Homework 1: Analyzing a single variable

Data Analytics and Visualization, Fall 2024

Prof. Gabe Hope

Learning objectives

- Understand how to locate and access a dataset of interest
- Load and preprocess a dataset with Pandas
- Visualize the distribution of a variable and compute summary statistics using Pandas and ggplot
- Create a univariate model and identify the relevant modeling assumptions
- Use a model to answer a question about a given population
- Publish a report using Quarto

Logistics

Timeline

- **Friday, September XX 11:59pm** Create a gradescope submission for your assignment with all group members.
- **Thursday, September XX 11:59pm** Submit your final assignment on Gradescope
- **Friday, September XX 11:59pm** Submit your reflections on Gradescope.
- **Thursday, September XX 11:59pm** Submit your peer feedback on Gradescope.

Groups

This assignment **must** be completed in groups of 2-3 students. The instructor reserves the right to add students to a group as needed.

Once you have chosen your group, please create a submission on Gradescope with all group memebers by **Friday, September XX**.

Assignment

Goal: Choose a public dataset and put yourselves in the shoes of the team that collected this data. You want to show why this data is interesting and illustrate examples of what can be learned from it!

To this end, you will create example visualizations of your dataset and perform some basic analysis to show what conclusions can be drawn from this data. For this assignment we are focusing on *single-variable analysis*, we'll look at how to visualize and model interactions between variables in future assignments.

Each member of your team should choose a varible in the dataset that they think may show something interesting. For each of these variables, create a visualization that shows the distribution of that variable. The exact type of visualization is up to you, but it should provide a *clear* and *accurate* picture of the data. A reader should also understand either from the visualization and/or from text, what some of the relevant descriptive statistics of the variable are (e.g. mean, median, min, max, mode etc., depending on what you think is most important to show).

For each variable perform either a predictive or inferential analysis that illustrates something that we can conclude from that variable. For example, this could mean estimating the likelihood of some future value or testing a hypothesis about the population mean of that variable.

Submit a 1-3 page (including figures) report about you dataset that includes your visualizations and the results of your analysis. Your report should be written for a general audience who you think should be interested in this data. This could be researchers, policy-makers or the general public; anyone whose descisions may be influenced by your findings. This means you should first motivate *why* your chosen dataset is interesting/exciting/important. Then explain what your visualizations show about the data. Finally discuss what your analysis shows and why it is exciting, intuitive or surprising. You do *not* need to include the details of your calculations, but it should be clear what assumptions are being made for each model and how reasonable these assumptions are.

> **Note**
>
> Depending on the dataset, you may decide to have multiple team members choose the same variable. In this case, you could have a second visualization illustrate the results of your analysis.

Requirements checklist

- □ Choose a public dataset to focus on. In your report motivate *why* you find this dataset interesting.
- □ Choose a variable for **each team member** to analyze. (Team members *may* choose the same variable). In your report motivate why each variable is worth analyzing.
- □ Create 1 visualization for **each team member**. Every chosen variable should have a visualization that shows its distribution.
- □ Perform some kind of **single-variable analysis** for **each team member**. In your report discuss the asumptions used in each analysis and the conclusion.
- □ Submit a report that: **introduces and motivates the chosen dataset**, **shows and discusses your visualizations**, and **discuses your analysis**.

Format and submission

Assignments should follow the homework Quarto template. Each submission should be rendered as a PDF and submitted through Gradescope.

Rubric

Both peer and intstructor feedback will use the following rubric for assesment.

Reflections

Individual reflections are due on Gradescope within 1 day of submitting the final assignment. Each reflection will ask the following questions.

- How long did you spend on this assignment?
- What did you contribute in working on this assignment? Is there anything you are proud of or think you could have done better?
- Did each group member contribute (roughly) equally to this assignment?
- Do you have any feedback for the instructor about how this assignment could be improved? (Or any other issues with the course?)
- Did you recieve any particular peer or instructor feedback that you felt helped you produce a better submission?

The goal of reflections is to help you think about your progress throughout this course and to help the instructor identify any potential issues. They will be much more helpful if you answer honestly. Reflections will only be graded based on completion, **not** on content and responses will **not** be shared with your peers. You should **not spend more than 5 minutes** completing the reflection.

Peer feedback

Once the assignment has been completed you will be assigned to provide feedback on the submission of one of your peers. Peer feedback must be submitted on Gradescope within 1 week of the assignment due date. You should aim to spend **10-20 minutes** providing feedback.

Please also refer to the guidelines for effective feedback.

> **Note**
>
> While your options are valued and may be referenced in instructor feedback, peer feedback will not be directly used to determine assignment grades.

Time estimates

Below is a (rough) outline of how much time each team member should aim to work on each part of this assignment. If you find that any part of this assingment is taking considerably longer, please reach out to the instructor.

- **Choosing a dataset and variables:** 30-60 minutes.
- **Creating a visualization:** up to 30 minutes.
- **Analyzing a variable:** 30-60 minutes.
- **Writing the report:** up to 60 minutes.
- **Reflections and peer feedback:** <30 minutes.

Total: 3-4 hours

Tips, tricks and suggestions

This assignment is **intentionally open-ended**. The goal is give you the oppurtunity to experiment and to explore topics that interest you. Remember that there generally isn't a single *correct* way to visualize or analyze data. Your goal should be to provide *useful* and *informative* analysis, while avoiding pitfalls that can lead to incorrect conclusions.

Don't panic! Grading will be focused on whether or not you follow the principles outlined in the rubric, *not* whether you picked the best possible dataset, visualization, model or report format, and grading will be especially forgiving for early assignments. You are encouraged to use this oppurtunity take risks and make mistakes. That way you can learn from the feedback that you get! As we progress through the semester and learn more about effective data analysis and visualization, we'll work together to formulate more specific guidelines for evaluating these assignements.

Below is a guide for how you *could* approach this assignment. You don't need to follow it exactly; you're free to do whatever visualizations and analysis you'd like as long as they fall within the requirements outlined above.

Part 1: Finding a dataset

Choose a source of data for your assignment. The data sources page on the course website is a good place to start! As you are choosing consider the following guidelines:

- **Choose data that interests you!** For example, if you are a baseball fan, you could consider a dataset of player stats from the Lehman database. If you're interested in public health, you could consider COVID data from sources like the CDC or CA/NY open data repositories.
- **Consider the format!** Data formated in the CSV (or similar like TSV, XLSX, etc.) format will be the easiest to work with for this assignment and many data sources provide data in this format. Data that requires specialized formats or access via an API may require more time than you'd want to spend.
- **Keep it a managable size!** In many cases you'll find datasets with thousands, millions or even possibly *billions* of observations. If the dataset you chose is too large, you may have quite a bit of difficulty visualizing and analyzing it with the tools we are using in this course. Try to restrict your data to have fewer than ~10,000 observations, and it's

completely fine if your dataset only has as few as 30-50 observations. For example, if you are considering using the Stanford open policing project, you might choose only the data from a particular small city such as Little Rock.
- **Choose multifaceted data!** For the rest of this assignment you'll be asked to analyze different types of variables within your chosen data. Make sure to choose a dataset that has enough diversity in its variables to answer these questions. Otherwise you may need to use more than one dataset.

Once you've chosen your data, download it an make sure every team member has access.

Load your data into Python using Pandas. You should be able to verify that it loaded correctly using something like `df.head()`. Then make sure you understand what the data represents. You should be able to answer the following questions: *How many variables are in the dataset? What is the type of each variable and what does it represent? How many observations are in the dataset and what does each one represent?*

> **Tip**
>
> **For example:** You might choose to look at Lebron James' per-game performance over the 2023-2024 NBA season. You might find that each observation in your data represents a single game and the varibles include quantitative statistics such as points scored, rebounds, steals etc. as well as nominal values such as the opposing team and whether it was a home or away game.

Part 2: Visualizing single variables

With your dataset in hand, it's time to start trying to understand it using visualization! For this assignment, we're only going to look at one variable at a time. Again, try to choose variables that you think may show an interesting property of your data.

Choose one *continuous* or *ordinal* variable from your dataset. Using Lets-Plot or PlotNine, create a visualization of the distribution of this variable. Exactly what type of visualization you use is up to you, but the visualization should be clear, well-labeled and provide an accurate picture of the data. Follow the relevant guidelines from the visualization rubric.

You should also compute key summary statistics for the variable. At minimum, the *mean*, *median*, *minimum* and *maximum* using Pandas. If appropriate, show them in the visualizaiton otherwise you may refer to them in your write-up.

And/or choose one *nominal* (categorical) variable from your dataset. Visualize the distribution of this variable. As before, the exact style of visualization is up to you, but the same standards of quality will apply.

> **Tip**
>
> **For example:** In your analysis of Lebron's games you may decide to look at the number of points score per-game, visualizing this distribution using a histogram. You may consider visualizing the proportion of home and away games, but decide that because you already know that the proportion is 50-50 that it's not worth analyzing.

Part 3: Analysis

Now that we understand our data a bit better having visualized it, let's see if we can use it to help us make useful decisions. To do this we'll first need to determine a specific piece of knowledge that we want to get from our data.

Pose a question. Consider continuous/ordinal variable that you visualized in the previous part. Determine a question that you could answer using a model of this variable. In this case we'll restrict it to asking about the likelihood of some future outcome.

> **Tip**
>
> **For example:** if your variable is the number of points Lebron James scores in each game, you could attempt to answer the question: *how likely is it that Lebron will score more than 50 points?*

Fit a distribution. Choose the most appropriate distribution out of those discussed in class to model your continuous/ordinal variable. (This could be a Normal, Exponential, Binomial, Negative Binomial or Poisson). Estimate the parameters of this distribution that best fits your data using the *maximum likelihood principle.*

As you complete this section, consider the *assumptions* that you are relying on. For example: Is your chosen distribution a good fit to the data? Is it reasonable to assume your observations are independent?

Answer the question. Use your model to attempt to answer this question (e.g. *under our model, Lebron has an 8% chance of scoring more than 50 points*)

> **Tip**
>
> **For example:** if your variable is the number of points Lebron James scores in each game, you could attempt to answer the question: *how likely is it that Lebron will score more than 50 points?*

Make sure each group member performs some analysis of this form.

Part 4: Writing up!

Finally, once your visualization an analysis is complete you’ll need to write up a report on what you found. Consider our class guidelines for writing good reports!

```
import numpy as np

from lets_plot import *
import lets_plot
import pandas as pd
import vega_datasets
import seaborn as sns

from lets_plot import *

LetsPlot.setup_html()

from IPython.display import SVG
class ggsvg(lets_plot.plot.core.PlotSpec):
    def __init__(self):
        pass

    def __radd__(self, plot):
        ggsave(plot, 'plot.svg', w=8, h=4, unit='in', dpi=300)
        return SVG(filename='lets-plot-images/plot.svg')



ggplot() + geom_point(x=0, y=0) + ggsize(200, 200)
```

```
Unable to display output for mime type(s): text/html
```

```
<lets_plot.plot.core.PlotSpec at 0x1074999a0>
```

```
Plot.rectY({length: 10000}, Plot.binX({y: "count"}, {x:
d3.randomNormal()})).plot()
```