Hollownest AB: Case Study

1 UNDERSTANDING THE TASK

2 DATA SANITIZATION AND VALIDATION

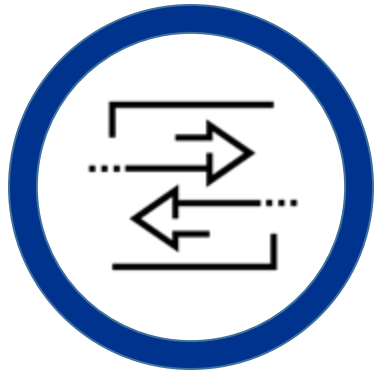
3 ANALYTICS & RESULTS

4 PRESENTATION

Data Investigation

Customer DB
Contains customer related information

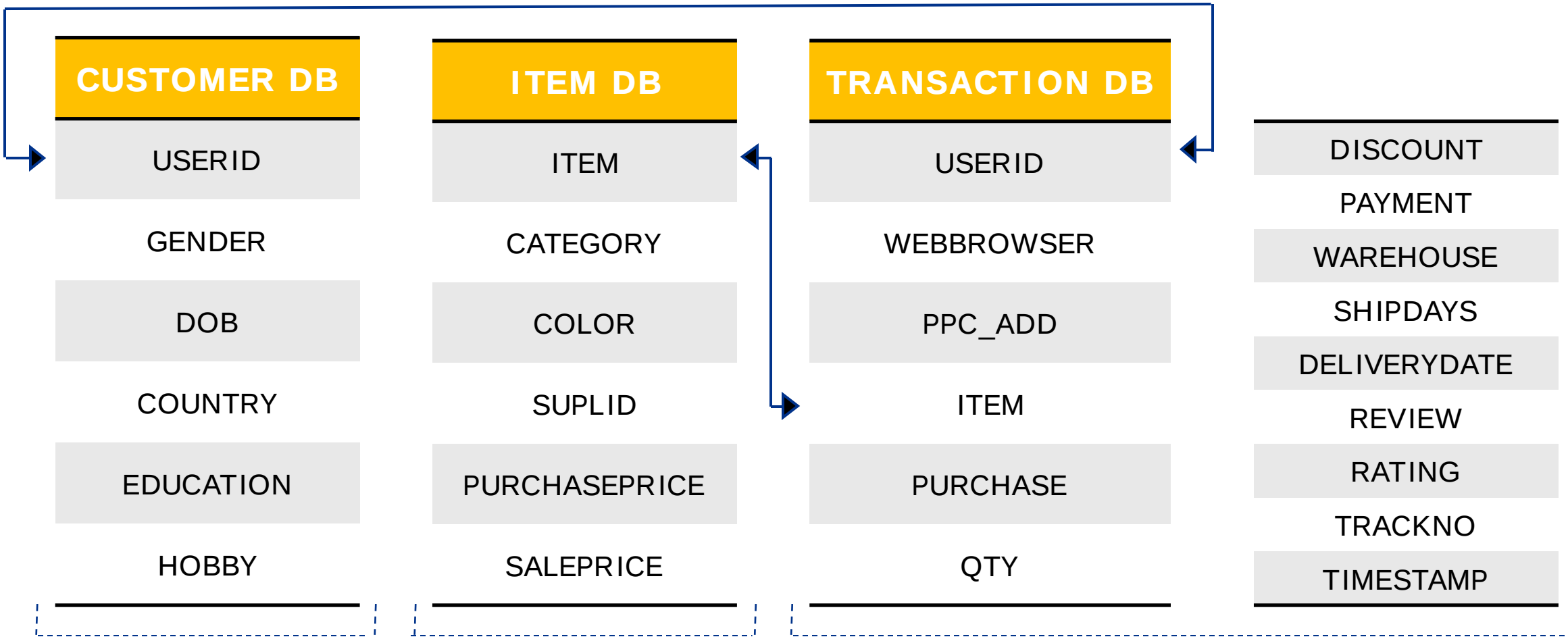
Item DB
Contains item/product related information



Transaction DB
Contains details of the transaction that the customer performed

- Customer DB has 574 unique **UserID** and Transaction DB has 574 unique **UserID** as well as 'NA'
- Item DB has 10,944 unique items in the column **Item** and each of the item is listed 200 times in the table whereas in the transaction DB, there are the same 10,944 unique items in the column **Item** but in this table each of the item occurring ranges from 140 - 266

Column Relationship in DB



Column Relationship in DB

- The data consists of 3 tables: **(1) Customer DB, (2) Item DB, (3) Transaction DB**

- In Customer DB, **UserID** is unique:
  - It can be used as a primary key to join with ‘Transactions Table’

- In Item DB, **Item** is not unique:
  - Thus, it cannot be used as a primary key to join with ‘Transactions Table’
  - In the ‘Items Table’, multiple values occur for one item and the table lacks a unique key identifier
  - Each item occurs 200 times in ‘Items DB’ while the ‘Transactions DB’ all the Items are present but not present 200 times each.

Understanding the Data

- The data has NA and empty values within variables. The file attached below contains detailed data investigation to understand the data (please click the icon to open it)

- To join the data, we need to have a unique identifier or a primary key in **'Item DB'.** I will propose a solution that help normalizes the tables and helps gets rid of these anomalies.

Manipulating ItemDB for joining the TransactionDB

- Investigating further led me to the conclusion that for each item in ItemsDB, category, color, purchase price and sale price are the same but Supplier ID is different

- Item column could be used as a primary key in joining with transaction table if Supplier ID column is excluded

	ITEM	CATEGORY	COLOR	SUPLID	PURCHASEPRICE	SALEPRICE
1	127521	PAJAMAS	PINK	41560EE	1867.09	2800
2	127521	PAJAMAS	PINK	18767GG	1867.09	2800
3	127521	PAJAMAS	PINK	43136JJ	1867.09	2800
4	127521	PAJAMAS	PINK	48094CC	1867.09	2800
5	127521	PAJAMAS	PINK	46396XX	1867.09	2800
6	127521	PAJAMAS	PINK	42596VV	1867.09	2800
7	127521	PAJAMAS	PINK	14656DD	1867.09	2800
8	127521	PAJAMAS	PINK	14656DD	1867.09	2800
9	127521	PAJAMAS	PINK	41560MM	1867.09	2800

Opportunities identified from the Data

- Data can be normalized with the presence of a **Unique Primary key** in all the tables and Supplier ID can become accessible for information analysis

- In many variables lot of NA and blanks are present. When joining databases, this can lead to troubles or result in loss of information

- **3803** values of UserID in TransactionDB were NA and merging the CustomerDB and TransactionDB could not make use of this recorded information. This information was unfortunately lost and could not be included in our analysis

Normalizing the Anomalies

*New Variables

*New Tables

CUSTOMER DB
USERID
GENDER
DOB
COUNTRY
EDUCATION
HOBBY

ITEM DB
ITEM
CATEGORY
COLOR
PURCHASEPRICE
SALEPRICE

ITEMSUPPLIER DB
ITEM ID*
ITEM
SUPLID

TRANSACTION DB
TRANSACTION ID*
WEBBROWSER
PPC_ADD
ITEM
PURCHASE
QTY

DISCOUNT
PAYMENT
WAREHOUSE
SHIPDAYS
DELIVERYDATE
REVIEW
RATING
TRACKNO
TIMESTAMP

LINKAGEDB
TRANSACTION ID*
USERID
ITEM ID

Recommendations

- I would recommend having five Tables/DB instead of three: Customer DB, Supplier DB, Item DB, ItemSupplier DB, Transaction DB, Linkage DB.
- Customer DB contains purely customer related data and Customer ID is the Primary Key
- Item DB contains purely item specific information and helps avoids data redundancy
- Item Supplier DB is created separately to relate item with supplier. Variable ‘Item ID’ is created to serve as unique item identity
- Transaction DB contain transaction specific information and a new column Transaction ID is created that also serves as primary key

Recommendations

- I would recommend recreating the ITEM table and make ‘SupplierItem’ table additionally.
  - Incase, we assume Purchase price and Sale Price to be only related to the ITEM and not the supplier
  - ItemsDB has Item, Category, Color, Sale Price, Purchase Price
  - SupplierItem table would have unique Item ID, Item, SupID

  - Incase, we assume Purchase price and Sale Price to be related to the item and as well as the supplier
  - ItemsDB has Item, Category and Color
  - SupplierItem table would have unique Item ID, Item, SupID, Sale Price, Purchase Price

- Data redundancy will be reduced and tables will be normalized:
  - ItemDB table will have one entry for each item and there will not be a need to repeat category or color reducing insertion anomalies.
  - LinkageDB a new table would help join the three tables instead of Transaction DB

Data Analysis

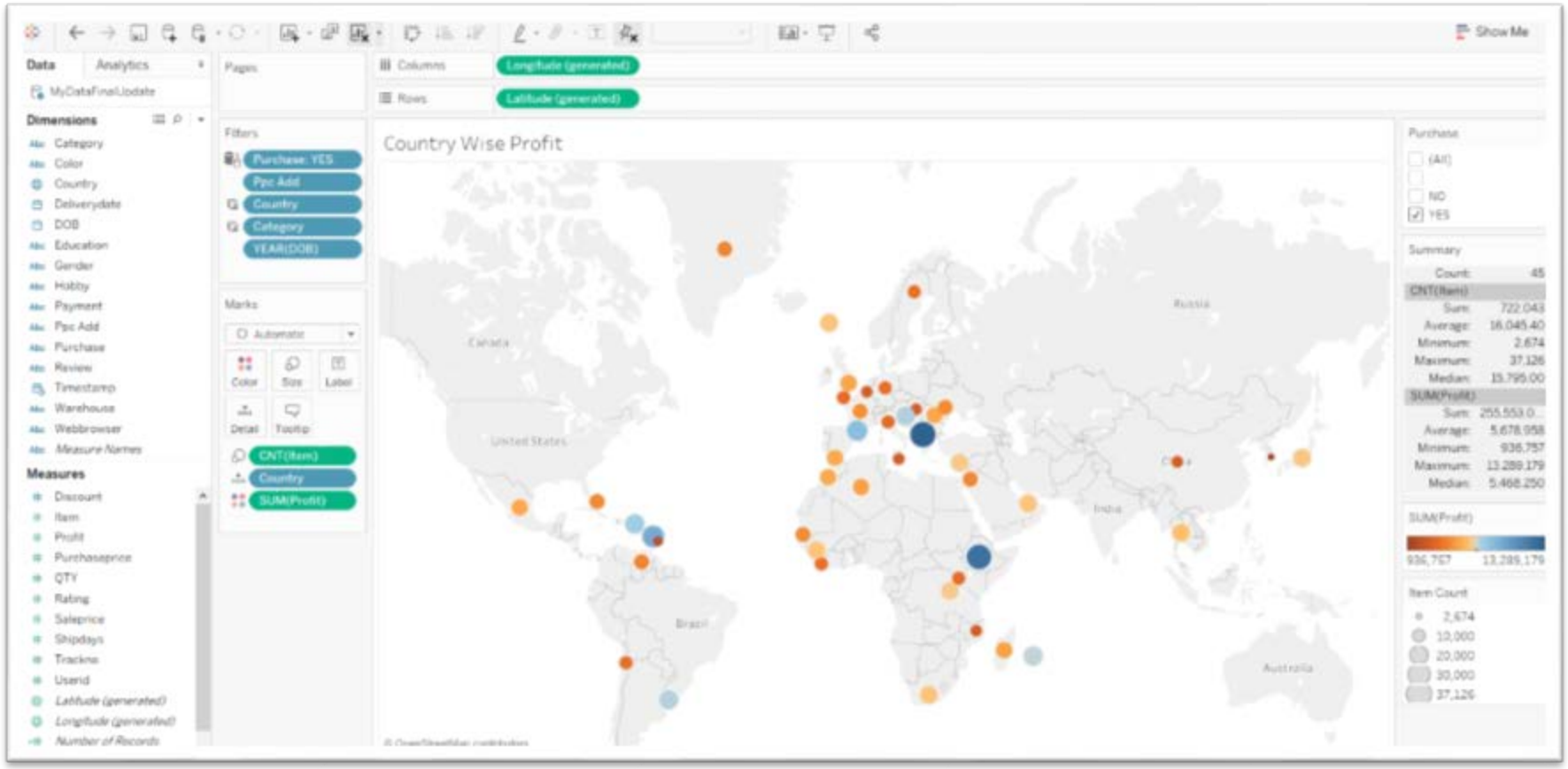
- Assumptions:
  - Discount taken as Percentage
  - Rating 1 assumed lowest and 5 as highest

