Hands-on Activity 11.1 Linear Regression Analysis

Submitted by: Dela Cruz, Eugene D.G.

Submitted to: Engr. Roman Richard

Import/download libraries and load dataset

```
!pip install hvplot
```

```
Collecting hvplot
  Downloading hvplot-0.9.2-py2.py3-none-any.whl (1.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 7.5 MB/s eta 0:00:00
Requirement already satisfied: bokeh>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (3.3.4)
Requirement already satisfied: colorcet>=2 in /usr/local/lib/python3.10/dist-packages (from hvplot) (3.1.0)
Requirement already satisfied: holoviews>=1.11.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (1.17.1)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from hvplot) (2.0.3)
Requirement already satisfied: numpy>=1.15 in /usr/local/lib/python3.10/dist-packages (from hvplot) (1.25.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from hvplot) (24.0)
Requirement already satisfied: panel>=0.11.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (1.3.8)
Requirement already satisfied: param<3.0,>=1.12.0 in /usr/local/lib/python3.10/dist-packages (from hvplot) (2.1.0)
Requirement already satisfied: Jinja2>=2.9 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (3.1.3)
Requirement already satisfied: contourpy>=1 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (1.2.1)
Requirement already satisfied: pillow>=7.1.0 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (9.4.0)
Requirement already satisfied: PyYAML>=3.10 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (6.0.1)
Requirement already satisfied: tornado>=5.1 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (6.3.3)
Requirement already satisfied: xyzservices>=2021.09.1 in /usr/local/lib/python3.10/dist-packages (from bokeh>=1.0.0->hvplot) (2024.4.0)
Requirement already satisfied: pyviz-comms>=0.7.4 in /usr/local/lib/python3.10/dist-packages (from holoviews>=1.11.0->hvplot) (3.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas->hvplot) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas->hvplot) (2023.4)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas->hvplot) (2024.1)
Requirement already satisfied: markdown in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (3.6)
Requirement already satisfied: markdown-it-py in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (3.0.0)
Requirement already satisfied: linkify-it-py in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (2.0.3)
Requirement already satisfied: mdit-py-plugins in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (0.4.0)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (2.31.0)
Requirement already satisfied: tqdm>=4.48.0 in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (4.66.2)
Requirement already satisfied: bleach in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (6.1.0)
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.10/dist-packages (from panel>=0.11.0->hvplot) (4.11.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from Jinja2>=2.9->bokeh>=1.0.0->hvplot) (2.1.5)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2->pandas->hvplot) (1.16.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-packages (from bleach->panel>=0.11.0->hvplot) (0.5.1)
Requirement already satisfied: uc-micro-py in /usr/local/lib/python3.10/dist-packages (from linkify-it-py->panel>=0.11.0->hvplot) (1.0.3)
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.10/dist-packages (from markdown-it-py->panel>=0.11.0->hvplot) (0.1.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (2024.2.2)
Installing collected packages: hvplot
Successfully installed hvplot-0.9.2
```

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import seaborn as sns
import hvplot.pandas

from sklearn.model_selection import train_test_split

from sklearn import metrics

from sklearn.linear_model import LinearRegression
%matplotlib inline



df = pd.read_csv('/content/Life Expectancy Data.csv')
```

```
!pip install scikit-learn
```

```
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.10/dist-packages (1.2.2)
Requirement already satisfied: numpy>=1.17.3 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (1.25.2)
Requirement already satisfied: scipy>=1.3.2 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (1.11.4)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (1.4.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (3.4.0)
```

Data wrangling/cleaning

