



# MOBILE APP TREND

**Google Play Store App Data Study from 2010 to 2018**

**Yan Yuet Jasmine Lam**



# CONTENTS

1. Background
2. Questions and Hypothesis
3. Approach and Analysis
4. Methodologies
5. Q1: Which category has the highest rating mean?
6. Q2: Which category has the highest number of installs?
7. Q3: Is there any relationship between high rating and number of installs?
8. Q4: Which app has: the highest rating, the highest reviews, the highest installs?
9. Q5: What is the trend of app built over time?
10. Deeper Analysis: Which machine learning model
11. Results



# BACKGROUND

**Why should we look at the trend of the play store app?**

- By 2025, approximately 7.5 billion people are mobile users**
- We spent most of the time in apps when we are using mobile**
- The mobile app revenue is projected to reach £501 billion by 2025, according to Statista**
- Mobile app development becomes important to generate revenue and build connections to customers for businesses**



# BACKGROUND

## Why Google Play Store?

- Google Play Store is the official app store in Android Market. Most users are required to install applications through Google Play Store.
- Google Play Store is available in various platform, such as Android, Chrome OS.



# QUESTIONS AND HYPOTHESIS

1. Which category has the highest rating mean?

*Hypothesis: Tools or productivity apps. Because the apps made should be specified to some customers' need. The apps would be beneficial to a specific group of installers.*

2. Which category has the highest number of installs?

*Hypothesis: Communication apps such as WhatsApp.*

3. Is there any relationship between high rating and number of installs?

*Hypothesis: Yes, high rating should increase the number of installs.*



# QUESTIONS AND HYPOTHESIS

4. Which app has highest rating, highest reviews and highest installs respectively?
5. What is the trend of app built over time?
6. Which machine learning model fits the most to predict the most successful app?



# APPROACH AND ANALYSIS

1. Raw data collected on Kaggle (year of dataset: 2010-2018)

<https://www.kaggle.com/datasets/lava18/google-play-store-apps>

2. Apply pandas in Python to perform data wrangling, data cleaning

3. Apply seaborn, matplotlib in Python to explore and visualise data

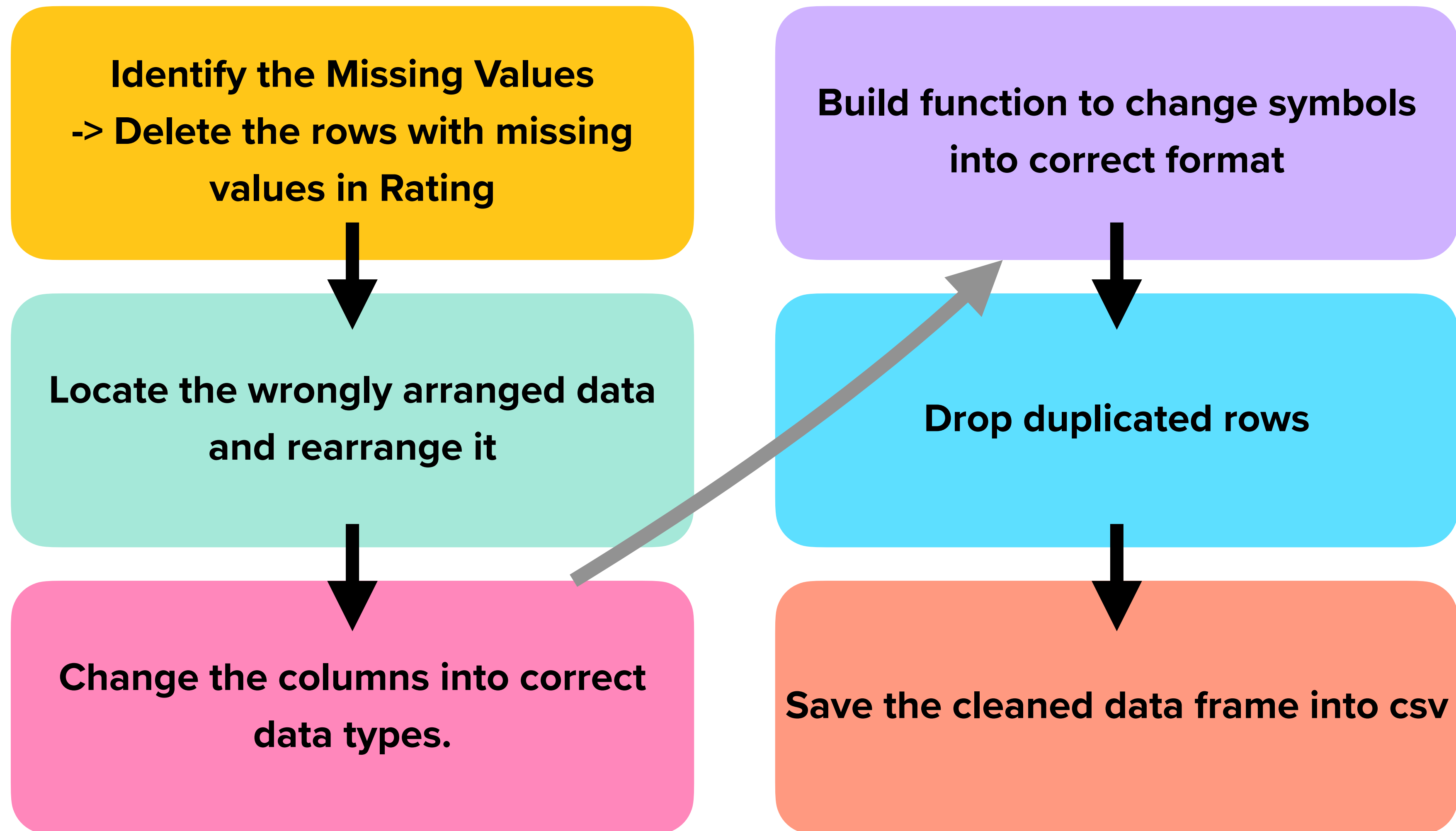
4. Apply Tableau to create interactive visual analytics

5. Utilise regression and classification models to find the best machine learning model





# METHODOLOGIES: DATA CLEANING







# METHODOLOGIES: EXPLORATORY DATA ANALYSIS AND VISUALISATION

**1.**

Change the specific columns into indicator variables



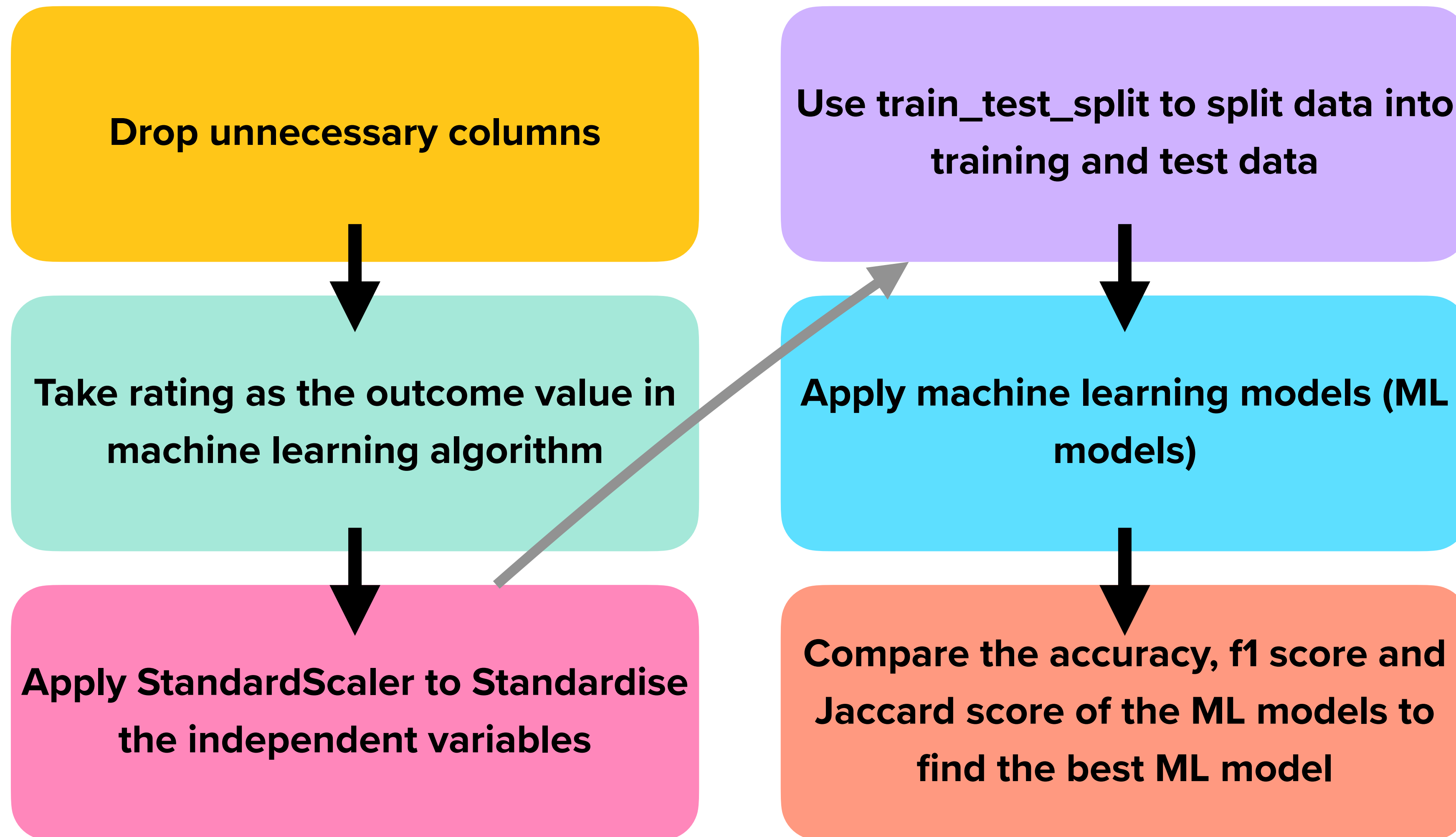
Use bar chart, box plot and scatter plot to explore the data

**2.**

Apply Tableau to create interactive data analytics

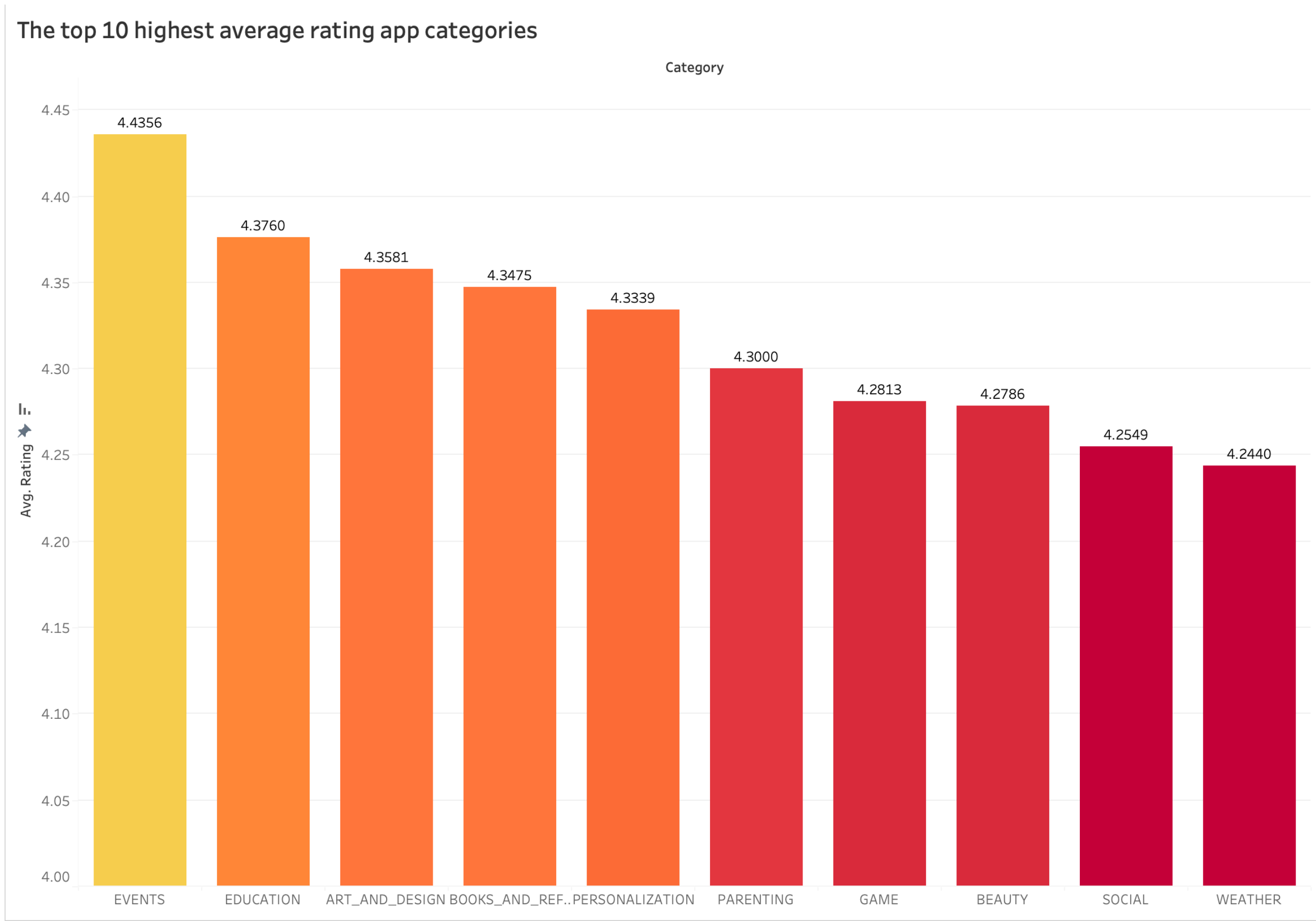


# METHODOLOGIES: MACHINE LEARNING





# Q1: WHICH CATEGORY HAS THE HIGHEST RATING MEAN?

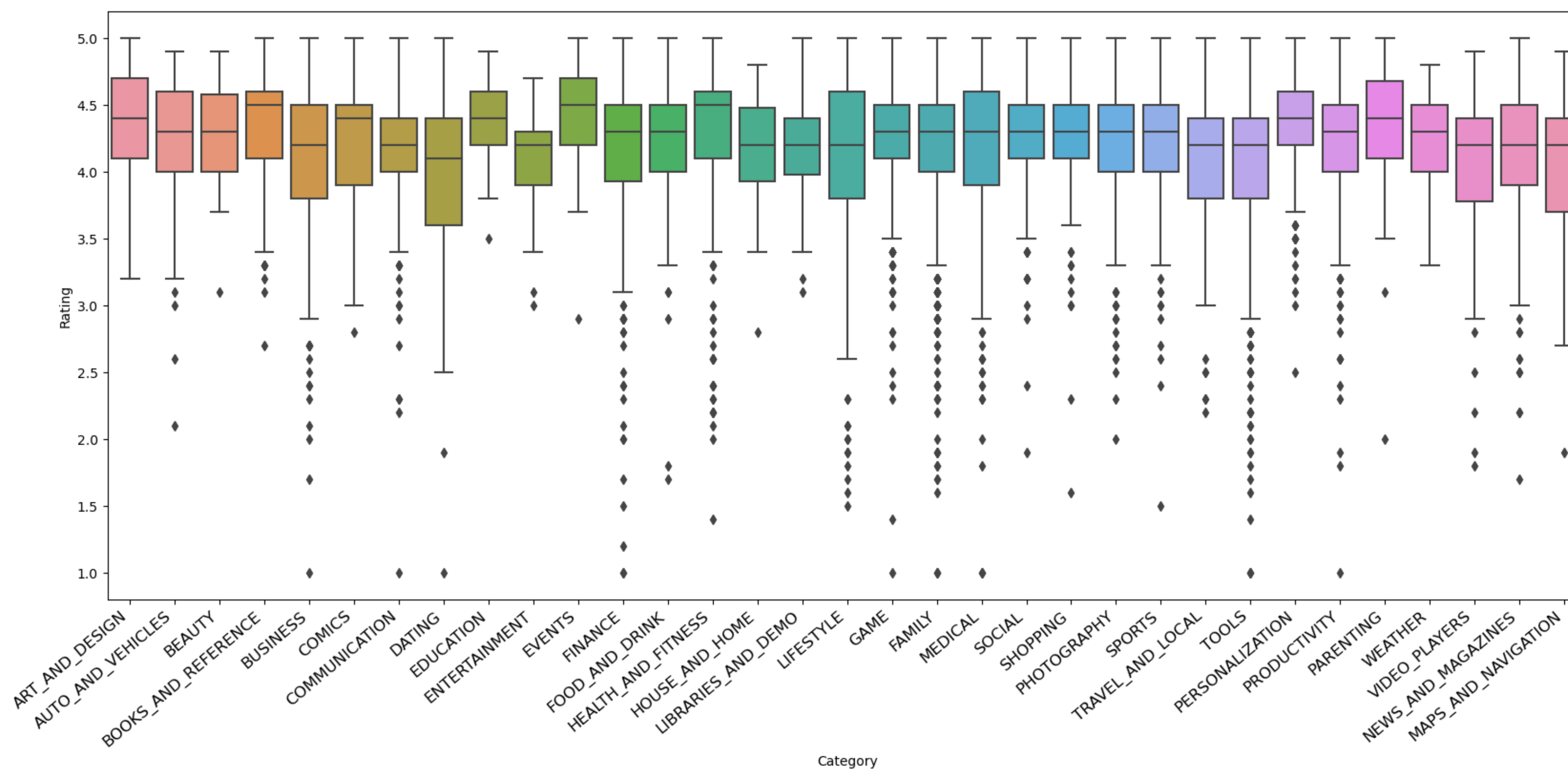


- The bar chart shows the top 10 categories with the highest rating mean
- Category 'Events' has the highest rating mean 4.44



# Q1: WHICH CATEGORY HAS THE HIGHEST RATING MEAN?

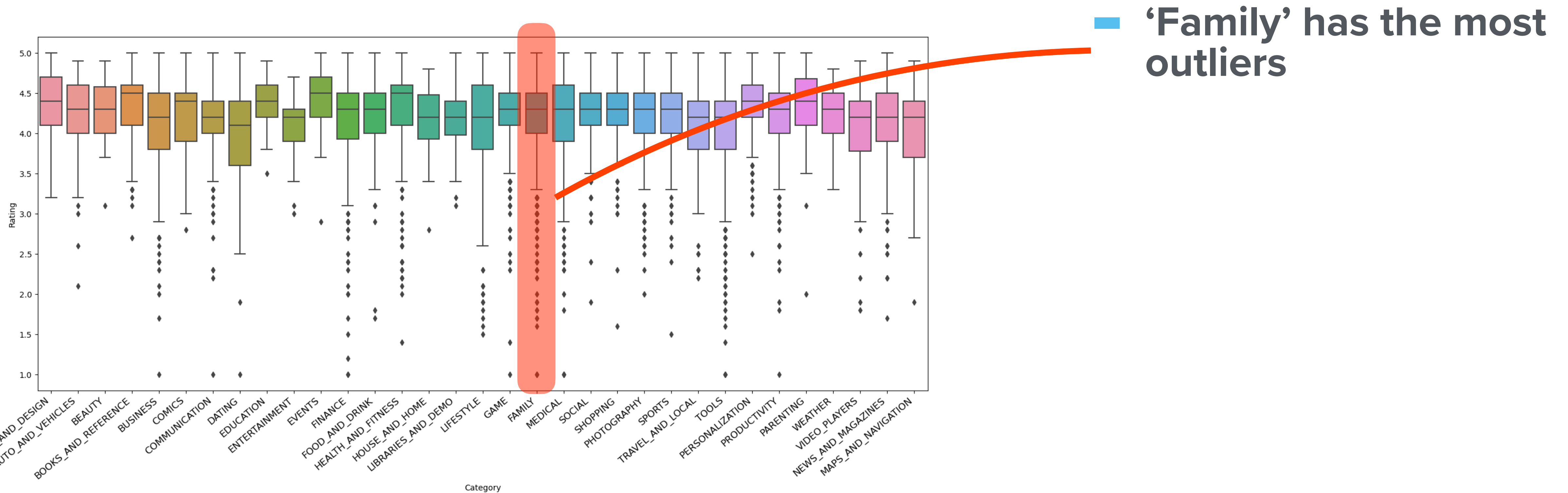
Box plot of the rating in each app category





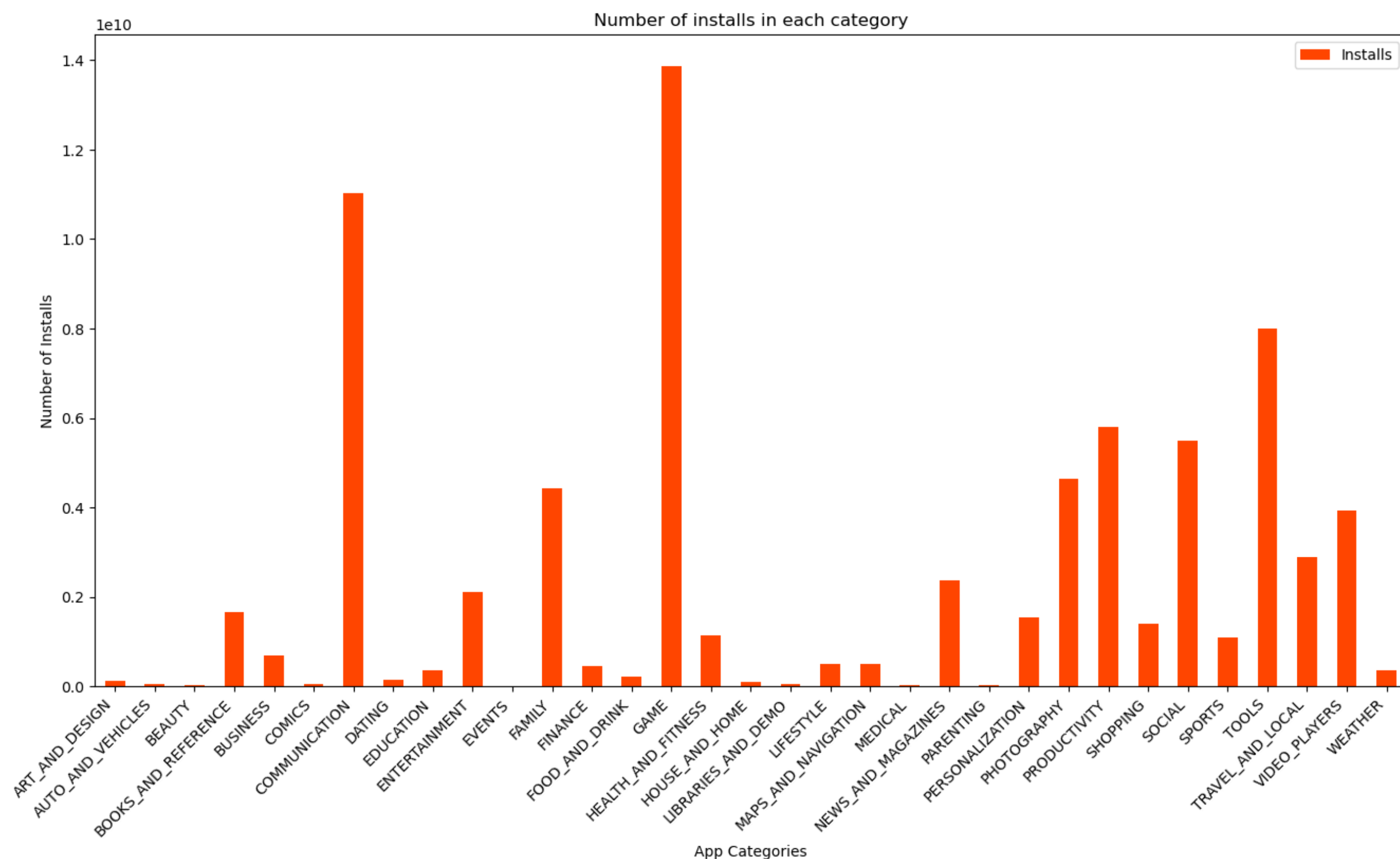
# Q1: WHICH CATEGORY HAS THE HIGHEST RATING MEAN?

Box plot of the rating in each app category





## Q2: WHICH CATEGORY HAS THE HIGHEST NUMBER OF INSTALLS?



- The bar chart shows the number of installs in each category
- It shows that the category 'Game' has the highest number of installs
- The number of installs in 'Game' category is **13878924415**

