FinalProjectFundamntals

Blackburn

December 13, 2017

library(ggplot2)  
library(visdat)

## Warning: package 'visdat' was built under R version 3.4.3

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(mosaic)

## Warning: package 'mosaic' was built under R version 3.4.3

## Loading required package: lattice

## Loading required package: ggformula

## Warning: package 'ggformula' was built under R version 3.4.3

##   
## New to ggformula? Try the tutorials:   
## learnr::run\_tutorial("introduction", package = "ggformula")  
## learnr::run\_tutorial("refining", package = "ggformula")

## Loading required package: mosaicData

## Warning: package 'mosaicData' was built under R version 3.4.3

## Loading required package: Matrix

##   
## The 'mosaic' package masks several functions from core packages in order to add   
## additional features. The original behavior of these functions should not be affected by this.  
##   
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.

##   
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':  
##   
## mean

## The following objects are masked from 'package:dplyr':  
##   
## count, do, tally

## The following objects are masked from 'package:stats':  
##   
## binom.test, cor, cor.test, cov, fivenum, IQR, median,  
## prop.test, quantile, sd, t.test, var

## The following objects are masked from 'package:base':  
##   
## max, mean, min, prod, range, sample, sum

library(gridExtra)

##   
## Attaching package: 'gridExtra'

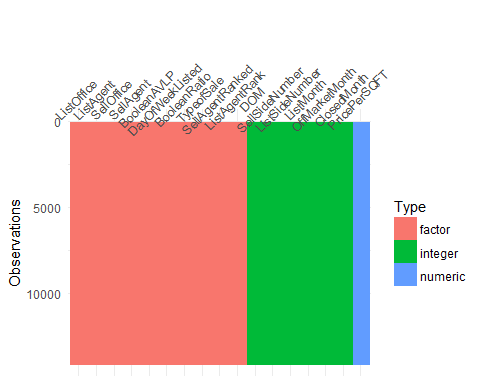
## The following object is masked from 'package:dplyr':  
##   
## combine

library(grid)  
library(RColorBrewer)

dat <- read.csv("FinalProject.csv", sep = ",", header = TRUE)  
head(dat)

## DOM ListOffice ListAgent  
## 1 68 Residential Properties, Ltd. Markham and Derentis Associates  
## 2 251 Residential Properties, Ltd. Jim DeRentis  
## 3 98 Residential Properties, Ltd. Markham and Derentis Associates  
## 4 82 Residential Properties, Ltd. Gerri Schiffman  
## 5 3 Coleman REALTORS, Inc. Michael Young  
## 6 1 Residential Properties, Ltd. Sally Lapides  
## SellOffice SellAgent  
## 1 Residential Properties, Ltd. Rebecca Rubin  
## 2 KELLER WILLIAMS RLTY NEWPORT Renee WELCHMAN  
## 3 Residential Properties, Ltd. Markham and Derentis Associates  
## 4 Residential Properties, Ltd. Gerri Schiffman  
## 5 Coleman REALTORS, Inc. Chris Healy  
## 6 Residential Properties, Ltd. Markham and Derentis Associates  
## BooleanAVLP PricePerSQFT DayOfWeekListed BooleanRatio TypeofSale  
## 1 True 155.302 Wednesday False Conv/Mkt Value  
## 2 False 195.000 Tuesday False Conv/Mkt Value  
## 3 False 155.302 Friday True Conv/Mkt Value  
## 4 False 155.302 Sunday False Conv/Mkt Value  
## 5 True 155.302 Monday False Conv/Mkt Value  
## 6 False 155.302 Thursday True Conv/Mkt Value  
## SellSideNumber SellAgentRanked ListSideNumber ListAgentRank ListMonth  
## 1 6 Low 58 High 10  
## 2 1 Low 39 High 7  
## 3 59 High 58 High 2  
## 4 73 High 99 High 3  
## 5 7 Low 5 Low 6  
## 6 59 High 2 Low 10  
## OffMarketMonth ClosedMonth  
## 1 5 5  
## 2 4 5  
## 3 7 7  
## 4 6 8  
## 5 7 7  
## 6 10 10

vis\_dat(dat)

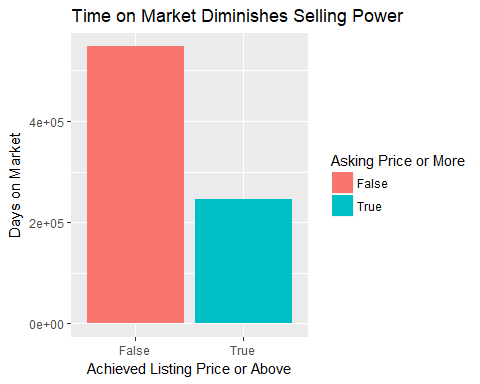


brewer.pal(n = 8, name = "Set2")

## [1] "#66C2A5" "#FC8D62" "#8DA0CB" "#E78AC3" "#A6D854" "#FFD92F" "#E5C494"  
## [8] "#B3B3B3"

I completed some exploration of the data, measuring and exploring various variables against each other to look for patterns in the visualizations, and potential changes to make in the Azure ML model.

#cbPalette <- c("#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")  
#ggplot <- function(...) ggplot2::ggplot(...) + #scale\_color\_brewer(palette="Spectral")  
ggplot(dat, aes(x=as.factor(BooleanRatio), y=DOM, fill = as.factor(BooleanRatio))) + geom\_bar(stat="identity")+  
labs(title = "Time on Market Diminishes Selling Power") +  
ylab("Days on Market") +  
xlab("Achieved Listing Price or Above")+  
guides(fill=guide\_legend(title="Asking Price or More"))



I take a look at what the standard deviation is in regard to DOM as even with clipping in Azure ML of the outliers, all of the plots were I examine the variables along Days on Market (DOM) the outliers are still quite evident.

sd(dat$DOM)

## [1] 52.80527

SDom <- scale(dat$DOM, center = TRUE, scale = TRUE)

The Standard Deviation is almost 53, since this is a direct correlation to days, that means practically 2 extra months of time on the market with this variable. I plot including the factor of the DOM with Standard Deviation, and the potential of capturing full asking price vs not does even out a bit, but there is naturally still a competitive differentiation.

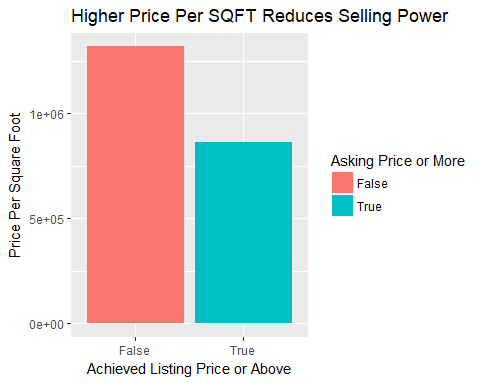
It makes sense to look at another numeric variable with the standard deviation taken into account.

sd(dat$PricePerSQFT)

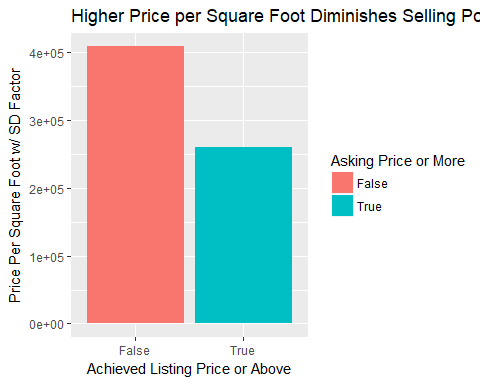
## [1] 47.25839

sdSQFT <- sd(dat$PricePerSQFT)

ggplot(dat, aes(x=as.factor(BooleanRatio), y=PricePerSQFT, fill = as.factor(BooleanRatio))) + geom\_bar(stat="identity")+  
labs(title = "Higher Price Per SQFT Reduces Selling Power") +  
ylab("Price Per Square Foot") +  
xlab("Achieved Listing Price or Above")+  
guides(fill=guide\_legend(title="Asking Price or More"))



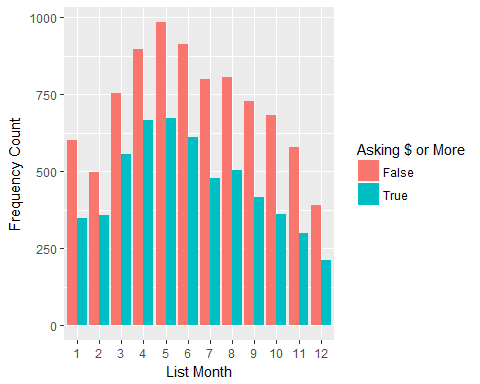
ggplot(dat, aes(x=as.factor(BooleanRatio), y=sdSQFT, fill = as.factor(BooleanRatio))) + geom\_bar(stat="identity")+  
labs(title = "Higher Price per Square Foot Diminishes Selling Power") +  
ylab("Price Per Square Foot w/ SD Factor") +  
xlab("Achieved Listing Price or Above")+  
guides(fill=guide\_legend(title="Asking Price or More"))



Lmonth<- factor(dat$ListMonth, levels = c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12"))  
Pmonth<-factor(dat$OffMarketMonth, levels = c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12"))  
Cmonth<-factor(dat$ClosedMonth, levels = c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12"))  
SellAgRank<- factor(dat$SellAgentRanked, levels = c("Low", "Medium", "High"))  
ListAgRank<- factor(dat$ListAgentRank, levels = c("Low", "Medium", "High"))  
Weekday <- factor(dat$DayOfWeekListed, levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))  
#Lmonths<- factor(dat$ListMonth, levels = c("January", "February", "March", "April", "May", "June", "July", "August", "September", "October", "November", "December"))

A rule that most real estate professionals stand by is that if one lists their property in the colder months, they will ultimately capture more revenue on their home sale. This is because of a basic supply and demand premise. The general idea that home sellers stand by is to list in the spring, when the grass is green the property looks its best. However, with a flood of inventory on the market, buyers have more options and generally should be able to negociate more off the listing price. Thusly, more supply, less demand. I did an exploration of the data to confirm or deny this premise, by looking at the listing months, pending months, and closed months. I did a side by side comparison of frequency count and proportionally in regard to our standard of measurement boolean - (did this sell for list price or higher?) The results were interesting, and seeminly contrary, at least for the past few years in this section of RI, to common thought in regard to listing real estate.

p2<- ggplot(dat, aes(x=Lmonth, fill = as.factor(BooleanRatio))) + geom\_bar(position="dodge")+  
xlab("List Month") +  
ylab("Frequency Count")+  
guides(fill=guide\_legend(title="Asking $ or More"))  
p2



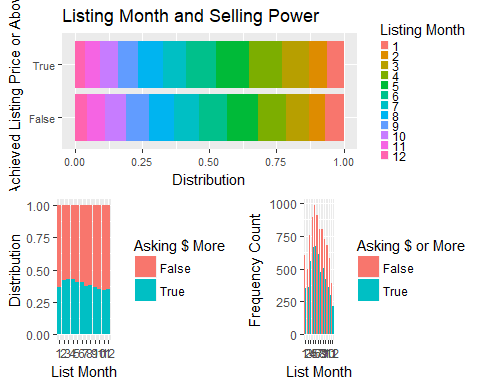
p1 <- ggplot(dat, aes(x=as.factor(BooleanRatio), fill = Lmonth))+   
geom\_bar(position ="fill")+  
labs(title = "Listing Month and Selling Power") +  
ylab("Distribution") +  
xlab("Achieved Listing Price or Above")+  
guides(fill=guide\_legend(title="Listing Month"))+  
theme(legend.key.size = unit(0.1, "in"))+  
theme(axis.text = element\_text(size = 8))+  
coord\_flip()

bla blba bla

p20<- ggplot(dat, aes(x=Lmonth, fill = as.factor(BooleanRatio))) + geom\_bar(position="fill")+  
xlab("List Month") +  
ylab("Distribution")+  
guides(fill=guide\_legend(title="Asking $ More"))

bla bla bla

require(gridExtra)  
grid.arrange(p1, arrangeGrob(p20,p2, ncol=2), ncol=1)



BLA BLA BLA

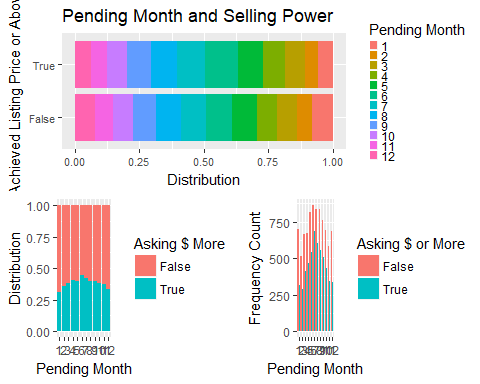
p3<- ggplot(dat, aes(x=Pmonth, fill = as.factor(BooleanRatio))) + geom\_bar(position="dodge")+  
xlab("Pending Month") +  
ylab("Frequency Count")+  
guides(fill=guide\_legend(title="Asking $ or More"))

p4 <- ggplot(dat, aes(x=as.factor(BooleanRatio), fill = Pmonth)) +   
geom\_bar(position = "fill")+  
labs(title = "Pending Month and Selling Power") +  
ylab("Distribution") +  
xlab("Achieved Listing Price or Above")+  
guides(fill=guide\_legend(title="Pending Month"))+  
theme(legend.key.size = unit(0.1, "in"))+  
theme(axis.text = element\_text(size = 8))+  
coord\_flip()

bla bla

p21<- ggplot(dat, aes(x=Pmonth, fill = as.factor(BooleanRatio))) + geom\_bar(position="fill")+  
xlab("Pending Month") +  
ylab("Distribution")+  
guides(fill=guide\_legend(title="Asking $ More"))

require(gridExtra)  
grid.arrange(p4, arrangeGrob(p21,p3, ncol=2), ncol=1)



bla bla

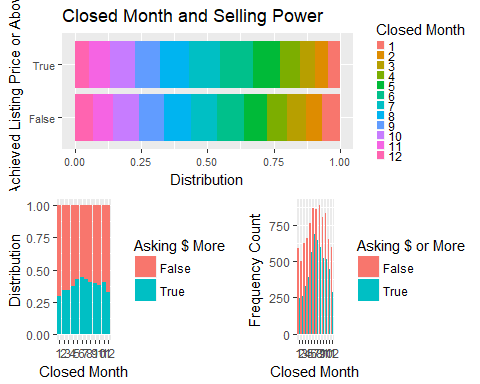
p5<- ggplot(dat, aes(x=Cmonth, fill = as.factor(BooleanRatio))) + geom\_bar(position="dodge")+  
xlab("Closed Month") +  
ylab("Frequency Count")+  
guides(fill=guide\_legend(title="Asking $ or More"))

p6 <- ggplot(dat, aes(x=as.factor(BooleanRatio), fill = Cmonth)) +   
geom\_bar(position = "fill")+  
labs(title = "Closed Month and Selling Power") +  
ylab("Distribution") +  
xlab("Achieved Listing Price or Above")+  
guides(fill=guide\_legend(title="Closed Month"))+  
theme(legend.key.size = unit(0.1, "in"))+  
theme(axis.text = element\_text(size = 8))+  
coord\_flip()

bla bla

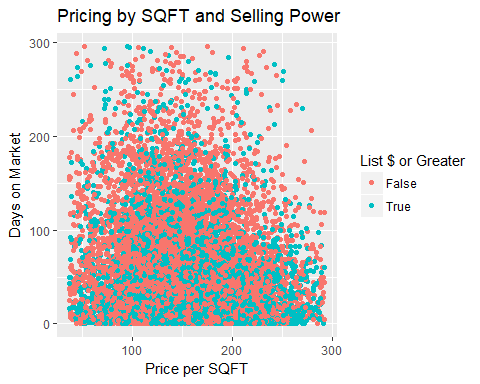
p22<- ggplot(dat, aes(x=Cmonth, fill = as.factor(BooleanRatio))) + geom\_bar(position="fill")+  
xlab("Closed Month") +  
ylab("Distribution")+  
guides(fill=guide\_legend(title="Asking $ More"))

require(gridExtra)  
grid.arrange(p6, arrangeGrob(p22,p5, ncol=2), ncol=1)

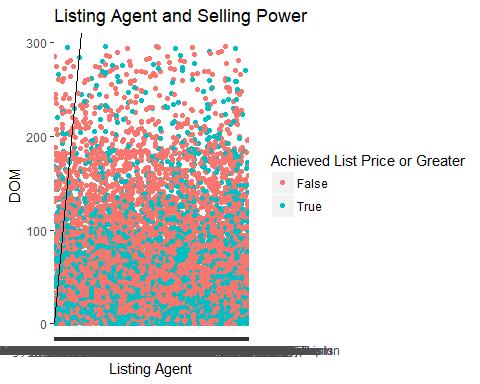


I wanted to look at pricing schema, against days on market and if they achieved the listing price or more. The graph is difficult to read because of a few outliers, so we address this in the next plot.

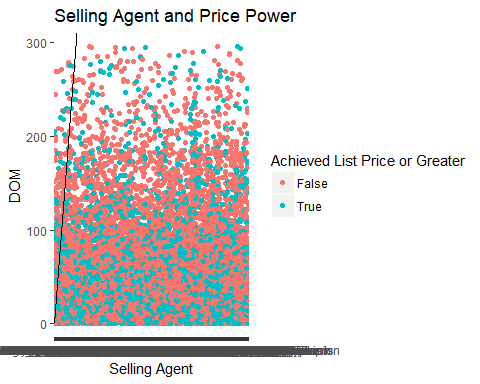
p7<- ggplot(dat, aes(x=PricePerSQFT, y= DOM, colour = as.factor(BooleanRatio))) + geom\_point()+  
labs(title = "Pricing by SQFT and Selling Power") +  
ylab("Days on Market") +  
xlab("Price per SQFT")+  
guides(colour=guide\_legend(title="List $ or Greater"))  
p7



p8<- ggplot(dat, aes(x= ListAgent, y = DOM, colour = as.factor(BooleanRatio))) + geom\_point()+  
labs(title = "Listing Agent and Selling Power") +  
xlab("Listing Agent") +  
ylab("DOM")+  
geom\_abline()+  
guides(colour=guide\_legend(title="Achieved List Price or Greater"))  
#theme(axis.text.x = element\_text(angle = 90, hjust = 1))  
p8



p9<- ggplot(dat, aes(x= SellAgent, y = DOM, colour = as.factor(BooleanRatio))) + geom\_point()+  
labs(title = "Selling Agent and Price Power") +  
xlab("Selling Agent") +  
ylab("DOM")+  
geom\_abline()+  
guides(colour=guide\_legend(title="Achieved List Price or Greater"))  
#theme(axis.text.x = element\_text(angle = 90, hjust = 1))  
p9



We clearly have an issue with the amount of variables in regard to listing agent and selling agent. Furthermore, on examination, we see that over the course of over almost 3 years, several agents had hardly ever sold a home at all. These are paths in our desicion trees that we may not need to follow. Therefore, we are going to attempt to put these agents in different buckets, visualized by histograms, on how many homes they actually sold. This begged for further investigation, and potential segmentation of the variable.

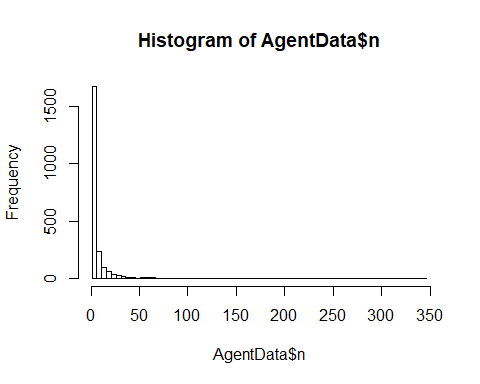
nrow(dat)

## [1] 14116

ListAgentSynched <- table(dat$ListAgent)  
ListAgentSynched <- as.data.frame(table(dat$ListAgent))  
nrow(ListAgentSynched)

## [1] 2204

AgentData <- dat %>%   
 count(ListAgent)   
hist(AgentData$n, breaks = seq(1, 350, by = 5))



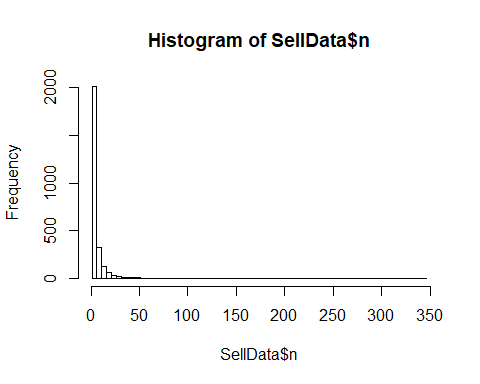
nrow(dat)

## [1] 14116

SellAgentSynched <- table(dat$SellAgent)  
SellAgentSynched <- as.data.frame(table(dat$SellAgent))  
nrow(SellAgentSynched)

## [1] 2614

SellData <- dat %>%   
 count(SellAgent)   
hist(SellData$n, breaks = seq(1, 350, by = 5))

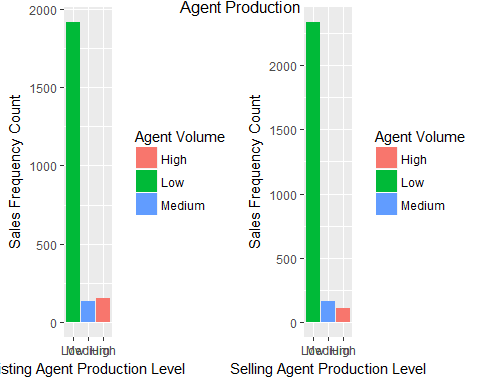


There is a huge disperity between those who are selling more than a handful of homes a year, and those who are not. I create different buckets, segmenting these agents into "low", "medium", and "high" producers.

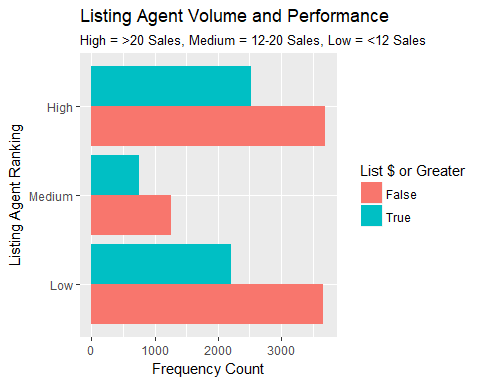
# High > 20, Medium > 12, Low  
AgentData <- AgentData %>% mutate(AgentVolume = ifelse(n < 12, "Low", ifelse(n < 20, "Medium", "High")))  
AgVol <- factor(AgentData$AgentVolume, levels = c("Low", "Medium", "High"))  
p10 <- ggplot(AgentData, aes(x=AgVol, fill = AgentVolume)) + geom\_bar()+  
xlab("Listing Agent Production Level") +  
ylab("Sales Frequency Count")+  
guides(fill=guide\_legend(title="Agent Volume"))

# High > 20, Medium > 12, Low  
SellData <- SellData %>% mutate(AgentVolume = ifelse(n < 12, "Low", ifelse(n < 20, "Medium", "High")))  
SAgVol <- factor(SellData$AgentVolume, levels = c("Low", "Medium", "High"))  
p11 <- ggplot(SellData, aes(x=SAgVol, fill = AgentVolume)) + geom\_bar()+  
xlab("Selling Agent Production Level") +  
ylab("Sales Frequency Count")+  
guides(fill=guide\_legend(title="Agent Volume"))

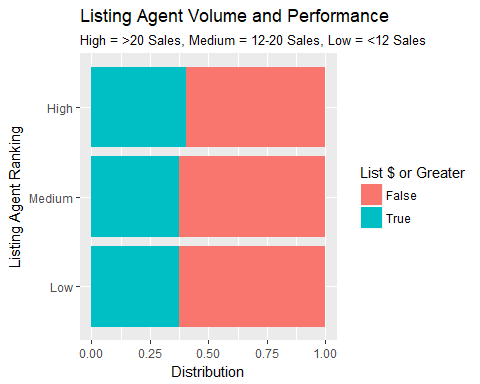
require(gridExtra)  
grid.arrange(p10, p11, ncol=2)  
pushViewport(viewport(layout = grid.layout(4, 2, heights = unit(c(2, 28, 28), "null"))))  
grid.text("Agent Production", vp = viewport(layout.pos.row = 1, layout.pos.col = 1:2))



p13<- ggplot(dat, aes(x=ListAgRank, fill =as.factor(BooleanRatio)))+  
geom\_bar(position = "dodge")+  
labs(title = "Listing Agent Volume and Performance",   
 subtitle = "High = >20 Sales, Medium = 12-20 Sales, Low = <12 Sales") +  
ylab("Frequency Count") +  
xlab("Listing Agent Ranking")+  
guides(fill=guide\_legend(title="List $ or Greater"))+  
coord\_flip()  
p13

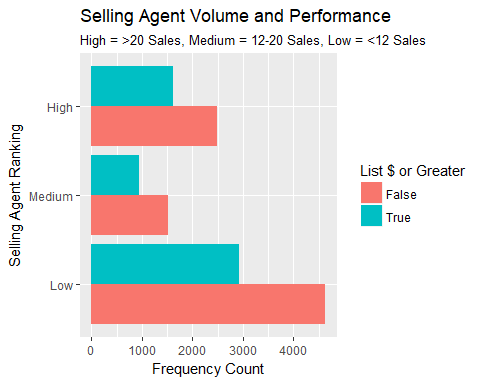


p14<- ggplot(dat, aes(x=ListAgRank, fill =as.factor(BooleanRatio)))+  
geom\_bar(position = "fill")+  
labs(title = "Listing Agent Volume and Performance",   
 subtitle = "High = >20 Sales, Medium = 12-20 Sales, Low = <12 Sales") +  
ylab("Distribution") +  
xlab("Listing Agent Ranking")+  
guides(fill=guide\_legend(title="List $ or Greater"))+  
coord\_flip()  
p14

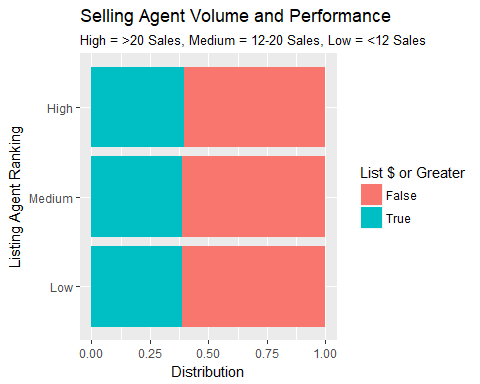


#require(gridExtra)  
#grid.arrange(p13, p14, ncol=2)

p15<- ggplot(dat, aes(x=SellAgRank, fill =as.factor(BooleanRatio)))+  
geom\_bar(position = "dodge")+  
labs(title = "Selling Agent Volume and Performance",   
 subtitle = "High = >20 Sales, Medium = 12-20 Sales, Low = <12 Sales") +  
ylab("Frequency Count") +  
xlab("Selling Agent Ranking")+  
guides(fill=guide\_legend(title="List $ or Greater"))+  
coord\_flip()  
p15



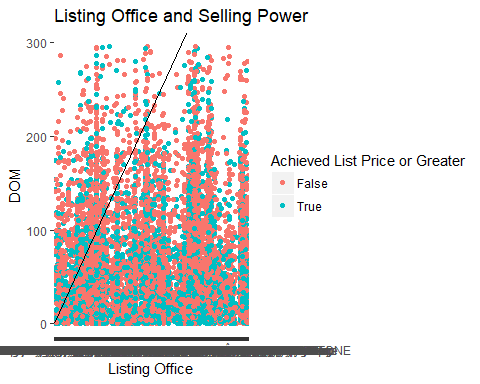
p16<- ggplot(dat, aes(x=SellAgRank, fill =as.factor(BooleanRatio)))+  
geom\_bar(position = "fill")+  
labs(title = "Selling Agent Volume and Performance",   
 subtitle = "High = >20 Sales, Medium = 12-20 Sales, Low = <12 Sales") +  
ylab("Distribution") +  
xlab("Listing Agent Ranking")+  
guides(fill=guide\_legend(title="List $ or Greater"))+  
coord\_flip()  
p16



#require(gridExtra)  
#grid.arrange(p15, p16, ncol=2)

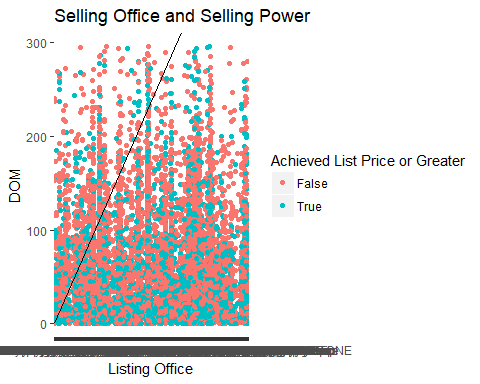
Now let's take a look at the incredible number of different observations within the Listing and Selling Agency variables.

p30<- ggplot(dat, aes(x=ListOffice, y = DOM, colour = as.factor(BooleanRatio))) + geom\_point()+  
labs(title = "Listing Office and Selling Power") +  
xlab("Listing Office") +  
ylab("DOM")+  
geom\_abline()+  
guides(colour=guide\_legend(title="Achieved List Price or Greater"))  
#theme(axis.text.x = element\_text(angle = 90, hjust = 1))  
p30



Now we will apply the same method to the selling offices.

p31<- ggplot(dat, aes(x=SellOffice, y = DOM, colour = as.factor(BooleanRatio))) + geom\_point()+  
labs(title = "Selling Office and Selling Power") +  
xlab("Listing Office") +  
ylab("DOM")+  
geom\_abline()+  
guides(colour=guide\_legend(title="Achieved List Price or Greater"))  
#theme(axis.text.x = element\_text(angle = 90, hjust = 1))  
p31



Again we see that there is something interesting going on with the plot - but there is too much information to be clear. Therefore, we'll try some binning.

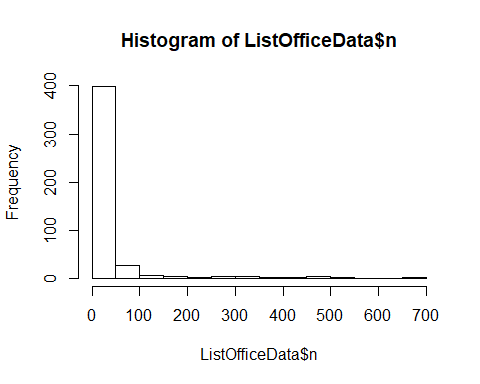
nrow(dat)

## [1] 14116

ListOfficeSynched <- table(dat$ListOffice)  
ListOfficeSynched <- as.data.frame(table(dat$ListOffice))  
nrow(ListOfficeSynched)

## [1] 455

ListOfficeData <- dat %>%   
 count(ListOffice)   
hist(ListOfficeData$n, breaks = 10)



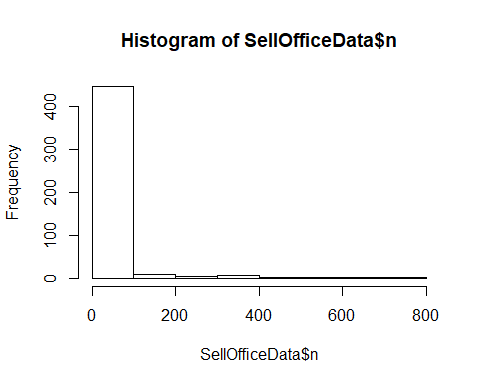
nrow(dat)

## [1] 14116

SellOfficeSynched <- table(dat$SellOffice)  
SellOfficeSynched <- as.data.frame(table(dat$SellOffice))  
nrow(SellOfficeSynched)

## [1] 473

SellOfficeData <- dat %>%   
 count(SellOffice)   
hist(SellOfficeData$n, breaks = 10)

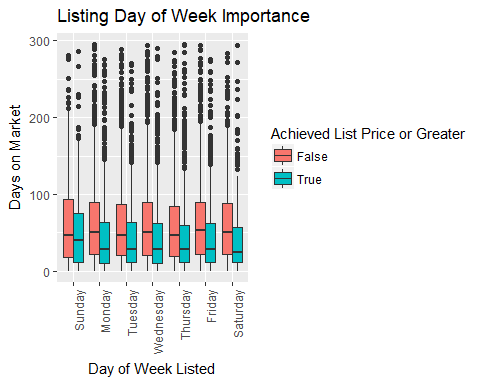


# High > 20, Medium > 12, Low  
ListOfficeData <- ListOfficeData %>% mutate(ListOfficeVolume = ifelse(n < 12, "Low", ifelse(n < 20, "Medium", "High")))  
LOfficeVol <- factor(ListOfficeData$ListOfficeVolume, levels = c("Low", "Medium", "High"))

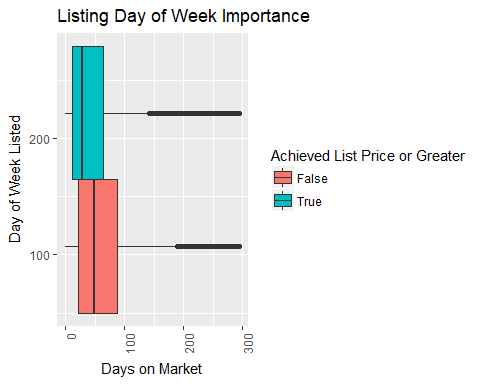
# High > 20, Medium > 12, Low  
SellOfficeData <- SellOfficeData %>% mutate(SellOfficeVolume = ifelse(n < 12, "Low", ifelse(n < 20, "Medium", "High")))  
SOfficeVol <- factor(ListOfficeData$ListOfficeVolume, levels = c("Low", "Medium", "High"))

# High > 20, Medium > 12, Low  
AgentData <- AgentData %>% mutate(AgentVolume = ifelse(n < 12, "Low", ifelse(n < 20, "Medium", "High")))  
AgVol <- factor(AgentData$AgentVolume, levels = c("Low", "Medium", "High"))  
p10 <- ggplot(AgentData, aes(x=AgVol, fill = AgentVolume)) + geom\_bar()+  
xlab("Listing Agent Production Level") +  
ylab("Sales Frequency Count")+  
guides(fill=guide\_legend(title="Agent Volume"))

p18<- ggplot(dat, aes(x=Weekday, y = DOM, fill = as.factor(BooleanRatio))) + geom\_boxplot()+  
labs(title = "Listing Day of Week Importance") +  
xlab("Day of Week Listed") +  
ylab("Days on Market")+  
guides(fill=guide\_legend(title="List $ or Greater"))+  
theme(axis.text.x = element\_text(angle = 90, hjust = 1))+  
guides(fill=guide\_legend(title="Achieved List Price or Greater"))  
p18



p25<- ggplot(dat, aes(x=PricePerSQFT, y = DOM, fill = as.factor(BooleanRatio))) + geom\_boxplot()+  
labs(title = "Listing Day of Week Importance") +  
xlab("Day of Week Listed") +  
ylab("Days on Market")+  
guides(fill=guide\_legend(title="List $ or Greater"))+  
theme(axis.text.x = element\_text(angle = 90, hjust = 1))+  
guides(fill=guide\_legend(title="Achieved List Price or Greater"))+  
coord\_flip()  
p25



ListOfficeData

## # A tibble: 455 x 3  
## ListOffice n ListOfficeVolume  
## <fctr> <int> <chr>  
## 1 36d5 Realty 3 Low  
## 2 A.R. Bilodeau Inc. 1 Low  
## 3 ABBOTT PROPERTIES 19 Medium  
## 4 Access Realty, Inc. 3 Low  
## 5 Acumen Group 30 High  
## 6 Adams & Adams Real Estate 1 Low  
## 7 Adams Realty 2 Low  
## 8 Adaptive Real Estate 1 Low  
## 9 Agnelli Real Estate 3 Low  
## 10 AJ Realty 2 Low  
## # ... with 445 more rows

SellOfficeData

## # A tibble: 473 x 3  
## SellOffice n SellOfficeVolume  
## <fctr> <int> <chr>  
## 1 1 Oak Real Estate Group 3 Low  
## 2 36d5 Realty 13 Medium  
## 3 401 Realty 1 Low  
## 4 ABBOTT PROPERTIES 20 High  
## 5 Access Realty, Inc. 2 Low  
## 6 Acumen Group 27 High  
## 7 Adams & Adams Real Estate 1 Low  
## 8 Adams Realty 1 Low  
## 9 Adaptive Real Estate 3 Low  
## 10 Adept Realty Group 3 Low  
## # ... with 463 more rows