Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Answer 1: Here many insights can be drawn from the plots:
1. Fall season has highest demand for rental bikes.
2. Increase in trend when it is comes to year on year. (2018 vs 2019)
3. Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing.
4. When there is a holiday, demand has decreased. as we see median value decreased
5. The good weathersit has highest demand.

Question 2: Why is it important to use **drop_first=True** during dummy variable creation?
Answer 2:

drop_first=True is important to use, as it helps in reducing the extra columns that gets created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Dropping the first columns as (p-1) dummies can explain p categories.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer 3: Looking at the pair-plot among the numerical variables, temp and atemp have the highest correlation (0.63) with the target variable (cnt).

Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer 4:

Residual Analysis:

Errors are normally distributed with a mean of 0. Actual and predicted result follow the same pattern. The error terms are independent of each other

R2 value for test predictions:

R2 value for predictions on test data (0.785) is almost in line with R2 value of train data (0.812). This is a good R-squared value; hence we can see our model is performing good even on unseen data (test data)

Homoscedasticity:

We can observe that variance of the residuals (error terms) is constant across predictions. i.e., error term does not vary much as the value of the predictor variable changes.

Plot Test vs Predicted value test:

The prediction for test data is very close to actuals.

Question 5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer 5: The top 3 features are:

1. yr
2. temp
3. weathersit

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.
Answer 1:

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.
When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.
Similarly, when there is only one dependent variable, it is considered Univariate Linear Regression, while when there are more than one dependent variables, it is known as Multivariate Regression.

Types of Linear Regression
There are two main types of linear regression:

Simple Linear Regression
This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:
$Y = \beta 0 + \beta 1X$
where:
Y is the dependent variable
X is the independent variable
$\beta 0$ is the intercept
$\beta 1$ is the slope

Multiple Linear Regression
This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:
$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \dots\dots\dots + \beta nXn$

where:
Y is the dependent variable
X1, X2, …, Xn are the independent variables
$\beta 0$ is the intercept
$\beta 1, \beta 2, \dots, \beta n$ are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

Question 2: Explain the Anscombe’s quartet in detail.
Answer 2:

Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. 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.

Question 3: What is Pearson's R?
Answer 3:

The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

- Pearson's r
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

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.

Question 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer 4:

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.

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-squared, etc.

Normalization/Min-Max Scaling:

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

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

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).

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

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer 5:

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

Answer 6:

The quantile-quantile (q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in linear regression to identify distributions.