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)
Answer:

- Demand was highest in fall and lowest in spring season
- Demand is increasing per year as its higher in 2019 as compared to 2018
- Throughout the week there is almost constant demand for bikes
- The demand is highest in the month of September
- There is a less demand on holidays
- In clear weather there is a high demand

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

Answer: If any categorical variable has 3 unique values and we were to create dummy columns to denote those values in binary fashion, at a time only one of the 3 variables will have value 1.

Which means,
If A = 1 & B = 0, C= 0
If A = 0 & B = 1, C = 0
If A = 0 & B = 0, C = 1

That means, at any point third column can be explained using other two.
drop_first=True helps in reducing the extra column created during dummy variable creation in order to reduces the correlations created among dummy variables. In this way we will always have n-1 dummy columns created. (n = number of unique categorical values)

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

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

- **Linear relationship between X and Y**: In the pair-plot there is a linear relationship between temp(X) and cnt (Y) variables. Validated the same from model coefficients where temp had coefficient value of 0.5300 and P value 0

- **Predictors are not highly correlated with each other:** VIF (Variance Inflation Factor) helps to identify the correlation between the predictor variables. Variables with VIF less that equal to 5 are considered to be good to build the linear regression model on**.**

- **(Residuals) Error terms are normally distributed (not x & y)**: In the residual analysis of train data, we calculated the difference between actual cnt from train data set and predicted cnt from model prediction result also known as Error Terms. When this difference was plotted, it resulted into a normally distributed plot centred at zero which means the error terms are normally distributed

- **(Residuals) Error Terms are independent of each other**: We plotted the residuals against the predicted values to see if there is any obvious pattern in the plot. With the absence of any pattern, it was concluded that the error terms are independent of each other.

- **(Residuals) Error Terms have constant variance:** Same plot used to identify if the error terms are independent of each other was used to check if the spread of residuals is roughly equal at every level of the predicted values and they are randomly distributed on both sides of 0. There was some variance which was not explained by the model but no obvious pattern in the variance was observed.

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:

Top 3 features explaining increase in demand for shared bikes: temp, yr, season_winter

Top 3 features explaining decrease in demand for shared bikes: windspeed, weathersit_Light, holiday

General Subjective Questions

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

Answer : Linear Regression is a supervised machine learning algorithm used to predict the continuous values of a variable based on one or more independent variables. Predictors are independent variables and the result is dependent variable.
The goal of linear regression is to find the best-fitting line for the data for which total prediction error is as small as possible.

For example,
If, weight loss depends on total number of calories burned. In this case, having known the actual values for both independent variable (calories burned) and dependent variable (weight loss) can be used to predict the weight loss for different set of values for burned calories considering there is a linear relationship between them. Prediction can be defined as a straight line denoted by following equation

$$yPred = b0 + b1x$$

b0 and b1 here are coefficients of predictor variable x.

The values for b1 and b0 needs to be in such way that there is minimum error between predicted value and actual value
If $b1 > 0$ then the predictor and prediction have a positive relationship whereas if $b1 < 0$ then the relationship is negative
b0 is the line intercept. It guarantees that the residuals will have mean zero which is one of the assumptions of linear regression model. If there is no b0 then the regression line will pass through origin and the prediction will become biased.
To reach the optimal values for b0 and b1 gradient decent method can be used where we use different values for b0 and b1 until we get the line that passes to maximum number of data points.

Null-Hypothesis and P-value
Null hypothesis is the initial claim that researcher specify using previous research or knowledge.
Low P-value: Rejects null hypothesis indicating that the predictor value is related to the response
High P-value: Changes in predictor are not associated with change in target

Linear model performance can be evaluated using R squared and sum of squared error value.
R-Square have values between 1 and 0 where 1 indicates that all the variability in the prediction is completely explained by the predictor and 0 indicates that the predictor variable does not explain any variation in the prediction.
Sum of squared error defines how much the actual value varies around the predicted value. It is denoted by following equation

Error = sum(actal y – predicted y)^2

There are some assumptions associated with linear regression

- There is a linear relationship between the variables
- Predictor variables are not highly corelated with each other
- Error terms are normally distributed
- Error terms are independent of each other
- Error terms have constant variance at all the levels of predictors

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

Answer: Anscombe's quartet are 4 datasets created by statistician Francis Anscombe.
All these 4 datasets have identical statistical summary such as Mean, Standard Deviation, Correlation Coefficients etc. But when we represent these datasets as scatter plots, they have completely different appearances.
When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.
It is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

Answer: Pearson's r measures the linear association between two continuous variables.
It quantifies the strength and direction of the linear relationship.
The value of r ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation.
RFE uses Pearson's R to assess the individual effect of each predictor variable on dependent variable to assign weights and select the features

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

Answer: Most of the times the continuous values in the dataset will have different scales. Scaling is used to normalise those values and transform them to the similar scale.

Scaling ensures that all features contribute equally to the model. Without scaling, features with larger values might dominate the model, leading to biased results.

Normalized (Min – Max) Scaling: It transforms all the features within the range of 0 – 1. Useful when there are no outliers, as it cannot handle them effectively.

Standardized Scaling: It transforms the features to have values that result in mean 0 and standard deviation 1. It handles outliers better than min max scaling.

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

Answer: When VIF is infinite, it means that the variable can be perfectly explained by the other variables in the model. It may occur if,

- The dataset has strongly correlated variables
- If there are a lot of variables in the dataset but very few observations

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

Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data 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.

It is used to check if two data sets —

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behaviour

In Q-Q plot, the quantiles of the first data set are plotted against the quantiles of the second data set. If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis then the two datasets have similar distribution and vice a versa.