In [139…
```
import pandas as pd
```

In [140…
```
import numpy as np
```

In [141…
```
#To import the train.csv dataset from my device into jupyter
df = pd.read_csv(r'C:\Users\hp\Desktop\Tech 1M\python assignment\train.csv', low_memory = False)
```

In [142…
```
#assigning the dataframe to be df_train to avoid confusion
df_train = df
```

In [143…
```
#To view the top 10 samples of the dataset
df_train.head(10)
```

Out[143]:

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
0	1	5	2015-07-31	5263	555	1	1	0	1
1	2	5	2015-07-31	6064	625	1	1	0	1
2	3	5	2015-07-31	8314	821	1	1	0	1
3	4	5	2015-07-31	13995	1498	1	1	0	1
4	5	5	2015-07-31	4822	559	1	1	0	1
5	6	5	2015-07-31	5651	589	1	1	0	1
6	7	5	2015-07-31	15344	1414	1	1	0	1
7	8	5	2015-07-31	8492	833	1	1	0	1
8	9	5	2015-07-31	8565	687	1	1	0	1
9	10	5	2015-07-31	7185	681	1	1	0	1

In [144…
```
#To view the columns and index (rows) of the dataset
df_train.shape
```

Out[144]:
```
(1017209, 9)
```

In [145…
```
#To view the datatypes
df_train.dtypes
```

Out[145]:
```
Store             int64
DayOfWeek         int64
Date             object
Sales             int64
Customers         int64
Open              int64
Promo             int64
StateHoliday     object
SchoolHoliday     int64
dtype: object
```

In [146…
```
#Installing pandas_profiling
!pip install pandas_profiling
```

```
Requirement already satisfied: pandas_profiling in c:\users\hp\anaconda3\lib\site-packages (3.3.0)
```

In [147…
```
#import profilereport
from pandas_profiling import ProfileReport
```

```
Requirement already satisfied: multimethod<1.9,>=1.4 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (1.8)
Requirement already satisfied: statsmodels<0.14,>=0.13.2 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (0.13.2)
Requirement already satisfied: missingno<0.6,>=0.4.2 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (0.5.1)
Requirement already satisfied: joblib~=1.1.0 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (1.1.0)
Requirement already satisfied: scipy<1.10,>=1.4.1 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (1.7.3)
Requirement already satisfied: tangled-up-in-unicode==0.2.0 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (0.2.0)
Requirement already satisfied: pandas!=1.4.0,<1.5,>1.1 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (1.4.2)
Requirement already satisfied: tqdm<4.65,>=4.48.2 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (4.64.0)
Requirement already satisfied: PyYAML<6.1,>=5.0.0 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (6.0)
Requirement already satisfied: pydantic<1.10,>=1.8.1 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (1.9.2)
Requirement already satisfied: seaborn<0.12,>=0.10.1 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (0.11.2)
Requirement already satisfied: visions[type_image_path]==0.7.5 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (0.7.5)
Requirement already satisfied: numpy<1.24,>=1.16.0 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (1.21.5)
Requirement already satisfied: matplotlib<3.6,>=3.2 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (3.5.1)
Requirement already satisfied: requests<2.29,>=2.24.0 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (2.27.1)
Requirement already satisfied: jinja2<3.2,>=2.11.1 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (2.11.3)
Requirement already satisfied: phik<0.13,>=0.11.1 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (0.12.2)
Requirement already satisfied: htmlmin==0.1.12 in c:\users\hp\anaconda3\lib\site-packages (from pandas_profiling) (0.1.12)
Requirement already satisfied: networkx>=2.4 in c:\users\hp\anaconda3\lib\site-packages (from visions[type_image_path]==0.7.5->pandas_profiling) (2.7.1)
Requirement already satisfied: attrs>=19.3.0 in c:\users\hp\anaconda3\lib\site-packages (from visions[type_image_path]==0.7.5->pandas_profiling) (21.4.0)
Requirement already satisfied: imagehash in c:\users\hp\anaconda3\lib\site-packages (from visions[type_image_path]==0.7.5->pandas_profiling) (4.3.1)
Requirement already satisfied: Pillow in c:\users\hp\anaconda3\lib\site-packages (from visions[type_image_path]==0.7.5->pandas_profiling) (9.0.1)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\hp\anaconda3\lib\site-packages (from jinja2<3.2,>=2.11.1->pandas_profiling) (2.0.1)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\hp\anaconda3\lib\site-packages (from matplotlib<3.6,>=3.2->pandas_profiling) (2.8.2)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\hp\anaconda3\lib\site-packages (from matplotlib<3.6,>=3.2->pandas_profiling) (1.3.2)
Requirement already satisfied: cycler>=0.10 in c:\users\hp\anaconda3\lib\site-packages (from matplotlib<3.6,>=3.2->pandas_profiling) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\hp\anaconda3\lib\site-packages (from matplotlib<3.6,>=3.2->pandas_profiling) (4.25.0)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\hp\anaconda3\lib\site-packages (from matplotlib<3.6,>=3.2->pandas_profiling) (3.0.4)
Requirement already satisfied: packaging>=20.0 in c:\users\hp\anaconda3\lib\site-packages (from matplotlib<3.6,>=3.2->pandas_profiling) (21.3)
Requirement already satisfied: pytz>=2020.1 in c:\users\hp\anaconda3\lib\site-packages (from pandas!=1.4.0,<1.5,>1.1->pandas_profiling) (2021.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in c:\users\hp\anaconda3\lib\site-packages (from pydantic<1.10,>=1.8.1->pandas_profiling) (4.1.1)
Requirement already satisfied: six>=1.5 in c:\users\hp\anaconda3\lib\site-packages (from python-dateutil>=2.7->matplotlib<3.6,>=3.2->pandas_profiling) (1.16.0)
Requirement already satisfied: idna<4,>=2.5 in c:\users\hp\anaconda3\lib\site-packages (from requests<2.29,>=2.24.0->pandas_profiling) (3.3)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\hp\anaconda3\lib\site-packages (from requests<2.29,>=2.24.0->pandas_profiling) (1.26.9)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\hp\anaconda3\lib\site-packages (from requests<2.29,>=2.24.0->pandas_profiling) (2021.10.8)
Requirement already satisfied: charset-normalizer~=2.0.0 in c:\users\hp\anaconda3\lib\site-packages (from requests<2.29,>=2.24.0->pandas_profiling) (2.0.4)
Requirement already satisfied: patsy>=0.5.2 in c:\users\hp\anaconda3\lib\site-packages (from statsmodels<0.14,>=0.13.2->pandas_profiling) (0.5.2)
Requirement already satisfied: colorama in c:\users\hp\anaconda3\lib\site-packages (from tqdm<4.65,>=4.48.2->pandas_profiling) (0.4.4)
Requirement already satisfied: PyWavelets in c:\users\hp\anaconda3\lib\site-packages (from imagehash->visions[type_image_path]==0.7.5->pandas_profiling) (1.3.0)
```

In [148…

```
#assigning profilereport as profile
profile = ProfileReport(df)
```

In [149…

```
#Running the profile report
profile.to_file(output_file = "train_profiling.html")
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
```

```
C:\Users\hp\anaconda3\lib\site-packages\scipy\stats\stats.py:4812: RuntimeWarning: overflow encountered in long
long_scalars

  (2 * xtie * ytie) / m + x0 * y0 / (9 * m * (size - 2)))
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
Export report to file:   0%|          | 0/1 [00:00<?, ?it/s]
```

In [150…

```
#Profile view
profile
```

Overview

Dataset statistics

Number of variables	9
Number of observations	1017209
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	69.8 MiB
Average record size in memory	72.0 B

Variable types

Numeric	4
Categorical	5

Alerts

`Date` has a high cardinality: 942 distinct values	**High cardinality**
`DayOfWeek` is highly correlated with `Open`	**High correlation**
`Sales` is highly correlated with `Customers` and 2 other fields (Customers, Open, Promo)	**High correlation**
`Customers` is highly correlated with `Sales`	**High correlation**
`Open` is highly correlated with `DayOfWeek` and 2 other fields (DayOfWeek, Sales, StateHoliday)	**High correlation**
`Promo` is highly correlated with `Sales`	**High correlation**

Out[150]:

In [ ]:

```
# from the profile overview Alerts, there is a high correlation between sales and 3 other variables (customers,
#hence i can easliy conclude from statisticalinference that sales is affected by number of customers which is i
#Therefore Promo increases the number of customers purchasing goods fromthe store and this increases sales and
```