CIS105 Final Project Report: **Investigating Walmart shopping trips with market-basket analysis**
James Wang | Dec 19th, 2016

**Goal:**

Deep dive into Walmart's shopping trip data to uncover insights on how item purchases vary by day, department, trip type, as well as products that are most frequently bought together.

**Data Description:**

The dataset contains 647054 observations and 7 features. Each instance represents an item purchased as part of a trip to the store. After data pre-processing (removing returned items), our dataset contains 631596 observations. There are 100102 unique items in the dataset, across 68 departments, grouped into 38 different types of trips.

The features are:

- TripType: Walmart has identified 38 types of trips that represent the reasons as to why customers go to Walmart
- VisitNumber: An ID that corresponds to a single trip by a single customer.
    - We can use VisitNumber to create "groups" of items purchased on a single trip.
- Weekday: The day of the week of the trip. Encoded as 1 = Monday, 2 = Tuesday, …, 7 = Sunday.
- UPC: "Universal Product Code" – the barcode of an item. This can be used to identify unique items purchased within a single trip by a single customer.
- ScanCount: The quantity of an item purchased on a single trip by a single customer. -1 represents a returned item.
- DepartmentDescription: The department from which the item purchased/returned was from.
- FinelineNumber: A more refined category for each of the items, created by Walmart. Not sure how useful this will be.

**Exploratory Data Analysis:**

What types of trips are the most frequent occurring?

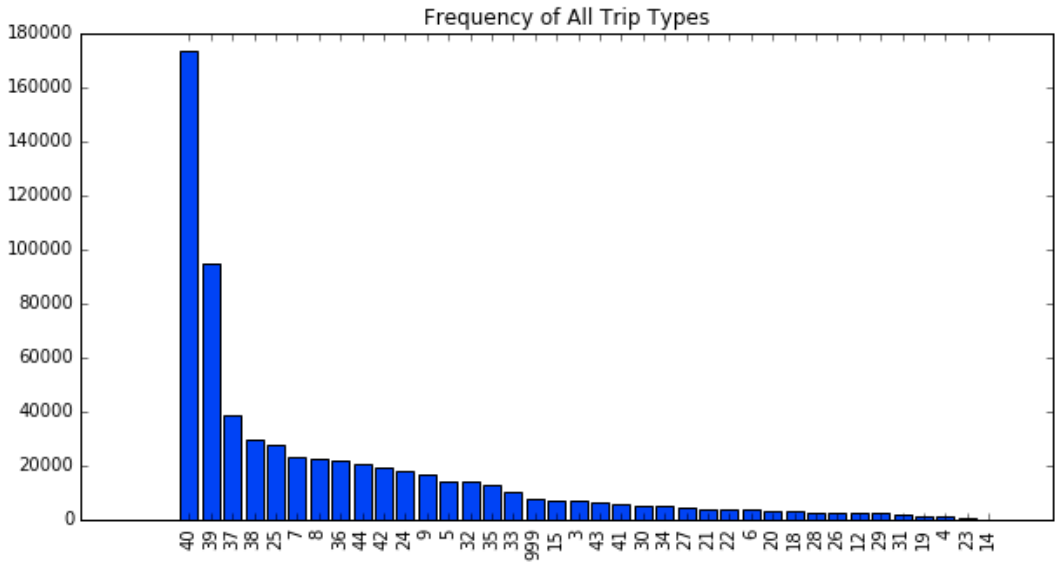


Figure 1.

*There are 38 unique trip types. While we do not know what Walmart labels these trip types, we can see that the majority of trips fall into category 40 and 39, with over 170000 trips and 90000 trips, respectively.*

## How many trips occur on different days of the week?

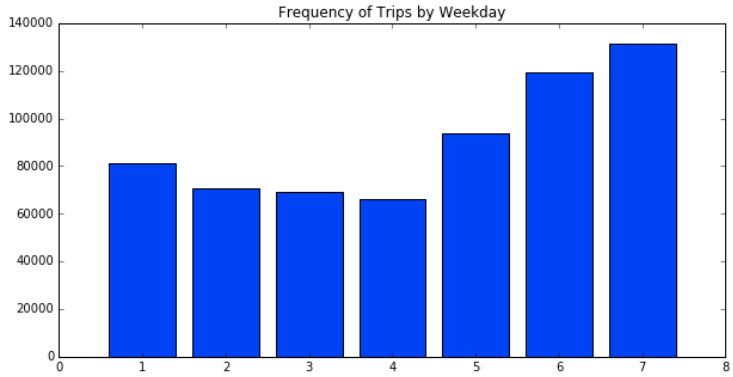


Figure 2.

*These results seem reasonable – most of the shopping occurs on Friday and the weekend (days 5, 6, and 7) when customers are off work.*

## What departments are shopped in the most?

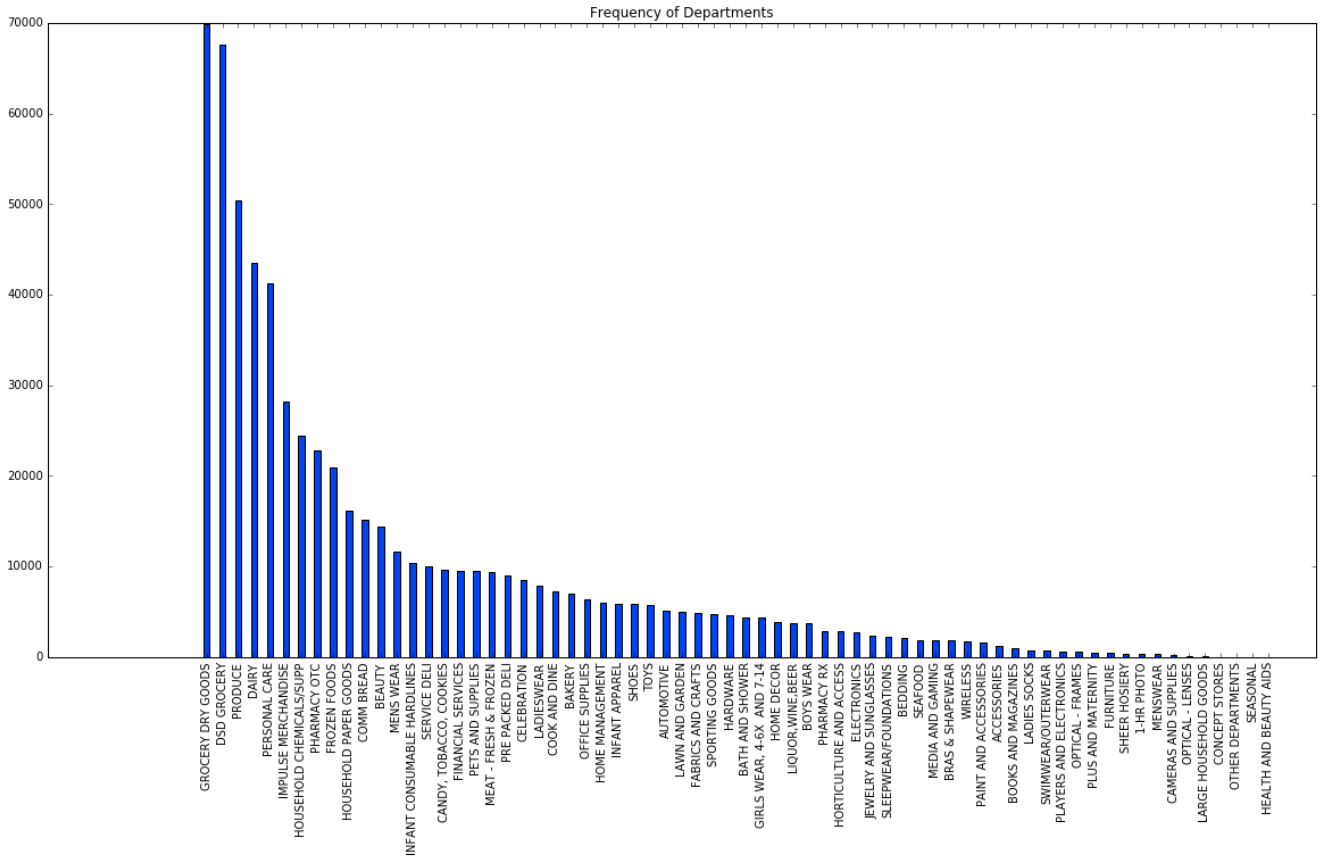


Figure 3.

*Groceries (including direct store delivery groceries) are the departments with the most frequently bought items during customer shopping trips. The next 2 items are also food-related: produce and diary. The data seems to suggest that most shopping trips to Walmart are utilized to buy food. From a personal point of view, this is surprising because when I was growing up, my family never went to Walmart for our grocery shopping. Also, to my surprise, impulse merchandise (items positioned near the checkout counters) are the 6$^{th}$ highest category of items bought. I never imagined people bought those items, but apparently, they do!*

*We can also turn our attention to the bottom of the spectrum – amongst the lowest shopped departments are "health and beauty aids", "seasonal", "large household goods", and "men's wear". One may be quick to assume that if Walmart had to cut some departments, they should start here. However, we must not ignore the interdependencies of certain departments. For example, while "seasonal" does not generate a lot of shopping trips, we cannot ignore the possibility that customers may come to Walmart to get "seasonal" items as their primary intention, and then buy other goods (such as groceries) out of convenience since they are already*

*there. We can further explore these interdependencies and co-occurrences of different items through the market-basket analysis below. It's also interesting to note that "men's wear" is one of the lowest-shopped departments, while "ladies wear" is in the upper half of the distribution. This may be reflective of Walmart's customer demographics.*

## By type of trip, what are the most frequently occurring week days (top plot) and departments (bottom plot)? (Selected Results)

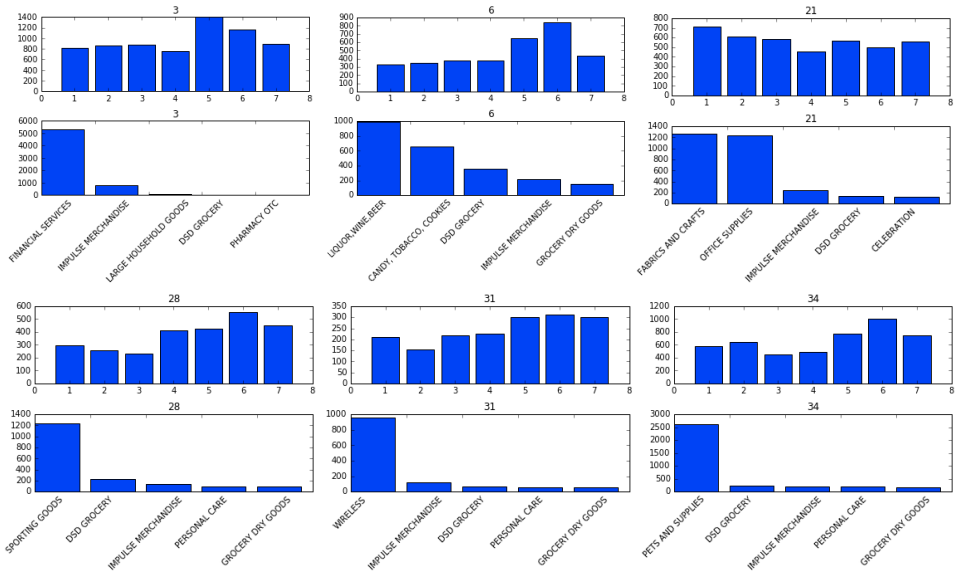


Figure 4.

*It is clear that certain items are assigned to certain trip types: financial services, sporting goods, wireless, and pets and supplies make up the vast majority of their respective trip types of 3, 28, 31, and 34. The distribution of shopping frequency across day of the week for each trip type generally seems to follow a similar shape to the aggregate distribution (Figure 2) in the sense that there is a dip in the middle of the week and peaks towards the beginning and end of the week. One can also see which trip type occurs the most on which day of week, which may be able to help Walmart structure their product positioning and promotions given a particular day of week.*

## What trip types do not vary by weekday, as measured by total variation distance?

| Most Stable Trip Types by Weekday | | | | |
|---|---|---|---|---|
| Trip Type | 7 | 5 | 21 | 8 |
| TVD measure | 0.04182005 | 0.04534496 | 0.048529729 | 0.048901637 |

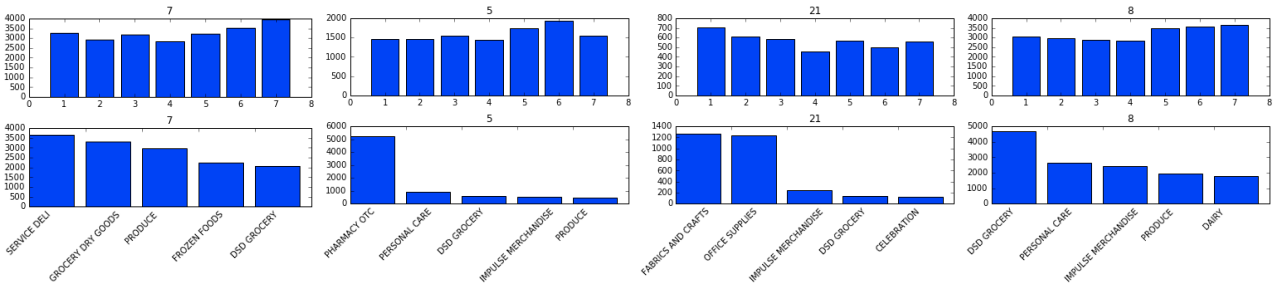


Figure 5.

*These trip types have weekday shopping frequency distributions that are most similar to a uniform distribution (i.e. each day has a shopping frequency of 14.285% of the total number of shopping trips for a given trip type). One can infer that the items bought during these trip types are the "most stable" with respect to day of the week, and have the same demand every day / bought in the same quantity every day. These items include those from the following departments: service deli, groceries, pharmacy, fabrics and crafts, office supplies, and personal care.*

| Least Stable Trip Types by Weekday | | | | |
|---|---|---|---|---|
| Trip Type | 41 | 18 | 23 | 12 |
| TVD measure | 0.181782293 | 0.208102055 | 0.281619654 | 0.291818119 |

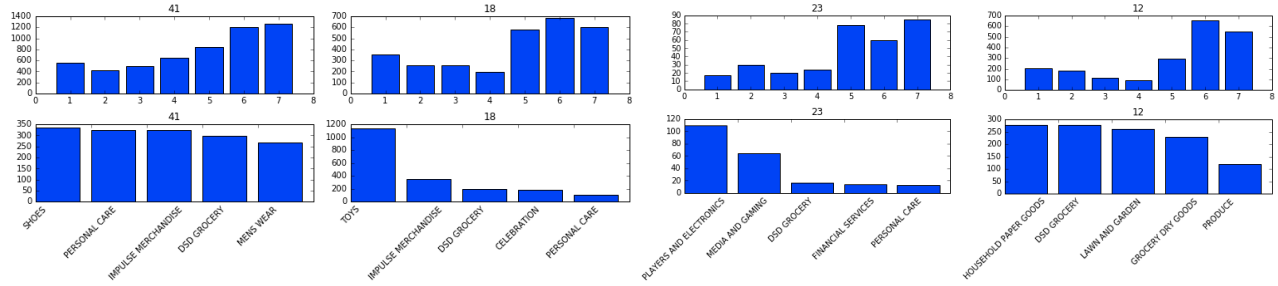


Figure 6.

*These trip types vary in shopping frequency the most from day to day, and seem to only spike up during Fridays and the weekend. These items include those from the following departments: shoes, personal care, impulse merchandise, toys, players and electronics / media and gaming, and lawn and garden. This means that if Walmart wants to increase sales for a particular department via trip type, then it should bolster promotions and positioning of these products to the weekend for the trip types above.*

## By day of week, what are the most frequently occurring trip types (top plot) and departments (bottom plot)?

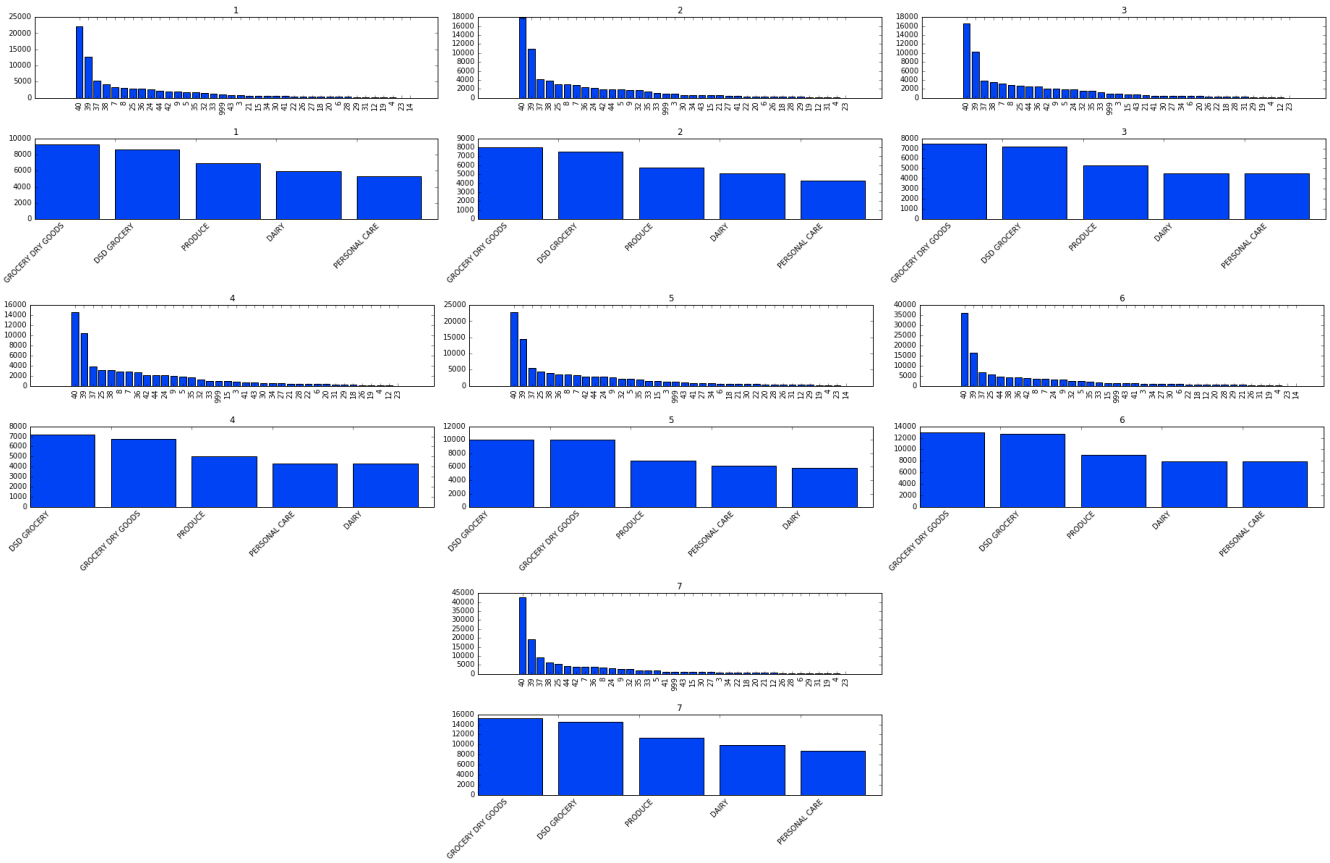


Figure 7.

*Looks like grocery and food-related items are the most bought items **every** day of the week! This is confirmed by both the trip type plots and the department plots. It seems on a day-to-day basis, the distribution of unique items bought is relatively the same. This means there's less of an obvious strategy for Walmart when it comes to, for example, increasing revenue from optimizing item placement based on different days of the week. The weak signal here will need to be teased out by other methods – hopefully basket analysis.*

**Market-Basket Analysis:**

*Introduction*

The purpose of market-basket analysis is to investigate what items are purchased frequently together. The methodology is based off probabilities and counts – which is good because it is easy to do, and bad because learning on the joint probability distribution is highly susceptible to overfitting (which can lead to bad predictions). However, when the dataset is large enough, the sample is arguably very representative of the population and has little noise (at least, from sampling) to overfit on. For complete *Methodology*, please see Appendix.

*Results*

**Top Results by Confidence, Interest, and Combined (Confidence + Interest): Items represented as UPC #**

| | Rule | **Confidence** | Rule | **Interest** | Rule | **Combined** |
|---|---|---|---|---|---|---|
| 1 | Given 83032400641.0, then 60538819035.0. | Confidence: 1.0 | Given 68113107941.0, then 60538807733.0. | Interest: 0.999911418052 | Given 60538807733.0, then 68113107941.0. | Combined: 1.99991141805 |
| 2 | Given 68113107939.0, then 60538887953.0. | Confidence: 1.0 | Given 60538807733.0, then 68113107941.0. | Interest: 0.999911418052 | Given 68113107941.0, then 60538807733.0. | Combined: 1.99991141805 |
| 3 | Given 60538896309.0, then 68113163351.0. | Confidence: 1.0 | Given 68113178252.0, then 68113178251.0. | Interest: 0.999726201252 | Given 60538807734.0, then 68113111868.0. | Combined: 1.99972620125 |
| | *26 rules meet the confidence threshold. 18 rules have confidence = 1.* | | *24 rules meet the interest threshold. 2 rules that met the confidence threshold did not meet interest threshold.* | | *Same 24 rules as those that meet the interest threshold.* | |

Figure 8.

**Unique Rules in the Top Combined Results: Replacing Item's UPC # with Item's Department**

| Rule | Combined | Frequency (within top 24 by Combined) |
|---|---|---|
| Given FINANCIAL SERVICES, then FINANCIAL SERVICES. | Combined: 1.99991141805 | 14 times |
| Given LARGE HOUSEHOLD GOODS, then IMPULSE MERCHANDISE. | Combined: 1.99963117698 | 1 time |
| Given WIRELESS, then WIRELESS. | Combined: 1.99945723425 | 2 times |
| Given IMPULSE MERCHANDISE, then FINANCIAL SERVICES. | Combined: 1.9989643965 | 3 times |
| Given PRODUCE, DAIRY, and COMM BREAD, then PRODUCE. | Combined: 1.88018993107 | 1 time |
| Given PRODUCE, PRODUCE, PRODUCE, and PRODUCE, then PRODUCE. | Combined: 1.8449040531 | 1 time |
| Given PRODUCE, PRODUCE, COMM BREAD, and DAIRY, then PRODUCE. | Combined: 1.84399083604 | 1 time |
| Given GROCERY DRY GOODS and DAIRY, then DAIRY. | Combined: 1.81159398739 | 1 time |

Figure 9.

*Discussion*

The first thing that jumped out at me was that amongst the top 24 rules (by combined confidence and interest measures), 14 of them are items relating to the financial services department! This means that if a customer comes into Walmart and purchases some items and one of those items is from the financial services department, then there is a high probability that another item they purchased is *also* from the financial services department! However, what might be more interesting is the items that are bought together which are not in the same department. A buyer of an item from large household goods will most likely buy an item from impulse merchandise (given they buy more items). This is a rather interesting phenomena of user behavior that should be explored further. Impulse merchandise purchases show up again – they seem to be quite predictive of financial services. And lastly, the other association rules are more intuitive. Buying multiple wireless items together makes sense, since there are usually accessories. Buying your groceries and food together also makes sense.

**Concluding Thoughts:**

One of the more surprising revelations from this project is the most of information that can be extracted from simply plotting the data. While I was doing this project, I had not even begun the market-basket analysis yet and I felt I had already extracted a great deal of insight from the data. Lastly, it was interesting to see that the hardest part about market-basket analysis is not the actual algorithm / problem itself, but more so a programming challenge of how to efficiently and effectively use Python's data structures and other algorithms to do the counts over the baskets. I was using the *time* package quite often to time my code and find faster implementations.

Appendix:

*Methodology*

1. Definitions:
    a. Basket: Group of items bought together on a single trip
    b. Itemset: Group or subgroup of items that are bought together
    c. Rule: "Given an itemset contains *items*, then it will most likely also contain *other item*."
2. Search dataset of baskets using A-Priori algorithm
3. Find itemsets with size up to 10 items per itemset
    a. Result: No itemsets of size greater than 5.
4. Determine a support threshold
    a. Support of itemset *i*: The number of baskets containing an itemset *i*
    b. Support Threshold = 50
        i. Result: There are about 1500 itemsets with a support greater than 50.
5. Find frequent itemsets using the Apriori algorithm
    a. Easiest algorithm to implement given computational time tradeoff
6. Calculate confidence, interest, and combined confidence and interest of each rule
    a. Confidence of rule *r*: Probability of item *i* given an itemset *j*.
        i. Confidence Threshold = 0.9
    b. Interest of rule *r*: Confidence – Probability of item *i* across all baskets
        i. Purpose: item *i* may be purchased frequently regardless of itemset *j* (i.e. item *i* is independent of itemset *j*), therefore a high confidence for rule *a* will be misleading.
        ii. Interest Threshold = 0.9

*Sources:*


- http://infolab.stanford.edu/~ullman/mmds/ch6.pdf
- http://mmds.org/mmds/v2.1/ch06-assocrules.pdf
- https://github.com/PacktPublishing/Learning-Data-Mining-with-Python