Lab Assignment 10: Exploratory Data Analysis, Part 1

DS 6001: Practice and Application of Data Science

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.

Problem 0

Import the following packages:

In [1]:
```
import numpy as np
import pandas as pd
import sidetable
import weighted # this is a module of wquantiles, so type pip install wquantile
from scipy import stats
from sklearn import manifold
from sklearn import metrics
import prince
from ydata_profiling import ProfileReport
pd.options.display.max_columns = None
```

Then load the GSS data with the following code:

In [2]:
```
%%capture
gss = pd.read_csv("https://github.com/jkropko/DS-6001/raw/master/localdata/gss2
                  encoding='cp1252', na_values=['IAP','IAP,DK,NA,uncodeable', 'N
                                                'DK', 'IAP, DK, NA, uncodeable',
```

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]

In [3]:

```
gss = gss[['id', 'wtss','sex','educ','region','age','coninc','prestg10','mapres
            'papres10','sei10','satjob','fechld','fefam','fepol','fepresch','mec
gss = gss.rename({'wtss': 'sample_weights',
                  'educ':'years_of_educ',
                  'coninc':'annual_income',
                  'prestg10':'prestige',
                  'mapres10':'mom_prestige',
                  'papres10':'dad_prestige',
                  'sei10':'socioeco_status',
                  'satjob':'job_satisfaction',
                  'fechld':'workingmom_bondswith_children',
                  'fefam':'women_stayathome',
                  'fepol':'men_suited4politics',
                  'fepresch':'toddler_misses_workingmom',
                  'meovrwrk':'men_tooworkfocused'}, axis=1)
```

In [4]:

```
gss['age'] = np.where(gss['age'] == "89 or older", 89, gss['age'])
gss['age'] = pd.to_numeric(gss['age'], downcast="float")
gss
```

Out[4]:

	id	sample_weights	sex	years_of_educ	region	age	annual_income	prestige	n
0	1	2.357493	male	14.0	new england	43.0	NaN	47.0	
1	2	0.942997	female	10.0	new england	74.0	22782.5000	22.0	
2	3	0.942997	male	16.0	new england	42.0	112160.0000	61.0	
3	4	0.942997	female	16.0	new england	63.0	158201.8412	59.0	
4	5	0.942997	male	18.0	new england	71.0	158201.8412	53.0	
...	...	...	...	...	...	...	...	...	
2343	2344	0.471499	female	12.0	new england	37.0	NaN	47.0	
2344	2345	0.942997	female	12.0	new england	75.0	22782.5000	28.0	
2345	2346	0.942997	female	12.0	new england	67.0	70100.0000	40.0	
2346	2347	0.942997	male	16.0	new england	72.0	38555.0000	47.0	
2347	2348	0.471499	female	12.0	new england	79.0	NaN	33.0	

2348 rows × 17 columns

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]

In [5]:

```
profile = ProfileReport(gss,
                        title='Pandas Profiling Report',
                        html={'style':{'full_width':True}},
                       minimal=False)
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]
```

Overview

Dataset statistics

Number of variables	17
Number of observations	2348
Missing cells	6276
Missing cells (%)	15.7%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	302.8 KiB
Average record size in memory	132.1 B

Variable types

Numeric	9
Categorical	8

Alerts

`years_of_educ` is highly overall correlated with `socioeco_status`	**High correlation**
`prestige` is highly overall correlated with `socioeco_status`	**High correlation**
`socioeco_status` is highly overall correlated with `years_of_educ` and 1 other fields (years_of_educ, prestige)	**High correlation**

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

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]

The column recording how much people agree or disagree with: "Most men are better suited emotionally for politics than are most women." has **849** missing values. This means **36.2%** of values are missing for this feature

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]

With a correlation value of **0.824**, the two distinct features that are most correlated are prestige and socioeconomic status.

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]

Without adjusting for weights, the average annual income for Americans in this study is found to be 49,973.96. This however is skewed by the extremely rich and extremely poor people, so by computing the average annual salary with the top and bottom 5% of earners removed, we see an average annual income of 46,657.48. This affirms the argument that "the average American today is making less than $48,000 a year."

In [6]:
```
round(gss.annual_income.mean(),2)
```

Out[6]:
```
49973.96
```

In [7]:
```
gss_temp = gss.loc[~gss.annual_income.isna()]
round(stats.trim_mean(gss_temp.annual_income, .05),2)
```

Out[7]:
```
46657.48
```

After adjusting for the sample weights, we see that the average annual income for Americans is 55,158.96 which affirms the argument that "the typical American makes more than $55,000 a year."

In [8]:
```
gss_temp = gss.loc[~gss.annual_income.isna()]
round(np.average(gss_temp['annual_income'], weights=gss_temp.sample_weights),2)
```

Out[8]:
```
55158.96
```

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]

In [9]:
```
gss.groupby('sex').agg({'annual_income':'mean'}).round(2)
```

Out[9]:

sex	annual_income
female	47191.02
male	53314.63

Table interpretation: We see that the average annual income for females is over $6,000 less than the average annual income for males. We will test our hypthesis that men and women have an equal average annual salary using an independent samples t-test below.

Hypothesis test: Let the mean annual income for females be μ_f and the mean annual income for men be μ_m. We will run an independent samples t-test in which our null and alternative hypotheses are:

$$H_0 : \mu_f = \mu_m$$

$$H_A : \mu_f \neq \mu_m$$

In [10]:
```
income_men = gss.query("sex=='male'").annual_income.dropna()
income_women = gss.query("sex=='female'").annual_income.dropna()

stats.ttest_ind(income_men, income_women, equal_var=False)
```

Out[10]:
```
Ttest_indResult(statistic=3.332824087618215, pvalue=0.000874955788153009)
```

Here the p-value is about 0.000875, which is the probability that under the assumption that men and women are paid equally, on average, that we could draw a sample with a difference between these two means of 3.3328 or higher. Because this probability is lower than .05, we can reject the null hypothesis that men and women are paid the same, and conclude that there is a statisitically significant difference between men and women in terms of average annual income.

Part b

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

In [11]:
```
gss.groupby('job_satisfaction').agg({'prestige':'mean'}).round(2)
```

Out[11]:

job_satisfaction	prestige
a little dissat	40.95
mod. satisfied	42.59
very dissatisfied	43.00
very satisfied	46.19

Table interpretation: We see that the average values of occupational prestige for the different levels of job satisfaction are all in the low to mid 40s. Those who are "very satisfied" with their jobs seem to have the most occupational prestege on average while those "a little dissatisfied" have the least. To see if there is a significant difference between any of these groups we will test below.

Hypothesis test: Let μ_{ld} be the average prestige for those a little dissatisfied with their jobs, μ_{ms} be the average prestige for those moderately satisfied with their jobs, μ_{vd} be the average prestige for those very dissatisfied with their jobs, and μ_{vs} be the average prestige for those very satisfied with their jobs. We will run an ANOVA test in which our null hypothesis is:

$$H_0 : \mu_{ld} = \mu_{ms} = \mu_{vd} = \mu_{vs}$$

In [12]:
```
stats.f_oneway(gss.query("job_satisfaction=='a little dissat'").age.dropna(),
               gss.query("job_satisfaction=='mod. satisfied'").age.dropna(),
               gss.query("job_satisfaction=='very dissatisfied'").age.dropna(),
               gss.query("job_satisfaction=='very satisfied'").age.dropna())
```

Out[12]:
```
F_onewayResult(statistic=19.15044841301385, pvalue=3.231685973302454e-12)
```

The p-value is very small (about .000000000003), and much smaller than .05, so we reject the null hypothesis that the four groups of job satisfaction have the same average values of occupational prestige.

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]

In [13]:
```
gss.loc[:, ['years_of_educ', 'socioeco_status', 'annual_income',
        'prestige', 'mom_prestige', 'dad_prestige']].corr()
```

Out[13]:

	years_of_educ	socioeco_status	annual_income	prestige	mom_prestige	dac
years_of_educ	1.000000	0.558169	0.389245	0.479933	0.269115	
socioeco_status	0.558169	1.000000	0.417210	0.835515	0.203486	
annual_income	0.389245	0.417210	1.000000	0.340995	0.164881	
prestige	0.479933	0.835515	0.340995	1.000000	0.189262	
mom_prestige	0.269115	0.203486	0.164881	0.189262	1.000000	
dad_prestige	0.261417	0.210451	0.171048	0.192180	0.235750	

We notice that prestige and socioeconomic status seem to have the highest correlation while annual income and mom's prestige have the least correlation.

In [14]:
```
gss_corr = gss[['years_of_educ', 'socioeco_status']].dropna()
stats.pearsonr(gss_corr['years_of_educ'], gss_corr['socioeco_status'])
```

Out[14]:
```
PearsonRResult(statistic=0.5581686004626786, pvalue=3.7194488100284597e-184)
```

The first number is the correlation coefficient, which is 0.558. The positive number means that the more years of education someone has, the higher their socioeconomic status tends to be. The p-value is the second number, which is so small that it rounds to 0 over 183 decimal places. The p-value is the probability that a random sample could produce a correlation as extreme as 0.558 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 the number of years of education and socioeconomic status.

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]

In [15]:
```
gss['age_groups'] = pd.cut(gss.age,
                           bins=[17,35,49,69,300],
                           labels=("18-35", "36-49", "50-69", "70+"))
gss[['age', 'age_groups']]
```

Out[15]:

	age	age_groups
0	43.0	36-49
1	74.0	70+
2	42.0	36-49
3	63.0	50-69
4	71.0	70+
...	...	...
2343	37.0	36-49
2344	75.0	70+
2345	67.0	50-69
2346	72.0	70+
2347	79.0	70+

2348 rows × 2 columns

In [16]:
```
gss['women_stayathome'] = gss['women_stayathome'].astype('category').cat.reorde
                                                                     'agree',
                                                                     'disagre
                                                                     'strongl
(pd.crosstab(gss.age_groups, gss.women_stayathome, normalize='index')*100).rour
```

Out[16]:

women_stayathome	strongly agree	agree	disagree	strongly disagree
age_groups				
18-35	3.94	14.04	47.54	34.48
36-49	4.79	17.46	46.48	31.27
50-69	4.63	20.85	48.07	26.45
70+	11.97	31.66	39.00	17.37

Hypothesis test: We will use a chi square hypothesis test where our null hypothesis is: H_0 : age group and 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." are not significantly linearly related.

In [17]:
```
crosstab = pd.crosstab(gss.age_groups, gss.women_stayathome)

stats.chi2_contingency(crosstab.values)
```

Out[17]:
```
Chi2ContingencyResult(statistic=69.24381761791811, pvalue=2.1419004733989943e-
11, dof=9, expected_freq=array([[ 23.23016905,  81.56957087, 186.89726918, 11
4.3029909 ],
       [ 20.31209363,  71.32314694, 163.42002601,  99.94473342],
       [ 29.63849155, 104.07152146, 238.45513654, 145.83485046],
       [ 14.81924577,  52.03576073, 119.22756827,  72.91742523]]))
```

The p-value is the second value listed, `3.21e-12`, which is 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 groups and 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."

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]

In [18]:
```
gss_cat = gss[['workingmom_bondswith_children',
               'women_stayathome',
               'men_suited4politics',
               'toddler_misses_workingmom',
```

```
                  'men_tooworkfocused']].dropna()
gss_cat
```

Out[18]:

	workingmom_bondswith_children	women_stayathome	men_suited4politics	toddler_misse
0	strongly agree	disagree	agree	s
2	strongly agree	disagree	disagree	
3	agree	disagree	disagree	
5	strongly agree	disagree	disagree	
8	disagree	strongly disagree	disagree	
...	...	...	...	
2341	disagree	strongly agree	agree	
2343	disagree	strongly disagree	disagree	s
2344	strongly agree	disagree	disagree	
2346	disagree	agree	disagree	
2347	strongly disagree	strongly agree	disagree	

1454 rows × 5 columns

In [19]:
```
mca = prince.MCA(
     n_components=2
)
mca = mca.fit(gss_cat)
```

In [20]:
```
mca.row_coordinates(gss_cat)
```

Out[20]:

	0	1
0	-0.202209	0.338284
2	-0.423361	-0.316907
3	-0.195576	-0.648699
5	-0.240092	-0.298095
8	0.341539	0.091187
...	...	...
2341	1.219019	0.567443
2343	-0.521777	0.384970
2344	-0.423361	-0.316907
2346	1.076900	0.642131
2347	1.440615	2.529641

1454 rows × 2 columns

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]

In [21]:
```
output = mca.column_coordinates(gss_cat).sort_values([0])
output.columns = ['agreement_index', 'survey_qs']
output
```

Out[21]:

	agreement_index	survey_qs
toddler_misses_workingmom_strongly disagree	-1.258059	0.886697
men_tooworkfocused_strongly disagree	-1.135403	1.283834
women_stayathome_strongly disagree	-0.922035	0.566805
workingmom_bondswith_children_strongly agree	-0.901119	0.472182
men_tooworkfocused_neither agree nor disagree	-0.480746	-0.163827
men_tooworkfocused_disagree	-0.228690	-0.242582
men_suited4politics_disagree	-0.180400	-0.063734
toddler_misses_workingmom_disagree	-0.067886	-0.529257
women_stayathome_disagree	0.022160	-0.572473
workingmom_bondswith_children_agree	0.080483	-0.586393
men_tooworkfocused_agree	0.358280	-0.187027
men_tooworkfocused_strongly agree	0.536780	1.292003
women_stayathome_agree	0.878984	-0.076577
workingmom_bondswith_children_disagree	0.918041	-0.010325
toddler_misses_workingmom_agree	0.919993	-0.036433
men_suited4politics_agree	1.131107	0.399614
workingmom_bondswith_children_strongly disagree	1.218706	2.005388
toddler_misses_workingmom_strongly agree	1.474181	2.233954
women_stayathome_strongly agree	1.564723	2.002720

This latent features is attempting to measure how much one agrees with the statments. This can be called an Agreement Index. The more negative the values are on the Agreement Index, the more one disagrees to the statements while more positive values relate to agreement.

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]

In [22]:
```
latentfeature = mca.row_coordinates(gss_cat)[[0]]
latentfeature.columns = ['agreement_index']
latentfeature
```

Out[22]:

	agreement_index
0	-0.202209
2	-0.423361
3	-0.195576
5	-0.240092
8	0.341539
...	...
2341	1.219019
2343	-0.521777
2344	-0.423361
2346	1.076900
2347	1.440615

1454 rows × 1 columns

In [23]:
```
gss = gss.join(latentfeature, how="outer")
gss
```

Out [23]:

	id	sample_weights	sex	years_of_educ	region	age	annual_income	prestige	n
0	1	2.357493	male	14.0	new england	43.0	NaN	47.0	
1	2	0.942997	female	10.0	new england	74.0	22782.5000	22.0	
2	3	0.942997	male	16.0	new england	42.0	112160.0000	61.0	
3	4	0.942997	female	16.0	new england	63.0	158201.8412	59.0	
4	5	0.942997	male	18.0	new england	71.0	158201.8412	53.0	
...	...	...	...	...	...	...	...	...	
2343	2344	0.471499	female	12.0	new england	37.0	NaN	47.0	
2344	2345	0.942997	female	12.0	new england	75.0	22782.5000	28.0	
2345	2346	0.942997	female	12.0	new england	67.0	70100.0000	40.0	
2346	2347	0.942997	male	16.0	new england	72.0	38555.0000	47.0	
2347	2348	0.471499	female	12.0	new england	79.0	NaN	33.0	

2348 rows × 19 columns

In [24]:

```
pd.crosstab(gss.age_groups, gss.sex,
            values=gss.agreement_index, aggfunc='mean')
```

Out [24]:

sex age_groups	female	male
18-35	-0.241140	-0.003774
36-49	-0.137001	-0.000686
50-69	-0.125059	0.222532
70+	0.129256	0.473005

Women have much smaller average latent values across every age group than men. In other words, women's Agreement Indexes are lower than men's at every agre group so women disagree with the statments more on average than men. Men agree with the statements more than women at every age group.