Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Final Project (Retail Forecasting)

01/16/2024

Gao Mo

Agenda

Problem Statement

Data Cleaning & Transformation

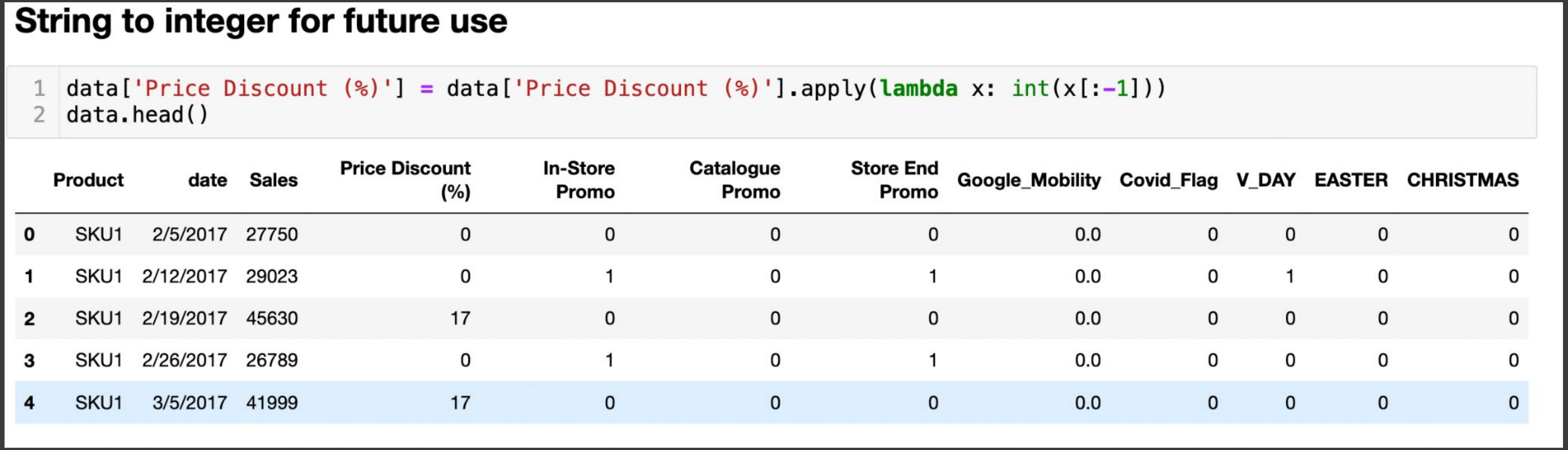
EDA

EDA Summary

Problem Statement

- A Company that sells many kinds of products requires us to replace the in-house software designed to predict retailing with an AI/ML driven method. The models that we implement should take seasonality and other additional factors into account.

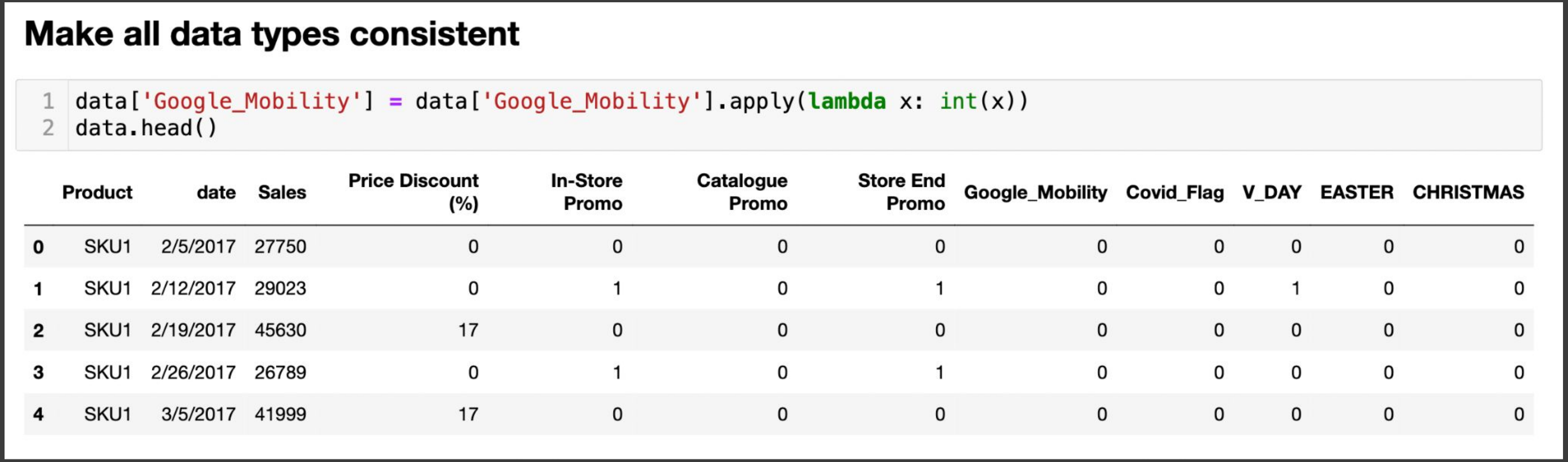
Data Cleaning & Transformation



String to integer for future use

```python
data['Price Discount (%)'] = data['Price Discount (%)'].apply(lambda x: int(x[:-1]))
data.head()
```

	Product	date	Sales	Price Discount (%)	In-Store Promo	Catalogue Promo	Store End Promo	Google_Mobility	Covid_Flag	V_DAY	EASTER	CHRISTMAS
0	SKU1	2/5/2017	27750	0	0	0	0	0.0	0	0	0	0
1	SKU1	2/12/2017	29023	0	1	0	1	0.0	0	1	0	0
2	SKU1	2/19/2017	45630	17	0	0	0	0.0	0	0	0	0
3	SKU1	2/26/2017	26789	0	1	0	1	0.0	0	0	0	0
4	SKU1	3/5/2017	41999	17	0	0	0	0.0	0	0	0	0

