data_exploration

Nikki Shintaku

4/15/2020

```
getwd()
```

```
## [1] "/Users/nikkishintaku/Desktop/Data_analytics_final_project/Data_Analytics_final_project"
```

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------ tidyv
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts --------------------------------------------------------------------------- tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggthemes)
library(knitr)

USGS_gage_height <- read.csv("./Data/Raw/USGS_tahoe_gage_height.csv")
NOAA_climate <- read.csv("./Data/Raw/NOAA_Tahoe_climate_data.csv")

#theme
mytheme <- theme_stata(base_size = 14, base_family = "sans", scheme = "s2mono") +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")

theme_set(mytheme)
```

```
#Changing Date on USGS Data
view(USGS_gage_height)
class(USGS_gage_height$datetime)
```

```
## [1] "factor"
```

```
USGS_gage_height$datetime <- as.Date(USGS_gage_height$datetime, format = "%m/%d/%y")

# We are formatting the data as year-2digit, month, day
USGS_gage_height$datetime <- format(USGS_gage_height$datetime, "%y%m%d")

#paste 19 if the input is greater than 181231 or 20 if it is less than
create.early.dates <- (function(d) {
       paste0(ifelse(d > 191231,"19","20"),d)
```

```
      })
#run the function on the USGS flow data for the datatime column
USGS_gage_height$datetime <- create.early.dates(USGS_gage_height$datetime)

#now reformat as a data in the format that we want
USGS_gage_height$datetime <- as.Date(USGS_gage_height$datetime, format = "%Y%m%d")

class(USGS_gage_height$datetime)
```

```
## [1] "Date"
```

```
summary(USGS_gage_height$gage_height) #3 NAs
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.260   3.950   6.350   5.866   7.700   9.400       3
```

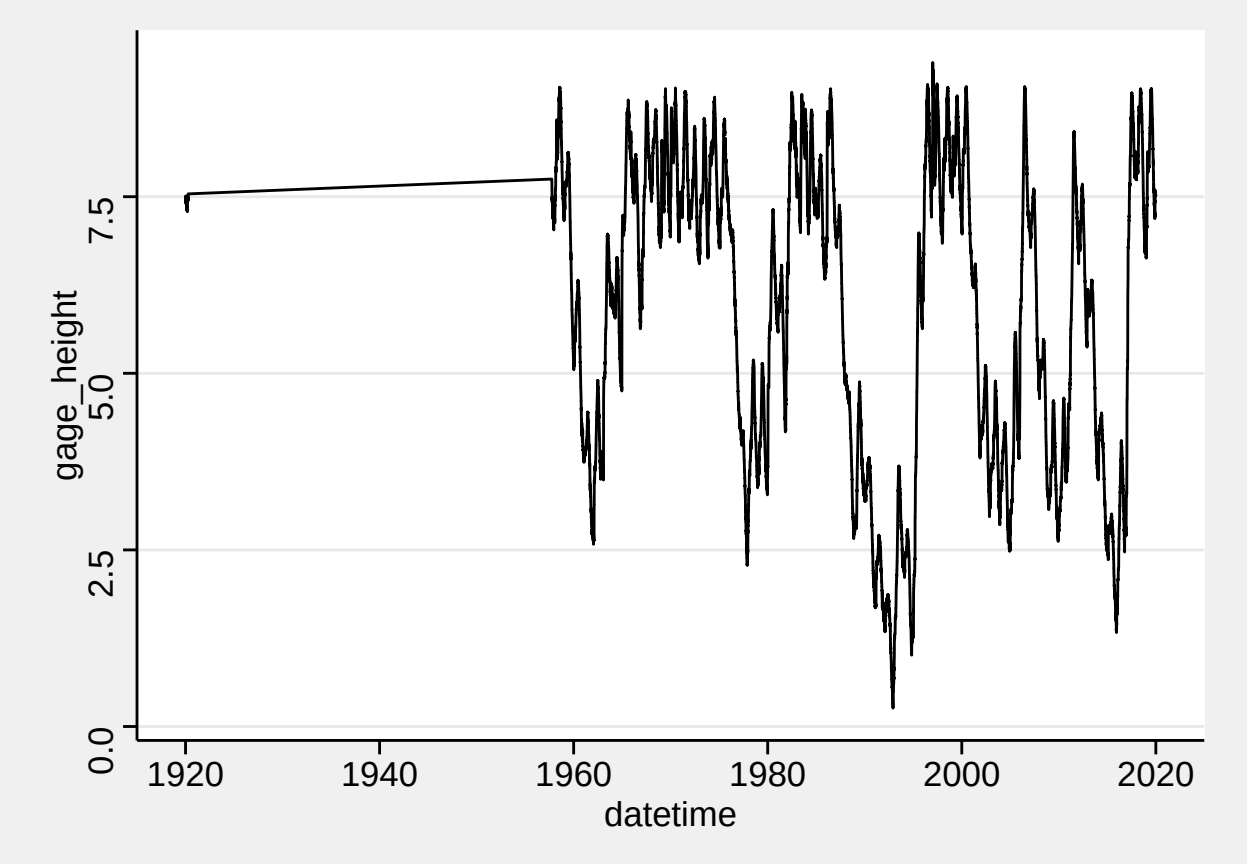
```
#changing date on NOAA Data
class(NOAA_climate$DATE)
```

```
## [1] "factor"
```

```
NOAA_climate$DATE <- as.Date(NOAA_climate$DATE, format = "%Y-%m-%d")
class(NOAA_climate$DATE)
```

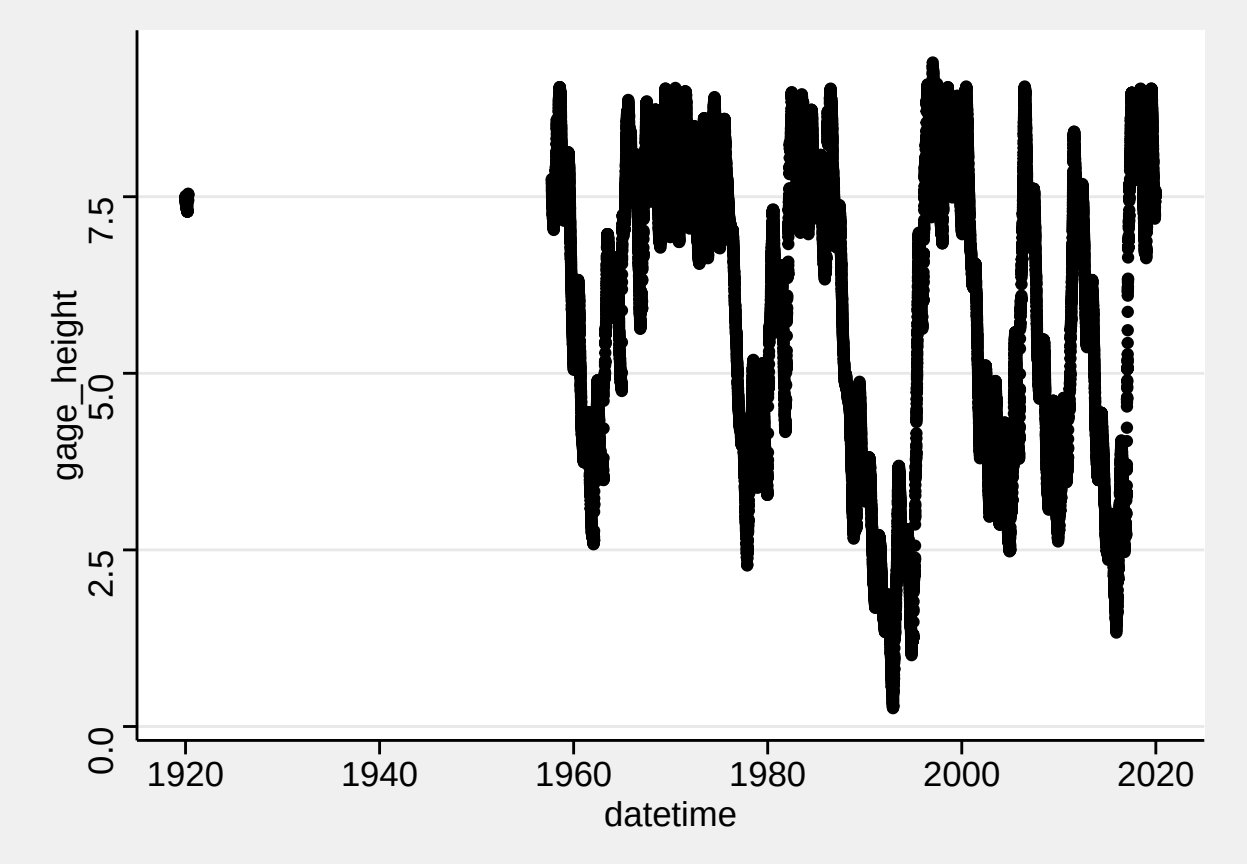
```
## [1] "Date"
```

```
#explore USGS data
ggplot(USGS_gage_height, aes(x = datetime, y = gage_height)) +
  geom_line()
```

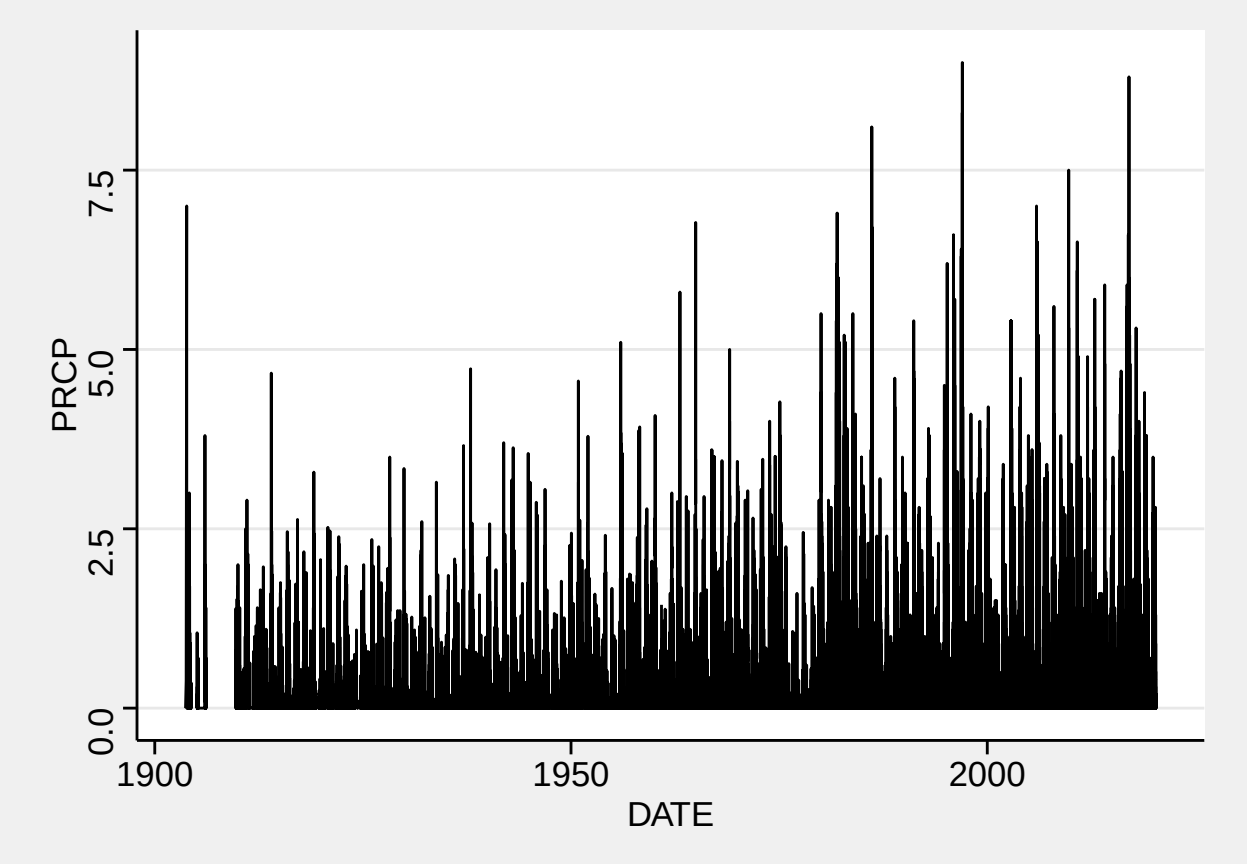



```
ggplot(USGS_gage_height) +
  geom_point(aes(x = datetime, y = gage_height))
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```



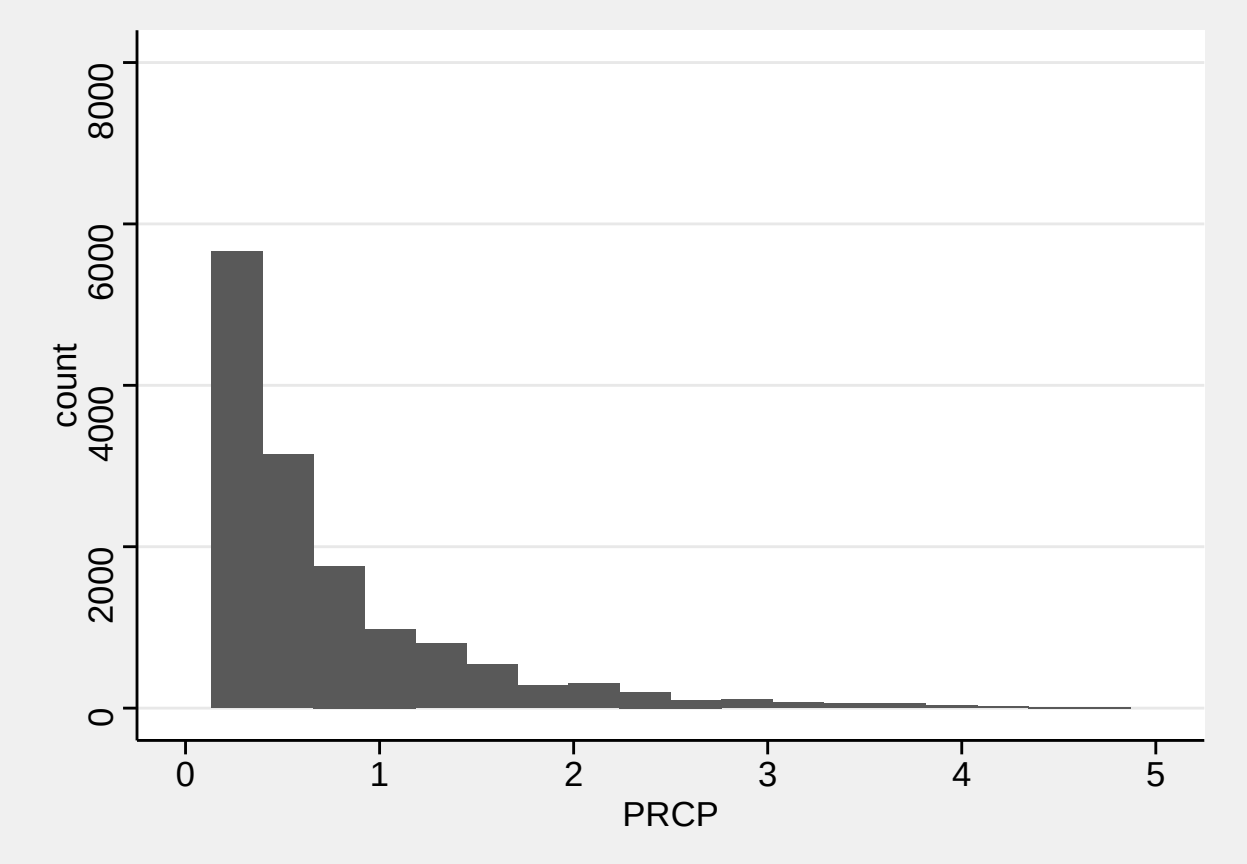

```
#explore NOAA data
ggplot(NOAA_climate, aes(x = DATE, y = PRCP)) +
  geom_line()
```




```
ggplot(NOAA_climate) +
  geom_histogram(aes(x = PRCP), bins = 20) +
  xlim(0,5) +
  ylim(0,8000)
```

```
## Warning: Removed 637 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

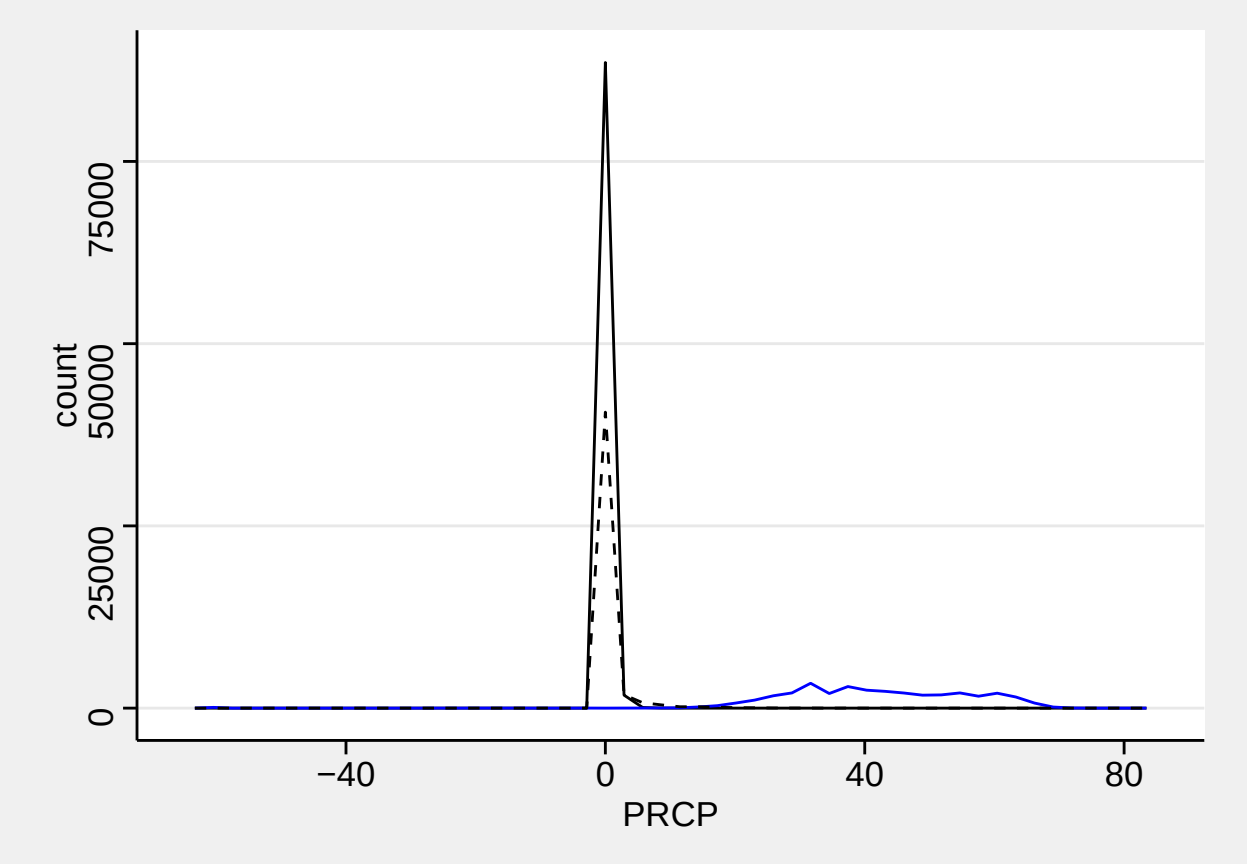



```
ggplot(NOAA_climate) +
  geom_freqpoly(aes(x = PRCP), bins = 50) +
  geom_freqpoly(aes(x = TAVG), bins = 50, color = "blue") +
  geom_freqpoly(aes(x = SNOW), bins = 50, lty = 2)
```

```
## Warning: Removed 582 rows containing non-finite values (stat_bin).
```

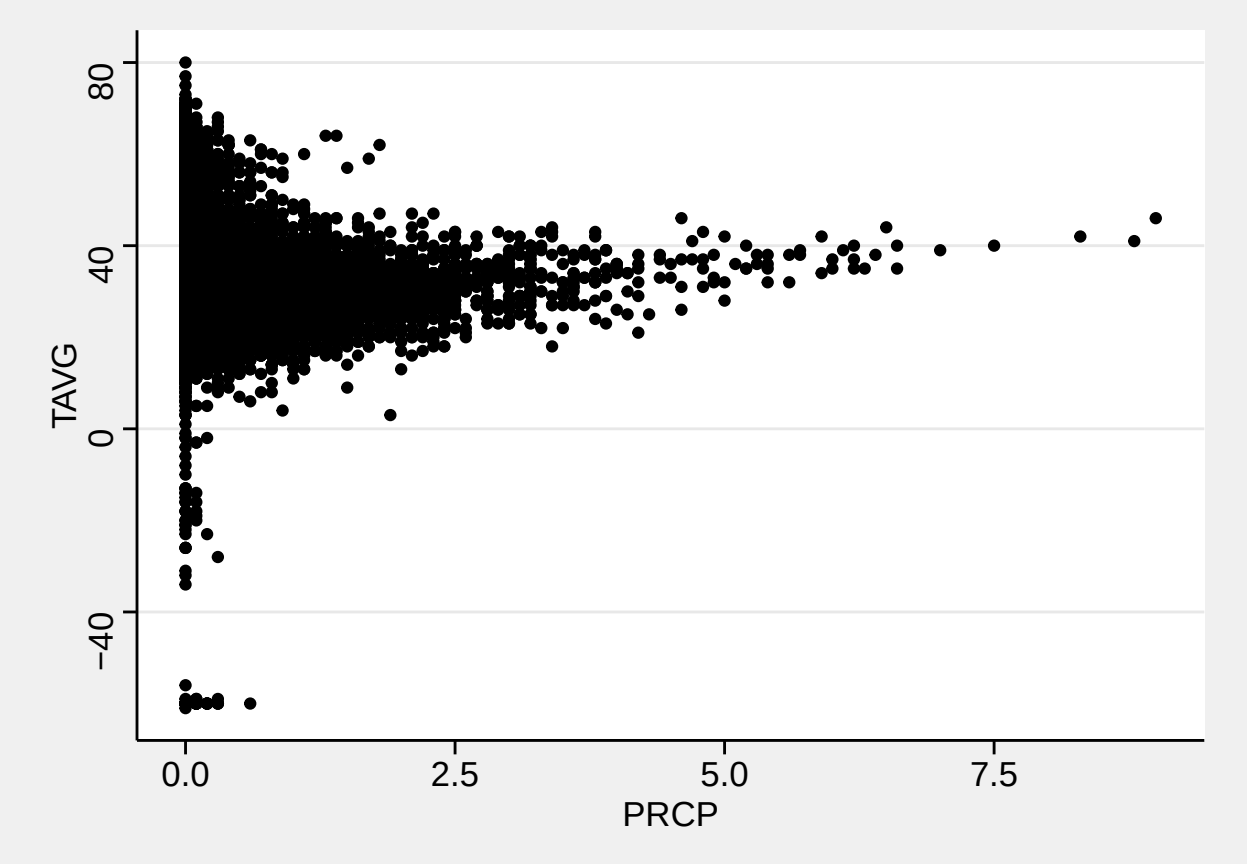
```
## Warning: Removed 57747 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 46752 rows containing non-finite values (stat_bin).
```




```
ggplot(NOAA_climate) +
  geom_point(aes(x = PRCP, y = TAVG))
```

```
## Warning: Removed 57749 rows containing missing values (geom_point).
```




```
summary(NOAA_climate$PRCP) #582 NAs
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.1255  0.0000  9.0000     582
```

```
summary(NOAA_climate$SNOW) #46752 NAs
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    0.00    0.00    0.56    0.00   49.00   46752
```

```
summary(NOAA_climate$TAVG) #57747 NAs
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -61.00   32.00   41.00   41.88   53.00   80.00   57747
```

```
summary(NOAA_climate$TMAX) #10883 NAs
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -61.00   43.00   54.00   55.43   69.00  452.00   10883
```

```
summary(NOAA_climate$SNWD) #34812 NAs
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    0.00    0.00   14.62   20.00  269.00   34812
```

```
summary(NOAA_climate$TMIN) #10882 NAs
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -61.00   24.00   31.00   30.96   39.00   99.00   10882
```

```
levels(NOAA_climate$NAME)
```

```
## [1] "SQUAW VALLEY G.C., CA US"   "SQUAW VALLEY, CA US"
## [3] "TAHOE CITY CROSS, CA US"    "TAHOE CITY, CA US"
## [5] "WARD CREEK NUMBER 3, CA US"
```

```
summary(NOAA_climate$NAME)
```

```
##   SQUAW VALLEY G.C., CA US        SQUAW VALLEY, CA US
##                      14481                       6691
##    TAHOE CITY CROSS, CA US          TAHOE CITY, CA US
##                      14426                      40638
## WARD CREEK NUMBER 3, CA US
##                      14792
```

DO SUMMARY STATS TABLE

#USGS Data

Parameter	Summary
Total Number of Samples	22,734
Start Date	1957-10-01
End Date	2019-12-31
Gage Height (ft) Mean	5.86
Gage Height (ft) Median	6.32
Gage Height (ft) Min	0.26
Gage Height (ft) Max	9.40

#NOAA Climate Data

```
NOAA_summary <- summary(NOAA_climate)

kable(NOAA_summary, caption = "Summary Table of NOAA Climate Raw Data")
```

Table 2: Summary Table of

STATION	NAME	LATITUDE	LONGITUDE	ELEVATION	
USC00048474: 6691	SQUAW VALLEY G.C., CA US :14481	Min. :39.14	Min. :-120.3	Min. :1899	
USC00048758:40638	SQUAW VALLEY, CA US : 6691	1st Qu.:39.17	1st Qu.:-120.2	1st Qu.:1899	
USS0020K25S:14792	TAHOE CITY CROSS, CA US :14426	Median :39.17	Median :-120.2	Median :1903	M
USS0020K27S:14426	TAHOE CITY, CA US :40638	Mean :39.17	Mean :-120.2	Mean :2035	

STATION	NAME	LATITUDE	LONGITUDE	ELEVATION	
USS0020K30S:14481	WARD CREEK NUMBER 3, CA US:14792	3rd Qu.:39.17	3rd Qu.:-120.1	3rd Qu.:2072	3
NA	NA	Max. :39.20	Max. :-120.1	Max. :2447	
NA	NA	NA	NA	NA	