DATA622 Homework #1

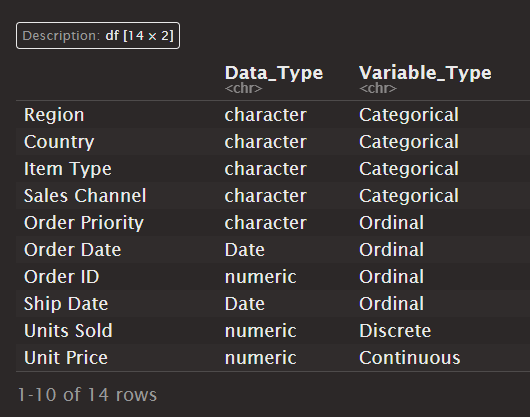
Daniel Craig

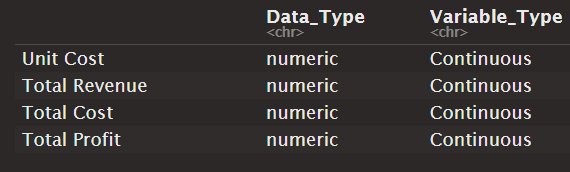
10/08/23

The two datasets chosen for this analysis were the 50,000 and 10,000 sales csv files from [here](https://excelbianalytics.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/) (https://excelbianalytics.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/). Both contain the below columns with column types ranging from categorical, string, date, and numeric.

**EDA**

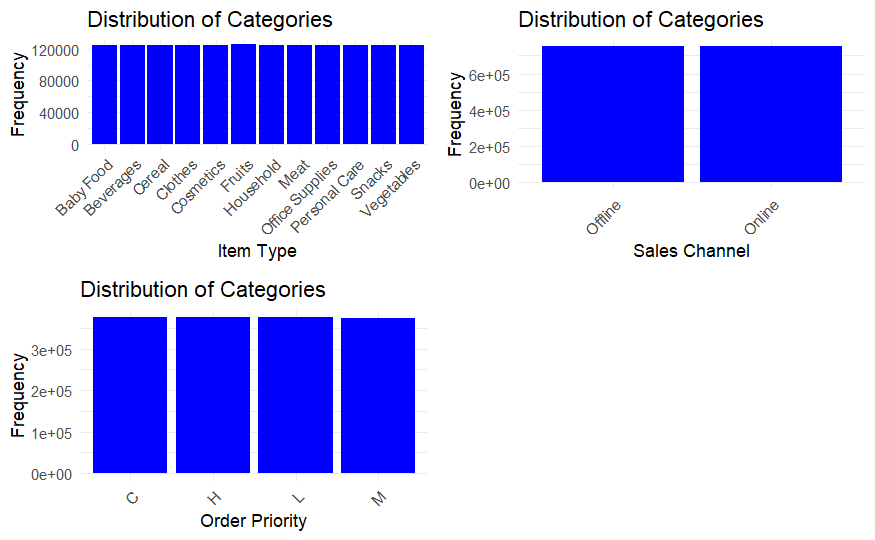
The data types, variable types and column names are as follows:

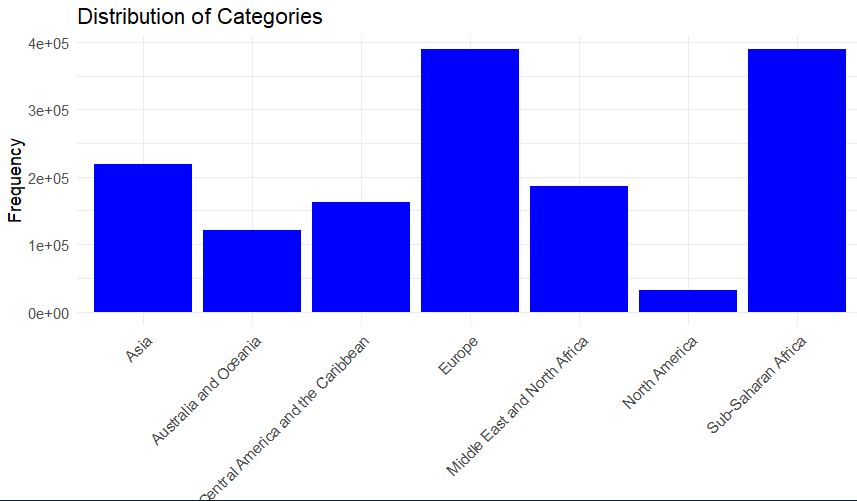




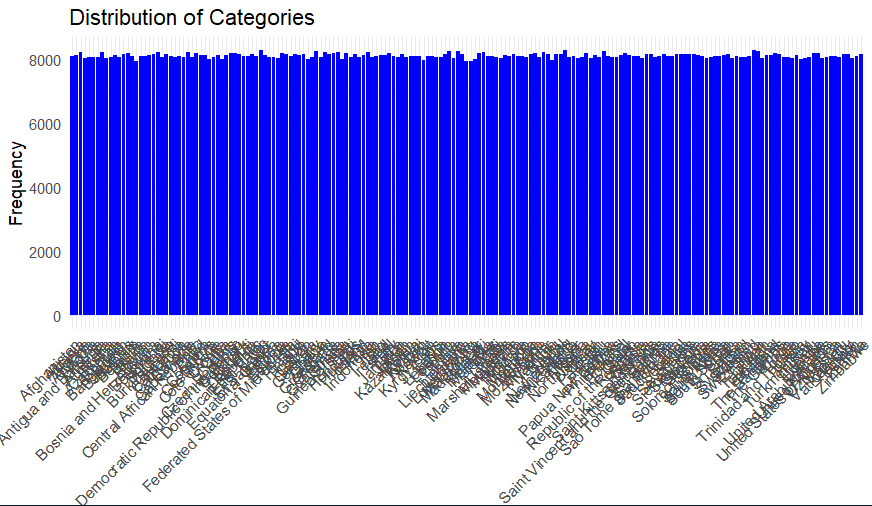
This data is a combination of categorical predictors based on geographic region and discrete/continuous predictors with pricing, cost, revenue and profit. The monetary variables will certainly be related due to the nature of cost, pricing, revenue, and profit but my interest will be more in a specific scenario that I’ve encountered in real life.

Before we get to that scenario, distributions of these variables are below.

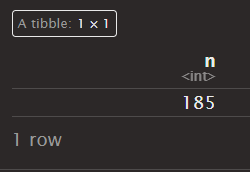
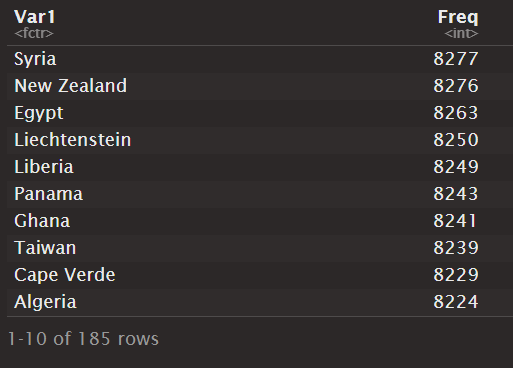


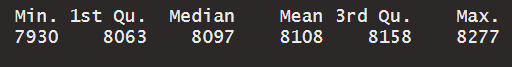


 The distribution of order numbers between regions are:

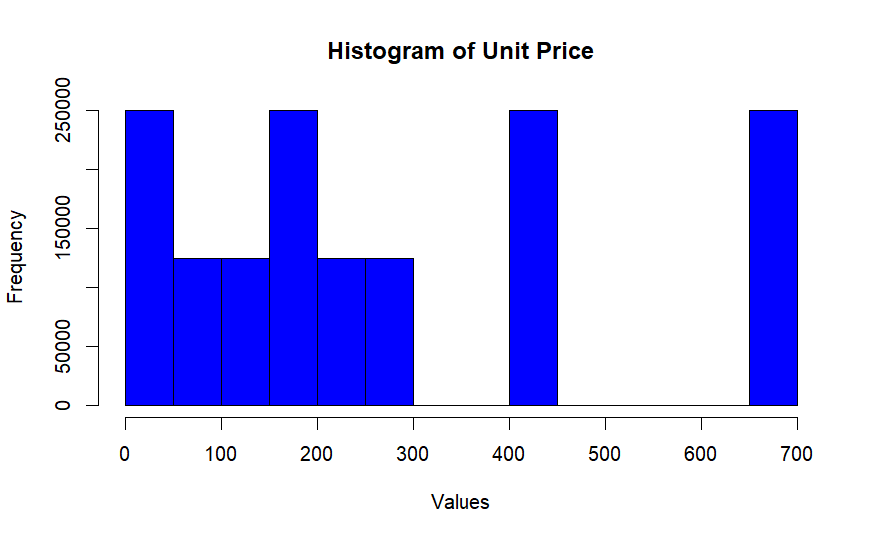


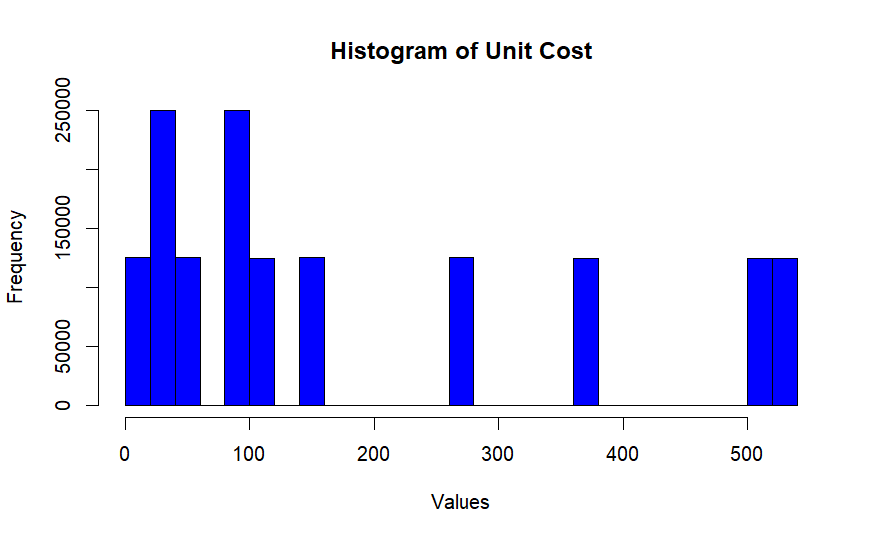
It seems that for everything outside of the geographic areas, the data is nearly evenyl distributed. To deal with the unreadable chart concerning countries above, I’ve also included a frequency table below to show the number of orders from each company. They are relatively tightly grouped with it being grouped between 7900 and 8300 orders from each country. Note that there were 185 different countries that submitted sales orders.

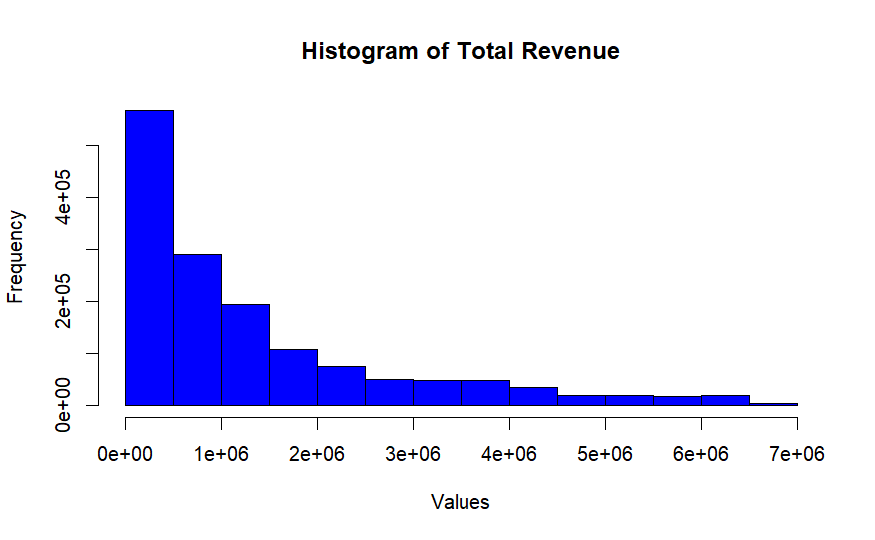


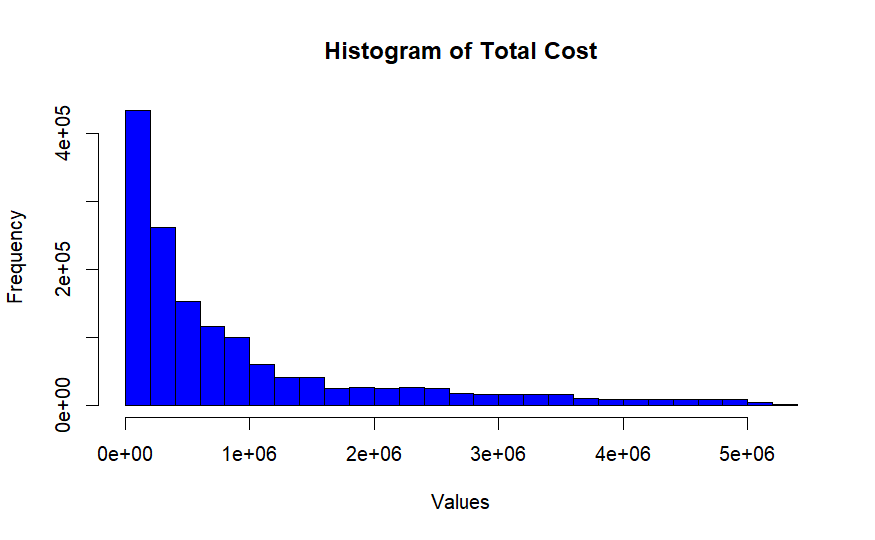


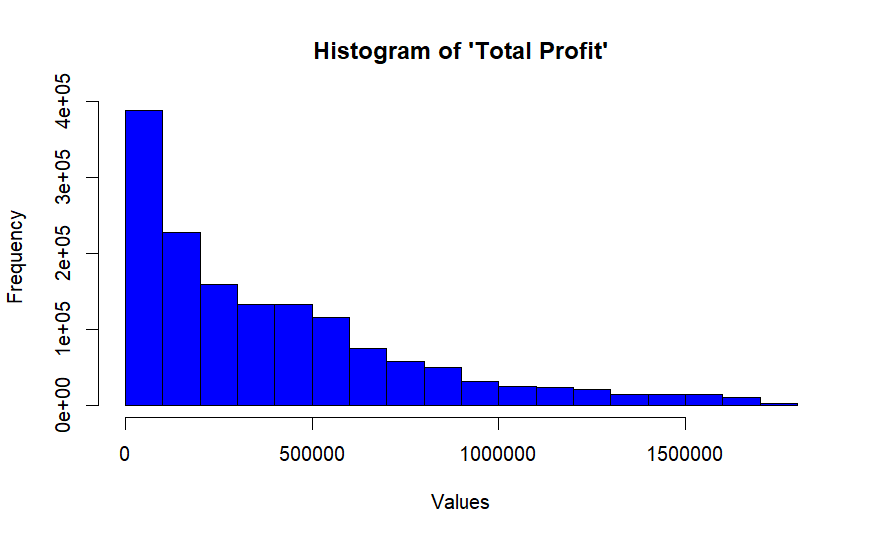
The discrete and continuous variables are bit more varied than their categorical counterparts.











It’s clear that there is a logarithmic presence in the revenue, cost, and profit variables that would most likely show a strong relationship between the three, unsurprisingly. There is certainly going to be a relationship between unit price, cost, revenue, and profit. What would be interesting to see if there are any relationships between geographical area and order numbers or item prices. Hopefully, we will see some of this appear in our models.

