labassignment12

April 29, 2023

1 Lab Assignment 12: Interactive Visualizations

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.

1.2 Problem 0

Import the following libraries:

```
[1]: import numpy as np
     import pandas as pd
     import plotly.graph_objects as go
     import plotly.express as px
     import plotly.graph_objects as go
     import plotly.figure_factory as ff
     import dash
     from jupyter_dash import JupyterDash
     import dash_core_components as dcc
     import dash_html_components as html
     from dash.dependencies import Input, Output
     external_stylesheets = ['https://codepen.io/chriddyp/pen/bWLwgP.css']
```

For this lab, we will be working with the 2019 General Social Survey one last time.

```
[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"]
     )
```

Here is code that cleans the data and gets it ready to be used for data visualizations:

```
[3]: mycols = ['id', 'wtss', 'sex', 'educ', 'region', 'age', 'coninc',
               'prestg10', 'mapres10', 'papres10', 'sei10', 'satjob',
               'fechld', 'fefam', 'fepol', 'fepresch', 'meovrwrk']
     gss_clean = gss[mycols]
     gss_clean = gss_clean.rename({'wtss':'weight',
                                   'educ':'education',
                                   'coninc':'income',
                                   'prestg10':'job_prestige',
                                   'mapres10':'mother_job_prestige',
                                   'papres10':'father_job_prestige',
                                   'sei10':'socioeconomic_index',
                                   'fechld':'relationship',
                                   'fefam':'male_breadwinner',
                                   'fehire':'hire_women',
                                   'fejobaff':'preference_hire_women',
                                   'fepol':'men_bettersuited',
                                   'fepresch':'child_suffer',
                                   'meovrwrk':'men_overwork'},axis=1)
     gss_clean.age = gss_clean.age.replace({'89 or older':'89'})
     gss_clean.age = gss_clean.age.astype('float')
```

The `gss_clean` dataframe now contains the following features:

- `id` - a numeric unique ID for each person who responded to the survey
- `weight` - survey sample weights
- `sex` - male or female
- `education` - years of formal education
- `region` - region of the country where the respondent lives
- `age` - age
- `income` - the respondent's personal annual income
- `job_prestige` - the respondent's occupational prestige score, as measured by the GSS using the methodology described above
- `mother_job_prestige` - the respondent's mother's occupational prestige score, as measured by the GSS using the methodology described above
- `father_job_prestige` -the respondent's father's occupational prestige score, as measured by the GSS using the methodology described above
- `socioeconomic_index` - an index measuring the respondent's socioeconomic status
- `satjob` - responses to "On the whole, how satisfied are you with the work you do?"
- `relationship` - 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."
- `male_breadwinner` - 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."
- `men_bettersuited` - agree or disagree with: "Most men are better suited emotionally for politics than are most women."
- `child_suffer` - agree or disagree with: "A preschool child is likely to suffer if his or her mother works."
- `men_overwork` - agree or disagree with: "Family life often suffers because men concentrate too much on their work."

1.3 Problem 1

Our goal in this lab is to build a dashboard that presents our findings from the GSS. A dashboard is meant to be shared with an audience, whether that audience is a manager, a client, a potential employer, or the general public. So we need to provide context for our results. One way to provide context is to write text using markdown code.

Find one or two websites that discuss the gender wage gap, and write a short paragraph in markdown code summarizing what these sources tell us. Include hyperlinks to these websites. Then write another short paragraph describing what the GSS is, what the data contain, how it was collected, and/or other information that you think your audience ought to know. A good starting point for information about the GSS is here: http://www.gss.norc.org/About-The-GSS

Then save the text as a Python string so that you can use the markdown code in your dashboard later.

It should go without saying, but no plagiarization! If you summarize a website, make sure you put the summary in your own words. Anything that is copied and pasted from the GSS webpage, Wikipedia, or another website without attribution will receive no credit.

(Don't spend too much time on this, and you might want to skip it during the Zoom session and return to it later so that you can focus on working on code with your classmates.) [1 point]

```
[4]: text = '''
According to the [Pew Research Center](https://www.pewresearch.org/
 ↪social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/), "Women␣
 ↪generally begin their careers closer to wage parity with men, but they lose␣
 ↪ground as they age and progress through their work lives, a pattern that has␣
 ↪remained consistent over time". According to [Forbes Magazine](https://www.
 ↪forbes.com/advisor/business/gender-pay-gap-statistics/), "There are two␣
 ↪types of gender pay gaps: the controlled and uncontrolled gap. The␣
 ↪controlled gap measures the difference in pay between men and women␣
 ↪performing the same job, with the same experience and qualifications. The␣
 ↪uncontrolled gap represents the overall difference in pay between men and␣
 ↪women, considering all the jobs and industries in which they work... When␣
 ↪comparing women and men with the same job title, seniority level and hours␣
 ↪worked, a gender gap of 11% still exists in terms of take-home pay."
```

```
In order to study the gender wage gap, we consider data from the General Social␣
 ↪Survey (GSS). According to the [National Opinion Research Center](https://
 ↪gss.norc.org/About-The-GSS), "For five decades, the General Social Survey␣
 ↪(GSS) has studied the growing complexity of American society. It is the only␣
 ↪full-probability, personal-interview survey designed to monitor changes in␣
 ↪both social characteristics and attitudes currently being conducted in the␣
 ↪United States. The General Social Survey (GSS) is a nationally␣
 ↪representative survey of adults in the United States conducted since 1972.␣
 ↪The GSS collects data on contemporary American society in order to monitor␣
 ↪and explain trends in opinions, attitudes and behaviors. The GSS has adapted␣
 ↪questions from earlier surveys, thereby allowing researchers to conduct␣
 ↪comparisons for up to 80 years. The GSS contains a standard core of␣
 ↪demographic, behavioral, and attitudinal questions, plus topics of special␣
 ↪interest. Among the topics covered are civil liberties, crime and violence,␣
 ↪intergroup tolerance, morality, national spending priorities, psychological␣
 ↪well-being, social mobility, and stress and traumatic events... The data␣
 ↪come from the General Social Surveys, interviews administered to NORC␣
 ↪national samples using a standard questionnaire." The data for this study␣
 ↪include values for sex, years of formal education, personal annual income,␣
 ↪occupational prestige, index of socioeconomic status, and agreement 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."
'''
text
```

[4]: '\nAccording to the [Pew Research Center](https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/), "Women generally begin their careers closer to wage parity with men, but they lose ground as they age and progress through their work lives, a pattern that has remained consistent over time". According to [Forbes Magazine](https://www.forbes.com/advisor/business/gender-pay-gap-statistics/), "There are two types of gender pay gaps: the controlled and uncontrolled gap. The controlled gap measures the difference in pay between men and women performing the same job, with the same experience and qualifications. The uncontrolled gap represents the overall difference in pay between men and women, considering all the jobs and industries in which they work… When comparing women and men with the same job title, seniority level and hours worked, a gender gap of 11% still exists in terms of take-home pay."\n\nIn order to study the gender wage gap, we consider data from the General Social Survey (GSS). According to the [National Opinion Research Center](https://gss.norc.org/About-The-GSS), "For five decades, the General Social Survey (GSS) has studied the growing complexity of American society. It is the only full-probability, personal-interview survey designed to monitor changes in both social characteristics and attitudes currently being conducted in the United States. The General Social Survey (GSS) is a nationally representative survey of adults in the United States conducted since 1972. The GSS collects data on contemporary American society in order to monitor and explain trends in opinions, attitudes

and behaviors. The GSS has adapted questions from earlier surveys, thereby
allowing researchers to conduct comparisons for up to 80 years. The GSS contains
a standard core of demographic, behavioral, and attitudinal questions, plus
topics of special interest. Among the topics covered are civil liberties, crime
and violence, intergroup tolerance, morality, national spending priorities,
psychological well-being, social mobility, and stress and traumatic events…
The data come from the General Social Surveys, interviews administered to NORC
national samples using a standard questionnaire." The data for this study
include values for sex, years of formal education, personal annual income,
occupational prestige, index of socioeconomic status, and agreement 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."\n'

1.4 Problem 2

Generate a table that shows the mean income, occupational prestige, socioeconomic index, and
years of education for men and for women. Use a function from a `plotly` module to display a
web-enabled version of this table. This table is for presentation purposes, so round every column
to two decimal places and use more presentable column names. [3 points]

```
[5]: data_frame = (gss_clean
        .drop(
            columns = [
                'region',
                'satjob',
                'relationship',
                'male_breadwinner',
                'men_bettersuited',
                'child_suffer',
                'men_overwork'
            ]
        )
        .groupby(
            [
                'sex'
            ]
        )
        .mean()
    )
    data_frame = data_frame[['income', 'job_prestige', 'socioeconomic_index',
     ↪'education']]
    data_frame = data_frame.round(2)
    data_frame = data_frame.rename(
        columns = {
            'income': 'mean annual income',
            'job_prestige': 'occupational prestige',
            'socioeconomic_index': 'index of socioeconomic status',
```

```
        'education': 'years of education'
    }
)
data_frame = data_frame.reset_index()
table = ff.create_table(data_frame)
table.show()
```

[6]:

sex	mean annual income	occupational prestige	index of socioeconomic status	years of education
female	47191.02	44.67	46.58	13.76
male	53314.63	44.7	47.38	13.69

1.5   Problem 3

Create an interactive barplot that shows the number of men and women who respond with each level of agreement to `male_breadwinner`. Write presentable labels for the x and y-axes, but don't bother with a title because we will be using a subtitle on the dashboard for this graphic. [3 points]

```
[7]: gss_clean['male_breadwinner'].value_counts().sort_index()
```

```
[7]: male_breadwinner
     agree               310
     disagree            714
     strongly agree       88
     strongly disagree   433
     Name: count, dtype: int64
```

```
[8]: gss_clean['male_breadwinner'] = gss_clean.male_breadwinner.astype('category')
     gss_clean['male_breadwinner'] = gss_clean.male_breadwinner.cat.
       ↪reorder_categories([
        'strongly agree',
        'agree',
        'disagree',
        'strongly disagree'
     ])
     sorted_value_counts = gss_clean.groupby(['sex', 'male_breadwinner']).size()
     sorted_value_counts = sorted_value_counts.reset_index()
     sorted_value_counts = sorted_value_counts.rename(
         columns = {
             'male_breadwinner': 'agreement',
             0: 'number of people'
         }
     )
     sorted_value_counts
```

6

```
[8]:        sex          agreement  number of people
      0  female     strongly agree                48
      1  female              agree               152
      2  female           disagree               377
      3  female  strongly disagree               286
      4    male     strongly agree                40
      5    male              agree               158
      6    male           disagree               337
      7    male  strongly disagree               147
```

```
[9]: bar_plot = px.bar(sorted_value_counts, x = 'agreement', y = 'number of people',␣
       ↪color = 'sex', barmode = 'group')
     bar_plot
```

[10]:
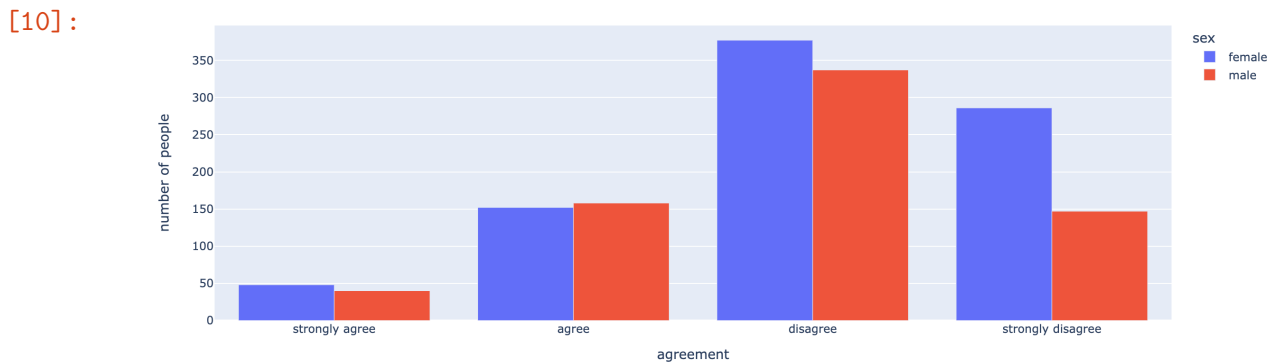
1.6   Problem 4

Create an interactive scatterplot with `job_prestige` on the x-axis and `income` on the y-axis.
Color code the points by `sex` and make sure that the figure includes a legend for these colors. Also
include two best-fit lines, one for men and one for women. Finally, include hover data that shows
us the values of `education` and `socioeconomic_index` for any point the mouse hovers over. Write
presentable labels for the x and y-axes, but don't bother with a title because we will be using a
subtitle on the dashboard for this graphic. [3 points]
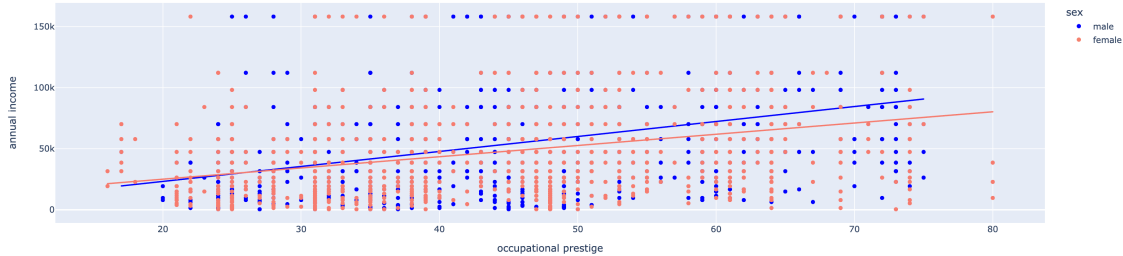
```
[11]: salmon = '#FA8072'
      scatter_plot = px.scatter(
          gss_clean,
          x = 'job_prestige',
          y = 'income',
          color = 'sex',
          color_discrete_map = {'male': 'blue', 'female': salmon},
          trendline = 'ols',
          labels = {
              'job_prestige':'occupational prestige',
              'income':'annual income'
```

```
        },
        hover_data = ['education', 'socioeconomic_index']
    )
    scatter_plot
```

[12]:

1.7 Problem 5

Create two interactive box plots: one that shows the distribution of `income` for men and for women, and one that shows the distribution of `job_prestige` for men and for women. Write presentable labels for the axis that contains `income` or `job_prestige` and remove the label for `sex`. Also, turn off the legend. Don't bother with titles because we will be using subtitles on the dashboard for these graphics. [3 points]

```
[13]: distributions_of_income_by_sex = px.box(
          gss_clean,
          x = 'sex',
          y = 'income',
          color = 'sex',
          color_discrete_map = {
              'male': 'blue',
              'female': salmon
          },
          labels = {
              'income': 'annual income'
          },
          width = 600,
          height = 600
      )
      distributions_of_income_by_sex = distributions_of_income_by_sex.update_layout(
          xaxis_title = None
      )
      distributions_of_income_by_sex = distributions_of_income_by_sex.update_layout(
          showlegend = False
      )
      distributions_of_income_by_sex
```
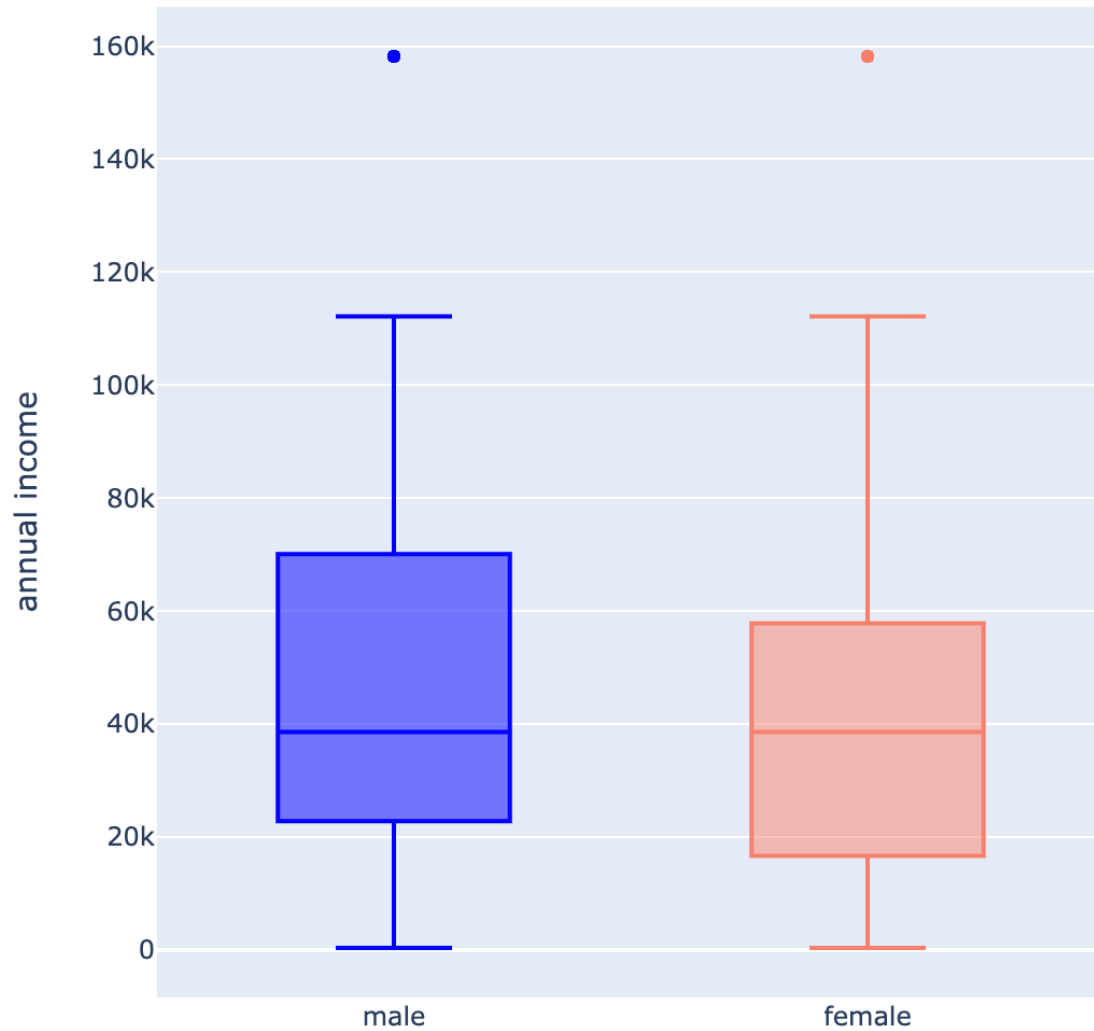
[14]:

```
[15]: distributions_of_occupational_prestige_by_sex = px.box(
          gss_clean,
          x = 'sex',
          y = 'job_prestige',
          color = 'sex',
          color_discrete_map = {'male': 'blue', 'female': salmon},
          labels = {'job_prestige':'ocuupational prestige'},
          width = 600,
          height = 600
      )
      distributions_of_occupational_prestige_by_sex =␣
       ↪distributions_of_occupational_prestige_by_sex.update_layout(
          xaxis_title = None
      )
```

```
distributions_of_occupational_prestige_by_sex =␣
 ↪distributions_of_occupational_prestige_by_sex.update_layout(
    showlegend = False
)
distributions_of_occupational_prestige_by_sex
```

[16]:
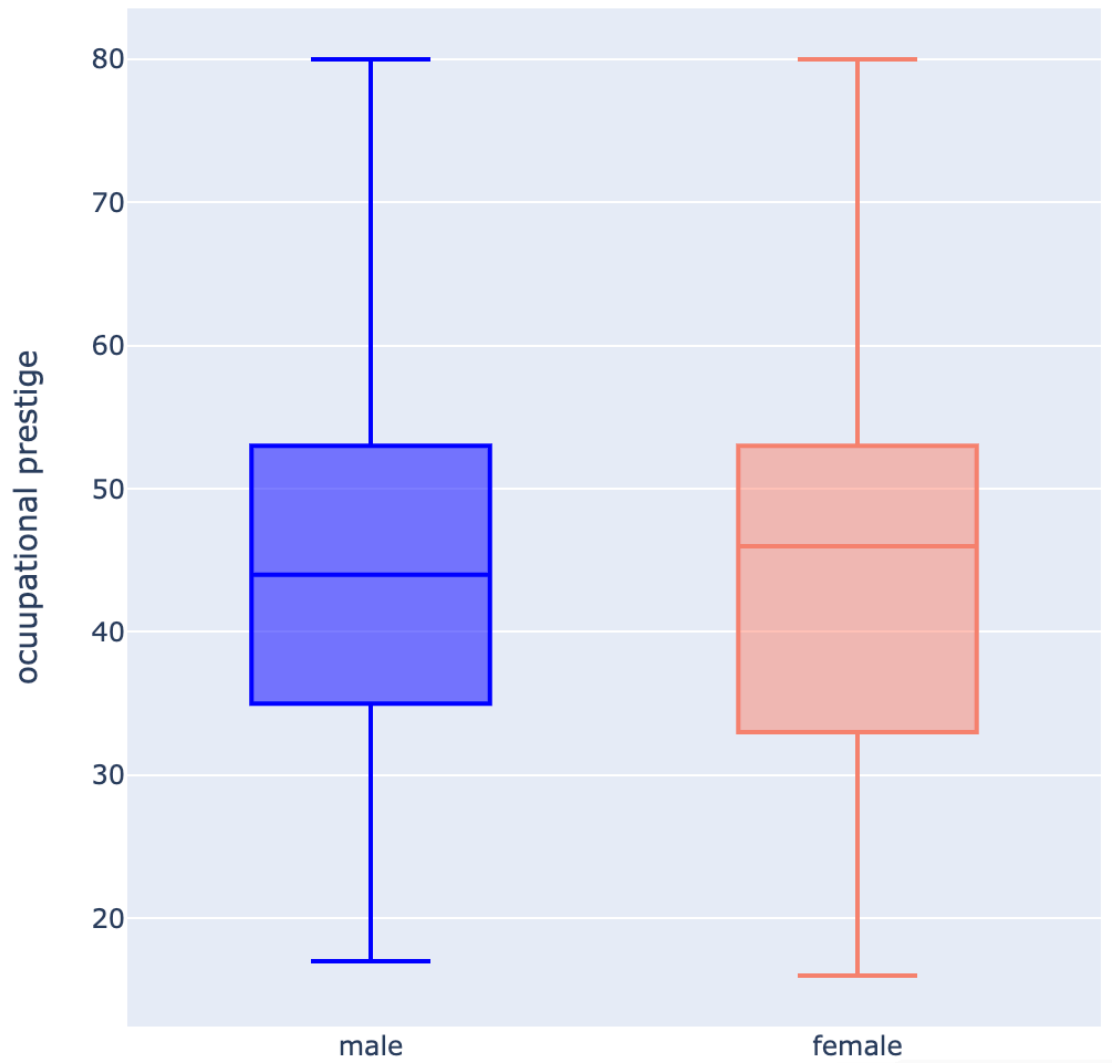
1.8 Problem 6

Create a new dataframe that contains only `income`, `sex`, and `job_prestige`. Then create a new feature in this dataframe that breaks `job_prestige` into six categories with equally sized ranges. Finally, drop all rows with any missing values in this dataframe.

Then create a facet grid with three rows and two columns in which each cell contains an interactive box plot comparing the income distributions of men and women for each of these new categories.

(If you want men to be represented by blue and women by red, you can include `color_discrete_map = {'male':'blue', 'female':'red'}` in your plotting function. Or use different colors if you want!) [3 points]

```
[17]: slice = gss_clean[['income', 'sex', 'job_prestige']]
      slice['job_prestige_binned'] = pd.cut(slice.job_prestige, bins = 6)
      slice = slice.dropna()
      slice = slice.sort_values(by = ['job_prestige_binned'])
      slice
```

```
[17]:           income     sex  job_prestige job_prestige_binned
      1      22782.5000  female          22.0   (15.936, 26.667]
      1721   47317.5000    male          25.0   (15.936, 26.667]
      1739    1402.0000    male          22.0   (15.936, 26.667]
      1746    9638.7500    male          24.0   (15.936, 26.667]
      1750   19277.5000  female          25.0   (15.936, 26.667]
      ...           ...     ...           ...                ...
      149    70100.0000  female          71.0     (69.333, 80.0]
      535    16648.7500  female          74.0     (69.333, 80.0]
      540    70100.0000  female          75.0     (69.333, 80.0]
      1277   31545.0000  female          74.0     (69.333, 80.0]
      369   158201.8412    male          72.0     (69.333, 80.0]

      [2075 rows x 4 columns]
```

```
[18]: facet_grid = px.box(
          slice,
          x = 'income',
          y = 'sex',
          color = 'sex',
          color_discrete_map = {'male': 'blue', 'female': salmon},
          facet_col = 'job_prestige_binned',
          facet_col_wrap = 2,
          labels = {'income':'annual income'},
          height = 500
      )
      facet_grid = facet_grid.for_each_annotation(lambda a: a.update(text=a.text.
       ↪replace("job_prestige_binned=", "Occupational Prestige: ")))
      facet_grid
```

```
[19]:
```
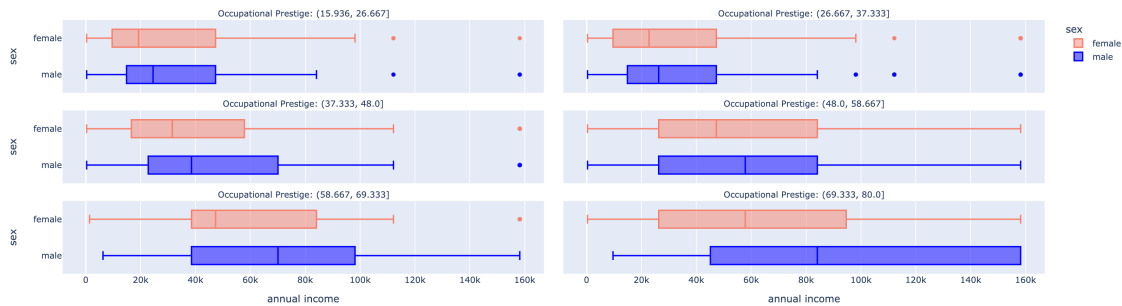
1.9 Problem 7

Create a dashboard that displays the following elements:

- A descriptive title
- The markdown text you wrote in problem 1
- The table you made in problem 2
- The barplot you made in problem 3
- The scatterplot you made in problem 4
- The two boxplots you made in problem 5 side-by-side
- The faceted boxplots you made in problem 6
- Subtitles for all of the above elements

Use `JupyterDash` to display this dashboard directly in your Jupyter notebook.

Any working dashboard that displays all of the above elements will receive full credit. [4 points]

```
[20]: external_stylesheets = ['https://codepen.io/chriddyp/pen/bWLwgP.css']
      app = JupyterDash(__name__, external_stylesheets = external_stylesheets)
      app.layout = html.Div(
          [
              html.H1('Gender Wage Gap'),
              html.H2('Introduction'),
              dcc.Markdown(children = text),
              html.H2('Socioeconomic Success By Sex'),
              dcc.Graph(figure = table),
              html.H2('Number Of People Who Support Men Achieving Outside Home And␣
          ↪Women Caring For Home'),
              dcc.Graph(figure = bar_plot),
              html.H2('Annual Income vs. Occupational Prestige'),
              dcc.Graph(figure = scatter_plot),
              html.H2('Distributions Of Income By Sex And Occupational Prestige'),
              dcc.Graph(figure = facet_grid),
              html.H2('Distributions of Income and Occupation Prestige By Sex'),
```

```
        html.Div(
            [
                dcc.Graph(figure = distributions_of_income_by_sex)
            ],
            style = {
                'width': '48%',
                'float': 'left'
            }
        ),
        html.Div(
            [
                dcc.Graph(figure =␣
  ↪distributions_of_occupational_prestige_by_sex)
            ],
            style = {
                'width': '48%',
                'float': 'right'
            }
        )
    ]
)
if __name__ == '__main__':
    app.run_server(mode = 'inline', debug = True, port = 8050)
```

Dash is running on http://127.0.0.1:8050/

<IPython.lib.display.IFrame at 0x16e163ad0>

[21]:

Gender Wage Gap

Introduction

According to the Pew Research Center, "Women generally begin their careers closer to wage parity with men, but they lose ground as they age and progress through their work lives, a pattern that has remained consistent over time". According to Forbes Magazine, "There are two types of gender pay gaps: the controlled and uncontrolled gap. The controlled gap measures the difference in pay between men and women performing the same job, with the same experience and qualifications. The uncontrolled gap represents the overall difference in pay between men and women, considering all the jobs and industries in which they work... When comparing women and men with the same job title, seniority level and hours worked, a gender gap of 11% still exists in terms of take-home pay."

In order to study the gender wage gap, we consider data from the General Social Survey (GSS). According to the National Opinion Research Center, "For five decades, the General Social Survey (GSS) has studied the growing complexity of American society. It is the only full-probability, personal-interview survey designed to monitor changes in both social characteristics and attitudes currently being conducted in the United States. The General Social Survey (GSS) is a nationally representative survey of adults in the United States conducted since 1972. The GSS collects data on contemporary American society in order to monitor and explain trends in opinions, attitudes and behaviors. The GSS has adapted questions from earlier surveys, thereby allowing researchers to conduct comparisons for up to 80 years. The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events... The data come from the General Social Surveys, interviews administered to NORC national samples using a standard questionnaire." The data for this study include values for sex, years of formal education, personal annual income, occupational prestige, index of socioeconomic status, and agreement 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."
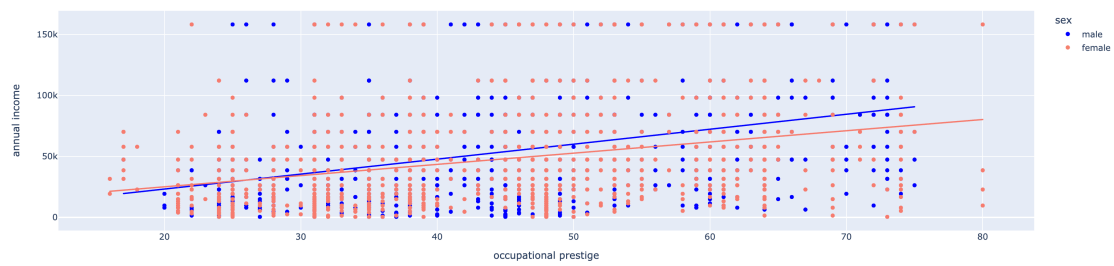
Socioeconomic Success By Sex

sex	mean annual income	occupational prestige	index of socioeconomic status	years of education
female	47191.02	44.67	46.58	13.76
male	53314.63	44.7	47.38	13.69

[22]:
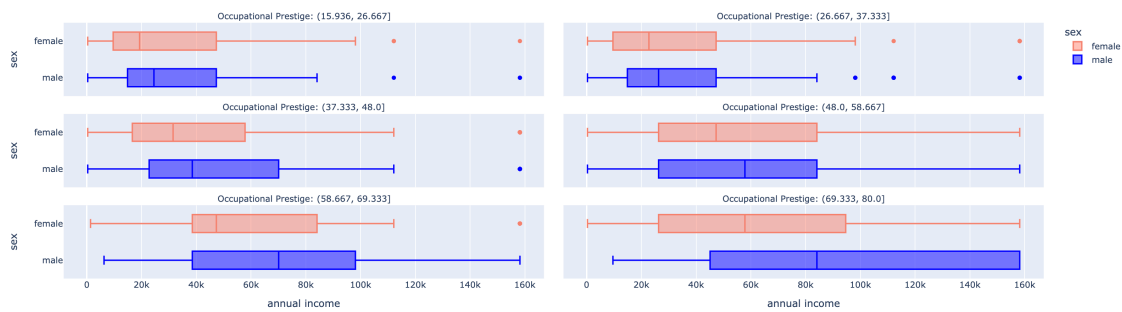
Number Of People Who Support Men Achieving Outside Home And Women Caring For Home
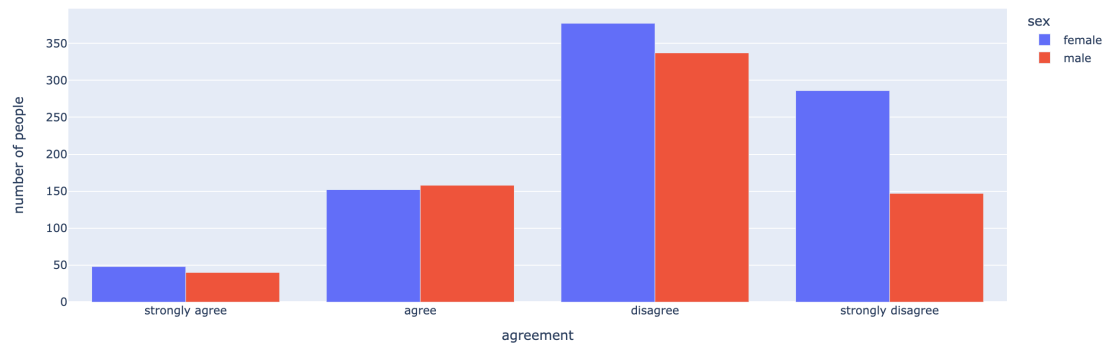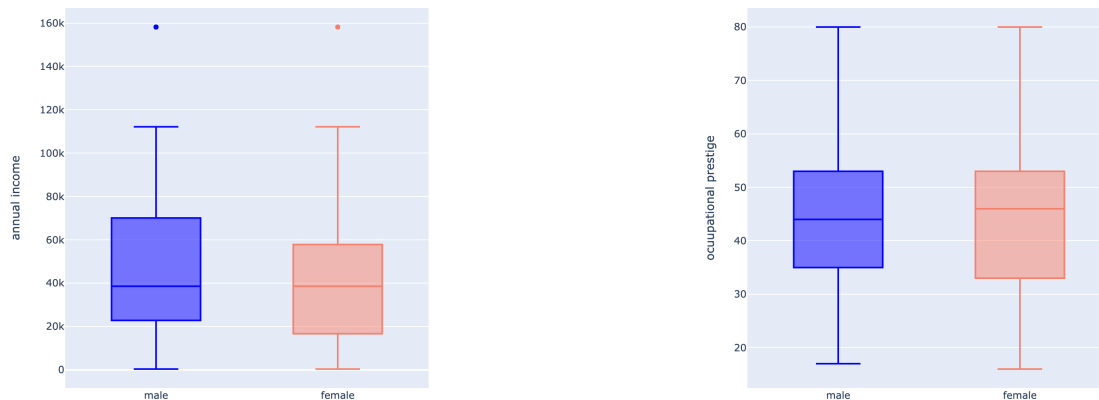

[23]:

Annual Income vs. Occupational Prestige

[24]:

Distributions Of Income By Sex And Occupational Prestige

[25]:

Distributions of Income and Occupation Prestige By Sex

1.10 Extra Credit (up to 10 bonus points)

Dashboards are all about good design, functionality, and accessability. For this extra credit problem, create another version of the dashboard you built for problem 7, but take extra steps to improve the appearance of the dashboard, add user-inputs, and host it on the internet with its own URL.

Challenge 1: Be creative and use a layout that significantly departs from the one used for the ANES data in the module 12 notebook. A good place to look for inspiration is the Dash gallery. We will award up to 3 bonus points for creativity, novelty, and style.

Challenge 2: Alter the barplot from problem 3 to include user inputs. Create two dropdown menus on the dashboard. The first one should allow a user to display bars for the categories of `satjob`, `relationship`, `male_breadwinner`, `men_bettersuited`, `child_suffer`, or `men_overwork`. The second one should allow a user to group the bars by `sex`, `region`, or `education`. After choosing a feature for the bars and one for the grouping, program the barplot to update automatically to display the user-inputted features. One bonus point will be awarded for a good effort, and 3 bonus points will be awarded for a working user-input barplot in the dashboard.

Challenge 3: Follow the steps listed in the module notebook to deploy your dashboard on Heroku. 1 bonus point will be awarded for a Heroku link to an app that isn't working. 4 bonus points will be awarded for a working Heroku link.

```
[26]:  !python3 App--Gender_Wage_Gap.py
```

```
/Users/tlever/Tom_Levers_Git_Repository/UVA/3--
Practice_And_Application_Of_Data_Science/12--
Interactive_Data_Visualization_And_Dashboards/App--Gender_Wage_Gap.py:7:
DtypeWarning: Columns (23,41,45,46,47,73,91,99,197,263,265,273,351,403,466,471,4
72,473,474,592,728,730,738,745,747,755,757,759,767,781,783,790,791,885,950,986,9
90,991,992) have mixed types. Specify dtype option on import or set
low_memory=False.
  gss = pd.read_csv(
```

```
Dash is running on http://127.0.0.1:8051/

 * Serving Flask app 'App--Gender_Wage_Gap'
 * Debug mode: on
/Users/tlever/Tom_Levers_Git_Repository/UVA/3--
Practice_And_Application_Of_Data_Science/12--
Interactive_Data_Visualization_And_Dashboards/App--Gender_Wage_Gap.py:7:
DtypeWarning: Columns (23,41,45,46,47,73,91,99,197,263,265,273,351,403,466,471,4
72,473,474,592,728,730,738,745,747,755,757,759,767,781,783,790,791,885,950,986,9
90,991,992) have mixed types. Specify dtype option on import or set
low_memory=False.
  gss = pd.read_csv(
```

1.10.1 See below link to Gender_Wage_Gap hosted by PythonAnywhere.

http://tslever.pythonanywhere.com