ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer 1. Insights that can be found from the categorical variables from the data sets are as follows:

1. Fall season shows higher demands .
2. Demand for 2019 year had higher demands.
3. Demand was increasing quite continously till June
4. Demand was higher in the month of September and lesser in the beginning and end of the year.
5. When weather is clear and during holiday too only when demand were higher for bikes .
6. When holiday is there demand decreased too
7. Weekdays are not providing clear picture for analysis .

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

Answer 2. It is important to use drop_first = True because it helps us in reducing the extra columns creation during dummy variable creation which reduces the correlation between dummy variables. In weathersit variable , first variable was not dropped as not to loose the info about severe weather situation.

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

Answer 3. The temp and atemp are the numerical variableshaving the highly positively correlated to each other with the target variable ‘cnt’ as 0.63.

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

Answer 4. We had validated the assumptions of Linear Regression after building model on the training test using following techiniques

1. **Residual Analysis:**
    a. Errors are normally distribured here with mean 0.
    b. So everything seems to be fine .
    c. Actual and Predicted result predicting almost the same pattern so this model seems to be correct
    d. Error terms are also independent from each other
    e. R2 same as we obtained from our final model
2. **R-Squared value for test predictions :**

a. R2 value for predictions on test data (0.815) is almost same as R2 value of train data(0.818).
b. This is a good R-squared value, hence we can see our model is performing good even on unseen data (test data)

3. **Homoscedacity:**

   Variance Residual is quite constant across the predictions Which meant to say that the value of the predictor variable does not changes as the error term changes or vary.

4. **Plot Test vs Predicted test values**

   Predictions of test data is quite very close to actual

5. **Plot Error Terms for test data**

   Error terms are randomly distributed as no pattern found which means the output is explained well by the model and there are no other parameters that can explain the model much better

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

Answer 5. Top 3 features are as follows :

1. yr(positive correlation)
2. temp(positive correlation)
3. weathersit(negative correlation)

GENERAL SUBJECTIVE QUESTIONS

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

Answer 1.

Linear regression is one of **the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables**. In the case of linear regression , as you can see the name suggests linear, **that means the two variables which are on the x-axis and y-axis should be linearly correlated.** Linear regression is a statistical regression method **used for predictive analysis and shows the relationship between the continuous variables.** Linear regression shows the linear relationship between the **independent variable (X-axis) and the dependent variable (Y-axis).**

If there is a single input variable (x), such linear regression is called **simple linear regression.** And if there is more than one input variable, such linear regression is called **multiple linear regression.**

The linear regression model gives a sloped straight line describing the relationship within the variables. The linear regression model can be represented by the following equation:

y= a0+a1x+ ε

The linear regression model provides a sloped straight line representing the relationship between the variables:

y= Dependent Variable (Target Variable)

x= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

The goal of the linear regression algorithm is to get the best values for a0 and a1 **to find the best fit line**. The best fit line should **have the least error means the error between predicted values and actual values should be minimized.**

The **cost function** helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points. The cost function is used to find the accuracy of the mapping function that maps the input variable to the output variable.

This mapping function is also known as the **Hypothesis function**.

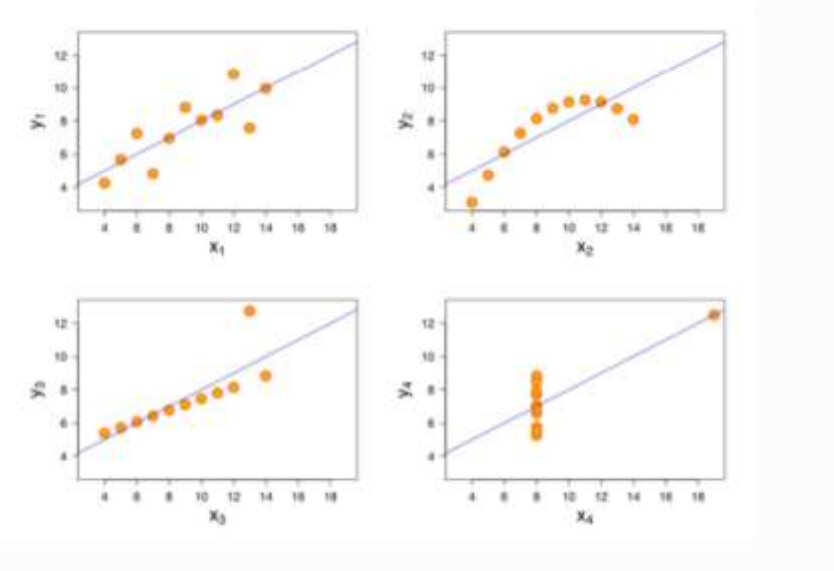
In Linear Regression, **Mean Squared Error (MSE) cost function** is used, which is the average of squared error that occurred between the predicted values and actual values.

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

Answer 2.

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 is known as **the Anscombe's quartet**.

1. It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.
2. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.
3. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.
4. They have very different distributions and appear differently when plotted on scatter plots.



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

Answer 3.

Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data.

It is **the covariance of two variables, divided by the product of their standard deviations; thus**, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

• r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

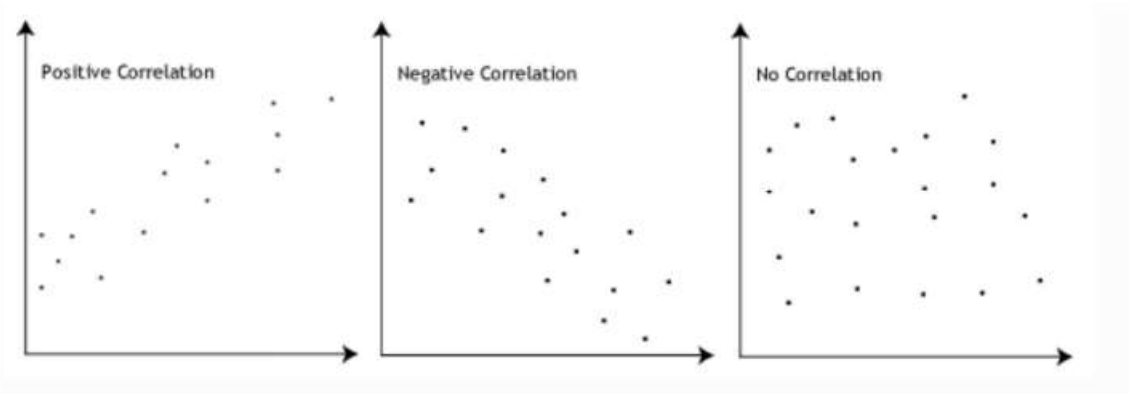
• r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

• r = 0 means there is no linear association

• r > 0 < 5 means there is a weak association

• r > 5 < 8 means there is a moderate association

• r > 8 means there is a strong association



Pearson r Formula Here,

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

• r =correlation coefficient
• xi =values of the x-variable in a sample
• $\bar{x}$=mean of the values of the x-variable
• yi =values of the y-variable in a sample
• $\bar{y}$ =mean of the values of the y-variable

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

Answer 4.

Scaling 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.

When we collect data set it 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 must do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling: • It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization 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).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

• One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Question 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer 5.

If there is perfect correlation, then VIF (Variance Inflation Factor) = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) = infinity. To solve this problem , we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Question 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer 6 .

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.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.

A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.