EPA 1315. Data Analysis and Visualization
Final Project Description

Due Date: 23:30 Friday 5th November 2019

1. The Task in Brief

The objectives of this assignment are to assess the following capabilities:

1. Problem Understanding. Developing data, graphics and analyses tailored to a given problem statement
2. Graphics and Visualization. Creating graphics or visualizations in R
3. Data and Preparation. Using R to clean and prepare data
4. Modeling. Using graphical models to formulate and implement data analyses
5. Reporting. Using the CRISP-DM process to document and manage your data science project

Demonstrating these objectives will require that you select, clean, visualize and model a dataset of your choice. You will also report on your findings using the CRISP-DM process. Further tips, tricks and advice in completing these tasks are shown below. A grading rubric is also provided.

The project is worth 29% of the grade, and is due as noted above. No late assignments are accepted, but project retakes in subsequent quarters may be permitted in the case of incompletes or course failures. You may work alone, or have up to four partners on this project. I require a short acknowledgment section where the tasks and responsibilities undertaken for all partners is listed.

2. Tasks and their Reporting

The following sections discuss the five assessment tasks as described above. The section discusses tips, tricks and reporting. The grading rubric is provided in the final section of this document.

2.1 Problem Understanding

A problem understanding section will be provided to you. You may develop a different business understanding report if necessary. You should demonstrate an understanding of your audience appropriate to the brief provided.

2.2 Graphics and Visualization

Several kinds of graphics may be appropriate here. The most essential graphic involves data exploration. You may also want to tailor a graphic to communicate the necessity or urgency of the problem to your decision-maker or their stakeholders. You may also want to tailor a graphic to show the geographic scope of the problem.

2.3 Data and Preparation

When completing this assignment, choose a data set from the World Bank open data set. I particularly recommend the World Development Indicators database. Choose a set of variables from this source. Other data sources may be permissible; please discuss with the instructor. A suitable selection might involve 50 national cases over a single year. There should be four to eight variables. I'm looking for a functional choice in the selection and preparation of the data, rather than complex choices.

2.4 Analysis

This must be a graphical model implemented in rjags. Choose one of the three most basic analyses as discussed in class – a chain, collider or fork. Other non-Bayesian analysis alternatives are not suitable for this assessment. (You may use other tools than rjags though, as you wish).

Be modest in introducing complexity to your model. If you are ambitious, please consider performing your modelling in two steps. Begin with a simple mode and lock-in your results. Only then add a second phase to your modelling exercise where you expand the model.

2.5 Reporting

This section is necessarily longer than the rest since it is where you show your work. This is also where you demonstrate your understanding of the CRISP-DM process. Document your results in a short brief of up to 6500 words.

This word count is only intended as a guideline. Your document may be more or less than this. Given the ten subsections of the report (numbered below) you will need to keep the report quite succinct. I don't expect extensive original research beyond the framing of the problem, and your statistical model.

In the material that follows I refer to the CRISP-DM manual. The manual refers to separate reports – here it is more appropriate to call these sections and subsections of your final project. Given our public sector and policy emphasis it is appropriate to rename "business understanding" to "problem understanding."

Given this the policy brief should contain the following sections and subsections:

1. *Business understanding section,* as specified in the CRISP-DM manual. This report will be provided to you. You may reference the provided report; it is not necessary that you include this in your final. This is a good place to include graphics to ease communication with the decision-maker or their stakeholders. Likewise it is a good place to include graphics to show the geographical boundaries of the problem. You may develop a different business understanding report if necessary.
2. *Data understanding section,* as specified in the CRISP-DM manual. This section has four subsections – the initial data collection subsection (1), the data description subsection (2), the data quality subsection (3), and the exploratory analysis subsection (4). Krushke asks the following Bayesian modelling question "identify the data relevant to the research questions." You will address that question here. This is the appropriate place to include your visualizations for data exploration.
3. *Data preparation section,* as specified in the CRISP-DM manual. This contains only one part, the dataset and dataset description subsection (5).
4. *Modeling section,* as specified in the CRISP-DM manual. This section contains five subsections – the test design (6), models (7), parameter settings (8), model description (9) and assessment subsections (10). Krushke asks three Bayesian modelling questions, which you should include here: define a descriptive model for the relevant data; specify a prior distribution on the parameters; use Bayesian inference to re-allocate credibility across parameter values.

The full CRISP-DM process also involves evaluation and implementation tasks and reports, but you are not responsible for these in this class.

3. Grading Rubric

The following grading rubric will be used when assessing final project. Each of the five tasks are of equal importance in assessing the final, although not necessary equal in required effort. The data and preparation task and the reporting task probably require most effort. The graphics and visualization and the analysis tasks are good places to invest extra effort and thereby excel.

	5 (or less)	6	7	8	9 (or more)
Task 1. Problem Understanding	A problem is not selected; the data and graphics are not adapted to the needs of the decision-maker	A problem is selected, but poorly described	A problem is compellingly described. Links are made by the appropriate selection of data or the appropriate selection of graphics	A problem is compellingly described. Both data and graphics are appropriately used.	A problem is compellingly described Both data and graphics are compellingly used.
Task 2. Graphics and Visualization	Not presented	Charts are presented using base R	Charts are presented and customized; the insight from the chart is described in text	Custom graphics are included using ggplot or other packages; the graphics are tailored to for decision-makers, analysts or stakeholders	Multiple effective graphics are shown throughout the report
Task 3. Data and Preparation	Data was not selected or prepared.	Data is selected, but its appropriateness for the problem is not discussed.	Data is selected and appropriately used. Inappropriate assumptions about the data are introduced.	Appropriate cleaning was performed and discussed. This is fully evidenced using the CRISP-DM reporting mechanisms.	Exceptional efforts were made to collect additional data or to fuse multiple data sets.
Task 4 Analysis	A graphical model in R is not presented.	The model selected is not discussed	The model selected is discussed	The model selected is strongly motivated; the Bayesian assumptions are fully discussed	The model selected is strongly motivated and additional insight is provided
Task 5. Reporting	The report is incomplete.	Discrepancies or limitations of reporting given the CRISP-DM framework are discussed.	The reporting is complete; the CRISP-DM framework is used in a productive manner.	The reporting is both concise and complete.	The report invites clear next steps for additional analysis or decision-making.

Revision Notes. 6 November 2018 -- Due date corrected (Friday is the 9th). Minor typos.