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:
I have done analysis on categorical columns using the bar plot.
Below are the few points we can infer from the visualization –

- Fall season appears to be the most popular for bookings.
- Across all seasons, there's a significant increase in bookings from 2018 to 2019.
- The months of May to October see the highest booking counts, with a trend of increasing bookings from the beginning of the year until mid-year, followed by a decline towards the year's end.
- Clear weather conditions are associated with higher booking numbers, which is to be expected.
- Thursdays, Fridays, Saturdays, and Sundays have higher booking numbers compared to the beginning of the week, indicating a preference for weekends.
- There's a relatively equal distribution of bookings between working days and non-working days.
- The year 2019 saw a notable increase in bookings compared to the previous year, reflecting positive growth in the business.

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

Answer:
Using `drop_first=True` in dummy variable creation helps prevent multicollinearity issues in regression analysis by excluding one level of the categorical variable as a reference category. This ensures that the dummy variables are linearly independent, improving the stability and reliability of the regression model. Without dropping one level, perfect collinearity can occur, leading to unstable estimates of regression coefficients and inflated standard errors.

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

Answer:
‘temp’ variable has the highest correlation 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)

I've validated the assumptions of the Linear Regression Model based on the following five criteria:

Normality of Residuals: Ensuring that the error terms follow a normal distribution.
Multicollinearity Check: Verifying that there's no significant multicollinearity among the predictor variables.
Homoscedasticity: Checking for absence of patterns in residual values, indicating consistent variance.

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

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes

- temp
- Winter
- September month
- Year

General Subjective Questions

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

Linear Regression is a machine learning algorithm which is based on supervised learning category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses Sum of Squared Residuals Method.
Linear regression is of the 2 types:
i. Simple Linear Regression: It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.
Formula for the Simple Linear Regression:

$Y=\beta 0+\beta 1X1 +\epsilon$

ii. Multiple Linear Regression: It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation

that can best determine the value of dependent variable Y for different values independent variables in X. It fits a ‘hyperplane’ instead of a straight line.
Formula for the Multiple Linear Regression:

$Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta pXp+\epsilon$

The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$ can be found by the following two methods:

· Differentiation

· Gradient descent

We can use statsmodels or SKLearn libraries in python for the linear regression.

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

Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.
The four datasets that make up Anscombe’s quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Purpose of Anscombe’s Quartet

Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

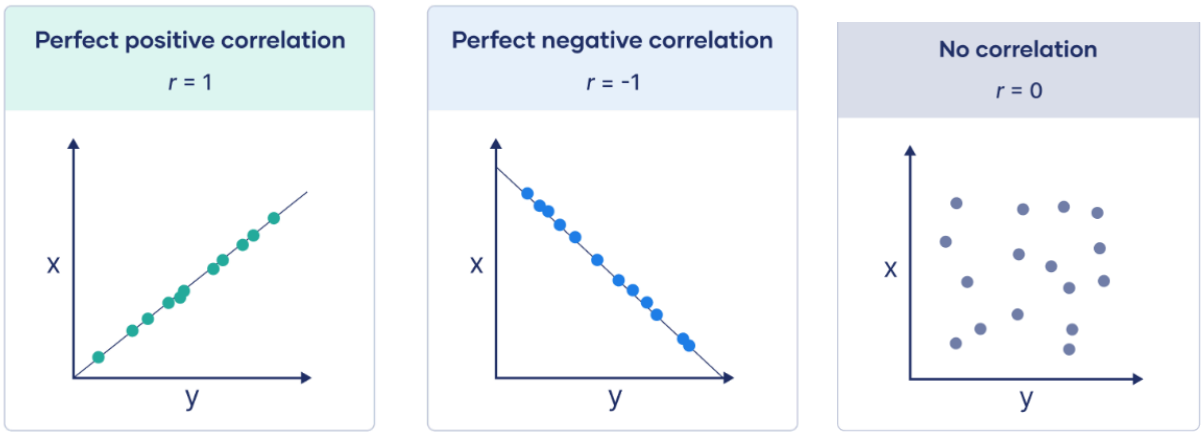
Anscombe’s Quartet Dataset

The four datasets of Anscombe’s quartet.

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

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

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
The Pearson correlation coefficient, r, can take a range of values from +1 to -1.
A value of 0 indicates that there is no association between the two variables.
A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.
A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.
This is shown in the diagram below:



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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

1.Normalized Scaling (Min-Max Scaling):

Normalized scaling transforms the values of features to a range between 0 and 1.
It preserves the relative distances between data points but does not handle outliers well.
The formula for normalized scaling is:

x-scaled=x-xmin/xmax-xmin

2.Standardized Scaling (Z-score Scaling):

Standardized scaling transforms the values of features to have a mean of 0 and a standard deviation of 1.
It is robust to outliers and maintains the shape of the distribution.
The formula for standardized scaling is:
x-scaled=x-mu/sigma

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

The occurrence of infinite values in the Variance Inflation Factor (VIF) is typically due to perfect multicollinearity among the predictor variables. Perfect multicollinearity happens when one or more of the independent variables in a regression model can be perfectly predicted by a linear combination of other variables.

When perfect multicollinearity exists, it means that one or more variables can be expressed as a perfect linear function of other variables, leading to a situation where the determinant of the correlation matrix of the independent variables becomes zero. This, in turn, results in infinite VIF values for those variables.

In practical terms, infinite VIF values indicate that the regression model cannot be estimated reliably because of the multicollinearity issue. To address this problem, one of the correlated variables should be removed from the model to improve its stability and interpretability.

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

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular probability distribution, such as the normal distribution.

In a Q-Q plot, the quantiles of the dataset are plotted against the quantiles of a theoretical distribution, typically the standard normal distribution. If the data points fall approximately along a straight line, it indicates that the dataset follows the assumed distribution. Deviations from the straight line suggest departures from the assumed distribution.

In linear regression, Q-Q plots are often used to evaluate the assumption of normality of residuals. The importance of Q-Q plots in linear regression lies in their ability to visually inspect whether the residuals (i.e., the differences between observed and predicted values) follow a normal distribution. If the residuals are normally distributed, they should approximately follow a straight line in the Q-Q plot. Departures from the straight line indicate potential violations of the normality assumption, which can affect the validity of the regression model's inference and predictions.

In summary, Q-Q plots are valuable tools in linear regression analysis for assessing the normality assumption of residuals. They provide a visual method to detect departures from normality, helping to identify potential issues in the regression model.