Introduction to Five College DataFest

Presented by Undergraduate Researchers Interested in Data

2/25/18

DataFest Background

- Five College DataFest is a nationally coordinated undergraduate competition hosted by the ASA.
- Groups of students are given one weekend to explore a large, complex dataset as they see fit, then present their findings to a panel of judges at the end of the weekend.
- **Start: Friday, March 23rd, 7:30PM**
- **End: Sunday, March 25th, 4PM**
- Location: Integrated Science Building (ISB)

Why should you participate?

- Great chance to meet people with similar interests from all around the Five College area.
- Develop the skills learned in your classes.
  - The best way to learn something is to surround yourself with people more knowledgeable/experienced than you!
- Gain experience using a real-world dataset to answer impactful business questions.
- It looks great on your resume/CV, whether you walk away with a medal or not.
- Free t-shirt, as well as catered snacks/meals all weekend.
- **Overall: just do it!**

Lots of participants!

Great prizes!

What to expect

- Large, “real world” dataset
  - Size - hundreds of thousands of rows, 50-100 columns
  - May include missing values
- Using the data to solve a business problem
  - e.g. “which groups of customers are not being advertised to well enough? What can we do about it to generate more conversions?”
  - Previous sponsors include Ticketmaster and Expedia.
- Weekend-long time commitment
  - Do homework beforehand, plan ahead to study for upcoming exams, etc.
  - You want to have the whole weekend to focus on your analysis.

General Tips: What to focus on in analysis

- Good - concise, actionable business insight.
  - e.g. "physical positioning of the teacher closer to the students increases student attention."
- Bad - lots of technical jargon with no clear benefit.
  - e.g. "we fit a generalized linear mixed model with a Poisson distribution using a log link and obtained findings that minimizing mean Euclidean distance between teacher and student increases student attentiveness by 5 points on a predefined 20-point scale."
- KISS - keep it simple, stupid!
  - Avoid unnecessary complexity if simpler, more easily interpretable methods are available.
- Remember, the main role of a data scientist to to turn *raw data* into useful **information**!

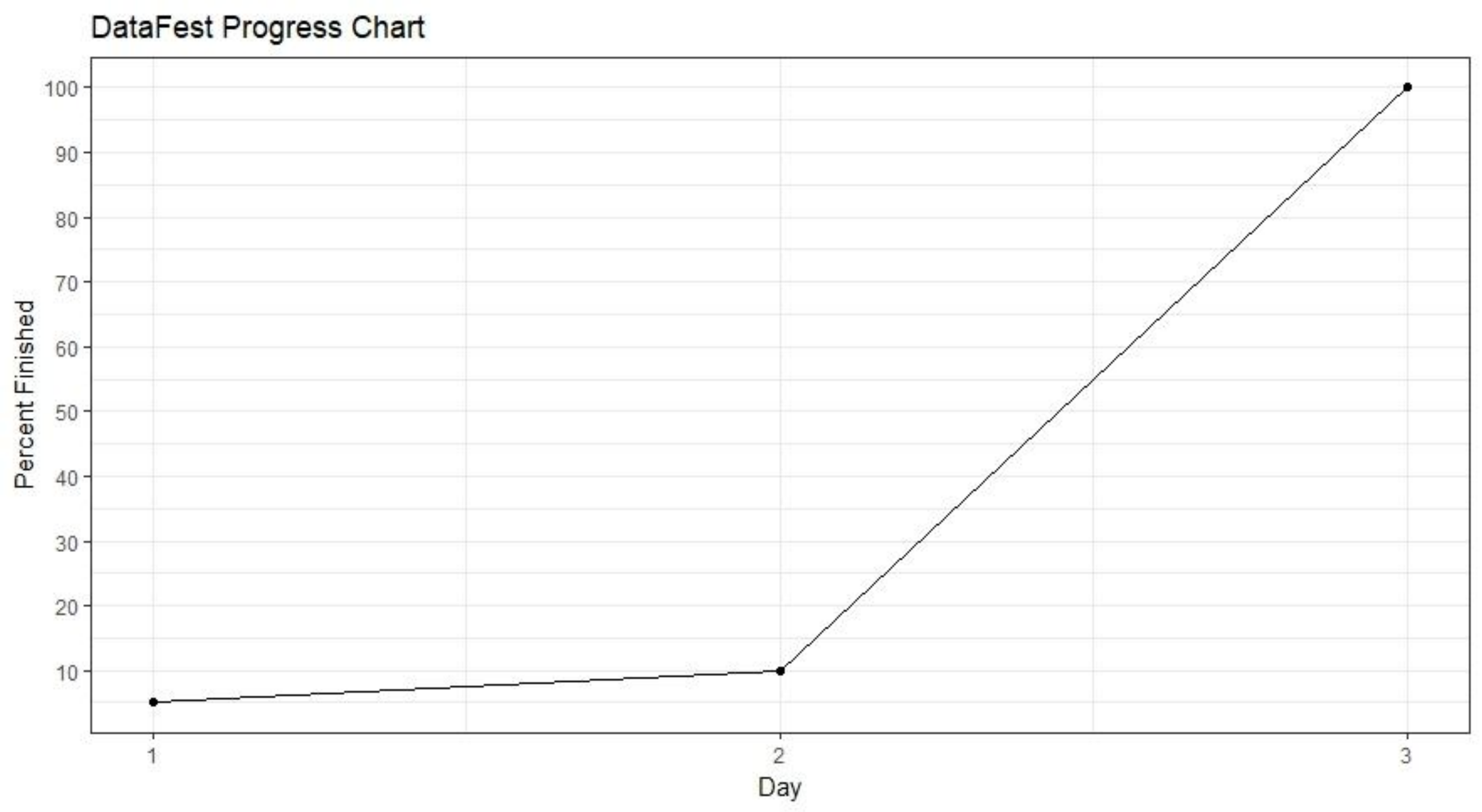
Examples courtesy of Alexander Bogdan

General Tips: Making a good impression

- Don’t underestimate the importance of your write-up and presentation.
  - This is where you get to succinctly and stylishly show off your combined efforts over the past weekend of work.
- **Main goal: convince the judges why your analysis would be useful to the business while highlighting the reasoning behind any assumptions made or techniques used.**
- Only focus on the most important aspects or examples from your analysis.
  - 5 minute/1 page time limit - consider introduction, methods, and results structure for both.
- Practice general public speaking skills
  - Every group member should talk as equally as possible.

General Tips: Effectively using your time

- First Day
  - Familiarize yourself with dataset - identify response variable, hypothesize relationships between covariates, performing initial EDA, and identify useful subsets of the data for later use.
- Second Day
  - Start performing more focused analysis once a solid direction has been identified. Don’t worry if this takes you a long time, or you try multiple leads before you find something promising - that’s normal!
  - Saturday is the most stressful day, so prepare yourself physically and mentally before coming.
- Third Day
  - Decide on a final, cohesive theme for your analysis and begin putting together a slideshow an write-up that reflect this. Must submit by noon for presentations and judging!


DataFest Progress Chart
Percent Finished
100
90
80
70
60
50
40
30
20
10
1
2
3
Day

Miscellaneous Tips

- Pace yourself.
    - You have all weekend, so don’t overwork yourself too early before the presentations.
- Come early to get a good spot.
    - Any conference room that has its own table (and sometimes even projector!) is very desirable.
- Check in frequently with other groups/experts for inspiration or feedback.
- Feel free to skip the workshops if you are confident in your knowledge.
    - They can be really helpful otherwise, though!
- Sample from dataset and test your plots/models on that first.
    - This will save you many headaches when your IDE/computer crashes from the full dataset.

Things to practice before

- General coding skills - language doesn’t matter, but R and Python are good.
  - Focus on data manipulation, then visualization, then modeling in that order.
- Working with “big data”.
  - More rows/columns than you are probably used to - that’s fine!
- Collaborative skills when working with a team to perform analysis.
- Organizing and presenting your findings to a live audience.

Introducing the Data

- To prepare for DataFest, we will be looking at a larger dataset together to get practice on the style of data we will be analyzing during the event.
- Head on over to Kaggle and download the housing prices dataset (register and make an account first if you haven't already).
- This dataset contains roughly 80 explanatory variables, but only about 1,500 rows. Expect many, many more observations at DataFest!
- Before getting started on the data, it helps to think about...
  - Hypothesized relationships between covariates (especially with the response variable)
  - Potential uses for analysis (e.g. how could this lead to actionable business insight?)
  - Ways to create new variables based on existing ones (feature engineering)
- Check out this post on our site for more tips on where to get started with a large dataset like this one.

Introducing the Data pt. 2

- Please feel free to explore this dataset as you see fit, or work on honing your statistical programming skills to prepare for DataFest.
  - Let us know if you want help learning R or Python!
- Potential goals of analysis:
  - Accurately predict housing prices.
  - Create visualizations or tables that show what features are common in certain types of houses (i.e. cheap, average, or expensive.)
  - Come up with suggestion for pricing houses based on analysis.
- To help with exploratory data analysis (EDA), we suggest trying the new DataExplorer package in R to easily facilitate high-level overviews of data missingness, variable distribution, and correlations.
- The stargazer R package is also useful for presenting clean, well-formatted looking tables of statistical results in your presentations.

Happy Analyzing!