Market-Basket Analysis of Job Skills

Piscitelli John Carlo
Master's in Data Science for Economics
Course: Algorithm for Massive Datasets
Professor: Dr. Malchiodi Dario

September 12, 2024

1 Introduction

This project aims to analyze job skill requirements using a market-basket analysis approach, focusing on identifying frequently co-occurring skills in job postings.

1.1 Goal

The primary objective of this project is to apply market-basket analysis techniques in order to identify the most frequently occurring combinations of skills, which can help in understanding the demand for specific skill sets in specific jobs field.

1.2 Dataset

The analysis utilizes the **LinkedIn Jobs & Skills dataset**, specifically the `job_skills.csv` file which contains job postings. It includes a column named `job_skills` that lists skills required for each job. The dataset comprises approximately 1.3 million rows, providing a full view of job skill requirements in different fields.

1.3 Methodology

To achieve the goal, the following methodology was employed:

1. **Data Preprocessing**:
    - **Data Cleaning**: The dataset was filtered to exclude rows with null or empty `job_skills` values.
    - **Skill Extraction**: Skills were extracted from the `job_skills` column, which contains skills separated by commas. Each job's skills were converted into a set to facilitate further analysis. Skills whose

number of characters were either shorter than 2 or longer than 20 were excluded.

- **Standardization**: Skills and skill pairs were standardized to ensure consistency in representation, which helps in accurate frequency counting. Also, duplicate skills per basket were removed.

2. **Market-Basket Analysis**:

   - **Apriori Algorithm**: The Apriori algorithm was applied to identify frequent single skills and skill pairs. This algorithm helps in discovering associations by generating frequent itemsets and pruning non-frequent itemsets iteratively.
   - **FP Growth**: The FPGrowth (Frequent Pattern Growth) algorithm is a powerful technique used in data mining to identify frequent itemsets in a transactional dataset. FPGrowth constructs a compact data structure called an FP-Tree to store frequent itemsets in a more efficient manner, making it particularly useful for large-scale data mining tasks.

3. **Analysis and Visualization**:

   - **Frequent Itemsets**: The frequent skill pairs and triplets were identified and analyzed to determine the most common combinations of skills.
   - **Visualization**: The results were visualized using horizontal bar charts to present the top frequent skill pairs, aiding in the interpretation of the most sought-after skill combinations.

By using these techniques, the project aims to provide insights into the skill requirements in the job market.

2 Apriori Algorithm for Market Basket Analysis

The Apriori algorithm is a method used to identify frequent itemsets within large datasets. It operates based on the principle that if an itemset is frequent, then all its subsets must also be frequent. This property is known as the *Apriori property.*

2.1 Mathematical Framework

Let D be a database of transactions, and let I be a set of items. The Apriori algorithm aims to find itemsets whose support exceeds a user-defined minimum support threshold.

2.2 Generating Frequent Itemsets

Definition: An itemset X is considered frequent if its support is greater than or equal to the minimum support threshold min_support, in our case, the minimum support of choice is 2000 occurrencies, roughly 0.15 percent of the entire dataset.

- **Support**: The support of an itemset X is defined as the proportion of transactions in which X appears.

$$\text{Support}(X) = \frac{\text{Number of Transactions Containing } X}{\text{Total Number of Transactions}}$$

Algorithm Steps:

1. **Initialization**: Start with single items and calculate their support.
2. **Iterative Process**:
   - **First Pass**: Count the frequency (support) of each item.
   - **Second Pass**: Define the frequency of itemsets of size two (pairs)
   - **Third Pass**: Continue computing the occurrency of itemsets of increasing size k from the frequent itemsets of size $k-1$ until no more frequent itemsets can be found.

2.3 Application of Apriori Algorithm

Implementation:

1. **Data Preparation**: Clean and preprocess the dataset.
2. **Apply Apriori**: Run the Apriori algorithm to find frequent itemsets.

