MSAN 593 – Exploratory Data Analysis
Instructor: Paul Intrevado
Course Syllabus
Summer 2016

SUMMARY INFORMATION
Office: SFH 525 (101 Howard, 5^{th} Floor)
Office Hours: open door policy, by appointment, videoconference or Slack
Mobile Phone: 765/418.6874
Office Phone: 415/422.2527
Email: pintrevado@usfca.edu

Class Location: 101 Howard Street, SFH 156 (Th) / SFH 527 (Fri)
Class Time: 10:00 - 11.50 / 13.00 - 14.50 Thursdays and Fridays

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 data types and coercion;
- Be able to subset and conditionally subset all data structures;
- Use `R`Markdown to create `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 standard numerical and visual summaries of data, including the five-number summary, box-and-whiskers plots, histograms, kernel density histograms;
- Understand how to correctly and effectively use the above-mentioned standard numerical and visual summaries of data;
- Employ the `dplyr` package for advanced data manipulation;
- Use the `magrittr` for writing code using piping notation;
- 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;
- Evaluate the performance of `R` code using the `microbenchmark` package;
- Employ regular expressions in character manipulation functions.

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. I am also the Director of the Practicum Program for MSAN as well as the Associate Director of the Data Institute. Please feel free to knock on my office door or call/text/Slack me at your convenience. I encourage you to video conference with me over Zoom. I am also happy to schedule an appointment with you.

ABOUT US. We will meet to discuss `R`, and the use of `R` for exploratory data analysis and applications from Thursday, July 14^{th}, 2016 through Friday, August 12^{th}, 2016. We will meet at the University of San Francisco's downtown campus at 101 Howard Street in SFH 126 on Thursdays and SFH 527 on Fridays.

ON TAs. There are two TAs assigned to assist this class with grading and providing students with guidance. Griffin Okamoto is a 2015 MSAN alum and Kimberly Siegler is a 2016 MSAN alum. Both can be reached via Slack on the main MSAN 593 group channel, as well as in private with direct messages in Slack. TAs will not respond to email.

ON COMMUNICATION. All formal course material such as the course syllabus, course notes, and data sets will be posted on Canvas. All grades are also posted exclusively on Canvas. All other forms of communication with the instructor will occur through Slack, either in the MSAN 593 group channel, or in private direct messages. You are required to check Slack daily and are responsible for any clarifications, changes and/or updates posted on the MSAN 593 group channel. Emails are discouraged.

ON TEXTBOOKS. There is no formal course textbook required. This course is custom-designed for the M.Sc. Analytics program at the University of San Francisco, and I have yet to find a singular reference that treats all of the topics we will discuss in MSAN 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 MSAN 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), do 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 Friday at 10.00 for **all** MSAN 593 students.

- If you are registered for the MSAN 593 10.00 section, you will take the quiz at 10.00 in SFH 527 with me (regular class location).
- If you are registered for the MSAN 593 13.00 section, you will take the quiz at 10.00 in SFH 529 with Professor Uminsky.

Quizzes will be 30 minutes in length and 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 five homework assignments. They will be posted on the Canvas course webpage. 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.

Homework is graded on a continuous scale. You are required to submit an RMarkdown file (`*.Rmd`) as well as a compiled `pdf` of your RMarkdown file for each homework assignment. Your RMarkdown file will be run by TAs on their local machines. If the code fails to run for any reason, you will loose 30% of your grade for that deliverable. You will subsequently have until the end of the module, i.e., August 12^{th} at 17.00 to submit a corrected version. If you fail to submit a corrected version, you will receive a grade of zero for that deliverable. If you submit a corrected version, you will be rescaled to the remaining 70%; **you will not automatically get the full 70% simply for submitting RMarkdown file that compiles**.

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, and should also be indexed, e.g., Figure 1 or Table 2. **There should be no orphaned figures or graphs**. Everything should be orderly and easy for the grader to read.
5. Homework may not be emailed to the instructor or the TAs. All homework should be uploaded to Canvas.
6. **I will not accept late deliverables under any circumstance.**

ON THE FINAL EXAMINATION. There will be no final examination in this course.

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 MSAN program. Grades will be assigned according to the following scale:

Letter Grade	A	A-	B+	B	B-	C+	C	C-
GPA Equivalence	4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7
Cutoff Grade	85	80	75	70	65	60	55	50

Your grade in this course will be computed according to the following weights:

Component	**Weight**
Quizzes	25% [1 × 1% + 4 × 6%]
Homework	75% [5 × 15%]

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 and ON CASE STUDIES sections 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. 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.

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.