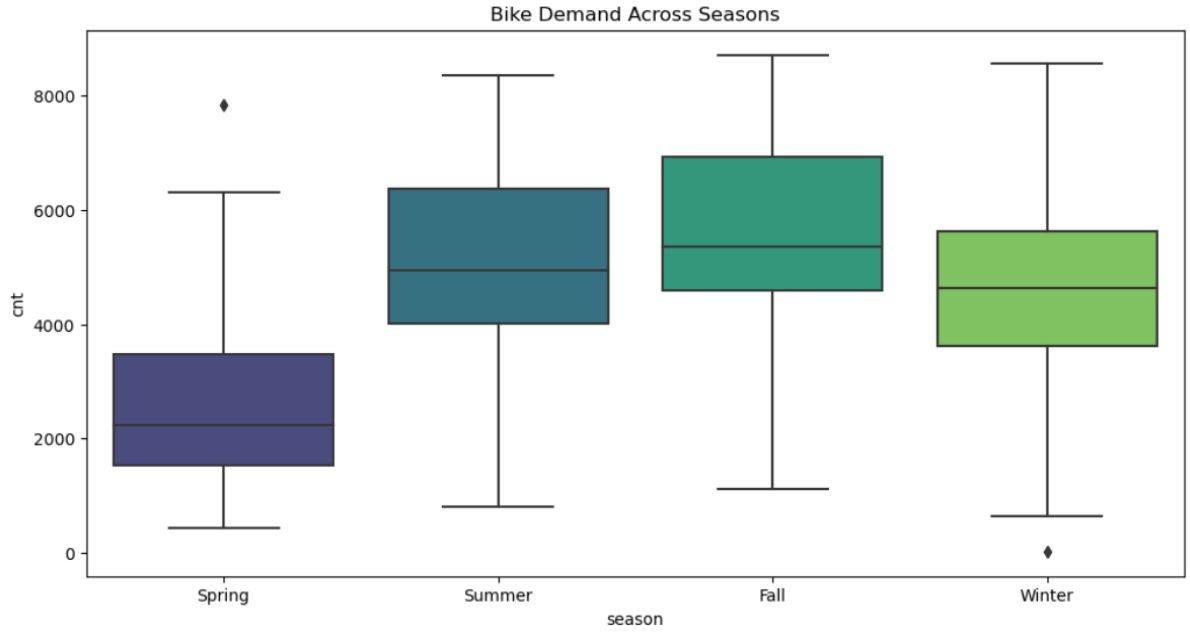
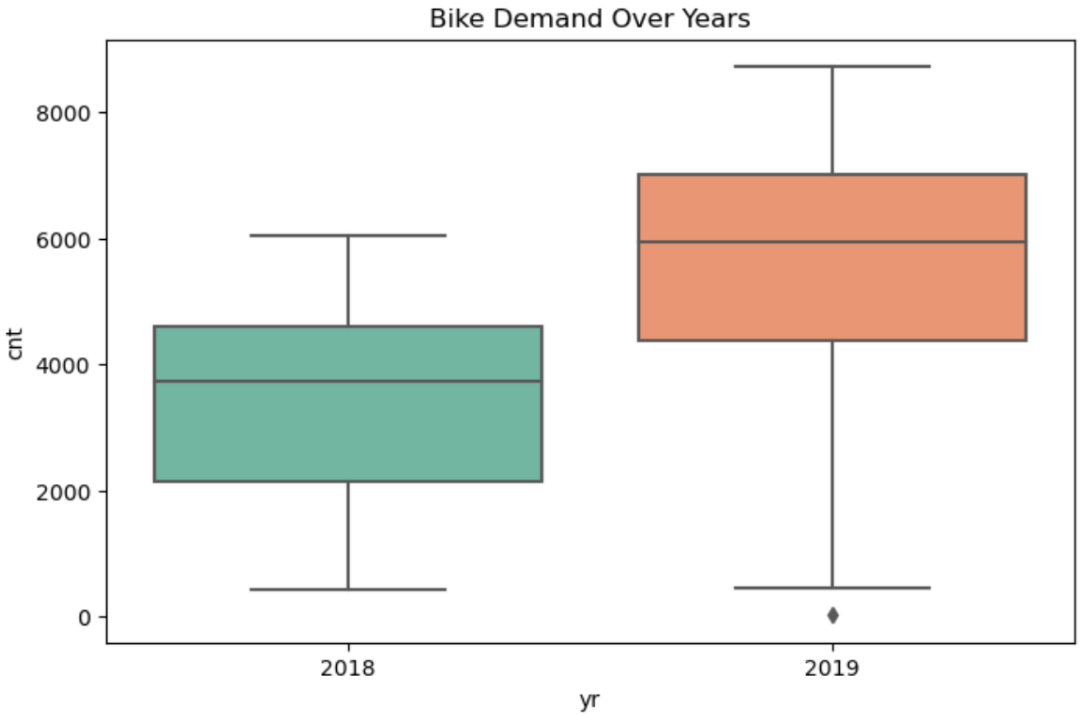
Assignment-based Subjective Questions

