Project Report

Submitted By: Lareb Amir (21K-4173) & Maham Sabir

# Objective of the Project:

As of the course, the objective of the project is to learn and implement the to tools and techniques of Data Science on the Data Set, and to observe and critically analysis the data to make it useful for the stakeholders, for further development and progress of consumers.

# Source of data

Kaggle, US E - Commerce 2020 data

# Brief description of the data

Data Comprise of Work-hold stuff.

* Having 3 main **Categories** in store, **Furniture, Technology** and **Office Supplies,** each having **Sub-Category** **and Product Line** respectively. Overall store has 17 **Sub-Category,** 9 in **Office Supplies,** 4 in **Furniture,** and 4 in **Technology.** And each Sub-Category has its product range.

|  |  |
| --- | --- |
| **Categories** | **Sub Categories** |
| Furniture | Chairs, Table, Furnishing, Book cases, |
| Technology | Accessories, Phones, Copier, Machines |
| Office Supplies | Paper, binder, Supplies, Art, Envelopes, Storages, Fasteners, Labels, Appliances |

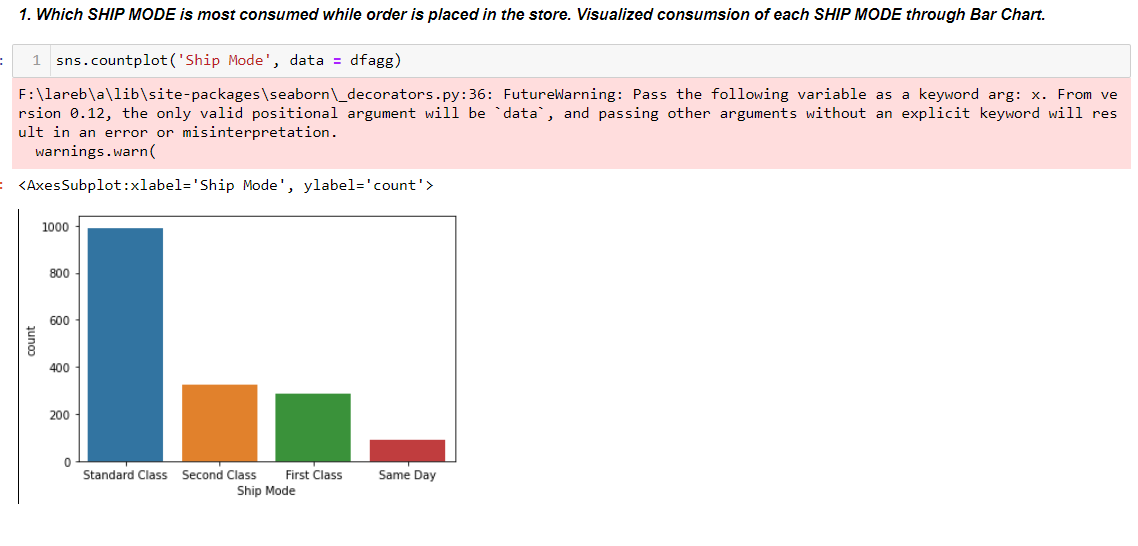
* The data Covers all the **Region** of US, Central, South, East and West, all the **States,** having all the **Cities.**

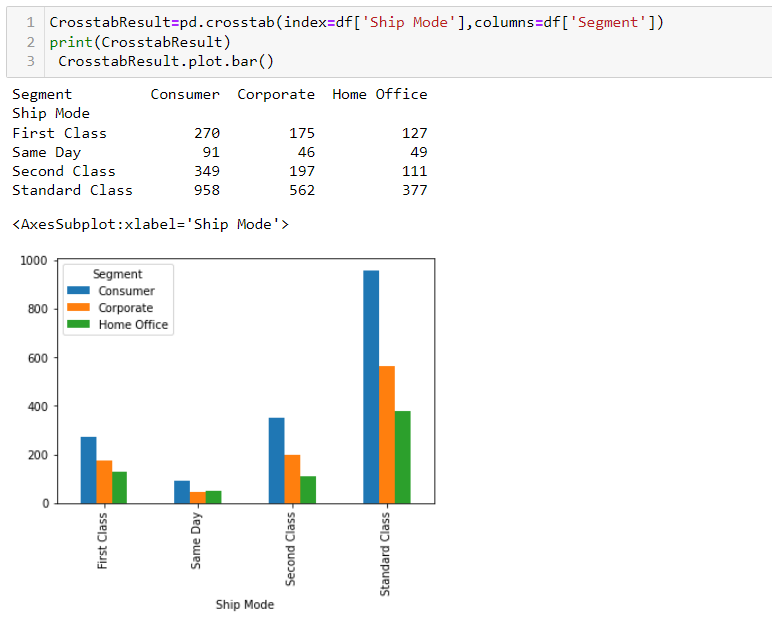
|  |  |  |
| --- | --- | --- |
| **Region** | **States** | **Cities** |
| Central,  East,  South,  West | Wisconsin, Texas ,Alabama  ,Florida ,Pennsylvania  ,District of Columbia  ,Missouri ,Tennessee  ,Indiana ,Massachusetts  ,Vermont ,Connecticut  ,Illinois ,Georgia  ,Kansas ,Iowa  ,Maryland ,Minnesota  ,New Jersey ,Oklahoma  ,Delaware ,Kentucky  ,Mississippi ,Arkansas  ,New Mexico ,Rhode Island  ,North Dakota ,Nebraska  ,West Virginia ,South Dakota | Aberdeen, Alexandria  Athens, Bowling Green  Bryan, Buffalo  Canton, Coachella  College Station  Eagan, Edinburg  Elyria, Escondido  Garden City  Gladstone, Greenwood  Grand Prairie  Grove City, Hagerstown  Hillsboro, Homestead  Jupiter, Las Cruces  Longview, Murfreesboro  Odessa, Ontario  Oxnard, Pasco  Renton, Riverside  Round Rock, Saginaw  Saint Louis, San Angelo  Santa Clara, Springdale  Summerville, Sunnyvale  Urbandale, Utica |

* As it is a E-Store, there are four **Shipping modes,** First Class, Standard Class, Same Day, and Second Class.
* And 3 types of **Segments** were there, one is **HonJHhhhHome office, Corporate** and **Consumers.**

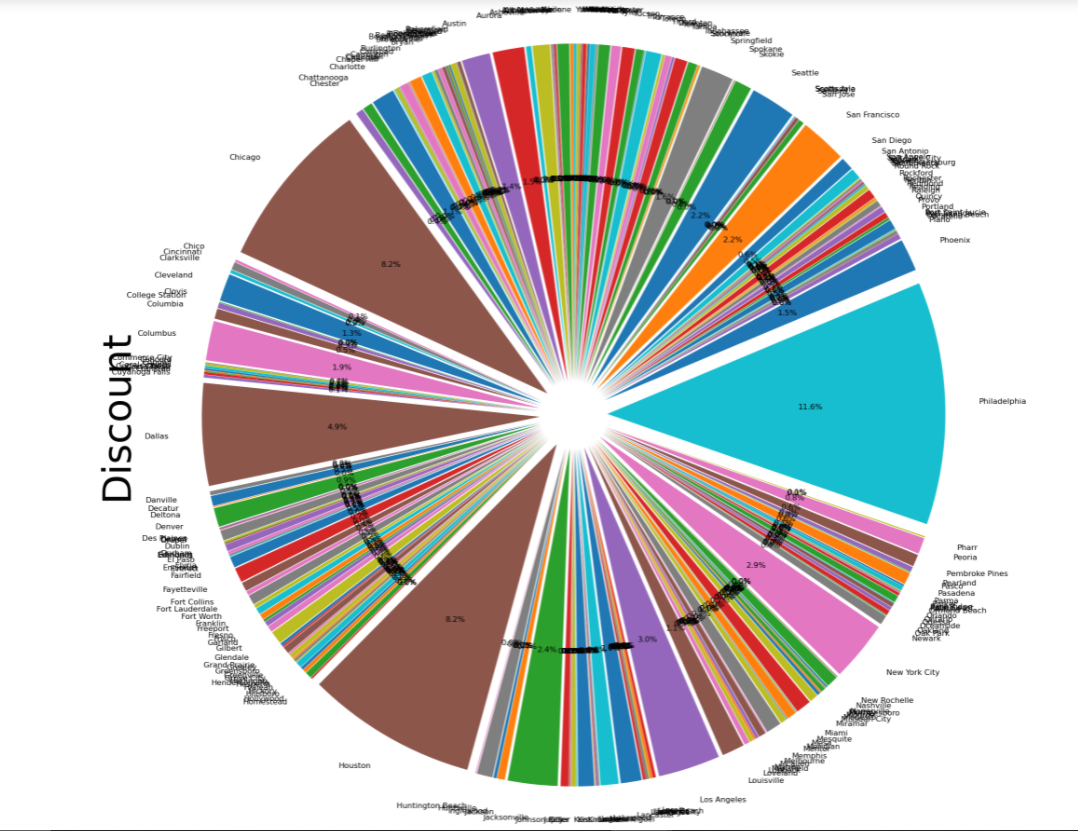
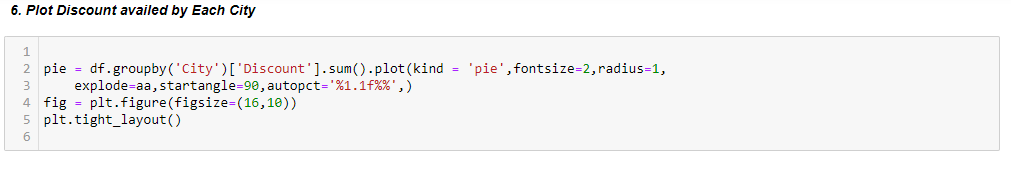
# EDA Techniques:

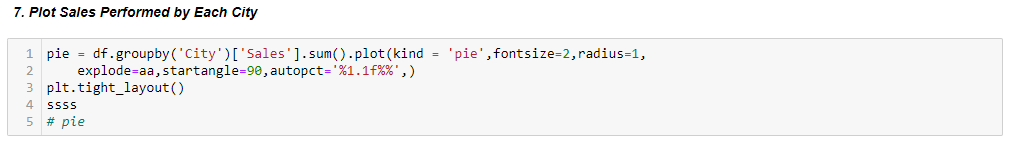
* Load Dataset.
* Fill empty data by similar field
* Cleanse data to get executed
* Use numpy , seaborn, matpltlib, sklearn, for loading manipulation and plotting
* Convert Categorical data to numerical data to perform modeling
* As data is unsupervised, so set Segment Feature as target variable.

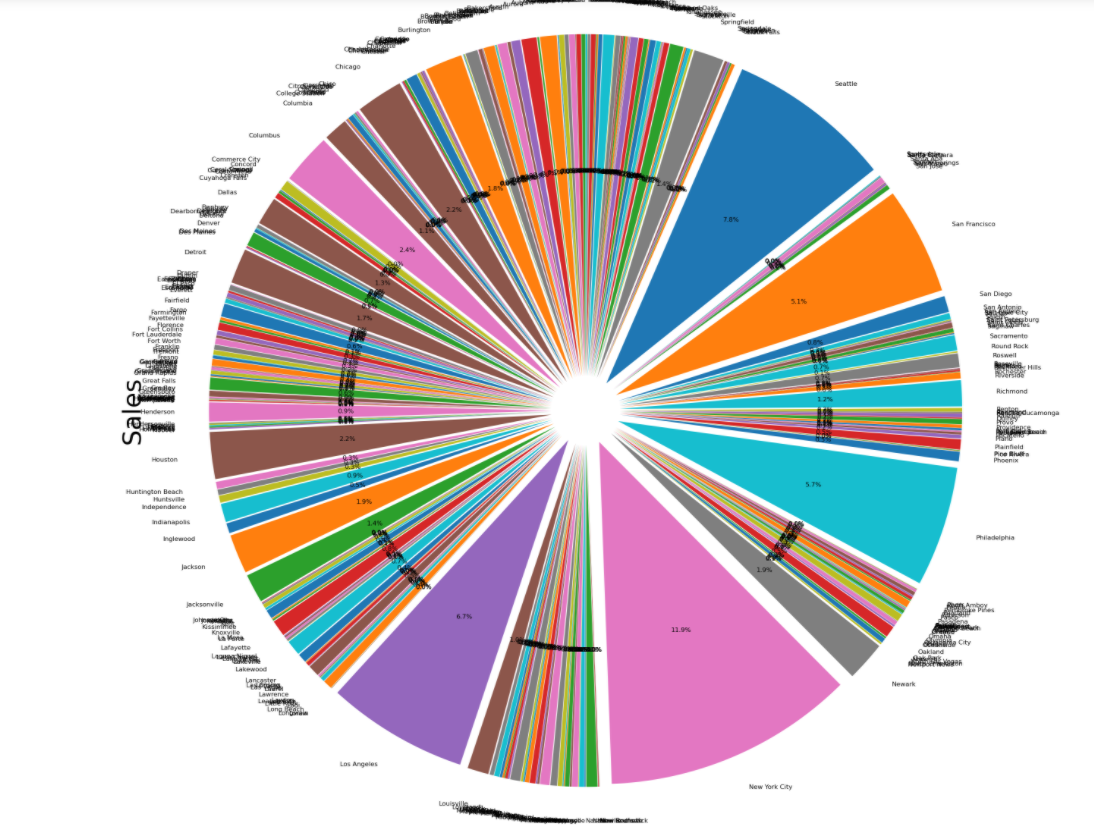


In this case we first identify the most consumed shipping mode then we come to know the consumer belong to which segment.





