Data Mining

Lecture-0

Course Introduction and Motivation

Dr. Salem Othman

Summer 2023

Topics to Be Covered

- About Me
- **Syllabus**: Required books, Grading, Schedule*, and more*. Please read it in detail and ask me if you have question.
- Required Books
- Data Sources
- Study Groups

About Me

Instructor: Prof. Salem Othman

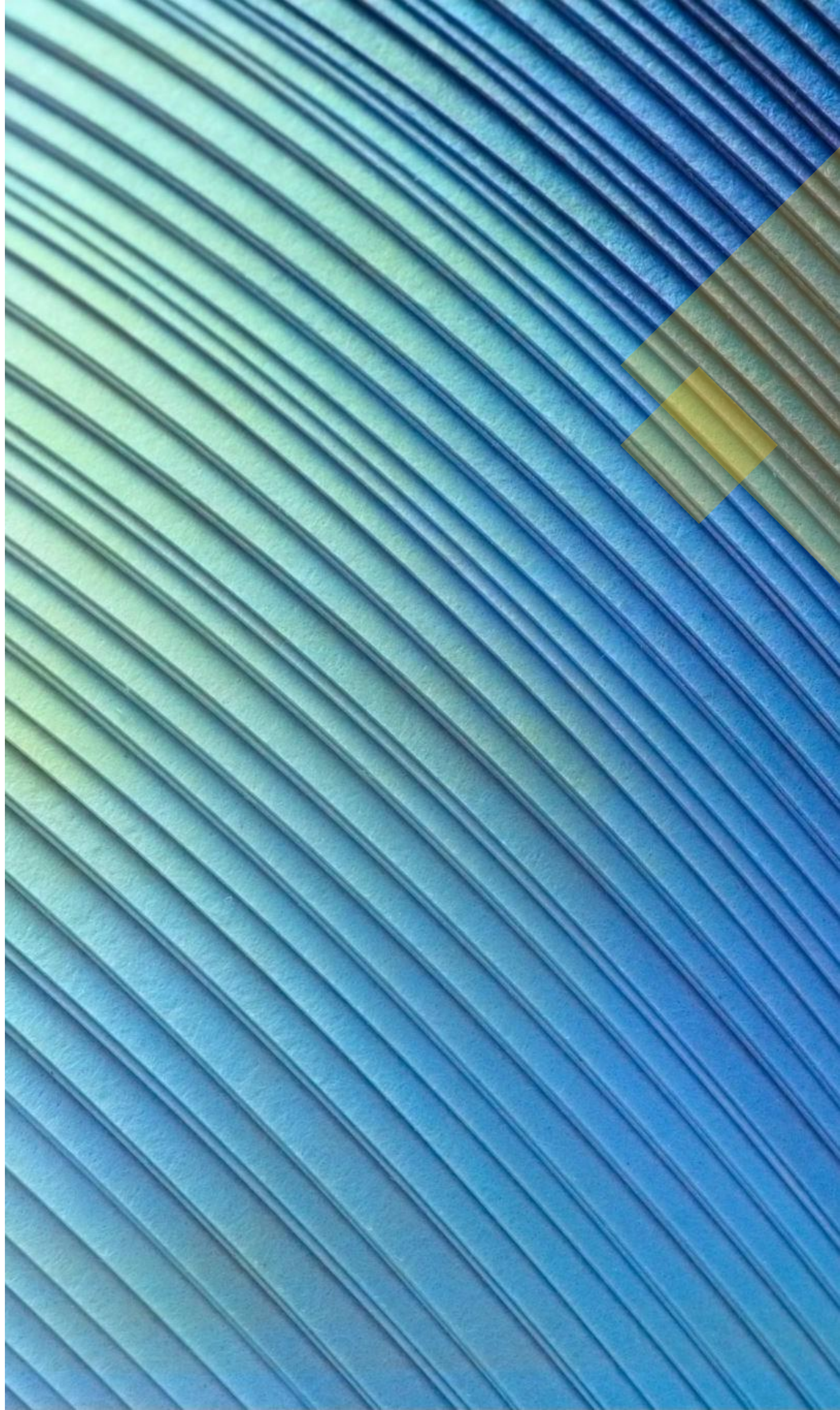
Lectures: T 5:00 PM- 6:20 PM

Office: Williston Hall - 110

Office Hours: By Appointment

Telephone: 617-989-4508 OR Zoom

Email: othmans1@wit.edu



Zoom Link

- Othman, Salem is inviting you to a scheduled Zoom meeting.
- Topic: Salem Othman's Zoom Meeting
- Time: This is a recurring meeting Meet anytime
- Join Zoom Meeting
- Join Zoom Meeting
- https://wentworth.zoom.us/j/345383795?pwd=cEUzOXk4V3ZnM2FNVTZ6UVJBTllsZz09

- Meeting ID: 345 383 795
- Passcode: 002109

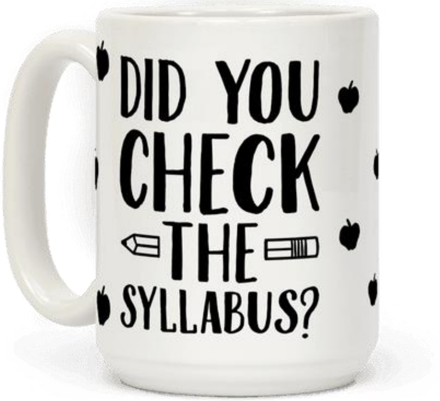
- One tap mobile
- +13052241968,,345383795# US
- +13092053325,,345383795# US

- Dial by your location
- +1 305 224 1968 US
- +1 309 205 3325 US
- +1 312 626 6799 US (Chicago)
- +1 646 558 8656 US (New York)
- +1 646 931 3860 US
- +1 301 715 8592 US (Washington DC)
- +1 346 248 7799 US (Houston)
- +1 360 209 5623 US
- +1 386 347 5053 US
- +1 507 473 4847 US
- +1 564 217 2000 US
- +1 669 444 9171 US
- +1 669 900 9128 US (San Jose)
- +1 689 278 1000 US
- +1 719 359 4580 US
- +1 253 205 0468 US
- +1 253 215 8782 US (Tacoma)
- Meeting ID: 345 383 795
- Find your local number: https://wentworth.zoom.us/u/kexycoCvrE

Syllabus

1. Open Brightspace.
2. Select the Content Area from the Course Menu that holds the syllabus.
3. Click on the syllabus.
4. It is your responsibility to read it.

brightspace
by D2L



What is Data Mining

- Data mining is a process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Wikipedia (https://en.wikipedia.org/wiki/Data_mining)
- Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more. SAS (https://www.sas.com/en_us/insights/analytics/data-mining.html)

Anomalies

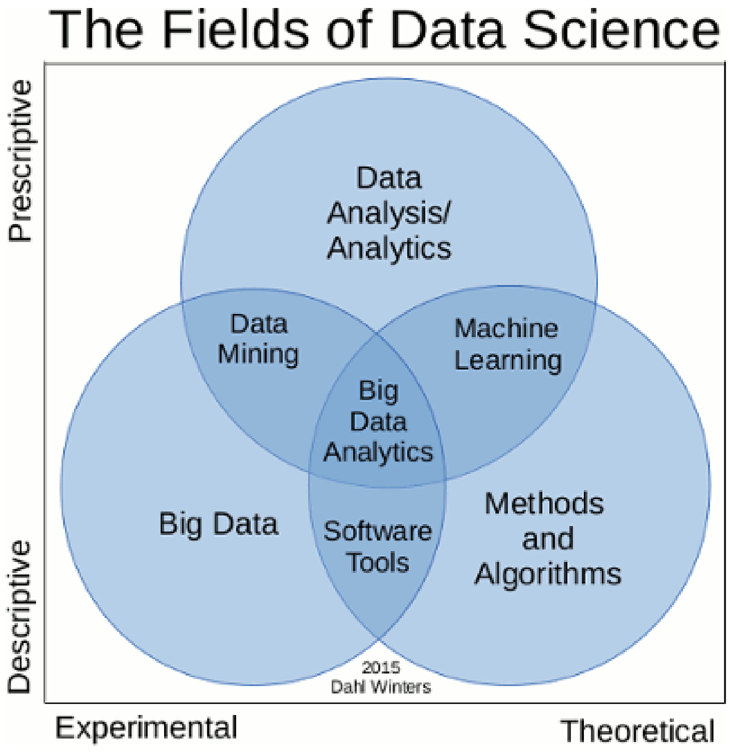
- These are also known as outliers, which are data points that significantly deviate from the rest of the data. They might indicate a problem or error, or they may represent an important, unusual occurrence that deserves further investigation.
- Real-life example: Credit card fraud detection. Here, anomaly detection algorithms are used to identify unusual patterns of transactions, such as unusually high charges or multiple transactions in a short time. These anomalies could indicate fraudulent activity. In this case, the bank could take preventive measures like blocking the card and contacting the customer to verify the transactions.

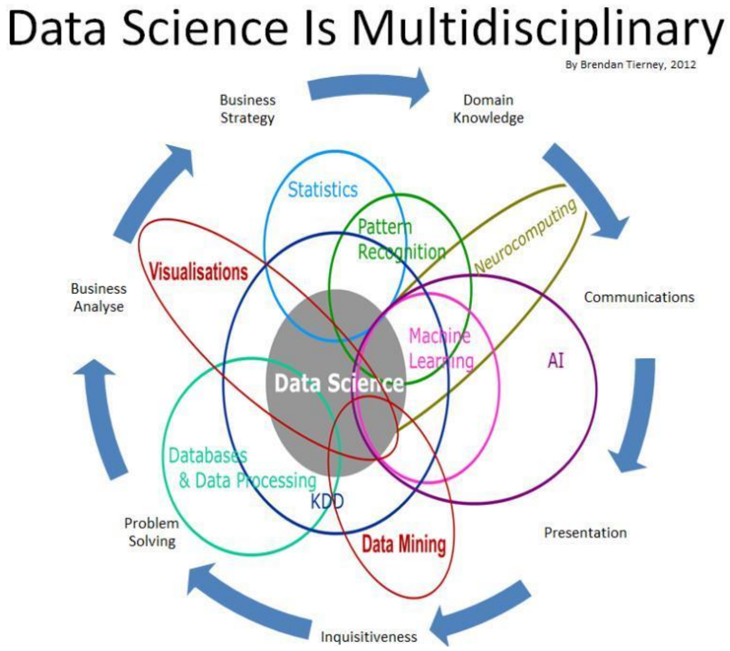
Patterns

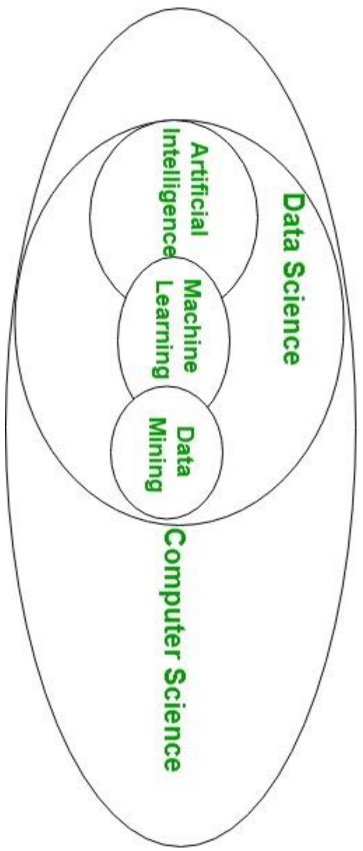
- These are recurring or systematic arrangements in the data that often reveal important relationships between variables.
- Real-life example: Market basket analysis in a supermarket. Patterns in data could reveal that customers who buy pasta also tend to buy pasta sauce and parmesan cheese. The supermarket can use this pattern to place these items near each other or suggest them as a bundle to increase sales.

Correlations

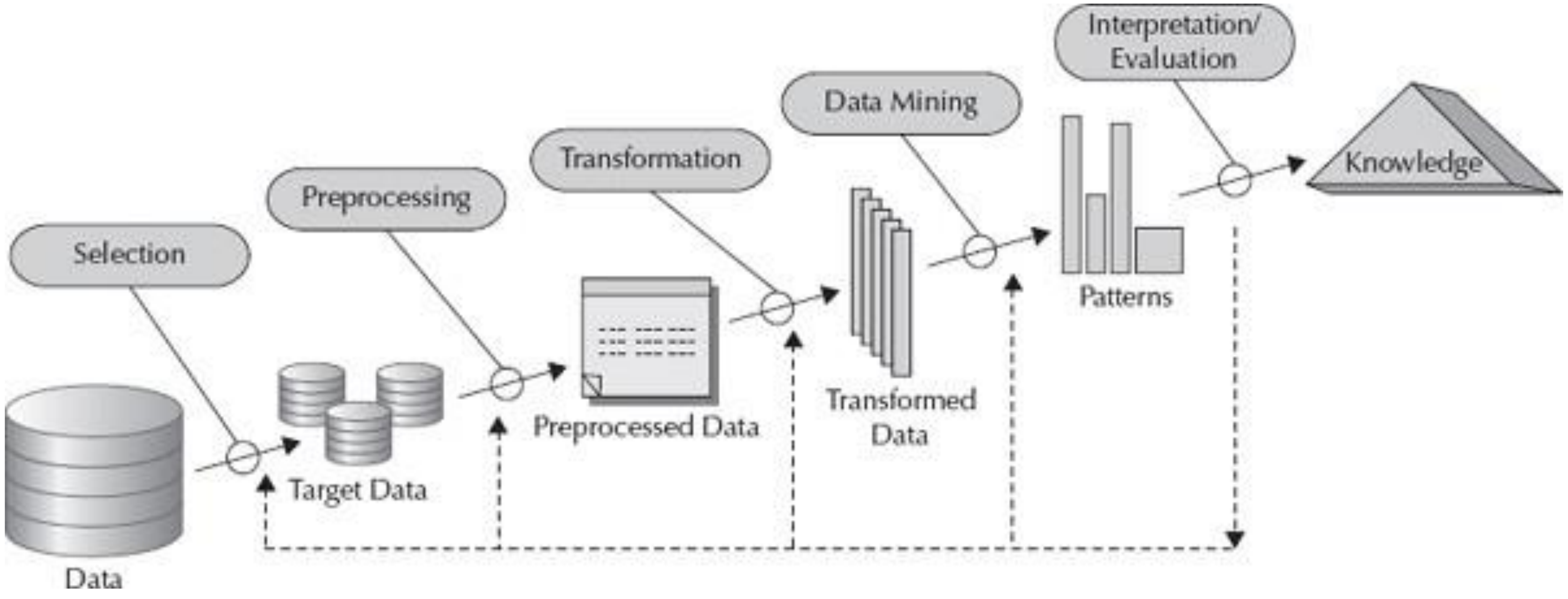
- These are statistical relationships between two or more variables or features in your dataset. A correlation could be positive (both variables increase or decrease together), negative (one variable increases while the other decreases), or neutral (no relationship).
- Real-life example: In healthcare, data mining might reveal a positive correlation between smoking and lung disease. This correlation can help doctors educate patients about the risks of smoking and persuade them to quit or never start smoking.







What is the difference between data mining and data science and machine learning?

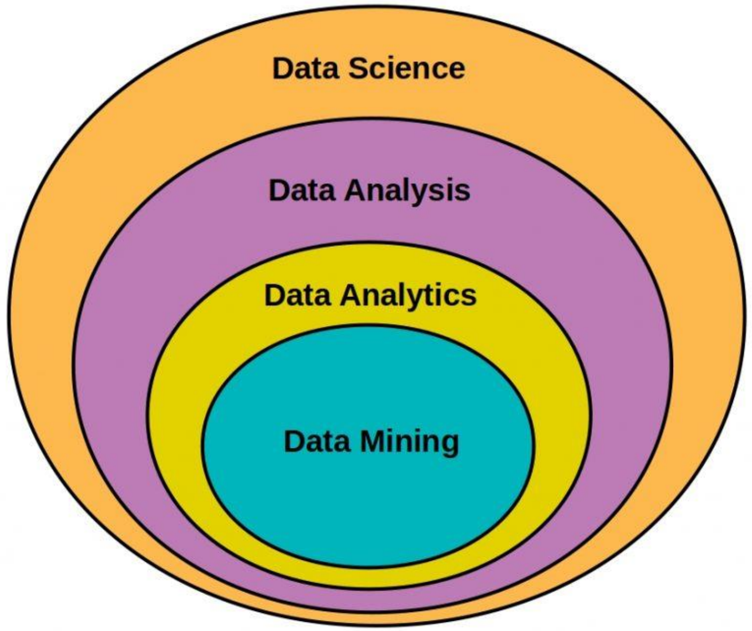


Data Mining

- Data mining is the process of discovering patterns, trends, and insights hidden in large datasets.
- It involves the use of various techniques, such as clustering, classification, association rule mining, and anomaly detection, to find useful information that can help in decision-making.
- Data mining is a subset of knowledge discovery in databases (KDD) and is primarily focused on extracting valuable insights from data.

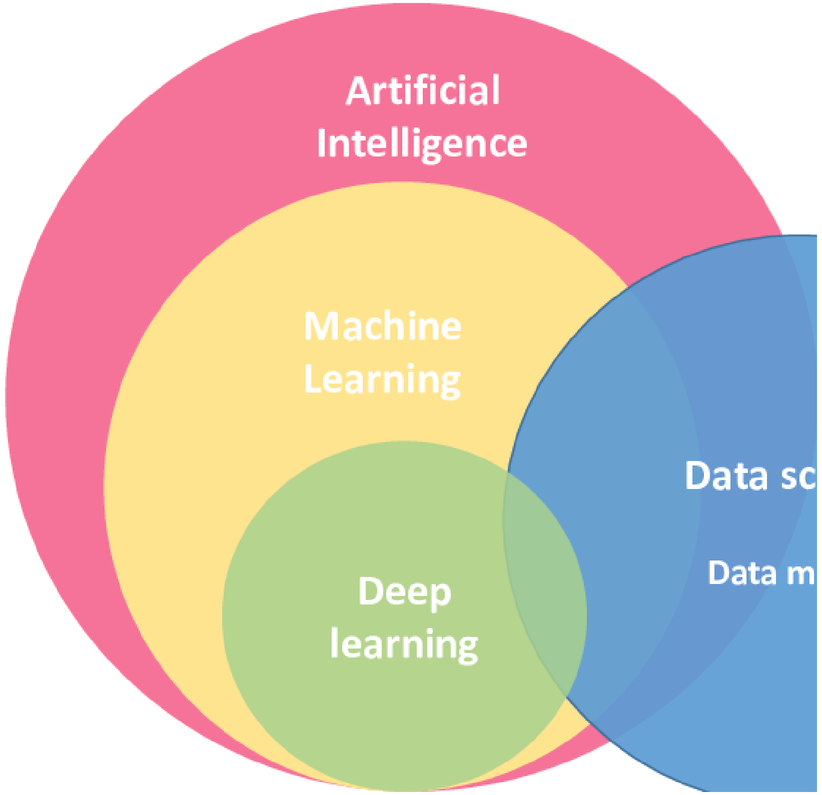
Data Science

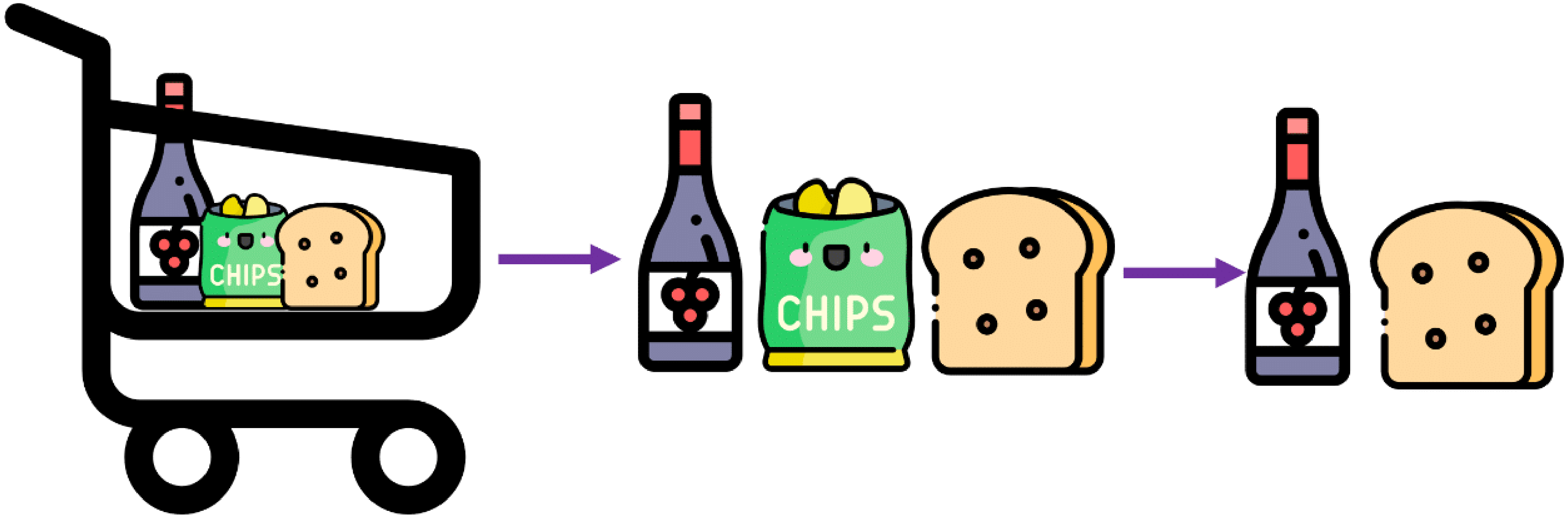
- Data science is an interdisciplinary field that combines techniques from statistics, computer science, and domain-specific knowledge to extract insights and knowledge from structured and unstructured data.
- It involves data collection, preprocessing, exploration, analysis, modeling, and visualization.
- Data mining is a part of data science, which also encompasses other techniques such as statistical modeling, machine learning, and big data processing.



Machine Learning

- Machine learning is a subfield of artificial intelligence (AI) and computer science that focuses on developing algorithms that can learn from and make predictions or decisions based on data.
- It involves creating models that can generalize patterns found in data to make predictions on unseen data.
- Machine learning can be applied in various tasks, such as image recognition, natural language processing, and recommendation systems.
- Machine learning is a key component of data science, and some data mining techniques are based on machine learning algorithms.

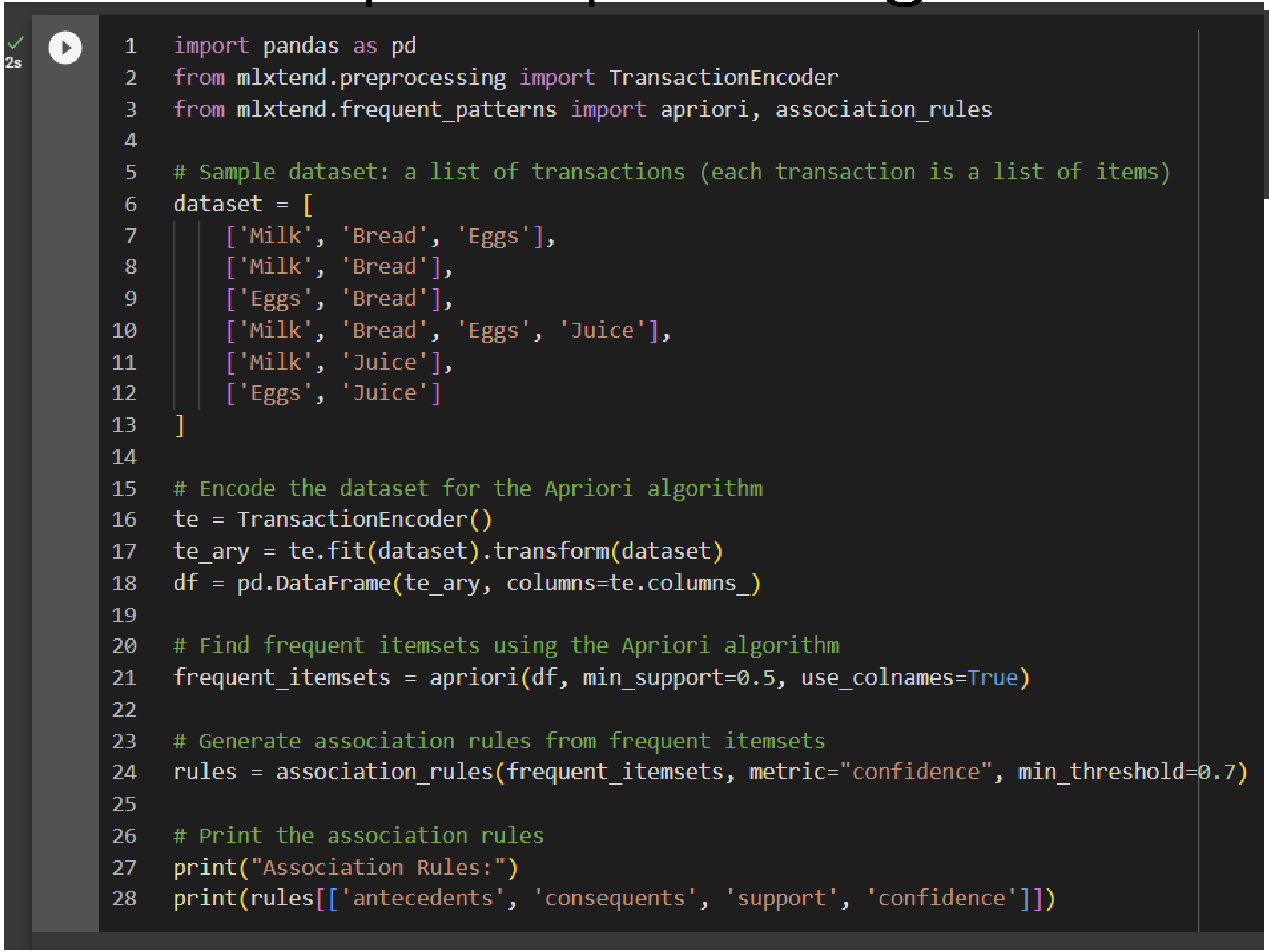




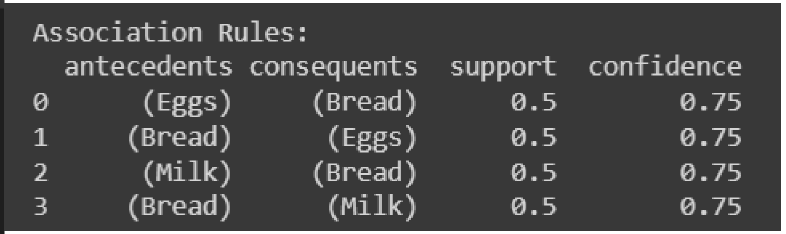
Example: Apriori algorithm

The following Python code demonstrates a simple example of data mining using the Apriori algorithm for discovering frequent itemsets and association rules in a dataset. This example uses the mlxtend library, which you can install using pip install mlxtend.

Example: Apriori algorithm



```python
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules

# Sample dataset: a list of transactions (each transaction is a list of items)
dataset = [
    ['Milk', 'Bread', 'Eggs'],
    ['Milk', 'Bread'],
    ['Eggs', 'Bread'],
    ['Milk', 'Bread', 'Eggs', 'Juice'],
    ['Milk', 'Juice'],
    ['Eggs', 'Juice']
]

# Encode the dataset for the Apriori algorithm
te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_ary, columns=te.columns_)

# Find frequent itemsets using the Apriori algorithm
frequent_itemsets = apriori(df, min_support=0.5, use_colnames=True)

# Generate association rules from frequent itemsets
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.7)

# Print the association rules
print("Association Rules:")
print(rules[['antecedents', 'consequents', 'support', 'confidence']])
```



```
Association Rules:
  antecedents consequents  support  confidence
0      (Eggs)     (Bread)      0.5        0.75
1     (Bread)      (Eggs)      0.5        0.75
2      (Milk)     (Bread)      0.5        0.75
3     (Bread)      (Milk)      0.5        0.75
```

Pattern

1. A pattern is a recurring, consistent, or systematic arrangement of data elements, often reflecting relationships or correlations between variables. Patterns can help identify underlying structures or behaviors in data.
2. A pattern represents a recurring relationship or structure in the data. It is often context-independent, meaning it may hold true across different situations or time periods. Patterns are usually observed by analyzing relationships between variables in a dataset.

Examples of Pattern

- Market Basket Analysis:
    - Technique: Association Rule Mining
    - Example: A retail store collects transaction data of customer purchases. By applying association rule mining, the store identifies a pattern that customers who buy diapers often also buy baby wipes.
    - Decision: The store can use this pattern to place diapers and baby wipes close together, create promotional offers or bundles, and increase sales of both products.
- Pattern (independent of time):
    - Technique: Clustering
    - Example: A company segments its customers into groups based on their purchase behavior, demographics, and preferences using a clustering algorithm. They find a pattern that customers in a specific cluster prefer eco-friendly products.
    - Decision: The company can use this pattern to target marketing efforts and promotions for eco-friendly products to this specific customer segment.

Trend

1. A trend is a general direction or long-term change in data over time. It indicates a consistent increase, decrease, or pattern in a dataset, which can help make predictions about future data points or understand the underlying dynamics of a process.
2. A trend is a specific type of pattern that refers to the long-term change in data over time. Trends are context-dependent, as they are associated with a specific time period or sequence of data points. Trends are typically observed in time-series data, where the data points are ordered chronologically.

Examples of trend

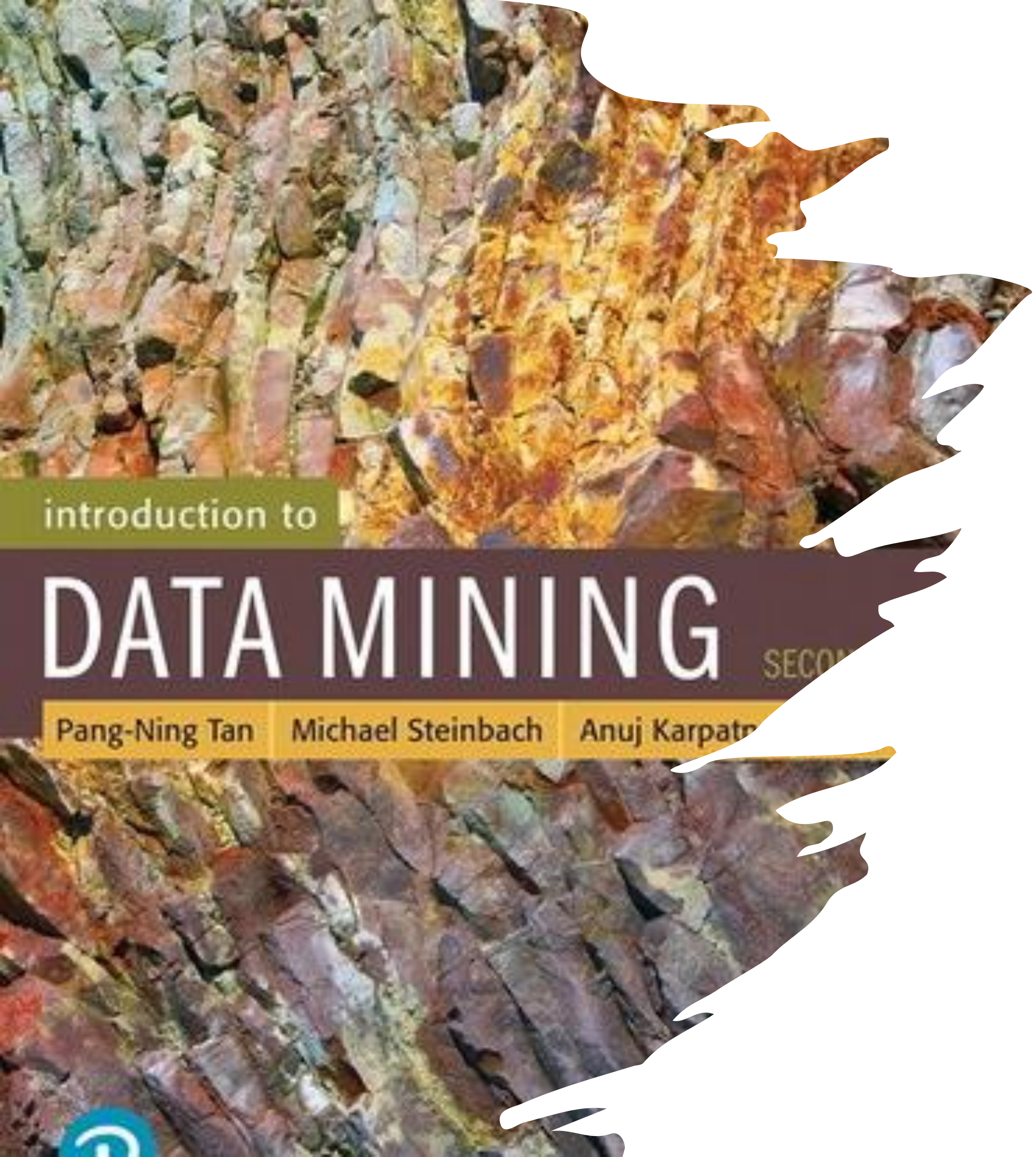
- Stock Market Analysis:
  - Technique: Time Series Analysis
  - Example: A financial analyst examines historical stock prices of a company and observes a consistent upward trend over the past five years.
  - Decision: The analyst could advise clients to invest in the company, expecting that the upward trend may continue in the future, leading to potential gains for investors.
- Trend (time-dependent):
  - Technique: Time Series Analysis
  - Example: A city planner analyzes the historical data of air quality in the city and observes a seasonal trend with consistently poor air quality during winter months.
  - Decision: The city planner can use this trend to implement preventive measures, such as increasing public transportation options or restricting high-emission vehicles during winter months to improve air quality.

Insight

1. An insight is a valuable piece of information or understanding derived from data analysis. Insights are often unexpected or non-obvious findings that can help drive decision-making, problem-solving, or strategy formulation.
2. An insight is the actionable information or understanding derived from the analysis of patterns or trends. Insights are the results of interpreting patterns and trends within a specific context, leading to valuable information for decision-making or problem-solving.

Example of Insight

- Customer Churn Prediction:
  - Technique: Classification
  - Example: A telecom company analyzes its customer data and applies a classification algorithm to predict which customers are at risk of churning. The analysis reveals that customers with limited data plans and high overage charges are more likely to churn.
  - Decision: The company can use this insight to proactively offer these customers better data plans, incentives, or discounts to retain them and reduce churn.
- Insight (resulting from interpretation of patterns or trends):
  - Technique: Text Mining
  - Example: A company analyzes customer reviews of their products using text mining techniques and discovers that customers consistently mention long delivery times as a major concern.
  - Decision: The insight gained from this analysis leads the company to revise its logistics and shipping processes, aiming to reduce delivery times and improve customer satisfaction.

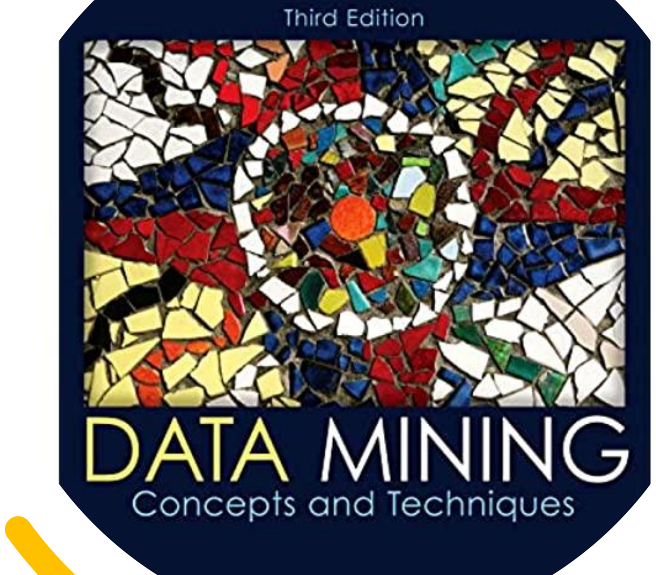


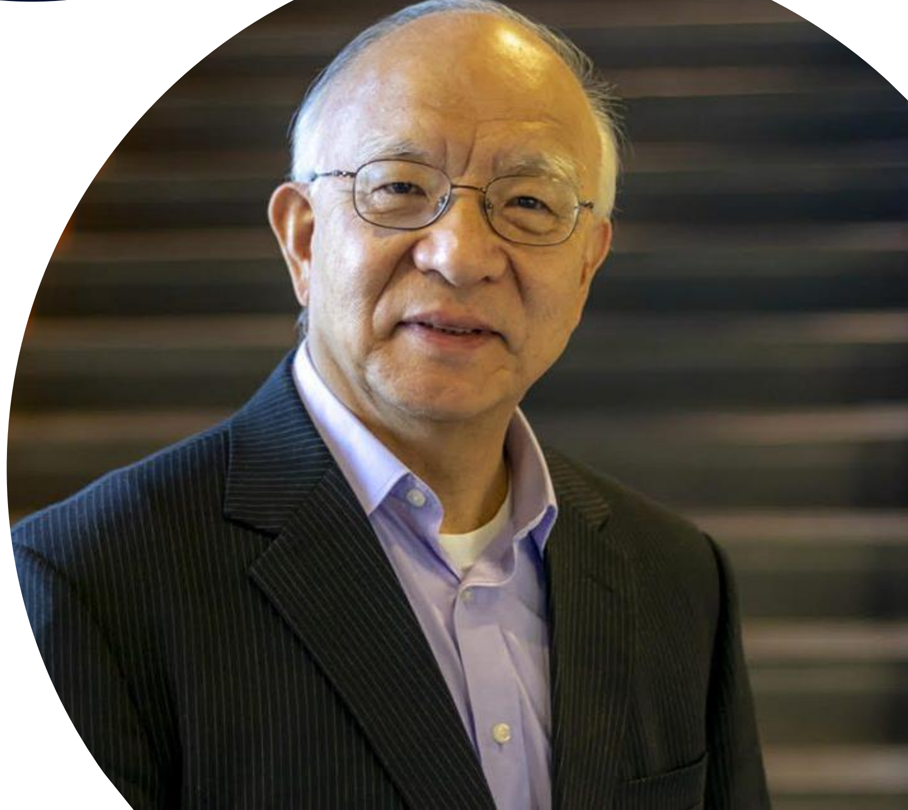
Required Book 1

- **Title**: Introduction to Data Mining (2nd Edition) (What's New in Computer Science) 2nd Edition
- **By**: Pang-Ning Tan (Author), Michael Steinbach (Author), Anuj Karpatne (Author), Vipin Kumar (Author)
- **ISBN-13: 978-0133128901**
- **ISBN-10: 0133128903**
- **Link**: https://www.pearson.com/store/p/introduction-to-data-mining/P100001265344/9780133128901

Required Book 2

- **Title**: Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) 3rd Edition
- **By:** Jiawei Han (Author), Micheline Kamber (Author), Jian Pei (Author)
- **ISBN**-13: 978-9380931913
- **ISBN**-10: 9780123814791
- **Link**: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

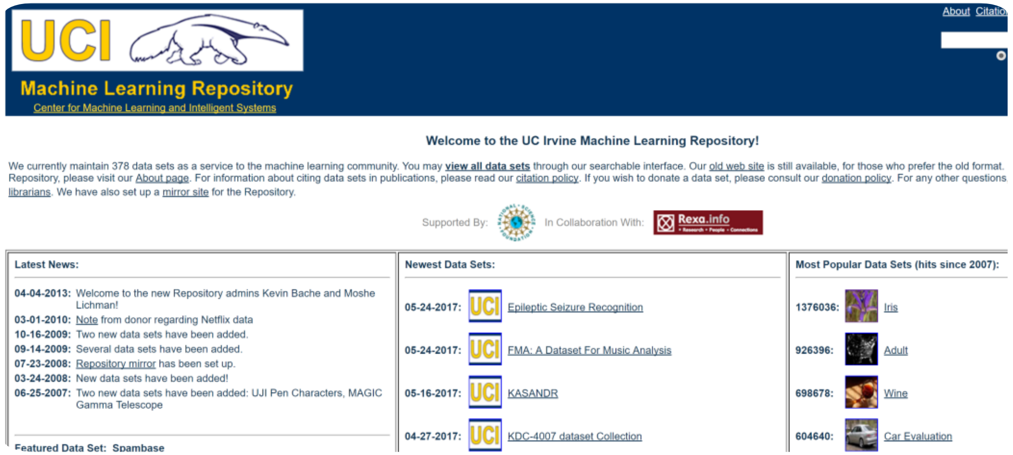




Data Sources

If you want to play around with algorithms, you can obtain many datasets from the Machine Learning Repository at University of California at Irvine (UCI). You can find it at:

http://archive.ics.uci.edu/ml



UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citatio

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 378 data sets as a service to the machine learning community. You may **view all data sets** through our searchable interface. Our old web site is still available, for those who prefer the old format. Repository, please visit our About page. For information about citing data sets in publications, please read our citation policy. If you wish to donate a data set, please consult our donation policy. For any other questions librarians. We have also set up a mirror site for the Repository.

Supported By: In Collaboration With: Rexa.info

Latest News:

04-04-2013: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!
03-01-2010: Note from donor regarding Netflix data
10-16-2009: Two new data sets have been added.
09-14-2009: Several data sets have been added.
07-23-2008: Repository mirror has been set up.
03-24-2008: New data sets have been added!
06-25-2007: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Featured Data Set: Spambase

Newest Data Sets:

05-24-2017: Epileptic Seizure Recognition
05-24-2017: FMA: A Dataset For Music Analysis
05-16-2017: KASANDR
04-27-2017: KDC-4007 dataset Collection

Most Popular Data Sets (hits since 2007):

1376036: Iris
926396: Adult
698678: Wine
604640: Car Evaluation

Study Groups

- A study group can help solidify and clarify course materials, leading to more promising classroom experiences, and potentially a better GPA.
- Study groups can help you develop as a student, person, and professional. Study groups encourage members to think creatively and build strong communication skills which also help in refining understanding of the material. https://www.fnu.edu/10-reasons-form-study-group/

Create a group of 2-3 students. Make sure to email me the people in your group one it is made.

Exercise 1: Anomalies

You are working as a data analyst in a bank. The bank has recently experienced a number of fraudulent credit card transactions. Use Python and a credit card transaction dataset (which includes columns such as transaction time, amount, cardholder name, and whether the transaction is fraudulent or not) to identify anomalous transactions that might be fraudulent. Use an anomaly detection algorithm of your choice (like Isolation Forest, Local Outlier Factor, or any other suitable method). Provide the Python code and a brief explanation of your approach and results.

Exercise 2: Patterns

You are a data scientist in a supermarket. The supermarket wants to understand the buying behavior of customers to increase sales. Use Python and a dataset of customer transactions (where each transaction includes the items bought) to identify frequent itemsets or patterns. You can use the Apriori algorithm or any other suitable method for this task. Provide the Python code and a brief explanation of your findings.

Exercise 3: Correlations

You are a healthcare data analyst. Your task is to understand the factors affecting heart disease in patients. Use Python and a dataset of patient records (which includes columns such as age, gender, cholesterol level, smoking status, and whether the patient has heart disease) to find correlations between different variables and heart disease. Provide the Python code and a description of the correlations you found.

Remember to include data preprocessing steps in your code and explain why you chose the particular algorithm for each task.

Q&A