STA5069Z - Topic Ideas

Sebastian Eide Aas

March 2025

Option 1

This option studies a dataset containing transaction records from a UK retailer. The research question is formulated as follows:

How can customer segmentation and dimensionality reduction techniques help analyze purchasing patterns and identify factors influencing sales trends?

The dataset consists of 541.909 rows and 8 columns, with the columns being

1. InvoiceNo: identifier for each transaction.
2. StockCode: product code.
3. Description: description of product.
4. Quantity: quantity of products sold.
5. InvoiceDate: date and time for the transaction.
6. UnitPrice: price of product.
7. CustomerID: customer ID for each transaction.
8. Country: country where the transaction took place.

Option 2

This option studies a dataset with crime rates from Boston. The research question is formulated as follows:

How can we use crime data to identify high-crime areas and predict the occurrence of crime in Boston?

The dataset consists of 501.070 rows and 17 columns, with the columns being

1. incidentnumber: unique identifier for each reported crime incident.
2. offensecode: numerical code representing the specific offense.
3. offensecodegroup: general category of the offense.
4. offensedescription: detailed description of the offense.
5. district: police district where the incident occurred.
6. reportingarea: specific area code within the district.

7. shooting: indicator if the incident involved a shooting.
8. occurredondate: date and time when the incident occurred.
9. year: year of the incident.
10. month: month of the incident.
11. dayofweek: day of the week when the incident occurred.
12. hour: hour of the day when the incident occurred.
13. ucrpart: uniform crime reporting classification.
14. street: street address where the incident occurred.
15. lat: latitude coordinate of the incident location.
16. long: longitude coordinate of the incident location.
17. location: longitude and latitude as a tuple.

I also have a similar dataset (crimes in Chicago) for the same kind of research question, which I would like to use in case the above dataset is very bad.

Option 3

This option studies a dataset with customer demographics and transactions data from an Indian bank. The research question is formulated as follows:

How can customer segmentation based on purchasing behavior and demographic factors help in understanding banking customer behavior?

The dataset consists of 1.048.567 rows and 9 columns, with the columns being

1. transactionid: unique identifier for each transaction.
2. customerid: unique identifier for each customer.
3. customerdob: date of birth of the customer.
4. custgender: gender of the customer.
5. custlocation: location of the customer.
6. custaccountbalance: current account balance of the customer.
7. transactiondate: date when the transaction was made.
8. transactiontime: time when the transaction was made.
9. transactionamount: amount involved in the transaction.