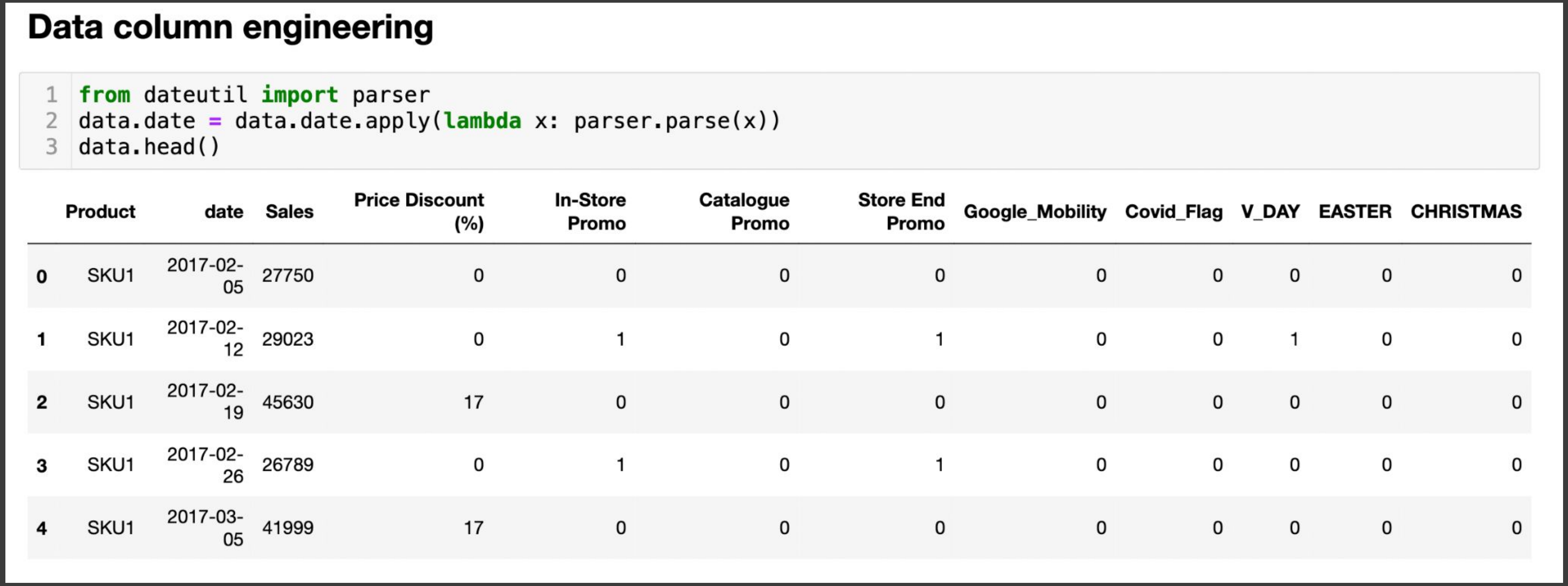


Make all data types consistent

```python
data['Google_Mobility'] = data['Google_Mobility'].apply(lambda x: int(x))
data.head()
```

	Product	date	Sales	Price Discount (%)	In-Store Promo	Catalogue Promo	Store End Promo	Google_Mobility	Covid_Flag	V_DAY	EASTER	CHRISTMAS
0	SKU1	2/5/2017	27750	0	0	0	0	0	0	0	0	0
1	SKU1	2/12/2017	29023	0	1	0	1	0	0	1	0	0
2	SKU1	2/19/2017	45630	17	0	0	0	0	0	0	0	0
3	SKU1	2/26/2017	26789	0	1	0	1	0	0	0	0	0
4	SKU1	3/5/2017	41999	17	0	0	0	0	0	0	0	0

Data Cleaning & Transformation

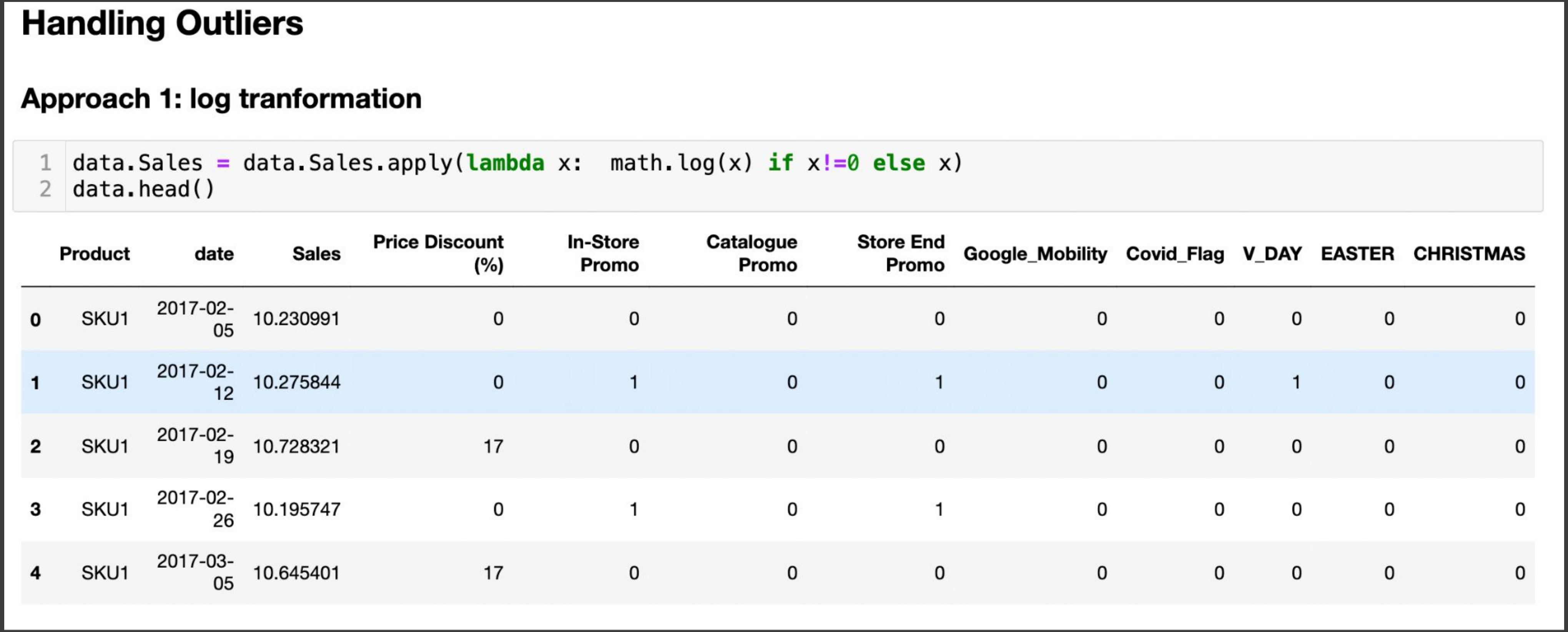


Data column engineering

```python
from dateutil import parser
data.date = data.date.apply(lambda x: parser.parse(x))
data.head()
```

	Product	date	Sales	Price Discount (%)	In-Store Promo	Catalogue Promo	Store End Promo	Google_Mobility	Covid_Flag	V_DAY	EASTER	CHRISTMAS
0	SKU1	2017-02-05	27750	0	0	0	0	0	0	0	0	0
1	SKU1	2017-02-12	29023	0	1	0	1	0	0	1	0	0
2	SKU1	2017-02-19	45630	17	0	0	0	0	0	0	0	0
3	SKU1	2017-02-26	26789	0	1	0	1	0	0	0	0	0
4	SKU1	2017-03-05	41999	17	0	0	0	0	0	0	0	0

Data Cleaning & Transformation



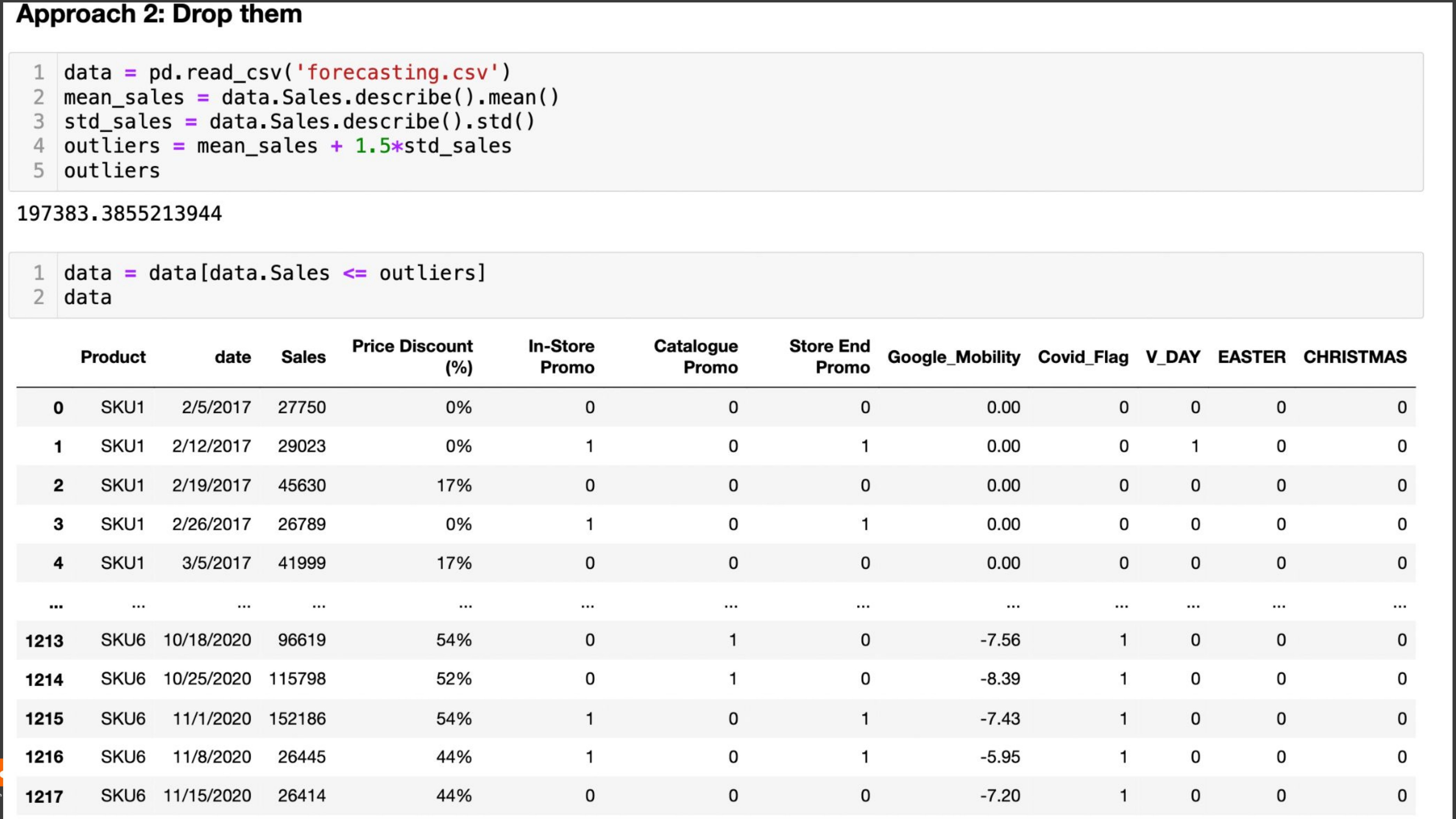
Handling Outliers

Approach 1: log tranformation

```
data.Sales = data.Sales.apply(lambda x:  math.log(x) if x!=0 else x)
data.head()
```

