PHP2511_HW1

Kuan-Min Lee

2023-02-11

Assignment 1 Simple Linear Regression

This Rmd file is created for Assignment 1 from PHP2511 course taught by Dr. Murillo in Spring 2023 semester at Brown University.

Exploratory Analysis Portion

In this portion of the assignment, the dataset from paper “Association of Highly Restrictive State Abortion Policies With Abortion Rates, 2000-2014” is utilized, and the exploratory data analysis will be performed on this dataset.

Data Previewing:

To start with this, we begin with loading the excel file of the dataset and viewing out the data structure of the dataset.

We start with loading all libraries that are necessary for loading the dataset:

The next step is to read in the excel file of the dataset and view the dimension and variable names of the dataset:

```
dataset_test <- read.csv("state_rep_laws.csv") # read in excel file for the dataset
dim(dataset_test) # get the dimension of the dataset
```

```
## [1] 2173    8
```

```
dataset_test_vars <- colnames(dataset_test) # get the variable names in the dataset
dataset_test_vars
```

```
## [1] "county"                        "state"
## [3] "women"                         "median_income"
## [5] "democrat_2008"                 "dist_to_closest_facility_miles"
## [7] "abortion_count_2010"           "highly_restrictive"
```

From above, we can see the dataset contains 2173 objects and 8 different datatypes.

(a)

Based on the description of the assignment, not all states have their abortion data available to the public. To find out the states publishes their data, the state variables are extracted out and view to see all available.

```
dataset_test_states <- dataset_test$state # extract out state variables
dataset_test_states <- unique(dataset_test_states) # eliminate duplicated state variables, show
up only unique ones
num_dataset_test_states <- length(dataset_test_states) # get the number of states that publish o
ut their data
dataset_test_states
```

```
##  [1] "alabama"        "arizona"        "california"     "colorado"      
##  [5] "delaware"       "florida"        "georgia"        "idaho"         
##  [9] "illinois"       "indiana"        "kansas"         "maine"         
## [13] "michigan"       "minnesota"      "mississippi"    "missouri"      
## [17] "nebraska"       "nevada"         "new york"       "north carolina"
## [21] "north dakota"   "ohio"           "oklahoma"       "oregon"        
## [25] "pennsylvania"   "south carolina" "south dakota"   "tennessee"     
## [29] "texas"          "utah"           "vermont"        "virginia"      
## [33] "washington"     "wisconsin"
```

```
num_dataset_test_states
```

```
## [1] 34
```

From the above analysis, it can be seen that only 34 states shown above have their abortion dataset available to public. There are still 17 states missing from the dataset. To show out the missing states, the following codes are implemented:

```
all_us_states = c("alabama", "alaska", "arizona", "arkansas",
                  "california", "colorado", "connecticut", "delaware", "district of columbia",
                  "florida", "georgia", "hawaii", "idaho", "illinois", "indiana",
                  "iowa", "kansas", "kentucky", "louisiana", "maine", "maryland",
                  "massachusetts", "michigan", "minnesota", "mississippi", "missouri",
                  "montana", "nebraska", "nevada", "new hampshire", "new jersey",
                  "new mexico", "new york", "north carolina", "north dakota", "ohio",
                  "oklahoma", "oregon", "pennsylvania", "rhode island", "south carolina",
                  "south dakota", "tennessee", "texas", "utah", "vermont", "virginia",
                  "washington", "west virginia", "wisconsin", "wyoming") # create a list of all
US state names
not_dataset_test_states <-setdiff(all_us_states,dataset_test_states) # show out the states that
are not available in the dataset
num_not_dataset_test_states <- length(not_dataset_test_states) # show out the number of states t
hat are not available in the dataset
not_dataset_test_states
```

```
##  [1] "alaska"               "arkansas"             "connecticut"
##  [4] "district of columbia" "hawaii"               "iowa"
##  [7] "kentucky"             "louisiana"            "maryland"
## [10] "massachusetts"        "montana"              "new hampshire"
## [13] "new jersey"           "new mexico"           "rhode island"
## [16] "west virginia"        "wyoming"
```

```
num_not_dataset_test_states
```

```
## [1] 17
```

The above states are the unavailable states, and there are 17 of them.

(b)

For this part of the assignment, a new variable names: abortions_per_1000_woman, which is defined as the number of abortions per 1000 women in each county.

To start with this, we first extract out only the data (variables) that will be useful for analysis.

```
dataset_test_counties <- dataset_test$county # extract out county names in the dataset
dataset_test_women <- dataset_test$women # extract out the number of women in each county
dataset_test_abort <- dataset_test$abortion_count_2010 # extract out the number of abortion in e
ach county
temp_dataset <- data.frame(county = dataset_test_counties,
                           num_women = dataset_test_women,
                           num_abort = dataset_test_abort) # form temporarily dataset for analys
is
```

Next, we will extract out the dataset and stored the final outcome into the temporarily dataset

```
abortions_per_women <- temp_dataset$num_abort/temp_dataset$num_women # calculation of the aborti
ons happen in each woman
abortions_per_1000_women <- abortions_per_women*1000 # calculation of the abortions happen in ev
ery 1000 women
abortions_per_1000_women <- data.frame(abortions_per_1000_women = abortions_per_1000_women) # cr
eate a dataframe for the outcome
temp_dataset <- data.frame(temp_dataset,abortions_per_1000_women) # concatenate the outcome back
to the temp table
dataset_test <- data.frame(dataset_test,abortions_per_1000_women) # concatenate the outcome back
to the original table
```

###(c) In this portion,a table will need to be created to view whether a state is a highly restrictive one on abortion or not.

To begin with, we first create a temporarily dataset table for storing only the data that will be utilized to analyze this part.

```
dataset_test_all_states <- dataset_test$state # pull out all states in the dataset
dataset_test_restrict <- dataset_test$highly_restrictive # pull out the highly restrictive, which is defined as a binary variable of 0 represents as not, and 1 represents as yes
temp_dataset_restrictive <- data.frame(state=dataset_test_all_states,restrict=dataset_test_restrict) # create a temporarily table for storing
tab_restrictive <- CreateTableOne(vars = names(temp_dataset_restrictive),strata=names(temp_dataset_restrictive)[2],data=temp_dataset_restrictive)
tab_restrictive
```

```
##                        Stratified by restrict
##                         0            1            p      test
##   n                     1486          687                    
##   state (%)                                       <0.001     
##      alabama               0 ( 0.0)    67 ( 9.8)             
##      arizona              15 ( 1.0)     0 ( 0.0)             
##      california           48 ( 3.2)     0 ( 0.0)             
##      colorado              1 ( 0.1)     0 ( 0.0)             
##      delaware              3 ( 0.2)     0 ( 0.0)             
##      florida              32 ( 2.2)     0 ( 0.0)             
##      georgia             159 (10.7)     0 ( 0.0)             
##      idaho                44 ( 3.0)     0 ( 0.0)             
##      illinois             26 ( 1.7)     0 ( 0.0)             
##      indiana               0 ( 0.0)    82 (11.9)             
##      kansas              105 ( 7.1)     0 ( 0.0)             
##      maine                16 ( 1.1)     0 ( 0.0)             
##      michigan              0 ( 0.0)    83 (12.1)             
##      minnesota            61 ( 4.1)     0 ( 0.0)             
##      mississippi           0 ( 0.0)    82 (11.9)             
##      missouri              0 ( 0.0)   113 (16.4)             
##      nebraska             93 ( 6.3)     0 ( 0.0)             
##      nevada                2 ( 0.1)     0 ( 0.0)             
##      new york             62 ( 4.2)     0 ( 0.0)             
##      north carolina      100 ( 6.7)     0 ( 0.0)             
##      north dakota         53 ( 3.6)     0 ( 0.0)             
##      ohio                  0 ( 0.0)    88 (12.8)             
##      oklahoma             77 ( 5.2)     0 ( 0.0)             
##      oregon               36 ( 2.4)     0 ( 0.0)             
##      pennsylvania          0 ( 0.0)    67 ( 9.8)             
##      south carolina        0 ( 0.0)    46 ( 6.7)             
##      south dakota          0 ( 0.0)    30 ( 4.4)             
##      tennessee            95 ( 6.4)     0 ( 0.0)             
##      texas               253 (17.0)     0 ( 0.0)             
##      utah                  0 ( 0.0)    29 ( 4.2)             
##      vermont              14 ( 0.9)     0 ( 0.0)             
##      virginia             93 ( 6.3)     0 ( 0.0)             
##      washington           39 ( 2.6)     0 ( 0.0)             
##      wisconsin            59 ( 4.0)     0 ( 0.0)             
##   restrict (mean (SD)) 0.00 (0.00)  1.00 (0.00)  <0.001     
```

From the above table, it can be seen that there are 1486 counties out of 2173 are not highly restrictive counties and 687 counties out of 2173 are highly restrictive counties. The ratio of the between the two are:

```
ratio_non_high <- 1486/2173*100
ratio_high <- 687/2173*100
ratio_non_high
```

```
## [1] 68.38472
```

```
ratio_high
```

```
## [1] 31.61528
```

To test out if the data distribution between the highly and non-highly restrictive are different, we can perform the t test between these two group.

```
grouped_data <- dataset_test %>% group_by(highly_restrictive) # group data based on whether or not it's highly restrictive county
rest_data <- grouped_data %>% filter(highly_restrictive==1) # filter out only highly restrictive data
non_rest_data <- grouped_data %>% filter(highly_restrictive==0) # filter out only non highly restrictive data
```

First, it’s the t test between the women residency:

```
t.test(rest_data$women,non_rest_data$women) # t test on the number of women residence
```

```
## 
##  Welch Two Sample t-test
## 
## data:  rest_data$women and non_rest_data$women
## t = -2.7134, df = 2147.5, p-value = 0.006713
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -31204.339  -5022.108
## sample estimates:
## mean of x mean of y 
##  45970.30  64083.52
```

Next, it’s the comparison between the median income:

```
t.test(rest_data$median_income,non_rest_data$median_income) # t test on the number of median income
```

```
##
##  Welch Two Sample t-test
##
## data:  rest_data$median_income and non_rest_data$median_income
## t = -7.1624, df = 1505.2, p-value = 1.236e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4650.578 -2650.937
## sample estimates:
## mean of x mean of y
##  45217.55  48868.30
```

Third, it’s the comparison between the vote for democrat in 2008:

```
t.test(rest_data$democrat_2008,non_rest_data$democrat_2008) # t test on the number of vote for democrat in 2008
```

```
##
##  Welch Two Sample t-test
##
## data:  rest_data$democrat_2008 and non_rest_data$democrat_2008
## t = 5.7684, df = 1562.4, p-value = 9.631e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02313921 0.04698392
## sample estimates:
## mean of x mean of y
## 0.4309675 0.3959059
```

The next analysis is regarding to abortion count in 2010:

```
t.test(rest_data$abortion_count_2010,non_rest_data$abortion_count_2010) # t test on the number of abortion in 2010
```

```
##
##  Welch Two Sample t-test
##
## data:  rest_data$abortion_count_2010 and non_rest_data$abortion_count_2010
## t = -2.9277, df = 2166.3, p-value = 0.00345
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -252.94342  -50.01374
## sample estimates:
## mean of x mean of y
##  190.0699  341.5485
```

The last is the distance to the closest facility providing abortions

```
t.test(rest_data$dist_to_closest_facility_miles,non_rest_data$dist_to_closest_facility_miles) #
t test on the the distance to the closest facility providing abortions
```

```
##
##  Welch Two Sample t-test
##
## data:  rest_data$dist_to_closest_facility_miles and non_rest_data$dist_to_closest_facility_mi
les
## t = -7.2016, df = 2049.5, p-value = 8.318e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -22.14620 -12.66621
## sample estimates:
## mean of x mean of y
##  58.03307  75.43928
```

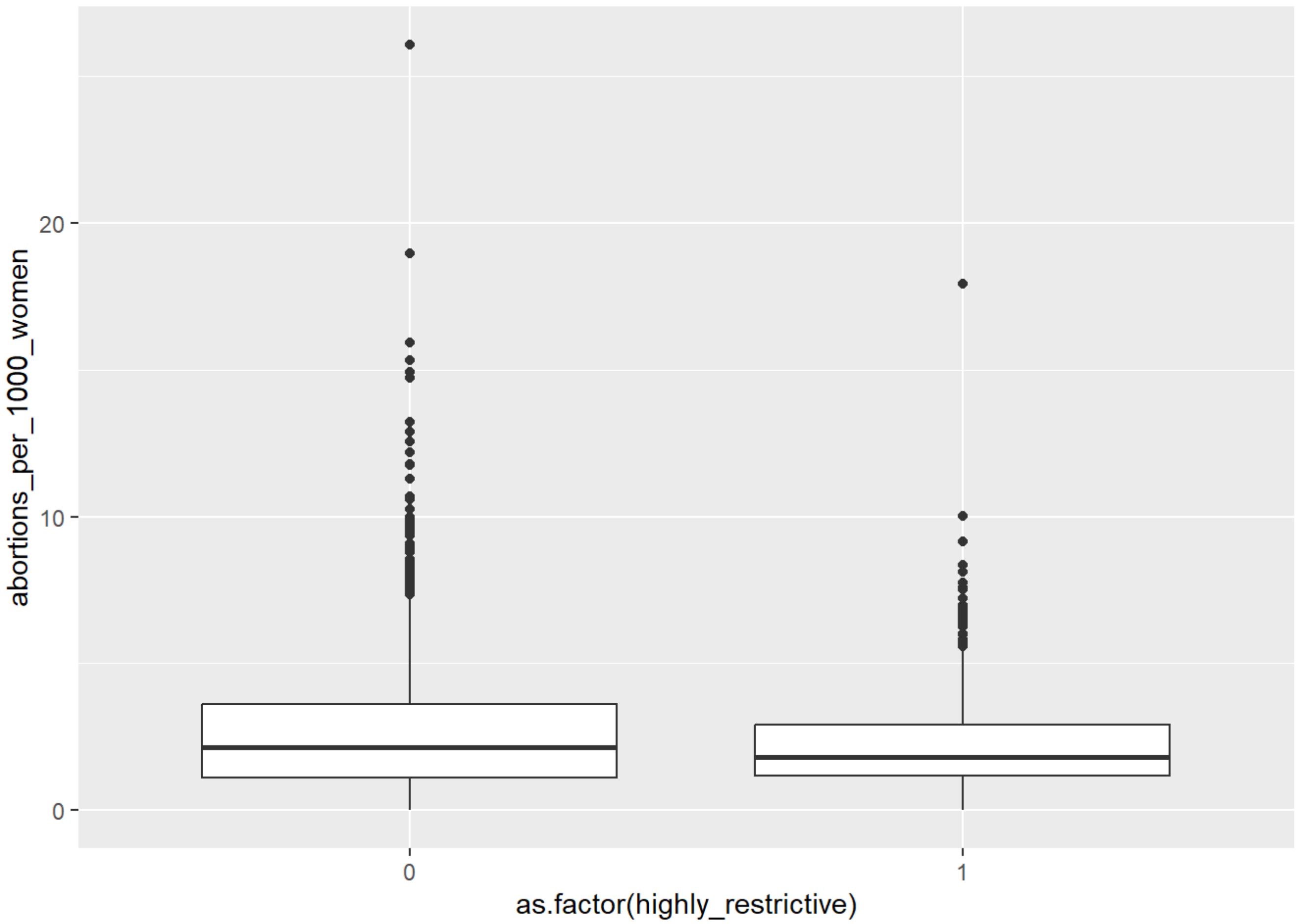
From the above, all above variables demonstrate significantly different data distribution between the highly restrictive and non highly restrictive counties.

(d)

For this portion of the problem, the plot for data distribution of abortion in every 1000 women in both original and log scale will be created and observed.

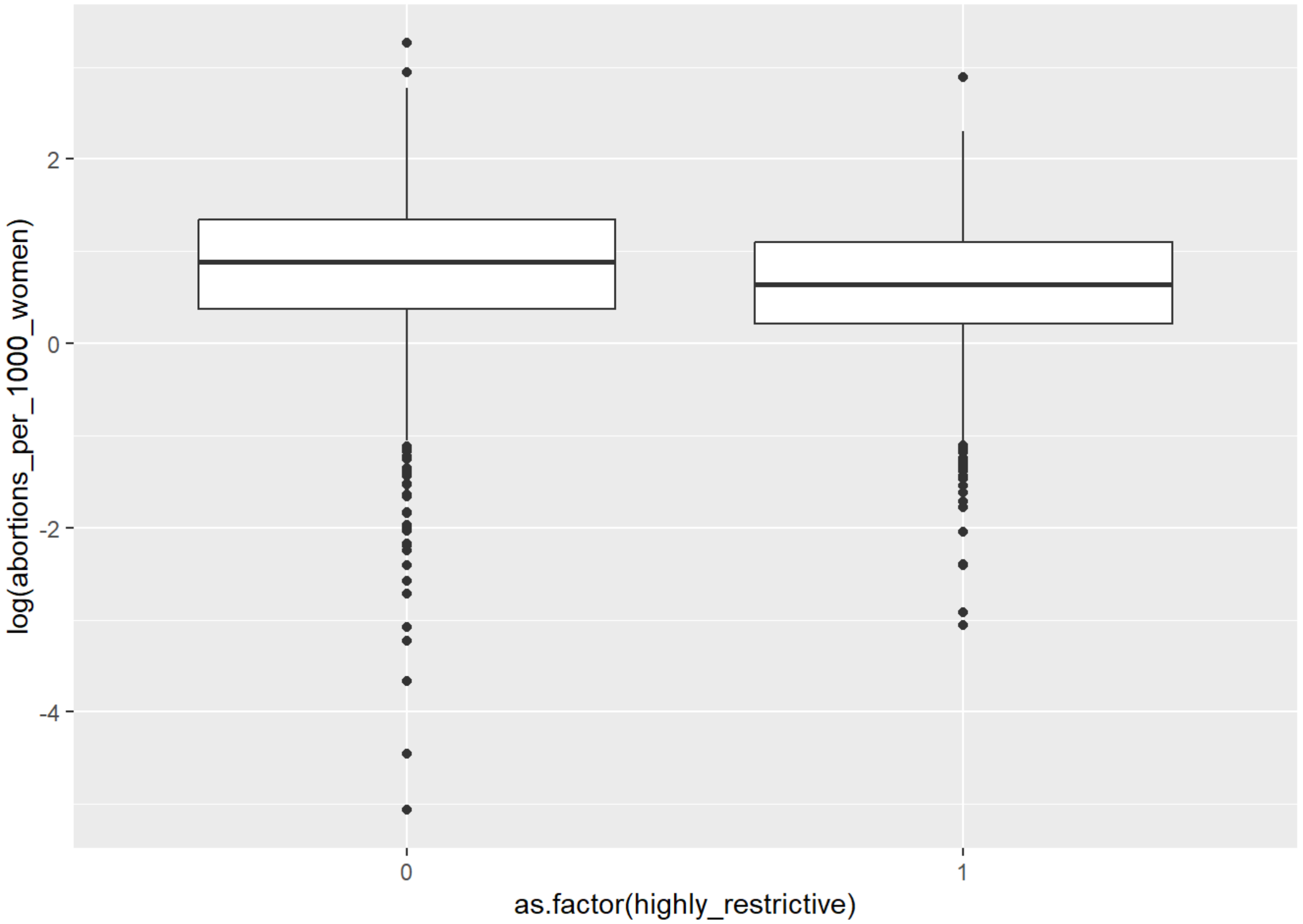
Since it’s categorical variable (highly restrictive, and non highly restrictive), a box plot can be utilized to visualize the data distribution between these two groups:

```
ggplot(dataset_test,aes(x=as.factor(highly_restrictive),y=abortions_per_1000_women))+geom_boxplo
t() # plot the boxplot for abortion per 1000 women based on whether it's a highly restrictive st
ate or not
```



```
ggplot(dataset_test,aes(x=as.factor(highly_restrictive),y=log(abortions_per_1000_women)))+geom_b
oxplot() # plot the boxplot for abortion per 1000 women based on whether it's a highly restricti
ve state or not
```

```
## Warning: Removed 189 rows containing non-finite values (`stat_boxplot()`).
```

