Rain Data

Lauren Temple

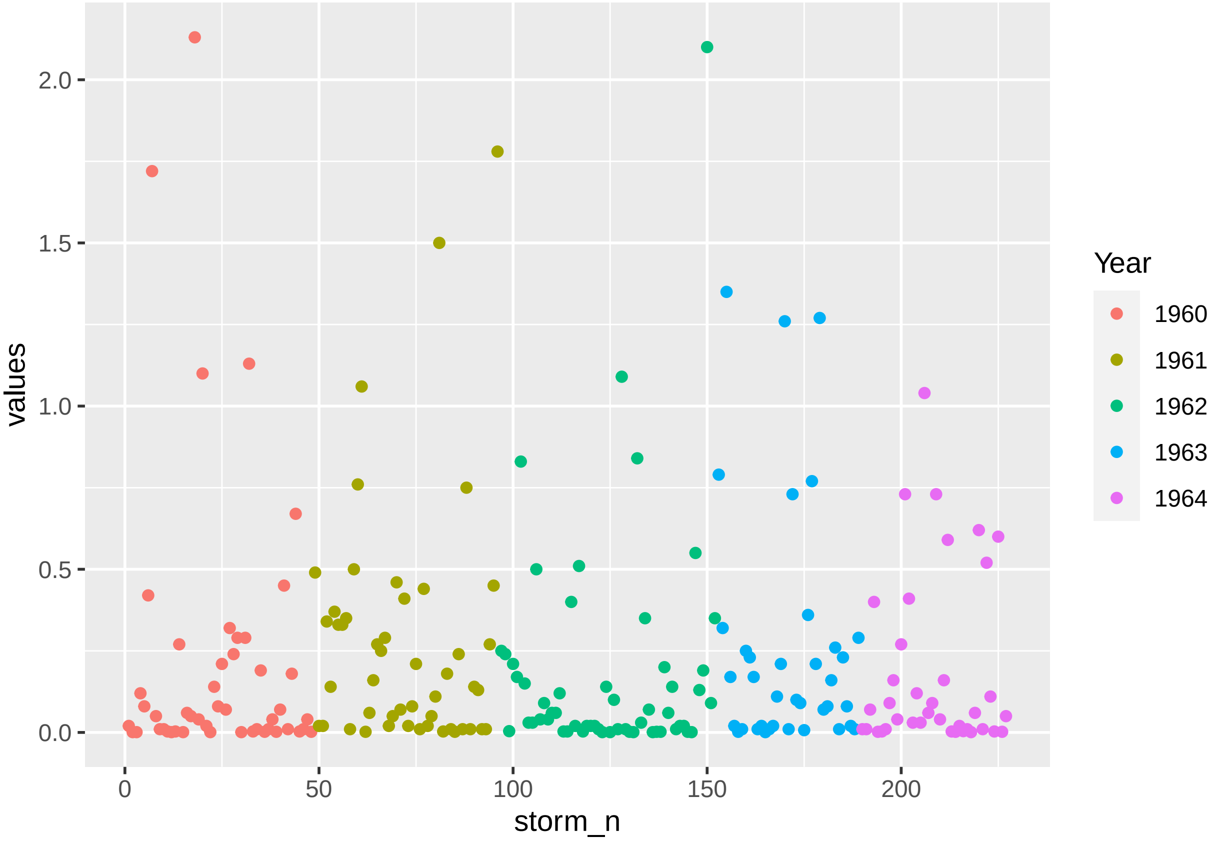
5/2/2022

Intro

This project is an exploration of the rainfall data collected in Southern Illionis between 1960 and 1964. In this report I will explore the distribution, fit a distribution to the data, simulate data using the parameters of the found distribution, and identify wet vs dry years.

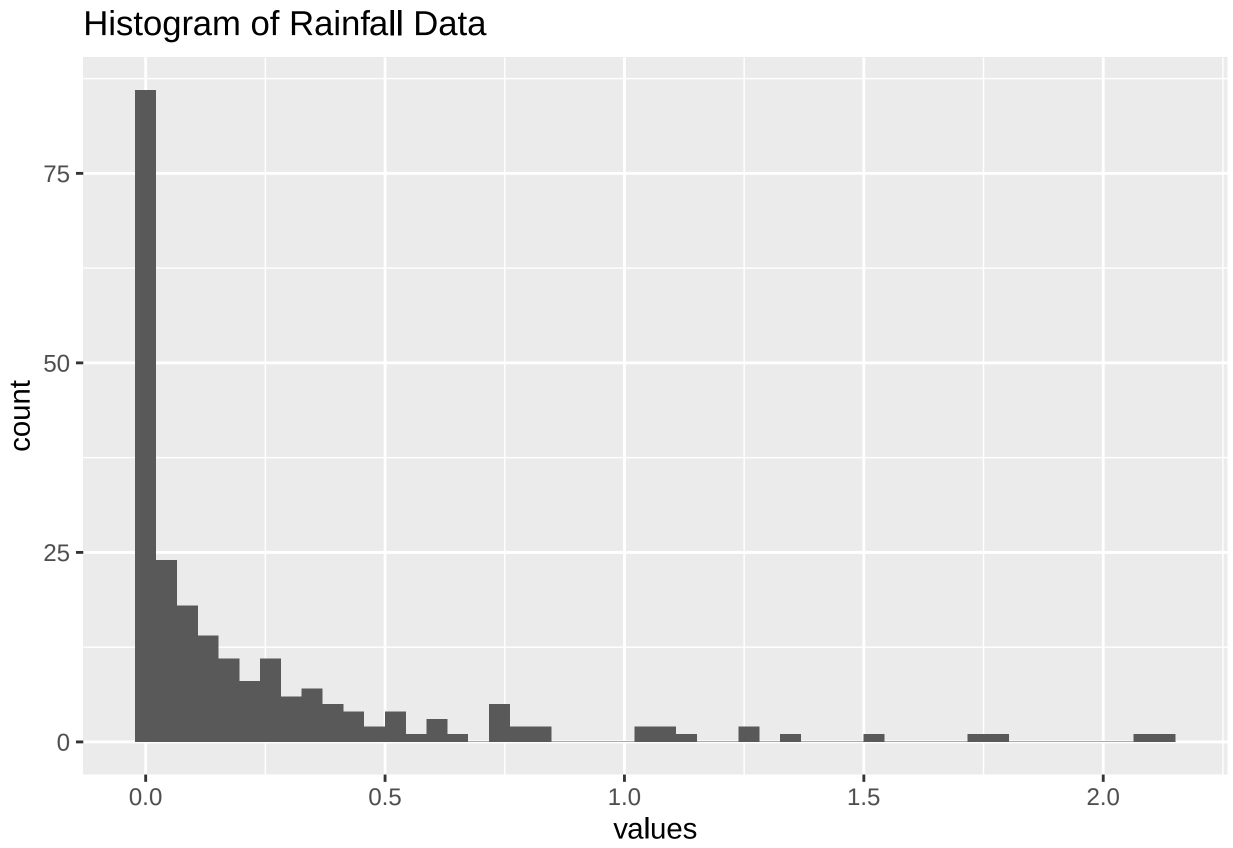
EDA

```
point
```



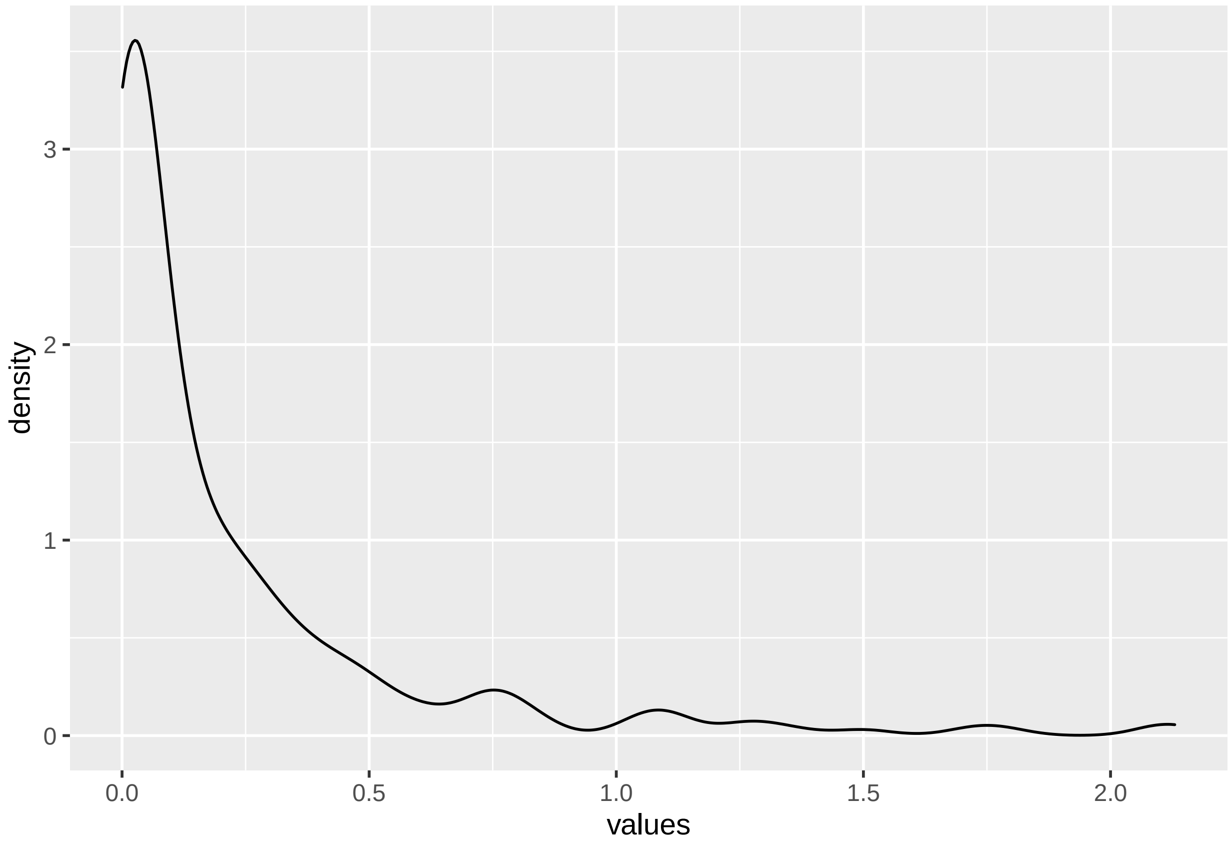
From this scatter plot we can see that the rainfall data across years appears to be similar. A majority of that data points are clustered at very low values close to zero, with a decrease in density as the rainfall value increases.

```
gghisto
```



We see this phenomenon reflected in our histogram of the rainfall values. The distribution is heavily right tailed with most values lying between 0.0 and 0.5.

```
ggdens
```

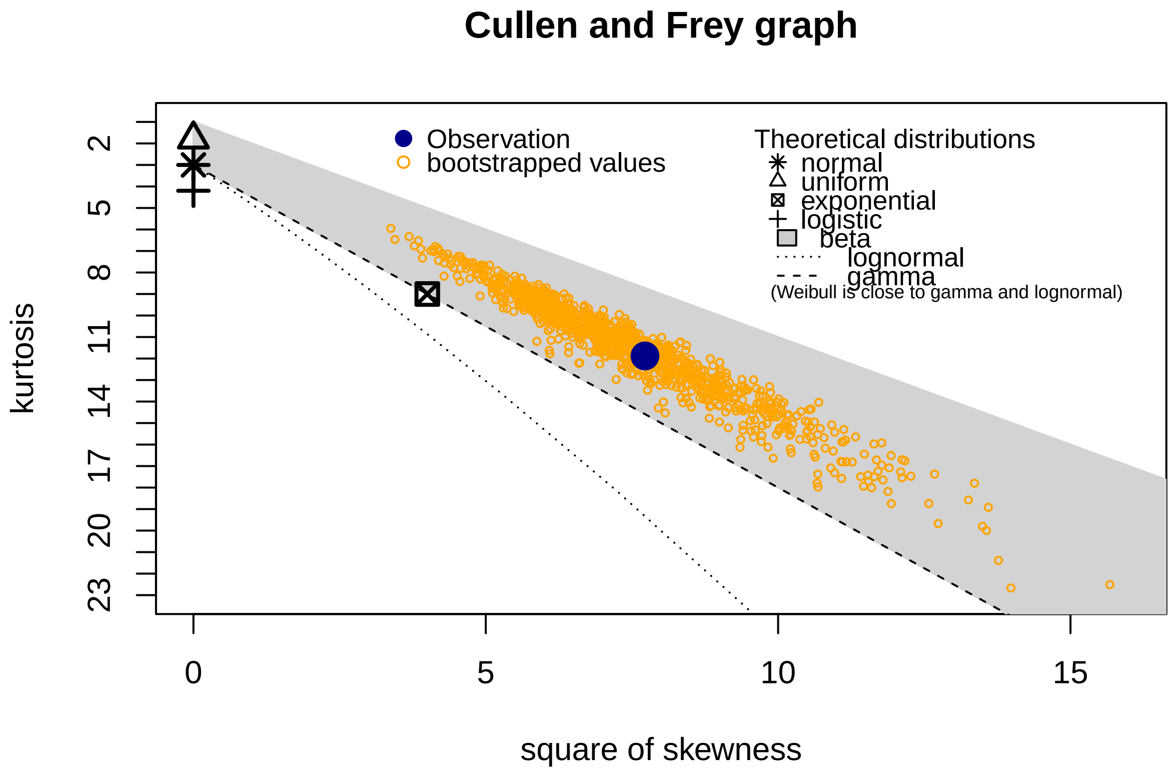


As we take a look at the density plot we can confirm our observations about the distribution of this dataset. We will need to find a distribution that is able to accurately represent this aspect of the data.

#Finding a distribution

I began by creating a Cullen and Frey graph to get an idea of which distributions might work well for this dataset.

```
descdist(tidy_data$values, boot = 1000)
```



```
## summary statistics
## ------
## min:  0.001   max:  2.13
## median:  0.07
## mean:  0.2243921
## estimated sd:  0.3658212
## estimated skewness:  2.778925
## estimated kurtosis:  11.87935
```

From this graph I decided to try Exponential, Gamma, and Weibull. I examined the goodness of fit statistics of each distribution, focusing on the KS and AD stats. I found that the Gamma distribution had the smallest KS and AD stats. I also examined the qqplot, pplot, density plot, and cdf plot for each distribution. The Gamma distribution also had the best fitting qq and pp plots. The exponential distribution had the best fitting density plot, however given the overall better fit of the Gamma distribution, I decided to go with that. Once I had decided on the Gamma distribution I did some research on what the Gamma distribution is typically used to model and found that rainfall data is one of its common uses. Below are the results from my Gamma distribution model fitting with fitdist.

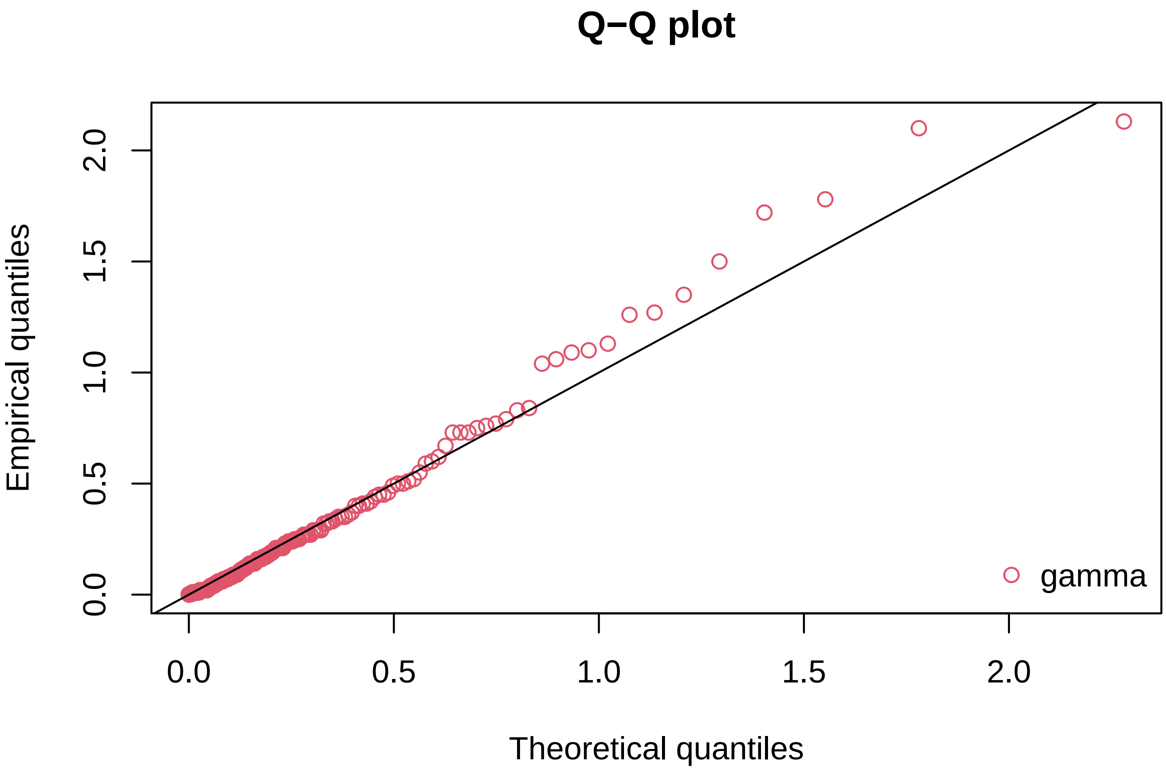
```
gModel <- fitdist(tidy_data$values, "gamma")
gofstat(gModel) ##lower KS and AD stat than exp
```

```
## Goodness-of-fit statistics
##                               1-mle-gamma
## Kolmogorov-Smirnov statistic   0.1110517
## Cramer-von Mises statistic     0.3573364
## Anderson-Darling statistic     2.3449731
##
## Goodness-of-fit criteria
##                                1-mle-gamma
## Akaike's Information Criterion   -366.6954
## Bayesian Information Criterion   -359.8455
```

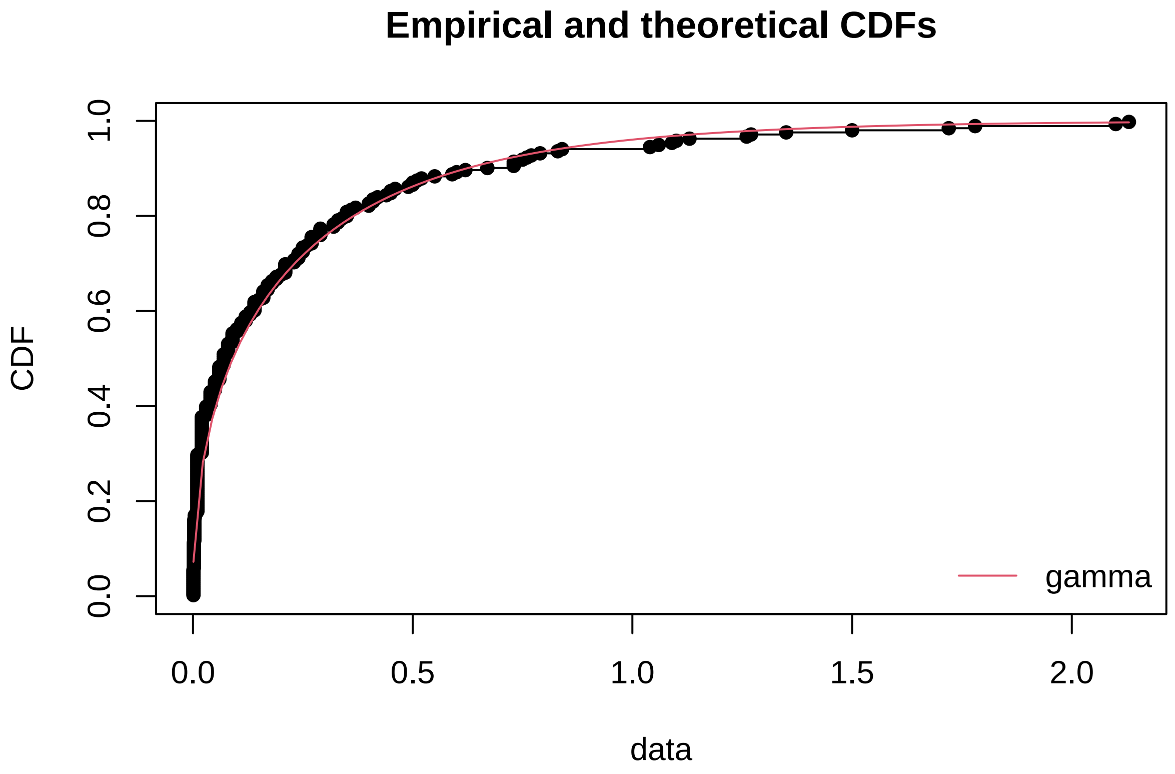
```
gModel$estimate
```

```
##     shape      rate
## 0.4408386 1.9648409
```

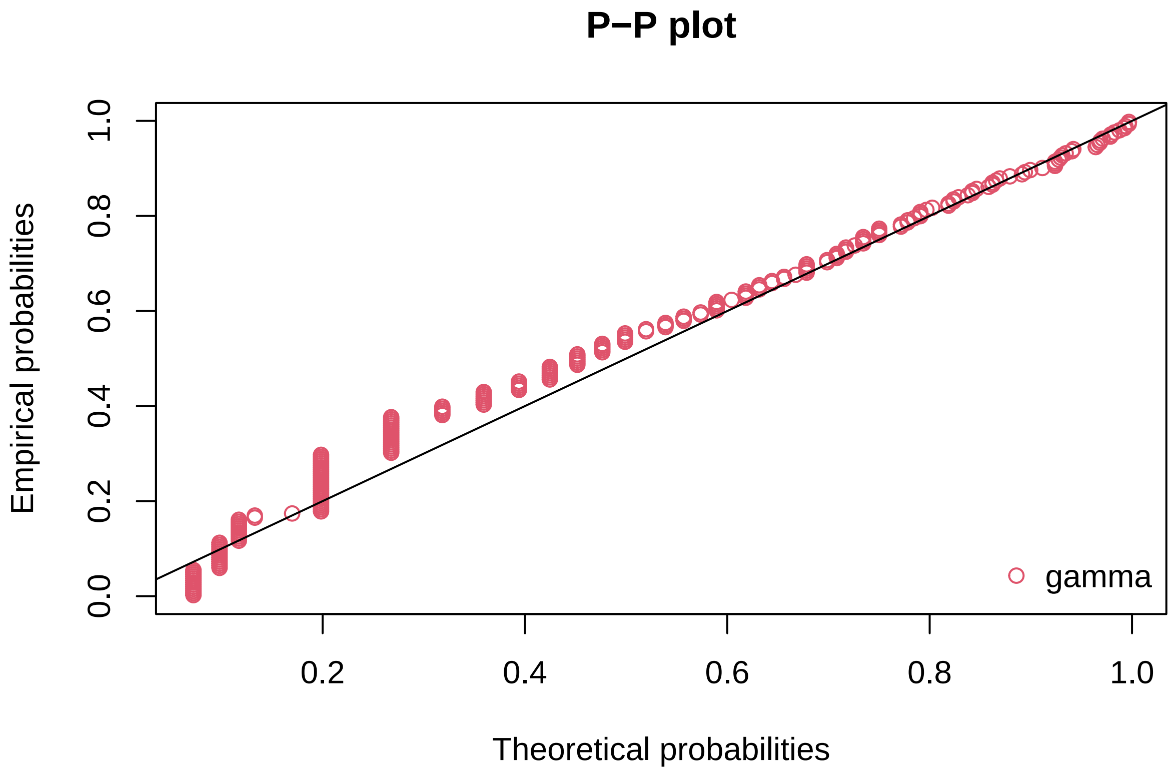
```
qqcomp(gModel)
```



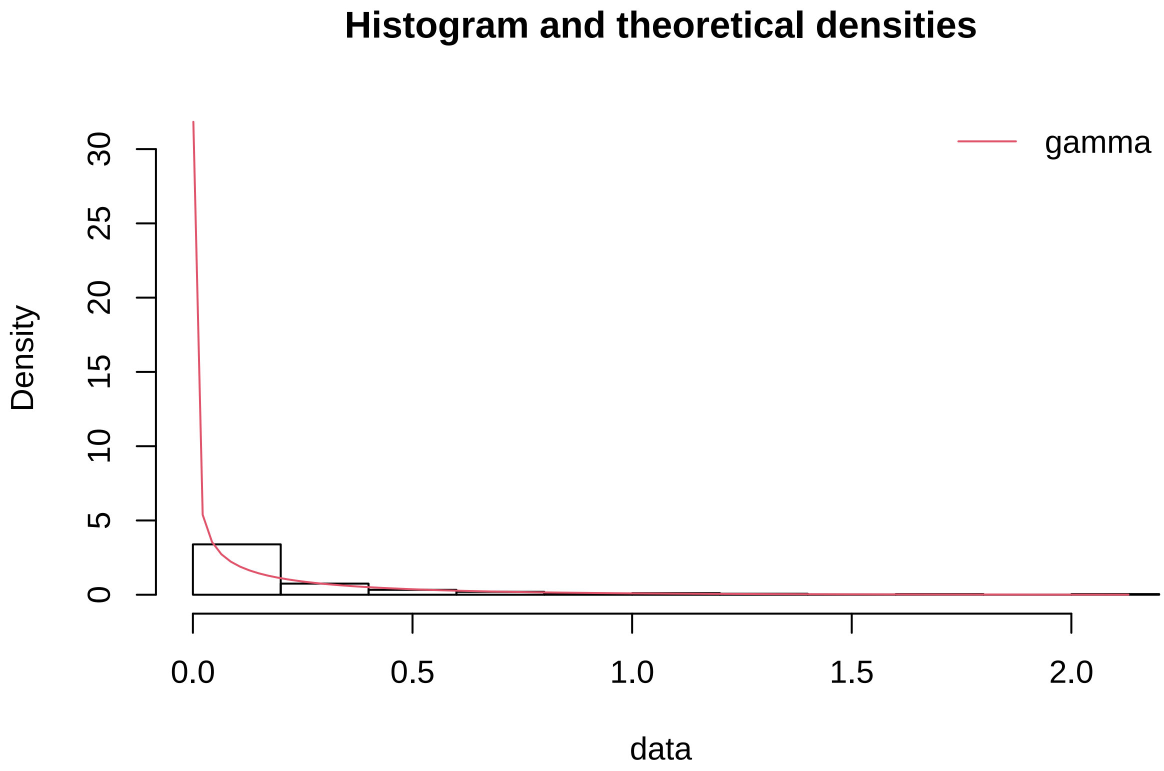
```
cdfcomp(gModel)
```



```
ppcomp(gModel)
```



```
denscomp(gModel)
```



From there I moved on to estimating the parameters of the distribution. I used mledist to do this. I found that the shape of the distribution is about 0.44 and the rate is about 1.96. I used these parameters to simulate some gamma distributed data and compare that with the actual rain data values.

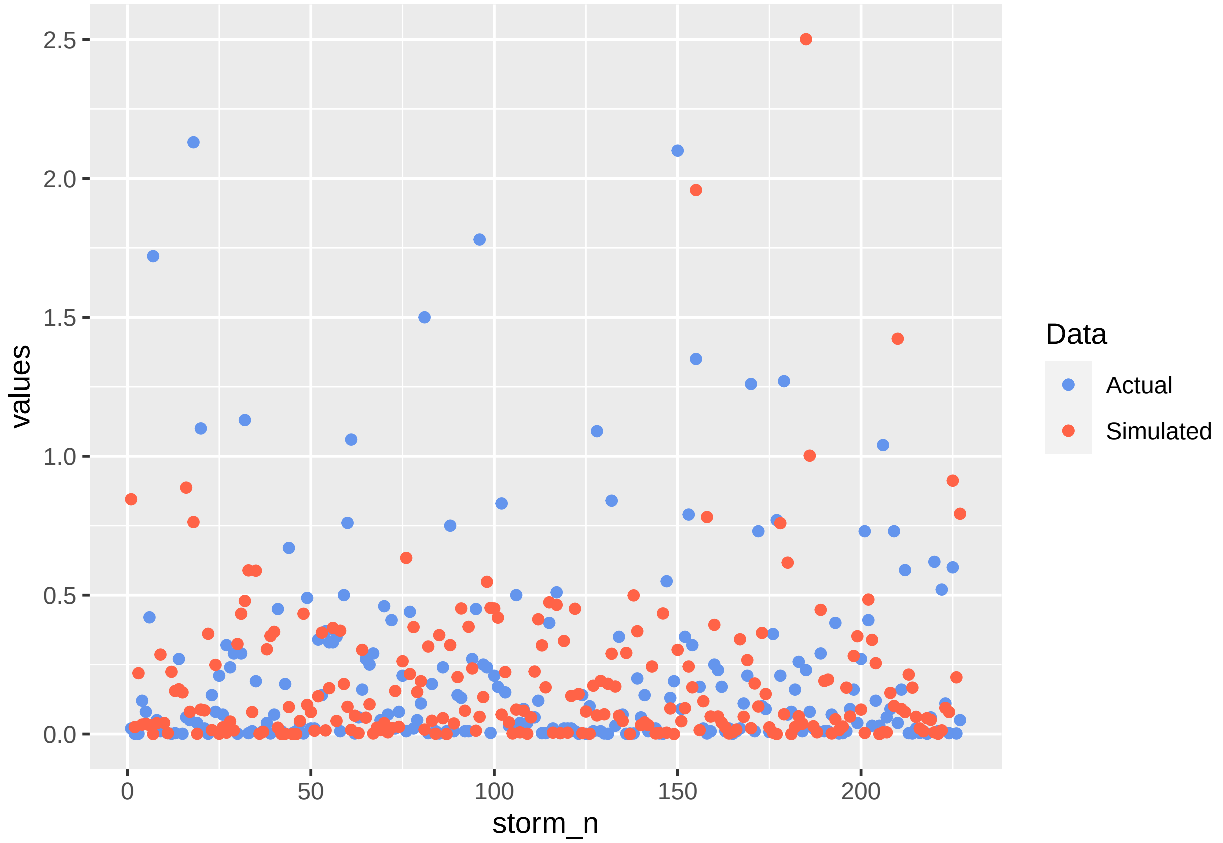
```
mledist(tidy_data$values, 'gamma')
```

```
## $estimate
##     shape      rate 
## 0.4408386 1.9648409 
## 
## $convergence
## [1] 0
## 
## $value
## [1] -185.3477
## 
## $hessian
##           shape       rate
## shape 1391.995 -115.53099
## rate  -115.531   25.92095
## 
## $optim.function
## [1] "optim"
## 
## $optim.method
## [1] "Nelder-Mead"
```

```
##
## $fix.arg
## NULL
##
## $fix.arg.fun
## NULL
##
## $weights
## NULL
##
## $counts
## function gradient
##       57       NA
##
## $optim.message
## NULL
##
## $loglik
## [1] 185.3477
```

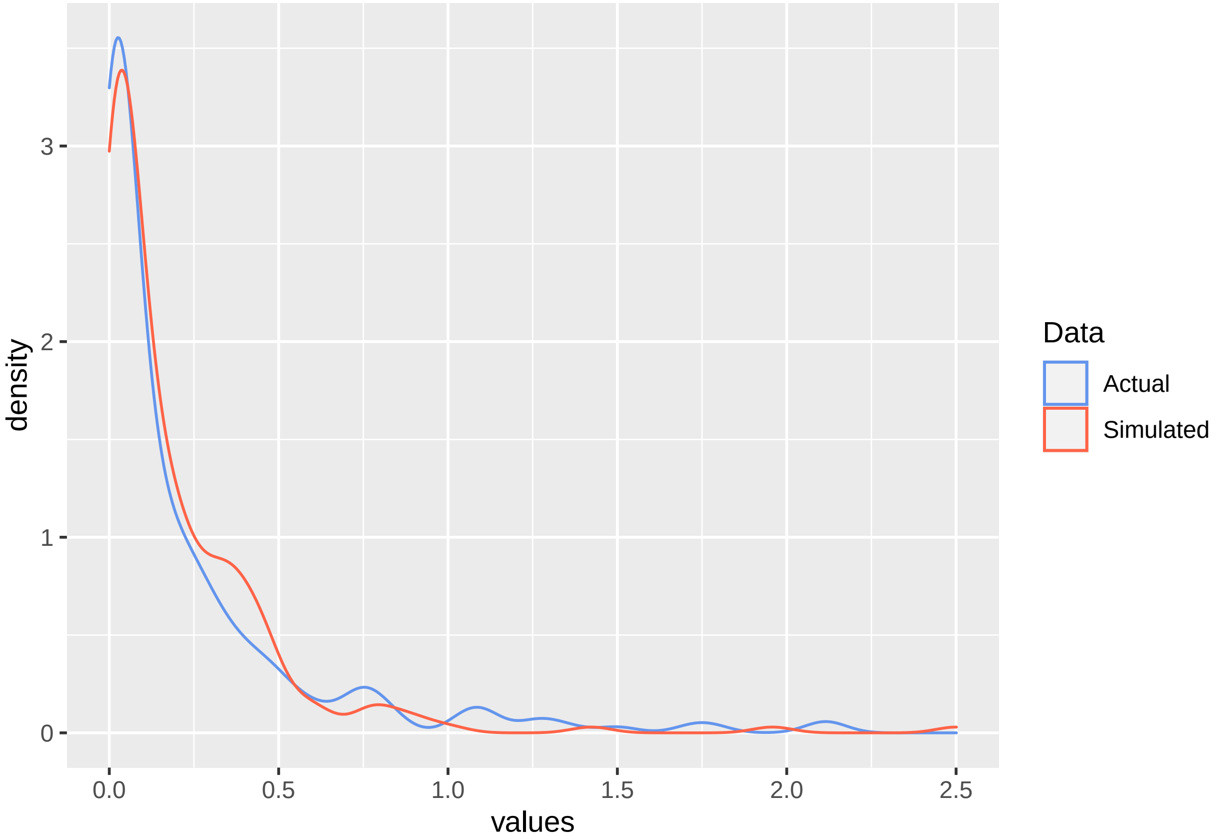
```
#estimate: parameter estimates
#Shape: 0.4408386
#Rate: 1.9648409
```

```
point_comp
```



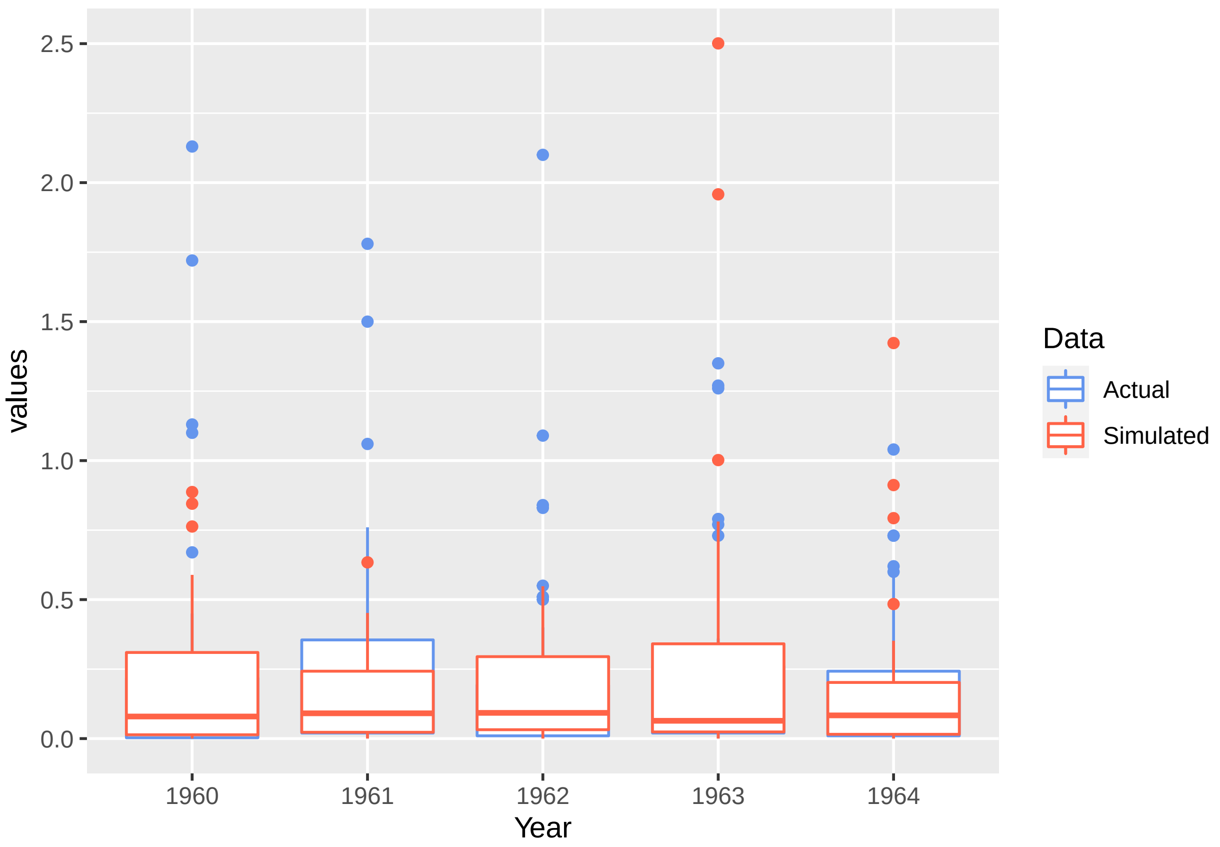
Here I plotted a scatter plot of the two distributions on top of each other to make sure that there was no obvious grouping of data in one dataset that is not present in the other. I found that the simulated data fits well among the actual rainfall data.

```
dense_comp
```



When looking at the comparison of the density plots between the actual rainfall data and simulated data we can see that the Gamma distribution does not quite capture the amount of zero values that are in the actual data but it does capture the general density of the data well.

```
box_comp
```



Here is a box plot comparison of the two datasets. The simulated Gamma data matches well with the actual rainfall data. A noteable exception is that the simulated data does have a maximum point of about 3 whereas the rainfall data only reaches values of about 2.

#Wet vs Dry years

Determining a wet vs dry year from a small dataset such as this one is difficult. I began examining Wet vs Dry years by looking at the summary statistics of the data. I started off with the tidy version of that data that I created.

```
summary(tidy_data)
```

```
##      values         Year       storm_n
##  Min.   :0.0010   1960:48   Min.   :  1.0
##  1st Qu.:0.0100   1961:48   1st Qu.: 57.5
##  Median :0.0700   1962:56   Median :114.0
##  Mean   :0.2244   1963:37   Mean   :114.0
##  3rd Qu.:0.2700   1964:38   3rd Qu.:170.5
##  Max.   :2.1300             Max.   :227.0
```

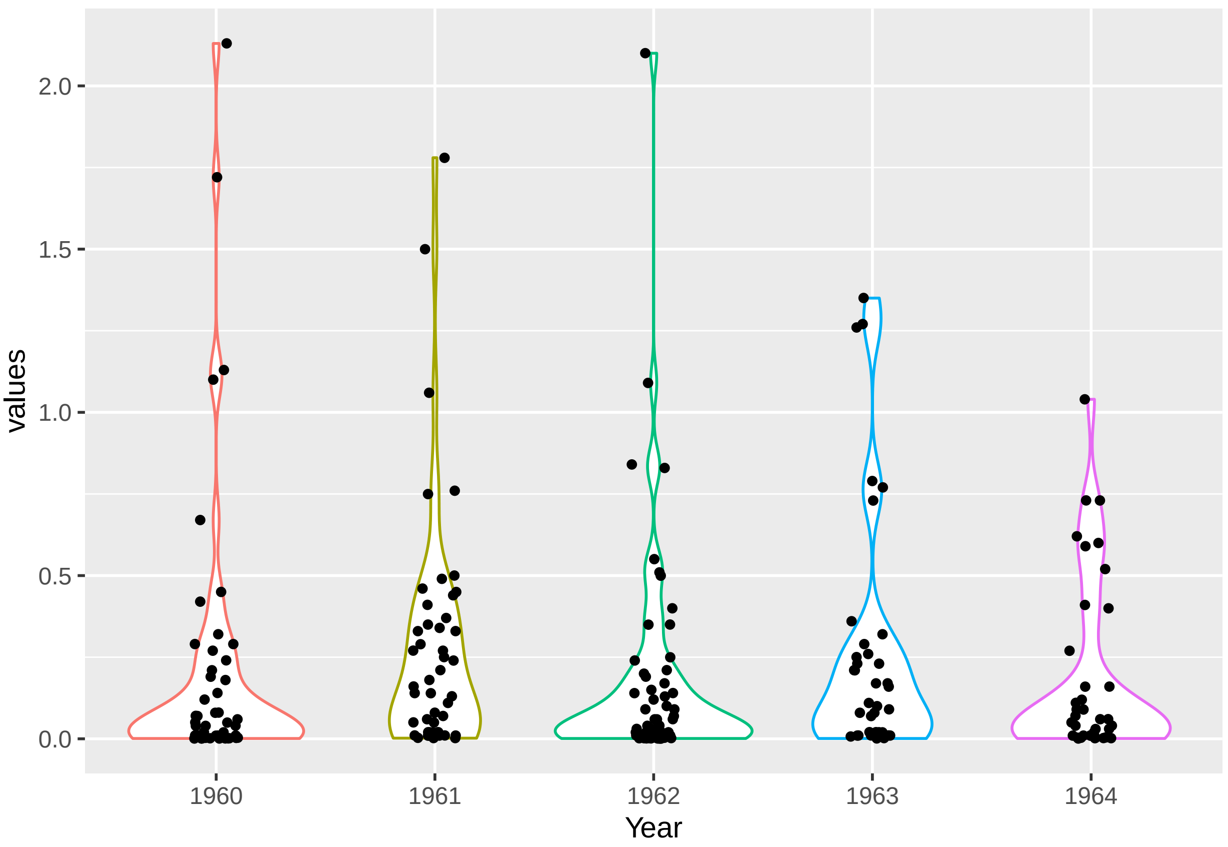
From this summary table I gathered that the mean rainfall across all five years was 0.2244. I can use that as a benchmark and judge the other years off of that average to determine if they are wet or dry years. I also was able to clearly see the number of storms per year. We can see that 1962 has the most storms and 1963 has the least number of storms. The storm_n column of this summary table is not meaningful as it is creating summary stats based off the cumulative number of the storm over the years. After looking at the tidy data I decided to go back to the original data set to get more in depth summary stats on each year of rainfall. I will keep in mind the overall average rainfall is 0.2244 and 1962 had the highest number of storms.

```
summary(rain_data)
```

```
##       1960            1961            1962            1963       
##  Min.   :0.0010   Min.   :0.0020   Min.   :0.0010   Min.   :0.0010  
##  1st Qu.:0.0030   1st Qu.:0.0200   1st Qu.:0.0100   1st Qu.:0.0200  
##  Median :0.0450   Median :0.1500   Median :0.0500   Median :0.1100  
##  Mean   :0.2203   Mean   :0.2749   Mean   :0.1847   Mean   :0.2624  
##  3rd Qu.:0.2175   3rd Qu.:0.3550   3rd Qu.:0.1925   3rd Qu.:0.2600  
##  Max.   :2.1300   Max.   :1.7800   Max.   :2.1000   Max.   :1.3500  
##  NA's   :8        NA's   :8                         NA's   :19      
##       1964       
##  Min.   :0.0010  
##  1st Qu.:0.0100  
##  Median :0.0550  
##  Mean   :0.1871  
##  3rd Qu.:0.2425  
##  Max.   :1.0400  
##  NA's   :18      
```

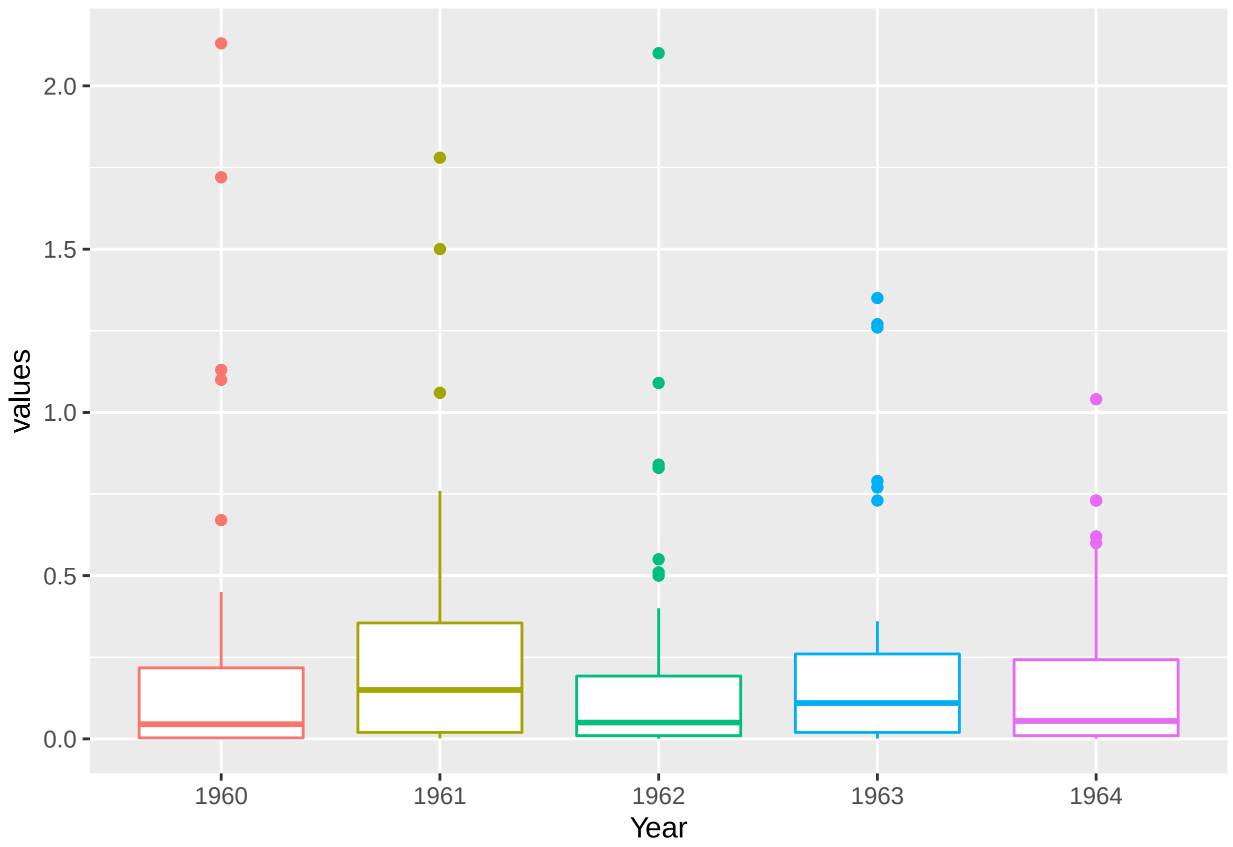
From this summary table we can see that 1961 and 1963 are the only two years with mean rainfall values over 0.2244. If we use that metric to determine wet vs dry years then 1961 and 1963 were wet years and 1960, 1962, 1964 were dry years. As a reminder, 1962 had the greatest number of storms, however this does not mean that they had the greatest amount of rainfall, in fact 1962 has the lowest mean value of rainfall. After looking at these summary stats I decided to dig deeper into the distribution of rainfall each year.

```
ggvio
```



I chose to examine the distribution of rainfall each year with a violin plot of the data. Looking at this plot we can see that a majority of the data points sit at low values each year. 1960, 1962, and 1964 have the largest density of low value rainfall points. This lines up with our finding that 1962 had the highest number of storms but not a high average value of rainfall. We can see that apart from one data point at about 2 inches, the rest of the rainfall data for 1962 lies below 1 inch and a majority of it lies below 0.5 inches. Another observation we can make it that all of the years share similar shapes of density distributions. We see the most similarity between 1960, 1962. We might assume that the rainfall these two years was similar.

```
ggbox
```



The boxplot representation of the data gives us an idea of the distribution and their possible outliers. We can see that all of the years are very similar with 1961 having the largest spread of data within its 1st through 3rd quartiles. We can also see that a majority of the data falls under 0.5.

```
rain_summarydf
```

```
##   Years    sum      mean storm_n
## 1  1960 10.574 0.2202917      48
## 2  1961 13.197 0.2749375      48
## 3  1962 10.346 0.1847500      56
## 4  1963  9.710 0.2624324      37
## 5  1964  7.110 0.1871053      38
```

Going back to a summary table, I created this one to show the total rainfall, the mean rainfall, and the number of storms each year. Given what we already know from looking at other summary tables and the violin plot we can confirm that 1961 is our best candidate for the wettest year. Although 1961 did not have the greatest number of storms it did have the most rain during the storms that did occur. I think that is a better measure of wet vs dry than just looking at the number of storms. I do not feel confident in labeling the other years wet or dry. I think they are all too similar to distinguish with the amount of data that we have.

Conclusion

I think that the results pertaining to the Gamma distribution will only generalize to other years of rainfall in Illinois. I also think that as we continue to get further from the original years in this data set that the results will generalize less and less due to the ever changing climate of this planet. Although these changes may be gradual they will impact the generalizability of the data analysis from this dataset. This project has provided me with the opportunity to dig into a novel dataset and discover its distribution as well as think critically about how well it applies to the actual data and how it can be used. I think the next steps would be to obtain more data and see if the parameters of this distribution are a good fit for that data as well. I think that as more data is collected the parameters will change and become better tuned to the actual dataset. I also feel that with more data we would be able to better determine wet vs dry years.