IS 525: Data Warehouse and BI

Professor Michael Wonderlich

Team Members:

Arundhati Raj (raj9)

Anveksha Vinod Pandey(avp6)

Jeet P Thakore (bhapkar2)

Mihika Bodke (mbodke2)

Yue Lei (yuelei4)

Last Updated : December 16th, 2024

Abstract

The scenario for this project is focused on developing an interactive, data-driven dashboard to explore two key datasets from the Author-ity 2009 collection. The goal is to offer comprehensive visual insights into the research and innovation landscape, particularly related to NIH/NSF grants, academic publications, patents, and researcher demographics.

Problem Statement

The main problem this project aims to solve is to provide a unified research analytics platform that can illuminate hidden patterns in research funding, institutional collaboration, and demographic representation across the academic sector. The dashboards will serve multiple stakeholders, including policymakers, academic institutions, and funding agencies, by delivering actionable insights to support data-driven decision making.

Intended Audience

The primary intended audience for this project includes:

- Policymakers - To provide evidence-based insights that can inform program development and resource allocation decisions
- Academic Institutions - To enable performance assessment and strategic planning through data-driven analytics
- Funding Agencies - To enhance portfolio analysis and impact measurement capabilities

Data Sources

The project utilizes two key datasets from the Author-ity 2009 collection:

- https://databank.illinois.edu/datasets/IDB-4370459

Dataset connecting researchers to records from NIH/NSF grants, USPTO patents, and academic publications

- https://databank.illinois.edu/datasets/IDB-9087546

Dataset adding demographic information, including ethnicity and gender predictions, for the researchers and inventors

- https://drive.google.com/drive/u/1/folders/1Wpq7Eks-mZ5hAtNYVYLNwRx2UXOUCI8D

All involved original and derived datasets, all packaged workbooks for 5 dashboards for this project.

Client And Perspective

There is no specific client for this project. The project is an independent effort to develop a comprehensive, interactive visualization platform that can serve the needs of multiple stakeholders in the academic ecosystem, including policymakers, academic institutions, and funding agencies. The goal is to enhance the understanding of the academic research and innovation landscape by providing data-driven insights.

Analysis and Discoveries

Some of the key analysis and discoveries made through this project include:

- Temporal trends in grant distributions, publication activities, and patent filings across different age groups and institutions
- Institutional and geographic analysis to identify key centers of research and innovation
- Demographic insights into the ethnic and gender diversity among researchers and inventors
- Relationship between principal investigator's age and award probability

Methods & Analysis

- Data Collection: The primary dataset were sourced from different datasets namely authorlink_nih.csv, authorlink_nsf.csv, authorlink_uspto.csv, uiuc_uspto.csv and genni-ethnea-authority2009.csv supplemented with NIH (National Institutes of Health) and NSF (National Science Foundation) grant data, USPTO patent linkage data,

Inventor data (detailed patent records) and EthnicSeer demographic predictions dataset.

- Data Cleaning and Transformation: Joins were used to produce a unified table. This consolidated dataset enabled efficient analysis and visualization. Missing and null values were removed or transformed to maintain consistency. Outliers were handled to ensure reliability.

Dataset Details

---

1. authorlink_nih.tsv

This dataset contains information about NIH grants and their linkage to authors in the Author-ity 2009 database. The detailed description of all fields in this dataset is shown in the following figure:

Field Name	Description	Data Type
`app_id`	Application ID for the grant.	Number (#)
`nih_full_proj_nbr`	Full project number of the grant.	Text (Abc)
`nih_subproj_nbr`	Sub-project number, if applicable (null for main projects).	Number (#)
`fiscal_year`	The fiscal year of the grant.	Number (#)
`pi_position`	Position of the principal investigator (PI).	Number (#)
`nih_pi_names`	Full names of the principal investigator(s).	Text (Abc)
`org_name`	Name of the organization receiving the grant.	Text (Abc)
`org_city_name`	City of the organization.	Text (Geographic)
`org_bodypolitic_code`	State or region of the organization.	Text (Abc)
`age`	Number of years since the investigator's first paper was published.	Number (#)
`prob`	Probability that the author matches the Author-ity 2009 database (`> 0.5` threshold).	Number (#)
`au_id`	Unique identifier for the author in the Author-ity 2009	Text (Abc)

2. authorlink_nsf.tsv

This dataset contains information about NSF grants and their linkage to authors in the Author-ity 2009 database. The detailed description of all fields in this dataset is shown in the following figure:

Field Name	Description	Data Type
`AwardId`	Unique award ID for the NSF grant.	Number (#)
`fiscal_year`	Fiscal year range of the grant (e.g., `1986–1986`).	Text (Abc)
`pi_position`	Position of the principal investigator (PI).	Number (#)
`PrincipalInvestigators`	Names of the principal investigator(s) (can include multiple names separated by `;`).	Text (Abc)
`Institution`	Name of the institution receiving the grant.	Text (Abc)
`InstitutionCity`	City where the institution is located.	Text (Abc)
`InstitutionState`	State where the institution is located.	Text (Abc)
`age`	Number of years since the investigator's first paper was published.	Number (#)
`prob`	Probability that the author matches the Author-ity 2009 database (`> 0.5` threshold).	Number (#)
`au_id`	Unique identifier for the author in the Author-ity 2009 database.	Text (Abc)

3. authorlink_uspto.tsv

This dataset links authors in the Author-ity 2009 database to inventors in the USPTO (United States Patent and Trademark Office). The detailed description of all fields in this dataset is shown in the following figure:

Field Name	Description	Data Type
`au_id`	Unique identifier for the author in the Author-ity 2009 database.	Text (Abc)
`inv_id`	Unique identifier for the inventor in the USPTO database.	Number (#)
`prob`	Probability that the author matches the inventor in the USPTO database (`> 0.5` threshold).	Number (#)

4. uiuc_uspto.tsv

This dataset contains information about disambiguated inventors in the USPTO database. The detailed description of all fields in this dataset is shown in the following figure:

Field Name	Description	Data Type
`inv_id`	Unique identifier for the inventor in the USPTO database.	Number (#)
`is_lower`	Binary flag indicating whether the inventor's name is lowercase in the database.	Number (#)
`is_upper`	Binary flag indicating whether the inventor's name is uppercase in the database.	Number (#)
`fullnames`	Full name of the inventor.	Text (Abc)
`patents`	List of patents associated with the inventor (separated by `	`).
`first_app_yr`	Year of the inventor's first patent application.	Number (#)
`last_app_yr`	Year of the inventor's last patent application.	Number (#)

5. genni-ethnea-authority2009.tsv

This dataset provides demographic information about authors, including ethnicity and gender predictions. The detailed description of all fields in this dataset is shown in the following figure:

Field Name	Description	Data Type
`auid`	Unique identifier for authors in the Author-ity 2009 database.	Text (Abc)
`name`	Full name of the author, used as input for ethnicity/gender predictions.	Text (Abc)
`EthnicSeer`	Predicted ethnicity from the EthnicSeer tool.	Text (Abc)
`prop`	Confidence score of the EthnicSeer prediction.	Number (#)
`lastname`	Last name of the author (used as input for Ethnea+Genni).	Text (Abc)
`firstname`	First name of the author (used as input for Ethnea+Genni).	Text (Abc)
`Ethnea`	Predicted ethnicity from Ethnea+Genni (detailed, e.g., `ENGLISH`, `SLAV-ENGLISH`).	Text (Abc)
`Genni`	Predicted gender (`M` for male, `F` for female).	Text (Abc)
`SexMac`	Predicted gender using a third-party tool (`female`, `male`, `mostly_female`, etc.).	Text (Abc)
`SSNgender`	Predicted gender based on US Social Security Name data (`F`, `M`, or `-`).	Text (Abc)

Dashboard 1

Problem Statement

This dashboard provides an **interactive overview** of research grant trends using data from **NSF** (National Science Foundation) and **NIH** (National Institutes of Health). It highlights **institutional performance**, demographic patterns of Principal Investigators (PIs), and grant success probabilities across career stages.

Data Source

Original Datasets:
authorlink_nih.tsv, authorlink_nsf.tsv

Derived Datasets:

Merged_file_NIH_NSF.xlsx

Join - Inner Join
NIH and NSF Tables

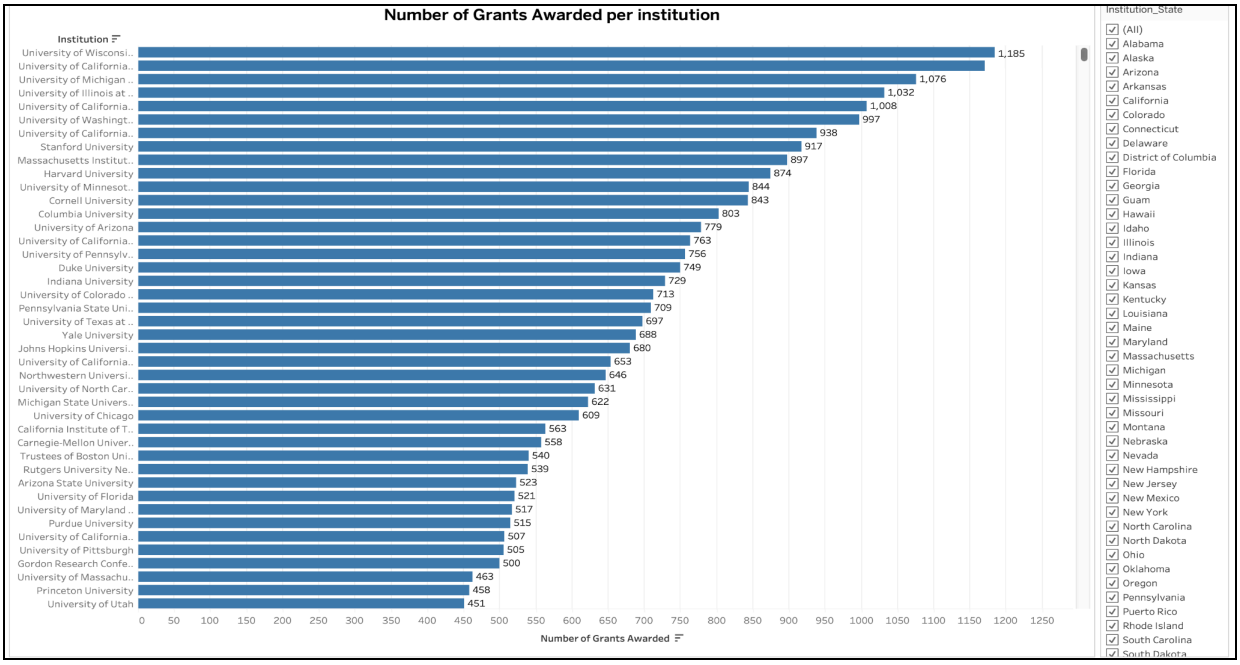
- Join Type: Inner Join to ensure all records are included.
- Join Condition: au_id (NIH) = au_id (NSF).
- Result: A combined table containing grant information from both NIH and NSF.

Challenges Encountered

- **Data Quality Issues:**
  Many fields contained null or inconsistent data.
  Solution: Applied filters to remove null values or replace them with placeholders using calculated fields.
- **Complex Joins:**
  Merging multiple datasets required careful join conditions to avoid data loss.
  Solution: Performed join (inner) to ensure data integrity.
- **Visualization Complexity:**
  Large datasets and multiple fields made creating effective visualizations challenging.
  Solution: Used Tableau's calculated fields and grouping features to simplify the data for clear storytelling.

Visualizations

Visualization 1: Grants per institution



Color Setting:
The bars represent the number of grants awarded per institution in blue.

Interactive Features:

- A filter list on the right allows users to select or deselect specific states to view grant data by state-level institutions.

Tooltips Include:

- Institution name
- Number of grants awarded

Interesting Discoveries from the Visualization

- University of Wisconsin leads with the highest number of grants awarded, totaling 1,185 grants.

- High-performing institutions include University of California campuses, University of Michigan, and University of Illinois. The University of California appears multiple times in the rankings, showing strong performance across its campuses like Berkeley, Los Angeles, and San Diego.

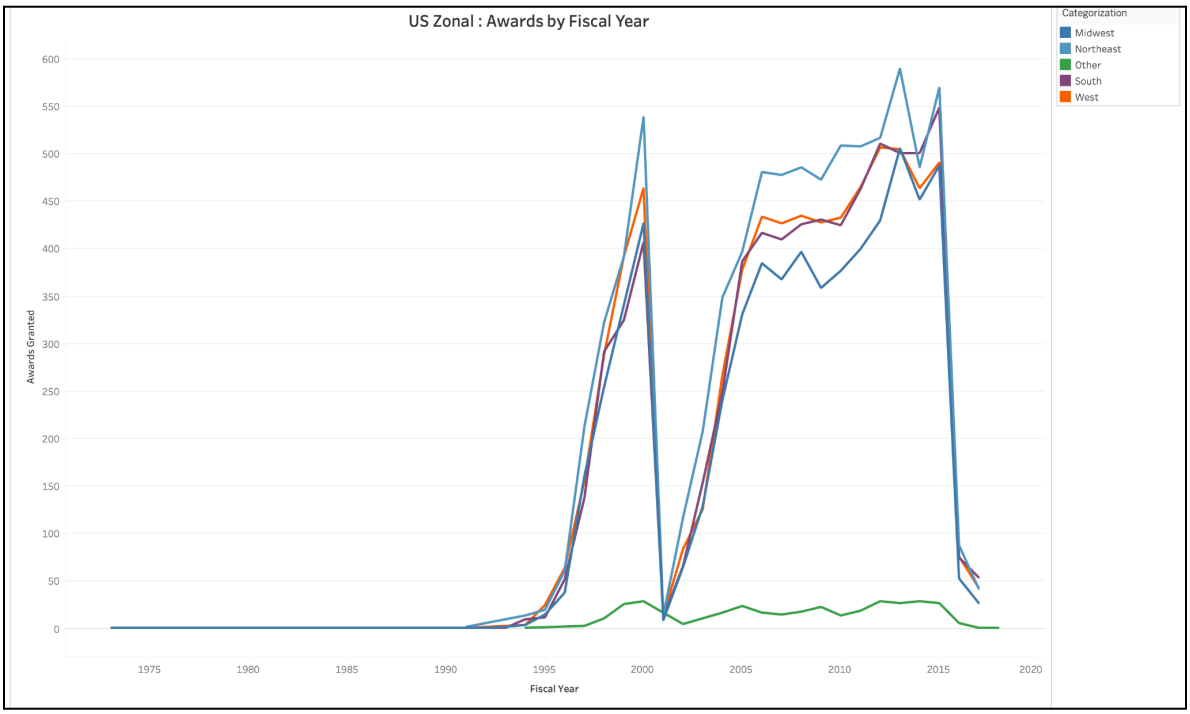
- The top institutions typically receive between 900 and 1,200 grants. Mid-tier institutions, such as Indiana University and Pennsylvania State University, have grant counts ranging from 700 to 800.

- Institutions from California are highly represented, reflecting a strong research ecosystem in the state. Ivy League schools, including Harvard University (874 grants) and Yale University (688 grants), also maintain high numbers of grants.

- Institutions like Princeton University, with 458 grants, and University of Utah, with 451 grants, appear near the lower end of the list among the top institutions.

Visualization 2: Awards by Fiscal Year

Visualization 2 : A time-series analysis showing the trends in grant awards over time across different US zones (Midwest, Northeast, South, West).



Color Setting:

- Blue represents the Midwest.
- Purple represents the Northeast.
- Green represents Other zones.
- Red represents the South.
- Orange represents the West.

Interactive Features:
A categorization legend allows users to focus on specific zones for detailed analysis.

Tooltips Include:

- Fiscal year
- Awards granted
- Zone categorization

Calculated field:

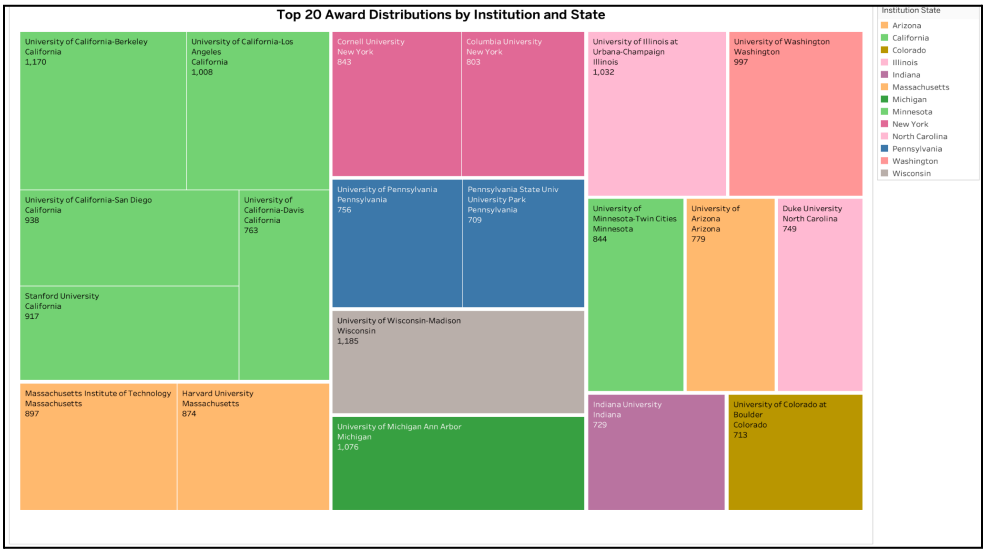
This calculated field has been created using a CASE statement to group states into regions: Northeast, Midwest, South, and West. States were assigned based on their geographic location, and any ungrouped states were labeled as "Other." This helped streamline region-based analysis and visualizations.

Interesting Discoveries from the Visualization

- **Awards Growth Over Time:**
  From 1995 to 2000, there was a **sharp increase** in awards granted across all zones, peaking.
  Dominant Zone:
  - The **Midwest** (blue) shows the highest number of awards consistently, peaking at over **550 awards** granted around 2010.
  - After 2015, the Midwest shows a significant decline in awards.
- **Regional Trends:**
  - The **West** (orange) and **South** (red) follow a similar trend with steady growth after 1995, peaking near **450–500 awards** around 2010.
  - The **Northeast** (purple) also reflects a growth trajectory parallel to the South and West but remains slightly lower.
- **Minimal Awards in Other Zones:**
  The **Other** category (green) shows consistently low awards granted, peaking marginally between 2000 and 2010 with fewer than 50 awards per year.
- **Drop After 2015:**
  There is a steep decline in awards granted across all zones post-2015, especially in the Midwest and West regions.

Visualization 3: Award Distribution by Institution and State

Visualization 3 : A tree map illustrating the geographic distribution of grants across institutions and states, with color coding for easy identification.



Color Setting:
The treemap uses different colors to represent states where the institutions are located:

- Green: California, Michigan, Minnesota
- Pink: Illinois, Washington, North Carolina
- Red: New York
- Orange: Massachusetts, Arizona
- Blue: Pennsylvania
- Purple: Indiana
- Yellow: Colorado
- Grey: Wisconsin

Interactive Features:
The treemap categorizes the top 20 institutions by state and awards granted, offering a clear breakdown of state-level contributions.

Tooltips Include:

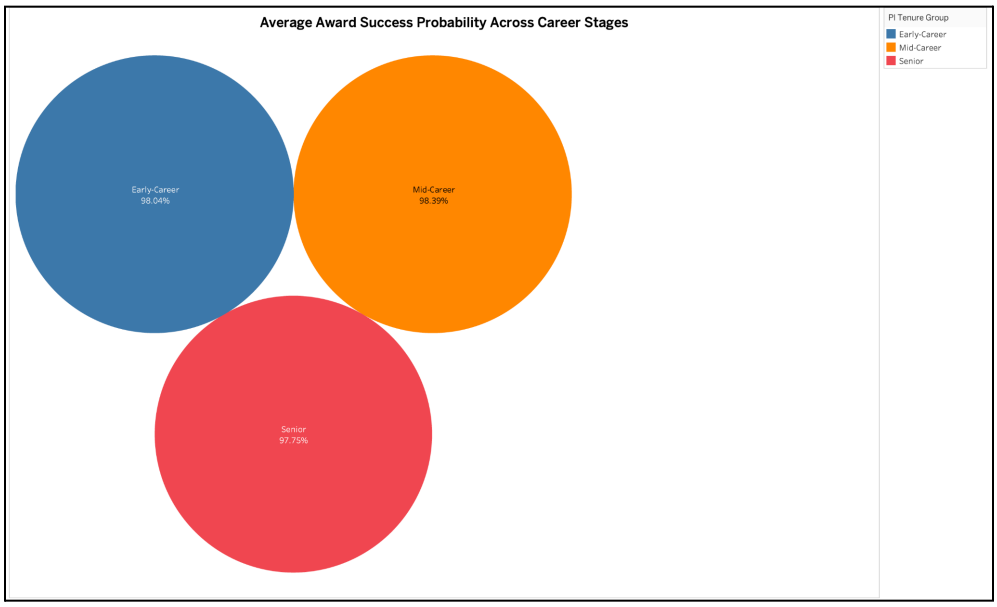
- Institution name
- State
- Awards granted

Interesting Discoveries from the Visualization

- **Top Performing Institution:**
    - The **University of Wisconsin-Madison** (Wisconsin) has the highest number of awards granted with **1,185 awards**.
- **Dominance of California:**
    - California institutions appear most frequently with 6 universities in the top 20.
    - The **University of California-Berkeley** leads among California institutions with **1,170 awards**, followed by **UCLA (1,008)** and **UC-San Diego (938)**.
- **Strong Performance Across States:**
    - **University of Michigan (1,076 awards)** in Michigan and **University of Illinois at Urbana-Champaign (1,032 awards)** are among the leading institutions outside of California.
    - Other high performers include **University of Washington** in Washington (997 awards) and **Stanford University** in California (917 awards).
- **Representation from the Northeast and Midwest:**
    - **Harvard University** (874 awards) and **MIT (897 awards)** represent Massachusetts.
    - New York's **Cornell University (843 awards)** and **Columbia University (803 awards)** also rank in the top 20.
    - Pennsylvania contributes with **University of Pennsylvania (756 awards)** and **Pennsylvania State University (709 awards)**.

- **Smaller Contributions:**
    - Institutions like **Duke University** (749 awards), **Indiana University** (729 awards), and **University of Colorado at Boulder** (713 awards) appear in the lower range of the top 20 list.

Visualization 4: Average Award Success Probability Across Career Stages



Color Setting:

- Blue represents **Early-Career** researchers.
- Orange represents **Mid-Career** researchers.
- Red represents **Senior** researchers.

Interactive Features:
The legend allows filtering by career stages for focused analysis of specific groups.

Tooltips Include:

- Career stage
- Average award success probability

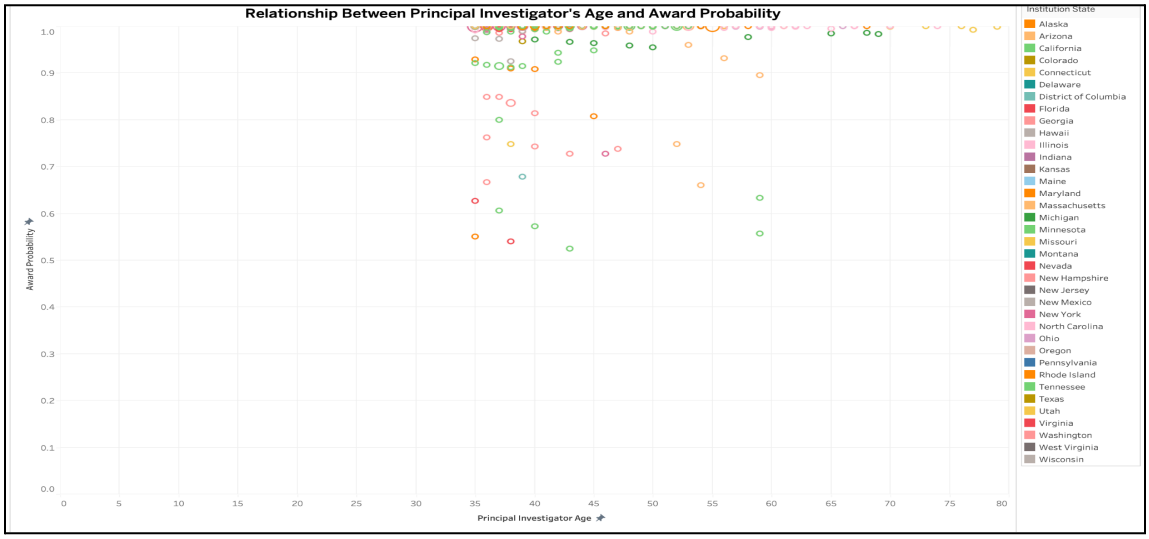
Calculated Field:

A calculated field called PI Tenure Group was defined to classify principal investigators based on their age. Using IF-ELSE IF logic, ages under 30 were labeled as Early-Career, ages between 30 and 45 as Mid-Career, and those above 45 as Senior. This segmentation enables clear analysis across different career stages.

Interesting Discoveries from the Visualization

- **Highest Success Probability:**
    - **Mid-Career** researchers have the highest average award success probability at **98.39%**.
- **Early-Career Success:**
    - **Early-Career** researchers closely follow, with a success probability of **98.04%**, showing strong performance despite being in earlier stages of their careers.
- **Senior Researchers:**
    - **Senior** researchers have the lowest success probability among the three groups, at **97.75%**. Although still high, it is slightly lower compared to Mid-Career and Early-Career stages.
- **Consistency Across Career Stages:**
    - All career stages exhibit success probabilities above **97.5%**, indicating a highly competitive and successful award environment across the board.

Visualization 5: Relationship between PI age and Prob



Color Setting:
Each dot color represents the **Institution State** of the Principal Investigator (PI), with a diverse palette assigned to each state.
Interactive Features:
The legend on the right allows filtering by specific states to analyze the relationship between PI age and award probability for institutions in that state.

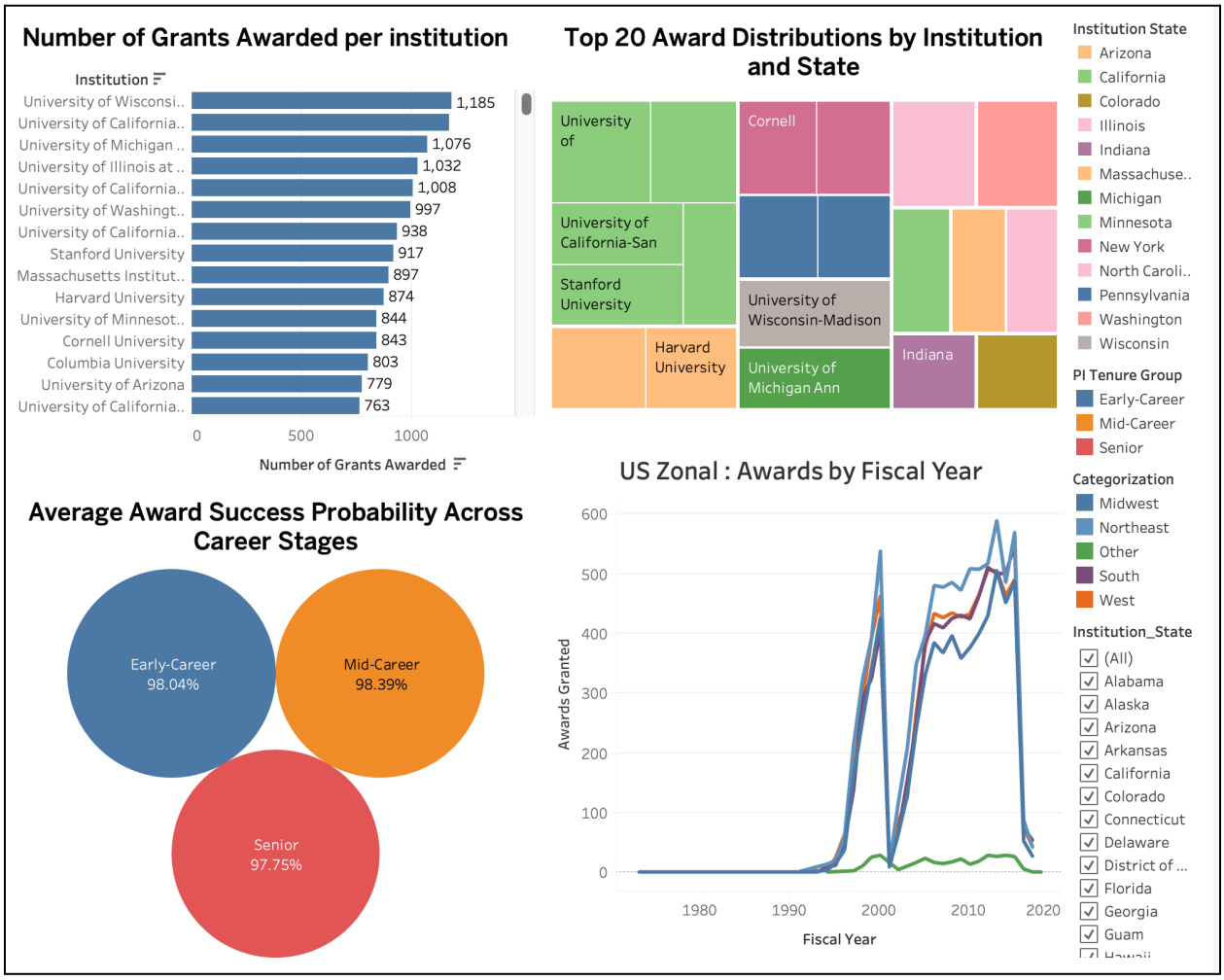
Tooltips Include:

- Principal Investigator Age
- Award Probability
- Institution State

Interesting Discoveries from the Visualization

- **High Award Probability Concentration:**
    - Most **Principal Investigators** exhibit an award probability close to **1.0**, regardless of age, indicating high success rates across ages.
- **Lower Award Probability Clusters:**
    - There is a cluster of **lower award probabilities** (between 0.5 and 0.8) for PIs in the **30–45 age range**, with a notable spread of probabilities below 0.6 for certain states.
- **Age Trends:**
    - Principal Investigators aged **50 and older** consistently achieve high probabilities (close to 1.0).
    - Younger PIs (below 40) show greater variability, with some achieving probabilities as low as **0.3**.
- **State Representation:**
    - States such as **California, Massachusetts, and New York** (represented with distinct colors) have a strong presence across the age spectrum.
    - Certain dots representing states like **Arizona and Colorado** show lower probabilities for younger PIs.
- **Gap Between Ages and Probabilities:**
    - There is no obvious linear relationship between **PI age** and award probability. While older PIs generally have higher success rates, a significant number of younger PIs also achieve probabilities close to **1.0**, suggesting that age alone is not a decisive factor.

Final Product



Dashboard 2

Problem Statement

The patent analysis dashboard aims to address the lack of comprehensive understanding of inventor demographics and patent filing trends. It seeks to uncover insights into gender disparities, ethnic diversity, and historical patterns in patent filings, with the ultimate goal of promoting diversity and innovation in the inventor community.

Intended Audience

The intended audience for this dashboard includes:

Patent office administrators

Policymakers in science and technology
Researchers studying innovation and diversity in STEM fields
Corporate R&D departments
Venture capitalists and investors interested in innovation trends

---

Data Source

The dashboard utilizes data from three main datasets:
Authorlink_USPTO: Contains author IDs and probabilities
UIUC_USPTO: Provides inventor IDs, first and last application years, and patent counts
genni-ethnea-authority2009: Includes author IDs, ethnicity data (EthnicSeer and Ethnea), and gender information (SSNgender)
These datasets are joined using author IDs and inventor IDs to create a comprehensive view of inventor demographics and patent activity.

Steps

Data Integration: Join the three datasets using author IDs and inventor IDs
Data Cleaning: Ensure consistency in gender and ethnicity classifications
Data Analysis: Calculate key metrics such as patent counts by ethnicity and gender
Visualization Creation: Develop charts and graphs to represent the analyzed data

---

Dashboard Assembly: Combine visualizations into a cohesive, interactive dashboard
Filtering Implementation: Add filters for year, ethnicity, and gender to enable dynamic data exploration

---

Methods and Analysis

The dashboard employs various analytical methods to provide insights:
Trend Analysis: A line chart displays patent filing trends over time, revealing historical patterns of innovation activity from the early 1900s through 20101

Demographic Analysis:

A **bubble chart** shows patent counts by ethnicity, highlighting the diversity among inventors

The **bar chart** and **box plot** illustrate gender diversity among inventors, emphasizing the significant gender disparity in patent filings

Top Performer Identification: A **leaderboard table** lists the top 10 inventors by patent count, showcasing individual achievements

The **line chart** shows the time-series trend of number of patents filed by the year of first application

---

Probability Analysis: A box plot displays patent probabilities by gender, providing insights into success rates across different demographic groups1

Comparative Analysis: The dashboard allows for comparison of patent activity across different ethnicities and genders through interactive filters1

Calculated Columns

The following calculated columns were created to enhance the analysis:
Top_10_PatentersByCounts: Identifies the top 10 inventors based on their total patent counts1

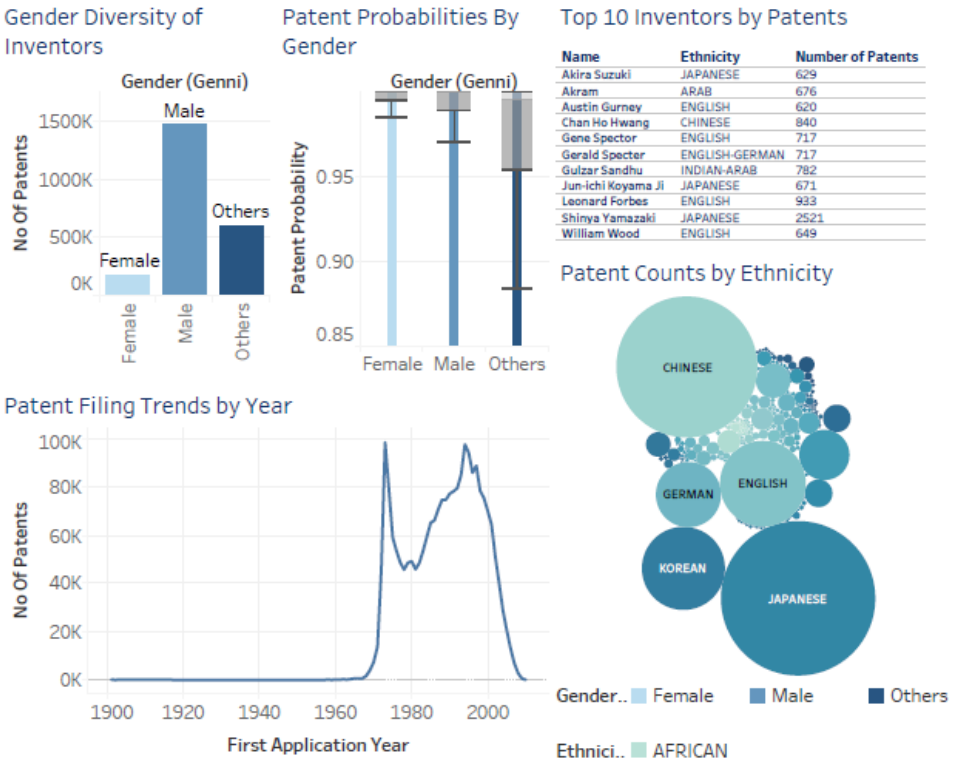
Name: Combines first and last names from the original datasets to create a full name field for easier identification1

Gender (Genni): Standardizes gender classification across the dataset, possibly using the SSNgender field from the genni-ethnea-authority2009 dataset1

DenseRank_Patents: Assigns a dense rank to inventors based on their patent counts, allowing for proper ordering in the top inventors list1

These calculated columns enable more sophisticated analysis and visualization of the patent data, providing a comprehensive view of inventor demographics and patent filing trends.

Final Product



DASHBOARD 3

Comprehensive Insights into Institutional Awards and Demographics
This theme reflects the dashboard's focus on providing a holistic view of awards distribution and demographic trends within academic institutions.

Dataset(s) Used: Merged Dataset of **authorlink_nih.tsv** and **authorlink_nsf.tsv**

Methods & Analysis:

To prepare the data for analysis, a structured approach was taken to clean and integrate multiple datasets seamlessly:

Step 1: Import Data

Loaded the both files (authorlink_nih.tsv and authorlink_nsf.tsv) into Tableau Prep.

Step 2: Perform Joins

1. NIH and NSF Tables

- Join Type: Inner Join to ensure all records are included.
- Join Condition: au_id (NIH) = au_id (NSF).
- Result: A combined table containing grant information from both NIH and NSF.

Step 3: Clean Data

- Filter Records:
  - Removed all negative age records to ensure consistent age data
- Standardize Field Names:
  - Created new fields to avoid any confusion(e.g.,Rank_Institution → Institution Rank).
- Aggregate Fields:
  - Combine fields like patents into (e.g., count(Number of Patents)).
- Export Final Dataset:
  - Saved the cleaned dataset as a .csv file for use in Tableau.
- Filtered and cleaned data in Tableau using calculated fields to replace nulls with placeholders like .
- Created calculated groups for Funding Agency to categorize records based on field conditions.

Analysis Overview:

The analysis used Tableau to visualize the data. Key steps included:

Visualizations Created:

Award Distribution Trends Across Age Groups Over the Years

A table displaying how NIH and NSF awards are distributed across different **age groups** over multiple **fiscal year ranges**. It highlights award trends for younger and older age groups.

Probability of Award Success Across Age Groups

A line chart showing the **average probability of award success** across different age groups. It helps identify which age group has the highest chances of securing awards.

Top 50 Cities Leading in Award Achievements

A horizontal bar chart showing the **top cities** where institutions have achieved the most awards. It ranks cities based on their total award counts.

Institution Distribution by State

A vertical bar chart representing the **distribution of institutions** across various states. It highlights the states with the highest number of distinct institutions.

Institution Rankings Based on Awards

A scatter plot ranking institutions by the **number of awards received**. It provides an overview of top-performing institutions across different states.

- **Tools Used:**
    - Tableau Prep: For initial data joins, merging, and cleaning.
    - Tableau Desktop: For advanced visualizations and further filtering.

Key Findings:

Age and Funding Patterns:

- NIH and NSF awards were heavily concentrated among **younger PIs** in the **20-25 and 25-30 age groups**, particularly during recent fiscal years.
- Award distribution declined among older age groups, showing a trend of funding preference toward early-stage researchers.

Geographic Trends:

- **East Coast**, **Midwest**, and **West Coast** regions dominated NIH and NSF funding distribution.
- States like **California, New York, Massachusetts, and Pennsylvania** emerged as the leading regions for award achievements.
- Funding was significantly clustered in cities like **Cambridge**, **New York**, and **Philadelphia**.

Institutional Rankings:

- Leading institutions were concentrated in **top-performing states**.
- Cities like **Cambridge** and **New York** ranked highest in award achievements, reinforcing their prominence as major hubs for research funding.

Demographic Insights:

- NIH and NSF awards showed **underrepresentation of older age groups**, with funding primarily concentrated among **younger PIs**.

Challenges Encountered

- **Data Quality Issues:**
    - Many fields contained null or inconsistent data(like negative age records).
    - Solution: Applied filters to remove null values or replace them with placeholders using calculated fields like Rank_Institution, No_Awards.
- **Complex Joins:**
    - Merging multiple datasets required careful join conditions to avoid data loss.
    - Solution: Performed stepwise joins (outer, inner, left) to ensure data integrity.
- **Visualization Complexity:**
    - Large datasets and multiple fields made creating effective visualizations challenging.
    - Solution: Used Tableau's calculated fields and grouping features to simplify the data for clear storytelling.

Final Product

1. Award Distribution Trends Across Age Groups Over the Years: Displayed the trends in award distributions across different age groups over time.
2. Top 50 Cities Leading in Award Achievements: Showed the top cities with the highest award achievements.
3. Institution Distribution by State: Mapped the distribution of institutions by state.

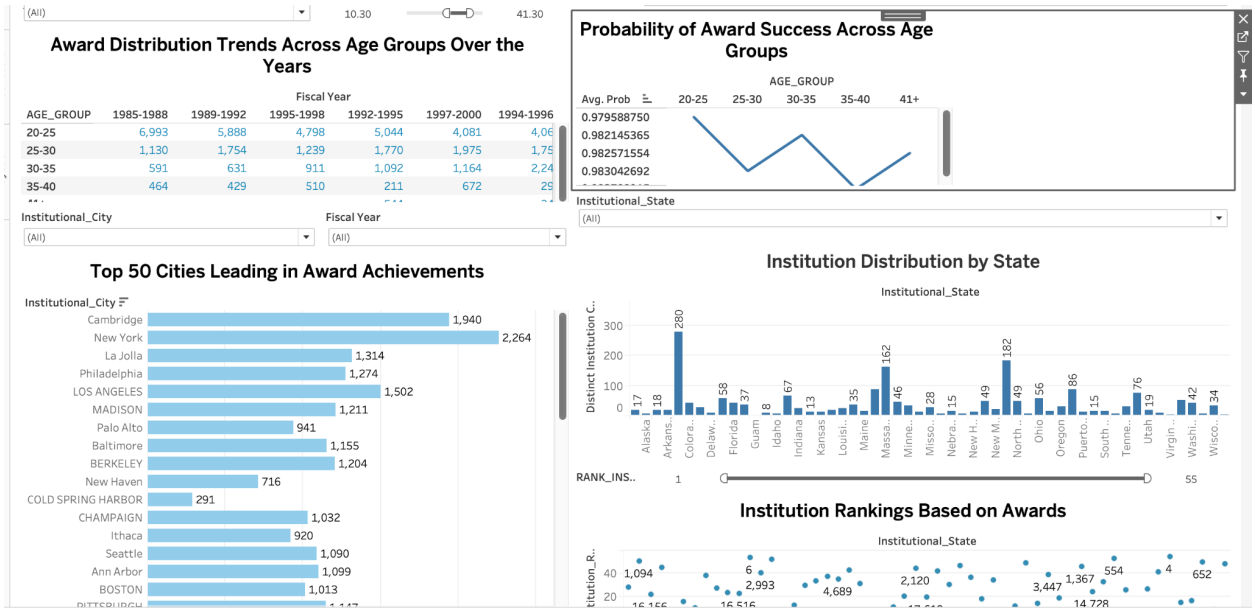
4. Institution Rankings Based on Awards: Visualized the rankings of institutions based on the awards received.
5. Probability of Award Success Across Age Groups

Conclusion

The analysis of NIH and NSF awards highlights significant patterns in award distribution across key demographics, regions, and institutions. Younger age groups consistently lead in award achievements, particularly among early-career researchers. Geographically, funding remains concentrated in top-performing cities and states, with notable dominance along the East Coast, Midwest, and West Coast regions. Institutional rankings reveal a strong correlation between location and success rates, with certain cities and institutions standing out as leaders in securing awards. Despite these insights, gender representation remains imbalanced, with male PIs receiving the majority of awards. These findings emphasize the need for targeted strategies to address demographic gaps and promote equitable funding opportunities across age, gender, and geographic lines.

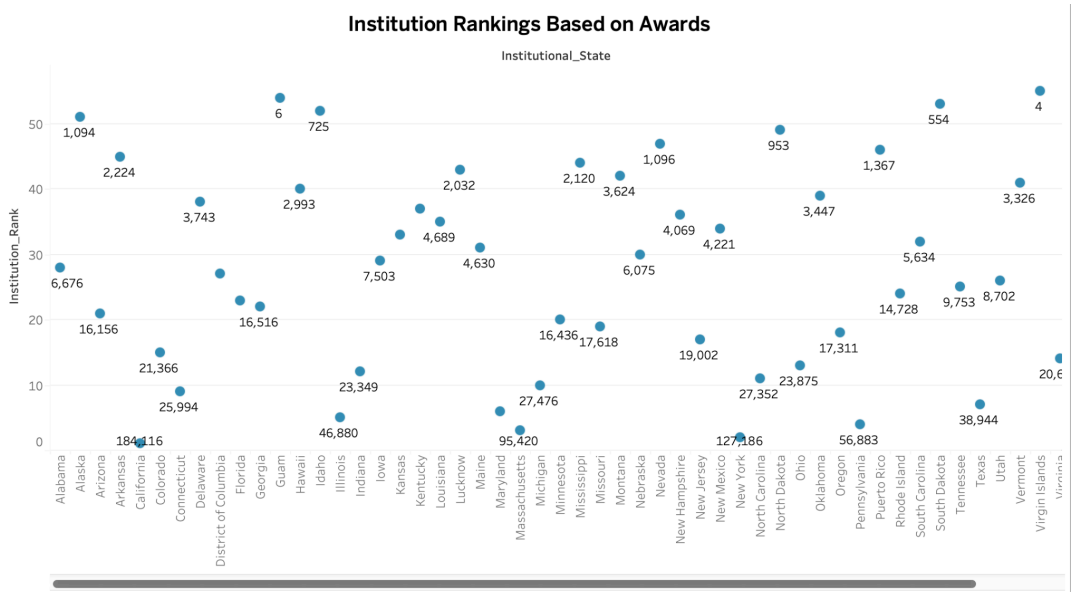
FINAL PRODUCT

FINAL DASHBOARD 3



Individual Visualizations-

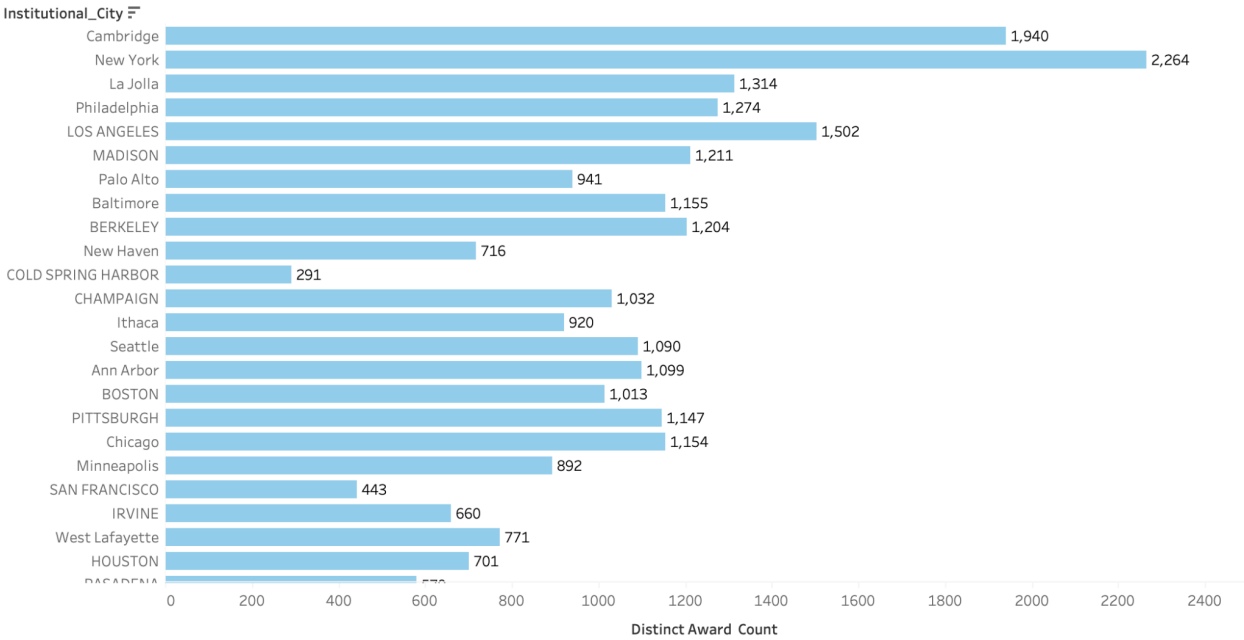
Institution Rankings Based on Awards



Award Distribution Trends Across Age Groups Over the Years

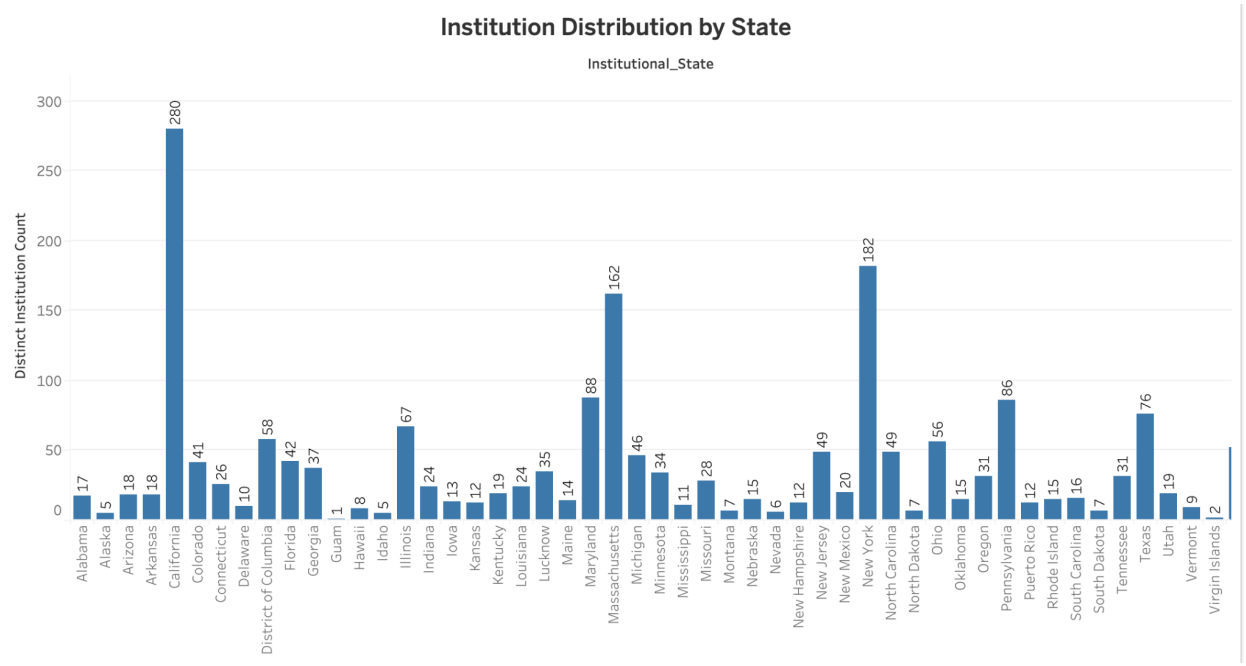
	Fiscal Year									
AGE_GROUP	1985-1988	1989-1992	1995-1998	1992-1995	1997-2000	1994-1996	2009-2013	2012-2015	2011-2015	2013-2016
20-25	6,993	5,888	4,798	5,044	4,081	4,060	3,054	2,424	1,833	1,945
25-30	1,130	1,754	1,239	1,770	1,975	1,752	1,038	1,168	727	740
30-35	591	631	911	1,092	1,164	2,246	1,260	1,117	858	905
35-40	464	429	510	211	672	291	1,298	867	839	864
41+				544		343				

Top 50 Cities Leading in Award Achievements



Probability of Award Success Across Age Groups

Avg. Prob	AGE_GROUP				
	20-25	25-30	30-35	35-40	41+
0.979588750					
0.982145365					
0.982571554					
0.983042692					
0.983722915					



Award Distribution Trends Across Age Groups Over the Years

	Fiscal Year									
AGE_GROUP	1985-1988	1989-1992	1995-1998	1992-1995	1997-2000	1994-1996	2009-2013	2012-2015	2011-2015	2013-2016
20-25	6,993	5,888	4,798	5,044	4,081	4,060	3,054	2,424	1,833	1,945
25-30	1,130	1,754	1,239	1,770	1,975	1,752	1,038	1,168	727	740
30-35	591	631	911	1,092	1,164	2,246	1,260	1,117	858	905
35-40	464	429	510	211	672	291	1,298	867	839	864
41+				544		343				

DASHBOARD 4

Methods & Analysis:

To prepare the data for analysis, a structured approach was taken to clean and integrate multiple datasets seamlessly:

Step 1: Import Data

Loaded the four files (authorlink_nih.tsv, authorlink_nsf.tsv, authorlink_uspto.tsv, uiuc_uspto.tsv) into Tableau Prep.

Step 2: Perform Joins

1. NIH and NSF Tables

- Join Type: Outer Join to ensure all grant-related records are included.
- Join Condition: au_id (NIH) = au_id (NSF).
- Result: A combined table containing grant information from both NIH and NSF.

2. Add USPTO Linking Data

- Join Type: Inner Join to include only authors linked to inventors.
- Join Condition: au_id (from the previous join) = au_id (USPTO linking dataset).
- Result: Merges grant and patent data.

3. Merge with Inventor Data

- Join Type: Left Join to retain all grant and patent data while adding inventor details.
- Join Condition: inv_id (USPTO linking dataset) = inv_id (uiuc_uspto.tsv).
- Result: Adds detailed patent information to the dataset.

4. Add Demographics from EthnicSeer

- Join Type: Inner Join to retain only records with demographic predictions.
- Join Condition: au_id (from the combined dataset) = auid (EthnicSeer).
- Result: Final unified dataset with grant, patent, and demographic information.

Step 3: Clean Data

- Filter Records:
    - Removed rows where prob (match probability) < 0.5.
- Standardize Field Names:
    - Renamed all fields for consistency (e.g., nih_full_proj_nbr → Project Number).
- Aggregate Fields:
    - Combine fields like patents into (e.g., count(Number of Patents)).
- Export Final Dataset:
    - Saved the cleaned dataset as a .csv file for use in Tableau.
- Filtered and cleaned data in Tableau using calculated fields to replace nulls with placeholders like 'Name Not Available'.
- Created calculated groups for Funding Agency to categorize records based on field conditions.

Analysis Overview:

Key Findings:

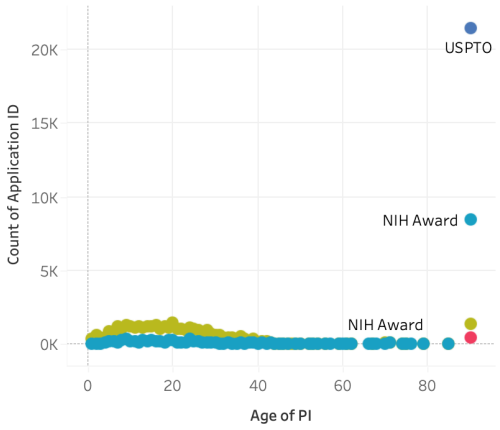
- Age and Funding Connections:
    - Award distributions varied across different age groups, with the 20-25 and 25-30 age groups consistently receiving the highest number of awards over the years.
    - The award success probability also differed across age groups, with the 'Early-Career' and 'Mid-Career' stages having the highest probabilities.
- Geographic Trends:
    - Award achievements were concentrated in a few key cities, with New York, Cambridge, and La Jolla leading the top 50 cities.
    - The institution distribution by state showed funding was primarily focused in a few states, with California, New York, and Massachusetts being the top recipients.
- Institutional Performance:
    - The institution rankings based on awards revealed a clear hierarchy, with top universities like the University of Wisconsin, University of California, and University of Michigan consistently receiving the most awards.

Challenges Encountered:

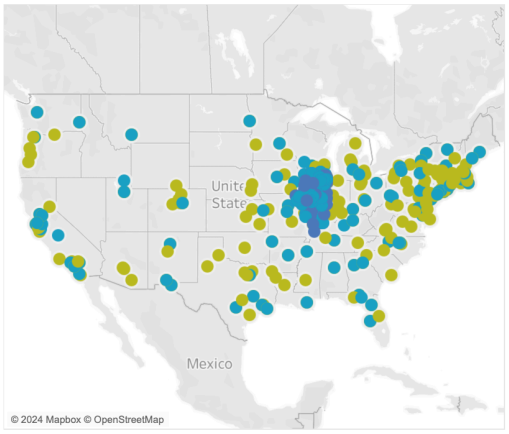
- Data Quality Issues:
    - The dataset contained null values and inconsistencies that required careful data cleaning and preparation.
    - Solution: Used Tableau Prep to identify and address data quality issues, such as replacing null values and harmonizing data types.
- Visualization Complexity:
    - The large number of data points and multiple dimensions made creating effective visualizations a challenge.
    - Solution: Leveraged Tableau's features like calculated fields, filters, and grouping to simplify the data and enhance the storytelling.

Final Product

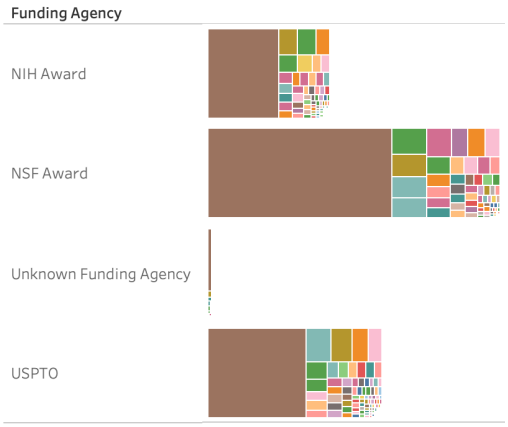
Patent and Grant Connections by Age



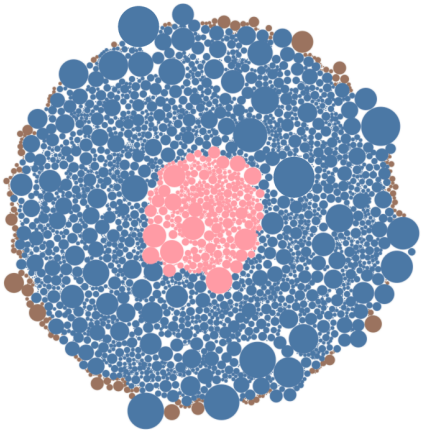
Geographic Funding Trends



Ethnic Distribution by Funding Agency



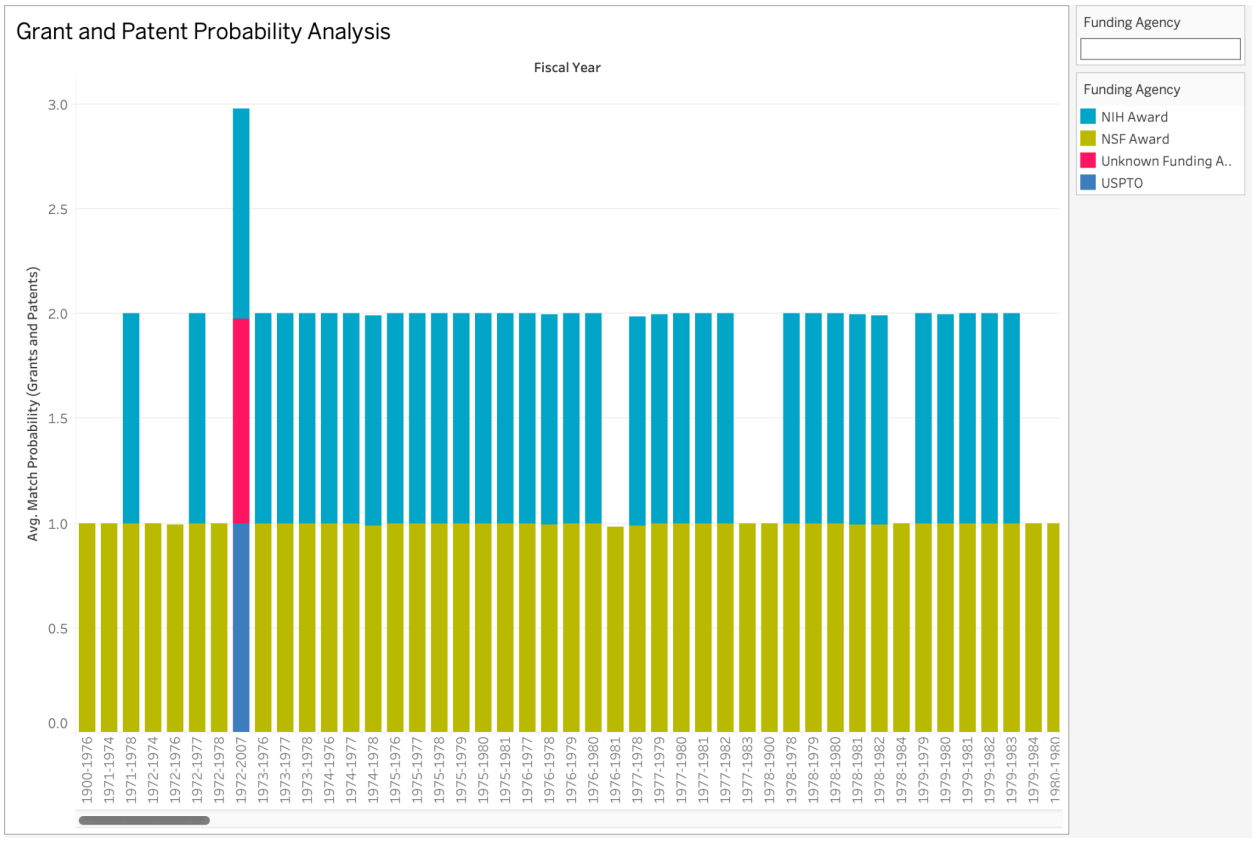
Gender vs. Patent Success Rate

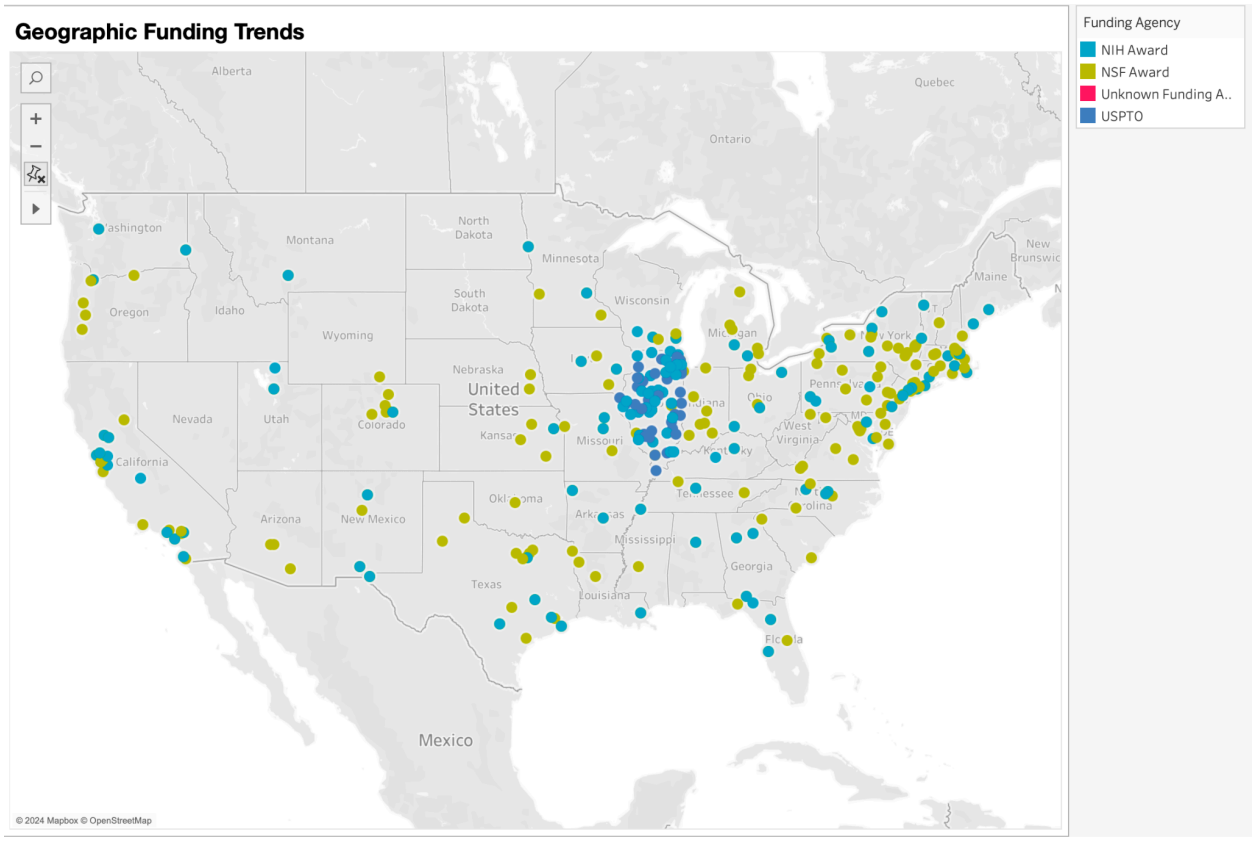


Ethnicity
- AFRICAN
- AFRICAN-CHINESE
- AFRICAN-ENGLISH
- ARAB
- ARAB-ENGLISH
- ARAB-FRENCH
- ARAB-INDIAN
- ARAB-ISRAELI
- BALTIC
- CHINESE
- CHINESE-BALTIC
- CHINESE-ENGLISH
- CHINESE-GERMAN
- CHINESE-KOREAN
- DUTCH
- DUTCH-ENGLISH
- DUTCH-FRENCH
- DUTCH-GERMAN
- ENGLISH
- ENGLISH-AFRICAN
- ENGLISH-ARAB
- ENGLISH-CHINESE
- ENGLISH-DUTCH
- ENGLISH-FRENCH
- ENGLISH-GERMAN
- ENGLISH-GREEK
- ENGLISH-HISPANIC
- ENGLISH-INDIAN

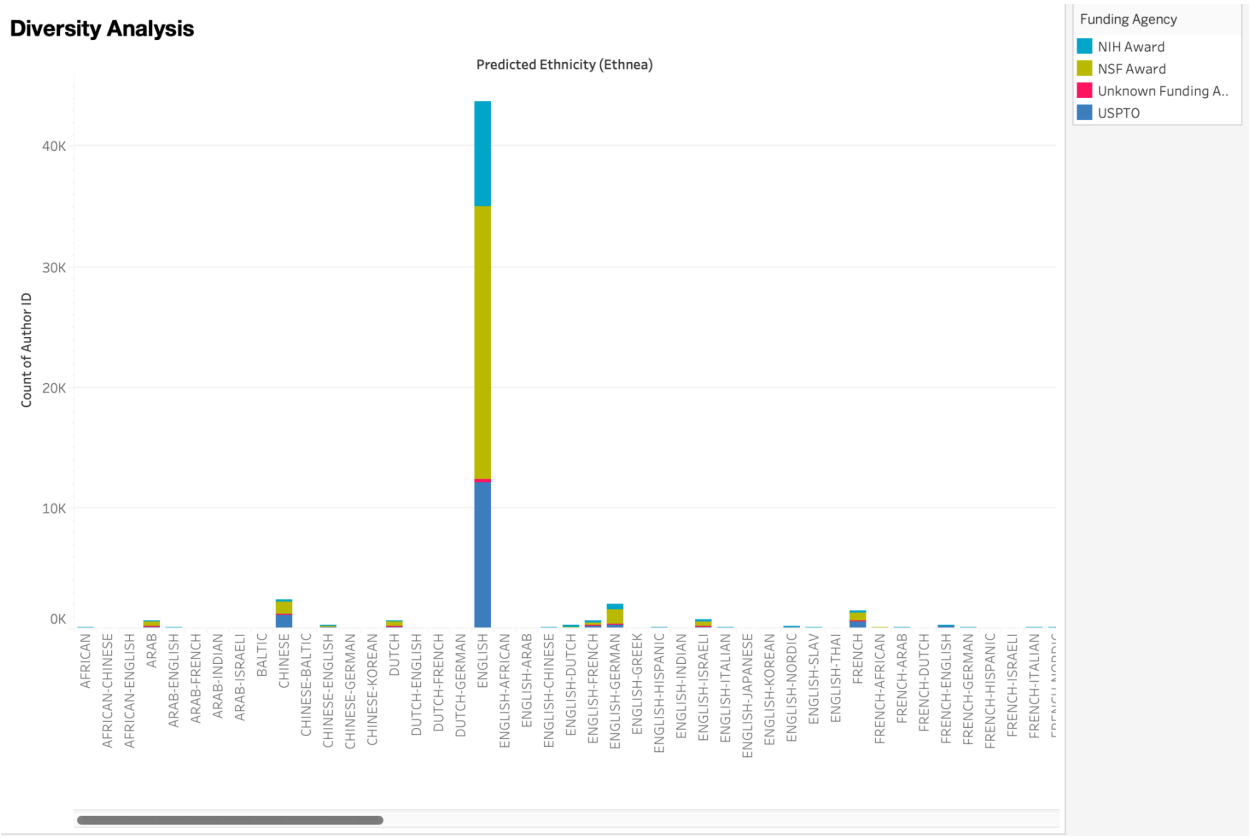
Funding Agency
- NIH Award
- NSF Award
- Unknown Funding A..
- USPTO

Gender
- F
- M
- Others

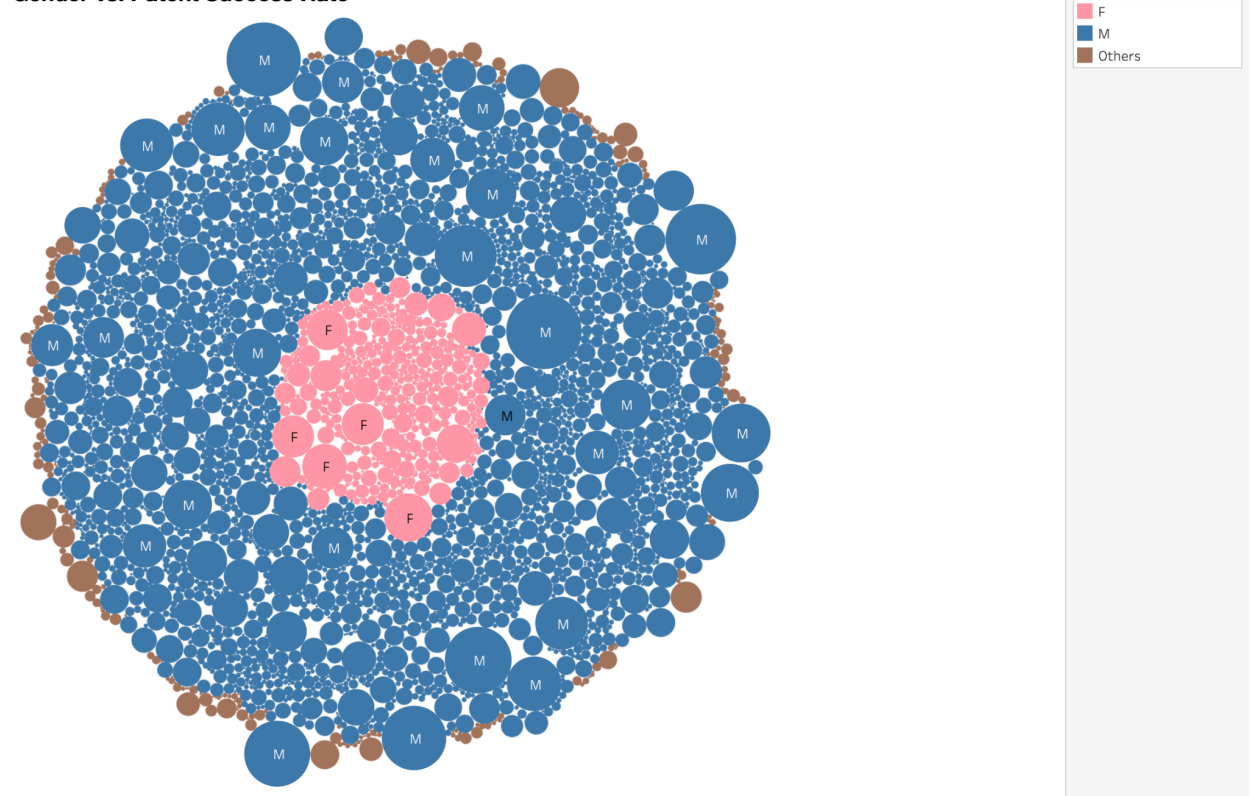

Grant and Patent Probability Analysis
Fiscal Year
Avg. Match Probability (Grants and Patents)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1900-1976
1971-1974
1971-1978
1972-1974
1972-1976
1972-1977
1972-1978
1972-2007
1973-1976
1973-1977
1973-1978
1974-1976
1974-1977
1974-1978
1975-1976
1975-1977
1975-1978
1975-1979
1975-1980
1975-1981
1976-1977
1976-1978
1976-1979
1976-1980
1976-1981
1977-1978
1977-1979
1977-1980
1977-1981
1977-1982
1977-1983
1978-1900
1978-1978
1978-1979
1978-1980
1978-1981
1978-1982
1978-1984
1979-1979
1979-1980
1979-1981
1979-1982
1979-1983
1979-1984
1980-1980
Funding Agency
Funding Agency
NIH Award
NSF Award
Unknown Funding A..
USPTO


Geographic Funding Trends
Alberta
Ontario
Quebec
Montana
North Dakota
Minnesota
Wisconsin
Maine
Oregon
Idaho
South Dakota
Wyoming
Nebraska
United States
Nevada
Utah
Colorado
Kansas
Missouri
West Virginia
California
Arizona
New Mexico
Oklahoma
Tennessee
Mississippi
Georgia
Texas
Louisiana
Mexico
© 2024 Mapbox © OpenStreetMap
Funding Agency
NIH Award
NSF Award
Unknown Funding A..
USPTO

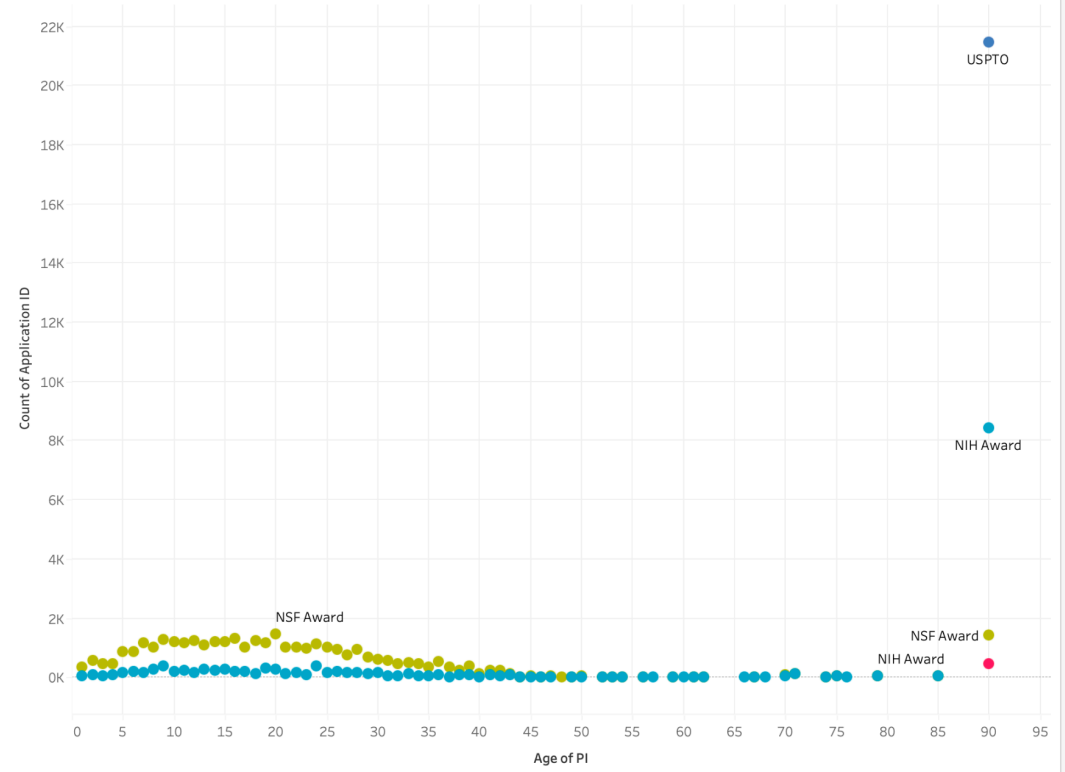
Diversity Analysis



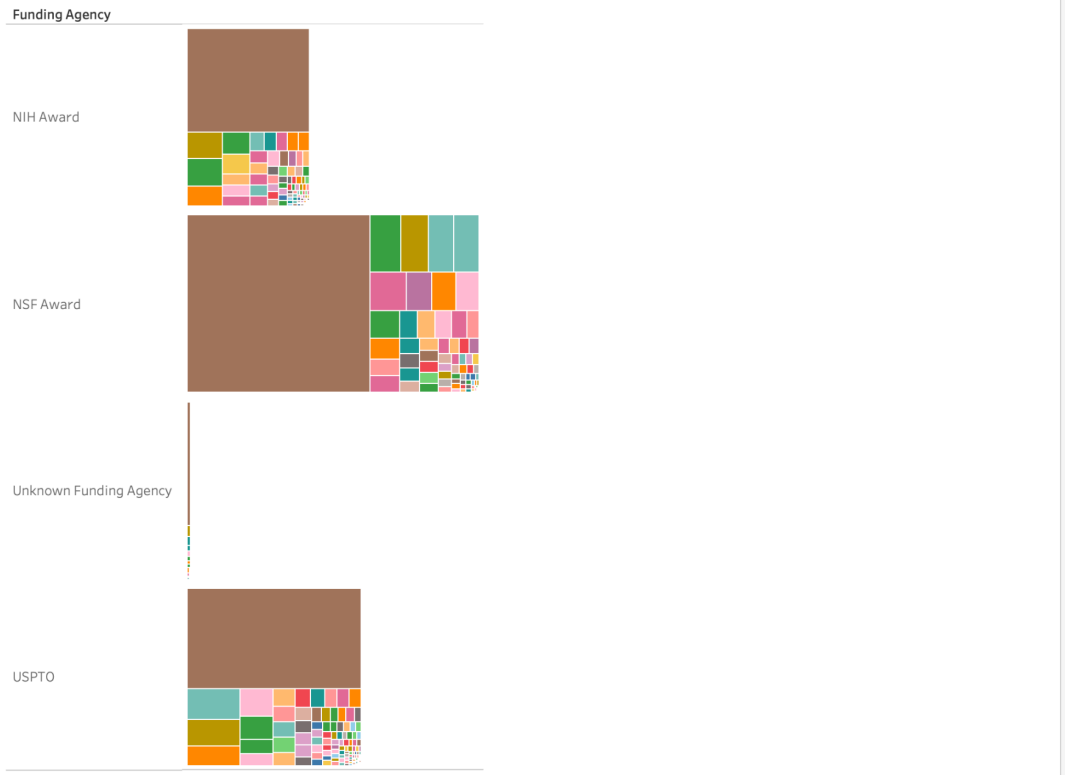
Gender vs. Patent Success Rate



Patent and Grant Connections by Age



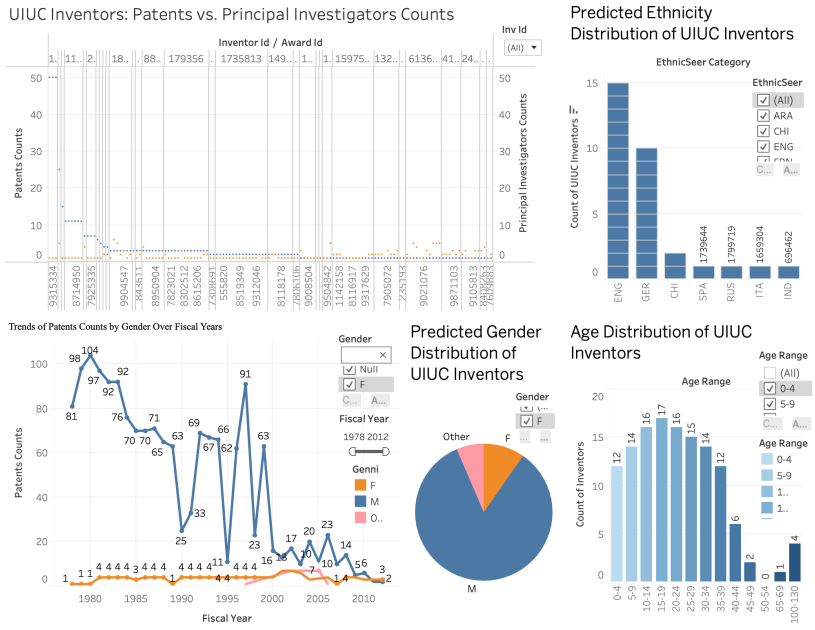
Ethnic Distribution by Funding Agency



DASHBOARD 5

Patent and Demographic Trends of UIUC Inventors

The overview of Dashboard 5 is shown in the following figure:



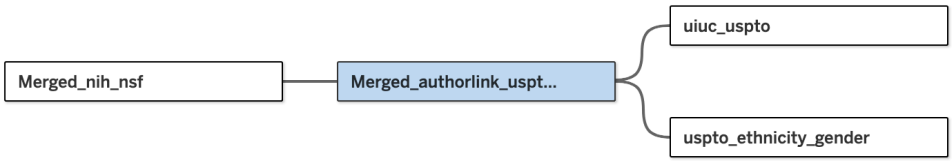
Datasets Used:

Original Datasets:
authorlink_nih.tsv,
authorlink_nsf.tsv,
authorlink_uspto.tsv,
uiuc_uspto.tsv,
genni-ethnea-authority2009.tsv

Derived Datasets:

Merged_nih_nsf.xlsx
Merged_authorlink_uspto_genni_ethnea.xlsx
uiuc_uspto.xlsx
uspto_ethnicity_gender.xlsx

The relationship between my data source is shown in the following figure:

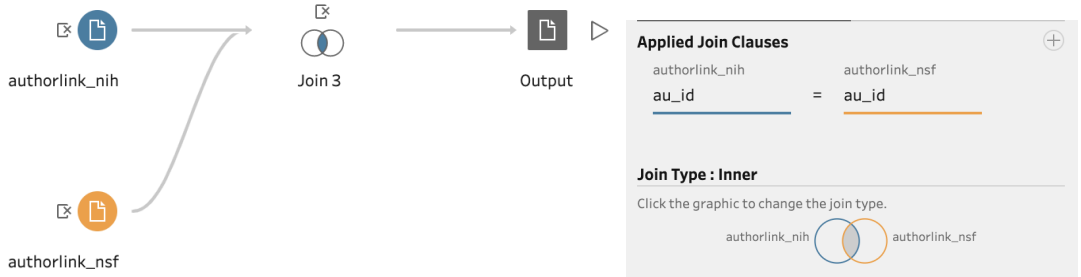


Theme:

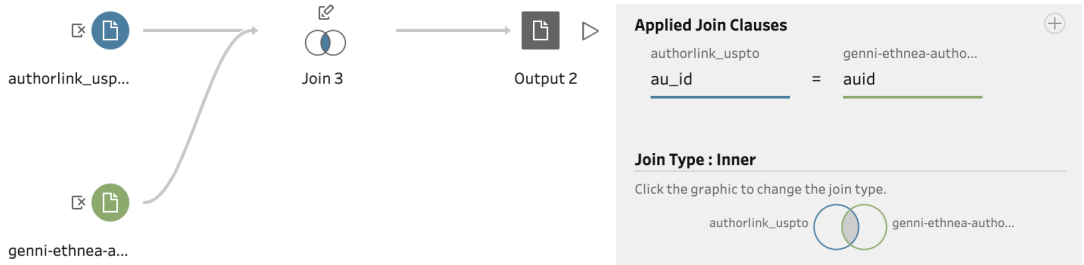
The theme of Dashboard 5 is to show data visualizations for users to explore interactively the patent activity and demographic characteristics (ethnicity, gender, and age) of inventors affiliated with the University of Illinois Urbana-Champaign (UIUC).

Data Preprocessing

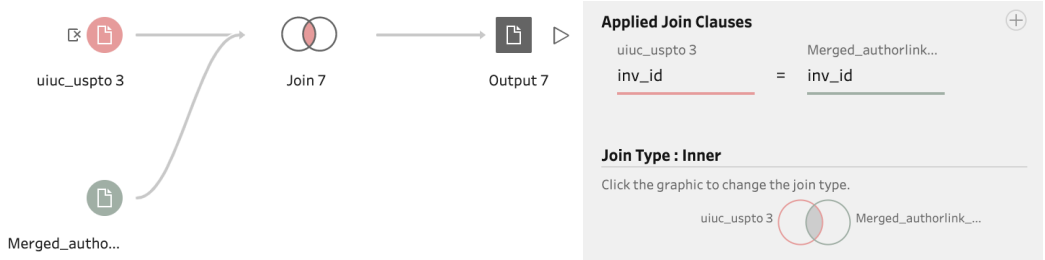
Join Type: Inner
The join operation for Merged_nih_nsf.xlsx is shown in the following figure:



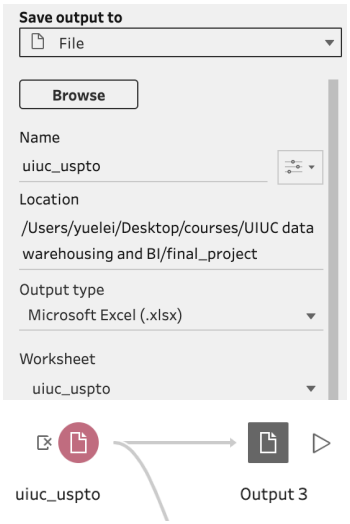
The join operation for Merged_authorlink_uspto_genni_ethnea.xlsx is shown in the following figure:



The join operation for uspto_ethnicity_gender.xlsx is shown in the following figure:

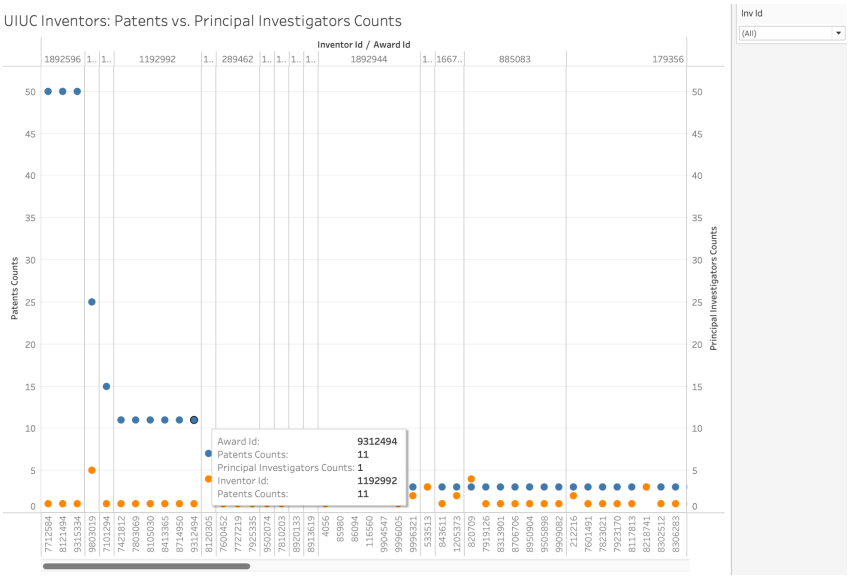


The uiuc_uspto.xlsx is created by output from the uiuc_uspto.tsv as shown in the following figure:

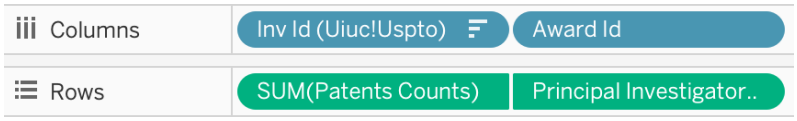


Visualization 1: Dual-Axis Scatter Plot

Visualization 1 "UIUC Inventors: Patents vs. Principal Investigators Counts" is shown in the following figure:



The table fields used in Columns and Rows of the visualization is shown in the following figure:



Information Shown in This Visualization

Award Ids are grouped by their corresponding Inv Id, splitting the x-axis into hierarchical levels.

Color Setting:

Blue dots represent Patents Counts for each inventor, and orange dots represent Principal Investigator Counts for each inventor at the grouped Award Id level.

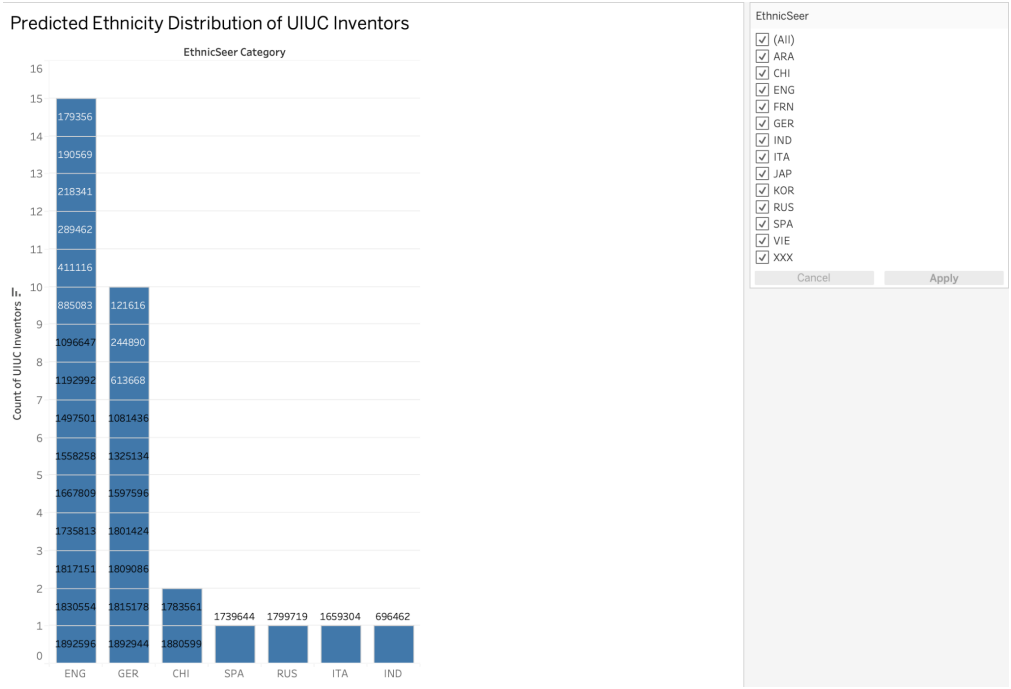
Interactive Feature:
A filter dropdown for Inv Id allows users to interactively select specific inventors for detailed analysis.

Tooltips Including:
Inventor Id, Award Id, Patents Counts, Principal Investigators Counts corresponding to that Inventor Id.

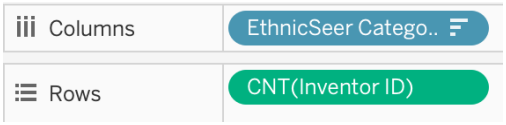
Interesting Discoveries from the Visualization
Most inventors have patents between 1 and 10, there are only a few inventors with relatively high patent counts (above 50). The inventor with Inventor ID 1892596 has the most 50 patents affiliated with UIUC.
While many inventors have a few Principal Investigators (between 0 and 5), some have higher Principal Investigators counts, indicating involvement in collaboration across multiple awards.
There is no obvious trend that could indicate any relationship between Patent Counts and Principal Investigators Counts. Certain inventors stand out with a significant gap between Patent Counts and Principal Investigators Counts, which may indicate further investigation into their research or innovation practices.

Visualization 2: Bar Chart

Visualization 2 "Predicted Ethnicity Distribution of UIUC Inventors" is shown in the following figure:



The table fields used in Columns and Rows of the visualization is shown in the following figure:



Information Shown in This Visualization

Color Setting:

All the bars are set to be in blue color.

Interactive Feature:

A multiple values list for EthnicSeer allows users to interactively select specific ethnicities for detailed analysis.

Tooltips Including:

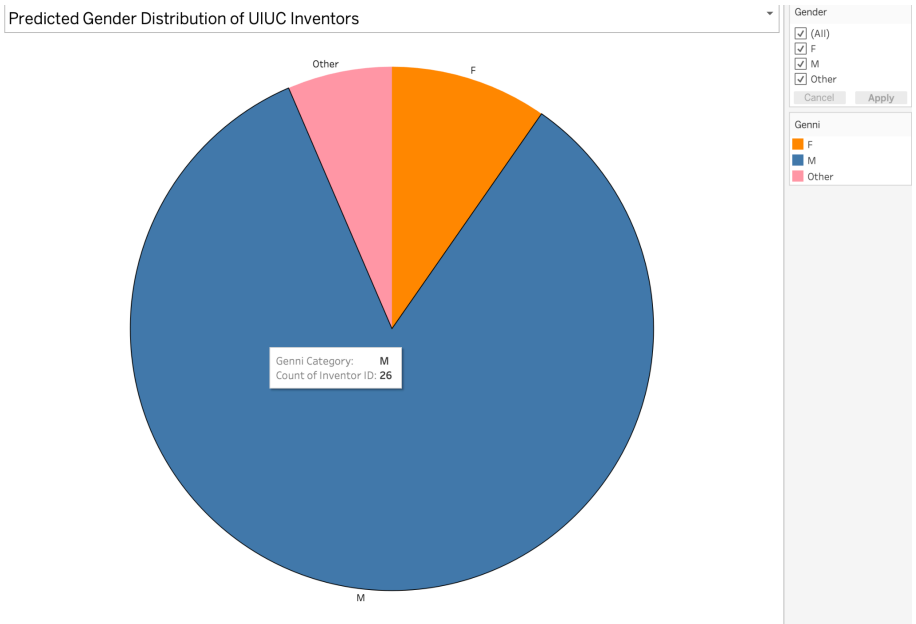
Every Inventor ID categorized into a specific EthnicSeer Category would show in place accordingly.

Interesting Discoveries from the Visualization

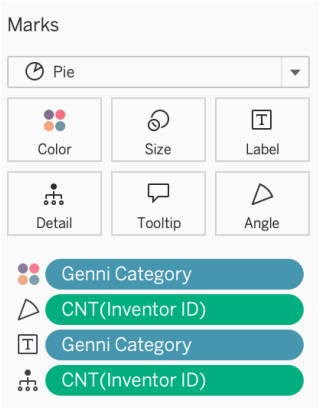
The EthnicSeer Category ENG has the highest number of inventors, with a count of 15. GER follows closely, with a count of around 10. Other EthnicSeer categories have significantly fewer inventors, with fewer than 3 individuals. Several EthnicSeer categories (e.g., JAP, KOR, VIE, XXX) appear in the filter but are not represented in the chart, suggesting no UIUC inventors fall under those categories.

Visualization 3: Pie Chart

Visualization 3 "Predicted Gender Distribution of UIUC Inventors" is shown in the following figure:



No table fields are used in Columns and Rows of the visualization. However, the Marks card is set as shown in the following figure:



Information Shown in This Visualization

Color Setting:

F: gender of female is shown in color orange.
M: gender of female is shown in blue orange.
Other: other gender is shown in pink orange.

Interactive Feature:

A multiple values list for Gender allows users to interactively select specific gender categories for detailed analysis.

Tooltips Including:

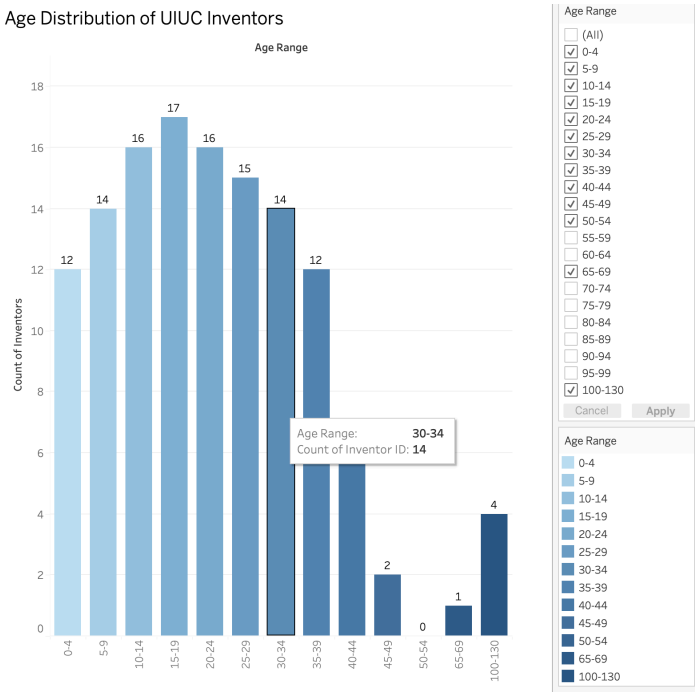
Genni Category and Count of Inventor ID for every gender category represented by a pie segment.

Interesting Discoveries from the Visualization

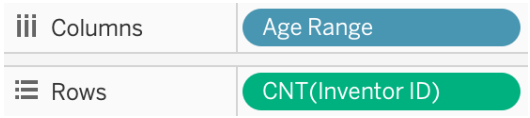
The pie chart segments represent the distribution of predicted genders among inventors affiliated with UIUC. The "Male" category occupies the largest portion of the pie chart, the "Female" category has a small portion of the pie with a count of only 3 inventors, the "Other" category has a slightly smaller segment compared to "Female" with a count of 2 inventors, indicating a significant gender imbalance and a skewed gender distribution among UIUC inventors in the dataset.

Visualization 4: Bar Chart

This visualization 4 "Age Distribution of UIUC Inventors" is shown in the following figure:



The table fields used in Columns and Rows of the visualization is shown in the following figure:



Information Shown in This Visualization

Color Setting:
All the bars are set to be in gradient blue colors, the older the age range is, the darker the blue color is.

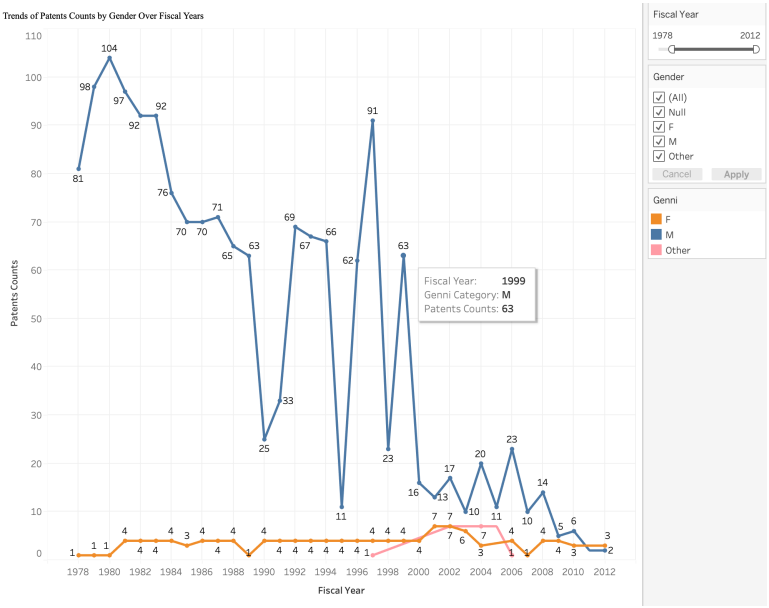
Interactive Feature:
A multiple values list for Age Range allows users to interactively select specific age ranges for detailed analysis.

Tooltips Including:
Age range and Count of Inventor ID under that age range.

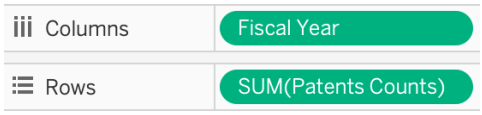
Interesting Discoveries from the Visualization
The visualization displays counts of unique inventors affiliated with UIUC under different age ranges including 0-4, 5-9, 10-14, ..., up to 100-130. The age group 15-19 has the highest number of inventors with a count of 17. Younger inventors seem to dominate the distribution of inventors in terms of age, they fall within the 0-34 age range. Surprisingly, there are 4 inventors in the age group 100-130

Visualization 5: Line Chart

Visualization 5 "Trends of Patents Counts by Gender Over Fiscal Years" is shown in the following figure:



The table fields used in Columns and Rows of the visualization is shown in the following figure:



Information Shown in This Visualization

Color Setting:
F: gender of female is shown in color orange.
M: gender of female is shown in blue orange.
Other: other gender is shown in pink orange.

Interactive Feature:
A slide of Range of Values for Fiscal year and a multiple values list for Gender allow users to interactively adjust specific fiscal years and toggle between different gender categories for detailed analysis.

Tooltips Including:
Fiscal Year, Genni Category and the total Patents Counts corresponding to that gender category in that year.

Interesting Discoveries from the Visualization
This visualization shows trends of the Sum of Patents Counts over a continuous timeline Fiscal Years (1978–2012). The trend line for Male in blue color shows significantly higher total patent counts compared to Female and Other throughout the fiscal years. However, it exhibits significant fluctuations, with sharp peaks and drops, particularly between 1980-2000. The trend line for females in orange remains relatively low, but it seems more steady or consistent. The

trend line for Other starts to show values after 1996, and seems to stop the patent activity in fiscal year 2006.

Challenges We Had:
We took a bunch of time to come up with more ideas that are different from the first 4 dashboards. After confirming that central idea for dashboard 5, we discussed actively on how to preprocess the original datasets by calculating more derived table fields.

Conclusion:
Dashboard 5 is a powerful analytical tool that could be applied to research analysis for institutional leaders, researchers, and industry stakeholders. It serves to promote research activity diversity, identify trends in research innovation, and support evidence-based strategies to enhance the University of Illinois Urbana-Champaign and other institutions' contributions to research and technological advancement.