In [2]:
```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
import os
```

In [3]:
```
df = pd.read_csv("Book1.csv")
df
```

Out[3]:

	Salesperson	Region Covered	February 2019 Sales	Cost of Sales	January 2019 Sales	Percent Change
0	Jeffrey Burke	Oklahoma	28,000	2,460	21,238	32%
1	Amy Fernadez	North Carolina	23,138	1,521	23,212	0%
2	Mark Hayes	Massachusesetts	25,092	1,530	20,454	21%
3	Judith Ray	California	21,839	1,923	24,619	-11%
4	Edward Graham	South Carolina	23,347	2,397	20,045	16%
5	Christina Foster	Delaware	23,365	1,500	17,537	33%
6	Judy Green	Texas	21,510	1,657	24,951	-14%
7	Paula Hall	Virginia	21,314	2,418	18,082	18%

In [4]:
```
df.head()
```

Out[4]:

	Salesperson	Region Covered	February 2019 Sales	Cost of Sales	January 2019 Sales	Percent Change
0	Jeffrey Burke	Oklahoma	28,000	2,460	21,238	32%
1	Amy Fernadez	North Carolina	23,138	1,521	23,212	0%
2	Mark Hayes	Massachusesetts	25,092	1,530	20,454	21%
3	Judith Ray	California	21,839	1,923	24,619	-11%
4	Edward Graham	South Carolina	23,347	2,397	20,045	16%

In [5]:
```
df.tail()
```

Out[5]:

	Salesperson	Region Covered	February 2019 Sales	Cost of Sales	January 2019 Sales	Percent Change
3	Judith Ray	California	21,839	1,923	24,619	-11%
4	Edward Graham	South Carolina	23,347	2,397	20,045	16%
5	Christina Foster	Delaware	23,365	1,500	17,537	33%
6	Judy Green	Texas	21,510	1,657	24,951	-14%
7	Paula Hall	Virginia	21,314	2,418	18,082	18%

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
Salesperson          0
Region Covered       0
February 2019 Sales  0
Cost of Sales        0
January 2019 Sales   0
Percent Change       0
dtype: int64
```

In [7]:
```
df.drop_duplicates()
```

Out[7]:

	Salesperson	Region Covered	February 2019 Sales	Cost of Sales	January 2019 Sales	Percent Change
0	Jeffrey Burke	Oklahoma	28,000	2,460	21.238	32%
1	Amy Fernadez	North Carolina	23,138	1,521	23,212	0%
2	Mark Hayes	Massachusesetts	25,092	1,530	20,454	21%
3	Judith Ray	California	21,839	1,923	24,619	-11%
4	Edward Graham	South Carolina	23,347	2,397	20,045	16%
5	Christina Foster	Delaware	23,365	1,500	17,537	33%
6	Judy Green	Texas	21,510	1,657	24,951	-14%
7	Paula Hall	Virginia	21,314	2,418	18,082	18%

In [8]:
```
df.isnull()
```

Out[8]:

	Salesperson	Region Covered	February 2019 Sales	Cost of Sales	January 2019 Sales	Percent Change
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	False	False	False	False
3	False	False	False	False	False	False
4	False	False	False	False	False	False
5	False	False	False	False	False	False
6	False	False	False	False	False	False
7	False	False	False	False	False	False

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
Salesperson          0
Region Covered       0
February 2019 Sales  0
Cost of Sales        0
January 2019 Sales   0
Percent Change       0
dtype: int64
```

In [10]:
```
df.dtypes
```

Out[10]:
```
Salesperson          object
Region Covered       object
February 2019 Sales  object
Cost of Sales        object
January 2019 Sales   object
Percent Change       object
dtype: object
```

In [11]:
```
df.describe(include = [np.object])
```

```
<ipython-input-11-bda6fcb54cc9>:1: DeprecationWarning: `np.object` is a deprecated alias for the builtin `object`. To silence this warning, use `object` by itself. Doing this will not modify any behavior and is safe.
Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  df.describe(include = [np.object])
```

Out[11]:

	Salesperson	Region Covered	February 2019 Sales	Cost of Sales	January 2019 Sales	Percent Change
count	8	8	8	8	8	8
unique	8	8	8	8	8	8
top	Jeffrey Burke	South Carolina	21,839	1,500	24,951	32%
freq	1	1	1	1	1	1

