Bivariate Analysis

Venukanan Subenthiran

2022-11-12

Bivariate Analysis of fraudTotal.db data frame

```
sum(is.na(fraudTotal.db))
```

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

```
str(fraudTotal.db)
```

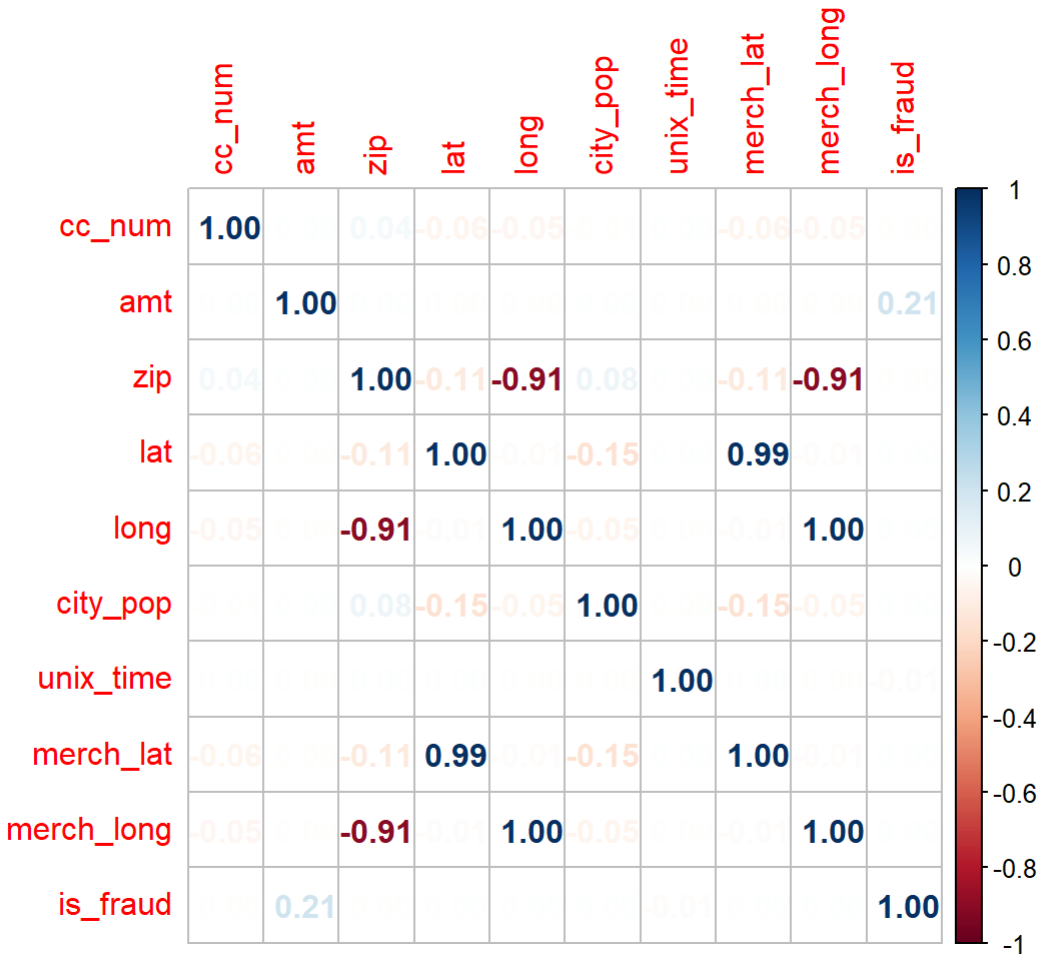
```
## 'data.frame':    1852394 obs. of  23 variables:
##  $ X                    : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ trans_date_trans_time: POSIXct, format: "2019-01-01 00:00:18" "2019-01-01 00:00:44" ...
##  $ cc_num               : num  2.70e+15 6.30e+11 3.89e+13 3.53e+15 3.76e+14 ...
##  $ merchant             : Factor w/ 693 levels "fraud_Abbott-Rogahn",..: 516 244 390 364 298
608 534 108 251 565 ...
##  $ category             : Factor w/ 14 levels "entertainment",..: 9 5 1 3 10 3 4 3 10 5 ...
##  $ amt                  : num  4.97 107.23 220.11 45 41.96 ...
##  $ first                : Factor w/ 355 levels "Aaron","Adam",..: 165 313 117 166 340 165 202
315 147 242 ...
##  $ last                 : Factor w/ 486 levels "Abbott","Adams",..: 19 162 387 469 154 85 365
473 73 4 ...
##  $ gender               : Factor w/ 2 levels "F","M": 1 1 2 2 2 1 1 2 1 1 ...
##  $ street               : Factor w/ 999 levels "000 Jennifer Mills",..: 577 440 611 946 423 4
78 897 229 697 218 ...
##  $ city                 : Factor w/ 906 levels "Achille","Acworth",..: 533 620 475 85 218 225
355 238 481 150 ...
##  $ state                : Factor w/ 51 levels "AK","AL","AR",..: 28 48 14 27 46 39 17 46 39 4
3 ...
##  $ zip                  : int  28654 99160 83252 59632 24433 18917 67851 22824 15665 37040
...
##  $ lat                  : num  36.1 48.9 42.2 46.2 38.4 ...
##  $ long                 : num  -81.2 -118.2 -112.3 -112.1 -79.5 ...
##  $ city_pop             : int  3495 149 4154 1939 99 2158 2691 6018 1472 151785 ...
##  $ job                  : Factor w/ 497 levels "Academic librarian",..: 373 432 309 331 117 4
83 30 128 378 332 ...
##  $ dob                  : Date, format: "1988-03-09" "1978-06-21" ...
##  $ trans_num            : Factor w/ 1852394 levels "00000ecad06b03d3a8d34b4e30b5ce3b",..: 803
27 227463 1169031 777910 1186867 177885 954104 789719 1824660 430621 ...
##  $ unix_time            : int  1325376018 1325376044 1325376051 1325376076 1325376186 1325376
248 1325376282 1325376308 1325376318 1325376361 ...
##  $ merch_lat            : num  36 49.2 43.2 47 38.7 ...
##  $ merch_long           : num  -82 -118.2 -112.2 -112.6 -78.6 ...
##  $ is_fraud             : int  0 0 0 0 0 0 0 0 0 0 ...
```

The correlation between numeric attributes

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

```
## corrplot 0.92 loaded
```

```
subset_fraudTotal.db <- subset(fraudTotal.db, select = c(3, 6, 13, 14, 15, 16, 20, 21, 22, 23))
corrplot(cor(subset_fraudTotal.db), method = "number")
```



Correlation Analysis of Numeric Variables

Correlation of zip, lat, long, merch_lat, and merch_long

```
cor.test(fraudTotal.db$zip, fraudTotal.db$lat, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  fraudTotal.db$zip and fraudTotal.db$lat
## t = -156.94, df = 1852392, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1159749 -0.1131326
## sample estimates:
##       cor
## -0.114554
```

```
cor.test(fraudTotal.db$zip, fraudTotal.db$long, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  fraudTotal.db$zip and fraudTotal.db$long
## t = -2983.3, df = 1852392, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9100423 -0.9095462
## sample estimates:
##       cor
## -0.9097946
```

```
cor.test(fraudTotal.db$zip, fraudTotal.db$merch_lat, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  fraudTotal.db$zip and fraudTotal.db$merch_lat
## t = -156.08, df = 1852392, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1153553 -0.1125126
## sample estimates:
##       cor
## -0.1139342
```

```
cor.test(fraudTotal.db$zip, fraudTotal.db$merch_long, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  fraudTotal.db$zip and fraudTotal.db$merch_long
## t = -2967.9, df = 1852392, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9092314 -0.9087309
## sample estimates:
##       cor
## -0.9089815
```

Correlation of is_fraud and amt

```
cor.test(fraudTotal.db$is_fraud, fraudTotal.db$amt, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  fraudTotal.db$is_fraud and fraudTotal.db$amt
## t = 291.33, df = 1852392, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2079305 0.2106844
## sample estimates:
##       cor
## 0.2093078
```

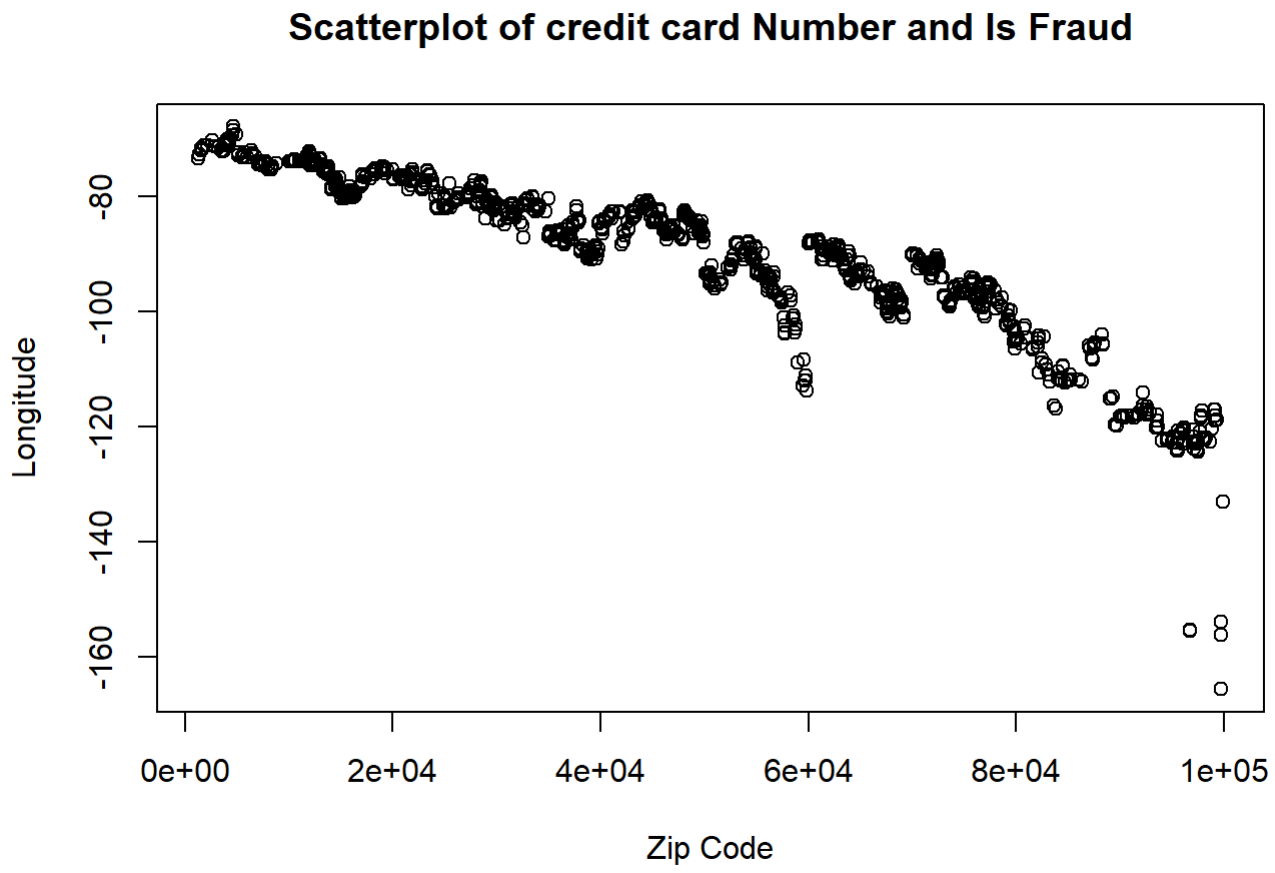
Correlation of is_fraud and city_pop

```
cor.test(fraudTotal.db$is_fraud, fraudTotal.db$city_pop, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  fraudTotal.db$is_fraud and fraudTotal.db$city_pop
## t = 0.4426, df = 1852392, p-value = 0.6581
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.001114869  0.001765256
## sample estimates:
##          cor
## 0.0003251944
```

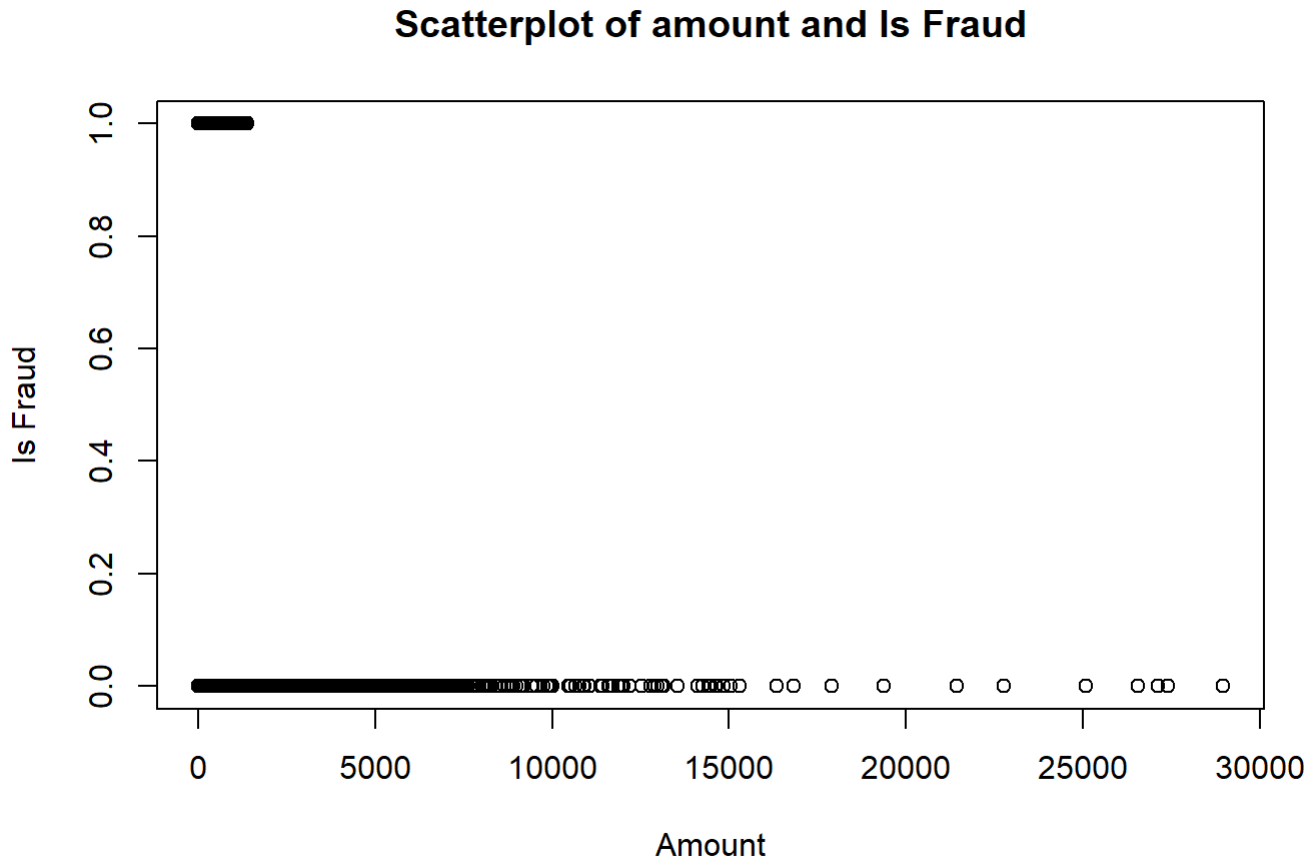
Scatterplot of zip and long

```
plot(fraudTotal.db$zip, fraudTotal.db$long, main = "Scatterplot of credit card Number and Is Fra
ud", xlab = "Zip Code", ylab = "Longitude")
```



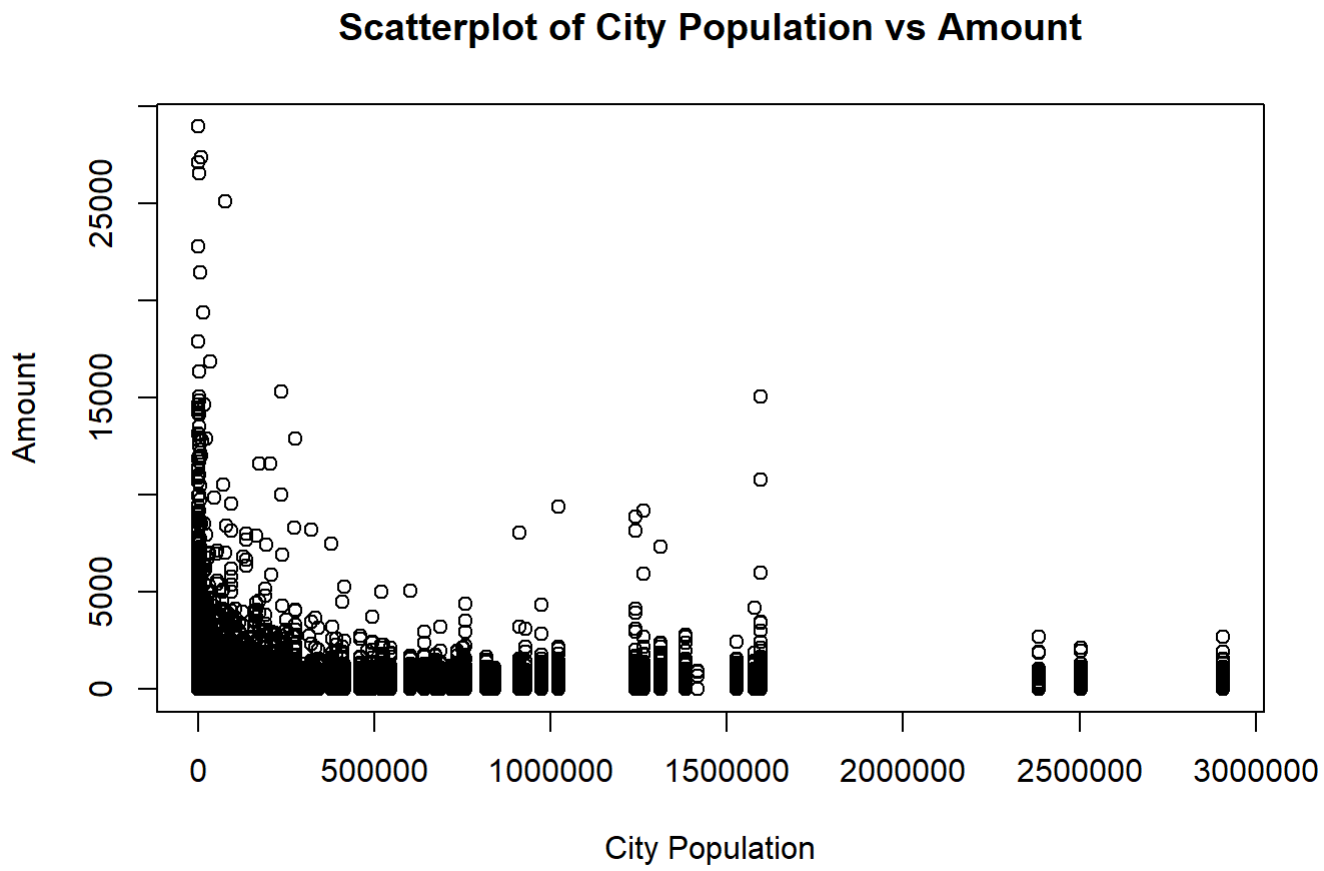
Scatterplot of amt and is_fraud

```
plot(fraudTotal.db$amt, fraudTotal.db$is_fraud, main = "Scatterplot of amount and Is Fraud", xla
b = "Amount", ylab = "Is Fraud")
```



Scatterplot of Amt and city_pop

```
plot(fraudTotal.db$city_pop, fraudTotal.db$amt, main = "Scatterplot of City Population vs Amoun
t", xlab = "City Population", ylab = "Amount")
```



Scatterplot of lat and is_fraud

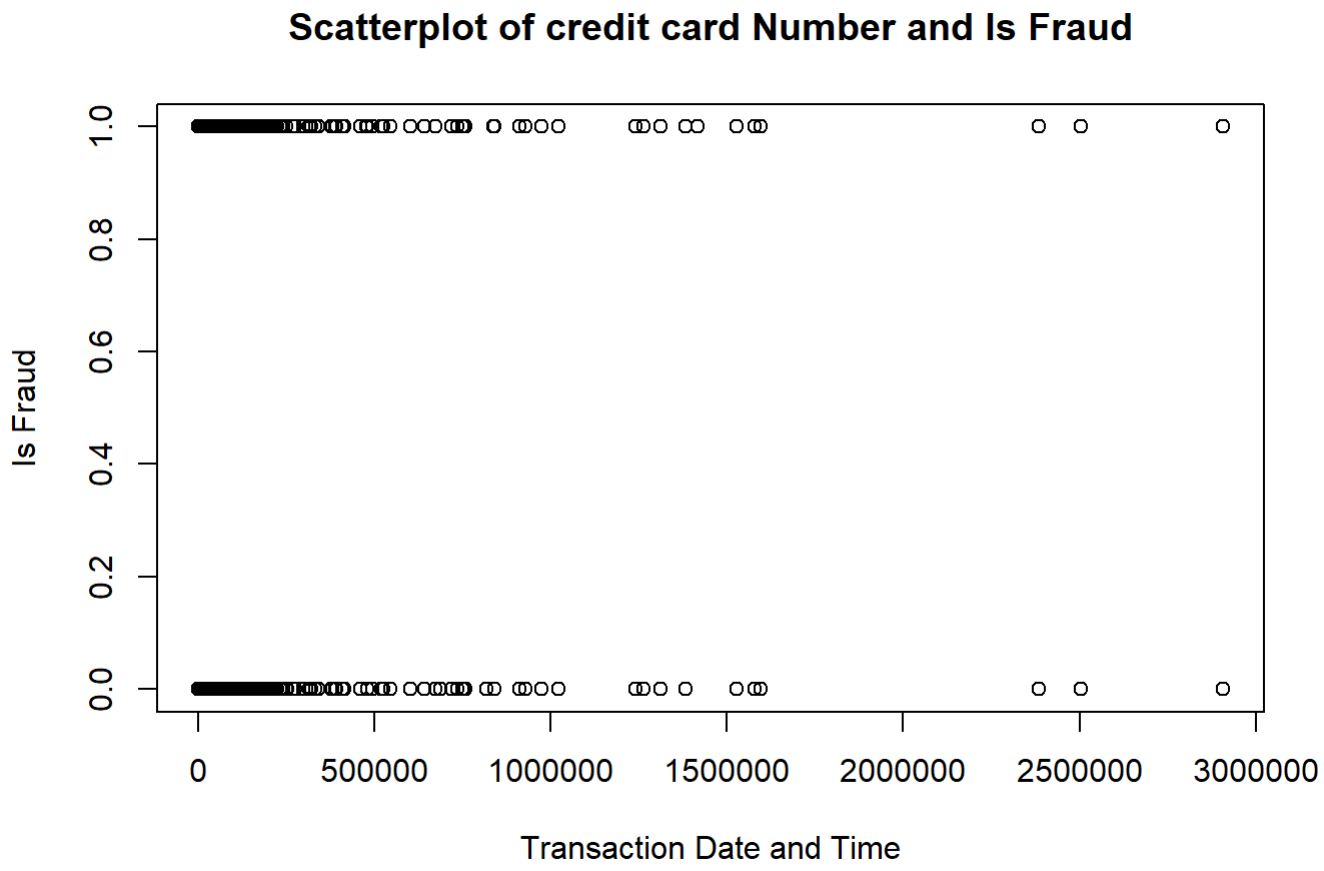
```
#plot(fraudTotal.db$lat, fraudTotal.db$is_fraud, main = "Scatterplot of credit card Number and I
s Fraud", xlab = "Transaction Date and Time", ylab = "Is Fraud")
```

Scatterplot of long and is_fraud

```
#plot(fraudTotal.db$long, fraudTotal.db$is_fraud, main = "Scatterplot of credit card Number and
 Is Fraud", xlab = "Transaction Date and Time", ylab = "Is Fraud")
```

Scatterplot of city_pop and is_fraud

```
plot(fraudTotal.db$city_pop, fraudTotal.db$is_fraud, main = "Scatterplot of credit card Number a
nd Is Fraud", xlab = "Transaction Date and Time", ylab = "Is Fraud")
```



Scatterplot of unix_time and is_fraud

```
#plot(fraudTotal.db$unix_time, fraudTotal.db$is_fraud, main = "Scatterplot of credit card Number and Is Fraud", xlab = "Transaction Date and Time", ylab = "Is Fraud")
```

Scatterplot of merch_lat and is_fraud

```
#plot(fraudTotal.db$merch_lat, fraudTotal.db$is_fraud, main = "Scatterplot of credit card Number and Is Fraud", xlab = "Transaction Date and Time", ylab = "Is Fraud")
```

Scatterplot of merch_long and is_fraud

```
#plot(fraudTotal.db$merch_long, fraudTotal.db$is_fraud, main = "Scatterplot of credit card Number and Is Fraud", xlab = "Transaction Date and Time", ylab = "Is Fraud")
```

Correlation of Categorical Variables

Chi Square test of merchant and category

```
chisq.test(fraudTotal.db$merchant, fraudTotal.db$category)
```

```
##
##  Pearson's Chi-squared test
##
## data:  fraudTotal.db$merchant and fraudTotal.db$category
## X-squared = 23821955, df = 8996, p-value < 2.2e-16
```

Chi Square test of first and last

```
chisq.test(fraudTotal.db$first, fraudTotal.db$last)
```

```
## Warning in chisq.test(fraudTotal.db$first, fraudTotal.db$last): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  fraudTotal.db$first and fraudTotal.db$last
## X-squared = 331405447, df = 171690, p-value < 2.2e-16
```

Chi Square test of first and gender

```
chisq.test(fraudTotal.db$first, fraudTotal.db$gender)
```

```
## Warning in chisq.test(fraudTotal.db$first, fraudTotal.db$gender): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  fraudTotal.db$first and fraudTotal.db$gender
## X-squared = 1852394, df = 354, p-value < 2.2e-16
```

Chi Square test of street and city

```
chisq.test(fraudTotal.db$street, fraudTotal.db$city)
```

```
## Warning in chisq.test(fraudTotal.db$street, fraudTotal.db$city): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  fraudTotal.db$street and fraudTotal.db$city
## X-squared = 1676416570, df = 903190, p-value < 2.2e-16
```

Chi Square test of city and state

```
chisq.test(fraudTotal.db$city, fraudTotal.db$state)
```

```
## Warning in chisq.test(fraudTotal.db$city, fraudTotal.db$state): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  fraudTotal.db$city and fraudTotal.db$state
## X-squared = 87829535, df = 45250, p-value < 2.2e-16
```

Chi Square test of category and job

```
chisq.test(fraudTotal.db$category, fraudTotal.db$job)
```

```
## Warning in chisq.test(fraudTotal.db$category, fraudTotal.db$job): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  fraudTotal.db$category and fraudTotal.db$job
## X-squared = 62987, df = 6448, p-value < 2.2e-16
```

Chi Square test of gender and job

```
chisq.test(fraudTotal.db$gender, fraudTotal.db$job)
```

```
## Warning in chisq.test(fraudTotal.db$gender, fraudTotal.db$job): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  fraudTotal.db$gender and fraudTotal.db$job
## X-squared = 1034523, df = 496, p-value < 2.2e-16
```

Chi Square test of category and trans_num

```
chisq.test(fraudTotal.db$category, fraudTotal.db$trans_num)
```

```
## Warning in chisq.test(fraudTotal.db$category, fraudTotal.db$trans_num): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  fraudTotal.db$category and fraudTotal.db$trans_num
## X-squared = 24081122, df = 24081109, p-value = 0.4992
```