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: Using bar plot I am able to compare how change in one variable affects dependent variable. Here are some inferences.

1. With using BoxPlot it is clear that , there is no much outliers in the given data set.
2. During Fall and Summer seasons, bike rental demand is more. It is also observed that rentals are more in Fall and Summer seasons of both years 2018 and 2019.
3. Demand is increased in 2019 when compare to 2018.
4. Count is more that 5000 during the months from May to October. Also highest bikes are counted in 2019 September.
5. Count in all weekdays is nearly same.
6. More rental bikes if whethersit is clear.

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

Ans: drop_first=True will remove the first dummy variable after creation, it is important to remove as we can represent n levels with n-1 dummy variable there is no need of extra dummy variable to build a model. If we miss to delete there will be a redundancy and multicollinearity get introduce which affects computation of coefficients.

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

Ans: temp

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

Ans: Firstly we can evaluate the model by examining difference between actual observed target values with predicted values. This difference is called as Residuals. Residuals can be analysed and interpret our assumptions with

1. Histogram plot. If histogram plot shows residual points are bell structure it means residuals are normally distributed.
2. Q-Q plot. If residual points are falling approximately on standard normal distribution points, indicates residuals are normally distributed.
3. Residuals vs fitted plot. This assesses if residuals are randomly scattered around 0 (there is no pattern) indicates model is appropriate.

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:

1. Temp
2. Light snow rain
3. Windspeed

General Subjective Questions

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

Ans: Linear regression algorithm well fit for continuous variables with linear relationship. Linear relationship means that dependent variables are gets change when there is a change in independent variable.

Algorithm:
Step 1: Understanding and Data visualization.
Step 2: Data Preparation.
Step 3: Data Split into Train data and Test data.
Step 4: Scaling Train data using MinMax scaling for 0/1 values.
Step 5: Building a linear model.
   a. Pop train target/dependent variable to y axis (y_test) and remaining independent variables to (x_test)
   b. Apply Automated approach RFE (Recursive Feature Elimination).
   c. Adding constant to X train linear model
   d. Applying Ordinary Least Squares.
   e. Evaluating P-Value and VIF for Multicollinearity in the model.
Step 6: Evaluating Model - Residual Analysis
   a. Error terms are normally distributed.
   b. Error terms are centred at zero.
   c. Examine difference between actual observed value vs predicted value

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

Ans: Anscombe's quartet is also known as A Tale of Four Datasets. Each consisting of 11 data points (x,y). While these datasets appear remarkably similar at first glance but they exhibit different statistical points when visualised.

Dataset		Key statistical properties (Example)	Reveals distinct patterns when visualised
1	A simple linear relationship between x and y.	Mean of x,y – [9, 7.5], Variance of x,y - [10, 4], Correlation coefficient – 0.816	A clear linear relationship between x and y.
2	A quadratic linear relationship between x and y, with single outlier.	Mean of x,y – [9, 7.5], Variance of x,y - [10, 4], Correlation coefficient – 0.816	A quadratic relationship with outlier pulling the line away from majority of points.
3	A linear relationship between x and y with constant x value for most points and a single outlier	Mean of x,y – [9, 7.5], Variance of x,y - [10, 4], Correlation coefficient – 0.816	A vertical line with an outlier pulling correlation coefficient towards 1.
4	A linear relationship between x and y with two outliers.	Mean of x,y – [9, 7.5], Variance of x,y - [10, 4], Correlation coefficient – 0.816	A linear relationship with two outliers, one pulling line towards up and another pulling line towards down.

Anscombe's quartet highlights the importance of data visualization in understanding the underlying relationships within data.

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

Ans: Pearson's R is a statistical measure that quantifies the linear relationship between two variables it ranges from -1 to 1.
r = 1 Perfect positive relation meaning variables increase or decrease together perfectly.
r = -1 Perfect positive relation meaning one variable increases as the other variable decreases perfectly.
r = 0 No correlation between the variables.

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

Ans: Scaling is a technique to transform the data within the fixed range. It is a crucial step which ensures that all variables contribute equally for model learning process.
Dominant features: When feature have significantly different scales, algorithms might give undue weight for larger values. This can lead to biased modelling.
Fair Contribution: Scaling ensures that all features contribute to the model's decisions based on their relative importance, rather than their magnitude.

Standardization, Scaling can standardize features to a common scale, making it easier to compare and interpret results.

Normalized Scaling, Scales features to a specific range (Ex: 0 to 1).

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

Ans: If the VIF value is infinite means there is a multicollinearity exists in the model. It means that predictor variable is highly correlated with other predictors which can lead to unstable and unreliable regression models. This can also happen if there is a numerical instability.

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

Ans: Quantile-Quantile plot is a visual comparison of distributions. It is particularly useful for assessing whether datasets follows a specific theoretical distribution (eg: Normal distribution, Uniform distribution, Exponential distribution).

Uses:
Normality Testing: To assesses if dataset follows a normal distribution.
Comparing Distributions: To compare two datasets or a datasets with a theoretical distribution.
Outlier Detection: To identify outlier in the dataset.

Interpretation:

Straight Line: If the points fall approximately on a straight line, it suggests that the two distributions are similar in shape.

Deviation from Line: Deviations from the line indicate differences in the distributions. For example, a curve shape might suggest a different distribution or the presence of outliers.