Assignment 3: Exploratory Data Analysis

In this assignment, you will identify a dataset of interest and perform an exploratory analysis to better understand the shape & structure of the data, investigate initial questions, and develop preliminary insights & hypotheses. Your final submission will take the form of a report consisting of captioned visualizations that convey key insights gained during your analysis.

Step 1: Data Selection

First, you will pick a topic area within the field of kinesiology that interests you and find a dataset that can provide insights into that topic. To streamline the assignment, we've pre-selected a number of datasets for you to choose from.

However, if you would like to investigate a different topic and dataset, you are free to do so. If working with a self-selected dataset, please check with the course staff to ensure it is appropriate for the course. Be advised that data collection and preparation (also known as data wrangling) can be a very tedious and time-consuming process. Be sure you have sufficient time to conduct exploratory analysis, after preparing the data.

After selecting a topic and dataset – but prior to analysis – you should write down an initial set of at least three questions you'd like to investigate.

Part 2: Exploratory Visual Analysis

Next, you will perform an exploratory analysis of your dataset using a visualization tool such as Tableau. You should consider two different phases of exploration.

In the first phase, you should seek to gain an overview of the shape & structure of your dataset. What variables does the dataset contain? How are they distributed? Are there any notable data quality issues? Are there any surprising relationships among the variables? Be sure to also perform "sanity checks" for patterns you expect to see!

In the second phase, you should investigate your initial questions, as well as any new questions that arise during your exploration. For each question, start by creating a visualization that might provide a useful answer. Then refine the visualization (by adding additional variables, changing sorting or axis scales, filtering or subsetting data, etc.) to develop better perspectives, explore unexpected observations, or sanity check your assumptions. You should repeat this process for each of your questions, but feel free to revise your questions or branch off to explore new questions if the data warrants.

Final Deliverable

Your final submission should take the form of a report – similar to a slide show or comic book – that consists of 10 or more captioned visualizations detailing your most important insights. Your "insights" can include important surprises or issues (such as data quality problems affecting your analysis) as well as responses to your analysis questions.

Each visualization image should be a screenshot exported from a visualization tool, accompanied with a title and descriptive caption (1-4 sentences long) describing the insight(s) learned from that view. Provide sufficient detail for each caption such that anyone could read through your report and understand what you've learned. You are free, but not required, to annotate your images to draw attention to specific features of the data. You may perform highlighting within the visualization tool itself, or draw annotations on the exported image. To easily export images from Tableau, use the Worksheet > Export > Image... menu item.

The end of your report should include a brief summary of main lessons learned.

Recommended Data Sources

To get up and running quickly with this assignment, we recommend exploring one of the following provided datasets:

- **OpenPowerlifting.** OpenPowerlifting is an open archive of powerlifting competition results from federations around the world, with about 4 million entries. Each row is one lifter at one meet, with sex, age, body weight, weight class, equipment, best squat, bench press and deadlift, total, and scoring points such as Dots and Wilks, plus the meet's date, federation and country. You might use this to investigate how strength changes with age and body size, how equipment changes performance, or how participation has grown over time. The complete dataset is a single CSV file in the public domain (about 160 MB zipped); the IPF-only file is smaller. Column descriptions are in the documentation.
- **Treadmill Maximal Exercise Tests, University of Málaga (PhysioNet).** This dataset contains 992 maximal graded treadmill tests from 857 people aged 10–63, recorded at the University of Málaga's Exercise Physiology and Human Performance Lab from 2008 to 2018. Heart rate, oxygen uptake (VO_2), carbon dioxide output (VCO_2), ventilation, breathing rate and treadmill speed are recorded breath by breath, and a second file gives each person's age, sex, height, weight, and the lab's temperature and humidity. You might use this to compare fitness across ages and sexes, examine how heart rate and VO_2 respond to rising speed, or look for data quality issues in breath-by-breath measurements. The data are two open-access CSV files that join on a shared test ID.
- **Crowd-sourced Fitbit Datasets, 2016 (Zenodo).** Thirty Fitbit users shared two months of tracker data, including daily, hourly and minute-level steps, activity intensity, calories and METs, heart rate every few seconds, and minute-level sleep, along with weight logs for some users. You might use this to investigate how active people are across the week and the day, how sleep relates to the next day's activity, or how tracker use varies between people. Because users wore different devices and tracked different things, missing and uneven data are part of what you should explore. The data are 18 CSV files per month in two ZIP archives under CC BY 4.0, linked by a shared user ID.
- **Equity in Athletics Data Analysis (EADA), U.S. Department of Education.** Every U.S. college with an intercollegiate athletics program and federal student aid reports its athletics participation and finances each year. The data include the number of men and women participating in each sport, coaching staff and salaries, and revenues and expenses by sport, along with institution details such as division and enrollment. You might use this to investigate gender equity in college sport, which sports generate or cost the most, or how athletics budgets differ across divisions. Full-year data files are available from the site's data download page, with a record per sport at each institution and a codebook for each year.

• **WHO Global Health Observatory: Physical Activity.** The World Health Organization tracks physical activity worldwide, including the share of adults and adolescents who are insufficiently active, by country, sex and year since 2000, and whether each country has national physical activity policies and guidelines. You might use this to compare countries and regions, look at trends over time, or relate activity levels to national policies. Each indicator can be downloaded as a CSV from its indicator page or through the GHO OData API; we encourage you to join several indicators by country and year for a sufficiently rich exploratory process.

Additional Data Sources

If you want to investigate datasets other than those recommended above, here are some possible sources to consider. You are also free to use data from a source different from those included here. If you have any questions on whether your dataset is appropriate, please ask the course staff ASAP!

- PhysioNet - Physiological and Clinical Research Data
- Zenodo and Figshare - Open Research Datasets (search for exercise, gait, biomechanics, wearables)
- CDC NHANES - National Health and Nutrition Examination Survey
- data.cdc.gov - CDC Open Data, including physical activity and chronic disease
- IPUMS.org - Integrated Census & Survey Data, including time use and health surveys
- data.gov - U.S. Government Open Datasets
- U.S. Census Bureau - Census Datasets
- OpenNeuro - Open Brain Imaging Datasets
- fivethirtyeight.com - Data and Code behind the Stories and Interactives, including sports
- 17 places to find datasets for data science projects

Visualization Tools

You are free to use one or more visualization tools in this assignment. However, in the interest of time and for a friendlier learning curve, we strongly encourage you to use Tableau. Tableau provides a graphical interface focused on the task of visual data exploration. You will (with rare exceptions) be able to complete an initial data exploration more quickly and comprehensively than with a programming-based tool.

- Tableau - Desktop visual analysis software. Available for both Windows and MacOS; register for a free student license.
- Data Transforms in Vega-Lite. A tutorial on the various built-in data transformation operators available in Vega-Lite.
- Data Voyager, a research prototype from the UW Interactive Data Lab, combines a Tableau-style interface with visualization recommendations. Use at your own risk!
- R, using the ggplot2 library or with R's built-in plotting functions.
- Jupyter Notebooks (Python), using libraries such as Altair or Matplotlib.

Data Wrangling Tools

The data you choose may require reformatting, transformation or cleaning prior to visualization. Here are tools you can use for data preparation. We recommend first trying to import and process your data in the same tool you intend to use for visualization. If that fails, pick the most appropriate option among the tools below. Contact the course staff if you are unsure what might be the best option for your data!

Graphical Tools

- Tableau Prep - Tableau provides basic facilities for data import, transformation & blending. Tableau Prep is a more sophisticated data preparation tool.
- OpenRefine - A free, open source tool for working with messy data.

Programming Tools

- JavaScript data utilities and/or the Datalib JS library.
- Pandas - Data table and manipulation utilities for Python.
- dplyr - A library for data manipulation in R.
- Or, the programming language and tools of your choice...

Grading

The assignment score is out of a maximum of 7 points. Each of the seven rubric components below is worth 1 point and is judged across your whole report. We will determine scores by judging the breadth and depth of your analysis, whether visualizations meet the expressiveness and effectiveness principles, and how well-written and synthesized your insights are.

We will use the following rubric to grade your assignment. Note, rubric cells may not map exactly to specific point scores.

Component	Excellent	Satisfactory	Poor
Breadth of Exploration	More than 3 questions were initially asked, and target substantially different portions/aspects of the data.	At least 3 questions were initially asked of the data, but there is some overlap between questions.	Fewer than 3 initial questions were posed of the data.
Depth of Exploration	A sufficient number of follow-up questions were asked to yield insights that helped to more deeply explore the initial questions.	Some follow-up questions were asked, but they did not take the analysis much deeper than the initial questions.	No follow-up questions were asked after answering the initial questions.
Data Quality	Data quality was thoroughly assessed with extensive profiling of fields and records.	Simple checks were conducted on only a handful of fields or records.	Little or no evidence that data quality was assessed.
Visualizations	More than 10 visualizations were produced, and a variety of marks and encodings were explored. All design decisions were both expressive and effective.	At least 10 visualizations were produced. The visual encodings chosen were largely effective and expressive, but some errors remain.	Several ineffective or inexpressive design choices are made. Fewer than 10 visualizations have been produced.
Data Transformation	More advanced transformation were used to extend the dataset in interesting or useful ways.	Simple transforms (e.g., sorting, filtering) were primarily used.	The raw dataset was used directly, with little to no additional transformation.
Captions	Captions richly describe the visualizations and contextualize the insight within the analysis.	Captions do a good job describing the visualizations, but could better connect prior or subsequent steps of the analysis.	Captions are missing, overly brief, or shallow in their analysis of visualizations.
Creativity & Originality	You exceeded the parameters of the assignment, with original insights or a particularly engaging design.	You met all the parameters of the assignment.	You met most of the parameters of the assignment.

Submission Details

This is an individual assignment. You may not work in groups.

Your completed exploratory analysis report is due by [time] on [day] [date]. Submit a link to your Google Doc report using [the submission form]. Please double check your link to ensure it is viewable by others (e.g., try it in an incognito window).

Resubmissions. Resubmissions will be regraded by the instructor, and you may earn back up to 50% of the points lost in the original submission. To resubmit this assignment, please use [the resubmission form] and follow the same submission process described above. Include a short 1 paragraph description summarizing the changes from the initial submission. Resubmissions without this summary will not be regraded. Resubmissions will be due by [time] on [day], [date]. Slack days may not be applied to extend the resubmission deadline. The teaching staff will only begin to regrade assignments once the Final Project phase begins, so please be patient.