- Detailed Data Visualization can be found in the Tableau file (attached)

Data Analysis

- Maximum profits came from **Macedonia, UAE, Ethiopia, Guinea, Martinique**

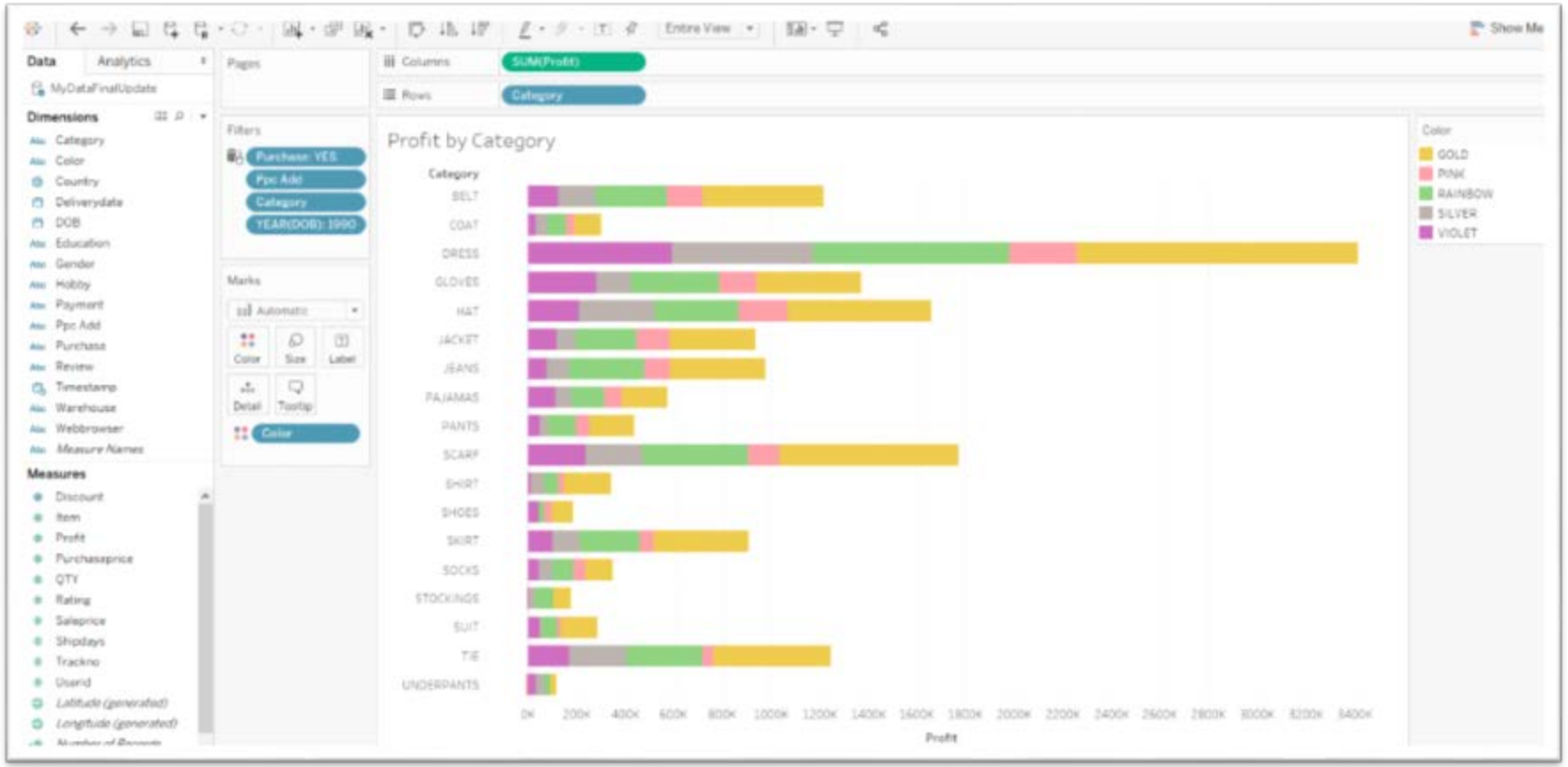
(For in depth view have a look at the chart Country Wise Profit from Tableau Visualizations)



Data Analysis

- **Dress** was the most sold item and **Gold** was the most popular color in the item purchase

