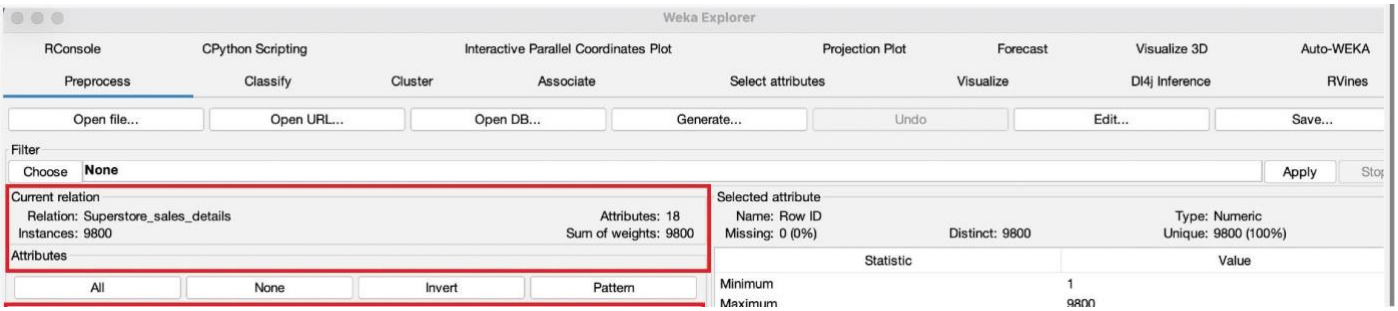
SUPERSTORE_DETAILS DATASET Student

Details:

Student Name and ID of the remaining members: **Jayachandra Jarajapu (1001964536)**

Loading the dataset:

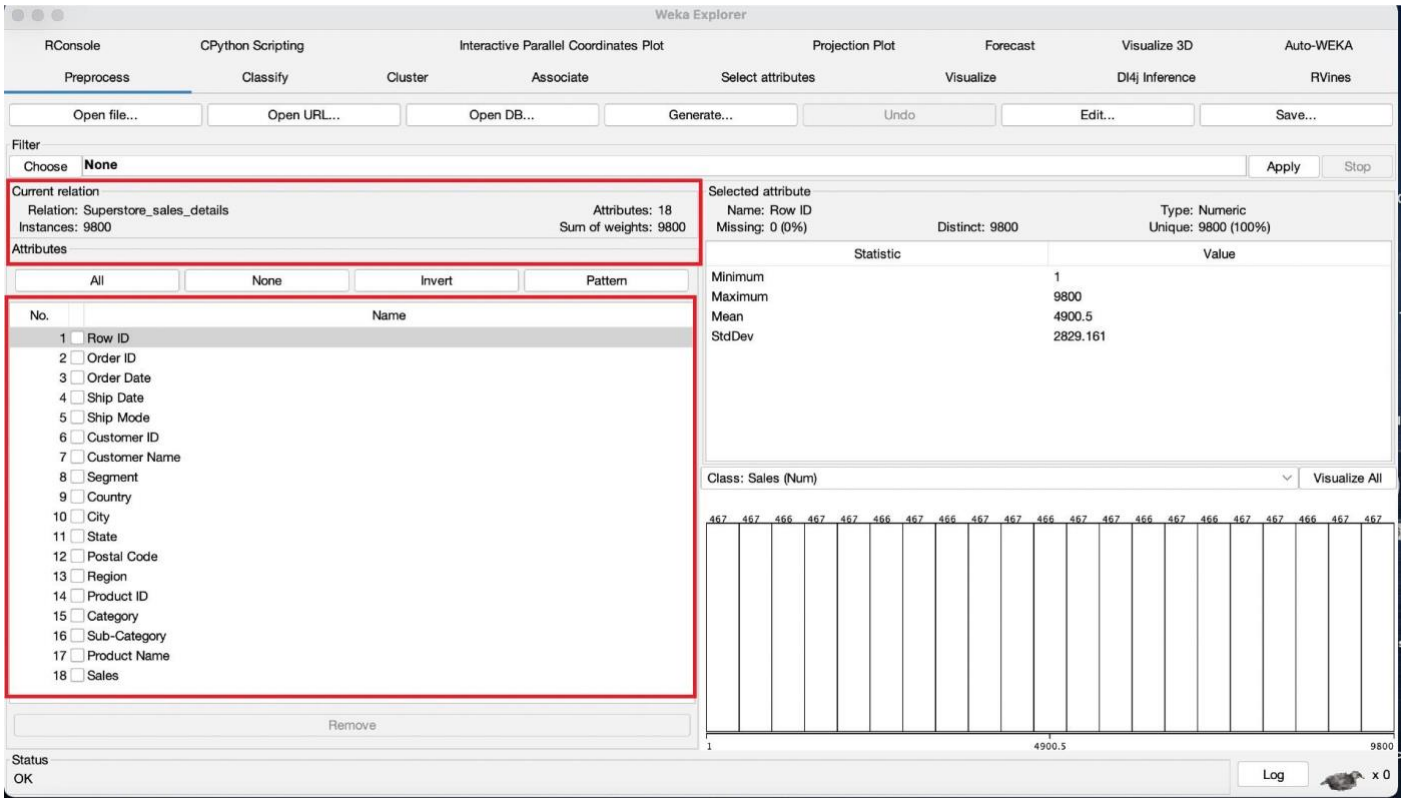
- The dataset which was given in the CSV format made it harder to load into WEKA and it contained special characters
- We had to remove few special characters without altering the fields
- The Materials data dataset is loaded into WEKA using the Weka console in explorer tab



TASK 1 - Statistical Exploratory Data Analysis

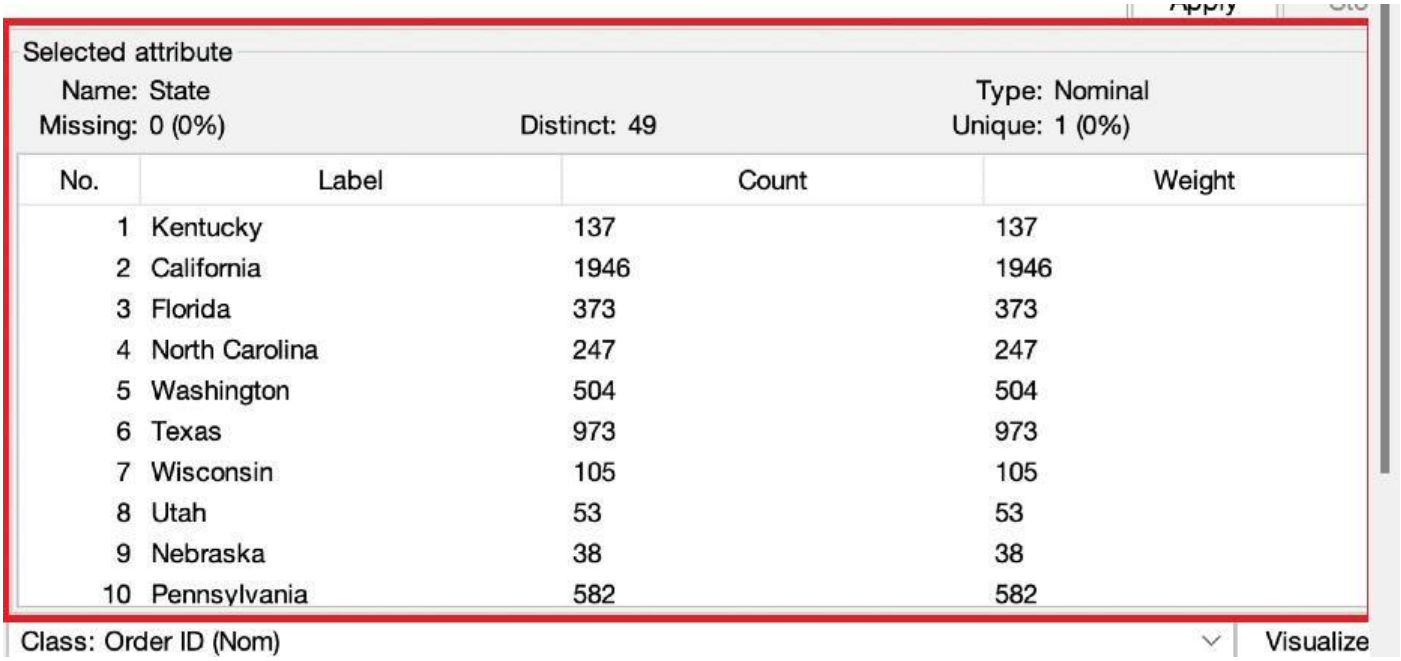
Printing the details of the data frame

- All the attributes of the dataframe are loaded into the Weka explorer



Task: Finding the number of rows and columns in dataset

- The length of the rows and columns are listed under the attributes field



Selected attribute

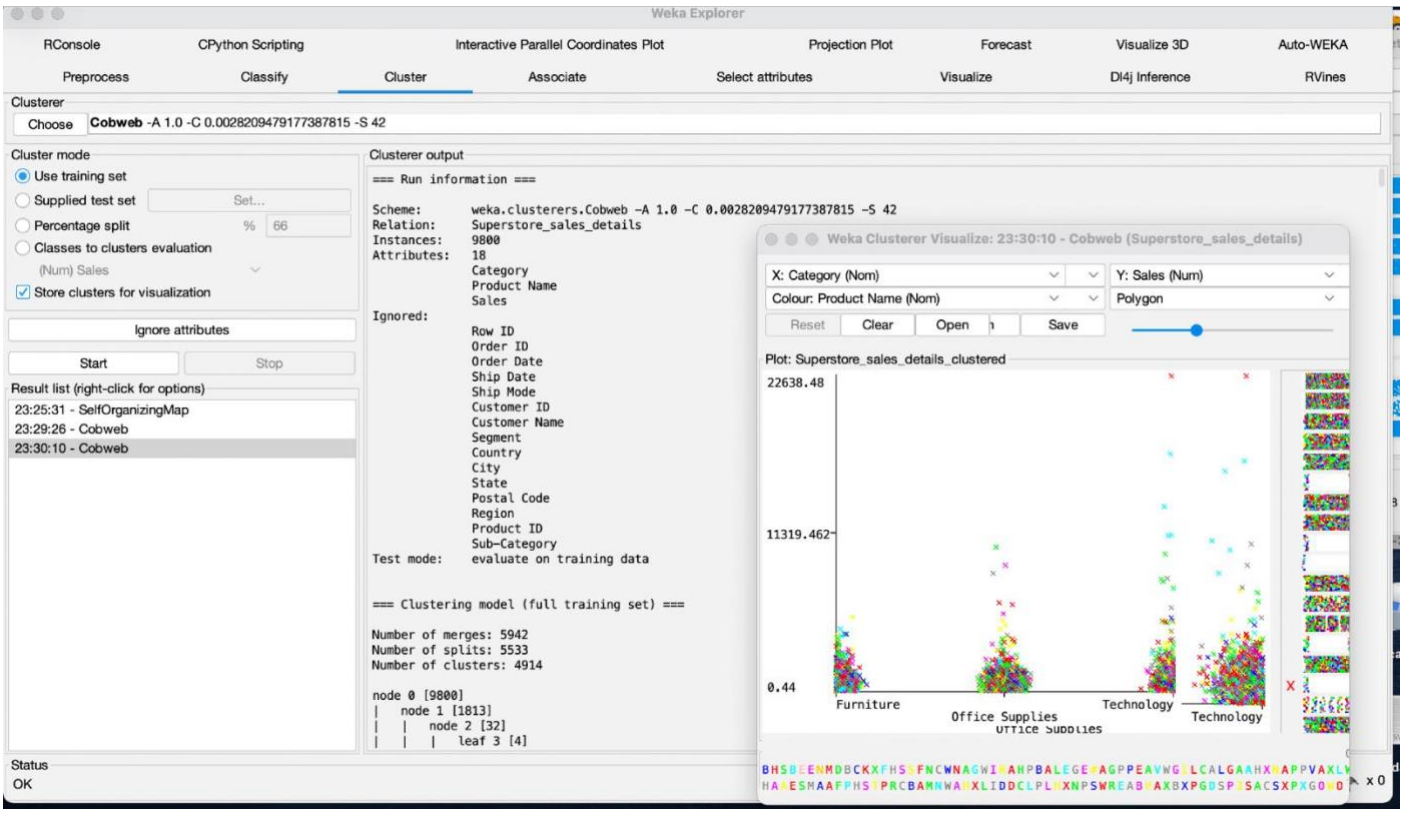
Name: State | Type: Nominal
Missing: 0 (0%) | Distinct: 49 | Unique: 1 (0%)

No.	Label	Count	Weight
1	Kentucky	137	137
2	California	1946	1946
3	Florida	373	373
4	North Carolina	247	247
5	Washington	504	504
6	Texas	973	973
7	Wisconsin	105	105
8	Utah	53	53
9	Nebraska	38	38
10	Pennsylvania	582	582

Class: Order ID (Nom) | Visualize

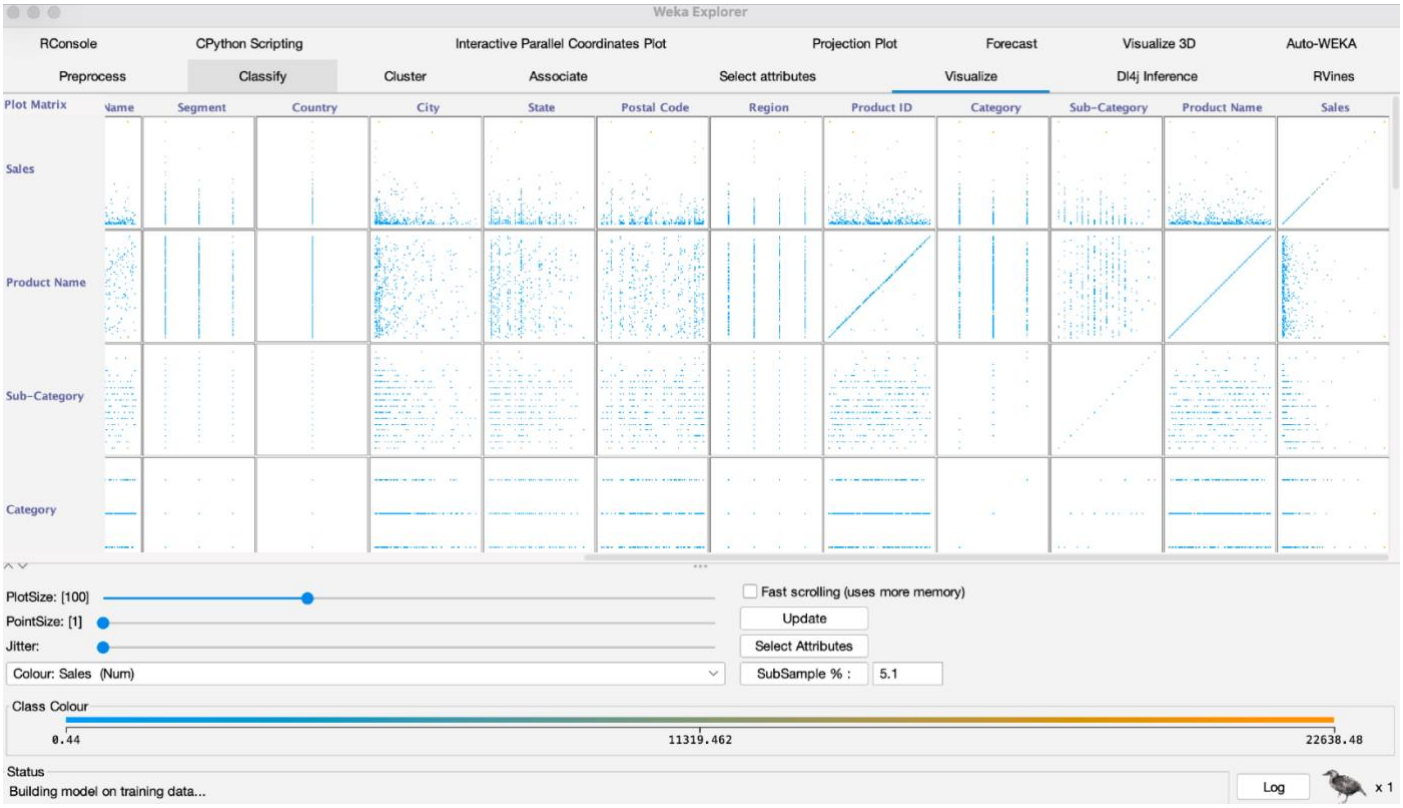
Task: Clustering between product and sales

- Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.
- We used clustering for the attribute sales and product name to plot in a nominal visualization • The Clusterer used here is selfOgranizingMap



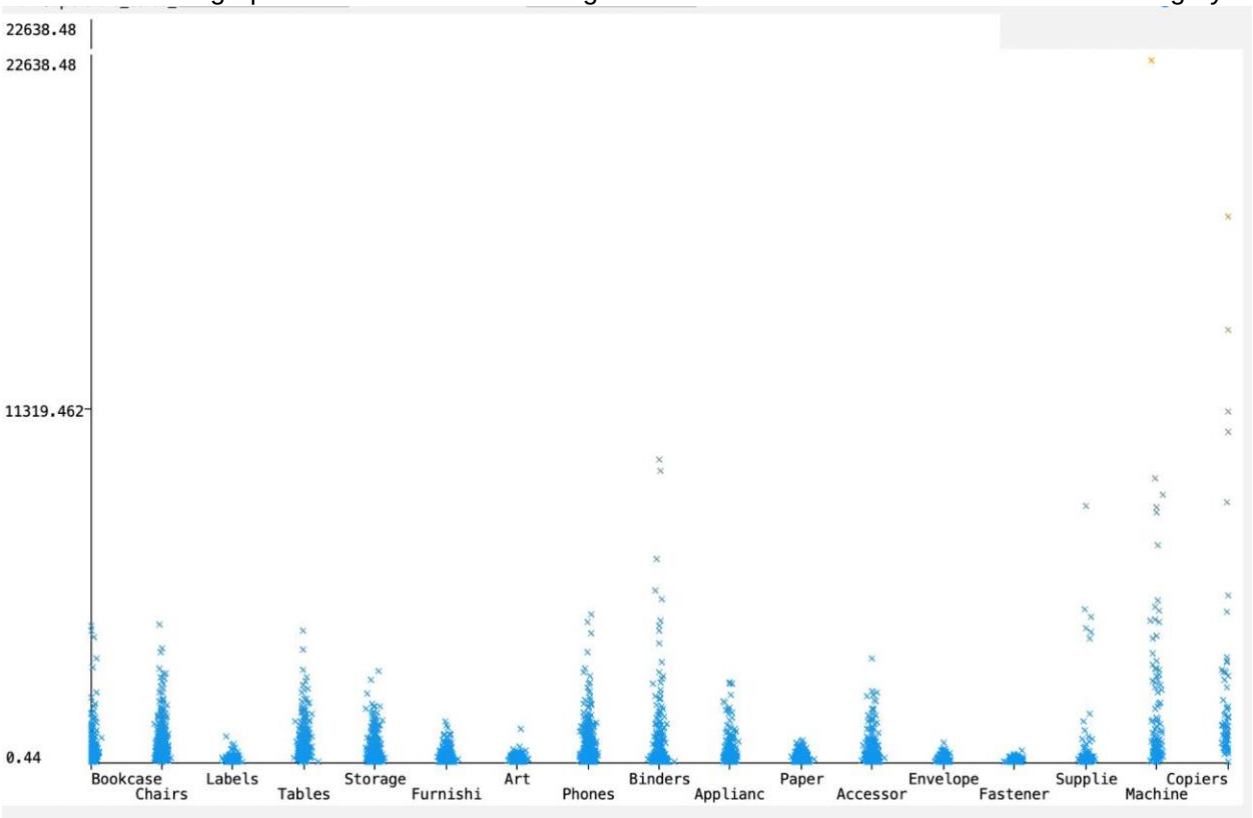
##TASK 3: VISUALIZATION

- The overview of all the data is represented using Visualization



Task 3-a: Plotting the comparison of sub categories with their sales

- Data Visualization weka package is used to plot the comparison between two columns
- we used 20% Jitter for the data as the size of the frame is more and visualized it accordingly
- From the graph we could see that the highest number of sales falls under machine category



Task 3-b: Decision tree classifying the state of the attribute using classifier J48.

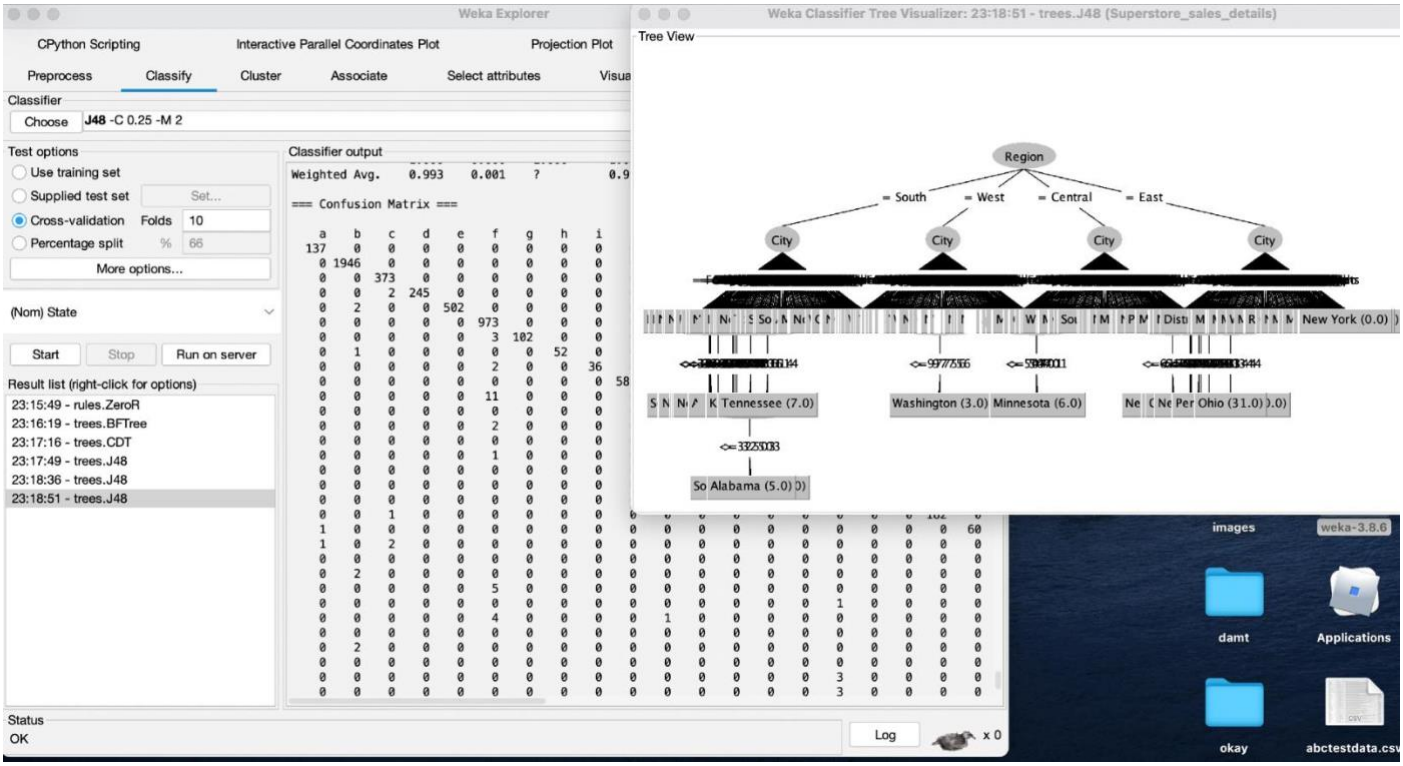
- Classification algorithm is used to visualize the tree for the state attribute and the classifier we used is the J48
- The tree weights each class by the means of order delivered in J48
- The tree is represented as

  State – reason

  Reason – city city

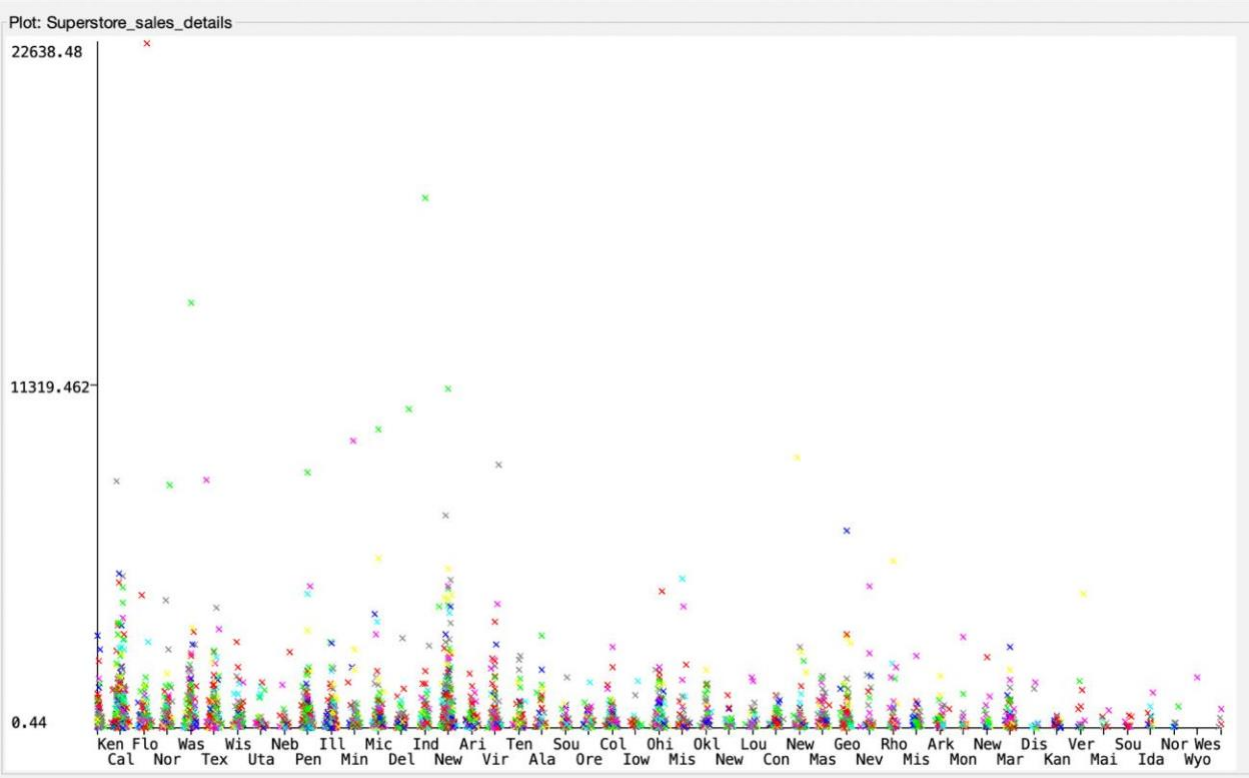
  – postal code

- We use the column ‘year_Birth’ to count the values using the summarise function
- The title is changed using the main() function
- We use geom_area function to plot the comparision of number of births with Year of Birth in each country
- We use ggplot to create a pie chart with size using labels and the total we counted earlier using the summarise function



Task 4: Pattern Visualization

- To find an interesting pattern we have used columns ‘sales” and “states” to find the pattern with number of sales aggregately in the plot
- We can use visualization trees such as decision tree for plotting this graph
- We have used 10% Jitter for this plot for all the sales for each state with best possible fit
- From the graph we could see that although Florida has top sales for a category, The Aggregate sales for California is higher than in Florida



References:

https://www.tutorialspoint.com/weka/index.htm

https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/

https://www.youtube.com/watch?v=B8UDaZupCZ8