**DSC 540 Data Preparation – Project Summary**

**Ganesh Kale**

**November 15, 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.

As part of this project, the data is collected from different sources and performed data wrangling steps such as – data cleaning, transforming, and merging all different sources 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 is 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 Kaggle 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. **Tabular data** – The neighborhood school information will be pulled from below websites, since the website has school information such as address, type of school, student per teacher ratio, ratings and percentile, star awards ratings etc. but we don’t need grade level rating so we are going to remove them and clean data and at last will be joined with main data set based in zip code value.

Links to data – <https://www.schooldigger.com/go/TX/city/Austin/search.aspx>

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>

**Learnings from this Course:**

As per the CRISP-DM methodology, data preparation is one of the key phases wherein we need to collect data, clean and transform data to make sure it is helpful to tackle the problem statement.

In real world not all data is ready to available for data analysis, we often need to search for right data source and collect data from such sources, as part of this course we got to learn different steps followed as part of data preparation phase.

* **Data Collection** - how to collect data from different sources– website tabular data, APIs, different Data Bases, file formats such as csv file, excel file, fixed width file, json files etc. Also, how to scrape data from websites not just tabular format data but in text format data.

While collecting data from different sources we need to make sure all the **PII** data is ignored or properly encrypted to adhere to the compliance policy and secure critical information. When web scraping, we often need to make sure we read thru the website policies about data scraping or the data is available for public.

Once data collection is done, then it is stored to the files or database tables.

* **Data Cleaning** – This is key step after collecting required data, because not all data is in clean shape, and it often needs to be cleaned before performing any analysis on the data. As part of this course, we learned what are the different methods, techniques, tools used to clean data such as identify null/na or missing values, find duplicates, outliers in dataset, transform the data, or change format of the data (for example date, currency, percentage etc). There are several techniques learned on how to identify and handle the data once na/null/duplicate values are identified. Also, learned different visualization to understand the data for missing values or outliers.
* **Data Storage –** The data is cleaned, formatted, and transformed, it is ready for next phase of cycle but before that to store such data is very crucial step since we may lose such cleaned form of data so storing them in database tables or flat files is required. We learned different packages to how to connect to different DBs and store data into tables. Also, learned on how fetch data from data base tables and consolidate them using different techniques. This are very helpful learning as in real time we often come across DBs wherein we need to store or fetch data from for data analysis.

**Steps Followed to Complete the Project:**

Since goal of the project is to collect, clean and transform all the data to make it ready format that will be suitable for machine learning model to build house prediction model.

We followed below steps:

1. Collected the data from different sources such as – API, Website and flat file.
2. Loaded all three datasets from different data sources into individual data frame using pandas package.
3. Cleaned the data, visualized the data - Checked the data summary and structure of the data - to understand what kind of variables are in the data and its types.
4. Checked for null, na or duplicated values and handled them appropriately.
5. Checked descriptive statistics for each dataset.
6. Added or removed unwanted features from the datasets.
7. Transformed the data - changed the categorical variable to numerical to fit to model, removed duplicates, performed aggregation to explore the data etc.
8. Explored and visualized the data - Plotted different charts to see the data distribution, outliers etc.
9. Checked 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. Finalized the features and made sure all are in correct format if not converted and standardized them so ML algorithm can fit on the data set.
11. Stored all the cleaned data into separate DB tables.

**Visualization and analysis:**

The data visualizations have been performed on the cleaned data to understand the how data is distributed, correlated etc.

* T**he House Price Distribution –** The histogram chart of home sale price shows the distribution of sale prices among the datasets. it shows that house prices are centrally distributed and don’t have many outliers.

**Chart, histogram

Description automatically generated`**

* **Correlation House Sale Price and Living Area –** The house sale price is the dependent variable and living area is exploratory variable, scatter plot shows the correlation between them and based on below chart we can see as living area increases as price of the house.

**Chart, scatter chart

Description automatically generated**

* **Correlation House Sale Price and House Age –** To see the correlation between sale price and house age, scatter plot is plotted and based on this chart we can say that as house age increase the sale price decreases.

**Chart, scatter chart

Description automatically generated**

Correlation - House Price Vs House Age

* **Correlation House Sale Price and Health Facility Area –** To see if any surrounding feature impact house price, here checked for health facility area with house price and it shows that bigger the health facility in that area house prices are higher.

**Chart, line chart

Description automatically generated**

* **Correlation House Sale Price and School Average Standard Score –** The house prices are also dependent on the school area so checked the correlation between sale price with school standard ratings and it does not show in correlation.

**Chart

Description automatically generated**

* **Correlation House Sale Price and Student/Teacher Ratio –** The house prices are also dependent on the school area so checked the correlation between sale price with student/teacher ratio and it does show that ratio increase the house price decreases.

**Chart, line chart

Description automatically generated**

* **Distribution of Number of bedrooms –** House price is also dependent on the number of bedrooms, so here we checked the distribution of house prices among different bedrooms, and we see that more the number of bedrooms higher the house price.

Chart, bar chart

Description automatically generated

* **Pair Plot -** This is basically scatter plot plotted for different pairs of features to see the collinearity between two variables. Based on this chart we can see that house prices are correlated to living area, age of house, number of bedrooms etc.

**A picture containing diagram

Description automatically generated**

**A picture containing table

Description automatically generated**

**Conclusion:**

As part of this project, I have implemented all the learning to make final dataset in ready format so that it will be passed to ML algorithms to make house price prediction model. There are few features kept in the data set such as urls, owner names, address for information purpose and can be ignored before feeding to ML algorithms.