ASSIGNMENT SIMPLE LINEAR REGRESSION

Name = Ankita Pandey PRN = 23060641070 Subject = Linear Model

PROBLEM STATEMENT =

The project analyzes factors affecting life expectancy from 2000 to 2015 for 193 countries. It includes health, economic, and social factors like immunization rates, mortality rates, GDP, and education levels. Handling missing data resulted in a dataset of 22 columns and 2938 rows. Key questions explore the impact of predictors on life expectancy, healthcare spending, and correlations with lifestyle factors. Inspired by uncovering these relationships, the study employs regression modeling to predict life expectancy, incorporating variables such as adult mortality rates. This approach offers insights for policymakers to enhance population health outcomes.

LINK = https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Code = IN PYTHON

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

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

	Country	Year	Status	Life expectancy	Adult Mortality
0	Afghanistan	2015	Developing	65.0	263.0
1	Afghanistan	2014	Developing	59.9	271.0
2	Afghanistan	2013	Developing	59.9	268.0
3	Afghanistan	2012	Developing	59.5	272.0
4	Afghanistan	2011	Developing	59.2	275.0
...	...	...	...	...	...
2933	Zimbabwe	2004	Developing	44.3	723.0
2934	Zimbabwe	2003	Developing	44.5	715.0
2935	Zimbabwe	2002	Developing	44.8	73.0
2936	Zimbabwe	2001	Developing	45.3	686.0

```
2937     Zimbabwe  2000  Developing              46.0            665.0


      infant deaths  Alcohol  percentage expenditure  Hepatitis B
Measles   \
0                62     0.01               71.279624         65.0
1154
1                64     0.01               73.523582         62.0
492
2                66     0.01               73.219243         64.0
430
3                69     0.01               78.184215         67.0
2787
4                71     0.01                7.097109         68.0
3013
...             ...      ...                     ...          ...
...
2933             27     4.36                0.000000         68.0
31
2934             26     4.06                0.000000          7.0
998
2935             25     4.43                0.000000         73.0
304
2936             25     1.72                0.000000         76.0
529
2937             24     1.68                0.000000         79.0
1483

      ...  Polio  Total expenditure  Diphtheria    HIV/AIDS
GDP  \
0     ...    6.0               8.16         65.0        0.1
584.259210
1     ...   58.0               8.18         62.0        0.1
612.696514
2     ...   62.0               8.13         64.0        0.1
631.744976
3     ...   67.0               8.52         67.0        0.1
669.959000
4     ...   68.0               7.87         68.0        0.1
63.537231
...   ...    ...                ...          ...        ...          ..
.
2933  ...   67.0               7.13         65.0       33.6
454.366654
2934  ...    7.0               6.52         68.0       36.7
453.351155
2935  ...   73.0               6.53         71.0       39.8
57.348340
2936  ...   76.0               6.16         75.0       42.1
548.587312
```

```
2937  ...   78.0                7.10          78.0       43.5
547.358878

      Population   thinness  1-19 years   thinness 5-9 years  \
0     33736494.0                   17.2                 17.3
1       327582.0                   17.5                 17.5
2     31731688.0                   17.7                 17.7
3      3696958.0                   17.9                 18.0
4      2978599.0                   18.2                 18.2
...          ...                    ...                  ...
2933  12777511.0                    9.4                  9.4
2934  12633897.0                    9.8                  9.9
2935    125525.0                    1.2                  1.3
2936  12366165.0                    1.6                  1.7
2937  12222251.0                   11.0                 11.2

      Income composition of resources  Schooling
0                                0.479       10.1
1                                0.476       10.0
2                                0.470        9.9
3                                0.463        9.8
4                                0.454        9.5
...                                ...        ...
2933                             0.407        9.2
2934                             0.418        9.5
2935                             0.427       10.0
2936                             0.427        9.8
2937                             0.434        9.8

[2938 rows x 22 columns]

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')

data = df[['Country','Adult Mortality','Life expectancy
']].set_index('Country')
data

             Adult Mortality  Life expectancy
Country
```

```
Afghanistan             263.0               65.0
Afghanistan             271.0               59.9
Afghanistan             268.0               59.9
Afghanistan             272.0               59.5
Afghanistan             275.0               59.2
...                       ...                ...
Zimbabwe                723.0               44.3
Zimbabwe                715.0               44.5
Zimbabwe                 73.0               44.8
Zimbabwe                686.0               45.3
Zimbabwe                665.0               46.0

[2938 rows x 2 columns]
```

```
data.columns
```

```
Index(['Adult Mortality', 'Life expectancy '], dtype='object')
```

```
data.shape
```

```
(2938, 2)
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2938 entries, Afghanistan to Zimbabwe
Data columns (total 2 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   Adult Mortality   2928 non-null   float64
 1   Life expectancy   2928 non-null   float64
dtypes: float64(2)
memory usage: 68.9+ KB
```

```
data.describe()
```

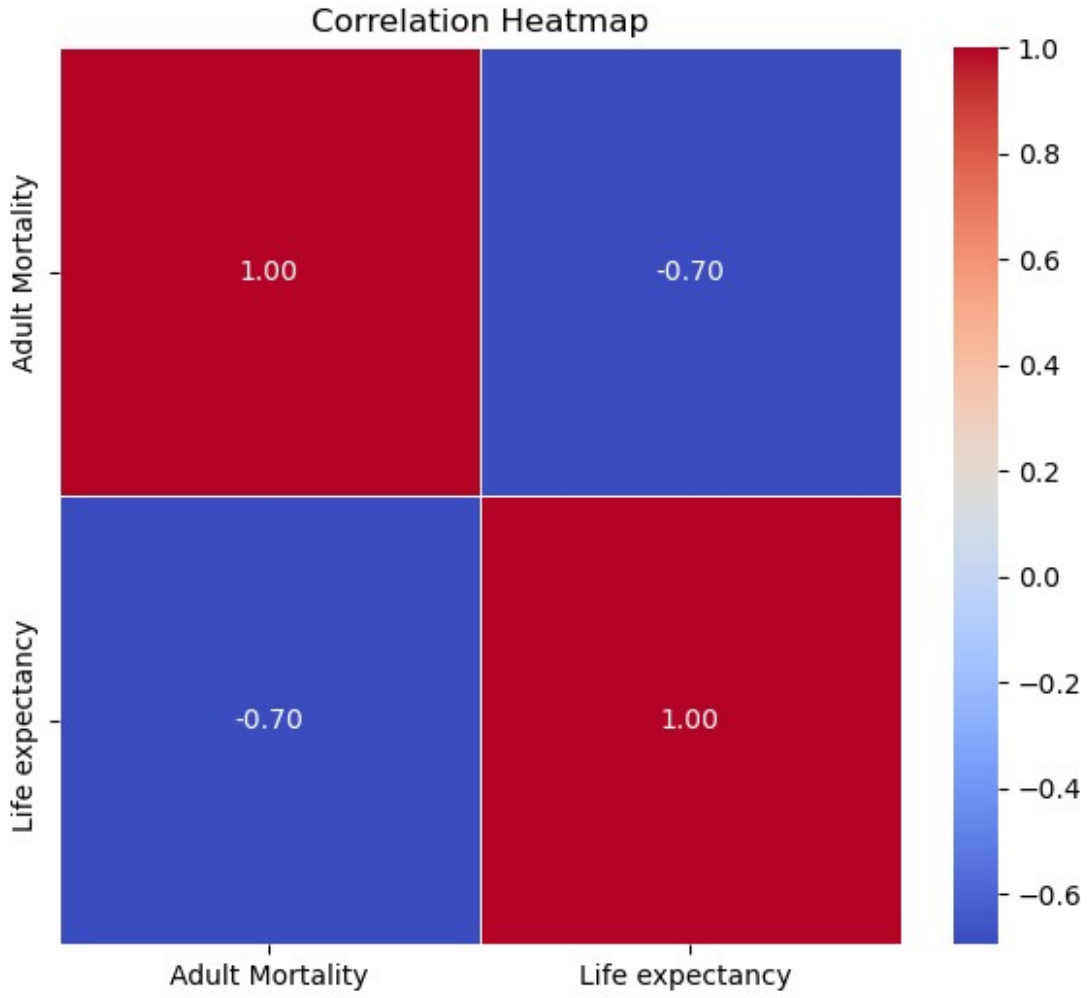
```
       Adult Mortality  Life expectancy
count      2928.000000      2928.000000
mean        164.796448        69.224932
std         124.292079         9.523867
min           1.000000        36.300000
25%          74.000000        63.100000
50%         144.000000        72.100000
75%         228.000000        75.700000
max         723.000000        89.000000
```

```
data.isnull().sum()
```

