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)

My Analysis on categorical columns using the boxplot. Below are the few points inferred from the visualization –

Season Variable: Fall season has more bookings and then followed by summer.

Month Variable: Bookings are more in the month of may, june, july, aug, sep and oct and increasing trend.

Weathersit variable: Clear weather attracted more rental bookings.

Weekday variable: Thu, Fir, Sat and Sun have a greater number of bookings as compared to the start of the week.

Holiday variable: Average number of bookings are comparatively less during holidays

Working day variable: Booking seemed to be almost equal either on working day or non-working day.

Year Variable: 2019 attracted a greater number of booking from the previous year

---

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)

Using **drop_first=True** when creating dummy variables is important for below reasons:

1) Avoiding the Dummy Variable Trap: Including all dummy variables can lead to multicollinearity, where one variable can be perfectly predicted by others, affecting model stability and interpretability.
2) Reducing Dimensionality: Dropping one category reduces the number of features from n to n−1, simplifying the model and decreasing computational costs, especially in high-dimensional datasets
3) Maintaining Interpretability: Dropped category acts as a baseline for comparison, making it easier to interpret the effects of other categories on the dependent variable.

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 and atemp both numerical columns has highest correlation with cnt target variable. However, as both are redundant removing atemp

---

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)

1. Linearity: Scatter Plots of Independent Variables, help to check for linear relationships between

each independent variable and the dependent variable

2. Independence of Errors: Used the Durbin-Watson statistic.Durbin-Watson Statistic: 2.030008269439887 which is close to 2. Values close to 2 indicate no autocorrelation.

3. Homoscedasticity: By plotting residuals versus fitted values.
Spread of residuals is constant across all levels of predicted values.No visible pattern observed from above plot for residuals.

4. Normality of Errors: Checked by using a Q-Q plot, points are closely follow the diagonal line.

5. No Multicollinearity: Calculate Variance Inflation Factor (VIF), all the independent variables has less the 5 VIF, which means No Multicollinearity

---

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
Temperature (temp): Coefficient: 0.531651 P-Value: 9.83e-86
Year (yr):Coefficient: 0.229323 P-Value: 2.52e-107
Weather Situation (weathersit_Light_snowrain): Coefficient: -0.247816 P-Value: 9.07e-20
These features have both low p-values and substantial coefficients, indicating their strong influence on bike demand.

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)

<Your answer for Question 6 goes here>

Linear Regression details
Linear regression is a statistical technique used to model and analyze the relationships between variables. The goal is to find a linear equation that best predicts a dependent variable based on one or more independent variables.
Types of Linear Regression

1. **Simple Linear Regression (SLR):**
    - Involves one independent variable.
    - Models the relationship between a single feature and a target variable.
2. **Multiple Linear Regression (MLR):**
    - Involves two or more independent variables.
    - Models the relationship between multiple features and a target variable.

Simple Linear Regression (SLR)

The equation for SLR can be represented as:
$Y=\beta0+\beta1X+\epsilon$
Where:

- Y = dependent variable (target)
- X = independent variable (feature)
- $\beta0$ = intercept of the line
- $\beta1$ = slope of the line (change in Y for a one-unit change in X)
- ϵ = error term (the difference between the predicted and actual values)

Estimation of Coefficients:

- The coefficients $\beta0$ and $\beta1$ can be estimated using the Ordinary Least Squares (OLS) method:

Multiple Linear Regression (MLR)
The equation for MLR can be represented as:
$Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta nXn+\epsilon$
Where:

- Y = dependent variable (target)
- X1,X2,…,Xn = independent variables (features)
- $\beta0$ = intercept
- $\beta1,\beta2,\ldots,\beta n$ = coefficients for each independent variable
- ϵ = error term

Estimation of Coefficients:

- The coefficients can also be estimated using OLS, which minimizes the sum of squared residuals:

Assumptions of Linear Regression

1. **Linearity:**
    - The relationship between the independent and dependent variables is linear.
2. **Independence:**
    - Observations are independent of each other.
3. **Homoscedasticity:**
    - The residuals (errors) have constant variance at all levels of the independent variables.
4. **Normality:**
    - The residuals should be normally distributed, especially for hypothesis testing.
5. **No Multicollinearity (for MLR):**
    - Independent variables should not be highly correlated with each other. High correlation can distort the estimates of the coefficients.

