Introduction to Data Science

Data Science Essentials

Mary van Valkenburg
Data Science Program Manager/Instructor
Nashville Software School

Goals for today

- Review last session coding tasks
- Review for-loops and list comprehensions
- Introduce the iterrows() method
- Tiptoe into next weeks session
  - variable types and missingness
  - Intro to matplotlib

Review last session coding tasks

week2_review notebook

List comprehensions

Task: you have a list of animals called *my_animals* and you want to get a list of the just the animals that start with the letter *t* from it

Creating the list with a for-loop:

```
# for loop to find animals in my_animals that start with the letter t
t_animals = []
for animal in my_animals:
    if animal.startswith('t'):
        t_animals.append(animal)
print(t_animals)

# list comprehension to find animals in my_animals that start with the letter t
t_animals2 = [animal for animal in my_animals if animal.startswith('t')]
print(t_animals2)
```

[*what to return* for *iterator* in *original list* if *conditional*]

***The iterator can be any word or letter**

Get Data → Process + Clean Data → Exploratory Data Analysis

Statistics and other info

df.describe() – **to get summary statistics about quantitative data**

df.info() – **to get information about the DataFrame**

df.isnull().sum() – **to get counts of missing values**

```
import matplotlib.pyplot as plt
import seaborn as sns
```

https://matplotlib.org/api/_as_gen/matplotlib.pyplot.plot.html#examples-using-matplotlib-pyplot-plot
https://seaborn.pydata.org/examples/index.html

matplotlib.pyplot and seaborn barplots, for-loops, list comprehensions, iterrows()

public_art3 notebook

Variable types

What kind of data is each variable?

Descriptive (Qualitative)

- **categorical**
  - **nominal**
  - **ordinal**

Numerical (Quantitative)

- **discrete**
- **continuous**

Categorical data

- value_counts()

- barplots

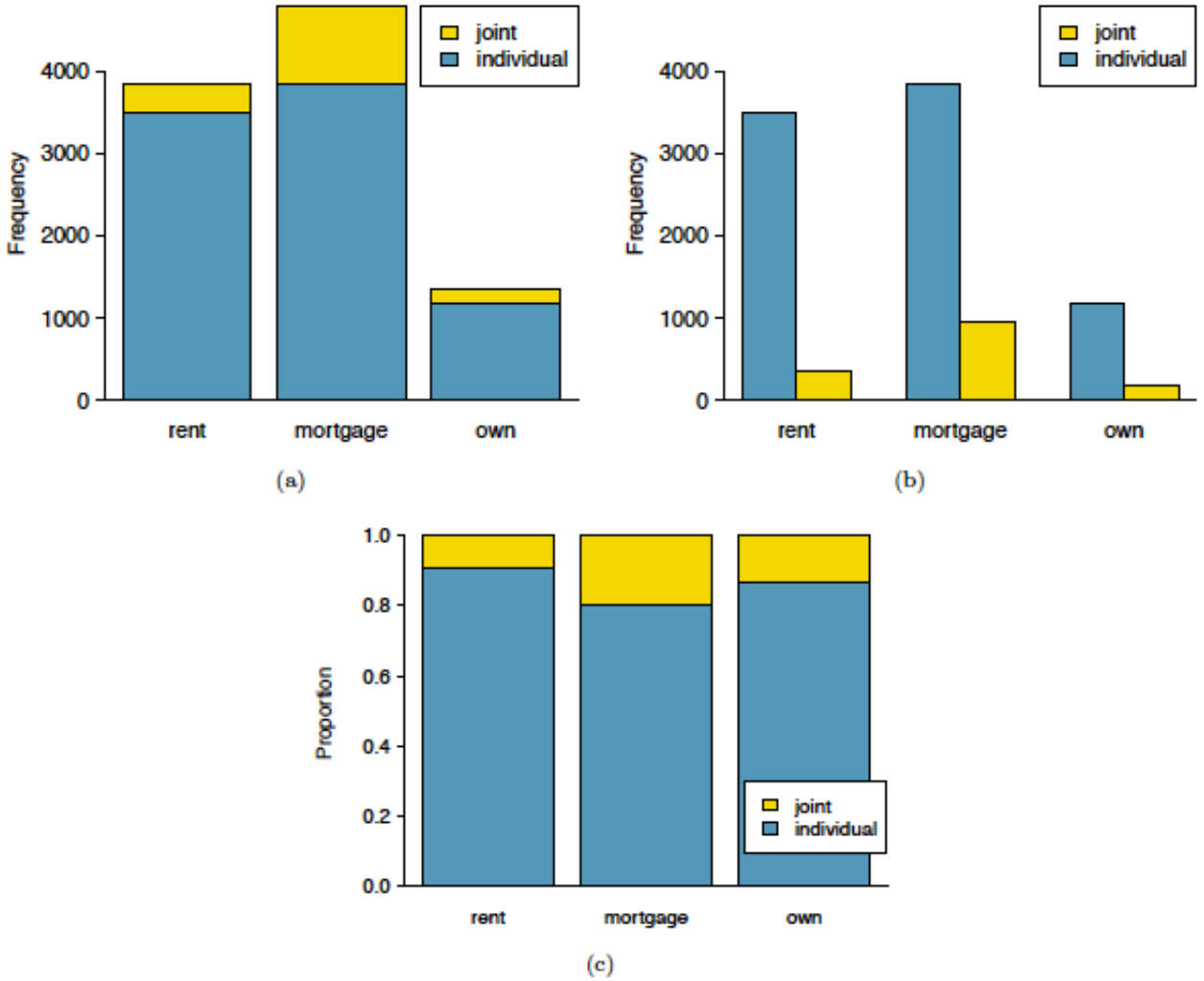


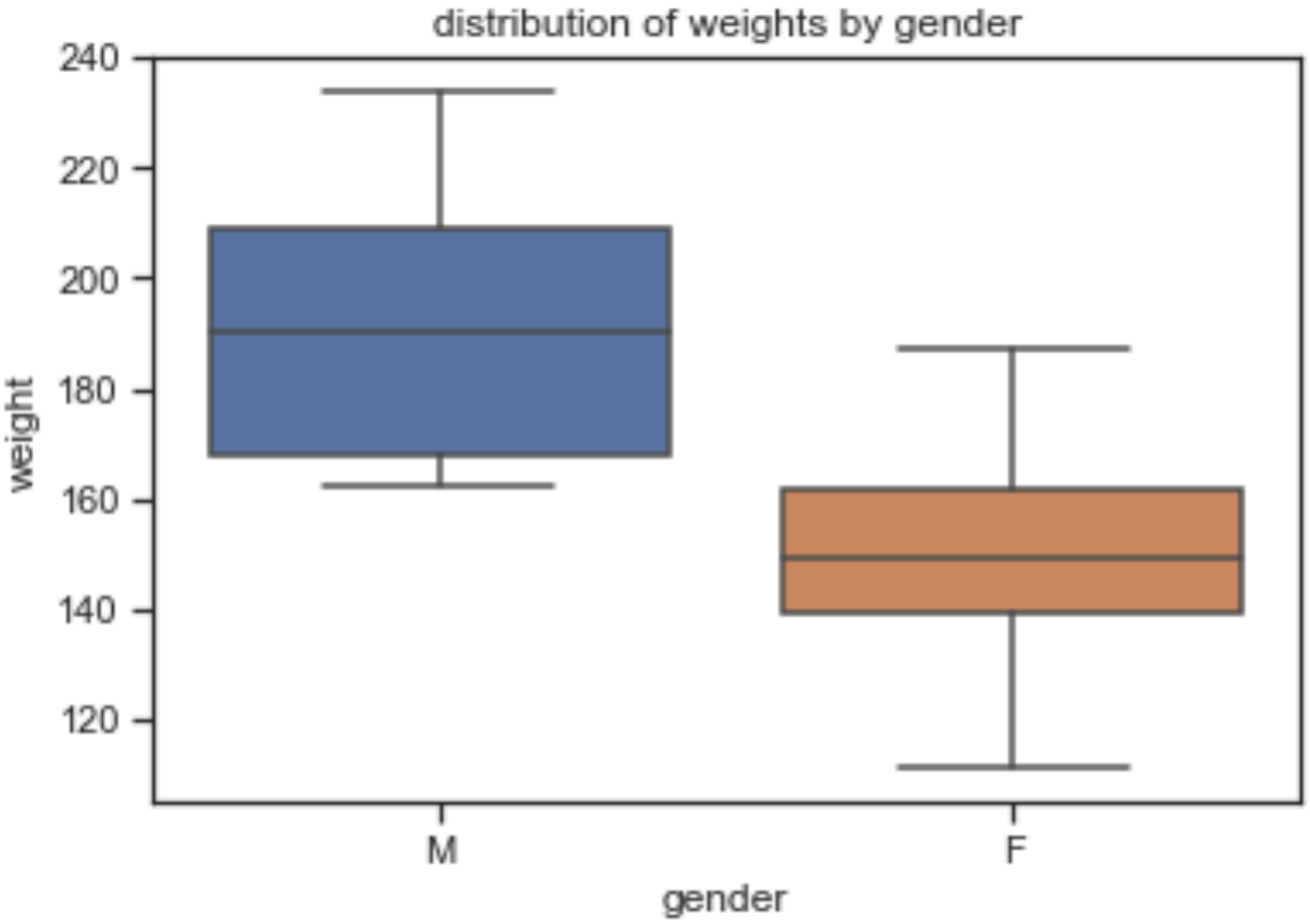
Figure 2.23: (a) Stacked bar plot for `homeownership`, where the counts have been further broken down by `app_type`. (b) Side-by-side bar plot for `homeownership` and `app_type`. (c) Standardized version of the stacked bar plot.

https://openintro.org

Numerical data

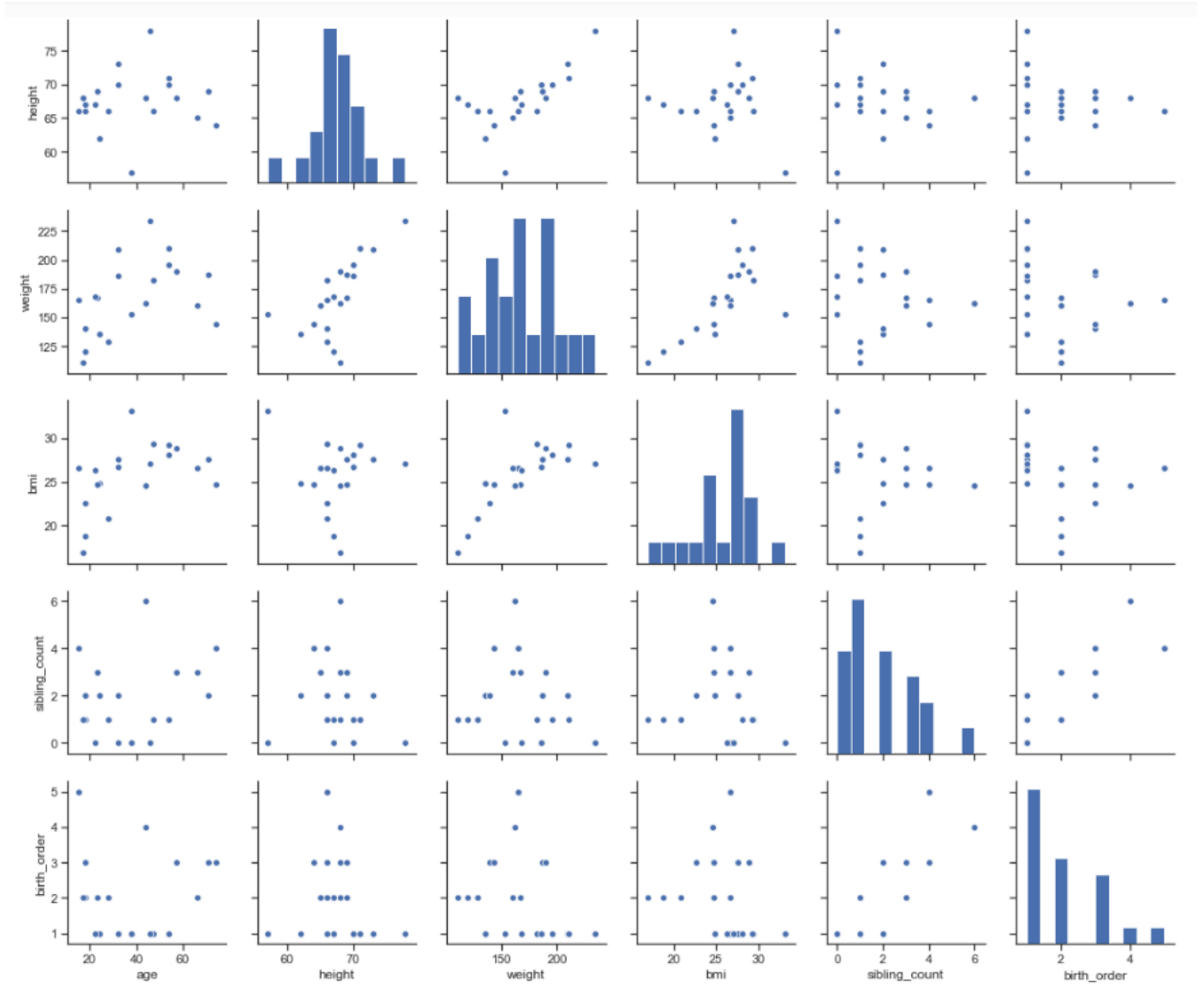
- describe()

- histograms
- boxplots
- swarmplots
- stripplots
- violin plots



Relationships

scatterplots
pairplot



Variable types and missingness

eda_workflow notebook

Reminders

- **Build upon your work in the same notebook each week (Github keeps a history)**
- **If the code in a cell did not run as expected, modify the code in that cell (not a new one)**
- **Remove any unused/un-useful cells**
- **Beware of the changing state of objects in your notebook**
  - **Example – if you create a df and drop 3 columns and then go back to add code to look at the head() in the same cell you are re-running the command to drop the 3 columns which are no longer there!**

Questions?