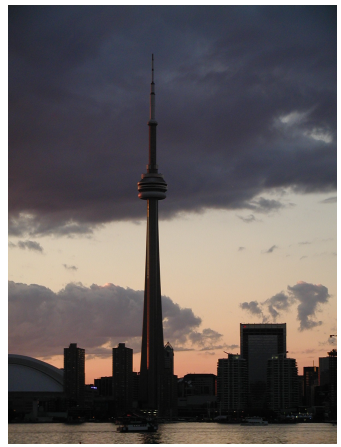
Lessons Learned from Six Years of Data Technologies Course Projects at Utah State University

Jürgen Symanzik*
*Department of Mathematics and Statistics
Utah State University
Logan, Utah, USA

e-mail: juergen.symanzik@usu.edu



https://www.math.usu.edu/~symanzik

https://www.math.usu.edu/~symanzik/talks/2023_IASE.pdf

July 11, 2023

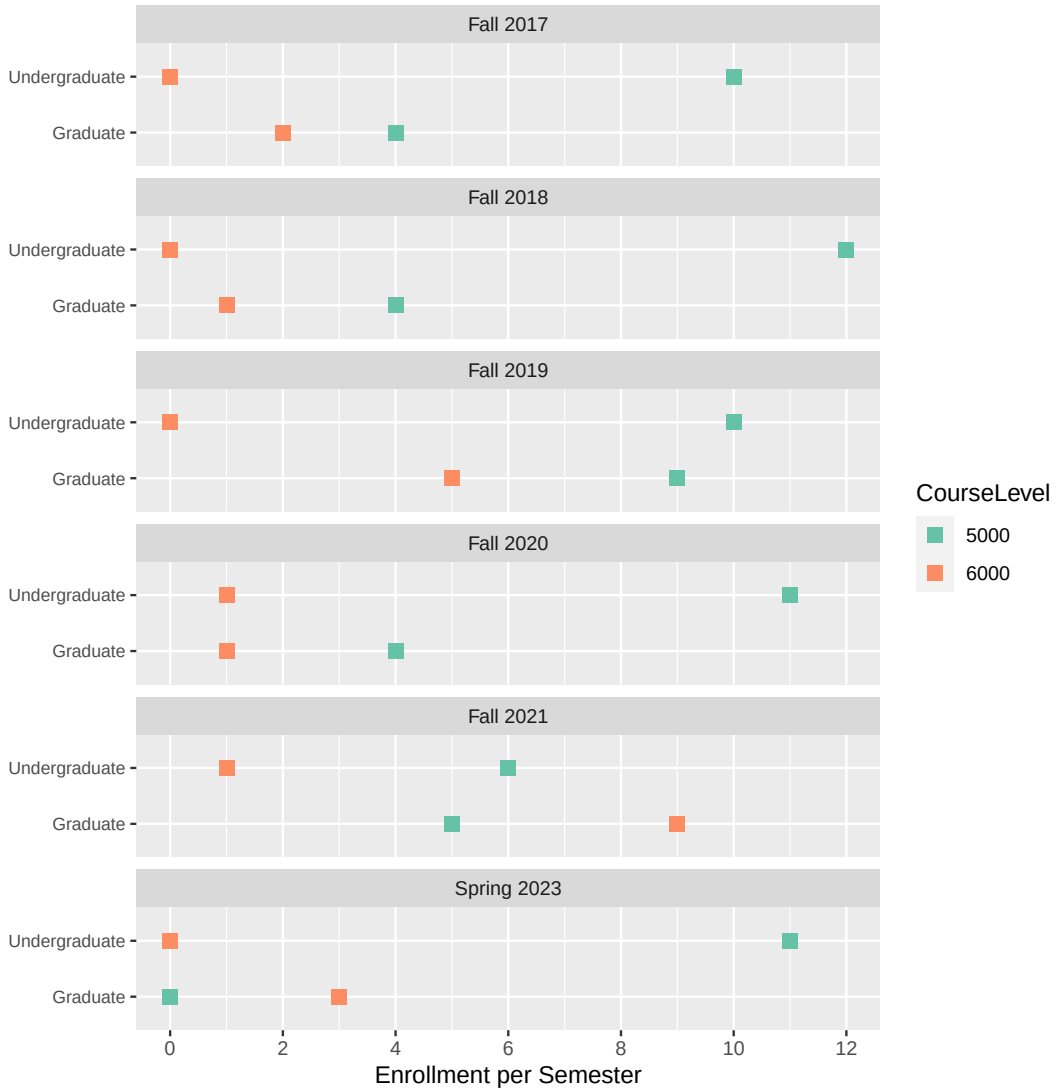
Outline

Course Format

- Data Technologies (DT) is a 2–credit (10–week) course with twenty 75 min lectures.
- Cross–listed for undergraduate[1] and graduate[2] students.
- Main difference is an additional 6000 Level course project (worth 30% of the final grade).
- Offered annually since 2017, usually in the Fall semester.
- Course syllabi accessible at `https://www.math.usu.edu/~symanzik/teaching/JSteaching.html`.

[1] 5000 Level: Stat 5810 [2017–2018] and 5080 [since 2019]
[2] 6000 Level: Stat 6910 [2017–2018] and 6080 [since 2019]

Course Enrollment



Total Enrollment: 14–24 Students.

