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?

Categorical variables yr,weathersit,season has major effect on dependent variable.

- The median count varies for each season.
- The count increases/decreases based on the weathersit.
- For each year the count seems to be increased.

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

We create dummy variables to give a numerical value to the categorical variables.

If there are n categories then n dummies variables can be created. But n-1 dummy variables are sufficient to recognize the categories.

Hence we use drop_first=True to create n-1 dummy variables.

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

Temp and atemp has the highest correlation(0.65) with Target variable.

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

Residual Analysis is performed.

Visualizing the residues by distplot to check if they are normally distributed.

Plotting the scatter plot of residues so that the homoscedasticity can be validated.

Checking if X and y have linear relationship.

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

1.Temp--->positive

2.year--->positive

3.weather_sit(Light snow/rain)--->negative

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables. It finds how the value of the dependent variable is changing according to the value of the independent variable. The linear regression model provides a sloped straight line representing the relationship between the variables.

There are 2 types of linear regression:

1. Simple Linear Regression

2. Multiple Linear Regression

Mathematical Representation:

$Y = \beta0 + \beta1.X1 + \beta2. X2 + \beta3. X3 + \beta4. X4+ \beta5. X5 + \beta5. X6 + \epsilon$

1. Regression Coefficient (or β1):

The Regression Coefficient in the above equation talks about the change in the value of dependent variable corresponding to the unit change in the independent variable.

2. Intercept (or β0):

Intercept is a constant value which tells us at what point in the x-y coordinate graph, should the regression line must start if it follows a linear regression.

3. Error Terms or Residuals (ϵ):

It is the difference between the actual and the predicted data point in the x-y coordinate graph.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson's R?

Pearson's R 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.

The Pearson correlation coefficient(Pearson's R) is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

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

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.

Why:

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.

Normalized scaling:

It brings all of the data in the range of 0 and 1.

Minmax scaling, x = x – min(x) / max(x) – min(x)

Standardized 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).

Standardization , x = x – mean(x) / standard deviation(x)

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 , that is very high correlation between the predictors

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

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.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.

c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis