Assignment-based Subjective Questions

Ques 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- The season and month variable have some positive correlation with the dependent variable but we cannot infer a solid conclusion of the effect on output variable.

However on checking the boxplot we can see the **season fall** has highest no of bikes and month September also had more of of bike rentals.

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

Answer - It is important as it helps in reducing extra column created during the variable creation. Hence it reduces the correlations create among dummy variables. If we do not drop the first dummy variable it can create multicollinearity issue.

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

Answer- The feature "temp" has highest correlation with the target variable.

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

Answer- After building the model we did residual analysis on the error terms and checked below points-

- Error terms are normally distributed with mean zero
- Checked linear relationship between temp and cnt columns.
- No multicollinearity was present among predicted variables.

Ques 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- The top features are-

a) Temperature- The demand of bike increase with rise in temperature.
b) Year – The target variable value increase with increase in year, therefore it seems when pandemic is over the demand would be good for bikes
c) Season spring – The spring season came out to be really good for the bike rentals.

General Subjective Questions

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

Answer- Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour 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.

Mathematically, we can write a linear regression equation as:

$y = bx + c$

Here x and y are 2 variables on the regression line

B = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

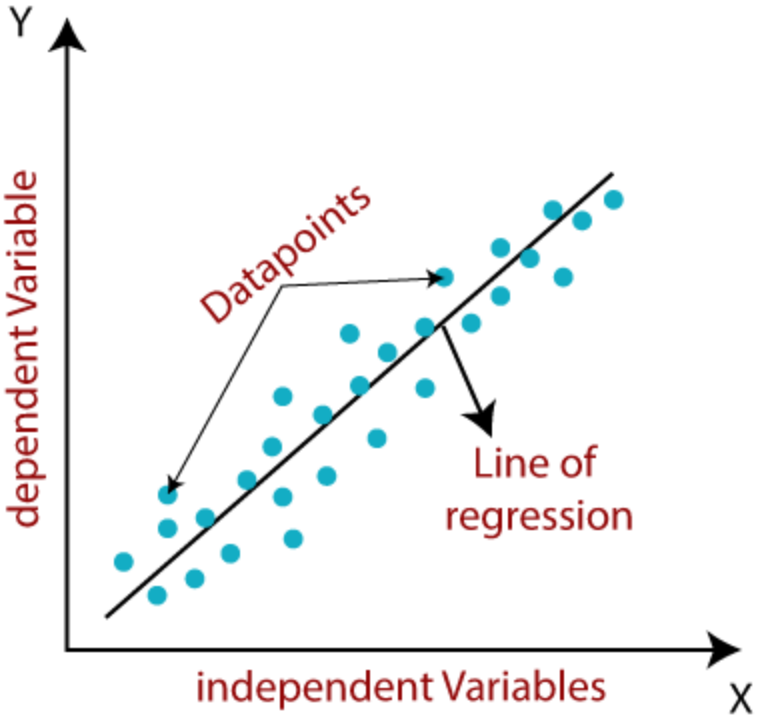
y = dependent variable from dataset.

Regression is not limited to two variables, we could have 2 or more variables showing a relationship. The results from the regression help in predicting an unknown value depending on the relationship with the predicting variables. For example, someone's height and weight usually have a relationship. Generally, taller people tend to weigh more. We could use regression analysis to help predict the weight of an individual, given their height.

When there is a single input variable, the regression is referred to as **Simple Linear Regression**. We use the single variable (independent) to model a linear relationship with the target variable (dependent). We do this by fitting a model to describe the relationship. If there is more than predicting variable, the regression is referred to as **Multiple Linear Regression**.

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

The different values for weights or the coefficient of lines (a0, a1) gives a different line of regression, so we need to calculate the best values for a0 and a1 to find the best fit line, so to calculate this we use cost function.



Cost function-

The different values for weights or coefficient of lines (a0, a1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.

Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.

We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

$$MSE = 1\frac{1}{N}\sum_{i=1}^{n}(y_i - (a_1x_i + a_0))^2$$

Where,

N=Total number of observation
Yi = Actual value
(a1xi+a0)= Predicted value.

Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If

the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Model Performance:

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by below method:

1. R-squared method:

R-squared is a statistical method that determines the goodness of fit.

It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.

The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.

It is also called a coefficient of determination, or coefficient of multiple determination for multiple regression.

It can be calculated from the below formula:

$$\text{R-squared} = \frac{\text{Explained variation}}{\text{Total Variation}}$$

Assumptions of Linear Regression

Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

a) **Linear relationship between the features and target:**
   Linear regression assumes the linear relationship between the dependent and independent variables.

b) **Small or no multicollinearity between the features:**
   Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.

c) **Homoscedasticity Assumption:**
   Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.

d) **Normal distribution of error terms:**
   Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.

It can be checked using the **q-q plot**. If the plot shows a straight line without any deviation, which means the error is normally distributed.

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

Answer-

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests 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 (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

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

Answer-

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (r) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

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

Answer-

Another important aspect to consider is feature scaling. 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:

1. **Ease of interpretation**
2. **Faster convergence for gradient descent methods**

We can scale the features using two very popular method:

1. **Standardizing**: The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

2. **MinMax Scaling**: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

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

Answer-

When there is a perfect correlation then VIF = Infinity . In case of perfect correlation we get R2 = 1 which leads to 1/1-R2 infinity.

(3 marks)

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

(3 marks)

Answer –

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. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

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.