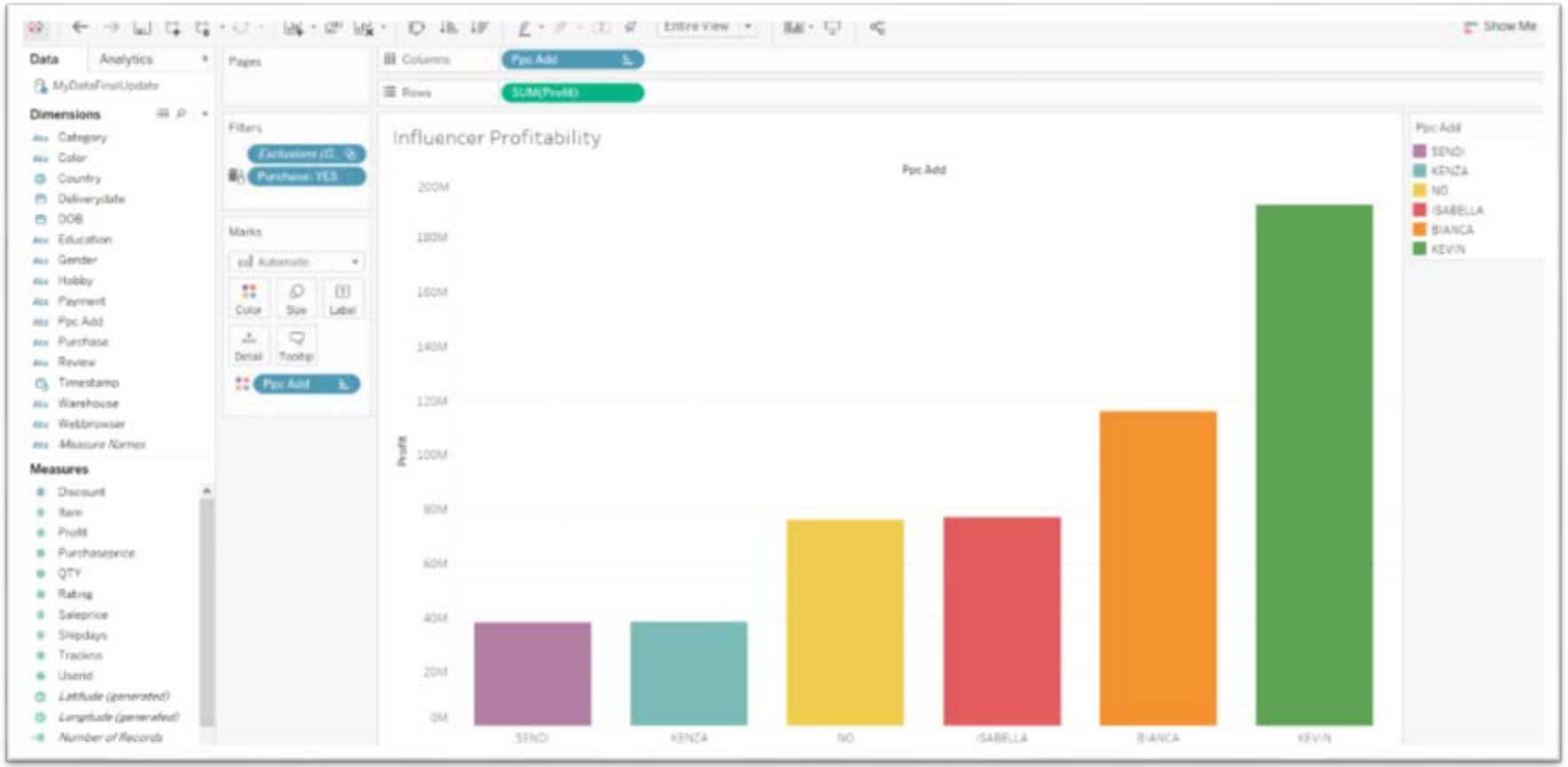
(For in depth view have a look at the chart Profit by Category from Tableau Visualizations)



Data Analysis

- **Kevin** leads the team influencing the sales and profitability followed by Bianca

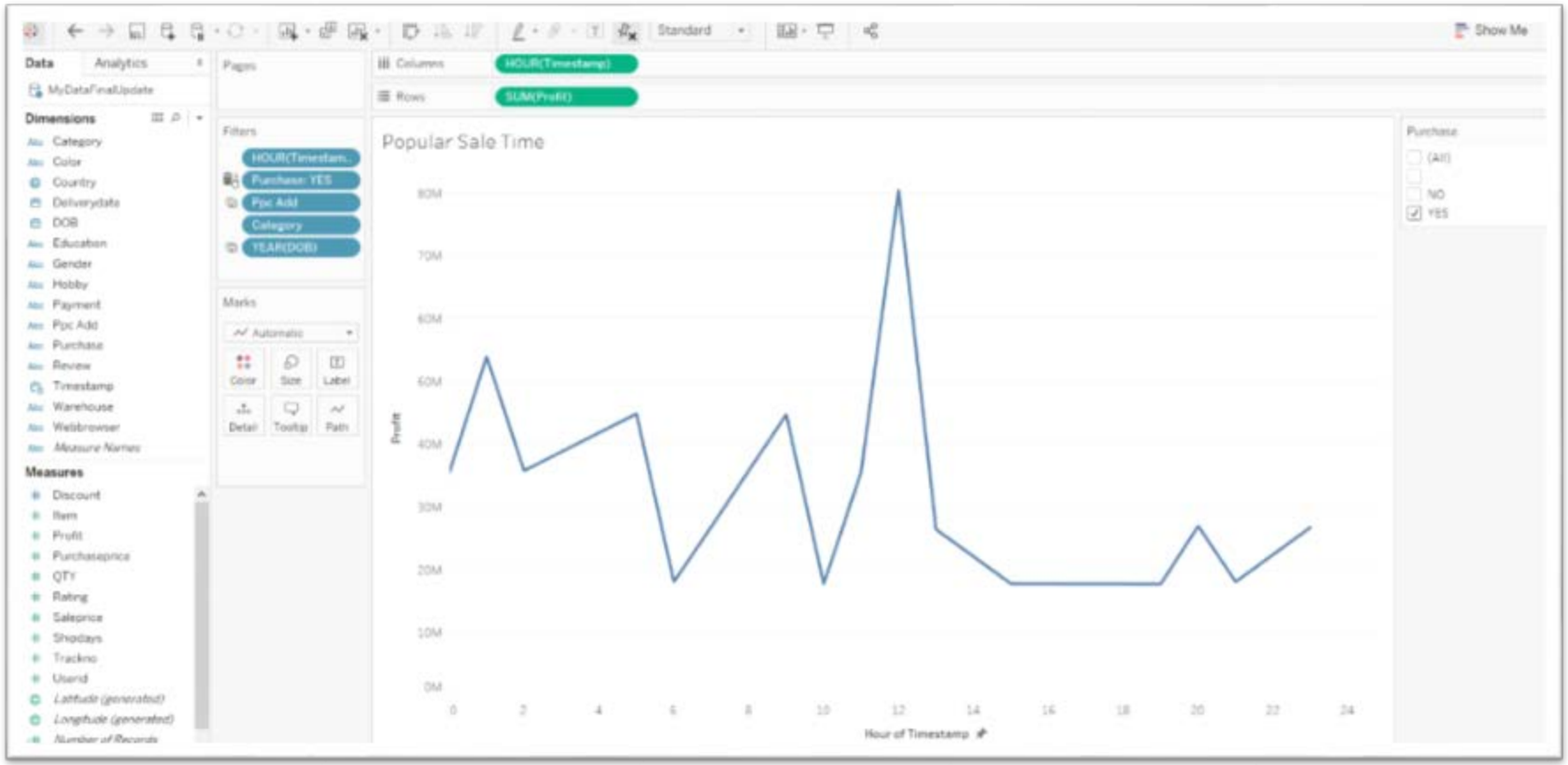
(For in depth view have a look at the chart Influencer Profitability from Tableau Visualizations)



Data Analysis

- **12 PM** is the most popular and profitable time. Profit and sales begin to appreciate **late night till noon**. Profit declines after noon

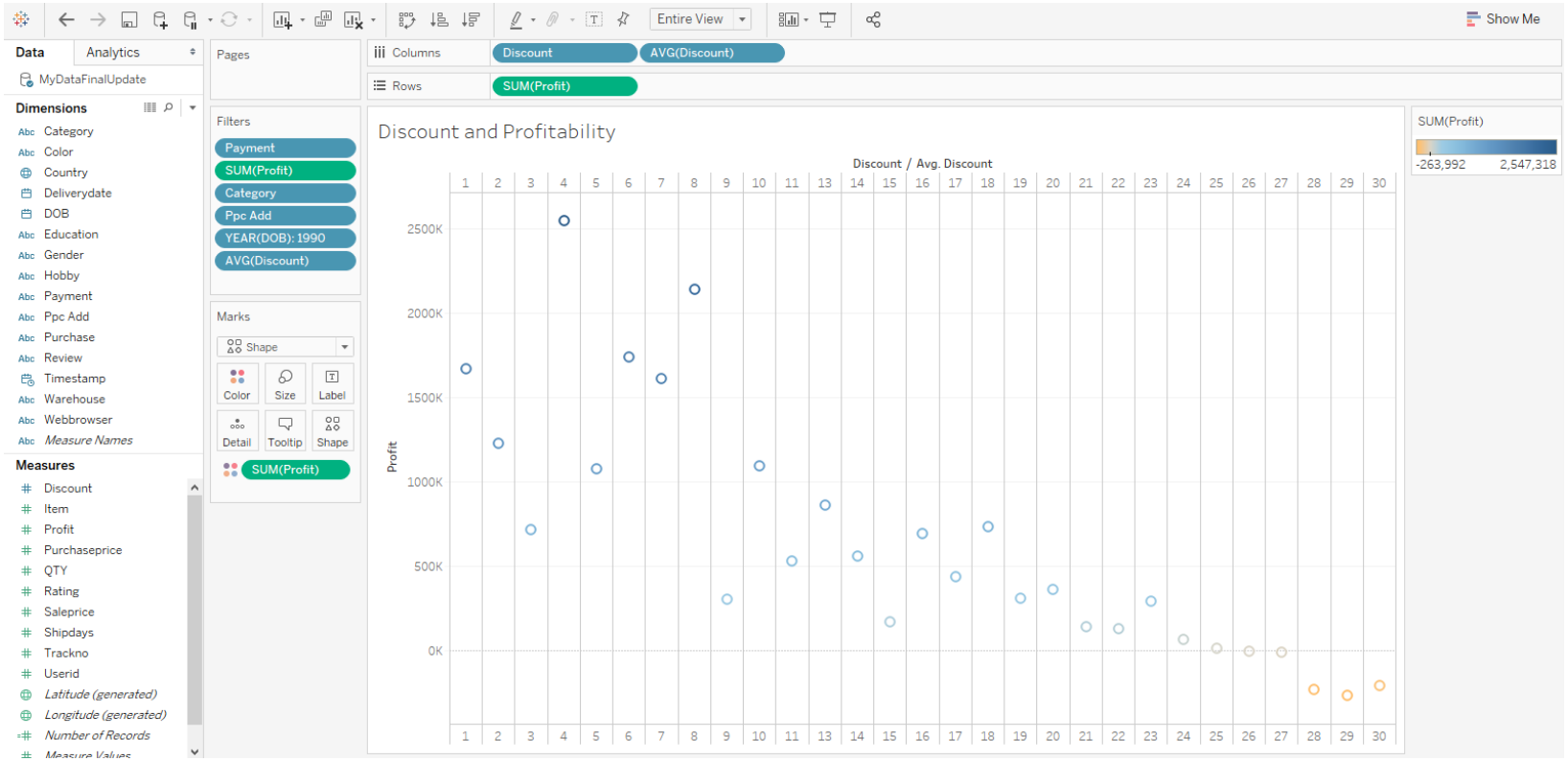
(For in depth view have a look at the chart Popular Sale Time from Tableau Visualizations)



Data Analysis

- There does not seem to be a positive correlation between increasing **discount and profitability**. Maximum profitability is at **4% discount** and any discount greater than that has lesser profitability