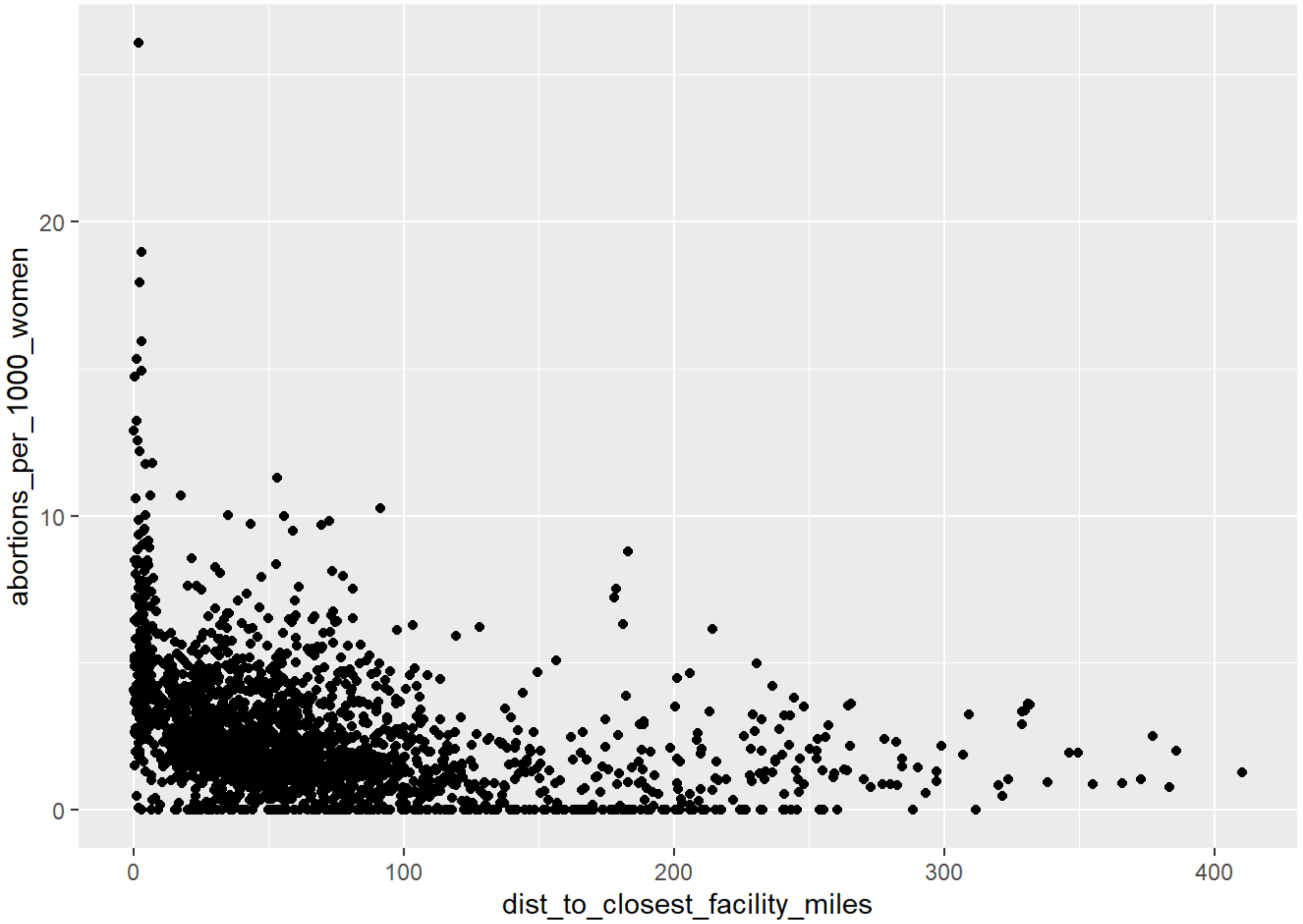


###

(e)

For this part, a scatter plot between the abortion per 1000 women and distance to the closest facility will be created and relationship between the two variables will be observed:

```
ggplot(dataset_test, aes(x=dist_to_closest_facility_miles, y=abortions_per_1000_women)) + geom_p
oint() # create a scatter plot between the two variables: abortion per 1000 women and distance t
o the closest facility
```



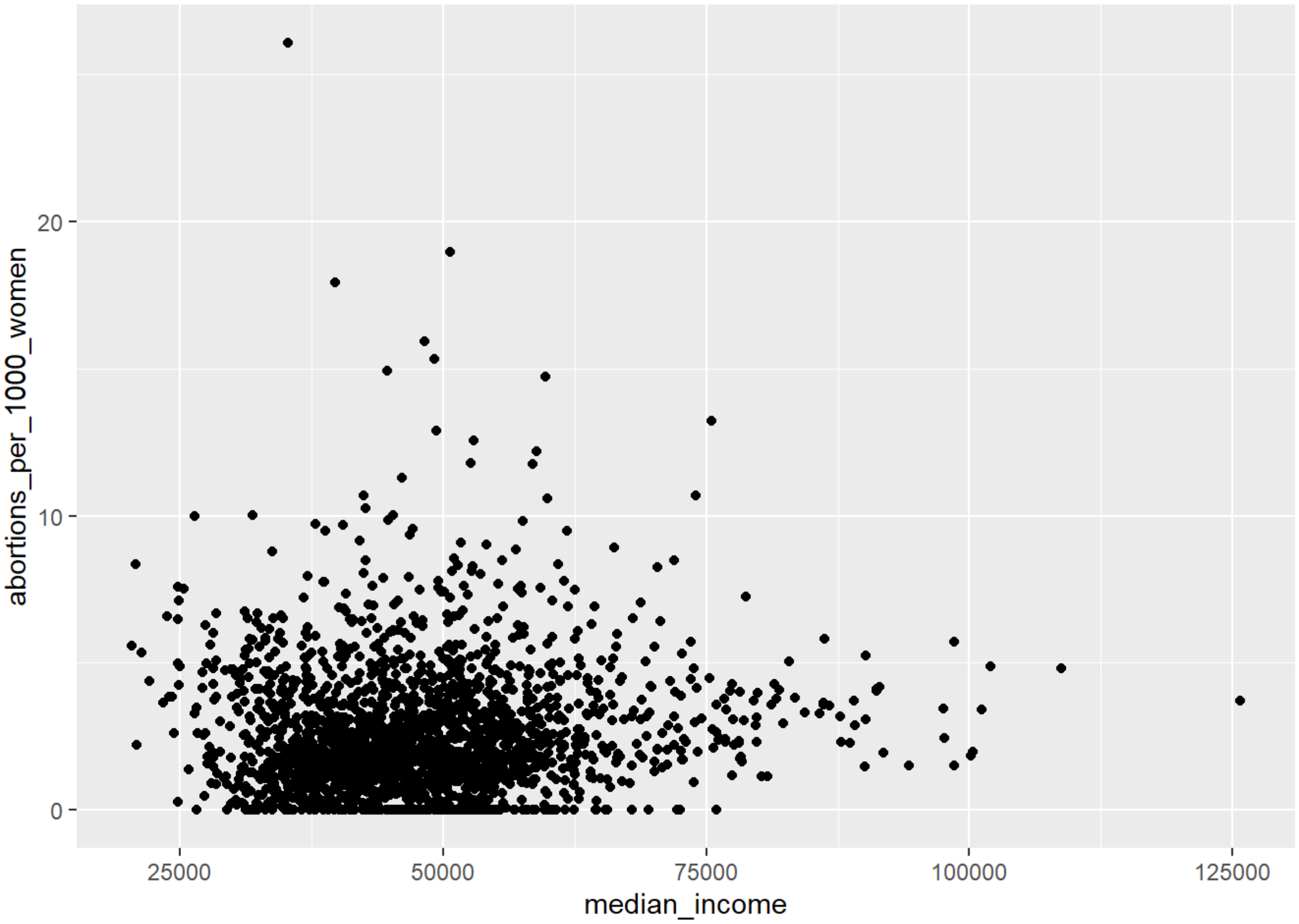
From the above scatter plot, it tends to show the relation that the closer the distance is, the more abortion will happen, and the relation between these two variables tend to follow an exponentail relationship.

(f)

The happening of abortion can be due to multiple aspects. One of them might be due to the financial circumstance. The poorer the family is, the more likely the family will conduct abortion.

To view this, another scatter plot is created to observe the relationship between these two variables:

```
ggplot(dataset_test, aes(x=median_income, y=abortions_per_1000_women)) + geom_point() # create a scatter plot between the two variables: abortion per 1000 women and median income
```



From the above graph, it does show the tendency that if the financial circumstance is worse, the more likely the family will conduct abortion.

Simple Linear Regression

In this portion of the assignment, two simple linear regression models will be built to further investigate the dataset that has been observed in the previous section:

(a)

In this part, a simple linear regression model will be built for observing the relationship between abortions per 1000 women and the highly restrictive indicator variable.

```
first_regression_mod <- lm(abortions_per_1000_women~factor(highly_restrictive), data=dataset_tes
t) # building up a regression model with highly restrictive indicator as predictor and abortions
_per_1000_women as the response variables
summary(first_regression_mod)
```

```
##
## Call:
## lm(formula = abortions_per_1000_women ~ factor(highly_restrictive),
##     data = dataset_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6240 -1.3611 -0.4520  0.9049 23.4286
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.62401    0.05603  46.833  < 2e-16 ***
## factor(highly_restrictive)1 -0.36197    0.09965  -3.633 0.000287 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.16 on 2171 degrees of freedom
## Multiple R-squared:  0.006041,   Adjusted R-squared:  0.005583
## F-statistic:  13.2 on 1 and 2171 DF,  p-value: 0.0002872
```

From the above outcomes, the intercept (or beta0) is calculated as 2.624 and the coefficient for highly_restrictive (or beta1) is calculated as -0.362, which indicates that if the state is highly restrictive on abortion, it will tend to have less abortion number compared to the non highly restrictive one.

(b)

The above linear regression model will be shown as:

abortion_per_1000_women = 2.624 + (-0.362) * highly_restrictive + residual

This linear regression is built based on the following assumptions:

1. It’s assumed that the relationship between the abortion per 1000 per women and highly restrictive indicator is linear.No nonlinear relationship should be observed.

2. The residual (or error) should follow normal distribution.

3. The indicator is an independent variable, which indicates that the relationship between highly restrictive and non highly restrictive states should be independent.

(c)

For this part of the assignment, the same procedure is repeated except for the predictor variable this time. This time distance to the closest facility replaces the highly restrictive variable to be the predictor for the model this time.

```
second_regression_mod <- lm(abortions_per_1000_women~dist_to_closest_facility_miles, data=datase
t_test) # building up a regression model with highly restrictive indicator as predictor and abor
tions_per_1000_women as the response variables
summary(second_regression_mod)
```

```
##
## Call:
## lm(formula = abortions_per_1000_women ~ dist_to_closest_facility_miles,
##     data = dataset_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3007 -1.2824 -0.5011  0.8674 22.7352
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.3421438  0.0649891   51.43   <2e-16 ***
## dist_to_closest_facility_miles -0.0119047  0.0006894  -17.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.031 on 2171 degrees of freedom
## Multiple R-squared:  0.1208, Adjusted R-squared:  0.1204
## F-statistic: 298.2 on 1 and 2171 DF,  p-value: < 2.2e-16
```

From the above, output, the intercept (or beta0) at this time is 3.342, and the coefficient (or beta1) at this time for distance to the closest facility is -0.012, which indicates that if the distance to the closest facility is larger, it will tend to have less abortion number compared to the non highly restrictive one.

The above linear regression model will be shown as:

abortion_per_1000_women = 3.342 + (-0.012) * Dist_to_closest_fact + residual

This linear regression is built based on the following assumptions:

1. It’s assumed that the relationship between the abortion per 1000 per women and highly restrictive indicator is linear.No nonlinear relationship should be observed.

2. The residual (or error) should follow normal distribution.