Final Project - Analyzing Sales Data

Date: 30 December 2021

Author: Kasidis Satangmongkol (Toy DataRockie)

Course: `Pandas Foundation`

```
# import data
import pandas as pd
df = pd.read_csv("sample-store.csv")
```

```
# preview top 5 rows
df.head()
```

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country/Region	City	...	Postal Code	Region	P I
0	1	CA-2019-152156	11/8/2019	11/11/2019	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	...	42420.0	South	F 1
1	2	CA-2019-152156	11/8/2019	11/11/2019	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	...	42420.0	South	F 1
2	3	CA-2019-138688	6/12/2019	6/16/2019	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	...	90036.0	West	C 1
3	4	US-2018-108966	10/11/2018	10/18/2018	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...	33311.0	South	F 1
4	5	US-2018-108966	10/11/2018	10/18/2018	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	...	33311.0	South	C 1

5 rows × 21 columns

```
# shape of dataframe
df.shape
```

```
(9994, 21)
```

```
# see data frame information using .info()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Row ID          9994 non-null   int64
 1   Order ID        9994 non-null   object
 2   Order Date      9994 non-null   object
 3   Ship Date       9994 non-null   object
 4   Ship Mode       9994 non-null   object
 5   Customer ID     9994 non-null   object
 6   Customer Name   9994 non-null   object
 7   Segment         9994 non-null   object
 8   Country/Region  9994 non-null   object
 9   City            9994 non-null   object
 10  State           9994 non-null   object
 11  Postal Code     9983 non-null   float64
 12  Region          9994 non-null   object
 13  Product ID      9994 non-null   object
 14  Category        9994 non-null   object
 15  Sub-Category    9994 non-null   object
 16  Product Name    9994 non-null   object
 17  Sales           9994 non-null   float64
 18  Quantity        9994 non-null   int64
 19  Discount        9994 non-null   float64
 20  Profit          9994 non-null   float64
dtypes: float64(4), int64(2), object(15)
memory usage: 1.6+ MB
```

We can use `pd.to_datetime()` function to convert columns 'Order Date' and 'Ship Date' to datetime.

```
# example of pd.to_datetime() function
pd.to_datetime(df['Order Date'].head(), format='%m/%d/%Y')
```

```
# TODO - convert order date and ship date to datetime in the original dataframe
```

```
# TODO - count nan in postal code column
```

```
# TODO - filter rows with missing values
```

```
# TODO - Explore this dataset on your owns, ask your own questions
```

Data Analysis Part

Answer 10 below questions to get credit from this course. Write `pandas` code to find answers.

```
# TODO 01 - how many columns, rows in this dataset
```

```
# TODO 02 - is there any missing values?, if there is, which colunm? how many nan values?
```

```
# TODO 03 - your friend ask for `California` data, filter it and export csv for him
```

```
# TODO 04 - your friend ask for all order data in `California` and `Texas` in 2017 (look at Order Dat
```

```
# TODO 05 - how much total sales, average sales, and standard deviation of sales your company make in
```

```
# TODO 06 - which Segment has the highest profit in 2018
```

```
# TODO 07 - which top 5 States have the least total sales between 15 April 2019 - 31 December 2019
```

```
# TODO 08 - what is the proportion of total sales (%) in West + Central in 2019 e.g. 25%
```

```
# TODO 09 - find top 10 popular products in terms of number of orders vs. total sales during 2019-202
```

```
# TODO 10 - plot at least 2 plots, any plot you think interesting :)
```

```
# TODO Bonus - use np.where() to create new column in dataframe to help you answer your own questions
```