Data Visualizations and Data Mining Questions of Sea Level Trends Over Time

CS-453 Data Mining Course Project Report

Author: Michael Cummings

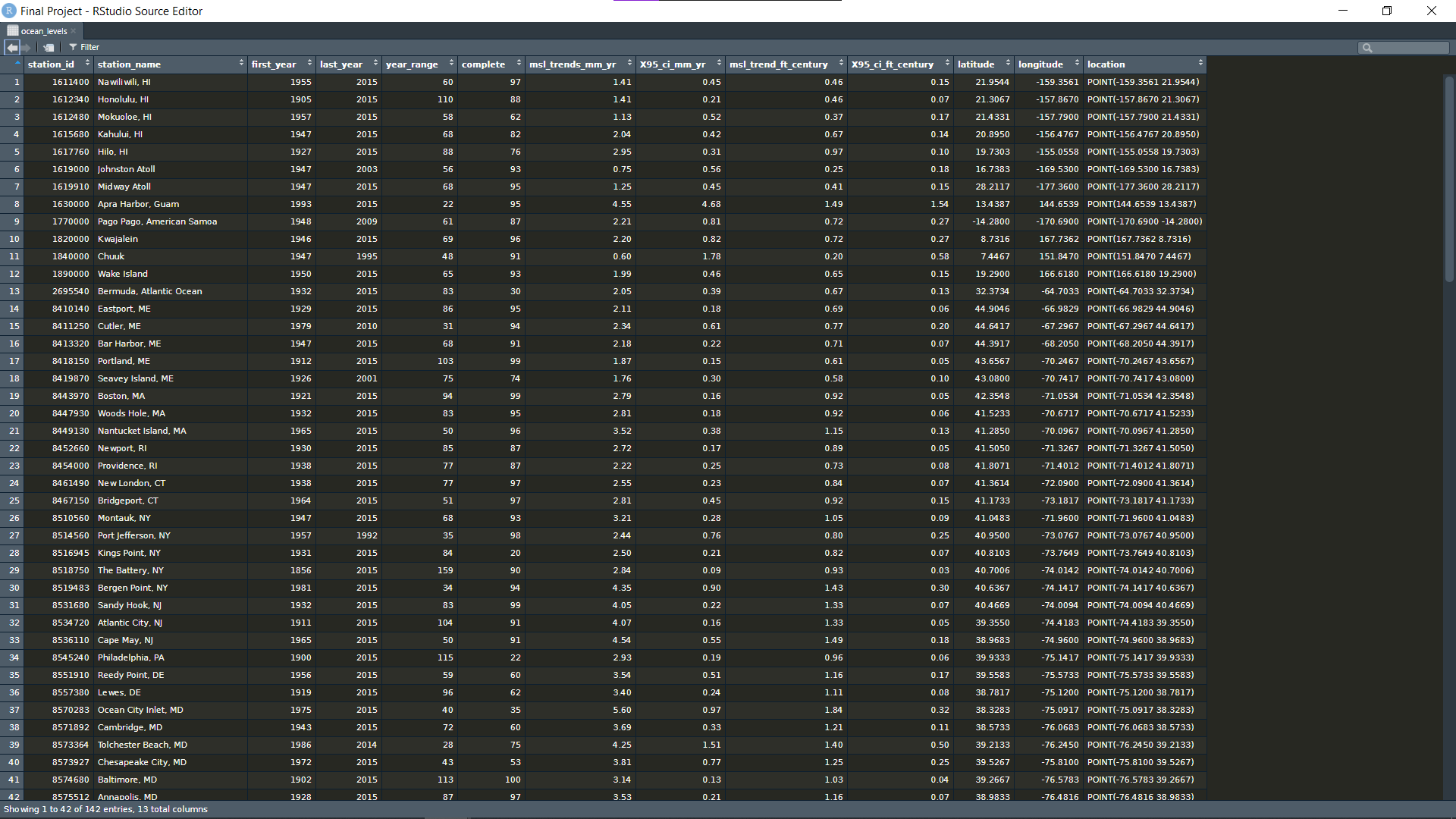
A picture containing chart

Description automatically generated

# Project Abstract:

The goal of this project was to answer meaningful data mining questions about a data set containing records of changes in sea level trend over time, collected by the National Oceanographic and Atmospheric Administration. Another objective of this project was to create meaningful data visualizations from the data set, using software packages in the R programming language. Using WEKA data mining tools in addition to the data visualization packages, I was able to produce informative graphics depicting the average change in sea level over time across the coastlines of the United States. Some data mining questions were answered more accurately than others by the tools in WEKA, but this data set may also be unsuitable for clustering and association rule questions because of its lack of nominal attributes. Numerical prediction could not be conducted with any accuracy using the data in this record, because the records of sea level change are averages over decades, which were beforehand averaged by month.

# Data Set:



Text

Description automatically generated

This data set was acquired from <https://data.world/noaa/sea-level-trends>

It contains 142 instances, which are composed of 13 attributes. There were no missing values.

* Nominal
  + Station\_id (unique)
  + Station\_name (unique)
* Numerical
  + First\_year (time)
  + Last\_year (time)
  + Year\_range (time)
  + Complete (continuous)
  + Msl\_trends\_mm\_yr (continuous)
  + X95\_ci\_mm\_yr (continuous)
  + Msl\_trends\_ft\_century (continuous)
  + X95\_ci\_ft\_century (continuous)
  + Latitude (Geospatial)
  + Longitude (Geospatial)
  + Location Coordinate Pair (Geospatial)

The oldest data that was used in the averages in this data set was recorded in the year 1856. The most recent year of records in this data set is 2015. Each instance in this data set has a minimum 30-year range of measurements recorded.

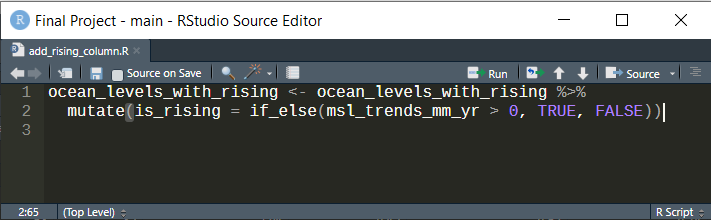
# Data Preprocessing:

The first issue I encountered with this data set was that multiple attributes would need to be excluded when using this data set in WEKA. The attributes station\_id, station\_name, first\_year, last\_year, and location would not be helpful in answering data mining questions. This left me with only 8 attributes to work with:

* the range of years of observations,
* the percentage of recorded observations compared to the range of years,
* the mean sea level trend change represented in millimeters per year
* the mean sea level trend change represented in feet per century
* the 95% confidence interval for mean sea level change in millimeters per year
* the 95% confidence interval for mean sea level change in feet per century
* Latitude
* Longitude

No nominal attributes remained, which would make working in WEKA difficult. To address this issue, I further altered the data set to contain two new nominal attributes:

* is\_rising, which was a logical value depending on whether an instance’s mean sea level trend over time was positive or negative
* state, which was a string value derived from the instance’s station\_name. The purpose of adding this attribute was to enable classification tasks in WEKA.



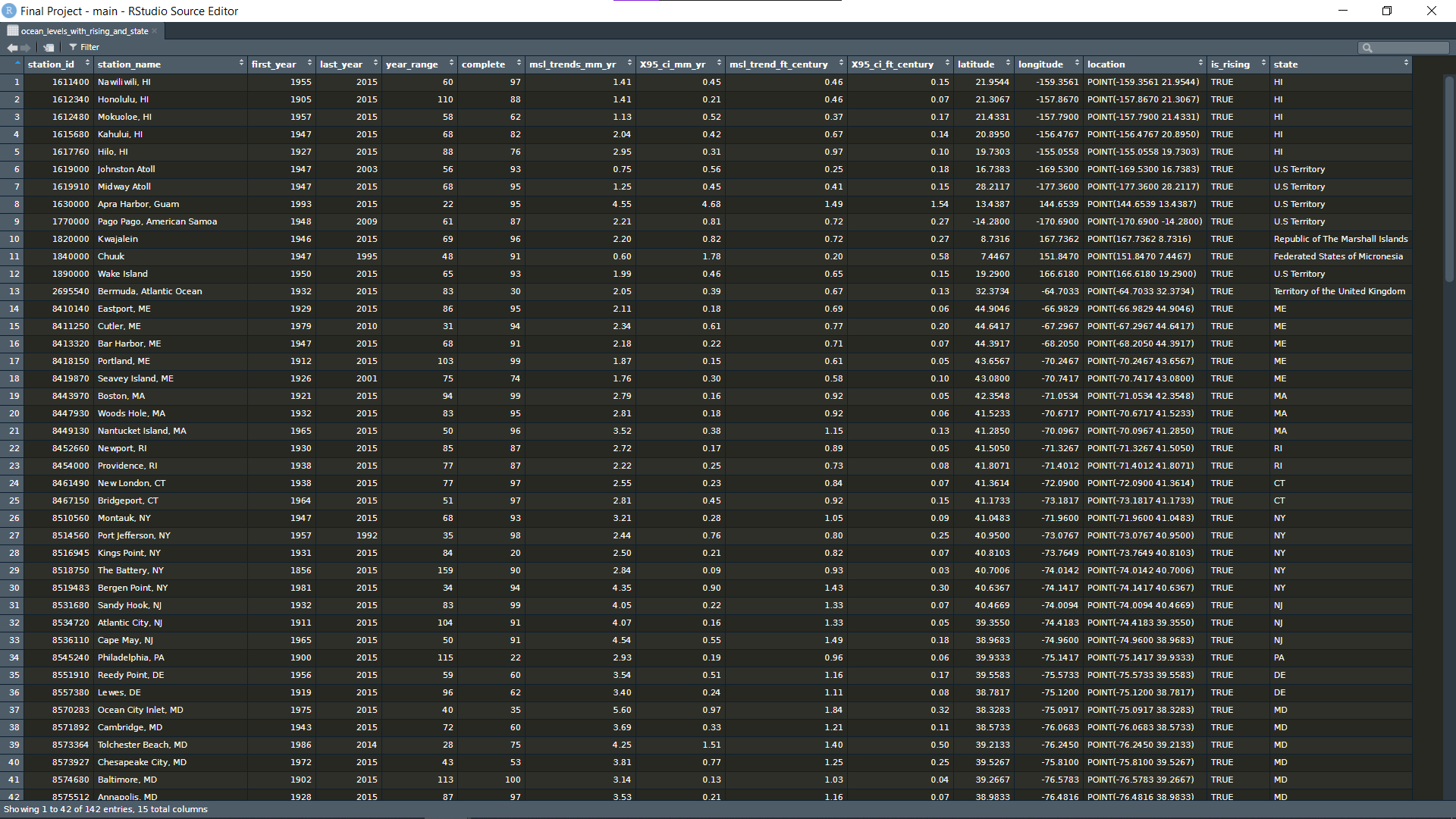
A computer screen capture

Description automatically generated with medium confidence

A computer screen capture

Description automatically generated with medium confidence

In addition to assigning stations in incorporated states to the correct state value, I also had to correctly assign those instances of stations which were not located in states in the United States. This required some outside research on my part. I learned that the instance of “Port Mansfield” is located in the state of Texas but is considered separate. I also learned that Guam, Bikini Atoll, and Midway Island are unincorporated U.S territories. I assigned each remaining instance to a value representing the government that has jurisdiction over the station’s location.

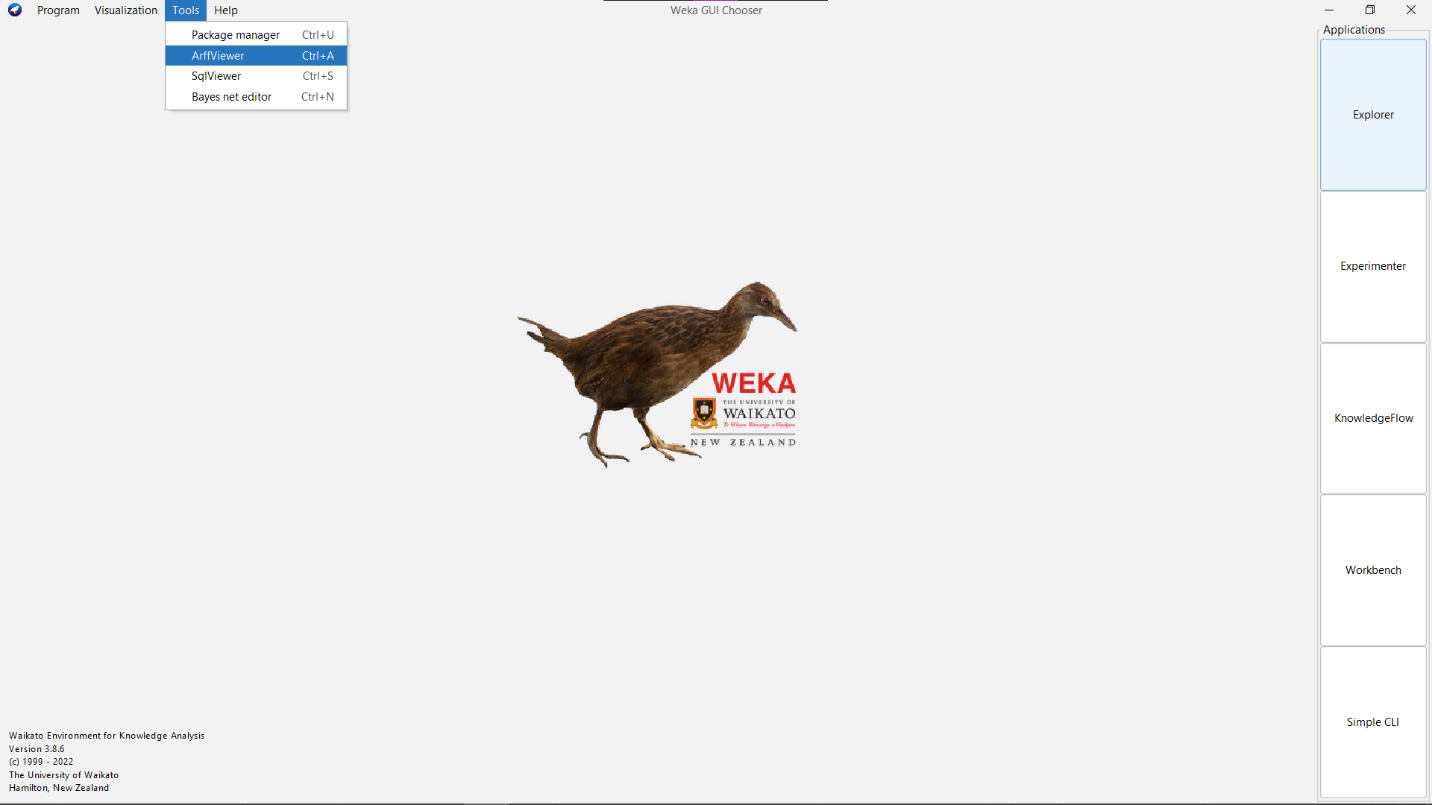


After the addition of these two attributes, the data set was ready to be converted for use in WEKA. The first step in this process was to write the data frame to a .csv file using the write.csv() function in R.

Graphical user interface

Description automatically generated

The next step was to convert the newly created .csv file into a .arff file that can be read by WEKA. There is a built-in utility within WEKA to perform this task.



Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generatedTable

Description automatically generatedA screenshot of a computer

Description automatically generated

# Data Mining Questions:

The questions I sought to answer with the tools in WEKA were:

* Can we classify instances by their state value?
* Can we classify instances based on whether their mean sea level trend is rising or falling?
* Can we cluster instances together depending on their mean sea level trend?
* Can we cluster instances together by state?
* Can we make numerical predictions on future changes in sea level?

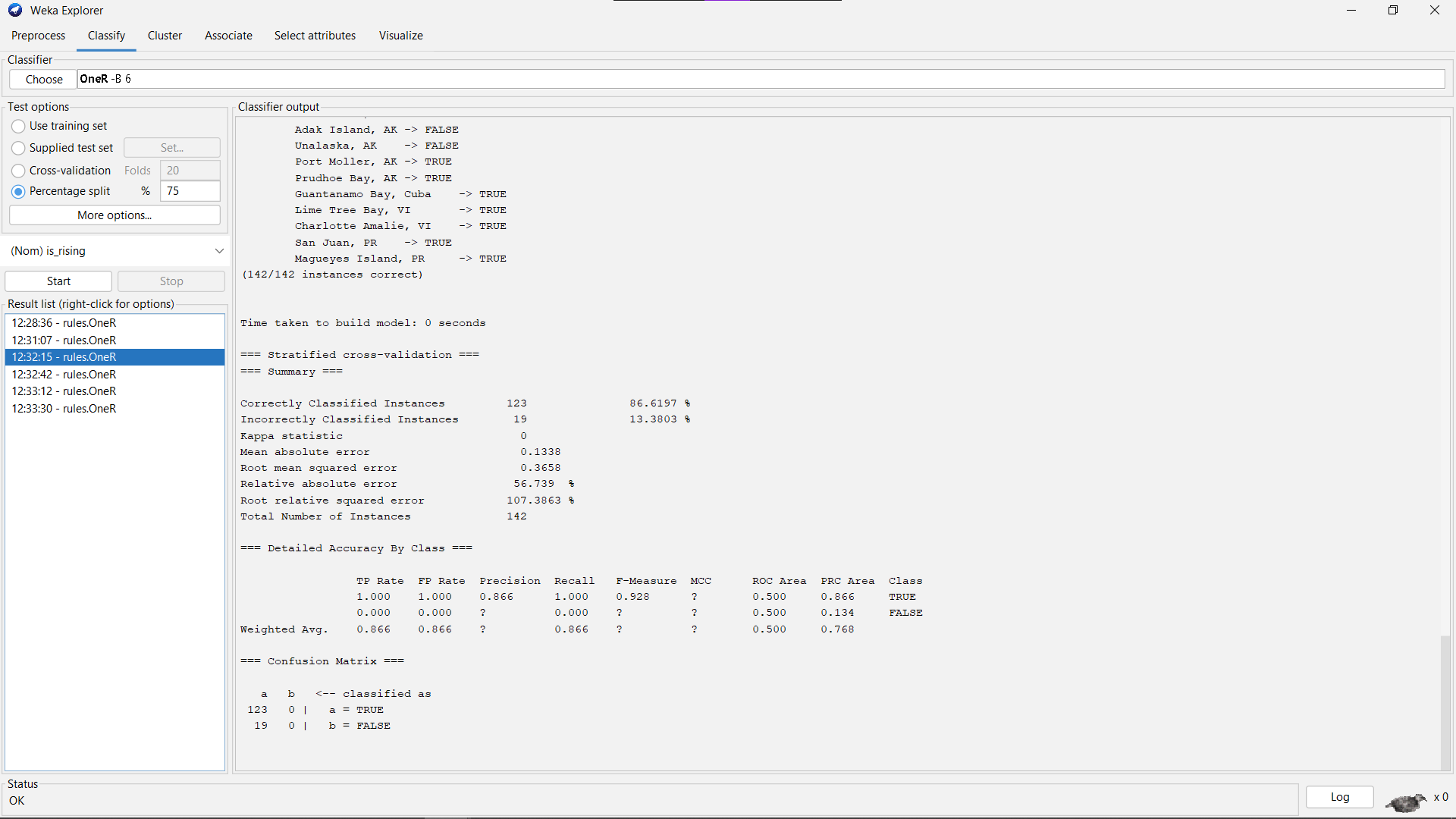
## Classification:

Classifying accuracy based on attribute “is\_rising”:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algorithm/ Test Option | OneR | ZeroR | **PART** | **J48** | REPTree | **RandomTree** | **Naïve Bayes** |
| Cross-Validation: 5 | 86.62%\* | 86.62%\* | **99.3%** | **99.3%** | 86.62%\* | **94.37%** | **95.07%** |
| Cross-Validation: 10 | 86.62%\* | 86.62%\* | **99.3%** | **99.3%** | 86.62%\* | **92.96%** | **93.66%** |
| Cross-Validation: 20 | 86.62%\* | 86.62%\* | **99.3%** | **99.3%** | 86.62%\* | **91.55%** | **93.66%** |
| Percent Split: 50% | 87.32%\* | 87.32%\* | **97.2%** | **97.2%** | 87.32%\* | 87.32% | **92.96%** |
| Percent Split: 66% | 87.5%\* | 87.5%\* | **100%** | **100%** | 87.5%\* | **100%** | **93.75%** |
| Percent Split: 75% | 82.85%\* | 82.86%\* | **100%** | **100%** | 82.86%\* | 85.71% | 88.57% |

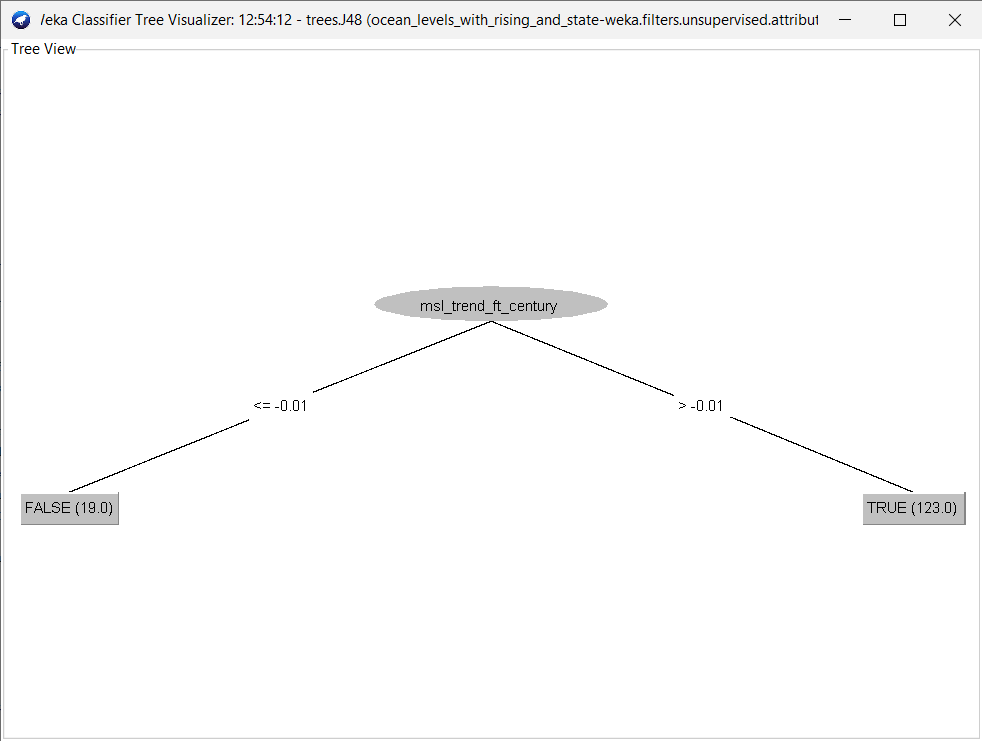
\* All “False” instances classified incorrectly

