EDA

Elinor chu

12/6/2021

```
theft_alldata=read_csv("../data/clean/dataclean.csv")
```

```
## Rows: 2293 Columns: 69
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): county, state
## dbl (67): fips, pertrump, permale, med_age, nevermarried, widowed, fromdifst...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
theft_train=read_csv("../data/clean/theft_train.csv")
```

```
## Rows: 1836 Columns: 69
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): county, state
## dbl (67): fips, pertrump, permale, med_age, nevermarried, widowed, fromdifst...
```

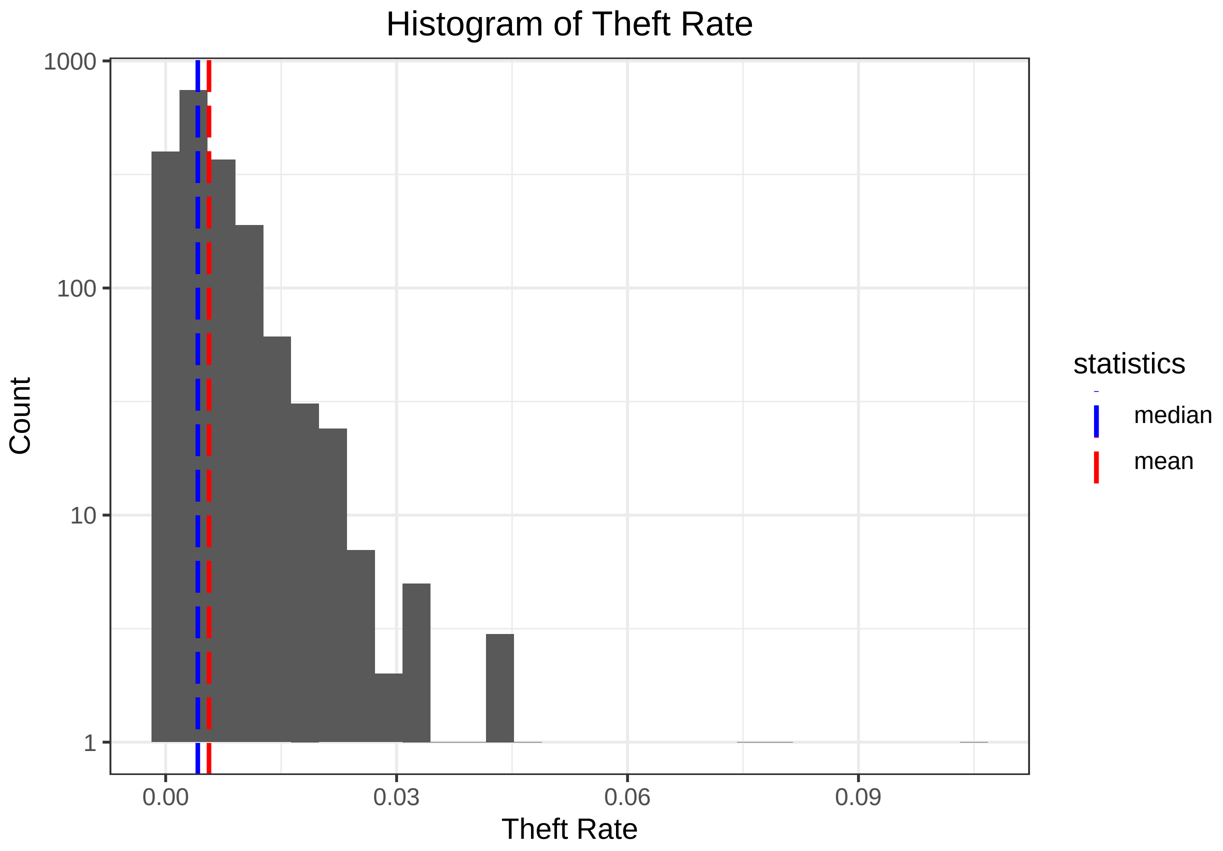
```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

EDA - Response Variable

Histogram of Response

```
theft_train %>% ggplot(aes(x = theftrate)) +
  geom_histogram()+
  labs(y = "Count",
       x = "Theft Rate",
       title = "Histogram of Theft Rate")+
  geom_vline(aes(xintercept = mean(theftrate),colour = "mean"),
             linetype ="longdash", size = .8)+
  geom_vline(aes(xintercept = median(theftrate),colour = "median"),
             linetype ="longdash", size = .8)+
  theme_bw()+
  scale_y_log10()+
  scale_color_manual(name = "statistics", values = c(median = "blue", mean = "red"))+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 13 rows containing missing values (geom_bar).
```



```
mean(theft_train$theftrate)
```

```
## [1] 0.0056293
```

```
median(theft_train$theftrate)
```

```
## [1] 0.004172272
```

To understand the distribution of the response variable, we first ploted a histogram of theft rate. As seen from Figure X, the data appears to be right-skewed, with some counties exceeding a theft rate of 4%. The mean county-level theft rate in 2020 is 0.563%; the median is 0.417%. There are a few outlier counties with very high theft rates.

Highest Theft Rate - Top10 Counties