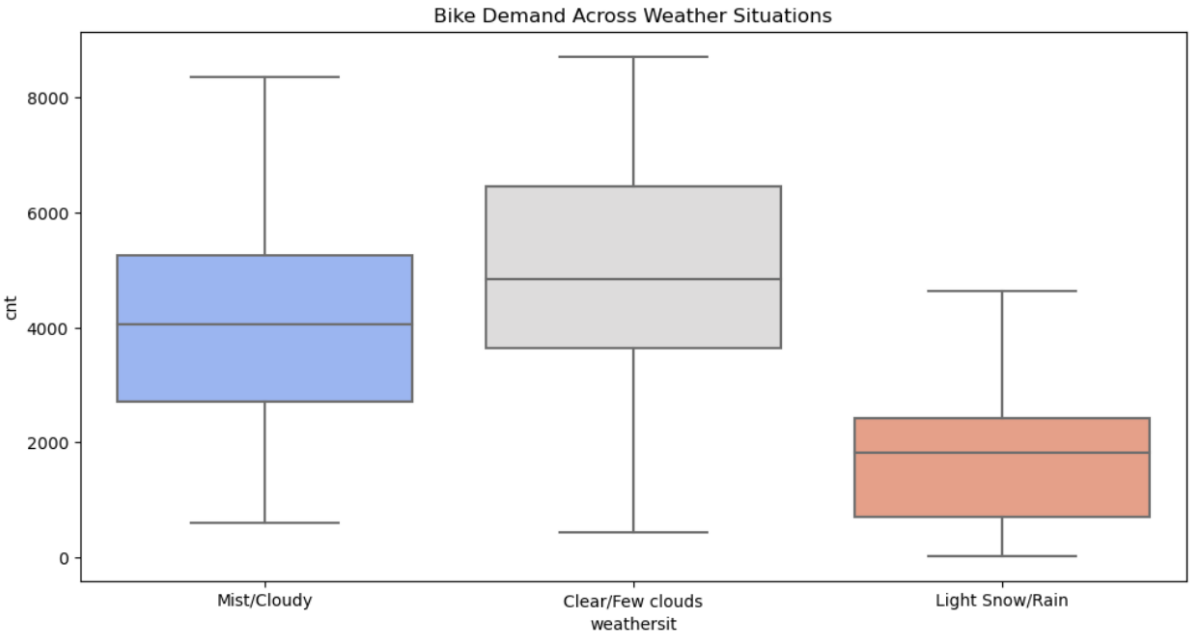
From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

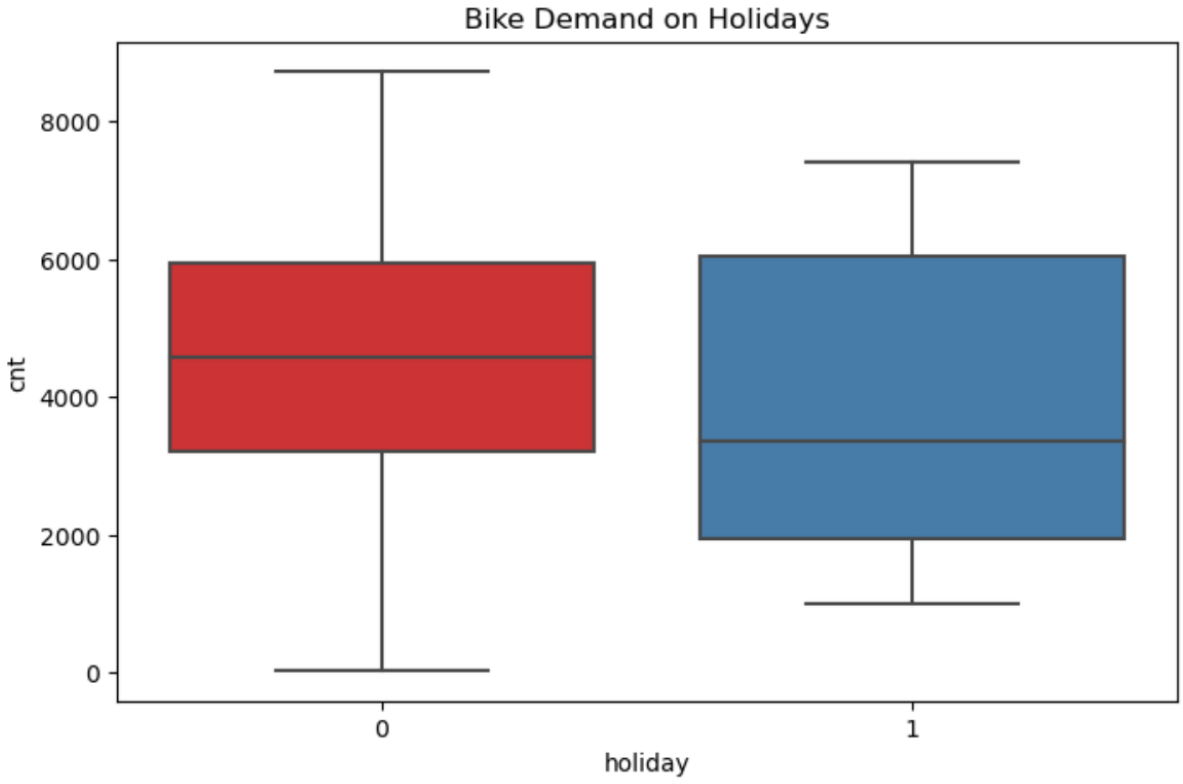
Some of the Categorical Variables have a strong effect on dependent variable 'cnt'.

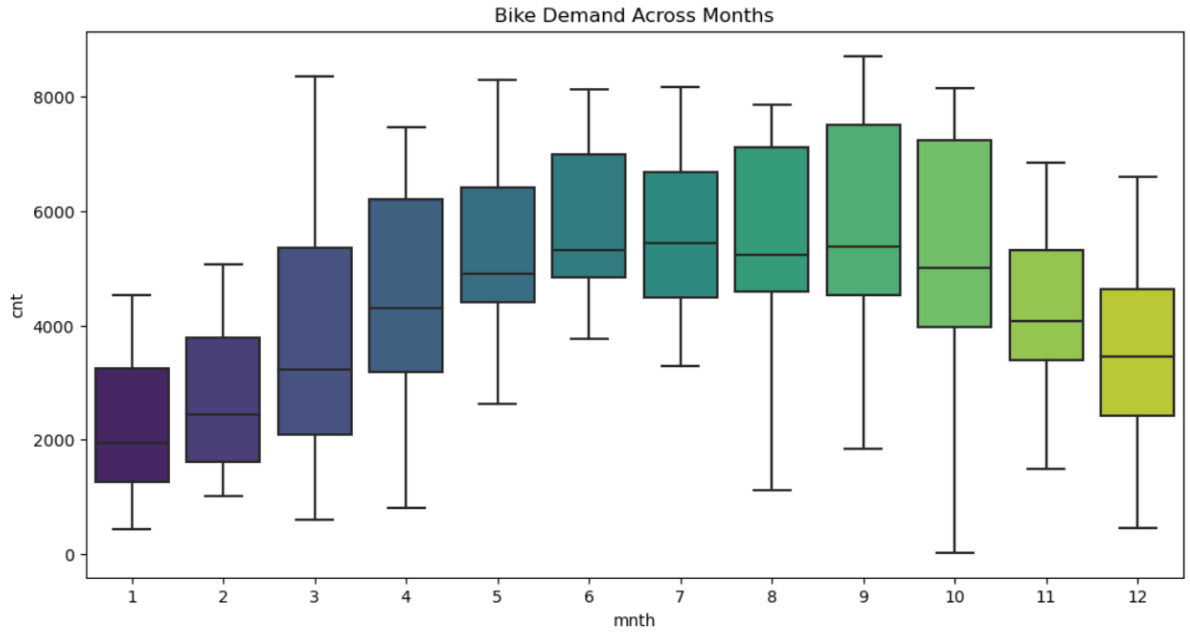
The Below Graphs show a clear Co-Relation of Seasons, Year, Weather Situation, Holiday, and Month.






Bike Demand Across Weather Situations
cnt
8000
6000
4000
2000
0
Mist/Cloudy
Clear/Few clouds
Light Snow/Rain
weathersit


Bike Demand on Holidays
cnt
8000
6000
4000
2000
0
0
1
holiday


Bike Demand Across Months
cnt
8000
6000
4000
2000
0
1
2
3
4
5
6
7
8
9
10
11
12
mnth

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

Parameter drop_first = True, this will drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level.

If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

Dropping 1 out of n variables helps with

Multicollinearity: Including all dummy variables in a regression model can introduce multicollinearity issues. Multicollinearity occurs when two or more independent variables are highly correlated, making it challenging for the model to distinguish their individual effects. In the context of dummy variables, if we include all levels, one level becomes a perfect predictor of the others.

