Linear Regression Assignment Questions - Sachin Karthik

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

→ The categorical variables found in the final model which help in determining the value of the dependent variable, demand (count), **are season, weather situation, weekday (or holiday).**

While the categorical variables related to weather and climate do have a considerable amount of influence on the dependent variable, the day of the week doesn't make a huge difference.

This was inferred by looking at the coefficients of the variables in the final qualifying model. The coefficient for a specific variable estimated the amount of rise in demand when the variable value is increased by one unit while all the other variables are held constant. The coefficients of the categorical variable in the qualifying model are

1. Season = 0.3227
2. Weather Situation = -0.1692
3. Weekday = 0.0749

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

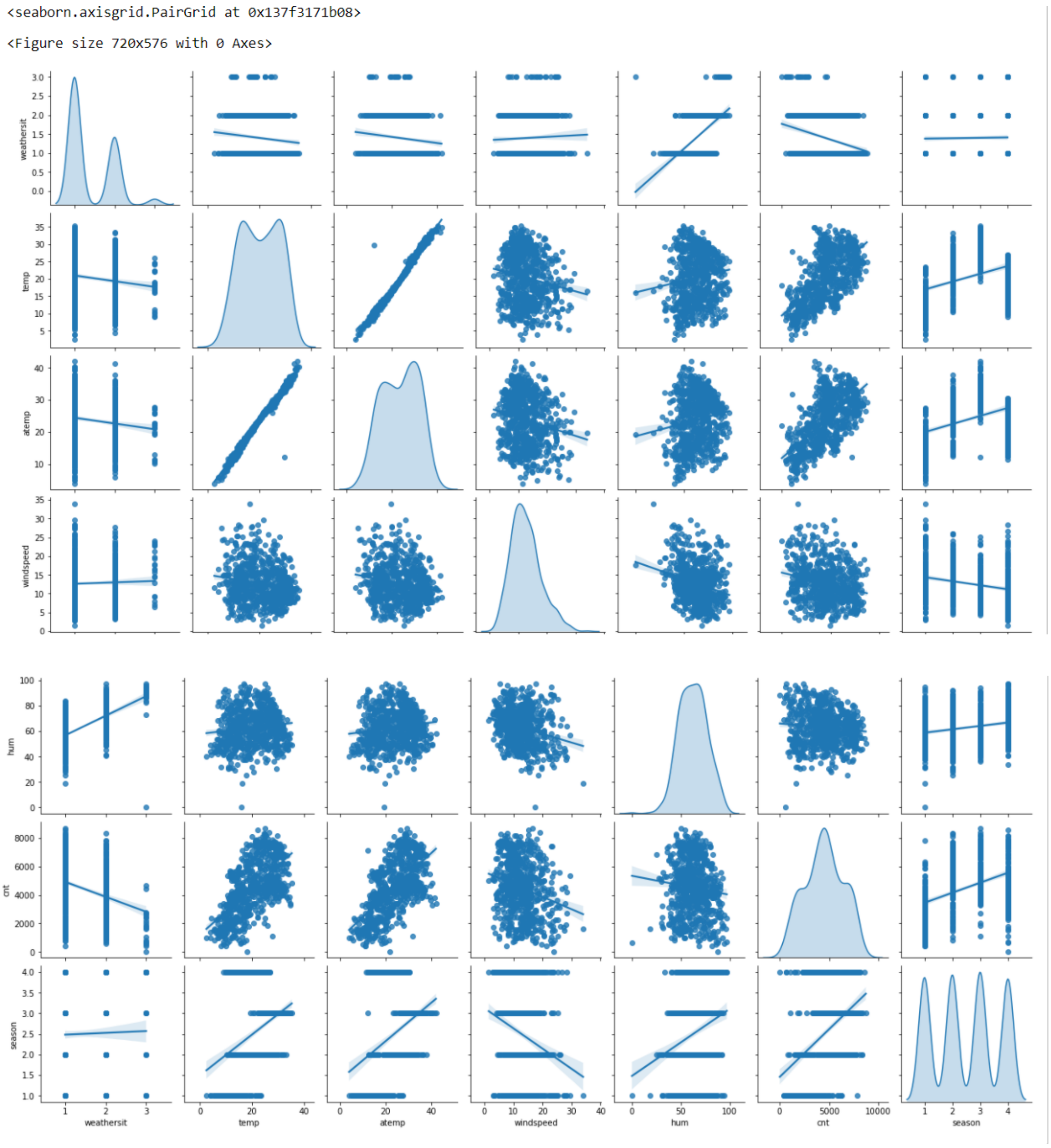
→ While there are k levels in a categorical variable, it is tempting to create k dummy columns (or variables) to represent each level with a distinct column, it is important to consider the multicollinearity issues which arise with that outlook towards dummy variable creation.

Multicollinearity is an issue that arises while building an ML model if two or more variables carry the same amount or very similar information. This essentially affects the interpretability of the model.

If we create k dummy variables for k levels of categorical data, the kth variable contains no new information.

This is why it is important to use drop_first = True when we create dummy variables, to avoid multicollinearity issues and maintain an interpretable model.

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



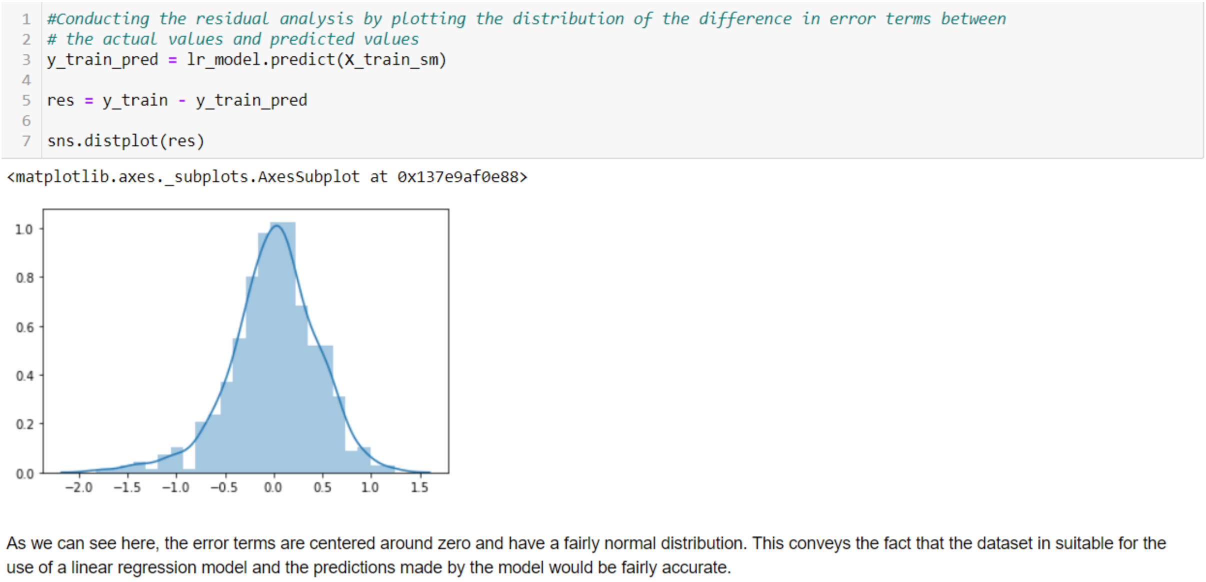
```
<seaborn.axisgrid.PairGrid at 0x137f3171b08>

<Figure size 720x576 with 0 Axes>
```

→ Looking at the pairplot, it is quite evident that there is a clear trend between the **temp, atemp variables** and the target variable, count.

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

→ While building the model, the Variance Inflation Factor was monitored closely to remove feature variables with very high correlation values to combat multicollinearity and preserve the interpretability of the model.

- After training the model with a good number feature variables and after arriving at satisfying values for adjusted R-squared, AIC and BIC, the model was evaluated on a test set that was separated from the main dataset before beginning the analysis. Out of the 2 models, the qualifying model gave an R-squared value of 0.778 on the evaluation set.
- Another assumption of Linear Regression was tested by conducting a residual analysis on the error terms of the fitted data. The end result of the residual analysis was a normal distribution of the error terms, centered around 0.



```
#Conducting the residual analysis by plotting the distribution of the difference in error terms between
# the actual values and predicted values
y_train_pred = lr_model.predict(X_train_sm)

res = y_train - y_train_pred

sns.distplot(res)
```

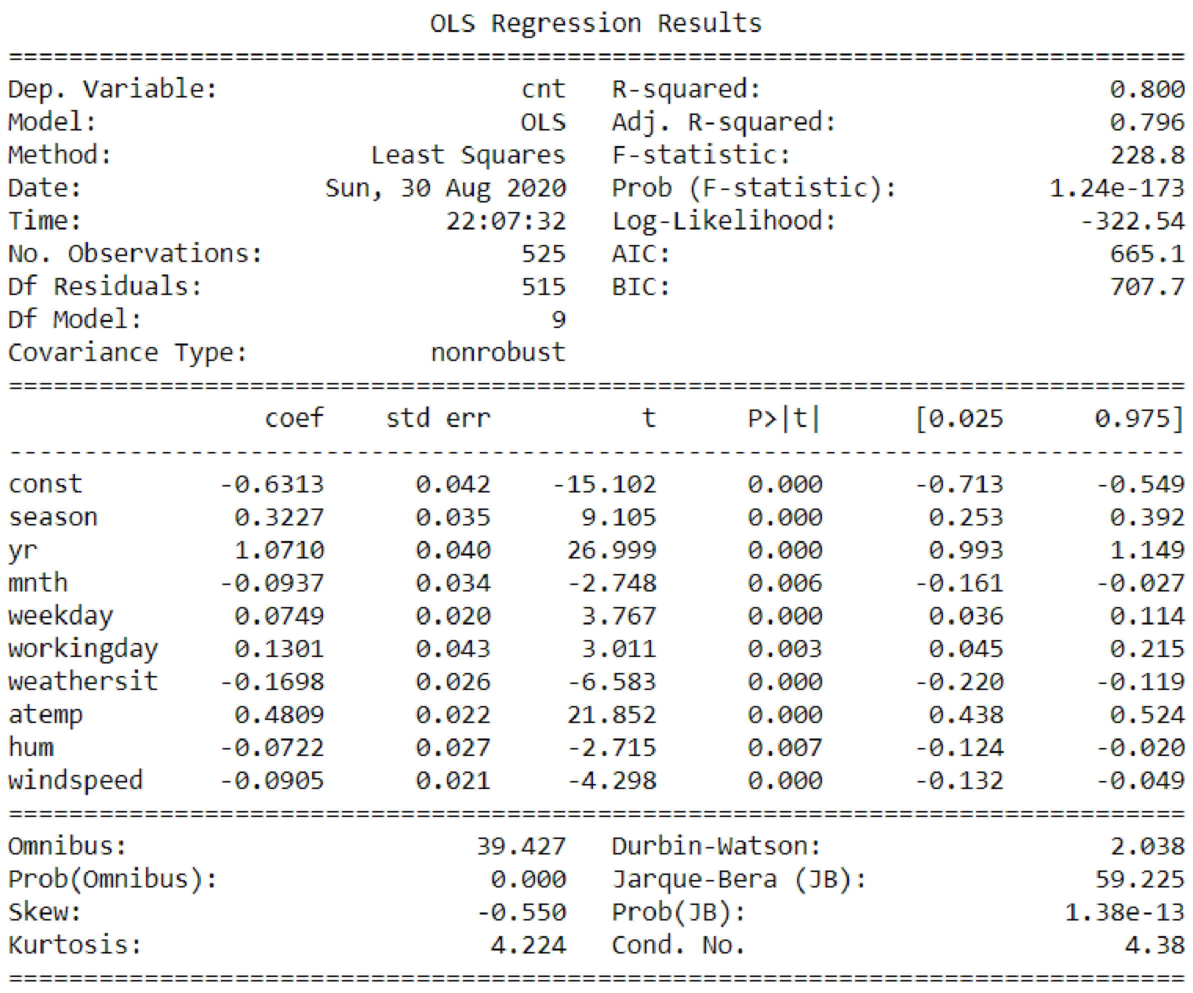
```
<matplotlib.axes._subplots.AxesSubplot at 0x137e9af0e88>
```

As we can see here, the error terms are centered around zero and have a fairly normal distribution. This conveys the fact that the dataset in suitable for the use of a linear regression model and the predictions made by the model would be fairly accurate.

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

→ Based on the coefficients of the final model, the top 3 features which explain the variability of the demand and are conducive to predicting the value of demand are

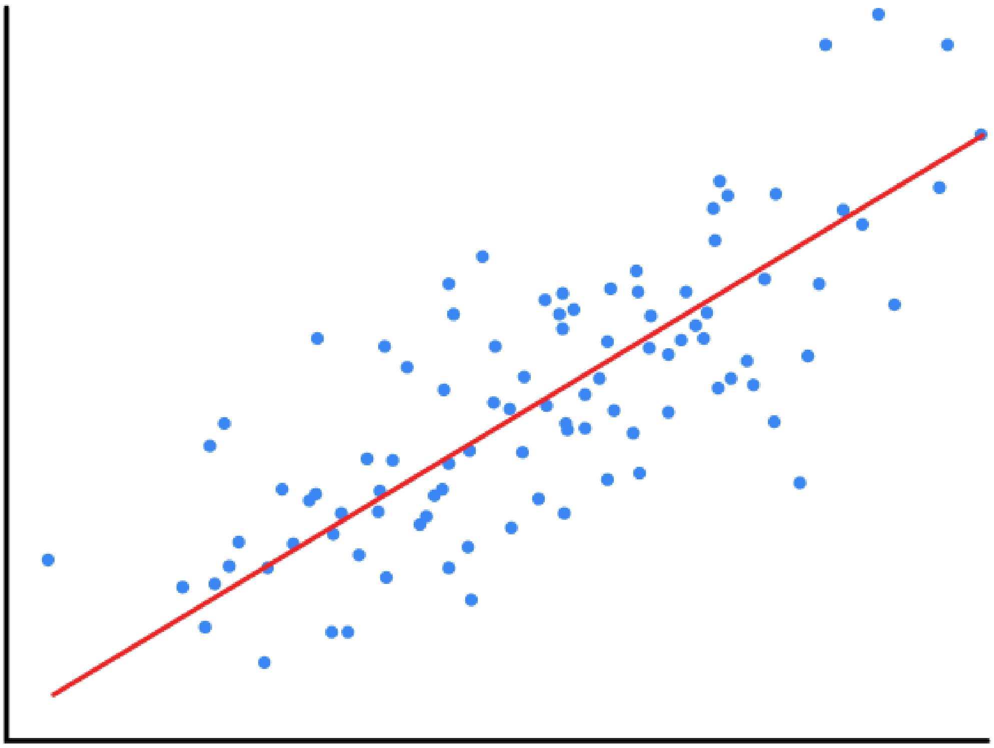
1. Year
2. Atmospheric Temperature (feeling temperature)
3. Weather Situation. The demand for cycles reduce when the weather situation worsens.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.800
Model:                            OLS   Adj. R-squared:                  0.796
Method:                 Least Squares   F-statistic:                     228.8
Date:                Sun, 30 Aug 2020   Prob (F-statistic):          1.24e-173
Time:                        22:07:32   Log-Likelihood:                -322.54
No. Observations:                 525   AIC:                             665.1
Df Residuals:                     515   BIC:                             707.7
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.6313      0.042    -15.102      0.000      -0.713      -0.549
season         0.3227      0.035      9.105      0.000       0.253       0.392
yr             1.0710      0.040     26.999      0.000       0.993       1.149
mnth          -0.0937      0.034     -2.748      0.006      -0.161      -0.027
weekday        0.0749      0.020      3.767      0.000       0.036       0.114
workingday     0.1301      0.043      3.011      0.003       0.045       0.215
weathersit    -0.1698      0.026     -6.583      0.000      -0.220      -0.119
atemp          0.4809      0.022     21.852      0.000       0.438       0.524
hum           -0.0722      0.027     -2.715      0.007      -0.124      -0.020
windspeed     -0.0905      0.021     -4.298      0.000      -0.132      -0.049
==============================================================================
Omnibus:                       39.427   Durbin-Watson:                   2.038
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               59.225
Skew:                          -0.550   Prob(JB):                     1.38e-13
Kurtosis:                       4.224   Cond. No.                         4.38
==============================================================================
```

Explain the linear regression algorithm in detail.

⟶ Linear Regression is a Machine Learning algorithm which stems from the statistical model Linear Regression. Linear Regression is used to find the Linear Relationship between 1 or more independent variables and a dependent variable.



