Assignment based Subjective Questions

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

We created dummy variables as the categorical variables might be hard to explain from the model. Dummy variables are the ones that take only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Dummy variables enable us to use a single regression equation to represent multiple groups so that we don't have to write out separate equation models for each subgroup. The dummy variables act like 'switches' that turn various parameters on and off in an equation. We worked on columns “weathersit” and “seasons” to create dummy variables.

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

A. drop_first=True removes first level to get k-1 dummies out of k categorical levels. It is used to remove the first column which is created for the first unique value of a column. It is important to use, as it helps in reducing the extra column created during dummy variable creation and hence reduce the correlations created among dummy variables. Retaining this extra column does not add any new information for the modelling process.

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

A. instant vs temp/atemp - S- shaped trend;

cnt vs temp/atemp are highlighly correlated.

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

A. Checked for in the pair plot - variables with non- linear trend were checked for correlations. Performed R squared test for model evaluation.

- There should be a linear relationship between dependent variable and independent variables.
- There should be no correlation between the error terms.
- The independent variables should not be correlated.

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

# 1. Yr/Month
# 2. Season_Fall
# 3. weathersit_Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

General Subjective Questions

Q. Explain the linear regression algorithm in detail

A. Linear regression is a supervised learning model. Linear regression assumes a linear relationship between the input variables (x) and the single output variable (y).
When there is a single input variable (x), the method is referred to as simple linear regression and when there are multiple input variables it is referred to as multiple linear regression.
Equation:

$$y = mx + c$$

y = output

x = input

m = slope

c = intercept
Once we find the best m and c values, we get the best fit line. When we are finally using our model for prediction, it will predict the value of y for the input value of x. By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum.

Cost function (J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value and true y value.

$$J = 1/n \sum (\text{pred}(y) - \text{true}(y))^2$$

Q. Explain the Anscombe's quartet in detail.

Anscombe's quartet is used to understand the importance of visualizing data before applying algorithms for building models. It comprises of four data-sets and each data-set consists of eleven (x,y) points. All these data-sets share the same descriptive statistics i.e mean, variance, standard deviation etc. but different graphical representation. This means the data features must be plotted to see the distribution of the samples that can help identify the various anomalies present in the data such as outliers, diversity of the data, linear separability of the data, etc. Moreover, the linear regression can only be considered a fit for the data with linear relationships and not any other.

Q. What is Pearson's R?

A. Pearson's R is a measurement of the strength of the relationship between two variables and their association with each other. The relation can be positive or negative (inversely proportional). The correlation coefficient formula finds out the relation between the variables and returns the values between -1 and 1. The stronger the association between the two variables, the closer the answer will incline towards 1 or -1. Attaining values of 1 or -1 signify that all the data points are plotted on the straight line of ‘best fit.’ It means that the change in factors of any variable does not weaken the correlation with the other variable. The closer the answer lies near 0, the more the variation in the variables.

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

A. It is applied to independent variables to normalize the data within a particular range. Sometimes, the collected data set contains features highly varying in magnitudes, units and range. We do scaling to bring all the variables to the same level of magnitude. It can make the analysis of coefficients easier and also prevent the model from being biased. If the features differ in scale then this may impact the resultant coefficients of the model and it can be hard to interpret the coefficients.

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

Standardized scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean zero and standard deviation one.

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

A. Infinite VIF (Infinity Variable Inflation) shows perfect correlation between two independent variables. VIF is used to identify the correlation of one independent variable with a group of other variables. VIF is a measure of the amount of multicollinearity in a set of multiple regression variables.

VIF = 1/(1-R2). When R2 is 1, VIF goes to infinity.

If the VIF =1.0 then the independent variables are orthogonal to each other i.e no collinearity.

We need ways to reduce dimensions. As a fix for infinite or very high VIF we need to drop one of the variables which is causing this perfect collinearity. There can be the case that one feature is correlated with many others and we might want to remove it.

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

A. 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.