

**Signature Assignment - Final Project Research Paperwork (Group Assignment)**

**Yelp Business Performance Analysis**

**ALY6110 – Data Management and Big Data**

**CRN: 80437**

**Group Delta**

**Likhitha Varakala (**[**varakala.l@northeastern.edu**](mailto:varakala.l@northeastern.edu)**)**

**Manmitha Pantangi (**[**pantangi.m@northeastern.edu**](mailto:pantangi.m@northeastern.edu)**)**

**Ayushi Walia (**[**walia.ay@northeastern.edu**](mailto:walia.ay@northeastern.edu)**)**

**Aditi Rajmane (**[**rajmane.a@northeastern.edu**](mailto:rajmane.a@northeastern.edu)**)**

**Zhiyuan Xue (**[**xue.zhiyu@notheastern.edu**](mailto:xue.zhiyu@notheastern.edu)**)**

**Summary**

The project aimed to analyze the Yelp dataset using PySpark, Spark SQL, and various data analysis techniques. The steps involved in the project included data preprocessing, exploratory data analysis, clustering, aggregation, and visualization. The purpose of the project was to gain insights into the Yelp dataset and uncover meaningful patterns and trends related to organizations, ratings, reviews, categories, cities, and states. The expected results included identifying the top-rated organizations, analyzing the distribution of ratings and reviews, exploring the most common categories, and visualizing geographic patterns. The clustering results can provide insights for decision-making, such as identifying groups of businesses with similar characteristics or targeting specific clusters for further analysis or marketing strategies. The project utilized PySpark's scalability and distributed processing capabilities, Spark SQL's SQL-like querying, and various data analysis techniques to derive valuable insights from the Yelp dataset.

**Content**

**Dataset:**

The dataset that we chose for the project is the Yelp dataset obtained from Kaggle. It consists of a collection of data from Yelp, a platform for user-generated reviews of businesses. The dataset contains approximately one million rows and 14 attributes. The attributes are as follows:

ID: The unique identifier for each entry in the dataset.

Time\_GMT: The time of the entry in GMT (Greenwich Mean Time).

Phone: The phone number associated with the business.

Organization: The name or organization associated with the business.

OLF: Unknown variable

Rating: The average rating given to the business.

NumberReview: The total number of reviews for the business.

Category: The category or type of business (e.g., restaurant).

Country: The country where the business is located.

CountryCode: The country code of the business location.

State: The state or province where the business is located.

City: The city where the business is located.

Street: The street address of the business.

Building: The building number or identifier associated with the business.

**Problem Statement:** The problem being addressed in the analysis of the Yelp dataset is to understand the factors that contribute to the success of businesses on Yelp. By exploring the dataset and performing various analyses, the goal is to uncover insights that can guide business owners in enhancing their performance, improving customer satisfaction, and ultimately achieving success on the Yelp platform.

**Benefits:**

1. Business Performance Improvement: The analysis helps businesses identify areas of strength and weakness, allowing them to focus on improving their products, services, or customer experience. Insights into high-performing categories or cities can guide business expansion or diversification strategies.
2. Customer Satisfaction Enhancement: Understanding the factors that drive positive ratings and reviews enables businesses to tailor their offerings to meet customer expectations and preferences. By addressing any identified pain points, businesses can enhance their overall customer satisfaction and reputation.
3. Competitive Advantage: Uncovering patterns and trends in the dataset provides businesses with a competitive edge by identifying opportunities or niches that have been overlooked. Businesses can differentiate themselves by leveraging insights to create unique value propositions or targeted marketing strategies.

**Drawbacks:**

1. Data Limitations: The analysis relies on the quality and completeness of the Yelp dataset. Inaccurate or missing data can impact the reliability of the findings.
2. External Factors: The analysis is limited to the Yelp dataset and may not account for external factors that influence business success, such as economic conditions or industry dynamics.

**Challenges:**

1. Big Data Processing: Handling a large dataset like Yelp's with millions of rows requires efficient big data processing techniques, such as distributed computing using tools like PySpark.
2. Data Preprocessing: Preparing the Yelp dataset for analysis involves dealing with missing values, data cleaning, and transforming the data into a suitable format.
3. Complex Relationships: Unraveling the complex relationships and interactions between various factors affecting business success on Yelp can be challenging.
4. Interpreting Results: Interpreting the analysis results and deriving meaningful insights from the data requires a deep understanding of the business context and domain expertise.

**Tools and Techniques used:**

In the project, several tools and techniques were used to analyze the Yelp dataset. These tools and techniques were chosen for their ability to handle large-scale data processing, provide efficient data analysis capabilities, and offer ease of use. Let's discuss the tools and techniques used and their respective justifications:

1. PySpark:
   * PySpark is the Python API for Apache Spark, a fast and distributed data processing framework. PySpark was chosen because it allows for scalable and distributed data processing, making it suitable for handling large datasets like the Yelp dataset.
   * PySpark provides a high-level API that integrates well with Python and enables the use of familiar Python libraries and syntax. This allows for efficient and expressive data manipulation and analysis.
2. Spark SQL:
   * Spark SQL is a module in Apache Spark that provides a programming interface for working with structured and semi-structured data. It allows you to query and analyze structured data using SQL-like syntax.
   * Spark SQL was used in the project to leverage its powerful SQL capabilities for data querying, filtering, and aggregation operations. It provides a convenient way to perform SQL-like operations on structured data within PySpark.
3. DataFrame API:
   * The DataFrame API in PySpark provides a tabular data structure that allows for easy manipulation and analysis of structured data. DataFrames are distributed and optimized for parallel processing, making them well-suited for big data scenarios.
   * DataFrames were used in the project to perform various data manipulation operations such as grouping, aggregating, filtering, and joining. The DataFrame API provides a high-level abstraction that simplifies the data analysis process.
4. Data Preprocessing Techniques:
   * Data preprocessing techniques were applied to clean and transform the raw Yelp dataset before analysis. These techniques included handling missing values, renaming columns, and converting data types as needed.
   * Data preprocessing is essential to ensure the quality and integrity of the data before performing analysis. It helps in removing inconsistencies, standardizing the data format, and preparing it for further analysis tasks.
5. Exploratory Data Analysis (EDA):
   * EDA techniques were used to gain initial insights into the Yelp dataset. These techniques included calculating summary statistics, visualizing distributions, and exploring relationships between variables.
   * EDA is crucial for understanding the characteristics of the data, identifying patterns, and discovering interesting insights. It helps in formulating hypotheses and guiding subsequent analysis tasks.
6. Aggregation and Statistical Functions:
   * Aggregation functions such as average (avg) and count were used to summarize and analyze the Yelp dataset. These functions allow for calculating summary statistics and aggregating data based on specific criteria.
   * Statistical functions provide valuable insights into the data distribution, central tendency, and other statistical properties. They help in understanding the overall patterns and characteristics of the dataset.
7. Clustering algorithm:

* The employed clustering technique in this analysis is the k-means algorithm, which identifies separate clusters in the dataset by considering restaurants' ratings and number of reviews. Each data point is assigned to a specific cluster, and the algorithm iteratively refines the cluster centers to minimize the distance between data points and their respective centers.
* Widely applied for unsupervised clustering, this algorithm offers valuable insights into the categorization and attributes of restaurants based on their ratings and reviews.

**Installations:**  
Here is a step by instep installing of Pyspark and Jupyter on the computer that already has Java Development kit and Spark.

1. Install PySpark:
   * Open a terminal or command prompt and install PySpark using pip code “pip install pyspark.”

A screenshot of a computer program

Description automatically generated with medium confidence

1. Install Jupyter Notebook:
   * Install Jupyter Notebook using pip code “pip install jupyter.”
2. Launch Jupyter Notebook:
   * Open a terminal or command prompt and navigate to the directory where you want to start Jupyter Notebook.
   * Run the following command to start Jupyter Notebook code “jupyter notebook.”
   * This will open Jupyter Notebook in your web browser.



1. Create a New Notebook:
   * In Jupyter Notebook, click on "New" and select "Python 3" to create a new notebook.
   * You can now write and execute PySpark code in the notebook cells.

A screenshot of a computer

Description automatically generated with medium confidence

**Analysis:**

**Data Loading and Preprocessing:**

* The Yelp dataset, containing information about organizations, ratings, review counts, categories, cities, and states, is loaded into PySpark.

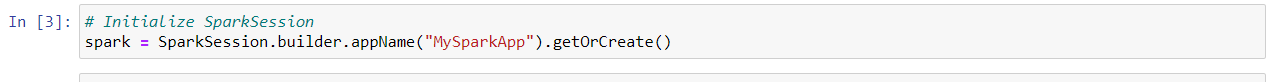
Imported libraries,

* SparkSession, desc, count and avg from pyspark.sql.
* VectorAssembler from pyspark.ml.feature
* KMeans from pyspark.ml.clustering
* Matplotlib.pyplot

A screenshot of a computer

Description automatically generated with medium confidence

Initialize the Spark Session:



Import the Yelp dataset:



Contents of the dataset:

A screenshot of a computer

Description automatically generated with medium confidence

**Exploratory Data Analysis (EDA):**

EDA techniques are applied to gain initial insights into the dataset.

This involved examining summary statistics, visualizing distributions of ratings and review counts, and exploring the category.

1. **Average Rating by Organization:**

The dataset is grouped by organization, and the average rating for each organization is calculated using the groupBy() and avg() functions. This analysis provides insights into the overall customer satisfaction level for each organization.

A screenshot of a computer

Description automatically generated with medium confidence

1. **Top 10 most reviewed businesses:**

Using the groupBy() function, the dataset is grouped by organization and sorted in decreasing order of the number of reviews received. For the purpose of displaying the top 10 most popular companies, these organizations are limited to a count of 10. With a total of 6956 reviews, Domino's Pizza is the most popular business, followed by Subway with 6775 reviews, according to the table.A screen shot of a computer

Description automatically generated with low confidence

1. **Popular Categories by Review Count:**

The dataset is analyzed to determine the most popular categories based on the number of reviews received. This analysis identifies categories that attract a significant number of reviews, indicating high customer engagement.

A picture containing text, font, line, screenshot

Description automatically generated

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

The bar graph makes it quite evident that the 'Plumbing' category has attracted the most reviews, with 51k, followed by 'Delivery' with 34k. The ‘Air Conditioning & Heating’ category has received the lowest number of reviews i.e., 14k, as compared to other two categories.

1. **Average rating for each category:**

The dataset is analyzed to determine the average rating associated to each of the categories. The dataset is grouped by category and average rating is calculated using groupBy() and agg(avg()) functions.

A picture containing text, font, line, screenshot

Description automatically generated

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

It is evident from the bar graph that the ‘Delivery’ category has the highest average rating of 3.5 as compared to other categories. It is to be noted that, even though the ‘Plumbing’ category has received the greatest number of reviews, it has the least average rating of 2.16. The ‘Air Conditioning & Heating’ category has recorded an average rating of 3.12.

1. **Distribution of Yelp Businesses by Category:**

The analysis involved aggregating the counts of businesses by category using the "yelp\_df" DataFrame and sorting the results in descending order. The resulting data was then plotted using Matplotlib and a Pandas DataFrame. The bar chart effectively displays the categories on the x-axis and the corresponding business counts on the y-axis. By providing a clear overview of the distribution, this visualization allows for easy identification of the categories with the highest and lowest business counts, enabling valuable insights for marketing, business strategies, and user preferences within the Yelp platform. Here, we can see that count of 'plumbing' category is 50k whereas for 'Air Conditioning and heating' Category it is approximately 15000.

A picture containing text, font, screenshot

Description automatically generated

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

1. **Distribution of Ratings:**

To analyze the distribution of Ratings, a subset of the dataset consisting of ‘rating’ column is taken and converted into Pandas for creating a histogram.

A picture containing text, font, screenshot

Description automatically generated

A picture containing text, diagram, screenshot, plot

Description automatically generated

The distribution of ratings across the dataset is shown in the histogram. The ratings range from 0 to 5, with 5 representing the highest rating. The graph clearly shows that the majority of the dataset's organizations either obtained the highest rating of 5.0 or the lowest rating of 0. About 30k organizations have received the highest rating of 5.0, while 25k have received a rating of 0.

1. **Top 20 organizations county by city:**

By using agg() and groupBy() functions, the count of organizations based on the city are calculated.

A screen shot of a computer code

Description automatically generated with low confidence

A picture containing text, screenshot, diagram, plot

Description automatically generated

The bar graph clearly shows the top 20 organizations count by city. The most organizations are in Los Angeles, where there are about 13.7k in total. With approximately 11.8k organizations overall, Chicago ranks as the second most popular city. With the fewest number of organizations—about 3800—Kansas City makes the top 10.

**Clustering Analysis**

We utilized the k-means algorithm to conduct clustering analysis.

The code includes the following imports:

* **VectorAssembler** is utilized to combine selected features into a single feature vector column.
* **KMeans** is the clustering algorithm employed for executing K-means clustering.
* **SparkSession** is used to create a SparkSession, which serves as the starting point for working with Spark.

A screenshot of a computer

Description automatically generated

In this code, the 'Rating' and 'NumberReview' features are chosen as the relevant ones for clustering. A VectorAssembler is employed to merge these features into a single feature vector. Subsequently, a KMeans instance is created with the desired number of clusters (k). The KMeans model is then trained on the data, resulting in cluster assignments for each business. Finally, the cluster assignments are displayed, and the coordinates of each cluster center are printed.

The k-means algorithm was employed to identify separate clusters within the dataset by considering the rating and number of reviews. Each data point was assigned to one of three clusters, labeled as Cluster 0, Cluster 1, and Cluster 2.

The clustered\_df DataFrame will display the cluster assignments for each business. It will contain the original features as well as an extra column indicating the assigned cluster for each business.

A screenshot of a computer

Description automatically generated

The cluster\_centers variable will store the coordinates of each cluster center. These coordinates indicate the average values of the selected features within each cluster.

A screenshot of a computer

Description automatically generated

Cluster 0 Center: [Rating: 2.72, Number of Reviews: 26.49]

Cluster 1 Center: [Rating: 4.06, Number of Reviews: 2314.81]

Cluster 2 Center: [Rating: 3.98, Number of Reviews: 523.50]

**Cluster Analysis: Rating vs. Number of Reviews**

A scatter plot was generated to visually represent the distribution of data points across different clusters, taking into account their rating and number of reviews. This scatter plot aids in comprehending how restaurants are grouped together and identifying any discernible patterns or relationships.

A picture containing text, screenshot, diagram, plot

Description automatically generated

In the scatter plot, each data point corresponds to a restaurant, positioned based on its rating (x-axis) and number of reviews (y-axis). Different clusters are visually distinguished by using distinct colors.

Based on the scatter plot, the following observations can be made:

**Cluster 0:** This cluster represents restaurants with moderate ratings and a relatively low number of reviews. On average, these restaurants have a rating of approximately 2.72 and around 26.49 reviews. They appear to be less popular compared to the other clusters.

**Cluster 1:** Restaurants in this cluster exhibit high ratings and significantly higher numbers of reviews compared to the other clusters. The average rating for this cluster is approximately 4.06, with an average number of reviews around 2314.81. These restaurants are likely to be popular and well-established.

**Cluster 2:** This cluster comprises restaurants with above-average ratings and a moderate number of reviews. The average rating for this cluster is approximately 3.98, with an average of about 523.50 reviews. These restaurants are relatively popular but not as widely reviewed as those in Cluster 1.

Overall, the cluster analysis provides insights into the distribution and characteristics of restaurants based on their ratings and number of reviews. The identified clusters allow for the segmentation of restaurants into distinct groups, facilitating targeted analysis, decision-making, and personalized recommendations.

**Comments**

**Pros:**

* Scalability: PySpark, being a distributed processing framework, allowed for efficient handling of large-scale datasets like the Yelp dataset. The distributed nature of PySpark enabled parallel processing, enabling faster data analysis.
* Expressive Data Manipulation: PySpark's DataFrame API provided a high-level abstraction for working with structured data. This allowed for easy and expressive data manipulation operations, such as filtering, grouping, and aggregating, making the analysis tasks more straightforward.
* Integration with Spark SQL: The integration of Spark SQL within PySpark offered the advantage of performing SQL-like operations on structured data. This flexibility enabled the use of SQL queries for data querying and aggregation, which is familiar and widely used.
* Data Preprocessing Capabilities: PySpark provided various functions and methods for data preprocessing tasks, such as handling missing values, renaming columns, and converting data types. These capabilities simplified the data cleaning process and ensured data quality.

**Cons:**

* Learning Curve: Working with PySpark, Spark SQL, and DataFrame API may have a steep learning curve for individuals new to the framework. Understanding the distributed nature of PySpark and its APIs required time and effort to grasp.
* Debugging and Error Handling: Debugging and error handling in PySpark can be challenging, as error messages may not always provide detailed information on the root cause. Identifying and resolving errors might require thorough understanding and troubleshooting.
* Resource Requirements: PySpark requires a significant number of computational resources, including memory and processing power, especially when working with large datasets. Ensuring sufficient resources for efficient execution is crucial.

**What Worked:**

* The PySpark framework provided efficient and scalable processing capabilities for handling the large-scale Yelp dataset.
* EDA techniques allowed for a comprehensive exploration of the dataset, providing initial insights into the data's characteristics.
* Analytical tasks, such as calculating average ratings, AND analyzing category distributions, were successfully executed.

**What Did Not Work and Why Not:**

* Data quality issues: Incomplete or inaccurate data in the Yelp dataset could impact the reliability of the analysis results.
* Complexity of relationships: Unraveling complex relationships between variables influencing business success may require more advanced analytical techniques or additional data sources.
* Limitations in external factors: The analysis may not fully account for external factors such as economic conditions or industry dynamics, which can influence business success.

**Conclusion**

Throughout the project, the experience of working with PySpark for analyzing the Yelp dataset was insightful and valuable. The findings and insights derived from the analysis provided meaningful information about the organizations, ratings, reviews, categories, cities, and states in the dataset. Here is a summary of the key conclusions:

1. Data Analysis and Insights:
   * The analysis revealed the top-rated organizations, allowing for recognition of businesses that consistently received positive reviews.
   * Understanding the distribution of ratings and reviews provided insights into customer sentiments and satisfaction levels.
   * Exploring the most common categories helped identify popular business types and industry trends.
   * The conclusions from the cluster analysis provide Yelp with actionable insights to enhance recommendations, targeted marketing, and market understanding for improved user experience and support for restaurant owners.
2. Limitations and Future Considerations:
   * The analysis focused on a specific dataset, and the findings are specific to the Yelp dataset used. Generalization to other datasets should be done cautiously.
   * Further analysis could include sentiment analysis of reviews using natural language processing techniques to extract more detailed insights.
   * Incorporating external data sources, such as demographic data or social media trends, could enrich the analysis and provide a broader context.
3. Technical Challenges and Learnings:
   * Working with PySpark and distributed computing frameworks presented challenges related to debugging, resource management, and performance optimization.
   * The project provided an opportunity to learn and understand the complexities of distributed data processing and how to address them effectively.

Working on this project reinforced the importance of data preprocessing, exploratory data analysis, and visualization techniques in deriving meaningful insights from large datasets. PySpark proved to be a powerful tool for scalable data processing and analysis, offering extensive functionality and integration with Spark SQL. It was a valuable experience to work with big data analytics tools and techniques and witness the potential they hold in extracting valuable information from vast datasets.

**References**

*Yelp Dataset - Contains 1 million rows*. (2022, January 28). Kaggle. <https://www.kaggle.com/datasets/abdulmajid115/yelp-dataset-contains-1-million-rows>

Naveen. (2023). Install PySpark in Anaconda & Jupyter Notebook. *Spark by {Examples}*. <https://sparkbyexamples.com/pyspark/install-pyspark-in-anaconda-jupyter-notebook/>

Kushwaha, B. (2021, December 12). How to visualize a dataset in Apache Spark — PySpark. *Medium*. <https://medium.com/@brijeshgzp05/how-to-visualize-a-dataset-in-apache-spark-pyspark-a32abd8ca4f4>

*Clustering - Spark 3.4.0 Documentation*. (n.d.). <https://spark.apache.org/docs/latest/ml-clustering.html>

Das, A. (2021, December 30). K Means Clustering using PySpark on Big Data - Towards Data Science. *Medium*. <https://towardsdatascience.com/k-means-clustering-using-pyspark-on-big-data-6214beacdc8b>

Kulkarni, P. (2021, August 10). *Tutorial : K-Means Clustering on Spark - Datasset to Mindset*. Datasset to Mindset. <https://www.data4v.com/tutorial-k-means-clustering-on-spark/>