Exploratory Data Analysis Project

Nutritional Supplement Sales for Multiple Regression

Nestele is a corporate company that has a broad portfolio of brands under its umbrella. Nestele Healthcare produces and commercialises Nutritional Supplements in many European Countries to date. The corporate wants to expand the presence of Peptamant Plus, one of their best-selling products, in a new nation and wants to forecast the amount of sales using data from their current market presence.

Please Note : The dataset provided is for learning purpose. Please don't draw any inference with real world scenario.

Summary of Attributes

There are 7 attributes inthe peptamant sales per country data including the target variable sales

- Country: Country of Sales
- Population: population of country in millions
- Sales: Peptamant Plus Sales in millions
- Sales_per_capita: Sales per capita
- GNP per capita: Gross National Product
- Unemployment rate: Unemployment rate as a function of GNP
- Healthcare spending: Cost of Healthcare as a function of GNP

Actions Performed

- Simple EDA
- View the First few rows
- Clean data to analyzable Format
- Pairplot for Features
- Correlation Matrix of Features
- Hypothesis Definition
- Conclusion and Discussion of Hypothesis Testing
- Overall Comments

In [1]:
```
# Import Necessary Libraries
import pandas as pd
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
```

In [14]:
```
peptamant_data = pd.read_csv('peptamant2.csv', index_col='Observation\n')
cols = ['Country', 'Population', 'Sales', 'Sales_per_capita', 'GNP_per_capita', 'Unemp
peptamant_data.columns = cols
peptamant_data.head()
```

Out[14]:

Observation	Country	Population	Sales	Sales_per_capita	GNP_per_capita	Unemployment_rate	Healthcare_sp
1	Austria	8,4	941,2	112,05	49600	4,2	
2	Belgium	10,5	1681,9	160,18	47090	8,1	
3	Bulgaria	7,6	154	20,26	6550	13,5	
4	Czech Rep.	10,2	1028,7	100,85	20670	6,6	
5	Denmark	5,5	935,4	170,07	62120	5,2	

In [38]:
```
peptamant_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 20 entries, 1 to 21
Data columns (total 7 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Country             20 non-null     object
 1   Population          20 non-null     object
 2   Sales               20 non-null     object
 3   Sales_per_capita    20 non-null     object
 4   GNP_per_capita      20 non-null     int64 
 5   Unemployment_rate   20 non-null     object
 6   Healthcare_spending 20 non-null     object
dtypes: int64(1), object(6)
memory usage: 1.2+ KB
```

Cleaning

- Change the data types to relevant ones

In [37]:
```
df = peptamant_data.copy()
new_cols = [col for col in df.columns if col != 'Country' and col != 'GNP_per_capita']
for col in new_cols:
    df[col] = df[col].str.replace(',', '.')

conv_dict = {col: 'float' for col in new_cols}
df = df.astype(conv_dict)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 20 entries, 1 to 21
Data columns (total 7 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Country             20 non-null     object 
 1   Population          20 non-null     float64
 2   Sales               20 non-null     float64
 3   Sales_per_capita    20 non-null     float64
 4   GNP_per_capita      20 non-null     int64  
 5   Unemployment_rate   20 non-null     float64
 6   Healthcare_spending 20 non-null     float64
dtypes: float64(5), int64(1), object(1)
memory usage: 1.2+ KB
```

3 Hypothesis to test

1. Peptamant Sales is affected by some variables e.g Population
2. Peptament Sales is not affected by some variables e.g Gross National Product
3. Peptament Sales is not a function of Nesteles current Market Presence

Testing Hypothesis 1: Peptamant Sales is affected by Population

Null Hypothesis: Peptamant Sales is not affected by Population i.e there is no correlation or causation between sales of peptamant and population

Alternative Hypothesis: There is a significant relationship between peptamant sales and the population of a country

The test statistic is correlation co-efficient and R-squared, If the null hypothesis is correct, observed correlation should be less than 0.5, Significance level to reject null is 0.05

In [43]:
```
display(df.corr()[['Sales']])

df.corr()
```

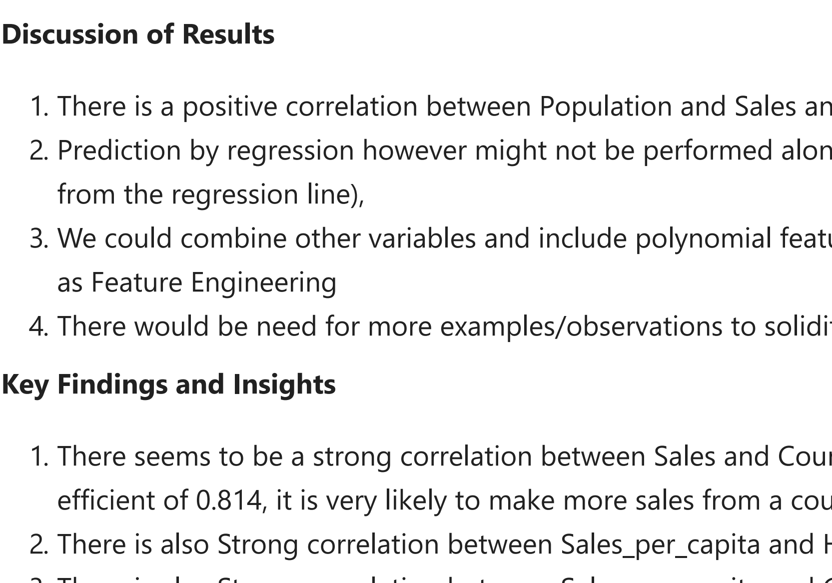
	Sales
Population	0.814402
Sales	1.000000
Sales_per_capita	0.174679
GNP_per_capita	0.162740
Unemployment_rate	0.179925
Healthcare_spending	-0.043954

Out[43]:

	Population	Sales	Sales_per_capita	GNP_per_capita	Unemployment_rate	Healthcare_s
Population	1.000000	0.814402	-0.243194	-0.147008	0.355227	-
Sales	0.814402	1.000000	0.174679	0.162740	0.179925	-
Sales_per_capita	-0.243194	0.174679	1.000000	0.807941	-0.472358	
GNP_per_capita	-0.147008	0.162740	0.807941	1.000000	-0.538765	
Unemployment_rate	0.355227	0.179925	-0.472358	-0.538765	1.000000	-
Healthcare_spending	-0.265028	-0.043954	0.706113	0.574156	-0.351221	

In [82]:
```
# Seaborn Regression Plot to view correlation between population and Sales
sns.regplot(df.Population, df.Sales)
plt.xlim(0, df.Population.max()+10)
plt.show()
```

```
C:\Users\toluo\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: P
ass the following variables as keyword args: x, y. From version 0.12, the only valid p
ositional argument will be `data`, and passing other arguments without an explicit key
word will result in an error or misinterpretation.
  warnings.warn(
```



Discussion of Results

1. There is a positive correlation between Population and Sales and it is greater than 0.5, around 0.8;
2. Prediction by regression however might not be performed alone because of a high residual(distance from the regression line),
3. We could combine other variables and include polynomial features and interactions between variables as Feature Engineering
4. There would be need for more examples/observations to solidify claims

Key Findings and Insights

1. There seems to be a strong correlation between Sales and Country Population, with corrrelation co-efficient of 0.814, it is very likely to make more sales from a country with greater population
2. There is also Strong correlation between Sales_per_capita and Healthcare_spending
3. There is also Strong correlation between Sales_per_capita and GNP_per_capita

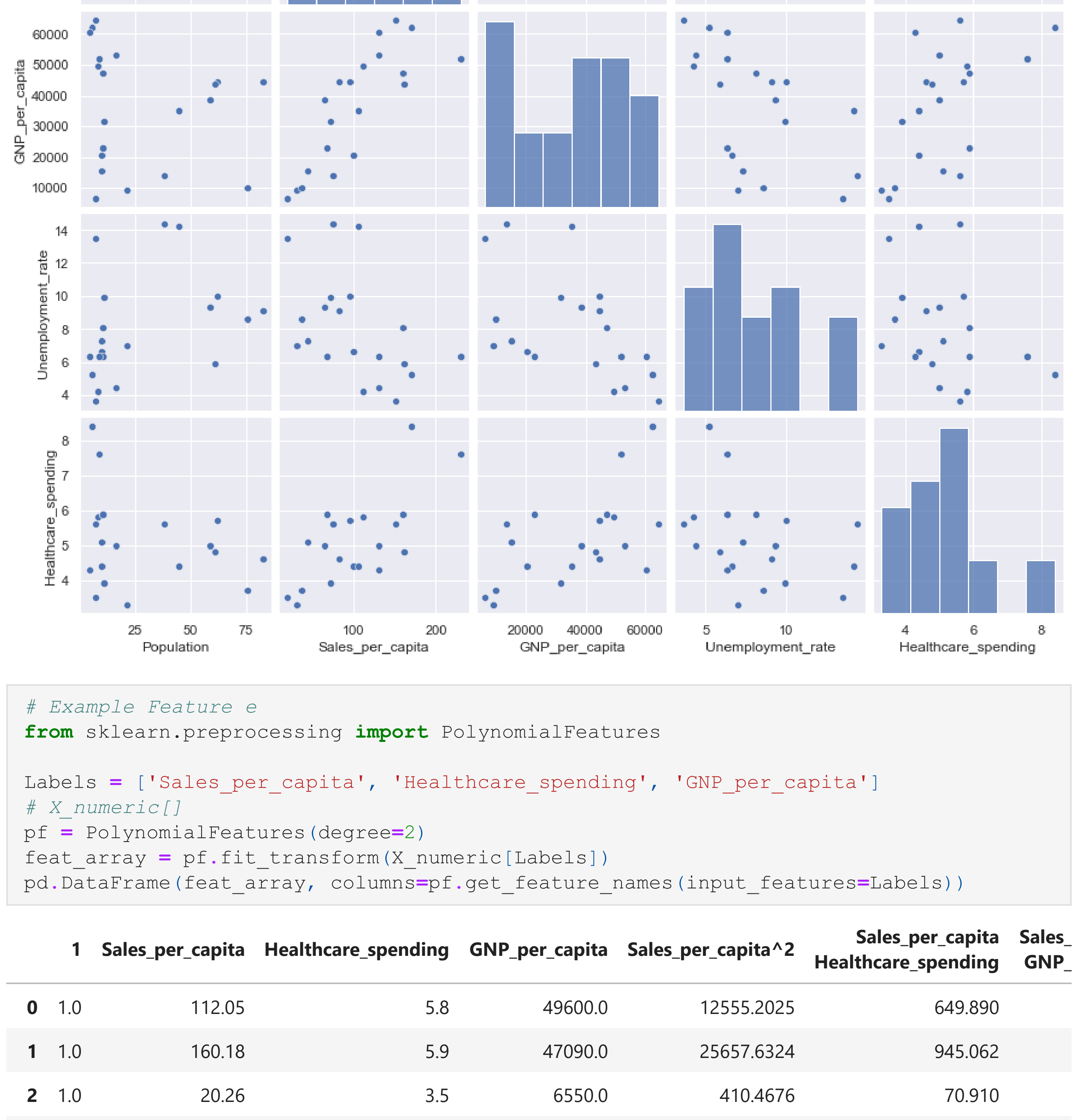
In [75]:
```
X_category = pd.get_dummies(df.Country, drop_first=True)
y = df.Sales
X_numeric = df.drop(['Sales', 'Country'], axis=1)

X_numeric.head()
```

