Exploratory Analysis

Venukanan Subenthiran

2022-11-12

Explorartoy Analysis

Sub-setting the Data

```
fraudTotal.db_fraud <- subset(fraudTotal.db, is_fraud == 1)
nrow(fraudTotal.db_fraud)
```

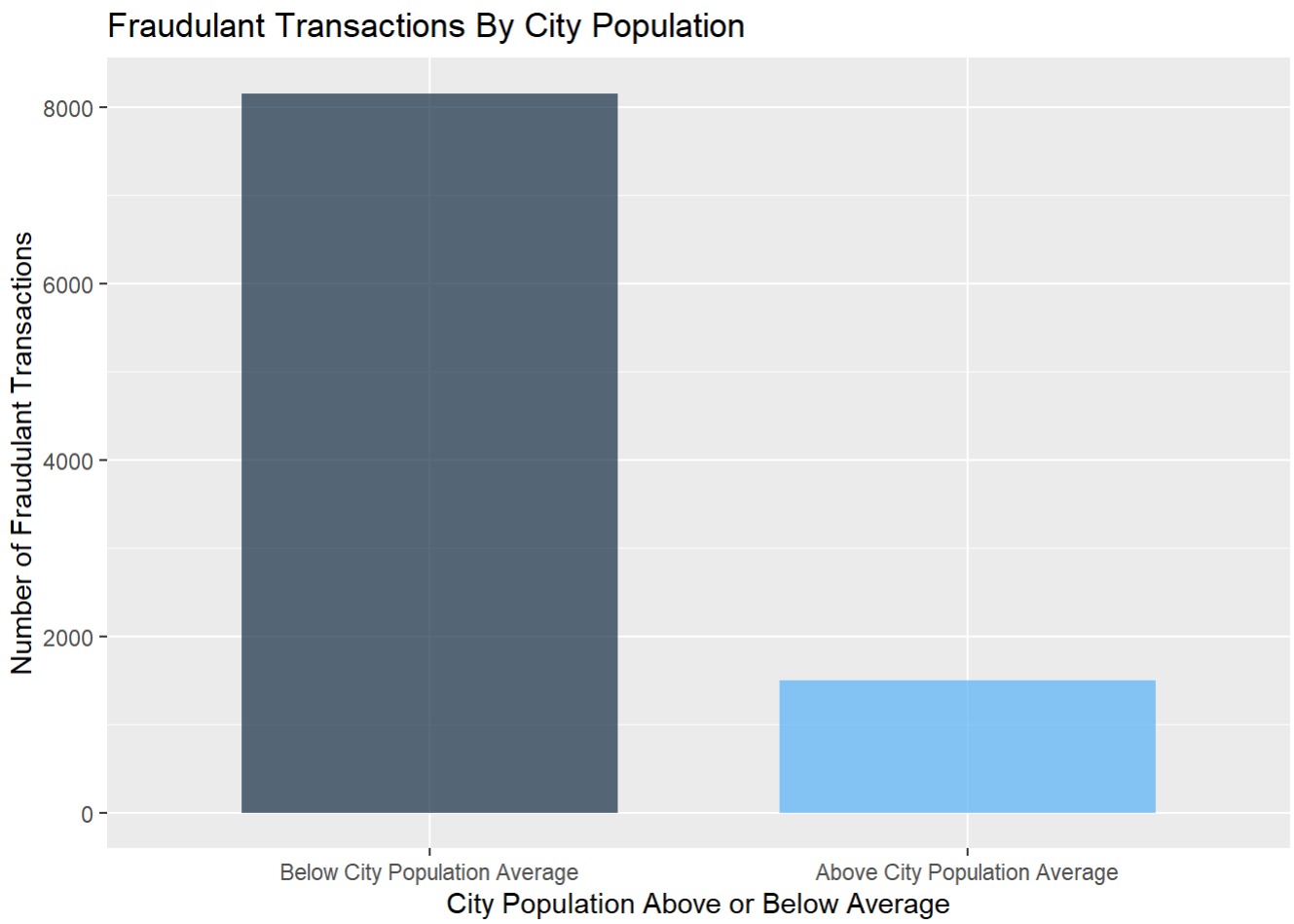
```
## [1] 9651
```

Does Having a Higher City Population Have More Fraud?

```
fraudTotal.db_fraud$city_pop <- ifelse(fraudTotal.db_fraud$city_pop < mean(fraudTotal.db_fraud$city_pop), 0, 1)

library(ggplot2)

fraud_based_on_city_pop <- ggplot(data = fraudTotal.db_fraud, aes(x = factor(city_pop), fill = city_pop)) + geom_bar(stat = "count", width = 0.7, alpha = 0.7) + ggtitle("Fraudulant Transactions By City Population") + xlab("City Population Above or Below Average") + ylab("Number of Fraudulant Transactions") + scale_x_discrete(labels = c("Below City Population Average", "Above City Population Average")) + theme(legend.position = "none")
fraud_based_on_city_pop
```



What type merchants receive more fraudulant transactions?

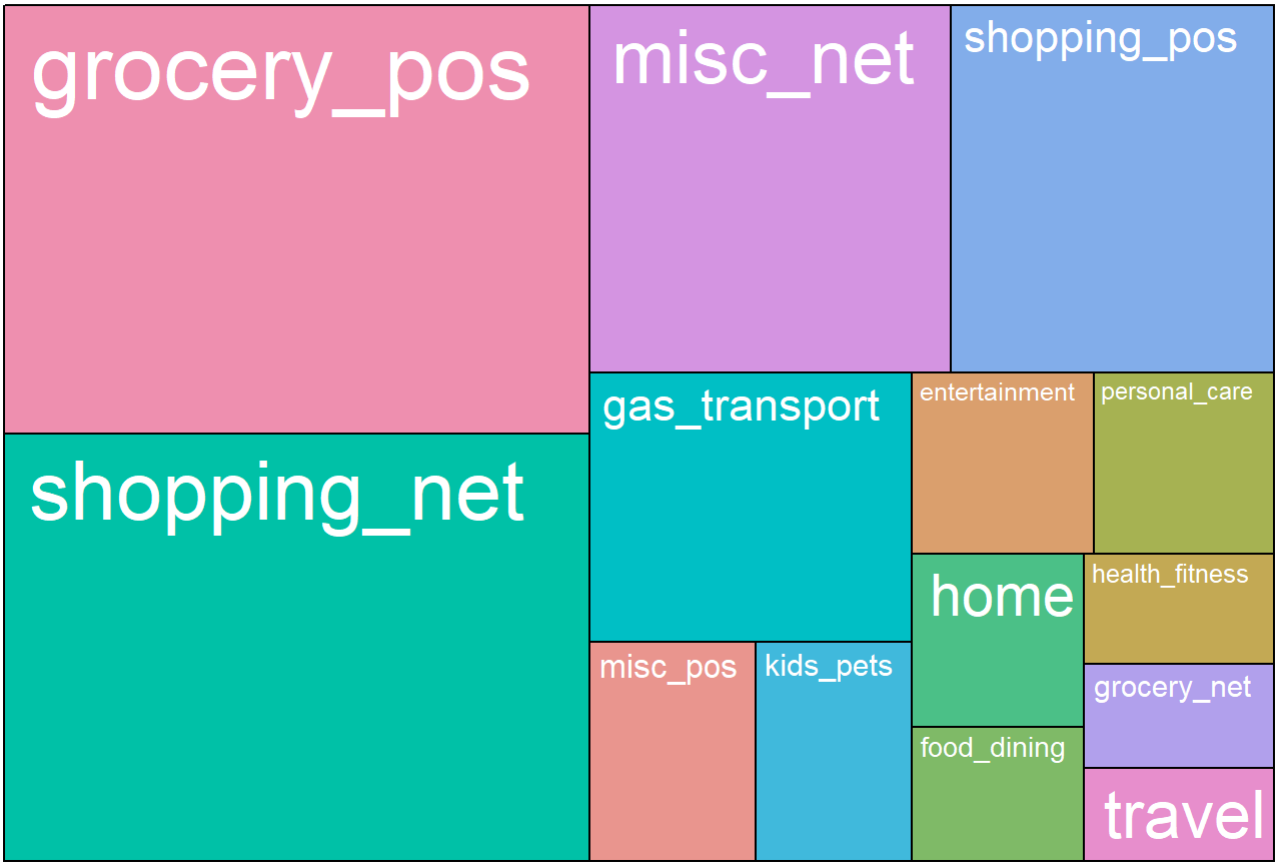
```
category_table <- table(fraudTotal.db_fraud$category)
category_table
```

```
##
##  entertainment    food_dining  gas_transport    grocery_net    grocery_pos
##            292            205            772            175           2228
## health_fitness           home      kids_pets       misc_net       misc_pos
##            185            265            304           1182            322
##  personal_care   shopping_net   shopping_pos         travel
##            290           2219           1056            156
```

```
category_fraud.db <- as.data.frame(category_table)
colnames(category_fraud.db)[1] <- "Category Type"

#install.packages("treemap")
library(treemap)
treemap(category_fraud.db, index = "Category Type", vSize = "Freq", type = "index", title = "Mer
chant Category Types in Fraudulant Transactions", palette = "HCL", border.col = c("black"), bord
er.lwds = 1, fontsize.labels = 10, fontcolor.labels = "white", fontface.labels = 1, bg.labels =
 c(0), align.labels = c("left", "top"), overlap.labels = 0.5, inflate.labels = T)
```

Merchant Category Types in Fraudulant Transactions



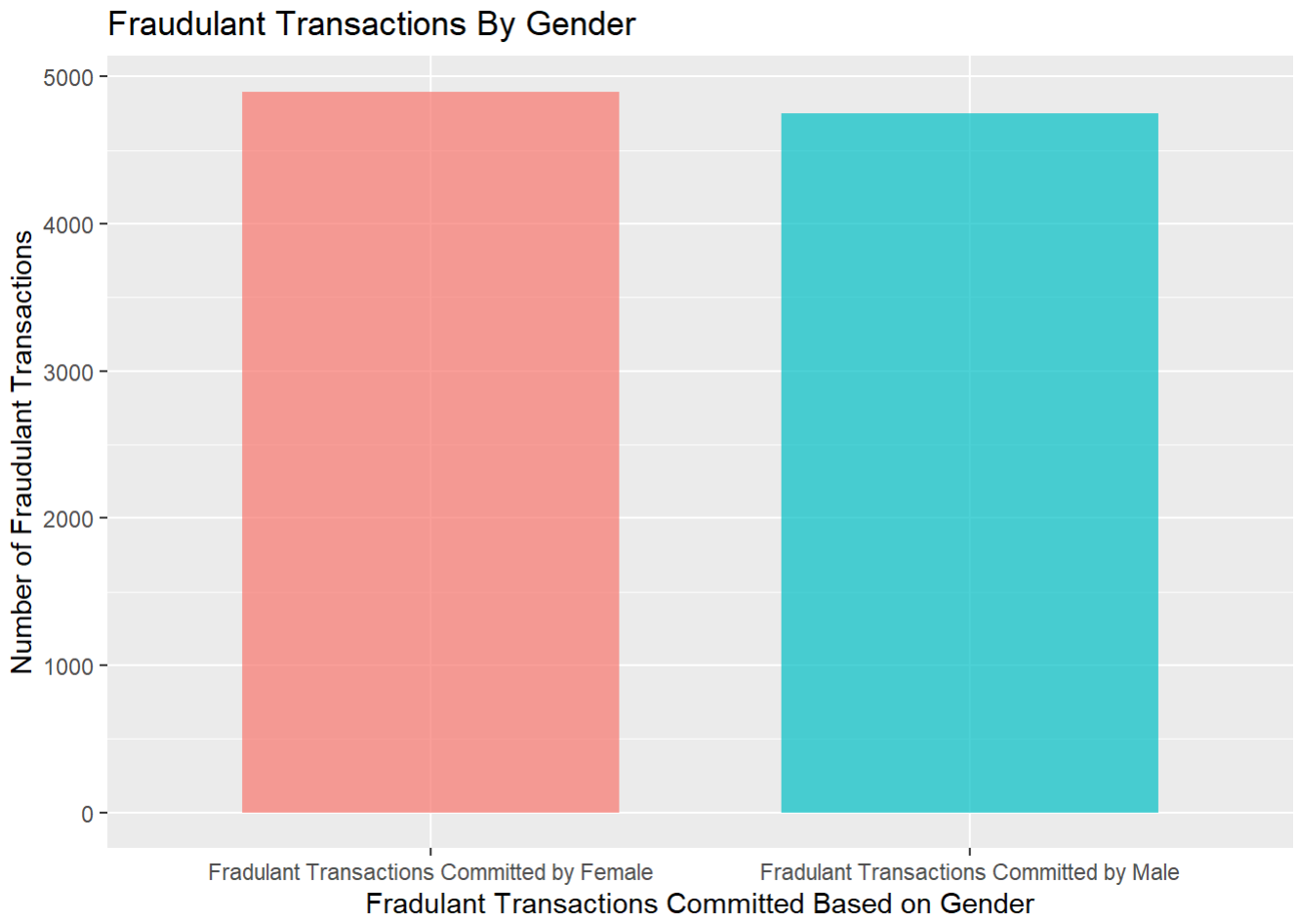
Are fraud transactions commited more by males of females?

```
gender_table <- table(fraudTotal.db_fraud$gender)
gender_fraud.db <- as.data.frame(gender_table)
colnames(gender_fraud.db)[1] <- "Gender"



library(ggplot2)

fraud_based_on_gender <- ggplot(data = gender_fraud.db, aes(x = Gender, y = Freq , fill = Gende
r)) + geom_bar(stat = "identity", width = 0.7, alpha = 0.7) + ggtitle("Fraudulant Transactions B
y Gender") + xlab("Fradulant Transactions Committed Based on Gender") + ylab("Number of Fraudula
nt Transactions") + scale_x_discrete(labels = c("Fradulant Transactions Committed by Female", "F
radulant Transactions Committed by Male")) + theme(legend.position = "none")
fraud_based_on_gender
```



Normalization of Numeric Values

Applying normalization to numeric variables

```
data2 <- fraudTotal.db[c(3, 6, 13, 14, 15, 16, 20, 21, 22, 23)]
summary(data2)
```

```
##      cc_num               amt                zip              lat
##  Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000361   1st Qu.:0.0002985   1st Qu.:0.2532   1st Qu.:0.3138
##  Median :0.0007054   Median :0.0016046   Median :0.4755   Median :0.4142
##  Mean   :0.0836052   Mean   :0.0023858   Mean   :0.4820   Mean   :0.3967
##  3rd Qu.:0.0009299   3rd Qu.:0.0028361   3rd Qu.:0.7174   3rd Qu.:0.4696
##  Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000   Max.   :1.0000
##       long           city_pop           unix_time        merch_lat
##  Min.   :0.0000   Min.   :0.0000000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.7048   1st Qu.:0.0002470   1st Qu.:0.2793   1st Qu.:0.3241
##  Median :0.8002   Median :0.0008326   Median :0.5021   Median :0.4196
##  Mean   :0.7720   Mean   :0.0304887   Mean   :0.5272   Mean   :0.4024
##  3rd Qu.:0.8751   3rd Qu.:0.0069856   3rd Qu.:0.7791   3rd Qu.:0.4729
##  Max.   :1.0000   Max.   :1.0000000   Max.   :1.0000   Max.   :1.0000
##    merch_long        is_fraud
##  Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.6997   1st Qu.:0.00000
##  Median :0.7945   Median :0.00000
##  Mean   :0.7666   Mean   :0.00521
##  3rd Qu.:0.8667   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000
```

```
#install.packages("caret")
library(caret)
```

```
## Loading required package: lattice
```

```
fraudTotal.db_process <- preProcess(as.data.frame(data2), method = c("range"))

fraudTotal.db_norm <- predict(fraudTotal.db_process, as.data.frame(data2))
```

Merging the fraudTotal.db and fraud.db_norm.

```
fraudTotal.db$cc_num <- fraudTotal.db_norm$cc_num
fraudTotal.db$amt <- fraudTotal.db_norm$amt
fraudTotal.db$zip <- fraudTotal.db_norm$zip
fraudTotal.db$lat <- fraudTotal.db_norm$lat
fraudTotal.db$long <- fraudTotal.db_norm$long
fraudTotal.db$city_pop <- fraudTotal.db_norm$city_pop
fraudTotal.db$unix_time <- fraudTotal.db_norm$unix_time
fraudTotal.db$merch_lat <- fraudTotal.db_norm$merch_lat
fraudTotal.db$merch_long <- fraudTotal.db_norm$merch_long
fraudTotal.db$is_fraud <- fraudTotal.db_norm$is_fraud
```