



**Module Code & Module Title**

**CC5067NP Smart Data Discovery**

**Assessment Weightage & Type**

**60% Individual Coursework**

**Year and Semester**

**2nd year, 2nd semester**

**Student Name: Vamsha Palja Tamu**

**Group: L2C3**

**London Met ID: 21050019**

**College ID:** **NP04CP4A210106**

*I confirm that I understand my coursework needs to be submitted online via Google Classroom under the relevant module page before the deadline in order for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a marks of zero will be awarded.*

**Table of Contents**

[1. Introduction 1](#_Toc133932569)

[2. Objectives 5](#_Toc133932570)

[3. Development 6](#_Toc133932571)

[ Data understanding 6](#_Toc133932572)

[ Data preparation 7](#_Toc133932573)

[ Data Analysis 12](#_Toc133932574)

[ Data Exploration 16](#_Toc133932575)

[4. Conclusion 24](#_Toc133932576)

[5. References 26](#_Toc133932577)

**Table of figures**

[Figure 1: image of data processing cycle 3](#_Toc133932578)

[Figure 2: merging all monthly reports into one final report 7](#_Toc133932579)

[Figure 3: checked if there was null values and removed them 8](#_Toc133932580)

[Figure 4: checking the data type of columns 9](#_Toc133932581)

[Figure 5: checked if there were any repeating headings 9](#_Toc133932582)

[Figure 6: removing the duplicated headers and converting the respective columns to integer 9](#_Toc133932583)

[Figure 7: splitting the month from ordered date and converted to integer 10](#_Toc133932584)

[Figure 8: extracting city name from purchased address 11](#_Toc133932585)

[Figure 9: statistics of sum, mean, standard deviation, skewness, and kurtosis 12](#_Toc133932586)

[Figure 10: calculating correlation of all variables 14](#_Toc133932587)

[Figure 11: showing correlation of all variables 15](#_Toc133932588)

[Figure 12: which month had the best sale and what was the earning 16](#_Toc133932589)

[Figure 13: best sale being represented on a graph 16](#_Toc133932590)

[Figure 14: graph on which month had the highest sale 17](#_Toc133932591)

[Figure 15: which city has highest product sold 18](#_Toc133932592)

[Figure 16: graph on which city had sold the most product 19](#_Toc133932593)

[Figure 17: which product was sold most being calculated 20](#_Toc133932594)

[Figure 18: graph on which product was sold the most 21](#_Toc133932595)

[Figure 19: calculating the numbers of products that were sold 22](#_Toc133932596)

[Figure 20: histogram graph on which product was sold the most 23](#_Toc133932597)

**Table of Tables**

[Table 1: describing the columns present in csv files 6](#_Toc133932598)

# Introduction

Smart data refers to the process of extracting meaningful insights from vast amounts of an existing raw, passive and unstructured form of data through the use of sophisticated algorithms and techniques to turn it into useful information that we can understand and improve and drive forward (Hosseinpour, et al., 2016).

Data discovery is the process of prying meaningful patterns from data. It does this by collecting data from a variety of sources and then applying advanced analytics to it to identify specific patterns or themes (TIBCO, n.d.). Data is being generated at an unprecedented rate, Data Discovery brings data sources together from different sources and analyses them to uncover deeper patterns, trends and insights, giving organizations a comprehensive view of their data. Such a view provides better insights into the available data, which in turn leads to better-informed decisions that help leaders become more agile in their day-to-day operations, helping them make better decisions that impact the business and help the business thrive toward sustained success and gain a competitive advantage in their industry by adapting to the dynamic market and making better decisions and gain a competitive edge in their industry.

Data could be broadly defined as a collection of raw facts and figures that have no inherent meaning or context. But in a business context, data is the collective information related to a company and its operations, which can be generated from a variety of sources including internal business systems, external databases, social media platforms, statistical information, raw analytical data, customer feedback data, sales figures and other information.

For example, a business may conduct surveys or focus groups to gather information about customer preferences or opinions. Another way to get data is through secondary research, which collects data from existing sources such as industry reports or government statistics.

Data could be broadly defined as a collection of raw facts and figures that have no inherent meaning or context. But in a business context, data is the collective information related to a company and its operations, which can be generated from a variety of sources including internal business systems, external databases, social media platforms, statistical information, raw analytical data, customer feedback data, sales figures and other information (Indeed Editorial Team, 2023). For example, a business may conduct surveys or focus groups to gather information about customer preferences or opinions. Another way to get data is through secondary research, which collects data from existing sources such as industry reports or government statistics.

Data processing refers to the conversion of raw data into meaningful information (Lahore College for Women University, n.d.). Data processing involves many steps: data gathering, data cleansing, data entry, data processing, and output. Data processing involves removing errors and inconsistencies from the datasets and combining data from multiple sources into a single dataset. Output involves using statistical methods and algorithms to identify patterns and trends in the data, and presenting the results of the analysis in a clear and concise manner.

Data processing is important for several reasons. First, it can help companies make informed decisions based on accurate and up-to-date information. By processing and analyzing data, companies can gain insights into customer behavior, market trends, and operational efficiencies that can inform organizations to develop better business strategies through strategic planning and decision-making to increase their competitive advantage over others. Finally, data processing can help companies improve their bottom line by identifying areas for cost savings or revenue increases.

Data processing is a crucial component of any successful business strategy. By using data to generate insights and make informed decisions, companies can gain a competitive advantage in their industry and achieve sustainable success. Therefore, investing in data processing tools and techniques is essential for companies looking to stay on top in today's data-driven economy.

The data processing cycle is a series of steps that are used to convert raw data into useful information (Mathur, 2022). These steps include data gathering, data preparing, entering data, data processing, outputting data, and storing data.

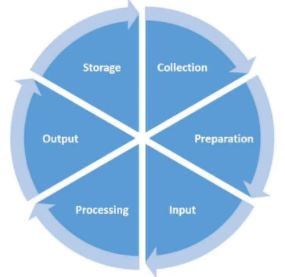


Figure 1: image of data processing cycle

In the data processing cycle, the first step is to collect data from various sources for processing. The second step is to clean the dataset to remove errors and inconsistencies from the dataset. The third step is to merge different datasets into one large dataset for easier processing. The fourth step is to enter the raw collection of the data set ready for further processing. The fifth step is to put the huge collection of raw data into a meaningful form. The sixth step is to get the output of valuable information from processing the data set. Finally, the output data set is stored for later future use, for easy access and retrieval of information when needed, as well as for immediate use as input in the next data processing cycle.

The data processing cycle is a critical component of any successful business strategy. By using data to generate insights and create better business strategies through strategic planning and decision making, companies can gain a competitive advantage in their industry and achieve sustained success. Therefore, investing in data processing tools and techniques is essential for companies looking to stay on top in today's data-driven economy.

# Objectives

* to learn how to prepare a dataset for Businesses Analytics and machine learners
* to develop critical thinking skills, to better solve problems that may arise during dataset preparation
* to work as a data scientist for companies, to analyze data for companies to increase their competitive advantage Others

# Development

## Data understanding

Using data from company ABC, it contained the sales data for each month of 2019, which products were sold to the customer, in what quantity, on what date and time, and where to ship them. It contained attributes such as order ID, product, order quantity, price per unit, order date and purchase address. There were some mixed null values ​​in the data set.

|  |  |  |  |
| --- | --- | --- | --- |
| s.no | Column name | Description | Data type |
| 1 | Order ID | This column contains the id number of orders that was made by the customer | Object |
| 2 | Product | This column contains the names of product that were sold | Object |
| 3 | Quantity Ordered | This column contains the quantity of products that was ordered by the customer | Object |
| 4 | Price Each | This column contains the price for that product | Object |
| 5 | Order Date | This column contains the date and time of when the order was made. | Object |
| 6 | Purchase Address | This column contains customer’s personal address that the product was to be delivered | Object |

Table 1: describing the columns present in csv files

## Data preparation

* Write a python program to merge data from each month into one CSV and read in updated dataframe.

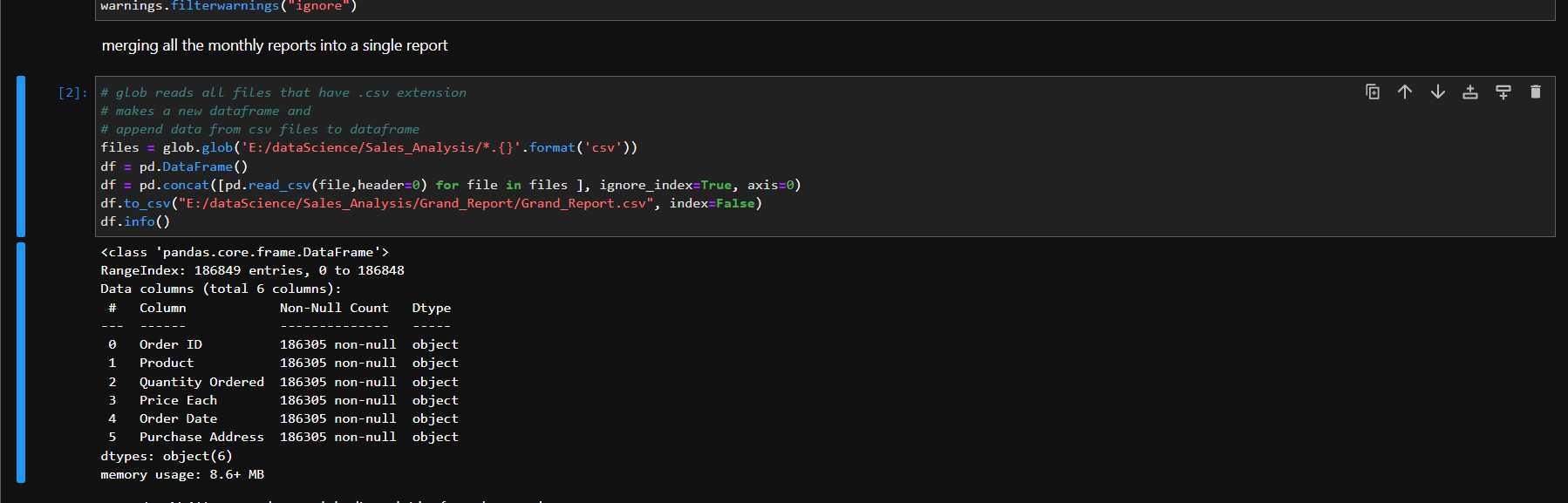


Figure 2: merging all monthly reports into one final report

As shown in the above picture, monthly reports are being converted into one single grand file named “Grand\_Report” and its format is CSV.

* Write a python program to remove the NaN missing values from updated dataframe.



Figure 3: checked if there was null values and removed them

In above block of code, null values are checked and it there are any then it is removed from the file. We could delete some rows that contain null values because we have data in abundance but for a smaller set of data that would not have been a good choice but we would have to replace data with some values such as mean, median or other way to surround that null data.

* Write a python program to convert **Quantity Ordered** and **Price Each** to numeric.

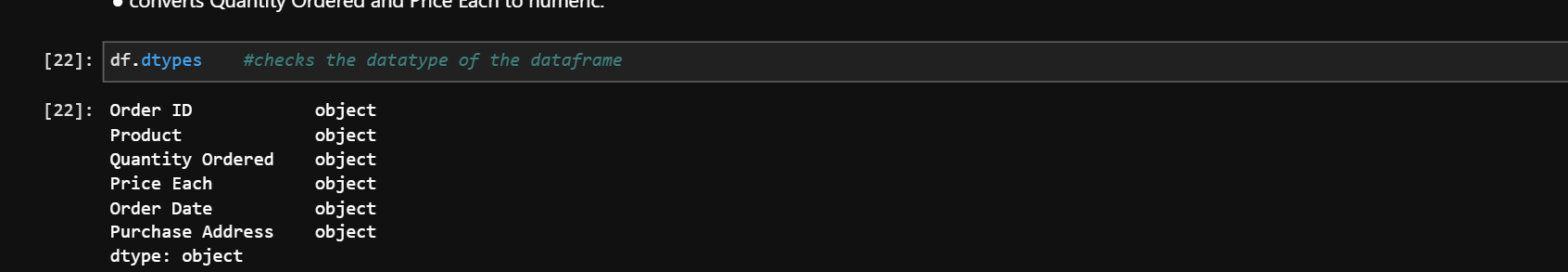


Figure 4: checking the data type of columns

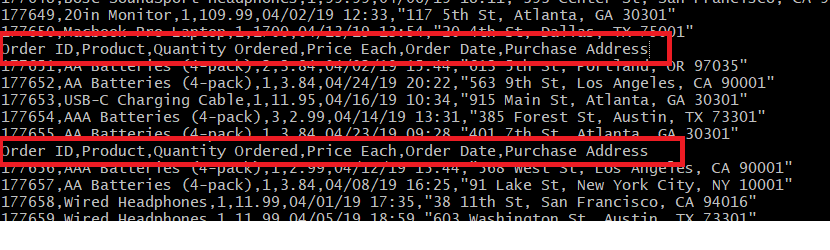


Figure 5: checked if there were any repeating headings

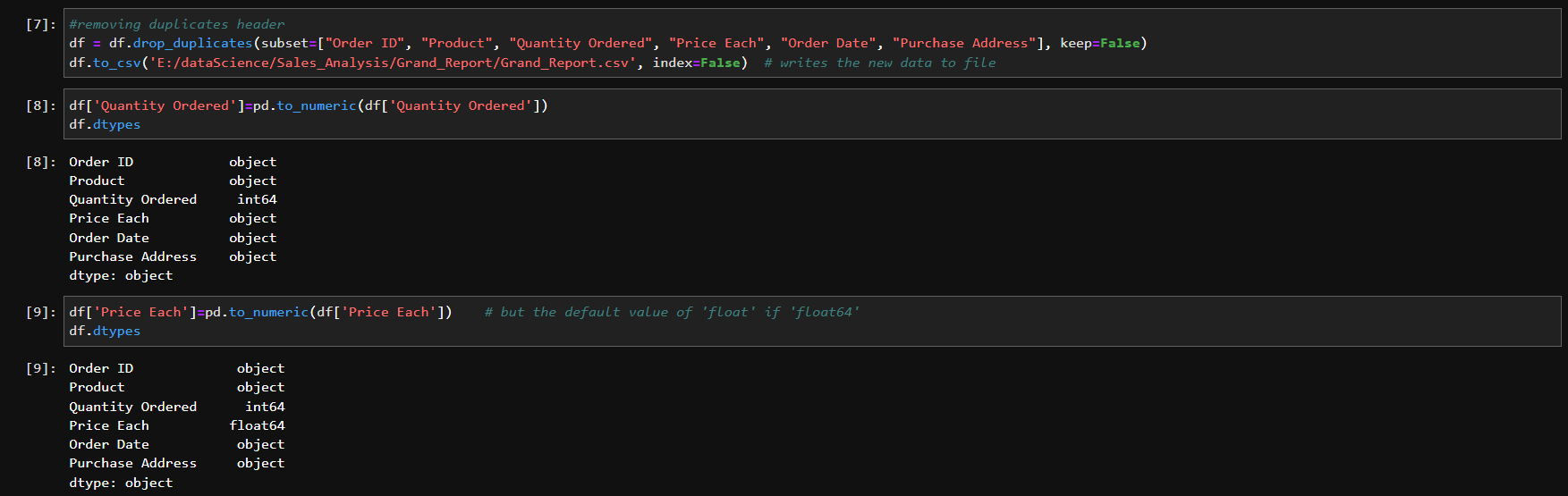


Figure 6: removing the duplicated headers and converting the respective columns to integer

Because there were duplicated heading kept as string, we needed to remove the duplicated headings first before converting the columns to integer as they were numbers. After conversion to integers, they are easier to perform the computation required for the analysis.

* Create a new column named **Month** from **Ordered Date** of updated dataframe and convert it to integer as data type.

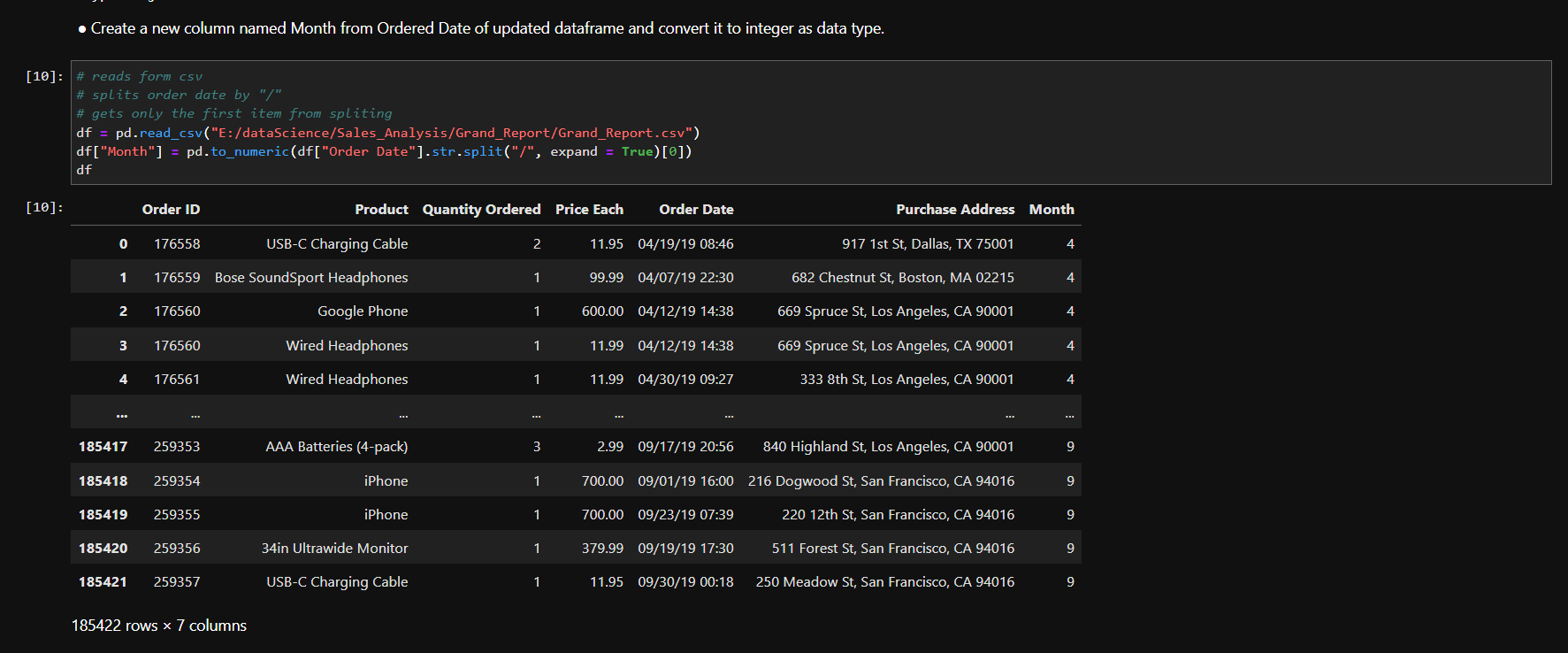


Figure 7: splitting the month from ordered date and converted to integer

In the above picture, month is extracted from “Ordered Date” and is transformed into integer as it indicates month for easy understanding and smoother analysis.

* Create a new column named **City** from **Purchase Address** based on the value in updated dataframe.

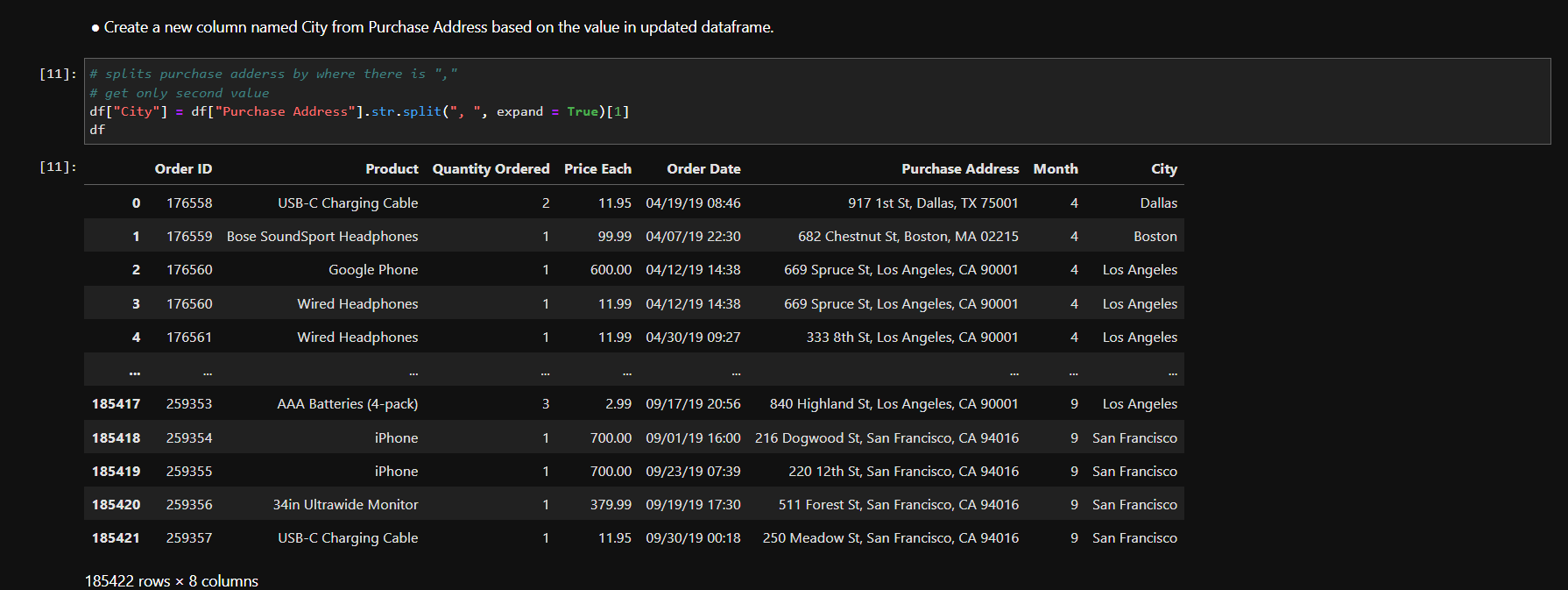


Figure 8: extracting city name from purchased address

In the above piece of code, the city name is extracted from “purchased address” and stored in another column to later find the city with the highest sales and the highest revenue.

## Data Analysis

* Write a Python program to show summary statistics of sum, mean, standard deviation, skewness, and kurtosis of any chosen variable.

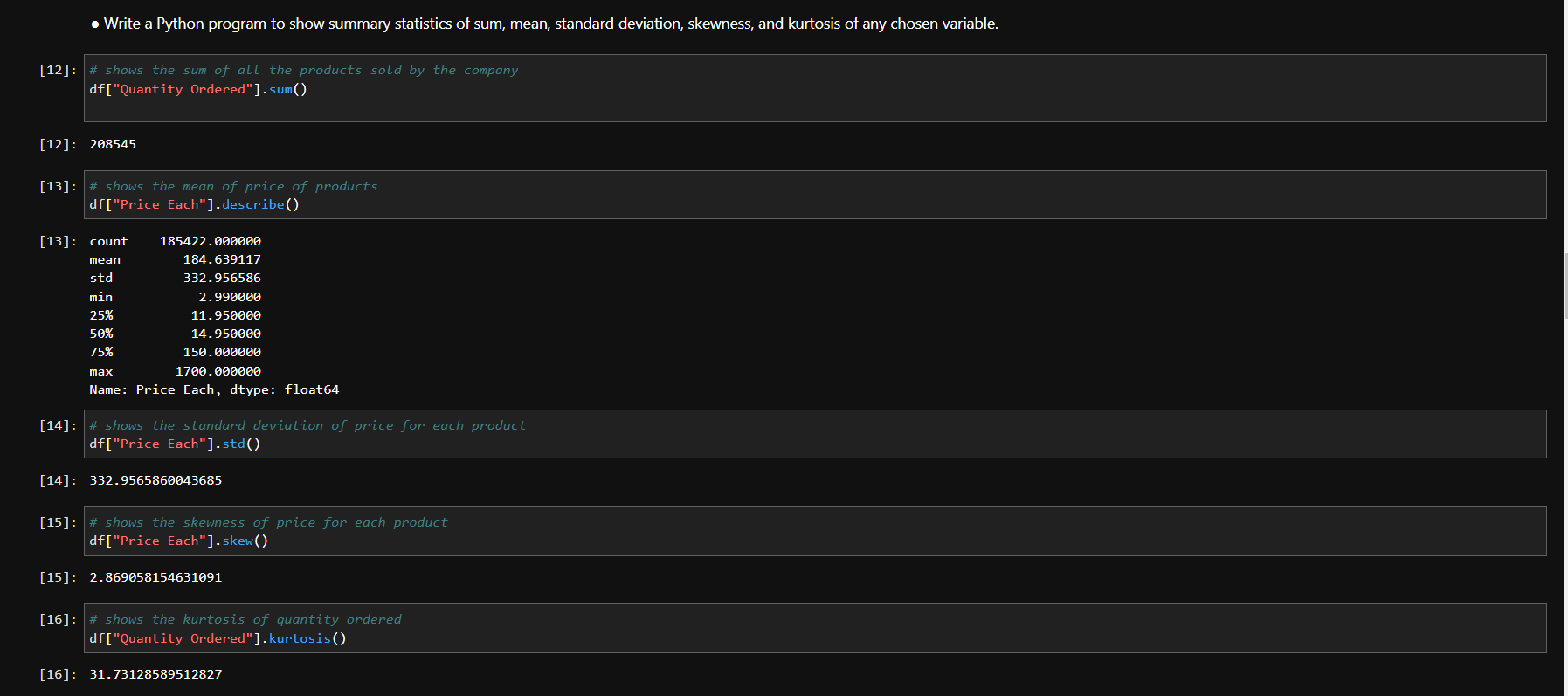


Figure 9: statistics of sum, mean, standard deviation, skewness, and kurtosis

In the above code, the number of products sold by the company in that year is expressed as follows: by adding all the numbers of products ordered by the customer. The mean of the data set is calculated as follows: The sum of the total value in the dataset is calculated as the mean the sum of the number of values. Mean provides a central tendency to represent typical values in the dataset.

Stander deviation measures the spread of data when it is plotted in a graph. It is calculated by dividing the variance by the mean square root of the two data points, which is the mean squared differences between each data point and the mean squared differences. A higher deviation value indicates that the spread of data is more and a lesser deviation value indicates the data are clustered tightly. We calculated the mean value of “quantity ordered” and the value came higher meant the data is spread widely.

Skewness represents the asymmetry of a distribution. A distribution is considered symmetric if both the right-hand side and the left-hand side have an equal number of data on both sides of the mean. We calculated the skewness value for “price each” and the value was found to be positive which meant majority of data in right-hand side of the mean to be heavy.

Kurtosis measures the distribution shape and describes the degree to which the data are peaked or flat. If kurtosis has a high value, then data peaked more than the normal distribution and if kurtosis has a low value, then data is flatter than the normal distribution. In this study, we calculated the kurtosis value of “quantity ordered” and we observed that the value increased as it is more peaked.

* Write a Python program to calculate and show correlation of all variables.

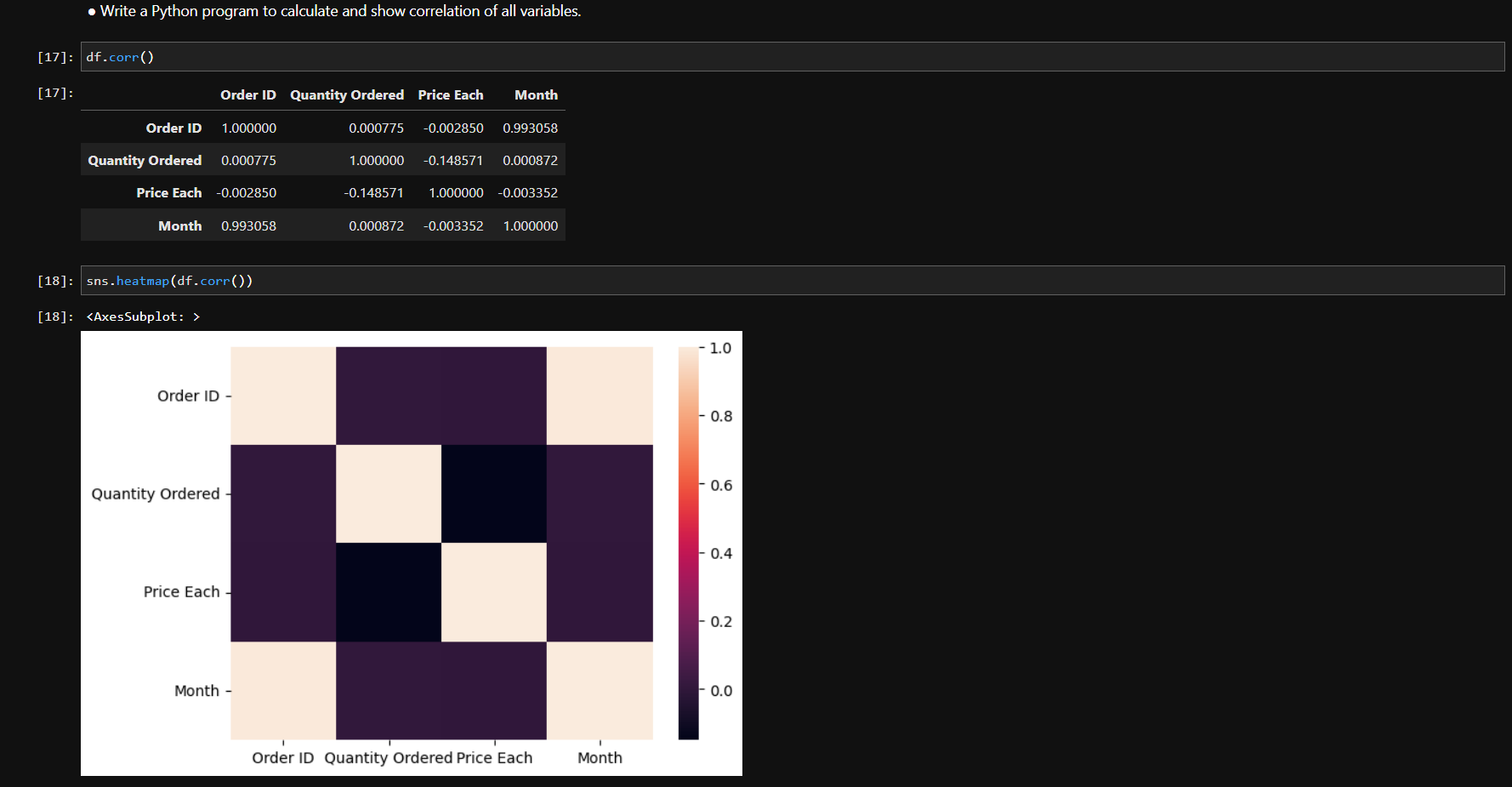


Figure 10: calculating correlation of all variables



Figure 11: showing correlation of all variables

In the above picture presented, correlation of all variables is being calculated and is being represented on a graph. Correlation can be decided as the measurement of relations between 2 or more variables in a dataset. At the time of the correlation calculation for the data frame, we observed that the returned range of values ranged from the neutral spectrum to the positive spectrum.

## Data Exploration

* Which Month has the best sales? and how much was the earning in that month? Make a bar graph of sales as well.

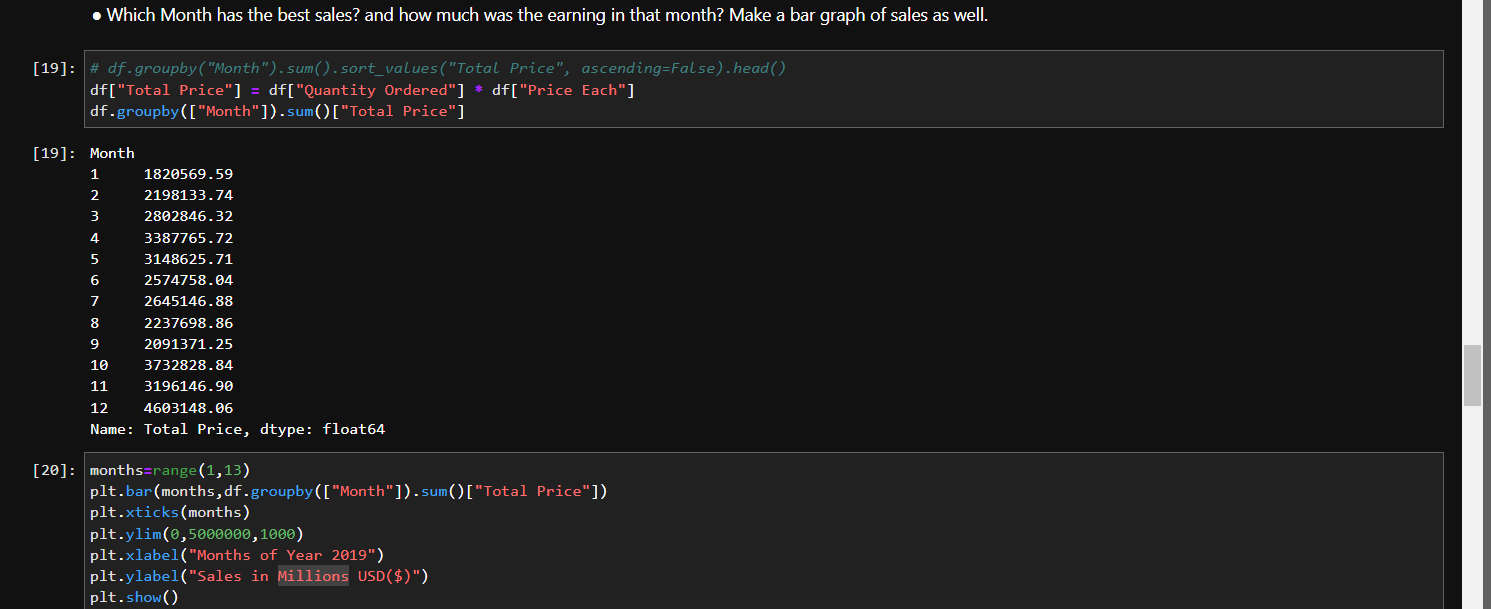


Figure 12: which month had the best sale and what was the earning

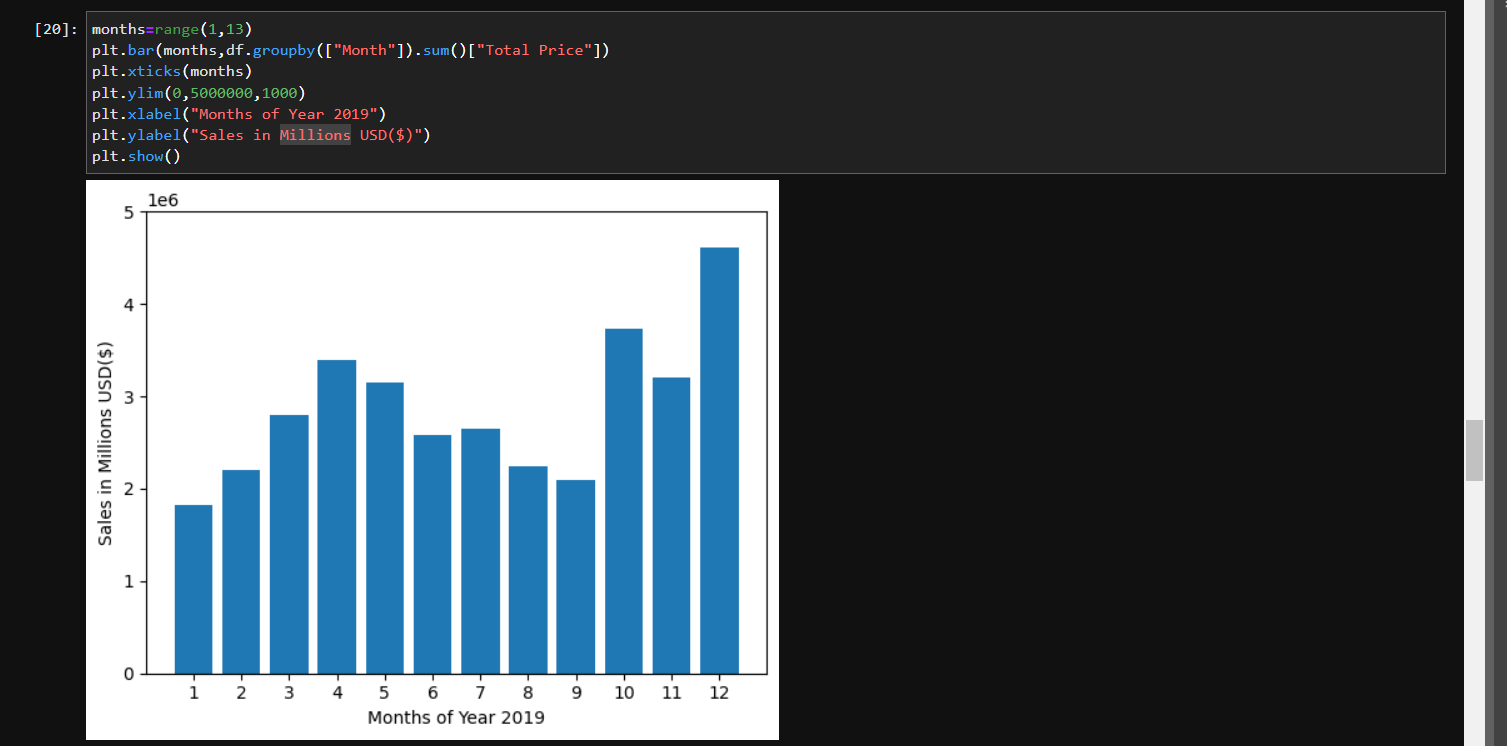


Figure 13: best sale being represented on a graph

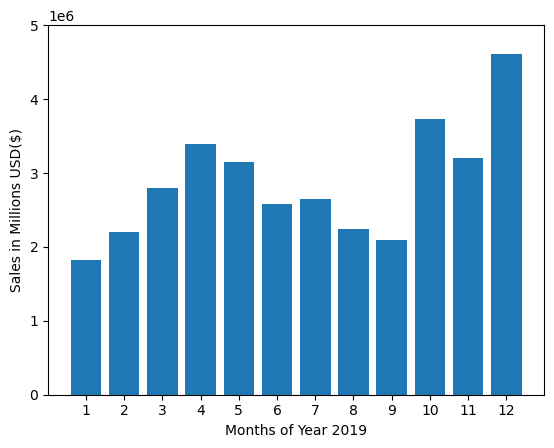


Figure 14: graph on which month had the highest sale

The image above calculates and graphs which month had the best sale. Revenues are shown vertically in millions of USD and months are shown separately horizontally. The chart showed the number of products sold in a different month.

* Which city has sold the highest product?

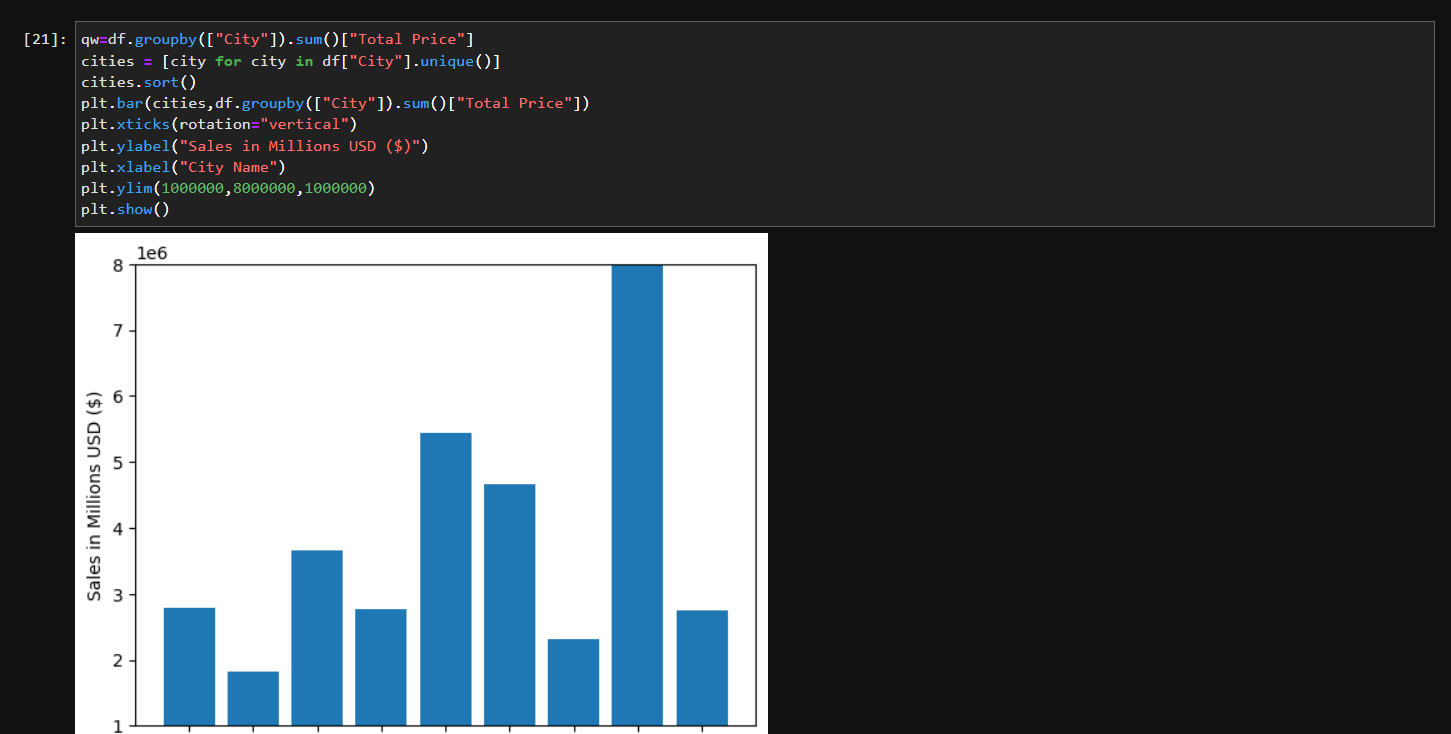


Figure 15: which city has highest product sold

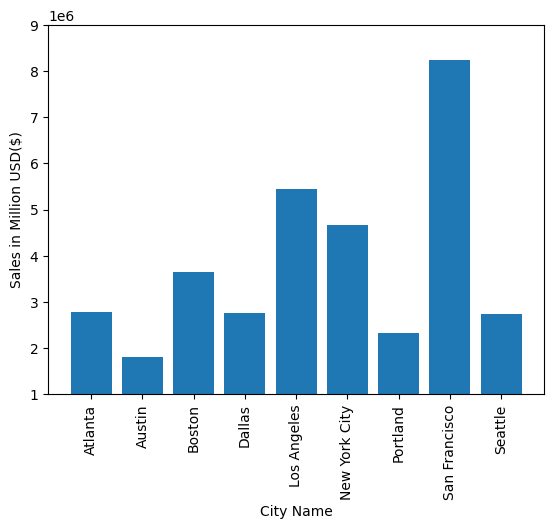


Figure 16: graph on which city had sold the most product

In the images above, the city with the highest product sales has been calculated and presented in a graph. Revenue is displayed vertically in millions of USD and city names are displayed horizontally. Different cities had different amounts of total sales where they sold the most and are shown in the chart

* Which product was sold the most in overall? Illustrate it through bar graph.

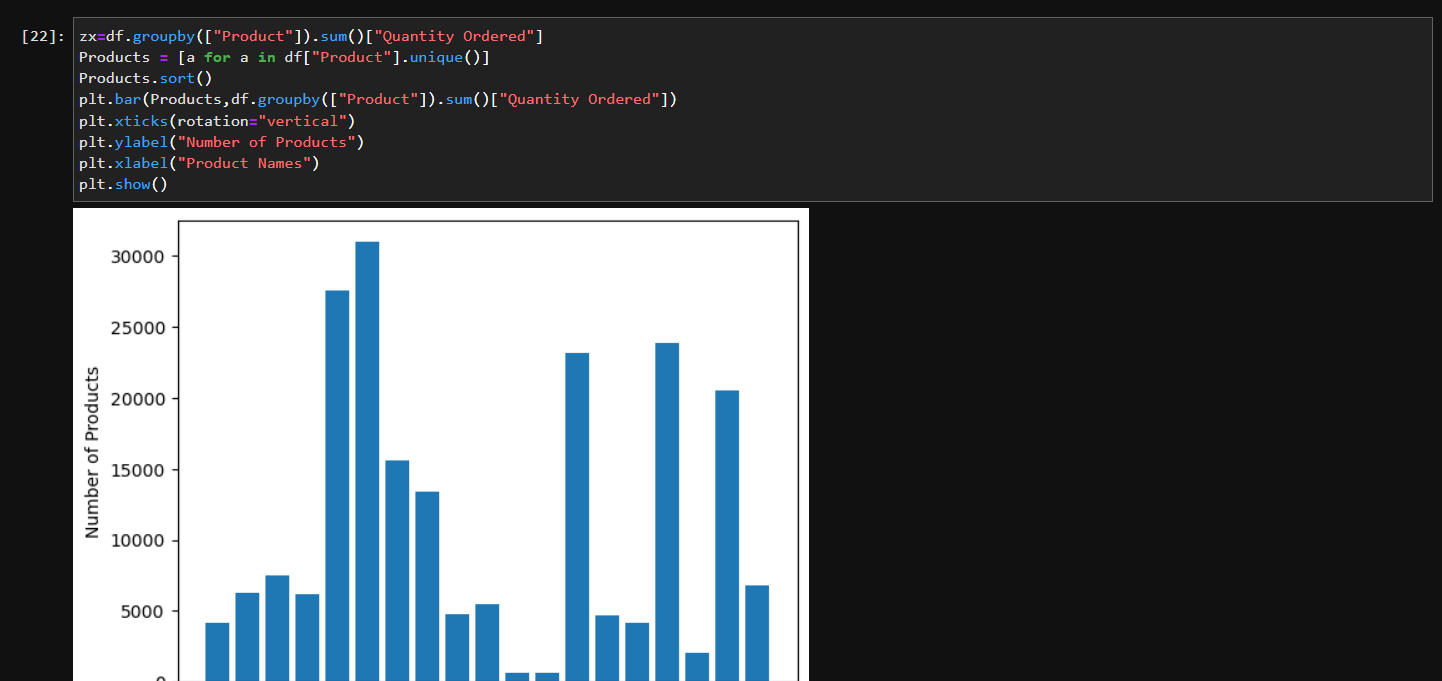


Figure 17: which product was sold most being calculated

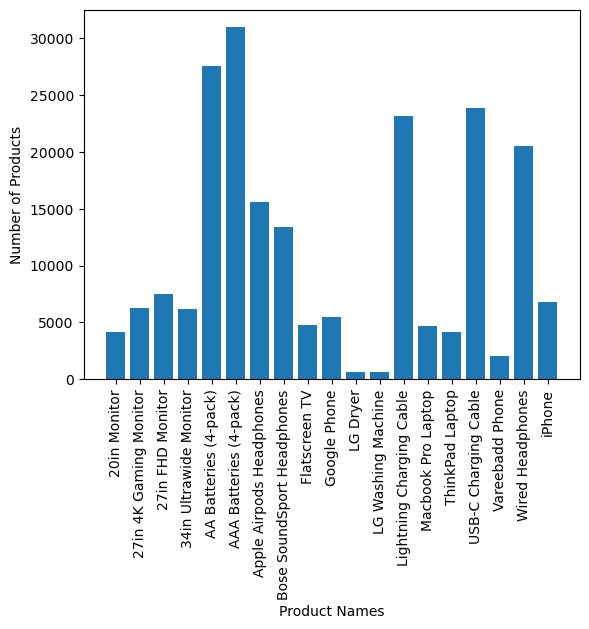


Figure 18: graph on which product was sold the most

In the images above, it is calculated which product sold the most and shown in the graph. The number of products sold is displayed vertically and the name of the products is displayed horizontally. Each type of product sold is recorded and in what quantity sold, they are plotted accordingly

* Write a Python program to show histogram plot of any chosen variables. Use proper labels in the graph.

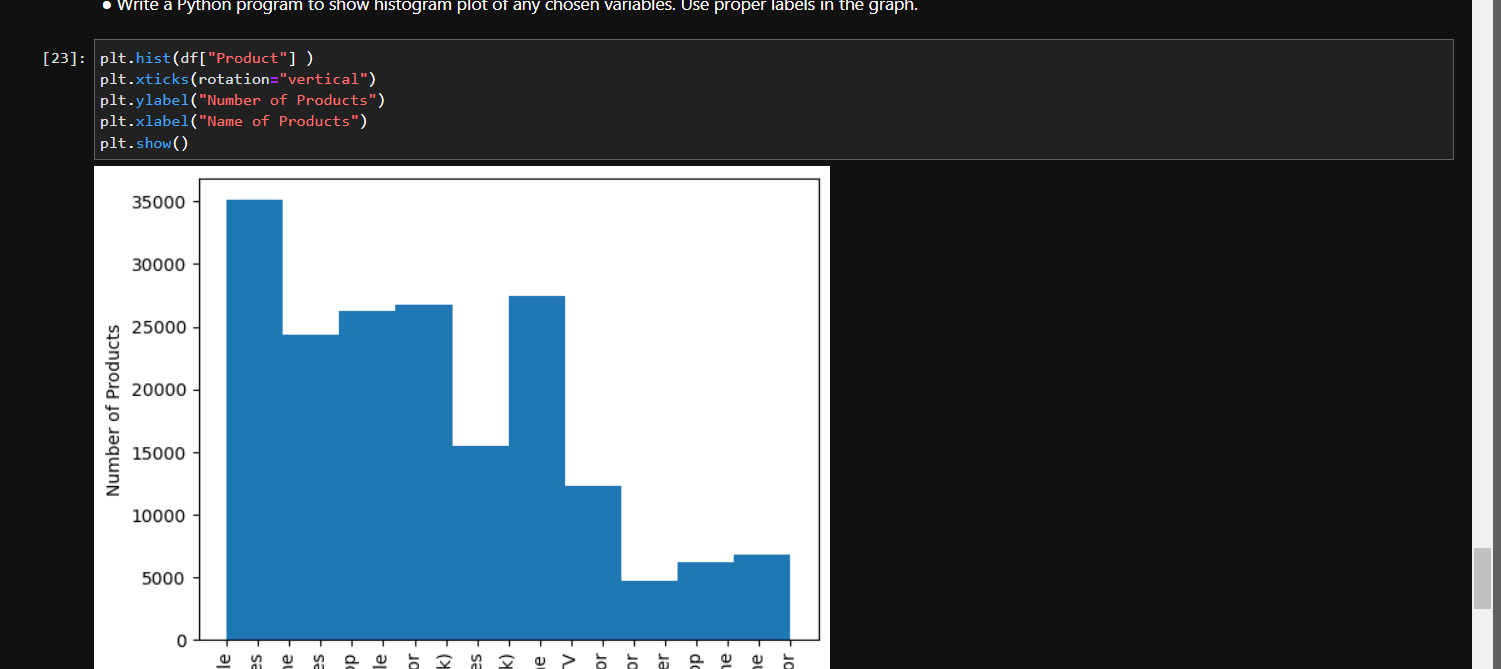


Figure 19: calculating the numbers of products that were sold

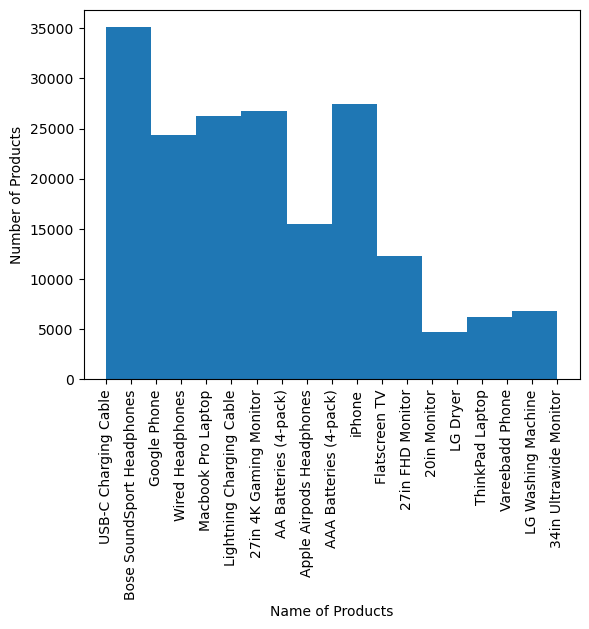


Figure 20: histogram graph on which product was sold the most

In the images above, it is calculated which product sold the most and shown in the graph. The number of products sold is displayed vertically and the name of the products is displayed horizontally. Each type of product sold is recorded and in what quantity sold, they are plotted accordingly in a histogram

# Conclusion

Through much effort, determination and research, the coursework was completed. This was an individual job that required us to process raw business data into information that was insightful for the company to achieve sustained success through an advantage over other competitors.

A lot of research was needed to get the right result. When I tried to convert the columns to integer and float respectively, I encountered a problem which meant that the column's data type could not be converted. I had searched the internet why I was getting this error when running a script on a Jupyter notebook and was also trying to figure out why. I had set up a few others who were also having the same problem and they were able to solve it online with the help of a stranger. I tried to implement the same script and still couldn't solve it. Unable to find the cause I dug into the CSV file and found that many headings kept repeating and since there was a string in the middle of the integer data and after removing the repeating heading I was able to convert the columns to integer and float.

While doing the scripting for the automation, I was able to get a new idea of ​​how I might use Python or other data processing programs. It also expands my knowledge of business analysis and how to use it to gain an advantage over other competitors. I've been able to learn a lot about Python, and this is what they are: I've been able to learn about various libraries that make computation easier, like Pandas and NumPy, learn about list comprehension, and implement it in scripting to see the imminent result it produces, and others.

In summary, the Data Processing and Business Analysis coursework was a valuable learning experience that required significant effort, determination, and research. It was an individual task that required processing raw business data into meaningful information that could be used to achieve competitive advantage and sustained success. There were challenges to overcome throughout the project, such as converting columns to integers and floats, but these have been overcome through research and creative thinking. The project also provided an opportunity to learn more about Python and its libraries like Pandas and Numpy, which made data processing more efficient. It expanded my knowledge of business analytics and how it can be used to make informed decisions that drive success in a competitive business landscape. Overall, this coursework demonstrated the importance of data processing and business analysis in achieving and sustaining success in the business world.

# References

Hosseinpour, F., Plosila, J. & Tenhunen, H., 2016. Smart Data: A New Perspective of Tackling the Big Data Phenomena Leveraging a Fog Computing System. *Smart Data: A New Perspective of Tackling the Big Data Phenomena Leveraging a Fog Computing System,* Volume 10, p. 124.

Indeed Editorial Team, 2023. *Indeed.* [Online]   
Available at: https://www.indeed.com/career-advice/career-development/data-in-business#:~:text=Business%20data%20is%20the%20collective,and%20other%20sets%20of%20information.  
[Accessed 3 April 2023].

Lahore College for Women University, n.d. *4DATAPROCESSING-converted.pdf.* [Online]   
Available at: https://ocd.lcwu.edu.pk/cfiles/Pharmacy%20Practice/Paper%20VIII/4DATAPROCESSING-converted.pdf

Mathur, V., 2022. *What is Data Processing and Why is it Important?.* [Online]   
Available at: https://www.analyticssteps.com/blogs/what-data-processing-and-why-it-important

TIBCO, n.d. *What is Data Discovery?.* [Online]   
Available at: https://www.tibco.com/reference-center/what-is-data-discovery  
[Accessed 12 04 2023].