Course Topics

1. Data.
2. Basics of simulation.
3. Representation of information.
4. Regular expressions.
5. HyperText Markup Language (HTML) and Web–scraping.
6. Text data from Portable Document Format (pdf) files and via Optical Character Recognition (OCR).
7. The *Tidyverse* (R packages for Data Science).
8. Data bases and Structured Query Language (SQL).
9. Extensible Markup Language (XML).
10. Resampling/bootstrap/others (outlook only).

Software

- Primarily R via RStudio (but students can use Python or others for the project).
- 6000 Level course requires LaTeX for all homeworks, the project report, etc. and the LaTeX Beamer document class for the project presentation.
- R code and templates for the project report and project presentation are provided in Canvas (a web–based learning management system).

Textbooks

- No textbook required.
- Slides and R code provided via Canvas.
- Main inspiration from:
  - Murrell, P. (2009) Introduction to Data Technologies, Boca Raton, FL: Chapman and Hall/CRC, https://www.stat.auckland.ac.nz/~paul/ItDT/.
  - Nolan, D. & Temple Lang, D. (2015) Data Science in R — A Case Studies Approach to Computational Reasoning and Problem Solving, Boca Raton, FL: CRC Press/Taylor & Francis, https://web.archive.org/web/20230124015039/https://rdatasciencecases.org/.
- Credits to Dr. Paul Murrell (University of Auckland), Dr. Duncan Temple Lang (UC Davis), Dr. Deborah Nolan (UC Berkeley), and Dr. Hadley Wickham (RStudio) for initially providing some of their course and training materials.

Project Motivation

- Chance article: Rundel, C. & Cetinkaya-Rundel, M. (2016) Taking a Chance in the Classroom: La Quinta is Spanish for "Next to Denny's", *Chance* 29(2): 53–57, https://doi.org/10.1080/09332480.2016.1181966.
- Article that describes a team–based homework assignment for undergraduate and graduate students in statistical computing courses at Duke University.
- Assignment makes use of web–scraping, visualization, statistical analyses, and introduces geospatial data.
- Tasks can be adapted to students with different backgrounds.

Project Stages and Timeline

- By Lecture 11: Preliminary Discussion of Project Proposal[3] (or Discussion of Project).
- By Lecture 14 (prior to 2–week break): Written Project Proposal[3].
- In Lectures 15, 17 & 19: Weekly Progress Reports.
- In Lecture 20 (or on Backup Lecture Date): Final Project Presentation.
- About 2 weeks after Final Project Presentation: Written Project Report.
- About 1 day after Written Project Report: Contributions to the Project.

[3] Topic proposed by student(s).

Overview of Project Topics 2017–2023

- Fall 2017: Findings from Released John F. Kennedy Documents [2 Students].
- Fall 2018: PDF Scraping of USDA National Honey Report Data [1][4].
- Fall 2019: Marvel Data Project: Web–scraping, Data Organization, and Interactive Interface [5].
- Fall 2020: Billboard Hot 100 Chart: Data from over 60 Years [2].
- Fall 2021: (i) Commissary Weekly Ads Analysis: 2018-2021 [4]; (ii) Data Collection and Analysis of Rancho Market Ads During the Pandemic [3]; (iii) Costco Grocery Ad Analysis: Pricing and Availability Pre & Post COVID-19 [3].
- Spring 2023: NFL Touchdown Passes Group Project [3][4].

[4] Topic proposed by student(s).

Fall 2017: Kennedy Documents Project Outline

- On July 24, 2017, the U.S. National Archives & Records Administration (NARA) started to release previously withheld John F. Kennedy assassination documents at `https://www.archives.gov/research/jfk/release`.
- By November 9, 2017, more than 31,000 documents were released in five batches.
- Main tasks were to obtain summary statistics about the documents and some idea about their content, e.g., agencies that authored a document (CIA, FBI, others), original dates of documents, type of document (text on paper, note, photo, audio, etc.), and sender/recipient information.
- Main DT components: Web–scraping, text extraction, processing, and OCR — visualization.

Fall 2017: Kennedy Documents Project Results

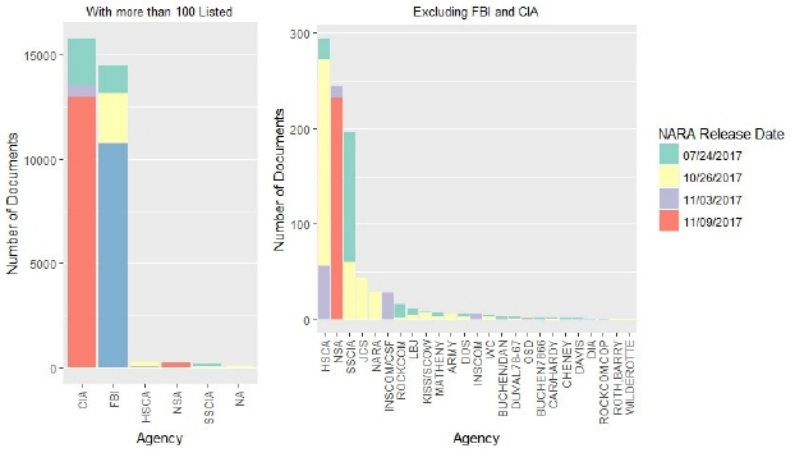


Figure 1: Number of documents released from each agency in descending order. Due to the extreme differences in document count, the left plot shows only the first five agencies (and NA), while the right plot shows the last 25 agencies. Note the change in vertical axis scales.

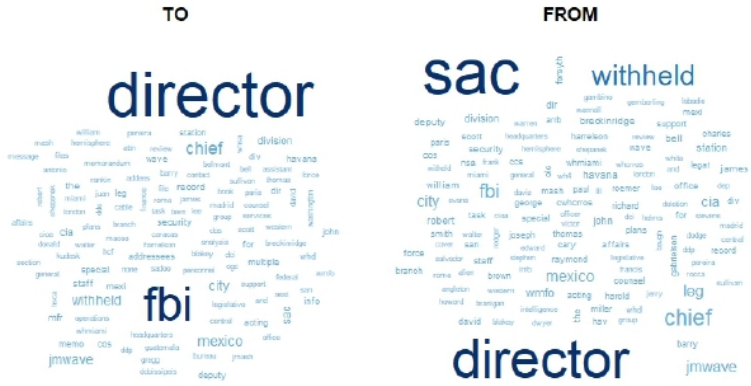


Figure 4: Two words clouds showing (left cloud) the relative frequency of who a document was directed to, and (right cloud) from.

Note: SAC stands for *Strategic Air Command*.

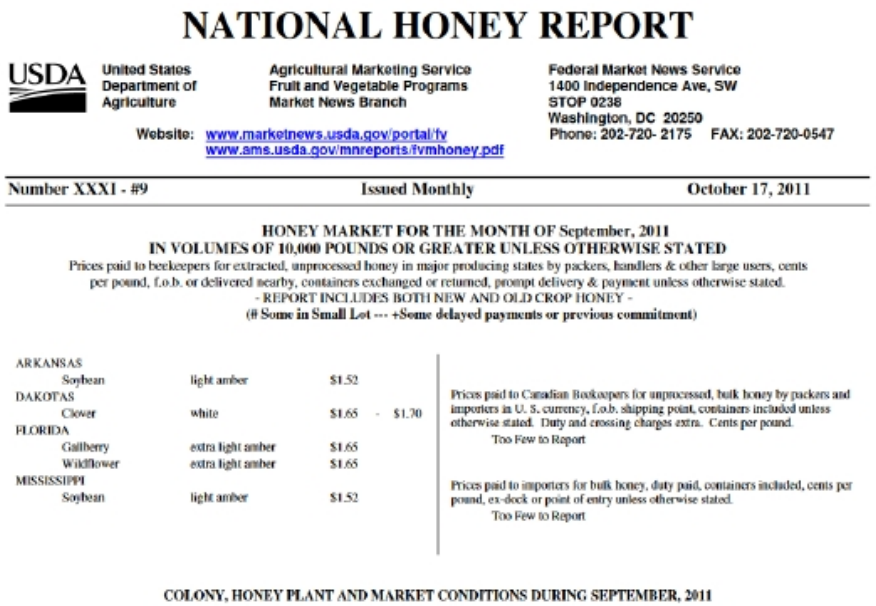
Fall 2017: Kennedy Documents Project Successes & Challenges

- First DT project successfully completed!
- NARA web page rather unexpectedly released four more batches of documents during the course of the project, often with a main focus on a single agency.
- R code had to be fully reproducible to work with the updated/extended NARA web page.

Fall 2018: USDA Honey Data Project Outline

- Starting in 2011, the United States Department of Agriculture (USDA) started to release monthly national honey reports in pdf format such as `https://www.ams.usda.gov/mnreports/fvmhoney.pdf`.
- Main tasks were to web crawl several months of reports and extract the honey production state, plant, honey type, and price range for a certain month from each pdf.
- Main DT components: Web–scraping, regular expressions, text extraction and processing.

Fall 2018: USDA Honey Data Project Results



NATIONAL HONEY REPORT

USDA United States Department of Agriculture

Agricultural Marketing Service
Fruit and Vegetable Programs
Market News Branch

Federal Market News Service
1400 Independence Ave, SW
STOP 0238
Washington, DC 20250
Phone: 202-720- 2175 FAX: 202-720-0547

Website: www.marketnews.usda.gov/portal/fv
www.ams.usda.gov/mnreports/fvmhoney.pdf

Number XXXI - #9 | Issued Monthly | October 17, 2011

HONEY MARKET FOR THE MONTH OF September, 2011
IN VOLUMES OF 10,000 POUNDS OR GREATER UNLESS OTHERWISE STATED

Prices paid to beekeepers for extracted, unprocessed honey in major producing states by packers, handlers & other large users, cents per pound, f.o.b. or delivered nearby, containers exchanged or returned, prompt delivery & payment unless otherwise stated.
- REPORT INCLUDES BOTH NEW AND OLD CROP HONEY -
(# Some in Small Lot --- +Some delayed payments or previous commitment)

ARKANSAS			
Soybean	light amber	$1.52	
DAKOTAS			
Clover	white	$1.65	$1.70
FLORIDA			
Gallberry	extra light amber	$1.65	
Wildflower	extra light amber	$1.65	
MISSISSIPPI			
Soybean	light amber	$1.52	

Prices paid to Canadian Beekeepers for unprocessed, bulk honey by packers and importers in U. S. currency, f.o.b. shipping point, containers included unless otherwise stated. Duty and crossing charges extra. Cents per pound.
Too Few to Report

Prices paid to importers for bulk honey, duty paid, containers included, cents per pound, ex-dock or point of entry unless otherwise stated.
Too Few to Report

COLONY, HONEY PLANT AND MARKET CONDITIONS DURING SEPTEMBER, 2011

Figure 1: Relevant portion of a report. Available at `https://search.ams.usda.gov/mndms/2011/10/FV20111017MHONEY.pdf`

	A	B	C	D	E	F	G
1	month	year	location	honeyPlant	honeyType	minPrice	maxPrice
2	9	2011	ARKANSAS	Soybean	light amber	1.52	1.52
3	9	2011	DAKOTAS	Clover	white	1.65	1.7
4	9	2011	FLORIDA	Gallberry	extra light amber	1.65	1.65
5	9	2011	FLORIDA	Wildflower	extra light amber	1.65	1.65
6	9	2011	MISSISSIPPI	Soybean	light amber	1.52	1.52
7							

Figure 2: Automatically generated Excel output from scraping the report shown in Figure 1

Fall 2018: USDA Honey Data Project Successes & Challenges

- Only one student: limited scope of the project.
- Format of pdf reports changed over time.
- No clear delimiters that identified the start/end of a table with data.
- One-column and two-column tables in pdf with no apparent formatting (as one typically finds in an HTML document).

Fall 2019: Marvel Cinematic Universe Project Outline

- Relevant information about the Marvel movies and the characters can be found in a specific wiki format at `https://marvelcinematicuniverse.fandom.com/wiki/Marvel_Cinematic_Universe_Wiki`.
- Main tasks were to web–crawl these pages (starting with the local sitemap) and extract all character names, their gender, their status (hero, neutral, villain), and the number of times each character is mentioned.
- Main DT components: Web–scraping, regular expressions, text extraction and processing, cloud storage via Box and boxr R package, SQL for local queries.

Fall 2019: Marvel Cinematic Universe Project Results

	Characters	Gender	Status	expressions	counts
1	a.i.m. president	M	Villain	[^-(__)[:alpha:]]a\\.i\\.m\\.(_| |.)?president[^-[:alpha:]]	264
2	abilisk	Unknown	Villain	[^-[:alpha:]]abilisk[^-[:alpha:]]	9467
3	abomination	M	Villain	[^-[:alpha:]]abomination[^-[:alpha:]]	62820
4	absorbing man	M	Hero-Villain	[^-(__)[:alpha:]]absorbing(_| |.)?man[^-[:alpha:]]	8880
5	abu bakaar	M	Villain	[^-(__)[:alpha:]]abu(_| |.)?bakaar[^-[:alpha:]]	1838
6	adolf hitler	M	Villain	[^-(__)[:alpha:]]adolf(_| |.)?hitler[^-[:alpha:]]	10473

Figure 2: Sample of the Final Table Including Population Counts and Character Regular Expressions

```
> question("Who is the most popular character?")
[1] "iron man"
> question("Who is the most popular hero?")
[1] "iron man"
> question("Who is the most popular villain?")
[1] "ultron"
> question("Who are the 15 most popular characters?")
 [1] "iron man"        "captain america" "thor"            "spider man"      "hulk"
 [6] "ultron"          "ant man"         "thanos"          "black panther"   "loki"
[11] "daredevil"       "black widow"     "captain marvel"  "hawkeye"         "doctor strange"
> question("What is the hero to villain ratio")
[1] "The ratio Hero to villain is 0.278"
> question("What is the female to male ratio")
[1] "The ratio Male to Female is 3.029"
> question("Who is the most popular female character?")
[1] "black widow"
> question("Who is the most popular female villain?")
[1] "witch"
> question("How many times does daredevil appear?")
[1] 753616
> question("How many times do daredevil and kingpin appear together?")
[1] 753616  12484
> question("How many times does batman appear?")
integer(0)
> question("Who are the 10 least popular female heros?")
 [1] "kate bishop"    "brigid o reilly" "gert yorkes"     "molly hernandez" "nico minoru"
 [6] "karolina dean"  "maria rambeau"   "crystal"         "medusa"          "ayo"
> question("How many times does groot appear?")
[1] 351147
> question("How many times does iron man appear?")
[1] 5052032
```

Figure 3: Output from User Interface

Fall 2019: Marvel Cinematic Universe Project Successes & Challenges

- Five students: most complex project overall.
- Large amount of data: 21,000 main web pages [in 2019] that extended to nearly 100,000 web pages (many with images or videos).
- Approximately 25 GB of local storage required.
- Fun part: During the project presentation, other students could propose some queries.

Fall 2020: Billboard Charts Project Outline

- Billboard Hot 100 Single Charts have been released since 1958 and can be accessed at `https://www.billboard.com/charts/hot-100/`.
- Main tasks were to web–crawl these pages (starting in 1958) and extract relevant information about the artist, song title, rank, etc. for all listed songs each week.
- Main DT components: Web–scraping, regular expressions, text extraction and processing.

