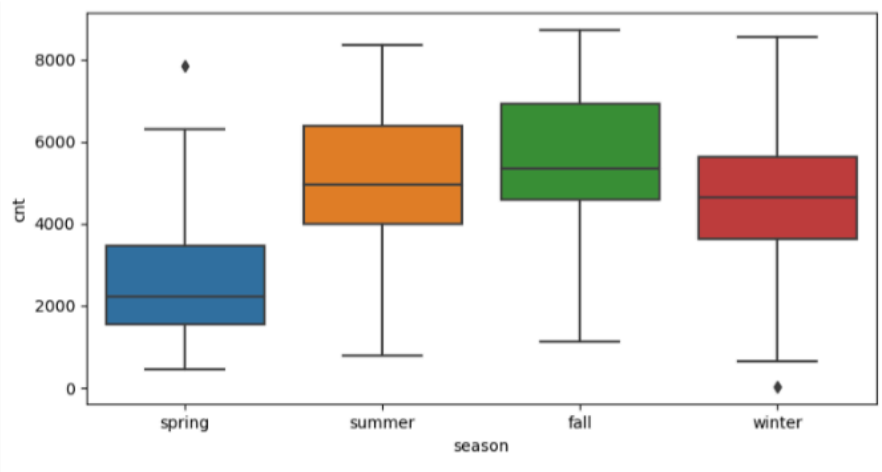
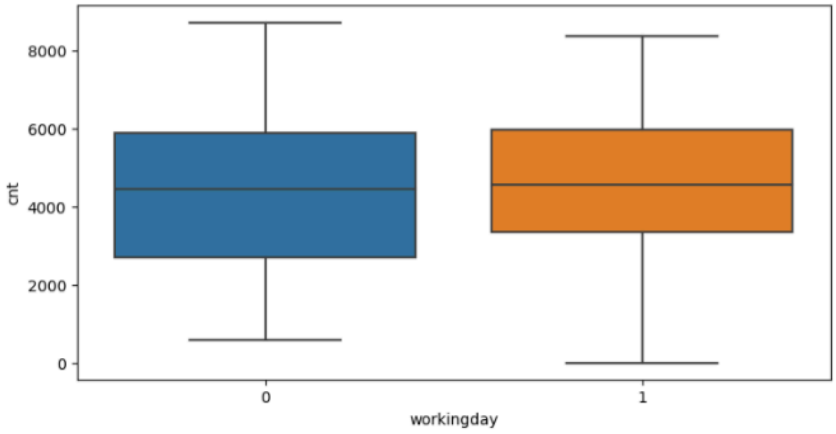
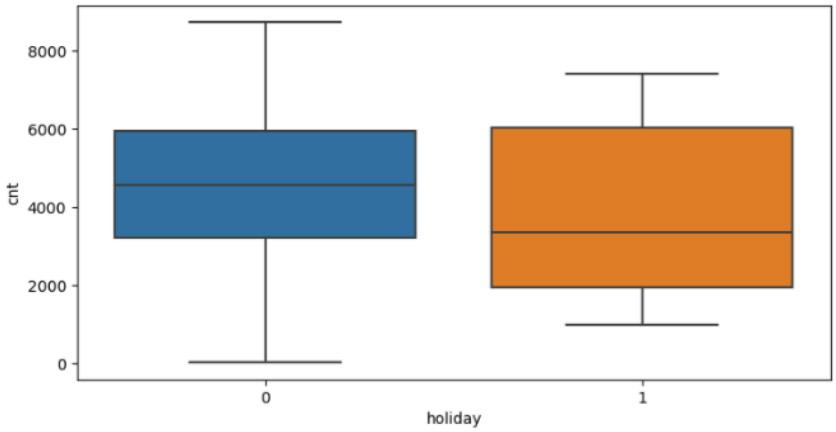
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