```
theft_train %>% select(state,county,theftrate) %>% arrange(desc(theftrate)) %>% head(10) %>%
  kable(format = "latex", row.names = NA,
        booktabs = TRUE,
        digits = 5,
        col.names = c("State", "County","Theft Rate"),
        caption = "This a table showing the top 10 counties with the highest theft rate.") %>%
  kable_styling(position = "center") %>%
  kable_styling(latex_options = "HOLD_position")
```

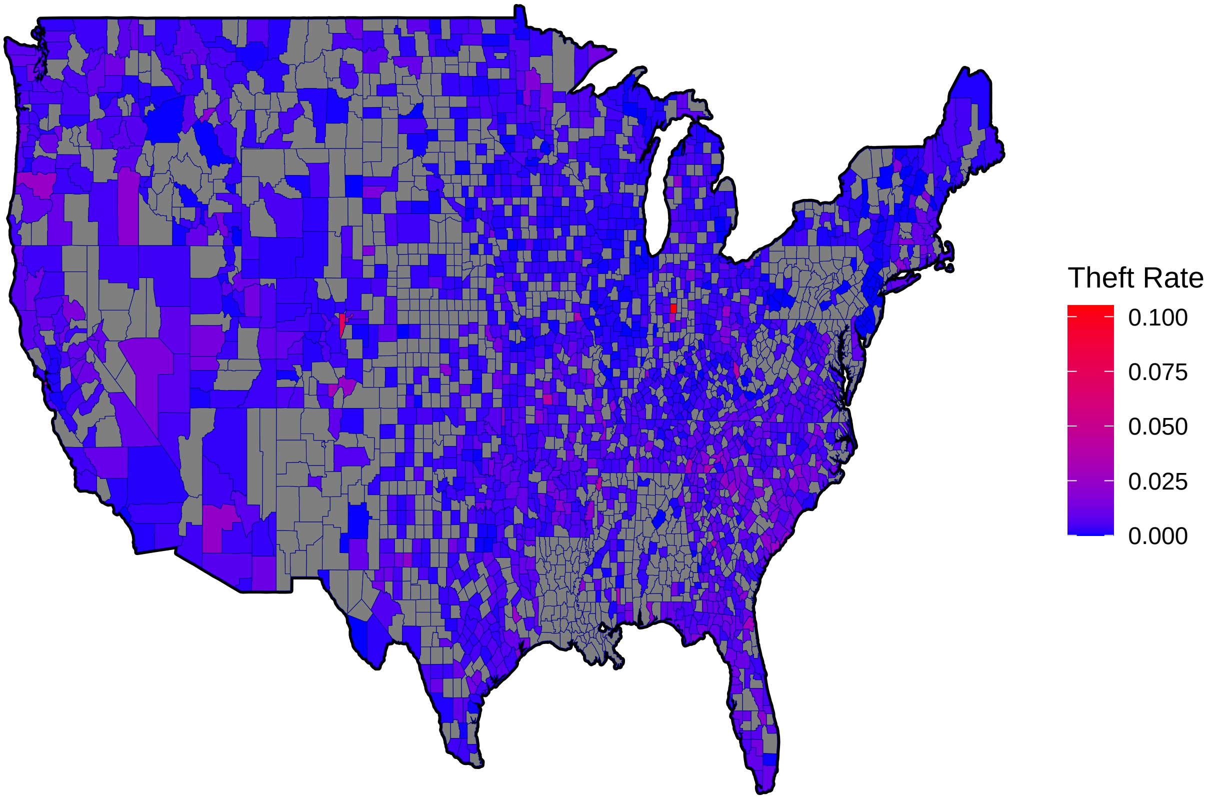
Table 1: This a table showing the top 10 counties with the highest theft rate.

State	County	Theft Rate
Indiana	Hamilton	0.10498
New York	New York	0.07927
Colorado	Jefferson	0.07638
Colorado	Denver	0.04860
Mississippi	Tunica	0.04465
California	San Francisco	0.04427
Missouri	Greene	0.04167
West Virginia	Wayne	0.04136
Missouri	Marion	0.03510
Georgia	Bibb	0.03394

We proceeded to determine which counties had extremely high theft rate in 2020 by looking at the sorted data. The sorted data in Table X shows that the top 6 counties with highest rates of theft incidence are Hamilton county in IN, New York county in NY, Jefferson county in CO, Denver county in CO, Tunica county in MS, and San Francisco county in CA.

Heat map of theft rate

```r
map_data("county") %>%
  as_tibble() %>%
  left_join(theft_train %>%
              rename(region = state,
                     subregion = county,
                     `Theft Rate` = theftrate) %>%
              mutate(region = str_to_lower(region),
                     subregion = str_to_lower(subregion)),
            by = c("region", "subregion")) %>%
  ggplot() +
  geom_polygon(data=map_data("state"),
               aes(x=long, y=lat, group=group),
               color="black", fill=NA,  size = 1, alpha = .3) +
  geom_polygon(aes(x=long, y=lat, group=group, fill = `Theft Rate`),
               color="darkblue", size = .1) +
  scale_fill_gradient(low = "blue", high = "red") +
  theme_void()
```

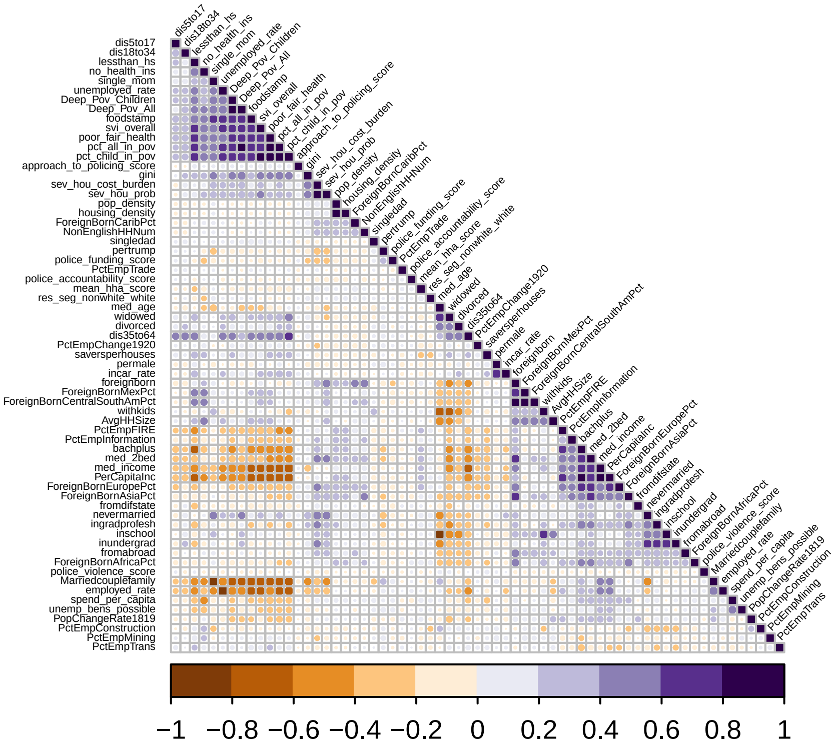


We also created a heat map for our response variable. As shown in Figure X, most of the counties included in this dataset have theft rates well below 2.5%. Very few states have theft rates about 7.5%. The grey areas indicate counties not included in our dataset.

#EDA for important features

corrplots of all features

```
theft_train_corrAll = theft_train%>% select(-fips, -state, -county,-theftrate)
M = cor(theft_train_corrAll)
corrplot(M, type = 'lower', order = 'hclust', tl.col = 'black',
         cl.ratio = 0.2, tl.srt = 45, col = COL2('PuOr', 10), tl.cex = 0.35)
```



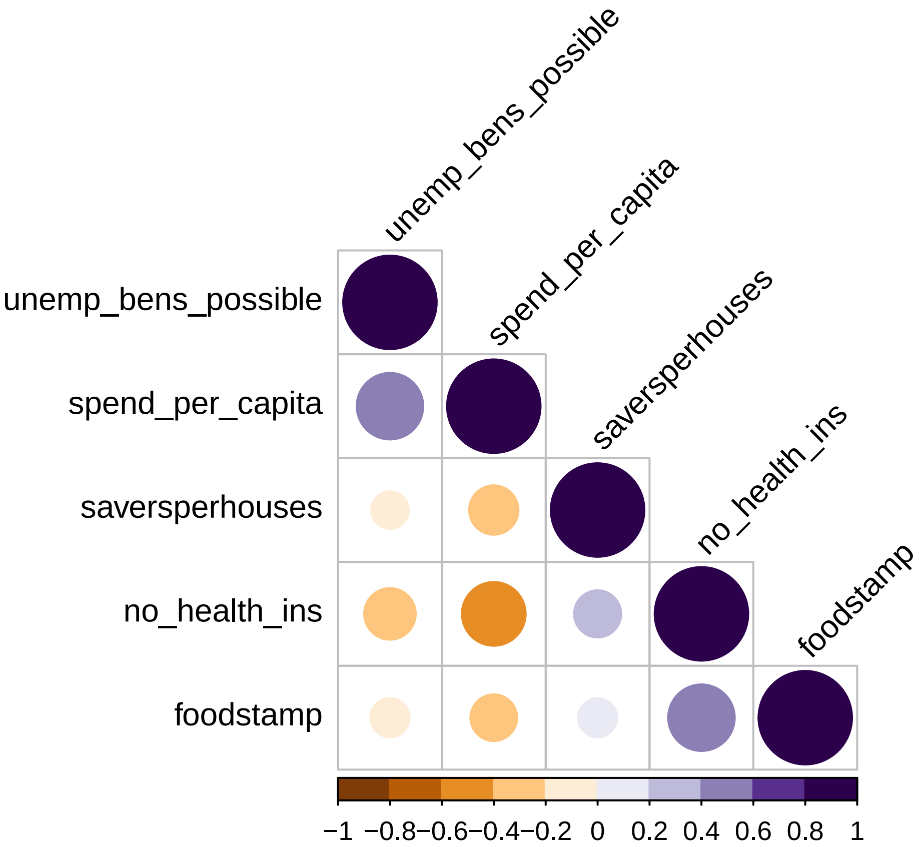
corrplots of 5 clusters of features

```
cluster_safetynet = theft_train%>% select(-fips, -state, -county) %>% select(unemp_bens_possible, spend_

cluster_criminaljustice = theft_train%>% select(-fips, -state, -county) %>% select(incar_rate, police_vi

cluster_health = theft_train%>% select(-fips, -state, -county) %>% select(mean_hha_score, poor_fair_heal

cluster_ses = theft_train%>% select(-fips, -state, -county) %>% select(lessthan_hs, bachplus, unemployed

cluster_demo= theft_train%>% select(-fips, -state, -county) %>% select(med_age,permale,divorced,widowed
```

