**Final Research Paper**

Zachery H. Baldwin

Colorado State University Global

MIS581: Capstone – Business Intelligence and Data Analytics

Dr. Goodwin

May 8, 2022

**Contents**

[Final Research Paper 3](#_Toc101967342)

[Introduction 4](#_Toc101967343)

[Problem Statement & Research Questions 4](#_Toc101967344)

[Proposal 4](#_Toc101967345)

[Overview of Study 5](#_Toc101967346)

[Topic of Research 5](#_Toc101967347)

[Further Research Opportunities 5](#_Toc101967348)

[Research Question and Hypothesis 6](#_Toc101967349)

[Research Questions 6](#_Toc101967350)

[Hypothesis 6](#_Toc101967351)

[Literature Review 7](#_Toc101967352)

[Research Design 8](#_Toc101967353)

[Tools 8](#_Toc101967354)

[Python 8](#_Toc101967355)

[SAS 9](#_Toc101967356)

[2018 MEPS Job Data Description 9](#_Toc101967357)

[Statistical Approaches/Research Methodology 11](#_Toc101967358)

[Limitations 12](#_Toc101967359)

[Ethics 12](#_Toc101967360)

[Week 5 Conclusion 13](#_Toc101967361)

[Findings 13](#_Toc101967362)

[Week 6 Conclusion 18](#_Toc101967363)

[Conclusion 19](#_Toc101967364)

[Recommendations 19](#_Toc101967365)

[References 20](#_Toc101967366)

Final Research Paper

This research project looks at why employees leave their jobs. The below details the proposal of the project as to why the research is important and how the research will be performed. It presents hypothesis for the dataset and gives background information about the data and statistical tests that will be applied. Also, presented below is a data dictionary to model the data, the assumed potential of the results the data may provide, statistical approaches that will be used to help derive insight, and future plans for the analysis of the projects progression. This research project will take a qualitative approach as this tries to deem insight from the data and show why and what causes employee churn in the cases pertaining to this dataset. Qualitative research is subject and tries to gleam insights from data versus measuring objectives (O′Leary, 2021). This has expanded on the previous weeks submissions and added sections for research limitations and more expanded ethical considerations. This also below presents the finding of the research in detail graphs and charts that are explained in detail below.

Introduction

Problem Statement & Research Questions

One major problem in all industries currently is employee turn as we have seen with the current pandemic that is still underway, the need to know how to keep employees has never been greater. What are the factors that drive employee turnover rates? Are there any demographical factors that can allude to when employee and how often employees may churn in industry? What are the risk factors that could identify employee churn? As this project progresses other problems and questions could be identified and arise throughout the research. Thus is the reasoning for attempting an agile approach to this research project, so that it will have the flexibility to change as needed during its progression.

Proposal

When it comes to employees there are four common mindsets insignificant, temporary, exclusive, and career focused (Rogers, 2020). These are the factors that in general go through an employee’s head prior to the beginning of the churn it is very important for businesses and organizations to be able to identify this change in thinking (Rogers, 2020). With natural disasters, pandemics, and general job market uncertainty there has always been an uncertain amount of risk that companies have faced. Risk will always exist for companies, but they will have to eval their reward versus risk prior to making decisions that with effect business situations regarding employee retention. When firms and organizations have field expertise working for them it can be critical to retain such talent and often times it can become very difficult when certain positions are in high demand, effort changes priorities, or polices change (Welch & Brantmeier, 2021). The following 2018 MEPS Job Data surveys attempt to showcase data that can lead to the knowledge of why employees exit careers choices. This will evaluate how the number of years an individual is employed prior to making any changes in their employment to understand what and if any correlation between the years and employment change coincides or has impact with each other. Benefit an organization to understand the rate of retention could potentially provide great insight for employers to understand what actions they may need to take to retain staff.

During data evaluation looking at the number of years of employment and how they either have a correlation affecting employment either directly or indirectly. Using a two-sample t-test and summary statistics this data will be able to show if there is a correlation between years of employment showing it affects employee churn. SAS and Python code will be used to assist in the gathering and evaluation of the data. SAS code examples for loading this job data set will be used from a GitHub repository that was originally written to load a 2002 jobs data file. It was originally used to show the employment numbers in early 2002 (Mitchell, 2019). These files for reading and writing will be used as examples for this project to assist in ensuring the data is correctly interrupted.

Overview of Study

Topic of Research

This research project is based on the topic of employee churn. What are the factors that make employees over time want to change their positions? Could companies not offering raises after a certain time frame, overall unhappiness, changes in management, or years worked ultimately be affecting employee turnover.

Further Research Opportunities

This research is affectively looking at the years worked for employees and trying to see if there are any factors or significance between years worked and employee churn. This topic could be further researched by using surveys via employees to capture feedback and to get their opinions on what factors would make them think about leaving their current position. Also, through research of this nature it can help companies to understand what they may be doing right that keeps their employees happy or shed light on weak areas of the company or things that the employees don’t like about the company that cause them to eventually depart.

Research Question and Hypothesis

Research Questions

Below will discuss the questions that pertain to deriving the hypothesis and that will need to be answered through the research of this project. These are the questions through which the research project will attempt to answer. One major problem in all industries currently is employee churn, as we have seen with the current pandemic that is still underway, the need to know how to keep employees has never been greater. What are the factors that drive employee turnover rates? Are there any demographical factors that can allude to when employee and how often employees may churn in industry? What are the risk factors that could identify employee churn? As this project progresses other problems and questions could be identified and arise throughout the research.

Hypothesis

Hypothesis testing involves a researcher making an opinion of an outcome and then performing statistical tests to understand and try to disprove their opinion/theory. The hypothesis is made up of two opinions the first is called the null hypothesis (H0) and the second is called the alternate hypothesis (HA) (Thakur, 2021). A hypothesis test is as easy as stating the one average value has no variation from 0 and the alternate would be that there is variation in the averages, then compares the scores of the test can tell one if there is any correlation between the variables (Thakur, 2021). This defines a process of decision making that is prepared prior to understanding or determining and outcome (De Santis & Gubbiotti, 2021). This project can have multiple hypothesis but a null hypothesis for this project is that the mean of the years worked pertaining to job change is not significant and the alternate is that it is significant and does provide insight that the number of years worked strongly affects the change in job result. The null would be that the means of the two data sets are equal and the alternate will be that the means vary and are not equal.

Literature Review

Regarding the topic of employee churn, what makes employees leave? Do employees get tired of the position/company, a matter of employees/company satisfaction, fears that technology is going to take over there career, or does it come down to the years worked variable? A problem in the United States (US) is the fact that there are a lot of unskilled workers which are seemingly tied to employee churn with the advances in technology and lack of job training (Vanian, 2020). Another factor that a lot of companies face in the US is burnout, which is a result of companies not having or being able to hire enough skilled workers to fill the positions, so the burden ends up falling on the over worked employees (Irei, 2020). We can see from the company CISCO which is a high demand networking company that is hurting for talented employees due to the high demand of the job and the lack of talent in the market place (Irei, 2020). Based from this it can be seen why understanding what factors go into employee churn is very important. On another note over ~27 million positions were vacated in 2021 in the first quarter. Of these ~9.1 million of the positions were termination and the other ~17.9 million all left of their own regard (Hulsey, 2021). These are some astounding numbers for one year, however data from 2020 & 2020 are kind of an anomaly due to the pandemic at the time that changed the dynamic of the work force. This two year period will make statistical understanding of employee churn very difficult as the period is an outlier. Companies employ persons not with the thought of when are they going to quit or churn but with thought to either increase or grow the business or replace an outgoing employee (Bratsberg et. al., 2021). It can be shown that there is a large need for this type of research as we see the churn within CISOC, European Labor Market, and other US companies. This has shown why this research has the potential to be very important to all industries. Taking the approach of looking at the times at which employees churn seems that it will lead or shed light on other factors that may be happening at the times of the churn.

Research Design

Tools

Below will discuss the tools and statistical methods that will be used in an attempt to showcase the data and research in an understandable way. The tools that will be discussed here at this point in the project will be Python and SAS, this will be the two main tools that will be used for the research. Other tools such as Tableau may be used as a visualization options and also Neo4j may be utilized as a data storage source and a way to represent relationships that lie within the data.

