MSBA 6130 Homework 1

Data Collection, Processing, and Visualization

Mochen Yang

*Please allocate sufficient time for all homework assignments as they usually take time to complete. Start early on assignments. If you have doubts, please email the instructor. This homework assignment is due on **October 11th before class** and it is worth 8 pts in total.*

1 Individual Part (3 pts)

This part of this homework should be completed on you own, without collaboration with other students. For all questions in this part, please complete them in an RMarkdown script, and submit the rendered PDF file on Canvas. Make sure that your R code and corresponding outputs are visible on the PDF file.

Question 1 (3 pts): This question is designed to help you practice basic data processing and visualization skills in R. You are encouraged (although not required) to use dplyr functions for relevant data processing tasks and ggplot2 functions for data visualization. You will use the ”hills.csv” dataset, which contains record times in 35 Scottish Hill Races. Each row contains the name, distance (in miles), climb (in feet), and record time (in minutes) of a given race. Specifically, please complete the following tasks:

a (0.25 pts) Import the ”hills.csv” dataset into R;

b (0.25 pts) Arrange data based on racing time in descending order, and print out the race name with longest racing time;

c (0.25 pts) Create a new variable, named “speed”, to record the record speed of each race (miles/minute);

d (0.25 pts) Create a new variable, named “High_Race”, that takes value 1 if a race has more than 1000 feet climb and 0 otherwise;

e (0.25 pts) Report the *average racing time* for high races (i.e., races with more than 1000 feet climb) and low races, respectively;

f (0.25 pts) Create a plot to show the distribution of record time for all races;

g (0.25 pts) Create a plot to show the relationship between race distance and record time;

h (0.25 pts) Create a plot to show the distributions of race distance for high races and low races respectively (on the same plot).

i (0.25 pts) Download ”hills2.csv” from Canvas. This dataset contains 94 Scottish Hill Races. Each row contains the name, distance (in kilometers), climb (in meters), and record time (in minutes) for both male and female contestants of a given race. Import this dataset into R (note that it has a delimiter that is not a comma);

j (0.25 pts) Remove rows that have missing values on record time;

k (0.25 pts) Merge hills2 data with the original hills data based on race name, keep all races in the original hills data (in database language, you should *left join* hills2 to the original hills data);

l (0.25 pts) Based on the merged dataset, create a plot to show the relationship between record time and climb (in feet) for male and female contestants respectively (on the same plot). Hint: to get the data into proper shape for visualization, you might want to take advantage of the “pivot_longer()” function in the “tidyr” package. Here is the documentation with examples: `https://tidyr.tidyverse.org/reference/pivot_longer.html`.

2 Group Part (5 pts)

This part of this homework should be completed in your assigned group. Each group should submit two deliverables: a technical document and a managerial document. Requirements for the two documents are specified after the problem description. Both documents should be submitted via Canvas. This part is worth 5 pts, but will be graded on a face value of 20 pts.

Question 3 (5 pts): XYZ is a nationwide retail chain that owns stores across most major cities in the U.S. Like many other businesses, XYZ has been severely affected by the decline in consumer foot-traffic due to COVID-19. The management team at XYZ is working to develop possible strategies that can mitigate the company's financial stress. For example, it is considering reducing the operations of some stores located in cities where COVID-19 cases are more widespread, and re-allocating the limited resources to other stores in cities where the pandemic is better controlled. To make these decisions, the management team needs relevant and timely information about COVID-19, such as how the cases are distributed across different geographic locations and change over time.

You are commissioned by XYZ to build a dashboard of data visualizations and provide evidence-based recommendations regarding what XYZ can do to survive the pandemic. There are a number of up-to-date public datasets that tracks various aspects of the COVID-19 (number of reported cases, deaths, and recovers; information about mask use; etc.), including but not limited to:

- Center for Systems Science and Engineering (CSSE) at Johns Hopkins University: https://github.com/CSSEGISandData/COVID-19;
- New York Times COVID data: https://github.com/nytimes/covid-19-data;
- COVID-19 Open Data: https://github.com/GoogleCloudPlatform/covid-19-open-data;
- The Economist: https://github.com/TheEconomist/covid-19-excess-deaths-tracker;
- Microsoft Bing COVID-19 data: https://github.com/microsoft/Bing-COVID-19-Data;
- COVID Tracking Data: https://github.com/COVID19Tracking/covid-tracking-data.

Specifically, please complete the following tasks:

a Determine what information of COVID-19, if properly visualized, can most effectively support XYZ's decision-making;

b Collect the relevant information using one or more of the public COVID-19 datasets. The links above represent a few high-quality resources. You are of course free to leverage other sources;

c Build a dashboard of data visualizations. This counts as the "technical document". See below for the specific requirements for this document.

d Present the key insights of your data visualization efforts as well as your recommendations to XYZ's management team. This counts as the "managerial document". See below for the specific requirements for this document.

Deliverable Requirements:

The **technical document** should be a PDF file or an HTML page that contains the dashboard of your data visualizations, generated by an RMarkdown script (as well as necessary dashboarding packages such as flexdashboard). The RMarkdown script should also be submitted as supplemental material. While the associated data file(s) should *not* be submitted, they should be available upon request. If you decide to build a "dynamic" dashboard with R Shiny and flexdashboard, you should submit the original RMarkdown script and some screenshots that demonstrate the dashboard.

For the **managerial document**, create 2 PowerPoint slides (excluding title slide, if any) that communicate a finding, insight, conclusion, or recommendation from your analysis. Each slide should:

- Include a clearly articulated headline that conveys information (not just a label) and **makes a point** based on your analyses. Remember you are advising the company on **what to do / what action to take**; you are not simply reporting the data.;
- Include a well-designed graph or chart to support your point;
- Include brief explanatory text in conjunction with the graph/chart;
- Demonstrate the six visual communication principles discussed in class (align form and content, focus attention, clear distractions, etc.)

Note that you *cannot* simply copy-paste your visualization dashboard into a few slides and submit them as the managerial document. You need to think about how to best organize and communicate your analyses and findings, and use the visuals from your dashboard in an informative and judicious manner.

On Canvas, you will find two examples of (reasonably good) managerial documents for this task. Please use these examples **only as a reference** — they do not reflect how the analyses *should* be conducted, nor are they the *only* ways to present the results. Of course, copying the examples entirely or modifying them only slightly as your own submission would be considered plagiarism and lead to a 0 pt score on this part.