```
M_safetynet = cor(cluster_safetynet)
M_criminaljustice = cor(cluster_criminaljustice)
M_health = cor(cluster_health)
M_ses = cor(cluster_ses)
M_demo = cor(cluster_demo)
```

social safety net

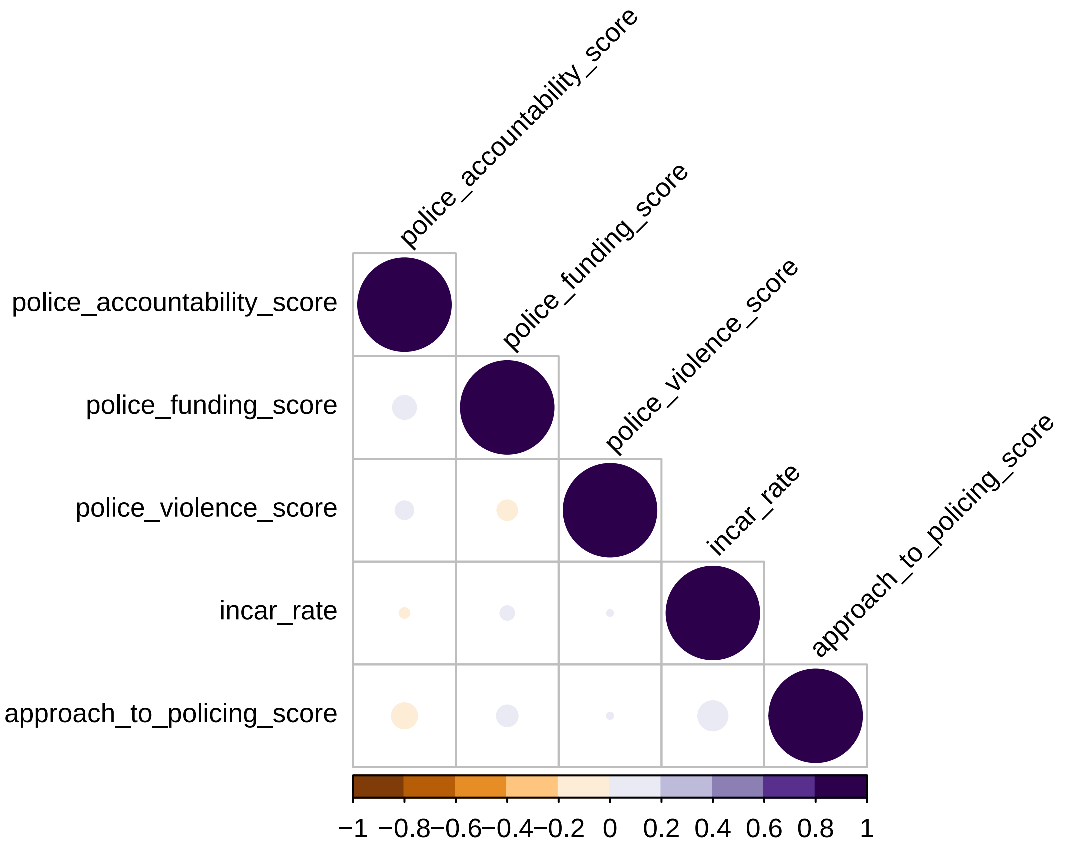
```
corrplot(M_safetynet, type = 'lower', order = 'hclust', tl.col = 'black',
         cl.ratio = 0.2, tl.srt = 45, col = COL2('PuOr', 10), tl.cex = 1)
```



We observed a positively correlation between State and local government spending on people and State unemployment insurance. Also, Percent of households qualifying for food stamps is positively correlated with No Health Insurance. Not surprisingly, No Health Insurance is negatively correlated with State and local government spending on people.

criminal justice response

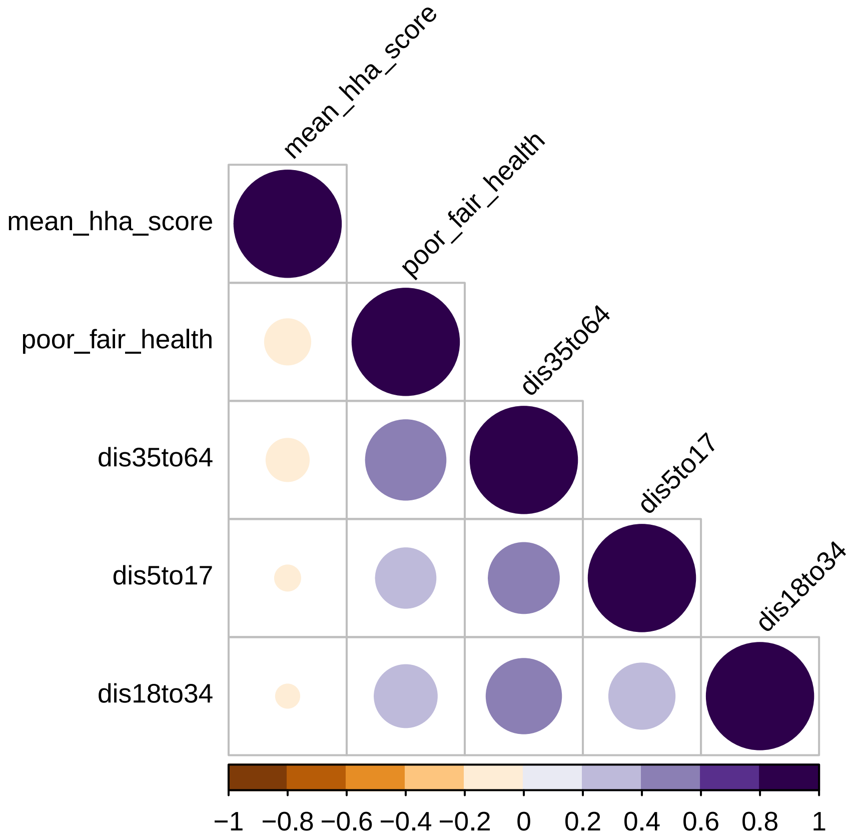
```
corrplot(M_criminaljustice, type = 'lower', order = 'hclust', tl.col = 'black',
         cl.ratio = 0.2, tl.srt = 45, col = COL2('PuOr', 10), tl.cex = 0.8)
```



We found no significant correlations among features belonging to the category of Criminal justice response.

##health-related factors

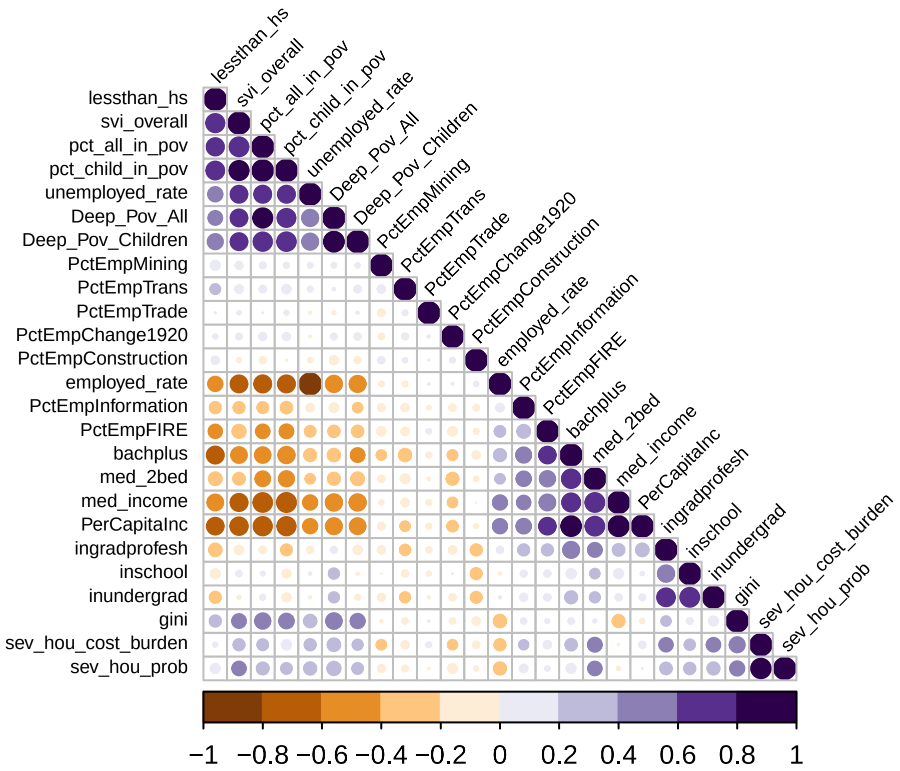
```
corrplot(M_health, type = 'lower', order = 'hclust', tl.col = 'black',
         cl.ratio = 0.2, tl.srt = 45, col = COL2('PuOr', 10), tl.cex = 0.8)
```



We observed positive correlations between Percent adults reporting poor or fair health and Percent of people with disability in each of the three age groups. Moreover, the percentages of people with disability for the three age groups are also positively correlated with each other.

SES

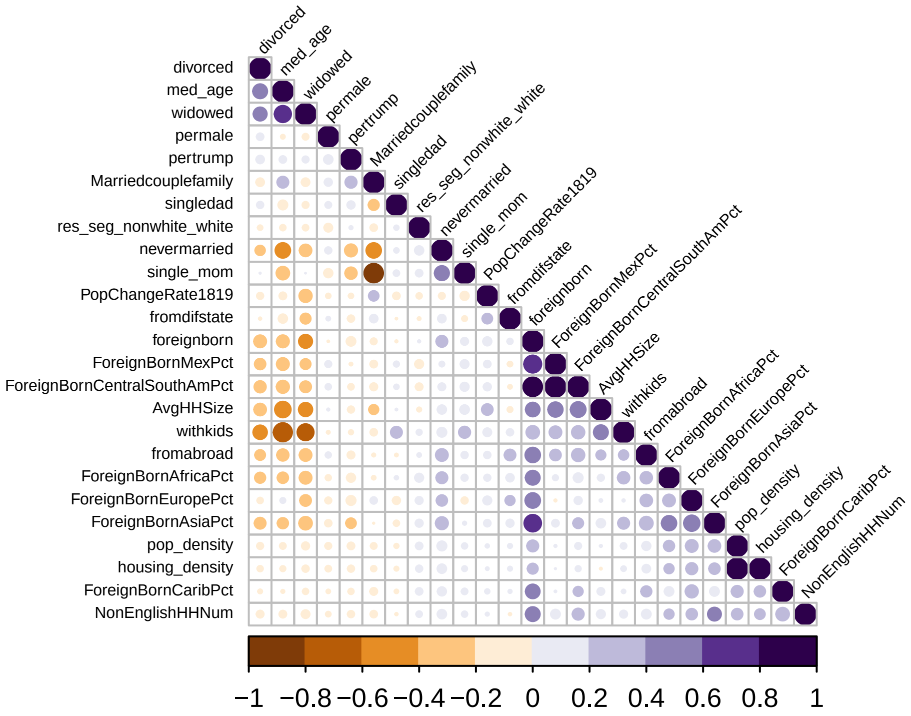
```
corrplot(M_ses, type = 'lower', order = 'hclust', tl.col = 'black',
         cl.ratio = 0.2, tl.srt = 45, col = COL2('PuOr', 10), tl.cex = 0.6)
```



We observed that there are positive correlations between the Social Vulnerability Index (SVI) and all poverty-related features. We also found that a group of features, including Employment rate, Percent with college or higher education, Median household income, and Per capita income in the past 12 months, that are negatively correlated with SVI and poverty-related features.

basic demographics

