Final Project

Benjamin Foglia

8/20/2019

## Discussion

For this report, I examined the dataset sales\_data.

## Initial Setup

## run important packages and read in data  
library(tidyverse)

## -- Attaching packages ------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.1 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 1.0.0  
## v tidyr 1.1.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ---------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(tidyr)  
library(ggplot2)  
library(readr)  
library(stringr)  
library(dplyr)  
  
Sales\_data <- read\_csv("F:/BAN 6003/sales\_data.csv")

## Parsed with column specification:  
## cols(  
## flag = col\_character(),  
## gender = col\_character(),  
## education = col\_character(),  
## house\_val = col\_double(),  
## age = col\_character(),  
## online = col\_character(),  
## customer\_psy = col\_character(),  
## marriage = col\_character(),  
## child = col\_character(),  
## occupation = col\_character(),  
## mortgage = col\_character(),  
## house\_owner = col\_character(),  
## region = col\_character(),  
## car\_prob = col\_double(),  
## fam\_income = col\_character()  
## )

## Metadata

**flag** **gender** **education** **house\_val** **age\_bracket** - *which age bin the person falls in (1=unknown, 2=lowest, 7=highest)* **age\_range** - *person’s age range (i.e. >65)* **online** **customer\_psy** **marriage** **children** **occupation** **mortgage** **house\_own** **region** **car\_prob** **fam\_income**

## Wrangling

The first thing I wanted to do with the initial dataset was change values like education and mortgage to numerical values. I also split age into 2 columns, one containting the number before the ’\_’, and one after. I thought this made it easier to read, and that using the first number would making grouping people together by age easier.

## Education  
Sales\_data$education <- str\_remove\_all(Sales\_data$education, " <HS")  
Sales\_data$education <- str\_remove\_all(Sales\_data$education, " HS")  
Sales\_data$education <- str\_remove\_all(Sales\_data$education, " Some College")  
Sales\_data$education <- str\_remove\_all(Sales\_data$education, " Bach")  
Sales\_data$education <- str\_remove\_all(Sales\_data$education, " Grad")  
  
Sales\_data$education <- extract\_numeric(Sales\_data$education)

## extract\_numeric() is deprecated: please use readr::parse\_number() instead

## Mortgage  
Sales\_data$mortgage <- str\_remove\_all(Sales\_data$mortgage, "low")  
Sales\_data$mortgage <- str\_remove\_all(Sales\_data$mortgage, "med")  
Sales\_data$mortgage <- str\_remove\_all(Sales\_data$mortgage, "high")  
  
Sales\_data$mortgage <- extract\_numeric(Sales\_data$mortgage)

## extract\_numeric() is deprecated: please use readr::parse\_number() instead

## Age  
Sales\_data <- separate(Sales\_data, age, c("age\_bracket", "age\_range"), sep = "\_")

## Transformation

From the original sales\_data, I created multiple new data sets by filtering for specific values.

## Data from all who bought the product  
buyers\_data <- Sales\_data %>% filter(flag == "Y")  
  
## Data for all homeowners whose houses had value  
house\_owners <- Sales\_data %>% filter(house\_val > 0, house\_owner == "Owner")  
  
## House value by region  
val\_by\_region <- Sales\_data %>% select(house\_val, region)

## Functions, Vectors, Iterations

Per Doc B’s request, I have included the function and for loop that categorizes and counts house values by “below average”, “average to expensive”, and “very expensive”.

mean\_val <- mean(Sales\_data$house\_val)  
very\_high <- 2\*sd(Sales\_data$house\_val) + mean\_val  
  
house\_expenses <- function(x){  
 if (x > very\_high) {  
 print("Very Expensive")  
 } else if (x > mean\_val){  
 print("Average to Expensive")  
 } else {  
 print("Below Average")}  
}  
# Test  
house\_expenses(1200000)

## [1] "Very Expensive"

## For loop  
  
input <- c(1000, 310000, 1200000)  
output <- vector("character", length(input))  
  
for (i in seq\_along(input)){  
 output[[i]] <- house\_expenses(input[[i]])  
}

## [1] "Below Average"  
## [1] "Average to Expensive"  
## [1] "Very Expensive"

input\_house\_val <- Sales\_data$house\_val

for (i in seq\_along(input\_house\_val)){  
 output[[i]] <- house\_expenses(input\_house\_val[[i]])  
}  
v=0  
a=0  
b=0  
  
for (i in seq\_along(output)){  
 if(output[[i]] == "Very Expensive") {v=v+1}  
 else if(output[[i]] == "Average to Expensive"){a=a+1}  
 else{b=b+1}  
}

house\_results <- c("Very Expensive" = v, "Average to Expensive" = a, "Below Average" = b)  
house\_results

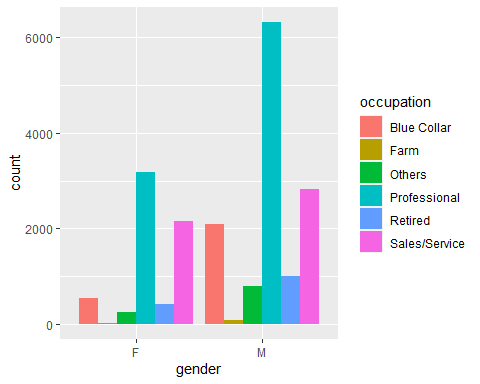
## Very Expensive Average to Expensive Below Average   
## 1203 12654 26143

## most houses are below average value

## Visualizations

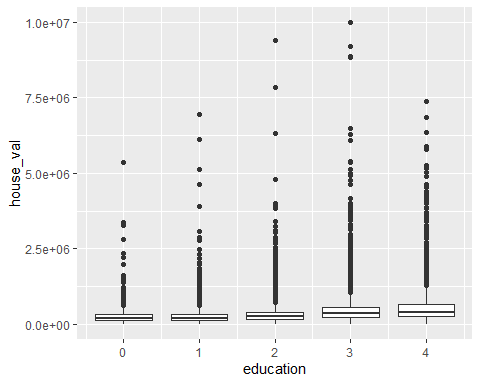
I focused on a couple of different things for my visualizations. First and foremost, I wanted to see what types of people were buying the product. Secondly, I did some of what Doc B asked for and looked for variations in in house values with respect to other variables.

## Using buyers\_data, I looked at which genders and professions were buying the product. To simplify, I only used data where the gender of the person was known.  
ggplot(buyers\_data %>% filter(gender != "U"))+  
 geom\_bar(aes(x=gender, fill = occupation), position = "dodge")

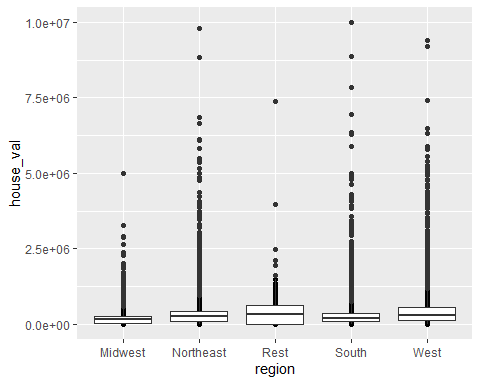


## For Doc B, I created a boxplot to compare education and house value using the house\_owner dataset  
ggplot(house\_owners, aes(x=education, y=house\_val))+  
 geom\_boxplot(aes(group = education))

## Warning: Removed 420 rows containing missing values (stat\_boxplot).



## based on the boxplot, while the most expensive houses are in education groups 2 & 3, there does seem to be evidence suggesting that someone with a higher education level will own a more expensive house.  
  
## Another Boxplot, this time comparing house value by region  
ggplot(val\_by\_region, aes(x=region, y = house\_val, group = region))+  
 geom\_point()+  
 geom\_boxplot()



## Final Notes

I didn’t want to make a second boxplot; I would have much preferred to make a third type of visualization. I tried to spread my val\_by\_region and have each region be its own variable, but kept getiing errors saying “Each row of output must be identified by a unique combination of keys. Keys are shared for 40000 rows”. I then would have wanted to use the geom\_map function to illustrate average price by region in the United States.