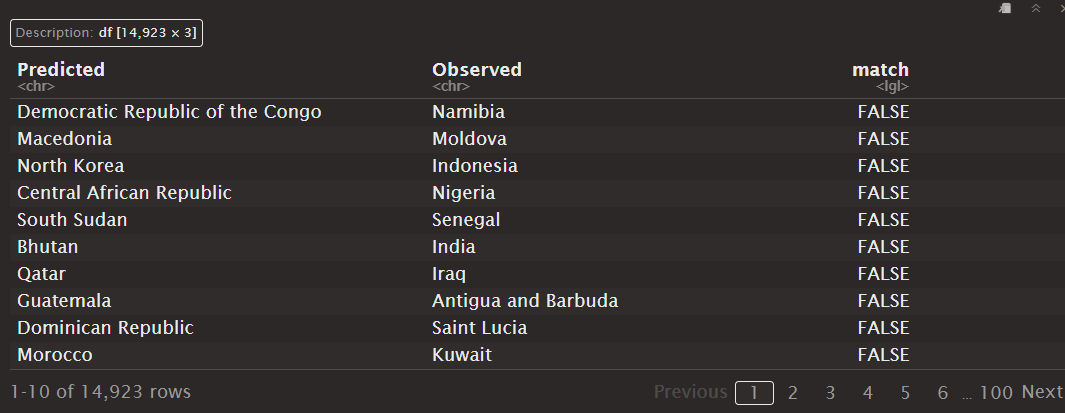
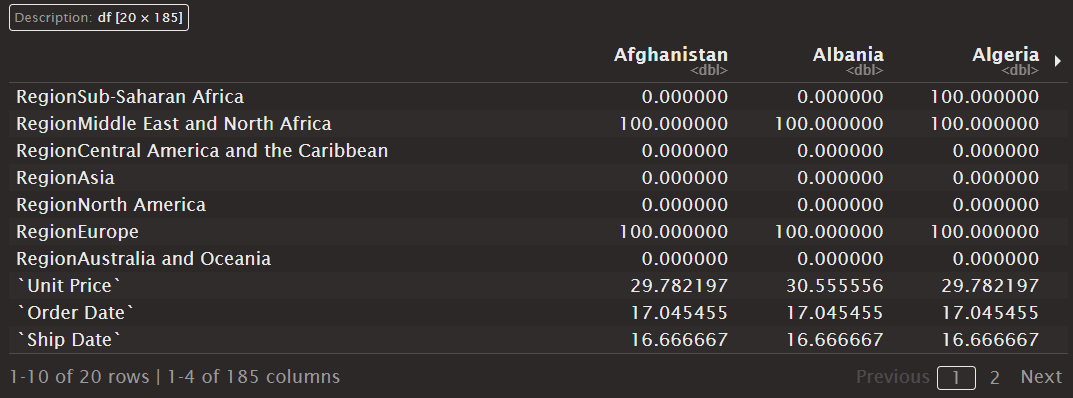
**KNN Model**

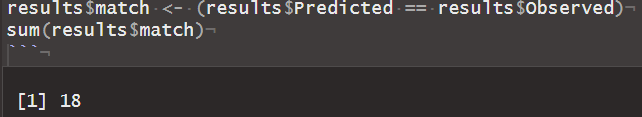
There are many methods to analyze the data depending on the use. In fact, I would argue the purpose for analysis is as much of a driver in technique choice as the data itself. For example, let's imagine a scenario where the company had lost portions of sales records leading to a blank customer ID column (there is not one in this dataset) or blank Country column.

To help narrow down potential values, this dataset was given to us a training set to create a prediction algorithm for the Country or Customer ID. This purpose would probably lend itself towards KNN, Decision Trees, or Discriminant Analysis. There are many classes that might not translate well to a purely linear modeling approach.

Let's try using kNN and Random Forests to predict what Region an order is for and see how they perform. Originally, I was going to attempt LDA/QDA instead of Random Forests, but since I run into an issue later with vector size errors due to the number of categorical variables being one shot encoded increasing the size of the data wildly, I chose to work with Random Forests. For the kNN model, I will use the 10,000 sample dataset.

Unfortunately, our modeling efforts to determine a Country based on the remaining columns were entirely in vain. The kNN model was only able to get 16 out of ~3,000 guesses correct. The first few countries variable importances are pasted below, as well as the guesses.

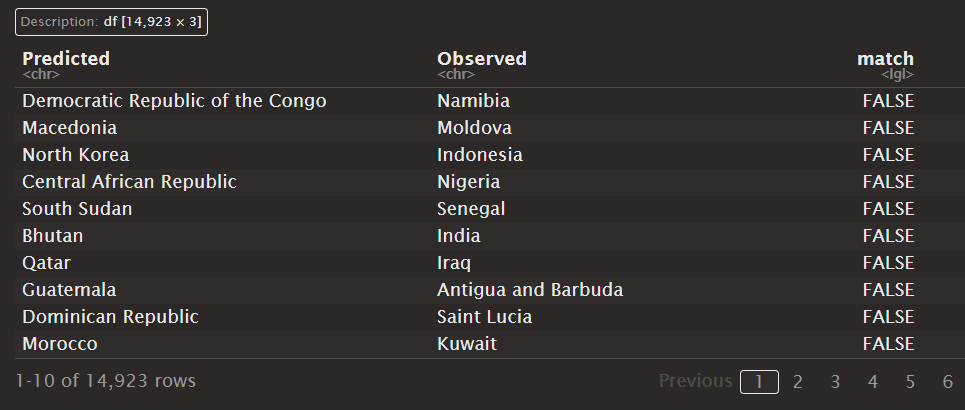


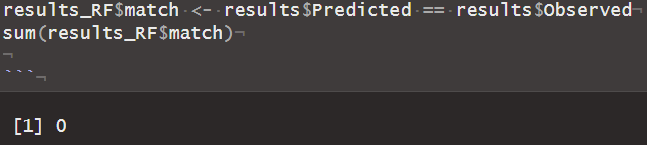


Only 18 correct predictions for KNN is quite poor and honestly a bit laughable. It makes me highly suspicious that I’ve either coded something incorrectly or missed an important fact about the data.

**Random Forest Model**

The random forest model performed even worse, not getting a single prediction correct. The random forest was trained on the 50,000 sample data set to see if maybe the sample size was causing issues with a model’s ability to predict. The random forest’s predictions are below:





With both models performing absolutely horrendously, I don’t think it’d be wise to make judgments off them. It is not a surprise that this was a difficult task, as it would’ve been much easier to go with attempting to predict the clearly correlated data in pricing and profit structure of the sales records, but nothing ventured nothing gained, and this would have been much more interesting had there been a promising relationship between Country and the rest of the information in the records like order priority or order revenue. These types of relationships could extend to interpret-able understanding such as a country having higher order priorities due to a national crisis, or order revenue due to an economic boon in the country that has lent to increased orders.

**Questions**  
  
**1. Are the columns of your data correlated?**  
  
Some of the data is clearly correlated. Order pricing, revenue, item price, and item stock all impact each other, as well as clearly the Region and Country columns. As far as correlation between these two groups, it seems unlikely as the models performed extremely poorly.  
  
  
**2. Are there labels in your data? Did that impact your choice of algorithm?**  
  
It does not seem like there are labels in the data either by manual inspection or by looking on the website. Hence it did not impact my choice in algorithms.  
  
**3. What are the pros and cons of each algorithm you selected?**  
  
Random Forests are great for increasing performance from regular forests, are less sensitive to near zero variance or overly powerful variables, and do not require heavy pre-processing. Typically, they generalize better than normal forests. For this increase in performance, they lose they're interpretability which is why I left its decision tree plot out of the analysis.   
  
KNN models are insensitive to outliers, are easy to understand, are well suited for multi-classing scenarios, and doesn't require specific distributions of data in terms of normality. These models do require some pre-processing, can take some computation time in very high dimensions, and different choices of 'k' (how many neighbors to include) can create drastically different answers. Luckily, models typically choose the 'k' that results in the least loss.  
  
**4. How your choice of algorithm relates to the datasets (was your choice of algorithm impacted by the datasets you chose)?**  
  
I specifically chose KNN due to how well it handled class distinctions in multi-class scenarios. I chose this one over LDA/QDA since I thought that the number of classes the sales orders could be attributed to were so numerous, LDA/QDA would not be able to model well.  
  
I chose Random Forests for the simplicity in pre-processing and easily understandable choices. Had it performed well, I would have looked into a normal regression based Decision Tree to see any variable importance or variable relationships.  
  
**5. Which result will you trust if you need to make a business decision?**  
  
Unfortunately both models performed awfully. I've actually never seen models perform this poorly and I would like to review this with someone else to see if I took a misstep somewhere. I could have chosen to predict total revenue or profit based on the variables, but that seemed like a problem easily solved by other methods than machine learning. I was more interested to see if item costs and prices had different relationships between countries.  
  
  
**6. Do you think an analysis could be prone to errors when using too much data, or when using the least amount possible?**  
  
I think that bigger datasets are no more error prone than smaller datasets as long as the coding framework stays relatively the same. If one dataset requires a distributed file system and the other requires loading a csv file, of course the larger dataset will be more error prone. In terms of making the wrong decisions, too much data can lead to over-specification, but that's easily managed by partitioning data.  
  
**7. How does the analysis between data sets compare?**  
  
Honestly, both are clear as mud. Neither model showed anything of interest. The only points of interest would be the logarithmic nature of the order pricing variables in their histograms and the geographic distribution of Europe and Sub-Saharan Africa dominating the number of orders.