```
corrplot(M_demo, type = 'lower', order = 'hclust', tl.col = 'black',
         cl.ratio = 0.2, tl.srt = 45, col = COL2('PuOr', 10), tl.cex = 0.5)
```

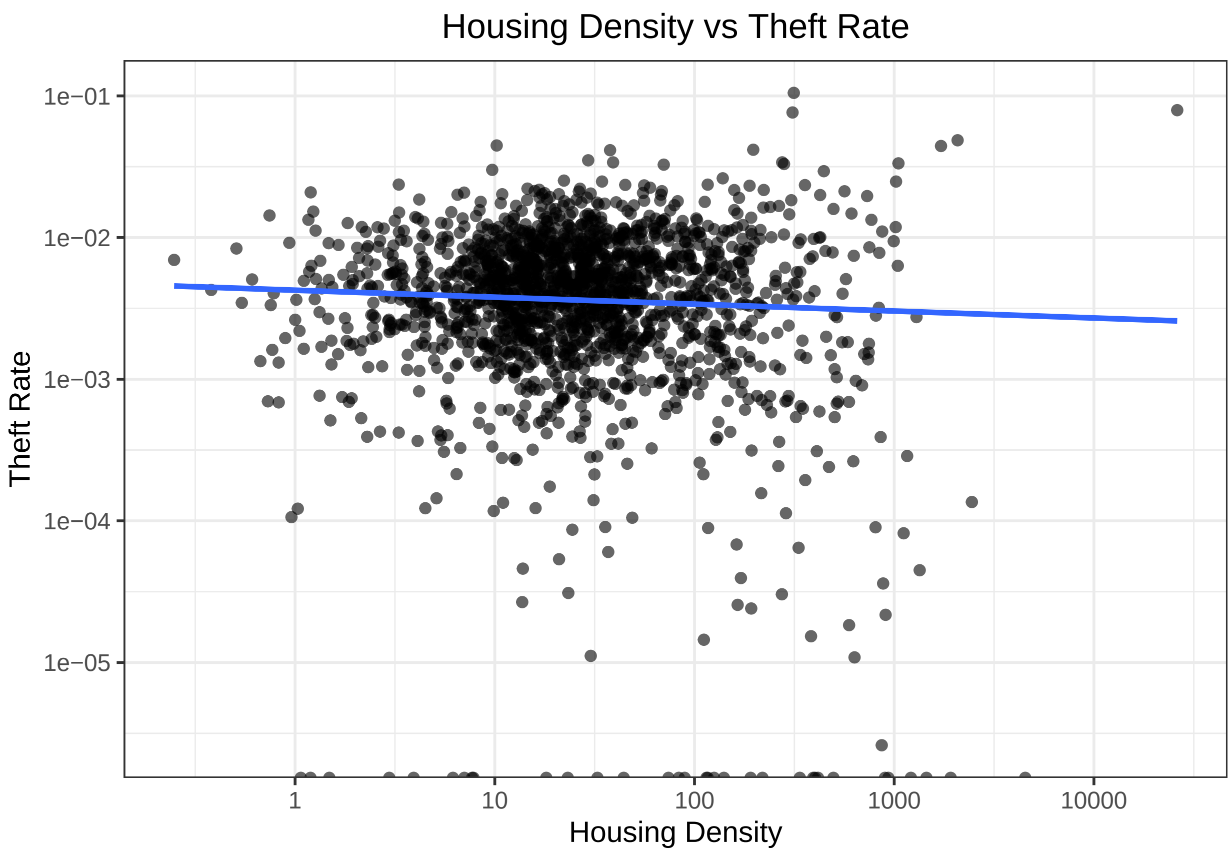


From Figure X, we observed that Percent of household with own children is negatively correlated with both Age (median) and Percent Divorced. Housing density has a significant positive correlation with the Population density.

Histogram for the Top7 important features (overlaps of Ridge & Lasso)

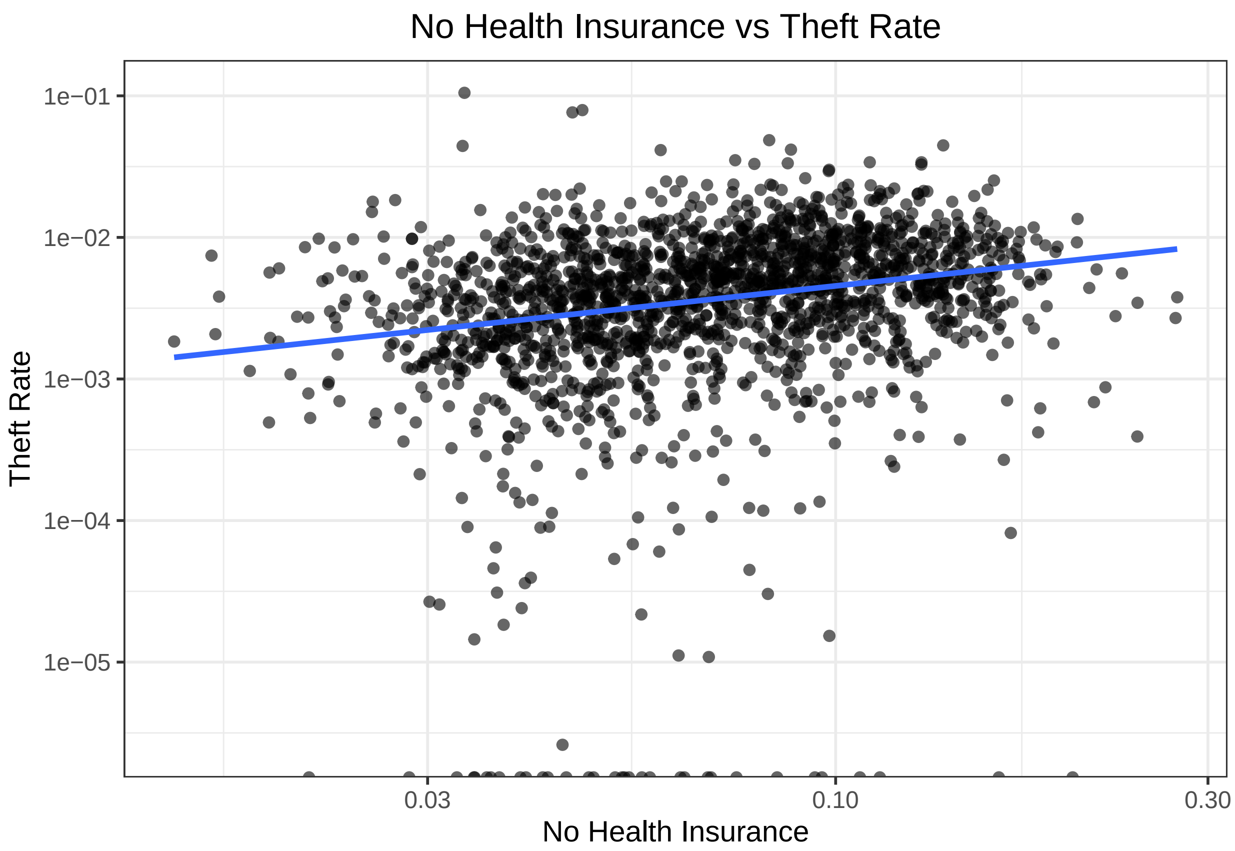
```r
# plot theftrate against housing_density
p1 = theft_train %>% select(-fips, -state, -county)%>%
  ggplot(aes(x = housing_density, y = theftrate)) +
  geom_point(alpha = 0.6) +
  scale_x_log10() +
  scale_y_log10() +
  geom_smooth(method = "lm", formula = "y~x", se = FALSE) +
  labs(x = "Housing Density",
       y = "Theft Rate",
       title = "Housing Density vs Theft Rate") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

p1
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 33 rows containing non-finite values (stat_smooth).
```



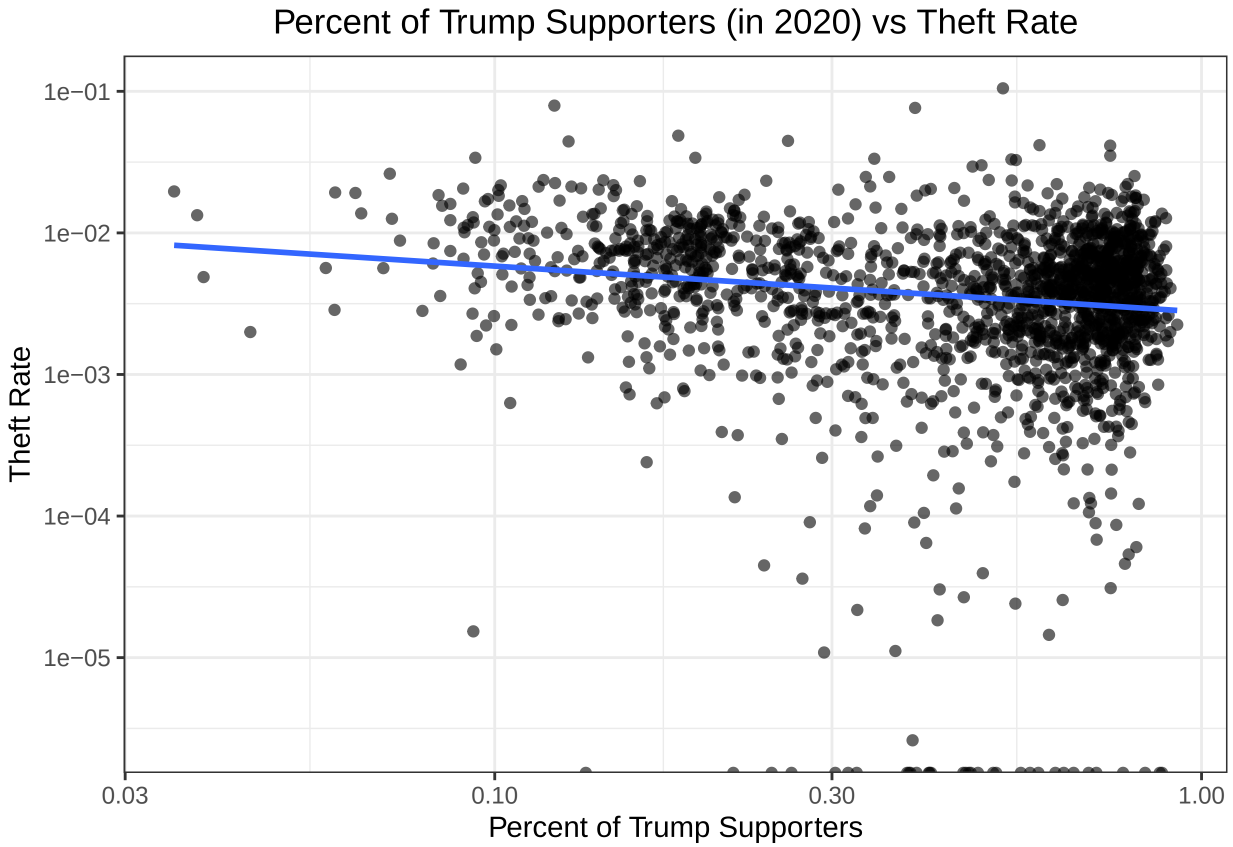
```r
# plot theftrate against no_health_ins
p2 = theft_train %>% select(-fips, -state, -county) %>%
  ggplot(aes(x = no_health_ins, y = theftrate)) +
  geom_point(alpha = 0.6) +
  scale_x_log10() +
  scale_y_log10() +
  geom_smooth(method = "lm", formula = "y~x", se = FALSE) +
  labs(x = "No Health Insurance",
       y = "Theft Rate",
       title = "No Health Insurance vs Theft Rate") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

p2
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 33 rows containing non-finite values (stat_smooth).
```



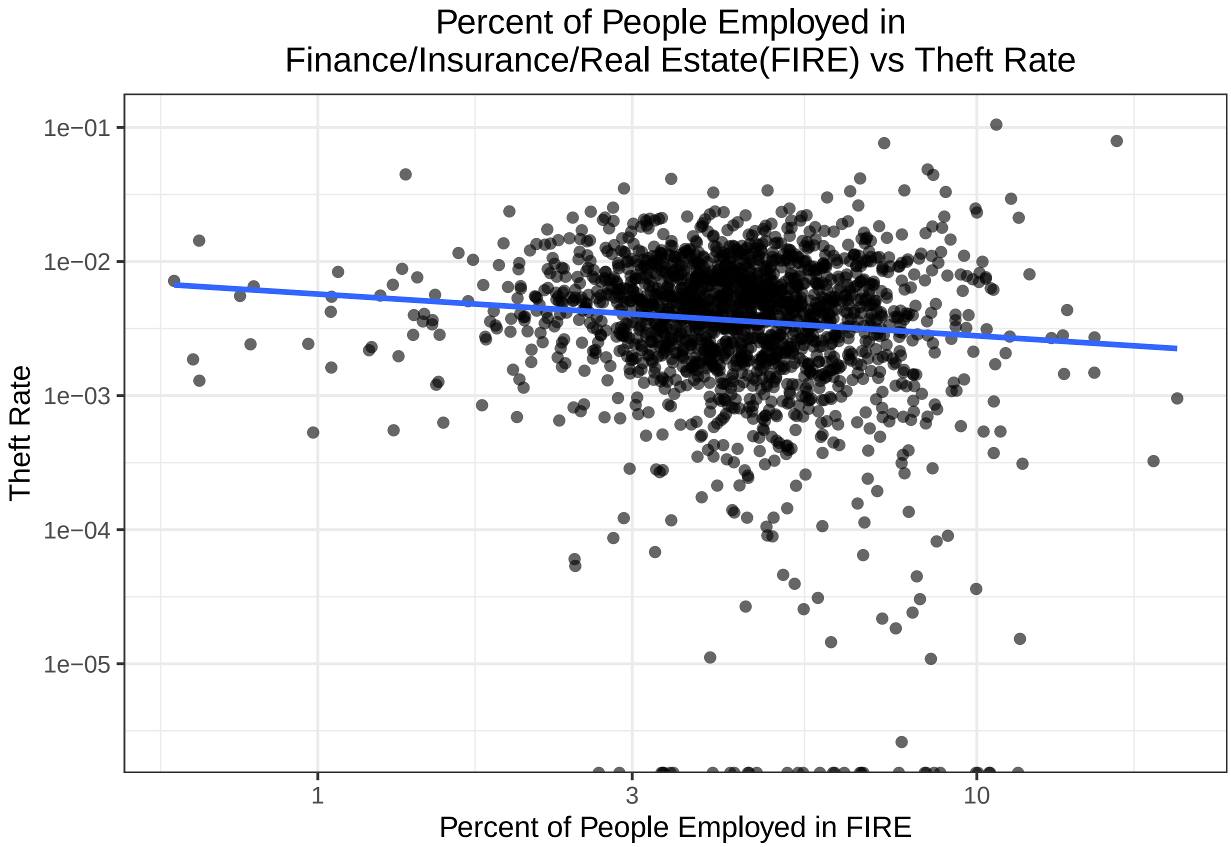
```
# plot theftrate against pertrump
p3 = theft_train %>% select(-fips, -state, -county)%>%
  ggplot(aes(x = pertrump, y = theftrate)) +
  geom_point(alpha = 0.6) +
  scale_x_log10() +
  scale_y_log10() +
  geom_smooth(method = "lm", formula = "y~x", se = FALSE) +
  labs(x = "Percent of Trump Supporters",
       y = "Theft Rate",
       title = "Percent of Trump Supporters (in 2020) vs Theft Rate") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

p3
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 33 rows containing non-finite values (stat_smooth).
```



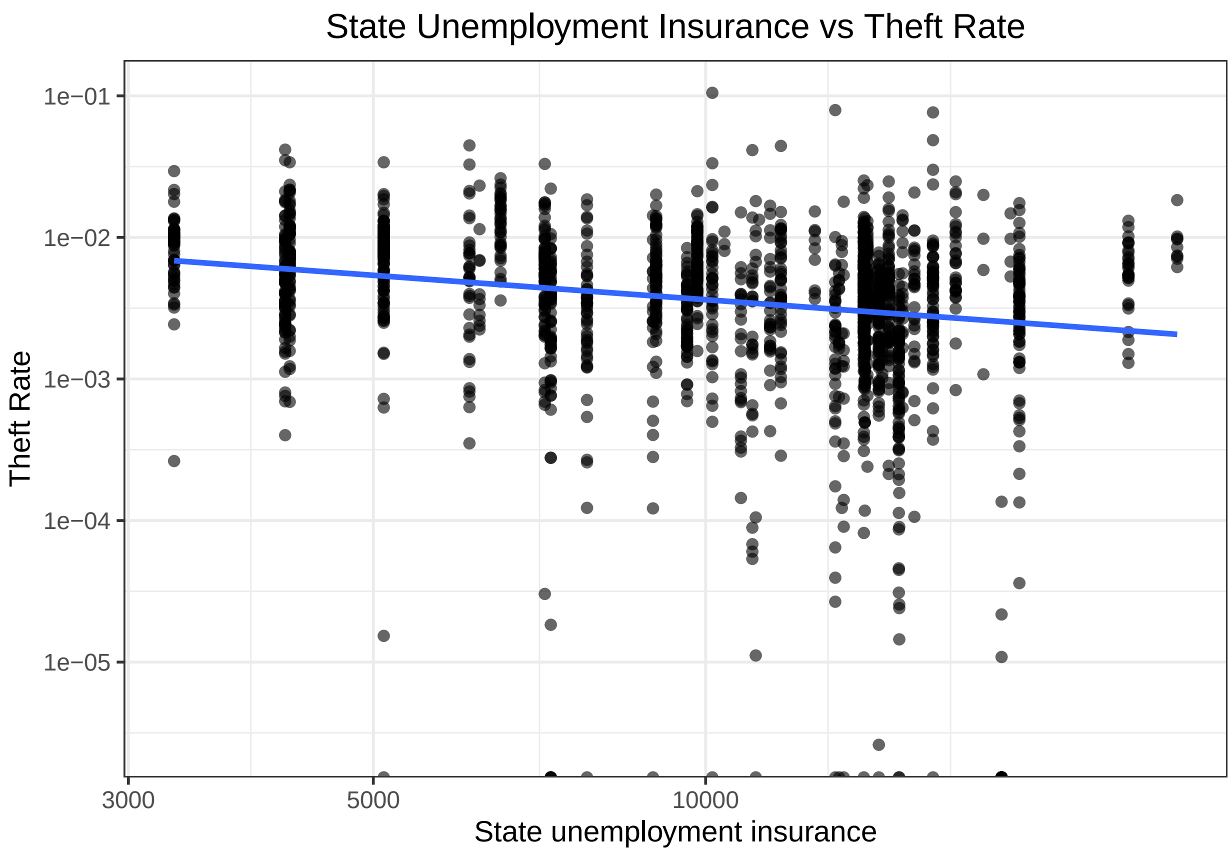
```r
# plot theftrate against PctEmpFIRE
p4 = theft_train %>% select(-fips, -state, -county)%>%
  ggplot(aes(x = PctEmpFIRE, y = theftrate)) +
  geom_point(alpha = 0.6) +
  scale_x_log10() +
  scale_y_log10() +
  geom_smooth(method = "lm", formula = "y~x", se = FALSE) +
  labs(x = "Percent of People Employed in FIRE",
       y = "Theft Rate",
       title = "Percent of People Employed in \n Finance/Insurance/Real Estate(FIRE) vs Theft Rate") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

