Regression Model Showcase

1. Installations

1.1 install the libraries

```
library(ggplot2)
library(car)
```

```
## Loading required package: carData
```

2.1 load the dataset

```
EV = read.csv("/Users/emilyluo/Desktop/2022 Fall/DSO 510/PROJECT/final/All_Data_Final-Overview.csv") # File Name: All_Data_Final - Overview.csv
```

2. Data Processing

2.1 change the column name

```
colnames(EV)[2] = 'EV_Registration_Num'
colnames(EV)[3] = 'Gas_Price'
colnames(EV)[4] = 'Population'
colnames(EV)[5] = 'Household_Income'
colnames(EV)[6] = 'Charging_Stations'
colnames(EV)[7] = 'Political_Party'
```

2.2 change the datatype

```
EV$EV_Registration_Num = as.numeric(gsub(',','', EV$EV_Registration_Num))

EV$Household_Income = gsub(',', "", EV$Household_Income)
EV$Household_Income = as.numeric(gsub("\\$", "", EV$Household_Income))

EV$Charging_Stations = as.integer(gsub(',', "", EV$Charging_Stations))

EV$Political_Party = as.factor(EV$Political_Party)

EV$Population = as.numeric(gsub(',','', EV$Population))
```

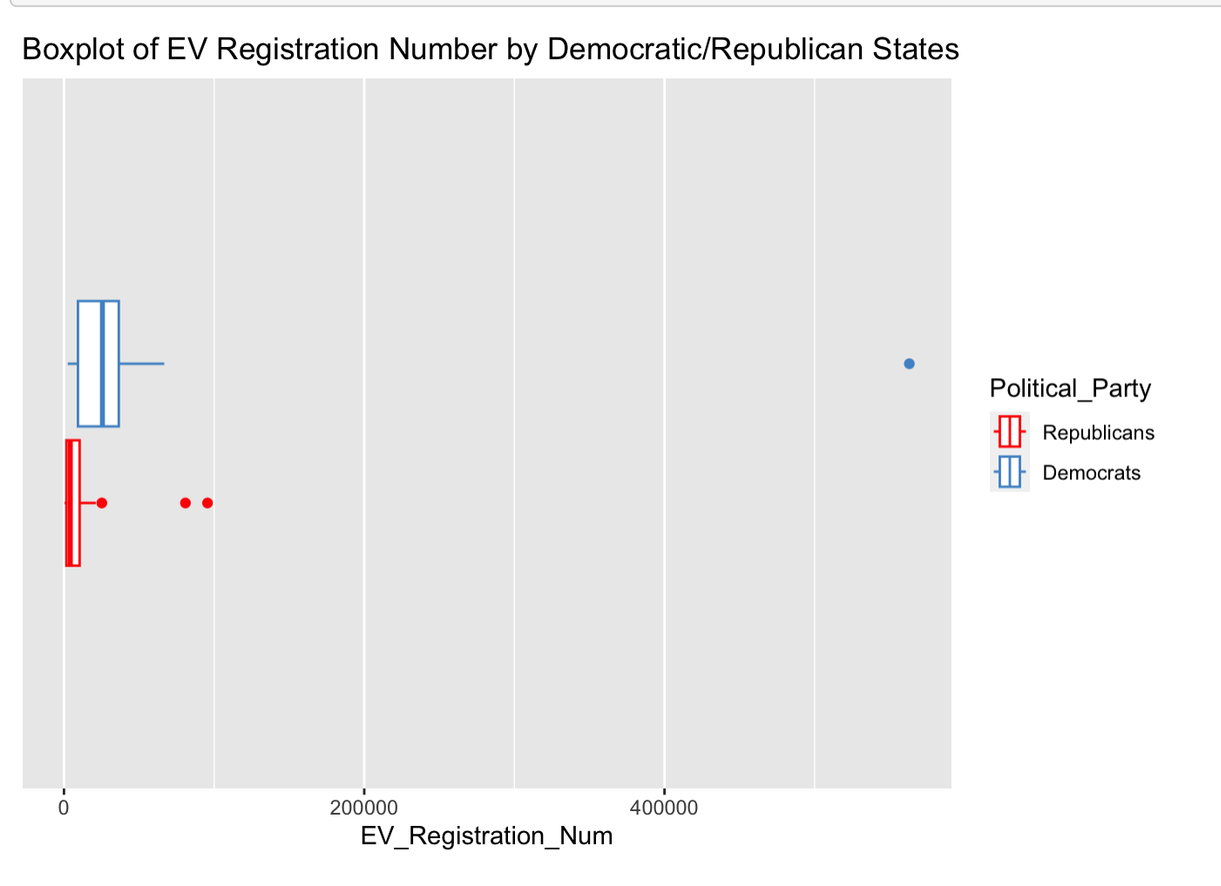
3. Visualization and Descriptive Analysis

- Original Graph

```
options(scipen = 999)
```

3.1 the box plot of the EV registration number by Democratic/Republican states

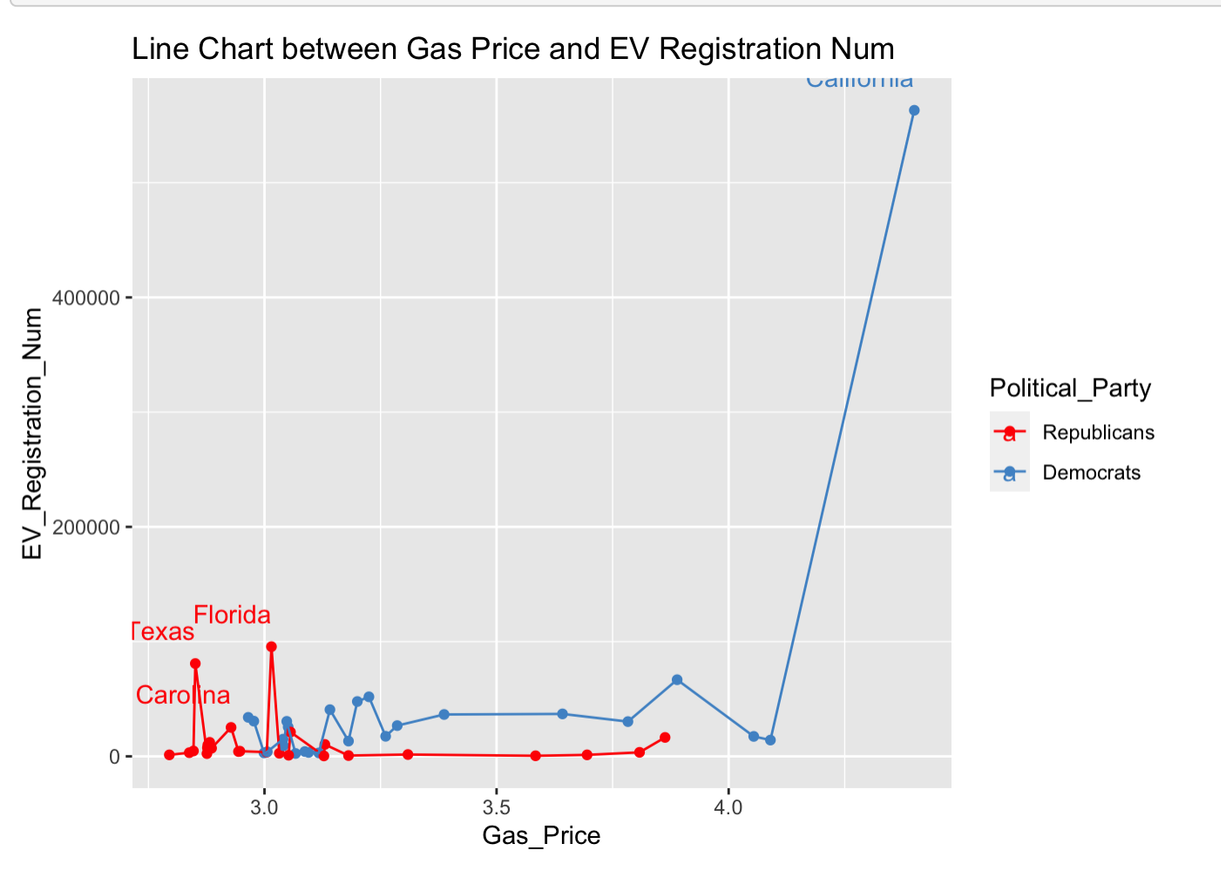
```
ggplot(EV, aes(EV_Registration_Num, color = Political_Party)) +
  geom_boxplot() +
  scale_y_discrete(breaks = "NULL") +
  scale_color_manual(values = c('red', 'steelblue3'), labels = c('Republicans', 'Democrats')) +
  labs(title = 'Boxplot of EV Registration Number by Democratic/Republican States')
```



3.2 the line chart between gas price and EV registration number

- Show the trend of EV registration number as the gas price gets higher
- Also IDENTIFY THE OUTLIERS STATES

```
ggplot(EV, aes(x = Gas_Price, y = EV_Registration_Num, color = Political_Party)) +
  geom_point() +
  geom_line() +
  scale_color_manual(values = c('red', 'steelblue3'), labels = c('Republicans', 'Democrats')) +
  geom_text(aes(label=ifelse((Political_Party == 1 & EV_Registration_Num>400000) | (Political_Party == 0 & EV_Reg
istration_Num>25000),States, ''), hjust=1,vjust=-1.3)) +
  labs(title = 'Line Chart between Gas Price and EV Registration Num')
```



3.3 exclude the outliers

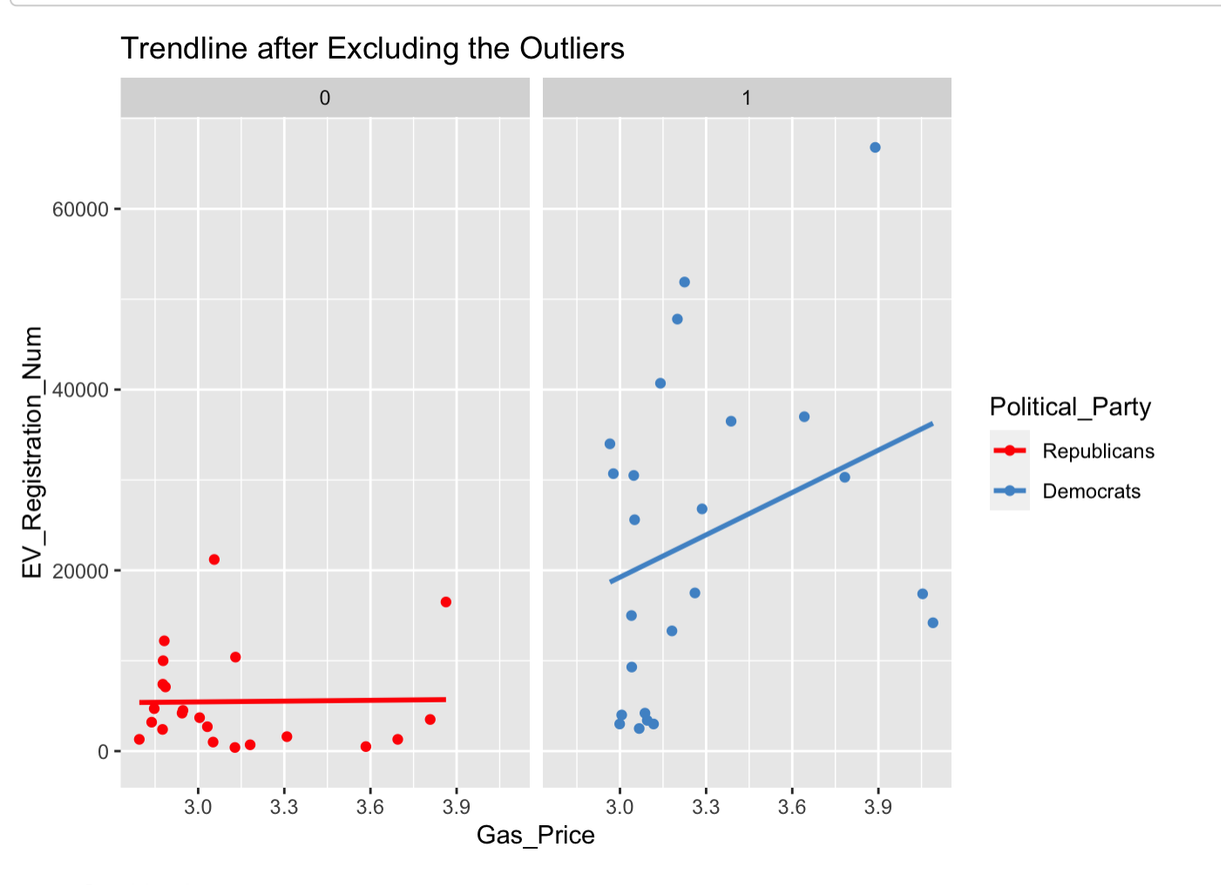
```
EV = EV[EV$States != 'California',] # Exclude the Democratic States Outlier
EVNEW = EV[!(EV$States %in% c('Florida','Texas','North Carolina')),] # Exclude the Republicans States Outliers
```

3.4 draw the linear trend line for both political states after excluding outliers

- Separate

```
ggplot(EVNEW, aes(x = Gas_Price, y = EV_Registration_Num, color = Political_Party)) +
  geom_point() +
  geom_smooth(method = 'lm', fill = NA) +
  scale_color_manual(values = c('red', 'steelblue3'), labels = c('Republicans', 'Democrats')) +
  facet_wrap(~Political_Party) +
  labs(title = 'Trendline after Excluding the Outliers')
```

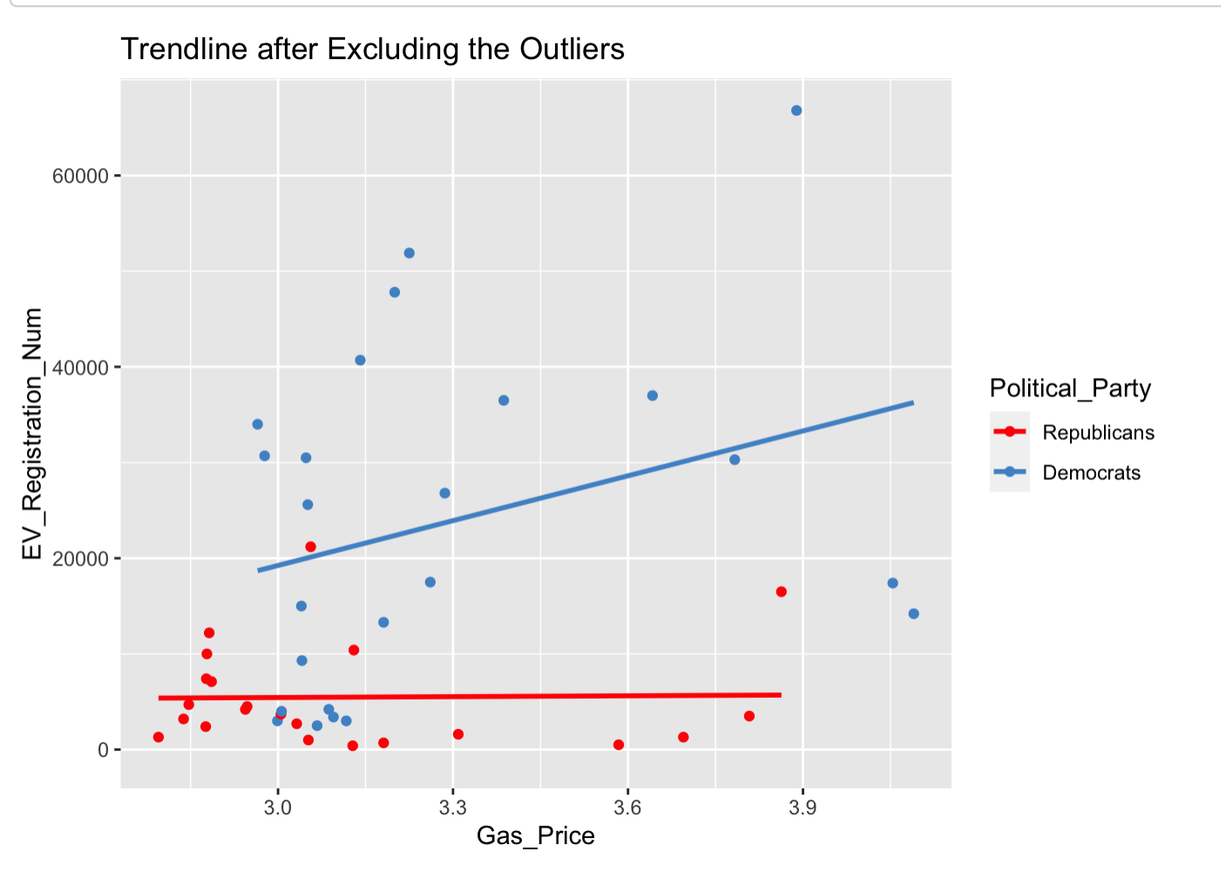
```
## `geom_smooth()` using formula = 'y ~ x'
```



