

Lab Assignment 12: Interactive Visualizations

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.

Problem 0

Import the following libraries:

```
In [1]: import numpy as np
import pandas as pd
import plotly.graph_objects as go
import plotly.express as px
import plotly.figure_factory as ff
import dash
from jupyter_dash import JupyterDash
from dash import dcc, 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.

```
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
```

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

```
In [3]: mycols = ['id', 'wtss', 'sex', 'educ', 'region', 'age', 'coninc',
                'prestg10', 'mapres10', 'papres10', 'seil0', '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',
        'fehires': '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."

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]

The gender pay gap is the systemic difference in the median earnings between men and women, that persists despite wage discrimination based on sex having been made illegal in the 60s. The specific figure that the gender pay gap refers to is that on average women earn 82 cents for every dollar than men earn. This pay gap has remained steady for the past few decades despite increased equity in education and job occupation, and is exacerbated by other areas of discrimination such as race, ethnicity, and disability.

The General Study Survey, or GSS, is a biannual survey of the United States of America that measures both the demographic information of its participants as well as their opinions on a variety of social, political, and moral topics. The aim of GSS is to create a reliable source of long-term, comprehensive sociological data that can be used to understand the current social landscape of the United States, and how it has changed over time.

Sources:

<https://www.americanprogress.org/article/quick-facts-gender-wage-gap/>

<https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/>

<https://www.pewresearch.org/short-reads/2023/03/01/gender-pay-gap-facts/>

<https://www.aauw.org/resources/research/simple-truth/>

<http://www.gss.norc.org/About-The-GSS>

```
In [4]: gss_summary = '''**The gender pay gap is the systemic difference in the medi
**The General Study Survey, or GSS, is a biannual survey of the United State
**Sources:**
**https://www.americanprogress.org/article/quick-facts-gender-wage-gap/**
**https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-
**https://www.pewresearch.org/short-reads/2023/03/01/gender-pay-gap-facts/**
**https://www.aauw.org/resources/research/simple-truth/**
**https://gss.norc.org/About-The-GSS**'''
```

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]

```
In [5]: gss_tab = gss_clean.groupby(['sex']).agg({'income': 'mean',
                                                'job_prestige': 'mean',
                                                'socioeconomic_index': 'mean',
                                                'education': 'mean'}).reset_index
gss_tab = gss_tab.rename(columns={'sex': 'Gender',
                                'income': 'Income',
                                'job_prestige': 'Occupational Prestige',
                                'socioeconomic_index': 'Socioeconomic Inde
                                'education': 'Education'})

fig1 = go.Figure(data=[go.Table(
    header=dict(values=list(gss_tab.columns),
                  fill_color='#4c78a8',
                  font_color='white'),
    cells=dict(values=[gss_tab.Gender, gss_tab.Income, gss_tab['Occupational
                        gss_tab['Socioeconomic Index'], gss_tab.Education],
                  fill_color='#f58518'))
])

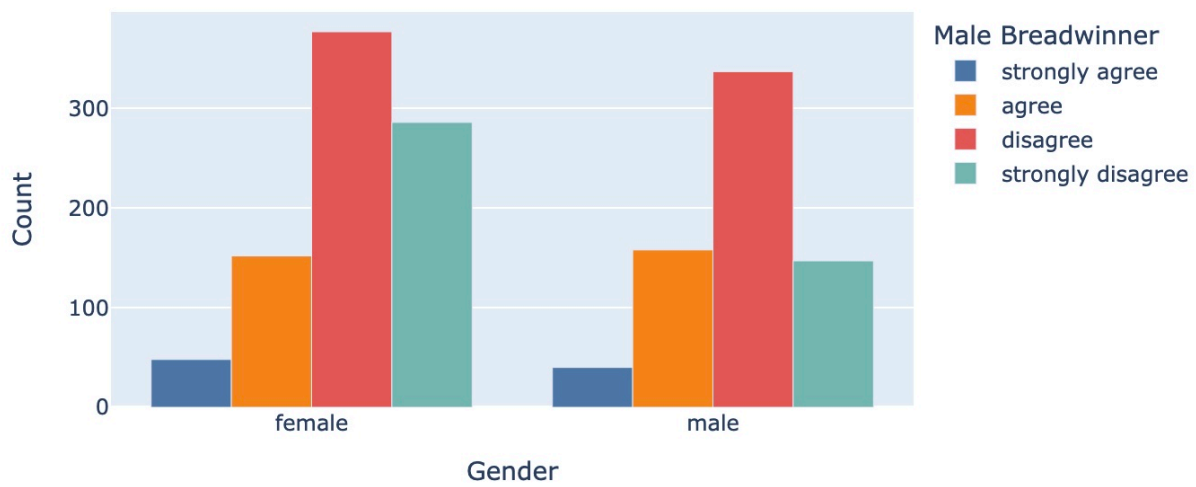
fig1.show()
```

Gender	Income	Occupational Prestige	Socioeconomic Index	Education
female	47191.02	44.67	46.58	13.76
male	53314.63	44.7	47.38	13.69

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]

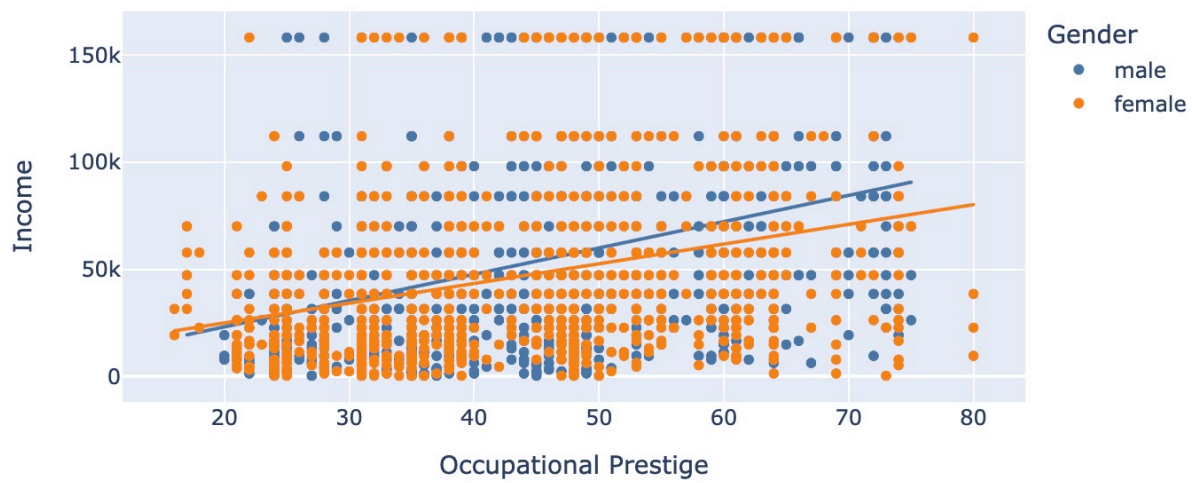
```
In [6]: gss_bar = gss_clean.groupby(['sex', 'male_breadwinner']).size().reset_index()
mb_categories = ['strongly agree', 'agree', 'disagree', 'strongly disagree']
gss_bar["male_breadwinner"] = pd.Categorical(gss_bar["male_breadwinner"], ca
gss_bar = gss_bar.sort_values(by = "male_breadwinner")
fig2 = px.bar(gss_bar, x='sex', y='Count', color='male_breadwinner',
              labels={'sex': 'Gender', 'male_breadwinner': 'Male Breadwinner'},
              barmode = 'group',
              color_discrete_sequence=px.colors.qualitative.T10)
fig2.update_layout(showlegend=True)
fig2.show()
```



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]

```
In [7]: fig3 = px.scatter(gss_clean,
    y = 'income',
    x = 'job_prestige',
    color = 'sex',
    labels={'sex': 'Gender',
            'income': 'Income',
            'job_prestige': 'Occupational Prestige',
            'socioeconomic_index': 'Socioeconomic Index',
            'education': 'Education'},
    hover_data = ["education", "socioeconomic_index"],
```

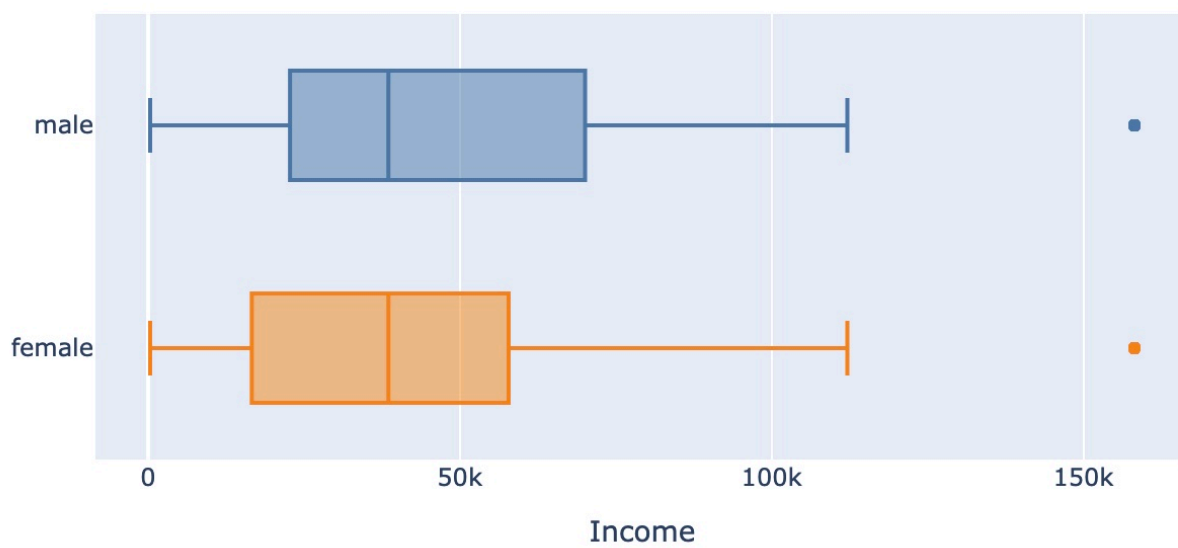



```
trendline = 'ols',  
color_discrete_sequence=px.colors.qualitative.T10)  
fig3.show()
```

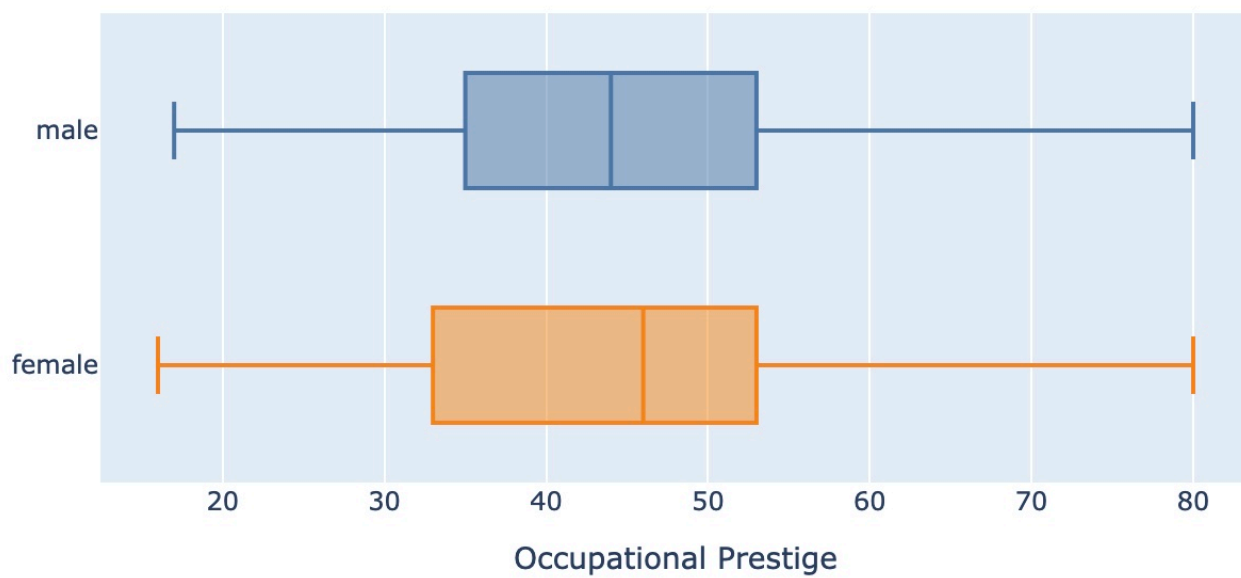
Problem 5

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]

```
In [8]: fig4 = px.box(gss_clean,  
                    x = 'income',  
                    y = 'sex',  
                    color = 'sex',  
                    labels={'income': 'Income'},  
                    color_discrete_sequence=px.colors.qualitative.T10)  
fig4.update_layout(showlegend = False)  
fig4.update_yaxes(title = '')  
fig4.show()
```



```
In [9]: fig5 = px.box(gss_clean,
                    x = 'job_prestige',
                    y = 'sex',
                    color = 'sex',
                    labels={'job_prestige': 'Occupational Prestige'},
                    color_discrete_sequence=px.colors.qualitative.T10)
fig5.update_layout(showlegend = False)
fig5.update_yaxes(title = '')
fig5.show()
```



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]

```
In [10]: gss_abr = gss_clean[['income', 'sex', 'job_prestige']]
gss_abr['job_bins'] = pd.cut(gss_abr.job_prestige, bins=6, labels = [1,2,3,4
gss_abr = gss_abr.dropna()
sex_categories = ['male', 'female']
gss_abr["sex"] = pd.Categorical(gss_abr["sex"], categories = sex_categories)
```

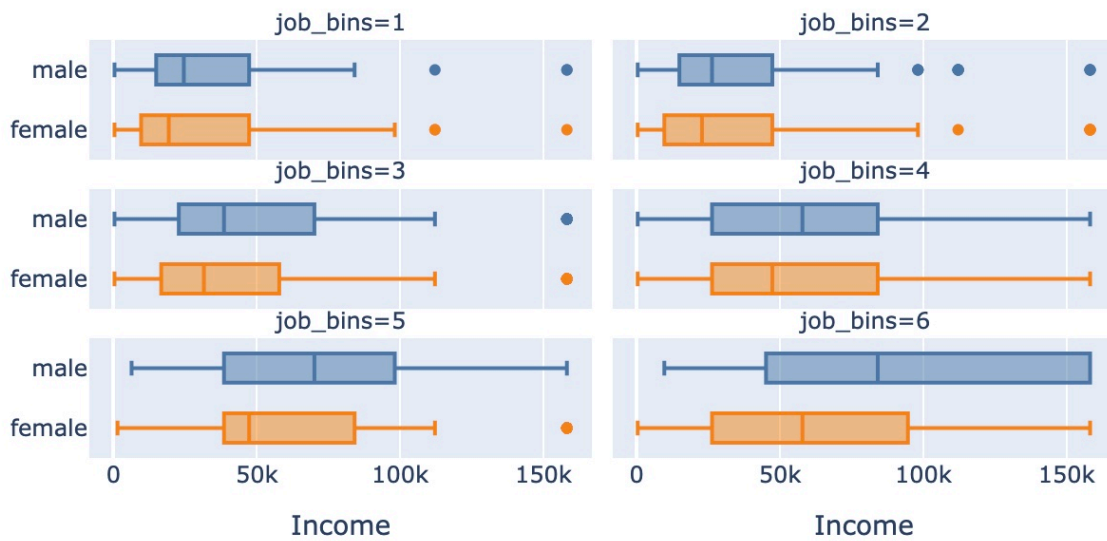
```
gss_abr = gss_abr.sort_values(by = ["sex", 'job_bins'])
```

```
In [11]: fig6 = px.box(gss_abr,
                        x='income',
                        y='sex',
                        color='sex',
                        labels={'income': 'Income'},
                        facet_col='job_bins',
                        facet_col_wrap=2,
                        color_discrete_sequence=px.colors.qualitative.T10)
fig6.update_layout(showlegend = False)
fig6.update_yaxes(title = '')
fig6.show()
```

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]

```
In [12]: app = JupyterDash(__name__, external_stylesheets=external_stylesheets)

app.layout = html.Div(
    [
        html.H1("GSS Gender Pay Gap Dashboard"),
        dcc.Markdown(children = gss_summary),

        html.H2("GSS Gender Data Average Table"),
        dcc.Graph(figure=fig1),

        html.H2("GSS Opinions on Men as the Primary Breadwinner by Gender"),
        dcc.Graph(figure=fig2),

        html.H2("GSS Income vs Occupational Prestige, by Gender"),
        dcc.Graph(figure=fig3),

        html.Div([
            html.H2("GSS Income Distribution by Gender"),
            dcc.Graph(figure=fig4)
        ], style={'width': '48%', 'float': 'left'}),
        html.Div([
            html.H2("GSS Occupational Prestige Distribution by Gender"),
            dcc.Graph(figure=fig5)
        ], style={'width': '48%', 'float': 'right'}),

        html.H2("GSS Job Group Income Distribution by Gender"),
        dcc.Graph(figure=fig6)
    ]
)

if __name__ == '__main__':
    app.run_server(mode='external', debug=True)
```

Dash app running on <http://127.0.0.1:8050/>

GSS Gender Pay Gap Dashboard

The gender pay gap is the systemic difference in the median earnings between men and women, that persists despite wage discrimination based on sex having been made illegal in the 60s. The specific figure that the gender pay gap refers to is that on average women earn 82 cents for every dollar than men earn. This pay gap has remained steady for the past few decades despite increased equity in education and job occupation, and is exacerbated by other areas of discrimination such as race, ethnicity, and disability.

The General Study Survey, or GSS, is a biannual survey of the United States of America that measures both the demographic information of its participants as well as their opinions on a variety of social, political, and moral topics. The aim of GSS is to create a reliable source of long-term, comprehensive sociological data that can be used to understand the current social landscape of the United States, and how it has changed over time.

Sources:

<https://www.americanprogress.org/article/quick-facts-gender-wage-gap/>

<https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/>

<https://www.pewresearch.org/short-reads/2023/03/01/gender-pay-gap-facts/>

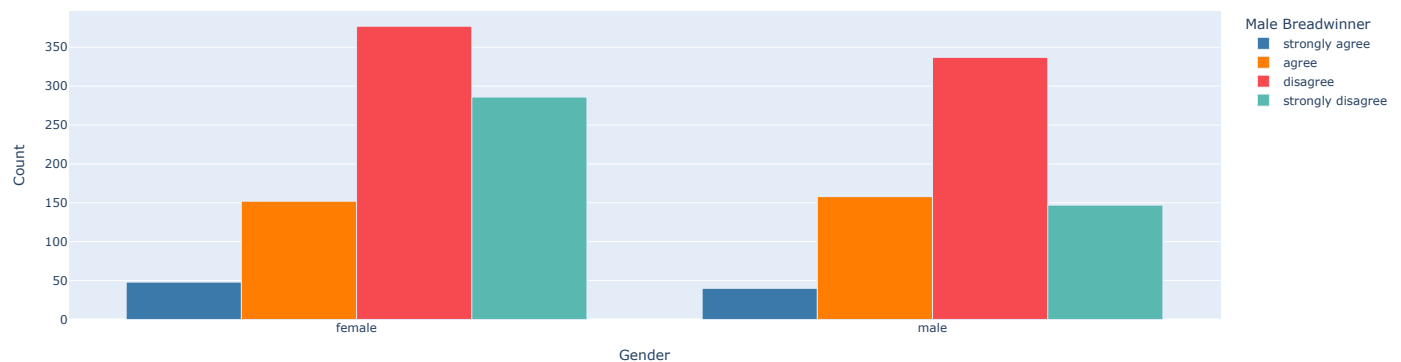
<https://www.aauw.org/resources/research/simple-truth/>

<https://gss.norc.umd.edu/About-The-GSS>

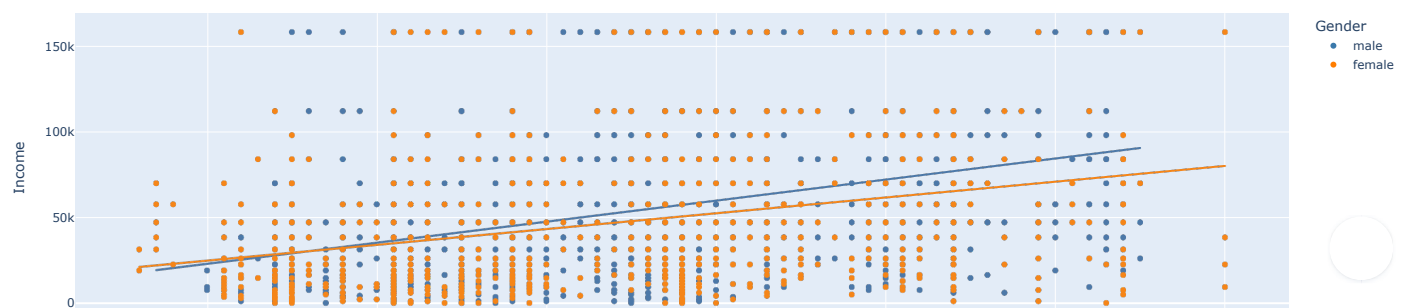
GSS Gender Data Average Table

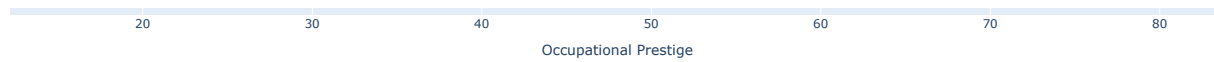
Gender	Income	Occupational Prestige	Socioeconomic Index	Education
female	47191.02	44.67	46.58	13.76
male	53314.63	44.7	47.38	13.69

GSS Opinions on Men as the Primary Breadwinner by Gender

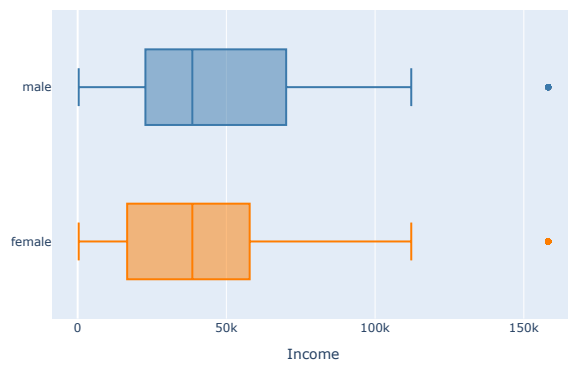


GSS Income vs Occupational Prestige, by Gender

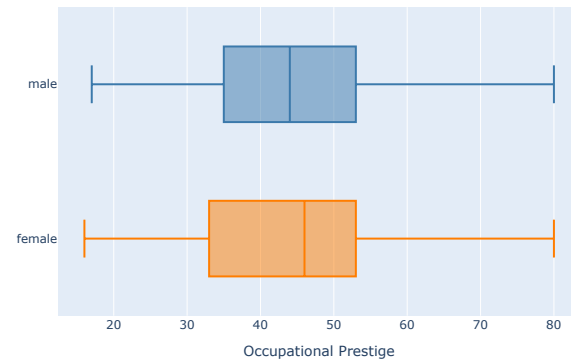




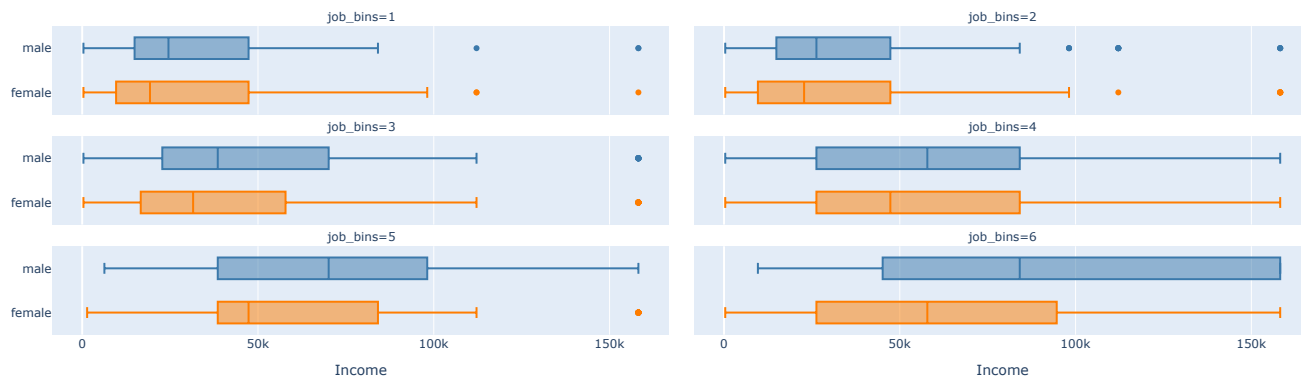
GSS Income Distribution by Gender



GSS Occupational Prestige Distribution by Gender



GSS Job Group Income Distribution by Gender



```
/opt/miniconda3/lib/python3.12/site-packages/dash/dash.py:556: UserWarning:
```

```
JupyterDash is deprecated, use Dash instead.  
See https://dash.plotly.com/dash-in-jupyter for more details.
```

Loading...

Extra Credit (up to 10 bonus points)

Dashboards are all about good design, functionality, and accessibility. 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 these steps to host the dashboard publicly on PythonAnywhere: https://docs.google.com/document/d/1IYxsRQ_J0IIM5Ztk0CN4c5JQkzlyr8dv6qMjcfLsMh0/edit?usp=sharing 4 bonus points will be awarded for a working PythonAnywhere link.

Note: I could not get the PythonAnywhere dashboard working due to a `ModuleNotFoundError` (as I explained in our class's Teams chat); I have included screenshots of my dashboard in the following pages to at least get partial credit on this extra credit assignment.

GSS Gender Pay Gap Dashboard

About the gender wage gap	About the GSS	Sources	Data Table	Opinions on Gender Roles	Income vs Occupational Prestige	Income/Job Prestige Distribution
---------------------------	---------------	---------	------------	--------------------------	---------------------------------	----------------------------------

The gender pay gap is the systemic difference in the median earnings between men and women, that persists despite wage discrimination based on sex having been made illegal in the 60s. The specific figure that the gender pay gap refers to is that on average women earn 82 cents for every dollar than men earn. This pay gap has remained steady for the past few decades despite increased equity in education and job occupation, and is exacerbated by other areas of discrimination such as race, ethnicity, and disability.

GSS Gender Data Average Table

Gender	Income	Occupational Prestige	Socioeconomic Index	Education
female	47191.02	44.67	46.58	13.76
male	53314.63	44.7	47.38	13.69

GSS Gender Pay Gap Dashboard

About the gender wage gap

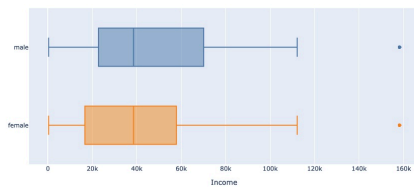
About the GSS

Sources

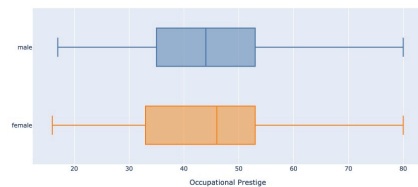
The gender pay gap is the systemic difference in the median earnings between men and women, that persists despite wage discrimination based on sex having been made illegal in the 60s. The specific figure that the gender pay gap refers to is that on average women earn 82 cents for every dollar that men earn. This pay gap has remained steady for the past few decades despite increased equity in education and job occupation, and is exacerbated by other areas of discrimination such as race, ethnicity, and disability.

Data Table	Opinions on Gender Roles	Income vs Occupational Prestige	Income/Job Prestige Distribution
------------	--------------------------	---------------------------------	----------------------------------

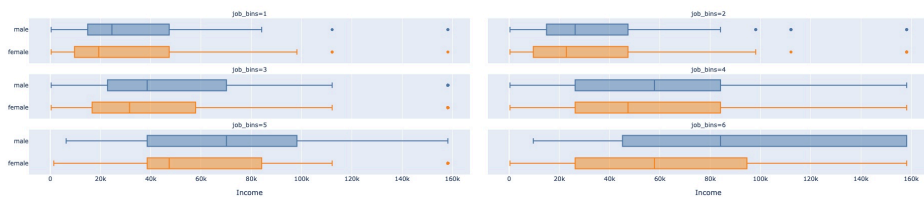
GSS Income Distribution by Gender



GSS GSS Occupational Prestige Distribution by Gender Job



Group Income Distribution by Gender



GSS Gender Pay Gap Dashboard

About the gender wage gap

About the GSS

Sources

The General Study Survey, or GSS, is a biannual survey of the United States of America that measures both the demographic information of its participants as well as their opinions on a variety of social, political, and moral topics. The aim of GSS is to create a reliable source of long-term, comprehensive sociological data that can be used to understand the current social landscape of the United States, and how it has changed over time.

Data Table

Opinions on Gender Roles

Income vs Occupational Prestige

Income/Job Prestige Distribution

Select a variable to display bars for:

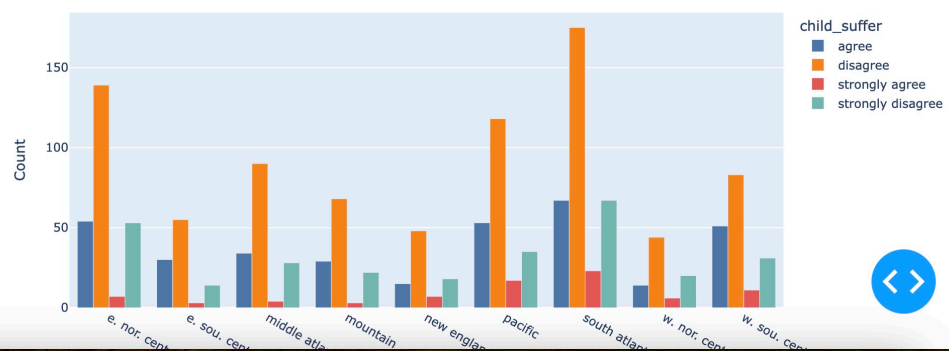
child_suffer

Select a variable to group the bars by:

4

No results found

GSS Opinions on Gender Roles



GSS Gender Pay Gap Dashboard

About the gender wage gap	About the GSS	Sources	Data Table	Opinions on Gender Roles	Income vs Occupational Prestige	Income/Job Prestige Distribution
---------------------------	---------------	---------	------------	--------------------------	---------------------------------	----------------------------------

<https://www.americanprogress.org/article/quick-facts-gender-wage-gap/>

<https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/>

<https://www.pewresearch.org/short-reads/2023/03/01/gender-pay-gap-facts/>

<https://www.aauw.org/resources/research/simple-truth/>

<https://gss.norc.berkeley.edu/About-The-GSS>

GSS Income vs Occupational Prestige, by Gender

