Homework Report

Basic Toolset

To complete this assignment, I am using the following tools

- Databricks on Azure
- Spark for data ingestion and cleaning
- Azure Volume for temporary file landing
- Google Looker Studio for generating the dashboard.

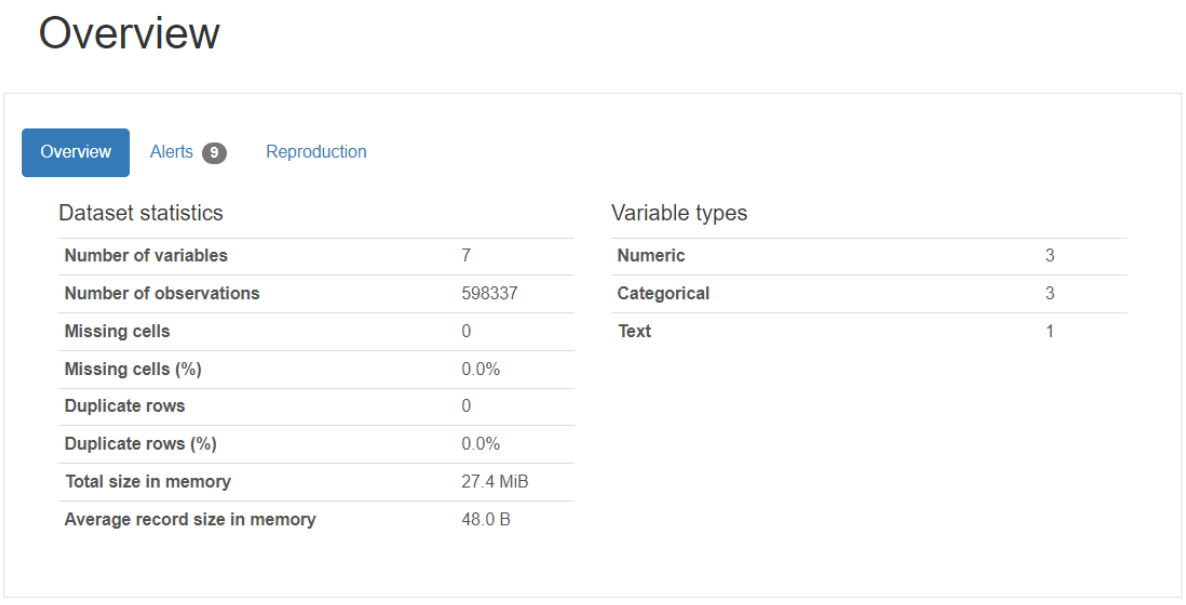
Step 1: Data Gathering & Ingestion

Due to a shortage of time, I am not following the standard practice in which we load the raw data into Bronze storage, after cleaning and transformation it is loaded in to the Silver storage and finally all sources are combined at the final stage of processing in the Gold storage which is then used for reporting and creating AI models.

Link to the Notebook - https://adb-7217844402021902.2.azuredatabricks.net/?o=7217844402021902#notebook/3211540213036274

To clean and load the data sets, I have performed the following steps

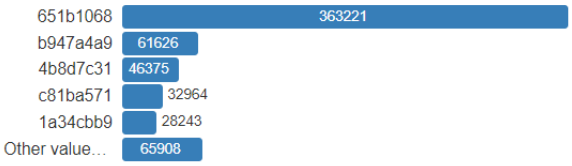
1. Loaded data files into Azure Volume, which I am using as a temporary landing space for this assignment. It can be automated using bash/batch script depending on the platform.
2. From files loaded the data sets into spark data frames.
3. As this is the first ingest so I have created a detailed profiling report to get an idea of how the data looks like and also to perform a brief sanity check. Attached below are the snippets of the profiling report.



department

Categorical

Distinct	21
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	4.6 MiB



More details

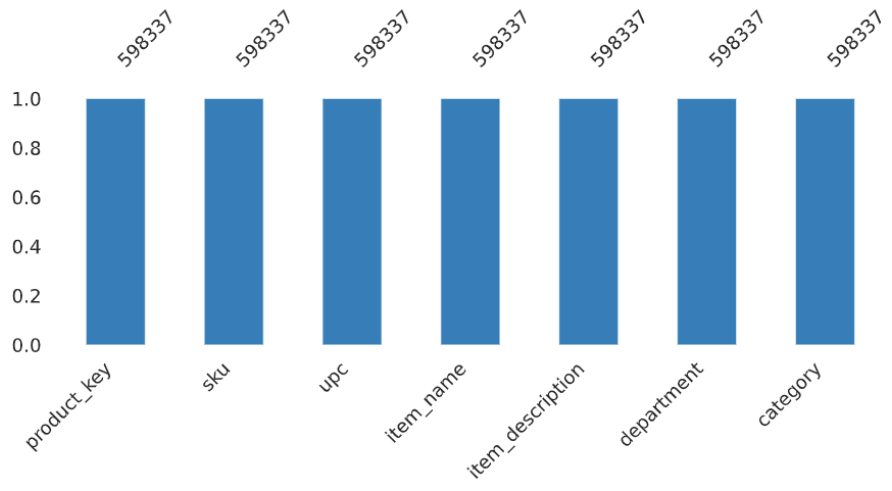
category

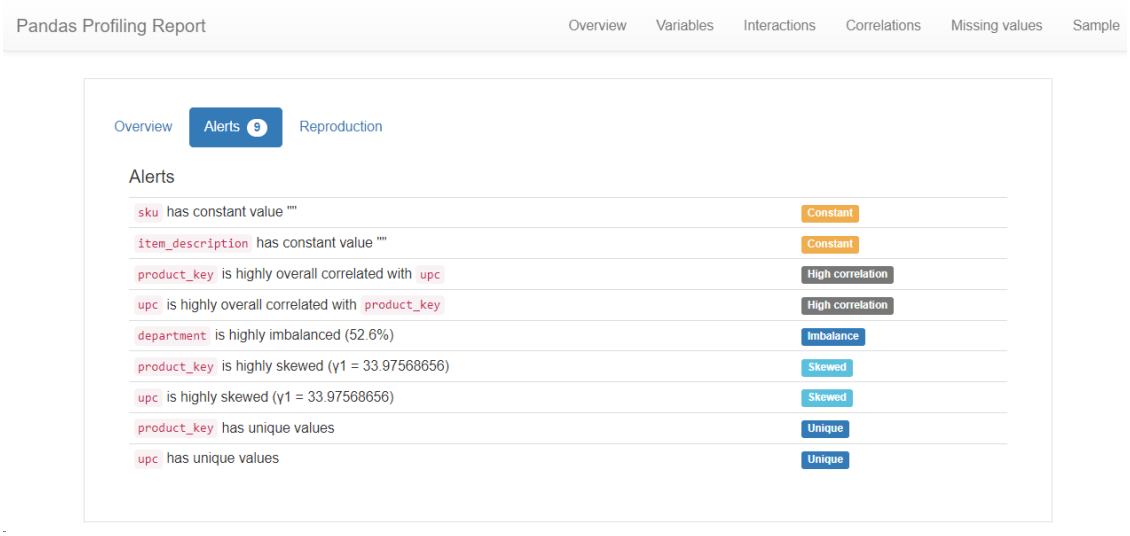
Text

Distinct	123
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	4.6 MiB

Missing values

Count | Matrix





4. Data set for **trans_fct** was a huge dataset and generating a report on that was very time and resource consuming so I have picked a random sample of 1% rows of the data set and generated a profiling report on the sample. You can find the sample report attached.

Step 2: Prepare & Cleanse the data in memory

For data preparation, I have loaded all 10 `trans_fct` files into spark data frame one by one because a few of them had different format or column sequence and loading those together in a frame meant column misplacement and data pollution. Also, replaced nulls in sales and units with 0 values.

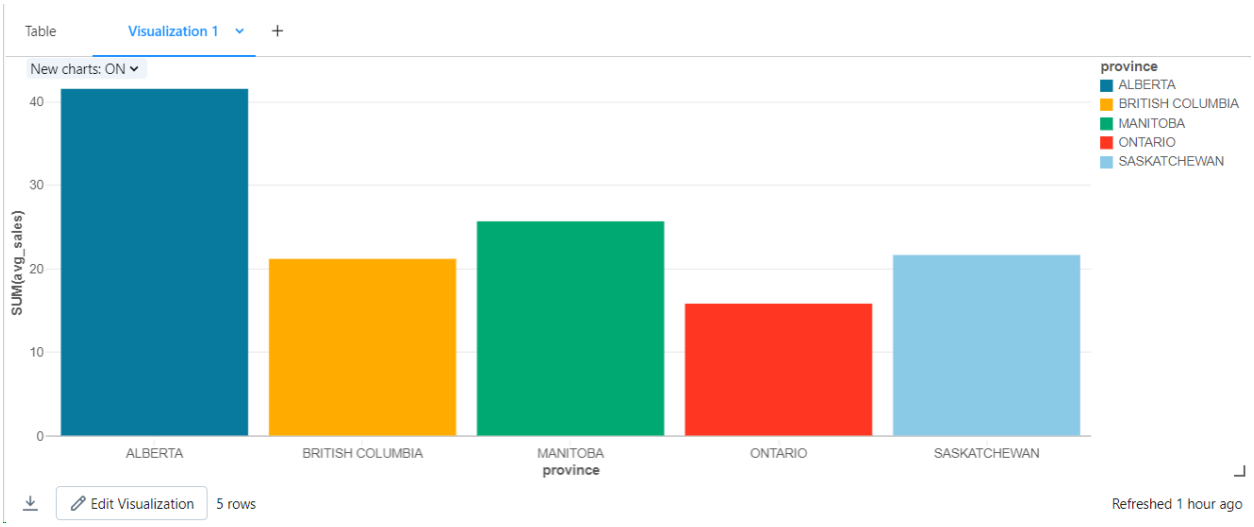
Some files also had the `trans_key` column coming in as Numeric with an exponent. It wasn't possible to recover the correct value for such transactions so either these files could be discarded entirely or they can be used as they were. I went with the latter option as the important values such as sales and units were still intact and invaluable. Also, because it seemed like an extraction error rather than erroneous data so made sense to still make use of it.

Step 3: Gather insights from data

Average Sale per Province

Table +

	province	avg_sales
1	ALBERTA	41.50833914212801
2	MANITOBA	25.64660321627838
3	SASKATCHEWAN	21.616420387198737
4	BRITISH COLUMBIA	21.152805926608313
5	ONTARIO	15.792577054113353



Average Sale per Store

Table +

	store_num	avg_sales
1	9807	157.11371912168363
2	9802	110.22534516765286
3	7125	63.80756264236902
4	7262	29.981304347826093
5	8185	26.45837209302325
6	7167	25.765927601809988
7	4823	25.703245883644424

Top Stores in each Province

Table +

	province	store_num	avg_sales
1	ALBERTA	9807	157.11371912168363
2	ALBERTA	9802	110.22534516765286
3	BRITISH COLUMBIA	7125	63.80756264236902
4	ALBERTA	7262	29.981304347826093
5	ONTARIO	8185	26.45837209302325
6	BRITISH COLUMBIA	7167	25.765927601809988
7	MANITOBA	4823	25.703245883644424

47 rows | 10.36 seconds runtime

Top Store performance Vs Median Store performance per province

Table | Visualization 1 | Visualization 2 | +

	province	store_num	avg_sales	median_province_sales	performance_vs_median
1	ALBERTA	9807	157.11371912168363	16.105377358490568	9.755357830150754
2	ALBERTA	9802	110.22534516765286	16.105377358490568	6.844008849599748
3	BRITISH COLUMBIA	7125	63.80756264236902	20.106973525872444	3.173404618067721
4	ALBERTA	7262	29.981304347826093	16.105377358490568	1.861571056701773
5	ONTARIO	8185	26.45837209302325	14.936387255785007	1.7714037296921092
6	BRITISH COLUMBIA	7167	25.765927601809988	20.106973525872444	1.2814423597194249
7	MANITOBA	4823	25.703245883644424	9.544074074074073	2.6931104771562713
8	ONTARIO	8187	25.54258426966293	14.936387255785007	1.7100911908781717
9	SASKATCHEWAN	7317	22.664568081991213	19.094680851063817	1.1869571562244
10	ONTARIO	8161	21.385181518151814	14.936387255785007	1.4317506068858201

47 rows | 3.82 seconds runtime

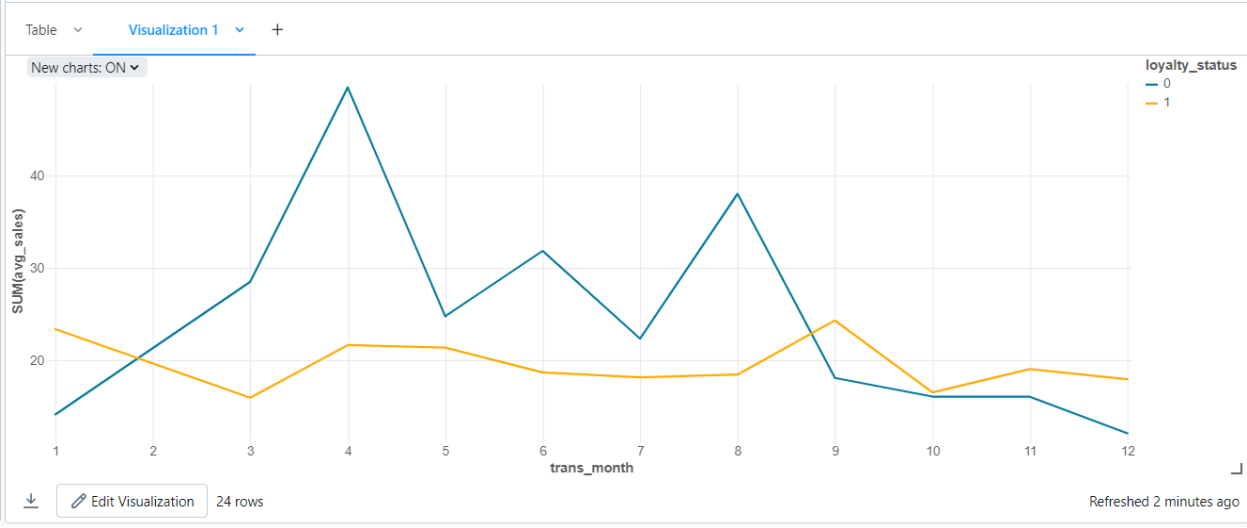
Top Store performance Vs Average Store performance per province

Table | Visualization 1 | Visualization 2 | +

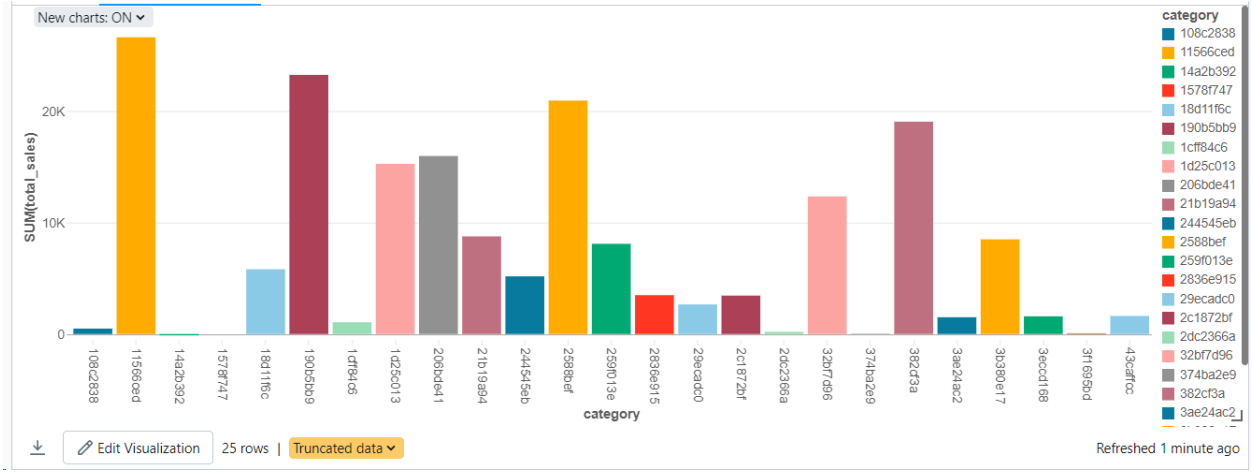
	province	store_num	avg_sales	avg_province_sales	performance_vs_avg
1	ALBERTA	9807	157.11371912168363	35.10966630010543	4.474941965518249
2	ALBERTA	9802	110.22534516765286	35.10966630010543	3.1394586386975107
3	ALBERTA	7262	29.981304347826093	35.10966630010543	0.8539330477127338
4	ALBERTA	7238	19.203431372549023	35.10966630010543	0.546955679054442
5	ALBERTA	7247	19.06821390374333	35.10966630010543	0.5431043901344648
6	ALBERTA	7240	17.167456140350875	35.10966630010543	0.4889666564646137
7	ALBERTA	7261	16.105377358490568	35.10966630010543	0.4587163324432453

47 rows | 4.62 seconds runtime

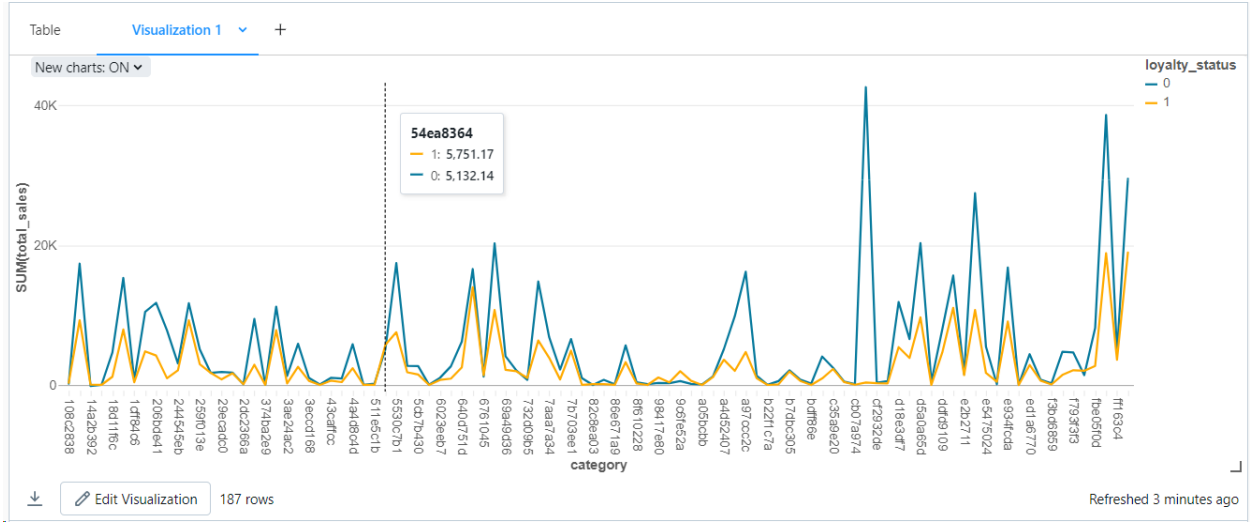
Loyalty vs Non-loyalty Sales Per Month



Sale per Category



Sale per category grouped by loyalty Status



Top 10 Categories per department

Table +

	department	top_categories
1	a461091	["687ed9e3"]
2	34a2a7e0	["1578f747"]
3	5bffa719	["a4d52407", "6c504249", "fa3a1bd8", "a121fb78", "3ae24ac2", "999b3a55", "8f610228", "108c2838", "2dc2366a", "3f1695bd"]
4	7569cb40	["cef3760b"]
5	b947a4a9	["382cf3a", "8b4f9982", "3b380e17", "29ecadc0", "6761045", "6023eeb7", "f2672c8c", "9db5a1ff"]
6	24d07cc8	["50c418ce"]
7	435ca98	["e0b38f5b", "21b19a94", "640d751d", "2836e915", "e8eeb80f", "a05bcbb"]
8	1a34cbb9	["fe148072", "ffcec4a7", "e49d14f1", "11566ced", "e934fcda", "7703921f", "a97ccc2c", "d18e3df7", "a8a688f9", "7b703ee1",
9	89d0c9d1	["511e5c1b"]

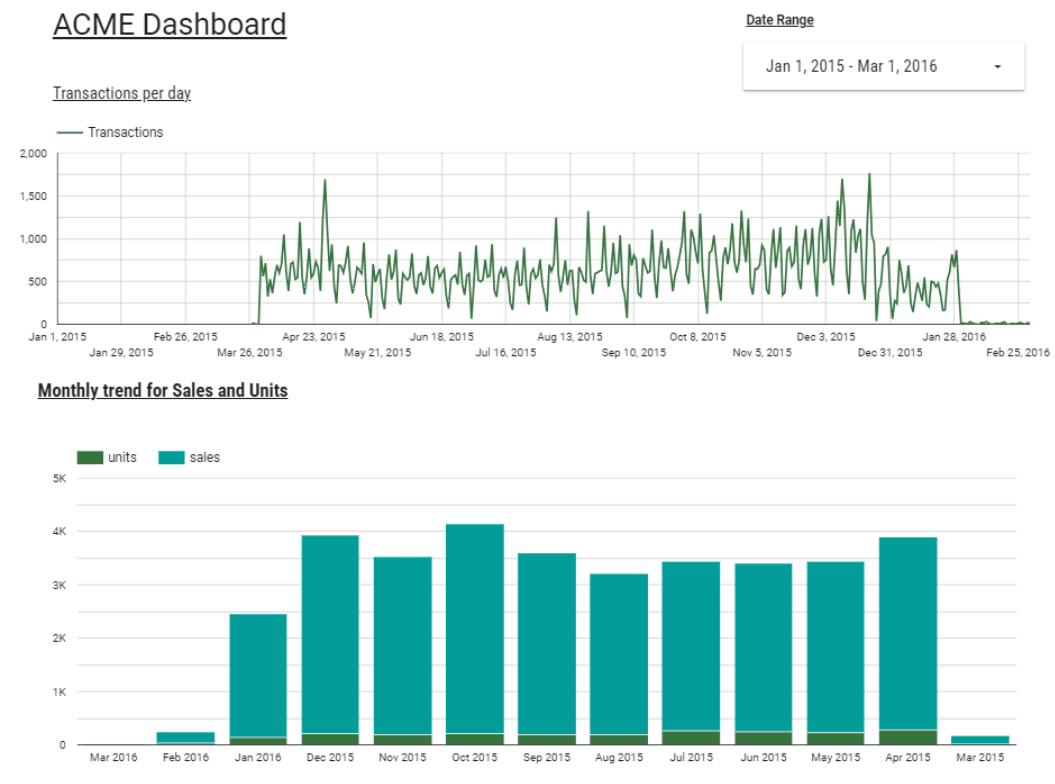
Top 5 stores per province

Table +

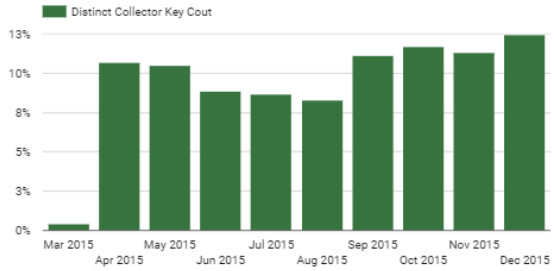
	store_num	province	sum(sales)	rank
1	9807	ALBERTA	1030351.7700000013	1
2	7296	ALBERTA	385435.04999999976	2
3	9802	ALBERTA	55884.25	3
4	7247	ALBERTA	17828.780000000013	4
5	7226	ALBERTA	9994.969999999994	5
6	7167	BRITISH COLUMBIA	113885.40000000015	1
7	7104	BRITISH COLUMBIA	101377.43000000011	2
8	7125	BRITISH COLUMBIA	28011.52	3
9	7194	BRITISH COLUMBIA	17304.700000000004	4
10	7175	BRITISH COLUMBIA	14096.529999999997	5
11	4823	MANITOBA	234156.5700000007	1
12	4861	MANITOBA	257.69	2
13	7403	MANITOBA	21.34	3

Step 4: Display your analytics (bonus points)

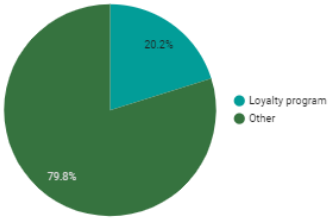
To display the analytics, I stored the data in SQL Datawarehouse and was trying to connect it with the Google Looker Studio directly but for some reason the Google connector wasn't working for Databricks. Then, I extracted the resultant dataset in the form of csv files and loaded those files in the looker Studio to generate a basic dashboard. Attaching the dashboard link in the email as well as it in the form of PDF at Github.



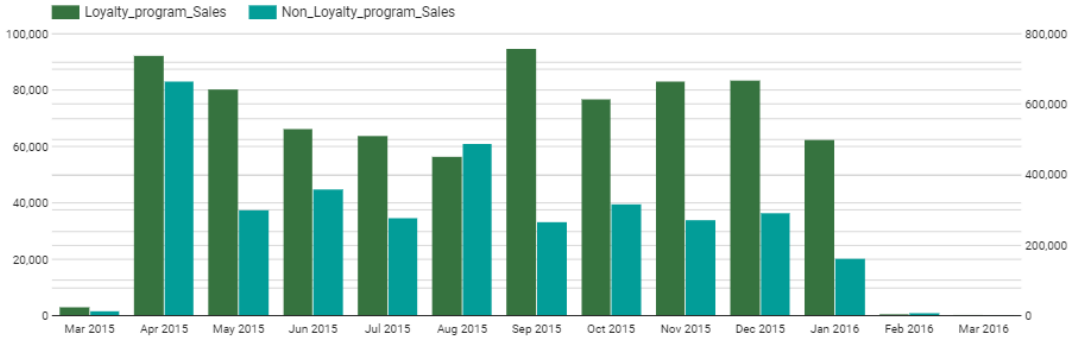
ACME Dashboard



Overall Transactions Breakdown



Monthly trend for Sales and Units



Please feel free to reach out to me if you have any questions. My email is summiyakhalid@gmail.com