DevAx Online Workshop Oct 2021- AWS Vietnam

Data Story-telling with AWS QuickSight

US E-COMMERCE DATASET

BY QUYEN DINH

GOAL

- Creating a dashboard
- Finding insights about future trends and improvements for the companies and/or products

TRACK

- Provided dataset
- US E-Commerce dataset

CONTENTS

1. EXPLORATORY DATA ANALYSIS

OVERVIEW ABOUT US E-COMMERCE DATASET

	Transaction_id	customer_id	Date	Product	Gender	Device_Type	Country	State	City	Category	Customer_Login_type	Delivery_Type	Quantity	Transaction Start	Transaction_Result	Amount US$	Individual_Price_US$	Year_Month	Time
0	40170	1348959766	14/11/2013	Hair Band	Female	Web	United States	New York	New York City	Accessories	Member	one-day deliver	12	1	0	6,910	576	13-Nov	22:35:51
1	33374	2213674919	05/11/2013	Hair Band	Female	Web	United States	California	Los Angles	Accessories	Member	one-day deliver	17	1	1	1,699	100	13-Nov	06:44:41
2	14407	1809450308	01/10/2013	Hair Band	Female	Web	United States	Washington	Seattle	Accessories	Member	Normal Delivery	23	1	0	4,998	217	13-Oct	00:41:24

- Provided file type: **.csv**
- Number of table: **1**
- Number of columns: **16**
- Number of rows: **65, 535**
- Header in table: **True**
- Data need to be cleaned up: **Yes**

SOME IMPORTANT ASSUMPTION ABOUT THE DATA

Since there is no communication to get more inputs from that E-commerce company, we need to make some assumption about the data for the further analysis:

- **Assumption 1:** In this context, cost is considered as Revenue

  "Amount US$": **Total cost** for the order => **Total revenue** from the order

  "IndividualPriceUS$": **Cost** for each item => **Revenue** on each item

- **Assumption 2:** After checking the value of the 2 columns ("Transaction_start" and "Transaction_result, we interpreted the Transaction as followed:

  Transaction_start: Total Transaction.

  Transaction_result:

  + Label 1: Completed Transactions
  + Label 0: Uncompleted/Abandoned/Return Order

```
df.Transaction_Start.value_counts()
```

```
1    65376
Name: Transaction_Start, dtype: int64
```

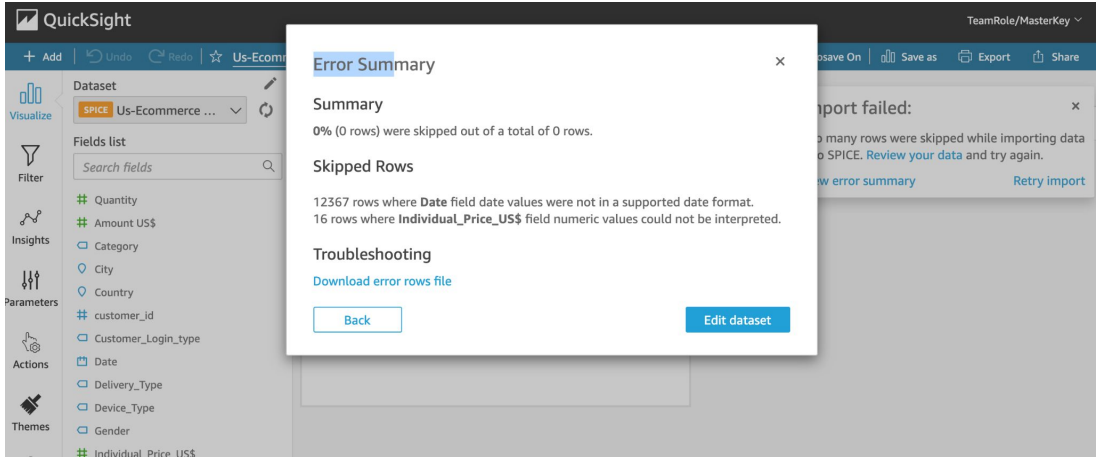
```
df.Transaction_Result.value_counts()
```

```
1    56709
0     8667
Name: Transaction_Result, dtype: int64
```

PROBLEM WITH THE DATA

DATASET IS NOT READY FOR VISUALIZATION ON AWS QUICKSIGHT

Problem 1. The format for Date in the dataset is **DD/MM/YYYY,** while the default in QuickSight is **MM/DD/YY** => data is not valid => Should change the QuickSight Date format to **DD/MM/YYYY** before loading the data into Visualization.



Problem 2. Some missing values in the Individual_Price_US$ column

Amount_US	Individual_Price_US
1,61,850	#VALUE!
1,26,834	#VALUE!

Problem 3. Some data values are not in a desirable format

For example, 2 unique values on “Delivery_Type” is 'one-day deliver' and 'Normal Delivery' => Need to capitalize the first letter for consistency.

EXPLORATORY DATA ANALYSIS WAS DONE USING PYTHON – JUPYTER NOTEBOOK

Notebook is available at my github link

REASON FOR PERFORMING EDA USING PYTHON/JUPYTER NOTEBOOK

- Account was not linked to AWS SageMaker
- More familiar with Python and Jupyter Notebook, didn't have enough time to explore other options on AWS

TASKS DONE

1. Convert the "Date" column to the standard format so that it could be ready to use in AWSQuicksight
2. Remove 159 rows that have Individual_Price_US is #VALUE! (0.24% of the data)
3. Clean up some values to get a better format and consistency:

```
spectacles' => 'Spectacles', 'vessels' => 'Vessels', 'one-day deliver' => 'One-day deliver'
```

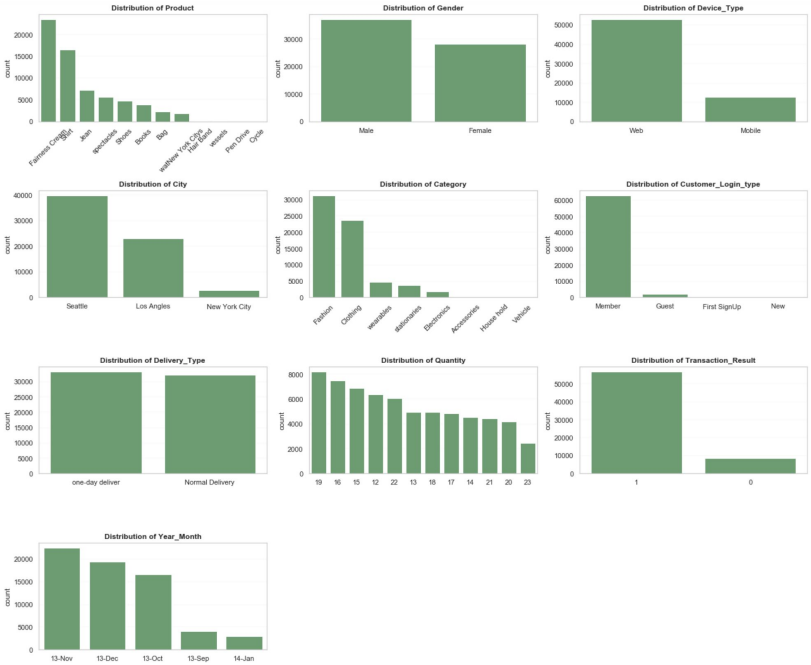
4. Add the column "Weekday": Convert Date to weekday 0 to 6, where 0 is Sunday and 6 is Saturday.
5. Change the "Time" column to only contain the hourly order information (00 to 23) for tracking daily pattern.

IMPORTANT NOTE:

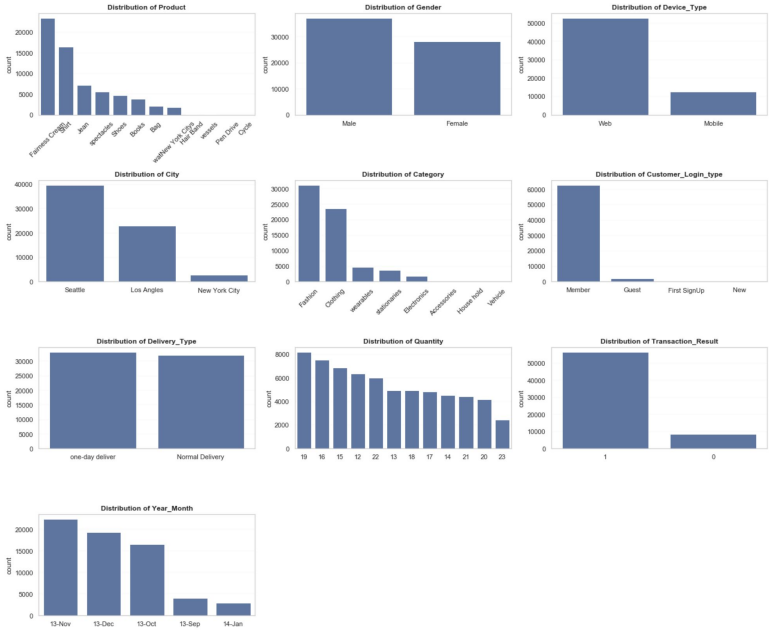
Remove 159 rows that have Individual_Price_US is #VALUE! (0.24% of the data) **didn't affect the data distribution.**

Brief overview of the data distribution on original and cleaned up data (Python Matplotlib library)

Original data



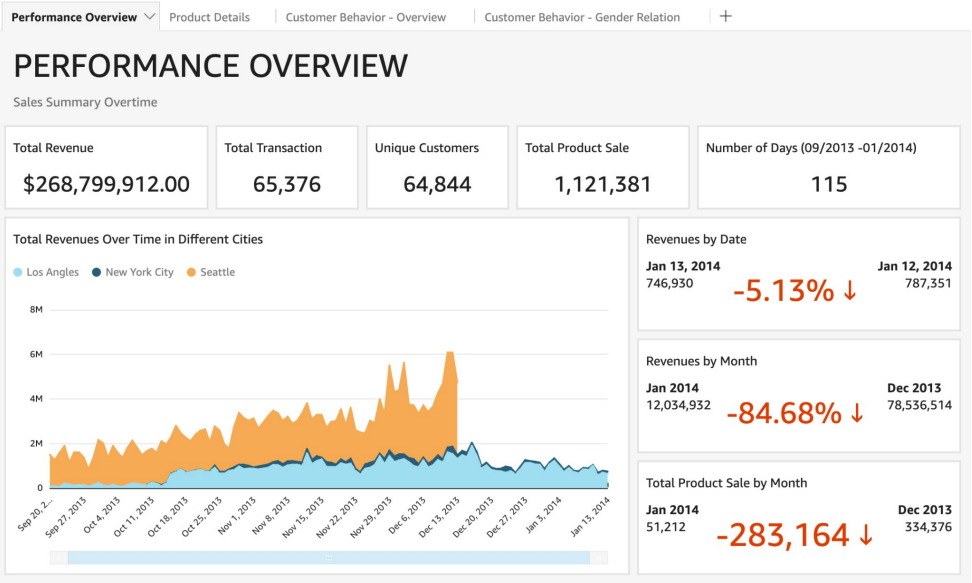
Cleaned-up data

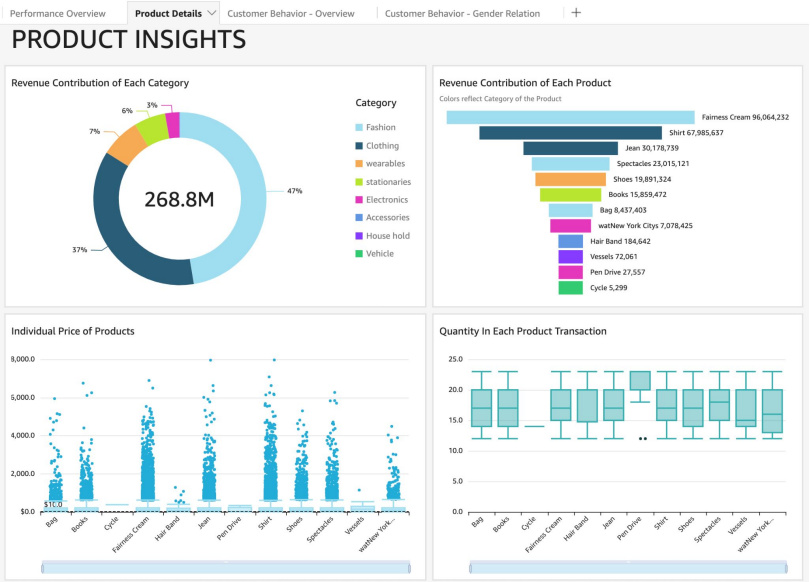


2. DASHBOARD OVERVIEW

BUILDING DASHBOARD WITH AWS QUICKSIGHT

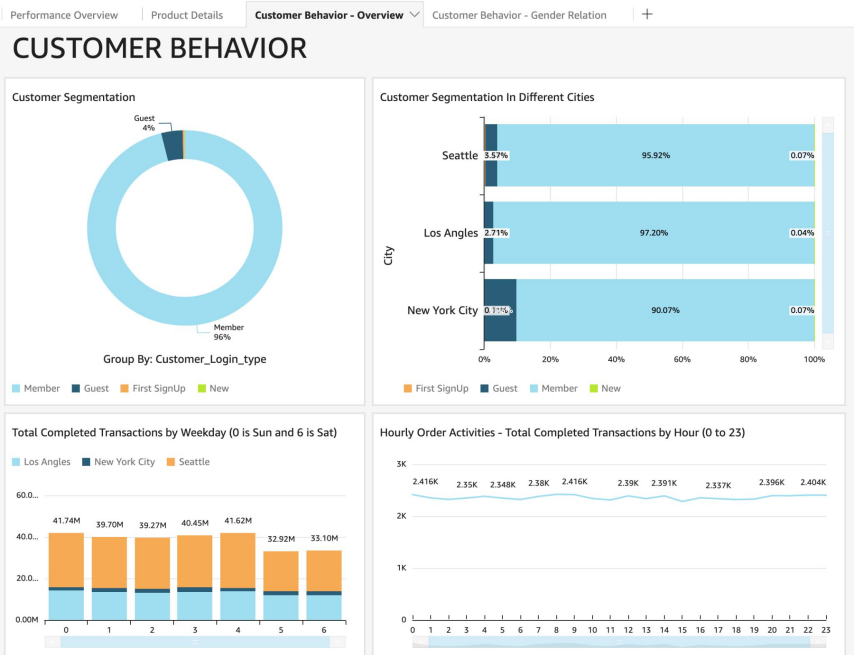
US-COMMERCE DASHBOARD SCREENSHOTS

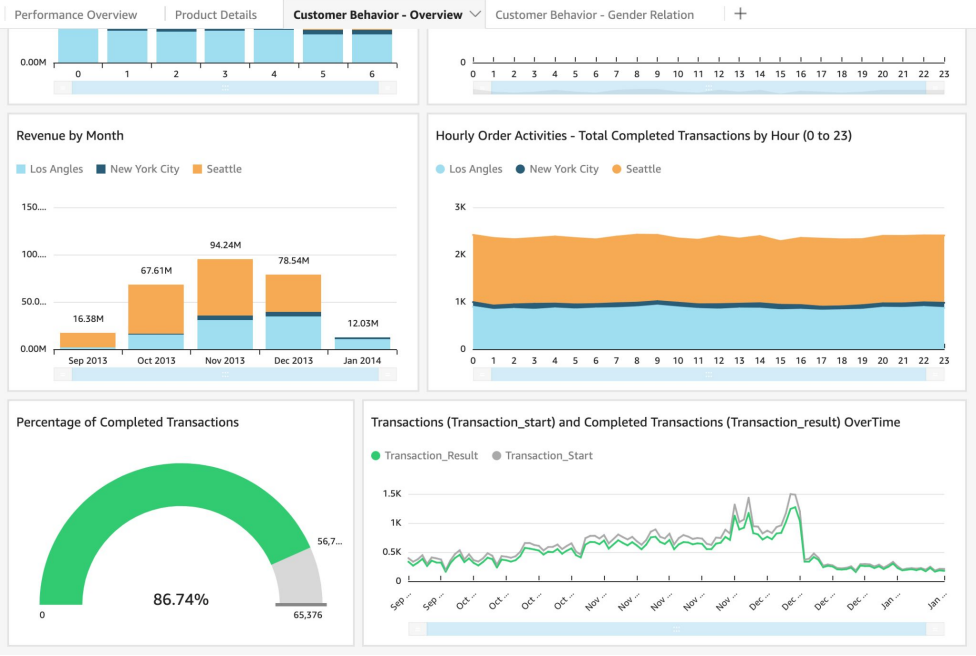




BUILDING DASHBOARD WITH AWS QUICKSIGHT

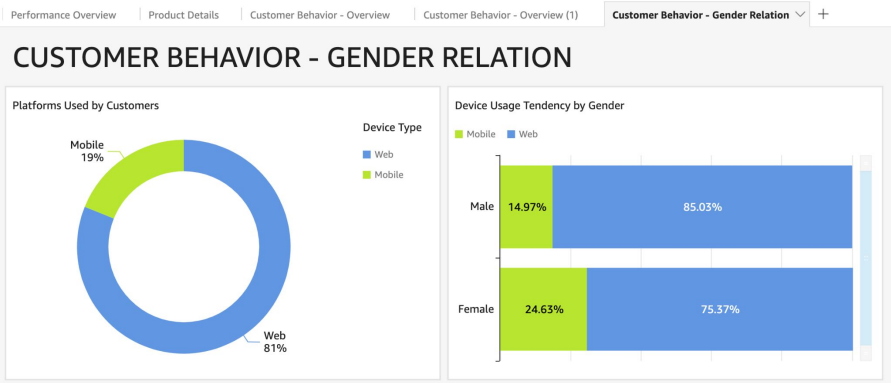
US-COMMERCE DASHBOARD SCREENSHOTS (continued)

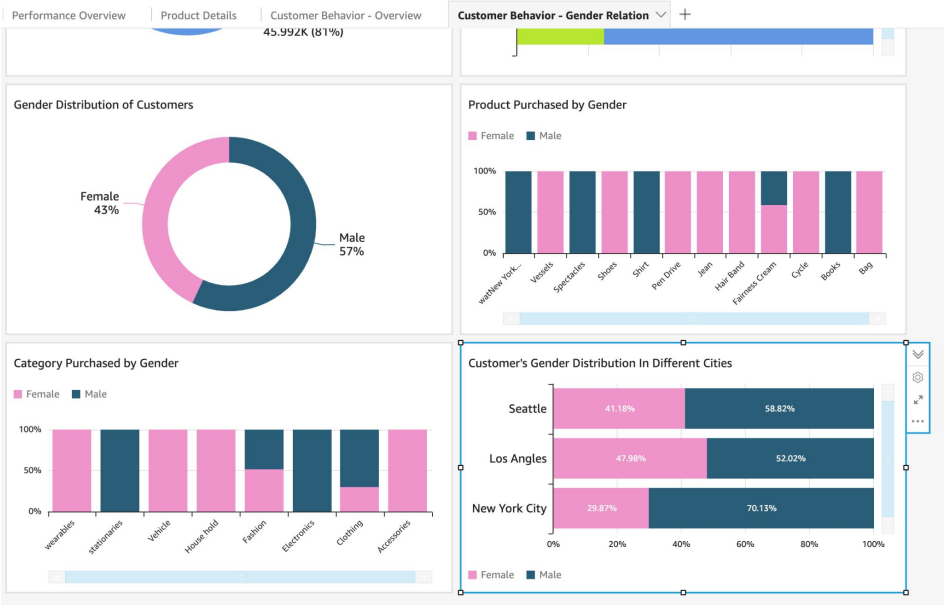




BUILDING DASHBOARD WITH AWS QUICKSIGHT

US-COMMERCE DASHBOARD SCREENSHOTS (continued)





3. US E-COMMERCE
SOME IMPORTANT INSIGHTS

PART 3 WOULD BE PRESENTED AS THE FOLLOWING FORMAT

1. Describe the data/graphs/charts and some insights.

2. Using green text box for discussing about future trends and improvements for the companies and/or products

US-COMMERCE PERFORMANCE OVERVIEW

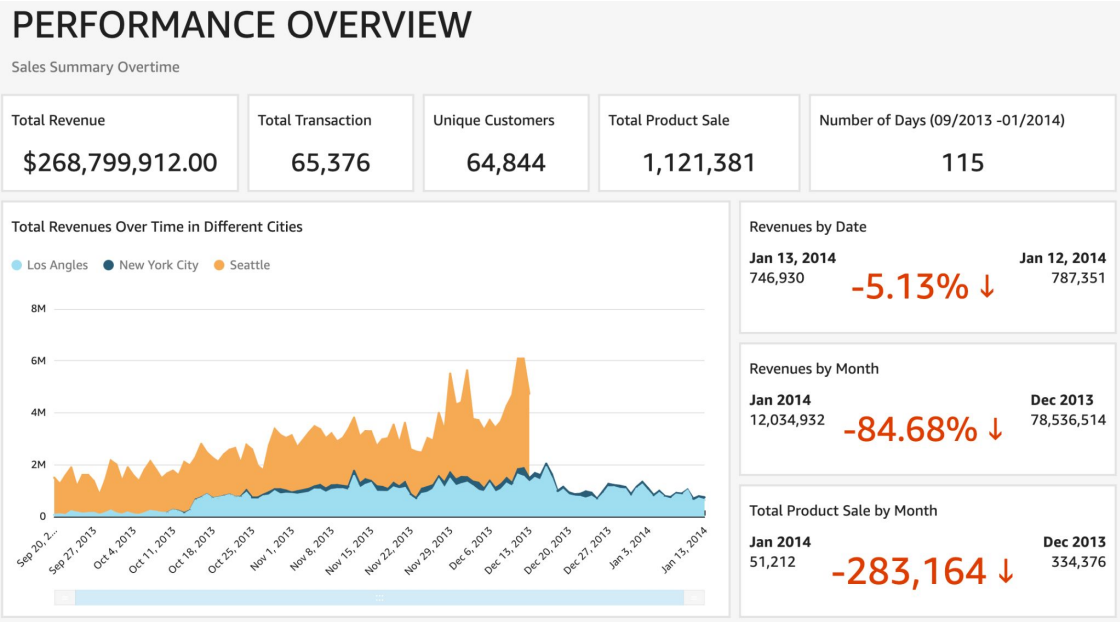
- The data ranges in 115 days (5 months): Sep 13, 2012 to Jan 14, 2014.

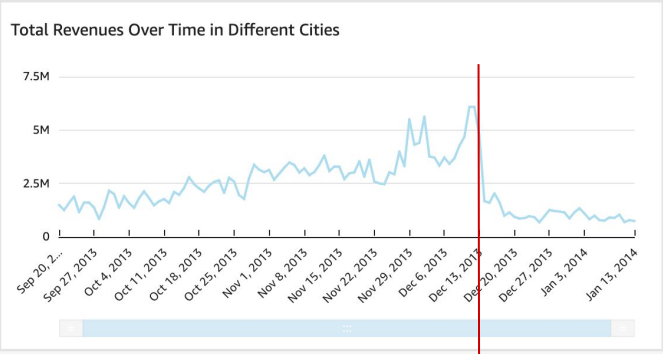
- Total revenue from all the orders was around $268.8M.

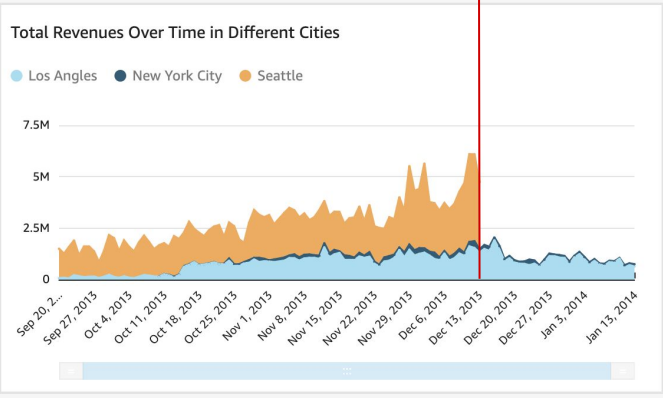
- Customers/Orders were from 3 cities Los Angeles (California), New York City (New York), and Seattle (Washington).

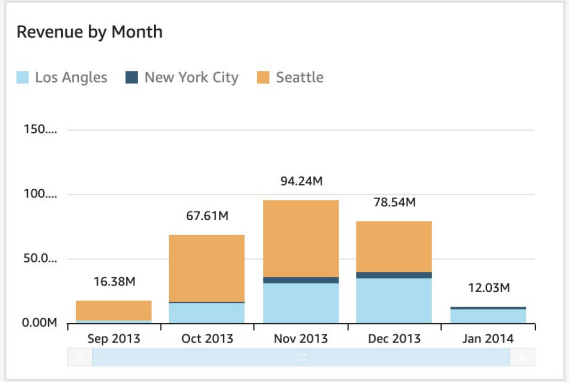
- The majority of revenue was from Seattle. But there was no order activities in Seattle since Dec 13, 2013.

- The revenue was significant decreased overtime.









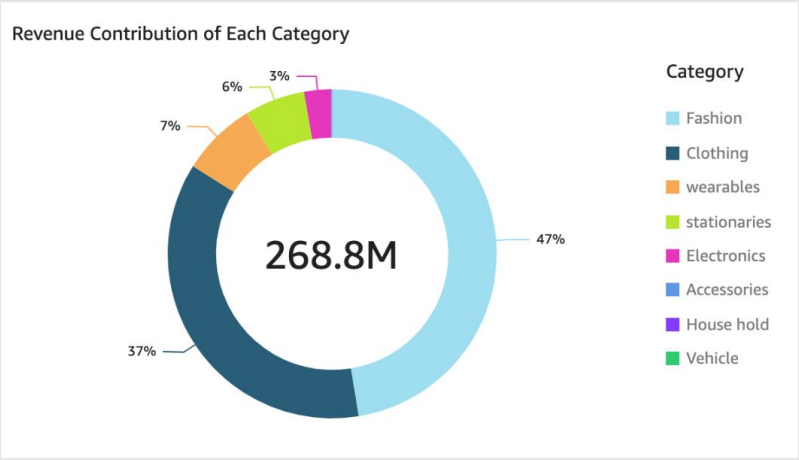
When looking at the **Revenue by Month,** except Seattle, the revenue/performance looks stable in New York city and LA. The sharp decrease in revenue could be related to the problem in Seattle.

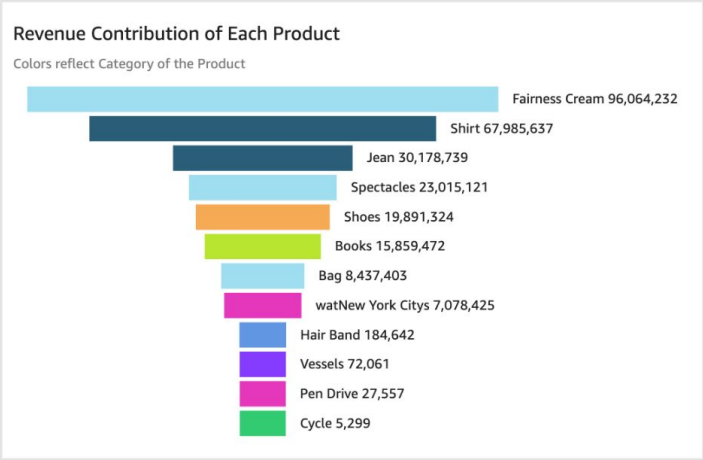
QUESTIONS/DISCUSSION NEEDED FOR PERFORMANCE IMPROVEMENT

1. What happened to the e-commerce activity in Seattle after Dec 13, 2013?
2. Why there was a great revenue decrease? Is it related to season? (Normally Nov- Dec is the high season comparing to other months of the year).

⇨ In such a short period of time (5 months), could not make any assumption on the revenue decrease. Need data on other periods to investigate more.

PRODUCT INSIGHTS





CATEGORY

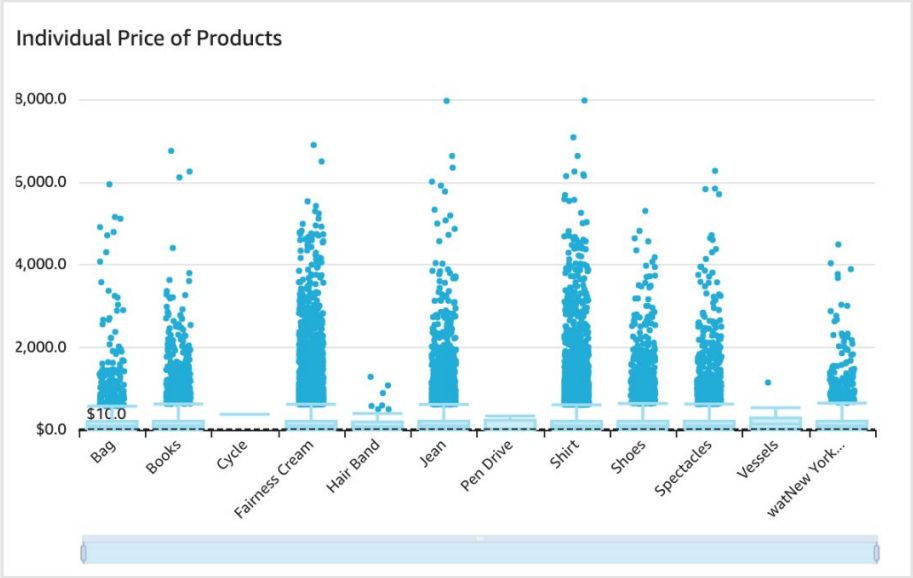
- 8 different product categories.

- **The top categories** that contributed to revenue is Fashion, Clothing, Wearables, and Stationaries.

PRODUCT

- 12 specifics products.

- **The top products** were **Fairness Cream (Fashion), Shirt, Jeans (Clothing), Spectacles (Fashion), Shoes (Wearables)**. Notes that in this product chart, the color reflects its own category.



