



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Report of Superstore Dataset

Meher Kiran Arasada
MSc Applied Computing
Data Mining Module, Assignment 3
20/05/2019
UWTSD, Swansea.



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

About Dataset

- The website URL of the sample dataset is:
<https://community.tableau.com/docs/DOC-1236>
- Created by Tableau online training team for their students
- This dataset consists of sales data of orders placed between 2014 and 2017 in various regions of the United States
- 21 attribute and 9994 records
- Row ID, OrderID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, Country, City, State, Postal Code, Region, Product ID, Category, Sub-Categories, Product Name, Sales, Quantity, Discount, Profit.



- 2 - Date Datatype, 6 - Numeric Datatype, 11- String Datatype, 2 – Alphanumeric
- Mean Value is 4997.5

Name: Row ID		Type: Numeric
Missing: 0 (0%)		Unique: 9994 (100%)
Distinct: 9994		
Statistic	Value	
Minimum	1	
Maximum	9994	
Mean	4997.5	
StdDev	2885.164	

- The dimensions available in this dataset are suitable for pictorial representations, which can make for easier understanding.



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Preprocess

- Starting point for knowledge exploration
- No attributes were empty
- In this process none of the attribute were deleted

Current relation

Relation: superstoredata

Attributes: 21

Instances: 9994

Sum of weights: 9994



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Preprocess

- ARFF header
- Assigned datatypes to the respective attributes.
- Able to visualise the data with this header

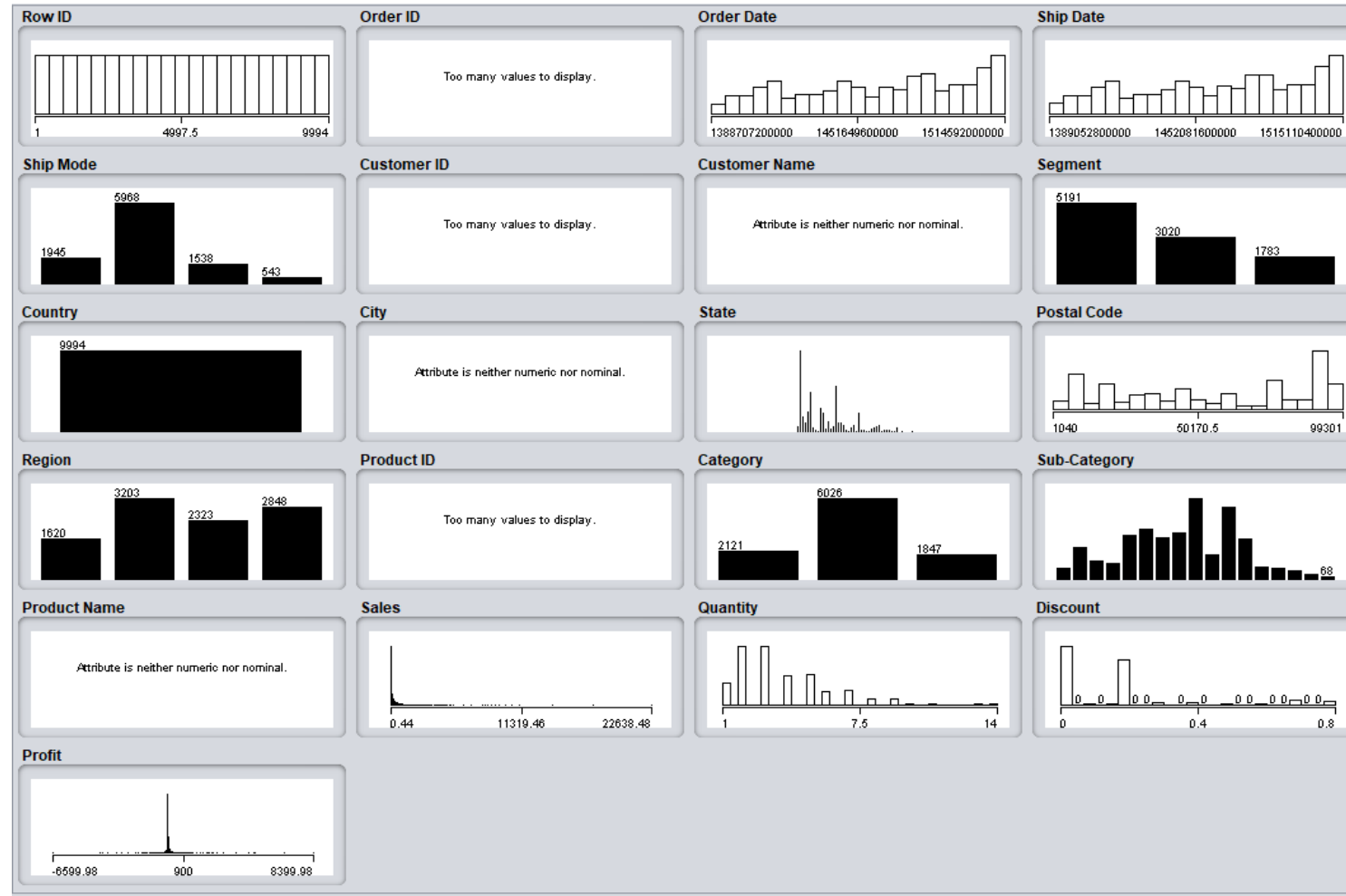
```
@relation exp
```

```
@attribute 'Row ID' numeric  
@attribute 'Order ID' {CA-2016-152156,CA-2016-138688,US-2015-108966,CA-2014-115812,  
@attribute 'Order Date' DATE "yyyy-MM-dd"  
@attribute 'Ship Date' DATE "yyyy-MM-dd"  
@attribute 'Ship Mode' {'Second Class','Standard Class','First Class','Same Day'}  
@attribute 'Customer ID' {CG-12520,DV-13045,SO-20335,BH-11710,AA-10480,IM-15070,HP-  
@attribute 'Customer Name' string  
@attribute Segment {'Consumer','Corporate','Home Office'}  
@attribute Country {'United States'}  
@attribute City string  
@attribute State {Kentucky,California,Florida,'North Carolina',Washington,Texas,Wis  
@attribute 'Postal Code' numeric  
@attribute Region string  
@attribute 'Product ID' {FUR-BO-10001798,FUR-CH-10000454,OFF-LA-10000240,FUR-TA-100  
@attribute Category {Furniture,'Office Supplies',Technology}  
@attribute Sub-Category string  
@attribute 'Product Name' string  
@attribute Sales numeric  
@attribute Quantity numeric  
@attribute Discount numeric  
@attribute Profit numeric
```

```
@data
```



Data Visualisation





Data Visualisation

- Row ID: It shows an equal vertical bar, which means data is equally distributed; it shows the mean value.
- Order date: Date-type has been assigned and pattern was generated.
- Ship Mode: We can see four vertical bars, one of which was steep (Standard Class), second best was second class, next was first class and last was Same day.
- Segment: There are three segments, pattern seems to be like steps. First come Consumers, next Corporate and Home Office.
- Country: One single bar was present because the data related to one country.
- State: It provides the information about the sales made in each state.
- Postal code: Shows small bars of sales occurred at those postal codes.
- Region: There are four regions and four vertical bars can be seen; this provides information on which regions have had more sales (West, East, Central and South).



Data Visualisation

- Category: There are three different categories. It shows which category has made more sales. Office supplies, Furniture and Technology
- Sub-category: These bar shows the sales of each sub-category. Best sold item was Binder (1523 units).
- Sales: One of the key information for this dataset, there is no specific pattern identified which can be used for marketing or forecast.
- Quantity: Fewer quantity of items per sales are more than the high number of items per sale.
- Discount: Only few items have discounts on the products, so we can not find any specific pattern.
- Profits: It show the company's ability for sales and periods of high demand.



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

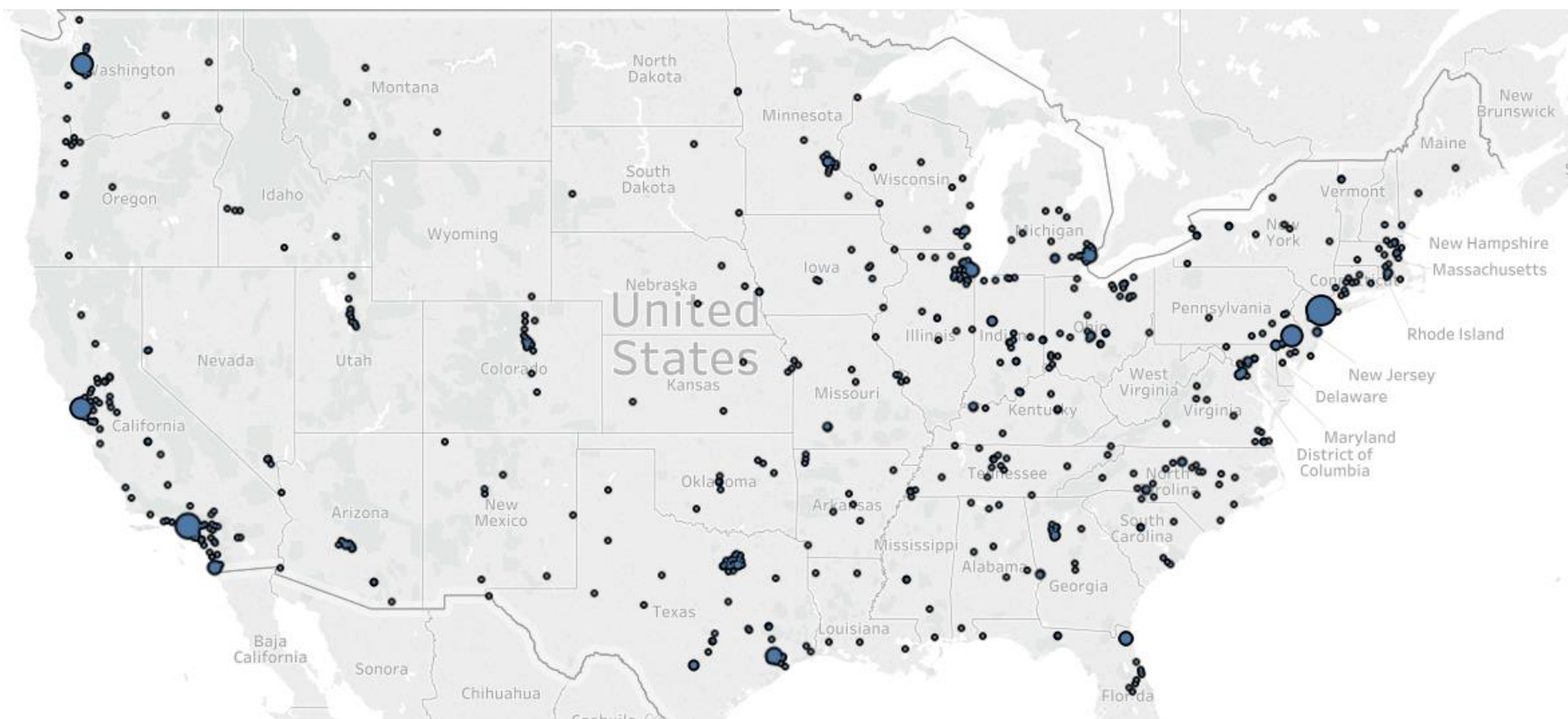
Questions

1. How does Tableau generate latitude and longitude?
2. How to measure classification errors?
3. How do you analyse the trend of sales?
4. What combination of products could be sold as a special offer?
5. What will the organization's sales look like in next one year?



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Geographical Data Visualisation





Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

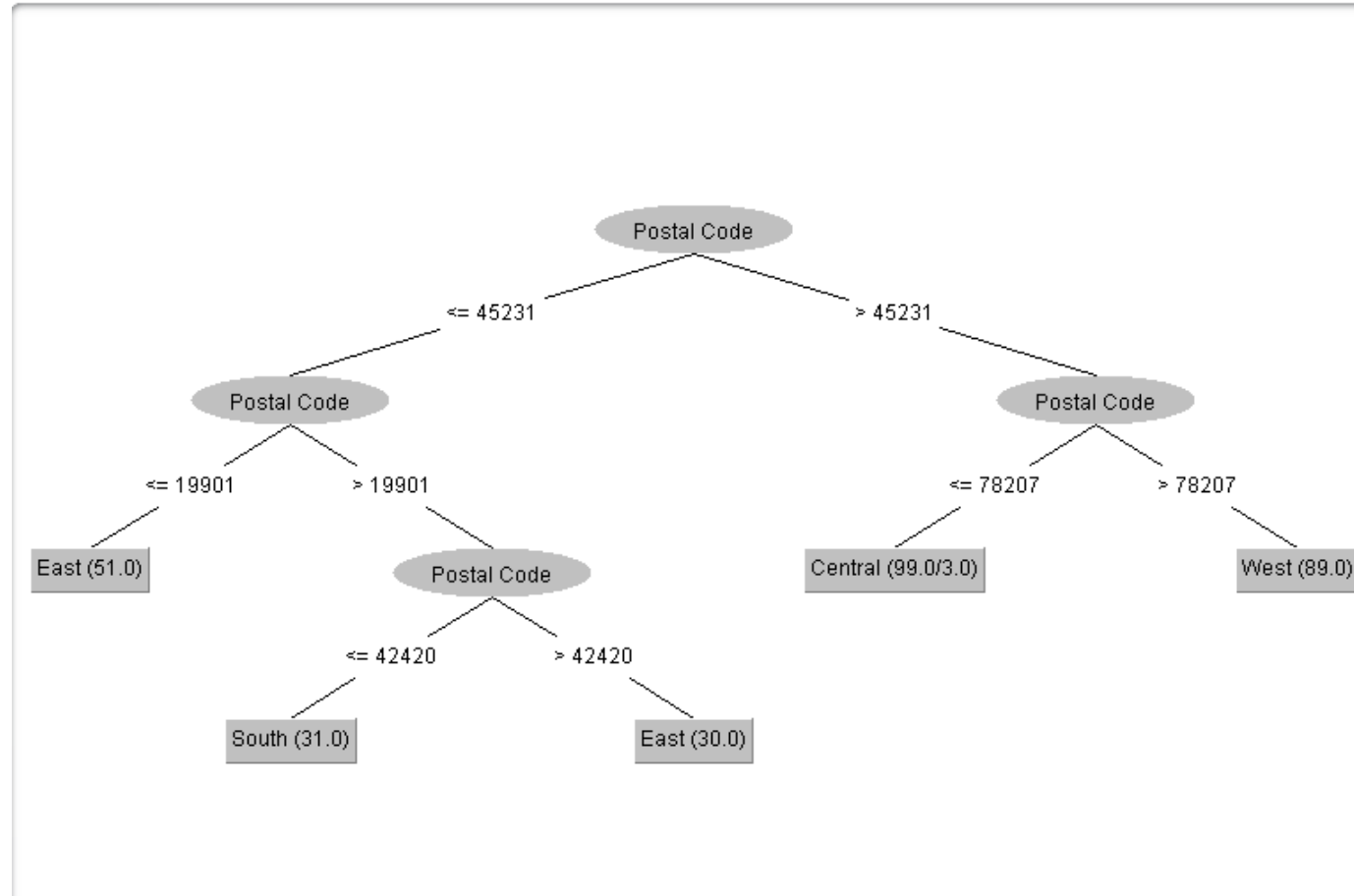
Data Visualisation

- Geographical data visualization
- Most of the blue spots identified towards West and Central part of US
- Big spots indicates more sales in those places.
- New York has the highest sales of \$256,368



Classifier

Tree View





Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Classifier

- J48 tree classifier
- minNumObj: 25
- Number records 300
- Errors can be seen in Confusion Matrix
- 3 errors in central
- Tree helps to identify the errors easily

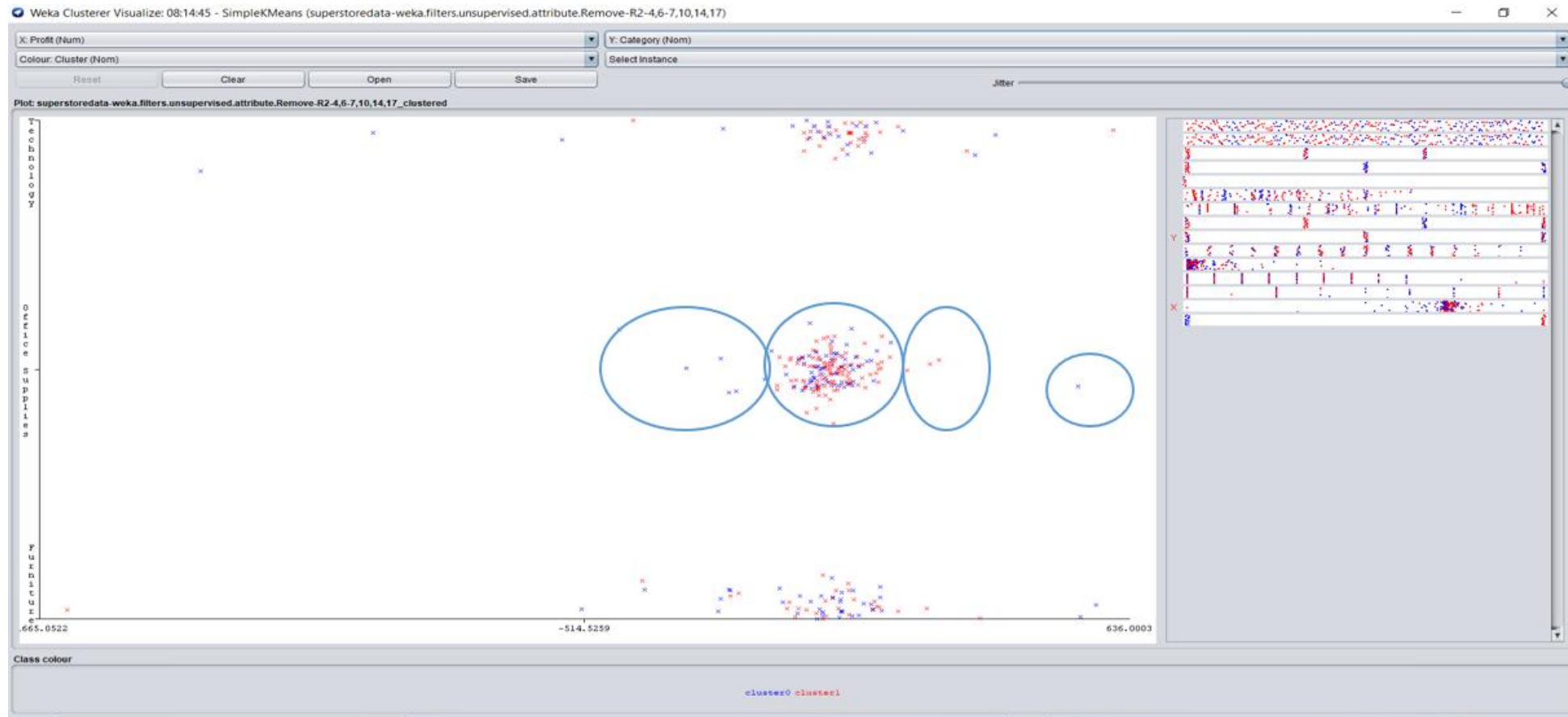
=== Confusion Matrix ===

a	b	c	d	<-- classified as
31	0	3	0	a = South
0	89	0	0	b = West
0	0	96	0	c = Central
0	0	0	81	d = East



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Clustering





Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Clustering

- SimpleKMeans type of the cluster
- Based on the results we can say, which category has more Profits
- Office supplies made more profits(\$496) in the year 2014
- Product is Honeywell Portable HEPA Air Cleaner





Association

Weka Explorer

Preprocess Classify Cluster **Associate** Select attributes Visualize Forecast

Choose Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Start Stop

Result list (right-click...)

