UC San Diego

Text and Interaction Based Analysis of Clothing Fit

Sumega Manadi, Bowen Chen, Junshu Li,
Kai Shi, Qiuhan Ye

Dataset Introduction

UC San Diego

The dataset contains measurements of clothing fit from *Rent The Runway*.

Basic statistics

	Modcloth	Renttherunway
Number of users:	47,958	105,508
Number of items:	1,378	5,850
Number of transactions:	82,790	192,544

Example


```
{
  "fit": "fit",
  "user_id": "420272",
  "bust size": "34d",
  "item_id": "2260466",
  "weight": "137lbs",
  "rating": "10",
  "rented for": "vacation",
  "review_text": "An adorable romper! Belt and zipper were a little
hard to navigate in a full day of wear/bathroom use, but that's to be
expected. Wish it had pockets, but other than that-- absolutely
perfect! I got a million compliments.",
  "body type": "hourglass",
  "review_summary": "So many compliments!",
  "category": "romper",
  "height": "5' 8\"",
  "size": 14,
  "age": "28",
  "review_date": "April 20, 2016"
}
```

UC San Diego

1. Interaction Based Data analysis

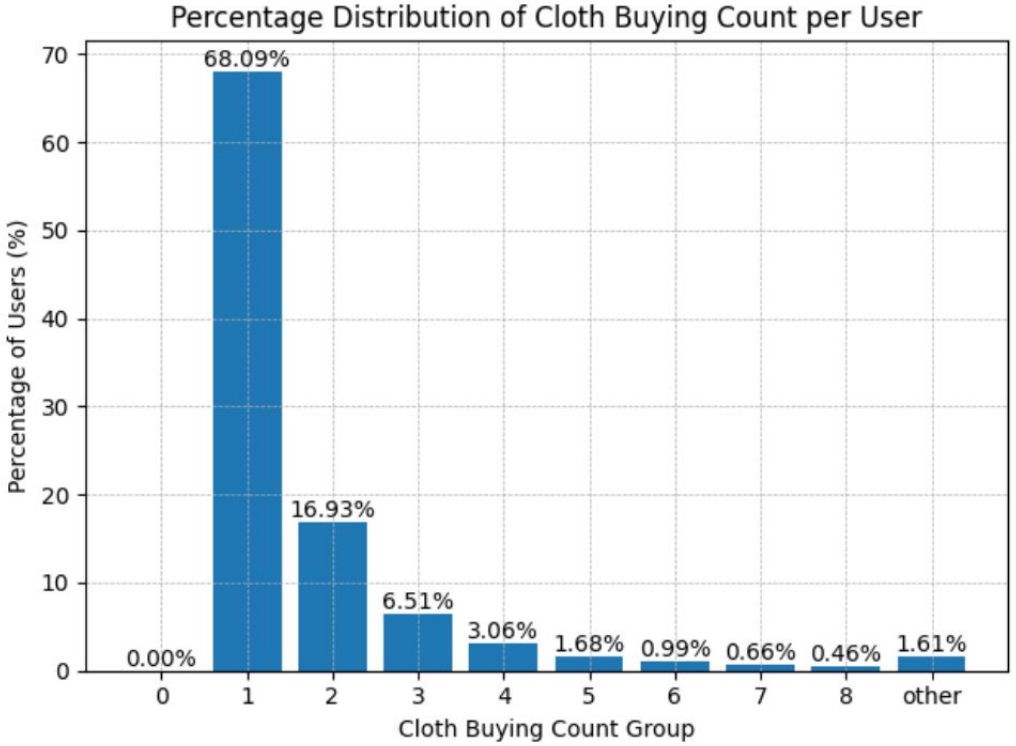
Processing and dataset cleaning

1. Calculated the number of unique users and clothing items based on the unique user and clothing IDs.

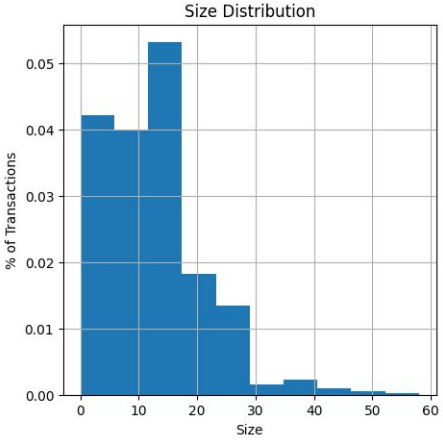
num_records	num_user	num_cloth
192544	105571	5850

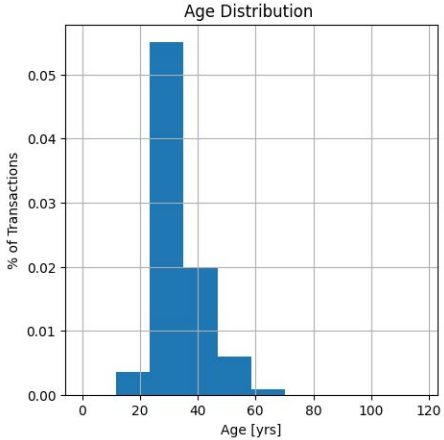
2. Define the columns to be analyzed.

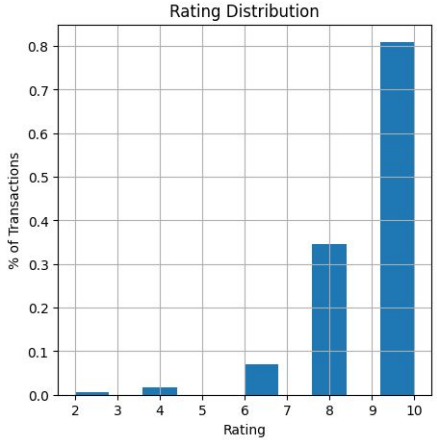
3. Visualize distribution of garments per transaction.

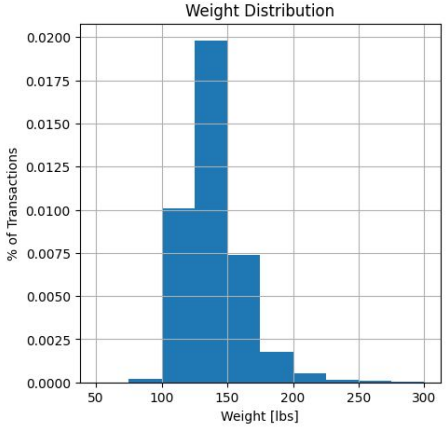


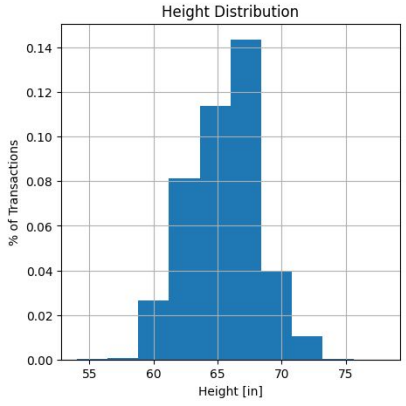
Histograms









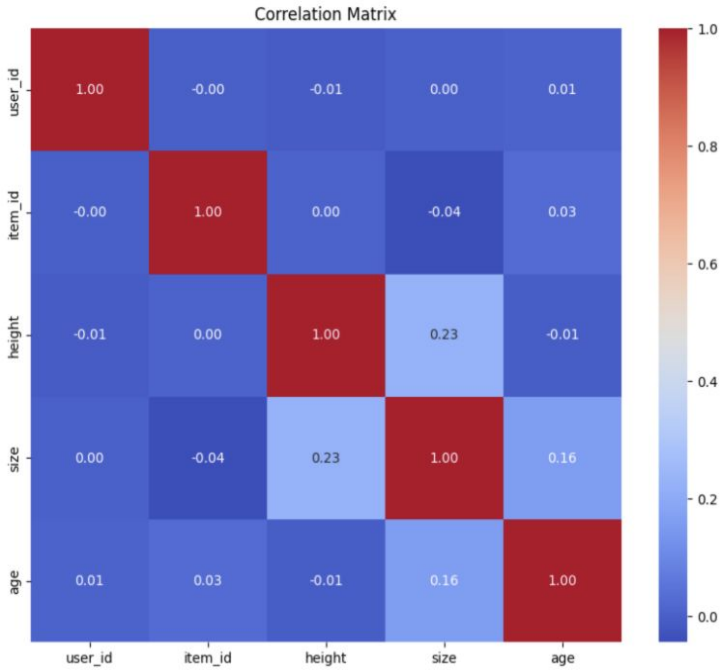


- Histograms for customer feature distributions
- Height, weight, age is close to normal distribution
- Rating Distribution shows most customers are satisfied

Correlation Matrix

UC San Diego

- Correlation matrix for clothing related features
- Features pairs (height-size), (age-size) demonstrate a positive linear correlation.



Pie Charts

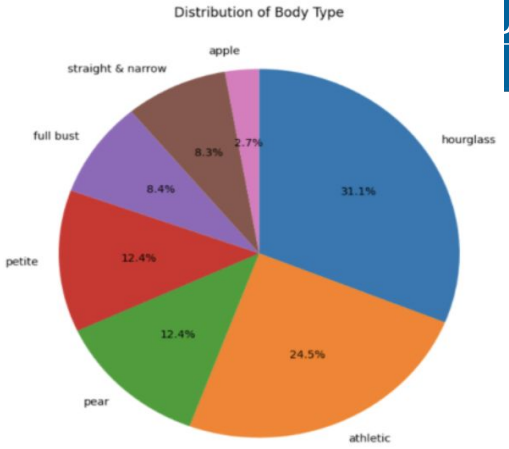
Top 3 Body types

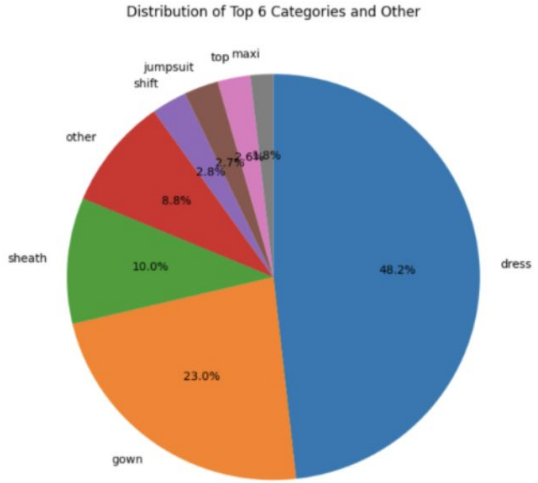
- Hourglass - 31.1%
- Athletic - 24.5%
- Pear - 12.4%

Top 3 Clothing Categories

- Dress - 48.5%
- Gown - 23%
- Sheath - 10%

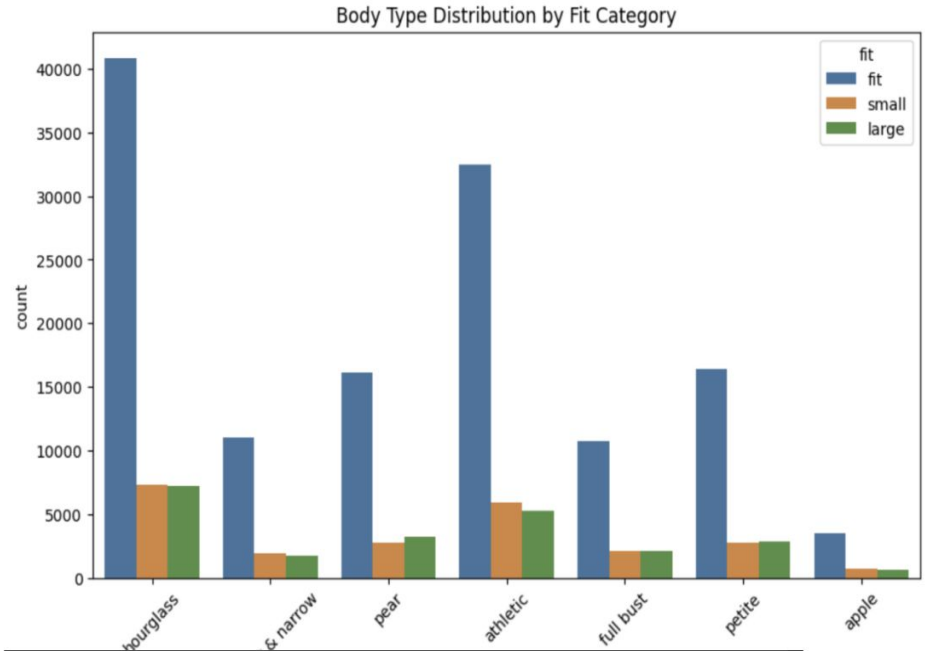
Least Popular Clothing Category - Maxi (<2%)

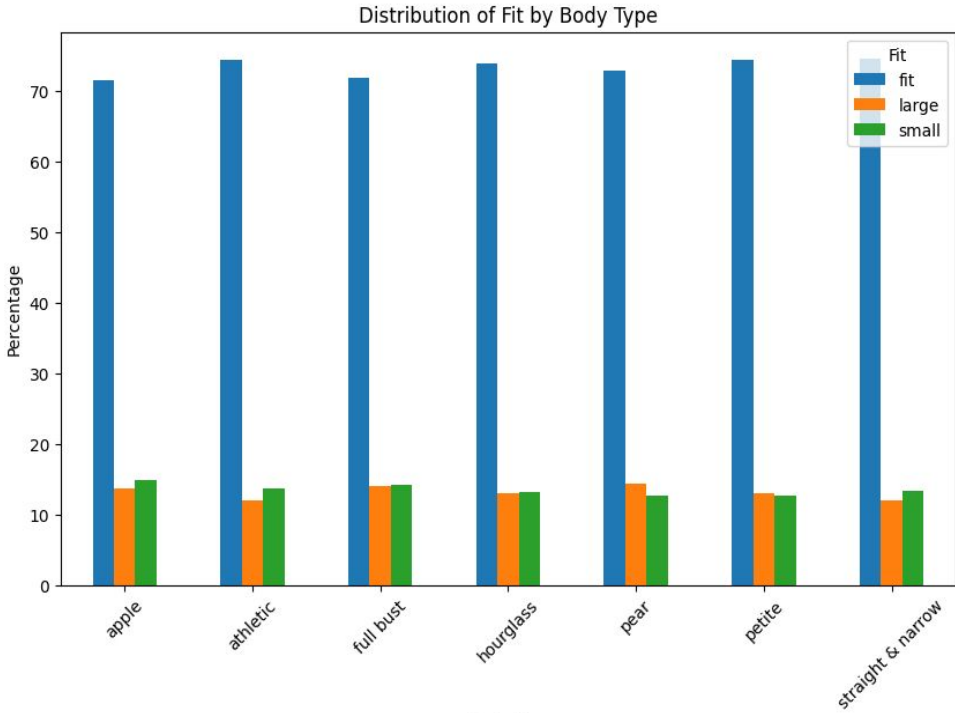




Fit Distribution by Body Type

UC San Diego



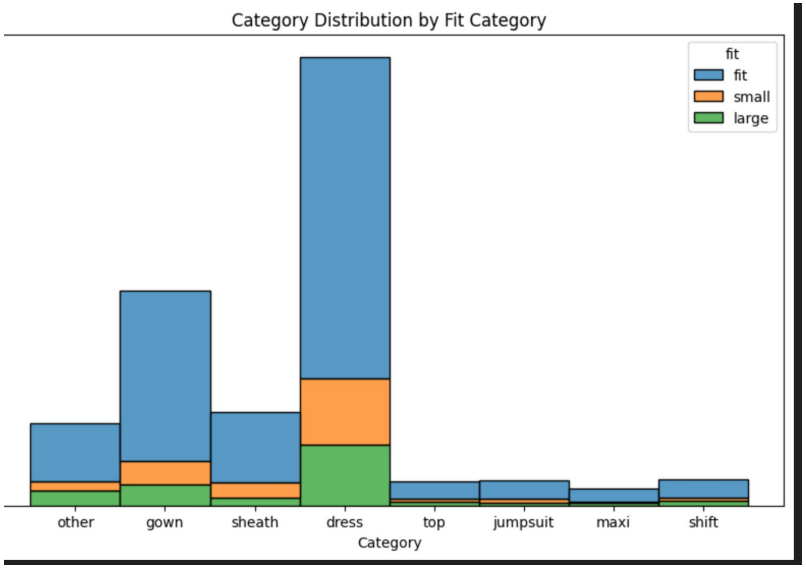


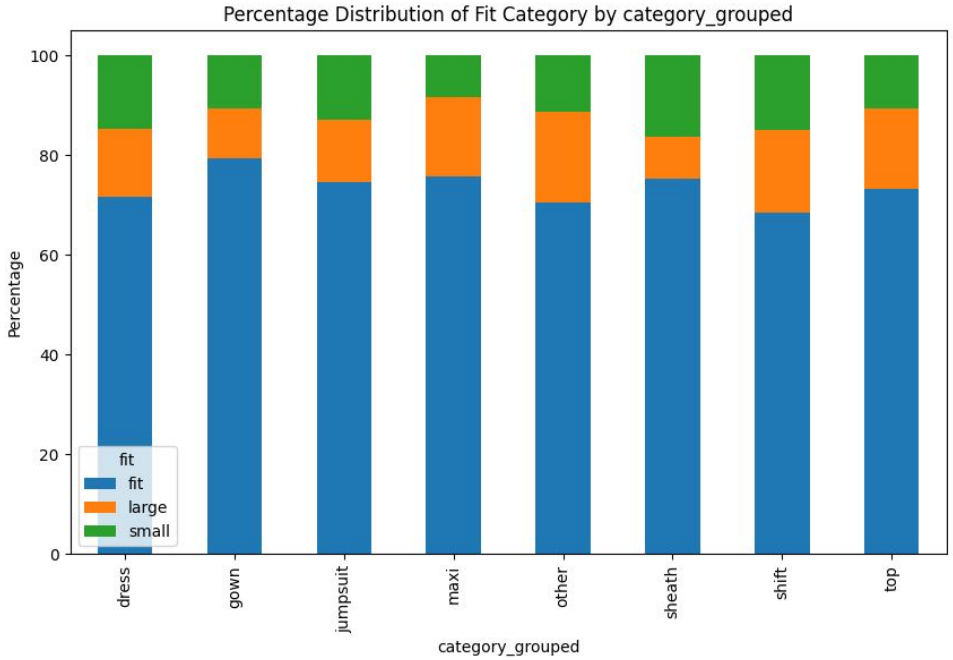
Normalized data shows no significant relationship between fit and body type

Fit Distribution by Clothing Category

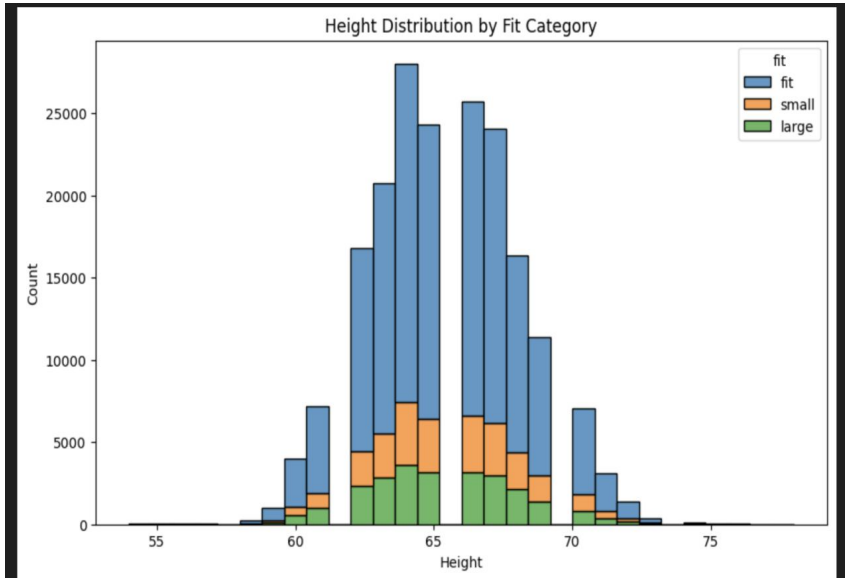
UC San Diego

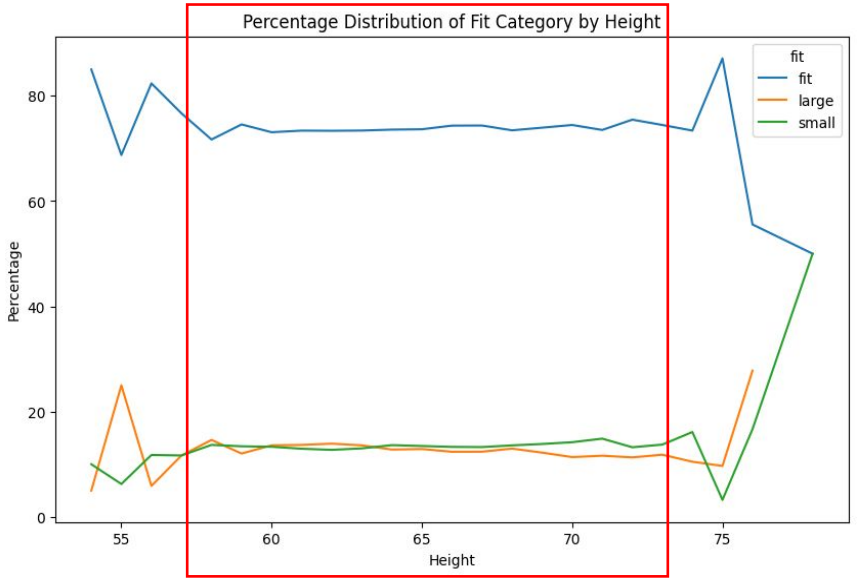
Category Distribution by Fit Category





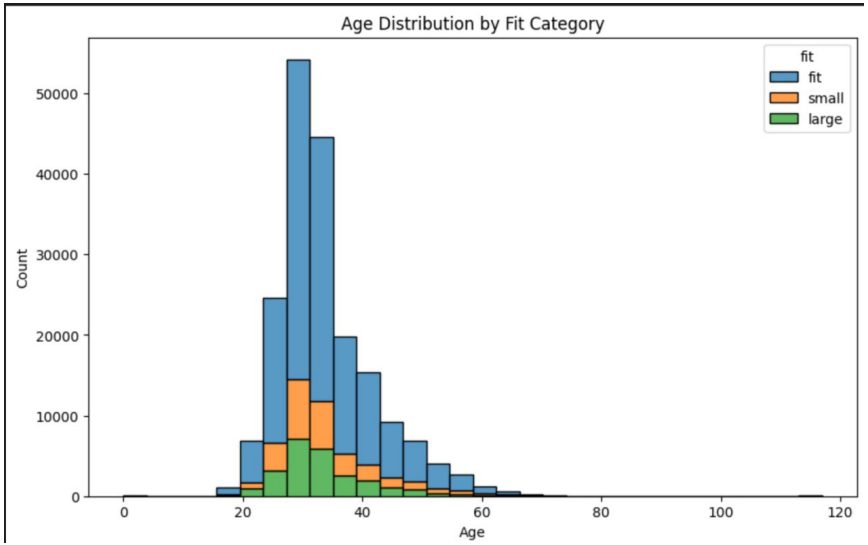
Fit Distribution by Height

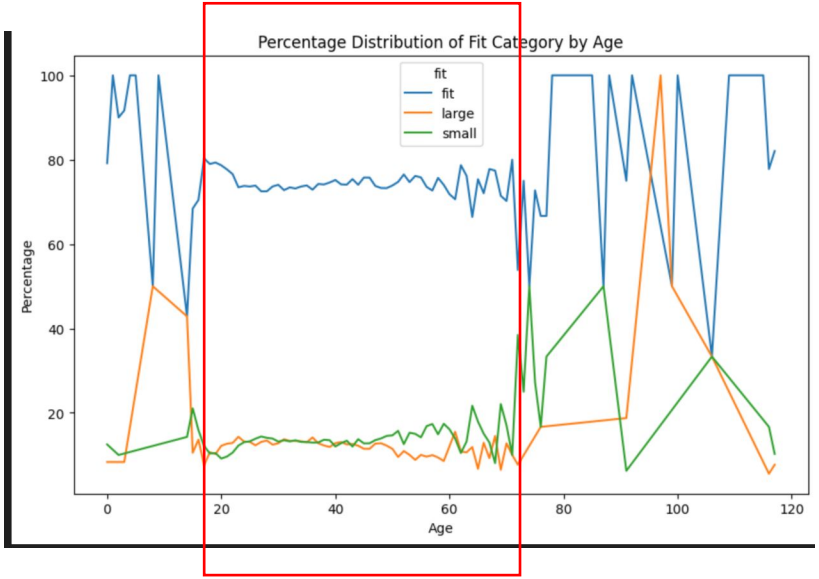




Normalized data shows no significant relationship between fit and height.

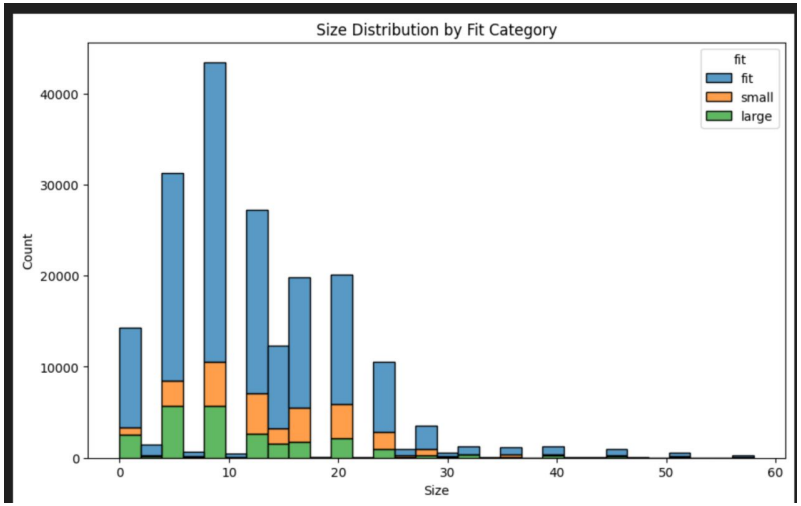
Fit Distribution by Age

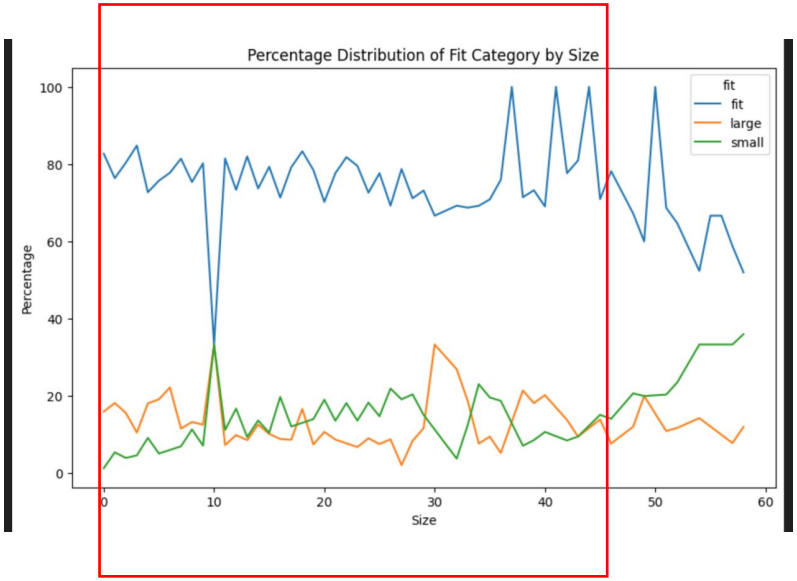




Normalized data shows no significant relationship between fit and age for young people between 20 and 40. Fit satisfaction decreases with age.

Fit Distribution by Size





Fit satisfaction decreases with very small or very large sizes

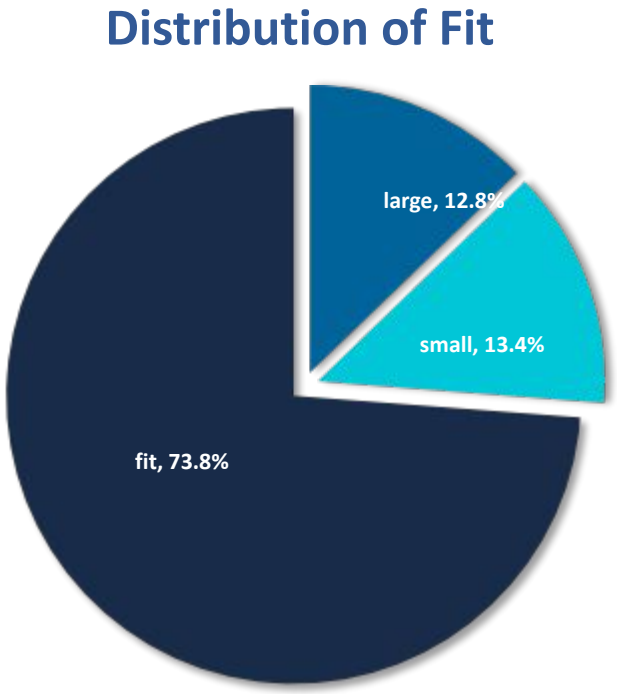
UC San Diego

2.Text-Based data analysis and models

Review Analysis

- User 691468 submitted the most comments
- The most frequency comment was “.”
- Most garments fit as expected

	count	unique	top	frequency
user id	192544	105571	691468	436
review text	192544	190951	–	63
fit	192544	3	fit	141995



Missing Values

Attribute	Null Percentile
body_type	7.60%
height	0.35%
age	0.49%
rating	0.04%
weight	15.57%
bust size	9.59%

Table2: Null value count

- Most frequent missing values were body type,weight, and bust size.
- Maybe users do not want to submit this personal information in their reviews.

Word Cloud Analysis

UC San Diego

“fit” Category

“large” Category

“small” Category

- Word clouds show most frequently used words in reviews separated by fit category.
- Most common word is dress because the data primarily consists of dress rentals.

CountVectorizer VS TfidfVectorizer

UC San Diego

model and feature	class	precision	recall	f1-score	acc
Logistic Regression CountVectorizer	fit	0.82	0.87	0.85	0.74
	large	0.49	0.39	0.43	
	small	0.41	0.33	0.37	
Logistic Regression TfidfVectorizer	fit	0.79	0.97	0.87	0.78
	large	0.72	0.27	0.4	
	small	0.6	0.2	0.3	

The accuracy and precision of TfidfVectorizer is higher than CountVectorizer.

Logistic Regression VS SVC

UC San Diego

model and feature	class	precision	recall	f1-score	acc
Logistic Regression CountVectorizer	fit	0.82	0.87	0.85	0.74
	large	0.49	0.39	0.43	
	small	0.41	0.33	0.37	
SVC CountVectorizer	fit	0.77	0.99	0.87	0.77
	large	0.78	0.18	0.3	
	small	0.87	0.08	0.14	

In Logistic Regression, the precision of the 'fit' class are significantly higher than those of other classes.
The recall of the 'large' and 'small' classes in the SVC model is relatively low.

UC San Diego

Thankyou!