St. Francis Institute of Technology, Mumbai-400 103

Department Of Information Technology

A.Y. 2020-2021

Class: TE-ITA/B, Semester: VI

Subject: **Business Intelligence Lab**

Experiment – 10: Business Intelligence Mini Project

1. Aim: Develop a Mini Project for Business Intelligence

2. Objectives: After study of this experiment, the students will be able to develop mini project

3. Outcomes:

CO6: Apply BI to solve practical problems : Analyze the problem domain, use the data collected in enterprise apply the appropriate data mining technique, interpret and visualize the results and provide decision support

4. Prerequisite: Study of BI Tools.

5. Requirements: Personal Computer, Windows XP operating system/Windows 7, Internet Connection, Microsoft Word, WEKA tool, BI Tool.

6. Theory:

Each group is assigned one new case study for this; A BI report must be prepared outlining the following steps:

a. Problem definition, Identifying which data mining task is needed
b. Identify and use a standard data mining dataset available for the problem. Some links for data mining datasets are WEKA site, UCI Machine Learning Repository, KDD site, KDD Cup etc.
c. Implement the data mining algorithm of choice
d. Interpret and visualize the results
e. Provide clearly the BI decision that is to be taken as a result of mining.

7. Laboratory Exercise: Study of BI Tools, Designing and development of Mini Project for BI and take Printout of Implementation along with coding and snapshot of output

Case Study on Mall Customer Segmentation using Data Mining

Introduction to Data Mining :

Data mining, also known as knowledge discovery in databases, is prompted by the need for new techniques to help analyze, understand or even visualize the large amounts of stored data gathered from business and scientific applications. It is the process of investigating knowledge, such as patterns, associations, changes, anomalies or significant structures from large amounts of data stored in databases, data warehouses, or other information repositories.

Data Mining refers to extracting or mining knowledge from large amounts of data. The term is actually a misnomer. Thus, data mining should have been more appropriately named as knowledge mining which emphasizes mining from large amounts of data. It is a computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Purpose of Data Mining:

- Basically, the information gathered from Data Mining helps to predict hidden patterns, future trends and behaviors and allows businesses to take decisions.
- Technically, data mining is the computational process of analyzing data from different perspectives, dimensions, angles and categorizing/summarizing it into meaningful information.
- Data Mining can be applied to any type of data e.g. Data Warehouses, Transactional Databases, Relational Databases, Multimedia Databases, Spatial Databases, Time-series Databases, World Wide Web.

Data Mining Techniques :

- **Data cleaning and preparation :** Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.Data preparation is the process of cleaning and transforming raw data prior to processing and analysis.

- **Tracking patterns:**Tracking patterns is a fundamental data mining technique. It involves identifying and monitoring trends or patterns in data to make intelligent inferences about business outcomes.

- **Classification:** Classification is a data mining function that assigns items in a collection to target

categories or classes. The goal of classification is to accurately predict the target class for each case in the data.

- **Association:** Association rule mining is a procedure which aims to observe frequently occurring patterns, correlations, or associations from datasets found in various kinds of databases such as relational databases, transactional databases, and other forms of repositories.

- **Outlier detection:** Outlier detection is a primary step in many data-mining applications. ... In presence of outliers, special attention should be taken to assure the robustness of the used estimators. Outlier detection for Data Mining is often based on distance measures, clustering and spatial methods.

- **Clustering:** Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

- **Regression:** Regression is a data mining machine learning technique used to fit an equation to a dataset. The simplest form of regression, linear regression , uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the value of y based upon a given value of x.Basically a Linear regression model is used to show or predict the relationship between two variables or factors.

Mall Customer Data

Introduction :

The shopping mall is mainly for customers to get out of the house for a while and do something entertaining. Shopping malls can provide the best shopping experiences such as social gatherings, entertainment, performances, product launches, promotions and festivals.Malls or shopping complexes are often indulged in the race to increase their customers and hence making huge profits.To achieve this task machine learning is being applied by many stores already.

Background :

In this case study, we gathered the mall's customer dataset which mentioned that People who own a supermarket mall through membership cards have some basic data about their customers like Customer ID, age, gender, annual income and spending score.Spending Score is something assign to the customer based on your defined parameters like customer behavior and purchasing data.This data set is created only for the learning purpose of the customer segmentation concepts , also known as market basket analysis .

Mall Customer Dataset :

The dataset is gathered from a data mining dataset site . It consists of six attributes. This dataset's aim is to make specific customer segmentation using data mining algorithms.The attributes that are used in this case study as follows:

- Customer ID (Numeric)
- Age(Numeric)
- Gender(Nominal)
- Annual income (Numeric)
- Spending score(Numeric)

Problem Definition :

People or Owner who own the mall and want to understand the customers like who can be easily converge i.e.target customers so that the sense can be given to marketing team and plan the strategy accordingly to get more customers to visit to the shopping malls or supermarkets.To determine and understand the marketing strategy works in real world.

Problem Solution :

To understand the customers behaviour and their interest in shopping malls and supermarkets this can be done by using a data mining algorithm in order to make marketing strategies to target customers. This can be done using unsupervised Machine Learning Technique KMeans Clustering Algorithm in the simplest form.

Data Mining Algorithms on Chosen Data set in Weka Tool :

Clustering : Clustering is a task of grouping objects into classes of similar objects. It is an unsupervised classification or partitioning of patterns into groups or clusters based on their locality and connectivity within an n-dimensional space. In this study, clustering has been used for finding clusters of customers with similar characteristics.

Weka Tool : It is a data mining tool for data analysis a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform.

The algorithms can either be applied directly to a dataset or called from your own Java code.

Results in Weka Tool :

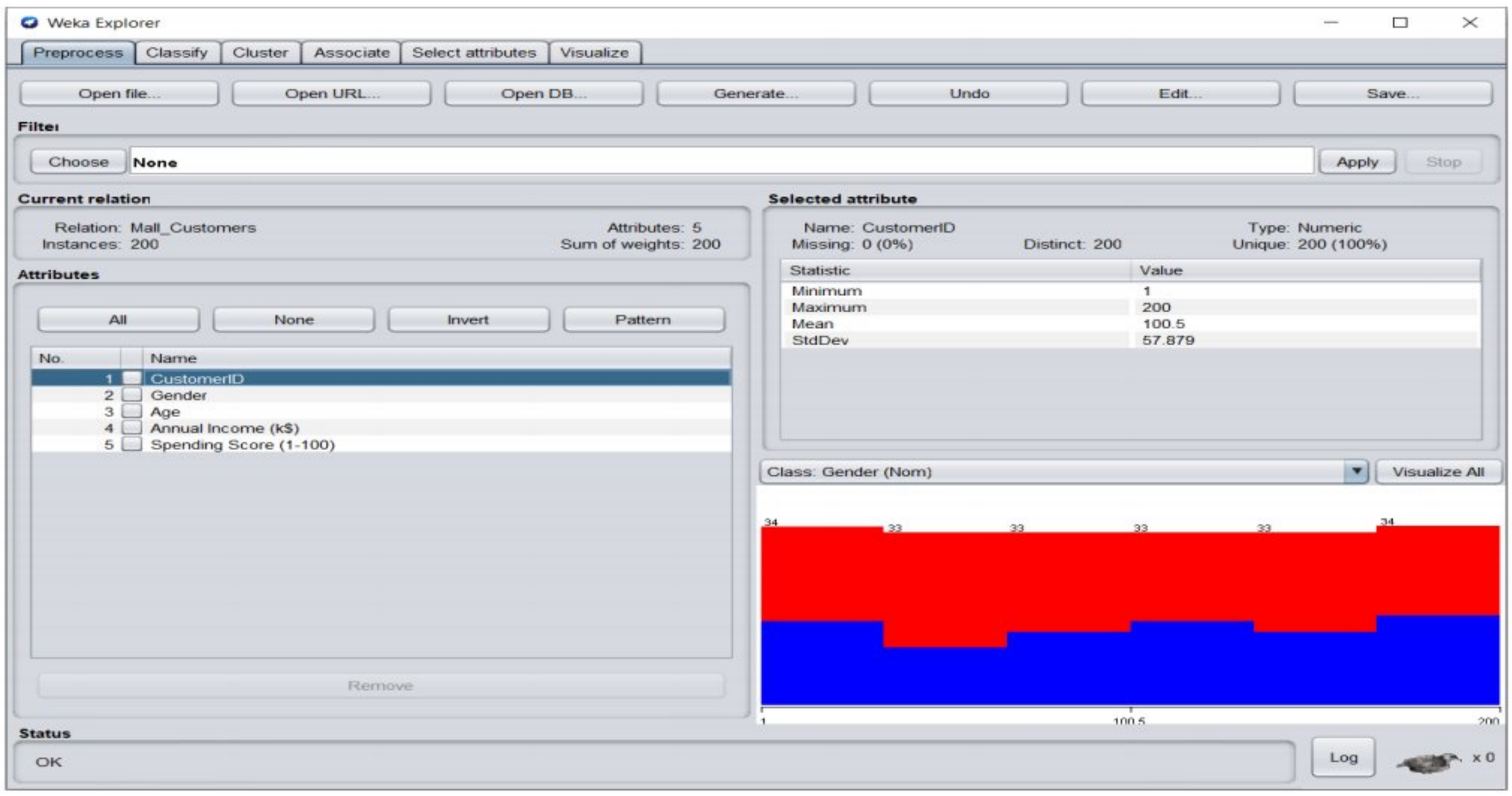


Fig.no.1

The Mall Customer Dataset is preprocessed for cluster algorithms.The above screenshot which shows that mall customer dataset is opened in the weka tool by clicking on the open file option.

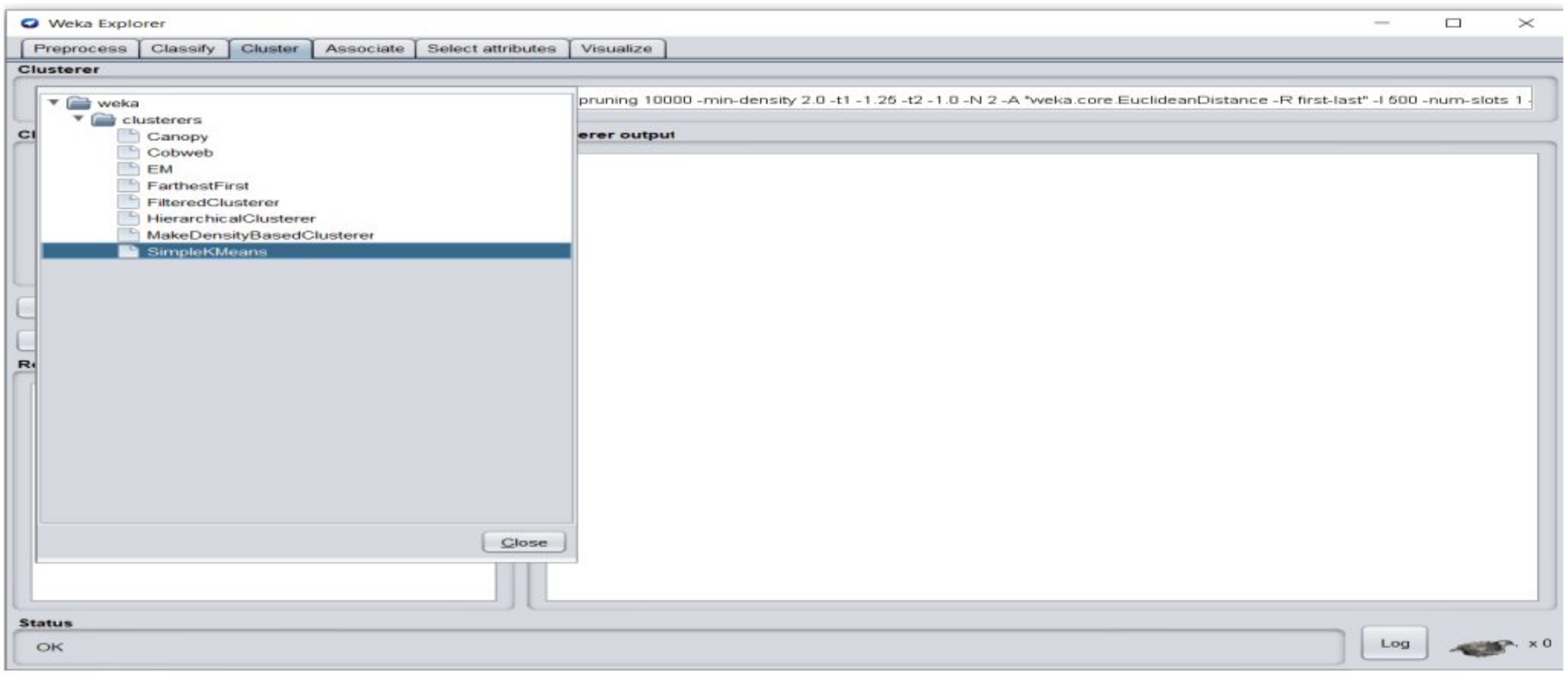


Fig.no.2

The above screenshot shows that the simple K-Means algorithm is chosen to apply a clustering algorithm on the dataset.

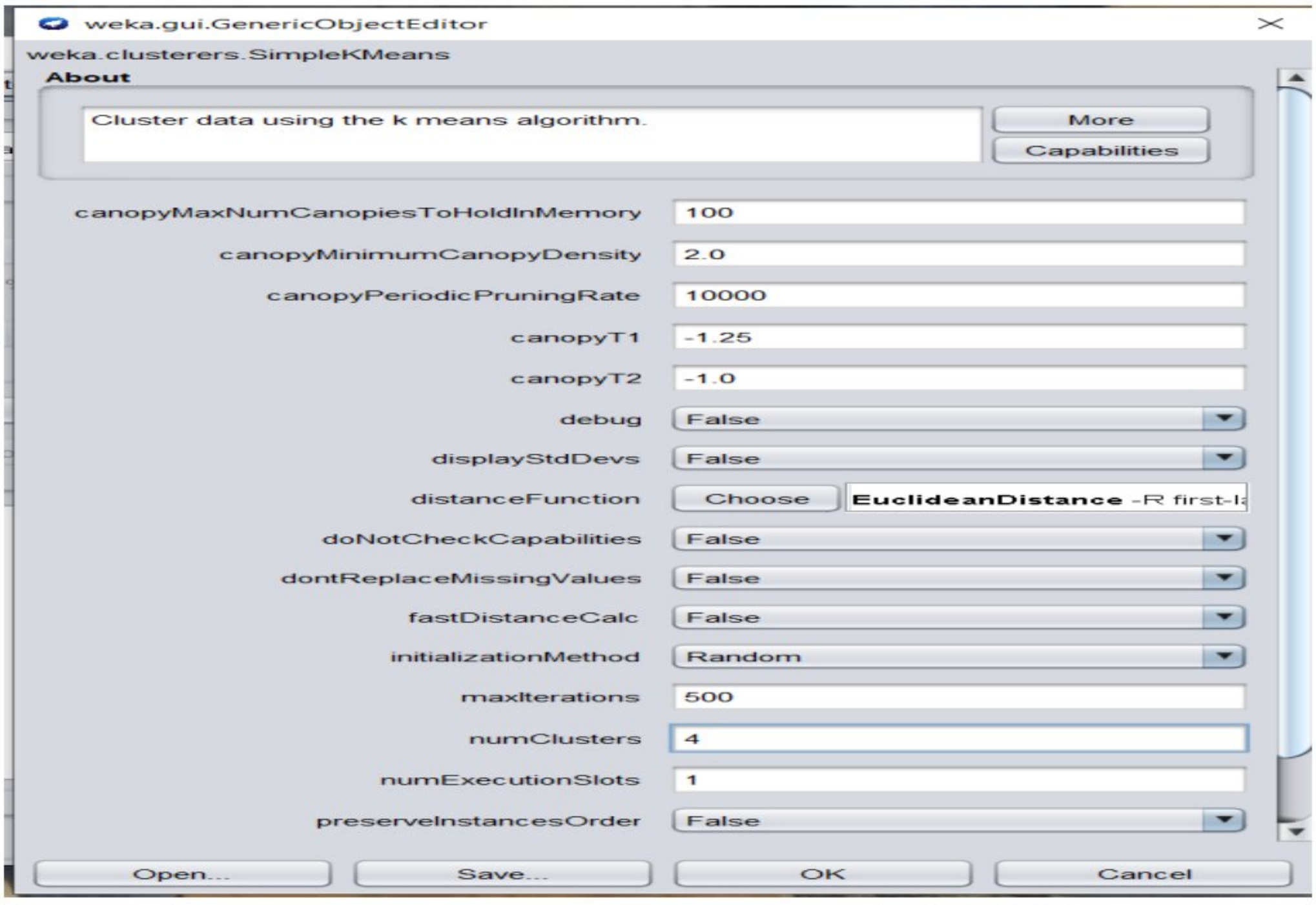


Fig.no.3

In the pop-up window screenshot ,we enter 4 as the number of clusters instead of the default values of 2. Generally , K-means is quite sensitive to how clusters are initially assigned. Thus, it is often necessary to try different values and evaluate the results.

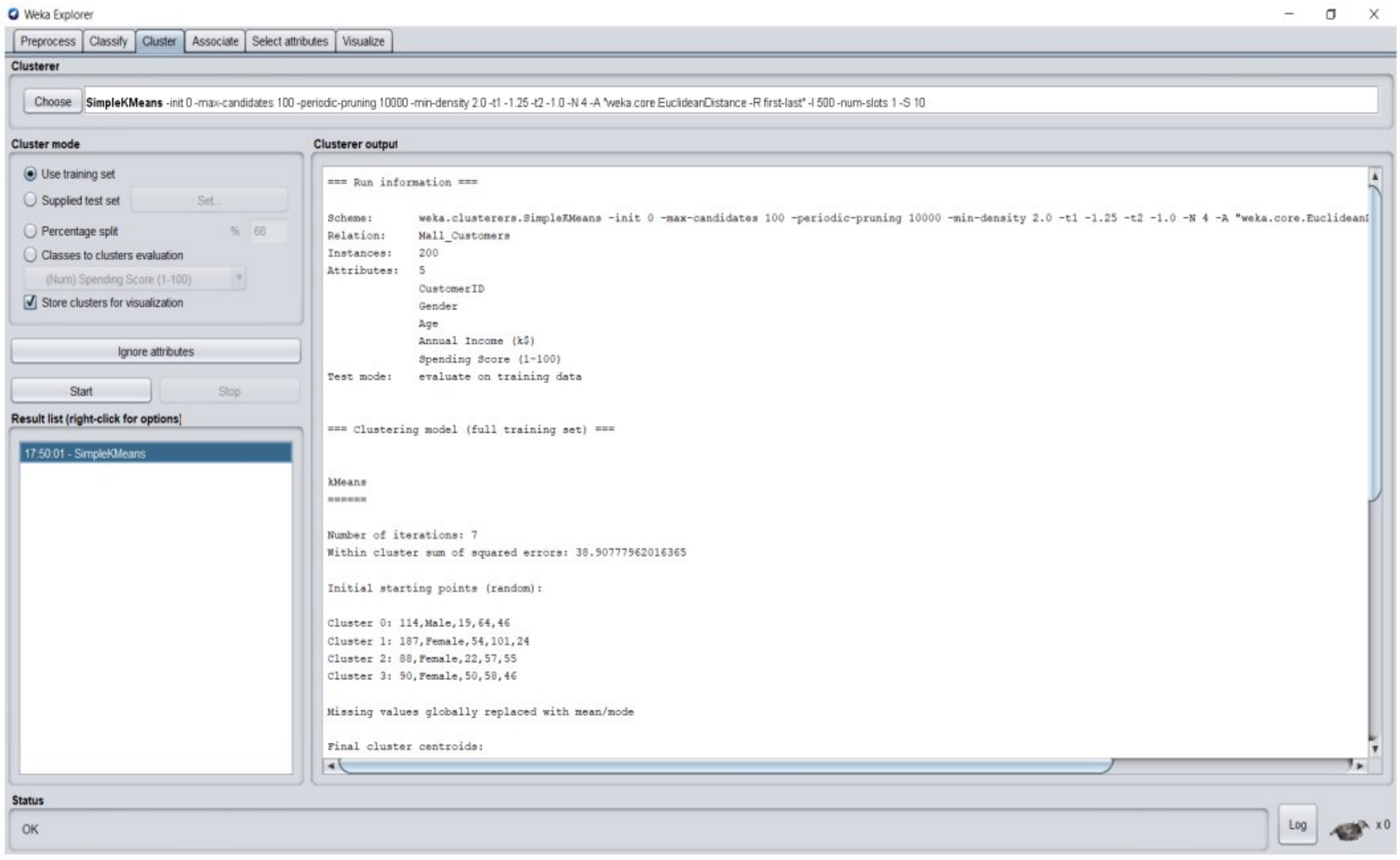


Fig.no.4

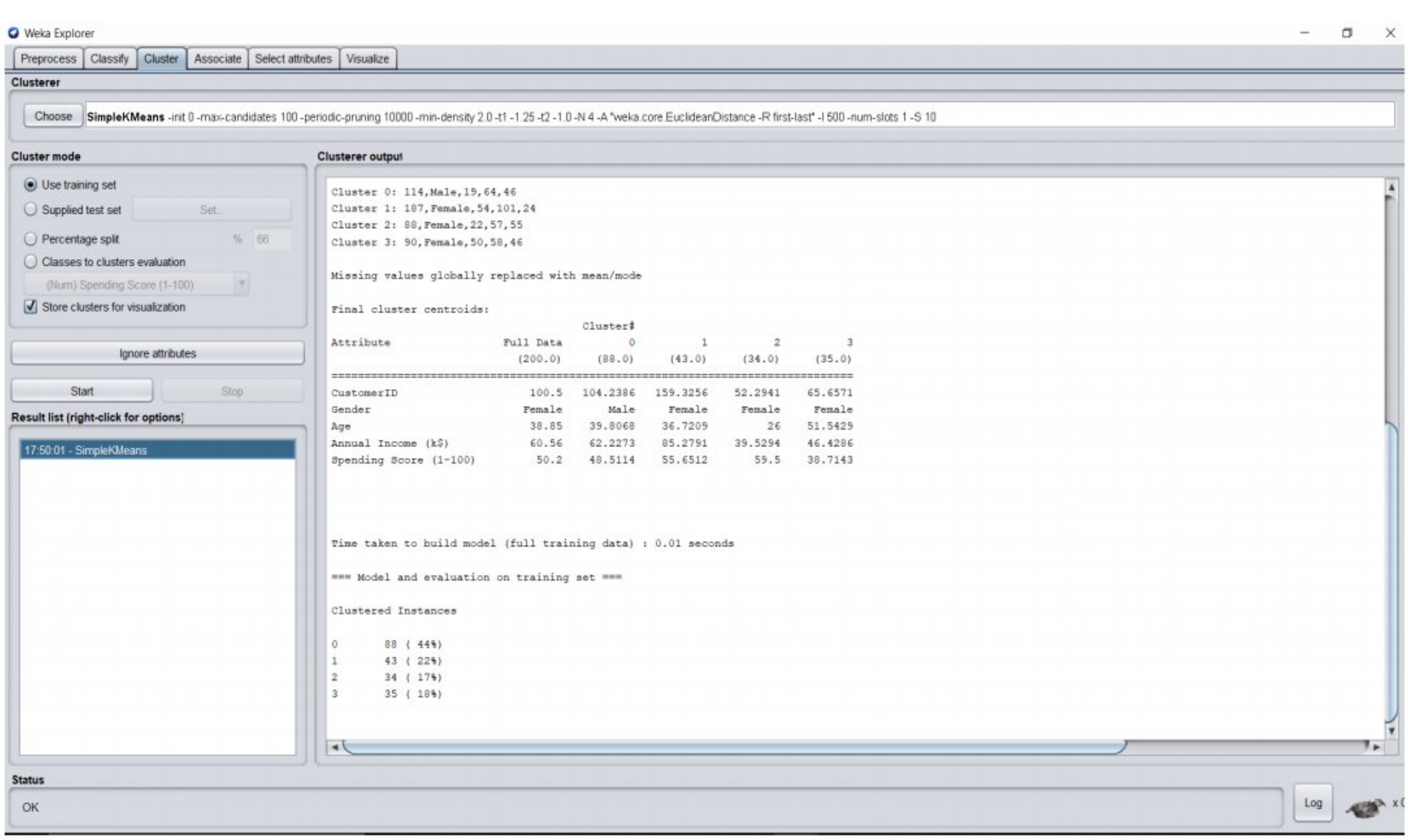


Fig.no.5

The above result window screenshot shows the centroid of each cluster as well as statistics on the number and percentage of instances assigned to different clusters. Cluster centroids are the mean vectors for each cluster; each dimension value in the centroid represents the mean value for that dimension in the cluster. Thus, centroids can be used to characterize the clusters.

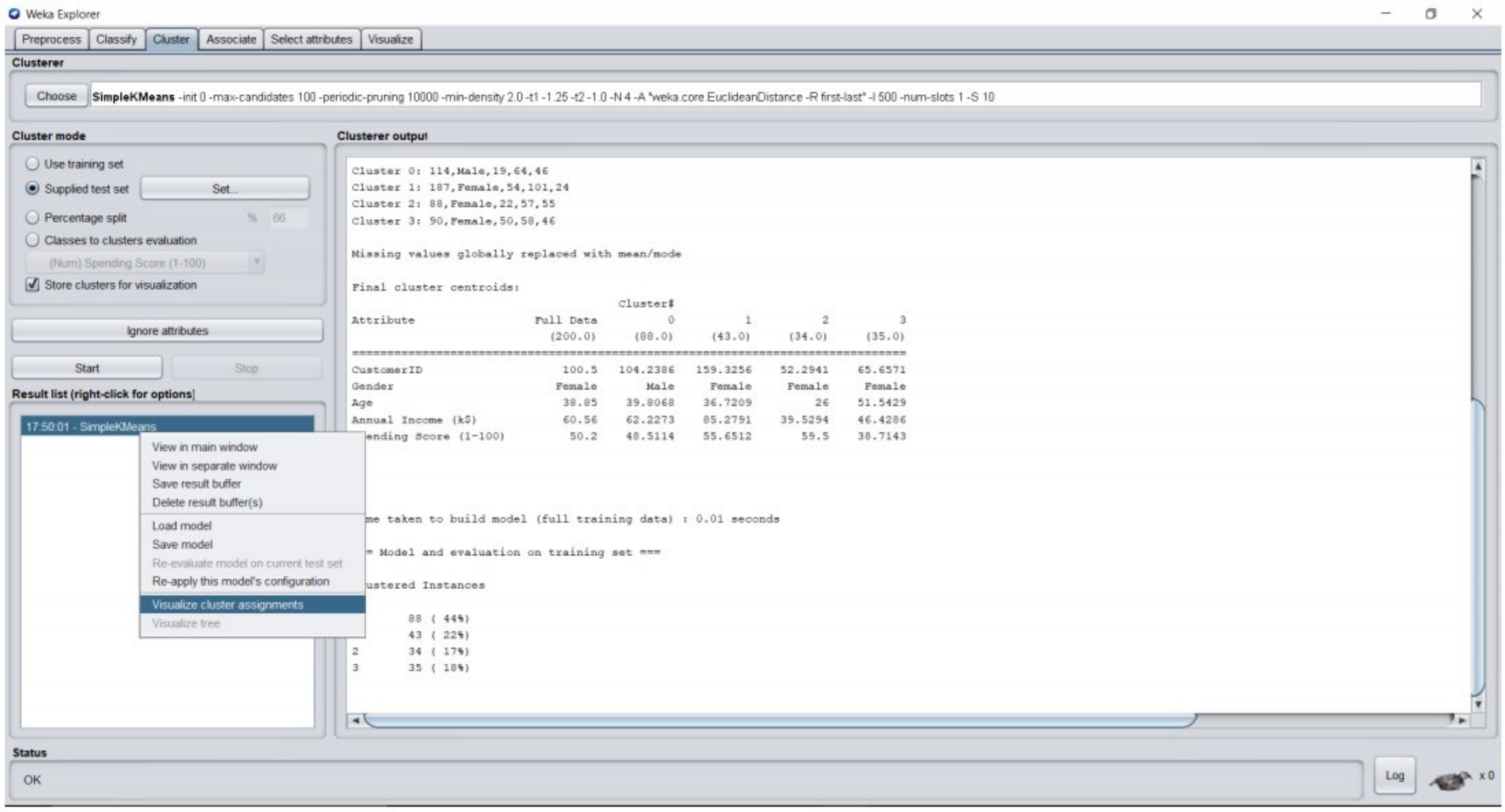


Fig.no.6

The above screenshot shows another way of understanding the characteristics of each cluster through visualization. This can be done by right-clicking the result set on the left "Result list" panel and selecting "Visualize Cluster assignments".

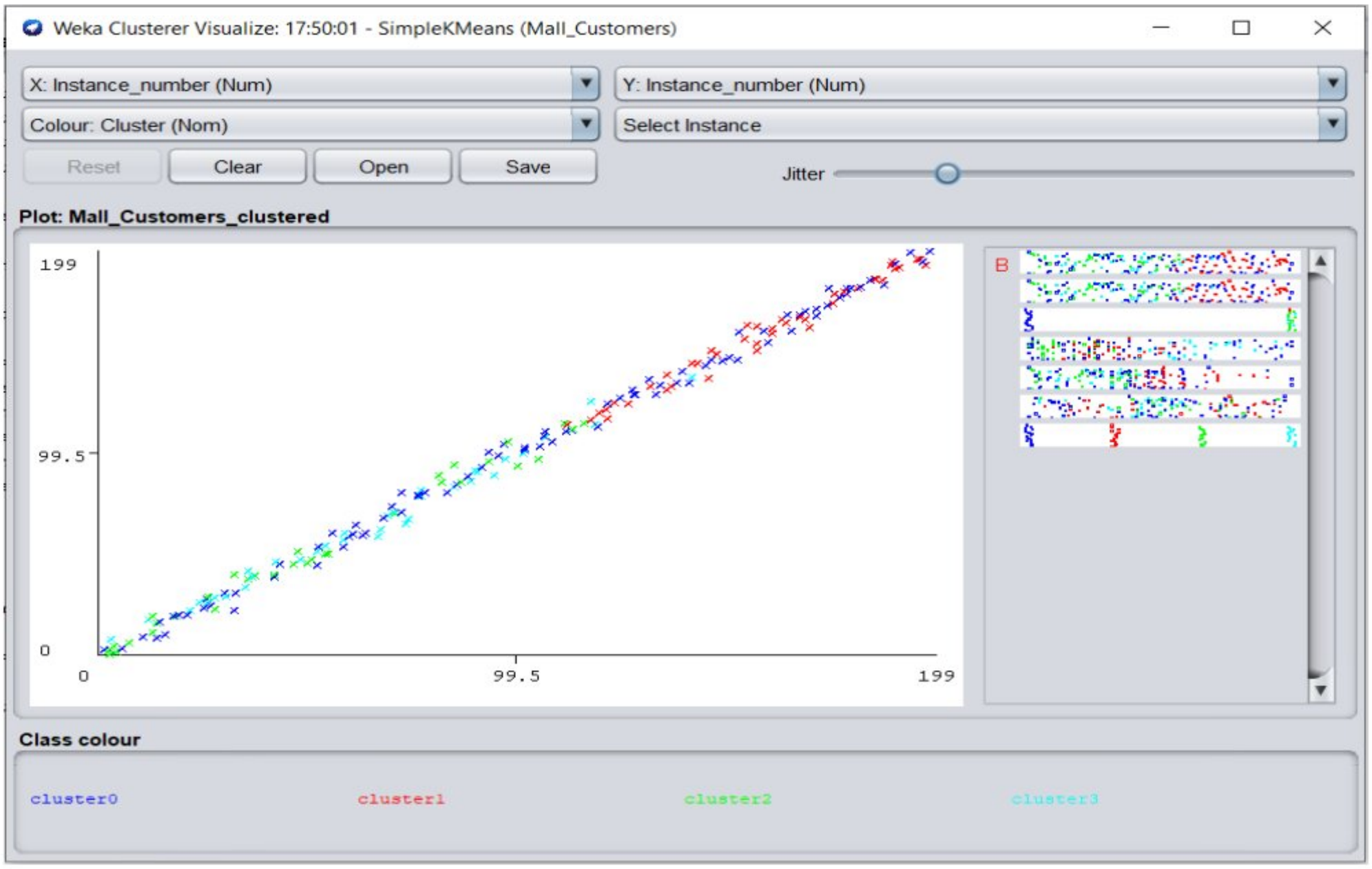


Fig.no.7

You can choose the cluster number and any of the other attributes for each of the three different dimensions available (x-axis, y-axis, and color). Different combinations of choices will result in visual rendering of different relationships within each cluster. In the above screenshot, we have chosen the instance number as the x-axis, the instance number (assigned by WEKA) as the y-axis,and the cluster as the color dimension. This will result in a visualization of the distribution of clusters .In this case, by changing the color dimension to other attributes, we can see their distribution within each of the clusters.

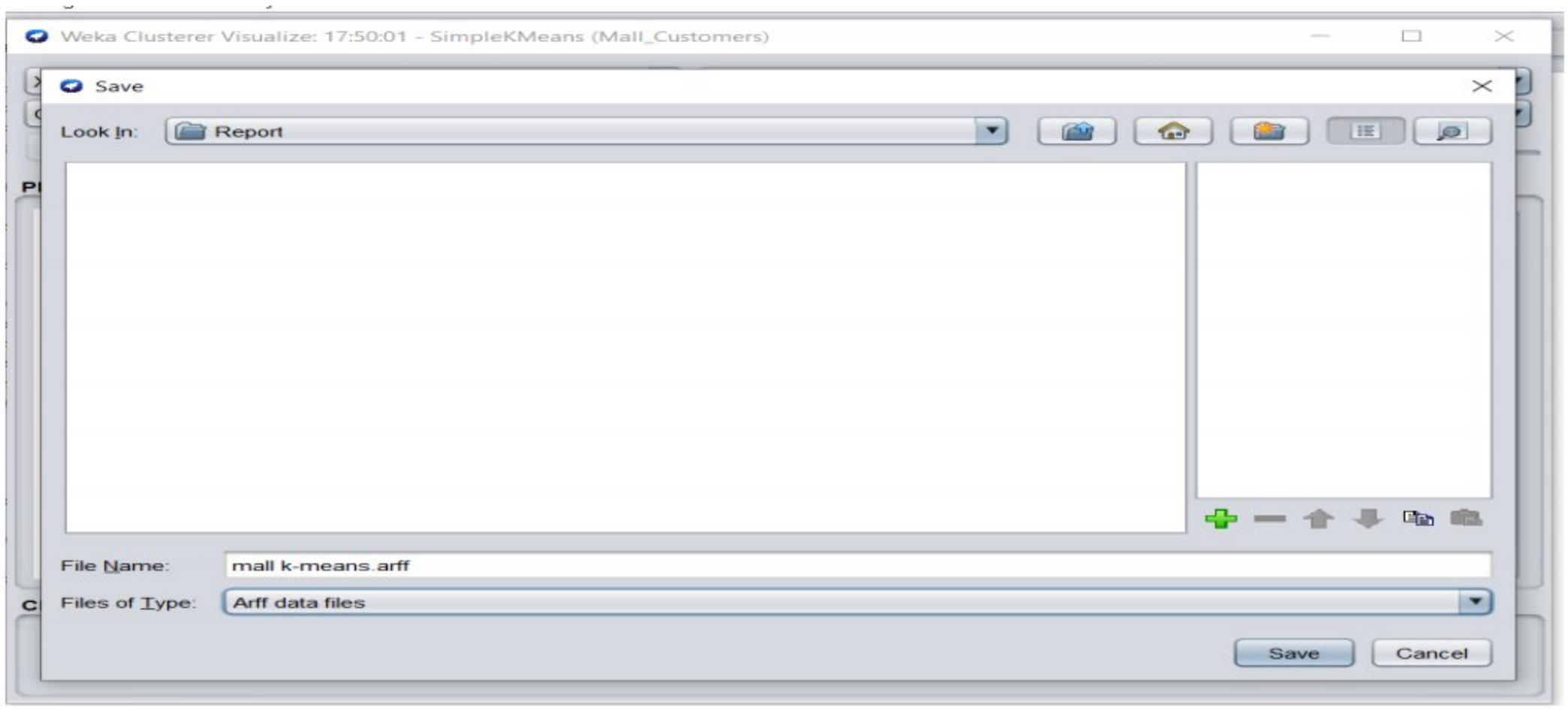


Fig.no.8

The above screenshot of saving results of K-means clusters which included each instance alongwith its assigned cluster.

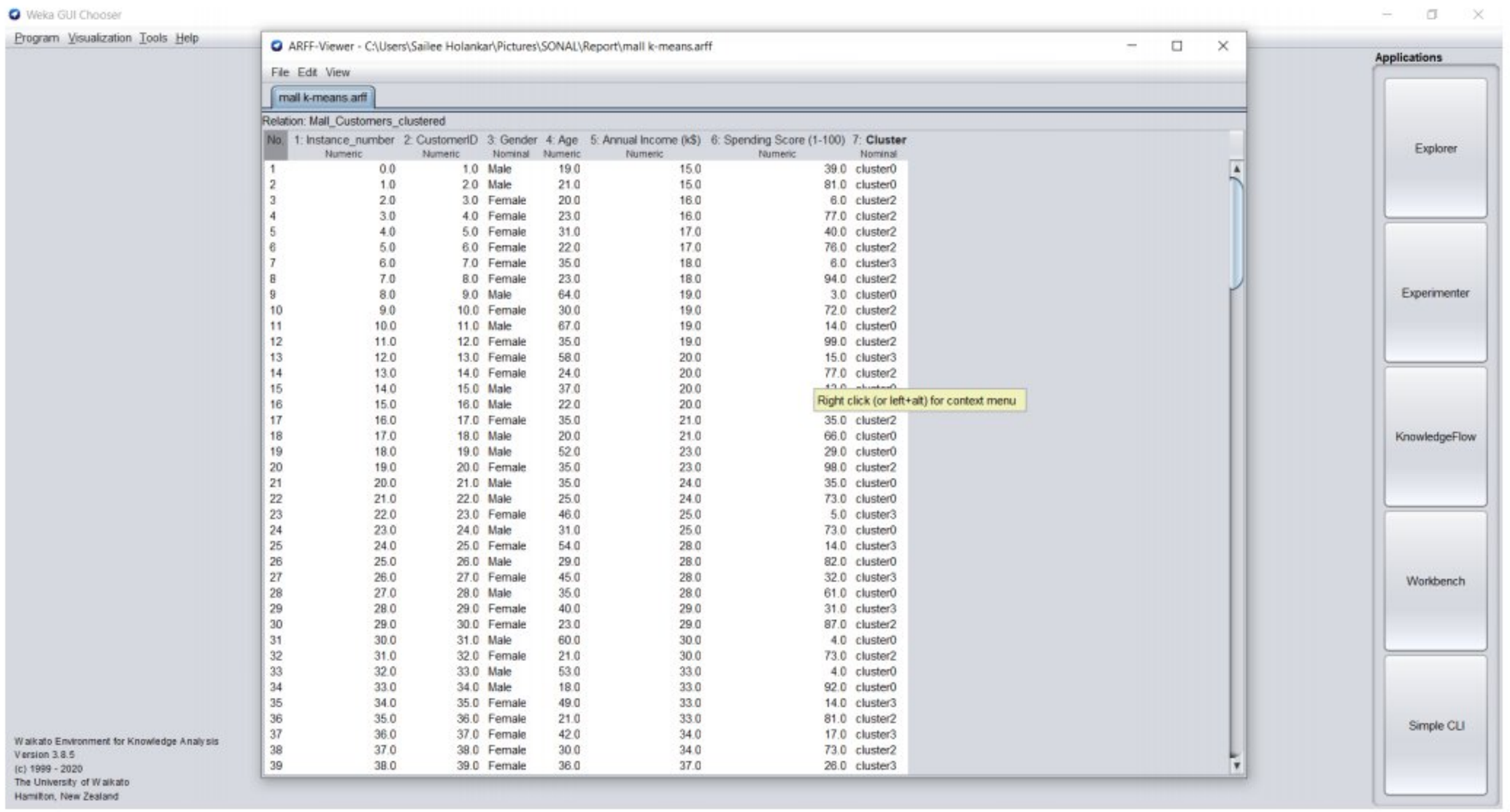


Fig.no.9

The above screenshot is of mall k-means.arff file which is opened in the ARFF viewer option from tools and it shows all customers or instances with assigned clusters.

Association & Classification Algorithms on Mall Customer Dataset :

Data preprocessing :It is a data mining technique which is used to transform the raw data in a useful and efficient format.

