NEXT TECH

AN ANALYSIS ON US TECH HUB CITIES

PROJECT OVERVIEW & REPORT

Contributors: Deep Patel, Joshua Coronel, Keana Mabilog, Stephano Castro

01

SYNOPSIS

Tech Hubs are locations primed for entrepreneurship and innovation. We have witnessed rapid growth in technological innovations in places such as Silicon Valley. This growth has continued and is now spreading across cities in the United States. We want to find out how and why

WHERE WILL THE NEXT TECH HUB EMERGE?

To answer this question, the project was broken down into 2 main criteria.

1) The use of **machine learning** to predict cities whose trends align with that of well establish tech hubs in the US.

2) **Data Visualizations** of existing tech hubs to give a clear idea of what the information means through visual context

02

TOOLS

LANGUAGES
- Python
- JavaScript
- HTML/CSS

FRAMEWORKS
- Flask
- PySpark

LIBRARIES
- Pandas
- Scikit-Learn
- Leaflet.js
- Plotly.js
- D3.js
- Google Charts

DATABASES
- MongoDB
- AWS S3

OTHER
- Heroku

DATA PROCESSING

In deconstructing how we would analyze a tech hub, we considered the following factors which may contribute to its success

- **Real Estate Value** - representative of a strong local economy
- **Income per Capita** - denotes the abundance of high-wage jobs
- **Education Levels** - access to talent is important
- **Public Transportation** - contributor of population density
- **Median Male/Female Age** - younger vs older population
- **Crime Rates** - how this might affect certain areas
- **Technical Occupation Count** - technical positions in demand

To source the data, we developed a Python script with APIs to request data from the Census, Bureau of Labor Statistics, Government OpenData, and Zillow. All the unstructured data is stored in an AWS S3 Bucket. PySpark is our distributed processing tool to clean and integrate the raw data from the three different data sources. Once merged, all the data was stored in the cloud data MongoDB Atlas.

03 PREDICTING EMERGING TECH HUBS

Goal: *Train a machine learning model to predict if a zip code is an emerging tech hub. In our approach, we need to define the criteria for a zipcode to be considered a tech hub.*

UNSUPERVISED LEARNING MODEL: K-MEANS

After feature preprocessing the dataset, we separated the data into 5 clusters by training a K-Means Model. We reviewed the 5 different clusters to identify which cluster contained cities we have identified as existing tech hubs (e.g. San Francisco, Austin, Chicago, New York City). With the cluster identified, we created a binary column with every zip code labeled as Yes or No for being a Tech Hub - this column will be used as the output to train against for our supervised learning model.

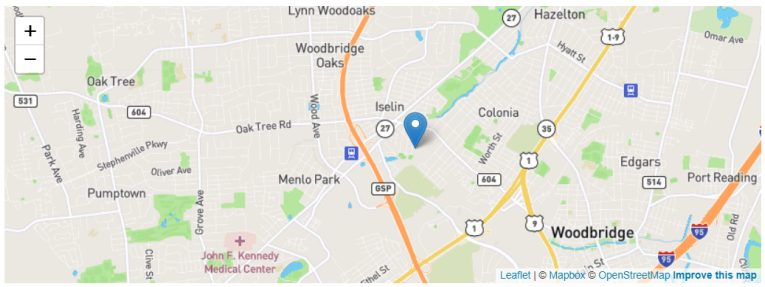
SUPERVISED LEARNING MODEL: LOGISTICAL REGRESSION

With a label for our output, we split 30% of our data into a test set and 70% into a training set. We fit and trained the model to produce an accuracy of 95%. The trained machine learning model was serialized into a .pickle file to be deployed in production. The deployed machine learning algorithm was deserialized in a Flask API to be called in the full-stack web application for the client to call.

RESULTS

The webpage allows the user to enter a zip code. Once entered, the webpage scrolls down and generates a table and leaflet map to represent the tech hub prediction with the different data points that resulted in the prediction

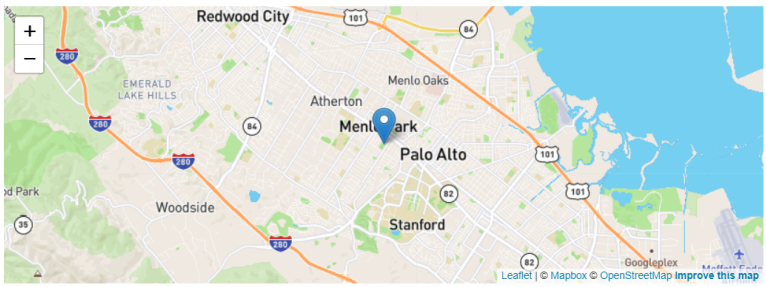
Status	City,State	Real Estate	Income per Capita	Bachelor's Degree	Public Transportation	Median Age Female	Median Age Male
No!	Colonia, NJ 08830	$309685.75	$40142.0	28.23%	17.88%	42.7	42.5



Entering a zip code that is not predicted as a tech hub will result in with a status text of "No" Here you can see Colonia, NJ (suburban town).

For a zip code that is predicted to be an emerging tech hub, the status text will result in a "Yes". The screenshot on the right is of Menlo Park, CA which is home of Facebook's HQ.

Status	City,State	Real Estate	Income per Capita	Bachelor's Degree	Public Transportation	Median Age Female	Median Age Male
Yes!	Menlo Park, CA 94025	$2654259.0	$80653.0	27.45%	5.85%	37.8	38.7



04 VISUALIZING THE DATA

Goal: *Create 3 interactive visualizations depicting the average home values with crime, wages for occupation, and technical occupational opportunities for each city*

LEAFLET MAP

The geo-mapped visualization is a multi-layer leaflet map to determine insights between real-estate values and crime activity in different tech hubs. We chose a Leaflet map over Plotly as this had better tools available and focused specifically on making maps. The map object within Leaflet was created using Mapbox.

BASE LAYER

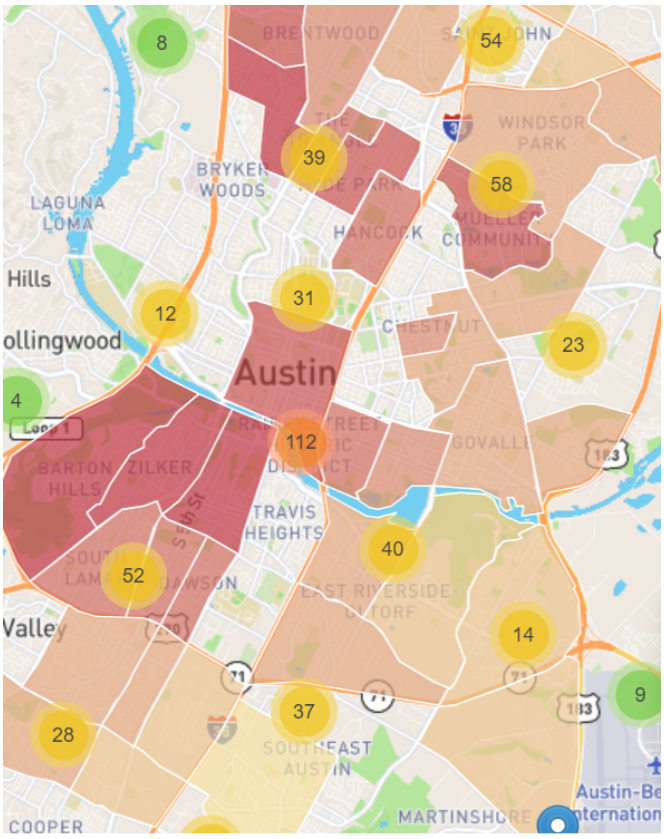
The base layer comes from a Mapbox API that enables a day and night toggle for the map. This can be further referenced on https://www.mapbox.com/maps/

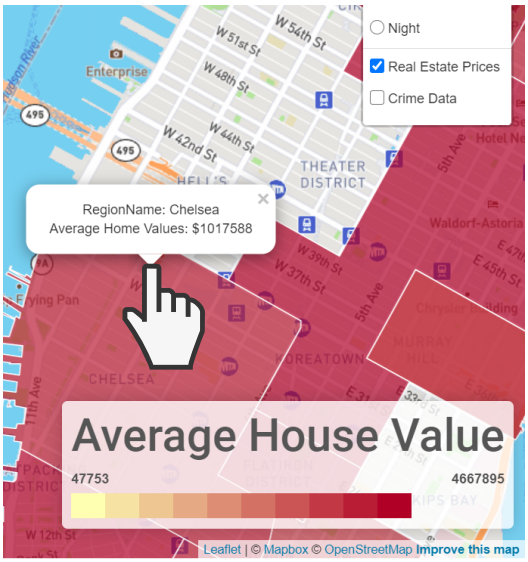
CHOROPLETH LAYER

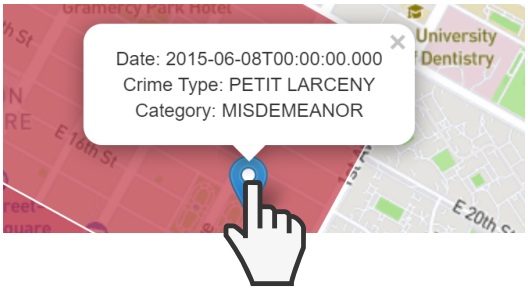
To set up the choropleth layer, we sourced polygon information (GeoJson) for each of the neighborhoods in our real estate data. This data creates the borders for each region that's shown on the map. ETL was used to merge this with our real estate data. We then used a D3.js function to parse the data for the choropleth objects. Each region includes a pop up to display the region's name and real estate value.

MARKER LAYER

For the marker layer, we created cluster groups that expand to individual markers at specific coordinates as the user zooms in. Following the same pattern as the choropleth layer, we created a function to read from data sources and append the corresponding markers to a ClusterGroup layer which was then added to an overlay layer in our map container. Each marker includes a pop up as well, displaying the crime's data and time, type, and category region.







05 VISUALIZING THE DATA - CONT.

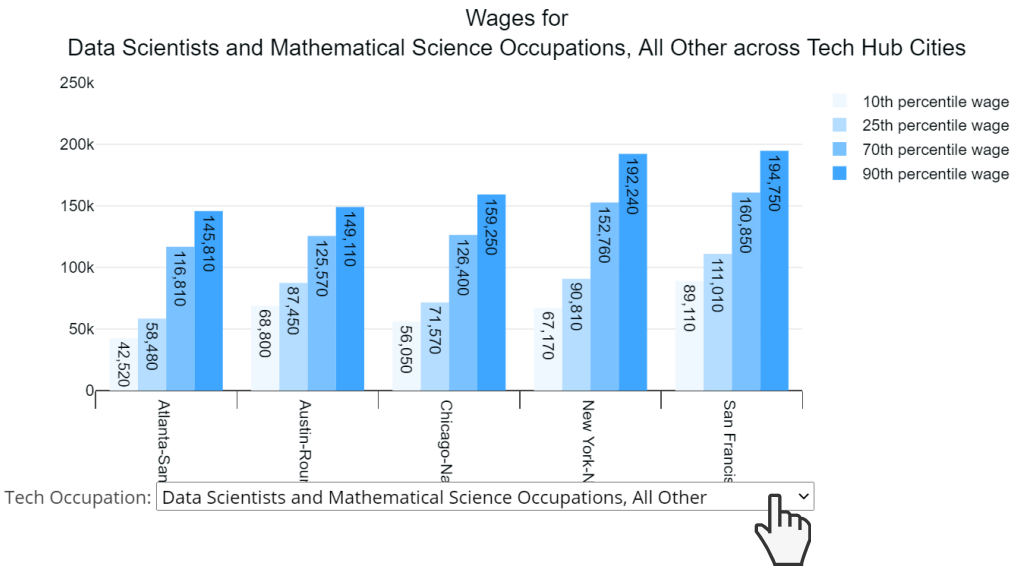
Goal: *Create 3 interactive visualizations depicting the average home values with crime, wages for occupation, and technical occupational opportunities for each city*

PLOTLY - GROUPED BAR CHART

The grouped vertical bar chart visualization separates tech salaries by percentiles to compare compensation in the different tech hubs. The data visualized is based on the tech occupation selected in the dropdown menu. D3.js was used to enter data into the Plotly bar chart object.

Interactive Features:

- Dropdown tech occupation selection
- Filter the wage percentiles groups
- Tooltip highlighted data

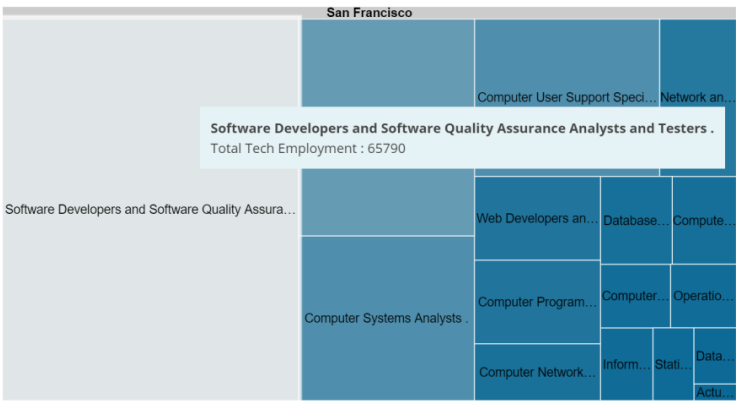


GOOGLE CHARTS - TREE MAP

The tech occupation treemap visualizes the proportional volumes of the tech occupations in many well-known cities. Once a city is selected, each occupation category is sized based on the employment count. The lighter the shade of blue, the higher the employment count.

Interactive Features:

- Tooltip with full occupation category and employment count
- Click into city/occupation



INSIGHTS AND TRENDS

WHAT WE HAVE FOUND OUT

Looking through the data and visualizations a few of the key elements we discovered about the selected tech hub cities are as follows

- Crime is consistent across the board relative to real estate prices. In areas of higher real estate prices top criminal activities are
  - Dangerous Drugs
  - Petit Larceny
  - Theft Related Offenses

- **Average housing values** across all hosting tech hub cities are approximately *$500,000*

- Of the 5 tech hubs cities analyzed, **NYC** holds the **highest employed** in tech at **38%**
  - San Francisco, CA (20%)
  - Chicago, IL (18%)
  - Atlanta, GA (16%)
  - Austin, TX (8%)

- The **most employed** tech occupations are *Software Developers and Software Quality Assurance Analyst*

- The **least employed** tech occupations are *Mathematicians and Actuaries*

07

REFERENCES

DATA SOURCES

- https://www.zillow.com/research/data/
- https://www.census.gov/data/developers/data-sets.html
- https://www.bls.gov/developers/
- https://www.data.gov/

SOURCE CODE

- https://github.com/joshcoronel/machine-learning-tech-hubs

DEMO LINK

- https://tech-hub-predictor.herokuapp.com/