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)

Please find the below points inferred from the categorical variables:
1. The demand of shared bikes is more in fall season and 2019 year.
2. The sales of shared bikes is more in Aug, Sep and Oct
3. The sales of shared bikes is more when the weather situation is clear.
4. The average sales of shared bikes in weekdays is quite close, only key difference is in the lower 25 quartile.

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

This option will help in removing one of the source of multicollinearity, the dropped category can be used as a reference category and its presence or absence can be inferred from the remaining categories, for example: let's say if we have a categorical column with values as "high", "low" and "medium", and hence the dropping of "low" value means the presence of "high" or "medium" can be inferred for the remaining rows.

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

with the target variable? (1 mark)

Registered

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

training set? (3 marks)

- We did the residual analysis of the error terms, and it comes out these are normally distributed with mean centred to 0.
- We have also plotted the predicted vs residuals graph to visualize the trends, and also the residuals are randomly distributed having constant variance.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- Year
- Temperature
- Winter Season

General Subjective Questions

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

Linear regression is a technique followed when the target variable is continuous or numeric in nature, and it follows a linear equation with the independent variables also called as features like:

y = beta0 + beta1x1 + beta2x2+...........+ betanxn + error

Where beta0, beta1, beta2 are the coefficients of the linear equation which implies that if there is beta1 change in one variable, and keeping other variables at constant it affects the dependent or the y variable by beta1 units.

Assumptions of Linear Regression:

For a model to be more fitted and logically working, it needs to follow following criteria:

- There should be linear relationship between X and y.
- The residual error terms between actual and predicted must follows normal distribution.
- The Error terms are independent of each other.
- The Error terms have constant variance.

Model Estimation:

The coefficients are estimated using a method such as Ordinary least square method (OLS). OLS minimizes the sum of the squared differences between the observed and the predicted values.

Model Evaluation:

There are common metrics which provides insights about the model performance or model evaluation to suggest best fitted model:

- R-squared: This metric is used to define the proportion of the variance in the dependent variables explained by the independent variables. Higher values suggest better fitted model.
- Adjusted R-squared: This metric is like R2, buts its adjusted based on the number of predictor variables.
- Root Mean Squared Error (RMSE): The square root of the mean squared differences between observed and predicted values. Its lower values indicate higher performances.

Applications of Linear Regression:

It has wider usage in the field of economics, finance and engineering. It uses for predicting sales based on advertising spend, estimating house prices based on property features etc.

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

The Anscombe's quartet is a set of 4 datasets with similar summary statistics like the mean, variance, correlation, and regression coefficients. It explains the importance of data visualization over summary statistics, as sometimes only relying on statistics summary can be misleading, and the actual trends or patterns can be visualized by plotting graphs. The Anscombe's quartet contains datasets like:

- Dataset I: This dataset forms a simple linear relationship between the x and y variables. The relationship is straightforward, with some scatter around the linear trend.
- Dataset II: This is like Dataset I; Dataset II also exhibits a linear relationship between the x and y variables. However, an outlier point significantly influences the slope and intercept of the regression line.
- Dataset III: This dataset appears to have a non-linear, quadratic relationship between the x and y variables. When plotted, it resembles a perfect quadratic curve, despite having the same summary statistics as the other datasets.

- Dataset IV: Dataset IV has one outlier data point that causes the correlation coefficient and linear regression line to change drastically compared to the other datasets.

It serves as a powerful reminder that visualizing data can provide valuable insights that may not be apparent from summary statistics alone.

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

Pearson's R is a metric used to denote the linear relationship between two continuous variables and it is denoted by (r). Its values range from -1 to +1 where,

- r=+1: This indicates there is a positive linear relationship between two variables, which means if one variable is increasing, then the other will also increase proportionally.
- r=-1: This indicates there is a strong negative linear relationship between two variables which means if one variable is increasing, then the other will decrease proportionally.
- r=0: It means there is no linear relationship between the variables.

It gives two more important information:

a) **Strength of the Relationship:** The value of r is between -1 to +1. The closer r to +1 or -1 indicates strong relationship between variables which can be positive or negative and explains the strength of the relationship.

b) **Direction of the Relationship:** The sign of r indicates whether its positively aligned or negatively aligned which means if the sign is positive the one variable is increasing then the other variable tends to increase, whereas if the sign is negative, the one variable is increasing then another variable tends to decrease. So, it denotes the direction if r is +1, both moves in the same direction, and if r is -1, then both moves in the opposite direction.

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

Scaling is a preprocessing technique used in data analysis and machine learning to transform the features of a dataset onto a similar scale. It involves adjusting the range of values of the features so that they have similar magnitudes. Scaling is performed to address issues that may arise when working with features that have vastly different ranges or units.

When we have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So, we need to scale features because of two reasons:

- Ease of interpretation
- Faster convergence for gradient descent methods

There are 2 popular ways of scaling up the features:

a) Normalized Scaling: It this technique the variables are scaled in a such a way that all the values lies between zero and 1, which is calculated using the maximum and minimum values in the data, it follows a formula as:
$x = x\text{-}x_{min} / X_{max} - X_{min}$

b) Standard Scaling: In this technique, the variables are scaled in a such way that their mean is 0 and standard deviation is 1, it follows a formula as:
$X = X - \text{mean}(X) / \text{sd}(x)$

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

(3 marks)

VIF basically helps explaining the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:

$VIF_i = 1/1-R_i^2$

Here are some scenarios where VIF might become infinite:

- **Duplicate Variables:** There might be a possibility of adding two identical variables in the model, it means there R^2 become 1 resulting in infinite VIF for the duplicated variable.
- **Linear Dependence:** When one predictor variable is a linear combination of other predictor variables in the model, it leads to perfect multicollinearity and infinite VIF.
- **Perfect Fit:** In some cases, a regression model may perfectly fit the data due to overfitting or other issues. This can also result in infinite VIF values.

When VIF becomes infinite, it indicates a severe problem with the regression model. It suggests that the predictor variables are not independent of each other, and the model may not be reliable for making predictions about the relationships between variables. In such cases, it's essential to carefully review the data and the model specification to identify and resolve the issue causing perfect multicollinearity.

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 (Quantile-Quantile) plot is a graph that helps us see if our data follows a particular pattern or distribution, like a normal distribution. It compares the distribution of our data to what we would expect if it followed a certain pattern, like being normally distributed. The following points explains the importance of Q-Q plot in linear regression:

- **Checking Assumptions:** Since the linear regression follows an assumption that the difference between the predicted and actual values follows a normal distribution, which we can verify using Q-Q plot. If our data points fall close to a straight line in the Q-Q plot, it suggests that our residuals are close to being normally distributed.
- **Spotting Problems:** Q-Q plots can help us spot any unusual patterns or outliers in our data. If there are any points that deviate significantly from the straight line, it could mean that there are problems with our regression model or that some of our data points are unusual and might need further investigation.
- **Improving Our Model:** By using Q-Q plots, we can identify areas where our model might need improvement. For example, if we notice a lot of points deviating from the straight line, we might need to reconsider our model or look for ways to better represent our data.