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?**

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

- Summer is the most popular season for renting bikes, followed by spring and winter, which is expected given that the weather is ideal for motorcycling.
- Additionally, the median number of bike rentals is increasing year over year, with 2019 having a higher median than 2018, potentially due to the growing popularity of bike rentals and people's increased environmental consciousness.
- Furthermore, the data suggests that the larger median for fall months is reflected in the overall spread in the month plot, which is likely due to the seasonal weather changes. In terms of rental patterns, people tend to rent more frequently on non-holiday days than on holidays, which could be attributed to a preference for spending time with family and using personal vehicles.
- Interestingly, while the overall median for bike rentals is consistent across all days, the spread of rentals is larger on Saturdays and Wednesdays.

- Finally, the data indicates that clear skies are optimal for renting bikes as they provide ideal temperate conditions with little humidity and cooler temperatures.

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 somethinglike 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 thethree tiers.

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

Relation Status	**In A Relationship**	**Marrie d**
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, thesubject 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 term “temp” exhibited the highest correlation with a coefficient of 0.63 with target variable “cnt”.

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

Answer: Based on the results, it can be observed from the residuals distribution diagram that the distribution is normal with a mean value of zero.

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:

- September
- Yr
- July

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: The linear regression algorithm uses a straight line to attempt toexplain 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 fromthe 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 ofthe 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 thevariable rather than a line. The anticipated variable looks like this:

$$Y= \beta 0+\beta 1x1+\beta 2x2+\beta 3x3+\cdots + \beta nx$$

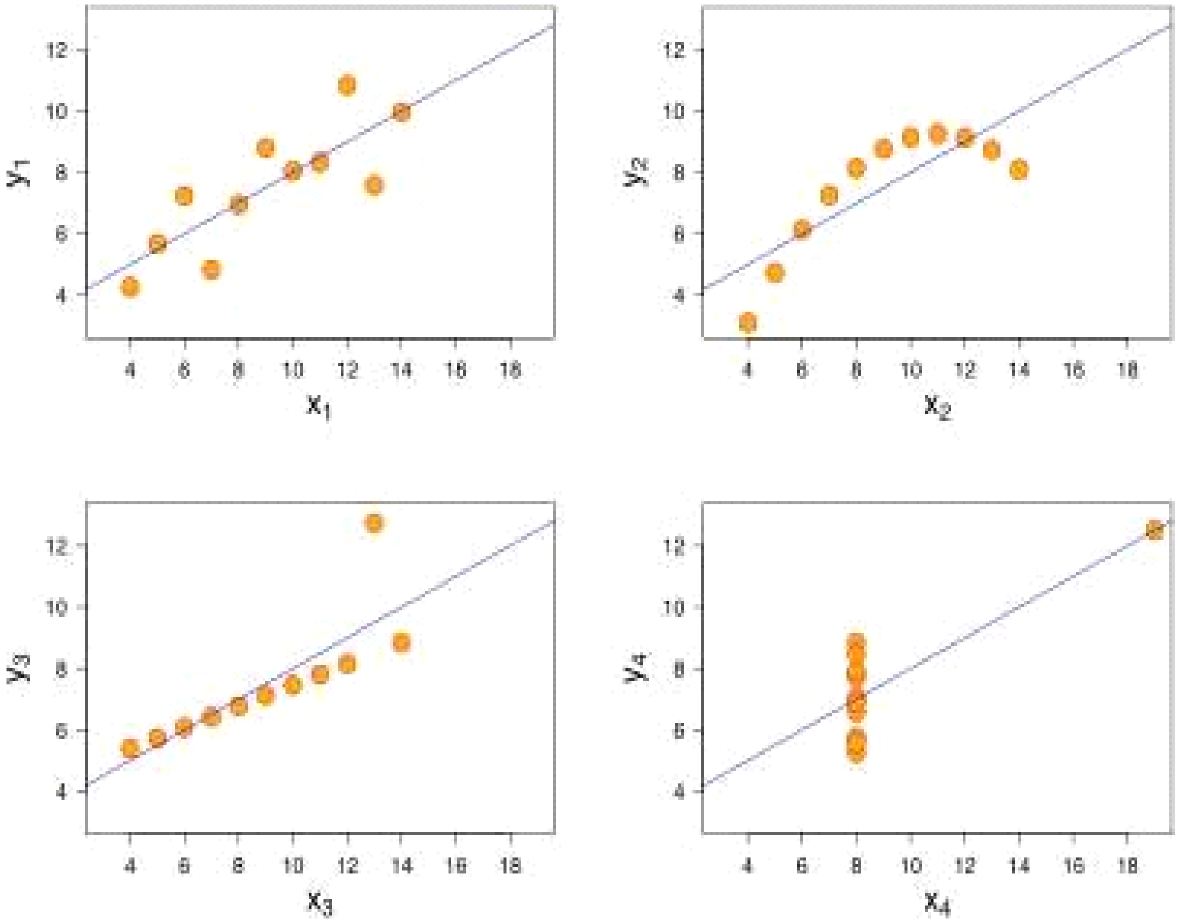
- The anticipated variable The predicted variable is than compared with testdata and assumptions are checked.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet consists of four data sets with virtually similarsimple descriptive statistics, but when represented graphically, the distributions are very different.

The mean, sample variation of x and y, correlation coefficient, linearregression 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 comparablestatistical features. The charts are displayed below:



- The first plot (top left) seems to represent a straightforward linearrelationship.

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

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

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

3. What is Pearson’s R?

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

- A value of 1 denotes a complete linear positive correlation. It implies that ifone 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 ifone variable rises, another will fall.

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

Answer: To keep a variable within a specified range, scaling is used. In a linearregression 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 inthe 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 ofVIF is infinite. Why does this happen?

Answer: The formula for VIF is

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

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 importanceof 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 distributionis its goal.
Data is being visually checked. If all of the data are from the same source, theplot will seem like a line. A Q-Q plot, short for a quantile-quantile plot, is a graphical method to compare the distribution of a sample of data to a theoretical distribution. The purpose of the Q-Q plot is to visually check if the distribution of the data matches the expected theoretical distribution, such as a normal distribution, and to identify any deviations from it.

In a Q-Q plot, the quantiles of the sample data are plotted against the corresponding quantiles of the theoretical distribution on a scatterplot. If the sample data follow the theoretical distribution, the points should fall along a straight line. However, if the sample data do not follow the theoretical distribution, the points will deviate from the straight line. In the context of linear regression, Q-Q plots are used to check the assumption of normality of the residuals, which are the differences between the observed values and the predicted values. The normality assumption is crucial for linear regression because it affects the validity of statistical inference, such as hypothesis testing and confidence interval estimation. To create a Q-Q plot for the residuals, the residuals are first standardized to have mean zero and standard deviation one, and then plotted against the corresponding quantiles of a normal distribution. If the residuals follow a normal distribution, the points should fall along a straight line. If the residuals deviate from the straight line, it indicates that the normality assumption may not be satisfied. In summary, Q-Q plots are important in linear regression because they provide a visual check for the normality assumption of the residuals, which is a crucial assumption for valid statistical inference.