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? (3 marks)

   Season: Fall season seemed to have the highest count of bike sharings

   Year: 2019 seems to have a higher amount of bike sharings

   Month: October month seems to have the highest bike sharings

   Holidays have a higher bike sharing Bike sharing is significantly higher during clear weather

   There is absolutely no bikes borrowed when there is heavy rains with thunderstorms.

   Nothing significant was observed if the bike was borrowed on a weekday or working day

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

   We need to use the drop_first=True during creating dummy variable to eliminate redundance. For n different values in categorical variable column, will only need 'n -1' dummy variables to represent the n different values.
   For example, we need to analysis a political party. It may be republic, democratic or independent.

1	0	Republic
0	1	Democratic
0	0	If not Above it is independent

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   The highest correlation with the target variable, cnt, is with the independent variable temp.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
(3marks)

The assumptions of Linear Regression are:

- ❖ There should be a linear relationship between dependent and independent variables. Using pairplot, we checked if there were any linear relationships that existed between the independent variable and target variable cnt
- ❖ Residuals must be normally distributed. In the Step 4 of Residual analysis, distribution plot for error terms show normal distribution with mean at 0.
- ❖ Error terms must be independent. In Step 4 of Residual Analysis, a scatter plot of residuals shows that there is no visible pattern between error terms and hence are independent.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
(2marks)

The top 3 features contributing significantly towards the demand of shared bikes are:
a. Temp: temperature – higher the temperature, higher the demand
b. Season: winter season
c. Month: October

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear Regression algorithm is a machine learning algorithm, based on the predictive modelling technique that helps us to find the relationship between input and the target variable. It is used to find the effect of input on the target variable and predict the upcoming trends. There are two types of Linear regression based on the number of independent variables.

1. Simple Linear Regression: where there is only one independent variable (x) and a target variable (y)

2. Multiple Linear Regression: where there can be two or more independent variables (x1, x2, x3, …, xn) and a target variable (y).

The independent variables are known as “predictor variables” and the dependent variables are known as “output” or “target” variables.

Linear regression at each X finds the best estimate for Y. As the model predicts a single value, there is a distribution of error terms.

As we are making inferences on the population using a sample, the assumption that variables are linearly dependent is not enough to generalize the results from sample to the population. So, we should have some assumptions to make inferences.

Assumptions of Linear regression:

1. There is a linear relationship between X and Y

2. Error terms are normally distributed with mean zero.

3. Error terms are independent of each other 4. Error terms have constant variance. (homoscedasticity)

When you fit a straight line through the data, you will obviously get the two parameters of the straight line: intercept (β0) and the slope (β1). You start by saying that β1 is not significant, i.e. there is no relationship between X and y.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet can be defined as a group of 4 datasets which look nearly identical in simple descriptive statistics but have different distributions and appear differently when plotted on scatter plots. These peculiarities in the data should be identified and handled without doing which these peculiarities fool the built regression model. This signifies the importance of plotting graphs and visualizing the data before analyzing it and building a model. Anscombe's quartet suggests that the variables in the dataset must be plotted to observe the sample distribution which help in identifying the anomalies in the data.

3. What is Pearson's R? (3 marks)

Pearson's R is the correlation coefficient that lies between -1 and +1.
R = -1 and +1 means the data is perfectly linear with negative and positive slopes respectively. R = 0 means there is no linear correlation in the data
0<R<R8 means a strong association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a data pre-processing step applied on predictor variables to normalize the data within certain range, and thus help in speeding up the calculations in an algorithm. Scaling is very important to bring all the variables to the same magnitude level. This issue arises when the dataset contains variables with high variations in units, ranges and magnitude. If scaling is not done, then the algorithm only takes magnitude into account and not units. This results in incorrect modelling. So scaling is very important to address this issue. "Normalized scaling" means to scale a variable to have values between 0 and 1. "Standardized Scaling" means the data is transformed to have a mean of 0 and standard deviation of 1
We either Normalize (also known as Minmax Scaling) or Standardize the data. Standardization brings all the data into a standard normal distribution with mean 0 and standard deviation/ variance = 1. MinMax scaling, brings all the data in the range of 0 and 1. The formulae in the background used for each of these methods are as given below:

Standardization: $\frac{\boldsymbol{X-X(mean)}}{\boldsymbol{standard\ deviation}}$

MinMax Scaling: $\frac{\boldsymbol{X-Xmin}}{\boldsymbol{Xmax-Xmin}}$

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

(3 marks)

VIF is a measure or index of multicollinearity of a variable in a given dataset. In a Regression model which has been built using several independent variables, some of these variables might be interrelated, due to which the presence of that variable in the model is redundant. To address this issue, we use the value of VIF to determine whether or not to retain that variable. The value of VIF can range in between 1 and Infinity.