p4
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 33 rows containing non-finite values (stat_smooth).
```



```
# plot theftrate against unemp_bens_possible
p5 = theft_train %>% select(-fips, -state, -county)%>%
  ggplot(aes(x = unemp_bens_possible, y = theftrate)) +
  geom_point(alpha = 0.6) +
  scale_x_log10() +
  scale_y_log10() +
  geom_smooth(method = "lm", formula = "y~x", se = FALSE) +
  labs(x = "State unemployment insurance",
       y = "Theft Rate",
       title = "State Unemployment Insurance vs Theft Rate") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

p5
```

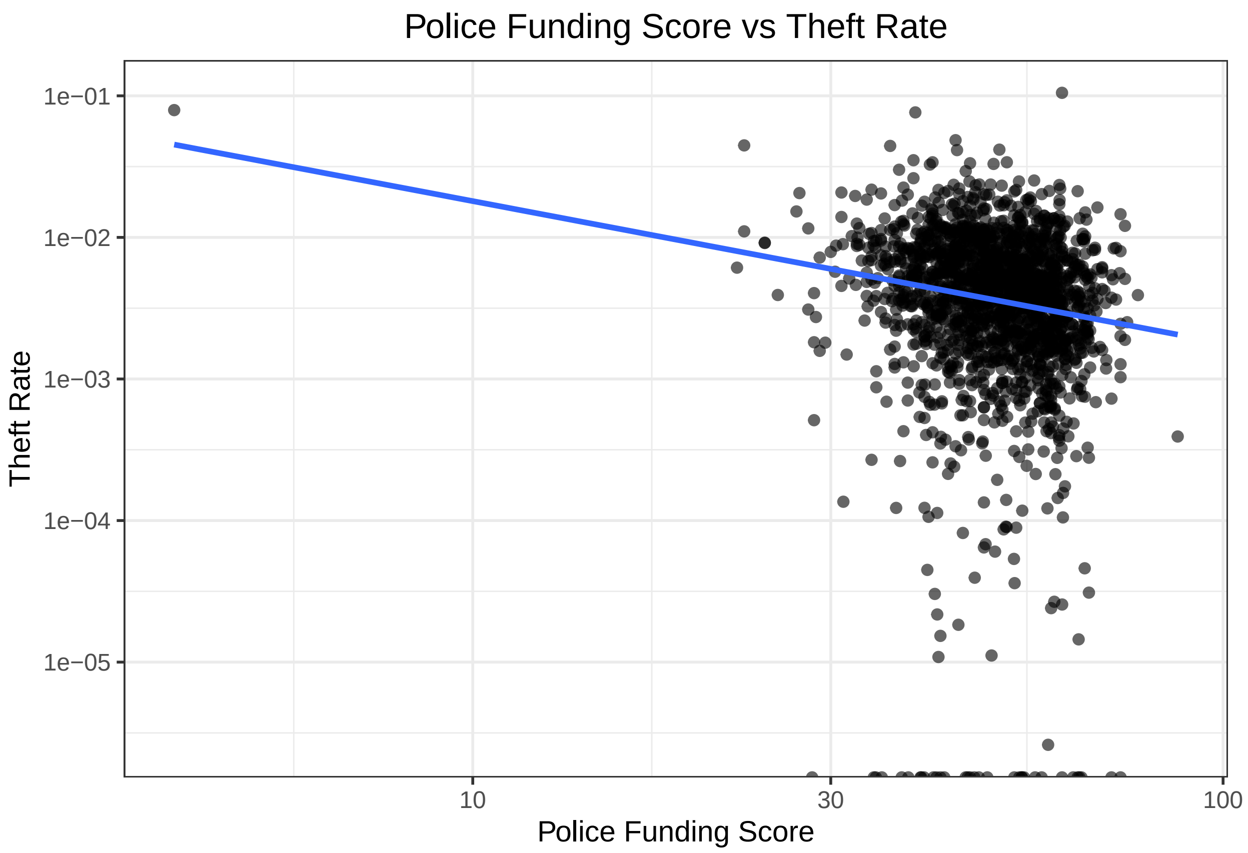
```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 33 rows containing non-finite values (stat_smooth).
```



```r
# plot theftrate against police_funding_score
p6= theft_train %>% select(-fips, -state, -county)%>%
  ggplot(aes(x = police_funding_score, y = theftrate)) +
  geom_point(alpha = 0.6) +
  scale_x_log10() +
  scale_y_log10() +
  geom_smooth(method = "lm", formula = "y~x", se = FALSE) +
  labs(x = "Police Funding Score",
       y = "Theft Rate",
       title = "Police Funding Score vs Theft Rate") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

p6
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 33 rows containing non-finite values (stat_smooth).
```


Police Funding Score vs Theft Rate
Theft Rate
1e−01
1e−02
1e−03
1e−04
1e−05
10
30
100
Police Funding Score