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?

   Here are the observations from above univariate analysis of categorical data

   - We have more number of days whether weather was clear
   - When compared with working days data, majority of them are for working days
   - Data with Holidays is less

   Here are the observations from above Bivariate analysis for categorical columns

   - In Summer and Fall seasons good amount of rentals are there when compared with other two seasons
   - In Y2019 there are more number of rentals
   - During April to October months there are more number of rentals
   - When weather is Clear, there are more number of rentals

   Here are the observations from above Univariate analysis for numerical columns

   - On Windspeed, good amount of data distributed between greater than 5 and less than 20
   - On humidity, good amount of data distributed between greater than 40 and less than 90
   - On temp and atemp data, around 10 to 30 its distributed.

   Here is the equation based on the final model where R-squared value is 0.838, with p-values for all variables/features is less than 0.05 along with VIF of them is less than 5.

   cnt = 0.2893 + temp * 0.4026 - windspeed * 0.1540 - season_Spring * 0.1034 + season_Winter * 0.0650 + yr_2019 * 0.2348 - mnth_Dec * 0.0510 - mnth_Jan * 0.0556 - mnth_Jul * 0.0643 - mnth_Nov * 0.0488 + mnth_Sep * 0.0537 - holiday_Yes * 0.0913 - weathersit_Light_Snow * 0.2949 - weathersit_Mist_Cloudy * 0.0812

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

   “drop_first=True” needs to be used while creation of a dummy variable. This will be helpful for the model to reduce a kind of overhead/correlation factor comes into picture with its presence. As per the concept if we have a categorical variable with values of count as N, then there will be N-1 columns will be getting added with usage of “pd.get_dummies()” passing “drop_first=True” parameter.

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

Based on the output that got displayed from EDA in the python code @ cell#1105, we can see temp and atemp are having a highest co-relatoin with the target variable (I,e 'cnt' column)

Using heatmap output which is present @ cell#1106 we can see a correlation value for temp and atemp with cnt as 0.63

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

   Using the below set of steps model which got generated with python code got evaluated to see whether it got trained well or not.

   As per Part#8 step captured in code@ cell#1139, carried out resuidal analysis on the train data set by taking the diff between predicted vs actual. By plotting histogram of the residual terms we can see the distribution of error terms. Distribution came out be centered around 0 which was shown as part of output @ cell#1139

   As per Part#9 step captured in code @ cell#1144, we can see the relation between y_test and y_pred where the regression line is linear with mean around 0.0

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

   Based on the output with code @ cell#1135 and cell#1136, below features are having significant impact towards demand of shared bikes

   - 'temp'
   - 'yr_2019'
   - 'weathersit' when it is "Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist"

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   There are three types of machine learning algorithms

   - Regression
   - Classification
   - Clustering

   Liner regression that is getting discussed here comes under “Regression” type of machine learning. It falls under the category of supervised learning. It is a process of estimating the relationship between variables. How the values of dependent variable changes where we change one unit of each of the independent variable. Here we change one variable at a time and see how dependent variables will vary.

   Simple linear regression is used for prediction a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y

   The mathematical formula for linear regression is Y = mX + b, where Y is the dependent variable, X is the independent variable, m is the slope of the line, and b is the Y-intercept.

   This same regression model can be extended for multiple features by extending equation for the number of variables available within the dataset. This is called as multiple linear regression.

   Linear regression is widely used in forecasting and prediction usecases.

   Regression guarantee's interpolation of data not extrapolation.

2. Explain the Anscombe’s quartet in detail.

   Anscombe's quartet is a set of four datasets that have nearly identical statistical properties yet look very different when plotted. This highlights the importance of visualizing data and not relying solely on summary statistics. The quartet consists of four sets of x and y values, each with 11 data points. When plotted, they reveal different patterns - one linear, one non-linear, one with an outlier, and one with a strong relationship except for one point. Anscombe's quartet is a set of four datasets that have nearly identical statistical properties yet look very different when plotted. This highlights the importance of visualizing data and not relying solely on summary statistics. Key takeaway of this concept is before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson’s R?

   Pearson's R => Known as the Pearson correlation coefficient.

   It is another statistical measure that quantifies the strength and direction of the linear relationship between two variables.

   As we data range for correlation, with Pearson's R too data ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship.

   Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit .

   The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively.

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

   Scaling is a data preprocessing technique that transforms variables to a specific range or distribution.

   Why do we need to scale features ?

   - Ease of interpretation.
   - Faster convergence for gradient descent methods.

   Scaling just affects the coefficients and none of the other parameter's gets affected.

   There are tow major methods for scaling the variables/features

   - Standardisation
     - Brings all of the data into a standard normal distribution with mean zero and standard deviation as one.
     - called as Z-Score Normalization.
     - less affected by outliers.
   - MinMax Scaling
     - Brings all of the data in range of 0 and 1
     - called as Scaling Normalization
     - affected by outliers.

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

VIF, known as Variance Inflation Factor use one of the technique used for dealing with multicollinearity. VIF measures how well a predictor variable can be predicted using all other predictor variables. VIF gets a value of infinity when there is a perfect multicollinearity between variables.

$VIF_i = 1/(1-R_i^2)$

closer the R^2 value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable. Dropping one of the correlated features will help in bringing down the multicollinearity between correlated features.

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.