Quantifying the Impacts of Building Retrofits in Energy Equity

Devan Addison-Turner

2022-05-27

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

```
#BIOS 238: Principles and Techniques for Data Visualization
#Instructor: Rebecca Gellman
#Spring 2022
#School of Medicine
#Stanford University

#California School (K-12) Building Retrofits in Energy Equity Project

#Research Question: What group consumes the most energy (kilowatt-hour)
#in public and secondary school (K-12) buildings across the entire state of California?

#The objective is to evaluate the annual energy consumed by school buildings in
#California compared to the energy consumed per pupil within the school buildings
#across 3 median income classes: high, low, and medium to identify any trends
#over a several year period (2013-2017).

#Note: The year 2015 is missing from this dataset.

#A pupil is described as a person or learner who is enrolled in an
#educational institution or school. It is also used to refer to someone
#who is under the direct supervision of a teacher because he is either a
#minor or has special needs. In most parts of the world, such as England and in
#Asia, the term "pupil" is used to refer to schoolchildren who are in the
```

```
#primary and elementary grades as well as those in secondary schools.

#Read more: Difference Between Pupil and Student #,http://www.differencebetween.net/language/words-lang
#-and-student/#ixzz7VW1ok8Gy

#Data: The data used was collected and aggregated from
#Pacific Gas & Electric Energy
#consumption (loads/rate of energy on kilowatts per hour),
#American Community Survey data 2015-2019 for the median income of (K-12)
#student populations, and Stanford Education Data Archive to
#compute a diversity index based on a formula from the U.S. Census Bureau
#according to race. Furthermore, I analyzed the average rate of energy consumed
#by school buildings and the average rate of energy consumed per pupil
#during a 4 year period. Two conditions were applied to test the
#variance of energy consumption using the diversity index and retrofit treatment.

#Data sample size (n = 2,306 observations)
#2,036 datapoints represent smart meters to measure energy peformance in school
#facilities (K-12) in California. There are 535 schools in this dataset.

#The first metric, income, is the median household income of the census tract
#in which the school is located. High, medium, and low-income subsets were
#created based on the California Department of Housing and Community Development
#definition of a low-income community having a median household income at or below
#80% of the statewide median household income. According to the ACS, for 2016-2020
#California's median household income was $78,672. Therefore, the cutoff
#for a school in a low-income community is a median household income at or below
#$62,937. The high-income community is defined as having a median income at or
#above 150% of the statewide median household income,or $118,008.

#The second metric we use is the Diversity Index of the student population.
#The Diversity Index is a measure (bounded between 0 and 1) of the probability
#that two people randomly chosen from the population will belong to
#different race and ethnicity groups. The US Census defines the Diversity Index (DI)
#as: DI = 1 - (H^2 + W^2 + B^2 + AIAN^2 + Asian^2), Where H is the proportion
#of the population who are Hispanic or Latino, W is the proportion of the
#population who are White alone, B is the proportion of the population who are
#Black or African American alone, AIAN is the proportion of the population
#who are American Indian or Alaska Native alone, and Asian is the proportion
#of the population who are Asian alone. The DI of each school was calculated,
#then calculated the quintiles of the DI of our population to set cutoffs
#for high- and low-diversity schools. High-diversity schools are in the
#top quintile (DI is greater than or equal to 0.58) and low-diversity schools
#are in the bottom quintile
#(DI is less than or equal to0.21). The diversity index is from 0 to 1,
#with 0 being everyone is the
#same race and 1 being everyone is a different race.

#The third metric is retrofit impact on energy consumption, 0 as being a
#non-retrofit school and 1 being a school that received retrofits treatment.
#In theory, retrofits should reduce energy consumption, which is validated in the graph.

#Now is the development of my linear regression model to perform the analysis:
```

```
#Pro-Tips: Make sure to select the PDF format when creating a
#new RMarkdown file and install tinytex package to Knit file to PDF.
#Install tinytex package to Knit RMarkdown File to PDF.

#install.packages('tinytex')
#install.packages("tinytex", repos = "http://cran.us.r-project.org")
#tinytex::install_tinytex()

library(tinytex)
#set working directory and specify file path from my computer
setwd("~devanaddisonturner/Documents/GitHub/devanaddisonturner.github.io/BIOS 238_Princ. and Tech. for
#imports impacts_PGE csv. file from my working directory
impacts_PGE <- read.csv("impacts_PGE.csv")
```

```
#install packages
#install.packages('RMySQL', repos='http://cran.us.r-project.org')
```

```
#install.packages("tidyverse")
#install.packages("viridis")
#install.packages("units")
#if (!require("RColorBrewer")) {
#install.packages("RColorBrewer")
library(RColorBrewer)
#}
#loads packages
library (viridis)
```

```
## Loading required package: viridisLite
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(units)
```

```
## udunits database from /Library/Frameworks/R.framework/Versions/4.1/Resources/library/units/share/udu
```

```
#This provides the description of what is contained within each
#individual variable in the original dataset impacts_PGE, shows first  6 rows and all columns
data()
data(impacts_PGE)
```

```
## Warning in data(impacts_PGE): data set 'impacts_PGE' not found
```

```
head(impacts_PGE)
```

```
##   X char_prem_id wea_stn_cd year impact_baseline       adj pct_shed_baseline
## 1 1   1026095425      LAUBT 2013        1.309408  35.69298          3.668530
## 2 2   1026095425      LAUBT 2013        1.309408  35.69298          3.668530
## 3 3   1032218427      LFATT 2013        1.701542  92.58118          1.837892
## 4 4   1070908192      LBFLT 2013       24.353049  99.00483         24.597838
## 5 5   1074845258      LMILT 2013        7.471712  48.46957         15.415263
## 6 6   1095638956      LCCRT 2013       10.804547 165.09669          6.544375
##   Annual_energy LONGITUDE LATITUDE      GEOID Median_income_households totenrl
## 1      149053.6 -121.1361 38.87205 6061020501                    86250     172
## 2      149053.6 -121.1361 38.87205 6061020501                    86250     221
## 3      253848.4 -119.7350 36.73783 6019002903                    23596     792
## 4      171393.2 -118.9614 35.34710 6029002302                    31900      NA
## 5      219842.1 -121.8011 37.30659 6085503336                   112554     639
## 6     1070635.2 -122.0968 37.90384 6013340002                   146838      NA
##      perfrl Diversity_Index kwh_person retrofit time
## 1 0.3604651       0.3104551   866.5907        0    0
## 2 0.1674208       0.1507383   674.4507        0    0
## 3 1.0000000       0.3734995   320.5157        0    0
## 4        NA              NA         NA        0    0
## 5 0.5555556       0.5599334   344.0409        0    0
## 6        NA              NA         NA        0    0
```

```
summary(impacts_PGE)
```

```
##        X           char_prem_id        wea_stn_cd             year
##  Min.   :   1.0   Min.   :3.084e+08   Length:2306        Min.   :2013
##  1st Qu.: 577.2   1st Qu.:3.044e+09   Class :character   1st Qu.:2014
##  Median :1153.5   Median :7.037e+09   Mode  :character   Median :2016
##  Mean   :1153.5   Mean   :6.169e+09                      Mean   :2015
##  3rd Qu.:1729.8   3rd Qu.:9.335e+09                      3rd Qu.:2017
##  Max.   :2306.0   Max.   :9.981e+09                      Max.   :2017
##
##  impact_baseline         adj           pct_shed_baseline  Annual_energy
##  Min.   :-87.1503   Min.   :  0.0065   Min.   :-79.2231   Min.   :      0
##  1st Qu.: -1.8540   1st Qu.: 19.7143   1st Qu.: -5.6708   1st Qu.: 111149
##  Median :  0.2981   Median : 49.0362   Median :  1.8100   Median : 252380
##  Mean   :  2.8029   Mean   : 83.7100   Mean   :  0.3507   Mean   : 447002
##  3rd Qu.:  5.1061   3rd Qu.: 95.5003   3rd Qu.:  8.7407   3rd Qu.: 473954
##  Max.   :162.4117   Max.   :933.6280   Max.   : 79.2724   Max.   :8526522
##  NA's   :101        NA's   :101        NA's   :101        NA's   :63
##    LONGITUDE         LATITUDE         GEOID           Median_income_households
##  Min.   :-123.7   Min.   :34.60   Min.   :6.001e+09   Min.   :     0
##  1st Qu.:-122.0   1st Qu.:36.74   1st Qu.:6.019e+09   1st Qu.: 43569
##  Median :-121.3   Median :37.33   Median :6.029e+09   Median : 65244
##  Mean   :-121.1   Mean   :37.42   Mean   :6.044e+09   Mean   : 73542
##  3rd Qu.:-119.8   3rd Qu.:38.08   3rd Qu.:6.082e+09   3rd Qu.: 94514
##  Max.   :-118.8   Max.   :40.64   Max.   :6.115e+09   Max.   :250001
##
##     totenrl           perfrl       Diversity_Index    kwh_person
```

```
##  Min.   :   7.0   Min.   :0.0097   Min.   :0.0000   Min.   :    0.06
##  1st Qu.: 416.0   1st Qu.:0.4156   1st Qu.:0.2687   1st Qu.:  231.01
##  Median : 561.0   Median :0.6914   Median :0.4781   Median :  474.15
##  Mean   : 613.3   Mean   :0.6258   Mean   :0.4136   Mean   :  790.27
##  3rd Qu.: 733.0   3rd Qu.:0.8817   3rd Qu.:0.5648   3rd Qu.:  762.75
##  Max.   :3096.0   Max.   :1.0000   Max.   :0.7427   Max.   :92629.42
##  NA's   :487      NA's   :487      NA's   :487      NA's   :539
##     retrofit          time
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000
##  Mean   :0.2936   Mean   :0.5009
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
##
```

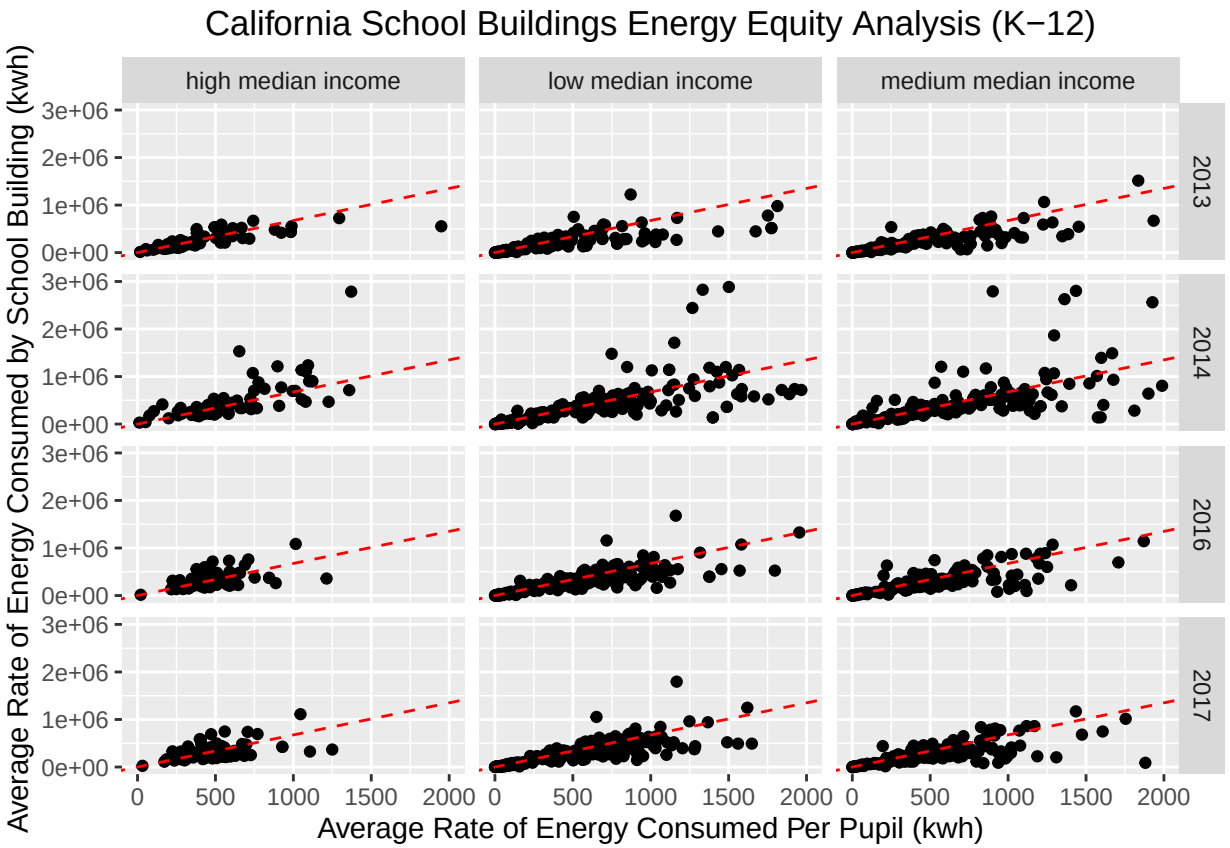
```
#Creates a new dataframe impacts_PGE from the original file called impacts_2
impacts_2 <- impacts_PGE
```

```
#Used high median income and low median income from the impacts 2 dataframe as
#references to create new columns with 3 subsets
high_income <- subset(impacts_2, Median_income_households >= 109773) %>%
  mutate(income = "high median income")
medium_income<- subset(impacts_2, Median_income_households > 58545 & Median_income_households < 109773)
  mutate(income = "medium median income")
low_income <- subset(impacts_2, Median_income_households <= 58545) %>%
  mutate(income = "low median income")
#created a new column (variable) for median income subset using the low median
#income threshold there is not a threshold for median income from original data,
#however, I am grouping every census tract  based on what is in
#between the median low median income households and high median income households
#created a new dataframe called impacts_3, using the rbind function.The rbind function
#combines the 3 new columns according to 3 median income profiles into a new dataframe
#called impacts 3.
impacts_3 <- rbind(high_income, medium_income, low_income)
```

```
#Metric 1 - Median Income

#American Community Survey median household income for 8,035 census tracts in California.
#High, medium, and low-income subsets were created based on the
#California Department of Housing and Community Development.
#A low-income community is defined as median household income at or below 80%
#of the statewide median household income.
#For 2016-2020 California's median household income was $78,672,
#and for a school in a low-income community is at or below $62,937.
#The high-income community is defined as having a median income at
#or above 15% of the statewide median household income or $118,008.
#The medium-income community is undefined, however, I created a third subset
#to group the households whose median income is between $118,008 and $62,937.


ggplot(impacts_3, aes(x=kwh_person, y=Annual_energy)) +
#geom specifies the type of graph points to be displayed to compares the
```

```
#annual energy consumed by student compared to the annual total energy
#consumed by the school buildings
  geom_point() + geom_abline(aes(intercept = 0, slope = 675), color = "red", linetype =2) +
  facet_grid(year ~income) +
  xlim(0, 2000) + ylim(0, 3e+06) +
#facet grid separates plots groups into separate bins on individual grids
#based on the annual energy consumed by each pupil starting from the
#year 2013 to 2017 period across the 3 median income profiles
#high, low, medium income.
#The limits for the range in both x and y directions are
#adjusted to zoom in on the graphs. I created a red, dashed trend line to
#display on the graph.
labs(x= "Average Rate of Energy Consumed Per Pupil (kwh)",
       y = "Average Rate of Energy Consumed by School Building (kwh)",
        title = "California School Buildings Energy Equity Analysis (K-12)") +
  theme(plot.title = element_text(hjust = 0.5)) # Center ggplot title
```

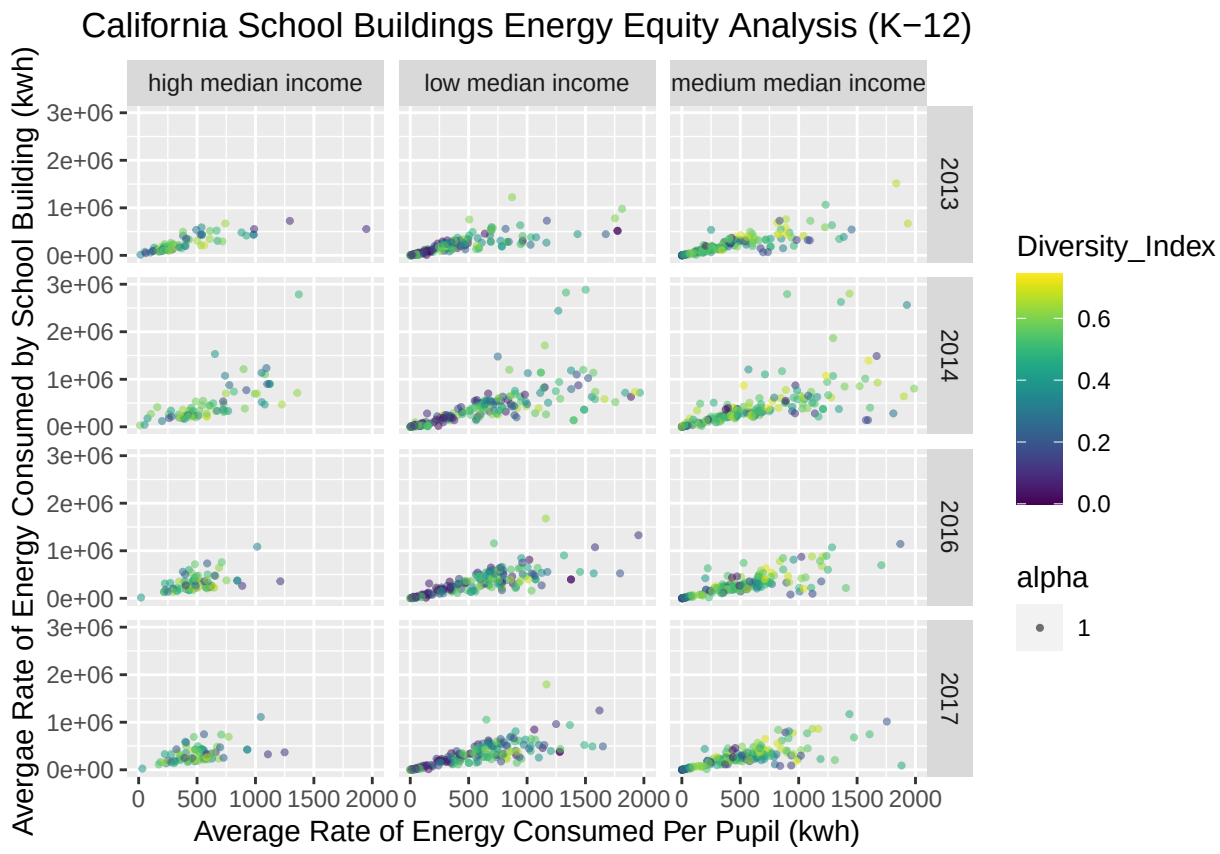
```
## Warning: Removed 613 rows containing missing values (geom_point).
```




```
#Results:Pupils in the lowest income classes consumed the most amount of energy,
#pupils in the medium income classes were ranked second for energy consumption,
#and the students in the high-income class consumed the least amount of energy.
#Schools from three median income classes consumed the most energy in the year 2014,
#although, consumed the least in 2013. After 2014, schools reduced '
#energy consumption through 2017.
```

```r
#Metric 2. Diversity Index Metric

#DI = measure (bounded between 0 and 1) of the probability that two people
#randomly chosen from the population will belong to
#different race and ethnicity groups. The US Census defines the Diversity Index (DI) as:
#DI = 1 - (H² + W² + B² + AIAN² + Asian²)
#H = proportion of the population who are Hispanic or Latino,
#W = proportion of the population who are White alone,
#B = proportion of the population who are Black or African American alone,
#AIAN = proportion of the population who are American Indian or Alaska Native alone,
#Asian = proportion of the population who are Asian alone
#We calculated the DI of each school, then calculated the quintiles
#of the DI of our population to set cutoffs for high and low-diversity schools.
#High-diversity schools are in the top quintile (DI  = 0.58)
#and low-diversity schools are in the bottom quintile (DI  = 0.21).

#Create 2-D scatterplots with color using ggplot function from impacts_3 dataframe
ggplot(impacts_3, aes(x=kwh_person, y=Annual_energy, color =  Diversity_Index, alpha = 1)) +
#geom specifies the type of graph points to be displayed
  geom_point(size = 0.75) + facet_grid(year ~income) + xlim(0, 2000) + ylim(0, 3e+06) + scale_color_viri
labs(x = "Average Rate of Energy Consumed Per Pupil (kwh)",
       y = "Avergae Rate of Energy Consumed by School Building (kwh)",
       title = "California School Buildings Energy Equity Analysis (K-12)") +
  theme(plot.title = element_text(hjust = 0.5))  # Center ggplot title
```

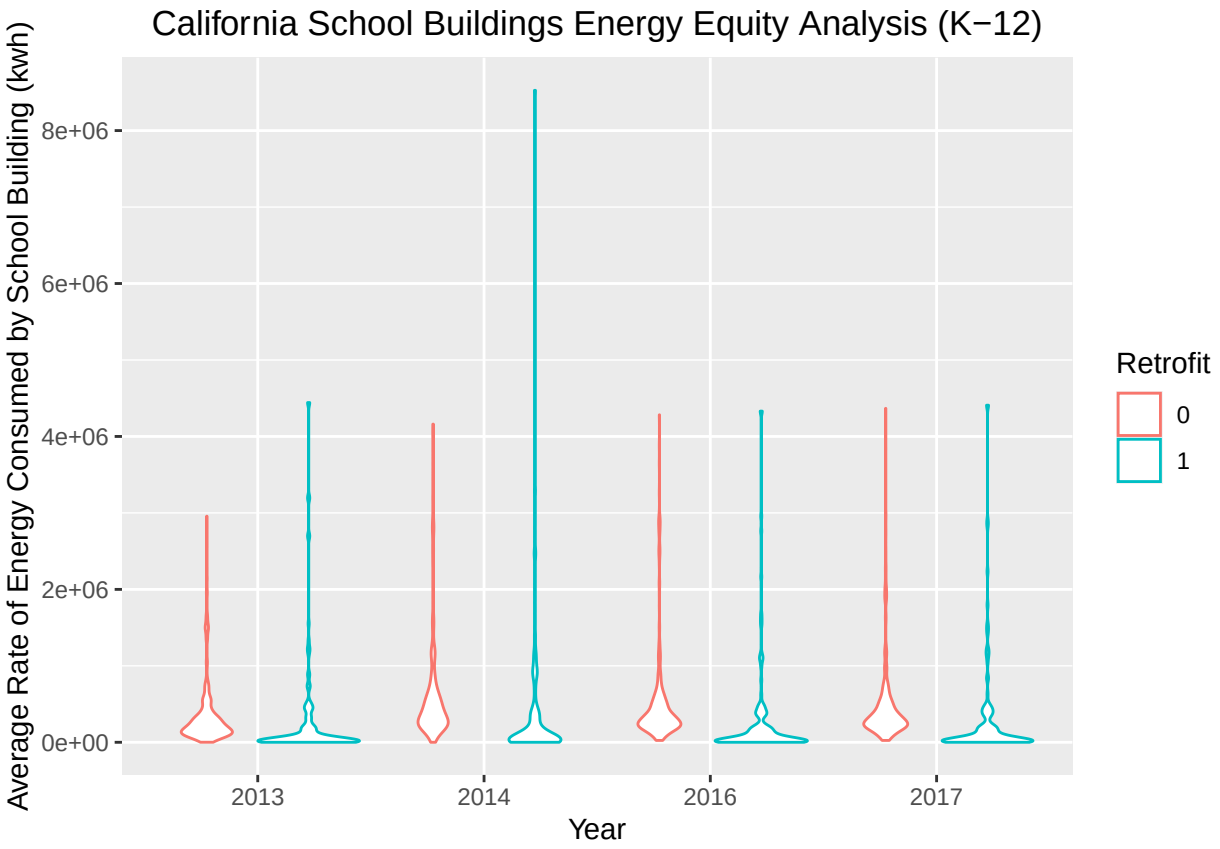
```
## Warning: Removed 613 rows containing missing values (geom_point).
```




```
#facet grid separates plots groups into separate categories on
#individual grids for based on the annual energy consumed by
#each pupil starting from the year 2013 to 2017 period across the
#3 median income profiles high, low, medium income.
#There is another factor used, Diversity Index to test variance.
#Applied the color palette viridis
#facet grid separates plots groups into separate bins on
#individual grids for based on the annual energy  consumed by
#each pupil starting from the year 2013 to 2017 period across the
#3 median income profiles high, low, medium income.
```

```
#Metric 3 - Retrofits Treatment

#Create 2-D line graphs with color using ggplot function from impacts_3 dataframe
ggplot(impacts_3, aes(x = as.factor(year), y = Annual_energy, color = as.factor(retrofit))) +
geom_violin() + # keep year as numeric
#geom specifies the type of graph points to be displayed
guides(color = guide_legend(title = "Retrofit")) +
labs(x = "Year", #Manually set labels
       y = "Average Rate of Energy Consumed by School Building (kwh)",
        title = "California School Buildings Energy Equity Analysis (K-12)") +
  theme(plot.title = element_text(hjust = 0.5))  # Center ggplot title
```

```
## Warning: Removed 63 rows containing non-finite values (stat_ydensity).
```

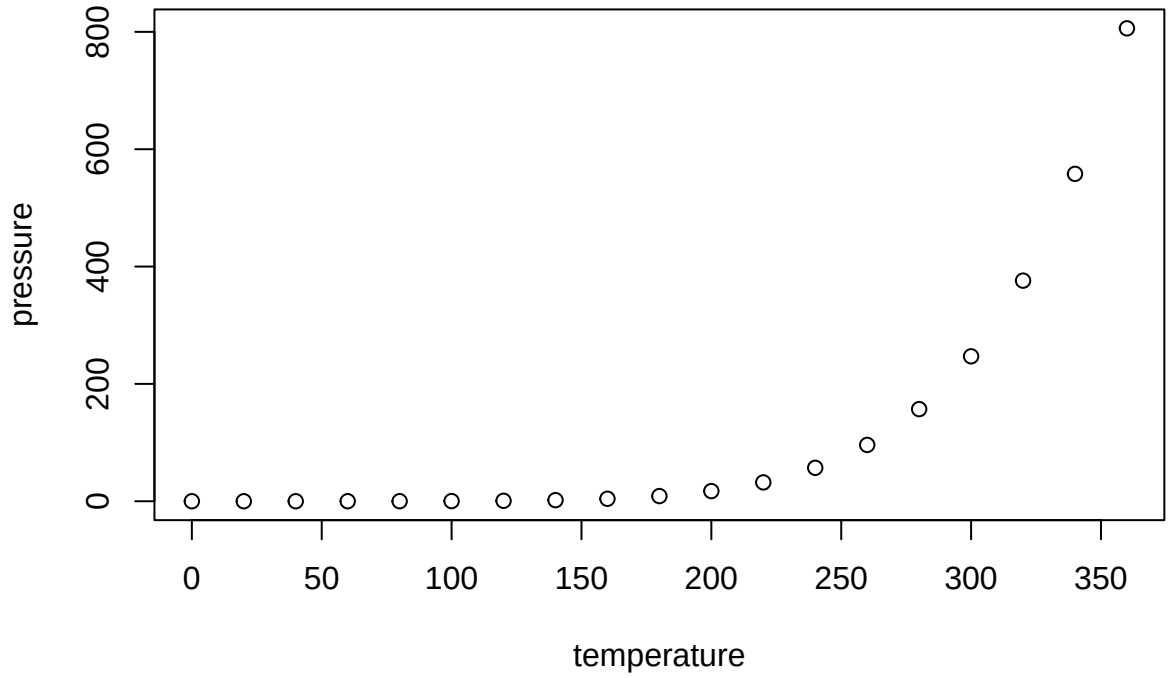



```
#Public and secondary schools in California consume less energy
#with pupils from higher median income classes each year
#from 2013 to 2017. However, schools consumed more energy
#with higher levels of diversity. Schools increased their
#diverse enrollment. There is a higher concentration
#of diverse students in both medium and low income classes
#who consumed the most energy. Across all three income subsets,
#the less diverse, higher income schools consumed the least amount of energy.

#Assumptions of Key Factors
#1. School districts are using different energy efficiency technology
#(e.g., heat pumps, HVAC, mixed ventilation, lighting).
#2. There is a Higher number of pupils enrolled and attending classes (in-person)
#by location, income level, and health factors, which impacts energy efficiency
#with less people occupying schools (physical campus).
#3. Year of construction (e.g., durability, structural integrity)
#4. Building materials (e.g., deterioration, thermal insulation)
#5. Size, sq footage (area), number of floors, space functions, equipment
#6. Geographic location of schools within school districts
#(e.g., land-use policies, climate zones)
#7. Proximity of schools near sources producing high pollution concentrations
#from buildings, transportation, and the natural environment
#8. Fuel source in schools could impact energy efficiency
#(e.g., oil & gas, renewable, electric)
```

```
#Results: In the first graph, the results show that pupils in the
#lowest median income groups consume the most amount of energy and the
#medium median income pupil group are ranked second as consuming the
#most energy by school buildings and per pupil that occupy the school buildings,
#while the high median income students consume the least amount of energy.
#All 3 median income classes consumed the most energy in the year 2014, although,
#consumed the least in 2013. After 2014, the average rate of energy consumed by
#both school buildings and students reduced over time. I assume that in 2015,
#the schools began to receive retrofits, decreasing the rate of energy consumed.

#In the second graph, my inference is that public and secondary schools in
#California consume less energy with students from higher median income classes
#each year from 2013 to 2017. However, schools consume more energy with higher
#levels of diversity. Based on the DI graph, more students of higher diversity
#consume more energy. There is a higher concentration of diverse students in
#both medium median and low median income classes. Across all three income subsets,
#the less diverse, higher income schools consumed the least amount of energy.

#Finally, the last graph is intended to compare the average energy consumption
#by school buildings annually and the impact of retrofits from 2013-2017.
#Also, the class setting (in-person, hybrid, or virtual) would impact energy performance
#in building. The shape sizes (length and width) are indicative of how much energy is
#consumed in the longitudinal direction, while the width is the number of schools
#within a cluster (proximity of individual schools distance to one another or
#the number of schools within each school district. In 2014, I speculate that
#more schools began to receive retrofit/upgrades during this time due to the
#this being the year of highest energy consumption during a several year period.
#As the average rate of energy consumed by buildings annually goes to down to 0 kwh,
#pupils were probably rezoned within school districts, so there may have been
#time periods where a number of schools were closed for renovations and construction,
#not using power. It is possible that schools were being shutdown permanently as well.

#Future Research: During what times of the day are students and buildings
#consuming the most energy? My hypothesis is that the energy efficient school buildings
#are using different technologies to reduce consumption of energy
#(e.g. heat pumps, (Mechanical and Natural), HVAC and lighting packages.
#Other considerations that may impact energy consumption are the year
#when the school buildings were constructed, building materials (insulation/thermal efficiency), geogra

#I intend to collect weather station data, test scores, pollution concentrations/sources including tran
#Finally, I will create maps to identify the location of the school districts most
#at risk based on the various factors aforementioned.
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.