**Clustering in Tableau**

**Use-case:**

You are a Data Scientist working in a laundry pickup services start-up *WeWashUSleep.* This is relatively small company, and they cannot compete with the big players in major cities. The company’s strategy is to build a vast network in the smaller cities.

*WeWashUSleep* already had a strong presence in 140 locations and recently opened stores in 10 new cities. Additionally, the company has two separate sales regions.

**Task 1:**

Identify which of the following sales regions is performing better (i.e., outperforms the other in 2 of the following 3 metrics):

* Average revenue per city
* Average marketing spends per city.
* Average ROMI (Return on Marketing Investment) per city *(Revenue / Marketing Spend)*

**Task 2**:

Identify which of the 10 new locations have the best potential for the company to invest more funds into marketing.

**Sheet 1 – Custom Territories Via Group**

* The state entity is added into the sheet which will show the map of United States. It is important to sales region as two different regions are present and hence sales column is added into the color.
* Add Revenue to the details tab and select Aevrage value which shows the average revenue per city.
* Select all the states of one particular region, right click -> Group -> All Dimensions. A group is created and remove the state feature and two custom territories which indicate two regions with average revenue where region 1 is performing better than region 2.
* Add marketing spend (Average) and Sales into the labels.
* Creating calcualted field ROMI = Revenue / Marketing Spend and add it to labels.
* Edit all the labels and 2 different regions with statistics can be observed.

Map

Description automatically generated

**Sheet 2 – Custom Territories Via Geographic Roles**

* Every state is associated with sales region and tableau can recognise it.
* Right click on Sales Region -> Geographic Role -> Create from -> State. Sales region is now available into State, City.
* Drag the sales region into the working area and same results can be achieved as done above. This method is more proper, solid process, and versatile.

**Sheet 3 – Highlighting**

* Add state into working area, click on the plus (+) sign, and both cities and states are available, and no un-known city is present.
* Add revenue to size and colour. Drag new expansion feature and add it to the detail. Click on drop down and select the highlighter by which new and old expansion can be checked. It is not intrusive as it is not using any colour, shape, or size.

**Chart

Description automatically generated**

**Sheet 4 – Clustering**

* Drag Revenue on Y-axis and Marketing on X-axis and add store-ID into the Detail. Two distinct groups have been created.
* The graph shows that when the company spends more on marketing, it generated more revenue. But these points do not confirm it and hence more analytical and statistical methods needs to be applied and hence clustering comes into the picture.

Chart, scatter chart

Description automatically generated

* Click on the Analytics -> Cluster (Drag It on chart). The sum of marketing sales and revenue are the two clusters created. The tableau analysed the data and created K-means clustering. The cluster helps to determine which store is better to derive the revenue on the total marketing spend.

Chart, scatter chart

Description automatically generated

* Navigate to Data tab, drag new expansion into the details and from drop down, select show highlighter and we can visualize new or old cities.

|  |  |
| --- | --- |
| *New Cities* | *Old Cities* |

* Looking at both the graphs, it seems that investing in the ‘Cluster 1’ will fetch more revenue that cluster 2. But the points which lie in between cluster 1 and 2, were more likely to be the part cluster 1, but tableau added them in cluster 2.

**Domain Concept Behind Clustering**

It is very important to understand the domain knowledge behind the use case. It is very much essential for any Data Scientist. The organisation/company which we are working on is not some financial service which gives to loan to another company or an industry which is produces some parts.

The company we are dealing with is the laundry service, which cleans the dirty clothes and brings it back. It deals with people and heavenly relies on them. The amount work they get is directly proportional to the amount of people that they can potentially serve with i.e., The term which we are looking here is the population. The amount of clothes is directly co-related to the population. So, population may play very important role for the analysis. So, we need an external dataset by which population can be analysed.

New csv file is added which contains the population of the cities in US. We can blend the files but here we are creating cross data join (joining of excel and csv file). The files have some duplicated data and hence both the files must be joined with ‘state’ column in addition with the ‘city ‘column.

Graphical user interface, text, application

Description automatically generated

**Sheet 5 –Advanced Clustering**

* Drag the population into the clusters and another cluster is created. The charts indicate three different clusters, and few data points/shops even overlap.

**Chart, scatter chart

Description automatically generated**

* Add ‘population’ into Details and analyse the population of each cluster.
* Cluster-3 (Red Colour) shows the population value around 100 – 110k.
* Cluster -2 (Blue Colour) shows the population value around 170 – 212k.
* Cluster -1 (Orange Colour) shows the population value around 100 – 148k.

|  |  |
| --- | --- |
| **Cluster** | **Population** |
| Cluster - 3 (Red Colour) | 100 – 110k |
| Cluster -2 (Blue Colour) | 170 – 212k |
| Cluster -1 (Orange Colour) | 100 – 148k |

* To enhance the trends and differences, navigate to analytics tab and drag trend lines on to the chart.
* Three different trend lines are seen which can be observed in the below figure.

Chart, scatter chart

Description automatically generated

* If we hover over the trend line, it shows three values. Revenue, R-square, and P-value.
* It shows a linear trend which is represented by the equation: *y=mx+c*
* Cluster-3 (Red colour) shows the trend that every 1$ spent on marketing, the revenue goes up by 0.94$.

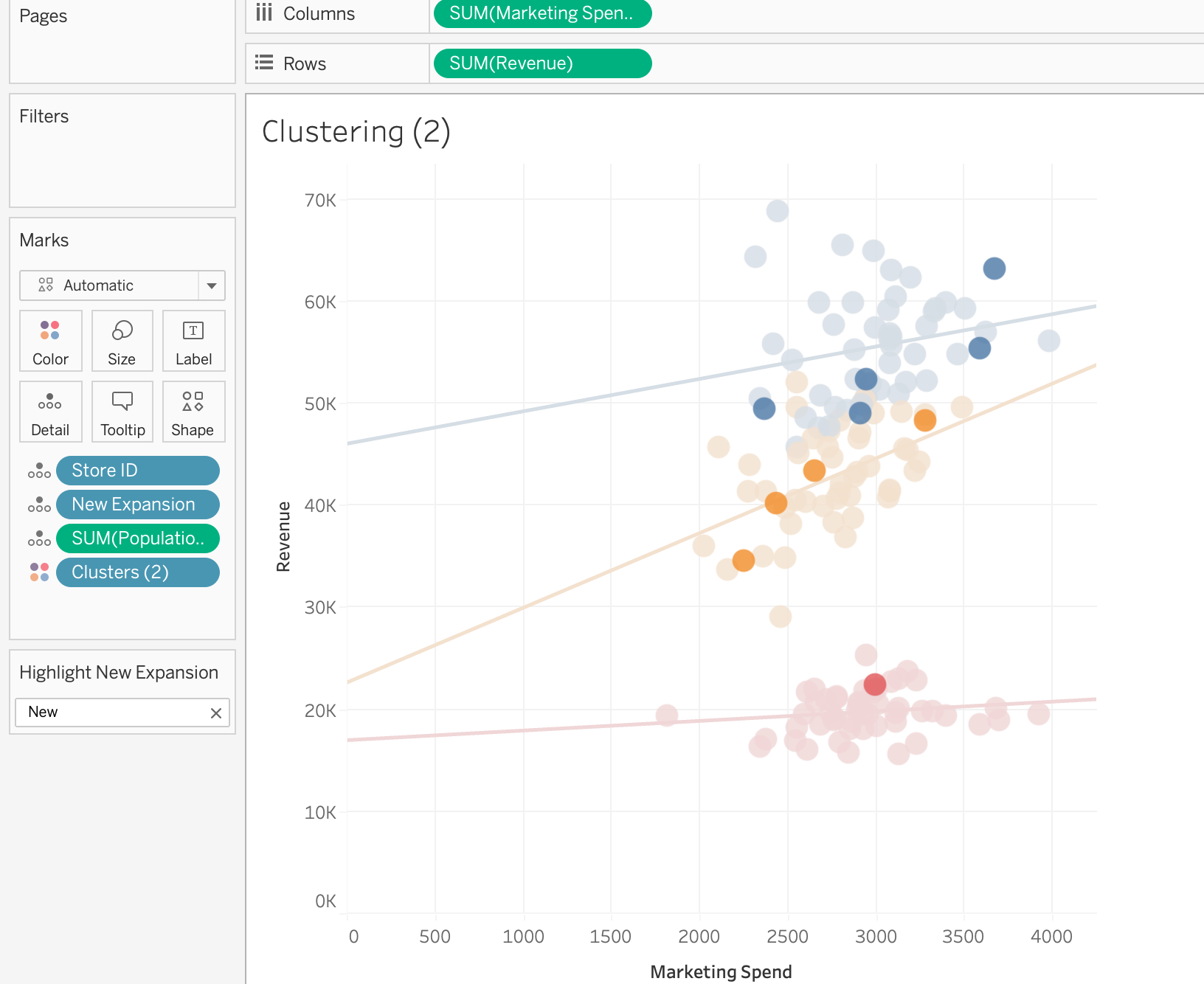
**Chart, scatter chart

Description automatically generated**

* Cluster-2(Blue colour) shows the trend that every 1$ spent on marketing, the revenue goes up by 3.17$.
* Cluster-1(Orange colour) shows the trend that every 1$ spent on marketing, the revenue goes up by 7.32$.

|  |  |
| --- | --- |
| Cluster 2 | Chart, scatter chart  Description automatically generated  Cluster 1 |

* This is very important and crucial part of the analysis required for the organisation to set-up new shops.
* In the highlighting option, select new and 10 new shops data will be observed



* There is one shop in the cluster 3(Red), and it can be predicted that if the company try to invest in marketing than nothing much will be impacted on the revenue and hence investing on this will be a bad idea.
* There are five shops in the cluster 2(Blue) which can be said as the great cities as they are generating much revenue but at-last we receive 3$ for every 1$ invested in the marketing.
* There are found shop in the cluster 1(Orange) which can be best location to invest in as the revenue generated is maximum when compared to other two clusters.

Hence the trend lines help for better analysis and interpretation of the data. Even though the cluster 2(Blue) indicates few shops with big revenue, but trend line indicates cluster 1 provides greater revenue if any new shop is opened in that city/region.

To view a better picture with the maps, the clusters have been added to the maps which indicate the clusters of different cities and states.

Map

Description automatically generated