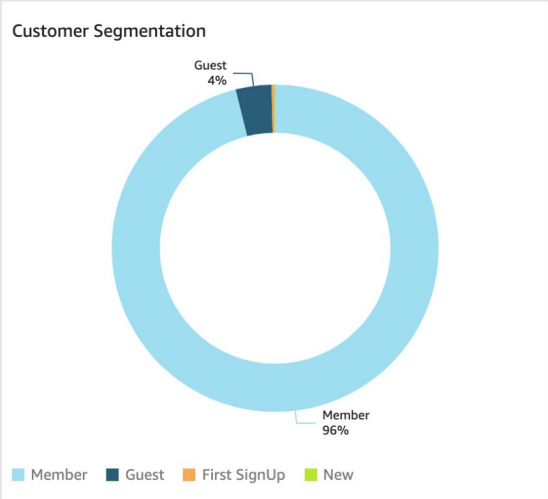
- There are the huge differences in **“Individual Price of Product” with a lot of outliers**

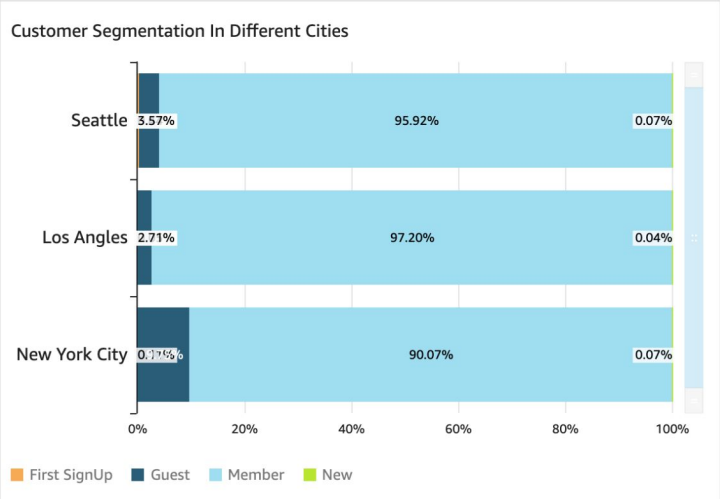
- The product quantity is not so “different”.

⇨ Need to perform Product Segmentation (normal, luxury, etc.) to see the performance of different product segments

⇨ Need to have further inputs from the company on how to do Product Segmentation

CUSTOMER SEGMENTATION





- Revenue came mostly from the “Member” customer, followed by Guest order.

- In **New York**, the behavior is slightly different: **more orders came from guest account** comparing to other cities.

⇨ More programs to “take care” of the “Member” customers.

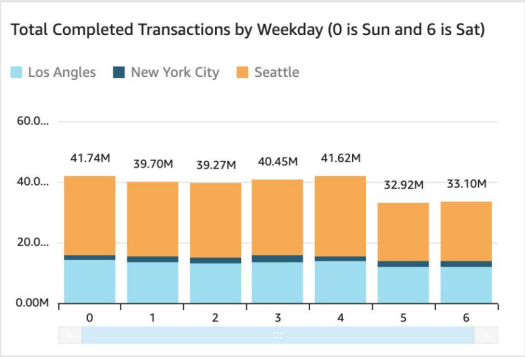
⇨ Program to convert other customer types to “Members”.

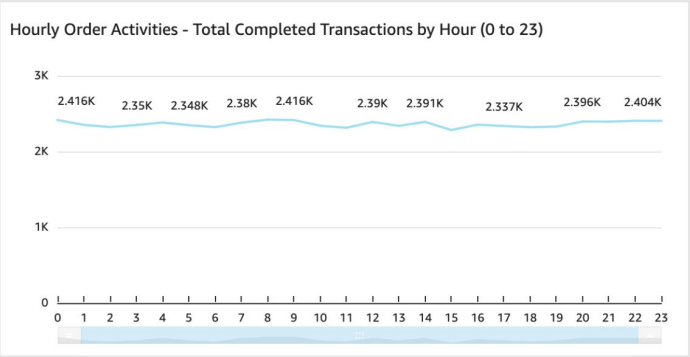
⇨ The approach for launching the loyalty program could be different in New York.

ORDER ACTIVITY – ORDERING PATTERN

- The order activities was slightly high on Sunday and Thursday, and low on Friday and Saturday.

⇨ More promotion and advertisement on Sunday and Thursday

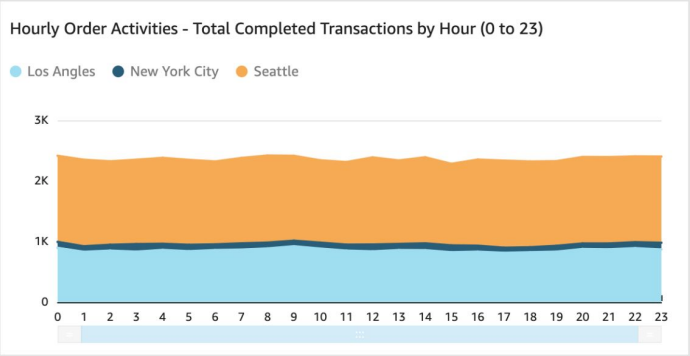




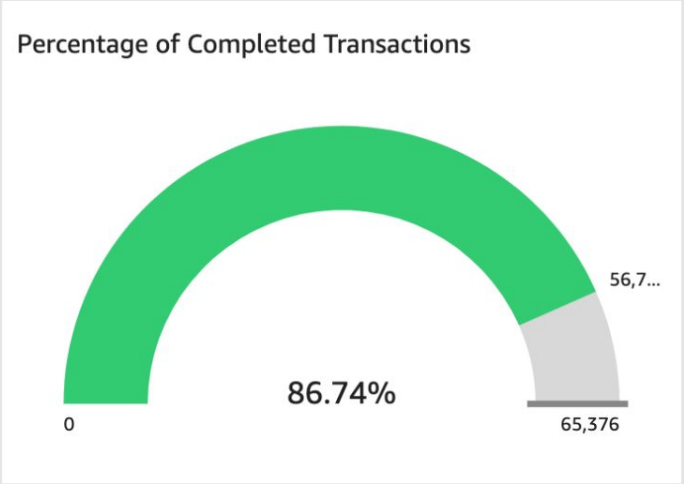
- Surprisingly, there was **no downtime in 24 hours of a day!!!**

