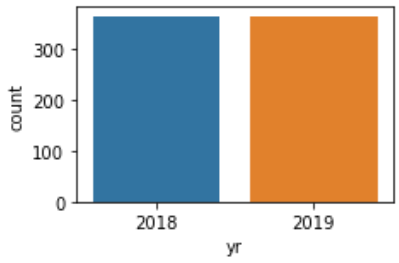
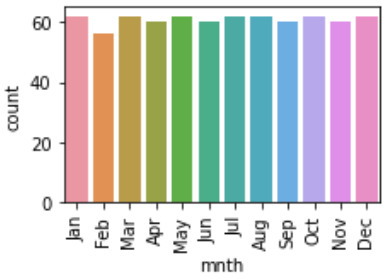
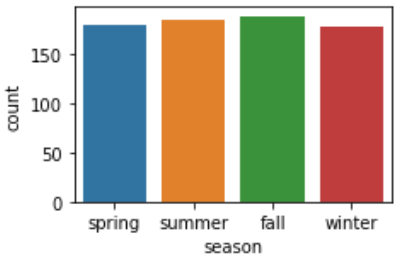
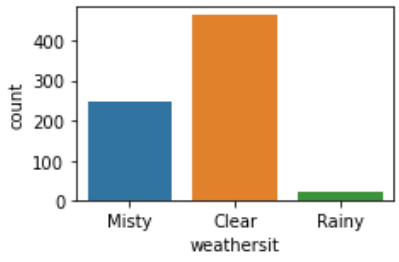
Assignment-based Subjective Questions

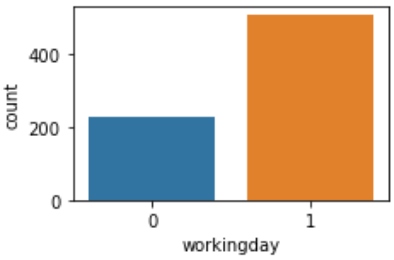
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

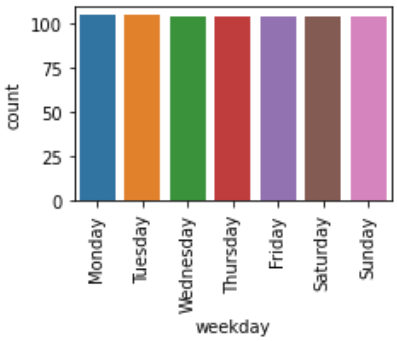


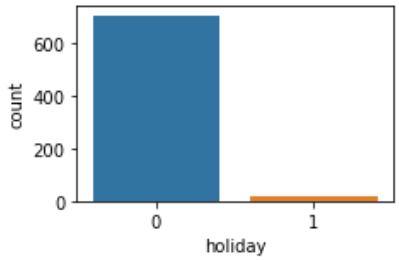


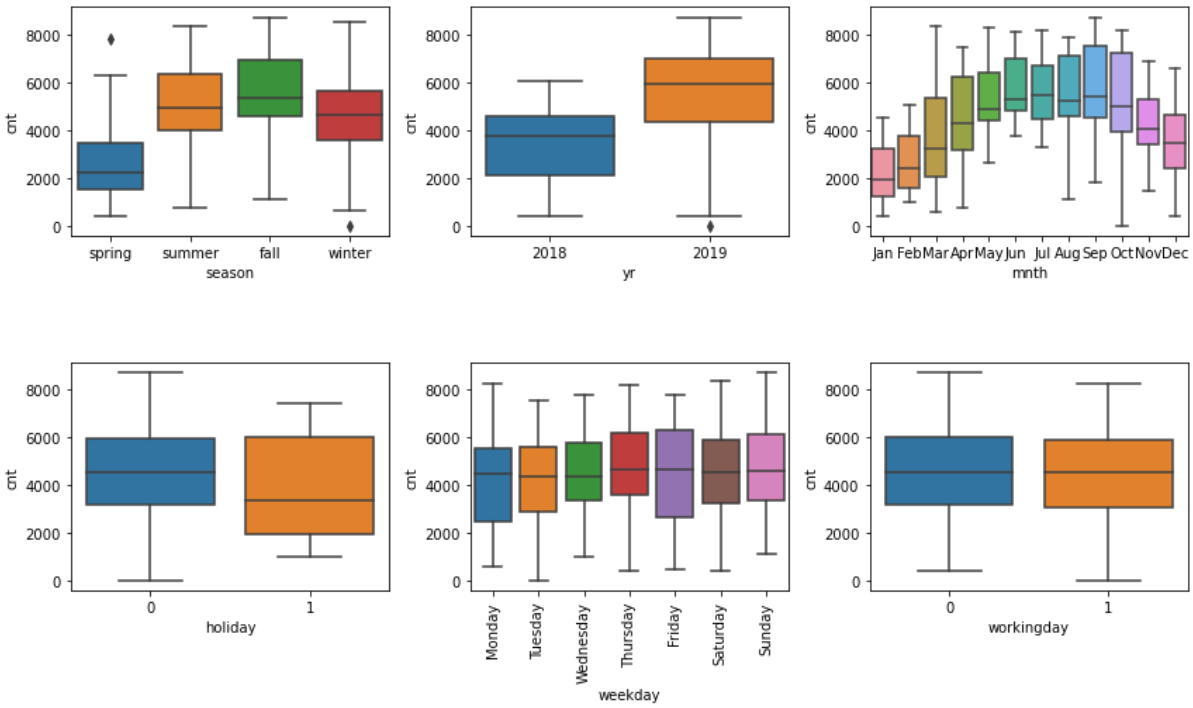












\\\

Inferences from the analysis of categorical variables:

1. From the first plot of the seasons, we can see there are outliers for the Spring and Winter seasons. Also, the most number of users are in Fall while surprisingly the least are Spring as the median is relatively lower compared to Winter which would likely be a less popular season for bike sharing.

2. There are higher numbers of users in 2019 compared to 2018 with a significantly higher median and wider distribution. Possibly due to popularity perhaps from marketing efforts,etc.

3. For the months boxplot, similar to the Seasons it follows a pattern of increase in fall and summer and spring have the least users.

4. Clear days have more users and wider distribution, with Rainy days having the least.

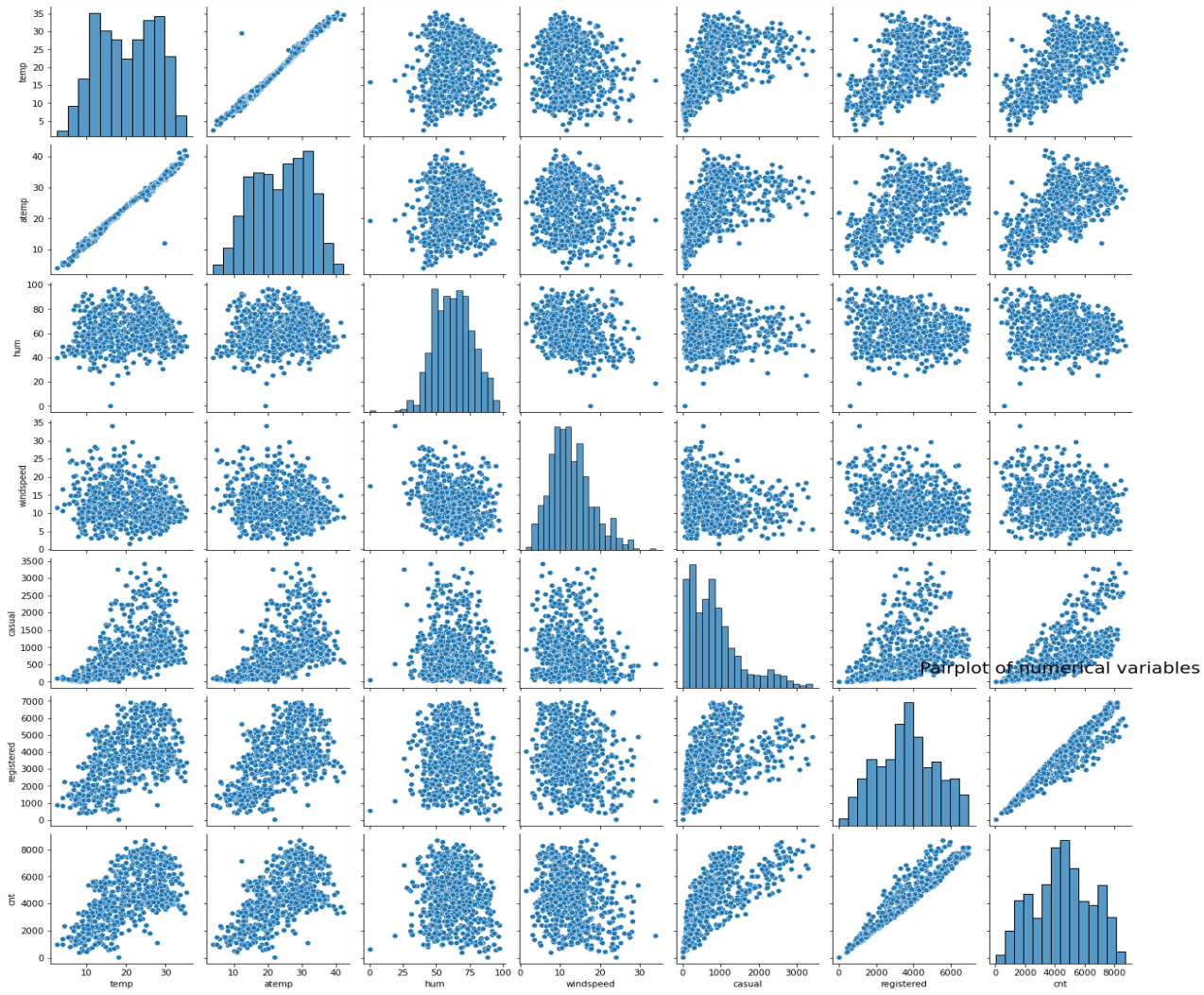
5. While the boxplots for working days have similar ranges, we can tell from the countplot that days that are not the weekend or a holiday (working day) have the most users.

6. The countplot for holiday shows that most users do not use the shared bikes on holidays, implying that perhaps a significant segment of users might be using the service on the way to work or such.

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

It is important to drop_first=True when creating dummy variables because the first column is an extra column which is correlated with the other created columns.
Dropping the first column will reduce the correlation among the dummy variables.

