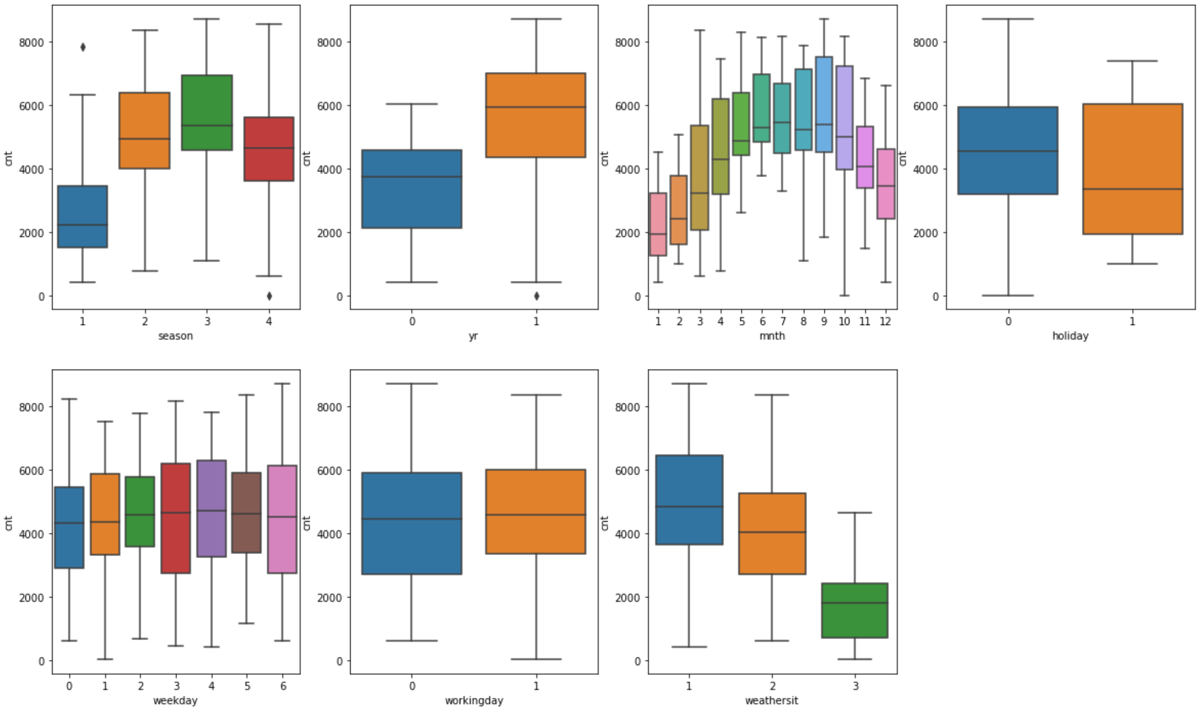
# Assignment-based Subjective Questions

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

   Following plots were used to understand the effect of categorical variables on the dependent variable



