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)

The analysis done on categorical data using bar plots for various independent variables has provided the following insights

- There is a variation in demand based on the season. Spring has the lowest demand whereas fall has the highest demand.
- There is a slight variation in demand based on whether a day being a working day or not. Working days are having more demand.
- The demand seems constant throughout the week, being slightly lower on Sunday and Monday.
- Weather is a major factor affecting the demand. On rainy days there are no customers. Clear weather attracts maximum demand.
- A holiday leads to a dip in demand.
- There is a year-on-year increase in demand for the Bike rental across all the variables impacting the demand.

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: Dummy variables are created for categorical data to convert nonbinary data into binary data. For a set of categorical variables, the 'n' values can be represented by 'n-1' dummy variables. The 'n-1' variables can predict the value for nth variable. If all the states of categorical variable are retained as dummy variables, one of the states will be deterministic by combination of all the remaining states. This will pose a problem of multicollinearity as one of the states will have high correlation with all other independent states.

To overcome this problem, drop_first=True is required to be used while creating dummy variables. Drop_first=True will drop the first column after creating the dummy variable. Thus, it eliminates one of the dummy variables ensuring the total number of dummy variables are 'n-1' when there are 'n' different states for the categorical 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)

The variable 'temp' 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: <Your answer for Question 4 goes below this line> (Do not edit)

The assumptions of the Linear Regressions are follows

1. Linearity: The assumption is that there exists a linear relation between dependent and independent variables.

   The assumption can be verified by

a. Scatter Plots: Plotting a scatter plot between independent variables and dependent variables. Data points on a straight line validates a linear relationship
b. Partial Residual Plots: These plots show the relationship between each predictor and the residuals after considering the effect of other variables.

2. Independence: There is no relation between residuals of the consecutive terms

   The assumption can be verified by

a. Durbin-Watson (DW) statistic test: The value between shall be between 0-4. A value midway i.e. around 2 indicates there is no auto correlation. A value close to 0 show a negative correlation and a value close to 4 shows a positive correlation.

3. Homoscedasticity: There is a constant variance across the entire range of independent variable

   The assumption can be verified by

a. Scatter Plots: Plotting a scatter plot between independent variables and variances. If the variances are evenly spread indicates that the residuals have the constant variance

4. Normality: The residuals shall have normal distribution

   The assumption can be verified by

a. Q-Q (Quantile-Quantile) plot: if the data points on the graph forms a straight diagonal line than the residuals have normal distribution
b. Histogram of Residuals: The histogram of residuals shall have a bell shape curve
c. Shapiro-Wilk test or Kolmogorov-Smironov or Anderson-Darling: These statistical method helps in determining that the distribution is normal

5. No Multicollinearity: The independent variables shall not be corelated

   The assumption can be verified by

a. VIF(Variance Influence Factor): A value below 4 indicates 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)

The top three factors affecting the demand are as follows

- Temp
- yr
- `weathersit_Snow`

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 is a supervised machine learning algorithms which computes a linear relationship between a dependent variable and a set of independent variables by establishing a linear equation between a dependent variable and a set of independent variables.

Simple Linear Regression: In the simplest form the univariant linear regression establishes a relationship between a dependent and one independent variable. The univariant linear regression equation is

Y = mX+c

Y = dependent variable

X = Independent variable

m = slope of the regression line which establishes the relationship between the dependent variable and the independent variable

c = constant which defines the value of Y when X =0

Multiple Linear Regression: The dependent variable is related to more than one independent variable. The equation for Multiple Linear Regression is

Y = m1X1+m2X2+……mnXn+c

Y = dependent variable

X1, X2,…..Xn = Independent variable

m1,m2…..mn= slope of the regression line which establishes the relationship between the dependent variable and the respective independent variable

c = constant which defines the value of Y when X =0

The Linear relationship can be

a. Positive Linear relationship
b. Negative linear relationship

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)

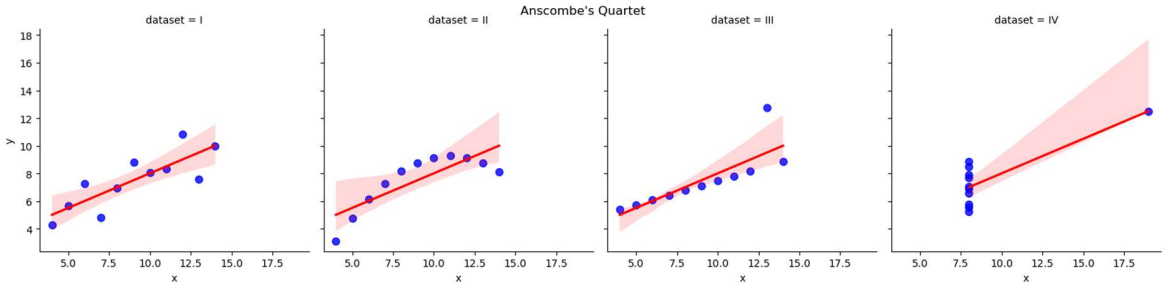
<Your answer for Question 7 goes here>

Anscombe's quartet is a powerful data synthesis which demonstrates the importance of necessity for data visualization. Developed by statistician Francis Anscombe in 1973 to explain how statistical parameters like mean, variance is some time misleading and visual analysis can provide a entire different interpretation of the data. The quartet consists of four datasets, each with 11 pairs of values for x and y.

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
				Summary Statistics				
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

As it is can be clearly visible from the table above that the statistical parameters are all identical for the data set.

The four data set are plotted, and the scatter plots are shown below



The following are interpreted from the above plots

Dataset 1: Show a linear relationship

Dataset 2: Show a nonlinear relationship

Dataset 3: Impact due to outliers

Dataset 4: Vertical line with outliers

The above plots and the table emphasize the fact that only statistical analysis is not enough and data visualization is a very import tool in understanding the data.

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 correlation coefficient is a statistical measure that evaluates the strength and direction of the relationship between two continuous variables. The Pearson's correlation coefficient is denoted by 'r' value varies between -1 and +1. A value of +1 shows a positive correlation between the variables whereas the value of -1 shows a negative correlation between the variables. 'r' having

a value 0 shows that the variables are not corelated.

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>

The data available for building a linear regression model is often in different ranges. Scaling is a pre procession step to normalize the values of input variables within a particular range.

Following are the major reasons to apply the scaling

a. To improve algorithm Performance: When features are on similar scales, it allows the algorithms to converge faster. Some algorithms such as gradient descent, k-nearest neighbors performs better when the data are in same range
b. To avoid dominance of one variable

Normalized Scaling: In this process the variables are brought into a specific range, usually 0 to 1.

Standardized scaling: It brings the data in the standard normal distribution by removing the mean. The values are replaced by its z score.

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>

VIF is a measure of multicollinearity between various independent variables in a data set. A high VIF shows that the variables are highly corelated.

When VIF is infinite it indicates there exists a perfect correlation between the variables. The VIF is computed by 1/(1-R2). When R2 is 1 the VIF becomes infinite.

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 plot (Q-Q) plot is a graphical technique to determine if a dataset is following a particular probability distribution. It provides an indication if two datasets are derived from the same population.

Use of Q-Q plot

Quantiles are points in a dataset that divide the data into intervals containing equal probabilities or proportions of the total distribution. They are often used to describe the spread or distribution of a dataset. The most common quantiles are (25th, 50th, and 75th percentiles): Quartiles divide the dataset into four equal parts. The first quartile (Q1) is the value below which 25% of the data falls, the second quartile (Q2) is the median, and the third quartile (Q3) is the value below which 75% of the data falls.

In a Q-Q plot, the x-axis represents the theoretical quantiles (from the assumed distribution, often the normal distribution), and the y-axis represents the sample quantiles (from your actual data).

A straight line at 45 deg is drawn on the Q-Q plot. If the data points fall along the line, the data is from same distribution.

On a broad level Q-Q plot is useful for

Checking the Normality of Residuals

Identifying Outliers

Assessing Homoscedasticity:

Evaluating Model Fit

Importance of Q-Q plot

Q-Q plot is an important tool to ascertain the normality of residuals.
