#Environment setup and making a connect between storage account and databricks cluster

storage\_account\_name = "rahulpysparkassignment"

storage\_account\_access\_key = "8FGtuM6vW6jrqpFiOnxjfcnwUBdRTP27MZCHCnLwpjmbTe+5aTUv2cvMj6w0ixtdWKdEC1n32b1g+ASt2zwLXg=="

container\_name = "pyspark"

dbutils.fs.mount(

  source = f"wasbs://{container\_name}@{storage\_account\_name}.blob.core.windows.net/",

  mount\_point = "/mnt/data",

  extra\_configs = {f"fs.azure.account.key.{storage\_account\_name}.blob.core.windows.net": storage\_account\_access\_key}

)

dbutils.fs.ls("/mnt/data")

#load parquet files with Pandas

import pandas as pd

transactions = pd.read.csv ('transactions.parquet')

#load csv files with Pandas

vips\_2020\_11\_01 = pd.read\_csv('vips\_2020-11-01.csv')

vips\_2020\_11\_15 = pd.read\_csv('vips\_2020-11-15.csv')

vips\_2020\_11\_25 = pd.read\_csv('vips\_2020-11-25.csv')

#load csv mapping files with Pandas

vip\_mapping = pd.read\_csv('umd\_vip\_to\_profile\_mapping.csv')