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)

- The demand for cycles has been increasing year on year
- The demand for cycles is highest in fall and least in spring
- Fall usually occurs in October - November, hence higher demand is observed in these months
- Clear weather favours ease of cycling. Hence higher demand when weather is clear

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

The variable can be uniquely identified with just n-1 numbers. Addition of another column will only become redundant, which is why it is important to use drop_first=True

Eg., we have a categorical column which has two levels – A and B

Cat_col
A
B

A	B
1	0
0	1

A
1
0

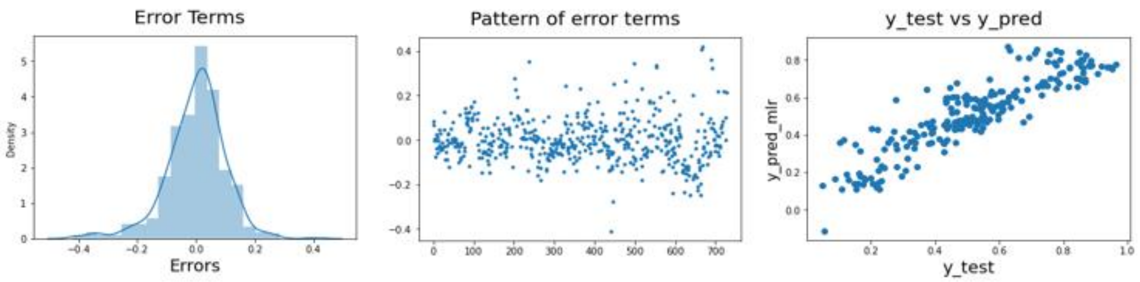
The final table which has just one column is enough to uniquely identify the variables. If A=0, which automatically means B.

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

'atemp'

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

- Checked for normal distribution of error terms (i)
- Checked for randomness in error – Showed no pattern (Homoscedasticity Assumption) (ii)
- Plotted a scatter plot for predicted vs actual, found a linear relationship (iii)
- Small or no multicollinearity between features or independent variables



i ii iii

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

- atemp
- yr
- Light_Snow

General Subjective Questions

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

- Linear regression is one of the supervised learning techniques which supports finding correlation between the variables and predicting values for continuous variables
- It shows a linear relationship between independent (x) and dependent variables (y) and fits a line through the datapoints such that the error is minimum using
    - Y = a0 + a1x; a0 : intercept, a1 : coefficient of the line
- The cost function helps in figuring the possible best values for a0 and a1 which provides the best fit line for the data points
- In Linear Regression, Mean Squared Error cost function is used, which is the average of squared error that occur between the predicted values and actual values
- To update a0 and a1 values and minimise the cost function, we make use of gradient descent which randomly selects coefficient values and then iteratively updates the values to reach the minimum cost function
- With low learning rate, the minimum cost function obtained and updated in the best fit model is used to make predictions for the continuous variable

Assumptions for linear regression to hold true –

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

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

Anscombe's quartet is a group of four datasets that are have nearly identical descriptive statistics – mean, median & variance but they have very different distributions and appear differently when plotted on a scatter plot.

It tells us the importance of data visualisation before applying any algorithm to make predictions.

For linear regression specifically, one of the assumptions is that the variables have to be linearly related. In such cases data visualization is very important to not violate the model's assumptions.

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

The Pearson's R correlation coefficient is a measure of strength of a linear association between two variables and is denoted by *r*. Pearson correlation coefficient will indicate how far away the data points are to the line of best fit.

It takes the value from +1 to -1.

- 0 indicates that there is no association between two variables.
- > 0 indicated that there is a positive association between variables, i.e., if one increases, the other also increases.
- <0 indicated that there is a negative association between variables, i.e., if one increases, the other decreases.

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

Scaling is the technique of bringing all the features/variables of the dataset into a fixed range. Most of the distance-based algorithms – linear regression, logarithmic regressions are all based on the kind/range of inputs we give. There might be cases when the units of features are different and it is not possible to compare. Scaling makes it possible to compare features of different units and not get influenced by units.

Normalized scaling – the technique re-scales features or observations with distribution value between 0 and 1

Standardized scaling – the technique rescales features or observations so that the distribution has 0 mean value and variance = 1

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

The formula or VIF = 1/(1 – r^2)

So when the variables are perfectly correlated, r value is 1 which makes the denominator 0 and VIF infinite

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

Q-Q plot (Quantile-Quantile) is a graphical tool that will help us understand if a set of data plausibly came from a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

In linear regression when we have training and test data set received separately, we can confirm using Q-Q plot that both the data sets are from populations with same distributions.