Black Friday Project

Dave Anderson

December 5, 2018

Abstract

Introduction

I find quantifying human behavior to be both incredibly interesting and powerful. One of the best ways to understand our society is to examine consumer data. Like it or not, money is a critical aspect of our lives, and how we choose to spend it is an important decision. I am facinated by the ability of companies to predict who there customers will be and what they will buy, and I would be honored to work in a position where I perform similar analysis to help a company succeed. To begin to understand this field, I chose to analyze a company's black friday sales report. The data includes about 500,000 transactions. Variables include a customer code, age category, gender, occupation category, product category, and city information. Most of the data is masked as we do not know what the store is or what the categories represent. At first, the missing information almost detered me from the topic, but then I decided it would be interesting to attempt to discover possible answers for the missing categories through the data. I began my investigation by understanding the dataset at a basic level.

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(coefplot)
library(kableExtra)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
blackfriday <- read_csv("BlackFriday.csv")
```

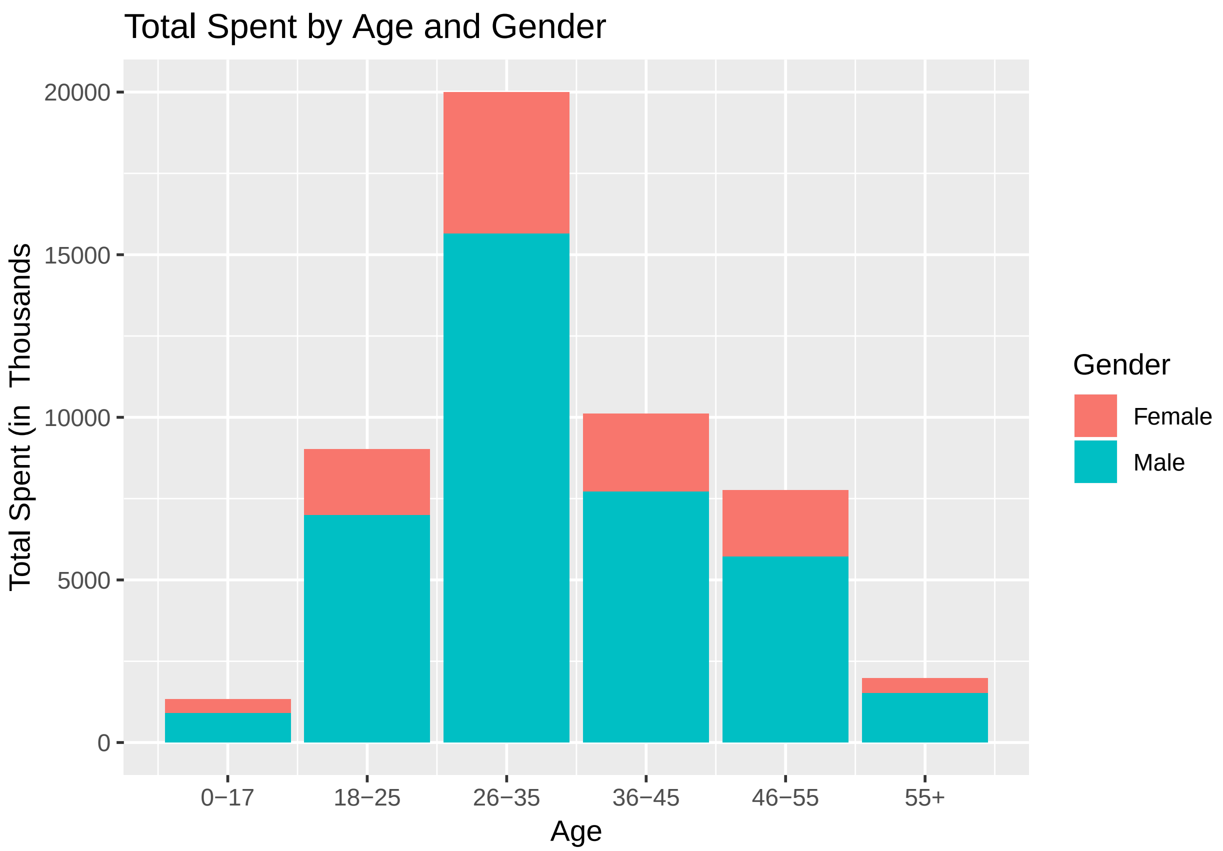
```
## Parsed with column specification:
## cols(
##   User_ID = col_integer(),
##   Product_ID = col_character(),
##   Gender = col_character(),
##   Age = col_character(),
##   Occupation = col_integer(),
##   City_Category = col_character(),
##   Stay_In_Current_City_Years = col_character(),
##   Marital_Status = col_integer(),
##   Product_Category_1 = col_integer(),
##   Product_Category_2 = col_integer(),
##   Product_Category_3 = col_integer(),
##   Purchase = col_integer()
## )
```

```
#Add column for number of purchases for each person
purchase <- blackfriday %>% group_by(User_ID) %>% summarise(Purchases=n())
blackfriday <- full_join(blackfriday,purchase,by = "User_ID")

#Add Age Category
blackfriday <- blackfriday %>% mutate(age_cat = ifelse(Age == '0-17',0,ifelse(Age == '18-25',1,ifelse(Ag

#Add Gender Binary
blackfriday <- blackfriday %>% mutate(gender_cat = ifelse(Gender == "M",1,0))

#Add City Category
blackfriday <- blackfriday %>% mutate(city_cat = ifelse(City_Category == "A",1,ifelse(City_Category ==

#Change to dollars
blackfriday$Purchase <- blackfriday$Purchase/100

#Individuals as single observation
unique <- blackfriday %>% group_by(User_ID) %>% summarise(Average = mean(Purchase),gender = min(gender_
```

EDA

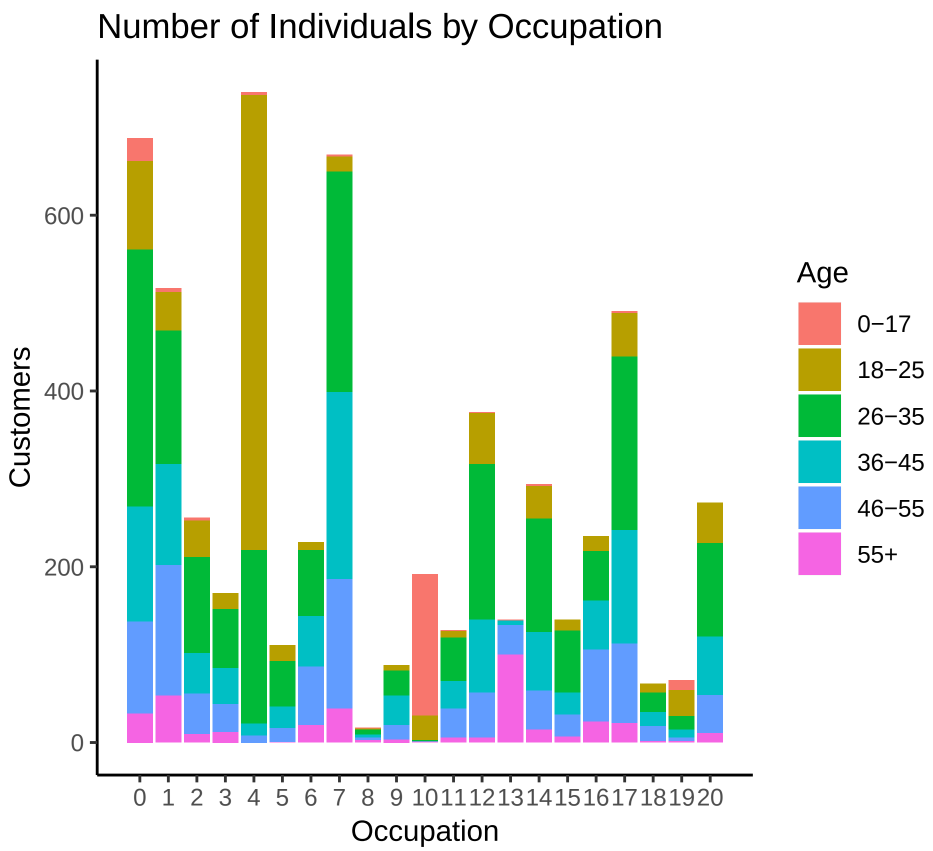
Demographics The original dataset will be useful in analyzing the consumer data by looking into what products certain people are buying. But first, I wanted to learn about who the customers are. I created a dataset with each individual as a row, including average purchase, total purchase amount, and number of purchases as variables. There are 5,891 customers with number of purchases ranging from 5 to 1,025. From the first plot, we can see that the target demographic of this store is 26-35 year old males. Males actually tend to buy more expensive products (95 to 88), more items (222 to 192), and there are many more male customers in general (4,225 to 1,666).

```
#Counts and Purchases by age, gender
ggplot(unique,aes(age,total/1000))+
  geom_col(aes(fill = factor(gender)))+
  labs(title = "Total Spent by Age and Gender",y = "Total Spent (in  Thousands",x = "Age")+
  scale_fill_discrete(name="Gender",breaks=c("0", "1"),
                          labels=c("Female", "Male"))+
  scale_x_continuous(breaks=c(0,1,2,3,4,5),labels = c("0-17","18-25","26-35","36-45","46-55","55+"))
```

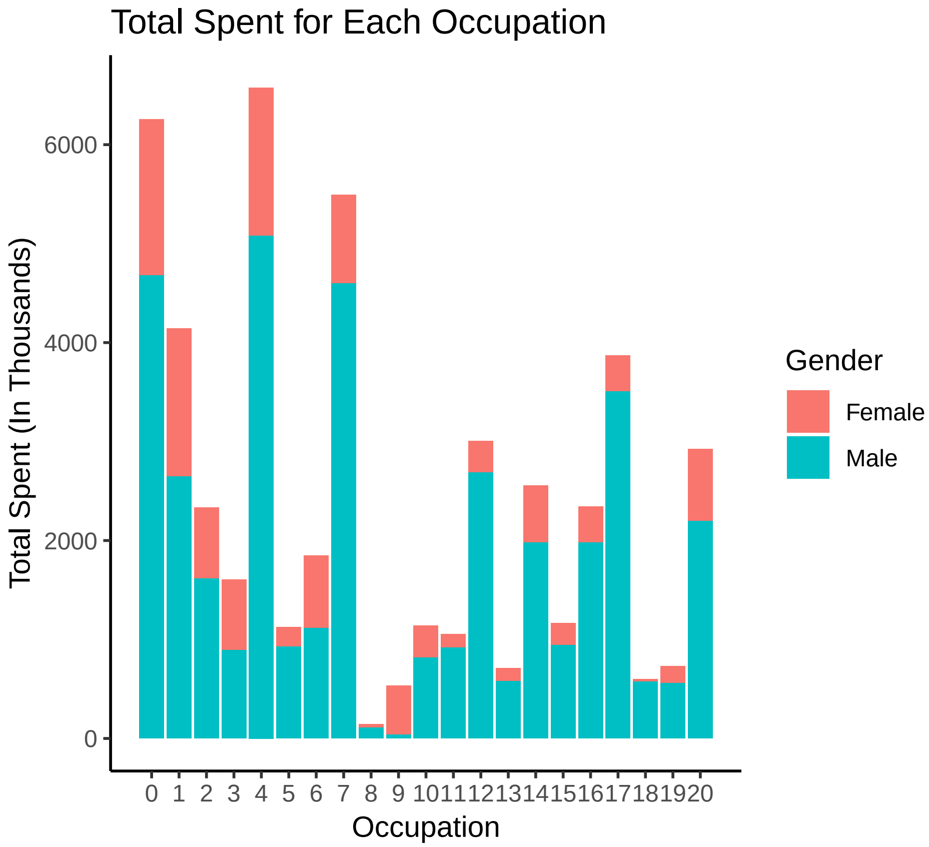


Occupations Occupation is most likely another key factor. From the two plots, we can see that occupations 0, 4 and 7 have the most customers and spend the most. Occupation 9 is the only one with a female majority. Occupation 10 is where most of the customers under the age of 18 are, which makes me believe this is unemployed or student. Occupation 4 could potentially be college students, with most of the 18-25 year olds in this category. The company's target age, 26-35, is present across multiple occupations.

```
par(mfrow = c(2,2))

#Occupation Numbers
ggplot(unique,aes(occupation))+
  geom_bar(aes(fill = factor(age)))+
  scale_x_continuous(breaks = 0:20)+
  labs(title = "Number of Individuals by Occupation", x = "Occupation", y = "Customers")+
   scale_fill_discrete(name="Age",breaks=c(0:5),
                         labels=c("0-17", "18-25","26-35","36-45","46-55","55+"))+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),panel.background = elemen
```



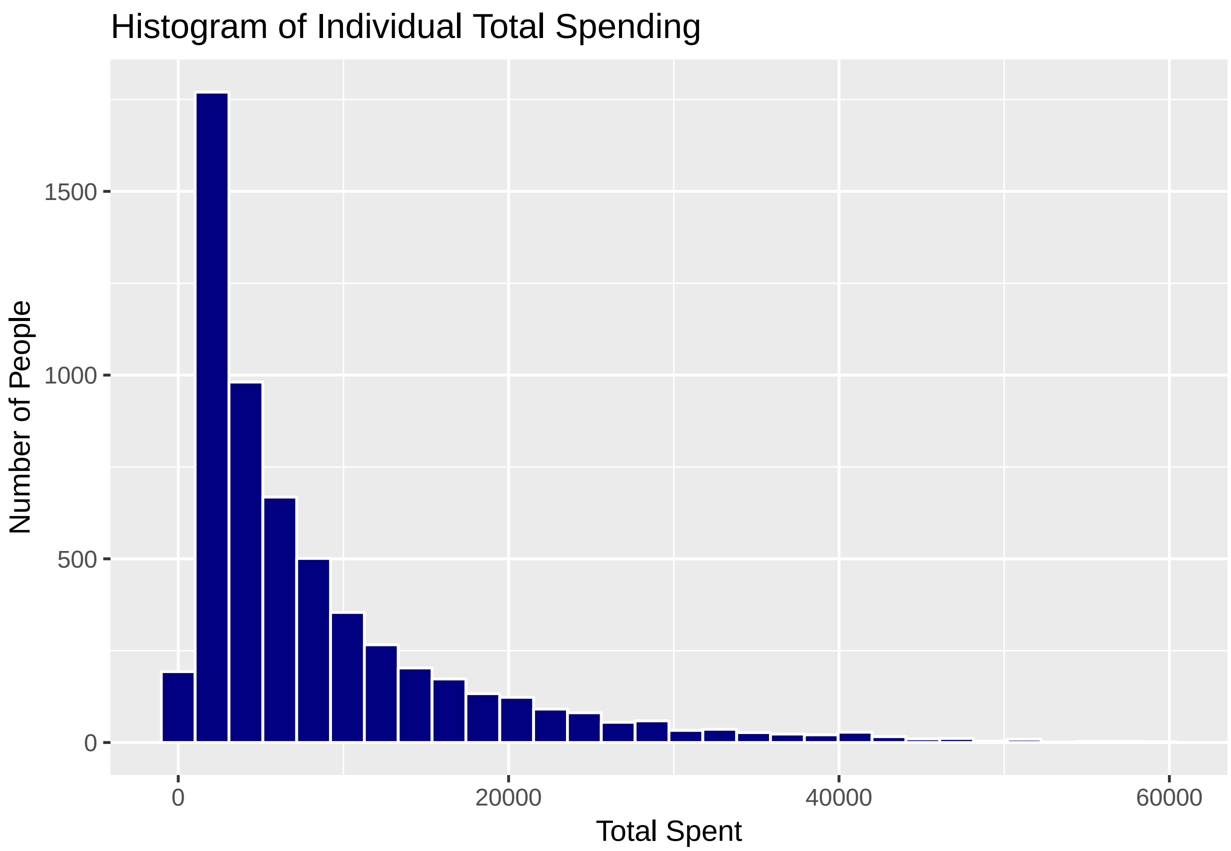
```
#Occupation Spending
ggplot(unique,aes(occupation,total/1000))+
  geom_col(aes(fill = factor(gender)))+
  scale_x_continuous(breaks = 0:20)+
  labs(title = "Total Spent for Each Occupation", y = "Total Spent (In Thousands)", x = "Occupation")+
   scale_fill_discrete(name="Gender",breaks=c("0", "1"),
                         labels=c("Female", "Male"))+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
panel.background = element_blank(), axis.line = element_line(colour = "black"))
```



Purchases at Individual Level As previously stated, the goal of studying this data are to help maximize profits for the company. Therefore, the key outcome variables for our models will be purchase amounts for the individual, number of purchases for each individual, and product purchases. To investigate these variables, I began with a histogram of total amount spent by each individual. For visual purposes, I excluded 13 individuals over 60,000, with the largest value being about 100,000. Plotting the distribution of the number of purchases for each customer shows a similar pattern.

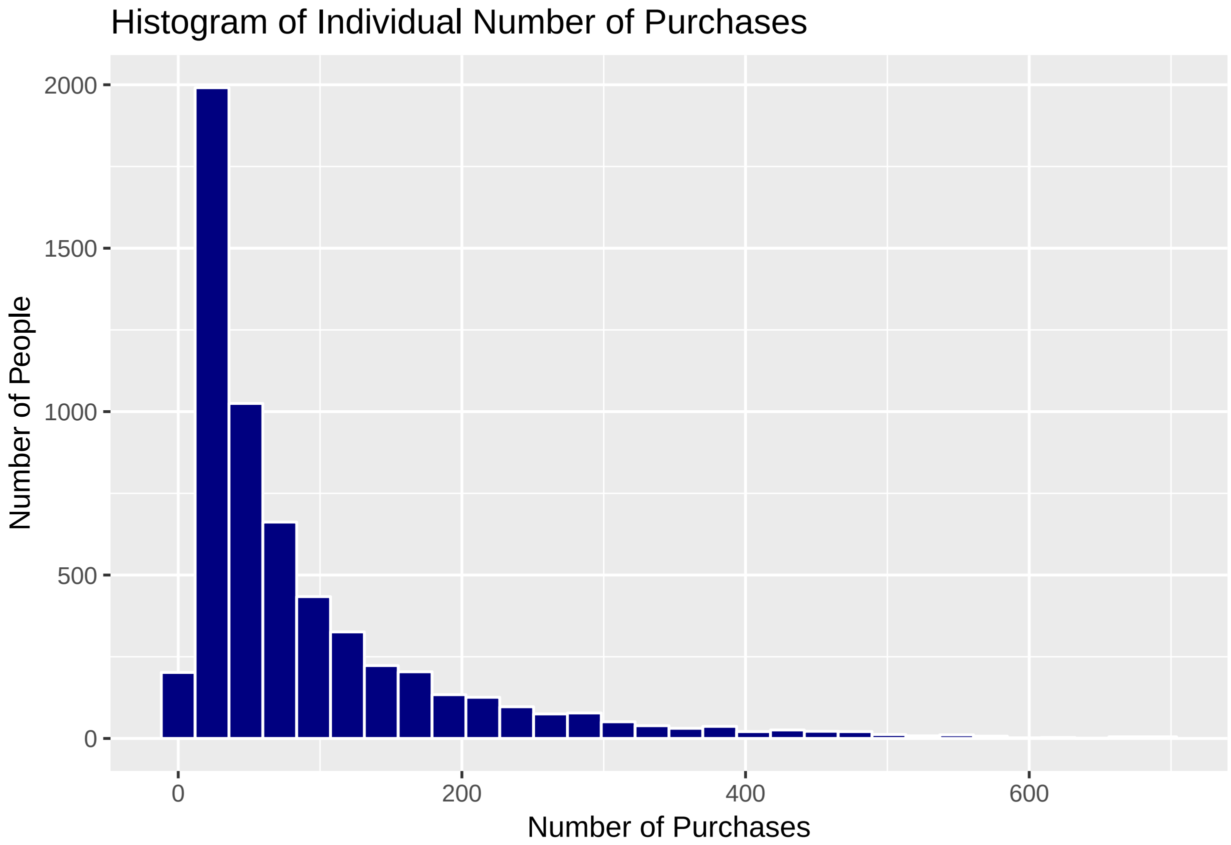
```
totals_hist <- unique %>% filter(total < 60000)
ggplot(totals_hist,aes(total))+geom_histogram(color = "white",fill = "navy")+
  labs(title = "Histogram of Individual Total Spending", x = "Total Spent", y = "Number of People")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
number_hist <- unique %>% filter(n < 700)
ggplot(number_hist,aes(n))+geom_histogram(color = "white", fill = "navy")+
  labs(title = "Histogram of Individual Number of Purchases", x = "Number of Purchases", y = "Number of
```

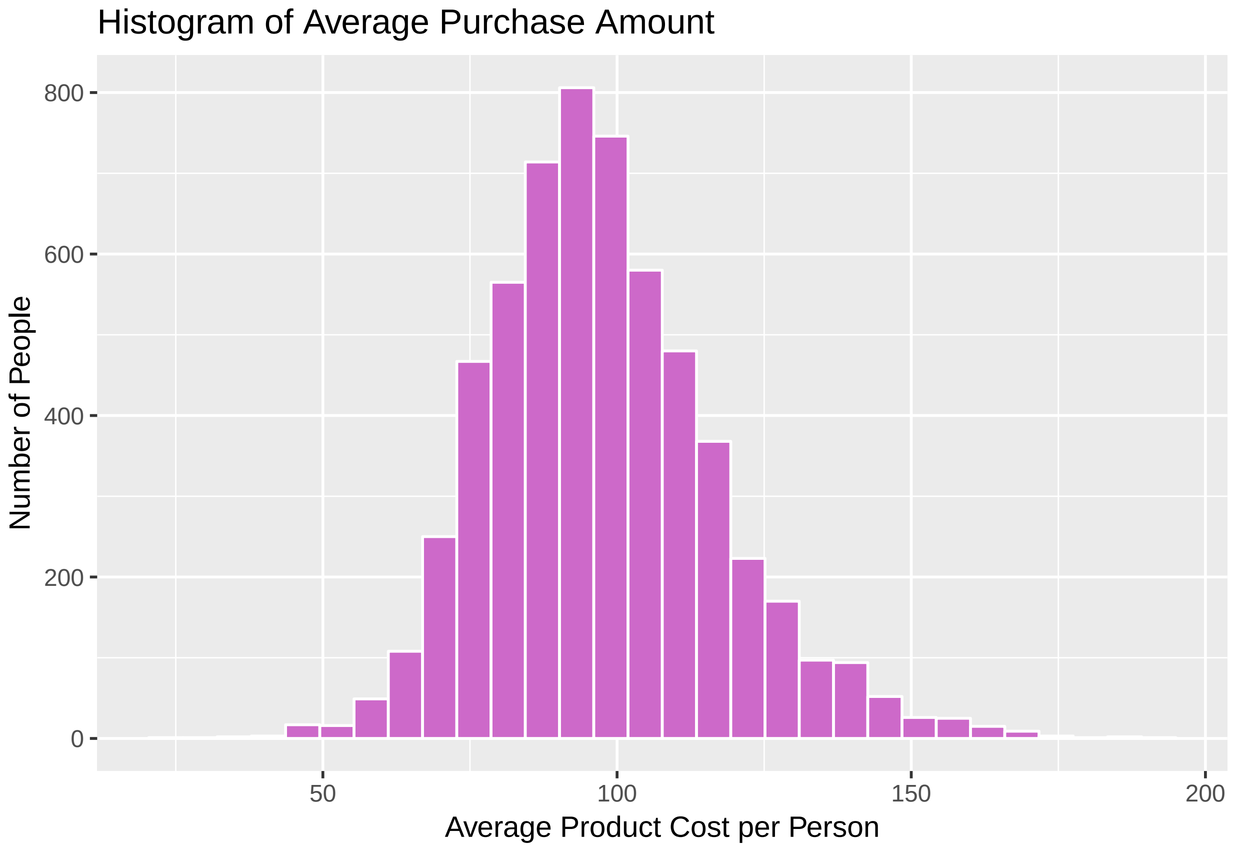
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Individual purchases

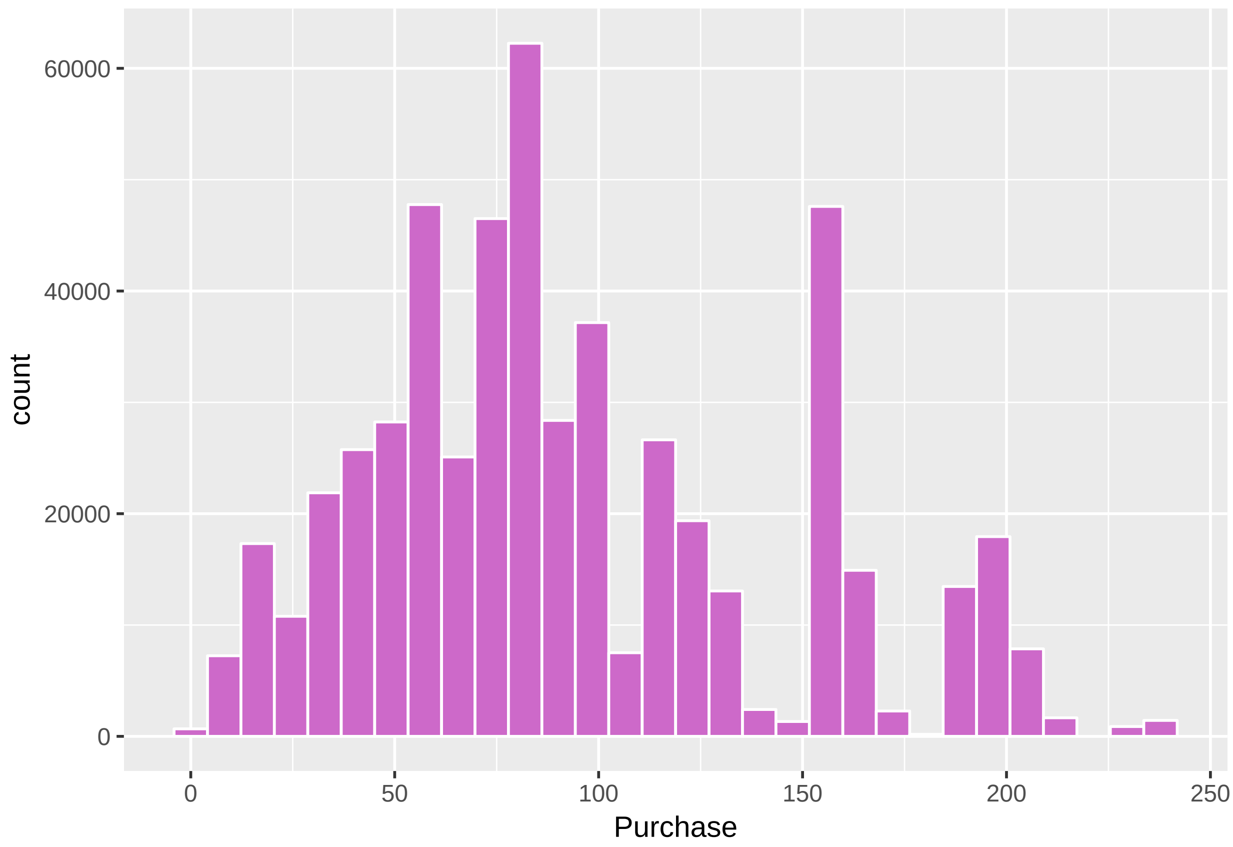
```r
ggplot(unique,aes(Average))+geom_histogram(color = "white", fill = "orchid 3")+
  labs(title = "Histogram of Average Purchase Amount", x = "Average Product Cost per Person", y = "Numbe
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



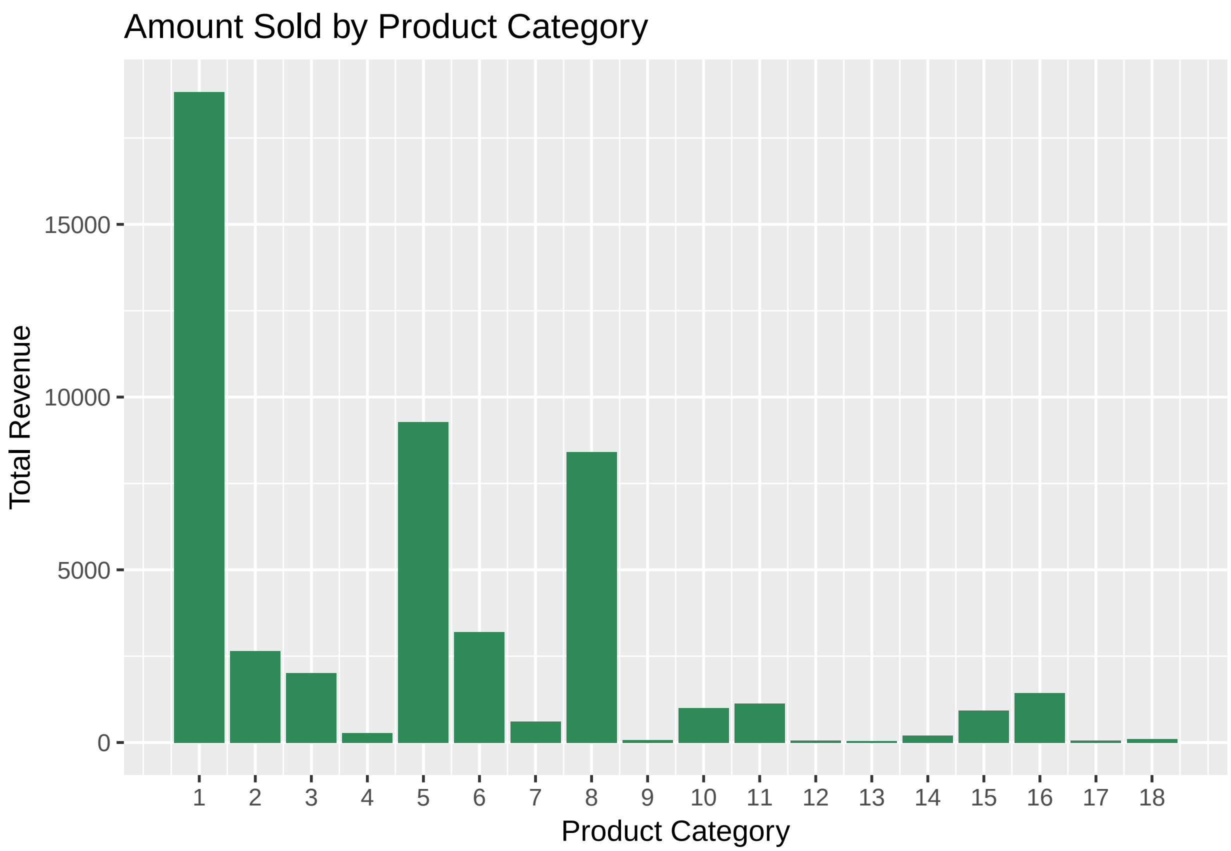
```r
ggplot(blackfriday,aes(Purchase))+geom_histogram(color = "white", fill = "orchid 3")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Products Product categories range from 1 to 18. Some observations have one category listed, while others have two or three. With so many categories of products, I am tempted to think this store is a large department store. On the other hand, we can see from the bar plot that most of the sales fall into a few categories. Grouping by product ID, we can see that there are many popular items. As we would expect, the average age and gender of these top-selling products is in line with our target demographic.

```
ggplot(blackfriday,aes(Product_Category_1,Purchase/1000))+
  geom_col(fill = 'seagreen4')+
  scale_x_continuous(breaks = 1:18)+
  labs(x = "Product Category",y = "Total Revenue", title = "Amount Sold by Product Category")
```

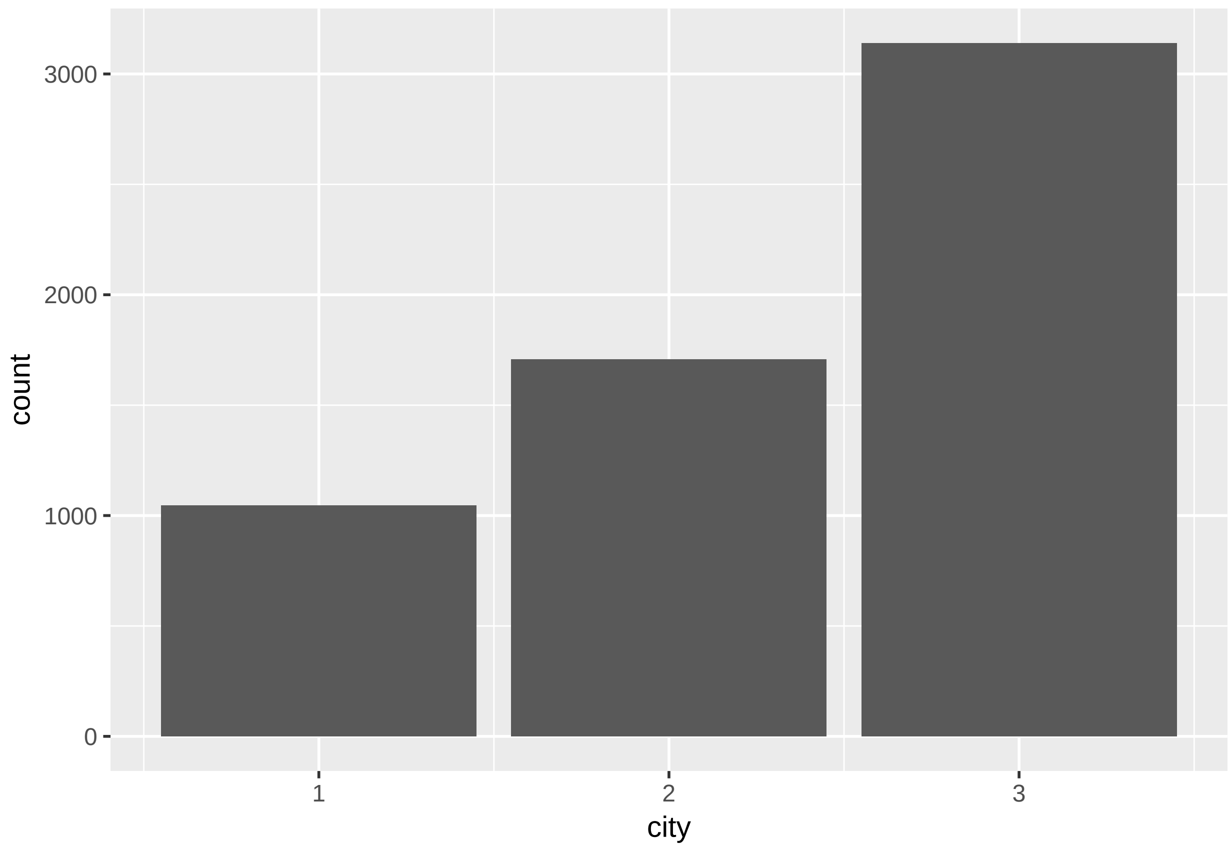


```r
top_products <- blackfriday %>% group_by(Product_ID) %>% summarise(Number=n(),'Male Ratio' = sum(gender

kable(top_products[1:5,1:4,6])
```

Product_ID	Number	Male Ratio	Average Age Category
P00265242	1858	0.7282024	2.382669
P00110742	1591	0.7756128	2.323696
P00025442	1586	0.7849937	2.409836
P00112142	1539	0.7842755	2.261858
P00057642	1430	0.8209790	2.316783

City Finally we have the city variables to investigate. I am not particularly interested since there are only three city categories, and we don't know what the categories represent.

```r
ggplot(unique)+geom_bar(aes(city))
```



```
#Correlation Map
#cor_map <- blackfriday
#cor_map <- mutate()
#cor_map$Purchase <- as.numeric(cor_map$Purchase)
#cor_map <- cor_map[,5:10]
#cor_map[is.na(cor_map)] <- " "
#cormap <- cor(cor_map)

#melted_cormap <- melt(cormap)
#ggplot(data = melted_cormap, aes(x=Var1, y=Var2, fill=value)) +
  #geom_tile()
```

Modeling

```
lm1 <- lm(total ~ factor(gender) + factor(age) + factor(occupation),unique)
summary(lm1)
```

```
##
## Call:
## lm(formula = total ~ factor(gender) + factor(age) + factor(occupation),
##     data = unique)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11237  -5798  -2907   2493  94441
##
## Coefficients:
```

```
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            5688.176   1065.405   5.339 9.70e-08 ***
## factor(gender)1        2268.533    278.881   8.134 5.01e-16 ***
## factor(age)1           1322.302   1046.321   1.264  0.20637    
## factor(age)2           2906.007   1047.580   2.774  0.00555 ** 
## factor(age)3           1938.371   1066.502   1.818  0.06919 .  
## factor(age)4            971.668   1073.721   0.905  0.36553    
## factor(age)5          -1255.676   1157.414  -1.085  0.27801    
## factor(occupation)1    -592.621    539.318  -1.099  0.27189    
## factor(occupation)2       7.308    672.504   0.011  0.99133    
## factor(occupation)3     940.929    790.600   1.190  0.23404    
## factor(occupation)4    -118.623    528.188  -0.225  0.82231    
## factor(occupation)5     690.610    939.927   0.735  0.46252    
## factor(occupation)6    -522.870    706.531  -0.740  0.45930    
## factor(occupation)7   -1162.998    505.070  -2.303  0.02133 *  
## factor(occupation)8    -366.700   2255.059  -0.163  0.87083    
## factor(occupation)9   -1555.711   1056.839  -1.472  0.14106    
## factor(occupation)10 -1436.659   1119.228  -1.284  0.19933    
## factor(occupation)11 -1198.081    886.624  -1.351  0.17666    
## factor(occupation)12 -1794.180    592.805  -3.027  0.00248 ** 
## factor(occupation)13 -1749.151    931.857  -1.877  0.06056 .  
## factor(occupation)14  -625.743    640.661  -0.977  0.32875    
## factor(occupation)15 -1222.681    853.098  -1.433  0.15185    
## factor(occupation)16  1031.702    700.772   1.472  0.14101    
## factor(occupation)17 -1772.625    547.866  -3.236  0.00122 ** 
## factor(occupation)18  -652.305   1178.287  -0.554  0.57987    
## factor(occupation)19  1382.771   1157.171   1.195  0.23215    
## factor(occupation)20  1459.936    657.946   2.219  0.02653 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 9177 on 5864 degrees of freedom
## Multiple R-squared:  0.03669,    Adjusted R-squared:  0.03242 
## F-statistic: 8.591 on 26 and 5864 DF,  p-value: < 2.2e-16
```

```
lm2 <- glm(n ~ factor(gender) + factor(age) + factor(occupation), data = unique, family = poisson)

summary(lm2)
```

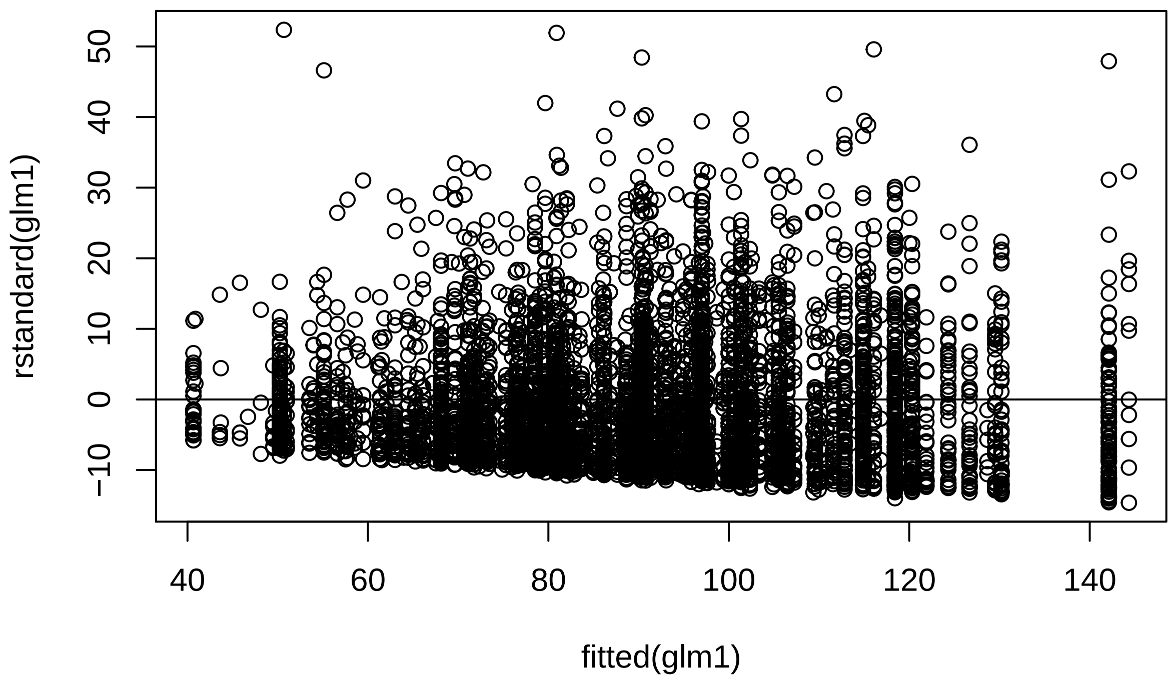
```
## 
## Call:
## glm(formula = n ~ factor(gender) + factor(age) + factor(occupation), 
##     family = poisson, data = unique)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -14.531   -7.946   -4.045    2.393   52.245  
## 
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)            4.212721   0.013018 323.619  < 2e-16 ***
## factor(gender)1        0.211473   0.003308  63.923  < 2e-16 ***
## factor(age)1           0.180900   0.012779  14.156  < 2e-16 ***
## factor(age)2           0.349946   0.012771  27.402  < 2e-16 ***
## factor(age)3           0.234926   0.012994  18.079  < 2e-16 ***
## factor(age)4          0.108932   0.013131   8.296  < 2e-16 ***
## factor(age)5         -0.232725   0.014654 -15.881  < 2e-16 ***
## factor(occupation)1  -0.048334   0.006090  -7.937 2.07e-15 ***
## factor(occupation)2   0.016024   0.007311   2.192  0.02840 *
## factor(occupation)3   0.089666   0.008552  10.485  < 2e-16 ***
## factor(occupation)4  -0.030325   0.005809  -5.220 1.79e-07 ***
## factor(occupation)5   0.048821   0.009916   4.924 8.50e-07 ***
## factor(occupation)6  -0.077894   0.008122  -9.591  < 2e-16 ***
## factor(occupation)7  -0.155414   0.005722 -27.160  < 2e-16 ***
## factor(occupation)8  -0.071874   0.025914  -2.774  0.00554 **
## factor(occupation)9  -0.205188   0.013523 -15.174  < 2e-16 ***
## factor(occupation)10 -0.203082   0.013820 -14.694  < 2e-16 ***
## factor(occupation)11 -0.136138   0.010173 -13.382  < 2e-16 ***
## factor(occupation)12 -0.269376   0.006933 -38.854  < 2e-16 ***
## factor(occupation)13 -0.274780   0.013055 -21.047  < 2e-16 ***
## factor(occupation)14 -0.106072   0.007230 -14.672  < 2e-16 ***
## factor(occupation)15 -0.202771   0.009986 -20.305  < 2e-16 ***
## factor(occupation)16  0.095040   0.007498  12.675  < 2e-16 ***
## factor(occupation)17 -0.265808   0.006396 -41.559  < 2e-16 ***
## factor(occupation)18 -0.059795   0.012995  -4.601 4.20e-06 ***
## factor(occupation)19  0.198059   0.011744  16.865  < 2e-16 ***
## factor(occupation)20  0.182531   0.006727  27.134  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 522424  on 5890  degrees of freedom
## Residual deviance: 498417  on 5864  degrees of freedom
## AIC: 532917
##
## Number of Fisher Scoring iterations: 5
```

Understanding how many items an individual will buy is an important factor when it comes to maximizing the profit of our company. We can clearly see that 26-35 year olds are the target demographic, but can we discover important trends as to which other groups are more likely to buy multiple items? I decided to use a poisson regression to attempt to model the count data of items purchased for each person.

```
glm1 <- glm(n ~ factor(age) + factor(gender)+ factor(occupation), data = unique, family = poisson)

summary(glm1)
```

```
##
## Call:
## glm(formula = n ~ factor(age) + factor(gender) + factor(occupation),
##     family = poisson, data = unique)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -14.531   -7.946   -4.045    2.393   52.245
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           4.212721   0.013018 323.619  < 2e-16 ***
## factor(age)1          0.180900   0.012779  14.156  < 2e-16 ***
## factor(age)2          0.349946   0.012771  27.402  < 2e-16 ***
## factor(age)3          0.234926   0.012994  18.079  < 2e-16 ***
## factor(age)4          0.108932   0.013131   8.296  < 2e-16 ***
## factor(age)5         -0.232725   0.014654 -15.881  < 2e-16 ***
## factor(gender)1       0.211473   0.003308  63.923  < 2e-16 ***
## factor(occupation)1  -0.048334   0.006090  -7.937 2.07e-15 ***
## factor(occupation)2   0.016024   0.007311   2.192  0.02840 *
## factor(occupation)3   0.089666   0.008552  10.485  < 2e-16 ***
## factor(occupation)4  -0.030325   0.005809  -5.220 1.79e-07 ***
## factor(occupation)5   0.048821   0.009916   4.924 8.50e-07 ***
## factor(occupation)6  -0.077894   0.008122  -9.591  < 2e-16 ***
## factor(occupation)7  -0.155414   0.005722 -27.160  < 2e-16 ***
## factor(occupation)8  -0.071874   0.025914  -2.774  0.00554 **
## factor(occupation)9  -0.205188   0.013523 -15.174  < 2e-16 ***
## factor(occupation)10 -0.203082   0.013820 -14.694  < 2e-16 ***
## factor(occupation)11 -0.136138   0.010173 -13.382  < 2e-16 ***
## factor(occupation)12 -0.269376   0.006933 -38.854  < 2e-16 ***
## factor(occupation)13 -0.274780   0.013055 -21.047  < 2e-16 ***
## factor(occupation)14 -0.106072   0.007230 -14.672  < 2e-16 ***
## factor(occupation)15 -0.202771   0.009986 -20.305  < 2e-16 ***
## factor(occupation)16  0.095040   0.007498  12.675  < 2e-16 ***
## factor(occupation)17 -0.265808   0.006396 -41.559  < 2e-16 ***
## factor(occupation)18 -0.059795   0.012995  -4.601 4.20e-06 ***
## factor(occupation)19  0.198059   0.011744  16.865  < 2e-16 ***
## factor(occupation)20  0.182531   0.006727  27.134  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 522424  on 5890  degrees of freedom
## Residual deviance: 498417  on 5864  degrees of freedom
## AIC: 532917
##
## Number of Fisher Scoring iterations: 5
```

```
plot(fitted(glm1),rstandard(glm1));abline(h=0)
```



```
pchisq(glm1$deviance, df=glm1$df.residual, lower.tail=FALSE)
```

```
## [1] 0
```