---

title: "Trends of Fraud in the US"

output: pdf\_document

---

Prepared by Davin Kaing, Nikhita Kandikonda, and Praneeth Tamvada

# Abstract

On average, fraud costs approximately 5% of an organization annual revenue [1]. This burden not only impacts the economy overall, it threatens the security of information. To better understand this issue, we examined the trend of various types of frauds, their economic costs, and the targetted age group. We built predictive models to forcast these factors into the future and present recommendations on how to best approach this problem.

# Background Research & Questions

As we enter the exciting era of big data, there is an enormous amount of information and tools that can help drive effective decisions. However, one of the biggest concerns is fraud. According to a survey conducted by Ernst & Young [2], one-third of global organizations aren't prepared for cyber breach. This poses a concern for the security of confidential data and it affects many organizations ranging from pharmaceutical industry to financial services. In this paper, we focus on the impact of fraud in the United States. We use linear regression to examine the trends of the last 6 years (2010-2015) of fraud data retrieved from Federal Trade Commission website [3]. We then used these trends to project into the next 6 years (2016-2021). We also examine the economic costs of fraud and the targeted age groups.

#S.M.A.R.T Questions

Our objective is to answer the following questions: What is the past and future trend of the different types of fraud for the different states? What is the past and future trend of the economic costs of fraud in the US? What age group(s) is most targeted in the future? Our questions follow the SMART model:

## Specific

- Who? People of the US

- What? Impact of frauds

- Where? U.S.

- When? Using the last 6 years (2010 - 2015) to predict the next 6 years (2016 - 2020).

## Measurable

- Data for the number frauds per state are available, along with economic loss data

## Attainable

- Data are available

## Realistic

- Data are collected by the Federal Trade Commission

## Timely

- Can be completed in a timely manner

# Tools Used

This project is completed in both R and Python. We apply linear regression to build our models. To visualize our results, we use ggplot for both R and Python. For our analysis in python, we use the following packages: pandas, numpy, statsmodels, ggplot, collections, matplotlib, and warnings.

# Analysis

##Fraud Data

```{r fraud data}

library(ggplot2)

fraud\_df <- read.csv("~/Documents/Master's Program/Assignments/Intro to Data Science First Project/fraud data.csv")

fraud\_df = fraud\_df[, c("year", "state", "categories", "complaints")]

# number of observations and variables

cat("The dimension of the fraud data is", dim(fraud\_df))

```

``` {r fraud data analysis}

# data is measuring the number of complaints by categories, state,

# and year

# mean, mode, median of each variable

summary(fraud\_df)

find\_mode <- function(vector) {

mode <- vector[which.max(table(vector))]

return(mode)

}

cat("Mode of complaints is ", find\_mode(fraud\_df$complaints))

```

``` {r fraud data plots}

# plot of complaints in California

CA <- fraud\_df[fraud\_df$state == "California",]

ggplot(CA, aes(x = year, y = complaints, color = categories)) +

geom\_line() +ggtitle("Plot of Complaints in California")

```

```{r fraud data without transformation}

# visualizing complaints without transformation

ggplot(fraud\_df,aes(complaints)) +geom\_histogram(binwidth = 10000,fill="navyBlue")+

ggtitle("Histogram of Complaints")

```

```{r fraud data with transformation}

# data transformation

new\_data <- fraud\_df[, c("year", "state", "categories", "complaints")]

new\_data$complaints <- log(new\_data$complaints)

ggplot(new\_data, aes(complaints))+geom\_histogram(binwidth = 0.5,fill="navyBlue")+

xlab("log(complaints)") +ggtitle("Histogram of Log of Complaints")

```

```{r}

new\_CA <- new\_data[new\_data$state== "California",]

ggplot(new\_CA, aes(x = year, y = complaints, color = categories))+

geom\_line()+ggtitle("Trend of Complaints in California")+

ylab("log(complaints)")

```

```{r fraud data regression model 1}

fraud\_model <- lm(complaints ~., data = new\_data)

summary(fraud\_model)$r.squared

```

```{r fraud data regression model plot 1}

ggplot(new\_CA, aes(x = year, y = complaints, color = categories)) +

geom\_point() + geom\_smooth(method = "lm", se = F) +

ylab("log(complaints)")

```

```{r fraud data regression model 2}

m\_2 <- lm(complaints ~ year + state+categories + year:categories,

data = new\_data)

summary(m\_2)$r.squared

```

```{r fraud data regression model plot 2}

predict\_data <- new\_data[,1:3]

predict\_data$year <- predict\_data$year + 6

## visualize predictions

model <- m\_2

predict\_data$complaints <- predict(model, predict\_data)

nap <- rbind(new\_data, predict\_data)

CA <- nap[nap$state == "California", ]

CA$complaints <- exp(CA$complaints)

ggplot(CA, aes(x = year, y = complaints, color = categories))+

geom\_line()+ggtitle("Projected Trends of Different Frauds")

```

```{r}

nap$complaints <- exp(nap$complaints)

nap\_summ <- aggregate(complaints~year, nap, sum)

ggplot(nap\_summ, aes(x = year, y = complaints/1e6, group = year)) +

geom\_bar(stat= "identity", fill = "navyblue", color = "darkred")+

ylab("complaints (in million)") +

ggtitle("Projected Complaints in the U.S.")

```

## Economic loss Data

```{r Economic cost data}

# economic cost of fraud

econ\_df<- read.csv("~/Documents/Master's Program/Assignments/Intro to Data Science First Project/Economic Loss.csv")

econ\_df <- econ\_df[,c(1,2,4)]

colnames(econ\_df) <- c("year", "state" , "cost")

econ\_df$cost <- as.numeric(as.character(econ\_df$cost))

summary(econ\_df)

cat("Mode of cost is ", find\_mode(econ\_df$cost))

cat("The dimension of the economic loss data is", dim(econ\_df))

```

```{r plotting}

ggplot(econ\_df, aes(cost/(1e6)))+geom\_histogram(binwidth = 10,fill="navyBlue")+

xlab("cost (in million)")+

ggtitle("Histogram of Economic Cost of Fraud")

```

```{r plot log(cost)}

econ\_df$cost <- log(econ\_df$cost)

ggplot(econ\_df, aes(cost))+geom\_histogram(binwidth = 0.5,fill="navyBlue")+

xlab("log(cost)") +ggtitle("Histogram of Log of Cost")

```

```{r regression model}

# economic loss model

econ\_model <- lm(cost~., data = econ\_df)

summary(econ\_model)$r.squared

```

```{r visualization}

econ\_pred <- econ\_df[,1:2]

econ\_pred$year <- econ\_pred$year +6

econ\_pred$cost <- exp(predict(econ\_model, econ\_pred))

econ\_df$cost <- exp(econ\_df$cost)

combined\_pred\_orig <- rbind(econ\_df, econ\_pred)

ggplot(combined\_pred\_orig, aes(year, cost/(1e6), color = state))+

ylab("cost (in million)") +geom\_line() +

ggtitle("Projected Economic Cost of Fraud")

```

## Frauds Complaints by Age data

```{r age data}

age <- read.csv("~/Documents/Master's Program/Assignments/Intro to Data Science First Project/fraud complaints by age.csv")

age <- age[,1:3]

colnames(age) <- c("year", "age", "complaints")

summary(age)

cat("Mode of complaints is ", find\_mode(age$complaints))

cat("The dimension of the age data is", dim(age))

```

```{r plot for age data}

ggplot(age, aes(factor(age),complaints))+geom\_bar(stat="identity",aes(fill=year))+

xlab("Age") +ggtitle("Histogram of Fraud complaints by age")

```

```{r regression model for age data}

age\_model <- lm(complaints~year+age+age\*year, data = age)

summary(age\_model)$r.squared

```

```{r visualization for age data}

age\_pred <- age[,1:2]

colnames(age\_pred) <- c("year", "age")

age\_pred$year <- age\_pred$year + 6

age\_pred$complaints <- predict(age\_model, age\_pred)

combined\_pred\_age <- rbind(age, age\_pred)

ggplot(combined\_pred\_age, aes(x = year, y = complaints, color = age))+

geom\_point()+geom\_smooth(method = "lm", se = F)+

ggtitle("Projected Trends in Targetted Age Groups")

```

# References

[1] 2016 ACFE Report to the Nations. (n.d.). Retrieved October 08, 2016, from [link](http://www.acfe.com/rttn2016/costs.aspx)

[2] EY Identifies Top Fraud and Corruption Trends for 2016. Retrieved October 08, 2016, from [link](http://www.ey.com/us/en/newsroom/news-releases/news-ey-identifies-top-fraud-and-corruption-trends-for-2016)

[3] Consumer Sentinel Network Reports. Retrieved October 08, 2016, from [Link](https://www.ftc.gov/enforcement/consumer-sentinel-network/reports)