Python

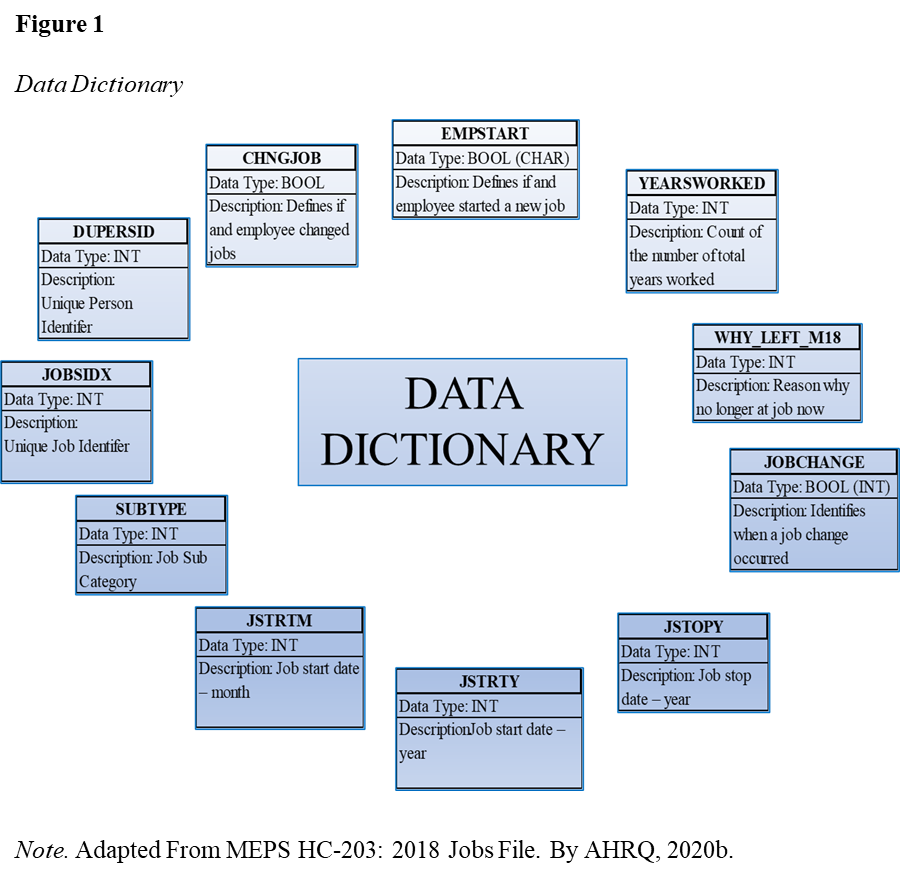
Python is an interpreted programming language with varying uses and applications. This is also one of the most widely used and preferred scripting languages on the market due to its ease of use, flexibility and over all general application to all fields. Python has such toolboxes called modules for statistical models, data cleaning, and data processing. The module that will be used to help clean and organize the data will be the Python Data Analysis Library (PANDAS) library/module. Pandas contains methods for reading/writing data, uses data frames i.e. tables for storage, and provides ample statistical functions and processes (Leong, 2019). Such methods that will be used are reading csv functions, describing statistical summaries, and creating figures to use as visual aids to assist in presenting the data in a more cohesive way for a better take-away they just presenting textual data. This package offers quite a bit of functionality and diversity for file reading and writing and offers the researcher’s options for cleaning filtering and sorting data, along with the ability to have multiple data types in a frame much like a database table. Other modules that Python offers are SCI-Kit Learn a statistical modeling toolbox, Scientific Processing (SCIPY), and Numerical Processing (NUMPY).

SAS

This software will be used to perform the statistical tests to test the future developed hypothesis for accuracy and validity. It will also be used to clean and sort the data into workable forms and to cut down on the amount of data that is being processed. Much like what will be done in Python, both applications have their own strengths of which will be levied and utilized throughout this research project. Using SAS to perform the different statistical tests that will be applied, of which are listed below under the statistical approaches section. SAS offers ease of use when it comes to testing statistical processes and applying them to data. It allows a user to not have to perform the down in the weeds calculations and allows the users to spend more time interpreting the results versus coding the entire processing by the use of its built in functions.

2018 MEPS Job Data Description

This section is carried forward from the project proposal. This section details out what the data is, where the data came from and how the data is organized. Data that will be used is 2018 employment data that was used for a multi-year study to focus on why employees end employment. 2018 job data provided by the Household Components (HC) MEPS all identifying personal information has been taken out. This data contains multiple variables, definitions and has been assessed as a complete collection of adequate data to be able to provide an accurate evaluation. The HC of the survey contains demographic information that will be detailed later through the project (AHRQ, 2020a). This data set is gathered during a two year period containing interviews of 5 rounds that were performed to capture information such as health and employment data (AHRQ, 2020a). This HC data collected through interviews, one individual pertaining to one household was questioned by using computer assisted personal interviewing techniques (CAPI) (AHRQ, 2020a). This survey process and use of computer aided software was of great benefit to analysts and the individuals who collected the data. CAPI allowed faster access to data and assisted greatly in the storing of the information, so that data allowing for archiving for later use. 2018 MEPS jobs data file contain to 3, 4, & 5 rounds of 22 panels and 1, 2, 3 rounds pertaining to panel number 23 for surveying the panels containing 5 rounds within a two-year duration (AHRQ, 2020a). This data covers the lengths of three interview rounds for observations pertaining to each panel. All data variables are laid out in detail in the MEPS HC-203 2018 Jobs File Codebook (2020b), this gives all variable descriptions and valid identifiers for the data that is being worked with throughout this project. In the below Figure 1 presents the data dictionary.



Statistical Approaches/Research Methodology

Statistical approaches are used to assist in interrupting information and for deriving insights from data, along with being able to provide a way to reproduce the study if necessary (Yu et. al., 2022). The statistical approaches that will be used are the t-test, summary statistics, analysis of variance (ANOVA), along with linear/logistic modeling. These approaching will assist in identifying relationship that may reside inside the data and add weight to variables that appear to have a correlation with one another. Thus potentially identifying factors that may represent flags that could potentially help to identify employee churn. By using these statistical methods and evaluating the years of employment prior to change we should be able to identify how strong that years worked may identify a job change. This would be able to assist companies when they would need to set anniversary dates, bonuses, and raises which all could assist in employee retention, this is a very simplified simple example.

Limitations

Limitations throughout this project were data and time. Time is said as a limitation due to this being a short turn around research project, thus the reason for only looking into the few variables and hypothesizes that were researched. Time being defined as the amount of hours in a day on top of a full time 40-45 hour a week engineering position on top of a 10 hour a week toward a small business intuitive research (SBIR) project, on top of performing the research for this project in competition of the Master’s program. This is not an excuse for anything just provided as for a more understanding of why the research doesn’t go more in-depth than it does. Another limitation that a lot of research projects face is a lack of data, this can be due to access limitations or data just not having been cataloged in enough detail. Other than time and data as limitations incorrect sample sizes play a big role in the lack of data and could contribute to a loss in time (Dorji, 2020). These can be overcame by understanding and managing the time that one does have and schedule things out accordingly and not biting off more than one can or not being to over confident in the limited time one has (Dorji, 2020).

Ethics

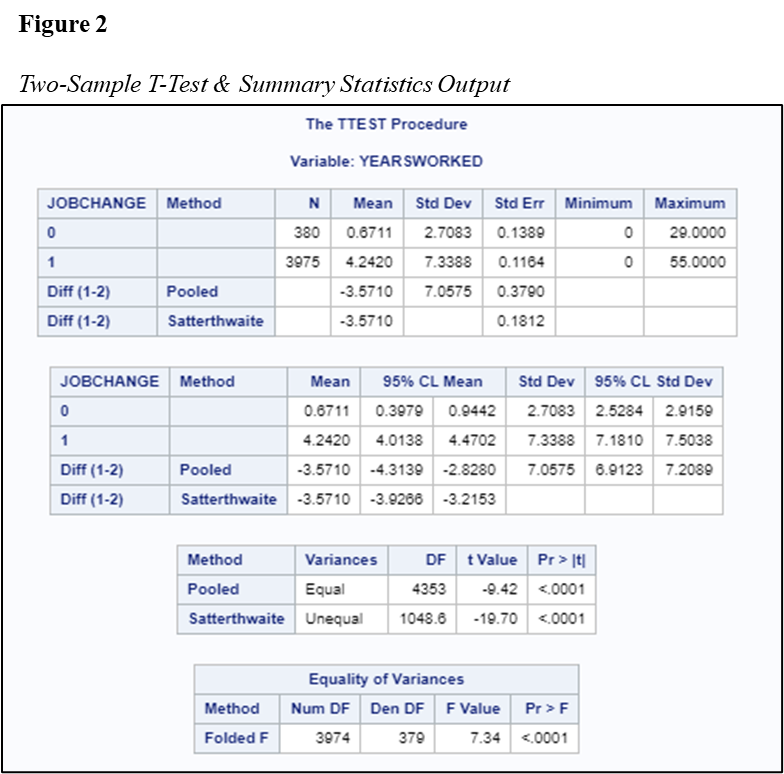
Protecting the rights of people and data should be at the top of the list. This can be done by ensuring that data collection is performed in a fair and just way, showing respect for personal data, keeping things unbiased, and being transparent (Klose et. al., 2020). Ethics & privacy protection boils down to the behavior of the researchers or data retrievers and how they view the data that is being retrieved. Data should always be understood as to its type and protected accordingly so that rights of individuals are upheld and not taken for granted. This data has been modified to protect the privacy of the individuals and personal information of the study participants. Things like name and age have been removed to keep the study unbiased toward any specific demographic and replaced with unique identifiers (AHRQ, 2020a). Researchers should always hold themselves to elevated standards when it comes to ethics such as posted by the Institute of Electrical and Electronics Engineers (IEEE). Example ethical responsibilities are integrity, reliability, professional conduct, and to strive to protect the privacy and the safety of other individuals (IEEE, n.d.).

Week 5 Conclusion

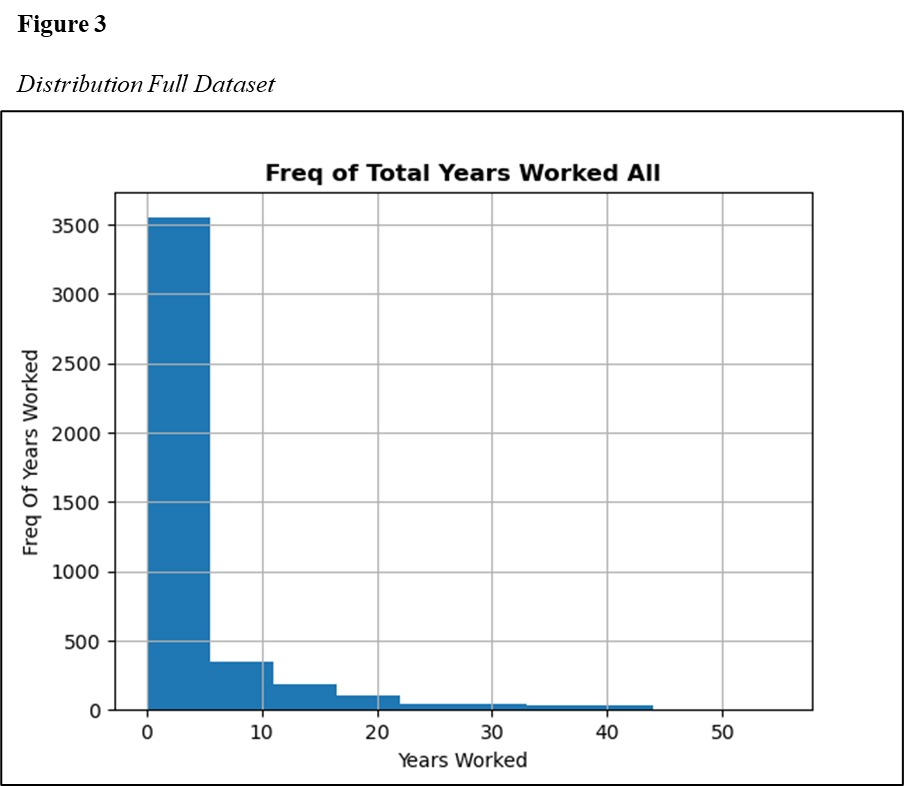
In conclusion this is the first draft of the final research document, this document pulls together the work that has been accomplished thus far through the progression of this course and research project. This project has been organized into the formatted project write-up document as requested and design for the completion of this research project. This has provided the information that will be presented in the final project with any updates thus far. Also, place holder sections have been implemented that will be for the upcoming future submissions.

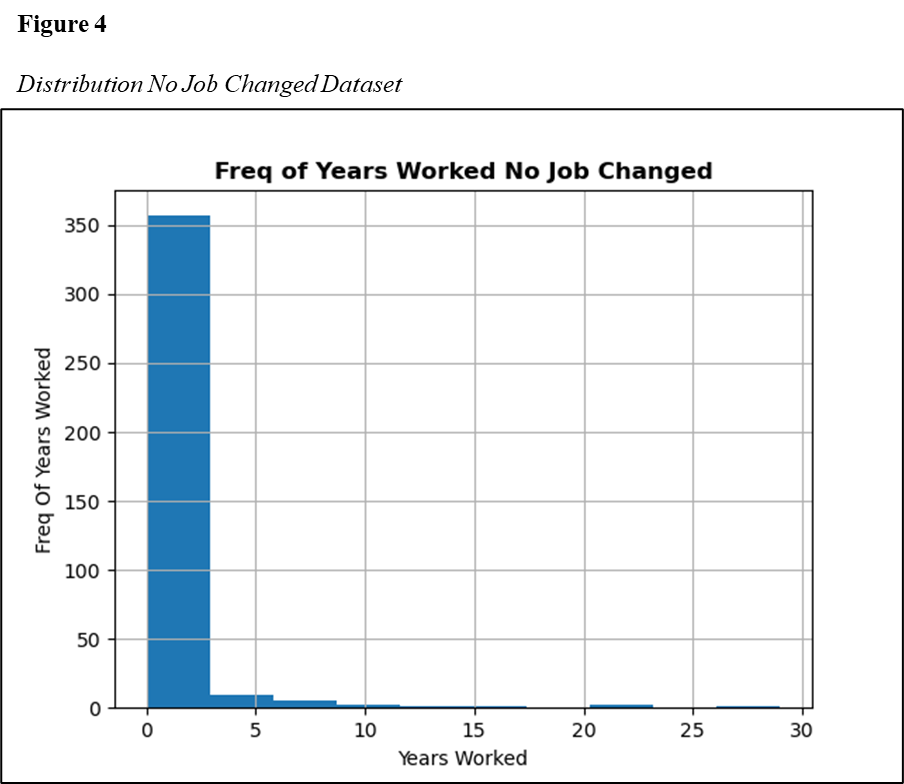
Findings

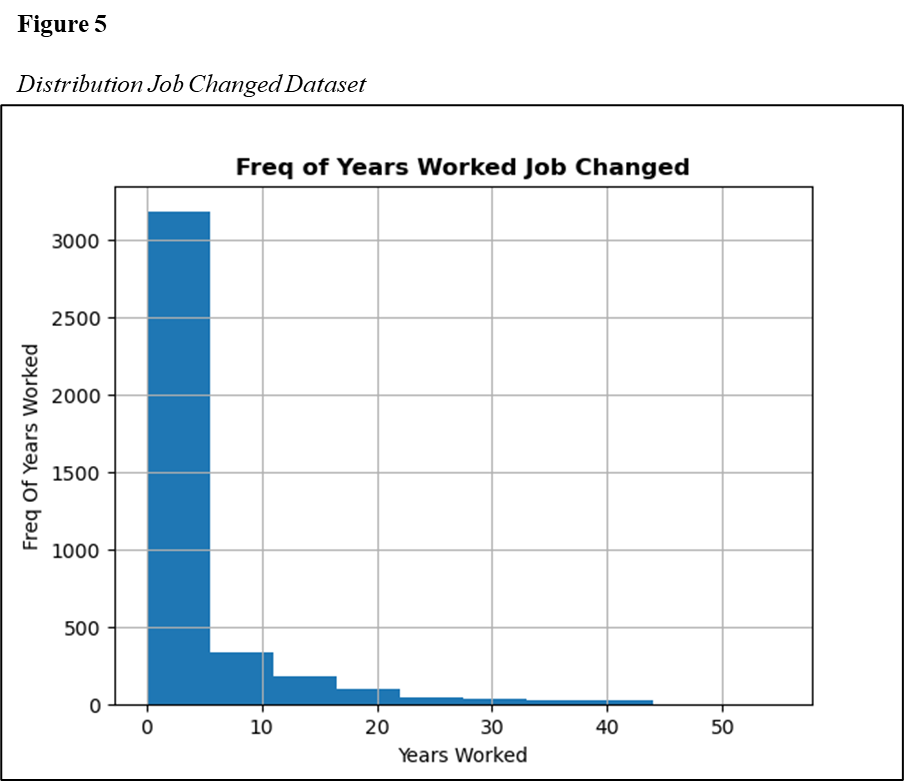
Figure 2 below shows the results from the two-sample t-test. When examining the hypothesis of equal variance an F test should be used, this test will be used to evaluate the p value compared to the significance value, which is 0.05. If Figure 2 is examined, it can be seen that the p value for the F test is much less than the significance value which indicates that a relationship appears to exist between years worked and job change. Then it can be stated that the null hypothesis can be rejected and stated that there is enough variance shown in the means to imply that years worked and change in position show a strong correlation.



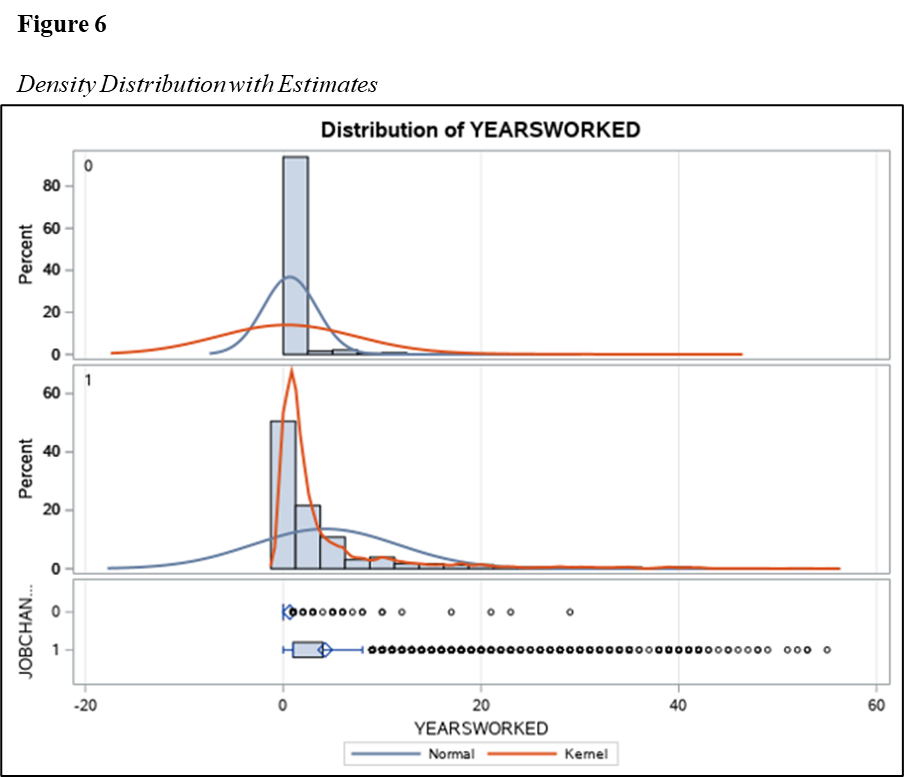
Figures 3, 4, & 5 below show the frequency distributions of years worked. Through reviewing these figures larges variance in the distribution can be seen between years worked with no job changed and years worked with job change. Specifically looking at the no job change Figure 4 and the job changed Figure 5 focusing on the 0-5 years period it shows that there is a very high rate at which jobs are changed. This is a focal area that would require more research and more data that would be able to provide great insight into why jobs are changed so quickly once started.



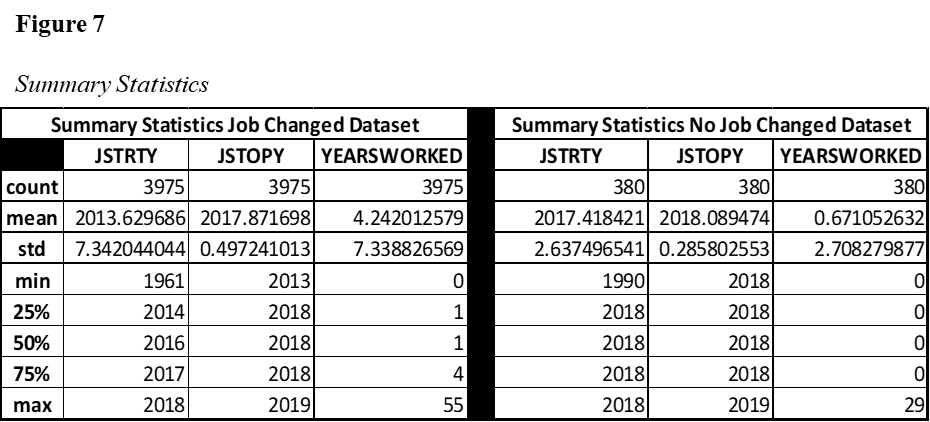




These results could have been influenced by many factors such as age. Age could indicate that the individuals were continuing through school and were changing careers paths at the time. There are many variables that can factor into employee churn, and it could be different for every industry and every person. Through further research of looking and reviewing more historical data, more patterns for variable, such as, age, career dislike, and/or career improvement could be evaluated through employee surveys. Surveys like this that bring in the employee would be more engaging for some and irritating for others but gain valuable insight into the minds of the employees. One other figure, Figure 6 provided insight into this data set are the density distributions estimator figure that shows the density estimation side by side and provides whisker figures that show how spread out the data becomes as the years grow and less employee’s churn.



It can be seen in Figure 6 that neither dataset follows a normal density estimation. The above shows two right skewed density distributions, if these both would have been normalized distributions the null hypothesis could have been accepted. The Kernel density estimation try’s to provide information toward a populations characteristics from a bounded dataset (DeepAI, 2020). Then if Figure 7 is evaluated looking at the summary statistics the distributions seem off as there is a lot more of a population under ~1 year captured for employees not changing positions. Hence the recommendation for analysis further to be performed on a much larger scale and more historic datasets. While it does appear that the number of years worked does has a significant relationship to employee churn, it would be more evident if more factors and data were analyzed.



Week 6 Conclusion

In conclusion this is the second draft of the final research document, this document pulls together the work that has been accomplished thus far through the progression of this course and research project. This week has added the findings and results from the research that was completed through this course. Thus further detailing the research and progression of this project. Also, the determination of the status of the hypothesis has been evaluated and presented. It has been shown through the visual aids that this data may not have enough distribution in the data to accurately depict the affect that years worked has on employee churn. It would be recommended for extended research to be completed with this data accompanied with more and additional historical data.

Conclusion

In conclusion, this project has been completed using statistical methods that have been learned over the years in this data analytics program. This project shows the culmination of the research and work that has been done throughout this capstone course. This paper has brought forth the results of the research and the evaluation of the hypothesis. Also, this has presented different visualization of the data in different ways to understand visually how the data is spread and also through textual summary statistics. Below are future recommendations for research into this topic going forward.

Recommendations

For future research and going forward it would be recommended to compare more variables, along with years worked to understand what other variables contribute strongly to terminating employment. Taking into account the facts of when the most people change employment or start to search for new opportunities, companies should work on understanding what they can do to thwart the churn. Things in the nature of incentives to maintain their position and place in the company. A good way that has been experienced in the past in retaining employees is work environment and climate. So it would be recommended to evaluate the climate in the office and understand what employees are and are not happy with regarding operation on a daily basis.

References

AHRQ. (2020a). *MEPS HC-203: 2018 Jobs File*. Retrieved March 23, 2022, from https://meps.ahrq.gov/data\_stats/download\_data/pufs/h203/h203doc.pdf

AHRQ. (2020b). *MEPS HC-203 2018 Jobs File Codebook*. March 23, 2022, from

https://meps.ahrq.gov/data\_stats/download\_data/pufs/h203/h203cb.pdf

Bratsberg, B., Raaum, O., & Røed, K. (2021). *Excess churn in integrated labor markets*. Journal of Population Economics, 34(3), 865–892. https://doi.org/10.1007/s00148-020-00795-1

De Santis, F., & Gubbiotti, S. (2021). *On the predictive performance of a non-optimal action in hypothesis testing*. Statistical Methods & Applications, 30(2), 689–709. https://doi.org/10.1007/s10260-020-00539-1

DeepAI. (2020). *Kernel density estimation*. https://deepai.org/machine-learning-glossary-and-terms/kernel-density-estimation

Dorji, J. (2020). *Facilitating students overcome their limitations through doing their own action research*. Theory and Practice in Language Studies, 10(10), 1240. https://doi.org/10.17507/tpls.1010.08

Hulsey, L. (2021). *Job churn includes more workers quitting by choice than being laid off*. Journal-News, The (Hamilton, OH).

IEEE. (n.d.). *IEEE code of ethics*. Institute of Electrical and Electronics Engineers. Retrieved April 5, 2022, from https://www.ieee.org/about/corporate/governance/p7-8.html

Irei, A. (2020). *CISO position burnout causes high churn rate*. Information Security, 21(2), 1–3.

Klose, M., Desai, V., Song, Y., & Gehringer, E. (2020). *EDM and privacy: Ethics and legalities of data collection, usage, and storage*. International Educational Data Mining Society.

Leong, N. (2021, December 12). *Python for data science - A guide to pandas - Towards data science*. Medium. Retrieved March 31, 2022, from https://towardsdatascience.com/python-for-data-science-basics-of-pandas-5f8d9680617e#:%7E:text=Python%20for%20Data%20Science-%20A%20Guide%20to%20Pandas,Cleaning.%20. . .%2010%20GroupBy.%20. . .%20More%20items. . .%20

Mitchell, E. (2019). *EM2.sas*. Retrieved March 23, 2022, from https://github.com/HHS-

AHRQ/MEPS/tree/master/SAS/older\_exercises\_1996\_to\_2006/Employment\_examples/EM2/EM2.sas

O′Leary, Z. (2021). *The essential guide to doing your research project* (Fourth ed.). SAGE Publications Ltd.

Rogers, C. (2020). *Retention: Key mindsets that retain top talent*. HigherLife Publishing.

Thakur, M. (2021). *Hypothesis testing formula*. EDUCBA. Retrieved April 6, 2022, from https://www.educba.com/hypothesis-testing-formula/

Vanian, J. (2020). *When it comes to A.I., worry about “job churn” instead of “job loss.”* Fortune.Com, N.PAG.

Welch, L., & Brantmeier, N. K. (2021). *Examining employee retention and motivation trends in research administration*. Journal of Research Administration, 52(2), 70–86.

Yu, Z., Guindani, M., Grieco, S. F., Chen, L., Holmes, T. C., & Xu, X. (2022). *Beyond t test and*

*ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research*. Neuron, 110(1), 21–35. https://doi.org/10.1016/j.neuron.2021.10.030