มหาวิทยาลัยอัสสัมชัญ
ASSUMPTION UNIVERSITY
of THAILAND

VINCENT MARY
SCHOOL of SCIENCE and
TECHNOLOGY

Project 3 Report

Semi-Structure Data Analytics

Saw Zwe Wai Yan 6318013

CSX 4205 Big Data Analytics (1/2023)

ABSTRACT

In this report, we embark on a data analytics journey, leveraging MongoDB and a set of real-world data from the "sample_supplies" collection. Our mission is to extract valuable insights, answer specific queries, and present our findings to illuminate the underlying trends and patterns. Our data pipeline is meticulous, consisting of five distinct stages: Ingestion, Transformation, Preparation, Processing, and Result. We employ a Python-based notebook tool for this analysis, ensuring our readiness for future data-driven challenges.

The culmination of our work is presented in a Jupyter Notebook, accessible on GitHub. This notebook serves as a comprehensive article, meticulously outlining our methodologies, code, and visualizations. It stands as a testament to our dedication to the art of data analysis.

Keywords: Data Analytics, MongoDB, Jupyter Notebook, GitHub, Ingestion, Transformation, Insights, Data Pipeline, Query Analysis.

Table of Contents

1 INTRODUCTION

In this report, we're diving into the world of data analysis. We're using a dataset called "sample_supplies" from MongoDB to find interesting insights and answer some important questions.

We'll follow a step-by-step process to analyze the data, and we're using a handy tool called a Jupyter Notebook to do this. The results will be shared in a notebook on GitHub, where we'll explain our findings

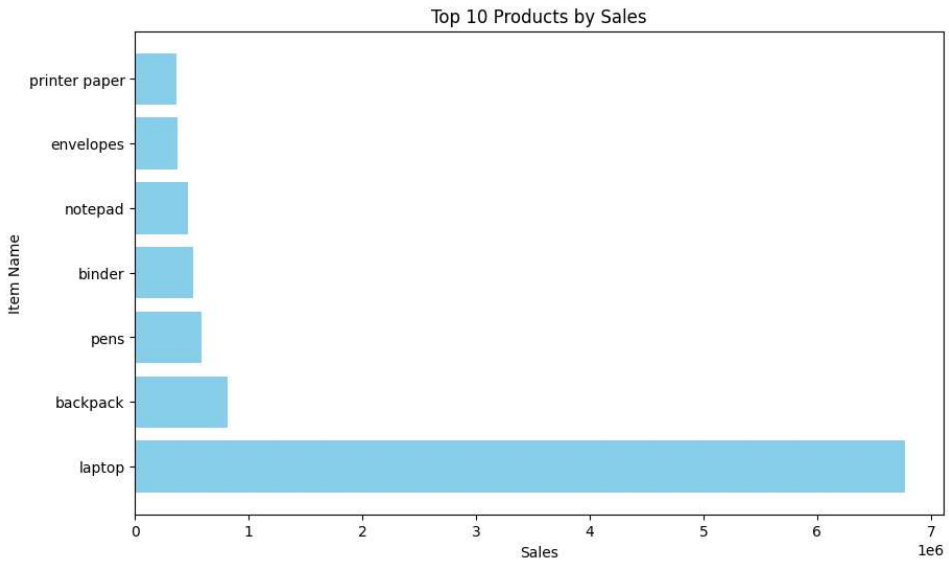
2 DATA PIPELINE

We get the Data from MongoDB sample Data, we use the MongoDB tools to get the data. We can choose between 2 format. CSV format and Json format. I have chosen the Json format since the original data is also in Json format. I use pandas to read the json format. I exported the file using mongo export. The command is

mongoexport --uri "<yourDatabaseString>/sample_supplies" --collection sales --out sales.json

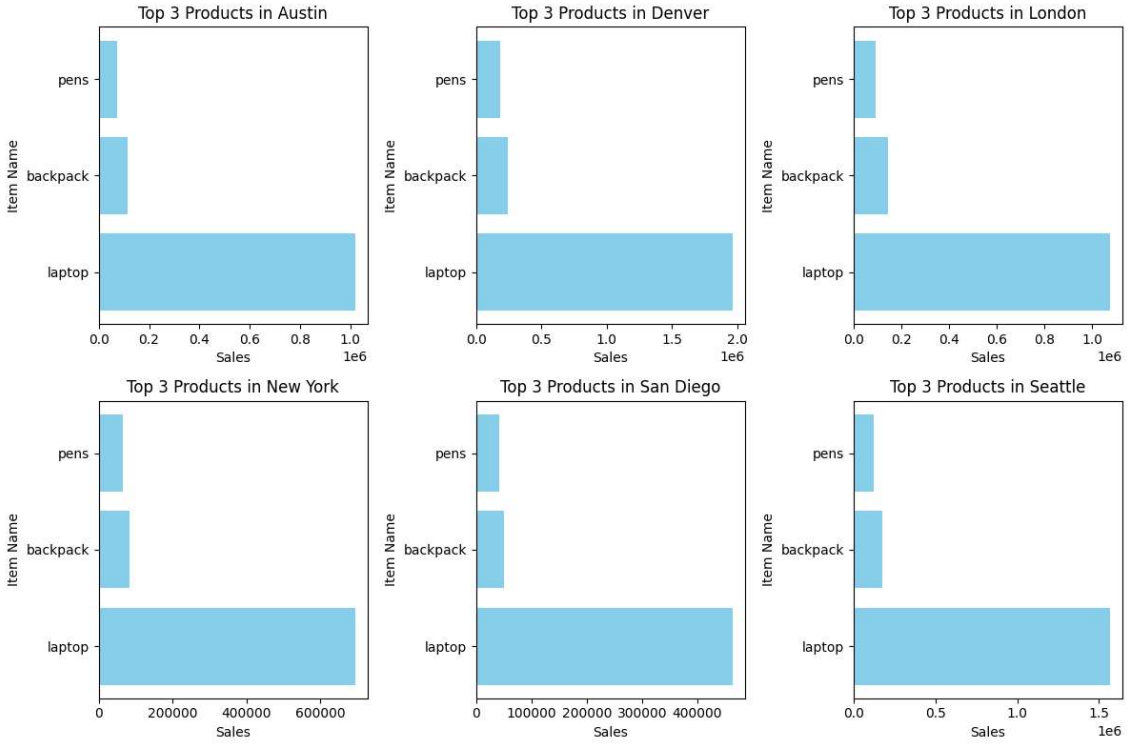
3 QUERY ANALYSIS

3.1 Show top 10 products (name) sales (quantity x price).



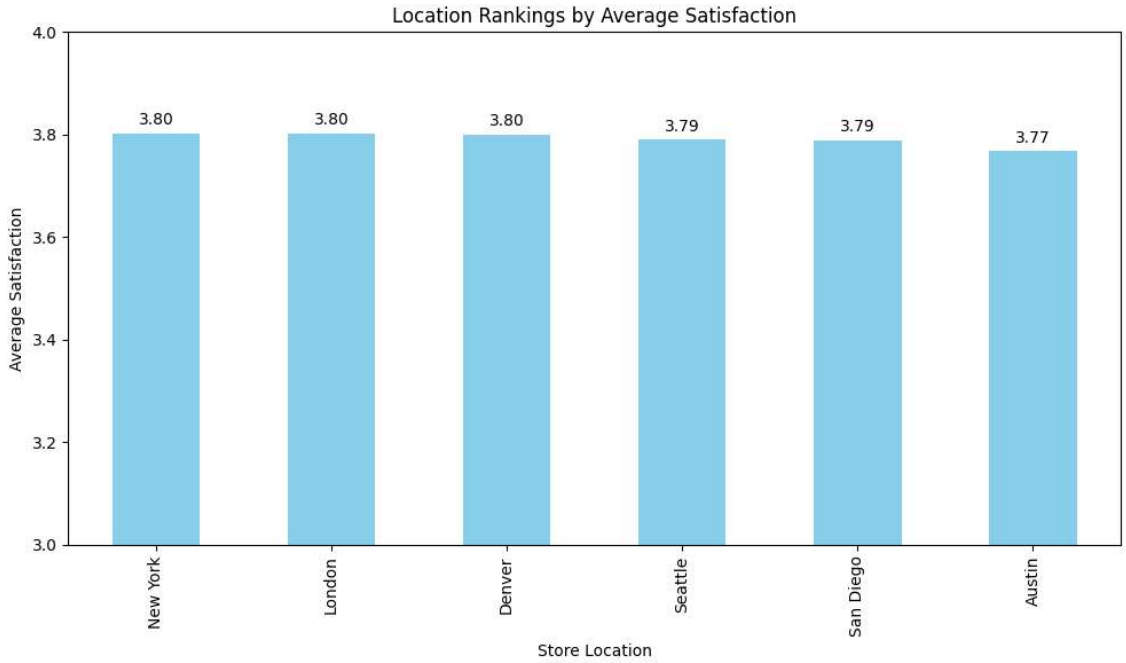
This query is to show the top 10 sales of product. There is no sale column thus we have to calculate for it. We can do that by multiplying quantity and price. We can see the visualization from the graph that, Laptop has the most significant sales compared to other items.

3.2 Show top 3 products (name) sales by store (location).



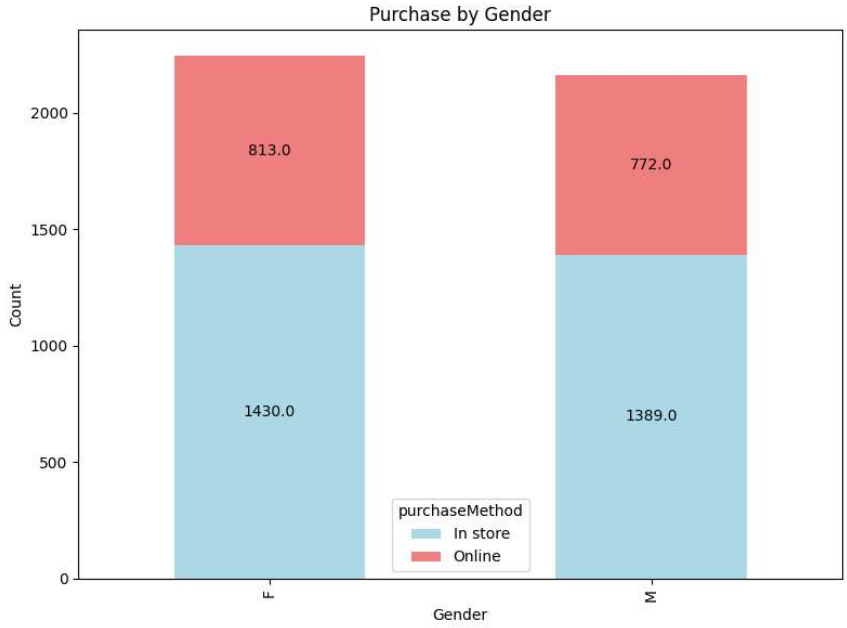
The graph represents each location and what their sales are. We can conclude that all locations perform well with the laptop, and next we have backpack and lastly we have pens. These are the top 3 products.

3.3 Show rankings of each store (location).



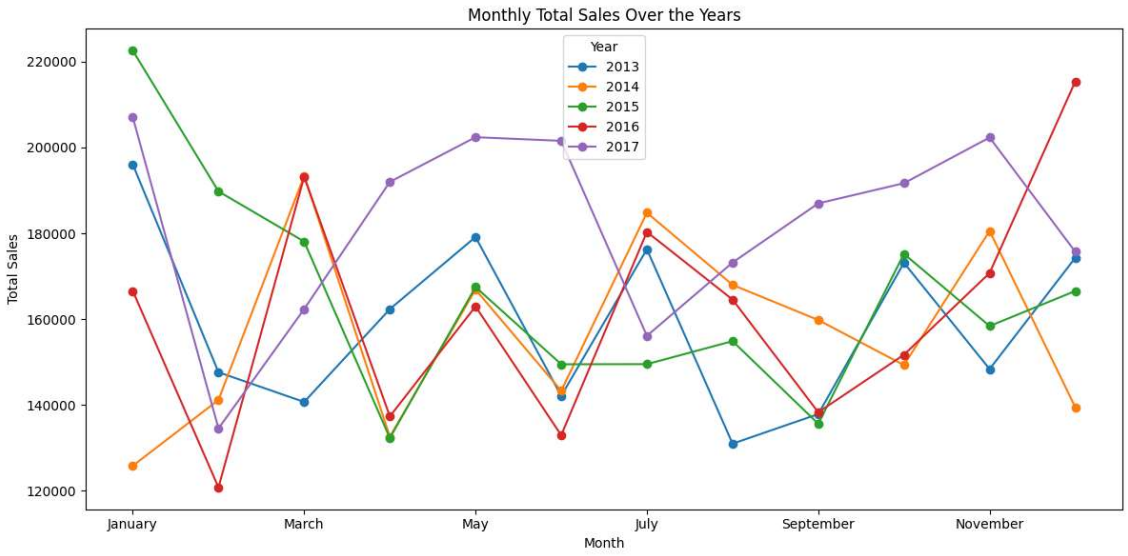
For the rankings there are 2 ways you can approach, either by the sales or by the satisfaction of the customer. I have chosen customer satisfaction, we can conclude that the New York, London and Denver are tided for the first place with an average rating of 3.8/5 and the lowest rating is from Austin with a rating of 3.77/5.

3.4 Show purchased method by gender table



This shows the demographic of people who purchases the items online or in store. We can conclude that for both genders, both of them prefer buying items in store rather than online.

3.5 Show monthly total sales



The graph represents how much each month made sales with the range of 5 years, 2013 – 2017. From the graph we can see that the sales performance generally rises during April or July. Meaning we can have better deals during that time.

4 VIDEO PRESENTATION

https://youtu.be/uBy5NwGCqZg

5 NOTEBOOK

https://github.com/sawzwe/bigdata_project3/blob/main/project_3.ipynb