Hands-on Activity 11.1 Linear Regression Analysis

Objective(s):

- This activity aims to demonstrate how to apply simple linear regression analysis to solve regression problem

Intended Learning Outcomes (ILOs):

- Demonstrate how to solve regression problems using simple linear regression
- Use the linear regression model to predict the target value

Resources:

- Jupyter Notebook

Fetching the data and importing important libraries

```
!pip install hvplot
```

```
Collecting hvplot
  Downloading hvplot-0.9.2-py2.py3-none-any.whl (1.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 9.8 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
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2->pandas->hvplot) (1.16.
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.
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.
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplo
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.
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->panel>=0.11.0->hvplot) (20
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 hvplot.pandas

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import LabelEncoder

from sklearn import metrics

from sklearn.linear_model import LinearRegression

%matplotlib inline

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

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	H
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	
...	...	...	...	...	...	...	...	...	
2933	Zimbabwe	2004	Developing	44.3	723.0	27	4.36	0.000000	
2934	Zimbabwe	2003	Developing	44.5	715.0	26	4.06	0.000000	
2935	Zimbabwe	2002	Developing	44.8	73.0	25	4.43	0.000000	
2936	Zimbabwe	2001	Developing	45.3	686.0	25	1.72	0.000000	
2937	Zimbabwe	2000	Developing	46.0	665.0	24	1.68	0.000000	

2938 rows × 22 columns

Data Cleaning and Data Wrangling

```
df.rename(columns=lambda x: x.strip(), inplace=True)
```

```
df.head()
```

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hep
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	

5 rows × 22 columns

```
df.shape
```

```
(2938, 22)
```

```
df.isnull().sum()
```

```
Country                             0
Year                                0
Status                              0
Life expectancy                     0
Adult Mortality                     0
infant deaths                       0
Alcohol                             0
percentage expenditure              0
Hepatitis B                         0
Measles                             0
BMI                                 0
under-five deaths                   0
Polio                               0
Total expenditure                 226
Diphtheria                         19
HIV/AIDS                            0
GDP                               448
Population                        652
thinness  1-19 years               34
thinness 5-9 years                 34
Income composition of resources   167
Schooling                         163
Country_encoded                     0
Status_encoded                      0
dtype: int64
```

```
df['Life expectancy'] = df['Life expectancy'].fillna(df['Life expectancy'].mean())
df['Adult Mortality'] = df['Adult Mortality'].fillna(df['Adult Mortality'].mean())
df['Alcohol'] = df['Alcohol'].fillna(df['Alcohol'].mean())
df['Hepatitis B'] = df['Hepatitis B'].fillna(df['Hepatitis B'].mean())
df['BMI'] = df['BMI'].fillna(df['BMI'].mean())
df['Polio'] = df['Polio'].fillna(df['Polio'].mean())
```

```
df.columns
```

```
Index(['Country', 'Year', 'Status', 'Life expectancy', 'Adult Mortality',
       'infant deaths', 'Alcohol', 'percentage expenditure', 'Hepatitis B',
       'Measles', 'BMI', 'under-five deaths', 'Polio', 'Total expenditure',
       'Diphtheria', 'HIV/AIDS', 'GDP', 'Population', 'thinness  1-19 years',
       'thinness 5-9 years', 'Income composition of resources', 'Schooling'],
      dtype='object')
```

```
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
```

```
categorical_columns = ['Country','Status']

label_encoder = LabelEncoder()
for column in categorical_columns:
    encoded_column_name = column + '_encoded'
    df[encoded_column_name] = label_encoder.fit_transform(df[column])
```

```
df
```

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109
...	...	...	...	...	...	...	...	...
2933	Zimbabwe	2004	Developing	44.3	723.0	27	4.36	0.000000
2934	Zimbabwe	2003	Developing	44.5	715.0	26	4.06	0.000000
2935	Zimbabwe	2002	Developing	44.8	73.0	25	4.43	0.000000
2936	Zimbabwe	2001	Developing	45.3	686.0	25	1.72	0.000000
2937	Zimbabwe	2000	Developing	46.0	665.0	24	1.68	0.000000

2938 rows × 24 columns

```
df.drop(['Country','Status'], axis=1)
```

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BM
0	2015	65.0	263.0	62	0.01	71.279624	65.0	1154	19.
1	2014	59.9	271.0	64	0.01	73.523582	62.0	492	18.
2	2013	59.9	268.0	66	0.01	73.219243	64.0	430	18.
3	2012	59.5	272.0	69	0.01	78.184215	67.0	2787	17.
4	2011	59.2	275.0	71	0.01	7.097109	68.0	3013	17.
...	...	...	...	...	...	...	...	...	.
2933	2004	44.3	723.0	27	4.36	0.000000	68.0	31	27.
2934	2003	44.5	715.0	26	4.06	0.000000	7.0	998	26.
2935	2002	44.8	73.0	25	4.43	0.000000	73.0	304	26.
2936	2001	45.3	686.0	25	1.72	0.000000	76.0	529	25.
2937	2000	46.0	665.0	24	1.68	0.000000	79.0	1483	25.

2938 rows × 22 columns

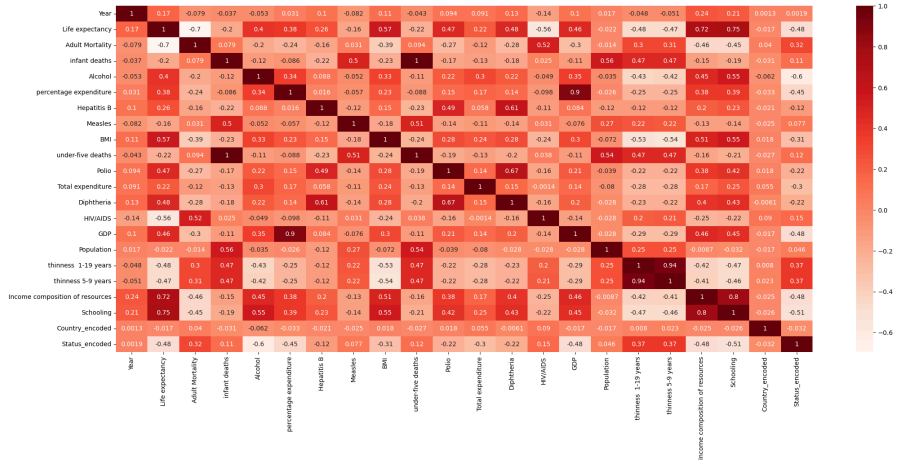
```
numeric_df = df.select_dtypes(include=[np.number])
numeric_df.corr()
```

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepa
Year	1.000000	0.170033	-0.079052	-0.037415	-0.052990	0.031400	0.1
Life expectancy	0.170033	1.000000	-0.696359	-0.196557	0.404877	0.381864	0.2
Adult Mortality	-0.079052	-0.696359	1.000000	0.078756	-0.195848	-0.242860	-0.1
infant deaths	-0.037415	-0.196557	0.078756	1.000000	-0.115638	-0.085612	-0.2
Alcohol	-0.052990	0.404877	-0.195848	-0.115638	1.000000	0.341285	0.0
percentage expenditure	0.031400	0.381864	-0.242860	-0.085612	0.341285	1.000000	0.0
Hepatitis B	0.104333	0.256762	-0.162476	-0.223566	0.087549	0.016274	1.0
Measles	-0.082493	-0.157586	0.031176	0.501128	-0.051827	-0.056596	-0.1
BMI	0.108974	0.567694	-0.387017	-0.227279	0.330408	0.228700	0.1
under-five deaths	-0.042937	-0.222529	0.094146	0.996629	-0.112370	-0.087852	-0.2
Polio	0.094158	0.465556	-0.274823	-0.170689	0.221734	0.147259	0.4
Total expenditure	0.090740	0.218086	-0.115281	-0.128616	0.296942	0.174420	0.0
Diphtheria	0.134337	0.479495	-0.275131	-0.175171	0.222020	0.143624	0.0
HIV/AIDS	-0.139741	-0.556556	0.523821	0.025231	-0.048845	-0.097857	-0.
GDP	0.101620	0.461455	-0.296049	-0.108427	0.354712	0.899373	0.0
Population	0.016969	-0.021538	-0.013647	0.556801	-0.035252	-0.025662	-0.1
thinness 1-19 years	-0.047876	-0.477183	0.302904	0.465711	-0.428795	-0.251369	-0.1
thinness 5-9 years	-0.050929	-0.471584	0.308457	0.471350	-0.417414	-0.252905	-0.1
Income composition of resources	0.243468	0.724776	-0.457626	-0.145139	0.450040	0.381952	0.1
Schooling	0.209400	0.751975	-0.454612	-0.193720	0.547378	0.389687	0.
Country_encoded	0.001342	-0.016763	0.039802	-0.030528	-0.062134	-0.032983	-0.0
Status_encoded	0.001864	-0.482136	0.315284	0.112252	-0.596660	-0.454261	-0.

22 rows × 22 columns

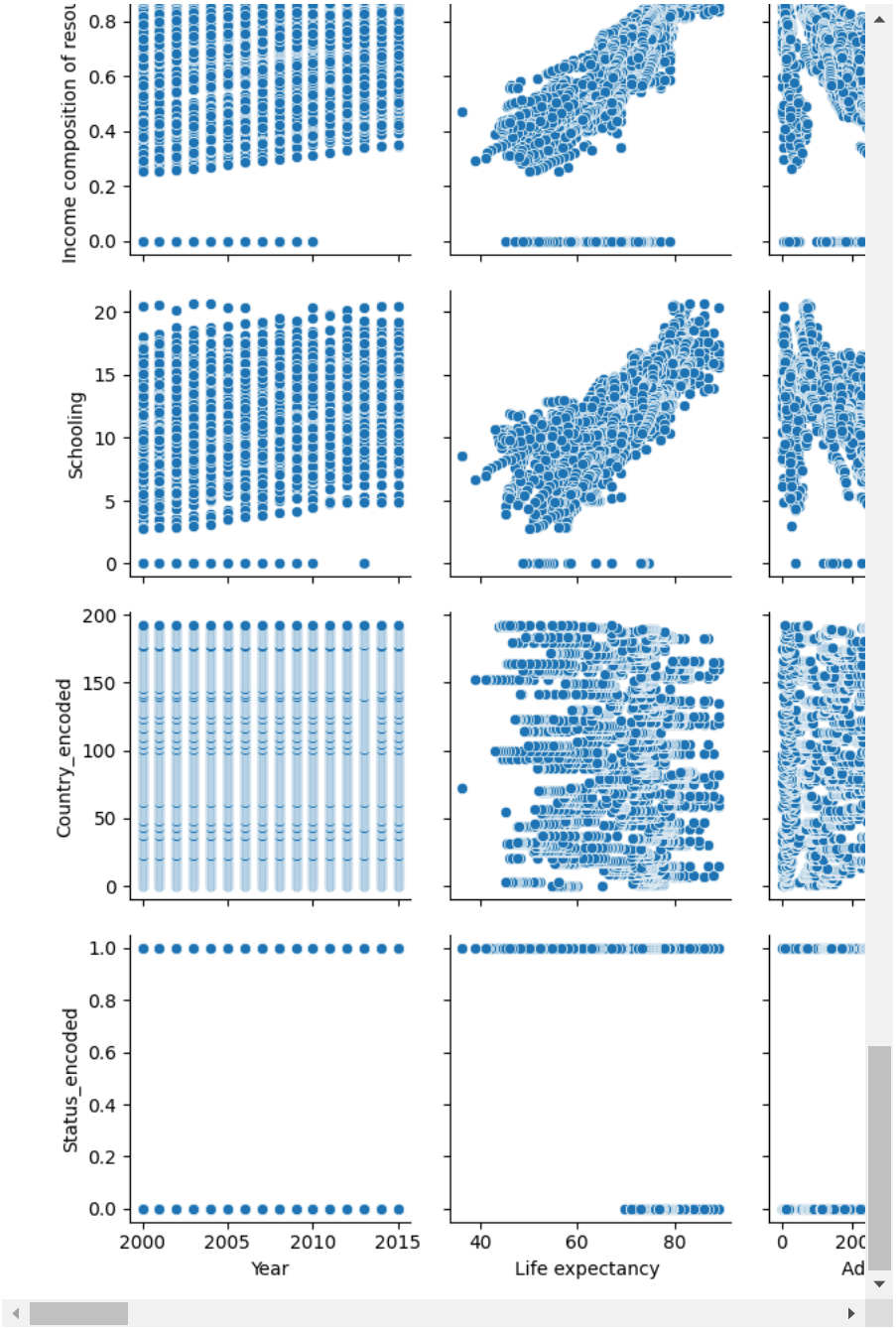
```
fig, ax = plt.subplots(figsize=(25, 10))
sns.heatmap(numeric_df.corr(), annot=True, cmap='Reds')
```

```
<Axes: >
```



Exploratory Data Analysis (EDA)

```
sns.pairplot(numeric_df)
```


Income composition of resou
Schooling
Country_encoded
Status_encoded
Year
Life expectancy
Ad