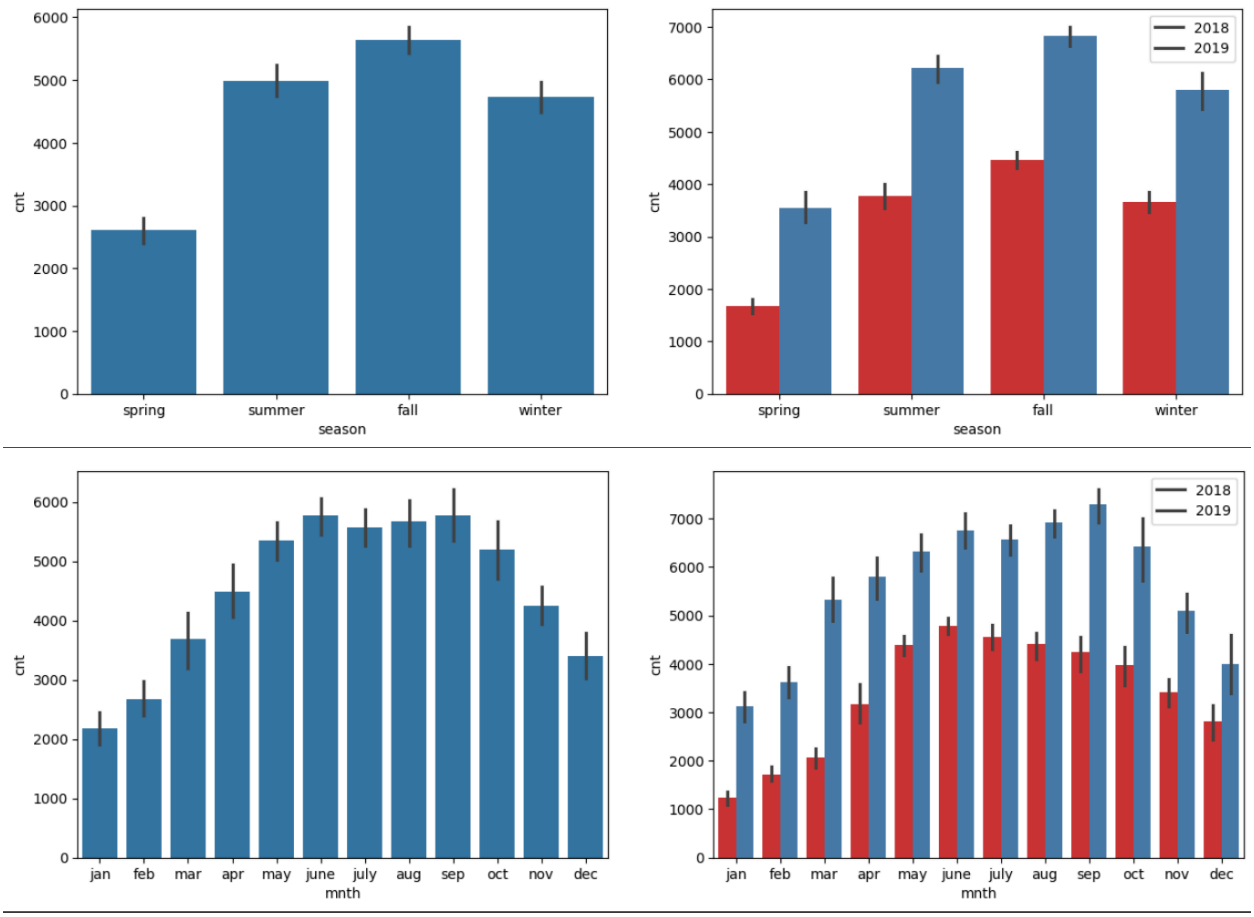
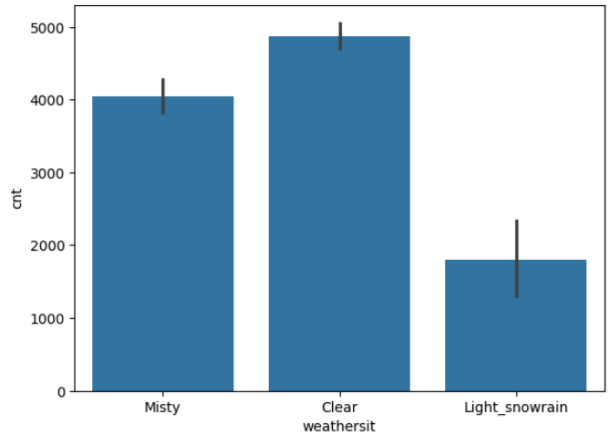
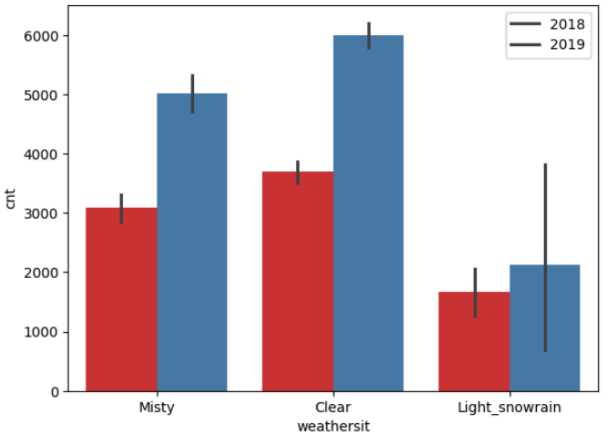
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

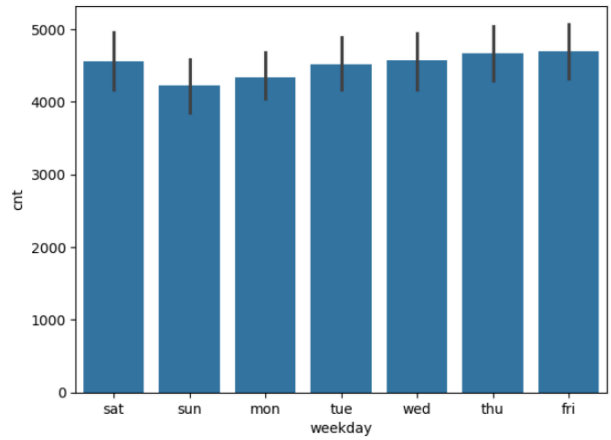
By doing a thorough analysis I was able to infer the following:

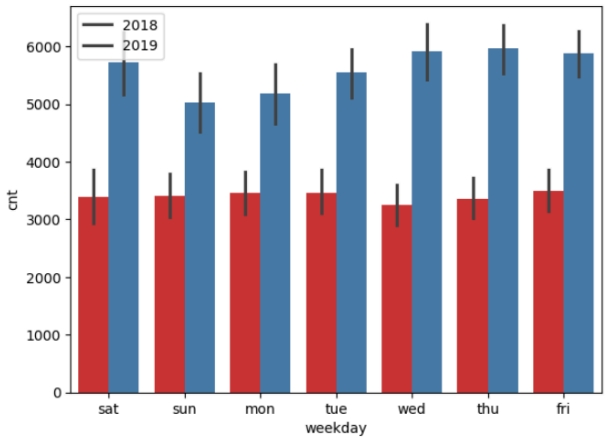
- The Fall Season brought the highest number of customers, i.e. the months of June, July, August and September.
- A clear weather resulted in more customers compared to misty or rainy.
- There were more renting as it got closer to the weekend and reduced on Sunday suggesting that's when they returned their vehicle.
- On non-holidays, there are more customers, suggesting that most users use the vehicle for work commutes.
- Booking seemed the same whether it was a non-working or working day.

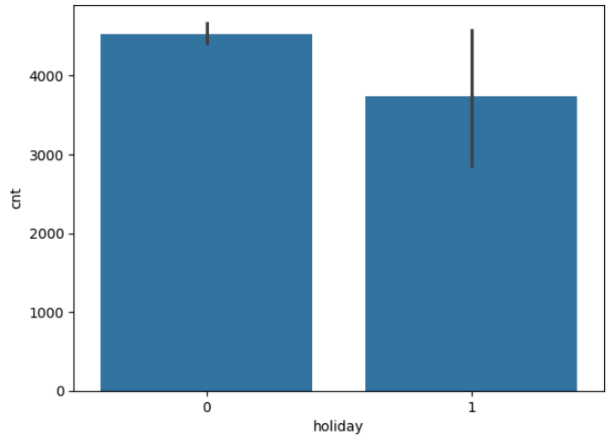


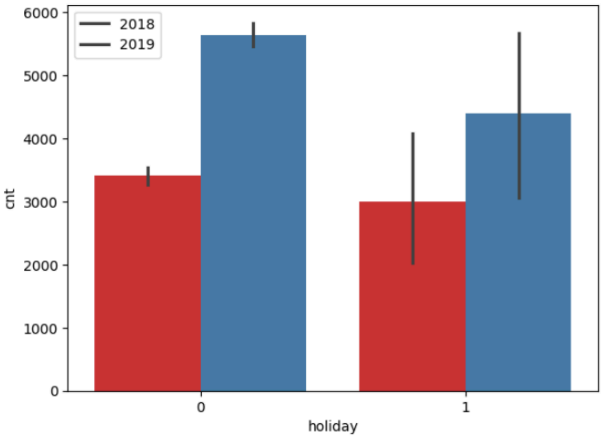












2. Why is it important to use drop_first=True during dummy variable creation?

Using drop_first=True during dummy variable creation is important to avoid multicollinearity, which can occur when you have redundant dummy variables.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
We can validate the assumption of the Linear Regression Model based on the following:

- Linearity: The relationship between the predictors and the response is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: The variance of the residuals is constant across all levels of the independent variables.
- Normality of Residuals: The residuals of the model are normally distributed.
- No Multicollinearity: Predictors are not highly correlated with each other.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
The top 3 features were year, temp and season_winter.

1. Explain the linear regression algorithm in detail.
Linear regression is a statistical method used to model the relationship between a dependent variable (also known as the response or target variable) and one or more independent variables (also known as predictors or features). The goal of linear regression is to find the best-fitting linear relationship between the dependent variable and the independent variables.

Mathematically the relationship can be represented with the help of the following equation − Y = mX + c Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

The following are some assumptions about dataset that are made by the Linear Regression model −
Multi-collinearity – The linear regression model assumes that there is very little or no multi-collinearity in the data. Multi-collinearity occurs when the independent variables or features have a dependency on them.
Auto-correlation – Another assumption the Linear regression model assumes is that there is very little or no auto-correlation in the data. Auto-correlation occurs when there is a dependency between residual errors.
Relationship between variables – The linear regression model assumes that the relationship between response and feature variables must be linear.
Normality of error terms – Error terms should be normally distributed
Homoscedasticity – There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

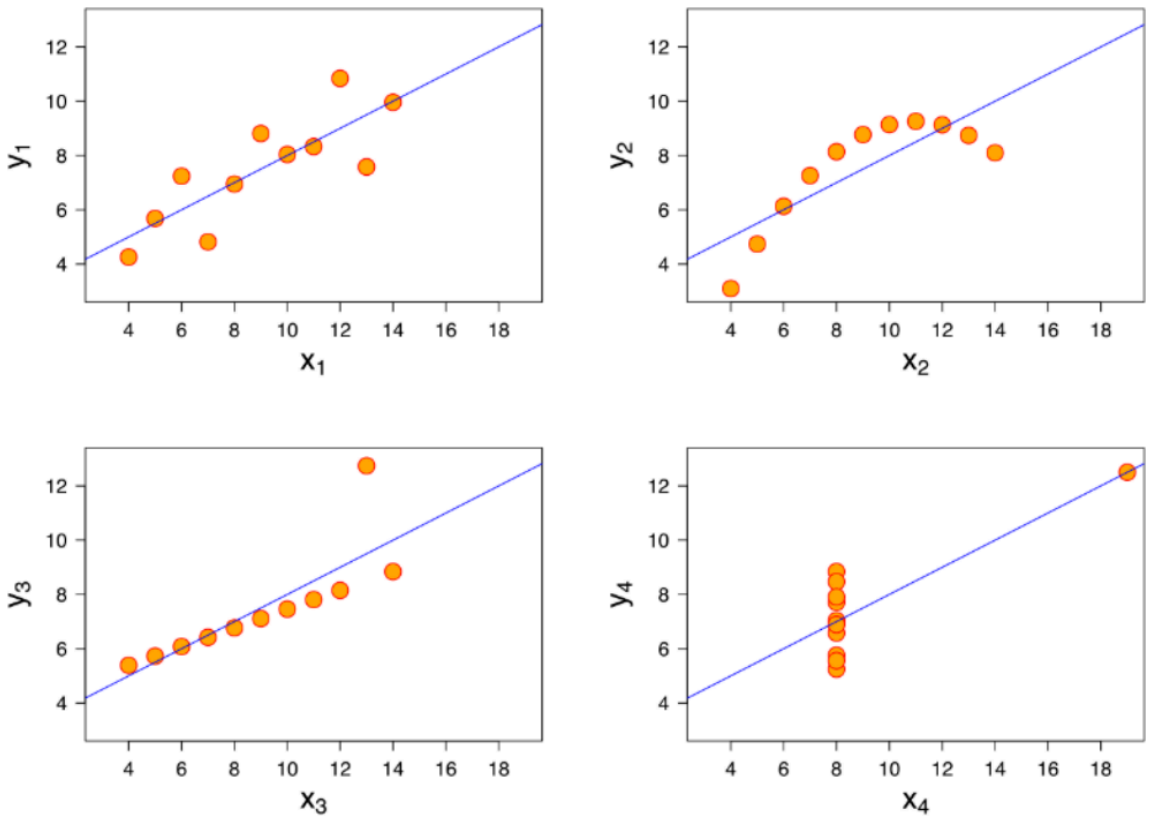
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

• Mean of x is 9 and mean of y is 7.50 for each dataset.

• Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

• The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



• Dataset I appears to have clean and well-fitting linear models.
• Dataset II is not distributed normally.
• In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
• Dataset IV shows that one outlier is enough to produce a high correlation coefficient. This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R?

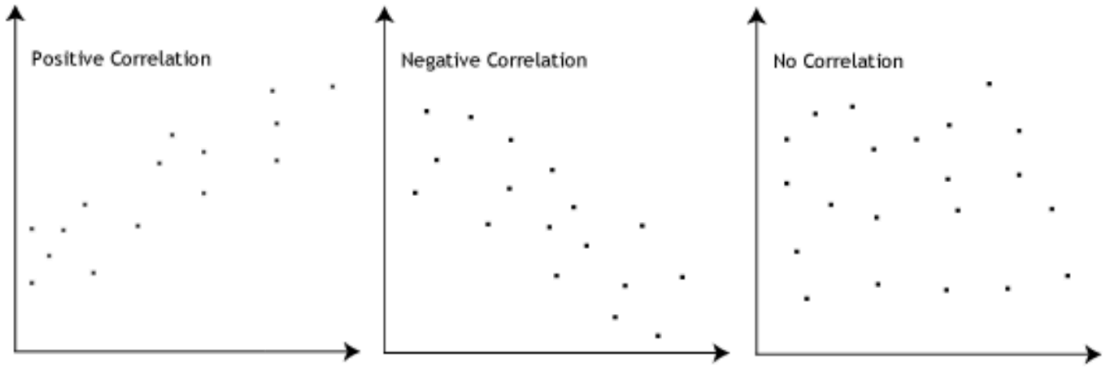
Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It quantifies the degree to which a linear relationship exists between two variables and ranges from -1 to 1.

The Pearson correlation coefficient is defined as:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

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. This is shown in the

diagram below:



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

Scaling is the process of adjusting the range of features or data to a specific scale. This is done to ensure that different features contribute equally to the model's learning process.

Why is Scaling Performed?

1. **Improves Convergence in Gradient Descent**: Features with different scales can lead to poor convergence or very slow convergence in gradient descent optimization. Scaling helps in faster and more efficient convergence.
2. **Prevents Dominance by Large-Scale Features**: Features with larger scales can dominate the learning process, causing the model to be biased towards them. Scaling ensures that each feature contributes proportionately.
3. **Enhances Model Performance**: Many machine learning algorithms (like k-nearest neighbours, SVM, etc.) perform better when features are on the same scale.
4. **Facilitates Fair Comparison**: Scaling allows for a fair comparison of different features, ensuring that the model considers all features equally.

Normalized Scaling (Min-Max Scaling)

- **Definition**: Normalization rescales the values of a feature to a fixed range, usually [0, 1] or [-1, 1].
- **Formula**:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Usage: It's useful when you want the features to have a specific bounded range. It's especially useful when the data has no significant outliers.

Pros: Keeps all the values within a fixed range.

Cons: Sensitive to outliers because they can skew the min and max values.

Standardized Scaling (Standardization or Z-score Normalization)

- **Definition**: Standardization rescales the values of a feature to have a mean of 0 and a standard deviation of 1.
- **Formula**:

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

Usage: It's useful when the data follows a Gaussian (normal) distribution. It helps in standardizing the input feature distributions.

Pros: Less sensitive to outliers compared to normalization. Suitable for algorithms that assume data is normally distributed.

Cons: Does not bound the values within a specific range.

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/(1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.