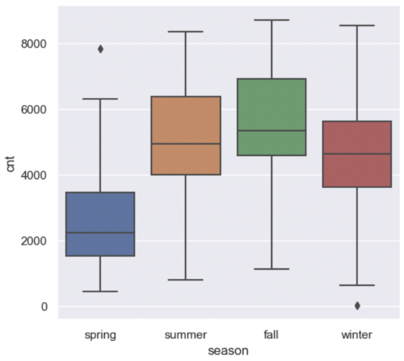
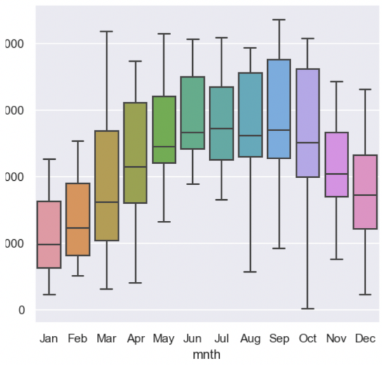
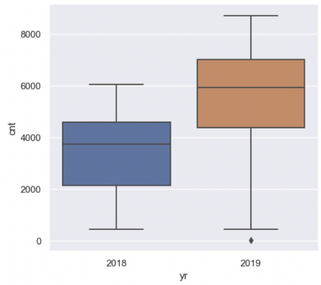
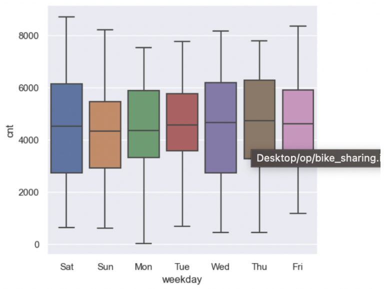
# Assignment-based Subjective Questions

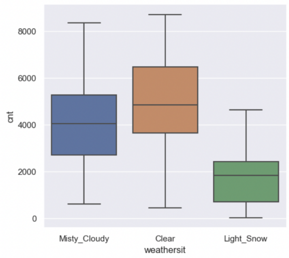
## Q1 From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

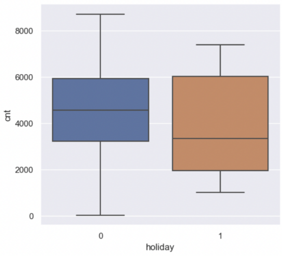












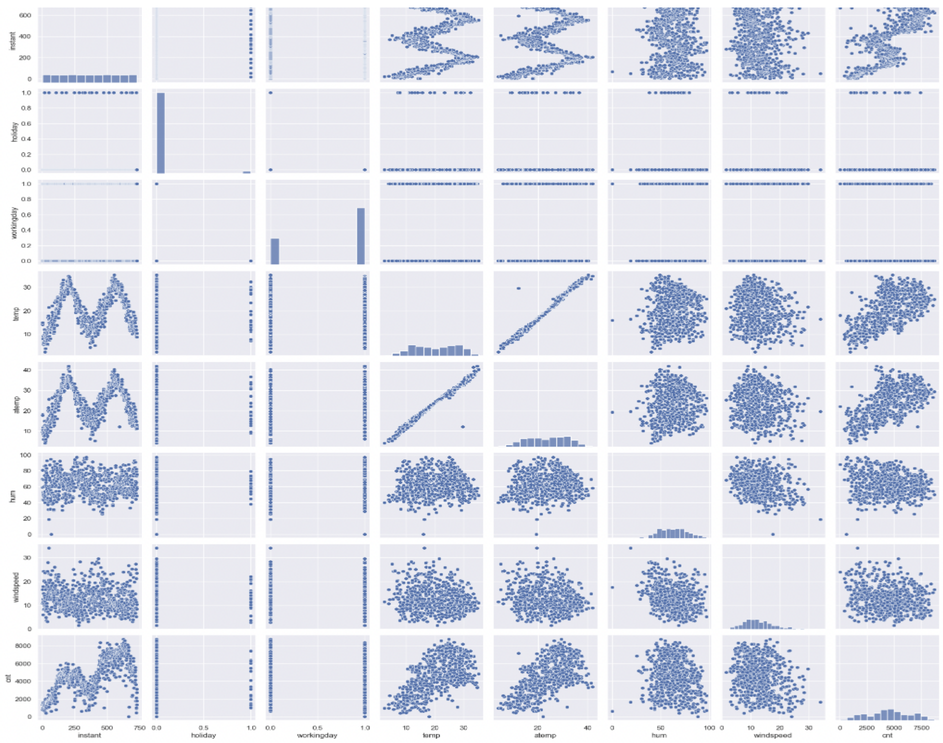
Here are some of the categorical variables from the provided day.csv dataset
The categorical variables, including season, month, year, weekday, working day, and weathersit, exhibit a significant influence on the dependent variable 'cnt.' The figure below visually represents the relationships or associations among these categorical variables and their impact on 'cnt.'