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

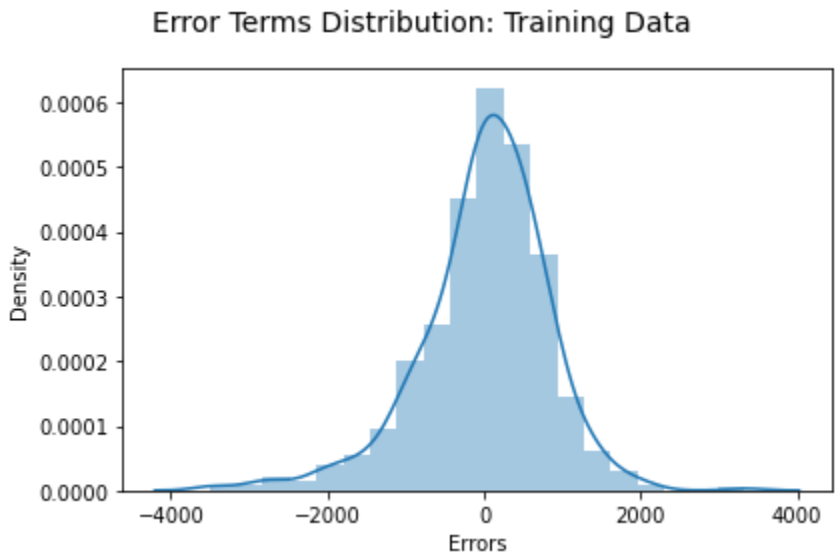


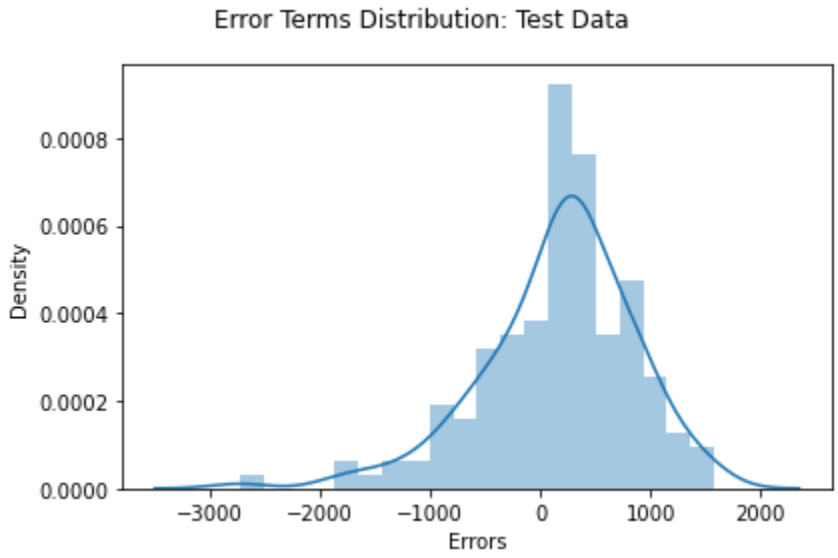
From the pairplot, we can see that ‘temp’(temperature in celsius) and ‘atemp’(feeling temperature in celsius) has the highest correlation among all variables. Followed by, ‘registered’ and ‘cnt’.

Indicating that one of the highly correlated columns should be dropped to avoid multicollinearity.

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

We validate the assumptions of Linear Regression during Residual Analysis where we check if the error terms are normally distributed. If the residuals are normally distributed, then the assumption of Linear Regression is valid.





We can see the residuals on both the training and test data indicate a normally distribution. Hence, validating our assumption of Linear Regression i.e. there is a linear relationship between X and y.

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

cnt =970.5 * temp - 213.98 * hum - 228.16 * windspeed - 259.29 * season_spring + 247.22 * season_summer + 393.50 * season_winter + 984.75 * yr_2019 + 221.16 * mnth_sep - 187.78 * weathersit_Misty - 344.11 * weathersit_Rainy

This is our equation for the best fit line. Based on our finale model, the top 3 features contributing significantly to the demand of the shared bikes:

1. Temperature
2. Year
3. Rainy

These three features increase by the coefficient (as mentioned in the equation) for each unit of our dependent variable(cnt).

Firstly, the cnt increases based on the temperature (i.e. warm weather is ideal for users to ride bikes).
Secondly, number of people has increased considerably from 2018 to 2019 and the coefficient indicates that as the year increases the number of users is to increase as well.
Thirdly, the negative coeffcieint for rainy weather days indicates that the number of users decrease on rainy days. The dataset has more entries of users in the Winter season and hence the coefficient indicates an increase in number of users in winter. This requires further investigation to determine the accuracy of the potential positive impact of winter (because winter months should generally have fewer users due to the difficult conditions for bike usage).

General Subjective Questions

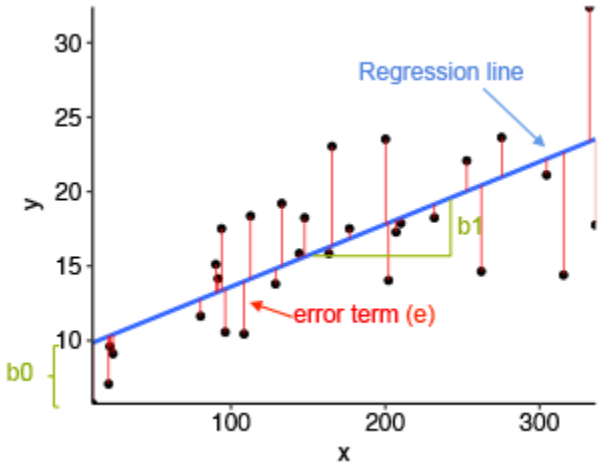
1. Explain the linear regression algorithm in detail.

Regression is a statistical method that falls under the Supervised Learning method, used to determine the relationship between two or more variables.

If two variables are linearly correlated, the correlation coefficient will lie between -1 and 1. We use linear regression in such cases, to predict the value of the dependent variable based on the other variables(s).

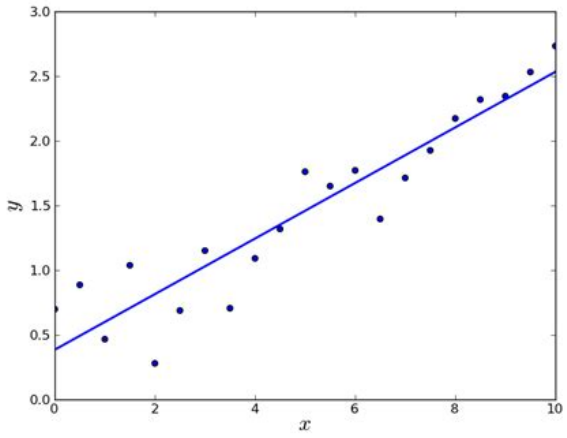
The correlation between two variables can be determined through a scatterplot, and if a linear relationship exists between both variables we can fit a line. Linear regression searches through the data to find the best fit line and reduces the total error of the model. The best fit line would the one where error terms are at a minimum (Residual sum of errors is least).

Linear Regression is used to determine how a dependent variable y (target variable) changes as the independent variable X (predictor(s) variable) changes.



There are two types of Linear Regression:

1. Simple Linear Regression: is used to determine the relationship between two quantitative variables (one dependent and one independent variable) using a straight line.

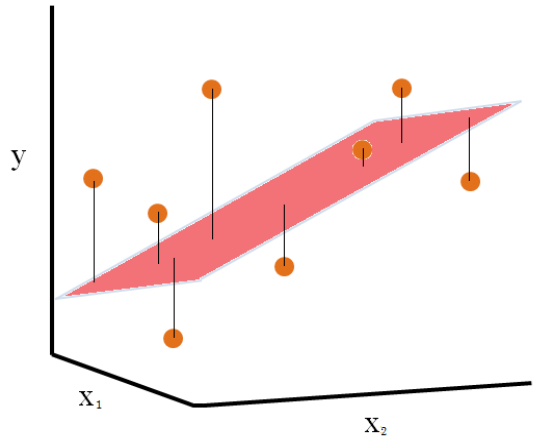


Formula:

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

- y is the predicted value of the dependent variable (y)
- B_0 is the y-intercept
- B_1 is the regression coefficient
- X is the independent variable
- e is the error of the estimate

2. Multiple Linear Regression: is used to determine the relationship between one independent variable (target variable) and two or more independent variables(predictor variables) using a straight line.



$$y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$$

- y is the predicted value of the dependent variable (y)
- B_0 is the y-intercept
- B_1X_1= the regression coefficient (B_1) of the first independent variable (X_1)
- B_nX_n = the regression coefficient of the last independent variable
- e is the model error

Linear Regression makes the following assumptions:

- There is a Linear Relationship between X and y
- The error terms are normally distributed
- The error terms have constant variance
- The error terms are independent of each other

2. Explain the Anscombe's quartet in detail.

Francis Anscombe created Anscombe's quartet to illustrate the need for graphing prior to analyzing data. Anscombe's quartet consists of 4 datasets with the same statistical properties (mean, variance, standard deviation, etc) but different graphical representations.

The following picture is the descriptive statistics for the four datasets:

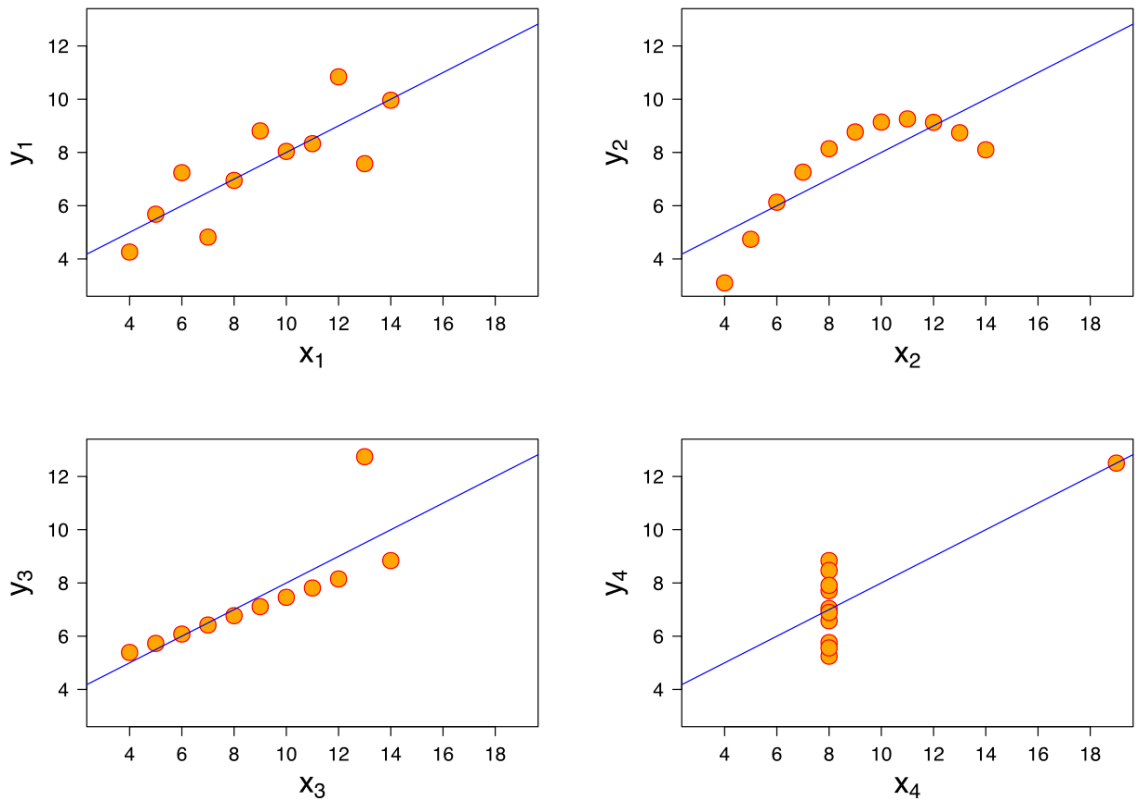
```
_____________________Mean of x_____________________
x1 : 9.000      x2 : 9.000      x3 : 9.000      x4 : 9.000

_____________________Mean of y_____________________
y1 : 7.501      y2 : 7.501      y3 : 7.500      y4 : 7.501

____________________Variance of x____________________
x1 : 11.000     x2 : 11.000     x3 : 11.000     x4 : 11.000

____________________Variance of y____________________
y1 : 4.127      y2 : 4.128      y3 : 4.123      y4 : 4.123

__________________Correlation of x & y__________________
x1/y1 : 0.816   x2/y2 : 0.816   x3/y3 : 0.816   x4/y4 : 0.817
```



Graph 1 : the data points represent a linear relationship

Graph 2 : the data points form a curve, which is not a linear relationship

Graph 3 : the data points indicate a tight linear relationship with one outlier

Graph 4 : the data points are gathered on one x value with a single outlier

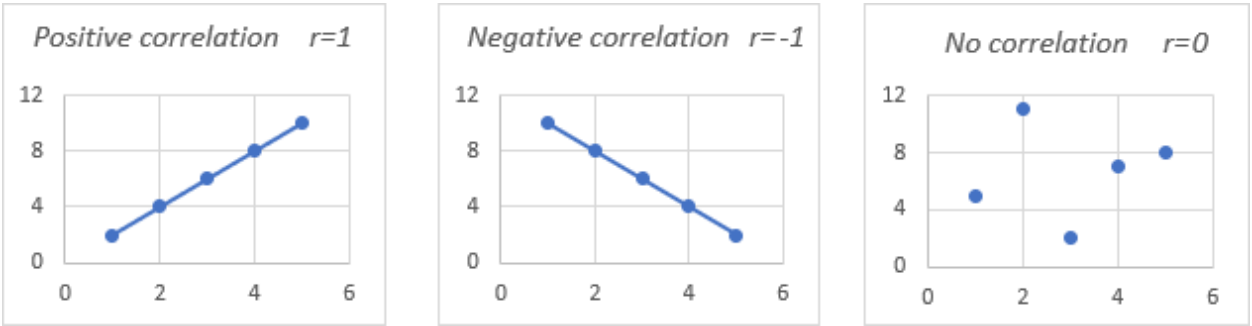
In addition to illustrating the importance of graphing, Anscombe's quartet explains the effect of outliers on data when graphed. The bottom two graphs above have an outlier each which is evident from the graph but cannot be observed from the descriptive statistics which are identical. Without these outliers, both graphs the descriptive statistics would be different.

Hence, Anscombe's quartet depicts the importance of graphical representation of data and the effect of outliers on datasets.

3. What is Pearson's R?

Pearson's R , also known as Pearson Correlation Coefficient, is used to measure the strength of a linear relationship between two variables.

The Pearson's correlation coefficient always lies between -1 and +1.



+1: Indicates a perfectly positive linear relationship
When the Pearson coefficient value is 1 or nearing 1, it implies that as the value of one variable x increases as the other variable y also increases.

-1: Indicates a perfectly negative linear relationship

When the coefficient is -1 or nearing 1, this means that one variable x increases as the other variable y decreases (or vice versa: y increases as x decreases).

0: This indicates that there is no linear relationship between both variables.

A Pearson coefficient value nearing 0 does not mean two variables have no relationship, as the Pearson coefficient only identifies the strength of a linear relationship between two continuous variables. It cannot identify non linear relationships or differentiate between dependent and independent variables.

Pearson's coefficient is the ratio of covariance of two variables to the product of their standard deviation, given by the following formula:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Here, X and Y are two variables and the sigma represents standard deviation.

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

Scaling is one of the steps in regression needed to normalize the independent variables within a range. In datasets, independent variables have differing ranges, magnitudes and units and without scaling the models could be biased towards the high values because some machine learning algorithms can be sensitive. To avoid this, we use feature scaling and bring variables within the same range to be treated as equal.

There are two types of scaling:

1. Normalization: also known as MinMax scaling, brings all the data within the range of 0 and 1.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where, X' is the normalized value, X is the actual value, Xmin is the minimum value in the feature (column) and Xmax is the maximum value in the feature.

2. Standardization: also known as z-score normalization, scales all features to have a 0 mean and 1 standard deviation. The values are centered around the mean with a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

Where, X' is the standardized value, mu is the mean of the feature and sigma is the standard deviation of the feature.

The difference between normalization and standardization is Standardization is useful when the data follows a normal distribution and normalization is good when the data does not follow a normal distribution. The normalized value is more sensitive to outliers (as seen in the formula: mean and standard deviation) and hence, standardization is more robust in cases of outliers as the data will not be affected by outliers when standardized.

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

Infinite VIF values indicates that a corresponding variable can be perfectly explained by the linear combination of other variables(which show a infinite VIF value as well).
When the VIF values shows 'inf' or is infinity, this means that there is a perfect correlation between two independent variables.

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

R squared is the coefficient of determination and lies between 0 and 1. It measures how much of the variation in a dependent variable can be explained by the independent variable.
When there's perfect correlation, one dependent variable can be completely explained by the independent variable. For perfect correlation, the R squared = 1and as per the formula the value of vif is infinity.
This value is used to indicate collinearity between variables, and if the vif is infinite, there is perfect multicollinearity and one of these variables needs to be dropped for linear regression.

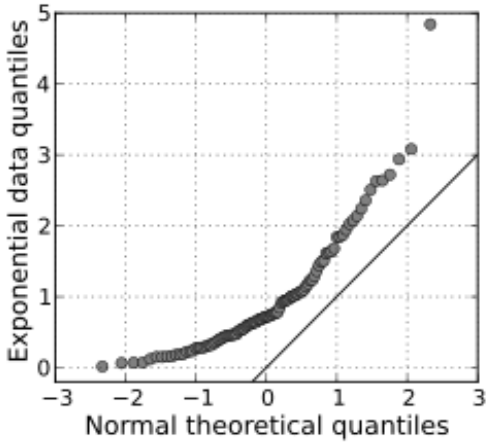
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q (Quantile-Quantile) plot is a graphical tool used to determine if two samples of data come from the same population. It is a scatterplot done by plotting the quantiles of the first dataset against the quantiles of the second dataset. A QQ plot can assess if a set of data plausibly came from a normal, exponential or uniform distribution. It can be useful in linear regression when training and test data are received separately and we can confirm if their datasets are from populations with the same distribution.

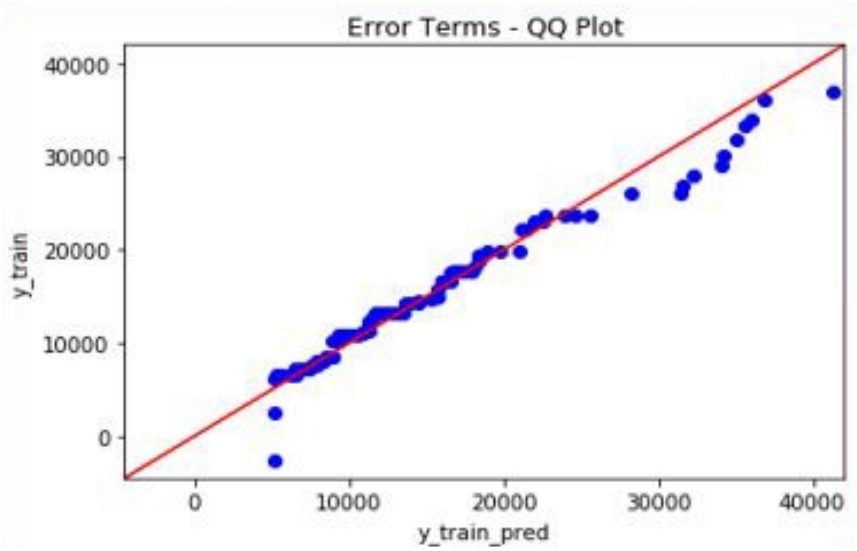
Advantages of QQ plots include that: 1. They can be used on sample sizes 2. It can detect shifts in locations, scales, changes in symmetry and the presence of outliers as well. 3. It can indicate If two datasets have the same distributional shapes and similar tail behaviour.

Interpretation of a QQ plot:

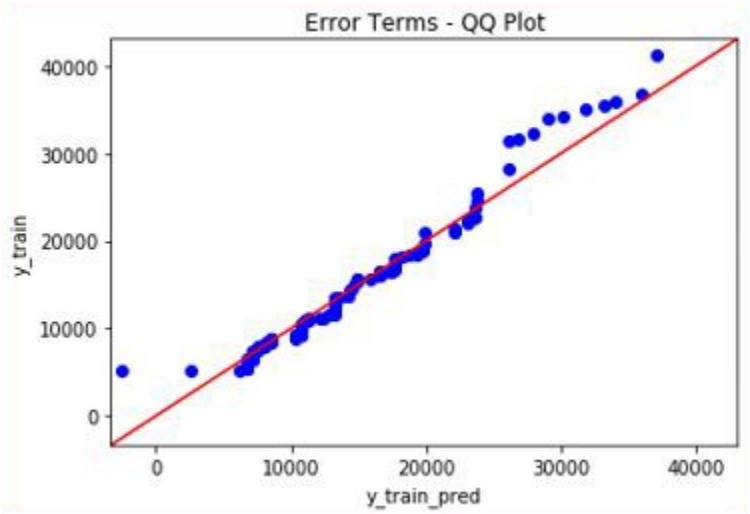
1. A similar distribution: if all quantile points lie on or close to the straight line at an angle of 45 degrees from the x axis.



2. Y values < X values: If Y quantile points are lower than X quantile points.



3. X values < Y values: If X quantile points are lower than Y quantile points.



4. Different distribution: if the quantile points lie away from the straight line at a 45 degree angle from the x axis.