February 9, 2018

Homework 1

Sanchit Singhal

CE 395R – Data Mining with Carlos Caldas

Spring 2018

University of Texas at Austin

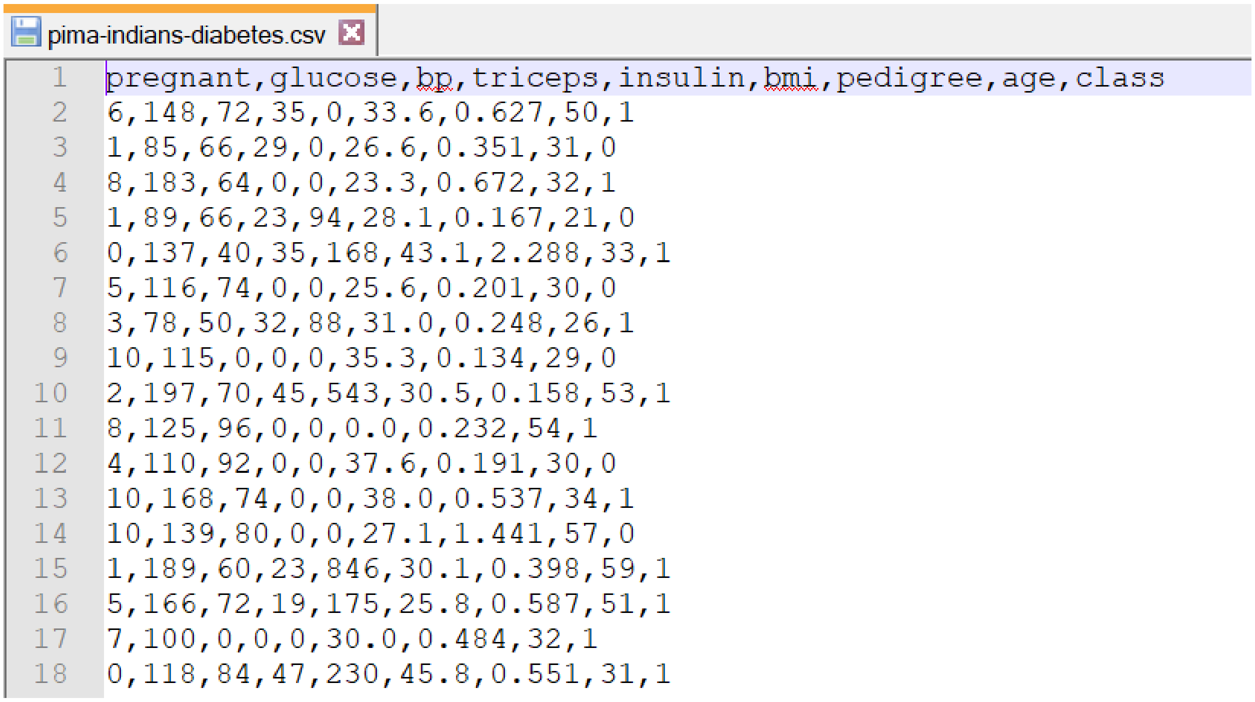
Pima Indians Diabetes Database

a) Data Selected

- Number of Attributes = 8 + class (all numeric)
  - Number of times pregnant
  - Plasma glucose centration
  - Diastolic blood pressure (mm Hg)
  - Triceps skin fold thickness (mm)
  - Serum insulin (mu U/ml)
  - Body Mass Index (kg/m^2)
  - Diabetes pedigree
  - Age (years)
  - Class variable (0 or 1) – Diabetes

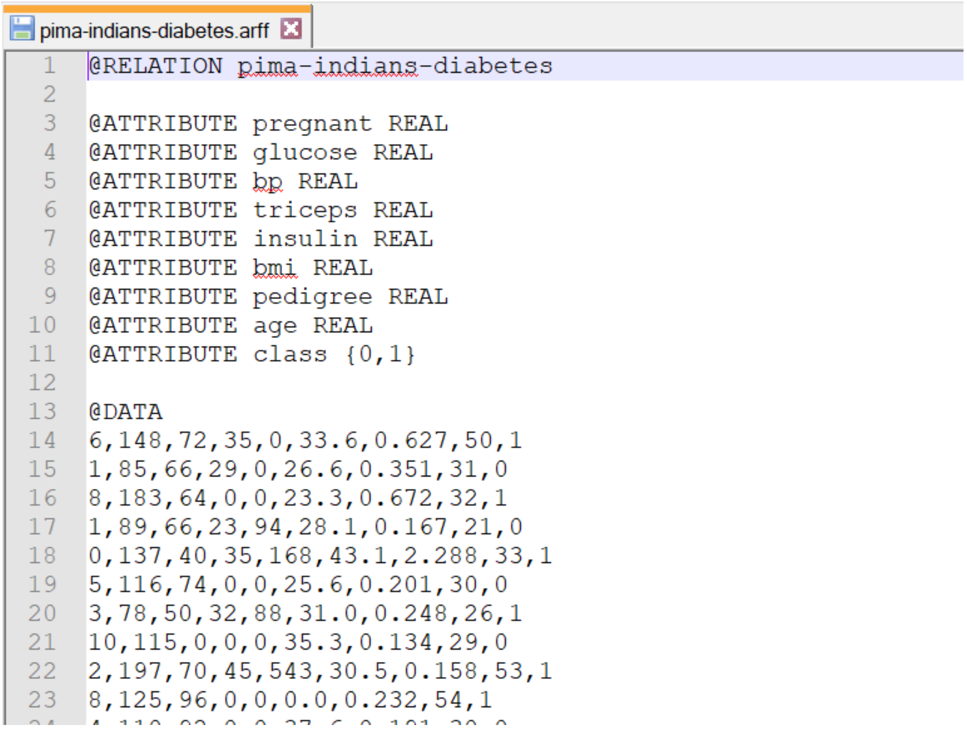
- Number of Instances = 768

I downloaded this dataset in .data and .name file types. I first converted this information into .csv:



```
pregnant,glucose,bp,triceps,insulin,bmi,pedigree,age,class
6,148,72,35,0,33.6,0.627,50,1
1,85,66,29,0,26.6,0.351,31,0
8,183,64,0,0,23.3,0.672,32,1
1,89,66,23,94,28.1,0.167,21,0
0,137,40,35,168,43.1,2.288,33,1
5,116,74,0,0,25.6,0.201,30,0
3,78,50,32,88,31.0,0.248,26,1
10,115,0,0,0,35.3,0.134,29,0
2,197,70,45,543,30.5,0.158,53,1
8,125,96,0,0,0.0,0.232,54,1
4,110,92,0,0,37.6,0.191,30,0
10,168,74,0,0,38.0,0.537,34,1
10,139,80,0,0,27.1,1.441,57,0
1,189,60,23,846,30.1,0.398,59,1
5,166,72,19,175,25.8,0.587,51,1
7,100,0,0,0,30.0,0.484,32,1
0,118,84,47,230,45.8,0.551,31,1
```

From my csv file I converted the file into .arff for ease of use with Weka 3.8:



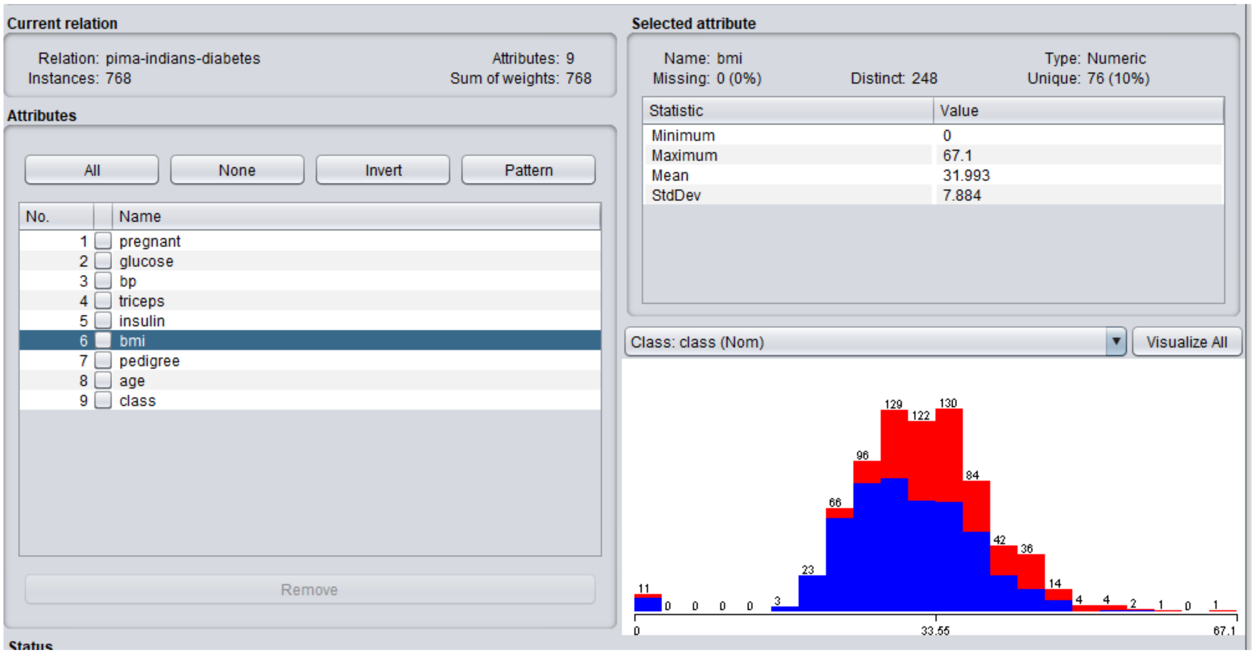
```
@RELATION pima-indians-diabetes

@ATTRIBUTE pregnant REAL
@ATTRIBUTE glucose REAL
@ATTRIBUTE bp REAL
@ATTRIBUTE triceps REAL
@ATTRIBUTE insulin REAL
@ATTRIBUTE bmi REAL
@ATTRIBUTE pedigree REAL
@ATTRIBUTE age REAL
@ATTRIBUTE class {0,1}

@DATA
6,148,72,35,0,33.6,0.627,50,1
1,85,66,29,0,26.6,0.351,31,0
8,183,64,0,0,23.3,0.672,32,1
1,89,66,23,94,28.1,0.167,21,0
0,137,40,35,168,43.1,2.288,33,1
5,116,74,0,0,25.6,0.201,30,0
3,78,50,32,88,31.0,0.248,26,1
10,115,0,0,0,35.3,0.134,29,0
2,197,70,45,543,30.5,0.158,53,1
8,125,96,0,0,0.0,0.232,54,1
```

b) Data Quality Problem Identified

After reviewing the data in the Explorer, I realized that the BMI attribute had a minimum of 0 and there were 11 such instances. This makes no sense as nobody can have zero mass. This tells me that this is actually just missing data:



Current relation

Relation: pima-indians-diabetes
Instances: 768
Attributes: 9
Sum of weights: 768

Attributes

No.	Name
1	□ pregnant
2	□ glucose
3	□ bp
4	□ triceps
5	□ insulin
6	□ bmi
7	□ pedigree
8	□ age
9	□ class

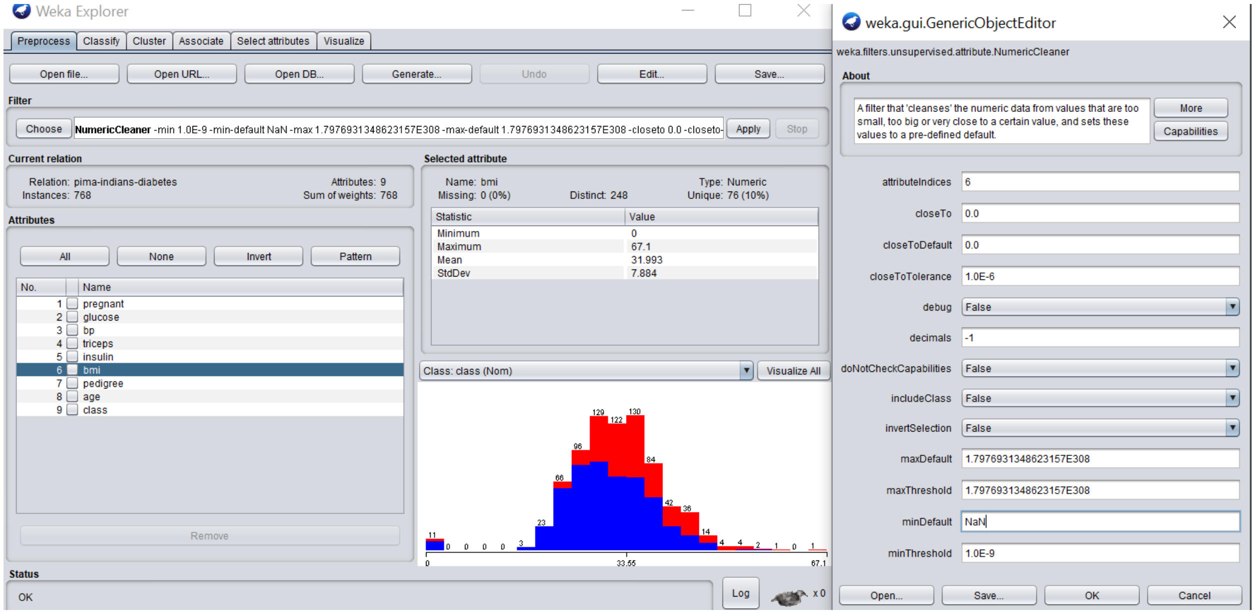
Selected attribute

Name: bmi
Missing: 0 (0%)
Distinct: 248
Type: Numeric
Unique: 76 (10%)

Statistic	Value
Minimum	0
Maximum	67.1
Mean	31.993
StdDev	7.884

c) Proposed Solution

The first step to clean this data would be to identify these instances as having missing values. I did this using the unsupervised.attribute.NumericalCleaner filter. I set the attributeIndicies to 6 to specify the bmi attribute and changed the minThreshold to slightly above 0. The minDefault was changed to NaN as that is unknown:



Filter

NumericCleaner -min 1.0E-9 -min-default NaN -max 1.7976931348623157E308 -max-default 1.7976931348623157E308 -closeto 0.0 -closeto-

Current relation

Relation: pima-indians-diabetes
Instances: 768
Attributes: 9
Sum of weights: 768

Attributes

No.	Name
1	pregnant
2	glucose
3	bp
4	triceps
5	insulin
6	bmi
7	pedigree
8	age
9	class

Selected attribute

Name: bmi
Missing: 0 (0%)
Distinct: 248
Type: Numeric
Unique: 76 (10%)

Statistic	Value
Minimum	0
Maximum	67.1
Mean	31.993
StdDev	7.884

Status

OK

weka.gui.GenericObjectEditor

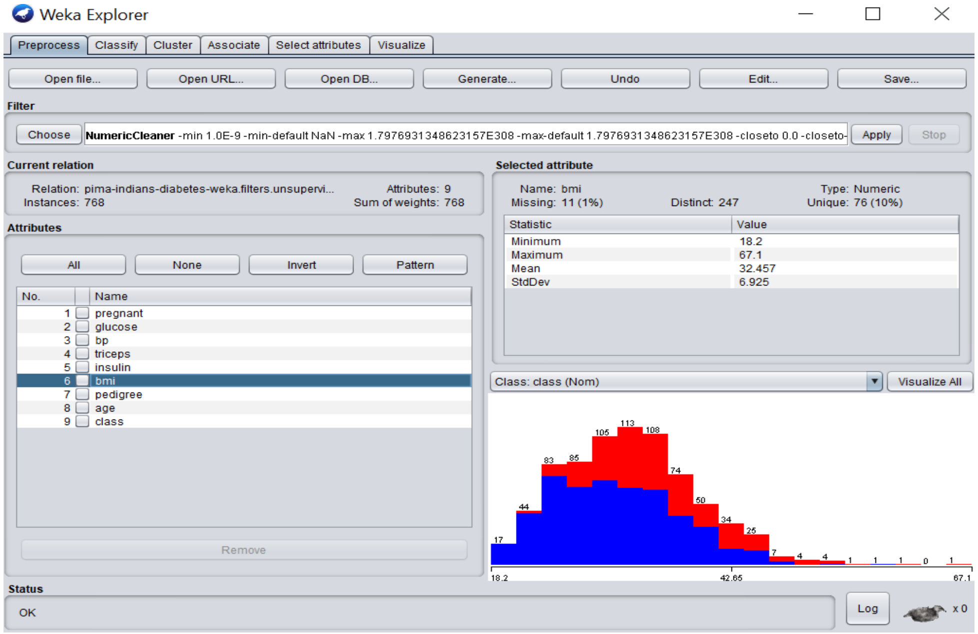
weka.filters.unsupervised.attribute.NumericCleaner

About

A filter that 'cleanses' the numeric data from values that are too small, too big or very close to a certain value, and sets these values to a pre-defined default.

attributeIndices	6
closeTo	0.0
closeToDefault	0.0
closeToTolerance	1.0E-6
debug	False
decimals	-1
doNotCheckCapabilities	False
includeClass	False
invertSelection	False
maxDefault	1.7976931348623157E308
maxThreshold	1.7976931348623157E308
minDefault	NaN
minThreshold	1.0E-9

After applying the filter, I was able to identify 11 attribute values as missing which were formally set to 0 :



Filter

NumericCleaner -min 1.0E-9 -min-default NaN -max 1.7976931348623157E308 -max-default 1.7976931348623157E308 -closeto 0.0 -closeto-

Current relation

Relation: pima-indians-diabetes-weka.filters.unsupervi...
Instances: 768
Attributes: 9
Sum of weights: 768

Attributes

No.	Name
1	pregnant
2	glucose
3	bp
4	triceps
5	insulin
6	bmi
7	pedigree
8	age
9	class

Selected attribute

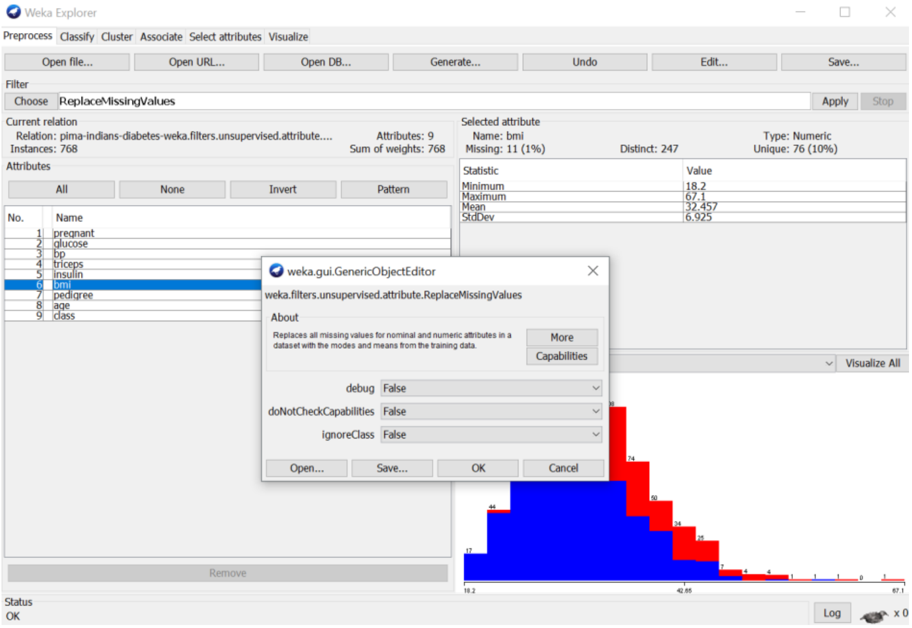
Name: bmi
Missing: 11 (1%)
Distinct: 247
Type: Numeric
Unique: 76 (10%)

Statistic	Value
Minimum	18.2
Maximum	67.1
Mean	32.457
StdDev	6.925

Status

OK

Now that I have the missing data identified, I must decide on a strategy to handle them. Two obvious methods would be to either remove them or impute them with the mean. Because we are dealing with population data, I think it would make sense to replace the missing values with the mean of the bmi attribute. This will ensure the results will not be affected much by the artificial values we impute. I did this using the unsupervised.attribute.ReplaceMissingValues:



Filter
Choose ReplaceMissingValues

Current relation
Relation: pima-indians-diabetes-weka.filters.unsupervised.attribute.... Attributes: 9
Instances: 768 Sum of weights: 768

Selected attribute
Name: bmi Type: Numeric
Missing: 11 (1%) Distinct: 247 Unique: 76 (10%)

Statistic	Value
Minimum	18.2
Maximum	67.1
Mean	32.457
StdDev	6.925

Attributes

No.	Name
1	pregnant
2	glucose
3	bp
4	triceps
5	insulin
6	bmi
7	pedigree
8	age
9	class

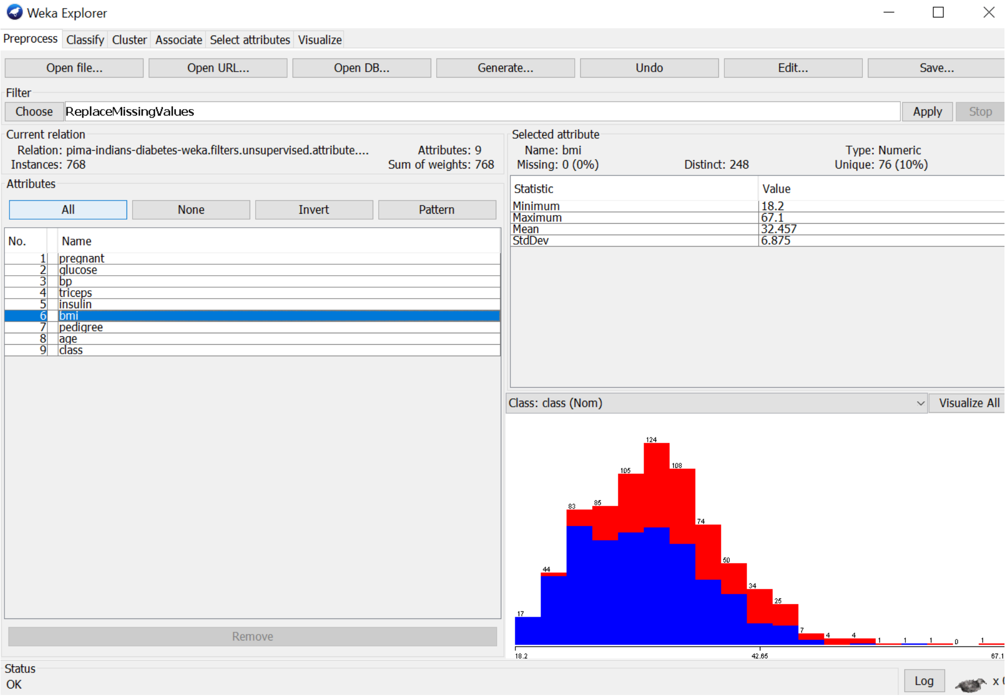
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.ReplaceMissingValues

About
Replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data.

debug False
doNotCheckCapabilities False
ignoreClass False

Status
OK

After applying the filter, I was able impute the mean for the missing values for bmi attribute:



Filter
Choose ReplaceMissingValues

Current relation
Relation: pima-indians-diabetes-weka.filters.unsupervised.attribute.... Attributes: 9
Instances: 768 Sum of weights: 768

Selected attribute
Name: bmi Type: Numeric
Missing: 0 (0%) Distinct: 248 Unique: 76 (10%)

Statistic	Value
Minimum	18.2
Maximum	67.1
Mean	32.457
StdDev	6.875

Attributes

No.	Name
1	pregnant
2	glucose
3	bp
4	triceps
5	insulin
6	bmi
7	pedigree
8	age
9	class

Class: class (Nom)

Status
OK

I exported this cleaned dataset into a new CSV.

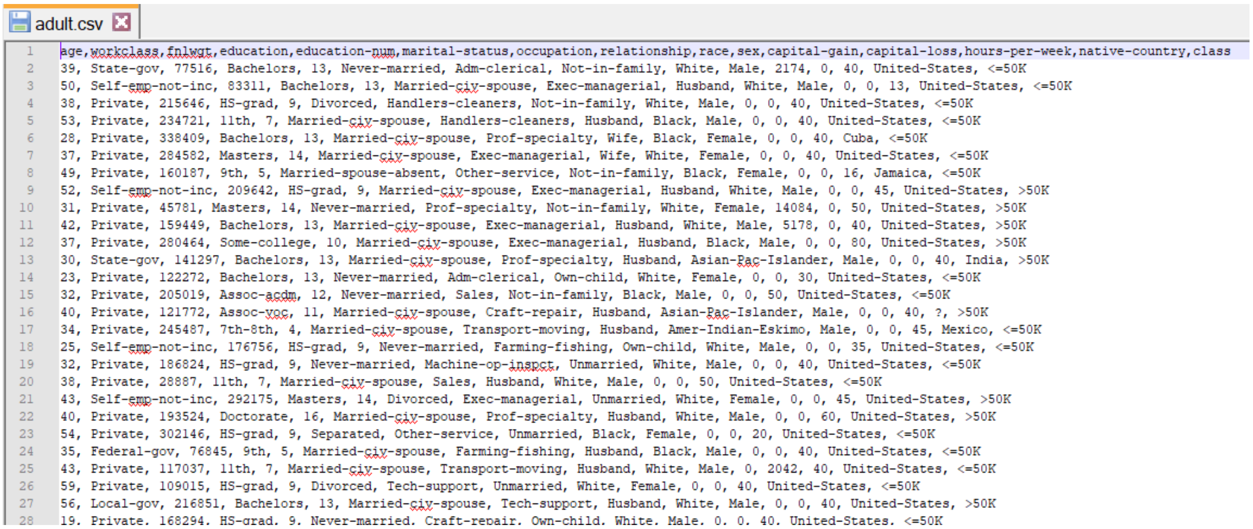
Adult Income Database

a) Data Selected

- Number of Attributes = 14 + class
  - Age (years)
  - Workclass (categorical)
  - Final Weight (continuous)
  - Education (categorical)
  - Education-num (continuous)
  - Martial-status (categorical)
  - Occupation (categorical)
  - Relationship (categorical)
  - Race (categorical)
  - Sex (categorical)
  - Capital-gain (continuous)
  - Capital-loss (continuous)
  - Hours-per-week (continuous)
  - Native-country (categorical)
  - Class (>50K, <=50K)

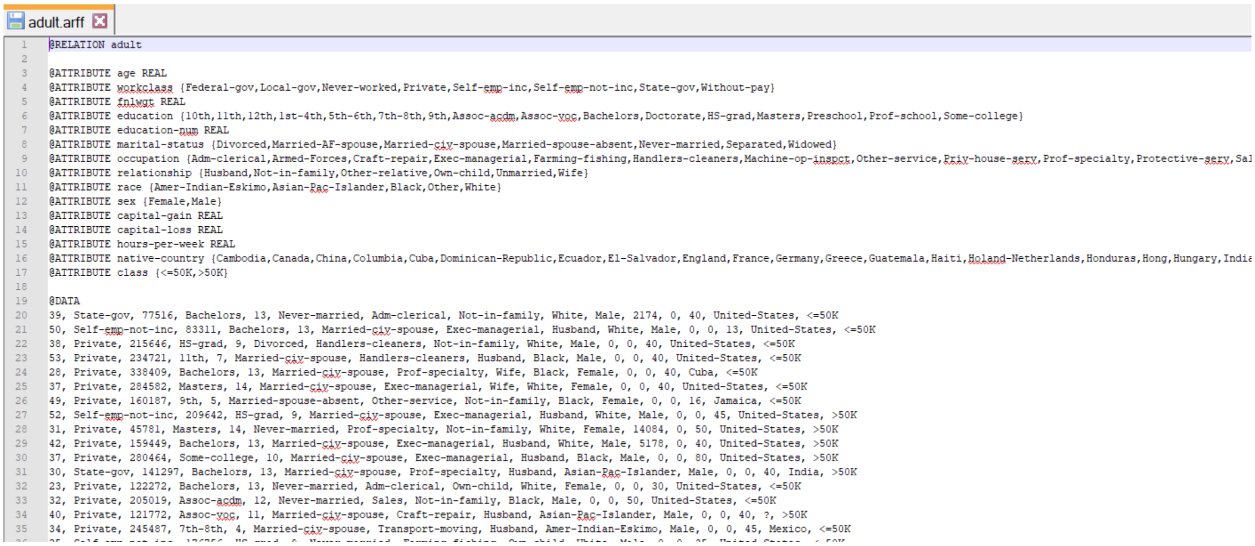
b) Number of Instances = 32561

I downloaded this dataset in .data and .name file types. I first converted this information into .csv:



```
age,workclass,fnlwgt,education,education-num,marital-status,occupation,relationship,race,sex,capital-gain,capital-loss,hours-per-week,native-country,class
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, <=50K
53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K
28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K
37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K
49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female, 0, 0, 16, Jamaica, <=50K
52, Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 45, United-States, >50K
31, Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, White, Female, 14084, 0, 50, United-States, >50K
42, Private, 159449, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 5178, 0, 40, United-States, >50K
37, Private, 280464, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Black, Male, 0, 0, 80, United-States, >50K
30, State-gov, 141297, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, Asian-Pac-Islander, Male, 0, 0, 40, India, >50K
23, Private, 122272, Bachelors, 13, Never-married, Adm-clerical, Own-child, White, Female, 0, 0, 30, United-States, <=50K
32, Private, 205019, Assoc-acdm, 12, Never-married, Sales, Not-in-family, Black, Male, 0, 0, 50, United-States, <=50K
40, Private, 121772, Assoc-voc, 11, Married-civ-spouse, Craft-repair, Husband, Asian-Pac-Islander, Male, 0, 0, 40, ?, >50K
34, Private, 245487, 7th-8th, 4, Married-civ-spouse, Transport-moving, Husband, Amer-Indian-Eskimo, Male, 0, 0, 45, Mexico, <=50K
25, Self-emp-not-inc, 176756, HS-grad, 9, Never-married, Farming-fishing, Own-child, White, Male, 0, 0, 35, United-States, <=50K
32, Private, 186824, HS-grad, 9, Never-married, Machine-op-inspct, Unmarried, White, Male, 0, 0, 40, United-States, <=50K
38, Private, 28887, 11th, 7, Married-civ-spouse, Sales, Husband, White, Male, 0, 0, 50, United-States, <=50K
43, Self-emp-not-inc, 292175, Masters, 14, Divorced, Exec-managerial, Unmarried, White, Female, 0, 0, 45, United-States, >50K
40, Private, 193524, Doctorate, 16, Married-civ-spouse, Prof-specialty, Husband, White, Male, 0, 0, 60, United-States, >50K
54, Private, 302146, HS-grad, 9, Separated, Other-service, Unmarried, Black, Female, 0, 0, 20, United-States, <=50K
35, Federal-gov, 76845, 9th, 5, Married-civ-spouse, Farming-fishing, Husband, Black, Male, 0, 0, 40, United-States, <=50K
43, Private, 117037, 11th, 7, Married-civ-spouse, Transport-moving, Husband, White, Male, 0, 2042, 40, United-States, <=50K
59, Private, 109015, HS-grad, 9, Divorced, Tech-support, Unmarried, White, Female, 0, 0, 40, United-States, <=50K
56, Local-gov, 216851, Bachelors, 13, Married-civ-spouse, Tech-support, Husband, White, Male, 0, 0, 40, United-States, >50K
19, Private, 168294, HS-grad, 9, Never-married, Craft-repair, Own-child, White, Male, 0, 0, 40, United-States, <=50K
```

From my csv file I converted the file into .arff for ease of use with Weka 3.8:



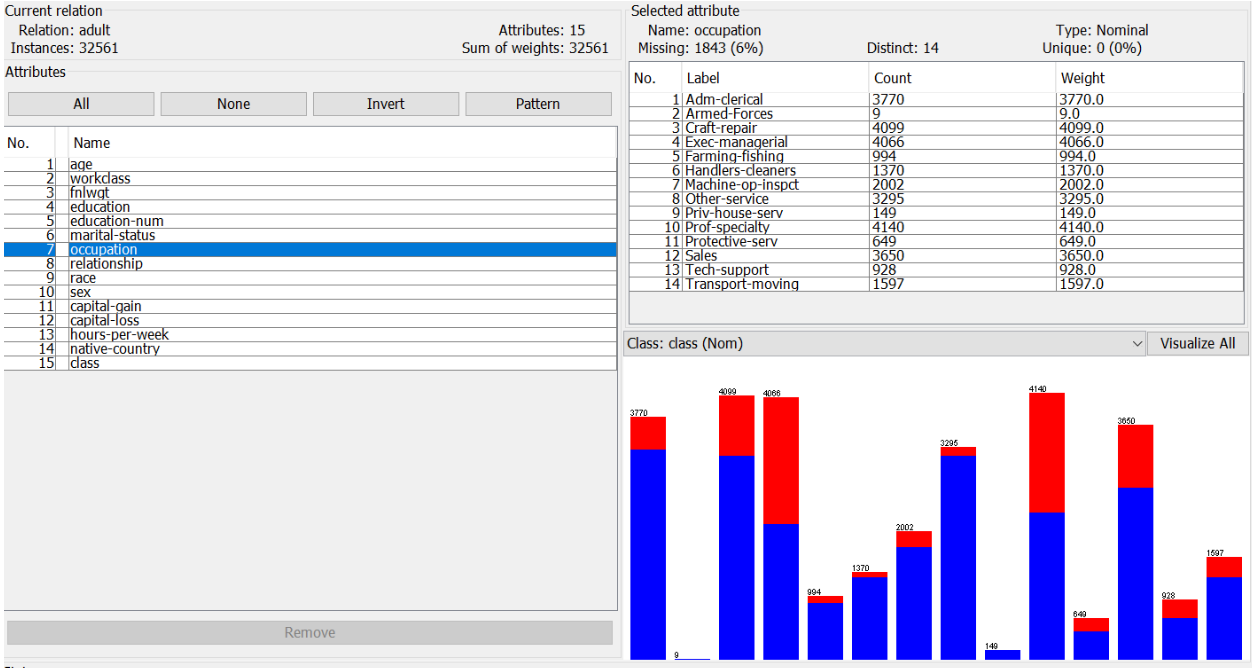
```
@RELATION adult

@ATTRIBUTE age REAL
@ATTRIBUTE workclass {Federal-gov,Local-gov,Never-worked,Private,Self-emp-inc,Self-emp-not-inc,State-gov,Without-pay}
@ATTRIBUTE fnlwgt REAL
@ATTRIBUTE education {10th,11th,12th,1st-4th,5th-6th,7th-8th,9th,Assoc-acdm,Assoc-voc,Bachelors,Doctorate,HS-grad,Masters,Preschool,Prof-school,Some-college}
@ATTRIBUTE education-num REAL
@ATTRIBUTE marital-status {Divorced,Married-AF-spouse,Married-civ-spouse,Married-spouse-absent,Never-married,Separated,Widowed}
@ATTRIBUTE occupation {Adm-clerical,Armed-Forces,Craft-repair,Exec-managerial,Farming-fishing,Handlers-cleaners,Machine-op-inspct,Other-service,Priv-house-serv,Prof-specialty,Protective-serv,Sal
@ATTRIBUTE relationship {Husband,Not-in-family,Other-relative,Own-child,Unmarried,Wife}
@ATTRIBUTE race {Amer-Indian-Eskimo,Asian-Pac-Islander,Black,Other,White}
@ATTRIBUTE sex {Female,Male}
@ATTRIBUTE capital-gain REAL
@ATTRIBUTE capital-loss REAL
@ATTRIBUTE hours-per-week REAL
@ATTRIBUTE native-country {Cambodia,Canada,China,Columbia,Cuba,Dominican-Republic,Ecuador,El-Salvador,England,France,Germany,Greece,Guatemala,Haiti,Holand-Netherlands,Honduras,Hong,Hungary,India
@ATTRIBUTE class {<=50K,>50K}

@DATA
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, <=50K
53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K
28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K
37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K
49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female, 0, 0, 16, Jamaica, <=50K
52, Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 45, United-States, >50K
31, Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, White, Female, 14084, 0, 50, United-States, >50K
42, Private, 159449, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 5178, 0, 40, United-States, >50K
37, Private, 280464, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Black, Male, 0, 0, 80, United-States, >50K
30, State-gov, 141297, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, Asian-Pac-Islander, Male, 0, 0, 40, India, >50K
23, Private, 122272, Bachelors, 13, Never-married, Adm-clerical, Own-child, White, Female, 0, 0, 30, United-States, <=50K
32, Private, 205019, Assoc-acdm, 12, Never-married, Sales, Not-in-family, Black, Male, 0, 0, 50, United-States, <=50K
40, Private, 121772, Assoc-voc, 11, Married-civ-spouse, Craft-repair, Husband, Asian-Pac-Islander, Male, 0, 0, 40, ?, >50K
34, Private, 245487, 7th-8th, 4, Married-civ-spouse, Transport-moving, Husband, Amer-Indian-Eskimo, Male, 0, 0, 45, Mexico, <=50K
```

b) Data Quality Problem Identified

After reviewing the data in the Explorer, I realized that the occupation attribute has 1842 missing values which is roughly 6% of the total data set:



Current relation

Relation: adult
Instances: 32561
Attributes: 15
Sum of weights: 32561

Attributes

No.	Name
1	age
2	workclass
3	fnlwgt
4	education
5	education-num
6	marital-status
7	occupation
8	relationship
9	race
10	sex
11	capital-gain
12	capital-loss
13	hours-per-week
14	native-country
15	class