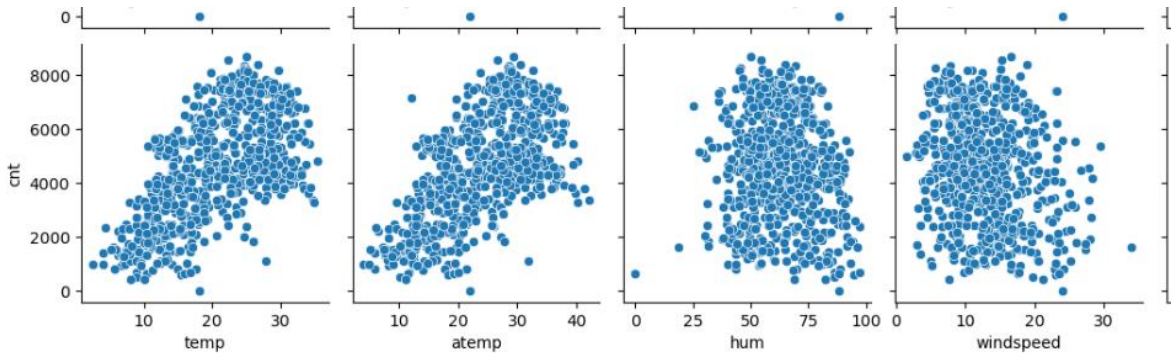
Interpretability: By setting drop_first=True, one level of the categorical variable is omitted, and the remaining levels are represented by the dummy variables. This helps avoid the dummy variable trap (perfect multicollinearity) and makes it easier to interpret the coefficients.

For a binary variable (e.g., gender with levels Male/Female), having one dummy variable (e.g., Male) is sufficient to capture the information. If Male = 0, it implies Female.

For a categorical variable with more than two levels (e.g., seasons - Spring, Summer, Fall, Winter), dropping one level (e.g., Spring) ensures that the coefficients for the remaining levels represent the change from the omitted level.

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

atemp has highest co-Relation with Target Variable. Temp also has same relation, but we drop it during Data-Engineering, as it is highly co-linear.



Correlation with target variable:

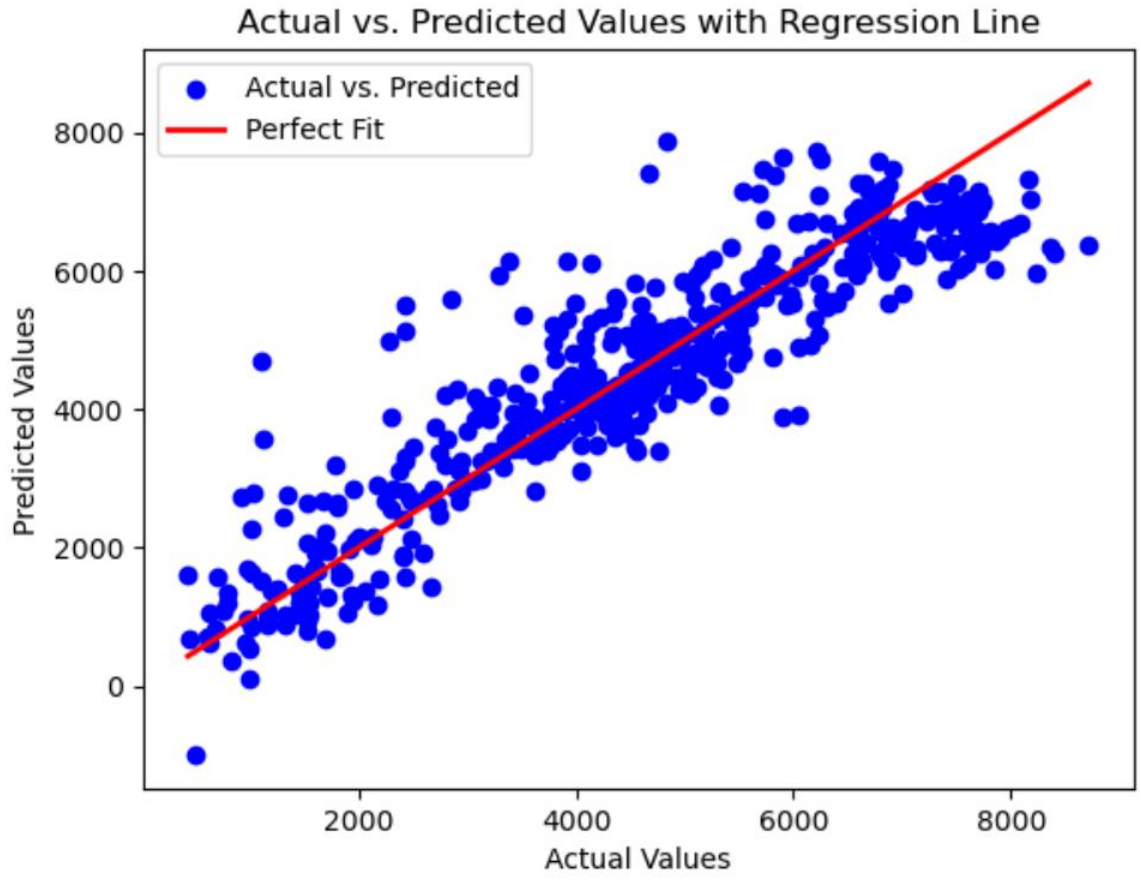
atemp	0.630685
yr	0.569728
mnth	0.278191
weekday	0.067534
workingday	0.062542
holiday	-0.068764
hum	-0.098543
windspeed	-0.235132

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

I checked for these aspects after Building the Model on Training DataSet.

 # Linearity: Check for linearity by plotting predicted values against actual values.

The points should be close to a diagonal line.



