Data Analytics Capstone Topic Approval Form

Student Name: Natalie Toler

Student ID: 011148248

Capstone Project Name: Chi-squared Test of Independence for Business Entity Dataset

Project Topic: Analyzing the survival rate of businesses registered in Colorado using chi-squared test of independence.

X This project does not involve human subjects research and is exempt from WGU IRB review.

Research Question: Is there a significant difference in survival rates among different businesses registered in Colorado?

Hypothesis: **Null hypothesis**(H_0): - There is no significant difference in survival rates among different businesses registered in Colorado.
Alternate Hypothesis(H_1): - There is a significant difference in the survival rates among different businesses registered in Colorado.

Context: This study contributes to the field of data analytics by analyzing factors contributing to business failure, focusing on Colorado startups since 2000. The choice of business structure is crucial for accessing tax credits and funding opportunities (U.S. Small Business Administration, 2024), and inadequate funding and lack of entrepreneur planning are common causes of failure linked to this choice (Amankwah-Amoah & Wang, 2019). The analysis aims to assess the relationship between business entity types and success rates, enhancing the understanding of business success determinants.

The Chi-Square Test is a robust statistical method used to compare categorical variables and assess the independence between groups (Dishant Salunke, 2024). Previous research has demonstrated the effectiveness of Chi-Square Tests in analyzing business survival rates across different industries and business forms (Kakkad, 2017). This study will help in understanding how business structure affect long-term success, contributing valuable insights to the field of entrepreneurship and business development.

Data: The data for this project is sourced from the Colorado Information Marketplace, an open-source data catalog. The dataset, titled "Business Entities in Colorado," was contributed by the Colorado Department of State and contains records of business entities from 1886 to 2024, with 2.71 million rows and 35 columns prior to cleaning (Colorado Information Marketplace, 2018). This dataset is maintained by the Colorado Information Marketplace and the office of the Colorado Secretary of state. The data is reliable, usable, and durable, having been made accessible for the purpose of solving business challenges and providing business insights (Business Intelligence Center, 2024).

https://data.colorado.gov/Business/Business-Entities-in-Colorado/4ykn-tg5h/about_data

Field	Data Type
entityid	Categorical
entityname	Categorical
principaladdress1	Categorical
principaladdress2	Categorical
principalcity	Categorical
principalstate	Categorical
principalzipcode	Categorical
principalcountry	Categorical
mailingaddress1	Categorical
mailingaddress2	Categorical

mailingcity	Categorical
mailingstate	Categorical
mailingzipcode	Categorical
mailingcountry	Categorical
entitystatus	Categorical
jurisdictonofformation	Categorical
entitytype	Categorical
agentfirstname	Categorical
agentmiddlename	Categorical
agentlastname	Categorical
agentsuffix	Categorical
agentorganizationname	Categorical
agentprincipaladdress1	Categorical
agentprincipaladdress2	Categorical
agentprincipalcity	Categorical
agentprincipalstate	Categorical
agentprincipalzipcode	Categorical
agentprincipalcountry	Categorical
agentmailingaddress1	Categorical
agentmailingaddress2	Categorical
agentmailingcity	Categorical
agentmailingstate	Categorical
agentmailingzipcode	Categorical
agentmailingcountry	Categorical
entityformdate	Categorical

Limitations: The dataset contains missing values, particularly in fields not applicable to all registered businesses. Additionally, the principal address is not always guaranteed to be the physical location of the business's main operation. The dataset is actively updated, meaning that the data downloaded may change over time (*Colorado Information Marketplace Data Changes*, 2024).

Delimitations: The dataset will be delimitated by removing records that have missing information in the Entity Type and Entity Status since these are the variables of interest. Additionally, columns for variables that are not relevant to the study will be removed to simplify the dataset to only the variables that will be used for the analysis and reporting. The study will also only be using records where entity form date is 2000 to 2024. Analyzing business structure on business success rates will add further nuance to the growing research on business success and failures (Zambrano Farias et al., 2021).

Data Gathering: The dataset will be downloaded as a CSV file from the Colorado Information Marketplace website. The data contains categorical variables stored as text strings. The "Entity Form Date" will be transformed into the datetime data format. All records earlier than 1/1/2000 will be dropped. Columns for variables that are not relevant to the study will be dropped, resulting in 10 remaining columns. If there are rows with null values in the columns "entityformdate", "entitystatus" and "entitytype" will be dropped to maintain data integrity (Ngugi, 2022). The overall data sparsity is 40.97% however the data sparsity with irrelevant columns removed is 1.28%. The cleaning process will be conducted using Python.

Data Analytics Tools and Techniques: Exploratory data analysis will be conducted to understand trends over time and across business types, guiding the structure of the statistical analysis. Only entities formed since 2000 will be included in the analysis, this choice is supported by data collected by the U.S. Bureau of Labor Statistics' Business Employment Dynamics data which shows that 20% of business fail in the first two years, 45% in the first five years, 65% in the first ten years, and only 25% of businesses will make it to the 15-year mark (Deane, 2022). Therefore using 24 years of data will account for the overall statistics of business success, this additionally gives padding for historic events that have affected business creation, such as the Covid-19 pandemic which led to a surge of startups (Haltiwanger, 2021).

With the appropriate variables and time frame a chi-square test of independence will be run. With entity status noting whether a business entity is still in good standing and entity type noting the structure of the business, e.g., LLC, Corporation, Sole proprietorship. The p-value and chi-square statistics will be analyzed to identify consistent patterns or significant differences in survival rates over time (Arshad, 2023). The chi-square test will also be performed for entity status and principal city, principal state, and jurisdiction of formation with the results compared to the results of entity type to further analyze the significance of entity type on business success.

The presentation of results will make use of tableau's visualizations to illustrate the significance of the results, as well as visualizations of the data itself to add context to the results.

Justification of Tools/Techniques:

- Python: Python is a versatile programming language with extensive libraries for data analysis, making it better than the programming language R for business-focused analyses (Learning Python for Data Analysis, 2021). Python is also open source and scalable, making it a better language than SAS for this analysis (*SAS vs R vs Python*, 2023).
- Jupyter Notebook: This tool allows for the integration of comments and code, creating a clean and readable format for analysis (Shafi, 2023).
- Tableau: Tableau will be used for visualizing the results, making the findings accessible and easy to interpret.

Project Outcomes: This analysis aims to build on previous studies of business lifecycles. These studies include Business Survival in the Construction Industry in Relation to Other Businesses which similarly uses the chi-square test to show that construction companies have the lowest survival rates (Kakkad, 2017). Analytics from The Economics Daily on business establishments still in operation by industry (*34.7 Percent of Business Establishments Born in 2013 Were Still Operating in 2023*, 2024). Support for the alternative hypothesis is found in (Worku, 2013) which found through the chi-square test that businesses failed due to lack of initial capital, failure to utilize finance in according with the business plan, among other significant financial and educational reasons.

Projected Project End Date: October 20, 2024

Sources:

Amankwah-Amoah, J., & Wang, X. (2019). Business Failures around the World: Emerging Trends and New Research Agenda. *Journal of Business Research*, *98*, 367-369. Retrieved 24, 2024 from https://doi.org/10.1016/j.jbusres.2019.02.064

Colorado Department of State (2014, March 19). *Business Entities of Colorado*. Colorado Information Marketplace. Retrieved September 24, 2024, from https://data.colorado.gov/Business/Business-Entities-in-Colorado/4ykn-tg5h/about_data

Colorado Secretary of State (n.d.). *Business Intelligence Center*. Retrieved September 24, 2024, from https://www.sos.state.co.us/pubs/BIC/home.html

Colorado Department of State (n.d.). *Colorado Information Marketplace Data Changes*. Colorado Information Marketplace. Retrieved September 24, 2024, from https://data.colorado.gov/stories/s/pbek-aaa3

Columbia University: The Fu Foundation School of Engineering and Applied Science (n.d.). *Learning Python for Data Analysis*. Columbia Engineering Bootcamps. Retrieved September 24, 2024, from https://bootcamp.cvn.columbia.edu/blog/learning-python-for-data-analysis/

Deane, M. T. (2024, June 1). *Top 6 Reasons New Businesses Fail*. Investopedia. Retrieved September 24, 2024, from https://www.investopedia.com/financial-edge/1010/top-6-reasons-new-businesses-fail.aspx

Decoding Data Science (n.d.). *Understanding the Chi-Square Test: A Comprehensive Guide*. Decoding Data Science. Retrieved September 24, 2024, from https://decodingdatascience.com/chi-square-test-learn/

Geeks for Geeks (2023, December 5). *SAS vs R vs Python*. Retrieved September 24, 2024, from https://www.geeksforgeeks.org/sas-vs-r-vs-python/

Haltiwanger, J. C. (2021). Entrepreneurship During the COVID-19 Pandemic: Evidence from the Business Formation Statistics. *National Bureau of Economic Research*. https://doi.org/10.3386/w28912

Kakkad, S. A. (2017). *Business Survival in the Construction Industry in Relation to Other Businesses: A Comparative Analysis* [Master's Thesis, Texas A&M University]. Electronic Theses, Dissertations, and Records of Study (2002–). Retrieved September 24, 2024, from https://hdl.handle.net/1969.1/161347

Ngugi, J. (2022, May 21). *Handling Missing Values - Data Science*. Medium. Retrieved September 24, 2024, from https://ngugijoan.medium.com/handling-missing-values-data-science-7b8e302264ee

Salunke, D. (2024, March 31). *Understanding the Chi-Square Test: A Comprehensive Guide*. Medium. Retrieved September 24, 2024, from https://medium.com/@dishant.salunke9/understanding-the-chi-square-test-a-comprehensive-guide-f3bece83b920

U.S. Small Business Administration (2024, August 18). *Choose a Business Structure*. U.S. Small Business Administration. Retrieved September 24, 2024, from https://www.sba.gov/business-guide/launch-your-business/choose-business-structure

Worku, Z. (2013). Analysis of Factors That Affect the Long-Term Survival of Small Businesses in Pretoria, South Africa. *Journal of Data Analysis and Information Processing*, *1*(4), 67-84.Retrieved September 24, 2024, from https://doi.org/10.4236/jdaip.2013.14008

Zambrano Farias, F., Valls Martínez, M. D., & Antonio, P. (2021). Explanatory Factors of Business Failure: Literature Review and Global Trends. *Sustainability*, *13*(18), 10154. Retrieved September 24, 2024, from https://doi.org/10.3390/su131810154

Course Instructor Signature/Date:

☒ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: 9/25/2024

Reviewed by:

Comments: Click here to enter text.