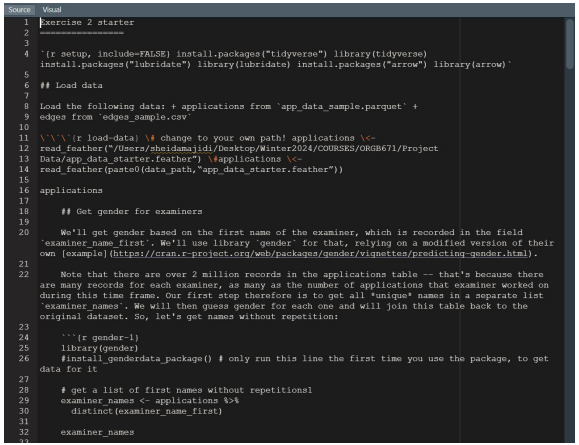
Talent Analytics
ORGB 671

Weekly Exercise #2
In-Class Presentation

Group 3

Joshua Poozhikala
Keani Schuller
Meriem Mehri
Niki Mahmood Zadeh
Sheida Majidi

R Code



```
Source  Visual
1  Exercise 2 starter
2  ==================
3
4  ```{r setup, include=FALSE} install.packages("tidyverse") library(tidyverse)
   install.packages("lubridate") library(lubridate) install.packages("arrow") library(arrow)```
5
6  ## Load data
7
8  Load the following data: + applications from `app_data_sample.parquet` +
9  edges from `edges_sample.csv`
10
11 \`\`\`{r load-data} \# change to your own path! applications \<-
12 read_feather("/Users/sheidamajidi/Desktop/Winter2024/COURSES/ORGB671/Project
13 Data/app_data_starter.feather") \#applications \<-
14 read_feather(paste0(data_path,"app_data_starter.feather"))
15
16 applications
17
18     ## Get gender for examiners
19
20     We'll get gender based on the first name of the examiner, which is recorded in the field
   `examiner_name_first`. We'll use library `gender` for that, relying on a modified version of their
   own [example](https://cran.r-project.org/web/packages/gender/vignettes/predicting-gender.html).
21
22     Note that there are over 2 million records in the applications table -- that's because there
   are many records for each examiner, as many as the number of applications that examiner worked on
   during this time frame. Our first step therefore is to get all *unique* names in a separate list
   `examiner_names`. We will then guess gender for each one and will join this table back to the
   original dataset. So, let's get names without repetition:
23
24     ```{r gender-1}
25     library(gender)
26     #install_genderdata_package() # only run this line the first time you use the package, to get
   data for it
27
28     # get a list of first names without repetitions1
29     examiner_names <- applications %>%
30       distinct(examiner_name_first)
31
32     examiner_names
```

- **Objective:** To analyze patent examiner data for insights into demographics, work patterns, and decision-making processes.
- **Data Preparation:** Data from app_data_sample.parquet and edges_sample.csv are loaded, Libraries like tidyverse, lubridate, and arrow are used for data manipulation and date handling.
- **Gender Identification:** Using gender library on first names from examiner_name_first field.
- Race Estimation: Utilizing wru package to estimate racial demographics based on surnames.
- **Professional Tenure Analysis:** Calculating tenure of each examiner in the organization by determining the time interval between their first and last observed application dates.
- **Quarterly Performance Analysis:** Aggregating data quarterly and examining various performance metrics like number of new, abandoned, allowed, and in-process applications.
- **Statistical Modeling:** Implementing linear and logistic regression models to predict factors influencing examiner turnover and changes in Art Units (AU).
- Utilizing **gtsummary** and **lme4** for model summaries and visualization.

Rationale

- Data-Driven Decisions: Better understanding of the workforce dynamics and decision-making patterns in the examination process.
- Inclusivity & Fairness: Demographic analysis (gender and race) ensures a diverse and equitable work environment.
- Efficiency Improvement: Insights from tenure and quarterly performance analysis can optimize resource allocation and process efficiency.
- Predictive Modeling: Regression models provide predictive insights for better planning and policy-making.

Results Interpretation

- Gender prediction: The gender of examiners was predicted from first names. This method is not foolproof, especially for unisex or culturally diverse names. The predictions should be treated as estimates, not absolute identifications.
- Race estimation: Racial categories were estimated from surnames, a method that has inherent limitations and may not accurately reflect the complex nature of racial identities. These estimations are broad and probabilistic.
- Merging demographic data: Gender and race estimations were merged into the main dataset, adding demographic dimensions. It's important to remember these are based on estimations and carry uncertainties.
- Tenure calculation: Examiner tenure was calculated from the range of observed application dates. This provides a proxy for the length of time examiners have been associated with the organization but may not precisely represent their actual employment period.
- Quarterly data aggregation: The data was transformed for quarterly trend analysis. This approach helps in understanding patterns over time but may miss finer details visible in a shorter time frame.
- “panel_data” for trend analysis: Panel data was created to analyze examiners’ performance and behavior over time. This aggregation allows for a broad view of trends but can generalize individual variations.

Underlying Assumptions (Code)

- The method of predicting gender using the `**gender**` package and race using the `**wru**` package based on names has inherent limitations, as discussed in Holland's "Causation and Race" Report. The accuracy of these predictions may vary due to cultural diversity and the evolving nature of names and racial identities.

- The calculation of tenure using `lubridate` is based on the range of observed application dates. This method provides a useful proxy for understanding an examiner's duration with the organization, but it may not accurately reflect their actual employment period. Implications of this approach on understanding workforce dynamics are explored in Rosenow's article.

- The use of `**gtsummary**` for creating descriptive tables is based on the assumption that summarizing complex data in a comprehensible format enhances the interpretability of the results.

Links to Readings & Course Materials

- The demographic analysis and performance metrics are assumed to contribute to a more inclusive, fair, and efficient work environment. These assumptions are grounded in contemporary organizational theories and practices, as suggested by the foundational principles in the regression readings.

- Quarterly aggregation of production data adds a time element to the analysis, underscoring the value of temporal insights in discerning work trends and decision-making.

- Utilizing regression models to predict turnover and mobility emphasizes the complexity of modeling human behavior and organizational dynamics as described in Biderman's paper on predicting turnover using alternative analytics.

References

Biderman, M. D., Swartout, K. K., Davison, H. K., & Newsome, M. (2003). Predicting Turnover Using Alternative Analytic Techniques. Paper presented at the 18th Annual Society for Industrial and Organizational Psychology Conference, Orlando, FL.

Holland, P. W. (2003). Causation and Race. Research & Development Division, Educational Testing Service, Princeton, NJ.

Rosenow, R. (2016, May 9). Analyzing Employee Turnover - Predictive Methods. LinkedIn. Retrieved from LinkedIn.