**Subjective Question**

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

- Temp, Year(current year rental is better than previous year) and months of july and september

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

- By compairing the values of adjusted r2 and r2 of prediction on test and train set.

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

- By analysing the pair plots of numerical variable , we can see that there are some independent variables which look positively correlated to the 'count' variable. Eg: Bike rentals are more correlated to temperature.

**4. Why is it important to use drop\_first=True during dummy variable creation?**

- It helps in reducing the extra column created during dummy variable creation and hence reduces the correlations created among dummy variables.

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

- The plots above shows the relationship between categorical variables and a Target variable.

Bike Rentals are more during the Fall season and then in summer

Bike Rentals are more in the year 2019 compared to 2018

Bike Rentals are more in partly cloudy weather

Bike Rentals are more on Saturday,wednesday and thursday.

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

Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value.

Linear regression is defined as an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events.

The following is an example of a resulting linear regression equation:

Linear regression equation

**Y** = dependent variable

**x1, x2,** = explanatory variables.

The coefficients (b1, b2, and so on) explain the correlation of the explanatory variables with the dependent variable.

The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated.

b0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

A diagram of a graph

Description automatically generated

**The Different Types of Linear Regression**

There are many different types of linear regression, but the most common is Ordinary Least Squares (OLS) regression. OLS regression is a statistical method that estimates the relationships between variables by minimizing the squared errors of the predictions.

Other types of linear regression include:  
-Ridge regression: This type of regression adds a penalty for large coefficients, which prevents overfitting and can improve the accuracy of the model.

-Lasso regression: This type of regression also adds a penalty for large coefficients, but it is more aggressive than ridge regression and can lead to more sparse models (models with fewer features).

-Elastic net: This type of regression combines both ridge and lasso penalties, providing a balance between them.

-Bayesian linear regression: This type of regression uses Bayesian inference to estimate the model parameters.

1. **Explain the Anscombe’s quartet in detail.**

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

**Anscombe's quartet** comprises four [data sets](https://en.wikipedia.org/wiki/Data_set) that have nearly identical simple [descriptive statistics](https://en.wikipedia.org/wiki/Descriptive_statistics), yet have very different [distributions](https://en.wikipedia.org/wiki/Probability_distribution) and appear very different when [graphed](https://en.wikipedia.org/wiki/Plot_(graphics)).

Snippet of Anscombe Dataset

A table of numbers and letters

Description automatically generated with medium confidence

Statistical summary of Anscombe data :

A screenshot of a math form

Description automatically generated

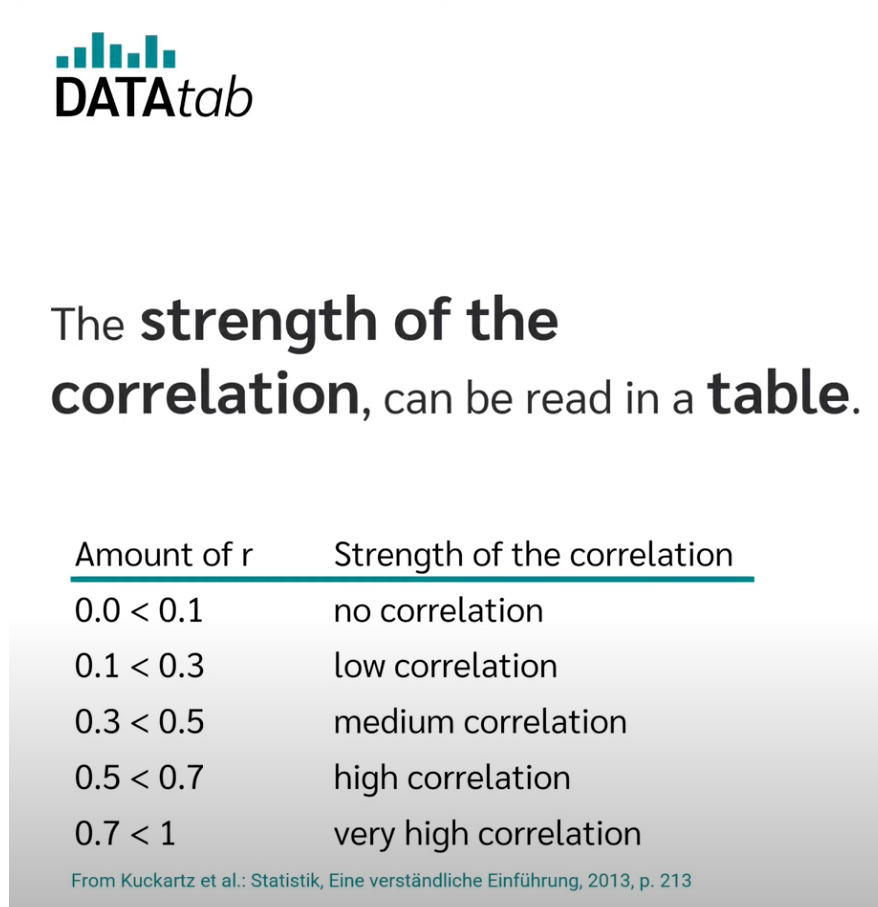
1. A graph of a graph of a function

   Description automatically generated with medium confidence**What is Pearson’s R?**

Pearson’s Correlation Coefficient, often denoted as r, measures the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

A diagram of mathematical equations

Description automatically generated



**Assumptions for a Pearson Correlation**

1. Data must be from random or representative samples for meaningful statistical inferences.
2. Both variables should be continuous and follow a normal distribution.
3. Homoscedasticity is crucial, ensuring similar variance around the line of best fit.
4. Extreme outliers, whether univariate or multivariate, impact the Pearson Correlation Coefficient. For instance, plotting age vs. loan amount reveals a correlation, but reversing the variables might yield inconclusive results. Consider these assumptions when interpreting correlations for robust analysis.
5. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

***Normalization/Min-Max Scaling:***

It brings all of the data in the range of 0 and 1.