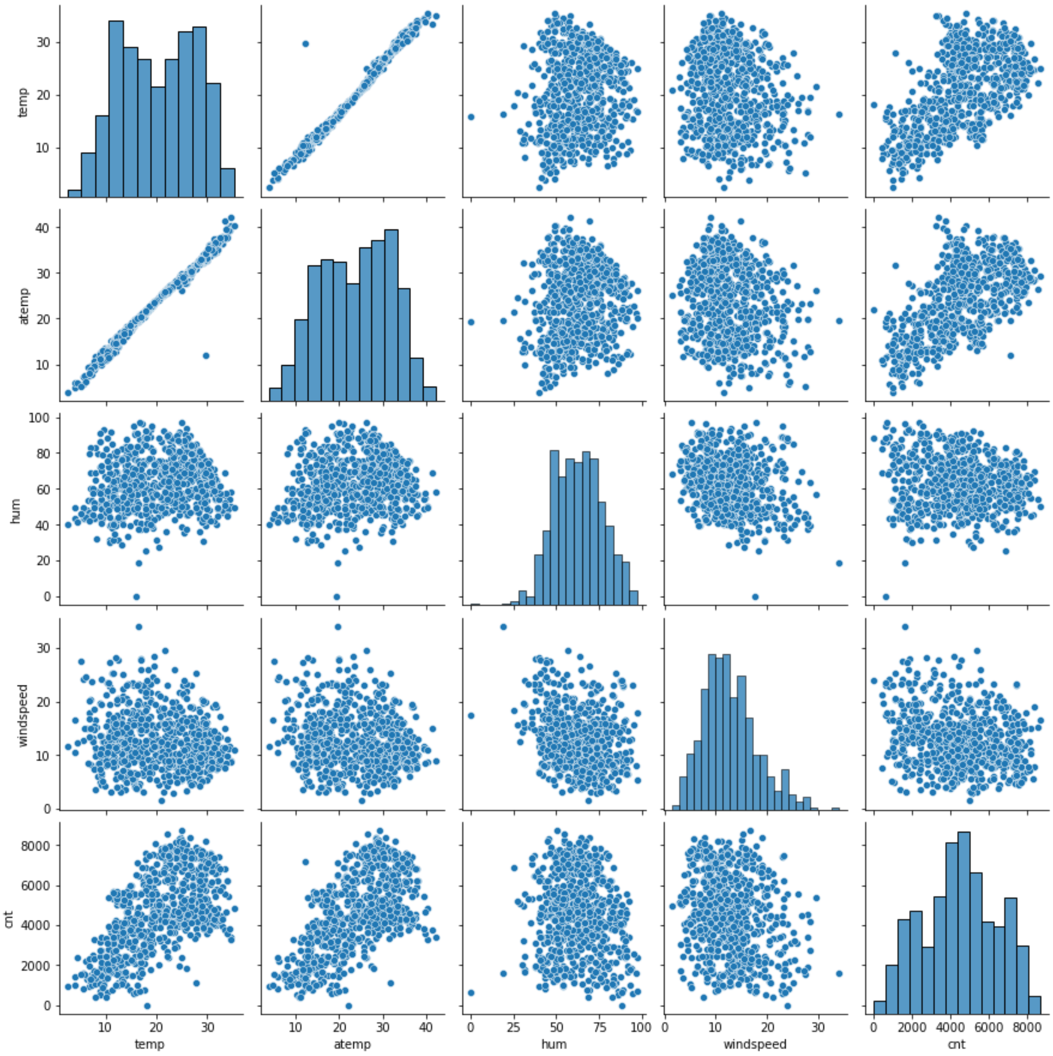
- Season: Spring (1) seems to clearly have very low ridership compared all other seasons. Fall (3) has the highest ridership
- Yr: 2019 had in general much higher ridership then 2018
- mnth: january has the lowest ridership and then it keeps gradually increasing every month until May from which it remains elevated until October after which it again tapers off
- holiday: mean ridership is slightly less on holidays but in general not much difference
- weekday: not much differentiation in ridership can be seen for different weekdays
- workingday: not much differentiation can be seen
- weathersit: for level 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) lowest ridership is seen. In general as weather situation becomes worse ridership falls, highest fall being in case of weather situation 3

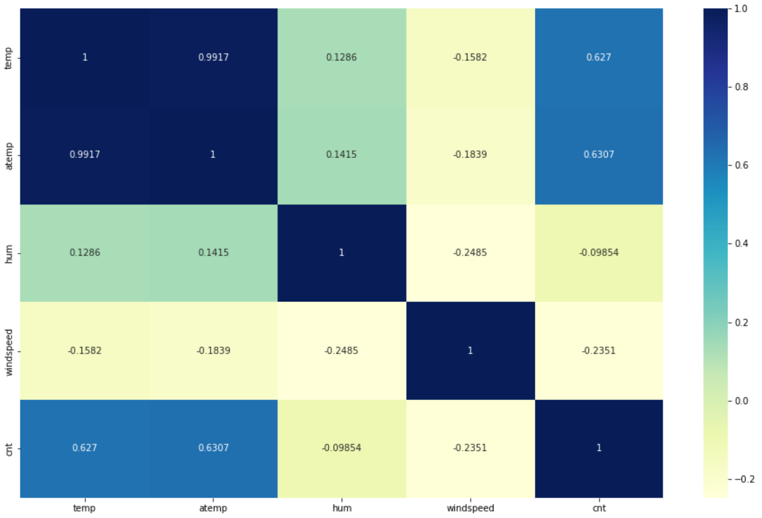
2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

   By default pandas.get_dummies will convert each categorical variable level into a separate dummy variable. So if a categorical variable has n levels n dummy variables will be created. However we only need n-1 dummy variables to fully represent n levels. This is achieved by using drop_first=True. Also if n dummy variables are used VIF would become infinity as a particular dummy variable could be fully determined by the rest of the dummy variables for that categorical variable.

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

Atemp (Feel temperature) has the highest correlation (0.6307) with target variable cnt which can be seen from the following pair-plot and correlation matrix.



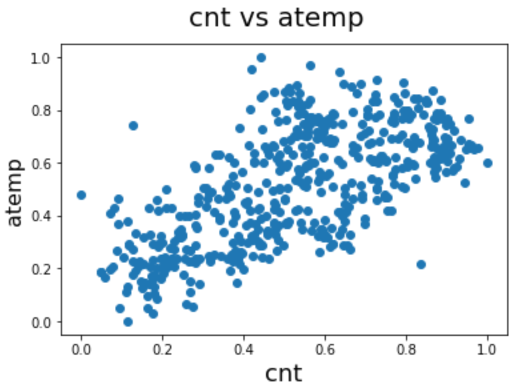


4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

the assumptions of simple linear regression are:

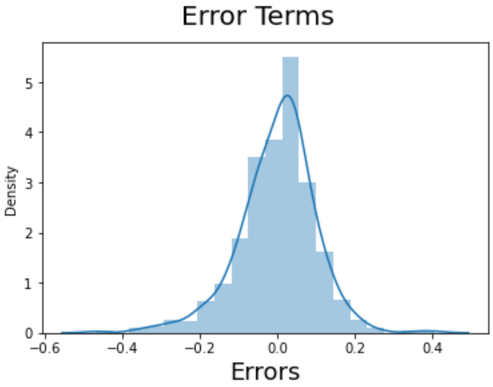
1. Linear relationship between X and Y

   As can be seen below atemp (X) has a linear relationship with cnt (Y). All other X variables are categorical.



2. Error terms are normally distributed

   For the final model residuals can be seen to be approximately normal

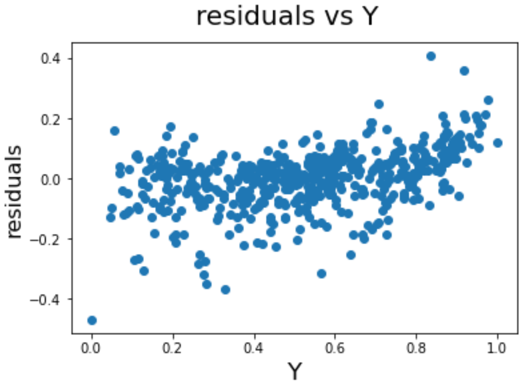


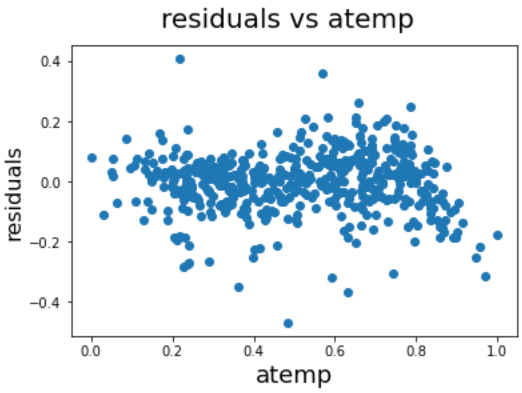
3. Error terms are independent of each other

   The Durbin Watson statistic of 2 implies no autocorrelation, the value of 2.135 implies only slight negative autocorrelation.

4. Error terms have constant variance (homoscedasticity)

   To check independence and constant variance for error terms / residuals we look at plot of residuals vs Y(cnt) and X (atemp)





It can be seen that the error terms appear to be evenly distributed noise around zero, though for higher X and Y a bit of downward and upward trend is seen. This implies constant variance and as no trends are observed error terms can be assumed to be independent.

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 regression equation is

   total ride count = 0.0318 + 0.631 \times atemp - 0.2632 \times weathersit_3 + 0.2348 \times year + 0.1371 \times season_4(winter) + 0.0971 \times mnth_9 + 0.0775 \times season_2 (summer) - 0.06869 \times weathersit_2

   Based on the scaled variables it can be seen that the top 3 features contributing significantly towards explaining the demand of the shared bikes are as follows

   1) Atemp (feel temperature)
   2) Weathersit_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
   3) Year

# General Subjective Questions

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

Linear Regression is a type of Supervised Learning algorithm where we try to predict the values of a dependent variable (Y) using independent / predictor variables (X). If there is only a single predictor variable it is called simple linear regression. In case of more than one predictor variables it is called Multiple Linear Regression.

In regression we are trying to fit the X and Y data to the following equation.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_3 + \ldots + \beta_k X_k$$

In case of simple linear regression the equation would represent a line while in case of multiple linear regression it would be a hyperplane.

The fitting can be performed by using the Ordinary Least Squares Method (OLS). In this approach the parameters (Betas) are estimated such that the sum of residuals (difference of actual Y and predicted Y using regression equation) is zero. The sum of residuals is denoted by RSS (Residual Sum of Squares). OLS method has a formula which uses X and Y values to determine the model parameters (betas) such that the RSS is zero.

The strength of linear regression can be assessed using $R^2$ (Coefficient of Determination). It's a number lying between 0 and 1 which signifies what proportion of variance in Y is explained by X. A value of 1 would mean 100% of the variance in Y is explained by the X variables used in regression.

For multiple linear regression adjusted $R^2$ is used as it penalizes for number of variables. This is because $R^2$ will always increase or remain same when a new variable is added to the regression even if does not have any explanatory power.

We only use the sample population of X and Y to fit the regression line. The assumption is that the sample regression line would be a good proxy of the actual population regression line. If this assumption is indeed true this will allow us to use the fitted linear regression equation to predict Y using X for unknown Y values.

To check if the Betas are statistically significant t test is used. The Null hypothesis being that the Beta is equal to zero. We generally use a p value of 0.05 to determine significance of the estimated betas.

F statistic is used to check if the overall model fit is significant or not.

To be able to use Linear regression there must be a linear relationship between X and Y. Additionally the residuals must meet the following criterion.

1. Residuals are normally distributed

2. Residuals are independent of each other

3. Residuals have constant variance

We could face the issue of Overfitting. Here the fitted model (sample regression line) is not a good proxy for the population regression line as it has sort of memorized the sample distribution instead of generalizing. One way to check for this is when $R^2$ on test data set is way lower than the $R^2$ on the train data set. Test data set comprises of independent data that is not used to train/fit the regression equation.

Another issue faced is Multicollinearity. This happens when some of the X's (predictor variables) might themselves be explained well by other predictor variables. While multicollinearity does not effect prediction accuracy it does make the coefficients and p values less reliable and will cause an issue with model interpretability. VIF is a measure used to check for multicollinearity. Pairwise correlations can also be used but it is only useful where relationship is one to one. VIF is a more reliable measure. Solution is to drop variables with high VIF after understanding which variable is causing the issue. VIF has been discussed in detail below.

Feature Scaling as discussed in detail below can be used to improve the interpretation of the regression equation.

Categorical variables need to be converted to dummy variables so that they can be included in the linear regression equation. For a categorical variable with n levels n-1 dummy variables should be created.

For selecting independent variables to be used in a regression a mix of manual and automated approaches like RFE need to be used. The general goal is to arrive at a regression equation in which all independent variables are significant at 0.05 p value, VIF is below 5, F statistic is significant, and the 4 Linear Regression assumptions mentioned above are satisfied. The regression equation is fitted on train data and then tested on unseen test data to rule out overfitting.

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

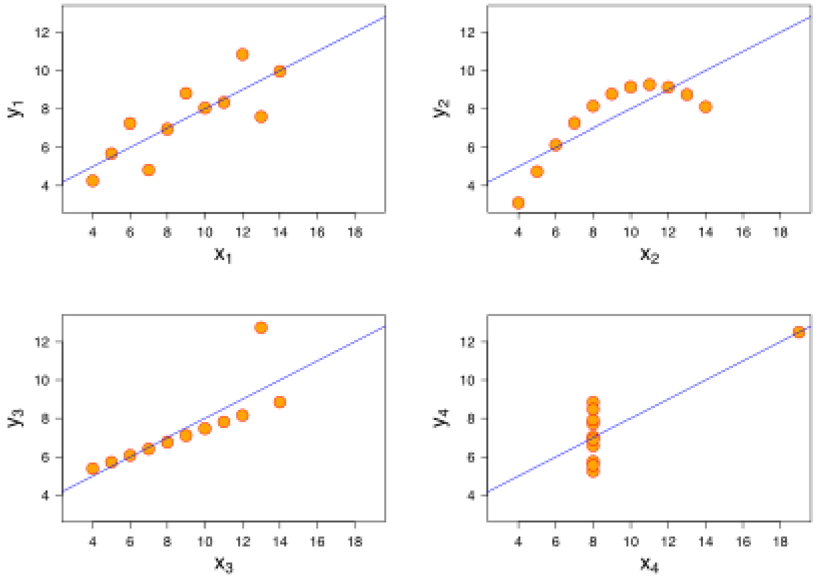
Anscombe’s quartet refers to the following 4 datasets with the same summary statistics but totally different  relationships between x and y. These were created by the statistician Francis Anscombe to demonstrate the importance of data visualization and also to show how outliers can impact statistical properties. These 4 datasets display how looking at standalone summary statistics can be misleading. Based on just the summary statistics which is exactly same for the 4 datasets it would seem that the same linear regression line is a good fit for all 4 datasets even though they actually display different relationships from each other.

| **Anscombe's quartet** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **I** | | **II** | | **III** | | **IV** | |
| *x* | *y* | *x* | *y* | *x* | *y* | *x* | *y* |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

**Summary Statistics**

| Property | Value |
|---|---|
| Mean of x | 9 |
| Sample variance of x | 11 |
| Mean of y | 7.50 |
| Sample variance of y | 4.125 |
| Correlation between x and y | 0.816 |
| Fitted Linear Regression line | y = 3.00 + 0.500x |
| R Square | 0.67 |

**ScatterPlot of X & Y for the 4 datasets**



*Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet*

It can be seen

1. For x1 and y1 it can be seen from the scatter plot that they indeed have a linear relationship and are being modelled correctly.
2. For x2 and y2 it can be seen from the scatter plot that x and y have a non linear relationship and hence linear regression is not appropriate for modelling even though correlation is high.
3. For x3 and y3 it can be seen from the scatter plot that while x and y have in general have a perfect linear relationship the regression line has been tilted higher than it should be due to

one outlier which has pulled the regression line up. Further if not for the one outlier the correlation would have been 100% for x3 and y3.

4. For x4 and y4 it can be seen that only due to 1 point the correlation coefficient value is high even though for all but one y the x = 8

**3. What is Pearson's R? (3 marks)**

Pearson's R is a method to calculate the correlation between two variables. Correlation is one metric to quantify the strength of the relationship between two variables. But as was seen above in the case of Anscombe's quartet it can be some times misleading. All the 4 data sets above had the same correlation of 0.816 though the relationship for each was very different.

It is calculated as follows for two variables x and y.

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}}$$

$r$ = correlation coefficient

$x_i$ = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

$y_i$ = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Pearson's correlation coefficient is always a value between -1 and 1. The sign signifies the direction of relationship. Positve sign means when one increases the other also increases and vice versa. Negative sign means when one decreases the other increases and vice versa. The absolute value represents the strength of the relationship, for example 1 and -1 would represent perfect positive and negative relationship.

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

Different variables will have different range of values making comparison difficult. For example a regression equation might have a distance variable in km and a temperature variable in Celsius. The coefficients of the unscaled variables in a regression equation would not be easily comparable. In Scaling we change the range of the different variables to allow for comparison between them. If scaled variables are used in regression than the regression coefficients can be directly compared. The magnitude of the scaled coefficients will signify the strength of the relationship between the independent and the dependent variable. Scaling also helps in faster convergence of the gradient descent algorithm which can be used to estimate the model parameters for linear regression and other ML algorithms.

In normalized scaling or Min Max Scaling the values are modified such that they are in between 0 and 1. Max value becomes 1 and min value becomes 0. The formula for that is as follows.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

In standardized scaling the values are adjusted such that the mean of the distribution is made 0 and standard deviation is made 1. This is done using the following formula.

$$x = \frac{x - mean(x)}{sd(x)}$$

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

The formula for VIF is given by

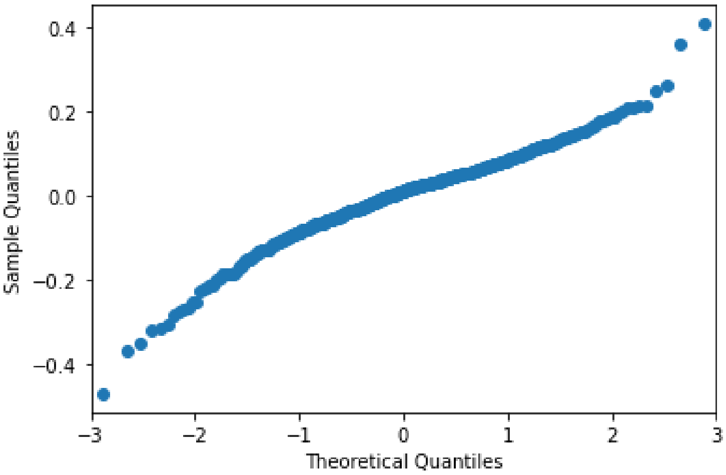
$$\mathrm{VIF}_\mathrm{i} = \frac{1}{1 - R_i^2}$$

It can be seen that if R=1 then VIF becomes infinite. This would happen when one of the predictors can be completely explained by another set of predictors, basically they have a deterministic relationship. An example of this is if we create n dummy variables instead of n-1 for a categorical variable with n levels. In this case the value for any dummy variable could be determined by the remaining n-1 dummy variables leading to R=1 and VIF = infinity.

**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Q-Q (quantile-quantile) plot is a X-Y line plot which can be used to compare any 2 distributions by plotting the different quantiles of the 2 distribution against each other. Each axis will have the quantiles for the respective distribution. If the two distributions are exactly same then the points on the Q-Q plot will lie on a straight line y = x. The interpretation is that closer the points are to the y=x line closer the 2 distributions are.

It is generally used to compare the sample distribution against a theoretical distribution like in case of Linear Regression we can use QQ plot to see how close the residual distribution is to a normal distribution. Normality of the residuals is an assumption of Linear Regression. The following is a QQ plot used by me to check the assumption of residual normality.


0.4
0.2
0.0
−0.2
−0.4
Sample Quantiles
−3
−2
−1
0
1
2
3
Theoretical Quantiles