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:

Analysis of categorical variables through boxplots and bar plots revealed the following insights:

1. **Seasonal Impact:**
    - The fall season observed the highest booking rates.
    - A significant increase in bookings was noted across all seasons from 2018 to 2019.
2. **Monthly Trends:**
    - Bookings peaked during May to October.
    - There was a steady rise in bookings at the start of the year, reaching a peak mid-year, followed by a decline towards year-end.
3. **Weather Influence:**
    - Clear weather conditions correlated with higher bookings, as expected.
4. **Weekly Patterns:**
    - Thursdays, Fridays, Saturdays, and Sundays recorded higher bookings compared to other weekdays.
5. **Holiday vs. Non-Holiday:**
    - Bookings were lower on non-holidays, as people may prefer spending holidays at home.
6. **Working Days:**
    - Bookings were nearly equal for working and non-working days.
7. **Yearly Comparison:**

The year 2019 witnessed more bookings than 2018, indicating growth in business

---

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:

drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level. Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C

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:
Temp variable has the highest correlation with the target variable.

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:

1. **Normality of Error Terms:**
    - Verified through a Q-Q plot and histogram of residuals.
    - Residuals aligned closely with the diagonal line in the Q-Q plot, confirming normality.
2. **Multicollinearity:**

    Checked using Variance Inflation Factor (VIF)
3. **Linearity:**
    - Scatter plots of residuals against predictors confirmed a linear relationship.
4. **Homoscedasticity:**
    - Residual plots showed no discernible patterns, indicating constant variance.

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: The top 3 features significantly explaining shared bike demand are:

Rank	Feature	Contribution
1	Year	High
2	Temp	Significant
3	Summer	Substantial

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer:
Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear correlation, meaning any change (increase or decrease) in the independent variable(s) directly influences the dependent

variable in a proportional manner. This relationship is expressed mathematically as:

Y = mX+c

Where:

- Y: The dependent variable we aim to predict.
- X: The independent variable(s) used for prediction.
- m: The slope of the regression line, indicating the impact of X on Y.
- c: The y-intercept, representing the value of Y when X=0.

Assumptions of Linear Regression:

For the model to be valid, it relies on the following assumptions:

1. **Multi-collinearity:**
    - The model assumes minimal or no dependency among the independent variables. High multi-collinearity (when predictors are highly correlated) can distort the regression results.
2. **Auto-correlation:**
    - Residuals (errors) should be independent. Auto-correlation, where residuals are dependent on each other, violates this assumption.
3. **Linearity:**
    - The relationship between the predictor variables and the response variable must be linear.
4. **Normality of Error Terms:**
    - Residuals should follow a normal distribution.
5. **Homoscedasticity:**
    - The variance of residuals should remain consistent across all levels of the independent variable(s). Patterns in residuals indicate heteroscedasticity, which violates this assumption.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer:

Anscombe's Quartet, introduced by statistician Francis Anscombe, consists of four datasets, each containing 11 pairs of x and y values. While these datasets share identical summary statistics, their visual representations reveal vastly different patterns. This underscores the critical importance of data visualization in analysis.

Key Summary Statistics (Identical Across All Four Datasets):

- **Mean of x:** 9.0
- **Mean of y:** 7.50
- **Variance of x:** 11.0
- **Variance of y:** 4.13
- **Correlation Coefficient (r) Between x and y:** 0.816

When plotted on an x-y coordinate plane, each dataset presents a unique story:

1. **Dataset I:** Displays a clean and well-fitting linear relationship.
2. **Dataset II:** Shows a non-linear pattern.
3. **Dataset III:** Features a linear distribution disrupted by a single outlier.
4. **Dataset IV:** Highlights how one significant outlier can produce a misleadingly high correlation coefficient.

Importance:

Anscombe's Quartet demonstrates that identical statistical summaries can mask significant differences in data behavior. Visualization reveals these differences, providing a clearer understanding of the data's underlying structure. It serves as a reminder to complement numerical analysis with graphical representations.

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 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)

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. S.NO. Normalized scaling Standardized scaling 1. Minimum and maximum value of features are used for scaling Mean and standard deviation is used for scaling. 2. It is used when features are of different scales. It is used when we want to ensure zero mean and unit standard deviation. 3. Scales values between [0, 1] or [-1, 1]. It is not bounded to a certain range. 4. It is really affected by outliers. It is much less affected by outliers. 5. Scikit-Learn provides a transformer called MinMaxScaler for Normalization. Scikit-Learn provides a transformer called StandardScaler for standardization.

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)

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
When the value of VIF is infinite it shows a perfect correlation between two independent

variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this 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)

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests

---