Homework 3

April 4, 2022

Data Science for Everyone - HW 3, **Name:** Suvir Wadhwa, **N Number:** N16395336

Question 1

(a) Dependent Variable: The approval of Mayor Adams

(b) Independent Variable: Replies from People.

(c) The researcher is conceptualizing the approval of Mayor Adams by understanding what people think of his policies. Hence, how supportive people are of the policy changes mayor adams made. The bettter they think, the more supportive they would be.

(d) The researcher is operationalization the approval of Mayor Adams by setting a metric to measure categorize what people think. The researcher does this by giving the interviewee the options: "very", "somewhat", "not really", "not at all", and "prefer not to say/no opinion/don't know."

(e) The strength of this measure is that it gives the researcher set categories of data to work with. It enables the researcher to easily organize responses and make inferences from them.

(f) The weakness of this measure is that it is very generic and isn't a thorough measure of peoples apporval of the mayor.

(g) Misshearing the participant in the study.

(h) Response Bias. Participants may not be aware of the policies imposed by Mayor Adams. Hence, they may assume or give responses without any assurance. It will likely Bias the result away from the approval of Mayor Adams.

(i) 1. High number of Mayor Adams supporters in Times Square 2. Participant may not be a resident of NYC

(j) Error of Validity

Question 2

[4]:
```
'(a)'
import numpy as np #Required Libraries
import pandas as pd
df = pd.read_csv("books_per_mm.csv")
df.head(15)
```

[4]:

	Entity	Code	Year	Book titles per capita (Fink-Jensen 2015)
0	Algeria	DZA	1953	10.596210

```
1   Algeria  DZA  1954                                  8.115622
2   Algeria  DZA  1963                                 12.596606
3   Algeria  DZA  1964                                 13.605609
4   Algeria  DZA  1965                                 10.950347
5   Algeria  DZA  1966                                 15.884413
6   Algeria  DZA  1967                                 20.218645
7   Algeria  DZA  1968                                 21.983427
8   Algeria  DZA  1979                                 14.097694
9   Algeria  DZA  1980                                 14.622945
10  Algeria  DZA  1982                                 25.157543
11  Algeria  DZA  1983                                 10.057549
12  Algeria  DZA  1984                                 33.644875
13  Algeria  DZA  1985                                 32.623833
14  Algeria  DZA  1991                                 19.231459
```

[5]:
```
'(b)'
df.rename(columns = {'Entity' : 'entity', 'Code' : 'code', 'Year': 'year',
                     'Book titles per capita (Fink-Jensen 2015)' : 'books per␣
 ↪capita'}, inplace = True)
df.head()
```

[5]:
```
    entity code  year  books per capita
0  Algeria  DZA  1953          10.596210
1  Algeria  DZA  1954           8.115622
2  Algeria  DZA  1963          12.596606
3  Algeria  DZA  1964          13.605609
4  Algeria  DZA  1965          10.950347
```

[6]:
```
'(c)'
df['books per capita'] = df['books per capita'].round(decimals = 2)
df
```

[6]:
```
           entity      code  year  books per capita
0         Algeria       DZA  1953             10.60
1         Algeria       DZA  1954              8.12
2         Algeria       DZA  1963             12.60
3         Algeria       DZA  1964             13.61
4         Algeria       DZA  1965             10.95
…           …       …    …                 …
8307  Yugoslavia  OWID_YGS  1987            473.32
8308  Yugoslavia  OWID_YGS  1988            536.91
8309  Yugoslavia  OWID_YGS  1989            500.96
8310  Yugoslavia  OWID_YGS  1990            425.67
8311  Yugoslavia  OWID_YGS  1991            175.30

[8312 rows x 4 columns]
```

[7]:
```
'(d)'
df.sort_values("year")
```

[7]:

	entity	code	year	books per capita
1776	Germany	DEU	1500	45.24
372	Belgium	BEL	1500	25.60
6639	Switzerland	CHE	1500	81.54
7163	Turkey	TUR	1500	0.00
5068	Poland	POL	1500	0.00
…	…	…	…	…
7035	Switzerland	CHE	2009	1460.42
4820	Norway	NOR	2009	3276.76
6638	Sweden	SWE	2009	2504.42
1193	Denmark	DNK	2009	2405.38
8052	United Kingdom	GBR	2009	2114.85

[8312 rows x 4 columns]

From '1500' is the longest ago observation in this dataset.

[8]:
```
print('(e) Data for all countries in 1909')
df_1909 = df[df['year'] == 1909]
df_1909.head(10)
```

(e) Data for all countries in 1909

[8]:

	entity	code	year	books per capita
707	Belgium	BEL	1909	361.46
1093	Denmark	DNK	1909	1503.34
4447	Netherlands	NLD	1909	736.61
4720	Norway	NOR	1909	476.98
6538	Sweden	SWE	1909	756.83
7953	United Kingdom	GBR	1909	179.67

[9]:
```
print('(e) Data for all countries in 2009')
df_2009 = df[df['year'] == 2009]
df_2009.head(10)
```

(e) Data for all countries in 2009

[9]:

	entity	code	year	books per capita
1193	Denmark	DNK	2009	2405.38
4547	Netherlands	NLD	2009	2628.86
4820	Norway	NOR	2009	3276.76
5696	Russia	RUS	2009	898.49
6638	Sweden	SWE	2009	2504.42
7035	Switzerland	CHE	2009	1460.42
8052	United Kingdom	GBR	2009	2114.85

[10]:
```
print("(f) Country will highest book production in 1909:")
df_1909[df_1909['books per capita'] == df_1909['books per capita'].max()]
```

```
(f) Country will highest book production in 1909:
```

[10]:

	entity	code	year	books per capita
1093	Denmark	DNK	1909	1503.34

[11]:
```
print("(f) Country will highest book production in 2009:")
df_2009[df_2009['books per capita'] == df_2009['books per capita'].max()]
```

```
(f) Country will highest book production in 2009:
```

[11]:

	entity	code	year	books per capita
4820	Norway	NOR	2009	3276.76

[12]:
```
print("(g) Total observations for 1909:")
df_1909.shape[0]
```

```
(g) Total observations for 1909:
```

[12]:
```
6
```

[13]:
```
print("(g) Total observations for 2009:")
df_2009.shape[0]
```

```
(g) Total observations for 2009:
```

[13]:
```
7
```

(g) There is a selection bias as the dataset only considers European Countries but claims to provide a 'global' metric.

Question 3

[14]:
```
print("(a)")
data = pd.read_csv("CIOdata.csv")
data.head()
```

```
(a)
```

[14]:

	scode	ccode	country	year	uiareg	un	ciob	cioc	ciod	ciotot	…	\
0	AFG	700	Afghanistan	1952	3	1	7	0	0	8	…	
1	AFG	700	Afghanistan	1957	3	1	9	0	0	10	…	
2	AFG	700	Afghanistan	1962	3	1	12	1	0	14	…	
3	AFG	700	Afghanistan	1967	3	1	12	2	0	15	…	
4	AFG	700	Afghanistan	1972	3	1	13	2	0	16	…	

	sartoc	sacu	sarccus	scsa	uambd	uar	upeb	udeac	unesco	unido

```
0     NaN   NaN      NaN   NaN    NaN   NaN   NaN    NaN       1    NaN
1     NaN   NaN      NaN   NaN    NaN   NaN   NaN    NaN       1    NaN
2     NaN   NaN      NaN   NaN    NaN   NaN   NaN    NaN       1    NaN
3     NaN   NaN      NaN   NaN    NaN   NaN   NaN    NaN       1    NaN
4     NaN   NaN      NaN   9.0    0.0   NaN   NaN    9.0       1    9.0

[5 rows x 256 columns]
```

(b) The unit of analysis: membership of countries and territories in Conventional Intergovernmental Organizations (CIO).

[15]:
```
print("(c) There are", data.shape[1], "Variables in this dataset")
```

```
(c) There are 256 Variables in this dataset
```

[16]:
```
data.dtypes
```

[16]:
```
scode       object
ccode        int64
country     object
year         int64
uiareg       int64
             …
uar        float64
upeb       float64
udeac      float64
unesco       int64
unido      float64
Length: 256, dtype: object
```

(c) Most of the variables seem to be of the type int. This is because each variable is just a count of participating nations. A count is always discrete so hence int.

(d) Conceptualized: Includes all non-profit international organizations that have a widespread, geographically-balanced membership, management and policy-control.

(e) Operationalized: The rule applied here is that there should be members in at least 60 countries, or else in more than 30 countries provided that the distribution between continents is well-balanced.

(f) The strength of this measure is that it includes a large number of variables, making the data more widespread and informative. The weakness is that the measure requires the "distribution between continents to be well-balanced". However, there is no specific way to check for balance.

(g) One possible source of selection bias in the data set is not selecting enough countries to represent each continent. Each country has different relations and problems and leaving out specific ones can lead to findings totally different from what actually may be possible. In this case it will bias the results of analysis to be higher as only the largely participating countries are being considered.

[17]:
```
mean_val = (data[data['year'] == 1962]['ciob']).mean()
print("Mean:", mean_val)
```

```
Mean: 13.567796610169491
```

(h) The mean value represents the average of all non-profit international organizations that have a widespread, geographically-balanced membership, management and policy-control that each country may have.

End of HW 3

[ ]: