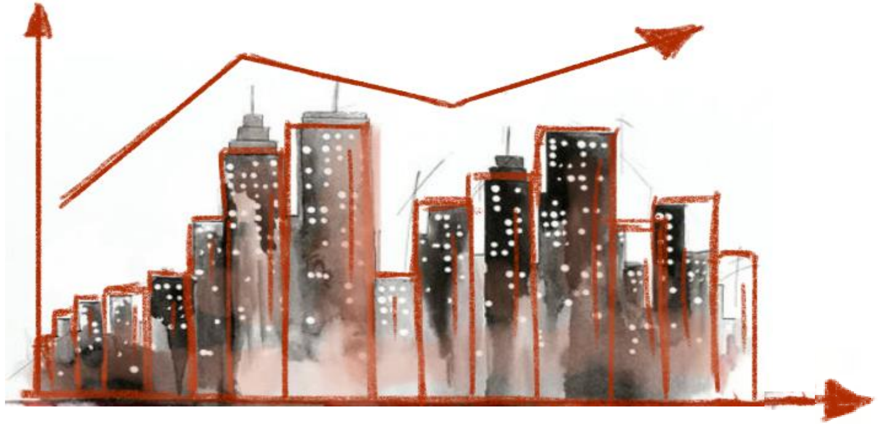
MTAT.03.319

Business Data Analytics

Lecture 8: Course Summary

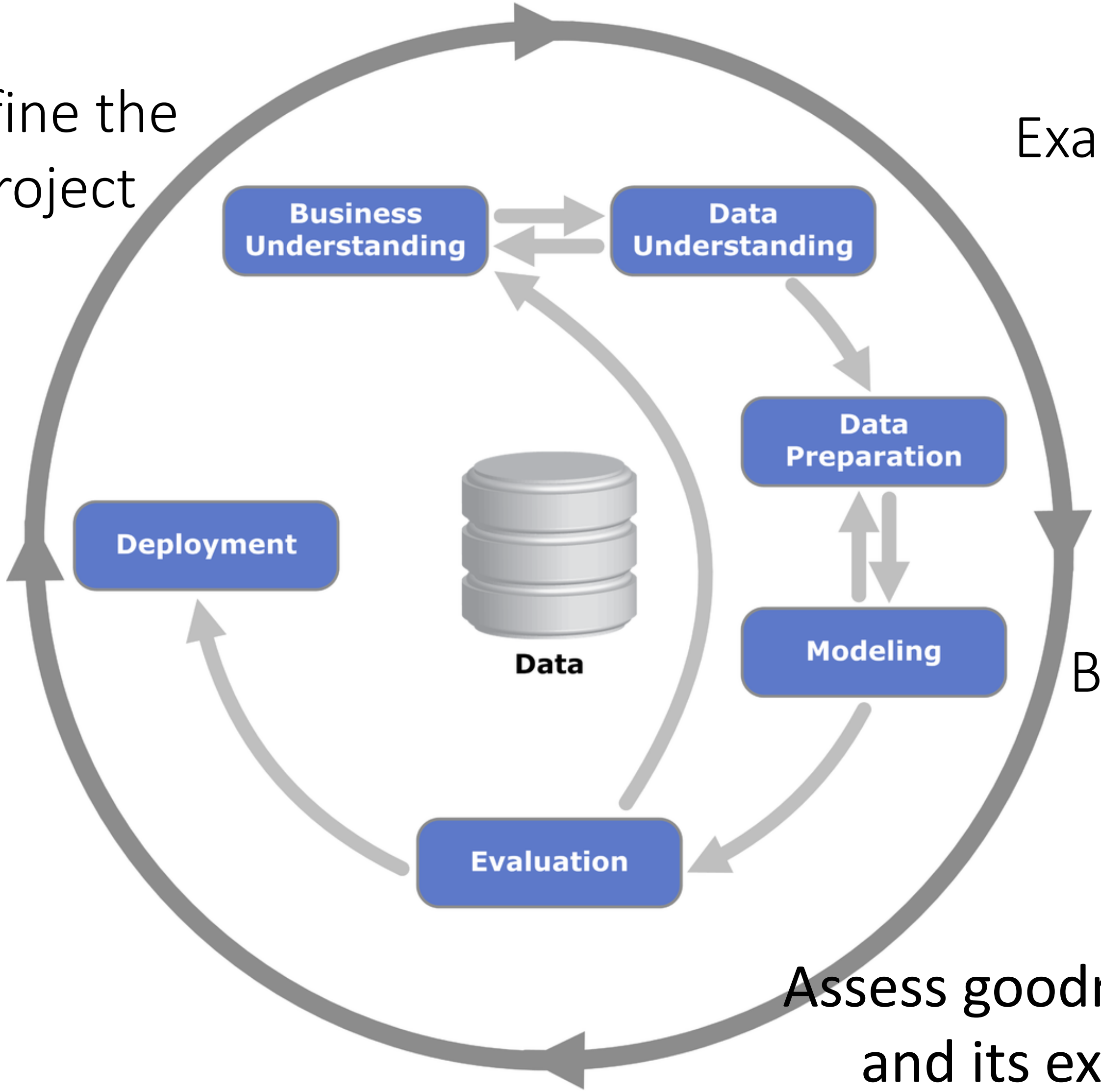


Rajesh Sharma

Recap

1. What is data analytics? Why we need it? How to approach it?
2. Data exploration: visualization & descriptive analysis
3. Customer segmentation
4. CLM – regression
5. CLM – classification (propensity, churn)
6. CLM – recommender systems (cross-sell/up-sell)
7. A/B Testing

Recap: CRISP-DM
Cross-Industry Standard for Data Mining



Define the project

Examine the data

Fix data issues

Build models

Assess goodness of the model and its expected impact

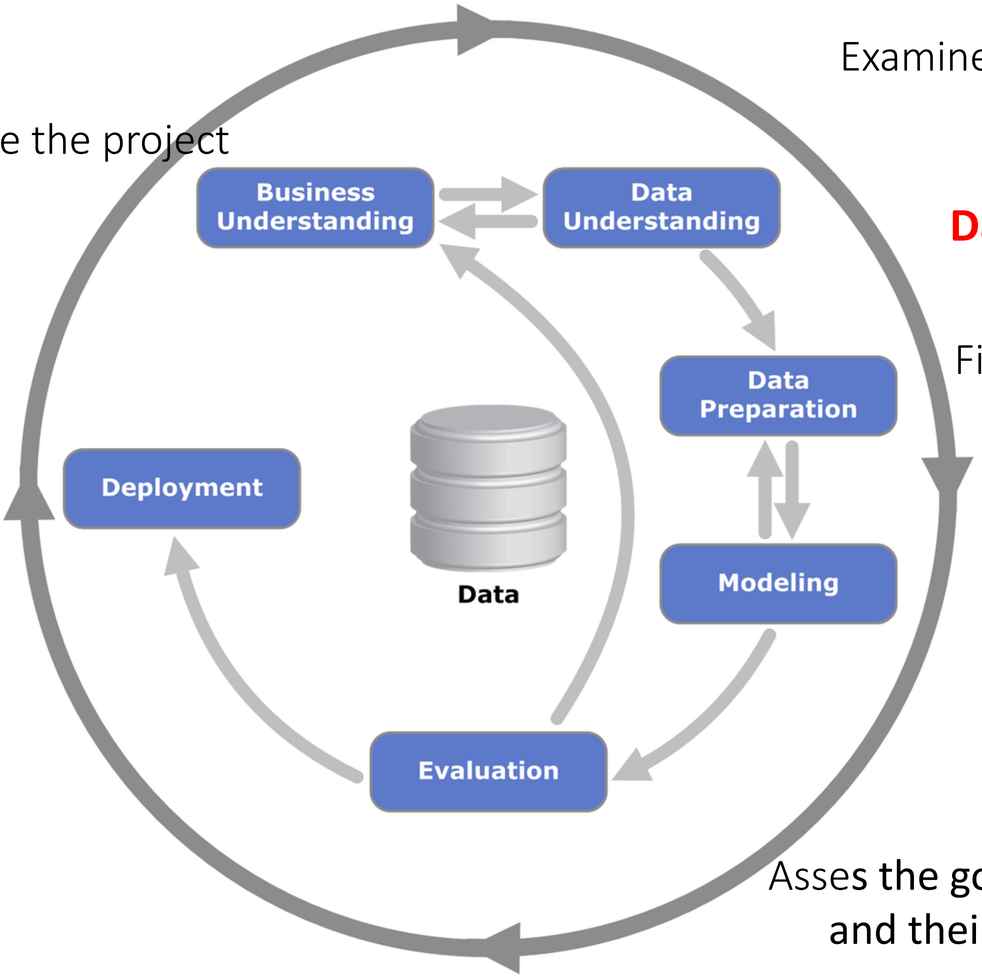
Put models and insights into use

Who is involved ?

Domain Experts

Business sponsor

Define the project



Data steward

Analytics expert

Examine the data; identify issues with data

Data steward

Analytics expert

Fix data issues; create derived variables (features)

Build models

Analytics expert

Asses the goodness of the model and their expected impact

Analytics expert

Put models and insights into use

Business Understanding

- Define the business objective
- Formulate the question(s)
- Identify target variable & attributes
- Define the success criteria
- Cost/benefit analysis

Who is involved?

- Business sponsor
- Domain expert(s)
- Analytics expert
- Data steward & DB expert

Data Understanding

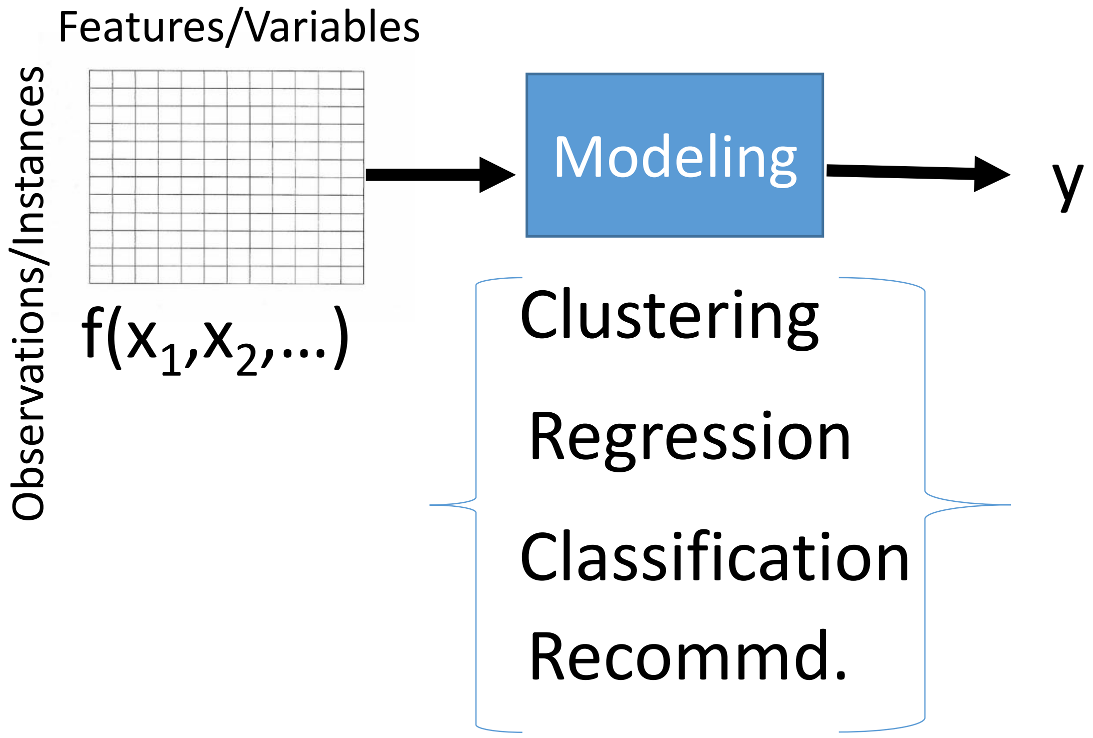
- Data Collection
  - Identify data sources
  - Write queries
- Data Description
  - Document data quality issues
  - Compute basic statistics
- Data Exploration
  - Simple univariate data plots/distributions
  - Investigate attribute interactions
  - Data Quality Issues
    - Missing Values
    - Skewed Distributions

Data Preparation (cont.)

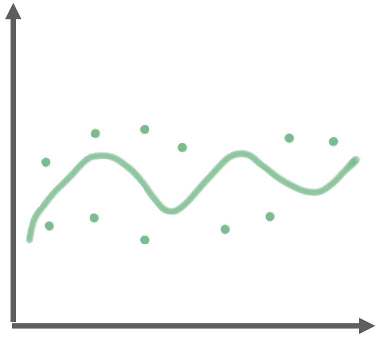
- Integrate Data
  - Joining multiple data tables/frames
  - Summarisation/aggregation of data
- Select Data
  - Attribute subset selection
  - Sampling (sometimes useful for large datasets)
- Transform data
  - Using functions such as log
  - Normalization/Discretisation/Binning
- Clean Data
  - Handling missing values/Outliers
- Enrich Data
  - Calculate derived attributes

Modeling

- Select modeling technique depending on type of problem/output
  - Supervised versus unsupervised
  - Regression versus classification
- Develop a testing regime
  - Select measures of model quality
  - Sampling (train versus test)
- Build Model
- Assess the model

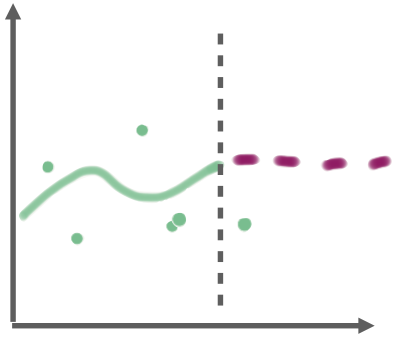


Types of Business Analytics



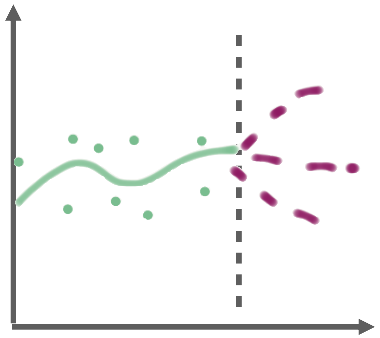
Descriptive

What has happened?
E.g. what top five
customer segments we have?
Which pairs of products
are bought together?



Predictive

What will happen?
E.g. Who will buy?
Who will churn?

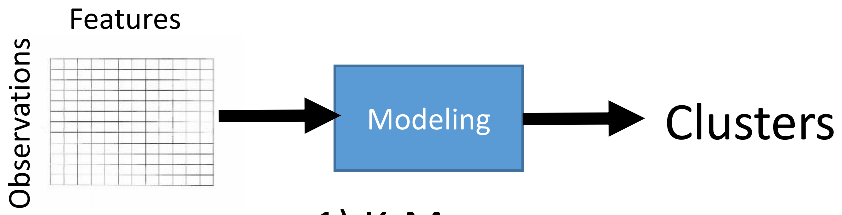


Prescriptive

What to do to achieve my goals?
When should I make my next
customer call, to whom
and what should I tell them?

Recap – Customer Segmentation

- RFM model
  - What does it stand for? What is it useful for? How can it be used to group customers?
- Clustering
  - K-means clustering and hierarchical clustering
    - What are they? What do they need as input? What they provide as output?
    - What are their relative advantages and drawbacks?
  - How do we determine the *k* in k-means clustering?


Features

Observations

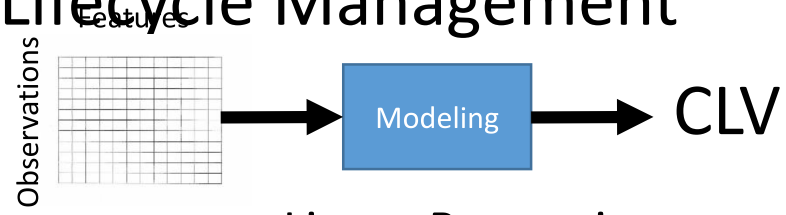
Modeling

Clusters

1) K-Means
2) Hierarchical Clustering
3) DBSCAN

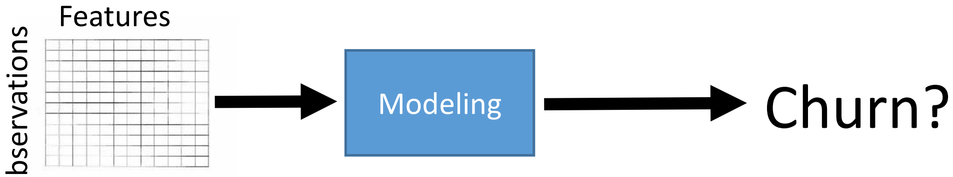
Recap – Regression in CLM

- What is CLV (or CLTV)?
- What is regression?
  - What is the input? What is the output?
- How do we train a regression model?
- How do we measure how good a regression model is?
- How can regression be used in Customer Lifecycle Management (CLM)?



Linear Regression

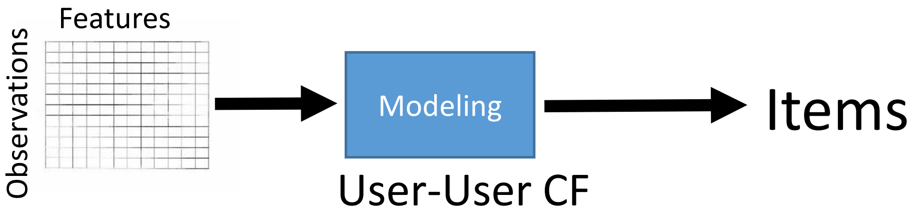
Recap – Classification in CLM



Logistic Regression
Random Forest
XgBoost

- What is classification?
  - What is the input? What is the output?
- How do we train a classification model?
  - Which methods are there? How to use them?
  - What is the difference between a white-box and a black-box classification?
- How do we measure how good a classification model is?
- What is over-fitting? How can we detect it?
- What is class imbalance? How does it impact classification?

Recommender systems for cross-and up-selling



User-User CF

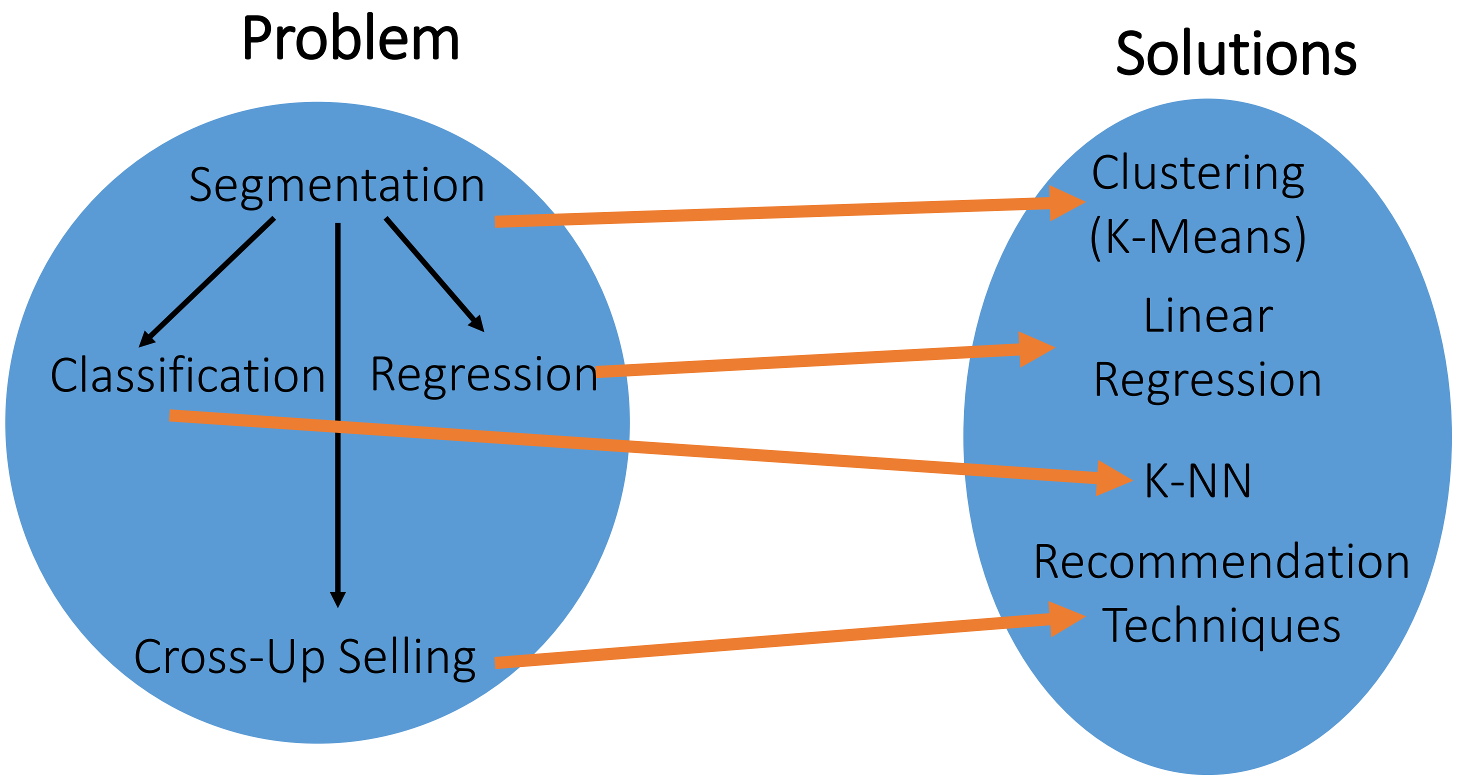
Item-Item CF

- Market-basket analysis
  - What is it? What is it useful for?
- What is the relation between market-basket analysis and association rule mining?
- What is the input of output of association rule mining?
- How do we measure the goodness of association rules?
- Collaborative filtering: user-based versus item-based
  - What is it? What is it useful for? What is the tradeoff between user-based versus item-based collaborative filtering
- Tradeoffs between market-basket analysis and collaborative filtering

A/B Testing

- Call to Actions -> Conversion rate
- Results you are seeing are just by chance or they are statistically significant?
- Hypothesis Testing: Version A (Null Hypothesis) is better than B (alternate hypothesis)?
- T-test
    - T-value: Difference between two distributions (observations)
    - p-value: used for Reject or Accept the null hypothesis compared to significance value.
    - significance value: Threshold (0.5 or 0.1)
- Test of proportions
    - If you are measuring the proportions rather than absolute values

At the end of the course: mesh structure



Exam Preparation

Exam structure

- 30 points, ~15 questions and 1 to ~4 points per question
- Time: 2 hours
- **NO** correlation between number of correct options and marks. For example, If a question has 3 points, does not mean it has 3 correct options.
- Two A4 sized cheat sheets allowed. No limit on the font size. But **do not** zoom more than 100%.
- All communications through email that you have registered on SIS (outcome of your exam).
- We need around 2 weeks to compile your total grades (after your exam).
- We will send the result of both exams together. So students who will be giving the first exam: You have to wait for ~ 3 weeks.

How Exam will be conducted online?

- Zoom (will be send before the exam) + Moodle (Login to Moodle)
- During Exam: NO discussion with your fellow classmates.
- Use of Mobile phones during the exam is NOT allowed.
- Ask us privately about your doubts (through zoom chat).
- You will do the exam by Moodle and also be online on Zoom.
- Open your camera and share your screen during the entire exam.
- Please make sure you have good internet.

Types of Questions

1) Multiple choice questions:
2) Fill in the blanks questions
3) Simple calculation questions (calculators, desktop/laptop allowed)
4) Analyzing/comparing plots: Answers the questions based on provides data/plots.
5) Problem-solving questions

NOTE: Some questions invite negative questions. Please read them carefully. When options are provided, they naturally invite negative marking.

Question 1: Simple multiple choice question

- (1 point) Mark the correct option(s)
  Supervised learning differs from unsupervised learning in that supervised learning requires:

1. At least one input feature.
2. Input features to be categorical.
3. At least one output feature.
4. Output features to be categorical.

Answer: 3 as both require input features and you can give any kind of data (off course you need to change the format depending on the algorithm)

NOTE: a wrong selection cancels out a correct selection, e.g. two correct selections and one incorrect = one correct selection.

Question 2: Interpreting a plot

(1 point) Both graphs shows the number of customers which bought a product in 3 months. Which graph is better ? Mark the correct option(s).

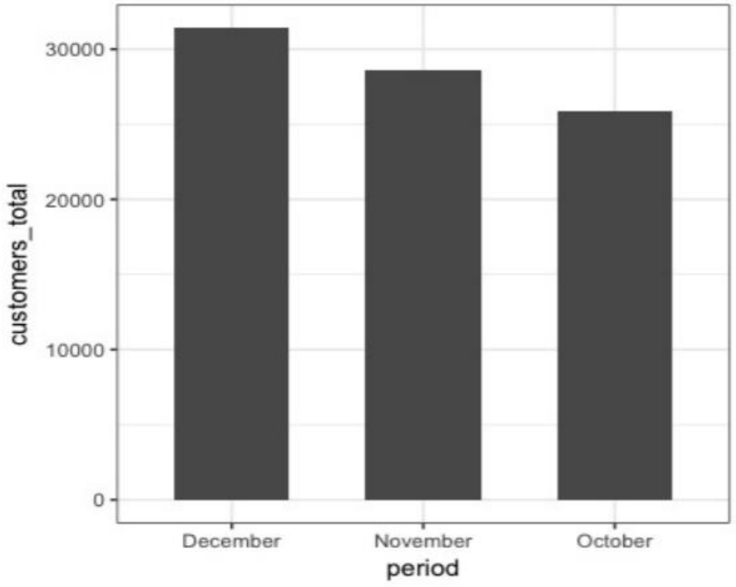
A. plot (a) as it does not skew the data showing the adequate difference

B. plot (b) as it highlights the difference between months

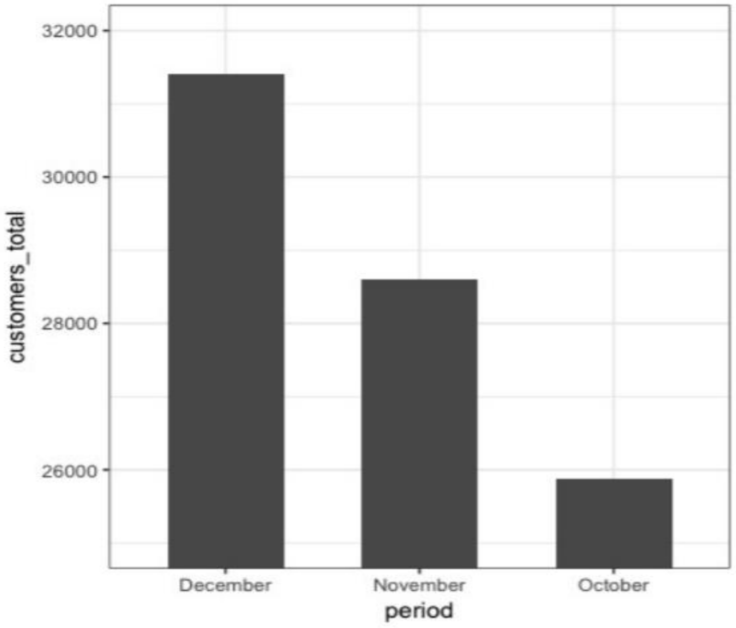
C. plot (a) as the last y-axis tick is closer to the maximum

D. Both plots are bad

Answer: A (but D is also acceptable as months are not ordered properly)



(a)



(b)

Question 3: Fill in the Blanks (simple, definitional question)

- (1 point) RFM analysis ranks customers by considering ...... of their orders.

Question 4: Simple calculation question

- (2 points) Consider the following confusion matrix below:

Prediction

Actual

	Positive	Negative	Total
Positive	50	30	80
Negative	25	15	40
total	75	45	120

- Calculate Precision and Recall based on these numbers.

Question 5: Problem-Solving (open ended question)

(3 points) Reflect on the following case. What modelling techniques to use and for what purpose? What features could be extracted to build the model?

- You are inventory manager in an e-commerce retail company that sells furniture products
- Your goal is to minimize:
  - Carrying cost (cost of holding inventory)
  - Lost sales revenue due to OOS (out-of-stock)
- The company has data about
  - All sales and all shipments for the past 5 years
  - All purchases from suppliers and all deliveries to the warehouse
- The number of Out-Of-Stock (OOS) events has increased by 5% in the past 2 years. The goal is to reduce OOS events, while capital inventory cost has been stable.