- 09:19:19 - Apriori
- 09:21:30 - Apriori
- 09:27:13 - Apriori
- 09:27:25 - Apriori
- 09:35:51 - Apriori

Associator output

```
=== Run information ===
Scheme:      weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:    superstoredata-weka.filters.unsupervised.attribute.Remove-R2-4,6-7,10,14,17-weka.filters.unsupervised.attribute.Remove-R5-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.un
Instances:    300
Attributes:   5
  Ship Mode
  Segment
  Region
  Category
  Sub-Category

=== Associator model (full training set) ===

Apriori
=====

Minimum support: 0.1 (30 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 16
Size of set of large itemsets L(2): 25
Size of set of large itemsets L(3): 7

Best rules found:

1. Sub-Category=Binders 37 ==> Category=Office Supplies 37 <conf:(1)> lift:(1.69) lev:(0.05) [15] conv:(15.05)
2. Sub-Category=Paper 37 ==> Category=Office Supplies 37 <conf:(1)> lift:(1.69) lev:(0.05) [15] conv:(15.05)
3. Sub-Category=Storage 33 ==> Category=Office Supplies 33 <conf:(1)> lift:(1.69) lev:(0.04) [13] conv:(13.42)
```

Status

OK Log x 0



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

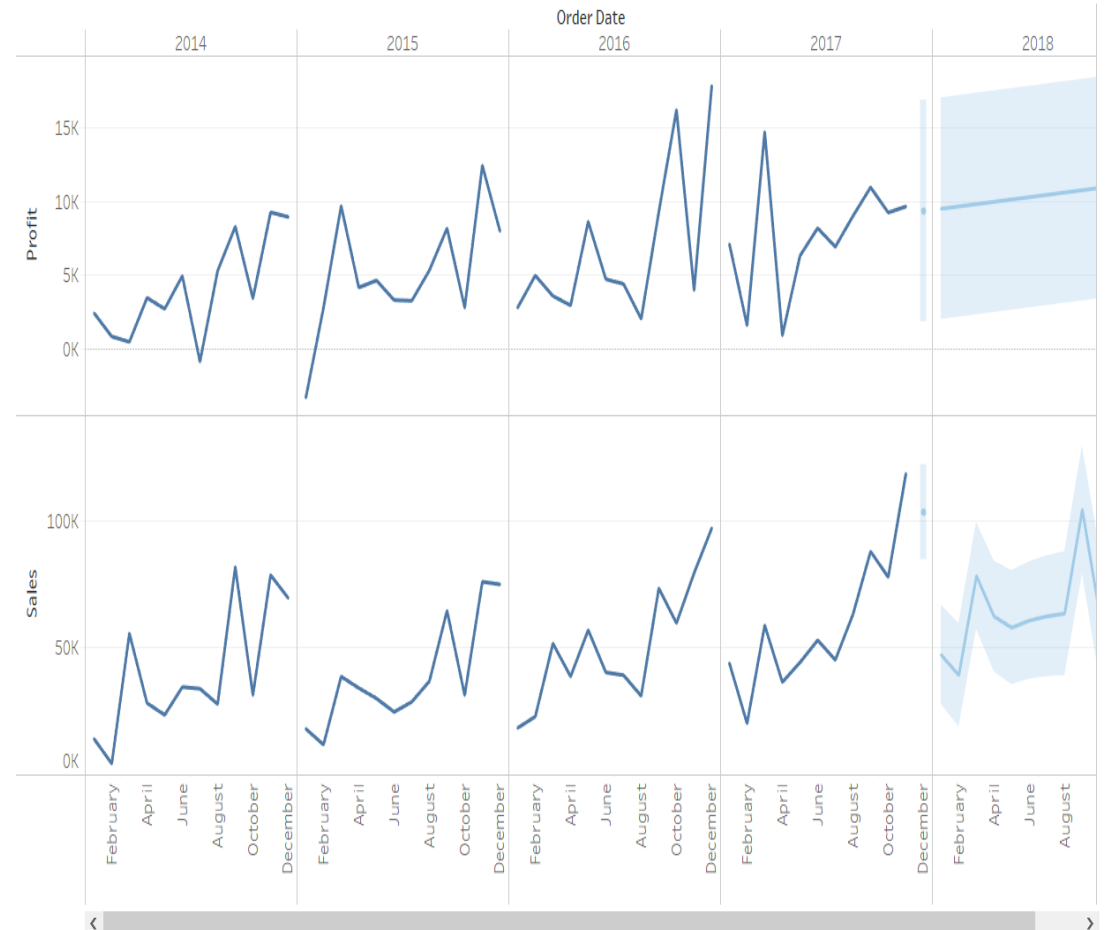
Association

- Apriori is the associator
- {Storage, Paper} → Art (Blind, Pencils), Binder, Envelopes
- {Table, Chair} → Study Case, Lamp
- Combination and deals will help in increase in sales



Forecast

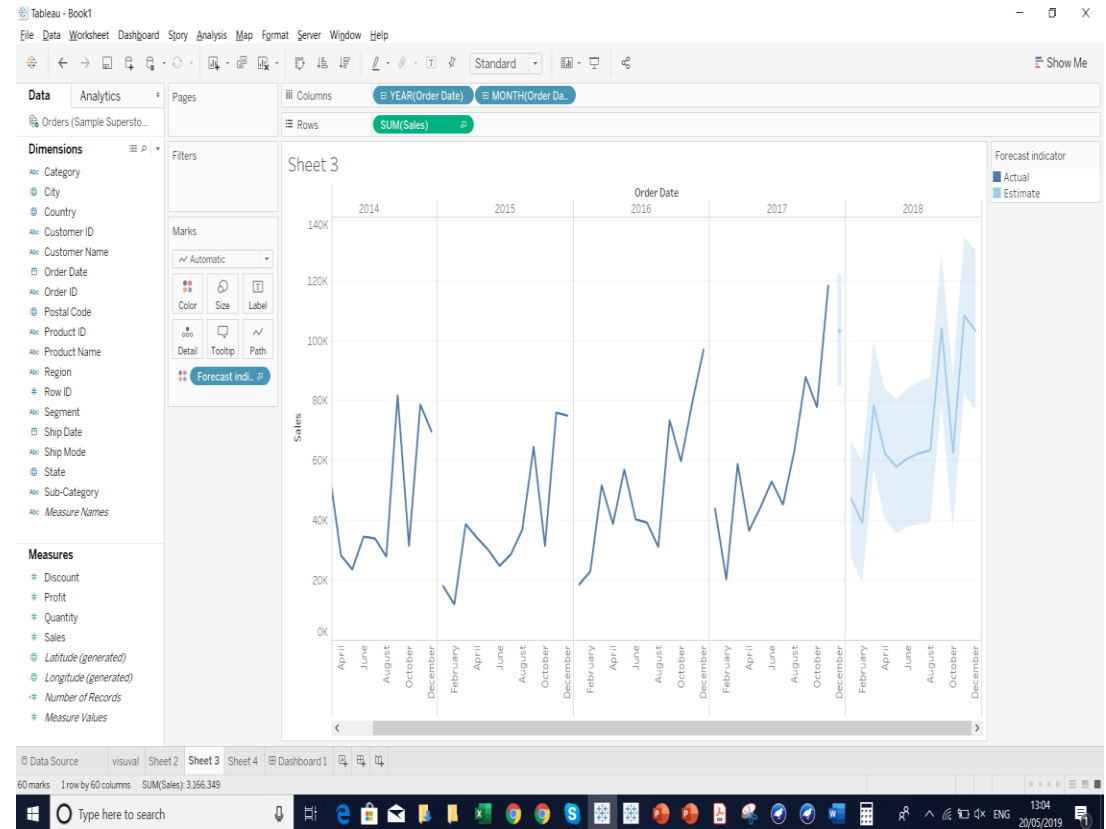
- Forecasted sales and profit based on order date. Used tableau.
- Forecast of profit shows a gradual increase without ups and downs





Forecast

- Sales wise forecast, slight drop down in the month of February and steep in the month March and from April average sales till August and steep in the September and dropdown of sale in October and increase of sales in November.
- Best sales month can be March, September and November





Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Reflection

- Most difficult thing was finding dataset and uploading the data to Weka.
- Initially no pattern appeared for Order date and Ship date, as date datatype is assigned patterns appeared.
- Acquired knowledge on data mining such as how to acquire the data, play with the data and generate the results from it.



Conclusion

- New York has the highest sales of \$256,368
- Best product Honeywell Portable HEPA Air Cleaner in the 2014, because of more profit.
- Placing some below combination, might increase in sales
 - {Storage, Paper} → Art (Blind, Pencils), Binder, Envelopes
 - {Table, Chair} → Study Case, Lamp
- Sales in March, September and November months look promising based on the forecast output.



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Thank You All



Prifysgol Cymru

Y Drindod Dewi Sant

University of Wales

Trinity Saint David

POSTER



REPORT OF SUPERSTORE DATASET – DATA MINING TECHNIQUES



Meher Kiran Arasada

1805462, UWTSD, Swansea.
1805462@student.uwtsd.ac.uk



Abstract

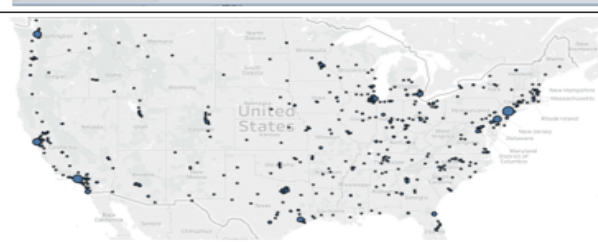
A report has been generated based on the sample dataset of a superstore. The dimensions available in this dataset are suitable for pictorial representations, which can make for easier understanding. In addition, the size of this superstore dataset is enough for performing data mining techniques. Data analysis helps the organization to understand their customers' buying trends, which will help the organization to market more effectively and ultimately increase sales.

Dataset

The information present in the superstore data was in relation to the sales done by an organization. The data set is created by Tableau online training team for their students' use. This dataset consists of sales data of orders placed between 2014 and 2017 in various regions of the United States. It has key measures such as Quantity Sales, Discount and Profit. It also has key dimensions such as Category, Sub-Category, Date, Region, City and State. It contains both numeric and descriptive data in this sample dataset. Overall there are 9994 records present in the dataset with 21 attributes. It is a single file, there are no joint files. It also has key dimensions such as Category, Sub-Category, Date, Region, City and State. It contains both numeric and descriptive data in this sample dataset. The dimensions available in this dataset are suitable for pictorial representations, which can make for easier understanding.

Data Pre-process

The preprocess is the starting point for the knowledge examination. Here the data will be loaded. There were no empty fields to handle and no data was deleted or added. The Excel file was converted into CSV format and loaded into Weka and saved the dataset in ARFF format. Assigned datatypes to the respective attributes. Initially, date data pattern was not generated, but after assigning the datatype to the date format pattern was generated. Proper ARFF header was obtained in this process. It shows the relation of the dataset, number of attributes, and number of instances.



Data Analysis

Analysed the data by using some data mining techniques like Classify, Cluster, Associate and Forecast. Classification errors experience the sales in the different regions. Algorithm applied to perform clustering technique is SimpleKMean. Cluster provides the information about the most profitable product in the data. Apriori algorithm was introduced to identify the item sets frequently and relevant association rules. This algorithm result makes the customer to identify the item combinations with ease and improve the sales of any marketing companies. Time series analysis was implemented to predict the profit, based on the sales. Forecast of profit and sales based on order date was performed by using Tableau software.

Results

- New York has the highest sales of \$256,368
- Best product "Honeywell Portable HEPA Air Cleaner" in the 2014, because of more profit.
- Sales in March, September and November months look promising based on the forecast output.

Source of Dataset

The website URL of the sample dataset is:
<https://community.tableau.com/docs/DOC-1236>