Assignment 5 - Lucas Lobo

Lucas Lobo

2025-02-24

Assignment 5: Lucas Lobo

The following is my submission for assignment 5 in a .qmd file. I will make note when I do each step throughout the document.

(1): was done above.

(2): Loading the dataset.

```
import pandas as pd
import wbgapi as wb

# Define the indicators to download
indicators = {
    'gdp_per_capita': 'NY.GDP.PCAP.CD',
    'gdp_growth_rate': 'NY.GDP.MKTP.KD.ZG',
    'inflation_rate': 'FP.CPI.TOTL.ZG',
    'unemployment_rate': 'SL.UEM.TOTL.ZS',
    'total_population': 'SP.POP.TOTL',
    'life_expectancy': 'SP.DYN.LE00.IN',
    'adult_literacy_rate': 'SE.ADT.LITR.ZS',
    'income_inequality': 'SI.POV.GINI',
    'health_expenditure_gdp_share': 'SH.XPD.CHEX.GD.ZS',
    'measles_immunisation_rate': 'SH.IMM.MEAS',
    'education_expenditure_gdp_share': 'SE.XPD.TOTL.GD.ZS',
    'primary_school_enrolment_rate': 'SE.PRM.ENRR',
    'exports_gdp_share': 'NE.EXP.GNFS.ZS'
}

# Get the list of country codes for the "World" region
country_codes = wb.region.members('WLD')
```

```python
# Download data for countries only in 2022
df = wb.data.DataFrame(indicators.values(), economy=country_codes, time=2022, skipBlanks=True

# Delete the 'economy' column
df = df.drop(columns=['economy'], errors='ignore')

# Create a reversed dictionary mapping indicator codes to names
# Rename the columns and convert all names to lowercase
df.rename(columns=lambda x: {v: k for k, v in indicators.items()}.get(x, x).lower(), inplace=

# Sort 'country' in ascending order
df = df.sort_values('country', ascending=True)

# Reset the index after sorting
df = df.reset_index(drop=True)

# Display the number of rows and columns
print(df.shape)

# Display the first few rows of the data
print(df.head(3))

# Save the data to a CSV file
df.to_csv('wdi.csv', index=False)
```

```
(217, 14)
       country  inflation_rate  exports_gdp_share  gdp_growth_rate  \
0  Afghanistan             NaN          18.380042        -6.240172
1      Albania        6.725203          37.197085         4.826688
2      Algeria        9.265516          30.808979         3.600000

   gdp_per_capita  adult_literacy_rate  primary_school_enrolment_rate  \
0      357.261153                  NaN                            NaN
1     6846.426143                 98.5                      96.371231
2     4961.552577                  NaN                     108.343933

   education_expenditure_gdp_share  measles_immunisation_rate  \
0                              NaN                       56.0
1                         2.744330                       86.0
2                         4.749247                       79.0
```

```
   health_expenditure_gdp_share  income_inequality  unemployment_rate  \
0                           NaN                NaN             14.100
1                           NaN                NaN             10.137
2                           NaN                NaN             12.346

   life_expectancy  total_population
0           62.879        40578842.0
1           76.833         2777689.0
2           77.129        45477389.0
```

(3): Exploratory Data Analysis:

The three items of analysis I will analyze are:

1. Descriptive statistics of the inflation_rate variable.
2. Correlation between unemployment_rate and life_expectancy.
3. An OLS regression of gdp_per_capita explained by life_expectancy, unemployment_rate, and education_expenditure_gdp_share

```python
# (1) Inflation statistics:
print(df['inflation_rate'].describe())
```

```
count    173.000000
mean      12.404067
std       19.467053
min       -6.687321
25%        5.518129
50%        7.930929
75%       11.665567
max      171.205491
Name: inflation_rate, dtype: float64
```

Figure 1

Three main takeaways:

1. The mean inflation rate of 12.404 is about 4 points higher than the median inflation rate of 7.931, perhaps demonstrating that a select few countries have rates of hyper-inflation that skew the distribution. Additionally, while the minimum inflation rate (0th percentile) is -6.687, the 25th percentile is 5.518, which may suggest that very few countries experience deflation, but far more experience inflation.

```
# (2) Correlation
coeff = df[['unemployment_rate', 'life_expectancy']].corr().iloc[0, 1]
print(f"Correlation between Unemployment Rate and Life Expectancy: {coeff:.4f}")
```

```
Correlation between Unemployment Rate and Life Expectancy: -0.2112
```

Figure 2

2. The scatterplot shows a general negative correlation between unemployment rate and life espectancy. However, in terms of sample data, most countries have unemployment rates less than 10 percent, with only 4-5 countries having a rate higher than 25 percent. So, it may be difficult to extrapolate this data outside of a specified interval. The correlation between Unemployment Rate and Life Expectancy is -0.2112. So as unemployment rate increases, life expectancy is expected to decrease. This is confirmed by our line of best fit and confidence interval.
3. The R-squared of our model is 0.440 – pretty high. Around 44% of the variation in gdp per capita can be explained by life expectancy, unemployment rate, and education expenditure by gdp share. Life expectancy is statistically significant at the $alpha = 0.05$ significance level, whereas unemployment rate and education expenditure are not.

(4): Visualizations:

The two visualization I will explore are:

1. A scatterplot between unemployment_rate and life_expectancy (like (2) above).
2. A boxplot of gdp_per_capita.

Figure 1. Scatterplot showing the relationship between unemployment rate and life expectancy. The red line represents a linear trend, and the shaded red area represents a 95% confidence interval of this estimate.

Source: World Development Indicators

Figure 1. Boxplot showing the summary statistics of the gdp_per_capita variable.

Source: World Development Indicators

(5): Table.

I will produce a table that highlights the count, mean, std, min, max, 25, 50, and 75th percentile of each numerical variable.

Table 1. Summary statistics of variables.

(6): Cross-references:

```python
# (3) Regression Analysis:
import statsmodels.api as sm
# Clean and define variables
df_clean = df.dropna(subset=['gdp_per_capita', 'life_expectancy', 'unemployment_rate', 'educa
X = df_clean[['life_expectancy', 'unemployment_rate', 'education_expenditure_gdp_share']]
y = df_clean['gdp_per_capita']
# Constant term:
X = sm.add_constant(X)
# Model + Summary.
model = sm.OLS(y, X).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         gdp_per_capita   R-squared:                       0.440
Model:                            OLS   Adj. R-squared:                  0.425
Method:                 Least Squares   F-statistic:                     30.33
Date:                Tue, 25 Feb 2025   Prob (F-statistic):           1.50e-14
Time:                        14:49:04   Log-Likelihood:                -1343.7
No. Observations:                 120   AIC:                             2695.
Df Residuals:                     116   BIC:                             2706.
Df Model:                           3
Covariance Type:            nonrobust
=====================================================================================
                                      coef    std err          t      P>|t|      [0.025
-------------------------------------------------------------------------------------
const                          -1.244e+05   1.57e+04     -7.949      0.000   -1.55e+05   -9
life_expectancy                 1929.9881    213.162      9.054      0.000    1507.795    23
unemployment_rate               -157.3729    290.267     -0.542      0.589    -732.284     4
education_expenditure_gdp_share  746.3388    952.471      0.784      0.435   -1140.150    26
==============================================================================
Omnibus:                       66.853   Durbin-Watson:                   1.864
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              263.203
Skew:                           2.013   Prob(JB):                     7.02e-58
Kurtosis:                       9.036   Cond. No.                         687.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 3


Unemployment Rate vs. Life Expectancy
Life Expectancy
Unemployment Rate
85
80
75
70
65
60
55
0
5
10
15
20
25
30
35

Figure 4

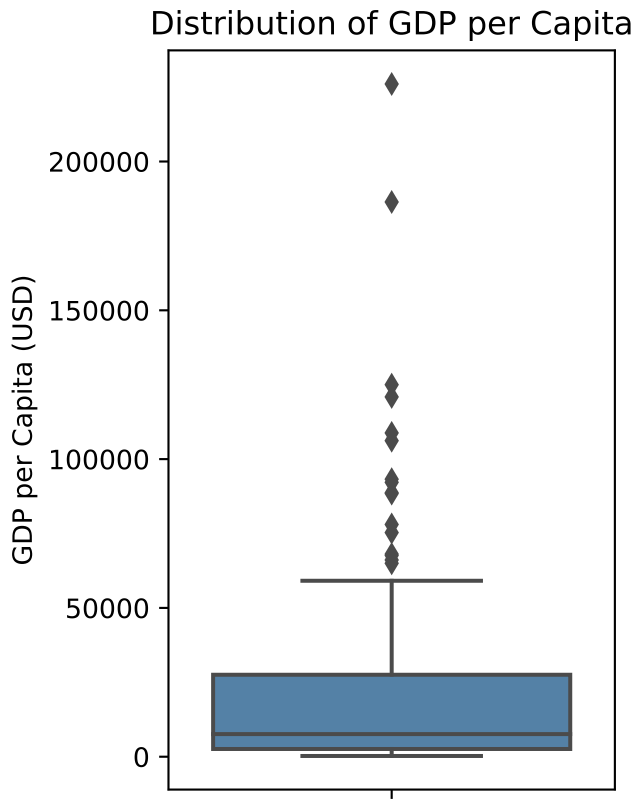

Figure 5

Table 1

Summary Statistics

Stat…	infl…	expo…	gdp_…	gdp_…	adul…	prim…	educ…	meas…	heal…	inco…	unemp…
count	173.…	179.…	206.…	207.…	54.00	156.…	137.…	193.…	20.00	28.00	186.00
mean	12.40	47.63	4.39	2052…	80.97	100.…	4.16	84.10	9.04	38.33	7.23
std	19.47	35.63	6.71	3064…	18.43	12.04	1.77	15.41	2.70	7.72	5.84
min	-6.69	1.57	-28.…	250.…	27.28	67.23	0.35	33.00	5.10	26.40	0.13
25%	5.52	24.36	2.55	2599…	74.76	94.70	2.95	76.00	7.26	32.90	3.48
50%	7.93	40.82	4.21	7606…	85.45	99.84	3.94	90.00	8.93	38.10	5.33
75%	11.67	59.74	6.20	2754…	95.88	104.…	4.96	96.00	10.63	43.12	9.26
max	171.…	211.…	63.33	2260…	100.…	156.…	10.70	99.00	16.57	54.80	35.36

As seen in **Table Table 1**, the count of each of the variables in inconsistent across indicators. This is because of missing data. However, when computing our coefficients for the **Regression Figure 3** and **Correlation Coefficient Figure 2**, empty values are automatically dropped from the dataset. While this may create some incomplete information, it is the only way to obtain these values. It may also explain the high t-score for education expenditure per gdp, since only 137 countries provide that information.

Additionally, our findings from **GDP Boxplot Figure 5** are further expressed in **Table Table 1**, as we see that the mean value is far less than the median/50% percentile, indicating there are far more countries will lower gdp per capita.

In addition to the wdi data, Patrick Hoang-Vu Eozenou and Pirlea (2023) highlights the ways in which new health discoveries in developing countries are helping achieve broader international sustainable development goals. Reports (2024) showcases a ranking of countries' Human Development Indeces: a metric that is similar to WDI data but weights certain factors like quality of life and degrees of oppression more heavily.

(7): Bibliography

---

bibliography: references.bib

Patrick Hoang-Vu Eozenou, Sven Neelsen, and Ana Florina Pirlea. 2023. "Universal Health Coverage as a Sustainable Development Goal." https://datatopics.worldbank.org/world-development-indicators/stories/universal-health-coverage-as-a-sustainable-development-goal.html.

Reports, Human Development. 2024. "Human Development Insights." https://hdr.undp.org/data-center/country-insights#/ranks.