Assignment-based Subjective Questions.

1. From your analysis of the categorical variables from the dataset, whatcould you infer about their effect on the dependent variable?

Answer: Here are a few conclusions I came to after studying categorical data from the dataset that related to the dependent variable (Count).

1. Autumn has the highest median, which is to be expected given that the weather is best for biking, followed by summer.
2. Median bike rentals are rising year over year, with 2019 having a higher median than 2018, which may be related to the growing popularity of bike rentals and people's increased environmental consciousness.
3. Because fall months have a larger median, the overall spread in the month plot reflects the season plot.
4. People rent more frequently on non-holiday days than on holidays, which could be because they prefer to spend time with their families and use their own vehicles rather than renting bikes.
5. While the overall median for all days is the same, the spread on Saturday and Wednesday is larger, as may be seen.
6. The median for working and non-working days is about the same, but the dispersion is larger on non-working days since some people may have plans and not wish to rent bicycles on those days.
7. Clear skies are best for renting bikes because they are temperate, have little humidity, and are cooler in temperature.

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

Answer: Dummy variables with the value n-1 can represent a variable with n levels. So, even without the first column, we can still express the data. If the first variable has a value of 1, then the variables from 2 to n must also have a value of 0.

For instance, if the term "Relationship" has three levels, "Single," "In a Relationship," and "Married," I would make a mock table looking something like this:

Relation Status	Single	In A Relationship	Married
Single	1	0	0
In A Relationship	0	1	0
Married	0	0	1

However, it seems obvious to me that there is no need to distinguish between three levels. Even if I dropped a level, let's say "Single," I could still describe the three tiers.

Dropping the dummy variable "Single" from the columns will reveal how the table will appear:

Relation Status	In A Relationship	Married
Single	0	0
In A Relationship	1	0
Married	0	1

If the dummy variables 'In a Relationship' and 'Married' are both equal to 0, the subject is unmarried. Finally, if "In a relationship" is zero and "Married" is one, the individual is married. If "In a relationship" is one and "Married" is zero, the person is in a relationship.

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

   Answer: The word "temp" showed the strongest correlation, 0.63.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   Answer: by means of a residuals distribution diagram. It has a normal distribution and a mean value of zero, according to the findings.

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

   Answer: The following are the top three qualities that significantly contribute to the need for shared bikes:

   - atemp (0.412)
   - yr (0.236)
   - weathersit (-0.275) Light rain

.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: The TA linear regression algorithm uses a straight line to attempt to explain the relationship between the independent and dependent variables. Only numerical variables are applicable.

When performing linear regression, the following actions are taken:
• Test and training data are separated from the dataset.
• Target (dependent) and features (independent) datasets are separated from the train data.
• The training dataset is used to fit a linear model. The gradient descent algorithm is used internally by the Python APIs to determine the coefficients of the best fit line. The cost function is minimised by the gradient descent process. Residual sum of squares is a common instance of a cost function.
• When there are numerous characteristics, a hyperplane is projected as the variable rather than a line. The anticipated variable looks like this:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x$$

 • The anticipated variableThe predicted variable is than compared with test data and assumptions arechecked.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet consists of four data sets with virtually similar simple descriptive statistics, but when represented graphically, the distributions are very different.
 The mean, sample variation of x and y, correlation coefficient, linear regression line, and R-Square value make up the simple statistics.
 Anscombe's Quartet demonstrates how graphing can nevertheless reveal significant differences between numerous data sets with many comparable statistical features. The charts are displayed below:
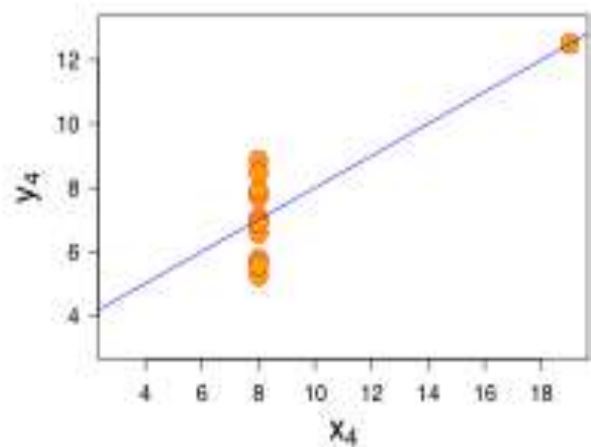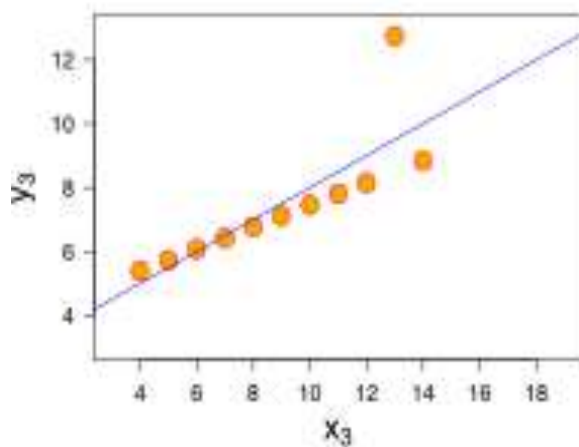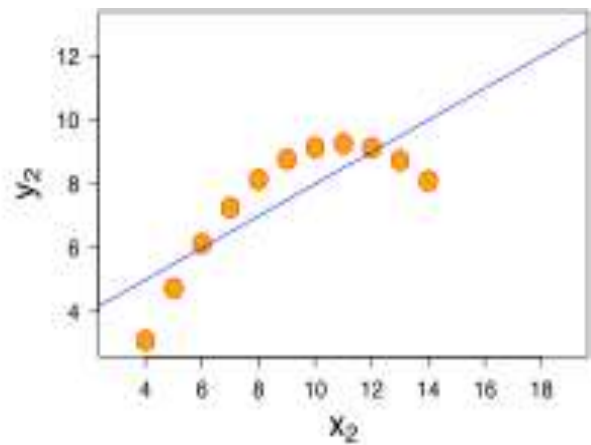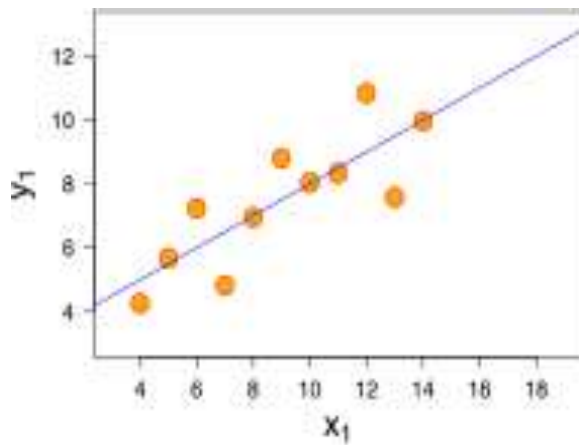
1. The first plot (top left) seems to represent a straightforward linear relationship.

2. The correlation coefficient is useless because the second figure (top right) depicts a nonlinear relationship and is not normally distributed.
3. The third plot is linear but uses a different regression line (bottom left). This is taking place as a result of the outliers in the data.

4. The fourth plot (bottom right) does not demonstrate a linear relationship, but the data were changed because of outliers.

To put it simply, it is better to visualise data and eliminate outliers before examining it.

3. What is Pearson's R?

Answer: The strength of a relationship between two variables is measured by Pearson's R.
It is calculated by dividing the covariance of two variables by the sum of their standard deviations. Its range of values is +1 to -1.

• A value of 1 denotes a complete linear positive correlation. It implies that if one variable rises, the others will follow suit.

• Zero indicates there is no association.

• A score of -1 indicates a completely negative association. It implies that if one variable rises, another will fall.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: To keep a variable within a specified range, scaling is used. In a linear regression study, scaling is a pre-processing step. To speed up the computation of gradient descent, we scale a variable.
The gradient descent process will take a very long time if the data contains both small variables (values in the range of 0–1) and big variables (values in the range of 0–1000).
The step size of gradient descent is typically modest for precision.

Normalised Scaling	Standardized scaling
Called min max scaling, scales the variable such that the range is 0-1	Values are centred around mean with a unit standard deviation
Good for non- gaussian distribution	Good for gaussian distribution
Value id bounded between 0 and 1	Value is not bounded
Outliers are also scaled	Does not affect outliers

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

Answer: The formula for VIF is

$$\text{VIFi} = \frac{1}{1 - R_i^2}$$

Basically, VIF becomes limitless if R square is 1.
It indicates that the characteristics perfectly align with one another.

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

Answer: A Q-Q plot is a scatter plot that compares two sets of quantiles.
To determine whether the two sets of data came from the same distribution is its goal.
Data is being visually checked. If all of the data are from the same source, the plot will seem like a line.