Fall term project: Expectations

Do people who type fast on a keyboard also type faster on their mobile phones?

There are two components to this project: (A) the statistical analysis and (B) the written report.

A – Analysis

When you get the data, you will begin by loading it into R and cleaning it so you can proceed with exploration and analysis. You will then conduct exploratory analyses, followed by a more thorough statistical analysis.

Your analysis must be reproducible, so make sure to submit the data file you are using in submissions A1 and A2 if you make any changes to the original file that is shared with you (in other words, if you make all of your data manipulation in R and include your data cleaning code in your .Rmd, you don't need to include your clean data file because it will be fully reproducible from the original data file). Also make sure to include any necessary calls to `library()` at the beginning of your .Rmd so it's clear what packages you're using.

Part A1: Exploratory Data Analysis (EDA)

You will complete an exploratory data analysis including the following:

- Data cleaning: A summary of the decisions you made when cleaning the data (~200 – 500 words, try to limit to about a page – bullet points are OK for this section).
- Examination of relevant variables: You will examine relevant variables using graphical and/or other summary methods to observe interesting, surprising, or unusual patterns in the data
- Preliminary insights: Produce **one key figure** which provides initial insight into the research question, as clearly as possible AND a description of the plot and what it means in context (100-200 words).

While this EDA submission will not be graded explicitly, it will be submitted and count towards your participation grade. In the next Thursday class, you should also have your EDA with you as we will discuss it in small groups and as a class to make sure everyone is in a good place to start their analysis.

Part A2: Statistical Analysis

You will conduct a thorough statistical analysis of the data with an aim to investigate the research question. You should be clear about the statistical method(s) you consider, and perform diagnostics to assess whether model assumptions are reasonable and whether model fit is adequate. While you will include your annotated .Rmd (and the resulting pdf file), you will also include a section at the beginning of your document which summarizes your final model(s) – you should not have more than one model addressing the same research question, but may have a model addressing the primary research question and another model addressing a secondary question. Your statistical analysis document should have the following main headings:

- Introduction: Brief summary of the research question and the context of the data, in full sentences (~100-200 words)
- Data cleaning: A summary of the decisions you made when cleaning the data (~200 – 500 words, try to limit to about a page – bullet points are OK for this section). *Note: You should also have this section completed from your EDA submission, but you may have made some changes since then, which should be reflected in this submission.*

- Method(s): Describe the statistical method(s) you used to investigate the research question(s). Describe the method(s) and explain why it is appropriate, including commenting on key assumptions (~100-200 words).
- Results: Report on the results of applying your selected model(s) to the data. Interpret the statistical findings in terms of the research context (~100-200 words). You should also include a visualization that supports communicating your results. This should not be a diagnostic plot, but rather a visualization that would be interesting/relevant to a collaborator. This may be similar to the visualization you produced in your EDA, but doesn't need to be. It must be compatible with the statistical model and results you are interpreting.
- Full statistical analysis: This section will include the R code you used to conduct your statistical analyses. This section of your document should make sense to the reader, without requiring the reader to go look at the full code itself (although all code must be included in the .Rmd document); it should be enough for the reader to read the written commentary (outside of the code chunks) and glance at the plots and output. The R code should also be well documented, so the reader can follow your work.

Audience for this submission: Your TA, who is familiar with advanced statistical methods and R.

B - Written Report

You will produce a written report, addressed to a researcher whose only experience with statistics is an introductory course (at the level of STA220). The audience of the report is the researcher, and it should be written as if he or she collected data to answer the research question. There are no rules for how long your report should be, but I expect that it will be about 5-7 pages long (including plots and tables). Note that you will submit a draft report on Friday November 18th. Approximately a week after that, you will receive feedback on your draft, and you'll have approximately one week to incorporate the feedback into your final version.

Refer to the evaluation rubric for more guidance about the expectations for your written report.

Key Dates

Component	Files to submit	Date	Weight
Your raw data	Your raw data (format to be decided in class)	Tuesday September 27	Participation
Full class data	N/A	Wednesday September 28	N/A
EDA	.Rmd + .pdf + data file	Wednesday October 12	Towards participation grade (but will be discussed in class on Oct 13)
Statistical analysis	.Rmd + .pdf + data file	Friday October 28	6%
Draft written report	.pdf	Friday November 18	4%
Final written report	.pdf	Friday December 2	5%