**Northumbria University, Newcastle**

**Faculty of Engineering and Environment**

**Department of Computer and Information Sciences**

**Big Data Analytics**

**Submitted By**

|  |  |
| --- | --- |
| **Name:** | **Busra Ecem Sakar** |
| **ID:** | **w22065255** |
| **Programme:** | **MSc Big Data and Data Science Technology** |
| **Semester:** | **2** |
| **Tutor:** | **Nitsa Herzog** |
| **Date:** | **15 May 2023** |
| **Word Count:** | **4.860 (include tables, figures names)** |

# **Table of Contents**

[**Table of Contents** 2](#_Toc134836227)

[**Table of Tables** 3](#_Toc134836228)

[**Table of Figures** 3](#_Toc134836229)

[**1.** **Big Data Analytics (Python)** 5](#_Toc134836230)

[**1.1.** **Problem Domain, Data Description, and Research Question** 5](#_Toc134836231)

[**1.1.1.** **Problem Domain** 5](#_Toc134836232)

[**1.1.2.** **Data Description** 5](#_Toc134836233)

[**1.1.3.** **Research Question** 8](#_Toc134836234)

[**1.2.** **Solution Exploration** 9](#_Toc134836235)

[**1.3.** **Solution Development** 11](#_Toc134836236)

[**1.3.1.** **Data Cleaning and Preparation** 12](#_Toc134836237)

[**1.3.2.** **Frequency Analysis and Normality Test** 16](#_Toc134836238)

[**1.3.3.** **Statistical Analysis; to solve research questions** 26](#_Toc134836239)

[**1.4 Evaluation and Future Development** 35](#_Toc134836240)

[**2.** **Business Intelligence (Tableau)** 37](#_Toc134836241)

[**2.1** **Task 1** 38](#_Toc134836242)

[**2.2** **Task 2** 41](#_Toc134836243)

[**2.3** **Task 3** 45](#_Toc134836244)

[**2.4** **Task 4** 46](#_Toc134836245)

[**2.5** **Task 5** 47](#_Toc134836246)

[**3.** **References** 49](#_Toc134836247)

# **Table of Tables**

[Table 1 Variable's Descriptions 6](#_Toc134836109)

[Table 2 Data type and descriptions 8](#_Toc134836110)

[Table 3 Python Libraries used in the analysis 11](#_Toc134836111)

# **Table of Figures**

[Figure 1 Creating Google Colaboraty 6](#_Toc134836050)

[Figure 2 Uploading Dataset and Calculate total number of rows and columns 7](#_Toc134836051)

[Figure 3 Concise Summary of the Dataframe 7](#_Toc134836052)

[Figure 4 Summary of big data-related cloud service models (Fan and Rana, 2021) 9](#_Toc134836053)

[Figure 5 Analysis of missing values in the data 12](#_Toc134836054)

[Figure 6 Percentage of missing values in total 12](#_Toc134836055)

[Figure 7 Removing Missing Values from Dataset 12](#_Toc134836056)

[Figure 8 Total Number of rows & columns, after removing Missing Values from the Dataset 13](#_Toc134836057)

[Figure 9 Count number of zeros in all columns of Dataframe 13](#_Toc134836058)

[Figure 10 Descriptive Statistics For Numeric Variables 14](#_Toc134836059)

[Figure 11 Descriptive Statistics without demographics.age<=0, new database 15](#_Toc134836060)

[Figure 12 Duplicated value analysis 15](#_Toc134836061)

[Figure 13 Frequency analysis of "year" variable 16](#_Toc134836062)

[Figure 14 Frequency analysis of "gender" variable 17](#_Toc134836063)

[Figure 15 Frequency analysis of "region" variable 17](#_Toc134836064)

[Figure 16 Frequency analysis of "wealth type" variable 18](#_Toc134836065)

[Figure 17 Frequency analysis of "wealth how category" variable 19](#_Toc134836066)

[Figure 18 Frequency analysis of "wealth how inherited" variable 20](#_Toc134836067)

[Figure 19 Frequency analysis of "wealth how idustry" variable 21](#_Toc134836068)

[Figure 20 Normality test for "rank" variable 23](#_Toc134836069)

[Figure 21 Normality test for "demographics.age" variable 23](#_Toc134836070)

[Figure 22 Normality test for "wealth.worth in billions" variable 25](#_Toc134836071)

[Figure 23 Python code for Research Question 1, step 1 26](#_Toc134836072)

[Figure 24 Kruscal Wallis test's Output for Research Question 1 27](#_Toc134836073)

[Figure 25 Perform Dunn test to analyse pairwise comparison between two groups, for Research Question 1 27](#_Toc134836074)

[Figure 26 Dunn Test’s Output, Research Question 1 28](#_Toc134836075)

[Figure 27 Python code for Research Question 2 29](#_Toc134836076)

[Figure 28 Kruscal Wallis test's Output for Research Question 2 29](#_Toc134836077)

[Figure 29 Perform Dunn test to analyse pairwise comparison between two groups, for Research Question 2 30](#_Toc134836078)

[Figure 30 Dunn Test’s Output, Research Question 2 30](#_Toc134836079)

[Figure 31 Python code for Research Question 3 32](#_Toc134836080)

[Figure 32 Chi-squared test's Output for Research Question 3 32](#_Toc134836081)

[Figure 33 Python code for Research Question 4 33](#_Toc134836082)

[Figure 34 Chi-squared test's Output for Research Question 4 33](#_Toc134836083)

[Figure 35 Python code for Research Question 5 34](#_Toc134836084)

[Figure 36 Spearman Correlation test's Output for Research Question 5 34](#_Toc134836085)

[Figure 37 Import data to Tableu 37](#_Toc134836086)

[Figure 38 Creating Parameter for top n 38](#_Toc134836087)

[Figure 39 Creating a calculated field 38](#_Toc134836088)

[Figure 40 Filtering top 10 sales product 39](#_Toc134836089)

[Figure 41 Duplicating sheets 39](#_Toc134836090)

[Figure 42 Filtering 10 low-selling product 40](#_Toc134836091)

[Figure 43 Top 10 selling product 40](#_Toc134836092)

[Figure 44 Top 10 low-selling product 40](#_Toc134836093)

[Figure 45 Split Columns to create a field\_1 41](#_Toc134836094)

[Figure 46 Split Columns to create a field\_2 42](#_Toc134836095)

[Figure 47 Rename fields 42](#_Toc134836096)

[Figure 48 Calculating Sales 43](#_Toc134836097)

[Figure 49 New databased with Sales and City Fields 43](#_Toc134836098)

[Figure 50 Creating Parameters for AVG and SUM of sales 43](#_Toc134836099)

[Figure 51 Create calculated fields as Sales Statistics 44](#_Toc134836100)

[Figure 52 Total/Average sales made in each city 44](#_Toc134836101)

[Figure 53 Step 1 for Task 3 45](#_Toc134836102)

[Figure 54 the stores’ performance on a weekly basis 45](#_Toc134836103)

[Figure 55 Create calculated field for Warranty End Date 46](#_Toc134836104)

[Figure 56 Warranty End Date 46](#_Toc134836105)

[Figure 57 Adding Actions to link the sheets together 47](#_Toc134836106)

[Figure 58 Choosing the sheets in the dashboard that want to link to 47](#_Toc134836107)

[Figure 59 Sales Details Dashboard 48](#_Toc134836108)

# **Big Data Analytics (Python)**

Python has grown in popularity as a big data research language because it has adaptability and a rich ecosystem of data analysis tools. In this research Pyhton library used for statistical analysis.

## **Problem Domain, Data Description, and Research Question**

### **Problem Domain**

The year 2023 is off to a bad start for the world's richest people economically. According to the list of the world's richest people published every year in Forbest 2023, the number of 10-digit wealth holders, which was 2,668 last year, decreased to 2,640 this year. Hong Kong and Macau are second with 562 billionaires valued at $2 trillion, followed by India with 169 billionaires valued at $675 billion (Peterson-Withorn, 2023).

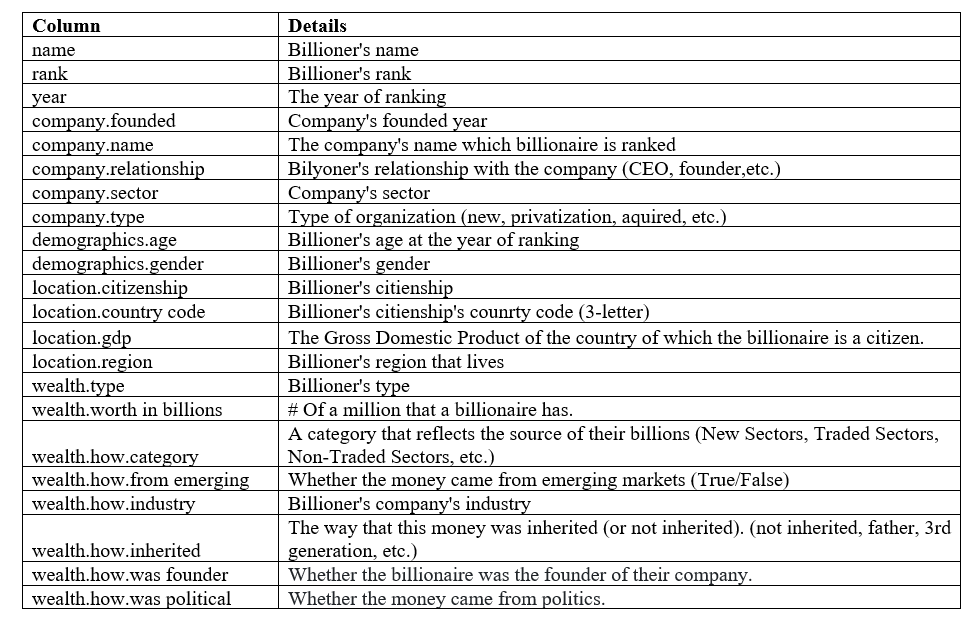
This research has a dataset that addresses statistics on the world's billionaires' net worth, age, gender, place of origin, and source of wealth in 1996, 2001, and 2014. The dataset consists of 2614 observations, each containing 22 variables, name is “***Billionaires.csv***”.

This dataset's usage field is most likely in economics, business, and finance. This information could be used by researchers to analyse patterns in wealth distribution and uncover differences between billionaires based on age, gender, geography, sector, and so on. With the results, businesses can use overall knowledge of global variances across billionaires to determine what has the most influence, or investors can make an overall investment decision.

### **Data Description**

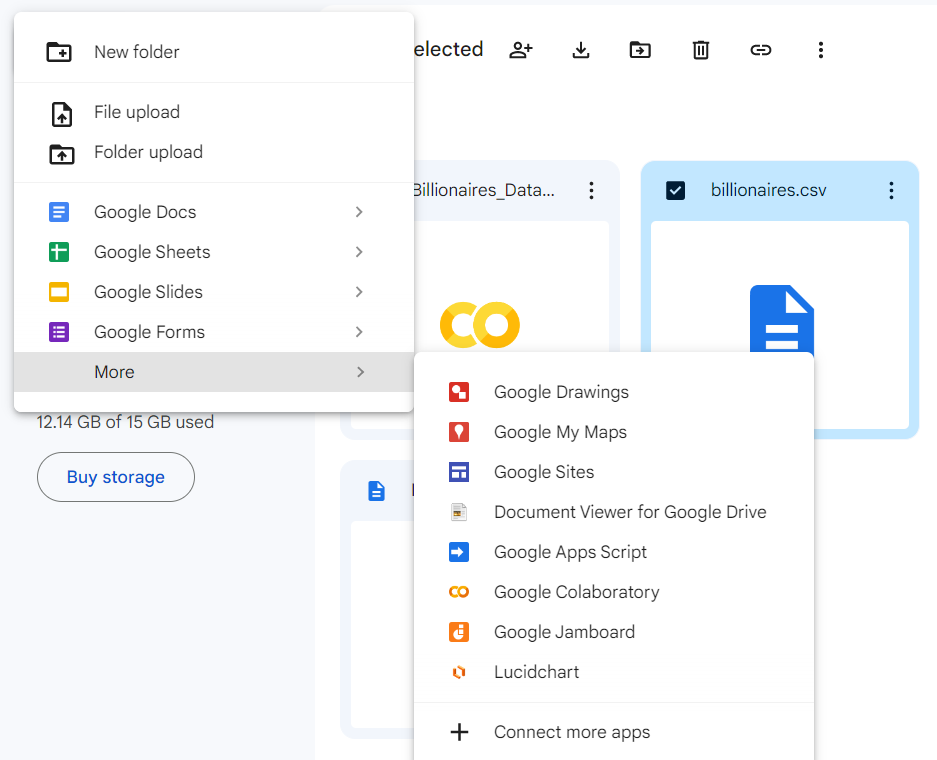
Details of the variables in the “Billionaires.csv” dataset are shown in Table1. Accordingly, this dataset contains both the data of the billionaires and the details of the companies that are billionaires. Based on these data, hypotheses of the research questions will be formed, and then appropriate statistical methods will be applied to answer these hypotheses.

**Table 1 Variable's Descriptions**

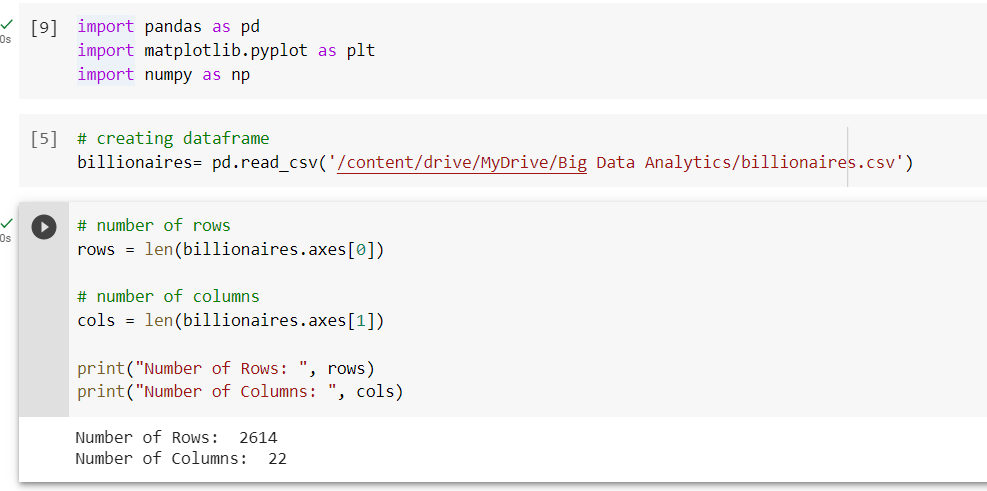


After uploading “***Billionaires.csv***” dataset to Google Drive, Colaboraty was created via Google Drive (Figure 1). The Colaboratory (Colab) product from Google Research is especially helpful for machine learning, data analysis, and education because it enables users to develop and run Python code through their browsers. (Google, n.d)

**Figure 1 Creating Google Colaboraty**



**Figure 2 Uploading Dataset and Calculate total number of rows and columns**

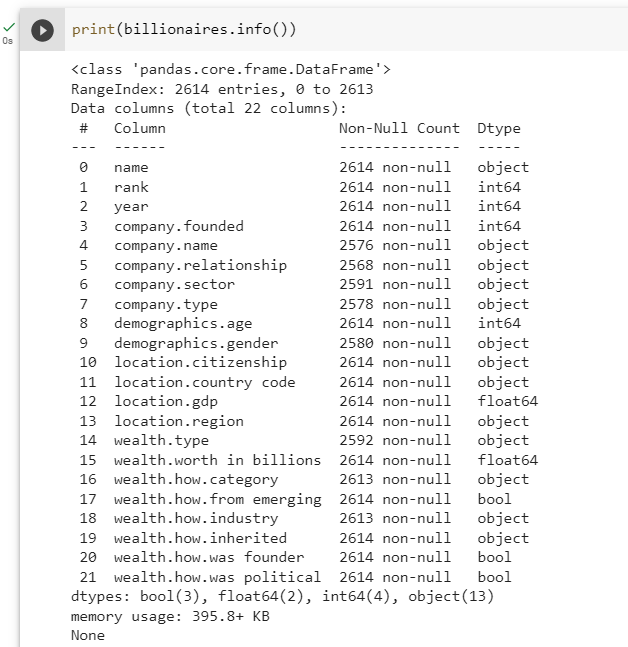


After uploading the “***Billionaires.csv***” dataset to Google Drive, the python code in figure 2 were run to get the total number of rows and columns of the dataset and to give the following output. Dataset has;

Number of Rows: 2614

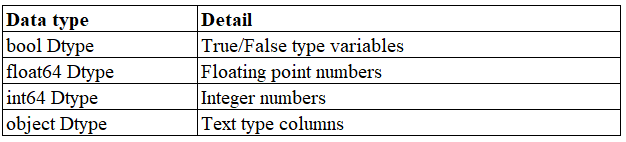
Number of Columns: 22

**Figure 3 Concise Summary of the Dataframe**



The Python code in Figure 3 was run to examine the variables and their types in the dataframe. Dataframe has 3 of bool, 2 of float64, 4 of int64, 13 of object type column which is; (Harrison and Petrou, 2020)

**Table 2 Data type and descriptions**



### **Research Question**

*Research Question 1: What are the top 10 countries with the highest number of billionaires?*

H0: There is no significant difference between the amount of wealth in the top 10 countries with the highest number of billionaires.

H1: There is a significant difference between the amount of wealth in the top 10 countries with the highest number of billionaires.

*Research Question 2: What industries/sectors are most successful?*

H0: There is no significant difference in Billionaire's rank based on industries.

H1: There is a significant difference in Billionaire's rank based on industries.

*Research Question 3: What are the main industries with the highest number of women billionaires?*

H0: There is no significant relationship between the gender of billionaires and industry.

H1: There is a significant relationship between the gender of billionaires and industry.

*Research Question 4: What age range represents the highest and lowest number of billionaires?*

H0: There is no significant relationship between success and age groups.

H1: There is a significant relationship between success and age groups.

*Research Question 5: Is there a relationship between age and wealth?*

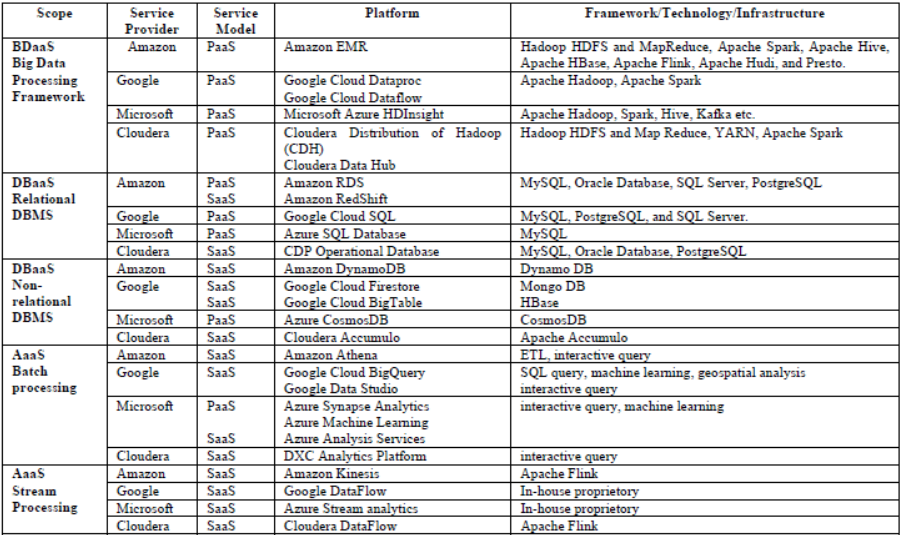
H0: There is no relationship between billionaries’s age and wealth.

H1: There is a relationship between billionaries’s age and wealth.

## **Solution Exploration**

Large and complicated datasets must be handled and processed in order to use big data applications. Big Data as a Service (BDaaS), Database as a Service (DBaaS), and Analytics as a Service (AaaS) are a few instances of cloud service brokerage-related big data analytics, and provides a comparison of various services provided by the major participants in the cloud computing sector (Fan and Rana, 2021).

**Figure 4 Summary of big data-related cloud service models (Fan and Rana, 2021)**



A cloud-based solution called "Big Data as a Service" (BDaaS) provides big data processing engines and frameworks in addition to the three fundamental cloud service types, (DBaaS) includes data storage services, Analytics as a service (AaaS) replaces expensive licenced business intelligence systems with more affordable analytical tools (Fan and Rana, 2021).

Using conventional databases like MySQL or Oracle is one strategy, but due to their constrained ability to store and analyse massive datasets, they might not be appropriate for handling big data. Utilising distributed computing frameworks like Apache Hadoop, which can manage enormous datasets over a cluster of computers, is an additional strategy (Warren and Marz, 2015).

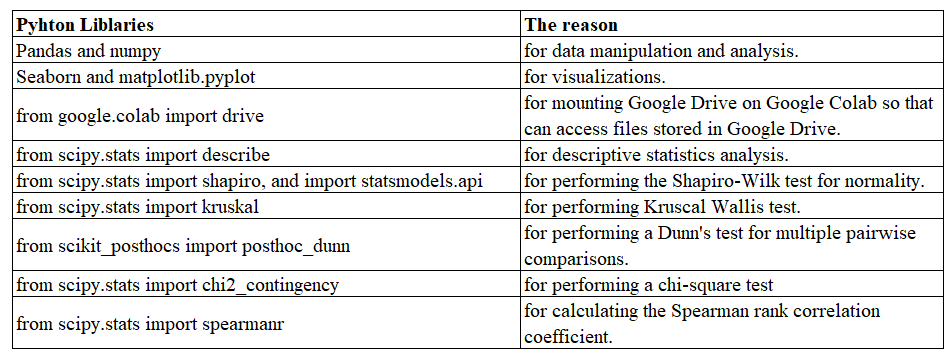
In our case, “Billionaires.csv” dataset is not a Big Data, that is why, The Google Colap Jupiter notebook was used. This app is provided free of charge by Google and is cloud-based. The data used was added to google drive beforehand, and by giving google drive access permission in the application, the data was used in the analysis. At the same time, thanks to this application, the Python code written in the study has been saved with the results and the link below can be used for access.

In the data cleaning and preparation step would involve cleaning up the Billionaires.csv dataset and preparing it for analysis. This could include removing duplicates, dealing with missing values, and transforming data as necessary. The data exploration and visualisation process will then: Exploratory data analysis methods will be applied in this step to better comprehend the data and analyse any potential patterns. Visualisations like scatter plots, histograms, and heat maps are used for the statistical analysis step for looking for trends and patterns in the data analysed, such as the top 10 countries with the most billionaires, the most lucrative industries or sectors, the age range with the highest and lowest number of billionaires, as well as other potential wealth-influencing factors. In the end, in the Communication of Results steps, the results will be presented of the analysis in a clear.

## **Solution Development**

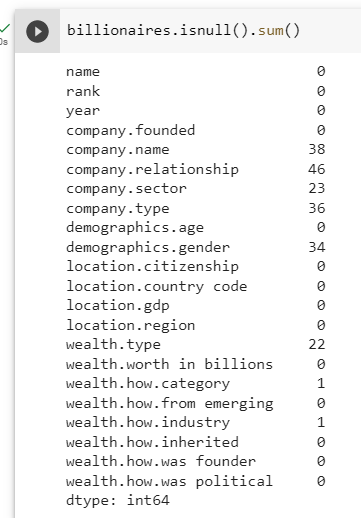
In the study, categorical data were summarised using frequency, and numerical data were summarised using mean, standard deviation, median, minimum, and maximum values. The Kruskal-Wallis test were used for compare the median between two groups because the numerical variables did not has a normal distribution, which was analyzed with the Shapiro-Wilk test. Bonferroni-corrected Dunn test used if there was a statistically significant difference in the median for pairwise median comparisons. Spearman Correlation analysis was applied to analyse the relationship between the numerical data. All analyses, data processing and statistical analysis, applied in the given scenario were made with Python. Python libraries used in the analysis are shown,

**Table 3 Python Libraries used in the analysis**

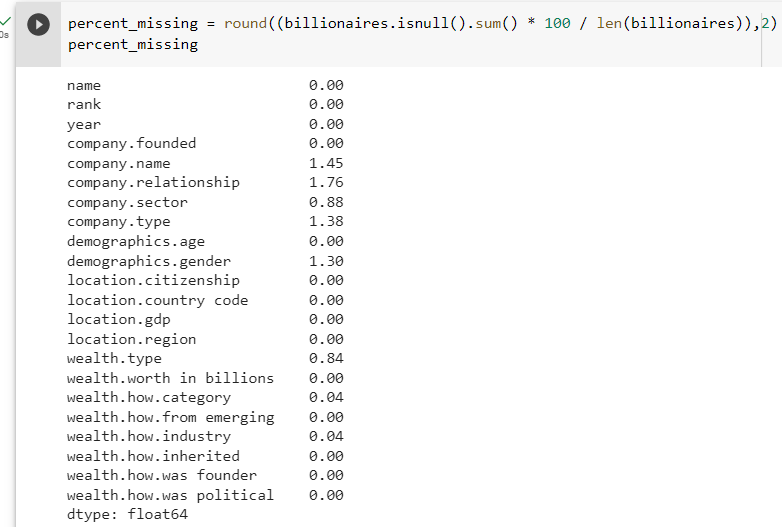


### **Data Cleaning and Preparation**

**Figure 5 Analysis of missing values in the data**

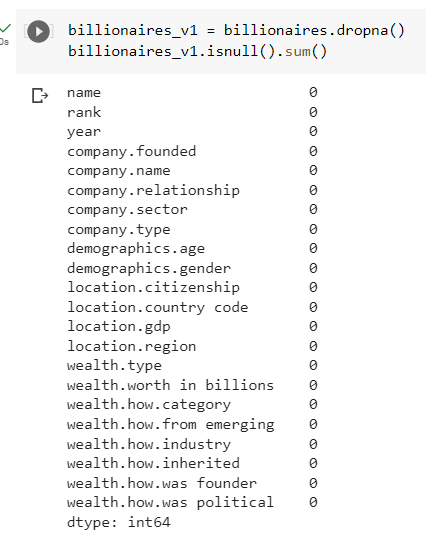


**Figure 6 Percentage of missing values in total**



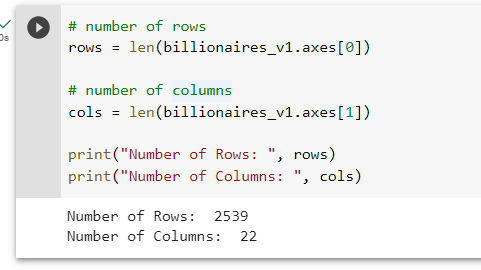
After examining the details and types of dataframe, the missing data of the variables were analyzed. In Figure 5, the variables that are missing are indicated. According to Figure 5, 8 out of 22 variables have missing values. In addition, in Figure 6, the percentage of these missing data in the total data was analyzed. The calculated percentages are in the range of min 0.04, max 1.76. Since these calculated percentages are very low, it was foreseen that they would not affect the analysis and it was decided to exclude them from the data.

**Figure 7 Removing Missing Values from Dataset**

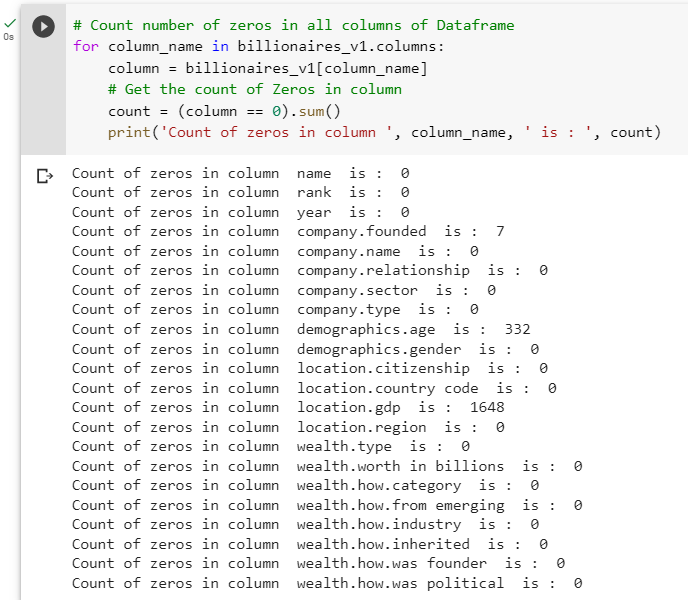


As indicated in Figure 8, missing data was removed from the data set. When we analyzed the number of rows and columns again after dropping null values (Figure 8), the number of new rows decreased from 2614 to 2539. In this case, a total of 75 columns were deleted.

**Figure 8 Total Number of rows & columns, after removing Missing Values from the Dataset**



**Figure 9 Count number of zeros in all columns of Dataframe**



After removing the missing data from the data set, the dataset was checked if is included "0" or not. According to Figure 9, the dataset has 7 “0” variables in the company.founded, 332 0” variables in the demographics.age, and 1648 0” variables in the location.gdp. In this case, instead of removing these values, it was decided that action should be taken if they are used in the analysis.

**Figure 10 Descriptive Statistics For Numeric Variables**



According to the descriptive statistics values specified in Figure 10,

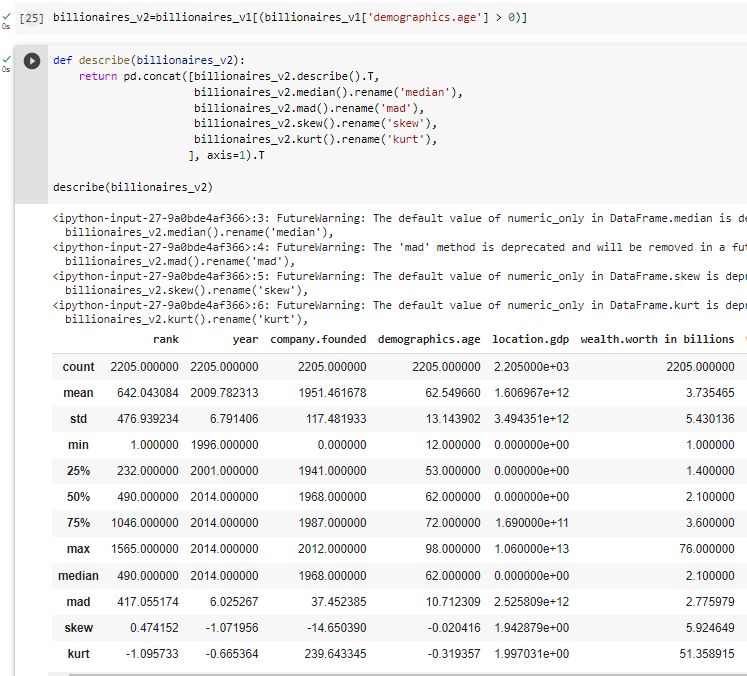
*For the Rank variable;* Their average is 606.1±469.3 and their ranking starts from min 1 to max 1565.

*For the Year variable;* The billioner data contains data up to a minimum of 1996, and a maximum of 2014.

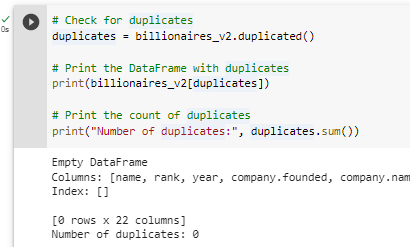
*For the variable wealth.worth in billions;* Its average is 3.57±5.15 billion and it contains min 1, max 76 billion wealth.

While these 3 data do not have 0 or negative values, the variables company.founded and demographics.age contain values <=0. The location.gdp value in the dataframe has no significance in the analysis. These descriptive can affect statistics for 'demographics.age' variables. For this reason, it is necessary to filter data for age in the column. The new database is created with the name as billionaires\_v2 which excludes 'demographics.age' <= 0 measurements (Figure 11).

**Figure 11 Descriptive Statistics without demographics.age<=0, new database**



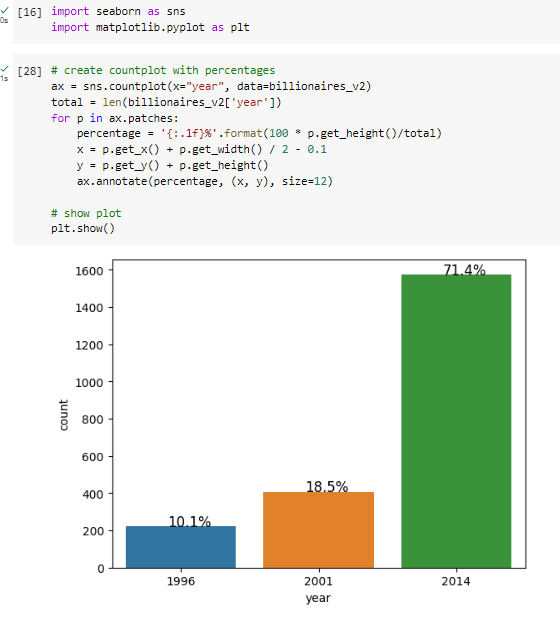
**Figure 12 Duplicated value analysis**



In the data from which missing and erroneous data were removed, the repetitive data were checked. As seen in Figure 12, there is no repeated data in the data.

### **Frequency Analysis and Normality Test**

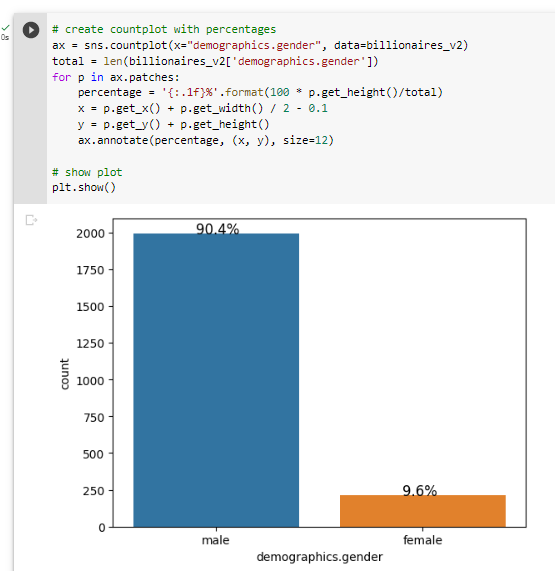
**Figure 13 Frequency analysis of "year" variable**



The Seaborn library in Python was used to analyze the frequency distribution of categorical data. A matplotlib-based Python data visualisation library called Seaborn offers a sophisticated drawing tool for creating eye-catching and educational statistical visuals. (Seaborn, n. d)

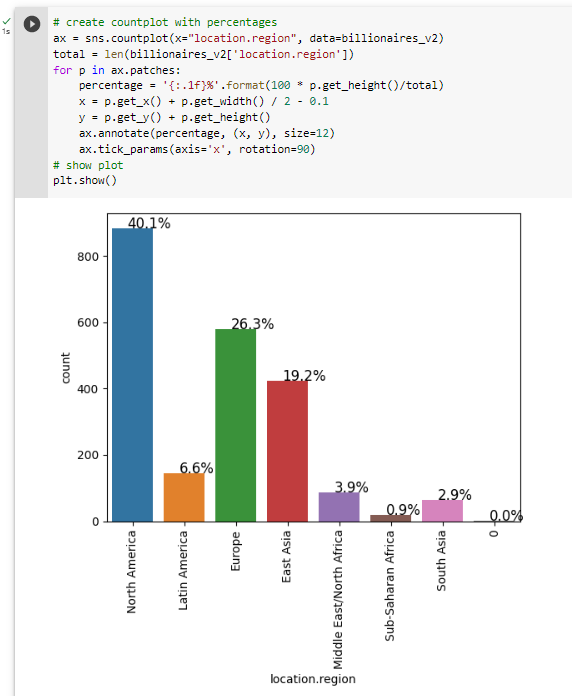
According to the barplot applied to see the frequency distribution of the year variable in the dataframe (Figure 13); 14.8% of the data includes the rankings for 1996, 20.8% for 2001, and 64.5% for 2014.

**Figure 14 Frequency analysis of "gender" variable**



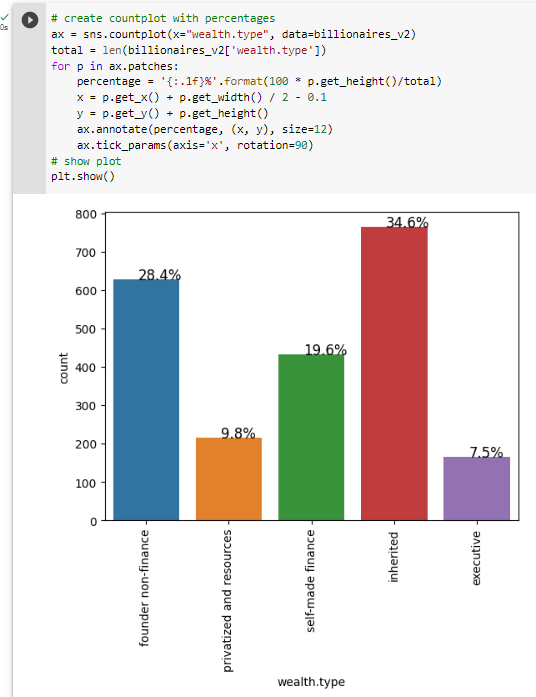
According to the barplot applied to see the frequency distribution of the gender variable in the data frame (Figure 14); 90.4% of the data is male, and 9.6% is female.

**Figure 15 Frequency analysis of "region" variable**



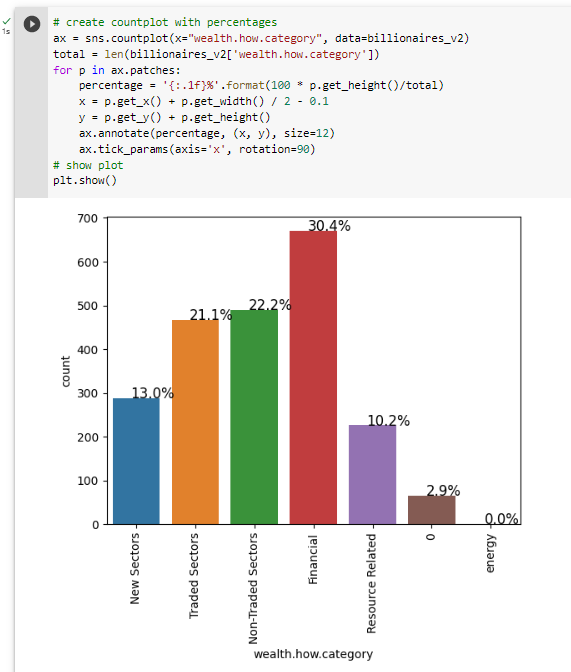
According to the barplot applied to see the frequency distribution of the region variable in the data frame (Figure 15); 40.1% of the billionaires in the data live in the North America region. Then, 26.3% live in Europe, 19.2% East Asia, 6.6% Latin America, 3.9% Middle East/North Africa, 2.9% South Asia, 0.9% Sub-Saharan Africa.

**Figure 16 Frequency analysis of "wealth type" variable**



According to the barplot applied to see the frequency distribution of the wealth type variable in the data frame (Figure 16); 34.6% of the billionaires’s wealth type in the data is inherited. Then, 28.4% is founder non-finance, 19.6% is self-made finance, 9.8% is privatized and resources, 7.5% is executive.

**Figure 17 Frequency analysis of "wealth how category" variable**



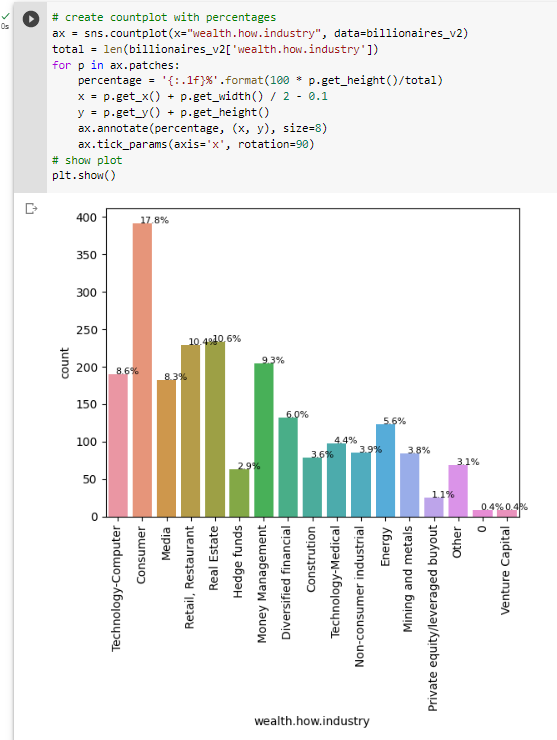
According to the barplot applied to see the frequency distribution of the wealth how category variable in the data frame (Figure 17); 30.4% of the billionaires’s the source of their billions in the data is financial. Then, 22.2% is non-trated sectors, 21.1% is traded sectors, 13% is new sectors, 10.2% is resource related. Data include “0” variables.

**Figure 18 Frequency analysis of "wealth how inherited" variable**



According to the barplot applied to see the frequency distribution of the wealth how inherited variable in the data frame (Figure 18); 65.5% of the billionaires’s the way that their money was inherited in the data is not inherited. Then, 21.7% is father, 6.9% is 3rd generation, 2.8% is 4th generation, 2.1% is spouse/widow, 1% is 5th generation or longer.

**Figure 19 Frequency analysis of "wealth how idustry" variable**



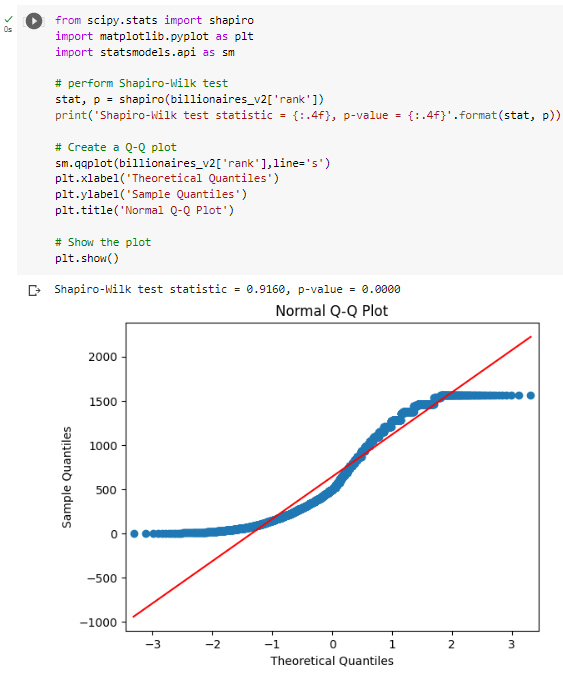
According to the barplot applied to see the frequency distribution of the wealth how industry variable in the data frame (Figure 19); the consumer industry is 17.8% of billionaires, which has the most billionaire’s industry.

After analyzing the frequency distribution of the categorical variables, the normal distribution test was performed before starting the analysis on the numerical variables.

While the Kolmogorov-Smirnov test is used for n≥50, the Shapiro-Wilk test is more appropriate for small sample sizes (n<50 samples), though it can also handle larger sample sizes. The null hypothesis is accepted and the data are referred to as normally distributed when p > 0.05 (Mishra, et. al ).

Q-Q plot is a graphical method to check the normality of the data. If the data is normally distributed, the points on the plot should follow a straight line, and If the points should not flow a straight line, it means that the data is not normally distributed (Holgersson, 2006).

**Figure 20 Normality test for "rank" variable**

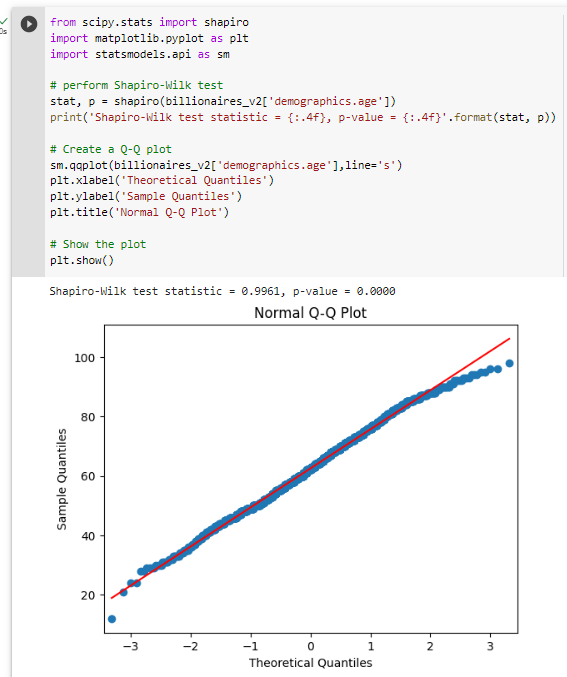


H0: “Rank” variable is normally distributed

H1: “Rank” variable is not normally distributed.

The p value calculated as a result of the Shapiro wilk test is <0.05. In this case, the null hypothesis is rejected statistically and it cannot be said that the rank variable has a normal distribution. As we can see in the Normal Q-Q Plot, the data did not follow a straight line, which means that the data is not normally distributed.

**Figure 21 Normality test for "demographics.age" variable**

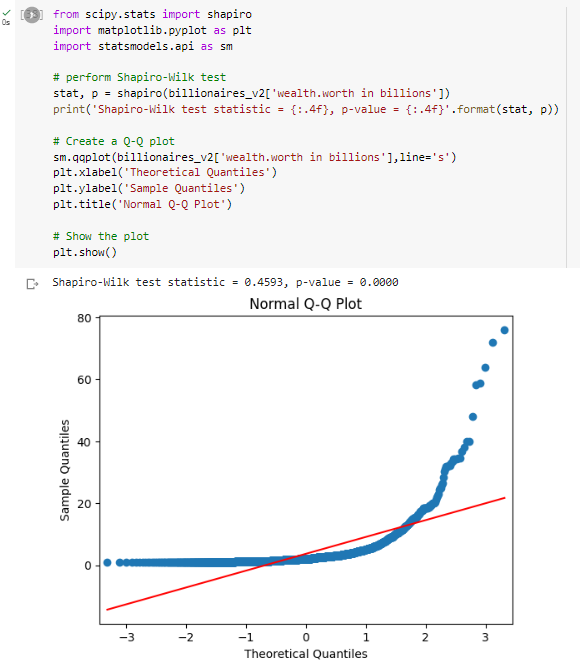


H0: “demographics.age” variable is normally distributed

H1: “demographics.age” variable is not normally distributed.

The p value calculated as a result of the Shapiro wilk test is <0.05. In this case, the null hypothesis is rejected statistically and it cannot be said that the age variable has a normal distribution. As we can see in the Normal Q-Q Plot, the data did not follow a straight line, which means that the data is not normally distributed.

**Figure 22 Normality test for "wealth.worth in billions" variable**



H0: “wealth.worth in billions” variable is normally distributed

H1: : “wealth.worth in billions”variable is not normally distributed.

The p value calculated as a result of the Shapiro wilk test is <0.05. In this case, the null hypothesis is rejected statistically and it cannot be said that the wealth worth in billions variable has a normal distribution. As we can see in the Normal Q-Q Plot, the data did not follow a straight line, which means that the data is not normally distributed.

The assumption of normality in numerical variables was checked with the Shapiro Wilk test, and the Kruscal Wallis test was used for comparisons with more than two groups, since they did not show normal distribution. Chi-square test (if n<5 fisher's exact test) was used to analyze the relationship between categorical data. Spearman Correlation analysis was applied to determine the relationship between the numerical data, depending on the results of non-normal distribution.

### **Statistical Analysis; to solve research questions**

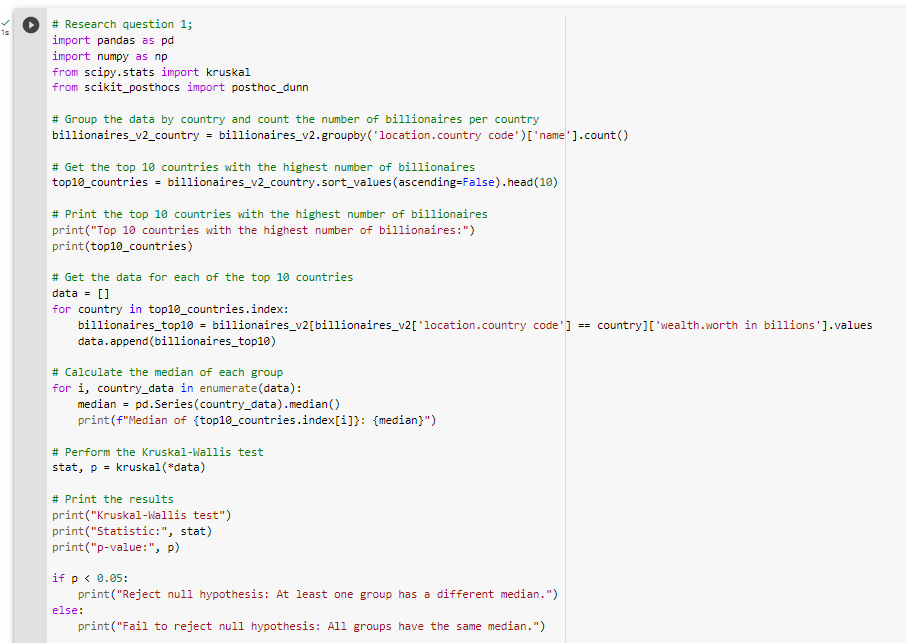
*Research Question 1: What are the top 10 countries with the highest number of billionaires?*

H0: There is no significant difference between the amount of wealth in the top 10 countries with the highest number of billionaires.

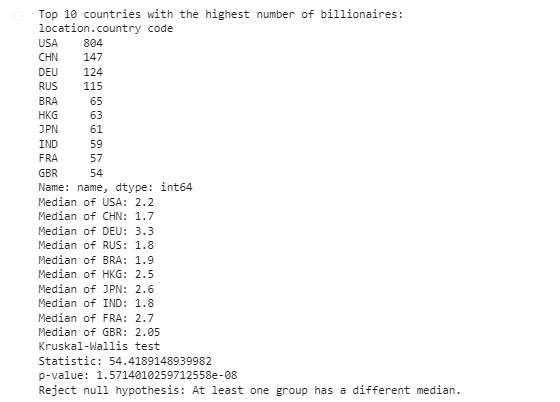
H1: There is a significant difference between the amount of wealth in the top 10 countries with the highest number of billionaires.

Note: This is a descriptive statistics question, not a hypothesis-testing question. That is why, first detect the top 10 countries that have the most billionaire from. Then compare these top 10 countries based on their wealth worth in billion.

**Figure 23 Python code for Research Question 1, step 1**



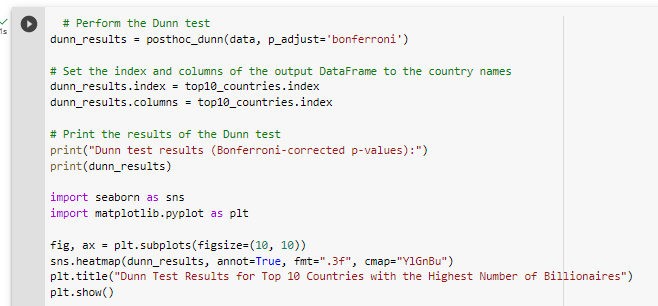
**Figure 24 Kruscal Wallis test's Output for Research Question 1**



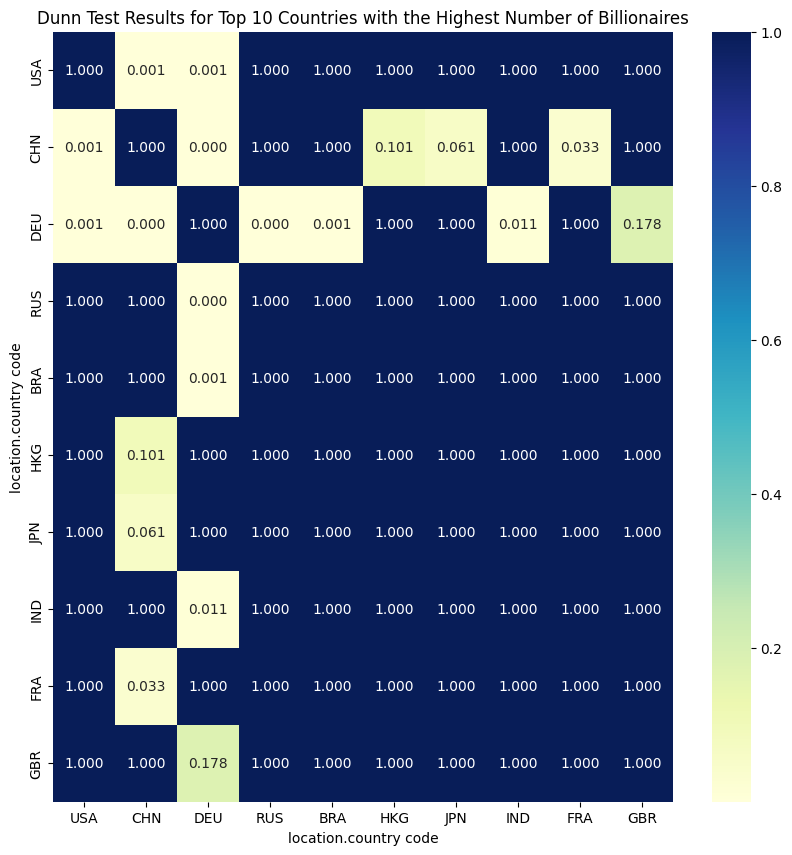
In the analysis of research question 1, the Python code that is shown in Figure 23 for firstly the top 10 countries with the highest number of billionaires was determined, and then applied Kruscall Wallis to analyse the hypothesis. The top 10 countries with the highest number of billionaires are, in order, the USA with 804 billionaires, CHN with 147 billionaires, DEU with124 billionaires, RUA with 115 billionaires, BRA with 65 billionaires, HKG with 63 billionaires, JPN with 61 billionaires, IND with 59 billionaires, FRA with 57 billionaires, and GBR with 54 billionaires. (Figure 24)

Afterwards, Kruscal Wallis non-parametric test, which is used to compare more than 2 groups from statistical methods, was applied to determine whether there is a difference in wealth worth in billions between these countries. The Python code used in Figure 23 is indicated. When we interpret the outputs of the analysis (Figure 24), p<0.05 means that, at least one of group has different median.

**Figure 25 Perform Dunn test to analyse pairwise comparison between two groups, for Research Question 1**



**Figure 26 Dunn Test’s Output, Research Question 1**



After the Kruskal Wallis test, Dunn's test with Bonferroni correction was applied for pairwise group comparisons (Figure 25). The output of the Dunn test is a matrix that shows the p-values for all pairwise comparisons between the groups (Figure 26).

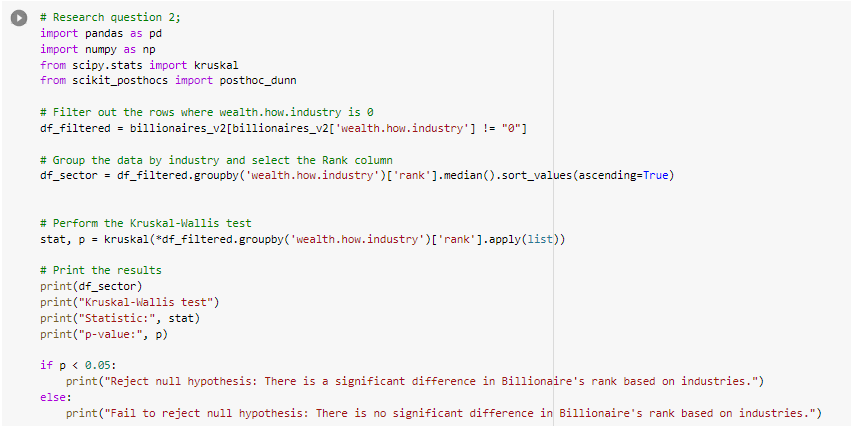
In order to determine the difference between the groups, according to the Bonferroni corrected Dunn test results applied after the Kruskal Wallis test (Figure 26); There is a significant difference between USA and CHN, DEU countries. The median wealth worth in billions of USA is lower than DEU and is higher than CHN. Likewise; There is a significant difference between CHN and DEU, HKG and FRA countries. The median wealth worth in billions of CHN is lower than DEU, HKG, and FRA. There is a significant difference between DEU and RUS, BRA, and IND countries. The median wealth worth in billions of DEU is higher than RUS, BRA, and IND.

*Research Question 2: What industries/sectors are most successful?*

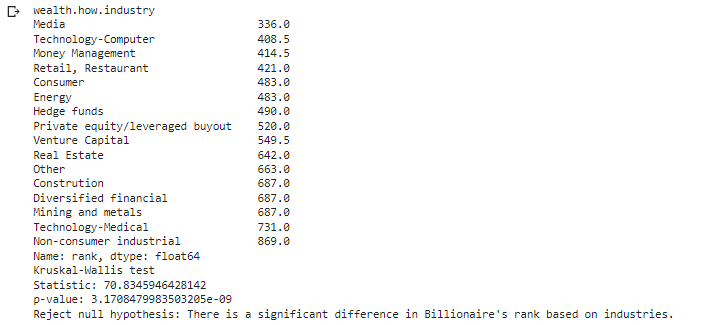
H0: There is no significant difference in Billionaire's rank based on industries.

H1: There is a significant difference in Billionaire's rank based on industries.

**Figure 27 Python code for Research Question 2**

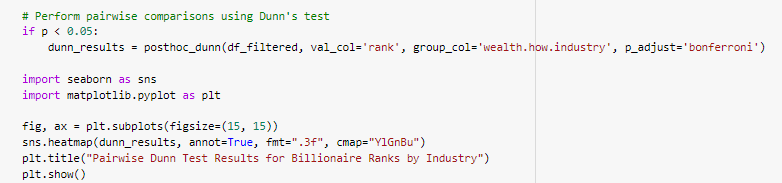


**Figure 28 Kruscal Wallis test's Output for Research Question 2**

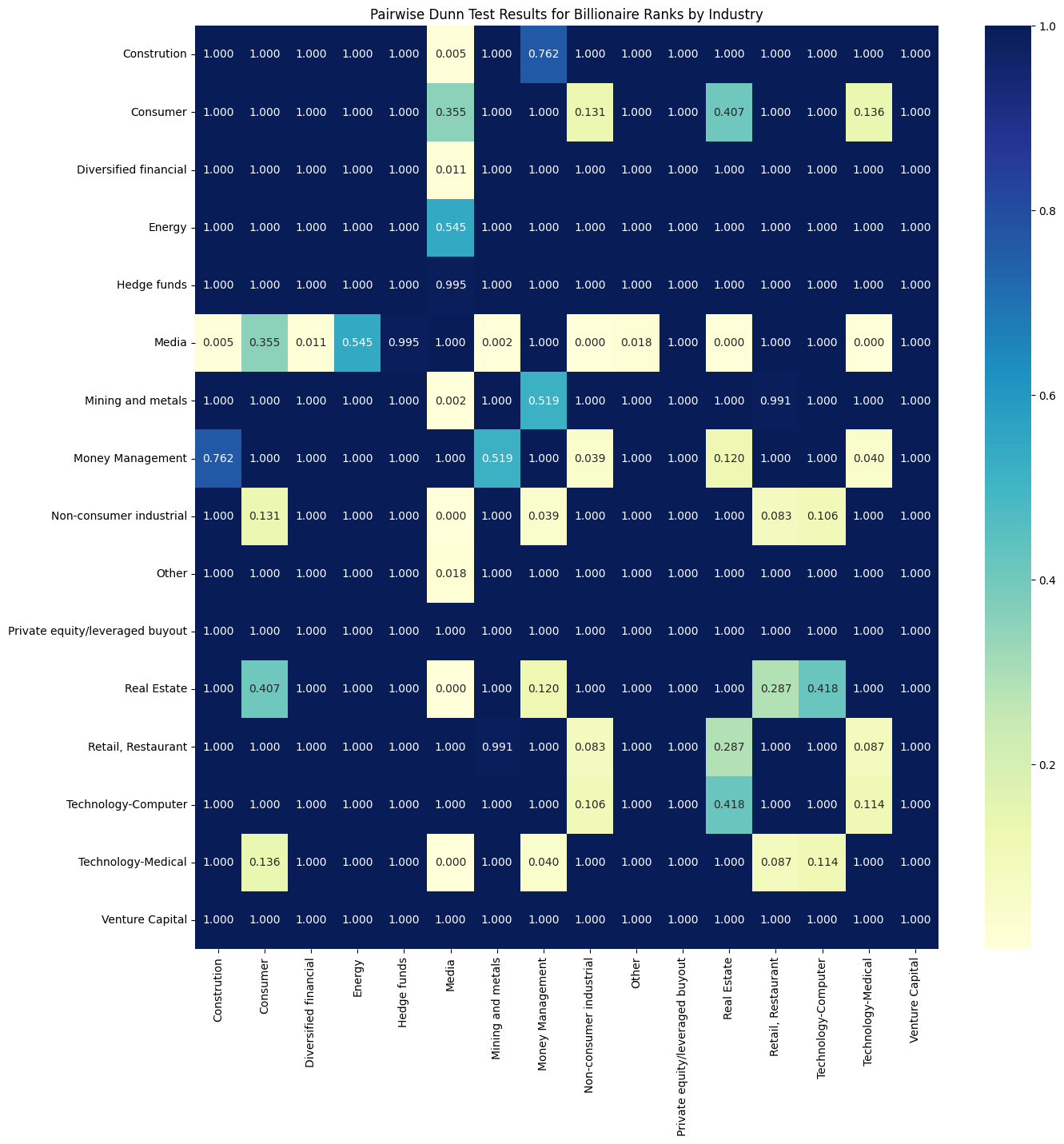


In the analysis of research question 2, Kruscal Wallis's non-parametric test, which is used to compare more than 2 groups from statistical methods, was applied to determine whether there is a difference in Billionaire's rank between industries. The Python code is used in Figure 27. When we interpret the outputs of the analysis (Figure 28), p<0.05 means that we can reject the null hypothesis, which means there is a significant difference in Billionaire's rank within industries. If we analyse the median of industries ranks, we can say that the most successful industry is Media, the second one is Technology-Computer and the third one is Money Management (Figure 28).

**Figure 29 Perform Dunn test to analyse pairwise comparison between two groups, for Research Question 2**



**Figure 30 Dunn Test’s Output, Research Question 2**



After the Kruskal Wallis test, Dunn's test with Bonferroni correction was applied for pairwise group comparisons (Figure 29). The output of the Dunn test is a matrix that shows the p-values for all pairwise comparisons between the groups (Figure 30). In this matrix, if p<0.05 we can say that, there is a significant difference in the medians of ranks between industries.

*Research Question 3: What are the main industries with the highest number of women billionaires?*

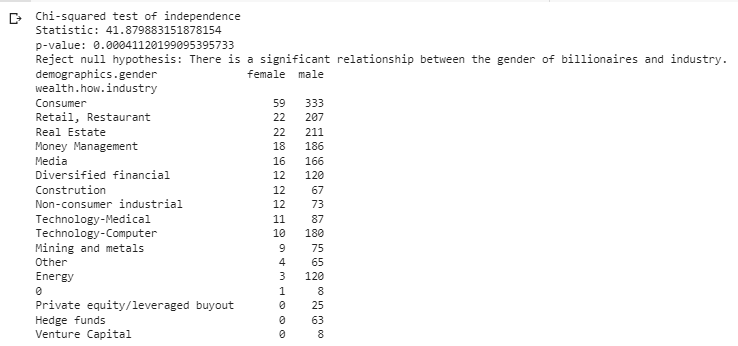
H0: There is no significant relationship between the gender of billionaires and industry.

H1: There is a significant relationship between the gender of billionaires and industry.

**Figure 31 Python code for Research Question 3**



**Figure 32 Chi-squared test's Output for Research Question 3**



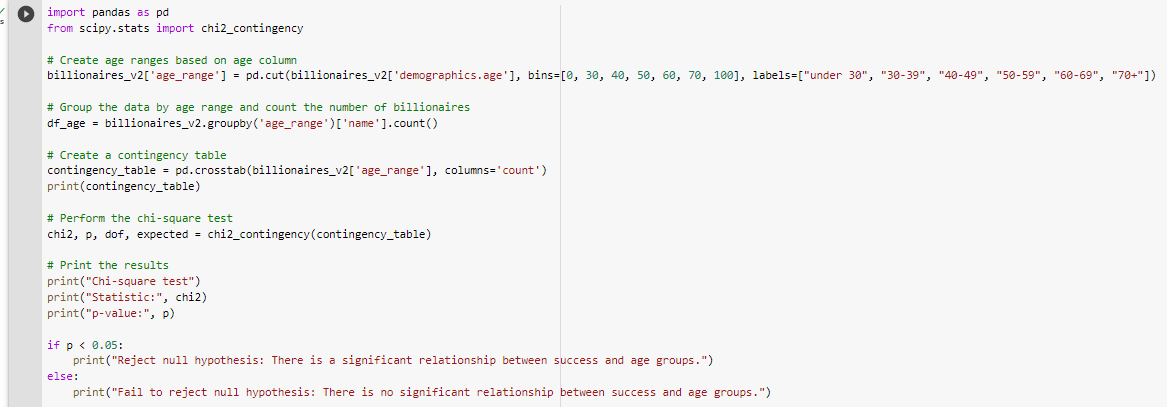
In the analysis of research question 3, the Chi-squared test, is used to determine whether there is a significant association between two categorical variables, billionaires' gender and industries. The Python code is used in Figure 31. When we interpret the outputs of the analysis (Figure 32), p<0.05 means that we can reject the null hypothesis, which means there is a significant difference with Billionaires' gender within their industries. If we analyse the frequency of gender, we can say that male billionaires are more successful than female billionaires. Based on the cross table shown in Figure 32, the Consumer industry has the highest number of women billionaires (n=59), in order, Retail, Restaurant and Real Estate have 22, Money Management has 18 and so on.

*Research Question 4: What age range represents the highest and lowest number of billionaires?*

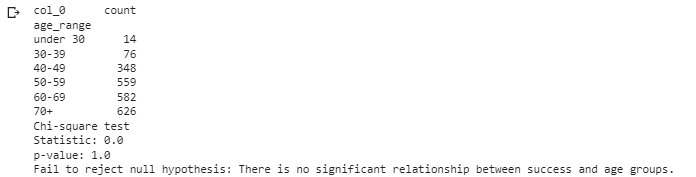
H0: There is no significant relationship between success and age groups.

H1: There is a significant relationship between success and age groups.

**Figure 33 Python code for Research Question 4**



**Figure 34 Chi-squared test's Output for Research Question 4**



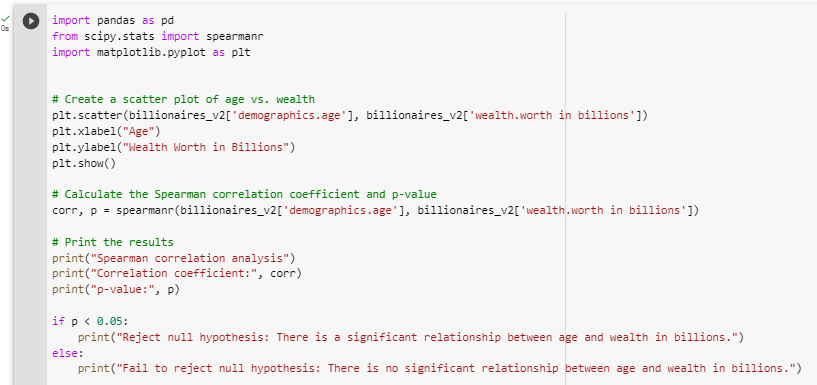
In the analysis of research question 4, First, the age group variable created based on age, than the Chi-squared test, is used to determine whether there is a significant association between two categorical variables, billionaires' success and age group. The Python code is used in Figure 33. When we interpret the outputs of the analysis (Figure 34), p>0.05 means that we can not reject the null hypothesis, which means there is no significant difference with Billionaires' age group within their success. If we analyse the frequency of age groups, we can say that the 70+ age billionaires are more successful (n=626) than other age groups, and we can say that those under 30 of billionaires are the lowest successful (n=14) than other age groups.

*Research Question 5: Is there a relationship between age and wealth?*

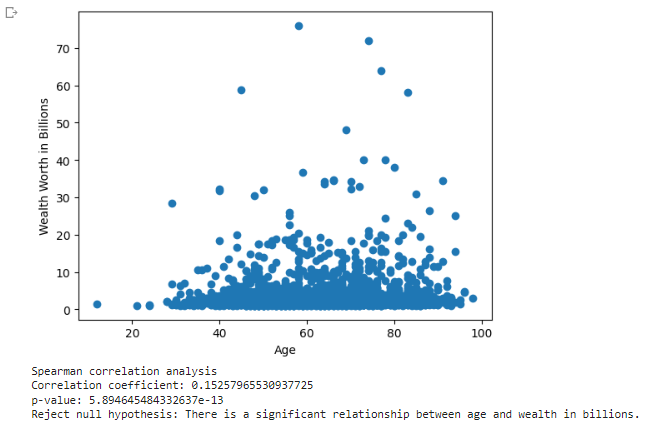
H0: There is no relationship between billionaries’s age and wealth.

H1: There is a relationship between billionaries’s age and wealth.

**Figure 35 Python code for Research Question 5**

****

**Figure 36 Spearman Correlation test's Output for Research Question 5**

****

In the analysis of research question 5, the Spearman Correlation test (non-parametric test for Pearson) is used to determine whether there is a significant relationship between two numeric variables, billionaires' age and wealth worth in billions. The Python code is used in Figure 35. When we interpret the outputs of the analysis (Figure 36), p<0.05 means that we can reject the null hypothesis, which means there is a significant relationship between billionaires' age and wealth worth in billions. If we analyse the Correlation coefficient:0.152, we can say that there is a weak positive correlation between age and wealth and as age increases, wealth also increases.

## **1.4 Evaluation and Future Development**

In this research “Billionaires.csv” dataset used to answer research questions, which are;

Research Questions;

*Research Question 1: What are the top 10 countries with the highest number of billionaires?*

*Research Question 2: What industries/sectors are most successful?*

*Research Question 3: What are the main industries with the highest number of women billionaires?*

*Research Question 4: What age range represents the highest and lowest number of billionaires?*

*Research Question 5: Is there a relationship between age and wealth?*

In order to reach the answers to these research questions, 5 different hypotheses were established and the necessary analyzes for each hypothesis were made in section 1.3.3. Pyhton codes and outcomes of the analyzes are also included in the same section.

Based on statistical analysis results, are shown in Section 1.3.3,

* For research question 1,

The top 10 countries with the highest number of billionaires are USA, CHN, DEU, RUA, BRA, HKG, JPN, IND, FRA, and GBR.

At least one of the countries has a different median of wealth worth in billions. (there is a difference among group distribution)

And based on the Bonferroni correction Dunn test;

The median wealth worth in billions of USA is lower than DEU and is higher than CHN, CHN is lower than DEU, HKG, and FRA, DEU is higher than RUS, BRA, and IND.

* For research question 2,

At least one of the industries has a different median of rank. (there is a difference among group distribution)

And based on the Bonferroni correction Dunn test;

We can say that the most successful industry is Media, the second one is Technology-Computer and th*e third one is Money Management.*

* For research question 3,

There is a significant association between billionaires' gender and industries.

And based on the frequency of gender, we can say that male billionaires are more successful than female billionaires. And the Consumer industry has the highest number of women billionaires (n=59.

* For research question 4,

There is no significant difference with Billionaires' age group within their success

The frequency of age groups, we can say that the 70+ age billionaires are more successful (n=626) than other age groups, and we can say that those under 30 of billionaires are the lowest successful (n=14) than other age groups.

* For research question 5,

There is a significant relationship between billionaires' age and wealth worth in billions.

Based on Correlation coefficient;

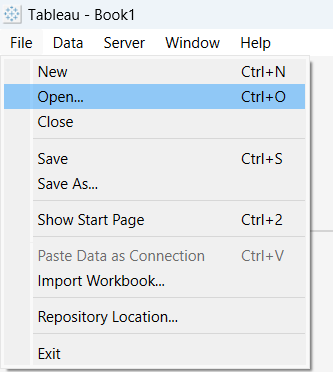
There is a weak positive correlation between age and wealth and as age increases, wealth also increases.

Overall, based on these outcomes,

Billionaires dataset, the top 10 countries with the highest number of billionaires are USA, CHN, DEU, RUA, BRA, HKG, JPN, IND, FRA, and GBR, and it was found that there is a difference in the distribution among these groups. The most successful industry was the Media industry, followed by the Technology-Computer industry and the Money Management industry. It was discovered that male billionaires were more successful than female billionaires, and the Consumer industry had the most female billionaires. In addition, it was discovered that billionaires 70 and older had the greatest success rates, with a total of 626 billionaires. However, there were the fewest billionaires among individuals under the age of 30, with only 14 billionaires. Additionally, there was a marginally positive relationship between age and wealth, indicating that as age rises, so does wealth.

# **Business Intelligence (Tableau)**

**Figure 37 Import data to Tableu**



As shown in Figure 37, “Electronic Sales” data was imported to Tableau.

## **Task 1**

Step 1. Parametre was created (Figure 38)

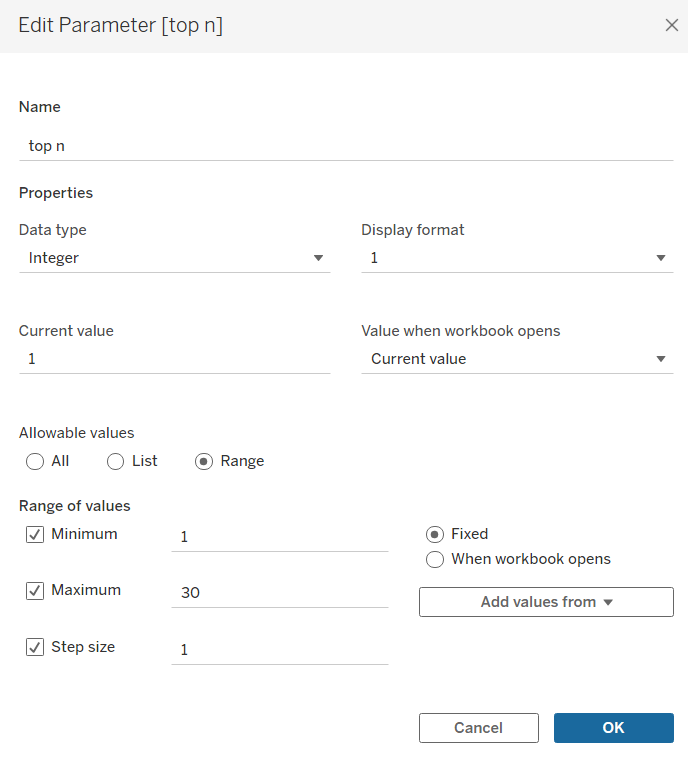
Step 2: Create a calculated field with the RANK function to rank the orders (Figure 39)

Step 3: Drop the calculated fields that name is Order to filter for calculating top 10 (Figure 40)

Step 4: Duplicate top 10 sheet (Figure 41)

Step 5: Filter below 10 (Figure 42)

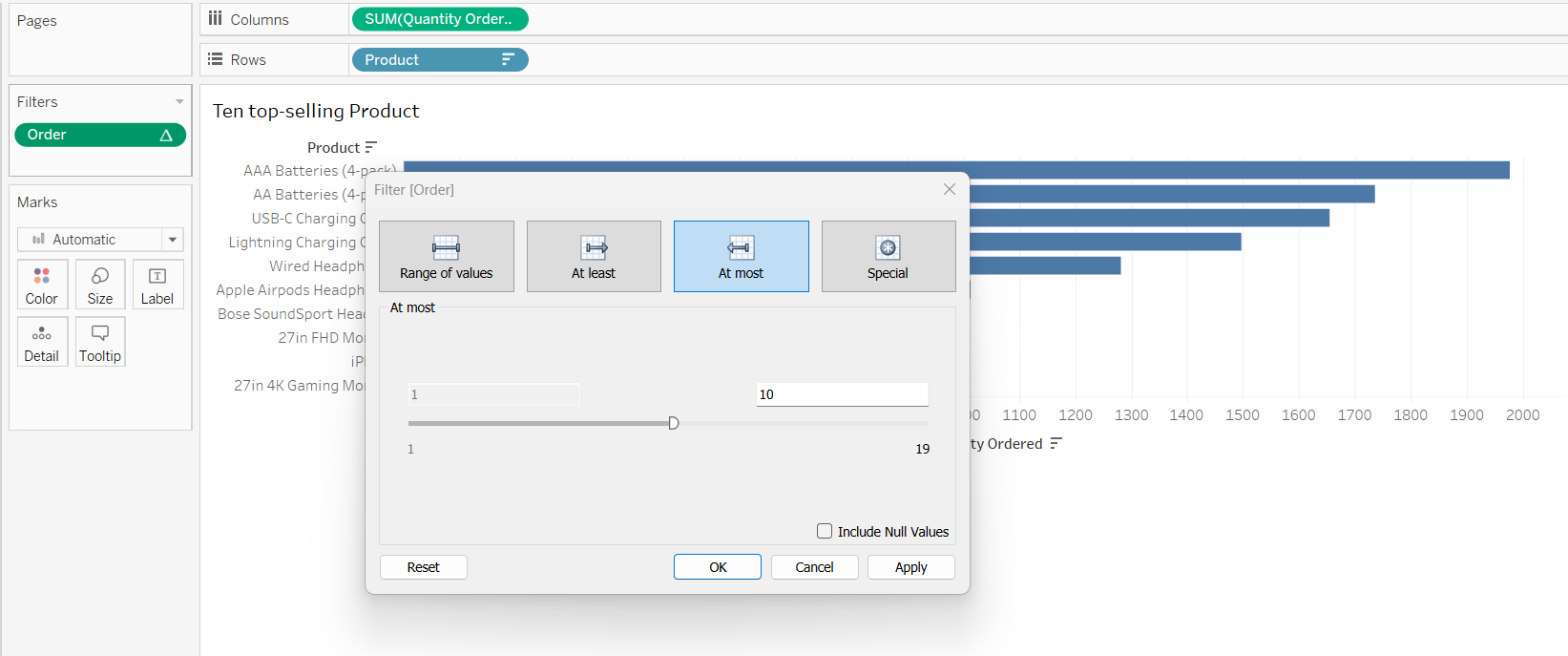
**Figure 38 Creating Parameter for top n**



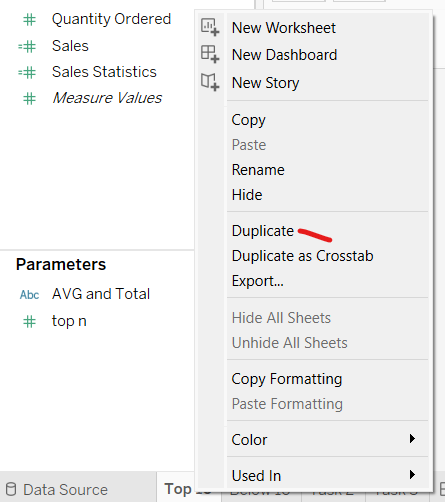
**Figure 39 Creating a calculated field**



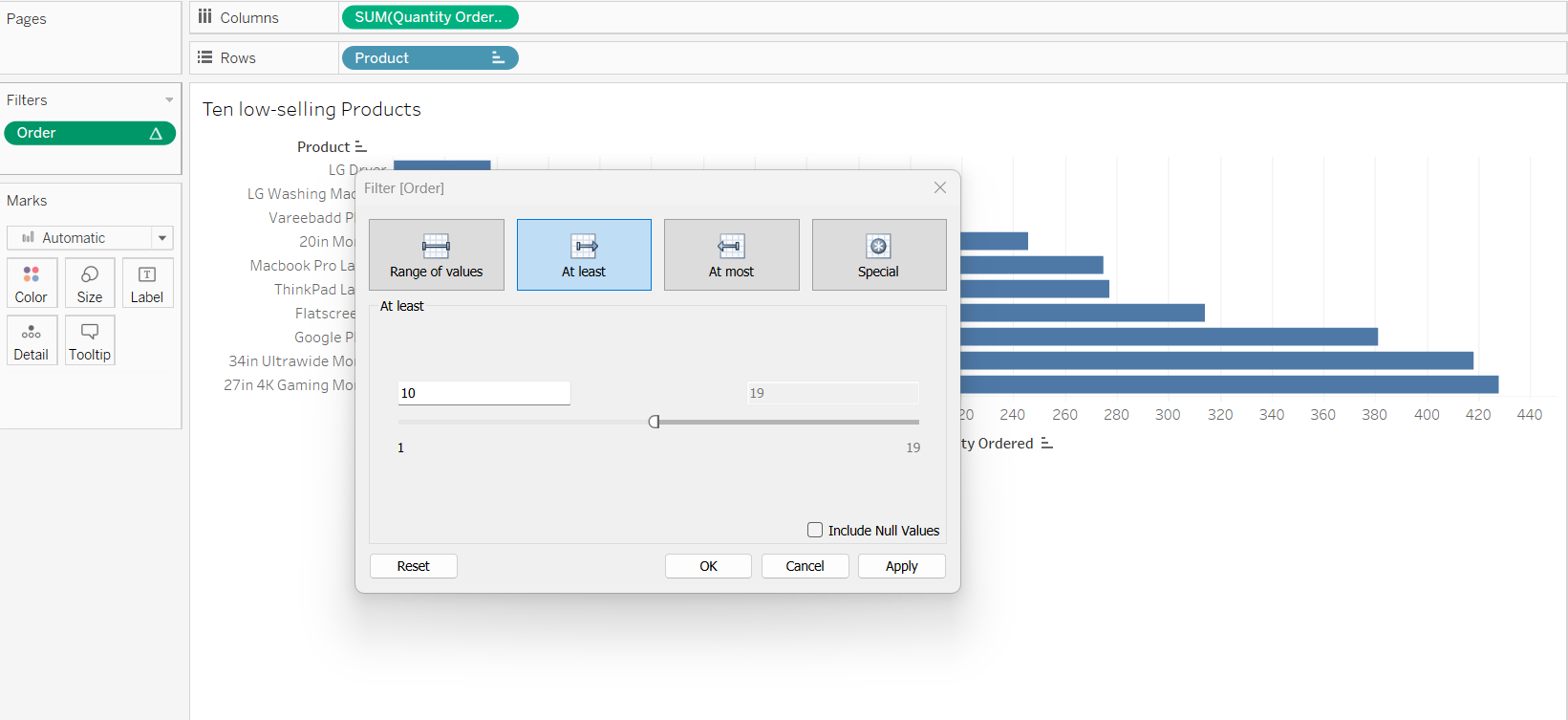
**Figure 40 Filtering top 10 sales product**



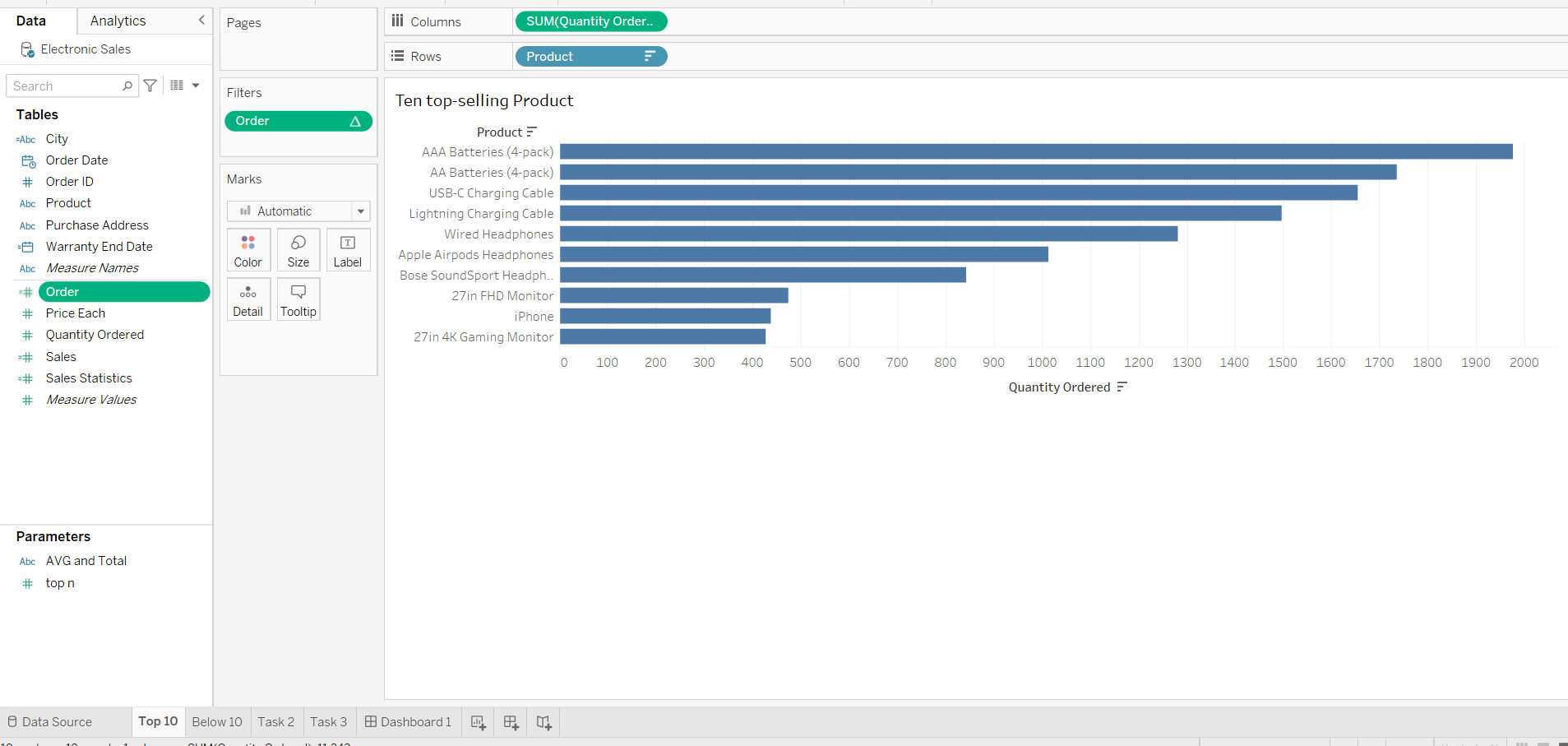
**Figure 41 Duplicating sheets**



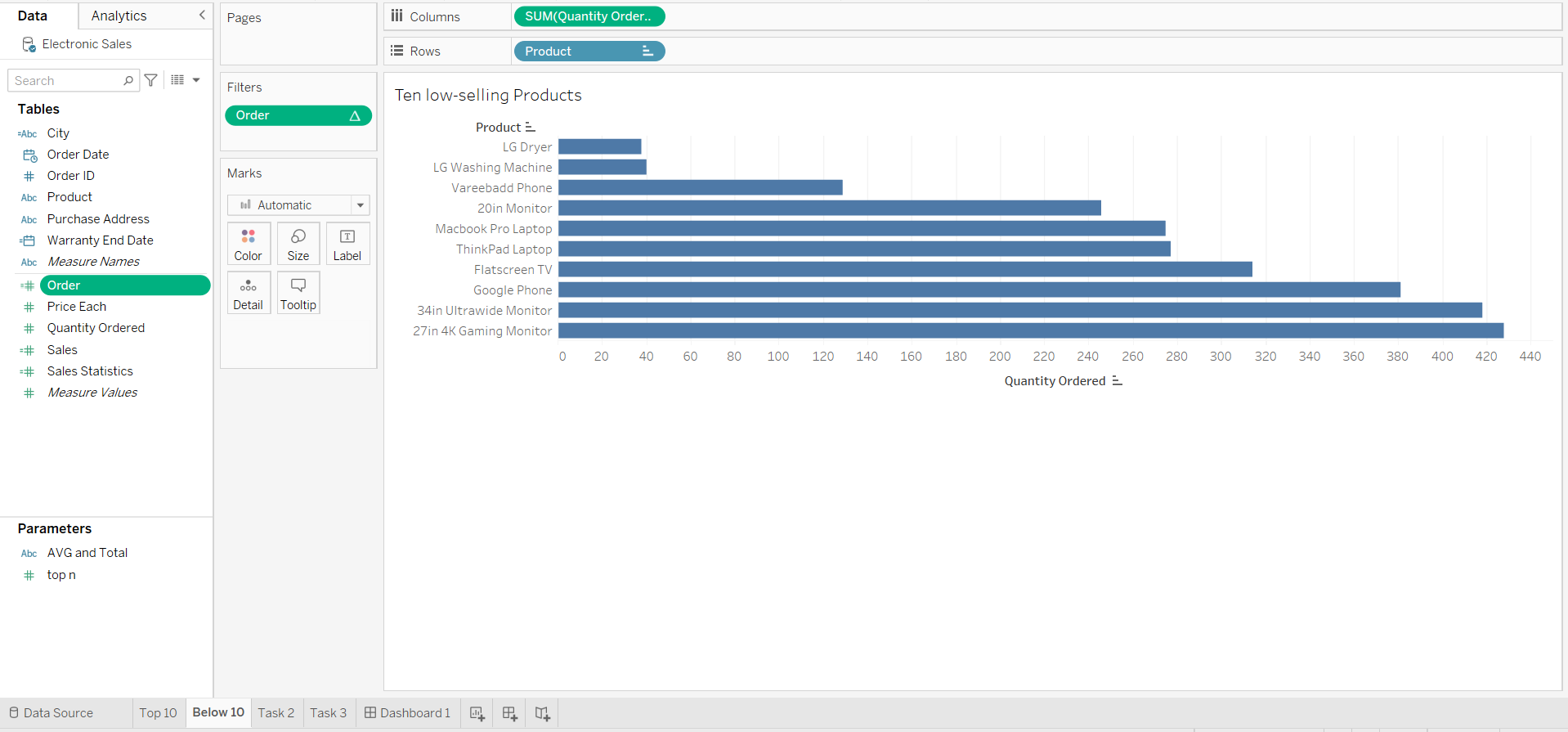
**Figure 42 Filtering 10 low-selling product**



**Figure 43 Top 10 selling product**



**Figure 44 Top 10 low-selling product**



## **Task 2**

Step 1: Using Custom Split data to create a new fields as City (Figure 45, 46)

Step 2: Rename custom split field as “City” (Figure 47)

Step 3: Calculate Sales with Quantity Ordered\*Price Each (Figure 48)

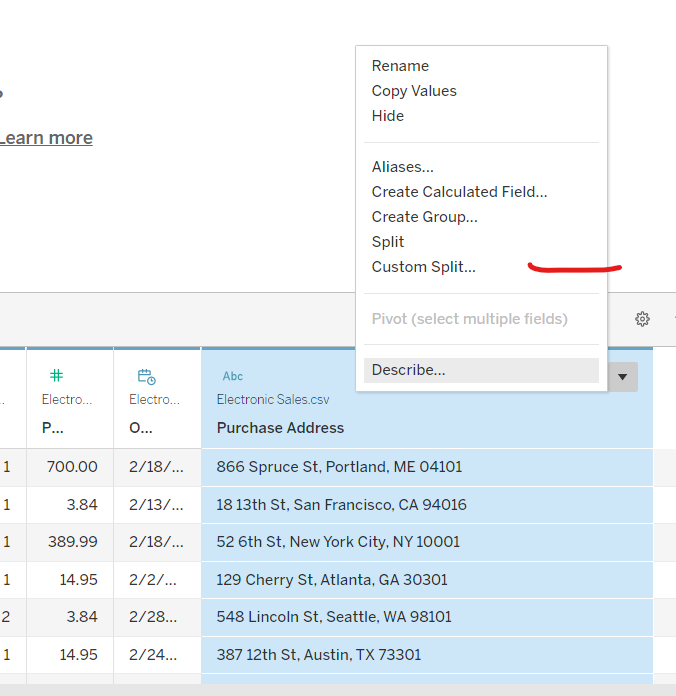
Step 4: Create Parameter for AVG and SUM (Figure 50)

Step 5: Create calculated fields as Sales Statistics (Figure 51)

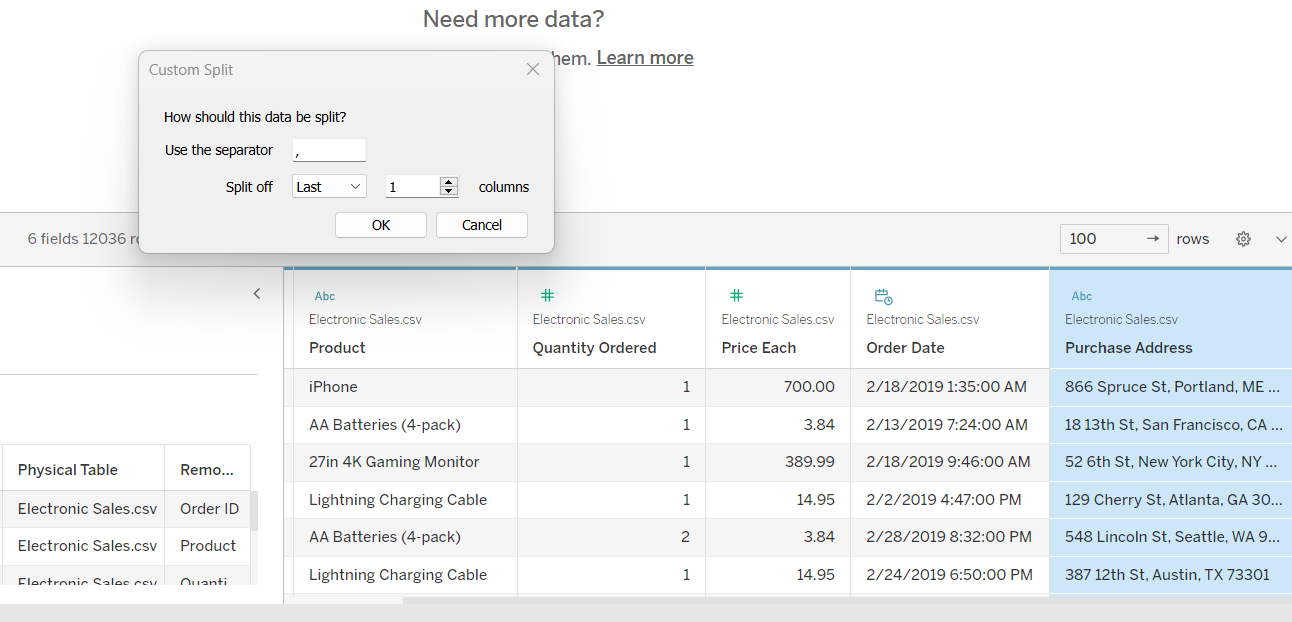
Step 6: Drop sales statistics to columns, City to row, and city to filter to not show null values

Step 7: With parameter, user can select at the right side, avg or total.

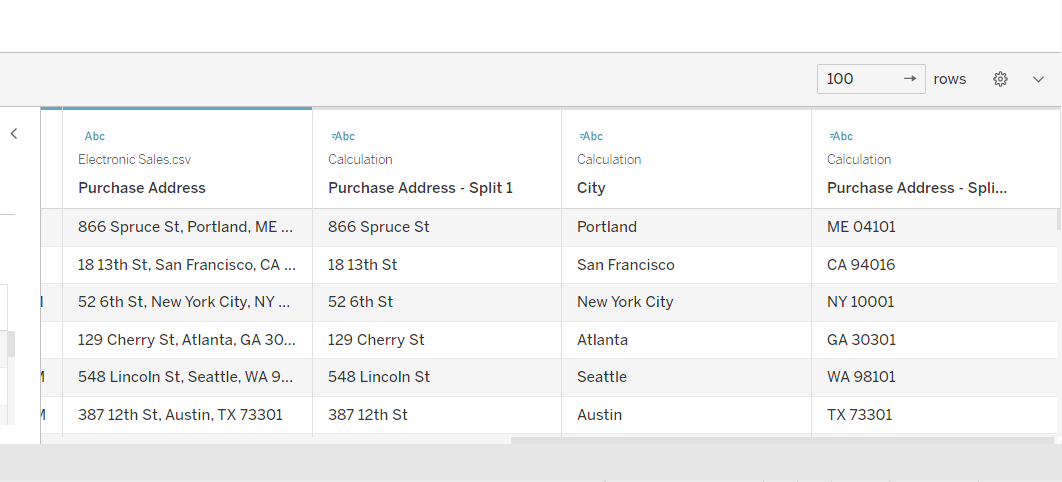
**Figure 45 Split Columns to create a field\_1**



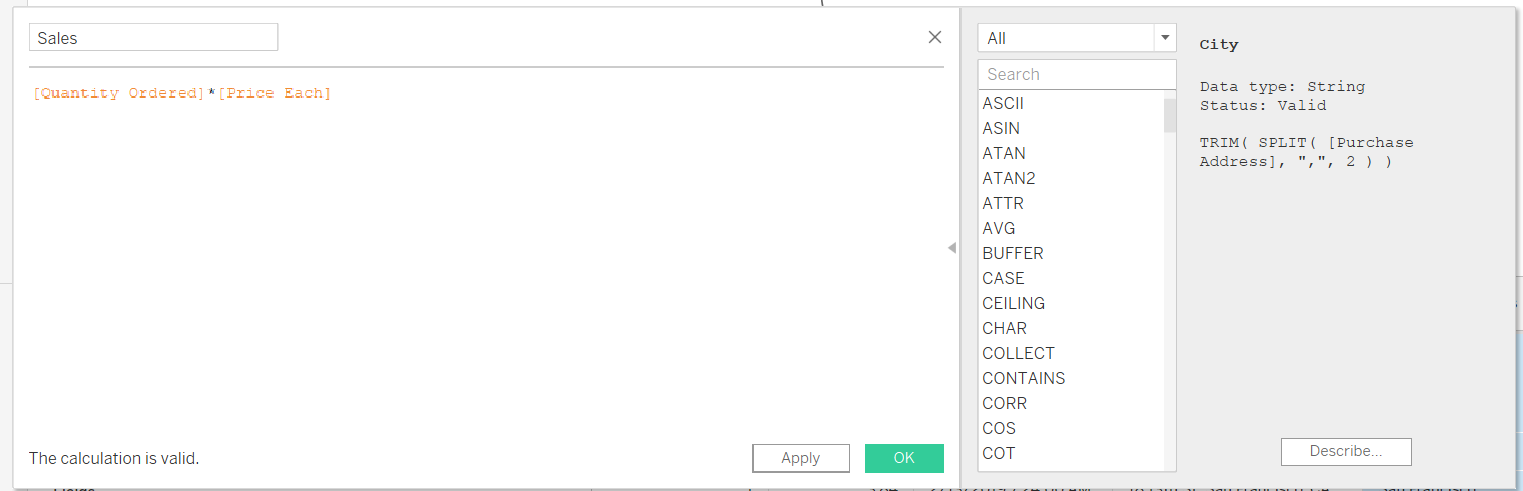
**Figure 46 Split Columns to create a field\_2**



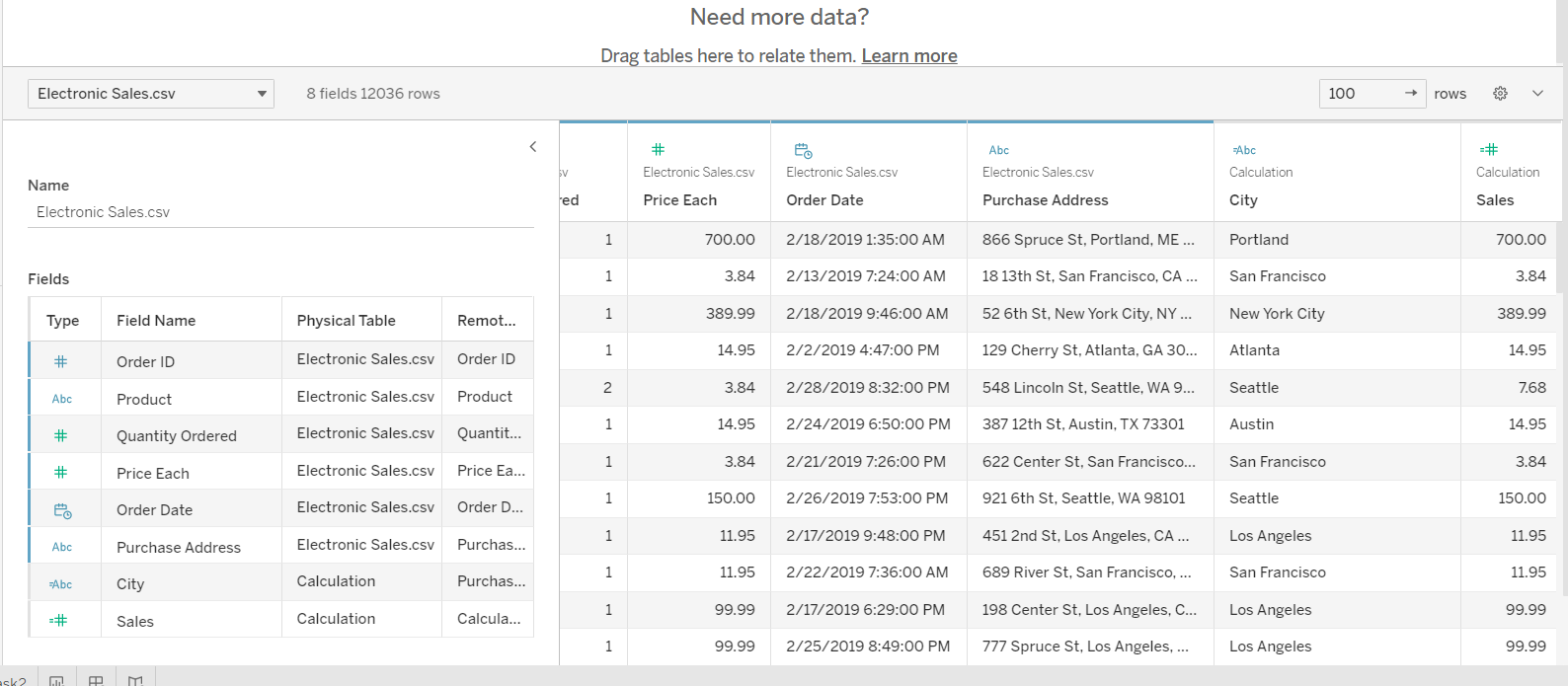
**Figure 47 Rename fields**



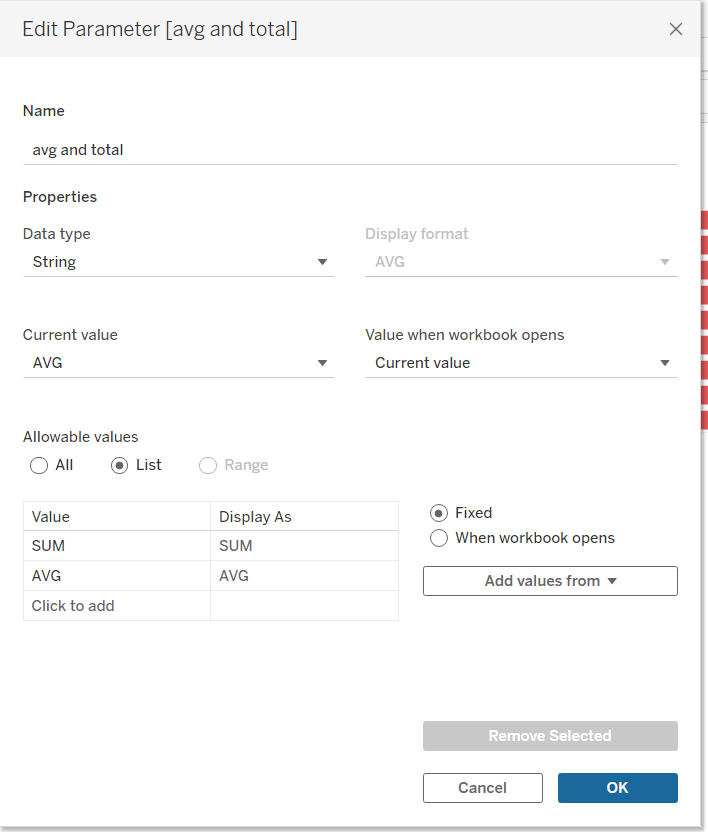
**Figure 48 Calculating Sales**



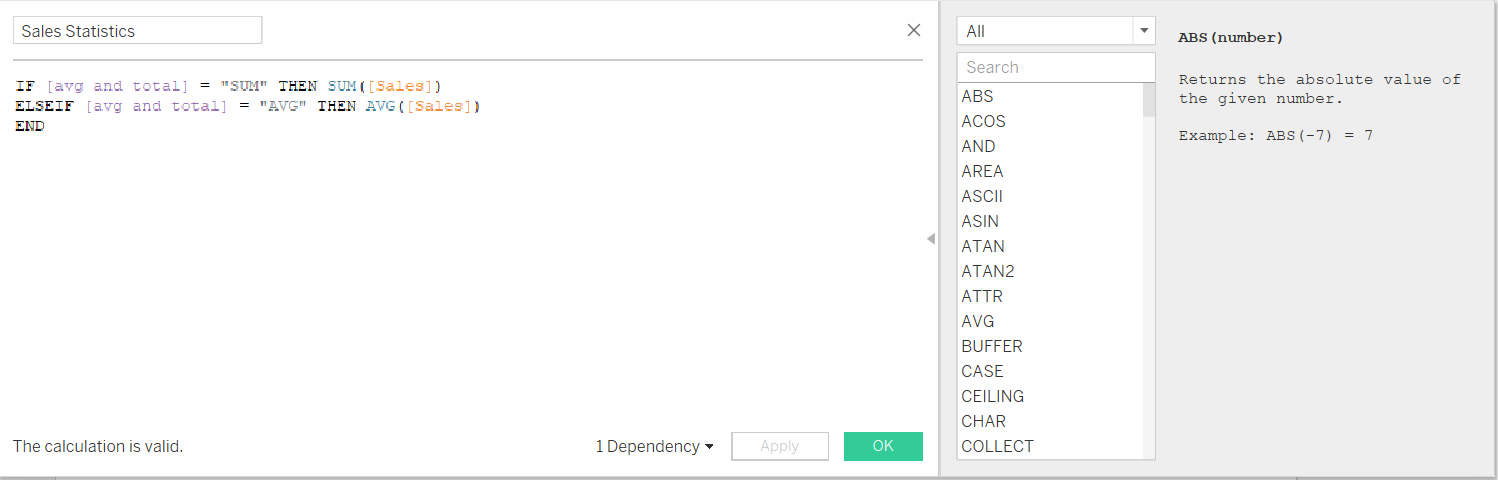
**Figure 49 New databased with Sales and City Fields**



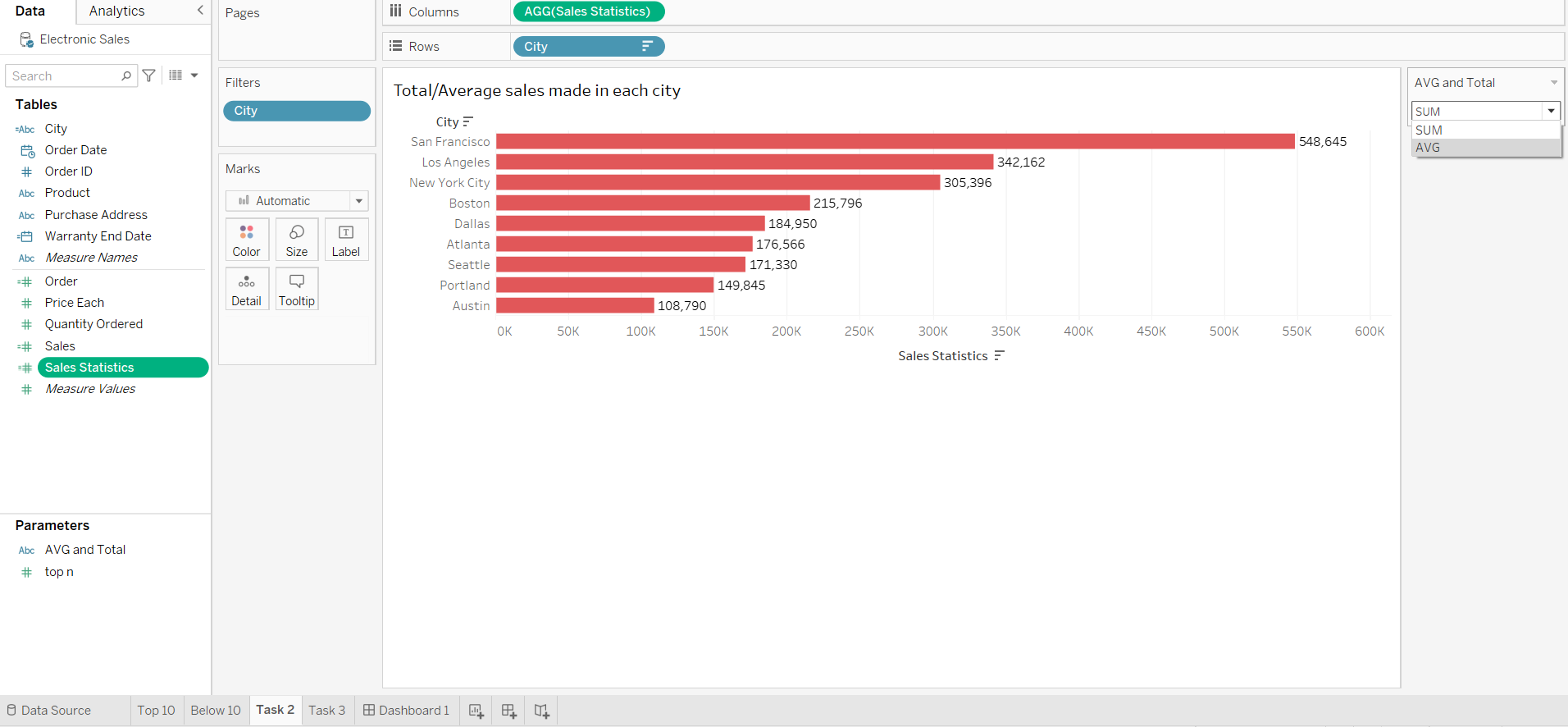
**Figure 50 Creating Parameters for AVG and SUM of sales**



**Figure 51 Create calculated fields as Sales Statistics**



**Figure 52 Total/Average sales made in each city**

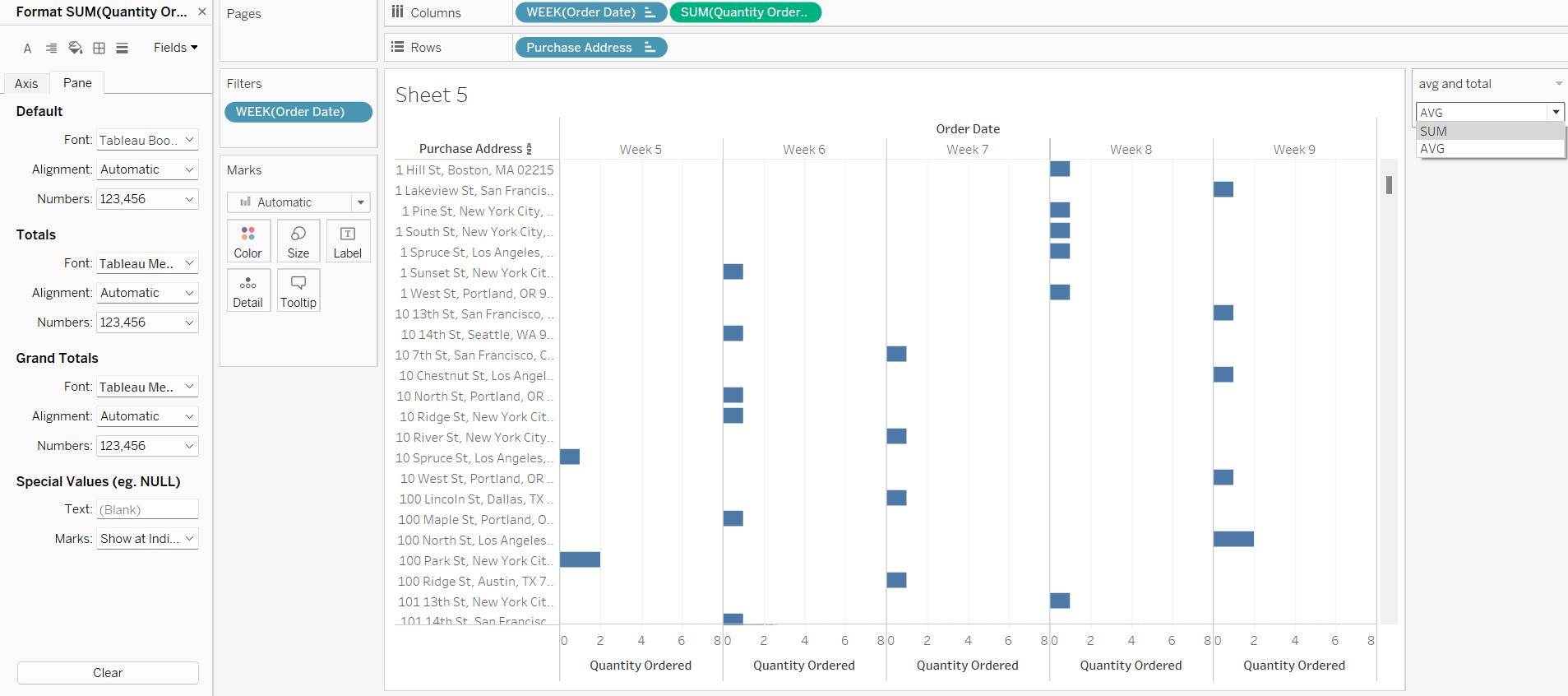


## **Task 3**

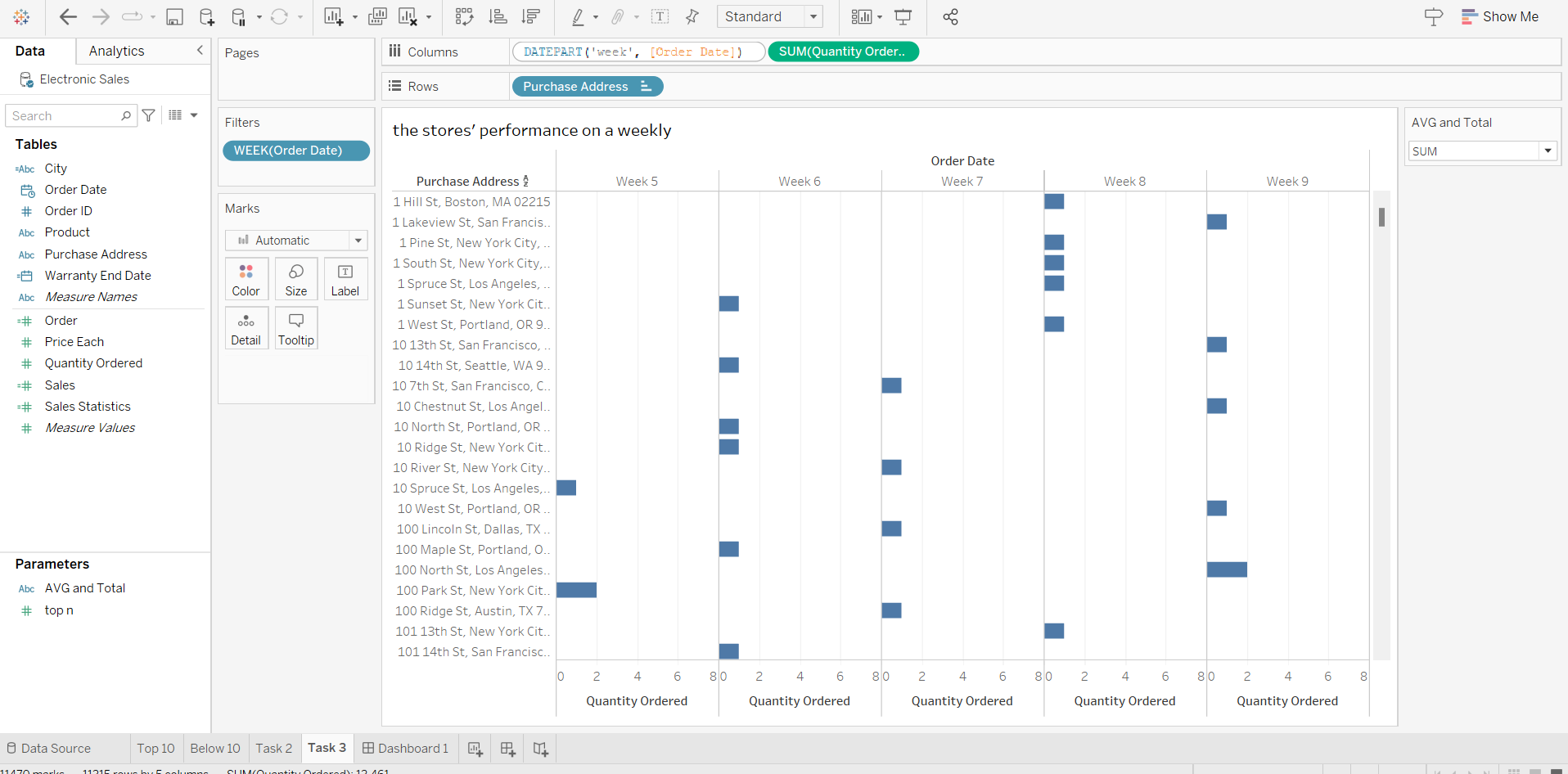
Step 1: Drop Order date, Quantity Order to Columns, and Purchase Address to Rows. And order date to filter to not show null values (Figure 53)

Step 2: Edit in Shelf for Order date with DATEPART function. (Figure 54)

**Figure 53 Step 1 for Task 3**



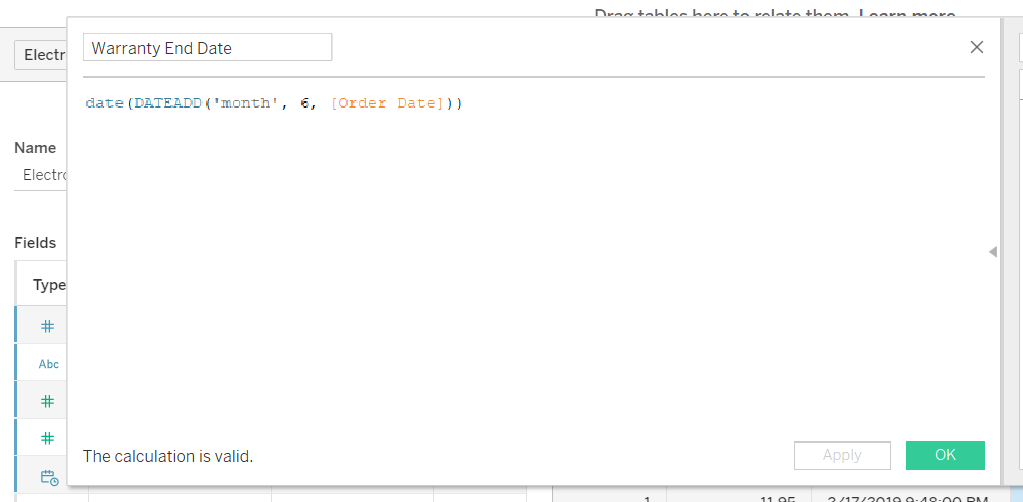
**Figure 54 the stores’ performance on a weekly basis**



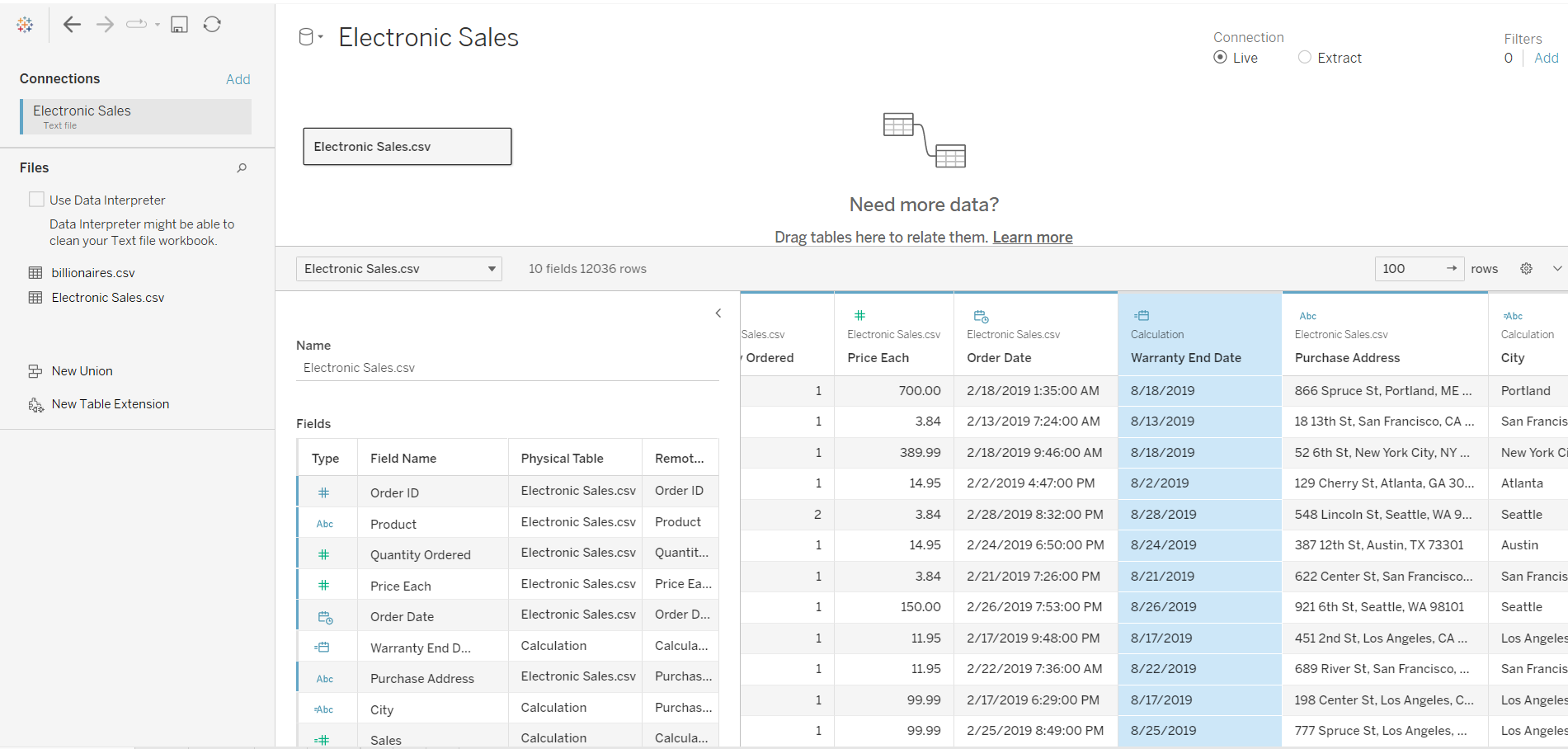
## **Task 4**

Step 1: Create calculated field with order date (Figure 55)

**Figure 55 Create calculated field for Warranty End Date**



**Figure 56 Warranty End Date**



## **Task 5**

Step 1: Create a new dashboard

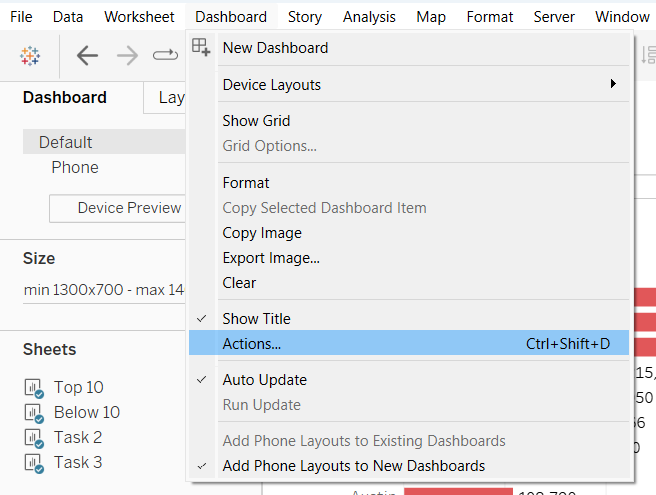
Step 2: Drag the four sheets that created for task 1,2,3 to the dashboard

Step 3: Arrange them in a layout that makes sense for the data.

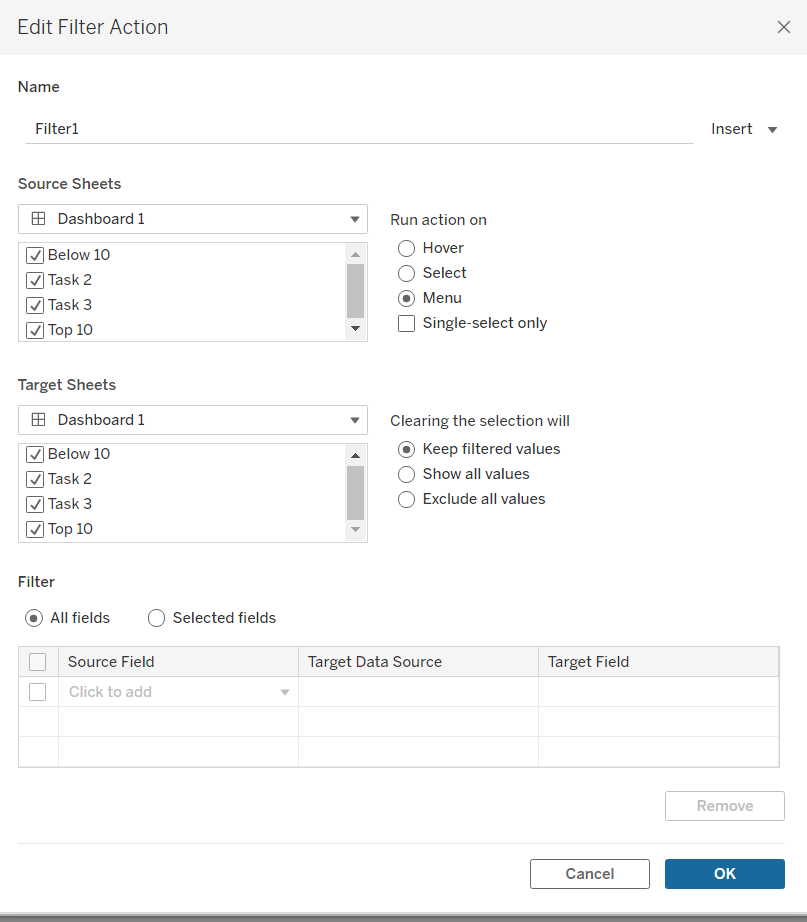
Step 4: Click on the "Actions" button and select "Add Action", than select "Filter". (Figure 57)

Step 5: Choose the sheets in the dashboard that want to link to. (Figure 58)

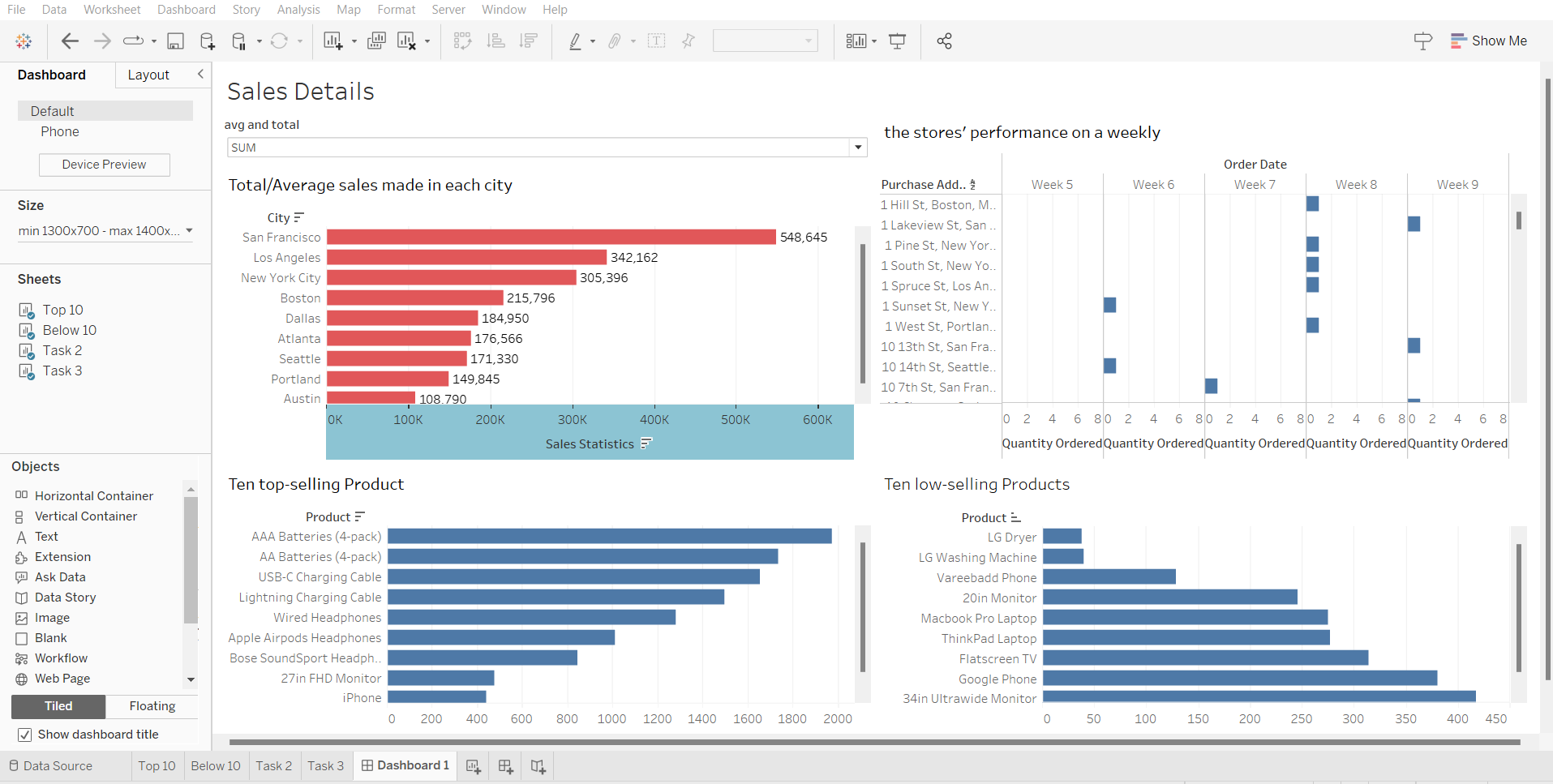
**Figure 57 Adding Actions to link the sheets together**



**Figure 58 Choosing the sheets in the dashboard that want to link to**



**Figure 59 Sales Details Dashboard**



# **References**

George, D. and Mallery, P. (2010) *SPSS for windows step by step: A simple study guide and reference, 17.0 update*. Boston, MA: Allyn & Bacon.

Google (no date). *Google colab*. Google. Available at: https://research.google.com/colaboratory/faq.html (Accessed: April 25, 2023).

Harrison, M. and Petrou, T. (2020) “1,” in *Pandas 1.x cookbook: Practical recipes for scientific computing, Time Series analysis and exploratory data analysis using Python*. Birmingham, UK: Packt Publishing, pp. 6–9.

Holgersson, H.E. (2006) “A graphical method for assessing multivariate normality,” *Computational Statistics*, 21(1), pp. 141–149. Available at: https://doi.org/10.1007/s00180-006-0256-9.

Kumar, A. and Panda, S.P. (2019) “A survey: How python pitches in IT-world,” *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* [Preprint]. Available at: https://doi.org/10.1109/comitcon.2019.8862251.

Mishra, P., Pandey, C.M., Singh, U., Gupta, A., Sahu, C. and Keshri, A., 2019. Descriptive statistics and normality tests for statistical data. *Annals of cardiac anaesthesia*, *22*(1), p.67.

Peterson-Withorn, C. (2023) *Forbes' 37th Annual World's billionaires list: Facts and figures 2023*, *Forbes*. Forbes Magazine. Available at: https://www.forbes.com/sites/chasewithorn/2023/04/04/forbes-37th-annual-worlds-billionaires-list-facts-and-figures-2023/?sh=172891b977d7 (Accessed: April 25, 2023).

Seaborn (no date). *Statistical Data Visualization#* (no date) *seaborn*. Available at: https://seaborn.pydata.org/ (Accessed: April 25, 2023).

Warren, J. and Marz, N., 2015. *Big Data: Principles and best practices of scalable realtime data systems*. Simon and Schuster.