To get started, get an idea of what is meant by "state fragility" by exploring the FFP website.

Next, download the 2020 version of the FSI dataset here. This is the data we will be working with for Questions 1 through 5.

You will also need to consult the codebook for the dataset, which can be found here.

(a) Import the 2020 FSI dataset as a .csv and show the first five rows.

[1]:
```
import pandas as pd
import numpy as np
readFile = pd.read_excel('fsi-2020.xlsx')
readFile.to_csv("fsi-2020.csv", index=None, header=True)
data = pd.DataFrame(pd.read_csv("fsi-2020.csv"))
data.head()
```

[1]:

	Country	Year	Rank	Total	C1: Security Apparatus	C2: Factionalized Elites	C3: Group Grievance	E1: Economy	E2: Economic Inequality	E3: Human Flight and Brain Drain	P1: State Legitimacy	P2: Public Services	P3: Human Rights	S1: Demographic Pressures	S2: Refugees and IDPs
0	Yemen	2020-01-01	1st	112.438694	9.700000	10.0	9.69887	9.400000	7.800000	7.000000	9.889823	9.500000	9.950000	9.800000	9.7
1	Somalia	2020-01-01	2nd	110.888959	9.811328	10.0	8.60000	9.100000	9.367151	8.900000	8.888107	9.100000	9.000000	9.981087	9.1
2	South Sudan	2020-01-01	3rd	110.752190	9.400000	9.7	9.10000	9.500000	9.200000	6.800000	9.944415	9.500000	9.000000	9.450357	9.7
3	Syria	2020-01-01	4th	110.749697	9.900000	9.9	10.00000	8.686367	7.200000	8.413343	9.950000	9.100000	10.000000	7.600000	10.0
4	Congo Democratic Republic	2020-01-01	5th	109.394621	8.500000	9.8	9.70000	8.000000	8.619842	6.900000	9.660971	9.464911	9.548897		

```
4                    9.800000                    10.0

   X1: External Intervention  Change from Previous Year
0                  10.000000                  -1.061306
1                   9.041286                  -1.311041
2                   9.457419                  -1.447810
3                   9.999986                  -0.850303
4                   9.400000                  -0.805379
```

(b) Rename the following columns as follows and show the first five rows of the revised dataset:

- 'Total': 'total'
- 'C3: Group grievance': 'grievance'
- 'E1: Economy': 'economy'

[2]:
```
data.rename(columns = {'Total':'total', 'C3: Group Grievance':'grievance', 'E1:␣
↪Economy':'economy'}, inplace = True)
data.head()
```

[2]:
```
                    Country        Year Rank       total  \
0                     Yemen  2020-01-01  1st  112.438694
1                   Somalia  2020-01-01  2nd  110.888959
2               South Sudan  2020-01-01  3rd  110.752190
3                     Syria  2020-01-01  4th  110.749697
4  Congo Democratic Republic  2020-01-01  5th  109.394621

   C1: Security Apparatus  C2: Factionalized Elites  grievance   economy  \
0                9.700000                      10.0    9.69887  9.400000
1                9.811328                      10.0    8.60000  9.100000
2                9.400000                       9.7    9.10000  9.500000
3                9.900000                       9.9   10.00000  8.686367
4                8.500000                       9.8    9.70000  8.000000

   E2: Economic Inequality  E3: Human Flight and Brain Drain  \
0                 7.800000                          7.000000
1                 9.367151                          8.900000
2                 9.200000                          6.800000
3                 7.200000                          8.413343
4                 8.619842                          6.900000

   P1: State Legitimacy  P2: Public Services  P3: Human Rights  \
0              9.889823             9.500000          9.950000
1              8.888107             9.100000          9.000000
2              9.944415             9.500000          9.000000
3              9.950000             9.100000         10.000000
4              9.660971             9.464911          9.548897
```

```
   S1: Demographic Pressures  S2: Refugees and IDPs  \
0                   9.800000                    9.7
1                   9.981087                    9.1
2                   9.450357                    9.7
3                   7.600000                   10.0
4                   9.800000                   10.0

   X1: External Intervention  Change from Previous Year
0                  10.000000                 -1.061306
1                   9.041286                 -1.311041
2                   9.457419                 -1.447810
3                   9.999986                 -0.850303
4                   9.400000                 -0.805379
```

(c) Generate a correlation matrix for all numeric variables. Describe the correlation between `total` and `economy` in as much detail as possible.

Note: You'll notice that the correlations between ***`total`*** *and the other variables are generally quite high. This is because* ***`total`*** *is a composite of all the other variables. But, we don't know how the creators of the dataset weighted or combined the other variables in order to generate the overall stability score (****`total`****). In this homework, we are going to explore the effects of just a few of these variables on the dependent variable* ***`total`*** *to see how useful they are as standalone explainers and predictors.*

[3]:
```
data.corr()
```

[3]:
```
                                   total  C1: Security Apparatus  \
total                           1.000000                0.887130
C1: Security Apparatus          0.887130                1.000000
C2: Factionalized Elites        0.874747                0.763882
grievance                       0.672733                0.638240
economy                         0.858549                0.722351
E2: Economic Inequality         0.866416                0.726518
E3: Human Flight and Brain Drain  0.779708              0.654861
P1: State Legitimacy            0.856040                0.741336
P2: Public Services             0.904541                0.788740
P3: Human Rights                0.842300                0.759301
S1: Demographic Pressures       0.882506                0.749146
S2: Refugees and IDPs           0.819482                0.720535
X1: External Intervention       0.827935                0.685118
Change from Previous Year       0.112451                0.175215

                                C2: Factionalized Elites  grievance  \
total                                           0.874747   0.672733
C1: Security Apparatus                          0.763882   0.638240
C2: Factionalized Elites                        1.000000   0.701263
grievance                                       0.701263   1.000000
economy                                         0.676032   0.416864
```

E2: Economic Inequality	0.674450	0.438727
E3: Human Flight and Brain Drain	0.583273	0.382782
P1: State Legitimacy	0.867237	0.605498
P2: Public Services	0.669605	0.453755
P3: Human Rights	0.797493	0.608289
S1: Demographic Pressures	0.660440	0.432959
S2: Refugees and IDPs	0.695177	0.656434
X1: External Intervention	0.697836	0.443671
Change from Previous Year	0.030375	0.092232

	economy	E2: Economic Inequality	\
total	0.858549	0.866416	
C1: Security Apparatus	0.722351	0.726518	
C2: Factionalized Elites	0.676032	0.674450	
grievance	0.416864	0.438727	
economy	1.000000	0.767501	
E2: Economic Inequality	0.767501	1.000000	
E3: Human Flight and Brain Drain	0.759385	0.711998	
P1: State Legitimacy	0.656312	0.685664	
P2: Public Services	0.836760	0.890326	
P3: Human Rights	0.598253	0.671398	
S1: Demographic Pressures	0.781763	0.879751	
S2: Refugees and IDPs	0.681460	0.616687	
X1: External Intervention	0.804255	0.666823	
Change from Previous Year	0.122451	0.172509	

	E3: Human Flight and Brain Drain	\
total	0.779708	
C1: Security Apparatus	0.654861	
C2: Factionalized Elites	0.583273	
grievance	0.382782	
economy	0.759385	
E2: Economic Inequality	0.711998	
E3: Human Flight and Brain Drain	1.000000	
P1: State Legitimacy	0.529542	
P2: Public Services	0.767005	
P3: Human Rights	0.513212	
S1: Demographic Pressures	0.721149	
S2: Refugees and IDPs	0.555292	
X1: External Intervention	0.757911	
Change from Previous Year	0.016465	

	P1: State Legitimacy	P2: Public Services	\
total	0.856040	0.904541	
C1: Security Apparatus	0.741336	0.788740	
C2: Factionalized Elites	0.867237	0.669605	
grievance	0.605498	0.453755	

economy	0.656312	0.836760
E2: Economic Inequality	0.685664	0.890326
E3: Human Flight and Brain Drain	0.529542	0.767005
P1: State Legitimacy	1.000000	0.671294
P2: Public Services	0.671294	1.000000
P3: Human Rights	0.892416	0.658843
S1: Demographic Pressures	0.653779	0.932785
S2: Refugees and IDPs	0.620620	0.728401
X1: External Intervention	0.634256	0.719994
Change from Previous Year	0.075583	0.144908

	P3: Human Rights	S1: Demographic Pressures	\
total	0.842300	0.882506	
C1: Security Apparatus	0.759301	0.749146	
C2: Factionalized Elites	0.797493	0.660440	
grievance	0.608289	0.432959	
economy	0.598253	0.781763	
E2: Economic Inequality	0.671398	0.879751	
E3: Human Flight and Brain Drain	0.513212	0.721149	
P1: State Legitimacy	0.892416	0.653779	
P2: Public Services	0.658843	0.932785	
P3: Human Rights	1.000000	0.691350	
S1: Demographic Pressures	0.691350	1.000000	
S2: Refugees and IDPs	0.620215	0.692495	
X1: External Intervention	0.605788	0.682386	
Change from Previous Year	0.065444	0.124930	

	S2: Refugees and IDPs	\
total	0.819482	
C1: Security Apparatus	0.720535	
C2: Factionalized Elites	0.695177	
grievance	0.656434	
economy	0.681460	
E2: Economic Inequality	0.616687	
E3: Human Flight and Brain Drain	0.555292	
P1: State Legitimacy	0.620620	
P2: Public Services	0.728401	
P3: Human Rights	0.620215	
S1: Demographic Pressures	0.692495	
S2: Refugees and IDPs	1.000000	
X1: External Intervention	0.672606	
Change from Previous Year	0.077670	

	X1: External Intervention	\
total	0.827935	
C1: Security Apparatus	0.685118	
C2: Factionalized Elites	0.697836	

```
grievance                                   0.443671
economy                                     0.804255
E2: Economic Inequality                     0.666823
E3: Human Flight and Brain Drain            0.757911
P1: State Legitimacy                        0.634256
P2: Public Services                         0.719994
P3: Human Rights                            0.605788
S1: Demographic Pressures                   0.682386
S2: Refugees and IDPs                       0.672606
X1: External Intervention                   1.000000
Change from Previous Year                   0.038646
```

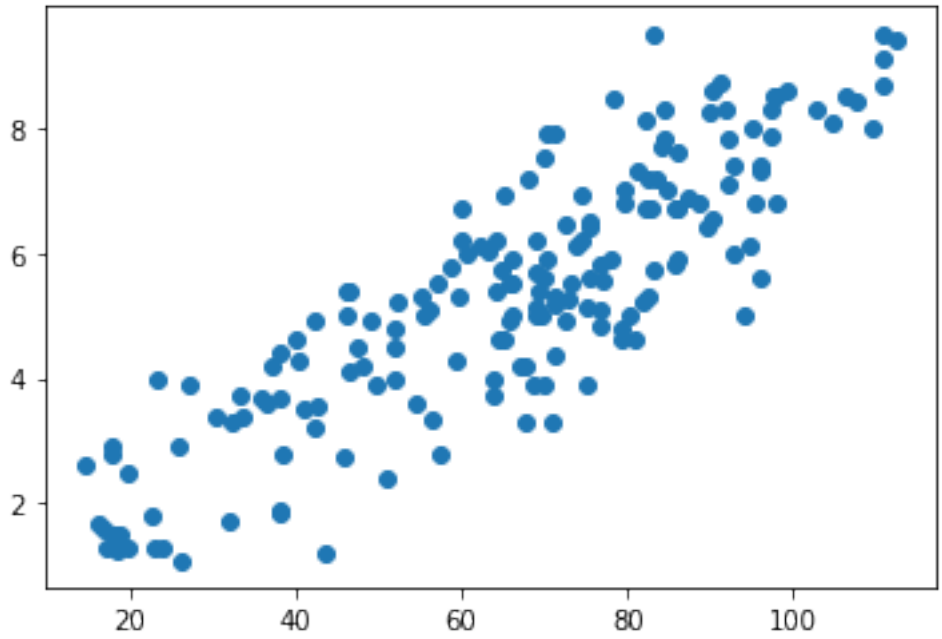
```
                                    Change from Previous Year
total                                               0.112451
C1: Security Apparatus                              0.175215
C2: Factionalized Elites                            0.030375
grievance                                           0.092232
economy                                             0.122451
E2: Economic Inequality                             0.172509
E3: Human Flight and Brain Drain                    0.016465
P1: State Legitimacy                                0.075583
P2: Public Services                                 0.144908
P3: Human Rights                                    0.065444
S1: Demographic Pressures                           0.124930
S2: Refugees and IDPs                               0.077670
X1: External Intervention                           0.038646
Change from Previous Year                           1.000000
```

[ ]:

(d) Generate a scatterplot between `total` and `economy`, where `total` is on the y-axis. What do you observe about the relationship between these two variables? Be as specific as possible.

[4]:
```
import matplotlib.pyplot as plt
plt.scatter(data.total, data.economy)
```

[4]:
```
<matplotlib.collections.PathCollection at 0x7f2cf0b31450>
```



The relationship between total and economy is positiveley correlated.

(e) We are going to conduct a simple linear regression to evaluate the effect of `economy` on `total`. What are our alternative and null hypotheses for a two-tailed test?

The null hypothesis is that economy would have a great effect on the total. The alternative hypothesis is that

(f) Conduct a simple OLS regression with the model from Question 1(e) using the statistical approach and present a table of the results.

[7]:
```
import statsmodels.api as sm
import statsmodels.formula.api as smf
results = smf.ols('total ~ economy', data=data).fit()
results.summary()
```

[7]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  total   R-squared:                       0.737
Model:                            OLS   Adj. R-squared:                  0.736
Method:                 Least Squares   F-statistic:                     493.5
Date:                Wed, 02 Dec 2020   Prob (F-statistic):           6.09e-53
Time:                        13:19:07   Log-Likelihood:                -700.39
No. Observations:                 178   AIC:                             1405.
Df Residuals:                     176   BIC:                             1411.
```

```
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     10.8796      2.648      4.108      0.000       5.654      16.106
economy       10.2765      0.463     22.214      0.000       9.364      11.189
==============================================================================
Omnibus:                        4.777   Durbin-Watson:                   1.543
Prob(Omnibus):                  0.092   Jarque-Bera (JB):                2.791
Skew:                           0.046   Prob(JB):                        0.248
Kurtosis:                       2.394   Cond. No.                         16.7
==============================================================================

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

(g) Write out the complete and correct interpretation of the intercept, including what the relevant variable(s) stand for, not just the variable names.

[ ]:

(h) Write out the complete and correct interpretation of the coefficient on `economy`, including what the relevant variable(s) stand for, not just the variable names.

[ ]:

(i) Write out the complete and correct interpretation of the p-value for `economy`, including what the relevant variable(s) stand for, not just the variable names.

[ ]:

(j) Write out the complete and correct interpretation of the confidence interval for `economy`, including what the relevant variable(s) stand for, not just the variable names.

[ ]:

(k) Write out the complete and correct interpretation of R-squared, including what the relevant variable(s) stand for, not just the variable names.

[ ]:

3 Question 2

(a) We will now conduct a multiple regression using OLS. We are going to add the variable `grievance` as a second independent variable. State the null and alternative hypotheses for this

new variable for a two-tailed test.

