Cornerstone International Community College of Canada

Big Data Project Report

Business
Location & Optimization

Class teacher: Ramya Kappagantu School subject: DS 203 - Big Data Essentials

Project by Otávio Londero; Amir Oliveira

Abstract: This project analyzes historical business license data from the City of Vancouver, covering 1997 to 2024. The combined datasets include over 1.7 million business license records across hundreds of business categories, neighbourhoods, and statuses. By examining trends in license issuance, regional concentrations, and business types, we aim to uncover patterns that can inform decisions on optimal business locations and sector growth over time. Our analysis will also assess the impact of major policy changes, such as the 2024 consolidation of license categories, and identify opportunities for business development based on historical patterns. The final insights will support entrepreneurs, analysts, and policymakers in making data-driven decisions about business planning in Vancouver.

TABLE OF CONTENTS

1. Executive Summary

The primary objective of this project is to identify patterns of business saturation and growth within the City of Vancouver, helping inform strategic decisions around where and what types of businesses are most viable. To achieve this, was analyzed the open data published by the City of Vancouver, covering over 25 years of business license records (from 1997 to 2024).

This project was built with the idea of creating a fictional scenario, but with a real purpose and a real database. The Group 6 Analysts extracted information to clarify the whole scenario and help the audience (both technical specialists and business investors) to provide comprehensive, data-driven insights to support strategic investment decisions.

Our technical approach leverages PySpark for efficient Big Data processing, SQL queries for structured data exploration across tables, and Jupyter Notebooks to document and organize the analysis workflow.

The **key takeaways** we aim to uncover include:

- Which local areas in Greater Vancouver demonstrate high potential for new business development, based on long-term growth trends and low market saturation?
- Which business types have shown consistent historical success and indicate strong future growth potential, guiding strategic investment and policy-making?

2. Methodology

2.1. Some Questions We Intend to Answer

To investigate business development trends across cities and local areas, we adopt a methodology combining distributed processing, transformation, and analysis using **Apache Spark SQL**.

The questions below guided the analysis workflow for the project:

2.1.1. Which neighbourhoods or local areas offer the greatest potential for new business development?

Necessary Data: localareas, businesstype, status, numberofemployees, issueddate, expireddate, businesssubtype, city.

2.1.2. What business type has the fastest average growth in the number of licences issued per year and city?

Necessary Data: businesstype, licencersn, issueddate, status, city

2.1.3. What are the most stable business types (least likely to close within 5 years)?

Necessary Data: businesstype, issueddate, expireddate, status.

2.1.4. *Which business types show strong historical performance and future growth potential?*

Necessary Data: businesstype, issueddate, status, numberofemployees.

2.1.5. *Which city is oversaturated with the same business types?*

Necessary Data: city, businesstype, businesssubtype, issueddate, status, licencersn.

2.1.6. *How does business longevity differ between cities or business categories?*

Necessary Data: issueddate, expireddate, businesstype, businesssubtype, city, status.

2.1.7. *Are there seasonal patterns in new business openings or closures?*

Necessary Data: issueddate, expireddate, status, businesstype, city, localarea.

2.1.8. *What proportion of licences are renewed vs. replaced with new ones each year?*

Necessary Data: licencersn, issueddate, businesstype, status, city.

2.1.9. *How often are business licences cancelled or made inactive, and what are the patterns by category?*

Necessary Data: businesstype, status, issueddate, expireddate, city, localarea.

2.1.10. *Are certain business categories more prone to being "Pending" for long durations?*

Necessary Data: businesstype, status, issueddate, expireddate, city.

2.2. Methodology Detailing

After defining the analytical questions, we ingested and preprocessed the datasets using PySpark. We explicitly structured schemas during the data ingestion to ensure type consistency. Data cleaning included standardizing city names for geographic focus on Greater Vancouver, removing null or invalid entries.

For the analytical process, we used Spark SQL to perform filtering, grouping, aggregations, conditionals, subqueries, window functions, joins and derived column generation. Queries were often multi-staged, combining multiple transformations to answer complex business questions.

2.3. Analytical Methods

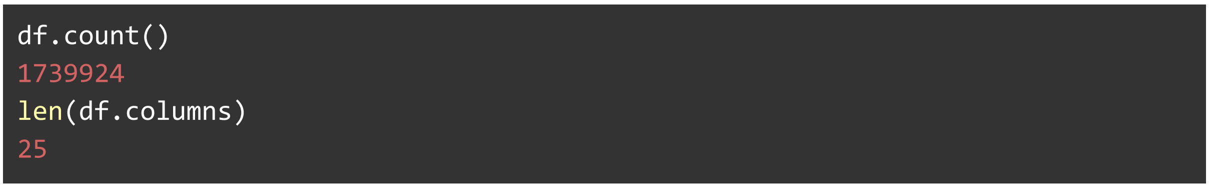
No.	Method	Formula/Concept	Purpose
1	Year-over-Year Growth	YOY % formula	Growth over time
2	Average Growth per Year	Total licences / active years	Business type expansion speed

3	Stability Rate	Survival over 5 years / total businesses	Business resilience
4	Saturation Index	Business type licences / total city licences	Market saturation identification
5	Seasonal Trend	Monthly group counts	Seasonality of openings/closures
6	Renewal vs. New Licence Rate	Renewal detection by LicenceSRN duplicates	New vs established businesses
7	Cancellation / Inactive Rate	Percentage of cancelled/inactive licences	Business risk patterns
8	Pending Duration	Days pending status	Regulatory bottleneck analysis

3. Dataset Overview

The project utilized two primary data sources — business_licenses-1997-to-2012 (209 MB) and business_licenses-2013-to-2024 (196 MB) — which together formed the original dataset for analysis.

Number of rows and columns in total:

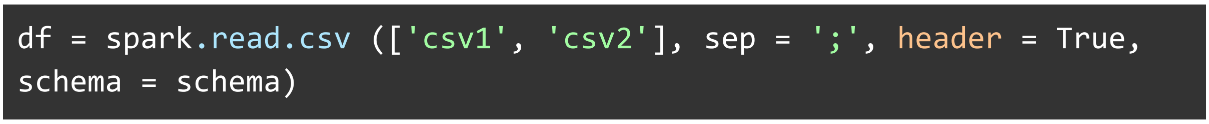


```
df.count()
1739924
len(df.columns)
25
```

4. ETL Process (Extract, transform, load)

4.1. Data Cleaning

4.1.1. Merging and standardizing both datasets



```
df = spark.read.csv (['csv1', 'csv2'], sep = ';', header = True,
schema = schema)
```

4.1.2 Dataframe cleaning, checking SCHEMA and POS count()

The ETL process required the group members to research how to drop unnecessary columns, which would help the Integrated Development Environments (IDEs) run queries faster and avoid errors. Additionally, to facilitate code sharing, they transformed all string data types to lowercase, ensuring uniformity. Afterwards, during the analysis and code writing phase, the team created a list of cities to filter only those from the Greater Vancouver area and to eliminate any anomalies.

Result:

```
df.count()
1675781

len(df.columns)
13
```

4.2. Data Normalization

Another ETL action was to normalize the dataframe with the datatypes in the schema.

Example:

```
schema = StructType([StructField("issueddate", TimestampType(),
True)])
```

Some columns, by default, after the Spark read CSV, came out wrong, then

5. Exploratory Data Analysis (EDA)

To conduct an effective exploratory analysis for the group, it was necessary to research general concepts of market dynamics and economics, as well as to gain a better understanding of business behaviours in Greater Vancouver. This research facilitated the development of the strategy and enabled the comprehension of the information and the data handling from the source more effectively.

All questions created during the process had the main objective of answering both of these questions.

Breaking it down and relating to the column names.

- 1. Which local areas in Greater Vancouver have high potential for new business development?
  - New business growth over time → issueddate
  - Business closures over time → expireddate
  - Survival rate (active after 3–5 years) → issueddate, expireddate, status
  - Number of businesses by area and type → localarea, businesstype
  - Business saturation in each area → localarea, businesstype

    - Compare openings vs. closures per area → localarea, issueddate, expireddate, status
- 2. Which business types show strong past performance and future growth?
    - Growth in licenses per year → issueddate, businesstype
    - Long-term survival by type → issueddate, expireddate, status, businesstype
    - Common business status trends → status, businesstype
    - Average business lifespan → issueddate, expireddate, businesstype
    - Fastest-growing business types → issueddate, businesstype
    - Business size indicator → numberofemployees, businesstype

6. Architecture & Tools

Our project was structured to ensure scalability, efficiency, and clear insights for business decision-making.

- Data Sources:
    - City of Vancouver Open Data Portal
        - Business Licences 1997–2012
        - Business Licences 2013–2024
- Big Data Processing Framework:
    - Apache Spark (PySpark)
        - Distributed data processing
        - Efficient in-memory transformations and aggregations
- Programming Language:
    - Python 3.11
        - Easy integration with Spark
        - Rich ecosystem for data analysis (Pandas, Matplotlib)
- Cluster Environment:
    - Local Mode (Simulated cluster for development) (can be seamlessly adapted to cloud clusters like AWS EMR if scaling is needed)
- Data Storage Format:
    - CSV Files (original source)
    - Transformed internally into Spark DataFrames for optimized queries
- Visualization and Reporting Tools:
    - Matplotlib: To generate trend graphs and distribution analyses
    - Google Docs: For project documentation
    - GitHub: Version control, project tracking, and deployment

7. Results

7.1. Question 01

Neighbourhoods such as Downtown, Fairview, Mount Pleasant, and Kitsilano show the greatest potential for new business development. These areas combine high numbers of recent licences with relatively strong growth ratios and manageable churn rates. Notably, Downtown leads in both total and recent licences, while Renfrew-Collingwood stands out with a remarkably high average number of employees, indicating capacity for larger business operations.

7.2. Question 02

The business type with the fastest average growth in the number of licences issued per year is Single Detached House in Vancouver, showing a remarkable average year-over-year growth of 205.15%. Other high-growth categories include Office (97.23%) and Secondary Suite – Permanent (62.61%). These top-performing categories indicate strong and consistent expansion in Vancouver's urban development, housing, and professional service sectors. Emerging trends also appear in Health Services, Restaurant Class 1, and Wholesale Dealer, reinforcing their potential for future investment.

7.3. Question 03

The most stable business types, with a 100% stability rate, include specialized and niche categories such as Peddler – Food, Vending Machines, Lumber Yard, Christmas Tree Lot, Bowling Alley, and Auto Wrecker, among others. These sectors had no recorded closures within 5 years of operation. Among high-volume categories, One-Family Dwelling stands out with a 96.3% stability rate across over 9,500 businesses, demonstrating its strong resilience. Other reliable types include Personal Care Home, Exhibitions/Shows/Concerts, and Duplex, all with stability rates above 93%, making them promising for long-term investment.

7.4. Question 04

Among all business types analyzed, Bingo Hall stands out as the most promising, showing both 100% stability and 100% average year-over-year growth, suggesting strong historical performance and future potential. In contrast, Auto Parking Lot/Parkade had a moderate stability rate (63.4%) and very low growth (3.8%), indicating limited expansion potential.

7.5. Question 05

Cities with the highest HHI index — such as Anmore (0.1639), Belcarra (0.1634), and West Vancouver (0.1523) — show a high concentration of the same business types, suggesting potential oversaturation and limited diversification. In contrast, Vancouver (0.0439) and Burnaby (0.0764) have the lowest HHI scores, indicating a more diverse business ecosystem and lower saturation risk.

7.6. Question 06

Long ranking, preferably to take a look at the notebook.

7.7. Question 07

There is a clear seasonal pattern in business activity. January and December show the highest average business openings, with January reaching over 11,400 and December peaking at 15,262. On the other hand, business closures surge in October and December, with December averaging 2,029 closures, and October showing an unusual spike (19.62 closures on average), suggesting year-end operational shifts or regulatory deadlines.

7.8. Question 08

Throughout most of the observed period (1998–2024), renewals consistently dominate, averaging around 83% to 87% of all issued licences annually. Only in 1997 did new licences represent 100%, due to it being the initial year with no prior renewals. In recent years, the proportion of new licences has remained stable between 13% and 19%, indicating that the business ecosystem is primarily driven by existing businesses renewing operations, rather than by new entrants.

7.9. Question 09

The cancellation and inactivity rates vary significantly by business category. In 2024, Exhibitions/Shows/Concerts had a remarkably high inactivity/cancellation rate of 94.12%, followed by Temporary Liquor Licence Amendments (43.75%) and Electrical-Temporary (Filming) (34.62%), indicating high turnover or short-term nature. Historic and niche sectors like Casino, Dating Services, and Private Hospitals also showed elevated inactivity percentages. In contrast, more stable categories such as School (Private) and Pest Control exhibited much lower rates, suggesting stronger operational continuity.

7.10. Question 10

The most resilient and popular business subtypes combine high average longevity and a high percentage of active licences. Notable examples include Dentist (0.98 avg. longevity, 98.15% active), Physician/Surgeon (0.97 longevity, 97.94%), and Accountant/Auditor (0.93 longevity, 97.30%), indicating strong long-term stability. Among high-volume subtypes, Barrister & Solicitor, Consultant, and Building Contractors also stand out with longevity above 0.88 and over 95% active licences, making them particularly resilient and well-established sectors in the market.

8. Challenges & Solutions

One of the main challenges of this project was data cleaning and effectively answering the analytical questions. A significant portion of the raw data was either inconsistent, incorrect, or irrelevant to our overall business analysis goals.

For example, we encountered many entries where the city names were incorrectly recorded or referred to locations outside the Greater Vancouver Area. The original dataset documentation did not provide clear explanations for these anomalies. As a result, we needed to manually identify, filter, and correct these inconsistencies, which ultimately led to the removal of 64,143 records to ensure that our analysis was geographically accurate and meaningful.

Another major difficulty was in formulating the queries necessary to answer the business questions. This process proved to be particularly challenging due to the high level of abstraction required, which was beyond the typical query formulation we were accustomed to. To address this, we had to combine multiple intermediate queries, aggregations, and conditional filters into a single, more complex SQL statement. This iterative approach of layering queries allowed us to provide comprehensive answers that aligned with the real business needs of the investor and technical specialists.

Through these challenges, we significantly strengthened our skills in data preprocessing, query optimization, and analytical problem-solving — all essential competencies for handling real-world big data projects.

9. Conclusion

This project provided valuable insights into business trends across the Greater Vancouver area, uncovering patterns of growth, stability, and saturation across decades of business licence data. From a business perspective, the analysis highlights key neighbourhoods and business types with strong development potential, supporting strategic decisions for investors seeking scalable, resilient opportunities.

On the technical side, the project reinforced the importance of data cleaning, schema validation, and query optimization when working with large public datasets. Apache Spark proved to be an efficient and scalable platform for distributed processing, enabling the execution of complex aggregations and window functions on over 1.7 million records.

The solution is highly adaptable and scalable: its modular design using Spark SQL and PySpark can easily be extended to larger datasets or real-time data streams via tools like Apache Kafka. Future iterations of this project could incorporate additional variables such as tax records or geographic zoning data, further enriching the decision-making toolkit for urban planners, entrepreneurs, and investors.

10. References

- City of Vancouver Open Data Portal. (2024). Business Licences 2013 to 2024 Dataset. Retrieved from https://opendata.vancouver.ca/explore/dataset/business-licences-2013-to-2024/
- City of Vancouver Open Data Portal. (2024). Business Licences 1997 to 2012 Dataset. Retrieved from https://opendata.vancouver.ca/explore/dataset/business-licences-1997-to-2012/
- Cornerstone International Community College of Canada (CICCC). (2025). Final Unpublished Project Instructions.
- Introduction to Apache Spark. (2025). Course Material. Retrieved from internal distribution.
- Spark DataFrame Guide. (2025). Course Material. Retrieved from internal distribution.