Homework Report

Basic Toolset

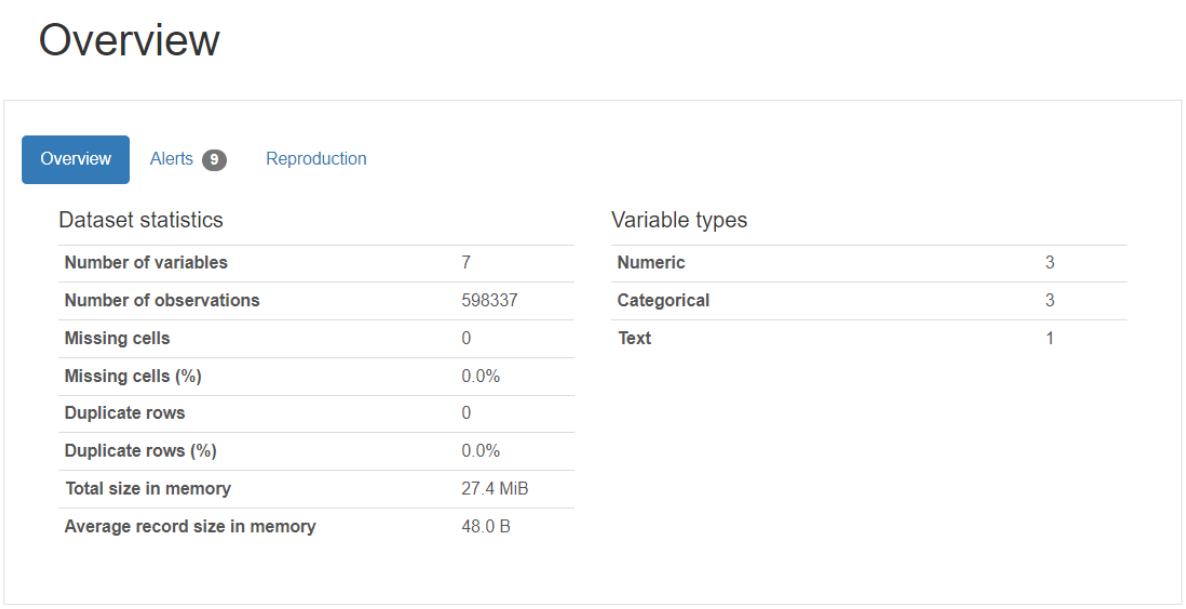
To complete this assignment, I am using the following tools

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

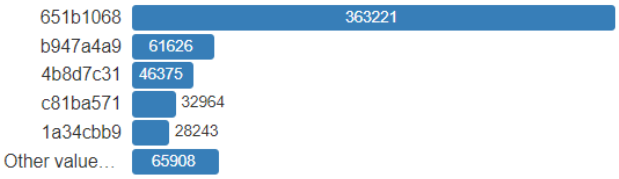
1. Loaded data files into Azure Volume, which I am using as a temporary landing space for this assignment. It can be automated using bash/batch script depending on the platform.
2. From files loaded the data sets into spark data frames.
3. As this is the first ingest so I have created a detailed profiling 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

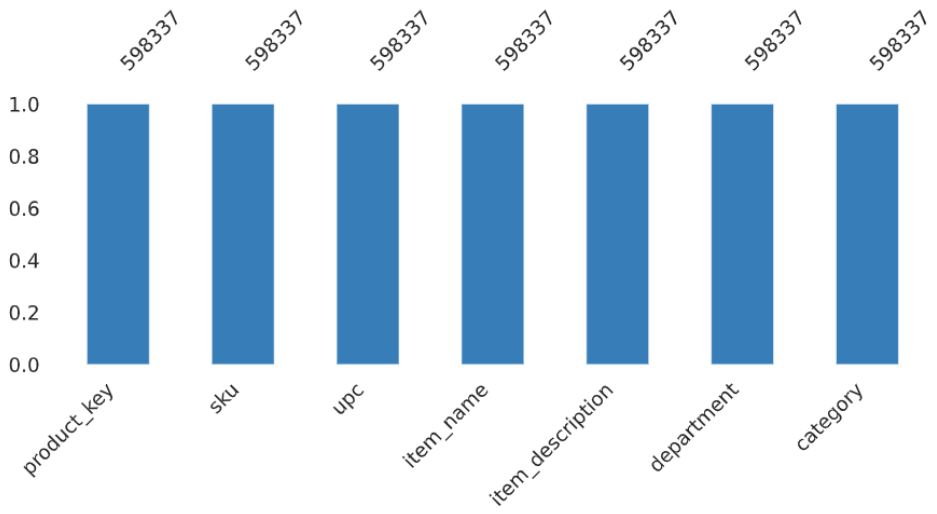
category

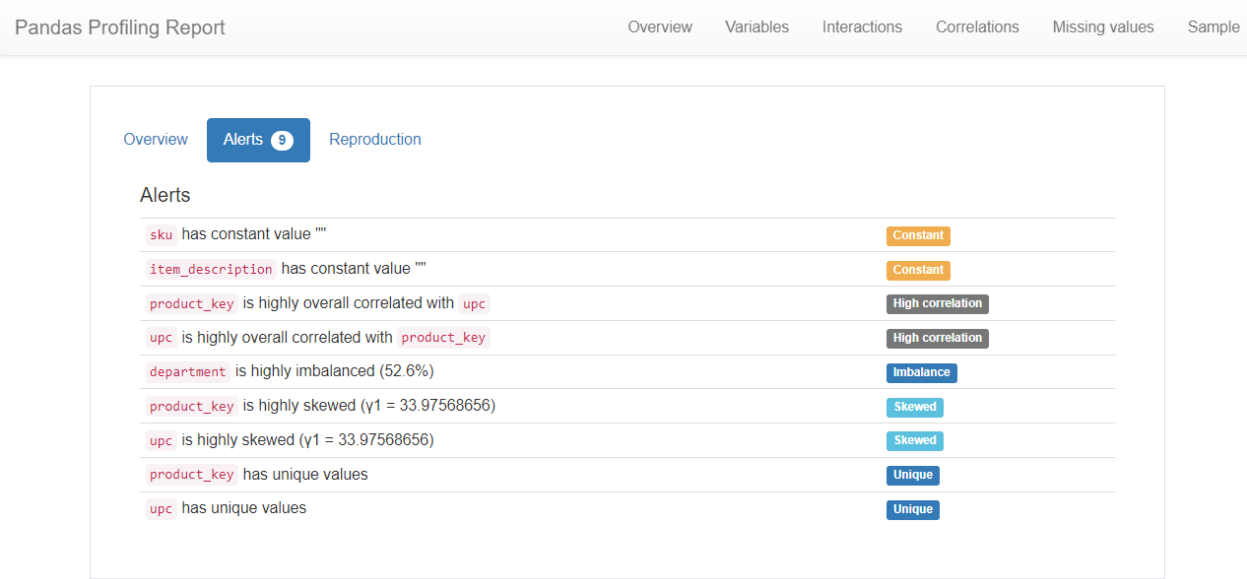
Text

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

You can find the insights in the Notebook.

Step 4: Display your analytics (bonus points)

To display the analytics, I was trying to store the data in SQL datalake and connect it with Google Looker Studio but for some reason the 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.

Please feel free to reach out to me for further questions at summiyakhalid@gmail.com