

# Yoochoose - RecSys Challenge 2015

<b>Summary</b>	Plotting insights with respect to Pricing, Recommendations, Search and Promotions
<b>URL</b>	<a href="https://recsys.yoochoose.net/challenge.html">https://recsys.yoochoose.net/challenge.html</a>
<b>Dashboard Link</b>	<a href="https://na139.lightning.force.com/wave/wave.app#dashboard/0FK4W000000cnbpWAA">https://na139.lightning.force.com/wave/wave.app#dashboard/0FK4W000000cnbpWAA</a>
<b>Category</b>	Web

[Yoochoose - RecSys Challenge 2015](#)

[Abstract](#)

[Dataset File Description](#)

[The training data comprises two different files:](#)

[Working on Datasets using XSV](#)

[Working on the dataset using Trifacta](#)

[Working on the dataset using Snowflakes](#)

[Creating Insights using Salesforce Einstein Analytics](#)

[Pricing:](#)

[Search & Recommendations:](#)

[Promotions:](#)

[Answered Questions related to Dashboards](#)

[Which columns are dimensions, which columns are measures?](#)

## Abstract

- In this assignment, we are provided with a sample dataset and asked to analyze and build an analytical dashboard as a Proof-of-concept to illustrate the value of data driven analytics.
- The themes to be considered include: Pricing, Promotion, Search, Recommendations

- We will analyze the data using tools (xsv, Trifacta), stage data using Snowflake and build a dashboard using Salesforce Einstein analytics.

## Dataset File Description

The training data comprises two different files:

**yoochoose-clicks.dat** - Click events. Each record/line in the file has the following fields:

- **Session ID** – the id of the session. In one session there are one or many clicks.
- **Timestamp** – the time when the click occurred.
- **Item ID** – the unique identifier of the item.
- **Category** – the category of the item.

**yoochoose-buys.dat** - Buy events. Each record/line in the file has the following fields:

- **Session ID** - the id of the session. In one session there are one or many buying events.
- **Timestamp** - the time when the buy occurred.
- **Item ID** – the unique identifier of the item.
- **Price** – the price of the item.
- **Quantity** – how many of this item were bought.

## Working on Datasets using XSV

XSV commands to Sample the dataset and other Wrangling operations like Filtering, Joining, Slicing, Search, cleaning, Sampling

1. **Headers:** xsv header command applied to list the **headers** of both the files

```
C:\Users\kaviy\Desktop\Sem3>xsv headers ycclick.csv
1 5
2 2014-04-07T17:13:46.713Z
3 214530776
4 0

C:\Users\kaviy\Desktop\Sem3>xsv headers ycbuy.csv
failed to open ycbuy.csv: The system cannot find the file specified. (os error 2)

C:\Users\kaviy\Desktop\Sem3>xsv headers ycbuys.csv
1 420374
2 2014-04-06T18:44:58.314Z
3 214537888
4 12462
5 1
```

## 2. Stats: Calculated the stats using xsv stats command of the datasets

```
C:\Users\kaviy\Desktop\Sem3>xsv stats ycclick.csv | xsv table
field      type      sum      min      max      min_length  max_length  mean      stddev
5          Integer  702230708465  5      7862270  1          7          669700.0295305282  423498.613216545
2014-04-07T17:13:46.713Z  Unicode  225305088072789  2014-04-01T03:00:08.250Z  2014-08-11T23:33:48.331Z  24         24         214867880.7646461  7858889.10714466
214530776  Integer  214507239      643078907  5          9          9          0.5777622010824196  1.1118696056571284
0          Unicode  0              S          1          2

C:\Users\kaviy\Desktop\Sem3>xsv stats yccbuys.csv | xsv table
failed to open yccbuys.csv: The system cannot find the file specified. (os error 2)

C:\Users\kaviy\Desktop\Sem3>xsv stats ycbuys.csv | xsv table
field      type      sum      min      max      min_length  max_length  mean      stddev
420374     Integer  5677494810300  11      10777833  2          8          5414486.145769381  3076926.2774466984
2014-04-06T18:44:58.314Z  Unicode  231738667292004  2014-04-01T03:05:31.743Z  2014-08-16T02:51:24.731Z  24         24         221003425.87988177  51270413.58802422
214537888  Integer  231738667292004  214507331  1178837797  9          10         0.5777622010824196  1.1118696056571284
12462     Integer  1309270866     0        334998    1          6          1248.6191888991875  4388.846050711053
1          Integer  605827         0        30        1          2          0.5777622010824196  1.1118696056571284
```

