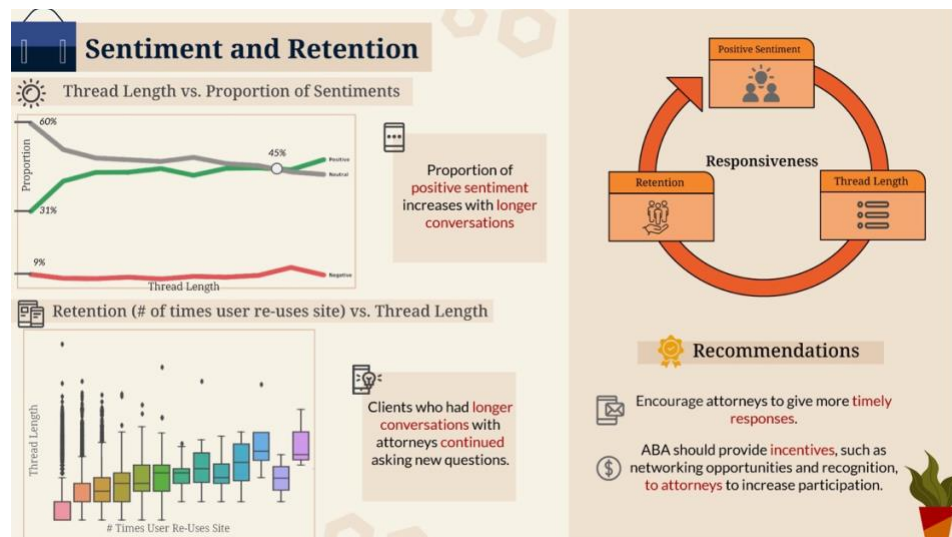
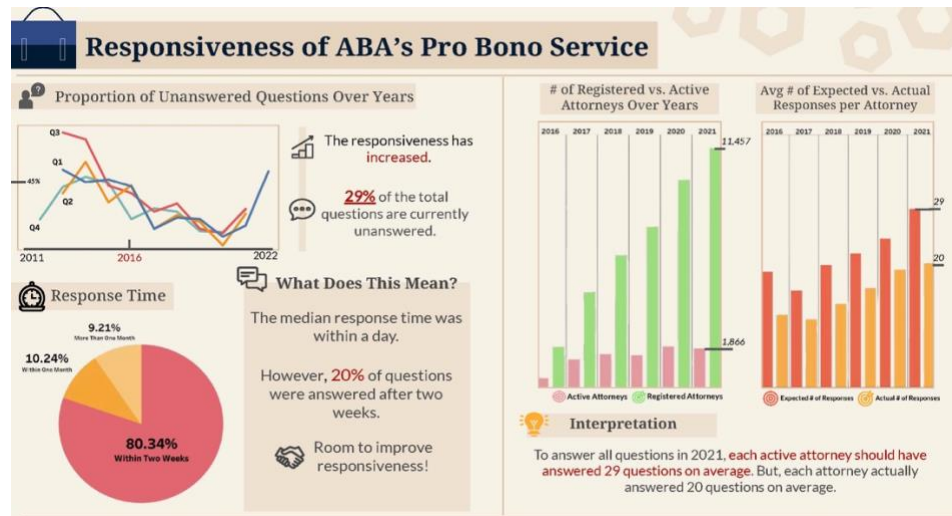


Table of Contents

UCLA DATAFEST HACKATHON SUBMISSION WINNER.....	2
REAL-TIME STOCK + TWITTER SENTIMENT TRACKER	4
POWER BI STORIES BASED ON NYC REAL ESTATE AND TAX CLAIMS	5
NETWORK ANALYSIS OF INDIVIDUAL CAREERS AND TRAINING GRANTS	6
AWARD SPREADSHEET LOOKUP + NETWORK	10
AWARD BI DASHBOARD	13
PUBLISHING IES DATA ONLINE (UNSTRUCTURED DATA CLEANING)	16
FIELD APP	17
BI ENGINEER SPECIAL PROJECTS.....	19
SEARCH, REVIEW, AND EXPLORE RESTAURANTS (WEBSITE)	20
LINKEDIN JOB FINDER.....	21
PORTFOLIO WEBSITE.....	22
BBC NEWS NLP REPORT	23
SQL CHALLENGE (LA CLIPPERS FINAL INTERVIEW)	25
ESPORTS EDA ASSESSMENT	26
SOCIALS DASHBOARD (GOOGLE DATA STUDIO).....	28
RECEIPT OCR WEBSITE.....	29
LOAN PREDICTOR APP	30
NFT MARKET ANALYSIS	31
STOCK ANALYZER	32
FOOD IMAGE RECOGNITION.....	33

UCLA DataFest Hackathon Submission Winner



We analyzed the responsiveness of ABA's Pro Bono Service that provides free legal counseling, as well as strategies for client retention through sentiment analysis.

A 64 year old man living in Natrona, Wyoming posted a question in November 2021 that didn't receive an answer until January. He then demanded why it took so long to get an answer. An attorney responded by saying they are taking time out of their day to answer. But the man said, "So are the poor people asking the questions. An answer more than a month later is of no help at all." This man was an active user who posted on 4 separate occasions before, but stopped posting completely after this incident. So, this demonstrates how responsiveness is a crucial factor for client retention.

The first plot (point at the plot) visualizes the proportion of unanswered questions over the years. Note that the data before 2016 is before the full launch of the service and 2022 only has one month of data. The different color lines represent quarters 1 to 4 just in case seasonality is a factor, but there seems to be a minimal effect. Overall, the decrease in the rate of unanswered questions indicates an increase in responsiveness. However, 29% of the total questions are still left

unanswered. The number of questions asked increased while the number of active attorneys did not, explaining this spike in 2021.

Let's look at the pie chart on the bottom left. The pie chart shows the distribution of response time. About 80% of the questions were responded within two business weeks, and the median response time was within one **business** day. However, for 20% of the questions, it took longer than 2 weeks to get a response, showing that there is room for improvement in response time.

Shifting our attention to the bar charts, we see that the first bar chart compares the number of registered attorneys to those who are *actually* answering questions for each year. In 2021, only 16% of around 11 thousand registered attorneys were noted as active.

We also calculated the average number of questions that should have been addressed by each active attorney to answer all of the questions asked in that year and looked at how many were actually answered per attorney on average. The second bar chart on the right shows this result and that the number of active attorneys should increase to reduce the workload of currently active ones.

This can help increase client satisfaction because high responsiveness will lead to longer, more active, and helpful conversations. Next, we analyzed the sentiment of clients and their retention rate with regards to the length of conversation thread.

When looking at the importance of sentiment and retention in attorney services, we noticed that clients' feedback gets more positive as they continue talking with their attorney. As shown on the top left, positive comments increased from 31% to over 45%, while negative comments stayed the same.

Now, taking a look at the number of times users returned to the site to create new question threads, we saw that these retained users tend to have longer conversations and ask more consecutive questions.

Therefore, positive sentiments, longer threads, and retention rates are all correlated with each other. This leads to a cyclical pattern that we can observe in the top right.

The cycle begins with a positive experience that leads to happy clients continuing the conversation and increasing thread length that leads to higher retention, where clients are more likely to re-use the site in the future.

The key variable that binds these 3 together is Responsiveness.

When there are more available attorneys responding more quickly, this cycle accelerates because the process is facilitated by response rate.

No response. No happy clients.

So overall, our data finds that the timely responsiveness of a lawyer is key to a client's experience. Slow or no response is limiting this cycle of user retention.

The recommendations our team came up with are to encourage more attorneys to respond in a more timely manner. The ABA committee can incentivize this by providing professional resources or running events such as award ceremonies that reward active attorneys with networking opportunities and recognition for their volunteer work.

This way future clients of the ABA Pro Bono Service won't receive the untimely, unhelpful, poor experience of our 64-year-old man, who waited 2 months from November to January for a response.

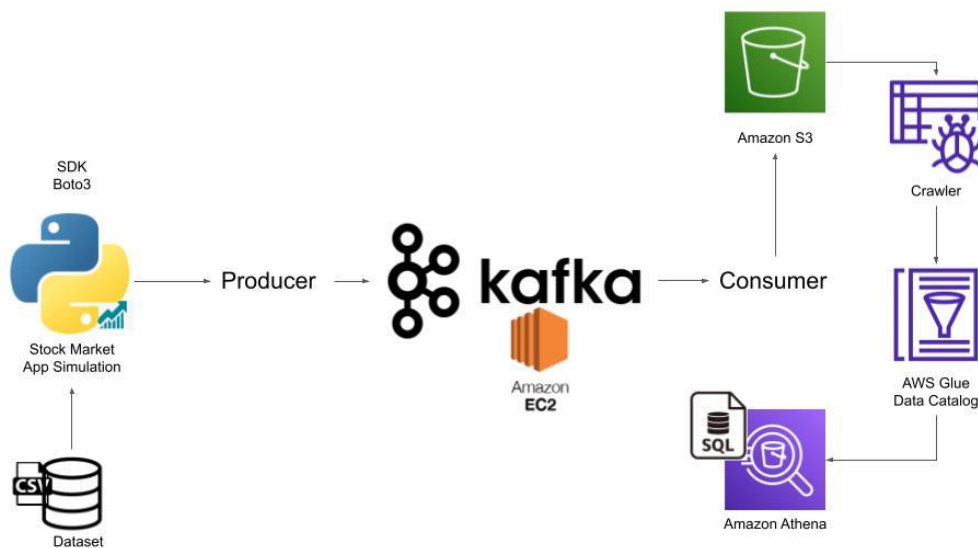
Real-time Stock + Twitter Sentiment Tracker

I developed a real-time stock analysis platform using various big data technologies for fun.

The app tracked real-time stock changes, cross-referenced with sentiment analysis of related tweets

- Apache Kafka facilitated real-time communication and processing the continuous flow of data
- Apache Spark enabled rapid analysis of the real-time numbers and text
- Cassandra connected the database with robust handling of big data across multiple servers, ensuring high availability.
- AWS for reliable and scalable cloud computing services
 - Amazon S3 for data storage and EC2's server

The project exemplified how these technologies can synergize to solve complex real-world problems. Given more time, I would probably include more data sources like Reddit or Facebook.



Power BI stories based on NYC real estate and tax claims

Home Price Disparity in NYC

NYC Housing

\$21,505,551,898
Estimated Total Value

4,722
Total # of Properties

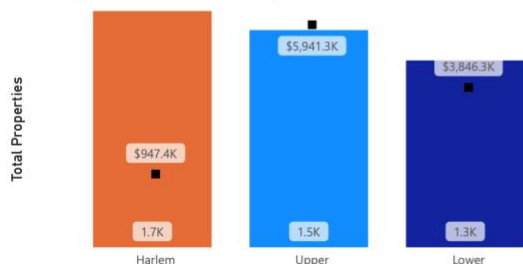
\$3,800,000
Median Sale Price

4.73%
Average Sale Turnover

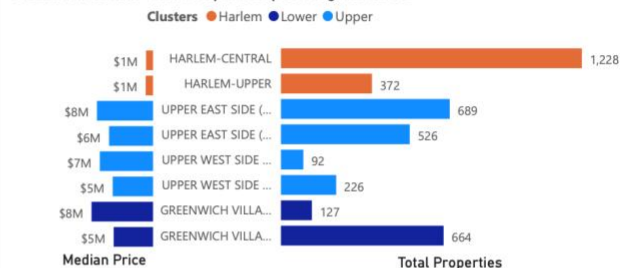
Estimated Total Value and Price Range per Cluster



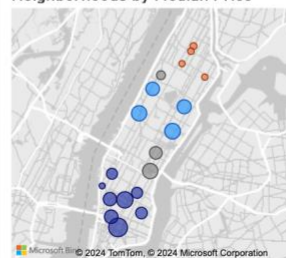
Total Properties and Median Price per Cluster



Median Price and Total Properties per Neighborhood



Neighborhoods by Median Price



Neighborhoods by # of Properties



How NYC EITC claims have changed since 2004

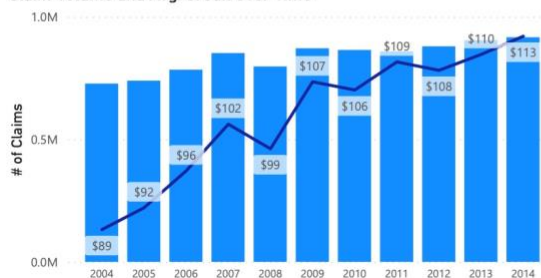
NYC EITC

\$951,336,090
Total Amount Claimed

9,199,688
Total # of Claims

\$103
Average Credit

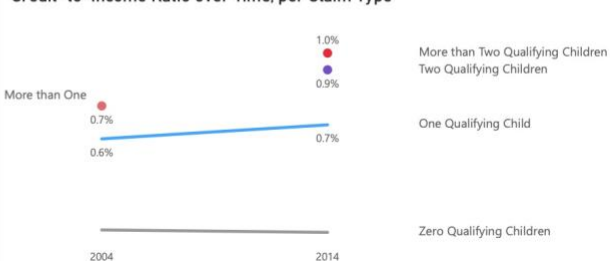
Claim Volume and Avg. Credit over Time



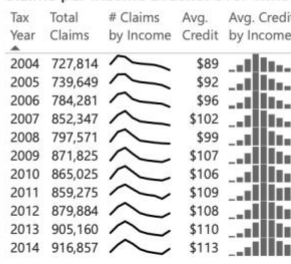
Claim Volume and Avg. Credit per Income Bracket



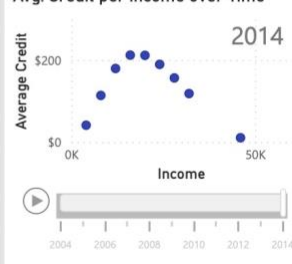
Credit-to-Income Ratio over Time, per Claim Type



Claims per Income Bracket over Time



Avg. Credit per Income over Time



Network Analysis of Individual Careers and Training Grants

“In terms of the network analysis, I wish I had a simple spreadsheet that I could share with you! In the absence of that, I wonder if there is a way to scrape/pull data from our abstracts to tell a story about how IES has built a community of scholars in the education sciences. Interestingly enough, when you search by an individual’s last name, it doesn’t show up in the search results when the person’s name is included in the tables included in the predoctoral table... So I’m not sure how much you can do with the data that we have, but at the very least, I’d love to be able to know all the instances of pre- and post-doctoral fellows who have continued to participate in IES projects as key personnel, co-PIs, or PIs and to understand the sequencing and timing of those fellows and which programs they were initially trained in.” --Task

Jumping off creating the network visualization for the effect of training grants and path of individuals after those grants end, I wanted to look into the retention rate of people who continue predoc/postdoc programs as PIs and co-PIs.

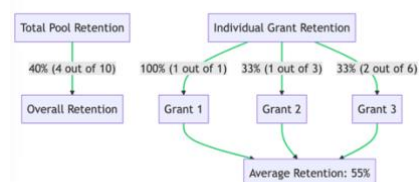
To do this, I segmented the problem into 3 steps:

1. Calculating individual properties
2. Aggregating per grant
3. Aggregating per PI and university affiliation

1. For individual properties, I defined retention, returnee, and projects impacted.
 - a. **Retention**: if individual has ever acted as PI or co-PI (led or co-led a project)
 - b. **Returnee**: if individual trained in both predoc and postdoc programs
 - c. **ProjectsImpacted**: # of projects the individual has been PI or co-PI for
2. For calculating retention, I chose to group for each grant THEN group by each PI/Affiliation because I wanted to equalize each grant. If chose to skip grouping by grants, then each individual would be weighed against each other. But I thought that weighing different grants against each other would make more sense because they tell the story on a higher level.

For instance, if I have 3 training grants consisting of 1, 3, and 6 predocs, respectively, and of those predocs, 1, 1, and 2 continue as PIs and are considered “retained.”

If we were to calculate the overall retention from a total pool perspective, we would get a 40% retention rate (4 out of 10). However, if we calculate the retention for each grant individually, we get retention rates of 100%, 33%, and 33% respectively. The average of these percentages gives us a 55% retention rate. This approach highlights the remarkable success of the first grant despite its smaller cohort size.



This approach provides a balanced and fair comparison between different grants. So smaller but successful grants are not overshadowed by larger ones, emphasizing the broader retention rates per grant. This calculation affects *CohortRetention* and *AvgImpact_cohort*.

3. For the PI-level and university level, I calculated cohort sizes, total impact, average impact, and cohort retention.

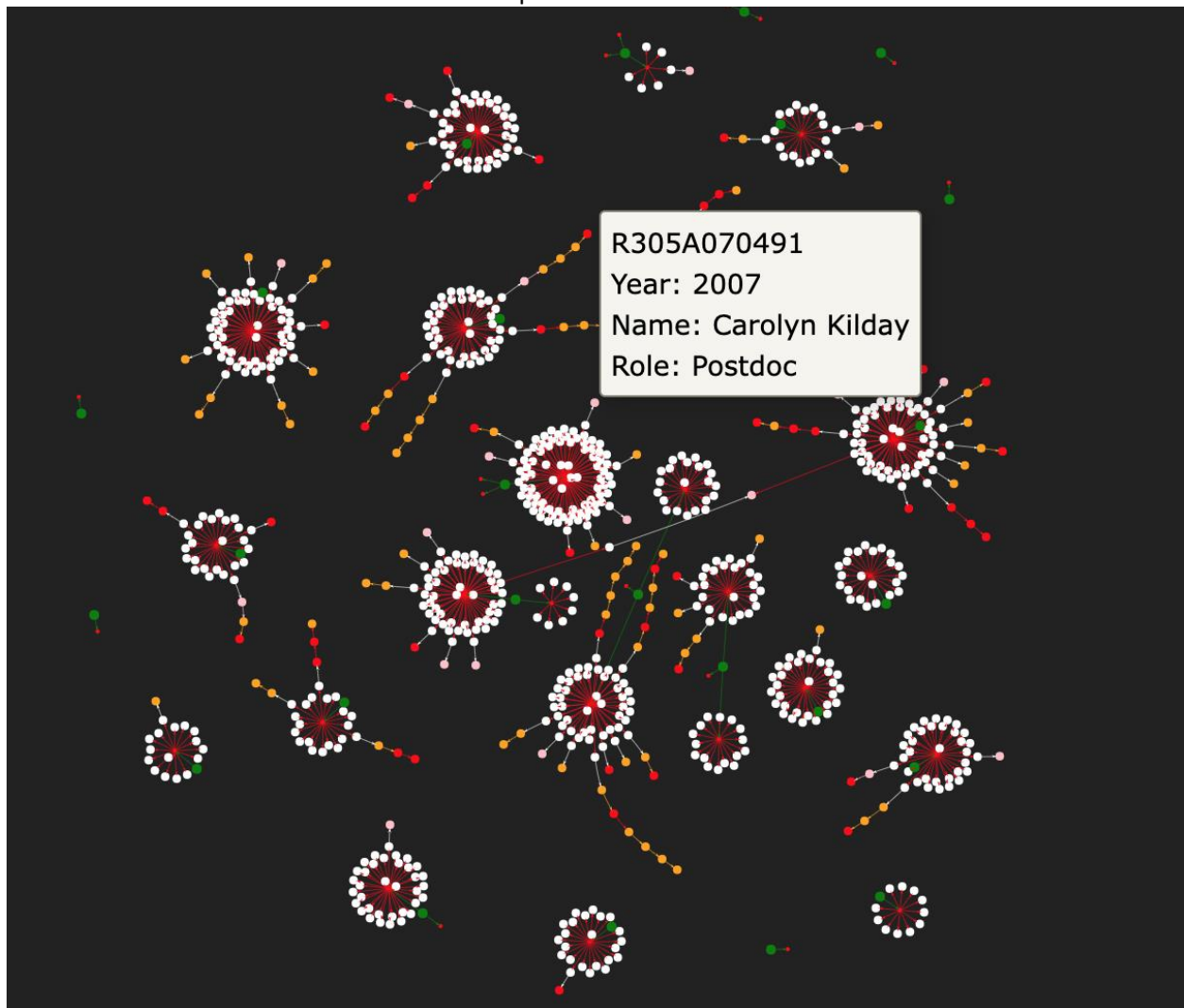
- a. **Cohort/Alum Size:** sum participants in the training grants
- b. **Total Impact:** sum effect of the individuals' work.
- c. **Average Impact:** mean effect of the work of individuals*
- d. **Retention:** mean retention rate between grants*

**where cohort mean is weighted per grant as explained above*

	Definition
Link	Link to IES website for the grant
ID	IES Database ID
AwardNum	Unique Award Number
Year	Award creation date
Names	Name of individual
Type	Position / Role
Title	Award title
PrincipalName	Award Principal Investigator
PrincipalAffiliationName	Award Principal Investigator's Affiliation
CenterName	Award's Center Name
FundTypeDesc	Grant or Contract
ProgramName	Award Program
GoalText	Award Goal
AwardAmt	Award Funding (in dollars)
AwardPer	Award Length
Period	Award Length (numeric)
ProjectsImpacted	# of projects the individual has been PI or co-PI for
Retained?	if individual ever acted as PI or co-PI
Returnee?	if individual was both predoc and postdoc
Predoc_cohort	# of predocs trained by this PI or grant (depending)
Postdoc_cohort	# of postdocs trained by this PI or grant (depending)
TotalImpact_cohort	sum of ProjectsImpacted for individuals in cohort
AvgImpact_cohort	mean of ProjectsImpacted for individuals in cohort*
PredocCohortRetention	% of predocs retained in cohort*
PostdocCohortRetention	% of postdocs retained in cohort*
PredocAlum	# of predocs trained with the university
PostdocAlum	# of postdocs trained with the university
Alum	# of individuals trained with the university
TotalImpact_affiliation	sum of ProjectsImpacted for individuals with the university
AvgImpact_affiliation	mean of ProjectsImpacted for individuals with the university
PredocAffiliationRetention	% of predocs retained with the university
PostdocAffiliationRetention	% of postdocs retained with the university

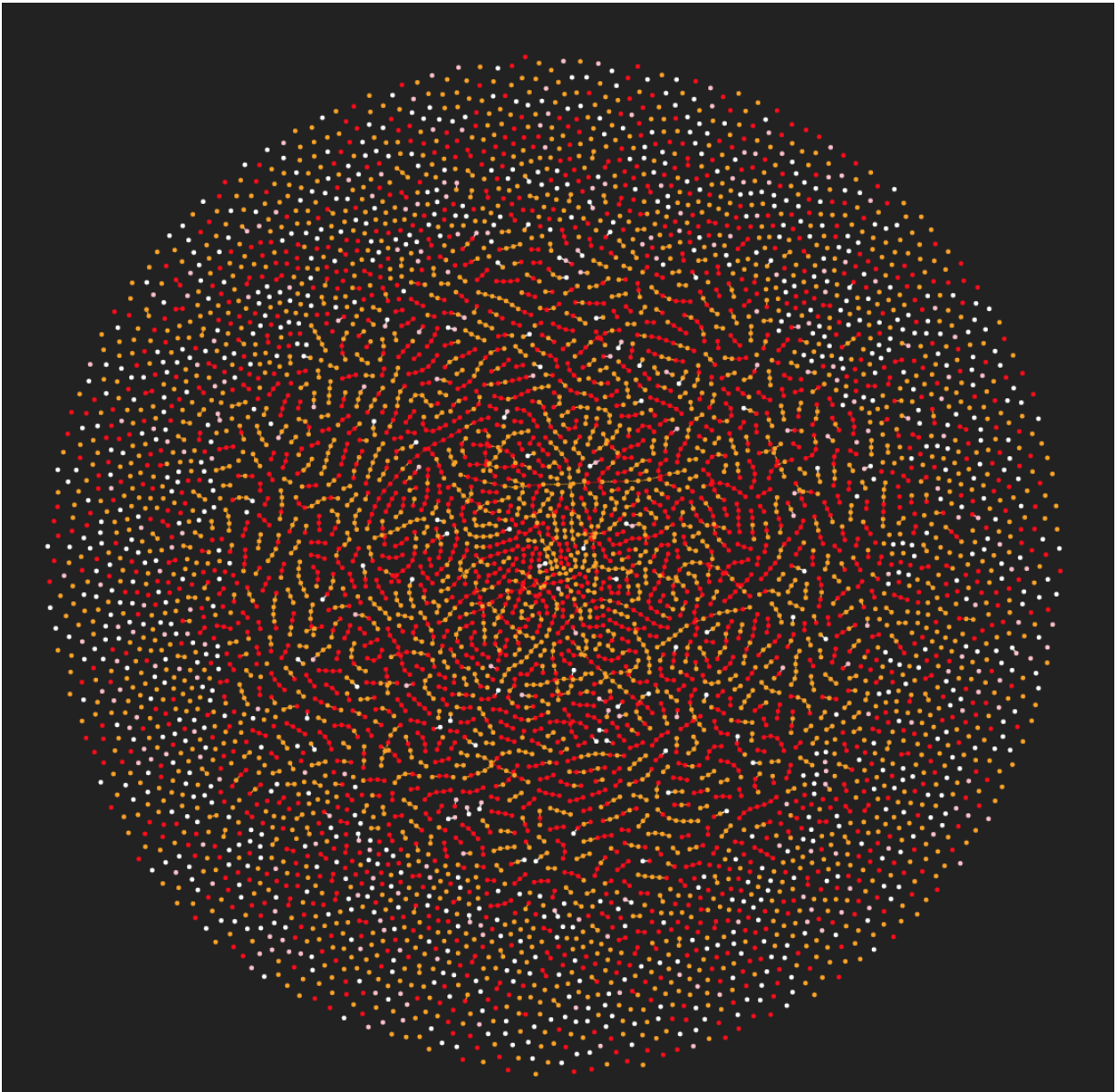
*Note: to view the networks in .html files, they should be downloaded and opened as an active tab for a few minutes.

predocs.html



*predoc network where each white ring is a training grant's predocs

nodes.html



*all individual timelines visualized per line

Award Spreadsheet Lookup + Network

Date Start/End	Title	Objective
11/08/2022 – 12/22/2022	Finding related projects and project families	Leverage existing ICER data to identify possible connections in a way that doesn't require technical expertise
<p>Issue: IES staff need to determine which projects (grants and/or contracts) are related to one another, but the existing raw data (in ICER) is incomplete. Parent-Child relations are not always captured, and "Grandparent"-Child relations are not always apparent. They are also difficult to read due to a residual suffix tag. Due to this, the related families are incorrect and incoherent within the grant search system.</p>		
<p>Approach: I propose to create an Excel file that would allow staff to look up awards and see whether the data suggests a possible parent-child relation or identifies possible award families. Staff are likely to have access to Excel and are familiar with the software; thus, it would be simpler for them to adopt.</p>		
<p>Planned Steps:</p> <ol style="list-style-type: none"> 1) Import ICER data into Excel and a pipeline to refresh the raw data 2) Create calculated columns that search the dataset for an award number based on simple rules to create a more comprehensive list of children and parents 3) Combine these 2 rule-based columns to form a general "Related" family column 		
<p>Outcome: I created an Excel document that calculates relations using search functions on the columns based on search rules. This file not only allows users to see the connections but also includes formatted and hyperlinked project titles. This allows users to quickly add information to abstracts.</p> <p>Below are descriptions of the different tabs within the Calculations file.</p> <ul style="list-style-type: none"> • <i>Data Import</i> - Paste exported data for use throughout the notebook. • <i>Table Data</i> - Table with calculated family search columns based on imported data. • <i>Tree</i> - Another table with transformed table, using an alternate method of finding child-parent relationships. • <i>Lookup</i> - Lookup information by grant award number (the primary tab that users leverage to search for information about an award) • <i>Fun Facts</i> - Interesting information we can see from the data • <i>Testing</i> - Preliminary functions I used to test functionality of the calculations • <i>Formulas</i> - Saved list of formulas in each sheet <p>Below are instructions for:</p> <p>Importing new ICER data into Calculations Excel file:</p> <ol style="list-style-type: none"> 1. Export/download data from ICER SharePoint 2. Open both excel files (ICER Data and Calculations) with the app (not in browser) 3. Select the table of exported data and copy it into the table in Calculations "Imported Data" sheet. <ol style="list-style-type: none"> a. In the owssvr sheet, select a cell in the table that is not a header like A2. Next, select all (except headers) with Ctrl+A. Copy with Ctrl+C. b. Go into the Calculations file and move to the first sheet tab called "Imported Data" and select A2. Paste with Ctrl+V. 4. At the top of the app, select Data and press the Refresh All button. (Ctrl+Alt+F5). 		

Development Information:

Files Added	Calculations.xlsx
-------------	-----------------------------------

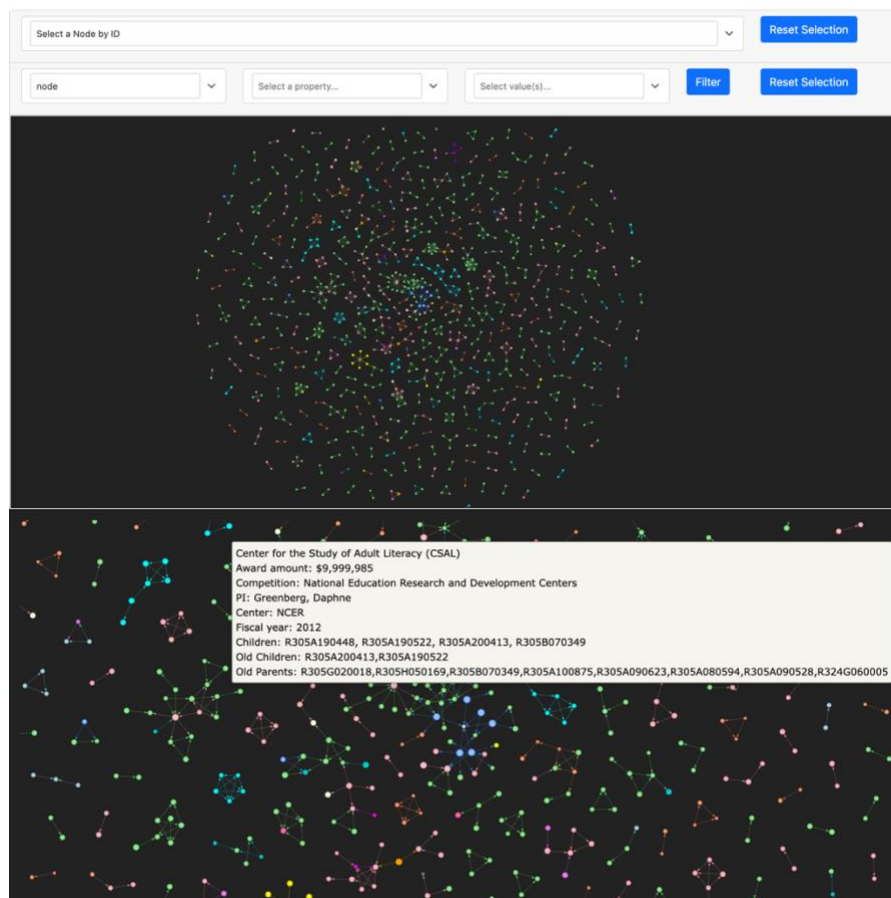
How the File Works:

I created a pair of check columns to find new child/parent matchings in the data and a pair of final columns to combine the pre-existing and calculated search matchings. With this, the co-parent is also extracted from the parents list. From the 2 final columns, a new “Related” column is created that combines them to form an immediate family. More levels of the family structure are found by continuing this search matching with the family columns: “R2”, “R3”, “R4”, and “R5.” Columns R4 and R5 are the same because there are no new levels found after R4.

Photos

Award	R305A110143
IES ID	1139
Fund type	Grant
Center name	NCER
Fiscal year	2011
Program Officer/CO	Elizabeth Albino
Award number	R305A110143
Title	A Toolkit for Identifying and Assessing Socially Rejected Children
Award amount	\$2,325,731.00
Supplemental funding	0
OrigAward+Supplement	2325731
Principal investigator	McKown, Clark
Grantee institution/CO	Rush University Medical Center
Grantee institution/CO	R
Grant search link	http://ies.ed.gov/funding/grantsearch/details.asp?ID=1139
Competition	Education Research
Topic/Strand	Social and Behavioral Context for Academic Learning
Project Type/Goal	Measurement: #9
Disability area	Not applicable: #30
Related	Extension or predecessor of another IES-funded project
Related children	R305A160053;#2253
Related parent(s)	
Modifications and act	
Updates and changes	0
Award period in months	48
Item Type	false
Path	Closed
Child Check	91990022C0040, R305A200220, R305A160053
Parent Check	
Child Final	91990022C0040, R305A160053, R305A200220
Parent Final	
Co-Parent	91990021C0028, R305A150189, R305A160053, R305A200463
R1 (Related)	91990021C0028, 91990022C0040, R305A150189, R305A160053, R305A200220, R305A200463
R2	91990021C0028, 91990022C0040, R305A110143, R305A150189, R305A160053, R305A200220, R305A200463
R3	91990021C0028, 91990022C0040, R305A110143, R305A150189, R305A160053, R305A200220, R305A200463
R4	91990021C0028, 91990022C0040, R305A110143, R305A150189, R305A160053, R305A200220, R305A200463
R5	91990021C0028, 91990022C0040, R305A110143, R305A150189, R305A160053, R305A200220, R305A200463

Elevated data quality by cleaning the incomplete input for the recorded children and parent awards, allowing for a lookup tool to be built.



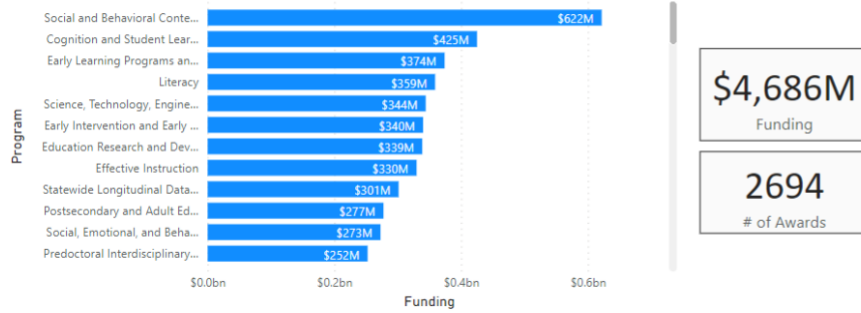
I created a visualization to present the award connections, so IES staff can easily see relationships and related awards to improve the ability to find similar awards and build off their findings in research as well as see how awards have sprouted from each other throughout time.