[ ]:

(b) Conduct a multiple OLS regression with the model from Question 2(a) using the statistical approach and present a table of the results.

[ ]:

(c) Write out the complete and correct interpretation of the coefficient on `grievance`, including what the relevant variable(s) stand for, not just the variable names.

[ ]:

(d) Is this model an improvement over the simple linear regression from Question 1? How did you come to this conclusion?

[ ]:

(e) Can we reject or fail to reject the null hypothesis for new variable? How do you know? Be as thorough as possible in your reasoning.

[ ]:

4 Question 3

(a) We're going to conduct a multiple linear regression through machine learning, using the exact same multiple regression model from Question 2. First, set up your X and y arrays using the variable names in your setup syntax.

[ ]:

(b) Divide your X and y into training and test sets. Make the test set 25% of your data, and set the `random_state` to 4.

[ ]:

(c) Train the algorithm.

[ ]:

(d) Print the intercept and coefficients and interpret them accordingly, including what the relevant variable(s) stand for, not just the variable names.

[ ]:

(e) Generate a table of the first 10 values of your actual vs. predicted values. How does it look like you did based on this?

[ ]:

(f) Generate the R-squared and interpret it.

[ ]:

(g) Remember that when we have more than one independent variable, we use the adjusted R-squared instead. Write code that will generate the adjusted R-squared (may require some research!) and interpret it.

[ ]:

(h) Compare your overall ML output with your statistical approach to multiple OLS in Question 2. How do the results compare?

[ ]:

(i) Go back and re-run this ML regression analysis, but this time with the test set as 90% of your data (keep the `random_state` the same). How does this influence your results, and why?

Note: You don't need to re-type all your code, just run the analysis, but ***make sure to change your code back to your original answer for Question 3(b)*** *before turning it in!*

[ ]:

5 Question 4

(a) We're going to conduct a k-NN classification. To do this, we need a categorical dependent variable. Create a new column called `total_cat` that takes the value "fragile" if the `total` score is greater than the median for `total`, and the value "stable" if the `total` score is less than or equal to the median for `total`. Show the first five rows of your updated data frame.

[ ]:

(b) Create your X and y arrays using the .iloc syntax in your setup.

[ ]:

(c) Divide your X and y into training and test sets. Make the test set 25% of your data and set the `random_state` to 4.

[ ]:

(d) Train your algorithm.

[ ]:

(e) Generate a table of the first 10 actual vs. predicted values. How does it look like you did?

[ ]:

(f) Print the confusion matrix and classification report. Interpret the confusion matrix.

[ ]:

(g) Create a plot of the mean error of the model for different values of k between 1 and 50.

[ ]:

(h) How does the model perform as k increases?

[ ]:

6 Question 5

(a) Looking back on our work across four models, how confident are you that we've explained state fragility with these two independent variables, and why?

[ ]:

(b) Is there a risk of endogeneity in our research? If yes, please provide examples for each relevant independent variable.

[ ]:

(c) Is there a risk of confounders? If yes, please provide examples for each relevant independent variable.

[ ]:

(d) How confident are you in terms of how the variables we've worked with are measured? Please describe your answer in terms of conceptualization and operationalization.

[ ]:

(e) What is a next step you might recommend to improve research on this topic?

[ ]:

7 Question 6

Read the article "From Tuskegee to a COVID Vaccine: Diversity and Racism Are Hurdles in Drug Trials" ("drug_trials.pdf") and answer the questions that follow.

(a) While trials for the effectiveness of a Covid vaccine involve more diverse participant pools than previous efforts, Prof Jackson (at MGH) notes that they have some way to go before they reach which benchmark? That is, what does Jackson want the trial subjects to be representative of?

[ ]:

(b) Which of the four ethical principles discussed in class does Jackson's concern speak to most? Briefly explain your answer.