Project 3

Donald Moratz*, Shanze Fatima Rauf†, Daniel Wang‡and Ruiwen Yang§

March 2025

An interactive version of many of our charts can be found at our GitHub site![1]

1 Project Summary

This project establishes a validated extension of the repository of research outputs leveraging Federal Statistical Research Data Centers (FSRDC). Under project 2, groups from UPenn's CIT5900 class collected potential FSRDC research outputs. Extant sources provide an account of some these outputs, but high levels of missingness makes it difficult to track the extent and impact of FSRDC-related work. This project seeks to validate and incorporate the outputs found under the auspices of project 2. We utilize public APIs (specifically OpenAlex) and advanced text-processing steps to systematically locate and confirm whether a reported publication from the groups of project 2 uses FSRDC microdata through the presence of restricted-data acknowledgments, RDC location references, and disclosure-review statements. Our approach employs both exact and fuzzy name matching to handle the inherent variability of author and project naming conventions, thereby minimizing duplicate entries.

A major contribution of this effort is a combined dataset of 1,662 unique publications, all validated via use of OpenAlex API. This validation process narrowed the potential sources down from a list of nearly 18,000 unique potential outputs. Throughout data collection, we pay particular attention to screening out invalid or incomplete entries, ensuring that only outputs meeting the FSRDC-use criteria remain. This rigorous curation helps shape a comprehensive record of the diverse scholarly outputs (articles, working papers, dissertations, etc.) generated under the FSRDC umbrella. In turn, the resulting dataset lays a firm groundwork for advanced analysis and serves as a valuable resource for stakeholders interested in the scope and evolution of FSRDC-supported research.

Our exploratory data analysis (Section 3) shows that the newly identified outputs nearly match the previously identified outputs in number, suggesting a significant portion of outputs from FSRDC supported research have gone previously unpublished. A graph of the counts over the years shows that in the past 10 years, an average of nearly 100 research outputs have been generated via FSRDC-related research. We also show the impact of specific FSRDC centers, highlighting Penn State and Yale as driving works with particular high average citation counts. Using H-index, we identify a handful of authors who drive citations of FSRDC research, including Miranda, Kerr, Haltiwanger, Foster and Holan.

*dmoratz@sas.upenn.edu
†sfrauf@sas.upenn.edu
‡wang50@seas.upenn.edu
§ruiweny@seas.upenn.edu

[1] `https://ruiweny316.github.io/CIT5900_project3/`

The analysis in Section 4 extends from descriptive statistics into regression modeling, relying on NLP-based topic classification to categorize each publication. First we examine the linkage between title length, time since project start and associated RDC to generate a logistic regression predicting research output type. In our analysis, we find evidence that even this basic data is sufficient to predict output type with an accuracy of roughly 67%. Our second analysis explores the dynamics behind citation counts. We identified in project 2 that papers concerning labor economics tend to have higher citation counts. Using this more extensive dataset, we utilize FLAN-T5, an NLP based classifier applied to titles to determine whether the title is related to labor economics. We attempt to narrow down parameters for our regression utilizing PCA, but identify no meaningful application of either PCA or K-Means clustering. Applying a simple linear regression, we estimate the relationship between citation count and several features of the data. We find an unsurprising relationship between years since published and citation count, and a slightly negative impact of title length, but unlike in our previous analysis, we find no evidence that publications related to labor economics experience a significant increase in citations. Finally, we extend our analysis by applying BERTopic to the keywords we found for each research output. Using the topics generated by this approach, we explored the relationship between topic and citation counts and were unable to find any meaningful connections.

This report is organized as follows. Section 2 provides details on the data processing techniques and project, output pairings. Section 3 describes the validation and data enrichment of the research collected via the groups in project 2. Section 4 discusses our initial exploratory data analysis, focusing on trends in the data. Section 5 delves into broader empirical findings—from topic-based citation patterns, logistic regression prediction, and dimensionality reduction efforts.

2 Data Processing

2.1 Data Integration

Since each group has different columns reported in their final outputs, we need to identify the columns that are unique to only one group, and the columns shared by at least two groups, not only by the column names, but also by the content of the columns due to different naming conventions might be used. For columns shared by at least two groups, we need to make sure that they are corrected with the same name so that later on they could be concatenated together without any problems.

At the same time, we will also drop some unnecessary columns that will not be helpful in later steps to reduce the size of combined data later, such as the column indicating whether the research output is the FSRDC output without any evidence provided, since for all reported research outputs, they were considered as valid FSRDC outputs, there is no extra benefit to include such columns.

We will also process each group's output a little to make sure the format is consistent across all group outputs:

- Convert all column names to lowercase except group 8 (most columns are exactly the column names used in 2024 research output excel file)
- Add group and idx columns in each group's output for easier retrieval of the original record in the group output
- Any other necessary processing to make sure shared columns have consistent data format

Group 1 Cleaned Output:

- Shared Columns to be Renamed (original name: new name):
  - title: OutputTitle
  - project_id: ProjID
  - project_pi: ProjectPI
  - agency: ProjectRDC
  - project_status: ProjectStatus
  - year: OutputYear
- Dropped Columns: keywords
- Additional Processing: None
- All Remaining Columns: OutputTitle, abstract, authors, source, URL, acknowledgments, data_descriptions, disclosure_review, rdc_mentions, dataset_mentions, ProjID, ProjectPI, ProjectRDC, ProjectStatus, OutputYear, doi, institution_display_names, raw_affiliation_strings, detailed_affiliations, group, idx

Group 2 Cleaned Output:

- Shared Columns to be Renamed (original name: new name):
  - title: OutputTitle
  - year: OutputYear
  - type_crossref: OutputType
  - author_id: openalex_id
  - location: ProjectRDC
- Dropped Columns: publication_date, cited_by_count, topics
- Additional Processing: Get rid of curly brackets, separated by semicolon in the authors column
- All Remaining Columns: OutputTitle, doi, abstract, OutputYear, authors, affiliations, source_display_name, OutputType, researcher, openalex_id, queried_dataset_terms, matched_dataset_terms, ProjectRDC, group, idx

Group 3 Cleaned Output:

- Shared Columns to be Renamed (original name: new name):
  - title: OutputTitle
  - rdc: ProjectRDC
  - start year: ProjectYearStarted
  - end year: ProjectYearEnded
  - pi: ProjectPI
  - publication_year: OutputYear
  - is_published: OutputStatus

  - type_crossref: OutputType
  - author: authors
  - author_affiliation: affiliations

- Dropped Columns: unnamed: 0, # of papers, # of fsrdc-relevant papers, fsrdc evidence, has fsrdc evidence?, cited_by_count, primary_topic, field, updated_date, author_address_locality, author_address_region, author_address_country, data_source, keywords, record_id
- Additional Processing: None
- All Remaining Columns: OutputTitle, ProjectRDC, ProjectYearStarted, ProjectYearEnded, ProjectPI, abstract, doi, openalex_id, OutputYear, host_organization_name, OutputStatus, OutputType, authors, orcid_id, affiliations, agency, group, idx

Group 4 Cleaned Output:

- Shared Columns to be Renamed (original name: new name):
  - title: OutputTitle
  - year: OutputYear
- Dropped Columns: citations, has_fsrdc_evidence
- Additional Processing: Get rid of double quotation marks and affiliations in parentheses in the authors column
- All Remaining Columns: researcher, OutputTitle, OutputYear, abstract, authors, source, fsrdc_acknowledgments_evidence, fsrdc_data_sources_evidence, fsrdc_disclosure_evidence, fsrdc_rdc_locations_evidence, group, idx

Group 5 Cleaned Output:

- Shared Columns to be Renamed (original name: new name):
  - pi: ProjectPI
  - year: OutputYear
  - title: OutputTitle
- Dropped Columns: title_clean
- Additional Processing: None
- All Remaining Columns: ProjectPI, OutputYear, doi, OutputTitle, group, idx

Group 6 Cleaned Output:

- Shared Columns to be Renamed (original name: new name):
  - title: OutputTitle
  - researchers: researcher
  - year: OutputYear
  - rdc: ProjectRDC

- Dropped Columns: no. researchers, keywords
- Additional Processing: None
- All Remaining Columns: OutputTitle, abstract, doi, researcher, OutputYear, ProjectRDC, dataset mentions, group, idx

Group 7 Cleaned Output:

- Shared Columns to be Renamed (original name: new name):
  - title: OutputTitle
- Dropped Columns: is_fsrdc
- Additional Processing: None
- All Remaining Columns: OutputTitle, abstract, source, group, idx

Group 8 Cleaned Output:

- Shared Columns to be Renamed (original name: new name):
  - DOI: doi
  - Authors: authors
  - Abstract: abstract
  - ProjectPI: researcher
- Dropped Columns: Unnamed: 0, ProjectID, ProjectStatus, ProjectRDC, ProjectStartYear, ProjectEndYear, OutputBiblio, OutputStatus, OutputVenue, OutputMonth, DoiExtract, NormTitle, MatchType, Uniqueness, FuzzScores, FSRDC_related
- Additional Processing: Use semicolon to separate values in researcher and authors columns
- All Remaining Columns: ProjectTitle, researcher, OutputTitle, OutputType, OutputYear, OutputVolume, OutputNumber, OutputPages, doi, authors, abstract, group, idx

(Note that Group 8 has all invalid Project ID and invalid OutputBiblio)

All Metadata:

- Extract unique projects (drop duplicated projects) based on columns Proj ID, Status, Title, RDC, Start Year, End Year, PI in all_metadata, save these information in project data

After cleaning all group output data, we construct a set consisting of all columns across all 8 groups of data. We add the column that does not exist in the group data with values of NaN to prepare for the concatenation of all 8 groups of data. Then we can concatenate all 8 groups of data since now they all have the same columns.

2.2 Data Cleaning

- **Clean DOI column:** There are some records with incomplete doi (such as 10.1086/701807) compared to other records (such as https://doi.org/10.1086/701807), we will complete the DOI for these records by adding https://doi.org/ to the front
- **Clean source and url columns:** There are multiple sources across all groups of data, but URLs only exist when the source is arXiv. So we can safely drop the source column and rename the column URL as the column arxiv_url
- **Clean authors and researcher columns:** We combine the column authors and researcher by using semicolon, split the combined data based on semicolon, and save the set as the column authors_all
- **Clean acknowledgement evidence columns:** We update acknowledgments column based on fsrdc_acknowledgments_evidence since fsrdc_acknowledgments_evidence column comes from another group of data than where the column acknowledgement comes from
- **Clean dataset evidence columns:** We update dataset_mentions column based on columns dataset mentions, fsrdc_data_sources_evidence, and matched_dataset_terms since these three columns come from three different groups of data than the column dataset_mentions
- **Clean disclosure evidence column:** We update rdc_mentions column based on the column fsrdc_rdc_locations_evidence since these two columns come from different groups of data
- **Clean Project columns:** We verified that all non-null project IDs are actually valid project IDs that exist in the all_metatdata.
  - **Clean records without project ID:** There are 287 records with invalid project titles for the corresponding project ID, which means these records are incorrect. We reset the project information of all these records to be NaN. We also verified that all RDC values for these records are valid, and all records with a start year also have an end year and corresponding RDC value.
  - **Clean records with project ID:** We populate the project information for all these records with Project ID by merging with project data extracted from the 2024 research output on the column project ID. Use project information in all_metadate to override existing project data for these records.

2.3 Data Deduplication

Across all groups of data, only group 1, 2, and 6 has evidence reported that determines whether the research output is a valid FSRDC output. Among these three groups, Group 2 has the most meaningful evidence provided instead of just having several boolean variables indicating each of the criteria (acknowledgement, dataset, disclosure, and RDC), followed by Group 1 and Group 6. Therefore, if certain research output is captured by all three groups, we will prioritize to keep records in Group 2 than Group 1 then Group 6 than any other groups. At this stage, we will keep all unique records in Group 2 no matter if it has DOI or not, but we only keep records with DOI for all other groups since Group 2 has the most complete records even though they don't have DOIs, we can still validate these records in some ways. So our deduplication process needs to be done in the following order:

1. Deduplicate Group 2 Research Outputs: For the records with DOI, we deduplicate based on the DOI column, we also keep records without DOI, and we deduplicate these records based on the output titles

2. Deduplicate Group 1 Research Outputs: For the records with DOI, we deduplicate based on the DOI column, we only keep records with unique DOIs that are not any DOIs in the previous step

3. Deduplicate Group 6 Research Outputs: For the records with DOI, we deduplicate based on the DOI column, we only keep records with unique DOIs that are not captured by the previous step

By concatenating the new results and the result from the previous step, we will get the deduplicated data at the end.

2.4 Project Information Matching

After getting the deduplicated data, we need to populate the project information for all the records in the duplicated data. We have already populated the project information for the records with Project ID, so we only need to deal with records that do not have a Project ID. Even though these records do not have Project IDs but some records have RDC (2068 records), start year (22 records), end year (22 records), and PI information (12982 records) that could be used to match a project:

1. **Match based on PI:** If the PI listed for that record has only one project according to project data, then this record definitely belongs to this listed project (6512 records left)

2. **Match based on PI and RDC:** If the PI listed for that record has only one project in the specific RDC, then this record definitely belongs to this listed project (6510 records left)

3. **Fuzzy Matching:** If a certain PI has more than one projects after filtering by other available project information, then we need to use fuzzy matching to calculate the similarity score to decide the best matches (Use threshold of 80 unless specified).

    (a) **Match based on RDC, Years, and PI:** We filter the project based on RDC, Years, and PI (if these information is available) as candidates to match, if there are more than one project listed, we calculated the similarity score based on project and output title, we pick the project with the highest similarity score as the matched project (6458 records left, all records only have RDC and PI information now)

    (b) **Match based on RDC, and authors:** We filter the project based on RDC, and calculate the similarity score between authors and PI, we pick the PI of the highest similarity score, and use that PI's project(s) as the filtered projects. If the PI has more than one projects, we calculate the similarity score based on the project and output title, and we pick the project with the highest similarity score as the matched project (5661 records left)

    (c) **Match based on authors:** We calculate the similarity score between authors and PI, we pick the PI of the highest similarity score, and use that PI's project(s) as the filtered projects. If the PI has more than one projects, we calculate the similarity score based on the project and output title, and we pick the project with the highest similarity score as the matched project (5523 records left)

(d) **Match based on titles with threshold:** We calculate the similarity score based on the project and output title with the use of threshold of 80, and we pick the project with the highest similarity score as the matched project (5523 records left)

(e) **Match based on titles without threshold:** We calculate the similarity score based on the project and output title without the use of threshold of 80, and we pick the project with the highest similarity score as the matched project (0 records left)

After this step, we populated all project information for the deduplicated data, saved in both populated_data.csv and populated_data.pkl files

2.5 API Data Enrichment and Validation

To expand and enrich our dataset of publications using Federal Statistical Research Data Centers (FSRDCs), we leveraged the OpenAlex API to search approximately 18,000 papers. Where available, we queried OpenAlex by DOI; in the absence of a DOI, we used the article title. For each identified paper, we retrieved associated metadata including the abstract, keywords, journal name, publication date, volume, issue, page numbers, and author list.

The first stage of processing involved validating papers by searching titles and abstracts for over 300 dataset-specific terms as well as common FSRDC-related references. This yielded approximately 1,600 papers that referenced one or more of 40 distinct datasets. Due to computational constraints, we were unable to search full text for all of the terms in our dataset. However, in a second phase, we narrowed our term list to those 40 dataset terms identified in the first round and applied full-text search across the full set of 18,000 papers to ensure we captured the entire potential set of FSRDC outputs. This more computationally intensive approach recovered roughly 80 additional papers.

Due to the careful application of terms related specifically to FSRDC outputs, we are confident that the resulting set of just under 1,700 papers constitutes a valid corpus of FSRDC-related scholarship.[2] All records have been enriched with available bibliographic metadata to the extent possible, including publication venue and structured citation information.

3 Descriptive

We use the compiled verified dataset *'output_matches_new.csv'*. As a first step we identify and remove duplicates based on whether the listed research output has the same title, publication date, bibliography, abstract, type of publication (e.g. journal article, working paper etc.), output venue (e.g. journal name etc), authors and project RDC. Additionally, as the Federal Statistical Research Data Centers (FSRDC) program started in 1989, we remove any output that is published before that.[3] We also identify duplicates in the FSRDC-verified Projects 2024 , which is accessible via: **FSRDC-verified Projects** based on the title, type of publication, authors, project RDC, and output venue. Table 1 shows the size of the datasets after removing duplicates.

Figure 1 presents the annual distribution of FSRDC research outputs compiled by all groups and verified using the methodology outlined in Section 2. The number of research outputs steadily increases from 2000, reaching a peak in 2018 with 126 outputs. Our dataset also includes research published through 2025, with 7 outputs recorded for that year. Based on our dataset the top 10 RDCs are geographically distributed across the U.S with the largest contributor being Boston with

[2]We acknowledge that these outputs may include papers that are simply FSRDC-related rather than reliant on exclusive FSRDC data.

[3]We have 18 cases with output year is before 1989

Dataset	No. of Res. Out.	Final No. of Res. Out.	Used for Analysis
output_matches_new.csv	1681	1662	Yes
FSRDC-verified Projects 2024	1735	1734	Yes

Table 1: Details of Datasets Used

169 research outputs (Figure 2). The second is Washington (163 research outputs) and Michigan (145 research outputs) while the research Triangle (143 research outputs) is the fourth.

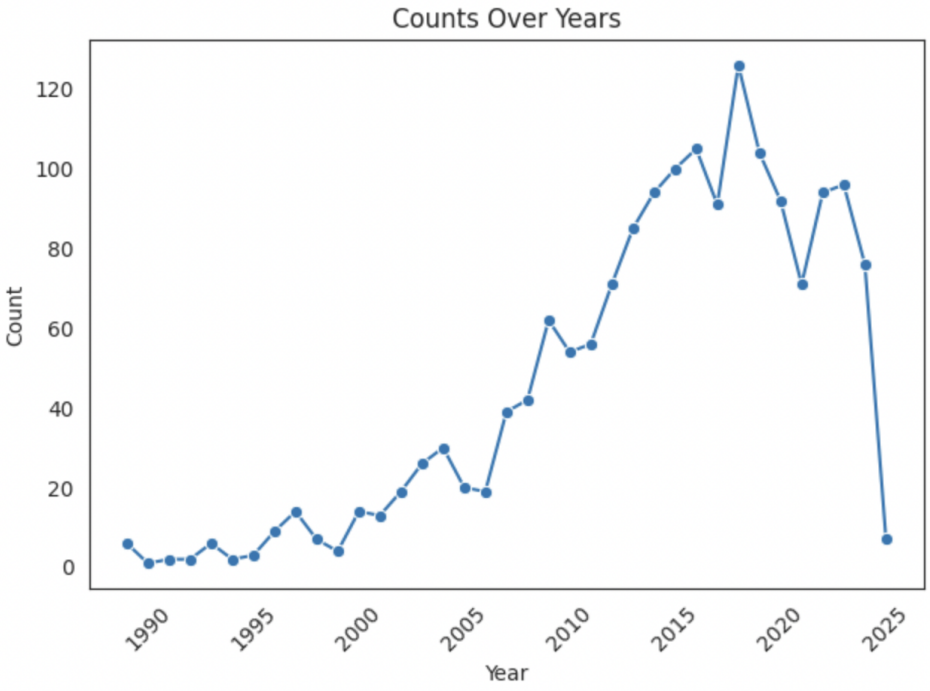


Figure 1: This figure shows the yearly count of research outputs in the output_matches_new.csv dataset.

In order to identify the most prolific authors, we employ two methods. Firstly, using all authors who are part of a research output, we create a author-level dataset where each row corresponds to an author part of particular research output. Using this author level dataset we create a count for each author counting the frequency of author name occurrence in our dataset. This tells us how many research outputs are linked to a particular author. Figure 3 plots the top-10 authors with the y axis denoting author names and x axis denoting count. Scott H. Holan is the author with the most research outputs in our dataset i.e. 40. and John Haltiwanger being the second. However, this methodology does not account for the impact factor of each research count. To incorporate that we calculate the H-index for each author. The H-index is a a measure that reflects a researcher's or scholar's influence and output by considering both their publications and the number of citations those works receive. Figure 4 plots the top ten authors with the highest H-index. Javier Miranda has the highest H-index being the most influential author in our dataset with William R. Kerr being the next. It is important to note that the order of 'prolific' authors identified here is different from the ones identified using a simple count.

We also analyze how the cite count varies since day of publication. Figure 5 plots this with the

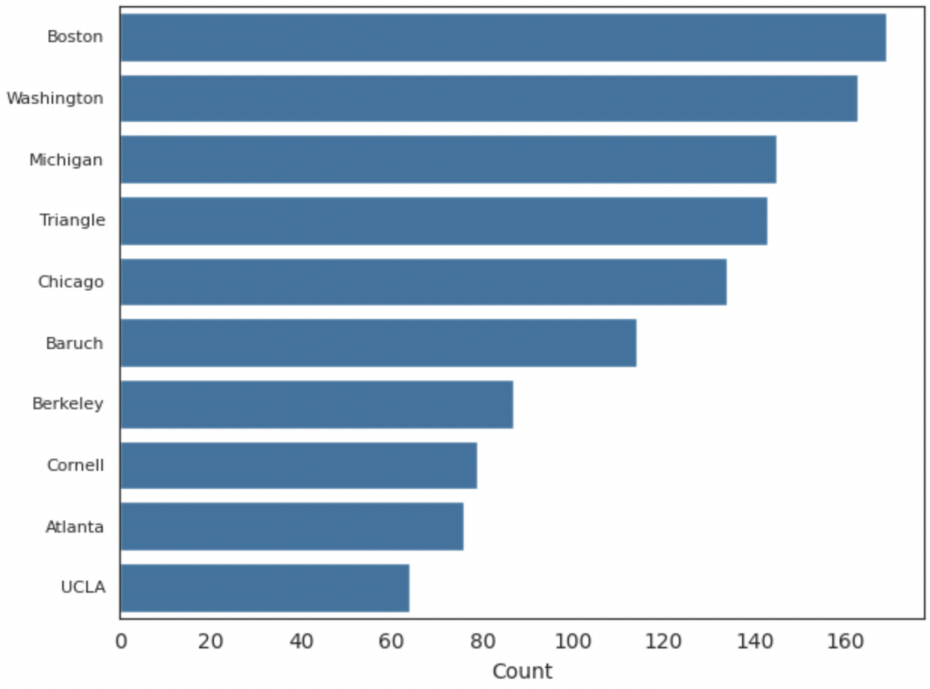


Figure 2: This figure shows the top 10 RDCs

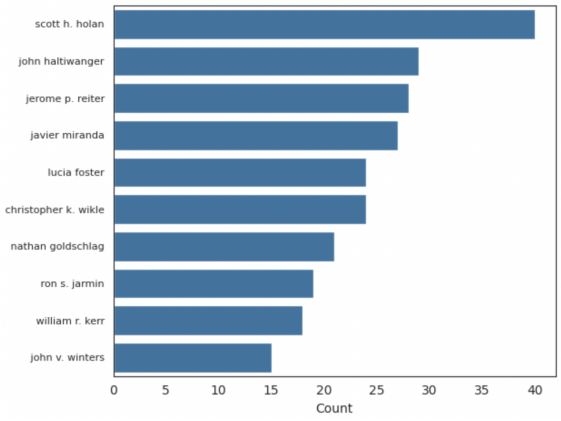


Figure 3: This figure shows the top 10 authors based on count of Research Outputs

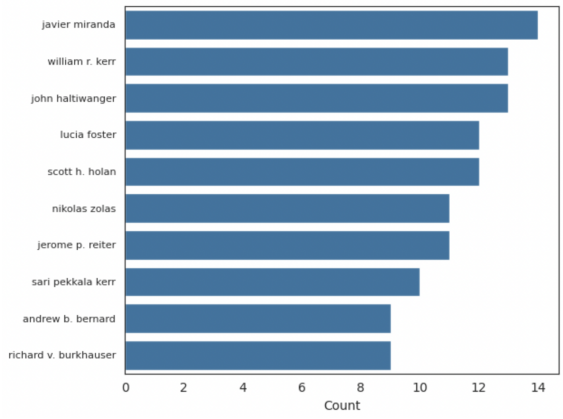


Figure 4: This figure shows the top 10 authors based on H-Index

x axis plotting days since publication[4] and y axis the cite count for each research output. Each dot therefore represents a research output. The plot does not show a clear linear relationship between cite count and days since publication. We explore how cite count varies across RDCs to identify the RDCs with the highest cite counts on average (Figure 6). Penn State and Yale on average have research outputs with the highest cite count. Interestingly, these RDCs are not the RDCs with the highest number of research outputs in our dataset. Additionally, on average the highest cite count is from output type classified as 'Other Publication' which includes reviews, letters etc and the lowest being from Graduate Research Outputs (Appendix A Table 1). We also plot how the average cite count by output type changes with days since publication (Appendix A Figure 1). We find that average cite count spikes especially for journal article publications as days since publication increase. We find no similar consistent pattern for other types of publications.

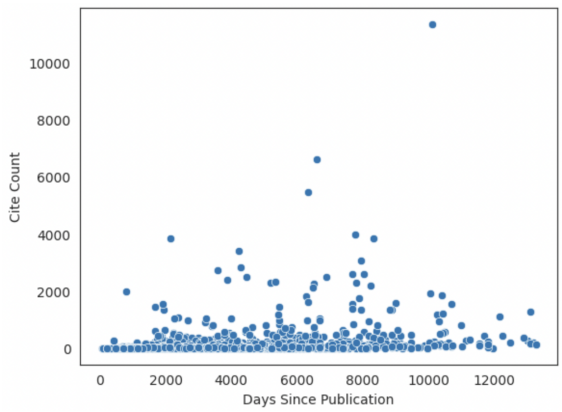


Figure 5: Cite Counts and Days Since Publication

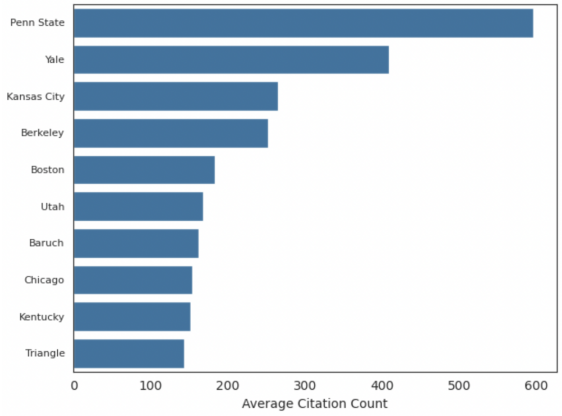


Figure 6: Top 10 RDCs with Highest Average Cite Count

Finally, we analyze which projects have been the most successful in driving citations. While total number of publications can be a useful metric, in many cases, whether or not an output is highly cited is more indicative of impactful work than the number of works published. We find that the project title "The Long-Run Determinants of Social, Demographic, and Economic Characteristics and Processes" led by PI Bryan A. Stuart has by far the most total citations with over 119,000 citations across all outputs. Table 2 shows the results of this analysis.

Table 2: Top 5 Projects by Total Citation Count

Project Title	Total Citations
The Long-Run Determinants of Social, Demographic,...	119,123
Organizations in the Digital Economy: Information...	84,724
Firm Organization Across Space	73,050
Innovation and Market Concentration	55,858
Establishment Human Resource Practices	50,390

[4]Reference day is 2025-05-06

4 Analysis

We conduct two primary analyses. First, we classify whether each research output is a journal article by leveraging a combined dataset consisting of FSRDC-verified Projects 2024 outputs and the *output_matches_new.csv* dataset. We construct a binary variable, 'JournalArticle', which equals 1 if the research output is a journal article and 0 otherwise. Given the binary nature of the dependent variable, we apply a regularized logistic regression model using the default regularization parameter.

The model includes the following characteristics: (1) the length of the title of the research output, based on the assumption that the journal articles usually adhere to the word limits set by the journals;[5] (2) the duration between the project start date and the publication date, under the assumption that the journal articles generally have longer publication timelines; and (3) the associated RDC of the project, encoded using one-hot encoding to generate binary indicators for each RDC location. There are 674 cases where the project start date is less than publication date. 644 are part of the new compiled dataset while 30 are part of the FSRDC-verified Projects 2024 dataset. We consider these as data collection errors and remove them from our classification analysis. We classify each research output as a 'Journal Article' if the probability is greater than 50 % as our classes are balanced (48 % of our research outputs are journal articles). Our classifier has an accuracy of 67.37 %. We also use the ROC curve to evaluate our model. Figure 7 shows the ROC curve for our model. The AUC score is 0.73 which means that we have a 73% chance that the model will correctly rank a randomly chosen positive instance higher than a randomly chosen negative one.

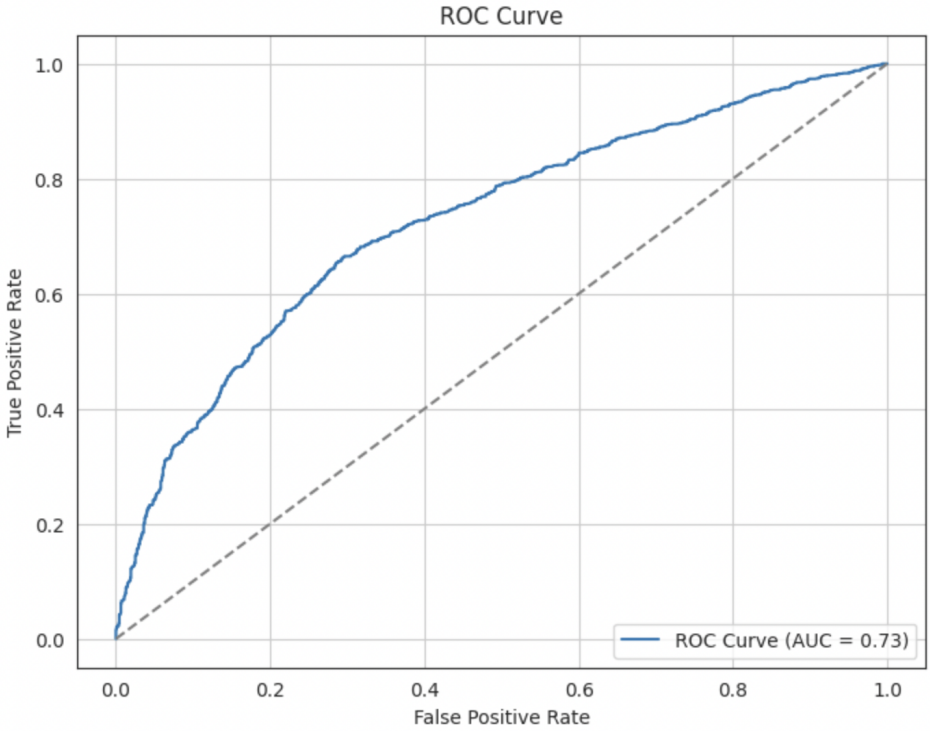


Figure 7: This figure shows the ROC Curve for our logistic classifier

In our second analysis we try to understand the dynamics behind cite counts. We enrich the FSRDC-verified Projects 2024 dataset with cite counts through OpenAlex API search. We combine

[5]This varies by journal

cite counts for both the new datasets and the FSRDC-verified Projects 2024 dataset to create our final dataset with 3396 observations. We create features that can help us understand which outputs are more likely to have a higher cite count. Based on our topic analysis in Project 2, one of the topic classifications of research outputs was "Human Capital, Labor Force and Productivity". Using this knowledge, we classify all outputs part of our dataset into whether the papers are about labor economics. We employ an NLP model i.e. T5, FLAN-T5 locally classifying all output titles into whether they are about labor economics or not. Only 1.1 % of our research outputs are about labor economics. Secondly, we create an indicator variable equal to '1' if the research output is part of a top 10 RDC i.e. Boston, Washington, Michigan, Triangle, Chicago, Baruch, Berkley, Cornell, Atlanta and UCLA, and is '0' otherwise. We also create a variable 'Years Since Publication' to count the years since the output was published till date. Additionally, we also use length of title and whether the output was a journal article.

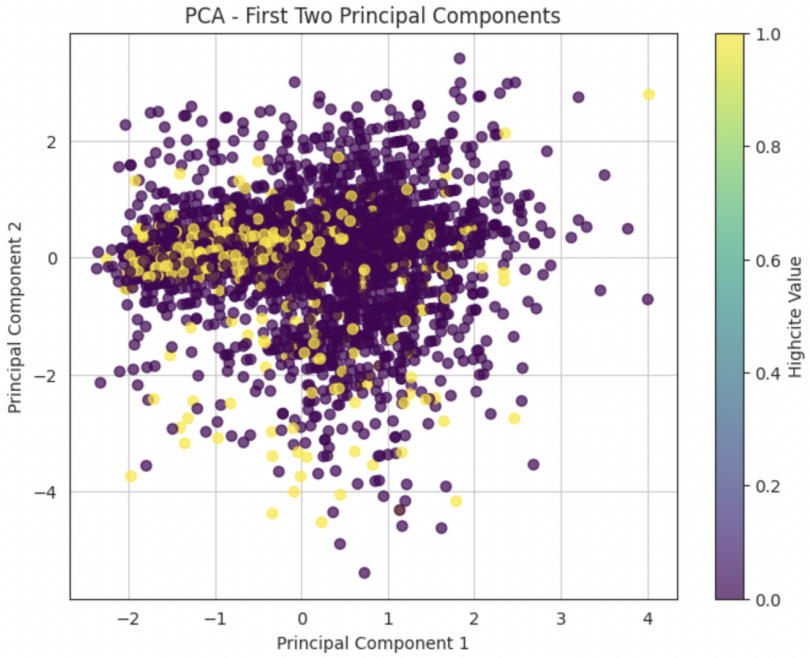


Figure 8: Clustering Identifies Little Delineation in PCs

As a first step, to reduce the dimensionality, we perform PCA combining whether the output is a journal article, years since start of project and length of title. As the point of dimension reduction is to reduce the number of inputs, we estimate two Principal components. The two principal components however do not explain a lot of the variance in the dataset. Principal component 1 explains 39.07 % of the variance while Principal component 2 explains 33.60 %. We also plot how the two PCs to see if any clusters any be identified based on cite counts. We classify research outputs as having 'high' cite count if their cite count is larger than the mean in the dataset i.e. 420. Unfortunately, no discernible clusters and patterns can be highlighted (Figure 8). We therefore, decide not to use the principal components in our analysis.

Our first OLS model uses a simple linear regression to estimate the relationship between cite count and our features. Before we estimate our model we plot the distribution of cite count. We find that the distribution is right-skewed and therefore we add a unit and log our cite count before analysis (Figure 9 and 10). Table 3 shows our results. Our model does not perform well as it only explains 5% of the variance. However, it is important to note that we find some significant

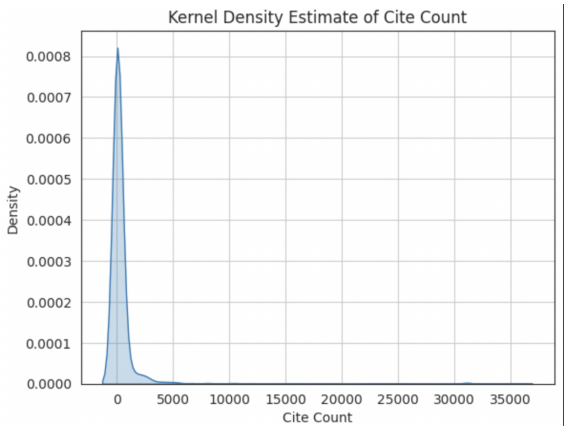


Figure 9: Citation Counts are Clustered near Zero, but with a Long Tail

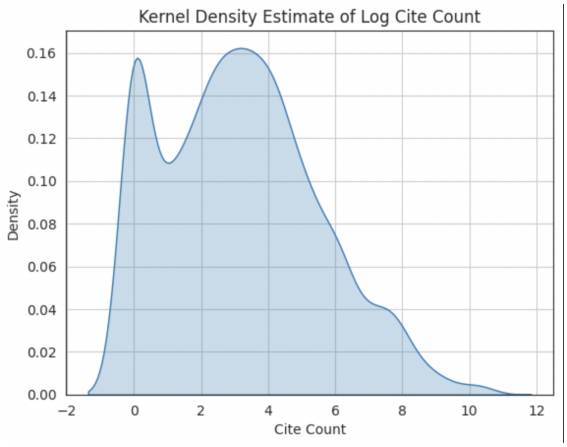


Figure 10: Logged Counts Show Many Works with Zero Citations

relationship in our features and logged cite count. For example, each additional year since publication is associated with approximately 8.3% more citations regardless of the output type, RDC and topic of output. Interestingly, holding everything else constant, each additional word in the title is associated with approximately 0.82% fewer citations.

Table 3: OLS Regression Results

Variable	**Coef.**	**Std. Err.**	**t**	**P>t**	**[0.025, 0.975]**
Intercept	3.3400	0.133	25.052	0.000	[3.079, 3.601]
top10	-0.2799	0.086	-3.263	0.001	[-0.448, -0.112]
YearSincePublication	0.0797	0.006	12.343	0.000	[0.067, 0.092]
TitleLen	-0.0082	0.001	-6.486	0.000	[-0.011, -0.006]
JournalArticle	-0.1564	0.080	-1.959	0.050	[-0.313, 0.000]
Economics_Labor_dummy	-0.1049	0.495	-0.212	0.832	[-1.075, 0.865]
JournalArticle:Economics_Labor_dummy	0.5093	0.752	0.677	0.498	[-0.966, 1.984]

Observations: 3396
R-squared: 0.056 *Adj. R-squared*: 0.054
F-statistic: 33.61 *Prob (F-statistic)*: 1.36e-39
AIC: 15150 *BIC*: 15200
Durbin-Watson: 1.645
Notes: Standard errors assume correctly specified errors. Large condition number (1.81e+03) may indicate multicollinearity. Interpret results with caution.

For our final analysis, we utilize BERTopic to attempt to extract meaningful information from the keywords we extracted from OpenAlex for each research output. We began by applying simple clustering analysis to the sparse data matrix generated by creating columns to indicate keyword presence, but the high sparsity of the resulting matrix means that traditional dimensionality reduction tools were ineffective at reducing the dimensionality of keywords. We tried both PCA and Truncated SVD, the graphs of which are in the appendix and neither succeeded in meaningful reductions of dimensionality. We also attempted to interpret the outputs with respect to location as a potential way to engender meaningful information without success. Instead we turned to BERTopic as a way to cluster our keywords into topics. We settled on 10 distinct topics, which

we then applied human coding to for interpretability. The topics and their respective counts are displayed in Table 4.

Table 4: Distribution of Outputs Across BERTopic-Inferred Topics

Topic	Number of Outputs
Urban Studies	1,403
Firms	1,104
General Economics	418
Healthcare	226
Data Estimation	125
None	42
Education	39
Environment	28
Crime	19
Community Development	11

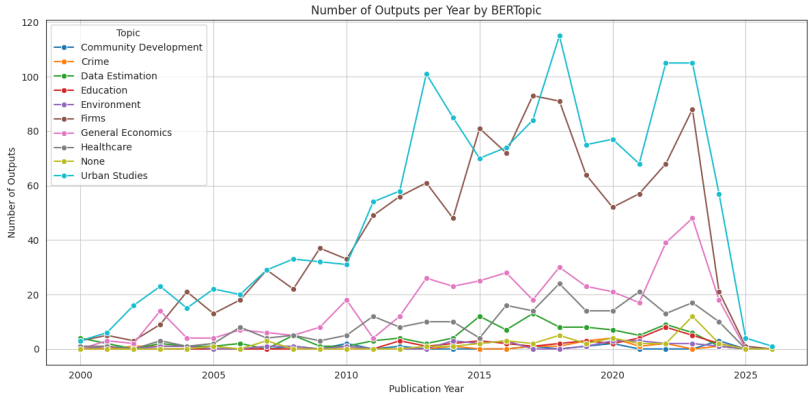


Figure 11: Outputs on Urban Studies, Firms, and General Economics Are Increasing Over Time

Figure 11 shows the trend of these topics over time. We can see that many topics have essentially seen no increase in publication rates since the year 2000, however, three topics (Urban Studies, Firms, and General Economics) have significantly increased in population over time. To study the impact of topic on citation counts, we once again employed OLS. We estimate the impact of a paper belonging to a specific topic on the log of citation counts. We control for Top 10 RDC and Years since publication and apply robust standard errors. This regression shows that there is not a strong correlation between any of our topics and citation count. In this analysis, we find that only Years Since Publication and Title Length have an impact on citation counts, with Years Since Publication having a roughly 5.1% increase in the number of citations for each year that passes and Title Length having a roughly -.3% decrease in citations for each additional character in the title.

Table 5: OLS Regression Predicting Log Citations from BERTopic Categories and Controls

	coef	std err	z	P> |z|	[0.025	0.975]
Intercept	3.0504	0.860	3.548	0.000	1.365	4.736
Crime	0.1711	0.922	0.186	0.853	-1.636	1.978
Data Estimation	-0.5055	0.875	-0.578	0.563	-2.220	1.209
Education	1.3112	1.076	1.218	0.223	-0.798	3.421
Environment	1.4264	0.975	1.463	0.143	-0.485	3.337
Firms	0.4010	0.859	0.467	0.640	-1.282	2.084
General Economics	0.6126	0.866	0.708	0.479	-1.084	2.310
Healthcare	0.3469	0.870	0.399	0.690	-1.358	2.052
None	-0.9911	0.986	-1.005	0.315	-2.923	0.941
Urban Studies	0.0925	0.859	0.108	0.914	-1.591	1.775
YearsSincePublication	0.0507	0.006	8.721	0.000	0.039	0.062
TitleLen	-0.0028	0.001	-2.026	0.043	-0.005	-0.0009
top10	-0.1181	0.087	-1.350	0.177	-0.290	0.053

R-squared:	0.038
Adjusted R-squared:	0.035
F-statistic:	11.17
Prob (F-statistic):	2.91e-22
No. Observations:	3,355

Notes: Standard errors are heteroskedasticity-robust (HC3).

5 Conclusion

This project underscores the challenges of tracing the scholarly footprints of FSRDC-based studies. From an initial prospective set of matches of nearly 18,000 outputs, our validation efforts via methodical data collection and rigorous text analysis narrow that set down to just under 1,700. By developing a pipeline that utilizes careful API retrieval combined with advanced pattern matching, we compiled 1,662 newly confirmed research outputs, thereby extending existing datasets and offering a richer foundation for future investigations. We utilize careful matching techniques to assign these outputs to the appropriate FSRDC project to allow for careful comparison across multiple features of the dataset. In addition, our regression results emphasize the relevance of publication timing while uncovering surprising findings related to the negative impact of title length - though many nuances, such as journal prestige and co-author networks, remain to be fully explored.

Despite these advances, our findings also suggest that a substantial volume of yet-undiscovered FSRDC research remains hidden across other repositories and databases. To a large extent, our initial efforts have been only able to uncover published works. However, many of the outputs originally reported as FSRDC-related are working papers, conference papers, or presentations, all of which are hard to find via traditional scholarly APIs.

There are several limitations to our approach. First, our reliance on OpenAlex and similar bibliometric APIs restricts our reach to outputs indexed in these databases, potentially missing grey literature or institutionally archived works. Additionally, some valid FSRDC outputs may fail to explicitly acknowledge data restrictions, RDC locations, or disclosure reviews, thus eluding our pattern-matching methods. While we employed an extensive set of search terms, the possibility

of false negatives remains, particularly for older or less standard publications. Furthermore, the absence of full-text search capabilities across all outputs likely constrained the identification of nuanced references to FSRDC data.

Future research can build on this work by expanding the scope of data sources, including integrating metadata from institutional repositories, conference proceedings, and preprint servers beyond arXiv. Incorporating machine learning approaches for classifying FSRDC-related content based on full-text content, when available, may also enhance recall. A more comprehensive integration with Census internal records could help identify valid outputs that lack formal acknowledgments or citations. Additionally, deeper bibliometric analysis on collaboration networks and institutional ties could reveal the structural dynamics of FSRDC-supported research.

For FSRDC stakeholders and policymakers, our results underscore the importance of standardized acknowledgment practices and metadata quality. Encouraging consistent citation of FSRDC datasets and disclosure statements would facilitate easier tracking and impact measurement. The findings also support the case for developing an internal FSRDC publication registry that researchers can opt into at the point of data approval. Such a system could substantially improve the discoverability, validation, and policy relevance of FSRDC research outputs, ultimately helping agencies demonstrate the public value of secure microdata access.

6 Appendix

Output Type	Mean Cite Count
Other Publication	340.636364
Journal Article Publication	140.517493
Book	127.04651
Working Paper	70.984211
Dataset	58.666667
Graduate Research Output	0.500000

Table 1

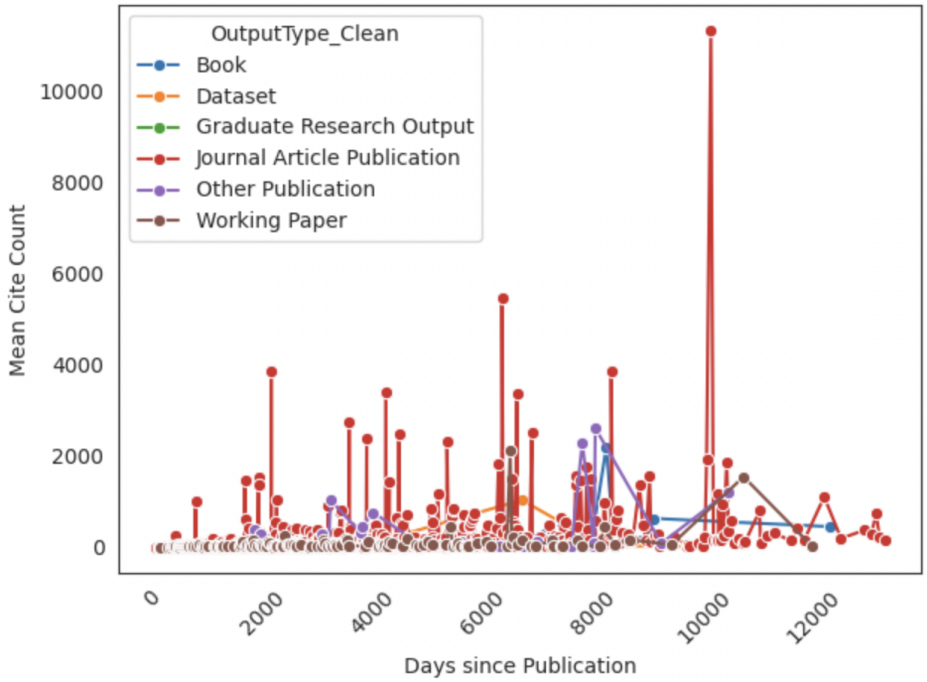
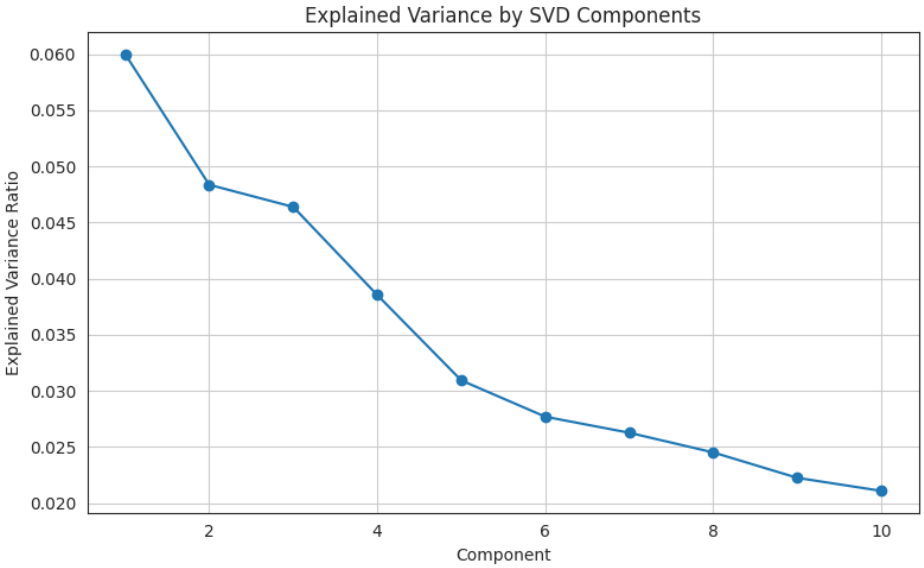
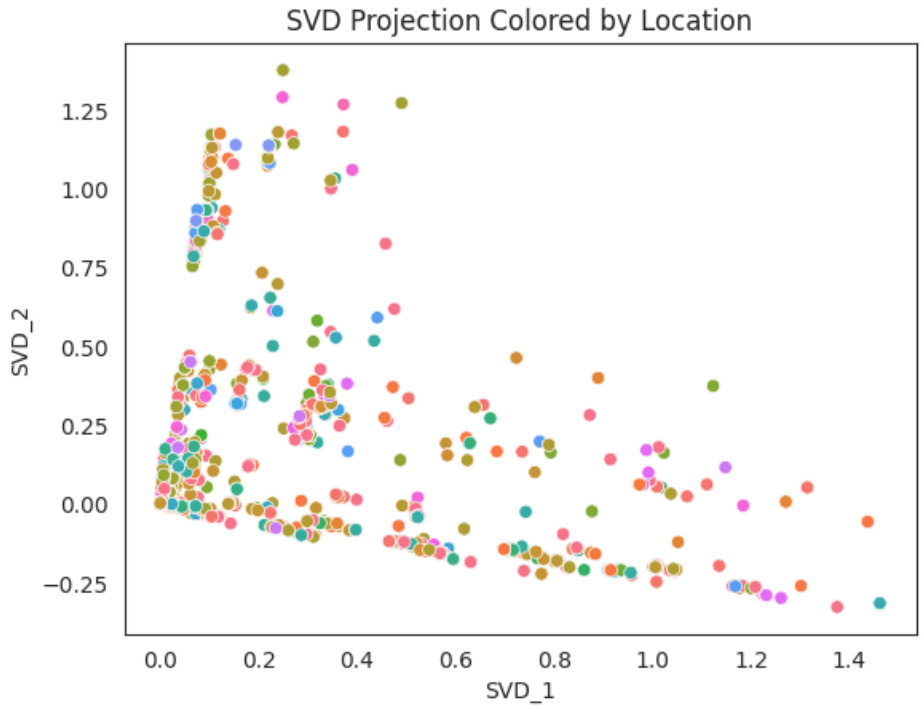


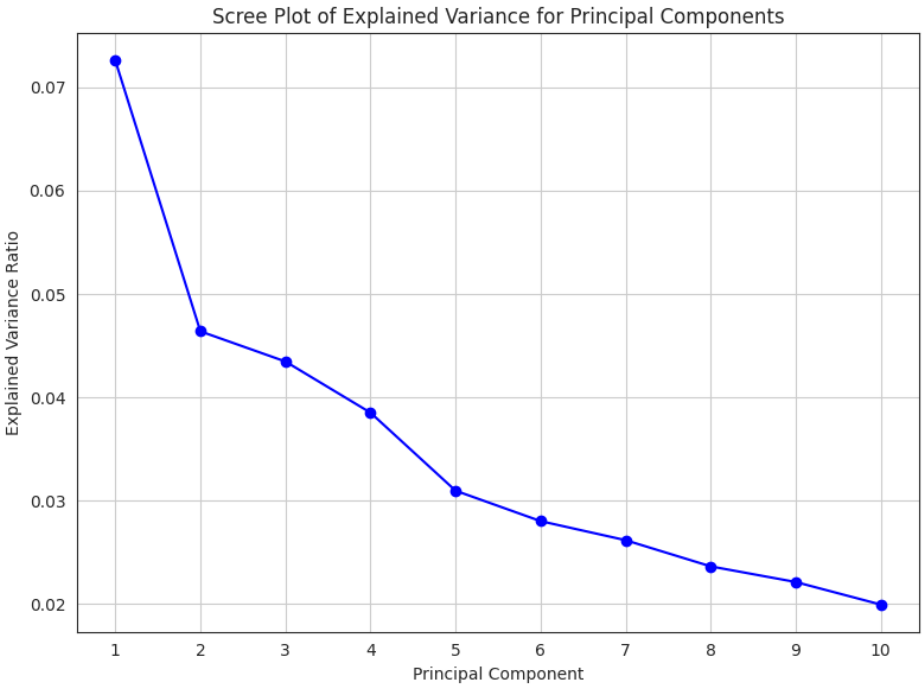
Figure 1



Appendix Figure 2



Appendix Figure 3



Appendix Figure 4