1. Introduction / Business Problem

Cleveland is a major city in the U.S. state of Ohio. It is located on the southern shore of Lake Erie and after its founding grew into a major manufacturing center due to its location on both a river and lake shore. The city is home to a vibrant arts and culture scene, internationally recognized hospital systems, abundant city parks, and several major league professional sports teams. The estimated 2019 population was 381,099 and ranks as the 53rd-largest city in the United States.

Like many cities once reliant on manufacturing, it is considered a 'rust belt' city, so-called because of the economic devastation caused by the disappearance of manufacturing jobs due to automation and offshoring. Cleveland has seen a resurgence in the recent decade; however, the economic prosperity widely differs between neighborhoods. Housing prices vary widely depending on the neighborhood. Some neighborhoods have robust amenities, while others sorely lack basic amenities such as grocery stores.

This analysis will focus on the ties between available amenities and the effect those amenities have on housing pricing, through the lens of neighborhoods. What types of amenities have the most influence on housing prices? Which neighborhoods are currently doing well and what can we learn from those neighborhoods to help those that are struggling? This information is beneficial to city planners and community development corporations as they aim to create more equitable neighborhoods to help advance the economic prosperity of Cleveland as a whole.

2. Data Collection

To consider the above state problems, the following data was collected:

- Housing Prices: Zillow provides smoothed, seasonally adjusted measures of the typical home value and market changes across a given region on a monthly basis. It reflects the typical value for homes in the 35th to 65th percentile range.

- Amenities: The Foursquare API provides lists of venues with location data that will be used to map an amenity to a neighborhood.

- Neighborhood Boundaries: For mapping purposes, a .geojson file of Cleveland neighborhoods was obtained from GitHub. These neighborhoods match the neighborhoods provided in the Zillow Housing Prices data.

3. Methodology

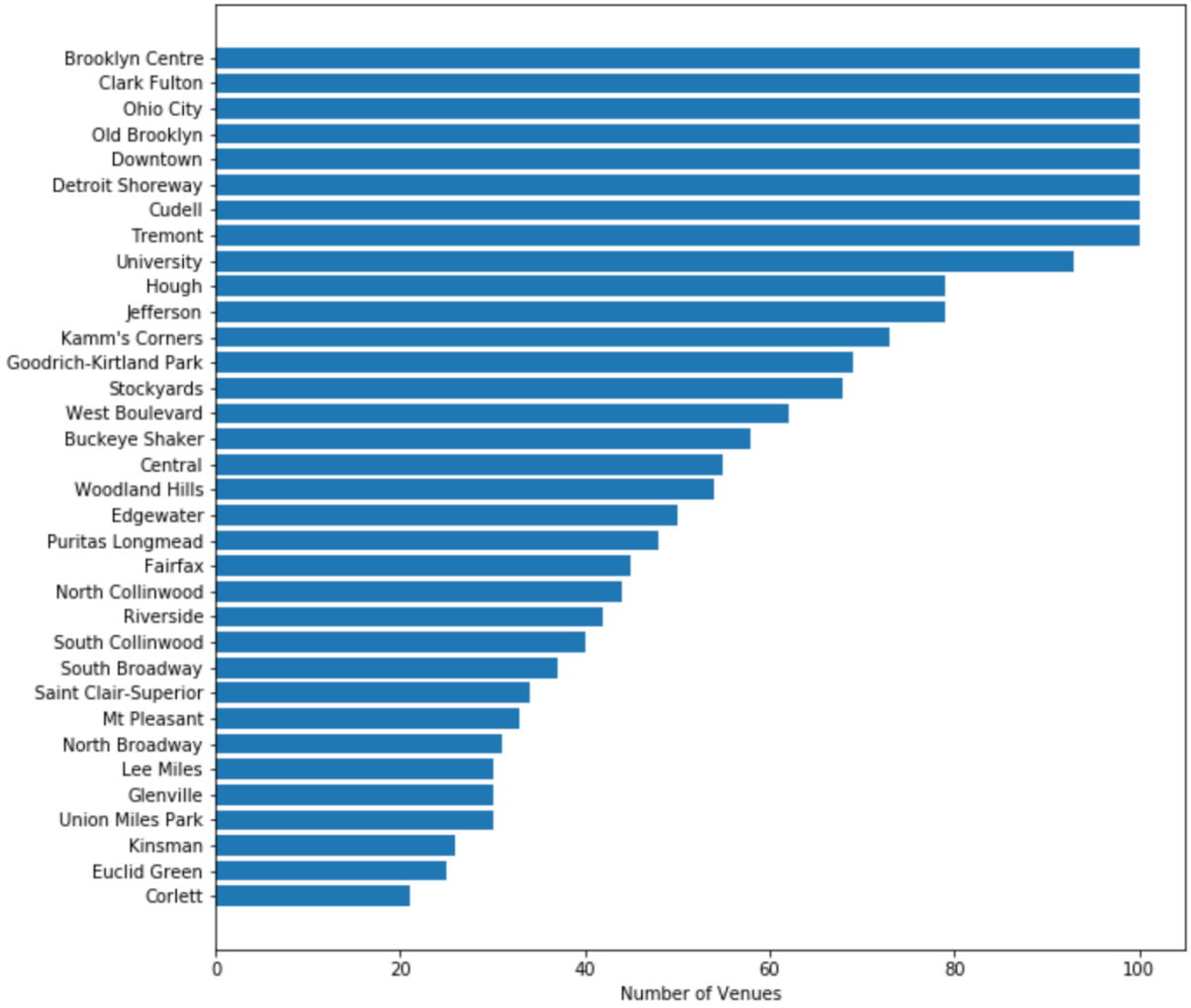
3.1. Data Preparation

Using the .geojson file of Cleveland neighborhoods, central points of each of the 34 neighborhoods were calculated and stored in a data frame. Once those points were obtained, I utilized the Foursquare API to generate a list of venues within 2000 meters of the central point for each neighborhood. The Foursquare API only allows 100 venues to be returned, so the venue data is not entirely complete based on this limitation. A total of 2,056 venues in 233 unique categories were returned and added to the existing data frame.

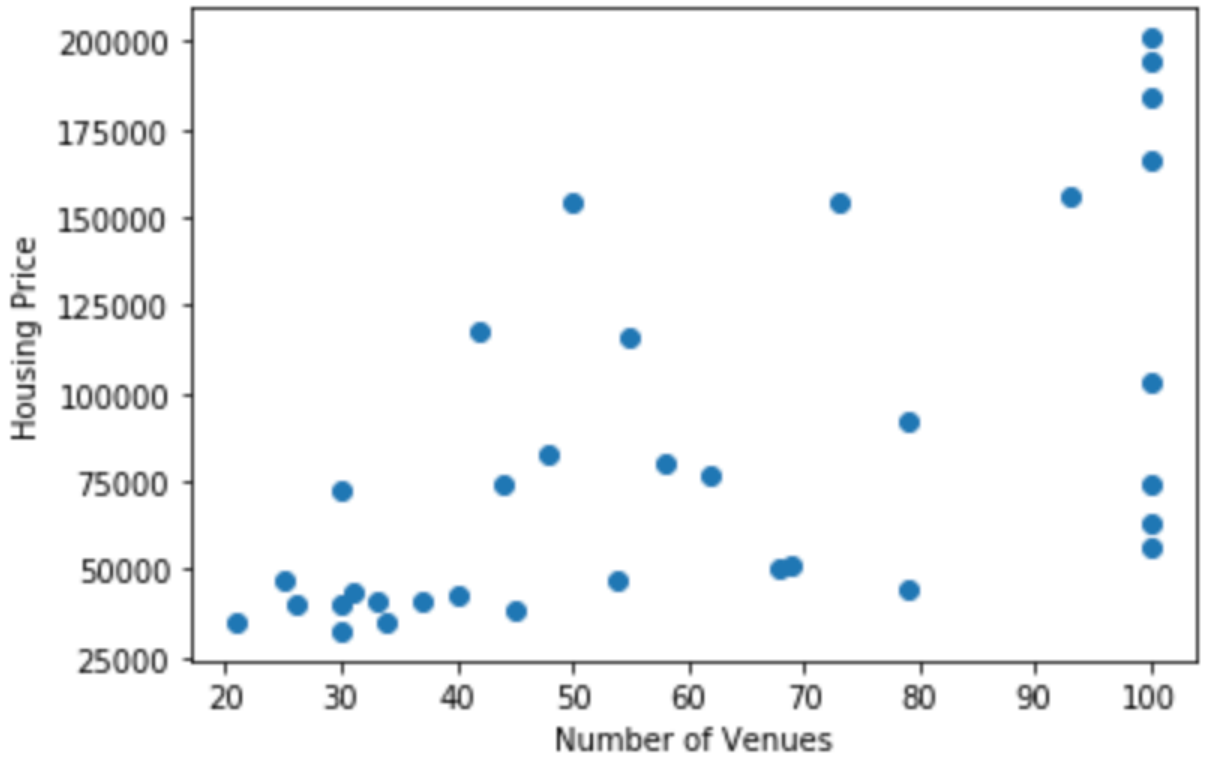
Housing pricing data was obtained from Zillow's Research hub for Cleveland neighborhoods in 2020. The data is provided by Zillow on a monthly basis, so to simplify the data the mean of 2020 was calculated and added as a column to the data frame. Monthly columns were then dropped, so all that remained was a data frame containing the name of the neighborhood and the mean housing value in that neighborhood for the year 2020.

3.2. Data Exploration

Basic data visualizations reveal the spread of how many venues are returned for each neighborhood. Eight neighborhoods returned the maximum number of venues allows by Foursquare, likely indicating that even more venues are located in that neighborhood. Some neighborhoods returned as few as 21 venues, with a wide amount of variance in between. This does limit the findings, as the same amount of data is not being compared for each neighborhood.



Creating a scatterplot comparing the number of venues in each neighborhood to the mean housing price shows that these two features do not have a strong correlation. Some neighborhoods with the maximum number of venues are among those with the lowest mean housing price, with a large amount of variance in between. This implies that the type or category of venue is likely important, rather than just the presence of a venue.



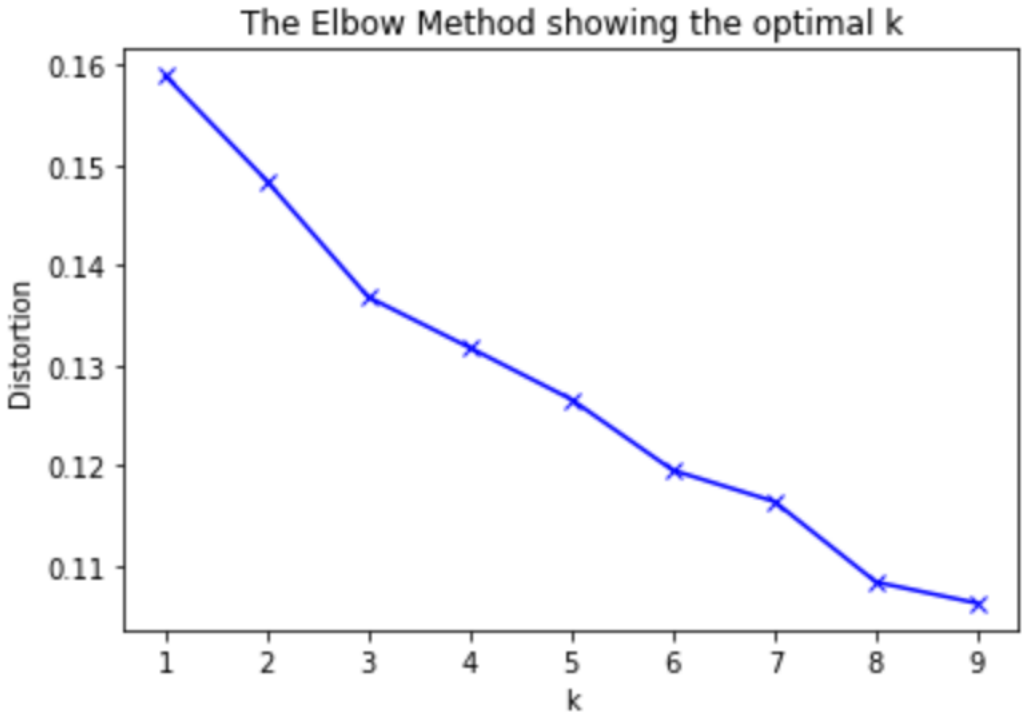
3.3. Analyzing Each Neighborhood

Common venue types for each neighborhood were obtained by using a combination of one-hot encoding, grouping, and sorting. This resulted in a data frame containing the name of the neighborhood, as well as a ranking of the most common venue types from 1 to 10:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Brooklyn Centre	Zoo Exhibit	Sandwich Place	Pizza Place	Pharmacy	Diner	Coffee Shop	Discount Store	Bar	Playground	Department Store
1	Buckeye Shaker	Light Rail Station	American Restaurant	Discount Store	Bank	Sandwich Place	Park	Pharmacy	Clothing Store	Southern / Soul Food Restaurant	Diner
2	Central	Grocery Store	Fast Food Restaurant	Sandwich Place	Gas Station	Fried Chicken Joint	Imported Food Shop	Chinese Restaurant	Discount Store	Coffee Shop	Performing Arts Venue
3	Clark Fulton	Zoo Exhibit	Sandwich Place	Diner	Department Store	Gas Station	Grocery Store	Pharmacy	Pizza Place	Dive Bar	Fast Food Restaurant
4	Corlett	Discount Store	Sandwich Place	Fast Food Restaurant	Gas Station	Grocery Store	Deli / Bodega	Donut Shop	Liquor Store	Restaurant	Cosmetics Shop

3.4. Clustering Neighborhoods

Using k-means clustering, neighborhoods were grouped according to similarities in common venue types. Prior to conducting the k-means clustering, the elbow method was used to determine the appropriate number of clusters. The diagram below shows that 3 clusters should be used for optimal results.

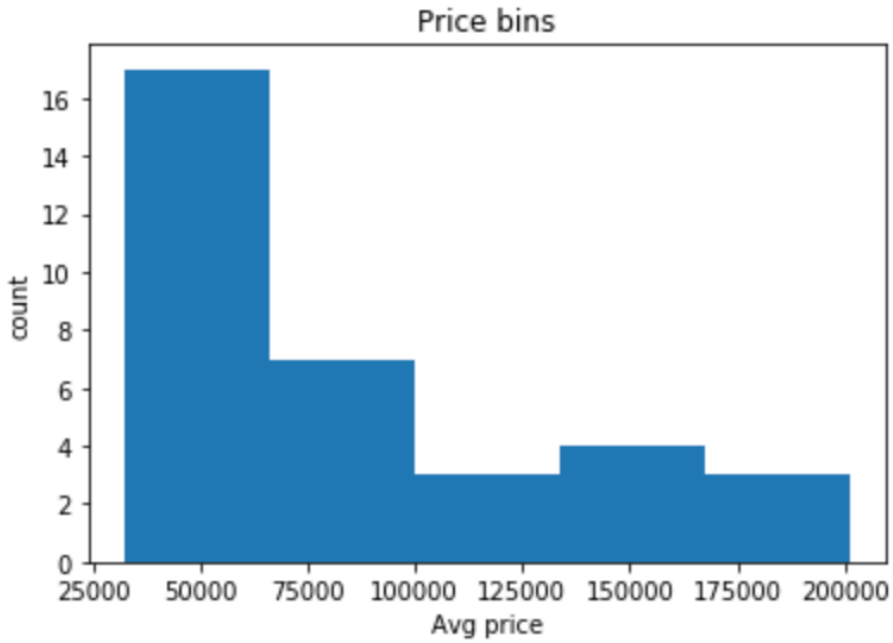


With 3 set as the optimal k value, 3 clusters of neighborhoods were returned. Examining these clusters, we can assign them the following labels:

- Cluster 1: Restaurants and Culture
- Cluster 2: Discount Stores and Fast Food
- Cluster 3: Delis and Convenience Stores

3.5. Binning Housing Prices

A histogram visualization helped create usable categories out of the mean housing prices for each neighborhood.



From this distribution, an array of values was stored and each category given the following names:

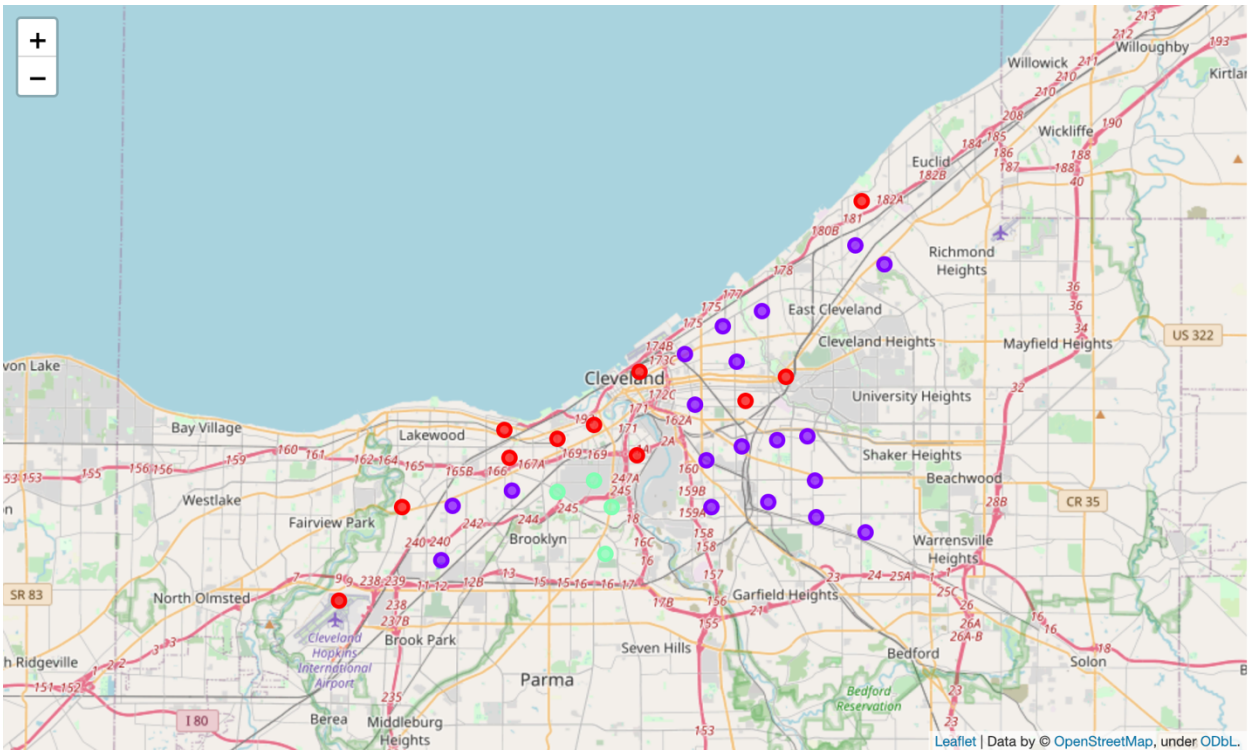
- Low Level 1
- Low Level 2
- Average Level
- High Level 1
- High Level 2

Adding these pricing categories to our data frame, the final data resembles the following:

	Neighborhood	housing_mean	latitude	longitude	Cluster Labels	Cluster Name	Price-Categories
0	Old Brooklyn	103011	41.433555	-81.704266	2	Delis and Convenience Stores	Average level
1	Kamm's Corners	154741	41.452543	-81.814195	0	Restaurants and Culture	High level 1
2	Jefferson	92120	41.452883	-81.786845	1	Discount Stores and Fast Food	Low level 2
3	Puritas Longmead	82996	41.430653	-81.792927	1	Discount Stores and Fast Food	Low level 2
4	West Boulevard	77171	41.459105	-81.755003	1	Discount Stores and Fast Food	Low level 2

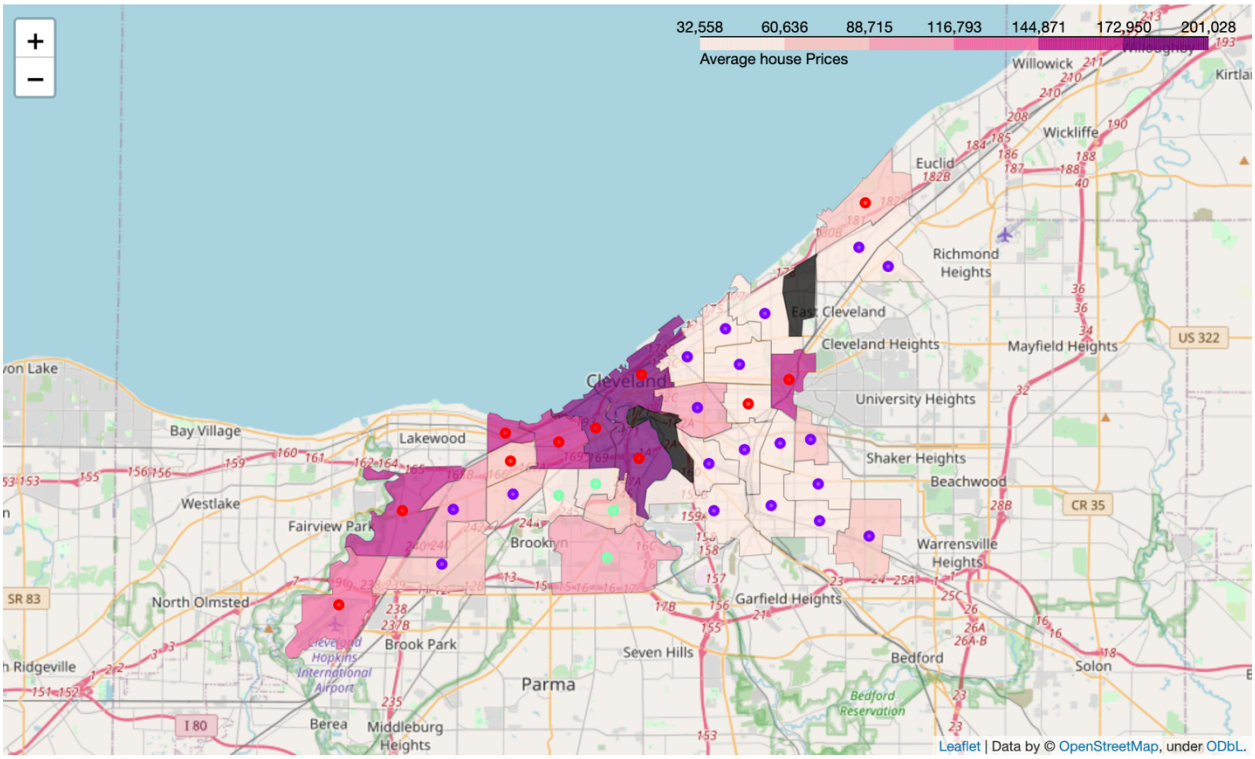
4. Results

Using folium, the results of the neighborhood clustering can easily be viewed and explored. The highest density of Restaurants and Culture Neighborhoods can be found on the northern border of the city along Lake Erie. The Discount Stores and Fast Food neighborhoods are largely found east of the Cuyahoga river and are physically clustered together as well.



Red: Restaurants and Culture
Purple: Discount Stores and Fast Food
Green: Delis and Convenience Stores

Layering on the housing values using Chloropleth, a correlation between neighborhood type and mean housing price can easily be seen.



5. Discussion + Limitations

All of the neighborhoods with high level housing prices fall into the Restaurants and Culture category. It is important to note that just because a neighborhood is categorized as Restaurants and Culture that it automatically has a high housing price. The Fairfax, North Collinwood, and West Boulevard neighborhoods fit this classification, but have some of the lower mean housing prices in the city.

The limited amount of venue data able to be obtained from Foursquare should also be considered when evaluating these results. Venues were obtained based on the geographically central point of a neighborhood, not necessarily where the most venues are located in a neighborhood. Only being able to pull 100 venues for a neighborhood could also drastically affect the clustering. In a neighborhood like Downtown, there are likely thousands of venues, but only the 100 provided by Foursquare were considered.

6. Conclusion

It is interesting to see that there is in fact a strong link between the types of amenities a neighborhood contains and the housing sale price for that neighborhood.

This analysis focused mostly on categorizing neighborhoods, but it would be worthwhile to see what types of venues and amenities have the greatest impact on housing sale prices. Are restaurants and breweries in fact the most influential venues you can add to a neighborhood? Does the quantity of each type of venue matter? Are there less common types of venues that actually have a greater effect on prices even though they occur less often? A more granular approach to venue and amenity analysis could help answer these questions.