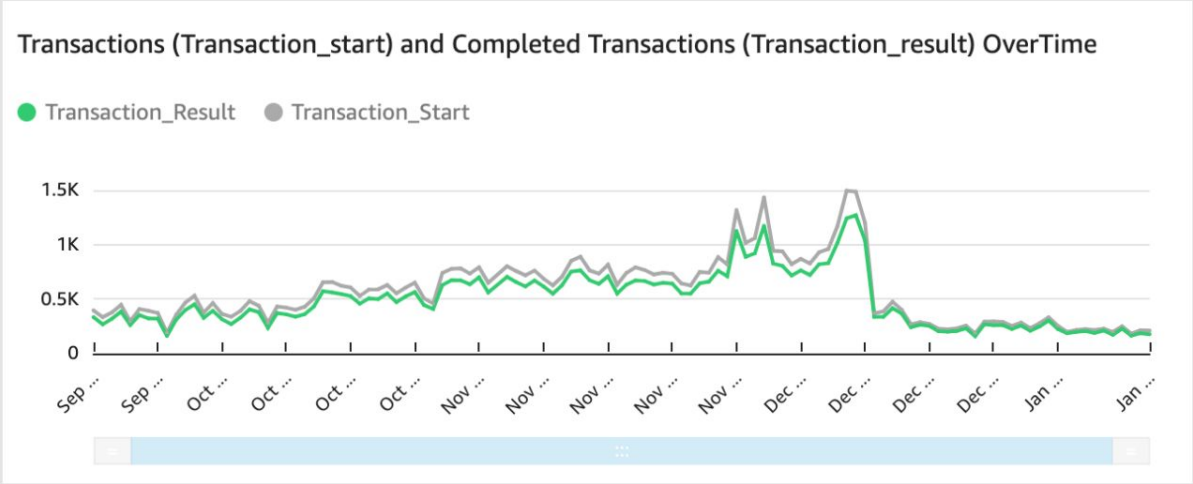
Everyone in the 3 big cities was just wide awake and did shopping day and night!!

⇨ Need to investigate more on the Timestamp data to see if there is any problem with data input/ETL process.



TOTAL TRANSACTION VERSUS COMPLETED TRANSACTION





- The percentage of completed transaction **(conversion rate) is 86.74%**, which is pretty good.

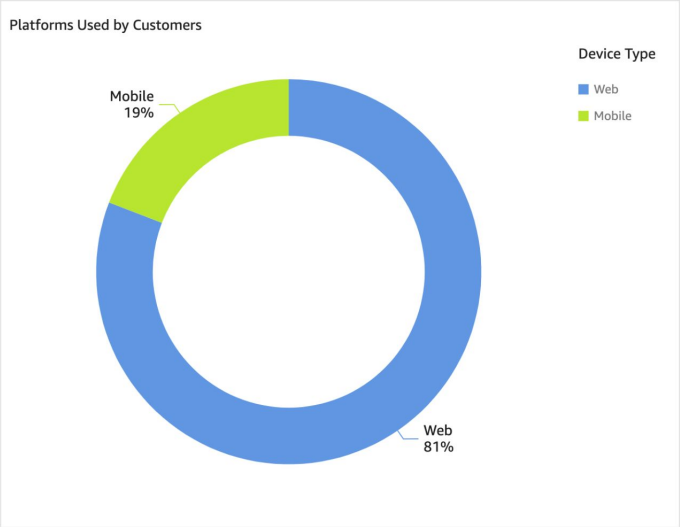
- When looking as the total transaction and completed transaction overtime, <u>the conversion rate looks stable</u> across 5 months.

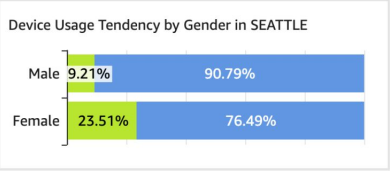
⇨ Need to improve it? Then it needs more input from the company.

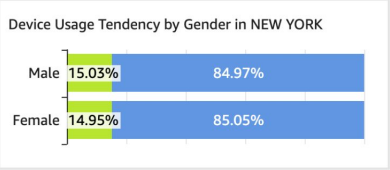
PLATFORMS USED BY CUSTOMERS

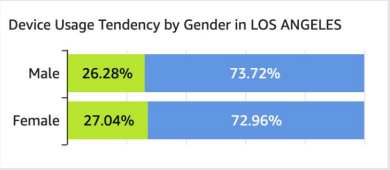
- 81% of the transaction was performed by **Web**.

- In New York and Los Angeles, the device usage tendency was equal in male and female. But it is not the same for **Seattle. More <u>female</u> used mobile for shopping.**







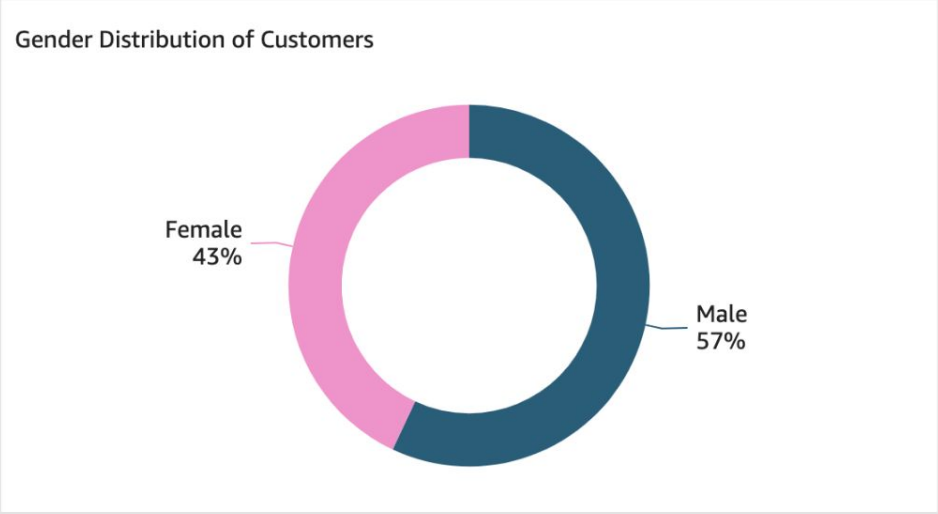


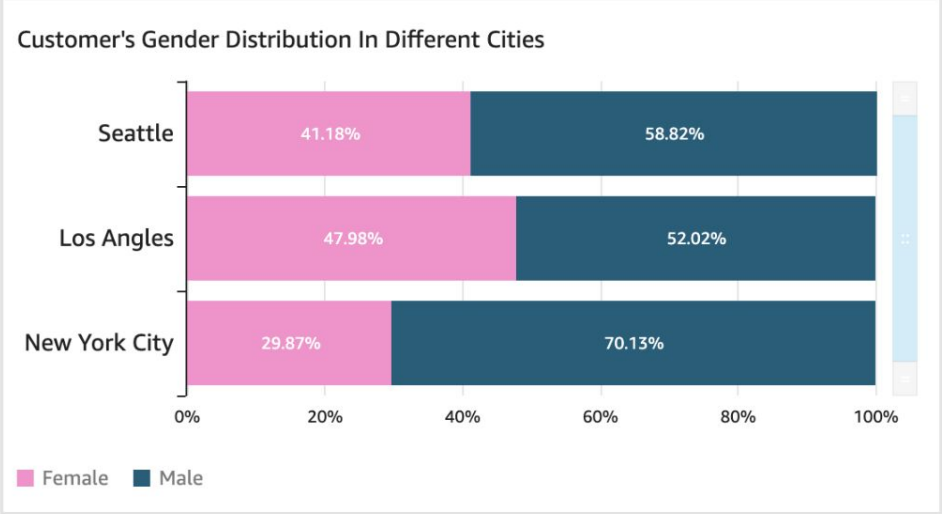
⇨ Need to discuss suitable actions for promotion/advertisements/ maintenance on Web.

⇨ Improve experience for Mobile app.

⇨ The approach could be different in Seattle since more female used mobile app comparing to other cities.

CUSTOMER BEHAVIOR BY GENDER

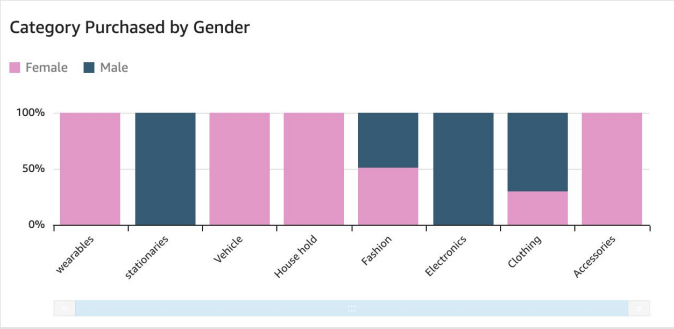


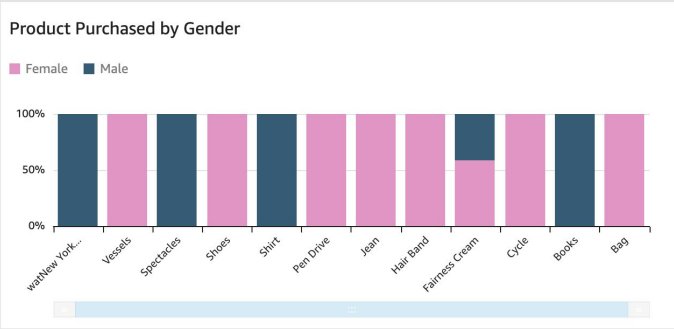


- Majority of customer (>50%) was Male. Significantly, in **New York, 70.13% of customer was Male.**

⇨ Need to discuss suitable actions for promotion/advertisements/ maintenance target on Gender.

CATEGORY/PRODUCT TENDENCY BY GENDER





Product Purchased by Gender

	Gender	
	Male	Female
Product	Count	Count
Bag		2,185
Books	3,810	
Cycle		1
Fairness Cream	9,525	13,939
Hair Band		48
Jean		7,216
Pen Drive		11
Shirt	16,527	
Shoes		4,739
Spectacles	5,593	
Vessels		20
watNew York ...	1,762	

- Interestingly, Female and Male Customer shared the same interest/need in buying **Fairness Cream**.

- Except that, all other products were purchased mostly either by Male or Female => Really **gender-specific product!**

⇨ Need to discuss suitable actions for promotion/advertisements/ maintenance target on Gender:

- Keep the product gender-specific like that, or change?

- The way to expand the product variety: what kinds of product to be expand?

4. CONCLUSIONS

CONCLUSION ON US E-COMMERCE DATASET

- Data science project is a repetitive cycle, which needs lots of inputs/feedback from the company for further actions (data analysis, finding more insights for improving and forecasting purposes).

⇨ Besides all the discussions/suggestions on other slides, E-commerce dataset needs more data and input for further analysis.

CONCLUSION ON AWS VIETNAM – DEVAX ONLINE WORKSHOP

(Oct 2021)

OVERALL

- Great contents, organization and participation! Great job! Thank you!

MY EXPERIENCE WITH AWS QUICKSIGHT

Since I could only join the 3rd and 4th session of this workshop, I could give my opinion on AWS Quicksight only.

- QuickSight is a potential platform for data visualization.

- I realized that I need more customized graphs, don't know whether QuickSight supports it, such as:

  + Combo charts: Stack more charts in the same figure without available fields.
  + Customize the length, width of the bars (on bar charts)
  + Adding a few label values (not all the values) on the existing charts.
  + Drill-down/Filtering features are not so user friendly (visualization result).
  + I still could not figure out how to modify/prepare, clean data on QuickSight. I still need to prepare/modify the data using other tools.

- Impression on community support: comparing to Tableau and other BI tools, I have the feeling that the available resources for QuickSight is not as much as other tools. When searching for some troubleshootings, there were few answers available. Many online questions/issues were left unanswered.

Thank you!