Extraction, Transformation, and Load Technical Report

Group 7 - Velkis Moreno, Aaron Zeanah, Jeremy Hartnett

**TABLE OF CONTENTS**

1. Introduction 3
   1. Summary 3
   2. Scope 3
   3. Technologies and resource contributions 3
   4. Definitions, Acronyms and Abbreviations 3
2. ETL Details 4
   1. Data Import/Extract Sources and Method 4
   2. Data Acquisition 4
   3. Data Transform 4
   4. Data Integrity 4
   5. Data Refresh Frequency 4
   6. Data Security 4
   7. Data Loading and Availability 5
3. Data Quality 6

|  |  |
| --- | --- |
| **1.** | **INTRODUCTION** |

*The purpose of the Extraction, Transformation, and Load (ETL) Technical Report is to capture details that pertain specifically to ETL portion of the data pipeline that is to be used in a data science project. This however does keep in mind the final target objective while performing the ETL.*

# 1.1 Summary

XYZ Real Estate Investment Corp (the client) is interested in making future real estate investments throughout the United States. As part of the things to consider as reasons to invest in certain cities, XYZ feels that air quality is one of the most important to consider and speculate this contributes to market “hotness”. The client has tirelessly searched for readily-available public data which shows the impact of real estate hotness market and air quality with no avail. Fortunately, they have asked VAJ Data Science, Inc. a leader in data science and analytics to deliver a data set that will allow them to understand the relationship between Real Estate Market "Hotness" and observed Air Quality nationwide. The client hopes that with the dataset provided by VAJ DATA Science, the analysis will show a positive correlation between air quality and housing market hotness that can be used in the future selection of real estate investments they have in mind.

# 1.2 Scope

The scope of this engagement is to take two datasets found on Realtor.com and the Environmental Protection Agency (EPA) website and join both datasets with the goal of creating a new master dataset for the client’s use. This data set must have the ability to be queried and contain all relevant data that will help determine two main things; correlation of air quality and market hotness, and their respective locations.

**In Scope:**

* Extraction of data sets from sources
* Transformation of housing data set (monthly data into annual data)
* Design of database to hold data
* Load of data into the database
* Design of basic queries for client

**Out of Scope:**

* Analysis of the data set
* Automated data refreshes

# 1.3 Technologies and resource contributions

There were three Data Scientists who contributed to this research: Velkis Moreno (Database Architect), Aaron Zeanah (Programmer), Jeremy Hartnett (Database Architect/Technical Writer).

The technologies used for this research were the following:

* **PgAdmin** - used to create the database and combine data sets sourced from the Environmental Protection Agency and Realtor.com websites.
* **Pandas** - used to aggregate monthly Realtor.com metrics into annual format, and to produce annual data sets for loading into database tables.
* **Word** - used to create the technical report

# 1.4 Definitions, Acronyms and Abbreviations

* **ETL** - Extract, Transform, Load
* **CBSA**- Core-based Statistical Area
* **Normalize** - Restructuring a relational database in accordance with a series of “normal” forms to improve data integrity
* **SQL** - Structured Query Language
* **RAW Data** - Unaltered data

|  |  |
| --- | --- |
| **2.** | **ETL DETAILS** |

# 2.1 Data Import/Extract Sources and Method

**Data Import:**

The data used for this project came directly from realtor.com and ags.epa.gov and were received as .csv files. The data was first imported into PANDAS so that we can summarize data by year as it was provided as monthly. It was then analyzed to see if we could identify any gaps in the data, inconsistencies and if the data had common factors that would allow for joins. After an initial run through PANDAS it was converted into new .csv files to be imported into PgAdmin (SQL) to then create database tables.

**Extract Sources:**

**Realtor.com - Market Hotness Index: July 2016 through April 2020**

<https://www.realtor.com/research/data/>

**EPA Daily AQI by CBSA - Daily data, July 2016 through April 2020 (converted to monthly data)**

<https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual>

**Method:**

Both datasets were then normalized from its raw data state by removing blanks and creating consistency for each column. Some of the issues with the raw data is that 2020 did not have a complete year worth of information so not all fields were populated. We also normalized the data to allow for joining of similar time periods and locations. Our goal was to prepare the data in a clean way that it could be delivered to the client in a format which would allow the client to then manipulate the data as they saw fit.

# 2.2 Data Acquisition

The group evaluated both data sets and determined that multi-year historical data; 2016 to present, as the oldest information from Realtor.com. As for EPA, their data went as far back as 1980. VAJ Data Science would like the client be mindful as they review the joined data, that real estate is a lot more volatile in shorter time period compared to air quality, and that any analysis should be made based on this understanding We also recommend that this same dataset be updated in another 5-10 years to get a better perspective or understanding of the relationship between market “hotness” and air quality as air-quality data does not change as quickly.

The data pulled was static/fixed data. There were no data changes or updates required for this data set as the information was an accumulation of past records from 2016 to present. For the Air Quality data, formatting was not required as the data was based on yearly information for each CBSA.

As for joining the dataset, VAJ Data Science felt that the client would benefit from the creation of a primary key column. In order to create the key column, the team did a concatenate of year and unique code. This seemed to be an extra step needed to create unique values between both datasets. In addition, the realtor.com dataset did need some additional formatting because the data provided was based on year and month. The team collectively felt that it was more data than what was actually needed, therefore through the use of PANDAS, the Developer consolidated the data by creating a programming script, which resulted in an annual data files from a master file. The result of this new data allowed the client to be able to now see data on either a monthly or annual basis. This data process can be repeated which allows us to examine the data structure and address any changes before combining with the master data set. To reduce complexity in future updates, all data from the files downloaded from websites was kept (no data was deleted). Data growth rate is expected to be relatively slow, given data can be updated only on a monthly or annual basis, there is a limit to the size of each update (fixed number of CBSAs), so we do not anticipate storage or transmission of files to be a problem in future updates. There were no dependencies in order to obtain both datasets as they were free downloads available to the general public, and given the history and size of the data providers, we expect data to be made available in similar fashion in the foreseeable future.

# 2.3 Data Transform

The data sets have a common field (CBSA Code) which allows for joining. However, the time periods in the downloaded data sets are different; the Realtor.com Market Hotness file provides data in monthly format for the period of July 2016 through April 2020, whereas the EPA air quality data is in daily format for the years . The EPA data was aggregated to produce a similar monthly air quality metric to allow for “apples to apples” comparison. Changes to the raw data download consisted of formatting (splitting date fields into year, month columns) only; no data was dropped in the course of this project. Derived fields were not used, so resulting data is as close to the source data as possible.

# 2.4 Data Integrity

EPA Air Quality data is directly sourced from the EPA website, which “provides access to outdoor air quality data collected from state, local and tribal monitoring agencies across the United States.” Data integrity appears to be high; data is sourced from the government agency that owns the process for producing the data (field monitoring stations), data fields are complete, etc. Data is updated on a daily basis, and can then be aggregated as needed to match corresponding data availability from other sources; an annual air quality metrics report is also available on this site. . As the real estate data is compiled on a monthly basis, the EPA data can also be updated on a monthly basis vs. daily. New data is posted to the website as it is made available to the public; there are no notifications provided.

Realtor.com data is updated on a monthly basis, ~4-5 days following month end; the website lists the next scheduled update date. New data can be appended to the data set as it is made available, but must be done manually vs. via an automated routine. In terms of data quality, the website states that “Data in this realtor.com library is based on the most comprehensive and accurate database of MLS-listed for-sale homes in the industry. We aggregate and analyze data from hundreds of sources and produce hundreds of metrics for multiple markets, and curate figures and trends where possible for reliability and comparability. However, as we continue to evolve our coverage and fine-tune our definitions, some data points may be too volatile or incomparable over time or across markets. This is particularly true for data in a) smaller geographies; b) markets with special or changing definitions of active inventory; c) markets with limited or partial listing; and d) coverage markets with limited or partial sales coverage. Where possible, these cases are annotated for individual metrics (see data dictionaries for more info). Also, every month, we reissue the full historical series, and past data points may change as we improve data breadth and accuracy, and/or re-state the data altogether.” New data is posted to the website as it is made available to the public; there are no notifications provided.

# 2.5 Data Refresh Frequency

For the purpose of this engagement, data refresh is not needed as this report itself along with the normalized and raw data is part of the deliverable package and based on the statement of work.

VAJ Data Science is however offering extended maintenance for an additional 5 years at the client’s option which will allow new data to be generated on either a daily, monthly, or yearly basis and based on the client’s specific needs.

# 2.6 Data Security

There are no data security requirements as all data is sourced from public websites that have no access controls, data is available to the public directly via the websites, data is not sensitive, etc. However, the client is responsible for portraying the data and any analysis in an ethical manner which complies with the use as stated in both realtor.com and EPA’s website.

For further details, please refer to the links provided below for data security:

<https://www.nar.realtor/data-privacy-security>

<https://www.epa.gov/home/epa-disclaimers>

# 2.7 Data Loading and Availability

Table layout consists of a master EPA data file, and annual tables for Real Estate hotness. Primary key is CBSA code. The client will access the data and perform queries via SQL queries.

|  |  |
| --- | --- |
| **3.** | **DATA QUALITY** |

**Data Validation** - The data was validated by running scripts, loading the data and ensuring the data provided expected results. VAJ Data Science tested the data a few times to mitigate any bugs, duplications, and errors resulting from using raw data.

**Data Quality** - checks can be performed by measuring CBSA totals vs. master CBSA lists that can be found online via multiple sources. Client can verify same, and can also check to ensure there is both an air quality metric and a real estate hotness score for each CBSA.

**Acceptance Testing** - VAJ Data Science created a test environment which provided 100% success results. They then recreated the data after creating the tables, loading the data for each table and confirming the data was loaded successfully. It was then that the team felt the new data could be uploaded on Github for client review and acceptance. Once acceptance was received, the data files were consolidated and the client was provided with one deliverable.

**Data Results** - The client should see approximately 2k +/- unique identifiers.