Assignment-based Subjective Questions

1. From the analysis of categorical variables in the dataset we can see that:

i. season has an impact on the target variable

ii. yr has an impact on the target variable

iii. mnth has an impact on the target variable

iv. holiday has an impact on the target variable

v. weekday does not have a significant impact on the target variable

vi. workingday does not have a significant impact on the target variable

2. During dummy variable creation using pandas.get_dummies() we get n dummy columns for a categorical column having n different categories while the same information can be represented using n-1 columns.

Using drop_first=True ensures we get n-1 dummy columns instead of n columns.

3. Looking at the pair-plot among the numerical variables, it can be seen that the column “registered“ has the highest correlation with the target variable (0.95).

4. The following points help validate the assumptions of Linear Regression –

i. The R2 score of the final model is 0.80 which shows a linear relationship between the predictor variables and the target variable.

ii. The residuals are normally distributed and centered around 0.

iii. Perfect multicollinearity is not there in the final model as the VIF of no predictor variable is greater than or equal to 5.

iv. The residual terms are spread out on either side of 0 when plotted against the predicted values.

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

i. Spring

ii. Light Snow & Rain

iii. yr

General Subjective Questions

1. 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. cat, dog).

Mathematically

$$f(x,y,z)=w1x+w2y+w3z$$

The variables x,y,z represent the attributes, or distinct pieces of information, we have about each observation.

Cost Function - It is a measure of how wrong the model is in terms of its ability to estimate the relationship between X and y.

Gradient descent – It is an efficient optimization algorithm that attempts to find a local or global minima of a function. It helps to minimize the cost function.

2. Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

Each dataset consists of eleven (x,y) points.

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Application - The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. Correlation is a bi-variate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. A value of ± 1 indicates a perfect degree of association between the two variables. The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a - sign indicates a negative relationship.

Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r.

Assumptions –

i. For the Pearson r correlation, both variables should be normally distributed.

ii. There should be no significant outliers.

iii. Each variable should be continuous.

iv. The two variables have a linear relationship.

v. The observations are paired observations.

vi. The error term is the same across all values of the independent variables.

4. Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

i. Normalized Scaling - Transform features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g., between zero and one. This Scaler shrinks the data within the range of -1 to 1 if there are negative values.

ii. Standardized Scaling - The Standard Scaler assumes data is normally distributed within each feature and scales them such that the distribution centered around 0, with a standard deviation of 1.

5. Variance Inflation Factors (VIFs) provide a one-number summary description of collinearity for each model term. Given an experiment with multiple factors, the variance inflation factor associated with the ith factor reflects the increase in the variance of the estimated coefficient for that factor compared to if the factors were orthogonal, and is defined as $VIF_i = 1/1\text{-}R^2_i$ where R^2_i is the coefficient of determination of a regression model where the ith factor is treated as a response variable in the model with all of the other factors. VIF_i can range from one to infinity. Values equal to one imply orthogonality, while values greater than one indicate a degree of collinearity between the ith factor and one or more other factors.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

6. The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

Importance - It 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.