Final Project Description

Advanced Probability and Statistics for Data Science

Data science professionals typically need to learn new techniques during their careers. This assignment gives you exposure to the skill of using what you know to facilitate learning something new.

The final project consists of a presentation explaining a new method, not covered in this course or in Introduction to Probability and Statistics for Data Science, incorporated into a full analysis of a question based on a dataset, a paper explaining the analysis beginning with a brief executive summary of the analysis, and R code for the analysis. When consistent with ethical obligations, the data set should accompany the project, either as a link or an uploaded csv or RData file.

The subject you investigate in your final project may be controversial. The methods you use and your data should be consistent a responsible application of data analysis and inference.

While several methods not covered in class may be used, the explanation should focus on one method not covered in class. The explanation of this method should include the principles on which the method is based, a purpose of the method, requirements for application, and basic diagnostics, where applicable. Any other new methods should be explained in the presentation and in the paper to the point that colleagues would understand their application. The analysis should be original, not a replication of a tutorial. If you would like to do a replication study, please discuss this with the instructor

Rubric for presentation

* Data source and definitions explained 10
* Appropriate exploratory data analysis performed 10
* Research question presented 10
* Method for addressing research question explained 10
* Primary method explained in principle 10
* Application of primary method explained 10
* Data satisfaction of requirements of method demonstrated 10
* Primary new method applied and interpreted correctly 10
* Supporting visualizations provided 10
* Presentation style shows sufficient preparation in organization 10

and familiarity with topics addressed

Rubric for paper (maximum of 6 pages)

The executive summary should start the paper. It should be meaningful to someone with familiarity with the subject area from which the data are drawn, but without data science expertise.

The body of the paper should be addressed to a colleague who is interested in your conclusions and is also interested in applying the primary method in other contexts in a well-informed process.

* Executive summary explains research question and conclusion 10
* Data source and definitions explained 10
* Appropriate exploratory data analysis performed 10
* Method for addressing research question explained 10
* Primary method explained in principle 10
* Application of primary method explained 10
* Data satisfaction of requirements of method demonstrated 10
* Primary new method applied and interpreted correctly 10
* Supporting visualizations provided 10
* Body of paper appropriate for entry level professional 10

Possible methods for use in project include but are not limited to the following:

Power Analysis

Robust Regression

High Dimensionality methods (model building when many explanatory variables are present)

-Regression trees

-Principal Component Analysis

-Partial Least Squares

-Random forests

` -Factor analysis

Clustering

- Latent Class Analysis

- DBSCAN

Non linear methods

- further study of generalized linear models or general additive models

-Spline estimation

-principal curves

Bayesian Analysis (Big topic, scratch the surface)

Sample Mean Tests for multivariate continuous responses

Time series data

Hierarchical Linear Models

Mixed models

ROC curves

Propensity Scores

Methods for data with dependent error structures (geospatial and longitudinal data)

Generalized estimating equations (GEE)

SEM (Structural Equation Modeling)

IRT (Item Response Theory)

Multinomial logit models (discrete choice analysis application)

Extreme Value Theory

Beta regression (Regression for proportion data)

Analysis of circular data (methods for data emphasizing magnitude and direction)

Data sources:

* "datasets" package in R using: require(datasets), help(package=datasets), then help for individual data set. You can use the data set directly by name. To have it in your environment, use data('data set name')
* <https://research.stlouisfed.org/fred2/>, Federal Reserve Data
* <http://www.kdnuggets.com/datasets/index.html>, collected data sets for data analysis and data mining
* <http://community.amstat.org/stats101/home>
* <http://wise.cgu.edu/helpful-links/data-sources/>, a master list of possibilities
* <http://webserv.jcu.edu/math/faculty/TShort/Bradstreet/index.html> , drug development data sets
* <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html>, index of data sets provided with R
* <http://www.nhtsa.gov/FARS>, FARS (Fatal Analysis Reporting System), National Highway Traffic Safety Administration: summary data and raw data for U.S. traffic fatalities, 1975–present. (alternative: <https://cdan.nhtsa.gov/> )
* <https://ocg.cancer.gov/programs/ctd2/data-portal>, cancer-related data
* <https://toolbox.google.com/datasetsearch>, a search tool for data sets
* <https://www.ipums.org/> large collection of US and international census and survey data. Requires login and data extraction
* <https://mappingpoliceviolence.org> data curated by activists
* <https://www.bjs.gov/> Bureau of Justice statistics
* <https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases> international covid data
* <https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv> USA data, county level
* <https://timss2019.org/international-database/> Trends in International Mathematics and Science Study (TIMSS)
* <https://nces.ed.gov/datalab/onlinecodebook/> National Center for Education Statistics
* <https://worldhappiness.report/> World Happiness Report
* <https://gssdataexplorer.norc.org/> The General Social Survey (GSS), project of NORC, University of Chicago
* <https://www.icpsr.umich.edu/web/pages/> ICPSR maintains a data archive of more than 250,000 files of research in the social and behavioral sciences
* <https://community.amstat.org/dataexpo/home> Data Challenge Expo, data and problems
* <https://www.pewresearch.org/download-datasets/> Pew Research Center Surveys
* <https://vincentarelbundock.github.io/Rdatasets/articles/data.html> Summary of R data sets
* <https://fatalencounters.org/> database of people killed during interactions with police
* <https://www.washingtonpost.com/graphics/investigations/police-shootings-database/>
* <https://www.apha.org/topics-and-issues/health-equity/racism-and-health> See “National Data and Useful Links” tab under “Resources and Tools”
* <https://www.cdc.gov/healthyyouth/data/yrbs/data.htm> The Center for Disease Control and Prevention Youth Risk Behavior Survey (YRBS) has survey responses for middle school and high school students including information about risky behavior.
* <https://www.cbioportal.org/study/summary?id=msk_met_2021> MSK-MET (Memorial Sloan Kettering - Metastatic Events and Tropisms) is an integrated pan-cancer cohort