	Product	date	Sales	Price Discount (%)	In-Store Promo	Catalogue Promo	Store End Promo	Google_Mobility	Covid_Flag	V_DAY	EASTER	CHRISTMAS
0	SKU1	2017-02-05	10.230991	0	0	0	0	0	0	0	0	0
1	SKU1	2017-02-12	10.275844	0	1	0	1	0	0	1	0	0
2	SKU1	2017-02-19	10.728321	17	0	0	0	0	0	0	0	0
3	SKU1	2017-02-26	10.195747	0	1	0	1	0	0	0	0	0
4	SKU1	2017-03-05	10.645401	17	0	0	0	0	0	0	0	0

Data Cleaning & Transformation



Approach 2: Drop them

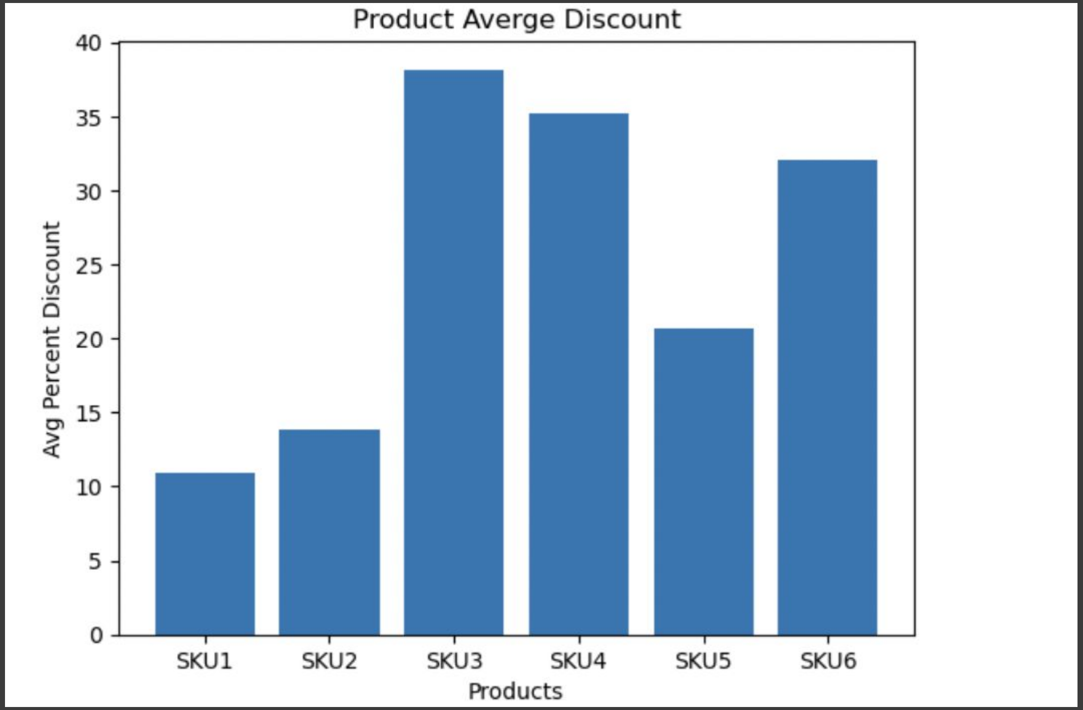
```
data = pd.read_csv('forecasting.csv')
mean_sales = data.Sales.describe().mean()
std_sales = data.Sales.describe().std()
outliers = mean_sales + 1.5*std_sales
outliers
```