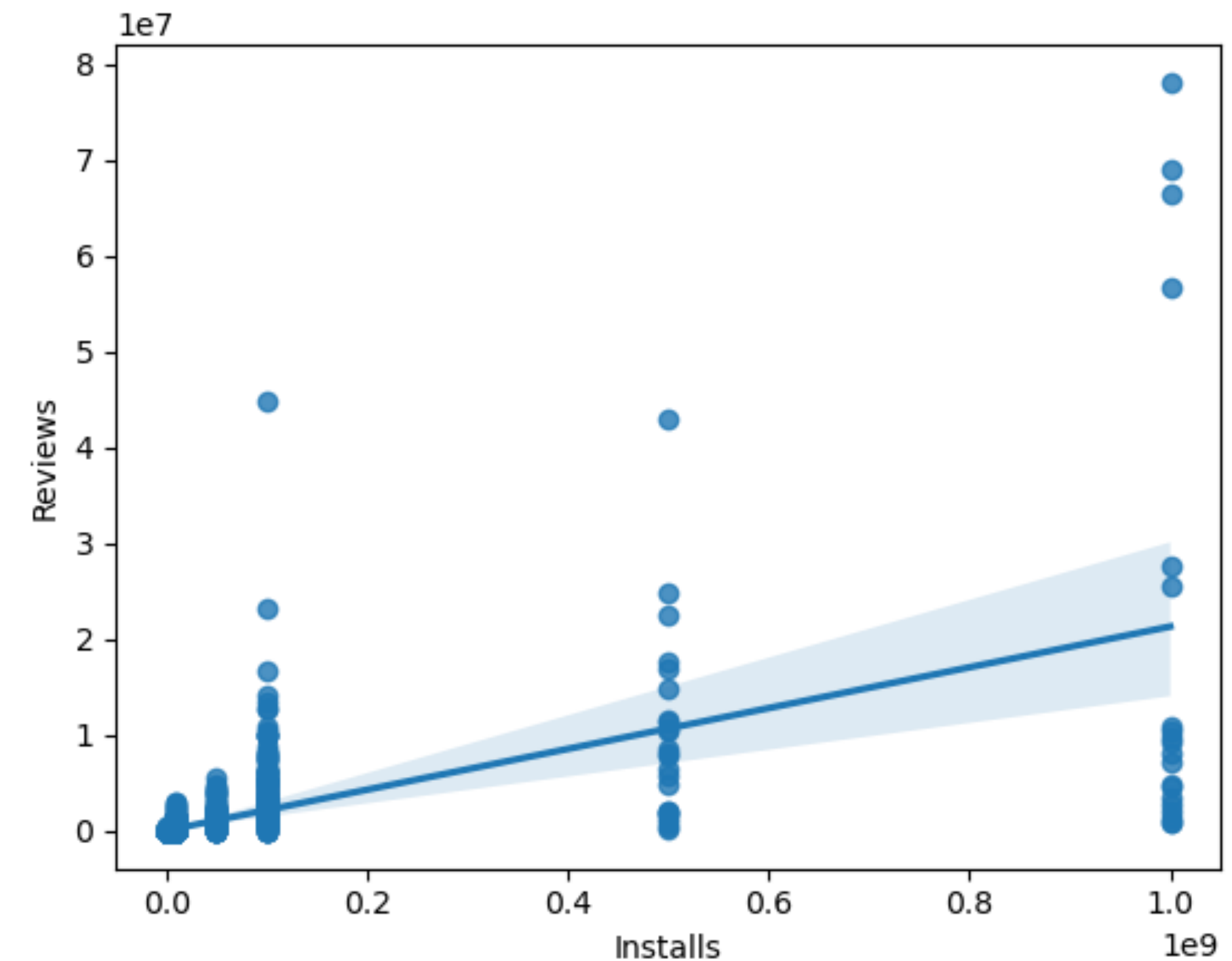




## Q3: IS THERE ANY RELATIONSHIP BETWEEN HIGH RATING AND NUMBER OF INSTALLS?

	Rating	Reviews	Size	Installs	Price
Rating	1.000000	0.055090	0.037936	0.040181	-0.021144
Reviews	0.055090	1.000000	0.037894	0.625149	-0.007599
Size	0.037936	0.037894	1.000000	-0.007466	-0.015048
Installs	0.040181	0.625149	-0.007466	1.000000	-0.009404
Price	-0.021144	-0.007599	-0.015048	-0.009404	1.000000

- The table shows the correlations between different columns
- It is observed that the number of installs and rating have a positive relationship, but they are not the most correlated
- ‘Installs’ and ‘Reviews’ have a significant positive relationship instead



- The scatter graph shows the relationship between number of reviews and number of installs





## Q4: WHICH APP HAS: THE HIGHEST RATING, THE HIGHEST REVIEWS, THE HIGHEST INSTALLS?

App	Category	Rating	Reviews	Size	Installs	Price
Ek Bander Ne Kholi Dukan	FAMILY	5.0	10.0	3000000	10000	0.0
CL Keyboard - Myanmar Keyboard (No Ads)	TOOLS	5.0	24.0	3200000	5000	0.0
Oración CX	LIFESTYLE	5.0	103.0	3800000	5000	0.0
Superheroes, Marvel, DC, Comics, TV, Movies News	COMICS	5.0	34.0	12000000	5000	0.0
Hojiboy Tojiboyev Life Hacks	COMICS	5.0	15.0	37000000	1000	0.0

- The table lists the top five apps with the highest rating and installs
- The top rating and installs app is 'Ek Bander Ne Kholi Dukan' in the category 'Family' with **10000** installs
- All of the top five are free apps



## Q4: WHICH APP HAS: THE HIGHEST RATING, THE HIGHEST REVIEWS, THE HIGHEST INSTALLS?

App	Category	Rating	Reviews	Size	Installs	Price
FHR 5-Tier 2.0	MEDICAL	5.0	2.0	1200000	500	2.99
Super Hearing Secret Voices Recorder PRO	MEDICAL	5.0	3.0	23000000	100	2.99
ADS-B Driver	TOOLS	5.0	2.0	6300000	100	1.99
P-Home for KLWP	PERSONALIZATION	5.0	4.0	12000000	100	0.99
Android P Style Icon Pack	PERSONALIZATION	5.0	1.0	60000000	100	0.99

- Now we look at the top five paid apps with the highest rating and installs
- The top rating paid app is 'FHR 5-Tier 2.0' in the category 'Medical' with 500 installs



## Q4: WHICH APP HAS: THE HIGHEST RATING, THE HIGHEST REVIEWS, THE HIGHEST INSTALLS?

App	Category	Rating	Reviews	Size	Installs	Price
Facebook	SOCIAL	4.1	78143257.0	0	1000000000	0.0
WhatsApp Messenger	COMMUNICATION	4.4	69114494.0	0	1000000000	0.0
Instagram	SOCIAL	4.5	66554892.0	0	1000000000	0.0
Messenger – Text and Video Chat for Free	COMMUNICATION	4.0	56644712.5	0	1000000000	0.0
Subway Surfers	GAME	4.5	27721321.2	76000000	1000000000	0.0

- This table shows the top five apps with the highest installs and reviews
- We can see that the top four are the popular social media and communication apps
- The highest installs and review app is 'Facebook'



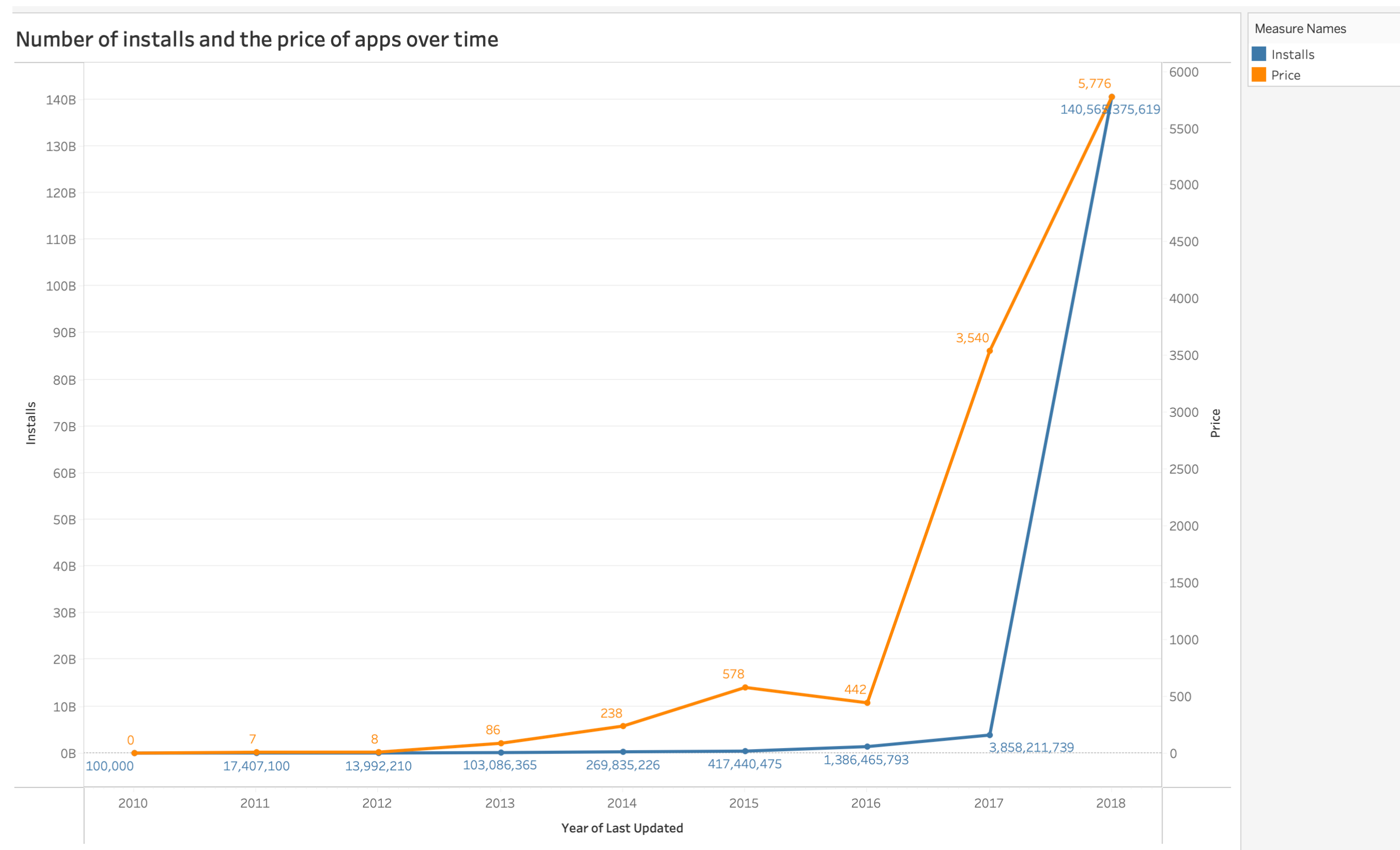
## Q5: WHAT IS THE TREND OF APP BUILT OVER TIME?

Year	Most apps were made in the category	Number of apps made in the corresponding category
2010	Family	1
2011	Tools	6
2012	Family	9
2013	Game	18
2014	Family	39
2015	Family	101
2016	Family	203
2017	Family	438
2018	Family	1024

- From 2014 to 2018, most of the app made are in the category 'Family'



# Q5: WHAT IS THE TREND OF APP BUILT OVER TIME?



- The graph shows the number of installs and the price of apps over time
- The rate of number of installs increase rapidly from 2017 to 2018 with approximately 3500% of increase rate
- The change of app price increase speedily from 2016 to 2018. It increase from a total amount of 442USD in 2016 to 5776 USD in 2018





# DEEPER ANALYSIS: WHICH MACHINE LEARNING MODEL

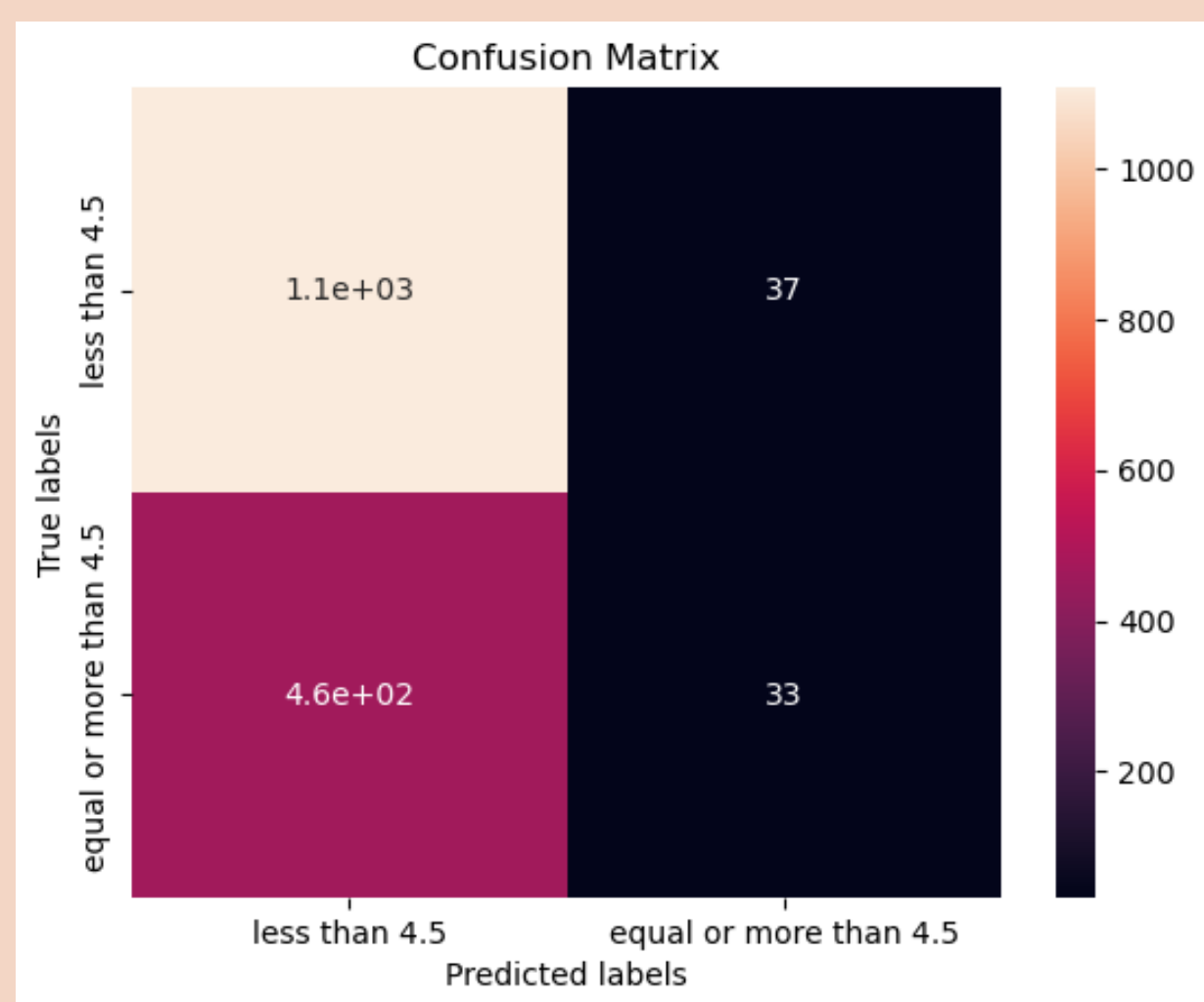
## Multiple Linear Regression Model

- Multiple linear regression model is used to observe how different variables affect the rating
- It is found that the number of reviews has the closest relationship with the rating
- The R-squared score is 0.02, it indicates a weak proportional relationship between rating and other variables
- However, we cannot control the number of reviews and installs. So I apply classification models to predict how to get a successful app (the app with rating greater or equal to 4.5)



# DEEPER ANALYSIS: WHICH MACHINE LEARNING MODEL

## Confusion Matrices of the Classification Models



### DECISION TREE

Best Parameter

Criterion: entropy

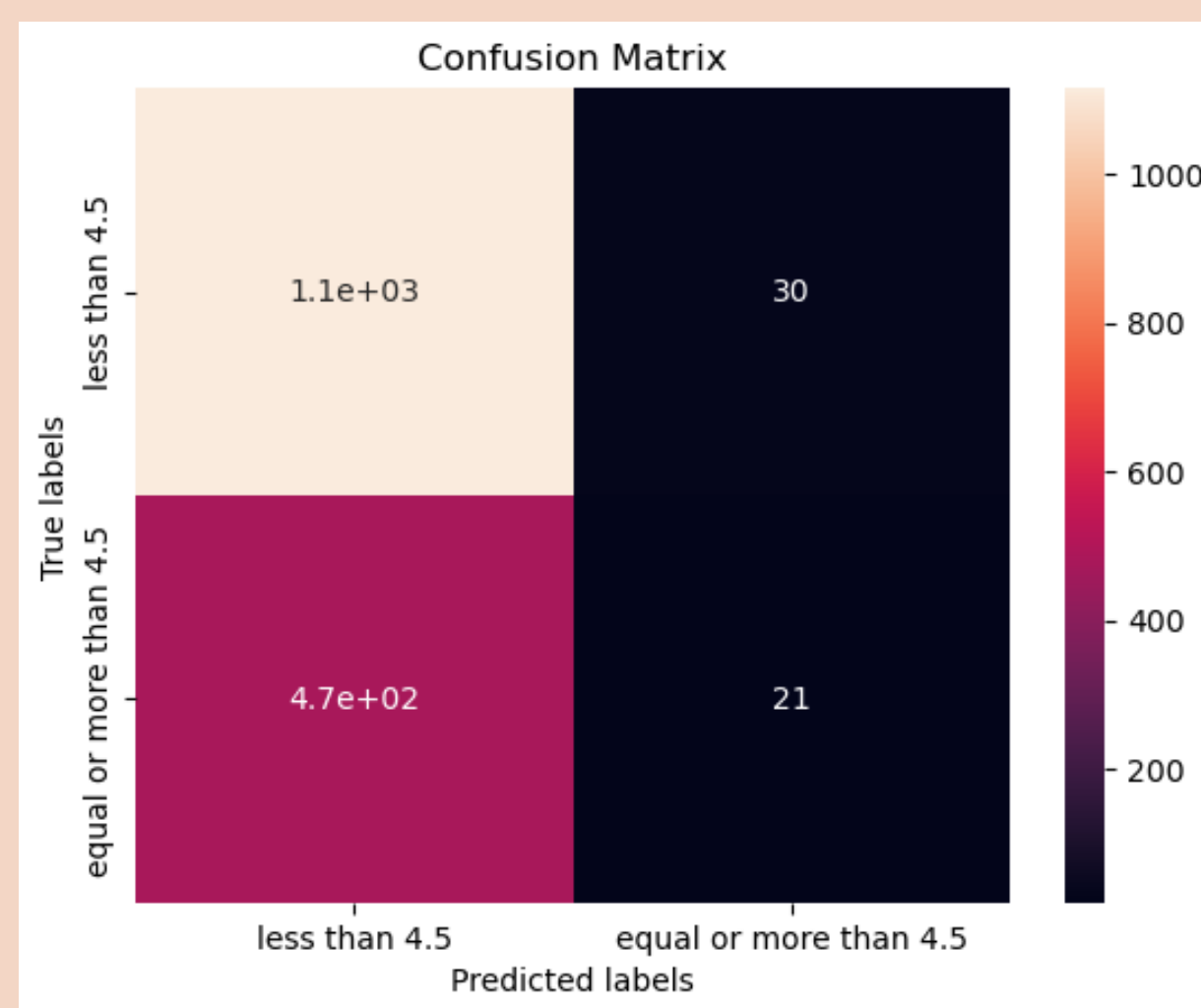
Max depth: 8

Max features: sqrt

Min samples leaf: 2

Min sample split: 2

Splitter: random



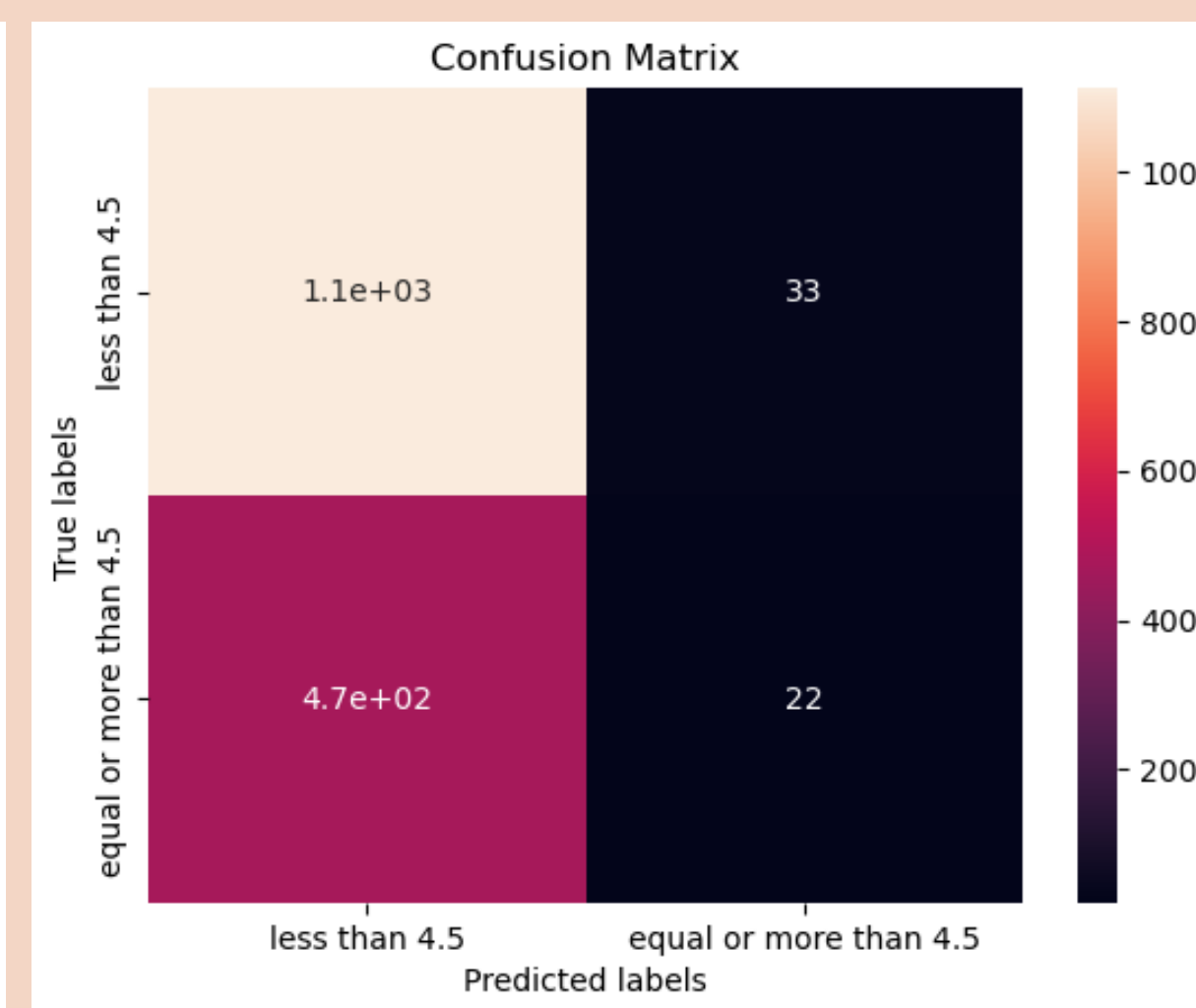
### LOGISTIC REGRESSION

Best Parameter

C: 1

Penalty: l2

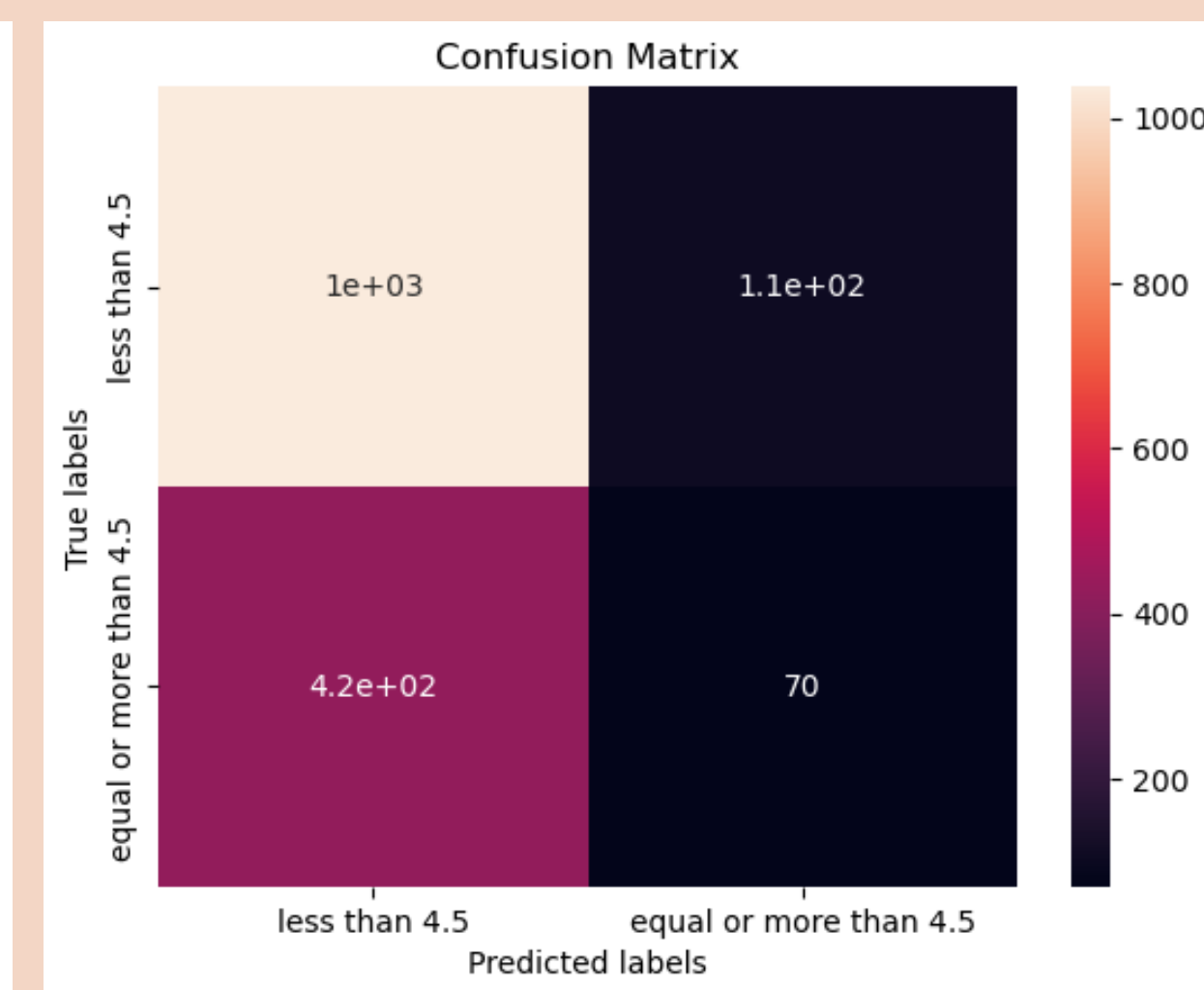
Solver: lbfgs



### SUPPORT VECTOR MACHINES (SVM)

Best Parameter

Kernel: linear



### K NEAREST NEIGHBORS (KNN)

Best Parameter

Algorithm: auto

N\_neighbors: 10

p: 2





# DEEPER ANALYSIS: WHICH MACHINE LEARNING MODEL

## Compare the metrics of different Classification Models


Table of the metrics

	Tree	LogReg	SVM	KNN
Accuracy	0.681098	0.681707	0.680488	0.678049
f1_score	0.592072	0.601191	0.601934	0.612246
Jaccard Score	0.485113	0.489656	0.489416	0.493183

Table of the class distribution

Number of unsuccessful apps in the test set	1135
Number of successful apps in the test set	505

- Logistic Regression has the highest accuracy: 0.682
- KNN has the highest f1 score and Jaccard Score
- KNN f1 score: 0.612
- KNN Jaccard Score: 0.493
- For an uneven class distribution, f1 score is a better metric than accuracy
- Therefore, KNN is chosen to be the best machine learning model



# RESULTS

**1. Which category has the highest rating mean?**

*The 'Events' category has the highest rating mean. While 'Family' category appears to have the most outliers under its rating mean*

**2. Which category has the highest number of installs?**

*The top 3 highest number of installs categories are:  
Game, Communication and Tools.*

**3. Is there any relationship between high rating and number of installs?**

*Yes, but the relationship is insignificant. There is a more significant positive relationship between the number of installs and reviews.*



# RESULTS

## 4. Which app has highest rating, highest reviews and highest installs respectively?

- *The top rating and installs app: 'Ek Bander Ne Kholi Dukan' in category 'Family' with 10000 installs*
- *The top rating paid app is 'FHR 5-Tier 2.0' in the category 'Medical' with 500 installs*
- *The highest installs and review app is 'Facebook' in the category 'Social'*

## 5. What is the trend of app built over time?

- *From 2014 to 2018, most of the app made are in the category 'Family'. The top 3 number of app categories build are: Family, Game and Tools*
- *The rate of number of installs increase rapidly from 2017 to 2018*
- *The change of app price increase speedily from 2016 to 2018*



# RESULTS

## 6. Which machine learning model fits the most to predict the most successful app?

- *In this project, I applied Decision Tree, Logistic Regression, Support Vector Machines and K Nearest Neighbors to train the data.*
- *If the app rating is equal or greater than 4.5, it is a successful app.*
- *KNN model has the highest f1 score and Jaccard Score (f1 score: 0.619, Jaccard Score: 0.497). This indicates KNN is the best classification model in predicting a successful app.*



# RESULTS

## IF WE HAVE TO BUILD A NEW APP, WHAT KIND OF APP IS MORE LIKELY TO SUCCESS?

The figure shows how top rating, top installs and top review app categories overlap.

We can conclude that an app that includes family and social content is more likely to success

