CSE-564 Final Project Report

Exploration and Prediction of Housing Prices in the United States — Based on Zillow Data and U.S. Census Data

Wentao Wu (112524704) Shengwei Li (112516088)

1. Summary

We have accomplished and implemented most of our idea and visualization in our initial proposal. We mainly explored four aspects of the housing data and population distribution data, and visualized them respectively (figure 1). The whole dimensions of the housing data were fully excavated for finding the most dominant and popular house styles in each State of the United States. As to the visualization of these dominances, users could move the mouse over the main block in the dashboard (which is a USA map), and the home styles would pop out over the State. The correlation between houses' types and their prices are also explored, and this is visualized by two principal components plotting. The housing prices levels of each State were compared and could be easily seen in the two principal components plot by coloring them differently. Our work also implemented an accurate prediction of the pricing trend for selected areas and specific time periods, which can be insightful reference for people who are thinking about buying house for living or investment purposes.

We spent around a week in total in finding and merging the data, analyzing data, implementing visualization, and at last presenting our work. Actually, the data collecting, integrating, and digging insights of the whole dataset cost most of our time in this project. Although we have finished it, there are some aspects either in breath or in depth could be improved. For instance, housing price or population densities could be shown on the U.S map, scaling and linking function could be added, and prediction curve could be shown but not only the predicated numbers that we provide in the dashboard, etc.

We did not complete the housing price and local industry analysis, which we included in our project proposal. The reason is that these dimensions are not strongly correlated regardless how hard we tried to find some interesting and valuable aspects. We did find some, but these findings are more related with population data, which is diverged from our research subject. Although this part of work was not done as descripted in our proposal, we focused more on the Zillow housing data and census data.

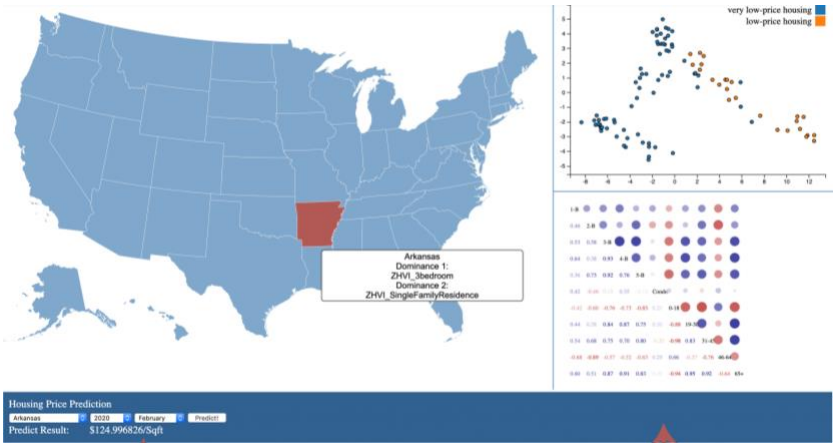

Figure 1. Dashboard of final project

2. Dominant Housing Styles Exploration

In this part, we analyzed the whole housing dataset for finding the most dominant house styles in each State of the U.S. The dominant home styles are the ones that affect and drive the whole housing price market to the most extend, indicating they are of the most popular house styles.

The main method we used in this part is PCA and the calculation of the highest attributes loading for the principal components of the dataset. Because there are more than sixty dimensions in our housing dataset, so before doing the PCA, we firstly cleaned the data and cut some dimensions out. For instance, the 'Median Price for All Homes' and 'Average Price for All Homes' and similar dimensions were deleted, while they were used as target attributes for coloring the plot clusters.

As the analyzing results, each of the State has its own characteristic in the first several dominant home styles. For instance, the Midwest States generally have one bedroom and two bedrooms as their dominant home styles, while in the northeast several States, the four-bedroom and five-bedroom houses are more popular. In addition, in Hawaii, the most popular house style are one-bedroom, condo, and four-bedroom houses, from this result we could conclude that those small houses might mainly for tourism. These dominant or popular home styles are correlated with the population distribution of different age groups, which we will discuss later in this report.

Michigan	ZHVI_4bedroom	ZHVI_CondoCoop
Minnesota	ZHVI_5BedroomOrMore	ZHVI_4bedroom
Mississippi	ZHVI_4bedroom	ZHVI_3bedroom
Missouri	MedianListingPricePerSqft_4Bedroom	MedianListingPrice_4Bedroom
NewYork	ZHVI_3bedroom	ZHVI_4bedroom
Hawaii	ZHVI_5BedroomOrMore	ZHVI_CondoCoop

Table 1. Screenshot of some State's dominant home styles

For each state selected by user, two of the most dominant housing styles will be shown (Dominance 1 and Dominance 2), as in figure 2. These are the housing styles that affect the whole housing price market to the most extend, indicating they are two of the most popular house styles in the state selected.

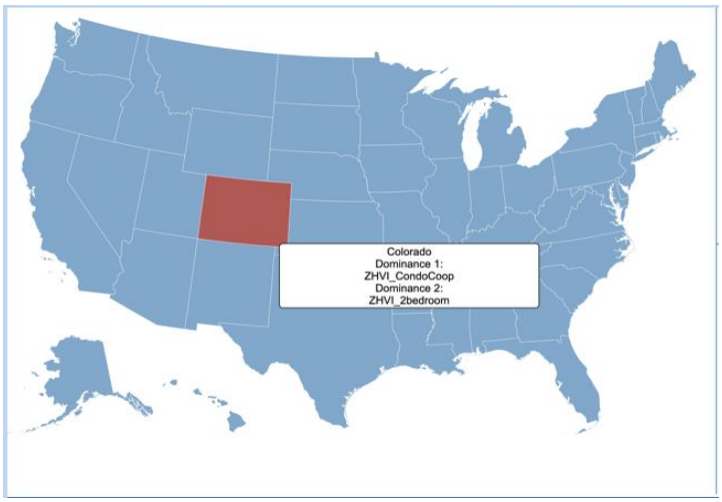

Figure 2. State selection for dominant house styles

3. Two principal components plot

This part of work is to visualize the data points in each State. From the graphs, we could find clearly the houses price distribution (Two plots, New York and Alabama are shown in figure 4). Two have a general idea of which price level the selected State is among the whole country, we classified 'median price for all home' dimension first into five groups according to the whole country's housing price, this is the target file. Then we cut off the 'median price for all home' dimension when doing the PCA analysis and the new coordinates creation for the new two principal components plotting. At last when load the created data to the D3, we add the target file again to the coordinate file so as to plot the data points in different colors according to their price levels.

```
def function_label(a):
    if a < 150000: return 'very low-price(<150k)'
    if a>=150000 and a< 250000: return 'low-price(150-250k)'
    if a>=250000 and a< 350000: return 'medium-price(250-350k)'
    if a>=350000 and a< 450000: return 'high-price(350-450k)'
    else:
        return 'very high-price(450k+)'
```

```
def function(s):
    State_df = pd.DataFrame()
    State_df = df[df['RegionName'] == s]
    State_df = State_df.fillna(State_df.mean().apply(np.round))
    #add the target column to the stritified data frame
    State_df['price'] = State_df.apply(lambda x: function_label(x.MedianListingPrice_AllHomes), axis = 1)
    target = State_df.loc[:,['price']].values
    State_df.to_csv(s,sep=',')
    return State_df
```

```
var circles = svg.selectAll('circle')
    .data(data)
    .enter()
  .append('circle')
    .attr('cx',function (d) { return 1000 + xScale(d.PC1) })
    .attr('cy',function (d) { return yScale(d.PC2) })
```

```
state_pca_data = {}
for i in range(0,len(state_pca_name)):
    state_pca_data[statenames[i]] = []
    csv_data = pd.read_csv('pca/'+state_pca_name[i]).to_dict(orient='records')
    for j in range(0,len(csv_data)):
        csv_data[j].pop('Unnamed: 0')
        state_pca_data[statenames[i]].append(csv_data[j])
```

Firgure 3. Code for making target file and PCA plot for each State

Figure 4 shows the visualization of this part's work; the chart is arranged at the upright of the main dashboard. This PCA plot shows the housing price data distribution of selected State among different time points. The more disperse the dots are, the greater price variation there was over the past several years. The different colors represent the relative price level of the selected State among the whole country's housing market.

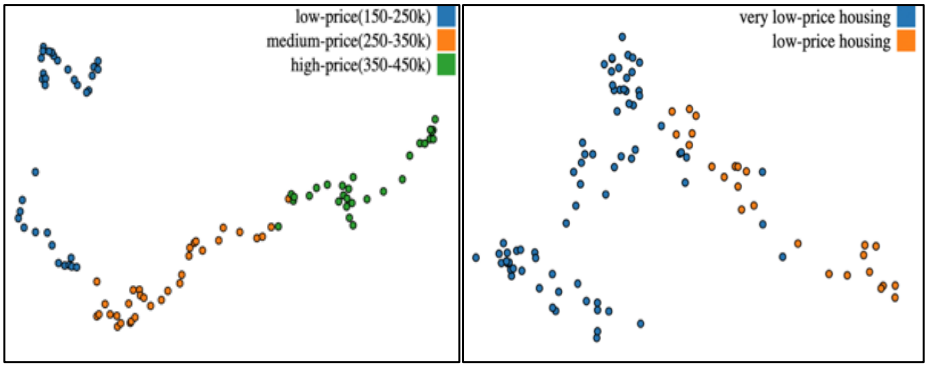

Figure 4. Visualization of data point distribution plotted into two principal components, New York and Alabama

4. Correlations Between House Styles and Age Groups

In this part, we merged the population distribution by age into each State and tried to find the correlations between age groups population fluctuation and home styles prices. Firstly, we classified the population into five different age groups due to the raw data from the U.S. Census.gov is recorded by single year old. We could not process those large number of dimensions if population of each year old be a single dimension. Therefore, we separated them into five, depending on people's age stage in which they may own different types of houses.

```
#divide age into groups
def pop_divide(a):
    if a>= 0 and a < 19: return '0-18'
    if a>=19 and a< 31: return '19-30'
    if a>=31 and a< 46: return '31-45'
    if a>=46 and a< 65: return '46-64'
    if a>=65 and a< 85: return '65+'
```

```
def state_pop(s):
    global pop_df
    State_pop_df = pop_df[pop_df['NAME'] == s]
    State_pop_df['AgeRange'] = State_pop_df.apply(lambda x: pop_divide(x.AGE), axis = 1)
    AgeRange = State_pop_df.loc[:,['AgeRange']].values
    State_pop_df.to_csv(s.replace(' ', '') + '_population',sep=',')
    return State_pop_df
```

Figure 5. Code for separating people into different age groups and redoing the statistics

The correlation matrix was built based on data collected from five different houses styles and five different age groups, we chose 1-5 bedrooms and condo as the main home styles. The visualization is a colored matrix as shown in figure 6, which is at the downright of the main dashboard. Blue and red color represents positive and negative correlations between attributes; and circle sizes indicate the strength of the correlation.

Figure 6 shows the correlations of New York and Colorado. In these two examples, in Colorado, populations of age 31-45 and 65+ have relatively higher positive relations with 3, 4 and 5 bedroom houses. However, in New York, the significant feature is that the 19-35 age groups affect almost

all the house styles prices, which might mean younger people drive the whole house market the most in New York. There are other interesting points could be found if we read more correlation matrix of different States and compare them. From these figures, we could have some insights of the home style price features along with the population distribution in specific State.

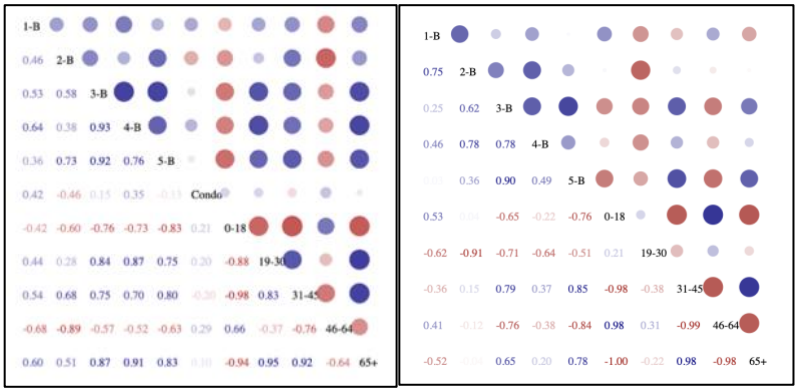

Figure 6. Correlation matrix of house style prices and age group population fluctuations New York and Colorado

5. Price Prediction

We have implemented the house price prediction based on the past ten years data from Zillow for each State. This part of work is arranged at the bottom of the main dashboard. Figure 7 shows the way how to use it. By choosing 3 key input factors -- State name, Year and Month, and click on the "Prediction!" button, the predicted housing price in $/Sqft will be shown below.

Figure 7. House price prediction

The prediction is based on model built by Keras[5]. We applied Keras deep learning package in building a house price model based on our exit data. The model will be created in a separate python file and loaded into our backend. Figure 8 is the code for loading and running the model, and answering to form which comes from our website.

```
json_file = open('houseprice.json', 'r')
model_json = json_file.read()
json_file.close()
model = model_from_json(model_json)
model.load_weights('houseprice.h5')
```

```
if request.method == 'POST' or request.method =='GET':
    predata = request.form.to_dict()
    prearray = np.asarray([[int(predata['Year']), int(predata['Month']), int(predata['States'])]])
    predictresult = model.predict(prearray)
    presult = "$" + str(predictresult[0][0]) + "/Sqft"
```

Figure 8. Code for house price prediction in backend

6. Conclusion

Here we have successfully found the most dominant housing styles in each State of the U.S. Using 2D PCA plot, we also illustrated the fluctuation and relative level among the US housing market. Moreover, we also presented the correlations between housing styles and different age populations. For future directions, more precise prediction model will need to be developed for specific home types.

References

[1] **https://www.zillow.com/**
[2] **https://www.census.gov/**
[3] **https://www.d3-graph-gallery.com/graph/correlogram_basic.html**
[4] **https://github.com/topojson/us-atlas**
[5] **https://www.kaggle.com/ironfrown/deep-learning-house-price-prediction-keras/notebook**

Final Project Preliminary Report

— Exploration and Prediction of Housing Prices in the United States

Wentao Wu(112524704) Shengwei Li(112516088)

1. Data Binding and Cleaning

Our data is from Zillow.com, most of the years' data, especially in the 1990s and early 2000s, was not recorded completely. In some year, some dimension's data was missed in a large scale. However, in the latest few years (from 2010 to 2017), numbers of different dimensions were detailed logged in each month. There is no 2018 and 2019 data in the CSV file. This is understandable because of almost no newest business data could be freely obtained. Therefore, in all the PCA and correlation analyzing processes, we used 2010-2017 year's data so as to avoid inaccuracy and also make our result up to date.

In PCA and attributes correlation analysis, because data items was monthly recorded in the initial CSV file. All the States' data was bound with time sequences, which is not directly useable for us. Therefore, we firstly bound the data entries into States. In this way, for each State, average values of specific dimensions were filled into the missing data points of the relevant dimensions. Because this project is primarily aim to exploration and prediction of housing prices, in this stage most of our work is to analyze the Zillow data, but not integrating the new age distribution data. We will try to add a block in our main dashboard in the future, to show the relations between age distribution and housing prices.

2. K-means Clustering

In this part, we firstly did K-means clustering by using all the dataset. In this way, we could get a more general look at the whole data. Here we got the appropriate k value is 5, so we separate each State's housing price into five levels. The code followed is to make cluster labels for data items.

```
strati_df = pd.DataFrame()
    n_clusters = 5
    persnt = 0.8
    for i in range(n_clusters):
        Clstr_i = np.where(kmeans.labels_ == i)[0].tolist()
        num_i = len(Clstr_i)
        sample_i = np.random.choice(Clstr_i, int(persnt*num_i))
        i_cluster_df = no_ch_df.loc[sample_i]
        strati_df = pd.concat([strati_df,i_cluster_df],axis = 0)
```

3. Principal Component Analysis

3.1 Target data frame creation

Target data frame is used to record the data items with their pricing levels, which was implemented by using the following code. We made a function based on 'MedianListingPrice_AllHomes' dimension so as to separate each State's houses into groups with the same standard. According to the clustering result, the group features are divided and named 'very low-price housing', 'low-price housing', 'medium-price housing', 'high-price housing', 'very high-price housing'. The following code is the implementing function, a new dimension named 'price' was created and integrated at the end of the columns, denoting which price level belongs to. The 'target' data frame is actually the price column of the 80% stratified data frame after adding price feature column.

```
# devide the 'MedianListingPrice_AllHomes' into 5 groups according to the price
level
    def function(a):
        if a < 180000: return 'very low-price housing'
        if a>=180000 and a< 250000: return 'low-price housing'
        if a>=250000 and a< 380000: return 'medium-price housing'
        if a>=380000 and a< 520000: return 'high-price housing'
        else: return 'very high-price housing'
    # add the target column to the stritified data frame
    strati_df['price'] = strati_df.apply(lambda x:
function(x.MedianListingPrice_AllHomes), axis = 1)
    target = strati_df.loc[:,['price']].values
```

3.2 PCA

By calculating the explained variance, we got four intrinsic components in our data set, the result is shown in figure 1. Here we used the whole data set in finding out the principal components' number. Actually, we tried doing PCA for every State at first. However, the result of the very first several States were the same, which indicated the intrinsic component number was four. Therefore, for unification in the future's visualization part, we choose four as our PCA result.

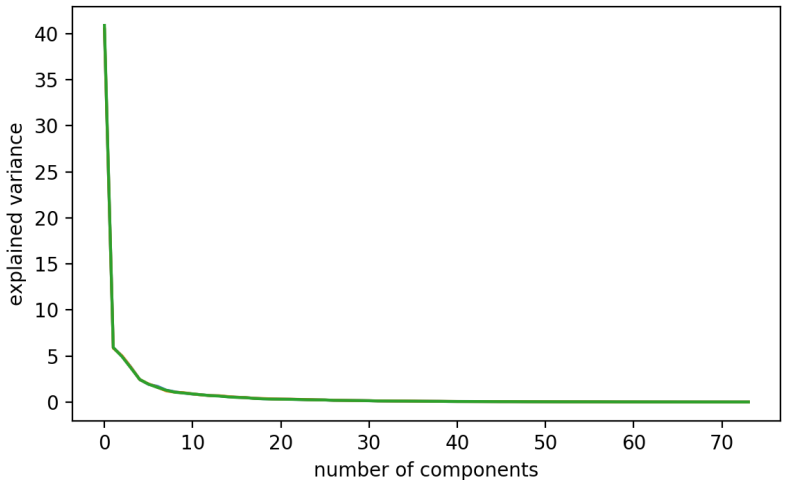

Figure 1. PCA explained variance plotting

4. Attributes selection and correlation matrix construction

We decided to choose two to three highest PCA loaded attributes for each State, these 2-3 attributes would represent the most significant factors influencing the housing price in that State. In other words, these are the most popular housings we are going to show in the visualization part, along with the two PCA coordinates plotting. And the price is shown along with the relevant attributes.

We have built up the attributes' correlation matrix. On the dashboard, the correlation network will probably be shown in the same area of the PCA plot showing area, interchangeable by clicking switch buttons. We only considered ten of the highest loading attributes to show. In our preliminary results in this part, take Alabama State for example, the median housing price has a high relation with the 3-bed room and 4-bed rooms' ZHVI pricing, and has low relation with the 'MedianRentalPricePerSqft_5BedroomOrMore'. Actually, we are now considering whether we should use the price per square feet of all homes as our main research subject, instead of median pricing of all homes. We will explore that whether our whole result would be optimized if we do so.

5. House Price Prediction

In this part, we applied Keras deep learning package in predicting future house prices. Data used here is temporary to simulate the complete data. First, we need to split data frame indices into train and valid part, we choose the seventy and thirty percentage.

```
# label_col is values we want to predict now
label_col = 'MedianListingPricePerSqft_AllHomes'
np.random.seed(1) # make np.random produce same result
data_train = np.random.permutation(data.index)[:int(0.7*len(data))]
data_valid = np.random.permutation(data.index)[int(0.7*len(data)):]
# x values are the parameters and y values is what to be predict
y_train = data.loc[data_train, [label_col]]
x_train = data.loc[data_train, :].drop(label_col, axis=1)
y_valid = data.loc[data_valid, [label_col]]
x_valid = data.loc[data_valid, :].drop(label_col, axis=1)
```

The next step is to Z-normalise the entire data frame and then convert them into NumPy arrays, which are used by Keras.

```
def z_score(col, df):
    mu = np.mean(df)
    s = np.std(df)
    newdf = pd.DataFrame()
    for c in col.columns:
        newdf[c] = (col[c]-mu[c])/s[c]
    return newdf
arr_x_train = np.array(z_score(x_train, data))
arr_y_train = np.array(y_train)
arr_x_valid = np.array(z_score(x_valid, data))
arr_y_valid = np.array(y_valid)
```

Then we began creating the Keras model, temporarily, we chose 3 layers and Adam optimizer to define our model.

```
def bulidmodel(x_size, y_size):
    t_model = Sequential()
    t_model.add(Dense(100, activation="tanh", input_shape=(x_size,)))
    t_model.add(Dense(50, activation="relu"))
    t_model.add(Dense(y_size))
    t_model.compile(loss='mean_squared_error',optimizer=Adam(),metrics=
[metrics.mae])
    return(t_model)
model = bulidmodel(arr_x_train.shape[1], arr_y_train.shape[1])
```

The last step is to fit and train our Keras model. Here we chose the batch size of 128 and 500 epochs and adopted the EarlyStopping, which watches the model measurements and stops fitting when no improvement. Besides, we recorded the training and validation history. As we specified EarlyStopping with patience=20, with luck the training will stop in less than 200 epochs. And last but most important we evaluated the performance of the trained model.

```
keras_callbacks = [EarlyStopping(monitor='val_mean_absolute_error', patience=20,
verbose=0)]
history = model.fit(arr_x_train, arr_y_train, batch_size=128, epochs=500,
shuffle=True, verbose=0, validation_data=(arr_x_valid, arr_y_valid),
callbacks=keras_callbacks)
train_score = model.evaluate(arr_x_train, arr_y_train, verbose=0)
valid_score = model.evaluate(arr_x_valid, arr_y_valid, verbose=0)
```

After obtaining the model, the price depend on the time and location could be predicted. However, for now the model is not as good as we expected. To get it better, first we will try to combine the model with more principal aspects derived from the PCA result. In such way our model could be improved by trainning with more data. Besides, we would try to make better use of the power of Keras, try to maintain more layers and use more concept such as regularisers and dropouts. Overall,

our predictions could be more specific and efficient, we will continue focusing on improving it.

6. Visualization

We have built up the frame of our dashboard, it will be an U.S. map implemented by D3, with the open source json data: U.S. Atlas TopoJSON . The most popular house style and its price will be shown by cilcking each state. There are also two blocks will be present by the side of the main graph, demonstrating the results we obtained obove. The code followed is how we implement it for now; and figure 2 showed the graph we are building on.

```
d3.json("https://cdn.jsdelivr.net/npm/us-atlas@2/us/10m.json", function(us) {
     svg.append("g")
        .attr("class", "states")
        .selectAll("path")
        .data(topojson.feature(us, us.objects.states).features)
        .enter().append("path")
        .attr("d", path);
    svg.append("path")
        .attr("class", "state-borders")
        .attr("d", path(topojson.mesh(us, us.objects.states, function(a, b) {
return a !== b; })));
  });
```

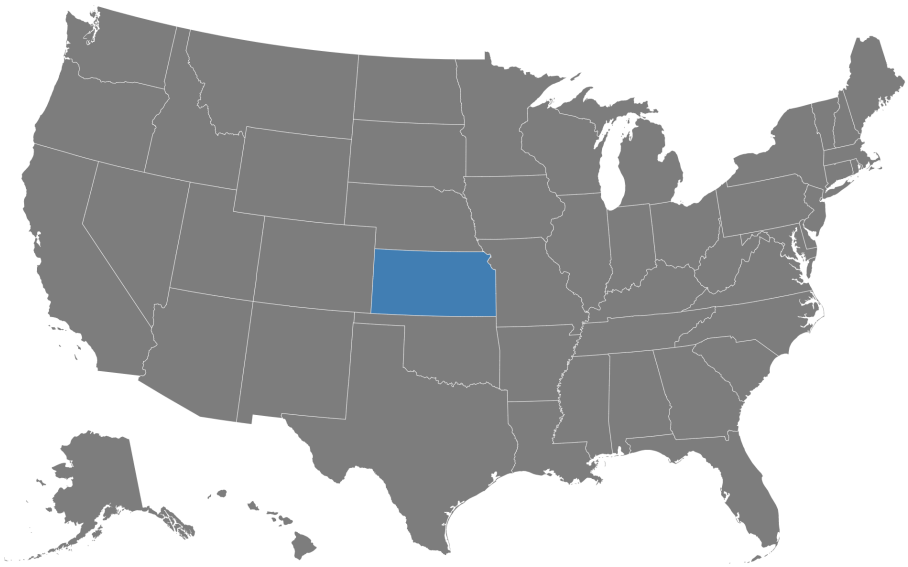

Figure 2. D3 US map

Exploration and Prediction of Housing Prices in the United States — Based on Zillow Data, Population Distribution and Economic Characteristics Data

Wentao Wu (112524704) Shengwei Li (112516088)

1. Introduction

Housing has always been one of the major concerns of most people's lives. People need places for living, working, exercising, and so on, by either owning or renting a housing property. In the United States, around 5 million different types houses are sold each year [1]. The pricing of houses can be affected by various factors, for instance, historical sale prices, size, appeal, locations, etc.; which are fairly obvious and well-perceived by people. While there are also some implicit factors that play equal or more important roles in regulating house pricings. These reasons could be the inner driver for the prices; and they are largely inter-correlated. Revealing of these inner driving forces requires deep and comprehensive exploration of the data available.

When it comes to buying a house, the more factors are taken into consideration, the more likely that one will choose the most suitable one. And the most important attribute of a house is its location and price. In our project, we will try to resolve part of these concerns with data analysis and an interactive visualization. Furthermore, based on our results and conclusion, we will attempt to give people some guidance in choosing the right house in the near future.

2. Executive Summary

This project is mainly aimed to help people understand what the most popular house styles in each state of the United States are. We will also explore and show how these best-selling/renting houses affect the overall house pricing in different states. By adding more dimensions into the data file, we will analyze the intrinsic interplays among the most popular houses, distribution of popularity (ie., age distribution), and economic characteristics (ie., major industry types), in the whole country as well as in each state. For instance, is there a correlation between the price of a 1-bedroom homes and the buyer's age; are loft sell/rental prices higher in those business intensive states; what are the major factors that affect the inventory and the price per square feet, and so on. In this project, we will explore and answer these types of questions by visualization of data. Because our original data set has a ten-year span, we want to visualize more by generating new data from it and make this work more valuable. The ideas would be illustrated in section 3, and the implementation approaches will be described in section 4. Each part of our work comes with a visualization. Because our data is mainly featured by area and time, most of this project is about correlating the data into these two aspects. Based on this feature, the main visualization work will be built on a

web page, which shows an US map. By interacting with the map, all the results will be shown with the form of graphs/plots.

3. Statement of Need

The main part of the data is from Zillow.com, which is about buying and/or renting prices of different housing types of in the fifty states of US plus District of Columbia. These data include multiple dimensions such as house inventories, 1-bedroom average sell/rent price, 2-bedroom average sell/rent price, loft average sell/rent price, whole average sell/rent price, etc. We will also try to mine more data from other source and use data integration to get a bigger dataset. For instance, we are planning to add the age distribution data and economical characteristics as new dimensions. Specifically, each age group would be a new variable, and so are the major industries. Each column of the major industries stores the economic size to the whole. The aim of merging these data is to try to find the correlations between the housing price, age distributions and the major industries.

We will also try to find the dominant house types. In another word, the house type that mainly contributes to the average of house selling/renting prices. Besides finding the most popular home types, we also plan to explore which types of the houses are even highly correlated with the population distribution and area major industries. Suppose there are strong correlations between them, we could conclude that home selling/renting prices does affect or can be affected by the people and economical distributions.

Merging data from different sources can be time consuming. Because in some analysis steps, we need to divide data items into different groups, such as different areas or time series, therefore merging new data into the raw may require a lot of binding work. In addition, analyzing the relations between different aspects of data source can be complicated, especially when trying to find the interplays between them. The visualization part is also important, which we expect will take at least half of our project time to finish. Although we have already had some ideas about how to perform the visualization work, to perfectly accomplish the entire project will still take time.

Apart from visualization of the existing data, we will also consider about its implications for the future. Although the changes of prices are affected by many attributes, which makes it difficult to simulate the precise price changes in the future, an inaccurate data prediction still can be insightful, and can be a reference for those people who are thinking about new investments. Therefore, according to the existing data, we will try to predict how the prices will change in the future as well. There are multiple ways to achieve this -- a convincing way would be machine learning, which is a powerful tool to predict future trend based on current data analysis. Build a close price mode can also be time consuming, and we anticipate using the other half of our project time to accomplish this part.

4. Project Description

4.1 Data cleaning

There are more than seventy thousand data items collected by Zillow company for house pricings in different states. Because a large part of data lack in the very early years from 1995 to 2000, we decided to ignore those data items for the purpose of maintaining consistence. Also, to make the analysis more representative and up to date, we will use the data from year 2010 to 2019, which are sufficient enough for this project. We will also try to fill the gaps between data by referring to the data from neighbor months in the corresponding state.

4.2 Housing price exploration

In this section, we conduct a clustering first based on the whole data sets. Noticing that we have more than 15 thousand data items, if the clustering runs slow, we will try to stratify half of the data from each year. K-means clustering and Density-based spatial clustering will be used in this part of work, and clustering results will be compared between the two methods. Although we plan to use both of the two methods, the K-means method would probably work better for our project based on our experience with the previous projects. The clustering result will show how many groups will all the data sets belong to. Assuming that there are five clusters, we will divide the housing price into five levels -- very low, low, medium, high and very high housing prices. We can also divide them into four or six groups based on how many clusters we get.

We will also perform dimension reduction to the dataset. The main method would be principal component analysis (PCA), which aims to find the intrinsic dimensions of the housing data. In this way, the most dominating dimensions in the raw data could also be found by calculating the highest loadings contributing to the principal components. The dominant dimensions will be some of the house style selling or renting prices, which therefore are the main factors that contribute to the housing price. From the results of this part work, we will get all the fifty states' most dominant factor(s). In the visitation part, a force directed US map will be built, therefore every states' situation will be shown by mouse interactions. In this way, people could easily get a general idea about the housing situations in different states of the US.

Besides PCA, we will also build correlation metrics by calculating dimension correlations. In this part of work, the age distribution data and economical characteristics data would be adding as new dimensions. Specifically, each age group would be a new variable, and so are the major industries. In each column of the major industries, the economic size to the whole are shown.

The graph below is an example of what will be shown in our main website page (Figure 1). A force directed US map will be built. By moving mouse over each state, general information such as the dominant housing price will be shown. By clicking them, you will jump into another page, which will show the correlations metrics/plots.

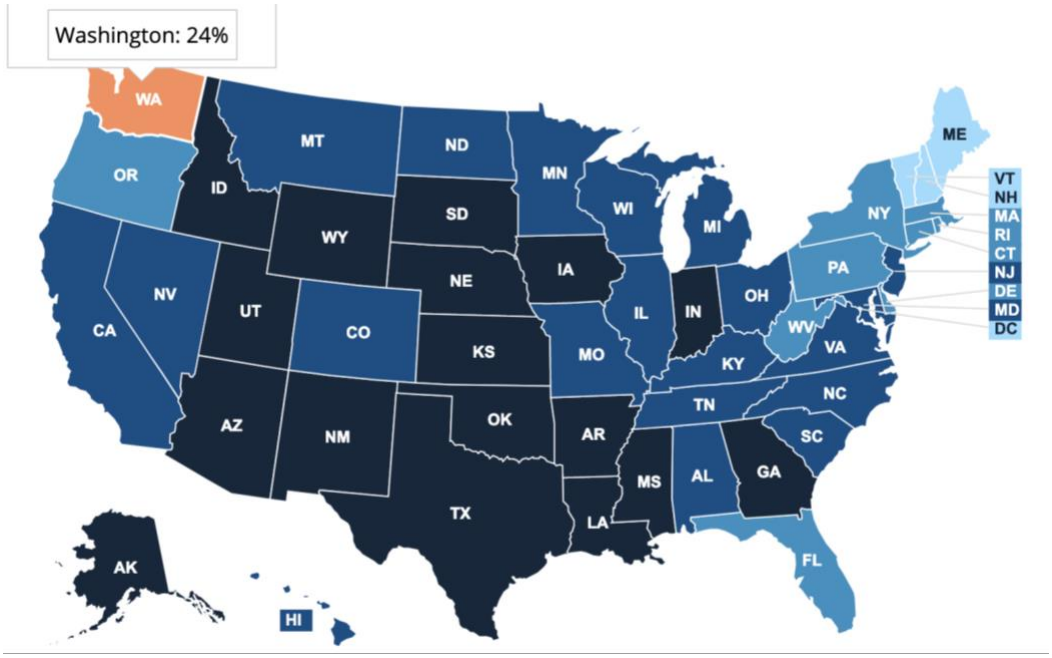

Figure 1. Visualization map sample [2]

4.3 Housing price prediction

We will draw a house price curve among different states. Then we will attempt to visualize these existing data in an interactive way -- with mouse click on each state, an interactive curve will show up, and the details can be seen with the mouse move (Like in Figure 2).

Figure 2. Sample of interactive curve [3]

Besides analyzing the existing data, we plan to make the data helpful for people who are looking for a house in the near future. To implement this, we will add a predict button, where people can type in a future time, and will get a predicted price change within certain time and areas. For the prediction part, we will use the neural network concept through the Keras[4] and TensorFlow[5].

To be more specific, we will normalize the some of our data first. Then split the dataset into features and labels, which are to be predicted. We will also attempt to build prediction models and train them for each state. Since the data size is limited, it is not a good idea to build a complicated architecture. Therefore, we will add less layer to avoid over fitting. The most important work in this part is to find the most suitable parameter, which is to fill in the model template and track the loss function so as to continuously maintain model reliability. The model building and training work will be time consuming, and it is unpractical to perform computations upon click on the button for prediction. As a result, we will build the model in advance and store them back in the website, which makes the result immediately showing up when interacting with it.

5. Conclusions

This project will help people understand the current overall housing market in the US, as well as get an idea of the future trend as a guidance for house-purchasing directions. To sum up, in this project, the dominant factors that influence the pricing of different house styles will be shown; the interrelations among house prices, population distributions and industry distributions will be displayed; and there will also be a price prediction according to different states and different time periods. Finally, the whole work will be presented with a US map-based visualization, which will be interactive and easy to understand.

6. References

[1]https://www.statista.com/statistics/226144/us-existing-home-sales/

[2]https://www.kff.org/other/state-indicator/distribution-by-age/?activeTab=map&currentTimeframe=0&selectedDistributions=children-0-18&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

[3]https://medium.freecodecamp.org/how-to-build-historical-price-charts-with-d3-js-72214aaf6ba3

[4]https://hackernoon.com/build-your-first-neural-network-to-predict-house-prices-with-keras-3fb0839680f4

[5]https://www.kdnuggets.com/2017/04/simple-understand-gradient-descent-algorithm.html