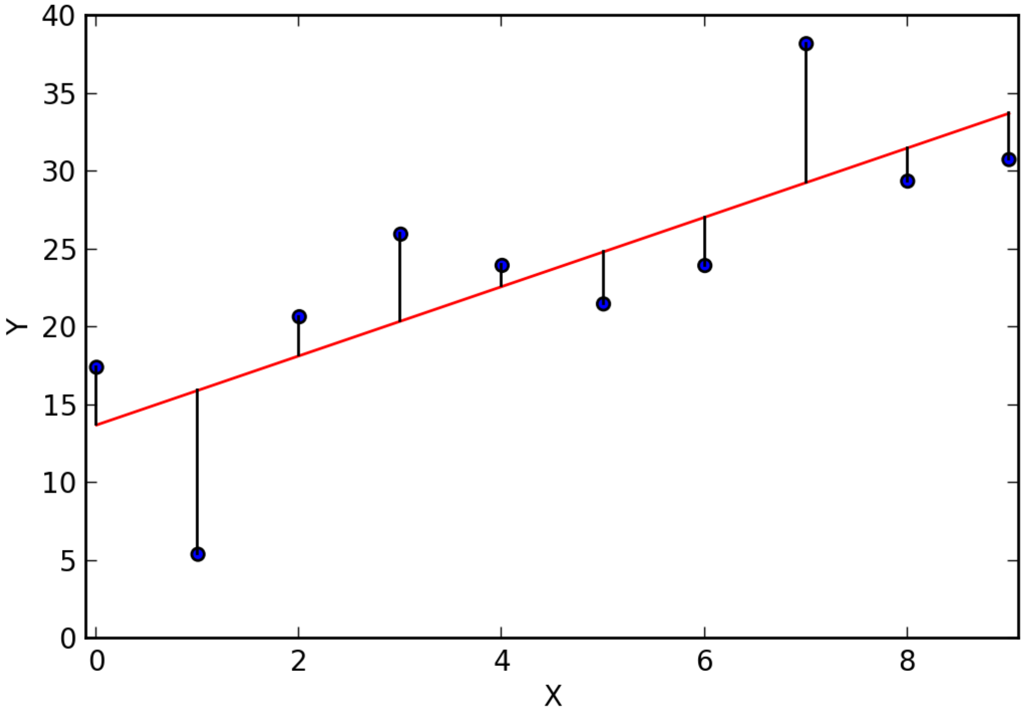
In the above image, if we consider the **variable X to be the balls faced by a batsman in a match** (say Virat Kohli) and the **dependent variable y to be the runs scored by him**, we could see that there is a linear trend between the balls faced and the runs scored in each match.

Once we spot this linear trend, we could try fitting a line through the data points in a way that the error between the data points and the predicted values on the line, for a given value of x in minimum.

The error scored and summed (we are squaring to avoid the values of the errors cancelling out due to the change in sign of the error term) is known as the residual of the line.

Residual Sum of Squares (RSS) = $\Sigma(yitrue - yipred)^2$

Where i ranges from 1 to n.



By repeating the RSS calculation for a variety of different lines and slopes with different intercept values, we can finally determine the best line for the given set of data points by selecting the line which has the least RSS value.

The general form of the best fit line equation is given by

$$y = \beta 0 + \beta 1X$$

Where β0 - The y intercept of the line

β1 - Slope of the line

One of the core assumptions of Linear Regression is that it is possible only if the error terms (Residuals) are normally distributed, else the model would give us inaccurate predictions.

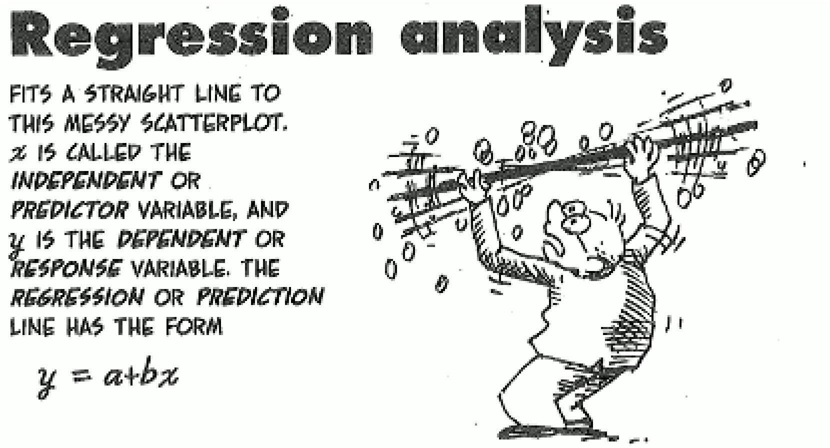
To help us calculate the best fit line for the given set of data points, we also use a parameter known as the R-squared, which tells us the fit between the data points and the line.

The R-squared value is the squared value of the pearson correlation coefficient.

We can also perform linear Regression with more than one independent variable but the dependent variable has to be only one.

In a nutshell, Linear Regression would be a simple and fairly accurate model if

1. There is a linear relationship between the independent variables and the dependent variable
2. The fit of the line with the data points gives us fairly accurate predictions, given the values of the dependent variables (which is measured using the R-squared score)
3. The error terms are normally distributed.


Regression analysis
FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT. x IS CALLED THE INDEPENDENT OR PREDICTOR VARIABLE, AND y IS THE DEPENDENT OR RESPONSE VARIABLE. THE REGRESSION OR PREDICTION LINE HAS THE FORM
y = a+bx

Explain the Anscombe's quartet in detail.

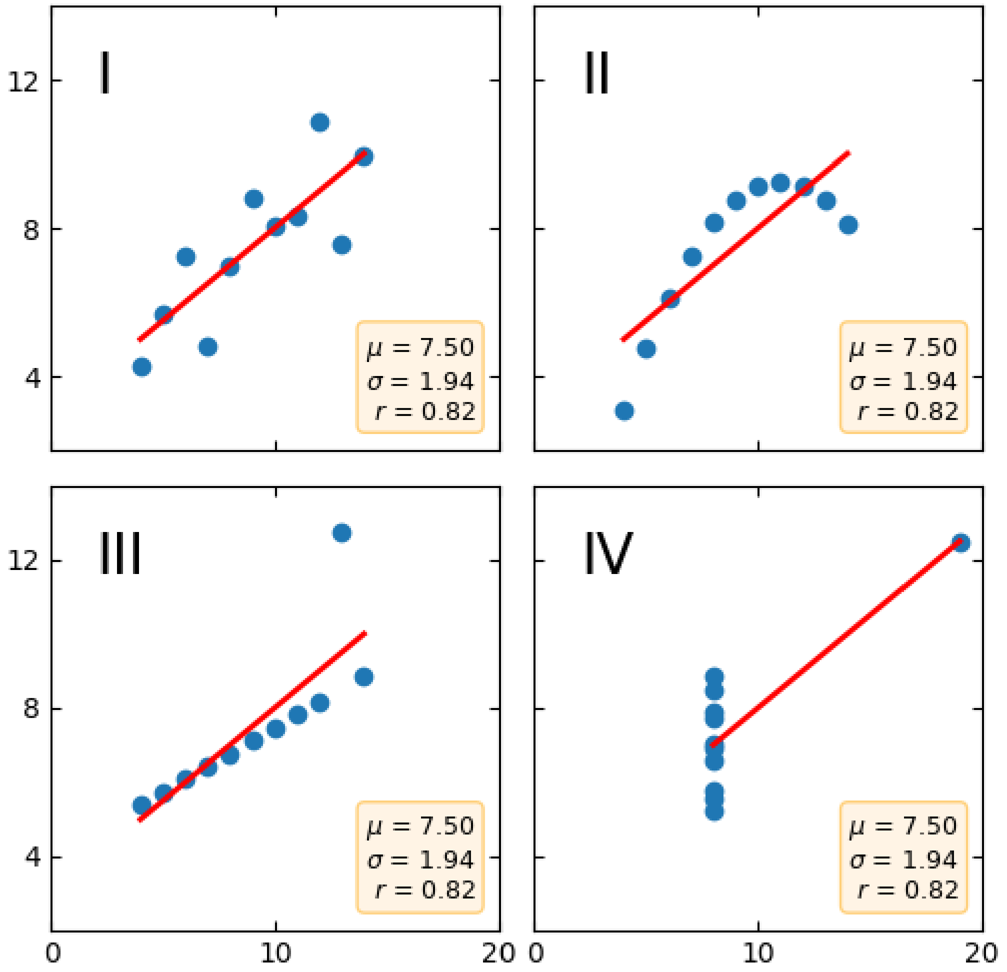
→ Anscombe's Quartet is a concept in Statistics which explains the importance of visualizing data in addition to looking at the Summary Statistics.

Summary Statistics of any data gives us the big picture rather than showing the value of each datapoint and allowing us to intuitively estimate the ranges and values of the data.

Summary Statistics consists of calculating values like the Average (mean), Median, Mode (if categorical), and the spread of data — minimum value, 25th percentile, 50th percentile (same as median), 75th percentile and maximum value — if it is a numerical column.

While it does give us a good idea about the data, it doesn't really help us look at the shape or the distribution of data.

Anscombe's Quartet is a collection of 4 different datasets with different individual data points have the same average values for the data yet widely different distributions (shape).



In the above picture, the summary statistics of the 4 distinct dataset have identical values - 7.50 for mean of the data, 1.94 for Standard Deviation and 0.82 for the Correlation Coefficient - yet they have vastly different distributions.

This is the crux of the concept of Anscombe's Quartet - How summary statistics don't give the full picture, literally!

What is Pearson's R?

→ Pearson's R is a statistical measure that is used to determine the measure of the strength of association between two numerical and Linear Variables.

Pearson's correlation coefficient is usually calculated by plotting the values of the independent variable of a sample on the x-axis and the corresponding values of the dependent variable of the sample on the y-axis. Note that, the variables strictly don't have to be dependent on one and another.

After plotting the values on the graph, the covariance values are calculated by the formula

$$Cov(x,y) = \Sigma(xi - xbar)(yi - ybar)/N - 1$$

where xi - X value of an individual data point

yi - y value of an individual data point

x bar - Mean of X

y bar - Mean of y

Once the covariance value is calculated, the Correlation coefficient is calculated by dividing the value of covariance with the standard deviation of X and Y.

$$R = Cov(x,y)/\sigma x\sigma y$$

σx - Standard Deviation of X

σy - Standard Deviation of y

The Correlation coefficient tells us if there is a strong positive relationship, strong negative relationship, weak positive relationship, weak negative relationship or no relationship between the 2 variables. The Correlation Coefficient value would only range in between -1 to1.

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

→ Scaling is a process of converting a feature variable or multiple feature variables into a standard scale or a common scale.

- The reason this is done is because as multiple feature variables come with their own scales of data and their distribution curves, the ML model which is using the features to predict the response variable would implicitly take the feature variable with a higher scale to be more important than a feature variable with a lower scale.
- Another reason why scaling is recommended is it makes the interpretation of the model really hard since the coefficients for different variables would have extremely high and low values, due to various scales.

Scaling prevents some of the above listed issues and makes the data much more interpret-friendly and it also expedites the process of finding the coefficients since gradient descent algorithm responds better to data in a constricted scale than in a large/ very small variable scale combinations.

The two common types of scalers used are

1. MinMax Scaler (Normalized Scaler)
2. Standard Scaler

MinMax Scaler or Normalized Scaler scales the data to fit it within the range of 0 to 1, no matter however big the range of the data is.

MinMax Scaler achieves this by learning the values of the minimum and maximum values initially, subtracting the value of each data point from the minimum value and dividing it by the range of the data.

MinMax Scaler = $(x - xmin)/(xmax - xmin)$

Standard Scaler converts the values of the variable to a format where it sets the average value of the data close to zero and the standard deviation of the data to 1. This method of scaling is quite beneficial if the variable which is about to be scaled follows a normal distribution.

Standard Scaler = $(x - \mu)/\sigma x$

Both the scalers are affected by outliers since the parameters used to scale the values are affected by outliers.

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

→ The formula for Variance Inflation Factor is

$$VIF = 1/1 - Ri^2$$

