ETL Report

Group 1:

Hayden Muscha, Kripanjali Dhungana,

Nathan Van Schyndel, Zheding Zhao

ETL Process Date: 7/16/2022

Introduction

The dataset that was given to us for this project was the census data which can be found at census.gov. While exploring the vast dataset, our group spent a majority of time understanding the dataset and brainstorming ideas and created hypothetical questions that we deemed interesting to conquer. While we looked through the datasets, we decided to use two specific datasets as they looked interesting and helped us find answers to our questions. However, we needed to transform our data first to get to the answers we were looking for which we will talk more about in the transformation section. Our interest specifically generated on looking at the technological usage in different industries which also led to linking salaries that technical employees make across the industries. Below are some of the areas we want to test out:

- Is there a specific kind of technology that certain industries prefer and not prefer?
- Industry specific comparison between employees and their payroll.
- Technology specific comparison between average pay and industries.
- Did the increased use in technology improve the employee salary?
- The type of technology that industries are willing to get into/testing phase.
- Do employees get paid more as their companies implement more technology?

Data Sources

We found our datasets on www.census.gov. We primarily accessed this on Friday morning, 7/15/2022 and have been accessing the data throughout the weekend as per our needs. Below is the website:

Bureau, U. S. C. (2021, October 14). *Annual Business Survey (ABS) APIS*. Census.gov. Retrieved July 15, 2022, from https://www.census.gov/data/developers/data-sets/abs.2019.html

Extraction

The data was retrieved from the US Department of Commerce as provided for this module's assessment. Below are the steps taken to retrieve the data:

1. We visited www.census.gov.
2. We explored the various datasets.
3. We decided to particularly work with the datasets Company Summary and Technology Characteristics of Business.
4. We extracted the data using a web API key.
5. We converted the data into json format.

Transformation

We did not use all of the data in the form we extracted. We used two separate datasets that we merged after we extracted them. Our transformation process was focused on our hypothetical questions that we wanted to find answers for.
Our GitHub has the file Census_Data.ipynb where our transformation process took place and is numbered in steps.

Our Github has the file Census_Data.ipynb where our code to extraction and transformation can be found.

Load

After transforming the data in the Jupyter Notebook, we loaded our data into Python's Integrated Development Environment (IDE): (Visual Studio Code (VSC)) where we performed our visualization analysis and process.

Conclusion

To conclude our ETL report, we initially started by exploring the datasets in the US Department of Commerce from which we chose two particular datasets, the first one was company summary and the second one was the technological characteristics of businesses. From these datasets we focused on areas such as levels of technology usage, average salary, industry wise technology preference, employee compensation and so on. After the process of extraction, transformation and loading, we prepared visualizations to answer our questions and queries described in the introduction section of this report. The answers will be discussed more in our detailed project report.