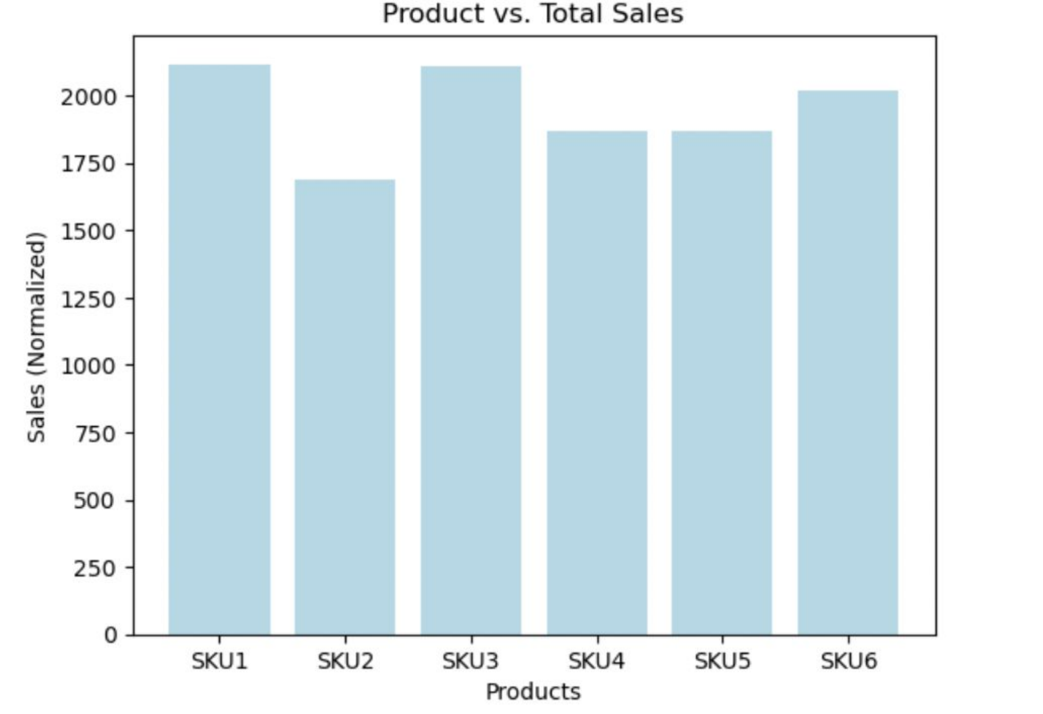
```
197383.3855213944
```

```
data = data[data.Sales <= outliers]
data
```

	Product	date	Sales	Price Discount (%)	In-Store Promo	Catalogue Promo	Store End Promo	Google_Mobility	Covid_Flag	V_DAY	EASTER	CHRISTMAS
0	SKU1	2/5/2017	27750	0%	0	0	0	0.00	0	0	0	0
1	SKU1	2/12/2017	29023	0%	1	0	1	0.00	0	1	0	0
2	SKU1	2/19/2017	45630	17%	0	0	0	0.00	0	0	0	0
3	SKU1	2/26/2017	26789	0%	1	0	1	0.00	0	0	0	0
4	SKU1	3/5/2017	41999	17%	0	0	0	0.00	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
1213	SKU6	10/18/2020	96619	54%	0	1	0	-7.56	1	0	0	0
1214	SKU6	10/25/2020	115798	52%	0	1	0	-8.39	1	0	0	0
1215	SKU6	11/1/2020	152186	54%	1	0	1	-7.43	1	0	0	0
1216	SKU6	11/8/2020	26445	44%	1	0	1	-5.95	1	0	0	0
1217	SKU6	11/15/2020	26414	44%	0	0	0	-7.20	1	0	0	0

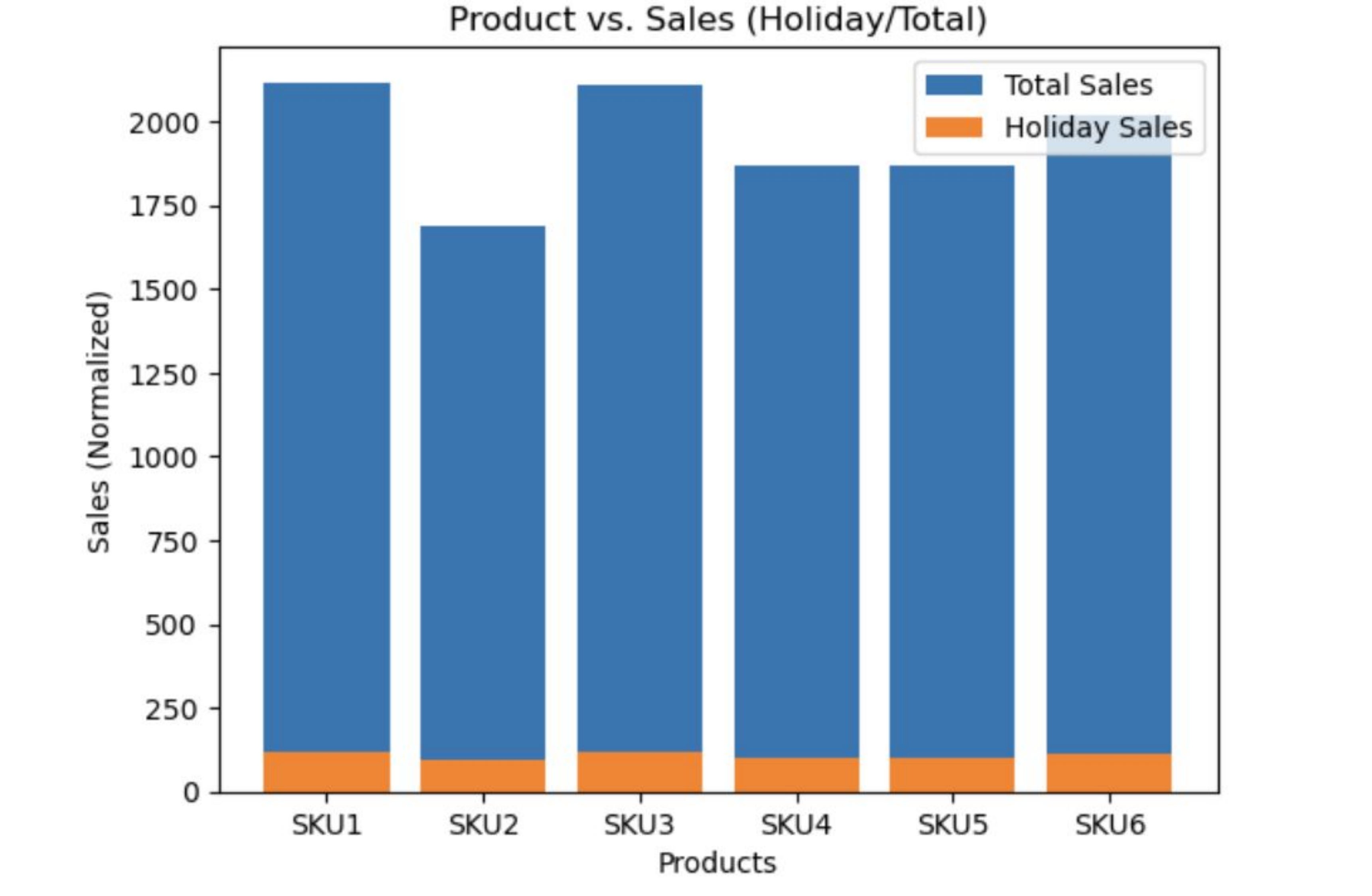
EDA





Data Glacier
Your Deep Learning Partner

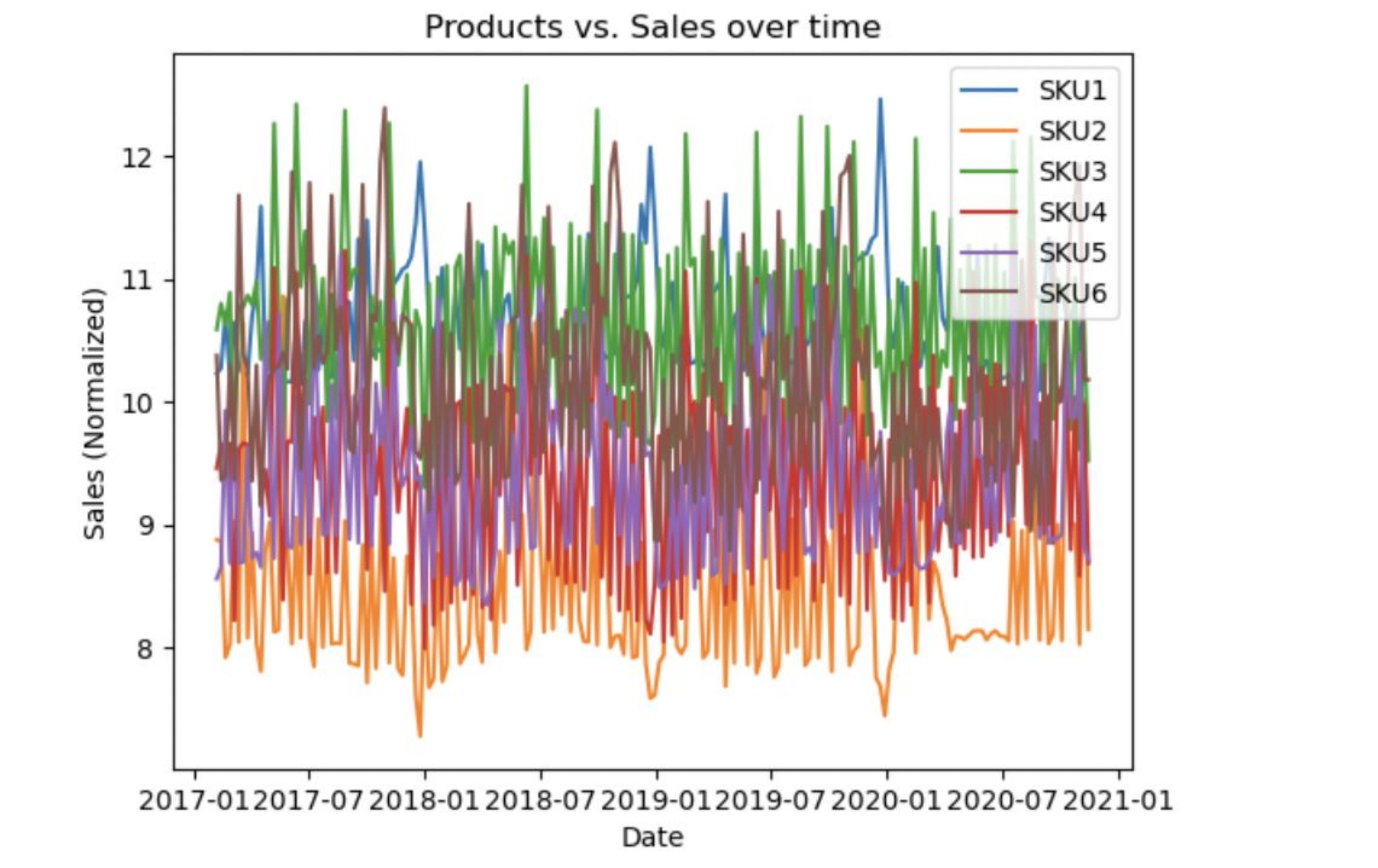
EDA



Data Glacier

Your Deep Learning Partner

EDA



Data Glacier

Your Deep Learning Partner

EDA Summary

- Products vary in the average amount of discount percentage due to seasonality & holidays
- Products vary a little in terms of sales made during specific holidays and seasonality, which shows that the total sales are not too influenced by seasonality and discount
- Product sales vary a lot over time

Thank You

Data Glacier

Your Deep Learning Partner