Extraction, Transformation, and

Load - Technical Report

LexCorp Data Analytics

Compiled by:

Andrew Boring

Vimal Mathew

**TABLE OF CONTENTS**

1. [Introduction 3](#_bookmark0)
   1. [Summary 3](#_bookmark1)
   2. [Scope 3](#_bookmark2)
   3. [Technologies and resource contributions 3](#_bookmark3)
   4. [Definitions, Acronyms and Abbreviations 3](#_bookmark4)
2. [ETL Details 4](#_bookmark5)
   1. [Data Import/Extract Sources and Method 4](#_bookmark6)
   2. [Data Acquisition 4](#_bookmark7)
   3. [Data Transform 4](#_bookmark8)
   4. [Data Integrity 4](#_bookmark9)
   5. [Data Refresh Frequency 4](#_bookmark10)
   6. [Data Security 4](#_bookmark11)
   7. [Data Loading and Availability 5](#_bookmark12)
3. [Data Quality 6](#_bookmark13)

**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.*

# Summary

This ETL report summarizes data accumulation and cleanup for initial estimates for feasibility study into Lex Luthor’s Plan.

The Plan outline is summarized here: http://github.com/andrewboring/lexcorp-feasibility-study

# Scope

The scope of this report is limited to the extraction and transformation of data obtained for initial estimated value of real estate property along the San Andreas Fault. This report does not include a summary of data for detailed estimations and ranked-order recommendations outlined as deliverables for the Plan itself.

# Technologies and resource contributions

Resources:

|  |  |  |
| --- | --- | --- |
| Source | Item(s) | Contributor |
| Realtor.com | Residential Database Historical Listings | Andrew Boring |
| Zillow Research | Zillow Home Value Index | Andrew Boring |
| Census.gov | Census data by zip code | Vimal Mathew |

Technologies uses in this ETL project include:

* Python 3, with Jupyter notebook and various Python data handling modules.

# Definitions, Acronyms and Abbreviations

List acronyms and terms that need to be defined in this section, such as ETL: Extract, Transform and Load

ZHVI

**2. ETL DETAILS**

*This section outlines a more detailed description of the processes utilized/proposed to achieve the objectives of this initiative.*

# Data Import/Extract Sources and Method

The following section outlines the data sources used and the mechanisms for obtaining them, along with any restrictions imposed by the suppliers.

**Realtor.com Residential Database**– This data was obtained from the Realtor.com Real Estate Data Library, located at <https://www.realtor.com/research/data/>. Files obtained here are free to use, but require attribution and a link to the source page. The historical data file is a direct link to the current dataset, and will update on a monthly basis. The file itself exceed software repository file limits, so the data\_assemble tool located in the repository contains a mechanism to download the current data file if a local copy is unavailable. A later iteration of this utility will include the ability to compare an md5 checksum of any local copies and compare them to the HTTP Etag header to determine if download is required.

**Zillow Home Value Index** - This data set was obtained from Zillow Research’s Data page, located at <https://www.zillow.com/research/data/>. The Zillow Home Value Index Current Month was the source material, which is provided through a download form.

Zillow does provide an API for direct data access, however it is somewhat undocumented and there does not appear to be any internal resources maintaining this service. Its use is not currently recommended.

**Census data** – This data set was obtained from factfinder.census.gov, a browser-based query tool for obtaining selected data from the US Census. The use of this tool is described in a separate document, *Obtaining Census Data Set,* located in the /docs folder in the software repository.

# Data Acquisition

Census data obtained was from Year 2000 census and based on a superset of the zip codes requested by zip code prefix. The data included Housing information, such as number of housing units per zip code, housing density (units per sq mile), and area of zip code in sq miles. This source was generated using an online query tool described in section 2.1, and documented separately with screenshots for validation. A URL query string was displayed for repeatable requests; however, this did fail in some instances and may require additional session information. An alternate method for programmatically obtaining this dataset might use a web scraping function with browser automation.

The Zillow Reserarch was obtained through a selectable download form, which in turn provided a download of the CSV file. The selection fields create a repeatable URL (http://files.zillowstatic.com/research/public/Zip/Zip\_Zhvi\_Summary\_AllHomes.csv) that can be queried periodically (monthly), or queried for Etag and compared to a local md5 checksum.

The downloaded data file includes current month ZHVI and zip code, along with additional information not used (eg, month-over-month change).

The realtor.com data file was 250MBs, and downloadable directly from their site through a simple, direct access link. The raw data set is not included as it exceed repository limits, so a download function is provided to obtain the data set. Like the Zillow Research data, this file can also be accessed periodically or through an Etag comparison. This file includes historical information by zip code, including median and average listing prices, along with additional data such as listing counts and m/m changes.

The pandas library will load large files in smaller increments, which can make datatype guessing inaccurate. As such, three columns required explicit type casting upon import.

# Data Transform

The Realtor.com data is downloaded, and imported into a Pandas Dataframe. The month column is renamed to Date, and then filtered to just 2018 data. The dataframe is then merged with a list of zip codes using an inner join to select only the rows with the desired zip codes.

The Realtor file does not contain information on all zip codes: 45 zip codes are not available, so the filtered data set only contains info on 39 of the 84 zip codes of interest.

The ZHVI file was imported and necessary columns renamed to match the Realtor data and unneeded columns dropped. This file was then merged with the Realtor file using two outer joins (an outer left and outer right on Date and Zip Code columns) so that any unrepresented zip codes in Realtor file would be present with an alternative ZHVI.

The Census data included a double-column header with the database column names, so some initial cleanup to provide readable column headers was the first task. Several columns were renamed to match date and zipcode column names in previous files, and rows with a non-five-digit zip code were dropped (these zip code fields included zip prefix aggregate data and other uses). This dataframe was then merged with the previous merges to provide our final data set.

# Data Integrity

This dataset does not include information on all zip codes, and indeed is not granular enough to provide parcel-level information. It suffices as an exercise to understand the financial scope of this project, but is only a first step in a more detailed process to obtain granular information.

Additionally, the combination of disparate data sets with different zip code inclusions result in many incomplete fields. Any aggregations should account for this or drop rows with empty fields.

Finally, the date fields have differing month boundaries between data sets (eg, February might be represented as Feb 01, 2019 or Feb 28, 2019); as a workaround, this may a transform to merge based on month/year prior to calculations/aggregations.

# Data Refresh Frequency

The census data is updated decennially, so would not require routine extraction until after the next census.

The real estate source data changes monthly, and includes all the previous data already obtained. While obtaining only the changes isn’t supported by the sources, the files can be overwritten on each extraction for minimal data growth of a few text rows per month.

# Data Security

The Census data is public or free to use, with the attribution restrictions referenced in section 2.1

# Data Loading and Availability

The data will be loaded into a MySQL database for long-term retention.

Users can access the data\_assemble notebook in the software repository for regenerating the source data and output, while a web dashboard will be provided for visualization of the final dataset.

**3. DATA QUALITY**

As a phase 1, initial estimate with a time-limited delivery, this data is of sufficient quality to assess the financial scale of the larger project. This data has a number of gaps, partly because the initial sources do not cover the entire geographic region.

Where gaps exist, the option is available for the analyst exclude data that contains gaps, to “fill in” gaps based on some methodology, or work around the missing data based on the analysts goals.

The phase 2 data estimates will overcome this limitation.