US SuperStore : Ne...

FINISHED

US SuperStore - New Opening Store Feasibility Analysis (Group5)

Took 0 sec. Last updated by anonymous at April 04 2023, 10:15:30 AM.

FINISHED

DataSets Overview

Two Datasets have been used - Superstore2.csv and citypopulation.csv.

The Superstore2.csv dataset contains sales data of Superstore in the US for 4 years(2015 - 2018).

The dataset includes the following columns - Orderid/Orderdate/Shipdate/ShipMode/Customerid/CustomerName/Segment/Country/City/State/PostalCode/

The Customers are divided into 3 segments : Customer, Corporate and Home Office.

Category is divided into 3 segments : Furniture, Office supplies and Technology.

The CityPopulation.csv has 3 columns - City/State/Population.

Took 0 sec. Last updated by anonymous at April 04 2023, 10:11:03 AM.

FINISHED

Feasibility Analysis

Look for the potential region to open new store next year.

Financial & Demographic Analysis

1) Top Sales by Region
2) Top 5 Cities with the highest sales, from the potential region

Sales Metrics (for potential region & cities)

1) Customer Segments
2) Top Product Category & Product Sales
3) Seasonal Sales Trends

Took 0 sec. Last updated by anonymous at April 04 2023, 10:13:17 AM.

Step1: Create DataFrame from CSV File

FINISHED

Took 0 sec. Last updated by anonymous at March 30 2023, 6:20:46 PM.

FINISHED

```
val Superstore2 = spark.read
.option("inferSchema", "true")
.option("header", "true")
.csv("/tmp/group5/Superstore2.csv")
```

```
Superstore2: org.apache.spark.sql.DataFrame = [RowID: int, OrderID: string ... 16 more fields]
```

Took 0 sec. Last updated by anonymous at April 04 2023, 8:16:27 PM.

FINISHED

```
val CityPopulation = spark.read
.option("inferSchema", "true")
.option("header", "true")
.csv("/tmp/CityPopulation.csv")
```

```
CityPopulation: org.apache.spark.sql.DataFrame = [City: string, State: string ... 1 more field]
```

Took 1 sec. Last updated by anonymous at April 01 2023, 10:12:39 PM.

Step2: Print the DataFrame Schema in a tree format

FINISHED

Took 0 sec. Last updated by anonymous at March 30 2023, 6:23:04 PM.

FINISHED

```
%spark2
Superstore2.printSchema()
```

```
root
 |-- RowID: integer (nullable = true)
 |-- OrderID: string (nullable = true)
 |-- OrderDate: timestamp (nullable = true)
 |-- ShipDate: string (nullable = true)
 |-- ShipMode: string (nullable = true)
 |-- CustomerID: string (nullable = true)
 |-- CustomerName: string (nullable = true)
 |-- Segment: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- City: string (nullable = true)
 |-- State: string (nullable = true)
 |-- PostalCode: integer (nullable = true)
 |-- Region: string (nullable = true)
 |-- ProductID: string (nullable = true)
 |-- Category: string (nullable = true)
 |-- Sub_Category: string (nullable = true)
 |-- ProductName: string (nullable = true)
```

Took 1 sec. Last updated by anonymous at April 04 2023, 8:16:49 PM.

FINISHED

```
%spark2
CityPopulation.printSchema()
```

```
root
 |-- City: string (nullable = true)
 |-- State: string (nullable = true)
 |-- Population: integer (nullable = true)
```

Took 0 sec. Last updated by anonymous at April 01 2023, 10:14:23 PM.

Step3: Convert DataFrame to TempView FINISHED

Took 0 sec. Last updated by anonymous at April 01 2023, 12:46:49 PM.

```
%spark2

Superstore2.createOrReplaceTempView("SuperstoreView")
```

FINISHED

Took 0 sec. Last updated by anonymous at April 01 2023, 1:15:25 PM.

```
%spark2

CityPopulation.createOrReplaceTempView("PopulationView")
```

FINISHED

Took 0 sec. Last updated by anonymous at April 01 2023, 10:15:06 PM.

Step4: Query the data from TempView to check whether the data is ready to use. FINISHED

Took 0 sec. Last updated by anonymous at March 31 2023, 4:29:55 PM.

```
%spark2.sql
SELECT *
FROM SuperstoreView
LIMIT 10
```

FINISHED

RowID ▼	OrderID ▼	OrderDate ▼	ShipDate ▼	ShipMode ▼	CustomerID ▼	Custor
1	CA-2017-152156	2017-11-08 00:00:00.0	11-11-2017	Second Class	CG-12520	Claire (
2	CA-2017-152156	2017-11-08 00:00:00.0	11-11-2017	Second Class	CG-12520	Claire (
3	CA-2017-138688	2017-06-12 00:00:00.0	16-06-2017	Second Class	DV-13045	Darrin '
4	US-2016-108966	2016-10-11 00:00:00.0	18-10-2016	Standard Class	SO-20335	Sean C
5	US-2016-108966	2016-10-11 00:00:00.0	18-10-2016	Standard Class	SO-20335	Sean C

Took 0 sec. Last updated by anonymous at April 01 2023, 1:15:49 PM.

```
%spark2.sql
```

```
SELECT *
FROM PopulationView
ORDER BY Population desc
LIMIT 10
```

FINISHED

City	State
California	California
Texas	Texas
Florida	Florida
New York	New York
Pennsylvania	Pennsylvania
Illinois	Illinois
Ohio	Ohio
Georgia	Georgia
North Carolina	North Carolina

Took 0 sec. Last updated by anonymous at April 04 2023, 10:14:17 AM.

FINISHED

Financial & Demographic Analysis: Top Sales by Region

Took 0 sec. Last updated by anonymous at April 04 2023, 10:46:25 AM.

```
%spark2.sql
SELECT
    Region,
    ROUND(SUM(Sales),2) as TotalSales,
    COUNT(*) as TotalRecords
FROM SuperstoreView
GROUP BY Region
ORDER BY ROUND(SUM(Sales),2) desc
```

FINISHED

settings ▲

All fields:

Region TotalSales TotalRecords

Keys

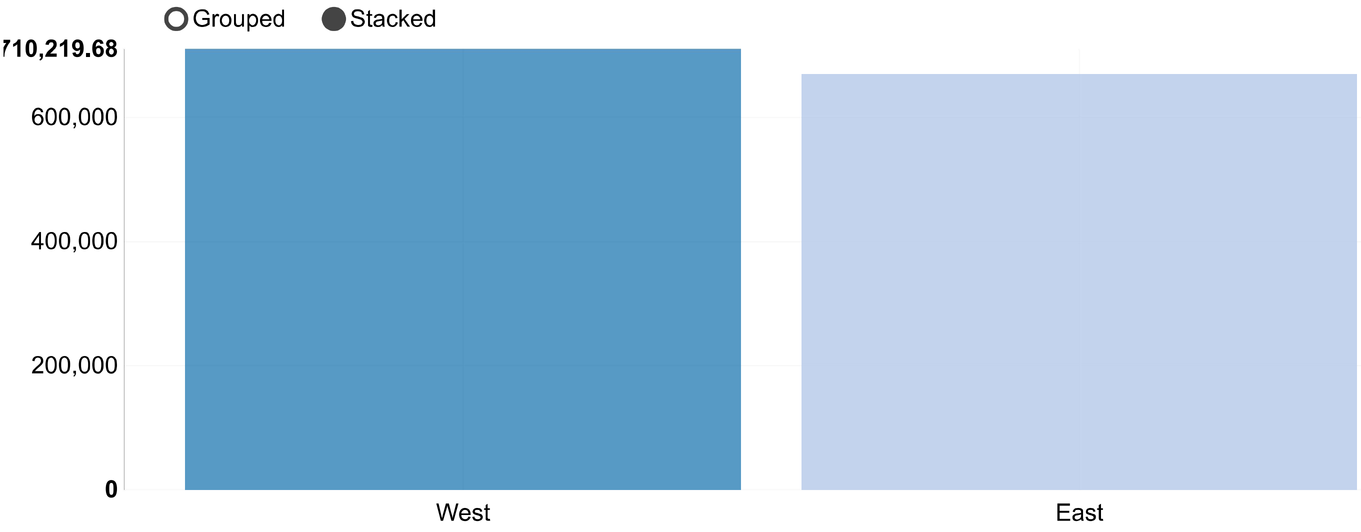
Region ✖

Groups

Region ✖

Values

TotalSales SUM ✖



Took 2 sec. Last updated by anonymous at April 01 2023, 1:16:18 PM. (outdated)

FINISHED

Financial & Demographic Analysis: Top 5 Cities with the highest sales, from the West region

Took 0 sec. Last updated by anonymous at April 04 2023, 10:46:39 AM.

FINISHED

```
%spark2.sql
SELECT
    Region,
    City,
    ROUND(SUM(Sales),2) as TotalSales
FROM SuperstoreView
GROUP BY Region, City
HAVING Region = 'West'
```

```
ORDER BY TotalSales desc
LIMIT 5
```

Region ▼	City
West	Los Angeles
West	Seattle
West	San Francisco
West	San Diego
West	Denver

Took 1 sec. Last updated by anonymous at April 03 2023, 2:43:02 PM.

FINISHED

Financial & Demographic Analysis: Sales per Capita

Took 0 sec. Last updated by anonymous at April 04 2023, 10:47:14 AM.

FINISHED

```
%spark2.sql
SELECT s.City, ROUND(SUM(s.Sales),2) AS TotalSales,
c.Population,
ROUND((SUM(s.Sales)/c.Population),2) AS SalesPerCapita
FROM SuperstoreView AS s
JOIN PopulationView c ON s.City = c.City
GROUP BY s.City, c.Population
ORDER BY SalesPerCapita DESC
LIMIT 5
```

settings ▲

All fields:

City TotalSales Population SalesPerCapita

Keys

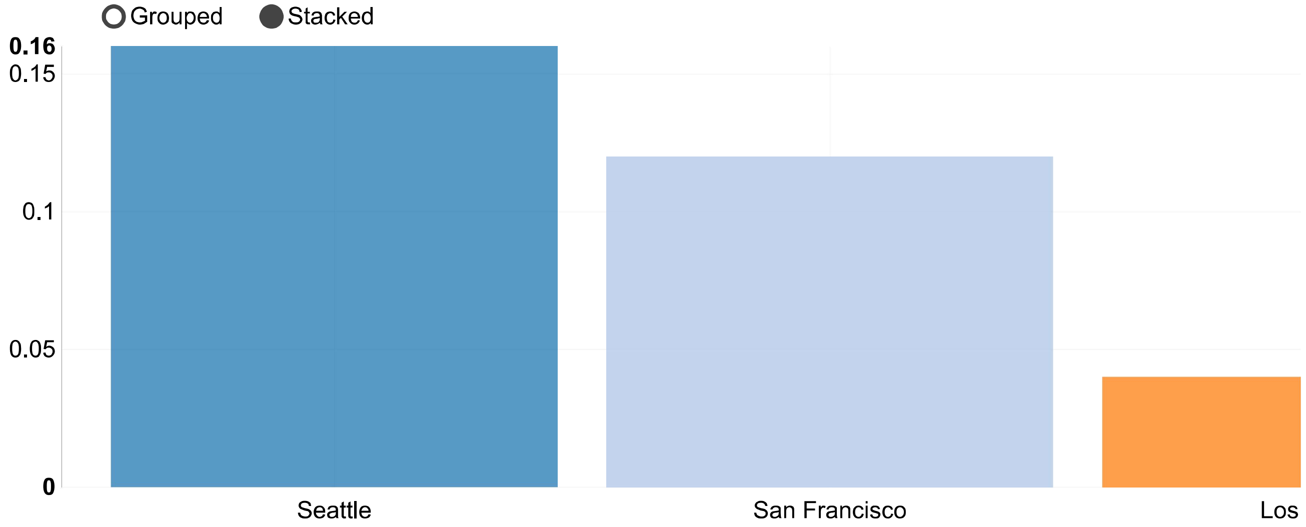
City ✖

Groups

City ✖

Values

SalesPerCapita SUM ✖



Took 2 sec. Last updated by anonymous at April 04 2023, 10:48:34 AM.

FINISHED

Sales Metrics: Customer Segments in Seattle

Took 0 sec. Last updated by anonymous at April 04 2023, 10:30:22 AM.

FINISHED

```
%spark2.sql
SELECT
    Segment as CustomerSegment,
    ROUND(SUM(Sales),2) as TotalSales
FROM SuperstoreView
WHERE City = 'Seattle'
GROUP BY Segment
ORDER BY TotalSales desc
```

settings ▲

All fields:

CustomerSegment TotalSales

Keys

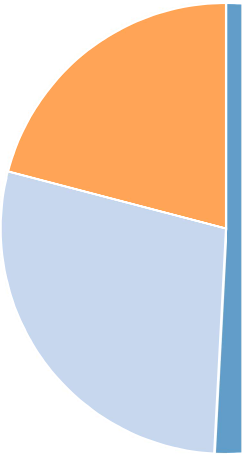
CustomerSegment ✖

Groups

CustomerSegment ✖

Values

TotalSales SUM ✖



Took 0 sec. Last updated by anonymous at April 03 2023, 2:43:34 PM. (outdated)

FINISHED

Sales Metrics Analysis: Top Product Categories in Seattle

Took 0 sec. Last updated by anonymous at April 04 2023, 10:32:44 AM.

```
%spark2.sql
SELECT
    Category as Product,
    ROUND(SUM(Sales),2) as TotalSales
FROM SuperstoreView
WHERE City = 'Seattle'
GROUP BY Category
ORDER BY TotalSales desc
LIMIT 10
```

FINISHED

settings ▲

All fields:

Product TotalSales

Keys

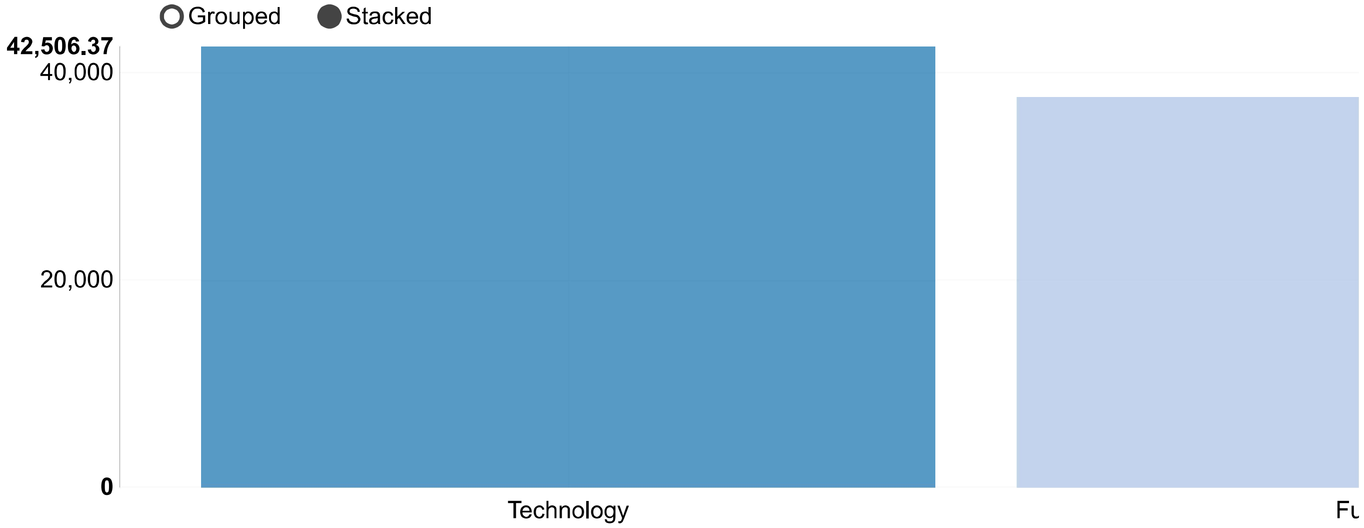
Product ✖

Groups

Product ✖

Values

TotalSales SUM ✖



Took 1 sec. Last updated by anonymous at April 03 2023, 2:58:28 PM. (outdated)

FINISHED

Sales Metric Analysis: Top 10 Product Sales in Seattle

Took 0 sec. Last updated by anonymous at April 04 2023, 10:36:33 AM.

FINISHED

```
%spark2.sql
SELECT
    Sub_Category as Product,
    ROUND(SUM(Sales),2) as TotalSales
FROM SuperstoreView
WHERE City = 'Seattle'
GROUP BY Sub_Category
ORDER BY TotalSales desc
LIMIT 10
```

settings ▲

All fields:

Product TotalSales

Keys

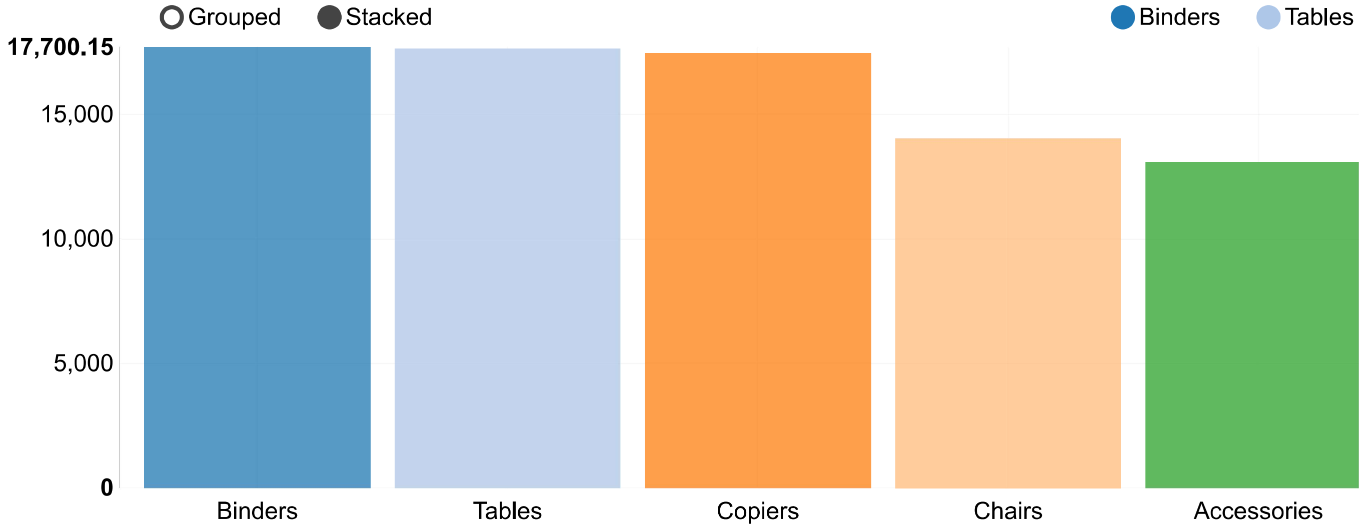
Product ✖

Groups

Product ✖

Values

TotalSales SUM ✖



Took 1 sec. Last updated by anonymous at April 03 2023, 11:26:30 PM. (outdated)

FINISHED

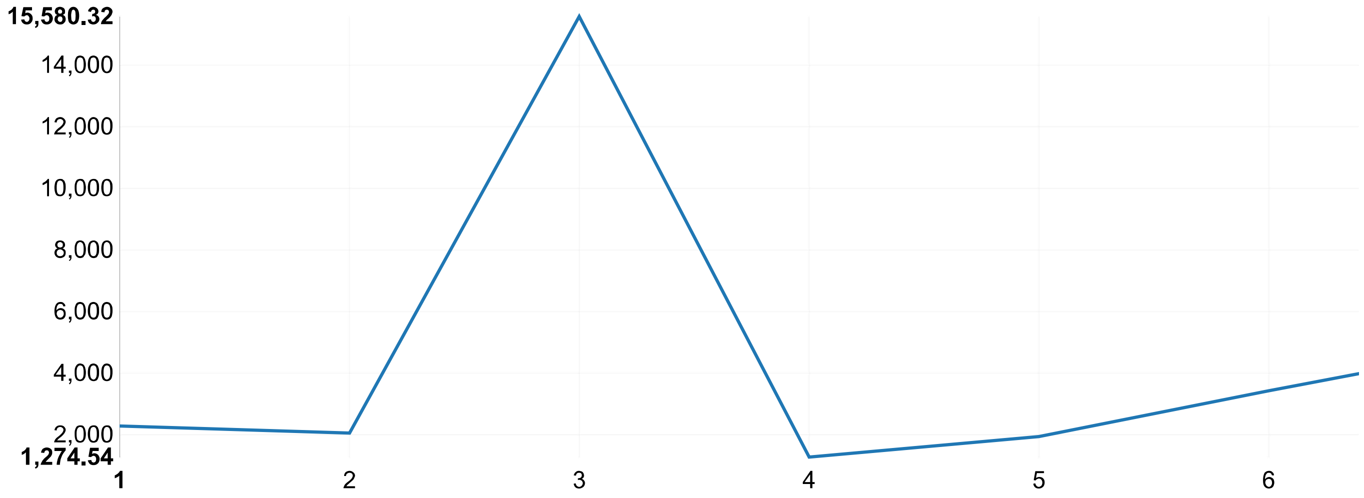
Sales Metrics: Seasonal Sales Trends in Seattle

Took 0 sec. Last updated by anonymous at April 04 2023, 10:38:11 AM.

FINISHED

```
%spark2.sql
SELECT
    Month(OrderDate) as Month,
    ROUND(SUM(Sales),2) as TotalSales
FROM SuperstoreView
WHERE YEAR(OrderDate) = '2018' and City = 'Seattle'
GROUP BY Month(OrderDate)
ORDER BY Month(OrderDate)
```

settings ▾



Took 2 sec. Last updated by anonymous at April 04 2023, 10:38:43 AM.

```
%spark2.sql
```

READY