CIS 4560 Tutorial

Authors: Jesus Barba, Lawrence Cho, Heriberto Corona, Luis Rodriguez, Freddy Romero
Instructor: Jongwook Woo
Date: 12/18/2022

Lab Tutorial

Jesus Barba
Lawrence Cho
Heriberto Corona
Luis Rodriguez
Freddy Romero

eCommerce Behavior Data Analysis using Spark

Objective

Ecommerce Behavior data is one of the most rapidly grown areas for Big Data adoption. In this project, you need to find a data set(2GB>), use hadoop, hive, and present the insights and the data analysis. Also include visualization.

Introduction

In this sample you will use HiveQL that analyzes eCommerce Behavior to get an insight into how customers are purchasing items.
In this tutorial you will learn how to use hadoop cluster to:

- Download and upload a csv file
- Create hive tables to query those ecommerce data
- Use Tableau to connect to Hadoop Cluster to retrieve the analyzed data
- User Tableau 3D map for 3D Visualization

Prerequisite

Everything you need to go through the queries and scripts is already provisioned with the cluster. To export the analyzed data to tableau , you must meet the following requirements:

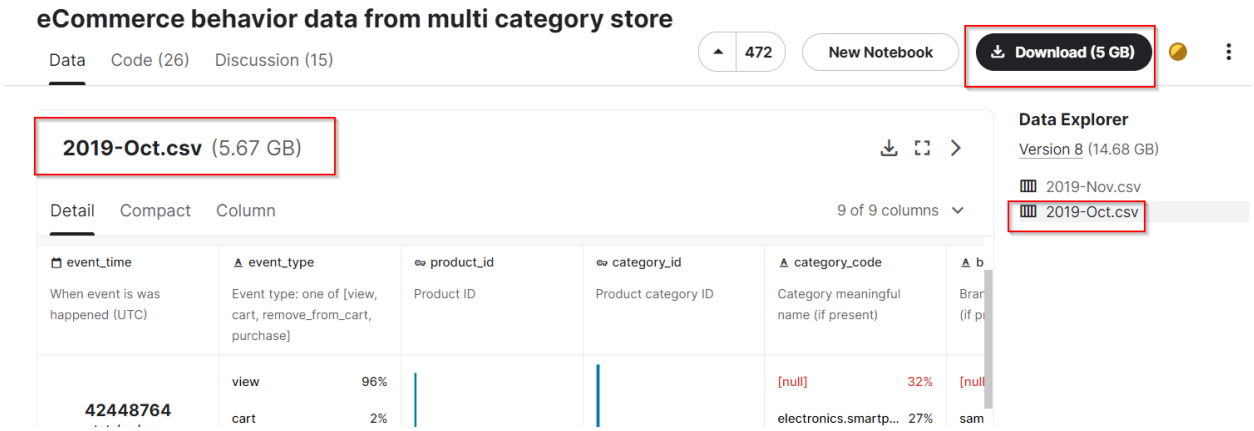
- You must have an account at Tableau site and tableau installed

Step 1: Downloading Data

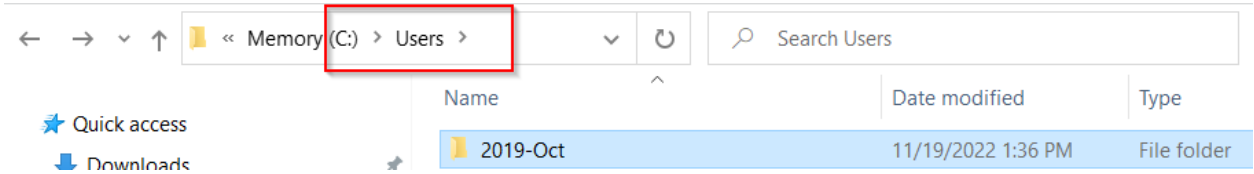
HOW TO DOWNLOAD DATASET

1. **Download dataset from Kaggle**
   Link:
   https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store

   Note: Make sure to click on 2019-Oct.csv (Not 2019-Nov.csv)

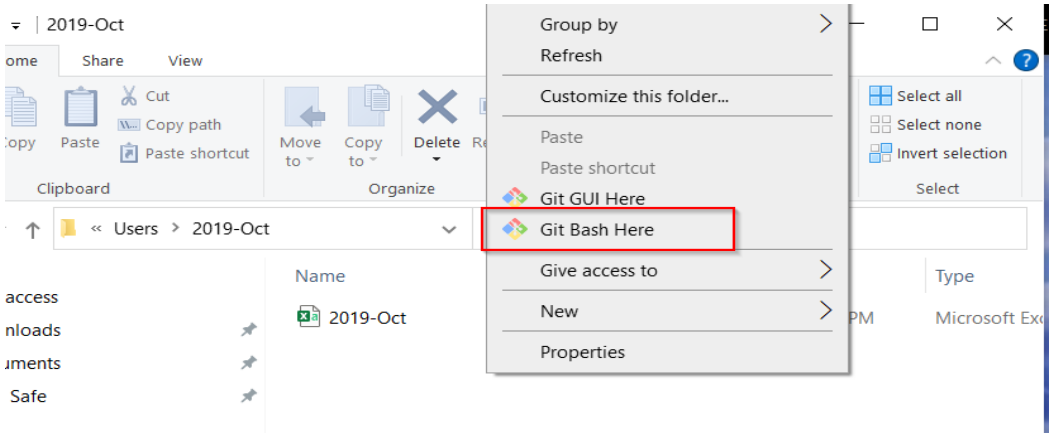


2. **Make a file on your computer:** C:\ > users > 2019-Oct (Create a folder with this name)

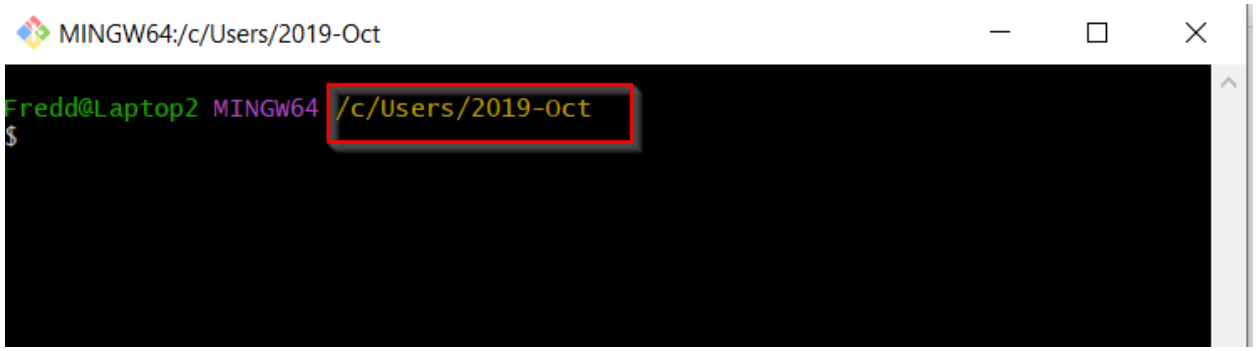


3. **Click and drag the dataset file to the folder.**

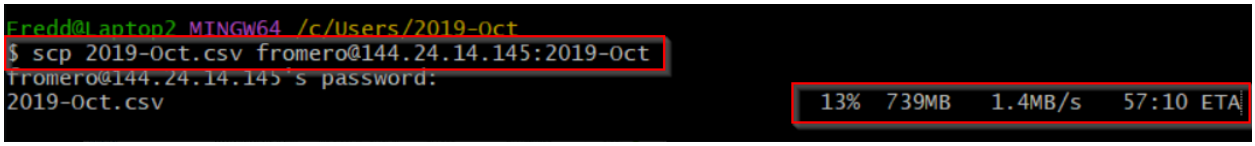
4. **Right click the file and click “Git Bash Here”**



- By doing so, we can open Bash directly in the folders directory
- Alternatively, you can use the cd commands



5. **On Git Bash, type “scp 2019-Oct.csv yourusername@144.24.14.145:2019-Oct**
   Note: Do not use “yourusername”, replace “yourusername” with your database username



- The data is now downloading. This process will take around 1 hour depending on your computer performance.

Step 2

CREATING DIRECTORY FOR DATASET

Open a new Git Bash terminal

Insert these commands

1. ssh yourusername@144.24.14.145

   Note: Replace “yourusername” with your database username

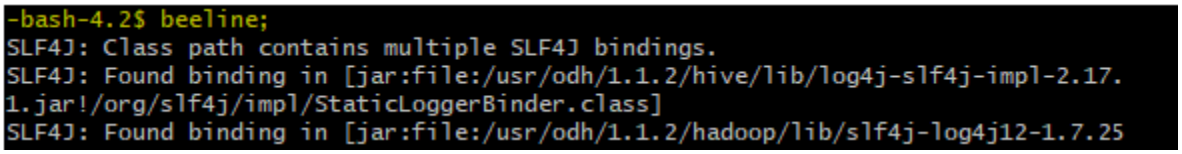
2. hdfs dfs -mkdir ecommerce
3. hdfs dfs -ls
4. hdfs dfs -put 2019-Oct ecommerce
5. hdfs dfs -ls ecommerce

BEELINE

Open a new Git Bash terminal

Insert these commands

1. ssh yourusername@144.24.14.145
2. Enter into beeline



```
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.292 seconds)
0: jdbc:hive2://bigdaiwn0.sub02180640120.trai>
```

3. use yourdatabase;
4. The command below is used to create the table.

```
DROP TABLE IF EXISTS ecommerce;

CREATE EXTERNAL TABLE IF NOT EXISTS ecommerce(
event_time STRING,
event_type STRING
product_id INT,
category_id BIGINT,
category_code STRING,
brand STRING,
price DECIMAL,
user_id BIGINT,
user_session STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/USE OWN DATABASE/ecommerce'
TBLPROPERTIES("skip.header.line.count"="1");
```

DATA ANALYSIS USING HIVEQL

Now you can run the following HiveQLs. Note: Due to the large dataset, each process may take up to 1-2 minutes depending on your computer performance.

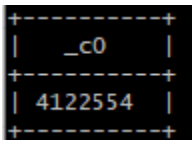
```
SELECT product_id,category_code,brand,price,user_id FROM ecommerce LIMIT 10;
```

This is the result that you should have.

product_id	category_code	brand	price	user_id
44600062		shiseido	36	541312140
3900821	appliances.environment.water_heater	aqua	33	554748717
17200506	furniture.living_room.sofa		543	519107250
1307067	computers.notebook	lenovo	252	550050854
1004237	electronics.smartphone	apple	1082	535871217
1480613	computers.desktop	pulser	909	512742880
17300353		creed	381	555447699
31500053		luminarc	41	550978835
28719074	apparel.shoes.keds	baden	103	520571932
1004545	electronics.smartphone	huawei	566	537918940

```
SELECT COUNT(brand) FROM ecommerce WHERE brand='apple';
```

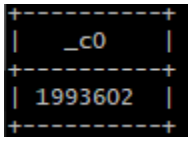
This is the result that you should have.



You may also try…

```
SELECT SUM(price) FROM ecommerce WHERE user_id=563459593;
```

This is the result that you should have.



Step 3

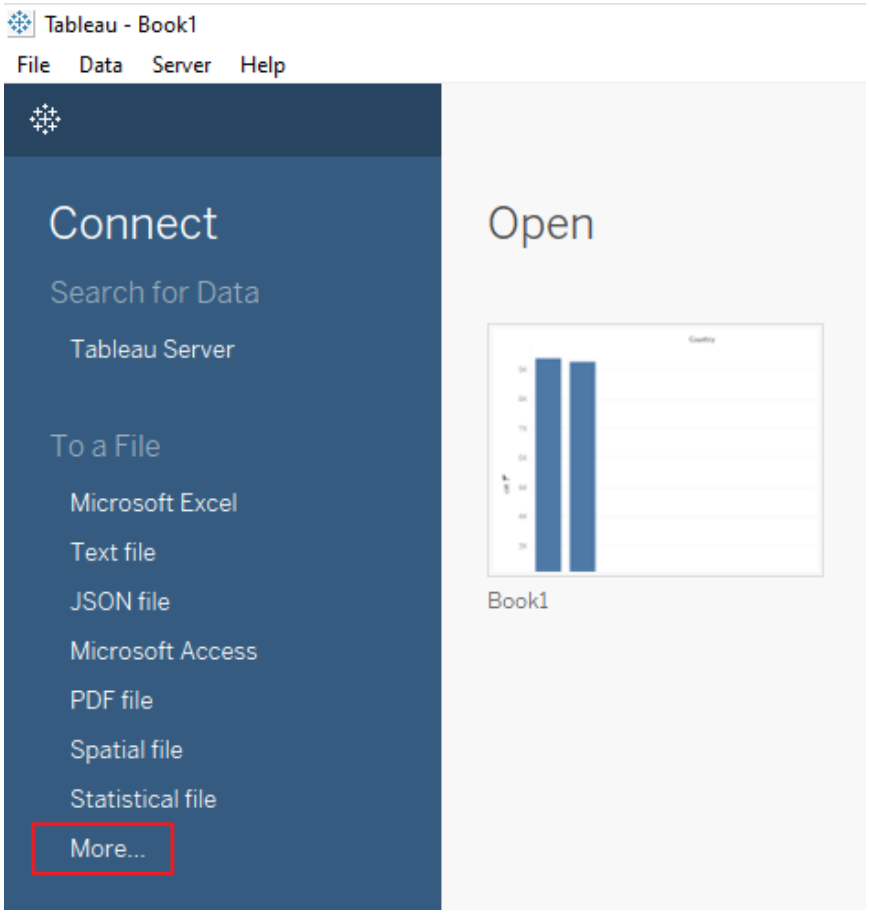
Using Tableau for Data Visualization

If you do not have Tableau, the link below is the free desktop version for students.

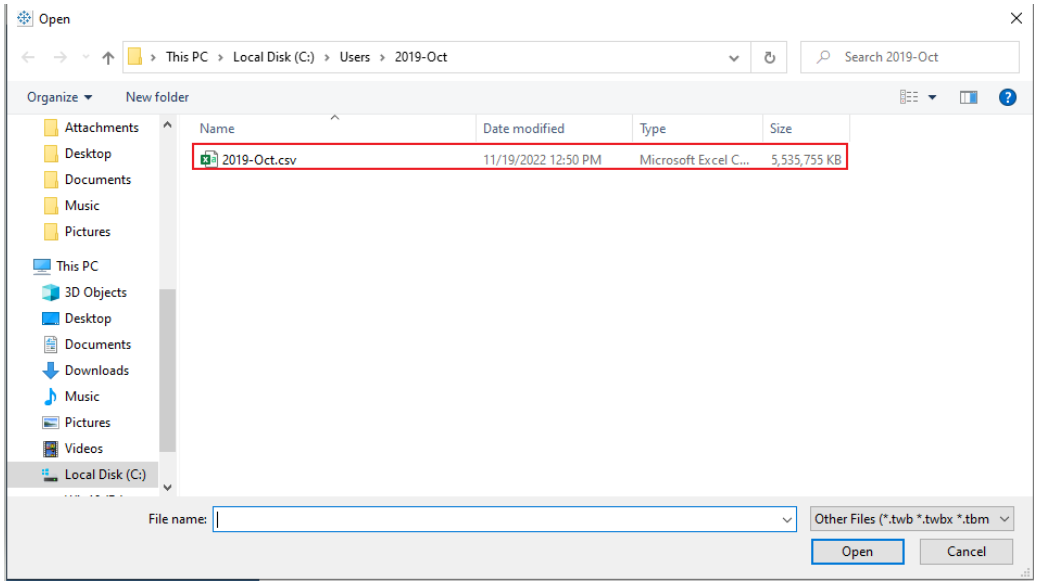
https://www.tableau.com/academic/students

Once you have Tableau set up, follow the directions below.

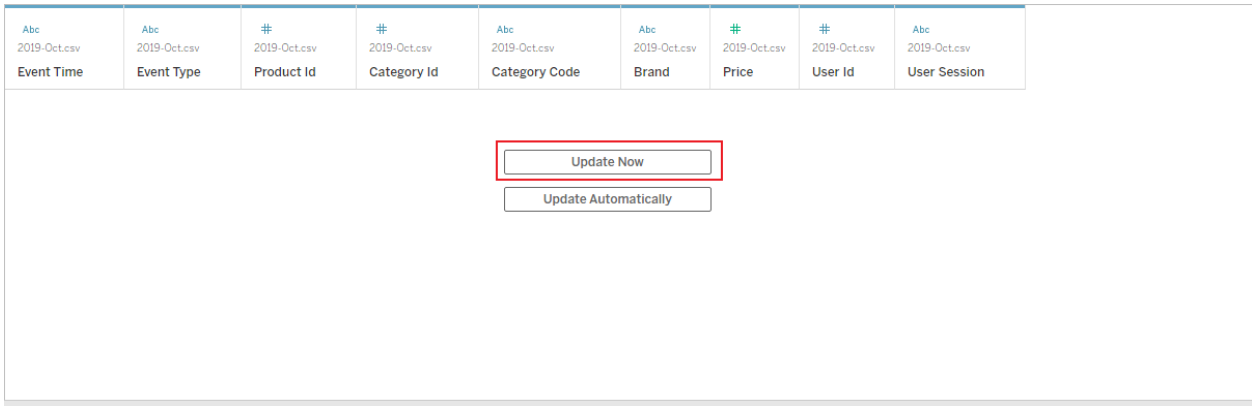
1. Open Tableau on your computer. You need to select **More…** to select the file you have downloaded in step 1.



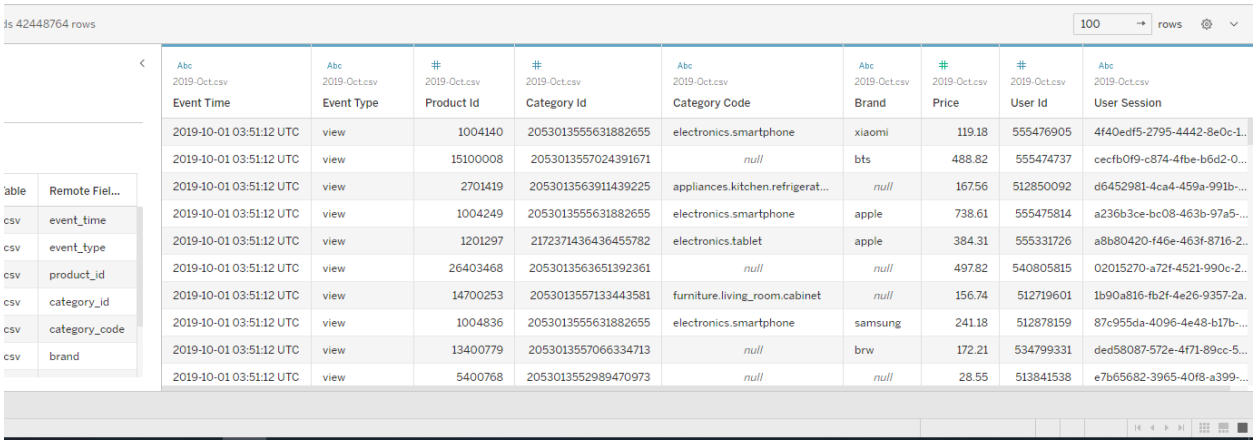
2. Select the file that you have downloaded.



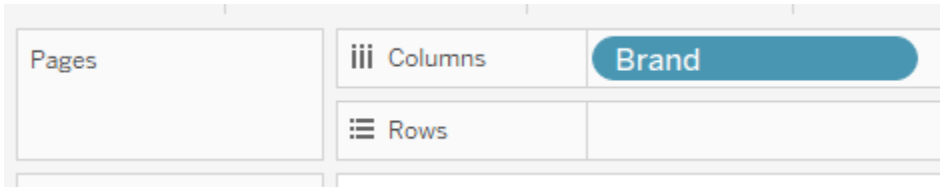
3. Click on **Update Now** located on Tableau



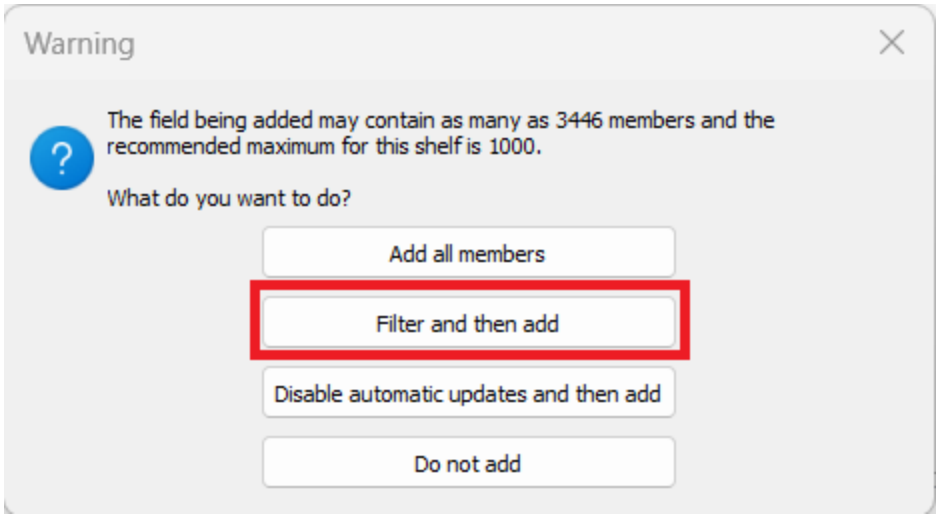
4. You should now see that all the data has loaded into the application.



5. Let’s start off with a basic visualization, on Tableau, showing the SUM of Price per Brand. We will start off by dragging “Brand” dimension under “Tables” and drag it over to the “Columns” section.

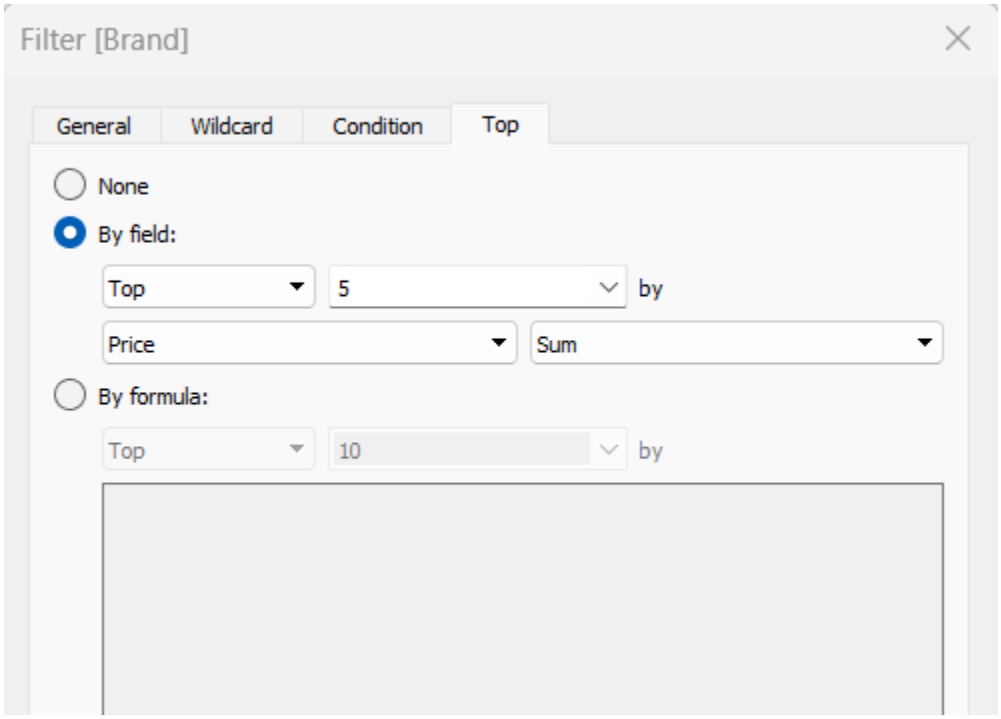


You might encounter the following warning:

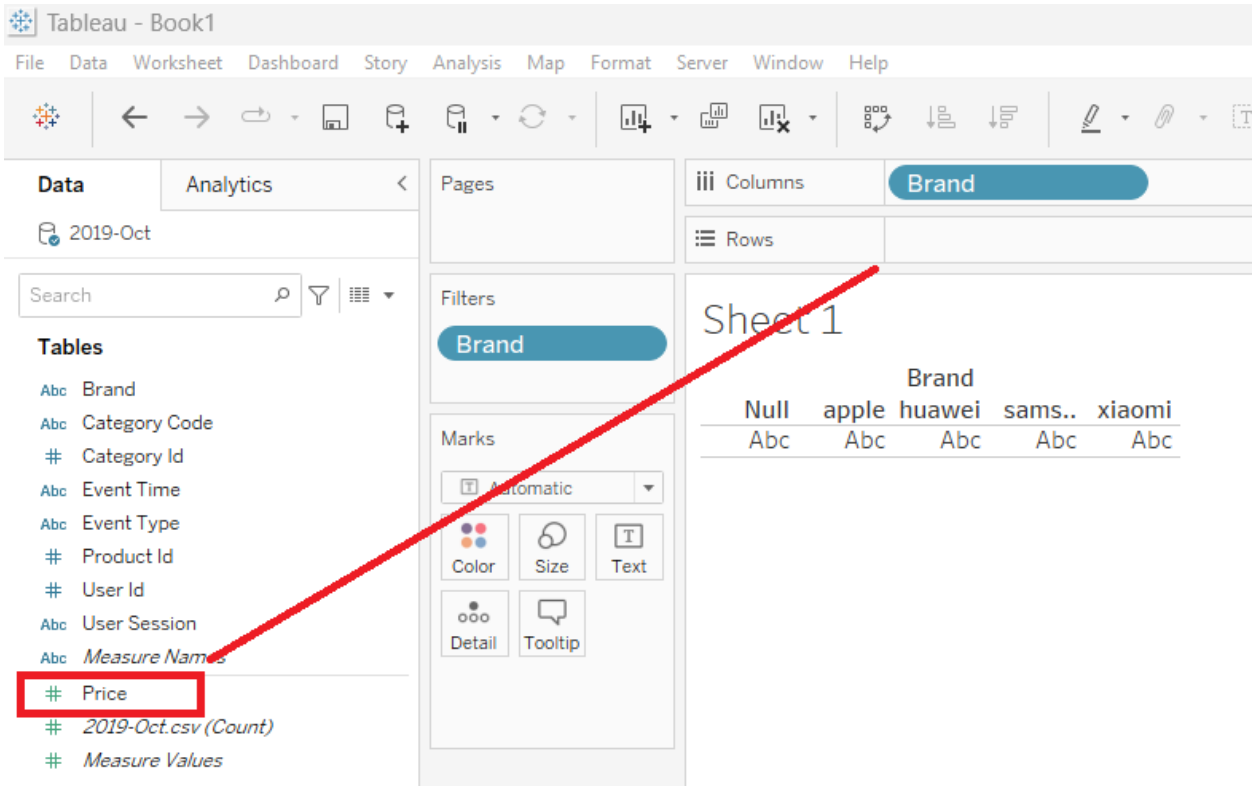


We're going to select "Filter and then add" option.

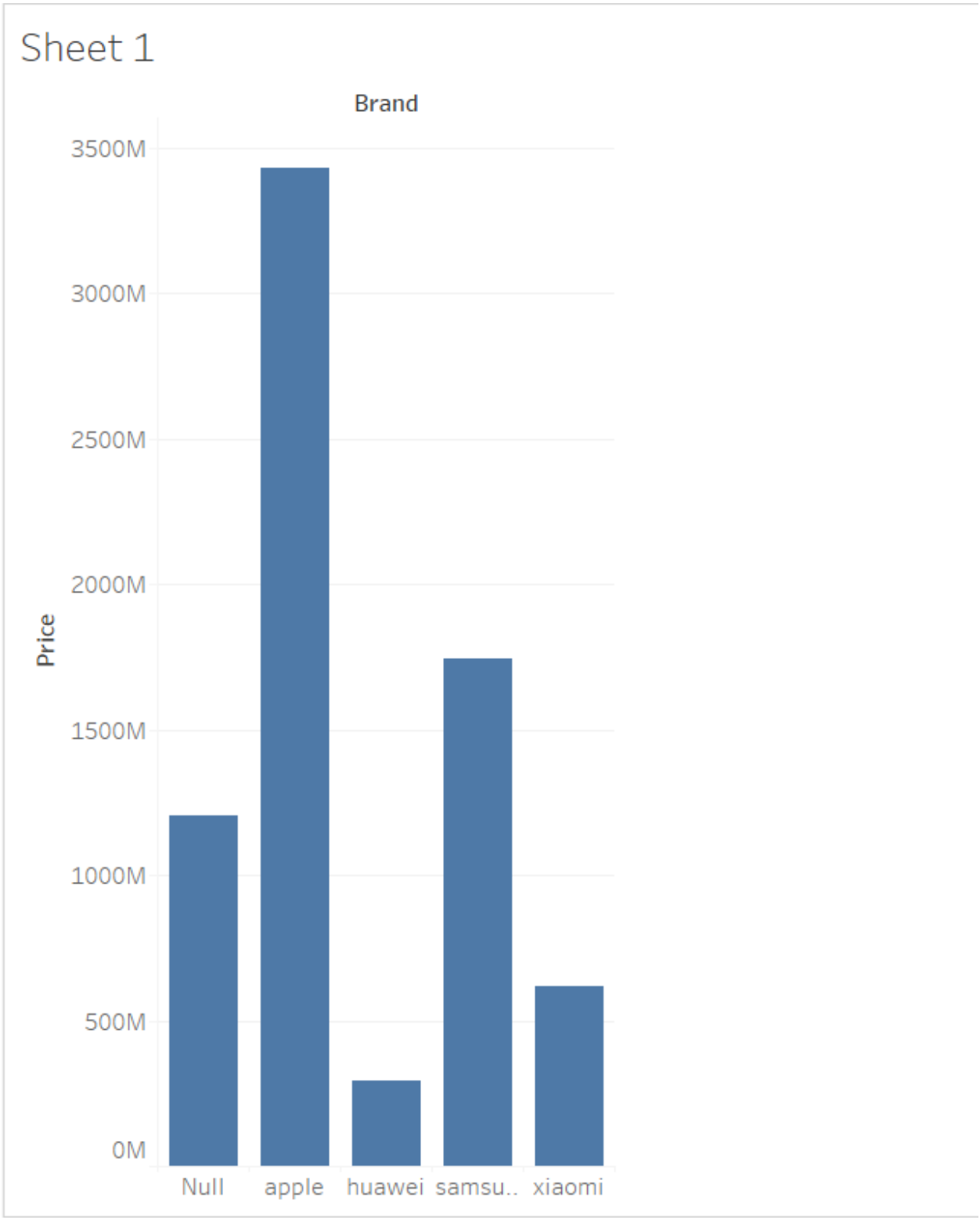
Next, we will select the Top option, By field, and filter the view for it to only show up the Top 5 brands based on Sum of Price.



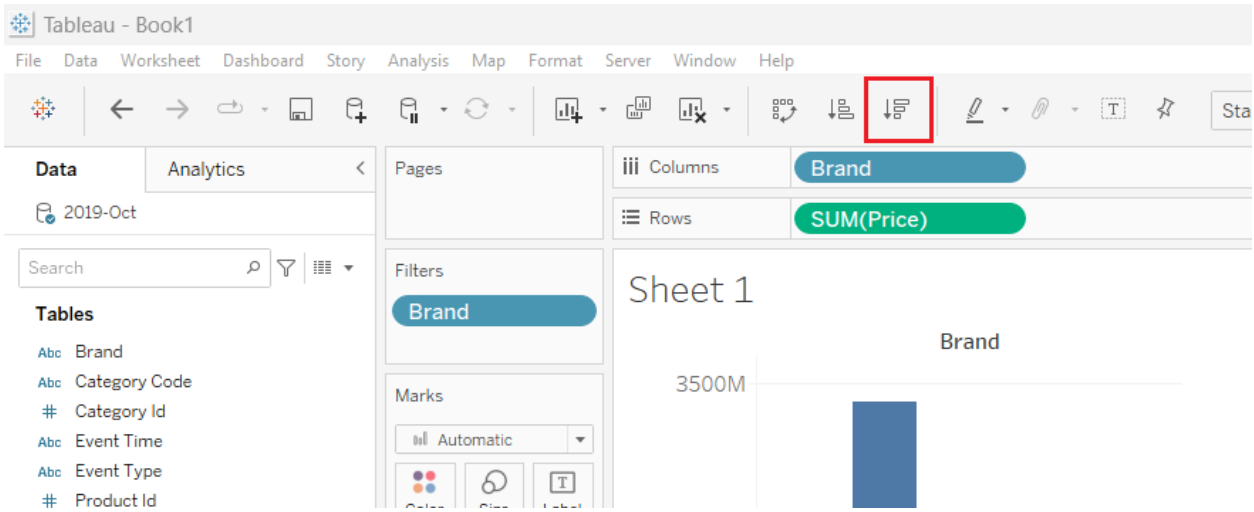
Next, we will drag the "Price" measure to our "Rows" section.



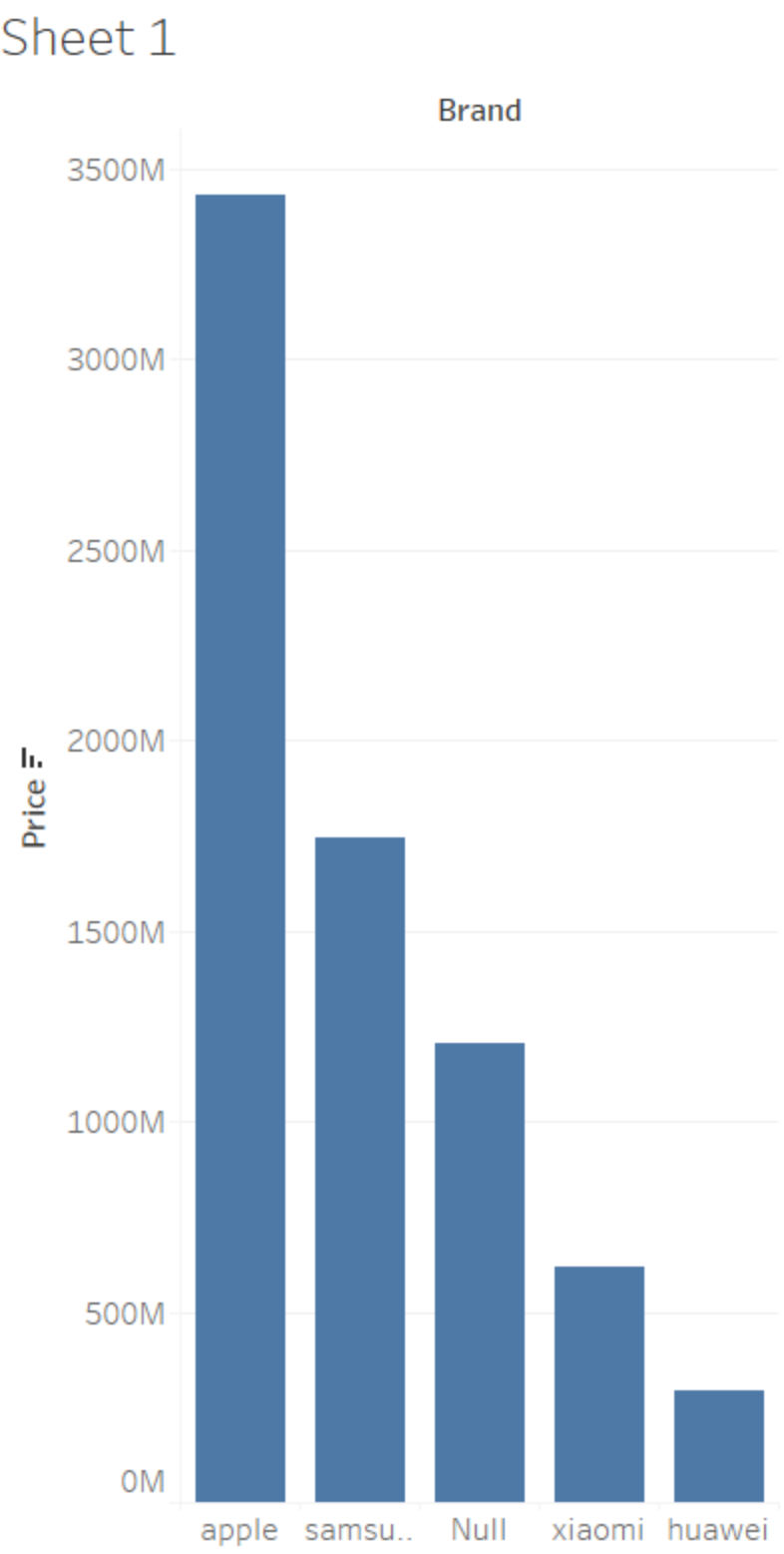
The result should be the following:



At this point, we have our desired results, but it might not be clear to the viewer what exactly is going on. We should organize the chart, in descending order from highest to lowest. We can easily do this, with a press of a button.

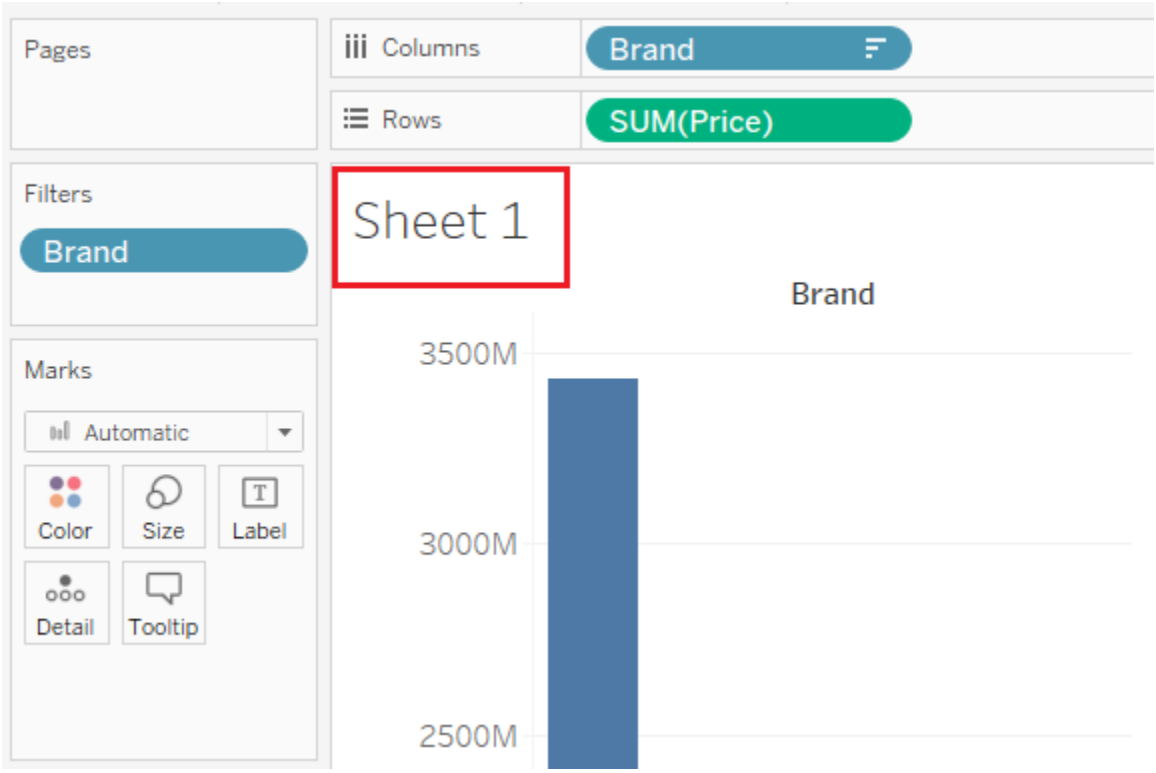


Now are results are starting to look better:

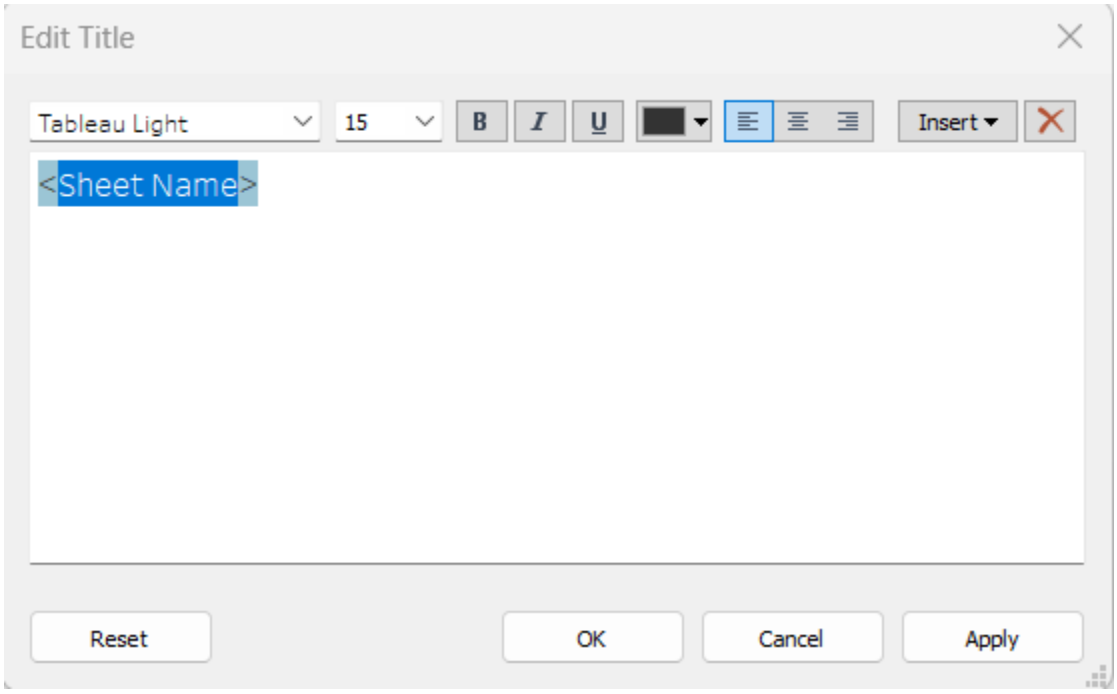


We can take things a step further to make this chart easier to view and read. Let's add our title and a caption to explain what is going on…

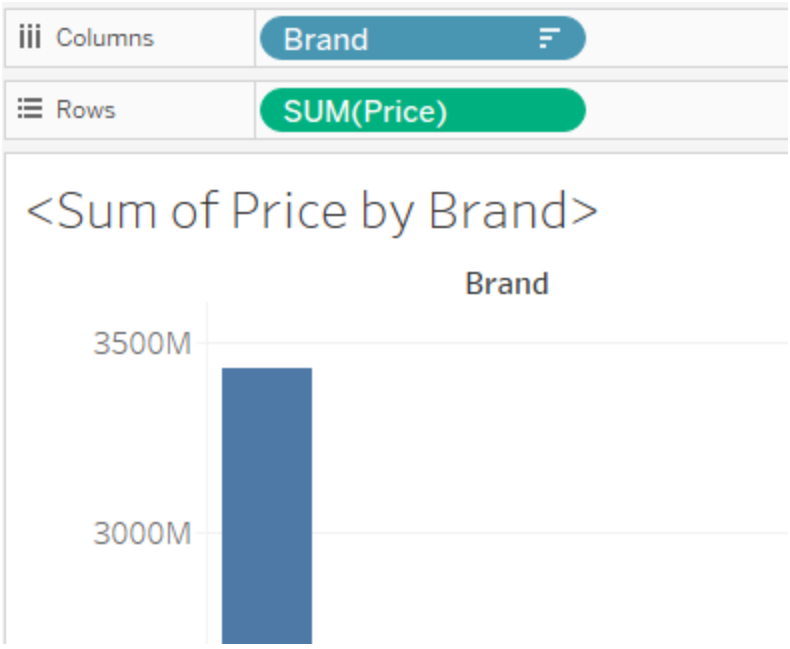
To add a title to our chart, we will double click where it says “Sheet 1”



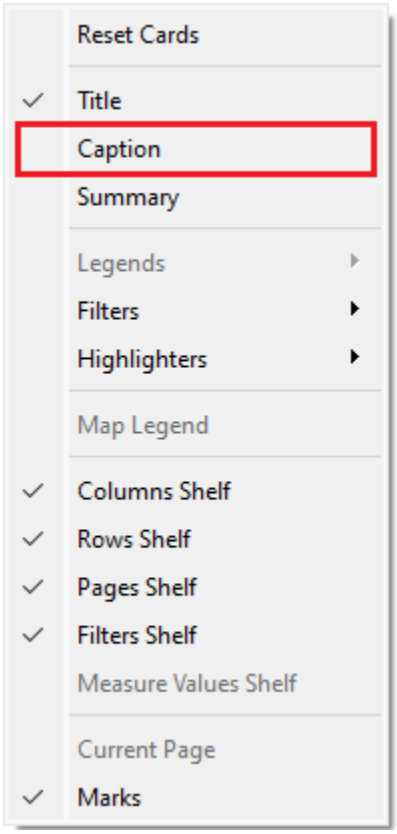
We are greeted with the following window, let’s highlight the inside of the two symbols and add our desired title.



Results:

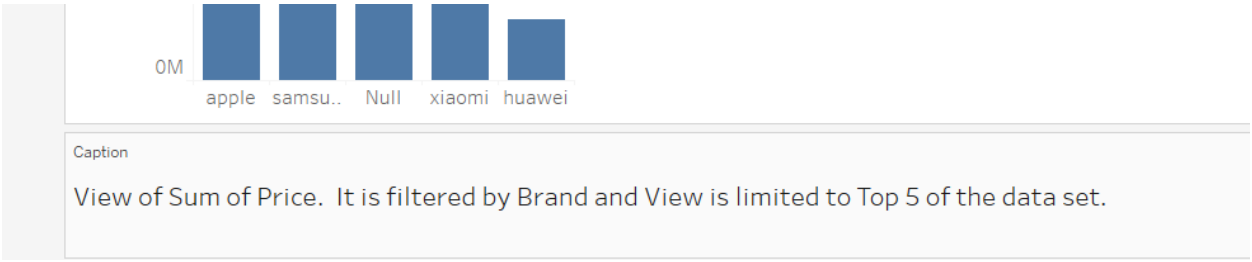


Now let's right click anywhere and select "Caption".

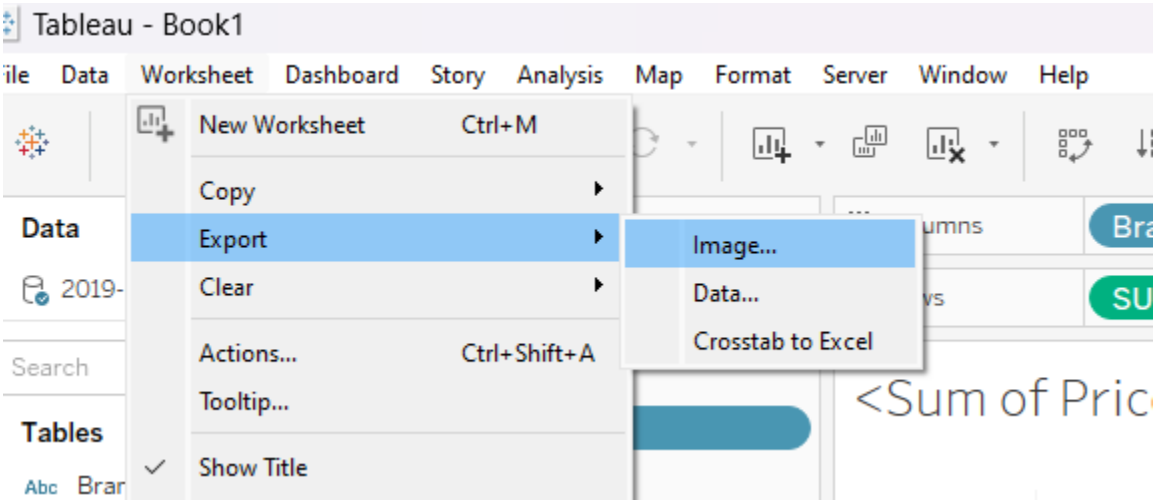


Let's add to the caption to explain to the viewer/reader what the graph is telling them.

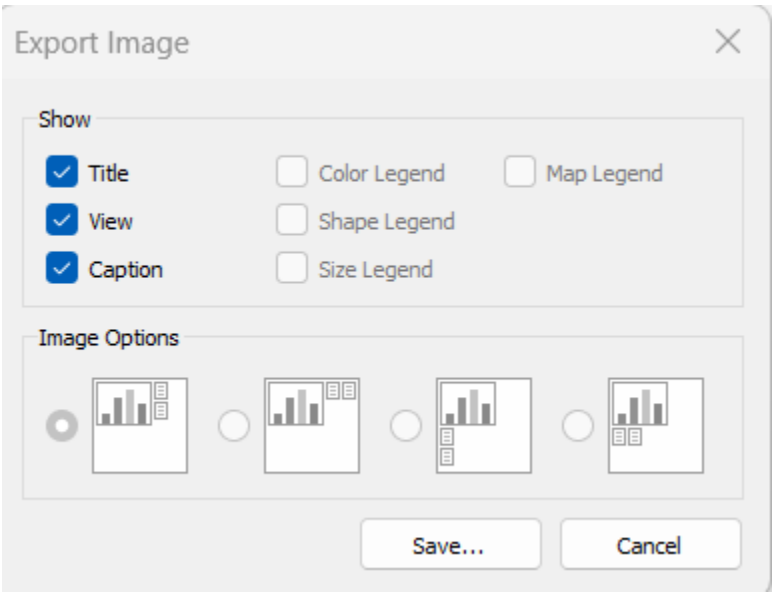
Result:



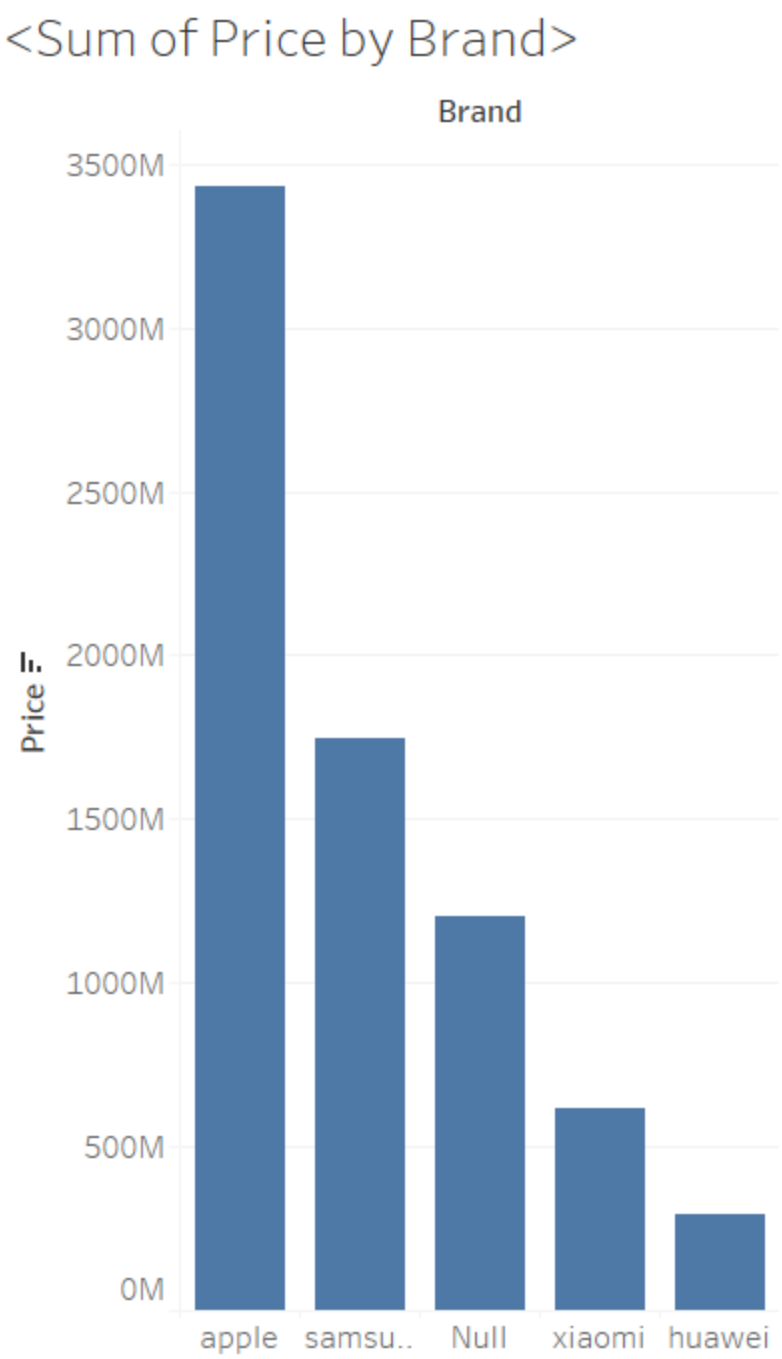
With that, our Sheet is complete. If we wanted to Export the image with everything we altered, we will go to Worksheet on the top left > Export > Image…



Select Save…



And the final results:



View of Sum of Price. It is filtered by Brand and View is limited to Top 5 of the data set.

With this in mind, you should feel free to move variables and measures to either Columns and Rows, Tableau has a variety of features that allow you to create several graphs to visualize data.