2.4 Findings

Overview of the Singletons

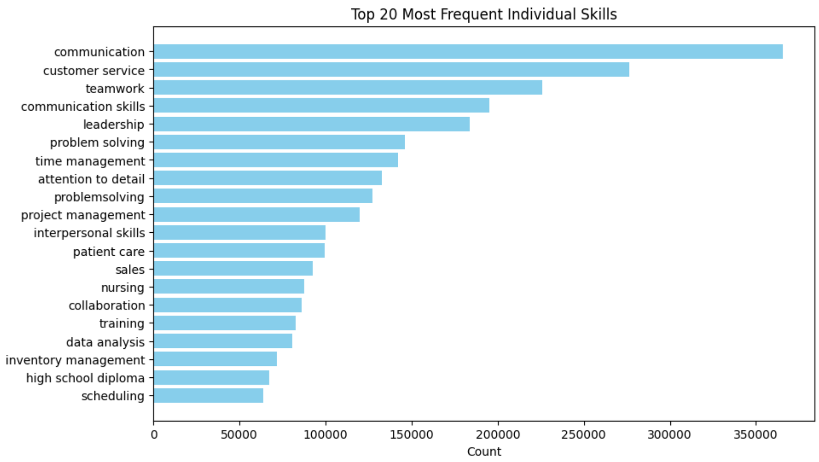

Figure 1: Top 20 Most Frequent Individual Skills

Firstly, we analyze the 20 most frequent skills appearing in job postings. Unsurprisingly, the highest positions are taken by transversal soft skills like *Communication*, *Teamwork*, and *Problem Solving*, underlining how social skills and the ability to work in a group are essential to any sector. *Data Analysis* is also present in the top 20, which suggests the growing importance of data-related skills across different sectors. Nursing and Patient Care also appear in the list, indicating a need for healthcare personnel.

Overview of the Skill Pairs

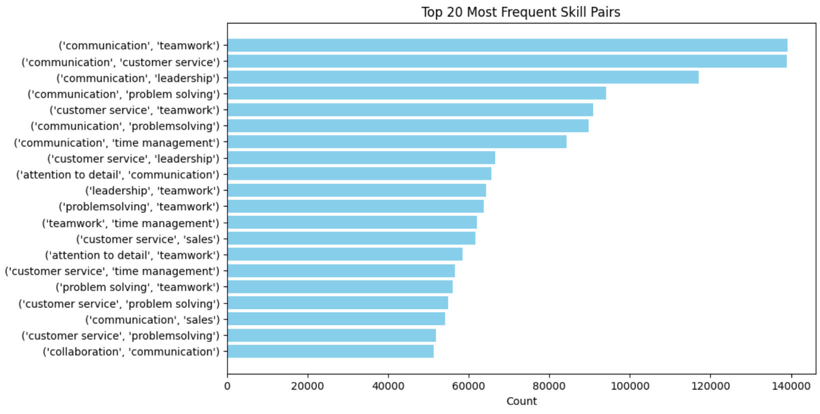

Figure 2: Top 20 Most Frequent Skill Pairs

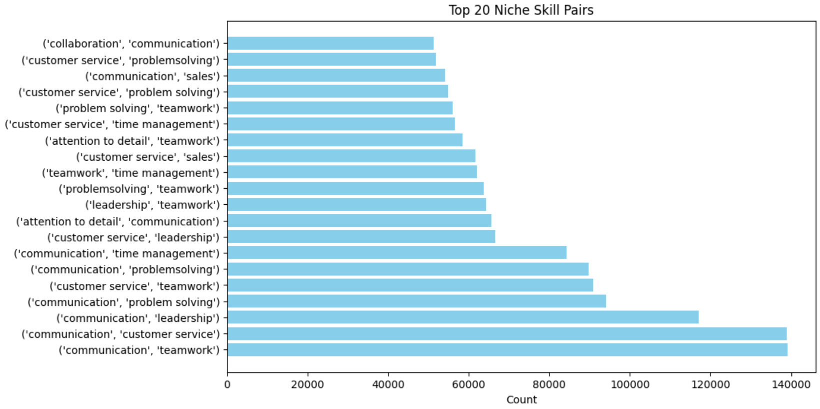

Figure 3: Top 20 Niche Skill Pairs

The overview of the frequently occurring pairs, determined by the Apriori algorithm, confirms the trend observed during the analysis of the singletons: the most sought-after skills are soft skills, specifically related to communicating to both colleagues and customers, as well as teamwork. This is evident in both the Top 20 most common pairs and the Top 20 niche pairs.

Overview of the Skill Triplets

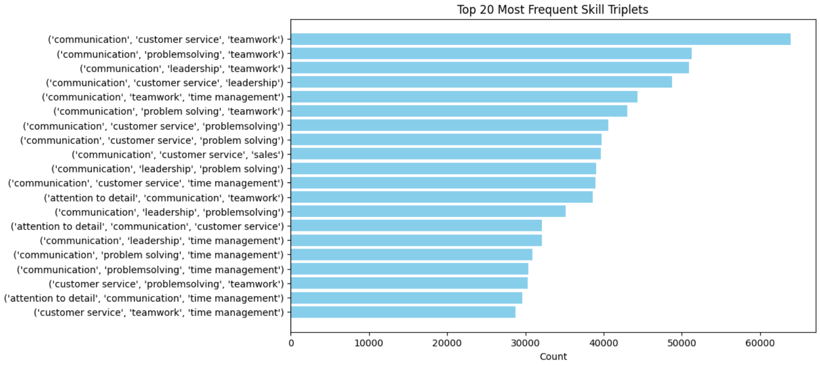

Figure 4: Top 20 Most Frequent Skill Triplets

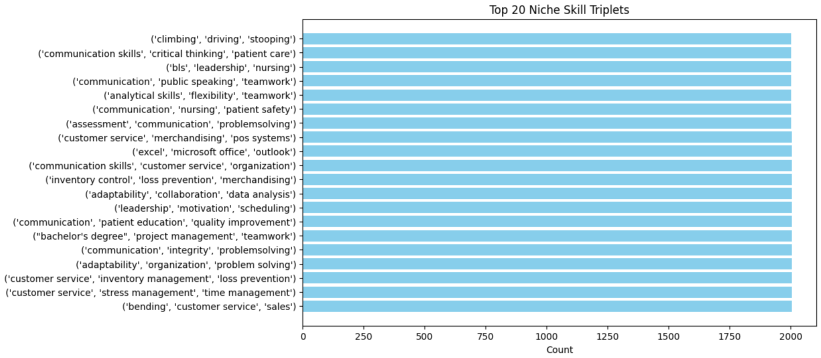

Figure 5: Top 20 Niche Skill Triplets

Similarly to the pairs, the most frequent triplets revolve around communication, customer service, and teamwork. These skills are highly valued across various job postings.

However, the analysis of niche triplets highlights some specific demands:

- **Climbing, Driving, Stooping**: This combination pertains to physical jobs, such as logistics or maintenance positions, where physical agility and strength are relevant.

- **Communication Skills, Critical Thinking, Patient Care**: This triplet suggests that in healthcare or service-related roles, being able to communicate effectively, think critically, and provide care to patients is highly valued.

- **Adaptability, Collaboration, Data Analysis**: This combination highlights the importance of flexibility in roles involving both teamwork and data analysis, possibly in dynamic business environments or research fields.

3 FP-Growth Algorithm for Market Basket Analysis

3.1 FP-Growth Algorithm and computational demand of the Apriori Algorithm

The FP-Growth (Frequent Pattern Growth) algorithm is an efficient alternative to the Apriori algorithm for identifying frequent itemsets in large datasets. Unlike Apriori, FP-Growth does not generate candidate itemsets. Instead, it uses a compressed data structure called the FP-Tree (Frequent Pattern Tree) to represent the dataset and recursively extracts frequent itemsets from this tree, which makes it faster and more scalable for large datasets.

One of the major drawbacks of the Apriori algorithm is its computational complexity, especially when dealing with large datasets and higher-order itemsets. As the number of itemsets grows, Apriori struggles to scale efficiently due to the generation of candidate itemsets at each iteration.

By implementing FP-Growth, we were able to explore higher-order itemsets to perform analysis which would be problematic to run using Apriori alone.

3.2 Mathematical Framework

Let D represent a database of transactions, where each transaction contains a set of items from a universal set of items I. The FP-Growth algorithm aims to find frequent itemsets that meet a user-defined minimum support threshold, denoted by *min support*.

- **Itemset:** A set of one or more items from I.
- **Support:** The support of an itemset $X \subseteq I$ is defined as the proportion of transactions in the dataset D in which the itemset X appears. Mathematically:

$$\text{Support}(X) = \frac{\text{Number of Transactions Containing } X}{\text{Total Number of Transactions}}$$

- **Frequent Itemset:** An itemset X is considered frequent if its support is greater than or equal to the minimum support threshold, i.e., $\text{Support}(X) \geq$ *min support*.

3.3 Generating Frequent Itemsets

FP-Growth leverages the FP-Tree, a compact representation of the dataset, which allows frequent itemsets to be generated without candidate generation. The key steps are as follows:

3.4 Building the FP-Tree

The FP-Growth algorithm starts by constructing a data structure known as the Frequent Pattern Tree (FP-Tree). The FP-Tree is a compressed representation of the transaction database that retains the frequency of individual items and the relationships between them. The construction of the FP-Tree involves several key steps:

1. **First Database Scan:** The algorithm first scans the transaction database to identify frequent items. Items that do not meet the minimum support threshold are discarded at this stage. The remaining items are sorted in descending order of their frequency, which helps in building a compact and efficient FP-Tree.

2. **Second Database Scan:** In the second scan, the algorithm processes each transaction in the dataset. For each transaction, it creates a path in the FP-Tree that links the items in the transaction. The items are added in the same frequency-based order determined in the first scan. If an item already exists in the current path, the algorithm increments the count of the existing node rather than creating a new node.

3. **Shared Prefixes:** One of the key features of the FP-Tree is that transactions that share common items will share the same path in the tree. This sharing of paths leads to significant compression of the dataset, as common patterns are stored only once rather than being duplicated. For example, if multiple transactions include the itemset $\{A, B, C\}$, they will all follow the same path in the FP-Tree, with the counts of the nodes being incremented accordingly.

3.5 Recursive Pattern Growth

Once the FP-Tree is built, the FP-Growth algorithm begins the process of mining frequent patterns. This is done recursively through the following steps:

1. **Mining Frequent Patterns:** Starting from the least frequent item in the header table, the algorithm traces all paths in the FP-Tree where this item appears. These paths are used to construct a conditional pattern base, which is essentially a subset of transactions that co-occur with the item. Based on this pattern base, a conditional FP-Tree is built.

2. **Recursive Mining:** The algorithm recursively mines each conditional FP-Tree to extract frequent itemsets. For each conditional FP-Tree, new frequent itemsets are found, and the process continues until no more frequent patterns can be extracted from the tree.

3. **Efficiency Gains:** By using conditional FP-Trees, FP-Growth avoids generating and testing candidate itemsets. The recursive nature of the algorithm allows it to focus on smaller subsets of the dataset in each iteration, which makes it significantly more efficient than Apriori for large datasets.

In fact, unlike the Apriori algorithm, which generates candidate itemsets and performs multiple scans of the database, FP-Growth compresses the dataset into an FP-Tree and performs a recursive search for frequent patterns. This reduces the computational cost and time, especially for large datasets.

3.6 Application of FP-Growth Algorithm

Implementation:

1. **Data Preparation:** Clean and preprocess the dataset.

2. **Apply FP-Growth:** Construct the FP-Tree and recursively extract frequent itemsets using the FP-Growth algorithm.

3.7 Findings

Overview of the Skill Quadruplets

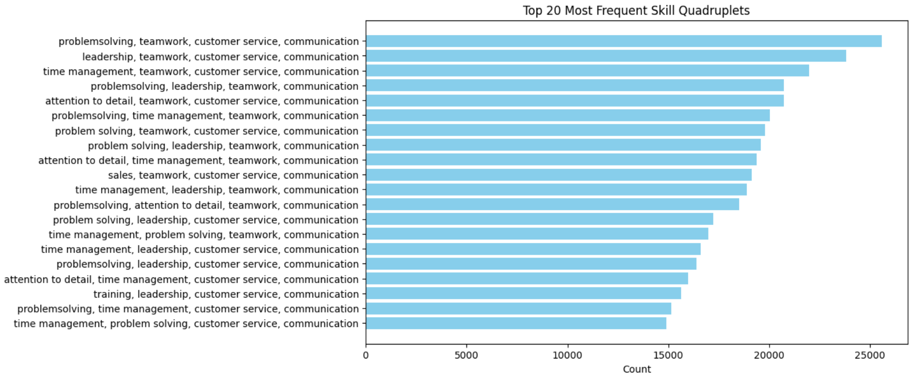

Figure 6: Top 20 Most Frequent Skill Quadruplets

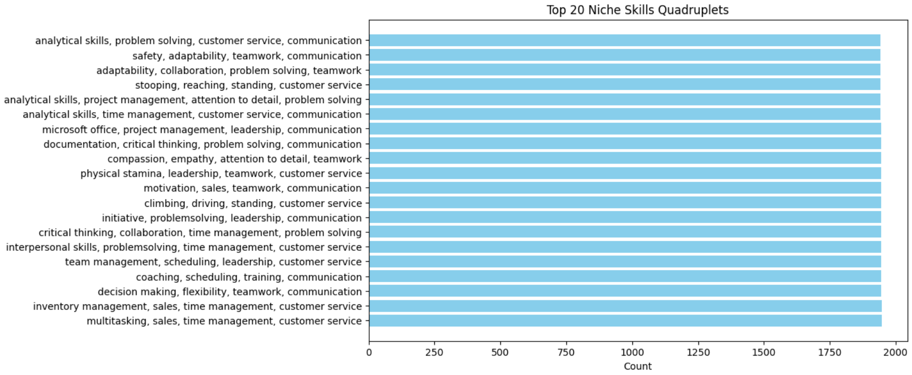

Figure 7: Top 20 Niche Skill Quadruplets

Not surprisingly, the overview of the quadruplets, emphasizes soft skills, with communication, teamwork, customer service, and problem-solving being the most frequent skills that appear in various combinations The chart depicting the Top 20 Niche Quadruplets however, hints at more specific roles that are sought after:

- **Analytical Skills, Problem Solving, Customer Service, Communication**: This combination suggests roles that require both analytical thinking and the ability to effectively communicate with clients.

- **Safety, Adaptability, Teamwork, Communication**: This combination could point to safety-critical industries, and offers an overview of the skillsets that is sought after.

- **Climbing, Driving, Standing, Customer Service**: This indicates a niche demand for roles that require physical stamina and skills, possibly in sectors like logistics, maintenance, or outdoor services, where interaction with customers is also expected.

Overview of the Skill Quintuplets

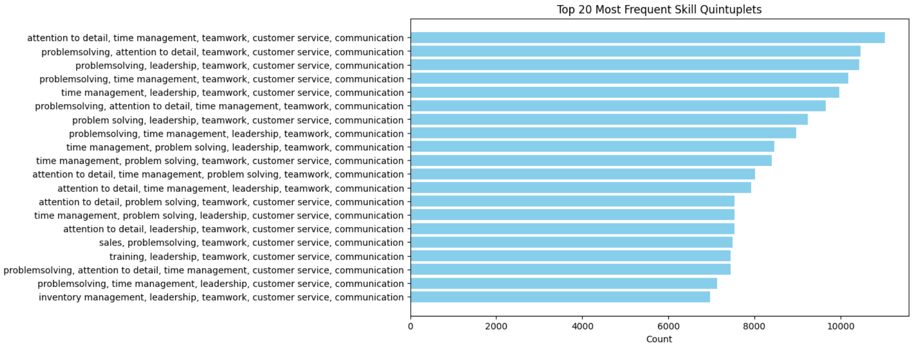

Figure 8: Top 20 Most Frequent Skill Quintuplets

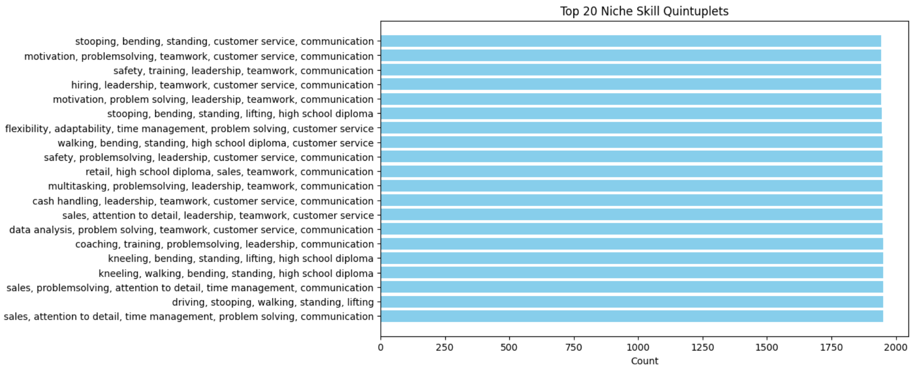

Figure 9: Top 20 Niche Skill Quintuplets

The Top 20 Most Frequent Quintuplets generated by the FP-Growth algorithm again reveal a tendency for employee to directly search for a wide range of soft skills related to the social and communicative sphere. The overview of the niche quintuplets, however, offers some insight regarding unique combinations of skills:

- **Stooping, Bending, Standing, Customer Service, Communication**: This quintuplet likely refers to physical demanding jobs where constant movement is required, as well as interacting with customers, for example retail or logistics.
- **Data Analysis, Problem Solving, Teamwork, Customer Service, Communication**: This combination highlights how, in Data Analytics positions, it is important to work in a team and have good communicative skills with customers.

Absence of Skill Sextuplets

The FP-Growth algorithm did not identify any sextuplet with frequency above support. This suggests that job postings focus on a limited set of skills, which, by our analysis, mostly concern transversal soft skills which are required at any level. Quintuplets provide an extensive view of what companies are looking for, and any additional skill would not be adding value to job postings.

4 Conclusion: An Overview of Data Analytics Job Offerings

In the concluding phase of our market-basket analysis, we focused on job skills related to the field of data analytics. The analysis was performed using the FP-Growth algorithm. We will be focusing on the most relevant pairs and quintuplets containing target skills related to data analysis, data science, and machine learning.

4.1 Focus on Pairs

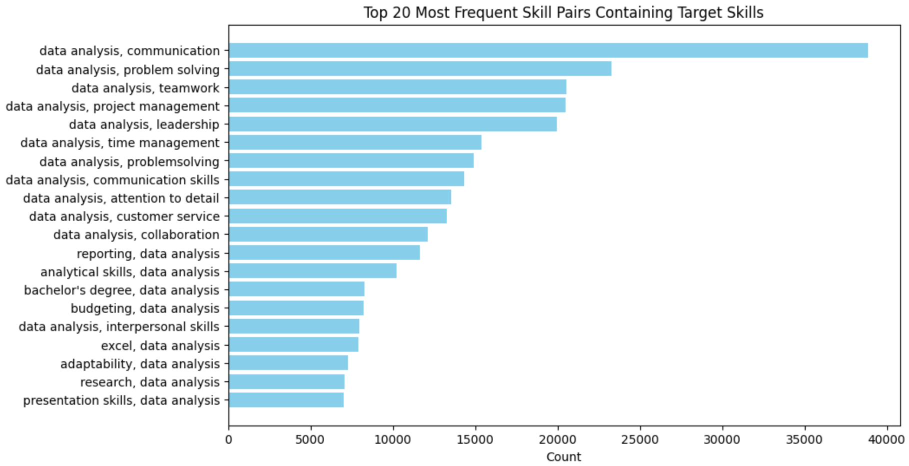

Figure 10: Top 20 Most Frequent Data Related Pairs

The overview of pairs provided significant insights into the most common skill combinations in the field of data analytics. The top pairs highlight how skills pertaining to the social sphere are sought after: it may indicate the relevancy of analyzing data and communicating the outcomes to external non-technical stakeholders. These pairs underline the multidisciplinary nature of roles in data analytics, where technical skills are equally as important as soft skills such as leadership and problem-solving abilities. Data Analytics offerings are also often paired with the requirement of having a Bachelor's Degree, defining how a higher education is necessary for the more technical skills.

4.2 Focus on Quintuplets

Figure 11: Top 20 Most Frequent Data Related Quintuplets

The exploration of quintuplets provide a more generic overview of the Data Analytics postings, once again underlining how a specific set of social transversal skills is relevant, perhaps for better communication with customers and stake holders. A focus on project management and leadership is also evident, highlighting how the ability of managing large scale projects is sought after by employees.

I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work, and including any code produced using generative AI systems. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.