Award BI Dashboard

*some information missing due to PII

Date Start/End	Title	Objective
01/10/23-01/24/23	Leveraging Project Description and Focus Area Tagging	Create an accessible tool to allow non-technical staff to run tag analyses
<p>Issue: IES staff need to be able to identify grants and contracts that have similar research foci or that share other features that IES staff need to report on or analyze.</p>		
<p>Approach: I propose to use Power BI dashboards because IES staff have access to this software and because it does not require advanced data or software skills to manipulate data directly. This also goes beyond simple Excel analysis because it is more portable, user-friendly, and clean. Best of all, no download is necessary and all displayed chart data can be exported. After cleaning the data, Power BI will take it and provide a refined report.</p> <p>Steps: I used Power Query to clean data from the ICER Project Description and Focus Areas datasets, so every column's text is readable and correctly formatted. Afterward, I created reference tables per column using the original dataset in order to split the clean lists from the newly formatted columns, so each tag is separated and properly shown in filters. Finally, I created a Power BI dashboard with the dataset with a collection of visualizations and filters available to enable free exploration.</p>		
<p>Outcome:</p> <p>I have created a Power BI dashboard as described in my approach outline. In the future, IES may also be able to embed the file into the SharePoint site for greater accessibility through the data pipeline I have made. Using this dashboard, staff will be able to filter data easily to ask and answer their own questions. The file is here: Dashboard URL</p> <p>Below are brief descriptions of the features of the Power BI dashboard:</p> <p>Table: A simple, complete list of awards and their tags, with filters for each unique property, a query search bar, and a title search bar (using the bottom left filters allows for cross-comparison filtering, e.g. filtering for awards under major focus area of Mathematics AND Science).</p> <p>Graph: A few demonstrative graphs to show the power and scope of Power BI visualizations</p> <p>Graph+: 2 graph templates that plot Student Age and Year versus any selectable award property (quantified by award amount)</p> <p>Q&A: A DIY, querying visualization that provides a data/visual/statistic upon receiving a user's written query request. A few suggested queries that answer plausible questions IES may ask are shown for didactic demonstration. (The examples in the existing tab are based on an analysis of IES's examples. See Appendix B for a table that shows examples of mapping from questions to queries.)</p> <p>Influencer: A built-in machine learning algorithm to find key relationships in the data and how award properties tend to impact award funding on average.</p> <p>Tree: An interactive breakdown of award amount and count by attribute</p> <p>Demo: A possible informational report that could be sent to staff (showcased in front of supervisors in presentation)</p> <p>IES may wish to revise, add, or delete some of these tabs in the dashboard. I developed them based on my understanding of the use cases and the examples given (see Appendix A). Additionally, IES may need to curate some training materials if staff need more documentation or support to understand Power BI features.</p> <p>How to Update Data for the Dashboard</p> <p>Because IES may have new or revised underlying data (i.e., the ICER data may be updated), IES may need to refresh the files that power the dashboard. There are a couple options for this as explained below, assuming the exported data retains the same column structure.</p> <p>Option 1: Manually Replace Data</p> <p>Download data and then in the Power BI file, go to File -> Options and settings -> Data source settings > Right click data sources and change source.</p> <p>Option 2: Locally Stored Data</p> <p>Download data and replace the original table in the Excel file with the newly downloaded data. Then hit refresh on the Power BI file on your computer and hit Publish.</p> <p>Option 3: Automatic Refresh</p> <p>Link the 2 SharePoint lists to the Power BI file or add them as datasets and schedule an automatic data refresh.</p>		

Most Funded Programs



Program Name
Search
☐ Select all
☐ Accelerating the Ac...
☐ AI-Augmented Lear...
☐ Arts in Education

Center Name
☐ NCEE
☐ NCER
☐ NCER, NCSE
☐ NCES
☐ NCSE

Principal Name
Search
☐ Select all
☐ Aaron Lyon
☐ Aaron Sojourner
☐ Aaron Thompson
☐ Abhishek Datta
☐ Abi Olukeye

Goal
☐ Select all
☐ Efficacy
☐ Evaluation
☐ Exploration
☐ Initial Efficacy
☐ Measurement

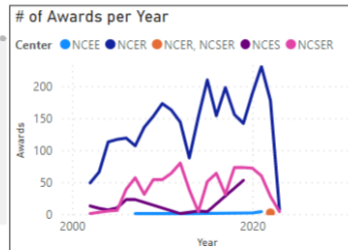
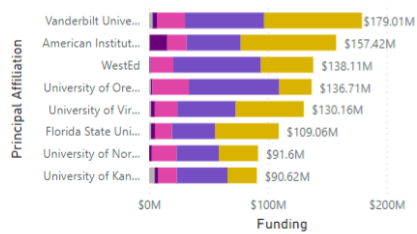
\$4,686M

2694

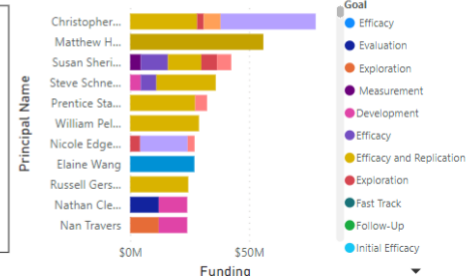
of Awards

Most Funded Affiliates over Award Period

AwardPer (bins) (Blank) 0 1 2 3 4 5

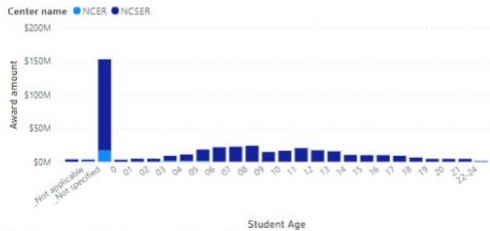


Most Funded Principals

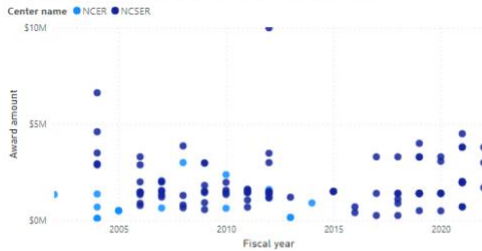


Select Attribute
Disability Area

Award amount by Student Age and Center name

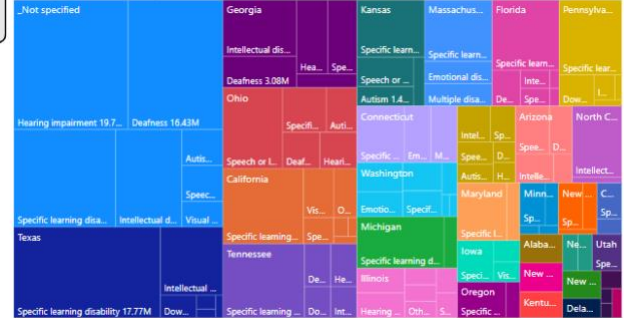


Fiscal year and Award amount by Award number and Center name

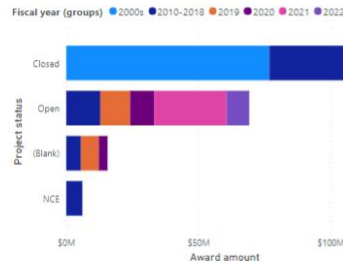


students with disabilities
Major focus area
Multiple selections

Funding by State/Territory and Value



Award amount by Project status and Fiscal year (groups)



\$211M

Award amount

Minor Focus Area
☐ Academic achievement policy
☐ Academic content
☐ Access
☐ Access to the general education c...
☐ Accountability
☐ Achievement gap
☐ Administrator effectiveness
☐ Adult education system

Appendix A: Background Materials from NCER/IES

Task 3: Analyzing Project Descriptions and Focus Area tagging concurrently

Problem:

Currently, there is no easy way to filter results inside ICER to export information from the ICER Catalog list based on values in the Focus Area list. In fact, there isn't even a way to view these data clearly in ICER. Even if we export the two lists and try to combine them (e.g., in Access or Excel), the data are so messy that most program officers cannot use them (e.g., all the hashtags and semicolons confuse them).

I was able to build (with a contractor's help) ways to filter one list base on the values of another list for internal viewing and leveraged this to be able to search by award to find the focus areas listed for that award:

<https://icer.ies.ed.gov/SitePages/All%20Focus%20Area%20And%20ICER.aspx>

However, the opposite type of view was not possible and still would not solve the problem because the filtered list is still not exportable.

Previous interns have tried to create different ways to combine the data such that people could get counts for projects, get some information about the award components (e.g., who the Principal Investigators were, what FY the projects were funded in, how much money for each award). But the tools we currently have are out of date and would need to be rebuilt anyway.

Request:

Help create some sort of file or portal or tool that would allow staff to run analyses that could address questions such as the ones above. This could be something in Excel or other offline software or maybe something SharePoint based, as long as it allows for the export of filtered data so that people can run analyses, create tables/charts, etc.

Ideally, the solution would include all of the data from the Interns View of the Project Description/ICER Catalog data along with the Focus Area data (including both Major and Minor foci).

Appendix B: Q&A Information

Q&A Visual: Question to Answer

Proposed Question	Q&A Query Input (Example)
For projects with a focus on postsecondary education, how many were development projects, exploration projects, or evaluations?	Count awards and project type goal when grade level is postsecondary
What is the distribution of math-related research across education settings (e.g., elementary school, high school, postsecondary)?	sum funding, mfa Mathematics, grade level
What program officers have overseen research relevant to early learning?	PO and major focus area Early Learning
In which states is NCER and NCSER development research take place?	Count award number, state, project type goal contains "Development"
What focus areas are common for grants funded under the Low-Cost, Short-Duration Evaluation competition?	Count "Low-Cost, Short-Duration" in competition per major focus area
How many awards are tagged under Vocabulary?	Minor focus area Vocabulary and award count
View distribution of minor focus areas within major focus areas.	Count award, major focus area, minor focus area in stacked bar

Publishing IES Data Online (Unstructured Data Cleaning)

Date Start/End	Title	Objective
02/14/23-03/07/23	Leveraging Project Description and Focus Area Tagging	Create an accessible tool to allow non-technical staff to run tag analyses
<p>Issue: The IES staff aim to evaluate if and how they can utilize unstructured data as structured data. This is in line with their goal of improving user-friendliness of public data while also identifying limitations in achieving this objective. The project's focus is on the Abstract field available publicly online, which currently contains a considerable amount of unstructured data that is challenging to analyze and update.</p>		
<p>Approach: Specifically, we are interested in the "Abstract" field from the publicly available IES website data. This field contains a lot of information formatted in HTML code, including co-principal investigators, associated publications, and project descriptions under different headings. To make this information more structured, I explored different methods to extract the information from this field and populate new fields.</p>		
<p>Steps: I used Python code to process the HTML and parse possible headers and corresponding text automatically.</p> <ol style="list-style-type: none"> 1. Load data from an Excel file using pandas and do initial reformatting of the dataset. 2. Extract years from the 'AwardPer' column and clean text in 'Title' and 'AwardAmt'. 3. Consolidate goal names in 'Goal'. 4. Merge columns and fix column names. 5. Extract data from HTML in 'Abstract' using BeautifulSoup library. <ol style="list-style-type: none"> a. Loop through each award's Abstract and parse the HTML content. Find all <p> tags. Loop through every <p> tag: Check if the <p> tag has a tag within it, and if so, extract the text from the tag as a potential header. If the <p> tag is the first in the list and does not have a tag within it, assume it's an introductory paragraph and assign to the "Intro" column. If the tag text meets certain criteria (e.g. length, capitalization), assume it is a header and add the corresponding text from the <p> tag as a value in the "row" dictionary. If the <p> tag contains other <p> tags nested within it, assume that they may contain additional headers and values. Extract the potential headers and loop through each one. For each potential header, loop through the <p> tag's contents (text and nested tags) and add any text that follows the header (up to the next header) to a list of potential values. Once all potential headers and values have been processed for the current <p> tag, add them to the table. Finally, clean the column names and stop. 6. Using fuzzy match, remove possible column headers in the detected texts. 7. Merge duplicate column header names and their underlying data 8. Remove columns without at least 1000 non-null values 9. Find "Investigator" in Abstract and create new column with more results than the automatic parse above 		
<p>Outcome:</p> <p>The parsing method in Python was not too clean and will need further refining if capturing every detail is necessary; however, it currently captures plenty of information (except outliers like tabled data). To increase data quality, I can search for specific key words in the Abstract to parse as headers and text. By searching "Co-Principal Investigator," I was able to get approximately 20 more results for the column with the list of co-principal investigators.</p> <p>However, the data is not balanced, and there is naturally empty data for many awards by construction of missing info from the source. For instance, the abstract of one award may have cost analysis while it may be unwritten but applicable in another. Balance might be important for the user to compare awards and seeing a significant number of empty rows might come as distressful. I can think of resolving this by filtering out null values. Another solution is to navigate the user to search within the rows. For instance, instead of looking at all names, the person can search for a specific person's name.</p> <p>Additionally, there are writing variations that increase the capacity for error. Going through the existing Abstracts and manually editing the more outlandish errors might be a route to consider.</p> <p>I used Power BI to show how much each center has spent on contracts versus grants, or how many grants were awarded to different institutions. I explored the existing data and created various visualizations to illustrate the requested information, although most of the data requires textual analysis to be properly represented because there are few numeric columns.</p>		

Field App

Screenshots of the field app I independently developed over the summer for safety forms, material requests, employee actions, and document viewing for 50+ employees.

Projects / Modules / Submissions

Projects

Search

City of Knowledge

19-029

Kindred Ontario VSI 3 & 4

19-036

Kindred SGV HVAC Replacement

20-017

Manhattan Beach Kinecta

20-097

5559 Burton Ave

21-034

La Brea Lofts

21-041

Park Reseda

21-048

Kindred Ontario Pyxis

21-055

Kinecta Gardena

21-058

Kindred Paramount (ARU)

21-059

What do you need to do?

Daily JHA

Weekly Safety Talks

Weekly Temp Power Checklist

Weekly Safety Inspection

Inspection (Request & Report)

Fall Protection Plan

Employee Actions

Safety Orientation

Safety Incident Report (SIR)

Lockout/Tagout

Field Impacts

11/17/2021 3:44 PM

Field Directive (T&M) - Change Work

ID: 168

Signed?

Yes

11/16/2021 3:45 PM

Field Directive (T&M) - Change Work

ID: 167

Signed?

Yes

11/16/2021 3:45 PM

Field Directive (T&M) - Change Work

ID: 166

Signed?

Yes

11/16/2021 3:38 PM

Field Directive (T&M) - Change Work

ID: 165

Signed?

No

11/16/2021 3:37 PM

Field Directive (T&M) - Change Work

ID: 164

Signed?

Yes

11/16/2021 3:14 PM

Field Directive (T&M) - Change Work

ID: 163

Signed?

No

11/16/2021 3:08 PM

Field Directive (T&M) - Change Work

ID: 162

Signed?

No

Forms and Actions

JHA

No Work

✓

Hazards

Search items

Cuts, crush, pinch, etc. during operation

Dark work areas when power turned off

Elevated noise level

Falls from ladders

Fire and Emergency response

Fire and explosion while refueling

Mitigation/Reduction

Cuts, crush, pinch, etc. during operation an...

Keep protective guards in place; disconnect from power source before servicing; use lockout-tagout; use PPE; keep away from power lines

Falls from ladders

Select proper ladder; maintain 4:1 slope ratio with straight ladders; seek assistance in dangerous and high traffic areas; don't carry tools while climbing; extend 36" if climbing over roof; do not use top step of stepladder and...

Daily Report

No Work

✓

* Project

20-097 Manhattan Beach Kinecta

* Report Date

6/2/2023

* Work Performed

Select item(s)

* Areas Worked

Select item(s)

* Crew

* Any Injuries

Meeting/Special Notes

21-092-001

Close

+

Search items

Isolation Point: Temp power panel

Adam C. Dome locked on 6/16/2022

Isolation Point: Temp power main

Adam C. Dome locked on 6/16/2022

Employee Actions

✓

* Action Type

Termination

* Action Date

12/31/2001

Last Date Worked

12/31/2001

* Full Name

Rikesh Patel

Employee Number

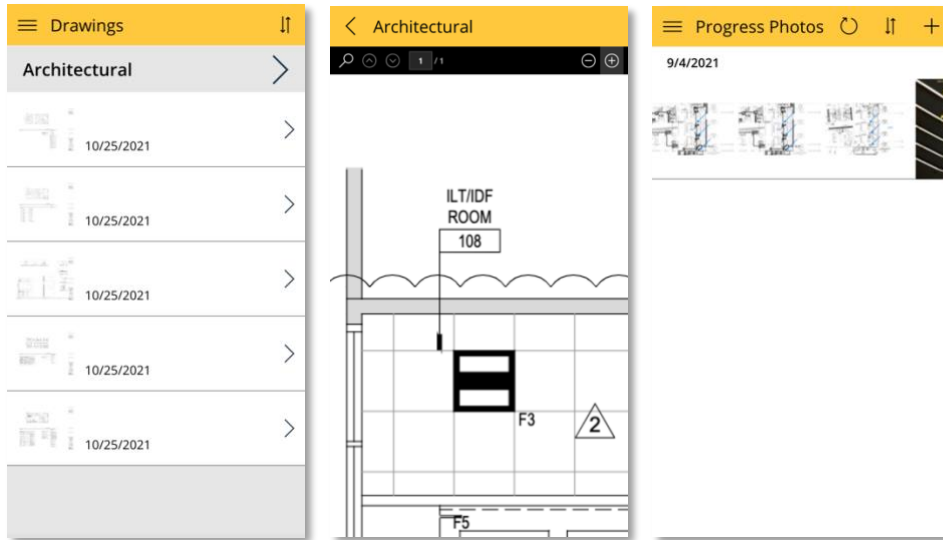
34

Employee Signature

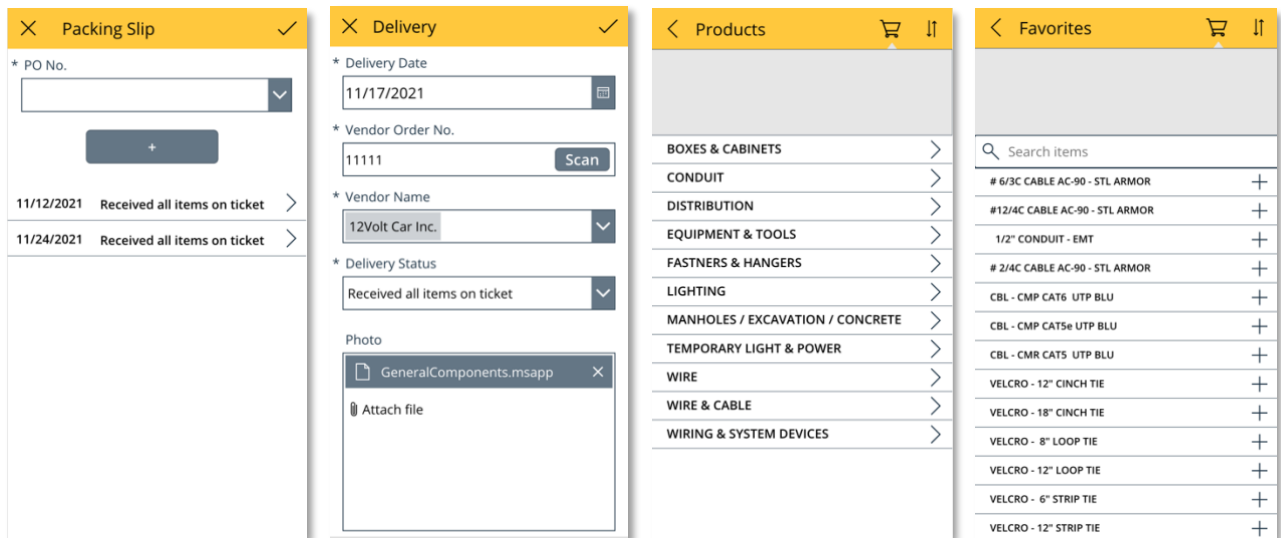
Rikesh Patel

Signature

Document Viewer and Photo Upload



Material Requests



BI Engineer Special Projects

During my path from intern to Business Intelligence Engineer, I have independently completed special projects:

- Finalized real-time A/R reports, integral to project management cash flow
- Cut costs by \$1000s/month for API development of cost-tracking BI dashboards
- Maximized safety form submissions by over 350% by publishing professional app
- Executed CEO's vision, automating HR and project estimation
- Conducted testing of GPT-4 transformer for contract chatbot integration
- Wrote ETL scripts in Python to receive HR and labor info into our structured database
- Simplified HR document signature requests through automated desktop flows
- Achieved full automation of HR employee onboarding process for efficiency

Search, Review, and Explore Restaurants (Website)

I was touring Greece and Rome and found that it was difficult for me to use Yelp to find nice places to eat. In frustration, I decided to create my own website to take advantage of reviews online as well as other data and piece it together to find the most suitable place to eat based on my query. I expanded this concept by implement NLP models, sentiment analysis, and adjective visualization to improve the user interaction.

Welp: Reviews and Ratings

by [Rikesh Patel](#)





Search Bar

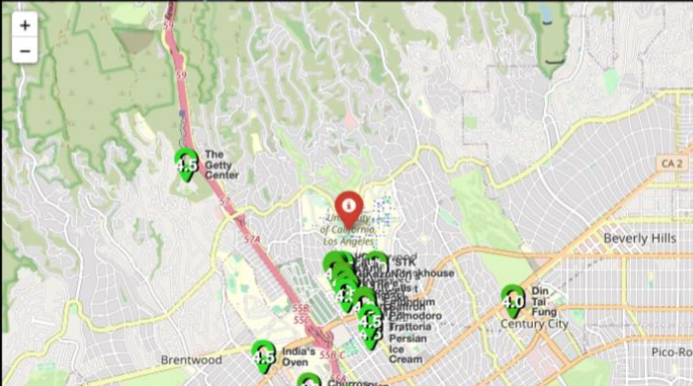
Type: ramen, tacos, pasta, burgers, chicken

UCLA

- University of California, Los Angeles, Westholme Avenue, Holmby Hills, Westwood, Los Angeles, Los Angeles County, CAL Fire Southern Region, California, 90095, United States
- UCLA, Marina del Rey, Los Angeles County, CAL Fire Southern Region, California, United States
- UCLA, Marina del Rey, Los Angeles County, CAL Fire Southern Region, California, 90296, United States
- UCLA, 4o Retorno del Roble, Villas del Álamo, Pachuquilla, Mineral de la Reforma, Hidalgo, 42184, México

Search

Name	Image	Reviews	Type	Rating	Transactions	Price	Phone	Miles	Address
Daddy Rice Cookies		5742	Desserts, Bakeries, Ice Cream & Frozen Yogurt	4.5	delivery, pickup	\$	(310) 208-0448	0.54	926 Broxton Ave, Los Angeles CA 900
In-N-Out Burger		891	Fast Food, Burgers	4.0	delivery	\$	(800) 786-1000	0.54	922 Gayley, Los Angeles CA 900
Bella Pita		830	Middle Eastern, Sandwiches, Mediterranean	4.0	delivery, pickup	\$	(310) 209-1050	0.57	960 Gayley Ave, Los Angeles CA 900
Pat Sals Deli		1406	Delis, Burgers, Sandwiches	4.0	delivery, pickup	\$\$	(310) 452-3220	0.58	972 Gayley Ave, Los Angeles



LinkedIn Job Finder

This project involves leveraging the power of LinkedIn and advanced NLP techniques to build a LinkedIn Job Finder system. This system will enable me to efficiently identify and target the most relevant job opportunities in the market.

Project Overview:

LinkedIn Integration: The project utilizes the "linkedin-jobs-scraper" package to seamlessly connect to the LinkedIn platform and extract job listings based on specific queries.

Data Extraction: The project extracts key details from job listings such as job titles, company names, posting dates, direct links to job postings, and insightful data about each job opportunity. This information is organized in a structured manner for further analysis.

2 event listeners are defined to handle the data extracted from the job listings and metrics information.

```
queries = [  
    Query(  
        query='Data Python',  
        options=QueryOptions(  
            locations=['California'],  
            limit=1000,  
            filters=QueryFilters(  
                relevance=RelevanceFilters.RELEVANT,  
                time=TimeFilters.DAY,  
                experience=[ExperienceLevelFilters.ENTRY_LEVEL]  
            )  
        )  
    )  
]  
scraper.run(queries)
```

Advanced Text Analysis: Leveraging the power of the popular "spaCy" library, the project applies advanced natural language processing techniques to analyze job descriptions. It extracts valuable information such as required education levels, required experience levels, and salary ranges. This automated extraction saves significant time and effort in manually analyzing each job description.

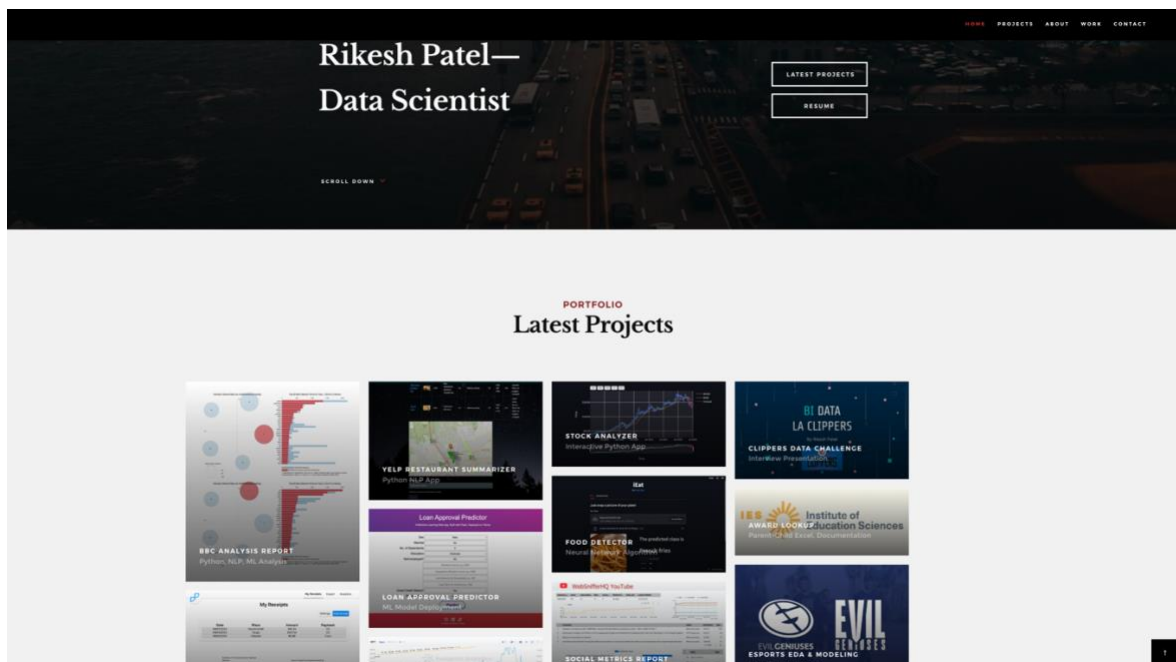
Duplicate Removal: The project employs fuzzy matching algorithms to remove duplicate job listings and spam. This ensures that you have a clean and unique dataset to work with, avoiding redundancy and confusion.

Overall, the LinkedIn Job Finder project offers a unique opportunity to enhance my job search strategy, save valuable time and resources, and make informed decisions based on comprehensive job market data. With its customizable queries, advanced text analysis, and interactive visualization capabilities, this project empowers me to stay ahead in the competitive job market.

Portfolio Website

Filled with anticipation and eagerness, I embarked on the creation of my very own personal website - a project I had been yearning to undertake for quite some time. Finally, the opportunity had presented itself, and I threw myself into the task with gusto, determined to bring my unique vision to life.

To do this, I used HTML, CSS, and JavaScript to set everything up. One of the most difficult features to implement was the changing backgrounds upon page refresh, but I persevered to get it to work, adamant on publishing the website only after I figured out how to set up this ostensibly trivial feature. It took much patience, but I finally found a solution. Then, I added mobile support for different screen resolutions and a contact screen page to accompany my resume.



BBC News NLP Report

Using a few models and different Python libraries, I took a look into a folder full of .txt files. In this report, I get to answer a few questions, analyze country name dropping, look at sentiment and polarity, classify articles into topics, and more.

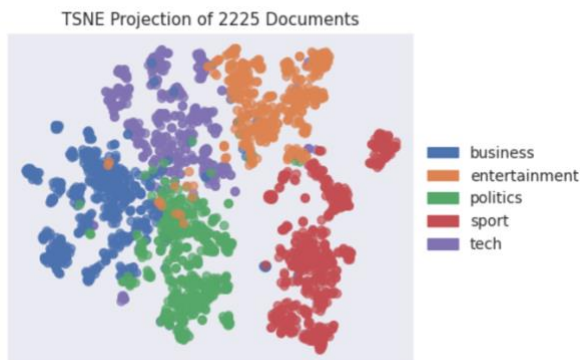
1. Describe the methodologies you used in your report.

In my report on identifying the most mentioned G20 country, I employed natural language processing (NLP) techniques to analyze the articles. Firstly, I utilized the Counter function to count the frequency of common word sequences in each article category after performing text cleaning. To address potential bias towards repetitive articles, I added an interim step to make every word unique per article. However, this approach may result in some two-letter phrases not being counted in certain rare cases. To obtain a count of repeated word usage, I tallied the number of times specific words appeared in the text, and I plotted the counts for each individual article. The final result was presented in a polished plot. Further exploration could be done to identify a method that avoids loss of two-letter phrases while still maintaining uniqueness.

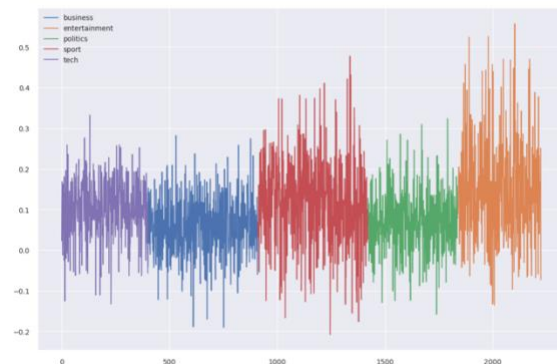
2. With more time, how would you strengthen your report?

If I had more time, I would integrate the geopolitical entity detection and geocoding I set up for fun with the simple country name counter to create a more accurate list of most mentioned countries out of all countries globally using the Google Maps API. I would also apply more topic models to check if they may have greater coherence. I would also like to stress test the script's speed by adding more news articles and articles in different topics to the dataset.

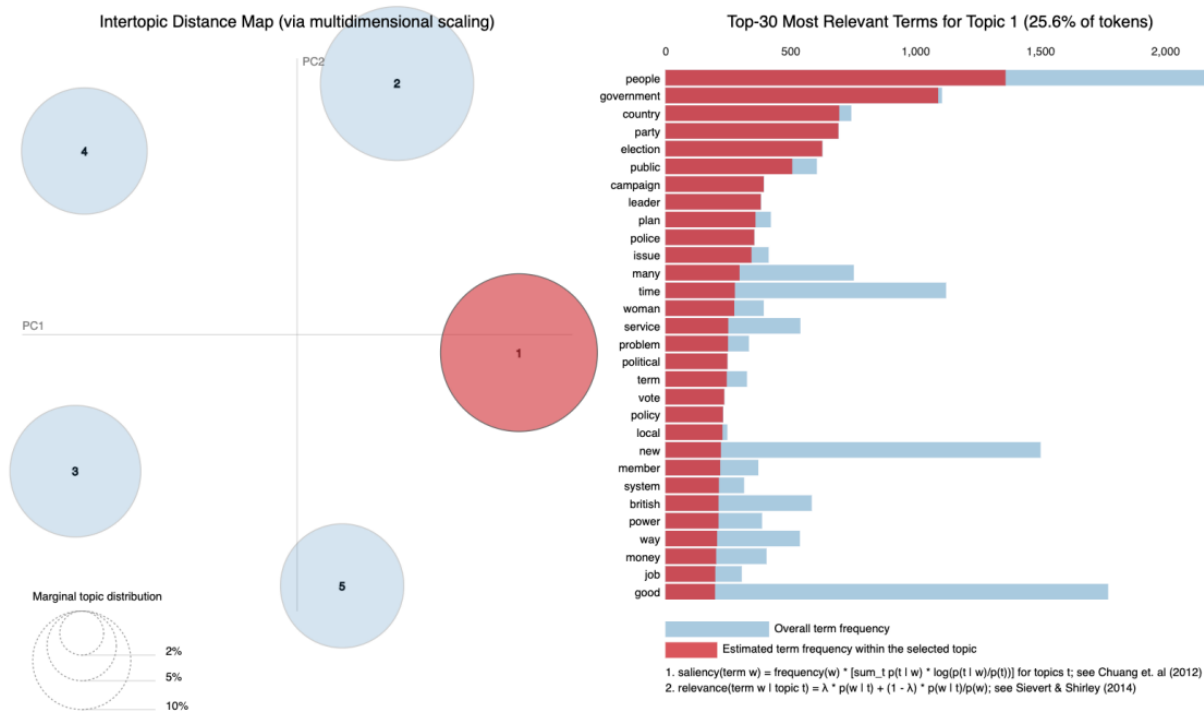
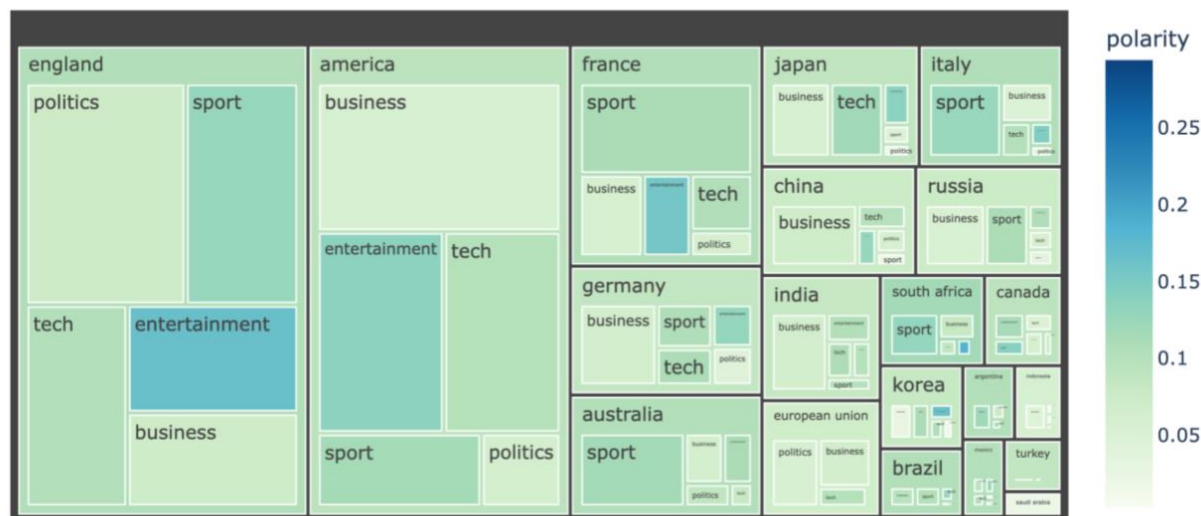
Article Similarity Distribution Visualization via t-SNE



Polarity by Topic










6 Treemap Distribution of Most Mentioned G20 Countries by Topic



SQL Challenge (LA Clippers Final Interview)

Before my L.A. Clippers final interview, I was given the challenge assignment of analyzing seating, tickets, and attendance data to answer some business queries and then package my solutions into a business presentation. After completing my presentation, I found many points to improve my skills on, and I know I gained a lot from the opportunity.

Ticket Package			Game Date & Time		Income Source		Recommendation		
Price Per Game			Home Game Data						
	Group	\$35-\$595	01	42 home games out of 82 games in total		Ticket Sells	Adjust promotion strategy based on key features Consider overselling tickets to ensure attendance		
	Mini Plan	\$90-\$390		02	9 games in the afternoon 33 games in the evening		Advertising	Charge for in-arena advertising based on headcount	
	Half Season	\$55-\$371			03	6 on Mon, 5 on Tue, 7 on Wed, 4 on Thu, 4 on Fri, 8 on Sat, 8 on Sun		Merchandise	Estimate the storage of merchandise and revenue from selling them
	Full Season	\$68-\$337							

2)[SQL]For all games, provide the average number of fans in arena 2hrs prior to start time, 1hr prior to start time, and at start time.

	Arrival	Avg_Attendance
0	2 Hours Early	4
1	1 Hour Early	705
2	At Start Time	3574
3	15 Minutes Over	4552
4	30 Minutes Over	5176
5	1 Hour Over	5613

Based on scanned tickets data, on average, there are 4 fans are seated 2 hours early, around 700 fans are seated 1 hour early, and nearly 3600 fans are seated by the start of each game.

Esports EDA Assessment

Determine if this dataset needs any preprocessing. If so, clean the dataset and document your steps. If not, explain how you came to that conclusion.

Multicollinearity has a negative impact on many popular ML models. Check if this dataset experiences any multicollinearity. If so, reduce the impact until an acceptable point. The data will definitely require some preprocessing due to the presence of null values, indicated by "?"s. I will also convert column types, remove outliers, and remove high VIF columns.

Determine what are the most important features that could help predict a player's rank?

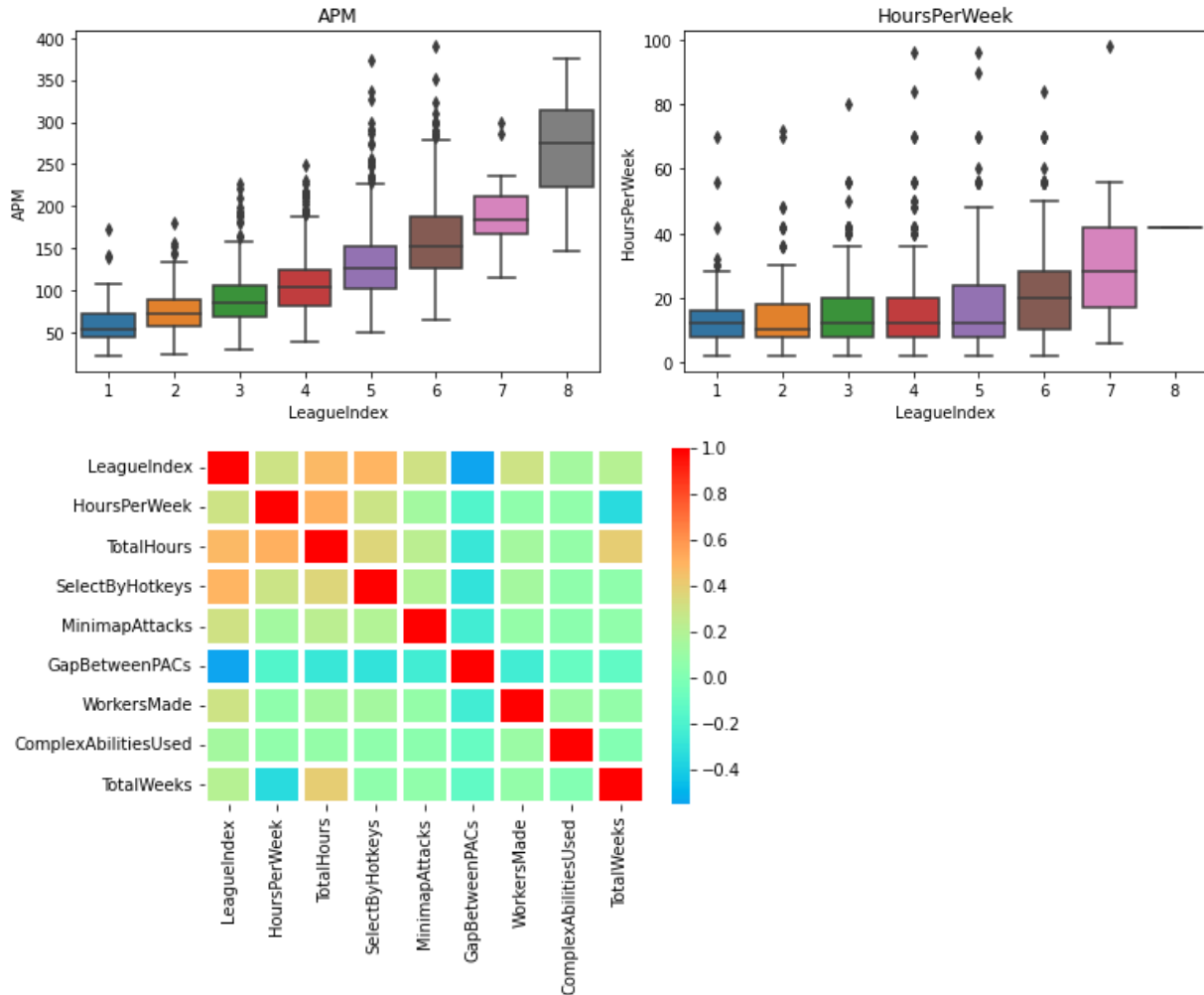
The features most strongly correlated to LeagueIndex are **APM, ActionLatency, and GapBetweenPACs**. Correlation is important because it measures the linear relationship between two factors. For instance, a higher APM and GapBetweenPACs is associated with a higher LeagueIndex. Conversely, a lower latency is associated with a higher league. TotalHours and SelectByHotkeys will also be shown as important features in the demo model I built. These results are natural because these features are the building blocks of what enables a player to improve: low latency and total hours. If latency is higher, a player will not be able to showcase all their skills and will also have lower action speed. **A minimum total hours is also necessary** to practice enough and beat the learning curve. A natural indicator of skill, albeit biased. Meanwhile, players in higher leagues tend to be more skilled, i.e. faster actions (APM, Hotkeys). A more nuanced measure of skill (strategic/calculative skills) is not provided in the dataset. A skill indicator like win rate would be valuable, but players are matched against those in the same skill bracket. Because of this, the win rate will approach 50% when there is 1 winner and 1 loser per game. Therefore, **action speed is the most prominent differentiator**. Raw speed is not normally an effective metric for competitive games, but StarCraft centers around multi-tasking, so performance will be stunted with fewer actions, regardless of strategy or skill.

Your team's Starcraft2 coaching staff loved your project! They think this is perfect for scouting rising stars. Using your discoveries from (3), create a function to find players who should be given a chance to become professionals. Explain why your set of players make sense.

By looking at the feature distribution grouped by each LeagueIndex, I saw a statistical difference between those in the professional league versus those in lower leagues. By taking the core stats of Professional League players, I believe I can filter the players and create a list of players with high potential. League 7 players have an average playtime of 31 hours/week (10 hours more than League 6 players and almost double League 5 players), which represents the time commitment required. There is also a baseline of at least 1000 total hours of playtime, which is flexible but serves as a foundation for players' skills. I believe there is further error in the total hours played input because higher league players may create second accounts after playing for a while. Continuing by analyzing the minimum, first quartile, third quartile, and maximum values, I decided to filter as shown below: $APM \geq 250$, $GapBetweenPACs \leq 23$, $ActionLatency \leq 40$ For a smaller list, also filter by $HoursPerWeek \geq 25$, $TotalHours \geq 1000$.

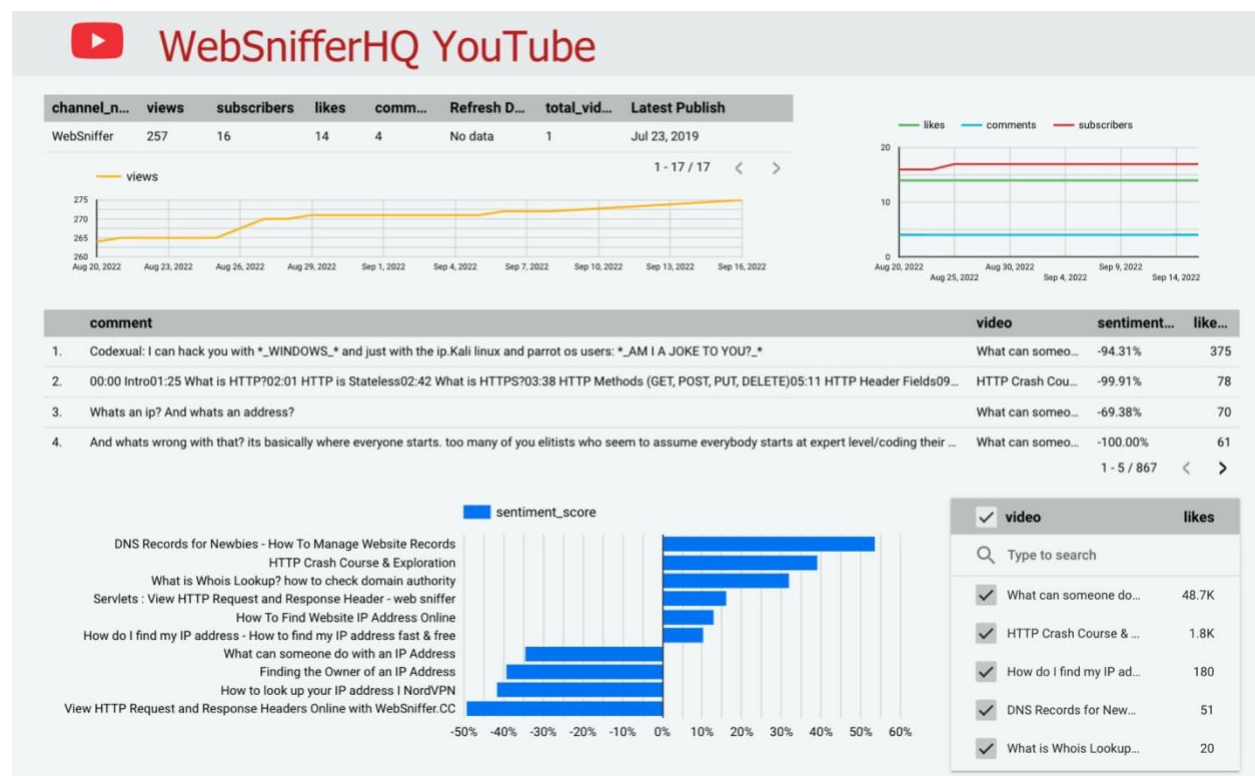
Hypothetically, if you were to move forward with creating a fully-fledged model to predict LeagueIndex, what model(s) would you consider and why? (Don't actually implement anything!)

Linear regression and SVM may have trouble due to the hyper-overlapping of the players between leagues in terms of the other features in the dataset, as seen in the plotted pair plot. The regression would not accurately place a line through the scatter plot while differentiating between leagues. KNN, Decision Tree, and Random Forest models are popular picks because they are based on robust algorithms that systematically succeed, so I would probably start with them to pick out the league for the similar-looking players. I did some preliminary research into trying to model the data already for Question 3, and I believe regression makes more sense in terms of accuracy and output than classification (although it depends on the goal).



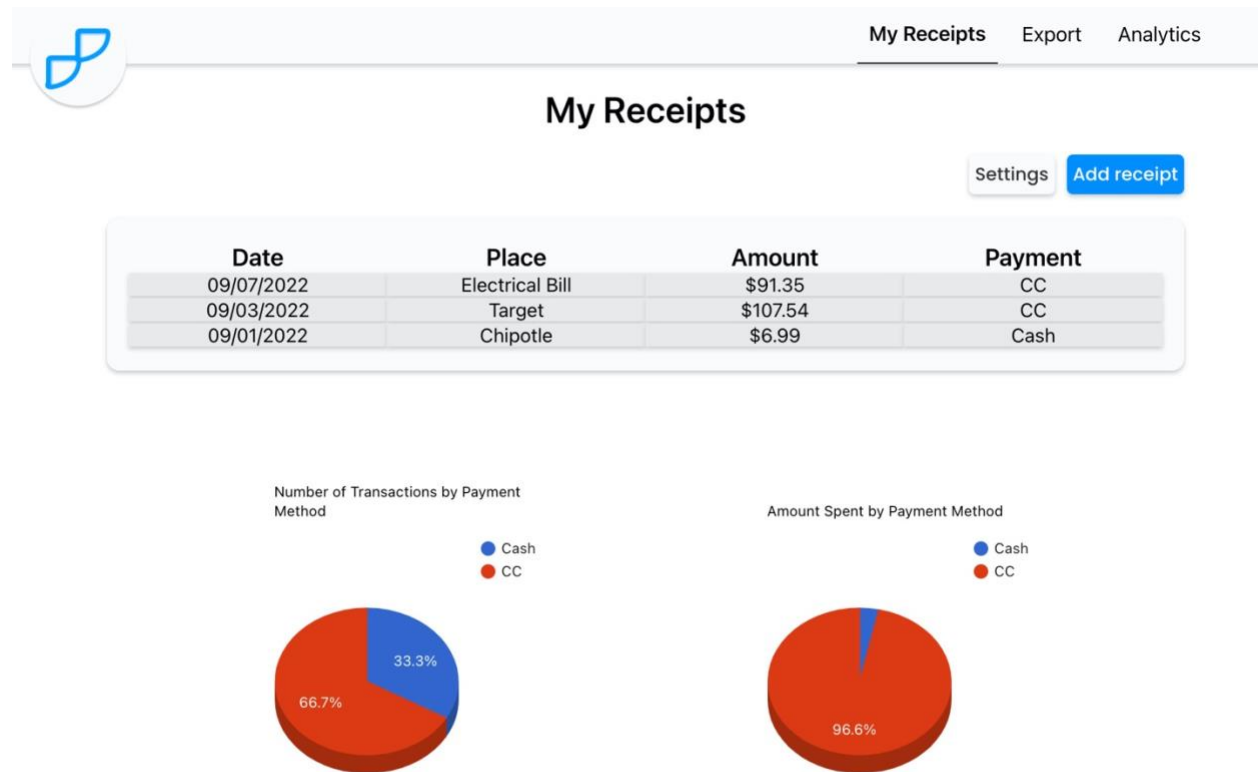
Socials Dashboard (Google Data Studio)

I was elected to be the Data Analytics Team Lead during my WebSniffer internship. I provided guidance to my team members while communicating with upper management. I helped my team by offering technical expertise and advice on data analysis, and I also worked with the team to identify and resolve any obstacles that arose. One of the key methods I helped the team achieve its goals was by providing clear communication: I ensured that everyone on the team understood their roles and responsibilities by working closely with the team to ensure that everyone had the resources they needed to be successful. Overall, my experience as a team lead taught me the value of collaboration and problem-solving in achieving team goals; I believe that these skills are essential for success in any team environment, and I am confident I can contribute these pillars to any future team I am accepted into.



Receipt OCR Website

Receipts are a transcript of all of our expenses, but many people choose to throw them away. I wanted to repurpose this nuisance of paper into useful statistical measurement of cost management. By creating a simple app to read receipts and transform their numbers into a report, I accomplished my initiative to add utility to something that has few immediate benefits.



Loan Predictor App

I created a simple framework to host my machine learning classification model. I thought that the finished product was a interesting display of the capabilities of machine learning, so I am happy to have created it.

Loan Approval Predictor

A Machine Learning Web App, Built with Flask, Deployed on Vercel.

Sex

Male

▼

Married

No

▼

No. of Dependents

0

▼

Education

Graduate

▼

Self-employed?

No

▼

Monthly Income e.g. 2400

Coapplicant Monthly Income eg. 2400

Loan Amount (in thousands) e.g. 300

Loan Term (in months) e.g. 360

Good Credit History?

Yes

▼

Property Area

Urban

▼

Predict

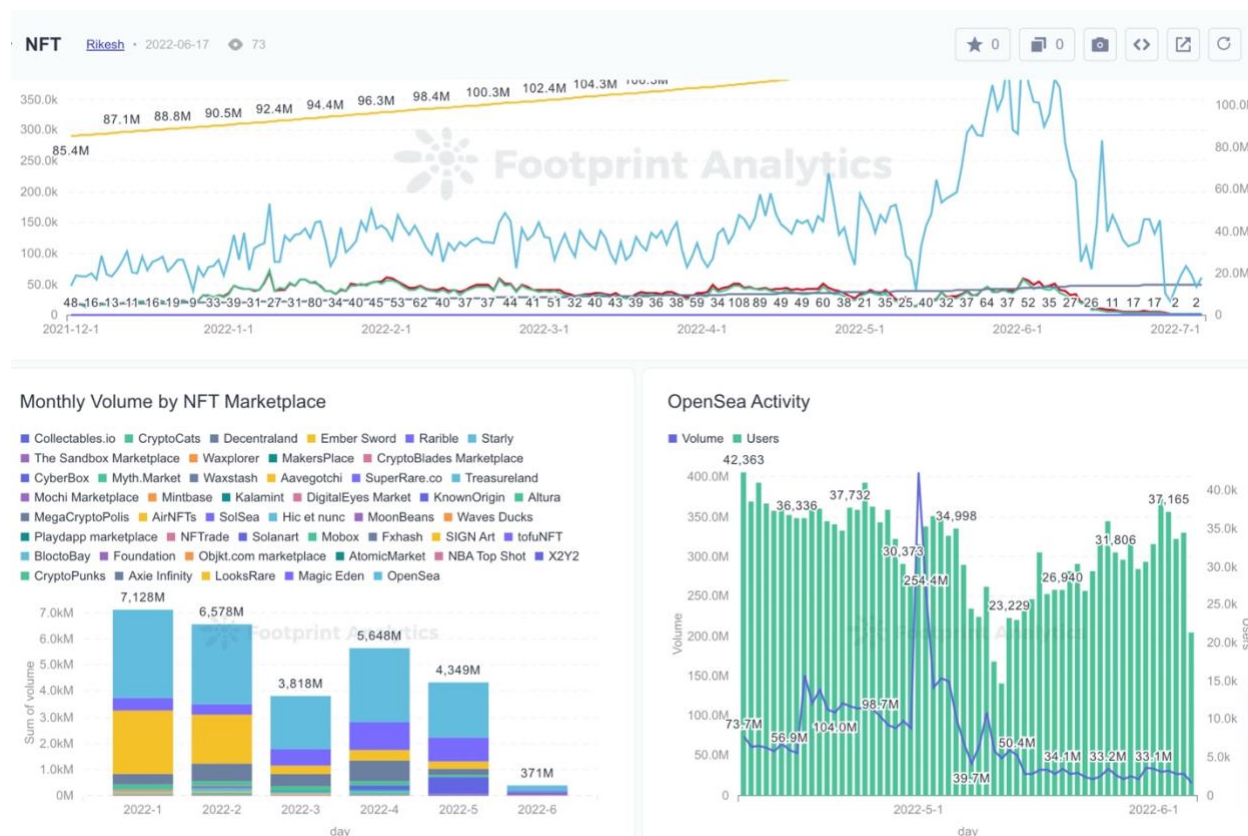


Created by Rikesh Patel.

NFT Market Analysis

While researching cryptocurrencies during my internship with Fayble Inc., I decided to create a dashboard on my own time to capture the state of the market. We had weekly presentations, and I thought this dashboard would serve well in tying my conclusions together, especially as a showcase of the market to my clients in the audience.

For instance, I found that the recent downward surge in prices during the end of summer 2022 was propelled by unfavorable macro structures: liquidation, derivatives, and loan markets. Most investors were liquidating their holdings, so there was less cash pumping the market and the prominent action became selling.



Stock Analyzer

I wanted to play around with real world financial data and try my hand at predicting the future of the stock market. Taking Yahoo finance data, I was able to construct a graph using historical stock data with a moving average and forecast with LSTM. I initially thought LSTM would be valuable in stock prediction to capture both macro and micro trends. Upon further research, I learned they can be fine-tuned, but this process is time-intensive and hyper-focused. In terms of efficiency, we can conclude that the LSTM network's performance fails to meet expectations when compared to other relevant models, like AR and ARIMA that better fit modeling stock price time series data. Ergo, LSTM is inferior in terms of extrapolating ability and requiring large and consistent data.

There are several options to visualize a stock or index via ticker symbol and a button to download the data shown.



Food Image Recognition

I needed to track my caloric intake and nutritional gain for my Nutrition class, and I thought of the perfect solution: Neural networks! I used a set of 100 common food images to train my model and built a Streamlit app to be its framework.

