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:

   - The demand for shared bikes in very low in spring season and highest in the fall season.
   - Compared to 2018, there is approximately 50% increase in the demand for bikes in 2019.
   - The demand for shared bikes increases gradually from January to September and then decreases till December.
   - Looking at the 25th and the 50th percentile, the demand for shared bikes is more on working days than holidays.
   - When the weather is clear the demand for bikes is more and it is the lowest when there is light rain or snow.

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

   Answer:

   - When creating dummy variables for a categorical column, when there are 'n' different values that the column has, then it is enough if we create 'n-1' dummy variables as all the 'n' different values can be represented using the 'n-1' dummy variables.
   - Eg: say a categorical column has A, B, C as the unique values. We create 'n-1' dummy variables: A & B and we omit C. The different values for these dummy columns are:

A	B
0	0
0	1
1	0

   - 00 represents C, 01 represents B and 10 represents A. So, it is enough if we have 'n-1' dummy variables.
   - Programmatically this is achieved in Pandas using the *drop_first=True* argument passed to the *get_dummies* method.

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

   Answer: The variables 'temp' (Temperature) and 'atemp' (Feeling temperature) has the highest correlation of 0.63 with the target variable (count).

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

   Answer:
   - Error terms are distributed normally: plotted a histogram of the error terms.
   - residual terms are independent: no auto-correlation: Durbin-Watson score calculated.
   - Homoscedasticity – no visible pattern in the residuals: plotted scatter plot of predicted values vs residuals.
   - No Multicollinearity – no correlation among the independent variables: computed VIF for all the selected variables and plotted a heatmap of the correlation of the variables.
   - Linear Relationship – independent variables must have linear relationship with the dependent variable: plotted ccpr plots for all the selected columns.

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:
   - temp has the highest correlation with the target variable (+0.64).
   - yr has the second highest correlation with the target variable (+0.54).
   - Spring season has the negative correlation of -0.54.

General Subjective Questions

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

   Answer:

   - Linear Regression supervised learning technique in machine learning. It involves building a model using some features (called the independent variables) to make predictions on a continuous variable (called as the dependent / target variable).
   - Linear regression tries to identify a linear relationship (positive or negative) between the independent and dependent variables and uses this to make future predictions.
   - Linear regression can be broadly classified into:
     - Simple linear regression – Only one dependent variable.
     - Multiple linear regression – two or more dependent variables.
   - Example: Let us consider a simple linear regression model with the following equation:

     $$y = mx + c$$

     Where y: dependent variable, x: independent variable, m: slope of regression,

     c: y intercept (constant).