Evaluation Metrics
To assess the performance of linear regression models, the following metrics are commonly used:

1. **R-squared (R2R^2R2):**
    - Measures the proportion of variance in the dependent variable that can be explained by the independent variables.
    - Ranges from 0 to 1, with higher values indicating a better fit.
2. **Adjusted R-squared:**
    - Adjusted for the number of predictors in the model. It is used to compare models with different numbers of independent variables.
3. **Mean Absolute Error (MAE):**
    - The average of absolute errors between predicted and actual values.
4. **Mean Squared Error (MSE):**
    - The average of squared errors.
5. **Root Mean Squared Error (RMSE):**

The square root of MSE, providing error in the same units as the dependent variable.

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)

<Your answer for Question 7 goes here>

Anscombe's Quartet:

Anscombe's Quartet is a set of four datasets that were constructed by the statistician Francis Anscombe in 1973. The datasets are notable for illustrating the importance of graphing data before analyzing it. Despite having nearly identical statistical properties, the datasets exhibit very different distributions and patterns when visualized. This highlights the potential pitfalls of relying solely on summary statistics.

The Datasets

Anscombe's Quartet consists of four datasets, each containing 11 pairs of (x,y) values. Here are the datasets:

Dataset	x values	y values
A	10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5	8.04, 6.58, 12.74, 9.96, 11.74, 12.74, 4.26, 3.19, 10.84, 5.68, 6.58
B	8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8	6.58, 5.76, 7.71, 7.14, 7.93, 6.58, 8.06, 6.80, 6.76, 7.82, 6.88
C	13, 8, 9, 11, 14, 8, 8, 8, 9, 14, 12	12.74, 9.96, 11.74, 12.74, 9.26, 12.74, 10.84, 11.14, 10.84, 5.68, 6.58
D	8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8	12.74, 4.26, 12.74, 9.96, 11.74, 12.74, 5.68, 4.26, 10.84, 5.68, 6.58

Statistical Properties

For all four datasets, the following statistical properties are nearly identical:

- **Mean of x**: Approximately 9
- **Mean of y**: Approximately 7.5
- **Correlation coefficient (r)**: Approximately 0.82
- **Regression line**: The regression line for all datasets has a slope of approximately 0.5 and an intercept of approximately 3.

Visualization

When plotted, each dataset reveals distinct characteristics:

1. **Dataset A**: A linear relationship with no outliers.
2. **Dataset B**: A vertical line of points (constant x), indicating that variations in y do not correspond to x.
3. **Dataset C**: A linear relationship with an outlier point that greatly affects the regression line.
4. **Dataset D**: Similar to Dataset B but with a point far away from the others, skewing the visual representation.

Importance of Anscombe's Quartet

Anscombe's Quartet serves several key purposes in statistical analysis:

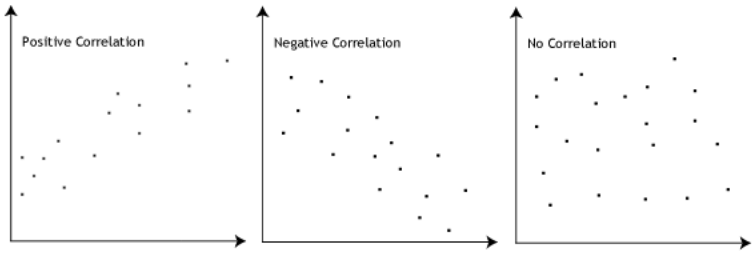
1. **Graphical Analysis**: It underscores the necessity of visualizing data before relying on summary statistics. Graphs can reveal trends, outliers, and patterns that numbers alone might obscure.
2. **Outlier Sensitivity**: It demonstrates how a single outlier can significantly influence statistical results, such as regression coefficients.
3. **Understanding Data Distribution**: The quartet illustrates that datasets can have identical statistical properties while exhibiting vastly different distributions and relationships.
4. **Teaching Tool**: Anscombe's Quartet is often used in statistics education to emphasize the importance of exploratory data analysis (EDA).

