Gurpinder Singh

STAT 108

11/9/2022

The research questions is to infer if there is a connection between quality of school and home prices inside of California for the year 2021.

Load all the followin library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(knitr)
library(skimr)
library(broom)
library(readr)
```

The following data tries to measure the quality of school. More specifically it takes into account the following variables: Absentness/Reason for absent, chronic Absentee, population, stability of student, suspension count, nation test results.

```
absentReason <- read.delim("data/schoolData/abreason2021.txt")
chronicAbsentee <- read.delim("data/schoolData/chrabs2021.txt")
cohort <- read.delim("data/schoolData/cohort2021.txt")
stabilityCount <- read.delim("data/schoolData/sr2021.txt")
suspeneded <- read.delim("data/schoolData/susp2021.txt")
test <- read.delim("data/schoolData/test/test.txt")
```

The response variable is housing price.

```
housing <- read_csv("data/housingData/housing.csv")
```

```
## Rows: 27424 Columns: 283
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr   (7): RegionName, RegionType, StateName, State, City, Metro, CountyName
## dbl (276): RegionID, SizeRank, 2000-01-31, 2000-02-29, 2000-03-31, 2000-04-3...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
absentReason <- absentReason %>%
  mutate(Average.Days.Absent=round(as.numeric(absentReason$Average.Days.Absent), digits = 0))
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
cohort <- cohort %>%
  mutate(Dropout..Rate.=round(as.numeric(cohort$Dropout..Rate.), digits = 0))
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
stabilityCount <-stabilityCount %>%
  mutate(Non.Stability.Rate..percent.=round(as.numeric(stabilityCount$Non.Stability.Rate..percent.), dig
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
suspeneded <-suspeneded %>%
  mutate(Suspension.Rate..Total.=as.numeric(suspeneded$Suspension.Rate..Total., digits = 0))
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

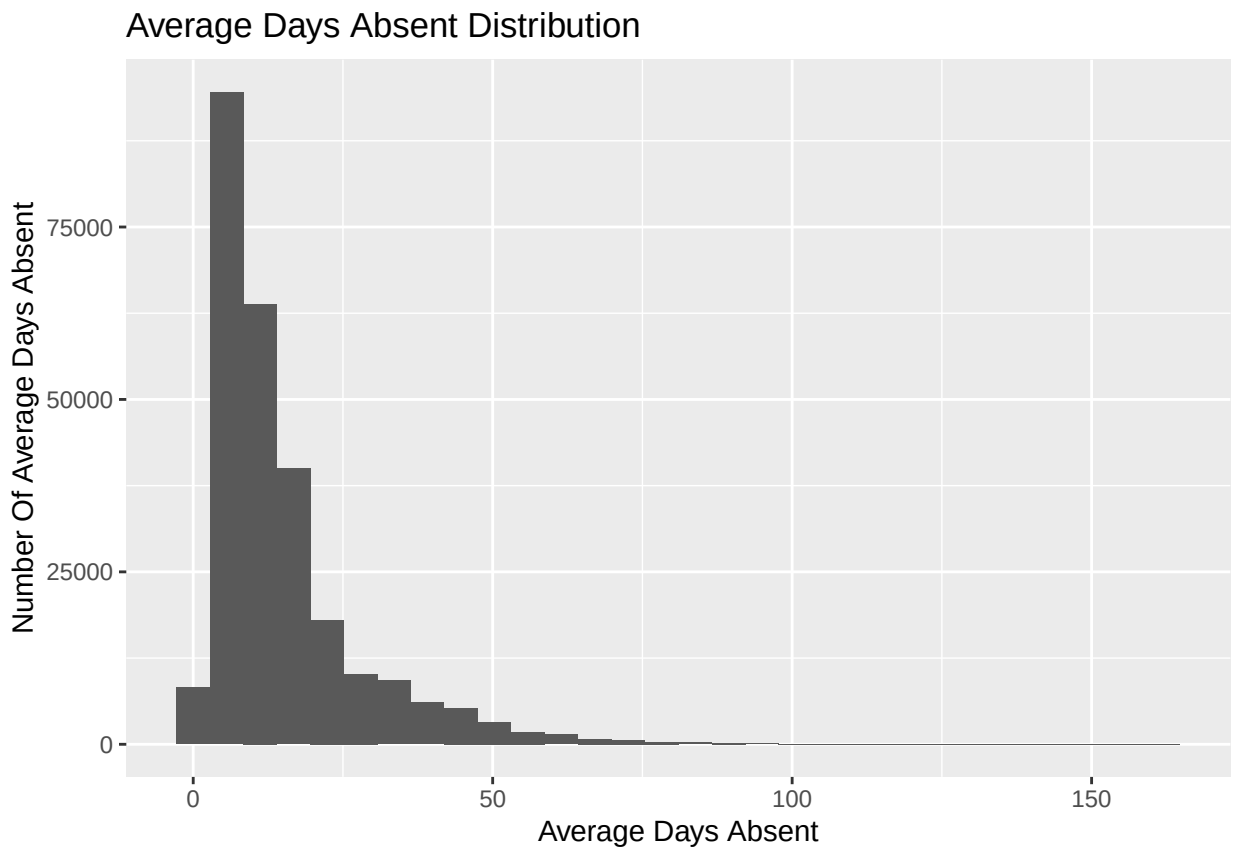
```
glimpse(test)
```

```
## Rows: 102,669
## Columns: 1
## $ County.Code.District.Code.School.Code.Filler.Test.Year.Student.Group.ID.Test.Type.Total.Tested.at.
```

```
ggplot(data = absentReason, aes(x =Average.Days.Absent)) +
  geom_histogram() +
  labs(x ="Average Days Absent",
       y = "Number Of Average Days Absent",
       title = "Average Days Absent Distribution")
```

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

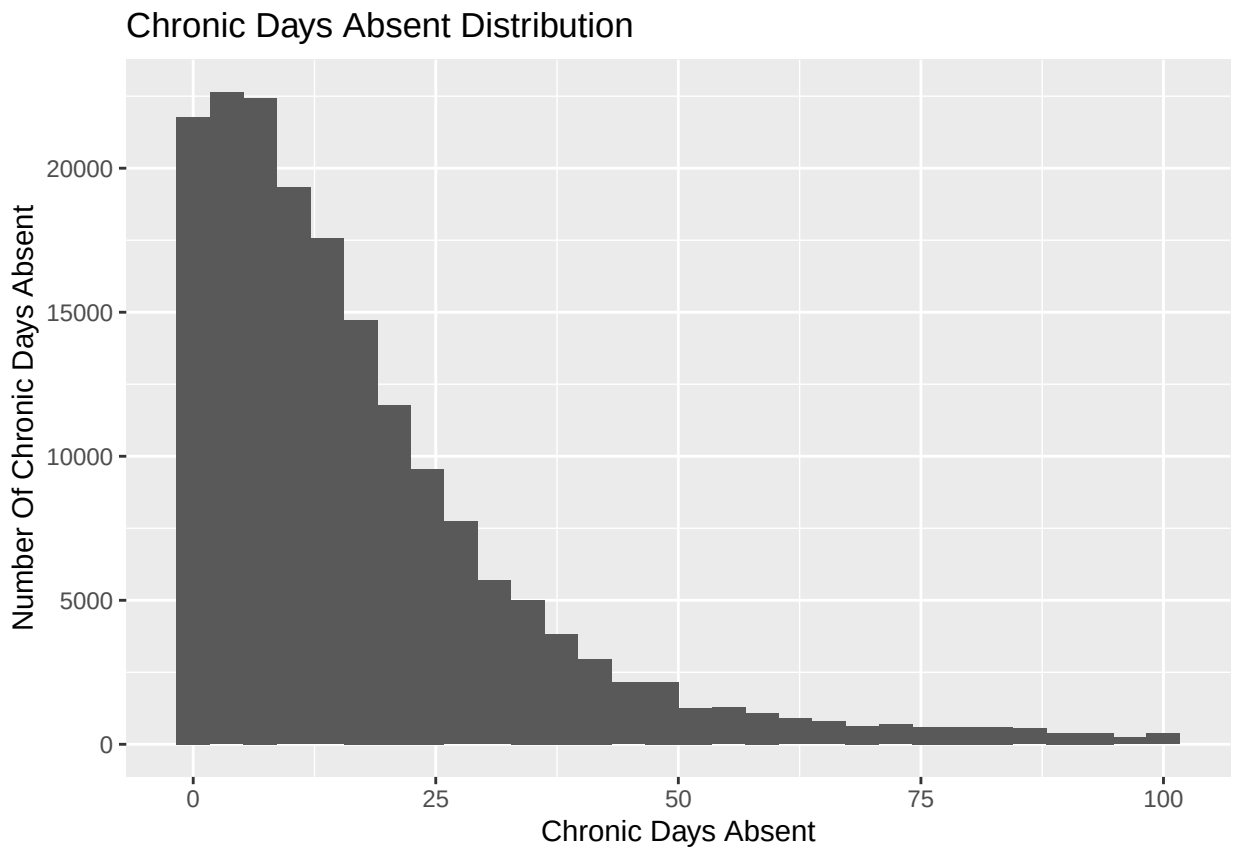
```
## Warning: Removed 107863 rows containing non-finite values (stat_bin).
```




```
ggplot(data = chronicAbsentee, aes(x =ChronicAbsenteeismRate)) +
  geom_histogram() +
  labs(x ="Chronic Days Absent",
       y = "Number Of Chronic Days Absent",
       title = "Chronic Days Absent Distribution")
```

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

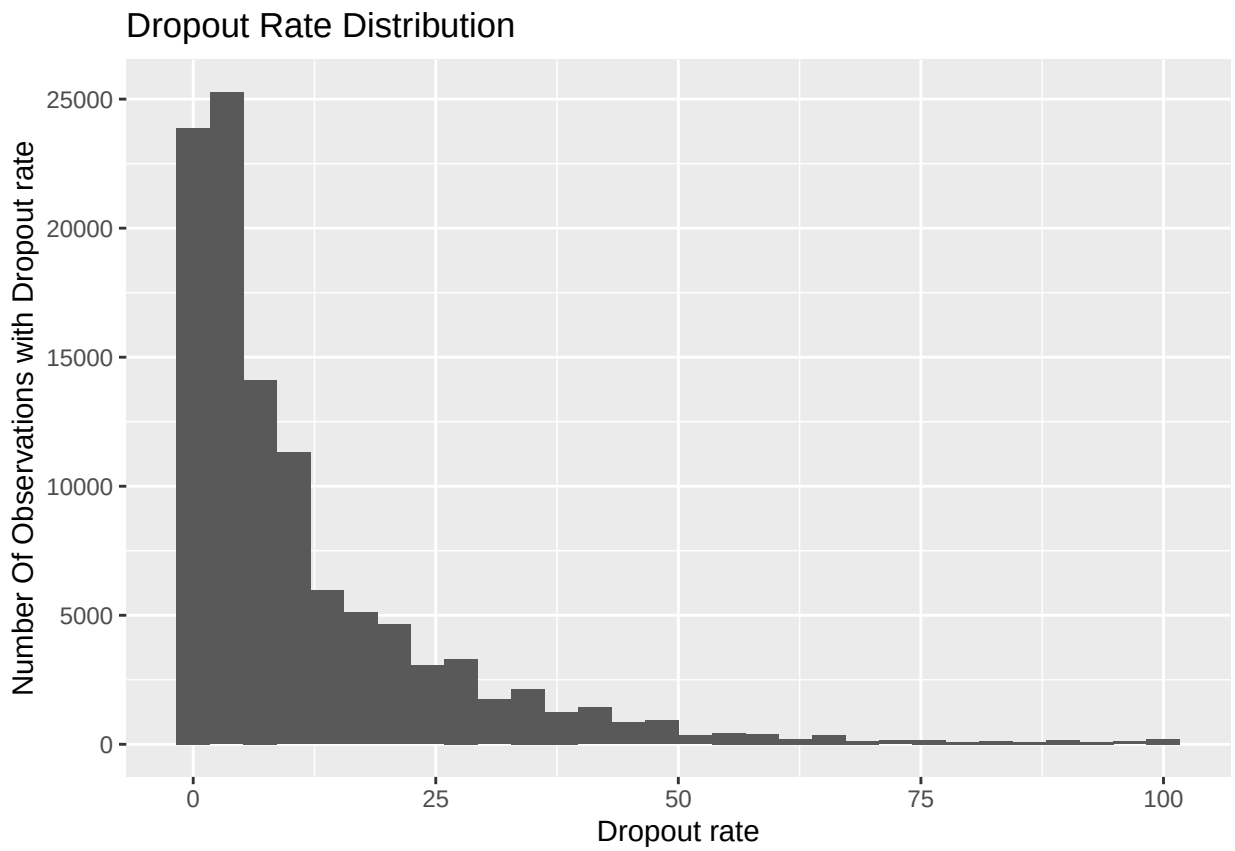
```
## Warning: Removed 83425 rows containing non-finite values (stat_bin).
```




```
ggplot(data = cohort, aes(x =Dropout..Rate.)) +
  geom_histogram() +
  labs(x ="Dropout rate",
       y = "Number Of Observations with Dropout rate",
       title = "Dropout Rate Distribution")
```

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

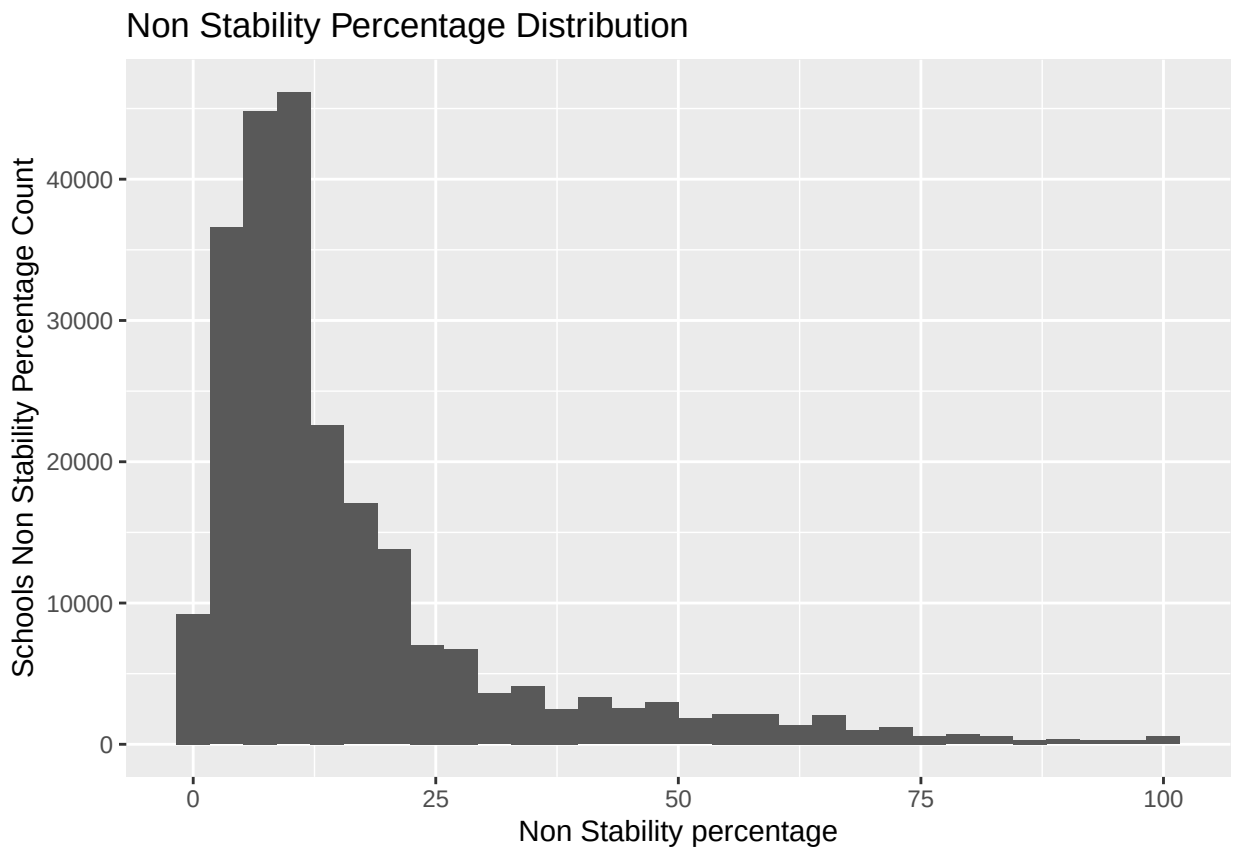
```
## Warning: Removed 146951 rows containing non-finite values (stat_bin).
```




```
ggplot(data = stabilityCount, aes(x =Non.Stability.Rate..percent.)) +
  geom_histogram() +
  labs(x ="Non Stability percentage",
       y = "Schools Non Stability Percentage Count",
       title = "Non Stability Percentage Distribution")
```

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

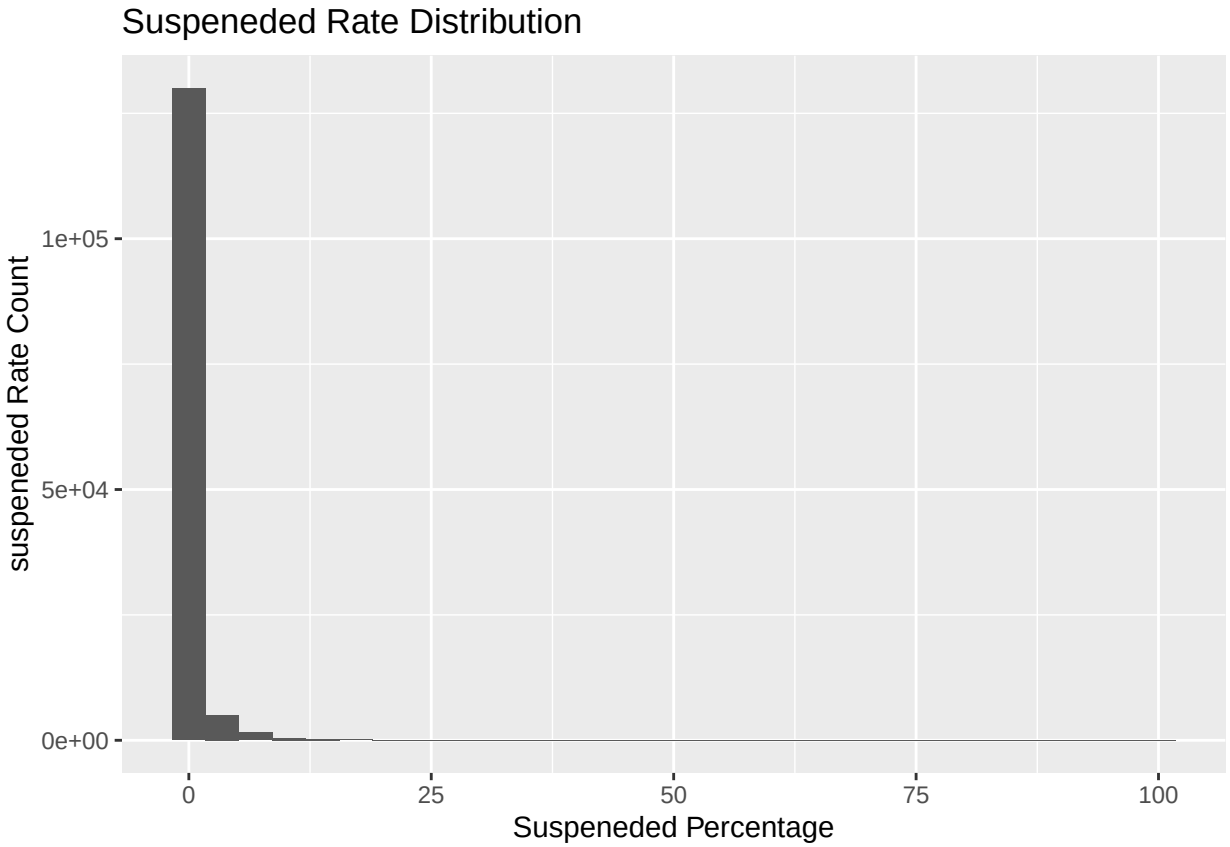
```
## Warning: Removed 112315 rows containing non-finite values (stat_bin).
```




```
ggplot(data = suspeneded, aes(x =Suspension.Rate..Total.)) +
  geom_histogram() +
  labs(x ="Suspeneded Percentage",
       y = "suspeneded Rate Count",
       title = "Suspeneded Rate Distribution")
```

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

```
## Warning: Removed 85218 rows containing non-finite values (stat_bin).
```


Suspeneded Rate Distribution
suspeneded Rate Count
1e+05
5e+04
0e+00
0
25
50
75
100
Suspeneded Percentage