

**INTRODUCTION**

We have taken dataset which is based on the products that are brought by different consumers in all regions in USA. The data set consists of three sheets Orders, People and Returns

* Link of dataset: <https://community.tableau.com/docs/DOC-1236>
* Author: Michael Martin
* Total Rows: 9995
* Total Attributes: 21
* Column Names: Here are some important columns below:

Order Id: Represents the Id for every product

Order Data: Represents when the product was ordered

Ship Date: Represents when the product was delivered

Sales: Represents how much sale of product was done in particular region

Quantity: Represents how many of same product were purchased.

Discount: Represents discount assigned to product

Profit: Represents profit earned from the order

Returned: Whether the product is returned or not

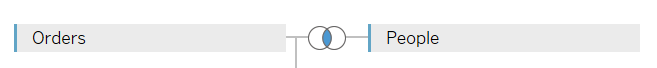
Our main aim is to achieve the following insights from it:

* What is common shipping mode preferred?
* What segmentation gives more profit to store?
* Which state(city) specifically should store invest more?
* Forecasting future profit estimates for different categories

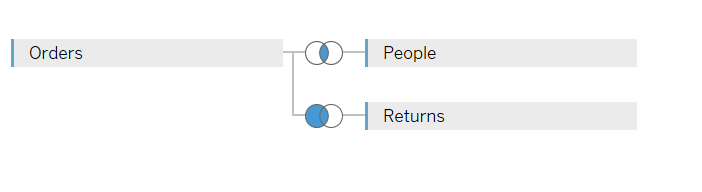
**TABLEAU INSIGHTS**

Firstly, there are three sheets and we need to join them and have a good knowledge on which join we should to combine so that we don’t have any multiple entries.

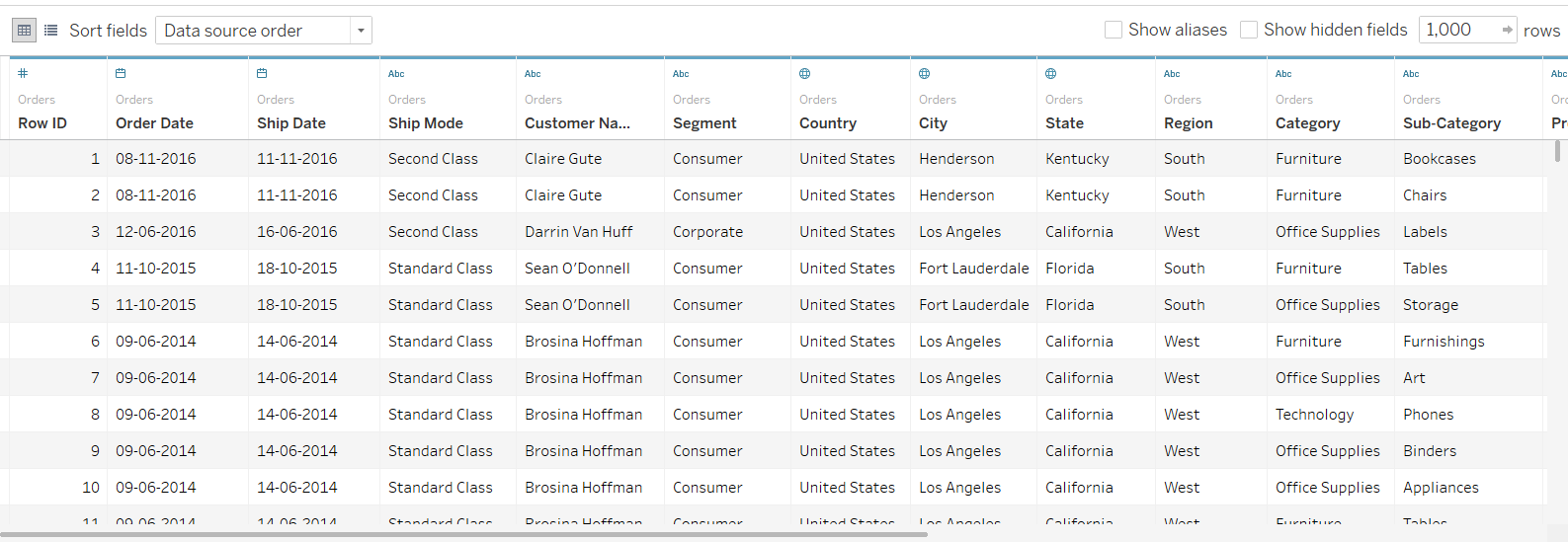
We first did inner join of Orders and People sheet



Then we did left outer join of Orders and Returns so our final data looks like these:

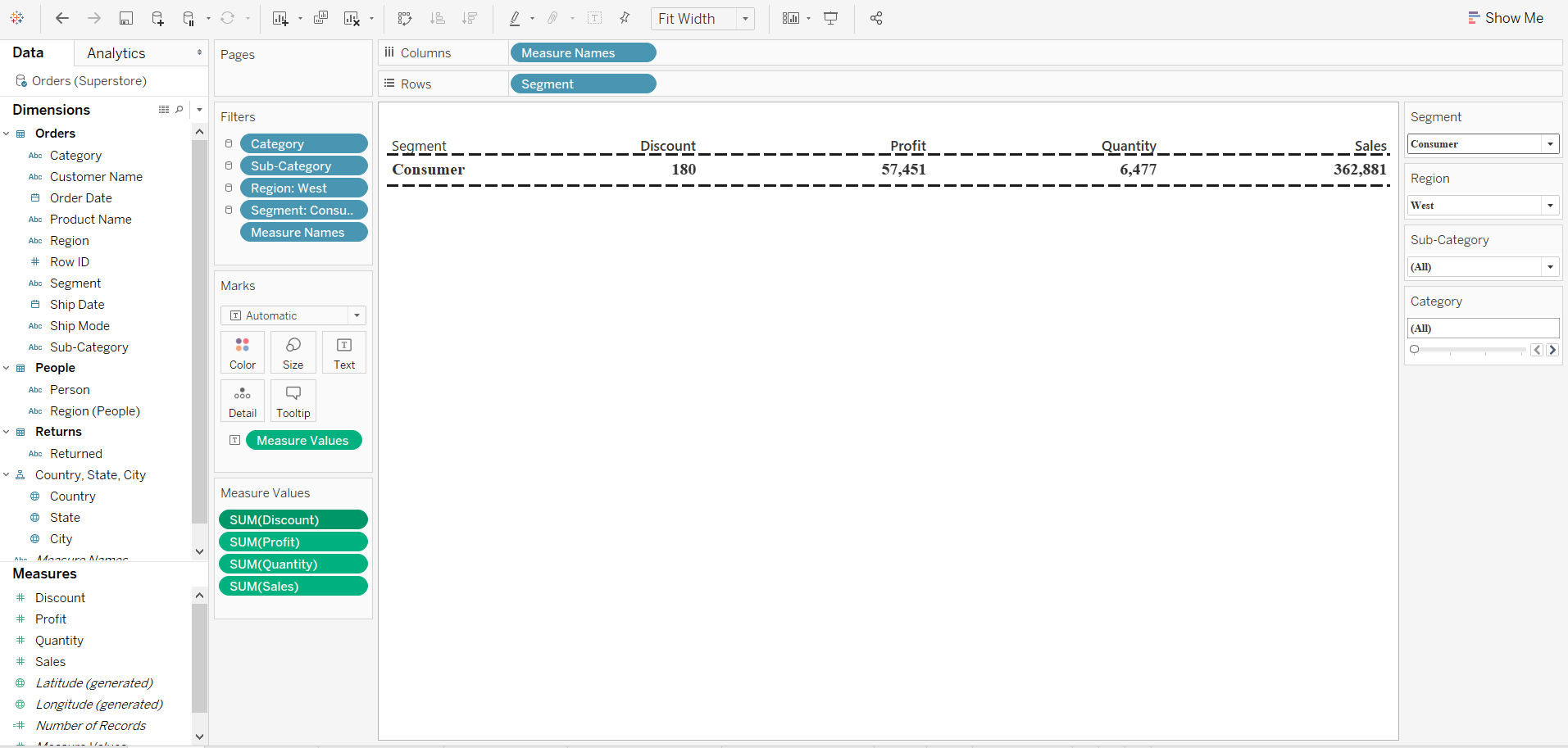


Dataset now looks like this:



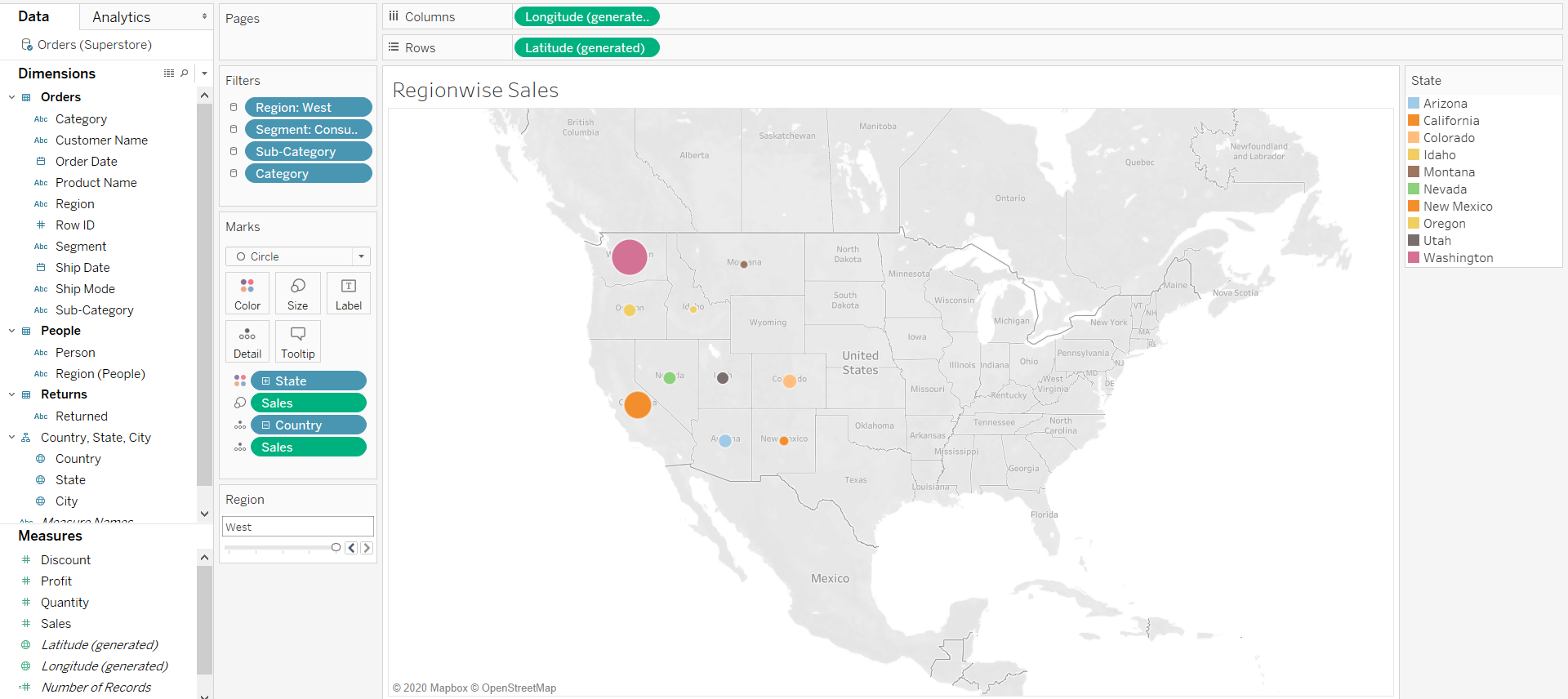
1. Summary

In this we have created a short summary on segments and provided the overall detail about the profit, discount, quantity and sales. We have performed analysis to meet the questions that are mentioned in the introduction.



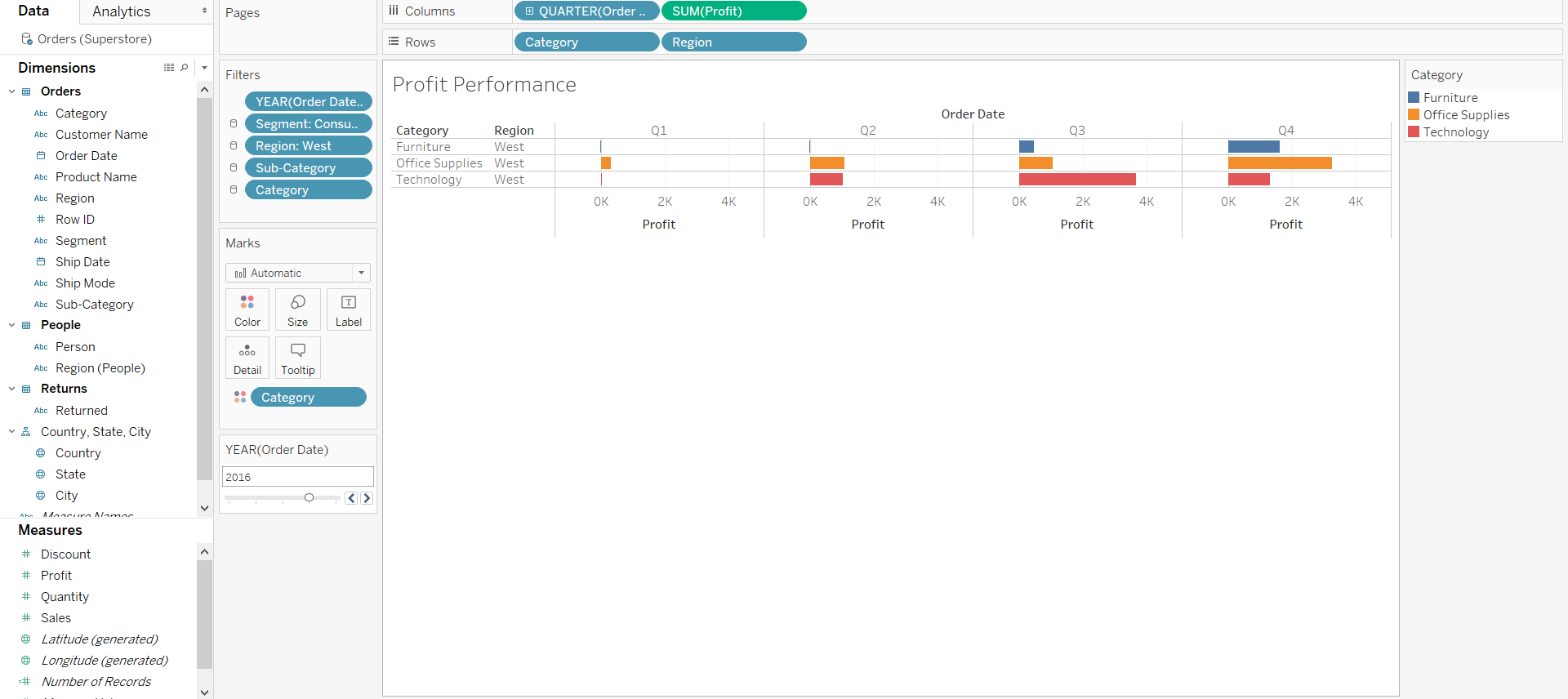
1. Region-wise Sales

In this we have used the map of US and shows the regions that have the count of sales when we hover over it. The size of the circle over the region represents the density of sales in the particular region



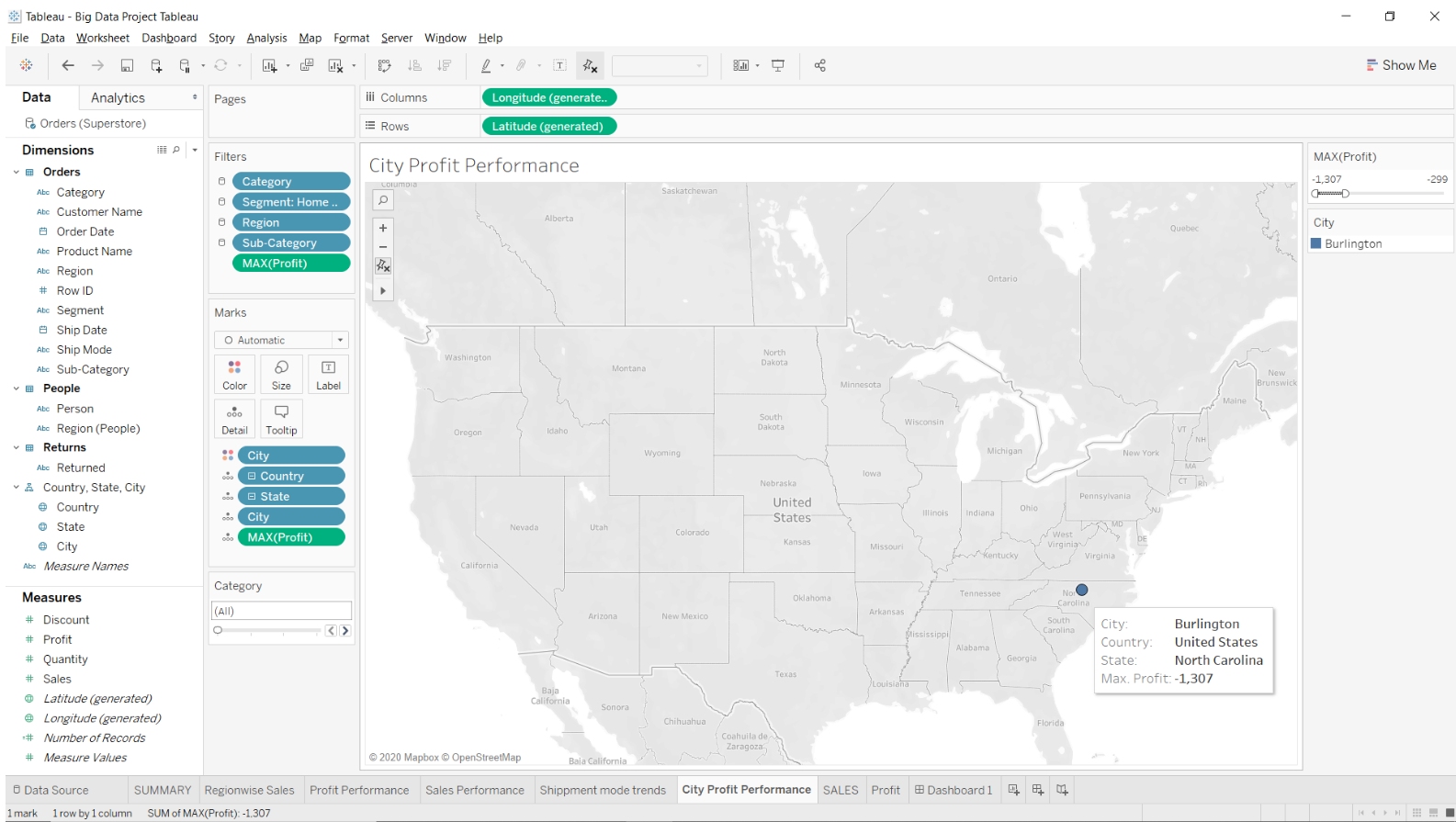
1. Profit Performance

We have here explained the quarterly earned profit on each category over the whole country. We can see that the profit in the first quarter is quite negligible compared to the last quarter.



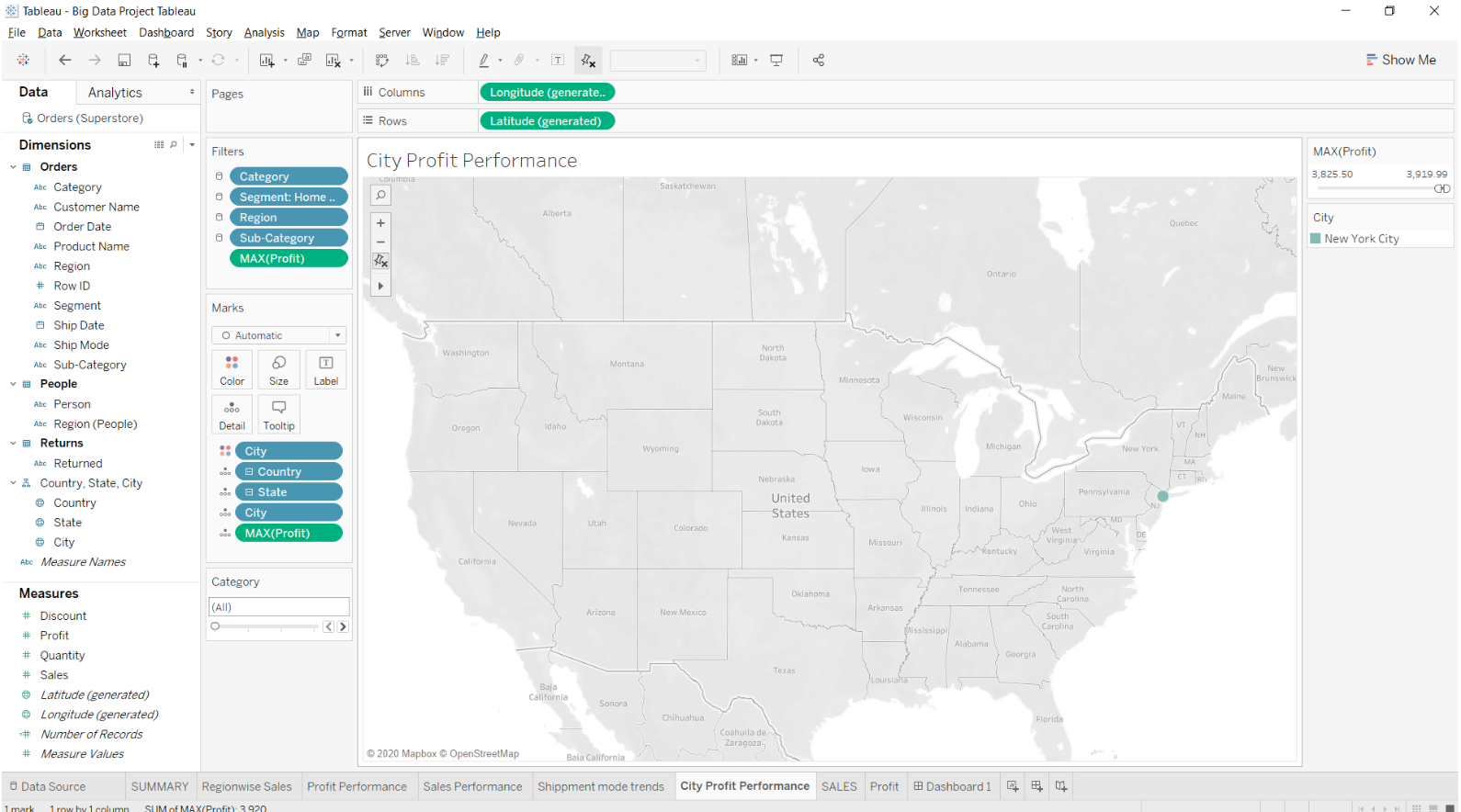
1. Least Profit

The least profit earned was from the city of Burlington and we can say that it suffered a loss of about 1,307 units.



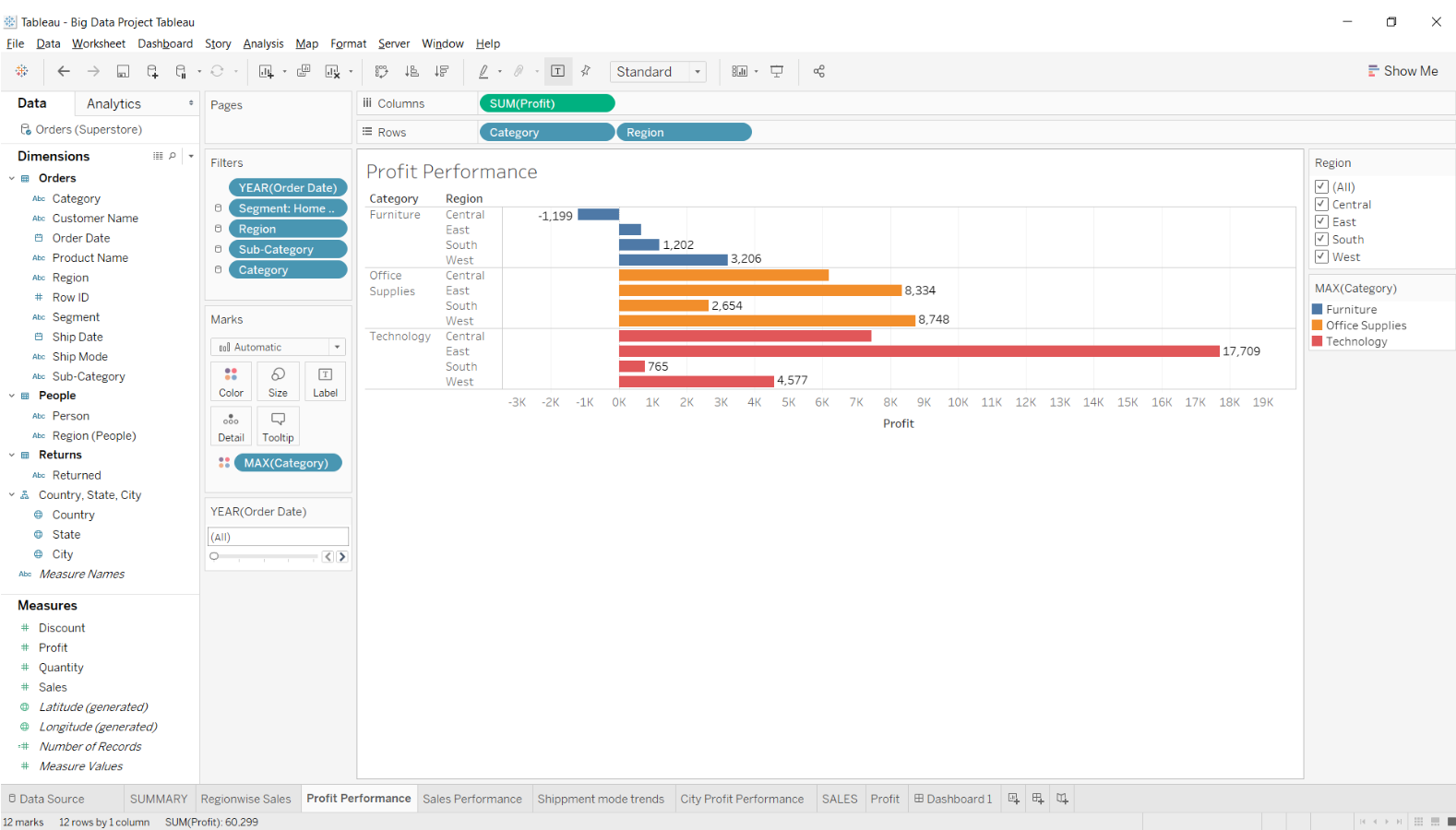
1. Maximum Profit

The maximum profit earned was from the city of New York. So investment in this state would lead to more sales and hence more profit.



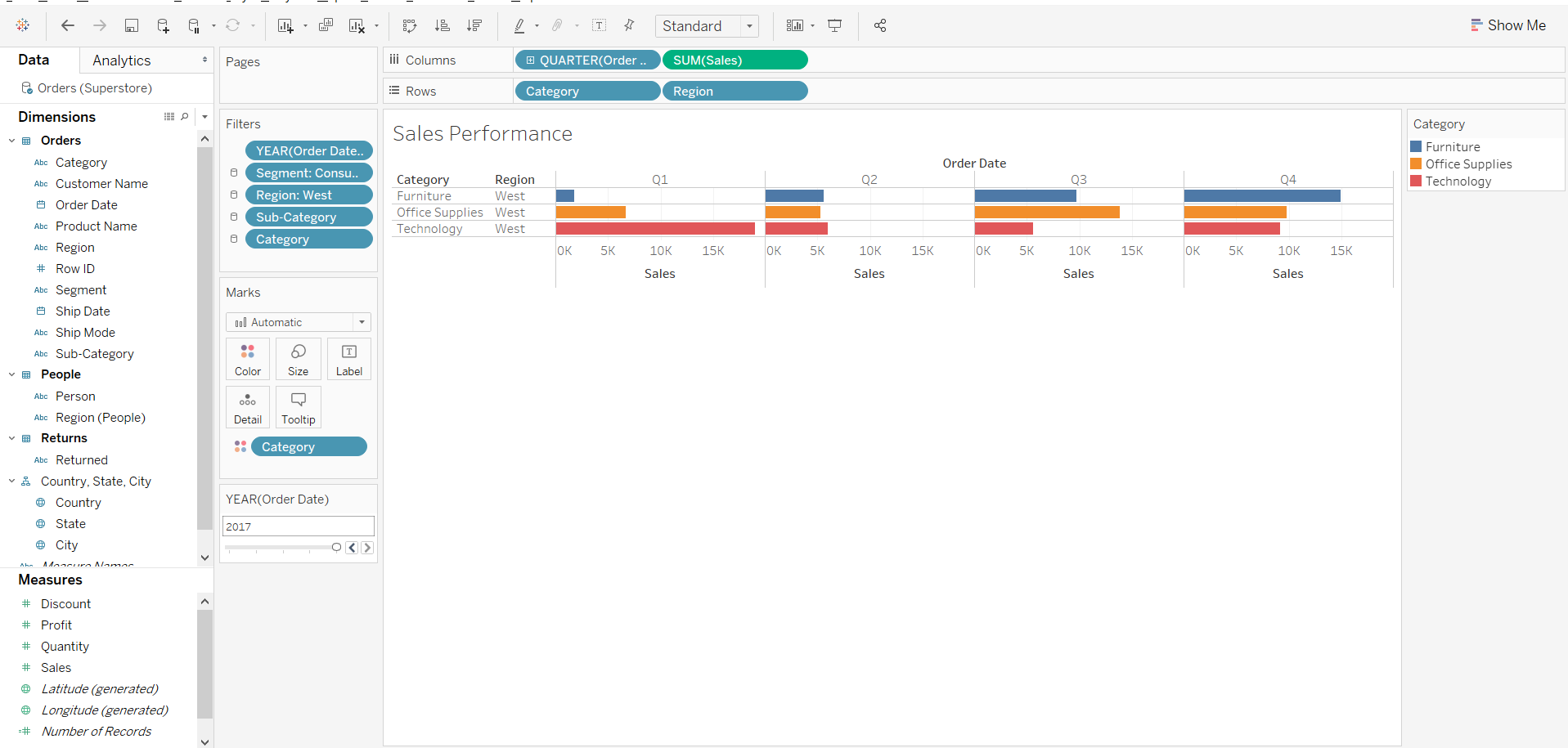
1. Region-wise Profit

We can see Technology sector has gained more profit in the East region of US. The West region of US has gained more profit in furniture sector compared to other regions. Office supplies is more demanding and gained more profit in the West region.



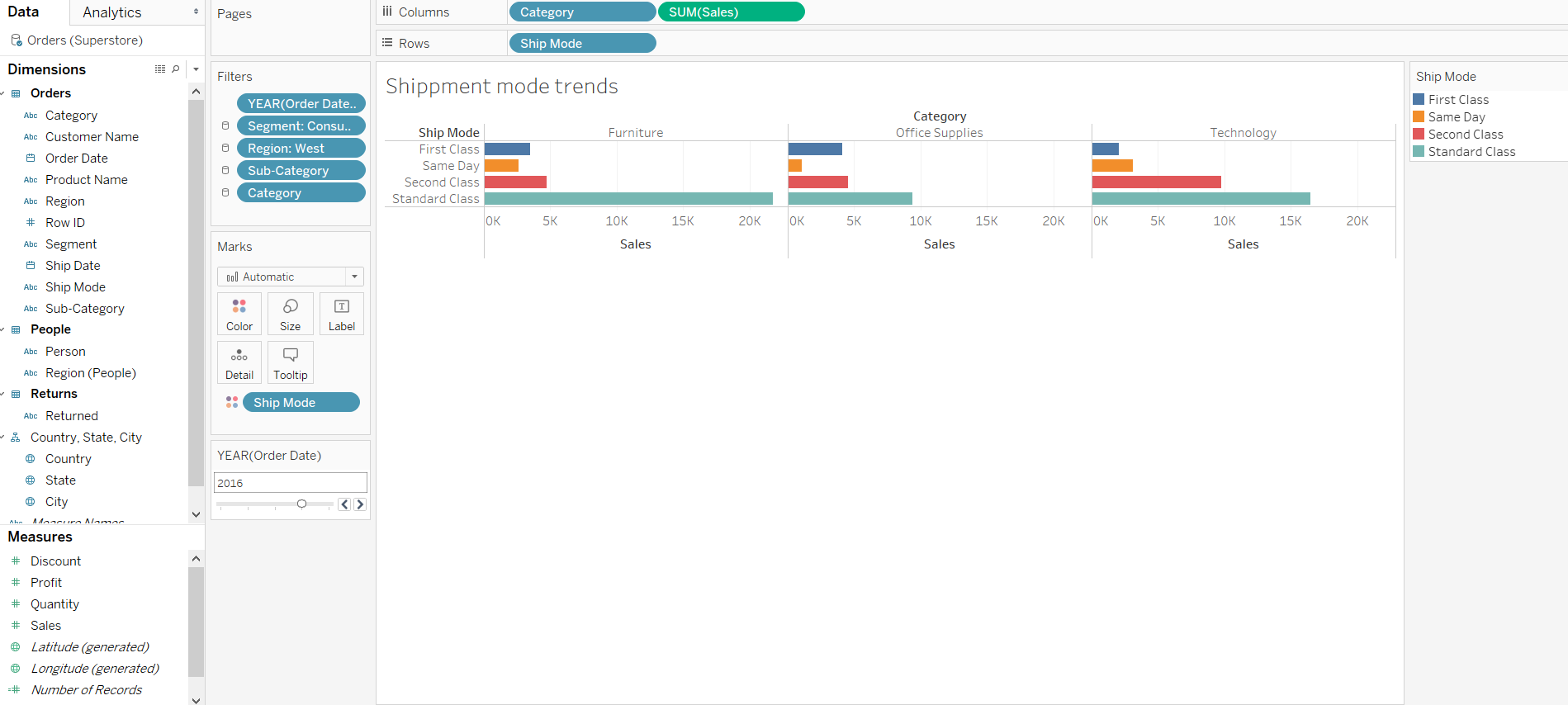
1. Sales Performance

We have analysed the quarterly sales in the different regions of US and found out that technology sector had more demand in the first quarter compared to the third quarter. Furniture had a significant increase in sales from first quarter and reached the highest sales in the end of the year.



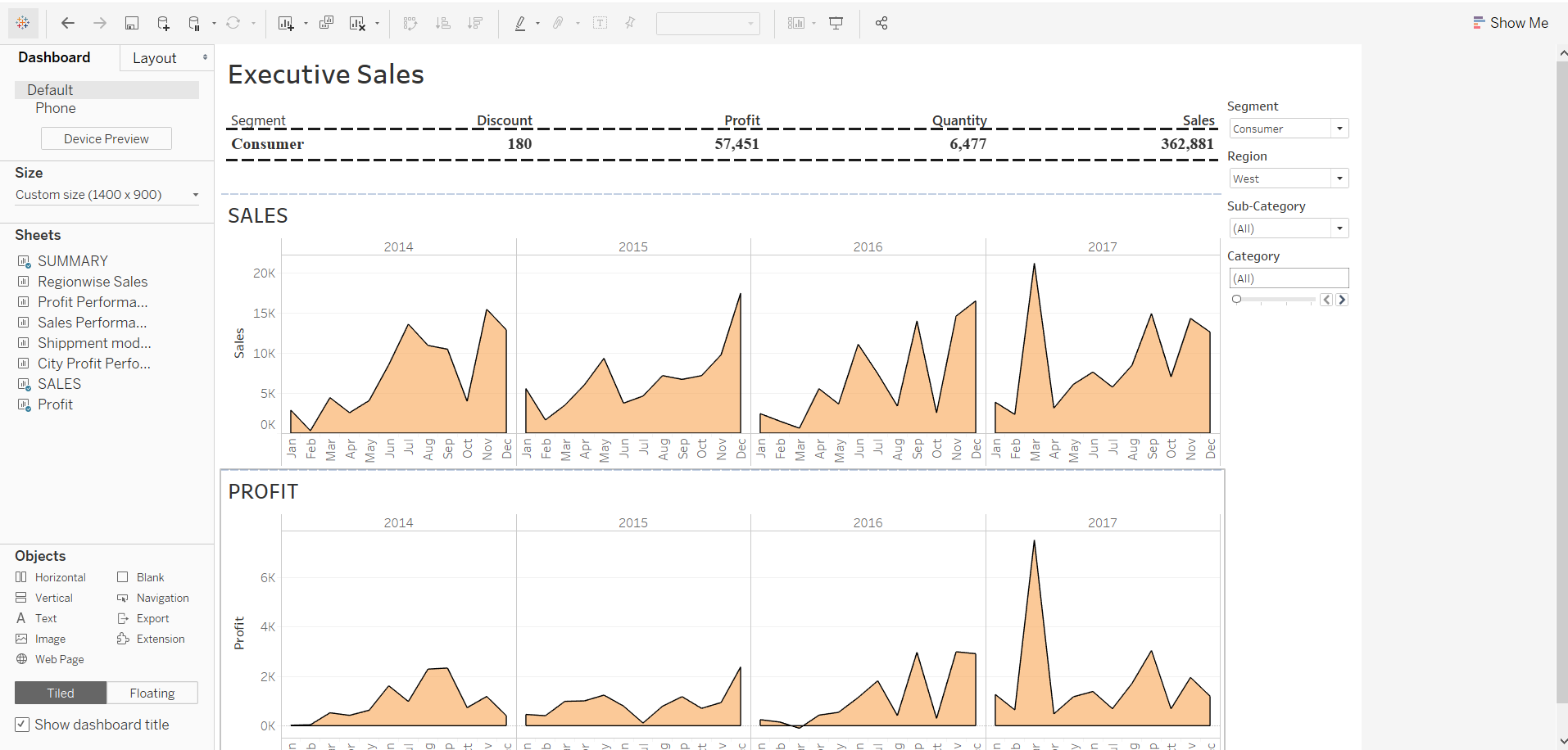
1. Shipment Mode trends

The most common shipping mode preferred is Standard Class as the price for shipping is low and most of the times is takes the same time as second class and sometimes it is also delivered on the same day.



1. Tableau Dashboard

We have shown the sales and profit of each year. We have provided filters so that we can see which segment had what profit and how much sales was obtained.



**ANALYSIS**

1. **Importing Packages:** We imported necessary packages like Pandas, NumPy, Matplotlib, etc. in order to perform analysis and present visualization.
2. **Pre-processing Data:** The loaded data has 999 rows and 21 columns. There are no null values in the dataset. We drew descriptive statistics on the data and then prepared the data for ‘Furniture’ Sub-type forecasting.

A screenshot of a cell phone

Description automatically generated

1. **Building Model:** 
   1. **Furniture Sales Forecast:** In the pre-processed data we drop all the columns except order date and sales. This was done to make the data in to a time series. Then the data was sorted by order date in an ascending order. Then the sorted time series was plotted in a line graph.

A picture containing shirt

Description automatically generated

Figure 1: Line plot of Furniture sales from Jan 2014 to Dec 2017

From the plotted data we can see that the **highest** sales were in the month of **Jan 2015** and the **least** sales made was in **February 2016**. Based on the trend in the graph we can say that the approximate average sales are 250.

We then decomposed the time series in 4 components namely, Trends, Seasonality, Residuals, Observed and plotted all 4 components.

A picture containing photo

Description automatically generated

Figure 2: Line plot of time series decomposition for Furniture Sales

We can see an overall **declining trend** in the sales of the furniture as the years go by. This is a result of people **renting** or **buying used furniture** instead of buying brand new furniture. This trend is also observed because of the DIY era where a lot of people prefer **making their own furniture** from scratch, hence not contributing much towards the sales of new furniture.

The seasonality shows that the furniture sales start **increasing** around **Thanksgiving, Christmas and New Year**, because people tend to redecorate houses and offices during these festive times which leads to a spike in the sales. These sales start **declining** soon after the first week of **Jan** as the **busiest month** kicks in, where not a lot of buying of furniture happens.

Next, we used ARIMA (1, 1, 1) x (1, 1, 0, 12) with an AIC of 263.9371084381265 as the best parameter for SARIMAX model. With this model, we generated the forecast for the future sales of furniture.

A picture containing shirt

Description automatically generated

Figure 3: Line plot of observed values of Furniture Sales with a forecast for year 2017

The one step-ahead forecast conditions on the training data set. Based on the forecasted sales value it can be concluded that the forecast and the observed values have quite a difference. This means that this is not the line of best fit. For the current dataset since the data is scattered widely around the regression line, we get a Mean Squared Error **(MSE) of 5911.26**

* 1. **Office Supplies Forecast:** We preprocess the original Data frame to extract only the order date and sales for office supplies. We then plot the sales from Jan 2014 to Dec 2017.

A picture containing monitor, screen, shirt, video

Description automatically generated

Figure 4: Line plot of Office Supplies sales from Jan 2014 to Dec 2017

From the plotted sales we can see that the highest sales were in Jan 2017 and least sales made were in Feb 2014. The overall average sales are 140. To further understand the trend in the sales of office supplies we decompose the time series.

A close up of a logo

Description automatically generated

Figure 5: Line plot of time series decomposition for Office Supplies Sales

From the decomposition we can see that there is a **valley** like **trend** in the sales of office supplies. It is **higher** towards the **end of year 2014**, **falls** down in the year **2015** through mid 2016 and starts **increasing** again from mid 2016 towards the end of **2017**.

The seasonality shows that there are **more** sales in the months of **November, December and January** which starts **decline** from the month of **February**.

Next, we used ARIMA (1, 1, 1) x (1, 1, 0, 12) with an AIC of 246.479666199784as the best parameter for SARIMAX model. With this model, we generated the forecast for the future sales of Office Supplies.

**A close up of text on a white background

Description automatically generated**

Figure 6: Line plot of observed values of Office Supplies Sales with a forecast for year 2017

With the one step-ahead forecast we will check the performance of the model. Based on the forecasted sales value, we can conclude that the forecast and the observed values have **some difference**, but it is not huge. This means that this is **not** the line of **best fit**. For the current dataset since the data is scattered widely around the regression line, we get a Mean Squared Error **(MSE) of 3393.63.**

1. **Analysis:** From the two forecasts we generated we can say that both the sales show a similar seasonality of increase in sales in the months of Nov, Dec & Jan and a Steep fall in the sales in the month of Feb. Apart from this, the model is better fitted for the office supplies sales than it is for the furniture sales.

**CONCLUSION**

From the forecasts we’ve seen for the two sales category we can say that the data is spread wide around the regression line making it difficult to predict accurate sales values. This can be improved with help of model boosting and further cleaning the data. Seasonal ARIMA forecasting uses time series data and this helps in avoiding problems that are associated with multivariate models. Seasonal ARIMA model was very useful in getting proper analysis and in getting better results. We could have improved the MSE value to a bit low by getting more data and achieved much better results.

**REFERENCES**

* Prabhakaran, S. (2020, April 28). ARIMA Model – Complete Guide to Time Series Forecasting in Python. Retrieved May 11, 2020, from[**https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/**](https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/)
* (n.d.). Retrieved May 15, 2020, from <https://community.tableau.com/docs/DOC-1236>
* Time Series Talk: ARIMA Model [**https://www.youtube.com/watch?v=3UmyHed0iYE**](https://www.youtube.com/watch?v=3UmyHed0iYE)