## 3. Count: Calculated the count of both the datasets

```
C:\Users\kaviy\Desktop\Sem3>xsv count ycclick.csv
1048575

C:\Users\kaviy\Desktop\Sem3>xsv count ycbuys.csv
1048575
```

## 4. Sampled: Sliced and Sampled ycclick.csv from 1048575 rows to 65000 rows and ycbuys.csv from 1048575 rows to 65000

```
C:\Users\kaviy\Desktop\Sem3>xsv slice ycbuys.csv -s 65000 | xsv table
```

```
C:\Users\kaviy\Desktop\Sem3>xsv slice ycclick.csv -s 65000 | xsv table
```

## Working on the dataset using Trifacta

Missing value analysis and data imputation

Feature Extraction: Event hour, Date, Day of Week, Timezone.

Applied some Wrangling operations like filtering, regex, Aggregate, Groupby.

Finally created the recipe for Transformation to Wrangle the data

### 1. List of Features extracted after joining the datasets - ycclick.csv and ycbuys.csv

The screenshot shows the Trifacta Wrangler interface for the 'ycbuys' dataset. The 'Recipe' tab is active, displaying a list of 10 wrangling steps:

- 1 Rename column2 to 'SessionId'
- 2 Rename column4 to 'ItemId'
- 3 Rename column5 to 'Price'
- 4 Rename column6 to 'Quantity'
- 5 Create SellingPrice from Price \* Quantity
- 6 Split column3 on delimiters matching 'T' into 2 columns
- 7 Split column2 on delimiters matching 'Z' into 2 columns
- 8 Rename column1 to 'shopDate'
- 9 Rename column3 to 'shopTimestamp'
- 10 Create dayofweek from WEEKDAY(shopDate)

The dataset summary at the bottom indicates 11 Columns, 65,003 Rows, and 4 Data Types.

### 2. Recipe of the datasets - wrangling performed on the datasets

### 3. Inner Join implemented with ItemId as a primary key

## Assignment 1 - Part 2

Full data

Join - Keys & Conditions

Join type: Inner

Join keys: # ItemId = (Equal to) # ItemId (90% match)


Results summary:  
Based on current samples  
Rows in Current: 65923  
Rows in Joined-in: 65003  
Rows in Output: 159599

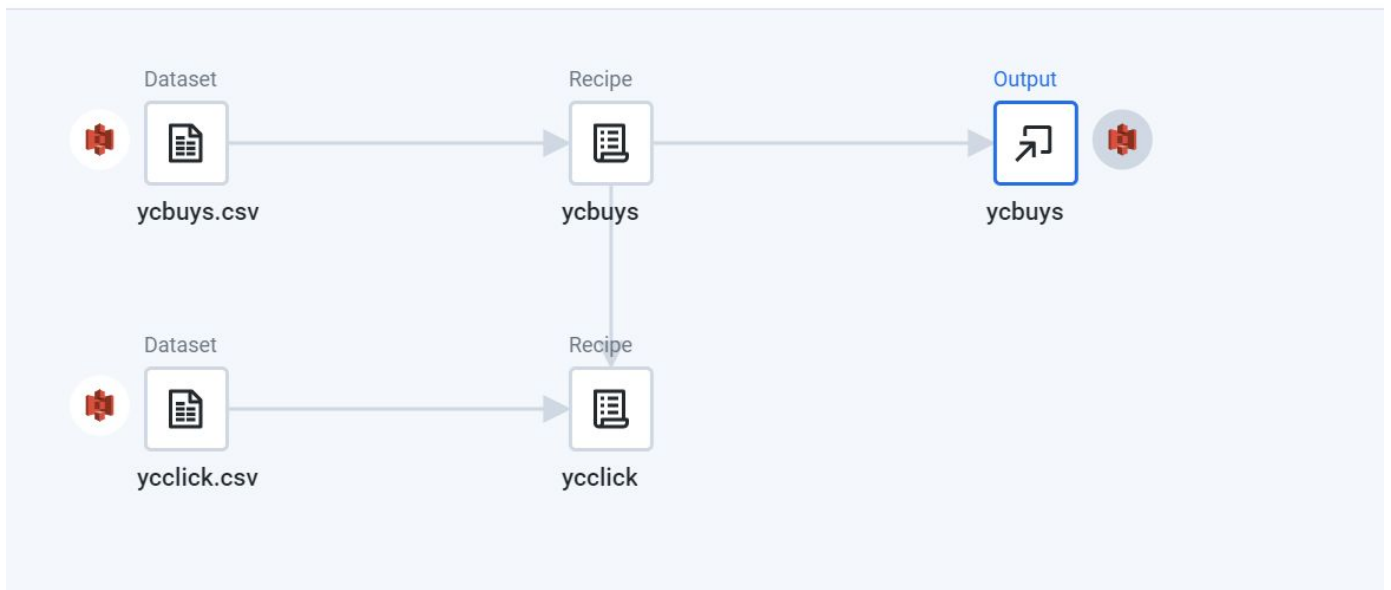
65,923 Rows in ● 65,003 Rows in ● 159,599 Rows in Output

