Updated Capstone Proposal 7/8/2024 / Wesley Hall

I did some thought on my proposal and have updated it to incorporate educational data. Do you think this will be too ambitious? I thought I may try it and if it is too complicated, will look at something else.

Objective

Predict the future spatial distribution of technology startups in Los Angeles based on business registration data, educational attainment metrics, and temporal factors.
Relevance:Understanding how educational demographics and temporal trends influence business clustering can inform targeted urban development and economic policies.

Data Description

Dataset 1: Business Registrations https://data.lacity.org/Administration-Finance/Listing-of-Active-Businesses/6rrh-rzua/data Contains business registrations categorized by NAICS codes, addresses with zip codes, geographical coordinates (latitude, longitude), and dates of registrations.

Dataset 2: Educational Attainment
https://data.census.gov/table?q=educational%20attainment%20by%20zip%20code&g=010XX00US_050XX00US06037Provides educational attainment metrics (e.g., percentage of population with bachelor's degrees, high school diplomas) by zip code in Los Angeles.

Methodology

Data Integration and Preparation

- Merge Datasets: Merge business registration data with educational attainment data based on zip codes to create a unified dataset. Since my educational data is aggregated annually by zip code, and the business registration data is at the individual business level with specific start dates, I can aggregate the business registration data by year to match the aggregation level of educational data. I will count the number of business registrations (or calculate other metrics like average or sum of certain attributes) within each zip code and year combination.
- Feature Engineering: create new features such as percentage of population with bachelor's degrees and high school diplomas for each zip code area.

Exploratory Data Analysis (EDA)

Summary Statistics: the distribution and summary statistics of business registrations, educational metrics, and temporal features.

- Correlation Analysis: Explore correlations between educational attainment levels, temporal trends, and business density or clustering patterns.

Feature Selection

Select Relevant Features: Identify features that are most predictive of future business clustering patterns based on initial EDA findings and domain knowledge.

Model Development and Prediction

K-means Clustering with Temporal Factors

- Apply K-means Clustering: Cluster current business locations based on geographical coordinates, educational variables, and temporal factors (dates of registrations); Determine the optimal number of clusters (K) using methods like the elbow method or silhouette score.

Supervised Learning with Temporal Features

- Train Supervised Learning Models: Use historical data on business registrations, educational variables, and temporal features to train supervised learning models (e.g., Random Forest, Gradient Boosting); Define the target variable as the likelihood of new technology startup clusters forming in specific neighborhoods over the next 5 years.

Model Evaluation and Validation

- Cross-validation: Validate the predictive models using techniques like K-fold cross-validation to ensure robustness and generalizability.
- Performance Metrics: Measure model performance using metrics such as accuracy, precision, recall, and F1-score tailored to prediction goals.

  Implementation Plan
- Data Preprocessing: Clean and preprocess datasets, handle missing values, normalize numerical features, and encode categorical variables as needed.
- Model Implementation: Implement K-means clustering and supervised learning models using Python libraries (e.g., scikit-learn, pandas).
- Visualization: Visualize clustering results, temporal trends, and predictive outcomes using interactive maps, time series plots, and other visualization tools.

Conclusion

- Prediction Results: Present spatial distribution maps and time series forecasts of technology startup clusters in Los Angeles considering educational and temporal dynamics.
- Insights: Discuss insights into how educational investments and temporal trends can shape future business clustering patterns and inform urban development strategies - refer to paper