Data Mining Project

Kevin Luth

11/3/2021

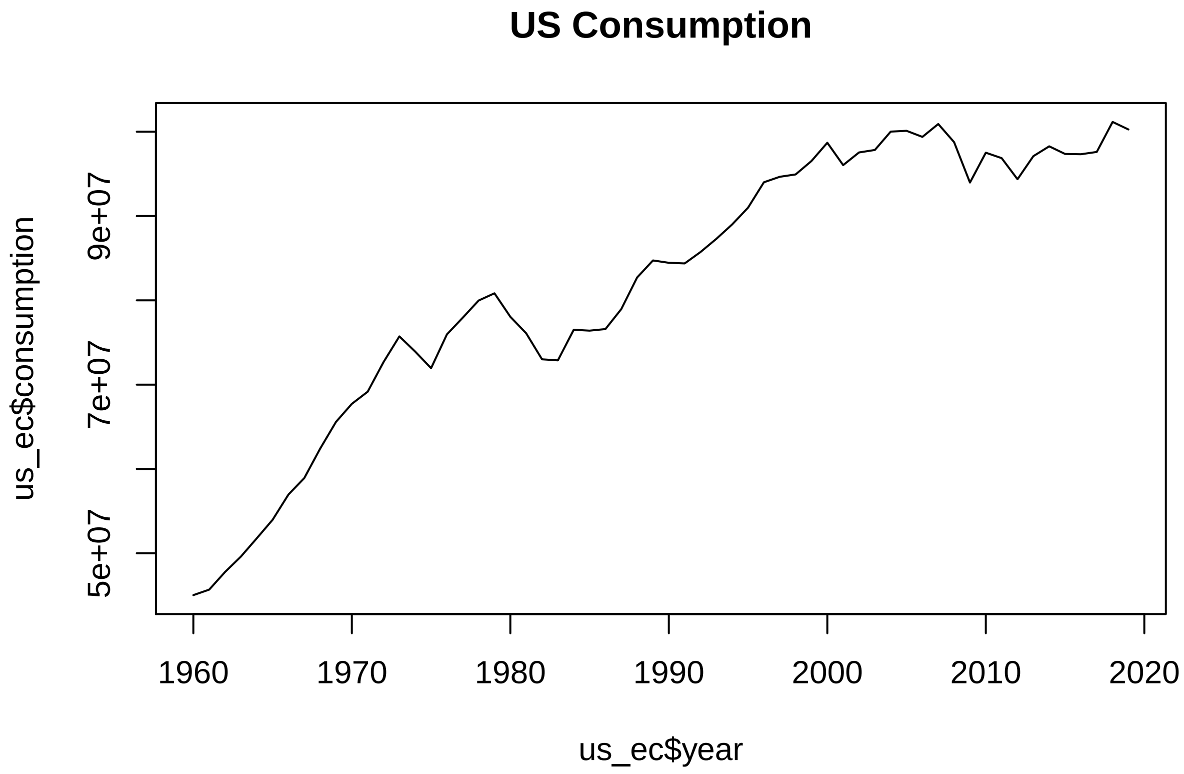
Initial Data Exploration

- We theorized that population is likely to have a strong positive relationship on a state’s consumption (can be seen in the line plot comparing population and consumption). To confirm this, we fit a linear regression model using these variables and confirmed that population does indeed strongly affect consumption in a positive manner. The R^2 value of the model is high at 0.78, meaning 78% of a state’s consumption variance can be explained by its population. This discovery means that in order to be able to uncover the true effects of other variables on consumption, we likely will need to normalize them for population.

```
#Consumption Data
energy_consump = read_xlsx("energy_consumption_by_state.xlsx", sheet = 2, skip = 2)
any(is.na(energy_consump)) #checks for missing values
```

```
## [1] FALSE
```

```
ec = energy_consump %>% pivot_longer(c(‘1960‘:‘2019‘), names_to = "year", values_to = "consumption") #t

#Isolates US totals
us_ec = energy_consump %>%
  pivot_longer(c(‘1960‘:‘2019‘), names_to = "year", values_to = "consumption") %>%
  filter(State == "US")

#Removes us from states data
ec = ec[!(ec$State == "US"),]

plot(us_ec$year, us_ec$consumption, type = "l", main = "US Consumption") #Shows overall consumption tre
```



```r
#Population Data
state_population = read_csv("total_population.csv")
```

```
## Warning: Missing column names filled in: 'X62' [62], 'X63' [63], 'X64' [64],
## 'X65' [65], 'X66' [66], 'X67' [67], 'X68' [68], 'X69' [69], 'X70' [70],
## 'X71' [71], 'X72' [72], 'X73' [73], 'X74' [74], 'X75' [75], 'X76' [76],
## 'X77' [77], 'X78' [78], 'X79' [79], 'X80' [80], 'X81' [81], 'X82' [82],
## 'X83' [83], 'X84' [84], 'X85' [85], 'X86' [86], 'X87' [87], 'X88' [88],
## 'X89' [89], 'X90' [90], 'X91' [91], 'X92' [92], 'X93' [93], 'X94' [94],
## 'X95' [95], 'X96' [96], 'X97' [97], 'X98' [98], 'X99' [99], 'X100' [100],
## 'X101' [101]
```

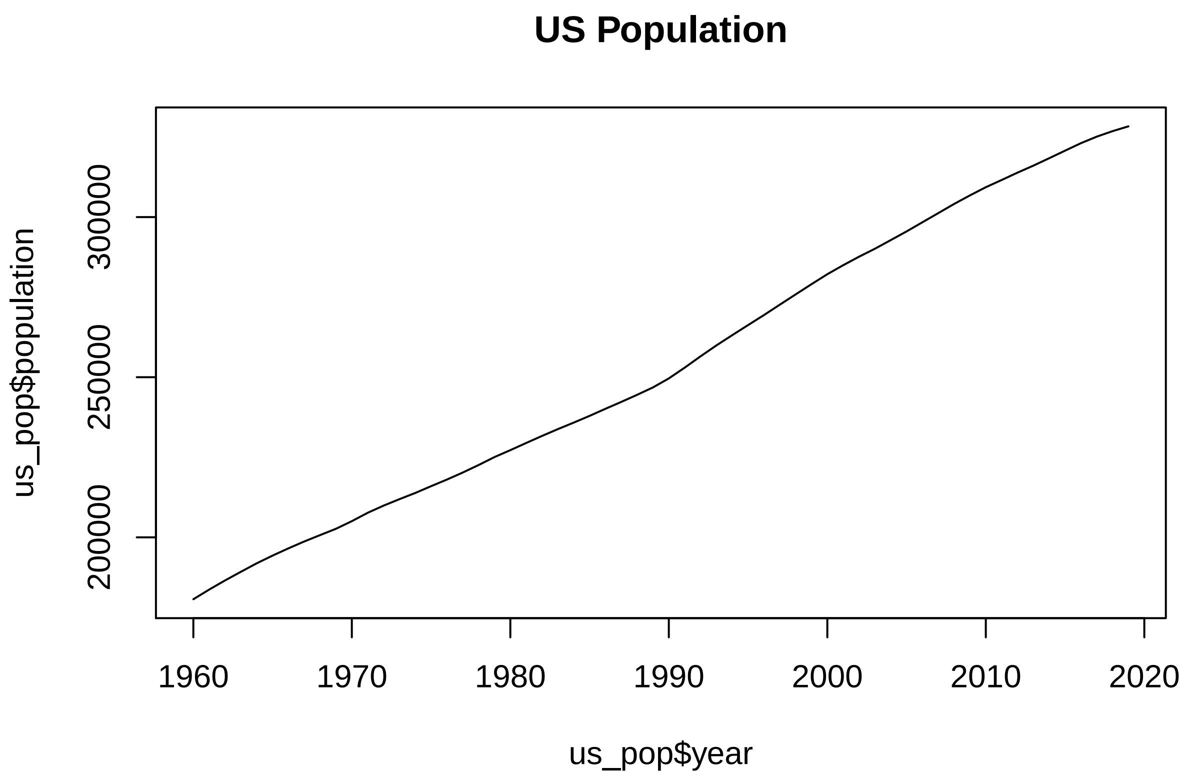
```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   State = col_character(),
##   X62 = col_logical(),
##   X63 = col_logical(),
##   X64 = col_logical(),
##   X65 = col_logical(),
##   X66 = col_logical(),
##   X67 = col_logical(),
##   X68 = col_logical(),
##   X69 = col_logical(),
##   X70 = col_logical(),
##   X71 = col_logical(),
##   X72 = col_logical(),
##   X73 = col_logical(),
##   X74 = col_logical(),
##   X75 = col_logical(),
##   X76 = col_logical(),
##   X77 = col_logical(),
##   X78 = col_logical(),
##   X79 = col_logical(),
##   X80 = col_logical()
##   # ... with 21 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
state_population = state_population %>% select(1:61)
any(is.na(state_population)) #checks for missing values
```

```
## [1] FALSE
```

```
spop = state_population %>% pivot_longer(c(‘1960‘:‘2019‘), names_to = "year", values_to = "population")

us_pop = state_population %>%
  pivot_longer(c(‘1960‘:‘2019‘), names_to = "year", values_to = "population") %>%
  filter(State == "US")
#Removes us from states data
spop = spop[!(spop$State == "US"),]

plot(us_pop$year, us_pop$population, type = "l", main = "US Population") #Shows overall population tren
```



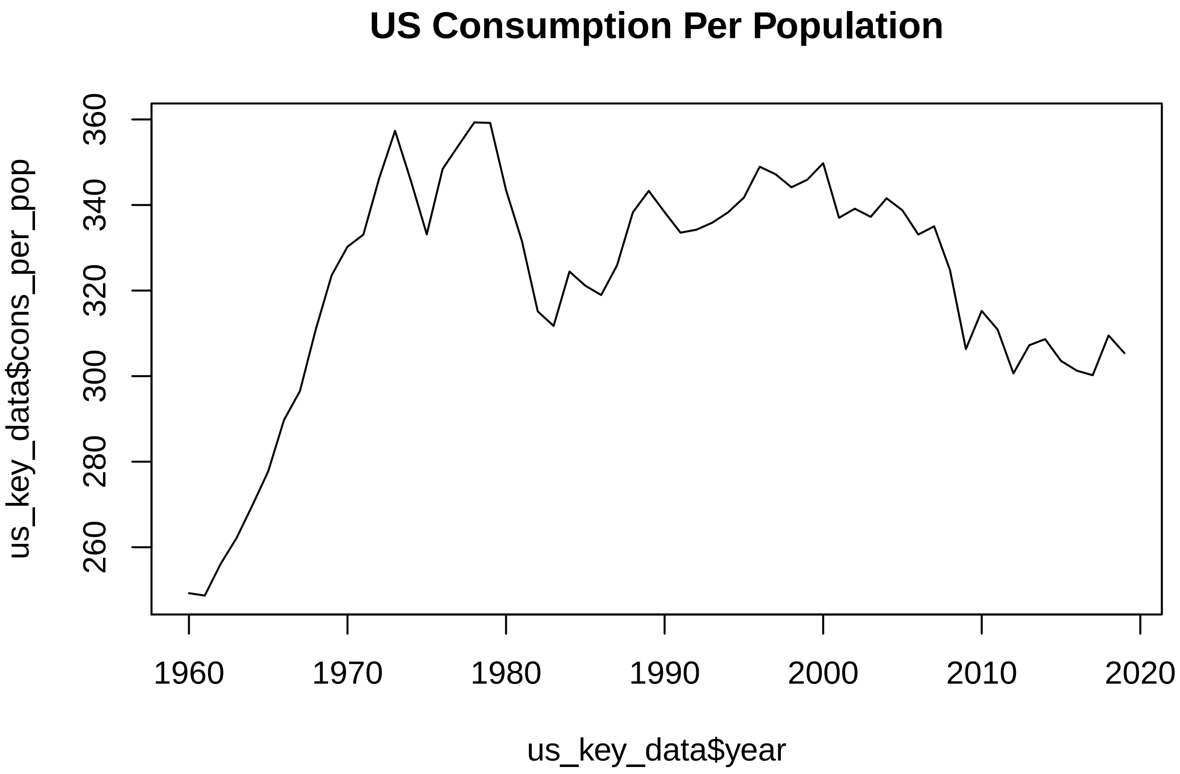
```
#Integrating Energy and population data
key_data = ec %>% inner_join(spop, by = c("State", "year"))
us_key_data = us_ec %>% inner_join(us_pop, by = c("State", "year"))
key_data = mutate(key_data, cons_per_pop = consumption / population)
us_key_data = mutate(us_key_data, cons_per_pop = consumption / population)

#Regress on population
#pop_fit = lm(consumption~population, key_data[key_data$State == "AK",]) #by indiv state
pop_fit = lm(consumption~population, key_data)
#pop_fit = lm(consumption~population, us_key_data)
#plot(us_key_data$population, us_key_data$consumption)
#abline(pop_fit)
summary(pop_fit)
```

```
##
## Call:
## lm(formula = consumption ~ population, data = key_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3665466  -242168  -111827   108462  6259548
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.032e+05  2.042e+04    9.95   <2e-16 ***
## population  2.827e+02  2.724e+00  103.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 845600 on 3058 degrees of freedom
## Multiple R-squared:  0.7789, Adjusted R-squared:  0.7788
## F-statistic: 1.077e+04 on 1 and 3058 DF,  p-value: < 2.2e-16
```

```
#plot(pop_fit) #diagnostic plots
```

- The following plot shows the trend of the population-adjusted energy consumption values over time. As can be seen, there is not a consistent, clear incline or decline overall in the values, but yet the short-term trends fluctuate a lot. This result tells us that there are other factors out there that affect energy consumption totals beside just population.

```
plot(us_key_data$year, us_key_data$cons_per_pop, type = "l", main = "US Consumption Per Population") #S
```



```
#Practice with ggplot
#ggplot(us_key_data) +
#  geom_line(mapping = aes(year, cons_per_pop, group = 1))
```

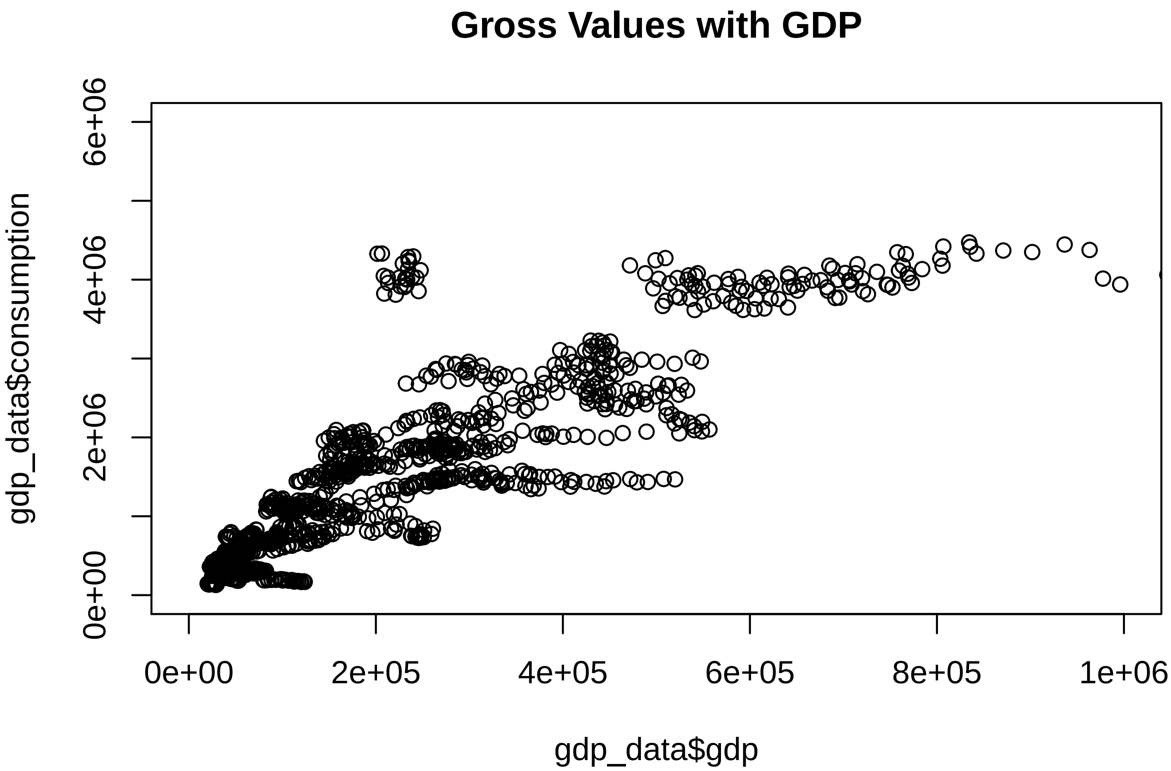
GDP Analysis

- According to the scatter plot of total consumption vs total gdp, there could be a positive relationship between the two. Performing a linear regression using these variables further supports this claim, as its R^2 value is relatively high at 0.7. However, upon viewing the model's diagnostic plots, it appears as though it does not satisfy all the proper assumptions for a good linear model. A possible explanation for this behavior is that by taking the total values of both variables, we are allowing the effect that population has on both to appear in our model. To account for this confounding variable, we normalize both consumption and gdp and then plot another scatter plot and fit a linear regression using these new variable values. Now it is evident in both the plot and the regression model that there is not much of a relationship between gdp and consumption. The plot shows no visual trend and the model has a very low R^2 value (0.002), indicating that a state's gdp per capita does not explain hardly any of the variance in its consumption per population.

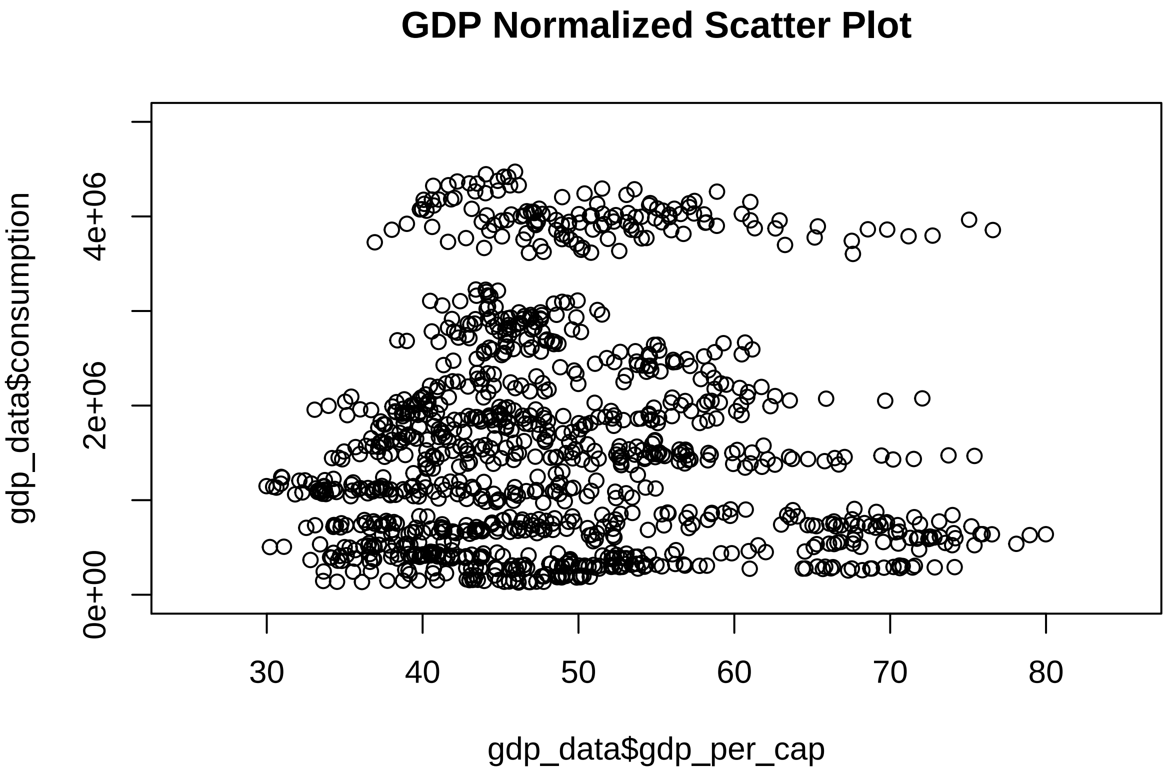
```
gdp = read_xlsx("energy_consumption_by_state.xlsx", sheet = 3, skip = 2)
gdp = gdp %>% select(-c(2:38))
any(is.na(gdp)) #checks for missing values
```

```
## [1] FALSE
```

```
#Isolates US totals
us_gdp = gdp %>%
  pivot_longer(c(`1997`:`2019`), names_to = "year", values_to = "gdp") %>%
  filter(State == "US")

#Removes us from states data
gdp = gdp[!(gdp$State == "US"),]

gdp = gdp %>% pivot_longer(c(`1997`:`2019`), names_to = "year", values_to = "gdp") #turns year headers

#Merge GDP and key_data
gdp_data = gdp %>% inner_join(key_data, by = c("State", "year"))
gdp_data = mutate(gdp_data, gdp_per_cap = gdp / population)

us_gdp_data = us_gdp %>% inner_join(us_key_data, by = c("State", "year"))
us_gdp_data = mutate(us_gdp_data, gdp_per_cap = gdp / population)

#GDP Regression
plot(gdp_data$gdp, gdp_data$consumption, xlim = c(0, 1000000), ylim = c(0, 6.0e+06), main = "Gross Valu
```



```r
#plot(gdp_data$gdp_per_cap, gdp_data$cons_per_pop, xlim = c(30, 85), main = "Normalized Scatter Plot wi
plot(gdp_data$gdp_per_cap, gdp_data$consumption, xlim = c(25, 85), ylim = c(0, 5e+06), main = "GDP Norma
```



```
gdp_fit = lm(consumption~gdp, gdp_data) #non-normalized fit
summary(gdp_fit)$r.squared #r-sq for non-normalized
```

```
## [1] 0.7018729
```

```
#gdp_fit = lm(cons_per_pop~gdp_per_cap, gdp_data) #normalized fit
#summary(gdp_fit)$r.squared #r-sq for normalized

gdp_fit = lm(consumption~gdp_per_cap, gdp_data) #gdp normalized fit
summary(gdp_fit)$r.squared #r-sq for normalized gdp only
```

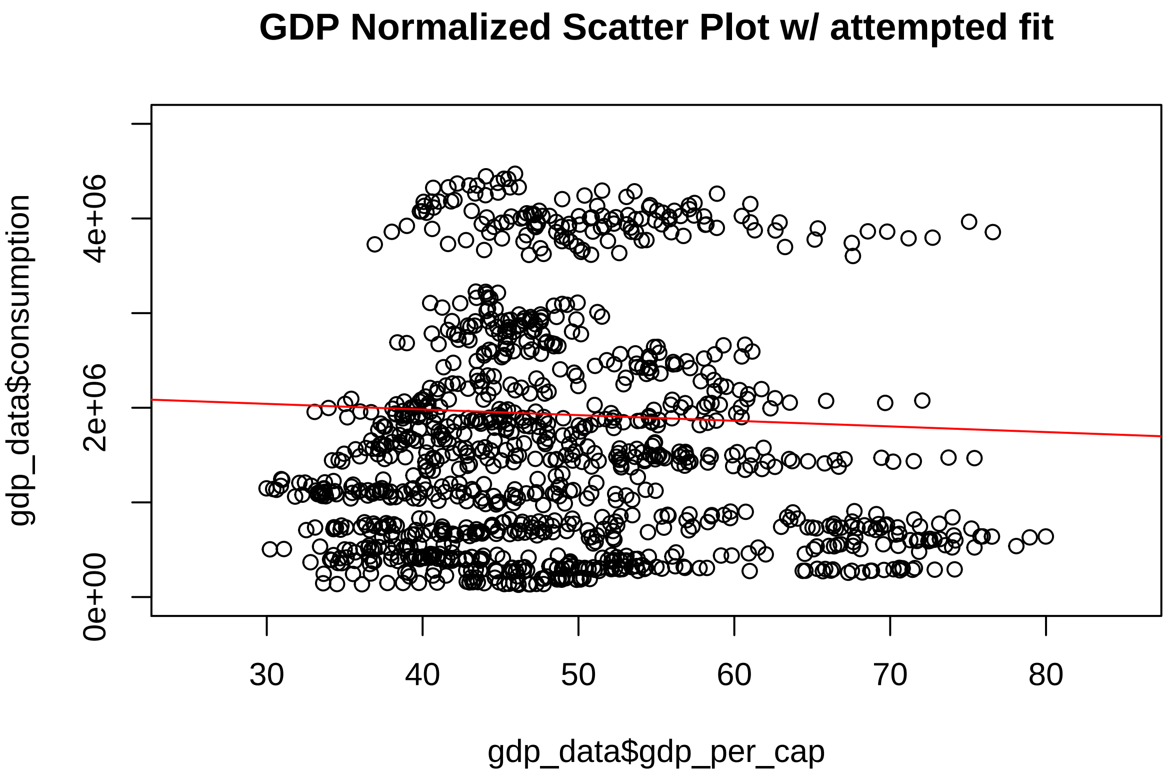
```
## [1] 0.003133728
```

```
summary(gdp_fit)
```

```
##
## Call:
## lm(formula = consumption ~ gdp_per_cap, data = gdp_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1876904 -1256977  -475369   539901 12370296
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2219844     169695  13.081   <2e-16 ***
## gdp_per_cap    -5959       3106  -1.919   0.0553 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2058000 on 1171 degrees of freedom
## Multiple R-squared:  0.003134,   Adjusted R-squared:  0.002282
## F-statistic: 3.681 on 1 and 1171 DF,  p-value: 0.05527
```

```
#plot(gdp_fit) #Diagnostic plots

plot(gdp_data$gdp_per_cap, gdp_data$consumption, xlim = c(25, 85), ylim = c(0, 5e+06), main = "GDP Norma
abline(gdp_fit, col = "red")
```


GDP Normalized Scatter Plot w/ attempted fit
gdp_data$consumption
0e+00
2e+06
4e+06
gdp_data$gdp_per_cap