Out[75]:

Observation	Population	Sales_per_capita	GNP_per_capita	Unemployment_rate	Healthcare_spending
1	8.4	112.05	49600	4.2	5.8
2	10.5	160.18	47090	8.1	5.9
3	7.6	20.26	6550	13.5	3.5
4	10.2	100.85	20670	6.6	4.4
5	5.5	170.07	62120	5.2	8.4

In [76]:
```
sns.pairplot(X_numeric)
plt.show()
```



In [81]:
```
# Example Feature e
from sklearn.preprocessing import PolynomialFeatures

Labels = ['Sales_per_capita', 'Healthcare_spending', 'GNP_per_capita']
# X_numeric[]
pf = PolynomialFeatures(degree=2)
feat_array = pf.fit_transform(X_numeric[Labels])
pd.DataFrame(feat_array, columns=pf.get_feature_names(input_features=Labels))
```

Out[81]:

	1	Sales_per_capita	Healthcare_spending	GNP_per_capita	Sales_per_capita^2	Sales_per_capita Healthcare_spending
0	1.0	112.05	5.8	49600.0	12555.2025	649.890
1	1.0	160.18	5.9	47090.0	25657.6324	945.062
2	1.0	20.26	3.5	6550.0	410.4676	70.910
3	1.0	100.85	4.4	20670.0	10170.7225	443.740
4	1.0	170.07	8.4	62120.0	28923.8049	1428.588
5	1.0	95.78	5.7	44510.0	9173.8084	545.946
6	1.0	82.72	4.6	44450.0	6842.5984	380.512
7	1.0	72.59	3.9	31670.0	5269.3081	283.101
8	1.0	44.90	5.1	15410.0	2016.0100	228.990
9	1.0	131.11	4.3	60460.0	17189.8321	563.773
10	1.0	65.50	5.0	38490.0	4290.2500	327.500
11	1.0	131.42	5.0	52960.0	17271.2164	657.100
12	1.0	74.92	5.6	13850.0	5613.0064	419.552
13	1.0	68.09	5.9	22920.0	4636.2481	401.731
14	1.0	32.26	3.3	9300.0	1040.7076	106.458
15	1.0	105.93	4.4	35220.0	11221.1649	466.092
16	1.0	150.72	5.6	64430.0	22716.5184	844.032
17	1.0	229.72	7.6	51950.0	52771.2784	1745.872
18	1.0	37.98	3.7	9940.0	1442.4804	140.526
19	1.0	162.09	4.8	43540.0	26273.1681	778.032

Overall Summary of Dataset and Results

The dataset was gotten from Kaggle and its a sample dataset for exploratory data analysis, and it is to determine the predictability of sales for expansion to new countries with a given number of independent variables. There is the need for more data to be able to substantiate claims from the hypothesis that htere is correlation between Peptament Sales and a Country's population

Next Steps

- Scale the numeric features both engineered and not-engineered with StandardScaler
- Concatenate Category Dataframe and numeric(without engineered features) dataframe, define a linear regression model and score
- Concatenate Category Dataframe and numeric(with engineered features) dataframe, define a linear regression model and score
- Then compare to see if there was improvement in scoring with new features