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)

In the above box plots, a few variables stand out as good univariate predictors of quality.

- weathersit, season, yr: huge difference between types so low correlated with each others
- holiday, workingday: the value range is not clear so maybe they are correlated with each others

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

Whether to get k-1 dummies out of k categorical levels by removing the first level. Because for a multiple value variable we just need to keep k-1 levels

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

- temp is highest correlation variable with the target variable

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

- We are based on R-squared and Adj. R-squared to define model is better or not
- Keep in mind the way to select features
    - High p-value: insignificant - High VIF
    - High - Low:
        - High p-value - low VIF: remove these first
        - Low p-value - high VIF: remove after the above
    - Low p-value and low VIF

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- Based on final model top three features contributing significantly towards explaining the demand are:
    1. Temperature: 0.493356
    2. Weathersit : - 0.31 (-0.25xlight_snow -0.06xmist)
    3. year (0.23)
    4.

General Subjective Questions

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

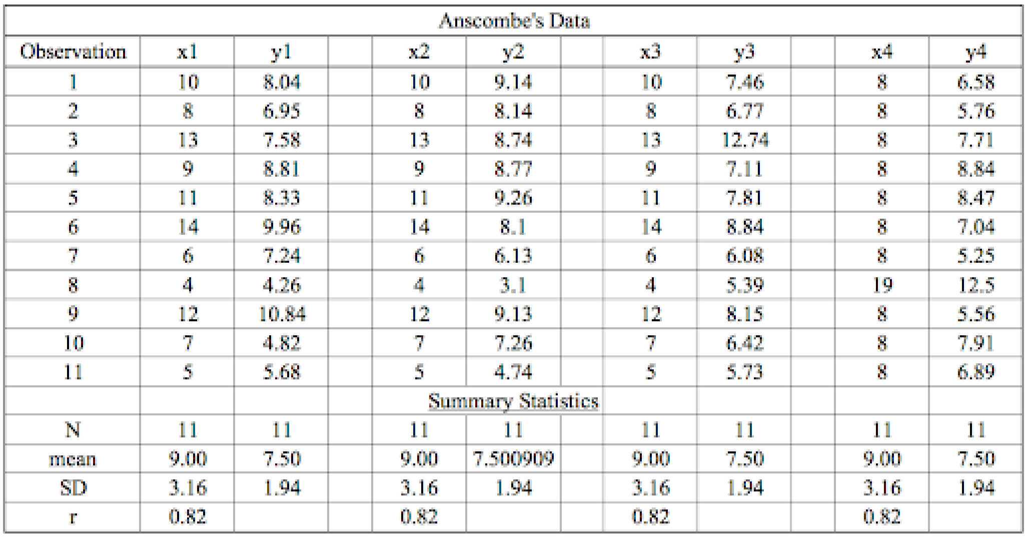
- Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.
- The equation for multiple linear regression is similar to the equation for a simple linear equation, i.e., y(x) = p0 + p1x1 plus the additional weights and inputs for the different features which are represented by p(n)x(n). The formula for multiple linear regression would look like.

$$y(x) = p0 + p1x1 + p2x2 + \ldots + p(n)x(n)$$

- Types of linear regression:
    - Simple linear regression: simple linear regression reveals the correlation between a dependent variable (input) and an independent variable (output)
    - Multiple linear regression: Multiple linear regression establishes the relationship between independent variables (two or more) and the corresponding dependent variable. Here, the independent variables can be either continuous or categorical.

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

- Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets.
- 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, linear separability of the data, etc.)
- Four datasets have the same statistic values:

- 

Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

- Anscombe's quartet four datasets:
    - Data Set 1: fits the linear regression model pretty well.
    - Data Set 2: cannot fit the linear regression model because the data is non-linear.
    - Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
    - Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.
- Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm

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

- 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.
- 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.

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

- Feature scaling is the process of normalizing the range of features in a dataset.

- Scaling is performed because real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.
- Normalization typically rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

- The VIF is equal to 1 if the regressor is uncorrelated with the other regressors, and greater than 1 in case of non-zero correlation.
- The greater the VIF, the higher the degree of multicollinearity.
- In the limit, when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity.

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

- A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight