Live Coding - Module 2

Rachel Holman

From this website, we will obtain the data file and code book for the following three datasets:

- Gallup Poll # 1936-0053: Teachers' Oath/Government Loans for Farmers/Employers Insurance Contributions/Presidential Candidates [Roper #31087039]
- Study Record: Longstanding Methods Collection ABC News/Ipsos Poll: January 2023 [Roper #31120091]
- USIA Poll # 2000-I20068: Economic Conditions/Government Approval/Security/Civilian Rule/International Relations/US [Roper #31086002]

In [12]:
```
import numpy as np
import pandas as pd
import os
# set working directory
os.chdir("/Users/rachelholman/Desktop/MSDS/MSDS-SummerCourses/DS6001 - Applicat
```

In [11]:
```
# read in the csv file
gallup = pd.read_csv("USAIPO1936-0053.csv")
```

In [8]:
```
# .head specifies how many of the first few rows to show
# .T = transpose (so columns are shown as rows and rows are shown as columns)

gallup.head(3).T
```

Out[8]:

	0	1	2
form	NaN	NaN	NaN
state	Indiana	Illinois	Michigan
region	East Central	East Central	East Central
female	Male	Male	Male
age	NaN	NaN	NaN
class	Av+	Av+	Av
OCCUPATION1	Skilled workers	Skilled workers	Business
OCCUPATION2	NaN	NaN	NaN
OCCUPATION3	NaN	NaN	NaN
black	NaN	NaN	NaN
size	Urban	Urban	Urban
education	NaN	NaN	NaN
AGE_3WAY	NaN	NaN	NaN
AGE40	NaN	NaN	NaN
OCC8	Labor	Labor	Professional
prof	Not Professional	Not Professional	Professional
REGION4	Midwest	Midwest	Midwest
EDU_RECODE	NaN	NaN	NaN
VOTE_PRO	Landon	Landon	Landon
VOTE_RETRO	Hoover	Hoover	Hoover
PHONE_RECODE	NaN	NaN	NaN
CAR_RECODE	NaN	NaN	NaN
ballot	53	53	53
Q1	Yes	Yes	No
Q2	Yes	NaN	No
Q3	Yes	Yes	Yes
Q4A	Roosevelt	Landon	Landon
Q4B	Roosevelt	Landon	Landon
Q4C	Landon	Landon	Landon
Q5A	Landon	Landon	Landon
Q5B	Yes, voted for Hoover	Yes, voted for Hoover	Yes, voted for Hoover
farm	Non-Farm	Non-Farm	Non-Farm
SIZE3	Urban	Urban	Urban
urban	Urban	Urban	Urban
StPOAbrv	in	il	mi

	0	1	2
SOUTH11	Non-South	Non-South	Non-South
SOUTH11xBLACK	NaN	NaN	NaN
SOUTH12	Non-South	Non-South	Non-South
SOUTH12xBLACK	NaN	NaN	NaN
south	Non-South	Non-South	Non-South
SOUTHxBLACK	NaN	NaN	NaN
year	1936	1936	1936
WtPubFeas	NaN	NaN	NaN
WtVotFeas	NaN	NaN	NaN

In [13]:
```
# see information about each column in data
gallup.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5515 entries, 0 to 5514
Data columns (total 44 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   form            0 non-null      float64
 1   state           5515 non-null   object 
 2   region          5515 non-null   object 
 3   female          5514 non-null   object 
 4   age             5169 non-null   object 
 5   class           4619 non-null   object 
 6   OCCUPATION1     5513 non-null   object 
 7   OCCUPATION2     0 non-null      float64
 8   OCCUPATION3     0 non-null      float64
 9   black           2058 non-null   object 
 10  size            5514 non-null   object 
 11  education       0 non-null      float64
 12  AGE_3WAY        0 non-null      float64
 13  AGE40           0 non-null      float64
 14  OCC8            5513 non-null   object 
 15  prof            5513 non-null   object 
 16  REGION4         5515 non-null   object 
 17  EDU_RECODE      0 non-null      float64
 18  VOTE_PRO        5382 non-null   object 
 19  VOTE_RETRO      5460 non-null   object 
 20  PHONE_RECODE    0 non-null      float64
 21  CAR_RECODE      2054 non-null   object 
 22  ballot          5515 non-null   int64  
 23  Q1              4410 non-null   object 
 24  Q2              4640 non-null   object 
 25  Q3              4553 non-null   object 
 26  Q4A             5124 non-null   object 
 27  Q4B             4988 non-null   object 
 28  Q4C             4923 non-null   object 
 29  Q5A             5382 non-null   object 
 30  Q5B             5460 non-null   object 
 31  farm            5514 non-null   object 
 32  SIZE3           5514 non-null   object 
 33  urban           5514 non-null   object 
 34  StPOAbrv        5515 non-null   object 
 35  SOUTH11         5515 non-null   object 
 36  SOUTH11xBLACK   2058 non-null   object 
 37  SOUTH12         5515 non-null   object 
 38  SOUTH12xBLACK   2058 non-null   object 
 39  south           5515 non-null   object 
 40  SOUTHxBLACK     2058 non-null   object 
 41  year            5515 non-null   int64  
 42  WtPubFeas       2055 non-null   float64
 43  WtVotFeas       2051 non-null   float64
dtypes: float64(10), int64(2), object(32)
memory usage: 1.9+ MB
```

In [21]:
```
# see frequency table
gallup['Q3'].value_counts()
```

Out[21]:
```
Yes    3105
No     1448
Name: Q3, dtype: int64
```

In [15]:
```
gallup['Q5A'].value_counts()
```

Out[15]:
```
Roosevelt       3044
Landon          2102
Lemke            167
Thomas            55
Other party       14
Name: Q5A, dtype: int64
```

In [22]:
```
# 2 variable frequency table
pd.crosstab(gallup['Q5A'], gallup['Q3'])
```

Out[22]:

Q3	No	Yes
Q5A		
Landon	921	820
Lemke	41	109
Other party	6	6
Roosevelt	442	2048
Thomas	10	41

In [19]:
```
gallup['age'].value_counts()
```

Out[19]:
```
45-54 yrs          1274
35-44 yrs          1240
25-34 yrs          1093
55 yrs and over     978
21-24 yrs           582
17-20                 2
Name: age, dtype: int64
```

In [20]:
```
pd.crosstab(gallup['age'], gallup['Q3'])
```

Out[20]:

Q3	No	Yes
age		
17-20	0	1
21-24 yrs	122	337
25-34 yrs	264	651
35-44 yrs	303	735
45-54 yrs	361	705
55 yrs and over	298	514

In [31]:
```
# Now let's look at a current ABC News Poll
# Note: this is a strata dataset
abcnews = pd.read_stata('31120091.DTA')

abcnews.head(3).T
```

Out[31]:

	0	1	2
id	12000001	12000002	12000003
xspanish	english	english	english
complete	qualified	qualified	qualified
ppage	70	85	63
ppeduc5	bachelor s degree	master s degree or above	some college or associate degree
ppeducat	bachelors degree or higher	bachelors degree or higher	some college
ppgender	female	male	male
ppethm	white, non-hispanic	white, non-hispanic	white, non-hispanic
pphhsize	2	2	3
ppinc7	$75,000 to 99,999$	$100,000 to 149,999$	$50,000 to 74,999$
ppmarit5	now married	now married	now married
ppmsacat	metro	metro	non-metro
ppreg4	south	midwest	west
pprent	owned or being bought by you or someone in you...	owned or being bought by you or someone in you...	owned or being bought by you or someone in you...
ppstaten	texas	michigan	utah
ppworka	retired	retired	retired
ppemploy	not working	not working	working part-time
xparent	not parents of children 0-17 yo	not parents of children 0-17 yo	not parents of children 0-17 yo
q1_a	approve	disapprove	disapprove
q1_b	disapprove	disapprove	disapprove
q1_c	disapprove	disapprove	disapprove
q1_d	disapprove	disapprove	disapprove
q1_f	approve	disapprove	disapprove
q1_g	disapprove	disapprove	disapprove
q1_i	approve	disapprove	disapprove
q1_j	disapprove	disapprove	disapprove
q1_k	disapprove	disapprove	disapprove
q1_l	disapprove	disapprove	disapprove
q1_m	disapprove	disapprove	disapprove
q2_a	disapprove	disapprove	disapprove
q2_b	disapprove	approve	approve
q3	inappropriately	inappropriately	appropriately
q4	inappropriately	inappropriately	inappropriately

	0	1	2
q5	both were about the same	both were about the same	what joe biden did was more a serious concern
q6	about the right amount	too much	too much
qpid	a republican	an independent	a republican
abcage	65+	65+	50-64
contact	no, i am not willing to be interviewed	no, i am not willing to be interviewed	yes, i am willing to be interviewed
weight_p	0.6558	0.4981	0.7332

In [28]:
```
abcnews['q1_j'].value_counts()
```

Out[28]:
```
disapprove    358
approve       168
skipped         6
Name: q1_j, dtype: int64
```

In [29]:
```
pd.crosstab(abcnews['ppinc7'], abcnews['q1_j'])
```

Out[29]:

q1_j	skipped	approve	disapprove
ppinc7			
less than $10,000	0	2	15
$10,000$ **to** $24,999	2	10	24
$25,000$ **to** $49,999	2	13	49
$50,000$ **to** $74,999	1	22	65
$75,000$ **to** $99,999	0	27	50
$100,000$ **to** $149,999	0	38	63
$150,000 or more	1	56	92

In [33]:
```
# Now for some REALLY unpleasant data
widths = pd.read_csv('USIA poll - Sheet1.csv')
```

In [34]:
```
widths['Width']
```

Out[34]:
```
0      4
1      3
2      1
3      1
4      1
      ..
131    2
132    1
133    1
134    1
135    1
Name: Width, Length: 136, dtype: int64
```

In [37]:
```
usia = pd.read_fwf('i20068.dat', widths=widths['Width'], header=None)
```

In [38]:
```
usia
```

Out[38]:

	0	1	2	3	4	5	6	7	8	9	...	126	127	128	129	13
0	1	155.0	2.0	1.0	2.0	108.0	12.0	14.0	2.0	3.0	...	1.0	5.0	6.0	2.0	4
1	2	155.0	2.0	1.0	2.0	120.0	9.0	13.0	1.0	1.0	...	1.0	6.0	6.0	1.0	Na
2	3	155.0	2.0	1.0	2.0	102.0	14.0	NaN	2.0	2.0	...	1.0	5.0	1.0	1.0	Na
3	4	155.0	2.0	1.0	2.0	129.0	9.0	13.0	2.0	3.0	...	1.0	6.0	5.0	1.0	Na
4	5	155.0	2.0	1.0	1.0	1.0	13.0	9.0	2.0	2.0	...	1.0	6.0	7.0	1.0	Na
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2033	2248	283.0	2.0	2.0	2.0	102.0	6.0	11.0	2.0	2.0	...	1.0	2.0	17.0	2.0	5
2034	2249	283.0	2.0	2.0	8.0	NaN	98.0	NaN	8.0	8.0	...	1.0	5.0	17.0	2.0	4
2035	2250	283.0	2.0	2.0	9.0	NaN	98.0	NaN	3.0	3.0	...	1.0	3.0	17.0	2.0	3
2036	2251	255.0	1.0	2.0	2.0	163.0	9.0	6.0	3.0	4.0	...	1.0	1.0	17.0	2.0	3
2037		NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	Na

2038 rows × 136 columns