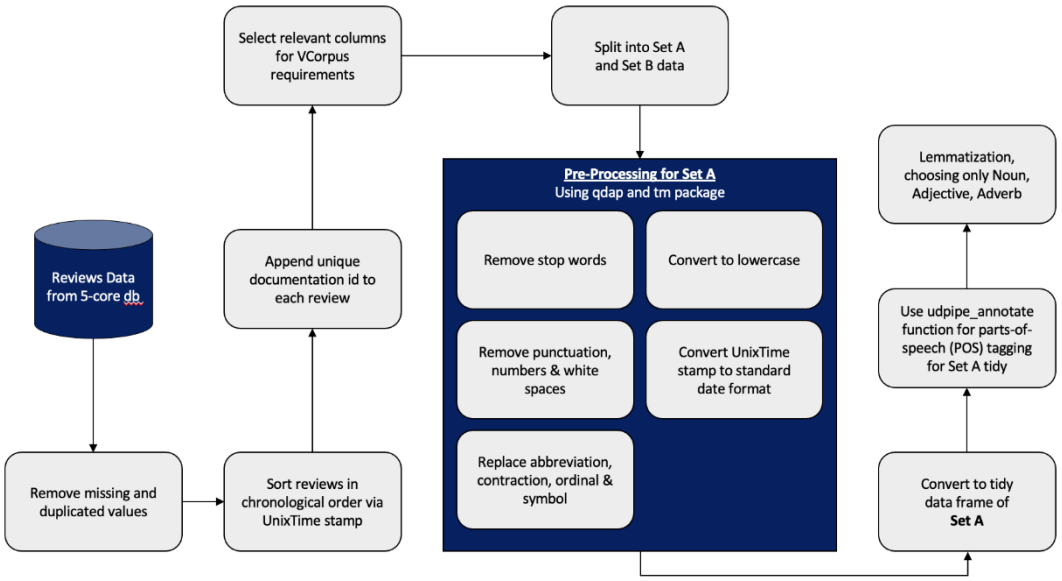
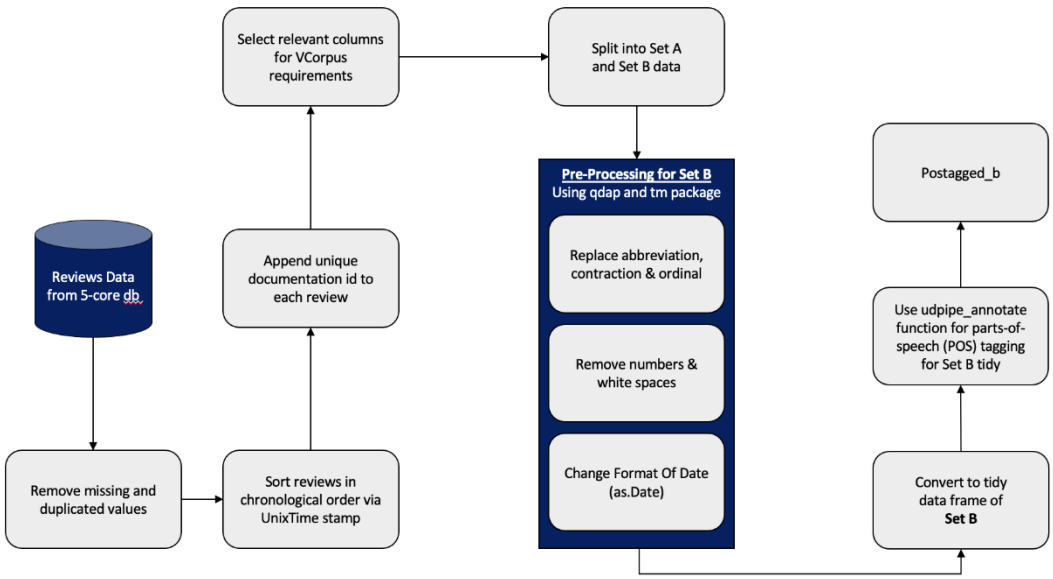
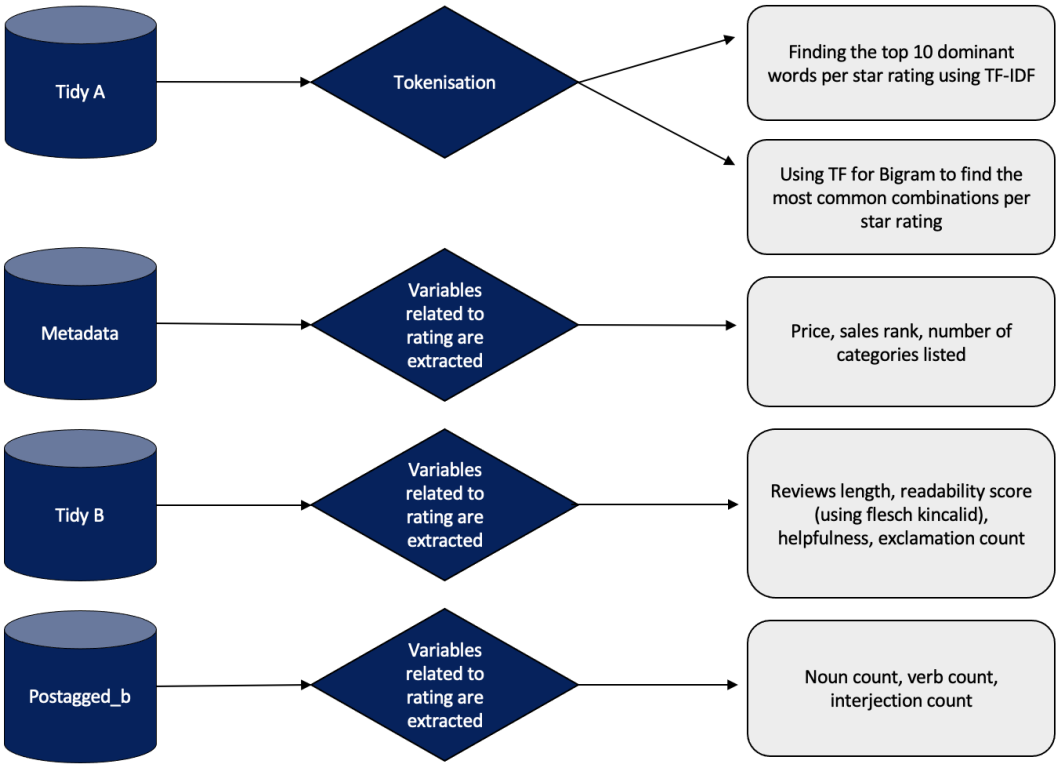
Data Staging

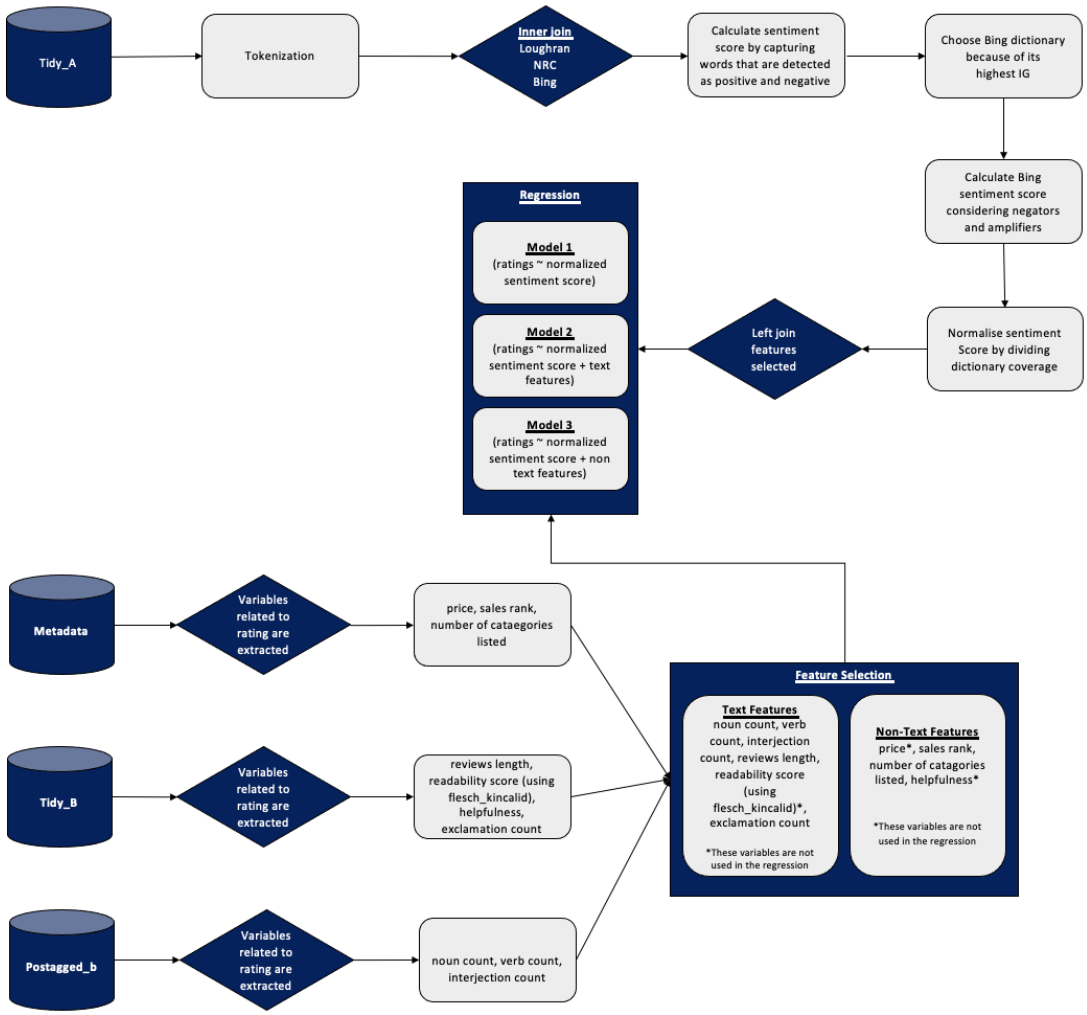




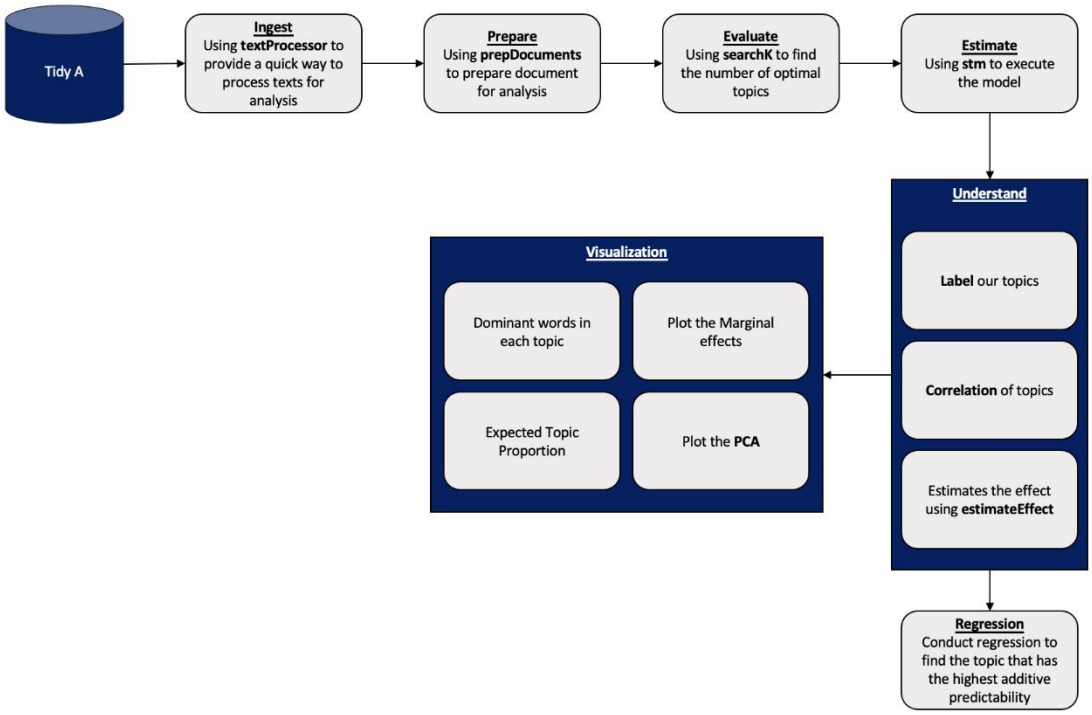
Part A


Tidy A
Tokenisation
Finding the top 10 dominant words per star rating using TF-IDF
Using TF for Bigram to find the most common combinations per star rating
Metadata
Variables related to rating are extracted
Price, sales rank, number of categories listed
Tidy B
Variables related to rating are extracted
Reviews length, readability score (using flesch kincalid), helpfulness, exclamation count
Postagged_b
Variables related to rating are extracted
Noun count, verb count, interjection count

Part B


Tidy_A
Tokenization
Inner join
Loughran
NRC
Bing
Calculate sentiment score by capturing words that are detected as positive and negative
Choose Bing dictionary because of its highest IG
Calculate Bing sentiment score considering negators and amplifiers
Normalise sentiment Score by dividing dictionary coverage
Left join features selected
Regression
Model 1
(ratings ~ normalized sentiment score)
Model 2
(ratings ~ normalized sentiment score + text features)
Model 3
(ratings ~ normalized sentiment score + non text features)
Metadata
Variables related to rating are extracted
price, sales rank, number of cataegories listed
Tidy_B
Variables related to rating are extracted
reviews length, readability score (using flesch_kincalid), helpfulness, exclamation count
Postagged_b
Variables related to rating are extracted
noun count, verb count, interjection count
Feature Selection
Text Features
noun count, verb count, interjection count, reviews length, readability score (using flesch_kincalid)*, exclamation count
*These variables are not used in the regression
Non-Text Features
price*, sales rank, number of catagories listed, helpfulness*
*These variables are not used in the regression

Part C


Tidy A
Ingest
Using textProcessor to provide a quick way to process texts for analysis
Prepare
Using prepDocuments to prepare document for analysis
Evaluate
Using searchK to find the number of optimal topics
Estimate
Using stm to execute the model
Understand
Label our topics
Correlation of topics
Estimates the effect using estimateEffect
Visualization
Dominant words in each topic
Plot the Marginal effects
Expected Topic Proportion
Plot the PCA
Regression
Conduct regression to find the topic that has the highest additive predictability