Show only: ☒ Included Rows ☐ Excluded Rows

#	ItemId	#	ItemId
214.51M - 643.08M	214819464	214.51M - 214.85M	214819464
	214819464		214819464
	214819464		214819464
	214819378		214819378
	214819378		214819378
	214819378		214819378
	214819380		214819380
	214819380		214819380
	214819380		214819380
	214843494		214843494
	214843494		214843494
	214843494		214843494
	214717867		214717867

4. Data profile - Created recipe of both datasets and then implemented join

 Yoochoose Ecommerce



5. Final dataset published in Trifacta

Untitled (deleted) > 2020\_Ecomm\_Sales\_Data - 3 (deleted)  
**Job 225470**  
 Finished 09/19/2020

[Overview](#) [Output Destinations](#) [Profile](#) [Dependencies](#)

### Completed stages

✓ **Transform with profile**  
 Completed 09/19/2020, started 09/19/2020 • Ran for 6 min  
 Environment Spark

100% valid values 0% mismatching values 0% missing values

[View steps and dependencies](#) [View profile](#)

✓ **Publish**  
 Completed 09/19/2020, started 09/19/2020 • Ran for <1 sec  
 Activity

2020_Ecomm_Sales_Data - 3.json	✓ <u>Completed</u>
--------------------------------	--------------------

## Working on the dataset using Snowflakes

The final dataset imported from trifecta after performing Data wrangling is staged on snowflakes.

The staged file will be used in Salesforce - Einstein Analytics to derive several insights

1. Creating Table: (Owner:SYSADMIN)

## Create Table

Table Name \* YC\_BuyClick

Schema Name PUBLIC

Comment

Columns \*



Add



Remove

Name	Type	Not Null	Default
Buy_SessionId	VARCHAR (50)	<input type="checkbox"/>	
shopDate	DATE	<input type="checkbox"/>	
dayofweek	VARCHAR (10)	<input type="checkbox"/>	
shopDays	VARCHAR (50)	<input type="checkbox"/>	
shopTimesta...	VARCHAR (50)	<input type="checkbox"/>	

[Show SQL](#)

Cancel

Finish

## 2. Loading data

Type			
VARCHAR			
VARCHAR			
VARCHAR			
VARCHAR			
VARCHAR			
VARCHAR(20)	true	NULL	
VARCHAR(50)	true	NULL	
DATE	true	NULL	

### Staging Files...

Encrypted Files



Staging Files



### Load Results

Loaded	File	Rows Parsed	Rows Loaded
✓	yoochoose.csv	103867	103867

OK

### 3. Preview Data: The dataset is staged successfully

New Worksheet

Find database objects

Starting with...

- DEMO\_DB
- SNOWFLAKE\_SAMPLE\_DATA
- UTIL\_DB
- YOOCHOOSE
  - INFORMATION\_SCHEMA
  - PUBLIC
    - Tables
      - YC\_BUYCLICK

No Views in this Schema

Run All Queries Saved 28 seconds ago

SYSADMIN COMPUTE\_WH (XS) Select Database Select Schema

Results Data Preview

Table: YOOCHOOSE.PUBLIC.YC\_BUYCLICK

Filter result...

