­

**Data Mining II**

**PA3 – Association Rules and Lift Analysis**

Jason Willis

College of Information Technology,

Western Governors University

Dr. Kesselly Kamara

July 6, 2022

Table of Contents for Each Rubric

Part I: Research Question

[Describe Purpose, Summarize Research Question and Define Objectives:](#RubricA) 3

***Part II: Market Basket Justification***

Explain Market Basket Analysis, Provide an Example & Summarize Assumptions4

***Part III: Data Preparation***

Transform Data, Generate Association Rules 6

Provide Support, Lift and Confidence Values, Identify Top Three Rules 6

Part IV: Data Summary and Implications

Summarize Support, Lift and Confidence Significance, Discuss findings, Provide Recommend Course of Action 8

Part V: Attachments

[Sources for Third-Party Code](#RubricG) 9

[Sources](#RubricH) 10

# Hospital Readmission Problem

For our chain of hospitals to lower readmission concerns, we need to identify patients who have increased risk of rehospitalization within a month of their release. According to Schuller (2020), non-obese adults were 21% less likely to be readmitted than obese adults. A readmission study by Gert, et. al. (2002) showed a correlation between longer initial hospital stays and readmission. Within the provided dataset, I’m leveraging these studies to help create my hypothetical question and shape my approach in finding potential patient groups with a statistically significant chance for readmission outcomes.

After viewing the provided medical\_clean.csv data set and accompanying data dictionary, there seems to be some patient groupings which are aligned with the research mentioned above. For instance, the following patient data fields: Initial patient admin days, Total Charges, and Initial Says (inpatient) both caught my attention and were underscored by the research mentioned above. While my initial feelings towards these variables might make them feel related, are they?

## A1 – Proposal of Question

Can we find the associations between medications that are frequently prescribed in our dataset?

**A2 – Defined Goal**

The goal of our analysis is to logically investigate the provided patient medication data set and, by leveraging market basket analysis techniques, understand the probability when a medication (*A*) is prescribed, then different, mutually exclusive medication is also prescribed (*B*), e.g. if *A* then *B. (A => B)*

**B1 – Explanation of Market Basket**

According to Larose (2019) “Association rules seek to uncover associations among the variables and take the form ‘If antecedent, then *consequent*,’ along with a measure of the support and confidence associated with the rule.” Given the dataset features in our scenario, this isn’t just a factorial problem of *n!* features due to the added dimensionality of various feature responses. As the number of data attributes grow, so would the rules associated. Enter Market Basket Analysis, which provides a technique to identify attribute set frequency. The probability of a medication (consequent) given an initial medication (antecedent) provides a measure of “confidence” while “lift” provides a measure of association strength between the antecedent and consequent. These “techniques” provide a means to construct useful recommendations based on findings, (Hull, 2022).

**B2 – Transaction Example**

Listed in Figure 1 – List of First Transaction, we can observe a python list that is sliced on row 1, displaying the first transaction. For instance, “Amlodipine” and “Albuterol Aerosol” are a subset within the first transaction record and could possibly have a complimentary relationship. Here it a complete transaction record below:  
Patient Transaction for Record 1 – Amlodipine, Albuterol Aerosol, Allopurinol, Pantoprazole, Lorazepam, Omeprazole, Mometasone, Fluconozole, Gabapentin, Pravastatin, Cialis, Losartan, Metoprolol Succinate XL, Sulfamethoxazole, Abilify, Spironolactone, Albuterol HFA, Levofloxacin, Promethazine, and Glipizide.

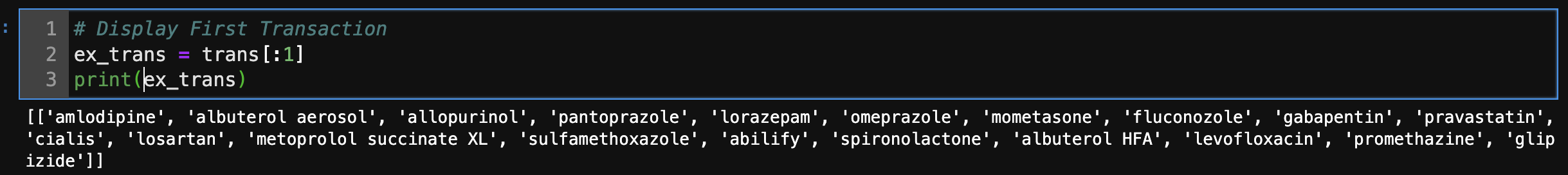


Figure - List of First Transaction

Graphical user interface, text, application

Description automatically generated

Figure 2 - Association Rules

**B3 – Market Basket Assumption**

Market Basket Analysis assumes there are complimentary relationships between associated items. Meaning, transactions (medications) have relationships between items; therefore, being prescribed certain meds directly leads to being prescribed other meds. This assumption isn’t always the case though. For example, while certain medications could be frequently prescribed together; they may not have a complementary relationship. They could be mutually exclusive while also being prescribed frequently, which may give an impression of association.

**C1** – **Transforming the Dataset**

The data set is transformed for market basket analysis and a cleaned version of the data frame is provided as: “cleaned\_df.csv”. (Figure 3)

Text

Description automatically generated

Figure 3 - Transform and Cleaned

**C2 – Code Execution**

The notebook provides code which executes to generate association rules with the Apriori algorithm. (Figure 2)

**C4 – Association Rules Table**

The submission includes a screenshot and accurately identifies the top 3 rules generated by the Apriori algorithm along with their summaries.

**C4 – Top Three Rules**

The data set accurately identifies the top 3 rules. (Figure 4)

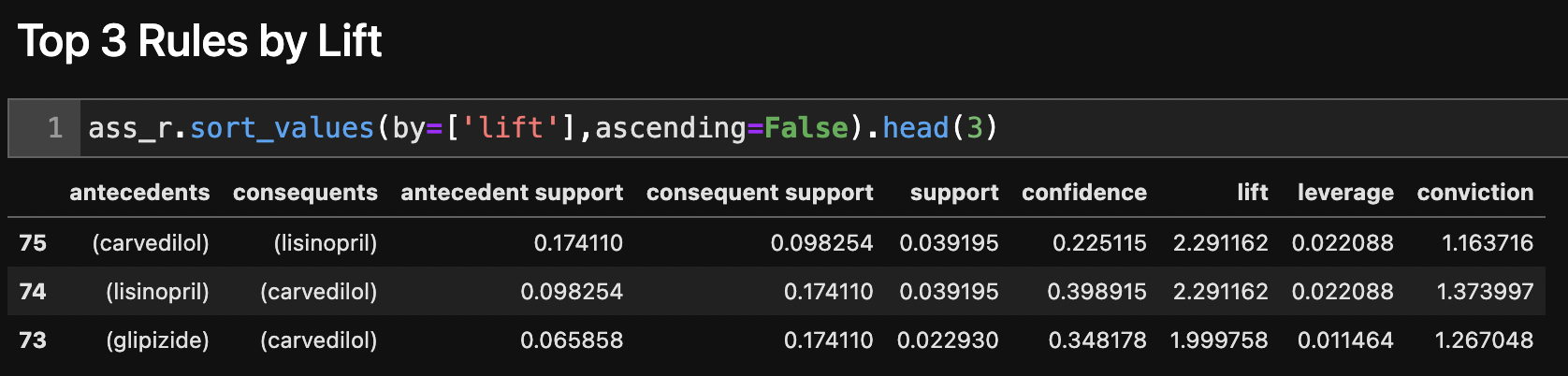


Figure 4 - Top 3 Rules

**D1 – Significance of Support, Lift and Confidence Summary**

The metrics used by the Apriori algorithm are:

* Support: The support column seen in Figure 2 provides a frequency value for a medication within our dataset.
  + Support Metric:
* Confidence: This column measures the association value if another medication is prescribed.
  + Confidence Metric:
* Lift: This column measures the level of importance for the specific rule, between zero and infinity.
  + Lift Metric:

**D2 – Practical Significance of Findings**

By filtering the overall data frame by ideal metric values, the final pruned list has 9 rules to focus our attention on. (Figure 5) The list was pruned by only returning rules which have support levels greater than 0.03, confidence levels greater than 0.2 and lift levels greater than 1.5.

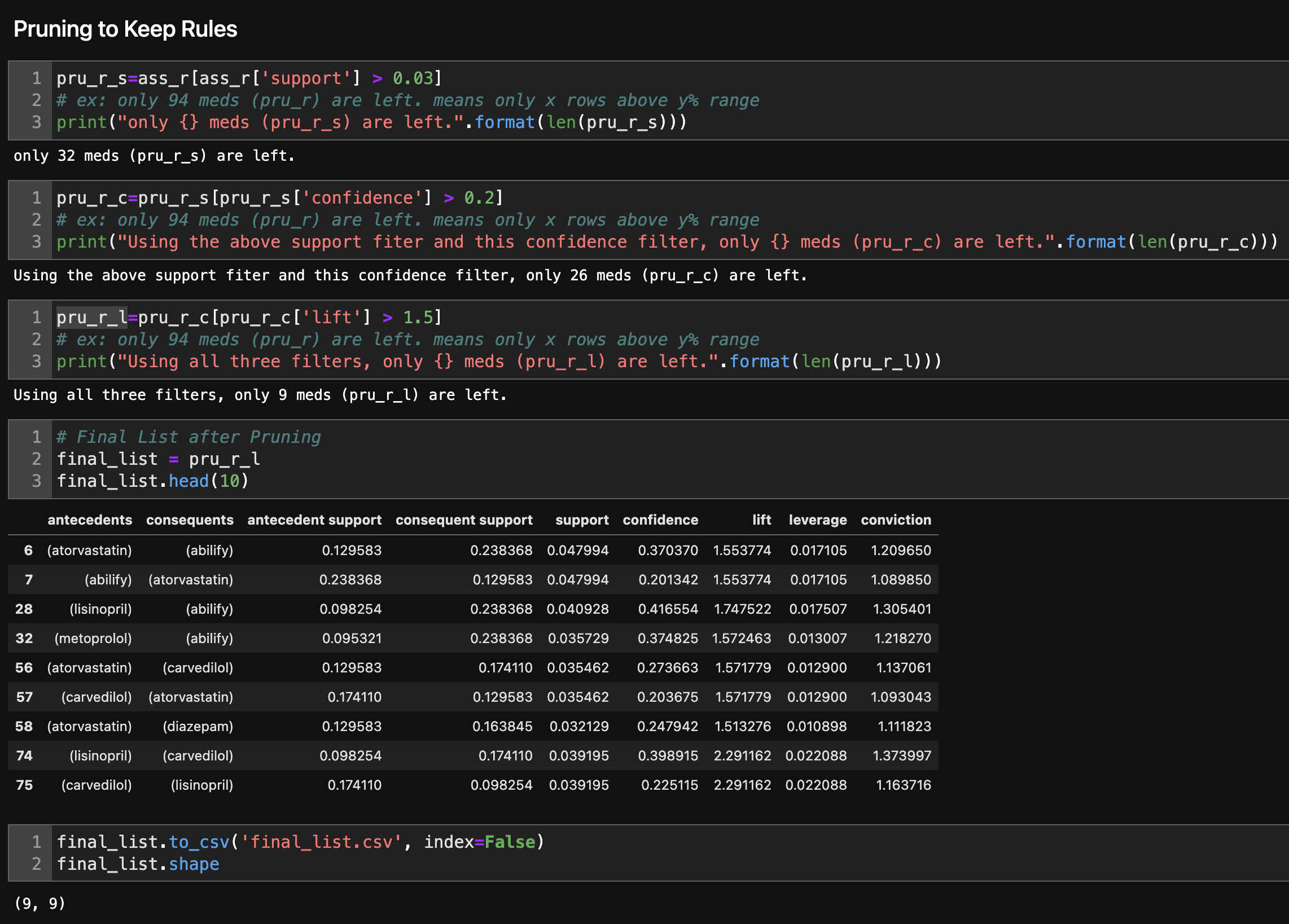


Figure 5 - Pruned List

**D3 – Course of Action**

We provided the following question in A1: “*From information about previous patients who were readmitted, can we ascertain the probability of certain medications (consequents) given a medication (antecedent) for our patients?”* Displayed in *Figure 2 – Association Rules*, we do indeed see a list of medications. Provided in the columns are confidence metrics which give values based on the association of a consequent given an antecedent. Furthermore, from the data analysis provided in *Figure 5 – Pruned List*, we can take this reduced list of 9 prescription data sets to focus on first. This list has the highest frequency of medications (support), the highest association values of consequents given antecedents (confidence) and the highest overall importance for this specific rules.

**E – Panopto Recording**

Panopto Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=5ae19ec5-0816-4a59-95d5-aec7004ba420>

**F – Web Sources**

* Help using Markdown: <https://www.markdownguide.org/basic-syntax/>
* Help to see ALL columns: <https://stackoverflow.com/questions/24524104/pandas-describe-is-not-returning-summary-of-all-columns>
* Help to create a better histogram design: <https://mode.com/example-gallery/python_histogram/>
* Matplotlib Help: <https://matplotlib.org/2.1.2/api/_as_gen/matplotlib.pyplot.plot.html>
* Multiple ways to conduct ANOVA: <https://www.marsja.se/four-ways-to-conduct-one-way-anovas-using-python/>
* Numpy Help: <https://numpy.org/doc/stable/>
* Pandas Help: <https://pandas.pydata.org/docs/user_guide/index.html#user-guide>
* Python Help: <https://docs.python.org/3.9/library/index.html>
* Scipy.stats Help: <https://docs.scipy.org/doc/scipy/reference/tutorial/stats.html>

## References

Gert P Westert, Ronald J Lagoe, Ilmo Keskimäki, Alastair Leyland, Mark Murphy,

An international study of hospital readmissions and related utilization in Europe and the USA, Health Policy, Volume 61, Issue 3, 2002, Pages 269-278, ISSN 0168-8510, <https://doi.org/10.1016/S0168-8510(01)00236-6>. (<https://www.sciencedirect.com/science/article/pii/S0168851001002366>)

Kamara, K. Market Basket Analysis - Data Mining II Lecture WGU. <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=5aefcf2b-73cd-41f5-b4d3-aea0011efd05>

Larose, D., C., & Larose, D., T. (2019). Data Science Using Python and R. Wiley. <https://www.wiley.com/en-us/Data+Science+Using+Python+and+R-p-9781119526810>

Hull, I. (2022) Market Basket Analysis in Python. DataCamp. Found Here: <https://campus.datacamp.com/courses/market-basket-analysis-in-python/>

Schuller K. A. (2020). Is obesity a risk factor for readmission after acute myocardial infarction?  *Journal of healthcare quality research*, *35*(1), 4–11. <https://doi.org/10.1016/j.jhqr.2019.09.002>