STAT108 Project - Exploratory Data Analysis

Yijia Sun

2022-11-19

Research Question

We want to examine the impact of state-level abortion restrictions on abortions rate and cross-state movement to obtain abortion care in the United States, 2017 - 2019. Out modeling objective is to infer the overall impact on abortion outcomes in the United States.

Data

Our data is collected from Centers for Disease Control and Prevention Abortion Surveillance System, Guttmacher Institute, and Advancing New Standards in Reproductive Health (ANSIRH) from University of California San Francisco.

We have the following variables in the dataset:

- `State` - 50 states in the United States and the District of Columbia
- `year` - 2017, 2018, 2019
- `policy_index` - Score of each states based on whether they had policies in effect in six categories of abortion restrictions and six categories of measures that protect or expand abortion rights and access (6 is most supportive, -6 is most restrictive)
- `abortion_rate_residence` - Number of abortions per 1,000 women aged 15 - 44, by state of residence
- `percentage_leaving` - Percentage of residents obtaining abortions who traveled out of state for care
- `facility_density` - The number of abortion-providing facility per 100,000 women of reproductive age 15 - 49
- `facilities_only_procedural` - Percentage of facilities offering only procedural abortion
- `facilities_only_medication` - Percentage of facilities offering only medication abortion
- `Facilities_both` - Percentage of facilities offering both procedural and medication abortion
- `gestational_limit_medication` - Mean gestational limit for medication abortion
- `gestational_limit_procedural` - Mean gestational limit for procedural abortion
- `cost_medication` - The median self-pay costs for abortion services, in U.S. dollars
- `cost_first_trimester` - The median self-pay costs for first trimester procedural abortion services, in U.S. dollars
- `cost_second_trimester` - The median self-pay costs for second trimester procedural abortion services, in U.S. dollars
- `accepts_insurance` - Percentage of abortion facilities accepting insurance
- `independent_clinics` - Percentage of independent clinics
- `planned_parenthoods` - Percentage of Planned Parenthoods
- `poverty` - Average percentage of people in poverty, 2019 - 2020

Our response variables are `abortion_rate_residence` and `percentage_leaving`.

Section 1 - Data Cleaning

Import Data

```
df <- read_csv("data/new_abortion_data.csv")
df <- df[,-c(1,2)]
```

Categorize state by policy index

We generate a new variable `policy_catog` to categorize states by their policy index. States with scores ranging from -6 to -2 are reported by Guttmacher to be hostile, -1 to +1 are neutral, and +2 to +6 are supportive.

```
df <- df %>% mutate(policy_catog =
                     case_when(policy_index <= -2 ~ "hostile",
                               policy_index <= 1 & policy_index >= -1 ~ "neutral",
                               policy_index >= 2 ~ "supportive")
)
```

Data Cleaning

Since states including California, Maryland, New Hampshire, and Wyoming don't report their data or only reported by occurrence, their abortion rate by state of residence aren't accurate so we remove these states from analysis. We also remove the District of Columbia from analysis, which was not included in the Guttmacher abortion policy report.

```
# filter out California, Maryland, New Hampshire, Wyoming, District of Columbia
df_filtered <- df[is.na(df$percentage_leaving) == FALSE & is.na(df$policy_index) == FALSE, ]
```

Section 2 - Exploratory data analysis

Before building the model, we first conduct exploratory data analysis.

Univariate

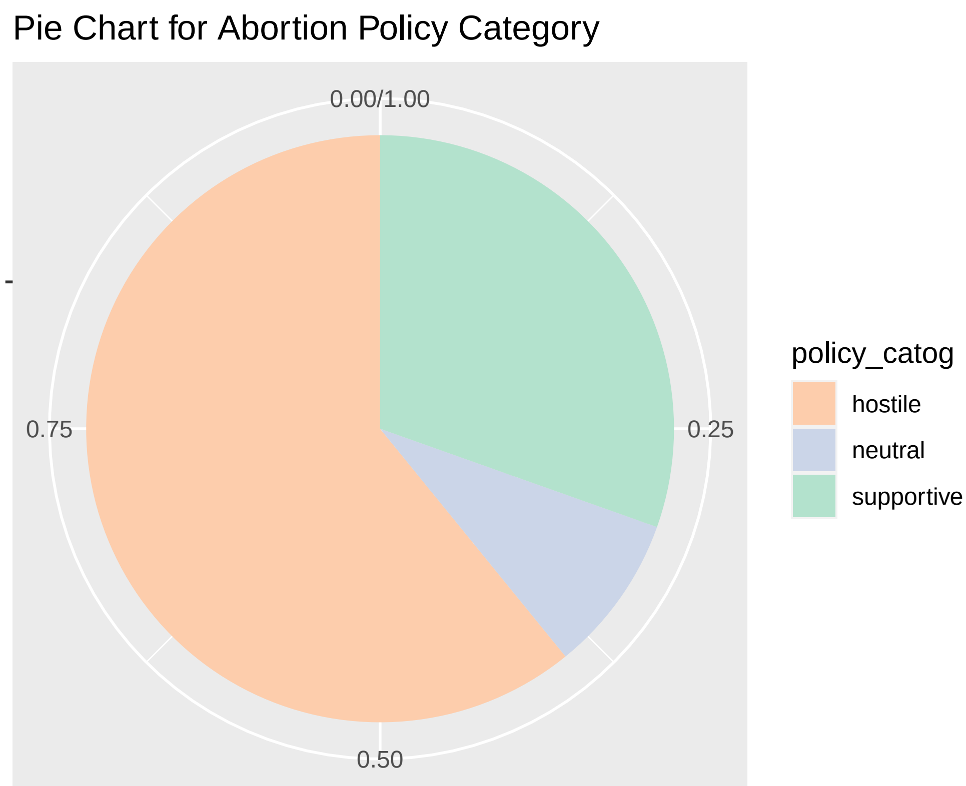
(1) Pie Chart

We create a pie chart and mao for the categorical variable `policy_catog`, in order to see the distribution of policy categories.

```
#pie chart
df_filtered %>%
  count(policy_catog) %>%
  mutate(proportion = n / sum(n))
```

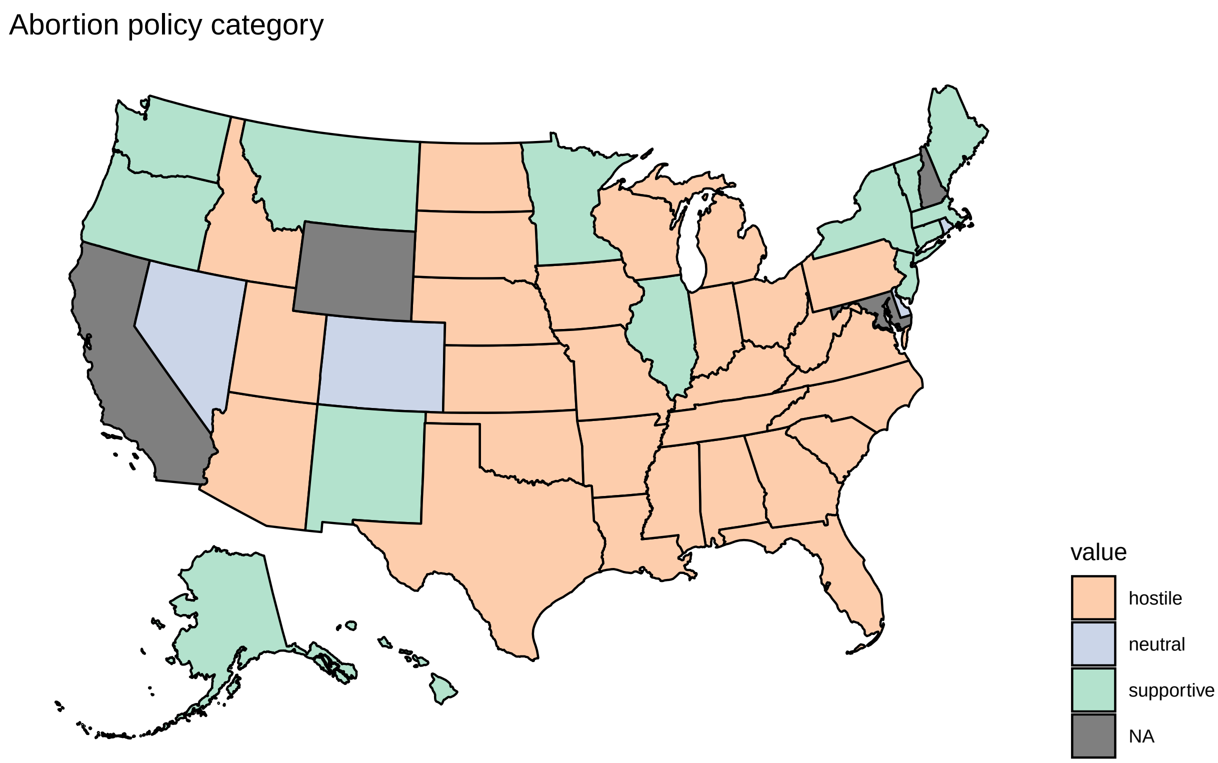
```
## # A tibble: 3 x 3
##   policy_catog     n proportion
##   <chr>        <int>      <dbl>
## 1 hostile         84     0.609
## 2 neutral         12     0.0870
## 3 supportive      42     0.304
```

```
df_filtered %>%
  ggplot(aes(x = "", fill = policy_catog)) +
  geom_bar(position = "fill", width = 1) +
  coord_polar(theta = "y") +
  labs(
    title = "Pie Chart for Abortion Policy Category",
    x = "",
    y = ""
  ) +
  scale_fill_manual(values=c("#fdcdac", "#cbd5e8", "#b3e2cd"))
```




```
#USMAP
#organzing data for state, transfer into abbr
map <- df_filtered[c(1,19)]
map <- map[order(df_filtered$State),]
map <- merge(map, statepop, by.x = "State", by.y = "full", all = TRUE)
map <- map[, -c(5)]
colnames(map) <- c('state', 'value', 'fips', 'abbr')
```

```r
#plot usmap
plot_usmap(data = map, values = "value") + ggtitle("Abortion policy category") + scale_fill_manual(valu
   theme(legend.position = "right")
```



From pie chart and map above, we find:

- `cost_second_trimester` have a large amont of missing value.
- Most states have hostile abortion policy.

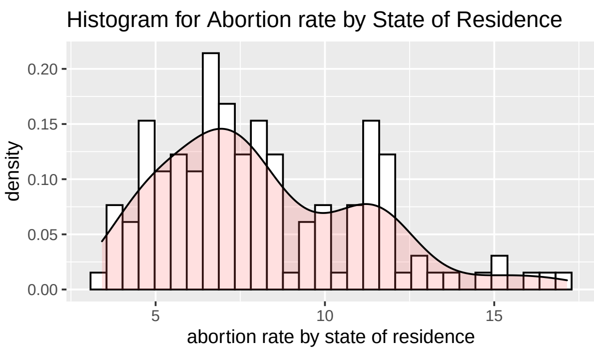
(2) Histograms

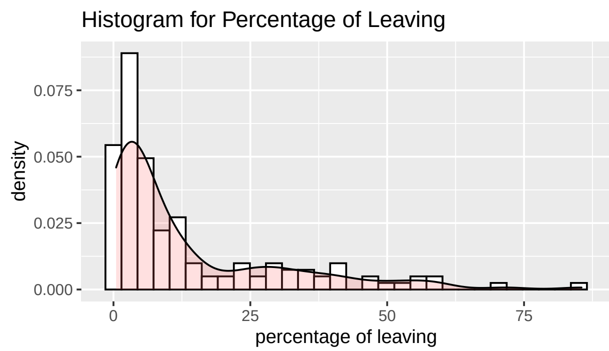
We create histograms to check the distribution of outcome variabels, `abortion_rate_residence` and `percentage_leaving`.

```r
his1 <- ggplot(df_filtered, aes(x = abortion_rate_residence)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666") +
  #facet_grid(year ~ ., scales = "free") +
  xlab("abortion rate by state of residence") + ggtitle("Histogram for Abortion rate by State of Reside

his2 <- ggplot(df_filtered, aes(x = percentage_leaving)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666") +
  #facet_grid(year ~ ., scales = "free") +
```

```
  xlab("percentage of leaving") + ggtitle("Histogram for Percentage of Leaving")

his1 + his2
```

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





From the histograms above, we find:

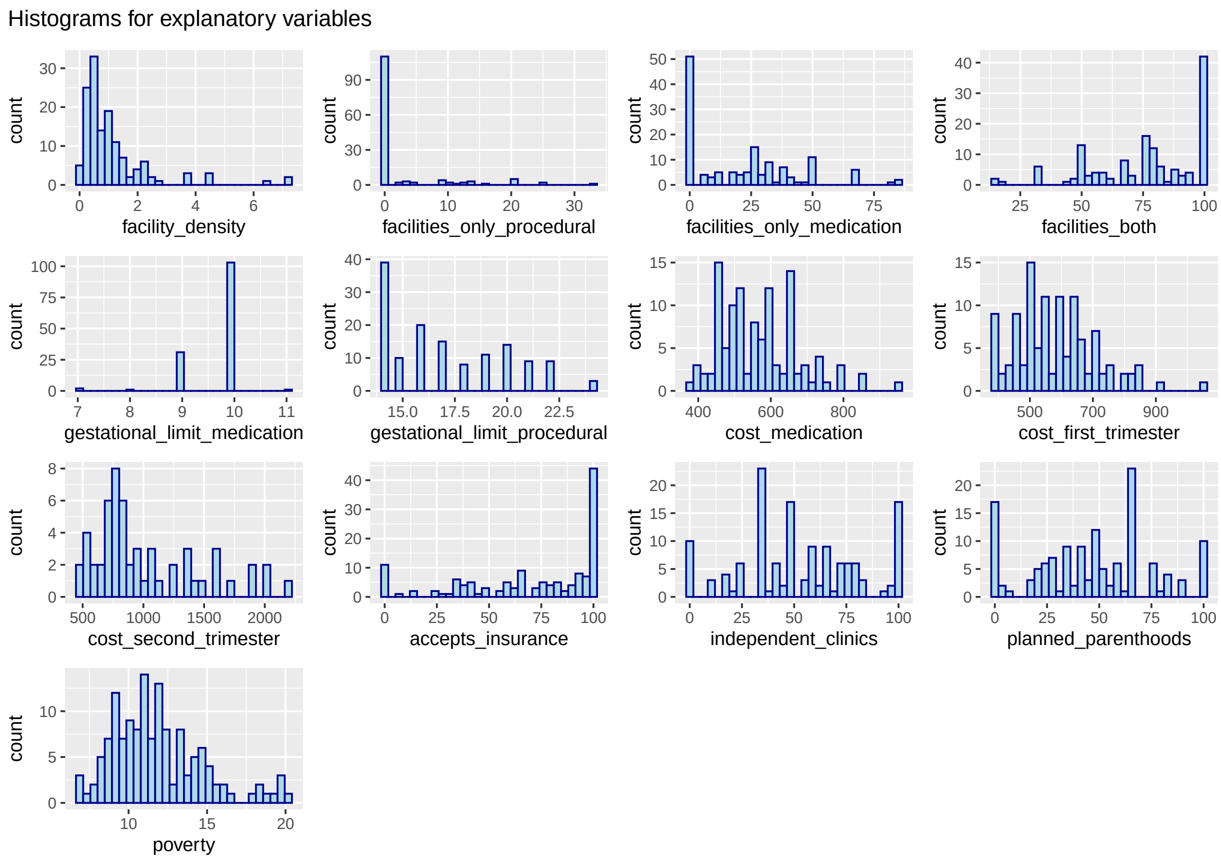
- Distribution for abortion rate is slightly right-skewed.
- Distribution for percentage of leaving is strongly right-skewed.

Then, for all other variables, we create histograms to check their distributions.

```
his3 <- ggplot(df_filtered, aes(x=facility_density)) +
  geom_histogram(color="darkblue", fill="lightblue")
  #+ xlab("facility density")

his4 <- ggplot(df_filtered, aes(x=facilities_only_procedural)) +
  geom_histogram(color="darkblue", fill="lightblue")
  #+ xlab("percentage of facilities with only procedural abortion")

his5 <- ggplot(df_filtered, aes(x=facilities_only_medication)) +
  geom_histogram(color="darkblue", fill="lightblue")
#+ xlab("percentage of facilities with only medication abortion")

his6 <- ggplot(df_filtered, aes(x=facilities_both)) +
  geom_histogram(color="darkblue", fill="lightblue")
#+ xlab("percentage of facilities with both")

his7 <- ggplot(df_filtered, aes(x=gestational_limit_medication)) +
  geom_histogram(color="darkblue", fill="lightblue")
#+ xlab("mean gestational limit of medication abortion")

his8 <- ggplot(df_filtered, aes(x=gestational_limit_procedural)) +
  geom_histogram(color="darkblue", fill="lightblue")
#+ xlab("mean gestational limit of procedural abortion")

his9 <- ggplot(df_filtered, aes(x=cost_medication)) +
```

```r
  geom_histogram(color="darkblue", fill="lightblue")
#+ xlab("mean cost of medication abortion")

his10 <- ggplot(df_filtered, aes(x=cost_first_trimester)) +
  geom_histogram(color="darkblue", fill="lightblue")
#+ xlab("mean cost of abortion at first trimester")

his11 <- ggplot(df_filtered, aes(x=cost_second_trimester)) +
  geom_histogram(color="darkblue", fill="lightblue")
#+ xlab("mean cost of abortion at second trimester")

his12 <- ggplot(df_filtered, aes(x=accepts_insurance)) +
  geom_histogram(color="darkblue", fill="lightblue")
#+ xlab("percentage of clinic accepting insurance")

his13 <- ggplot(df_filtered, aes(x=independent_clinics)) +
  geom_histogram(color="darkblue", fill="lightblue")
#+ xlab("percentage of independent clinics")

his14 <- ggplot(df_filtered, aes(x=planned_parenthoods)) +
  geom_histogram(color="darkblue", fill="lightblue")
#+ xlab("percentage of planned parenthoods")

his15 <- ggplot(df_filtered, aes(x=poverty)) +
  geom_histogram(color="darkblue", fill="lightblue")
#+ xlab("percentage of poverty")


(his3 + his4 + his5 + his6 + his7 + his8 + his9 + his10 + his11 + his12 + his13 + his14 + his15) + plot_
```

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



From the histograms above, we find:

- `facility_density` and `cost_second_trimester` are more right-skewed.
- `facility_both` and `accepts_insurance` are more left-skewed.
- Most states have 0 facility with only procedural or medication.

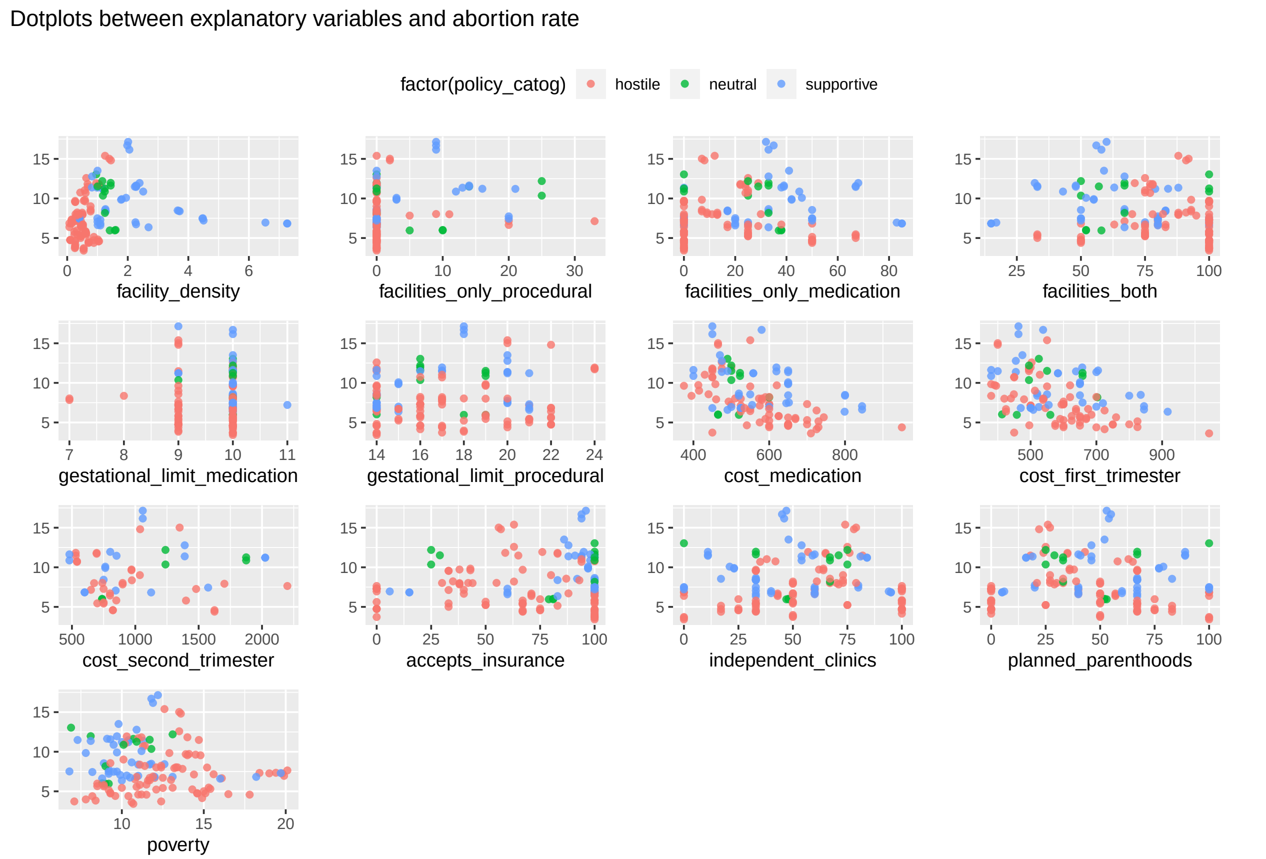
Bivariate

Then, we use dotplots to analyze the relationship between all variables with each outcome variables, grouped by the policy category.

(1) Dotplots for abortion rate

```
p1 <- ggplot(data=df_filtered, aes(x=facility_density,y=abortion_rate_residence, color = factor(policy_c
  geom_point(alpha=0.8)  + ylab(" ")

p2 <- ggplot(data=df_filtered, aes(x=facilities_only_procedural,y=abortion_rate_residence, color = facto
  geom_point(alpha=0.8)  + ylab(" ")

p3 <- ggplot(data=df_filtered, aes(x=facilities_only_medication,y=abortion_rate_residence, color = facto
  geom_point(alpha=0.8)  + ylab(" ")
```

```
p4 <- ggplot(data=df_filtered, aes(x=facilities_both,y=abortion_rate_residence,
                                   color = factor(policy_catog))) +
  geom_point(alpha=0.8)  + ylab(" ")

p5 <- ggplot(data=df_filtered, aes(x=gestational_limit_medication,y=abortion_rate_residence, color = fa
  geom_point(alpha=0.8)  + ylab(" ")

p6 <- ggplot(data=df_filtered, aes(x=gestational_limit_procedural,y=abortion_rate_residence, color = fa
  geom_point(alpha=0.8)  + ylab(" ")

p7 <- ggplot(data=df_filtered, aes(x=cost_medication,y=abortion_rate_residence, color = factor(policy_c
  geom_point(alpha=0.8)  + ylab(" ")

p8 <- ggplot(data=df_filtered, aes(x=cost_first_trimester,y=abortion_rate_residence, color = factor(pol
  geom_point(alpha=0.8)  + ylab(" ")

p9 <- ggplot(data=df_filtered, aes(x=cost_second_trimester,y=abortion_rate_residence, color = factor(po
  geom_point(alpha=0.8)  + ylab(" ")

p10 <- ggplot(data=df_filtered, aes(x=accepts_insurance,y=abortion_rate_residence, color = factor(polic
  geom_point(alpha=0.8)  + ylab(" ")

p11 <- ggplot(data=df_filtered, aes(x=independent_clinics,y=abortion_rate_residence, color = factor(pol
  geom_point(alpha=0.8) + ylab(" ")

p12 <- ggplot(data=df_filtered, aes(x=planned_parenthoods,y=abortion_rate_residence, color = factor(pol
  geom_point(alpha=0.8)  + ylab(" ")

p13 <- ggplot(data=df_filtered, aes(x=poverty,y=abortion_rate_residence,
  color = factor(policy_catog))) +
  geom_point(alpha=0.8)  + ylab(" ")

guide_area() + (p1+p2+p3+p4+p5+p6+p7+p8+p9+p10+p11+p12+p13) +
  plot_layout(guides = "collect",
              nrow = 2, heights = c(1,10)) +
  plot_annotation(title = "Dotplots between explanatory variables and abortion rate") &
  theme(legend.position = "top")
```



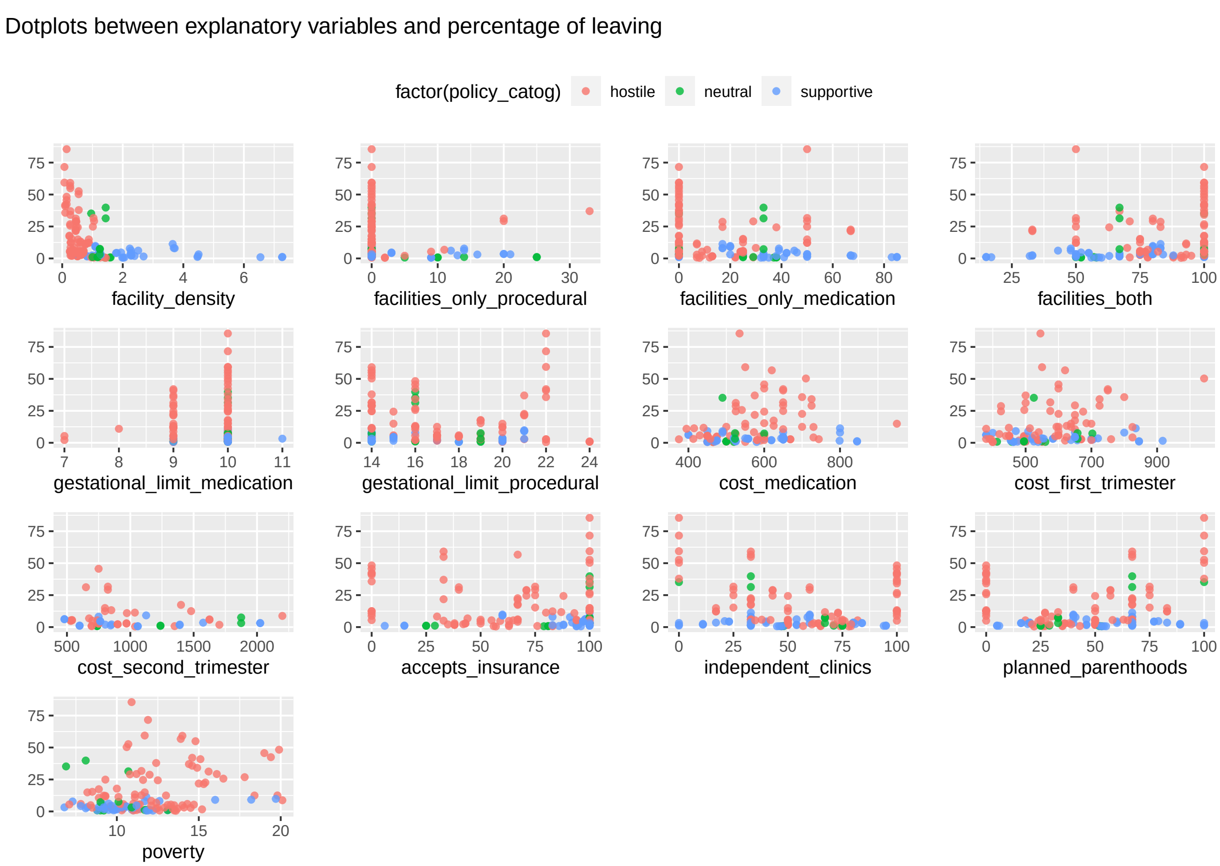
From the dotplots above, we find:

- Hostile states with higher `facility_density` tend to have relatively lower abortion rate.
- `cost_medication` and `cost_fisrt_trimester` both have negative association with abortion rate for all types of states.

(2) Dotplots for percentage of leaving

```
d1 <- ggplot(data=df_filtered, aes(x=facility_density,y=percentage_leaving, color = factor(policy_catog

d2 <- ggplot(data=df_filtered,aes(x=facilities_only_procedural,y=percentage_leaving, color = factor(pol

d3 <- ggplot(data=df_filtered, aes(x=facilities_only_medication,y=percentage_leaving, color = factor(po

d4 <- ggplot(data=df_filtered, aes(x=facilities_both,y=percentage_leaving, color = factor(policy_catog)

d5 <- ggplot(data=df_filtered, aes(x=gestational_limit_medication,y=percentage_leaving, color = factor(

d6 <- ggplot(data=df_filtered, aes(x=gestational_limit_procedural,y=percentage_leaving, color = factor(

d7 <- ggplot(data=df_filtered, aes(x=cost_medication,y=percentage_leaving, color = factor(policy_catog)

d8 <- ggplot(data=df_filtered, aes(x=cost_first_trimester,y=percentage_leaving, color = factor(policy_c
```

```
d9 <- ggplot(data=df_filtered, aes(x=cost_second_trimester,y=percentage_leaving, color = factor(policy_c

d10 <- ggplot(data=df_filtered, aes(x=accepts_insurance,y=percentage_leaving, color = factor(policy_cate

d11 <- ggplot(data=df_filtered, aes(x=independent_clinics,y=percentage_leaving, color = factor(policy_ca

d12 <- ggplot(data=df_filtered, aes(x=planned_parenthoods,y=percentage_leaving, color = factor(policy_ca

d13 <- ggplot(data=df_filtered, aes(x=poverty,y=percentage_leaving,
  color = factor(policy_catog))) + geom_point(alpha=0.8)  + ylab(" ")

guide_area() + (d1 + d2 + d3 + d4 + d5 + d6 + d7 + d8 + d9 + d10 + d11 + d12 + d13) +
  plot_layout(guides = "collect",
              nrow = 2, heights = c(1,10)) +
  plot_annotation(title = "Dotplots between explanatory variables and percentage of leaving") &
  theme(legend.position = "top")
```



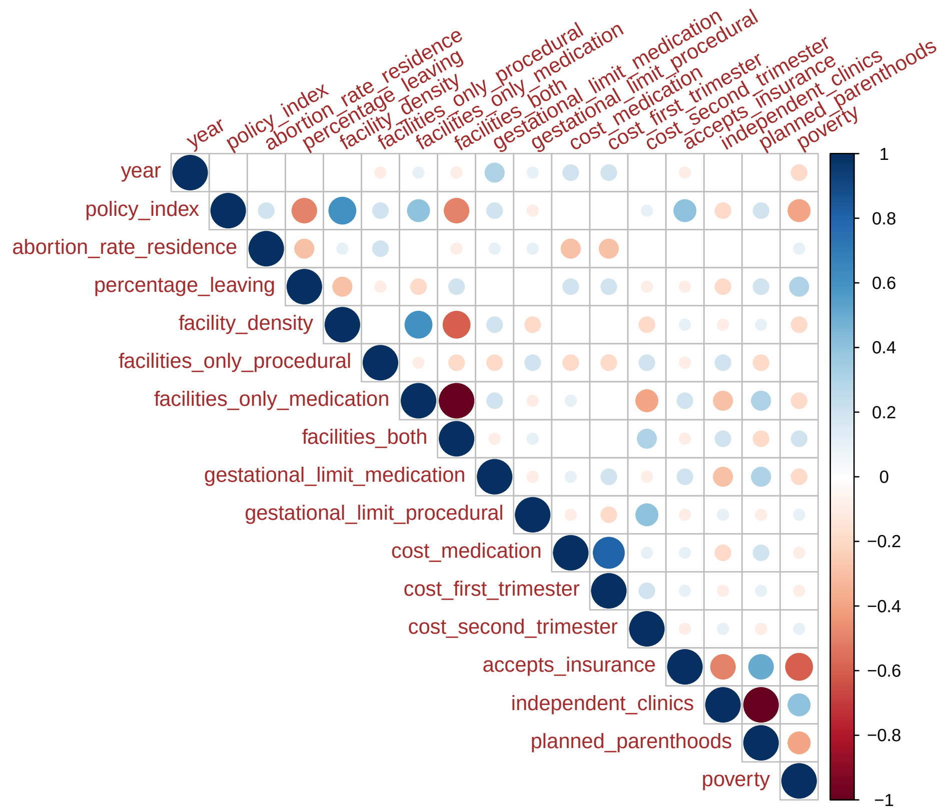
From the dotplots above, we find:

- Hostile states with higher `facility_density` tend to have relatively higher percentage of leaving.
- Hostile states with `gestational_limit_medication` of 9 or 10 tend to have relatively higher percentage of leaving.

Multivariate

Then, to analyze potential interaction terms or multicollinearity, we use correlation matrix to see the correlation coefficient between all variables.

```
correlation <- df[, -c(1,19)]
corr <- round(cor(correlation,use="pairwise.complete.obs"), 1)
corrplot(corr, tl.col = "brown", bg = "White", tl.srt=30, tl.cex =1,type = "upper")
```



From the correlation matrix above, we find:

- `policy_index` & `facility_density`, `accepts_insurance` & `poverty` have relatively strong negative correlation.

- `facilities_only_procedural` & `facilities_both`, `independent_clinics` & `planned_parenthoods` have very strong negative correlation.
- `cost_medication` & `cost_first_trimester` have very strong positive correlation.