Report - Assignment 2

Data - **Google ecommerce**

Team 1

Abhishek Gargha Maheshwarappa
Nuid -001375462
Shubam Sharma
Nuid - 001447366
Sonali Vinodkumar Singh
Nuid - 001393262

Objective

To build an analytical dashboard as a Proof-of-concept to illustrate the value of data driven analytics.

Analyze the data using tools

Data Wrangling

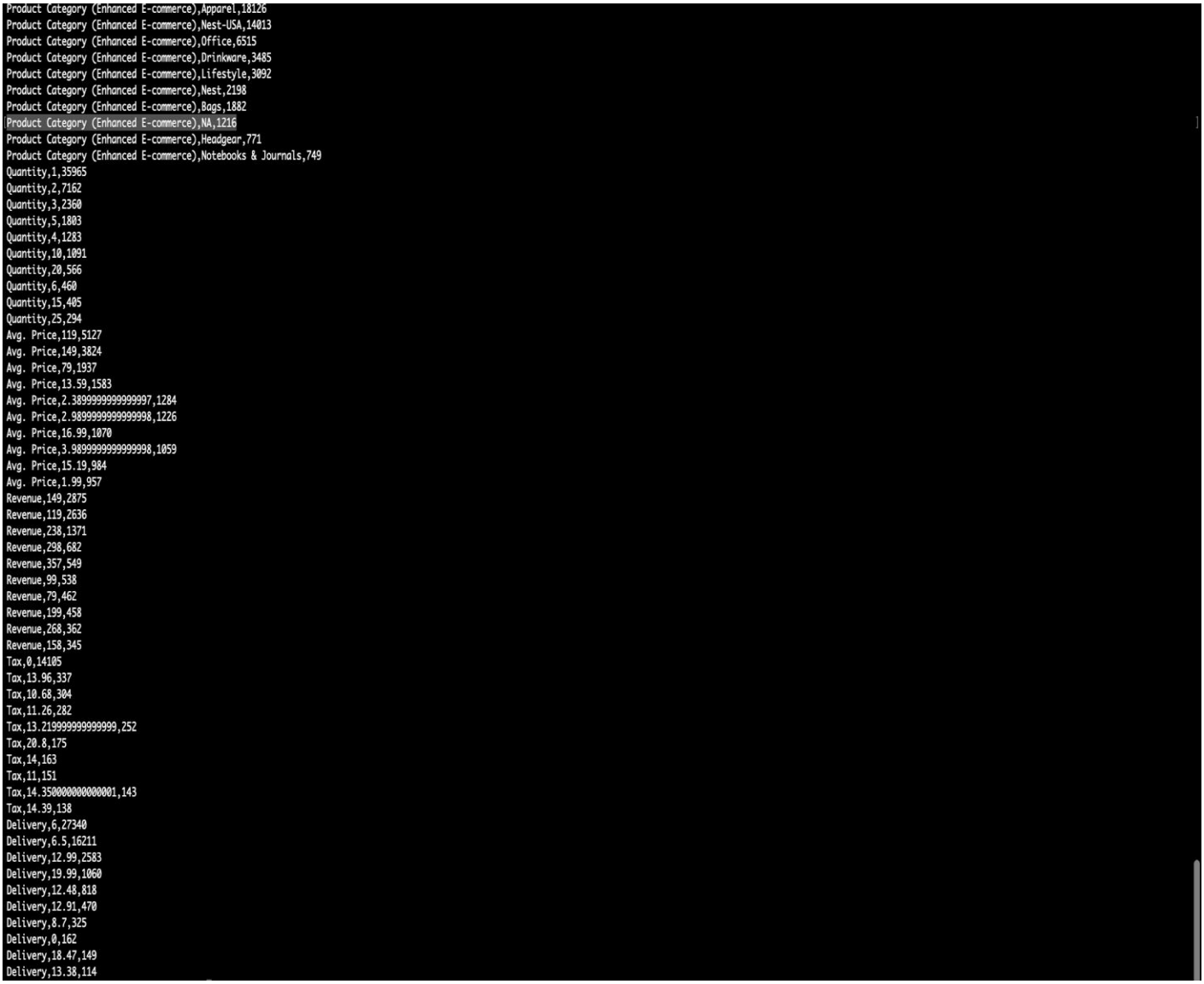
- Successful analysis relies upon accurate, well-structured data that has been formatted for the specific needs of the task at hand. Yet, today's data is bigger and more complex than ever before.
- It's time-consuming and technically challenging to wrangle it into a format for analysis.
- Data wrangling is the process you must undergo to transition raw data source inputs into prepared outputs to be utilized in analysis and various other business purposes.

XSV

- XSV is a command line tool which is used for joining , slicing , analysing CSV files.
- **What we liked about the tool**
    - Tool is very fast function like searching , joining for big csv file take ~ 6-7 sec
    - Tool has help function which gives description of every command of xsv
    - Found Commands liks frequency, stats, table really helpful to get a overview of data
- **Disadvantage of tool**
    - No user Interface only command line which may appear boring for few people
    - Commands are limited to joining, slicing and getting statistics of data
    - Limited resources available for this tool

```
shubhams-MacBook-Pro:marketing shubham$ xsv stats Online.csv | xsv table
field                                type     sum                 min                  max                                 min_length  max_length  mean                stddev
Transaction ID                       Integer  1750856793          16679                48497                               5           5           32337.04183289049   8633.735137517308
Date                                 Integer  1092121199824       20170101             20171231                            8           8           20170678.188239127  332.8962106342856
Product SKU                          Unicode                      GGOEA0CH077599       GGOEYOLR080599                      12          14
Product                              Unicode                      1 oz Hand Sanitizer  YouTube Youth Short Sleeve Tee Red  8           59
Product Category (Enhanced E-commerce) Unicode                    Accessories          Waze                                2           20
Quantity                             Unicode                      1                    NA                                  1           3
Avg. Price                           Float    2787544.020000651   0.39                 355.74                              1           19          51.48389516843935   63.588661979256415
Revenue                              Float    14998591.69000093   0.5                  23945.56                            1           18          277.0129966385967   902.5797217262942
Tax                                  Float    797640.1000000129   0                    754.14                              1           19          14.731828088061567  26.841075809003122
Delivery                             Float    572878.6099999635   0                    521.3599999999999                   1           18          10.580648086584043  19.690382757048294
```

Above screenshot is of Stats Command which gives us an overview of data like min, max, mean and standard deviation.



Above the screen shot we have used the frequency of the command which gives us a frequency map of each column.

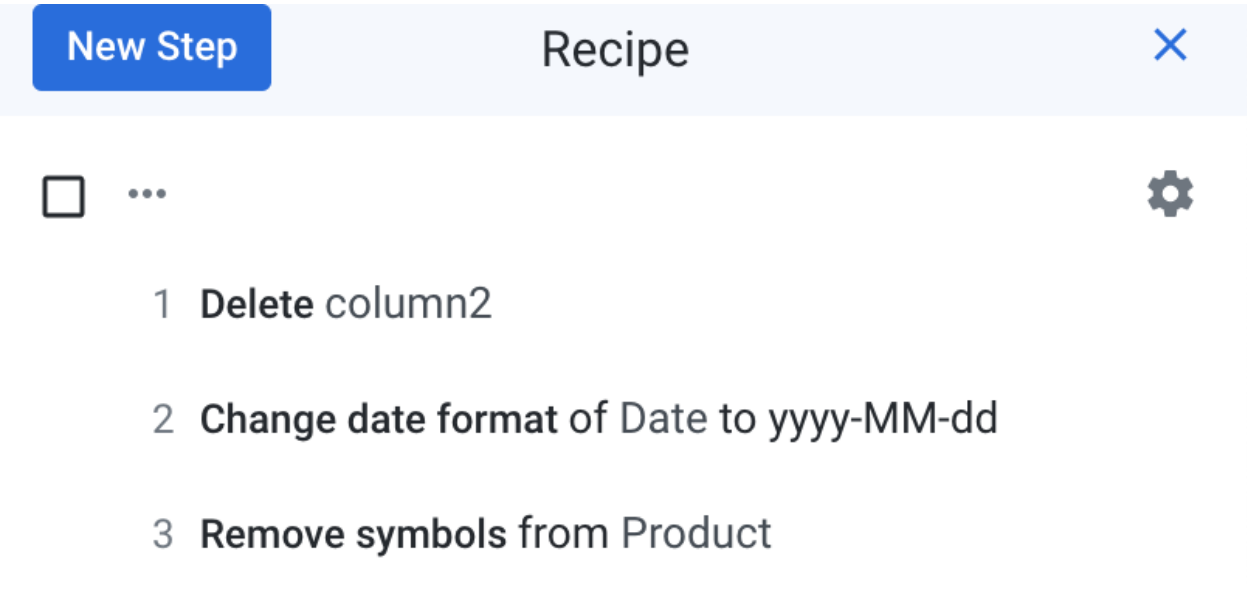
Trifacta

- software that helps individuals and organizations more efficiently explore, transform and join together diverse data for analysis.

- Whether you're working with files on your desktop, disparate data in the cloud or within large-scale data lake environments, Trifacta will accelerate the process of getting data ready to use.
- **What we liked about the tool**
    - Its Interactive User Interface, Easy to use No prior coding or Technical expertise required
    - We can create Recipe which can be used multiple time on multiple data sets
    - Its AI powered features help us in structuring, validating and cleaning data
    - Data profiling feature gives us a visual downloadable report of our data within minutes to analyse the scope of our data

- **Disadvantage of tool :**
    - Cannot download Data over 1GB on free version

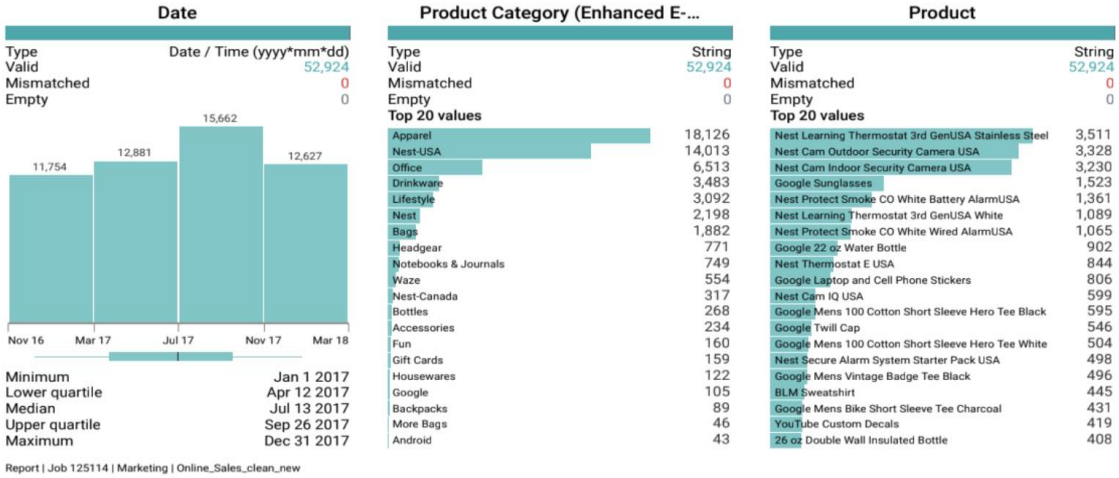
Recipes that we made to clean and structure our data

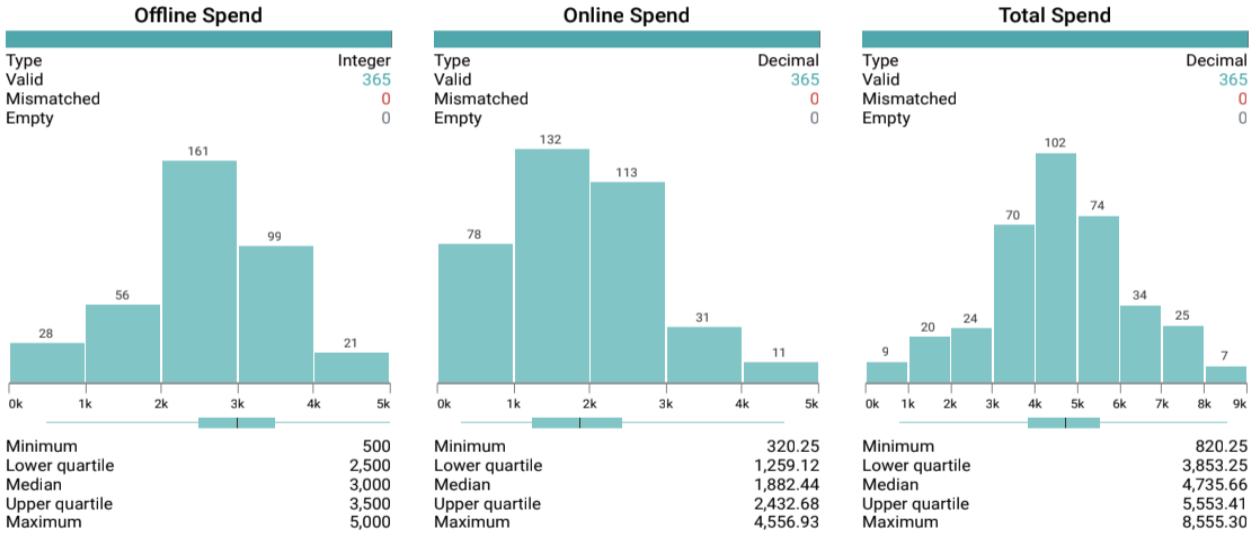


New Step

Recipe

1. **Inner join** with KEY_SKU.csv on StockCode == StockCode
2. **Inner join** with Online_Sales_clean_new on {Product SKU} == {Product SKU}
3. **Delete** StockCode
4. **Delete** Product SKU
5. **Create** counta_Product Category (Enhanced E-commerce) from COUNTA({Product Category (Enhanced E-commerce)}) grouped by 3 columns





Reports

Trifacta has this feature where they generate these customs reports having these statistics and data histograms which provide overall visibility into the quality of your transformation recipe.

1. Simple tasks should be easy.
2. Performance trade offs should be exposed in the CLI interface.
3. Composition should not come at the expense of performance

Pandas

- Pandas is a software library written for the Python programming language for data manipulation and analysis.

```
[ ] missing_percentage = Online.isnull().sum() / Online.shape[0] * 100
    missing_percentage
```

```
CustomerID                                0.000000
Transaction ID                            0.000000
Date                                      0.000000
Product SKU                               0.000000
Product                                   0.000000
Product Category (Enhanced E-commerce)    2.245863
Quantity                                  0.009235
Avg. Price                                0.000000
Revenue                                   0.000000
Tax                                       0.000000
Delivery                                  0.000000
dtype: float64
```

```
[ ] missing_percentage = Key_code.isnull().sum() / Key_code.shape[0] * 100
    missing_percentage
```

```
Product SKU    0.0
StockCode      0.0
dtype: float64
```

```
[ ] missing_percentage = Marketing_Spends.isnull().sum() / Marketing_Spends.shape[0
    missing_percentage
```

```
Unnamed: 0       0.0
Offline Spend    0.0
Online Spend     0.0
dtype: float64
```

- There were 1221Row which had been missing in the Product Category Column. The data cleaning process deleted these rows as there were more than 50000 so it would make very little difference to the analysis.
- The date formatting was different in each file so had to bring it into one common format - yyyy-mm-dd
- Pandas was later used for many Data Wrangling for many analysis like
    - Used for calculating the product and product category which appeared many times in a year using -Value counts and sort
    - Used for calculating the product and product category which sold in highest number of quantities in a year using -group by product and quantity
- The total quantity of product sold, total revenue, total cost of delivery and total tax for each day of the year was calculated using pandas quite easily.
- The above calculated table was merged with marketing spends each day of the year.

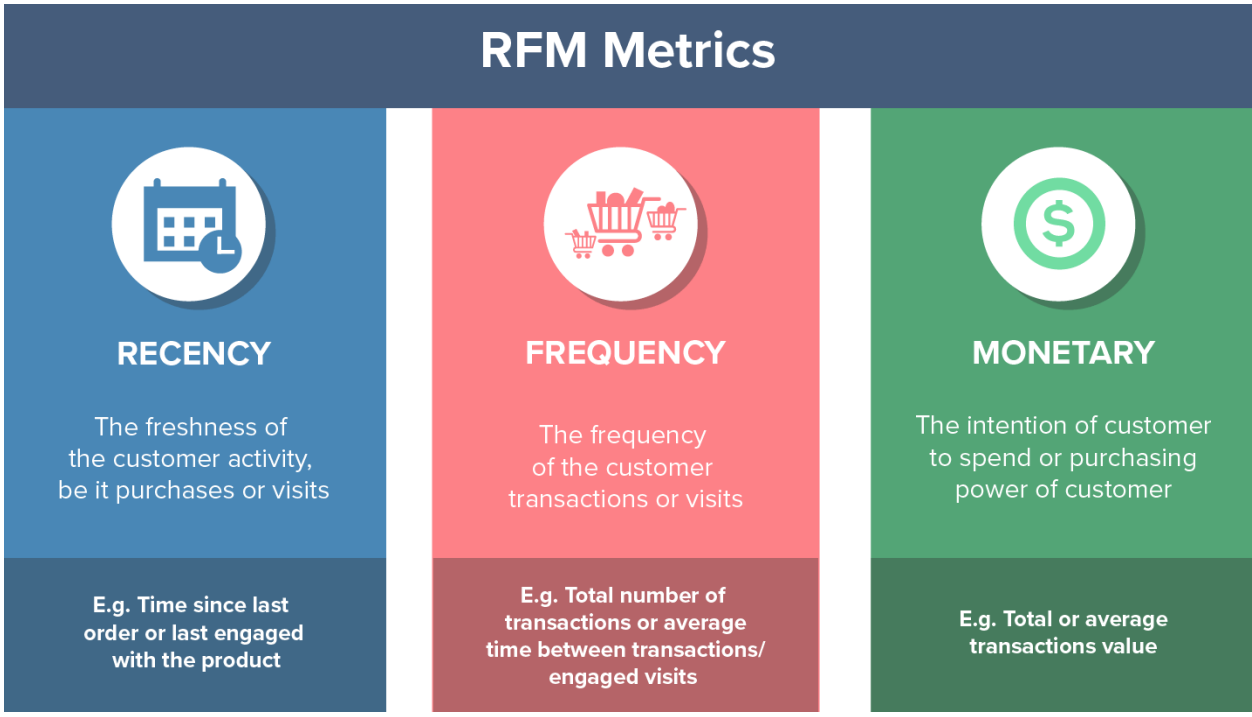
	Quantity	Date	Avg. Price	Tax	Delivery	Revenue
0	149.0	2017-01-01	1116.34	356.35	138.07	5112.21
1	104.0	2017-01-02	1222.45	786.42	53.16	7831.19
2	249.0	2017-01-03	1587.63	1255.11	630.07	11076.02
3	213.0	2017-01-04	1695.15	864.54	132.12	10049.54
4	1183.0	2017-01-05	1655.56	1144.23	610.51	13404.21
...	...	...	...	...	...	...
360	91.0	2017-12-27	1325.99	832.05	45.48	8334.89
361	25.0	2017-12-28	1256.72	694.46	41.19	6845.49
362	15.0	2017-12-29	1486.65	699.64	54.18	6974.16
363	25.0	2017-12-30	1314.52	454.77	54.18	5169.61
364	15.0	2017-12-31	1346.62	589.79	45.48	7221.17

- 365 rows x 6 columns
- Later it was used to do some preprocessing for RFM and Chorot analysis which is properly shown in the notebook.

Analysis

1. **Product that was bought most frequently -**
    - **a.** This is calculated by finding the product which appeared most number of time ignoring the quantity
    - b. Nest® Learning Thermostat 3rd Gen-USA - Stainless Steel was bought most frequently in year

2. **Product that was bought most number times considering quantity**
    a. This is calculated by finding the product which appeared most number of time with the quantity
    b. Maze Pen was brought in highest quantity
3. **Table containing - Date Wise data of - Quantity, Tax, Delivery, Revenue, Offline Spending, Online Spending and Total Spending.**
    a. This was used to compare the spendings and revenue for visualization.
    b. It gives the how spending is impacting the revenue whether there is increase or decrease in revenue.
4. **Plot of month against number of order**
    a. It gives how number order varies with different month
    b. December has the highest number of orders because of new year nearby and christmas.
5. **Plot of Day against number of order**
    a. It gives how number order varies with different days
    b. Monday has the highest number of orders because people want the product to be delivered by the weekend.
6. **cohort analysis**
    a. Cohort analysis in e commerce means to monitor your customers' behavior based on common traits they share – the first product they bought, when they became customers, etc. – to find patterns and tailor marketing activities for the group.
    b. According to the cohort analysis table January has the highest rate of retention that is 16 percent
7. **Recency, Frequency and Monetary Value** calculation
    a. **RFM** is an acronym of recency, frequency and monetary. Recency is about when the last order of a customer. It means the number of days since a customer made the last purchase. If it's a case for a website or an app, this could be interpreted as the last visit day or the last login time.
    b. **Frequency** is about the number of purchases in a given period. It could be 3 months, 6 months or 1 year. So we can understand this value as for how often or how many customers used the product of a company. The bigger the value is, the more engaged the customers are. Could we say them as our VIP? Not necessary. Cause we also have to think about how much they actually paid for each purchase, which means monetary value.
    c. **Monetary** is the total amount of money a customer spent in that given period. Therefore big spenders will be differentiated with other customers such as MVP or VIP.
    d. **T** represents the age of the customer in whatever time units chosen (weekly, in the above dataset). This is equal to the duration between a customer's first purchase and the end of the period under study.



e.
f. Average order value = Revenue / Transaction per customer
g. **Profit Margin** Profit margin is the commonly used profitability ratio. It represents how much percentage of total sales has earned as the gain.
h. **Purchase Frequency** is the average number of purchases made by a customer over a defined period of time (typically one month or one year). It is the sum of total number transactions divided by total number customers.
i. **Repeat rate** shows you the percentage of your current customer base that has come back to shop again.
j. **Churn Rate** is the annual percentage rate at which customers stop subscribing.
k. **Customer lifetime value,** lifetime customer value, or life-time value is a prediction of the net profit attributed to the entire future relationship with a customer.

Lifetime Value

The LTV is an important building block in campaign design and marketing mix management. Although targeting models can help to identify the right customers to be targeted, LTV analysis can help to quantify the expected outcome of targeting in terms of revenues and profits. The LTV is also important because other major metrics and decision thresholds can be derived from it. For example, the LTV is naturally an upper limit on the spending to acquire a customer, and the sum of the LTVs for all of the customers of a brand, known as the customer equity, is a major metric for business valuations. Similarly to many other problems of marketing analytics and algorithmic marketing, LTVmodeling can be approached from descriptive, predictive, and prescriptive perspectives.

Customer lifetime value can also be defined as the monetary value of a customer relationship, based on the present value of the projected future cash flows from the customer relationship.Customer lifetime value is an important concept in that it encourages firms to shift their focus from quarterly profits to the long-term health of their customer relationships. Customer lifetime value is an important metric because it represents an upper limit on spending to acquire new customers.For this reason it is an important element in calculating payback of advertising spent in marketing mix modeling.