is_rising ZeroR with 10-fold cross-validation

Graphical user interface, text

Description automatically generated



Diagram

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, application

Description automatically generatedGraphical user interface, table

Description automatically generatedGraphical user interface

Description automatically generated with medium confidence

Classifying based on attribute “state”:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algorithm/ Test Option | OneR | ZeroR | PART | J48 | REPTree | RandomTree | Naïve Bayes |
| Cross-Validation: 5 | 3.52% | 11.27% | 66.9% | 69.01% | 65.49% | 63.38% | 64.08% |
| Cross-Validation: 10 | 3.52% | 11.27% | 64.79% | 66.9% | 64.79% | 59.86% | 64.79% |
| Cross-Validation: 20 | 3.52% | 11.27% | 63.38% | 66.9% | 64.79% | 59.86% | 64.79% |
| Percent Split: 50% | 2.82% | 9.86% | 36.62% | 40.84% | 32.39% | 42.25% | 42.25% |
| Percent Split: 66% | 4.17% | 14.58% | 43.75% | 58.33% | 54.17% | 39.58% | 52.08% |
| Percent Split: 75% | 2.86% | 11.43% | 71.43% | 65.71% | 40% | **80%** | 57.14% |

Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

Text

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generatedA picture containing text

Description automatically generated

Graphical user interface, application

Description automatically generatedGraphical user interface, text, application

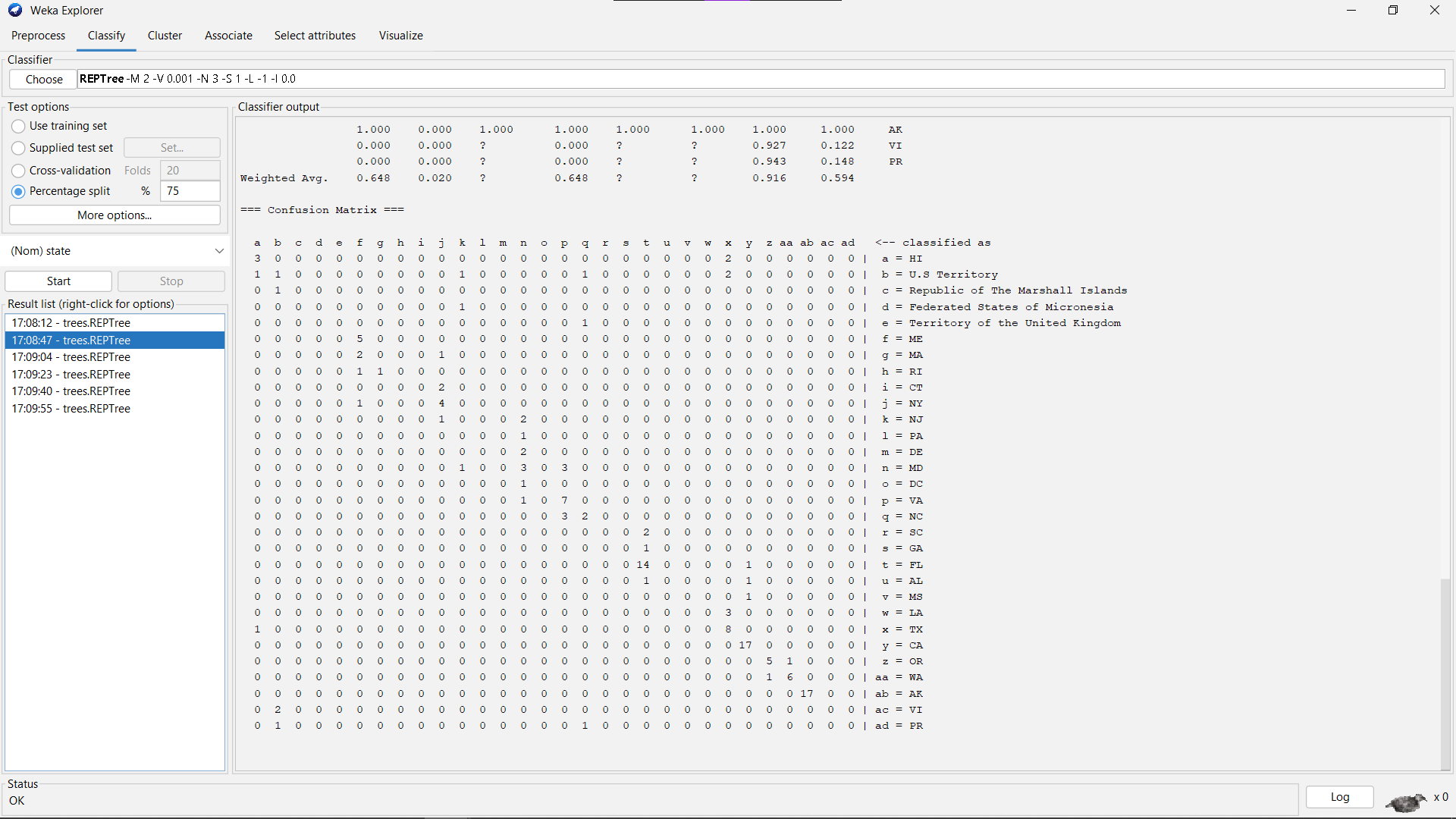
Description automatically generatedText

Description automatically generated with low confidence

Diagram

Description automatically generated

Graphical user interface, text, application

Description automatically generatedDiagram

Description automatically generated

Graphical user interface, application, table

Description automatically generatedTable

Description automatically generated

A picture containing table

Description automatically generated

## Classification Conclusions:

Classifying instances on the logical is\_rising attribute was incredibly accurate through some algorithms, while others like OneR and ZeroR simply assigned all instances to the majority class. Algorithms like PART and J48 performed very well on classification based on a logical attribute, however. The task of classifying instances by state was handled less accurately. The highest rate of accuracy reached was through a 75% split on the RandomTree algorithm. Through other testing options, the RandomTree algorithm performed similarly to the others, with the exception of ZeroR and OneR. These algorithms were not able to perform this task accurately, with a peak accuracy rate of 14.58% through ZeroR.

## Clustering:

Clustering instances based on “is\_rising” class:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Cobweb | EM | SimpleKMeans = 3 | SimpleKMeans = 4 | SimpleKMeans = 6 |
| Accuracy | 2.11% | 35.2% | 62.67% | 43.66% | 40.85% |
| No. Clusters | 198 | 8 | 3 | 4 | 6 |
| Hierarchal/Flat? | Hierarchal | Flat | Flat | Flat | Flat |
| Deterministic/Probabilistic? | Deterministic | Deterministic | Deterministic | Deterministic | Deterministic |
| Disjoint/Overlapping? | Overlapping | Disjoint | Disjoint | Disjoint | Disjoint |

Diagram, timeline

Description automatically generated

Graphical user interface, text, application, email

Description automatically generatedA computer screen capture

Description automatically generated with medium confidenceGraphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Clustering instances based on “state” class:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Cobweb | EM | SimpleKMeans = 3 | SimpleKMeans = 6 | SimpleKMeans = 10 |
| Accuracy | [Error] | 24.65% | 13.38% | 26.06% | 29.58% |
| No. Clusters | 192 | 5 | 3 | 6 | 10 |

## Clustering Conclusions:

The clustering tools in WEKA were only somewhat accurate when clustering instances together by the is\_rising attribute. The SimpleKMeans algorithm provided the highest accuracy at 62.67%, with a clustering option of k = 3. Clustering instances together by state was not accurate, and was in fact impossible through the Cobweb algorithm. I increased the number of clusters generated by the SimpleKMeans algorithm compared to the “is\_rising” evaluation to accommodate the large number of different values possible in the state attribute.

## Association Rules:

Graphical user interface, text, application, email

Description automatically generated

I was unable to conduct association rule testing through the Apriori algorithm.

## Numerical Prediction:

Numerical prediction based on “msl\_trend\_mm\_year” class:

|  |  |
| --- | --- |
|  | Linear Regression |
| Cross-Validation: 5 | 58% |
| Cross-Validation: 10 | 59% |
| Cross-Validation: 20 | 60% |
| Percent Split: 50% | 51% |
| Percent Split: 66% | 48% |
| Percent Split: 75% | 50% |

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text

Description automatically generated

Numerical prediction based on “msl\_trend\_ft\_century” class:

|  |  |
| --- | --- |
|  | Linear Regression |
| Cross-Validation: 5 | 58% |
| Cross-Validation: 10 | 59% |
| Cross-Validation: 20 | 60% |
| Percent Split: 50% | 51% |
| Percent Split: 66% | 50% |
| Percent Split: 75% | 50% |

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text

Description automatically generated

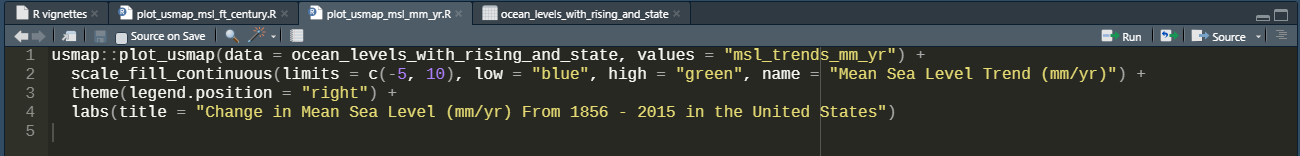
## Numerical Prediction Conclusions:

The linear prediction model in WEKA was only somewhat accurate in modeling the mean sea level trend attributes. Though similar rates of accuracy were achieved through both mean sea level measurements, the models constructed for each were quite different. The weights and their assigned variables differ greatly between the two evaluated variables.

# Data Visualizations:

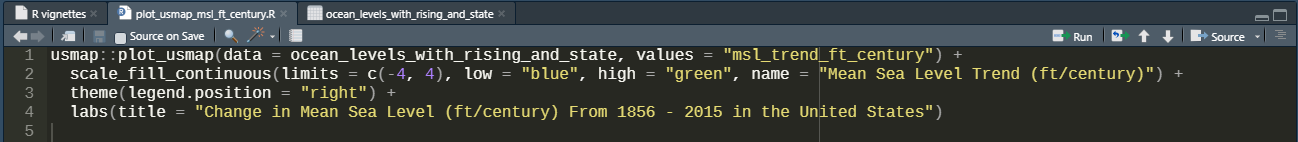
A picture containing chart

Description automatically generated



A picture containing map

Description automatically generated



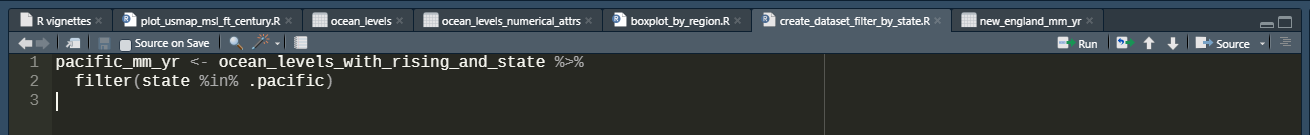
Statistical summary of mean sea level trend (mm/yr) across the United States:

|  |  |
| --- | --- |
| Minimum | -17.59 |
| 1st Quartile | 1.403 |
| Median | 2.285 |
| Mean | 1.773 |
| 3rd Quartile | 3.530 |
| Maximum | 9.650 |

Standard deviation for mean sea level trend (mm/yr) across the United States: 3.768

Text

Description automatically generated

Visualizations by region were made possible by the built-in groups of states within the usmap package, as well as functions provided in the dplyr package. I was able to create new data frames grouped together by geographical region. 

A computer screen capture

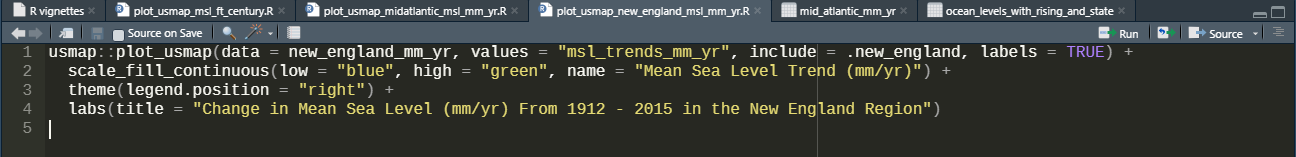
Description automatically generated with low confidence

A computer screen capture

Description automatically generated with low confidence

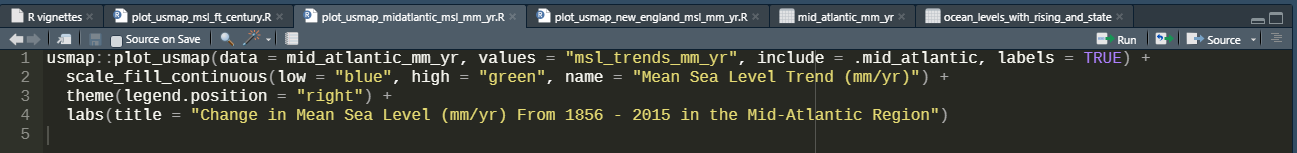
Map

Description automatically generated



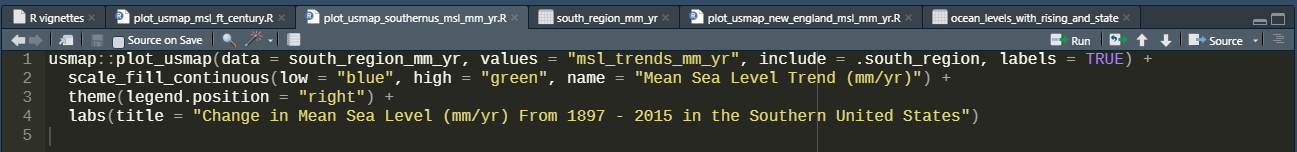
Chart

Description automatically generated



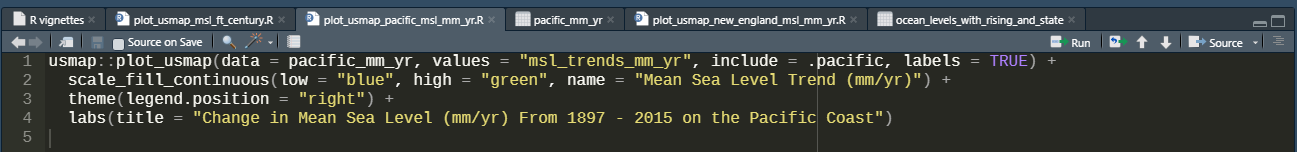
A picture containing diagram

Description automatically generated



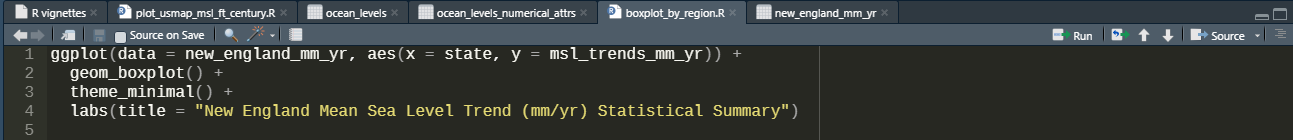
Map

Description automatically generated



Chart, box and whisker chart

Description automatically generated



Statistical Summary for Mean Sea Level Trend (mm/yr) in New England:

|  |  |
| --- | --- |
| Minimum | 1.760 |
| 1st Quartile | 2.163 |
| Median | 2.445 |
| Mean | 2.473 |
| 3rd Quartile | 2.795 |
| Maximum | 3.520 |

Standard deviation for mean sea level trend (mm/yr) in New England: 0.492

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Text

Description automatically generated

Statistical summary for Mean Sea Level Trend (mm/yr) in the Mid-Atlantic region:

|  |  |
| --- | --- |
| Minimum | 2.440 |
| 1st Quartile | 2.840 |
| Median | 3.210 |
| Mean | 3.437 |
| 3rd Quartile | 4.070 |
| Maximum | 4.540 |