The formula for VIF is given by

$$VIF = \frac{1}{1-R^2}$$

VIF will have the value of Infinity when the value of R-squared = 1, this means that the variable is highly correlated with other variables in the data set.

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

(3 marks)

A Q-Q plot, or Quantile-Quantile plot is a graphical tool that helps us in assess and infer if the data came from some theoretical distribution, such as normal or exponential or uniform distribution.

In Linear Regression, if the training and test data set is obtained separately, we can use the Q-Q plot to confirm if both data sets are from the population with common distribution.

Q-Q plots take our sample data, sorts it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. If the points fall pretty closely along the line, the data are normal.

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.

To run the code that creates a Q-Q plot, you need to install and load the package stats.

Advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in

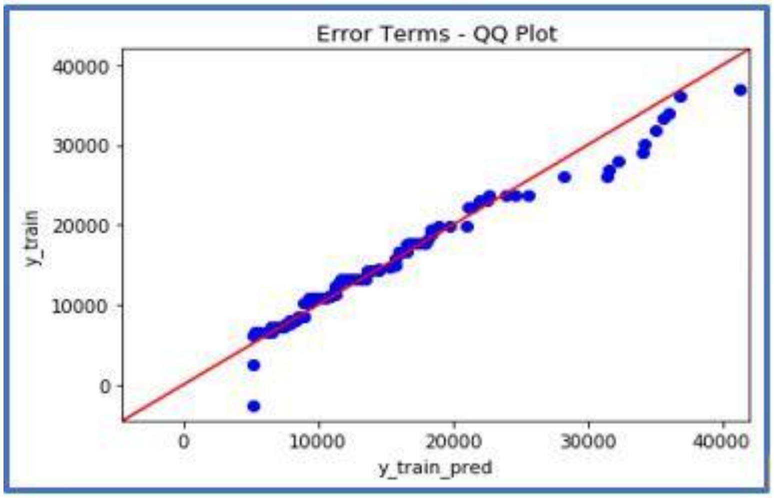
symmetry, and the presence of outliers can all be detected from this plot.
It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

Interpretation:

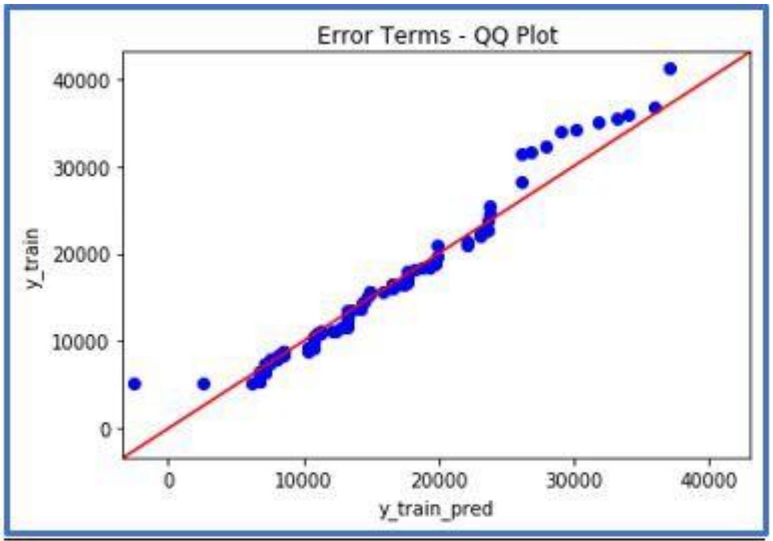
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.

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.


Error Terms - QQ Plot
y_train
y_train_pred