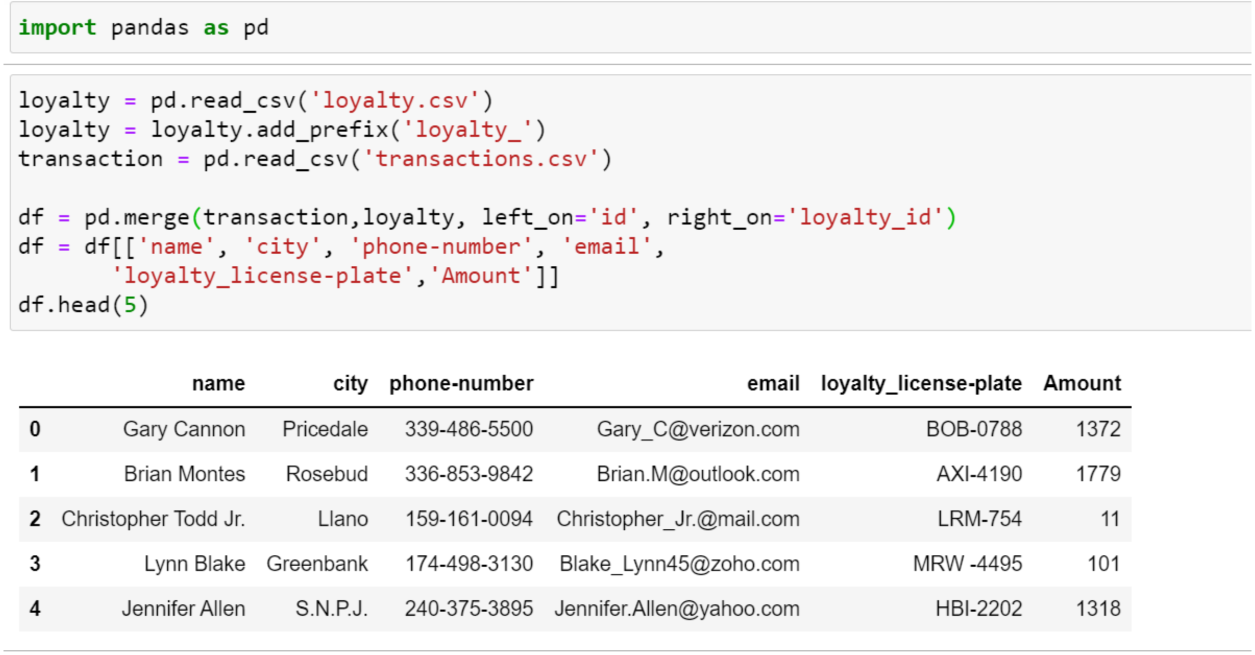
Take home test - Data science/ analytics/ engineering

Challenge 1: Data Cleaning, Transformations and ETL pipeline architecture

1) Loyalty dataset is a dimension table while transactions dataset is a fact table. Both datasets can be linked by a column 'id'. After linked and cleaned the data, the final dataset become as following:



```
import pandas as pd
```

```
loyalty = pd.read_csv('loyalty.csv')
loyalty = loyalty.add_prefix('loyalty_')
transaction = pd.read_csv('transactions.csv')

df = pd.merge(transaction,loyalty, left_on='id', right_on='loyalty_id')
df = df[['name', 'city', 'phone-number', 'email',
       'loyalty_license-plate','Amount']]
df.head(5)
```

	name	city	phone-number	email	loyalty_license-plate	Amount
0	Gary Cannon	Pricedale	339-486-5500	Gary_C@verizon.com	BOB-0788	1372
1	Brian Montes	Rosebud	336-853-9842	Brian.M@outlook.com	AXI-4190	1779
2	Christopher Todd Jr.	Llano	159-161-0094	Christopher_Jr.@mail.com	LRM-754	11
3	Lynn Blake	Greenbank	174-498-3130	Blake_Lynn45@zoho.com	MRW -4495	101
4	Jennifer Allen	S.N.P.J.	240-375-3895	Jennifer.Allen@yahoo.com	HBI-2202	1318

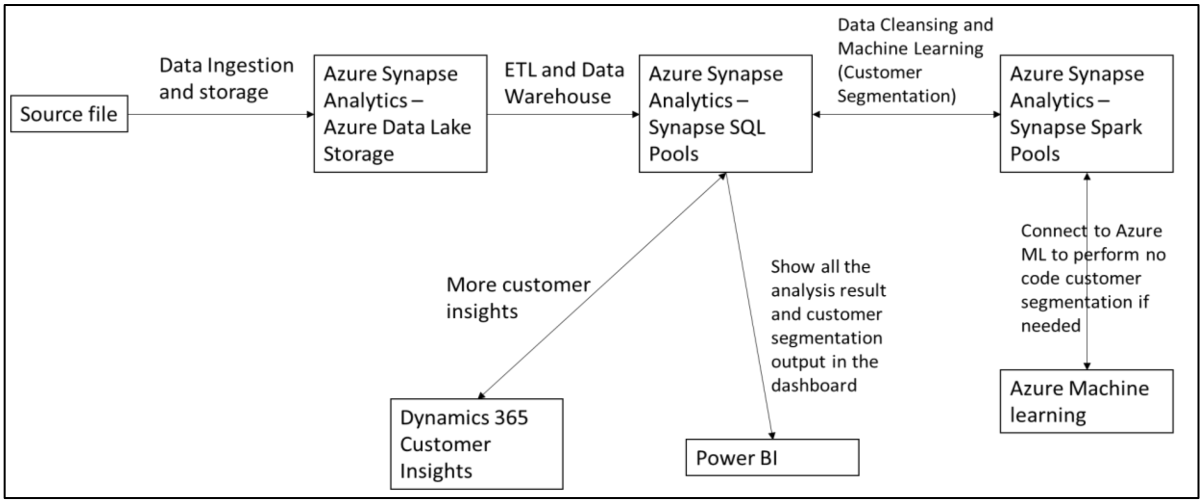
2)

a) 5 reasons for joining the datasets in this way

- The data in the loyalty table are very messy and we able to get more accurate details and amount data from transactions table by linking the column 'id'
- Easy to read and use as it is just referring to this final table
- Speed up the execution time and save the space for dropping some useless columns
- Able to perform the data analysis faster as we know the relationship between the variable
- To get more insights between the variable

b) Business use case – Assume the dataset is about the amount spending of the customer for a company. Based on the dataset, we can cluster the customer into different clusters based on their spending amount. After that, we can see the what customer under which cluster group and use different reward program or specific marketing strategy for different clusters based on the priority. The purpose of the reward program or marketing strategy is to improve the customer satisfaction, increase the sales and customer retention. Besides that, we also can analysis the spending power of each city and apply different marketing strategy to target different city.
(Please refer to the python file in Github for sample k-means clustering)

c) Microsoft Cloud pipeline architecture


Source file
Data Ingestion and storage
Azure Synapse Analytics – Azure Data Lake Storage
ETL and Data Warehouse
Azure Synapse Analytics – Synapse SQL Pools
Data Cleansing and Machine Learning (Customer Segmentation)
Azure Synapse Analytics – Synapse Spark Pools
Connect to Azure ML to perform no code customer segmentation if needed
Azure Machine learning
More customer insights
Show all the analysis result and customer segmentation output in the dashboard
Dynamics 365 Customer Insights
Power BI

Firstly, the data will be import and store in the Azure Data Lake of Azure Synapse. After that, the data will be performed ETL and store into data warehouse (SQL pools) as a table. Next, the Azure Synapse Spark Pools will be use to do the data transformation and data cleansing and develop machine learning which is customer segmentation. If needed, Azure Machine learning also can be connected to perform no code machine learning or auto machine learning and send back the result to the Azure Synapse. After that, all the results from data analysis and machine learning will be store into SQL Pools or Data Lake. Power BI dashboard will be built to show all the customer segment and analysis. Alternatively, Dynamics 365 Customer Insights can be used to perform more customer related analysis and find the insights as well as CRM.

The reason for using Microsoft platform is Microsoft having a product called Azure Synapse which one platform that can store the data, perform ETL, data analysis and machine learning development as well as connect with other Microsoft product easily. The Azure Data Lake Storage in the Azure Synapse is a data lake that able to store the data either structure and unstructured which it will be better for future if unstructured data to import. Azure Synapse also consisted of SQL data warehouse to store the data like sql server, write sql query and store procedure to run like a SSIS.

d) Analytics Validation

- We can evaluate the result of customer segmentation based on the spending amount or the sales from time to time after applying the marketing strategy / reward program for the cluster group by comparing the spending amount / sales with historical data.

Challenge 2: Customer engagement

- Present using dashboard
  - I will create a sample dashboard by connecting all the results and analysis from the customer segmentation. Dashboard is a good way to present all the outcomes with customer as it is more interactive and beautiful. In the dashboard, we can use different filters like cluster filters to see what customers are belong to which group and from which city as well as their spending amount. Other than that, it also will show what reward program / marketing strategy are good for this group and estimate how much spending amount / sales will be generated from it. Hence, the customer will be impressed by the beautiful and interactive dashboard when comparing to the static report.
- Present using web platform
  - I will create a web to integrate with the results and analysis from the customer segmentation. The concept is similar to the dashboard but it is in the web. It also can use the filters to interact with other graphics and map to see the analysis and details.
- Integrate with customer platform
  - If the customer having their own platform and customer would like to integrate the results to their platform. If this is the case, I will show how our system design and how to call the result from our end. Hence, the customer able to click the button inside their platform and get the output from us.