- Combined

```
ggplot(EVNEW, aes(x = Gas_Price, y = EV_Registration_Num, color = Political_Party)) +
  geom_point() +
  geom_smooth(method = 'lm', fill = NA) +
  scale_color_manual(values = c('red', 'steelblue3'), labels = c('Republicans', 'Democrats')) +
  labs(title = 'Trendline after Excluding the Outliers')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



3.5 correlations between multiple potential numeric factors

```
cor(EVNEW[, c(2,3,4,5,6)])
```

```
##                     EV_Registration_Num   Gas_Price  Population
## EV_Registration_Num           1.0000000  0.30945536  0.74091241
## Gas_Price                     0.3094554  1.00000000 -0.08034107
## Population                    0.7409124 -0.08034107  1.00000000
## Household_Income              0.4473394  0.45259553  0.07518320
## Charging_Stations             0.8448172  0.13992779  0.79670675
##                     Household_Income Charging_Stations
## EV_Registration_Num        0.4473394         0.8448172
## Gas_Price                  0.4525955         0.1399278
## Population                 0.0751832         0.7967068
## Household_Income           1.0000000         0.3704219
## Charging_Stations          0.3704219         1.0000000
```

- Conclusion:
  - 1. Charging Station and Population are strongly and positively correlated with EV Registration No.(iv~dv)
  - 2. Charging Station and Population are correlated (iv~iv) - **need to check the VIF No. in the regression model

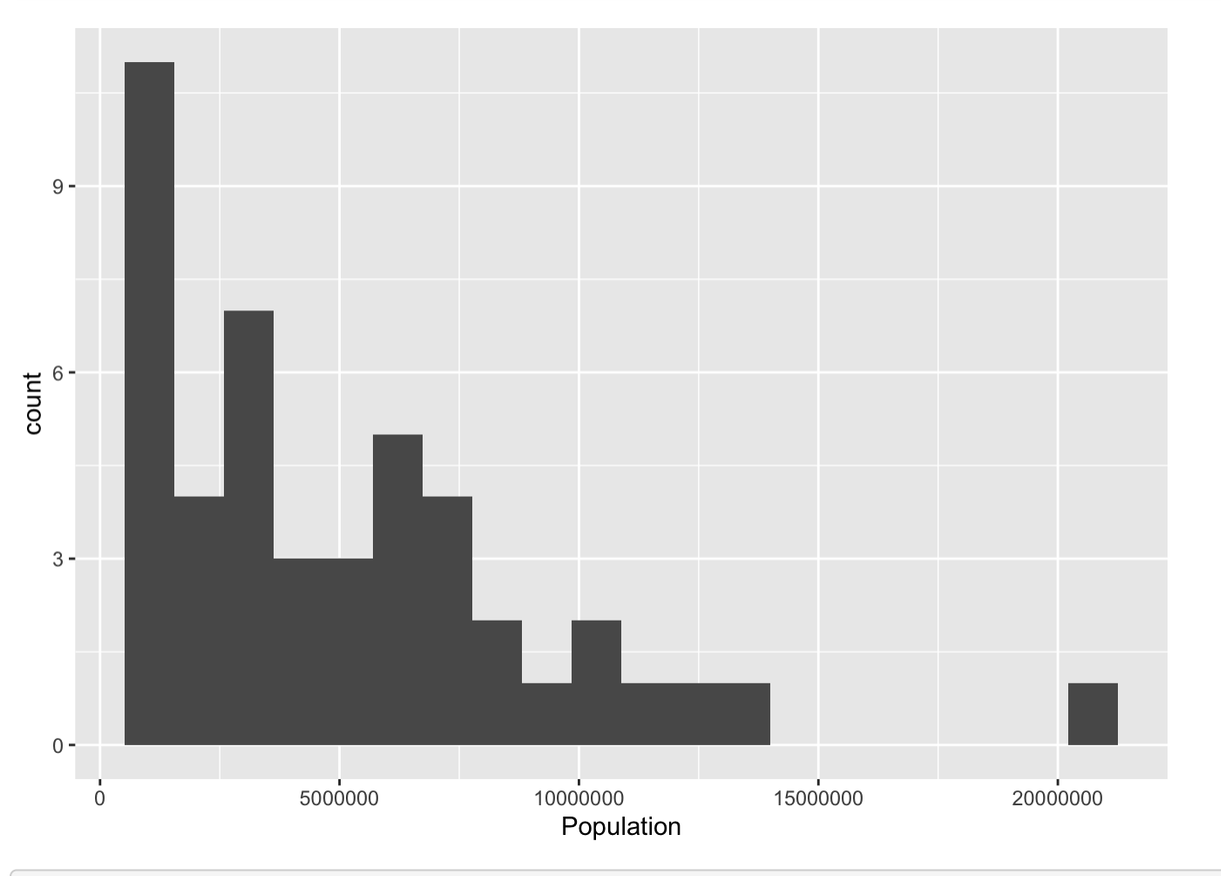
4. Regression Model

4.1 LOG functions:

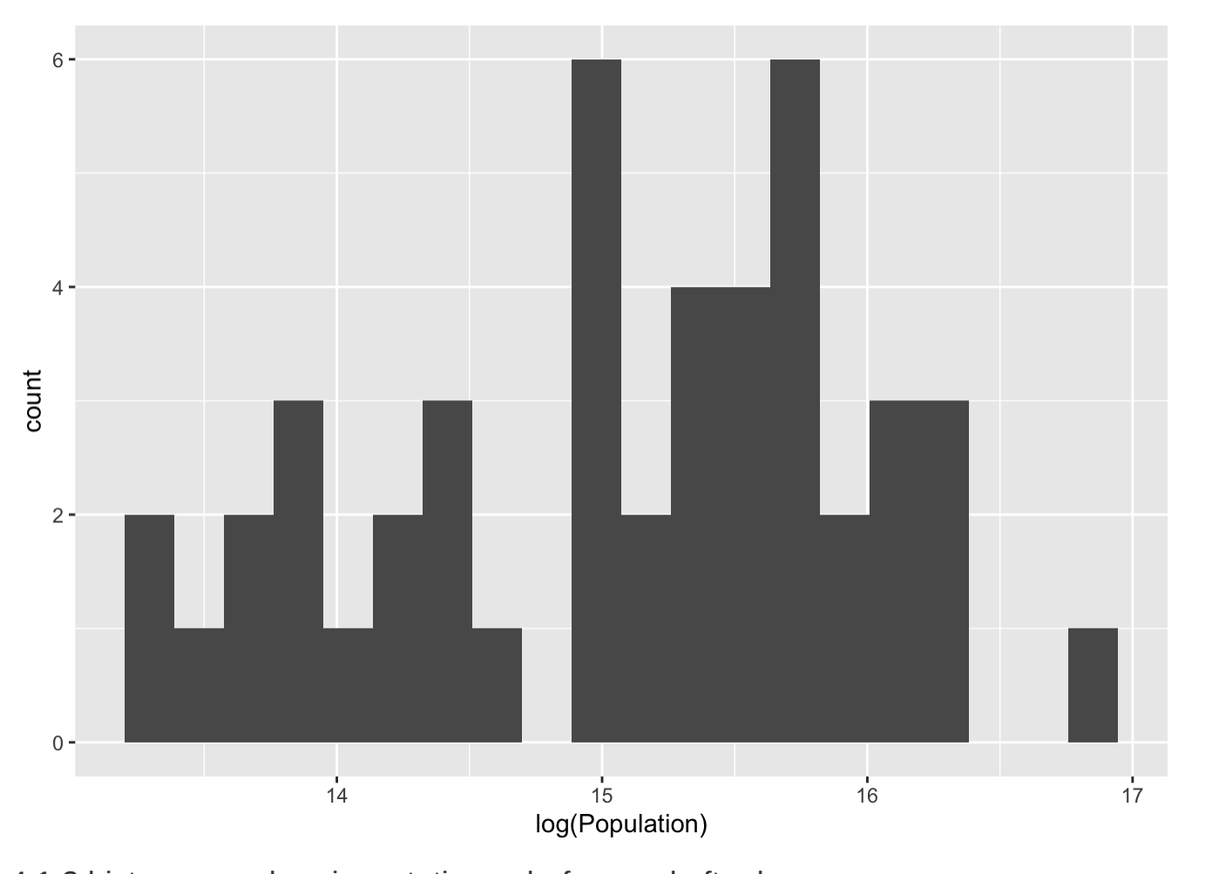
Because we check the histograms for each numeric independent variables and find out all variables except Household_Income are in skewed distribution. For better modeling, we will take the log for all numeric variables except Household Income.

4.1.1 histogram - population - before and after log

```
ggplot(EVNEW, aes(x = Population)) +
  geom_histogram(bins = 20)
```

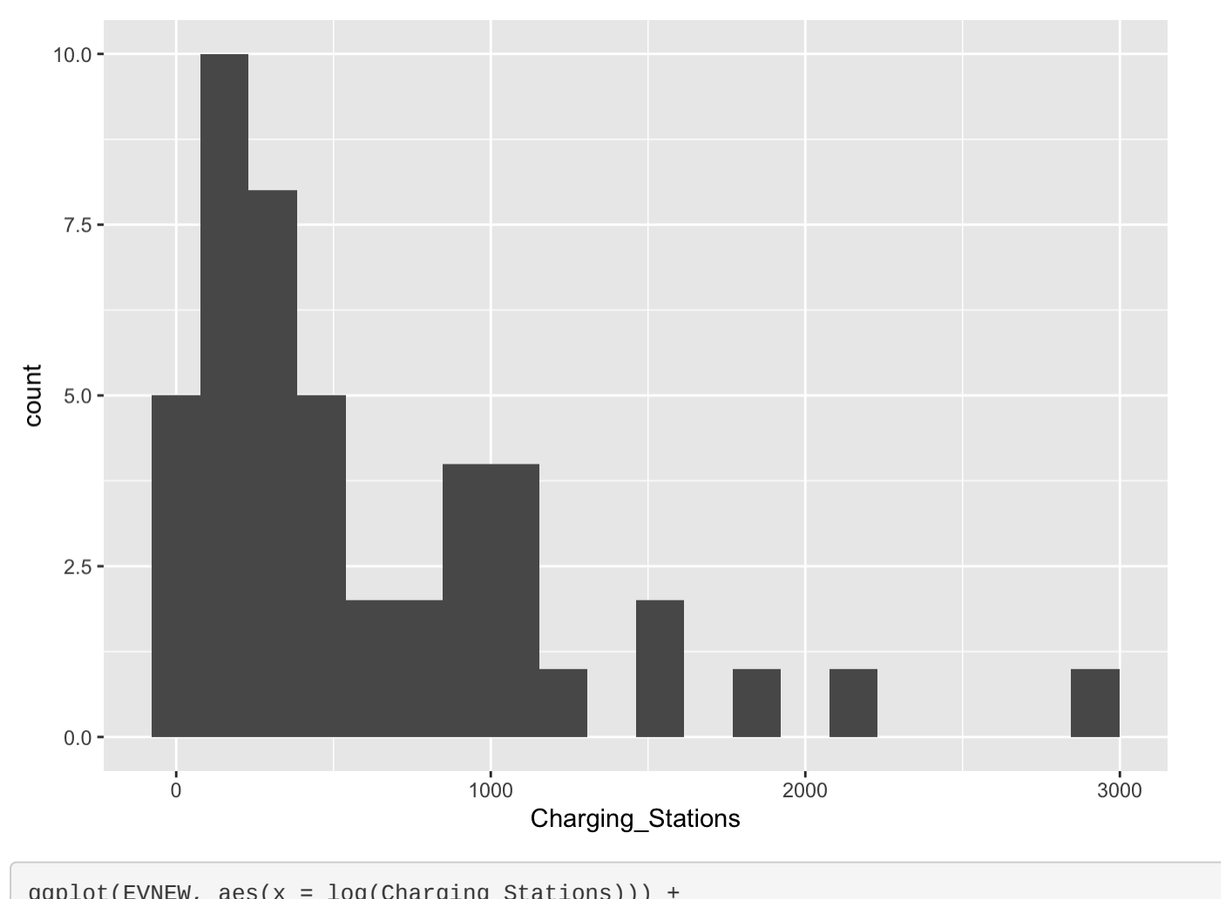



```
ggplot(EVNEW, aes(x = log(Population))) +
  geom_histogram(bins = 20)
```

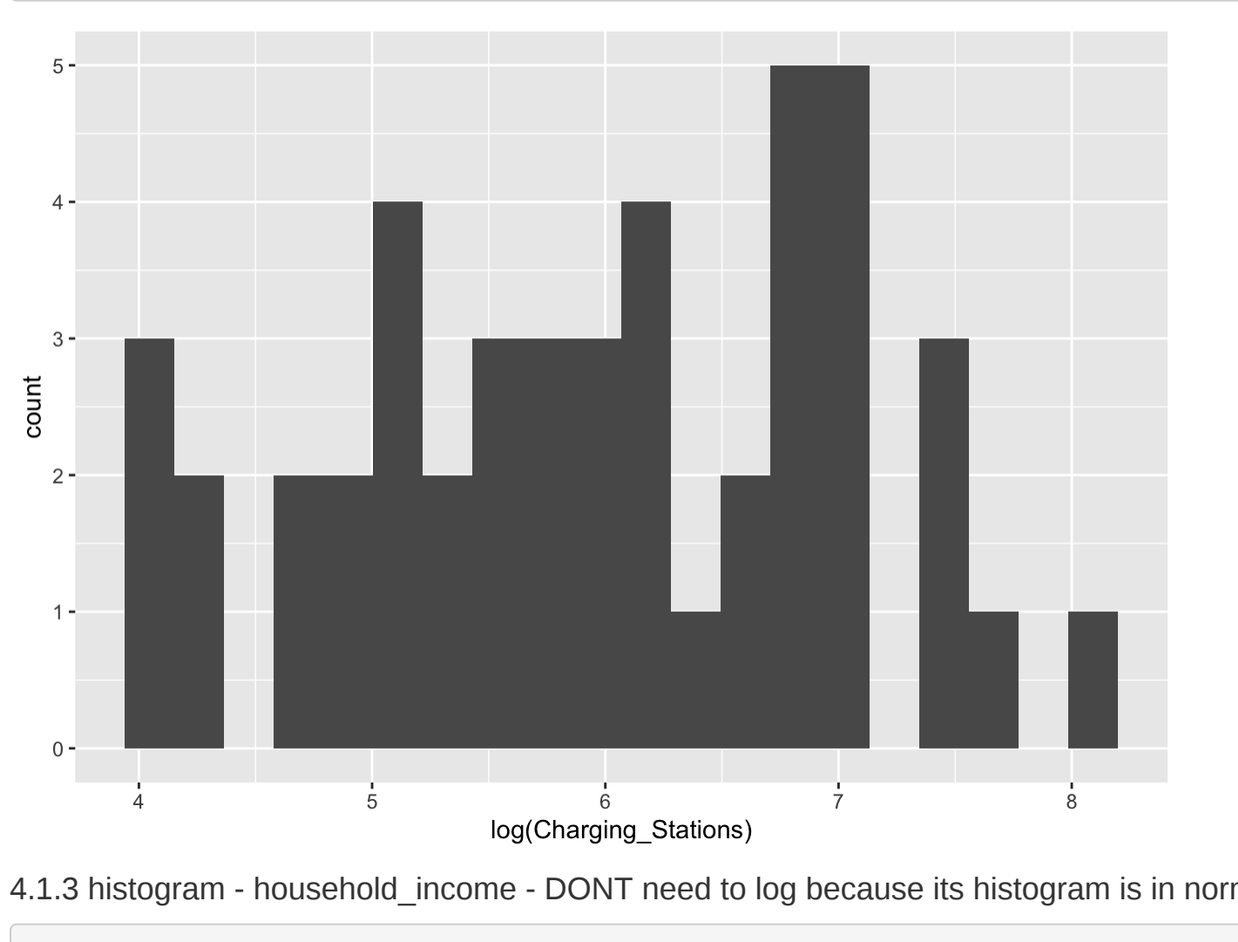


4.1.2 histogram - charging_stations - before and after log

```
ggplot(EVNEW, aes(x = Charging_Stations)) +
  geom_histogram(bins = 20)
```

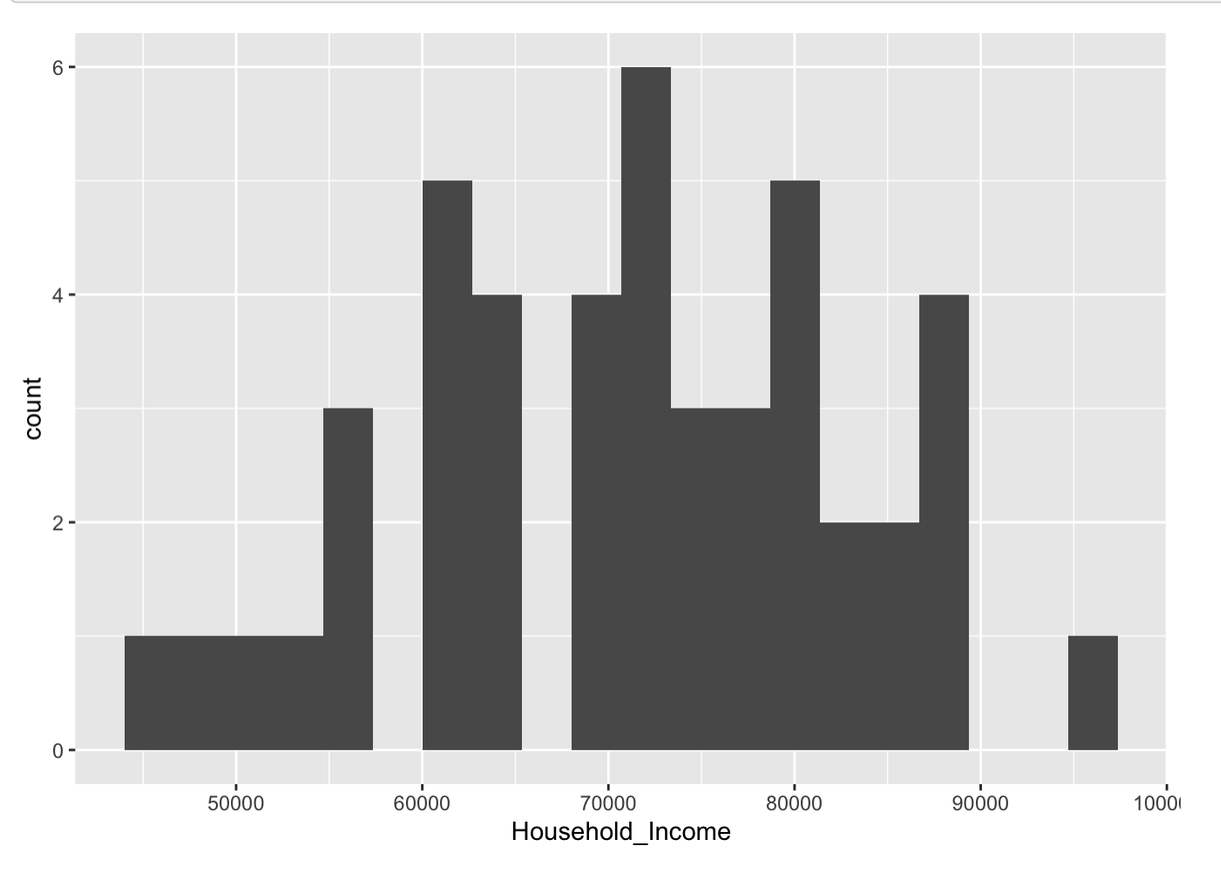



```
ggplot(EVNEW, aes(x = log(Charging_Stations))) +
  geom_histogram(bins = 20)
```



4.1.3 histogram - household_income - DONT need to log because its histogram is in normal distribution

```
ggplot(EVNEW, aes(x = Household_Income)) +
  geom_histogram(bins = 20)
```



5. Regression Model - Simple and Multiple

5.1 regression model 1 - single variable + dummy variable

- LOG(EV_Registration_Num) = b0 + b1*Democrats + b2*LOG(Gas Price) + b3*Democrats*LOG(Gas Price)
  - Democrat States: (b0 + b1) + (b2 + b3)*LOG(Gas Price) - Hypothesis: b3=0?
  - Republican States: b0 + b2*LOG(Gas Price)

```
l_dummy_s = lm(log(EV_Registration_Num)~Political_Party*log(Gas_Price), EVNEW)
summary(l_dummy_s)
```

```
##
## Call:
## lm(formula = log(EV_Registration_Num) ~ Political_Party * log(Gas_Price),
##     data = EVNEW)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0863 -0.8759  0.1276  0.7782  2.0408
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       10.279      2.638   3.897 0.000344 ***
## Political_Party1                  -4.911      3.717  -1.321 0.193599
## log(Gas_Price)                    -1.931      2.320  -0.832 0.410090
## Political_Party1:log(Gas_Price)    5.571      3.203   1.739 0.089373 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.058 on 42 degrees of freedom
## Multiple R-squared:  0.408,  Adjusted R-squared:  0.3657
## F-statistic: 9.647 on 3 and 42 DF,  p-value: 0.00005749
```

- Conclusion:
  - 1. D - P Increase, EV Increase, R - P Increase, EV Decrease - Different than our expectation
  - 2. R-square: 0.40, IV P Value: not low enough to be statistically significant

5.2 regression model 2 - multiple variables + dummy variable

- LOG(EV_Registration) = b0 + b1*Democrats* + *b2*LOG(Gas Price) + b3*LOG(Population)* + *b4*Income + b5*LOG(Charging Stations)* + *b6*Democrats*LOG(Gas Price)
  - Democrat States: (b0 + b1) + (b2 + b6)*LOG(Gas Price)* + *(b3*LOG(Population) + b4*Income* + *b5*LOG(Charging Stations)) - controlling variables
  - Republican States: b0 + b2*LOG(Gas Price)* + *(b3*LOG(Population) + b4*Income* + *b5*LOG(Charging Stations)) - controlling variables
  - Hypothesis: b6 = 0?

```
l_dummy_m_9 = lm(log(EV_Registration_Num)~Political_Party*log(Gas_Price)+log(Population)+Household_Income+log(Cha
rging_Stations), EVNEW)
summary(l_dummy_m_9)
```

```
##
## Call:
## lm(formula = log(EV_Registration_Num) ~ Political_Party * log(Gas_Price) +
##     log(Population) + Household_Income + log(Charging_Stations),
##     data = EVNEW)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68411 -0.23758  0.04389  0.21576  0.63935
##
## Coefficients:
##                                  Estimate  Std. Error t value    Pr(>|t|)
## (Intercept)                     -7.01098366  1.70616364  -4.109    0.000197 ***
## Political_Party1                -0.84978919  1.24623359  -0.682    0.499341
## log(Gas_Price)                   1.41577633  0.85970521   1.647    0.107632
## log(Population)                  0.67310027  0.11460264   5.873 0.000000777 ***
## Household_Income                 0.00001261  0.00000585   2.155    0.037355 *
## log(Charging_Stations)           0.50187501  0.11490059   4.368 0.000089797 ***
## Political_Party1:log(Gas_Price)  1.12108726  1.07396286   1.044    0.302969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3154 on 39 degrees of freedom
## Multiple R-squared:  0.9511, Adjusted R-squared:  0.9436
## F-statistic: 126.5 on 6 and 39 DF,  p-value: < 0.00000000000000022
```

- Conclusion:
  - 1. both D and R: P Increase, EV Increase
  - 2. R-square: 0.95, b6 P Value: not low enough to be statistically significant - accept the null hypothesis
  - 3. Population, Charging Station, and Income are statistically significant. But the business implication for Income is not significant ($1 increase in income, 0.0012% increase in EV).

5.3 check the multicollinearity between IV

- Conclusion: IVs VIF < 10, they are not highly correlated and should not be excluded from the model

```
vif(l_dummy_m_9)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                Political_Party                 log(Gas_Price)
##                     179.226239                       3.453960
##                log(Population)               Household_Income
##                       4.962864                       2.165594
##         log(Charging_Stations) Political_Party:log(Gas_Price)
##                       6.687973                     188.540908
```

5.4 the relationship between each IV and DV, controlling other IVs

```
avPlots(l_dummy_m_9)
```