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)

Ans: Below are a few of the inferences from analyzing categorical variables:

- Season: Demand for rental bikes is highest in the Fall season
- Yr : Demand has increased in 2019
- Weekday does not show a direct relation with demand of rental bikes
- Demand increases from March - Aug and then decreases for the rest of the month
- Misty and Clear Weather has higher demand as compared to Light snowrain

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

Ans: drop_first = True is important as it helps reduce the one extra variable created during creation of dummy variables thus reducing multicollinearity amongst the features

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

Ans: temp & atemp have the highest correlation with the target variable but only one of them is kept within the model as they are highly collinear with each other

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

Ans: Methods used to validate the assumptions of Linear Regression

- Multicollinearity of predictors was reduced by keeping VIF of all the variables under 5 and ensuring p-value is close to 0 suggesting that the predictors are significant
- Residual Analysis : Distribution of error terms(difference between predicted value and actual value of target variable) seemed normal
- Homoscedasticity : Distribution of error terms with the target variable seems to have less 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)

Ans: Predictor variables that impact the most:

1. 'yr' with a coefficient +0.2480
2. 'weathersit_Light_snowrain' with a coeff of (-0.3023)

3. 'season_spring' with a coeff of (-0.2581)

General Subjective Questions

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

Linear Regression algorithm is a supervised learning algo. It is a statistical method to make predictions of continuous variables.

The linear regression model provides a sloped straight line representing the relationship between the predictors and the target variable

Y = a0 + a1*x1 + a2*x2…

Where y = target variable

X1,x2… = predictor variables

a0= intercept of the line

a1,a2 = coefficient of x1,x2

There are two types of Linear regression:

1. Simple Linear Regression: It has a single predictor variable vs target variable
2. Multiple Linear Regression: It has multiple predictor variables with target variable

The goal is to find the best fit line in linear regression that has the minimum difference of predicted and actual value of target variable.

This minimum difference can be measured through MSE (mean squared error)

Assumptions of Linear regression:

1. Features and target variables have a linear relationship
2. Error terms are normally distributed
3. Homoscedasticity : There is no variance in error terms
4. Small or no multicollinearity between the features

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

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

- In the first one, the scatter plot seems to be a linear relationship between x and y.
- In the second one, there is a non-linear relationship between x and y.
- In the third one, there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one, there is one high-leverage point that is enough to produce a high correlation coefficient.

3. What is Pearson’s R? (3 marks)

It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

It is the covariance of two variables, divided by the product of their standard deviations;

- 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
- 0< r < 5 means there is a weak association
- 5< r < 8 means there is a moderate association
- r > 8 means there is a strong association

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

It 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.

Most of the times, collected data set 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 have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

- It brings the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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)

-

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

This means there is a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.