# TalkingData

## Predicting consumer demographics

Vivian Wu

# Problem

- Businesses need to be smart to survive in the competitive business world
- Have to, at the bare minimum, be mindful of:
  - Knowing what their clients need
  - Being able to meet that demand
  - Attracting new clients
- Need to properly interpret and make use of the data

# Project Description

- TalkingData
  - China's largest data intelligence solution provider
  - Founded in 2011

- Using their data, predict a mobile user's demographic group (gender & age group)

# Data Acquisition

- Data acquired from a Kaggle competition
- Data source comprised of 8 separate CSV files:
  - app_events.csv
  - app_labels.csv
  - events.csv
  - gender_age_test.csv
  - gender_age_train.csv
  - label_categories.csv
  - phone_brand_device_model.csv
  - sample_submission.csv

# Data Wrangling

- Most of the datasets were too large to process on a single machine
- Dask
  - Large datasets spread over multiple nodes
  - Enables parallel computation
  - Full support for Python (unlike Spark, a popular alternative)

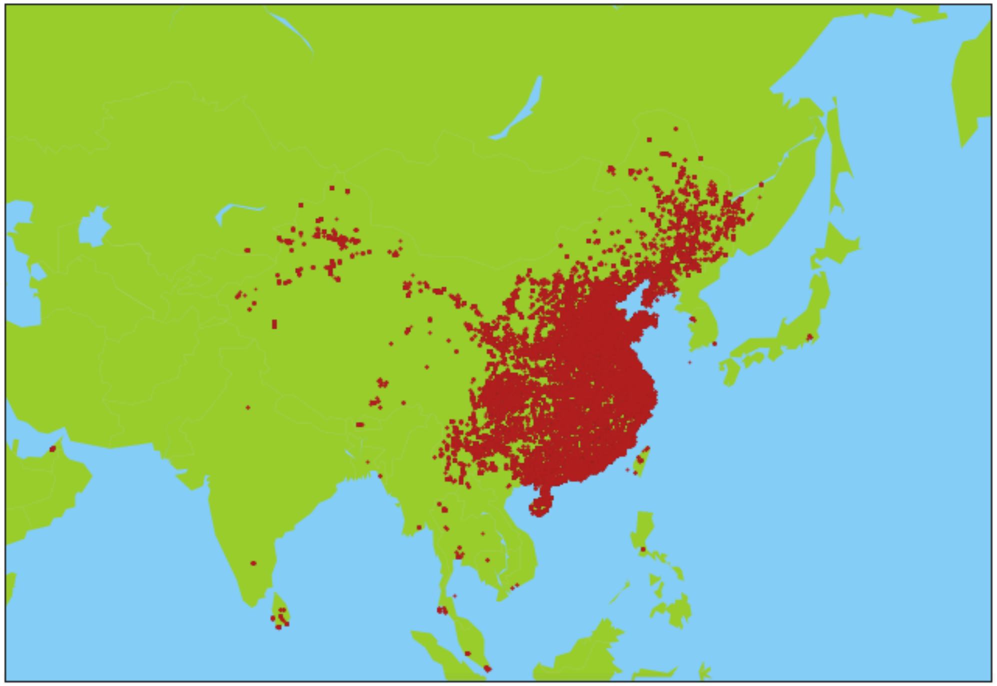
```
app_events = dd.read_csv('talkingdata/app_events.csv')
app_labels = dd.read_csv('talkingdata/app_labels.csv')
events = dd.read_csv('talkingdata/events.csv')
train = dd.read_csv('talkingdata/gender_age_train.csv')
label_categories = dd.read_csv('talkingdata/label_categories.csv')
phone_brand_model = dd.read_csv('talkingdata/phone_brand_device_model.csv')
```

# Data Wrangling

- No missing values
- Create new column: "app_count"
  - Only attainable through merging multiple tables
  - Grouped by "device_id" after each merge to remove duplicates and get app count

| | device_id | gender | age | group | app_count | phone_brand | device_model |
|---|---|---|---|---|---|---|---|
| **0** | -8260683887967679142 | M | 35 | M32-38 | 53 | 小米 | MI 2 |
| **1** | 7477216237379271436 | F | 37 | F33-42 | 26 | 华为 | 荣耀6 plus |
| **2** | 6352067998666467520 | M | 32 | M32-38 | 19 | 华为 | 荣耀畅玩4X |
| **3** | 8026504930081700361 | M | 25 | M23-26 | 31 | 小米 | MI 4 |
| **4** | -7271319853104672050 | M | 27 | M27-28 | 34 | 三星 | Galaxy Note 3 |

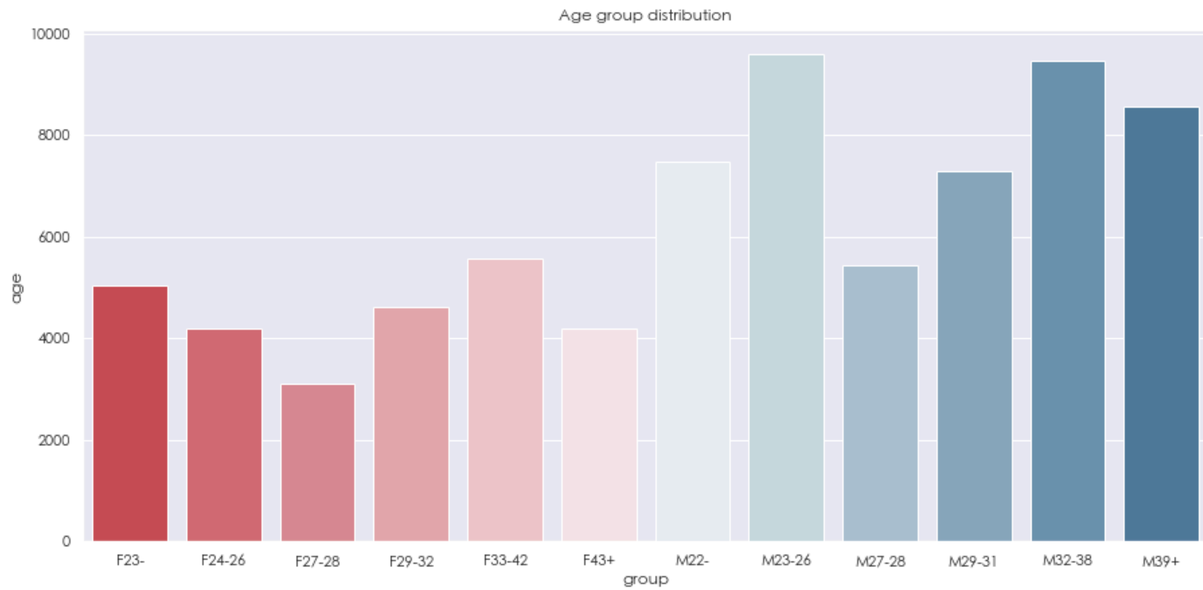
# Exploratory Data Analysis



- Majority of data gathered from eastern region of China
  - 5 most populous cities (Shanghai, Beijing, Tianjin, Shenzhen, Guangzhou) are located in eastern China

# Exploratory Data Analysis

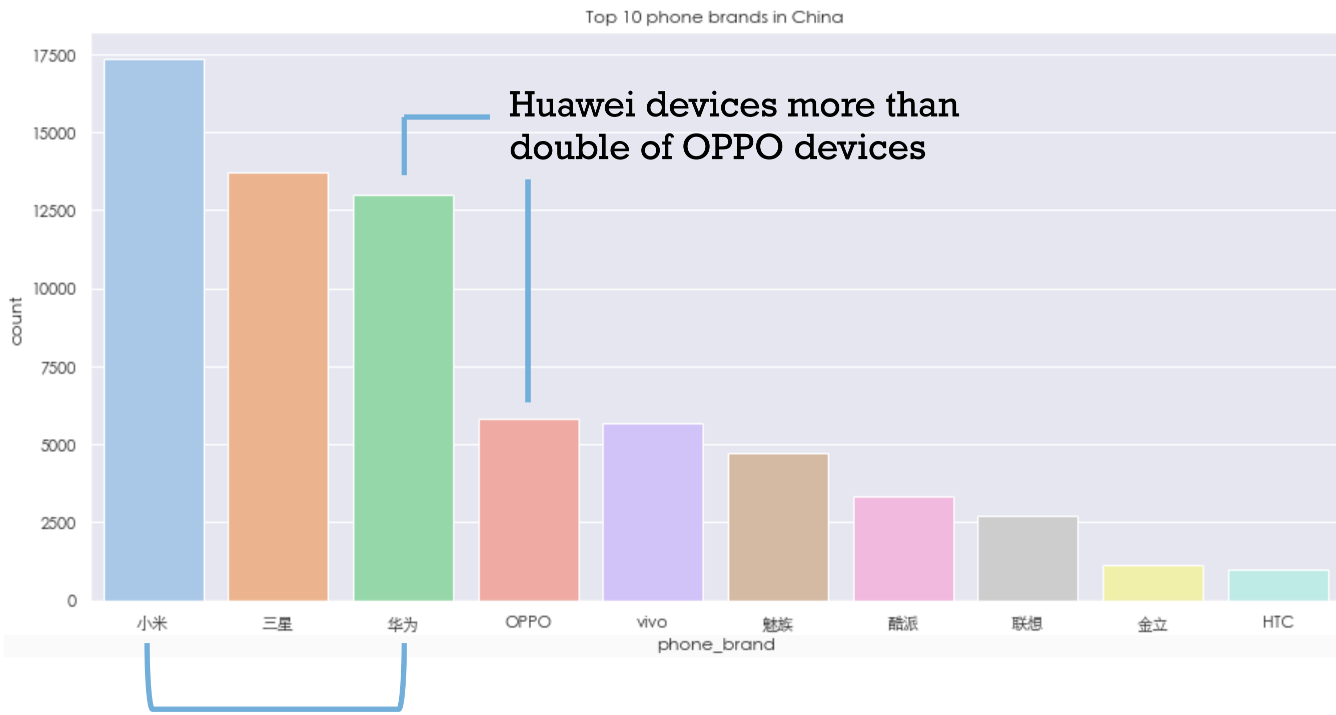
Bar chart of age group distribution




- Top two age group (female): under 23 and 33-42
- Top two age group (male): 23-26 and 32-38
- Gender imbalance in dataset, but *is* representative of the Chinese population because male population outnumbers female population by at least 30 million

# Exploratory Data Analysis

- Top 10 mobile phone brands in China

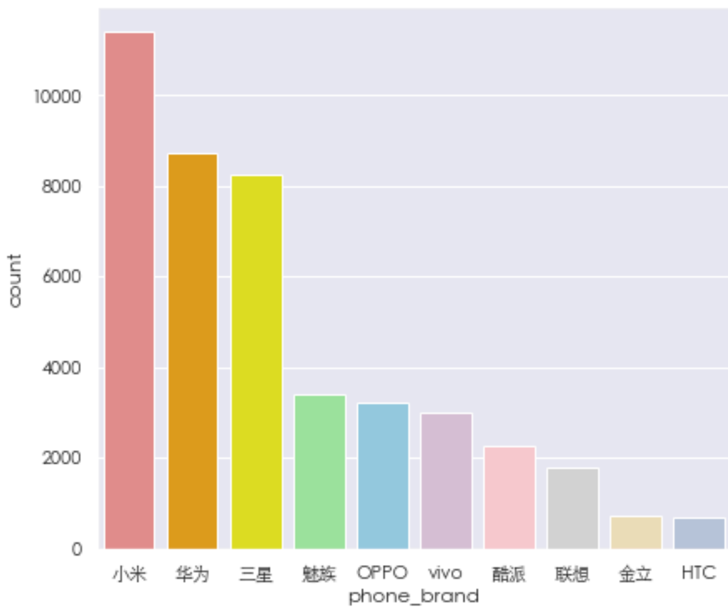


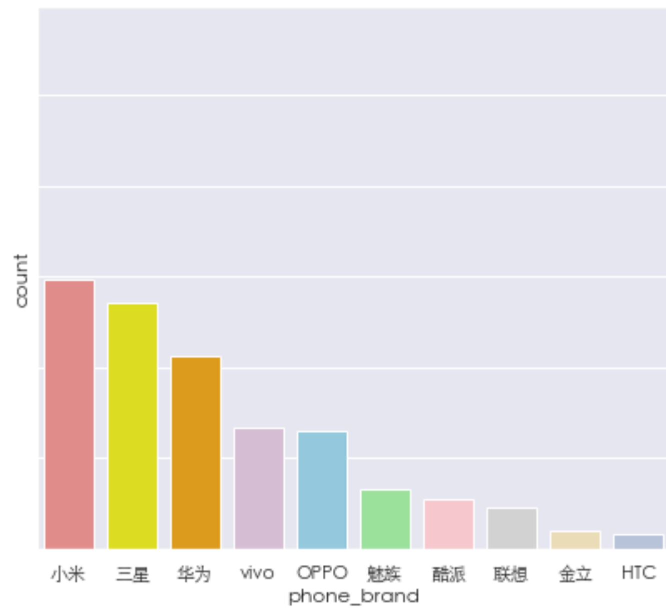

Leading by a staggering amount

| | phone_brand | count |
|---|---|---|
| 47 | Xiaomi | 17336 |
| 14 | Samsung | 13706 |
| 29 | Huawei | 13001 |
| 6 | OPPO | 5802 |
| 12 | vivo | 5658 |
| 117 | Meizu | 4710 |
| 106 | Coolpad | 3349 |
| 91 | Lenovo | 2695 |
| 109 | Gionee | 1124 |
| 1 | HTC | 1015 |

# Exploratory Data Analysis

- Chinese men and women have slightly different preferences when it comes to mobile devices

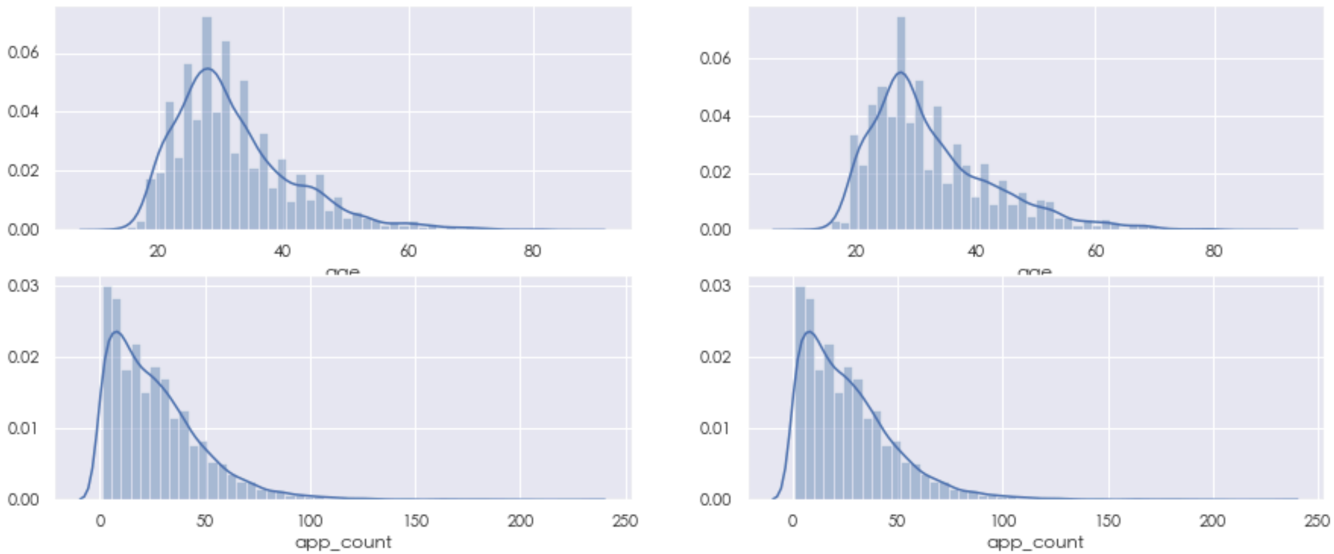



- Huawei & Meizu devices more popular to men

- Chinese women more likely to get a Samsung over Huawei

# Inferential Statistics

- Most, if not all, features in this dataset have skewed distributions
  - 

- Mann-Whitney U test
  - Tests whether two samples are likely to derive from the same population

# Inferential Statistics

- Null hypotheses:
  - The distributions of age between genders are the same
  - The distributions of number of apps per device by gender are the same

- Significance level: 0.05

```
Mann Whitney U p-value (age): 0.05365886631268518
```

1. Distributions of age between genders are the same
   - P-value is slightly larger than 0.05
   - Fail to reject null hypothesis

```
Mann Whitney U p-value (number of apps): 1.4108724351737035e-13
```

2. Distributions of number of apps are *not* the same
   - P-value is significantly smaller than 0.05
   - Reject null hypothesis

# Machine Learning

- Random Forest Classifier
- Combine brand and model into one column

| phone_brand | device_model |
|---|---|
| 小米 | MI 2 |
| 华为 | 荣耀6 plus |
| 华为 | 荣耀畅玩4X |
| 小米 | MI 4 |
| 三星 | Galaxy Note 3 |

| phone |
|---|
| 小米 MI 2 |
| 华为 荣耀6 plus |
| 华为 荣耀畅玩4X |
| 小米 MI 4 |
| 三星 Galaxy Note 3 |

  - Men and women seem to have slightly different preferences in mobile phone brands
  - Use phone to predict demographic group

# Machine Learning

- Machine learning algorithms have difficulty working with categorical data
  - Use a method called Feature Hashing to convert each unique phone value to a hash value

| | app_count | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | ... | 880 | 881 | 882 | 883 | 884 | 885 | 886 | 887 | 888 | 889 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 53 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1** | 26 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **2** | 19 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **3** | 31 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **4** | 34 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

# Machine Learning

- Result of model performance

```
RANDOM FOREST CLASSIFIER
------------------------
F1 score: 0.07346938775510203
```

- F1 score is a metric that measures accuracy
  - F1 score close to 1 indicates predictions are quite accurate
  - Looking at just the kind of phone and number of apps is not enough

# Conclusion

- Poor model performance
- Ideas to improve performance (NEXT STEPS):
  - Utilize text mining and/or Natural Language Processing (NLP) techniques to analyze mobile app label categories
    - 900+ unique label categories in dataset
    - Different gender/age group possibly interested in different kinds of mobile apps
    - App categories should improve model performance
  - Analyze app usage based on time of day