User Profile Analysis

Team 11

OUR TEAM

Vince Pan

Phoenix Wang

Peter Mankiewich

AGENDA

1

Project Overview

What is our goal? Where did we get the data?

2

Exploratory Data Analysis

What did we find out about the users?

3

Text Analysis & Cluster Analysis

How did the users answer the open-ended questions?
What types of users can you expect to see?

4

Insights & Challenges

How can OkCupid benefit from our insights?

PROJECT OBJECTIVE

We aimed to understand the types of users that are on the online dating app OkCupid.
By creating distinct clusters of users, we can help people make better decisions while swiping, leading to better results for match seekers.

ABOUT THE DATASET

The dataset was taken from Kaggle.
Each row represents an anonymous user.
It contains 59,946 rows and 31 columns,
including structured personal information and
written answers to 10 open-ended questions.

02

EXPLORATORY DATA ANALYSIS

DATA CLEANING & PREPROCESSING

Missing Values

- Replace with mean
- Use logical answers
- Fill in "unknown”

Categorical Features

- Group values
- Convert to ordinal
- Create dummy variables

Textual Data

- Replace meaningless content (e.g., links)
- Concatenate strings

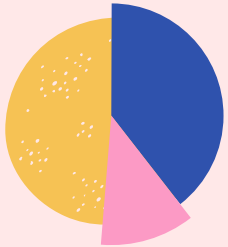
USERS OVERVIEW

GENDER

Male 60%

Female 40%

AGE



46% — < 30 yrs

35% — 30-40 yrs

19% — > 40 yrs

STRICTLY FOLLOWING DIET

Halal (23.4%)

Veggie (19.4%)

Kosher (15.7%)

TAKING RELIGION SERIOUSLY

Christianity

Islam

Hinduism

INTERESTING FACTS

1

Relationship Status

3.4% are in a relationship
0.5% are married

3

Sexual Orientation

9.3% are homosexual
4.6% are bisexual

2

Education Level

52.5% - college/university
3.5% - law or med school

4

Language Skills

Majority of Asian, Hispanic, and Middle Eastern users are bilingual

03

TEXT ANALYSIS

ESSAY QUESTIONS

About me...

The six things I could never do without...

The most private thing I am willing to admit...

What do these responses tell us about a user, and their level of commitment?

How can we use this information to make intelligent matches?

NAMED-ENTITY RECOGNITION

hello! i enjoy traveling, watching movies, and hanging out with friends. my # 1 MONEY rule for traveling is to go to places where where i can drink the tap water. =p i've been to japan GPE , washington dc GPE , new york, london GPE , paris GPE , china GPE (an exception to rule # 1 MONEY) and many other places. for the year DATE , i haven't decided yet. maybe toronto and montreal. my taste in movies (and television) fall into three CARDINAL genres: comedies, action, and drama. it doesn't matter if the film is in english or sub-titled (i can't stand dubbed), but it has to be well-written. aside from the traveling and movies, i also enjoy broadway shows (4 CARDINAL shows in 5 nights DATE in vegas a couple years ago DATE), bay area sports (except the sharks), online gaming, good food, and keeping up with current events. my friends would describe me as easy-going, goofy, thoughtful, consistent, honest, and dependable. i can get serious when i have to be. overall, i like to think that i lead my life by doing what is "right"._i'm a civil engineer, who enjoys helping the citizens of san san francisco._- looking at things objectively - getting things done (but others may disagree) =p - remaining calm in any situation_i'm ORG quiet until i get used to the environment (but isn't that normal?)._last book: "game change". movies: bourne series, action, smart comedies (in contrast to slapstick), must have a decent plot shows: local sports, ncis, good wife, top chef, currently exploring japanese NORP dramas music: koit... food: korean bbq, japanese NORP , chinese NORP , thai NORP , italian, burgers, almost anything_- iphone - friends and family - internet - bay area sports - humor - point & shoot camera_aside from work, how to improve my home._out enjoying friendly conversation over dinner._please let me think about this more._we have similar interests.

HOBBY ANALYSIS

- What are users interested in, and are they serious about the app?
- Benefits/problems identified
  - Marks important and relevant interests
  - The algorithm could be improved for more accurate identification

'acting'	'travel'	'traveling'	'traveling'
'acting'	'arts'	'drama'	'games'
'gaming'	'internet'	'shooting'	'sports'
'traveling'			

SENTIMENT & SUBJECTIVITY

Step 1

Apply TextBlob to each column to gather sentiment & subjectivity scores

Step 2

Calculate the average sentiment & subjectivity scores

Step 3

Create two columns in the dataset to save each user's scores

04

CLUSTER ANALYSIS

PCA/t-SNE & K-Means

Step 1

Standardize numeric features in the dataset

Step 2

Implement **PCA** & get components for 95% explained variance ratio

Implement **t-SNE** & get back 2 output columns/variables

Step 3

Record the inertia and silhouette scores for different K's

Step 4

Select the optimal K value for K-Means & examine each cluster's descriptive statistics

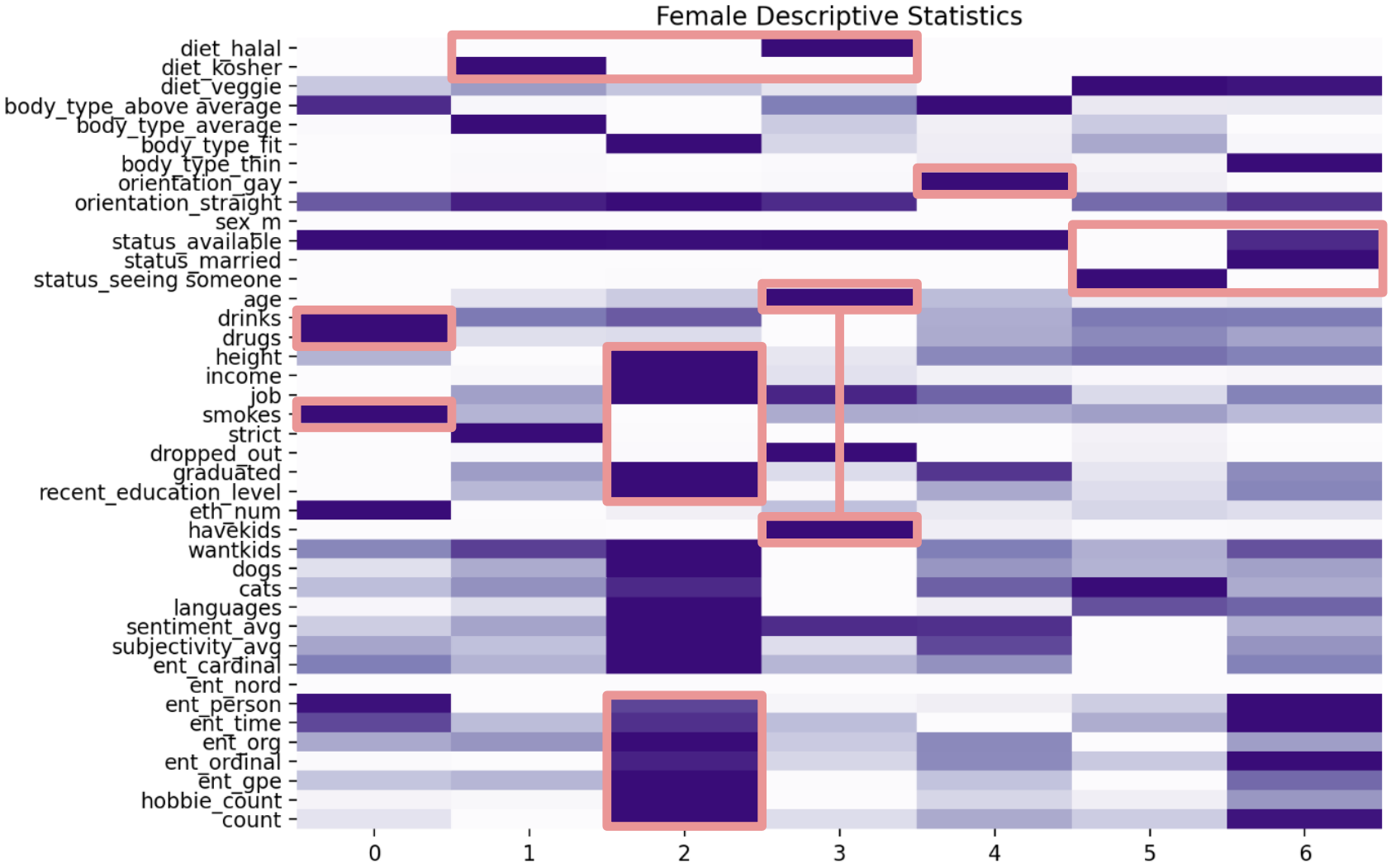
K-MEANS CLUSTERING

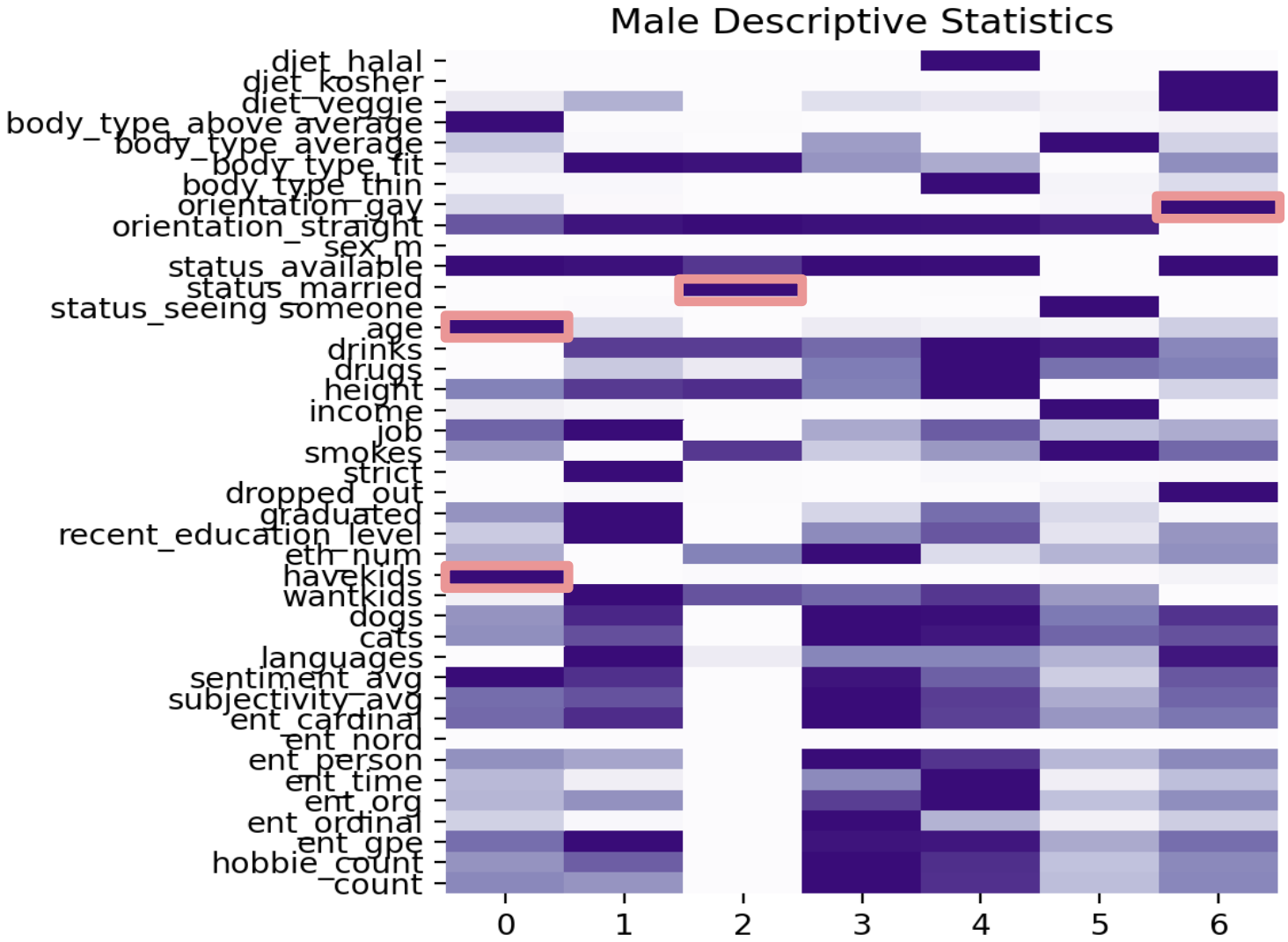
Male

Female

Below 30

Above 30


Female Descriptive Statistics
diet_halal
diet_kosher
diet_veggie
body_type_above average
body_type_average
body_type_fit
body_type_thin
orientation_gay
orientation_straight
sex_m
status_available
status_married
status_seeing someone
age
drinks
drugs
height
income
job
smokes
strict
dropped_out
graduated
recent_education_level
eth_num
havekids
wantkids
dogs
cats
languages
sentiment_avg
subjectivity_avg
ent_cardinal
ent_nord
ent_person
ent_time
ent_org
ent_ordinal
ent_gpe
hobbie_count
count
0
1
2
3
4
5
6


Male Descriptive Statistics
diet_halal
diet_kosher
diet_veggie
body_type_above_average
body_type_average
body_type_fit
body_type_thin
orientation_gay
orientation_straight
sex_m
status_available
status_married
status_seeing_someone
age
drinks
drugs
height
income
job
smokes
strict
dropped_out
graduated
recent_education_level
eth_num
havekids
wantkids
dogs
cats
languages
sentiment_avg
subjectivity_avg
ent_cardinal
ent_nord
ent_person
ent_time
ent_org
ent_ordinal
ent_gpe
hobbie_count
count
0
1
2
3
4
5
6


AgeBelow30 Descriptive Statistics
diet_halal
diet_kosher
diet_veggie
body_type_above average
body_type_average
body_type_fit
body_type_thin
orientation_gay
orientation_straight
sex_m
status_available
status_married
status_seeing someone
age
drinks
drugs
height
income
job
smokes
strict
dropped_out
graduated
recent_education_level
eth_num
havekids
wantkids
dogs
cats
languages
sentiment_avg
subjectivity_avg
ent_cardinal
ent_nord
ent_person
ent_time
ent_org
ent_ordinal
ent_gpe
hobbie_count
count
0
1
2
3
4
5
6


AgeAbove30 Descriptive Statistics
diet_halal
diet_kosher
diet_veggie
body_type_above average
body_type_average
body_type_fit
body_type_thin
orientation_gay
orientation_straight
sex_m
status_available
status_married
status_seeing someone
age
drinks
drugs
height
income
job
smokes
strict
dropped_out
graduated
recent_education_level
eth_num
havekids
wantkids
dogs
cats
languages
sentiment_avg
subjectivity_avg
ent_cardinal
ent_nord
ent_person
ent_time
ent_org
ent_ordinal
ent_gpe
hobbie_count
count
0
1
2
3
4
5
6

05

INSIGHTS & CHALLENGES

BUSINESS VALUE FOR okc

Internal User Rating

- Detect scam accounts
- Identify serious match seekers based on written answers

Recommendation System

- Recommend profiles to users based on their swiping behaviors
- Recommend activities based on the users' characteristics

Filter Optimization

- Combine structured information and essays to create intelligent filters
- Generate prompts to inspire profile updates

CHALLENGES

Missing Values

Users chose to not fill out certain fields when creating their profiles.

1

2

Mixed Data Types

Future work:

- Explore alternative techniques, e.g., multiple correspondence analysis, factor analysis, etc.

Computational Cost

The computational cost of creating a distance matrix is large.

3

THANK YOU
&
HAPPY SWIPING