Innovating Patent Analytics

Harnessing AI & Behavioral Sciences to Transform USPTO Operations

Contributing Authors[1]

Meriem Mehri
Sheida Majidi
Keani Schuller
Niki Mahmood Zadeh
Joshua Poozhikala

[1]McGill University, Desautels Faculty

Presented to

Professor Roman Galperin

Project Overview: Our project integrates a comprehensive analysis of USPTO data to investigate how organizational and social factors, including gender, race, and ethnicity, influence patent application review times and examiner attrition. This endeavor is twofold: it begins with an empirical study pinpointing specific aspects of patent prosecution duration and attrition causes, followed by a critical evaluation of existing people analytics tools—highlighting generative AI and behavioral science innovations—for their effectiveness in tackling these issues. We aim to offer well-founded recommendations to improve the USPTO's patent processing efficiency and fairness, ensuring equitable patent rights access. Furthermore, we propose developing an innovative tool employing OpenAI technologies to refine talent analytics, thereby enhancing the USPTO's crucial role in fostering economic growth through fair and unbiased patent grant processes.

Objectives: This project sets out to enhance the USPTO's efficiency and fairness in patent processing. Our goals include:

- Conducting a detailed empirical analysis to understand how factors like gender, race, and ethnicity influence patent review times and examiner attrition at the USPTO.
- Identifying key research questions regarding the duration of patent prosecutions and the factors contributing to examiner turnover.
- Assessing the effectiveness of current analytics tools, especially those incorporating generative AI and behavioral science, in overcoming the identified challenges.
- Formulating practical recommendations to improve patent processing speed and reduce biases.

Problem Statement:

Our report examines USPTO data to understand how factors like gender, race, and ethnicity affect patent review times and examiner turnover. We split our analysis into two parts: a detailed study of patent processing times and turnover, and an evaluation of the latest analytics tools, focusing on AI and behavioral science innovations. Our goal is to offer clear recommendations and introduce a new tool using OpenAI technologies to make the patent process more efficient and fairer. This work aims to ensure the USPTO operates transparently, supports innovation, and grants patents without bias.

Data Sources/Datasets

The dataset compiles information on patent applications managed by the United States Patent and Trademark Office (USPTO). The data is arranged in rows and columns, with each row corresponding to a distinct patent application. Columns present various details including the application number, the date the application was filed, and the names of the examiners—categorized by last and first names. This dataset serves as a resource for analyzing the workflow of patent examiners and the progression of patent applications over time.

Part 1: Analyzing the Impact of Organizational and Social Dynamics on USPTO Patent Processing and Examiner Attrition

In Part 1, we conduct an empirical analysis on how organizational and social factors, particularly gender, race, and ethnicity, affect patent processing times and examiner attrition at the USPTO. We analyze USPTO data using statistical and machine learning methods to detect significant trends and make projections. Our aim is to offer recommendations to improve the USPTO's efficiency, fairness, and inclusivity, fostering a more equitable patent system.

Analytical Approach: Our analytical approach for Part 1 involves a structured methodology to explore the impact of various factors on patent application review times and examiner attrition at the USPTO. This approach includes data collection, preprocessing, statistical analysis, and model building. The methodology emphasizes

identifying and quantifying the influence of gender, race, and ethnicity among other variables. By employing regression analysis and machine learning models, the study aims to uncover significant predictors of review times and attrition rates, providing a nuanced understanding of how organizational and social dynamics affect patent processing outcomes.

Methodology: Our methodology (*Refer to our Code File*) highlights steps such as data type correction, feature creation, and handling missing values. It also details the estimation of examiner gender and race through the use of specific libraries and functions within R, showcasing an innovative approach to enrich the dataset. This comprehensive methodology aims to prepare the dataset meticulously for further statistical analysis and machine learning modeling to explore the impact of organizational and social factors on patent processing outcomes.

1. **Data Loading and Preparation:** The process begins with importing a comprehensive dataset from the USPTO, which includes detailed information about patent applications. The preparation phase involves converting date formats and calculating the length of the prosecution for each application to derive meaningful metrics for analysis.
2. **Data Cleaning and Exploration:** Subsequent data cleaning addresses missing values and inconsistencies, ensuring the dataset's integrity for analysis. An exploratory data analysis (EDA) phase follows, aimed at understanding the dataset's structure, identifying anomalies, and gaining initial insights.
3. **Estimation of Examiner Demographics:** A novel aspect of the methodology is the estimation of examiners' gender and race, utilizing advanced algorithms to infer these demographics from names, enhancing the dataset with crucial variables for analysis.
4. **Analyzing Influencing Factors:** The core analytical component investigates how various factors, including examiner demographics and organizational units, influence patent application outcomes. This analysis utilizes logistic regression to model the relationship between these factors and the outcomes, providing a quantitative assessment of their impacts.
5. **Indirect Analysis for Examiner Attrition:** Given the absence of direct attrition indicators, the methodology adapts to explore patterns indicative of attrition-like behavior. This includes analyzing tenure and application outcomes to infer potential attrition signals indirectly.

Organizational & Social Factors Associated with the Length of Patent Application Prosecution

This portion of our analysis was conducted to understand the interplay between various organizational and social factors and the length of patent application prosecution. Through exploratory visualizations and regression modeling, factors such as examiner's art unit, USPC class/subclass, gender, and race are scrutinized to uncover patterns and disparities in prosecution length.

Exploratory Analysis on Tenure by Gender and Race: This phase aimed to uncover patterns in tenure duration across different gender and racial groups among examiners. Using visualizations like box plots, the analysis provided insights into the distribution of tenure days, highlighting potential disparities or trends related to gender and race. This approach is crucial for understanding how these demographic factors might influence, or be influenced by, career longevity and engagement within the USPTO.

Prosecution Length by Gender: (*Appendix 1*)

- The red box plot represents female examiners, while the blue box plot represents male examiners.
- The median prosecution length for both genders is around the same level, with no gender showing a clear median that is significantly higher or lower than the other. Thus, stating that the median for male examiners is higher is not supported by the visual data.

- The variability within each gender is indicated by the length of the boxes and the whiskers. The spread of the data points (interquartile range) seems similar for both genders, suggesting that both female and male examiners experience a similar range of prosecution lengths.
- Outliers, or individual points beyond the whiskers, are present for both genders, indicating that there are examiners in both groups with prosecution lengths significantly longer and shorter than the typical range.

The data assesses whether there are systematic differences in the prosecution lengths handled by female and male examiners. However, based on the box plots, it appears that gender is not a distinguishing factor in prosecution length, as both genders show similar medians and spreads in the data. The presence of outliers in both groups suggests that factors other than gender may have a more significant impact on the length of prosecutions. It would be important to investigate other variables that might influence these lengths, such as case complexity, geographical location, cultural ethnicity, or examiner experience.

Prosecution Length by Race: (*Appendix 2*)

- All racial groups have a similar median prosecution length, which is indicated by the line within each box. There doesn't appear to be a significant difference in the median length of prosecution among the groups based on the data shown.
- The interquartile range, which represents the middle 50% of the data, appears to be fairly consistent across the groups, although the 'Other' category has a slightly wider interquartile range, indicating more variability in prosecution length within this group.
- The presence of outliers, indicated by the individual points beyond the whiskers, is seen in all racial groups, suggesting that there are cases with exceptionally long or short prosecution lengths across all categories.
- The 'whiskers' of the box plots, which indicate the range of prosecution length excluding outliers, are similar for the Asian, Black or African American, and Hispanic groups. The 'Other' and 'White' categories show slightly longer whiskers, which could suggest a broader range of prosecution lengths within those groups.

This figure suggests that race, as categorized here, does not show significant differences in the median length of prosecution. However, the presence of outliers and the range of prosecution lengths within the 'Other' and 'White' categories might warrant further investigation to understand the factors contributing to the wider spread. It's also important to consider the context behind the data, such as the type of cases, jurisdictional differences, and other socio-economic factors that might influence the length of prosecutions.

Note: The visual and statistical outputs provided by our code offer a wealth of information regarding the length of patent prosecution and the role of gender, race, and ethnicity in this process.

Interpretation of Logistic Regression Results: The logistic regression model quantifies the effects of various factors, including gender and race, on the length of prosecution for patent applications. In this model, positive coefficients indicate an increased likelihood of a longer prosecution length, while negative coefficients suggest a shorter length relative to the baseline category. This analysis offers insights into how examiner demographics may influence the patent prosecution process and highlights areas where biases may exist. *(Appendix 3)*

- Positive coefficients, such as for 'raceblack' (2.526e-01), suggest that being in the Black or African American group is associated with a higher likelihood of a longer prosecution length compared to the baseline category, which is likely 'racewhite' given its negative coefficient.

- Negative coefficients, such as for 'gender.xfemale' (-1.572e-01), imply that female examiners are associated with a shorter prosecution length compared to the baseline, which may be 'gender.xmale' given its positive coefficient.
- The significance of the p-values, denoted by stars, indicates the strength of evidence that these factors are statistically significant predictors of prosecution length. For instance, 'tenure_days' has a highly significant negative coefficient (-3.982e-07), indicating that with each additional day of tenure, the likelihood of a longer prosecution length slightly decreases.
- The variable 'examiner_art_unit' has a small but positive and highly significant coefficient (3.068e-04), suggesting that differences in art units (which may represent different technological areas) are associated with variations in prosecution length.

This interpretation clarifies our model's findings, relating each variable to the likelihood of longer prosecution length while identifying the direction and significance of each effect.

Interaction Effects Analysis: Further, the study explored interaction effects among examiner art units, gender, and race on application outcomes. This advanced analysis aimed to identify complex interdependencies and nuanced influences that single-factor analyses might overlook. By examining how these factors interact, the study sought to provide a more comprehensive understanding of the dynamics at play in patent application decisions. Our interpretations provide insights into the operational dynamics of the USPTO, revealing how demographic factors might interact with organizational processes. However, these findings should be considered in the context of the broader system and not taken as evidence of causation without further analysis. These results can inform recommendations for addressing potential biases and improving efficiency within the USPTO.

Business Implications & Recommendations

Our empirical analysis has revealed that gender, race, and ethnicity do play roles in patent application processing times and examiner attrition rates. For instance, certain demographic groups might experience systematically different outcomes, such as longer processing times or higher attrition rates.

The regression analysis indicates that specific variables, such as 'raceblack' and 'gender.xfemale', are statistically significant, suggesting that these groups have distinct experiences within the USPTO, which could be attributed to a variety of organizational or systemic factors.

Impact on the USPTO:

- These disparities could affect the USPTO's operational efficiency and the fairness of the patent application process. If certain demographic groups are systematically disadvantaged, it could lead to delays in patent issuance and could potentially discourage innovation among underrepresented inventors.
- High attrition rates, particularly if concentrated in certain demographic groups, can lead to a loss of experienced examiners, increased training costs, and disruption in the patent examination process.

Broader Economic and Social Implications:

- The integrity of the patent system is crucial for economic growth, as it protects intellectual property and encourages investment in research and development. Any perceived biases within the system could undermine public trust and inhibit diverse contributions to innovation.

Strategic Recommendations

- **Bias Mitigation Training**: Implement training programs focused on unconscious bias for all USPTO examiners and staff to mitigate any potential biases in the patent examination process.

- **Diversity and Inclusion Initiatives**: Enhance diversity and inclusion initiatives to ensure a representative and equitable work environment. This could involve revisiting hiring practices, promotion criteria, and providing mentorship programs for underrepresented groups.
- **Process Standardization**: Standardize processes to minimize the impact of individual examiner differences on patent prosecution outcomes. This could involve developing clear guidelines that limit the discretion available to examiners in making decisions on patent applications.
- **Further Research**: Conduct further research to understand the root causes of the disparities identified. This might involve qualitative studies, such as interviews or focus groups, to understand the experiences and challenges faced by examiners of different demographics.
- **Monitoring and Evaluation**: Establish a system for ongoing monitoring and evaluation of patent prosecution processes to continuously assess the impact of implemented changes and the progress of diversity and inclusion efforts.
- **Development of AI Tools**: Leverage AI technology, such as OpenAI's offerings, to assist in the decision-making process, ensuring a data-driven approach that could help reduce the impact of individual biases. These tools should be transparent, explainable, and continuously monitored for fairness and effectiveness.

Part 2: Evaluating & Proposing Cutting-Edge People Analytics Solutions for USPTO Challenges

In Part 2, we evaluated the latest people analytics technologies, focusing on those incorporating generative AI and behavioral science, to tackle patent processing challenges. Our objective was to gauge how these technologies can boost the USPTO's efficiency and equity. After comparing all solutions, we decided iMocha was the best suited for the USPTO’s challenges we identified in our analysis.

Introduction to iMocha's AI-powered Skills Intelligence Cloud

iMocha, originally known as Interview Mocha, is leading the charge in revolutionizing talent management, acquisition, and development with its AI-powered Skills Intelligence Cloud. This platform is dedicated to fostering a skills-first approach within organizations, empowering them to adapt to the changing demands of the workforce. By creating job role taxonomies and customized skills inventories for each employee, iMocha effectively assesses and enhances workforce proficiencies. The platform's strength lies in its multi-channel skills validation and extensive library featuring over 2,500 skill sets, complemented by AI technology, providing vital insights for strategic workforce planning, internal mobility, skills gap management, and employee development.

Leveraging iMocha for the USPTO

The USPTO can utilize iMocha's skills assessment tool to identify skill gaps specific to each art unit. By customizing learning paths based on the needs of the examiners, the USPTO can ensure that all examiners within their art unit have the necessary skills to efficiently process patent applications. The insights the platform provides can also help the USPTO in addressing disparities in patent issuance. They can highlight areas in the USPTO that may suffer more heavily from issues such as implicit biases in certain art units or lacking employee benefit strategies. By correlating these outcomes with changes in patent issuance rates across different demographic groups, the USPTO can develop targeted interventions to address any disparities and ensure a fair and equitable patent examination process. Beyond understanding biases and their effects on decision-making, it’s important that the USPTO is as objective as possible when assigning patent work. iMocha's platform can contribute to creating a more equitable work environment by leveraging their skill-based assessment tool to distribute work objectively.

Construct Validity Concerns
Despite the benefits iMocha offers, some concerns arise from a research design and evidence quality perspective. The platform's reliance on skills validation for Learning & Development programs and skills assessment based on learning paths raises questions about construct validity. Construct validity, crucial for accurate measurement of claimed skills or competencies, necessitates rigorous validation studies to ensure reliability in predicting job performance. Without such validation, there's a risk of misinterpreting the data generated by the platform.

Effectiveness of Personalized Development Paths
While personalized development paths hold promise, their effectiveness hinges on the accuracy of underlying algorithms and data. Algorithms must be based on robust, empirical research accounting for individual learning styles, career aspirations, and job roles. Without this foundation, recommended paths may not align with the most effective developmental interventions for each employee.

Gamification Strategy
The use of certifications and badges for upskilling and employee engagement hints at a gamification strategy. However, evaluating such strategies requires experimental designs to isolate their impact. Implementing these tools without rigorous control groups or randomized controlled trials (RCTs) risks like Hawthorne effects or selection bias, potentially inflating the platform's perceived effectiveness.

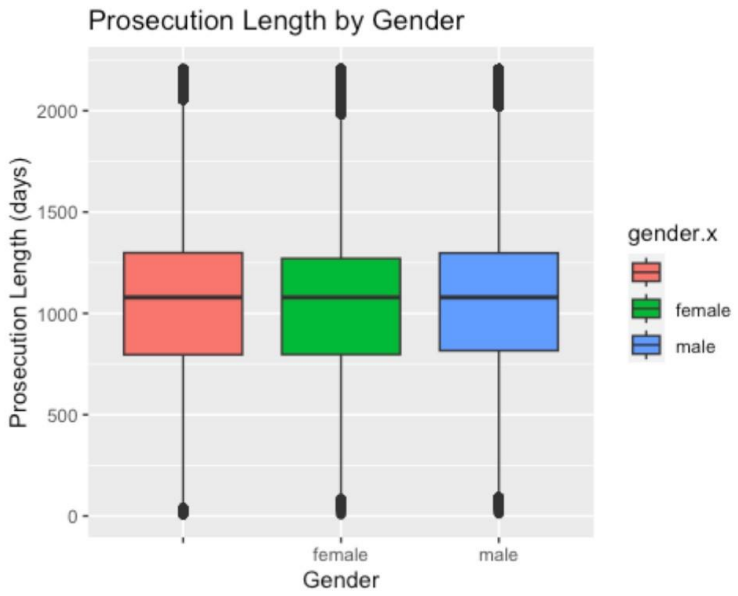
Scrutinizing Predictive Analytics
Predictive analytics features within iMocha's platform, such as forecasting future skill gaps or predicting the success of development paths, must undergo scrutiny for methodological rigor. Biases in training data can lead to inaccurate predictions, and assumptions may not hold true across different organizational contexts or as job roles evolve.

Ethical Considerations in Talent Management
Lastly, the ethical aspects of talent management tools, particularly concerning biases in skill assessments and learning recommendations, demand attention. Data or algorithmic biases can unintentionally exacerbate disparities within an organization, underscoring the need for vigilance.

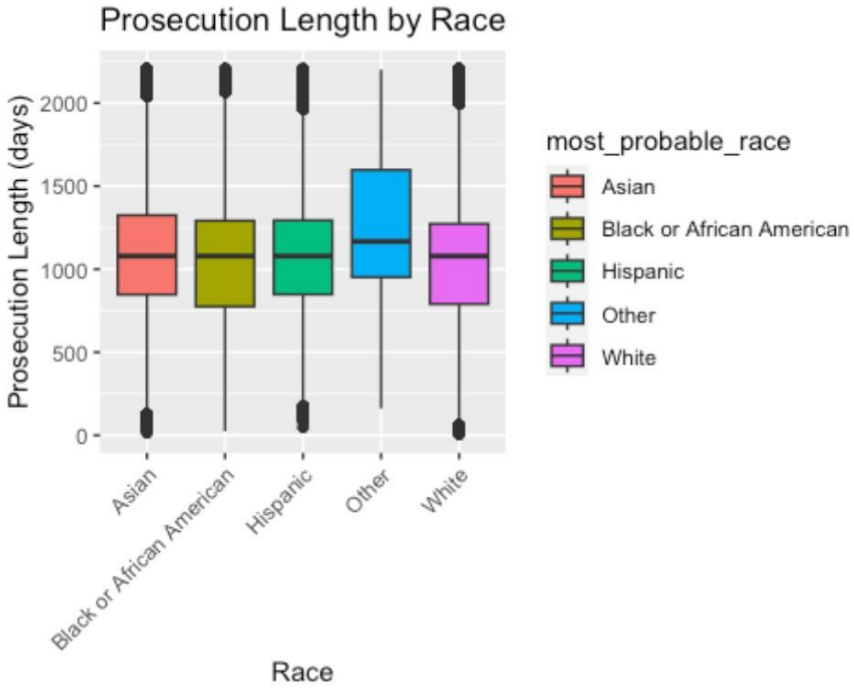
In conclusion, while iMocha's talent management solution offers promising tools, evaluating them requires considering research design and evidence quality. Ensuring effectiveness and fairness demands ongoing validation studies, robust data handling practices, and an awareness of predictive analytics limitations.

Appendix

Appendix 1. Prosecution Length by Gender



Appendix 2. Prosecution Length by Race



Appendix 3. Logistic Regression Results

```
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -4.089e-01  1.110e-02 -36.841  < 2e-16 ***
## examiner_art_unit  3.068e-04  5.085e-06  60.328  < 2e-16 ***
## tenure_days       -3.982e-07  1.771e-08 -22.487  < 2e-16 ***
## gender.xfemale    -1.572e-01  4.849e-03 -32.416  < 2e-16 ***
## gender.xmale       5.963e-02  4.450e-03  13.400  < 2e-16 ***
## raceblack          2.526e-01  7.596e-03  33.260  < 2e-16 ***
## raceHispanic      -1.539e-01  8.968e-03 -17.163  < 2e-16 ***
## raceother          2.226e-01  4.981e-02   4.469 7.86e-06 ***
## racewhite         -6.838e-02  3.486e-03 -19.613  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2627056  on 1900660  degrees of freedom
## Residual deviance: 2613941  on 1900652  degrees of freedom
## AIC: 2613959
##
## Number of Fisher Scoring iterations: 4
```

Based on the output of **summary(glm_outcomes)**, the logistic regression model's coefficients for gender and race can be interpreted as follows:

Gender (gender.x): The coefficients for gender variables, such as 'gender.xfemale' and 'gender.xmale', quantify the impact of an examiner's gender on the probability of a patent application being granted. Specifically:

- A positive coefficient for 'gender.xmale' (+5.963e-02) indicates that male examiners are associated with an increased likelihood of a patent application resulting in issuance, compared to the baseline gender category, which in this case is female due to the negative coefficient for 'gender.xfemale'.
- Conversely, a negative coefficient for 'gender.xfemale' (-1.572e-01) implies that female examiners are associated with a decreased likelihood of patent issuance compared to male examiners.

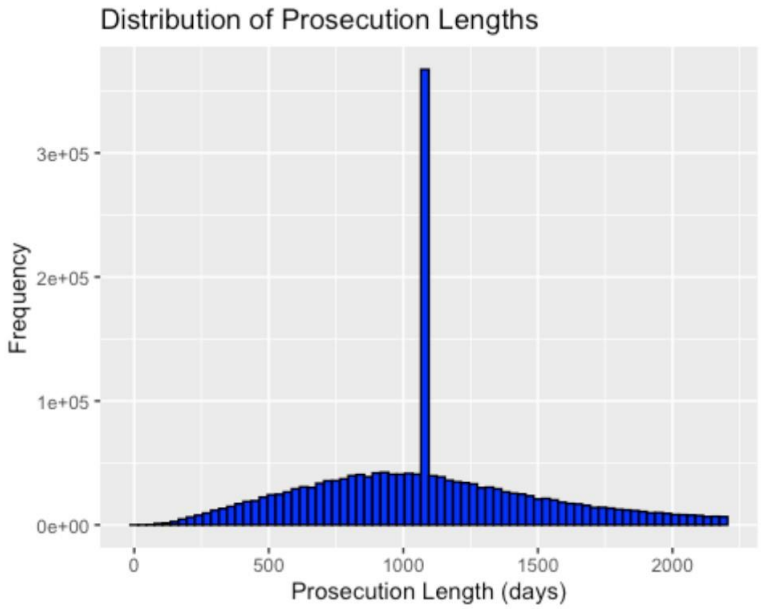
Race (race): The race coefficients reveal how the race of an examiner may influence the patent application process. For example:

- Positive coefficients, such as the one for 'raceblack' (+2.526e-01), suggest that Black or African American examiners have a higher likelihood of the applications they handle being issued compared to the baseline race category.
- A negative coefficient for 'racewhite' (-6.838e-02) suggests that White examiners are associated with a lower likelihood of issuance compared to the baseline category.

These interpretations allow us to understand the direction and magnitude of the influence that gender and race of examiners might have on patent application outcomes, as modeled in this analysis. It is crucial to note that

these interpretations are made within the context of the model's assumptions and the data used; they do not imply causation but rather an association observed in the dataset.

Appendix 4. Distribution of Prosecution Lengths

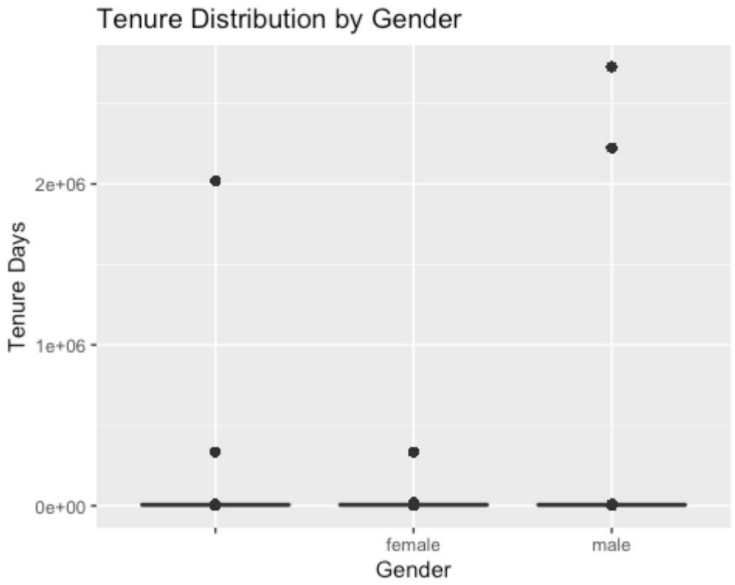


The histogram shows the frequency distribution of prosecution lengths:

- The distribution is highly skewed to the right, with a peak at the lower end of the prosecution length scale. This indicates that a large number of prosecutions are completed within a shorter time frame.
- The x-axis represents the length of the prosecution in days, while the y-axis represents the frequency of cases. The range of prosecution lengths extends to over 2000 days, but the vast majority of cases are clustered under 500 days.
- There is a very sharp peak close to the origin, which suggests that a substantial number of prosecutions are completed in a very short amount of time compared to the rest.
- As the prosecution length increases, the frequency sharply decreases, which is typical of a positively skewed distribution.
- There are very few cases with a prosecution length over 1000 days, as evidenced by the low frequency bars extending towards the right end of the x-axis.

This visualization could be used to analyze the efficiency of case management and identify potential areas of improvement. The peak at the lower end of the prosecution length might indicate a standard processing time for the majority of cases, while the long tail could signal that some cases are taking disproportionately longer to prosecute, which might require further investigation to understand the underlying causes.

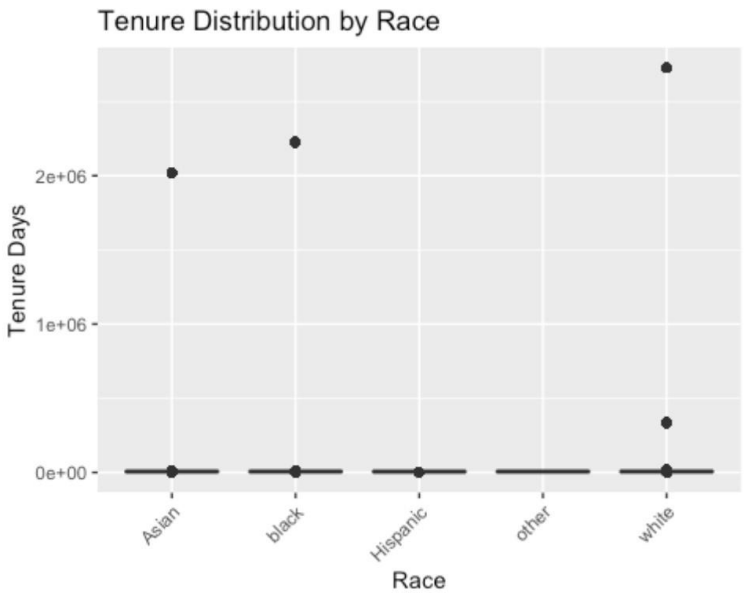
Appendix 5. Tenure Distribution by Gender



- The tenure days are plotted on the y-axis, and they range from 0 to what appears to be over 2,000,000 days, which is likely a data entry error since this number of days would equate to several thousands of years. This could be a case of extreme outliers or incorrect data.
- For both genders, the box plot components—such as the median (the line within the box), the interquartile range (the height of the box), and the whiskers (the lines extending from the box)—are not visible. This absence suggests that the median tenure and the bulk of tenure data for both genders might be very low or close to zero, as the scale of the plot is likely skewed by the presence of extreme values.
- The plot shows individual outlier points for both female and male categories. These outliers are quite distant from zero, indicating that there are tenure values that are much higher than the rest of the data points.

The box plot illustrates the tenure distribution for different genders. Due to the presence of extreme outliers, the detailed distribution for the majority of data points cannot be discerned at this scale. Both female and male categories show outliers with unusually high tenure days, which suggests potential data quality issues. A more detailed analysis, perhaps with the outliers removed or with a log-transformed scale, would be necessary to accurately compare the typical tenure distribution across genders.

Appendix 6. Tenure Distribution by Race



- The tenure days are plotted on the y-axis, with the scale again ranging from 0 to over 2,000,000 days. As with the previous gender-based plot, the extremely high tenure days suggest potential data errors or extreme outliers.
- Similar to the gender plot, the actual 'boxes' of the box plots are not visible, which suggests that the majority of tenure lengths are clustered near the bottom of the scale, close to zero.
- Each racial category shows outliers with unusually high tenure days, significantly distorting the visual representation of the distribution.
- There is no visible difference in the median tenure days between the racial categories because of the scale distortion caused by the outliers.

This box plot aims to show the tenure distribution among examiners of different races. However, due to the presence of extreme outliers, the typical tenure lengths are not distinguishable at this scale. All racial categories display outliers with tenure days much higher than the majority of data points, indicating a need for data cleaning or a more detailed analysis excluding these outliers. Without adjusting for these extreme values, it is difficult to draw any definitive conclusions about differences in tenure distribution across racial groups.

Appendix 7. Further Recommendations

To bolster the effectiveness and efficiency of patent examination processes, we propose the following advanced strategies aimed at enhancing the United States Patent and Trademark Office's operational outcomes.

- **Enhance Data Granularity**: Future analyses could benefit from more detailed data, including specific reasons for patent application delays and rejections, examiner workload details, and more granular demographic data to better understand the impact of diversity on examination outcomes.

- **Incorporate External Factors**: Consideration of external factors such as changes in patent law, technological advancements, and economic conditions could provide additional insights into trends in patent examination times and outcomes.
- **Leverage Advanced Analytical Techniques**: Employing machine learning models to predict examination outcomes and identify factors influencing examiner attrition could offer predictive insights, helping the USPTO to mitigate potential issues proactively.

Appendix 8. Considerations & Limitations

While our analysis provides valuable insights into the patent examination landscape, it is important to acknowledge the constraints and potential biases, which could influence the applicability of our findings.

- **Data Scope**: The dataset may not capture all factors influencing examination outcomes, such as examiner expertise level or applicants' legal representation quality.
- **Potential Bias**: There may be inherent biases in the data, especially concerning gender, race, and ethnicity estimations, which could affect the analysis' accuracy and fairness.
- **Temporal Validity**: The findings are subject to change over time as the USPTO's practices evolve, and external factors influencing patent examination processes vary.

References

Davenport, T. H., Harris, J., & Shapiro, J. (2010). Competing on talent analytics. Harvard Business Review, 88(10), 52-58.

De Bartolo, G., & Stranges, M. (2008). Demography and Turnover. In Applied Demography in the 21st Century (pp. 271-284). Springer.

Salganik, M. (2017). Bit by bit. Retrieved from https://www.bitbybitbook.com/en/1st-ed/preface/

In-Class Exercises

Galperin, R. (2023):

- Exercise 1: More on experiments [Class exercise]. Talent Analytics, McGill University.
- Exercise 2: NAs and data dictionaries [Class exercise]. Talent Analytics, McGill University.
- Exercise 3: DiD, IV, matching [Class exercise]. Talent Analytics, McGill University.
- Exercise 4: Model selection [Class exercise]. Talent Analytics, McGill University.
- Exercise 5: Project work [Class exercise]. Talent Analytics, McGill University.