**---------------------------------------------------------------------------------------------**

**DWM EXPERIMENT NO. : 06 (A)** **Roll no :50**

**---**-**-----------------------------------------------------------------------------------------**

**Problem definition :**

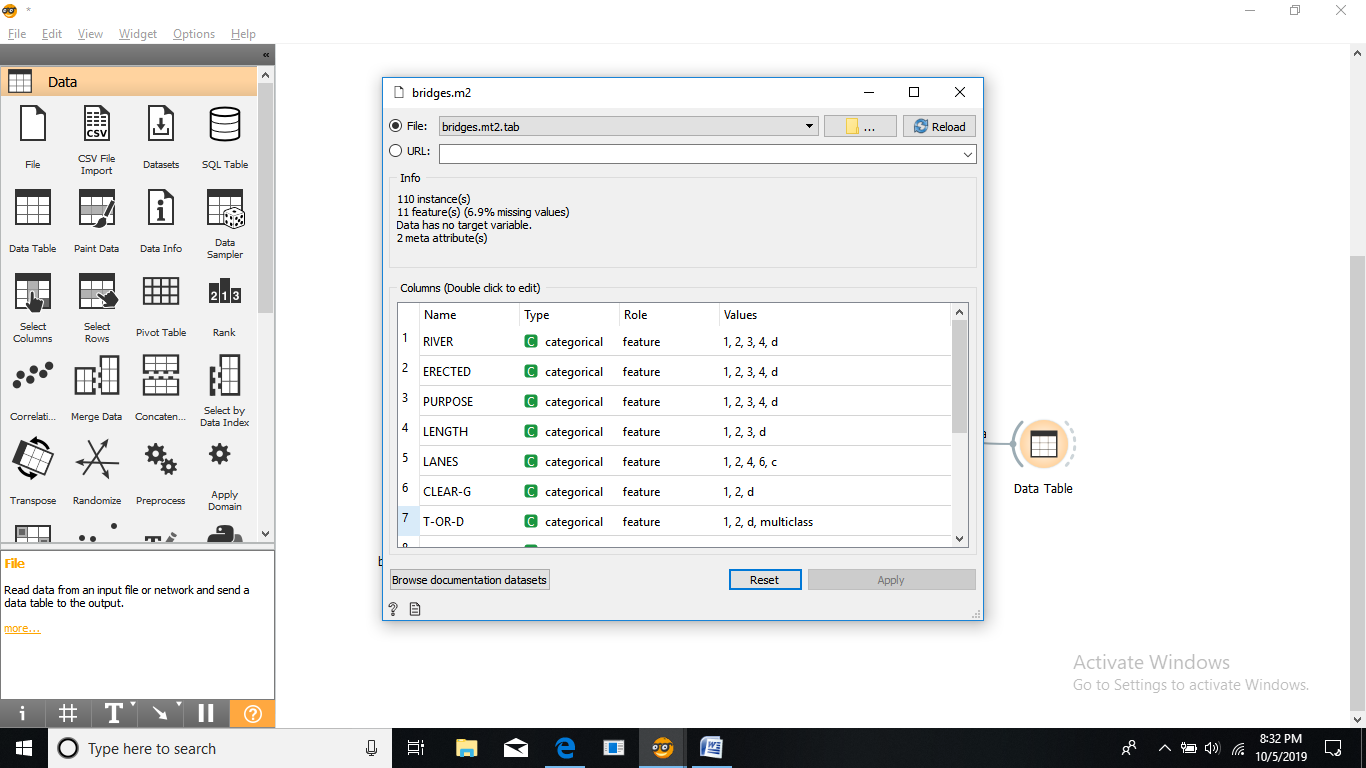
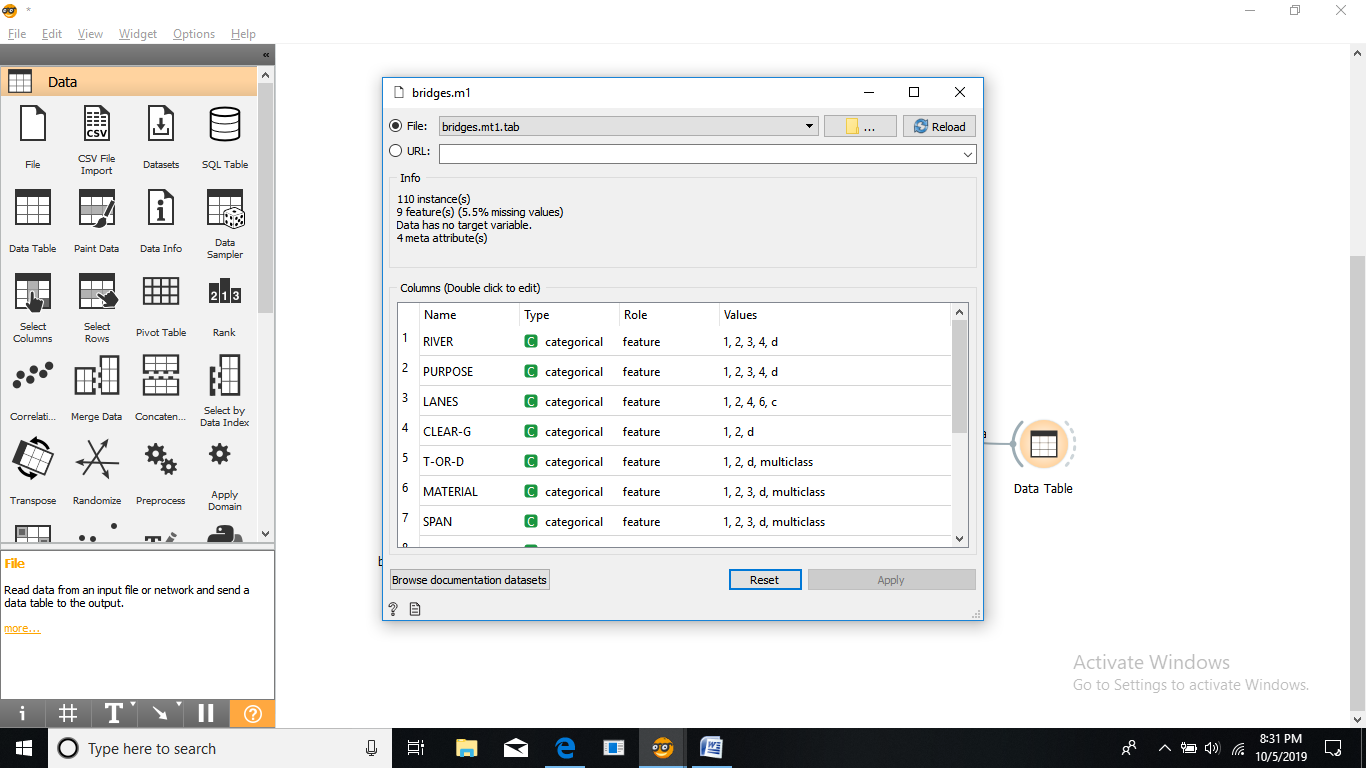
To perform ETL process for building a data warehouse using ORANGE tool.

---------------------------------------------------------------------------------------------

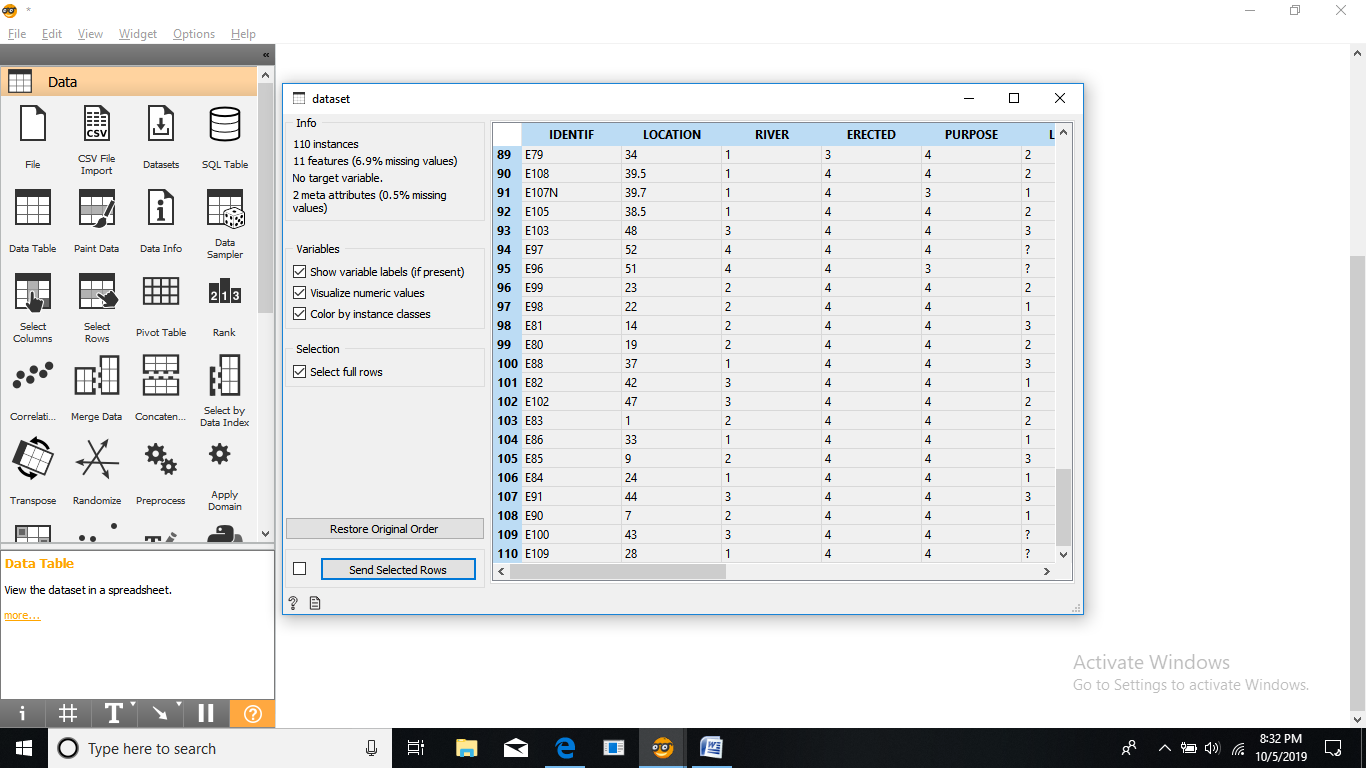
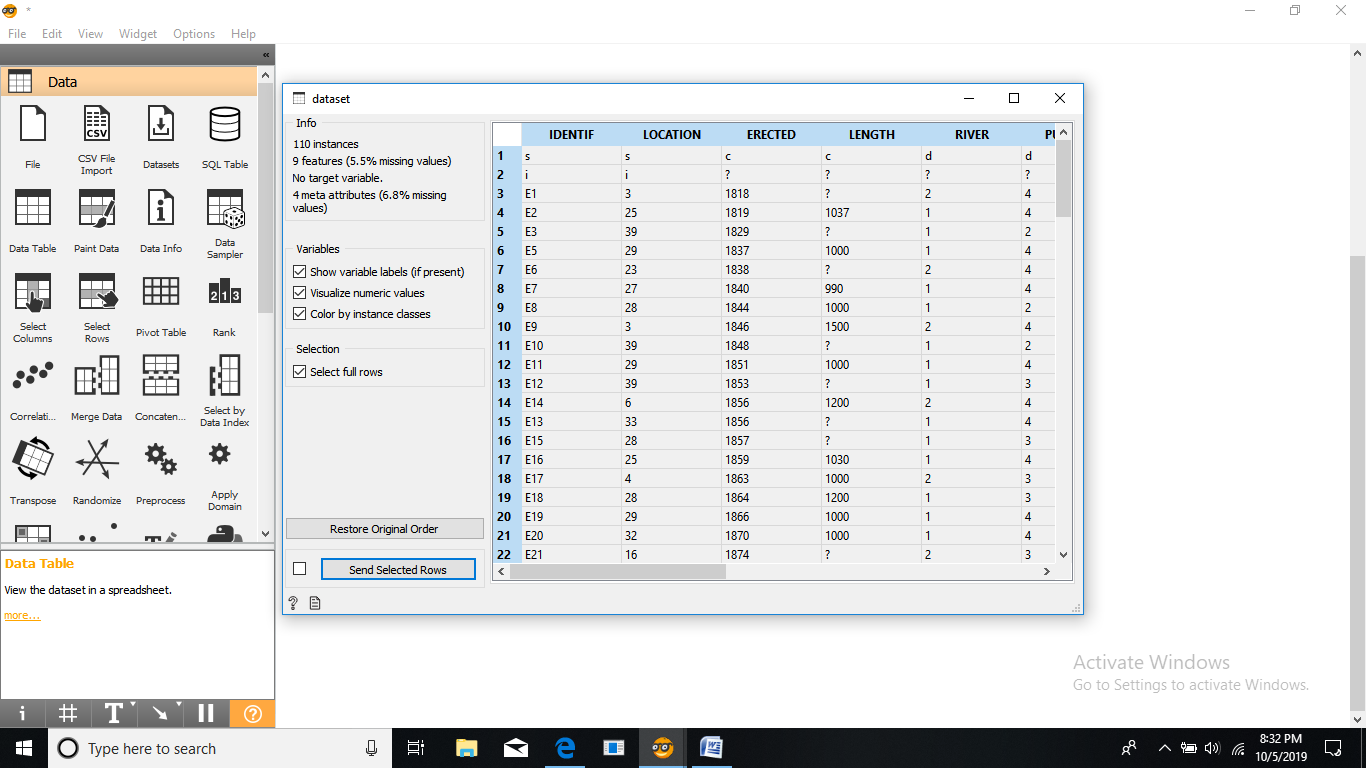
1. Consider bridges.mt1.tab and bridges.mt2.tab files as input and do as directed

* Concatenate the files.
* Apply preprocessing for missing values
* Save the new file
* Open the new file and view it.

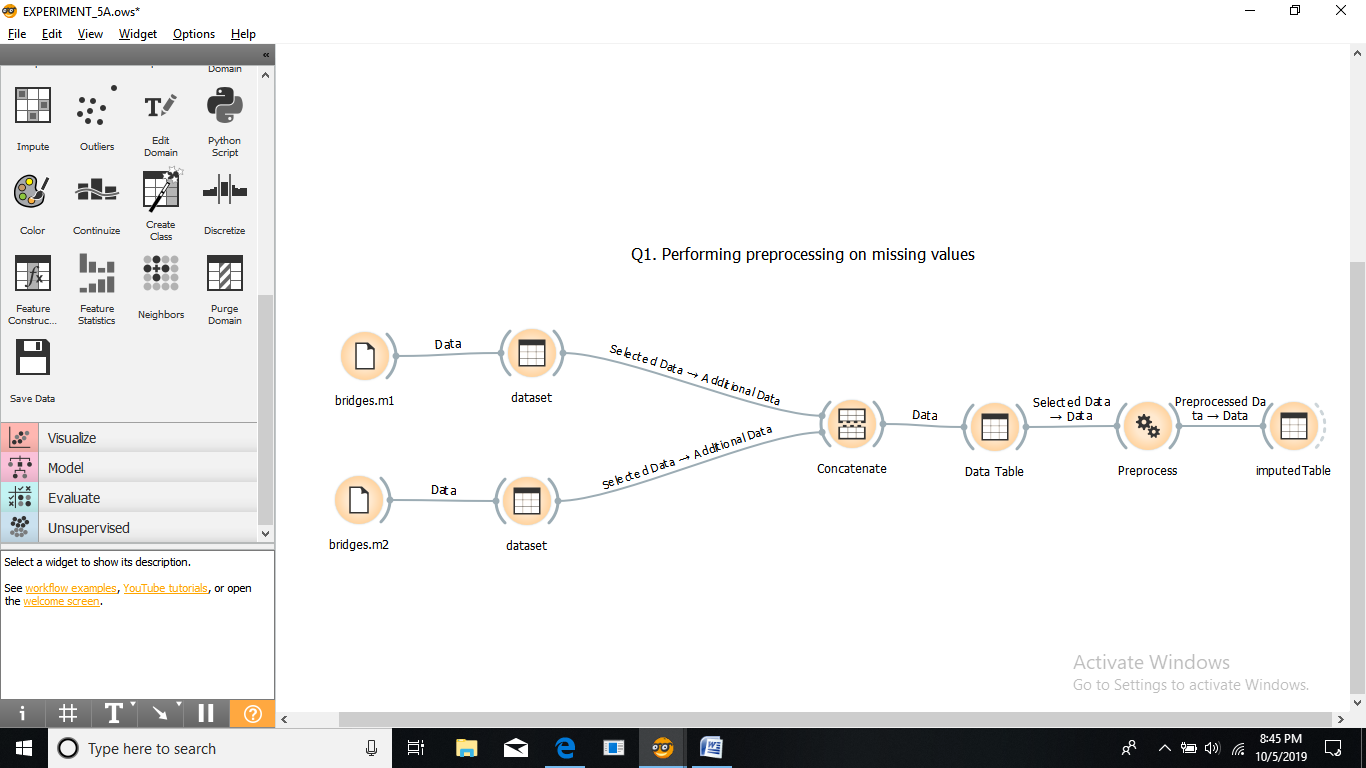
1. **Loading the .tab files.**



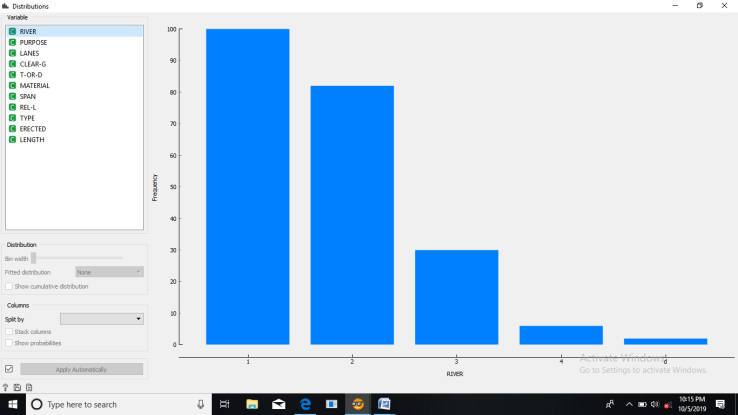
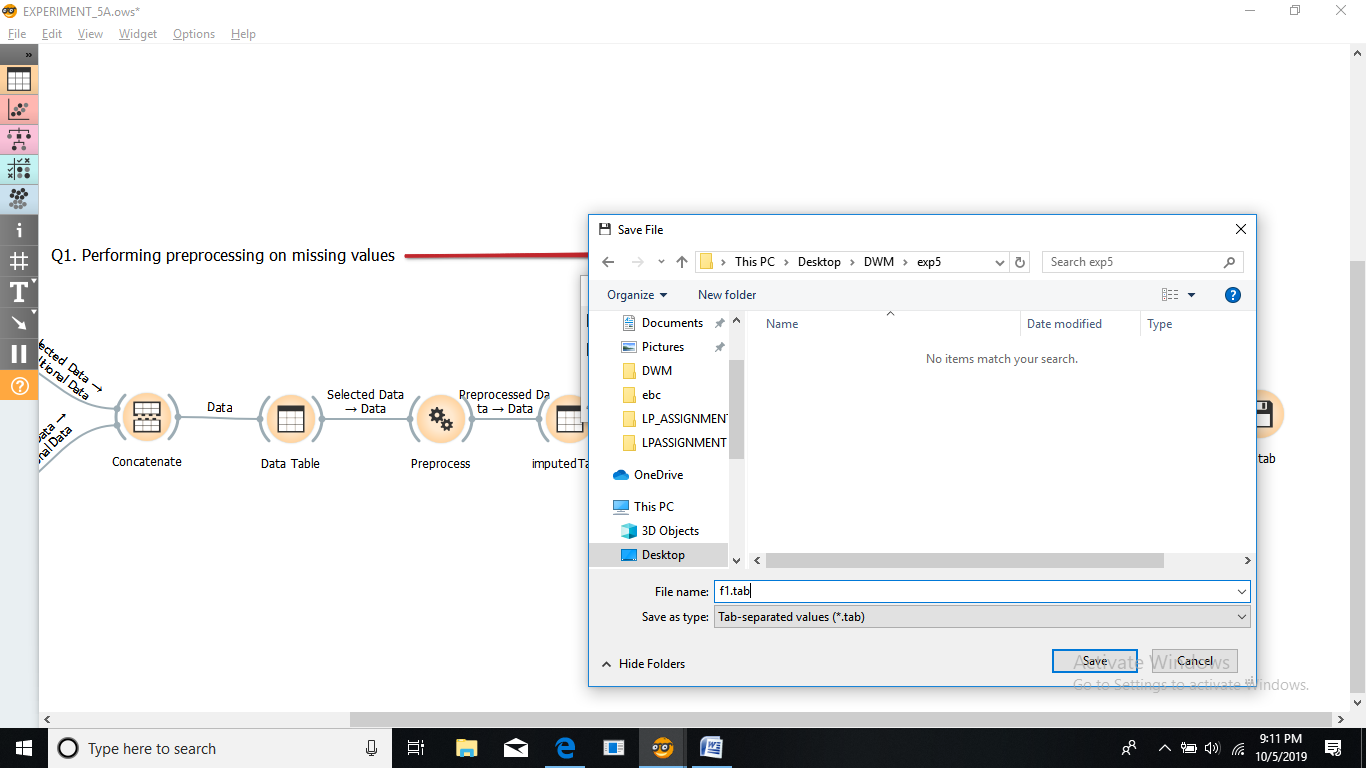
1. **Checking datasets of files.**



1. **Structure of the tools to perform preprocessing.**

****

1. **Saving the result to a tab file & visualizing the result set.**

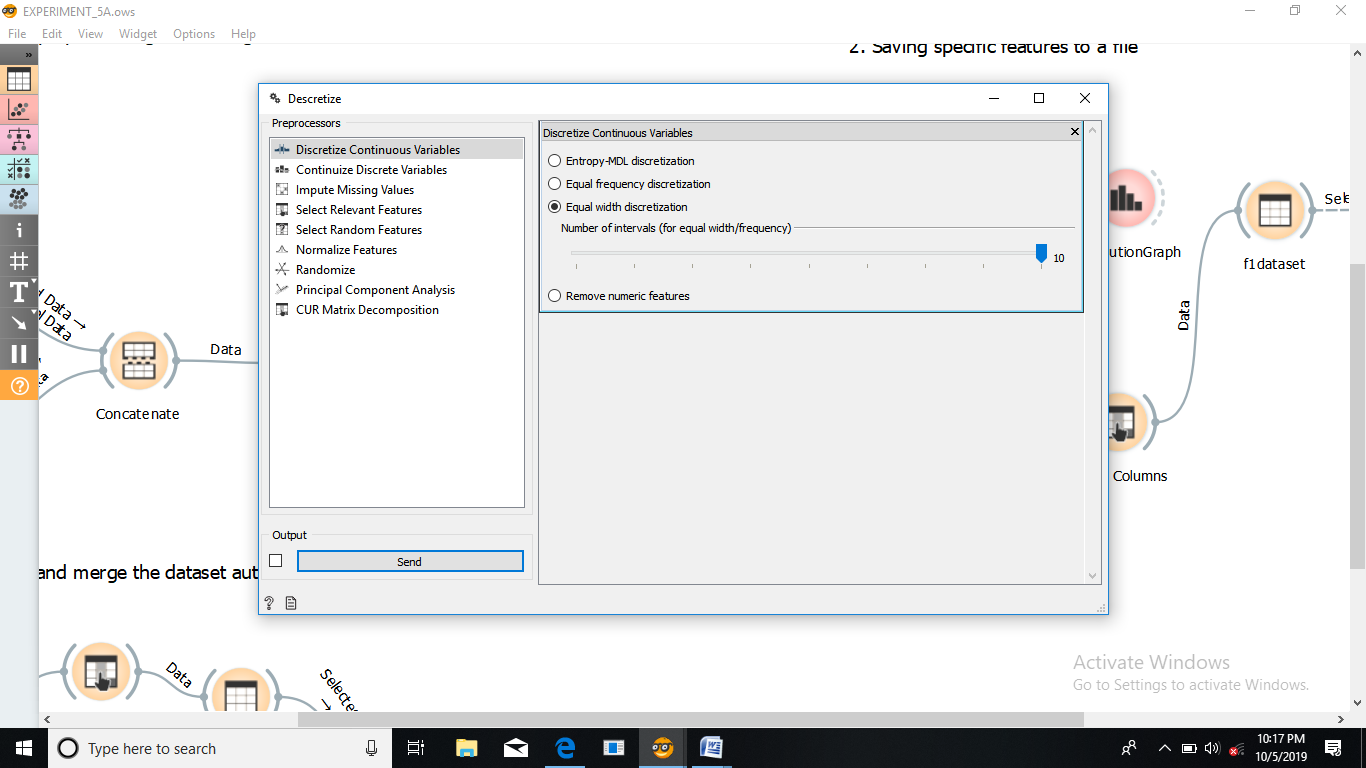
****

**------------------------------------------------------------------------------------**

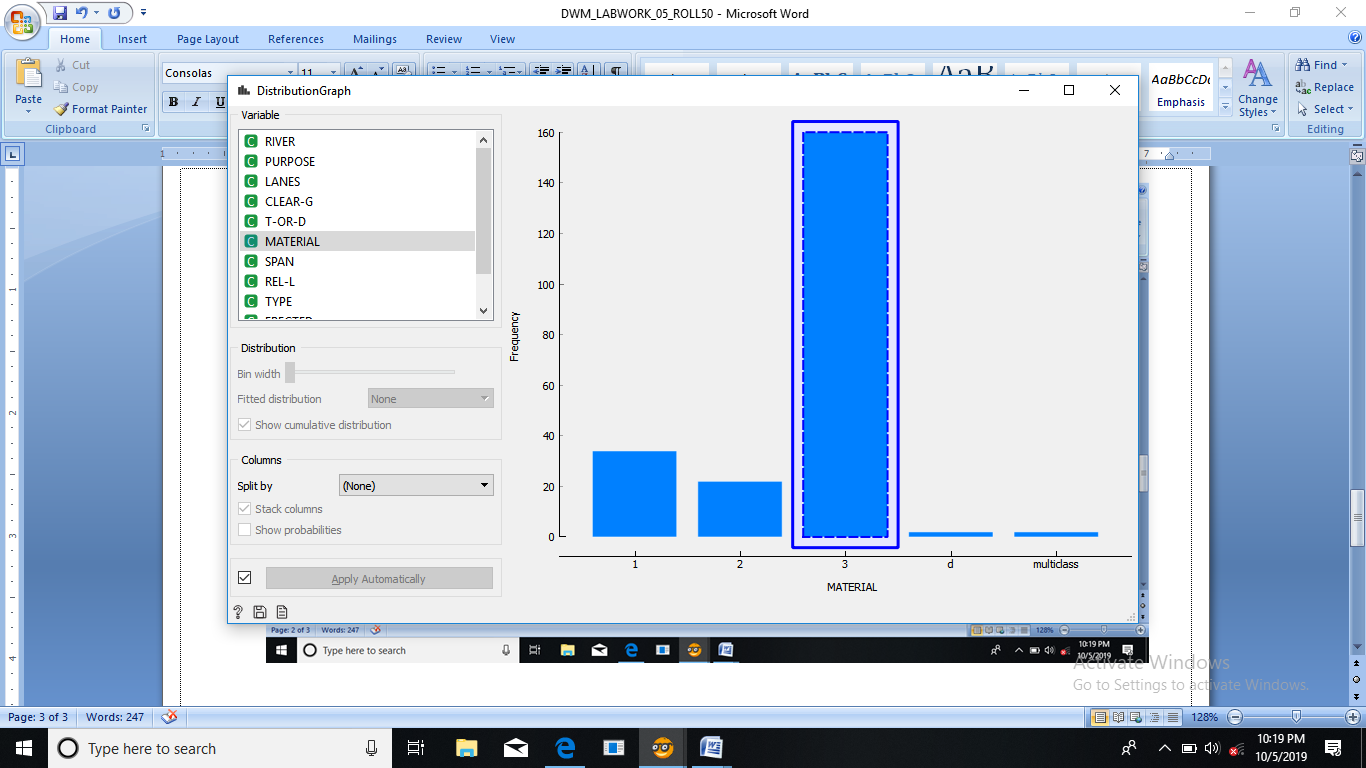
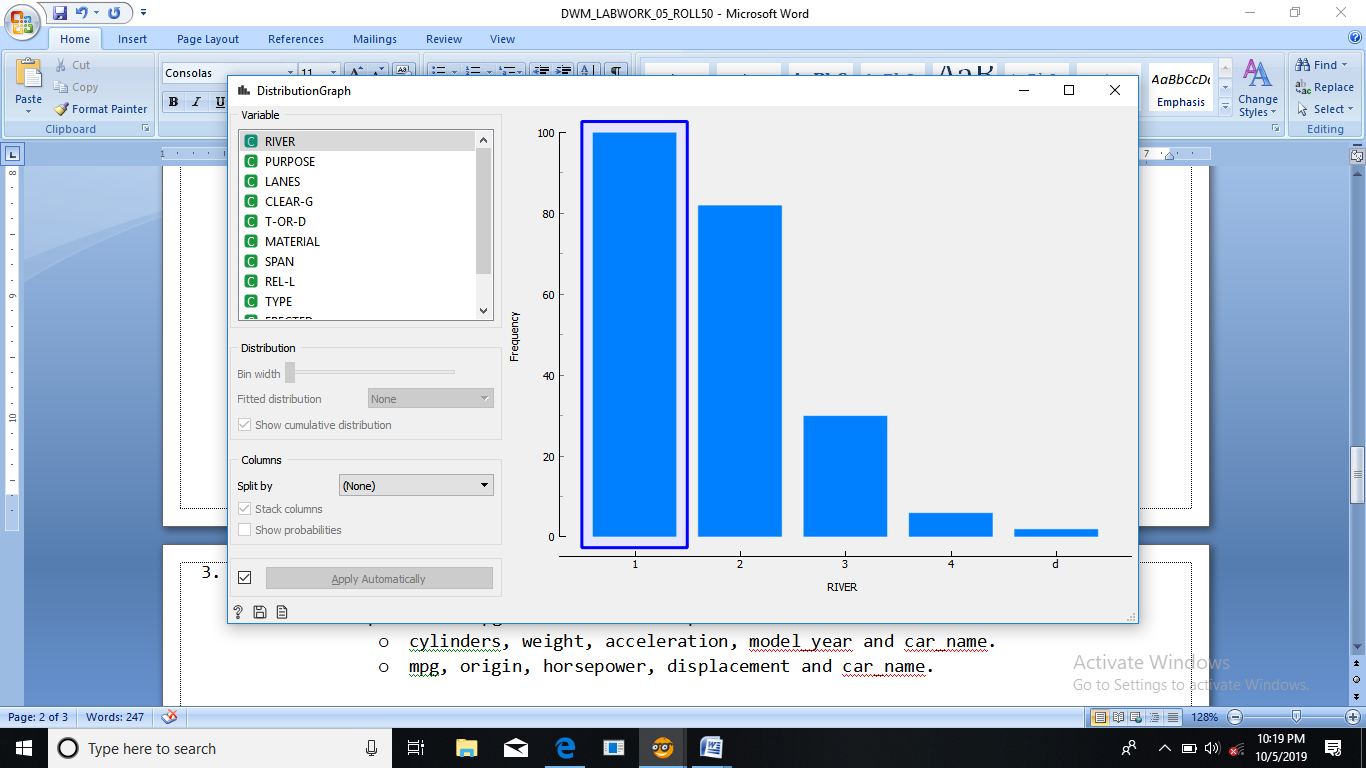
2. Use the above file to Concatenate the files.

* Discretize the variables and find the type of river that is maximum in number, which type of material is majorly used?
* Consider only erected, type, river, purpose and length columns and save it as file- F1.tab

1. **Discretizing the variable.**

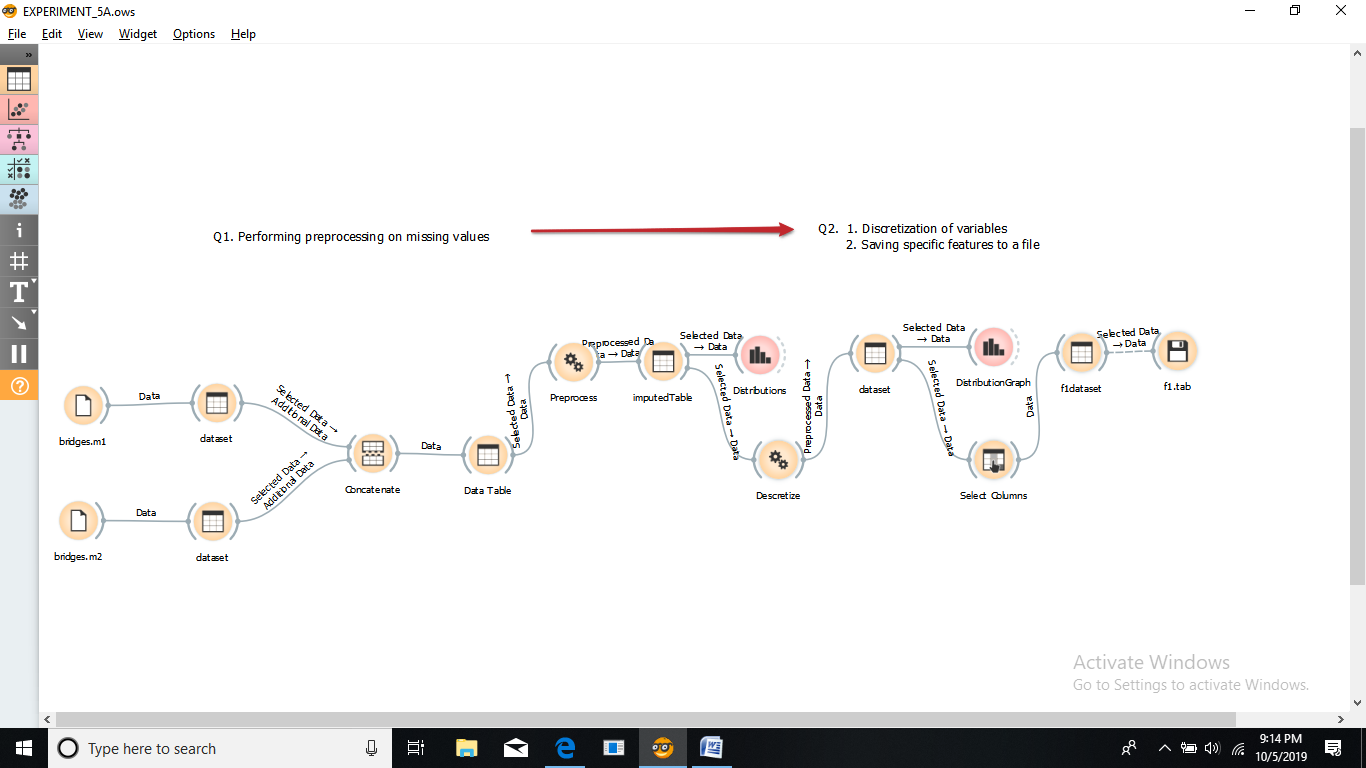


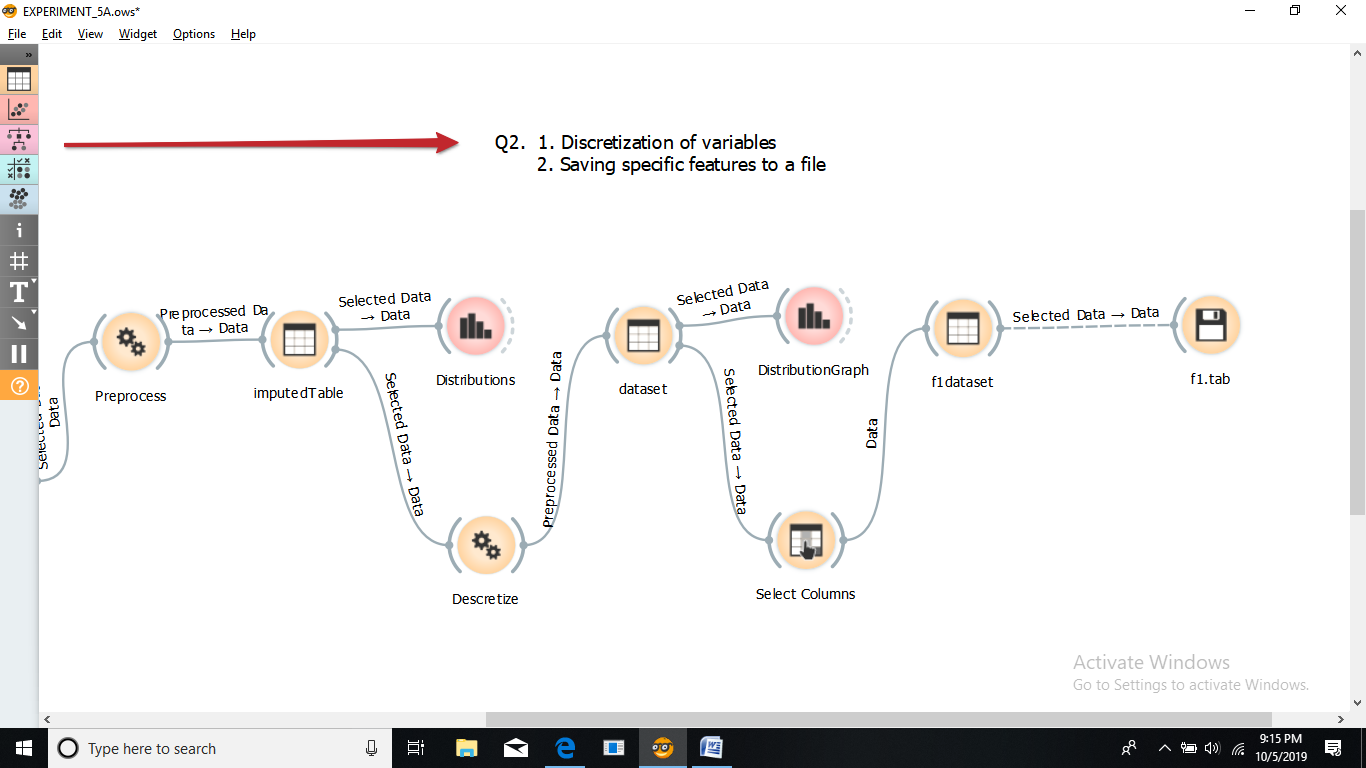
1. **Checking the maximum value in result of River and Material feature.**



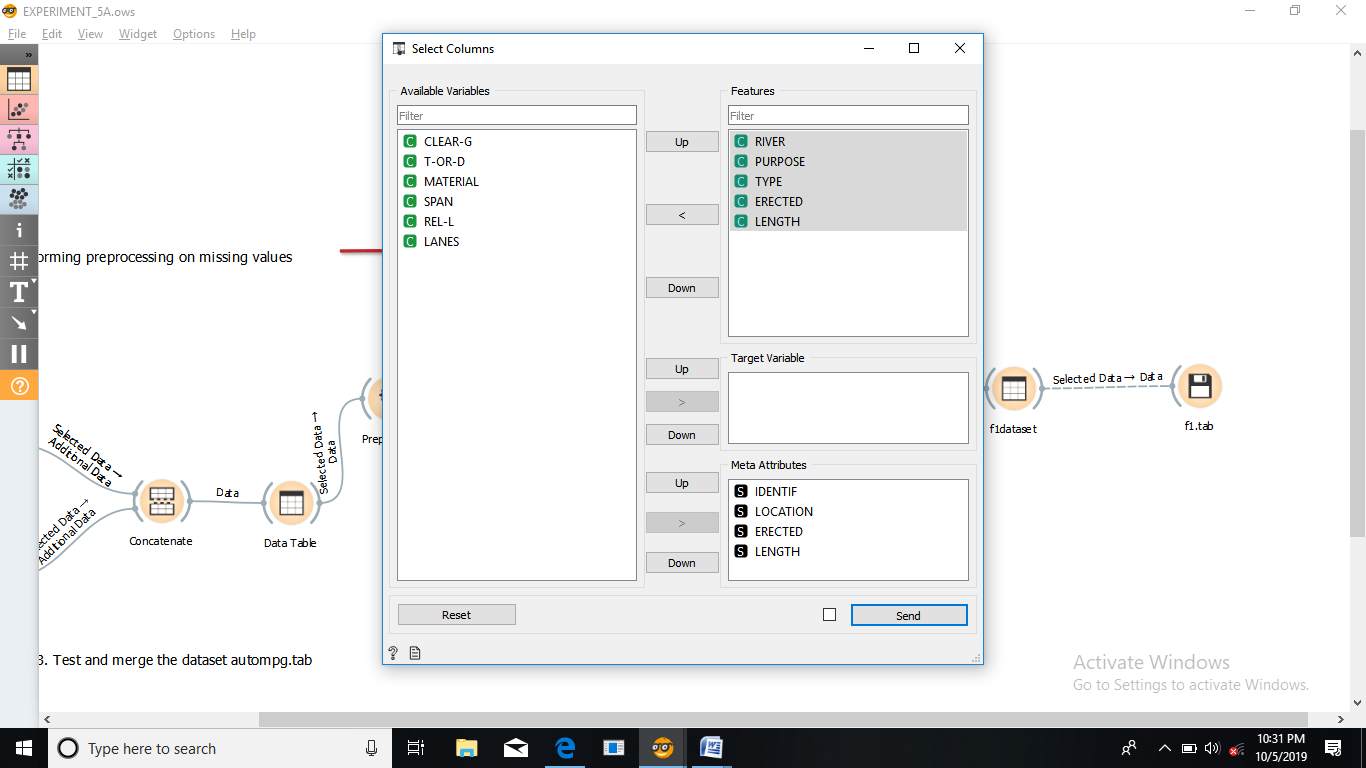
* The type of river that is maximum in number is River “A”.
* Material used majorly is WOOD.

1. **Structure of the tools in this case.**





1. **Choosing the specific features from all the features.**

**** 

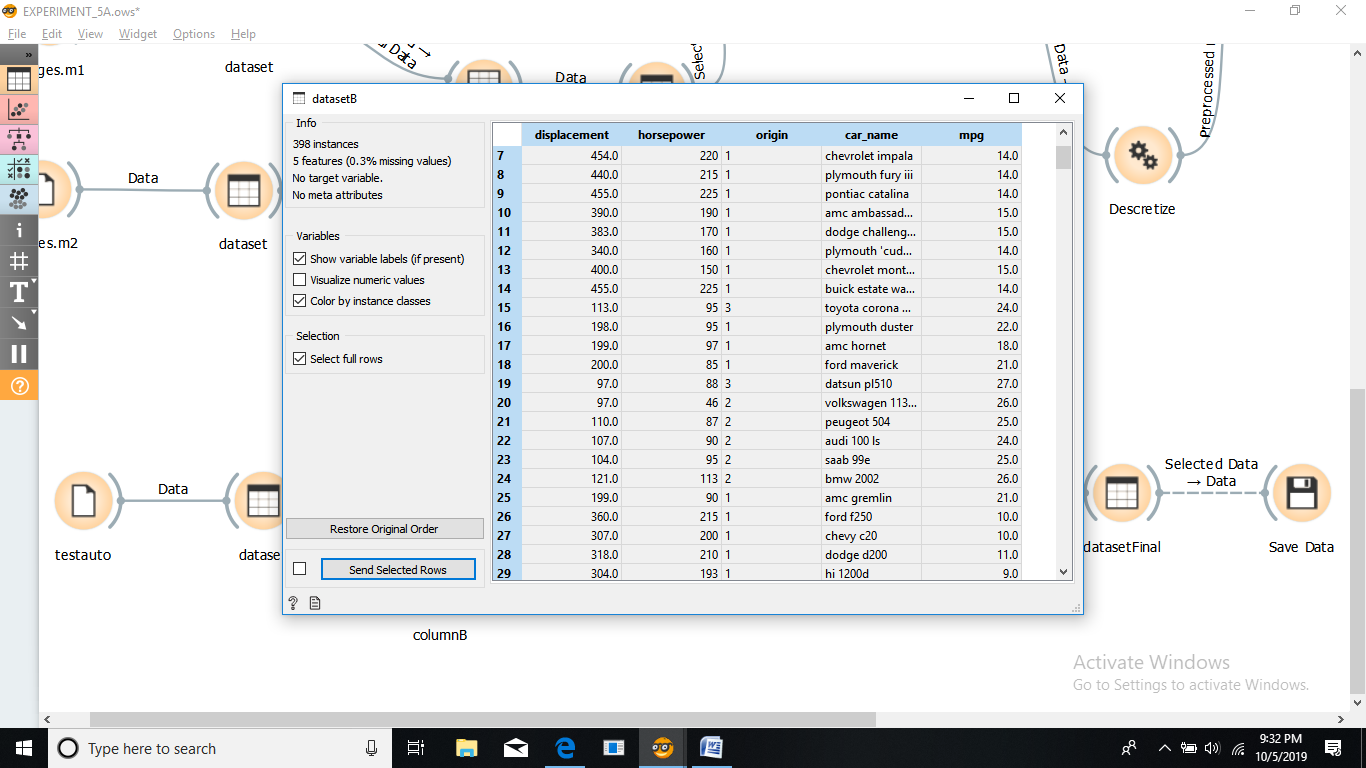
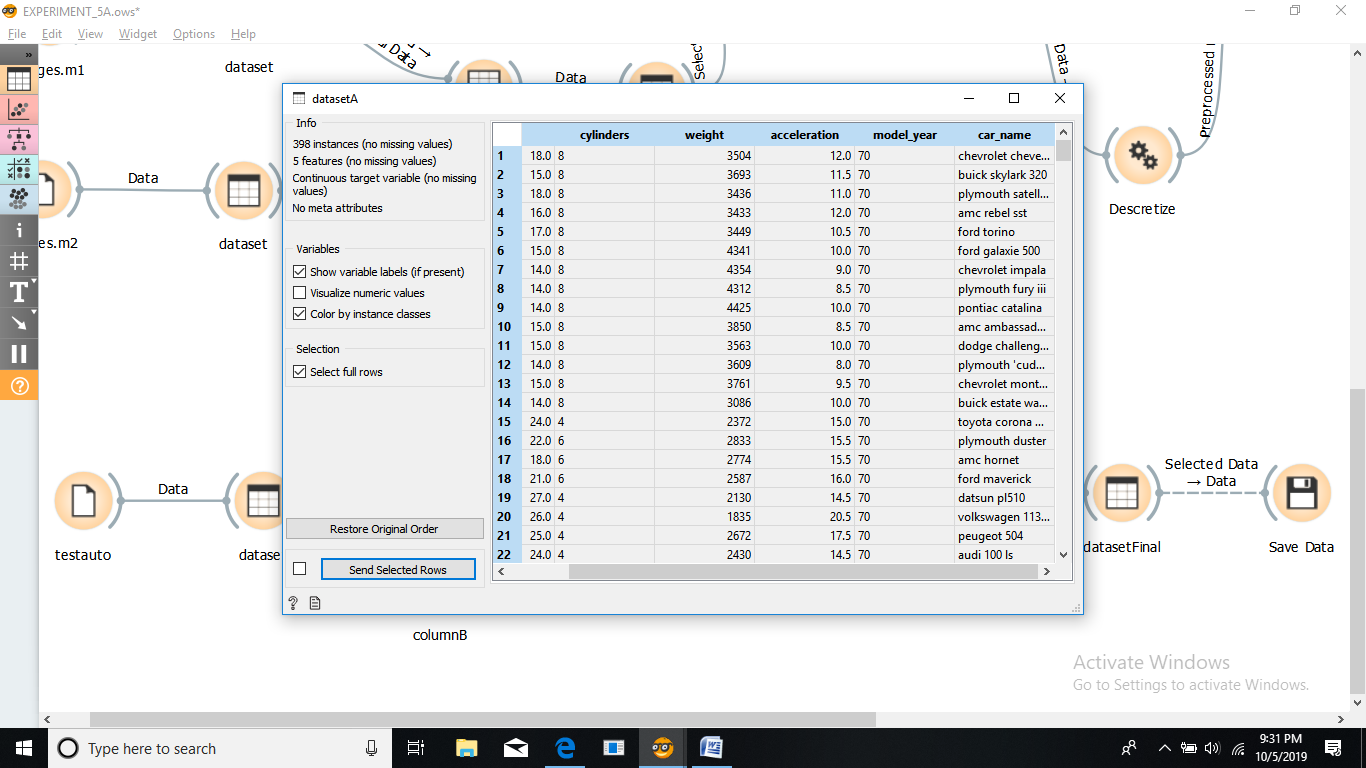
**------------------------------------------------------------------------------------**

3. Do as directed

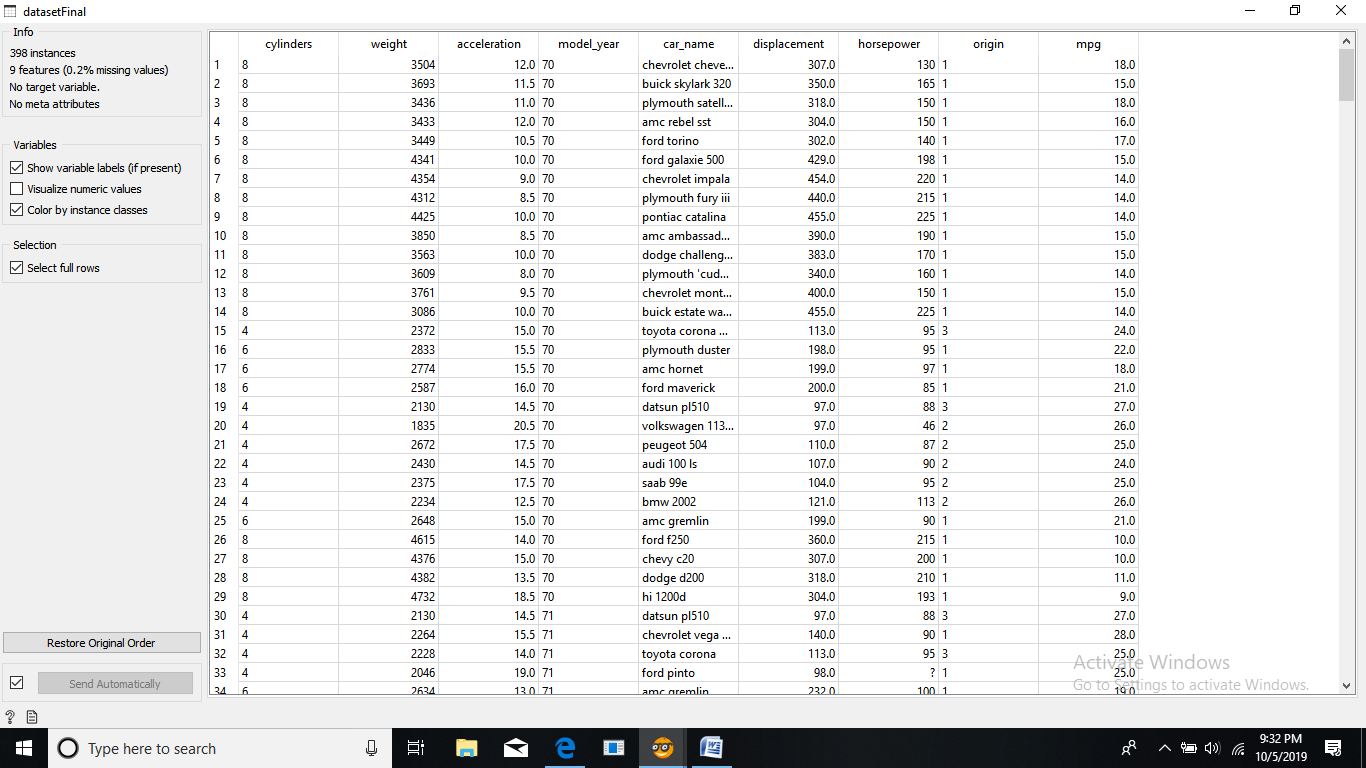
* Import auto-mpg.tab file. Make 2 parts of the file columns.
  + cylinders, weight, acceleration, model\_year and car\_name.
  + mpg, origin, horsepower, displacement and car\_name.

* Consider these 2 parts and merge to get the original file.

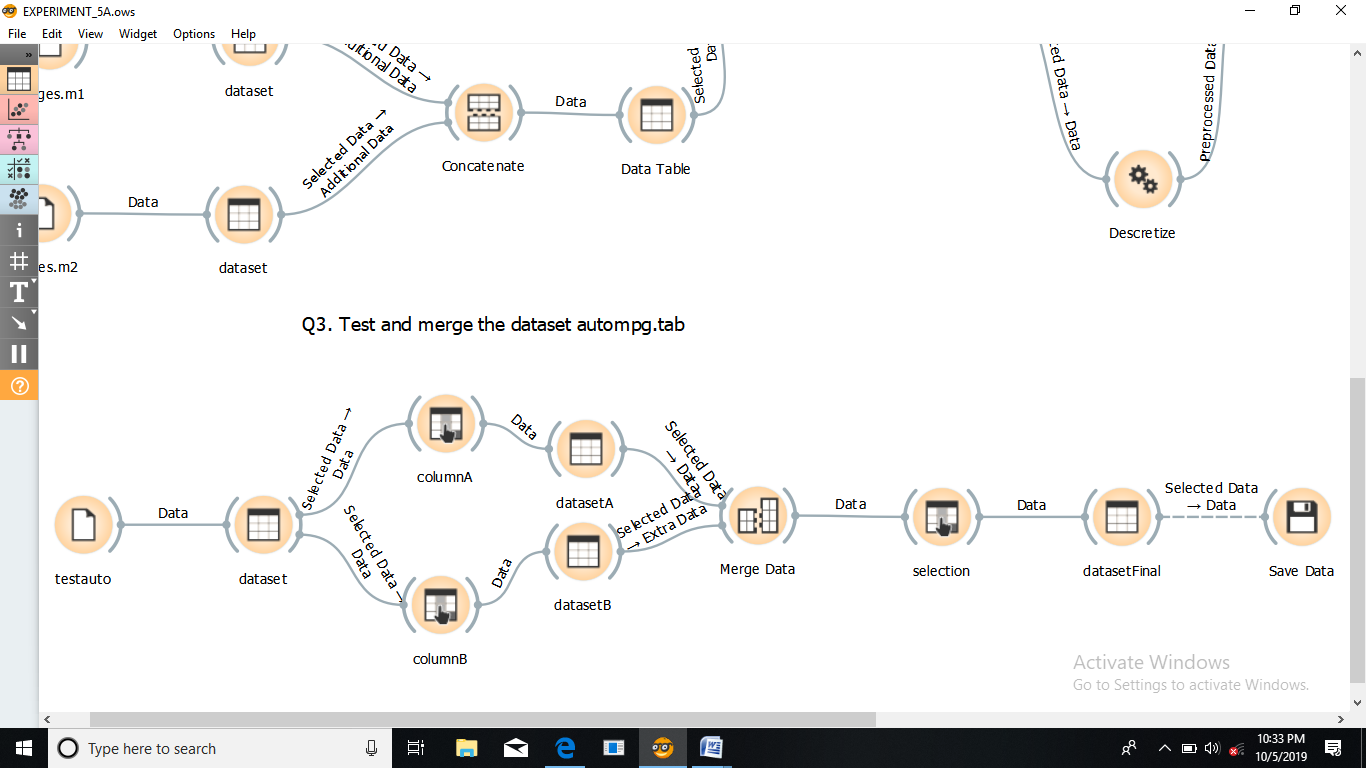
1. **Selecting specific features from the file.**



1. **Merging the files.**



1. **Structural arrangement of the tools in this scenario.**



**-------------------------------------------------------------------------------------**

**Answer the following questions:**

1. In which year maximum cars were built?

**Ans.** In 1970th year.

1. Write a python script to display first 3 rows of auto-mpg.

**Ans.**

indexes = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']

df = pd.DataFrame(auto-mpg , index=indexes)

print("First three rows of the data frame:")

print(df.iloc[:3])

**output**

a 23.6 4 140.0 ? 2905 14.3 80 1 ford mustang cobra

b 32.4 4 107.0 72 2290 17.0 80 3 honda accord

c 27.2 4 135.0 84 2490 15.7 81 1 plymouth reliant

1. Write a python script to display first 3 rows of auto-mpg and to find the maximum displacement, its index and the corresponding row.

**Python Script:**

**Script**

**import Orange**

**data= Orange.data.Table("auto-mpg")**

**print(data.domain)**

**for d in data[:3]:**

**print(d)**

**p=[1]**

**for i in range(0,398):**

**p.append(data[i]['displacement'])**

**#for i in range(0,3):**

**# print(p)**

**print('max value is:')**

**print(max(p))**

**print('index of max value is:')**

**i=p.index(max(p))**

**print(i)**

**data= Orange.data.Table("auto-mpg")**

**print(data.domain)**

**for q in data[i-1]:**

**print (q)**

**Output**

[cylinders, displacement, horsepower, weight, acceleration, model\_year,origin,car\_name| mpg]

[8, 307.000, 130.000, 3504.000, 12.000, 70, 1, chevrolet chevelle malibu | 18.000]

[8, 350.000, 165.000, 3693.000, 11.500, 70, 1, buick skylark 320 | 15.000]

[8, 318.000, 150.000, 3436.000, 11.000, 70, 1, plymouth satellite | 18.000]

max value is:

455.000

index of max value is:

9

[cylinders, displacement, horsepower, weight, acceleration, model\_year, origin, car\_name | mpg]

4.0

455.0

225.0

4425.0

10.0

0.0

0.0

241.0

14.0

1. Write a python script to find the maximum displacement, its index and the corresponding row.

**Ans.**

#to find max displacement and index

import operator

index, value = max(enumerate(auto-mpg), key=operator.itemgetter(1))

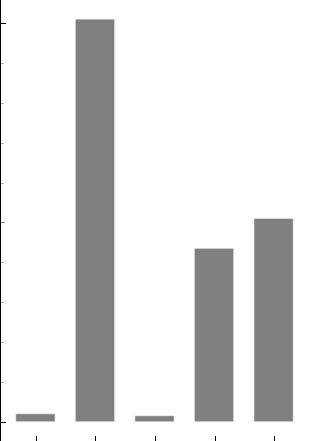
12.0 8 455.0 225 4951 11.0 73 1 buick electra 225 custom

14.0 8 455.0 225 4425 10.0 70 1 pontiac catalina

14.0 8 455.0 225 3086 10.0 70 1 buick estate wagon (sw)

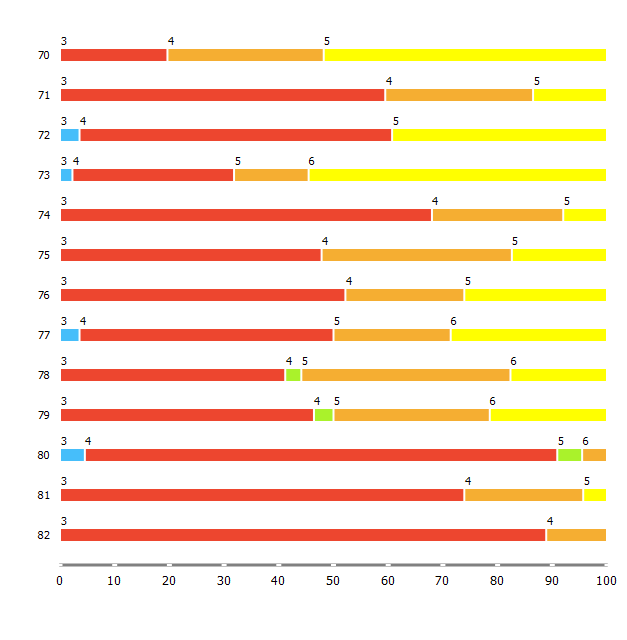
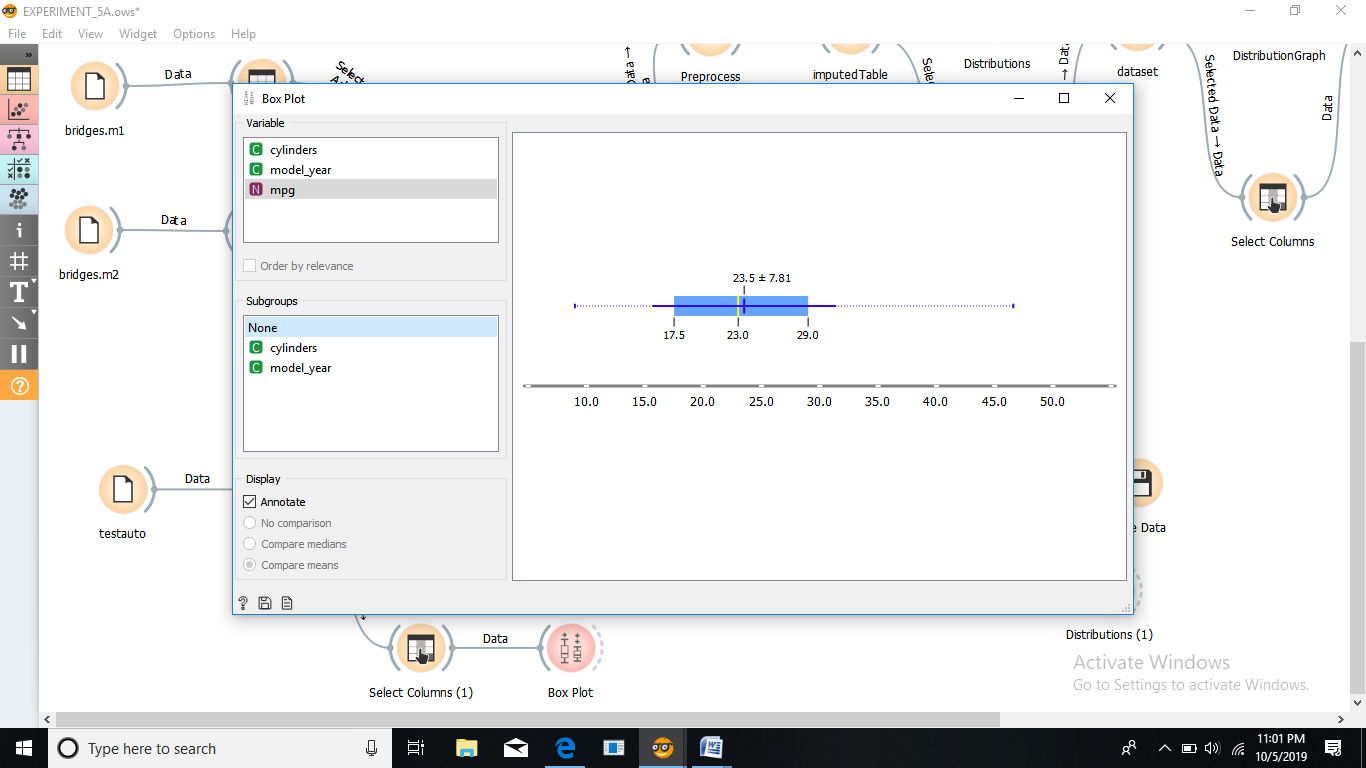
1. Conclude about the number of cylinders. Add grouping of model\_year in box plot to make conclusions.

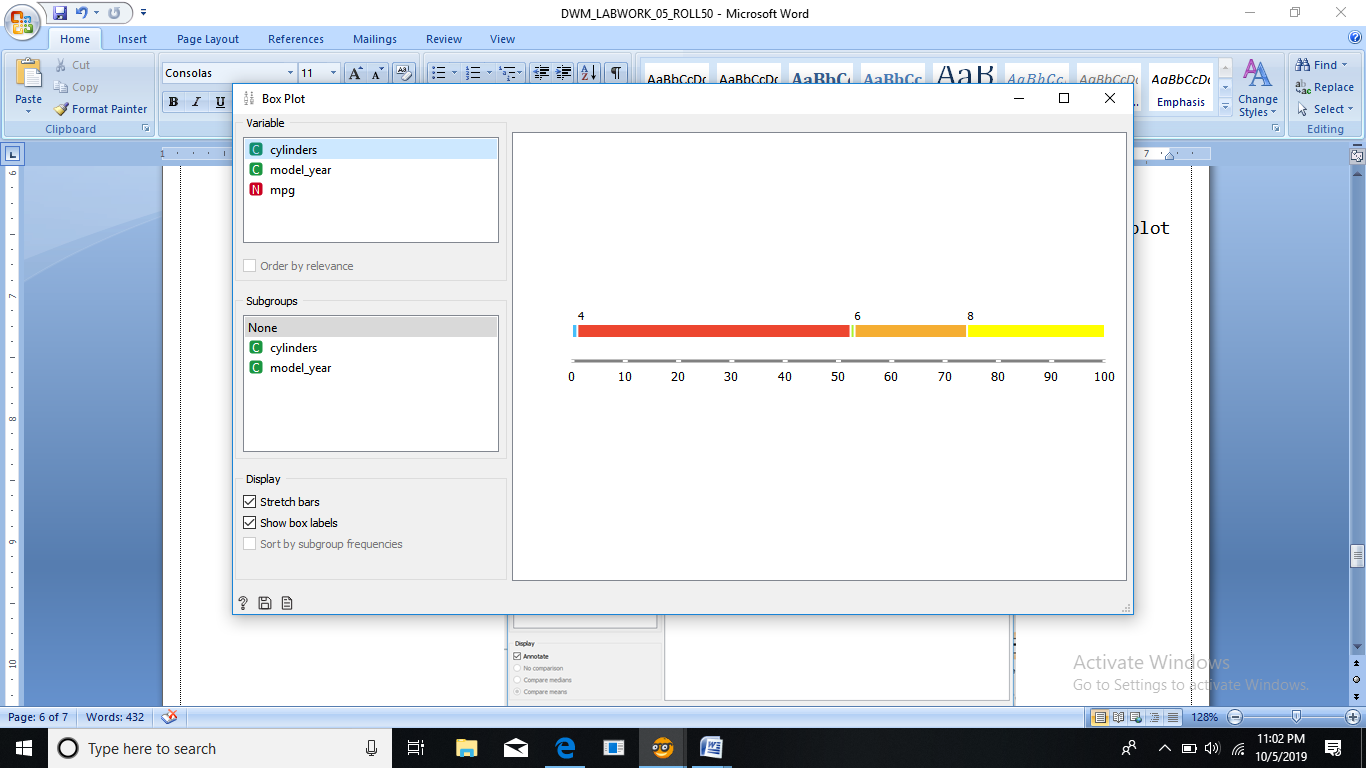
**Ans.**

****

It can be seen that for lower Cylinders, the mpg is far great. Hence we should definitely consider performing tests with Cylinder.

1. Show analysis using box plot.





**---------------------------------------------------------------------------------------------**

**DWM EXPERIMENT NO. : 06 (B)** **Roll no :50**

**---**-**-----------------------------------------------------------------------------------------**

**Problem definition :**

Introduction to the Weka machine learning toolkit.

---------------------------------------------------------------------------------------------

**1. Press the Explorer button on the main panel and load the weather dataset and answer the following questions**

* **How many instances are there in the dataset?**

14

* **State the names of the attributes along with their types and values.**

|  |  |  |
| --- | --- | --- |
| **NAME** | **TYPE** | **VALUE** |
| Outlook | Nominal | Sunny, overcast, rainy |
| Temperature | Numeric | Hot, mild, cool |
| Humidity | Numeric | High, normal |
| Windy | Nominal | True, false |
| Play | Nominal | Yes, no |

* **What is the class attribute?**

A class attribute is a target attribute. It is the attribute for which all the hypotheses are validated. In this case, class attribute is **Play**.

* **In the histogram on the bottom-right, which attributes are plotted on the X,Y-axes? How do you change the attributes plotted on the X,Y-axes?**

On X axis, label is marked whereas on Y axis, count is marked.

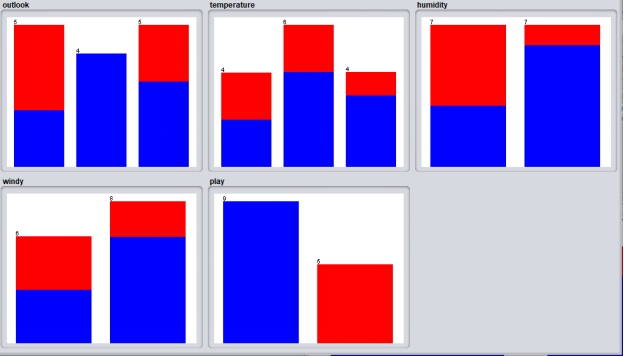
We can change the attributes on XY axes by changing the classes.

* **How will you determine how many instances of each class are present in the data?**

Instance count is equal to the number of distinct values. In this dataset, its 14. We can count the distinct values.

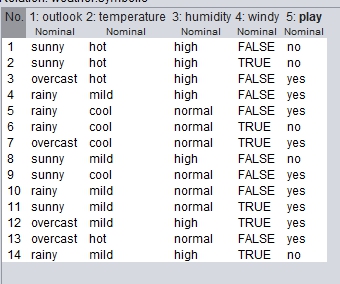
* **What happens with the Visualize All button is pressed?**

Histogram of all attributes get visible when visualize all is pressed.



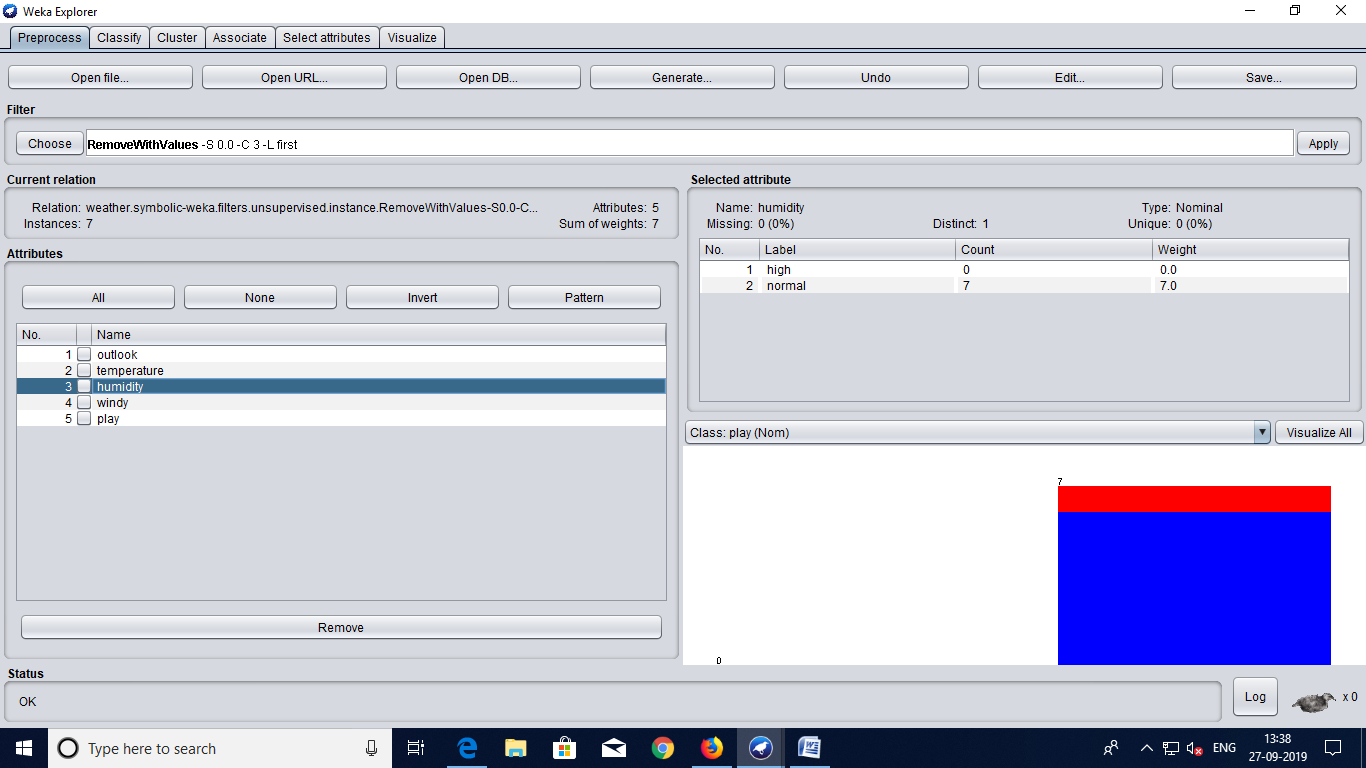
* **How will you view the instances in the dataset? How will you save the changes?**

By clicking the edit button, the viewer tab gets opened. Clicking OK saves the changes of active thread.



**2. Load the weather dataset and perform the following tasks:**

**1. Use the unsupervised filter RemoveWithValues to remove all instances where the attribute ‘humidity’ has the value ‘high’?**

****

**2. Undo the effect of the filter.**

**Answer the following questions:**

* **What is meant by filtering in Weka?**

Filtering is the process in weka which allows the data engineer to apply some attributes likeable to a criteria as per requirement to make some decision.

* **Which panel is used for filtering a dataset?**

Filter panel.

* **What are the two main types of filters in Weka?**

1. Supervised filter.
2. Unsupervised filter.

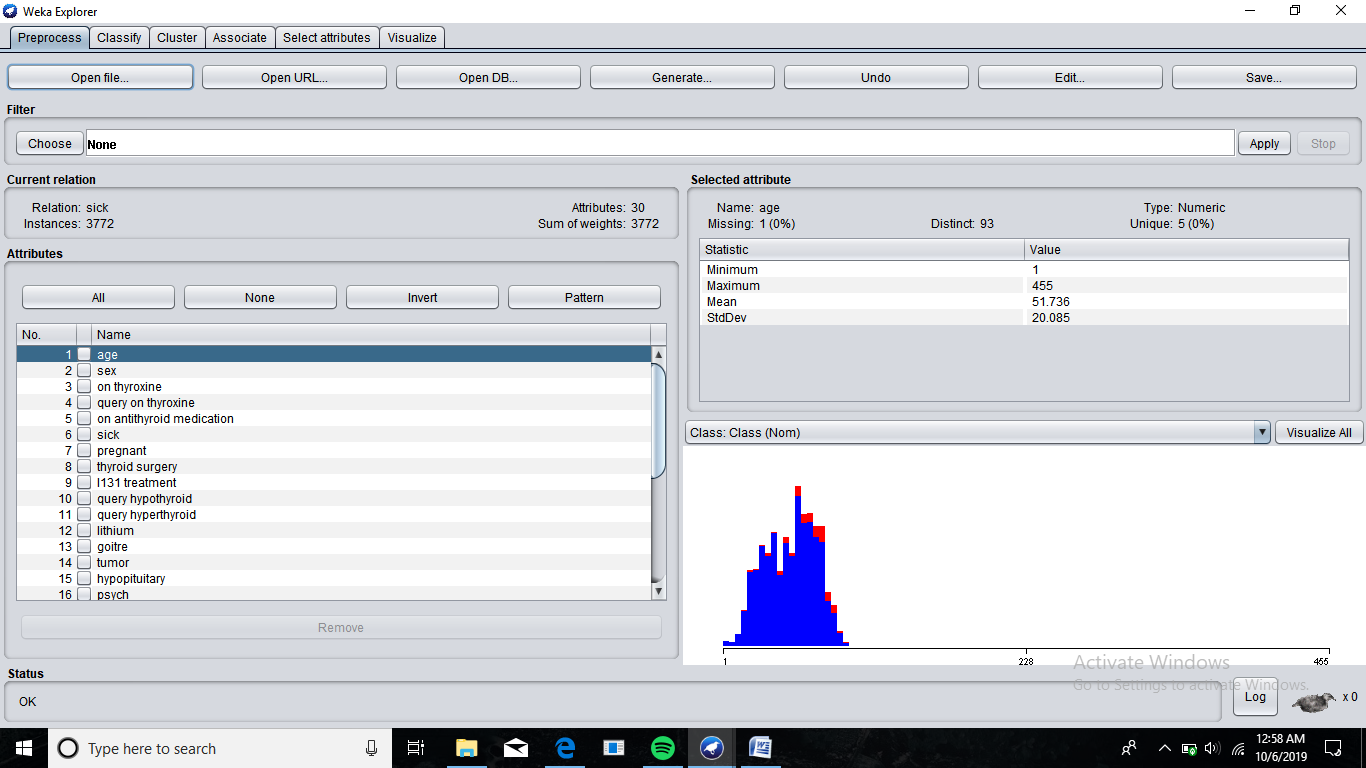
* **What is the difference between the two types of filters? What is the difference between and attribute filter and an instance filter?**

1. Supervised filters in Weka are filters that take the class distribution into account. If the data you are filtering is not classified or you don't want to use the classifications of the data points in the filter process, you'd want an "unsupervised filter".
2. An instance filter that creates a new attribute by applying a mathematical expression to existing attributes. An instance filter that adds an ID attribute to the dataset.A supervised attribute filter that can be used to select attributes. Converts the values of nominal and/or numeric attributes into class conditional probabilities. Changes the order of the classes so that the class values are no longer of in the order specified in the header.

**Part A: Application of Discretization Filters**

**1. Perform the following tasks**

1. **Load the 'sick.arff' dataset ?**

****

1. **How many instances does this dataset have?**

3772 instances.

1. **How many attributes does it have?**

30 attributes.

1. **Which is the class attribute and what are the characteristics of this attribute?**

Health is the target or class attribute.

**Characteristics :**

1. It can contain two values, either sick or negative.
2. It has a low cardinality.
3. It has nominal datatype.
4. **How many attributes are numerics? What are the attribute indexes of the numerical attributes?**

7 numeric attributes in all.

1, 18,20,22,24,26 and 28 indexed attributes are numerical.

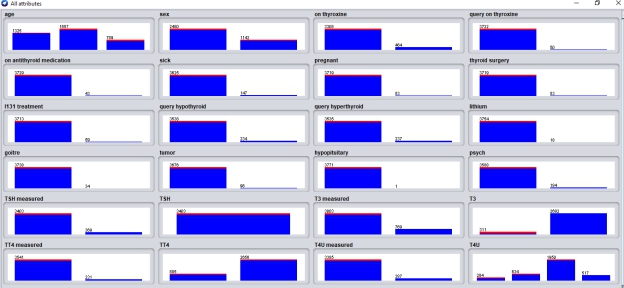
1. **Apply the Naive Bayes classifier. What is the accuracy of the classifier?**

92.9745% is the accuracy.

**2. Perform the following tasks:**

**1. Load the 'sick.arff' dataset.**

**2. Apply the supervised discretization filter.**

****

**3. What is the effect of this filter on the attributes?**

It discretizes a range of numeric attributes in the dataset into nominal attributes. The main beneﬁt of this is that some classiﬁers can only take nominal attributes as input, not numeric attributes. Another advantage is that some classiﬁers that can take numeric attributes can achieve improved accuracy if the data is discretized prior to learning.

**4. How many distinct ranges have been created for each attribute?**

Age – 3, TSH-1,T3-2,TT4-2,T4U-4,FTI-1,TBG-1

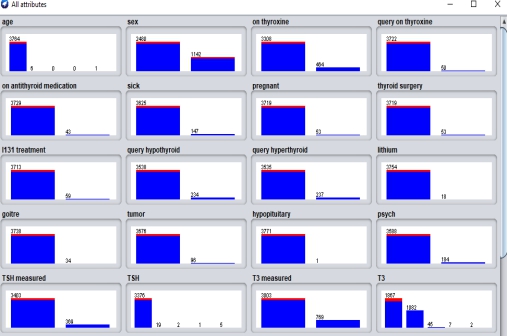
**5. Undo the filter applied in the previous step.**

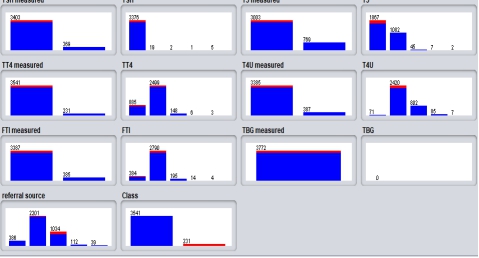
**6. Apply the unsupervised discretization filter. Do this twice:**

**1. In this step, set 'bins'=5**

**2. In this step, set 'bins'=10**

**3. What is the effect of the unsupervised filter filter on the datset?**

****

****

**7. Run the the Naive Bayes classifier after apply the following filters**

**1. Unsupervised discretized with 'bins'=5**

Correctly Classified Instances 3455 91.596 %

Incorrectly Classified Instances 317 8.404 %

1. **Unsupervised discretized with 'bins'=10**

Correctly Classified Instances 3654 96.8717 %

Incorrectly Classified Instances 118 3.1283 %

**3. Unsupervised discretized with 'bins''=20.**

Correctly Classified Instances 3662 97.0838 %

Incorrectly Classified Instances 110 2.9162 %

**8. Compare the accuracy of the following cases**

**1. Naive Bayes without discretization filters**

Correctly Classified Instances 3493 92.6034 %

Incorrectly Classified Instances 279 7.3966 %

**2. Naive Bayes with a supervised discretization filter**

Correctly Classified Instances 3662 97.0838 %

Incorrectly Classified Instances 110 2.9162 %

**3. Naive Bayes with an unsupervised discretization filter with different values for the 'bins attributes.**

**1. Unsupervised discretized with 'bins'=5**

Correctly Classified Instances 3455 91.596 %

Incorrectly Classified Instances 317 8.404 %

**2. Unsupervised discretized with 'bins'=10**

Correctly Classified Instances 3654 96.8717 %

Incorrectly Classified Instances 118 3.1283 %

**3. Unsupervised discretized with 'bins''=20.**

Correctly Classified Instances 3662 97.0838 %

Incorrectly Classified Instances 110 2.9162 %

***Part II: Attribute Selection***

**1. Perform the following tasks:**

**1. Load the 'mushroom.arff' dataset**

**2. Run the J48, 1Bk, and the Naive Bayes classifiers.**

J48 Correctly Classified Instances 8124 100 %

1Bk Correctly Classified Instances 8124 100 %

Incorrectly Classified Instances 0 0 %

**3. What is the accuracy of each of these classifiers?**

J48 Correctly Classified Instances 8124 100 %

IBK Correctly Classified Instances 8124 100 %

Incorrectly Classified Instances 0 0 %

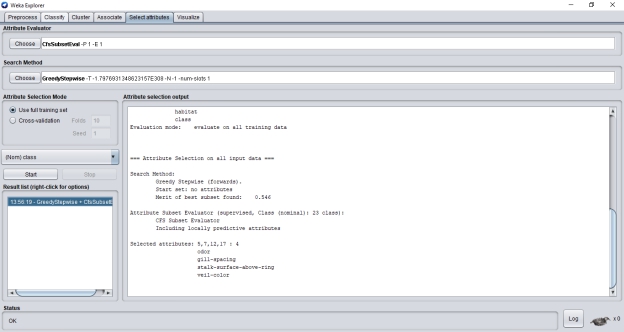
**2. Perform the following tasks:**

**1. Go to the 'Select Attributes' panel**

**2. Set attribute evaluator to CFSSubsetEval**

**3. Set the search method to 'Greedy Stepwise'**

**4. Analyze the results window**

****

**5. Record the attribute numbers of the most important attributes**

5,7,12,17

* odor
* gill-spacing
* stalk-surface-above-ring
* veil-color

**6. Run the meta classifier AttributeSelectedClassifier using the following:**

**1. CFSSubsetEval**

**2. GreedStepwise**

**3. J48, 1Bk, and NaiveBayes**

Results are same as above

**7. Record the accuracy of the classifiers**

**8. What are the benefits of attribute selection?**

* Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
* Improves Accuracy: Less misleading data means modelling accuracy improves.
* Reduces Training Time: Less data means that algorithms train faster.

**Part C**

**1. Perform the following tasks:**

1. Load the 'vote.arff' dataset.

2. Run the J48, 1Bk, and Naive Bayes classifiers.

3. Record the accuracies.

**J48**

Correctly Classified Instances 419 96.3218 %

Incorrectly Classified Instances 16 3.6782 %

**1Bk**

Correctly Classified Instances 434 99.7701 %

Incorrectly Classified Instances 1 0.2299 %

**Naïve Bayes**

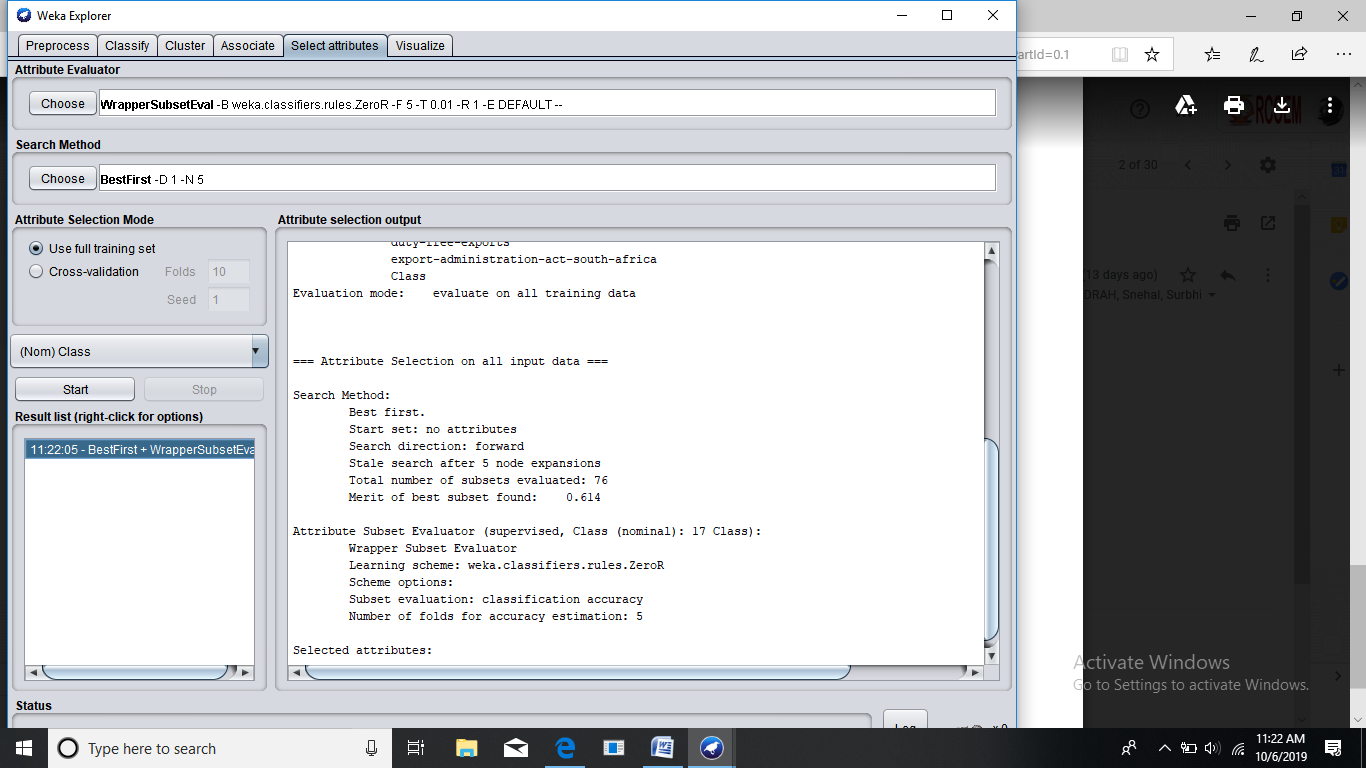
Correctly Classified Instances 393 90.3448 %

Incorrectly Classified Instances 42 9.6552 %

**2. Perform the following tasks:**

1. Go to the 'Select Attributes' panel

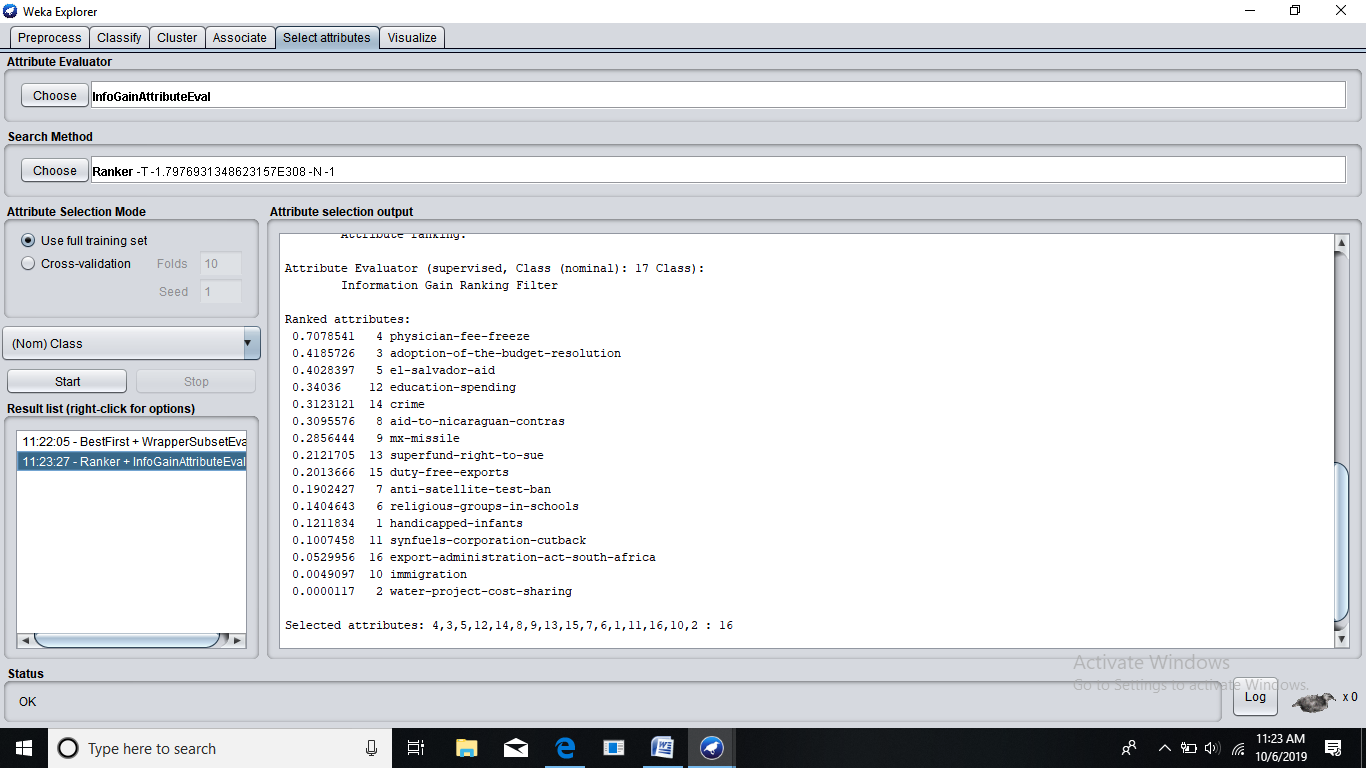
2. Set attribute evaluator to 'WrapperSubsetEval



**3. Set search method to ''RankSearch'**

**4. Set attribute evaluator to 'InfoGainAttributeEval'**

**5. Analyze the results**

****

**6. Run the metaclassifier AttributeSelectedClassifier using the following:**

**1. WrapperSubsetEval**

Correctly Classified Instances 416 95.6322 %

Incorrectly Classified Instances 19 4.3678 %

**2. RankSearch**

Correctly Classified Instances 416 95.6322 %

Incorrectly Classified Instances 19 4.3678 %

**3. InfoGainAttributeEval**

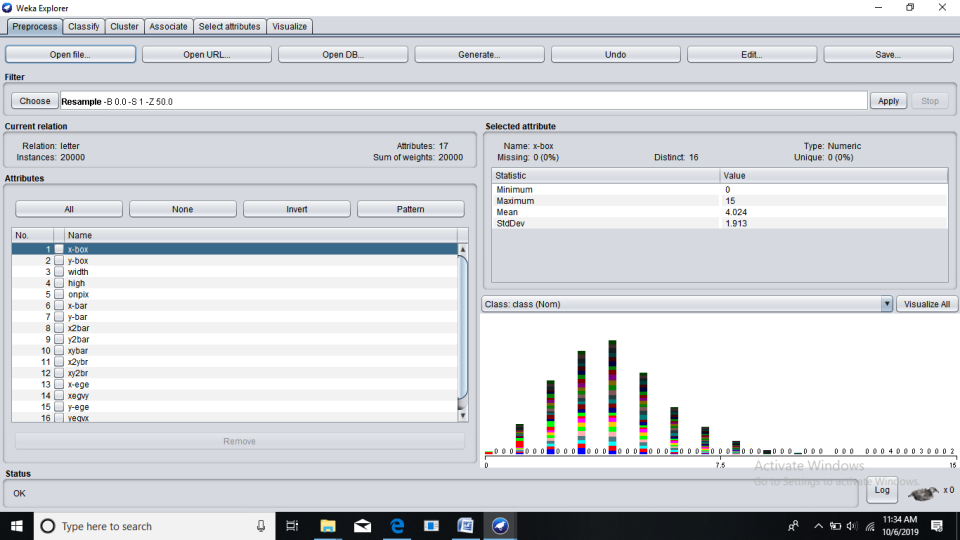
Correctly Classified Instances 416 95.6322 %

Incorrectly Classified Instances 19 4.3678 %

**7. Sampling**

**1. Load the 'letter.arff' dataset**

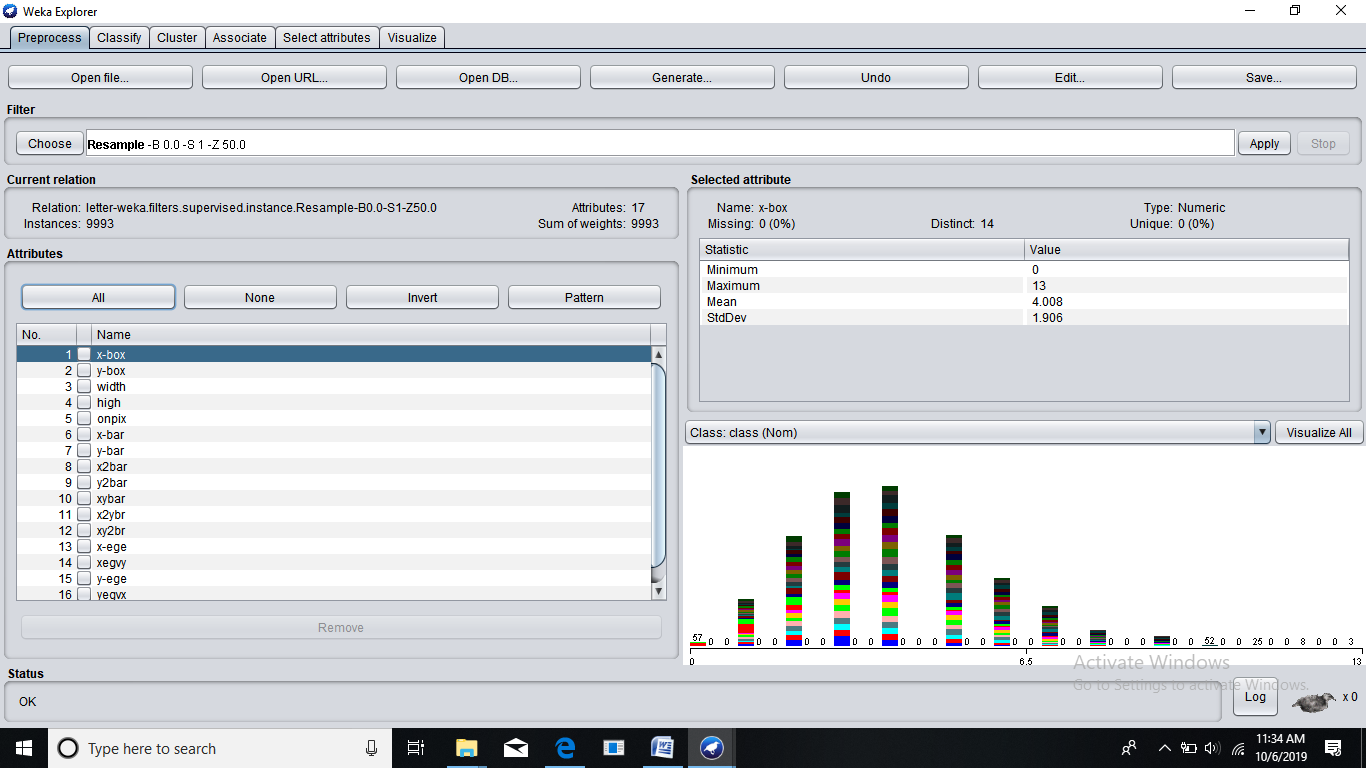
**2. Take any attribute and record the min, max, mean, and standard deviation of the attribute**



**3. Apply the Resample filter with 'sampleSizePercent' set to 50 percent**

**4. What is the size of the filtered dataset. Observe the min, max, mean, and standard deviation Of the attribute that was selected in step 2. What is the percentage change in the values?**

0.36% change resulted in Standard deviation values.



1. **Give the benefit of sampling a large dataset.**

Sampling can be particularly useful with data sets that are too large to efficiently analyze in full, for example, in data science applications or surveys. Identifying and analyzing a representative sample is more efficient and cost-effective than surveying the entirety of the data or population.