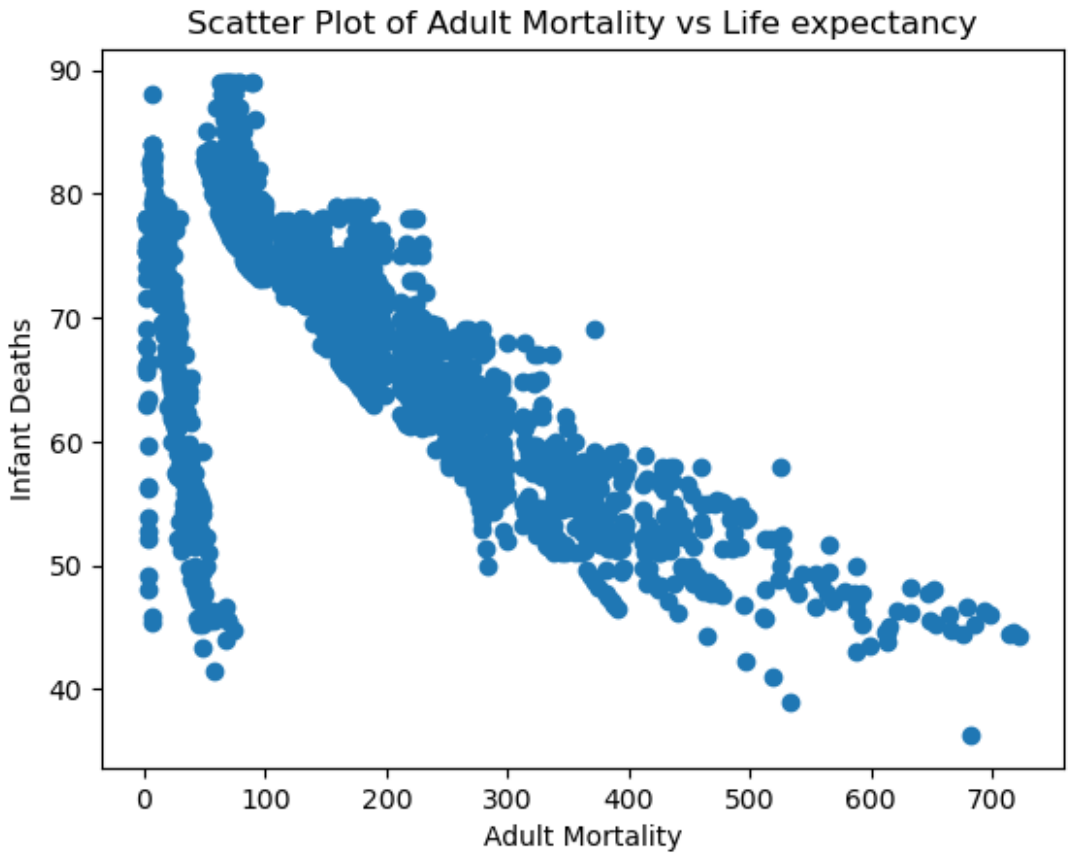
```
Adult Mortality     10
Life expectancy     10
dtype: int64
```

```
datas = data.corr()

plt.figure(figsize=(7, 6))
sns.heatmap(datas, annot=True, cmap='coolwarm', fmt=".2f",
linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()
```




```
#plt.scatter(data['Adult Mortality'],data['Life expectancy'])
plt.scatter(data['Adult Mortality'],data['Life expectancy '])
plt.xlabel('Adult Mortality')
plt.ylabel('Infant Deaths')
plt.title('Scatter Plot of Adult Mortality vs Life expectancy')
plt.show()
```




```
df = data.fillna(data.mean(),inplace=True)

data.isnull().sum()
```

```
Adult Mortality     0
Life expectancy     0
dtype: int64
```

```
data.head()
```

Country	Adult Mortality	Life expectancy
Afghanistan	263.0	65.0
Afghanistan	271.0	59.9
Afghanistan	268.0	59.9
Afghanistan	272.0	59.5
Afghanistan	275.0	59.2

```
X = data.iloc[:,0:1]
Y = data.iloc[:,-1]

X
```

```
             Adult Mortality
Country
Afghanistan            263.0
Afghanistan            271.0
Afghanistan            268.0
Afghanistan            272.0
Afghanistan            275.0
...                      ...
Zimbabwe               723.0
Zimbabwe               715.0
Zimbabwe                73.0
Zimbabwe               686.0
Zimbabwe               665.0

[2938 rows x 1 columns]
```

Y

```
Country
Afghanistan    65.0
Afghanistan    59.9
Afghanistan    59.9
Afghanistan    59.5
Afghanistan    59.2
               ...
Zimbabwe       44.3
Zimbabwe       44.5
Zimbabwe       44.8
Zimbabwe       45.3
Zimbabwe       46.0
Name: Life expectancy , Length: 2938, dtype: float64
```

```
from sklearn.model_selection import train_test_split,cross_val_score

X_train, X_test, Y_train, Y_test = train_test_split(X, Y,
test_size=0.2, random_state=2)

from sklearn.linear_model import LinearRegression

model_lr = LinearRegression()

model_lr.fit(X_train,Y_train)

LinearRegression()
```

X_test

```
                     Adult Mortality
Country
Malawi                         491.0
Philippines                    219.0
```

```
Cabo Verde                      126.0
Comoros                         241.0
Tajikistan                      194.0
...                               ...
Sudan                           251.0
Poland                          144.0
Russian Federation              242.0
Haiti                           259.0
Lebanon                           1.0

[588 rows x 1 columns]
```

Y_test

```
Country
Malawi                 51.5
Philippines            68.1
Cabo Verde             72.3
Comoros                62.2
Tajikistan             65.5
                       ...
Sudan                  61.8
Poland                 75.0
Russian Federation     69.4
Haiti                  62.3
Lebanon                74.1
Name: Life expectancy , Length: 588, dtype: float64
```

```
from sklearn.metrics import mean_squared_error, r2_score
import scipy.stats as stats

# Create a pipeline with feature scaling and linear regression
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
pipeline = make_pipeline(StandardScaler(), LinearRegression())
pipeline.fit(X_train,Y_train)
```

```
Pipeline(steps=[('standardscaler', StandardScaler()),
                ('linearregression', LinearRegression())])
```

```
# Perform k-fold cross-validation (k=5)
cv_scores = cross_val_score(pipeline, X_train, Y_train, cv=5,
scoring='r2')
print("Cross-Validation R-squared Scores:", cv_scores)
print("Mean Cross-Validation R-squared:", cv_scores.mean())
```

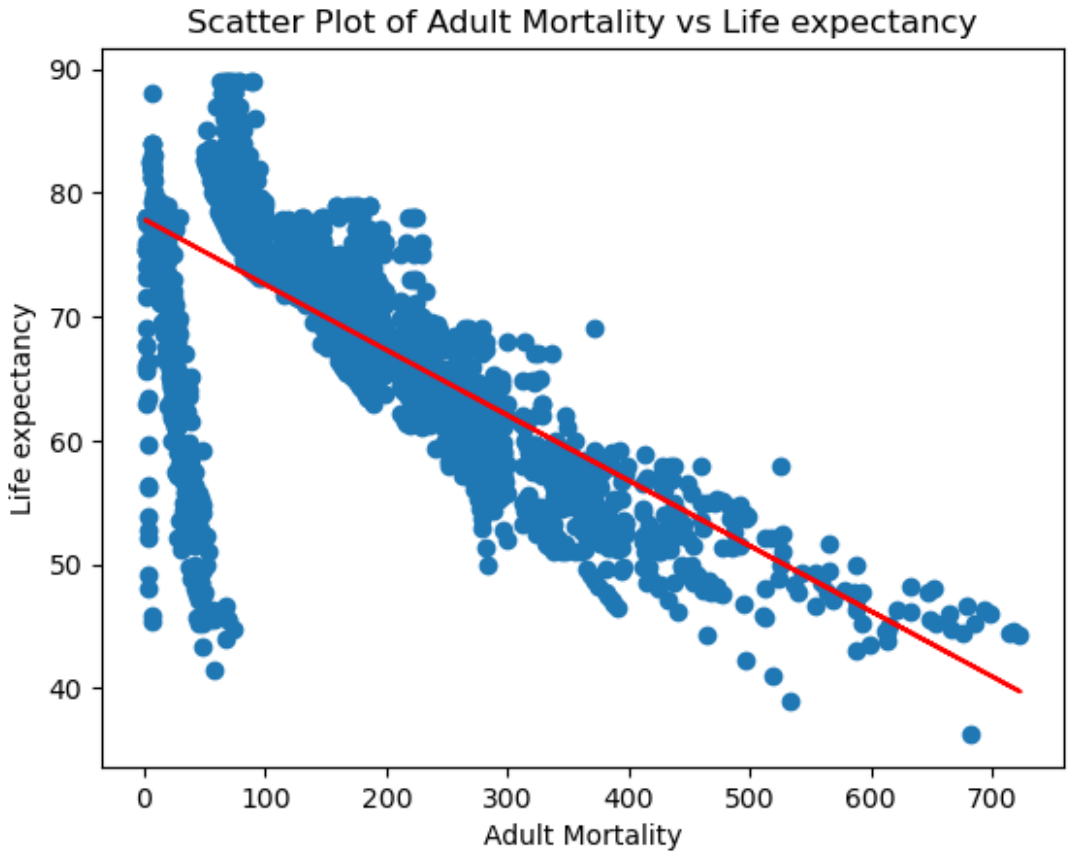
```
Cross-Validation R-squared Scores: [0.40672874 0.43924603 0.46742553
0.52833469 0.53476606]
Mean Cross-Validation R-squared: 0.47530021232734143
```

```
Y_pred = pipeline.predict(X_test)
```

```
# Model Evaluation
mse = mean_squared_error(Y_test, Y_pred)
r2 = r2_score(Y_test, Y_pred)
print("Mean Squared Error:", mse)
print("R-squared:", r2)
```

```
Mean Squared Error: 48.550861043179076
R-squared: 0.5059610190750128
```

```
plt.scatter(data['Adult Mortality'], data['Life expectancy '])
plt.plot(data['Adult Mortality'], model_lr.predict(data[['Adult
Mortality']]), color='red')
plt.xlabel('Adult Mortality')
plt.ylabel('Life expectancy')
plt.title('Scatter Plot of Adult Mortality vs Life expectancy')
plt.show()
```




```
m = model_lr.coef_
m
```

```
array([-0.05273425])
```

```
b = model_lr.intercept_
b
```

```
77.84967190237839
```

```
                                    Life Expectancy Prediction
Documentation
DESCRIPTION


Introduction:
        This documentation presents a comprehensive overview of
predicting life expectancy using linear regression modeling
based on the provided dataset. It encompasses data preprocessing,
model training, evaluation, and analysis of the results.



Data Exploration and Preprocessing:
    The dataset was loaded from the provided CSV file using pandas,
and its structure was examined using various methods
such as df.head(), df.columns, df.info(), and df.describe(). Missing
values were observed in the 'Adult Mortality' and
'Life expectancy' columns, which were then handled by replacing them
with the mean of their
respective columns using data.fillna(data.mean(), inplace=True).
where,
Data visualization techniques, including heatmaps and scatter plots,
were employed to visualize the correlation
between 'Adult Mortality' and 'Life expectancy'.
Data visualization techniques, such as heatmaps and scatter plots,
were used to visually assess the correlation
between 'Adult Mortality' and 'Life expectancy'.

Model Training and Evaluation:
              The dataset was split into training and testing sets
using train_test_split, and a linear regression model
was trained using LinearRegression() from scikit-learn.
Cross-validation was utilized to assess the model's performance with a
5-fold approach.
The model's performance was evaluated on the testing set using mean
squared error (MSE) and R-squared metrics.
The linear regression model revealed a negative relationship between
'Adult Mortality' and 'Life expectancy', with a coefficient of
approximately -0.0527. This suggests that as adult mortality
increases, life expectancy tends to decrease.

The intercept of the model was approximately 77.85, representing the
estimated value of 'Life expectancy' when
'Adult Mortality' is zero.
```

These findings underscore the significant impact of adult mortality on life expectancy, highlighting the importance of
addressing factors contributing to adult mortality to improve population health outcomes and life expectancy.

In conclusion, the linear regression analysis provides valuable insights into the relationship between adult mortality
and life expectancy, emphasizing the need for targeted interventions to reduce adult mortality and enhance life expectancy
in populations.

COFFECIENT = array([-0.05273425])
INTERCEPT = 77.84967190237839

Results:
    The trained model achieved an average cross-validation R-squared score of approximately 0.48,
indicating moderate predictive capability. The mean squared error on the testing set was approximately 48.55, suggesting reasonable model accuracy.
The coefficient of 'Adult Mortality' was approximately -0.0527, indicating a negative relationship with 'Life expectancy'.

Conclusion:
        This analysis demonstrates the feasibility of predicting life expectancy using 'Adult Mortality'
as a predictor variable.
The model provides valuable insights into the impact of adult mortality on life expectancy and could be further
refined with additional features and data. It underscores the importance of understanding and leveraging data-driven
approaches to inform public health interventions and policies.