**ETL Report Guide**

Greg Wagner, Sawyer Tucker, Daniel Brickman, and Chris Nash  
ETL Process 1/14/2022

**Introduction**

As the new millennium as gone on, one concerning trend that many people have noticed is the increasing wealth disparity. This, coupled with our more sophisticated understanding of and increasing calls for social justice and equity, prompted several questions from us. There are four major questions that we are concerned with, they are as follows:

1. What is the demographic breakdown of business owners?
2. Which demographic has the most employees, and how do the demographics compare to each other?
3. Do the differences in demographics lead into any disparities in metrics such as payroll?
4. What are the differences in technology uses between different demographics of businesses?

To solve these questions, we used US census data from the year 2019. There are four main tables that we referenced, called Company Summary, Characteristics of Businesses, Characteristics of Business Owners, and Technology pf Businesses. As the data stands, however, it is clunky and difficult to use. There is a lot of information that does not concern us wrapped up in the data, and we also wish to find aggregate data. Therefore, we need to clean our tables.

**Data Sources**

All of our data was accessed originally from the census.gov website, which acts as a sort of hub. All sources were originally accessed on 1/12/2022.

US Department of Commerce. (n.d.). 2018 - census.gov. Annual Business Survey (ABS) APIs. Retrieved January 12, 2022, from https://api.census.gov/data/2018/abscs?get=GEO\_ID,NAME,NAICS2017,NAICS2017\_LABEL,SEX,SEX\_LABEL,ETH\_GROUP,ETH\_GROUP\_LABEL,RACE\_GROUP,RACE\_GROUP\_LABEL,VET\_GROUP,VET\_GROUP\_LABEL,EMPSZFI,EMPSZFI\_LABEL,YEAR,FIRMPDEMP,FIRMPDEMP\_F,RCPPDEMP,RCPPDEMP\_F,EMP,EMP\_F,PAYANN,PAYANN\_F,FIRMPDEMP\_S,FIRMPDEMP\_S\_F,RCPPDEMP\_S,RCPPDEMP\_S\_F,EMP\_S,EMP\_S\_F,PAYANN\_S,PAYANN\_S\_F&for=us:\*

US Department of Commerce. (n.d.). 2018 - census.gov. Annual Business Survey (ABS) APIs. Retrieved January 12, 2022, from https://api.census.gov/data/2018/abscb?get=GEO\_ID,NAME,NAICS2017,NAICS2017\_LABEL,SEX,SEX\_LABEL,ETH\_GROUP,ETH\_GROUP\_LABEL,RACE\_GROUP,RACE\_GROUP\_LABEL,VET\_GROUP,VET\_GROUP\_LABEL,QDESC,QDESC\_LABEL,BUSCHAR,BUSCHAR\_LABEL,YEAR,FIRMPDEMP,FIRMPDEMP\_F,FIRMPDEMP\_PCT,FIRMPDEMP\_PCT\_F,RCPPDEMP,RCPPDEMP\_F,RCPPDEMP\_PCT,RCPPDEMP\_PCT\_F,EMP,EMP\_F,EMP\_PCT,EMP\_PCT\_F,PAYANN,PAYANN\_F,PAYANN\_PCT,PAYANN\_PCT\_F,FIRMPDEMP\_S,FIRMPDEMP\_S\_F,FIRMPDEMP\_PCT\_S,FIRMPDEMP\_PCT\_S\_F,RCPPDEMP\_S,RCPPDEMP\_S\_F,RCPPDEMP\_PCT\_S,RCPPDEMP\_PCT\_S\_F,EMP\_S,EMP\_S\_F,EMP\_PCT\_S,EMP\_PCT\_S\_F,PAYANN\_S,PAYANN\_S\_F,PAYANN\_PCT\_S,PAYANN\_PCT\_S\_F&for=us:\*&QDESC\_LABEL=SPOUSES

US Department of Commerce. (n.d.). 2018 - census.gov. Annual Business Survey (ABS) APIs. Retrieved January 12, 2022, from https://api.census.gov/data/2018/abscbo?get=GEO\_ID,NAME,NAICS2017,NAICS2017\_LABEL,OWNER\_SEX,OWNER\_SEX\_LABEL,OWNER\_ETH,OWNER\_ETH\_LABEL,OWNER\_RACE,OWNER\_RACE\_LABEL,OWNER\_VET,OWNER\_VET\_LABEL,QDESC,QDESC\_LABEL,OWNCHAR,OWNCHAR\_LABEL,YEAR,OWNPDEMP,OWNPDEMP\_F,OWNPDEMP\_PCT,OWNPDEMP\_PCT\_F,OWNPDEMP\_S,OWNPDEMP\_S\_F,OWNPDEMP\_PCT\_S,OWNPDEMP\_PCT\_S\_F&for=us:\*&QDESC\_LABEL=YRACQBUS

US Department of Commerce. (n.d.). 2018 - census.gov. Annual Business Survey (ABS) APIs. Retrieved January 12, 2022, from https://api.census.gov/data/2018/abstcb?get=GEO\_ID,NAME,NAICS2017,NAICS2017\_LABEL,SEX,SEX\_LABEL,ETH\_GROUP,ETH\_GROUP\_LABEL,RACE\_GROUP,RACE\_GROUP\_LABEL,VET\_GROUP,VET\_GROUP\_LABEL,NSFSZFI,NSFSZFI\_LABEL,FACTORS\_P,FACTORS\_P\_LABEL,YEAR,FIRMPDEMP,FIRMPDEMP\_F,FIRMPDEMP\_PCT,FIRMPDEMP\_PCT\_F,RCPPDEMP,RCPPDEMP\_F,RCPPDEMP\_PCT,RCPPDEMP\_PCT\_F,EMP,EMP\_F,EMP\_PCT,EMP\_PCT\_F,PAYANN,PAYANN\_F,PAYANN\_PCT,PAYANN\_PCT\_F,FIRMPDEMP\_S,FIRMPDEMP\_S\_F,FIRMPDEMP\_PCT\_S,FIRMPDEMP\_PCT\_S\_F,RCPPDEMP\_S,RCPPDEMP\_S\_F,RCPPDEMP\_PCT\_S,RCPPDEMP\_PCT\_S\_F,EMP\_S,EMP\_S\_F,EMP\_PCT\_S,EMP\_PCT\_S\_F,PAYANN\_S,PAYANN\_S\_F,PAYANN\_PCT\_S,PAYANN\_PCT\_S\_F&for=us:\*

**Extraction**

This data came from the United States census. This information is gathered every year to keep track of a wide array of metrics. This data specifically was gathered using an API and reading the data into a JSON file. This JSON file was the foundational data that we would summarize and craft into the various visualizations that we used to answer the questions we posed. This was done four times, once for each of the required data packages that would answer our questions. The Company Summary data set, once cleaned, is used in the third data set, the Characteristics of Business Owners, so make sure to do this first.

**Transformation**

In the first data set, the Company Summary, there were two transformations to get the data into a useable state. First, we dropped all the columns that did not hold relevant information. We dropped the following columns: 'GEO\_ID','NAME','NAICS2017','NAICS2017\_LABEL','SEX','ETH\_GROUP','RACE\_GROUP','VET\_GROUP', 'EMPSZFI','EMPSZFI\_LABEL','YEAR','FIRMPDEMP','FIRMPDEMP\_F','RCPPDEMP','RCPPDEMP\_F','PAYANN','PAYANN\_F','FIRMPDEMP\_S','FIRMPDEMP\_S\_F','RCPPDEMP\_S','RCPPDEMP\_S\_F','EMP\_S\_F','PAYANN\_S','PAYANN\_S\_F','NAICS2017','us'.

Then, we changed the ‘EMP’ column to a numeric data type, so that it could be aggregated. To specifically obtain the data shown in the visualizations, we had three further instances of the table. First, we grouped by ‘SEX\_LABEL’, summing the ‘EMP’ column, and removing all ‘SEX\_LABELS’ that were named ‘Classifiable’, ‘Total’, Unclassifiable’, or ‘Equally male/female’, as these were not demographics we cared to explore, or caused a modicum of overlap. In the second instance, we grouped on ‘RACE\_GROUP\_LABEL’, summing ‘EMP’. We eliminated the rows containing ‘Classifiable’, ‘Total’, Unclassifiable’, ‘Nonminority’, ‘Equally minority/nonminority’, or ‘Minority’. In the third instance, we grouped by ‘VET\_GROUP\_LABEL’, summing ‘EMP’. We eliminated the rows named ‘Classifiable’, ‘Total’, Unclassifiable’, or ‘Equally veteran/nonveteran’.

In the second table, the Characteristics of Businesses, we only kept the following columns: 'NAICS2017\_LABEL','RACE\_GROUP\_LABEL','SEX\_LABEL','PAYANN','EMP’.

We also eliminated any rows that did not include "Nonminority", "Minority", "Equally minority/nonminority" from the ‘RACE\_GROUP\_LABEL’ column or "Classifiable", "Total", "Unclassifiable" from the ‘SEX\_LABEL’ column.

In the third table, the Characteristics of Business Owners, we removed all the unnecessary columns, which are as follows:

'GEO\_ID','NAME','NAICS2017','OWNER\_SEX','OWNER\_ETH','OWNER\_RACE','OWNER\_VET','QDESC','QDESC\_LABEL','OWNCHAR','YEAR','OWNPDEMP','OWNPDEMP\_F','OWNPDEMP\_PCT’, 'OWNPDEMP\_PCT\_F','OWNPDEMP\_S','OWNPDEMP\_S\_F','OWNPDEMP\_PCT\_S','OWNPDEMP\_PCT\_S\_F','us'.

We removed all the rows that did not have non-specific race values, which are 'All owners of respondent firms', 'Nonminority’, and 'Minority'. We also removed all non-specific Sex and Veteran Status Labels, which are ‘All owners of respondent firms’ for both ‘OWNER\_SEX\_LABEL’ and ‘OWNER\_VET\_LABEL’ columns. Finally, we inner merge on the Company Summary data set on ‘RACE\_GROUP\_LABEL’.

For the fourth table, the Technology Characteristics of Businesses, we use only the following columns: 'RACE\_GROUP’, ’RACE\_GROUP\_LABEL’, ‘IMPACTWF\_P’, ‘TECHUSE’, ‘TECHUSE\_LABEL’, and ‘RCPPDEMP'.

Then, we make sure to cast ‘RCPPDEMP’ as a float type so we can aggregate it. Finally, we renamed all the values in the ‘TECHUSE\_LABEL’, so that the values no longer included the text segment ‘: Did not use’.

**Load**

If we were to load the data into SQL, we would need to create a new file to be used after all of the transformation steps. In each case, we would save the cleaned data to a JSON file or a CSV, whichever would work, and then import this data to an SQL server. Each table stands apart, and while share some categories, they are actually different values collected from different sources. As they do not share any keys, there is no relationship between the tables, and the files can be uploaded separately in any order desired. If the user wished to replicate the visualizations we created or aggregate as we did, however, they need to continue to cast the columns that we specified in the transformation process as numeric.

**Conclusion**

With the data that is collected and cleaned in the way that we specify, the user should be able to obtain insight into the disparities between demographics, pay, and technology in the United States today. This data is also easily transformed and usable in SQL or excel, and aggregation is quite easy so long as steps are properly followed.