EDA

Elinor chu

12/6/2021

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

```
## Rows: 1769 Columns: 70
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): state, county
## dbl (68): pertrump, permale, med_age, nevermarried, widowed, fromdifstate, f...
```

```
##
## 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

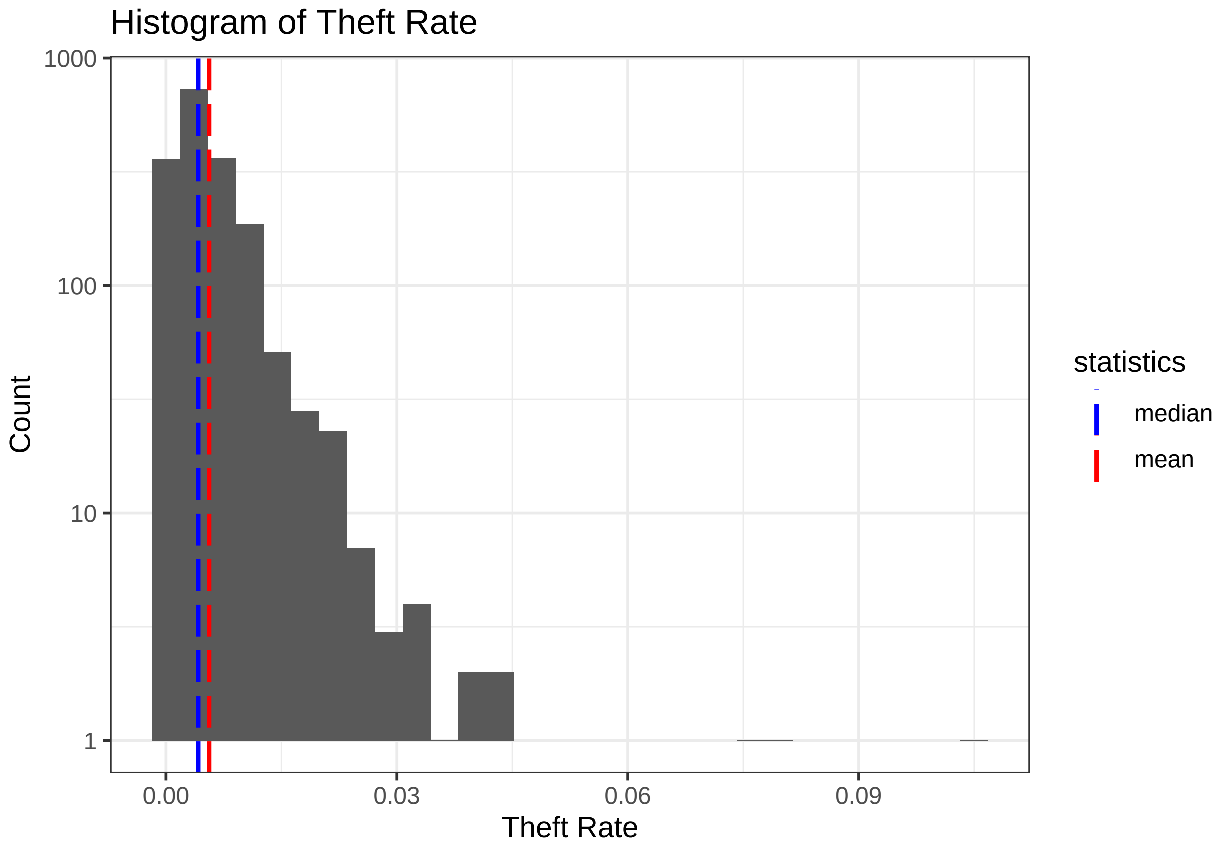
Histogram of Response

```
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"))
```

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

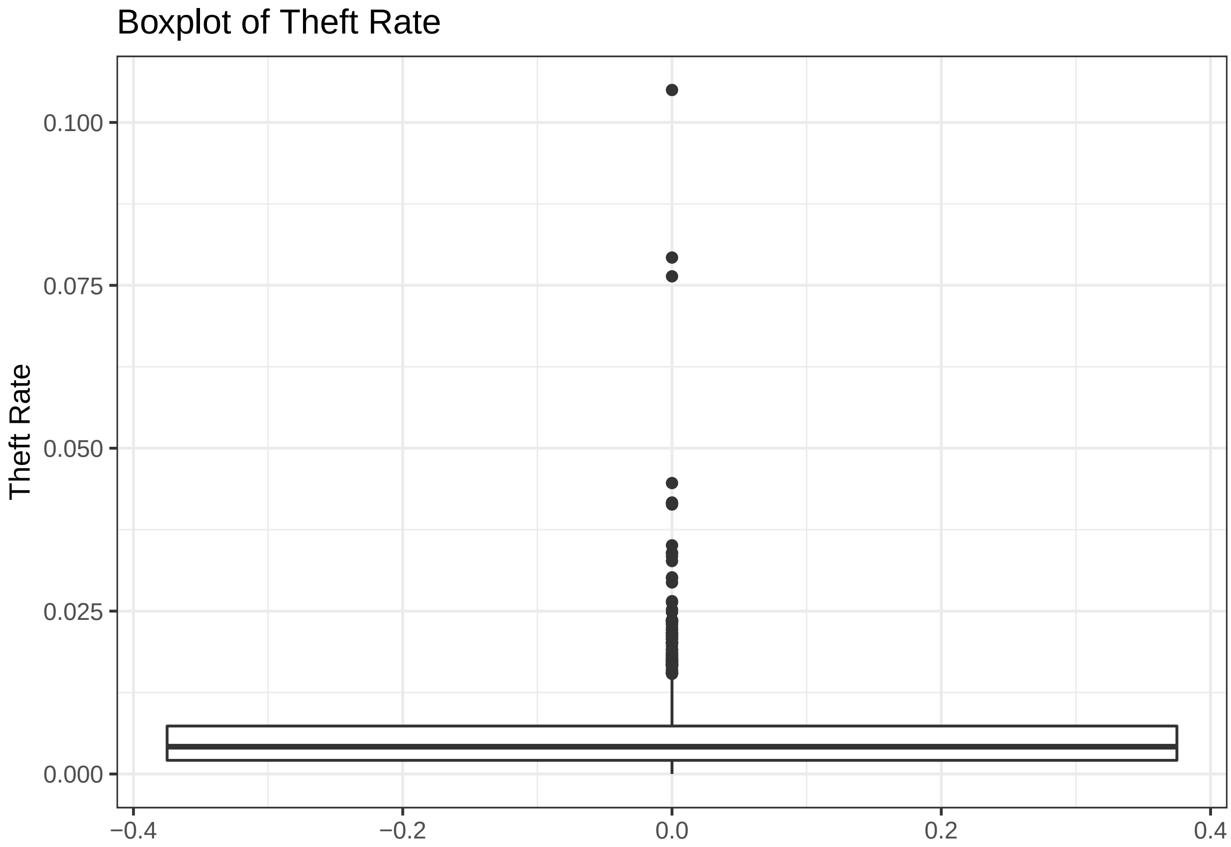
```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 14 rows containing missing values (geom_bar).
```



## Boxplot of Response