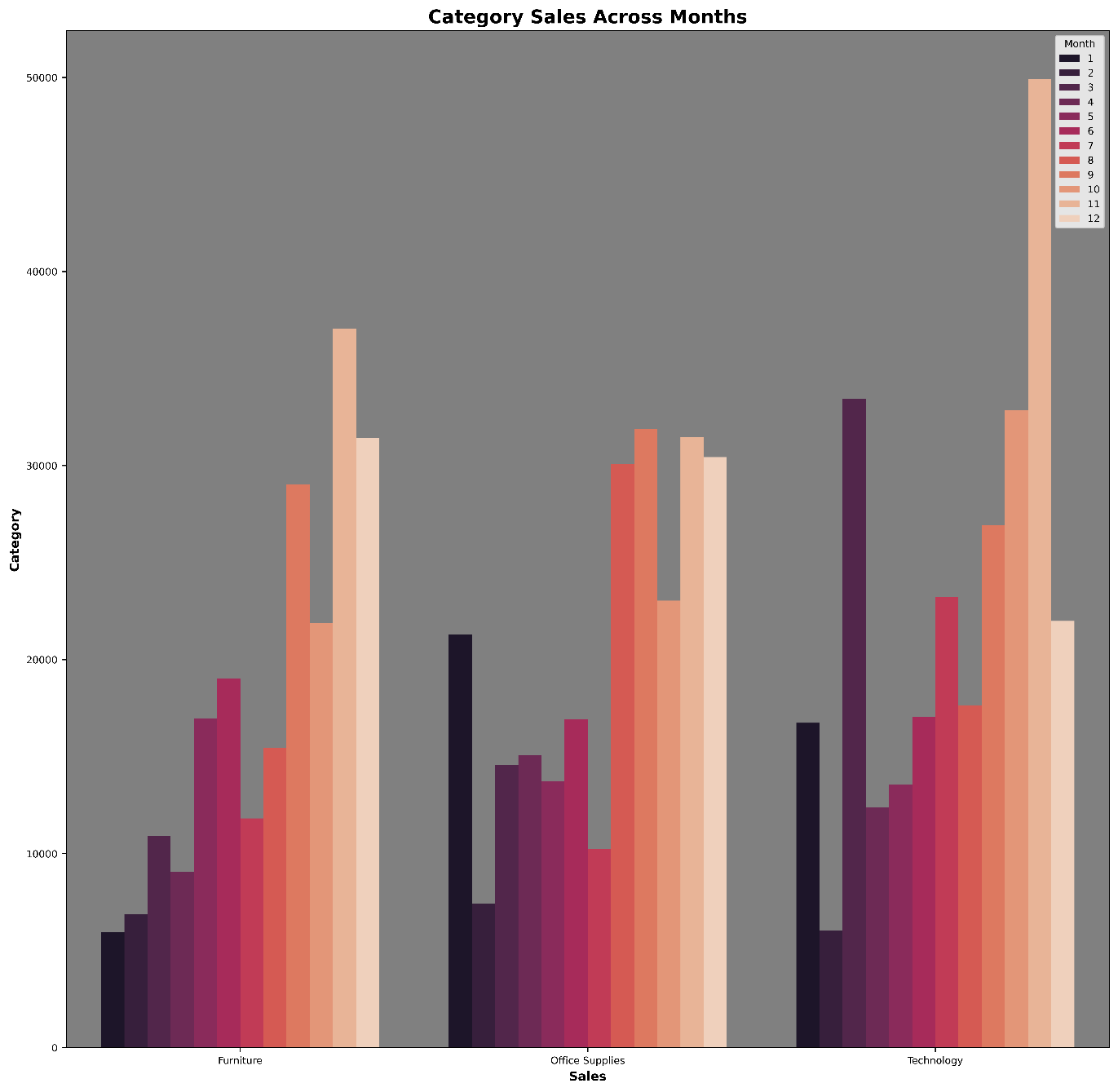
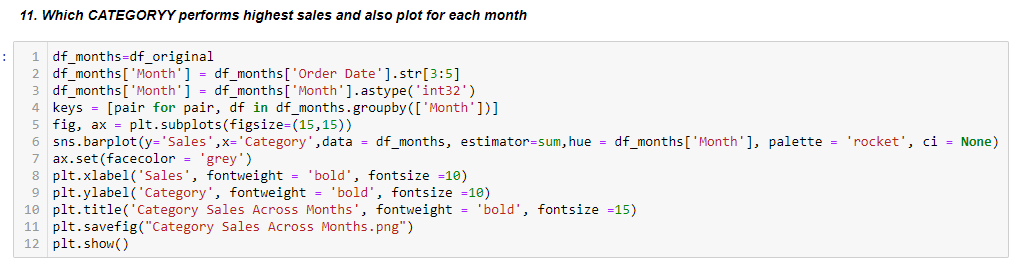


By the above two graphs we can visualize, the cities who generates high sales also avail discounts, like NEW YORK CITY performs highest sales and do avail discounts but not highest if the marketing is done accordingly maybe sales may be boost in the region, similarly in San Francisco, Los Angeles, Seattle.

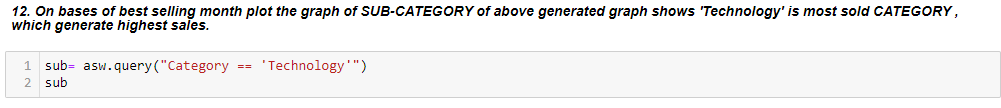
Whereas Chicago Huston and Philadelphia consumed highest discounts but doesn’t perform high sales.

