labassignment10

August 5, 2022

1 Lab Assignment 10: Exploratory Data Analysis, Part 1

1.1 DS 6001: Practice and Application of Data Science

1.1.1 Instructions

Please answer the following questions as completely as possible using text, code, and the results of code as needed. Format your answers in a Jupyter notebook. To receive full credit, make sure you address every part of the problem, and make sure your document is formatted in a clean and professional way.

In this lab, you will be working with the 2018 General Social Survey (GSS). The GSS is a sociological survey created and regularly collected since 1972 by the National Opinion Research Center at the University of Chicago. It is funded by the National Science Foundation. The GSS collects information and keeps a historical record of the concerns, experiences, attitudes, and practices of residents of the United States, and it is one of the most important data sources for the social sciences.

The data includes features that measure concepts that are notoriously difficult to ask about directly, such as religion, racism, and sexism. The data also include many different metrics of how successful a person is in his or her profession, including income, socioeconomic status, and occupational prestige. These occupational prestige scores are coded separately by the GSS. The full description of their methodology for measuring prestige is available here: http://gss.norc.org/Documents/reports/methodological-reports/MR122%20Occupational%20Prestige.pdf Here's a quote to give you an idea about how these scores are calculated:

> Respondents then were given small cards which each had a single occupational titles listed on it. Cards were in English or Spanish. They were given one card at a time in the preordained order. The interviewer then asked the respondent to "please put the card in the box at the top of the ladder if you think that occupation has the highest possible social standing. Put it in the box of the bottom of the ladder if you think it has the lowest possible social standing. If it belongs somewhere in between, just put it in the box that matches the social standing of the occupation."

The prestige scores are calculated from the aggregated rankings according to the method described above.

1.1.2 Problem 0

Import the following packages:

[1]:
```
import numpy as np
import pandas as pd
import sidetable
import weighted # this is a module of wquantiles, so type pip install␣
↪wquantiles or conda install wquantiles to get access to it
from scipy import stats
from sklearn import manifold
from sklearn import metrics
import prince
from pandas_profiling import ProfileReport
pd.options.display.max_columns = None
```

Then load the GSS data with the following code:

[2]:
```
%%capture
gss = pd.read_csv("https://github.com/jkropko/DS-6001/raw/master/localdata/
↪gss2018.csv",
                  encoding='cp1252', na_values=['IAP','IAP,DK,NA,uncodeable',␣
↪'NOT SURE',
                                                'DK', 'IAP, DK, NA, uncodeable',␣
↪'.a', "CAN'T CHOOSE"])
```

1.1.3 Problem 1

Drop all columns except for the following: * `id` - a numeric unique ID for each person who responded to the survey * `wtss` - survey sample weights * `sex` - male or female * `educ` - years of formal education * `region` - region of the country where the respondent lives * `age` - age * `coninc` - the respondent's personal annual income * `prestg10` - the respondent's occupational prestige score, as measured by the GSS using the methodology described above * `mapres10` - the respondent's mother's occupational prestige score, as measured by the GSS using the methodology described above * `papres10` -the respondent's father's occupational prestige score, as measured by the GSS using the methodology described above * `sei10` - an index measuring the respondent's socioeconomic status * `satjob` - responses to "On the whole, how satisfied are you with the work you do?" * `fechld` - agree or disagree with: "A working mother can establish just as warm and secure a relationship with her children as a mother who does not work." * `fefam` - agree or disagree with: "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family." * `fepol` - agree or disagree with: "Most men are better suited emotionally for politics than are most women." * `fepresch` - agree or disagree with: "A preschool child is likely to suffer if his or her mother works." * `meovrwrk` - agree or disagree with: "Family life often suffers because men concentrate too much on their work."

