Description:

```
[3] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import statsmodels.api as sm
    import scipy.stats as st
    import seaborn as sns
    import statsmodels.api as sm
    import matplotlib.pyplot as plt
    from sklearn.model_selection import train_test_split
    from statsmodels.graphics.regressionplots import plot_ccpr
    from sklearn.metrics import classification_report
    from sklearn.metrics import mean_squared_error
    df = pd.read_csv("Gender_Inequality_Index.csv")# upload data
```

This data set compares each country's Human Development, Gender Inequality Index (GII), Country Rank, percentage of maternal mortality, percentage of adolescent birth rate, percentage of seats in parliament held by women, percentage of males with at least some secondary education (ages 25 and older), percentage of females with at least some secondary education (ages 25 and older), and percentage of female in the labor force (ages 15 and older). The data set includes over 190 countries for the year 2021.

Furthermore, the Gender Inequality Index (GII) is defined by the United Nations Development Programme as a composite measure reflecting health, empowerment, and participation in the labor market. The United Nations measured the health dimension by maternal mortality ratio and the adolescent fertility rate. They also measured empowerment by how many seats in parliament were held by each gender and by how many people in each gender received secondary education. Furthermore, the participation in the labor market dimension was measured by women's participation in the labor market. GII varies between 0 and 1. 0 being that men and women are fairly equal and 1 is that men or women are poorly compared to the other gender in all situations. So the higher the GII, the greater the disparity is between the genders.

This is the data set we used: https://www.kaggle.com/datasets/gianinamariapetrascu/gender-inequality-index. For question 2 we are joining the GII dataset with a population data set to get the population of each country. This data set was taken from this https://www.kaggle.com/datasets/rsrishav/world-population. This data set includes 224 rows and 9 columns (features). These features include an ISO Code (internationally recognized code for each country, Country Name, the last updated population in 2021, the last updated population in 2020, the area in km, density per square km, growth rate, world population percentage, and country rank. For question 4 we are using this dataset: https://ourworldindata.org/grapher/gender-inequality-index-from-the-human-development-report.

```
df = pd.read_csv("Gender_Inequality_Index.csv")
df.head()
```

	Country	Human_development	GII	Rank	Maternal_mortality	Adolescent_birth_rate	Seats_parliament	F_secondary_educ	M_secondary_educ	F_Labour_force	M_Labour_force
0	Switzerland	Very high	0.018	3.0	5.0	2.2	39.8	96.9	97.5	61.7	72.7
1	Norway	Very high	0.016	2.0	2.0	2.3	45.0	99.1	99.3	60.3	72.0
2	Iceland	Very high	0.043	8.0	4.0	5.4	47.6	99.8	99.7	61.7	70.5
3	Hong Kong	Very high	NaN	NaN	NaN	1.6	NaN	77.1	83.4	53.5	65.8
4	Australia	Very high	0.073	19.0	6.0	8.1	37.9	94.6	94.4	61.1	70.5

```
[ ] #data set : https://www.kaggle.com/datasets/rsrishav/world-population
    df1= pd.read_csv("2021_population.csv")
    df1= df1.rename(columns={"country":"Country"})
    df2= df1[["Country","2021_last_updated"]]
    df3 =pd.merge(left=df, right=df2,left_on='Country',right_on='Country')
    df3["2021_last_updated"] = df3["2021_last_updated"].str.replace(',', '')
    df3["2021_last_updated"]  = df3["2021_last_updated"].astype({"2021_last_updated": 'int32'}).round(-5)
    df3.head()
```

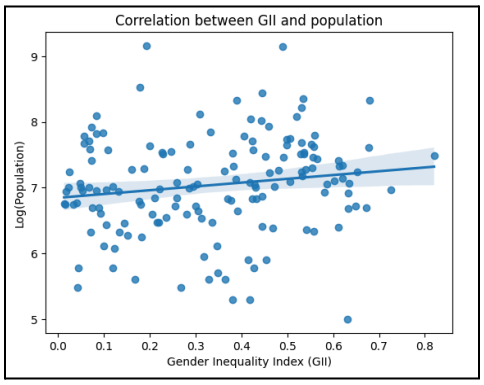
	Country	Human_development	GII	Rank	Maternal_mortality	Adolescent_birth_rate	Seats_parliament	F_secondary_educ	M_secondary_educ	F_Labour_force	M_Labour_force	2021_last_updated
0	Switzerland	Very high	0.018	3.0	5.0	2.2	39.8	96.9	97.5	61.7	72.7	8800000
1	Norway	Very high	0.016	2.0	2.0	2.3	45.0	99.1	99.3	60.3	72.0	5500000
2	Iceland	Very high	0.043	8.0	4.0	5.4	47.6	99.8	99.7	61.7	70.5	300000
3	Hong Kong	Very high	NaN	NaN	NaN	1.6	NaN	77.1	83.4	53.5	65.8	7600000
4	Australia	Very high	0.073	19.0	6.0	8.1	37.9	94.6	94.4	61.1	70.5	26000000

This data set contains 181 countries/regions, but one must note that there are 193 countries recognized by the United Nations. The data set above gives time series data from 1990-2021. It is important to note that while most of the countries contain data from 1990, a rare amount do not. For example, Afghanistan began collecting Gender Inequality Index from the year 2005. There are 5205 rows and 4 columns in the dataset. The rows include 'Entity, 'Code', 'Year', and 'Gender Inequality Index'. The Entity is the country or region being represented. The Year references the year (between 1990 and 2021). The Code is the country/region code. The last column, Gender Inequality Index, was explained above.

The overall goal of this data analysis is to test the effects of different factors on the Gender Inequality Index. The factors that we are including are the Adolescent Birth Rate, whether or not a country is considered developed, the proportion of males with secondary education, and policies that put more women in leadership roles. The target group of people that care about the answer to these questions are lawmakers, healthcare policymakers, those who work in education, activist groups, and organizations whose main goal is creating gender equality. These groups care about the answers to our questions because we tested a policy that the United Nations put into place to establish more female leadership in the peacemaking process. Furthermore, lawmakers are actively trying to have a greater gender equality impact with the laws and policies that they are making. Healthcare policymakers care about these questions because we are testing the adolescent birth rate and how it contributes to gender inequality. Furthermore, those who work in education care about our question because we see the effect that different proportions of males in secondary education are related to the gender inequality index. In addition, activist groups and organizations looking to improve gender equality care about our questions because we are testing factors that they can show support for if we show a significant difference in gender equality.

Questions:

1. What Is the Correlation between GII and Population?
2. How do all features of our dataset correlate with the GII, and which features are most significant?
3. Can GII effectively predict the classification of a country as part of the developed world?
4. How does GII correlate to adolescent birth rate across countries?
5. How does GII correlate with the percentage of males with at least some secondary education?
6. What kind of effect do worldwide policies that increase the number of women in the peacemaking process have on GII?



Question 1: What Is the Correlation between GII and Population? On the X-axis, Gender Inequality Index (GII) ranges from 0 to 1, and on the Y-axis the log10 of the Population. Log10 was used for display purposes, since there are some very small and large countries listed in the data, this would normally cause the data to be flattened together. Each point on the scatter plot signifies a country.

A regression line was added to show how the population might correlate with the GII, it shows there may be a positive correlation.

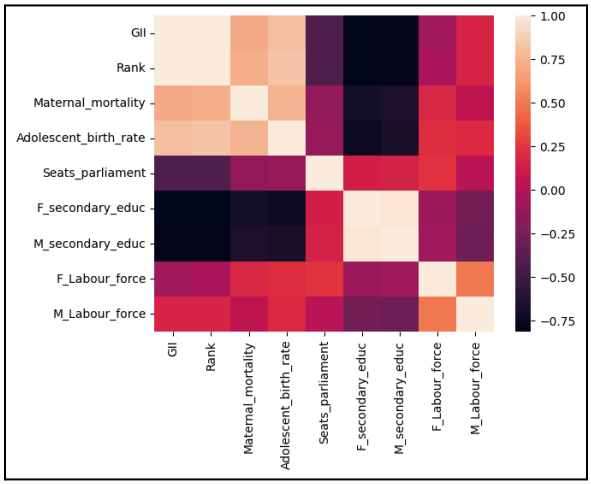
Code for graph:

```
x = df3["GII"]
y = np.log10(df3["2021_last_updated"])

#create basic scatterplot
sns.regplot(x=x, y=y)
plt.title('Correlation between GII and population')
plt.xlabel('Gender Inequality Index (GII)')
plt.ylabel('Log(Population)')
```

Question 2: How do all features of our dataset correlate with the GII, and which features are most significant?

```
qmore = df.drop(columns=["Country", "Human_development"])
sns.heatmap(qmore.corr())
```



The key to the right shows us that the lighter colors mean that there is a stronger correlation. In creating this heat map we dropped the columns Country and Human_development. We felt as though those two columns would not add as much valuable information to the graph. Furthermore, we found that GII is most closely correlated with Rank, Maternal_Mortality, and Adolescent_birth_rate. Rank and GII have a Pearson's correlation coefficient of around 1.0. Maternal_Mortality and Adolescent_birth_rate with GII have a Pearson's correlation coefficient of around 0.75, with Adolescent_birth_rate being slightly higher. This shows that our findings in question 1 are correct. In addition, our analysis from question 2 is consistent with our results here. We can see that GII and Male Secondary Education have a negative correlation with a Pearson's Correlation Coefficient of about -0.75. Some other features that are greatly negatively correlated are GII and Female Secondary Education, Rank and Female Secondary Education, Rank and Male Secondary Education, Maternal Mortality with Female Secondary Education and Male Secondary Education, and Adolescent Birth Rate with Female Secondary Education and Male Secondary Education.

Question 3: Can GII effectively predict the classification of a country as part of the developed world?

The concept of the developed world is often associated with a higher standard of living and opportunities. However, we want to investigate whether it is truly better for men and women in terms of the Gender Inequality Index (GII). To examine this relationship, we will employ logistic regression to determine if GII can effectively predict the classification of a country as part of the developed world.

```
Log_df = df[["Human_development","GII"]]# drop unneeded columns
Log_df["Human_development"].unique()
# For this we are going to convert human devolpment in to binary
Log_df["Human_development"] = np.where((df['Human_development'] == 'Very high')
                              | (df['Human_development'] == 'High'), 1, 0)
Log_df["Human_development"].unique() # OUTPUT: array([1, 0])
Log_df = Log_df.dropna() # drop null values
# graph
y = Log_df["Human_development"].value_counts()
print("the fraction of 'devolped countries' is "+
      str(y/len(Log_df["Human_development"])))
# the fraction of 'devolped countries' is     0.611765
```

Prediction: It is expected that there will be a statistically significant correlation between countries considered developed and higher GII values. This correlation aligns with our intuition, as countries tend to expand the rights of their citizens, including marginalized groups such as women, as they progress in development. This progress often involves implementing policies

and initiatives that aim to promote equal access to education, healthcare, employment opportunities, and political representation. As these measures are implemented, the GII may rise due to increased awareness and measurement of gender inequality. (Code for our prediction is on the left.)

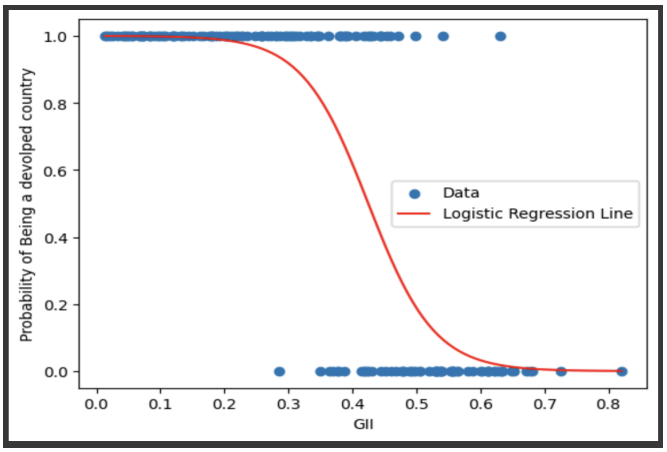
Why It's important: relating human development to GII:

```
# run logsitical regression on entire data set
import statsmodels.api as sm
logit_model=sm.Logit(Log_df["Human_development"],sm.add_constant(Log_df["GII"]))
result=logit_model.fit()
print(result.summary())
intercept, slope = result.params # Get the estimated coefficients
x_line = np.linspace( Log_df["GII"].min(), Log_df["GII"].max(), 100)
X_line_with_constant = sm.add_constant(x_line.reshape(-1, 1))
y_line_pred = result.predict(X_line_with_constant)

plt.scatter(Log_df["GII"], Log_df["Human_development"], label='Data')
plt.plot(x_line, y_line_pred, color='red', label='Logistic Regression Line')
plt.xlabel('GII')
plt.ylabel('Probability of Being a devolped country')
plt.legend()
plt.show()
```

Understanding the relationship between human development and the Gender Inequality Index (GII) holds significant importance. By examining this relationship, we can gain insights into the extent of gender inequality within societies and its impact on overall human development. This analysis allows us to assess the effectiveness of existing development measures and policies in promoting gender equality. By investigating the link between human development and GII through rigorous statistical methods like logistic regression, we can uncover valuable insights to inform policies and interventions aimed at reducing gender disparities and fostering inclusive and sustainable development.

```
Optimization terminated successfully.
         Current function value: 0.263192
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:      Human_development   No. Observations:                  170
Model:                          Logit   Df Residuals:                      168
Method:                           MLE   Df Model:                            1
Date:                Tue, 16 May 2023   Pseudo R-squ.:                  0.6060
Time:                        15:57:00   Log-Likelihood:                -44.743
converged:                       True   LL-Null:                       -113.55
Covariance Type:            nonrobust   LLR p-value:                 8.832e-32
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          8.2430      1.382      5.964      0.000       5.534      10.952
GII          -19.4034      3.176     -6.110      0.000     -25.628     -13.179
==============================================================================
```



Analysis of Regression: Fitting the logistic regression model and analyzing the relationship between the Gender Inequality Index (GII) and the classification of countries as developed or not developed, a clear trend is seen. Countries with higher GII values are more likely to be categorized as not developed.

Interpretation of the coefficients: The GII coefficient of -19.403 suggests that higher values of GII are associated with a lower likelihood of belonging to the Human_development category. We can also examine the p values of the Regressor GII, it is determined to be statistically significant according to our model. Suggesting that it is a valid predictor for our data. **The Pseudo R-squared** of 0.6060 indicates a relatively good fit of the model to the data. However, it should be interpreted with caution in logistic regression models.

```
X_train, X_test, y_train, y_test = train_test_split(sm.add_constant(Log_df["GII"])
, Log_df["Human_development"], test_size=.3, random_state=0)
reg = LogisticRegression().fit(X_train, y_train)
y_pred = reg.predict(X_test)
print(classification_report(y_test, y_pred))
```

Overfitting: When building a regression model, it's important to be cautious about overfitting. To address this concern, a common practice is to perform a train-test split on the data. The model will be trained on 70% of the data and tested on 30%. By evaluating the model on the test set, you can gain insights into its accuracy and assess whether it's overfitting.

```
              precision    recall  f1-score   support

           0       0.94      0.73      0.82        22
           1       0.82      0.97      0.89        29

    accuracy                           0.86        51
   macro avg       0.88      0.85      0.85        51
weighted avg       0.87      0.86      0.86        51
```

Support: We can see that the test data had 22 non-developed cases and 29 developed cases.

Precision: The ratio of correctly predicted data divided by its status of development. The precision is 0.94, indicating that 94% of the instances predicted not developed were truly not developed. The precision is 0.82, meaning that 82% of the instances predicted Developed were actually Developed.

Accuracy: The proportion of correctly classified instances out of the total instances in the test set. In this case, the accuracy is 0.86, meaning that the model correctly predicted 86% of the instances in the test set.

Staggered Training: While the train-test split is a good indicator of overfitting a more robust test for overfitting is by shuffling the data when testing. By shuffling the data and comparing the model performance across multiple iterations, such as by examining the accuracy, we can assess the consistency of the model's performance and determine if overfitting is a concern.

```
from sklearn.metrics import accuracy_score
Log_df = Log_df.sample(frac=1, random_state=42)
train_ratio = 0.6
step = 0.1  # Step size for staggered training
train_index = int(len(Log_df) * train_ratio) # Calculate the split indices
test_index = int(train_index + len(Log_df) * step)
accuracy_array= [] # create array to see how accuracy varies over steps
while test_index < len(Log_df):
    # Split the data into train and test sets
    train_data = Log_df[:train_index]
    test_data = Log_df[train_index:test_index]

    X_train = train_data["GII"]
    y_train = train_data["Human_development"]
    X_test = test_data["GII"]
    y_test = test_data["Human_development"]
    logit_model = sm.Logit(y_train, sm.add_constant(X_train)).fit()
    y_pred = result.predict(sm.add_constant(X_test))
    y_pred_binary = (y_pred > 0.5).astype(int)
    accuracy = accuracy_score(y_test, y_pred_binary)
    accuracy_array.append(accuracy)
    # Move to the next iteration
    train_index = test_index
    test_index = int(train_index + len(Log_df) * step)
print(accuracy_array) # [0.9411764705882353, 0.8823529411764706, 0.8235294117647058]
```

Staggered test accuracy

We implemented shuffling of the data to address any potential correlation between the position of data in the dataframe and its developed status and GII. By shuffling the data, we introduce more variability and randomness, which can lead to a more accurate staggered test-train split.

After shuffling the data and performing the staggered test-train split, we obtained the following accuracy results from each iteration: [0.9411764705882353, 0.8823529411764706, 0.8235294117647058].

By observing high accuracy scores across each iteration, we can conclude that the model generalizes well and is not overfitted. This provides reassurance that the models can effectively determine whether a country is developed or not based on its GII.

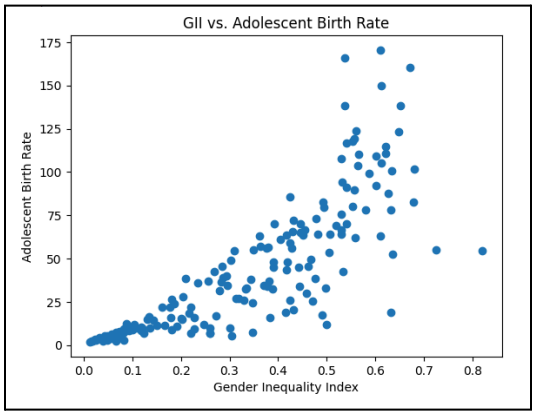
Conclusion: In conclusion, we have performed logistic regression analysis to explore the relationship between the Gender Inequality Index (GII) and the classification of countries as developed or not developed. Our findings indicate that there is a statistically significant correlation between higher GII values and countries being classified as not developed. This aligns with our intuitive understanding that as countries become more developed, they tend to prioritize expanding the rights and opportunities for their citizens, including marginalized groups such as women. This can lead to a higher GII, reflecting increased awareness and measurement of gender inequality. Furthermore, our logistic regression model demonstrated good predictive accuracy, with an overall accuracy of 0.86 on the test set. We also concluded the model wasn't overfitting its data using staggered test sets.

Problems to consider: One limitation of our analysis is the relatively small size of the dataset, which consists of only 170 entries with non-null values. Another important consideration is the classification of countries as "Developed" or "Not Developed" within the dataset. Unfortunately, the database does not provide explicit information about the source or methodology used for this classification. It is reasonable to speculate that these classifications might have been transferred from another experiment or study that employed regression techniques and indicators such as economic development, GDP, government spending, and other relevant factors.

Question 4: How does GII correlate to adolescent birth rates across countries?

```
[ ]  #Question1
     x=df['GII']
     y=df['Adolescent_birth_rate']
     plt.xlabel('Gender Inequality Index')
     plt.ylabel('Adolescent Birth Rate')
     plt.title('GII vs. Adolescent Birth Rate')
     plt.scatter(x,y)
```

This scatter plot (code on left, see figure on right) compares the Gender Inequality Index on the X-axis to the Adolescent Birth Rate on the Y-axis. We can see that there is a positive correlation between the Gender Inequality Index of a country with its adolescent birth rate. This is a general trend, and as the GII increases, the plots on the scatter plot become more and more spread out. This indicates that the correlation coefficient becomes weaker as the GII increases.

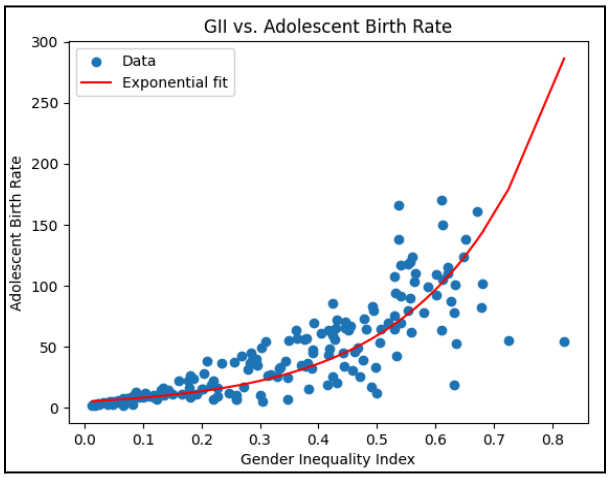


The original scatter plot compares the Gender Inequality Index on the X-axis to the Adolescent Birth Rate on the Y-axis. This further analysis (code and regression results on left) is intended to interpret the relationship between GII and Adolescent Birth Rate across the 190+ countries in the world. From this analysis, we can see that there is a positive correlation between the Gender Inequality Index of a country with its adolescent birth rate. This is a general trend, and as the GII increases, the plots on the scatter plot become more and more spread out. This indicates that the correlation coefficient becomes weaker as the GII increases.

```
X = df['GII']
y = np.log(df['Adolescent_birth_rate'])
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
print(model.summary())
plt.scatter(df['GII'], df['Adolescent_birth_rate'], label='Data')
plt.xlabel('Gender Inequality Index')
plt.ylabel('Adolescent Birth Rate')
plt.title('GII vs. Adolescent Birth Rate')
df_sorted = df.sort_values(by='GII')
plt.plot(df_sorted['GII'], np.exp(model.predict(sm.add_constant(df_sorted['GII']))))
        \ , color='red', label='Exponential fit')
plt.legend()
plt.show()
```

To improve this analysis, we have decided to also include an exponential regression. An exponential regression is the same as a linear regression but uses a natural logarithm to linearize the data. By applying the natural log to the dependent variable, we can then fit a linear regression model to the transformed data. This new regression provides a much better fit, as the R-squared value is now 0.759. As the gender inequality index increases, the adolescent birth rate exponentially increases. This is likely because, in more developed countries, where the inequality is less, women are more likely to take on careers and have fewer children. This analysis helps address the predictive question of how the birth rate is affected by rising or declining gender inequality. Our exponential regression graph is visualized on the right:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Adolescent_birth_rate   R-squared:                       0.759
Model:                               OLS   Adj. R-squared:                  0.757
Method:                    Least Squares   F-statistic:                     528.1
Date:                   Tue, 16 May 2023   Prob (F-statistic):           9.76e-54
Time:                           02:20:15   Log-Likelihood:                -138.46
No. Observations:                    170   AIC:                             280.9
Df Residuals:                        168   BIC:                             287.2
Df Model:                              1
Covariance Type:               nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.6147      0.085     18.987      0.000       1.447       1.783
GII            4.9290      0.214     22.980      0.000       4.506       5.352
==============================================================================
Omnibus:                       20.383   Durbin-Watson:                   1.776
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               23.723
Skew:                          -0.863   Prob(JB):                     7.06e-06
Kurtosis:                       3.607   Cond. No.                         5.71
==============================================================================
```



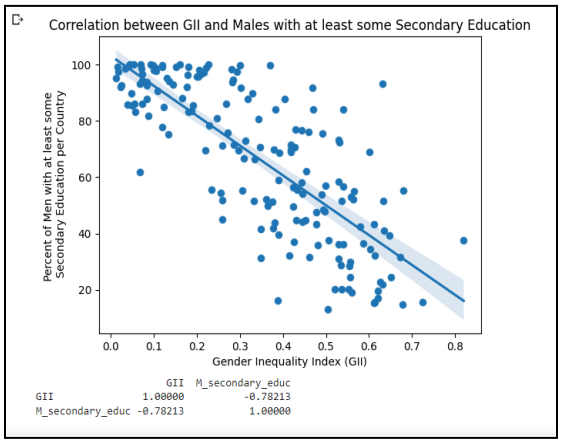
Question 5: How does GII correlate with the percentage of males with at least some secondary education?

```
df2 = df[['GII', 'M_secondary_educ']]
corr = df2.corr(method='pearson')

sns.scatterplot(data=df2, x='GII', y='M_secondary_educ')
sns.regplot(data=df2, x='GII', y='M_secondary_educ')

plt.title('Correlation between GII and Males with at least some Secondary Education')
plt.xlabel('Gender Inequality Index (GII)')
plt.ylabel('Percent of Men with at least some \n Secondary Education per Country')

plt.show()
print(corr)
```




```
                    GII  M_secondary_educ
GII             1.00000          -0.78213
M_secondary_educ -0.78213         1.00000
```

This scatter plot (code and visualization on left) effectively illustrates the correlation between the Gender Inequality Index (GII) and the proportion of male individuals who have completed at least some secondary education. The x and y-axes represent the GII and male secondary education percentages, respectively. Each data point found on the scatter plot signifies an individual country from our data set. The linear regression line depicts the direct negative correlation in the data, and the correlation coefficient shows the strength and direction of the correlation between the two variables. A nation from the dataset is represented by each point on the scatter plot.

The scatter plot's regression line illustrates the trend in the data; its negative slope reveals that as the GII rises, the proportion of men with at least some secondary education declines. This could suggest that inequality could be heavily accentuated in certain countries when the entirety of its inhabitants are devoid of ample educational resources.

Concerning Pearson correlation analysis, the calculation of its corresponding coefficient determines the magnitude and orientation of correlation exhibited between two variables. A value of 1 signifies perfect positive association while minus 1 is indicative of the opposite - perfect negative correlation. This Pearson Correlation test has a coefficient of -.78213, which indicates that there is a strong, negative relationship between the variables.

```
df3 = df[['GII','M_secondary_educ']].dropna()
X = sm.add_constant(df3['GII'])

model = sm.OLS(df3['M_secondary_educ'], X).fit()
print(model.summary())
```

Analyzing the regression line is necessary to understand the relationship between the variables. Examining the outputs of the OLS may allow us to accurately predict the percentage of male secondary education attainment based on the respective nation's GII. The assumptions of the linear regression will be tested. Our code is on the left and the results are on the next page.

The OLS analysis results show the **number of observations**, indicating that 20 of the original 190 countries weren't included in the model due to their unwillingness to disclose data for the GII, which could affect the accuracy of the results. The analysis indicates the **y-intercept** of the model is 103.0947. In context, this means the hypothetical percent of men that receive secondary education is somehow above 100% when the GII is 0; this is because no nation has a perfect GII. Assuming the validity of this model, then this would hypothetically indicate a nation of perfect equality amongst gender is one where everyone is educated. (According to the data source, the GII considers education rates for both genders comparatively, so individuals of all genders are probably educated at a high rate based on this.)

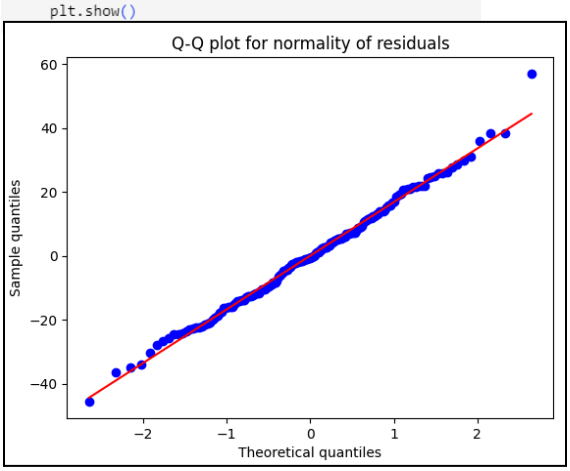
```
                            OLS Regression Results
==============================================================================
Dep. Variable:       M_secondary_educ   R-squared:                       0.612
Model:                           OLS   Adj. R-squared:                  0.609
Method:                Least Squares   F-statistic:                     264.7
Date:               Mon, 15 May 2023   Prob (F-statistic):           2.41e-36
Time:                       22:00:52   Log-Likelihood:                -718.99
No. Observations:                170   AIC:                             1442.
Df Residuals:                    168   BIC:                             1448.
Df Model:                          1
Covariance Type:           nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        103.0947      2.587     39.858      0.000      97.988     108.201
GII         -106.1288      6.523    -16.269      0.000    -119.007     -93.251
==============================================================================
Omnibus:                        1.573   Durbin-Watson:                   1.501
Prob(Omnibus):                  0.455   Jarque-Bera (JB):                1.215
Skew:                           0.187   Prob(JB):                        0.545
Kurtosis:                       3.177   Cond. No.                         5.71
==============================================================================
```

The **slope of the relationship** between the GII and rates at which men attain secondary education is -106.1288, which means that for every unit decrease in the GII, the chance of a man attaining advanced education decreases by that amount, or--in more relevant terms--a one percent increase of a nation's GII means that the chances of a man residing within said country receiving secondary education decreases by 1.06%.

The **95% CI** for the population's coefficients indicates that there's a 95% chance that the true populations' y-intercept would fall between 97.988 and 108.201, and between -93.251 and -119.007 for the slope. This shows that our model is generalizable to other nations, as given these metrics, it's safe to conclude that GII has a negative correlation with the rates of secondary educational attainment for men, as the 95% confidence variable for the negative slope doesn't include 0, nor any positive numbers. Similarly, the **standard error** measures the uncertainty of the estimated coefficients, as it showcases the possible variability of the estimates from one sample to another. A smaller SE indicates the model's reliability and preciseness, and the GII's coefficient has a SE of 6.523. This value indicates that the model is relatively precise.

The regression analysis significance tests show the statistical significance of the coefficients. The **t-statistics** of the coefficients, calculated by dividing the coefficients by their SE, are 39.858 and -16.269, both of which are significant. The t-statistics' respective **p-values**, or the probability of observing a t-statistic of this or greater extremity assuming the null hypothesis (that being that there's no influence between the variables) are both 0. The high **F-statistic** of 264.7 with its accompanying p-value of 2.41e-36 further confirms the results' statistical significance. Therefore, there's strong evidence against the possibility of the variables not influencing one another. Therefore, the relationship between the GII and male secondary educational attainment isn't due to chance.

```
st.probplot(model.wresid, dist="norm", plot=plt)
plt.xlabel('Theoretical quantiles')
plt.ylabel('Sample quantiles')
plt.title('Q-Q plot for normality of residuals')
plt.show()
```



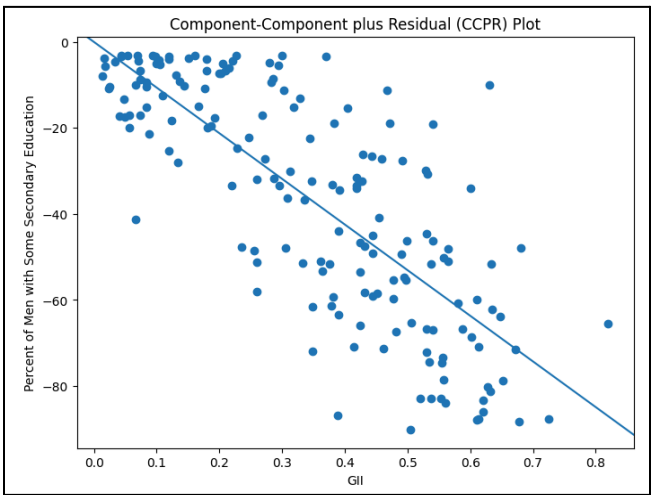
The **R-squared** value shows how much variance in predicted variables can be explained by the variance in independent variables. An R-squared value of 0.612 means that 61.2% of the variance in male educational attainment rates can be explained by the variance of GII values, indicating a decent fit for the data. However, it's important to check the assumptions of linear regression to ensure the model's accuracy, given the model's moderately successful fit. Influencing the R-squared value above, The values of residuals play a crucial role in linear regression as they determine the accuracy of the model. To ensure a good fit of the model, it's essential to have residuals that are non-constant (ensuring a good fit for the linearity of the model), homoscedastic, normally distributed, and independent observations. Any deviation from these factors can affect the statistics and analysis based on residuals.

The linear regression analysis tests the residuals, with the **Durbin-Watson** value of 1.501 indicating no autocorrelation (a value between 1 and 2 is ideal). The test confirms that the observations aren't dependent on each other, as autocorrelation happens when observations affect each other. The **Omnibus and JB** tests both returned values close to 1, with high p-values, indicating that the residuals are approximately normally distributed but with low certainty. The residual distribution is somewhat skewed to the right, as indicated by the value of 0.187. Finally, the kurtosis of 3.177 indicates the residuals have a "mesokurtic" distribution, i.e., There are slightly more points in the tails and slightly fewer in the center of the distribution compared to a normal distribution. Given these concerns about the normality of the residuals' distribution, a **Q-Q plot** can help confirm normality (see figure and code on left):

A Q-Q plot compares the distribution of the model's residuals to a normal distribution. In this plot, the residuals generally align with the normal distribution, indicating that they're relatively normally distributed.

```
fig, ax = plt.subplots(figsize=(8,6))
plot_ccpr(model, 'GII', ax=ax)
ax.set_xlabel('GII')
ax.set_ylabel('M_Secondary Education Values')
ax.set_title('Component-Component plus Residual (CCPR) Plot')
plt.show()
```

To test for heteroskedasticity and linearity, the following CCPR plot helps visualize the residuals (see figure on right):



The **component-component plus residual plot** visualizes the distribution of the residuals around a diagonal line of best fit; showing that the points that are randomly

distributed around the line to ensure that the residuals aren't heteroskedastic, and are also linear. This can be seen in the graph as the residuals are visually randomly distributed at all points of the visualization, and that the data is linear. With all of that considered, the assumptions of linear regressions are established, and the model is thus accurate.

On a concluding note; the analysis of the model shows that a higher national GII has a strong, negative effect on the educational attainment of men in that country, the results of which are statistically significant and are generalizable to other nations in the study. This leads to numerous theoretical sociological implications and questions:

- Is a society where gender equality is perfect a result of everyone being educated, or do higher education rates result from equal access to educational resources?
- Wealthier nations are more likely to have better educational systems and less gender inequality, indicating that financial status is a confounding variable to this model. Other social identities that affect financial status can also impact education rates accordingly. Apart from wealth, what specific social factors affect higher secondary education rates, and could also serve as confounding variables to this model?
- External factors such as poor environmental or political conditions may also negatively impact maternal mortality rates, which alters the GII data. Which factors in this case could play a confounding role in the model?

The findings from this analysis are particularly relevant to policymakers and educational boards, as the analysis shows that reducing gender inequality correlates to higher rates of advanced education for men. This correlation proves that improving gender equality won't harm men's success in education, which some opponents of feminism argue. This information is relevant to feminists in particular, as the data in this model directly disproves unpersuaded people concerned about how feminists' goals of equality may negatively affect educational opportunities for men.

```
timeseries = pd.read_csv('gender-inequality-index-from-the-human-development-report.csv')
X_train, X_test, Y_train, Y_test = train_test_split( \
sm.add_constant(timeseries['Year']), timeseries['Gender Inequality Index'], \
test_size=0.3, shuffle = False) #break up test and train data
timeseries['treatment'] = (timeseries['Year'] >= 2015).astype(int) # binary code
X_test = X_test.join(timeseries['treatment'])
X_train = X_train.join(timeseries['treatment']) # give binary code to train/test
model3 = sm.OLS(Y_train, X_train).fit()
y_pred = model3.predict(X_test)
print(model3.summary())
mse = mean_squared_error(Y_test, y_pred)
rmse = np.sqrt(mse)
print("Root Mean Squared Error:", rmse)
print("Mean Squared Error:", mse)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Gender Inequality Index   R-squared:                       0.051
Model:                                 OLS   Adj. R-squared:                  0.051
Method:                      Least Squares   F-statistic:                     98.81
Date:                     Tue, 16 May 2023   Prob (F-statistic):           1.62e-42
Time:                             18:33:13   Log-Likelihood:                 811.90
No. Observations:                     3643   AIC:                            -1618.
Df Residuals:                         3640   BIC:                            -1599.
Df Model:                                2
Covariance Type:                 nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.2136      1.024      9.979      0.000       8.207      12.220
Year          -0.0049      0.001     -9.546      0.000      -0.006      -0.004
treatment     -0.0030      0.011     -0.281      0.778      -0.024       0.018
==============================================================================
Omnibus:                     1445.698   Durbin-Watson:                   0.068
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              234.746
Skew:                          -0.297   Prob(JB):                     1.06e-51
Kurtosis:                       1.907   Cond. No.                     6.40e+05
==============================================================================
```

Question 6: What kind of effect do worldwide policies that increase the number of women in the peacemaking process have on GII?

In **2015** the security council of the United Nations adopted **resolution 2242**. This resolution pushed to double the number of women in all forms of the peace-making process. A few ways they tried to do this was to ensure that there was a gender analysis and technical gender expertise in all parts of the mission planning and execution. The gender analyst was needed to make sure that the number of women in the lawmaking process was at least doubled. Furthermore, there was training along with an investigation with all the United Nations peacemakers to prevent sexual exploitation and abuse that multiple members of the United Nations had been accused of.

This resolution was adopted in **2015**. More can be read about it here: https://press.un.org/en/2015/sc12076.doc.html. The purpose of this data analysis method is to compare the actual data of the gender inequality index with the predicted data if the resolution wasn't put into place. To do this I gathered time series data on each country and their GII from 1990 to 2021.

The source is Our Data Our World: https://ourworldindata.org/grapher/gender-inequality-index-from-the-human-development-report.

To test this, I made a train test split. I set shuffle equal to False. Shuffle being set to false allows us to set what portion of the data is being trained/tested. Here we are only training data before the policy was put into place (before 2015) and testing it on data when the resolution was adopted.

To do this I made another column ['treatment'] where it was set to 1 if the year was 2015 or greater and everything that has a year of 2014 or below was set to 0. Furthermore, I created a model of our trained data and created a y predicted using that model. In addition, I calculated the mean squared error and the root mean squared error.

1: R^2

We can analyze the results of the model summary by first looking at the R^2 value. The R^2 value looks at the proportion of variance in our predicted Gender Inequality Index compared to the true data. We have an R^2 of 0.051. The closer the R^2 value is to 0, the greater the

variation there is between our predicted values and the true values. Because 0.051 is close to 0, it tells us that the model explains 5.1% of the data. So our predicted and true values have a large amount of variation between them, and the model does not fit our data well.

2: Mean Squared Error and Root Mean Squared Error

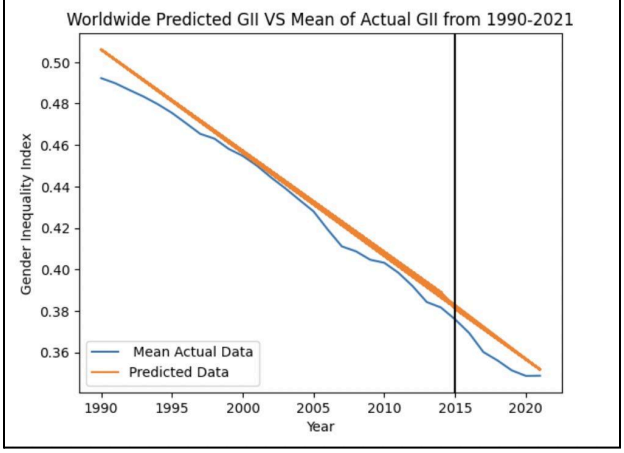
Here our mean squared error is 0.034811279225376864 and our root mean squared error is 0.1865778101098222. This is relatively high considering that GII values are between 0 and 1. This means that the model can predict the Gender Inequality Index with a **high error rate**. This means that the model does not do a very good job of predicting our data.

In conclusion, our R^2, MSE, and RMSE tell us that the model that we made of the trained data (the GII of each country from 1990-2014) does not accurately reflect the trained data (the GII of each country from 2015-2021). This means that from data before 2015, we would not be able to make useful decisions about Gender Inequality Index, since the predictions would not be accurate.

It can be shown that this policy affected the universal gender inequality index. Though this method alone would not stand to prove a causal relationship, we can hypothesize that the resolution caused a decrease in GII. The next step to prove a causal relationship would be to consider several hidden confounders. Some of these hidden confounders could be other policies made around the same time. Another could be the economy of certain countries. In conclusion, we have shown that after resolution 2242 was put into place in 2015, the correct gender inequality index was not able to be predicted using the data before 2015. To investigate this further, one should look into hidden confounders.

```
averageGII = timeseries.groupby('Year')['Gender Inequality Index'].mean()
plt.plot(averageGII.index, averageGII.values, label=' Mean Actual Data')
plt.plot(X_test['Year'], y_pred, label='Predicted Data')

plt.title('Worldwide Predicted GII VS Mean of Actual GII from 1990-2021')
plt.xlabel('Year')
plt.ylabel('Gender Inequality Index')
plt.legend()
plt.axvline(x=2015, color='black')
plt.show()
```



In continuation of the previous question, we made a line graph showing the predicted data against the mean actual data. We took the mean of each year so that we could easily compare it to the predicted value instead of having 180 lines on the same graph. Due to this, the orange line does not cut through the blue line. The black line shows the point where we started creating testing data from the trained data. We decided to take the mean of the actual data because each country has a different Gender Inequality Index for each year. To be able to compare the trend to the predicted trend, we took the mean of the gender inequality index for each year.

This line graph affirms the conclusion from the train test split above. After 2015 we can see that there is a gap between the orange and the blue line. Although a similar dip occurred in 2005, this one appears to dip lower and for a longer period of time. The graph emphasizes that after resolution 2242 in 2015, the gender inequality index varies from the predicted gender inequality index.

Conclusion:

Our analysis shows that the Gender Inequality Index (GII) is fairly reliable as a predictor for classifying whether a country is developed or not. Our findings show a pattern: as the GII rises, the likelihood of being categorized as developed decreases, indicating a negative correlation between gender inequality and development status. Furthermore, the in-depth analysis of the relationship between GII and Adolescent Birth Rate concludes that as GII increases, Adolescent Birth Rate also increases exponentially. This relationship can likely be explained by the fact that in less developed countries, where the GII is greater, women are often not allowed into the workforce. In turn, women are encouraged to stay home and raise families, which explains an increase in the birth rate. On the other hand, in countries where GII is lower, women enter the workforce and are less likely to have multiple children due to time constraints and career goals that can be delayed by having children.

We conclude that a higher national GII has a negative effect on the educational attainment of men in a country. This has sociological implications, raising questions about the relationship between education and gender equality, as well as the impact of socioeconomic factors on education rates. The findings are relevant to activists, policymakers, and educational authorities, as they show that reducing gender inequality leads to higher rates of advanced education for men, disproving those who argue that feminism harms men's educational opportunities. Lastly, we have shown that after resolution 2242 in 2015, the previous gender inequality index (GII) data would not be sufficient in predicting GII. We can hypothesize that policies that double the number of women in the lawmaking process reduce the gender inequality index. To show this causal relationship, further investigation needs to be done with confounding variables. The results of our data analysis across all 4 questions are very valuable to lawmakers, leaders, activists, healthcare policymakers, those who work in education, and organizations whose focus is on equality.

By: Jasmine Samadi, Yusef Rahimzada, Joey Guenoun, Conor Mervyn (NetIds: JNS222, YR238, JEG335, CDM226)