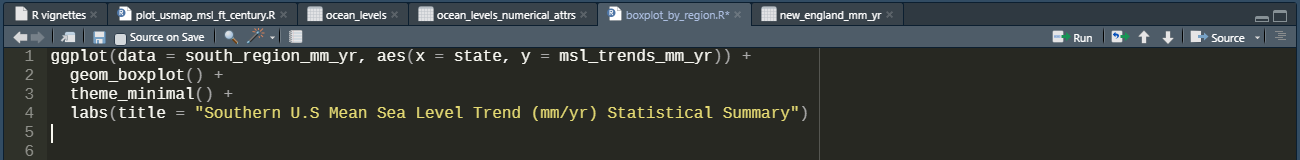
Standard deviation for mean sea level trend mm/yr in Mid-Atlantic: 0.819

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generated



Statistical summary of mean sea level trend (mm/yr) in Southern United States:

|  |  |
| --- | --- |
| Minimum | 1.920 |
| 1st Quartile | 2.803 |
| Median | 3.550 |
| Mean | 3.834 |
| 3rd Quartile | 4.442 |
| Maximum | 9.650 |

Standard deviation for mean sea level trend (mm/yr) in the Southern United States: 1.586

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Graphical user interface, text

Description automatically generated

Statistical summary of mean sea level trend (mm/yr) on U.S Pacific Coast:

|  |  |
| --- | --- |
| Minimum | -17.590 |
| 1st Quartile | -0.912 |
| Median | 0.935 |
| Mean | -0.9412 |
| 3rd Quartile | 1.85 |
| Maximum | 4.60 |

Standard deviation for mean sea level trend (mm/yr) on U.S Pacific Coast: 4.834

Text

Description automatically generated

Correlation plot of numerical variables:

Chart, scatter chart, bubble chart

Description automatically generated



Linear regression problems:

|  |  |  |
| --- | --- | --- |
| Coefficients | Intercept | Estimate |
| Complete ~ year\_range | 78.62 | 0.058 |
| Msl\_trends\_mm\_yr ~ latitude | 7.495 | -0.156 |
| Msl\_trends\_mm\_yr ~ longitude | 3.982 | 0.023 |
| X95\_ci\_mm\_yr ~ year\_range | 1.60 | -0.015 |

Text

Description automatically generated

Chart, histogram

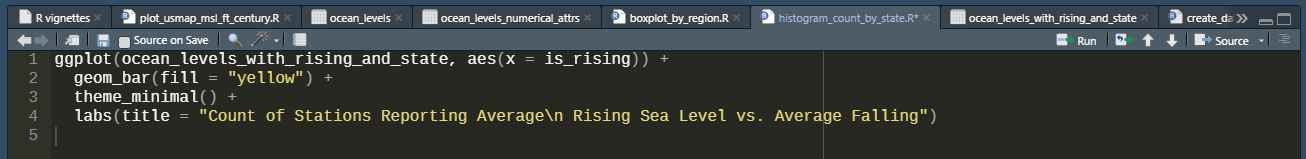
Description automatically generated

Graphical user interface, text

Description automatically generated

Chart, bar chart

Description automatically generated



# Conclusions:

I initially chose to work with this data set because I wanted to build my project around data that was collected from the natural world, and was illustrative of the continuing global climate crisis. At first, I sought out data sets containing information on sharks, such as a record of individual animals and their measurements, or a record of the movements of the animals over a given time. Finding a suitable data set on that subject proved too difficult, so I broadened my scope to center my project on the physical changes of the ocean. I wanted to construct meaningful and informative visualizations that reflected the physical changes we’ve seen on our planet over the last century. I also wanted to learn if reliable numerical prediction models could be constructed and used to prepare for future rises in mean sea level.

Working with the tools in WEKA and RStudio was an enlightening experience for me, and I would be interested to continue working with this software in the future. I was especially impressed by the ggplot2 and usmap packages.

I found that the nature of the data set made it difficult to construct an accurate numerical prediction model in WEKA. The data set is relatively small at 142 instances, and almost all of its meaningful attributes are continuous numeric values. In addition, the values for mean sea level trend were first averaged by month, then those averages were combined into yearly averages, and finally those values were combined into the lifetime averages recorded in the data set. The range of a stations’ years of observations varied greatly as well.

In the future, the information in this data set could be combined with other topics, like changes in the price of housing depending on the region’s change in sea level over time. I would like to find more detailed data on the change in mean sea level over time, as well as data from outside the United States and Western hemisphere.

# References

Data set: <https://data.world/noaa/sea-level-trends>

Tools:

* RStudio & R Programming Language
  + Corrplot
  + Tidyverse (ggplot2, dplyr, stringr, tibble)
  + Usmap: https://cran.r-project.org/web/packages/usmap/
* WEKA

Resources:

* R Reference Manual: <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>
* Adding Columns Based on Other Column Values in Dplyr: <https://www.marsja.se/r-add-column-to-dataframe-based-on-other-columns-conditions-dplyr/>
* StringR help: <https://evoldyn.gitlab.io/evomics-2018/ref-sheets/R_strings.pdf>
* Tidyverse for Beginners Cheat Sheet: <https://datacamp-community-prod.s3.amazonaws.com/c1fae72f-d2d7-4646-9dce-dd0f8fb5c5e8>
* Linear Regression in R Tutorial: <https://www.datacamp.com/tutorial/linear-regression-R>
* DataCamp Data Visualization with R and ggplot Courses
* United States State Abbreviations and Postal Codes: <https://www.bls.gov/respondents/mwr/electronic-data-interchange/appendix-d-usps-state-abbreviations-and-fips-codes.htm>