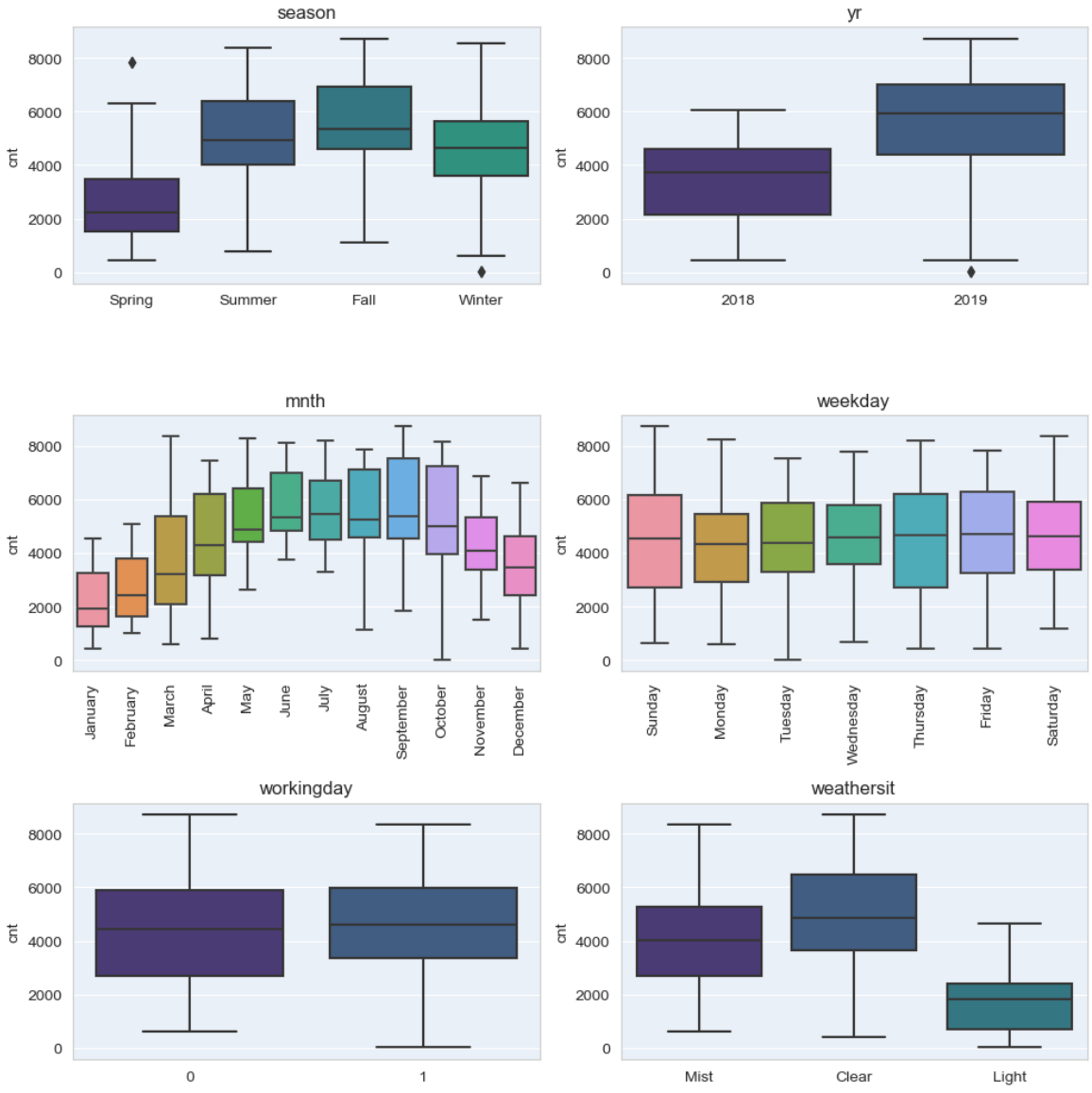
Assignment-based Subjective Questions

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

The effect of categorical variables on the dependent variable is important and noticeable and this can be visualized using the following boxplots:



The conclusions that could be drawn from above plots is that the count of total rental bikes is affect in following manner:

- Is higher around Summer and Fall and lower in Spring and Winter
- Is higher in year 2019 compared to 2018
- Is higher in June to November months compared to other months
- Is higher on Sundays
- Is higher during holidays
- Is higher during clear days

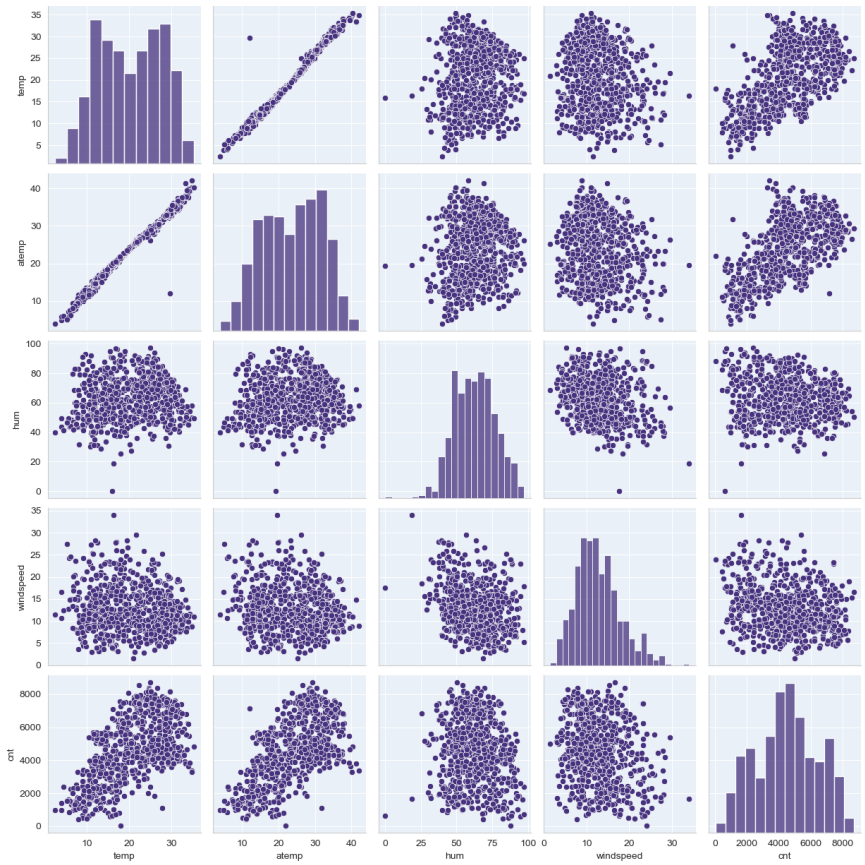
2. *Why is it important to use drop_first=True during dummy variable creation?*

To completely represent an original categorical variable, the number of dummy variables needed is typically n-1, where n represents the number of distinct categories in the variable.

Hence, it is important to use drop_first=True which effectively omits explicit dummy variable for one of the categories as it will be deduced from the values of other dummy variables. If n dummy variables are created instead of n-1 dummy variables then it could lead to a problem known as dummy variable trap. This is where a perfect multicollinearity occurs between dummy variables.

In conclusion, to avoid the problem multicollinearity occurring between dummy variables where values of one dummy have been fully inferred from the values of other dummy variables, we should drop one of the dummy variables by using drop_first=True. Note: we can choose and drop any one dummy variable. So it doesn't necessarily need to be the first dummy variable as long as we are dropping one of them.

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

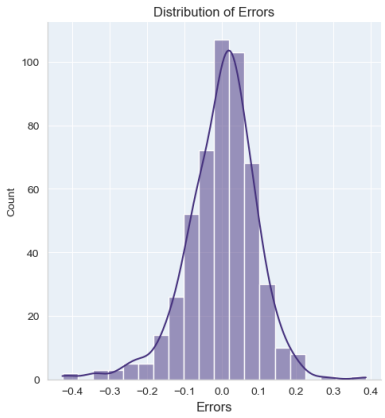


As displayed above 'temp' (also strongly correlated atemp) variable appears to have the highest correlation with the target variable according to the pair-plot.

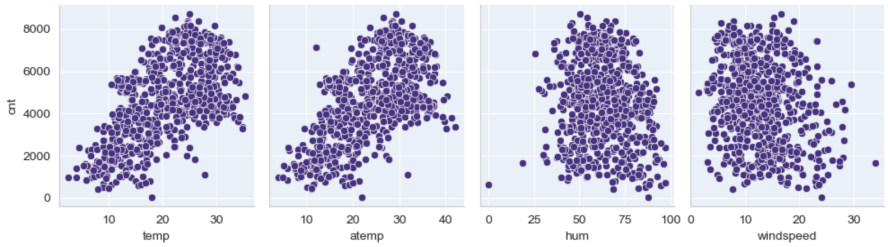
4. *How did you validate the assumptions of Linear Regression after building the model on the training set?*

   The assumptions which are required for Linear Regression are listed below along with steps we used to validate it on the training set.

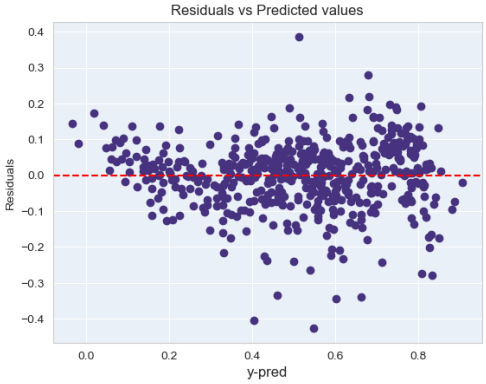
   i. <u>Normality of error terms</u>
   We validated this assumption by plotting a histogram of residuals and found the residuals to be normally distributed.



   ii. <u>Linear relationship</u>
   We use scatter plot between dependent and independent variables to test this assumption. This scatter plot shows a fairly linear correlation pattern between some of the features e.g. *cnt* and *temp*.



   iii. <u>Independence of residuals</u>
   We can use scatterplot of residuals vs predicted values validate the independency of residuals. There is no discernible pattern indicating independence of residuals.



iv. Homoscedasticity

Error terms must be homoscedastic in other words they must have constant variance. We use both the scatterplot of residuals vs predicted values (as shown above) and the scatterplot of predicted vs actual values on the training set to test for homoscedasticity.

v. Multicollinearity check

We calculated VIF values of each features to validate this assumption. All features in our final model exhibit VIF below 5 indicating that these features do not show significant Multicollinearity.

	Features	VIF
1	temp	4.22
5	yr_2019	2.06
3	season_Summer	1.94
6	mnth_July	1.58
4	season_Winter	1.57
9	weathersit_Mist	1.55
2	season_Spring	1.40
7	mnth_September	1.34
8	weathersit_Light	1.07
0	holiday	1.04

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

   Our final model is represented by the following linear equation:

$$0.503 \times temp - 0.299 \times weathersit\_Light + 0.233 \times yr\_2019 - 0.1 \times holiday + 0.083 \times season\_Winter + 0.081 \times mnth\_September - 0.078 \times weathersit\_Mist - 0.077 \times season\_Spring - 0.052 \times mnth\_July + 0.037 \times season\_Summer + 0.15$$

   The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

   i. **Temperature** – Higher temperature correlated with higher demand
   ii. **Year** – 2019 has higher demand
   iii. **Weathersit** – Light and Mist weathersit negatively correlated with demand

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear regression algorithm attempts to establish a linear relationship between dependent and one or more independent variables. It is a type of supervised machine learning algorithm. There are two types of linear regression models which are both based on the concept of representing the relationship between dependent and independent variable using a linear equation.

   1. Simple linear regression
   2. Multiple linear regression

   <u>Simple linear regression</u>

   In simple linear regression, we try to find the best-fitting straight line that represents the relationship between a dependent variable and an independent variable. In other words, we attempt to identify the linear equation that minimizes the differences between the observed dependent variable values and the values predicted by the line.

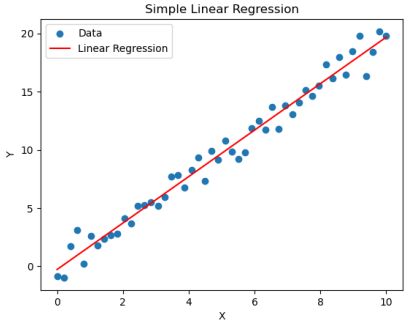


Figure 1: Linear regression line

Mathematically, equation for a straight line is given by:

$$Y = mx + c$$

In case of simple linear regression model, we can represent the best fitting line in by using equation of straight line in following terms:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Where,

- Y = Dependent variable
- X = Independent variable
- β_0 = the y-intercept
- β_1 = slope of the line or coefficient of feature X
- ε = the error term or residual

When the best fit lines is computed, the model can then be used to predict Y for unseen values of X. Linear regression models are effective for interpolation of data but not for extrapolation. Interpolation is the process of estimating or predicting the value of a dependent variable for independent values that fall within the range of available data.

Method of least squares

The values predicted by linear regression model may not actually be exactly equal to the observed value and the difference between predict and observed value of the dependent variable is called residual. The process of determining the best fit involves estimating the best values for β_0 and β_1 so as to minimize the sum of squared differences between observed and predicted values.

To optimize the regression coefficients, the sum of squared errors can be minimized by using mathematical techniques but typically in case of machine learning, they are computed using algorithms such as gradient decent.

Multiple linear regression

For multiple linear regression, the values of the dependent variable are predicted by deriving a linear relationship between the dependent variable and multiple independent variables. Like simple regression model, the aim again is to find the best-fit linear equation that minimizes the difference between the observed values of the dependent variable and the predicted values based on the independent variables.

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn + \varepsilon$$

Where,

- Y = Dependent variable
- X = Independent variable
- β0 = the y-intercept
- B1 = coefficient of feature X1
- B2 = coefficient of feature X2
- ε = the error term or residual

Linear regression algorithm can only be used when certain criteria are met. Among other critters the primary one is that there needs to exist a linear relationship between the dependent and independent variables. These are called assumptions of linear regression and all of them are listed below:

i. Normality of error terms: Residuals should be normally distributed.
ii. Linear relationship: There must be linear relationship between dependent and independent variables.
iii. Independence of residuals: Residuals should be independent of other variables.
iv. Homoscedasticity: Error terms must be homoscedastic in other words they must have constant variance.
v. No Multicollinearity: Intendent variables should not be highly corelated with other independent variables.

Main steps of Linear regression algorithm

In practice, linear regression algorithm entails the following steps:

i. **Reading and Understanding Data:** Use of data dictionary and preview different sections of data to understand its features.
ii. **Data Visualization & Exploratory Data Analysis:** Carry out various Data visualization, univariate, bivariate and multivariate analysis to check if the data contains linear relationship.
iii. **Data Preparation:** Create dummy variables where required, drop irrelevant columns.
iv. **Splitting Data into Training and Test sets:** Split data into train and test sets using for example 70% for training and 30% for testing.
v. **Building Linear Model:** Use either automated (such as RFE), manual or hybrid approach to select features for building the linear model. Ensure the features do not exhibit a high multicollinearity
vi. **Residual Analysis of the Train data**: Test if the residuals are normally distributed, independent and homoscedastic
vii. **Model Prediction**: Scale the test dataset and use the model to predict the dependent variable for the test dataset.
viii. **Model Evaluation:** Evaluate the performance of the model by using metric such as R-Squared and Adjusted to test if the model is either underfitted or overfitted.
ix. **Communicate the findings**: Based on the parameters of the model, now the findings effect of independent variables in the dependent variables can be communicated to the stakeholders.

2. Explain the Anscombe's quartet in detail.
   Anscombe's quartet is a group of four data sets which was developed by statistician Francis Anscombe in 1973 to demonstrate the significance of plotting data before analysis it statistically.

	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
Mean	**9.00**	**7.50**	**9.00**	**7.50**	**9.00**	**7.50**	**9.00**	**7.50**
Variance	**11.00**	**4.13**	**11.00**	**4.13**	**11.00**	**4.12**	**11.00**	**4.12**
PMCC	**0.82**		**0.82**		**0.82**		**0.82**	

Figure 2.1: Anscombe's quartet Dataset

The table contains four data sets and all of them feature identical descriptive statistics (mean, variance, standard deviation and so on) but has significantly different distribution and also appear very different when plotted.

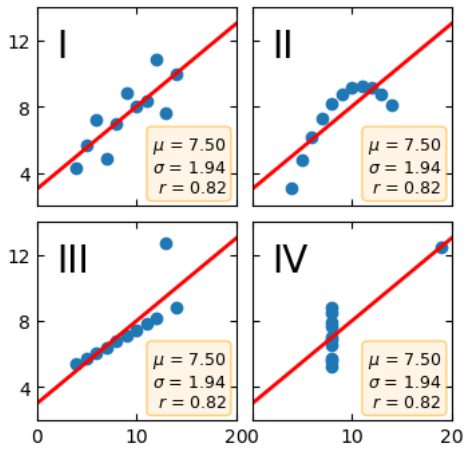


Figure 2.2: Graphical Representation of Anscombe's quartet. Source: (Hunter, John; Dale, Darren ; Firing, Eric; Dro, Michael; Matplotlib Development Team, 2012–2023)

However, when those four datasets are graphically represented, contrary to what their descriptive statistics suggest the four datasets are seem to exhibit very dissimilar patterns.

Hence, Anscombe's quartet demonstrates the limitations of solely replying on summary statistics and emphasizes the importance of visual exploration for accurate data understanding and interpretation.

3. What is Pearson's R?

   Person's R also known as Pearson correlation coefficient is quantification of the linear correlation between two datasets. It is a number between -1 and 1 that measures both the strength and direction of relationship between two variables.

   It is calculated by taking ratio between their covariance and product of their standard deviations. The following formula is given for calculating Pearson's R coefficient.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Pearson correlation coefficient	**Correlation type**	**Explanation**
Between 0 and 1	Positive correlation	When one variable change the other variable change in same direction
0	No correlation	No relationship between variables
Between 0 and -1	Negative correlation	When one variable change the other variable change in opposite direction

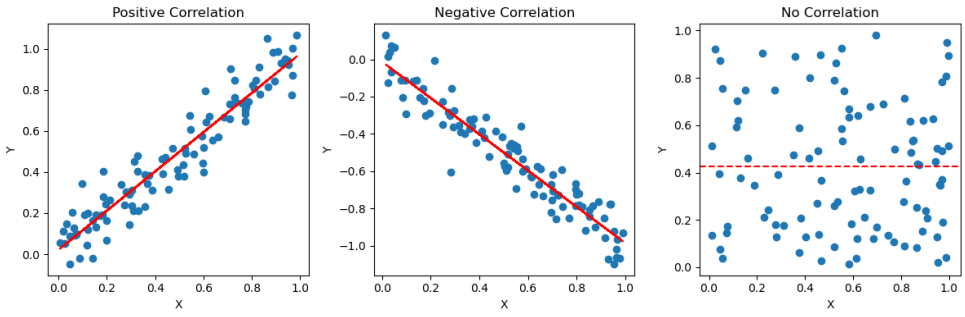


Figure 3: An example of Positive, Negative and No Correlation

It is important to note that the correlation coefficient measures the linear relationship between variables but does not imply causation. It's important to interpret the value of the coefficient in the context of the presented data.

Hence, the by calculating Pearson's correlation coefficient we can get an understanding of the strength as well as direction of linear relationship between two variables.

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

   Scaling is the process of adjusting numerical variables to a different more desirable range or distribution. It is considered as an important step for building a model from given dataset. It is mainly performed for reasons including:

   i. Avoiding Bias
      Scaling adjusts values for a feature to a range that is comparable to other features in the dataset. This prevents features with larger values to dominate the analysis and ensure all variable contribute to the analysis on same scale.

   ii. Optimized Model convergence
      Algorithms like gradient descent converge faster when the features are on similar scale.

   iii. Improve Interpretation of Coefficients
      Scaling make it easy and more intuitive to compare coefficients in a model. When all variable are in same scale comparison between coefficients becomes straight forward. The impact of each variable can be compare straight away by quantitively comparing the value of their coefficients.

   Normalize scaling and standardized scaling and two distinct types of scaling methods and the main difference among them are listed below.

Normalize Scaling	**Standardized Scaling**
Scales the value to a specific range, typically 0 to 1	Scales the value to be cantered around mean and have unit standard deviation
It is also referred to as Min Max scaling	It is also referred to as z-score scaling
Keeps the shape of original distribution	Alters the shape of original distribution
Sensitive to outliers	Less sensitive to outlier
Formula for Normalized Scaling given as: $$x' = \frac{(x - x_{\{min\}})}{(x_{\{max\}} - x_{\{min\}})}$$ *Where, $x_{\{min\}}$is min value for X and $x_{\{max\}}$ is max value for X.*	Formula for Standardized Scaling is given as: $$x' = \frac{(x - \mu)}{\sigma}$$ *Where, μ is the mean and σ is the standard deviation.*

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   VIF or variance inflation factor is used for estimating the degree of multicollinearity of multiple predictor variables in multiple linear regression.

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

   where:

R_i^2 = unadjusted coefficient of determination for regressing the i^{th} independent variable on the remaining ones

   As given by VIF formula if the value of R-squared is 1, which happens when there is perfect multicollinearity.

   Hence, the VIF can be infinite when there is perfect multicollinearity in the model. Perfect multicollinearity is when one or more variables in the dataset can be represented completely as linear combination of the other variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   The Q-Q plot, or quantile-quantile plot, is a scatter plot of two sets of quantiles against each other. It is a graphical technique for determining if two data sets come from the same population or not.

   Some important features of Q-Q plot are:
   - The sample sizes can be different as it plots the quantiles against each other
   - Shifts in location, scale, changes in symmetry, and the presence of outliers can all be detected from single Q-Q plot (National Institute of Standards and Technology, 2012)

   One of the assumptions of linear regression model is that the error terms are normally distributed. Q-Q plot can be used to test if the distribution of the residuals is normal. When residuals are plotted on a Q-Q plot it compares the quantiles of residuals to a theoretical normal distribution. If the residuals follow the a normal distribution, then a roughly straight line should be formed.

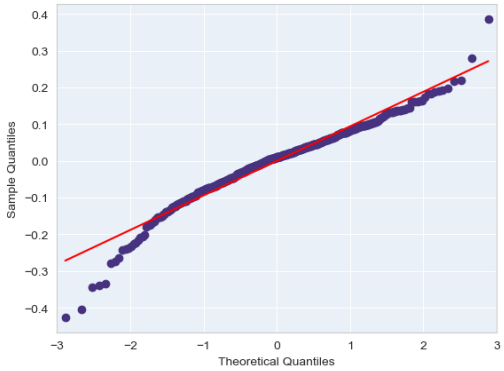


Figure 6: Example of Q-Q plot

Bibliography

Hunter, John; Dale, Darren ; Firing, Eric; Dro, Michael; Matplotlib Development Team. (2012–2023). *Matplotlib*. Retrieved from https://matplotlib.org/stable/gallery/specialty_plots/anscombe.html

National Institute of Standards and Technology. (2012). Retrieved from nist.gov: https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm