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)

Some of the categorical variable have significant impact on the dependent variable while others don't. such variables are,

- Season: Summer and Fall have positive impact on bike rental count. Spring has the least amount of bike rental bookings.
- Month: Similar trends were observed in months as well, warmer months (April to October) have higher rental count.
- Weather: There is a positive correlation between how good the weather is and bike rental. Days with rain have least number of rentals.
- Holiday: Though this does not have major significance but there is a slight drop in number of rentals on holiday which is an interesting insight.

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 to true drops the first dummy column. As we only need n-1 columns for a given categorical field which supports n different types of values.

For example, for a categorical column supporting three categorical values (A, B, C), drop_first = True will only create two columns, B and C and drop A.

Interpretation:

B = 0, C = 0 => A

B = 1, C = 0 => B

B = 0, C = 1 => C

The default value for drop_first is False.

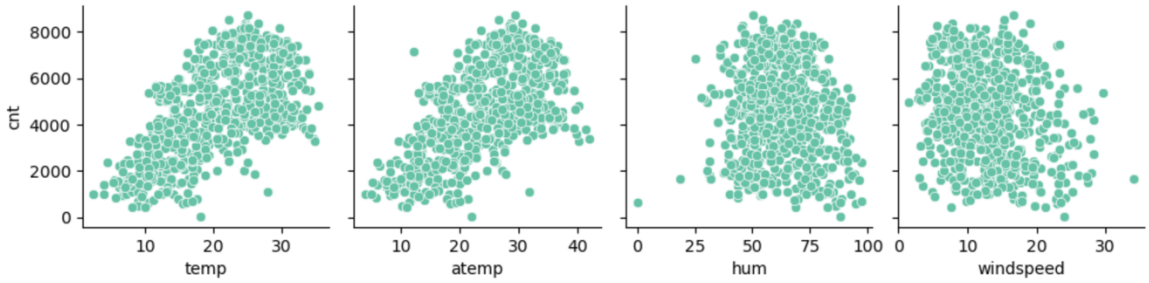
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

By looking at the pair-plot both temp and atemp seems to have highest correlation with the target variable

Pair Plot:



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 assumption of linear regression model has been validated be performing following steps.

- Validation of Linear relationship between X and Y
- Normal distribution of Error terms verified using the distribution plot.
- Independence of Error terms verified using residual analysis
- Constant Variance (homoscedasticity): Using the regplot visualization.

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)

Temp (coff: 0.449)
Light_Rain (coff: -0.285)
Yr (coff: 0.234)

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 statistical method for predicting a continuous dependent variable based on one or more independent variable. It assumes that there is a linear relationship between the independent variable (aka: predictor variable) and dependent variable (aka: target variable).

Formula:

For Simple Linear Regression:

$Y=\beta_0+\beta_1 X+\epsilon Y$

For Multiple Linear Regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \ldots + \beta_n X_n + \epsilon$$

Where,

Y is the dependent variable.

X is the independent variable

β_0 is the intercept which represent values of Y when X is 0. (Pronounced as "Beta Not")

$\beta_1, \beta_2, \beta_3 \ldots \beta_n$ is the slope, which represent the change in Y for a unit change is $X_1, X_2, X_3 \ldots X_n$ respectively.

ϵ (pronounced as epsilon) is the error term (residual).

For a linear regression to work effectively, following assumption must be true,

a. There is a linear relationship between the dependent and independent variables.
b. The residual error should be independent of each other.
c. Error should have constant variance.
d. Error terms should be normally distributed.
e. When there are multiple predictors in play, they shouldn't be highly correlated with each other.

The goal of linear regression is to find the best fitting line that minimizes the error. The most common method for estimating the parameters β_0 (intercept) and β_1 (slope) is Ordinary Least Squares (OLS).

$$RSS = (y_1 - \hat{y}_1)2 + (y_2 - \hat{y}_2)2 + (y_3 - \hat{y}_3)2 + \ldots + (y_n - \hat{y}_n)2$$

or, $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Where,

RSS is Residual Sum of Square

y_i is the actual observed value.

$\hat{y}_i$ is the predicted value for observation.

Once the model is fitted, it is crucial to evaluate how well it performs, common metric used for for evaluating Linear Regression models are,

a. R-Squared: It gives us the proportion of variance in the dependent variable that is explained by the independent variables (s). It ranges from 0 and 1 where 0 means the model explain none of the variance and 1 means the model explains all variances.

Formula:

$$\text{R-squared} = 1 - \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right) / \left(\sum_{i=1}^{n} (y_i - \bar{y}_i)^2\right)$$

Where $\bar{y}$ (y bar) is the mean of observed value.

b. Adjusted R-Squared: This value penalizes model which has higher number of predicators and observations.

$$\text{Adjusted } R^2 = 1 - ((1 - R^2)(n-1) / (n-p-1))$$

Where:
- n is the sample size meaning the number of rows you'd have in the dataset
- p is the number of predictor variables.

c. Mean Square Error (MSE): MSE is another common metric that quantifies the average squared difference between predicted and actual values.

$$MSE = 1/n(^{n}\sum_{i=1} (y_i - \hat{y}_i)^2)$$

d. Residual Analysis: By plotting the residuals, we can visually inspect if the assumptions of the linear regression model are met.

Following are some of the limitations of Linear Regression

- Linear regression assumes a linear relationship between the dependent and independent variables, if the relationship is non-linear, the model may not perform well.
- Linear regression is sensitive to outliers.
- When independent variables are highly correlated, it can make it difficult to estimate the coefficients accurately.
- In multiple regression, using to many predictors can lead to overfitting, where the model performs well on training data but poorly on unseen data.

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 refers to a set of four distinct datasets that have nearly identical simple descriptive statistics (such as mean, variance, and correlation) but exhibit very different patterns when plotted visually.
It was introduced by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphical analysis in understanding data and to highlight the potential pitfalls of relying solely on summary statistics, such as mean and correlation, to describe data.

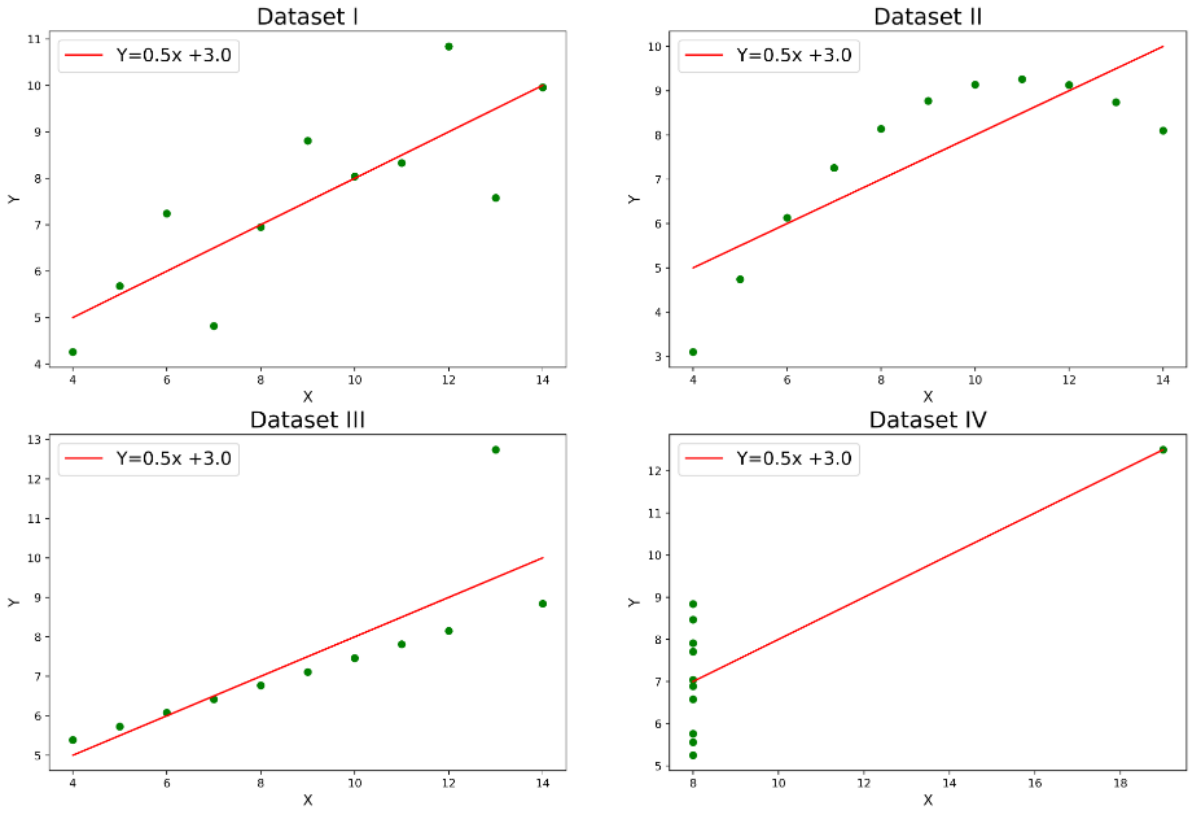
Dataset:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Result:

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

Visual Representation:



Dataset I: This dataset exhibits a nearly perfect linear relationship between X and Y. The points follow a straight line with some small random noise.

Dataset II: This dataset shows strong non-linear relationship. Even though the summary statistics are like Dataset I, the pattern is clearly non-linear

Dataset III: Shows almost perfect linear relationship between X and Y except 1 outlier. This shows despite having similar statistical value the presence of the outlier skeys the result of linear regression making it misleading.

Dataset IV: Perfect vertical alignment of points except one outlier. The correlation may suggest a strong relationship, but the relationship is not meaningful for linear regression.

Conclusion: Although summary statistics like mean, variance, and correlation give a quick overview of the data, they can be misleading without a visual inspection. The scatter plot can reveal patterns, outliers, and relationships that summary statistics cannot capture.

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 statistical measure that describes the strength and direction of the linear relationship between two variables. It quantifies how well the variation in one variable can be explained by the variation in the other.

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2]\ \ [N\sum y^2 - (\sum y)^2]}}$$

Where,

N = Number of pairs of scores

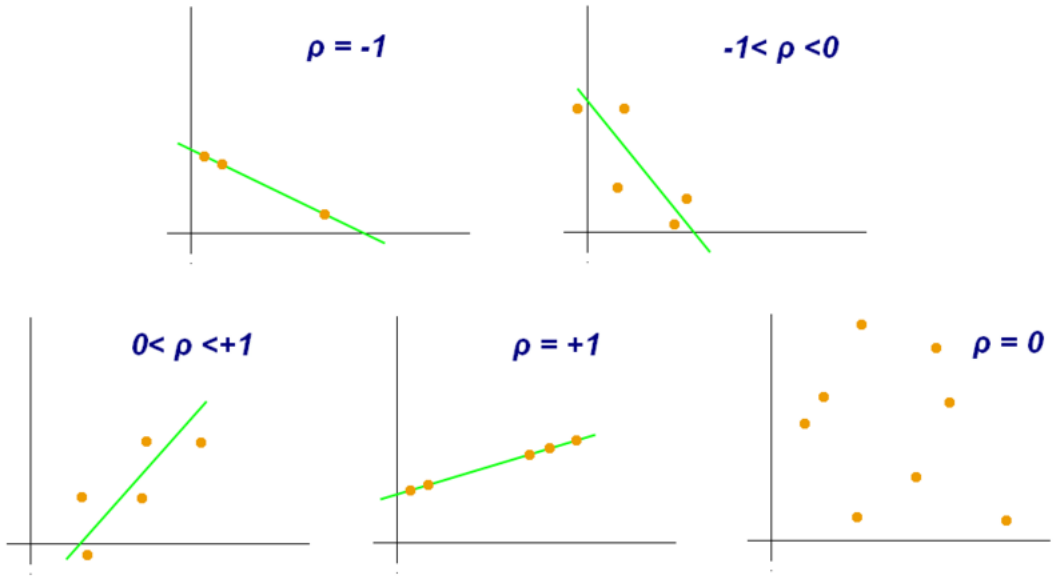
$\sqrt{x}$= Sum of x Scores

$\sqrt{y}$= Sum of y scores

$\sqrt{xy}$= sum of the products of paired scores

$\sqrt{x^2}$= sum of the squared x scores

$\sqrt{y^2}$= sum of the squared y scores



The value of Pearson's r ranges from -1 to 1, where:

- r = 1: Perfect positive linear relationship. As X increases, Y increases in a perfectly straight line.
- r = -1: Perfect negative linear relationship. As X increases, Y decreases in a perfectly straight line.
- r = 0: No linear relationship between the variables. There may still be a non-linear relationship, but Pearson's r only captures linear relationships.
- r > 0: Positive correlation. As one variable increases, the other tends to increase.
- r < 0: Negative correlation. As one variable increases, the other tends to decrease.

For Pearson's r to be a valid measure of correlation, certain assumptions need to be met:

1. Linearity: The relationship between the two variables should be linear.
2. Normality: Both variables should be approximately normally distributed.
3. Homoscedasticity: The variability of one variable should be roughly consistent across the values of the other variable. If the variance in Y increases as X increases (or vice versa), the data may

violate this assumption.
4. No outliers: Pearson's r is sensitive to outliers. A single extreme data point can drastically change the value of r and may mislead the interpretation of the relationship.

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>
Feature scaling refers to the process of transforming the feature of a dataset so that they have a common scale.

Scaling is done to bring all feature to same scale. Without scaling, features with larger numerical ranges or different units can dominate the model's learning process, potentially leading to biased results or poor performance.

Normalized Scaling v/s Standardized Scaling

- Normalized scaling scales the data to a fixed range between 0 and 1 whereas, standardized scaling transforms the data so that each feature has a mean of 0 and a standard deviation of 1.
- Normalized scaling gets significantly affected by outlier which is not the case in standardized scaling.

Normalized Scaling (also called Min-Max scaling)

$$= (X_i - X_{min})/(X_{max} - X_{min})$$

Standardized Scaling (also called Z-Score Normalization)

$$= (X_i - \mu)/ \sigma$$

Where,

- X_{min} and X_{max} are the min and max values of the feature
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

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>

An infinite Variance Inflation Factor (VIF) signifies a perfect multicollinearity between one or more independent variables.

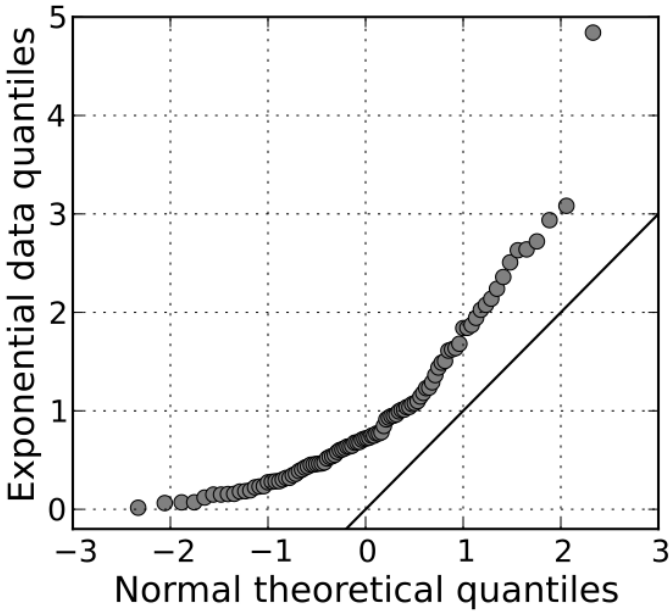
VIF is calculated using $1/(1-R^2)$, VIF will be infinite if $R^2 = 1$,

This can be solved by identifying and dropping independent variable causing 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)

<Your answer for Question 11 goes here>

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as a normal distribution, or to compare the distributions of two datasets.



In a Q-Q plot, we plot:
- Theoretical Quantiles on the x-axis.
- Empirical Quantiles (from the data) on the y-axis.

Where,

Quantiles are points that divide the data into intervals with equal probabilities.

Theoretical Quantiles: These are the quantiles of the theoretical distribution we want to compare the data to.

Empirical Quantiles: These are the quantiles of the actual dataset. In a Q-Q plot, these quantiles are plotted on the vertical axis.

Q-Q plot in linear regression can be interpreted as either,

Points on the line: If the points lie close to the line, the residuals are approximately normal, and the normality assumption is likely satisfied. Or,

Systematic deviations:

- Heavy tails (leptokurtic): If the points deviate upwards and downwards (away from the line at both ends), this suggests that the residuals have heavier tails than the normal distribution, which could indicate outliers or non-normality in the distribution of residuals.
- Skewness: If the points curve away from the line in one direction (e.g., upward or downward), it suggests that the residuals are skewed—either positively or negatively.
- S-shaped pattern: A pattern where the points curve away from the line in an S-shape might indicate a bimodal distribution or other complex distributional issues in the residuals.