**PRODUCT SALES ANALYSIS**

**Team members**

* **511921104008 - Arshvanth .V**
* **519921104040 -Musharaf Sharif .M**
* **511921104037 - Moha Rishi .S**
* **511921104017 -Dharneshwaran .H**

**Phase 4 submission document**

**What city sold the most product.**

To answer this question, obviously we need to create a new column called “City” column. How do we get that? As usual, we’re gonna check the top 5 data in our dataframe to figure out where can we get our “City” column using *.head()* method.

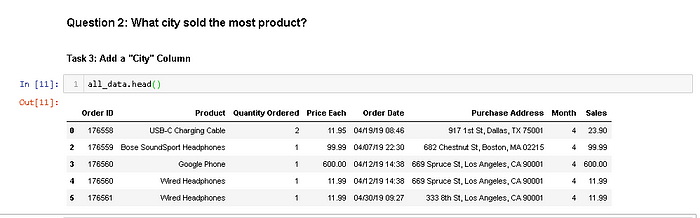


Figure 16. Showing Our Top 5 Updated Dataframe

As you can see at Figure 16, the “Purchase Address” Column contain the city. We can’t get it directly, we need to extract the data. We can use one of most useful function in pandas, *.apply()* method.

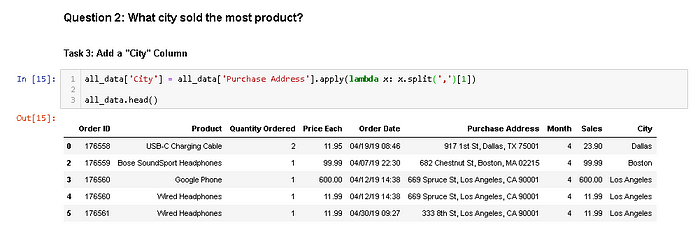


Figure 17. Using .apply() Method to Extract The Data

To make it neater, we can use this

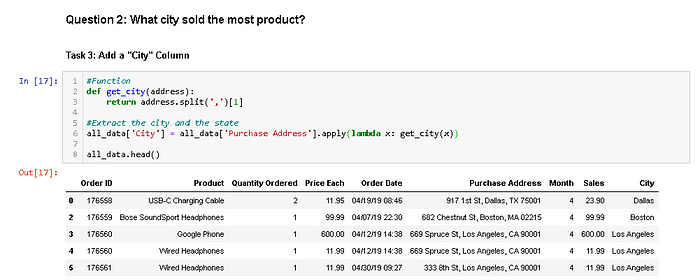


Figure 18. Using def function to Make The Code Neater

apply and lambda usually used to create new column based on other column (example .apply(lambda x: x\*2. It means every input x in other column will be changed to x\*2 in a new column). In this case we create “City” column based on “Purchase Address” column and we split the data into 3 part. The first one is before the first comma (index = 0), the second one is between the commas (index = 1), and the third on is after the last comma (index = 2). As we need to extract the city data, we use [1] to state it to index 1.

As you can see at Figure 17, we successfully created a “City” column. So are we ready to answer the second question? Not yet. We get an issue here. It’s not error, it’s the value of the “City” Column. This is just a rare case when there are 2 cities are named exactly the same. Example someone in New England and someone in West Coast would think Portland in different way. Someone in New England thinks Portland as Portland Maine and someone in West Coast thinks Portland as Portland Oregon. So in our dataset we actually had the overlapping cities between these two. So, we should also grab the state.

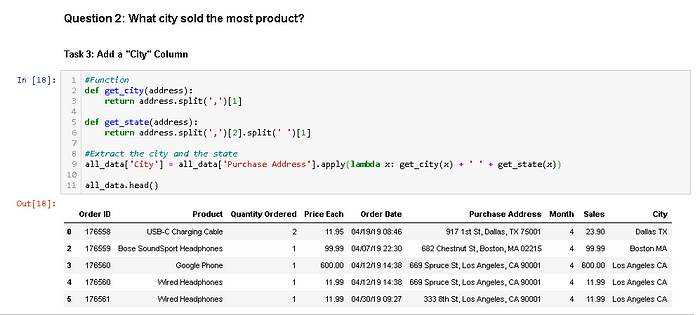


Figure 19. Extracting the state to “City” Column

The function get\_state() basically works as explained before. But in this function, we seperate the data again become three parts. The first one is before the whitespace (index = 0), the second one is between the whitespaces (index = 1), and the third one is after the last whitespace (index = 2). So that’s why we use *.split(‘ ‘)[1*] in the second split.

Now we’re ready to answer the second question, **what city sold the most product**? As we did before, we’re gonna group it by the city and summing all the values based on the group.

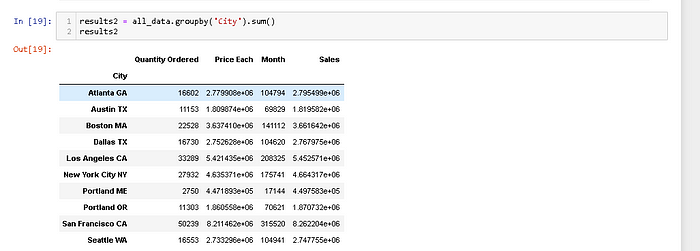


Figure 20. Grouping The Dataframe by The City

It’s too messy, but if you look carefully you can see that San Fransisco is the highest sold product of all cities with approximately $8,200,000. We clearly need to visualize it because it’s so hard to conclude anything just based on that numbers and also it will make our bussiness partner easier to understand.

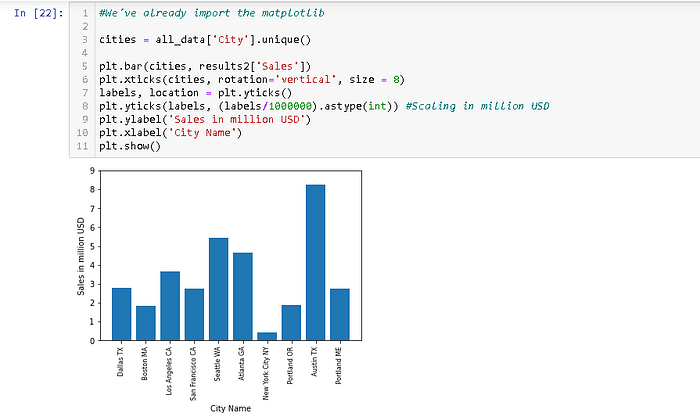


Figure 21. Plotting the Sales Grouped by Cities.

Now we successfully plot it. But there is a big issue here. If you notice that the values (Figure 20) and the plot (Figure 21) are not synchronized. The highest sales should be San Fransisco. What’s wrong with our code?

There’s an issue between *.unique()* method and *plt.bar()*. Their cities order are different. we’re gonna syncronized the order by simply fixing the variable ‘cities’.

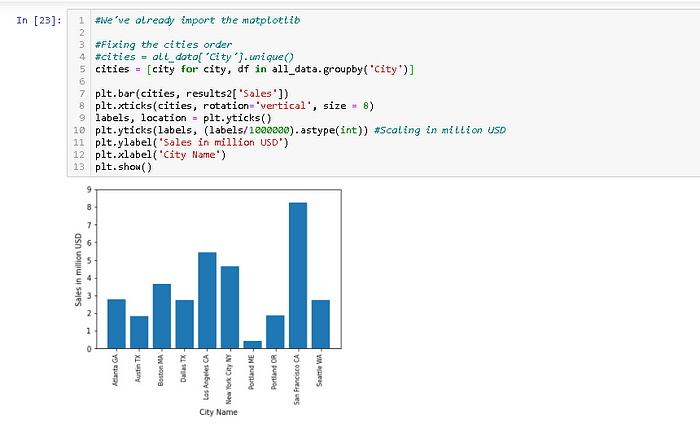


Figure 22. Fixing The Code

Now, we fix the issue and successfully plot it. As a data scientist, we need to figure out why San Fransisco is the highest sale compare to other cities. Maybe Sillicon Valley need more electronic products. Maybe the advertisement is better in San Fransisco.

We can use this data to improve the sales of bussiness.

**What Time Should We Display Advertisements to Maximize Likelihood of Customer’s Buying Product.**

As usual, to remind what our data look like, we use the .head() method to show the top 5 of our updated dataframe.

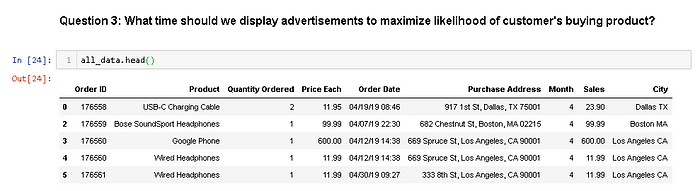


Figure 23. Showing our Top 5 updated dataframe

If we’re gonna use our data to answer this question, we need to aggregate the period in 24 hours distribution. Look carefully at Figure 23. In “Order Date” column, there are times data. We could extract it like we did before.

But to make it more consistent, we need to convert the “Order Date” Column into date time object.

We’re gonna use *pd.to\_datetime()* method.

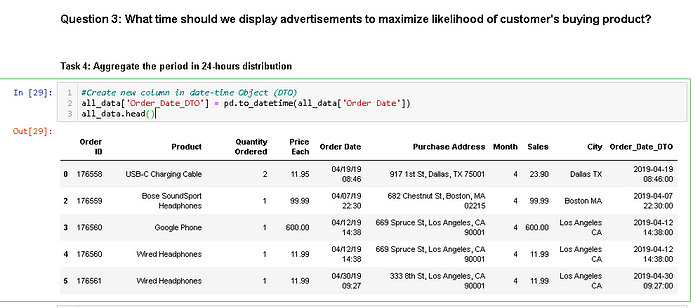


Figure 24. Converting the “Order Date” Column into Date-Time Object

It will take a little bit longer because of the heavy calculation. Now we can create a new column called “Hour” contain the extraction of “Order\_Date\_DTO” data.

We only need the hours data, so we can extract them by doing this.

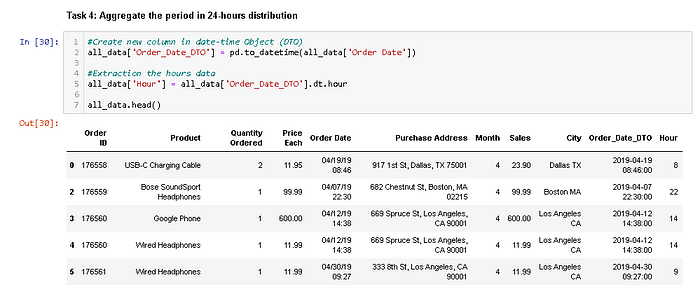
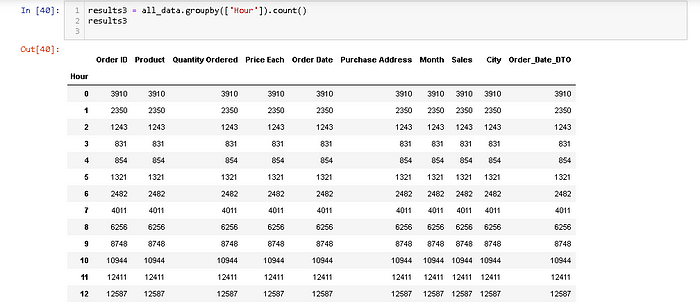


Figure 25. Extracting The Hours Data Into The New Column

Now we can answer the third question, **what time should we display advertisements to maximize likelihood of customer’s buying product?**To answer this, we’re gonna group it by the hours and counting all of the orders.



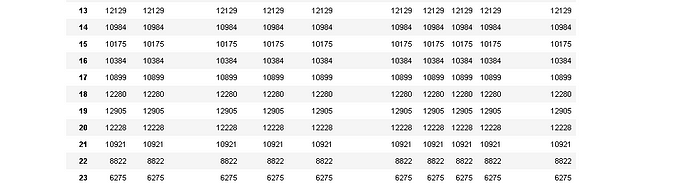


Figure 26. Grouping the data by the hours

If we want to answer the third question, we only need the “Quantity Ordered” column. Now let’s visualize it. We want it to be the line chart because this spesific data (hours) are more logical to show using line chart than bar chart because the data has to be continue.

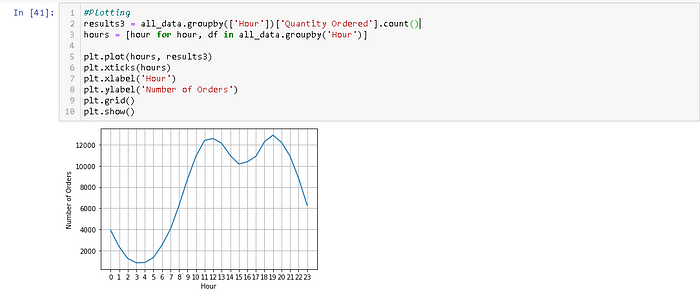


Figure 27. Visualizing the Number of Orders in 24 hours format

As you can see from Figure 27, there are approximately 2 peaks at the data. They are 12 (12 PM) and 19 (7 PM). It makes sense since most people shopping during the day.

From this data, we can suggest to our bussiness partner to advertise their product right before 12 PM and/or 7 PM. It could be 11.30 AM and/or 6.30 PM.

Remember, this chart is the total orders of **all cities**. Maybe you could make a spesific chart for a spesific city and planning the advertisement better for that city.

**Conclusion :**

In this phase ,which city may sold most products and What Time Should We Display Advertisements to Maximize Likelihood of Customer’s Buying Product. for further queries we may answer in next phase which is gonna final phase in the project. In that final phase this section you will document the complete project and prepare it for submission.