Where Ri^2 - The R-squared value of the fit between a feature variable as the dependent variable and the other feature variables as independent variables.

When the Ri-squared value tends to 1, the denominator of the equation tends to zero, which in turn causes the value of the Variance Inflation Factor to tend to infinity.

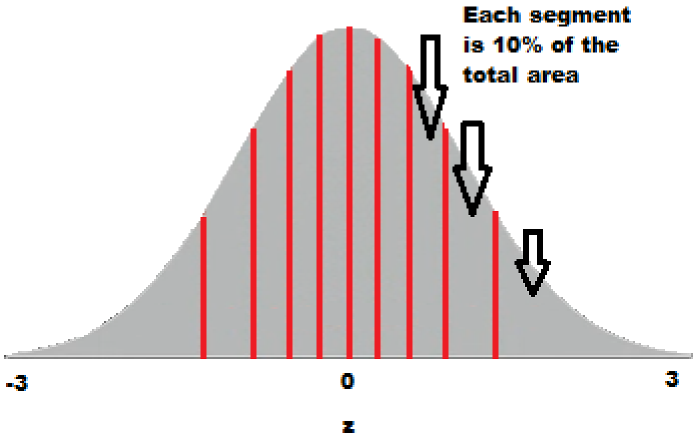
This effectively tells us that, when a feature variable is absolutely correlated with the other feature variables, the value of the variance inflation factor tends to become infinite.

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

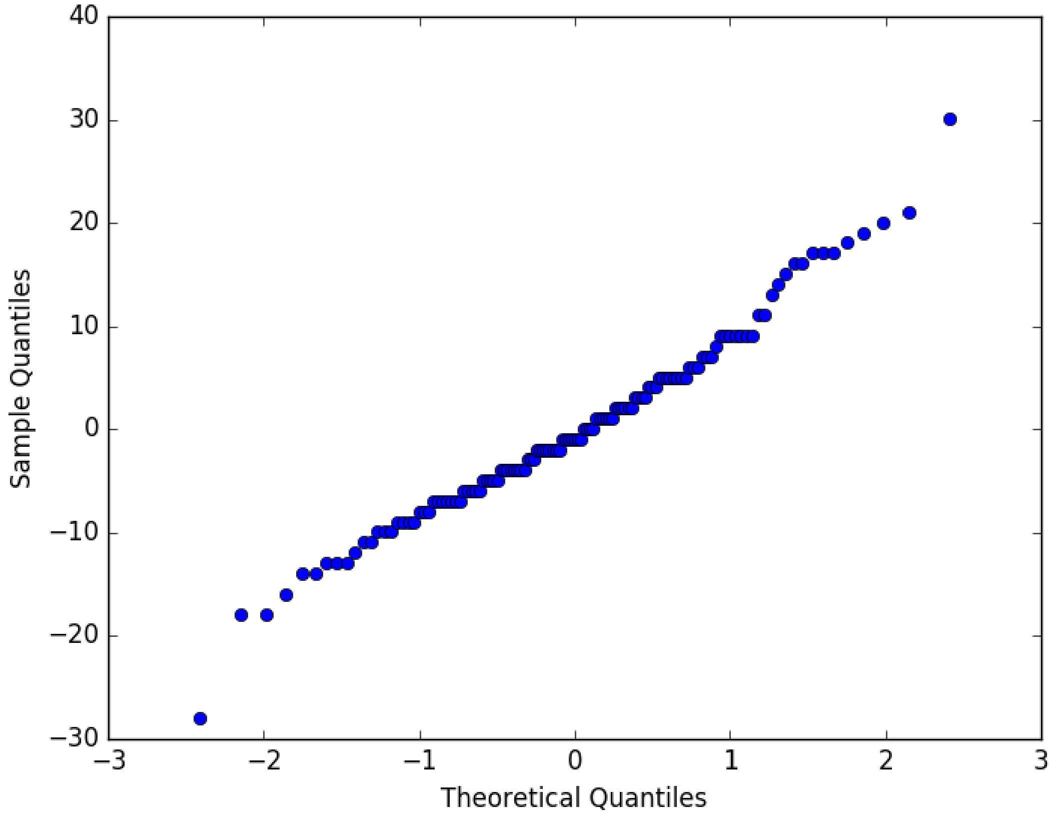
Quantile-Quantile plot or Q-Q plot is a method to figure out the distribution of an unknown sample dataset. This is done by repeating a series of simple steps.

Consider we have a set of data points on the amount of salt content in a biscuit, from a batch of biscuits from a biscuit manufacturing company. We want to figure out whether the salt content follows a normal distribution or uniform distribution.

1. To do this, first we need to order the values of salt content and find out the quantiles or percentiles values for the set of data points.
2. Next, we need to take a standard normal distribution curve and divide the curve into the same number of parts as the number of values in our salt content column. To account for the variable probability of a normal distribution curve, the lines at the end would have more gap between each other and the lines in the middle of the curve would be close to each other.



3. Once we are done with these 2 steps, we need to plot the values of the salt content in the x-axis of the Q-Q plot and the values of the normal distribution curve on the y-axis of the Q-Q plot. The data points of the Q-Q plots would emerge where the percentile values of the salt content and the normal distribution curve line values coincide.



4. After all the points are plotted at the intersection point of the values of the x-axis and y-axis variables, a line is drawn to fit the data. If the points sit pretty well on the line, the distribution follows a normal distribution.

5. If the data points don't fit well with the line, the process is repeated for a uniform distribution curve.

This process doesn't work only for normal of uniform distributions but any other kind of theoretical distribution like left skewed, right skewed, etc.