```
theft_train %>% select(-fips, -state, -county) %>%
ggplot(aes(y=theftrate)) +
  geom_boxplot()+
  labs(y = "Theft Rate",
       title = "Boxplot of Theft Rate")+
  theme_bw()
```



## Highest Theft Rate - Top10 Counties

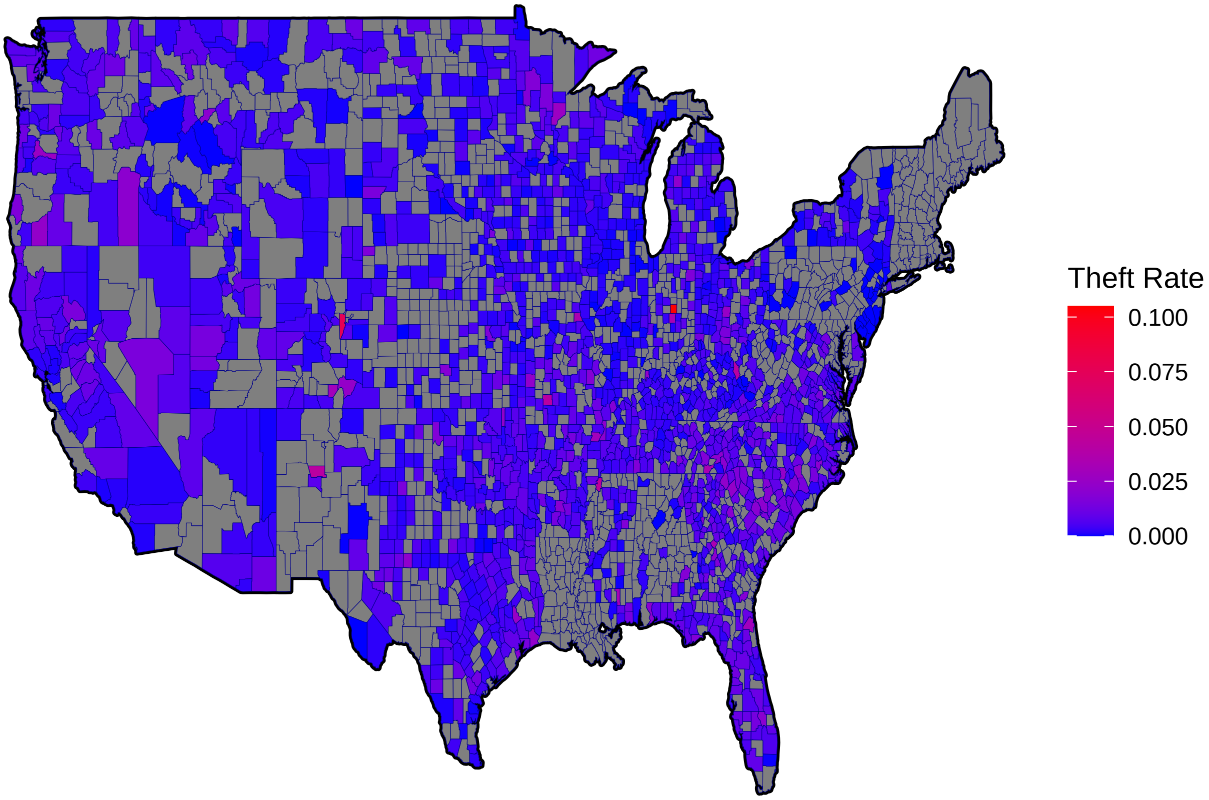
```
theft_train %>% select(state,county,theftrate) %>% arrange(desc(theftrate)) %>% head(10) %>%
  kable(format = "latex", row.names = NA,
        booktabs = TRUE,
        digits = 3,
        col.names = c("State", "County","Theft Rate"),
        caption = "This a table showing the top 10 counties with 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	0.105
New York	New York	0.079
Colorado	Jefferson	0.076
Mississippi	Tunica	0.045
Missouri	Greene	0.042
New Mexico	Bernalillo	0.042
West Virginia	Wayne	0.041
Missouri	Marion	0.035
Georgia	Bibb	0.034
North Carolina	Cherokee	0.034

Heat map of theft rate

```
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()
```

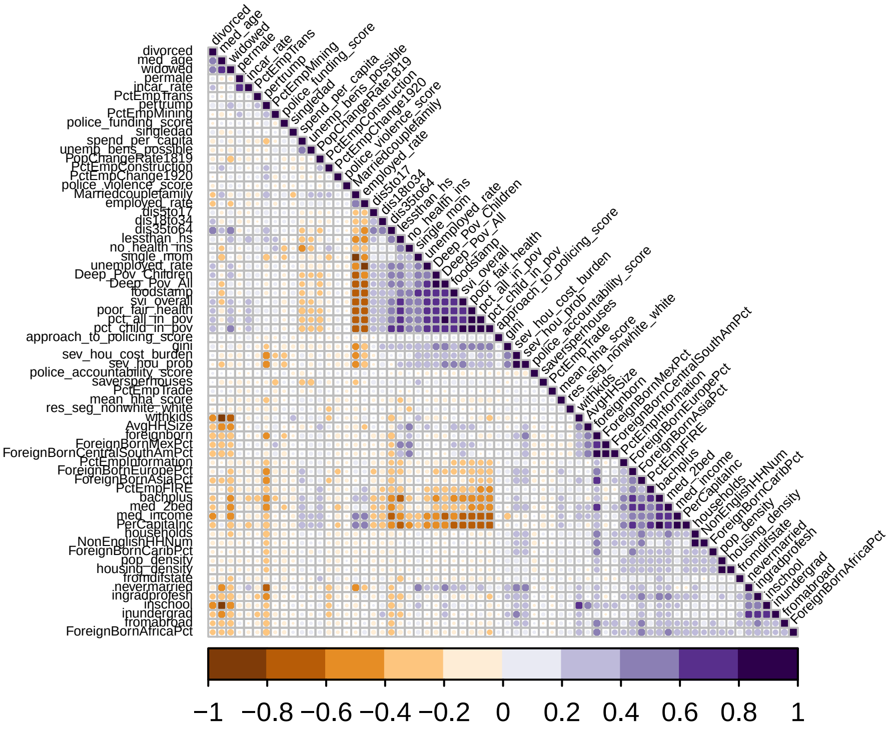


#EDA for important features

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 = 0.4)
```

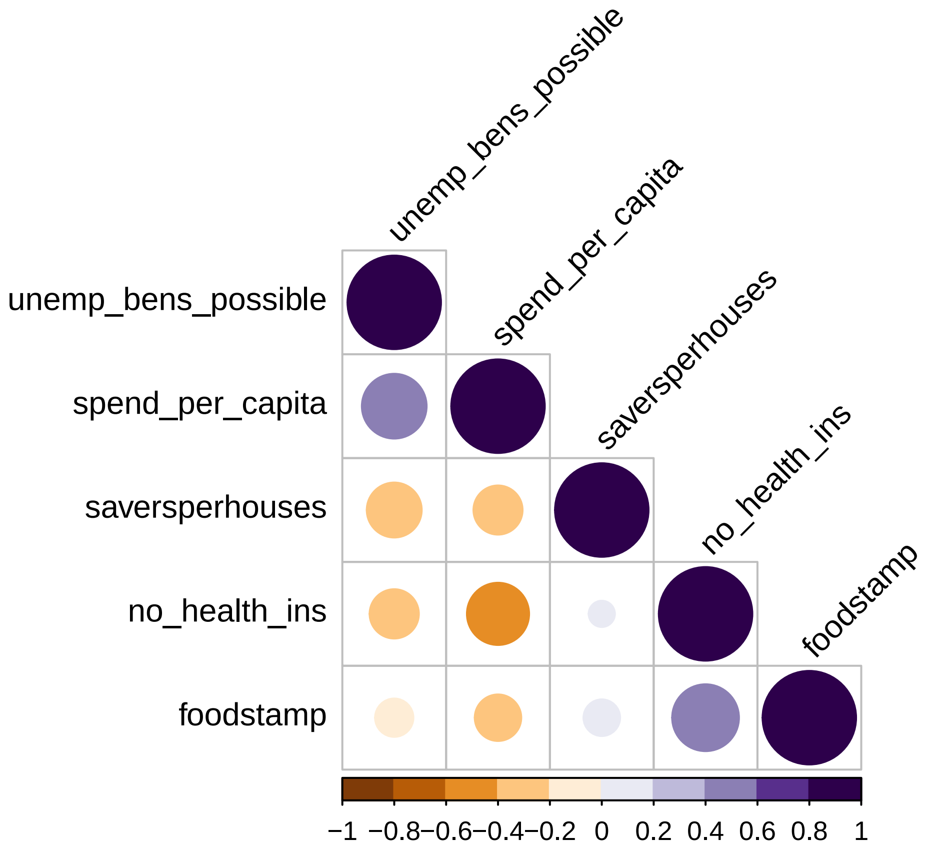


corrplots of 5 clusters of features

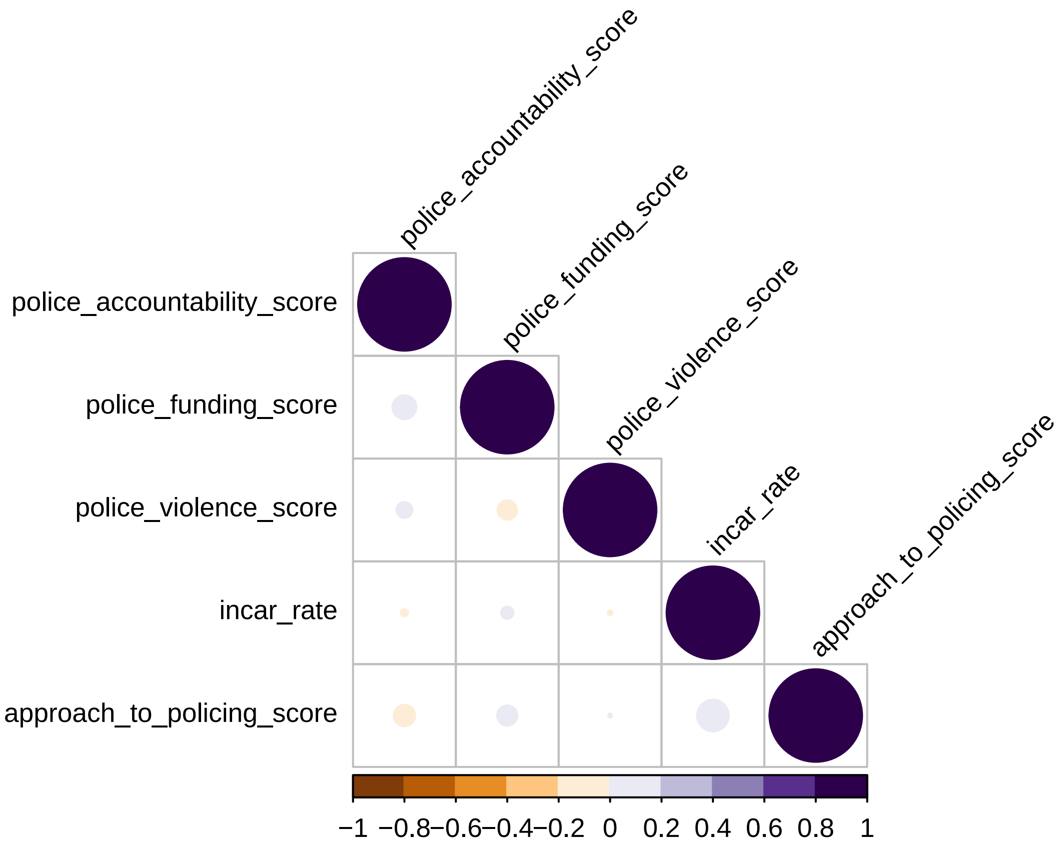
```
cluster_safetynet = theft_train%>% select(-fips, -state, -county) %>% select(unemp_bens_possible, spend_

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)
```

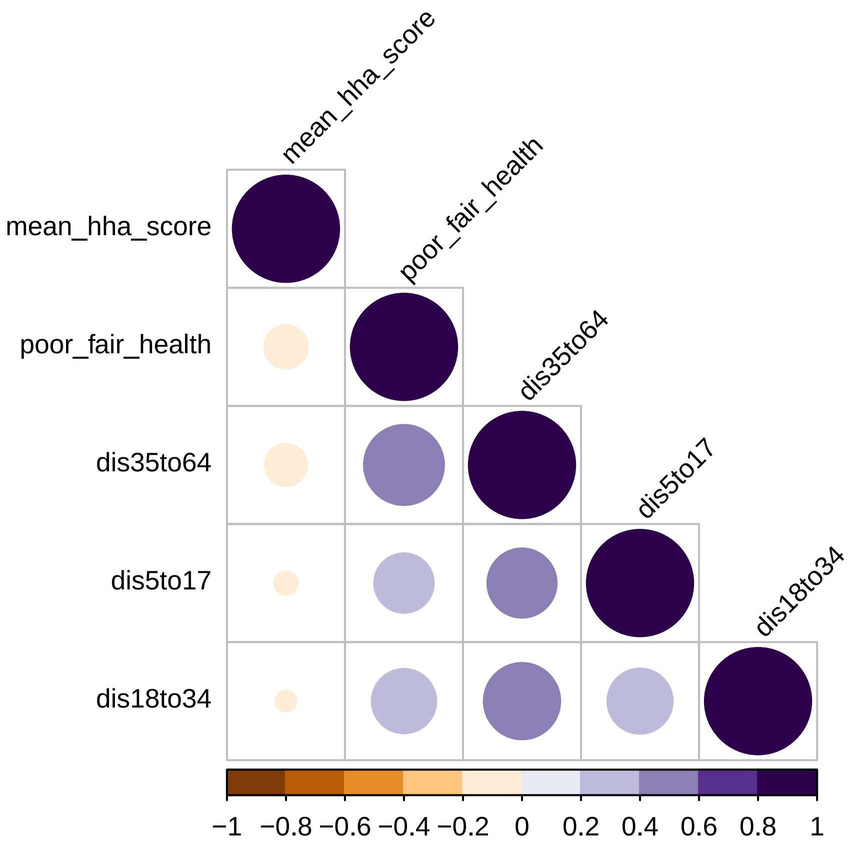
```
corrplot(M_safetynet, type = 'lower', order = 'hclust', tl.col = 'black',
         cl.ratio = 0.2, tl.srt = 45, col = COL2('PuOr', 10), tl.cex = 1)
```



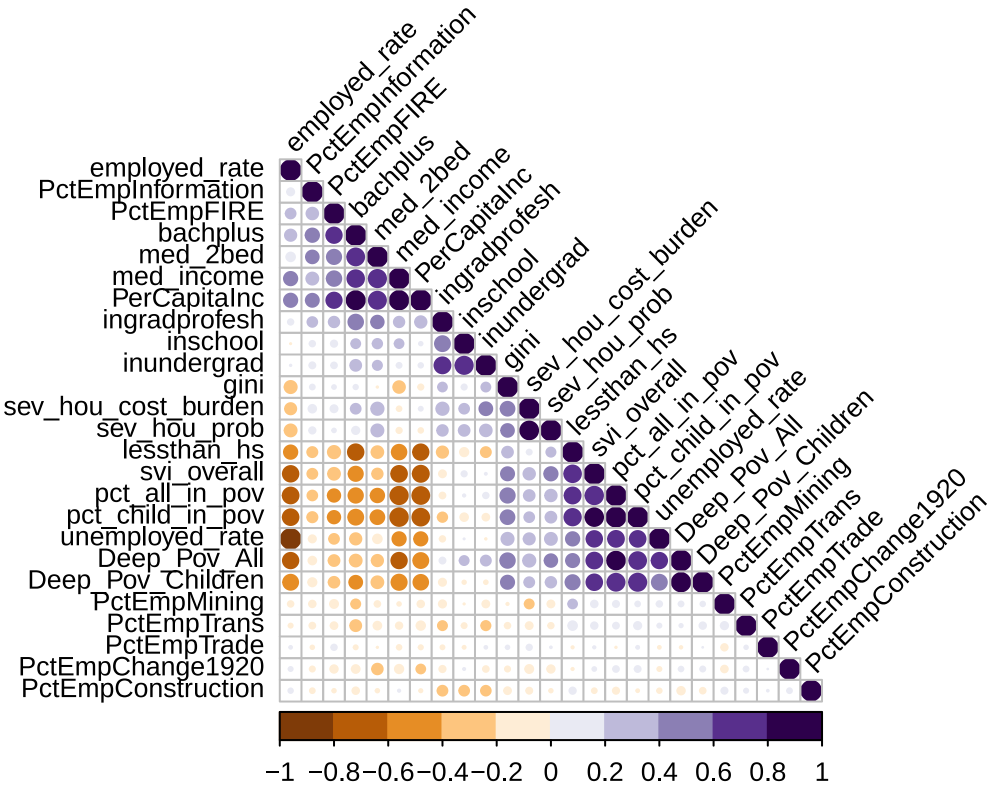
```
corrplot(M_criminaljustice, type = 'lower', order = 'hclust', tl.col = 'black',
         cl.ratio = 0.2, tl.srt = 45, col = COL2('PuOr', 10), tl.cex = 0.8)
```



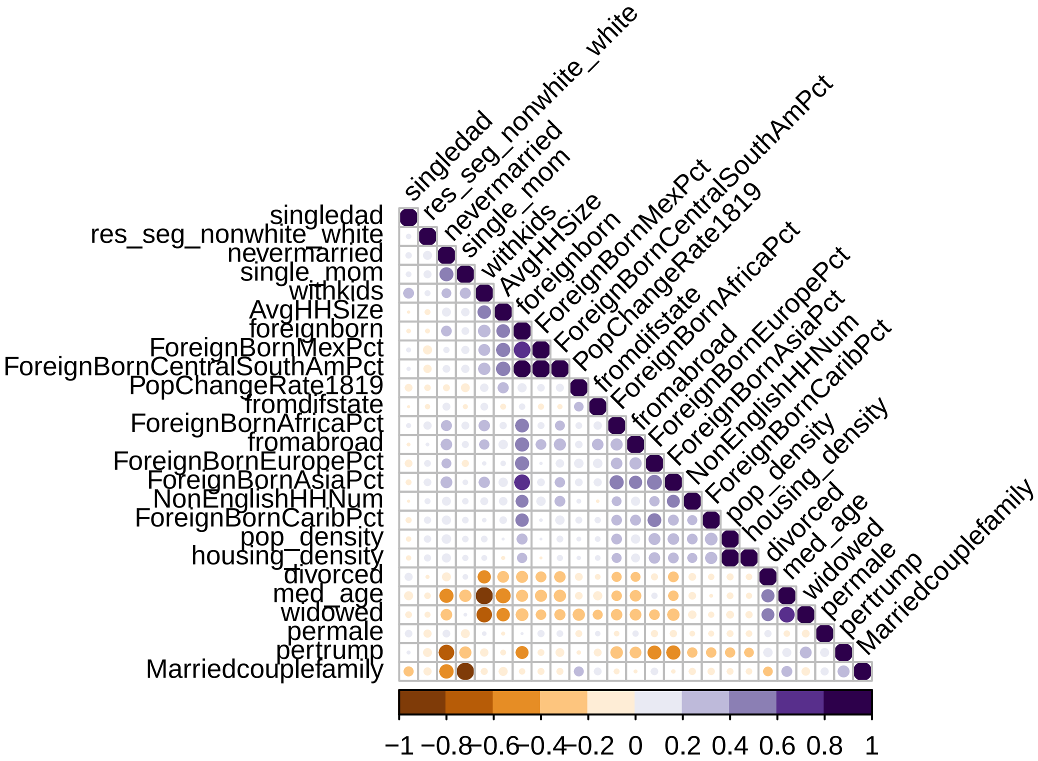
```
corrplot(M_health, type = 'lower', order = 'hclust', tl.col = 'black',
         cl.ratio = 0.2, tl.srt = 45, col = COL2('PuOr', 10), tl.cex = 0.8)
```



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



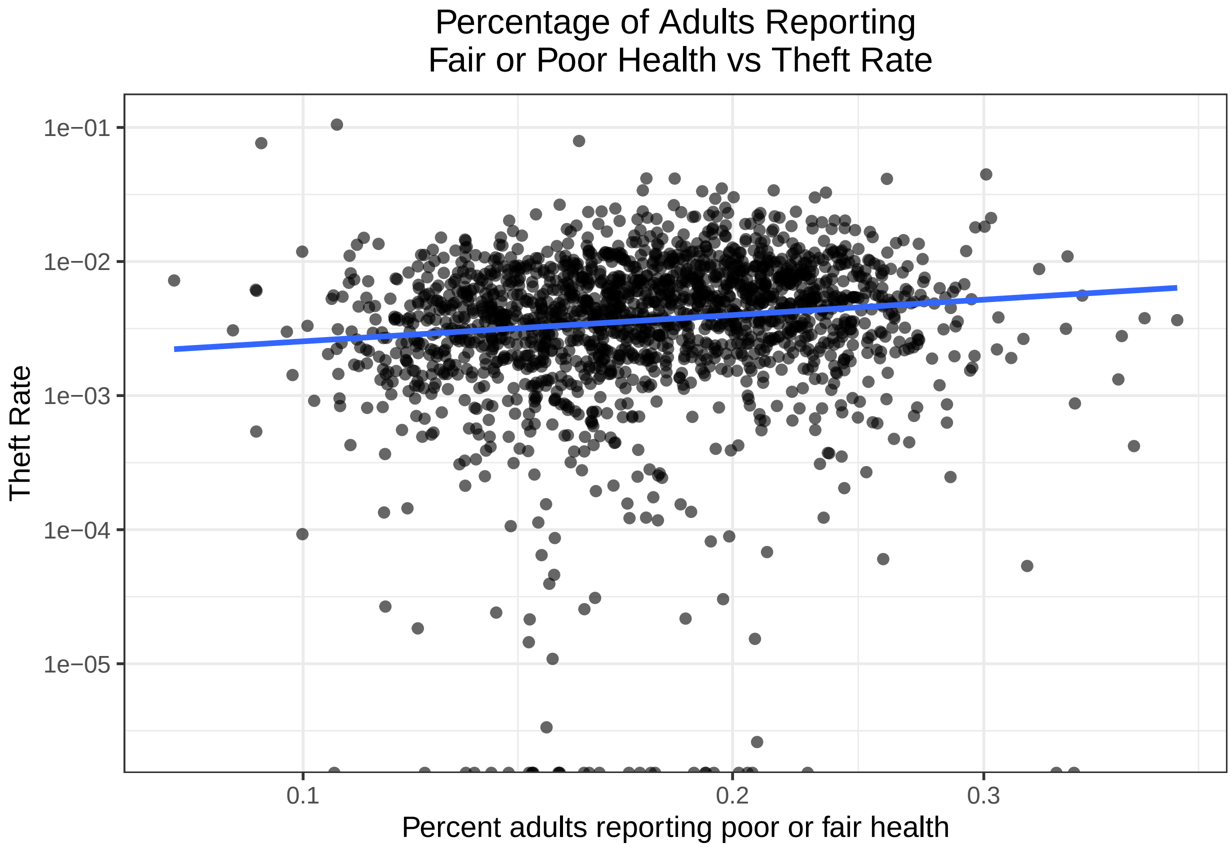
```
corrplot(M_demo, type = 'lower', order = 'hclust', tl.col = 'black',
         cl.ratio = 0.2, tl.srt = 45, col = COL2('PuOr', 10), tl.cex = 0.8)
```



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

```
# plot theftrate against poor_fair_health
p1 = theft_train %>% select(-fips, -state, -county) %>%
  ggplot(aes(x = poor_fair_health, 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 adults reporting poor or fair health",
       y = "Theft Rate",
       title = "Percentage of Adults Reporting\n Fair or Poor Health 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 30 rows containing non-finite values (stat_smooth).
```

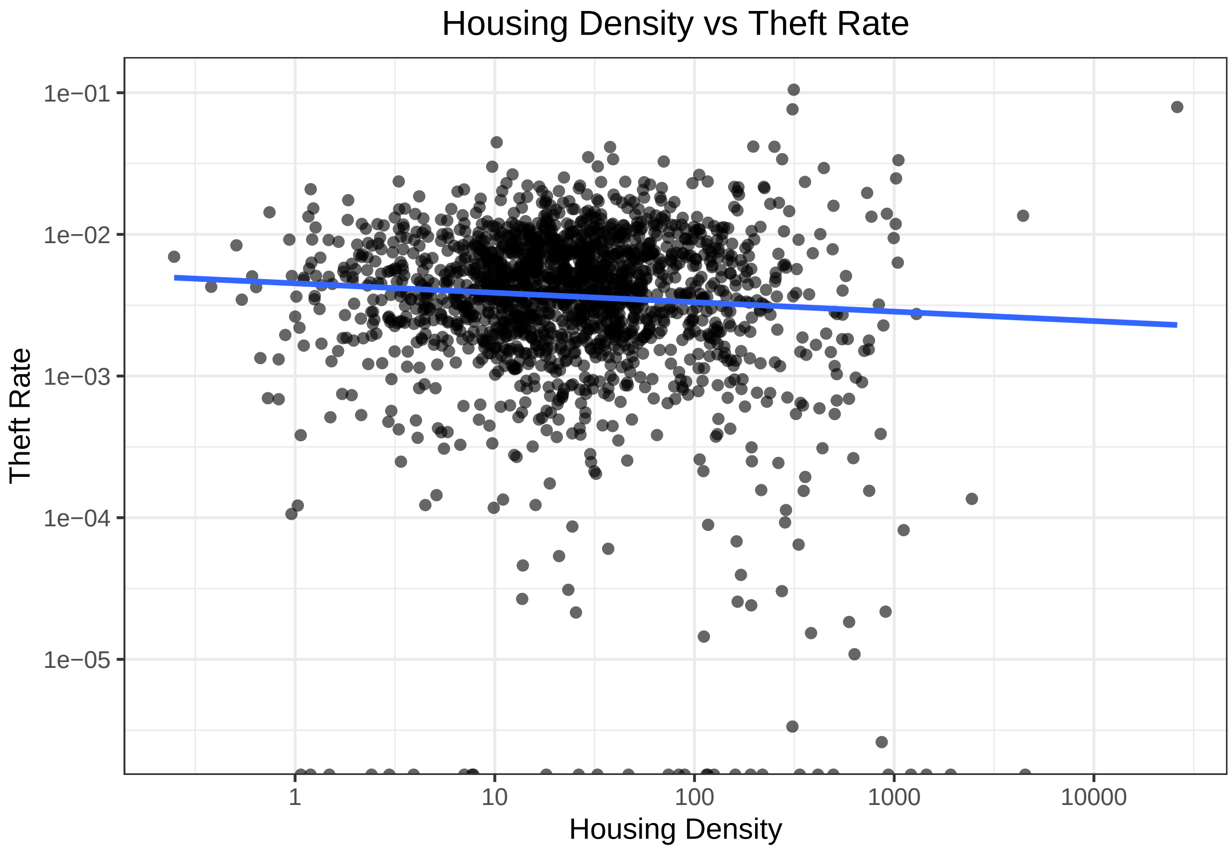


```
theft_train %>% arrange(desc(poor_fair_health))%>%head(5)
```

```
## # A tibble: 5 x 70
##   state       county  pertrump permale med_age nevermarried widowed fromdifstate
##   <chr>       <chr>      <dbl>   <dbl>   <dbl>        <dbl>   <dbl>        <dbl>
## 1 Texas       Zavala     0.340   0.517    32.9        0.297  0.0550      0.0247
## 2 Texas       Starr      0.471   0.487    28.8        0.270  0.0486      0.00257
## 3 Texas       Brooks     0.402   0.533    29.7        0.351  0.0462      0.0381
## 4 Texas       Willacy    0.440   0.540    33          0.310  0.0400      0.00371
## 5 Mississippi Claibo~    0.135   0.469    33.9        0.449  0.0494      0.0136
## # ... with 62 more variables: fromabroad <dbl>, divorced <dbl>,
## #   foodstamp <dbl>, households <dbl>, Marriedcouplefamily <dbl>,
## #   single_mom <dbl>, inschool <dbl>, inundergrad <dbl>, ingradprofesh <dbl>,
## #   lessthan_hs <dbl>, bachplus <dbl>, med_income <dbl>, gini <dbl>,
## #   singledad <dbl>, withkids <dbl>, med_2bed <dbl>, foreignborn <dbl>,
## #   unemployed_rate <dbl>, employed_rate <dbl>, no_health_ins <dbl>,
## #   dis5to17 <dbl>, dis18to34 <dbl>, dis35to64 <dbl>, fips <dbl>, ...
```

```
# plot theftrate against housing_density
p2 = 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))

p2
```

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

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

## Warning: Removed 30 rows containing non-finite values (stat_smooth).
```

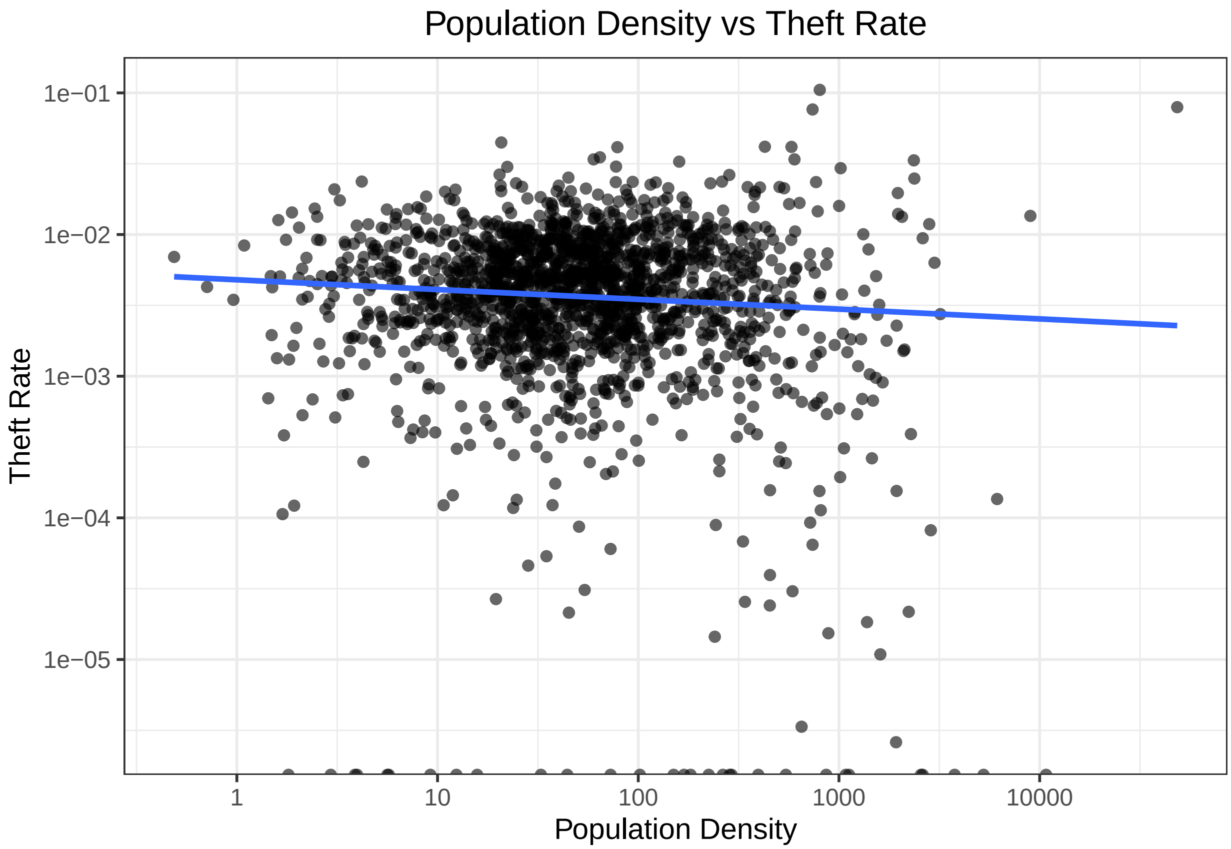


```
theft_train %>% arrange(desc(housing_density))%>%head(5)
```

```
## # A tibble: 5 x 70
##   state      county   pertrump permale med_age nevermarried widowed fromdifstate
##   <chr>      <chr>       <dbl>   <dbl>   <dbl>        <dbl>   <dbl>        <dbl>
## 1 New York   New York    0.123   0.473    37.5        0.433  0.0406       0.0369
## 2 New Jersey Hudson      0.262   0.497    35.3        0.334  0.0391       0.0293
## 3 Virginia   Arlingt~    0.171   0.500    34.7        0.369  0.0266       0.0637
## 4 New Jersey Essex       0.219   0.481    37.6        0.344  0.0447       0.0171
## 5 New Jersey Union       0.315   0.488    38.7        0.297  0.0492       0.0144
## # ... with 62 more variables: fromabroad <dbl>, divorced <dbl>,
## #   foodstamp <dbl>, households <dbl>, Marriedcouplefamily <dbl>,
## #   single_mom <dbl>, inschool <dbl>, inundergrad <dbl>, ingradprofesh <dbl>,
## #   lessthan_hs <dbl>, bachplus <dbl>, med_income <dbl>, gini <dbl>,
## #   singledad <dbl>, withkids <dbl>, med_2bed <dbl>, foreignborn <dbl>,
## #   unemployed_rate <dbl>, employed_rate <dbl>, no_health_ins <dbl>,
## #   dis5to17 <dbl>, dis18to34 <dbl>, dis35to64 <dbl>, fips <dbl>, ...
```

```
# plot theftrate against pop_density
p3 = theft_train %>% select(-fips, -state, -county) %>%
  ggplot(aes(x = pop_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 = "Population Density",
       y = "Theft Rate",
       title = "Population Density 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 30 rows containing non-finite values (stat_smooth).
```

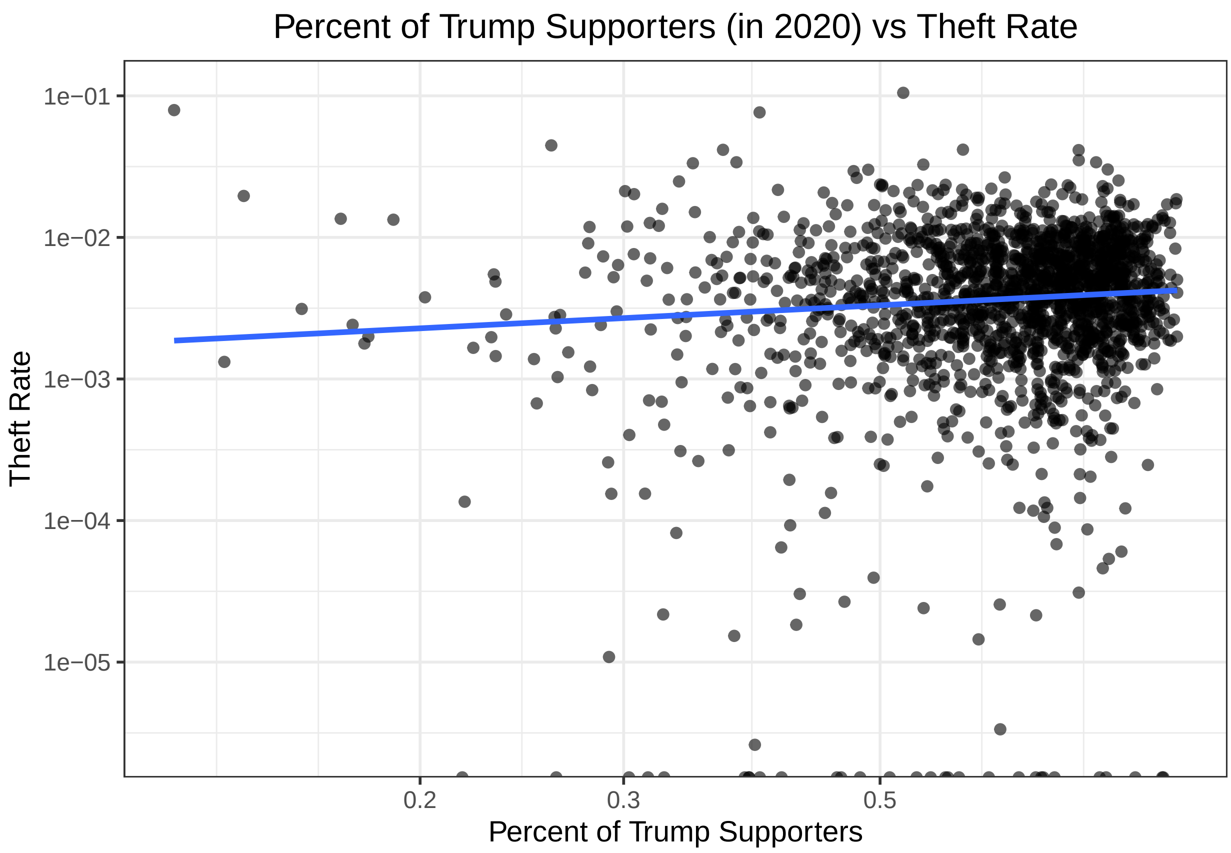


```
theft_train %>% arrange(desc(pop_density))%>%head(5)
```

```
## # A tibble: 5 x 70
##   state      county   pertrump permale med_age nevermarried widowed fromdifstate
##   <chr>      <chr>       <dbl>   <dbl>   <dbl>        <dbl>   <dbl>        <dbl>
## 1 New York   New York    0.123   0.473    37.5         0.433  0.0406       0.0369
## 2 New Jersey Hudson      0.262   0.497    35.3         0.334  0.0391       0.0293
## 3 Virginia   Arlingt~    0.171   0.500    34.7         0.369  0.0266       0.0637
## 4 New Jersey Essex       0.219   0.481    37.6         0.344  0.0447       0.0171
## 5 New Jersey Union       0.315   0.488    38.7         0.297  0.0492       0.0144
## # ... with 62 more variables: fromabroad <dbl>, divorced <dbl>,
## #   foodstamp <dbl>, households <dbl>, Marriedcouplefamily <dbl>,
## #   single_mom <dbl>, inschool <dbl>, inundergrad <dbl>, ingradprofesh <dbl>,
## #   lessthan_hs <dbl>, bachplus <dbl>, med_income <dbl>, gini <dbl>,
## #   singledad <dbl>, withkids <dbl>, med_2bed <dbl>, foreignborn <dbl>,
## #   unemployed_rate <dbl>, employed_rate <dbl>, no_health_ins <dbl>,
## #   dis5to17 <dbl>, dis18to34 <dbl>, dis35to64 <dbl>, fips <dbl>, ...
```

```r
# plot theftrate against pertrump
p4 = 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))

p4
```

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

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

## Warning: Removed 30 rows containing non-finite values (stat_smooth).
```

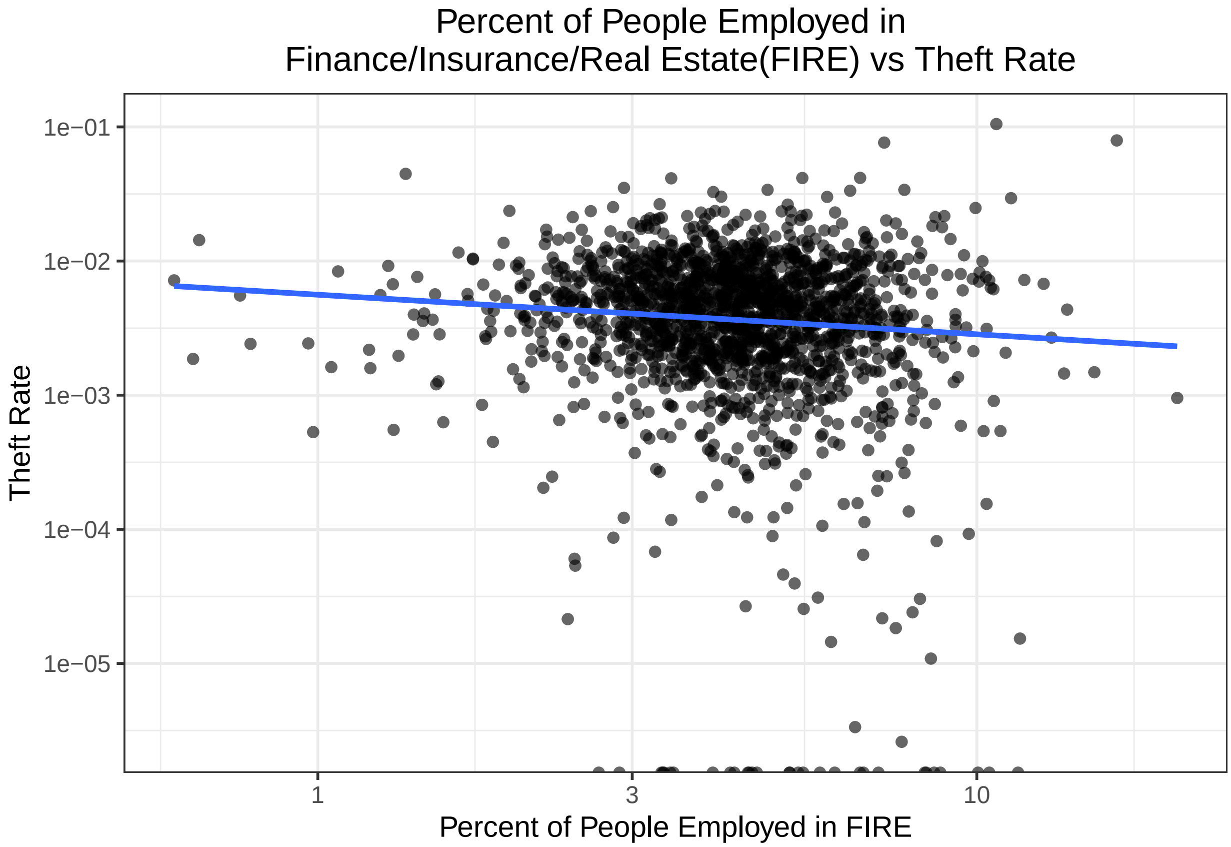


```r
theft_train %>% arrange(desc(pertrump))%>%head(5)
```

```
## # A tibble: 5 x 70
##   state    county   pertrump permale med_age nevermarried widowed fromdifstate
##   <chr>    <chr>       <dbl>   <dbl>   <dbl>        <dbl>   <dbl>        <dbl>
## 1 Texas    Jack        0.904   0.571    39.6        0.239  0.0441      0.00452
## 2 Oklahoma Beaver      0.904   0.503    39          0.179  0.0560      0.0253
## 3 Texas    Hansford    0.903   0.510    35.1        0.197  0.0563      0.0319
## 4 Georgia  Brantley    0.902   0.490    41.1        0.218  0.0713      0.0179
## 5 Oklahoma Ellis       0.901   0.481    44.1        0.166  0.0847      0.0184
## # ... with 62 more variables: fromabroad <dbl>, divorced <dbl>,
## #   foodstamp <dbl>, households <dbl>, Marriedcouplefamily <dbl>,
## #   single_mom <dbl>, inschool <dbl>, inundergrad <dbl>, ingradprofesh <dbl>,
## #   lessthan_hs <dbl>, bachplus <dbl>, med_income <dbl>, gini <dbl>,
## #   singledad <dbl>, withkids <dbl>, med_2bed <dbl>, foreignborn <dbl>,
## #   unemployed_rate <dbl>, employed_rate <dbl>, no_health_ins <dbl>,
## #   dis5to17 <dbl>, dis18to34 <dbl>, dis35to64 <dbl>, fips <dbl>, ...
```

```r
# plot theftrate against PctEmpFIRE
p5 = 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))

p5
```

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

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

## Warning: Removed 30 rows containing non-finite values (stat_smooth).
```

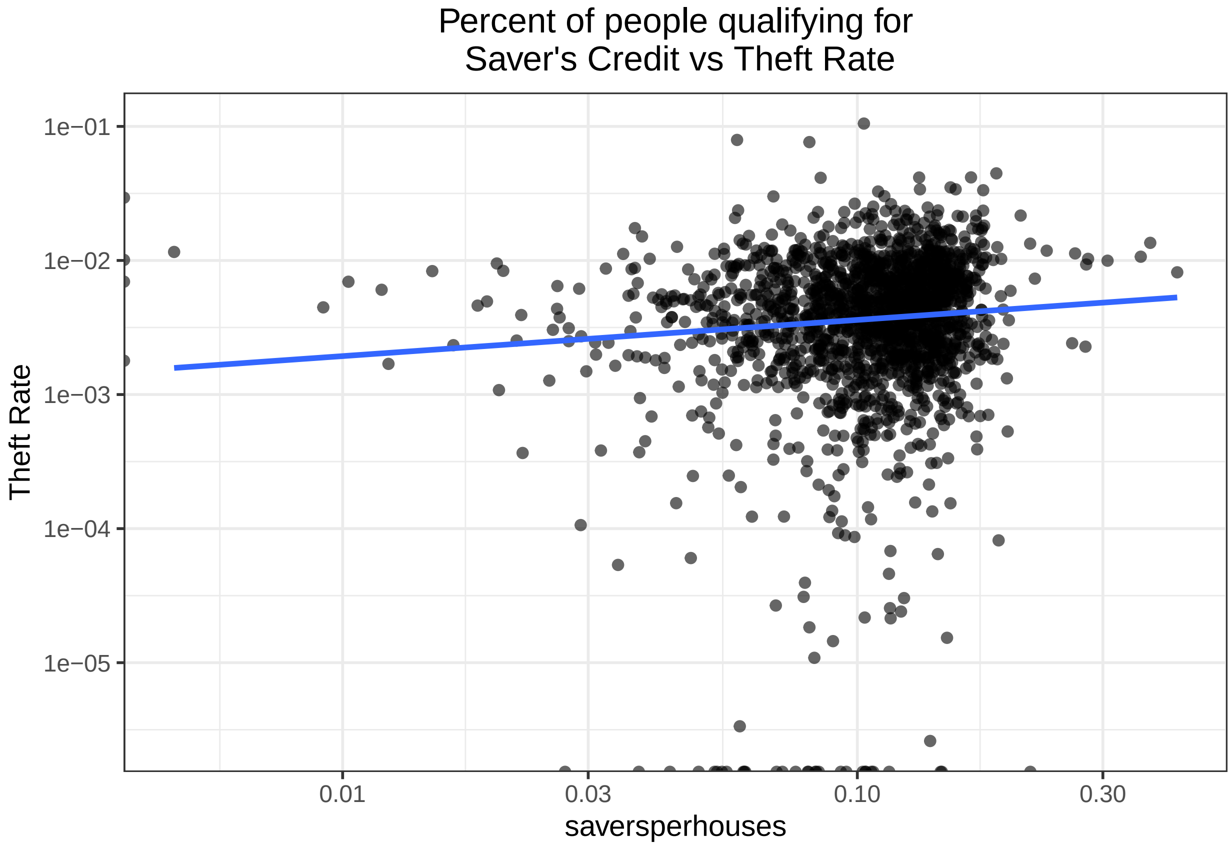


```
theft_train %>% arrange(desc(PctEmpFIRE))%>%select(state, county, PctEmpFIRE, theftrate)%>%head(5)
```

```
## # A tibble: 5 x 4
##   state    county   PctEmpFIRE theftrate
##   <chr>    <chr>         <dbl>     <dbl>
## 1 Iowa     Dallas         20.1  0.000953
## 2 New York New York       16.3  0.0793
## 3 Iowa     Polk           15.1  0.00148
## 4 Ohio     Delaware       13.7  0.00435
## 5 Colorado Pitkin         13.6  0.00145
```

```
# plot theftrate against saversperhouses
p6 = theft_train %>% select(-fips, -state, -county)%>%
  ggplot(aes(x = saversperhouses, y = theftrate)) +
  geom_point(alpha = 0.6) +
  scale_x_log10() +
  scale_y_log10() +
  geom_smooth(method = "lm", formula = "y~x", se = FALSE) +
  labs(x = "saversperhouses",
       y = "Theft Rate",
       title = "Percent of people qualifying for\n Saver's Credit vs Theft Rate") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

p6
```

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

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

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

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

## Warning: Removed 34 rows containing non-finite values (stat_smooth).
```

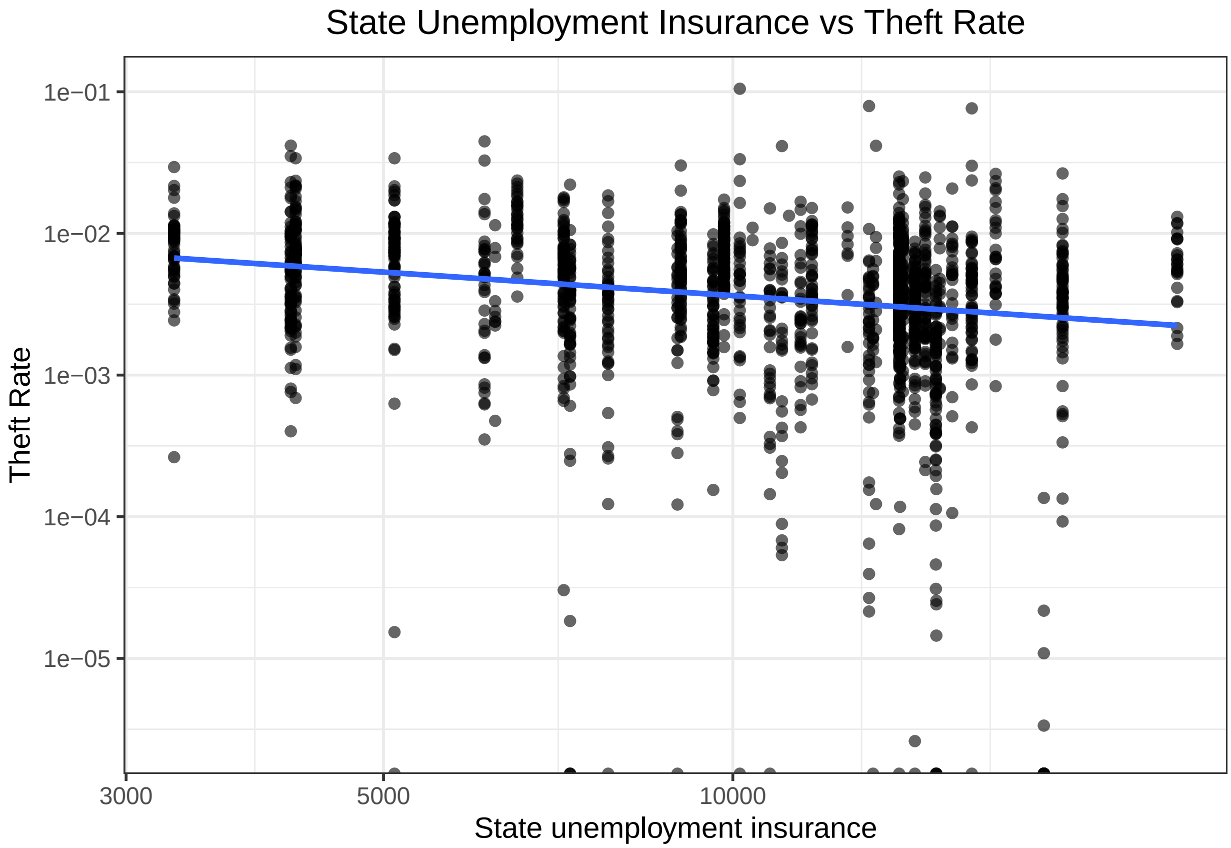


```
theft_train %>% arrange(desc(saversperhouses))%>%select(state, county, saversperhouses, theftrate)%>%hea
```

```
## # A tibble: 5 x 4
##   state    county      saversperhouses theftrate
##   <chr>    <chr>                 <dbl>     <dbl>
## 1 Virginia King George           0.418   0.00816
## 2 Virginia Arlington             0.371   0.0135
## 3 Virginia York                    0.355   0.0107
## 4 Florida  Walton                  0.306   0.00996
## 5 Florida  Bay                     0.281   0.0103
```

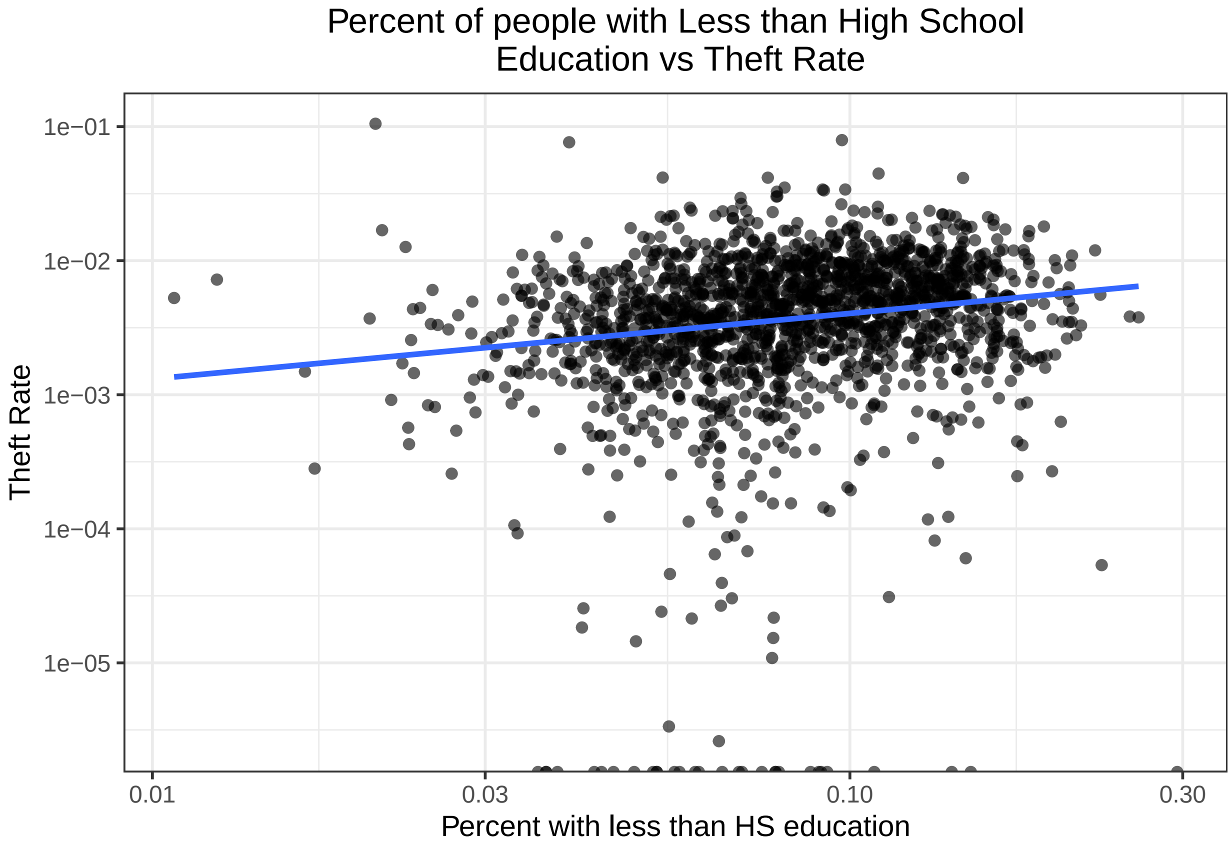
```
# plot theftrate against unemp_bens_possible
p7 = theft_train %>% select(-fips, -state, -county)%>%
  ggplot(aes(x = unemp_bens_possible, y = theftrate)) +
  geom_point(alpha = 0.6) +
  scale_x_log10() +
  scale_y_log10() +
  geom_smooth(method = "lm", formula = "y~x", se = FALSE) +
  labs(x = "State unemployment insurance",
       y = "Theft Rate",
       title = "State Unemployment Insurance vs Theft Rate") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

p7
```

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

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

## Warning: Removed 30 rows containing non-finite values (stat_smooth).
```



```
# plot theftrate against lessthan_hs
p8= theft_train %>% select(-fips, -state, -county)%>%
  ggplot(aes(x = lessthan_hs, 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 with less than HS education",
       y = "Theft Rate",
       title = "Percent of people with Less than High School\n Education vs Theft Rate") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

p8
```

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

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

## Warning: Removed 30 rows containing non-finite values (stat_smooth).
```



```
theft_train %>% arrange(desc(lessthan_hs))%>%select(state, county, lessthan_hs, theftrate)%>%head(5)
```

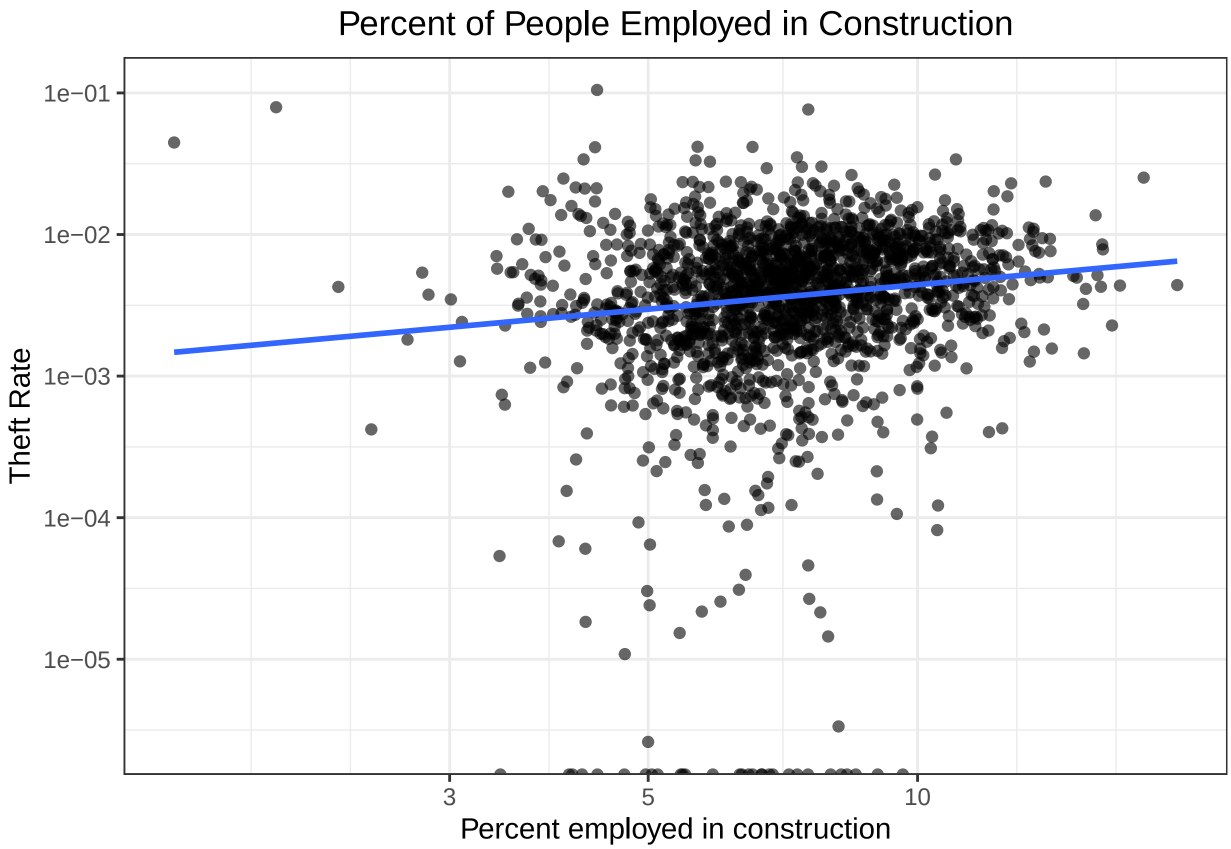
```
## # A tibble: 5 x 4
##   state         county   lessthan_hs theftrate
##   <chr>         <chr>          <dbl>     <dbl>
## 1 Texas         Presidio       0.295 0
## 2 Texas         Starr          0.259 0.00378
## 3 Kentucky      Clay           0.252 0.00383
## 4 West Virginia McDowell       0.230 0.0000536
## 5 Texas         Maverick       0.229 0.00557
```

```
# plot theftrate against PctEmpConstruction
p9= theft_train %>% select(-fips, -state, -county)%>%
  ggplot(aes(x = PctEmpConstruction, 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 employed in construction",
       y = "Theft Rate",
       title = "Percent of People Employed in Construction") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

p9
```

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

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

```
## Warning: Removed 30 rows containing non-finite values (stat_smooth).
```



```
theft_train %>% arrange(desc(PctEmpConstruction))%>%select(state, county, PctEmpConstruction, theftrate
```

```
## # A tibble: 5 x 4
##   state     county       PctEmpConstruction theftrate
##   <chr>     <chr>                     <dbl>     <dbl>
## 1 Texas    Gaines                        19.5   0.00439
## 2 Texas    San Jacinto                   17.9   0.0252
## 3 Wyoming  Lincoln                       16.8   0.00436
## 4 Texas    Caldwell                      16.5   0.00228
## 5 Virginia Mathews                       16.1   0.00785
```