# Project Pseudo Code and Outline

Johns Hopkins University — 625.725.81.SU22(Theory of Statistics I)

**Jack Moody**

**jmoody11@jhu.edu**

# Project Pseudo Code and Outline



## 1 Project Pseudo Code Requirements by Instructors

Pseudo-code and Outline: to help with your planning and document your progress to this date. Use this to help you determine what parts will be the most challenging to complete on-time.

## 2 Topics to be tackled

1. Template to use
   (a) I have decided to use this template, I feel it is clean and professional.
2. What data to use
   (a) I am still deciding between census and heart-disease/ NHANES . I will likely use heart-disease and NHANES cause I have no experience with it yet.
3. Possible papers to pull from
   (a) Sharma, Pratyush, Marko Sarstedt, Galit Shmueli, Kevin H. Kim, and Kai Oliver Thiele. "PLS-based model selection: The role of alternative explanations in information systems research." Journal of the Association for Information Systems20, no. 4 (2019): 4.
   (b) Jung, Yoonsuh. "Multiple predicting K-fold cross-validation for model selection." Journal of Nonparametric Statistics 30, no. 1 (2018): 197-215.
   (c) Methods and Criteria for Model Selection by Joseph B. Kadane and Nicole A. Lazar
   (d) Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning by Sebastian Raschka
   (e) Statistical Analysis based Hypothesis Testing Method in Biological Knowledge Discovery by Md. Naseef-Ur-Rahman Chowdhury, Suvankar Paul, and Kazi Zakia Sultana
4. Coding language, packages, IDE to use
   (a) Language: Python and perhaps R
   (b) Packages:
      i. Python $\longrightarrow$ Scipy, Seaborn, Numpy, Matplotlib, Pandas, Statsmodels
      ii. R $\longrightarrow$ ggplot2, data.table, dplyr, tidyr, plotly, knitr

# Project Pseudo Code and Outline

(c) IDE: Jupyter Notebook, I want to make this as interactive and human readable as possible for ease of reproduciblility and demonstration. Recall that Jupyter notebook supports both python and R. This will be put into a github repository for accessibility.

5. Anticipated deliverables by order of deadline
   (a) Bibliography
   (b) Full Methods
      i. What approaches/ tests I chose, and which papers influenced those choices and why.
      ii. Jupyter Notebook with reproducible and readable code
      iii. Cleaned data with steps for how I cleaned and organized it
   (c) Final report
6. Anticipated timeline/ struggles
   (a) I believe that a good amount of time needs to be dedicated to the bibliography to ensure I have enough source material. After that, I will need to spend the most amount of time on the methods since that is where the bulk of working with data and actually trying things out will happen. I will likely have to work that concurrently with the bibliography to ensure I have enough time. While doing the methods, it will be paramount to take good notes so I can ensure I don't forget anything that I did during the writing process. This will pose a challenge because it is always easy to let note taking fall to the wayside. Overall, I believe the full methods will be the hardest to complete on time. If I do enough work on that, then the full write up will be easier to manage and write.