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)

Through an analysis of categorical features using boxplots and bar charts, I have observed several key trends:

- **Seasonal Influence**: The demand for bikes peaked during the fall season, with a noticeable rise in bookings from 2018 to 2019.
- **Monthly Trends**: The highest bookings occurred between May and October, showing an increasing trend in the first half of the year, followed by a decline towards the year's end.
- **Weather Conditions**: Clear weather conditions led to a higher number of rentals, which is expected as unfavorable weather can deter outdoor activities.
- **Day of the Week Impact**: Bookings were higher on Thursdays, Fridays, Saturdays, and Sundays, indicating increased demand toward the weekend.
- **Holiday Effect**: Holidays showed a decrease in bike rentals, likely because people prefer to stay home or engage in family activities rather than commute.
- **Working vs. Non-Working Days**: There was no significant difference in demand between working and non-working days, suggesting consistent rental behavior across weekdays and weekends.
- **Yearly Growth**: The year 2019 recorded a higher number of bookings compared to the previous year, reflecting a positive trend in business growth.

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)

Setting **drop_first=True** while creating dummy variables is crucial as it helps in eliminating redundancy and reducing multicollinearity among categorical variables.

By default, when dummy variables are created, they generate ***k*** new columns for a categorical variable with ***k*** unique values. However, one of these columns can be inferred from the others. By setting ***drop_first=True***, we remove the first category, ensuring that only ***k-1*** columns are retained, which helps in avoiding the dummy variable trap (a scenario where one column is a linear combination of others, leading to multicollinearity).

Example:
If a categorical variable has three categories: **A, B, and C**, the dummy encoding creates three new columns. However, if we know that a row does not belong to A or B, it must belong to C. Thus, we only need two columns to represent all three categories, making the third unnecessary.
By dropping the first category, we maintain the same information while improving the model's efficiency.

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)

Based on the pair-plot analysis among numerical variables, the '**temp**' variable exhibits the highest correlation with the target variable '**cnt**'. This indicates that as temperature increases, the demand for bikes tends to rise, suggesting a strong positive relationship between these two variables.

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)

After building the model on the training set, the following assumptions of Linear Regression were validated:

1. **Normality of Error Terms**
   The residuals were analyzed to ensure they follow a normal distribution.

2. **Multicollinearity Check**
   Variance Inflation Factor (VIF) was calculated to confirm that there is no significant multicollinearity among independent variables.

3. **Linear Relationship Validation**
   Component and Component Plus Residual (CCPR) plots were used to verify linearity between predictors and the target variable.

4. **Homoscedasticity (Constant Variance of Residuals)**
   A residual vs. fitted values plot was examined to check for any visible patterns, ensuring homoscedasticity.

5. **Independence of Residuals**
   The Durbin-Watson test was considered to detect autocorrelation in residuals, ensuring independence.

By validating these assumptions, we confirmed that the model is reliable and meets the requirements of Linear Regression.

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)

Based on the final model, the three most significant features influencing the demand for shared bikes are:

1. **Temperature (temp)** – A higher temperature is strongly associated with increased bike demand.
2. **Year(year)** – The growing trend in bike usage over time indicates an increasing preference for shared bikes.
3. **Winter (winter)** – The winter season positively impacts bike usage, likely due to comfortable weather conditions.

And,

4. **September (sep)** – The demand is notably higher in September, indicating seasonal trends in bike rentals.

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 Algorithm:
Linear regression is a statistical modeling technique used to establish a linear relationship between a dependent variable and one or more independent variables. It predicts how changes in independent variables influence the dependent variable.

Mathematical Representation:
The linear regression equation is expressed as:

Y = mX + c

Where:
Y is the dependent variable (the target we aim to predict).
X is the independent variable(s) (the predictors used for prediction).
m represents the slope of the regression line, indicating the impact of X on Y.
c is the intercept, which is the predicted Y value when X = 0.

Types of Linear Regression:

1. Simple Linear Regression – Involves a single independent variable.
2. Multiple Linear Regression – Involves multiple independent variables affecting the dependent variable.

Nature of Linear Relationships:

1. **Positive Linear Relationship** – As the independent variable increases, the dependent variable also increases.
2. **Negative Linear Relationship** – As the independent variable increases, the dependent variable decreases.

Key Assumptions of Linear Regression:

1. **Multicollinearity:** The model assumes minimal or no correlation between independent variables. If multicollinearity is present, it can distort the relationship among variables.
2. **Autocorrelation:** Residual errors should not be correlated with each other; otherwise, predictions may become unreliable.
3. **Linearity:** The relationship between independent and dependent variables should be linear.
4. **Normality of Error Terms:** The error terms should follow a normal distribution.
5. **Homoscedasticity:** The variance of residuals should be constant across all levels of independent variables, meaning no visible pattern should exist in residual values.

Linear regression is widely used for trend analysis, forecasting, and predictive modeling in various domains, including finance, healthcare, and business analytics.

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)

Anscombe's Quartet:
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. However, when visualized through graphs, they tell entirely different stories. Each dataset exhibits unique patterns despite having identical summary statistics.

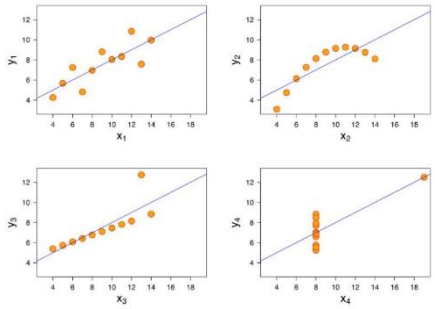
Key Statistical Properties of Anscombe's Quartet

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics of these four datasets are identical:

- Mean of x: 9
- Mean of y: 7.50
- Variance of x: 11
- Variance of y: 4.13
- Correlation coefficient (between x and y): 0.816
- The regression line equation remains the same across all datasets.

Interpretation of the Four Datasets



Dataset I: Appears to have a clean and well-fitting linear model.
Dataset II: Displays a non-normal distribution.
Dataset III: Although linearly distributed, the regression is thrown off due to an outlier.
Dataset IV: A single outlier significantly influences the correlation coefficient and regression line.

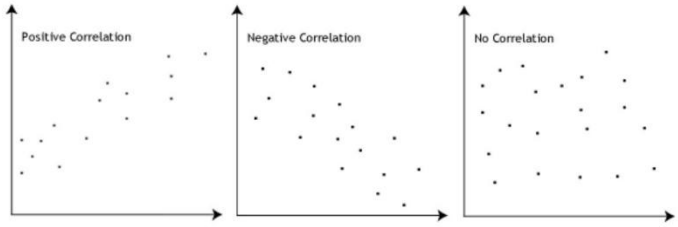
Significance of Anscombe's Quartet

Anscombe's Quartet emphasizes the importance of data visualization in analysis. While summary statistics may appear identical, visualizing the data through scatter plots reveals key differences in structure and trends. This underscores the need to go beyond numerical metrics and incorporate graphical exploration to gain deeper insights into a dataset.

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 R
Pearson's r is a numerical measure that summarizes the strength and direction of a linear relationship between two variables. If two variables increase or decrease together, the correlation coefficient is positive. If one variable increase while the other decreases, the correlation coefficient is negative.
The Pearson correlation coefficient, r, can take values between -1 and +1:

- **r = 0** indicates no correlation between the variables.
- **r > 0** signifies a positive correlation, meaning as one variable increases, the other tends to increase.
- **r < 0** signifies a negative correlation, meaning as one variable increases, the other tends to decrease.



This metric is useful in identifying relationships between variables, but it does not imply causation. Graphical representation is often used alongside Pearson's r to ensure accurate interpretation of data relationships.

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)

Feature Scaling is a preprocessing technique used to adjust the range of independent variables in a dataset. It ensures that all features have comparable scales, preventing algorithms from assigning undue importance to larger values simply due to their magnitude.

Why is Scaling Necessary?

Without scaling, machine learning models may misinterpret numerical values. For instance, an algorithm could consider 3000 meters to be significantly larger than 5 kilometers, even though they represent the same distance. By applying scaling, we bring all features to a uniform scale, improving model accuracy and stability.

Difference Between Normalized Scaling and Standardized Scaling:

Aspect	**Normalized Scaling**	**Standardized Scaling**
Method	Uses minimum and maximum values for scaling	Uses mean and standard deviation for scaling
When to Use?	When features are on different scales	When we need zero mean and unit variance
Range of Values	Scales values between [0,1] or [-1,1]	Not bounded to a fixed range
Effect of Outliers	Highly affected by outliers	Less affected by outliers
Implementation	MinMaxScaler in Scikit-Learn	StandardScaler in Scikit-Learn

---

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)

Variance Inflation Factor (VIF) measures multicollinearity among independent variables in a

dataset. When there is **perfect correlation** between two or more variables, the VIF value becomes **infinite**. This occurs because the **R-squared (R^2) value reaches 1**, making the denominator of the VIF formula **1 / (1 - R^2) approach infinity**.

A high VIF indicates that a variable is highly correlated with others, leading to instability in regression models. If VIF is infinite, it means that two independent variables provide redundant information. To resolve this issue, one of the correlated variables should be removed to eliminate perfect multicollinearity.

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 **Quantile-Quantile (Q-Q) plot** is a graphical tool used to assess whether a dataset follows a specific theoretical distribution by comparing its quantiles against the quantiles of a reference distribution. It is commonly used in statistics to check the normality of data, which is a key assumption in many statistical models, including **linear regression**.

Use of Q-Q Plot:

A Q-Q plot plots the quantiles of the observed dataset on the **y-axis** against the quantiles of a theoretical distribution (such as a normal distribution) on the **x-axis**. A **quantile** represents the proportion of data points that fall below a given value. For instance, the 30% quantile is the value below which 30% of the data points lie.

- A **straight 45-degree line** is drawn as a reference.
- If the data follows the expected distribution, the points will align closely with this reference line.
- **Deviations from this line indicate deviations from the assumed distribution**, suggesting skewness, heavy tails, or other non-normal characteristics.

Importance of Q-Q Plot in Linear Regression:

In **linear regression**, one key assumption is that the residuals (errors) should follow a **normal distribution**. A Q-Q plot is useful for validating this assumption:

1. **Detecting Normality** – If the residuals follow a normal distribution, the Q-Q plot will show points aligning along the reference line.
2. **Identifying Skewness** – A curve in the Q-Q plot suggests the data is skewed (left or right).
3. **Recognizing Heavy Tails or Outliers** – If points deviate at the ends, it indicates extreme values or heavy-tailed distributions.

By analyzing a Q-Q plot, we can determine whether **normality transformations** or adjustments are needed, ensuring that the regression model produces reliable and unbiased predictions.