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)

<Your answer for Question 8 goes here>

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:



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)

<Your answer for Question 9 goes here>

Scaling adjusts the range of feature values in a dataset to ensure that all features contribute equally to model training, particularly in distance-based algorithms.

Why is Scaling Performed?

1. **Improve Model Performance:** Enhances convergence speed and accuracy.
2. **Handle Different Units:** Normalizes features measured in different scales.
3. **Avoid Bias:** Prevents larger features from dominating.
4. **Enhance Interpretability:** Makes coefficients easier to understand.

Types of Scaling

1. **Normalization (Min-Max Scaling):**
   - Rescales values to a range (e.g., 0 to 1).
   - Formula:
   - Xnormalized= (X−Xmin) **/** (Xmax−Xmin)
   - Best for distance-based algorithms.
2. **Standardization (Z-Score Scaling):**
   - Centers values to have a mean of 0 and a standard deviation of 1.
   - Formula:
   - Xstandardized =(X−μ)**/**σ
   - Best when data is normally distributed.

Key Differences

Feature	Normalization	Standardization
Range	Fixed range (0 to 1)	Mean 0, Std Dev 1
Sensitivity	Sensitive to outliers	Less sensitive
Use Cases	Distance metrics	Normality assumptions

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)

<Your answer for Question 10 goes here>

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity among independent variables. A VIF value indicates how much a variable is correlated with other variables in the model.

Reasons for Infinite VIF

1. **Perfect Multicollinearity:**
   - This occurs when one independent variable is an exact linear combination of other independent variables. For example, if you have two features, X1X_1X1 and X2X_2X2, where X2=2×X1X_2 = 2 \times X_1X2=2×X1, X1X_1X1 and X2X_2X2 are perfectly correlated, leading to an infinite VIF.
2. **Redundant Features:**
   - If there are multiple features that essentially convey the same information, this redundancy can cause multicollinearity, resulting in an infinite or extremely high VIF for those features.
3. **Singular Matrix:**
   - In matrix algebra, if the design matrix (which includes the independent variables) is singular (non-invertible), this can lead to an infinite VIF. This situation arises when the determinant of the correlation matrix is zero.

Implications

When VIF values are infinite, it indicates severe multicollinearity, making it impossible to determine the individual effect of correlated variables on the dependent variable. This can lead to unreliable coefficient estimates and decreased model interpretability.

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)

<Your answer for Question 11 goes here>

Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, commonly the normal distribution. It compares the quantiles of the sample data against the quantiles of a specified distribution.

How a Q-Q Plot Works

1. **Quantile Calculation:** It calculates the quantiles of the sample data.
2. **Theoretical Quantiles:** It calculates the quantiles from the theoretical distribution (e.g., normal distribution).
3. **Plotting:** The quantiles of the sample data are plotted on the y-axis against the theoretical quantiles on the x-axis.

Interpretation

- **Straight Line:** If the points in the Q-Q plot fall approximately along a straight diagonal line, it suggests that the sample data follows the specified distribution (e.g., normality).
- **Curvature:** Deviations from the straight line indicate departures from the assumed distribution. For instance:
    - Points bending upward suggest a distribution with heavier tails than the normal distribution (e.g., more outliers).
    - Points bending downward suggest a distribution with lighter tails.

Use and Importance in Linear Regression

1. **Normality of Residuals:**
    - In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps assess this assumption.
2. **Model Validation:**
    - By checking the normality of residuals, analysts can validate the appropriateness of the linear regression model. Non-normally distributed residuals can indicate issues such as:
        - Model misspecification
        - Presence of outliers
        - Non-linearity in the relationship
3. **Guiding Transformations:**
    - If the Q-Q plot indicates that residuals are not normally distributed, transformations (e.g., log, square root) may be necessary to meet the assumption.
4. **Robustness of Results:**
    - Ensuring normality of residuals contributes to the robustness of hypothesis tests (like t-tests for coefficients), which assume that residuals are normally distributed.