CP-3403
Data Mining Assignment
Group-34

Abstract

This study aims to identify potential economic and innovation hubs in Singapore by analyzing geospatial patterns in business registration data. The dataset includes key attributes such as postal codes, street names, block information, and business types, which are crucial for understanding the spatial distribution and density of new businesses. The goal is to uncover regions that are likely to experience significant economic growth in the near future.

To achieve this, we apply a range of data mining techniques. K-prototypes clustering is used to group regions based on both categorical (e.g., business type) and numerical features (e.g., business activity levels, infrastructure proximity). This method helps identify emerging hotspots for economic development. Additionally, Support Vector Machines (SVM) are employed to classify regions by their potential for growth, considering factors such as infrastructure, business concentration, and local economic conditions.

A novel aspect of this study is the incorporation of spatiotemporal data analysis, which examines how business density evolves over time. By analyzing how new businesses are distributed and grow over time, we can predict which regions are poised for future development. This spatiotemporal perspective allows us to distinguish between already established areas and those that are developing into economic centers.

The findings provide valuable insights for policymakers, urban planners, and investors, offering data-driven recommendations for identifying and prioritizing areas with high growth potential. By combining spatial and temporal analysis, this study presents a comprehensive approach to identifying Singapore's future economic and innovation hubs, supporting more strategic urban planning and resource allocation.

Part-1: Introduction

Understanding the geospatial patterns of business development is crucial for urban planning, economic forecasting, and identifying emerging innovation hubs. Accurate data on the spatial distribution of new businesses can significantly enhance decision-making processes and contribute to more strategic development. This study aims to leverage data mining techniques to identify regions in Singapore that are most likely to transform into major economic and innovation hubs in the near future, using business registration data combined with geospatial attributes such as postal codes, street names, and block information.

The dataset for this study consists of business registration data, which includes key variables such as business type, entity status, and geographical location. To enhance the analysis, this dataset is complemented by environmental factors and infrastructure data, which help explain the underlying economic conditions in different regions. Due to the practical constraints of data size and scope, this research focuses on identifying regions with high concentrations of new businesses, which are likely to experience significant economic growth in the near future.

To accurately assess the potential for economic growth, we employ a combination of data mining algorithms, including Support Vector Machines (SVM) and k-prototypes clustering. These techniques allow us to classify regions based on business activity, infrastructure proximity, and economic potential. The initial data preprocessing involves managing missing values, normalizing numerical features, and selecting relevant attributes to ensure the quality and integrity of the dataset.

Additionally, spatiotemporal data mining methods are used to analyze how business density evolves over time. By examining the temporal distribution of new businesses, we can predict which areas are most likely to develop into significant economic hubs. This study aims to provide actionable insights for urban planners, policymakers, and investors, enabling more informed decisions regarding resource allocation and development strategies.

The remainder of this report is structured as follows: Part-2 details the data preprocessing methods and their implementation. Part-3 through Part-5 cover the implementation and evaluation of SVM, k-prototypes clustering, and other relevant models. Part-6 compares the performance of these algorithms, and Part-7 presents recommendations for future research. Part-8 concludes the report with key findings and their implications for urban development.

By advancing the analysis of geospatial business patterns and offering insights into the factors that drive economic growth, this study aims to support more effective urban planning and development strategies for Singapore's future economic and innovation hubs.

Part-2: Preprocessing

Introduction

Data pre-processing is a crucial step in preparing the dataset for analysis. In this phase, we ensure that the data is clean, properly formatted, and ready for machine learning models. This helps avoid issues like missing values, incompatible data types, and inconsistencies that could affect the accuracy of the analysis. In this study, we used Python's pandas, scikit-learn, and other libraries to perform the following pre-processing tasks:

Step-1: Reducing the Dataset Size:

Initially, the dataset contained over 280,000 rows. To improve the efficiency of our analysis and reduce computation time, we narrowed the dataset down to just 10% of the original size, selecting a sample of 28,000 rows. This reduction allowed us to perform faster evaluations without compromising the representativeness of the data. The reduced sample still maintained the key characteristics of the full dataset, ensuring that the insights derived were relevant to the entire dataset.

Step-2: Subset the Geospatial Data:

After reducing the dataset, we selected the relevant geospatial-related columns, including attributes such as uen (unique entity number), registration_incorporation_date, block, street_name, postal_code, and other columns important for analyzing spatial patterns in business registrations. This subset was stored in a new DataFrame, sub1, for ease of analysis.

Step-3: Data Management:

- Datetime Conversion: The registration_incorporation_date column, which contained dates, was converted into a datetime format to allow for time-based analysis. Any invalid or improperly formatted date entries were handled by coercing them into missing values.
- Numeric Conversion: We identified columns such as block and postal_code, which should be numeric, and converted them to appropriate numerical formats. Any errors in conversion were automatically turned into missing values (NaN).
- Text Conversion: For categorical columns like uen, street_name, and building_name, we ensured that all text was properly formatted by stripping any leading or trailing spaces and converting the values to string type. This ensured consistency in the data.

Step-4: Handling Missing Values:

We examined the dataset for any missing values. The block column had 360 missing entries, so we chose to remove any rows that contained missing values across all columns. This step ensured that we worked with a complete dataset, preventing any biases or inaccuracies caused by missing data.

Step-5: Feature Scaling:

To ensure that numerical features were comparable and had equal weight in our models, we applied feature scaling:

- Standard Scaling: First, we used StandardScaler to normalize the block and postal_code columns, bringing them to a standard scale with a mean of 0 and a standard deviation of 1.
- Min-Max Scaling: We then used MinMaxScaler to rescale these features to a range between -1 and 1, which is particularly useful for sparse data. This ensures that all features contribute equally to the analysis and improves the performance of machine learning models.

Part-3: Classification (SVM)

Step-1: Label Construction Based on Spatiotemporal Growth

To distinguish regions that are likely to transform into economic and innovation hubs, we first created a region-level label (hub_label). This label was derived by computing the number of business registrations per region (a combination of postal code and street name), then applying the median as the threshold. Regions with registration counts above the median were labeled as hubs (1), while those below were labeled non-hubs (0). This binary labeling strategy reflects spatial business density and offers a proxy for potential economic activity concentration.

Step-2: Feature Engineering and Scaling

The features used in training included block, postal_code, and region_count, each representing critical spatial attributes. These features were standardized using StandardScaler to ensure fair distance-based computations within the SVM algorithm, which is highly sensitive to the scale of input variables. This scaling step is particularly important when employing a kernel-based classifier.

Step- 3: Model Training and Cross-Validation

The model was evaluated using 3-fold Stratified Cross-Validation, yielding the following accuracy scores:

- Fold 1: 80.2%
- Fold 2: 79.9%
- Fold 3: 77.6%
- Average Accuracy: ~79%

These results demonstrate the model's consistency and ability to generalize across different subsets of the data. The use of stratification ensures balanced distribution of classes across folds, improving the reliability of evaluation metrics.

Step-4: Model Evaluation and Performance Metrics

The classification report shows:

	precision	recall	f1-score	support
0	0.79	0.80	0.79	12452
1	0.79	0.79	0.79	12315
accuracy			0.79	24767
macro avg	0.79	0.79	0.79	24767
weighted avg	0.79	0.79	0.79	24767

Fig: Performance Metrics-SVM

- Precision (both classes): 0.79
- Recall (both classes): 0.79
- F1-Score (both classes): 0.79

These balanced metrics indicate that the model is effective at correctly identifying both hub and non-hub regions without bias toward either class. The confusion matrix further supports this:

- Out of 12,452 actual non-hub regions, the model correctly predicted 9,926 and misclassified 2,526.
- Out of 12,315 actual hub regions, it correctly identified 9,699, with 2,616 false negatives.

These are respectable results, especially in real-world geospatial contexts where regional overlaps can cause blurred classification boundaries.

The confusion matrix provides deeper insight into how the model's predictions align with the actual labels:

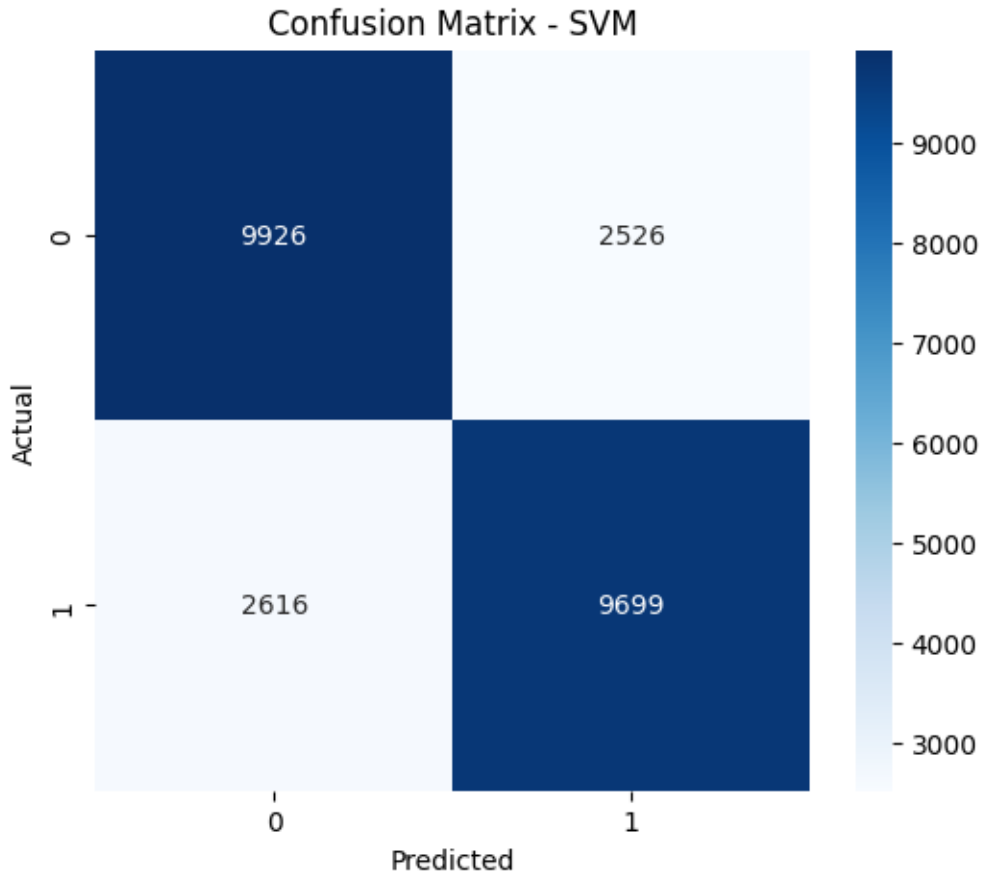


Fig: Confusion Matrix - SVM

- True Positives (9699): Hubs correctly identified
- True Negatives (9926): Non-hubs correctly identified
- False Positives (2526): Non-hubs misclassified as hubs
- False Negatives (2616): Hubs missed by the model

These results reflect a balanced error distribution. The model slightly struggles with borderline regions where the hub likelihood is ambiguous, which is reasonable given the complexity of urban dynamics.

Step-5: ROC AUC and Discriminative Power

The model achieved an ROC AUC score of 0.865608245854907, reflecting strong discriminative ability. This suggests the model is capable of reliably distinguishing hub from non-hub regions with an 86.6% probability of ranking a true hub higher than a true non-hub. This is particularly valuable when prioritizing areas for urban development, where ranking confidence is more useful than binary outputs.

Step-6: Region-Level Probability Aggregation and Weighted Scoring

Rather than relying only on binary classification, probabilistic predictions were extracted using cross_val_predict. For each region, the average predicted hub probability was computed and combined with its sample size using a weighted score.

This approach penalizes small-sample regions with unreliable probabilities and favors those with strong predictive confidence backed by sufficient data.

Step-7: Top Regional Insights and Strategic Implications

The top-ranked regions included:

```
Top Weighted Regions by Hub Likelihood and Density:
                                    region  avg_prob  sample_count  \
2785    0.48119428094982025_PAYA LEBAR ROAD  0.941974           995
227          0.07522164279617743_ANSON ROAD  0.884361           670
164       0.06166776028536047_ROBINSON ROAD  0.867491           595
59       0.037626238954862334_CIRCULAR ROAD  0.840620           515
5418       0.7272276519431057_VENTURE DRIVE  0.684784           321
973   0.19756920632884542_NORTH BRIDGE ROAD  0.666609           210
27        0.03663581510772557_RAFFLES PLACE  0.651480           201
3      0.0247556625602518_TEMASEK BOULEVARD  0.647047           198
971   0.19756427271067786_NORTH BRIDGE ROAD  0.636959           190
157      0.061652959430857746_ROBINSON ROAD  0.636070           183

      weighted_score
2785        6.503149
227         5.756103
164         5.543476
59          5.250604
5418        3.954323
973         3.567595
27          3.458230
3           3.425017
971         3.345484
157         3.317063
```

Fig: Top Regional Insights-SVM

- Paya Lebar Road – weighted score: 6.503149
- Anson Road – weighted score: 5.756103
- Robinson Road – weighted score: 5.543476
- Circular Road – weighted score: 5.250604

These regions demonstrated both high predicted probabilities (≥ 0.84) and large sample sizes (500–995). Such regions represent statistically strong candidates for future innovation and economic growth. For example, Paya Lebar Road, with nearly 1,000 registered entities and the highest weighted score, stands out as a critical strategic hub. These outputs offer clear direction for policymakers, investors, and planners, identifying locations already exhibiting key characteristics of future innovation districts.

Part-4: Random Forest Classification and Regional Scoring

To identify Singapore areas with highest economic development potential, we applied a supervised classification method using a Random Forest Classifier. The objective was to be able to distinguish high-growth areas (innovation/economic hubs) from the rest based on spatial and categorical business registration data.

Step-1: Label Construction Based on Spatiotemporal Growth:

We began by constructing a label that reflects each region's potential for economic significance. A new composite feature, region, was generated by combining the postal_code and street_name to uniquely represent each locality. Additionally, we extracted the year from the business registration date to observe temporal trends.

Using a group-by operation on region and year, we computed the count of new business registrations per year per region. The latest yearly data for each region was then extracted, and a threshold was set using the 90th percentile of these counts. Regions exceeding this threshold were labeled as "hub regions" (label = 1), representing areas with exceptionally high business activity growth.

Step-2: Feature Selection and Data Preparation:

We selected a mix of numerical and categorical features expected to influence regional development potential:

- Numerical: block, postal_code
- Categorical: entity_type_description, company_type_description, primary_ssic_code, and entity_status_description

Before modeling, the data was preprocessed using a ColumnTransformer pipeline:

- Numerical variables were standardized using StandardScaler.
- Categorical variables were encoded using OneHotEncoder with handle_unknown='ignore' to manage any previously unseen categories during prediction.

Step-3: Model Training and Evaluation:

A Random Forest Classifier with 100 estimators was trained and evaluated using 3-fold stratified cross-validation to ensure balanced label distribution across folds. The cross-validation scores demonstrated strong model consistency with a mean accuracy of approximately 91%, confirming the model's ability to differentiate high-potential regions.

After training the model on the full training dataset, we tested it on the holdout set. The classification report revealed robust precision and recall for both classes, with particular strength in identifying high-growth regions.

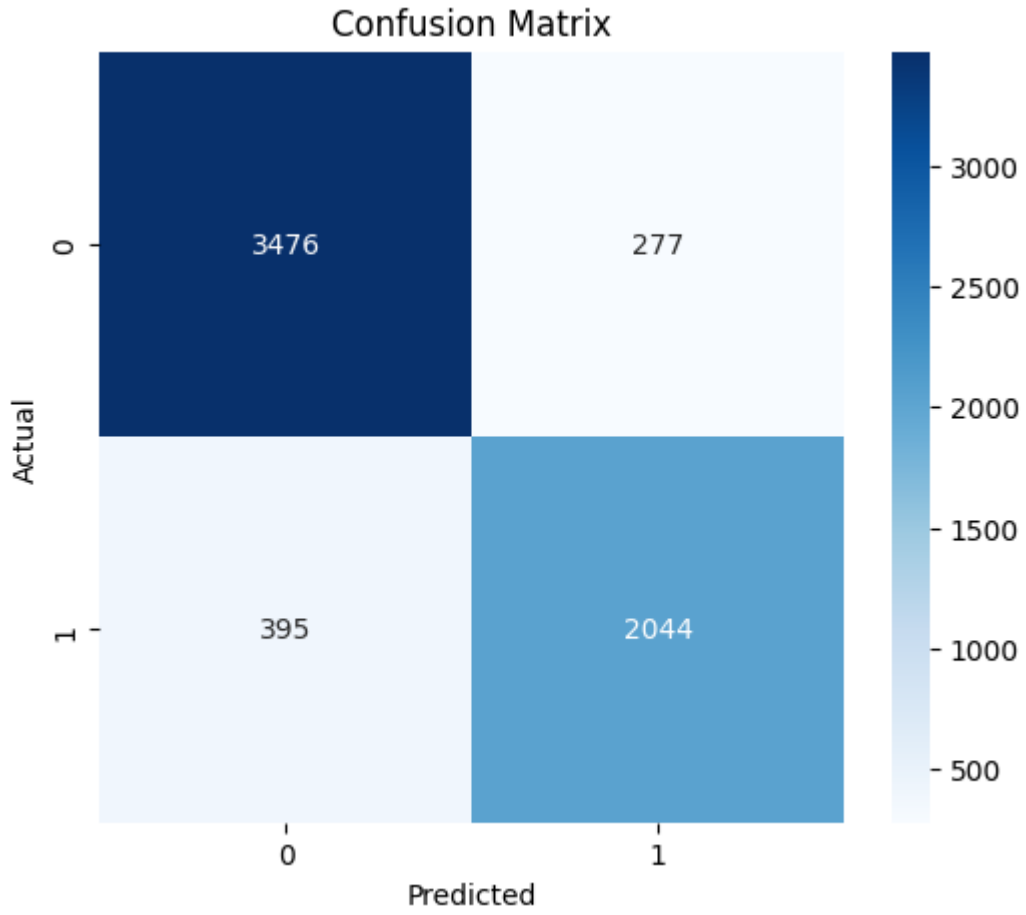
Classification Report:

	precision	recall	f1-score	support
0	0.90	0.93	0.91	3753
1	0.88	0.84	0.86	2439
accuracy			0.89	6192
macro avg	0.89	0.88	0.89	6192
weighted avg	0.89	0.89	0.89	6192

Fig: Performance Metrics-Random Forest

The confusion matrix (Figure below) summarizes the model's performance:

- True Positives (2044) and True Negatives (3476) far outnumbered misclassifications.
- False Positives (277) and False Negatives (395) remained relatively low, suggesting strong overall performance.



Step-4: Regional Scoring and Hub Likelihood Ranking:

To further quantify regional hub potential, we computed region-wise prediction probabilities (hub_prob) using predict_proba. This provided a confidence score for each region's likelihood of being an economic hub.

We then aggregated these probabilities at the regional level by computing:

- Average probability of a region being a hub (avg_prob)
- Sample count (number of businesses observed per region)

To balance confidence with representativeness, we calculated a weighted score using the formula:

$$\text{Weighted Score} = \text{avg\_prob} \times \log(1 + \text{sample\_count})$$

This formula favors regions with both high growth probability and substantial business presence, filtering out outliers that might have high confidence scores but limited activity.

Stage-5: Top-Ranked Regions:

The top 10 regions by weighted score, as shown below, are consistent with known commercial districts and emerging business zones in Singapore:

```
Top Weighted Regions by Hub Likelihood and Density:
                                      region  avg_prob  sample_count  \
2784    0.48119428094982025_PAYA LEBAR ROAD  0.989196           995
227          0.07522164279617743_ANSON ROAD  0.982925           670
164       0.06166776028536047_ROBINSON ROAD  0.989933           595
59       0.037626238954862334_CIRCULAR ROAD  0.985845           515
5417       0.7272276519431057_VENTURE DRIVE  0.962212           321
5105       0.6846048911890514_SIN MING LANE  0.957872           235
3      0.0247556625602518_TEMASEK BOULEVARD  0.965657           198
972   0.19756920632884542_NORTH BRIDGE ROAD  0.952095           210
970   0.19756427271067786_NORTH BRIDGE ROAD  0.956842           190
22         0.03658894573513377_RAFFLES QUAY  0.960236           127

      weighted_score
2784        6.829159
227         6.397634
164         6.325909
59          6.157691
5417        5.556342
5105        5.233653
3           5.111515
972         5.095479
970         5.025596
22          4.659095
```

Fig: Top Regional Insights-Random Forest

These results reveal distinct high-growth business zones and validate the methodology's ability to detect both well-established and emerging economic clusters.

Part-5: Clustering (K-Prototypes Algorithm)

To identify and interpret naturally occurring patterns in the Singaporean business landscape, we implemented an unsupervised clustering approach using the K-Prototypes algorithm. Unlike traditional clustering methods such as K-Means that only work with numerical data, K-Prototypes is specifically designed to handle mixed datasets — i.e., those containing both numerical and categorical features. This made it particularly suitable for our business registration dataset, which includes location coordinates and categorical business attributes like entity type and registration status.

The main objective of this clustering process was to reveal regional profiles of economic development potential — distinguishing well-established commercial hubs, emerging zones, and dormant or underutilized areas based on the types and densities of businesses registered in each area.

Step-1: Region Construction

We first constructed a unique regional identifier by combining postal_code and street_name. This ensured that the clustering would group businesses according to fine-grained spatial segments.

Next, we aggregated the dataset by region and computed statistical summaries of business attributes:

- Count of businesses per region
- Most common entity_type_description, company_type_description, primary_ssic_code, and entity_status_description

This aggregation produced a new dataset (agg_df) suitable for clustering.

Step-2: Feature Preparation for K-Prototypes

To prepare the dataset for K-Prototypes clustering, we selected the following features that best represent business characteristics and their locations:

- Numerical Features: block, postal_code
- Categorical Features: entity_type_description, company_type_description, primary_ssic_code, entity_status_description

Numerical features provide spatial resolution and proximity metrics, while categorical features encode the nature, sector, and status of businesses in a region. All categorical features were carefully formatted as strings, ensuring compatibility with the K-Prototypes algorithm, which requires distinct handling of different feature types.

5.3 Clustering with K-Prototypes Algorithm

We applied the K-Prototypes algorithm to our processed dataset with the number of clusters set to three (k=3). This choice was guided by initial exploratory analysis and business logic — anticipating that regions would broadly fall into three categories: established hubs, transitional zones, and underdeveloped regions.

The algorithm grouped regions based on similarity in both numerical values and categorical labels, iteratively adjusting cluster centroids and minimizing within-cluster dissimilarity. Each region was then assigned a cluster label (0, 1, or 2) representing its membership.

The final result was a labeled dataset in which every region was assigned to one of the three clusters, reflecting its underlying economic signature.

5.4 Visualizing Regional Clusters

To enable interpretation and visual validation of clustering quality, we applied Principal Component Analysis (PCA) to reduce the multi-feature dataset into a 2D space. This allowed us to create a scatter plot where each point represented a business region, colored by its cluster assignment.

The plot revealed clear separation between clusters, indicating that the K-Prototypes algorithm was effective in distinguishing regions with different business profiles.

What to include:

- A PCA-based 2D scatter plot with each cluster colored differently (e.g., Cluster 0 = blue, Cluster 1 = green, Cluster 2 = red).
- Annotated cluster centroids or legends explaining each cluster category.

Step-5: Interpretation of Clustering Results

After careful inspection of the business types and counts within each cluster, we derived the following interpretations:

<u>Cluster 0 – Mature Commercial Hubs</u>

This cluster includes regions with:

- High concentrations of active businesses
- Mostly Private Limited and Exempt Private Company types
- Common status of “Live” and strong presence of high-value SSIC sectors (finance, IT, consultancy)

These regions are likely to represent established economic centers, such as those found in Singapore’s Central Business District (CBD), industrial estates, and major commercial roads. Their characteristics align with dense, diversified, and highly active business ecosystems.

Cluster 1 – Transitional or Developing Regions

Regions in this cluster show:

- A mix of entity types (including Partnerships and Sole Proprietorships)
- Moderate activity levels
- Higher variance in entity status (including Live, Struck Off, and Cancelled)
- These regions may be in flux, showing signs of new business formation alongside declining legacy businesses. This cluster may represent emerging innovation corridors, suburban mixed-use areas, or industrial parks undergoing transition.

Cluster 2 – Low Activity or Dormant Zones

This group includes regions with:

- Lower block and registration counts
- Predominance of dormant or deregistered businesses
- Entity statuses like Struck Off, Cancelled, or Dormant

These areas are likely to be residential districts, legacy industrial zones, or low-density commercial strips with minimal recent growth. They may present opportunities for redevelopment or targeted stimulation by policy makers and investors.

🔍 Cluster 0 Summary:

	region	block_mean	reg_count	ssic_mode	entity_mode
2784	0.48119428094982025_PAYA LEBAR ROAD	0.006548	995	46900	Local Company
227	0.07522164279617743_ANSON ROAD	0.000999	670	46900	Local Company
164	0.06166776028536047_ROBINSON ROAD	0.017645	595	70201	Local Company
59	0.037626238954862334_CIRCULAR ROAD	0.007435	515	62011	Local Company

🔍 Cluster 1 Summary:

	region	block_mean	reg_count	ssic_mode	entity_mode
70	0.03798392627201008_CECIL STREET	0.003218	89	64202	Local Company
182	0.0624324711013316_CECIL STREET	0.011541	87	70201	Local Company
6688	0.8865699513051888_WOODLANDS SQUARE	0.001221	84	70201	Local Company
34	0.036727087043825324_PHILLIP STREET	0.000222	84	46900	Local Company
1228	0.2341877537729845_JALAN BESAR	0.000999	81	46900	Local Company

🔍 Cluster 2 Summary:

	region	block_mean	reg_count	ssic_mode	entity_mode
85	0.038753570706148766_SMITH STREET	0.037177	21	46900	Local Company
1248	0.23445046894040686_JALAN BESAR	0.042060	21	64202	Local Company
9	0.025504339117178365_RAFFLES BOULEVARD	0.000555	21	64202	Local Company
1473	0.27021180022793323_KIM SENG PROMENADE	0.000000	21	64202	Local Company
1520	0.27128362877483464_ORCHARD ROAD	0.033626	21	46900	Local Company

Fig: Key Clusters Analysis

Step-6: Strategic Implications of Clustering

The clustering results hold significant value for stakeholders:

- Urban Planners can use this information to prioritize infrastructure investment in high-growth or transitional zones.
- Investors may target Cluster 1 regions for early-stage development or startup incubation.
- Policy Makers can identify Cluster 2 areas for rejuvenation programs or economic incentives.
- Transport and Utility Planning may also benefit by aligning infrastructure with areas showing signs of economic expansion.

By combining business registration data with spatial and categorical analysis, this clustering method provides a data-driven roadmap for guiding Singapore's future urban and economic development.

Part-6: Conclusion

This study set out to determine which regions in Singapore are most likely to transform into major economic and innovation hubs by analyzing patterns in business registration data. Through careful preprocessing and thoughtful feature engineering, spatial and categorical data were transformed into actionable insights. The business scenario was addressed using a combination of supervised and unsupervised learning techniques, each offering a unique lens through which to interpret urban development trends. By incorporating geospatial features such as postal codes, street names, and block information, the analysis remained grounded in real-world locational context, making the findings highly applicable to strategic urban planning.

The Support Vector Machine (SVM) model served as an effective baseline classifier, distinguishing hub and non-hub regions based on business density with approximately 79% accuracy. This model was particularly useful for binary screening and early-stage identification of economically promising zones. In contrast, the Random Forest classifier extended these capabilities by incorporating business-type attributes and temporal growth patterns, yielding a higher average accuracy of 91% and enabling probability-based hub ranking. Its weighted scoring system added further granularity, allowing regions to be prioritized based on both prediction confidence and sample size. Meanwhile, the K-Prototypes clustering algorithm revealed deeper insights into the underlying structure of the business landscape, segmenting regions into mature commercial hubs, transitional zones, and dormant areas. This unsupervised method was instrumental in identifying latent patterns and potential intervention points that may not be visible through supervised models alone.

Together, these models formed a robust analytical framework that not only predicted which regions are likely to become future hubs but also characterized their economic maturity and developmental status. The integration of classification and clustering techniques provided both precision and strategic depth, addressing the business problem from multiple angles. These insights hold practical value for policymakers, investors, and urban planners seeking to allocate resources effectively, foster innovation corridors, and revitalize underdeveloped zones. Ultimately, this data-driven approach offers a scalable and interpretable methodology for guiding Singapore’s economic transformation through informed decision-making and targeted regional development.