In [12]:
```
df = pd.read_csv("Book1.csv")
df.drop_duplicates(inplace = True)
X = df.iloc[:,:-1].values
Y = df.iloc[:, -1]
print(X)
print(Y)
df
```

```
[['Jeffrey Burke' 'Oklahoma' '28,000' '2,460' '21,238']
 ['Amy Fernadez' 'North Carolina' '23,138' '1,521' '23,212']
 ['Mark Hayes' 'Massachusesetts' '25,092' '1,530' '20,454']
 ['Judith Ray' 'California' '21,839' '1,923' '24,619']
 ['Edward Graham' 'South Carolina' '23,347' '2,397' '20,045']
 ['Christina Foster' 'Delaware' '23,365' '1,500' '17,537']
 ['Judy Green' 'Texas' '21,510' '1,657' '24,951']
 ['Paula Hall' 'Virginia' '21,314' '2,418' '18,082']]
0     32%
1      0%
2     21%
3    -11%
4     16%
5     33%
6    -14%
7     18%
Name: Percent Change, dtype: object
```

Out[12]:

	Salesperson	Region Covered	February 2019 Sales	Cost of Sales	January 2019 Sales	Percent Change
0	Jeffrey Burke	Oklahoma	28,000	2,460	21,238	32%
1	Amy Fernadez	North Carolina	23,138	1,521	23,212	0%
2	Mark Hayes	Massachusesetts	25,092	1,530	20,454	21%
3	Judith Ray	California	21,839	1,923	24,619	-11%
4	Edward Graham	South Carolina	23,347	2,397	20,045	16%
5	Christina Foster	Delaware	23,365	1,500	17,537	33%
6	Judy Green	Texas	21,510	1,657	24,951	-14%
7	Paula Hall	Virginia	21,314	2,418	18,082	18%

In [13]:
```
df.shape
```

Out[13]:
```
(8, 6)
```

In [14]:
```
dummy1 = pd.get_dummies(df['Region Covered'])
dummy1
```

Out[14]:

	California	Delaware	Massachusesetts	North Carolina	Oklahoma	South Carolina	Texas	Virginia
0	0	0	0	0	1	0	0	0
1	0	0	0	1	0	0	0	0
2	0	0	1	0	0	0	0	0
3	1	0	0	0	0	0	0	0
4	0	0	0	0	0	1	0	0
5	0	1	0	0	0	0	0	0
6	0	0	0	0	0	0	1	0
7	0	0	0	0	0	0	0	1

In [15]:
```
df = pd.concat([dummy1,df], axis = 1)
df
```

Out[15]:

	California	Delaware	Massachusesetts	North Carolina	Oklahoma	South Carolina	Texas	Virginia	Salesperson	Region Covered	February 2019 Sales
0	0	0	0	0	1	0	0	0	Jeffrey Burke	Oklahoma	28,000
1	0	0	0	1	0	0	0	0	Amy Fernadez	North Carolina	23,138
2	0	0	1	0	0	0	0	0	Mark Hayes	Massachusesetts	25,092
3	1	0	0	0	0	0	0	0	Judith Ray	California	21,839
4	0	0	0	0	0	1	0	0	Edward Graham	South Carolina	23,347
5	0	1	0	0	0	0	0	0	Christina Foster	Delaware	23,365
6	0	0	0	0	0	0	1	0	Judy Green	Texas	21,510
7	0	0	0	0	0	0	0	1	Paula Hall	Virginia	21,314

In [16]:
```
df['Cost of Sales'].describe()
```

Out[16]:
```
count         8
unique        8
top       1,500
freq          1
Name: Cost of Sales, dtype: object
```

In [17]:
```
df['Cost of Sales'].min()
```

Out[17]:
```
'1,500'
```

In [18]:
```
df['Cost of Sales'].max()
```

Out[18]:
```
'2,460'
```

In [ ]:
```
sns.pairplot(data=df, vars=['February 2019 Sales','January 2019 Sales','Cost of Sales'])
plt.show()
```

In [ ]:
```
plt.hist(df['Cost of Sales'])
```

In [ ]:
```
auto = pd.read_csv('Book1.csv')
```

In [ ]:
```
plt.figure(figsize=(12,8))
sns.lineplot(x = 'Cost of Sales', y = 'February 2019 Sales', data = auto, linewidth = 5)
plt.title('Cost Vs sales')
```

In [ ]:
```
plt.figure(figsize = (8,8))
sns.scatterplot(x = 'Cost of Sales', y = 'January 2019 Sales', data = auto, color = 'green')
```

In [ ]:
```
plt.figure(figsize = (20,6))
sns.lineplot(data = auto['Cost of Sales'], linewidth = 1.5, label = 'Cost of Sales')
sns.lineplot(data = auto['January 2019 Sales'], linewidth = 1.5, label = 'January 2019 Sales')
```

In [ ]:
```
```