Fall 2020: Billboard Charts Project Results

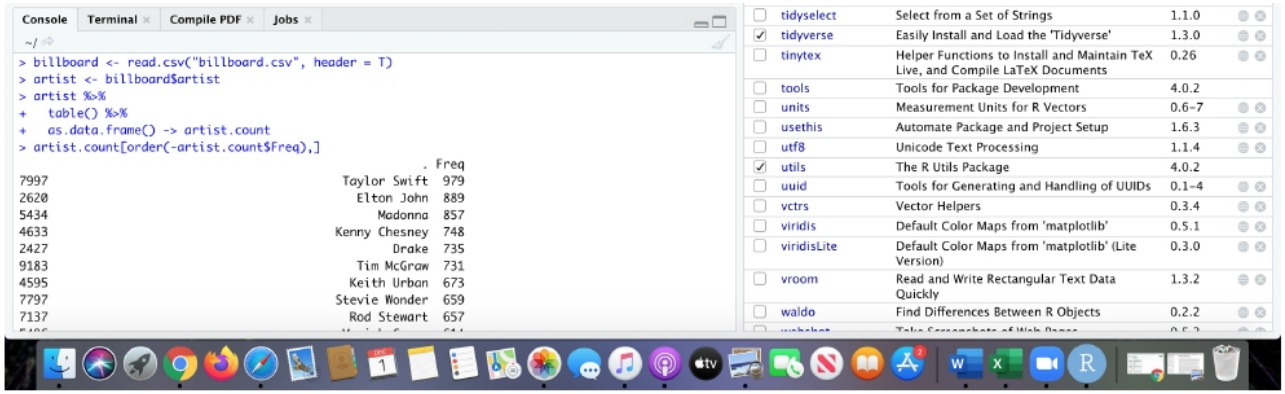


Figure 5: The most featured artist on the charts.

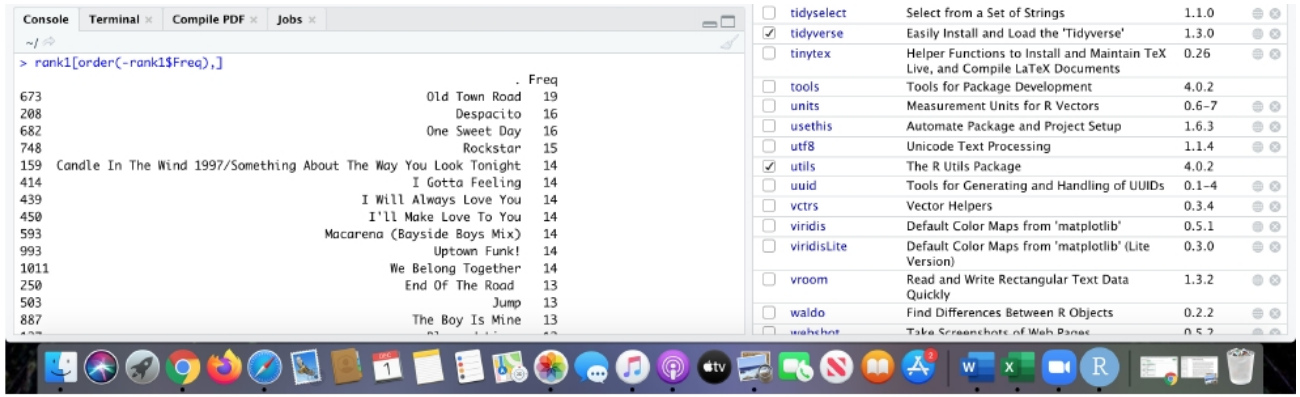


Figure 7: The song that spent the longest at the number one position.

Fall 2020: Billboard Charts Project Successes & Challenges

- Held during the Covid–19 pandemic: Zoom lectures and only remote interaction.
- Only two students: similar to 2019 project, but much less complex.
- About 60 years $\times$ 50 weeks of data (about 3,000 web pages overall).
- Fun part: During the project presentation, other students could propose some questions, e.g., which song was number one in the week of their birthday.

Fall 2021: Weekly Online Ads Project Outline

- Many grocery stores publish weekly (or monthly) online ads, e.g., Commissary (`https://weekly-ads.us/commissary-ads`), Rancho Markets (`https://weekly-ads.us/archive/rancho-markets`), and Costco (`https://weekly-ads.us/archive/costco`).
- Main tasks were to extract weekly ads for a particular store, identify products and prices, and do some statistical analyses of pre–Covid and post–Covid products, sales prices, and discounts that appear in the ads.
- Main DT components: Web–scraping, regular expressions, text extraction and processing, and OCR — statistical analyses and visualization.

Fall 2021: Weekly Online Ads Project Results

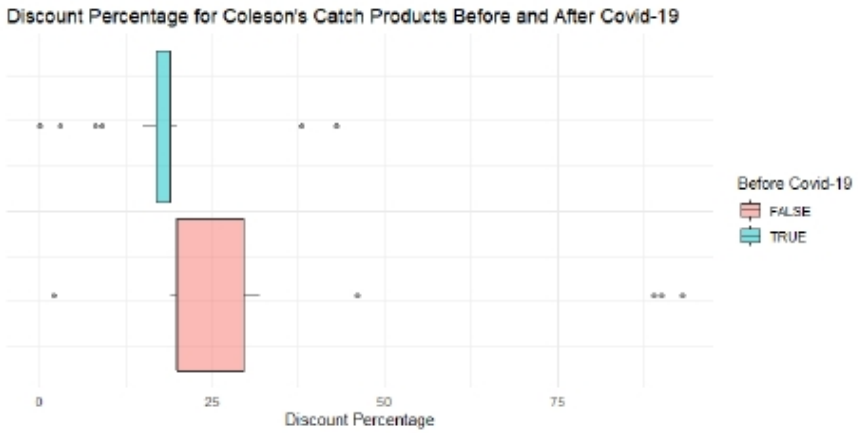


Figure 2: Discounts for Coleson's Catch in Relation to COVID-19

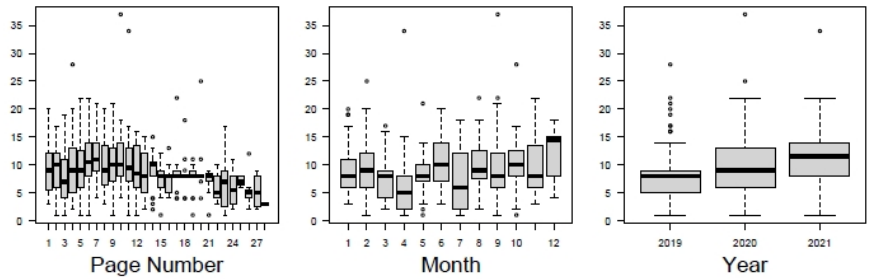


Figure 3: The distributions of item counts by page, month, and year. The first few pages typically have much higher variation in the number of items listed, but also tend to have a higher number of listed items. Most ads had at least 14 pages, but some had as many as 28. Looking at the months, April tended to have the least amount of advertised items. Their appears to be an upward trend by year from 2019 - 2021.

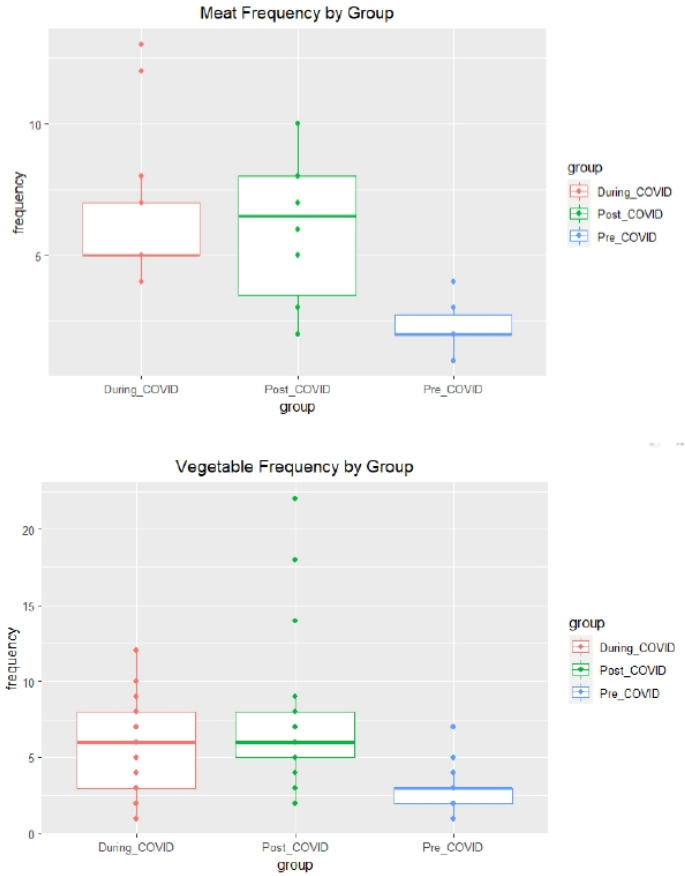


Figure 3: Distribution of the frequency of meat and vegetable items pre, during, and post COVID-19.

Fall 2021: Weekly Online Ads Project Successes & Challenges

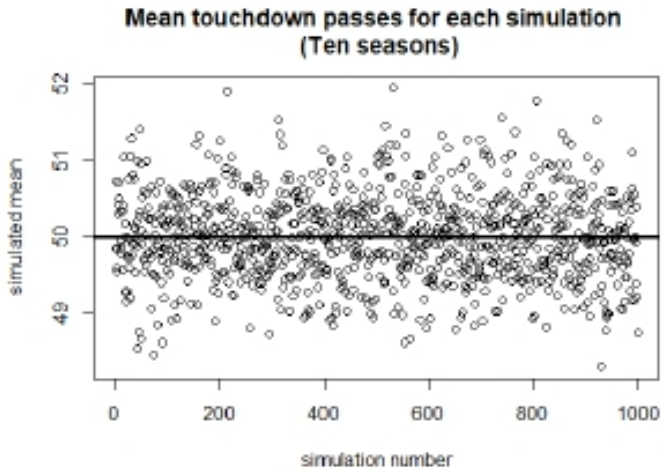
- First full in–person semester after the Covid–19 pandemic.
- Ten students, resulting in three similar projects with different data sources.
- Often, only the current ad was accessible in pdf format — older ones were archived in jpg format which required an additional OCR step.
- Cloudflare website security prevented systematic web–scraping so that only a limited number of ads could be downloaded manually (various Python packages did not overcome the problem).

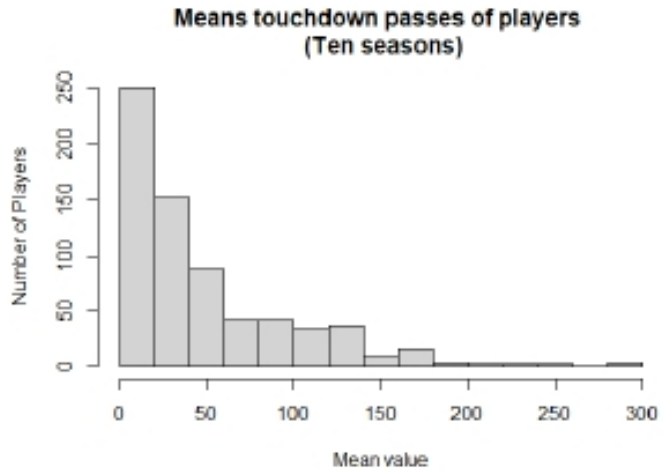
Spring 2023: NFL Touchdown Passes Project Outline

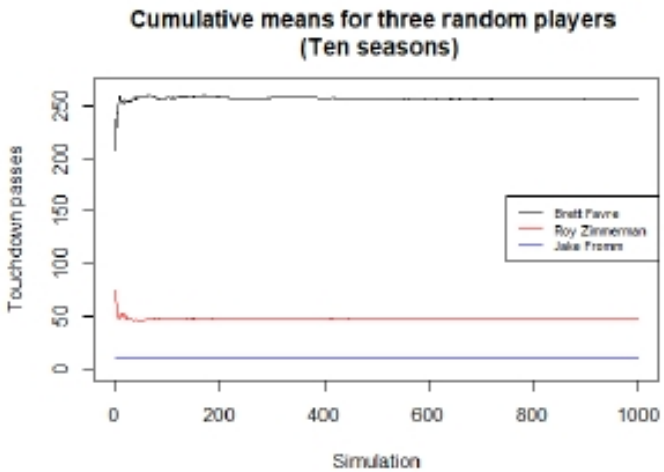
- Pro Football Reference provides access to statistics, history, scores, standings, playoffs, schedules and records since the 1920s at `https://www.pro-football-reference.com/`.
- Main tasks were to web crawl these pages (starting with the player sitemap) and extract all player names, their positions, and game statistics. Identify all quarterbacks, conduct what–if simulations of touchdown passes (e.g., no injuries, retired after most successful season, etc.) of the top–12 players and compare with other quarterbacks.
- Main DT components: Web–scraping, regular expressions, text extraction and processing, and simulation — visualization.

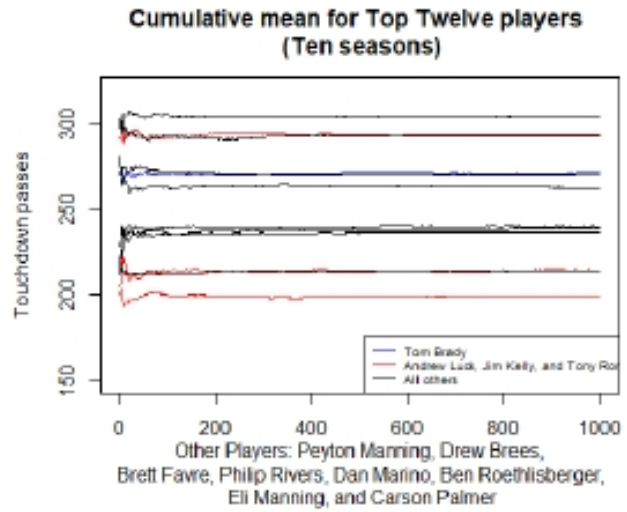
Spring 2023: NFL Touchdown Passes Project Results

B Visualizations









Spring 2023: NFL Touchdown Passes Project Successes & Challenges

- First project with a major simulation component.
- One student dropped out of graduate school during the semester.
- All students only had limited experience with visualization prior to the course.

Main Grading Categories (Out of 100 Points)

- Preliminary Discussion of Project Proposal (2 / — Points)[5].
- Written Project Proposal (7 / — Points).
- Progress Reports (9 / 9 Points).
- In-Class Presentation (40 / 40 Points).
- Written Project Report (40 / 50 Points).
- Contributions & Breakdown (2 / 1 Points).

[5] First number if topic proposed by student(s), second number if topic proposed by instructor

Grading Categories for the In–Class Presentation (Out of 40 Points)

- How good was the presentation overall? (10 Points)
- How good/meaningful/complete were the slides? (10 Points)
- What about the timing (i.e., close to 20min)? (4 Points)
- How useful/relevant were the presented graphs? (8 Points)
- How well were the results presented? (8 Points)

Grading Categories for the Written Project Report (Out of 40 / 50 Points)

- Length ($\leq$ 6 pages). (2 / 3 Points)[6]
- Title & Authors. (2 / 3 Points)
- Abstract & Keywords. (3 / 4 Points)
- Introduction. (5 / 6 Points)
- Main Text. (8 / 10 Points)
- Figures, Tables & Captions. (7 / 8 Points)
- Conclusion & Future Work. (7 / 8 Points)
- References (incl. R & R Packages). (3 / 4 Points)
- Appendix. (3 / 4 Points)

[6] First number if topic proposed by student(s), second number if topic proposed by instructor

Personal Reflections

- Six years of course project (Fall 2017 – Spring 2023): Challenging, but quite successful from my perspective.
- Ongoing learning experience on my side.
- Necessary to adapt to unexpected and new scenarios (Covid–19 and 3 groups instead of just 1 group).
- Permanent fine–tuning of project descriptions and grading criteria.

Recommendations for Instructors

- Let students find/propose a topic that is of interest to them, or suggest a topic that likely is of interest to many of the students.
- Do some initial exploration/screening of a possible topic, but do not expect to identify all possible problems and pitfalls. Consider this to be similar to a faculty research project.
- Adjust scope of the project and workload to the number and skill level of the students working in the group(s).
- Be flexible to the needs and requests of the students and do not shy away from dropping something that was initially proposed. If necessary (and possible), extend deadlines for project presentations and final reports if students ask for it.
- Provide students with as much information as possible how to work on such a project, e.g., clear list of deliverables and a timeline, templates for the presentation and final report, and detailed grading criteria.

Student Feedback

- Limited feedback in official course evaluations:
  - Fall 2019: *"The project was nice to learn to work with others and nice to complete something that would have been hard to do alone."*
  - Spring 2023: *"It was great as an undergraduate to see the projects of the graduate students and see their progress in their projects and see what we are doing in class and see them applied to real projects of the grad students."*
- Some students indicated in conversions after the course how valuable the course project has been for them with respect to their current job.

Possible Topics for Future Projects (1)

- Data from popular movie/TV web pages, e.g., Star Wars (`https://starwars.fandom.com/wiki/Main_Page`), Star Trek (`https://memory-alpha.fandom.com/wiki/Portal:Main`), or IMDb (`https://www.imdb.com/`).
- Data from music sites, e.g., Discogs (`https://www.discogs.com/`).
- Data from social media sites, e.g., TikTok (`https://www.tiktok.com/en/`).
- Data from sports pages, e.g., soccer/football (`https://data.world/datasets/soccer`), tennis (`https://data.world/datasets/tennis`), or basketball (`https://www.kaggle.com/datasets/wyattowalsh/basketball`).

Possible Topics for Future Projects (2)

- Data from literature data bases (`https://www.gale.com/databases/literature`).
- Data from popular magazines or newspapers, e.g., Time (`https://time.com/vault/`) or New York Times (`https://archive.nytimes.com/www.nytimes.com/ref/membercenter/nytarchive.html`).
- Location data of stores, recreation places, public buildings, churches, sports arenas, etc.
- "What–if" simulations, based on web–scraped data related to climate change, world population growth, elections, etc.
- **Question: Any suggestions from the audience?!?**

Acknowledgment of Participating Students

- Fall 2017: Lacy Christensen & Eric McKinney.
- Fall 2018: Jared Hansen.
- Fall 2019: Joanna Coltrin, Adelyn Fleming, Mina Hossain, Johnny Medri & Colby Wight.
- Fall 2020: Sydney Geisler & Tristan Peterson.
- Fall 2021: (i) Tyler Antoloci, Ragan Astle, Nate Nellis & McKade Thomas; (ii) Scout Jarman, Tyler Clayson & Kinspride Duah; (iii) Jake Rhodes, Matthew Lister & Kristen Sohm.
- Spring 2023: Brian Nalley, Gideon Perry & Tyson Ashcraft.

Original Project Descriptions and Grading Criteria

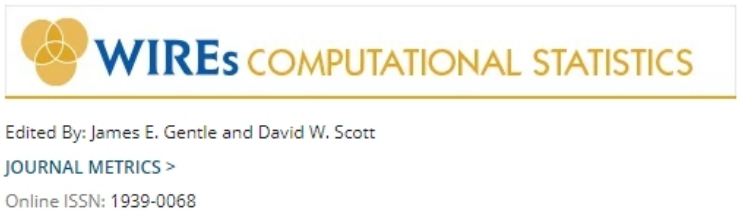
- Fall 2017: https://www.math.usu.edu/~symanzik/teaching/2017_stat5810_003/DTProject_2017.pdf
- Fall 2018: https://www.math.usu.edu/~symanzik/teaching/2018_stat5810_004_fall/DTProject_2018.pdf
- Fall 2019: https://www.math.usu.edu/~symanzik/teaching/2019_stat5080/DTProject_2019.pdf
- Fall 2020: https://www.math.usu.edu/~symanzik/teaching/2020_stat5080/DTProject_2020.pdf
- Fall 2021: https://www.math.usu.edu/~symanzik/teaching/2021_stat5080/DTProject_2021.pdf & https://www.math.usu.edu/~symanzik/teaching/2021_stat5080/Student_Project_Evaluation_Fa2021.tex
- Spring 2023: https://www.math.usu.edu/~symanzik/teaching/2023_stat5080/DTProject_Sp2023.pdf & https://www.math.usu.edu/~symanzik/teaching/2023_stat5080/Student_Project_Evaluation_Sp2023.tex

Further Reading

- Fall 2019: Project–like group homework assignment, required for all undergraduate and graduate students (24 overall).
- Students from the Stat 6080 course project served as heads of five groups.
- Further reading: Fleming, A., Coltrin, J. D., Medri, J., Hilyard, C., Tellez, R., Symanzik, J. (2022) Results and Student Perspectives on a Web–Scraping Assignment from Utah State University's Data Technologies Course to Evaluate the African Activity in the Statistical Computing Community, *Computational Statistics*, `https://doi.org/10.1007/s00180-022-01222-7`.
- Section 4 of that article discussed student observations and feedback for that particular group homework.

Advertisement

- I recently joined as a Commissioning Editor of Wiley's *WIREs Computational Statistics* review journal (`https://wires.onlinelibrary.wiley.com/hub/journal/19390068/about/editorialboard`).
- Reach out to me if you have an idea for a possible review article with a focus on statistical/data science education in the context of computational statistics, data science, and visualization.



- This presentation:

  `https://www.math.usu.edu/~symanzik/talks/2023_IASE.pdf`

- **Questions ?!? —**

- or e-mail: juergen.symanzik@usu.edu