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

The purpose of creating dummy variables for a categorical variable with 'n' levels is to generate 'n-1' new columns, each indicating the presence or absence of a specific level using binary values (0 or 1). By setting drop_first=True, we ensure that the resulting set of dummy variables corresponds to 'n-1' levels, effectively dropping one level to avoid multicollinearity or high correlation among the dummy variables.

For instance, if there are 3 levels, enabling drop_first will omit the first column among the dummy variables, helping to reduce correlation issues

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



Among all the variables considered in relation to the target variable 'cnt,' the variables 'temp' and 'atemp' exhibit the strongest correlation.

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

validating a Linear Regression model involves checking if the relationship between variables is linear, if there's no pattern in the errors, if the residuals follow a normal distribution, if the spread of residuals is consistent, and if there's no problematic correlation between predictor variables. These criteria ensure the model's reliability and appropriateness for making predictions

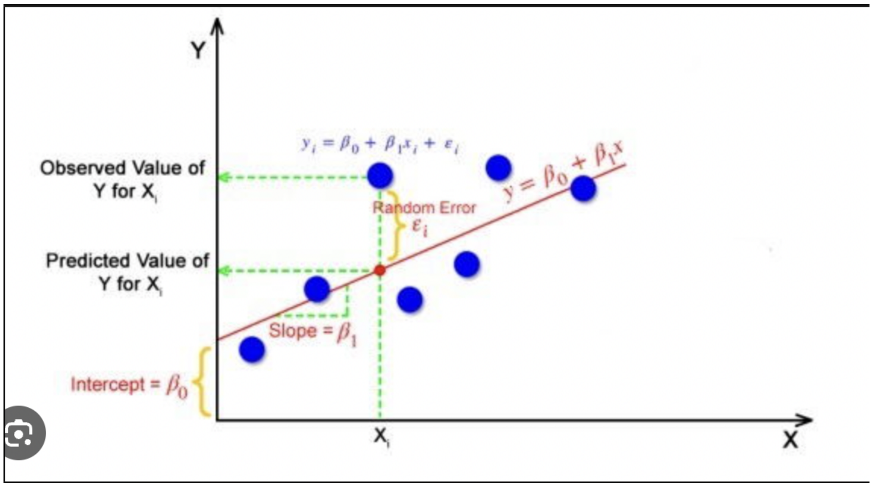
And particularly based on Linearity,No auto-correlation,Normality of error,Homoscedasticity, and Multicollinearity.

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

The three most influential factors in explaining the shared bike demand are temperature, year, and season. These variables play a significant role in understanding and predicting the demand for shared bikes.

