

First Mid-Term Project

Information: Read-Me Document

The first section of the course has focused on turning data into *information* about the city, its people, and its neighborhoods. We have also learned how to clean and modify certain variables to make the data set easier to analyze. This section will conclude with the first midterm project, which will have two pieces: the individual portion will be a Read-Me Document that acts as a first view into the contents of the data set and how it might be used; the group portion will be an updated version of the data set, loaded into the Dataverse, with accompanying updated data dictionary and R syntax.

Project Value: 15% of final grade

Due: End of Module 04

Individual Portion:

Think of the Read-Me Document as something someone visiting a public data portal—browsing for interesting data, downloading them for the first time, etc.—would read as a quick look into the contents of the data and the analytic possibilities they might offer. It should be both informative and enticing, and should include:

- **A brief overview of what the data set contains.** This should be a *short* paragraph composed of short sentences describing the source of the data and the type of information it holds. This will look a lot like the opening paragraph of the Data Dictionary that you have, though possibly even shorter.
- **“Fun facts.”** Provide about 5-10 pieces of information that are notable and illustrate the nature of the data set. These should be both engaging and potentially useful to someone looking to analyze the data. For example: “The 311 system categorizes requests for service into 218 types, the most common being General Request.” “The neighborhood with the most 911 calls in 2013 was Dorchester.”
- **Visualization.** Include at least two visualizations that illustrate some aspect of the data set that you find particularly noteworthy, complementing the “fun facts.”
- **No more than two pages.** Remember that this is something that a visitor will want to read quickly to decide whether it is of interest.

Rubric (Total 10 pts.)

Communication of Data Content: 3 pts.

Facts: 2 pts.

Figures: 2 pts.

Structure: 2 pts.

Details (Grammar, etc.): 1 pt.

Group Portion:

The group will work together to combine all the modifications made by group members into an updated data set and documentation. It will include three basic parts that will be uploaded to the Dataverse:

- The record-level file with all new and modified variables included.
- An updated Data Dictionary that describes the new variables and, if necessary, modifies the description of any already existing variables.
- Annotated R syntax clearly articulating steps for all data cleaning and variable creation. This should be efficient and complete such that the code could be run all at once on raw data to recreate the updated data set.

Rubric (Total 5 pts.)

Data: 2 pts.

Syntax: 2 pts.

Updated Documentation: 1 pt.