CSCE 474 Assignment 1 Report

Note: all corrections to the original report are made in **bold** font.

List of Figures and Tables

- Examine the data file in a text editor and summarize the dataset in terms of its purpose, number of attributes, their types, number of records, etc.

  The purpose of the dataset seems to be to record numerous attributes about eucalyptus across a select region of New Zealand, as found by examining the locality attribute. Each record has 20 attributes, all either nominal or numeric type. The attributes are information about the locale such as locality, altitude, longitude, rainfall, or information about the specific eucalyptus sample such as stem form, crown form, utility, height, and year with 736 records.

- Open the file in Weka Explorer and examine each attribute in more detail.
  - For each nominal attribute, note the number of different labels and their distribution. **See table 1 and figures 1 - 6.**

Table 1

Nominal Attribute	**Number of Labels**	**Distribution**	**Percent Missing**
Abbrev	16 unique	Max: Puk (84) Min: WSh (5)	**0%**
Locality	8 unique	Max: Southl_Wairapa (210) Min: Central_Poverty_Bay (6)	**0%**
Map_Ref	14 unique	Max: N158_344/626 (147) MIn: N151_922/226 (5)	**0%**
Latitude	12 unique	Max: 40__57 (192) Min: 82__32 (6)	**0%**
Sp	27 unique	Max: nd (86) Min: ro (2)	**0%**
Utility	5 unique	Max: good (214) Min: best (105)	**0%**

Name: Abbrev
Missing: 0 (0%)
Distinct: 16
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Cra	30	30.0
2	Cly	24	24.0
3	Nga	22	22.0
4	Wai	70	70.0
5	K81	65	65.0
6	Wak	73	73.0
7	K82	45	45.0
8	WSp	59	59.0
9	K83	49	49.0
10	Lon	53	53.0
11	Puk	84	84.0
12	Paw	55	55.0
13	K81a	33	33.0
14	Mor	63	63.0
15	Wen	6	6.0
16	WSh	5	5.0

Abbrev Attribute
Figure 1

Selected attribute

Name: Latitude
Missing: 0 (0%)
Distinct: 12
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	39__38	30	30.0
2	39__00	24	24.0
3	40__11	22	22.0
4	39__50	70	70.0
5	40__57	192	192.0
6	41__12	73	73.0
7	40__36	64	64.0
8	41__08	53	53.0
9	41__16	84	84.0
10	40__00	55	55.0
11	39__43	63	63.0
12	82__32	6	6.0

Latitude Attribute
Figure 2

Selected attribute

Name: Locality
Missing: 0 (0%)
Distinct: 8
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Central_Hawkes_Bay	93	93.0
2	Northern_Hawkes_Bay	24	24.0
3	Southern_Hawkes_Bay	86	86.0
4	Central_Hawkes_Bay_(coastal)	70	70.0
5	Central_Wairarapa	192	192.0
6	South_Wairarapa	210	210.0
7	Southern_Hawkes_Bay_(coastal)	55	55.0
8	Central_Poverty_Bay	6	6.0

Locality Attribute
Figure 3

Selected attribute

Name: Map_Ref
Missing: 0 (0%)
Distinct: 14
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	N135_382/137	30	30.0
2	N116_848/985	24	24.0
3	N145_874/586	22	22.0
4	N142_377/957	70	70.0
5	N158_344/626	147	147.0
6	N162_081/300	73	73.0
7	N158_343/625	45	45.0
8	N151_912/221	59	59.0
9	N162_097/424	53	53.0
10	N166_063/197	84	84.0
11	N146_273/737	55	55.0
12	N141_295/063	63	63.0
13	N98_539/567	6	6.0
14	N151_922/226	5	5.0

Map_Ref Attribute
Figure 4

Selected attribute

Name: Sp
Missing: 0 (0%)
Distinct: 27
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	co	27	27.0
2	fr	52	52.0
3	ma	3	3.0
4	nd	86	86.0
5	ni	31	31.0
6	ob	50	50.0
7	ov	62	62.0
8	pu	39	39.0
9	rd	37	37.0
10	si	9	9.0
11	mn	3	3.0
12	ag	9	9.0
13	bxs	17	17.0
14	br	28	28.0
15	el	12	12.0
16	fa	52	52.0
17	jo	9	9.0
18	ka	19	19.0
19	re	82	82.0

Sp Attribute
Figure 5

Selected attribute

Name: Utility
Missing: 0 (0%)
Distinct: 5
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	none	180	180.0
2	low	107	107.0
3	average	130	130.0
4	good	214	214.0
5	best	105	105.0

Utility Attribute
Figure 6

- For each numeric attribute, note their min, max, mean and variance. What percent of the values is missing? **See Table 2**

Table 2

Numeric Attribute	**Min**	**Max**	**Mean**	**Variance**	**Percent Missing**
Rep	1	22	2.026	1.104	0
Altitude	70	300	172.024	59.213	0
Rainfall	850	1750	1095.938	144.83	0
Frosts	-3	-2	-2.584	0.493	0
Year	1980	1986	1982.141	1.589	0
PMCno	1	3275	2054.739	618.44	1%
DBH	0.58	42085	72.947	1551.78	0
Ht	1.12	21.79	9.295	4.105	0
Surv	1.5	100	59.675	30.914	13%
Vig	0.5	5	3.076	1.013	9%
Ins_res	0	4.5	2.897	0.817	9%
Stem_Fm	0	5	2.996	0.714	9%
Crown_Fm	0	5	3.204	0.751	1%
Brnch_Fm	0	5	2.841	0.788	1%

- Add new attribute that is the index of the instances. Now examine the distribution of each feature. Select a numeric (“Ht”) and a nominal (“Locality”) and visualize their distribution as a function. **See Figure 7-9**

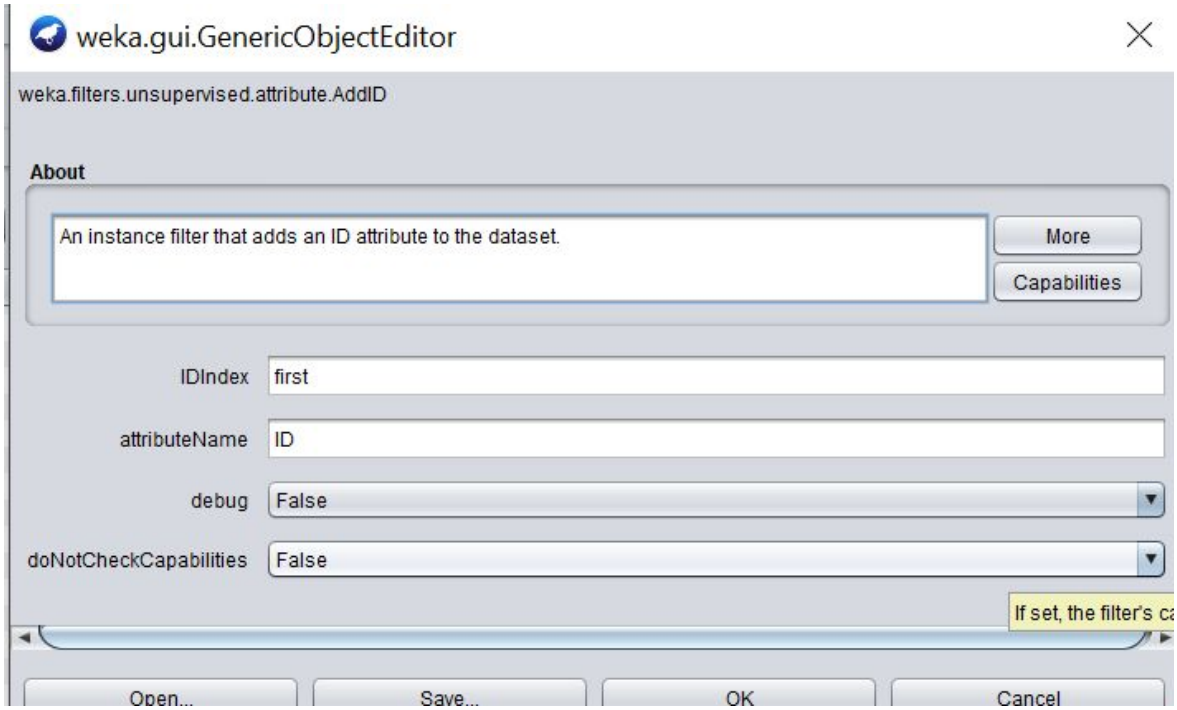

Figure 7

To add a new attribute that is the index of the instances, I used the addID filter. The filter added an attribute ID that indexed all of the data instances.

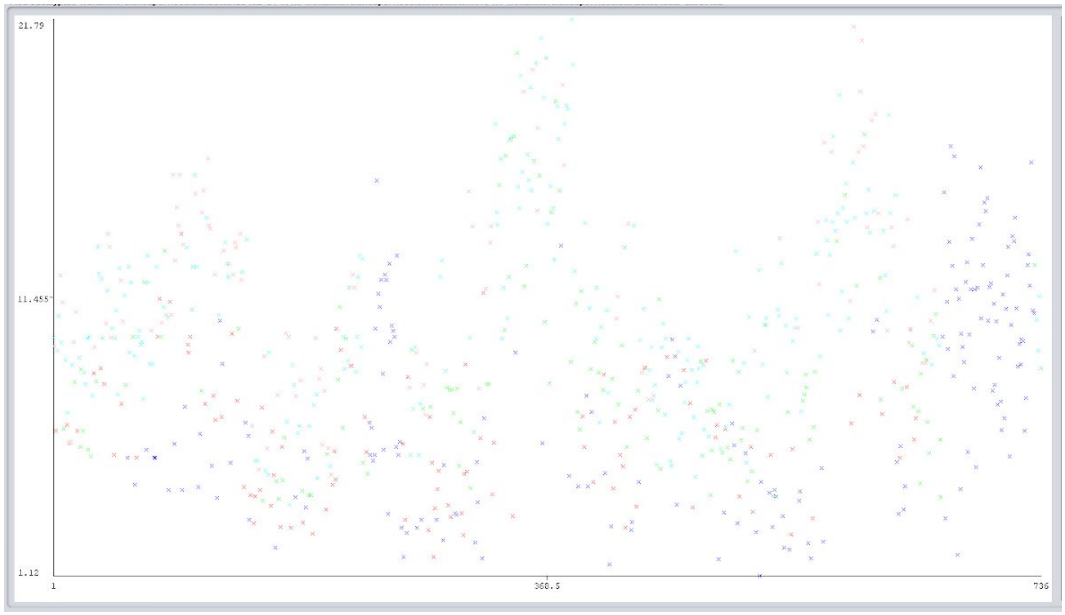

Figure 8
Height Plot

Figure 9
Locality Plot

- Also visualize the cross tabulations for pairwise attributes. Include them in your report. Examine the “Rainfall x Locality” and “Crown_Fm x Surv” cross tabs and summarize them briefly. **See figure 10 and 11**

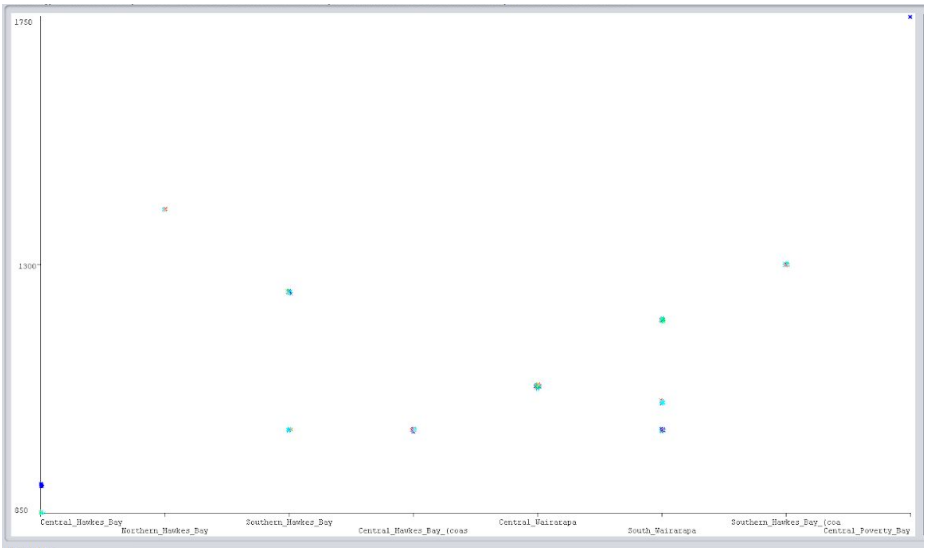

Figure 10
Rainfall x Locality

Through this plot we can notice that most locales have below-mean rainfall amounts with the outlier of central poverty bay having very high rainfall amounts. Also to note is that utility seems best around mid rainfall levels, with no utility at both extremes. (High utility plotted as light blue/green)

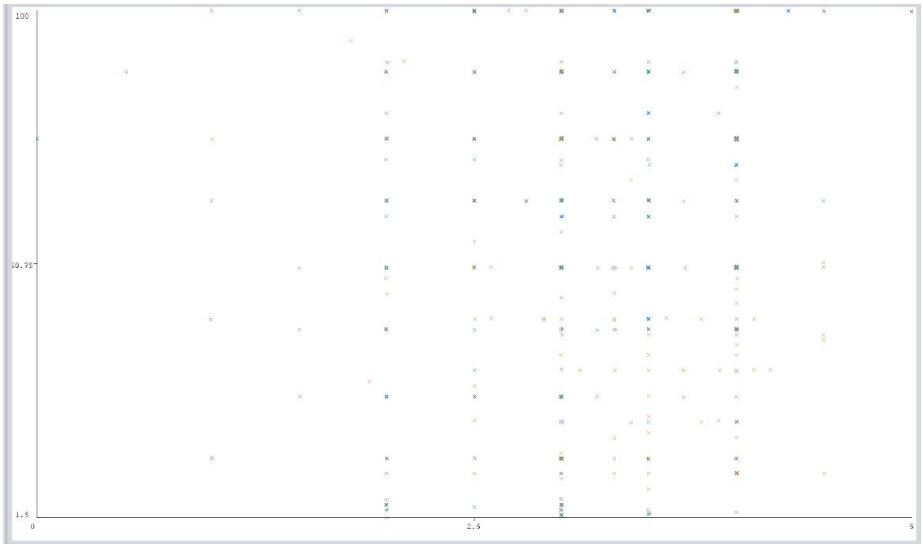

Figure 11
Crown_Fm x Surv

Here it is more difficult to see any direct relation from crown form against the survival attribute. We can see that almost all given values of crown form have a wide variance of low to high survival.

- Remove “Latitude” attribute of the dataset and then undo it.

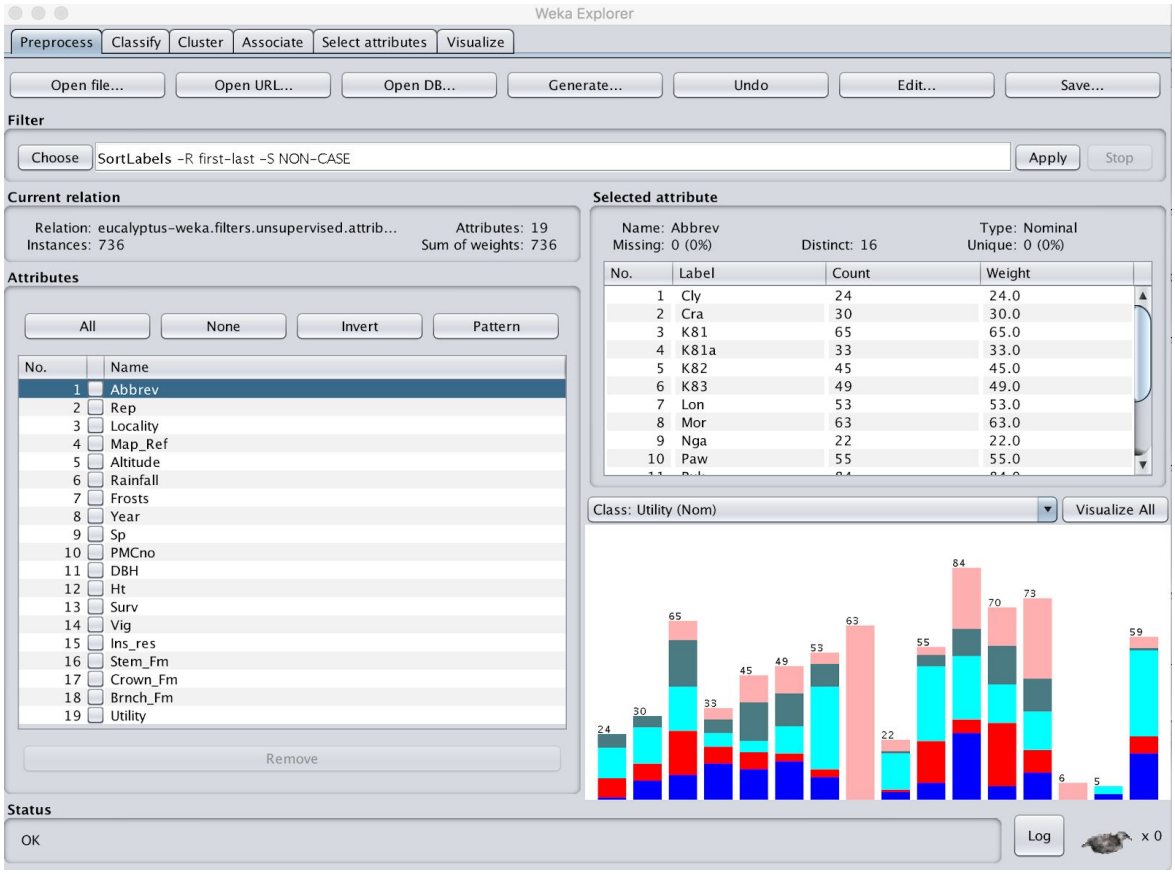

Figure 12

- Create new attribute of that is the log of Rainfall and call it “LogRainfall”. Examine its properties
    - I first copied the Rainfall column using a “Copy” **attribute filter**, found the column “Copy of Rainfall” via regular expression and replace using the MathEpxression filter, I applied log(A) to take the log of every object in the attribute. Result:

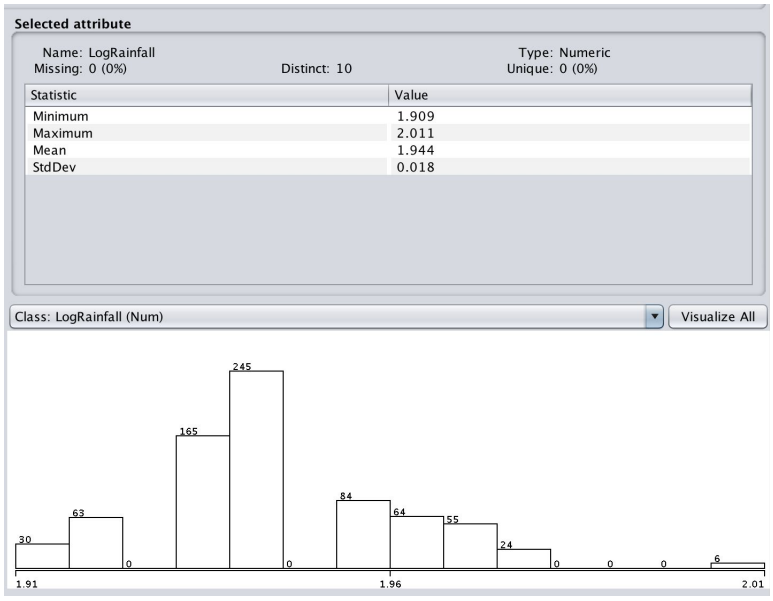

Figure 13

- Replace any annual precipitation values above 1500mm to 1400mm. Examine the summaries before and after this step.
    - **I used the NumericCleaner attribute filter to select values for which would replace values within the range I was looking for.**
        - **maxDefault was set to 1400, anything greater than this number would be written as 1400 (replacing all 1500mm rainfalls).**
        - **maxThreshold was set to 1500, triggering the replace event**
    - Figure 14 contains the summary of the precipitation values before the transformation, while Figure 15 contains the summary after.

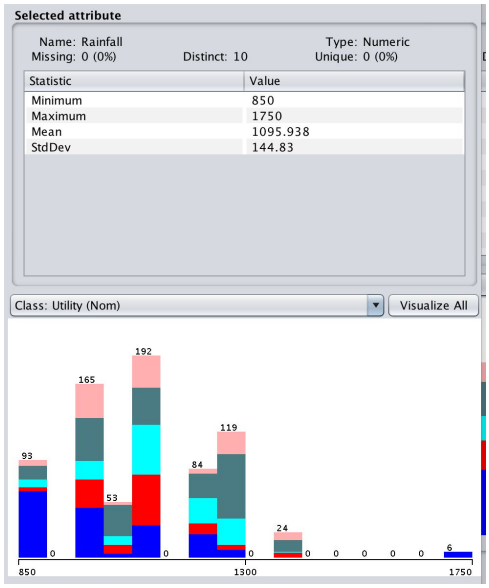

Figure 14
Thirty annual precipitation values were replaced as 1400.

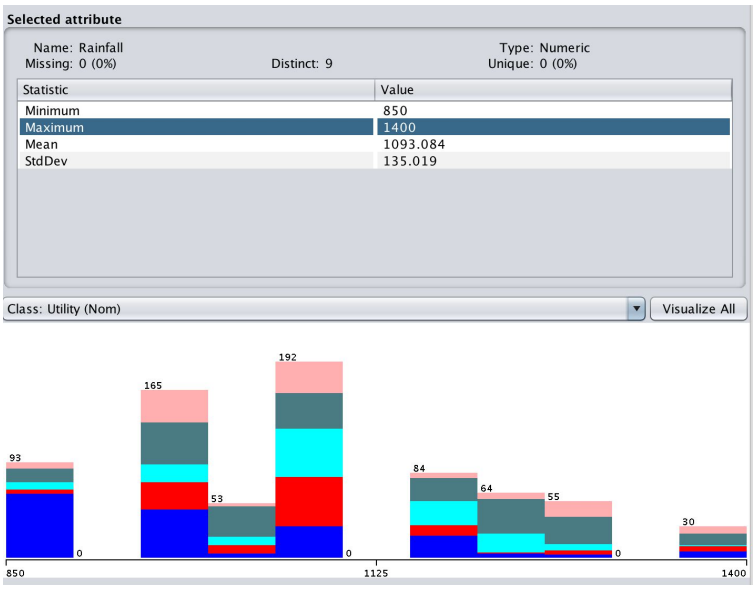

Figure 15
Original data before transformation

- Remove all records/instances that have any missing values. How many instances are left in the relation?
    - **First we used ReplaceMissingWithUserConstant, an unsupervised attribute filter that allows us to flag missing values with a default option**
        - **The options used for this filter were**
            - **nominalStringReplacementValue = MISSING**
            - **numericReplacementValue = 123456789999**
    - **With the missing instances now flagged, we filtered them using the unsupervised instance filter RemoveWithValues, which accepts Java style regex matching in support of finding and replacing values we set earlier.**
        - **matchMissingValues = True; this ensures that we match missing values not checked by us before**
        - **splitPoint = 12345679999; used to select missing numeric instances**
    - After removing all missing records, using RemoveWithValues (attributeIndex and matchMissingValues flags set) there are now 635 instances left in the relation.

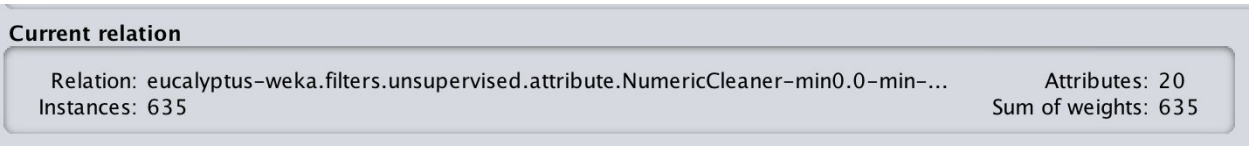

Figure 16

- **Save the new file as Eucalyptus New.arff and examine the changes**
    - **The resulting file was named Eucalyptus New.arff and saved. The result was a database with no missing values, and correctly ends at 635 relations.**

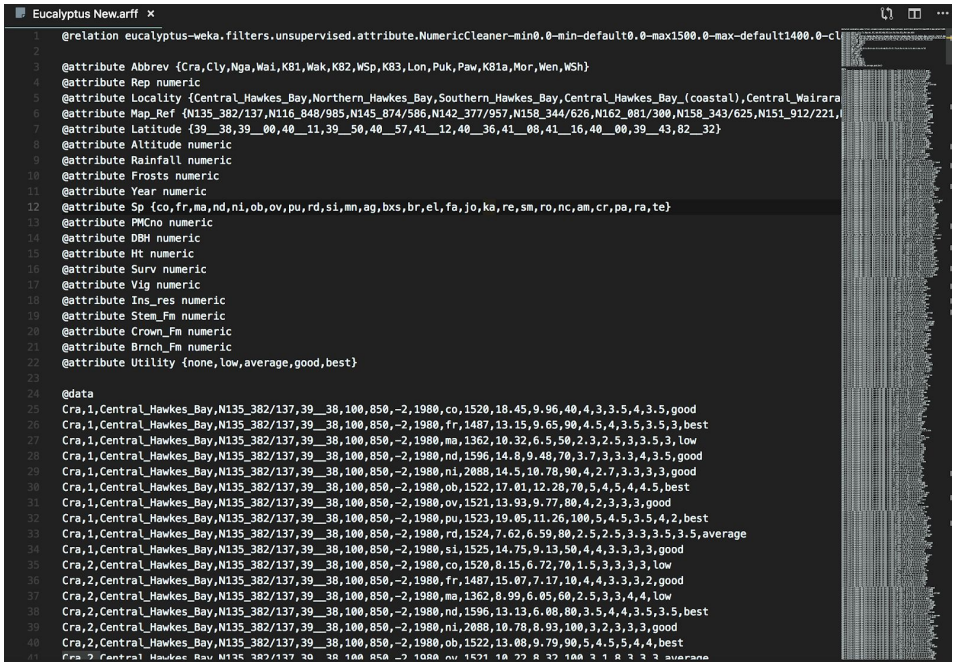

Figure 16a Eucalyptus New.arff opened in Visual Studio Code

- Choose two filters and describe them
    - **NumericCleaner,** attribute filter: This attribute filter allows for users to input regex-style expressions to find and replace values that are either too large, too small, or close to a certain “threshold”. This is useful when scanning attributes to tag for further processing, or simply replacing values.

    - **RemoveWithValues**, instance filter: This instance filter allows for users to input certain attribute indices, as well as a missing value parameter, that matches default missing value patterns, and can have a numeric split point (instances less than this point will be removed by default). This is useful when tagging and finding missing values within a relation.

- Convert each numeric attribute into no more than 5 discrete classes and save the data into a file called “Eucalyptus Nominal”.
    - I used the Discretize filter with 5 bins to convert all numeric attributes into nominal ones with 5 discrete classes.
        - **Using the original eucalyptus.arff file, on the choose filter option, I went through filters -> unsupervised -> attribute -> Discretize. Then set the bins parameter to 5. See Figure 17**
    - **After applying the filter all numeric attributes now nominal and only had 5 different values based on binning process similar to what was covered in class.**

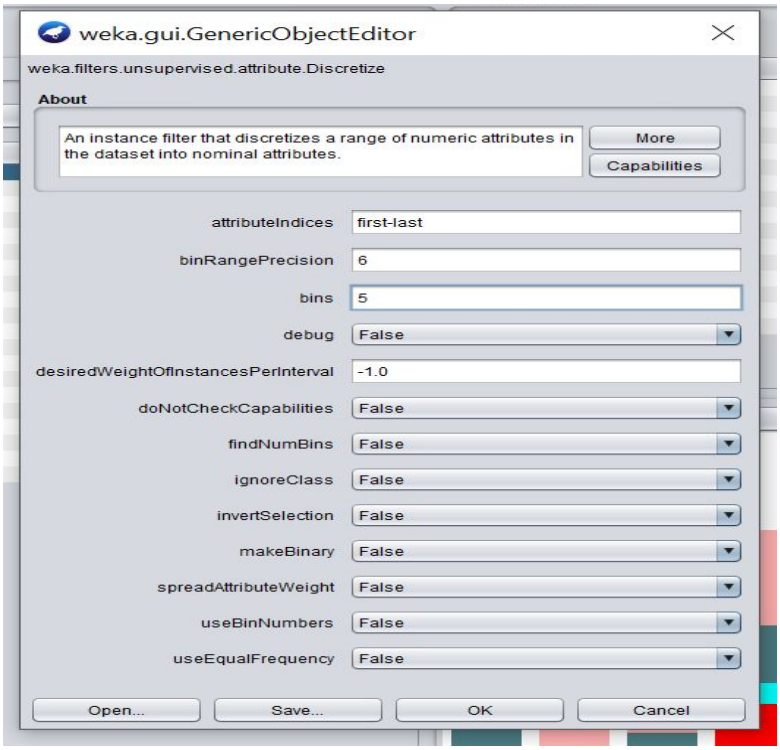

Figure 17

- Convert all the attributes (now nominal) into binary attributes, summarize your observations and save the file as “Eucalyptus Binary”

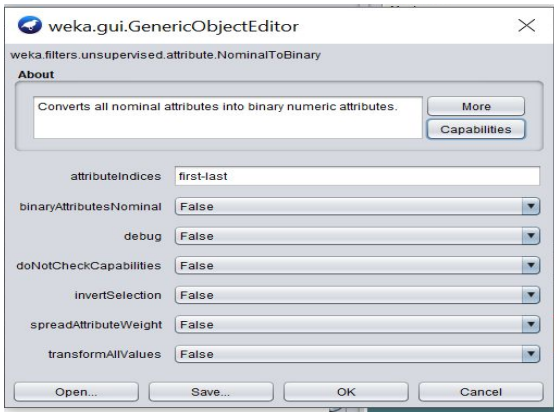

Figure 18

- I used the NominalToBinary filter with the default parameters to convert each now nominal attribute to binary.
    - **Using the modified arff file from the previous step, on the choose filter option, I went through filters -> unsupervised -> attribute -> NominalToBinary. I did not modify any of the default parameters. See Figure 18**

- The filter converted all attributes into new attributes splitting by their label into a new attribute with no more than 2 distinct values. This resulted in now 148 attributes. Previously numeric attributes that were converted to nominal from the previous step were split into binary attributes by their 5 discrete classes. Interesting to note was there while most attributes that were converted had 2 distinct values, 6 attributes are listed as one distinct value because there was actually no data available within the attribute (mean, max, etc… were listed as 0).

Contribution:

David focused on examining the original data, including summarizing the data for numeric attributes, examining distributions and visualizing the cross tabulations for pairwise attributes. Luis focused on Data Cleanup and Preprocessing, including creating the LogRainfall column, replacing annual precipitation values using the NumericCleaner filter, finding missing values and removing them, describing filters and adding the README to the project. David also completed association analysis.

David - 50%
Luis - 50%