Data Science for Everyone - HW 4

April 20, 2022

HW4: Data Science for Everyone

Name: Suvir Wadhwa, **N-Number:** N16395336

Question 1

(a) De-anonymization can be a concern regardless of the number of features associated with each observation. As seen in the article, the researchers from UT Austin only needed a few features to be able to identify a person. We can identify people with minimal data at times.

(b) Aggregation, i.e. intersectional data, protects privacy. De-aggregration, i.e. working with non-intersectional data, risks privacy. Especially for minority populations.

(c) One of the ethical concerns of conducting research in the digital age is that data from the research may be accessible by anyone. It will also enable people such as hackers to break past barriers to get personal information about people.

Belmont Principle: Respect for Persons. In case of the netflix data, this is a potential concern because the people do not agree to their indentities being revealed but in this case, it is possible for that to happen.

(d) No matter how sensitive data may be, it is not okay if it can be re-identifiable. As for ethical and privacy concerns, the smallest amounts of data can have greater and unknown impacts. If the source of data does not agree on sharing its details, then the data must not be identifiable.

(e) Inputing data with social biases will skew the findings of an alogrithm towards one side. It will restrict the algorithm from finding results about its true population. This type of data could be successfully de-biased if more inputs of data are added to the set. These inputs however, must be from non-biased sources.

Question 2

[24]:
```
import pandas as pd
import numpy as np

df = pd.read_csv("human-development.csv")
df.head()
```

[24]:
```
        Entity      Code  Year  Human Development Index (UNDP)  \
0     Abkhazia  OWID_ABK  2015                             NaN
1  Afghanistan       AFG  1980                           0.228
2  Afghanistan       AFG  1985                           0.273
```

```
3  Afghanistan       AFG  2002                           0.373
4  Afghanistan       AFG  2003                           0.383

   Corruption Perception Index - Transparency International (2018)  \
0                                                NaN
1                                                NaN
2                                                NaN
3                                                NaN
4                                                NaN

   Population (historical estimates) Continent
0                                NaN      Asia
1                         13356500.0       NaN
2                         11938204.0       NaN
3                         22600774.0       NaN
4                         23680871.0       NaN
```

[25]:
```
df.columns = ['entity', 'code', 'year', 'hdi', 'cpi', 'population', 'continent']
df.head()
```

[25]:
```
        entity      code  year    hdi  cpi  population continent
0     Abkhazia  OWID_ABK  2015    NaN  NaN         NaN      Asia
1  Afghanistan       AFG  1980  0.228  NaN  13356500.0       NaN
2  Afghanistan       AFG  1985  0.273  NaN  11938204.0       NaN
3  Afghanistan       AFG  2002  0.373  NaN  22600774.0       NaN
4  Afghanistan       AFG  2003  0.383  NaN  23680871.0       NaN
```

[45]:
```
df = df[df['continent'].isnull()] #Clear all rows with values in the continent␣
 ↪column
values = ['Asia', 'Africa', 'Antartica', 'Oceania', 'Europe', 'South America',␣
 ↪'North America', 'World']
df = df[df.entity.isin(values) == False] #clear every row with continent or␣
 ↪world in its entity column.
df.tail()
```

[45]:
```
         entity code  year hdi  cpi  population continent
55733  Zimbabwe  ZWE  1988 NaN  NaN   9849129.0       NaN
55734  Zimbabwe  ZWE  1989 NaN  NaN  10153852.0       NaN
55735  Zimbabwe  ZWE  2019 NaN  NaN  14645473.0       NaN
55736  Zimbabwe  ZWE  2020 NaN  NaN  14862927.0       NaN
55737  Zimbabwe  ZWE  2021 NaN  NaN  15092171.0       NaN
```

[46]:
```
df_2017 = df[df['year'] == 2017]
df_2017.head()
```

[46]:
```
                entity code  year    hdi   cpi  population continent
18         Afghanistan  AFG  2017  0.498  15.0  36296111.0       NaN
```

```
549          Albania  ALB  2017  0.785  38.0   2884169.0        NaN
806          Algeria  DZA  2017  0.754  33.0  41389174.0        NaN
1147  American Samoa  ASM  2017    NaN   NaN     55617.0        NaN
1169         Andorra  AND  2017  0.858   NaN     76997.0        NaN
```

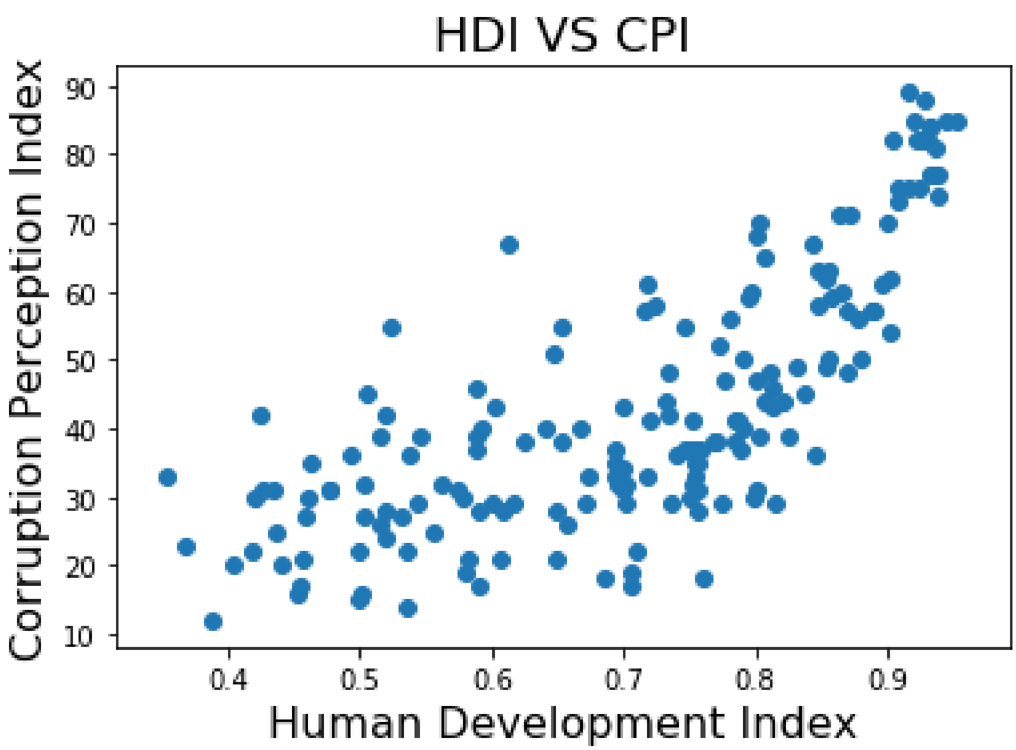
[47]:
```python
df_2017.shape #Number of rows and columns
```

[47]:
```
(239, 7)
```

Question 3

[48]:
```python
import matplotlib.pyplot as plt
```

[49]:
```python
plt.scatter(df_2017['hdi'], df_2017['cpi'])
plt.xlabel('Human Development Index', size=16)
plt.ylabel('Corruption Perception Index', size=16)
plt.title('HDI VS CPI', size=18)
plt.show()
```



(b) The relationship is almost a postive non-linear relationship. It seems like its going in the postive direction quadratically, its strength is quite strong as most of the points are in clusters.

(c) Human development is conceptualized as a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living.

It is operationalized by measuring:

– Life expectancy at birth

– Mean years of schooling & Expected years of schooling

– GNI per capita (in PPP adjusted international)

(d) Corruption is conceptualized as Annual ranking of countries by their perceived levels of corruption, as determined by expert assessments and opinion surveys.

Operatiinalized on a scale is from 100 (very clean) to 0 (highly corrupt).

[50]:
```
matrix = df_2017.corr() #(e)
matrix
```

[50]:

	year	hdi	cpi	population
year	NaN	NaN	NaN	NaN
hdi	NaN	1.000000	0.743760	-0.004584
cpi	NaN	0.743760	1.000000	-0.031604
population	NaN	-0.004584	-0.031604	1.000000

(f) The Pearson's correlation coefficent between the two is 0.743760. A measure of around 0.74 shows that there is a postive relationship between the cpi and hdi. This means that there is a postive relationship between the corruption in a country and the standards of living for those people in the country.

