{\text{Xmax-Xmin}}$$

b. Standardization- Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.
   Formula for standardization is - $\text{X}' = \frac{\text{X-}\mu}{\sigma}$

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

Ans – VIF is infinite or displayed as **inf,** this is due to the reason that there is perfect correlation between 2 independent variables. In such cases the $R^2 = 1$ and therefore $1/(1 - R^2) = \propto$. To solve this problem one of the variables from the dataset that is causing the multicollinearity is dropped.

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

Ans- A Q-Q plot also known as Quantile-Quantile plot is used to determine if the data comes from theoretical distribution like a Normal, Exponential or Uniform distribution. It also helps to determine if 2 datasets come from populations with common distributions.

With a QQ-plot, the quantiles of the sample data are on the vertical axis, and the quantiles of a specified probability distribution are on the horizontal axis. The plot consists of a series of points that show the relationship between the actual data and the specified probability distribution. If the elements of a dataset perfectly match the specified probability distribution, the points on the graph will form a 45-degree line.

In linear regression when the training and test data set is received separately then using Q-Q plot it can be confirmed of both are from the population with same distribution.