STAT 341: Assignment 1

DUE: Friday January 24 by 11:59pm EST

NOTES

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically in .pdf format via Crowdmark. This means that your responses for different questions should be in separate .pdf files. Your .pdf solution files must have been generated by R Markdown. Additionally:

- For mathematical questions: your solutions must be produced by LaTeX (from within R Markdown). Handwritten and scanned/photographed solutions will not be accepted and you will receive zero points.
- For computational questions: R code should always be included in your solution (via code chunks in R Markdown). If code is required and you provide none, you will received zero points.
- For interpretation question: plain text (within R Markdown) is fine.

Organization and comprehensibility is part of a full solution. Consequently, points will be deducted for solutions that are not organized and incomprehensible.

QUESTION 1: Evaluating the population mid-range [20 points]

Consider the population $\mathcal{P} = \{y_1, \ldots, y_N\}$. The population mid-range is the midpoint of the range

$$a(\mathcal{P}) = a(y_1, \ldots, y_N) = \frac{y_{(1)} + y_{(N)}}{2}$$

and hence a measure of center. In this question you will investigate several of its properties.

(a) [3 points] Determine whether the mid-range is location invariant, location equivariant, or neither.

(b) [3 points] Determine whether the mid-range is scale invariant, scale equivariant, or neither.

(c) [3 points] Determine whether the mid-range is location-scale invariant, location-scale equivariant, or neither.

(d) [3 points] Determine whether the mid-range is replication invariant, replication equivariant, or neither.

(e) [3 points] Derive the sensitivity curve for the mid-range, given a population $\{y_1, y_2, \ldots, y_{N-1}\}$.

(f) [3 points] For the population below, plot the sensitivity curve from part (e) for $y \in [-7, 7]$. You may find the `sc()` function from class useful.

```
set.seed(341)
pop <- rnorm(10000)
```

(g) [2 points] Given all that you have learned in parts (a) - (f), state one thing that is *good* about thie mid-range attribute and one thing that is *bad* about the mid-range attribute.

QUESTION 2: Write a plot-making function [5 points]

Write a function called `matrix.plot()` that takes in a single input (called `df`), that is an $N \times m$ data frame containing *numeric* data. This function should produce as its output an $m \times m$ matrix of plots where:

- the diagonal plots contain histograms of the columns of `df`
- the upper triangle of plots are scatter plots between all pairs of columns of `df`
- the lower triangle of plots report the correlation coefficients between the pairs of columns of `df`
- all plots should be labelled with the headings provided in `df`

QUESTION 3: Spotify Top 30 Analysis [25 points]

Spotify, the popular music streaming service, organizes and classifies songs based on a wide range of properties (variates):

Variate	Description
`genre`	the genre of the track
`year`	the release year of the recording (note that due to vagaries of releases, re-releases, re-issues and general madness, sometimes the release years are not what you'd expect)
`bpm`	beats per minute - the tempo of the song
`energy`	the higher the value the more energetic the song
`danceability`	the higher the value the easier it is to dance to the song
`loudness`	the higher the value the louder the song
`liveness`	the higher the value the more likely the song is a live recording
`valence`	the higher the value the more positive the mood of the song
`duration`	the duration of the song (in seconds)
`acousticness`	the higher the value the more acoustic the song is
`speechiness`	the higher the value the more spoken words the song contains
`popularity`	the higher the value the more popular the song is

Available for us to study is the population of $N = 300$ Billboard Top 30 songs from 2010 - 2019 (inclusive). In addition to the song's title and artist, measurements on each of the 12 variates listed in the table above have also been recorded for each of these songs. This data is available in the `spotify.csv` file.

(a) [2 points] Using the `matrix.plot()` function you developed in Question 2, produce the summary graphic for `energy`, `danceability`, `acousticness`, and `popularity`.

(b) [3 points] Considering all variates (except `genre` and `year`), which three are most strongly correlated with popularity? For each variate, explain the nature of its linear relationship with popularity.

(c) [1 point] Using R, determine which are the Top 10 most popular songs.

(d) [2 points] Using R, determine which song is the shortest and which is the longest.

(e) [4 points] Let y denote the beats per minute (`bpm`) of a song, and let $a(\mathcal{P}) = \bar{y}$ be the attribute of interest. Define the influence of song u on $a(\mathcal{P})$ to be:

$$\Delta(a, u) = |a(y_1, \ldots, y_{u-1}, y_u, y_{u+1}, \ldots, y_N) - a(y_1, \ldots, y_{u-1}, y_{u+1}, \ldots, y_N)|$$

Construct an influence plot of Δ vs. observation number and identify the song with the largest influence on the average `bpm` attribute. Why is this song in particular more influential than all of the others?

(f) [3 points] Using R, determine which artists have appeared in the Billboard Top 30 five or more times. For each of these artists state the number of times they have appeared and calculate the average popularity score of their songs.

(g) [5 points] Construct the following plot:

- Make a scatter plot of `duration` vs. `year`, but where `year` has been jittered slightly
- Add to this plot 10 red triangles indicating the average song duration in each year.
- Connect these 10 red triangles with red line segments.
- Add appropriate labels and a title to the plot.

Does the duration of popular songs seem to have changed over time? If so, in which direction?

(h) [5 points] Construct a 1×2 plot which contains histograms of `energy` and `acousticness`. Using the `powerfun()` function from class determine a range of powers (values of α) for each variate which make its distribution more symmetric. Plot another 1×2 plot containing histograms of the transformed `energy` and `acousticness` variates using what you feel is the *best* value of α in each case. Make sure all of your plots are appropriately titled and labelled.