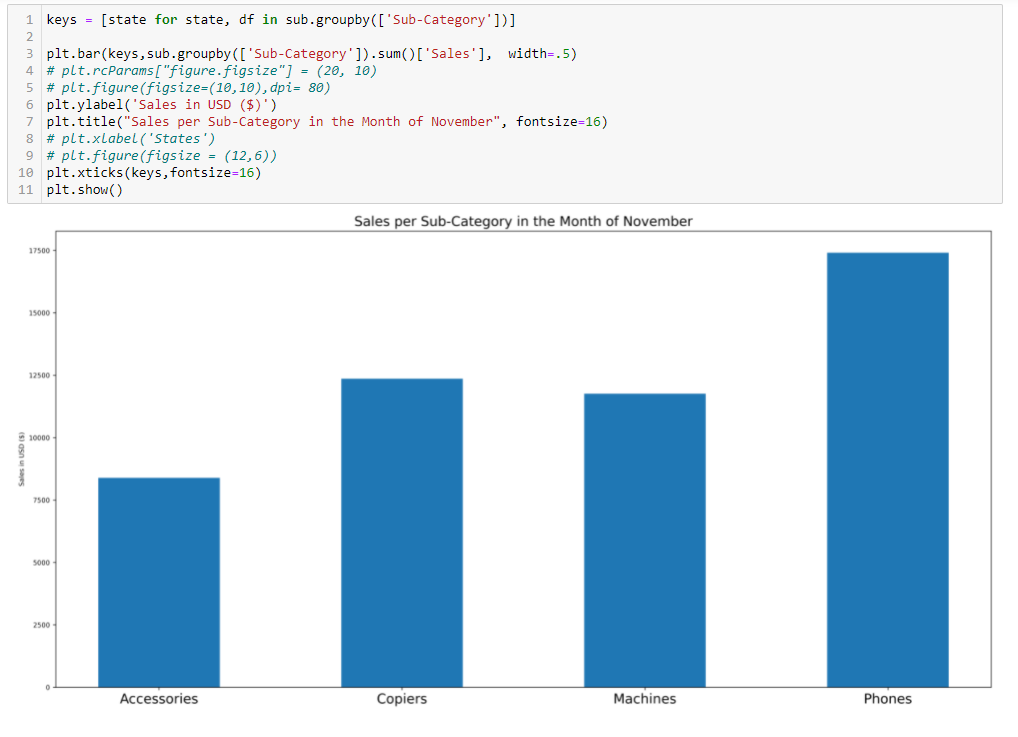


This graph shows the most selling products are from the category of Office Supplies and Is BINDER and ACCESSORIES.



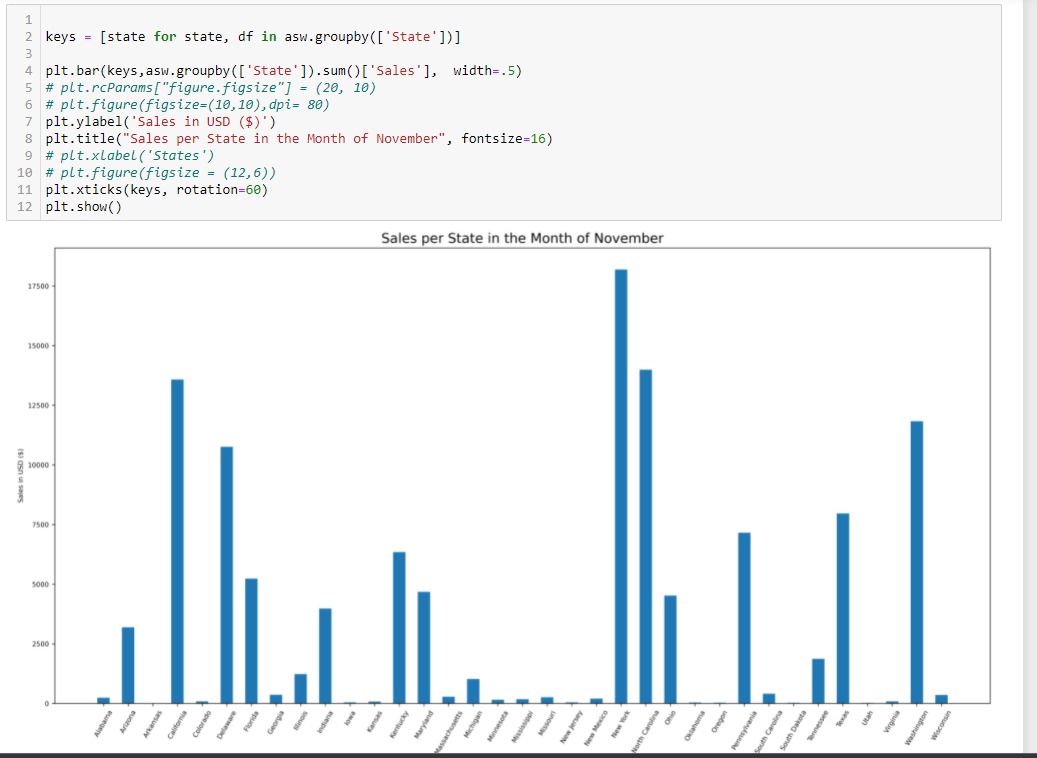
In the month of November each category has growth in its sales, technology got rose to $50000, which was at its best, while it technology also does it best in the month of March.



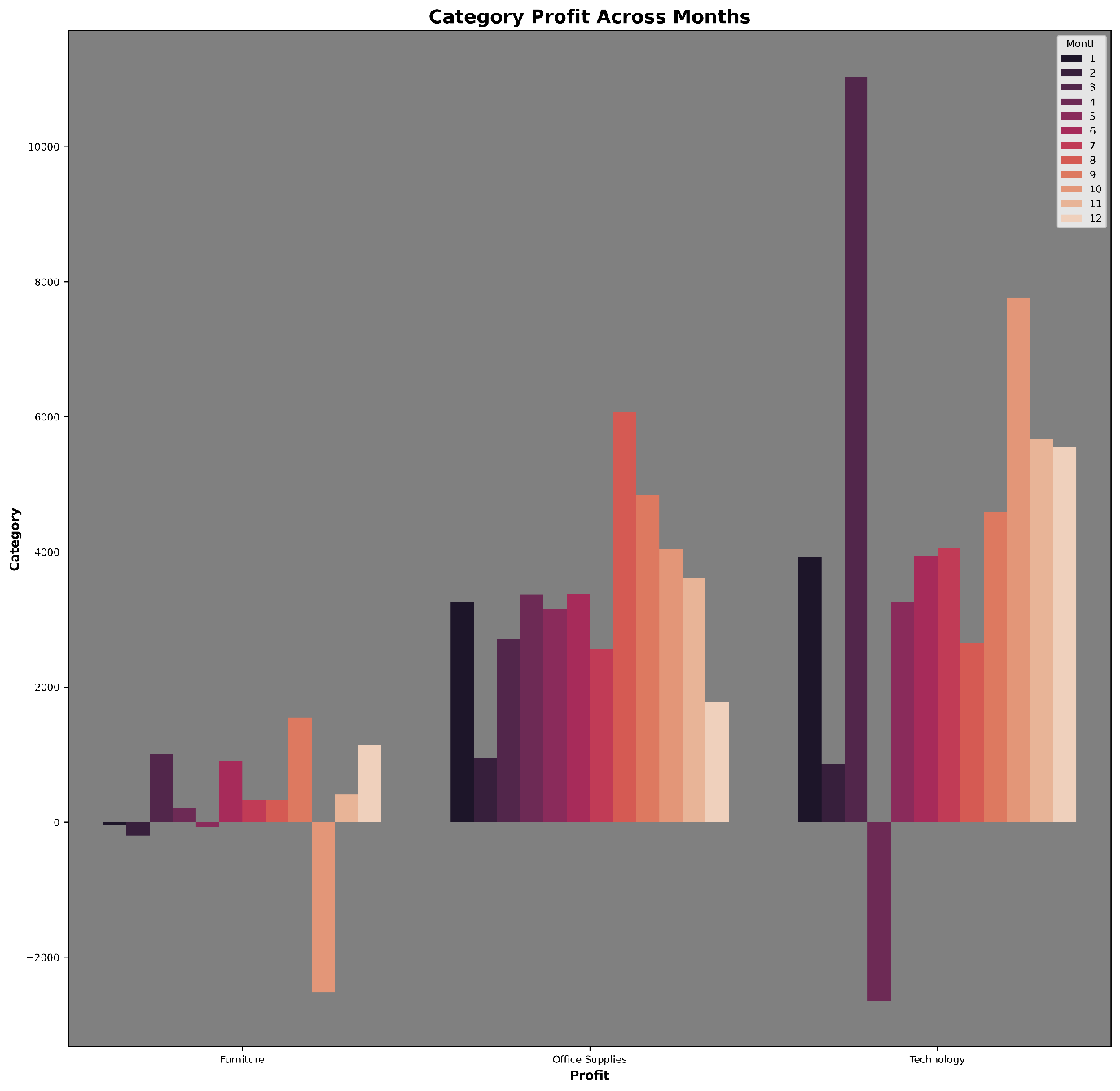
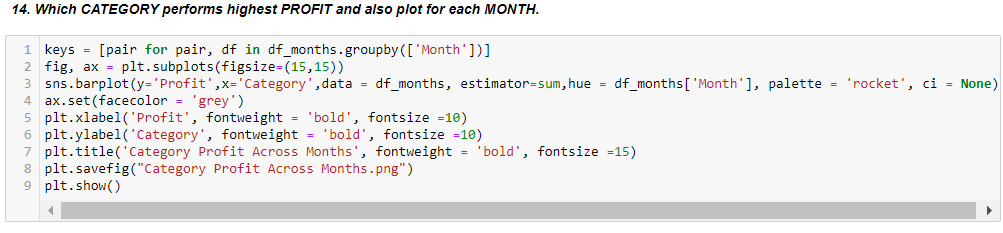


On the Bases of previous graph, we identify two things that technology performs well in the month of November, So in this graph we drill down it to sub category and find out which sub category performs highest sells.

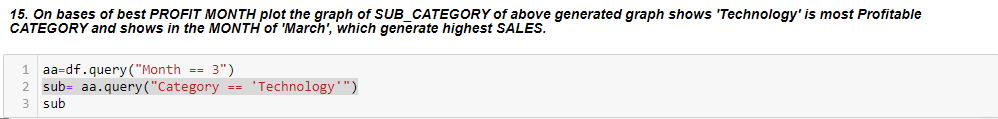


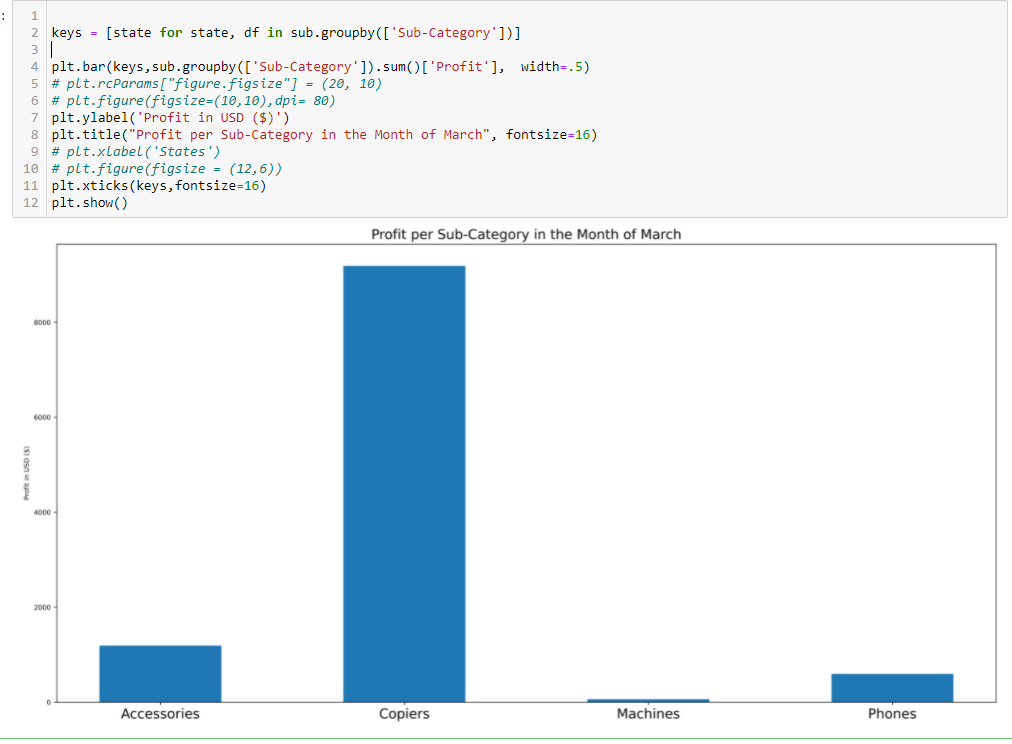


And On the Bases of Sub Category we also find out which state generates sales.

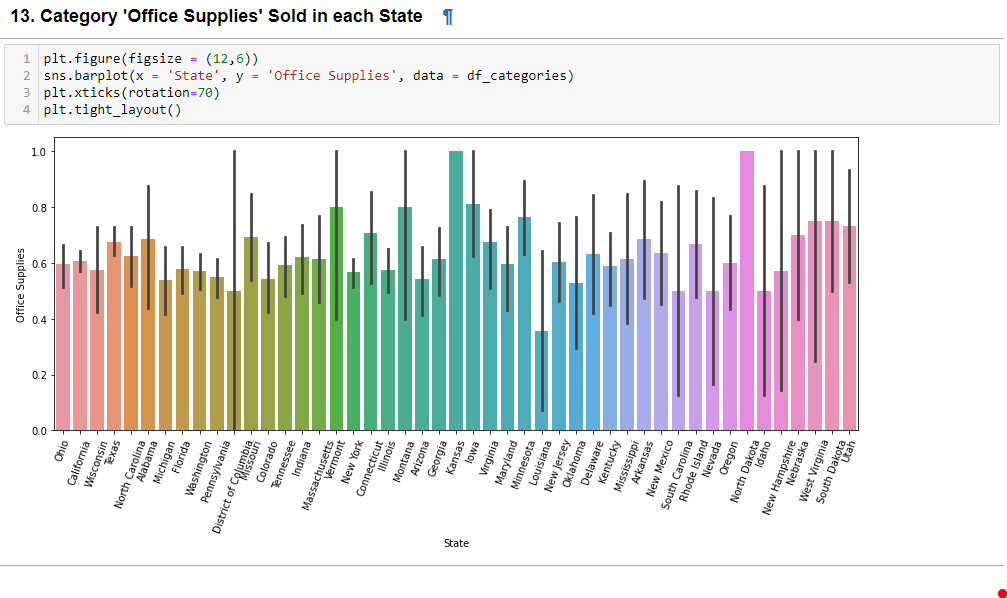
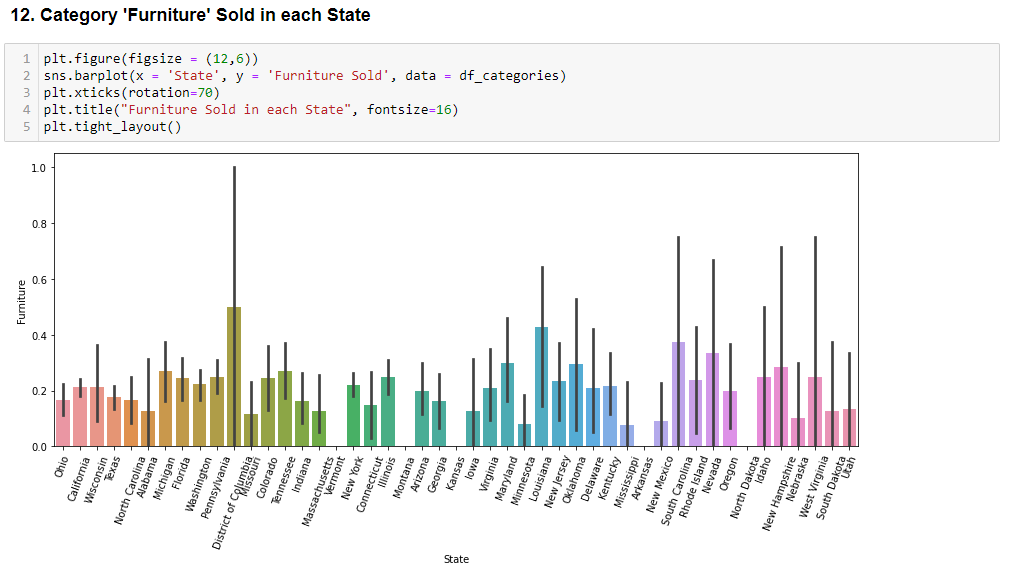


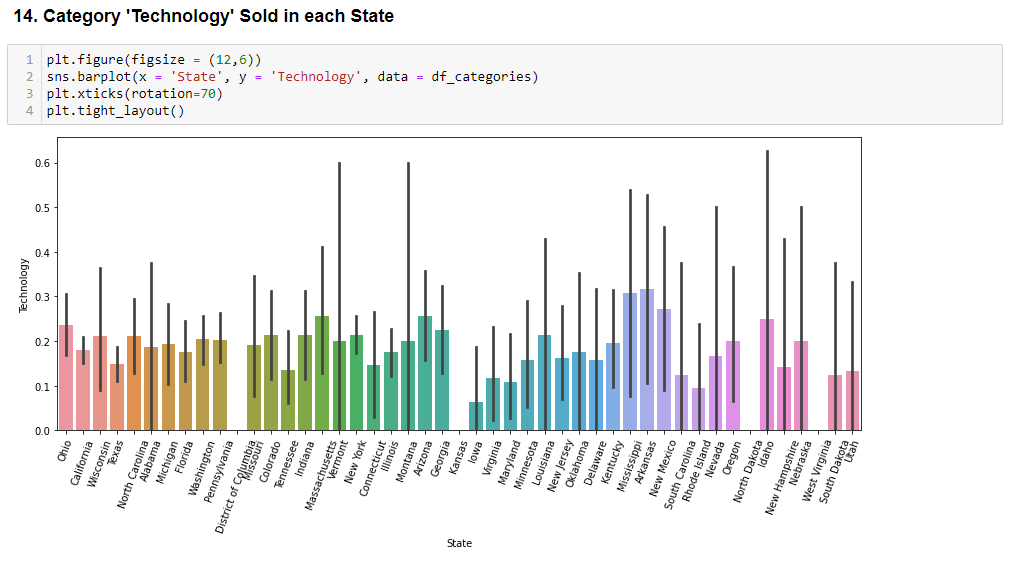
On bases of each month, we find out that in the month of march technology provide highest profit upto $10000 and in April it lost with $2000 in the Same category.

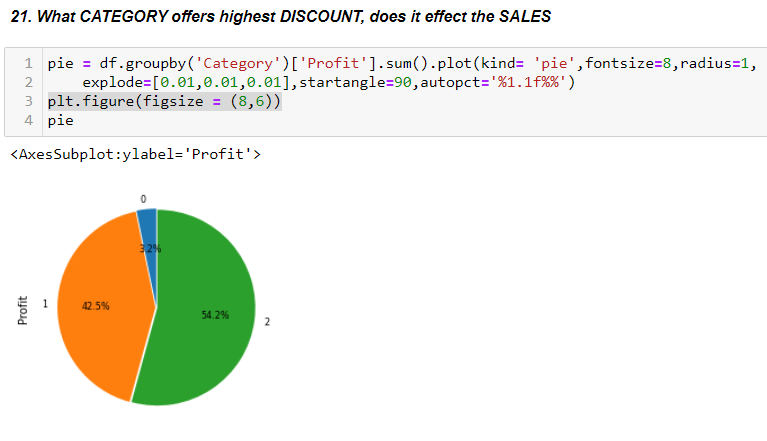
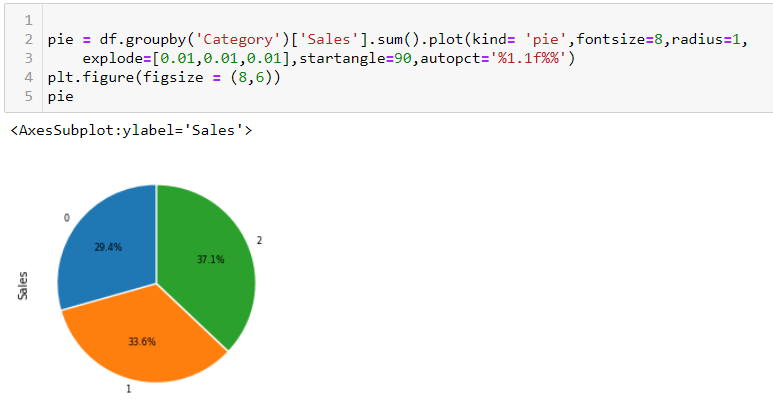




As March was the most profitable month and the category was technology to generate the highest profit, we find out that the sub category was copies to generate this much profit in the category.



While technology performs the highest sales so the profit, bot furniture generate the least profit but perform the reasonable sales, so it it might be said that I technology stuff there is a great margin of profit then of furniture.

# Model Prediction:

