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 wquant
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
```

```
/var/folders/ds/qp3gbx7n3tz0738b8w4wxs580000gn/T/ipykernel_1708/3521480416.p
y:9: DeprecationWarning: `import pandas_profiling` is going to be deprecated
by April 1st. Please use `import ydata_profiling` instead.
  from pandas_profiling import ProfileReport
```

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/g
                  encoding='cp1252', na_values=['IAP','IAP,DK,NA,uncodeable',
                                                'DK', 'IAP, DK, NA, uncodeabl
```

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",
    "mapres10",
    "papres10",
    "sei10",
    "satjob",
    "fechld",
    "fefam",
    "fepol",
    "fepresch",
    "meovrwrk"]]

gss = gss.rename({"wtss": "weight",
    "educ": "education",
    "coninc": "income",
    "prestg10": "prestige_score",
    "mapres10": "mother_prestige_score",
    "papres10": "father_prestige_score",
    "sei10": "socioeconomic_status",
    "satjob": "job_satisfaction",
    "fechld": "working_mother_relationship",
    "fefam": "family_gender_roles",
    "fepol": "political_gender_roles",
    "fepresch": "preschool_working_mother",
    "meovrwrk": "overwork_relationship"}, axis = 1)

gss["age"] = gss["age"].replace({"89 or older": 89})
gss["age"] = gss["age"].astype('float64')
gss
```

Out[3]:

	id	weight	sex	education	region	age	income	prestige_score
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 [4]:
```
profile = ProfileReport(gss,
                        title = "General Social Survey",
                        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

Brought to you by YData (https://ydata.ai/?utm_source=opensource&utm_medium=ydataprofiling&utm_campaign=report)

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	312.0 KiB
Average record size in memory	136.1 B

Variable types

Numeric	9
Categorical	8

Alerts

`education` is highly overall correlated with `socioeconomic_status`	**High correlation**
`prestige_score` is highly overall correlated with `socioeconomic_status`	**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 political_gender_roles (fepol) feature has the highest number of missing values (36.2%).

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]

The features with the highest correlation are socioeconomic_status (sei10) and prestige_score (prestg10).

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]

In [5]:
```
gss.income.median()
```

Out[5]:
```
38555.0
```

In [6]:
```
weighted.median(gss.income, gss.weight)
```

Out[6]:
```
47317.5
```

In [7]:
```
gss_temp = gss.loc[~gss.income.isna()]
np.average(gss_temp['income'], weights=gss_temp.weight)
```

Out[7]:
```
55158.96280421564
```

Taking the weighted median of gss.income allows one to make the claim that the average American makes less than `$48,000` per year (the unweighted median also suggests this), while the weighted mean suggests the average American makes more than `$55,000` per year.

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 [8]:
```
gss.groupby('sex').agg({'income':'mean'}).round(2)
```

Out[8]:

sex	income
female	47191.02
male	53314.63

According to the GSS data, there is a gender wage gap of over $6000.

Part b

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

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

Out[9]:

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

According to the GSS data, there are different average values of occupational prestige for different levels of job satisfaction. However, it is worth noting that job satisfaction does not linearly correlate to occupational prestige (e.g. very dissatisfied jobs have a higher average prestige than those moderately satisfied or a little dissatisfied).

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 [10]:
```
gss[['education', 'income',
       'prestige_score', 'mother_prestige_score', 'father_prestige_score',
       'socioeconomic_status']].corr()
```

Out[10]:

	education	income	prestige_score	mother_prestige_score	f
education	1.000000	0.389245	0.479933	0.269115	
income	0.389245	1.000000	0.340995	0.164881	
prestige_score	0.479933	0.340995	1.000000	0.189262	
mother_prestige_score	0.269115	0.164881	0.189262	1.000000	
father_prestige_score	0.261417	0.171048	0.192180	0.235750	
socioeconomic_status	0.558169	0.417210	0.835515	0.203486	

In [11]:
```
gss_corr = gss[['education', 'socioeconomic_status']].dropna()
stats.pearsonr(gss_corr['education'], gss_corr['socioeconomic_status'])
```

Out[11]:
```
PearsonRResult(statistic=0.5581686004626788, pvalue=3.719448810025752e-184)
```

The extremely low p-value ($ < 0.05$) means we can reject the null hypothesis that education and socieconomic status are uncorrelated, and assert that there is a correlation of 0.56 between these features.

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 [12]:
```
gss["age_group"] = pd.cut(gss["age"], bins=[18,35,49,69,89], labels=("18-35"
crosstab = (pd.crosstab(gss.age_group, gss.family_gender_roles, normalize='i
'agree', 'disagree', 'strongly disagree']]
crosstab
```

Out[12]:

family_gender_roles	strongly agree	agree	disagree	strongly disagree
age_group				
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

In [13]:
```
stats.chi2_contingency(crosstab.values)
```

Out[13]:
```
Chi2ContingencyResult(statistic=22.243414732165085, pvalue=0.00813872010547
9037, dof=9, expected_freq=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 low p-value ($ < 0.05$) means we can reject the null hypothesis that age group and response to traditional gender roles in the family are uncorrelated, and assert that there is a statistically significant relationship between these features.

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 [36]:
```
gss_stmnt = gss[['working_mother_relationship', 'family_gender_roles',
       'political_gender_roles', 'preschool_working_mother',
       'overwork_relationship']].dropna()
```

In [37]:
```
mca = prince.MCA(n_components=2)
mca = mca.fit(gss_stmnt)
```

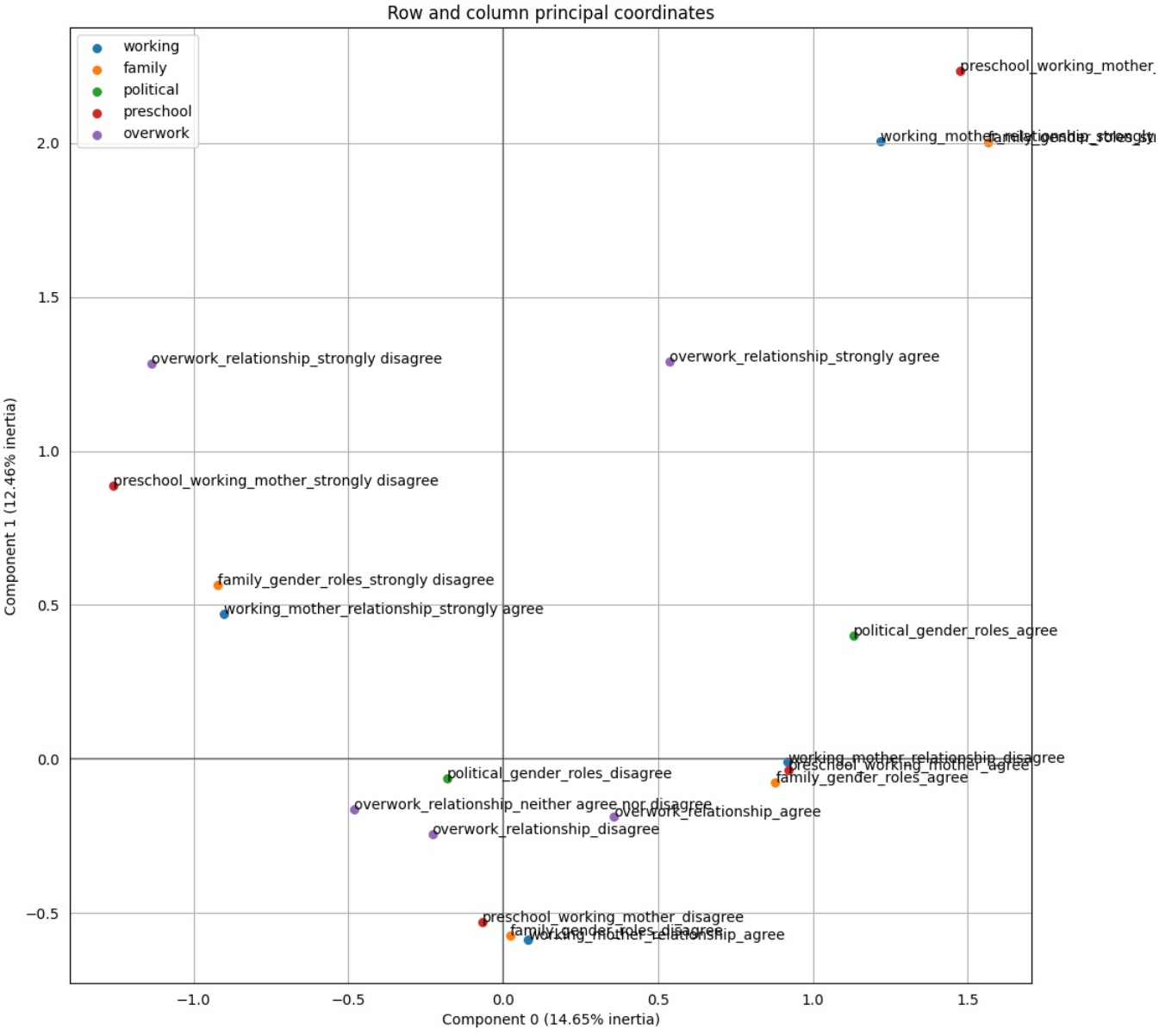
In [38]:
```
ax = mca.plot_coordinates(
     X=gss_stmnt,
     ax=None,
     figsize=(12, 12),
     show_row_points=False,
     row_points_size=10,
     show_row_labels=False,
     show_column_points=True,
     column_points_size=30,
     show_column_labels=True,
     legend_n_cols=1
 )
ax.get_figure().savefig('mca_coordinates.png')
```

```
/opt/miniconda3/lib/python3.12/site-packages/prince/mca.py:121: FutureWarnin
g: Series.__getitem__ treating keys as positions is deprecated. In a future
version, integer keys will always be treated as labels (consistent with Data
Frame behavior). To access a value by position, use `ser.iloc[pos]`
  ax.annotate(label, (x[mask][i], y[mask][i]))
```

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 [39]:
```
mca.column_coordinates(gss_stmnt).sort_values(0)
```


Row and column principal coordinates
working
family
political
preschool
overwork
preschool_working_mother
overwork_relationship_strongly disagree
overwork_relationship_strongly agree
preschool_working_mother_strongly disagree
family_gender_roles_strongly disagree
working_mother_relationship_strongly agree
political_gender_roles_agree
working_mother_relationship_disagree
preschool_working_mother_agree
family_gender_roles_agree
political_gender_roles_disagree
overwork_relationship_neither agree nor disagree
overwork_relationship_agree
overwork_relationship_disagree
preschool_working_mother_disagree
family_gender_roles_disagree
working_mother_relationship_agree
Component 1 (12.46% inertia)
Component 0 (14.65% inertia)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Out[39]:

	0	1
preschool_working_mother_strongly disagree	-1.258060	0.886698
overwork_relationship_strongly disagree	-1.135403	1.283828
family_gender_roles_strongly disagree	-0.922035	0.566810
working_mother_relationship_strongly agree	-0.901118	0.472179
overwork_relationship_neither agree nor disagree	-0.480746	-0.163825
overwork_relationship_disagree	-0.228690	-0.242581
political_gender_roles_disagree	-0.180400	-0.063736
preschool_working_mother_disagree	-0.067886	-0.529259
family_gender_roles_disagree	0.022159	-0.572468
working_mother_relationship_agree	0.080483	-0.586394
overwork_relationship_agree	0.358280	-0.187027
overwork_relationship_strongly agree	0.536780	1.292004
family_gender_roles_agree	0.878984	-0.076587
working_mother_relationship_disagree	0.918040	-0.010323
preschool_working_mother_agree	0.919993	-0.036429
political_gender_roles_agree	1.131106	0.399627
working_mother_relationship_strongly disagree	1.218706	2.005408
preschool_working_mother_strongly agree	1.474181	2.233956
family_gender_roles_strongly agree	1.564724	2.002692

One end of this feature strongly disagrees with the statements regarding traditional gender roles, while the other end strongly agrees with them. As such, I would call this feature a measurement of "adherence to traditional gender roles".

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 [40]:
```
latentfeature = pd.DataFrame(mca.row_coordinates(gss_stmnt)[0])
```

In [42]:
```
gss = gss.join(latentfeature, how="outer")
gss
```

Out[42]:

	id	weight	sex	education	region	age	income	prestige_score
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 [47]:
```
crosstab = pd.crosstab(gss.age_group, gss.sex, values = gss[0], aggfunc='mea
crosstab
```

Out[47]:

sex	female	male
age_group		
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

The relatively low (as in, close to 0) values in the table suggest that the age/sex demographics do not on average lean too heavily one way or the other in adherence to traditional gender roles. Males ages 18-49 in particular seem to be evenly split on the subject. The demographic that most rejects traditional gender roles is female ages 18-35, and the demographic that most adheres to traditional gender roles is males older than 50 (and especially older than 70).

In [ ]: