Yang Jiao

Lab 3

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
import statsmodels.formula.api as smf
import scipy.stats as stats
```

```
url = 'https://www.qogdata.pol.gu.se/data/qog_bas_cs_jan24.xlsx'
df = pd.read_excel(url)
```

```
df.head()
```

	ccode	cname	ccode_qog	cname_qog	ccodealp	ccodecow	version	ajr_settmor
0	4	Afghanistan	4	Afghanistan	AFG	700.0	QoGBasCSjan24	4.54009
1	8	Albania	8	Albania	ALB	339.0	QoGBasCSjan24	Na
2	12	Algeria	12	Algeria	DZA	615.0	QoGBasCSjan24	4.35927
3	20	Andorra	20	Andorra	AND	232.0	QoGBasCSjan24	Na
4	24	Angola	24	Angola	AGO	540.0	QoGBasCSjan24	5.63478

5 rows × 337 columns

Q1: Run a simple bivariate regression, and interpret your results. (Did the results fit your expectations? Why? Why not?)

Dependent Variable: gii_gii (UNDP Gender Inequality Index)

```
df[['gii_gii']].describe()
```

	gii_gii
count	170.000000
mean	0.344159
std	0.196390
min	0.013000
25%	0.171250
50%	0.366000
75%	0.503500
max	0.784000

Independent Variable: undp_hdi (UN Human Development Index)

```
df[['undp_hdi']].describe()
```

	undp_hdi
count	189.000000
mean	0.720349
std	0.149286
min	0.386000
25%	0.604000
50%	0.736000
75%	0.830000
max	0.959000

I expect that countires with higher human development index tend to have lower gender inequality index.

mediator:wdi_litradf(=female literacy rate)

```
HDI_GII = smf.ols(formula = 'gii_gii~undp_hdi', data = df, subset=df['wdi_litradf'].no
print (HDI_GII.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                gii_gii   R-squared:                       0.801
Model:                            OLS   Adj. R-squared:                  0.800
Method:                 Least Squares   F-statistic:                     492.5
Date:                Wed, 22 Oct 2025   Prob (F-statistic):           1.19e-44
Time:                        04:02:52   Log-Likelihood:                140.87
No. Observations:                 124   AIC:                            -277.7
Df Residuals:                     122   BIC:                            -272.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
```

```
Intercept      1.1534      0.035     32.979      0.000       1.084       1.223
undp_hdi      -1.1097      0.050    -22.192      0.000      -1.209      -1.011
==============================================================================
Omnibus:                        1.973   Durbin-Watson:                   1.988
Prob(Omnibus):                  0.373   Jarque-Bera (JB):                1.481
Skew:                          -0.030   Prob(JB):                        0.477
Kurtosis:                       2.468   Cond. No.                         10.5
==============================================================================

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

Coefficient for Human Development Index is -1.11, indicating that for every one-unit increase in the Human Development Index, the Gender Inequality Index decreases by approximately 1.11 units. The model is statistically significant with $p < 0.001$. The R-squared of 0.80 suggests that 80% of the variation in gender inequality across countries can be explained by differences in human development, indicating a strong association.

Indeed, the regression suggests a relationship where contries perceived as more human developed tend to have less gender inequality condition.

Q2: Add an additional variable that might mediate or partly "explain" the initial association from that simple regression above -- and explain your results. Did it work out? Yes? No?

```
df[['wdi_litradf']].describe()
```

	wdi_litradf
count	130.000000
mean	81.012996
std	22.346673
min	18.870001
25%	70.012503
50%	91.134998
75%	98.122498
max	100.000000

How do my two independent variables correlate? They correlate meaningfully (+0.84).

```
df_filtered = df[['undp_hdi', 'wdi_litradf']].dropna()
stats.pearsonr(df_filtered['undp_hdi'], df_filtered['wdi_litradf'])
```

```
PearsonRResult(statistic=np.float64(0.8351999304110216),
pvalue=np.float64(1.6144047589117065e-34))
```

```
HDI_GII_Actual = smf.ols(formula='gii_gii~undp_hdi +  wdi_litradf', data=df).fit()
print(HDI_GII_Actual.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 gii_gii   R-squared:                       0.802
Model:                             OLS   Adj. R-squared:                  0.798
Method:                  Least Squares   F-statistic:                     244.7
Date:                 Wed, 22 Oct 2025   Prob (F-statistic):           3.04e-43
Time:                         04:03:37   Log-Likelihood:                140.96
No. Observations:                  124   AIC:                            -275.9
Df Residuals:                      121   BIC:                            -267.5
Df Model:                            2
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.1555      0.035     32.601      0.000       1.085       1.226
undp_hdi      -1.1412      0.092    -12.439      0.000      -1.323      -0.960
wdi_litradf    0.0002      0.001      0.411      0.682      -0.001       0.001
==============================================================================
Omnibus:                        1.517   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.468   Jarque-Bera (JB):                1.260
Skew:                          -0.035   Prob(JB):                        0.533
Kurtosis:                       2.511   Cond. No.                     1.13e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specif
[2] The condition number is large, 1.13e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The Human Development Index (HDI) has a coefficient of –1.141, indicating that, controlling for female literacy rate, a 1-unit increase in HDI (on a 0–1 scale) is associated with a 1.14-point decrease in the Gender Inequality Index (GII)

The coefficient for female literacy rate (wdi_litradf) is 0.0002, but it is not statistically significant (p = 0.68). This indicates that, once human development is accounted for, literacy alone does not explain additional variation in gender inequality across countries. In other words, the positive but nonsignificant coefficient suggests that female literacy is already strongly embedded within the broader HDI measure, which includes education and health components.

The adjusted R^2 of 0.798 is nearly identical to the bivariate model (0.800), showing that adding literacy rate contributes little to improving the explanatory power of the model. This implies that female literacy does not mediate or absorb the strong relationship between human development and gender inequality.

These results suggest that HDI is still the best way to forecast gender inequality, even when education (as measured by literacy rate) is taken into account. Countries with higher human development tend to have far less gender inequality. However, when looking at other factors that affect development, the influence of female literacy seems to be very little.

Q3: More on extreme combinations. Find the top 5 entities that are ranked at the top on one variable and ranked at the bottom on another variable. Interpret your results.

I am going to look to countries that are very high in their gdp , but do not appear to have very much corruption in the country.

```
df['wealth_rank'] = df['wdi_gdpcapcon2015'].rank(ascending=False)
df['corruption_rank'] = df['ti_cpi'].rank(ascending=False)

extreme_comb = df[['cname', 'wealth_rank', 'corruption_rank']].copy()
extreme_comb['rank_difference'] = abs(extreme_comb['wealth_rank'] - extreme_comb['corrup

extreme_sorted = extreme_comb.sort_values(by='rank_difference', ascending=False).head(5)
extreme_sorted
```

	cname	wealth_rank	corruption_rank	rank_difference
144	Rwanda	170.0	50.0	120.0
16	Bhutan	129.0	23.0	106.0
178	Turkmenistan	78.0	166.0	88.0
98	Libya	85.0	172.0	87.0
151	Senegal	154.0	67.5	86.5

Next steps: Generate code with `extreme_sorted` New interactive sheet

These results show that wealth and corruption do not always align. Rwanda and Bhutan have lower GDP per capita but rank well for low corruption, suggesting strong governance despite limited resources. In contrast, Turkmenistan and Libya have higher income but more corruption, indicating that wealth alone doesn't ensure clean governance. Overall, strong institutions matter more than economic prosperity in reducing corruption.