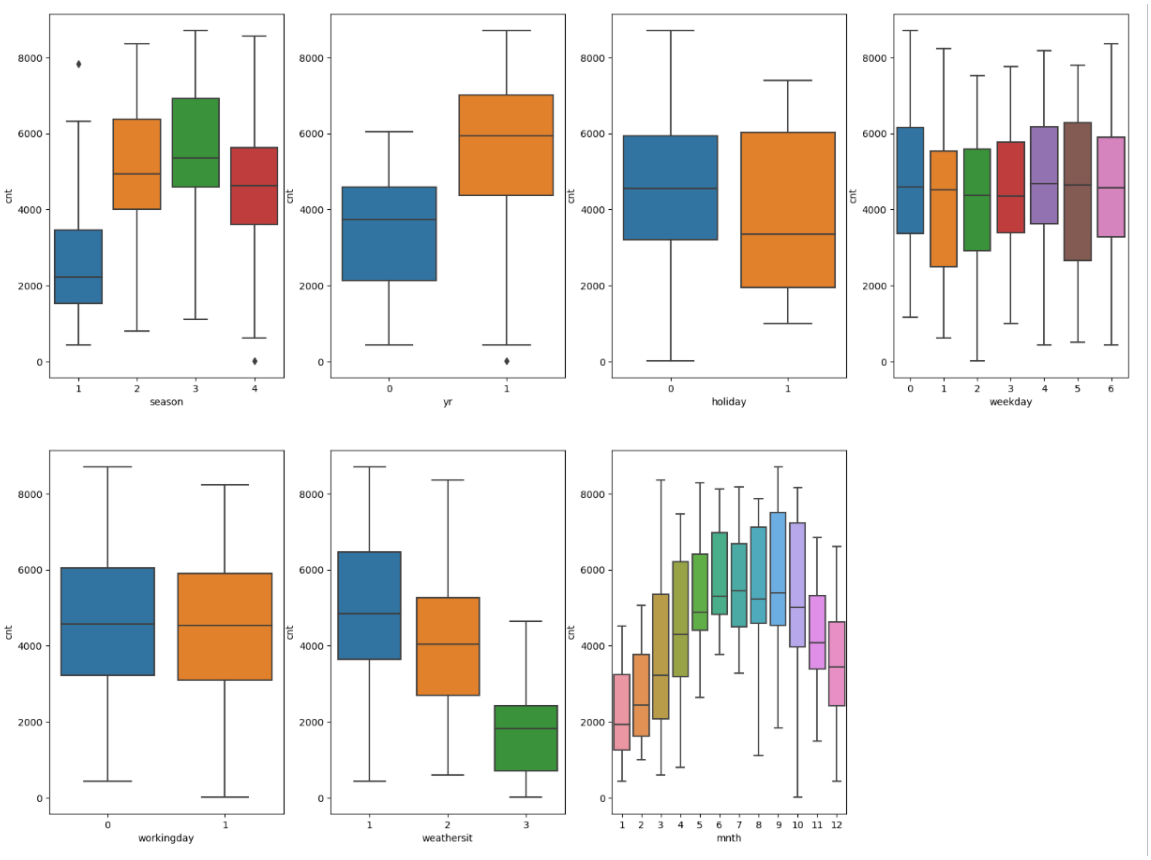
Assignment-based Subjective Questions

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

Ans:



- The dataset contains several categorical variables: season, year, holiday, weekday, working day, weathersit, and month. These variables were visualized using a boxplot (see attached figure), and they have different effects on our dependent variable. Here's a simpler explanation of the findings:
- Season: The boxplot shows that the spring season had the lowest number of rentals, while fall had the highest number. Summer and winter had a moderate number of rentals.
- Weathersit: When there is heavy rain or snow, there are no users renting bikes, indicating that this weather condition is extremely unfavorable. The highest number of rentals occurred when the weather was clear or partly cloudy.
- Year: There were more rentals in 2019 compared to 2018.
- Holiday: The number of rentals decreased during holidays.
- Month: September had the highest number of rentals, while December had the lowest. This observation aligns with the weathersit variable, as December usually experiences heavy snowfall, which may have resulted in fewer rentals.
- Weekday: The number of rentals remained relatively consistent throughout the week.
- Workingday: The median number of users renting bikes remained constant throughout the week.

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

Ans: It is important to use drop_first=True during dummy variable creation to avoid redundancy.

- When we create dummy variables from categorical variables, a separate column is created for each category. However, including all the columns can lead to redundancy because one category can be represented by a combination of the other categories.
- By using drop_first=True, we drop one of the dummy variable columns, which helps to eliminate redundancy. This ensures that each category is represented independently without introducing unnecessary duplication or redundancy in the data.

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

Ans: From the pair-plot analysis of the numerical variables, it was determined that the "temp" variable has the strongest correlation with the target variable.

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

Ans: The assumptions of Linear Regression were validated after building the model on the training set by conducting several checks. These checks included examining the variance inflation factor (VIF) to assess multicollinearity, analyzing the distribution of residuals to verify the assumption of error distribution, and assessing the linearity between the dependent variable and each feature variable.

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

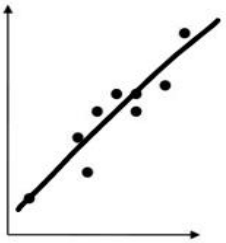
Ans: Based on the final model, the three most significant features that contribute to explaining the demand for shared bikes are the temperature, the year, and the holiday variables.

General Subjective Questions

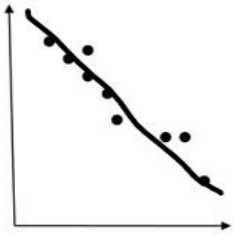
1. **Explain the linear regression algorithm in detail.?**

Ans: Linear regression is a statistical model that examines the linear relationship between a dependent variable and a set of independent variables. This means that when the values of the independent variables change (increase or decrease), the value of the dependent variable also changes accordingly (increase or decrease).

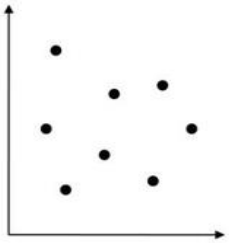
- The relationship can be represented by the equation:
- $Y = mx + c$
- Positive Linear Relationship: A linear relationship is considered positive when both the independent and dependent variables increase. This can be visualized in a graph where the line slopes upwards.
- Negative Linear Relationship: A linear relationship is considered negative when the independent variable increases and the dependent variable decreases. This can be visualized in a graph where the line slopes downwards.



Positive correlation



Negative correlation



little or no correlation

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans: Anscombe's quartet is a collection of four datasets that have similar statistical properties, such as means, variances, and correlations, but differ significantly in their distributions and visual appearance. Each dataset consists of eleven data points. The main purpose of Anscombe's quartet is to highlight the importance of examining data graphically before drawing conclusions or performing statistical analysis. Merely relying on summary statistics can be misleading, as visually inspecting the data allows for a more accurate understanding of the patterns and relationships within the datasets.

	Features	VIF
0	const	82.48
1	season_Spring	5.27
13	temp	4.42
3	season_Winter	3.83
2	season_Summer	2.76
14	hum	1.93
7	mnth_Nov	1.76
5	mnth_Jan	1.68
10	weathersit_Mist & Cloudy	1.57
4	mnth_Dec	1.49
6	mnth_Jul	1.49
8	mnth_Sep	1.34
9	weathersit_Light Snow & Rain	1.26
15	windspeed	1.21
11	yr	1.04
12	holiday	1.03

	Features	VIF
13	hum	30.94
12	temp	17.80
14	windspeed	4.72
0	season_Spring	4.37
2	season_Winter	4.06
1	season_Summer	2.81
9	weathersit_Mist & Cloudy	2.32
10	yr	2.09
6	mnth_Nov	1.83
4	mnth_Jan	1.75
5	mnth_Jul	1.59
3	mnth_Dec	1.55
7	mnth_Sep	1.41
8	weathersit_Light Snow & Rain	1.28
11	holiday	1.06

	Features	VIF
12	temp	5.17
13	windspeed	4.67
2	season_Winter	2.94
0	season_Spring	2.89
1	season_Summer	2.23
10	yr	2.07
6	mnth_Nov	1.80
4	mnth_Jan	1.66
5	mnth_Jul	1.59
9	weathersit_Mist & Cloudy	1.56
3	mnth_Dec	1.46
7	mnth_Sep	1.35
8	weathersit_Light Snow & Rain	1.09
11	holiday	1.06

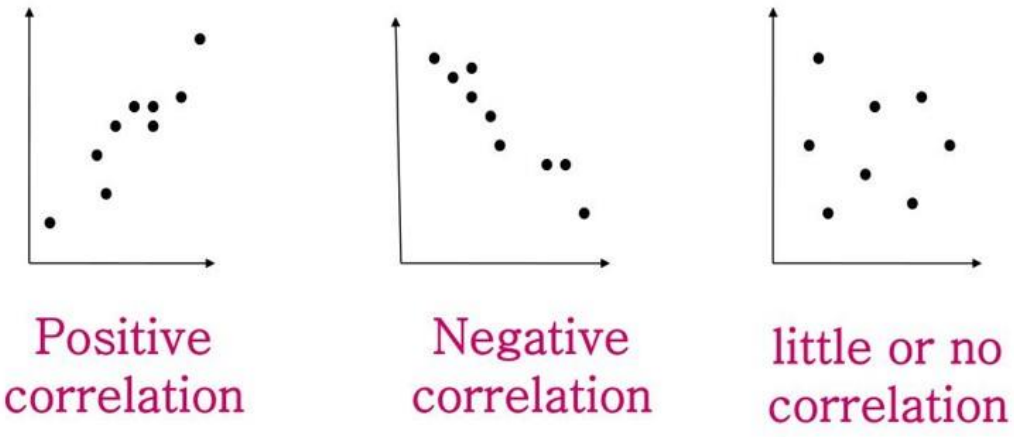
	Features	VIF
12	temp	5.17
13	windspeed	4.67
2	season_Winter	2.94
0	season_Spring	2.89
1	season_Summer	2.23
10	yr	2.07
6	mnth_Nov	1.80
4	mnth_Jan	1.66
5	mnth_Jul	1.59
9	weathersit_Mist & Cloudy	1.56
3	mnth_Dec	1.46
7	mnth_Sep	1.35
8	weathersit_Light Snow & Rain	1.09
11	holiday	1.06

3. What is Pearson's R?

Ans: Pearson's correlation coefficient, also known as Pearson's R, is a statistical measure used to determine the strength and direction of the linear relationship between two variables. It provides a numerical value between -1 and +1, indicating the extent of the correlation.

- A positive value close to +1 indicates a strong positive correlation, meaning that as one variable increases, the other variable also tends to increase.
- A negative value close to -1 indicates a strong negative correlation, meaning that as one variable increases, the other variable tends to decrease.
- A value close to 0 indicates a weak or no linear relationship between the variables.

- In simple terms, Pearson's R helps us understand how closely two variables are related and in what direction (positive or negative) that relationship exists.



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

Ans: Scaling is a preprocessing technique used to standardize or normalize the independent feature variables in a dataset to a consistent range.

- Scaling is performed to address issues that may arise due to varying magnitudes and units of different features in the dataset. If scaling is not performed, features with larger magnitudes can dominate the modeling process, leading to biased results.
- Normalized scaling, also known as Min-Max scaling, transforms the values of each feature to a range between 0 and 1. This is achieved by subtracting the minimum value of the feature and dividing by the range (the difference between the maximum and minimum values).
- Standardized scaling, also known as Z-score scaling, transforms the values of each feature to have a mean of 0 and a standard deviation of 1. It replaces the values with their respective Z-scores, which are calculated by subtracting the mean and dividing by the standard deviation.
- The main difference between normalization and standardization is the resulting range of values. Normalization brings all the data points within the range of 0 to 1, while standardization transforms the values to have a mean of 0 and a standard deviation of 1. The choice between the two scaling methods depends on the specific requirements of the modeling task and the nature of the data.

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

Ans: Sometimes, the value of VIF (Variance Inflation Factor) can be infinite. This happens when there is a complete and perfect correlation between two independent variables. In other words, the two variables are so strongly related that one can perfectly predict the other. In such cases, the R-squared value, which represents the goodness of fit of the regression model, is equal to 1.

- The VIF is calculated as 1 divided by (1 minus the R-squared value). Since the R-squared value is 1 in this scenario, the denominator becomes 1 minus 1, which is 0. Division by 0 results in infinity. Therefore, the VIF value becomes infinite.
- This occurrence of infinite VIF suggests a problem called multicollinearity. Multicollinearity happens when there is high correlation among independent variables in a regression model. It can cause issues in interpreting the effects of individual variables and lead to unstable and unreliable coefficient estimates.
- To address the issue of multicollinearity, one of the correlated variables should be removed from the model. By eliminating one of the variables, it becomes possible to establish a well-defined regression model that avoids the problems caused by perfect correlation.

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

Ans: A Q-Q plot, short for quantile-quantile plot, is a type of graph used to determine if two sets of data come from populations that share a common distribution.

- To use a Q-Q plot, the quantiles (which represent the fractions or percentages of points below a given value) of the first data set are plotted against the quantiles of the second data set. The plot also includes a reference line at a 45-degree angle. If the two data sets come from populations with the same distribution, the points on the plot should roughly follow this reference line. However, if there is a significant departure from the reference line, it suggests that the two data sets may come from different distributions.
- The Q-Q plot is important because it helps assess whether the assumption of a common distribution is valid when working with two data samples. If the assumption holds, it allows for combining the data sets to obtain estimates of the common location and scale. On the other hand, if the two samples differ in terms of their distribution, the Q-Q plot provides insights into the nature of these differences. It can offer a better understanding of the variation between the data sets compared to analytical methods like chi-square and Kolmogorov-Smirnov 2-sample tests.