DSUA-111 Project Example

Name: Niel chonoolal

netid: ndc338

Part 1: Pre-Step

Q1: Describe briefly the question you would like to answer or the topic you would like to explore. Essentially, what do you hope to learn from your analysis?

A course that I absolutely admire is biology, especially the fundamental parts of biology. A key concept of biology that I learned was how much impact a species population has on their enviroment. An increase of population means that a species is starting to reach closer to their "carrying capacity", which is the maximum amount of indviduals that can be sustained in a given enviroment. This can mean that as a species grows in population competiton and problems may increase due for the need space, food and water. I would like to see if the change in the human population overtime has any significant relationship with a growth in competition, such as: large unporprotional inequality, crime rates and other social problems that may arrise.

Part 2: Data

Q2: Find a dataset that may help you explore at least some of these questions. First, describe where you found the data set. Second, describe how you found it. Third, describe at least two variables in the dataset that are relevant to the analysis you described above. Finally, describe the unit of observation (individual, city, etc.).

My first step was to find a dataset that has the population of every country which include a time frame such as years. I did this by searching "World Population dataset" on Google and found a very large dataset at https://data.worldbank.org/indicator/sp.pop.totl (https://data.worldbank.org/indicator/sp.pop.totl), which was created by the World Bank; however the only problem was that the all the years in the dataset has their own column instead of all the years to be placed in a single column. So after more looking on Google I found a second dataset that is actually the same exact dataset at https://github.com/datasets/population (https://github.com/datasets/population) which was just a cleaner version where all the years were placed into a single column.

The second dataset was to be mainly about a social problem, so my first instinct was to search up "world crime dataset". The first dataset that I felt was to most accurate was by the United Nation at https://dataunodc.un.org/GSH_app (https://dataunodc.un.org/GSH_app), where I was able to choose homicide rates in every country and I was able to select the amount of years I would like to be presented in the dataset.

The two variables that I was mainly focusing on were:

1. The Population in the Population dataset which records the countries and years as well
2. The homicide rate in the Homicide Rate date set which also records countries and years

where the unit of observation were countries and year.

Q3: If you could change this dataset in one way to make it better for your analysis, what would that change be and how could it improve your analysis?

One change that can better the Population dataset is to in fact to be a bit simpler in terms of Country. The population Dataset presents every country in Alphabetically order however it also places geographic areas, such as Arab World, Carribean Small States, Central Europe and many other areas under country name where it should've been placed into a seperate column. This can make it a bit hard to really concentrate on each individual country.

Q4: Import the dataset into Jupyter using any method you like and show the first five observations. If you had to do any pre-work to get the data into an uploadable format please describe it briefly. (If you didn’t, please say so as well.)

In [1]:
```
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
```

In [2]:
```
#data = pd.read_csv(os.path.expanduser('~/your_materials/homiciderates.csv'))
data1 = pd.read_csv(os.path.expanduser('~/your_materials/population.csv'))
data1.head()
```

Out[2]:

	Country Name	Country Code	Year	Value
0	Arab World	ARB	1960	92197753
1	Arab World	ARB	1961	94724510
2	Arab World	ARB	1962	97334442
3	Arab World	ARB	1963	100034179
4	Arab World	ARB	1964	102832760

This dataset is pretty simple however I think it would be best if we remove the column 'Country Code' because it really isn't needed

In [3]:
```
population = data1.filter(['Country Name', 'Year', 'Value'])
population.head()
```

Out[3]:

	Country Name	Year	Value
0	Arab World	1960	92197753
1	Arab World	1961	94724510
2	Arab World	1962	97334442
3	Arab World	1963	100034179
4	Arab World	1964	102832760

Renaming the columns "Country Name" to "Country" can be helpful in the merging process and renaming "Value" to "Population" can also be helpful

In [4]:
```
population.rename(columns = {'Country Name':'Country', 'Value':'Population'}, inplace = True)
population.head()
```

Out[4]:

	Country	Year	Population
0	Arab World	1960	92197753
1	Arab World	1961	94724510
2	Arab World	1962	97334442
3	Arab World	1963	100034179
4	Arab World	1964	102832760

In [5]:
```
df = pd.read_csv(os.path.expanduser('~/your_materials/homiciderates.csv'))
df.head()
```

Out[5]:

	Country	Region	Subregion	Indicator	2011	2012	2013	2014	2015	2016	2(
0	Afghanistan	Asia	Southern Asia	Homicide Total Count	1231.0	1948.0	NaN	NaN	3367.0	2289.0	251
1	Afghanistan	Asia	Southern Asia	Homicide Rate	4.1	6.3	NaN	NaN	10.0	6.6	
2	Albania	Europe	Southern Europe	Homicide Total Count	142.0	157.0	124.0	117.0	82.0	99.0	6
3	Albania	Europe	Southern Europe	Homicide Rate	4.9	5.4	4.2	4.0	2.8	3.4	
4	Algeria	Africa	Northern Africa	Homicide Total Count	280.0	523.0	480.0	577.0	542.0	NaN	N

The dataset present two indicators "Homicide rate" and "Homicide Total Count", for this project I would prefer to work with homicide rate because it would be very simple.

In [6]:
```
df1 = df.drop(index=df[df['Indicator'] == 'Homicide Total Count'].index)
df1.head()
```

Out[6]:

	Country	Region	Subregion	Indicator	2011	2012	2013	2014	2015	2016	2017
1	Afghanistan	Asia	Southern Asia	Homicide Rate	4.1	6.3	NaN	NaN	10.0	6.6	7.1
3	Albania	Europe	Southern Europe	Homicide Rate	4.9	5.4	4.2	4.0	2.8	3.4	2.3
5	Algeria	Africa	Northern Africa	Homicide Rate	0.8	1.4	1.3	1.5	1.4	NaN	NaN
7	American Samoa	Oceania	Polynesia	Homicide Rate	9.0	3.6	5.4	5.4	7.2	5.4	NaN
9	Andorra	Europe	Southern Europe	Homicide Rate	1.2	0.0	0.0	0.0	0.0	NaN	NaN

Now that the datset is a bit more simplier, I will place Year into its own column

In [7]:
```
df2 = pd.melt(df1, id_vars=["Country", "Region", "Subregion", "Indicator"], var_name="Year", value_name= "Homicide_Rate")
df2.head()
```

Out[7]:

	Country	Region	Subregion	Indicator	Year	Homicide_Rate
0	Afghanistan	Asia	Southern Asia	Homicide Rate	2011	4.1
1	Albania	Europe	Southern Europe	Homicide Rate	2011	4.9
2	Algeria	Africa	Northern Africa	Homicide Rate	2011	0.8
3	American Samoa	Oceania	Polynesia	Homicide Rate	2011	9.0
4	Andorra	Europe	Southern Europe	Homicide Rate	2011	1.2

Using pd.melt() I was able to establish columns where the year is placed into a single variable as well as the values of the homicide rates, which can be simplfied further by filtering the "Indicator" column.

In [8]:
```
#We can also filter Region, Subregion and Indicators just to really co
ncentrate on the country, year and homicide rate
#This is the final cleaned dataset for homicide rate
homicide = df2.filter(["Country", "Year", "Homicide_Rate"])
homicide.head()
```

Out[8]:

	Country	Year	Homicide_Rate
0	Afghanistan	2011	4.1
1	Albania	2011	4.9
2	Algeria	2011	0.8
3	American Samoa	2011	9.0
4	Andorra	2011	1.2

Now that all the datasets are cleaned it's time to merge

In [9]:
```
#Due to an unexpected error code I had to make sure both Year in both
dataset are int not strings
population['Year'] = population['Year'].astype(int)
homicide['Year'] = homicide['Year'].astype(int)
```

In [10]:
```
#merged_df = population.merge(homicide, on=['Country'], how='right',in
dicator=True)
merged_df = population.merge(homicide, on=['Country','Year'])
```

In [11]:
```
merged_df.head()
```

Out[11]:

	Country	Year	Population	Homicide_Rate
0	Afghanistan	2011	30117413	4.1
1	Afghanistan	2012	31161376	6.3
2	Afghanistan	2013	32269589	NaN
3	Afghanistan	2014	33370794	NaN
4	Afghanistan	2015	34413603	10.0

Part 3: Initial analysis

Q5. Conduct at least two different manipulations of your now-ready table that help you un- derstand something of interest about the dataset (e.g., you might explore options like sort, shape, value counts, groupby, etc.). Why did you choose these two, and what have you learned? (Hint: You may need to do a bit work to get the data into a format that is usable for you – e.g., renaming columns, changing data types, etc. If any of this was necessary, show your code and briefly explain why you made these changes)

The first manipulation would be to using shape to have a better assessment of how many rows and columns are in the dataframe

```
In [12]: merged_df.shape
```

```
Out[12]: (1296, 4)
```

The second manipulation will be value counts for homicide rates, to understand what homicide rates appear more frequently

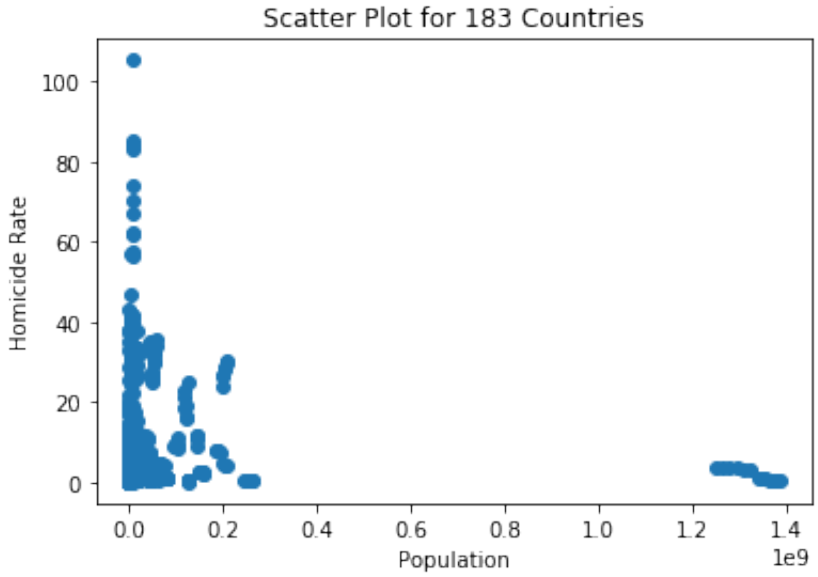
```
In [13]: merged_df['Homicide_Rate'].value_counts()
```

```
Out[13]: 0.9     30
         0.8     29
         0.7     28
         1.1     26
         0.6     20
                 ..
         31.1     1
         27.2     1
         33.7     1
         24.8     1
         27.9     1
         Name: Homicide_Rate, Length: 209, dtype: int64
```

With the second manipulation the most frequent homicide rate is 0.9 while the least frequent homicide rates are pretty high such as 31.1, 27.2, 33.7, 24.8 and 27.9. This tells me that most countries have realtively low and similar homicide rates while very few countries have a higher homicide rates

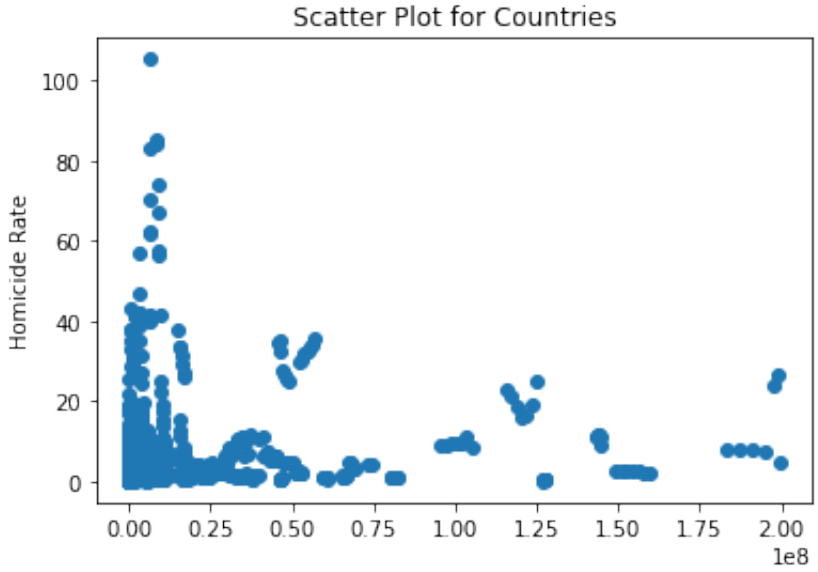
Q6. Generate two different types of graphs of any kind that are useful to you to better un- derstand what you're interested in. They don't need to be formatted particularly beautifully, but you do need to use two different types of graphs (e.g., a bar chart and a scatterplot) and explain what you hoped to understand, why you chose these graphs, and whether they're useful in improving your understanding.

