**EFFECT OF COVID ON SALES AND E-COMMERCE**

1. **INTRODUCTION**
2. **BUSINESS QUESTIONS**
3. Does increasing covid cases affect cosmetic sales?
4. Does increasing covid cases affect jewelry sales?
5. What are the most affected brands?
6. **FRAMEWORK USED**

We used Hadoop + Spark for our project.

Jupyter Notebook

Data Processing

Apache Spark

Hadoop YARN

Resource Management

Hadoop HDFS

Data Storage

File System

Data Source (.csv)

**Why we chose this framework?**

Jupyter notebook on Spark is a very nice combination for big data processes and data analytics. It provides a good user interface and combines multiple works into a single file. We could do everything from data processing and analysis to data visualization.

1. **DATASET**

We selected our dataset from Kaggle,

1. eCommerce in Cosmetic Shop:

<https://www.kaggle.com/datasets/mkechinov/ecommerce-events-history-in-cosmetics-shop>

1. eCommerce in Jewelry Shop:

<https://www.kaggle.com/datasets/mkechinov/ecommerce-purchase-history-from-jewelry-store>

1. Covid Dataset:

<https://www.kaggle.com/datasets/gauravduttakiit/covid-19>

eCommerce (Cosmetic) dataset consists of data from Oct 2019 – Feb 2020. eCommerce (Jewelry) dataset consists of data from Dec 2018 – Dec 2021. The covid dataset consists of data from Jan 2020 – Apr 2022. Total size of the datasets is around 3 GB.

1. **DATA STRUCTURE**

**E-commerce (Cosmetic):**

**E-commerce (Jewelry):**

**Covid:**

1. **DATA PREPARATION AND CLEANING**
2. Loaded the e-com dataset as a spark data frame in jupyter notebook.
3. Converted the datatype of the variables as desired.
4. Combined five datasets (500MB each) into single data frame. This is our e-Commerce Cosmetic dataset.
5. Loaded covid dataset as a spark data frame in jupyter notebook.
6. Converted the datatype of the variables as desired.
7. We processed the covid dataset to get covid cases on each day.
8. Finally, we are also using eCom – Jewelry dataset to find effect of covid on gold sales
9. Now we have two data frames: 1. eCom for Cosmetic 2. Covid dataset 3. eCom for Jewelry
10. **HOW OUR DATASET LOOKS NOW?**

**ECOM DATSET:**

**Table

Description automatically generated**

**COVID DATASET:**

**Graphical user interface, text

Description automatically generated**

**ECOM JEWELRY DATASET:**

**Graphical user interface, text, application

Description automatically generated**

1. **QUERIES AND RESULTS**
2. **WHAT IS THE TOP 5 PRODUCTS SOLD EACH MONTH?**

**OCTOBER 2019**

**Graphical user interface, application, Word

Description automatically generated**

**NOVEMBER 2019**

**Graphical user interface, text, application

Description automatically generated**

**DECEMBER 2019**

**Graphical user interface, text, application

Description automatically generated**

**JANUARY 2020**

**Graphical user interface, application

Description automatically generated**

**FEBRUARY 2020**

**Graphical user interface, text, application

Description automatically generated**

From the above tables, we could see that the products ‘5560754’, ‘5751422’, ‘5809910’ are one among the top 5 each month. The performance of these products is consistent throughout the five months. From this pattern, we couldn’t find much difference on sales before and during covid.

1. **WHAT ARE THE TOTAL SALES EACH MONTH?**

**Graphical user interface, application

Description automatically generated**

From the above report, we could see that the sales are more in November and less in December. From this, we don’t find the effect of covid on sales.

1. **WHAT ARE THE SALES ON EACH DAY OF THE MONTH?**

**Table

Description automatically generated**

From analyzing each day sales, we could observe spikes of sales on each month, except February. Also, we observe declining sales in February, which will be clearer in visualization. These declining sales may be the effect of Covid.

1. **WHAT ARE THE GOLD SALES EACH MONTH OF 2019, 2020, 2021?**

**Graphical user interface, application

Description automatically generated**

From the above table, we could observe that the gold sales keep increasing from 2019 to 2021. It is not just a slight change but could see a drastic increase of sales during covid which is unbelievable. Because of this increase in gold sales, we thought of analyzing the gold price during these periods, that might be the influencing factor.

1. **HOW DOES PRICE OF GOLD PRODUCTS CHANGES EACH YEAR FROM 2019 TO 2021?**

**Graphical user interface, table

Description automatically generated**

From the above table, we could observe that price of the product remaining the same for each product from 2019 to 2021.

1. **VISUALIZATION**
2. **HOW COVID INCREASES EVERYDAY?**

**Chart, histogram

Description automatically generated**

1. **WHAT ARE THE COSMETIC SALES EACH MONTH?**

**Chart

Description automatically generated**

1. **HOW COVID AFFECTS SALES IN FEBRUARY?**

**Graphical user interface, application

Description automatically generated**

1. **FINDINGS AND SOLUTIONS**
2. **CHALLENGES FACED AND PROCESS ATTEMPTED**
3. **Kafka implementation** – We tried implementing Kafka to process the data as streaming data. We started facing space issues with a very big dataset and very small storage in the ubuntu machine.
4. **AWS EMR** **+ Spark + Jupyter**– To overcome the above issue, we thought to use cloud data storage and integrate with Spark and Jupyter. But with the limited free tier limit, we were unable to keep the service running and had to terminate the service.
5. **Hadoop + Spark + Jupyter** – Finally, we decided to use Hadoop + Spark + Jupyter framework, which we found to be an easy and powerful combination for our project needs.
6. **LIMITATIONS**
7. **Dataset:** The main limitation is that we were unable to find the e-commerce dataset for Jan 2019, our dataset contains records starting from Oct 2019 to Feb 2020. To accurately find the sales and covid relationship, we needed the dataset for Jan 2019. In that case, we would have been clearly able to compare and correlate the sales in Jan 2019 and Jan 2020.
8. **FUTURE PLANS**