Done a **Recursive Feature Elimination** (RFE) to check if reducing any more features is possible, and if the model can be improved.

```
Baseline Model - MSE: 654483.89934685  Adjusted R-Square: 0.8124305592761027
Updated Model - MSE: 751868.8977358668  Adjusted R-Square: 0.7855668739185054
Improvement in MSE: -97384.99838901684
Improvement in Adjusted R-Square: -0.02686368535759731
```

This implied that no more feature needs to be dropped further.

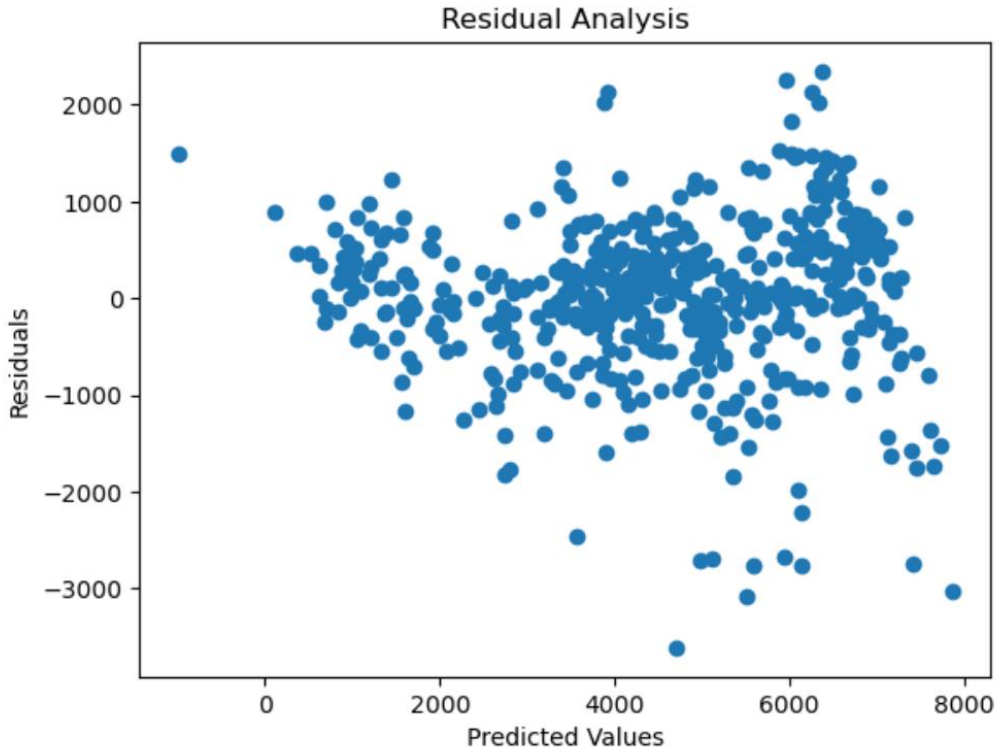
 # Standard Tests

R-Squared on testing set: 0.8236158928972526

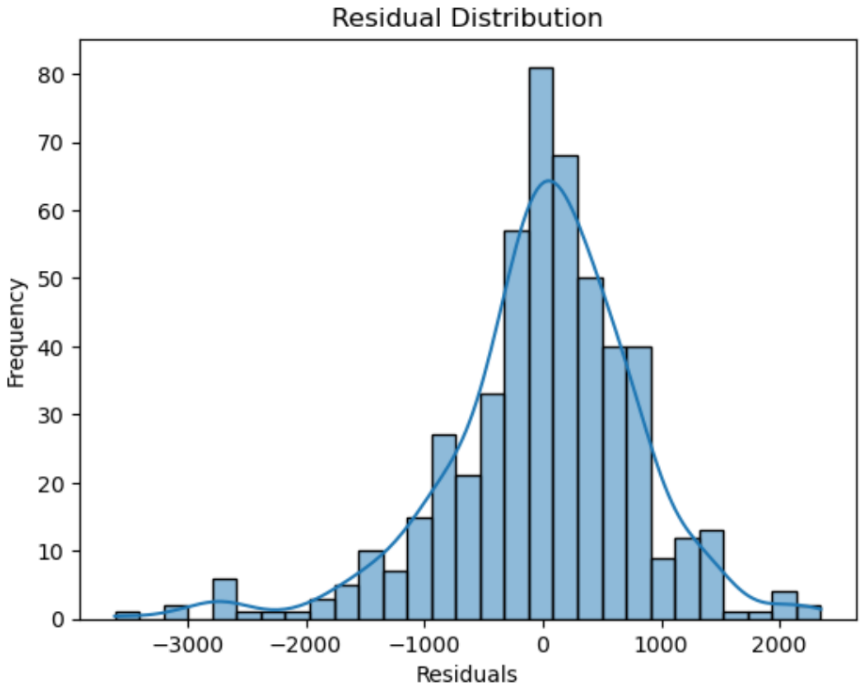
Adjusted R-Squared on testing set: 0.8124305592761027

Mean Squared Error on testing set: 654483.89934685

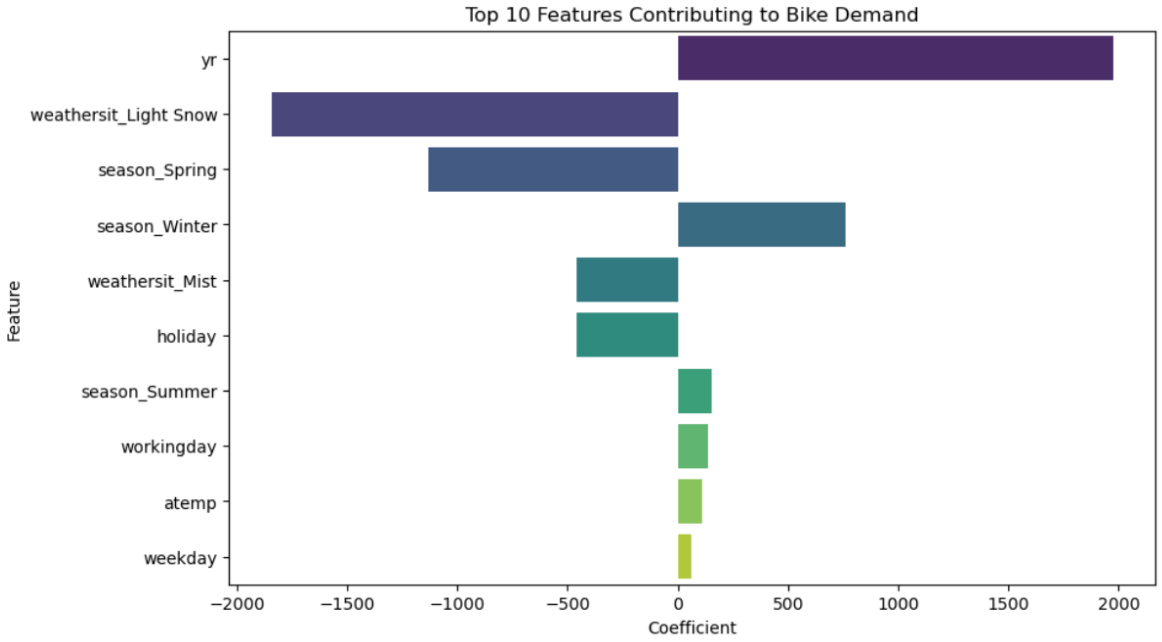
 # Residual Analysis: Plot the residuals (the differences between actual and predicted values) against predicted values. Check for a random pattern in the residuals, which indicates homoscedasticity. If there is a clear pattern or cone shape, it suggests heteroscedasticity.



 # Normality of Residuals: Plot a histogram of the residuals and check for a normal distribution. A normal distribution suggests that the model assumptions are met.



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

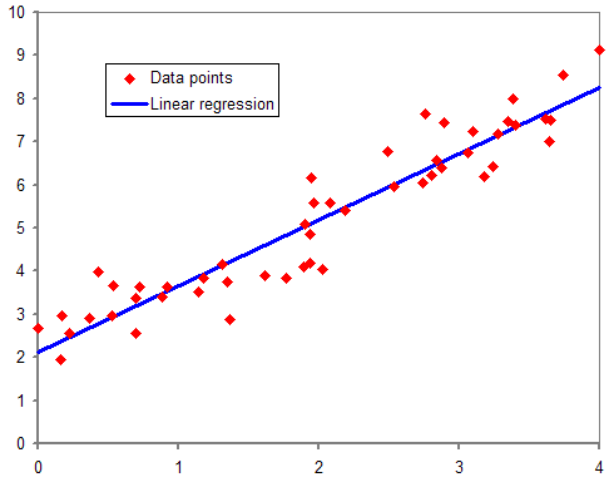


Top 10 Features:

Feature	Coefficient
yr	1976.382025
weathersit_Light Snow	-1844.126225
season_Spring	-1130.073083
season_Winter	761.982435
weathersit_Mist	-460.180797
holiday	-458.722554
season_Summer	151.925101
workingday	139.760049
atemp	112.084150
weekday	60.058733

General Subjective Questions

Explain the linear regression algorithm in detail.



Linear Regression is like drawing the best-fit straight line through a scatter plot of points. It helps us understand the relationship between two variables – one we want to predict (dependent variable) and another we use to make predictions (independent variable).

Imagine you have a bunch of data points (x, y), where 'x' is your input (independent variable) and 'y' is what you want to predict (dependent variable). Linear Regression tries to find the line (y = mx + b) that minimizes the distance between the predicted 'y' values on the line and the actual 'y' values in your data.

'm' (slope): It shows how much 'y' changes when 'x' changes. For example, if 'm' is 2, it means for every increase of 1 in 'x', 'y' increases by 2.

'b' (y-intercept): It's the point where the line intersects the 'y' axis. It gives the baseline value of 'y' when 'x' is 0.

Generally, for a single variable equation, it can be represented as y = mx + b.

The goal is to find the best 'm' and 'b' values so that our line is as close as possible to all the data points. This line helps us make predictions for new values of 'x'.

Example:

Let's say we're predicting a student's exam score ('y') based on the number of hours they studied ('x'). If we plot the data points and apply Linear Regression, the line will help us predict scores for any given study hours.

Equation: Score = 2 * Study Hours + 30

Here, '2' is the slope (meaning for every extra hour studied, the score goes up by 2), and '30' is the baseline score when the student didn't study at all.

Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear very different when graphed. It was created by the statistician Francis Anscombe to emphasize the importance of graphing data before analyzing it and to demonstrate the effect of outliers on statistical properties.

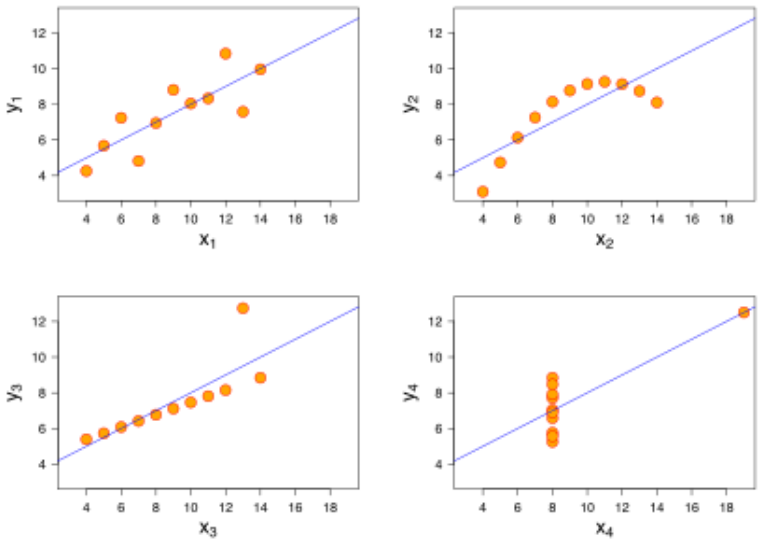
Let's break it down with an example:

Dataset I:

Simple linear relationship.

Suitable for simple linear regression.

A clear trend when plotted.



Dataset II:

Non-linear relationship.

Regression line is the same, but the distribution is different.

Emphasizes the importance of not relying only on regression statistics.

Dataset III:

Perfectly linear, except for one outlier.

The outlier significantly influences the regression line and correlation coefficient.

Dataset IV:

Perfectly horizontal line, except for one outlier.

The outlier has a significant impact on correlation and regression analysis.

Explanation:

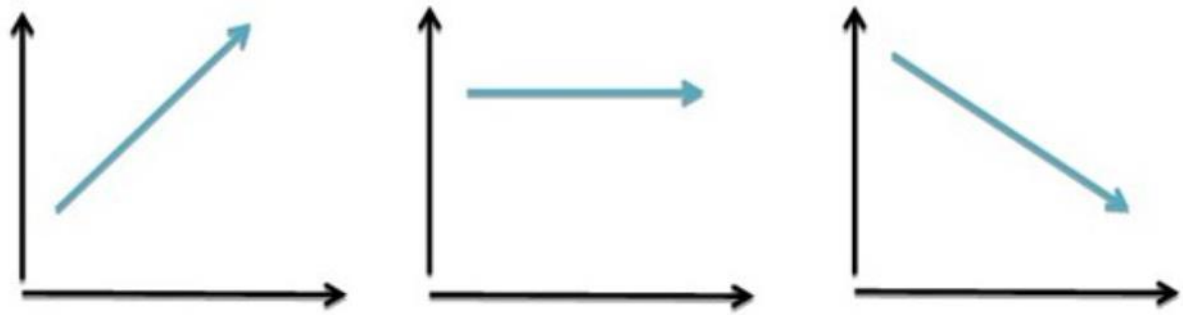
Anscombe's quartet underscores the limitations of summary statistics and the importance of visualization. Descriptive statistics like mean, variance, and correlation can be misleading, and relying solely on them without visualizing the data may lead to incorrect conclusions.

For example, Dataset I and Dataset II have the same mean, variance, and correlation coefficient, but their distributions are different. This highlights that numerical summary alone may not capture the essence of the data.

In summary, Anscombe's quartet emphasizes the need for graphical exploration of data to gain a comprehensive understanding, especially when working with statistical analyses like linear regression.

What is Pearson's R?

Pearson's correlation coefficient, denoted as r, is a measure of the linear relationship between two variables. It quantifies the strength and direction of a linear association between two continuous variables.



The formula for Pearson's correlation coefficient between variables

X and Y is given by:

The formula for Pearson's correlation coefficient between variables X and Y is given by:

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

Here:

- X_i and Y_i are individual data points.
- $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y, respectively.

Perfect Positive Correlation (r = 1): If the correlation is 1, it means that when one set of numbers goes up, the other always goes up too. It's like a perfect friendship – if one friend is happy, the other is always happy too.

No Correlation (r = 0): If the correlation is 0, it means there's no particular pattern. One set of numbers doesn't depend on the other. It's like having a friend, but whether you're happy or sad, it doesn't affect your friend.

Perfect Negative Correlation (r = -1): On the other hand, if the correlation is -1, it means that when one set of numbers goes up, the other always goes down. It's like a seesaw – when one side goes up, the other goes down.

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

Scaling makes different variables comparable or putting them on the same scale. It helps for these two cases.

Equality: It helps to treat different features (like height and weight) equally. If you don't scale them, one might have a larger impact just because its values are naturally bigger.

Algorithm's Preference: Many algorithms, like linear regression, work better when all features are on a similar scale. It's like ensuring everyone speaks the same language.

Generally, there are two types of scaling.

Normalized Scaling (MinMax Scaling):

Idea: Imagine your height and weight values are percentages. Normalization makes sure they're all between 0% and 100%.

Formula: (value − min value) / (max value − min value)

Result: All values will be between 0 and 1.

Standardized Scaling (Z-score Scaling):

Idea: Imagine your height and weight are scores. Standardization makes them look like how far they are from the average.

Formula: (value − average value) / (standard deviation)

Result: Most values will be around 0, and you'll know how many standard deviations each value is from the average.

In simple terms, normalization is like putting everything in percentages, while standardization is like talking in terms of average and standard deviations. Both make your data easier for the algorithm to understand.

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

Variance Inflation Factor (VIF): In the world of linear regression, VIF helps us understand how much the variance (or uncertainty) of an estimated regression coefficient increases if our predictors (independent variables) are correlated.

Infinite VIF Occurrence: When the VIF is infinite for a particular predictor, it's usually a sign of perfect multicollinearity. This happens when one variable can be perfectly predicted from the others.

Let's say you have two variables: "Number of Wheels" and "Number of Bicycles."

If you know the number of wheels, you can perfectly predict the number of bicycles, because each bicycle has a fixed number of wheels.

Consequence for VIF:

VIF involves calculating how much the variance of a regression coefficient increases when your variable is added to a model containing other variables.

If a variable can be perfectly predicted from others, it means its variance is essentially zero.

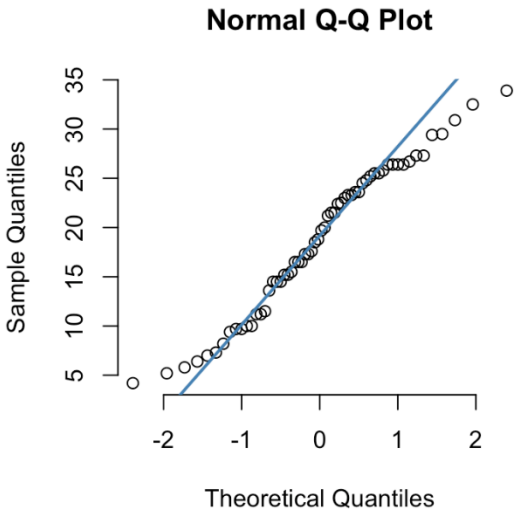
Dividing by nearly zero results in a very large number, and practically, we say the VIF is infinite.

When a variable can be perfectly predicted from other variables, it's like having redundant information, and VIF goes wild because mathematically it can't handle dividing by almost zero. This situation is rare in real-world scenarios but is crucial to be aware of when interpreting VIF values.

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

Q-Q Plot Explained:

A Quantile-Quantile (Q-Q) plot is like a visual aid that helps us check if a dataset follows a particular theoretical distribution. In the context of linear regression:



Use of Q-Q Plot in Linear Regression:

Normality Check: In linear regression, we often assume that the residuals (the differences between predicted and actual values) follow a normal distribution. Q-Q plots help us visually assess if this assumption holds.

How It Works:

- The plot compares the quantiles of our residuals to the quantiles of a theoretical normal distribution.
- If the points on the Q-Q plot roughly form a straight line, it suggests our residuals are close to normally distributed.

Importance:

- **Normal Residuals Assumption:**
  - Many statistical techniques rely on the assumption that residuals are normally distributed.
  - If this assumption holds, it implies that our statistical inferences (like confidence intervals and hypothesis tests) are more reliable.
- **Detecting Outliers:**
  - Outliers can be seen as points deviating from the expected straight line on the Q-Q plot.
  - Identifying outliers is crucial as they might indicate data points affecting the model's performance.

Example:

Imagine you have a set of residuals. You create a Q-Q plot, and if the points mostly align in a straight line, it's like saying, "Hey, our residuals are behaving like we expect in a perfect world where everything is normal." It's a handy tool to visually check assumptions and the health of your regression model.