Territory Distributions

Keith Hickman

October 26, 2017

# Intro:

This analysis presents several options for geographic distribution of territories using Revenue, Number of Accounts, and Types of Account by Industry. The data is loaded via three csv files.

See Domo cards for Geographic Distribution.

Begin with loading required packages:

Summary of our first dataset:

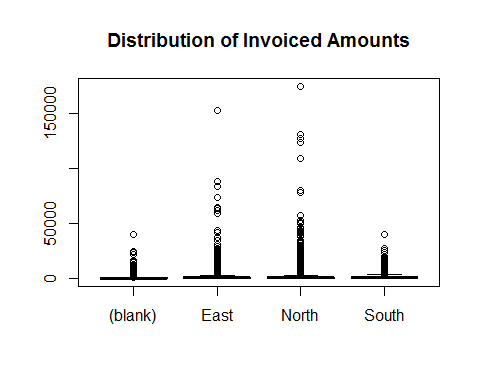
summary(territorydist)

## territorycode netivcamt standardized   
## Length:24587 Min. : 0.0 Min. :-0.31817   
## Class :character 1st Qu.: 137.7 1st Qu.:-0.27949   
## Mode :character Median : 391.1 Median :-0.20833   
## Mean : 1132.9 Mean : 0.00000   
## 3rd Qu.: 1028.2 3rd Qu.:-0.02938   
## Max. :174225.4 Max. :48.61492   
## NA's :1 NA's :3

# Revenue Analysis:

We have 24587 observations of three variables as mentioned above. The netivcamt variable is highly skewed (right-tailed) as shown by a mean and third quantile values much higher than the median, and the Max value several orders of magnitude higher than the 3rd quantile. Because the mean is sensitive to these higher values, we will use median and quantiles as the statistics of centrality.

territorydist <- na.omit(territorydist)  
boxplot(netivcamt ~ territorycode, territorydist, main="Distribution of Invoiced Amounts")



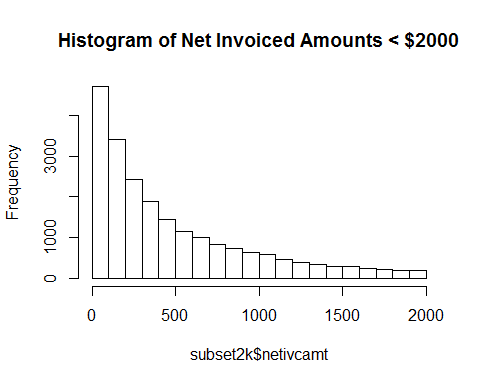
This box plot looks more like a bar chart, which reflect the outlier discussion above, but there's still some very interesting information here. The distribution of outliers (any values greater than 1.5x the Interquartile Range above or below ) clearly shows a much greater concentration of large invoiced amounts in the North and East Territories. The South territory looks more like out-of-territory than East or North.

**Dealing with outliers:** Since this is such a skewed distribution, filtering rows (invoices) above a certain threshold in order to analyze more data is preferable. Typically, we would define outliers as given above. In this case, where even normalizing numbers does not provide a suitable distribution, we can create two classes and analyze those separately. Consider that most of our values (transactions) fall between 0 and 2000 dollars, which will represent the breakpoint for our classes. This still leaves us with 21,476 out of ~24,000 observations in the class under 2k.

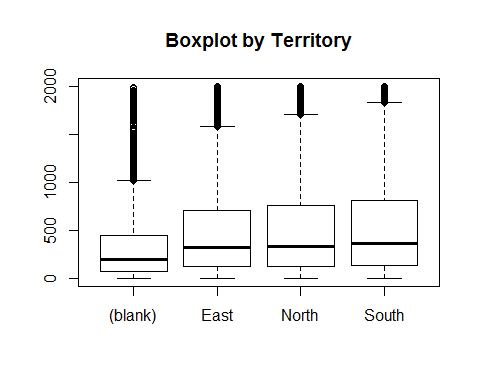
subset2k <- subset(territorydist, netivcamt<2000)  
subset2k

## # A tibble: 21,476 x 3  
## territorycode netivcamt standardized  
## <chr> <dbl> <dbl>  
## 1 East 0.00 -0.3181689  
## 2 East 1.00 -0.3178880  
## 3 East 1.64 -0.3177083  
## 4 East 1.77 -0.3176718  
## 5 East 2.04 -0.3175959  
## 6 East 2.10 -0.3175791  
## 7 East 2.16 -0.3175622  
## 8 East 2.40 -0.3174948  
## 9 East 2.61 -0.3174359  
## 10 East 2.67 -0.3174190  
## # ... with 21,466 more rows

hist(subset2k$netivcamt, main="Histogram of Net Invoiced Amounts < $2000")



boxplot(subset2k$netivcamt ~ territorycode,subset2k, main="Boxplot by Territory")

 The in-territory divisions look to be similarly distributed with respect to the , median, and values. The higher invoice amounts and overall number of transactions will likely explain the difference in the sums. The median of all in-territory values tends to be the same, which would indicate that most of the transactions that happen across all three territories is the same, around $250. Can the difference in total amounts can be explained by more opportunities and/or projects (over `$2k) in the North/East, and more regular, MRO-type orders (under $2k) in the South?

The out-of-territory invoices tend to be significantly smaller, with a under 500 USD.

**Total transactions**:

Time series data is also informative. Here, we'll examine monthly sales data from 1/1/2014 through 10/26/2017.

Read in the data:

territorytime <- read\_csv("C:/Users/khickman/Desktop/Personal/IUMSDS/AppliedDataMining/timeseries.csv")

## Parsed with column specification:  
## cols(  
## Year = col\_integer(),  
## Month = col\_integer(),  
## East = col\_double(),  
## North = col\_double(),  
## South = col\_double(),  
## OOT = col\_double()  
## )

summary(territorytime)

## Year Month East North   
## Min. :2014 Min. : 1.000 Min. : 464047 Min. :507030   
## 1st Qu.:2014 1st Qu.: 3.250 1st Qu.: 541310 1st Qu.:578918   
## Median :2015 Median : 6.000 Median : 605006 Median :650957   
## Mean :2015 Mean : 6.283 Mean : 625801 Mean :661857   
## 3rd Qu.:2016 3rd Qu.: 9.000 3rd Qu.: 670575 3rd Qu.:703714   
## Max. :2017 Max. :12.000 Max. :1047061 Max. :989420   
## South OOT   
## Min. : 82705 Min. : 52975   
## 1st Qu.:157676 1st Qu.: 80949   
## Median :195065 Median : 95181   
## Mean :196677 Mean :104256   
## 3rd Qu.:234719 3rd Qu.:113116   
## Max. :298438 Max. :253198

R treats Year and Month columns as integers, but they're esssentially categories here. To modify the column data types:

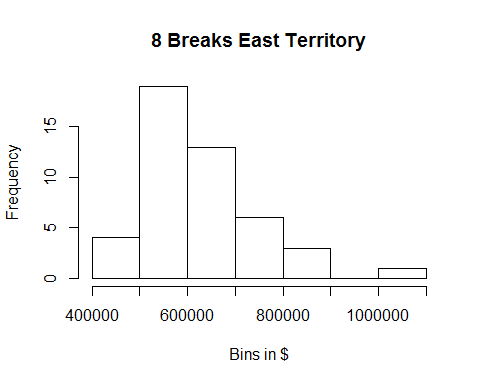
territorytime$Year <- as.factor(territorytime$Year)  
territorytime$Month <- as.factor(territorytime$Month)  
summary(territorytime)

## Year Month East North   
## 2014:12 1 : 4 Min. : 464047 Min. :507030   
## 2015:12 2 : 4 1st Qu.: 541310 1st Qu.:578918   
## 2016:12 3 : 4 Median : 605006 Median :650957   
## 2017:10 4 : 4 Mean : 625801 Mean :661857   
## 5 : 4 3rd Qu.: 670575 3rd Qu.:703714   
## 6 : 4 Max. :1047061 Max. :989420   
## (Other):22   
## South OOT   
## Min. : 82705 Min. : 52975   
## 1st Qu.:157676 1st Qu.: 80949   
## Median :195065 Median : 95181   
## Mean :196677 Mean :104256   
## 3rd Qu.:234719 3rd Qu.:113116   
## Max. :298438 Max. :253198   
##

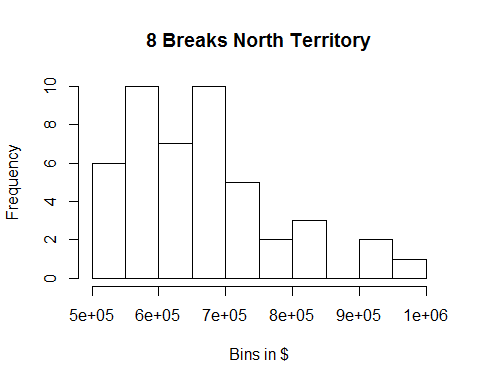
Examining each territory, we can see they are a bit closer to normal distributions, though the . Again the South is lagging in terms of monthly invoiced amounts, and trails the other two territories by a wide margin. Note the highest invoiced month for the South territory was 298,000, compared to ~990,000 and ~1,050,000 for both the North and East respectively. This

There are several ways to examine the data. Let's look at each territory's distribution of sales months. The bins represent

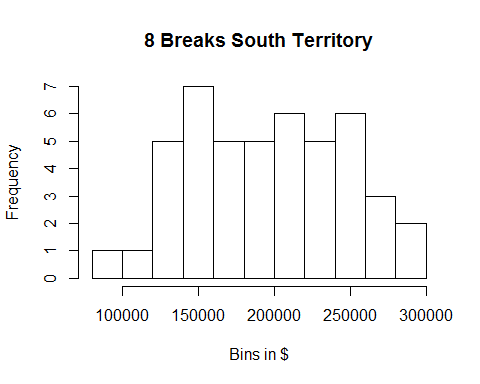
tt\_east <- territorytime$East  
tt\_north <- territorytime$North  
tt\_south <- territorytime$South  
tt\_oot <- territorytime$OOT  
  
hist(tt\_east,breaks=8, main="8 Breaks East Territory",xlab="Bins in $")



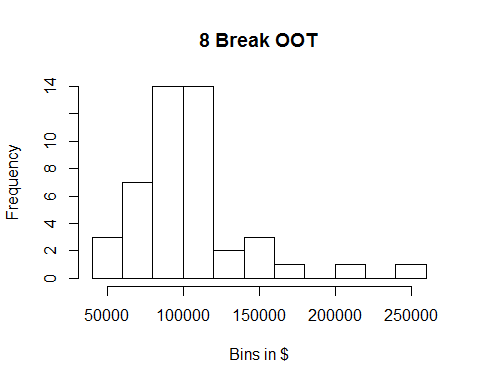
hist(tt\_north, breaks=8, main = "8 Breaks North Territory",xlab="Bins in $")



hist(tt\_south, breaks=8,main="8 Breaks South Territory",xlab="Bins in $")



hist(tt\_oot,breaks=8, main="8 Break OOT",xlab="Bins in $")



**A history of monthly sales by territory.** *Insert excel chart here* There has been a steady decline in the South territory, as well as a recent dip in sales for the East territory. No territory has had monthly sales of over $800,000 since late 2016. Beginning in mid-2016, the South and East have declined markedly, while North has remained steady.

mean(tt\_east)

## [1] 625800.5

mean(tt\_north)

## [1] 661856.8

mean(tt\_south)

## [1] 196676.9

mean(tt\_oot)

## [1] 104256.3

The data files are territorydist, which contains a list of invoices and the respective territory codes; timeseries, which contains month and year invoiced amounts, and industry, which contains invoiced amounts by industry. Columns in this dataset include 'territorycode', netivcamt, and standardized. territorycode represents the current tagged geo location based on county. netivcamt is the amount of each transaction, with a row or observation representing one transaction (invoice). standardized is the normalized value of the netivcamt column.  
Additionally, this data was extracted from Domo using the Account Master Zips and Fips dataset, with a filter applied to aggregate by transaction, and a date filter of > 1/1/2016 applied.