Extraction, Transformation, and Load Technical Report

NY State of Mind

Members: Kim Campbell, Lyle Sweet, Nancy Culley

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Summary

Our client wants to open a new radio station with an experimental genre of music in NY state and wants to see which locations might be more receptive to this new genre. Our hypothesis is that cities in NY state with higher populations have a more diverse array of music genres played on their radio stations.

1.2 Scope

Data

In order to drill down to potential markets for a new radio station, we pulled current NY radio stations, with city and genre (format) by scraping the following website:

https://radio-locator.com/cgi-bin/finder?sr=Y&s=T&state=NY

We also wanted the population of the surrounding area to the radio stations for potential listeners. We pulled population by county, assuming that a radio station could reach the whole county, from the Census Bureau API:

https://api.census.gov/data/2010/dec/sf1?get=H001001,NAME,H003001&for=county:&in=state:01&key=59a41ecf0dff091bd991b35c87176ffd56b8107d*

In order to join the two datasets, we needed a city to county lookup, which we scraped from two wiki pages for town to county and city to county:

https://en.wikipedia.org/wiki/List_of_cities_in_New_York_(state)

https://en.wikipedia.org/wiki/List_of_towns_in_New_York_(state)

Created a CSV to group formats from 101 in the radio-locator datasource to 17 more generic formats: formats.csv

Additional data that may be helpful, but was **out of scope** for this project could include radio stations statistics, such as number of average listeners, station revenue/profit, most popular music by format...

1.3 Technologies and resource contributions

This project used two different data-extraction techniques as well as a SQLlite integration from Python.

Nancy Culley: Used her API expertise to load New York population data by county from the US Census Bureau API

Kim Campbell: owned the object relational management section. She wrote the code to translate our Pandas dataframes into SQL tables.

Lyle Sweet: Extracted the radio station data from tables on the web using pandas.

All three team members took part in data-munging - cleaning merging and formatting the data sources into easily readable and actionable data.

1.4 Definitions, Acronyms and Abbreviations

ETL: Extract, Transform and Load
ORM: Object Relational Management
API: Application Programming Interface
SQL: Structured Query Language

2. ETL DETAILS

2.1 Data Import/Extract Sources and Method

We extracted data using a combination of Web Scraping and API. Our Web scraping sources were the following:

- **Radio Locator Page:** https://radio-locator.com/cgi-bin/finder?sr=Y&s=T&state=NY: This tells us which stations and music formats exist in New York State, as as the "city of license" for the radio stations that were represented
- **Wikipedia Pages of cities and towns in New York:** This told us which cities and towns our city of licenses corresponded to. This was critical in helping us see the county each location belonged to.
  - https://en.wikipedia.org/wiki/List_of_cities_in_New_York_(state)
  - https://en.wikipedia.org/wiki/List_of_towns_in_New_York_(state)
- **CensusAPI:** *https://api.census.gov/data/2010/dec/sf1?get=H001001,NAME,H003001&for=county:*&in=state:01&key=59a41ecf0dff091bd991b35c87176ffd56b8107d*
  - This told us the population for each of the counties in our data set.

All of these data sets were open and required no other permissions

2.2 Data Acquisition

The census bureau only performs collection every 10 years. Currently we're using 2010 information which is slightly outdated. Our team recommends a refresh of census data once new information is available in 2020.

The list of radio stations on Wikipedia is updated from http://worldradiomap.com/us-ny/. We are scraping the Wikipedia pages which should include fresh data with every scrape.

The only other data pulled into our project comes from manually creating a short-list of radio-formats. There were +100 unique formats in the original list of 600+ radio station with lots of slight spelling differences. More information can be found below in the "Data Transform" section.

2.3 Data Transform

There were several data transformations that helped us get to our final product.

1. **Recategorizing the radio formats**: We realized that many of formats were part of the same category but were just slightly different from one another. For instance Catholic, Christian, and Religious were all recategorized as Religious. We created a csv with 2 columns , one with the existing categories, and the other with the modified categories. We condensed the total number of formats from 101 to 17.
2. **Strip out unnecessary spaces and characters**: Simplifying and reformatting header values in our data frames were also helpful. They enabled us to successfully push the final dataframe into an sqlite database without issue. The characters created challenges with the initial database push before cleaning
3. **Aligning formats for the "county" column in several tables for a merge:** The counties from the census API were represented as the county name with a comma and the state New York(ie: Genesee County, New York). However the county value from the radio data frame only represented the county name (ie Genesee). We stripped out everything after the county's name in the API values so that they would match with the radio dataframes. This facilitated a successful join between the two data frames.
4. **Multi-level index for nested groupbys** : The goal was to produce a count of each format for each county. In order to accomplish this, we had to produce a multi-level index that first grouped the data by county, then by radio format. We then used .unstack to turn the series into a dataframe that could be manipulated for further analysis.

2.4 Data Integrity

We feel very confident in the US Census Bureau data taken via API. There are many datasets with geographic and demographic information. The only limiting factor here is how often the census is conducted - every 10 years.

The data scraped from Wikipedia we are also confident in. Data seems to be provided from http://worldradiomap.com/us-ny/ which is a personal project by radio enthusiast, Mikhail Shcherbak (a.k.a. Predavatel). There are paid sources for this data as well, however, it appears to exactly resemble the tables provided by Wikipedia. We know this from a free test trial.

2.5 Data Refresh Frequency

Everytime this python code is run fresh values will be pulled from Wikipedia and the US Census Bureau. Updates will rely solely on those entities. Details are provided in the above section “Data Integrity”

2.6 Data Security

All data in this project is pulled from free, publicly available datasets. There is zero personally identifiable information included. There are no disclosures required to view or use this information.

2.7 Data Loading and Availability

All information gathered is available via standard SQLite queries. We have chosen SQLite as our preferred database due to its simplicity. Standard SQLite queries samples can be found here => https://www.sqlite.org/lang.html.

3. DATA QUALITY

End result - Successfully provided a usable database for our client to cross reference county population and radio station formats in each of those counties.

Testing quality - To test our sql output, we reproduced the table in Excel and were able to replicate the same results.

If the client wants to test our output, would suggest they use the radio-locator website to validate a sample of the data.

Future scope - would offer client our services to build interactive app or website where they could choose a format and a NY heatmap would visualize how many locations around the state have those formats, with a secondary measure for population.