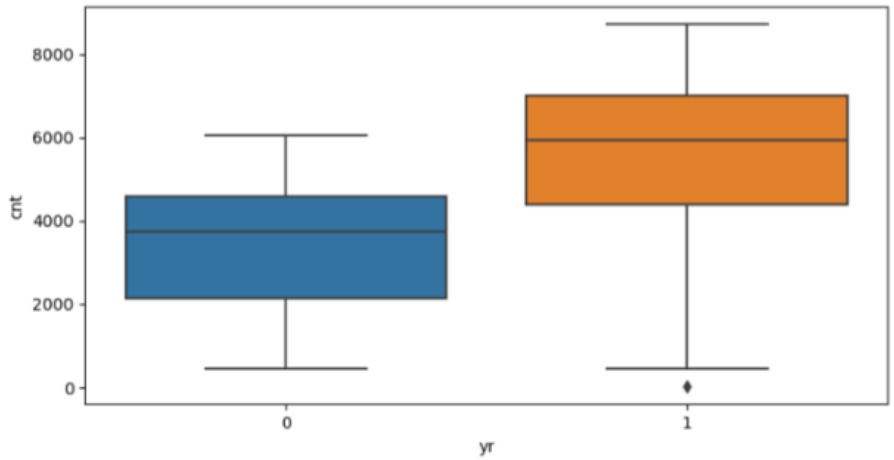


- Season plot shows highest rental of bikes happed in fall, followed by summer season

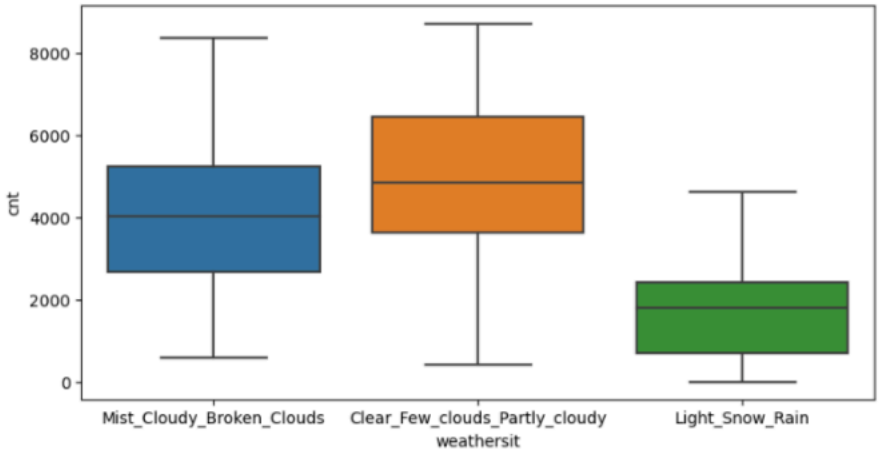




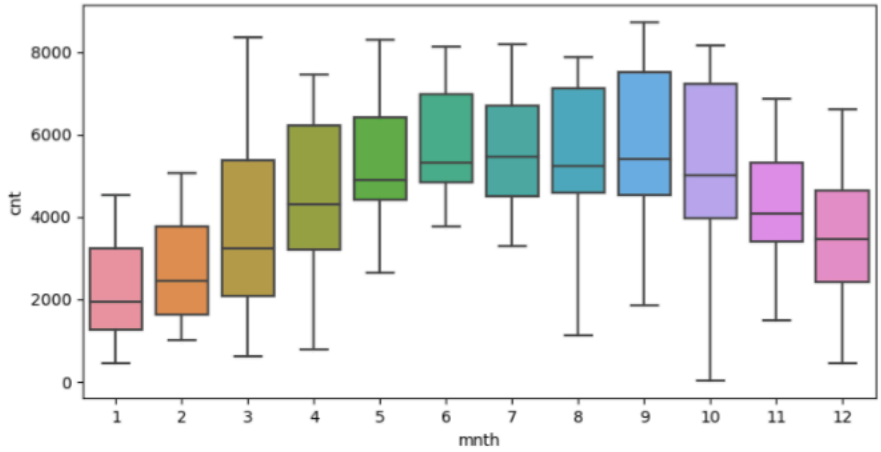
- Working day and holiday plots tell us that that bike rental is more on working days rather than weekends or holidays

