ALY 6000 Project 4
Independent Data Analysis

Instructions

Preliminary Work – Getting Started

Locate a dataset of interest. The following sites may help you identify data of interest to you and your work:

- The R Project for Statistical Computing
- Kaggle
- Data.gov

Your dataset should have at least 700 and no fewer than 6,000 records, as well as at least eight (8) variables.

Before beginning your exploratory data analysis, develop 3–4 data questions. These may or may not change as you explore your data in greater depth, but they will provide you with direction to begin your analysis. The following steps will walk you through a typical exploratory data analysis. Note that your analysis may differ based on the specific dataset you selected.

Part I – Exploring

1. Review any written description of your dataset. This is often referenced as the data dictionary.
2. Clean your data. Cleaning involves any task that prepares the dataset for analysis. This might include the following tasks:
    a. Renaming columns
    b. Managing NAs
    c. Correcting data types
    d. Removing columns or rows
    e. Manipulating strings

f. Reorganizing the data

g. Other steps that prepare your data

3. Determine descriptive statistics for interesting variables.

4. Produce visualizations from the raw data that identify and highlight interesting aspects. These can include bar charts, histograms, line graphs, scatter plots, etc. Be sure the chosen graph best represents the information.

Part II – Expanding

1. Create new variables that better capture the data you want to report. For example, if the data shows yearly sales by month, you might calculate the month-to-month increase or decrease in sales as a new column.

2. Group, summarize, rank, arrange, count, or perform any other useful operations to create new data frames that provide access to different views of the data.

3. Extract the most interesting data results and produce visualizations that best communicate these results.

Part III – Communicating

1. Report what you have learned about your data. Identify 3–5 observations or follow-up questions that you could explore in the future.

2. Complete all data management tasks in R.

Submission Guidelines

- Follow the above instructions to produce a slide deck that tells the story of your data in 5–8 slides (not including Title and reference list slide) through the use of **descriptive statistics and visualizations**. **Properly cite** all sources using **APA style**. Remember:
  - Visualizations are the **primary vehicle** to convey information in an analytics presentation.
  - Include written information in the **Notes section** on each slide that **connects** to the visualization's key points in a **concise** manner.
- Submit two (2) files under the assignment in Canvas with the following filename conventions:

A slide deck: LastName-FirstName-Project4.pptx or LastName-FirstName-Project4.pdf

An R Markdown file: LastName-FirstName-Project4.Rmd