**CSE 564 Visualization Project Two Short Report**

My project is to evaluate the factors that affect the American housing price. The 2016 and 2017 all 50 states and District of Columbia house price indexes were analyzed in this project. There are 27 dimensions in my data frame, including house inventory, house price with different bed rooms, rent price for all types of houses, and cut price, etc. However, the intrinsic dimensionality is 4 after principal components analysis. To show different plots, I made 4 html files that are linked to each other in the front end. The index.html includes the three PCA scree plots (raw, randomly and stratified sampled data), the 2dplot.html includes the two PCA vectors via 2D scatterplot, the mds.html includes both the Euclidian & correlation distance calculated 2D MDS scatterplots, and the attri\_matix.html is used to show the scatterplot matrix of the three highest PCA loaded attributes. Data will be passed into the front end when requested.

**Random Sampling and Stratified Sampling**

There are around 1200 data points in my CSV file, so after some data cleaning process, I randomly sampled 50% of them.

A screenshot of a cell phone

Description automatically generated

In stratified sampling process, the optimal of k value is 5 after using the Elbow method. Knowing the cluster number is 5, then the K-means method was used clustering all my data points. To implement the stratified sampling, I used the following method: store the data points into a list, randomly choose 50% of each cluster, and then concatenate the chosen data point into a new data frame named strati\_df.

A close up of a map

Description automatically generated A close up of text on a black background

Description automatically generated

**Intrinsic dimensionality**

Before doing the principal components analysis, I divided the ‘house median price’ into 5 levels which are ‘very high-price housing’, ‘high-price housing’, ‘medium-price housing’, ‘low-price housing’, and ‘very low-price housing’. Therefore, when project the data points into the two PC vector scatter plot, the different levels could be plot in different colors.

A screenshot of a cell phone

Description automatically generated

To do the PCA scree plot, eigenvalue were calculated by fitting the raw, random sampled and stratified sampled data. The following lists the implementation method. After ploting the graph,

I found my intrinsic dimentionality is 4. So in the next step, I use 4 components to calculate the 3 top-loading attributes.

A picture containing object

Description automatically generated

A close up of a logo

Description automatically generated

A picture containing object, clock

Description automatically generated

A close up of a sign

Description automatically generated

A picture containing object

Description automatically generated

A picture containing object

Description automatically generated

**Three highest PCA loadings attributes**

I used the following code to print out the result of the PCA loadings of all the initial attributes.

A screenshot of a cell phone

Description automatically generated

The result was sorted, and the three highest PCA loadings attributes are MedianListingPrice\_SingleFamilyResidence, MedianListingPrice\_3Bedroom, and MedianListingPrice\_2Bedroom. Therefore, I used these three attributes in the task 3 when visualizing the scatterplot matrix.

A close up of a newspaper

Description automatically generated

**Plot the MDS (Euclidian & correlation distance) in 2D scatterplots**

The following code is to implement the euclidian and correlation distance based scatterplots. Because my data is not very good clustered, so in the Euclidian distance based MDS plot, the low-priceand medium-price housing was not clearly seperated. But the high and very high price housing were greatly seperated.

A screenshot of a cell phone

Description automatically generated A close up of text on a black background

Description automatically generated

**Scatterplot matrix of three highest PCA loaded attributes**

The attributes MedianListingPrice\_SingleFamilyResidence, MedianListingPrice\_3Bedroom, and MedianListingPrice\_2Bedroom were chosen to plot the matrix. The plots were also colored. As my stratified data used in the last steps were around 800 points, so I did not use the raw to do the MDS and scatterplot matrix. However, I found stratified data is good enough to finish the task.

A picture containing object

Description automatically generated

The data frame used in this part is selected from the stratified data frame adding the ‘price’ column which is the target.

**Multi-page Implementation of Front End**

In each html file, I made several buttons and links so to implement different plots show.

The following code is in one of the html files.

A close up of a logo

Description automatically generated

After doing this project, I really got a deep understanding of data frame and how to process big size data. Furthermore, I could also visualize some of the key features of my data, which is cool and could make other people easily understand my data too.

You can find my video at <https://youtu.be/q5lSeDyhvec>. Thanks for watching!