Data preprocessing procedure :

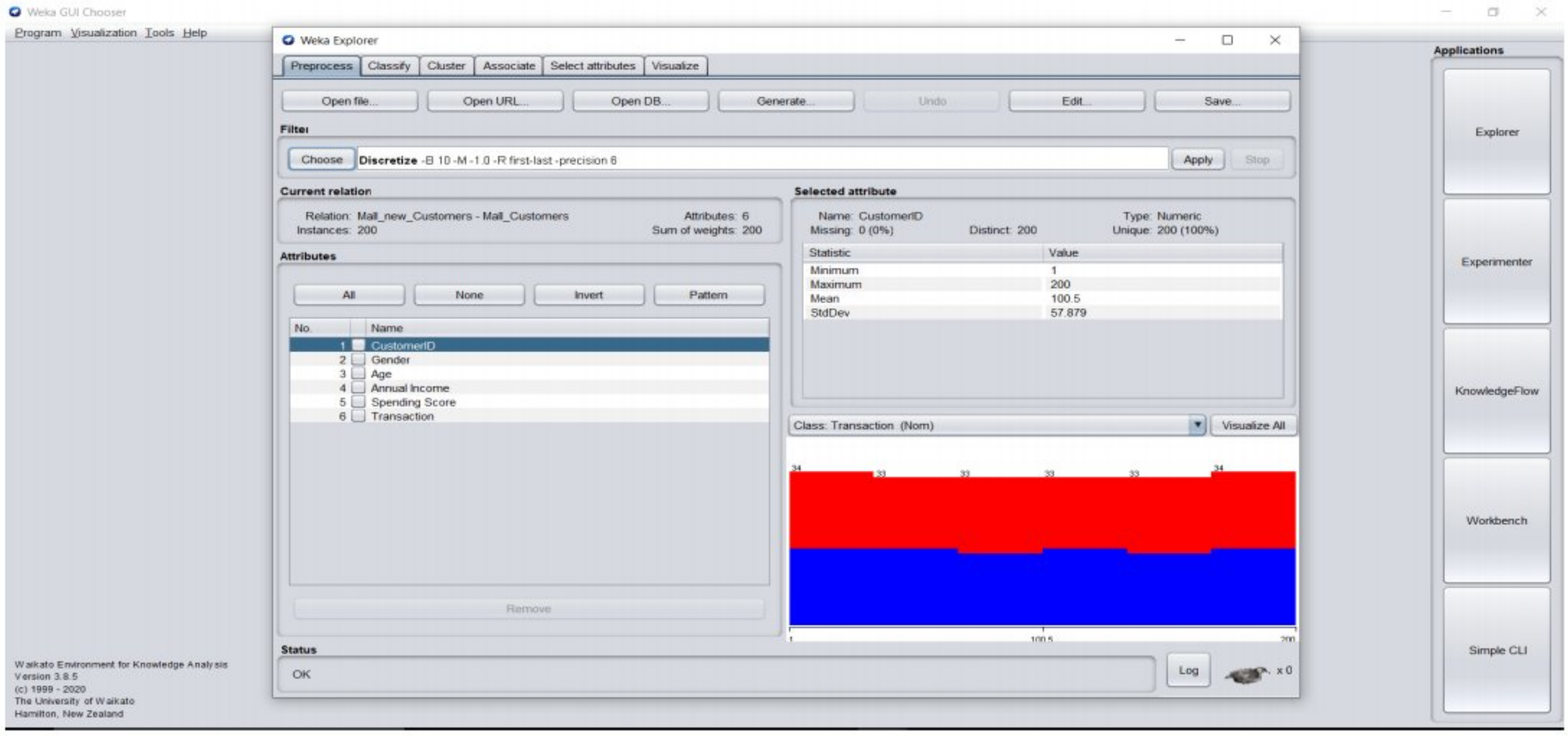


Fig.no.10

The above screenshot is of the Mall Customer dataset which is opened in the weka tool in order to preprocess data. The attributes like CustomerID, Gender, Spending score, Annual Income of dataset with their required types in order to perform steps on the data set. Transaction attribute is added to perform other various algorithms like classification, association etc. Therefore the dataset has to be preprocessed.

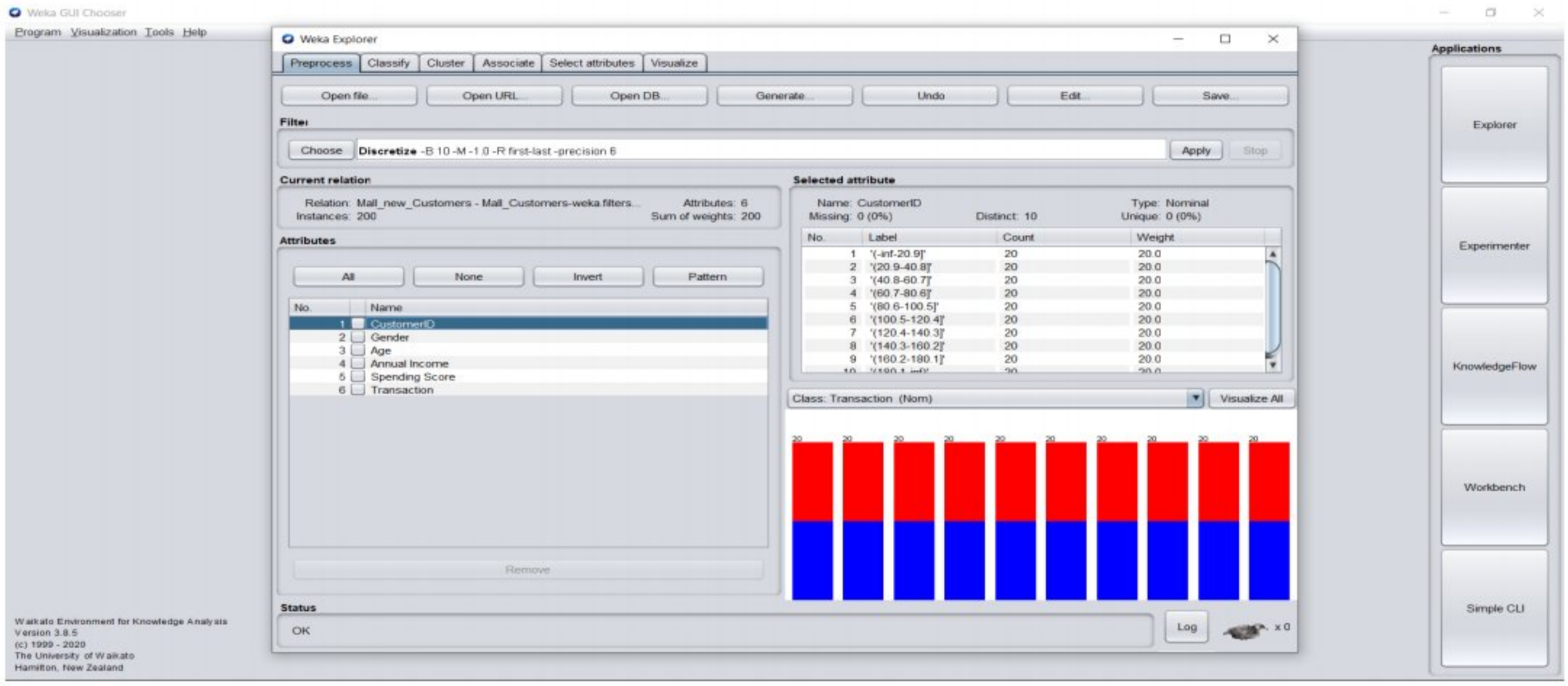


Fig.no.11

In the above screenshot, we applied a discretize task on the mall customer dataset.Some techniques, such as association rule mining, can only be performed on categorical data. This requires performing discretization on numeric or continuous attributes.

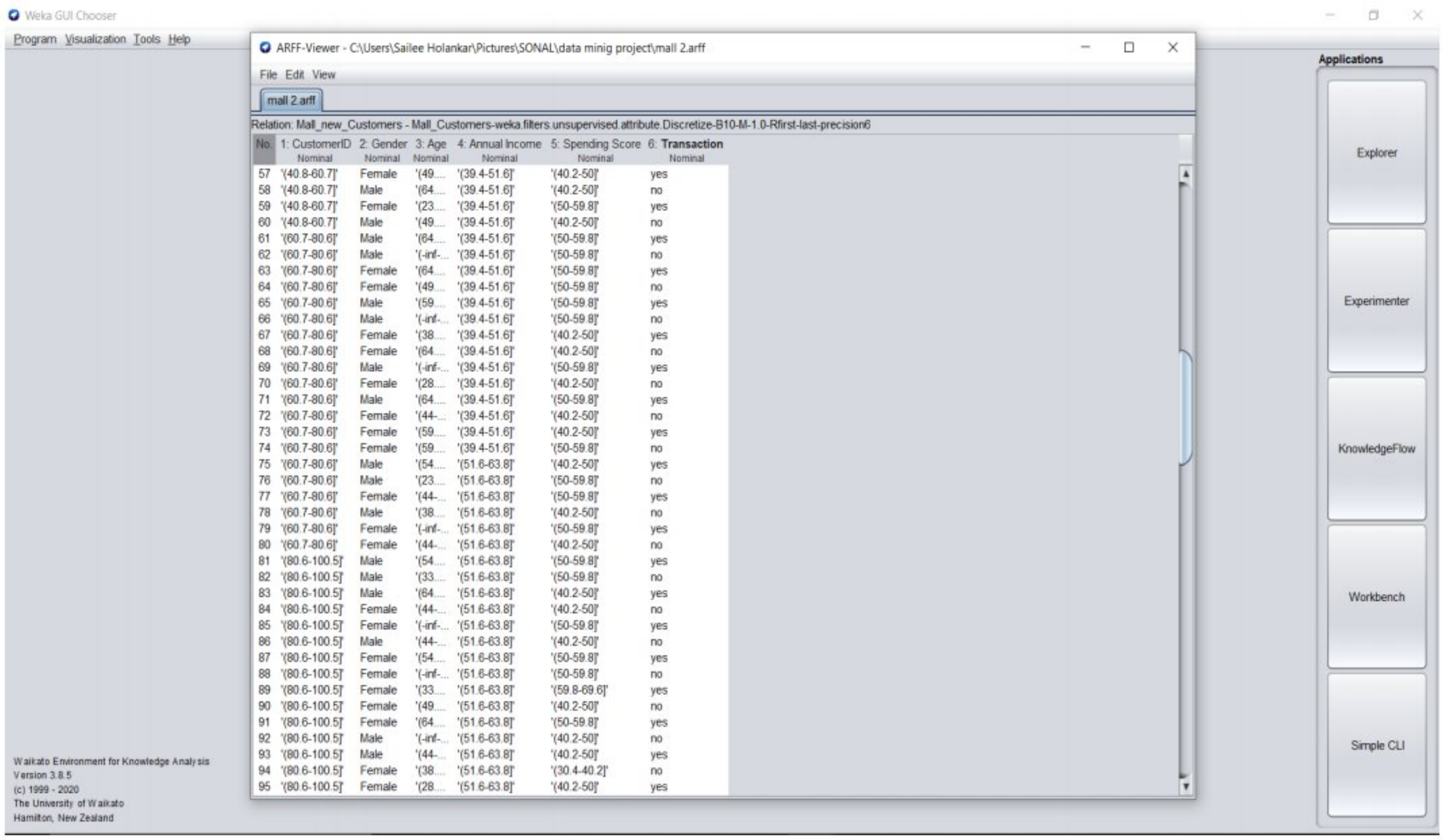


Fig.no.12

The above screenshot shows the attribute's values are replaced with bin values which is clearly shown when the dataset file is opened in the ARFF viewer tool

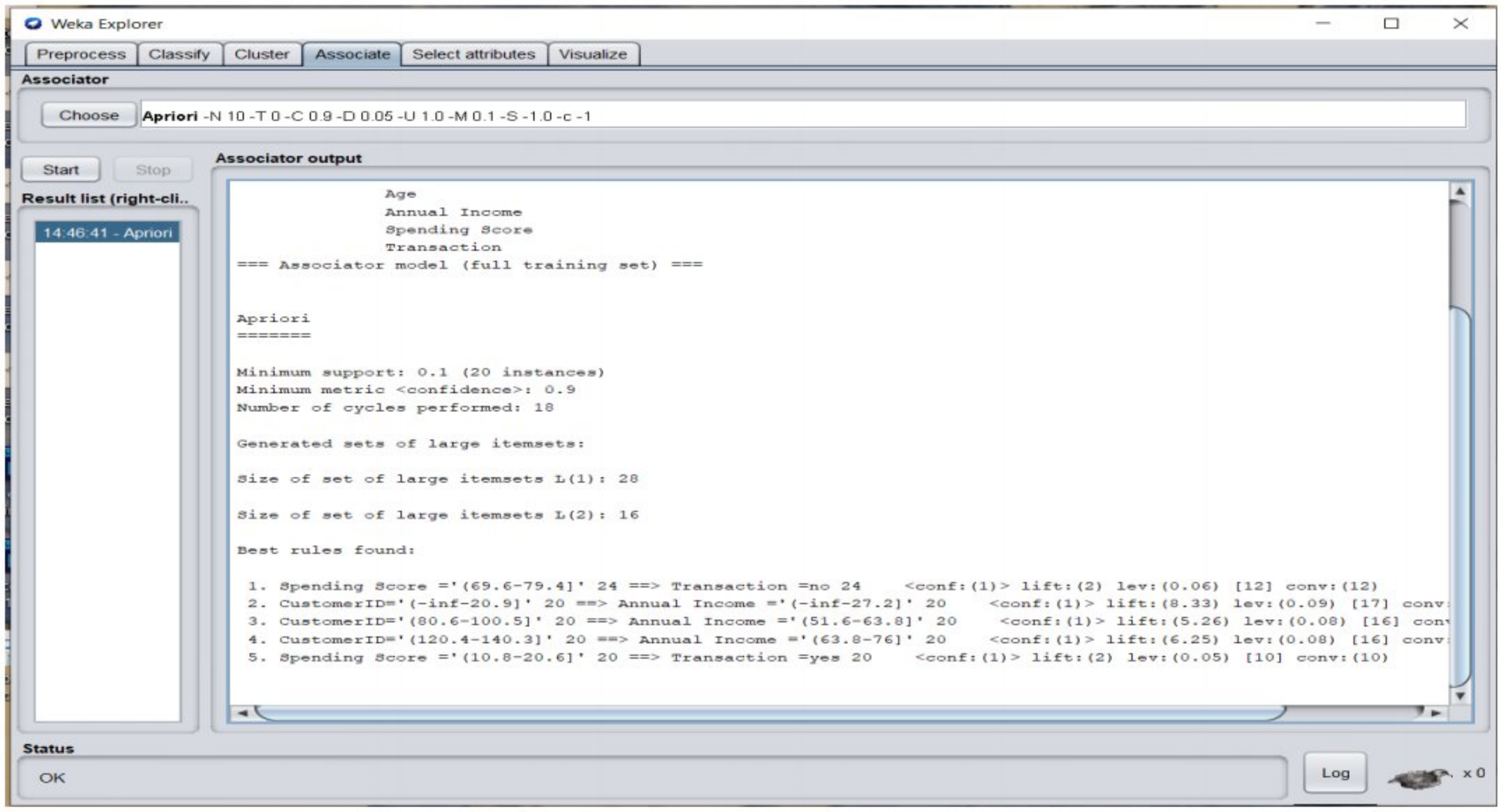


Fig.no.13

The above screenshot of association algorithm is performed on a mall customer dataset after applying discretization on the dataset. It shows some basic rules that can be achieved on the dataset. Sometimes dataset is not made for all the algorithms then we have to apply more preprocessing tasks on the dataset.

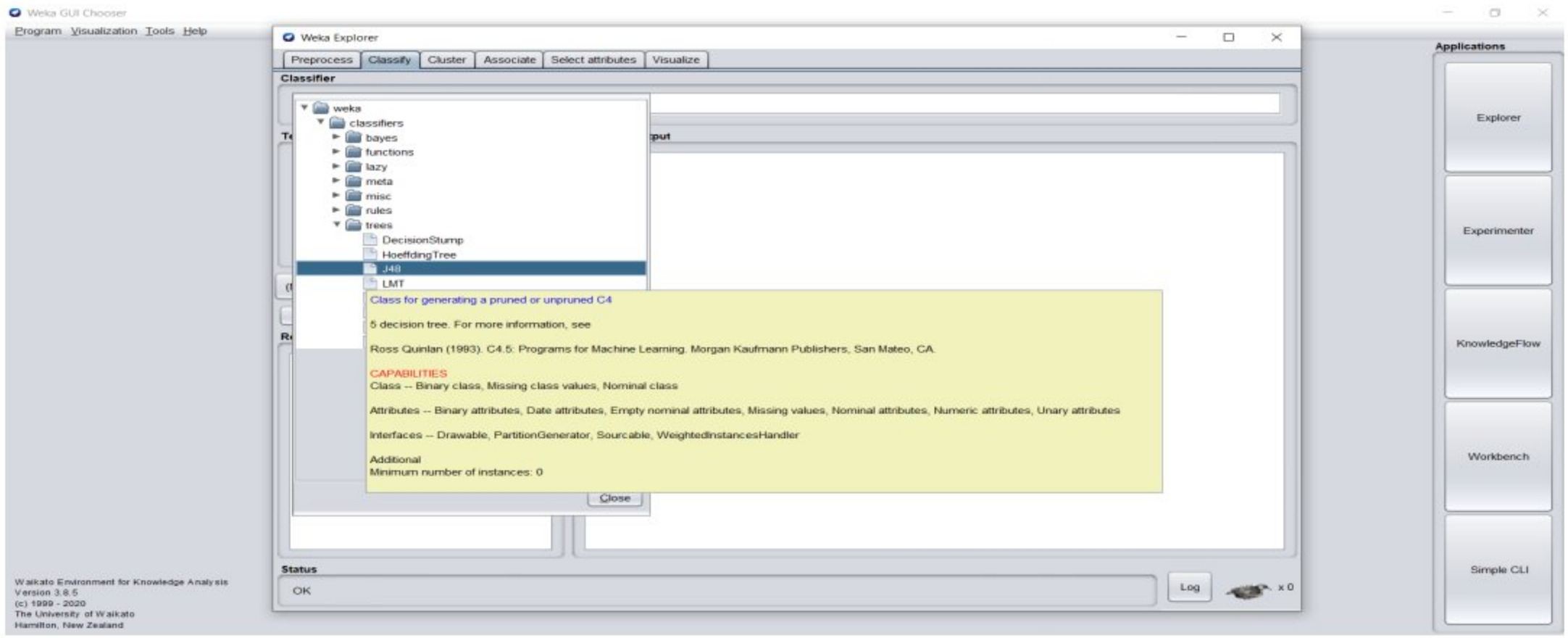


Fig.no.14

In the above screenshot, to perform classification on the dataset we have to click on the classify tab and choose the J48 classifier as an option from the option list.

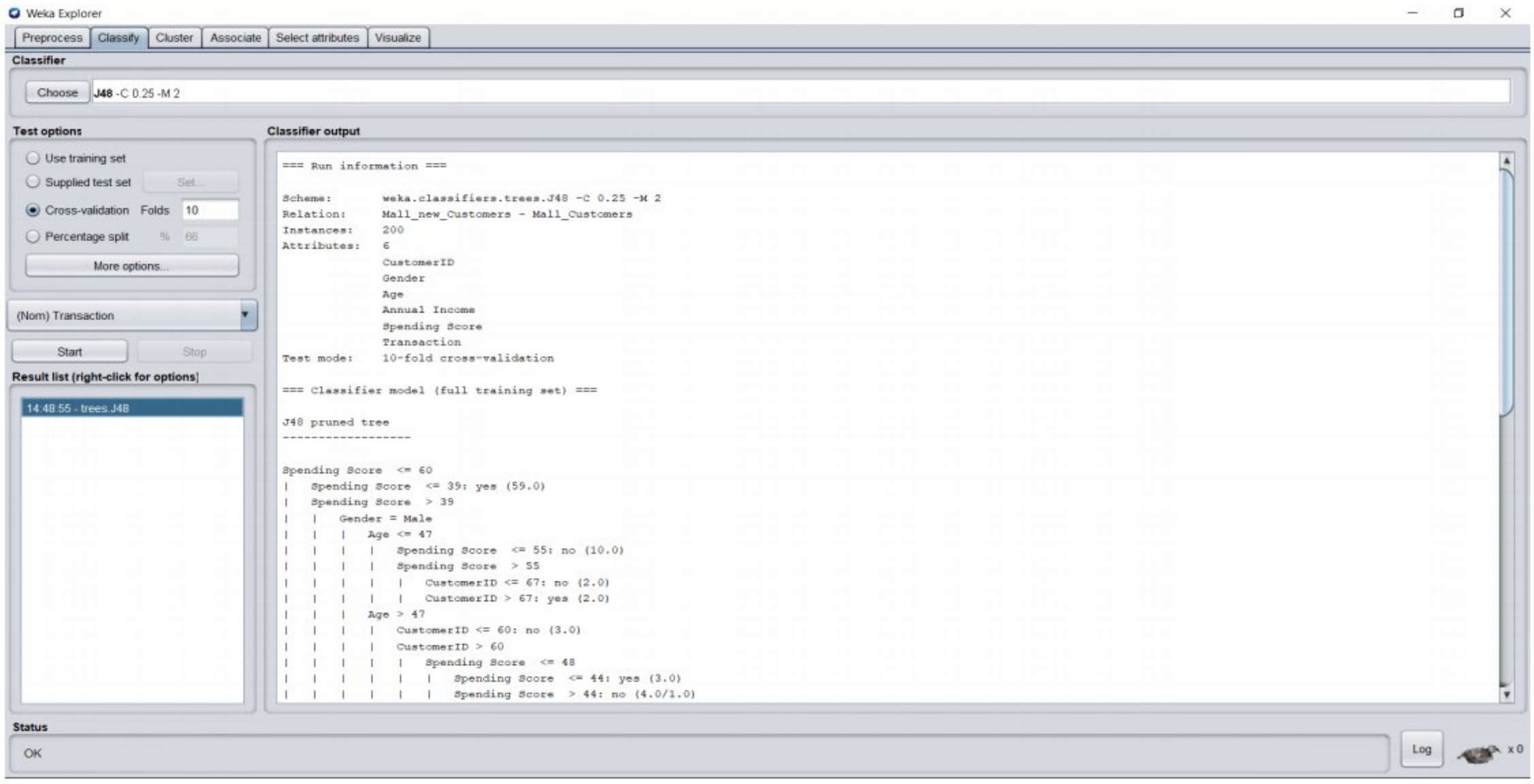


Fig.no.15

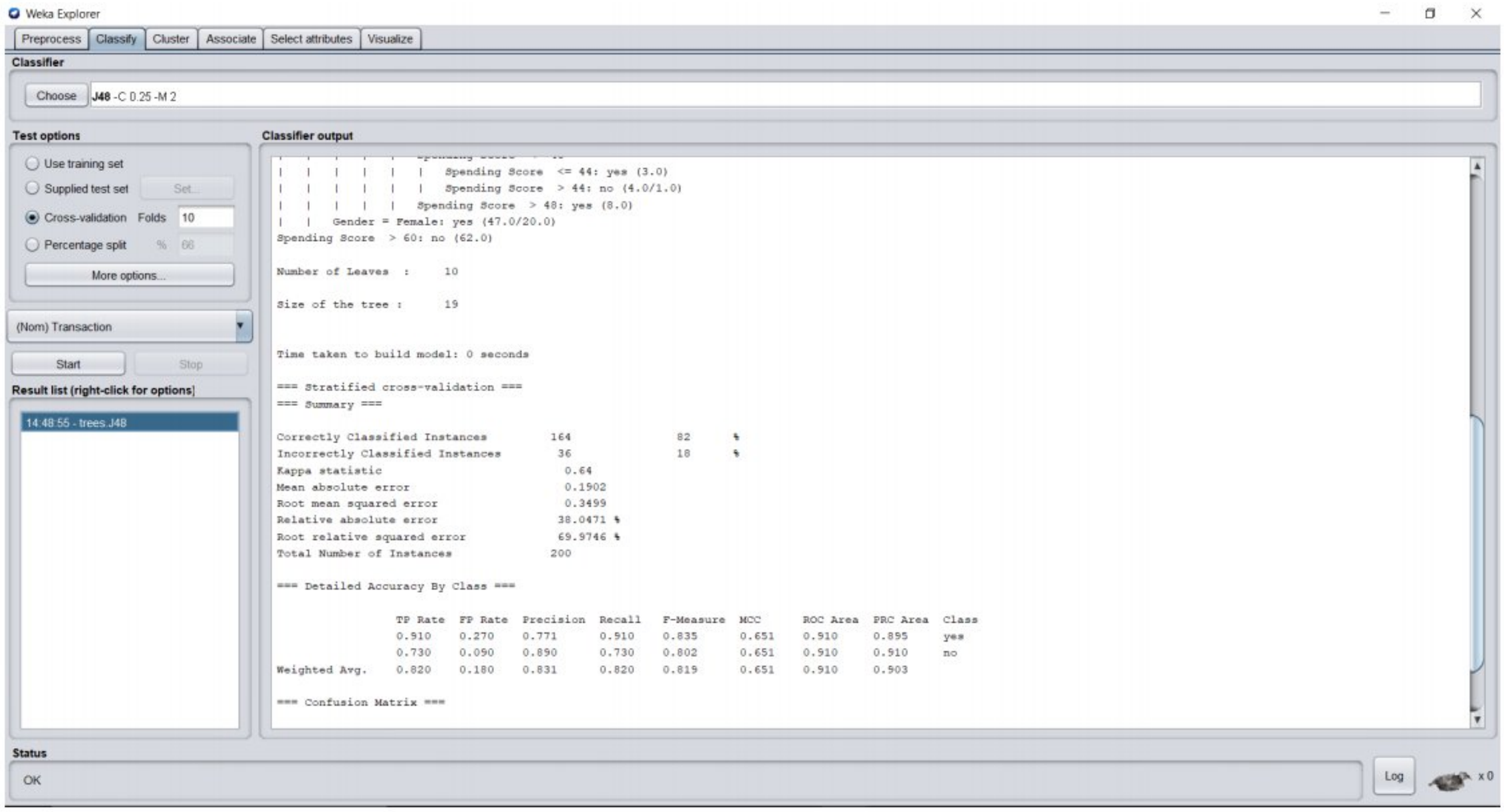


Fig.no.16

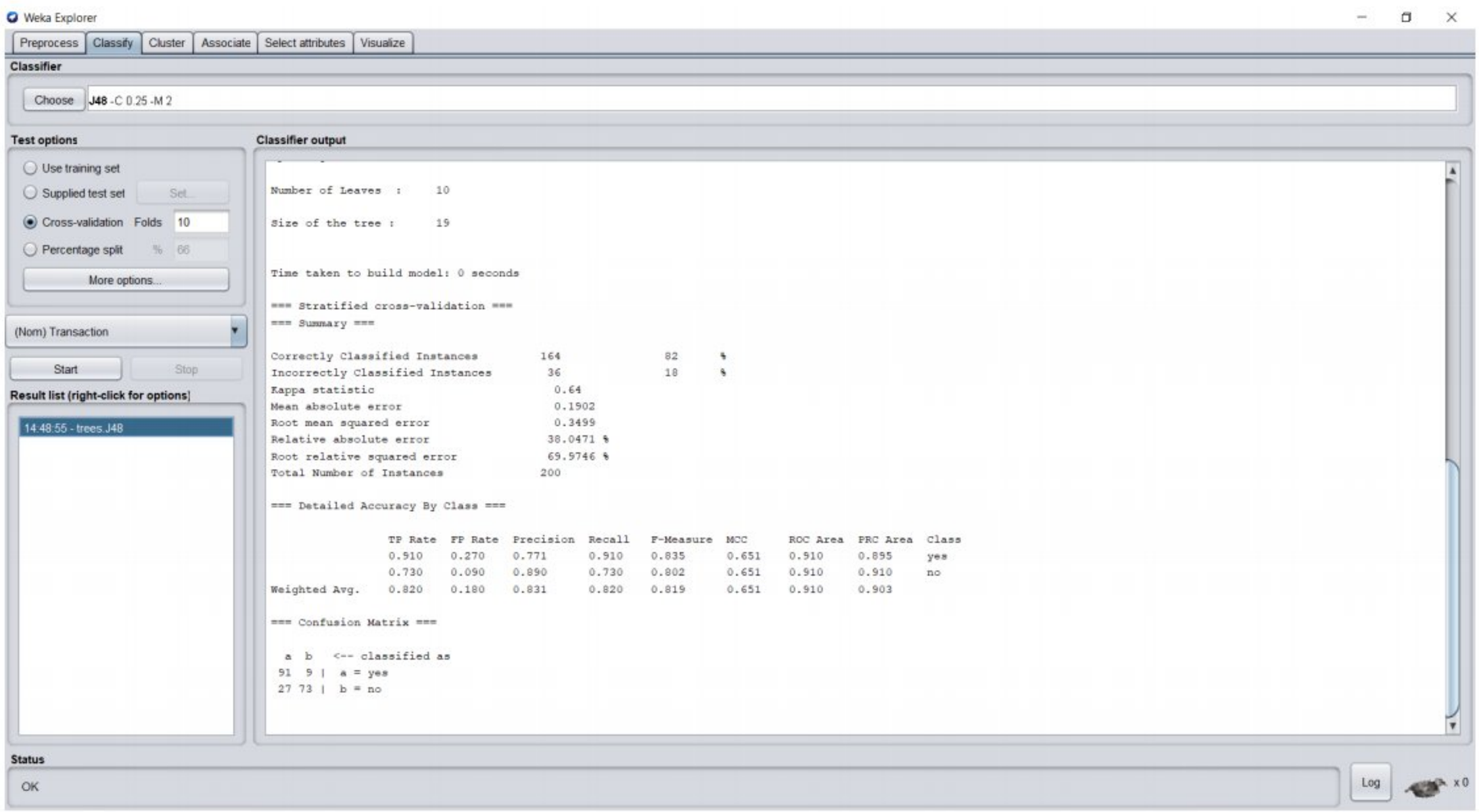


Fig.no.17

The above screenshots are of a classification result of a dataset under the Test options in the main panel we select 10-fold cross-validation as our evaluation approach. We can see the outputs in a separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model. These screenshots show the information about the detailed accuracy , confusion matrix , correctly classified and unclassified instances of model .

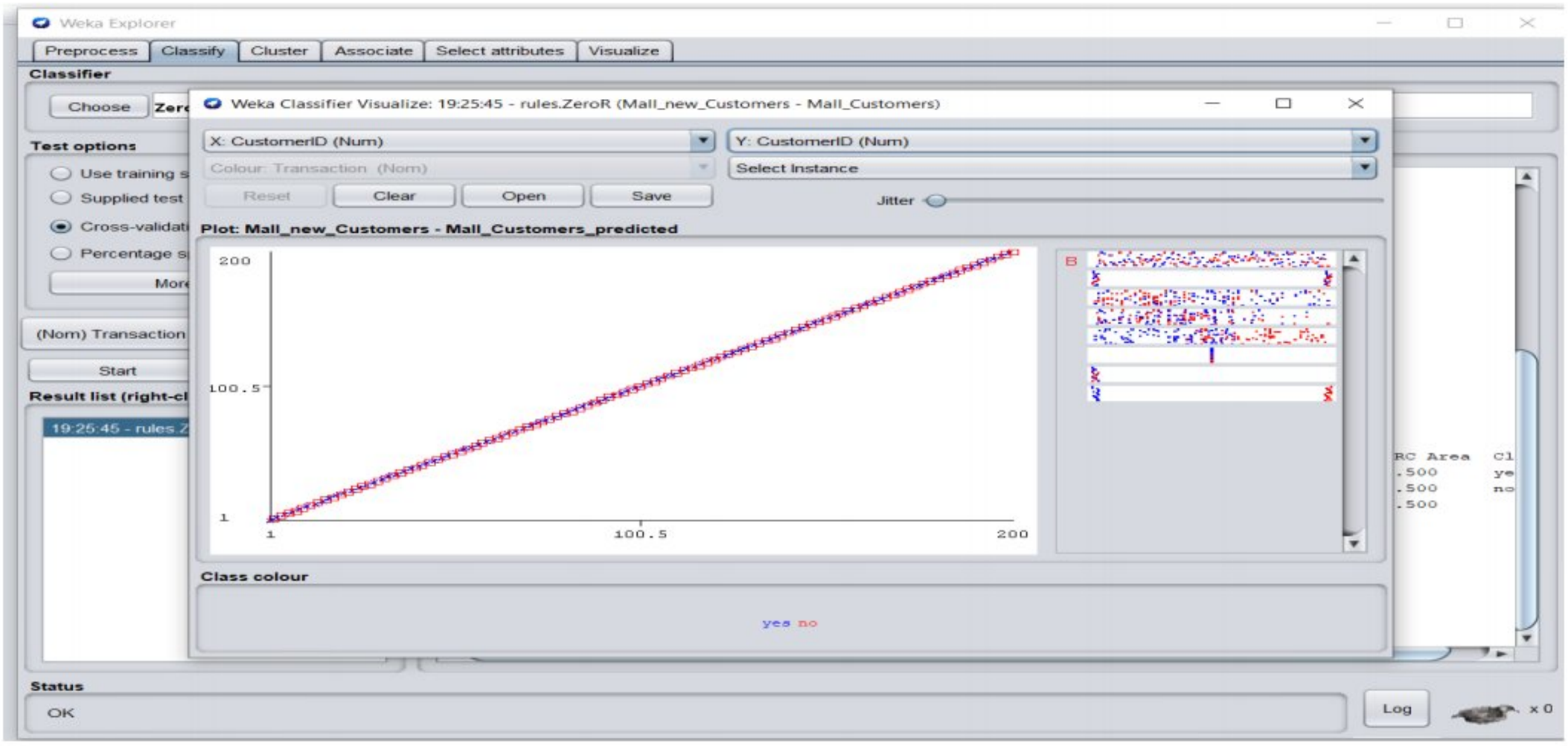
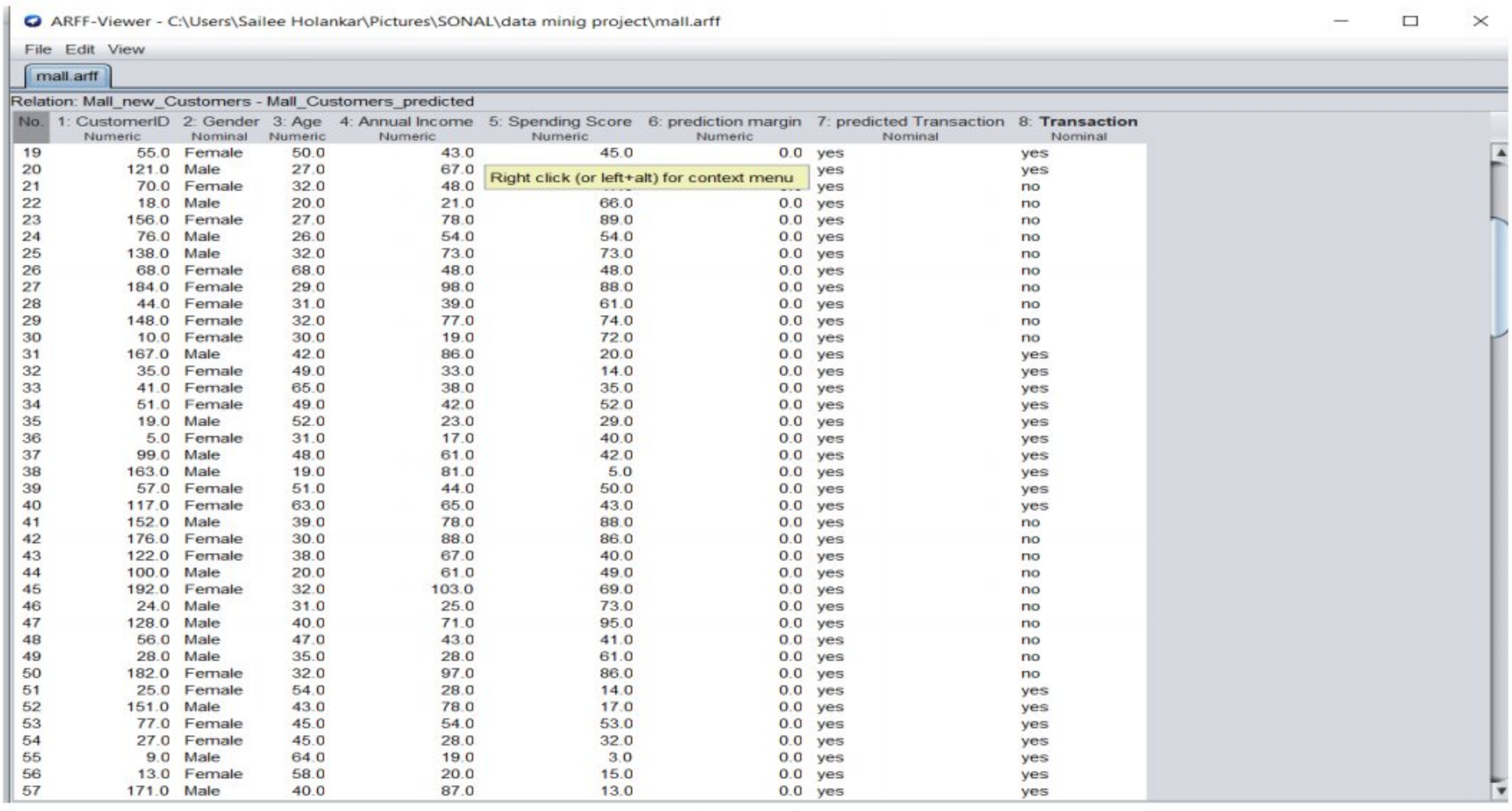


Fig.no.18



No.	1: CustomerID Numeric	2: Gender Nominal	3: Age Numeric	4: Annual Income Numeric	5: Spending Score Numeric	6: prediction margin Numeric	7: predicted Transaction Nominal	8: **Transaction** Nominal
19	55.0	Female	50.0	43.0	45.0	0.0	yes	yes
20	121.0	Male	27.0	67.0			yes	yes
21	70.0	Female	32.0	48.0			yes	no
22	18.0	Male	20.0	21.0	66.0	0.0	yes	no
23	156.0	Female	27.0	78.0	89.0	0.0	yes	no
24	76.0	Male	26.0	54.0	54.0	0.0	yes	no
25	138.0	Male	32.0	73.0	73.0	0.0	yes	no
26	68.0	Female	68.0	48.0	48.0	0.0	yes	no
27	184.0	Female	29.0	98.0	88.0	0.0	yes	no
28	44.0	Female	31.0	39.0	61.0	0.0	yes	no
29	148.0	Female	32.0	77.0	74.0	0.0	yes	no
30	10.0	Female	30.0	19.0	72.0	0.0	yes	no
31	167.0	Male	42.0	86.0	20.0	0.0	yes	yes
32	35.0	Female	49.0	33.0	14.0	0.0	yes	yes
33	41.0	Female	65.0	38.0	35.0	0.0	yes	yes
34	51.0	Female	49.0	42.0	52.0	0.0	yes	yes
35	19.0	Male	52.0	23.0	29.0	0.0	yes	yes
36	5.0	Female	31.0	17.0	40.0	0.0	yes	yes
37	99.0	Male	48.0	61.0	42.0	0.0	yes	yes
38	163.0	Male	19.0	81.0	5.0	0.0	yes	yes
39	57.0	Female	51.0	44.0	50.0	0.0	yes	yes
40	117.0	Female	63.0	65.0	43.0	0.0	yes	yes
41	152.0	Male	39.0	78.0	88.0	0.0	yes	no
42	176.0	Female	30.0	88.0	86.0	0.0	yes	no
43	122.0	Female	38.0	67.0	40.0	0.0	yes	no
44	100.0	Male	20.0	61.0	49.0	0.0	yes	no
45	192.0	Female	32.0	103.0	69.0	0.0	yes	no
46	24.0	Male	31.0	25.0	73.0	0.0	yes	no
47	128.0	Male	40.0	71.0	95.0	0.0	yes	no
48	56.0	Male	47.0	43.0	41.0	0.0	yes	no
49	28.0	Male	35.0	28.0	61.0	0.0	yes	no
50	182.0	Female	32.0	97.0	86.0	0.0	yes	no
51	25.0	Female	54.0	28.0	14.0	0.0	yes	yes
52	151.0	Male	43.0	78.0	17.0	0.0	yes	yes
53	77.0	Female	45.0	54.0	53.0	0.0	yes	yes
54	27.0	Female	45.0	28.0	32.0	0.0	yes	yes
55	9.0	Male	64.0	19.0	3.0	0.0	yes	yes
56	13.0	Female	58.0	20.0	15.0	0.0	yes	yes
57	171.0	Male	40.0	87.0	13.0	0.0	yes	yes

Fig.no.19

In the above screenshots, we chose visualize classifier errors by right clicking on the result list option.This brings up a separate window containing a two-dimensional graph. The predicted transaction attribute is added to the mall customer dataset file.

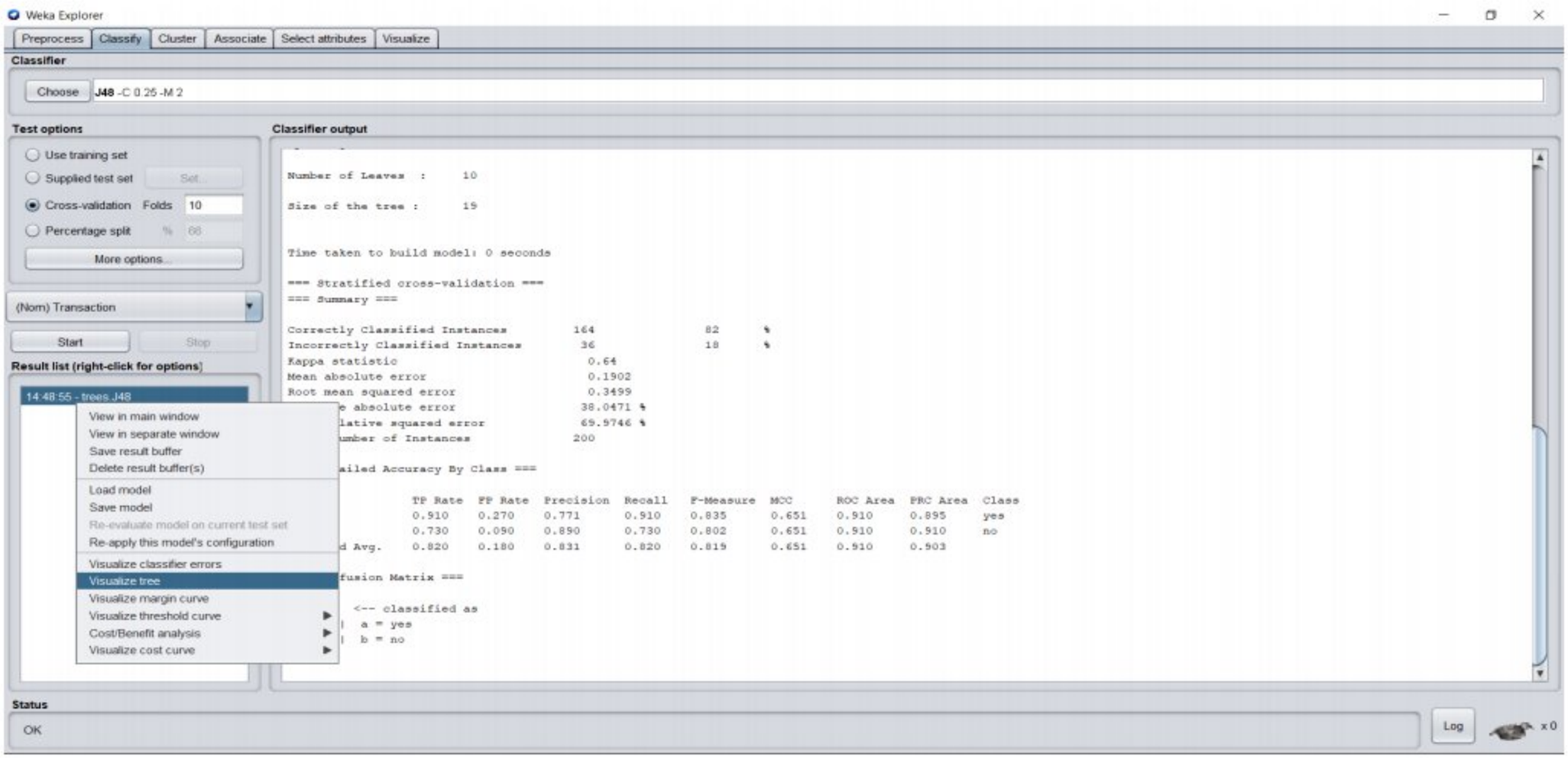


Fig.no.20

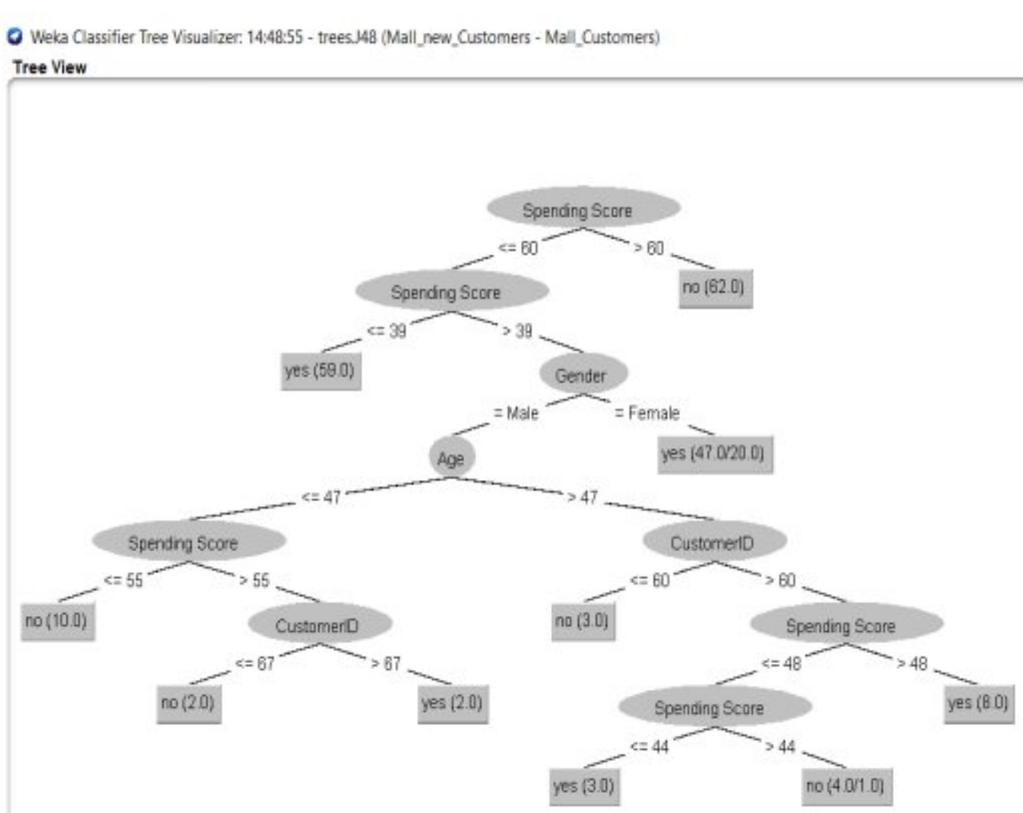


Fig.no.21

The above screenshots are of a graphical rendition of the classification tree which can be done by right clicking the last result set Visualize tree from the pop-up menu.The tree for this example is depicted Note that by resizing the window and selecting various menu items from inside the tree view using the right mouse button, we can adjust the tree view to make it more readable.

Introduction to Business Intelligence :

Business intelligence comprises the strategies and technologies used by enterprises for the data analysis of business information. BI technologies provide historical, current, and predictive views of business operations.Business intelligence leverages software and services to transform data into actionable insights that inform an organization's strategic and tactical business decisions.BI tools access and analyze data sets and present analytical findings in reports, summaries, dashboards, graphs, charts and maps to provide users with detailed intelligence about the state of the business.

Business Intelligence Tool :

- Business Intelligence (BI) tools are tools which utilize a set of methodologies and technologies to prepare, present and help analyze data. Through this process, data is turned into actionable business information which helps decision makers and end users to make more effective data-driven decisions.
- The set of methodologies and technologies used by business intelligence is widely diverse depending on the purpose of the solution. Some tools focus on the data preparation side of things and may include things like an ETL (Extract, Transform, Load) layer to better organize and utilize data.
- Some tools focus on wider enterprise use and may focus on data mashup to help businesses make organizational decisions based on information from disparate departmental systems. Some tools focus more on self-service capabilities and end-user experience.

Tableau Public For Business Intelligence :

Overview :

Tableau Software is an American interactive data visualization software company focused on business intelligence.Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. Tableau helps create the data that can be understood by professionals at any level in an organization. It also allows non-technical users to create customized dashboardsTableau is a visual analytics engine that makes it easier to create interactive visual analytics in the form of dashboards. These dashboards make it easier for non-technical analysts and

end users to convert data into understandable, interactive graphics.

Importance of Tableau Software :

As the market-leading choice for modern business intelligence, analytics platform makes it easier for people to explore and manage data, and faster to discover and share insights that can change businesses and the world.Everything is driven by tableau software. Tableau software's mission is to help people see and understand data, which is why tableau products are designed to put the user first—whether they're an analyst, data scientist, student, teacher, executive, or business user. From connection through collaboration, Tableau is the most powerful, secure, and flexible end-to-end analytics platform.

Tableau Software With Various Products:

Tableau Public: Tableau Public is essentially a free version of Tableau visualization software. It allows you to use most of the software functions.

Tableau Desktop: Tableau Desktop, largely used for business intelligence is a data visualization software that transforms huge amounts of data (mostly statistical data) into interactive visual representations, such as graphs and charts.Tableau Public does not allow you to save your workbooks locally.

Tableau Server: Tableau Server users will be able to access up-to-date content and gain quick insights without relying on static distributed content.As a Tableau Server administrator you will control who has access to server content to help protect sensitive data. Administrators can set user permissions on projects, workbooks, views, and data sources.

Tableau Online: Tableau Online is a hosted version of Tableau Server.Tableau Online makes business intelligence faster and easier than ever before. Publish dashboards with Tableau Desktop and share them with colleagues, partners or customers.

Business Intelligence Results with Data set :

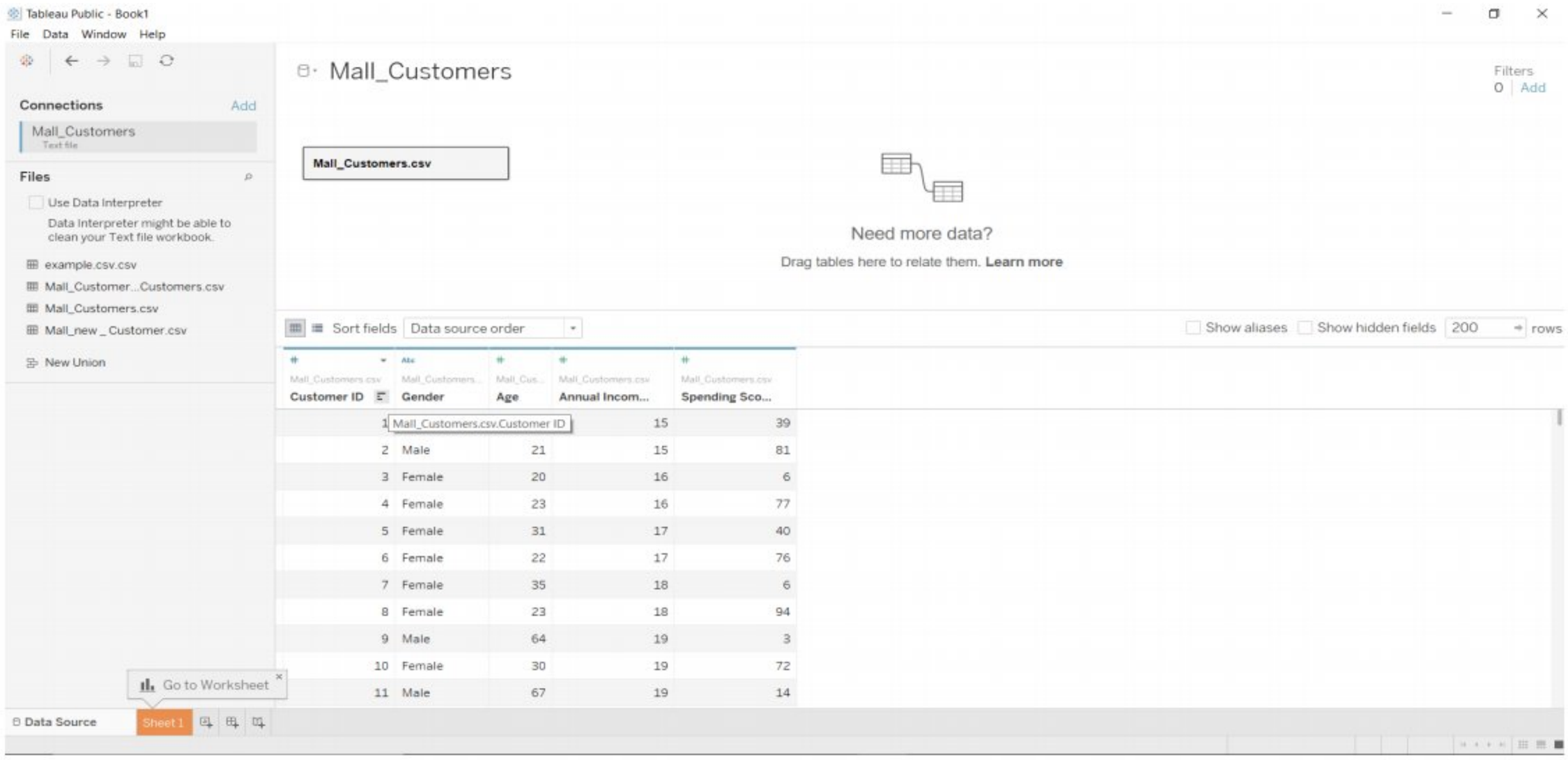


Fig.no.23

The above mentioned screenshot is of the main or first window of tableau public software.We loaded Mall Customer.csv file format in the tableau software. The tableau software does not support .arff file extension.

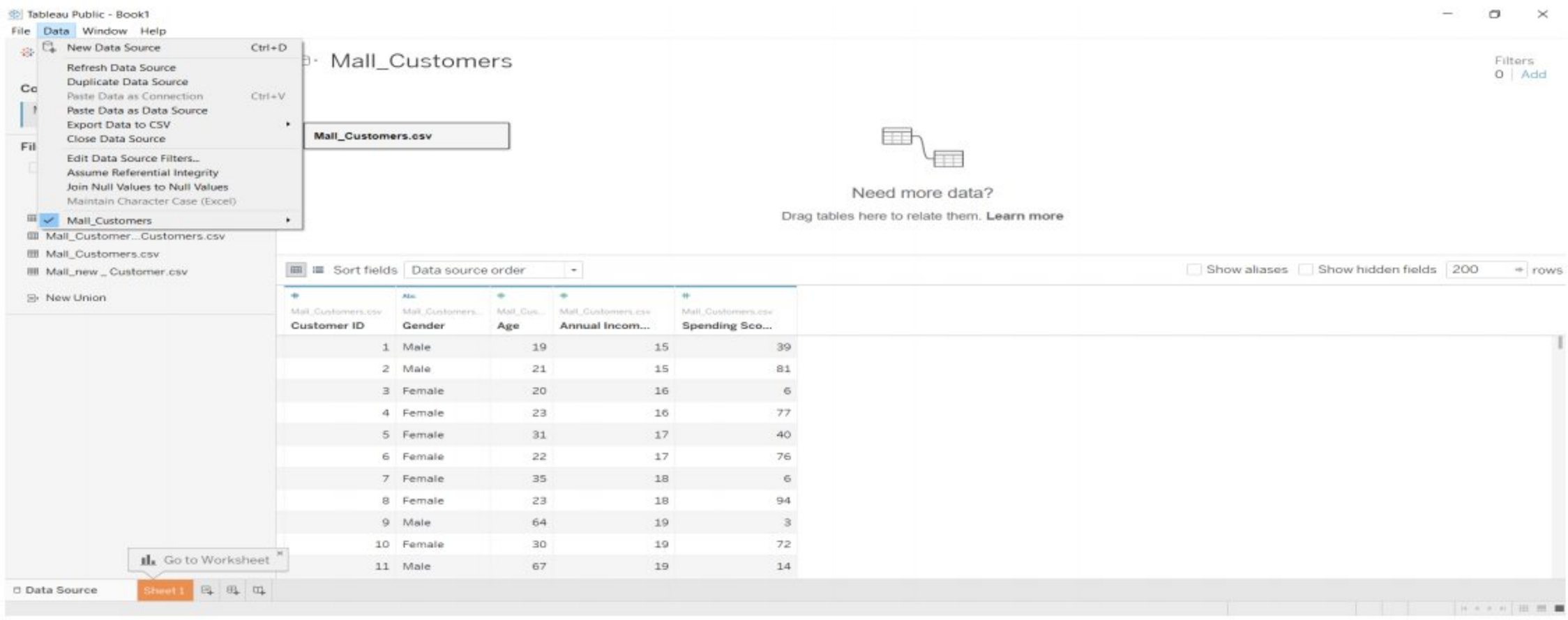


Fig.no.24

The above screenshot shows the various menus and options are provided by the multiple menus in the tableau public.

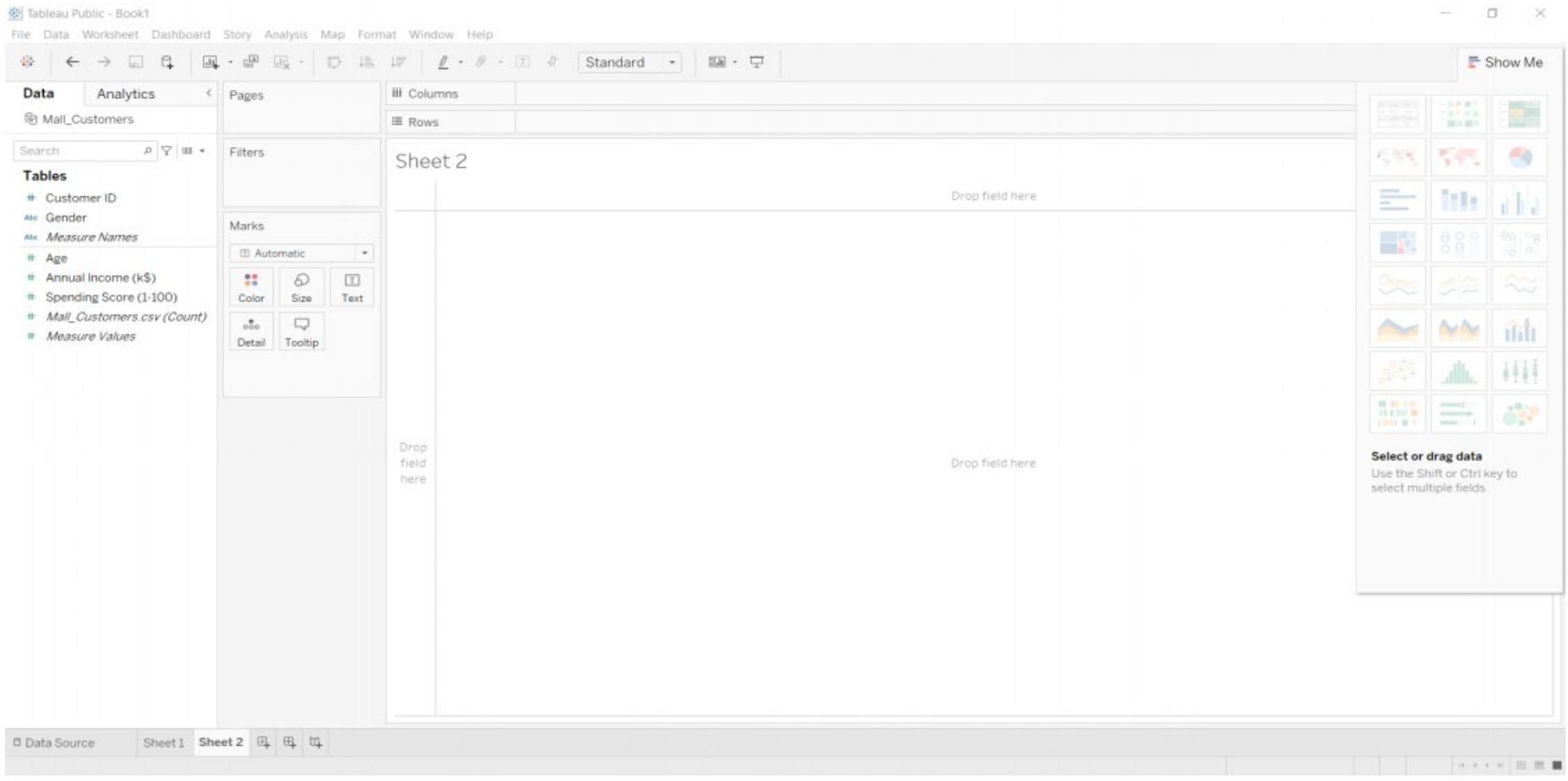


Fig.no.25

The above screenshot is of a sheet which is used to drag attributes in the columns and rows and it provides suggestions on which data visualization options are appropriate to show attribute details.

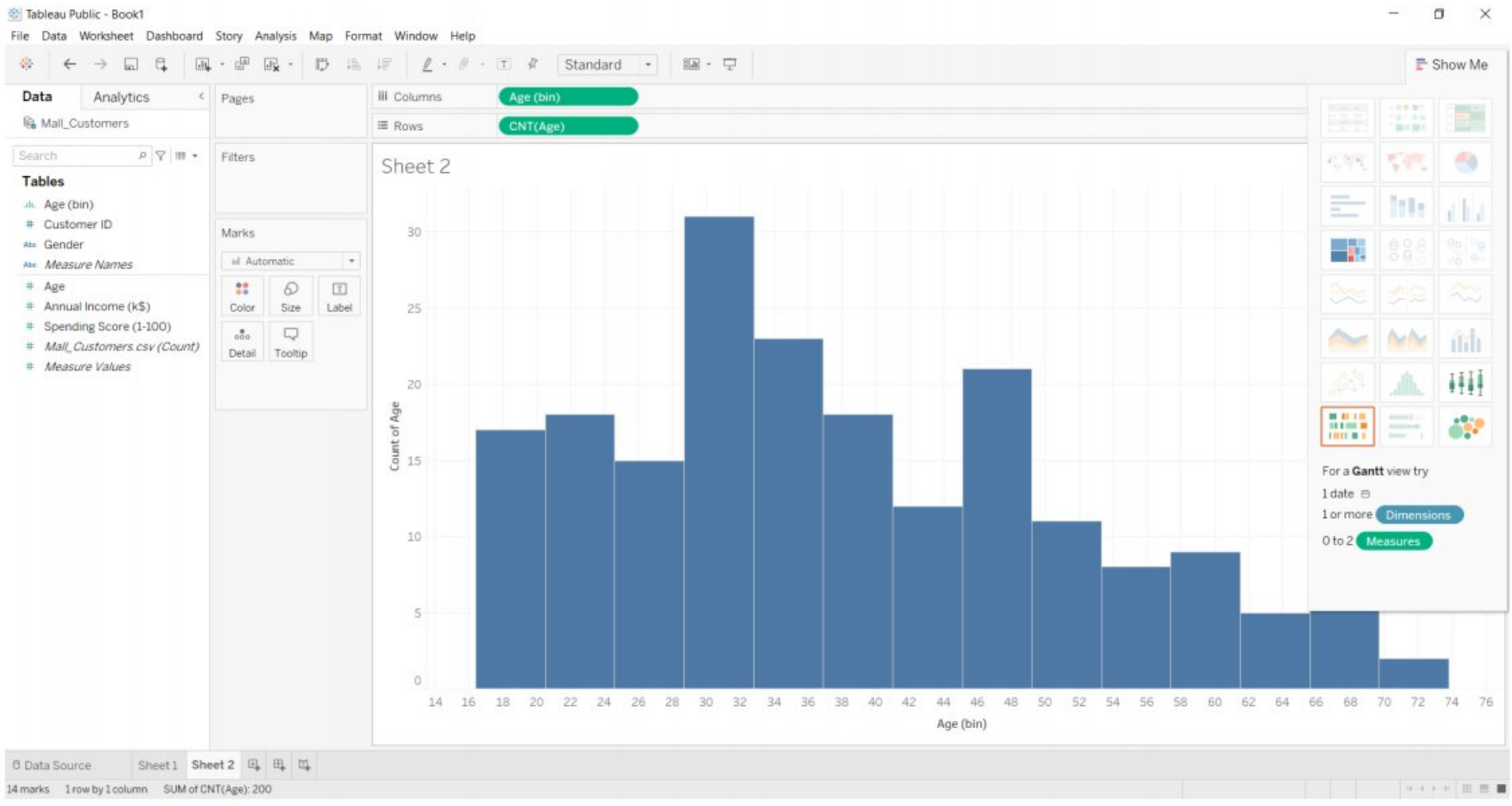


Fig.no.26

This above screenshot is of histogram graph which is applied on age attribute and it shows or visualizes data distributions.

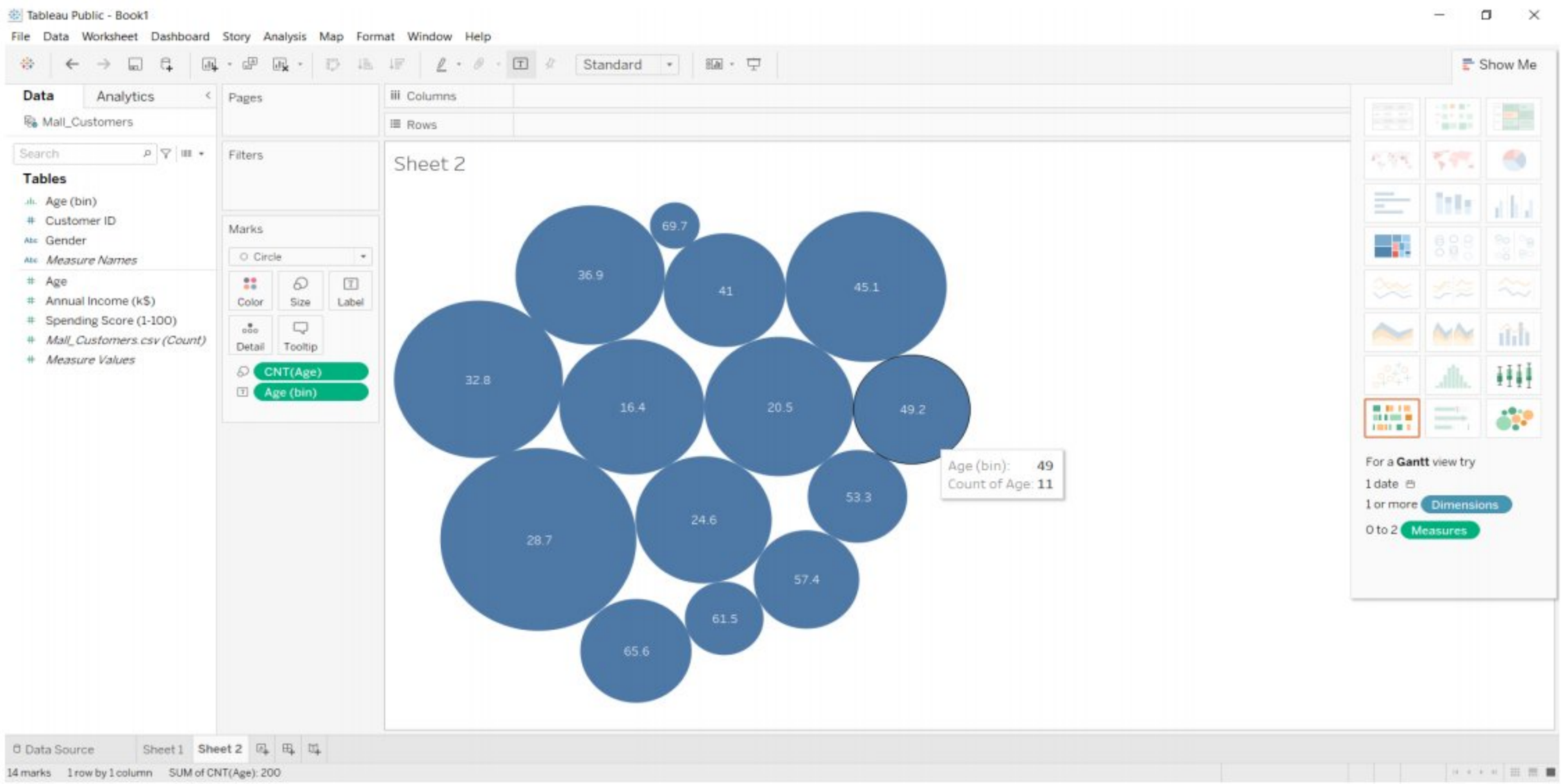


Fig.no.27

The above screenshot is of a packed bubble chart that is used on age attribute which is also used to represent data with occurrence.

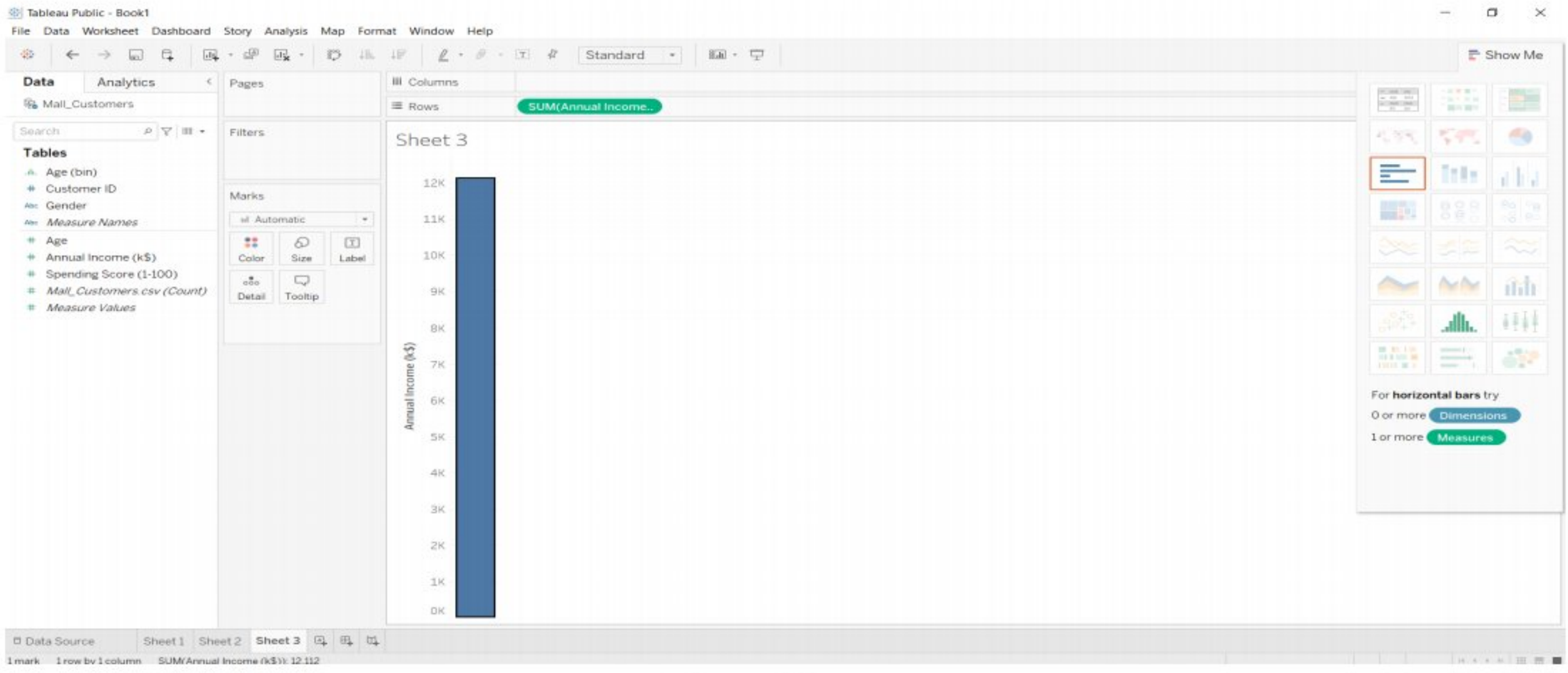


Fig.no.28

The above mentioned screenshot is of horizontal bars used on an annual income attribute to represent the data horizontally.

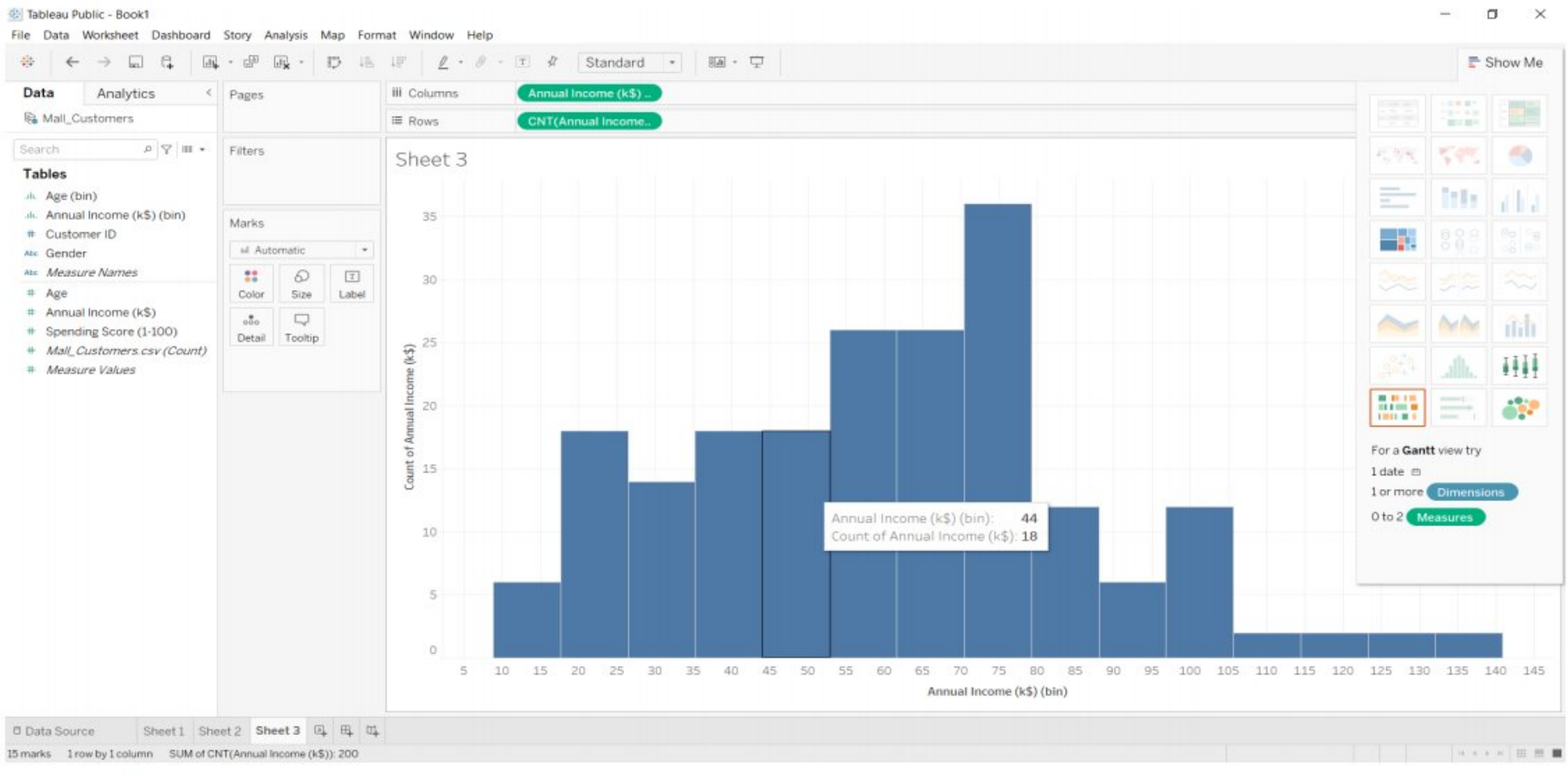


Fig.no.29

The above mentioned screenshot is of a histogram of annual income attribute whose function is to visualize data distribution.

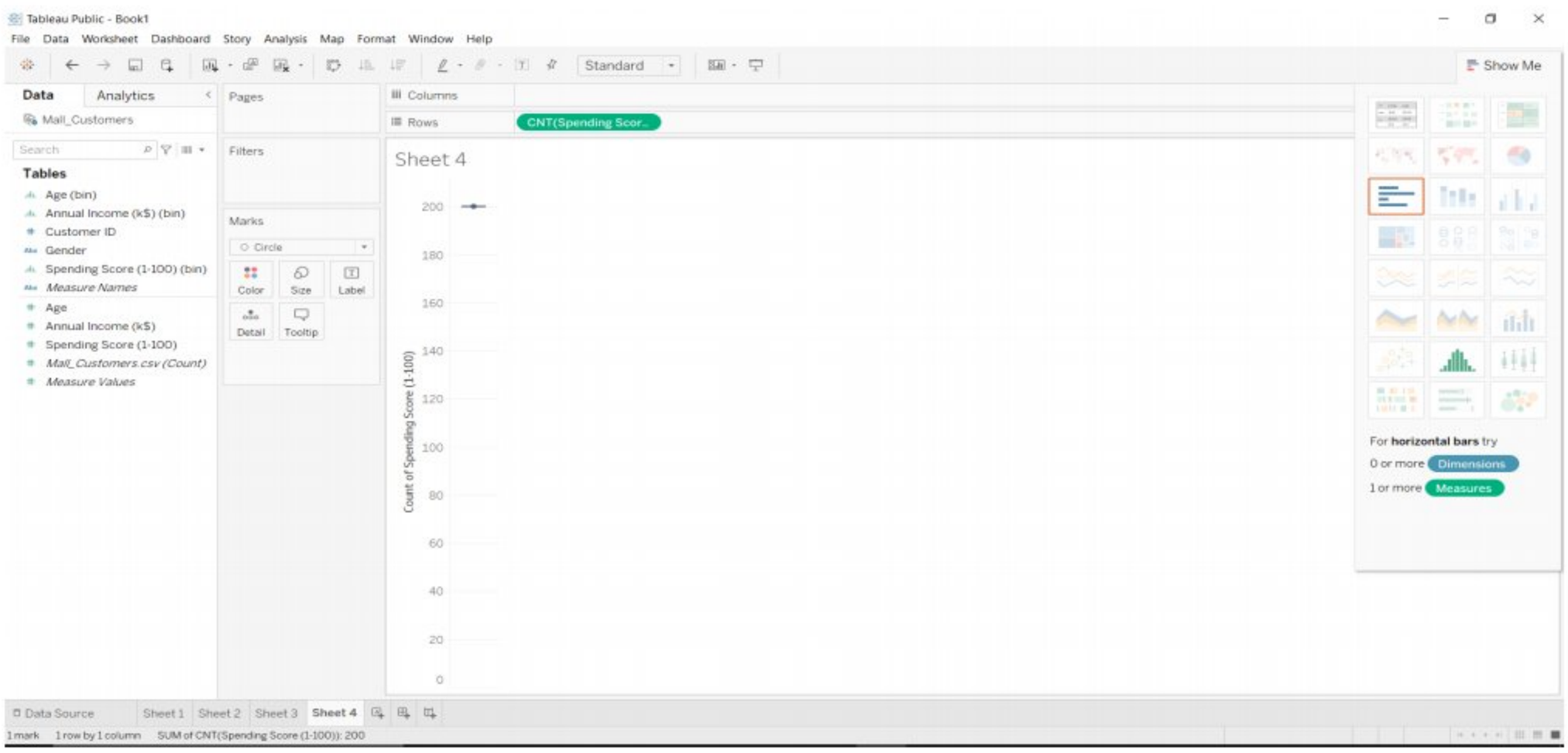


Fig.no.30

In the above mentioned screenshot , box-whisker plot data visualization technique applied on the spending score attribute of dataset.

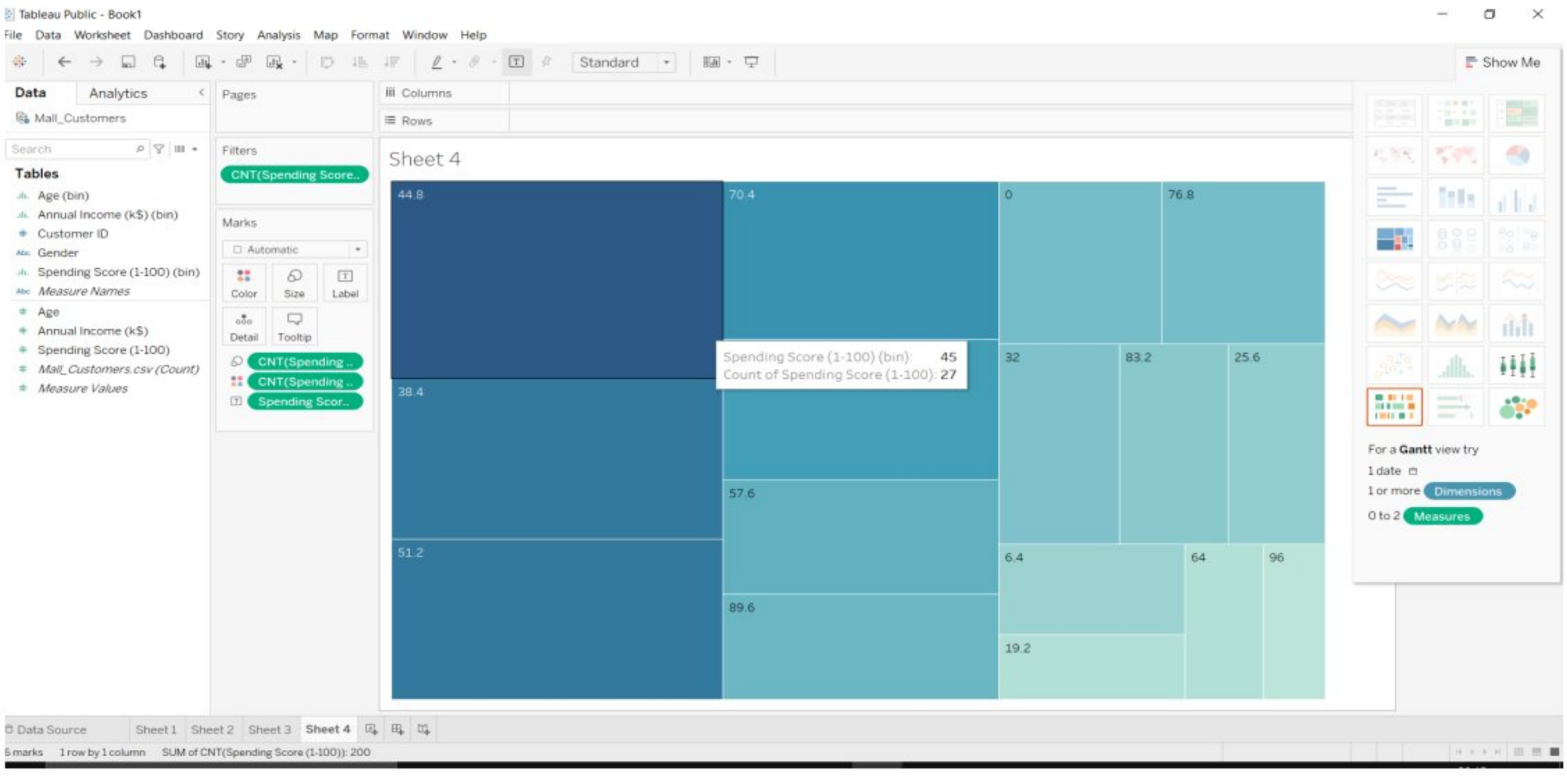


Fig.no.31

In the above mentioned screenshot, treemap is a data visualization technique applied on spending score attribute that is used to display hierarchical data using nested rectangles.The treemap chart is used for representing hierarchical data in a tree-like structure.

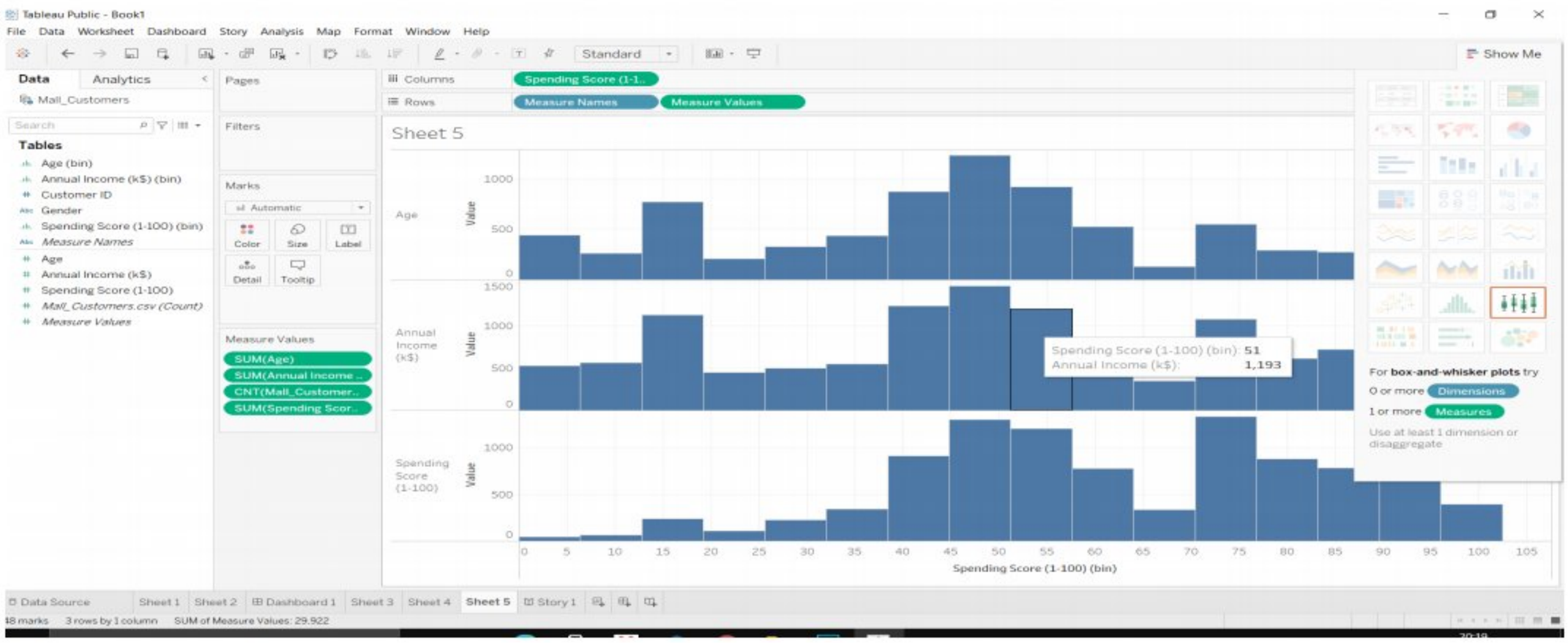


Fig.no.32

The above screenshot is histograms of attributes which are age, annual income and spending score which shows the whole histogram with multiple attributes.

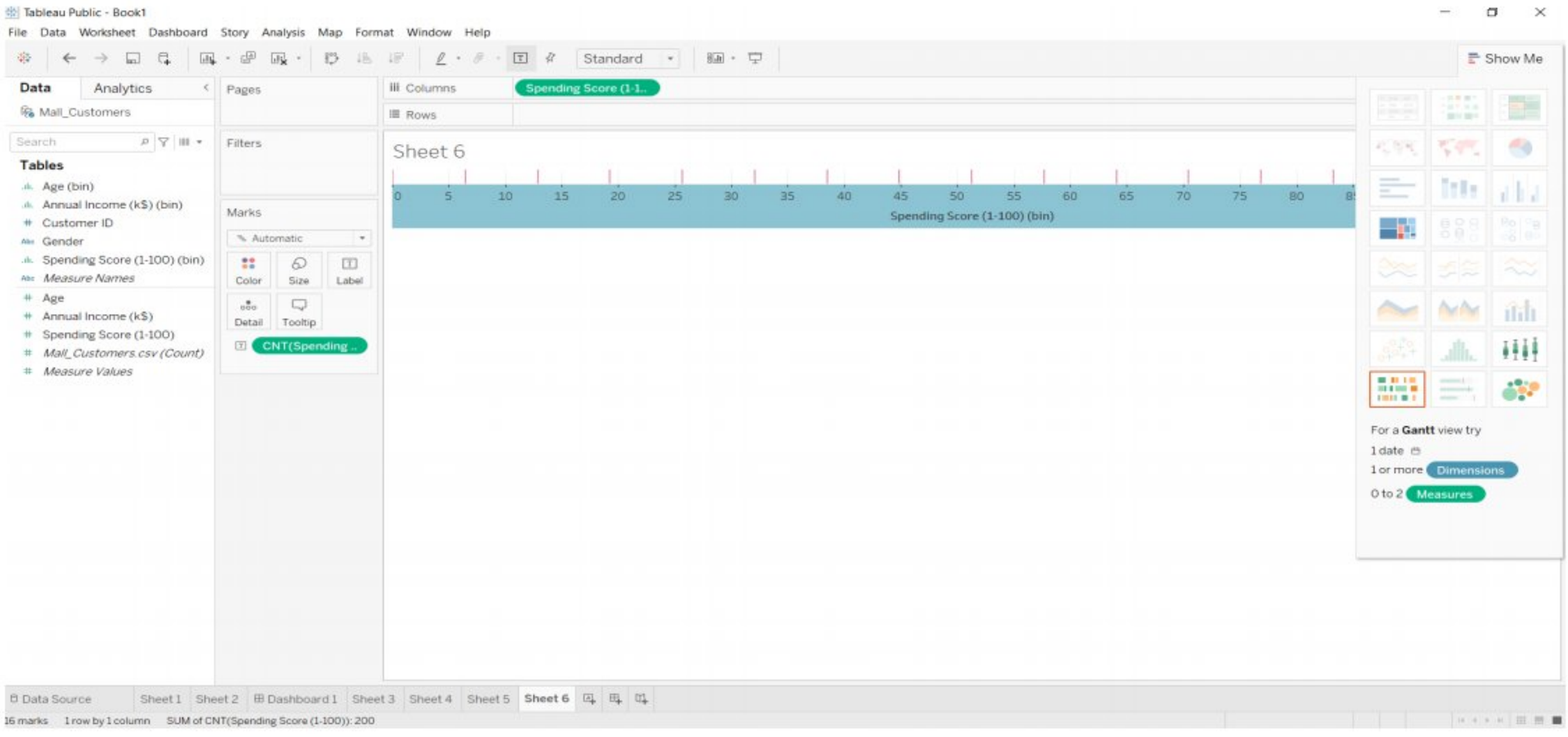


Fig.no.33

The above screenshot shows the gantt chart of spending score attribute which depicts the progress of the value of a spending score attribute.

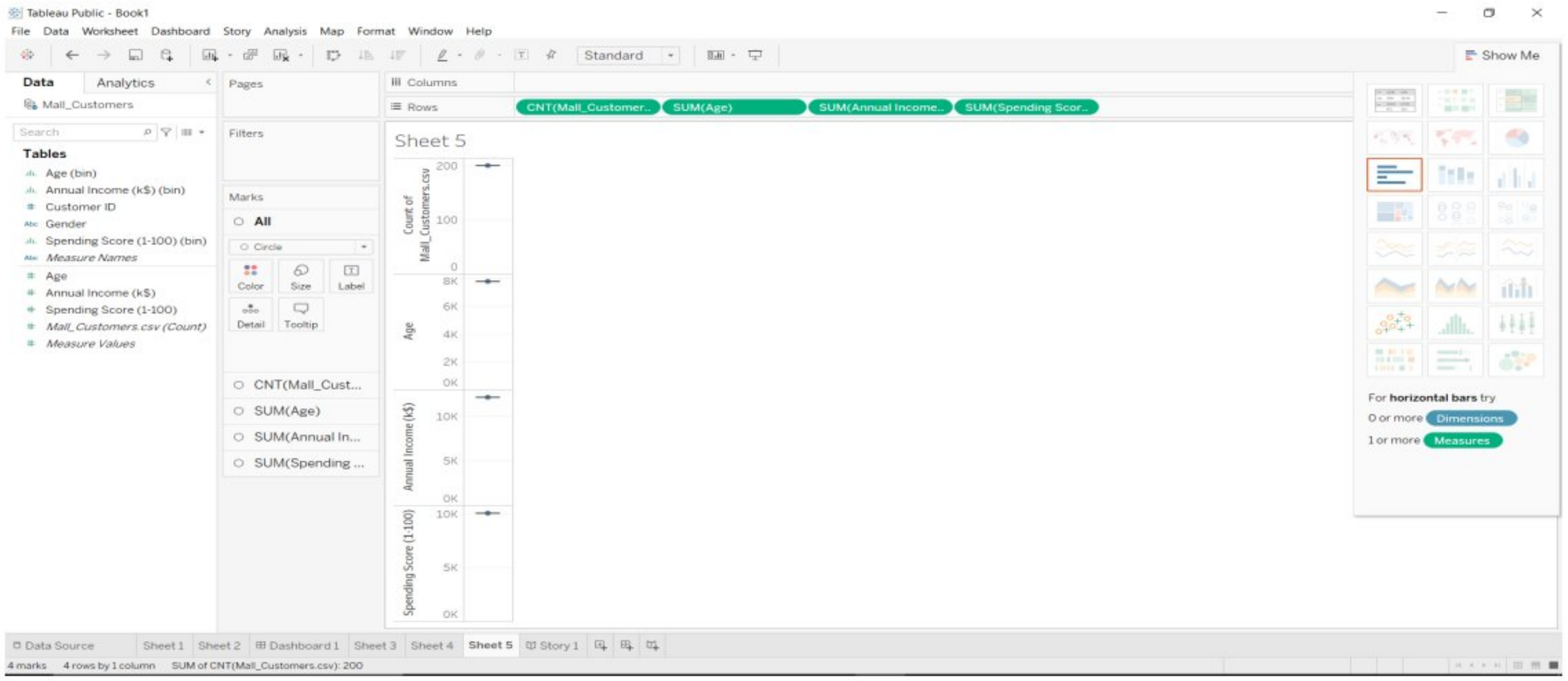


Fig.no.34

The above mentioned screenshot is box-whisker data visualization technique applied on multiple attributes which are age,annual income and spending score.

Stories and Dashboards in Tableau Software :

- In the Tableau, a story is a sequence of visualizations that work together to convey information. You can create stories to tell a data narrative, provide context, demonstrate how decisions relate to outcomes, or to simply make a compelling case.
- A story is a sheet, so the methods you use to create, name, and manage worksheets and dashboards also apply to stories. At the same time, a story is also a collection of sheets, arranged in a sequence. Each individual sheet in a story is called a story point.
- When you share a story for example, by publishing a workbook to Tableau Public, Tableau Server, or Tableau Online—users can interact with the story to reveal new findings or ask new questions of the data.
- The dashboard is a collection of several views, letting you compare a variety of data simultaneously. For example, if you have a set of views that you review every day, you can create a dashboard that displays all the views at once, rather than navigate to separate worksheets.
- Like worksheets, you access dashboards from tabs at the bottom of a workbook. Data in sheets and dashboards is connected; when you modify a sheet, any dashboards containing it change, and vice versa. Both sheets and dashboards update with the latest available data from the data source.

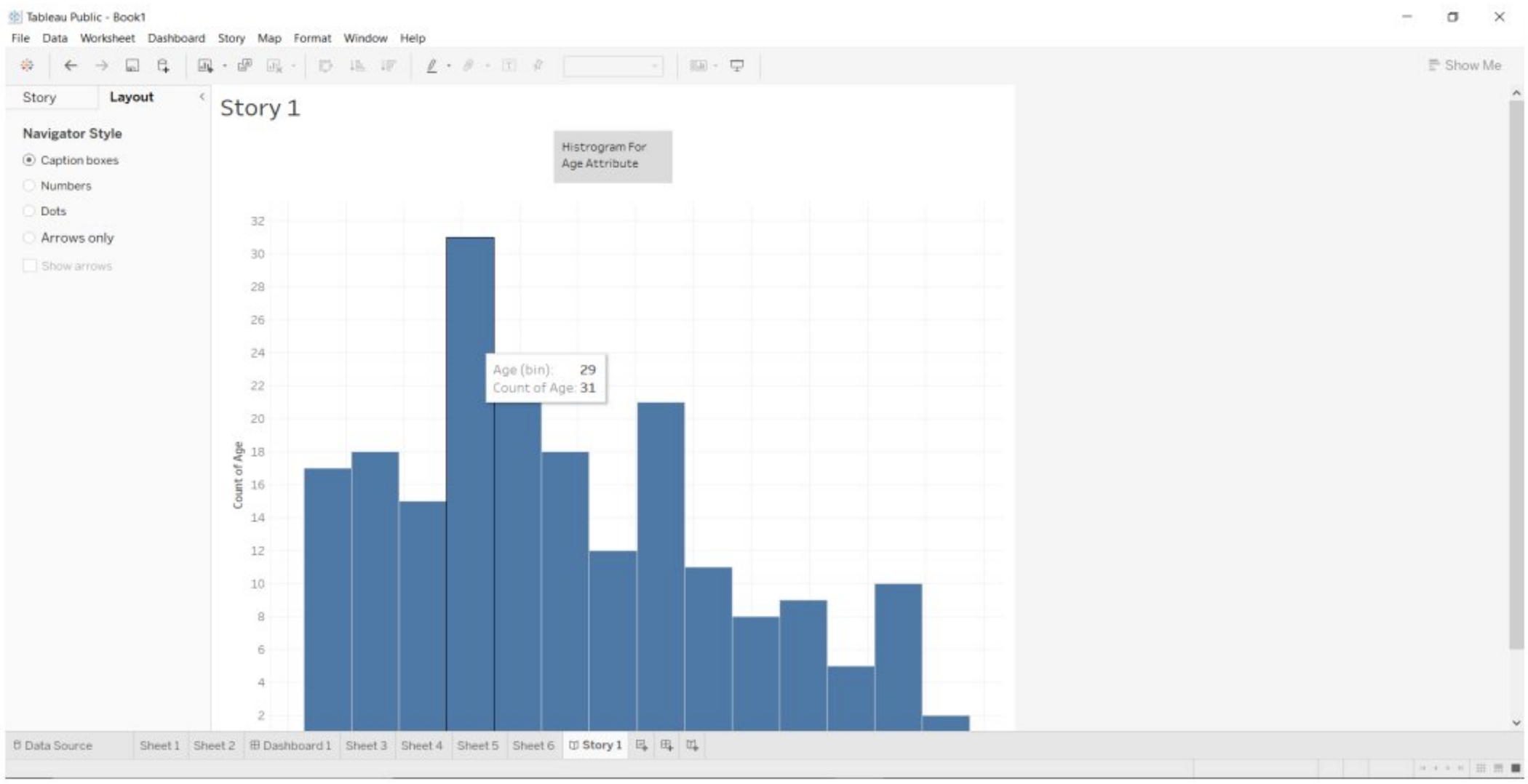


Fig.no.35

The above screenshot is a story of histogram of age attribute of the dataset.The caption to the story can be given to the stories.

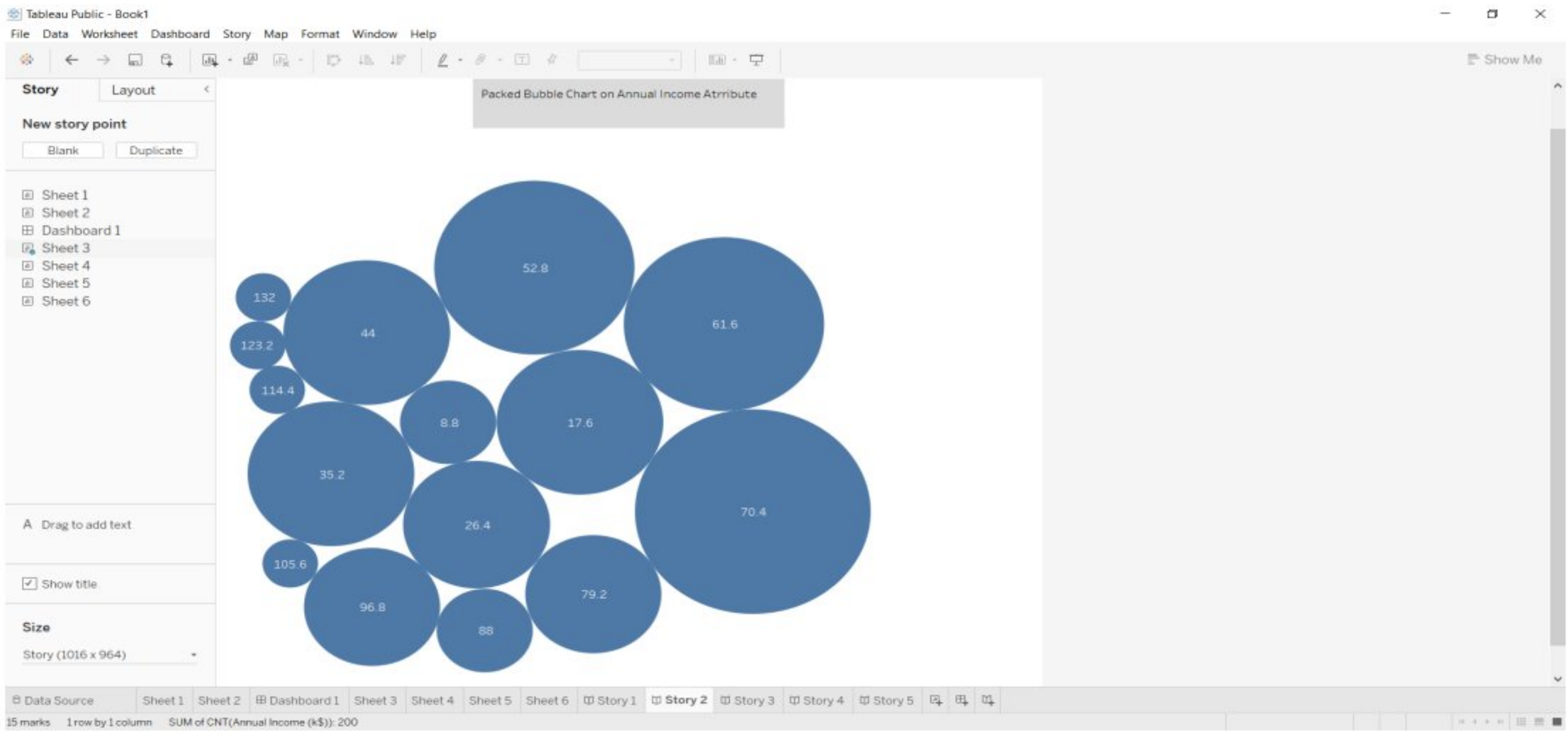


Fig.no.36

The above mentioned screenshot is a story of a packed bubble chart applied on the annual income attribute.

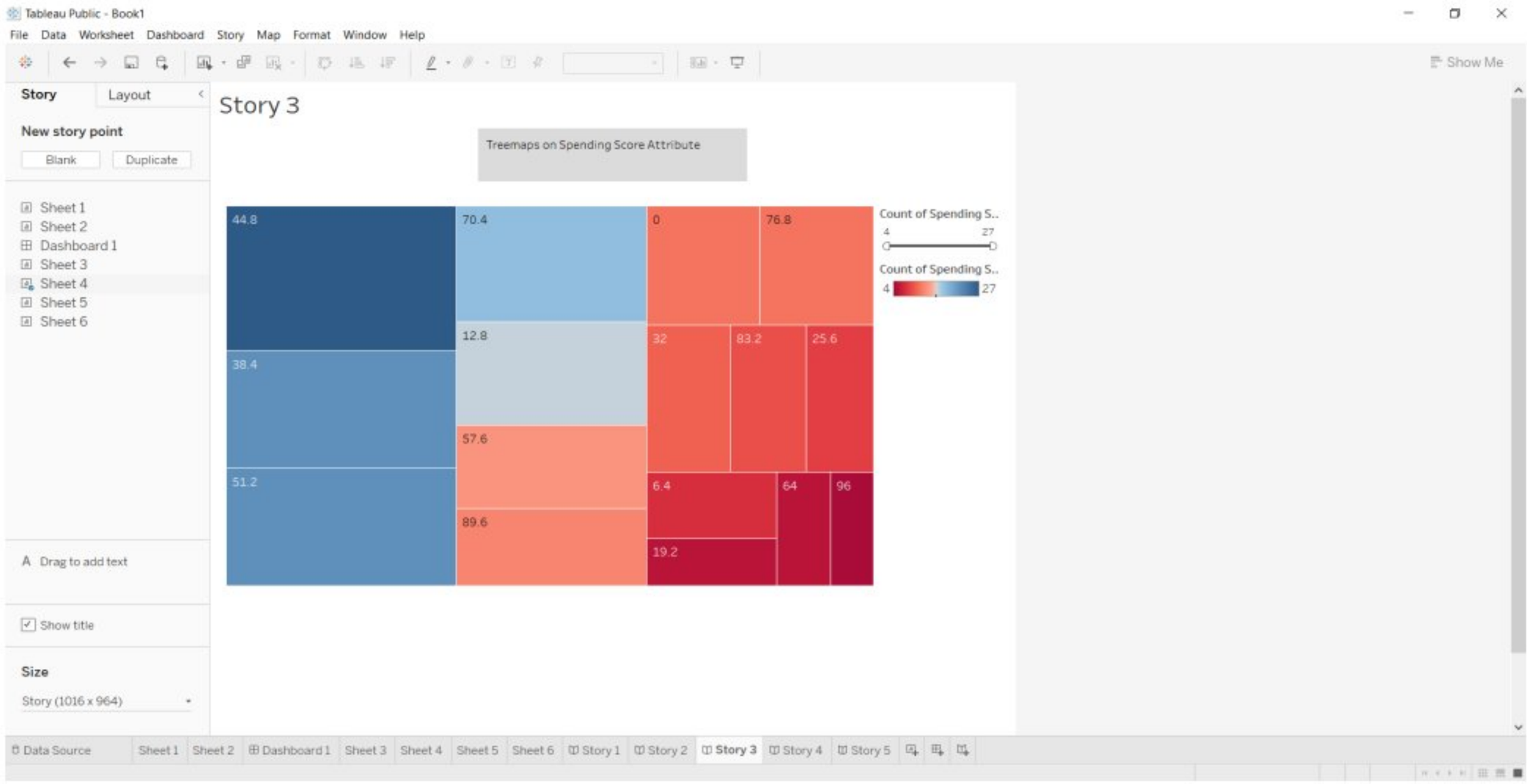


Fig.no.37

The above screenshot is a story of a treemap applied on the spending score attribute of the dataset.

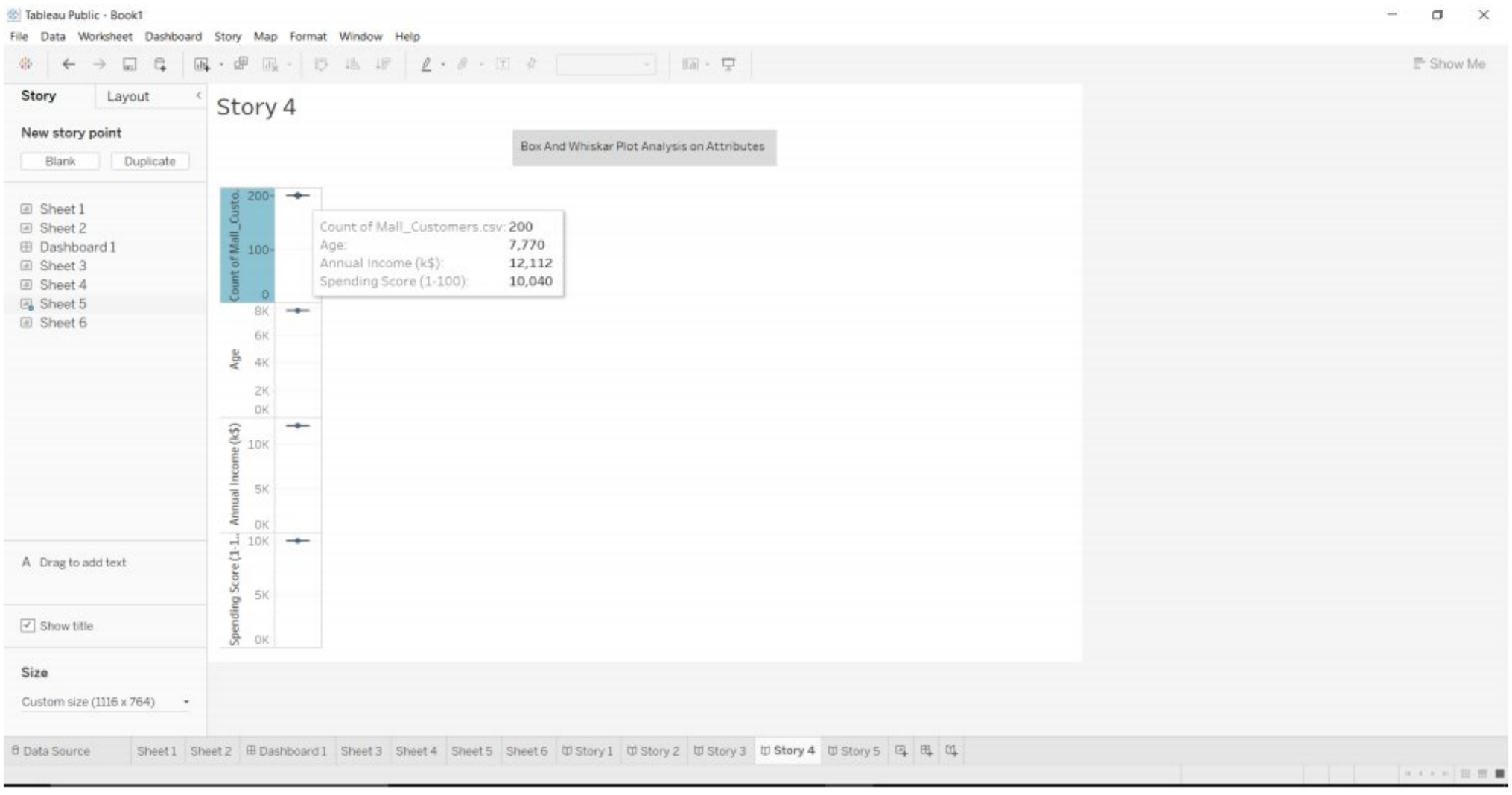


Fig.no.38

The above mentioned screenshot is of a story of box-whisker analysis applied on the multiple attributes.

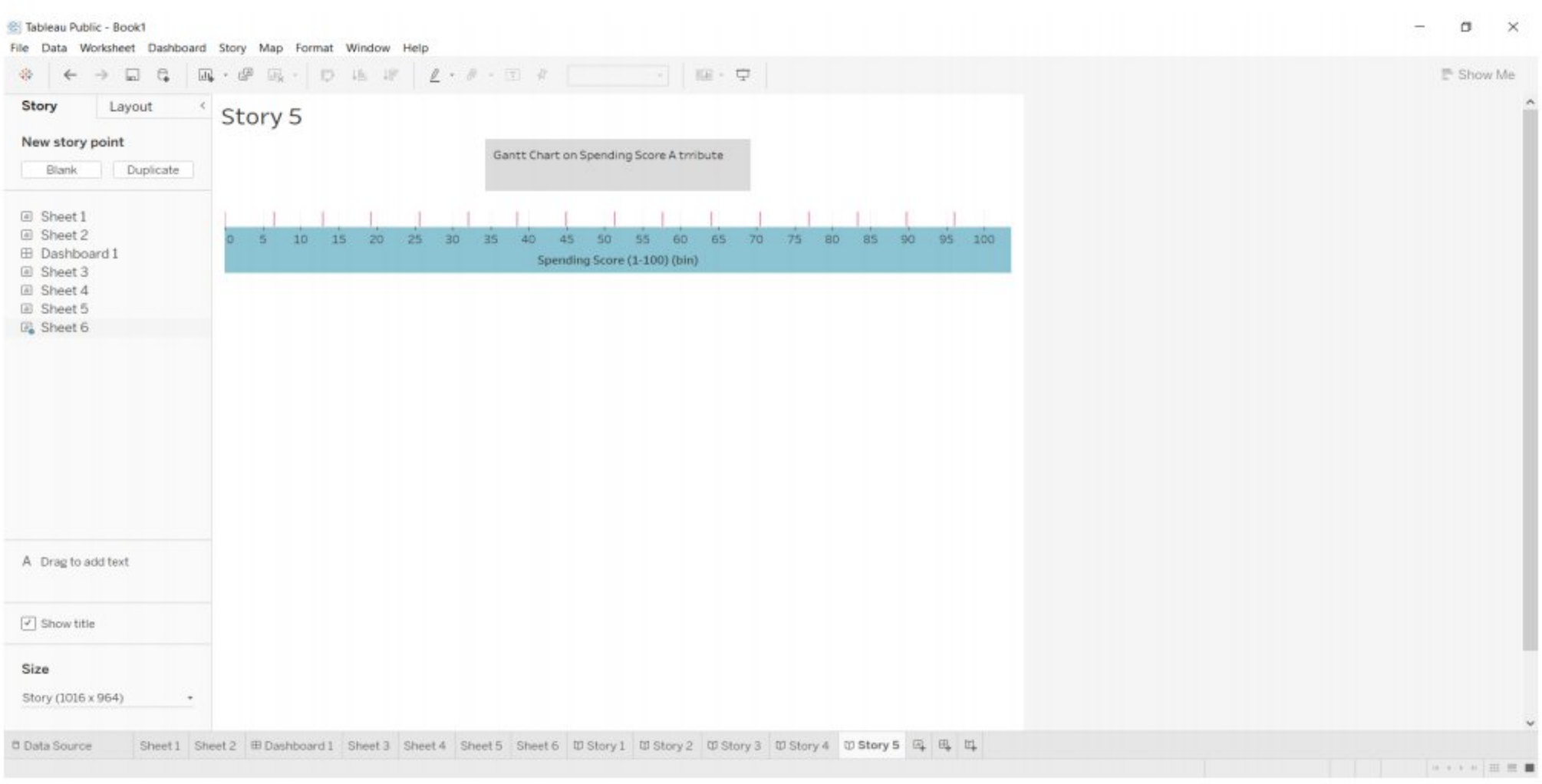


Fig.no.39

The above screenshot is of a story of a gantt chart applied on the spending score attribute of the dataset.

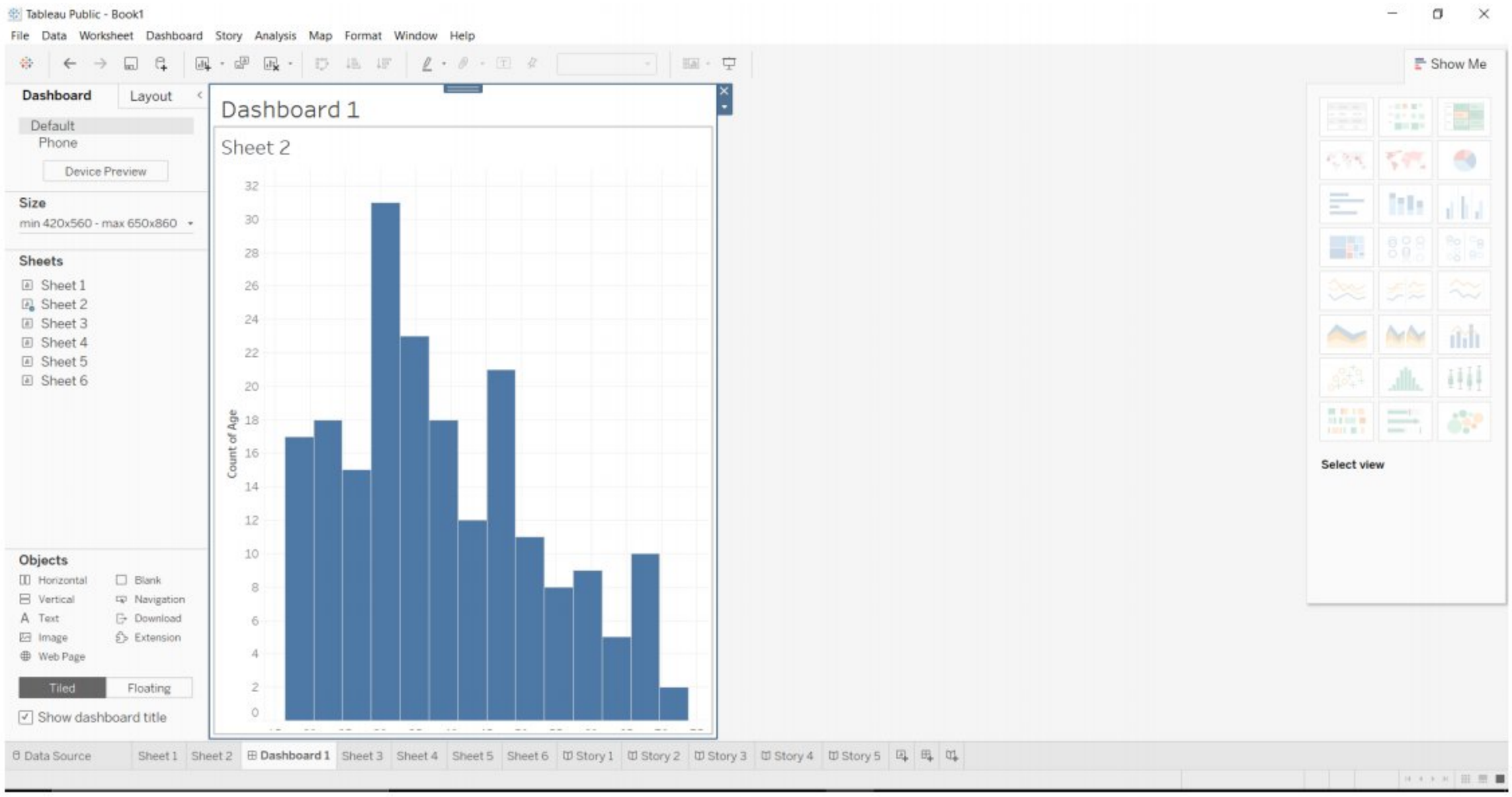


Fig.no.40

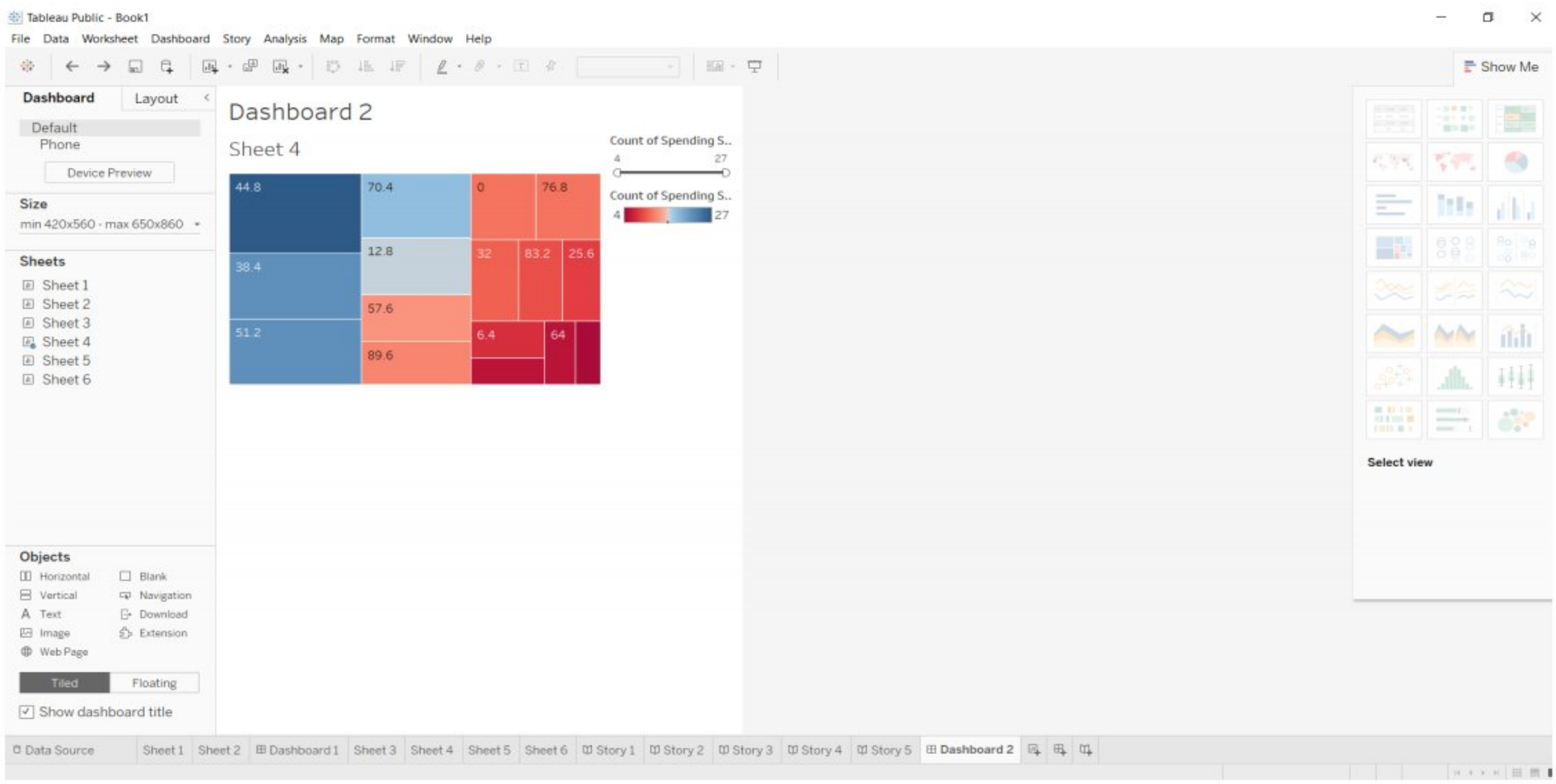


Fig.no.41

The above mentioned screenshots are of created dashboards of sheets which included the data visualization techniques applied on the attributes of the dataset.

8. Post-Experiment Exercise

a. Conclusion:

- Summary of Experiment
- Importance of Experiment
- Application of Experiment

9. Reference: Business Intelligence: Data Mining and Optimization for Decision Making by Carlo Vercellis, Wiley India Publications

Experiment NO : 10

Aim :- Develop a Mini Project for Business Intelligence

Conclusion :-

In this experiment, we studied and understood the case study of mall customer segmentation using data mining and business intelligence techniques. We download mall customer data.csv file for our case. We applied various data mining algorithms like clustering using k-means and classification via decision tree(J48) in the weka tool. For business intelligence we used tableau pubilc software for the data visulizations. We applied various graphs on our data sets' attributes for our chose casestudy we got desired outputs in the weaka tool as well as tableau public software.

References :-

1. https://www.tableau.com/products/cloud-bi
2. https://www.kaggle.com/datasets.