Assignment-based Subjective Questions

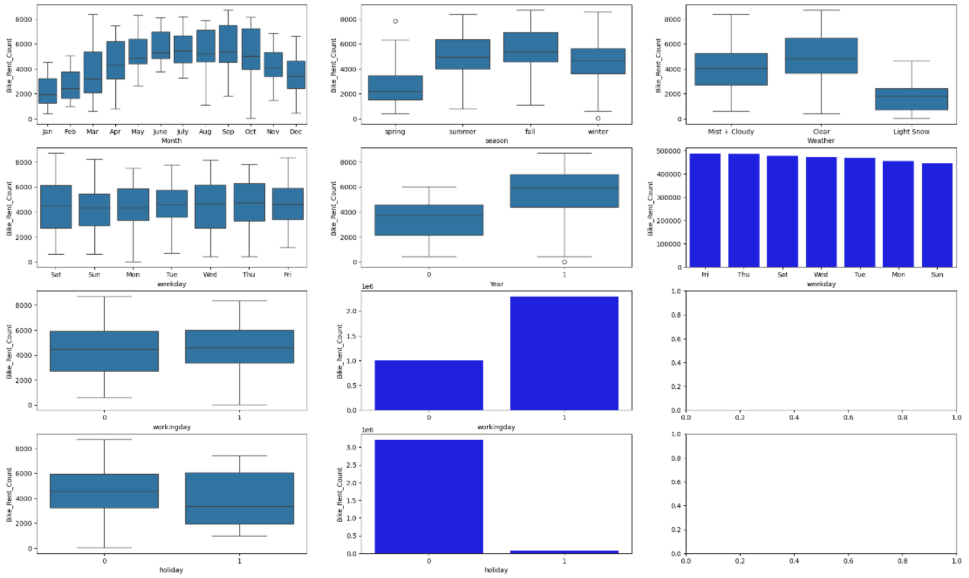
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

As per the graph below of categorical variables, it is evident that

1. 1st row of the graph on **month**, **season** and **weather** it is clear that bike is rented more in **warm weather** which is correlated with **feelsliketemperature**
2. 2nd-row weekday average rent is almost the same across the days hence, with almost no impact on the dependent variable, hence will be dropping this column
3. 2019 has seen a spike in bike rental
4. Workday has higher rented bikes than non-working days. The same trend is reflected for Holiday



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)

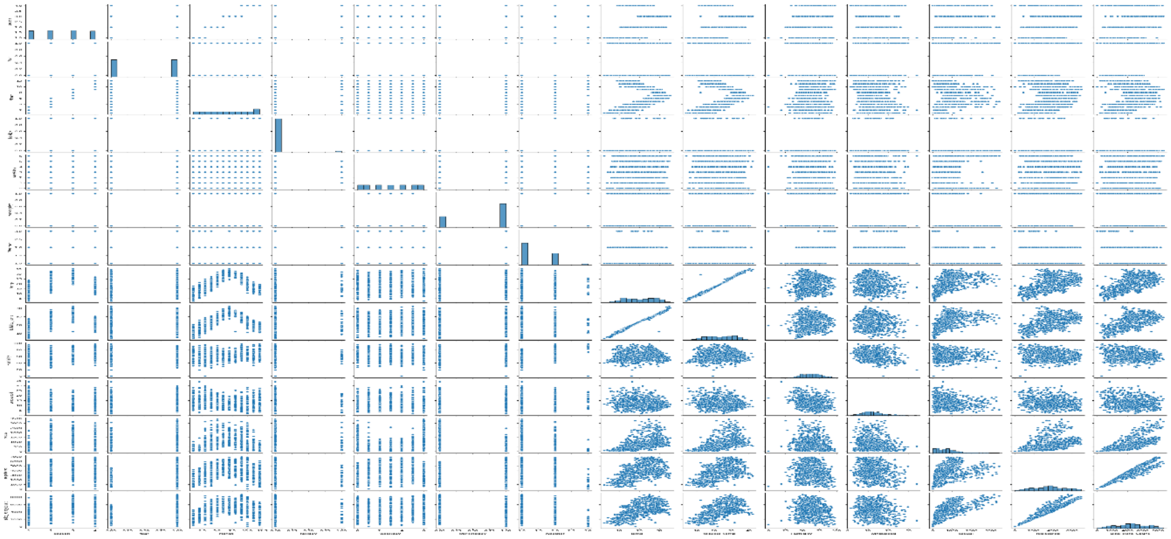
drop_first = true removes the first column which is created for the first unique value of a column. This avoids redundancy.

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)

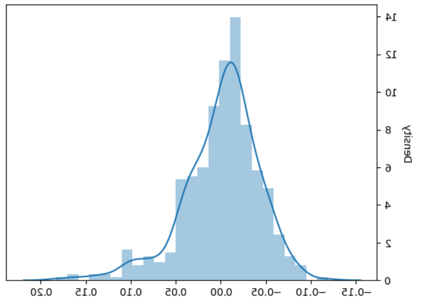
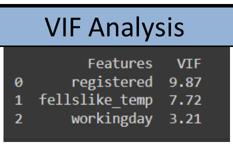
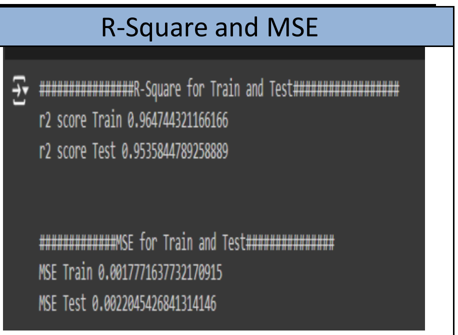
Based on Pair plot it is clear that, registered, causual, Temp and feelslikeTemp has clear pattern of linear regression



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)

Assumptions validated by plotting normal distribution curve of error, VIF analysis, MSE analysis and R-Square Analysis

NDC of Residual	VIF Analysis	R-Square and MSE

	`Features VIF`
	`0 registered 9.87`
	`1 fellslike_temp 7.72`
	`2 workingday 3.21`
		`##############R-Square for Train and Test################`
		`r2 score Train 0.964744321166166`
		`r2 score Test 0.9535844789258889`
		`############MSE for Train and Test#############`
		`MSE Train 0.0017771637732170915`
		`MSE Test 0.0022045426841314146`

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)

Top 3 features contributing significantly over demand are, Registered Users,Feels Like Temp, and workingday.

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>

A Linear equation is the equation of the First Order. It predicts the target or dependent variable based on given or identified features It's of two types simple linear equation (y= $\beta_0 + \beta_1$ X) and multivariable linear equation (y= $\beta_0 + \beta_1$ X + β_2 X + β_3 X + ………+ β_n X). Simple linear equation predicts with one variable however multivariable linear equation has many variables. A linear line showing the relation between the target and the feature is called a regression line. Positive linear equation when target variable increases with X and negative linear equation target variable decreases with increase of value in X

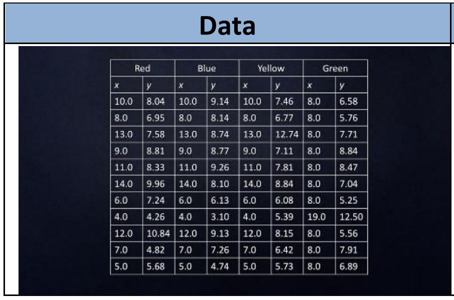
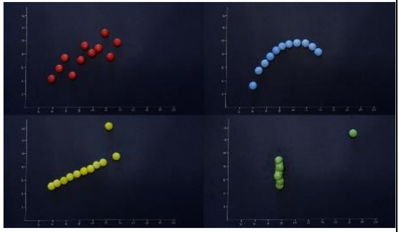
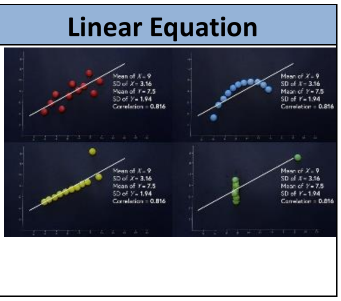
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Quartet shows how four entirely different data sets can be reduced down to the same summary metrics. Anscombe's Quartet 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 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.

Data	Scatter Graph	Linear Equation
		

Red		Blue		Yellow		Green	
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

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>

The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

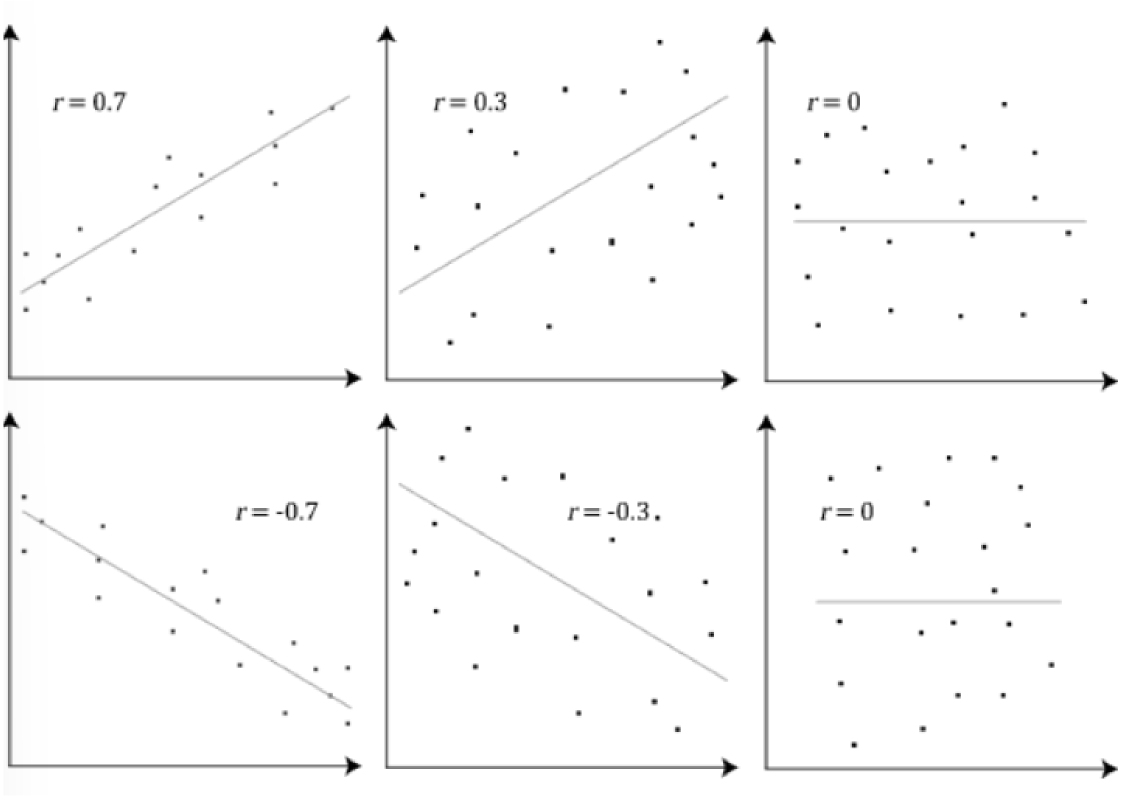
- Pearson’s r
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

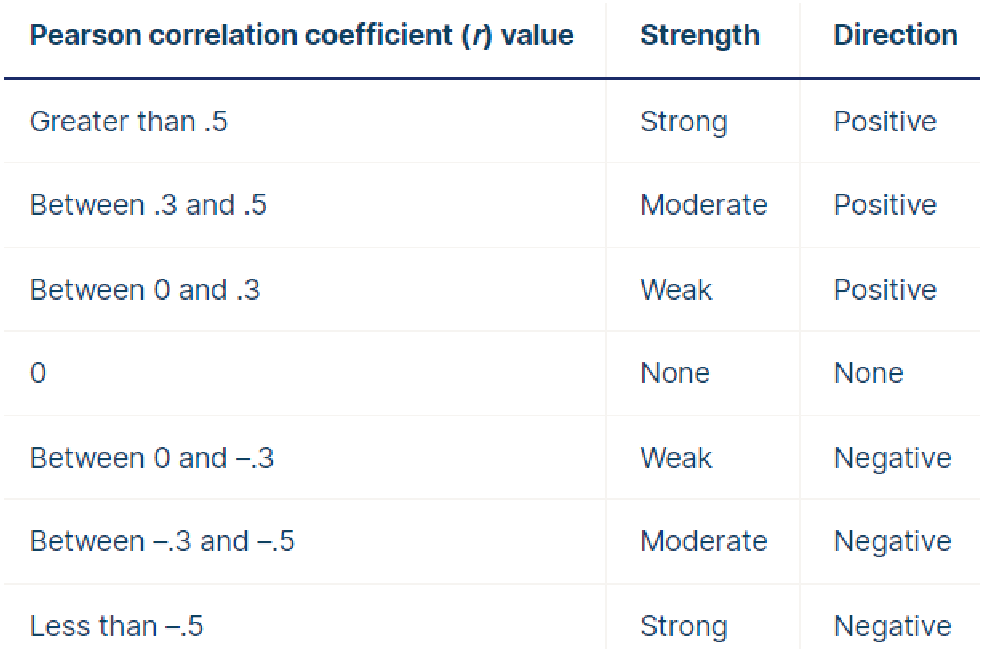
Pearson correlation coefficient (PCC)[a] is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

Formula is

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$





Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

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 technique used during preprocessing in ML analysis. This is the process of transforming data values to a similar range typically between 0 to 1. This ensures that all features contribute to the model equally.
Scaling is done to improve the performance of the model by preventing features with large values dominating the performance.
Normalization/min-max scaling: Aims to scale data between 0 to 1 by subtracting the minimum value and diving by range (max-min). This can be affected by outliner
Standardization: Scale data to have mean as 0 and standard deviation as 1, making it less sensitive towards outliners. It is often preferred when data follows normal distribution curve.

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 value is considered infinite when there is perfect multicollinearity, meaning one independent variable in the regression model can be perfectly predicted by a linear combination of other independent variables, creating an exact duplicate. To address this, remove one independent variable.

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 is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.