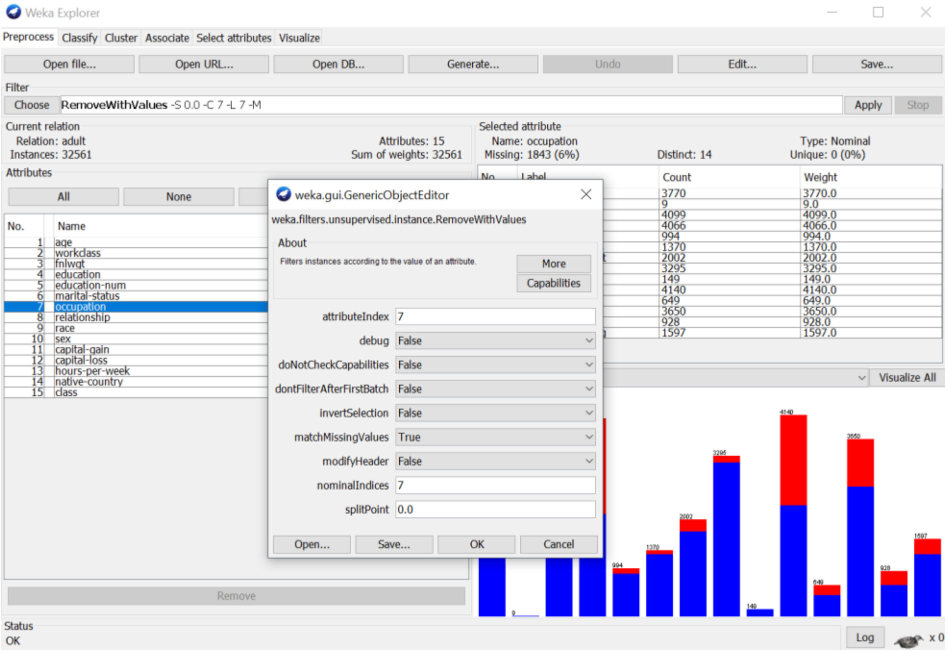
Selected attribute

Name: occupation
Missing: 1843 (6%)
Distinct: 14
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Adm-clerical	3770	3770.0
2	Armed-Forces	9	9.0
3	Craft-repair	4099	4099.0
4	Exec-managerial	4066	4066.0
5	Farming-fishing	994	994.0
6	Handlers-cleaners	1370	1370.0
7	Machine-op-inspct	2002	2002.0
8	Other-service	3295	3295.0
9	Priv-house-serv	149	149.0
10	Prof-specialty	4140	4140.0
11	Protective-serv	649	649.0
12	Sales	3650	3650.0
13	Tech-support	928	928.0
14	Transport-moving	1597	1597.0

c) Proposed Solution

Because this is a nominal attribute, it does not make sense to try and impute them like the other database and therefore I decided to simply remove these records. I did this using the unsupervised.instance.RemoveWithValues filter. I set the attributeIndicies and nomialIndicies to 7, the index of the occupation attribute. I marked 'True' for the matchMissingValues option as we have the instances marked in the dataset already:



Filter: RemoveWithValues -S 0.0 -C 7 -L 7 -M

Current relation
Relation: adult
Instances: 32561
Attributes: 15
Sum of weights: 32561

Selected attribute
Name: occupation
Missing: 1843 (6%)
Distinct: 14
Type: Nominal
Unique: 0 (0%)

Attributes

No.	Name
1	age
2	workclass
3	fnlwgt
4	education
5	education-num
6	marital-status
7	occupation
8	relationship
9	race
10	sex
11	capital-gain
12	capital-loss
13	hours-per-week
14	native-country
15	class

No.	Label	Count	Weight
		3770	3770.0
		9	9.0
		4099	4099.0
		4066	4066.0
		994	994.0
		1370	1370.0
		2002	2002.0
		3295	3295.0
		149	149.0
		4140	4140.0
		649	649.0
		3650	3650.0
		928	928.0
		1597	1597.0

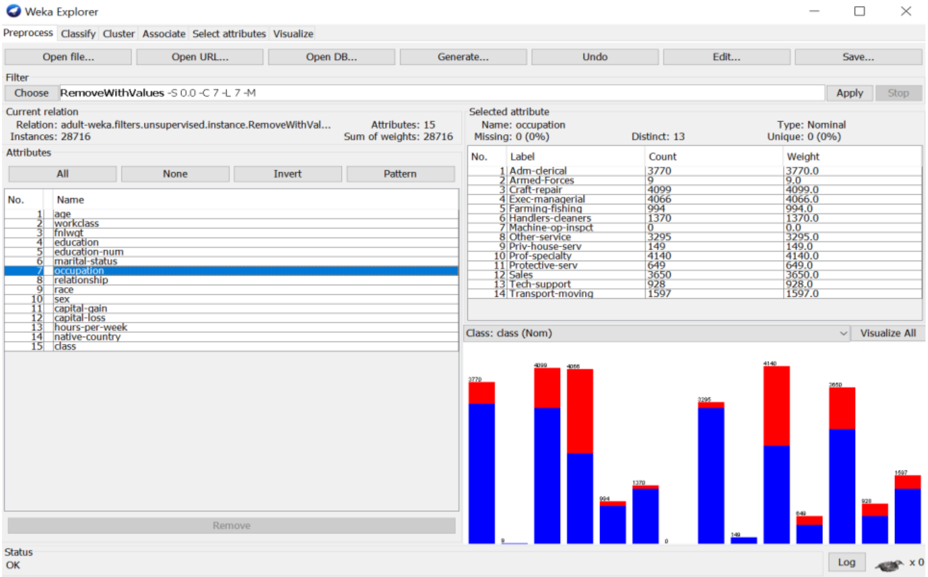
weka.gui.GenericObjectEditor
weka.filters.unsupervised.instance.RemoveWithValues

About
Filters instances according to the value of an attribute.

attributeIndex	7
debug	False
doNotCheckCapabilities	False
dontFilterAfterFirstBatch	False
invertSelection	False
matchMissingValues	True
modifyHeader	False
nominalIndices	7
splitPoint	0.0

Status
OK

After applying the filter, I was able to get rid of the missing values for occupation attribute:



Filter: RemoveWithValues -S 0.0 -C 7 -L 7 -M

Current relation
Relation: adult-weka.filters.unsupervised.instance.RemoveWithVal...
Instances: 28716
Attributes: 15
Sum of weights: 28716

Selected attribute
Name: occupation
Missing: 0 (0%)
Distinct: 13
Type: Nominal
Unique: 0 (0%)

Attributes

No.	Name
1	age
2	workclass
3	fnlwgt
4	education
5	education-num
6	marital-status
7	occupation
8	relationship
9	race
10	sex
11	capital-gain
12	capital-loss
13	hours-per-week
14	native-country
15	class

No.	Label	Count	Weight
1	Adm-clerical	3770	3770.0
2	Armed-Forces	9	9.0
3	Craft-repair	4099	4099.0
4	Exec-managerial	4066	4066.0
5	Farming-fishing	994	994.0
6	Handlers-cleaners	1370	1370.0
7	Machine-op-inspct	0	0.0
8	Other-service	3295	3295.0
9	Priv-house-serv	149	149.0
10	Prof-specialty	4140	4140.0
11	Protective-serv	649	649.0
12	Sales	3650	3650.0
13	Tech-support	928	928.0
14	Transport-moving	1597	1597.0

Class: class (Nom)

Status
OK

Like before, I exported this clean dataset into a new csv file. Obviously, this new file will contain 1842 fewer records than the original database since we have removed some records.