Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Lab Manual

Subject: Foundations of Data Analysis Laboratory (DJ19DSL303)

Semester: III **Experiment 4** **(Data Visualization)**

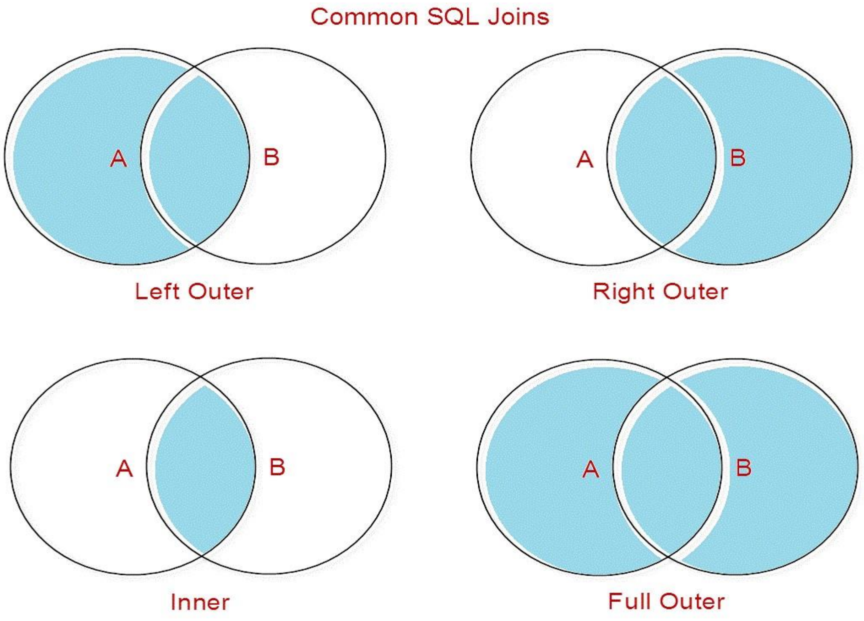
Name: Dev Patel **SAP ID:60009200016**

Batch: K/K1 **Date: 07/12/2021**

Aim: Perform joins, blends and create dual axis charts.

Theory:

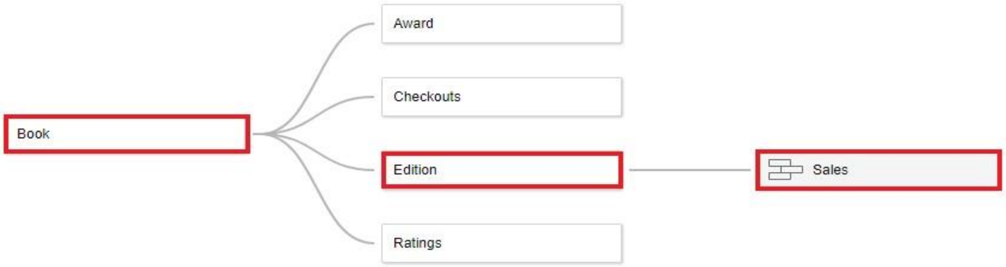
Data joining is a common requirement in any data analysis. You may need to join data from different tables in a single source or join data from multiple sources. A join means combining columns from one or more tables in a relational database. It also creates a set that can be saved as a table, or it can be used as it is. Joins are a concept taken from relational databases like SQL and may be very useful in the future.



Types of joins.

A blend is a smart join on the go. It is always a left join. Blends work for databases of different sources as well. Tableau may automatically suggest/use blends if fields are having the same name. If not, fields can be blended manually. Tableau Relationships

Relationships are a flexible way to combine data for multi-table analysis in Tableau. Think of a relationship as a contract between two tables. When you are building a viz with fields from these tables, Tableau brings in data from these tables using that contract to build a query with the appropriate joins.



When to use Joins & Blends:

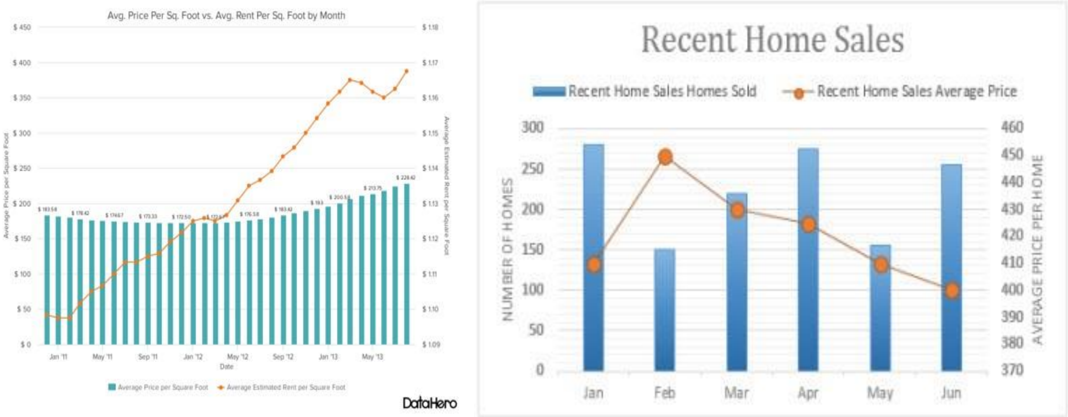
Joins: Combining data at row level

Blends: When data sources have different level of granularity or when data sources are from different sources

Relationships: Tableau connects the data at the right level of aggregation. It is a sort of smart on the fly method of connecting tables. It's more flexible than a join. Whenever in doubt, use a relationship. Use a Join or blend if 100% sure that it is required.

Dual Axis Charts:

A dual axis chart is a great way to easily illustrate the relationship between two different variables. They illustrate a lot of information with limited space and allow you to discover trends you may have otherwise missed if you're switching between graphs. Synchronizing Dual Axis charts is very important in most cases.



Examples of Dual Axis Charts

Lab Assignments to complete in this session

Use the given datasets and perform the following tasks:

Tool Used:

Tableau Public 2021.3

Datasets:

1) https://community.tableau.com/s/question/0D54T00000CWeX8SAL/sample-superstore-sales-excelxls

   Sample Superstore Dataset:

   This is a sample superstore dataset, a kind of data analysis to deliver insights on how the company can increase its profits while minimizing the loss.

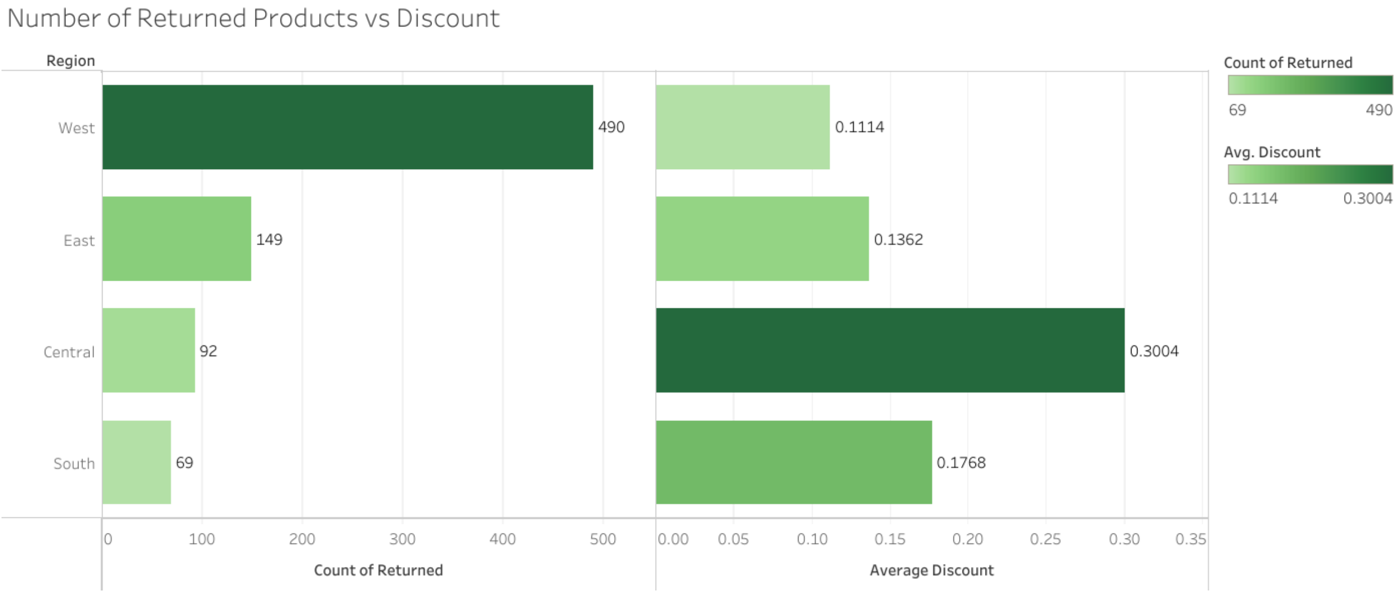
2) https://data.world/2918diy/coffee-chain/workspace/file?filename=Coffee+Chain.txt

Coffee Chain Dataset:

This is a coffee chain dataset, a kind of data analysis to deliver insights on how the company can increase its profits while minimizing the loss.

Visualizations:

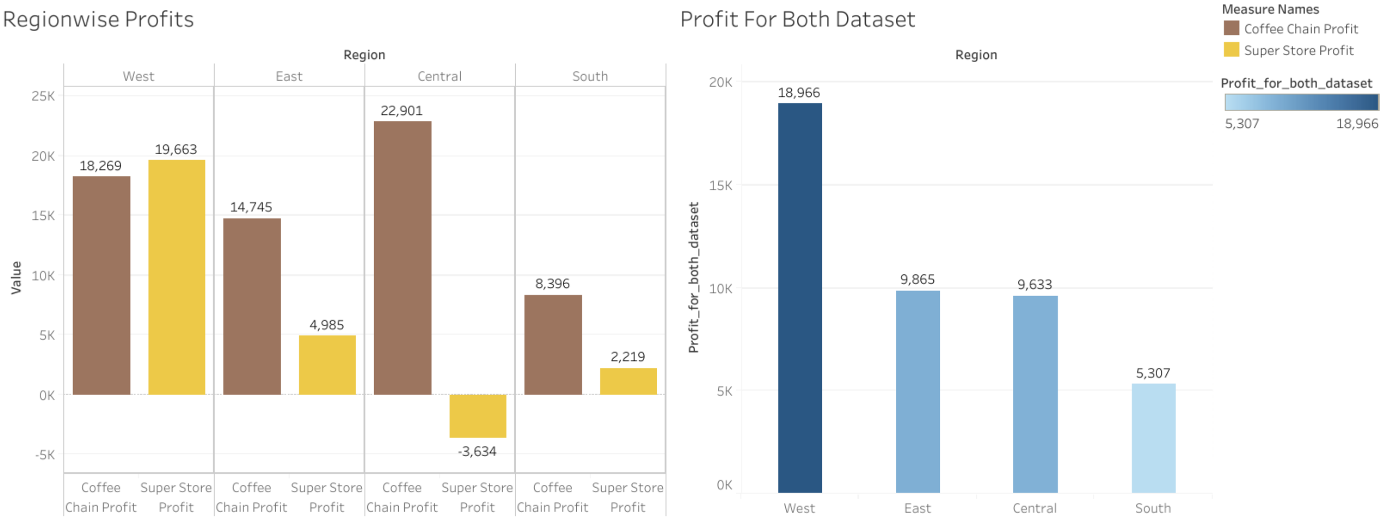
a) **The company wants to analyse region wise return of products from superstore dataset. Is there any relationship of return of product with the discounts offered on it?**
On X-axis: Count of Returned Products and average discount
On Y-axis: Regions
To analyse any relationship between number of returned products and the discount that was provided, we have visualised the data as follows.



From the visualisation, we can see that while the Central region had the highest average discount, the number of products returned by them are the second lowest. Also, while West region had the highest number of products returned, the region had the lowest average discount for any region. Similarly, South region had least number of products returned and had the second highest average discount. While there is no concrete relationship between number of returned products and the discount offered it seems like they lowkey follow an inverse relationship.

b) **Compare the region wise profits of superstore and coffee chain (Market means region). Find which region has performed well for both.**

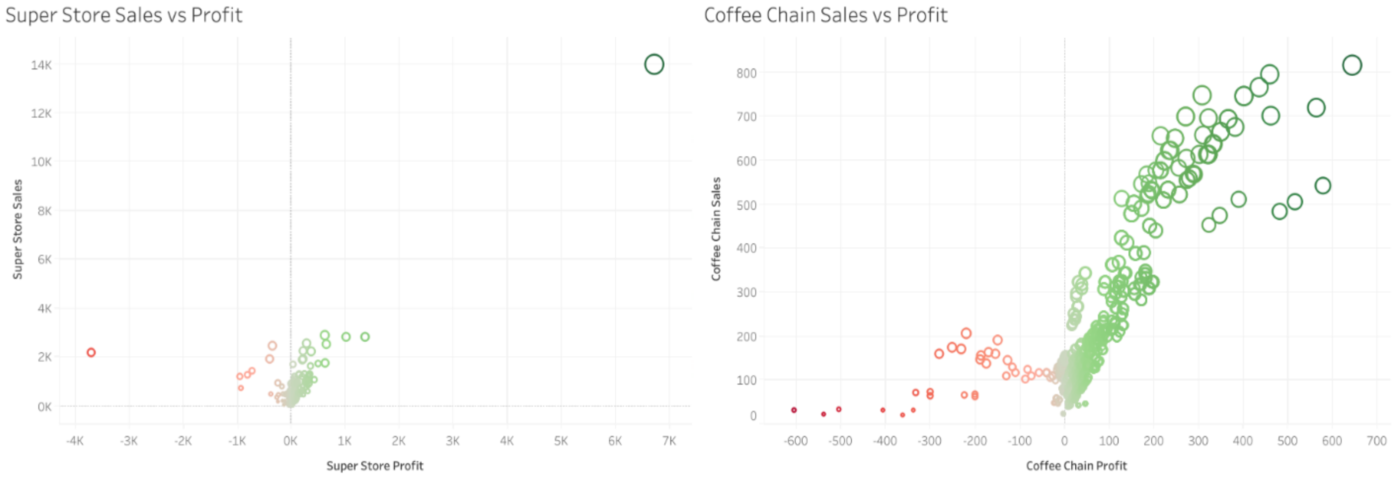
On dashboard:
Region-wise profits for both Superstore and Coffee chain datasets.
The profit for both datasets combined.



From the visualization, we can conclude that the West region has performed considerably well in both the organisations in comparison to its other counterparts. While the Central region shows the biggest profit for Coffee Chain, it incurs a loss for the super store.

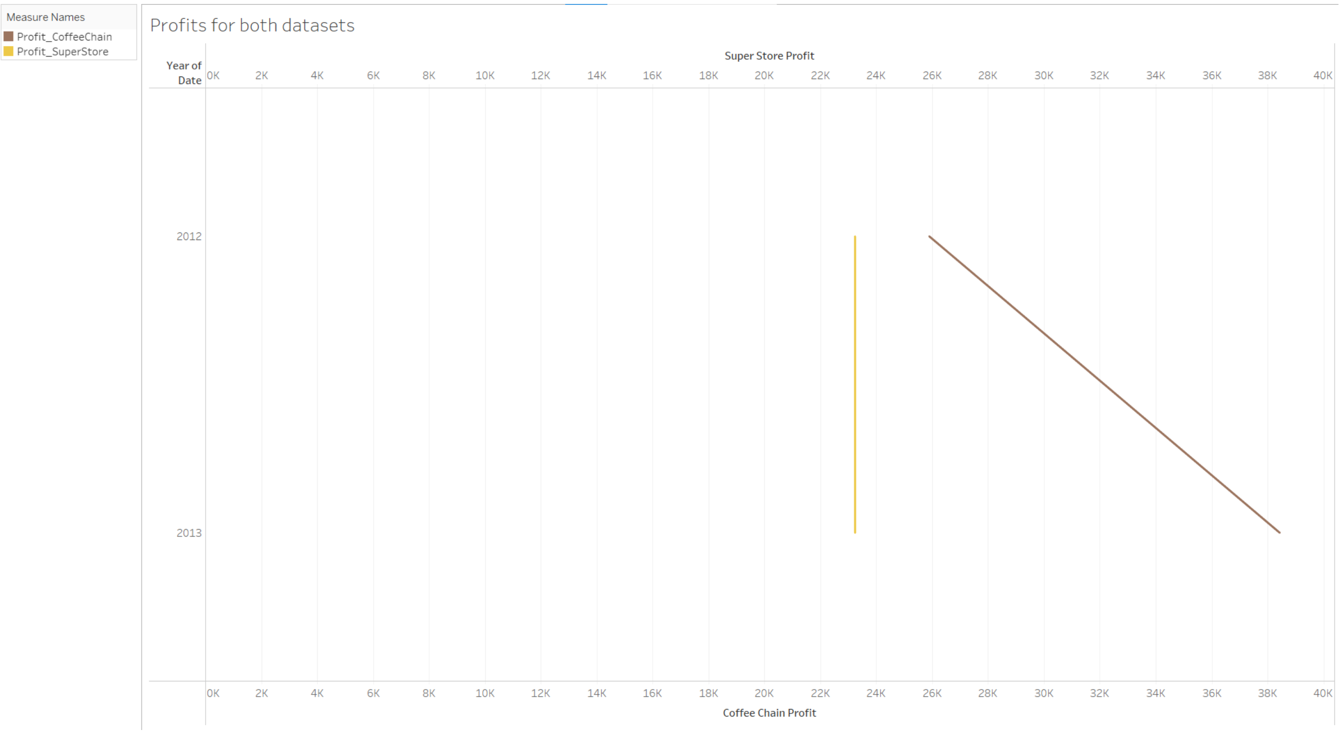
c) **Show sales and profit together for both the datasets choosing an appropriate visualization.**
On dashboard:
Sales vs Profit for Super Store and Sales vs Profit for Coffee Chain

The below visualisation shows the sales and profit together for both datasets in a scatterplot. For Coffee Chain, the highest profit made is of 646 and the highest sales is of 815. For Super store, the highest profit made is of 6720 and the highest sales is of 14000.



d) **Analyse year wise profit for both the datasets and illustrate the limitations of blending.**

The below visualisation shows the year-wise profit for both the organisations. The axis has been synchronised for a better comparison of profits for each of the dataset.



Limitations of blending:

For this dataset as both the datasets have different values in columns i.e., years are different so when we blend, the years don't match and we don't see the sales of one of the datasets. In general, the limitations of blending are as follows:

1. Non-additive aggregates like MEDIAN and COUNT have data blending issues.
2. Publishing the blended data source is complicated. You need to publish each data source and then blend the published data source together.
3. Secondary data sources are always calculated and aggregated.
4. Cube data sources must be the primary data source, always.