DATA 200 Graduate Project
Topic 1: Dataset A

Anya Michaelsen (3034964414)

Contents

1 Report Requirements

The narrative notebook should include figures sparingly to support specific claims. It can include runnable components, but it should not have large amounts of code. The length of the report should be 8+/-2 pages when it is printed as a PDF, excluding figures and code.

2 Background

COVID background information... vaccine timeline, US response

2.1 Research Questions

Clearly stated research questions and why they are interesting and important. You must include at least one research question involving at least one or more datasets from one of the topics we provided, but you may include additional research questions about each individual dataset. At least one of your research questions has to include a modeling component, e.g., can we build a model using climate data to predict growth in COVID-19 cases accurately?

2.2 Literature Review

A brief survey of related work on the topic(s) of your analysis and how your project differs from or complements existing research.

3 Methodology

Methodology: carefully describe the methods you use and why they are appropriate for answering your search questions. It must include

- a brief overview of causal inference, which should be written in a way such that another student in Data 100 who has never been exposed to the concept can carry out the analyses involving the datasets in your project.
- a detailed description of how modeling is done in your project, including inference or prediction methods used, feature engineering and regularization if applicable, and cross-validation or test data as appropriate for model selection and evaluation.

3.1 The Data

If applicable, descriptions of additional datasets that you gathered to support your analysis.

3.2 Data Processing

Lorem Ipsum

3.2.1 Data Cleaning

Lorem Ipsum

3.2.2 Feature Engineering

Lorem Ipsum

3.2.3 Causal Inference

a brief overview of causal inference, which should be written in a way such that another student in Data 100 who has never been exposed to the concept can carry out the analyses involving the datasets in your project.

https://data102.org/sp20/assets/notes/notes13.pdf

want to determine effects... but can have confounding effects (example with graphs) want to control for these confounding variables, but not mediators (intermediate variables creating a pathway of effect from one variable to another).

Can hold constant confounding variables and look at the effects within this population?

3.2.4 Modeling

ex-ante vs post-ante forecasting article:
https://otexts.com/fpp2/forecasting-regression.html#ex-ante-versus-ex-post-forecasts

4 Results

4.0.1 Dataset Findings

Interesting findings* about each dataset when analyzed individually. Include visualizations and descriptions of data cleaning and data transformation necessary to perform the analysis that led to your findings.

Interesting findings* involving your datasets. Include visualizations and descriptions of data cleaning and data transformation necessary to perform the analysis that led to your findings.

* Examples of interesting findings: interesting data distributions and trends, correlations between different features, the relationship between the data distribution for the general population and specific datasets (e.g., the gender distribution in the census dataset vs. in the mental health dataset), specific features that are notably effective/ineffective for prediction.

4.0.2 Model Findings

Analysis of your findings to answer your research question(s). Include visualizations and specific results. If your research questions contain a modeling component, you must compare the results using different inference or prediction methods (e.g., linear regression, logistic regression, or classification and regression trees). Can you explain why some methods performed better than others?

4.0.3 Limitations and Future Research

An evaluation of your approach and discuss any limitations of the methods you used.

Describe any surprising discoveries that you made and future work.

5 Appendices

6 References