1. ***sklearn.preprocessing.MinMaxScaler****helps to implement normalization in python.*

***Standardization Scaling:***

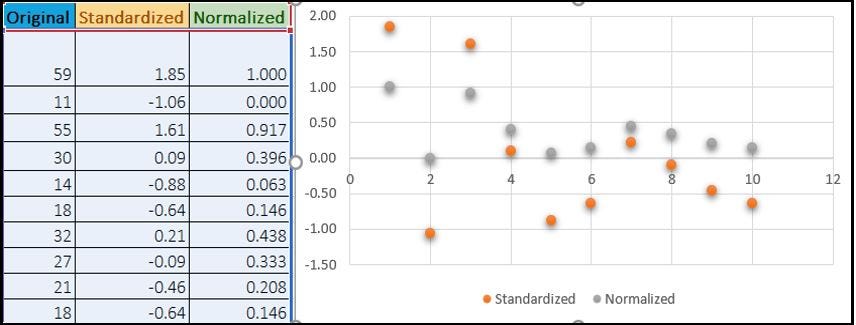
* Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

A black text on a white background

Description automatically generated

* **sklearn. preprocessing. scale** helps to implement standardization in python.

Below shows example of Standardized and Normalized scaling on original values.

* 

*One disadvantage of normalization over standardization is that it****loses****some information in the data, especially about****outliers****.*

Below shows example of Standardized and Normalized scaling on original values.

A screenshot of a computer screen

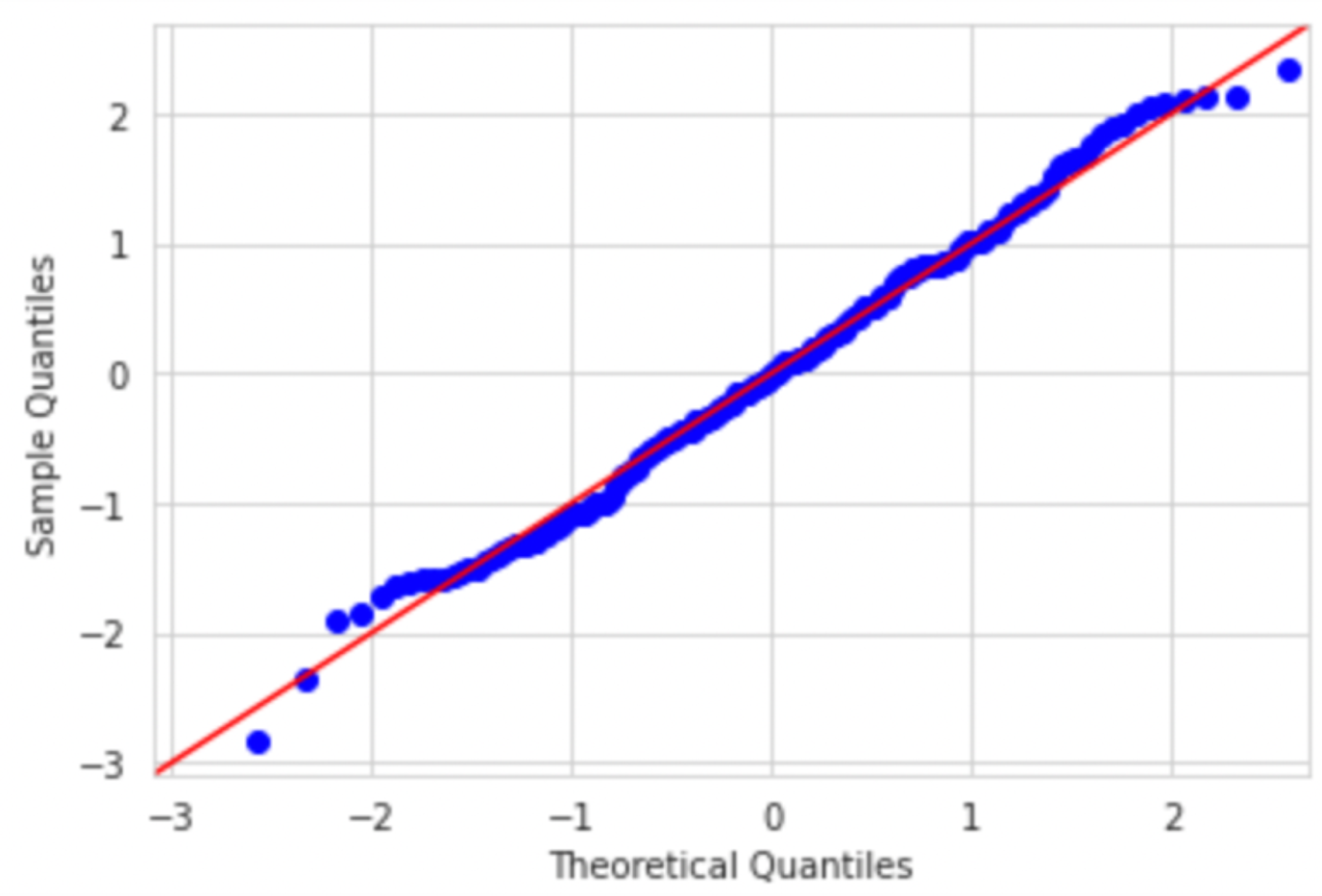
Description automatically generated

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

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line.



The power of Q-Q plots lies in their ability to summarize any distribution visually.

QQ plots is very useful to determine

* If two populations are of the same distribution
* If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
* Skewness of distribution