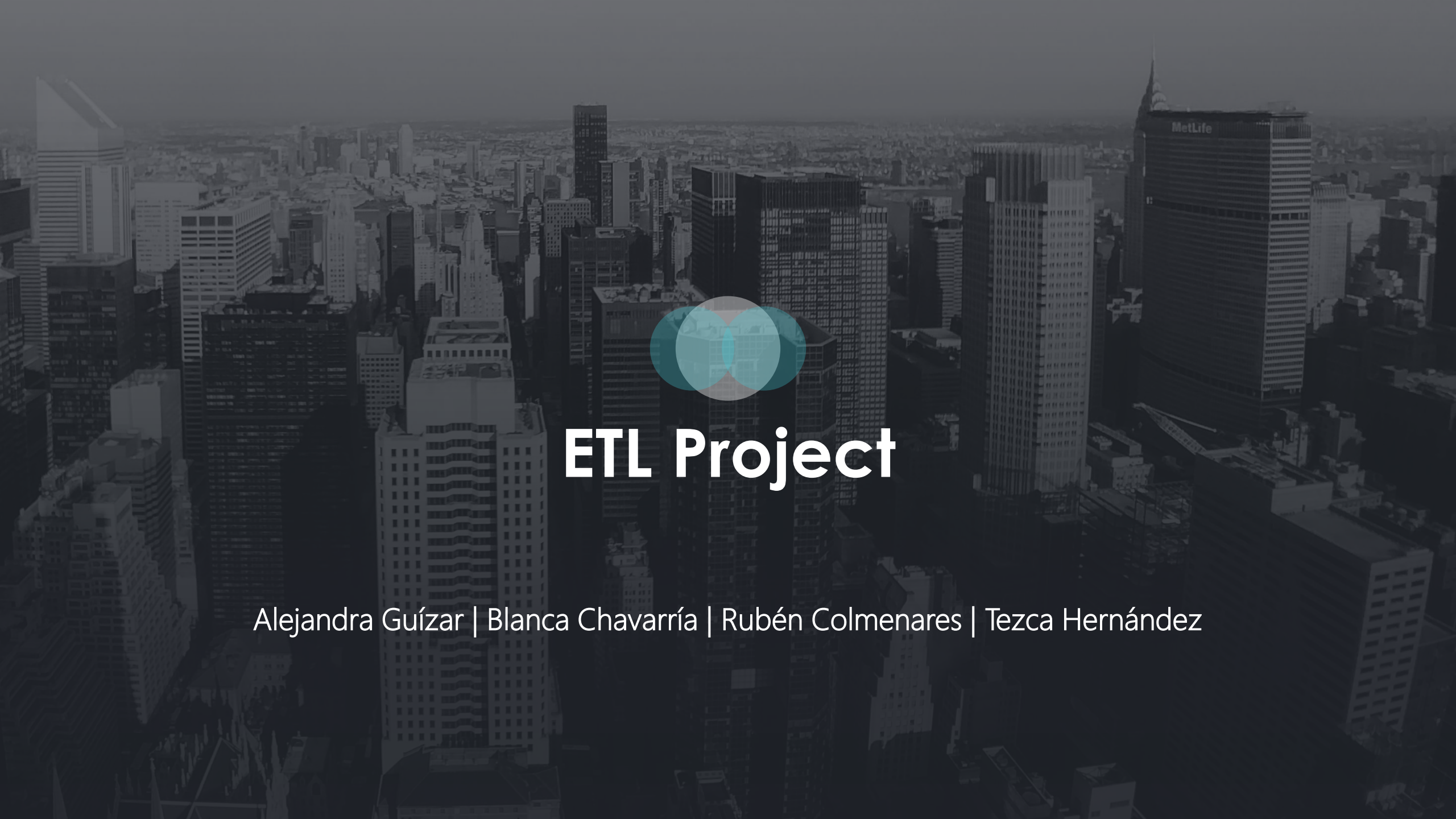


ETL Project

Alejandra Guízar | Blanca Chavarría | Rubén Colmenares | Tezca Hernández

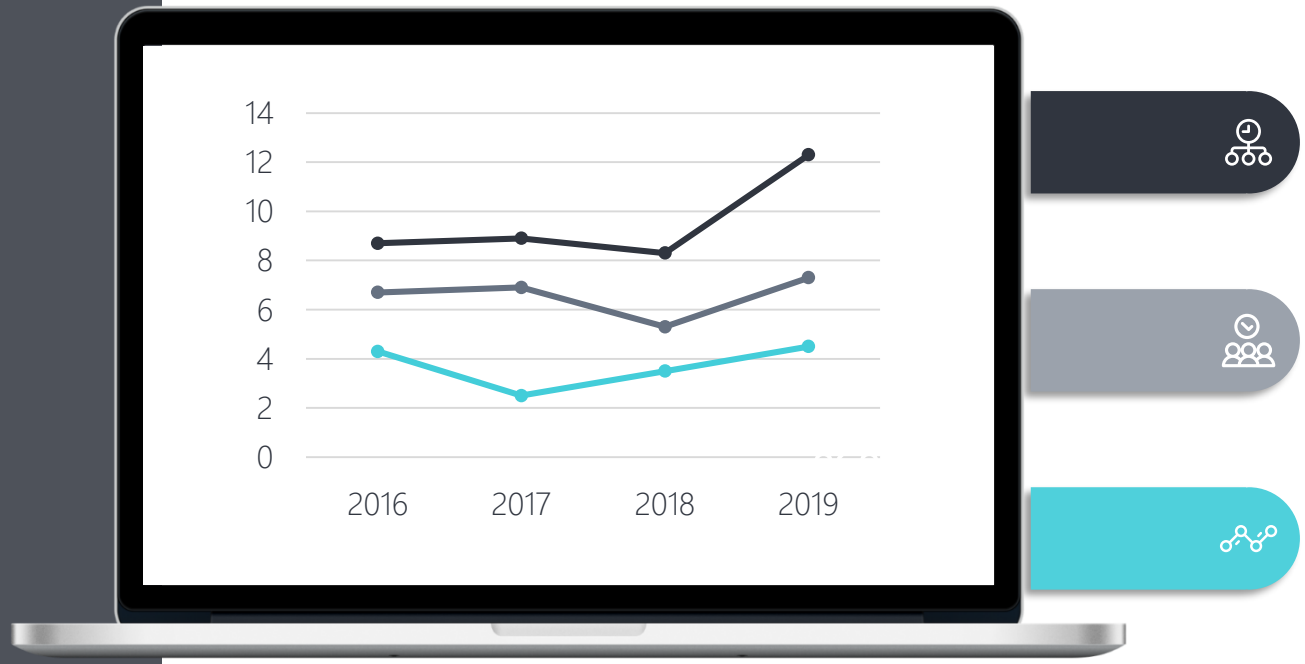
BACKGROUND

Our company is conducting a research about **Gentrification** in Mexico City.

Our data analysts asked us to assemble some initial datasets. They want to replicate other countries experiences about the demonstration that data from digital platforms have the potential to improve the understanding of gentrification.

We Extracted, Transformed and Loaded data sources and created requests to Google Places and Yelp. Then we stored the information in Mongo.

gen·tri·fi·ca·tion
noun
1.the process of renovating and improving a house or district so that it conforms to middle-class taste.
2.the process of making a person or activity more refined or polite.



Gentrification benefits:

- Better planning - urbanism
- Cities improvements
- Quality of life and services
- Social inclusion
- Safety and security

ZIP CODES – INEGI AND GEONAMES

EXTRACT
Read Data

- Started project ideation and researched available sources, databases and information.
- Chose our sources based on requirements: Zip Codes, Latitude, and Longitude for Mexico City only.
- Data was extracted from INEGI (zip codes) and coordinates from Geonames.

TRANSFORM
Clean & Structure

- Cleaned datasets.
- Removed errors, null and duplicated values.
- Selected relevant information for Mexico City : Zip codes and Coordinates.
- Merged two data frames (INEGI's and coordinates).

LOAD
Upload

- Response saved as JSON.
- The data was stored permanently in a Mongo cluster.

GOOGLE PLACES

EXTRACT
Read Data

- Used the merged data frames (INEGI's and coordinates) stored in Mongo.

TRANSFORM
Clean & Structure

- Search and target definition.
- Geo Coordinates and parameters delimitation: "Cafeteria" (keyword); "Restaurant" (type); 8,000 meters (radius).
- Using **Google Places API** we got business information.

LOAD
Upload

- The data was stored permanently in a **Mongo** cluster.

YELP

EXTRACT
Read Data

- Search in Yelp.com
- Filter: *Coffee and Tea businesses near Mexico City area.*
- Using **web scraping** got business name and location from each HTML page.

TRANSFORM
Clean & Structure

- Using **Python, Splinter,** and **Beautiful Soup** we identified the tag elements that contained business information (name, street and neighborhood).
- Using **Pandas** we grouped the information by neighborhood and did some aggregates for the data analysts.

LOAD
Upload

- The data was stored permanently in a **Mongo** cluster.

in addition to

COLLECTIONS IN MONGO

YELP

Documents with business information: name, street, and neighborhood

778

GOOGLE PLACES

Documents with business Information

+1,000*

ZIP CODES

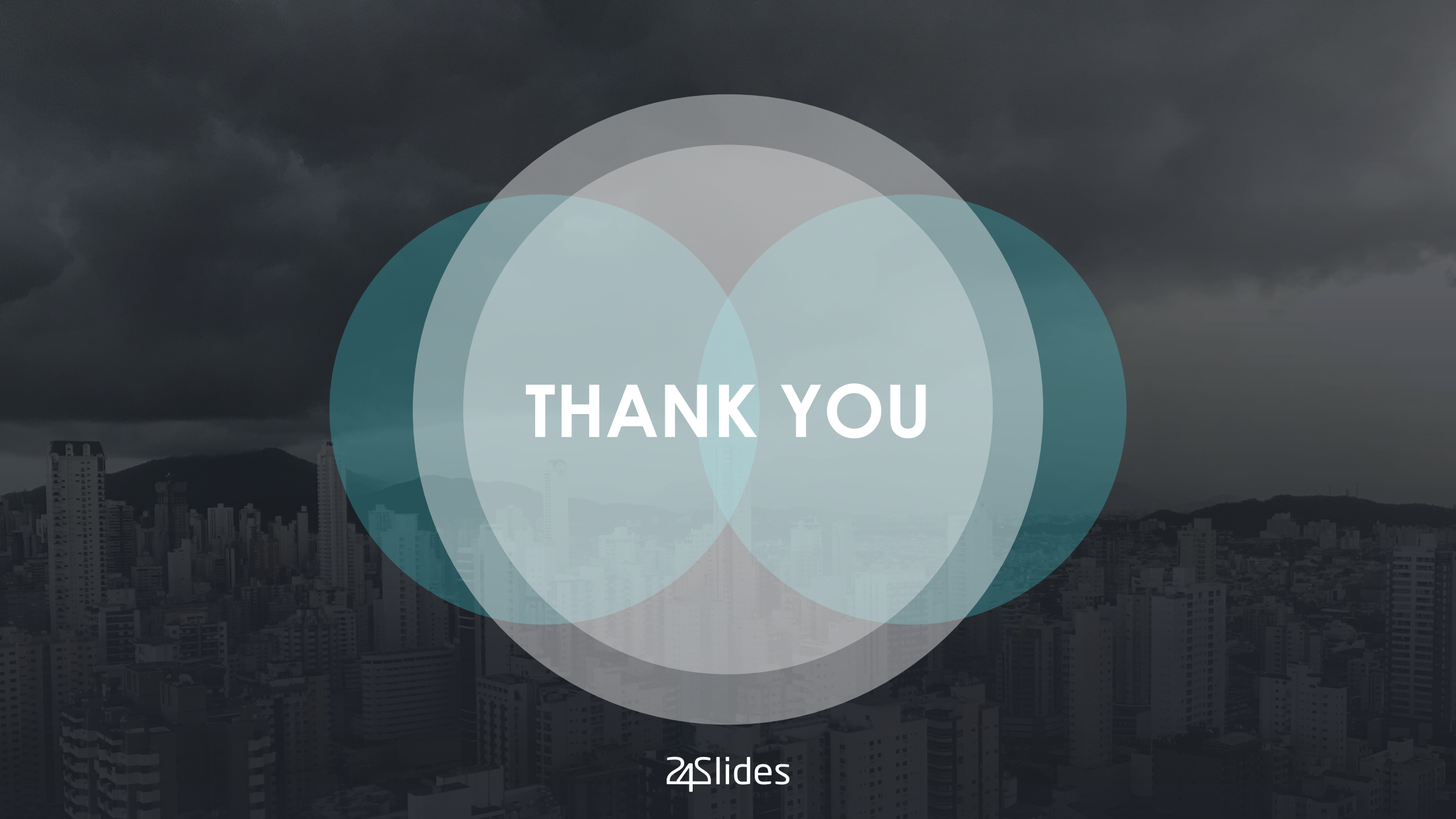
Mexico City Zip Codes by INEGI plus Latitude and Longitude by Geonames

+1,300

Why Mongo?

We chose a Mongo cluster to easily store the information obtained from Google and Yelp, in addition to its ability to handle large unstructured data.

* We only run a small test – 20 zip codes


THANK YOU
24Slides

OTHER COUNTRIES EXPERIENCES

Measuring Gentrification: Using Yelp Data to Quantify Neighborhood Change
Edward L. Glaeser, Hyunjin Kim, Michael Luca
NBER Working Paper No. 24952
Issued in August 2018
NBER Program(s):Economic Fluctuations and Growth, Productivity, Innovation, and Entrepreneurship

We demonstrate that data from digital platforms such as Yelp have the potential to improve our understanding of gentrification, both by providing data in close to real time (i.e. nowcasting and forecasting) and by providing additional context about how the local economy is changing. Combining Yelp and Census data, we find that gentrification, as measured by changes in the educational, age, and racial composition within a ZIP code, is strongly associated with increases in the numbers of grocery stores, cafes, restaurants, and bars, with little evidence of crowd-out of other categories of businesses. We also find that changes in the local business landscape is a leading indicator of housing price changes, and that the entry of Starbucks (and coffee shops more generally) into a neighborhood predicts gentrification. Each additional Starbucks that enters a zip code is associated with a 0.5% increase in housing prices.

OTHER COUNTRIES EXPERIENCES

Predicting gentrification through social networking data
https://www.cam.ac.uk/research/news/predicting-gentrification-through-social-networking-data
Data from location-based social networks may be able to predict when a neighbourhood will go through the process of gentrification, by identifying areas with high social diversity and high deprivation.

The Cambridge researchers, working with colleagues from the University of Birmingham, Queen Mary University of London, and University College London, used data from approximately 37,000 users and 42,000 venues in London to build a network of Foursquare places and the parallel Twitter social network of visitors, adding up to more than half a million check-ins over a ten-month period. From this data, they were able to quantify the 'social diversity' of various neighbourhoods and venues by distinguishing between places that bring together strangers versus those that tend to bring together friends, as well as places that attract diverse individuals as opposed to those which attract regulars.

When these social diversity metrics were correlated with wellbeing indicators for various London neighbourhoods, the researchers discovered that signs of gentrification, such as rising housing prices and lower crime rates, were the strongest in deprived areas with high social diversity. These areas had an influx of more affluent and diverse visitors, represented by social media users, and pointed to an overall improvement of their rank, according to the UK Index of Multiple Deprivation.