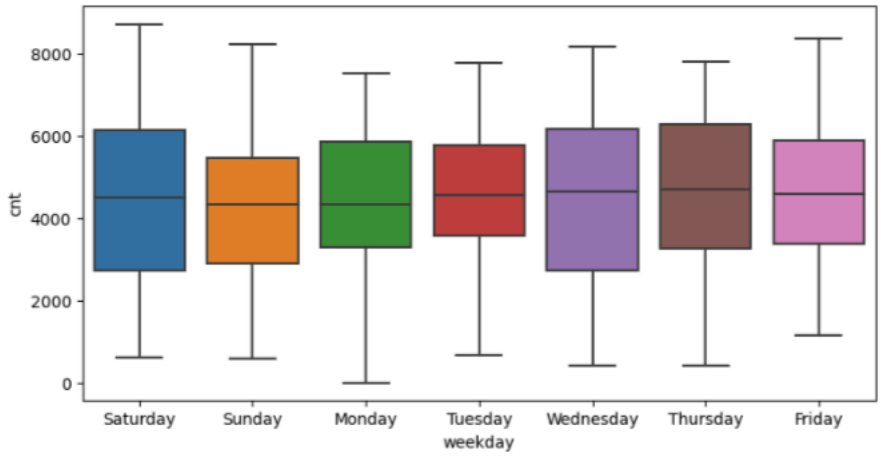


- Based on year box plot, more bikes were rented in year 2019 than 2018
- Based on weathersite parameter, when the weather was clear or partly cloudy on those days bike rental was more



- Based on below month and weekday plots, it can be inferred that, in the month of September the bike rented were more and on Saturdays more bikes were rented

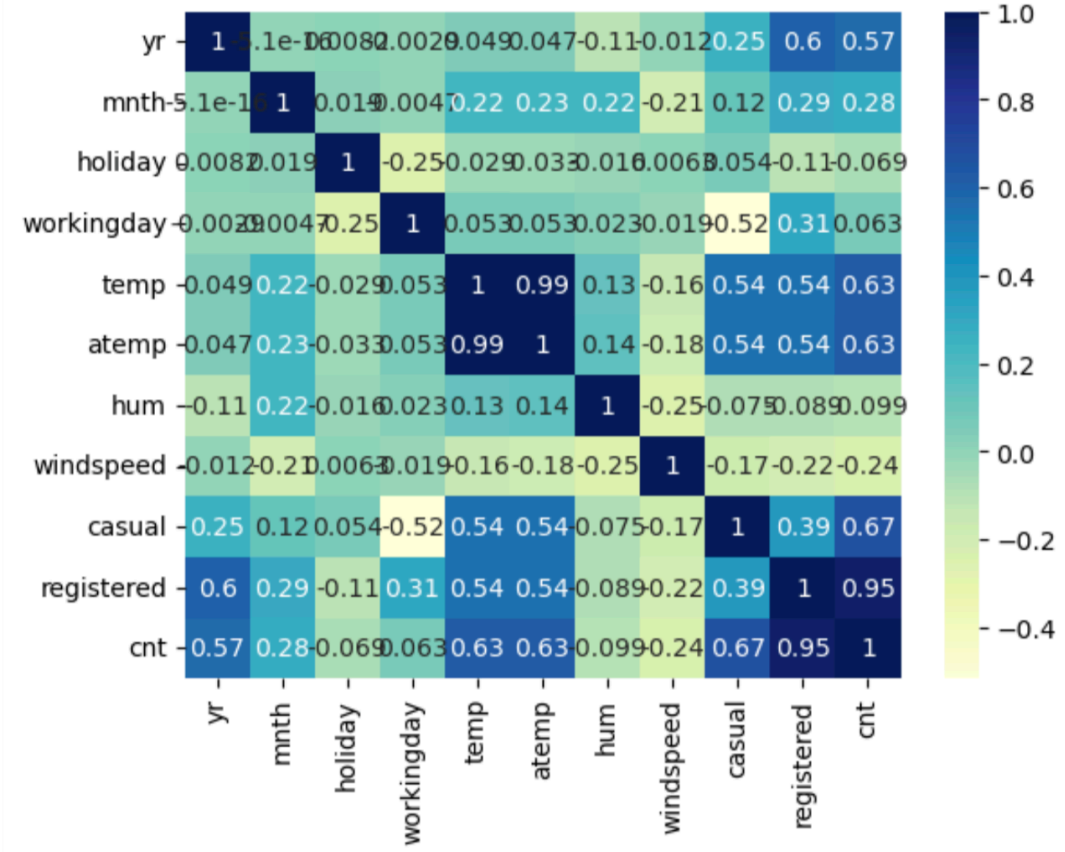

8000
6000
4000
2000
0
cnt
1
2
3
4
5
6
7
8
9
10
11
12
mnth


8000
6000
4000
2000
0
cnt
Saturday
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
weekday

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
    - While creating dummy variables, for n number of values n columns get created
    - However, n-1 columns are enough to determine the results, say columns with YES and NO, so if something is YES that means other column value is NO or NOT YES
    - By using drop_first= True, we delete the redundant column by reducing total number of columns by 1
    - Which in turn reduces the corelation created amoung the dummy variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

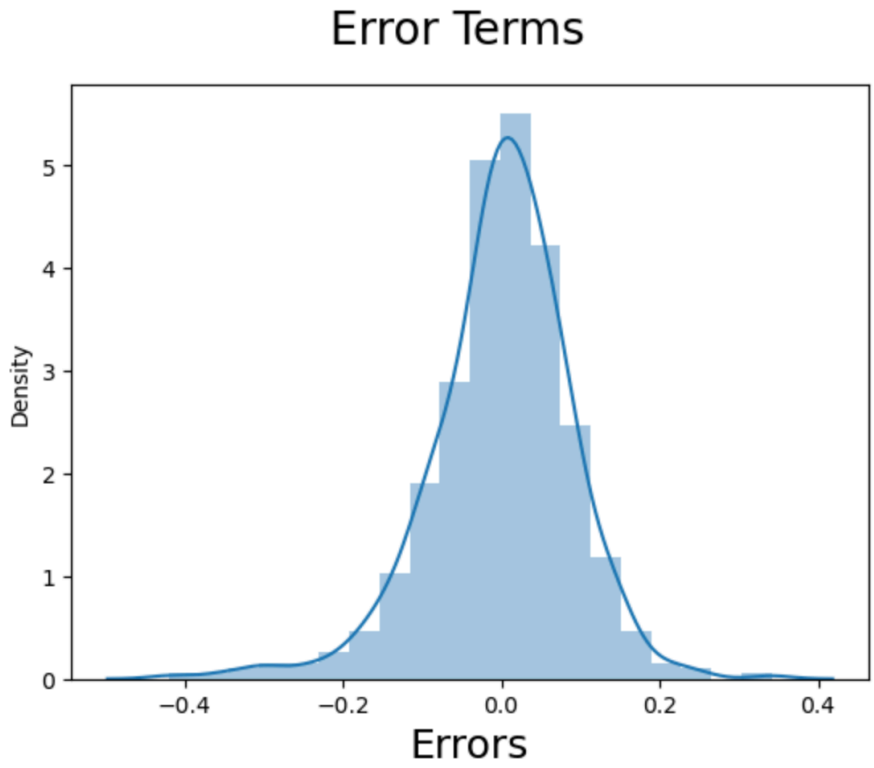
   Based on the heat map below, temp and atemp both have 0.63 corelation with cnt. (One of them can be hence dropped)



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The following tests were done to validate the assumptions of linear regression:

- First, linear regression needs the relationship between the independent and dependent variables to be linear. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not. Refer to the notebook for more details.
- Secondly, Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not. The diagram below shows that
  the residuals are distributed about mean = 0.



- Thirdly, linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to get the quantitative idea about how much the feature variables are correlated with each other in the new model.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   The top 3 features are:

   1. temp - coefficient : 0.4920
   2. yr - coefficient : 0.2336
   3. weathersit_Light Snow & Rain - coefficient : -0.2904

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

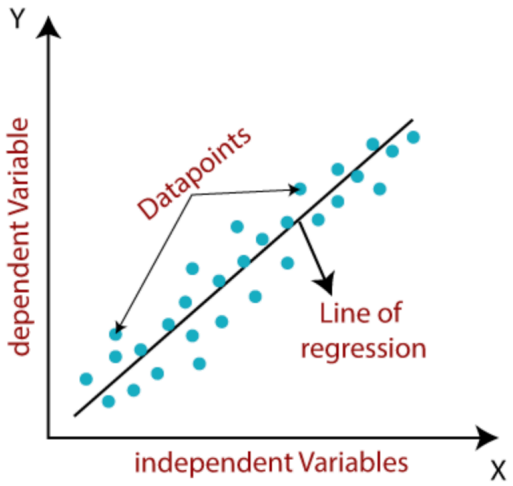
Based on number of input variable liner regression can be classified into two categories

- Simple Linear Regression
  a single independent variable is used to predict the value of a numerical dependent variable
- Multiple Linear Regression
  more than one independent variable are used to predict the value of a numerical dependent variable

There are many ways to train the data, one out of those mostly used is Ordinary least square.

The linear regression algorithm originates from statistics and is used in Machine learning. This is a most popular algorithm to do predictive data analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematical representation of any straight line is

y= b0+b1x+ e

Here,

y – target variable/ dependent variable

x – independent variable / predictor variable

b0 – y-axis intercept , gives additional degree of freedom

b1 – slop of a line / linear regression coefficient

e – random error

Assumptions of Linear Regression.

These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

1. Linear relationship between the features and target
2. Small or no multicollinearity between the features
3. Homoscedasticity Assumption
4. Normal distribution of error terms
5. No autocorrelations

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.
Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.
Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson's R? (3 marks)
   - Pearson's correlation coefficient, also known as Pearson's r, is a statistic that measures the linear relationship between two variables. It is the most common way to measure a linear correlation.
   - Value ranges from -1 and 1
   - Known by multiple names
     - Pearson's r
     - Bivariate correlation
     - Pearson product-moment correlation coefficient (PPMCC)
     - The correlation coefficient
   - It is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.
   - It can be used to test statistical hypotheses, we can test whether there is a significant relationship between two variables.
   - Pearson correlation coefficient (r) is as a measure of how close the observations are to a line of best fit. Also tells whether the slope of the line of best fit is negative or positive.
     - Slope negative, r negative
     - Slope positive, r negative
     - When r = 1/-1, all the points fall exactly on straight line

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   - What is scaling?

Scaling is a data pre-processing step, which is applied on independent variables for data normalization to fit data within particular range. Helps in speeding up the calculations in an algorithm.

- Why is scaling performed?
    - *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.*
- Difference between normalized scaling and standardized scaling
    1. Standardization centers data around a mean of zero and a standard deviation of one, while normalization scales data to a set range, often [0, 1], by using the minimum and maximum values
    2. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python
    3. sklearn.preprocessing.scale helps to implement standardization in python
    4. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

   If there is a perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

   What to do if VIF is large?

   If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression. Here are the various options:

   - One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.
   - A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.
   - The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
   - The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.
   - Finally, you can use a different type of model call ridge regression that better handles multicollinearity.

In conclusion, when you are building a multiple regression model, always check your VIF values for your independent variables and determine if you need to take any corrective action before building the model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   - A quantile-quantile (Q-Q) plot is a scatterplot that compares the quantiles of two distributions. It's also known as a Q-Q plot.
   - A Q-Q plot can help determine if a set of data came from a theoretical distribution, such as normal or exponential. It can also help determine if a dataset follows a particular type of probability distribution, like uniform or exponential.
   - A Q-Q plot plots the quantiles of a sample distribution against quantiles of a theoretical distribution. A quantile is the fraction or percent of points below a given value.
   - A Q-Q plot can be useful in parametric tests because they assume normality.
   - Q-Q plots can help
     - Determine if two samples are from the same population
     - Determine if two samples have the same tail
     - Determine if two samples have the same distribution shape
     - Determine if two samples have common location behavior
     - Determine if a dataset follows any particular type of probability distribution
   - Advantages
     - helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions
     - *It can be used with sample sizes also*
     - *Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.*

Reference:

https://medium.com/@premal.matalia/what-is-scaling-why-is-scaling-performed-normalized-vs-standardized-scaling-5113c86688f8
https://builtin.com/data-science/anscombes-quartet
https://www.geeksforgeeks.org/anscombes-quartet/
https://www.scribbr.com/statistics/pearson-correlation-coefficient/
https://www.investopedia.com/terms/p/pearsoncoefficient.asp
https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/#:~:text=How%20is%20Standardization%20different%20from,the%20minimum%20and%20maximum%20values.
https://www.sigmamagic.com/blogs/what-is-variance-inflation-factor/#:~:text=If%20there%20is%20perfect%20correlation,to%20the%20presence%20of%20multicollinearity.
https://medium.com/@premal.matalia/q-q-plot-in-linear-regression-explained-ab040567d86f