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 1=>

   - The demand of bike is less in the month of spring when compared with other seasons
   - The demand bike increased in the year 2019 when compared with year 2018.
   - Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
   - Bike demand is less in holidays in comparison to not being holiday.
   - The demand of bike is almost similar throughout the weekdays.
   - There is no significant change in bike demand with working day and non-working day.
   - The bike demand is high when weather is clear and few clouds however demand is less in case of Light-snow and light-rainfall. We do not have any data for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog so we cannot derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   Answer 2 =>
   - When you have categorical variables with say ‘n’ levels the idea of dummy variable creation is to build ‘n-1’ variables indicating levels. For a variable say ‘relationship’ with 3 levels namely “single” married””/in a relationship ”

Relationship status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

   - You can clearly see that there is no need of the defining 3 different levels ,If you drop a level, say single you would still be able to explain the 3 levels

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Answer 3 =>
   - Looking at the pair plot seems Temp Variable has highest Linear relationship with Cnt variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Answer 4=>
   - Firstly we checked for the Linear relationship between all independent Variables and Target variable using Pair plot, we found out that some of their variables are linearly related with Target Variable
   - Then we checked if the Error term is normally Distributed with mean Zero using displot, which came out that they are normally distributed with mean 0

- Then We ensured that Error terms are independent of Each other
- Then we also ensured that Error terms have constant variance(homoscedastic)

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 5 =>
   - Holiday
   - Temp
   - Windspeed

General Subjective Questions

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

Answer 1=>

- The Linear Regression Algorithm tries to explain the relationship between independent and dependent variable using a straight line. It is applicable=le to numerical variables only. Following steps are performed while doing linear regression
- The Dataset is divided in to test and training data
- Train data is divided into features and target dataset
- A linear model is fitted using the training dataset. Internally the api's from python is gradient descent algorithm to find the coefficients of the best fit line. The gradient descent algorithm works to minimizing the cost function .A typical example of cost function is residual sum of squares
- In case of multiple features, the predicted variable is a hyperplane instead of line. The predicted variable takes the following form
- Y=B0+ B1x1+B2x2+B3x3+….+Bnxn
- The predicted variable is than compared with test data and assumptions are checked

2. Explain the Anscombe's quartet in detail.
   Answer 2=>

- Ascombe'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.
- ii> What Is the Purpose of Anscombe's Quartet in Data Visualization? => 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 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.

3. What is Pearson's R? (3 marks)
   Answer 3=>
   - 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

Pearson correlation coefficient (r)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is no relationship between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

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

   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.
   - b) Why scaling is performed?
     - ❖ 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.
   - c) The Difference between Normalization and Standardization scaling
     - ❖ 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.
     - ❖ 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)

   Answer 5=>

   - The Formula for VIF is VIF(i)=1/1-Rsquare(i).Basically if R square is 1 then VIF becomes infinite. It means that there is perfect correlation between features

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

(3 marks)

Answer 6=>

- The QQ 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.
- Uses=>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. Q–Q plots can be used to compare collections of data, or theoretical distributions.
- Importance in Linear Regression=> in 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.