```r
#DATA EXPLOR
library(tidyverse)
data <- readRDS('data_final.rds')

#see how many cities we are looking at
length(unique(data$cityid)

#we have 37 cities we are looking at here are all of our cities
unique(data$city)

#here is the number of weather stations we are looking at
length(unique(data$station))

#let's see what in general the weather tends to stay around and how frequently each block occurs
ggplot(data) + geom_histogram(mapping = aes(x = value), binwidth = 0.5)

#let's see how each city maps compared to eachother
ggplot(data, mapping = aes(x = value, colour = city)) + geom_freqpoly(binwidth = 0.1)

#ANOVA box plot to look at variability of temperature per city
ggboxplot(data, x = "city", y = "value", color = "city", ylab = "temperature", xlab = "City") +
 theme(axis.text.x = element_blank())
#looking at this it is clear then

#let's check to see that we have all consisteint coverage for each city (we want to see that every single one
 has the same amount of values)
vtree(data, "city")
#clearly, while we like what we see, this is not perfect since there are a few cities in which there were days
 that tempeture data was either not collected or not uploaded, these cities are: Yucca Valley, susanville,
 soledad, and red bluff

#now let's see for each city within that just how many stations they have, and how it is distributed.
vtree(data, c("city", "station"))

#this tree is so large that it is basically unreadable so let's look at the ones that peaked interest on an
 individual level...
#let's seperate eachh city and learn a bit about them
Yucca_Valley=data %>% filter(data$city =='Yucca Valley, CA US')
Soledad=data %>% filter(data$city =='Soledad, CA US')
Santa_Ana=data %>% filter(data$city =='Santa Ana, CA US')
Sacramento=data %>% filter(data$city =='Sacramento, CA US')
Oxnard=data %>% filter(data$city =='Oxnard, CA US')
Los_Angeles=data %>% filter(data$city =='Los Angeles, CA US')
Yuba_City=data %>% filter(data$city =='Yuba City, CA US')
Simi_Valley=data %>% filter(data$city =='Simi Valley, CA US')
San_Luis=data %>% filter(data$city =='San Luis Obispo, CA US')
Rosamond=data %>% filter(data$city =='Rosamond, CA US')
Oceanside=data %>% filter(data$city =='Oceanside, CA US')
Long_Beach=data %>% filter(data$city =='Long Beach, CA US')
Visalia=data %>% filter(data$city =='Visalia, CA US')
Santa_Rosa=data %>% filter(data$city =='Santa Rosa, CA US')
San_Jose=data %>% filter(data$city =='San Jose, CA US')
Riverside=data %>% filter(data$city =='Riverside, CA US')
Oakland=data %>% filter(data$city =='Oakland, CA US')
Vallejo=data %>% filter(data$city =='Vallejo, CA US')
Santa_Maria=data %>% filter(data$city =='Santa Maria, CA US')
San_Francisco=data %>% filter(data$city =='San Francisco, CA US')
Ridgecrest=data %>% filter(data$city =='Ridgecrest, CA US')
Napa=data %>% filter(data$city =='Napa, CA US')
Ukiah=data %>% filter(data$city =='Ukiah, CA US')
Santa_Cruz=data %>% filter(data$city =='Santa Cruz, CA US')
San_Diego=data %>% filter(data$city =='San Diego, CA US')
Redding=data %>% filter(data$city =='Redding, CA US')
Monterey=data %>% filter(data$city =='Monterey, CA US')
Susanville=data %>% filter(data$city =='Susanville, CA US')
```

```
Santa_Clarita=data %>% filter(data$city =='Santa Clarita, CA US')
San_Bernardino=data %>% filter(data$city =='San Bernardino, CA US')
Red_Bluff=data %>% filter(data$city =='Red Bluff, CA US')
Modesto=data %>% filter(data$city =='Modesto, CA US')
Stockton=data %>% filter(data$city =='Stockton, CA US')
Santa_Barbara=data %>% filter(data$city =='Santa Barbara, CA US')
Salinas=data %>% filter(data$city =='Salinas, CA US')
Palm_Springs=data %>% filter(data$city =='Palm Springs, CA US')
Merced=data %>% filter(data$city =='Merced, CA US')
#looked up count of unique stations each

#let's see what in general the weather tends to stay around and how frequently each block occurs
ggplot(data) + geom_histogram(mapping = aes(x = data$value), binwidth = 0.5)
```