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: The Demand of Boom bikes is less in the month of spring when compared with other seasons. The demand of Boom bikes increased in the year 2019 when it compared with year 2018.

2. Why is it important to use drop_ first=True during dummy variable creation?

Ans: The drop_first=True is helps in reducing the extra column and 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?

Ans: The numerical variable of 'registered' has the highest correlation with the target variable 'cnt', if we consider all the features. But after data preparation, when we drop registered due to multicollinearity the numerical variable 'atemp' has the highest correlation with the target variable 'cnt'.

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

Ans: The linearity assumption can best be tested with scatter plots, the following two examples depict two cases, where no and little linearity is present.

Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

The Four Assumptions of Linear Regression

- Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.

- Independence: The residuals are independent. ...

- Homoscedasticity: The residuals have constant variance at every level of x.

- Normality: The residuals of the model are normally distributed.

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

Ans: Based on the model the top 3 Features are:

1. Temperature = 0.4911

2. Year = 0.2333

3. Winter = 0.0612

The final model will be having the highest demand in 2019 and lowest demand in 2018.In weekdays demand increases on Sunday. In month the sales increases in September and decreases in January.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. The Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

A Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e. g. sales, price) rather than trying to classify them into categories (e.g., cow, goat). There are two main types: Simple regression and Multivariate regression.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). The Linear Regression Equation

The equation has the form Y= a + bX, where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e., it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed. Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets. The most important insight of Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit. In each panel, the Pearson correlation between the x and y values is the same, r = .816. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values.

3. What is Pearson’s R?

Ans: The Pearson r is a measure to determine the relationship (instead of difference) between two quantitative variables (interval/ratio) and the degree to which the two variables coincide with one another—that is, the extent to which two variables are linearly related: changes in one variable correspond to changes in another variable.

Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables. the correlation coefficient r measures the strength and direction of a linear relationship between two variables on a scatterplot. The value of r is always between +1 and –1. To interpret its value, see which of the following values your correlation r is closest to: Exactly –1.

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

Ans: Scaling is a technique to standardize the independent features present in the data in a fixed range. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Scaling is essential for machine learning algorithms that calculate distances between data. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

The Difference between Normalization and Standardization is:

Normalization is often called as Scaling Normalization.

Standardization is often called as Z-Score Normalization.

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

Ans: An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation

The Variance Inflation Factor (VIF) is a measure of collinearity among predictor variables within a multiple regression. It is calculated by taking the ratio of the variance of all a given model's betas divide by the variance of a single beta if it were fit alone. Between the variables. A general rule of thumb is that if VIF > 10 then there is multicollinearity.

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

Ans: The quantile-quantile (q-q) plot is a graphical technique for determining of two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, it means the fraction or percent of points below the given value. That is, the 0.3 or 30% quantile is the point at which 30% percent of the data fall below and 70% fall above that value. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.