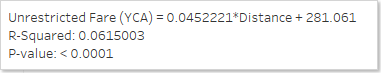
# Linear Regression



### Formula

* Simply stated, when we don’t travel any distance (i.e., X is 0) we pay $281.06 to the airline
* If we travel one mile:
* If we travel 3000 miles:

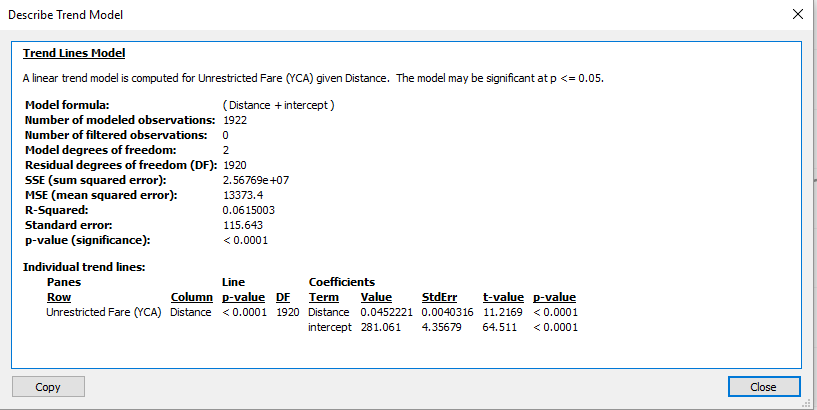
**Note**: In the formula, typically, you’ll see the dependent variable (what we’re trying to predict) represented as (or y-hat), this simply means that our results are an estimation of the value (i.e., the model is in no way 100% accurate for every value of X (or independent variable).

* Simply stated, the value tells us roughly how much of our data fell within the results of the line formed by the regression equation (closer to 1, we are achieving a perfect match, closer to 0, we have many data points that won’t agree with the model formula)
* 0.0615003 or 6.15% (roughly 6.15% of our data points can be explained well with the formula)
  + Is this good or bad? Why?

### p-value

* Simply stated, will a change in one variable (X) affect a change in the other (Y)
* If p-value > 0.05 then you can accept the null hypothesis (that there isn’t a relationship)
* Conversely, a p-value ≤ 0.05 then we can reject the null hypothesis and conclude there is an affect
* p-value = < 0.0001

**Go back to Tableau and click on the Trend Line, select Describe Model**



### Sum of Squared Error (SSE)

* Simply stated, it’s the actual values minus the predictive values, squared
* Squared because we want to make all values positive, and also emphasize large variations
* Our goal is for a low SSE which would indicate that actual and predicted values match up
* Our model SSE is 25,676,900

### Standard Error

* Absolute difference to the trend line, on average
* Our goal is for a low Standard Error
* Our model Standard Error is 115.643, for this model we can expect that there is an additional $115.64 variation in the model formula that should be attributed to error

Question: What is the general theme can we conclude from this model?

### Alternatives to Consider

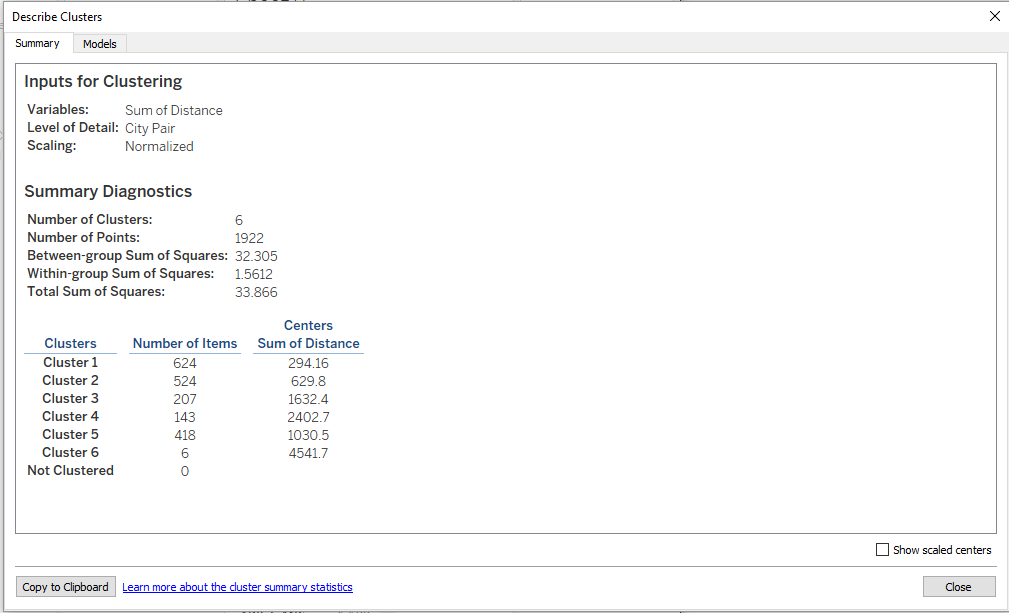
* Weakness – Simple linear regression only accounts for a single independent variable
* Weakness – Data may not be linear (e.g., what if it’s curvilinear), you may want to apply other type of model formulas to regression analysis
* In R or Python consider running multiple regression and review the affect of each coefficient’s p-value

# K-Means

### Algorithm

The general steps for k-means are:

* Randomly select cluster centers (given required number of clusters)
* Assign each instance to the nearest center
* Recalculate the new cluster centers
* Reassign each instance to the new closest cluster center (I believe Tableau is using the Euclidean [straight line] distance)
* The process stops either when no instances are reassigned to a different cluster or when the specified number of maximum iterations is reached
  + This is a good animated GIF regarding how the process works [https://en.wikipedia.org/wiki/K-means\_clustering#/media/File:K-means\_convergence.gif](https://en.wikipedia.org/wiki/K-means_clustering%23/media/File:K-means_convergence.gif)



### Between-group sum of squares

* A metric quantifying the separation between clusters as a sum of squared distances between each cluster’s center (average value), weighted by the number of data points assigned to the cluster, and the center of the data set
* The larger the value, the better the separation between clusters (i.e., the cleaner the cluster are segmented, limited overlap)

### Within-group sum of squares

* A metric quantifying the cohesion of clusters as a sum of squared distances between the center of each cluster and the individual marks in the cluster
* The smaller the value, the more cohesive the clusters (i.e., the tighter the cluster groups)

### Total sum of squares

* Totals the between-group sum of squares and the within-group sum of squares.
* The ratio (between-group sum of squares)/(total sum of squares) gives the proportion of variance explained by the model.
* Values are between 0 and 1; larger values typically indicate a better model. However, you can increase this ratio just by increasing the number of clusters, so it could be misleading if you compare a five-cluster model with a three-cluster model using just this value.

### Ideal K-clusters?

* Weakness -You must define how many clusters you want to include in your data
* In R or Python consider the Elbow Method, which plots the lift for each additional cluster in your dataset