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)

I conducted an analysis of the categorical variables using both boxplots and bar plots. The following key insights were derived from the visualizations:

- **Seasonal Trends**: The fall season appears to have attracted the highest number of bookings. Additionally, the number of bookings increased significantly across all seasons from 2018 to 2019.
- **Monthly Trends**: The majority of bookings occurred during June, July, August, September, and October. The trend showed an upward trajectory from the beginning of the year, peaking mid-year, and declining towards the year-end.
- **Day of the Week**: Average bookings were spread equally on all days with slight increase on Wednesday and Thursdays compared to the earlier days of the week
- **Working Days**: The average number of bookings was almost equal on working and non-working days with higher variance on non working day
- **Weather Situation**: Clear weather conditions has attracted a higher number of bookings followed by mist.
- **Holiday Impact**: Holiday periods saw slightly fewer bookings, likely because people prefer to spend time at home or with family.
- **Yearly Comparison**: The year 2019 recorded a higher number of bookings compared to 2018, indicating significant growth in business performance.

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** is important as it eliminates the additional column generated during dummy variable creation. This in turn helps reduce the multicollinearity among the 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.

Example: in the above assignment, weather is a categorical variable with four level/values, namely spring, summer, fall and winter. If weather is not winter, summer or fall, then it's obvious that the weather is spring, in this way we don't have to create dummy variable for spring, therefore we use **drop_first=True,** to avoid creation of spring column

---

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’ & ‘atemp’ has the highest correlation with the target variable ‘cnt’

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)
I validated the model that was built against following main assumptions :

- N**ormally distributed of Error terms**: plotted a graph for y_train_actual vs y_train_predicted, and examined to be distributed normally
- **Linear relationship validation**: predictor variables were showing linearly relationship with target variable
- **Multicollinearity validation :** made sure that predictor variables are not having highly co-related , Ex VFI for all predictor variables < 5
- **Homoscedasticity check :** error terms had constant variance
- **Independency of Error terms :** plotted a graph for y_train_predicted - y_train_actual against y_train_actual and confirmed there is no visible pattern.

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)
The top 3 features contributing significantly to the demand of shared bikes are :

- Temp (Temperature)
- year (yr)
- Weathers specifically Light-snow or Light rain

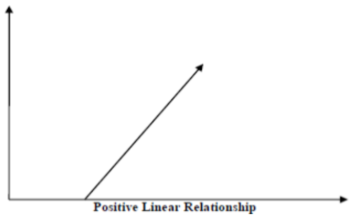
General Subjective Questions

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>

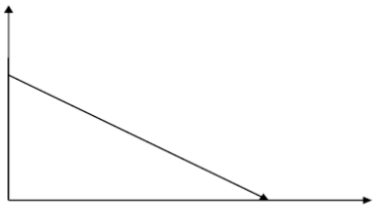
- Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables.
- Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
- Mathematically the relationship can be represented with the help of following equation −
    - $Y = mX + c$
    - Here, Y is the dependent variable we are trying to predict.
    - X is the independent variable we are using to make predictions.
    - m is the slope of the regression line which represents the

      effect X has on Y c is a constant, known as the Y-intercept.
- Furthermore, the linear relationship can be positive or negative in nature as explained below−
    - <u>Positive Linear Relationship:</u>
        - A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



Positive Linear Relationship

    - <u>Negative Linear relationship:</u>
        - A linear relationship will be called negative if independent increases and dependent variable decreases. It can be understood with the help of following graph −



Negative Linear Relationship

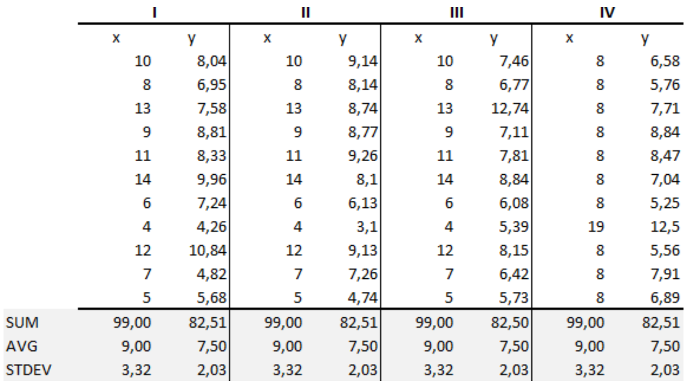
- Linear regression is of the following two types −
    - Simple Linear Regression
    - Multiple Linear Regression

- Assumptions -
    - The following are some assumptions about dataset that is made by Linear Regression model
    - Multi-collinearity
        - Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
    - Auto-correlation
        - Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
    - Relationship between variables
        - Linear regression model assumes that the relationship between dependent and independent variables must be linear.
    - Normality of error terms
        - Error terms should be normally distributed
    - Homoscedasticity
        - Error terms have constant variance.

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 is a set of four datasets created by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and the limitations of relying solely on summary statistics.
- 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. But things change completely, when they are graphed.
- Each graph tells a different story irrespective of their similar summary statistics.

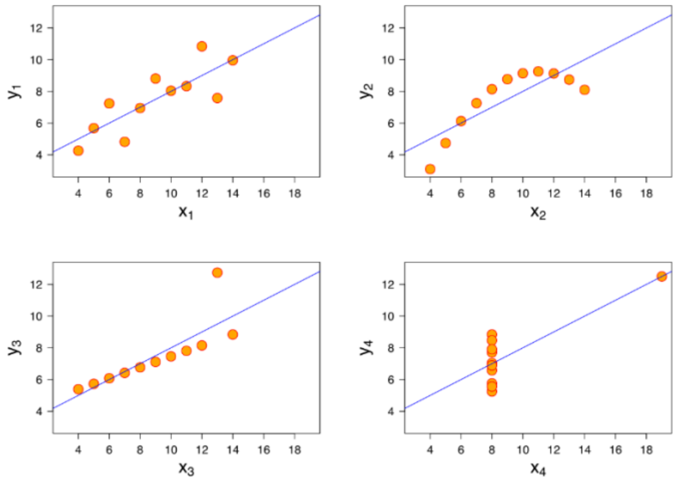


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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

- Remarkably, all four datasets share nearly identical statistical properties:
    - Same mean for x (9.0)
    - Same mean for y (7.5)

    - Same variance for x (11.0)
    - Same variance for y (4.1)
    - Same correlation coefficient (0.816)
    - Same linear regression line (y = 3 + 0.5x)

- However, when plotted, each dataset reveals dramatically different patterns:



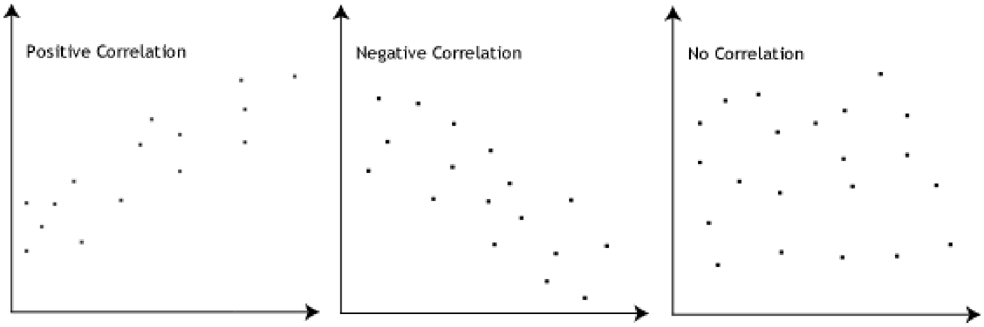
    - Dataset I: Shows a typical linear relationship
    - Dataset II: Shows a clear curvilinear relationship
    - Dataset III: Shows a linear relationship with one outlier
    - Dataset IV: Shows a case where a single outlier dramatically influences the regression line
- The main lessons from Anscombe's quartet are:
    - The importance of visualizing data before drawing conclusions
    - The danger of relying solely on summary statistics

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.



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 is a preprocessing technique in data analysis and machine learning where features are transformed to a similar scale or range.
- It involves adjusting the values of numeric variables to ensure they are comparable and prevent certain features from dominating the analysis due to their larger magnitudes.
- Scaling is performed for several imporatant reasons:
    - To ensure all features contribute equally to the model
    - To improve the convergence of machine learning algorithms
    - To prevent features with larger values from dominating the objective function
    - To enhance the numerical stability of many machine learning algorithms
- 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.
- Here's the difference between normalized scaling and standardized scaling :

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.

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) becomes infinite when there is perfect multicollinearity among predictor variables in a regression model.
- This occurs when one independent variable can be expressed as an exact linear combination of other independent variables, creating perfect linear dependencies.
- In the case of perfect correlation, we get R-squared (R2) =1, which lead to VIF= 1/ (1-R2) infinity. Since denominator becomes 0 .
- This happens due to several reasons:
    - Perfect correlation between two or more independent variables
    - Redundant variables in the dataset (one variable is a multiple of another)
    - Dummy variable trap, where one categorical variable's dummy is perfectly predictable from others
- To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- For example: Consider three variables: X1, X2, and X3 , If X3 = 2X1 + 3X2, then X3 is perfectly predictable from X1 and X2, leading to infinite VIF.

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>

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution
- It is a graphical diagnostic tool that compares two probability distributions by plotting their quantiles against each other.
- In linear regression, it specifically compares the distribution of residuals against a theoretical normal distribution to assess the normality assumption.
- The use and importance of Q-Q plots in linear regression include:
    - Assessing Normality Assumption:
        - Helps verify if residuals follow a normal distribution
        - Points falling along the 45-degree reference line indicate normality
        - Deviations from the line suggest non-normality
    - Detecting Specific Distribution Issues:
        - Heavy tails: Points curve away from the line at the ends
        - Skewness: Points show systematic deviation above or below the line
        - Outliers: Points at the extremes deviate significantly from the line
    - Practical Applications:
        - Validates the assumptions required for reliable regression analysis
        - Guides decisions about data transformations if needed
        - Assists in model diagnostics and improvement