In [14]:
```
plt.scatter(x=merged_df["Population"], y=merged_df["Homicide_Rate"])
plt.xlabel("Population")
plt.ylabel("Homicide Rate")
plt.title(f"Scatter Plot for 183 Countries")
plt.show()
```



The reason why I choose a scatter plot because since there are multiple values for each independent and dependent variable the scatter plot will show an accurate representation of the data. With this scatter plot there seems to be a concentration of countries that have generally low populations have lower homicide rates, however there are a few outliers that are shown. The first outlier is a few countries that have a low population but high homicide rate, however there also seems to be an outlier that is also a few countries with higher population that have very low homicide rates. With this refining the population to a lower range so we can have a closer look at the concentrated area to have a better understanding of the relationship

In [15]:
```
merged_df1 = merged_df[merged_df['Population']<= (2.0 * (10**8))]
plt.scatter(x=merged_df1["Population"], y=merged_df1["Homicide_Rate"])
plt.ylabel("Homicide Rate")
plt.title(f"Scatter Plot for Countries")
plt.show()
```



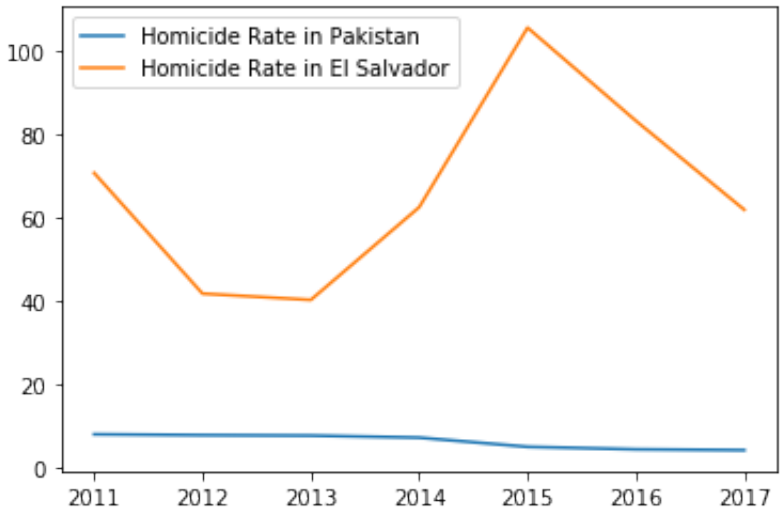
In [16]:
```
print(merged_df1.loc[merged_df1['Population'] == max(merged_df1.Popula
tion)])

print(merged_df1.loc[merged_df1['Homicide_Rate'] == max(merged_df1['Ho
micide_Rate'])])
#find country with highest population lowest population and then the c
ountry with the highest homicide rate and lowest crime rate
```

```
      Country  Year  Population  Homicide_Rate
915  Pakistan  2015   199426964            5.0
         Country  Year  Population  Homicide_Rate
375  El Salvador  2015     6325124          105.4
```

In this dataset Pakistan has the largest population seen in the dataset. El Salvador has the highest homicide rate. Creating a graph that compares the population and the homicide rate can show a relationship. Also both of the data returned is in the year of 2015.

In [17]:
```
pak = merged_df[merged_df.Country == 'Pakistan']
el = merged_df[merged_df.Country == 'El Salvador']
plt.plot(pak.Year, pak['Homicide_Rate'])
plt.plot(el.Year, el['Homicide_Rate'])
plt.legend(['Homicide Rate in Pakistan', 'Homicide Rate in El Salvador
'])
plt.show()
```



This line graph shows that at 2015 although El Salvador has a lower population than Pakistan, it has the higher murder rate than Pakistan which had the highest population.

Part 4: Hypothesis formation

Q7. What is your dependent variable and independent variable? Briefly describe how they are measured in this dataset. (Remember, they'll both need to be continuous variables.)

The independent variable is the population and the dependent variable is the homicide rate. The reason why is because we want to see if changes in population can affect the homcide rate. This is measured by the population in the dataset for each country and the homicide rate for each country as well.

Q8. Calculate the correlation coefficient between your two variables and interpret the result.

In [18]:
```
correlation_matrix = merged_df1[["Population", "Homicide_Rate"]].corr(
)
correlation_matrix
```

Out[18]:

	Population	Homicide_Rate
Population	1.000000	0.003063
Homicide_Rate	0.003063	1.000000

In [19]:
```
correlation_matrix.loc["Population", "Homicide_Rate"]
```

Out[19]: 0.00306322652831533

The correlation seems to be .306% which is not very strong and very minimal. This suggest that Population doesn't really have an effect on homicide rates.

Q9. Write out your regression model as an equation.

The equation will be:

```
Homicide Rate = Beta0 + Beta1 * Population
```

Beta0 is the y-intercept of the estimated regression line and Beta1 is the regression coefficent

Q10. Write out your null and alternative hypotheses.

The null hypothesis: Beta1 = 0 - Population change does't have an affect on crime rate

The alternative hypothesis: Beta1 != 0 - Population change does have an affect on crime rate

Part 5: Regression analysis

Q11. Estimate the regression equation you specified above and show the regression output.

In [26]:
```
import statsmodels.formula.api as sm
results_df = sm.ols(formula="Homicide_Rate ~ Population", data=merged_
df1).fit()
print(results_df.summary())
```

```
                            OLS Regression Results
==============================================================================
==========
Dep. Variable:          Homicide_Rate   R-squared:
0.000
Model:                            OLS   Adj. R-squared:
-0.001
Method:                 Least Squares   F-statistic:
0.007169
Date:                Fri, 08 May 2020   Prob (F-statistic):
0.933
Time:                        17:02:48   Log-Likelihood:
-2991.0
No. Observations:                 766   AIC:
5986.
Df Residuals:                     764   BIC:
5995.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
==========
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
----------
Intercept      7.3593      0.507     14.510      0.000       6.364
8.355
Population  1.086e-09   1.28e-08      0.085      0.933   -2.41e-08
2.63e-08
==============================================================================
==========
Omnibus:                      607.627   Durbin-Watson:
0.406
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
10430.580
Skew:                           3.569   Prob(JB):
0.00
Kurtosis:                      19.609   Cond. No.
4.62e+07
==============================================================================
==========

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.
[2] The condition number is large, 4.62e+07. This might indicate tha
t there are
strong multicollinearity or other numerical problems.
```

Q12 What do the results in the regression output tell you? Interpret the coefficient, p-value, and confidence interval for your independent variable (you don't have to do the intercept) and the R2.

In [21]:
```
results_df.params
```

Out[21]:
```
Intercept     7.359318e+00
Population    1.086016e-09
dtype: float64
```

In [22]:
```
beta1 = results_df.params['Population']
print("The increase of the population by", beta1, "is associated with
a an increase in homicide rate")
```

```
The increase of the population by 1.0860160199847005e-09 is associat
ed with a an increase in homicide rate
```

The next step is to find the p_value

In [23]:
```
results_df.pvalues
```

Out[23]:
```
Intercept     2.563652e-42
Population    9.325462e-01
dtype: float64
```

In [31]:
```
#results_df.pvalues['Population']
'%f' % (results_df.pvalues['Population'])
```

Out[31]:
```
'0.932546'
```

The p-value is saying that there is a significance of 0.932546 with our result. Due to this, the p-value is indicating that the null hypotheisis that states "Population change does't have an affect on crime rate" might be valid.

The 95 % interval for population seems to be [-2.41e-08 2.63e-08], which means that if we run samples like this multiple times 95% percent of the tumes the mean will be within these values.

In [25]:
```
'%f' % results_df.rsquared
```

Out[25]:
```
'0.000009'
```

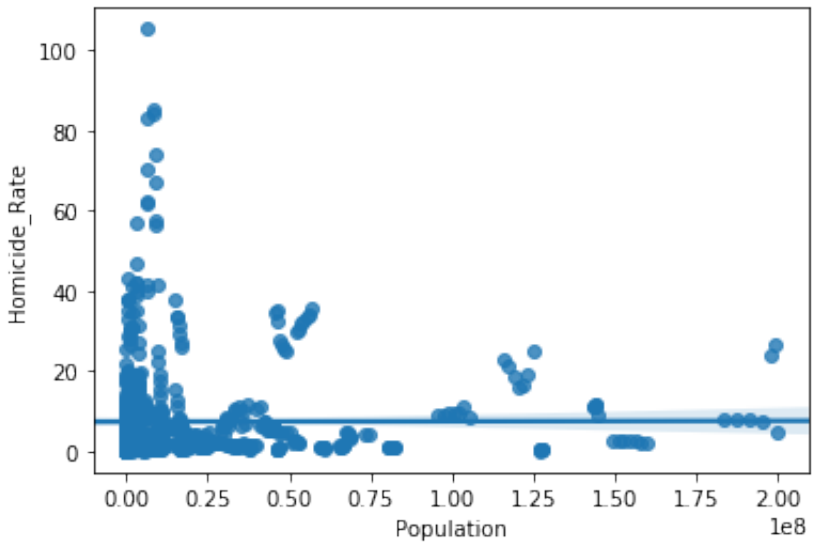
The r-squared value is 0.000009, which means that .0009% of the homicide rates can be explained by the population of a country, This value is extremely low and it really pushes us to reject the alternative hypothesis.

Q13. Which hypothesis do you reject and fail to reject, and why?

The hypothesis that will be rejected is the alternative hypothesis that states "Population change does have an affect on homicide rates". One reason why is that the r-squared value is saying that .0000 % of the time the homicide rates can be explained by population. This percentage is very low and it rejects the alternative hypothesis. Another reason why is the p-value supports the null hypothesis by producing an p-value of 0.932546 which is greater the 0.05 which supports the null value. The confidence interval describes a very small mean sample and it is furthering disproving any correlation between population and homicide rates. Therefore we can reject the alternative hypotheisis and fail to reject the null hypotheisis that sates "Population deosn't have an affect on homicdie rates"

Q14. Generate the residual plot and comment on any heteroskedasticity. What does this imply for your inference?

In [35]:
```
import seaborn as sns
sns.regplot(x='Population', y='Homicide_Rate', data = merged_df1)
plt.show()
```



Part 6: Conclusions

Q15. What biases might be present in the sample itself that could be affecting the outcome? Discuss at least two sources of bias.

One bias that may arrise is a selection bias of the homicide rates. There is not really a stable convinction of what homicide is in each country and some countries that may be corrupted may be under-represented or over-represented. Another bias that may arrise is information bias. The data did come from reliable sources however there seems to be many flaws when it comes to certain countries by placing NaN for population or homicide rate.

Q16. Considering all the work you've done, including the regression output, the results of your hypothesis tests, and any biases present in the data, what conclusions, however tentative, can you draw from your analysis about the relationship between your two variables of interest?

The conclusion that I can draw between Population and Homicide Rate is that there seems to be no correlation between them. The reason why is because in the data countries with high population doesn't have higher crime rates than countries that have lower population. This also means that there is not an inverse relationship because there are many cofounders that are in play. Cofounders such as government, laws, police force and inequality all play a factor in terms of homicide rates so there is no concrete evidence to lead us to believe that population affects homicide rates

Q17. What is your analysis's greatest weakness? In other words, what are the best reasons to be cautious about what we can learn from it?

The analysis greatest weakness is multiple data problems that arrised. Within the dataset some countries are very underrepresented in the dataset in terms of both Population and Homicide Rates. Although both datasets came from reliable sources such as the worldbank and the United Nations there still seems to be a large inconsitiicy for certain areas. Some countries have many NaN or 0 for population and/or homicide rates. This affects the regression process and is not helpful in determining what actually causes homicide rates. What can be learned from this is that there should be an intitative done to make sure these areas are not under-represented. Learning what causes homicide rates can help data scientist find ways how to create peace in the world and by taking measures to actually collect proper data for smaller overlooked countries.

In [ ]: