ETL Report Guide

Salma Tahlil
ETL Process Date

Introduction

The problem we are trying to solve here is what factors impact quality of life and can that quality of life vary from area to area. Some of the data that is being used to solve this problem is the CDC data behavioral risk factor. This data set gives us questions from a survey as well as the answers to those questions. Some of those questions include "Mean days of activity limitation" or "Mean mentally unhealthy days". All of the questions are within the topic of activity limitation, mental health, general health, and physical health which are all aspects of quality of life. This dataset also provides age ranges, race, and gender to see the differences in those fields. It also provides other factors such as state and year. Another dataset we used was the census data set. This provided all the different zip codes within the states, gender, population, min and max ages, city, as well as longitude and latitude. The first dataset provided only had zip codes and very few information that I could tie into my main census table which is why I used the other zip codes dataset from Kaggle with the states included. Lastly, after doing some research I found my last dataset which is state annual personal income from a government site. The dataset contained personal income per capita, states, population, and years dating back all the way to the early 1900s. This dataset was useful to see the differences in income especially in the early 2000s to see what other outside factors contribute to quality of life.

After analyzing my four datasets, CDC and my 2000 and 2010 census, and my personal income dataset I noticed a few things that needed to be done. The datasets needed to be cleaned and transformed to be able to make my connections and work with my data. Things such as mean and percentage being in the same column, some columns not having any or very few information, and other things like renaming, grouping, and filtering. My personal income dataset needed to have fewer rows and columns to make connections to my other tables.

Data Sources

As mentioned previously, my datasets were taken from Kaggle which is my Census data and CDC for my behavioral risk factors. My outside dataset was found off of a government website and was accessed on Friday December 2nd, 2022. I had access to my other datasets from the very beginning of my mastery project Monday November 28th, 2022.

Behavioral risk factor HRQOL - dataset by CDC. data.world. (2017, February 2). Retrieved November 28th, 2022, from https://data.world/cdc/behavioral-risk-factor-hrqol

State personal income: Revised estimates for 2010. BEA. (n.d.). Retrieved December 2nd, 2022, from https://apps.bea.gov/regional/histdata/releases/0911spi/index.cfm

US Census Bureau . (n.d). US Population by Zipcode, Version 2. Retrieved November 28th, 2022 from https://www.kaggle.com/datasets/census/us-population-by-zip-code

Extraction

The first dataset I retrieved was my CDC dataset from data.world. The dataset was given through the mastery project requirements, and it was in csv format when I received it.

1. The first step was to download my csv from the website and uploading it onto excel.
2. The second step I took was going to my power query editor and going to my sources to directly import my csv in my power query.
3. The next step I took was analyzing my data to see what columns/rows I wanted to keep. What made sense to put together or extract.

The second and third datasets were the census data from Kaggle. These datasets were also provided by the mastery project requirements which is how I got the data.

1. I downloaded both the 2000s and 2010 zip code datasets which were both in csv format.
2. The first blocker I encountered was how the zip code dataset was far too large for excel and it could only be imported through power query launcher and not directly on the workbook.
3. After analyzing the data, I realized I did not have enough information from my zip code data and went back to Kaggle and noticed the same data set but with more useful information like state and city and downloaded that csv from Kaggle and uploaded it to power query.

The last dataset I acquired was my government-based dataset of state income per capita. I found this dataset by googling income datasets and doing some research to find a more normalized dataset. This was one of the more normalized datasets I found and was super easy to analyze, clean, and transform.

1. I directly downloading the csv file from the website and loaded it onto my excel. It was not a huge file, so it was on my workbook.
2. To start my transforming process, I uploaded it on my power query by going through the same process as my previous datasets by using sources and uploading csv in power query.

Transformation

After analyzing my behavioral risk table, I noticed that I needed to do a lot of transforming and cleaning. I did not end up using all of the data that was provided with the original table.

1. My first step was to make a reference table so that if I encountered errors or messed up I could go back to my original table and start again.
2. The next thing I did was filtered my year column so I would not get any nulls in my years or blanks. I also filtered my locationabbr column so I would not have the US so it would not mess up my data.

3. Afterwards I added a new column and extracted my geolocation column so it would separate my values by the delimiter. I did this so I could have longitude and latitude in separate columns.
4. Then I removed columns like the data value unit, data sources, and other columns that didn't have necessary information or a lot of blanks or nulls.
5. I then separated my Data value type so that I could have my average number of days in one column and percentage in another.
6. I then deleted those columns that were given previously.
7. Next, I deleted more columns like high and low confidence and the ids only because I made my own ids in SQL. I did not think the low and high confidence made a difference in my analysis which is why it was removed.
8. I also changed column names to match my entity relationship diagram.

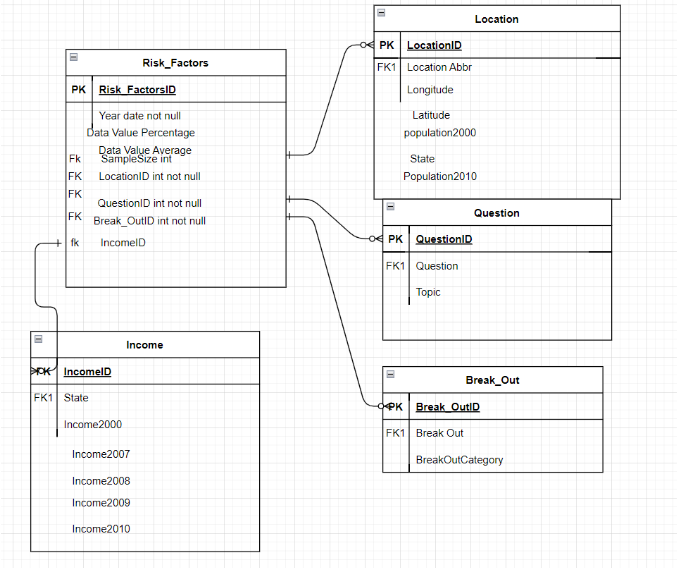
The next table I worked on was my zip codes. I looked at the information that would be helpful in my analysis which I originally thought to be my zip codes and in the end they were not. It took me awhile to make the conclusion that zip codes were too much and when I finished cleaning it would be hard to transfer into SQL.

1. Which is why I deleted my zip code column and decided to look at total state population. I made sure to make gender null to get total zip code population then grouped by to get total state population. I was then only left with state and population 2000 and 2010. I then merged my 2000 and 2010 in one table by doing a inner join.
2. I also changed my column names to match my entity relationship diagram.

My last dataset was mostly normalized which made it easier to work with.

1. I deleted all the columns besides 2000, 2007, 2008, 2009, and 2010 and the state.
2. The only cleaning I needed to do was delete things such as westside or Midwest in my states column to only have my states.
3. I also needed to replace values for Hawaii because it was written out Hawaii 3//.
4. I also needed to filter my per capita column to only get the income per capita for the states and get rid of the things such as the population.
5. I renamed my columns to be like my zip code table.

Since your end goal will be to load your data in SQL Server, include table mappings that identify the source data and its destination.



Load

After finalizing my datasets and all the steps I took to make them more normalized but not too much I took my final steps to load.

1. I closed and loaded my tables into my worksheet. I could do this with my zip code too since I reduced it to only 51 rows.
2. After that I saved my tables as csv's and went into SQL to import them.
3. I made a new database in SQL and then I right clicked it to use import wizard.
4. This would then lead me into finding where my csvs were and changing all of my ints to ints and not what they were previously. I also made it allow nulls for everything.
5. After this step it would import into SQL.

Conclusion

After gathering my datasets and going through the steps to extract, transform, and load and importing them into SQL I got more and more comfortable with my data. At first it was daunting to make those connections to answer the made questions of this project. I also ran into a lot of blockers and errors which was very stressful, but I took the steps to ask the questions and using resources to get through them.