



Data Science and Economics
Department of Economics, Management and Quantitative Methods
Department of Computer Science "Giovanni degli Antoni"
Università degli Studi di Milano

ALGORITHMS FOR MASSIVE DATA MODULE

MARKET BASKET ANALYSIS ON LINKEDIN JOB SKILLS

Soudabeh Masoudisoltani - 961506

JUNE 2024

Declaration:

“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 offenses 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.”

Link to the repository:

Soudabeh Masoudisoltani:

[https://github.com/pnightsore/Market Basket Analysis](https://github.com/pnightsore/Market_Basket_Analysis)

Table of Contents:

Abstract:	4
Chapter 1: Introduction	5
Chapter 2: Algorithm components and some essential theories	6
2.1. Market Basket Analysis	6
2.2. Frequent Pattern Growth algorithm	7
Chapter 2: Dataset	8
2.1. Data	8
2.2. Data pre-processing	9
Chapter 3: experimental methods	10
3.1. Sample Frequency	10
3.2. Sample Association rules	11
3.3. Experimenting with different supports and confidences value	11
3.4. Frequency and association rules for the full data	12
3.4.1 Association rules visualization for full data	14
Chapter 4: Experimenting with new values for support and confidence	15
4.1. The new association rules visualization	16
Chapter 5: Conclusion	17

Abstract:

The aim of this project is to study the co-occurrences of frequent items, also called Market Basket Analysis through a system implementation.

This project considers the study of LinkedIn job skills, in particular job skills as items and strings contained in the 'job_skills' column of our dataset as baskets through the application of Frequent Pattern Growth Algorithm to identify frequent itemsets and association rules.

Chapter 1: Introduction

one needs to be constantly in touch with the new skills required in the job market as well as updating oneself considering skills that are required in concurrence with each other.

In summary, we would like to explore the landscape of job skills that are published on LinkedIn. For this aim, we will be using Market Basket Analysis, which is a powerful data mining tool used for discovering the interesting hidden relationships in large datasets.

Traditionally, MBA is commonly used in retail and considers every transaction made by consumers. It helps identify items that are bought together to help the business owner make informed decisions about different situations.

Here the same concept is innovatively applied to LinkedIn job skills data.

In this context, we define our ‘basket’ as the set of skills possessed by an individual and ‘market’ as the entire LinkedIn user base.

Our aim is to uncover possible hidden associations and patterns between different skills.

This project is able to answer many questions and to shed light on some interesting curiosities. For example, primarily we want to identify similar and frequent items, i.e., job skills that tend to appear together on LinkedIn profiles. Job seekers, recruiters and researchers can gain valuable insights from this study.

Job seekers and newly graduates can use the understanding of skills associations we gain from the results of this study to effectively plan their skill building path and their marketability.

HR professionals can take advantage of these results to help identify skill sets that define successful recruits.

Our goal is to be able to contribute to more informed career decision making and strategy setting by shedding light on the complex interplay of skills in the job market.

We are living in a world where professional skills required to start a career rapidly change and



Chapter 2: Algorithm components and some essential theories























2.1. Market Basket Analysis

Also called association analysis, was traditionally used for improving the decisions made by retailers in a way to increase their profit. In the context of the retailer, this method answers many questions such as:

- How likely is it that a customer buys product A if they have already product B in their shopping cart? Answering the question “which products and goods are often bought together?”

To answer this question, a list of all past purchases is required to understand which products were bought together. Each row of the list should include one customer's shopping cart at one time and each row, as we will also use the same terminology here, is called a “transaction”.

Figure 2.1. shopping cart transactions example

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 

In order to be able to interpret our technical results later, we need to explain a few concepts that are essential for evaluating the strength and significance of the associations briefly before moving forward to the experiment chapter:

1. Frequency

How often a product A and B occur together in a transaction.

2. Support

Support tells us how frequently an itemset appears in the dataset. In other words it is the proportion of transactions that contain a particular item or itemset.

$$\text{Support}(A) = \frac{\text{Number of transactions containing } A}{\text{Total number of transactions}}$$

3. Confidence

Confidence tells us if product A is already purchased, how likely it is that product B is then also in the shopping cart in the traditional example. Calculated by dividing the number of transactions that contain both items by transactions that contain product A.

$$\text{Support}(A) = \frac{\text{Number of transactions containing } A}{\text{Total number of transactions}}$$

4. Lift

It indicates how much more likely it is that item B is purchased when item A is already purchased compared to the chance of purchasing item B independently.

By taking advantage of these metrics, businesses can identify strong associations among products and make informed decisions, for example product placement and promotions. For example if bread and butter appear together in transactions with high confidence and lift, it is a smart choice to place them near each other to encourage combined purchases.

2.2. Frequent Pattern Growth algorithm

This algorithm is a popular method for association rule learning in transaction databases.

This algorithm works by compressing the dataset into a structure called FP-tree.

The itemsets information is kept in FP-tree. To reveal frequent items, the algorithm mines the FP-tree. FP Growth is scalable, making it able to handle large datasets efficiently.

As a solution to scaling up, Using PySpark primitives, every algorithm and solution used in this project have been implemented. This is especially true for loading data, exploring, and putting methods into practice. This allows for the execution of the resulting operation and inspection without using up all of the system RAM, even when the dataset size is changed.

Chapter 2: Dataset

2.1. Data

To conduct this Market Basket Analysis research and to study the association rules in finding the frequent items tasks, we have used an available dataset from [LinedIn jobs & skills](#)

LinkedIn is a popular platform for professional career networking. It is home to millions of job advertisements.

This particular dataset includes 1.3 Million job listings which were gatherers from LinkedIn in 2024.

The dataset has 2 columns, the first one shows the link to the job listing and the second column consists of several job skills separated from each other by a comma.

Here is a random visualization for one row of the dataset:

Figure 2.1. Random visualization of the LinkedIn job skills

https://www.linkedin.com/jobs/view/%E2%80%8Bassistant-general-manager-buffalo-wild-wings-beckley-at-...	Management experience, Teamwork, Communication, Customer service, Problemsolving, Food safety, Resta...
https://www.linkedin.com/jobs/view/administrative-officer-at-first-division-consulting-inc-379987056...	Administrative support, Document management, Process improvement, Action tracking, SharePoint, Teams...
https://www.linkedin.com/jobs/view/housekeeper-part-time-flexible-daytime-hours-at-northbridge-compa...	Housecleaning, Cleaning Protocols, Laundry, Linen Care, Trash Removal, Maintenance Reporting, Proble...
https://www.linkedin.com/jobs/view/full-time-assistant-manager-at-universidade-de-mogi-das-cruzes-36...	Cinemark Assistant Manager, Management Foundations, Annual Certifications, MAP Core, Safe Alcohol Se...

2.2. Data pre-processing

The pre-processing of the data for this project is rather brief, we have used the spark library to be able to deal with our large scale dataset.

Apache Spark is a powerful open source library that is helpful for analyzing large scale data processing. It helps data mining through its key features. In our case, we are going to utilize Apache Spark for finding the association rules by using the F-P Growth algorithm to find frequent itemsets and generate association rules from large datasets.

- We have checked for possible duplicates, and none was found.
- Then the dataset was checked for null values and has considered those without Null values from study
- Finally a new column, “skills”, has been added to the dataset that consists of all the skills as strings split into lists. Below is a representation of one of skills’ cells:

[Building Custodial Services, Cleaning, Janitorial Services, Materials Handling, Housekeeping, Sanitation, Waste Management, Floor Maintenance, Equipment Maintenance, Safety Protocols, Communication Skills, Attention to Detail, Physical Strength, Experience in Housekeeping]

Chapter 3: experimental methods

To conduct this study, we performed the task first on a sample of 5% of the dataset with minimum support 0.01 and minimum confidence 0.5 and then on the whole data.

In our preprocessing section, we first removed null values and duplicates and finally created 3 columns:

- Job link: containing the link to the job advertisement.
- Job skills: all skills required by a job separated by comma.
- Skills: a new column that consists of arrays of job skills as strings in each row.

Before applying the FP Growth algorithm, stop words, such as ‘are’, ‘is’, ‘and’ are removed and a new column “transactions” has been created that contains the arrays of skills without stopwords.

Finally our model is made by fitting an FP Growth algorithm with minimum support 0.005 and minimum confidence 0.4 on the transactions column. By setting the thresholds we ensure that the rules generated will have a strong implication and are useful for decision making.

3.1. Sample Frequency

After having fit the fp growth algorithm on the sample data, we can observe the frequency of items in descending order and itemsets and the association rules.

From the results of the sample, we can say that the highest frequency among all skills belong to “Communication” followed by “Teamwork” and “Leadership” for both.

Figure 3.1. Item Frequency table for sample

items	freq
[Communication]	18165
[Teamwork]	11315
[Leadership]	8949
[Customer service]	8269
[Teamwork, Communication]	6974
[Communication skills]	5826
[Leadership, Communication]	5697
[Customer Service]	5448
[Problem Solving]	5002
[Sales]	4651

3.2. Sample Association rules

Association rules help identify the relationships between items in our dataset.

The table below shows the association rules table for the sample.

Lets look at rule 1 as an example:

- The 0.11 support shows that the itemset [Time management, Attention to detail] and item [Teamwork] were together in 11% of transactions.
- The 0.59 confidence shows that the 59% of the transactions that contain the itemset [Time management, Attention to detail] also contain item [Teamwork].
- There is a 40% increase in the likelihood of requiring [Teamwork] as a skill by a recruiter when itemset [Time management, Attention to detail] is already required.

Figure 3.5. First few Association rule for the sample

antecedent	consequent	confidence	lift	support
[Time management, Attention to detail]	[Teamwork]	0.598014888337469	3.410138280176625	0.01120530663484339
[Time management, Attention to detail]	[Communication]	0.6153846153846154	2.1858773211374367	0.011530771972784898
[Problem solving, Communication]	[Customer service]	0.5062362435803375	3.9501609801105473	0.010693861103792446
[Collaboration]	[Communication]	0.584237165582068	2.075240001918622	0.03756799900810564
[Collaboration, Leadership]	[Communication]	0.7486818980667839	2.6593560203117583	0.013204593710769803
[Time management, Customer service]	[Teamwork]	0.5497076023391813	3.1346693438560314	0.013111603614215086
[Time management, Customer service]	[Communication]	0.6705653021442495	2.38188191523553	0.01599429660741131
[Coaching]	[Leadership]	0.5179220779220779	3.7342592729652733	0.015451854377508796
[Coaching]	[Communication]	0.52	1.847066336361134	0.01551384777521194
[Attention to Detail, Problem Solving]	[Teamwork]	0.549831081081081	3.1353734727878564	0.010089425476186785

3.3. Experimenting with different supports and confidences value

To have a better balance, the number of rules generated are very important for us, as we wouldn't like to have to handle a large number of rules and prefer to use a sensible technique that delivers us just enough useful information without being overwhelming or cluttered.

So we have checked for some combination of support and confidence and we have chosen a minimum support of 0.005 and a minimum confidence of 0.4 producing fewer than 1000 rules that appear to be effective

Figure 3.6. Experimenting different metrics values

```
Support: 0.001, Confidence: 0.2, Number of Rules: 160891
Support: 0.001, Confidence: 0.4, Number of Rules: 146906
Support: 0.001, Confidence: 0.6, Number of Rules: 132452
Support: 0.005, Confidence: 0.2, Number of Rules: 1572
Support: 0.005, Confidence: 0.4, Number of Rules: 979
Support: 0.005, Confidence: 0.6, Number of Rules: 485
Support: 0.01, Confidence: 0.2, Number of Rules: 376
Support: 0.01, Confidence: 0.4, Number of Rules: 222
Support: 0.01, Confidence: 0.6, Number of Rules: 100
```

3.4. Frequency and association rules for the full data

The same pre-processing techniques and modifications as the sample data have been performed on the complete dataset. Followed by implementing the minimum support of 0.005 and minimum confidence of 0.4 to derive the association rules. Interesting to mention that same as before the most frequent items/itemsets taking all the data into consideration are [communication], [Teamwork] and [Leadership]

Figure 3.2. Itemset distribution among top 10 itemsets

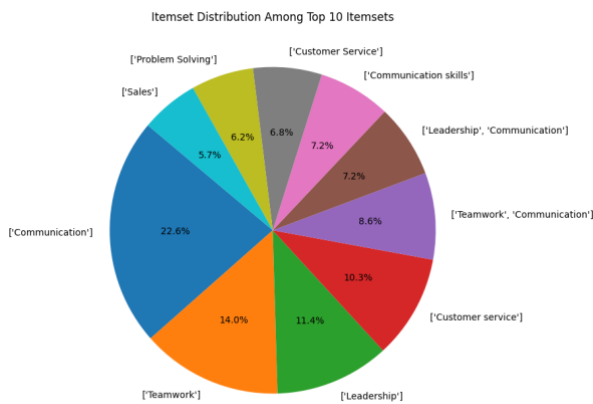


Figure 3.3. Word Cloud of itemsets



Figure 3.4. Top 10 most frequent itemsets

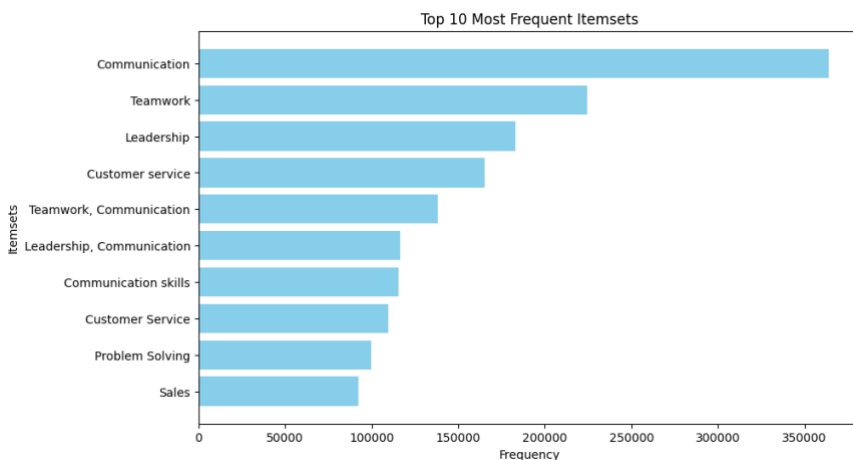


Figure 3.5. Full data Frequency table

items	freq
[Communication]	364027
[Teamwork]	224629
[Leadership]	182796
[Customer service]	165403
[Teamwork, Communication]	138347
[Leadership, Communication]	116574
[Communication skills]	115642
[Customer Service]	109616
[Problem Solving]	99866
[Sales]	92324

Together, these visualizations draw attention to the essential abilities that predominate in the LinkedIn Jobs & abilities dataset's job.

The most commonly mentioned skill is "communication," as the bar and pie charts both plainly demonstrate, highlighting its crucial significance in the workplace. Similarly, "Leadership" and "Teamwork" also show up frequently, suggesting that the ability to collaborate and manage others is highly regarded.

Figure 3.7. First few Association rules with min support 0.005 and min confidence 0.4 on all items

antecedent	consequent	confidence	lift	support
[Time Management, Leadership, Communication]	[Teamwork]	0.5580324691058881	3.2155363695981585	0.007116953832508996
[Time Management, Leadership, Communication]	[Customer Service]	0.4614732251029804	5.449194864522014	0.005885470505433515
[Time Management, Leadership, Communication]	[Problem Solving]	0.6312696874242791	8.181954522961908	0.008050996079958343
[Planning, Leadership]	[Communication]	0.7403269754768392	2.632387126657798	0.006297252571513334
[Dental benefits]	[Vision benefits]	0.8986114749803511	167.1417343463453	0.005299859236974785
[Dental benefits]	[Medical benefits]	0.9276919046371496	146.31171943131545	0.0054713707166553095
[Walking, Standing, Lifting]	[Bending]	0.6223194372962773	43.60494231207338	0.005605026058928872
[Time management, Problemsolving, Teamwork, Communication]	[Leadership]	0.44158706833210876	3.1268672180206623	0.0051074882530087904
[Time management, Problemsolving, Teamwork, Communication]	[Customer service]	0.5530024714447933	4.327563713922256	0.006396142073311114
[Time management, Problemsolving, Teamwork, Communication]	[Attention to detail]	0.4887449068198517	8.462220765938612	0.005652925661362172
[Conflict resolution, Communication]	[Leadership]	0.5847723704866562	4.140758836496943	0.005180110230891535
[Conflict resolution, Communication]	[Customer service]	0.5664573521716378	4.432855926191252	0.005017869642004552

The goal here is to extract meaningful and generalizable insights from the data.

To achieve the goal, we would want to capture a wide variety of patterns and to focus on those that are most significant, we have done this by setting the mentioned values for minconfidence and minsupport

In other words, a strong likelihood of requiring a particular skill when one other skill is already among the required skills is desirable. The high confidence values averaging around 61% evidence the meaningfulness of our associations.

For our future work, depending on what we need, we can adjust our approach by:

- Lowering minSupport and minConfidence to see a broad variety of patterns.
- Increase minSupport and minConfidence to focus on more insightful patterns.

Figure 3.8. Distribution of Confidence and Support in association rule

summary	confidence	summary	support
count	984	count	984
mean	0.613020902091151	mean	0.009203508816095815
stddev	0.14453877254360148	stddev	0.007128414414189697
min	0.4006851571249753	min	0.005000873008883059
max	0.9915927136851939	max	0.10688332738451174

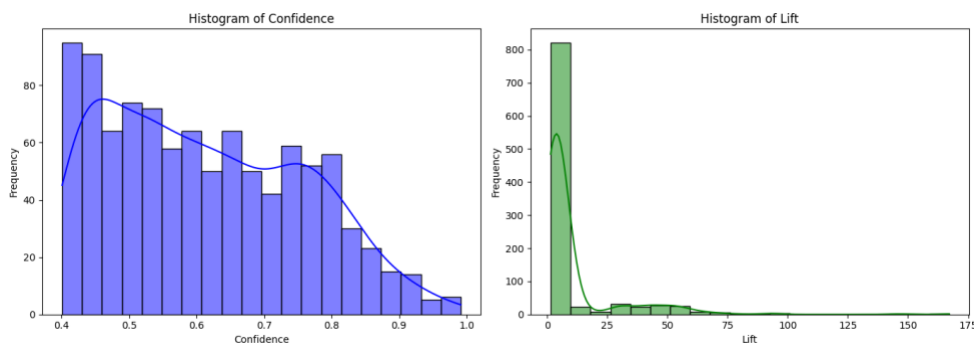
3.4.1 Association rules visualization for full data

The majority of the rules, according to the confidence histogram below, have a confidence level between 0.4 and 0.9, however they are primarily concentrated in the lower end of this range. That being said, very high confidence levels are quite uncommon, even if many rules have some degree of reliability.

The lift values primarily cluster at the lower end of the lift histogram, with many of them being quite near to zero. This suggests that many Association Rules aren't all that different from what would occur by chance. It is rare to find truly strong and meaningful rules with higher lift values.

The variation in confidence and lift in particular indicates that the most effective outcomes may be obtained by selecting the appropriate rules, that is, those with both high confidence and very high lift. Furthermore, the variations in support level.

Figure 3.9. Histograms of Confidence and Lift



Chapter 4: Experimenting with new values for support and confidence

To solve the problem from previous part, and to gain more specific association rules that are more insightful instead of general ones we have performed the project again, having increase both metrics to $\text{minSupport} = 0.01$ and $\text{minConfidence} = 0.6$

These new values give us 106 Association rules which seems to be just about enough for making meaningful decisions.

Figure 4.1. first few rows of new Association rules

antecedent	consequent	confidence	lift	support
[Time management, Attention to detail]	[Communication]	0.6237104961519567	2.2177328872479043	0.01177094101086703
[Time management, Customer service]	[Communication]	0.6816374195177377	2.423704157248919	0.01668528570567703
[Collaboration, Leadership]	[Communication]	0.7640491254186825	2.7167361834827686	0.014274854099356136
[Project Management, Problem Solving]	[Communication]	0.6699751861042184	2.382236651507887	0.01001256205702525
[Attention to Detail, Problem Solving]	[Communication]	0.7097155897392554	2.523541953627503	0.012974611665561886
[Teamwork]	[Communication]	0.6158910915331502	2.189929361591667	0.10688332738451174

The resulting collection of rules and itemsets is narrowly focused but insightful thanks to the parameters of $\text{minSupport} = 0.01$ and $\text{minConfidence} = 0.6$. These configurations appear to offer a decent trade-off between guaranteeing the association rules' dependability and relevance.

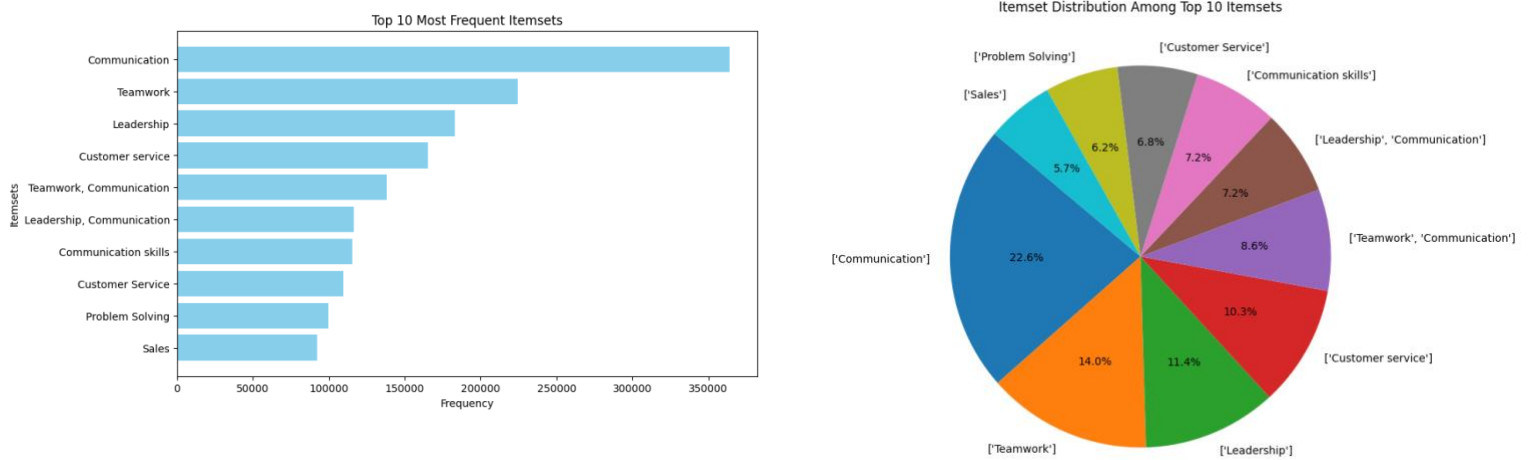
The rules are fairly dependable, as indicated by their average confidence of 0.719. This high score indicates that there's a good probability that another particular skill will also be required if one skill is indicated.

A high lift indicates that the correlation between talents in a rule is significantly stronger than that which would result from chance alone, even though we don't have exact statistics for every lift value. High lift rules are quite helpful since they demonstrate a strong connection between the talents.

Since the average support value is approximately 0.017, these rules are not limited to unusual or exceptional cases; rather, they are based on patterns that occur frequently enough to be dependable and helpful in a variety of contexts.

The frequency tables and graphs below show that the top most frequent itemsets have not change after changing the metrics:

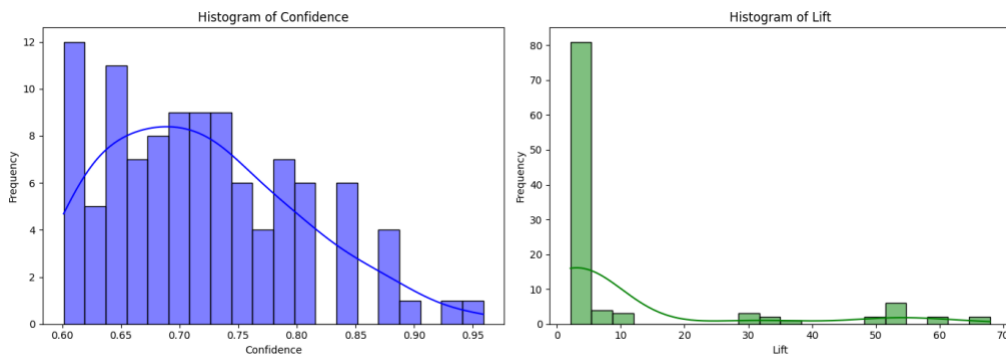
Figure 4.2. Top 10 most frequent itemsets



4.1. The new association rules visualization

Many of the rules are just at the threshold, as indicated by the confidence histogram's peak near the selected minimum. This suggests that tightening the confidence might exclude less certain but potentially helpful rules. The lift histogram is biased towards lower values, indicating the remarkable value of the few rules that exhibit unusually high lift.

Figure 4.3. Confidence and Lift histogram



Chapter 5: Conclusion

The aim of this study is to implement a system of finding frequent itemsets.

The study of finding frequent item sets has been performed on a LinkedIn jobs dataset from Kaggle that consists of about 1.3 Million job links and their respective skills required.

We have observed that among all skills required by recruiters on LinkedIn, 'Communication', 'Teamwork' and 'Leadership' seem to be the most highly demanded among all the others.

Then a Basket creation and definition is performed for running the Frequent Pattern Growth Algorithm.

We then defined both the most frequent items as well as the association rules among them using the FP Growth Algorithm that resulted in very interesting findings such as:

If the itemset [Time management, Attention to details] is required by a recruiter, there is a 62% likelihood that [Communication] is also required. And similarly for other skills.

Finally, we applied a comparison to understand how the distribution of Confidence and Support changes while setting different values for minSupport and minConfidence and how the number of rules generated also change. This observation allowed us to conclude that using the metric values that generate just the right number of rules is preferred since it provides us with fewer yet more meaningful and specific rules that can guide us in future decision making.

This study and similar ones, can be a very useful tool both for recruiters and job seekers.

Job seekers can understand which skills are highly in demand in the job market, how to plan their resume, and which skills are worth developing in oneself to secure a proper job after graduation.

Moreover, Recruiters can learn which skills have proven to be more useful in employees and which skills are required by more experienced businesses.