Exam 2 Guide

IST 3420, Fall 2017 Chen

Rules:

1. The exam 2 is 50 minutes long (from 11:00 AM to 11:50 AM on November 3rd). Try to be in the classroom at least 5 minutes before the exam starts.
2. The exam is a closed book exam. Textbooks, notes, and the Internet are **NOT** allowed to be used during the exam.
3. During the exam, electronic devices of any kind will **NOT** be allowed. You need to turn off computer and mobile phone.

About the Exam

1. The exam contains two kinds of questions: (1) multiple choice (only one best answer); and (2) short answers.
2. Use course slides to guide your review of the course content and reading materials.
3. Review in-class exercises and homework assignments would be helpful for exam preparation.
4. During the exam, use your time properly. If you get stuck in one question, you need to move on and come back later.

Coverage of Exam 2

Note: The exam may cover other issues NOT mentioned on this guide.

Module	Content to Cover
#4: Cleansing and Manipulating Data	▪ Understand why data cleansing is important for a data analytics project ▪ Be able to apply useful methods in dplyr and tidyr packages to cleanse a dataset ▪ Be able to apply methods used to manipulate strings such as tolower(), toupper(), nchar(), trimws(), grep(), sub(), gsub(), substr(), strsplit(), paste(), paste0() ▪ Understand the basic syntax of regular expression (refer to the reading assignment) ▪ Be able to apply forward pipe operator to a data analytics project ▪ Understand and be able to apply the following methods (in base R, dplyr, and tidyr packages) to manipulate datasets: o Create, recode, and rename variables o Convert data structures o Sort o Subset o Merge o Aggregate o Reshape

#5: Data Summarization and Visualization	▪ Be able to choose appropriate tabular and basic graphic methods for different types of data (qualitative vs. quantitative) ▪ Understand tabular and basic graphic methods and be able to interpret these visualization results ▪ Be able to use ggplot2 to visualize data ▪ Understand spatial data structure ▪ Be able to construct advanced visualization such as spatial plots, hexagon binning, mosaic plot, heat map
#6: Data Exploration	▪ Understand methods (listwise deletion, and imputation) used to deal with missing data ▪ Be able to detect outliers in a dataset by using multiple methods such as boxplot rule, z-score, and density-based local outlier ▪ Understand the difference between covariance and correlation ▪ Be able to visualize correlation relationships (scatter plot, scatter plot matrix, correlation matrix) ▪ Understand the distinction between population and sample and the concept of hypothesis testing ▪ Understand the meaning of p-value and be able to interpret p-value correctly ▪ Be able to conduct one-sample t-test and two-sample t-test
#7: Regression Analysis	▪ Understand the function of regression analysis ▪ Be able to conduct simple and multiple linear regression analysis and correctly interpret results ▪ Understand the issue of multicollinearity for multiple regression ▪ Be able to conduct logistic regression analysis and correctly interpret results