Assignment Based Subjective Questions:

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans1. In the dataset following categorical variable was provided:

1. Season (Spring, Summer, Fall and Winter)
    a. Based on my analysis and model coefficient I can infer that season has an impact on the demand for rental bike.
    b. During Spring the usage is likely to drop in comparison with other seasons
2. Year – Although the attribute year has only two values i.e. 0 and 1 where 0 represent year of 2018 and 1 represent year of 2019. Since this categorical variable has only 2 values, there was no need to create dummy variable for it. In 2019 the demand was higher than 2018, which suggest that company might have taken some steps in 2019 which resulted in higher demand.
3. Month – It contain values from 1 to 12, however based on my analysis the information is better captured via variable season and hence I have not used this variable
4. Holiday, weekday and working day, these are three categorical variables where
    a. Holiday is a binary variable which indicates if it's a holiday or not
    b. Weekday has 7 values from 0 to 6, one for each day of week
    c. Working day is a binary variable its 1 if it is a working day and it is 0 if it is holiday or weekend
    d. Upon further analysis we found that the categorical variable working day represents the information of both holiday and weekday and hence we dropped the variable holiday and weekday.
    e. Inference based on working day – Demand is more during a working day as compared to a non-working day
5. Weather situation (Clear, Cloudy and Rainy)
    a. Weather situation have an impact on demand for rental bike
    b. Rainy weather impact the demand negatively whereas clear weather impact the demand positively and cloudy weather sits somewhere in middle

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans2. When we use the pandas getdummies() function, it converts a categorical variable into a one hot encoding data frame, with one binary column for each unique value of the categorical variable. However if a categorical variable has N unique values then it can be represented by N-1 binary variable and we do not need to have N binary variable. Also if we use all N binary variable then one of the variable can be fully explained by other N-1 variable leading to much higher VIF and making the overall interpretation harder.

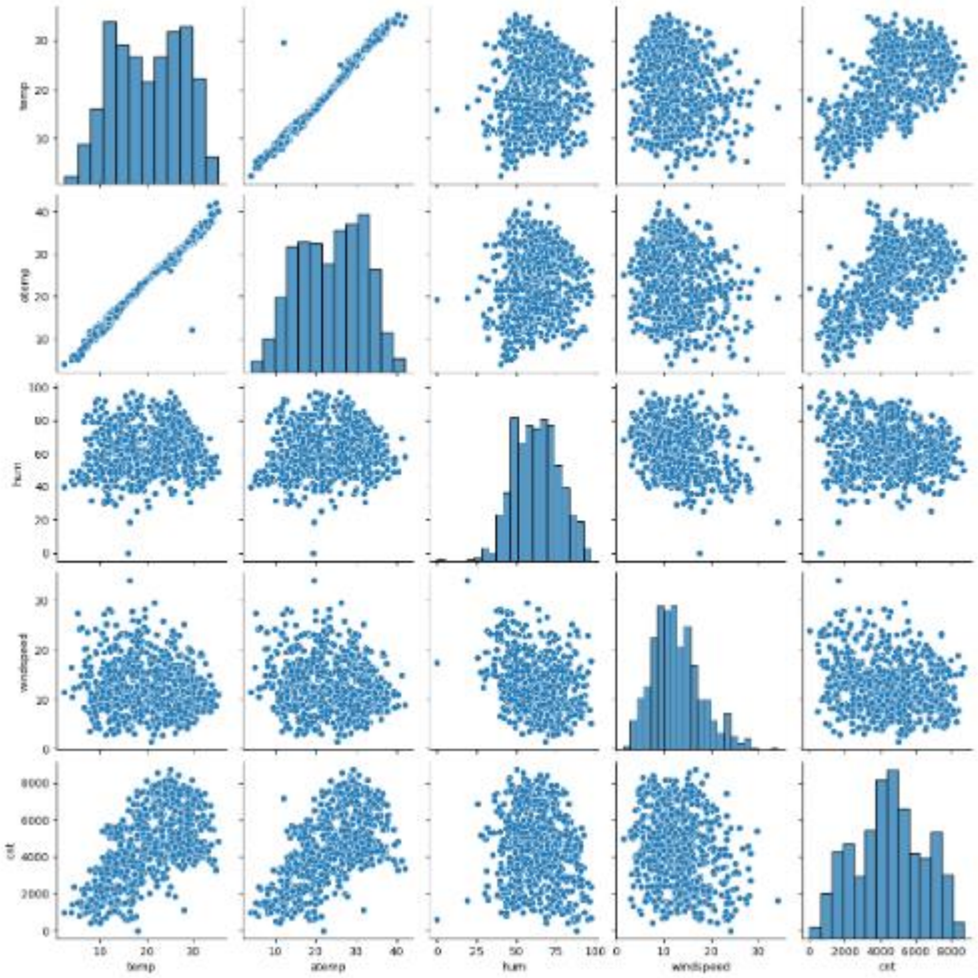
Example:

Categories	Representing Categories with 4 (N) binary variables				Representing Categories with 3 (N-1) binary variables		
	A	B	C	D	B	C	D
A	1	0	0	0	0	0	0

B	0	1	0	0	1	0	0
C	0	0	1	0	0	1	0
D	0	0	0	1	0	0	1

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans3. Temperature (Temp) and Feeling Temperature (Atemp) has the highest correlation.

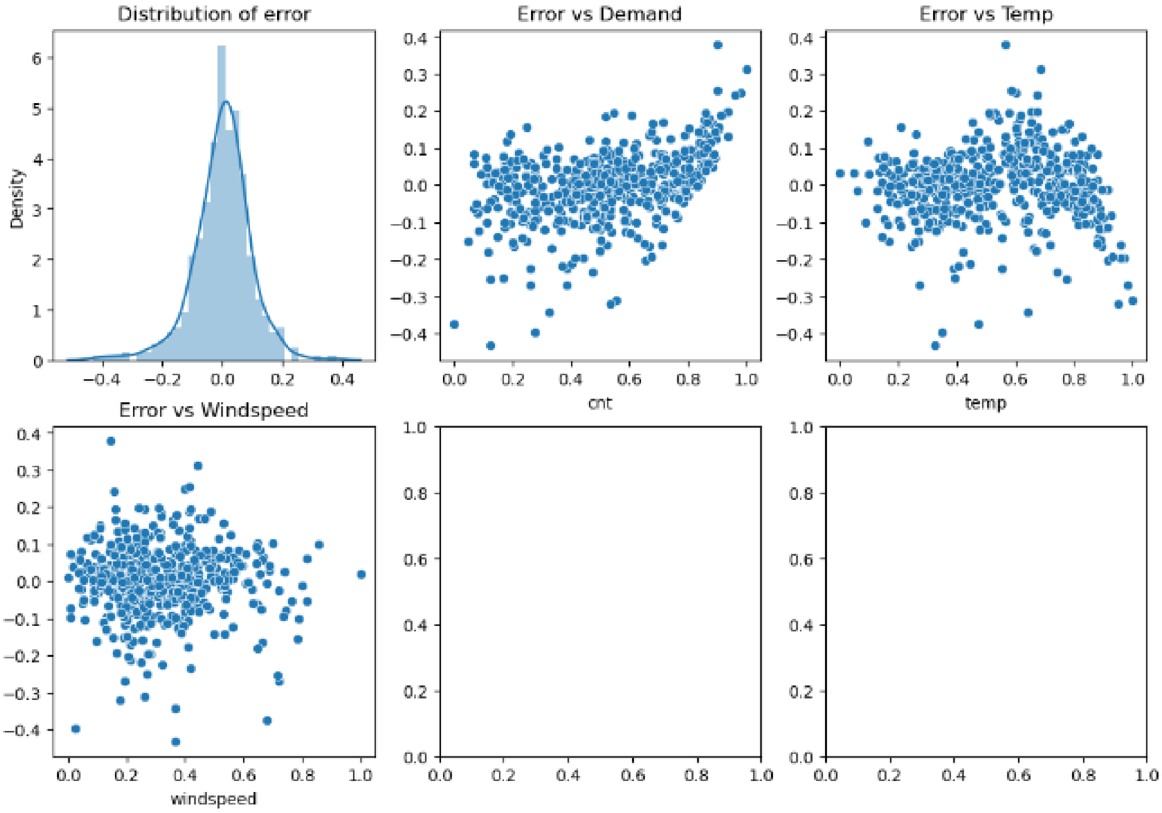


Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans4. Assumptions for linear regression:

1. There should be some linear relationship between dependent and independent variable. I have validated this through pairplot and heatmap
2. Error should be normally distributed with mean of 0, this I have validated by plotting the distribution of error
3. Error should be independent of each other, this I have validated by scatter plot of dependent variable on x-axis and error on y-axis
4. Error should have constant variance, this I have validated by scatter plot of independent variable on x-axis and error on y-axis

5. Plots used for assumption validation:



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans5. Top 3 features of my final model are:

1. Temperature of the day (Coefficient : 0.466252)
2. Rainy day (Coefficient : -0.277670)
3. Year (Coefficient: 0.233794)

General Subjective Questions:

Q1. Explain the linear regression algorithm in detail.

Ans1. [Reference used: Upgrade material and AWS website]

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data. When there is only one independent feature, it is known as Simple Linear Regression, and when there is more than one feature, it is known as Multiple Linear Regression.

Simple Linear Regression equation: $y=\beta_0+\beta_1 X1$

Multiple Linear Regression equation: $y = \beta_0+\beta_1 X1+\beta_2 X2+\ldots\ldots\ldots\beta_n Xn$

Where:

- Y is the dependent variable
- X1, X2, …, Xn are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line. The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables.

The best fit line is obtained by minimising a cost function or loss function, for linear regression we use OLS or Ordinary Least Square function to find the coefficient of best fit line.

Q2. Explain the Anscombe's quartet in detail

Ans2. [Reference used: Geek for Geeks]

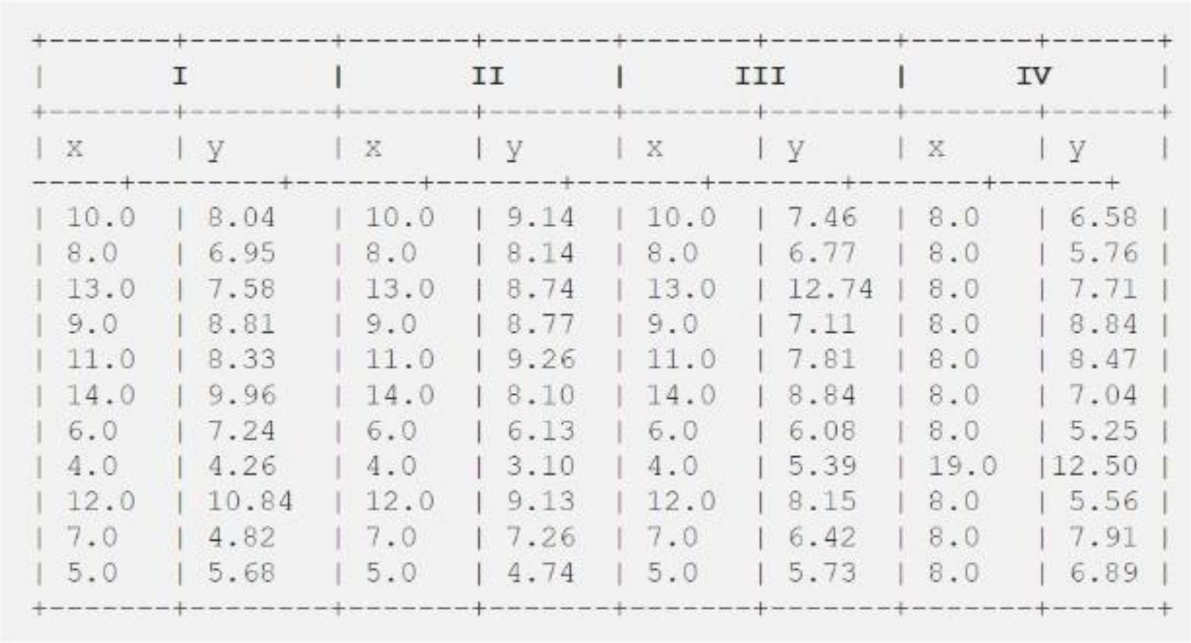
Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

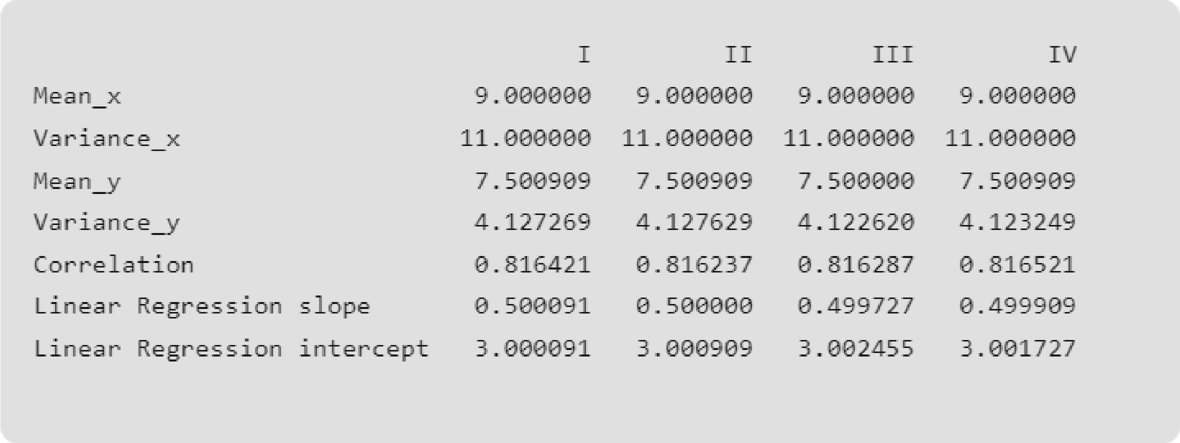
Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe’s Quartet Dataset:



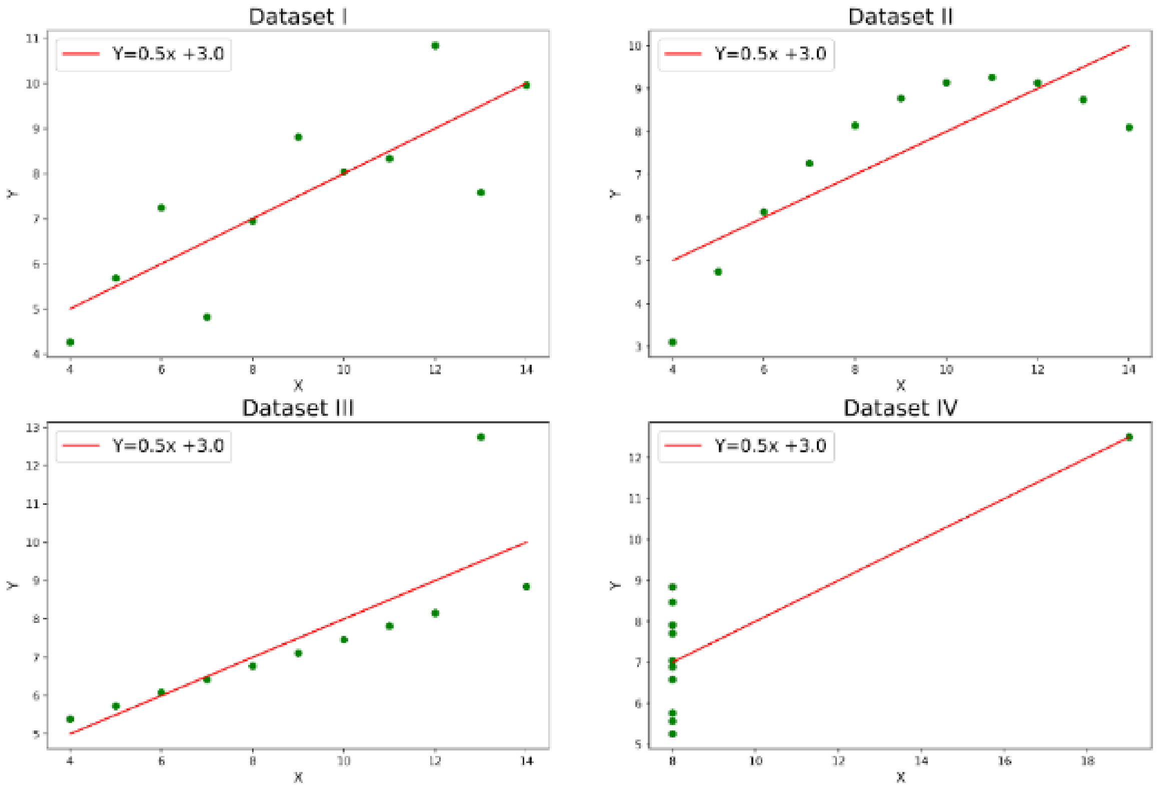
```
+-------+--------+-------+-------+--------+-------+-------+------+
|       I        |       II      |       III      |      IV      |
+-------+--------+-------+-------+--------+-------+-------+------+
| x     | y      | x     | y     | x      | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 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 |
+-------+--------+-------+-------+--------+-------+-------+------+
```

Statistical summary:



```
                                    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
```

The scatter plot and linear regression line for each datasets:



Anscombe's quartet Plot

Conclusion:

While the descriptive statistics of Anscombe's Quartet may appear uniform, the accompanying visualizations reveal distinct patterns, showcasing the necessity of combining statistical analysis with graphical exploration for robust data interpretation.

Q3. What is Pearson's R?

Ans3. [Reference Used: Statistics laerd]

Pearson correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r. Basically, Pearson correlation coefficient attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

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.

The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively.

Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for r between +1 and -1 (for example, r = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of r to 0 the greater the variation around the line of best fit.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans4. [Reference Used: Wikipedia]

Scaling:

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

Why it is used:

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence making the model less interpretable. Also it has been observed that scaling helps reach the minimum point of a cost function quicker there by making learning faster and more cost effective. Hence to make model more interpretive and achieve efficient learning, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalized Scaling or Min Max scaling It brings all of the data in the range of 0 and 1.

Min_max_scaling(x) = { x – min(X) } / { max(X) – min(X) }

Standardization Scaling replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ)

Standardize x = { x – mean(X) } / SD(X)

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans5. It happens when the variability in one of the independent variable can be completely explained by other variables in the dataset.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

Ans6. A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, most commonly the normal distribution. It compares the quantiles of the data to the quantiles of a specified theoretical distribution. If the data comes from the theoretical distribution, the points on the Q-Q plot will approximately lie on a straight line. In the context of linear regression, the Q-Q plot is primarily used to assess the normality of the residuals. Normality of residuals is an important assumption in linear regression, especially for hypothesis testing and constructing confidence intervals.