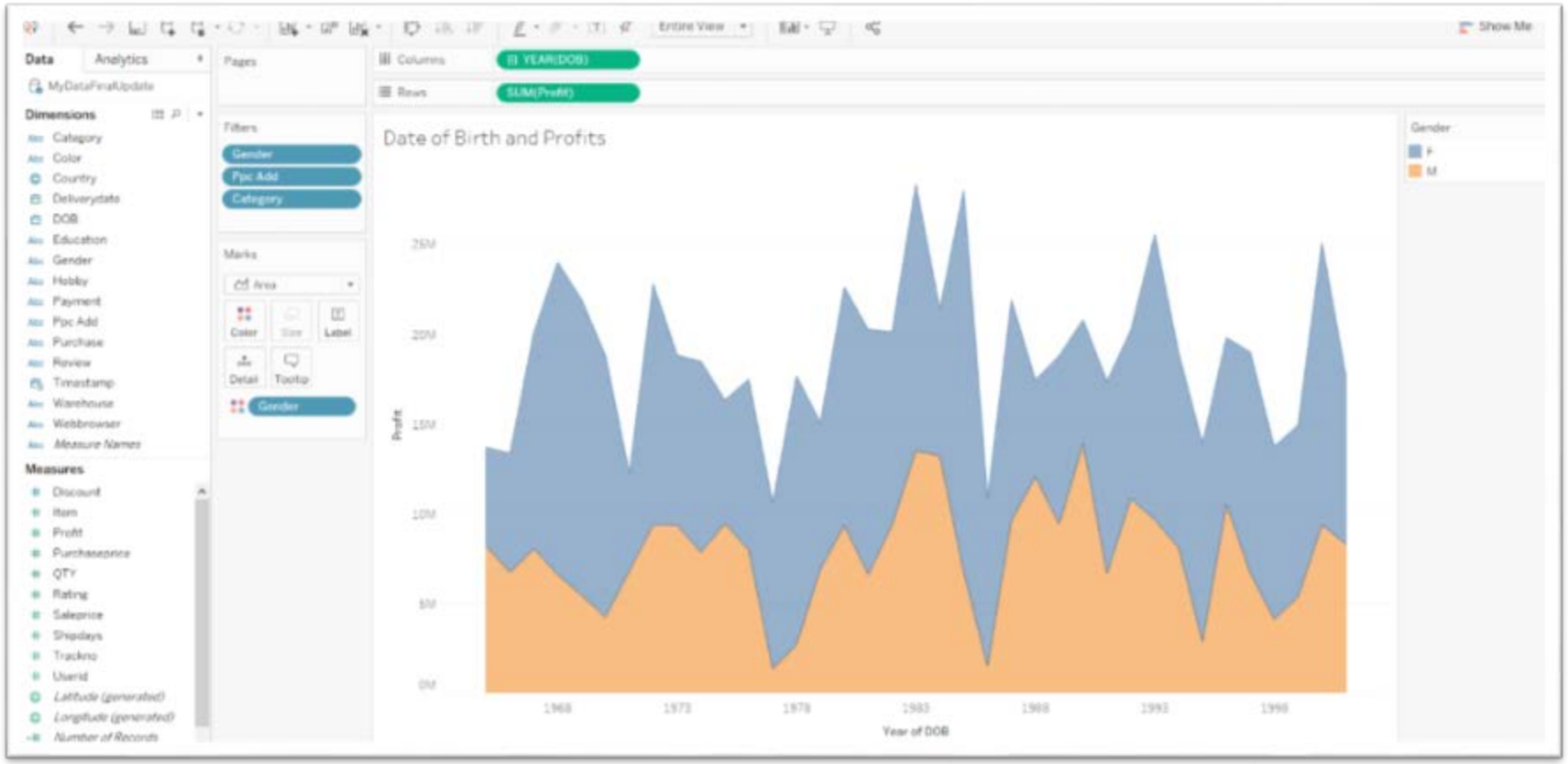
(For in depth view have a look at the chart Discount and Profitability from Tableau Visualizations)



Data Analysis

- Across all date of births, old and young, profitability has been greater contributed by **females**

(For in depth view have a look at the chart Date of Birth and Profits from Tableau Visualizations)



Suggestions based on Data Analysis

- Noon was identified as the time when maximum sales were made. Try selecting this hour for persuasive advertisement or new deals or new products

- In the analysis we found dress as the most sold item and people with hobby of embroidery proved most profitable. Perhaps what could be interesting insight from this analysis. Introduce items that are related with the hobbies to boost the sales and profitability.

- Offering discount beyond 10% has shown to be a unwise strategy. Best range of discount look to be between 1-10%. 4% looked most profitable