```
df.head()
```

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	expe
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Country                          2938 non-null   object
 1   Year                             2938 non-null   int64
 2   Status                           2938 non-null   object
 3   Life expectancy                  2928 non-null   float64
 4   Adult Mortality                  2928 non-null   float64
 5   infant deaths                    2938 non-null   int64
 6   Alcohol                          2744 non-null   float64
 7   percentage expenditure           2938 non-null   float64
 8   Hepatitis B                      2385 non-null   float64
 9   Measles                          2938 non-null   int64
 10   BMI                             2904 non-null   float64
 11  under-five deaths                2938 non-null   int64
 12  Polio                            2919 non-null   float64
 13  Total expenditure                2712 non-null   float64
 14  Diphtheria                       2919 non-null   float64
 15   HIV/AIDS                        2938 non-null   float64
 16  GDP                              2490 non-null   float64
 17  Population                       2286 non-null   float64
 18   thinness  1-19 years            2904 non-null   float64
 19   thinness 5-9 years             2904 non-null   float64
 20  Income composition of resources  2771 non-null   float64
 21  Schooling                        2775 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

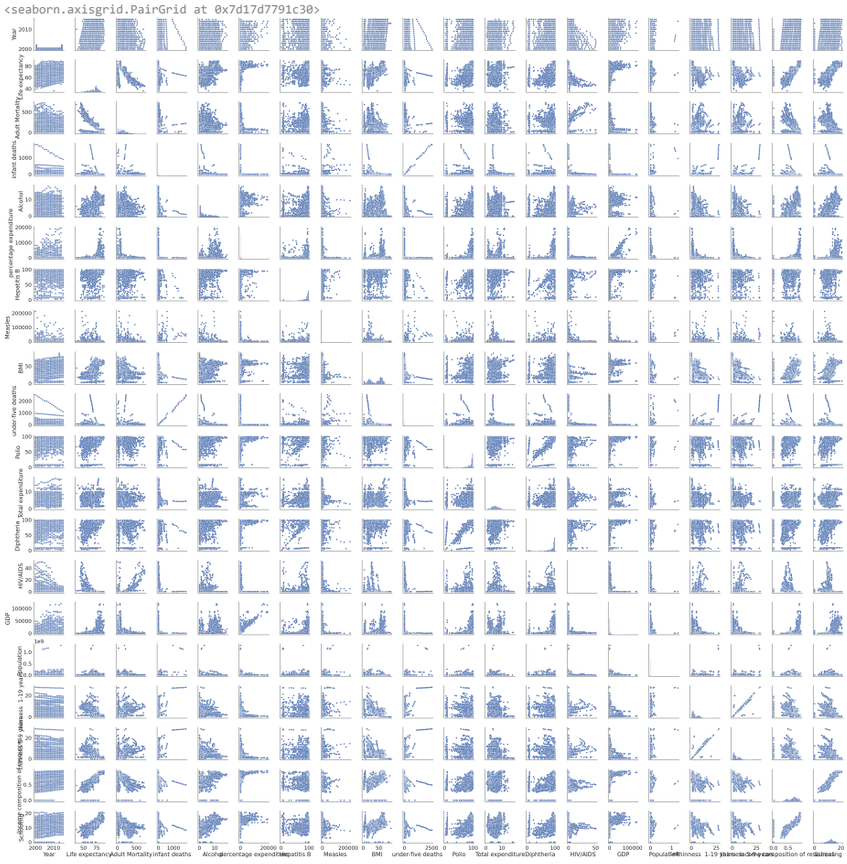
```
df.describe(include='all')
```

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B
count	2938	2938.000000	2938	2928.000000	2928.000000	2938.000000	2744.000000	2938.000000	2385.000000
unique	193	NaN	2	NaN	NaN	NaN	NaN	NaN	NaN
top	Afghanistan	NaN	Developing	NaN	NaN	NaN	NaN	NaN	NaN
freq	16	NaN	2426	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	2007.518720	NaN	69.224932	164.796448	30.303948	4.602861	738.251295	80.940461
std	NaN	4.613841	NaN	9.523867	124.292079	117.926501	4.052413	1987.914858	25.070016
min	NaN	2000.000000	NaN	36.300000	1.000000	0.000000	0.010000	0.000000	1.000000
25%	NaN	2004.000000	NaN	63.100000	74.000000	0.000000	0.877500	4.685343	77.000000
50%	NaN	2008.000000	NaN	72.100000	144.000000	3.000000	3.755000	64.912906	92.000000
75%	NaN	2012.000000	NaN	75.700000	228.000000	22.000000	7.702500	441.534144	97.000000
max	NaN	2015.000000	NaN	89.000000	723.000000	1800.000000	17.870000	19479.911610	99.000000

11 rows × 22 columns

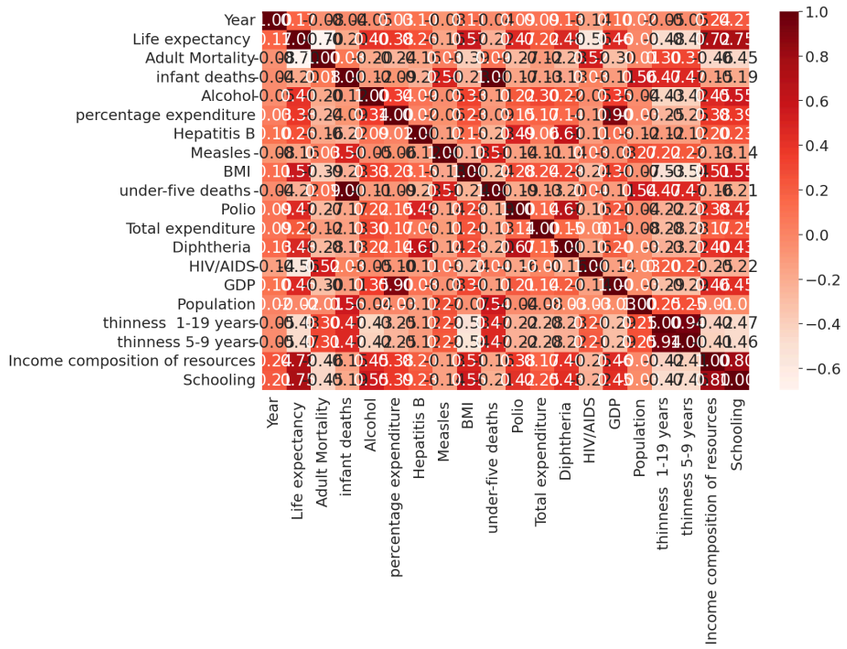
```
sns.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x7d17d7791c30>
```




```
numval = df.select_dtypes(include=[np.number])

correlation_matrix = numval.corr()
plt.figure(figsize=(16, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='Reds', fmt=".2f")
plt.show()
```




```
numval = numval.dropna() #remove missing values
```

```
numval.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1649 entries, 0 to 2937
Data columns (total 20 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Year                             1649 non-null   int64  
 1   Life expectancy                  1649 non-null   float64
 2   Adult Mortality                  1649 non-null   float64
 3   infant deaths                    1649 non-null   int64  
 4   Alcohol                          1649 non-null   float64
 5   percentage expenditure           1649 non-null   float64
 6   Hepatitis B                      1649 non-null   float64
 7   Measles                          1649 non-null   int64  
 8    BMI                             1649 non-null   float64
 9   under-five deaths                1649 non-null   int64  
 10  Polio                            1649 non-null   float64
 11  Total expenditure                1649 non-null   float64
 12  Diphtheria                       1649 non-null   float64
 13   HIV/AIDS                        1649 non-null   float64
 14  GDP                              1649 non-null   float64
 15  Population                       1649 non-null   float64
 16   thinness  1-19 years            1649 non-null   float64
 17   thinness 5-9 years             1649 non-null   float64
 18  Income composition of resources  1649 non-null   float64
 19  Schooling                        1649 non-null   float64
dtypes: float64(16), int64(4)
memory usage: 270.5 KB
```

Training

```
numval
```

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Di
0	2015	65.0	263.0	62	0.01	71.279624	65.0	1154	19.1	83	6.0	8.16	
1	2014	59.9	271.0	64	0.01	73.523582	62.0	492	18.6	86	58.0	8.18	
2	2013	59.9	268.0	66	0.01	73.219243	64.0	430	18.1	89	62.0	8.13	
3	2012	59.5	272.0	69	0.01	78.184215	67.0	2787	17.6	93	67.0	8.52	
4	2011	59.2	275.0	71	0.01	7.097109	68.0	3013	17.2	97	68.0	7.87	
...	...	...	...	...	...	...	...	...	...	...	...	...	
2933	2004	44.3	723.0	27	4.36	0.000000	68.0	31	27.1	42	67.0	7.13	
2934	2003	44.5	715.0	26	4.06	0.000000	7.0	998	26.7	41	7.0	6.52	
2935	2002	44.8	73.0	25	4.43	0.000000	73.0	304	26.3	40	73.0	6.53	
2936	2001	45.3	686.0	25	1.72	0.000000	76.0	529	25.9	39	76.0	6.16	
2937	2000	46.0	665.0	24	1.68	0.000000	79.0	1483	25.5	39	78.0	7.10	

Next steps: View recommended plots

```
X = numval.drop('infant deaths',axis=1)
y = numval['infant deaths']
```

```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3,random_state=101)
```

```
X_train.shape
```

```
(1154, 19)
```

```
y_train.shape
```

```
(1154,)
```

```
X_test.shape
```

```
(495, 19)
```

```
y_test.shape
```

```
(495,)
```

Linear Regression

```
model = LinearRegression()
```

```
model.fit(X_train, y_train)
```

```
LinearRegression
LinearRegression()
```

Evaluation

```
MC = model.coef_
```

```
pd.DataFrame(MC, X.columns, columns=['Coefficients'])
```

	Coefficients
Year	-1.143588e-01
Life expectancy	4.699094e-01
Adult Mortality	-8.910728e-04
Alcohol	-5.618026e-01
percentage expenditure	-1.466678e-04
Hepatitis B	1.108800e-02
Measles	2.712676e-04
BMI	-9.306181e-03
under-five deaths	7.033567e-01
Polio	2.039183e-02
Total expenditure	-1.114730e-01
Diphtheria	2.822234e-02
HIV/AIDS	2.005750e-01
GDP	-8.976582e-06
Population	5.567986e-08
thinness 1-19 years	3.031496e-02
thinness 5-9 years	1.682059e-01
Income composition of resources	1.810060e+00
Schooling	-1.887552e-01

Prediction for Model

```
y_pred = model.predict(X_test)
```

```
MAE = metrics.mean_absolute_error(y_test,y_pred)
MSE = metrics.mean_squared_error(y_test,y_pred)
RMSE = np.sqrt(MSE)
```

```
MAE
```

```
4.10039656169458
```

```
MSE
```

```
80.82112730395033
```

```
RMSE
```

```
8.990057135744484
```

```
df['infant deaths'].mean()
```

```
30.303948264125257
```

Residual Histogram

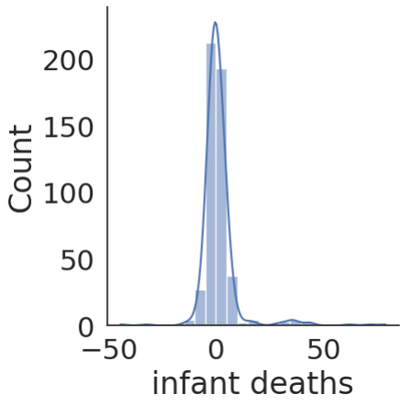
```
test_resid = y_test - y_pred
```

```
pd.DataFrame({'Error Values' : (test_resid)}).hvplot.kde()
```

```
/usr/local/lib/python3.10/dist-packages/holoviews/core/util.py:1585: PanelDeprecationWarning: 'param_value_if_widget' is deprecated and will be removed in version 1.4, use 'transform_r
  value = param_value_if_widget(value)
```

```
sns.displot(test_resid, bins=25, kde=True)
```

```
<seaborn.axisgrid.FacetGrid at 0x7d17e6ae5540>
```




```
sns.scatterplot(x=y_test, y=test_resid)

plt.axhline(y=0, color='r', ls='--')
```

```
<matplotlib.lines.Line2D at 0x7d17d7fabd90>
```