## General Subjective Questions

**Q1 Explain the linear regression algorithm in detail.**

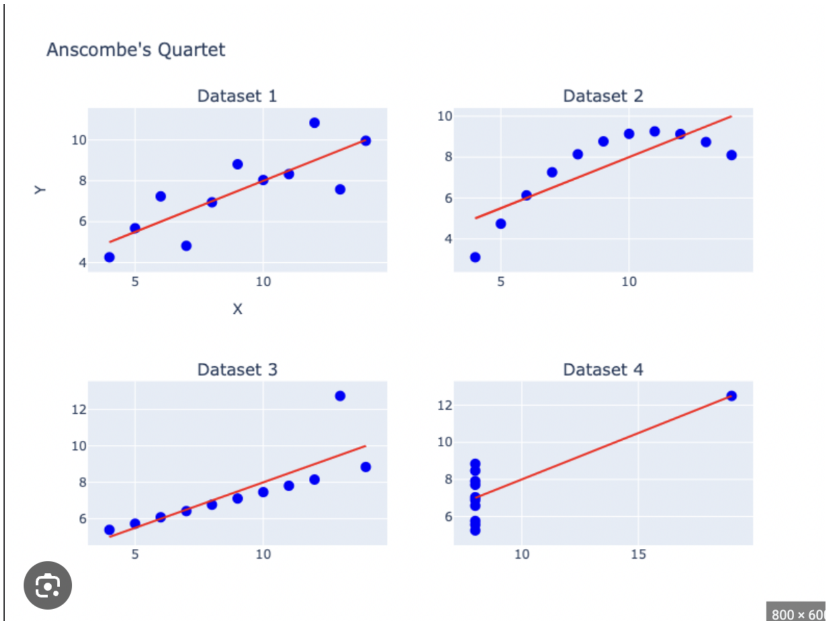

Y
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Observed Value of Y for $X_i$
Random Error $\varepsilon_i$
$y = \beta_0 + \beta_1 x$
Predicted Value of Y for $X_i$
Slope = $\beta_1$
Intercept = $\beta_0$
$X_i$
X

Linear regression is a predictive modeling method that elucidates the connection between a dependent variable (also known as the target variable) and independent variables (known as predictors). It focuses on identifying a linear relationship, revealing how changes in the dependent variable correspond to shifts in the independent variable's values. When there's only one input variable (x), this is referred to as simple linear regression. If multiple input variables are involved, it becomes multiple linear regression. The outcome of a linear regression model is a straight line that characterizes the association between the variables.

This regression line can manifest as either a Positive Linear Relationship or a Negative Linear Relationship. The primary objective of the linear regression algorithm is to determine the most appropriate values for a0 and a1, thereby establishing the best-fitting line that minimizes errors.

In Linear Regression, methods like Recursive Feature Elimination (RFE), Mean Squared Error (MSE), or cost functions are employed to identify the optimal values for a0 and a1, enabling the identification of the best-fit line for the given data points. These techniques aim to minimize errors and create an accurate model.

**Q2 Explain the Anscombe’s quartet in detail?**



Anscombe's Quartet comprises four distinct datasets that share similar basic statistical properties, such as means and variances of both the X and Y variables. However, these datasets exhibit unique characteristics that can potentially mislead a regression model if one were to blindly apply it. When visualized on scatter plots, they reveal starkly different patterns and distributions.

The main purpose behind creating Anscombe's Quartet was to underscore the critical importance of plotting data and examining visualizations before diving into statistical analysis or model building. It serves as a powerful reminder of how the presence of outliers, non-linear relationships, and influential data points can significantly affect the statistical properties and suitability of regression models.

To elaborate on the four datasets:

1. The first dataset appears to follow a linear relationship between X and Y, making it a suitable candidate for a linear regression model.

2. In contrast, the second dataset lacks a discernible linear relationship between X and Y, suggesting that linear regression is not an appropriate choice for modeling this data.

3. The third dataset displays outliers, which can pose challenges for linear regression models as they can unduly influence the model's predictions and assumptions.

4. The fourth dataset contains a high leverage point, indicating that it has a significant impact on the correlation coefficient. This underscores how influential individual data points can be in linear regression.

In essence, Anscombe's Quartet serves as a powerful cautionary tale, highlighting that regression algorithms can be misled by superficially similar statistical properties. Hence, it underscores the importance of thorough data visualization and exploration before embarking on machine learning model development.

## Q3 . What is Pearson’s R?

Pearson's r is a numerical value that ranges from -1 to 1, offering insights into the strength and direction of the relationship between two variables. When Pearson's r is close to 1, it signifies a strong positive linear correlation, indicating that as one variable increases, the other tends to increase as well. Conversely, when it is close to -1, it suggests a strong negative linear correlation, signifying that as one variable increases, the other tends to decrease. A value close to 0 indicates a weak or no linear relationship, implying that changes in one variable do not correspond to predictable changes in the other.

This correlation coefficient is a valuable tool in various fields, including economics, psychology, biology, and many others. Researchers and analysts use it to explore connections between variables, making it essential for hypothesis testing, model building, and drawing insights from data. It's important to note that while Pearson's r captures linear relationships well, it may not detect non-linear associations, which may require other correlation measures or data analysis techniques for proper examination.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

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

Scaling involves the transformation of data to fit it within a specific range or scale, and it plays a crucial role in data preprocessing. This process is essential for ensuring that data used in algorithms are uniform in scale and can lead to more efficient computations. When we collect data, it often includes features with varying magnitudes, units, and ranges. If we don't perform scaling, algorithms may assign undue importance to high-magnitude values while neglecting other parameters, potentially leading to inaccurate modeling.

There are two primary methods of scaling data: Normalizing Scaling and Standardize Scaling, each with distinct characteristics:

1. In Normalized Scaling, we use the minimum and maximum values of the features to scale them. On the other hand, Standardized Scaling employs the mean and standard deviation for scaling.

2. Normalized Scaling is ideal when dealing with features that have different scales, while Standardized Scaling is used to achieve a standard mean of zero and a unit standard deviation.

3. Normalized Scaling confines values within the range of (0,1) or (-1,1), whereas Standardized Scaling doesn't impose such boundaries.

4. Normalized Scaling can be influenced by outliers in the data, while Standardized Scaling is less affected by them.

5. Normalized Scaling is suitable when the data distribution is unknown, whereas Standardized Scaling is preferable when the distribution is approximately normal.

6. Normalized Scaling is often referred to as Scaling Normalization, while Standardized Scaling is commonly known as Z-Score Normalization.

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

The Variance Inflation Factor (VIF) serves as a valuable metric for understanding how one independent variable relates to all the other independent variables in a dataset. The formula for calculating VIF is provided below:

When assessing VIF values, a value exceeding 10 is considered notably high, and values greater than 5 should also raise concerns and prompt further investigation. A significantly elevated VIF indicates a strong correlation between two independent variables. In cases of perfect correlation, the R-squared value ($R^2$) equals 1, which leads to a VIF calculation of $1/(1-R^2)$ that approaches infinity. To address this issue of perfect multicollinearity, it becomes necessary to remove one of the variables from the dataset, as they are causing this extreme level of correlation.

$$VIF_i = \frac{1}{1-R_i^2}$$

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

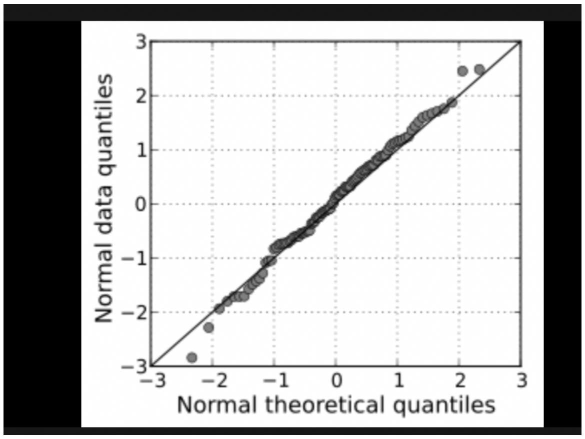
The Q-Q plot, short for Quantile-Quantile plot, serves as a graphical technique for comparing two probability distributions by juxtaposing their quantiles. It aids in assessing whether a given dataset could plausibly originate from a particular theoretical distribution, such as the Normal, exponential, or Uniform distribution.

Furthermore, the Q-Q plot offers a means to evaluate the similarity between two distributions. When the distributions are closely aligned, the Q-Q plot tends to display a more linear pattern. To rigorously test the linearity assumption, scatter plots are often employed. Additionally, in linear regression analysis, the multivariate normality assumption for all variables can be effectively examined using either a histogram or a Q-Q plot.

The significance of the Q-Q plot in linear regression lies in its ability to confirm whether both the training and test datasets adhere to the same population distribution. This validation ensures that the data samples from both sets share a common underlying distribution.

Advantages of the Q-Q plot include its applicability to datasets of varying sizes and its capacity to detect various distributional characteristics such as shifts in location, scale, symmetry changes, and the presence of outliers. The Q-Q plot is employed in the examination of two datasets to determine if they stem from a population with a common distribution, share a common location

Example


Normal data quantiles
Normal theoretical quantiles