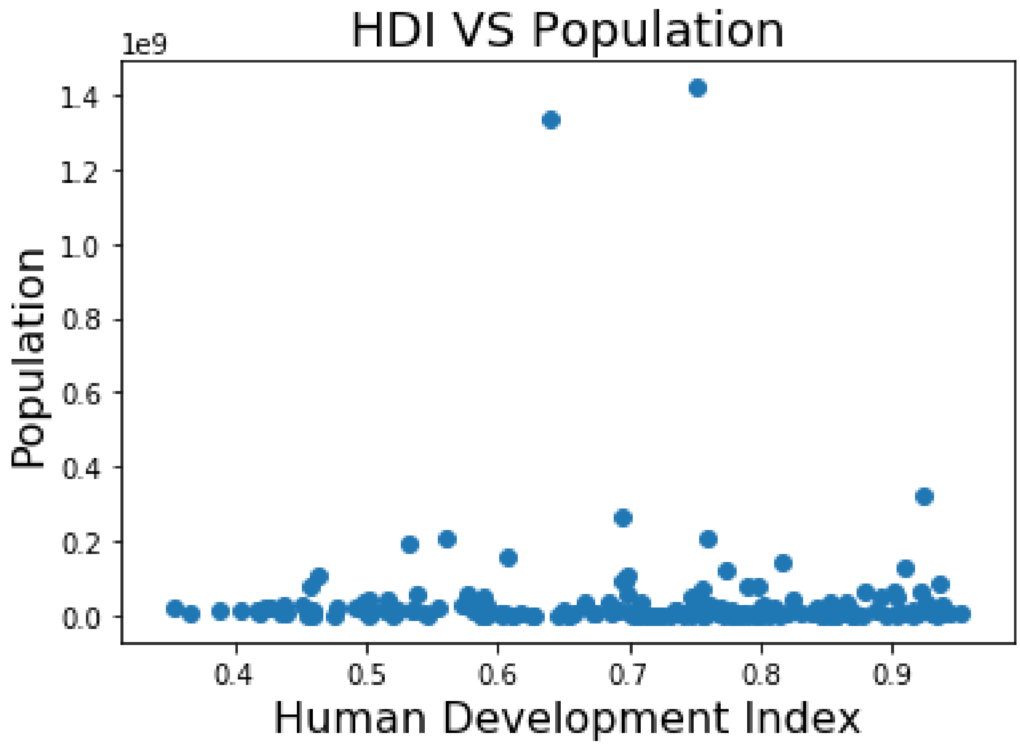
(g) Symmetric Correlation

(h) The correlation between hdi and population is weaker than the correlation between cpi and population.

(i) Correlation analysis is Unitless

Question 4

[51]:
```
plt.scatter(df_2017['hdi'], df_2017['population'])
plt.xlabel('Human Development Index', size=16)
plt.ylabel('Population', size=16)
plt.title('HDI VS Population', size=18)
plt.show()
```



[52]:
```
df_2017.sort_values(by=['population'], ascending=False)
```

[52]:

	entity	code	year	hdi	cpi	population	continent
10286	China	CHN	2017	0.752	41.0	1.421022e+09	NaN
22332	India	IND	2017	0.640	40.0	1.338677e+09	NaN
52664	United States	USA	2017	0.924	75.0	3.250848e+08	NaN
22591	Indonesia	IDN	2017	0.694	37.0	2.646510e+08	NaN
37351	Pakistan	PAK	2017	0.562	32.0	2.079062e+08	NaN
…	…	…	…	…	…	…	…
35689	Niue	NIU	2017	NaN	NaN	1.612000e+03	NaN
49997	Tokelau	TKL	2017	NaN	NaN	1.297000e+03	NaN
53988	Vatican	VAT	2017	NaN	NaN	7.980000e+02	NaN
25725	Korea	NaN	2017	0.903	NaN	NaN	NaN
25731	Kosovo	OWID_KOS	2017	NaN	39.0	NaN	NaN

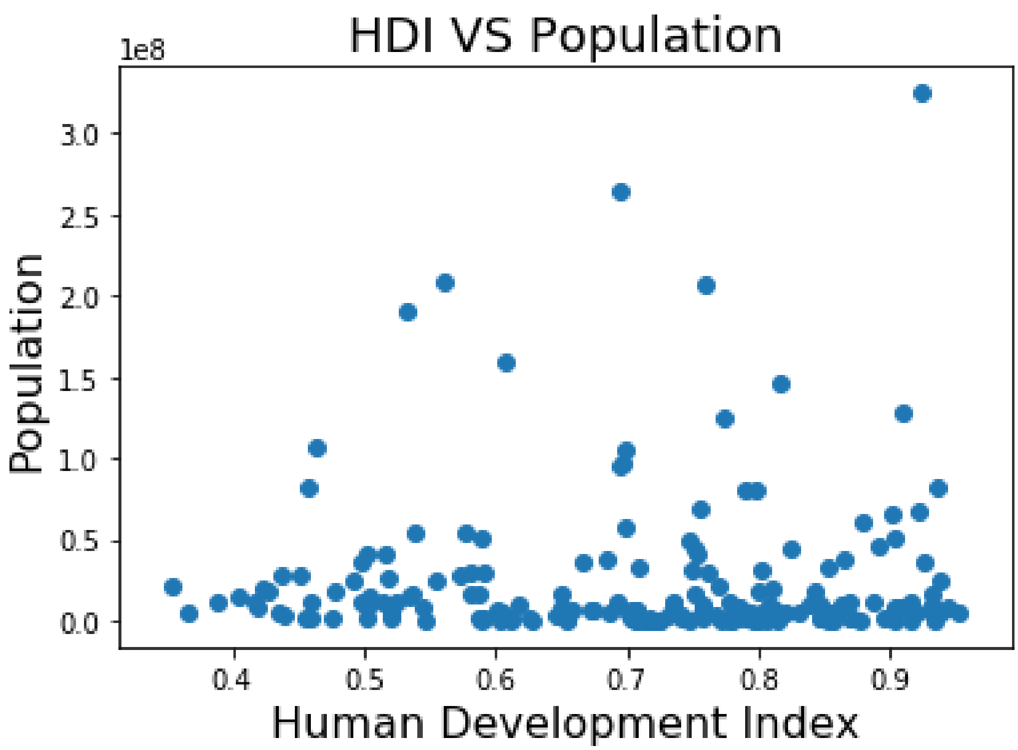
[239 rows x 7 columns]

(a) The two countries/outliers are China and India.

(b) Outliers should be kept in this analysis because that is the only way every country can be represented. China and India are two of the worlds most populated countries and this should be included in our findings to reflect accurate results. Outliers shouldn't be used because in this case it will be hard to work with the data. It could also skew the data in direction that's not expected.

[53]:
```
val = ['India', 'China']
df_new = df_2017[df_2017.entity.isin(val) == False]#df_2017[df_2017.entity !=␣
↪'China', 'India']
df_new.sort_values(by=['population'], ascending=False)

#Scatter Plot
plt.scatter(df_new['hdi'], df_new['population'])
plt.xlabel('Human Development Index', size=16)
plt.ylabel('Population', size=16)
plt.title('HDI VS Population', size=18)
plt.show()
```



(c) No relation at all. It seems like it is positive but it can't be concluded as there aren't enough points. It also is linear but with changes is HDI and a constant population.

[54]:
```
correlation = df_new['hdi'].corr(df_new['population'])
print("The correlation between the hdi and the population, without the␣
↪outliers, is {:2f}".format(correlation))
```

```
The correlation between the hdi and the population, without the outliers, is
0.009228
```

(e) The correlation has become stronger without the outliers as data is less skewed hence enabling us to predict a relationship more accurately.

(f) I would show both of the results as the outliers may play an important role in understanding this data. It will also give viewers an understanding of the impact the outliers had.

Question 5

(a) Null Hypothesis: There is no relation between corruption and human development.(the coefficient = 0)

Alternative Hypothesis: There is a relationship between corruption levels and human development.(the coefficient 0)

[55]:
```python
import statsmodels.formula.api as smf
results = smf.ols('hdi ~ cpi', data= df_2017).fit()     # simple linear regression
results.summary()
```

[55]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                    hdi   R-squared:                       0.553
Model:                            OLS   Adj. R-squared:                  0.551
Method:                 Least Squares   F-statistic:                     215.4
Date:                Wed, 20 Apr 2022   Prob (F-statistic):           2.95e-32
Time:                        01:51:31   Log-Likelihood:                 148.06
No. Observations:                 176   AIC:                            -292.1
Df Residuals:                     174   BIC:                            -285.8
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.4404      0.020     22.240      0.000       0.401       0.480
cpi            0.0062      0.000     14.677      0.000       0.005       0.007
==============================================================================
Omnibus:                        7.358   Durbin-Watson:                   1.977
Prob(Omnibus):                  0.025   Jarque-Bera (JB):                7.546
Skew:                          -0.480   Prob(JB):                       0.0230
Kurtosis:                       2.673   Cond. No.                         118.
==============================================================================

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

(c) The intercept is 0.4404 The coefficient on CPI is 0.0062

Interpretations:

When the CPI score is 0, the HDI is 0.4404

A one-unit increase in the cpi is associated with a 0.0062 increase in the HDI

(d) Yes, it does seem to be substantiviely significant as the error is low and the percentage of data within the bounds of the tails is minimal.

(e) The p-value for the coefficent on corruption is 0.000. This means the probability of seeing a coefficient of 0.4404 if the null hypothesis is true is 0.00. Hemce, When the p-value is 0, it's clearer if something is statistically significant.

(f) If the 95% confidence interval contains 0, then we have statistically significant coefficients. Hence, our alternate hypothesis is usually true.

(g) The R-squared for this regression model is 0.553

(h) The model is endogenous because the gradient by which cpi impacts hdi is similar to the gradient for hdi and cpi.

(i) Bad relations with other countries. When countries have bad relationships with other countries they're usually more corrupt. Hence, they also keep all of their resources to themselves and focus on strenghting only the development of their people.

(j) Yes, it is fair to say we have done a good job, although more could be done. We have checked for confounders, endoginity, accuracy of values, significance of values and different applicaitions.

Question 6

[56]:
```
results_2 = smf.ols('hdi ~ population + cpi', data= df_2017).fit()    #␣
↪multiple linear regression
results_2.summary()
```

[56]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                    hdi   R-squared:                       0.554
Model:                            OLS   Adj. R-squared:                  0.549
Method:                 Least Squares   F-statistic:                     107.4
Date:                Wed, 20 Apr 2022   Prob (F-statistic):           4.82e-31
Time:                        01:51:32   Log-Likelihood:                 148.18
No. Observations:                 176   AIC:                            -290.4
Df Residuals:                     173   BIC:                            -280.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.4390      0.020     21.896      0.000       0.399       0.479
population  2.609e-11   5.25e-11      0.497      0.620   -7.76e-11     1.3e-10
cpi            0.0062      0.000     14.654      0.000       0.005       0.007
```

```
==============================================================================
Omnibus:                        7.178   Durbin-Watson:                   1.983
Prob(Omnibus):                  0.028   Jarque-Bera (JB):                7.311
Skew:                          -0.469   Prob(JB):                       0.0258
Kurtosis:                       2.660   Cond. No.                     3.97e+08
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 3.97e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

(b) Null Hypothesis: There is no relationship between population and corruption levels on human development.(the coefficients = 0)

Alternative Hypothesis: There is a relationship population and corruption levels on human development.(the coefficients   0)

(c) The coefficent on population is 2.609e-11.

(d) No, the coefficient on population is greater than the $p<0.05$ level. The means the null hypothesis is true.

(e) No, the model is not an improvement becuase the adj r-squared value here is 0.549 which is lower than the previous value which was 0.551. This means, we have just added variables with no relation to the data. Hence, in this case, the adj R-Squared value adjusts.

———**END OF HW**———

[ ]: