Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Based on my analysis of categorical variables on dependent variable using Bar chart, below inferences I have got:

1. Fall Season seems to have more bookings and bookings have been increased from 2018 to 2019 for each season.

2. Most of the Booking are in May, Jun, Jul till Oct. The trend of increasing the booking started in the initial time of the year and started decreasing at the end of the year. Bookings have been increased from 2018 to 2019 for all over the months.

3. Clear weather attracted more bookings, and it increased in 2019 compared to 2018.
4. It seems obvious that bookings are more on holidays compared to working days as people prefer to spend time with their families on these days.
5. Thursday, Friday and Saturday have more bookings compared to other week starting days.
6. Booking have been increased from year 2018 to 2019.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Drop_first = True is important because it helps reduce the extra columns generated during Dummy variable creation.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

"Temp" variable has the highest correlation.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Validated the assumptions of Linear Regression by below assumptions:

1. Normality of Error Terms – My error term is normally distributed
2. Multicollinearity Check
3. Linear relationship validation among variables
4. Independence of residuals – No Auto Correlation

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)
Top 3 features for explaining the demand are:
Temp
Season_Winter
yr

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a statistical method used to model the relationship between one dependent variable (target) and one or more independent variables (predictors). The goal is to find the best-fitting line (or hyperplane in higher dimensions) that predicts the target variable based on the predictors.

Linear Regression Assumptions:
Linearity: The relationship between predictors and the target variable is linear.
Independence of Errors: Residuals are independent and uncorrelated with each other.
Homoscedasticity: Residuals have constant variance across all levels of predictors.
Normality of Errors: Residuals are normally distributed.
No Multicollinearity: Predictors are not highly correlated with each other.

Linear Regression is of 2 types:
Simple Linear Regression - One independent variable and one dependent variable.
The equation for a simple linear regression (with one predictor) is:
$y=\beta0+\beta1x+\epsilon$
Where:
y: Target variable (dependent variable)
β0: Intercept (value of y when x=0)
β1: Slope coefficient (rate of change in y for a unit change in x)
x: Predictor variable (independent variable)
ϵ: Error term (accounts for variability not explained by the model)

Multiple Linear Regression – More than one independent variable and one dependent variable.
$y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\epsilon$

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

The Anscombe's Quartet, introduced by statistician Francis Anscombe in 1973, is a group of four datasets that have nearly identical statistical properties (mean, variance, correlation, and regression line) but appear very different when visualized. It illustrates the importance of visualizing data rather than relying solely on summary statistics.

Each dataset in the quartet has:

1. Same mean for both x and y.
2. Same variance for both x and y.
3. Same linear regression equation: y=3+0.5xy = 3 + 0.5xy=3+0.5x.
4. Same correlation coefficient: ~0.816.
5. Same coefficient of determination (R^2): ~0.67.

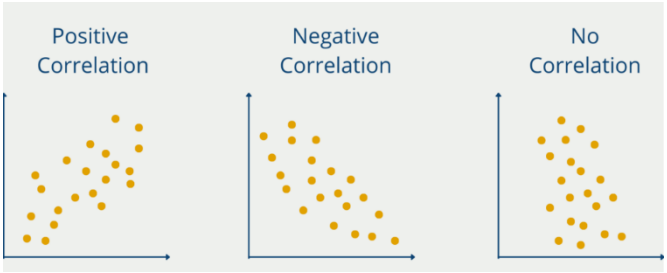
Below are the data characteristics that differentiate them:

1. Dataset 1:
    - A typical linear relationship between x and y.
    - Data points follow the regression line closely.
    - Perfectly suitable for linear regression.
2. Dataset 2:
    - A curvilinear relationship between x and y.
    - The regression line does not represent the data well.
    - Highlights the limitation of applying linear regression to nonlinear data.
3. Dataset 3:
    - A linear relationship with one extreme outlier.
    - The regression line is distorted due to the outlier.
    - Shows the impact of outliers on statistical summaries.
4. Dataset 4:
    - Data points are all vertically aligned except for one.
    - The regression line is heavily influenced by one single data point.
    - Demonstrates how a single point can dominate a statistical model.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's Correlation Coefficient (or simply Pearson's R) is a statistical measure that evaluates the strength and direction of a linear relationship between two continuous variables. It's one of the most widely used metrics to assess how changes in one variable are associated with changes in another.

Correlation measures the strength and direction of the relationship between two variables. A positive correlation means that as one variable increases, the other also increases (e.g., height and weight), with Pearson's r>0. A negative correlation indicates that as one variable increases, the other decreases (e.g., speed and travel time), with Pearson's r<0. When there is no correlation, the variables have no linear relationship, and Pearson's r=0



Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling is the process of transforming the values of features in a dataset so that they fall within a specific range or have similar magnitudes. It ensures that no single feature dominates the model due to its scale.

	Normalized Scaling	Standardized Scaling
Definition	Rescales data to a fixed range, usually [0, 1] or [-1, 1].	Centers data to have a mean of 0 and scales to unit variance (standard deviation = 1).
Output Range	Fixed, typically [0, 1] or [-1, 1].	No fixed range; depends on data distribution.
Sensitivity to Outliers	Highly sensitive to outliers as they can distort the min-max range.	Less sensitive to outliers as it uses mean and standard deviation.
When to Use	When data needs to be scaled to a specific range (e.g., pixel values or probabilities).	When data has a Gaussian distribution or standard normal distribution is required.
Example Algorithms	k-NN, Neural Networks.	Logistic Regression, SVM, PCA.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity among the predictor variables. This occurs when one or more predictor variables are exact linear combinations of others, making the denominator of the VIF formula zero:

Here, Ri2 represents the coefficient of determination for a regression where the i-th predictor is predicted using all other predictors. If Ri2=1, indicating perfect multicollinearity, the denominator becomes zero, causing the VIF to approach infinity. This typically signals a serious issue with the dataset's feature design, requiring corrective action such as removing or combining collinear variables.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, usually the normal distribution. The plot displays the quantiles of the data on the y-axis against the quantiles of the theoretical distribution on the x-axis. If the data follows the theoretical distribution, the points will roughly form a straight diagonal line.

Use in Linear Regression

In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot is used to visually assess this assumption by plotting the quantiles of the residuals against the quantiles of a standard normal distribution.

Importance of Q-Q Plot in Linear Regression

1. Checking Normality Assumption: It helps determine whether the residuals of the regression model are normally distributed, which is critical for valid hypothesis testing and confidence intervals.
2. Identifying Deviations: Deviations from the diagonal line in the Q-Q plot suggest that the residuals may have skewness, kurtosis, or heavy tails.
3. Model Diagnostics: It aids in identifying outliers or systematic errors in the model, indicating that the model might need improvements, such as transformations or different variable selections.