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)**

When performing Exploratory Data analysis during the model build, following insights were obtained,

- The demand for bike rentals follows a distinct pattern, peaking at the beginning of the year and gradually declining towards the end.
- September stands out as the month with the highest demand for bike rentals among all other months.
- Clear weather conditions are preferred, indicating that people are more inclined to hire bikes when the weather is clear.
- During holidays, there is an observed trend of people preferring to stay indoors, resulting in reduced demand for bike rentals during these periods.
- A significant increase in bookings occurred in 2019 compared to 2018, suggesting a notable shift in demand between the two years.
- The fall season attracts a greater number of bookings compared to other seasons.
- Thursdays and Fridays exhibit slightly higher bike rental numbers compared to other days of the week.

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

Setting drop_first=True in the get_dummies method during dummy variable creation is crucial. When encoding categorical features, get_dummies would typically create n dummy variables for n categories.

However, these dummy variables are correlated, leading to multicollinearity issues, making it challenging for the model to discern the impact of each variable. By using drop_first=True, the first category is omitted during dummy variable creation, resulting in n-1 variables.

This helps mitigate multicollinearity and enhances the consistency and interpretability of machine learning models.

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

Upon examining the pair-plot among numerical variables, it becomes evident that the Temp variable exhibits the highest positive correlation with the target variable.

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

Multicollinearity Check:

- Utilized the Variance Inflation Factor (VIF) to assess multicollinearity in the data.

- Ensured that VIF factors were within the acceptable range, typically below 5, indicating minimal multicollinearity issues.

Homoscedasticity of Residuals:

- Conducted scatter plot analysis of residuals to verify homoscedasticity.
- Checked for consistency in error terms across variables, ensuring no visible patterns that might indicate heteroscedasticity.

Residual Errors:

- Examined the distribution plot of residual errors.
- Verified that the mean of residual values was centered around zero, ensuring a balanced distribution of errors.

Linearity:

- Ensured linearity between predictors and the target variable, a prerequisite for linear regression models.
- Utilized scatter plots to visually inspect and confirm the linear relationship between predictors and the target variable.

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

- Temp, yr, mnth_sep are the top 3 features which contributes to the bike demand of shared bikes.

General Subjective Questions

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

Linear Regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.

If a value of independent variables increase/decrease resulting in the increase/decrease of dependent or target variables forming a linear relationship. This is explained by the equation,

$$y = mx + c$$

y = dependent variable
x = independent variable/predictors
m = slope of the regression line
c = constant, Y-intercept

For more than one predictor, this can be represented as,

$$y=\beta 0+\beta 1x1+\beta 2x2+...+\beta nxn+\epsilon$$

y is the dependent variable
$\beta 0$ is the intercept

βi are the coefficients for each independent variable xi
ϵ represents the error term.

There are few assumptions on the data set for following a Linear Regression model,

- Linearity: Assumes a linear relationship between the dependent and independent variables.
- Independence: Assumes that the residuals (the differences between observed and predicted values) are independent.
- Homoscedasticity: Assumes constant variance of residuals across all levels of the independent variables.
- Normality: Assumes that the residuals follow a normal distribution.
- Multi collinearity – The independent variables are not highly correlated with each other.

Linear Regression is a fundamental and widely used algorithm due to its simplicity, interpretability, and effectiveness in capturing linear relationships between variables.

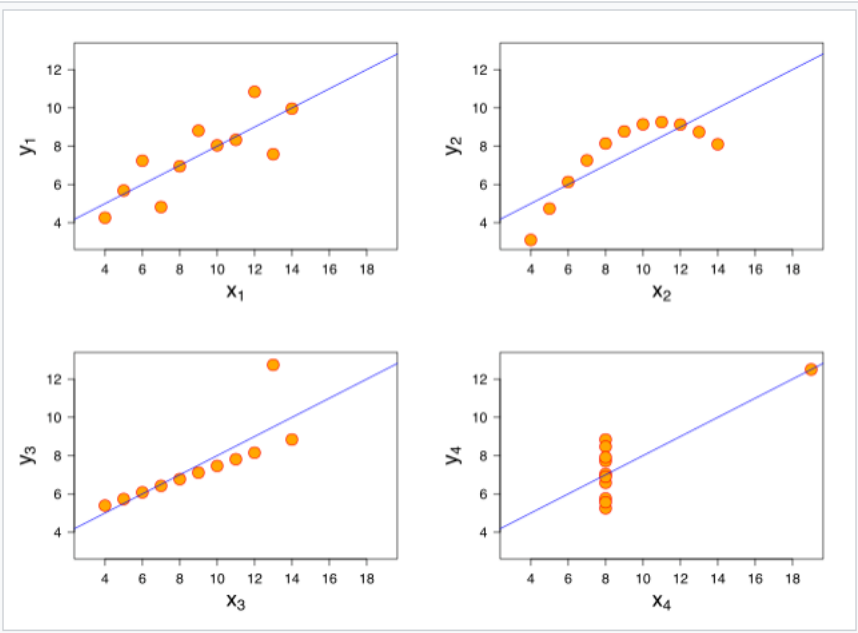
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet refers to a set of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear very different when graphed. The quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphical data exploration and the potential pitfalls of relying solely on summary statistics.

Each dataset in Anscombe's quartet consists of 11 data points, and they share the same mean, variance, correlation coefficient, and linear regression line. However, when graphed, these datasets exhibit distinct patterns that highlight the limitations of relying solely on summary statistics.

I		**II**		**III**		**IV**	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47

14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



All the dataset shares the same mean, SD, variance – 9, 11, 7.5 respectively.

Despite having similar summary statistics, these datasets have vastly different patterns when visualized. Anscombe's quartet serves as a cautionary example, demonstrating that relying solely on summary statistics may lead to an oversimplified understanding of the data and the potential to overlook important patterns or outliers.

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

Pearson's correlation coefficient, often denoted as *r* or Pearson's *r*, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

- *r*=1: Perfect positive linear correlation.
- *r*=−1: Perfect negative linear correlation.
- *r*=0: No linear correlation.

The formula for Pearson's correlation coefficient between two variables X and Y with n data points is:

.

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

Xi and *Yi* are individual data points
$\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively.

Pearson's correlation coefficient measures the degree to which the points in a scatter plot cluster around a straight line.

- If *r* is positive, it indicates a positive linear relationship (as one variable increases, the other tends to increase).
- If *r* is negative, it indicates a negative linear relationship (as one variable increases, the other tends to decrease).

It is important to note that Pearson's correlation coefficient assumes that the relationship between the variables is linear and that the data is approximately normally distributed.

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

Scaling is a preprocessing step in data analysis and machine learning that involves transforming the features of a dataset to a standard range. The goal is to bring all the features to a similar scale, preventing certain features from dominating or having an undue influence on the analysis due to their larger magnitude

Scaling is performed mainly due to below reasons,

1. Equal Weightage - Scaling ensures that all features contribute equally to the analysis.
2. Convergence in optimization - In optimization algorithms like gradient descent, having features on a similar scale helps the algorithm converge faster.
3. Distance based algorithm - For algorithms that rely on distances between data points (e.g., k-nearest neighbors), scaling is crucial to avoid biased results.
4. Regularization - Scaling is often necessary when using regularization techniques to avoid penalizing certain features disproportionately.

Difference between Normalized Scaling and Standardized Scaling:

1. **Normalized Scaling (Min-Max Scaling):**
    - **Formula:** normalized=X-Xmin/Xmax-Xmin
    - **Range:** Scales the features to a range between 0 and 1.
    - **Advantages:** Preserves the relative relationships between values.
    - **Concerns:** Sensitive to outliers, especially if the dataset has extreme values.
2. **Standardized Scaling (Z-score normalization):**
    - **Formula:** standardized=X-$\overline{X}$/SD(X)
    - **Range:** Scales the features to have a mean $\overline{X}$ of 0 and a standard deviation (σX) of 1.
    - **Advantages:** Less sensitive to outliers, works well with algorithms that assume normally distributed features.
    - **Concerns:** May not preserve the original distribution of the data.

In summary, both normalized and standardized scaling are methods to bring features to a similar scale. Normalized scaling is suitable when preserving the relative relationships between values is crucial, while standardized scaling is more robust to outliers and works well with algorithms that assume normally distributed features.

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

Infinite VIF values often occur when there is perfect multicollinearity, meaning one or more predictors in the model are perfectly correlated (i.e., one predictor is a perfect linear combination of others). This results in a situation where one predictor can be exactly predicted using the others.

To handle issues of multicollinearity, it's essential to identify highly correlated predictors and consider strategies such as dropping one of the correlated variables, combining them, or using regularization techniques in regression models

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution. It is commonly used to check whether the distribution of residuals (the differences between observed and predicted values) from a statistical model, such as linear regression, matches the expected normal distribution.

In a Q-Q plot, the quantiles of the observed data are plotted against the quantiles of the expected theoretical distribution (often the standard normal distribution). If the data follows the theoretical distribution, the points in the Q-Q plot should fall approximately along a straight line. Deviations from the line indicate departures from the assumed distribution.

Use and Importance in Linear Regression:

1. **Normality of Residuals:**

- A Q-Q plot helps assess whether the residuals from a linear regression model are approximately normally distributed. Normally distributed residuals are a key assumption for valid statistical inference, hypothesis testing, and confidence interval construction.

2. **Identification of Outliers:**
    - Outliers in the residuals can be detected by observing points that deviate significantly from the expected straight line in the Q-Q plot. Outliers can influence the validity and reliability of regression model results.
3. **Model Assumption Checking:**
    - Checking the normality of residuals is one of the diagnostic checks to verify the assumptions of a linear regression model. Departures from normality may suggest model misspecification or the presence of influential observations.

Interpretation:

- If the points in the Q-Q plot lie close to the expected line, it indicates that the residuals are approximately normally distributed.
- If there are systematic deviations, bends, or patterns in the plot, it suggests non-normality in the residuals.
- Outliers may be identified as points that deviate substantially from the straight line.