**DSC 540 Data Preparation – Project Milestone#1**

**Ganesh Kale**

**September 25, 2021**

**Austin TX House Market**

**Project Overview:**

House Prices in USA are booming, and house prices will continue to race ahead, at nearly twice the pace predicted before this year. This is what we hear or read when talk about housing market in USA. Buying house is very critical job, one should be aware of lots of things before buying house and when buying house nobody sure about when is the right time to buy house and wants to have some tool that would consider all the factors determining house price and predict the house price. Predicting the house price is challenging but doable and with help of machine learning algorithms this can be achieved.

This is the topic chosen for the final project to make data ready by collecting from different sources and cleaning, transforming, and merging all these data to make it final dataset in the ready format for machine learning algorithms so predicting house price model can be developed by training and validating on housing market dataset.

To build predicting house price model, we should consider different factors such as house information, and facilities available in neighborhood such school, hospitals etc.

As part of this exercise, Austin, Texas housing market data will be collected from 3 difference sources such as csv file with house related information, json file for hospital in neighborhood and tabular data for school information from website.

**Data Information**: The data for this project will be collected from different sources and different format and all data will be cleaned, transformed, and merged into one final dataset. Below is the information about each data set and source where it will be collected from:

1. **CSV File** – The csv file is acquired from kaggele website, this data set have Austin TX area house information, such as address, year built, sale price, No. of bedroom, bathrooms, lot area etc. Please refer data dictionary for more information about data.

Link to Data Set - (<https://www.kaggle.com/ericpierce/austinhousingprices>)

1. **Json File** – The neighborhood public health locations information is pulled from data.gov using API. The data is in json format, and it will have public health locations information such as facility name, address, hours of operation, website etc. Data dictionary has all the detailed information about this dataset.

Link to data set – <https://catalog.data.gov/dataset/austin-public-health-locations>

1. **Tabular data** – The neighborhood school information will be pulled from below websites, since the first website has school information such as address, type of school, student per teacher ratio etc. but does not have information such ratings, score etc. for that we are going to use second source of tabular data, and both will be merged later.

Links to data –

1. <https://www.greatschools.org/texas/austin/schools/?gradeLevels%5B%5D=e&view=table>
2. <https://rptsvr1.tea.texas.gov/cgi/sas/broker?_service=marykay&_debug=0&_program=perfrept.perfmast.sas&ptype=P&level=district&search=distnum&prgopt=2019/acct/campus_list.sas&namenum=227901>

**Relationship** **Among Datasets** –

To make one final dataset which will have all the information related house, public health, and school, we are going to merge them, to merge all these different data sources we need to have some relationship among them.

The relationship among all these three data is zip code. We have housing data set with house address and zip code, public health location dataset with address and zip code and school information dataset with address and zip code, hence zip code will be used to merge all these datasets into one.

**What needs to be done to complete project** – Since goal of the project is clean and transform all the datasets to make it format that will be suitable for machine learning model to build house prediction model.

We are going to follow below steps:

1. Load all three datasets from different data sources into individual data frames using pandas package.
2. Check the data summary and structure of the data - to understand what kind of variables are in the data and its types.
3. Check for null, na or duplicated values and handle them appropriately.
4. Check descriptive statistics for each dataset.
5. Add or remove unwanted features from the datasets.
6. For school dataset split the address into street name, city, and zip codes.
7. Transform the data - change the categorical variable to numerical to fit to model, remove duplicates, perform aggregation to explore the data etc.
8. Explore and visualize the data - Plot different charts to see the data distribution, outliers etc.
9. Check the data Distribution and Correlation - checking data distribution and correcting it by removing outliers if any and checking the correlation among the features to decide which one to keep and which one to remove.
10. Finalize the features and make sure all are in correct format if not convert and standardize them so ML algorithm can fit on the data set.

**Data Dictionary** –

C**SV File – Houses Information:**

The Austin TX housing data was originally obtained from Zillow and used in Kaggle competition. The data consists of around 29 different features of houses and have around 10K houses information.

**Dataset feature details** -

1. zpid - Zillow Property Id

2. city - City name

3. streetAddress - Address of House/Property

4. zipcode - Zip code of property

5. description - Property Description

6. latitude - Coordinates - location

7. longitude - Coordinates - location

8. propertyTaxRate - Property Tax rate in that area

9. garageSpaces - How many cars parking spaces in garage

10. hasAssociation - Is HOA there or not

11. hasCooling - AC units are installed in house or not

12. hasGarage - Property has garage or not

13. hasHeating - Property has heating system or not

14. hasSpa - Property has spa or not

15. hasView - Property has view or not

16. homeType - Property is single family or apartment or townhouse etc

17. parkingSpaces - How many parking spaces

18. yearBuilt - What year property was built

19. latestPrice - WHat is latest house price

20. numPriceChanges - How many times property prices changes since listed 21. latest\_saledate - Date of last sold

22. latest\_salemonth - Month of last sold

23. latest\_saleyear - Year of the last sold

24. latestPriceSource - The party provided the price of the property

25. lotSizeSqFt - Lot size in sq ft

26. livingAreaSqFt - Living area in sq ft

27. numOfBathrooms - Number of bathrooms

28. numOfBedrooms - Number of bedrooms

29. numOfStories - number of stories

**Json File – Public Health Information:**

This data set is pulled using api from data.gov and it will have below information:

1. Facility Name – Name of health facility

2. Address – Street address of facility

3. Zip Code – Zip code of facility

4. Hours – Hours of operations of facility

5. Website – Website of the facility

6. Phone Number – Phone Number of the facility

7.Other Phone – Other phone of the facility

8. Building ID – Facility building identifier

9. Ownership Status – The facility is leased or rented etc.

10. Owner – Owner of the facility

11. Occupying division – Kind of facility such community service or disease prevention etc

12. Occupancy Type – Facility is clinic or hospital etc

13. Area - Area of facility

14. Year Built – What year facility was built.

**Tabular data from website – School Information:**

The school information is one of the key factors that drives house price, so we are collecting it from above website wherein table is provides with school information.

1.address - street address of school

2. zip code – Zip code of school

3. Name. – School Name

4. Type – Type of school such as public or charter

5. Grades – What grades are in that school are taught.

6. Total Student – Total number of students enrolled in that school

7. School Type – School is elementary or middle or high school

8. Rating – school rating such A, B or C

9. Score – school score out of 100

10. Student Per Teacher – ratio of students per teacher