Then rename any columns with names that are non-intuitive to you to more intuitive and descriptive ones. Finally, replace the "89 or older" values of `age` with 89, and convert `age` to a float data type. [1 point]

[3]:
```
gss =␣
↪gss[['id','wtss','sex','educ','region','age','coninc','prestg10','mapres10','papres10','sei
        'satjob','fechld','fefam','fepol','fepresch','meovrwrk']]
```

```
gss = gss.rename(columns={'coninc':'income','sei10':'socioeconstat'})
gss.age = gss.age.replace("89 or older","89").astype(float)
```

1.1.4 Problem 2

Part a Use the `ProfileReport()` function to generate and embed an HTML formatted exploratory data analysis report in your notebook. Make sure that it includes a "Correlations" report along with "Overview" and "Variables". [1 point]

[4]:
```
profile = ProfileReport(gss,
                        title='Pandas Profiling Report',
                        html={'style':{'full_width':True}},
                        correlations={
        "pearson": {"calculate": True},
        "spearman": {"calculate": True},
        "kendall": {"calculate": True},
        "phi_k": {"calculate": True},
        "cramers": {"calculate": True}},
                       minimal=True)
profile.to_notebook_iframe()
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

<IPython.core.display.HTML object>
```

Part b Looking through the HTML report you displayed in part a, how many people in the data are from New England? [1 point]

124 people are from New England

Part c Looking through the HTML report you displayed in part a, which feature in the data has the highest number of missing values, and what percent of the values are missing for this feature? [1 point]

fepol has the highest number of values missing with 849, or 36.2% of the values.

Part d Looking through the HTML report you displayed in part a, which two distinct features in the data have the highest correlation? [1 point]

prestg10 and socieoconstat have the highest correlation

1.1.5 Problem 3

On a primetime show on a 24-hour cable news network, two unpleasant-looking men in suits sit across a table from each other, scowling. One says "This economy is failing the middle-class. The average American today is making less than \$48,000 a year." The other screams "Fake news! The typical American makes more than $55,000 a year!" Explain, using words and code, how the data

can support both of their arguments. Use the sample weights to calculate descriptive statistics that are more representative of the American adult population as a whole. [1 point]

[5]:
```
weighted.median(gss.income,gss.wtss) # Weighted Median
```

[5]:
```
47317.5
```

[6]:
```
gss_temp = gss.loc[~gss.income.isna()]
np.average(gss_temp['income'], weights=gss_temp.wtss)
# Average without NAs included increases the mean value
```

[6]:
```
55158.96280421564
```

The data can support both arugements becuase the first person is referring to the number that is calculated by the weighted mean and the second person is referring to the number calculted by the weighted average.

1.1.6 Problem 4

For each of the following parts, * generate a table that provides evidence about the relationship between the two features in the data that are relevant to each question, * interpret the table in words, * use a hypothesis test to assess the strength of the evidence in the table, * and provide a **specific and accurate** intepretation of the p-value associated with this hypothesis test beyond "significant or not".

Part a Is there a gender wage gap? That is, is there a difference between the average incomes of men and women? [2 points]

[7]:
```
gss.groupby('sex').agg({'income':'mean'})
```

[7]:

sex	income
female	47191.021452
male	53314.626187

Males average about 6,000 dollars more than the average Females

[8]:
```
income_men = gss.query("sex=='male'").income.dropna()
income_women = gss.query("sex=='female'").income.dropna()
```

[9]:
```
stats.ttest_ind(income_men, income_women, equal_var=False)
```

[9]:
```
Ttest_indResult(statistic=3.332824087618215, pvalue=0.0008749557881530089)
```

Here the p-value is about .000875, which is the probability that under the assumption that men and women make an equal amount, on average, that we could draw a sample with a difference between these two means of 3.333 or higher. Because this probability is lower than .05, we can reject the null hypothesis and conclude that there is a statisitically significant difference between men and women in terms of how much money they make.

Part b Are there different average values of occupational prestige for different levels of job satisfaction? [2 points]

[10]:
```
gss.groupby('satjob').agg({'prestg10':'mean'}).sort_values('prestg10')
```

[10]:
```
                     prestg10
satjob
a little dissat     40.946429
mod. satisfied      42.589984
very dissatisfied   43.000000
very satisfied      46.189320
```

As job satisfaction increases there seems to be no relationship with average value of occupational prestige

[11]:
```
stats.f_oneway(gss.query("satjob=='a little dissat'").prestg10.dropna(),
               gss.query("satjob=='mod. satisfied'").prestg10.dropna(),
               gss.query("satjob=='very dissatisfied'").prestg10.dropna(),
               gss.query("satjob=='very satisfied'").prestg10.dropna())
```

[11]:
```
F_onewayResult(statistic=12.205403153509732, pvalue=6.676686425029878e-08)
```

The p-value is very small, and much smaller than .05, so we reject the null hypothesis that the four groups of job satisfaction have the same average of occupational prestige.

1.1.7 Problem 5

Report the Pearson's correlation between years of education, socioeconomic status, income, occupational prestige, and a person's mother's and father's occupational prestige? Then perform a hypothesis test for the correlation between years of education and socioeconomic status and provide a **specific and accurate** intepretation of the p-value associated with this hypothesis test beyond "significant or not". [2 points]

[12]:
```
gss.columns
```

[12]:
```
Index(['id', 'wtss', 'sex', 'educ', 'region', 'age', 'income', 'prestg10',
       'mapres10', 'papres10', 'socioeconstat', 'satjob', 'fechld', 'fefam',
       'fepol', 'fepresch', 'meovrwrk'],
      dtype='object')
```

[13]:
```
cols = ['educ','socioeconstat','income','prestg10','mapres10','papres10']
gss.loc[:,cols].corr()
```

[13]:

	educ	socioeconstat	income	prestg10	mapres10	papres10
educ	1.000000	0.558169	0.389245	0.479933	0.269115	0.261417
socioeconstat	0.558169	1.000000	0.417210	0.835515	0.203486	0.210451
income	0.389245	0.417210	1.000000	0.340995	0.164881	0.171048
prestg10	0.479933	0.835515	0.340995	1.000000	0.189262	0.192180
mapres10	0.269115	0.203486	0.164881	0.189262	1.000000	0.235750

```
papres10       0.261417       0.210451  0.171048  0.192180  0.235750  1.000000
```

[14]:
```
gss_corr = gss[['educ', 'socioeconstat']].dropna()
stats.pearsonr(gss_corr['educ'], gss_corr['socioeconstat'])
```

[14]:
```
(0.5581686004626779, 3.719448810021532e-184)
```

The first number is the correlation coefficient, which is 0.56. The positive number means that the higher someones education, the higher their socioeconomic status, which is not surprising. The p-value is the second number, which is so small that it rounds to 0 over 16 decimal places. The p-value is the probability that a random sample could produce a correlation as extreme as 0.56 in either direction assuming that the correlation is 0 in the population. Because the p-value is so small, we reject the null hypothesis that these two features are uncorrelated and we conclude that there is a nonzero correlation between years of education and socioeconomic status.

1.1.8 Problem 6

Create a new categorical feature for age groups, with categories for 18-35, 36-49, 50-69, and 70 and older (see the module 8 notebook for an example of how to do this).

Then create a cross-tabulation in which the rows represent age groups and the columns represent responses to the statement that "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family." Rearrange the columns so that they are in the following order: strongly agree, agree, disagree, strongly disagree. Place row percents in the cells of this table.

Finally, use a hypothesis test that can tell use whether there is enough evidence to conclude that these two features have a relationship, and provide a specific and accurate intepretation of the p-value. [2 points]

[15]:
```
gss['age_cat'] = pd.cut(gss['age'], bins=[18, 35, 49, 69,np.inf],␣
↪labels=['18-35', '36-49', '50-69','70 and older'])
gss['fefam'] = gss['fefam'].astype('category').cat.
↪reorder_categories(['strongly agree','agree','disagree','strongly disagree'])
crosstab = (pd.crosstab(gss.age_cat, gss.fefam, normalize='index')*100).round(2)
crosstab
```

[15]:
```
fefam        strongly agree  agree  disagree  strongly disagree
age_cat
18-35                  4.07  13.74     48.35              33.84
36-49                  4.79  17.46     46.48              31.27
50-69                  4.63  20.85     48.07              26.45
70 and older          11.97  31.66     39.00              17.37
```

[16]:
```
stats.chi2_contingency(crosstab.values)
```

[16]:
```
(22.243414732165085,
 0.008138720105479042,
```

```
9,
array([[ 6.365 , 20.9275, 45.475 , 27.2325],
       [ 6.365 , 20.9275, 45.475 , 27.2325],
       [ 6.365 , 20.9275, 45.475 , 27.2325],
       [ 6.365 , 20.9275, 45.475 , 27.2325]]))
```

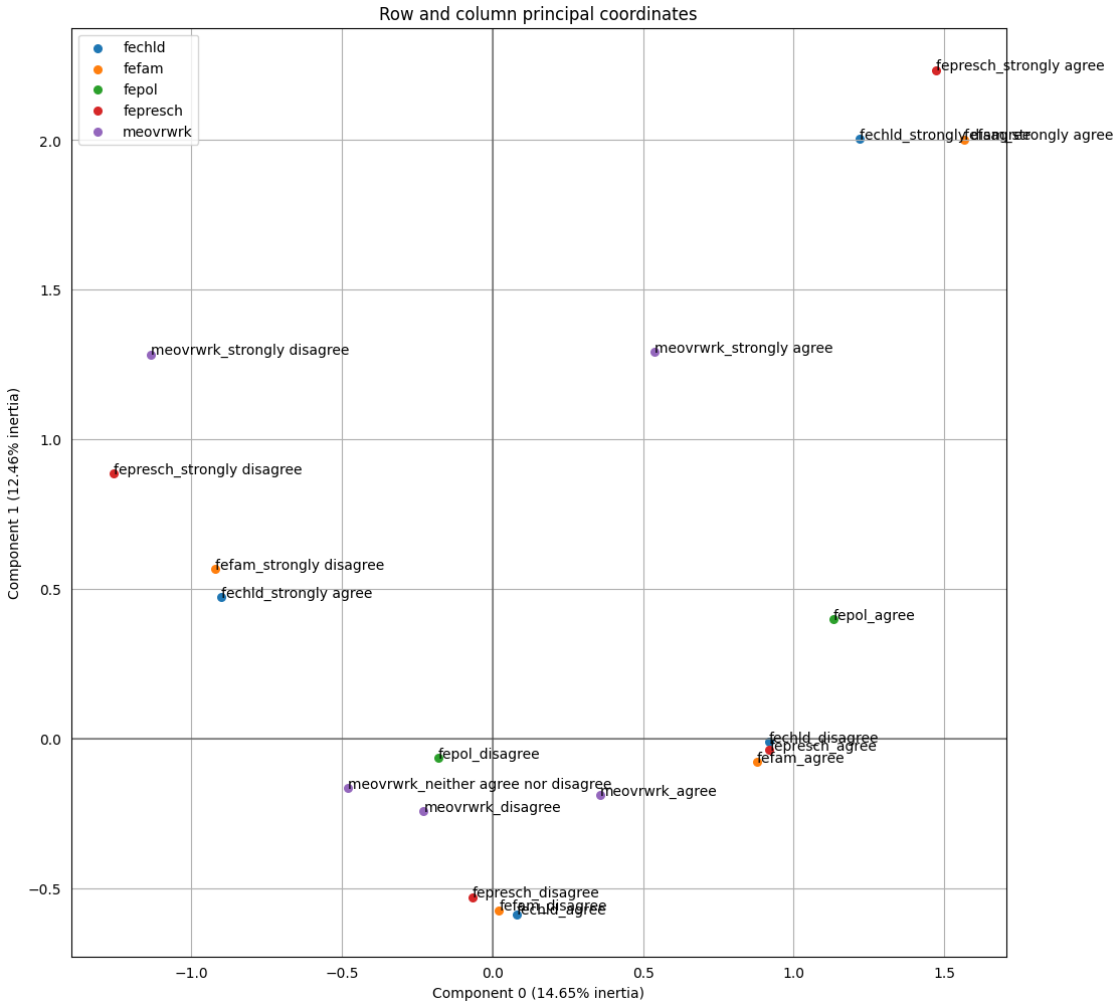
The p-value is the second value listed, 0.00814, which is very, very small and much less than .05. The p-value represents the probability that a cross-tab with row-by-row (or column-by-column) differences as extreme as the ones we see can be generated by a random sample if we assume that these two features are independent in the population so that the row percents should be constant across rows (and column percents should be constant across columns). Because the p-value is so small, we reject this null hypothesis and conclude that there is a statistically significant relationship between age category and strength of belief if the man should be the achiever outaide the house.

1.1.9 Problem 7

For this problem, you will conduct and interpret a correspondence analysis on the categorical features that ask respondents to state the extent to which they agree or disagree with the statements: * "A working mother can establish just as warm and secure a relationship with her children as a mother who does not work." * "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family." * "Most men are better suited emotionally for politics than are most women." * "A preschool child is likely to suffer if his or her mother works." * "Family life often suffers because men concentrate too much on their work."

Part a Conduct a correspondence analysis using the observed features listed above that measures two latent features. Plot the two latent categories for each category in each of the features used in the analysis. [2 points]

[17]:
```
beliefs = gss[['fechld', 'fefam','fepol', 'fepresch', 'meovrwrk']].dropna()
MCA = prince.MCA(n_components=2)
MCA = MCA.fit(beliefs)
ax = MCA.plot_coordinates(
     X=beliefs,
     ax=None,
     figsize=(12, 12),
     show_row_points=False,
     row_points_size=10,
     show_row_labels=False,
     show_column_points=True,
     column_points_size=30,
     show_column_labels=True,
     legend_n_cols=1
 )
```



Part b Display the latent features for every category in the observed features, sorted by the first latent feature. Describe in words what concept this feature is attempting to measure, and give the feature a name. [2 points]

[18]:
```
MCA.column_coordinates(beliefs).sort_values(0)
```

[18]:

	0	1
fepresch_strongly disagree	-1.258059	0.886697
meovrwrk_strongly disagree	-1.135403	1.283823
fefam_strongly disagree	-0.922035	0.566818
fechld_strongly agree	-0.901119	0.472175
meovrwrk_neither agree nor disagree	-0.480746	-0.163824
meovrwrk_disagree	-0.228690	-0.242580
fepol_disagree	-0.180400	-0.063738
fepresch_disagree	-0.067886	-0.529262
fefam_disagree	0.022159	-0.572466

```
fechld_agree                    0.080484 -0.586393
meovrwrk_agree                  0.358280 -0.187028
meovrwrk_strongly agree         0.536780  1.292002
fefam_agree                     0.878984 -0.076597
fechld_disagree                 0.918041 -0.010321
fepresch_agree                  0.919993 -0.036425
fepol_agree                     1.131106  0.399636
fechld_strongly disagree        1.218706  2.005412
fepresch_strongly agree         1.474181  2.233969
fefam_strongly agree            1.564723  2.002670
```

The feature is attemping to measure traditional views vs nontradition view of a man being the bread winner and a women tending to the childern and house. (traditional)

Part c We can use the results of the MCA model to conduct some cool EDA. For one example, follow these steps:

1. Use the `.row_coordinates()` method to calculate values of the latent feature for every row in the data you passed to the MCA in part a. Extract the first column and store it in its own dataframe.
2. To join it with the full, cleaned GSS data based on row numbers (instead of on a primary key), use the `.join()` method. For example, if we named the cleaned GSS data `gss_clean` and if we named the dataframe in step 1 `latentfeature`, we can type

```
gss_clean = gss_clean.join(latentfeature, how="outer")
```

3. Create a cross-tabuation with age categories (that you constructed in problem 5) in the rows and sex in the columns. Instead of a frequency, place the mean value of the latent feature in the cells.

What does this table tell you about the relationship between sex, age, and the latent feature? [2 points]

[19]:
```
traditional_index = MCA.row_coordinates(beliefs).loc[:,0]
```

[20]:
```
gss = gss.join(traditional_index, how="outer")
gss = gss.rename(columns={0:'fam_work_index'})
```

[23]:
```
crosstab = (pd.crosstab(gss.age_cat, gss.sex,values=gss.traditional_index,␣
 ↪aggfunc='mean').round(2))
crosstab
```

[23]:
```
sex           female  male
age_cat
18-35          -0.24 -0.00
36-49          -0.14 -0.00
50-69          -0.13  0.22
70 and older    0.13  0.47
```

This table tells us that males are either in the middle or display traditional. Furthermore, it shows that both male and female traditional beielfs increase the older the person's age.