A Statistical Analysis of Streamflow data in Oklahoma

By: Jake Mammen

Abstract

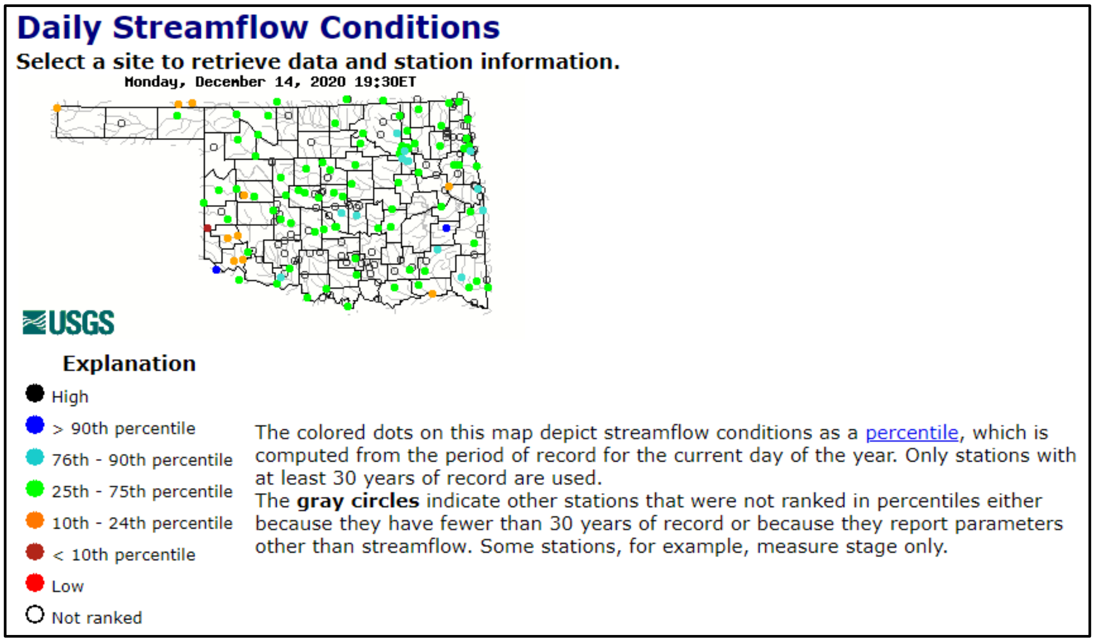
As the climate continues to change and human populations increase, the demands for water will also grow. For many years, local and state officials have debated ways on how to manage the way water is used. Research suggests agriculture and metropolitan areas account for the majority of water use, resulting in a steady decline of available water from the water-tables in which they pull from. In other words, nature can't keep up with how fast water is being discharged. This study aims to compare the differences in streamflow of the North Canadian River across two locations in state of Oklahoma through the use of statistical analysis. The statistical analysis will consist of multiple tests that will include: one and two sample T-tests, Chi-squared tests, ANOVA tests, correlation and regression tests, PCA tests, and lastly clustering to examine quantitative and qualitative data. Streamflow data can be messy to try and analyze given the nature of how the data is measured. Overall, this study found that the streamflow data across both locations had a lot more similarities than differences. However, there was enough information to suggest that there were differences in discharge over time. Similar studies like this can help bring forth useful information, to aid decision makers on how to manage water use in the future.

Introduction

The purpose of this study is to examine the spatial variability of streamflow data across two locations within the state of Oklahoma and compare to one another. I find it very interesting how water moves throughout the land. In addition, its intriguing to me to learn how much water, where it comes from, where it goes, and then how it gets replenished regarding water as a resource. Water is one of the most important resources to mankind. It is used for so many different things such as, in agriculture, power/electricity, and drinking water. According to a study conducted by Bureau Reclamation, the pressures on water resources in the southwestern United States is increasing as the populations continue to grow and the temperatures continue to rise (Reclamation, 2016). Water resources across the Great Central Plains can mainly be attributed to agriculture (Lehner et. al, 2017). More specifically and for the purpose of understanding the nature of water resources within Oklahoma, a study published through the USGS, Wahl et. al (1997) suggest that releases from Canton lake supply the North Canadian River and account for public-water supply withdrawals for half of the city of Oklahoma City. In addition, about ninety-two percent of the total withdrawals of surface water in the basin upstream of Oklahoma City come from the Beaver-North Canadian River and its tributaries acting as the primary source of public-water supply (Lurry et. al, 1996). The High plains water-table aquifer is primarily an aquifer that gets recharged through precipitation, according to Wahl et. al (1997). With that said back in the back in the 1960's that water-table was at equilibrium. It is no longer at that position possibly due to the fact that more water gets taken out for agricultural purposes, with no natural recharge (Havens et. al, 1984). Due to the severity of this problem the Boyle Engineering Corp, along with Wahl et. al argued that these severe land-use practices and decline in ground water levels, has contributed to the decreases in discharge of the Beaver-North Canadian River (Boyle Engineering Corp., 1987; and Wahl and Wahl, 1988).

The fight for water will certainly be something to watch for in the coming years as we see the impacts that climate change may have on streamflow's across the United States. However, before that time occurs, we need to understand the streamflow data means and represents so that we can then make decisions based of those results. In a study conducted by Dinpashoh et. al (2019), they suggest that streamflow's of most rivers have been altered or have changed in recent years. Ultimately analyses of streamflow have come in a variety of forms. Not only have they happened on a local level but all the way up to the national level in order to infer how the climate may be impacting streamflow data (Wimbrow, 2012). Several different studies, Guastini (2019) and Kao (2016), both argue that studying and understanding the characteristics of how streamflow response varies in time and at different spatial scales is important for assessing runoff. In addition, understanding water quality and water availability is crucial to making decisions to counter streamflow deficiencies (Kao, 2016).

Figure A. A Map of Streamflow stations across Oklahoma



Daily Streamflow Conditions

Select a site to retrieve data and station information.

Monday, December 14, 2020 19:30ET

Explanation

● High

● > 90th percentile

● 76th - 90th percentile

● 25th - 75th percentile

● 10th - 24th percentile

● < 10th percentile

● Low

○ Not ranked

The colored dots on this map depict streamflow conditions as a percentile, which is computed from the period of record for the current day of the year. Only stations with at least 30 years of record are used.

The **gray circles** indicate other stations that were not ranked in percentiles either because they have fewer than 30 years of record or because they report parameters other than streamflow. Some stations, for example, measure stage only.

Source: United States Geological Survey

The focus area for this study lies within the state of Oklahoma. One location up in northwestern Oklahoma and one location in central Oklahoma. This area really encompasses the full spectrum of land-use practices and should provide a good comparison between streamflow datasets.

Data and Methods

The data used and analyzed for this project comes from the United States Geological Survey. The USGS was created by Congress back in 1879 and has evolved over the last 125 years dedicated to pushing the knowledge of science and technology beyond its limits. USGS's natural science expertise and archives of hundreds of thousands earth and biological data holdings are what keep it at the fore front of scientifical research all around the world (Who We Are, 2020).

The data I chose for this research is related to streamflow. More specifically, I pulled streamflow data from the North Canadian River in the state of Oklahoma. The data represents streamflow measurements from the middle North Canadian River and lower North Canadian River. I chose two locations to analyze, those being, the North Canadian River in Woodward, Oklahoma and the North Canadian River in Harrah, Oklahoma (OKC). The reason I chose these locations is because I wanted to see if I could look at the differences between streamflow from the North Canadian River further upstream in northern Oklahoma and compare it to the differences in streamflow closer to the Oklahoma, City metro or central Oklahoma. The Woodward, Oklahoma streamflow data was pulled from the USGS 07237500 North Canadian River at Woodward, OK in Woodward County. The hydrologic unit code for this location is 11100301 with a drainage area of approximately 11,883 square miles, and a contributing drainage area of 8,386 square miles. The Harrah, Oklahoma streamflow data was pulled from the USGA 07241550 North Canadian River near Harrah, OK in Oklahoma County. The hydrologic unit code for this location is 11100302 with a drainage area of approximately 13,775 square miles, and a contributing drainage area of 10,278 square miles.

In order to look at the differences in streamflow data and compare between the two locations, I had to go in and pull the field measurements data with channel information included. For Woodward, Oklahoma, there are 308 measurements spanning from 1979-2020. In Harrah, Oklahoma, there are 311 measurements over the period of 1980-2020. There are a total of 32 variables in each dataset. The variables are as follows:

- Agency code
- Site number
- Measurement number
- Date/time
- Time zone codes
- Measurement used
- Who
- Agency
- Streamflow (cubic feet per second)
- Gage height (feet)
- Gage height change (feet)
- Measurement duration (decimal hours)
- Measurement rating (quality of measurement)

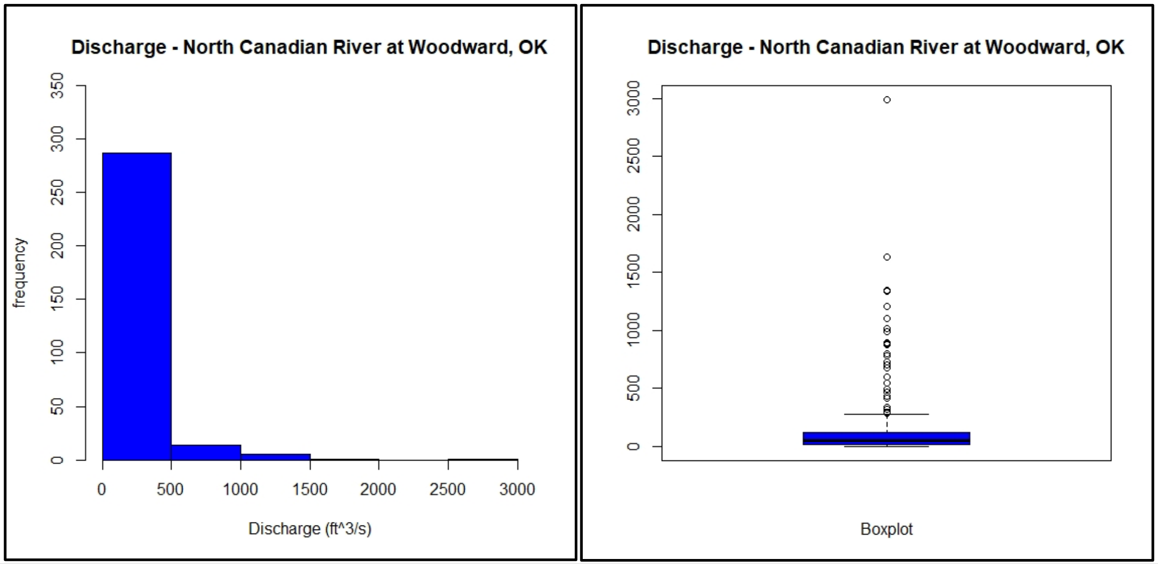
- Control type (condition)
- Flow adjustment code (adjustment code)
- Channel number
- Channel name
- Measurement type
- Streamflow method
- Velocity method
- Channel flow (cubic feet per second)
- Channel width (feet)
- Channel velocity (feet per second)
- Channel area (square feet)
- Channel stability
- Channel material
- Channel evenness (from bank to bank)
- Longitudinal velocity description
- Horizontal velocity description
- Vertical velocity description
- Channel location code
- Channel location distance

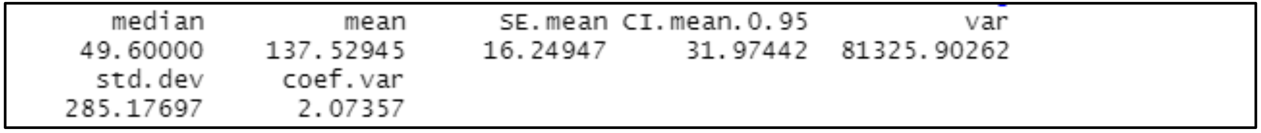
The next step was to download each dataset, then import them into R. From there I can run the data through multiple statistical tests such as: one and two sample T-tests, Chi-squared tests, ANOVA tests, correlation and regression tests, PCA tests, and lastly clustering. The reason multiple statistical tests will be needed is because both datasets contain quantitative and qualitative data. It was important that I pick two locations with similar measurement counts and time ranges to minimize error and improve representation across both samples. For the purpose of the research, I only chose variables that were the most representative of measuring streamflow.

Results and Analysis

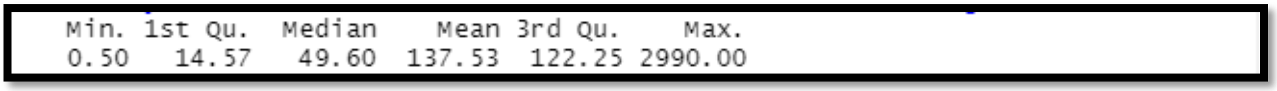
In this section I will discuss the results I found through the multiple types of statistical tests and analyze how the streamflow data may differ across both locations. After looking at the data from both locations, I realized out of 32 total variables not every one of those variables could be used. I determined what variables may be useful and ran those variables through a series of tests. In order to understand the nature of both datasets we need to visual the data, first.

Figure 1. Distribution of Discharge from the N. Canadian River at Woodward, OK

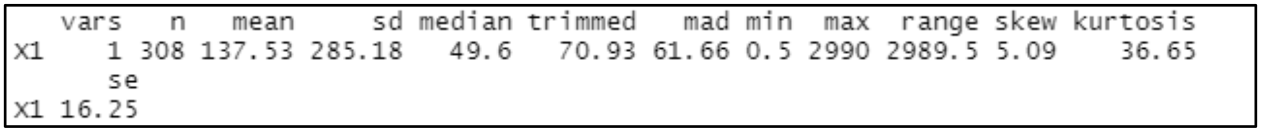




```
      median         mean      SE.mean CI.mean.0.95          var
    49.60000    137.52945     16.24947     31.97442  81325.90262
     std.dev     coef.var
   285.17697      2.07357
```



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.50   14.57   49.60  137.53  122.25 2990.00
```



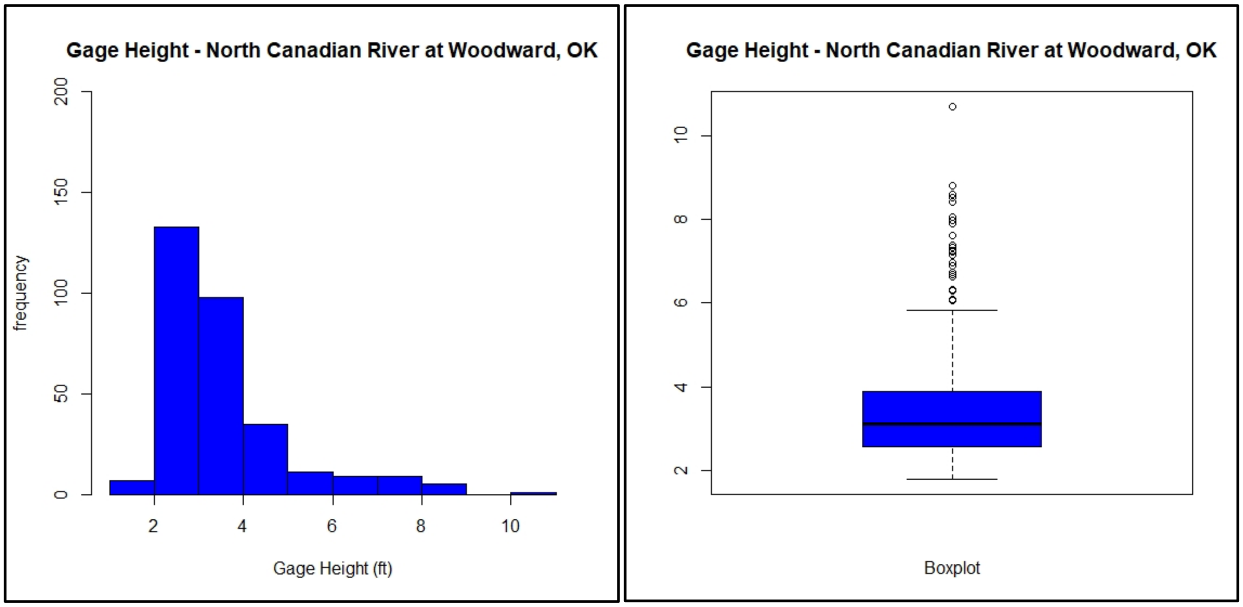
```
   vars   n   mean     sd median trimmed   mad min  max  range skew kurtosis
X1    1 308 137.53 285.18   49.6   70.93 61.66 0.5 2990 2989.5 5.09    36.65
      se
X1 16.25
```

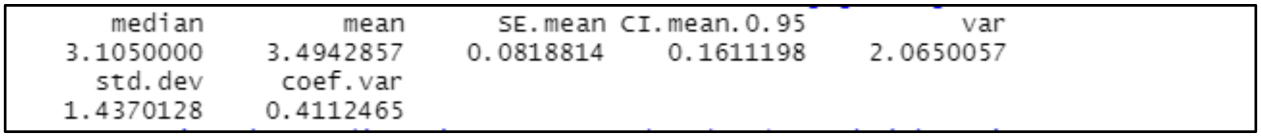
In the above figure you will notice the distribution is skewed right. To better understand the central tendency and variability of this data I had to use the five-number summary, which is expressed in the heavier outlined box above. Furthermore, we can express the variability by finding the interquartile range which is:

$$IQR = Q3 - Q1 = 122.25 - 14.57 = 107.68$$

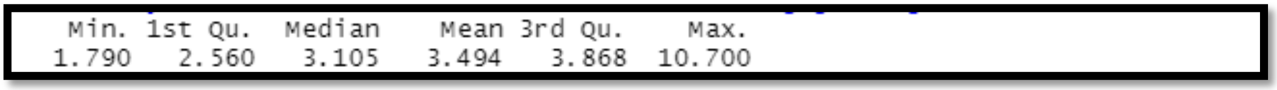
After calculating the interquartile range, we get 107.68 which explains the variability of this variable. While the IQR shows us how much of the data lies within a certain range, the boxplot indicates there are a handful of possible outliers present within the discharge dataset. However, it's hard to get an accurate representation due to dramatic fluctuations within the dataset and I don't necessarily think it should be thrown out.

Figure 2. Distribution of Gage Height from the N. Canadian River at Woodward, OK

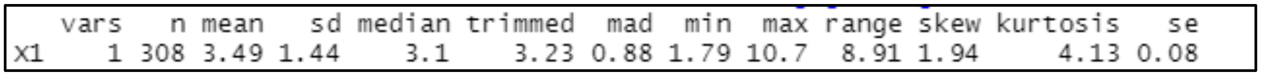




```
    median         mean      SE.mean CI.mean.0.95          var 
 3.1050000    3.4942857    0.0818814    0.1611198    2.0650057 
   std.dev     coef.var 
 1.4370128    0.4112465 
```



```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 1.790   2.560   3.105   3.494   3.868  10.700 
```



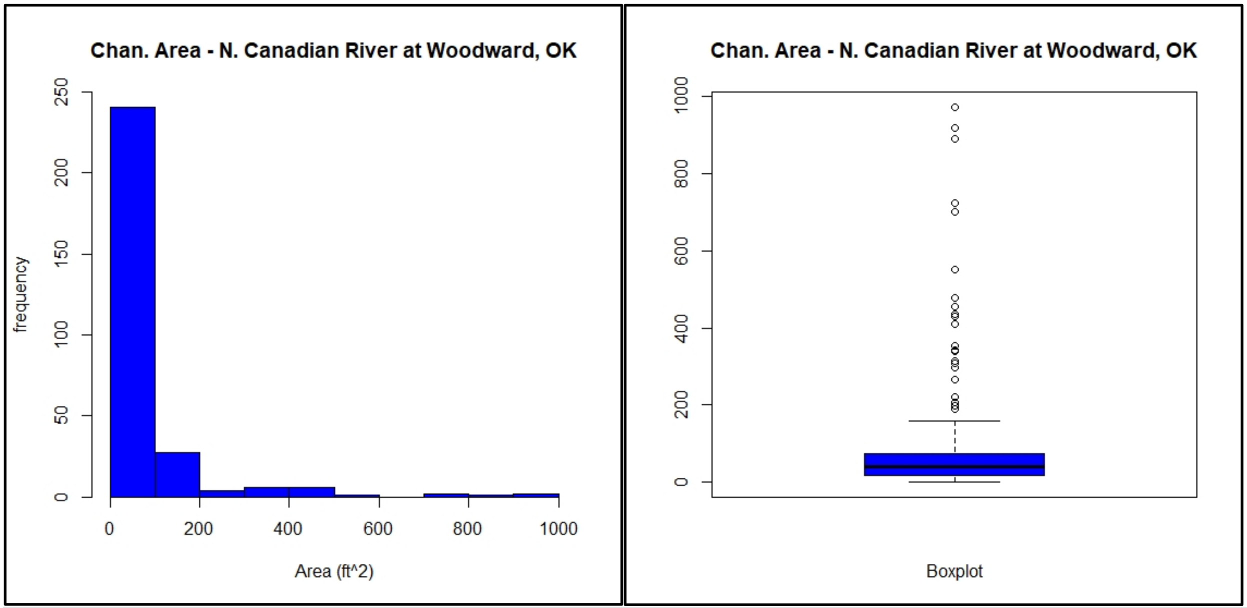
```
   vars   n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 308 3.49 1.44    3.1    3.23 0.88 1.79 10.7  8.91 1.94     4.13 0.08
```

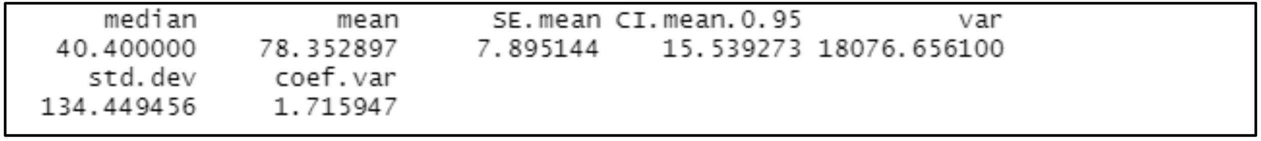
In this figure you will notice the distribution is also skewed right but not as extreme. Similar to analyzing the discharge data, to better understand the central tendency and variability of this data I had to use the five-number summary, which is expressed in the heavier outlined box above. Furthermore, we can express the variability by finding the interquartile range which is:

$$IQR = Q3 - Q1 = 3.868 - 2.560 = 1.308$$

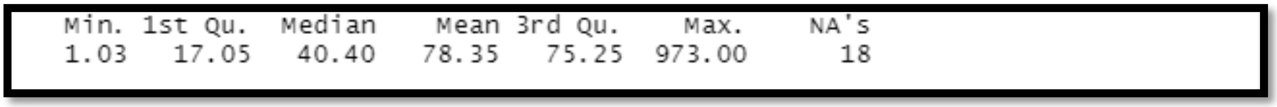
After calculating the interquartile range, we get 1.308 which explains the variance of this variable. While the IQR shows us how much of the data lies within a certain range, the boxplot indicates there are a handful of possible outliers present within the discharge dataset. However, it's hard to get an accurate representation due to dramatic fluctuations within the dataset and I don't necessarily think it should be thrown out.

Figure 3. Distribution of the Channel Area for the N. Canadian River at Woodward, OK

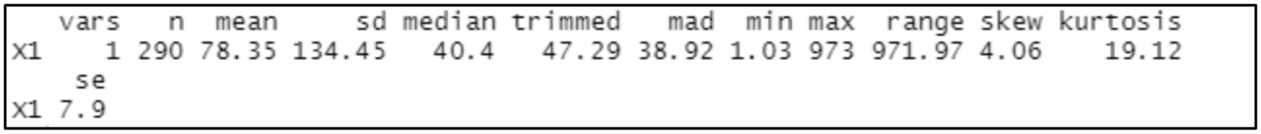




```
     median          mean       SE.mean CI.mean.0.95          var
  40.400000     78.352897      7.895144    15.539273 18076.656100
    std.dev      coef.var
 134.449456      1.715947
```



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1.03   17.05   40.40   78.35   75.25  973.00      18
```



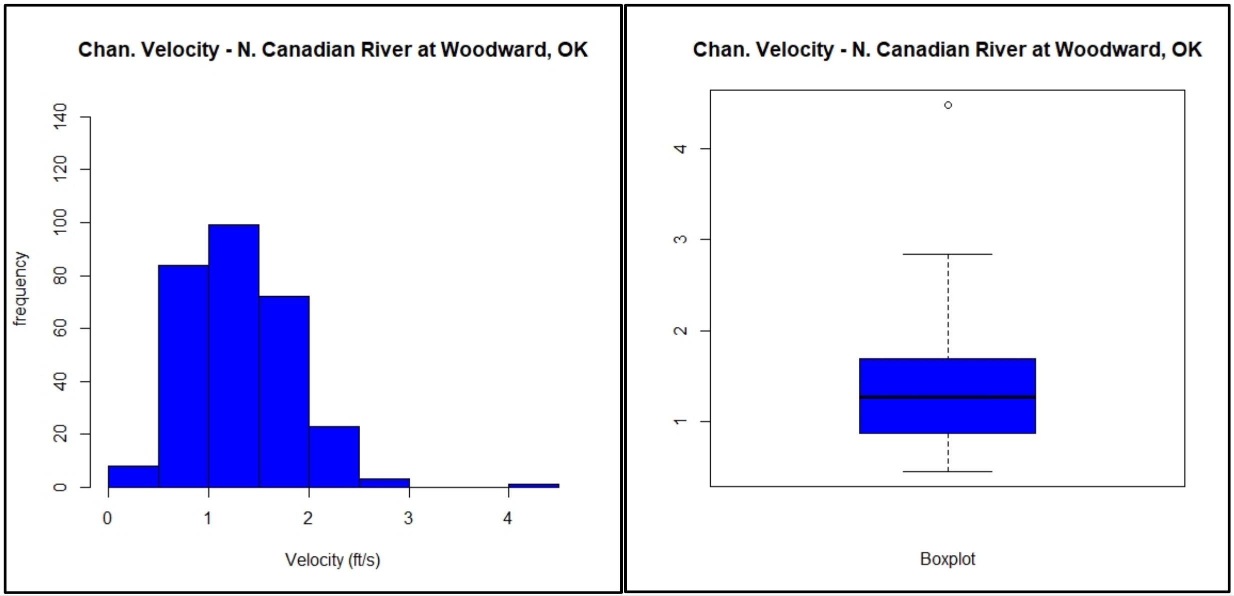
```
   vars   n  mean     sd median trimmed   mad  min max  range skew kurtosis
X1    1 290 78.35 134.45   40.4   47.29 38.92 1.03 973 971.97 4.06    19.12
     se
X1 7.9
```

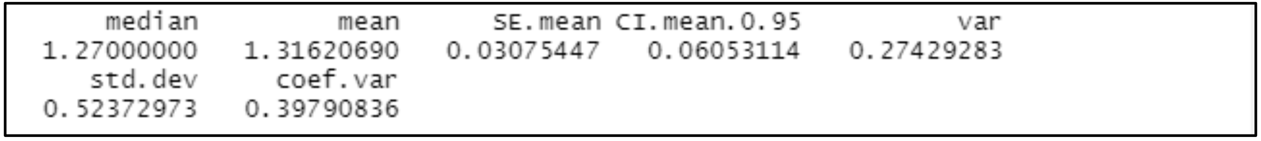
In the figure above you will see that the distribution is skewed right. Using the five-number summary helps us better understand the central tendency and variability of this data, which is expressed in the heavier outlined box above. Furthermore, we can express the variability by finding the interquartile range which is:

$$IQR = Q3 - Q1 = 75.25 - 17.05 = 58.2$$

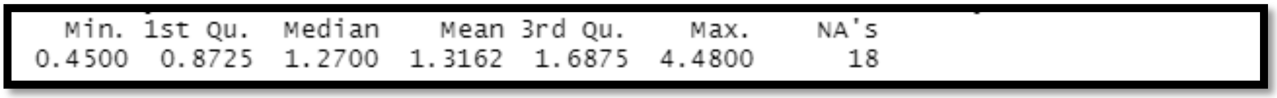
After calculating the interquartile range, we get 58.2 which explains the variance of this variable. While the IQR shows us how much of the data lies within a certain range, the boxplot indicates there are a handful of possible outliers present within the discharge dataset. However, it is hard to get an accurate representation due to dramatic fluctuations within the dataset and I don't necessarily think it should be thrown out.

Figure 4. Distribution of Channel Velocity for the N. Canadian River at Woodward, OK

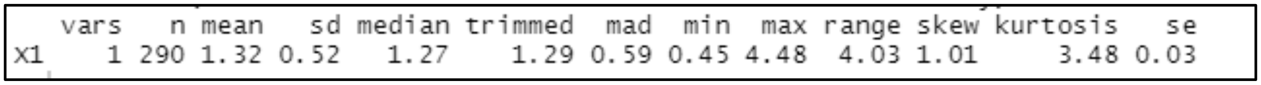




```
      median           mean       SE.mean CI.mean.0.95          var
 1.27000000     1.31620690    0.03075447   0.06053114    0.27429283
     std.dev       coef.var
 0.52372973     0.39790836
```



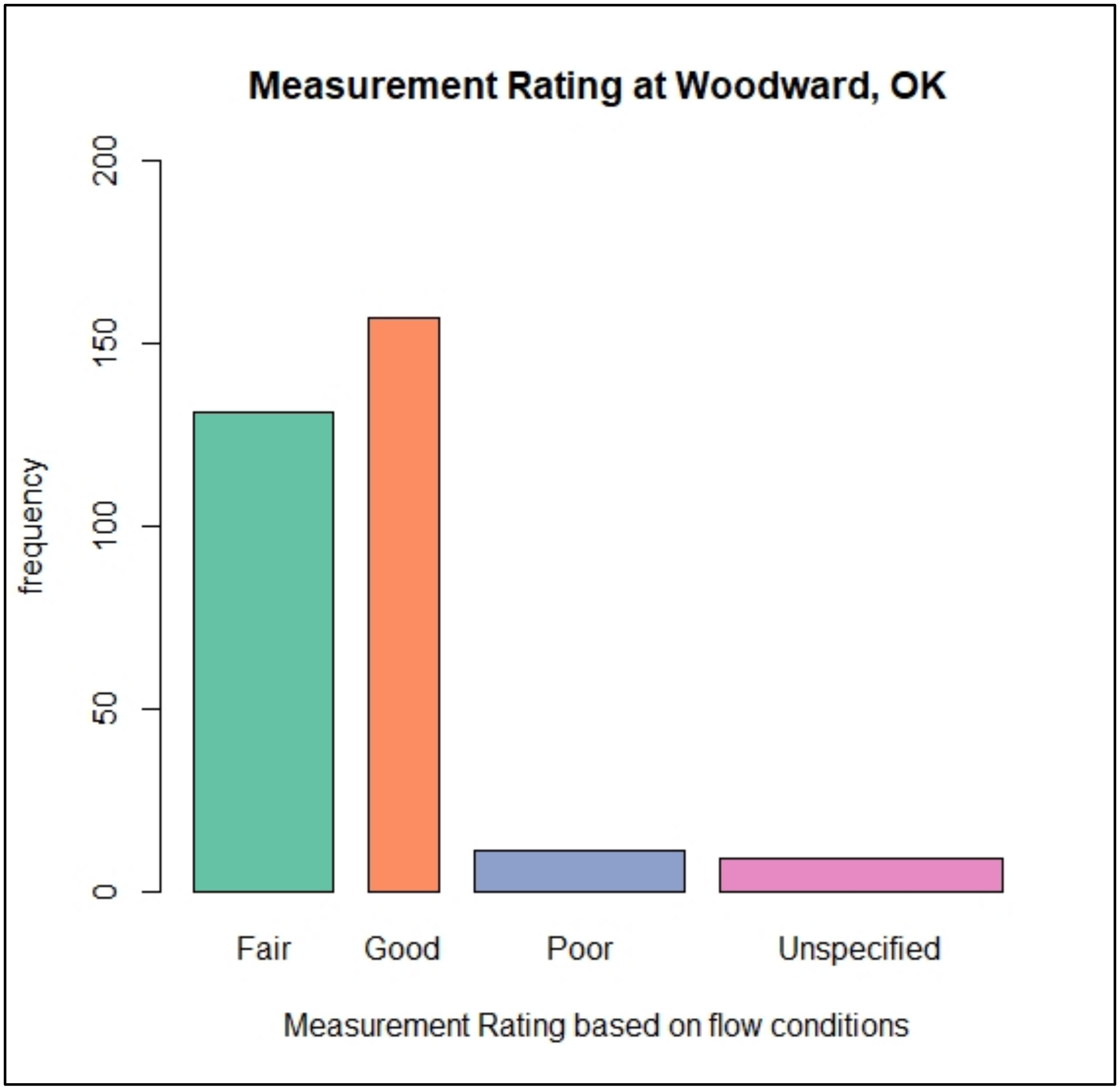
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.4500  0.8725  1.2700  1.3162  1.6875  4.4800      18
```



```
   vars   n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 290 1.32 0.52   1.27    1.29 0.59 0.45 4.48  4.03 1.01     3.48 0.03
```

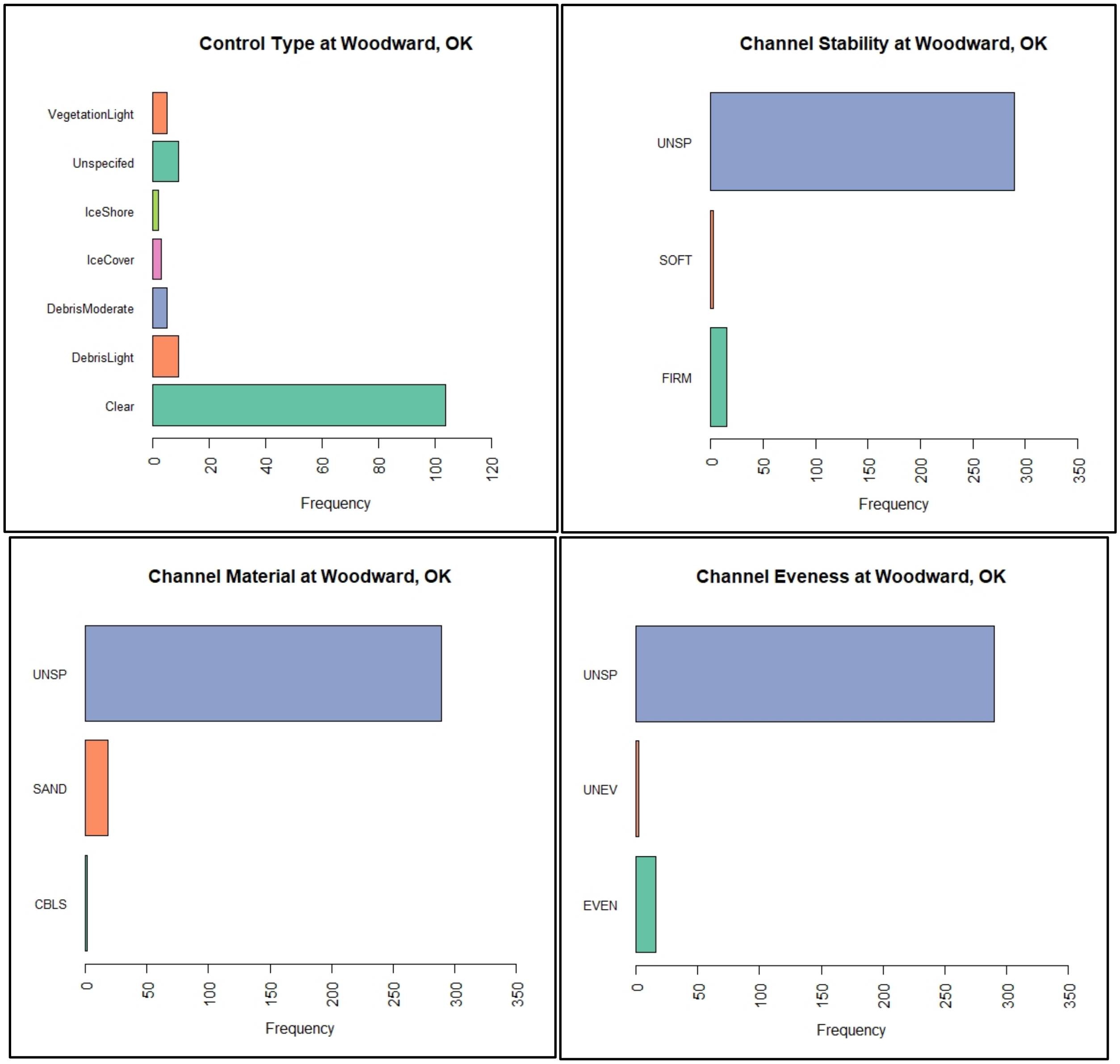
In the figure above you will see that the distribution is approximately normal. For understanding this distribution we can use the mean and standard deviation in the heavier outlined box above, for the central tendency and variability. Furthermore, we can also look at the skewness number which is approximately one. When examining the skewness number, usually you want to see a number between 0 and 1 which would indicate that your data is normally distributed. For the purpose of this study, we will go ahead and assume that our data is normally distributed above. The boxplot also confirms a normal distribution as there is only one or so outliers showing for this variable.

Figure 5. Barplot of Measurement Ratings for the N. Canadian River at Woodward, OK



In the figure above we can see that overall, of the 308 measurements taken from 1979-2020, the majority of the measurements were fair or good based on flow conditions. To me this would indicate that the data is representative and possibly help minimize or eliminate error.

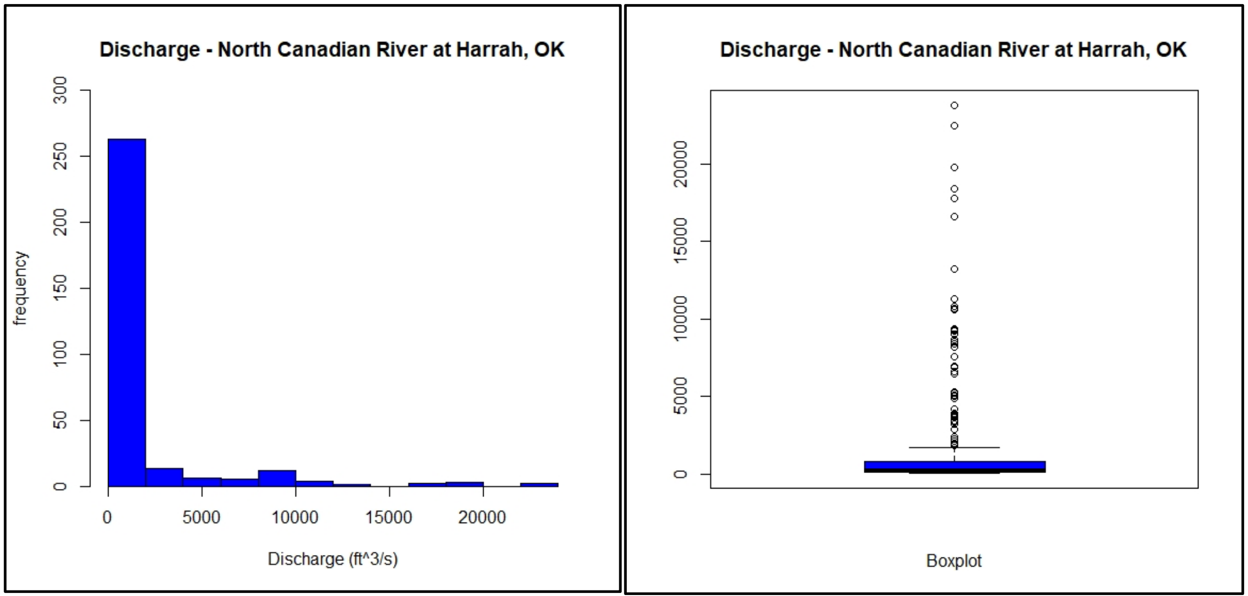
Figure 6. Barplots of Categorical Streamflow data for the N. Canadian River at Woodward, OK

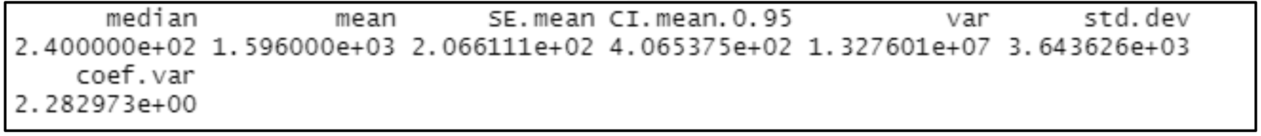


In the above figures, we have a look at the control type variable representing the condition of the rating control at the time of the measurement. Most of the control type data for the N. Canadian

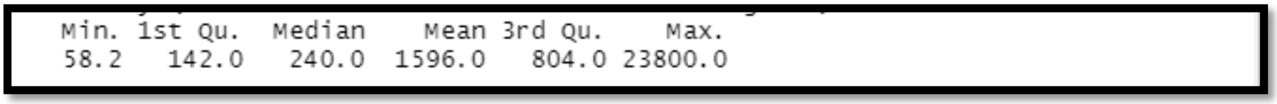
River at Woodward, OK ended up being clear or with nothing impeding the measurement. In terms of channel stability, which is the stability of the channel material, we can see that most of the measurements fell under the unspecified category, while a couple were either soft or firm. Additionally, channel material is pretty self-explanatory it describes the material in the channel. Lastly, channel evenness describes how level the channel is from bank to bank. We see that for both channel material and channel evenness, much of the data was described as unspecified, with only a few of the variables falling under the other specified categories.

Figure 7. Distribution of Discharge from the N. Canadian River at Harrah, OK

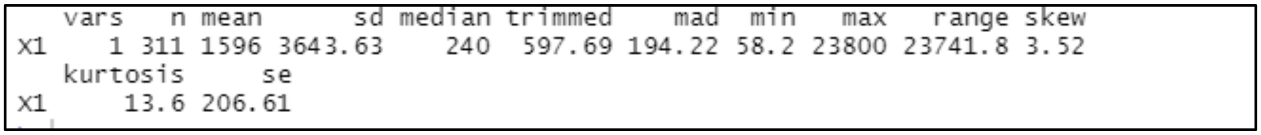




```
     median           mean       SE.mean CI.mean.0.95          var      std.dev
2.400000e+02 1.596000e+03 2.066111e+02 4.065375e+02 1.327601e+07 3.643626e+03
    coef.var
2.282973e+00
```



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   58.2   142.0   240.0  1596.0   804.0 23800.0
```



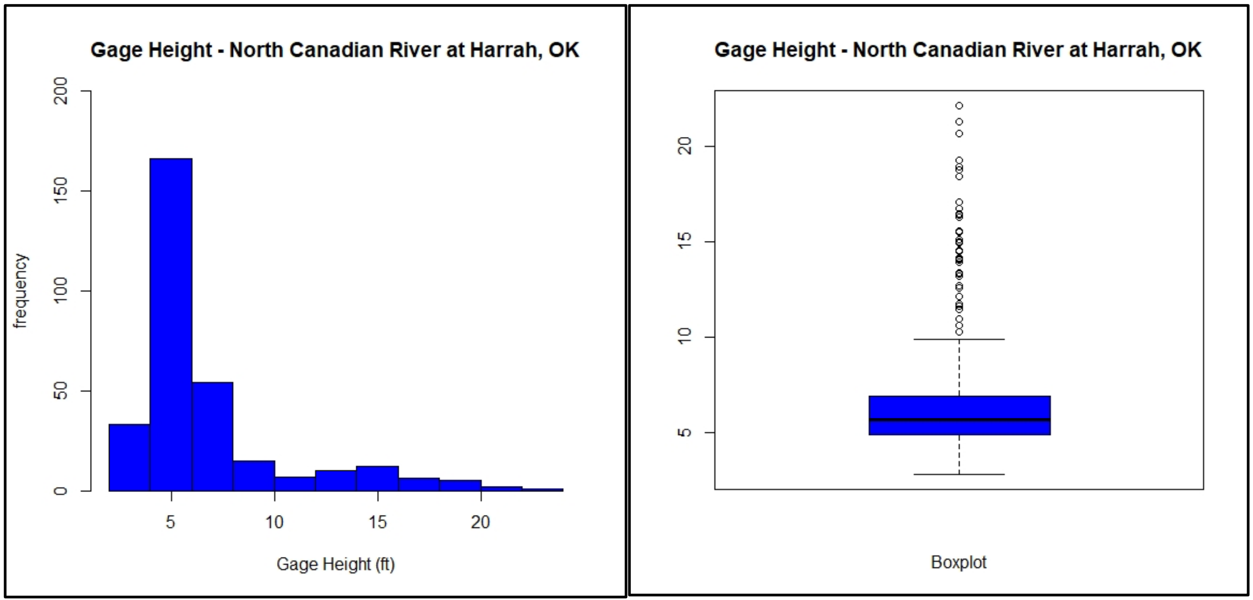
```
   vars   n mean      sd median trimmed    mad  min   max   range skew
X1    1 311 1596 3643.63    240  597.69 194.22 58.2 23800 23741.8 3.52
   kurtosis     se
X1     13.6 206.61
```

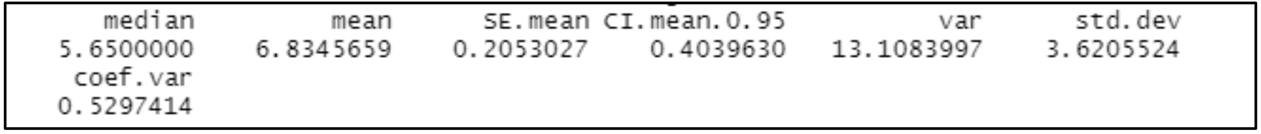
In the above figure you will notice the distribution is skewed right. To better understand the central tendency and variability of this data I had to use the five-number summary, which is expressed in the heavier outlined box above. Furthermore, we can express the variability by finding the interquartile range which is:

$$IQR = Q3 - Q1 = 804.0 - 142.0 = 662$$

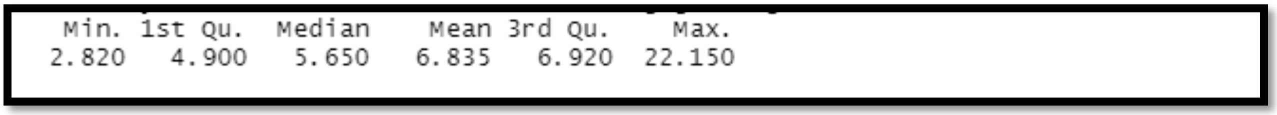
After calculating the interquartile range, we get 2.02 which explains the variance of this variable. While the IQR shows us how much of the data lies within a certain range, the boxplot indicates there are a handful of possible outliers present within the discharge dataset. However, it's hard to get an accurate representation due to dramatic fluctuations within the dataset and I don't necessarily think it should be thrown out.

Figure 8. Distribution of Gage Height from the N. Canadian River at Harrah, OK

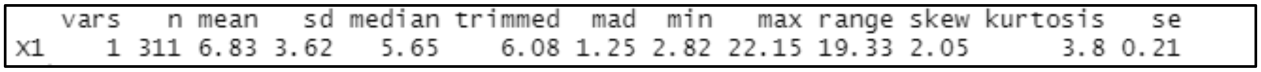




```
     median          mean      SE.mean CI.mean.0.95          var       std.dev
  5.6500000     6.8345659    0.2053027    0.4039630   13.1083997     3.6205524
   coef.var
  0.5297414
```



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.820   4.900   5.650   6.835   6.920  22.150
```



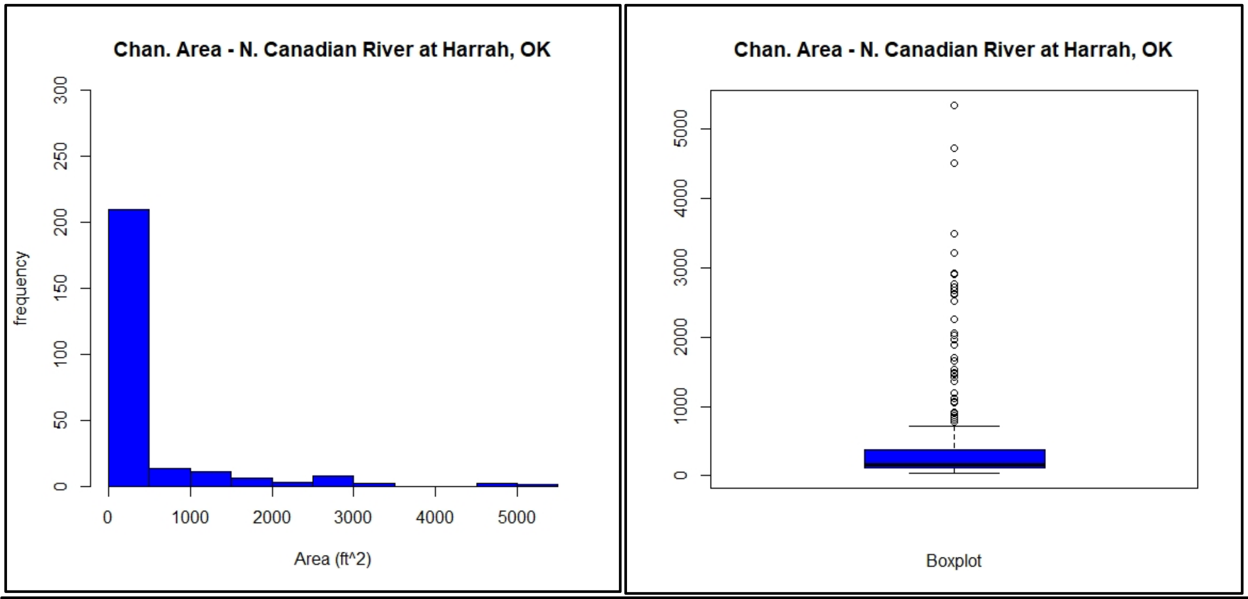
```
   vars   n mean   sd median trimmed  mad  min   max range skew kurtosis   se
X1    1 311 6.83 3.62   5.65    6.08 1.25 2.82 22.15 19.33 2.05      3.8 0.21
```

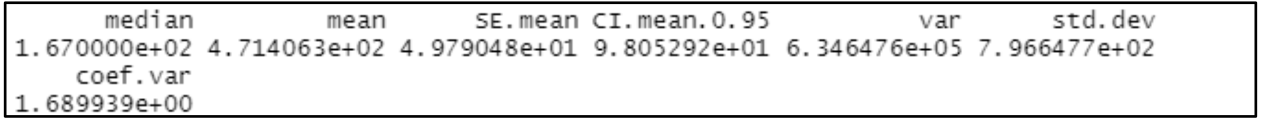
In this figure you will notice the distribution is also skewed right but not quite as extreme. Similar to analyzing the discharge data, to better understand the central tendency and variability of this data I had to use the five-number summary, which is expressed in the heavier outlined box above. Furthermore, we can express the variability by finding the interquartile range which is:

$$IQR = Q3 - Q1 = 6.920 - 4.900 = 2.02$$

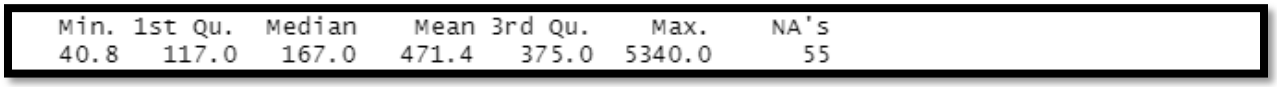
After calculating the interquartile range, we get 2.02 which explains the variance of this variable. While the IQR shows us how much of the data lies within a certain range, the boxplot indicates there are a handful of possible outliers present within the discharge dataset. However, it’s hard to get an accurate representation due to dramatic fluctuations within the dataset and I don’t necessarily think it should be thrown out.

Figure 9. Distribution of Channel Area from the N. Canadian River at Harrah, OK

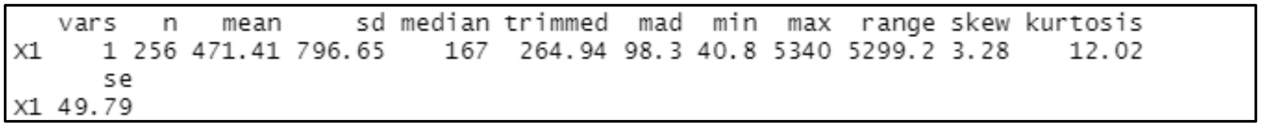




```
      median          mean       SE.mean CI.mean.0.95           var       std.dev
1.670000e+02 4.714063e+02 4.979048e+01 9.805292e+01 6.346476e+05 7.966477e+02
    coef.var
1.689939e+00
```



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   40.8   117.0   167.0   471.4   375.0  5340.0      55
```



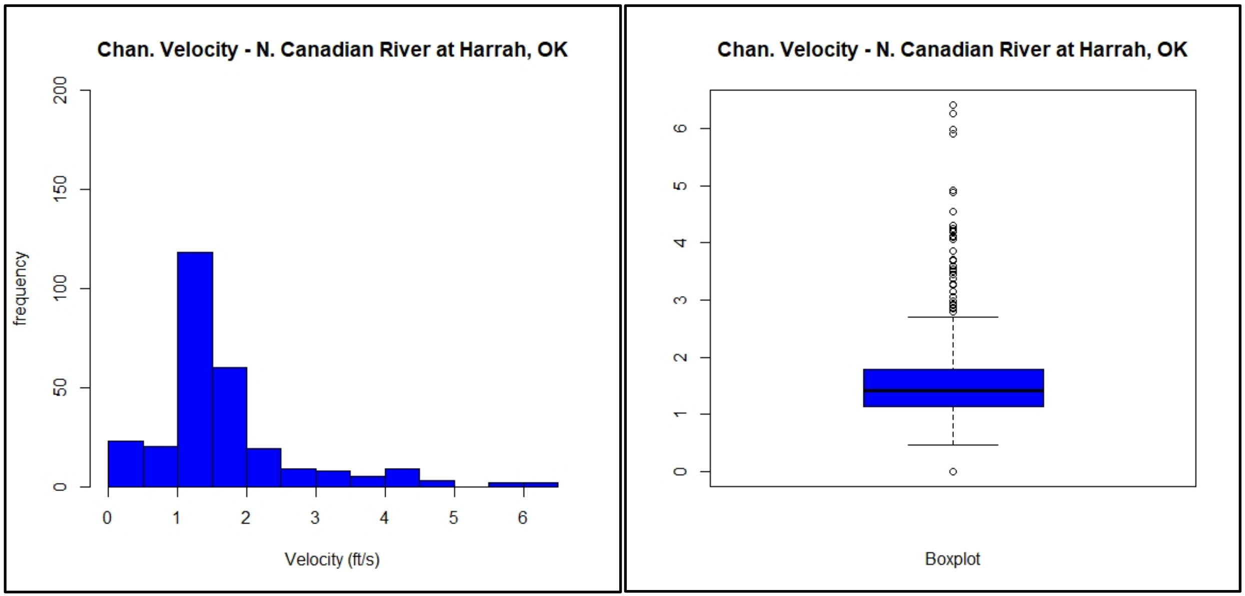
```
   vars   n   mean     sd median trimmed  mad  min  max  range skew kurtosis
X1    1 256 471.41 796.65    167  264.94 98.3 40.8 5340 5299.2 3.28    12.02
      se
X1 49.79
```

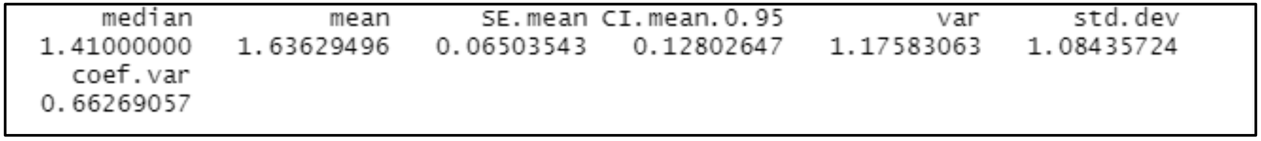
In the figure above you will see that the distribution is skewed right. Using the five-number summary helps us better understand the central tendency and variability of this data, which is expressed in the heavier outlined box above. Furthermore, we can express the variability by finding the interquartile range which is:

$$IQR = Q3 - Q1 = 375.0 - 117.0 = 258$$

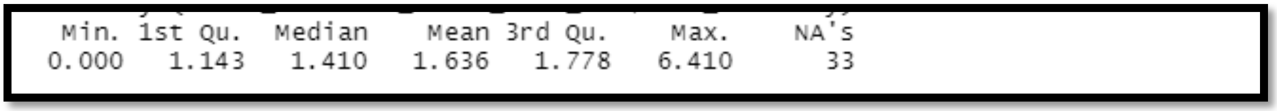
After calculating the interquartile range, we get 258 which explains the variance of this variable. While the IQR shows us how much of the data lies within a certain range, the boxplot indicates there are a handful of possible outliers present within the discharge dataset. However, it is hard to get an accurate representation due to dramatic fluctuations within the dataset and I don't necessarily think it should be thrown out.

Figure 9. Distribution of Channel Velocity from the N. Canadian River at Harrah, OK

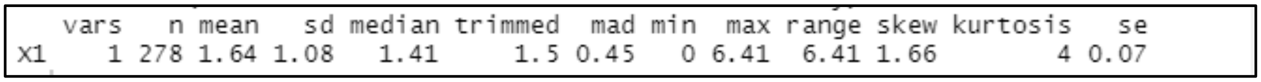




```
     median         mean      SE.mean CI.mean.0.95          var      std.dev
 1.41000000   1.63629496   0.06503543   0.12802647   1.17583063   1.08435724
   coef.var
 0.66269057
```



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  0.000   1.143   1.410   1.636   1.778   6.410      33 
```



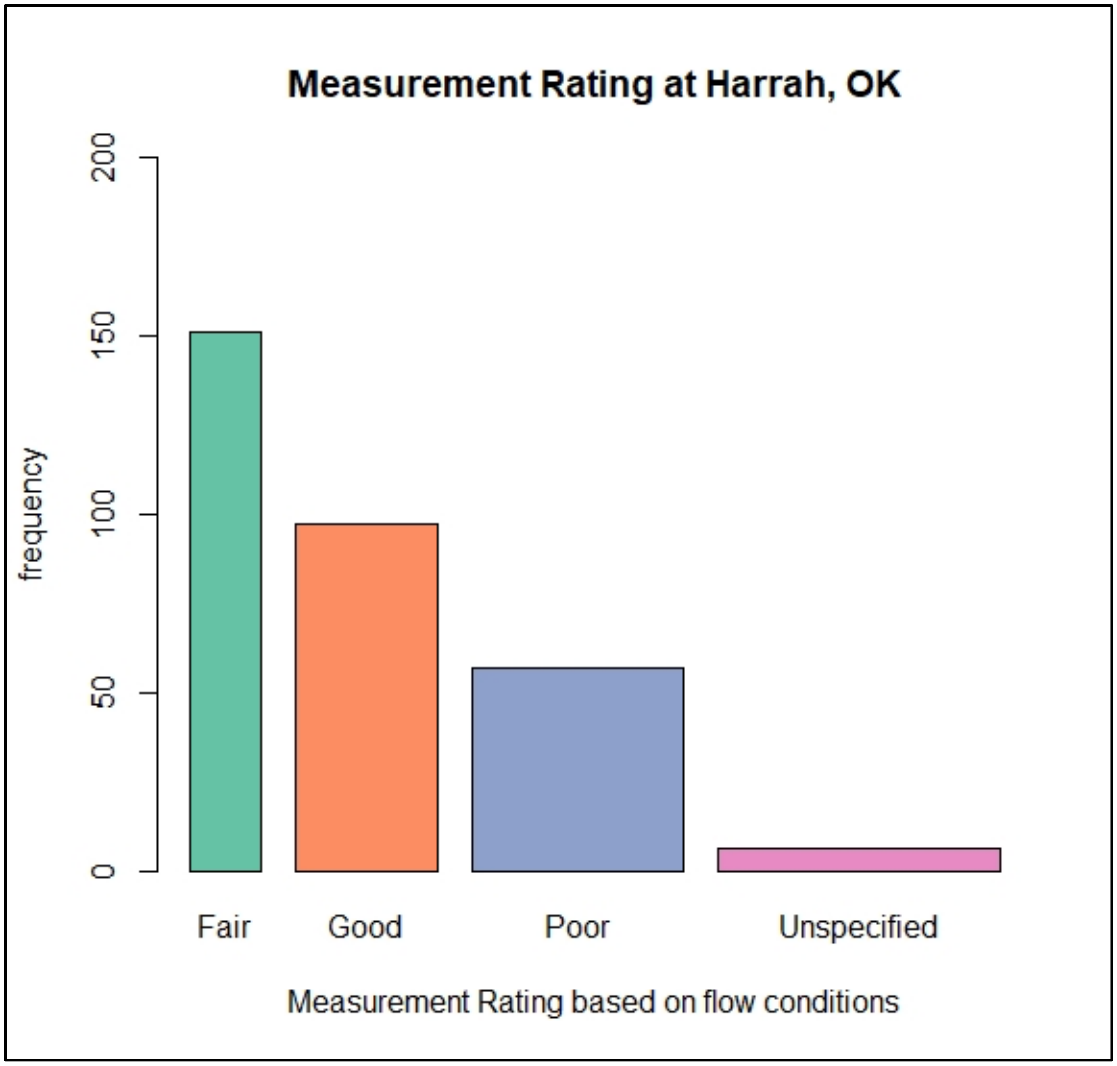
```
   vars   n mean   sd median trimmed  mad min  max range skew kurtosis   se
X1    1 278 1.64 1.08   1.41     1.5 0.45   0 6.41  6.41 1.66        4 0.07
```

In the figure above you will see that the distribution is skewed right. Using the five-number summary helps us better understand the central tendency and variability of this data, which is expressed in the heavier outlined box above. Furthermore, we can express the variability by finding the interquartile range which is:

$$IQR = Q3 - Q1 = 1.778 - 1.143 = .635$$

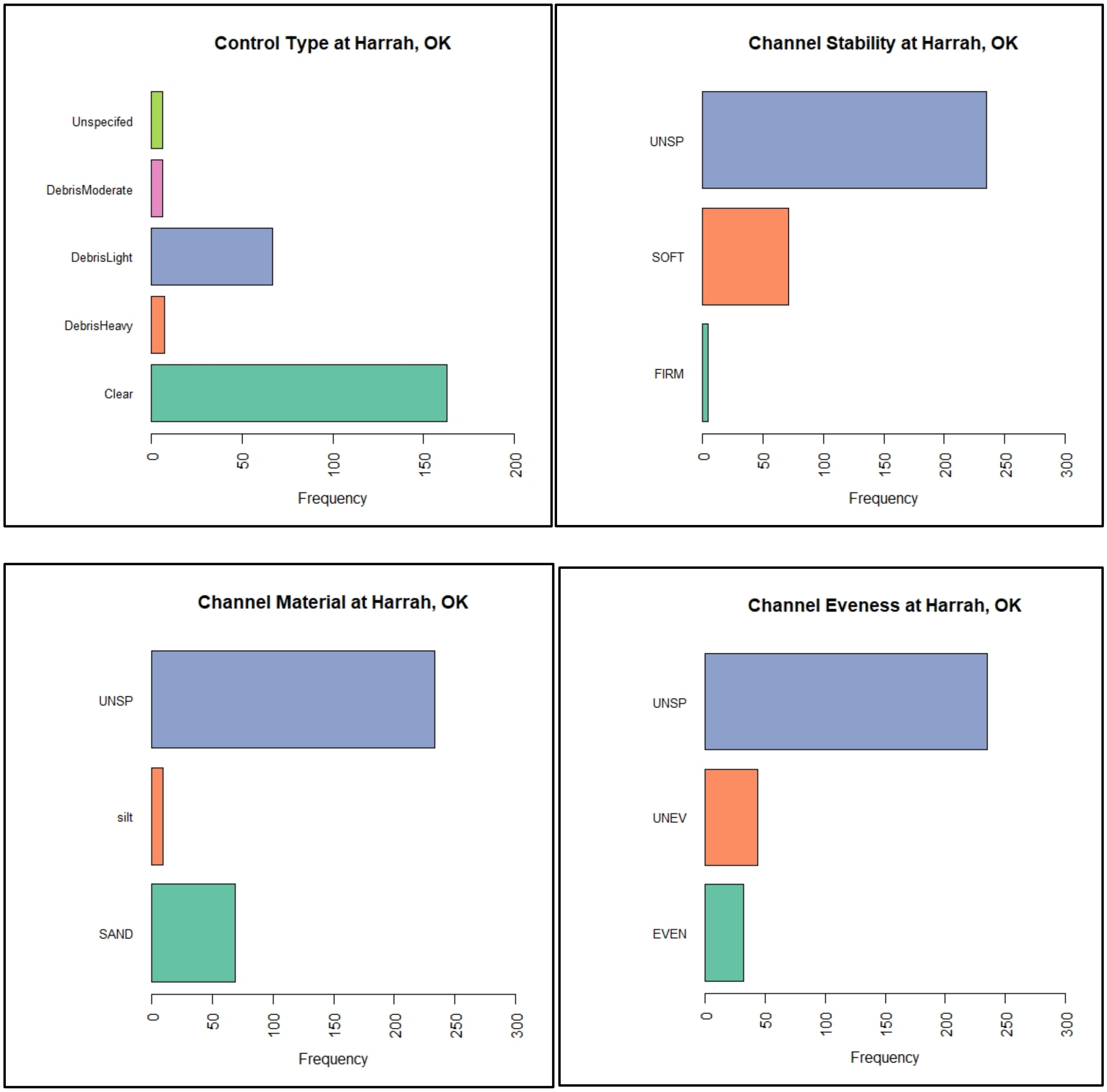
After calculating the interquartile range, we get .635which explains the variance of this variable. While the IQR shows us how much of the data lies within a certain range, the boxplot indicates there are a handful of possible outliers present within the discharge dataset. However, it is hard to get an accurate representation due to dramatic fluctuations within the dataset and I don’t necessarily think it should be thrown out.

Figure 10. Barplot of Measurement Ratings for the N. Canadian River at Harrah, OK



In the figure above we can see that overall, of the 311 measurements taken from 1980-2020, the majority of the measurements were fair or good based on flow conditions. To me this would indicate that the data is representative and possibly helping minimize or eliminate some of the error.

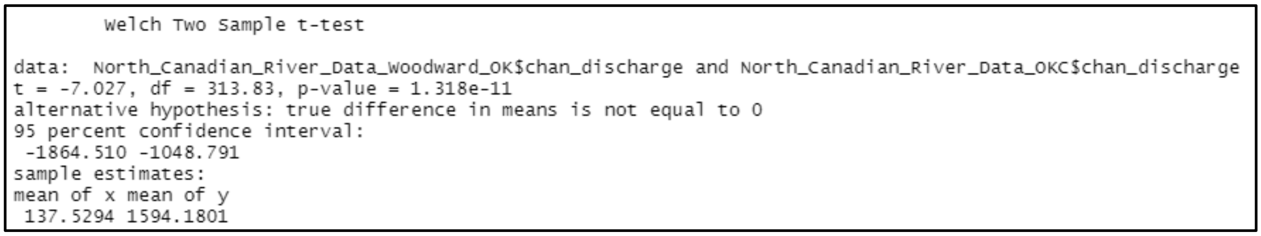
Figure 11. Barplots of Categorical Streamflow data for the N. Canadian River at Harrah, OK



In the above figures, we have a look at the control type variable representing the condition of the rating control at the time of the measurement. Most of the control type data for the N. Canadian River at Harrah, OK ended up being rather clear or with nothing impeding the measurement. In terms of channel stability, we can see that most of the measurements fell under the unspecified category, while the soft category has the second most data points. We see that for both channel material and channel evenness, much of the data was described as unspecified, with only a few of the variables falling under the other specified categories.

After visualizing the data, it helps stablish a deep understanding of the nature of the data and can now start performing proper statistical tests to then be able to interpret the data. For the purpose of this paper, I'm going to assume that most of my data is normally distributed, allowing me to run tests such as One sample T-Tests and Independent 2-Group T-Tests to compare the means of streamflow data across both locations. First, I looked at an Independent 2-Group T-Test which tests the mean of streamflow from the N. Canadian River at Woodward, OK to the mean of the of the N. Canadian River at Harrah, OK.

Figure 11. Independent 2-Group T-Test



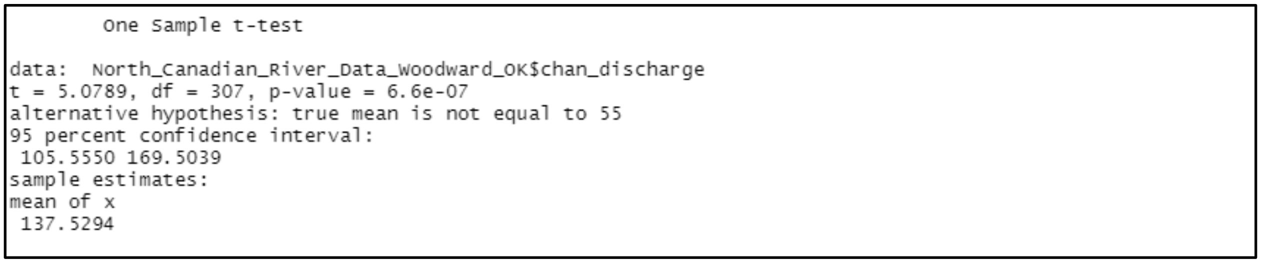
```
        Welch Two Sample t-test

data:  North_Canadian_River_Data_Woodward_OK$chan_discharge and North_Canadian_River_Data_OKC$chan_discharge
t = -7.027, df = 313.83, p-value = 1.318e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1864.510 -1048.791
sample estimates:
mean of x mean of y
 137.5294 1594.1801
```

Independent 2-Group T-Test Results
H_o: True difference in means is equal to 0.
H_a: True difference in means is not equal to 0.
Result: Reject H_0 so the true difference in means is not equal to 0.
Sig. value: 1.318e-11

In the above figure, you can see that the means of streamflow across both locations is statistically significantly different. We can state this because the p-value is very small, therefore we can reject the null hypothesis and accept the alternative hypothesis. This states that by comparing the means across both locations you can infer that the true difference is not equal to 0. We can also infer that the average streamflow is perhaps different over time across both locations.

In the below tests, I look to examine how the means of the measurements in both datasets, compare to the actual annual mean which was found on the USGS site. The annual mean for the Woodward, OK location streamflow is 55. For the Harrah, OK location stream flow it was 358.

Figure 12. One Sample T-Test on Streamflow at Woodward, OK

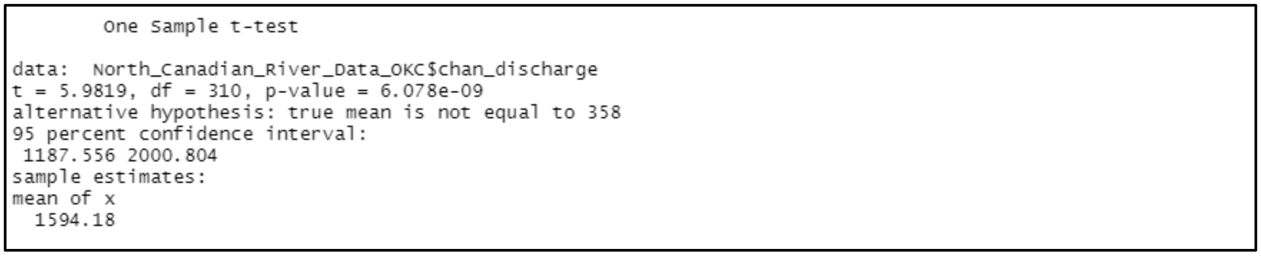


```
        One Sample t-test

data:  North_Canadian_River_Data_Woodward_OK$chan_discharge
t = 5.0789, df = 307, p-value = 6.6e-07
alternative hypothesis: true mean is not equal to 55
95 percent confidence interval:
 105.5550 169.5039
sample estimates:
mean of x
 137.5294
```

One Sample T-Test Results
H_o: μ = 55
H_a: μ ≠ 55
Result: Reject H_0, the true mean is not equal to the target mean of 55.
Sig. value: 6.6e-07

The goal for the One Sample T-Test above was to take the mean of streamflow across the appropriate time period for the N. Canadian River at Woodward, OK and compare it to the annual average streamflow over 42 years at this station location. We can see after we run that test the p-value is very small, so we can reject the null hypothesis and accept the alternative. We can infer that the true mean of the streamflow for the N. Canadian River at Woodward, OK is statistically significantly different.

Figure 13. One Sample T-Test on Streamflow at Harrah, OK

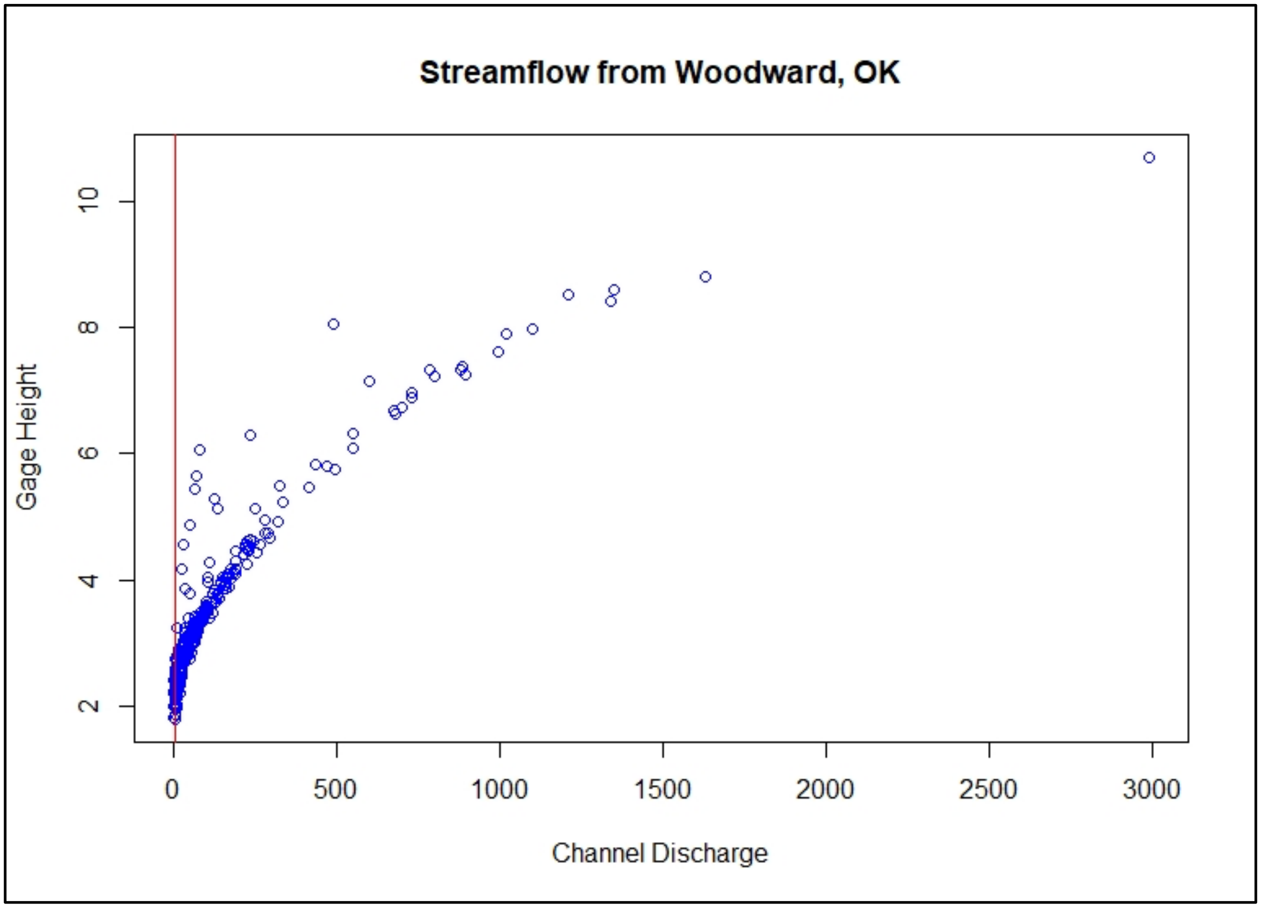


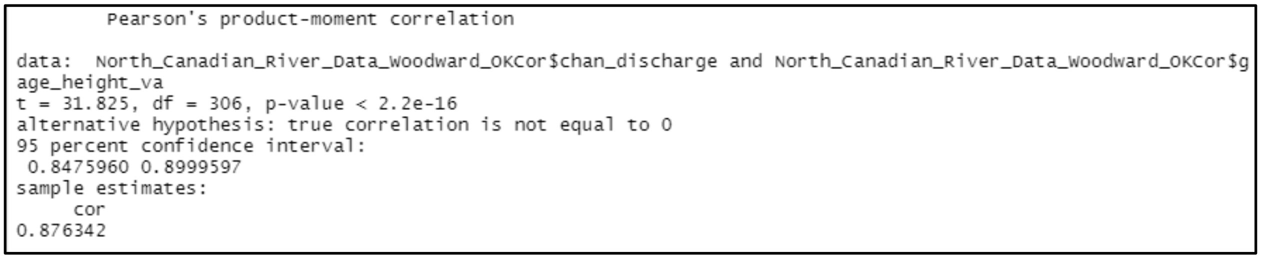
```
        One Sample t-test

data:  North_Canadian_River_Data_OKC$chan_discharge
t = 5.9819, df = 310, p-value = 6.078e-09
alternative hypothesis: true mean is not equal to 358
95 percent confidence interval:
 1187.556 2000.804
sample estimates:
mean of x
  1594.18
```

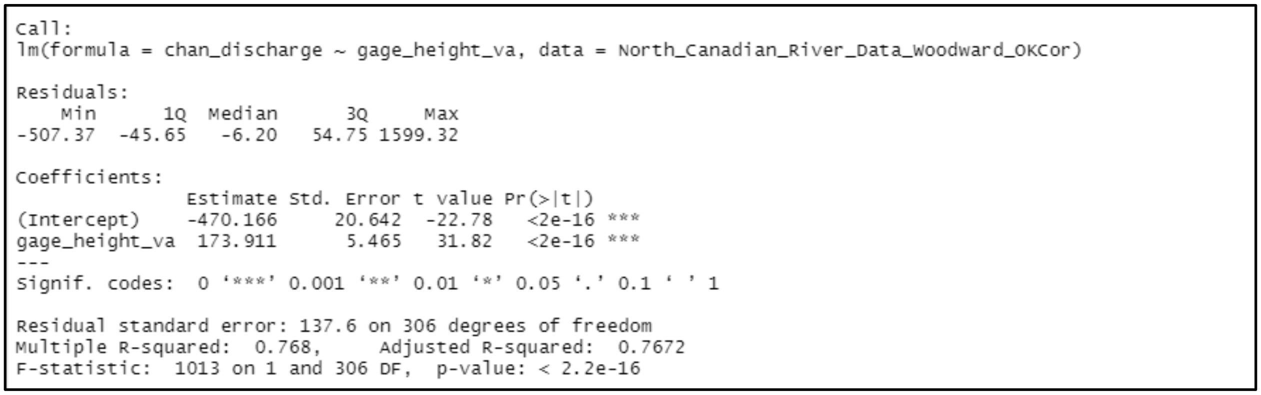
One Sample T-Test Results
H_o: μ = 358
H_a: μ ≠ 358
Result: Reject H_0, the true mean is not equal to the target mean of 358.
Sig. value: 6.078e-09

The goal for the One Sample T-Test above was to take the mean of streamflow across the appropriate time period for the N. Canadian River at Harrah, OK and compare it to the annual average streamflow over 41 years at this station location. We can see after we run that test the p-value is very small, so we can reject the null hypothesis and accept the alternative. We can infer that the true mean of the streamflow for the N. Canadian River at Harrah, OK is statistically significantly different.

Figure 14. Correlation on Streamflow at Woodward, OK



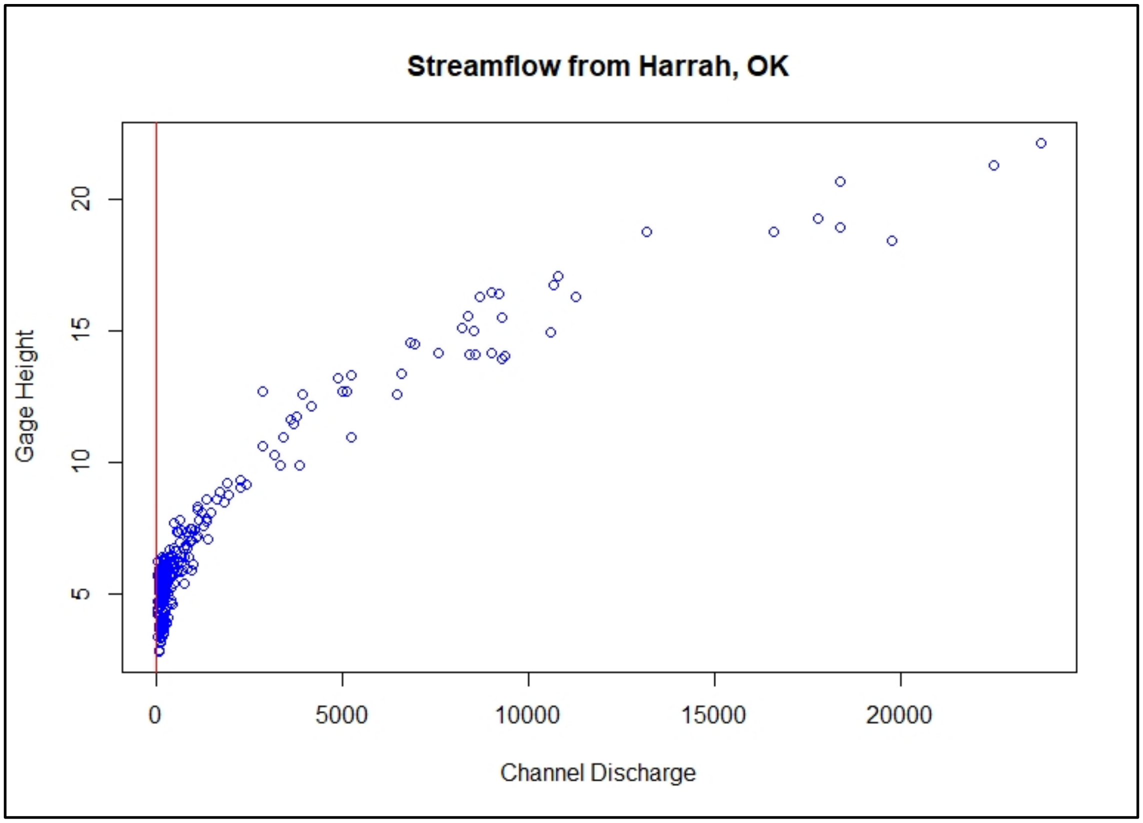

```
        Pearson's product-moment correlation

data:  North_Canadian_River_Data_Woodward_OKCor$chan_discharge and North_Canadian_River_Data_Woodward_OKCor$g
age_height_va
t = 31.825, df = 306, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8475960 0.8999597
sample estimates:
     cor
0.876342
```

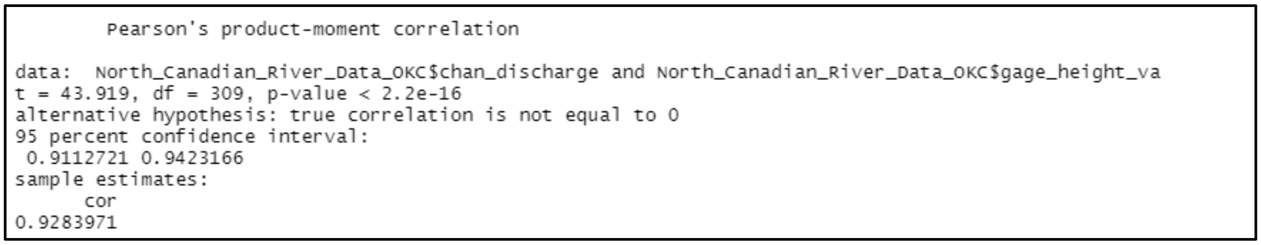

```
Call:
lm(formula = chan_discharge ~ gage_height_va, data = North_Canadian_River_Data_Woodward_OKCor)

Residuals:
    Min      1Q  Median      3Q     Max
-507.37  -45.65   -6.20   54.75 1599.32

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -470.166     20.642  -22.78   <2e-16 ***
gage_height_va  173.911      5.465   31.82   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 137.6 on 306 degrees of freedom
Multiple R-squared:  0.768,     Adjusted R-squared:  0.7672
F-statistic:  1013 on 1 and 306 DF,  p-value: < 2.2e-16
```

Correlation and Regression Results
H_o: The true correlation is equal to 0
H_a: The true correlation is not equal to 0
Result: Reject H_0, the true mean is not equal to 0.
Sig. value: 2.2e-16

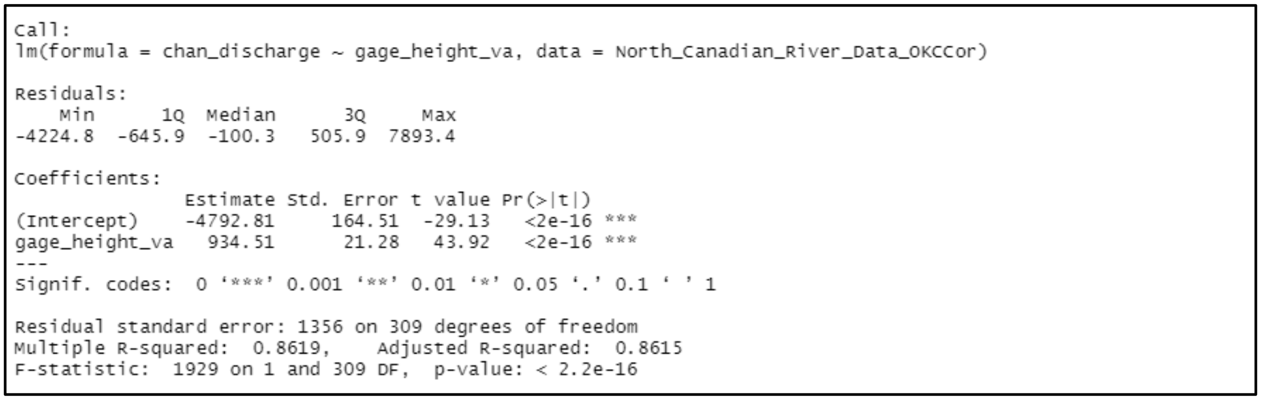
In the above figure we test the correlation between the streamflow (discharge) and the gage height of the N. Canadian River at Woodward, OK. If we look at the scatterplot, we can see that the correlation or relationship is fairly strong and positive. It is also a non-linear correlation. After running the correlation test, we see that the p-value is very small so we can reject the null hypothesis that the true correlation is equal to 0. Therefore, we accept the alternative hypothesis inferring that the true correlation is not equal to 0 or that the r value is statistically different from 0. Then we run the linear regression model for the channel discharge and gage height variables to see how much variability is present. The test yields a multiple r-squared value of 0.768 which helps us interpret approximately 77% of the variability from the linear regression line.

Figure 15. Correlation on Streamflow at Harrah, OK





```
        Pearson's product-moment correlation

data:  North_Canadian_River_Data_OKC$chan_discharge and North_Canadian_River_Data_OKC$gage_height_va
t = 43.919, df = 309, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9112721 0.9423166
sample estimates:
      cor
0.9283971
```

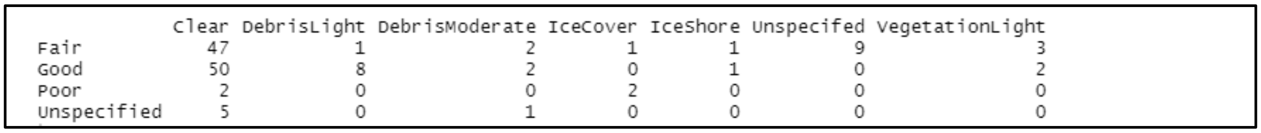


```
Call:
lm(formula = chan_discharge ~ gage_height_va, data = North_Canadian_River_Data_OKCCor)

Residuals:
    Min      1Q  Median      3Q     Max
-4224.8  -645.9  -100.3   505.9  7893.4

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4792.81     164.51  -29.13   <2e-16 ***
gage_height_va   934.51      21.28   43.92   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1356 on 309 degrees of freedom
Multiple R-squared:  0.8619,    Adjusted R-squared:  0.8615
F-statistic:  1929 on 1 and 309 DF,  p-value: < 2.2e-16
```

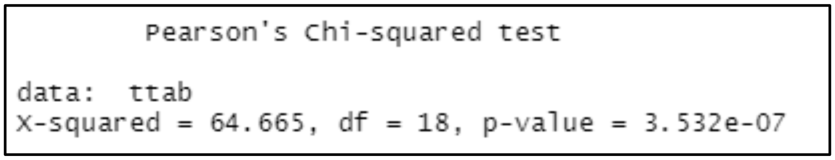
Correlation and Regression Results
H_o: The true correlation is equal to 0
H_a: The true correlation is not equal to 0
Result: Reject H_0, the true mean is not equal to 0.
Sig. value: 2.2e-16

In the above figure we test the correlation between the streamflow (discharge) and the gage height of the N. Canadian River at Harrah, OK. If we look at the scatterplot, we can see that the correlation or relationship is fairly strong and positive similar to that of the Woodward, OK correlation scatterplot. It is also a non-linear correlation. After running the correlation test, we see that the p-value is very small so we can reject the null hypothesis that the true correlation is equal to 0. Therefore, we accept the alternative hypothesis inferring that the true correlation is not equal to 0 or that the r value is statistically different from 0. Then we run the linear regression model for the channel discharge and gage height variables to see how much variability is present. The test yields a multiple r-squared value of 0.8619 which helps us interpret approximately 86% of the variability from the linear regression line.

Figure 16. Chi-Squared Tests on Categorical Streamflow data from the N. Canadian River at Woodward, OK



```
             Clear DebrisLight DebrisModerate IceCover IceShore Unspecifed VegetationLight
  Fair          47           1              2        1        1          9               3
  Good          50           8              2        0        1          0               2
  Poor           2           0              0        2        0          0               0
  Unspecified    5           0              1        0        0          0               0
```

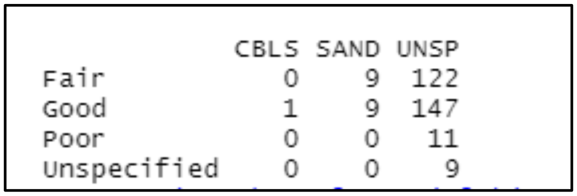


```
	Pearson's Chi-squared test

data:  ttab
X-squared = 64.665, df = 18, p-value = 3.532e-07
```

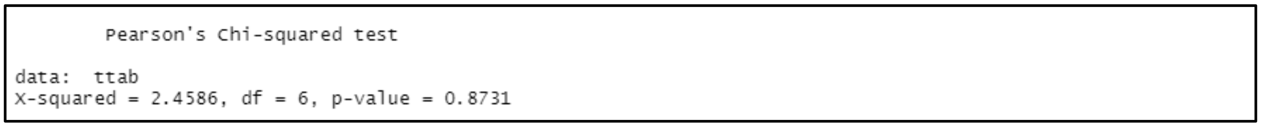
Chi-Squared Results
H_o: There is no relationship between measurement rating and control type.
H_a: There is a relationship between measurement rating and control type.
Result: Reject H_0, there is a relationship between measurement rating and control type.
Sig. value: 3.532e-07

I took the categorical or qualitative data from the Woodward, OK dataset and produced a contingency table of the measurement rating data and the control type data. Then I used a Chi-Squared test to determine if there is any relationship between the two variables. After performing the test, we get a p-value of 3.532e-07 which is very small. We can therefore, reject the null hypothesis and accept the alternative hypothesis stating that there is a statistically significant relationship between these two variables.

Figure 17. Chi-Squared Tests on Categorical Streamflow data from the N. Canadian River at Woodward, OK



```
             CBLS SAND UNSP
Fair            0    9  122
Good            1    9  147
Poor            0    0   11
Unspecified     0    0    9
```

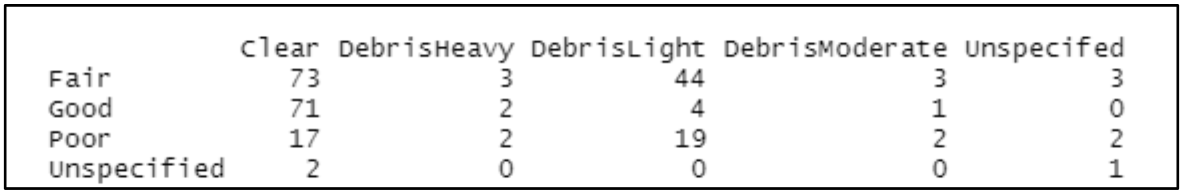


```
        Pearson's Chi-squared test

data:  ttab
X-squared = 2.4586, df = 6, p-value = 0.8731
```

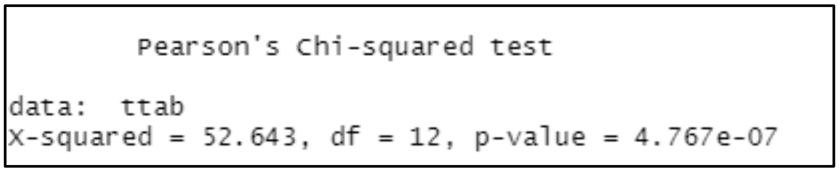
Chi-Squared Results
H_o: There is no relationship between measurement rating and channel material.
H_a: There is a relationship between measurement rating and channel material.
Result: Reject H_a, there is now relationship between measurement rating and channel material.
Sig. value: 0.8731

This time I wanted to look to see if there happened to be any relationship between measurement rating and channel material. Next, I produced a contingency table of the measurement rating data and the channel material data. Then I used a Chi-Squared test to determine if there is any relationship between the two variables. After performing the test, we get a p-value of 0.8731 which isn't small. We can therefore, fail to reject the null hypothesis and reject the alternative hypothesis stating that there is not a statistically significant relationship between these two variables.

Figure 18. Chi-Squared Tests on Categorical Streamflow data from the N. Canadian River at Harrah, OK



```
            Clear DebrisHeavy DebrisLight DebrisModerate Unspecifed
Fair           73           3          44              3          3
Good           71           2           4              1          0
Poor           17           2          19              2          2
Unspecified     2           0           0              0          1
```

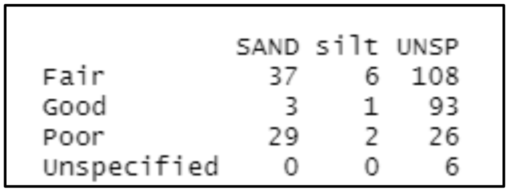


```
        Pearson's Chi-squared test

data:  ttab
X-squared = 52.643, df = 12, p-value = 4.767e-07
```

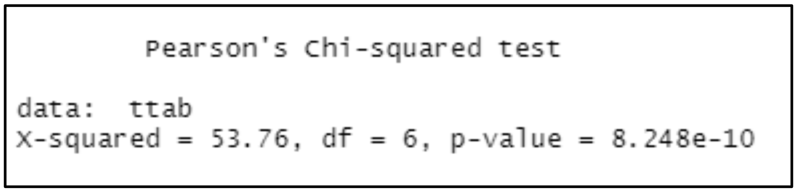
Chi-Squared Results
H_o: There is no relationship between measurement rating and control type.
H_a: There is a relationship between measurement rating and control type.
Result: Reject H_0, there is a relationship between measurement rating and control type.
Sig. value: 4.767e-07

I performed the same procedure for the categorical or qualitative data from the Harrah, OK dataset and produced a contingency table of the measurement rating data and the control type data. Then I used a Chi-Squared test to determine if there is any relationship between the two variables. After performing the test, we get a p-value of 4.767e-07which is very small. We can therefore, reject the null hypothesis and accept the alternative hypothesis stating that there is a statistically significant relationship between these two variables for the Harrah, OK location.

Figure 19. Chi-Squared Tests on Categorical Streamflow data from the N. Canadian River at Harrah, OK



```
            SAND silt UNSP
Fair          37    6  108
Good           3    1   93
Poor          29    2   26
Unspecified    0    0    6
```

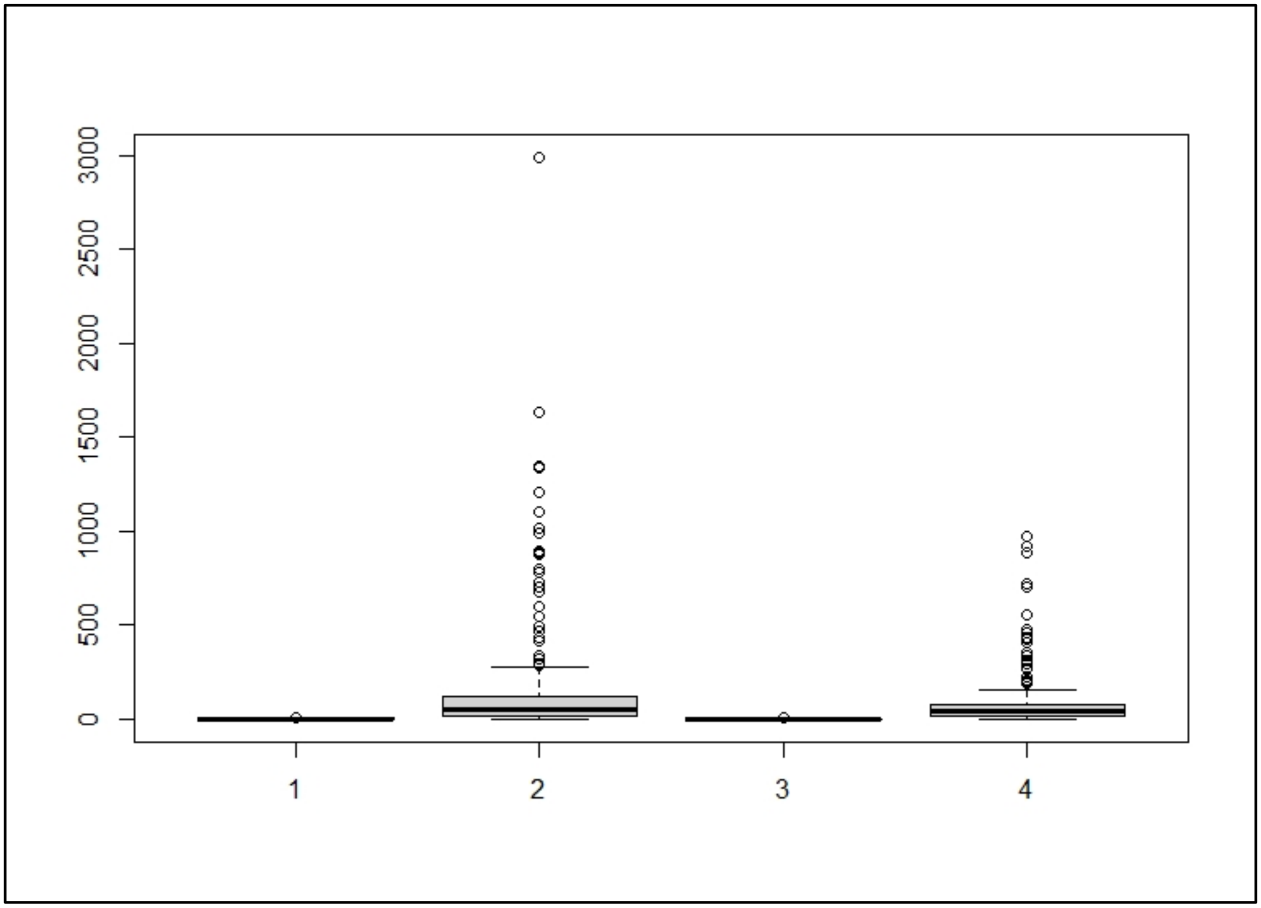


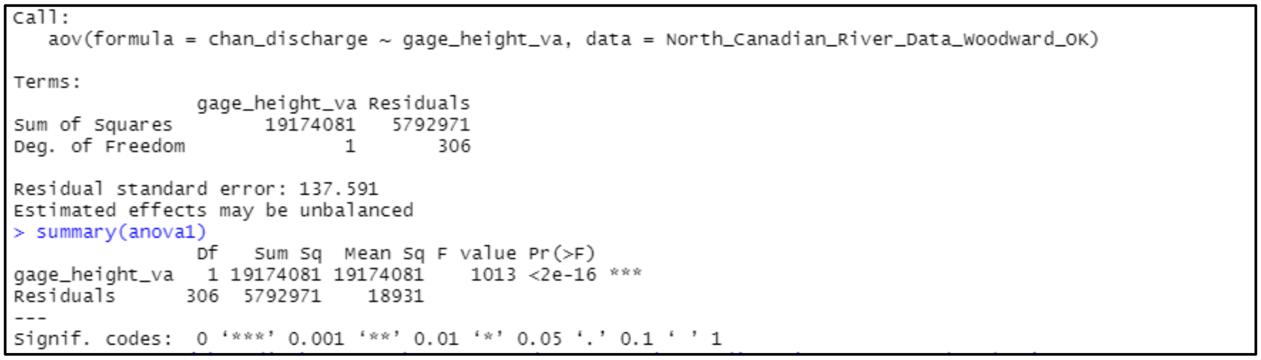
```
        Pearson's Chi-squared test

data:  ttab
X-squared = 53.76, df = 6, p-value = 8.248e-10
```

Chi-Squared Results
H_o: There is no relationship between measurement rating and channel material.
H_a: There is a relationship between measurement rating and channel material.
Result: Reject H_a, there is now relationship between measurement rating and channel material.
Sig. value: 8.248e-10

Additionally, I wanted to look to see if there happened to be any relationship between measurement rating and channel material for Harrah, OK dataset. I produced a contingency table of the measurement rating data and the channel material data. Then I used a Chi-Squared test to determine if there is any relationship between the two variables. After performing the test, we get a p-value of 8.248e-10 which is very small. We can therefore, reject the null hypothesis and accept the alternative hypothesis stating that there is a statistically significant relationship between these two variables. This result differs from that of the Woodward, OK streamflow data set. This allows us to infer that the measurement rating for the Harrah, OK streamflow data is influenced by channel material, unlike the Woodward, OK streamflow measurement rating.

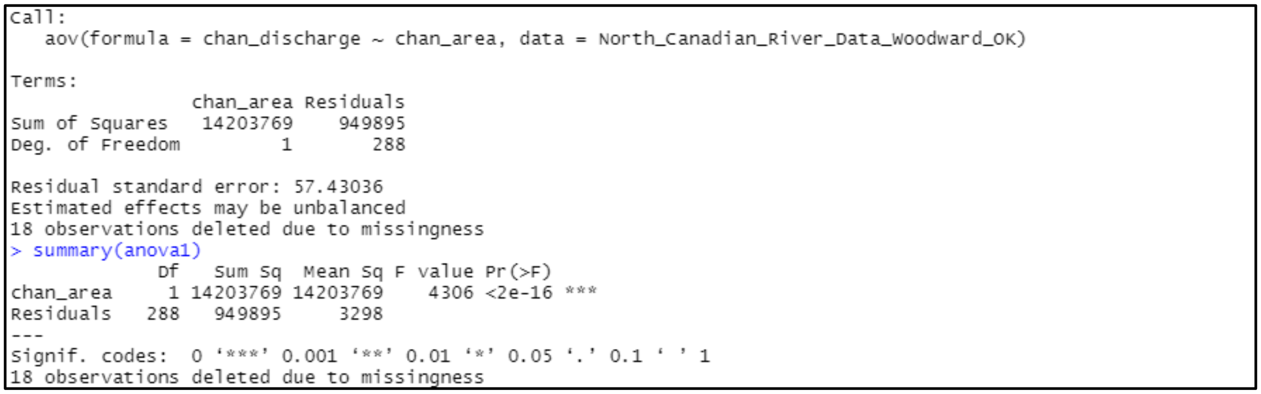
Figure 20. ANOVA tests for streamflow data at Woodward, OK





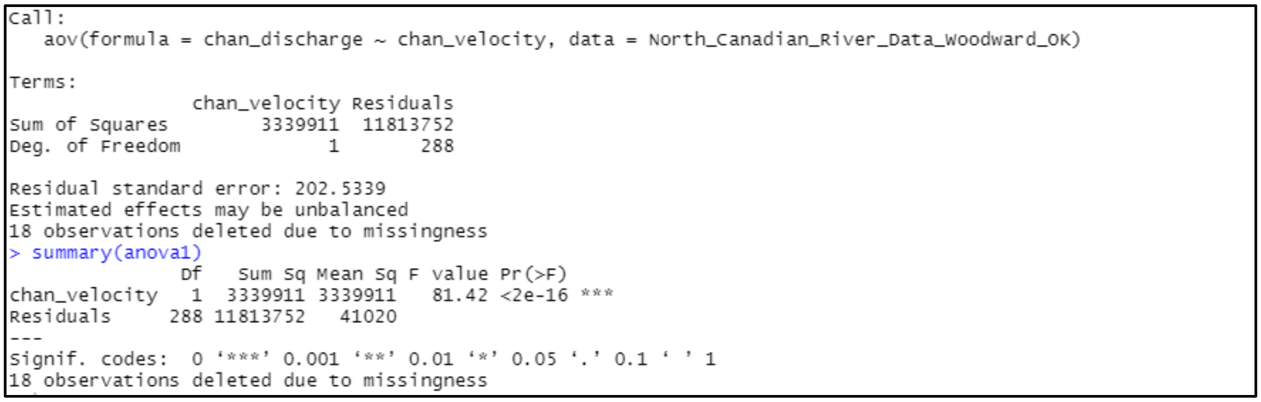
```
Call:
   aov(formula = chan_discharge ~ gage_height_va, data = North_Canadian_River_Data_Woodward_OK)

Terms:
                gage_height_va Residuals
Sum of Squares        19174081   5792971
Deg. of Freedom              1       306

Residual standard error: 137.591
Estimated effects may be unbalanced
> summary(anova1)
                Df   Sum Sq  Mean Sq F value Pr(>F)    
gage_height_va   1 19174081 19174081    1013 <2e-16 ***
Residuals      306  5792971    18931                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA Test Results
H_o: μ1 = μ2 = μ3 ….
H_a: At least one of the means is different from the others.
Result: Reject H_o, at least one of the means is different from others.
Sig. value: <2e-16



```
Call:
   aov(formula = chan_discharge ~ chan_area, data = North_Canadian_River_Data_Woodward_OK)

Terms:
                chan_area Residuals
Sum of Squares   14203769    949895
Deg. of Freedom         1       288

Residual standard error: 57.43036
Estimated effects may be unbalanced
18 observations deleted due to missingness
> summary(anova1)
             Df   Sum Sq  Mean Sq F value Pr(>F)    
chan_area     1 14203769 14203769    4306 <2e-16 ***
Residuals   288   949895     3298                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
18 observations deleted due to missingness
```

ANOVA Test Results
H_o: μ1 = μ2 = μ3 ….
H_a: At least one of the means is different from the others.
Result: Reject H_o, at least one of the means is different from others.
Sig. value: <2e-16

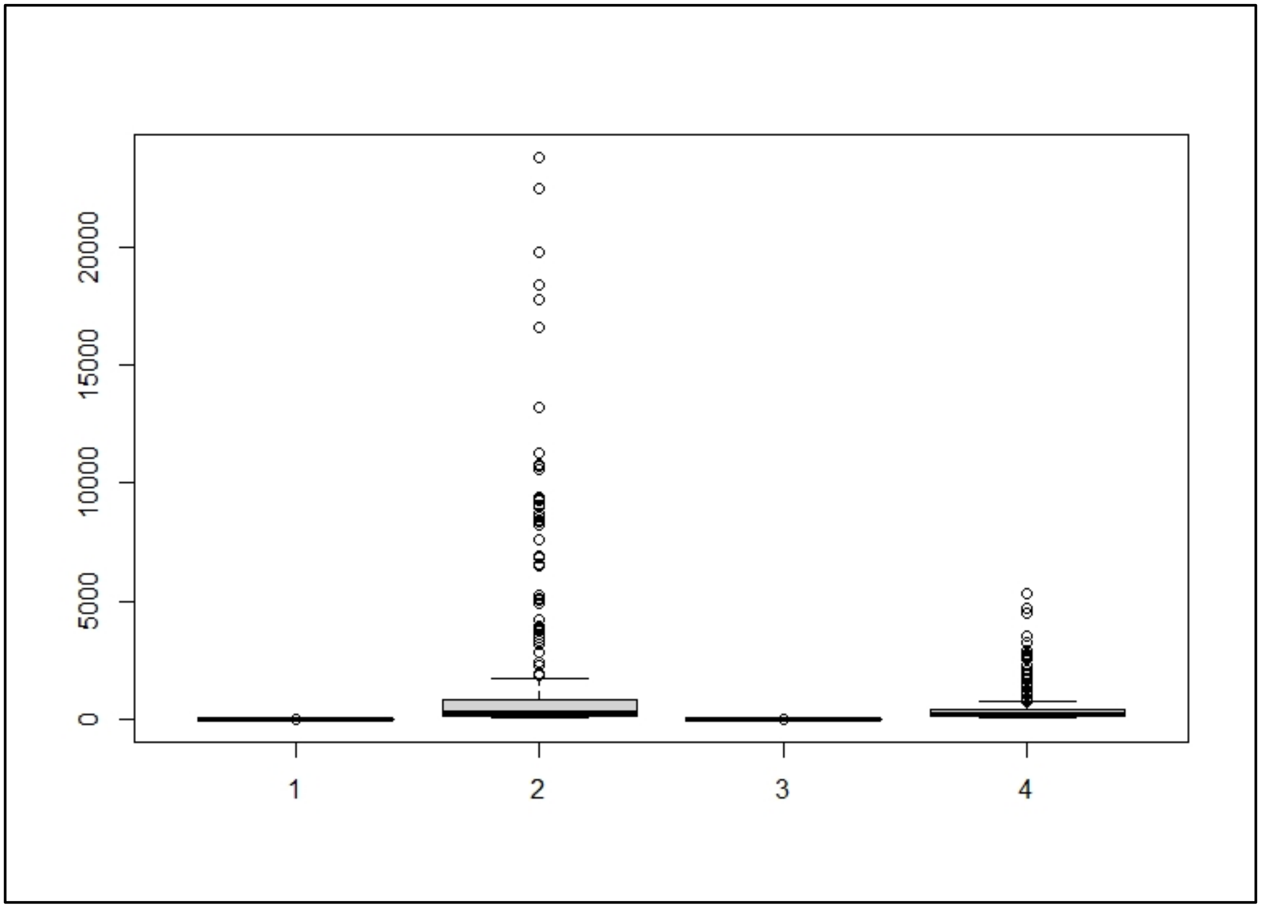


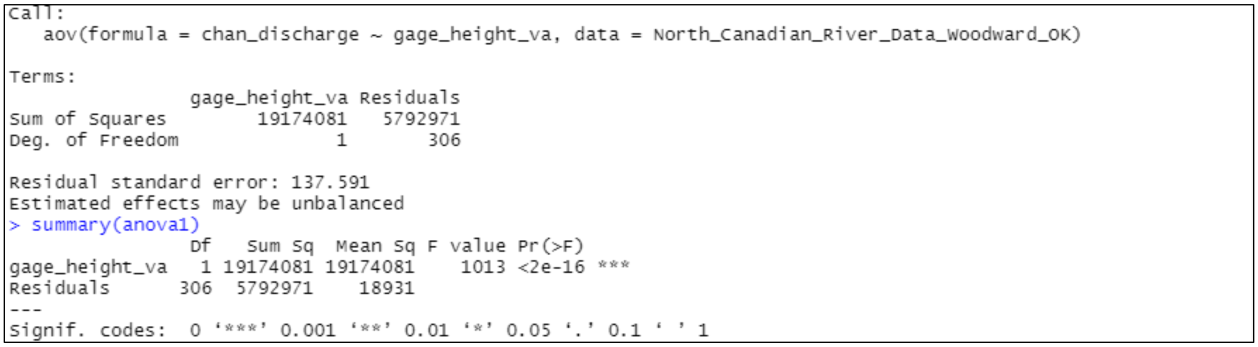
```
Call:
   aov(formula = chan_discharge ~ chan_velocity, data = North_Canadian_River_Data_Woodward_OK)

Terms:
                chan_velocity Residuals
Sum of Squares        3339911  11813752
Deg. of Freedom             1       288

Residual standard error: 202.5339
Estimated effects may be unbalanced
18 observations deleted due to missingness
> summary(anova1)
               Df   Sum Sq Mean Sq F value Pr(>F)    
chan_velocity   1  3339911 3339911   81.42 <2e-16 ***
Residuals     288 11813752   41020                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
18 observations deleted due to missingness
```

ANOVA Test Results
H_o: μ1 = μ2 = μ3 ….
H_a: At least one of the means is different from the others.
Result: Reject H_o, at least one of the means is different from others.
Sig. value: <2e-16

After looking at the statistical significance of each variables means, I wanted to look at how the means of certain variables vary from each other. First, we take a look at the boxplots for to understand the variability and central tendency. Then we run a One-Way ANOVA test to compare the means. It turns out that for the above tests the p-value is the same, which is very small, therefore we can reject the null hypotheses for all the above ANOVA tests. This allows us to accept the alternative hypothesis and infer that at least one of the means is different from the others, for all variables tested above.

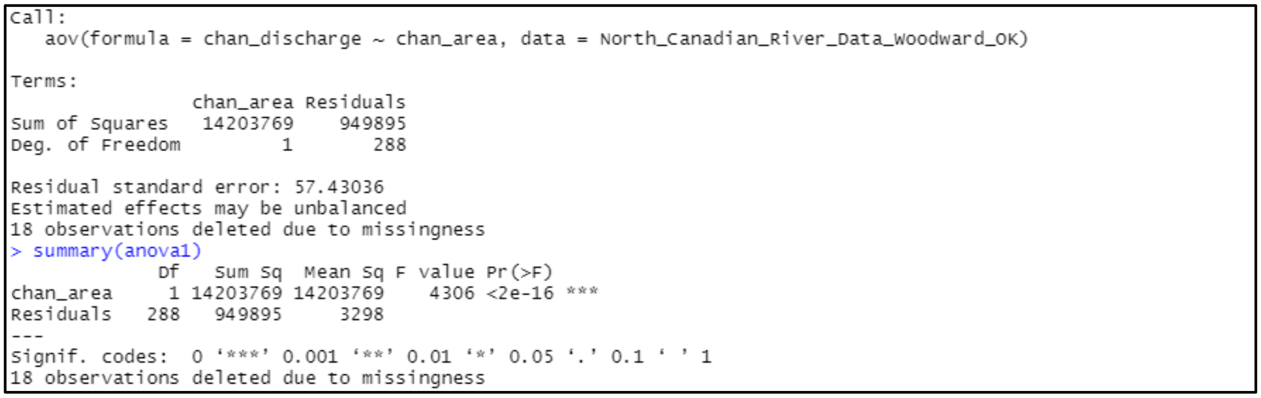
Figure 21. ANOVA tests for streamflow data at Harrah, OK





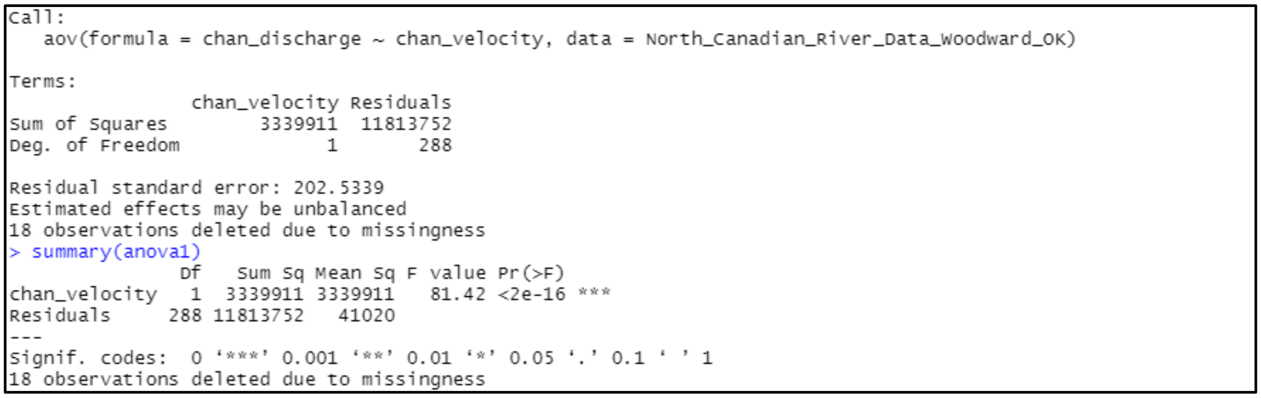
```
Call:
   aov(formula = chan_discharge ~ gage_height_va, data = North_Canadian_River_Data_Woodward_OK)

Terms:
                gage_height_va Residuals
Sum of Squares        19174081   5792971
Deg. of Freedom              1       306

Residual standard error: 137.591
Estimated effects may be unbalanced
> summary(anova1)
                Df   Sum Sq  Mean Sq F value Pr(>F)    
gage_height_va   1 19174081 19174081    1013 <2e-16 ***
Residuals      306  5792971    18931                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA Test Results
H_o: μ1 = μ2 = μ3 ….
H_a: At least one of the means is different from the others.
Result: Reject H_o, at least one of the means is different from others.
Sig. value: <2e-16



```
Call:
   aov(formula = chan_discharge ~ chan_area, data = North_Canadian_River_Data_Woodward_OK)

Terms:
                chan_area Residuals
Sum of Squares   14203769    949895
Deg. of Freedom         1       288

Residual standard error: 57.43036
Estimated effects may be unbalanced
18 observations deleted due to missingness
> summary(anova1)
             Df   Sum Sq  Mean Sq F value Pr(>F)    
chan_area     1 14203769 14203769    4306 <2e-16 ***
Residuals   288   949895     3298                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
18 observations deleted due to missingness
```

ANOVA Test Results
H_o: μ1 = μ2 = μ3 ….
H_a: At least one of the means is different from the others.
Result: Reject H_o, at least one of the means is different from others.
Sig. value: <2e-16

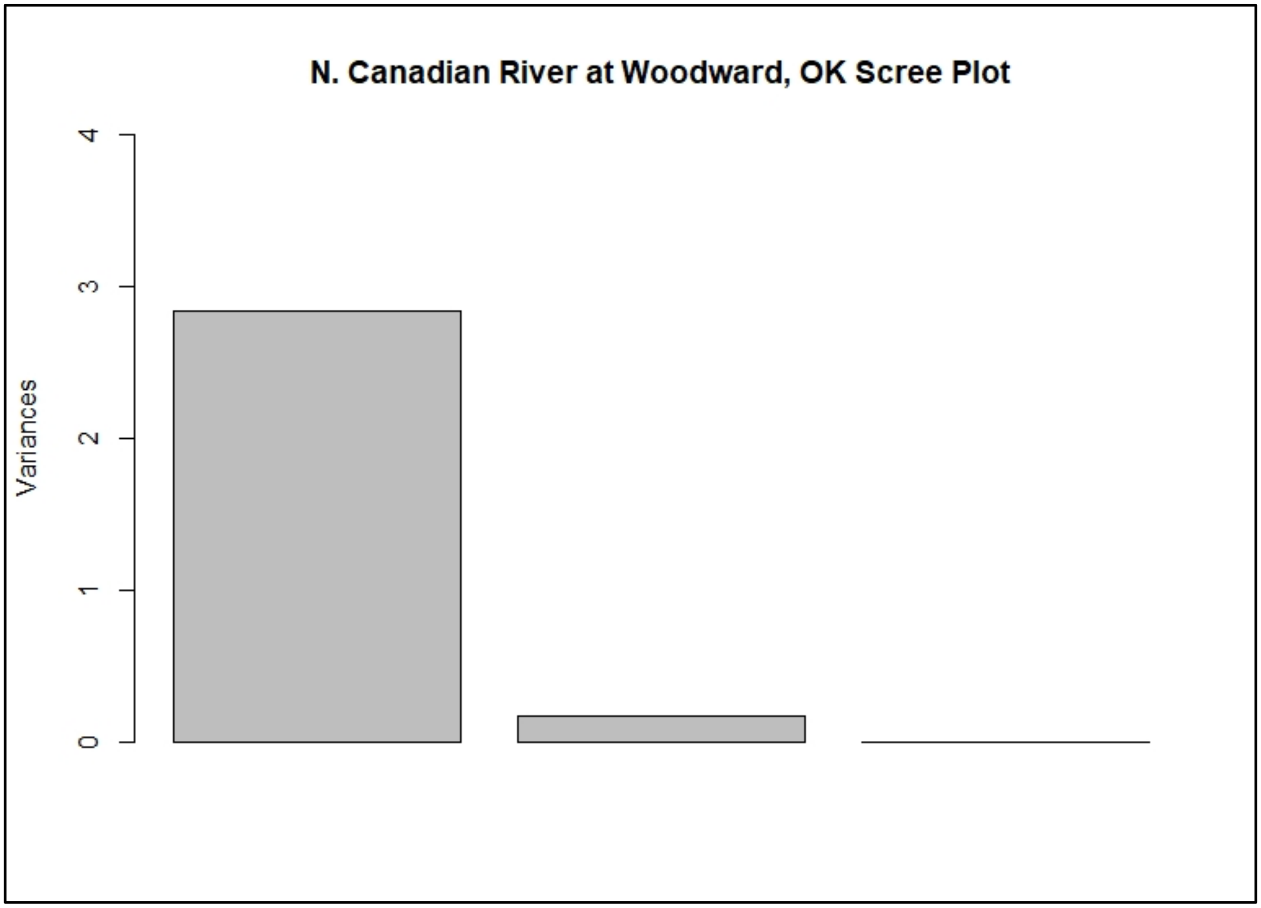


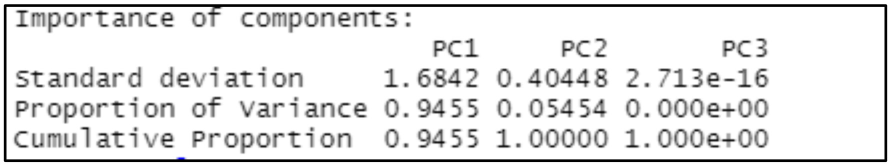
```
Call:
   aov(formula = chan_discharge ~ chan_velocity, data = North_Canadian_River_Data_Woodward_OK)

Terms:
                chan_velocity Residuals
Sum of Squares        3339911  11813752
Deg. of Freedom             1       288

Residual standard error: 202.5339
Estimated effects may be unbalanced
18 observations deleted due to missingness
> summary(anova1)
               Df   Sum Sq Mean Sq F value Pr(>F)
chan_velocity   1  3339911 3339911   81.42 <2e-16 ***
Residuals     288 11813752   41020
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
18 observations deleted due to missingness
```

ANOVA Test Results
H_o: $\mu1 = \mu2 = \mu3$ ….
H_a: At least one of the means is different from the others.
Result: Reject H_o, at least one of the means is different from others.
Sig. value: <2e-16

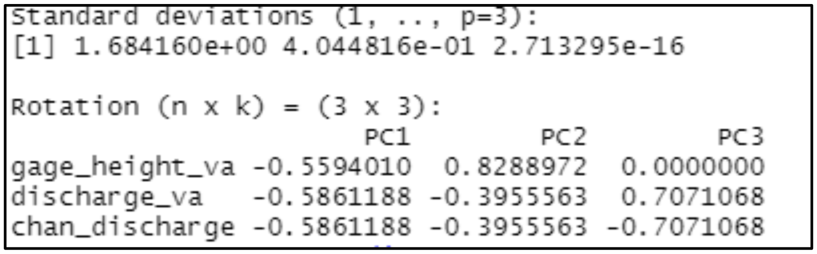
It turns out that the One-Way ANOVA tests for the same variables in the streamflow data from Harrah, OK yields the same result as the Woodward, OK streamflow data. The p-value is the same, which is very small, therefore we can reject the null hypotheses for all the above ANOVA tests. This allows us to accept the alternative hypothesis and infer that at least one of the means is different from the others, for all variables tested above.

Figure 22. Principle Components Analysis tests for streamflow data at Woodward, OK





```
Importance of components:
                          PC1     PC2       PC3
Standard deviation     1.6842 0.40448 2.713e-16
Proportion of Variance 0.9455 0.05454 0.000e+00
Cumulative Proportion  0.9455 1.00000 1.000e+00
```

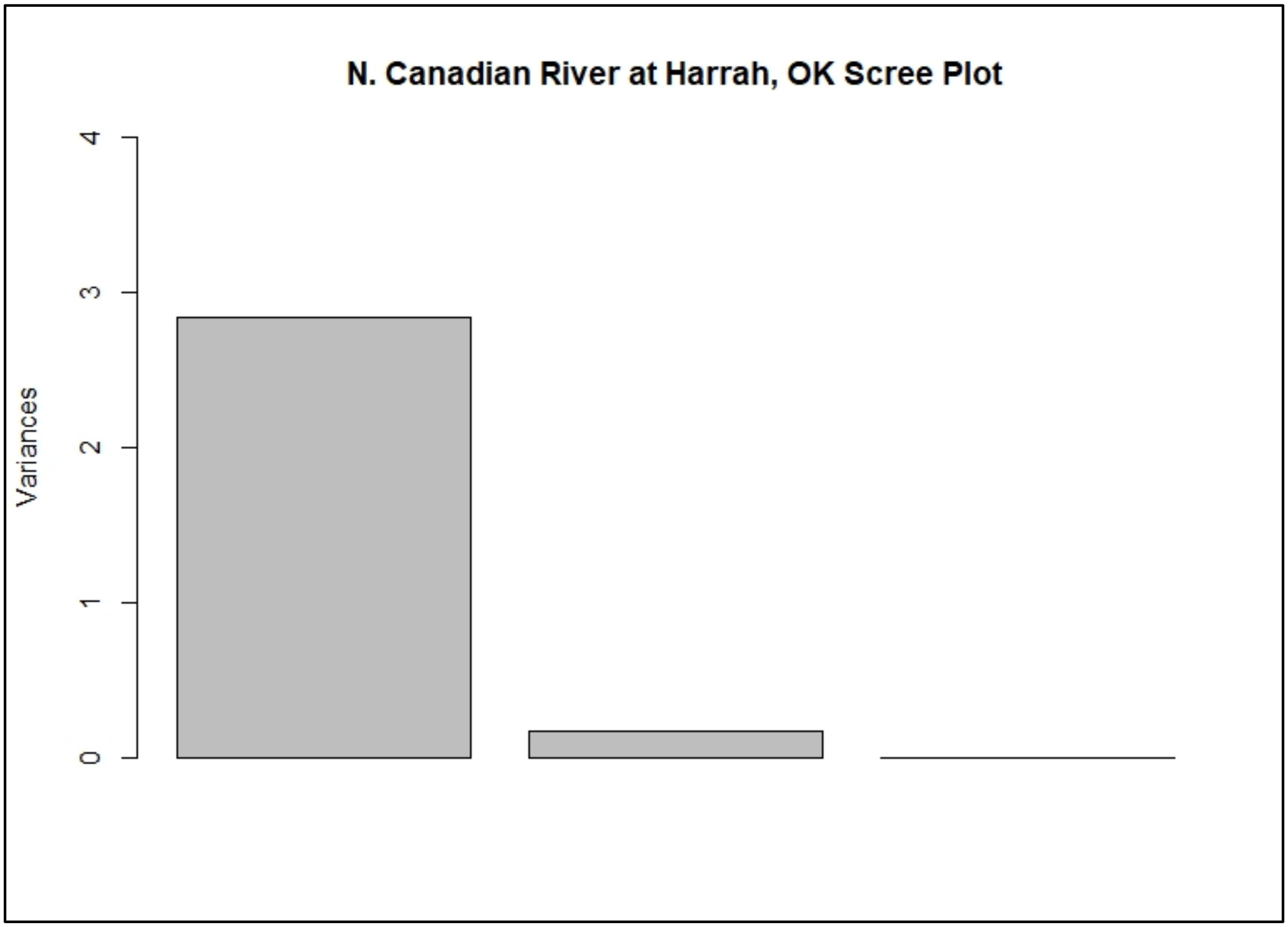


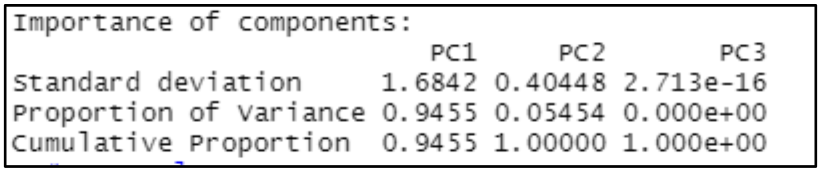
```
Standard deviations (1, .., p=3):
[1] 1.684160e+00 4.044816e-01 2.713295e-16

Rotation (n x k) = (3 x 3):
                     PC1        PC2        PC3
gage_height_va -0.5594010  0.8288972  0.0000000
discharge_va   -0.5861188 -0.3955563  0.7071068
chan_discharge -0.5861188 -0.3955563 -0.7071068
```

The last several tests I chose to run on the streamflow data is a Principle Components Analysis Test. Unfortunately given the nature of streamflow data, it can be very messy data to begin with. So I had to work around those issues in order to run the streamflow data from

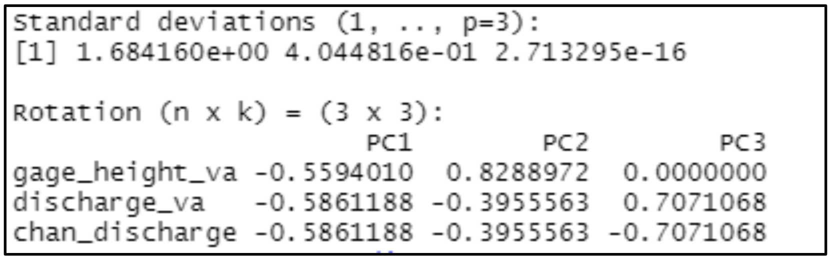
Woodward, OK through a PCA test. With that said I was able to retain tree of the six variables due to the fact that the other three variables had missing values. Those six variables used are as follows: gage height, discharge, channel flow, channel width, channel area, and channel velocity. Out of those six the three variables I had to eliminate for missing data are as follows: channel area, channel width, and channel velocity. If you notice from the scree plot, this example gives us a good idea of how many components or variables we want to keep and use. Given those results we were only able to keep the gage height variable. So by looking at PC1 we can state that we know 94% of the variability just by looking at that one principle component.

Figure 23. Principle Components Analysis tests for streamflow data at Harrah, OK





```
Importance of components:
                          PC1     PC2       PC3
Standard deviation     1.6842 0.40448 2.713e-16
Proportion of Variance 0.9455 0.05454 0.000e+00
Cumulative Proportion  0.9455 1.00000 1.000e+00
```



```
Standard deviations (1, .., p=3):
[1] 1.684160e+00 4.044816e-01 2.713295e-16

Rotation (n x k) = (3 x 3):
                     PC1        PC2        PC3
gage_height_va -0.5594010  0.8288972  0.0000000
discharge_va   -0.5861188 -0.3955563  0.7071068
chan_discharge -0.5861188 -0.3955563 -0.7071068
```

As it turns out, the Harrah, OK stream flow data yields the same result as does the Woodward, OK streamflow data when ran through a One-Way ANOVA test. I wanted to try and perform a cluster analysis. However, because I was only able to retain the gage height variable data, the clustering was very messy. Therefore, I chose to keep cluster analysis out of this form of research.

Conclusions

The overall goal of this study was to perform a statistical analysis on streamflow data from the North Canadian River at Woodward, Oklahoma and compare it to streamflow data from Harrah, Oklahoma. The majority of the data was skewed, I think in part to the extreme fluctuations of measurements. I would also argue that streamflow data is just messy in general. However, when visualizing the data via boxplots and the use of the five-number summary to determine variability and central tendency, helped understand the nature of the streamflow datasets. It was clear that the data had a lot more similarities than differences. However, by performing T-Tests on the discharge means I was able to infer that there were differences in discharge or streamflow across both locations. Furthermore, you could see strong non-linear correlations across both locations in streamflow and gage height. One possible inference I may make is that when gage height increases, discharge, or streamflow increases. To me this indicates that when there has been precipitation present or runoff into those rivers then you can expect discharge to increase. The other interesting relationship that is worth noting is that of the measurement ratings and control types across both locations. Through Chi-Squared testing we were able to determine that there are relationships between measurement ratings and control types. Why is this important, well, I would argue that if the majority of the measurement ratings are fair or good based off control type then they are representative samples of streamflow data.

Looking back at this data, I would state that it was slightly difficult to work with because most of my distributions weren't necessarily normal. However, I was still able to perform ample statistical tests to then draw conclusions on the similarities and differences across both locations. If further research happened to be conducted, I would suggest finding datasets that didn't have any missing values to help eliminate some of the error. Given cleaner data, you would be able to consider multiple linear regression models, a more successful principle components analysis, and clustering analysis. Overall, I think that this type of research is important to understand how streamflow can be impacted by a changing climate and population increase.

References

Asadieh, Behzad, & Krakauer, Nir Y. (2017). Global change in streamflow extremes under climate change over the 21st century. *Hydrology and Earth System Sciences, 21*(11), 5863-5874.

Bureau of Reclamation (2016). Climate Change Adaptation Strategy. (33 pp.). Retrieve from

Davids, Jeffrey C, Rutten, Martine M, Pandey, Anusha, Devkota, Nischal, Van Oyen, Wessel David, Prajapati, Rajaram, & Van de Giesen, Nick. (2019). Citizen science flow – an assessment of simple streamflow measurement methods. *Hydrology and Earth System Sciences, 23*(2), 1045-1065.

Dinpashoh, Yagob, Singh, Vijay P, Biazar, Seyed Mostafa, & Kavehkar, Shahab. (2019). Impact of climate change on streamflow timing (case study: Guilan Province). *Theoretical and Applied Climatology, 138*(1-2), 65-76.

E. P. Maurer, I. T. Stewart, C. Bonfils, P. B. Duffy, & D. Cayan. (2007). Detection, attribution, and sensitivity of trends toward earlier streamflow in the Sierra Nevada. *Journal of Geophysical Research - Atmospheres, 112*(D11), D11118-N/a.

Guastini, Enrico, Zuecco, Giulia, Errico, Alessandro, Castelli, Giulio, Bresci, Elena, Preti, Federico, and Penna, Daniele. "How Does Streamflow Response Vary with Spatial Scale? Analysis of Controls in Three Nested Alpine Catchments." Journal of Hydrology (Amsterdam) 570 (2019): 705-18. Web.

Holtschlag, D., Michigan. Department of Natural Resources Environment, & Geological Survey. (2011). *Use of instantaneous streamflow measurements to improve regression estimates of index flow for the summer month of lowest streamflow in Michigan by David J. Holtschlag ; prepared in cooperation with the Michigan Department of Natural*

Resources and Environment. (Scientific investigations report ; 2010-5236). Reston, Va.: U.S. Dept. of the Interior, U.S. Geological Survey.

Kao, Wen-Hsiung, Kincannon, Don F., Gaudy, A. F., Jr., and Graves, Quintin B. Statistical Study of the Relationship Between Inorganic Quality of River Water and Streamflow (2016). Web.

Lehner, Flavio, Wood, Andrew W, Llewellyn, Dagmar, Blatchford, Douglas B, Goodbody, Angus G, & Pappenberger, Florian. (2017). Mitigating the Impacts of Climate Nonstationarity on Seasonal Streamflow Predictability in the U.S. Southwest. *Geophysical Research Letters, 44*(24), 12,208-12,217.

Leta, Olkeba, El-Kadi, Aly, & Dulai, Henrietta. (2018). Impact of Climate Change on Daily Streamflow and Its Extreme Values in Pacific Island Watersheds. *Sustainability (Basel, Switzerland), 10*(6), 2057.

Lurry, D.L., and Tortorelli, R.L., 1996, Estimated freshwater withdrawals in Oklahoma, 1990: U.S. Geological Survey Water-Resources Investigations Report 95-4276, 2 sheets.

Teuling, Adriaan J, De Badts, Emile A. G, Jansen, Femke A, Fuchs, Richard, Buitink, Joost, Hoek van Dijke, Anne J, & Sterling, Shannon M. (2019). Climate change, reforestation/afforestation, and urbanization impacts on evapotranspiration and streamflow in Europe. *Hydrology and Earth System Sciences, 23*(9), 3631-3652.

Wahl, Kenneth L., and Robert L. Tortorelli. Changes in Flow in the Beaver-North Canadian River Basin Upstream from Canton Lake, Western Oklahoma. U.S. Dept. of the Interior, U.S. Geological Survey, 1997.

Who We Are. (n.d.). United States Geological Survey. Retrieved December 12, 2020, from https://www.usgs.gov/about/about-us/who-we-are

Wimbrow,George H.,,II. (2012). A statistical analysis of streamflow trends for the State of Michigan (Order No. 1514548). Available from ProQuest Dissertations & Theses Global. (1030437543). Retrieved from https://login.ezproxy.lib.ou.edu/login?url=https://www-proquest-com.ezproxy.lib.ou.edu/dissertations-theses/statistical-analysis-streamflow-trends-state/docview/1030437543/se-2?accountid=12964