MSDS 593 – Exploratory Data Analysis with `R`
Instructor: Paul Intrevado
Course Syllabus
Summer 2018

SUMMARY INFORMATION
Offices: Shared Office (Downtown) / McLaren Hall, Room 103 (Main Campus)
Office Hours: Wednesdays, 10.00 - 11.30 and 13.45 - 15.15 in Shared Office
Office Phone: 415/422.2527
Email: `pintrevado@usfca.edu`

Class Time: 10:00 - 12:00 / 13:00 - 15:00 Thursdays and Fridays
Final Exam: 13.00 - 15.00 Friday, August 10^{th}, 2018
Quizzes: 09:15 - 10:00 Thursdays

ON COURSE GOALS. Any student who successfully completes this course should:

- Understand the strengths and weaknesses of using `R`;
- Be able to search for, download, install and maintain packages;
- Use functions from specific packages, whether those packages are loaded or not;
- Understand the pitfalls associated with functional masking;
- Confidently create and manipulate `R` data structures;
- Understand `R` data types and coercion;
- Be able to conditionally subset all data structures;
- Use RMarkdown to create `html` and `pdf` documents for consumption by a general audience
- Confidently use control flow tchniques (`for`, `while`, `repeat`, `if`, `if else`, `ifelse`, `switch`);
- Be able to import `csv`, `txt`, `JSON`, and delimited data;
- Be able to use `R` to generate and interpret standard numerical and visual summaries of data, including the five-number summary, box-and-whiskers plots, histograms, kernel density histograms;
- Employ the `dplyr` package for advanced data manipulation;
- Use the `magrittr` for writing code using piping notation;
- Reshaping data from long to wide (and vise versa) using the `tidyr` package;
- Create advanced graphics using the `ggplot2` package;
- Understand lexical scoping and be able to write functions (including anonymous functions) in `R`;
- Be able to write robust code that includes condition handling and defensive programming techniques;
- Intelligently employ functionals (e.g., `apply()`) in lieu of loops, as well as variants of the `map()` function from the `purrr` package;
- Evaluate the performance of `R` code using the `microbenchmark` package;
- Manipulate dates using the `lubridate` package;
- Employ regular expressions in character manipulation functions using the `readr` package.

ABOUT ME. My name is Paul Intrevado. Please call me Paul. I'm an Assistant Professor of Analytics in the Department of Mathematics & Statistics, in the College of Arts & Sciences.

ABOUT US. We will meet to discuss the use of `R` for exploratory data analysis and applications from Thursday, July 12^{th}, 2018 through Thursday, August 10^{th}, 2018. We will meet at the University of San Francisco's downtown campus at 101 Howard Street.

ON COMMUNICATION. All formal course material such as the course syllabus, course notes, and data sets will be available online via github. All grades will be available through Canvas, and digital homework submissions are to be made exclusively through Canvas. All other forms of communication with the instructor will occur through Slack (either in the MSDS 593 group channel, or in private, direct messages on Slack), or via email. You are required to check Slack daily, and are responsible for any clarifications, changes and/or updates posted on the MSDS 593 group channel.

ON TEXTBOOKS. There is no formal course textbook required. This course is custom-designed for the M.S. Data Science program at the University of San Francisco, and have yet to find a singular reference that treats all of the topics we will discuss in MSDS 593. The material contained in this course is sourced from various sources, including over a dozen different textbooks, many of which are available in our MSDS library.

ON `R`. `R` is a powerful open-source scripting language and software environment for statistical computing and graphics. The `R` language is used by many professional statisticians and is making deep inroads in industry as well. `R` is equipped with a wide variety of statistical and graphical techniques. **The use of `R` for exploratory data analysis is a course objective, therefore you are not permitted to use any other scripting or programming language for this course**.

ON ATTENDANCE. Formal attendance will not be taken, nor will it be required. You are all graduate students and are expected to be mature enough to manage your time intelligently. If you miss lecture(s), you need not explain or excuse yourself to me. My objective as a course instructor is to ensure that you understand the material to be covered in this course. If you are already familiar with the material or choose to learn it on your own time, that is your prerogative.

ON QUIZZES. Quizzes will be administered every Thursday at 09.15 for **all** MSDS 593 students. There will be a total of four quizzes in this course. Quizzes will be administered in a paper/pencil format. Failure to show for the quiz will result in the forfeiture of the associated grade. Under no circumstances will make-up quizzes be administered.

ON HOMEWORK. You will be required to complete four homework assignments. You must work on homework **individually** and **submit your own, individualized deliverable(s)** (unless otherwise specified). You may consult with other students in the class regarding homework, but each student should complete all parts of the assignment successfully without assistance. Significant differences between homework scores and test scores may be subject to investigation.

ON TAs. The class has two TAs, Neerja Doshi and Shikhar Gupta. They will hold office hours in the $5th$ floor agora on Wednesdays from 10.00 - 12.00 and from 13.00 - 15.00, and on Thursdays and Fridays from 15.00 - 17.00. Yoanceu can also ping them on Slack for assistance.

Homework is graded on a discrete scale as follows:

Description	Grade
All questions attempted and completed with the proper diligence, attention to technical detail, and clarity of presentation.	100%
All questions attempted with minor or serious problems in **one** of the following areas: proper diligence, attention to technical detail, or clarity of presentation.	80%
All questions attempted with minor or serious problems in **more than one** of the following areas: proper diligence, attention to technical detail, or clarity of presentation.	60%
Not all homework questions attempted **OR** not submitted on time.	0%

Homework is due every Thursday at 09:15 *before* the quiz. To complete a homework submission you must execute the following by the submission time:

1. submit a paper copy of your homework using `RMarkdown` to the instructor in class before the quiz begins; you can generate a `pdf` or `html` and print it to paper, double-sided and stapled in the top-left corner

2. upload your `RMarkdown` source file (`lname_fname_hw_X.Rmd`) to Canvas (aforementioned naming convention **must** be used)

3. upload your knitted `RMarkdown` file (`lname_fname_hw_X.pdf`) (**only** `pdf` will be accepted, aforementioned naming convention **must** be used)

All homework is subject to the following rules:

1. All code should be commented in a neat, concise fashion, explaining the objective(s) of individual lines of code.

2. When making reference(s) to *summary* results, include all relevant output in text of the deliverable where it is being discussed, not in an appendix at the back of the deliverable.

3. Do not include a copy of the raw data in the body of the deliverable unless there is a compelling reason.

4. `R` can generate hundreds of graphs and statistical output extremely easily. Only include *relevant* graphs and output in the deliverable. All graphs and statistical output included in the deliverable should be referenced in the text of the deliverable. **There should be no orphaned figures or graphs**. Everything should be orderly and easy for a grader to read.

5. All code should be visible in the submitted paper and `pdf` versions of the homework, i.e., for each code chunk, be sure to set `echo = TRUE`

6. Homework **may not be emailed to the instructor**. All homework should be submitted in class *and* uploaded to Canvas.

7. **I will not accept late deliverables under any circumstance.**

ON THE FINAL EXAMINATION. There will a final, cumulative exam on Friday, August 11th from 13.00 to 15.00 for all students. Student must pass the final exam to pass the class.

ON GRADING. Part of my job as an instructor is to assign grades fairly and in a manner that reflects the high academic standards at the University of San Francisco and in the MSDS program. Your grade in this course will be computed according to the following weights:

Component	Weight
Quizzes	40% [4 × 10%]
Homework	20% [4 × 5%]
Final Exam	40% [1 × 40%]

Final letter grades are assigned on a relative scale. Historically, final grades are approximately distributed as follows:

Letter Grade	A's	B's	C's
Approximate % of Students	15%	65%	20%

This year's grade distribution will necessarily be different. I do not have a predefined letter grade distribution or quota of grades or grade distribution to which I must adhere. The A grade range indicates distinguished performance and competence; the B grade range indicates strong to adequate performance and competence; the C grade range demonstrates weak understanding of the material. A grade of C- is the minimal passing grade. A grade of F is given for performance that insufficiently demonstrates academic competence.

ON LAPTOPS. Bring a laptop to lecture and have `R` installed on it. You will be expected to use `R` in a lecture setting for in-class examples and labs. I expect you to use your laptops judiciously, refraining from surfing the web or engaging in any other distracting behavior during lecture.

ON CHEATING. As a Jesuit institution committed to *cura personalis*—the care and education of the whole person—the University of San Francisco has an obligation to embody and foster the values of honesty and integrity. The university upholds standards of honesty and integrity from all members of the academic community, including faculty, students, and staff. All students are expected to know and to adhere to the university's honor code. You can find the full text of the code online at `http://www.usfca.edu/catalog/policies/honor/`. You are also bound by the terms of the MSAN Code of Conduct that you signed prior to matriculating in the analytics program. Refer to ON HOMEWORK section for details regarding student collaboration on each category of deliverable. Plagiarism consists of copying *any* material from *any* source and submitting it as your own original work, regardless of where that material was sourced: the Internet, a book, textbook, or from deliverables perviously submitted by other students. All students involved in any cheating or plagiarized deliverables, i.e., the cheater as well as the person(s) who willfully enabled or facilitated the act of cheating, will be reported to the MSAN Program Director and the Dean of the College of Arts and Sciences. If you ever have questions about what constitutes plagiarism, cheating, or academic dishonesty in this course, I am happy to discuss with you at your convenience.

ON DISABILITIES. If you are a student with a disability or disabling condition, or if you think you may have a disability, please contact USF Student Disability Services (SDS) at 415/422.2613 within the first week of class, or immediately upon onset of the disability, to speak with a disability specialist. If you are determined eligible for reasonable accommodations, please meet with your disability specialist so they can arrange to have your accommodation letter sent to me, and we will discuss your needs for this course. For more information, please visit `http://www.usfca.edu/sds/` or call 415/422.2613.