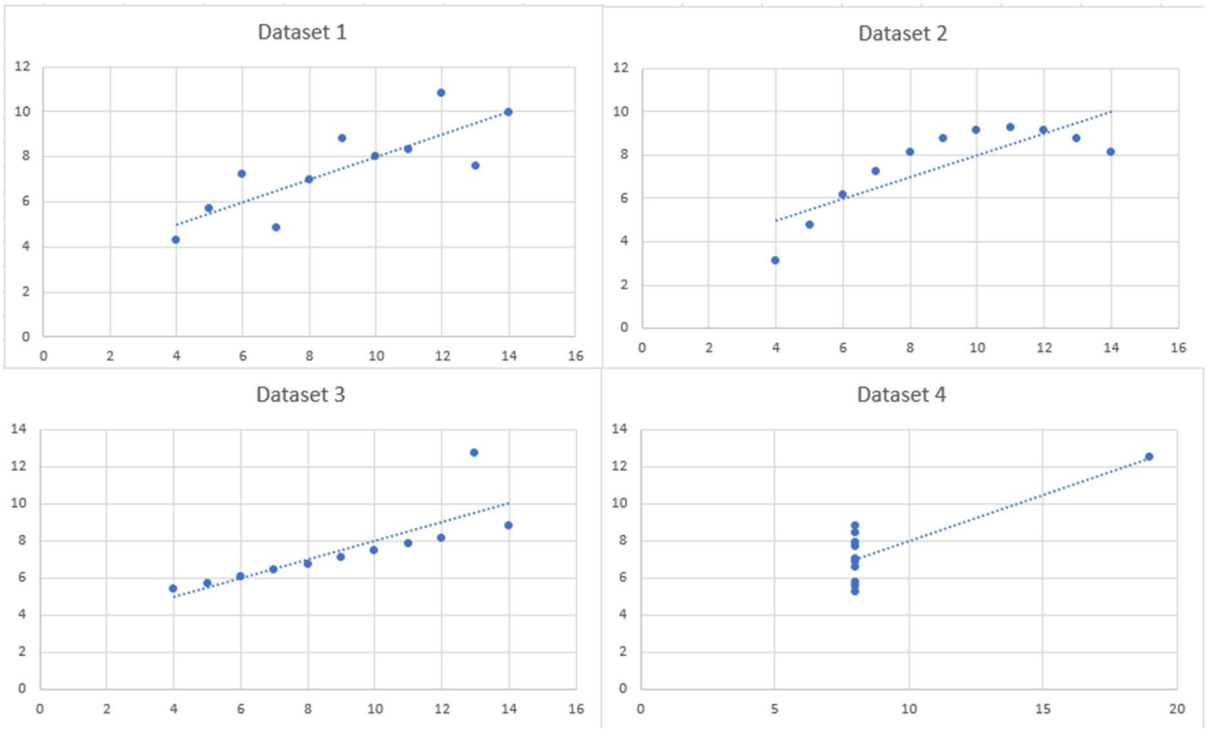
     Linear regression ultimately boils down to finding the optimal value of ‘m’ which will reduce the model’s errors on the prediction.

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

Answer:

- The Anscombe's quartet was constructed by statistician Francis Anscombe in 1973.
- It illustrates the importance of plotting the data and not blindly believing in summary statistics before building a model.
- The Anscombe's quartet consists of 4 dataset that have very similar summary / descriptive statistics (similar mean, standard deviation, etc.) but when the data is plotted, they are very different.

	Anscombe's quartet							
	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
	Summary Statistics							
Count	11	11	11	11	11	11	11	11
Mean	9	7.500909	9	7.500909	9	7.5	9	7.500909
SD	3.316625	2.031568	3.316625	2.031657	3.316625	2.030424	3.316625	2.030579
Sum	99	82.51	99	82.51	99	82.5	99	82.51



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

Answer:

- Pearson's R or the Pearson correlation coefficient measures the linear correlation between two variables in a dataset.
- The Pearson correlation coefficient can range between [-1, 1].
- Here, -1 represents a strong negative correlation, 0 represents no correlation and 1 represents a strong positive correlation.
- The Pearson's correlation coefficient is fount out using this formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

Answer:

- Scaling is done before the dataset is fed into a machine learning model.
- Often data in the real world will be present in different ranges and magnitudes. If this data is fed to the ML model directly, it might take a long time for the model to converge on the optimal model parameters.
- Scaling helps the model to learn and understand the relationships in the data easily.
- Difference between normalized scaling and standardized scaling:

Normalized Scaling	**Standardized Scaling**
Formula: $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$	Formula: $z = \frac{x - \mu}{\sigma}$
Uses the minimum and maximum values of the feature.	Uses the mean and standard deviation of the feature.
Useful when the feature's distribution is unknown.	Useful when the feature is normally distributed.
Outliers greatly affect this as it depends on the min and max values.	Outliers do not affect much.
It scales values between [-1 to 1] or [0 to 1].	The scaled values are not bounded to a fixed range.

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

Answer:

- When the VIF is infinite, it means that there is a perfect (100%) correlation between the current variable and all the other variables.
- It means that the variable is completely explained by some combination of the other variables.
- This is perfect multicollinearity and the variable has to be dropped.

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

(3 marks)

Answer:

- Q-Q plot or the Quantile-Quantile plot is used to compare two different probability distributions by plotting their quantiles against each other.
- It helps us to assess if the data came from the same distribution or not.
- A Q-Q plot is a scatter plot where we plot the quantiles of the one dataset and quantiles of the other dataset against each other. If their quantiles have come from the same distribution, the plot would resemble a rough line.