CIST 2500: SEMESTER PROJECT INSTRUCTIONS

The semester project for this class is entirely individual (there's some exceptions to sharing of datasets found below). The goal of the project is to allow each student to apply the concepts learned in class to the real-world. After all, statistics is highly applied, and learning concepts and theories in class is necessary but insufficient for actually performing these same tasks on your own in future courses or later in your career. To accommodate the wide variety of student interests, prior knowledge, and priorities, students will be allowed to select from 2 main options for completing this project. For each option, the deadlines stay the same, and the workload is approximately equal, controlling for prior knowledge. I encourage you to choose the option that you believe will be most beneficial for your future concentration in your major, or anticipated career trajectory.

Option 1: Applying Statistical Tests to a Real-World Dataset

For this option, students will select a pre-existing dataset (secondary data) to work from. The following is a list of sources for getting good secondary datasets, but you may choose from another source if appropriate. All students will get feedback on their chosen dataset (Project Checkpoint 1), so you will get a chance to finetune your selection or choose a different one if you desire. Note that some datasets may be in a .sav or other file format. Contact Dr. Toh if you need help converting these files to a .csv or .xlsx:

- Pew Research: https://www.pewresearch.org/internet/datasets/ (click on each title to access the report, and then click on "download dataset" on the toolbar on the right.
- Data.gov: https://catalog.data.gov/dataset?q=&sort=views_recent+desc
- Global Health Observatory (WHO): https://apps.who.int/gho/data/node.home
- FBI Crime Data Explorer: https://cde.ucr.cjis.gov/LATEST/webapp/#
- NOAA Climate Datasets: https://www.ncei.noaa.gov/cdo-web/datasets (access is through FTP browsing, and datasets are very large. You'll need to sift through quite a bit to find something usable).

Activities to Complete

Once you have selected the dataset you would like to use for your project, perform the following tasks:

1. Review the dataset by taking note of the type of data contained in the dataset, the variables included in the dataset and their associated levels or values noted in any corresponding documentation, the size of the dataset (the number of entries or rows).
2. Read any associated reports or materials referencing the dataset. This is important for understanding how the data was collected, the limitations of the dataset, as well as any existing analyses that have already been conducted.
3. Based on #2, develop ***at least two*** of your own new research questions that are not found in the associated reports or materials. These questions can be inspired by these reports or materials, other methods in different domains, your own personal interest in the topic, or through discussion with Dr. Toh.
4. **Project Checkpoint 1:** Once you have successfully chosen a dataset for analysis and developed preliminary research questions, write a brief report due at the start of the semester outlining your goals for your project, and preliminary tasks needed to clean up the dataset and make it ready for analysis. Describe how your choice of dataset and research questions has been informed by your interests, academic concentrations/minors, and your intended career path.
5. ***Perform the same analysis methods or statistical tests that were conducted for the following Practice Portfolios.*** The choice of test and analysis methods depend on the type of data that you have (#1 and #2) and the research questions that you have about this dataset (#3):
    a. Chapter 2: Descriptive Statistics: Tabular and Graphical Displays
    b. Chapter 3: Descriptive Statistics: Numerical Measures

c. *(Optional, depending on the size of your dataset)* Chapter 7: Sampling and Sampling Distributions

d. Chapter 9: Hypothesis Tests

e. Any ***two*** of the tests:

i. Chapter 10 Inference about Means and Proportions with Two Populations

ii. Chapter 11: Inferences about Population Variances

iii. Chapter 12: Tests of Goodness of Fit, Independence, and Multiple Proportions

iv. Chapter 13: Experimental Design and Analysis of Variance

6. **Project Checkpoint 2:** Write a report summarizing activities #1 through #4 above. Notably, your choice of dataset and research questions may have changed in Project Checkpoint 1, so describe how these evolved over the course of the project. You are free to use any font, formatting, page length, images, tables, diagrams, screenshots for your report. However, at a minimum you should include the following sections in your report:

a. Narrative regarding the process of choosing your dataset, the properties of the dataset (#1), and summarize the methods that were used to develop the dataset (#2).

b. Your research questions developed in #3. Describe the rationale or your theory behind these questions, and use the information found in the associated reports and materials as motivation (#2).

c. The procedure and results of the analysis methods or statistical tests performed for Chapters 2, 3, 7 (if you're selecting a subset of datapoints from a large dataset), 9, and two out of Chapters 10 through 13.

d. Summarize the findings of all your results from #4. Perform a comprehensive interpretation of your results following guidance from classes and assignments. Simply rehashing the results of the statistical tests is not sufficient. You need to use the results of these tests to answer the research questions that you formulated in #3, describing in your own words what the results of your project mean.

e. Reflect on your experience using real-world dataset of your choice. Compare this with your experience working on in-class datasets. What types of questions can we ask of different datasets, and how can we make that decision during the process of analysis? What aspects of your chosen dataset do you still have questions about or need help with? What's next for you in terms of your development of your statistical knowledge?

Option 2: Learning how to use R or Jupyter Notebook for Statistics

The focus of this option is to teach yourself how to use modern scripting languages and platforms for statistical analysis. This is a great option for someone interested in pursuing Data Science as a concentration or career path, since scripting languages are the de facto standard for data science applications (since we work with very large datasets, and platforms like Excel are simply not practical). In many ways, after acquiring a solid understanding of the concepts and theories covered in class using a simple graphical interface like Excel, learning a scripting language to perform the same tests is not typically very challenging, especially if you have some baseline programming knowledge and experience. However, learning the basics first is important since these scripting languages and platforms can obscure the step-by-step tasks and parameters needed to run these tests, since the focus of these platforms is on efficiency and resource savings, not on learning.

How to Choose between R or Jupyter Notebook?

In general, the ***R Statistical Package*** has been the standard choice for data science platform for the last few decades. It is an Open Source statistical package that is maintained by the Cran open source project and the participating community members. Anyone can submit a change or refinement to this project, and through the years, a very comprehensive list of statistical packages and functions have been developed on this platform. Researchers publishing work in peer-reviewed journals tend to lean towards using R because of this very large and validated set of statistical packages that have been "vetted" by the community. R is a type of scripting language, which means it is very useful for working very large datasets that will be tedious to manipulate and explore in a tabular-based platform like Excel. The language itself is very easy to use and you will download R Studio as the editor and GUI for working with R. R Studio is well suited to exploring large datasets and has been refined over the decades to have features most useful for advanced data science applications.

Jupyer Notebook, on the other hand, is a platform that can be used to deploy a variety of scripting languages such as Python (and also R!). Python is the most typical scripting language used for data science outside or R—for much the same reasons I provided above (it can process a large amount of datapoints and recurring functions or operations very quickly). However, since Python is used by a wider range of computing professionals on applications outside of data science, it has enjoyed widespread adoption for those looking to get started with data science. The language itself was not built specifically for data science, so the syntax may be less intuitive or easy to learn compared to R, but if you're already familiar with Python, then the learning curve is fairly gentle to get started using Jupyter Notebook. Jupyter notebook is simply a platform for organizing your Python scripts, storing and accessing your datasets, and managing the installed packages on your machine. It is fairly user friendly and modern compared to R Studio, but there are less "quality of life" amenities that make data science tasks easy to perform. Another consideration in choosing between these two languages is the relative age of the two packages. R is very established and most any test you want to run can be done at a standard that is acceptable for peer-reviewed publications. On the other hand, Python is fairly new as medium for data science, so some statistical packages are still under development, have fairly frequent updates to them (resulting in different numerical results sometimes!) and documentation is less consolidated compared to R. This means you may be spending more time searching online for the specifics behind installed packages (e.g., numpy, pandas, etc) and the mechanics behind how the tests are conducted, than say, if you chose R. The one upside to learning Python for statistics is that there's a greater flexibility in what you can do with your scripts and outputs since connectivity with other applications is very high. R, on the other hand, is "standalone".

Activities to Complete

1. Use the following resources (or others that you find) to install and set up your chosen platform on your machine:
    a. **Install and Learn R:**
        i. Download and install R and R Studio on your machine: https://rstudio-education.github.io/hopr/starting.html
        ii. R for Data science online textbook: https://r4ds.had.co.nz/index.html
    b. **Install and Learn Jupyter Notebook:** Tutorial for starting your own Jupyter Notebook and installing the Anaconda distribution: https://www.dataquest.io/blog/jupyter-notebook-tutorial/
2. **Project Checkpoint 1:** Once you have successfully installed a working copy of R or Python on your machine, and have verified that it functions as expected, write a brief report due at the start of the semester outlining your goals for your project, and preliminary tasks needed to set up your workspace with your chosen platform. Describe how this choice has been informed by your interests, academic concentrations/minors, and your intended career path.
3. ***Perform similar tests to what was performed in class for the Practice Portfolios*** for the following chapters (use the same datasets provided for these practice problems in your chosen platform):
    a. Chapter 2: Descriptive Statistics: Tabular and Graphical Displays
    b. Chapter 3: Descriptive Statistics: Numerical Measures
    c. Chapter 10 Inference about Means and Proportions with Two Populations
    d. Chapter 11: Inferences about Population Variances
    e. Chapter 12: Tests of Goodness of Fit, Independence, and Multiple Proportions
    f. Chapter 13: Experimental Design and Analysis of Variance
4. **Project Checkpoint 2:** Write a report summarizing activities #1 and #2 above. You are free to use any font, formatting, page length, images, tables, diagrams, screenshots for your report. However, at a minimum you should include the following sections in your report:
    a. Narrative regarding the process of deciding between R and Jupyter Notebook. Describe how this choice has been informed by your interests, academic concentrations/minors, and your intended career path (repeat, or change from Project Checkpoint 1)
    b. The procedure and results of the analysis methods or statistical tests performed for Chapters 2, 3, 10, 11, 12, and 13. Include code blocks, screenshots, visuals produced in R or Jupyter Notebook, and any markdown files you created for organizing your results.

c. Summarize the findings of all your results from #2. Perform a comprehensive interpretation of your results following guidance from classes and assignments. Simply rehashing the results of the statistical tests is not sufficient. Compare and contrast any differences in results obtained from using Excel for these same tests. Hypothesize why these differences might exist based on online documentation for the corresponding python package or distribution notes.

d. Reflect on your experience using a scripting language for statistical analysis. Compare this with your experience using Excel in class. What application areas or dataset types are best for using excel, and which are best for using R or Python? What aspects of your chosen platform do you still have questions about or need help with? What's next for you in terms of your development of your statistical knowledge?