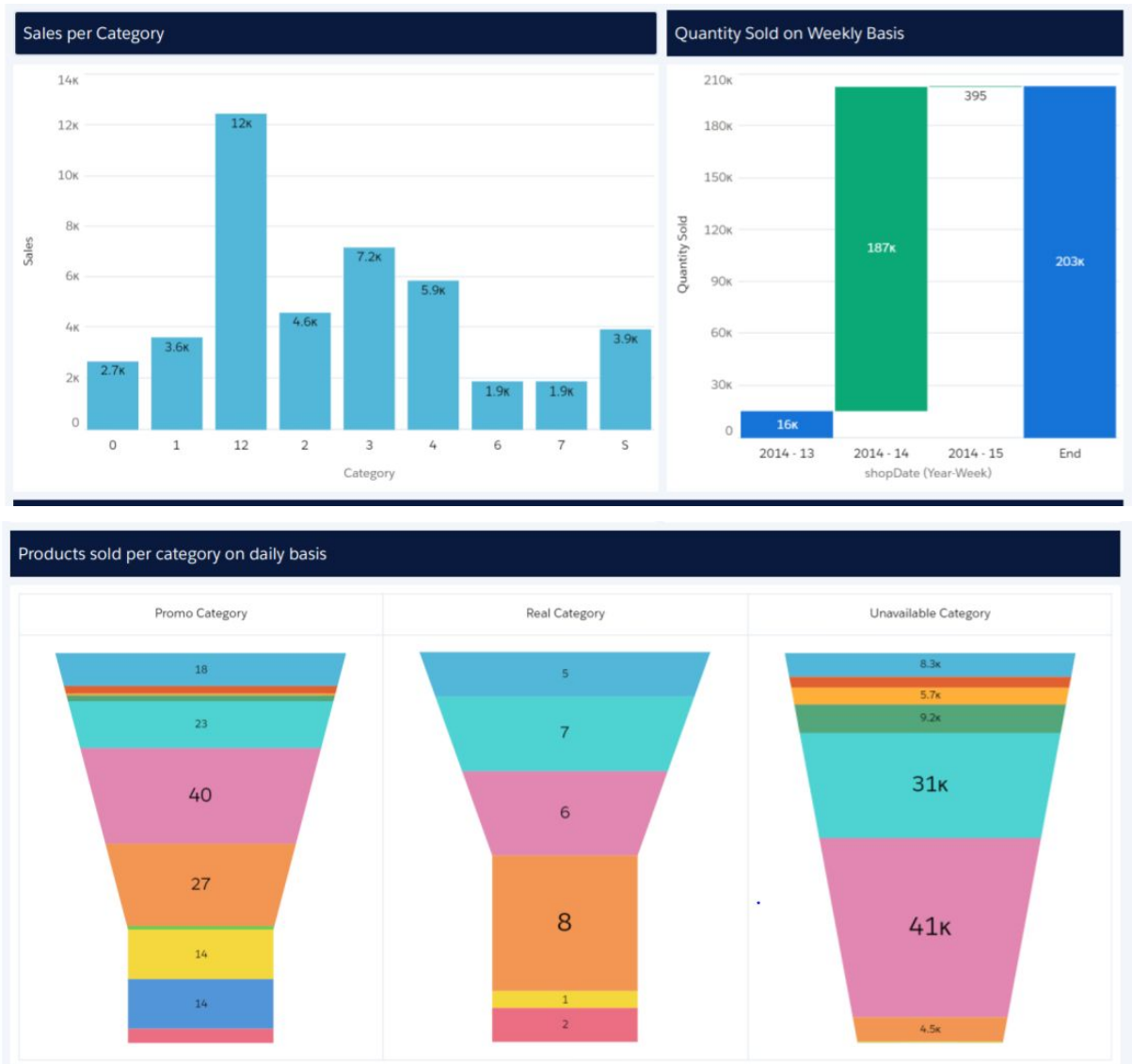
Row	BUY_SESSIONID	SHOPDATE	DAYOFWEEK	SHOPDAYS	SHOPTIMESTAMP	ITEMID	PRICE	QUANTITY	SELLINGPRICE
1	420374	2014-04-06	7	Sunday	44:58.3	214537888	12462	1	12462
2	420374	2014-04-06	7	Sunday	44:58.3	214537888	12462	1	12462
3	420374	2014-04-06	7	Sunday	44:58.3	214537888	12462	1	12462
4	420374	2014-04-06	7	Sunday	44:58.3	214537888	12462	1	12462
5	420374	2014-04-06	7	Sunday	44:58.3	214537888	12462	1	12462
6	420374	2014-04-06	7	Sunday	44:58.3	214537888	12462	1	12462
7	105892	2014-04-06	7	Sunday	12:54.1	214537888	12462	1	12462
8	105892	2014-04-06	7	Sunday	12:54.1	214537888	12462	1	12462

Columns

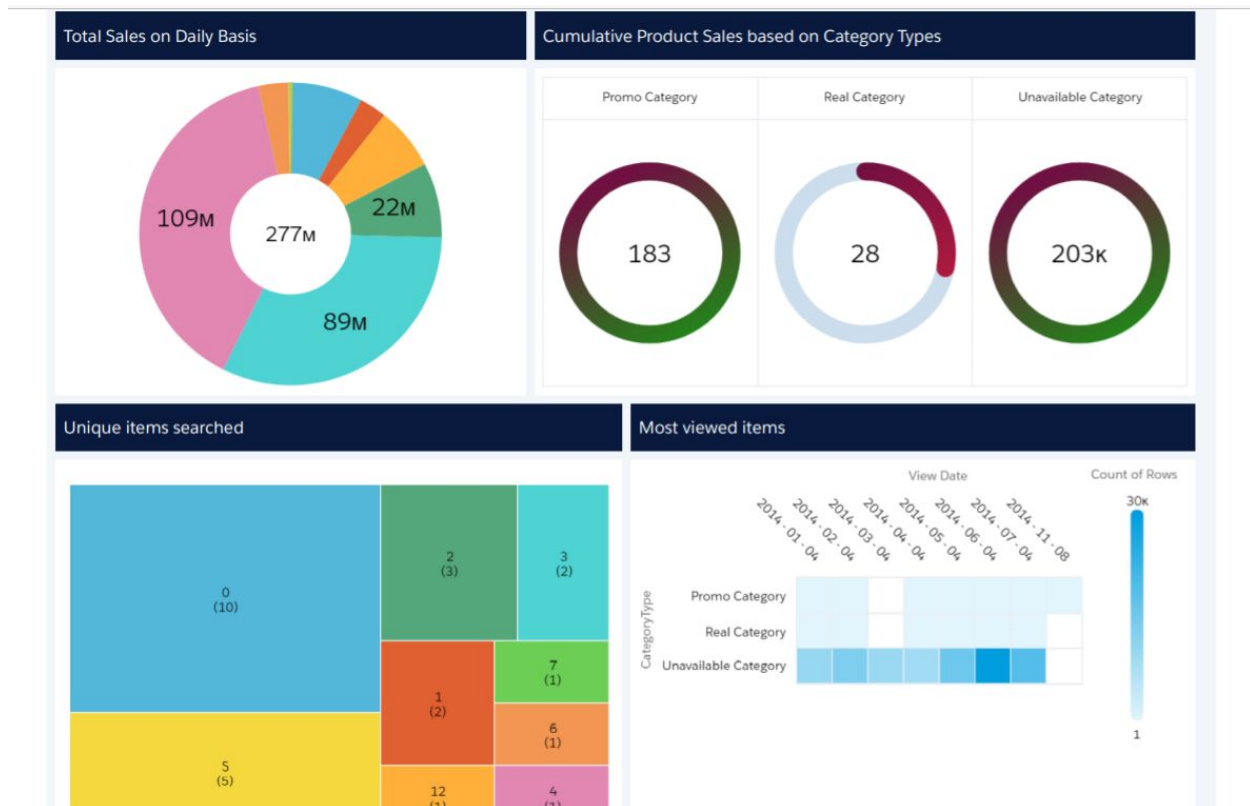


# Creating Insights using Salesforce Einstein Analytics

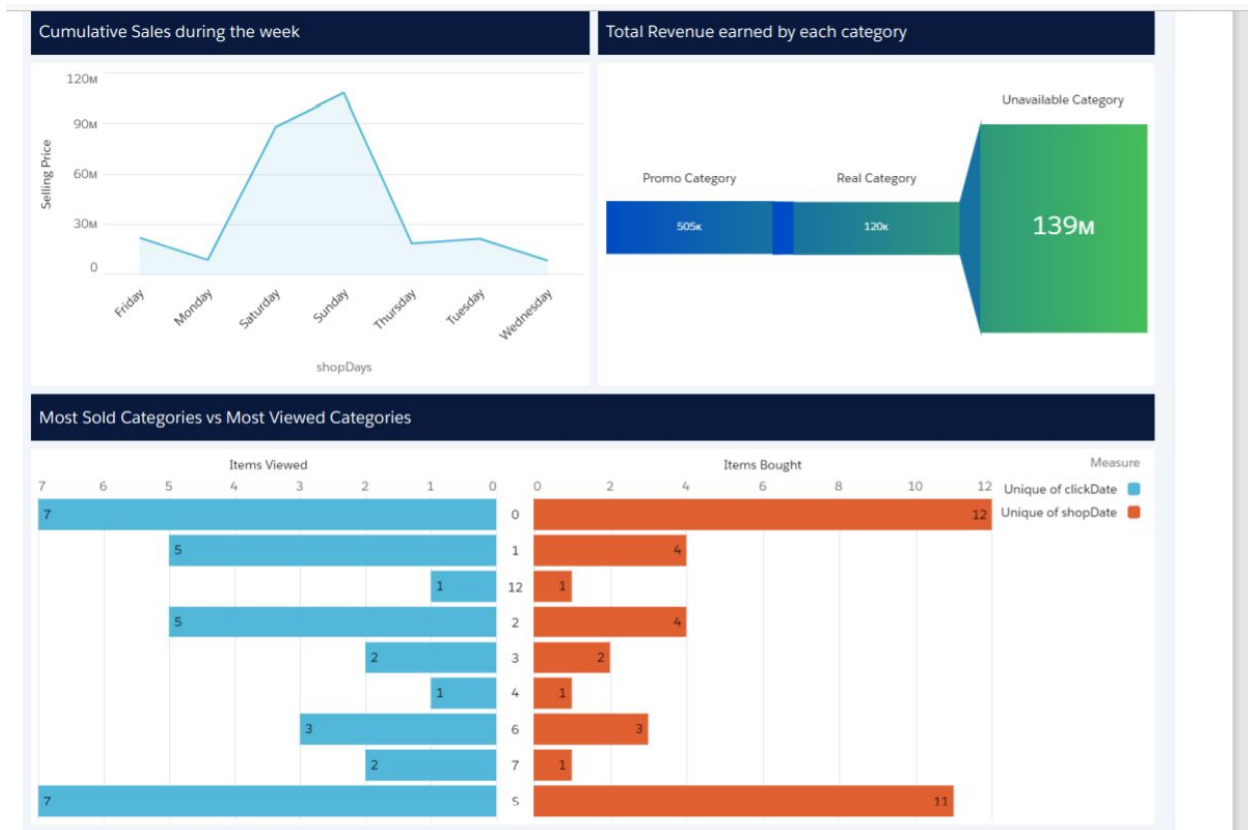
## Pricing:



## Search &amp; Recommendations:



## Promotions:



## Answered Questions related to Dashboards

Which columns are dimensions, which columns are measures?

Dimensions -

1. SessionID
2. ItemID
3. Category
4. CategoryType
5. clickTimestamp
6. shopTimestamp

Time -

1. shopDate
2. clickDate

Measures -

1. Price
2. Quantity

How would you generate new dimensions? What will you do to summarize measures?

1. New dimensions are created using wrangling tools like Trifacta, XSV tool, google DataPrep, tabula etc or can be performed in Python using libraries like Numpy, Pandas, Theano and R using libraries - Purrr, Dplyr, JOOnline
2. We have summarized the measures in form of intuitive graphs and tables

Who would use this dashboard?

1. Business Analyst
2. Data Analyst
3. Data Scientist
4. Marketing Manager

What value would be generated using this dashboard ?

1. Helps business to decide the outcomes that needs to be achieved, in order to achieve their goals
2. making marketing decisions and mitigate risks
3. Identifies why expected results are not achieved by looking down into data chain