```
# This is formatted as code
```

1. Google Play Store apps and reviews

Mobile apps are everywhere. They are easy to create and can be lucrative. Because of these two factors, more and more apps are being developed. In this notebook, we will do a comprehensive analysis of the Android app market by comparing over ten thousand apps in Google Play across different categories. We'll look for insights in the data to devise strategies to drive growth and retention.

Let's take a look at the data, which consists of two files:

- `apps.csv` : contains all the details of the applications on Google Play. There are 13 features that describe a given app.
- `user_reviews.csv` : contains 100 reviews for each app, most helpful first. The text in each review has been pre-processed and attributed with three new features: Sentiment (Positive, Negative or Neutral), Sentiment Polarity and Sentiment Subjectivity.

In [ ]:
```
# Read in dataset
import pandas as pd
apps_with_duplicates = pd.read_csv("datasets/apps.csv")

# Drop duplicates from apps_with_duplicates
apps = apps_with_duplicates.drop_duplicates(subset = ["App"], keep= False)

# Print the total number of apps
print('Total number of apps in the dataset = ', apps.shape[0])

# Have a look at a random sample of 5 rows
print(apps.head())

# we can see that the "Last Updated" column should be in the proper date format
# and not in what it is in now
```

```
Total number of apps in the dataset =  9659
   Unnamed: 0                                          App  \
0           0     Photo Editor & Candy Camera & Grid & ScrapBook
1           1                                Coloring book moana
2           2  U Launcher Lite – FREE Live Cool Themes, Hide ...
3           3                              Sketch - Draw & Paint
4           4               Pixel Draw - Number Art Coloring Book

         Category  Rating  Reviews  Size     Installs  Type Price  \
0  ART_AND_DESIGN     4.1      159  19.0      10,000+  Free     0
1  ART_AND_DESIGN     3.9      967  14.0     500,000+  Free     0
2  ART_AND_DESIGN     4.7    87510   8.7   5,000,000+  Free     0
3  ART_AND_DESIGN     4.5   215644  25.0  50,000,000+  Free     0
4  ART_AND_DESIGN     4.3      967   2.8     100,000+  Free     0

  Content Rating                     Genres      Last Updated  \
0       Everyone               Art & Design   January 7, 2018
1       Everyone  Art & Design;Pretend Play  January 15, 2018
2       Everyone               Art & Design    August 1, 2018
3           Teen               Art & Design      June 8, 2018
4       Everyone    Art & Design;Creativity     June 20, 2018

          Current Ver   Android Ver
0               1.0.0  4.0.3 and up
1               2.0.0  4.0.3 and up
2               1.2.4  4.0.3 and up
3  Varies with device    4.2 and up
4                 1.1    4.4 and up
```

In [ ]:
```
print(apps.Installs.unique(), '\n')
print(apps.Price.unique(), '\n')
print(apps.info())

# note that Installs and Price contain characters ",","$", and "+" that prevents
# the them from being numerical values
# date is also improperly formatted.
# Columns "Rating","Size","Current ver", and "Android ver" have missing values
```

```
['10,000+' '500,000+' '5,000,000+' '50,000,000+' '100,000+' '50,000+'
 '1,000,000+' '10,000,000+' '5,000+' '100,000,000+' '1,000,000,000+'
 '1,000+' '500,000,000+' '50+' '100+' '500+' '10+' '1+' '5+' '0+' '0']

['0' '$4.99' '$3.99' '$6.99' '$1.49' '$2.99' '$7.99' '$5.99' '$3.49'
 '$1.99' '$9.99' '$7.49' '$0.99' '$9.00' '$5.49' '$10.00' '$24.99'
 '$11.99' '$79.99' '$16.99' '$14.99' '$1.00' '$29.99' '$12.99' '$2.49'
 '$10.99' '$1.50' '$19.99' '$15.99' '$33.99' '$74.99' '$39.99' '$3.95'
 '$4.49' '$1.70' '$8.99' '$2.00' '$3.88' '$25.99' '$399.99' '$17.99'
 '$400.00' '$3.02' '$1.76' '$4.84' '$4.77' '$1.61' '$2.50' '$1.59' '$6.49'
 '$1.29' '$5.00' '$13.99' '$299.99' '$379.99' '$37.99' '$18.99' '$389.99'
 '$19.90' '$8.49' '$1.75' '$14.00' '$4.85' '$46.99' '$109.99' '$154.99'
 '$3.08' '$2.59' '$4.80' '$1.96' '$19.40' '$3.90' '$4.59' '$15.46' '$3.04'
 '$4.29' '$2.60' '$3.28' '$4.60' '$28.99' '$2.95' '$2.90' '$1.97'
 '$200.00' '$89.99' '$2.56' '$30.99' '$3.61' '$394.99' '$1.26' '$1.20'
 '$1.04']

<class 'pandas.core.frame.DataFrame'>
Int64Index: 9659 entries, 0 to 9658
Data columns (total 14 columns):
Unnamed: 0        9659 non-null int64
App               9659 non-null object
Category          9659 non-null object
Rating            8196 non-null float64
Reviews           9659 non-null int64
Size              8432 non-null float64
Installs          9659 non-null object
Type              9659 non-null object
Price             9659 non-null object
Content Rating    9659 non-null object
Genres            9659 non-null object
Last Updated      9659 non-null object
Current Ver       9651 non-null object
Android Ver       9657 non-null object
dtypes: float64(2), int64(2), object(10)
memory usage: 1.1+ MB
None
```

2. Data cleaning

Data cleaning is one of the most essential subtask any data science project. Although it can be a very tedious process, it's worth should never be undermined.

By looking at a random sample of the dataset rows (from the above task), we observe that some entries in the columns like `Installs` and `Price` have a few special characters (`+` `,` `$`) due to the way the numbers have been represented. This prevents the columns from being purely numeric, making it difficult to use them in subsequent future mathematical calculations. Ideally, as their names suggest, we would want these columns to contain only digits from [0-9].

Hence, we now proceed to clean our data. Specifically, the special characters `,` and `+` present in `Installs` column and `$` present in `Price` column need to be removed.

It is also always a good practice to print a summary of your dataframe after completing data cleaning. We will use the `info()` method to acheive this.

In [ ]:
```
# finally found a way to take out all the unnessecary chars in a string
# with a single line

apps["Installs"] = apps["Installs"].str.replace("\+|\,", "")
print(apps["Installs"])
```

```
0          10000
1         500000
2        5000000
3       50000000
4         100000
5          50000
6          50000
7        1000000
8        1000000
9          10000
10       1000000
11       1000000
12      10000000
13        100000
14        100000
15          5000
16        500000
17         10000
18       5000000
19      10000000
20        100000
21        100000
22        500000
23        100000
24         50000
25         10000
26        500000
27        100000
28         10000
29        100000
          ...
9629         100
9630        1000
9631       10000
9632       50000
9633      500000
9634         100
9635      100000
9636       10000
9637        5000
9638        1000
9639          50
9640          10
9641         100
9642       10000
9643         100
9644     5000000
9645        5000
9646       10000
9647       10000
9648      100000
9649        5000
9650      100000
9651        1000
9652         500
9653          10
9654        5000
9655         100
9656        1000
9657        1000
9658    10000000
Name: Installs, Length: 9659, dtype: object
```

In [ ]:
```
apps["Installs"] = apps["Installs"].str.replace("\+|\,|\$", "")
apps["Price"] = apps["Price"].str.replace("\+|\,|\$", "")

# apps = apps.astype({"Installs":int,"Price":float}) -> changes the dtypes
# into something more appropriate

print(apps.Installs.unique(), '\n')
print(apps.Price.unique(), '\n')
print(apps.info())
```

```
['10000' '500000' '5000000' '50000000' '100000' '50000' '1000000'
 '10000000' '5000' '100000000' '1000000000' '1000' '500000000' '50' '100'
 '500' '10' '1' '5' '0']

['0' '4.99' '3.99' '6.99' '1.49' '2.99' '7.99' '5.99' '3.49' '1.99' '9.99'
 '7.49' '0.99' '9.00' '5.49' '10.00' '24.99' '11.99' '79.99' '16.99'
 '14.99' '1.00' '29.99' '12.99' '2.49' '10.99' '1.50' '19.99' '15.99'
 '33.99' '74.99' '39.99' '3.95' '4.49' '1.70' '8.99' '2.00' '3.88' '25.99'
 '399.99' '17.99' '400.00' '3.02' '1.76' '4.84' '4.77' '1.61' '2.50'
 '1.59' '6.49' '1.29' '5.00' '13.99' '299.99' '379.99' '37.99' '18.99'
 '389.99' '19.90' '8.49' '1.75' '14.00' '4.85' '46.99' '109.99' '154.99'
 '3.08' '2.59' '4.80' '1.96' '19.40' '3.90' '4.59' '15.46' '3.04' '4.29'
 '2.60' '3.28' '4.60' '28.99' '2.95' '2.90' '1.97' '200.00' '89.99' '2.56'
 '30.99' '3.61' '394.99' '1.26' '1.20' '1.04']

<class 'pandas.core.frame.DataFrame'>
Int64Index: 9659 entries, 0 to 9658
Data columns (total 14 columns):
Unnamed: 0        9659 non-null int64
App               9659 non-null object
Category          9659 non-null object
Rating            8196 non-null float64
Reviews           9659 non-null int64
Size              8432 non-null float64
Installs          9659 non-null object
Type              9659 non-null object
Price             9659 non-null object
Content Rating    9659 non-null object
Genres            9659 non-null object
Last Updated      9659 non-null object
Current Ver       9651 non-null object
Android Ver       9657 non-null object
dtypes: float64(2), int64(2), object(10)
memory usage: 1.1+ MB
None
```

3. Correcting data types

From the previous task we noticed that `Installs` and `Price` were categorized as `object` data type (and not `int` or `float`) as we would like. This is because these two columns originally had mixed input types: digits and special characters. To know more about Pandas data types, read this.

The four features that we will be working with most frequently henceforth are `Installs` , `Size` , `Rating` and `Price` . While `Size` and `Rating` are both `float` (i.e. purely numerical data types), we still need to work on `Installs` and `Price` to make them numeric.

In [ ]:
```
apps = apps.astype({"Installs":float,"Price":float})

# Checking dtypes of the apps dataframe
print(apps.dtypes)
```

```
Unnamed: 0          int64
App                object
Category           object
Rating            float64
Reviews             int64
Size              float64
Installs          float64
Type               object
Price             float64
Content Rating     object
Genres             object
Last Updated       object
Current Ver        object
Android Ver        object
dtype: object
```

4. Exploring app categories

With more than 1 billion active users in 190 countries around the world, Google Play continues to be an important distribution platform to build a global audience. For businesses to get their apps in front of users, it's important to make them more quickly and easily discoverable on Google Play. To improve the overall search experience, Google has introduced the concept of grouping apps into categories.

This brings us to the following questions:

- Which category has the highest share of (active) apps in the market?
- Is any specific category dominating the market?
- Which categories have the fewest number of apps?

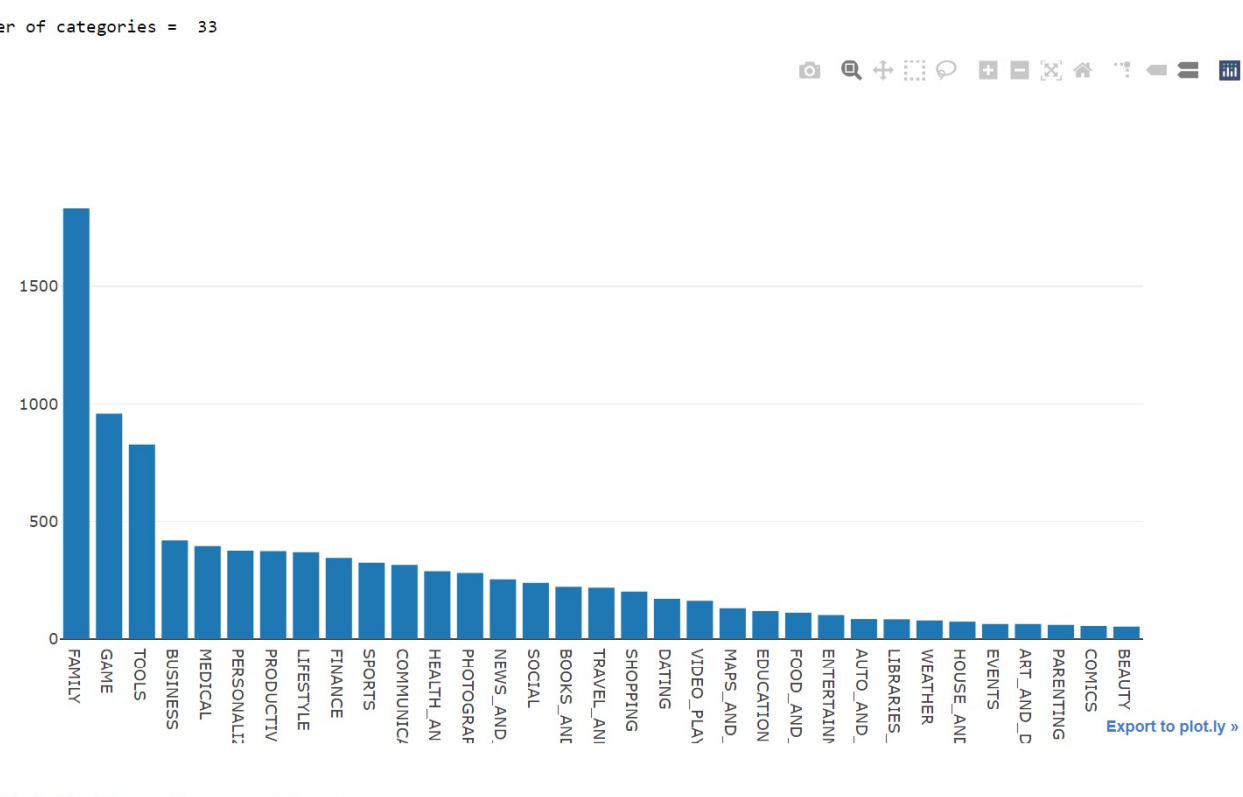
We will see that there are `33` unique app categories present in our dataset. *Family* and *Game* apps have the highest market prevalence. Interestingly, *Tools*, *Business* and *Medical* apps are also at the top.

In [ ]:
```
print(apps.Category.value_counts().sort_values(ascending = False))
```

```
FAMILY                 1832
GAME                    959
TOOLS                   827
BUSINESS                420
MEDICAL                 395
PERSONALIZATION         376
PRODUCTIVITY            374
LIFESTYLE               369
FINANCE                 345
SPORTS                  325
COMMUNICATION           315
HEALTH_AND_FITNESS      288
PHOTOGRAPHY             281
NEWS_AND_MAGAZINES      254
SOCIAL                  239
BOOKS_AND_REFERENCE     222
TRAVEL_AND_LOCAL        219
SHOPPING                202
DATING                  171
VIDEO_PLAYERS           163
MAPS_AND_NAVIGATION     131
EDUCATION               119
FOOD_AND_DRINK          112
ENTERTAINMENT           102
AUTO_AND_VEHICLES        85
LIBRARIES_AND_DEMO       84
WEATHER                  79
HOUSE_AND_HOME           74
EVENTS                   64
ART_AND_DESIGN           64
PARENTING                60
COMICS                   56
BEAUTY                   53
Name: Category, dtype: int64
```

In [ ]:
```
# both imports are really nice since they provide additional functions
# that matplotlib does not seem to have

import plotly
plotly.offline.init_notebook_mode(connected=True)
import plotly.graph_objs as go

# Print the total number of unique categories
num_categories = len(apps["Category"].unique())
print('Number of categories = ', num_categories)

# Count the number of apps in each 'Category'.
num_apps_in_category = apps.Category.value_counts()

# Sort num_apps_in_category in descending order based on the count of apps in each category
sorted_num_apps_in_category = num_apps_in_category.sort_values(ascending = False)

data = [go.Bar(
        x = num_apps_in_category.index, # index = category name
        y = num_apps_in_category.values, # value = count
)]

plotly.offline.iplot(data)
```

```
Number of categories =  33
```



Export to plot.ly »

5. Distribution of app ratings

After having witnessed the market share for each category of apps, let's see how all these apps perform on an average. App ratings (on a scale of 1 to 5) impact the discoverability, conversion of apps as well as the company's overall brand image. Ratings are a key performance indicator of an app.

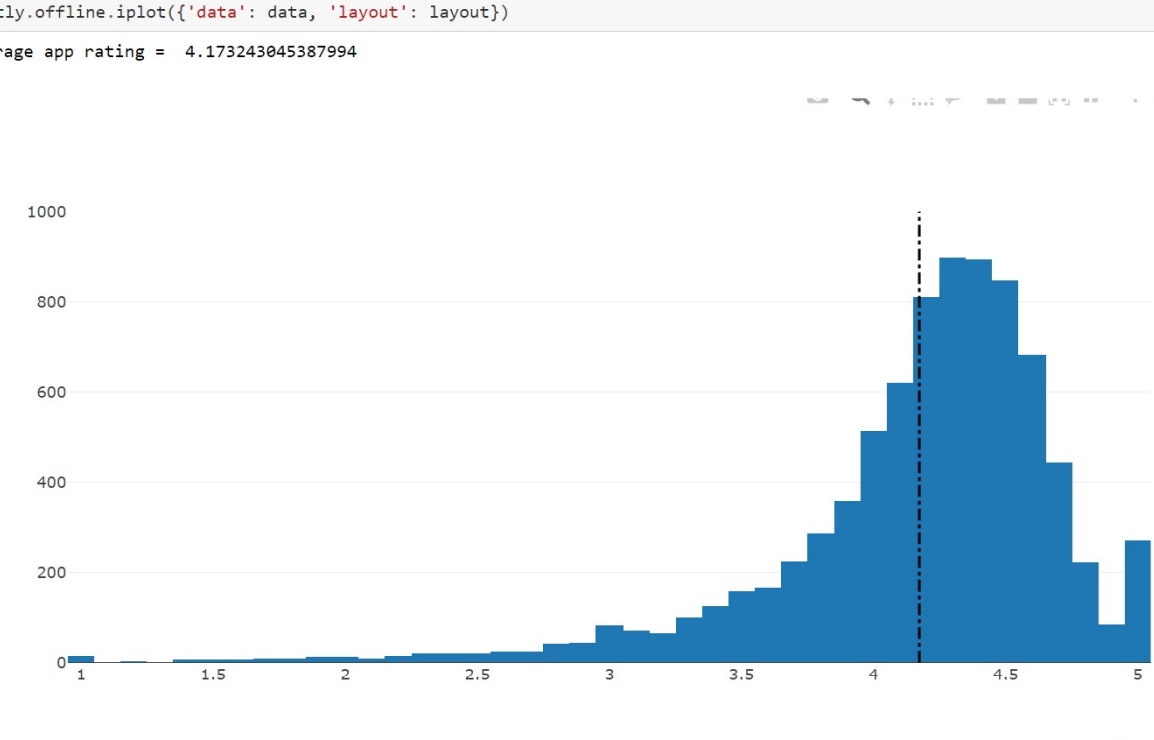
From our research, we found that the average volume of ratings across all app categories is `4.17` . The histogram plot is skewed to the left indicating that the majority of the apps are highly rated with only a few exceptions in the low-rated apps.

In [ ]:
```
print(apps["Rating"].mean())
```

```
4.173243045387994
```

In [ ]:
```
# Average rating of apps
avg_app_rating = apps["Rating"].mean()
print('Average app rating = ', avg_app_rating)

# Distribution of apps according to their ratings
data = [go.Histogram(
        x = apps['Rating']
)]

# Vertical dashed line to indicate the average app rating
# x0 and y0 are the origin of the dashed line
# the y1 and x1 is where the end of the dashed line should be
# so the line goes from (4.17, 0) to (4.17, 1000).
# type defines what the structure is and 'line' specifies what kind of
# line is being made. The dict key 'dash' is type that specifies
# what kind of dashed line it should be which is dashdot
# the shapes key determine what figure could exist on the graph and its
#  value determines the characteristics of that figure. we can add more
# figures in the graph by adding more key:value pairs of figures.
layout = {'shapes': [{
              'type' :'line',
              'x0': avg_app_rating,
              'y0': 0,
              'x1': avg_app_rating,
              'y1': 1000,
              'line': { 'dash': 'dashdot'}
          }]
          }

# data is given the histogram to be graphed and the layout gives the extra
# figures the user has added.
plotly.offline.iplot({'data': data, 'layout': layout})
```

```
Average app rating =  4.173243045387994
```



Export to plot.ly »

6. Size and price of an app

Let's now examine app size and app price. For size, if the mobile app is too large, it may be difficult and/or expensive for users to download. Lengthy download times could turn users off before they even experience your mobile app. Plus, each user's device has a finite amount of disk space. For price, some users expect their apps to be free or inexpensive. These problems compound if the developing world is part of your target market; especially due to internet speeds, earning power and exchange rates.

How can we effectively come up with strategies to size and price our app?

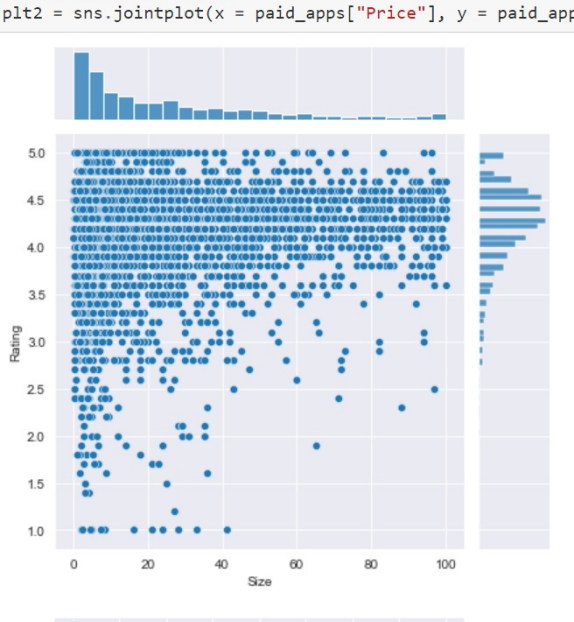
- Does the size of an app affect its rating?
- Do users really care about system-heavy apps or do they prefer light-weighted apps?
- Does the price of an app affect its rating?
- Do users always prefer free apps over paid apps?

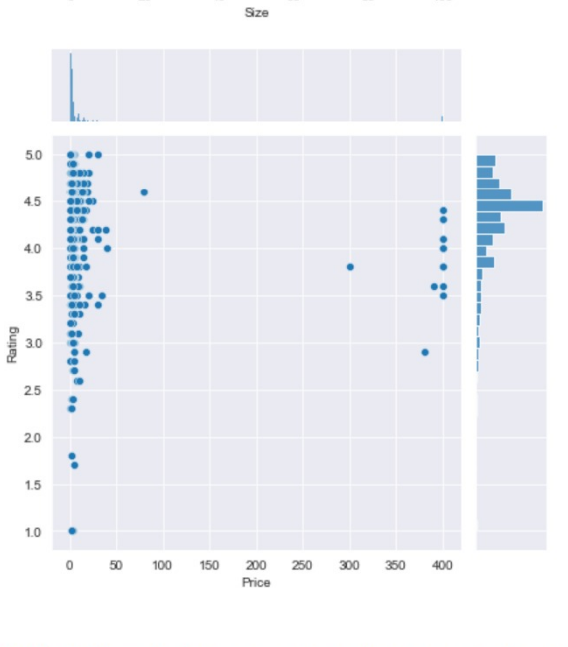
We find that the majority of top rated apps (rating over 4) range from 2 MB to 20 MB. We also find that the vast majority of apps price themselves under $10.

In [ ]:
```
lolipop = apps[(apps["Rating"].isna() == False)&(apps["Size"].isna() == False)]
large_categories = lolipop.groupby("Category").filter(lambda x: len(x) >= 250)

print(lolipop.groupby("Category").filter(lambda x: len(x) >= 250))
```

```
      Unnamed: 0                                            App   Category  \
839         1050                              Citibanamex Movil    FINANCE
841         1052                                    KTB Netbank    FINANCE
842         1053                                Mobile Bancomer    FINANCE
843         1054                                  Nedbank Money    FINANCE
844         1055                                       SCB EASY    FINANCE
848         1059                                      Itau bank    FINANCE
849         1060                                         Nubank    FINANCE
851         1062                                            IKO    FINANCE
852         1063                                       Cash App    FINANCE
853         1064                   Standard Bank / Stanbic Bank    FINANCE
854         1065                              Bualuang mBanking    FINANCE
858         1069                                    MyMo by GSB    FINANCE
860         1071                         Ecobank Mobile Banking    FINANCE
861         1072                                  Banorte Movil    FINANCE
862         1073                             Wells Fargo Mobile    FINANCE
864         1075                                 BZWBK24 mobile    FINANCE
865         1076                                         PayPal    FINANCE
866         1077                            Capital One® Mobile    FINANCE
867         1078                         Zenith Bank Mobile App    FINANCE
869         1080                                      Post Bank    FINANCE
870         1081                                          İşCep    FINANCE
871         1082                                  People's Bank    FINANCE
873         1084                                       Transfer    FINANCE
874         1085                                    T-Mobile in    FINANCE
875         1086                               TrueMoney Wallet    FINANCE
877         1088                                 Bank of Brazil    FINANCE
878         1089                   WiseBanyan - Invest For Free    FINANCE
880         1091                       Wells Fargo Daily Change    FINANCE
881         1092     Even - organize your money, get paid early    FINANCE
882         1093                 Digit Save Money Automatically    FINANCE
...          ...                                            ...        ...
9599       10781                           Modern Strike Online       GAME
9600       10782                        Trine 2: Complete Story       GAME
9601       10783   Modern Counter Terror Attack – Shooting Game       GAME
9602       10784                                     Big Hunter       GAME
9603       10785                                   sugar, sugar     FAMILY
9604       10786                                  ChopAssistant      TOOLS
9605       10787                Modern Counter Global Strike 3D       GAME
9607       10789             Modern Counter Global Strike 3D V2       GAME
9609       10791                              Winter Wonderland       GAME
9610       10792                        Soccer Clubs Logo Quiz       GAME
9611       10793                                      Sid Story       GAME
9613       10795                       Reindeer VPN - Proxy VPN      TOOLS
9614       10796  Inf VPN - Global Proxy & Unlimited Free WIFI VPN    TOOLS
9615       10797                          Fuel Rewards® program  LIFESTYLE
9618       10800                                      FR Roster      TOOLS
9619       10801                            Fr Ignacio Outreach     FAMILY
9620       10802                            FR: My Famous Lover     FAMILY
9621       10803                   Fatal Raid - No.1 Mobile FPS       GAME
9622       10804                                   Poker Pro.Fr       GAME
9623       10805                                  Scoreboard FR  LIFESTYLE
9627       10809            Castle Clash: RPG War and Strategy FR     FAMILY
9630       10812                                  Fr Agnel Pune     FAMILY
9632       10814                            FR: My Secret Pets!     FAMILY
9635       10817                           HTC Sense Input - FR      TOOLS
9638       10820                                Fr. Daoud Lamei     FAMILY
9645       10827                             Fr Agnel Ambarnath     FAMILY
9652       10834                                  FR Calculator     FAMILY
9654       10836                               Sya9a Maroc - FR     FAMILY
9655       10837                Fr. Mike Schmitz Audio Teachings     FAMILY
9658       10840  iHoroscope - 2018 Daily Horoscope & Astrology  LIFESTYLE

      Rating  Reviews   Size     Installs  Type  Price Content Rating  \
839      3.6    52306   42.0    5000000.0  Free   0.00       Everyone
841      3.8    42644   19.0    5000000.0  Free   0.00       Everyone
842      4.2   278082   70.0   10000000.0  Free   0.00       Everyone
843      4.2     6076   32.0     500000.0  Free   0.00       Everyone
844      4.2   112656   93.0    5000000.0  Free   0.00       Everyone
848      4.2   957973   40.0   10000000.0  Free   0.00       Everyone
849      4.7   130582   24.0    5000000.0  Free   0.00       Everyone
851      4.7   167168   20.0    1000000.0  Free   0.00       Everyone
852      4.0    34428   15.0   10000000.0  Free   0.00       Everyone
853      3.6    15247   28.0    1000000.0  Free   0.00       Everyone
854      4.0    48445   10.0    5000000.0  Free   0.00       Everyone
858      4.2    21996   14.0    5000000.0  Free   0.00       Everyone
860      4.0    12073   42.0    1000000.0  Free   0.00       Everyone
861      4.1   111632   65.0    1000000.0  Free   0.00       Everyone
862      4.4   250706   37.0   10000000.0  Free   0.00       Everyone
864      4.5    64959   39.0    1000000.0  Free   0.00       Everyone
865      4.3   659741   47.0   50000000.0  Free   0.00       Everyone
866      4.6   510392   79.0   10000000.0  Free   0.00       Everyone
867      4.3     7215   18.0    1000000.0  Free   0.00       Everyone
869      4.5    60449  100.0    1000000.0  Free   0.00       Everyone
870      4.5   381788   32.0   10000000.0  Free   0.00       Everyone
871      3.6    10697   46.0    1000000.0  Free   0.00       Everyone
873      4.1    31804    9.8    5000000.0  Free   0.00       Everyone
874      3.5     3856    8.2    1000000.0  Free   0.00       Everyone
875      4.4   199684   18.0    5000000.0  Free   0.00       Everyone
877      4.5  1336246   39.0   10000000.0  Free   0.00       Everyone
878      4.2      257   44.0      10000.0  Free   0.00       Everyone
880      4.2      283    5.0     100000.0  Free   0.00       Everyone
881      4.8    12304   21.0     100000.0  Free   0.00       Everyone
882      4.5     8188   22.0     100000.0  Free   0.00       Everyone
...      ...      ...    ...          ...   ...    ...            ...
9599     4.3   834117   44.0   10000000.0  Free   0.00           Teen
9600     3.8      252   11.0      10000.0  Paid  16.99           Teen
9601     4.2      340   72.0      50000.0  Free   0.00     Mature 17+
9602     4.3   245455   84.0   10000000.0  Free   0.00   Everyone 10+
9603     4.2     1405    9.5      10000.0  Paid   1.20       Everyone
9604     4.2      455    2.8      50000.0  Free   0.00       Everyone
9605     4.1      297   48.0      50000.0  Free   0.00           Teen
9607     4.0      368   48.0      50000.0  Free   0.00   Everyone 10+
9609     4.0     1287   38.0      50000.0  Free   0.00       Everyone
9610     4.2    21661   16.0    1000000.0  Free   0.00       Everyone
9611     4.4    28510   78.0     500000.0  Free   0.00           Teen
9613     4.2     7339    4.0     100000.0  Free   0.00       Everyone
9614     4.7    61445    7.8    1000000.0  Free   0.00       Everyone
9615     4.6    32433   46.0    1000000.0  Free   0.00       Everyone
9618     4.1      174   12.0       5000.0  Free   0.00       Everyone
9619     4.9       52   19.0       1000.0  Free   0.00       Everyone
9620     4.0      185   28.0      10000.0  Free   0.00           Teen
9621     4.3    56496   81.0    1000000.0  Free   0.00           Teen
9622     4.2     5442   17.0     100000.0  Free   0.00           Teen
9623     4.3        3   15.0        100.0  Free   0.00       Everyone
9627     4.7   376223   24.0    1000000.0  Free   0.00       Everyone
9630     4.1       80   13.0       1000.0  Free   0.00       Everyone
9632     4.0      785   31.0      50000.0  Free   0.00           Teen
9635     4.0      885    8.0     100000.0  Free   0.00       Everyone
9638     5.0       22    8.6       1000.0  Free   0.00           Teen
9645     4.2      117   13.0       5000.0  Free   0.00       Everyone
9652     4.0        7    2.6        500.0  Free   0.00       Everyone
9654     4.5       38   53.0       5000.0  Free   0.00       Everyone
9655     5.0        4    3.6        100.0  Free   0.00       Everyone
9658     4.5   398307   19.0   10000000.0  Free   0.00       Everyone

             Genres        Last Updated         Current Ver         Android Ver
839         Finance       July 27, 2018              20.1.0          5.0 and up
841         Finance       June 28, 2018                8.18          4.2 and up
842         Finance       July 30, 2018               10.78        4.0.3 and up
843         Finance       July 28, 2018       4.1.1-0-prod          4.2 and up
844         Finance       July 19, 2018               3.9.0          4.1 and up
848         Finance       July 30, 2018               6.5.7          4.2 and up
849         Finance      August 2, 2018  Varies with device  Varies with device
851         Finance        July 4, 2018            3.115.27          4.1 and up
852         Finance      August 2, 2018              2.31.1          5.0 and up
853         Finance       July 13, 2018               3.0.5        4.0.3 and up
854         Finance       July 16, 2018               2.6.0        4.0.3 and up
858         Finance       July 19, 2018              1.17.2        4.0.3 and up
860         Finance        May 25, 2018                 3.3          4.2 and up
861         Finance        June 9, 2018       5.10.1.40699          4.4 and up
862         Finance       July 31, 2018           6.8.0.109          5.0 and up
864         Finance       June 18, 2018               5.0.1          4.1 and up
865         Finance       July 18, 2018              6.28.0          4.4 and up
866         Finance      August 1, 2018              5.38.1          5.0 and up
867         Finance       July 23, 2018               2.7.4          4.0 and up
869         Finance       July 23, 2018              2.9.12          4.0 and up
870         Finance      August 2, 2018              3.22.0          4.1 and up
871         Finance       July 18, 2018              3.22.1          4.1 and up
873         Finance       July 27, 2018            3.1.6.12          4.0 and up
874         Finance       July 31, 2018               3.2.1          4.4 and up
875         Finance       July 19, 2018  Varies with device  Varies with device
877         Finance      August 3, 2018            6.39.1.4          4.1 and up
878         Finance       July 31, 2018              2.2.12          4.4W and up
880         Finance      April 26, 2018            2.5.0.26          5.0 and up
881         Finance      August 2, 2018              6.0.16          5.0 and up
882         Finance       July 31, 2018             2.29.01          4.1 and up
...             ...                 ...                 ...                 ...
9599         Action       July 30, 2018              1.25.4          4.1 and up
9600         Action   February 27, 2015                2.22          5.0 and up
9601         Action    October 27, 2017                   1          4.1 and up
9602         Action        May 31, 2018               2.8.6          4.0 and up
9603         Puzzle        June 5, 2018                 2.7          2.3 and up
9604          Tools   February 28, 2017                 1.6          6.0 and up
9605         Action      March 28, 2018                 1.2          4.1 and up
9607         Action      March 28, 2018                 1.7          4.1 and up
9609           Word   December 18, 2013                   1          2.2 and up
9610         Trivia        May 24, 2018              1.3.81          4.0 and up
9611           Card      August 1, 2018               2.6.6        4.0.3 and up
9613          Tools        May 10, 2018                1.74          4.1 and up
9614          Tools       July 26, 2018             1.9.734          4.1 and up
9615      Lifestyle       June 26, 2018               2.9.1          5.0 and up
9618          Tools       July 30, 2018                6.04          4.4 and up
9619      Education    January 19, 2018                   1          4.4 and up
9620  Entertainment      August 6, 2015               1.3.0          3.0 and up
9621         Action      August 7, 2018             1.5.447          4.0 and up
9622           Card        May 22, 2018               4.1.3          2.3 and up
9623      Lifestyle      August 7, 2018                 2.1          4.2 and up
9627       Strategy       July 18, 2018               1.4.2          4.1 and up
9630      Education       June 13, 2018              2.0.20        4.0.3 and up
9632  Entertainment        June 3, 2015               1.3.1          3.0 and up
9635          Tools    October 30, 2015          1.0.612928          5.0 and up
9638      Education       June 27, 2018               3.8.0          4.1 and up
9645      Education       June 13, 2018              2.0.20        4.0.3 and up
9652      Education       June 18, 2017               1.0.0          4.1 and up
9654      Education       July 25, 2017                1.48          4.1 and up
9655      Education        July 6, 2018                   1          4.1 and up
9658      Lifestyle       July 25, 2018  Varies with device  Varies with device

[4039 rows x 14 columns]
```

In [ ]:
```
%matplotlib inline
import seaborn as sns
sns.set_style("darkgrid")
import warnings
warnings.filterwarnings("ignore")

# Select rows where both 'Rating' and 'Size' values are present (ie. the two values are not null)
apps_with_size_and_rating_present = apps[(apps["Rating"].isna() == False)&(apps["Size"].isna() == False)]

# Subset for categories with at least 250 apps
large_categories = apps_with_size_and_rating_present.groupby("Category").filter(lambda x: len(x) >= 250)

# Plot size vs. rating
plt1 = sns.jointplot(x = large_categories["Size"], y = large_categories["Rating"])

# Select apps whose 'Type' is 'Paid'
paid_apps = apps_with_size_and_rating_present[apps_with_size_and_rating_present["Type"] == "Paid"]

# Plot price vs. rating
plt2 = sns.jointplot(x = paid_apps["Price"], y = paid_apps["Rating"])
```





7. Relation between app category and app price

So now comes the hard part. How are companies and developers supposed to make ends meet? What monetization strategies can companies use to maximize profit? The costs of apps are largely based on features, complexity, and platform.

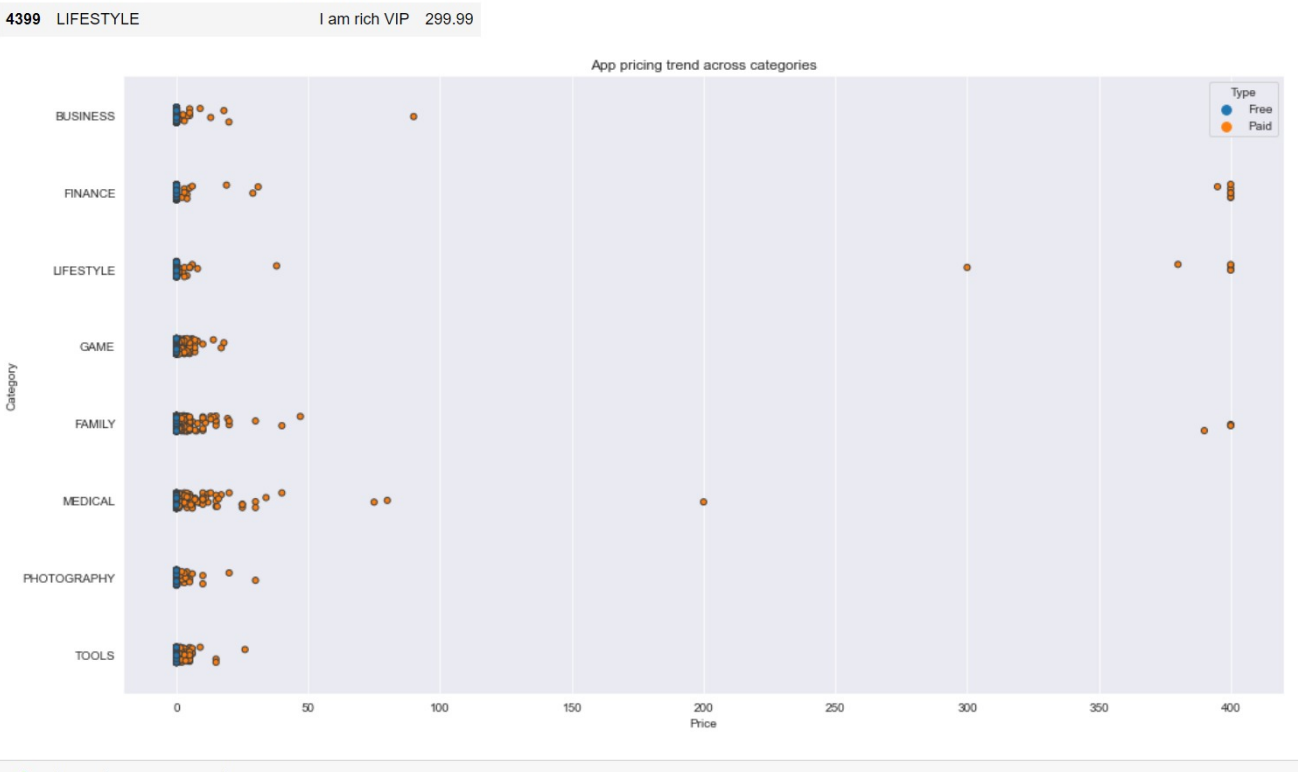
There are many factors to consider when selecting the right pricing strategy for your mobile app. It is important to consider the willingness of your customer to pay for your app. A wrong price could break the deal before the download even happens. Potential customers could be turned off by what they perceive to be a shocking cost, or they might delete an app they've downloaded after receiving too many ads or simply not getting their money's worth.

Different categories demand different price ranges. Some apps that are simple and used daily, like the calculator app, should probably be kept free. However, it would make sense to charge for a highly-specialized medical app that diagnoses diabetic patients. Below, we see that *Medical and Family* apps are the most expensive. Some medical apps extend even up to $80! All game apps are reasonably priced below $20.

In [ ]:
```
import matplotlib.pyplot as plt
fig, ax = plt.subplots()
fig.set_size_inches(15, 8)

# Select a few popular app categories
popular_app_cats = apps[apps.Category.isin(['GAME', 'FAMILY', 'PHOTOGRAPHY',
                                            'MEDICAL', 'TOOLS', 'FINANCE',
                                            'LIFESTYLE','BUSINESS'])]

# Examine the price trend by plotting Price vs Category
ax = sns.stripplot(x = popular_app_cats["Price"], y = popular_app_cats["Category"], hue = popular_app_cats["Type"],  jitter=Tru
e, linewidth=1)
ax.set_title('App pricing trend across categories')

# Apps whose Price is greater than 200
apps_above_200 = apps[apps["Price"] > 200]
apps_above_200[['Category', 'App', 'Price']].sort_values("Price", ascending = False)
```

Out[ ]:

	Category	App	Price
3469	LIFESTYLE	I'm Rich - Trump Edition	400.00
3327	FAMILY	most expensive app (H)	399.99
3465	LIFESTYLE	💎 I'm rich	399.99
4417	FINANCE	I AM RICH PRO PLUS	399.99
4413	FINANCE	I am Rich	399.99
4408	FINANCE	I am rich (Most expensive app)	399.99
4406	FAMILY	I Am Rich Pro	399.99
4403	FINANCE	I am rich(premium)	399.99
4402	FINANCE	I am Rich!	399.99
4400	FINANCE	I Am Rich Premium	399.99
4398	FAMILY	I am Rich Plus	399.99
4396	LIFESTYLE	I am rich	399.99
8780	LIFESTYLE	I'm Rich/Eu sou Rico/أنا غني/我很有錢	399.99
8763	FINANCE	Eu Sou Rico	394.99
4410	FAMILY	I Am Rich	389.99
4401	LIFESTYLE	I am extremely Rich	379.99
4399	LIFESTYLE	I am rich VIP	299.99



In [ ]:
```
print(popular_app_cats)
```

```
      Unnamed: 0                                              App  \
187          187                      Visual Voicemail by MetroPCS
188          188                                Indeed Job Search
189          189                                      Uber Driver
190          190                             ADP Mobile Solutions
191          191                              Snag - Jobs Hiring Now
192          192                       Docs To Go™ Free Office Suite
193          193                               Google My Business
194          194                   OfficeSuite : Free Office + PDF Editor
195          195                                      USPS MOBILE®
196          196                          Job Search by ZipRecruiter
197          197   Curriculum vitae App CV Builder Free Resume Maker
198          198                                    Google Primer
199          199              Alba Heaven - Alvarez Job Portal Services
200          200                                     SuperLivePro
201          201                           Facebook Pages Manager
202          202                           OfficeSuite Pro + PDF (Trial)
203          203                        My Space - Employment Center
204          204                                              Box
205          205                               Polaris Office for LG
206          206                                     Call Blocker
207          207                            Jobs in Alabama - Jobs in Alba
208          208                               Square Point of Sale - POS
209          209                                    Plugin:AOT v5.0
210          210                                      Kariyer.net
211          211                                  SEEK Job Search
212          212                     Become a Job - Find a job or advertise
213          213                                ZOOM Cloud Meetings
214          214                          Easy Installer - Apps On SD
215          215                            Facebook Ads Manager
216          216     IndiaMART: Search Products, Buy, Sell & Trade
...          ...                                              ...
9614       10796    Inf VPN - Global Proxy & Unlimited Free WIFI VPN
9615       10797                            Fuel Rewards® program
9616       10798                               Word Search Tab 1 FR
9618       10800                                        FR Roster
9619       10801                              Fr Ignacio Outreach
9620       10802                              FR: My Famous Lover
9621       10803                     Fatal Raid - No.1 Mobile FPS
9622       10804                                     Poker Pro.Fr
9623       10805                                    Scoreboard FR
9624       10806                                   SnakeBite911 FR
9625       10807                                        My FR App
9626       10808                               lesparticuliers.fr
9627       10809              Castle Clash: RPG War and Strategy FR
9628       10810                                Fr Lupupa Sermons
9630       10812                                    Fr Agnel Pune
9631       10813                                   DICT.fr Mobile
9632       10814                              FR: My Secret Pets!
9634       10816                                FieldBi FR Offline
9635       10817                             HTC Sense Input - FR
9636       10818                               Gold Quote - Gold.fr
9638       10820                                  Fr. Daoud Lamei
9639       10821                                          Poop FR
9642       10824                                        Cardio-FR
9645       10827                               Fr Agnel Ambarnath
9652       10834                                    FR Calculator
9653       10835                                         FR Forms
9654       10836                                 Sya9a Maroc - FR
9655       10837                  Fr. Mike Schmitz Audio Teachings
9656       10838                             Parkinson Exercices FR
9658       10840    iHoroscope - 2018 Daily Horoscope & Astrology

       Category  Rating  Reviews  Size      Installs  Type  Price  \
187    BUSINESS     4.1    16129   NaN   10000000.0  Free   0.00
188    BUSINESS     4.3   674730   NaN   50000000.0  Free   0.00
189    BUSINESS     4.4  1254730   NaN   10000000.0  Free   0.00
190    BUSINESS     4.3    85185  29.0    5000000.0  Free   0.00
191    BUSINESS     4.3    32584   NaN    1000000.0  Free   0.00
192    BUSINESS     4.1   217730   NaN   50000000.0  Free   0.00
193    BUSINESS     4.4    70991   NaN    5000000.0  Free   0.00
194    BUSINESS     4.3  1002861  35.0  100000000.0  Free   0.00
195    BUSINESS     3.9    16589   9.1    1000000.0  Free   0.00
196    BUSINESS     4.8   148945  25.0    1000000.0  Free   0.00
197    BUSINESS     4.5     4458   3.9     500000.0  Free   0.00
198    BUSINESS     4.4    62272  18.0   10000000.0  Free   0.00
199    BUSINESS     4.0     8941  12.0    5000000.0  Free   0.00
200    BUSINESS     4.3    46353  21.0    1000000.0  Free   0.00
201    BUSINESS     4.0  1279184   NaN   50000000.0  Free   0.00
202    BUSINESS     4.2    88073   NaN   10000000.0  Free   0.00
203    BUSINESS     4.5    67000   NaN    1000000.0  Free   0.00
204    BUSINESS     4.2   159872   NaN   10000000.0  Free   0.00
205    BUSINESS     4.2    30847  55.0    5000000.0  Free   0.00
206    BUSINESS     4.6   188841   3.2    5000000.0  Free   0.00
207    BUSINESS     4.1    11622   NaN    5000000.0  Free   0.00
208    BUSINESS     4.6    95912   NaN    5000000.0  Free   0.00
209    BUSINESS     3.1     4034   0.0     100000.0  Free   0.00
210    BUSINESS     3.9    45964  16.0    1000000.0  Free   0.00
211    BUSINESS     4.3    14955   NaN    1000000.0  Free   0.00
212    BUSINESS     4.1     6903  14.0    1000000.0  Free   0.00
213    BUSINESS     4.4    31614  37.0   10000000.0  Free   0.00
214    BUSINESS     4.1    23055   NaN    5000000.0  Free   0.00
215    BUSINESS     4.1    19023   NaN    1000000.0  Free   0.00
216    BUSINESS     4.5   207372  11.0    5000000.0  Free   0.00
...         ...     ...      ...   ...          ...   ...    ...
9614      TOOLS     4.7    61445   7.8    1000000.0  Free   0.00
9615  LIFESTYLE     4.6    32433  46.0    1000000.0  Free   0.00
9616     FAMILY     NaN        0   1.0         50.0  Paid   1.04
9618      TOOLS     4.1      174  12.0       5000.0  Free   0.00
9619     FAMILY     4.9       52  19.0       1000.0  Free   0.00
9620     FAMILY     4.0      185  28.0      10000.0  Free   0.00
9621       GAME     4.3    56496  81.0    1000000.0  Free   0.00
9622       GAME     4.2     5442  17.0     100000.0  Free   0.00
9623  LIFESTYLE     4.3        3  15.0        100.0  Free   0.00
9624    MEDICAL     NaN        1  42.0        500.0  Free   0.00
9625      TOOLS     NaN        2   4.2        100.0  Free   0.00
9626  LIFESTYLE     NaN       96   1.0      50000.0  Free   0.00
9627     FAMILY     4.7   376223  24.0    1000000.0  Free   0.00
9628   BUSINESS     4.8       19  21.0        100.0  Free   0.00
9630     FAMILY     4.1       80  13.0       1000.0  Free   0.00
9631   BUSINESS     NaN       20   2.7      10000.0  Free   0.00
9632     FAMILY     4.0      785  31.0      50000.0  Free   0.00
9634   BUSINESS     NaN        2   6.8        100.0  Free   0.00
9635      TOOLS     4.0      885   8.0     100000.0  Free   0.00
9636    FINANCE     NaN       96   1.5      10000.0  Free   0.00
9638     FAMILY     5.0       22   8.6       1000.0  Free   0.00
9639     FAMILY     NaN        6   2.5         50.0  Free   0.00
9642    MEDICAL     NaN       67  82.0      10000.0  Free   0.00
9645     FAMILY     4.2      117  13.0       5000.0  Free   0.00
9652     FAMILY     4.0        7   2.6        500.0  Free   0.00
9653   BUSINESS     NaN        0   9.6         10.0  Free   0.00
9654     FAMILY     4.5       38  53.0       5000.0  Free   0.00
9655     FAMILY     5.0        4   3.6        100.0  Free   0.00
9656    MEDICAL     NaN        3   9.5       1000.0  Free   0.00
9658  LIFESTYLE     4.5   398307  19.0   10000000.0  Free   0.00

     Content Rating         Genres        Last Updated          Current Ver  \
187        Everyone       Business       July 30, 2018   Varies with device
188        Everyone       Business        May 21, 2018   Varies with device
189        Everyone       Business      August 3, 2018   Varies with device
190        Everyone       Business       July 17, 2018                3.4.2
191        Everyone       Business         May 4, 2018   Varies with device
192        Everyone       Business       April 2, 2018   Varies with device
193        Everyone       Business       July 24, 2018       2.19.0.204537701
194        Everyone       Business      August 2, 2018            9.7.14188
195        Everyone       Business       June 14, 2018               4.9.10
196        Everyone       Business       July 19, 2018                5.2.8
197        Everyone       Business        July 8, 2018                  3.3
198        Everyone       Business       June 26, 2018              3.550.2
199        Everyone       Business       July 31, 2018               4.6.30
200        Everyone       Business      April 13, 2016                  2.8
201        Everyone       Business      August 2, 2018   Varies with device
202        Everyone       Business        July 9, 2018   Varies with device
203        Everyone       Business       July 17, 2018   Varies with device
204        Everyone       Business       July 31, 2018   Varies with device
205        Everyone       Business    February 8, 2018               7.3.21
206        Everyone       Business       June 21, 2018               1.1.13
207        Everyone       Business       July 26, 2018   Varies with device
208        Everyone       Business       July 30, 2018   Varies with device
209        Everyone       Business  September 11, 2015  3.0.1.11 (Build 311)
210        Everyone       Business       July 18, 2018                5.1.5
211        Everyone       Business       July 30, 2018   Varies with device
212        Everyone       Business       July 10, 2018                2.3.6
213        Everyone       Business       July 20, 2018        4.1.28165.0716
214        Everyone       Business        June 8, 2018   Varies with device
215        Everyone       Business      August 1, 2018          99.0.0.35.75
216        Everyone       Business      August 2, 2018               12.2.4
...             ...            ...                 ...                  ...
9614       Everyone          Tools       July 26, 2018              1.9.734
9615       Everyone      Lifestyle       June 26, 2018                2.9.1
9616       Everyone         Puzzle    February 6, 2012                  1.1
9618       Everyone          Tools       July 30, 2018                 6.04
9619       Everyone      Education    January 19, 2018                    1
9620           Teen  Entertainment      August 6, 2015                1.3.0
9621           Teen         Action      August 7, 2018              1.5.447
9622           Teen           Card        May 22, 2018                4.1.3
9623       Everyone      Lifestyle      August 7, 2018                  2.1
9624       Everyone        Medical     October 9, 2017                  1.2
9625       Everyone          Tools       April 9, 2018         1.283.0037
9626       Everyone      Lifestyle   November 25, 2014                  1.5
9627       Everyone       Strategy       July 18, 2018                1.4.2
9628       Everyone       Business       June 12, 2018                    1
9630       Everyone      Education       June 13, 2018               2.0.20
9631       Everyone       Business       July 17, 2018               2.1.10
9632           Teen  Entertainment        June 3, 2015                1.3.1
9634       Everyone       Business      August 6, 2018                2.1.8
9635       Everyone          Tools    October 30, 2015           1.0.612928
9636       Everyone        Finance        May 19, 2016                  2.3
9638           Teen      Education       June 27, 2018                3.8.0
9639       Everyone  Entertainment        May 29, 2018                    1
9642       Everyone        Medical       July 31, 2018                2.2.2
9645       Everyone      Education       June 13, 2018               2.0.20
9652       Everyone      Education       June 18, 2017                1.0.0
9653       Everyone       Business  September 29, 2016                1.1.5
9654       Everyone      Education       July 25, 2017                 1.48
9655       Everyone      Education        July 6, 2018                    1
9656       Everyone        Medical    January 20, 2017                    1
9658       Everyone      Lifestyle       July 25, 2018   Varies with device

             Android Ver
187   Varies with device
188   Varies with device
189   Varies with device
190           5.0 and up
191   Varies with device
192   Varies with device
193           4.4 and up
194           4.1 and up
195         4.0.3 and up
196           5.0 and up
197           4.1 and up
198           4.1 and up
199         4.0.3 and up
200           1.5 and up
201   Varies with device
202   Varies with device
203           4.3 and up
204   Varies with device
205           4.1 and up
206           4.0 and up
207   Varies with device
208   Varies with device
209           2.2 and up
210           4.1 and up
211   Varies with device
212           4.1 and up
213           4.0 and up
214   Varies with device
215           4.1 and up
216           4.0 and up
...                  ...
9614          4.1 and up
9615          5.0 and up
9616          3.0 and up
9618          4.4 and up
9619          4.4 and up
9620          3.0 and up
9621          4.0 and up
9622          2.3 and up
9623          4.2 and up
9624          4.1 and up
9625        2.3.3 and up
9626          2.3 and up
9627          4.1 and up
9628          4.4 and up
9630        4.0.3 and up
9631          4.1 and up
9632          3.0 and up
9634          4.1 and up
9635          5.0 and up
9636          2.2 and up
9638          4.1 and up
9639        4.0.3 and up
9642          4.4 and up
9645        4.0.3 and up
9652          4.1 and up
9653          4.0 and up
9654          4.1 and up
9655          4.1 and up
9656          2.2 and up
9658  Varies with device

[5428 rows x 14 columns]
```

8. Filter out "junk" apps

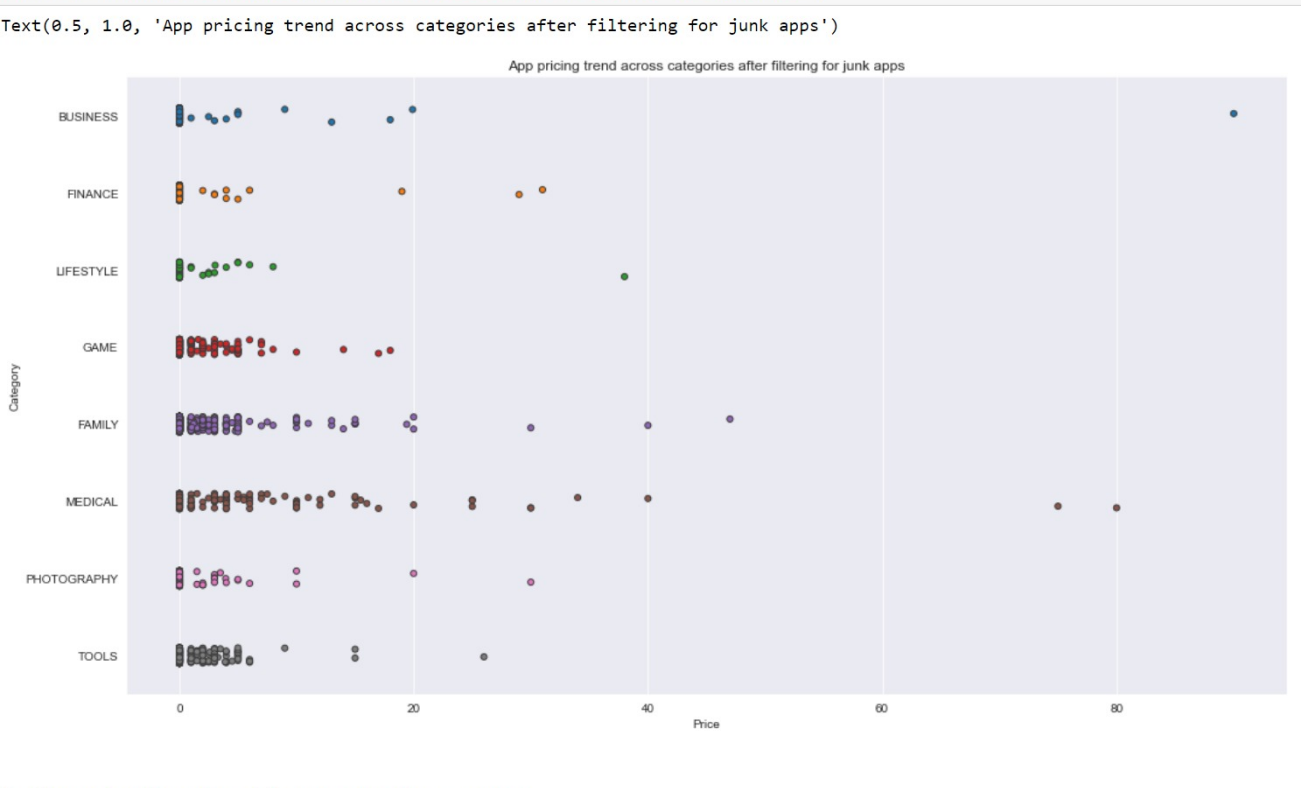
It looks like a bunch of the really expensive apps are "junk" apps. That is, apps that don't really have a purpose. Some app developer may create an app called *I Am Rich Premium* or *most expensive app (H)* just for a joke or to test their app development skills. Some developers even do this with malicious intent and try to make money by hoping people accidentally click purchase on their app in the store.

Let's filter out these junk apps and re-do our visualization.

In [ ]:

In [ ]:
```
# Select apps priced below $100
apps_under_100 = popular_app_cats[popular_app_cats["Price"] < 100]

fig, ax = plt.subplots()
fig.set_size_inches(15, 8)

# Examine price vs category with the authentic apps (apps_under_100)
ax = sns.stripplot(x = apps_under_100["Price"], y = apps_under_100["Category"], data = apps_under_100, jitter = True, linewidth
= 1)
ax.set_title('App pricing trend across categories after filtering for junk apps')
```

Out[ ]:
```
Text(0.5, 1.0, 'App pricing trend across categories after filtering for junk apps')
```



9. Popularity of paid apps vs free apps

For apps in the Play Store today, there are five types of pricing strategies: free, freemium, paid, paymium, and subscription. Let's focus on free and paid apps only. Some characteristics of free apps are:

- Free to download.
- Main source of income often comes from advertisements.
- Often created by companies that have other products and the app serves as an extension of those products.
- Can serve as a tool for customer retention, communication, and customer service.

Some characteristics of paid apps are:

- Users are asked to pay once for the app to download and use it.
- The user can't really get a feel for the app before buying it.

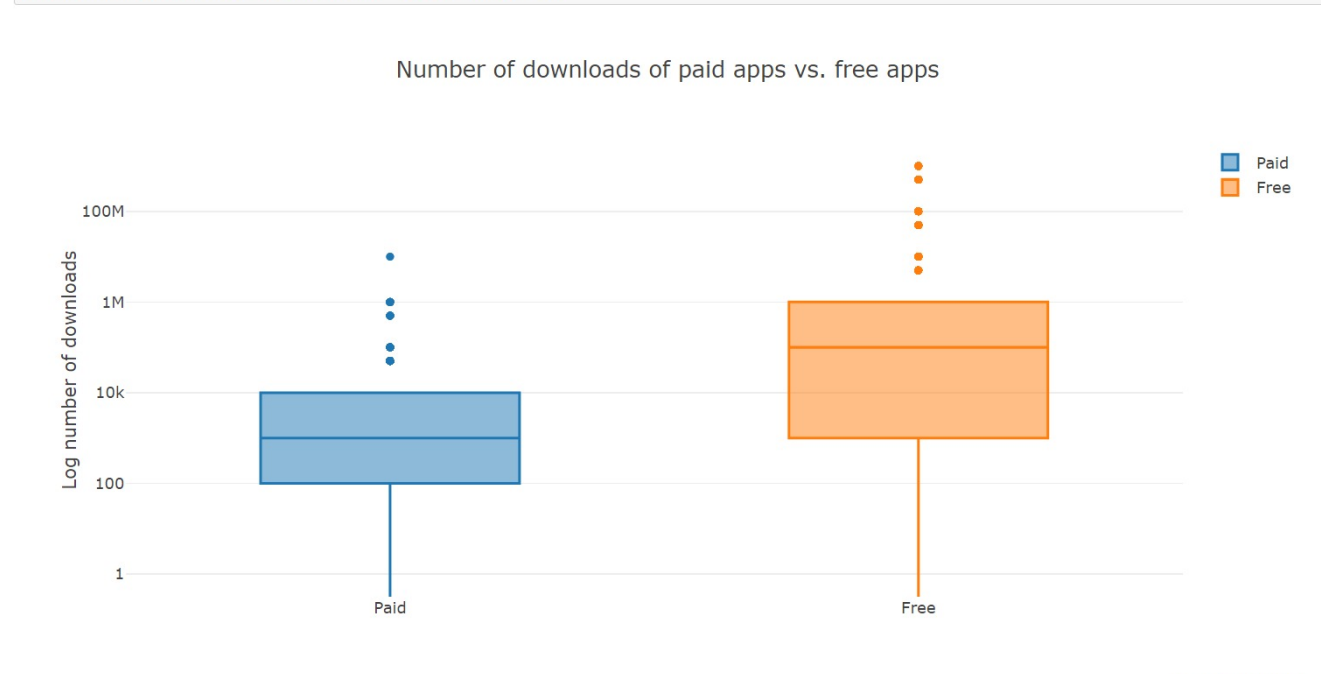
Are paid apps installed as much as free apps? It turns out that paid apps have a relatively lower number of installs than free apps, though the difference is not as stark as I would have expected!

In [ ]:
```
apps[apps['Type'] == 'Paid'][['Category','Installs']].head()
```

Out[ ]:

	Category	Installs
233	BUSINESS	100000.0
234	BUSINESS	100000.0
372	COMMUNICATION	100000.0
406	DATING	10000.0
407	DATING	1000.0

In [ ]:
```
trace0 = go.Box(
    # Data for paid apps
    y = apps[apps['Type'] == 'Paid']['Installs'],
    name = 'Paid'
)

trace1 = go.Box(
    # Data for free apps
    y = apps[apps['Type'] == 'Free']['Installs'],
    name = 'Free'
)

layout = go.Layout(
    title = "Number of downloads of paid apps vs. free apps",
    yaxis = dict(title = "Log number of downloads",
                type = 'log',
                autorange = True)
)

# Add trace0 and trace1 to a list for plotting
data = [trace0, trace1]
plotly.offline.iplot({'data': data, 'layout': layout})
```



Export to plot.ly »

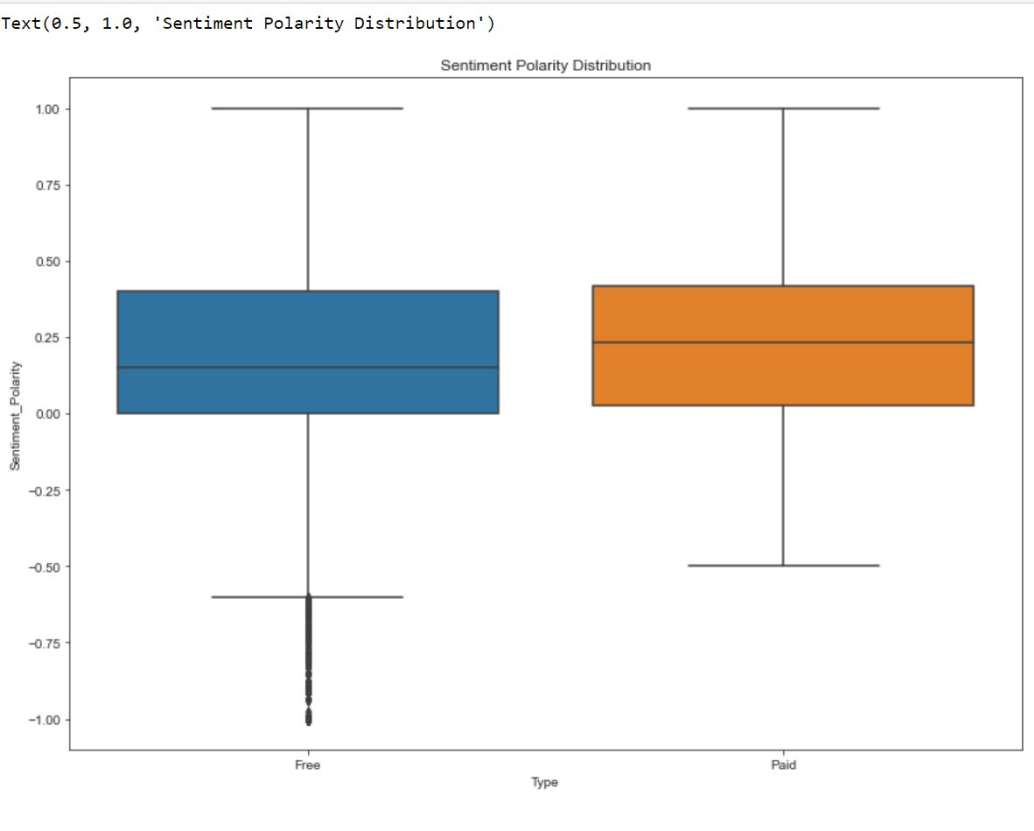
10. Sentiment analysis of user reviews

Mining user review data to determine how people feel about your product, brand, or service can be done using a technique called sentiment analysis. User reviews for apps can be analyzed to identify if the mood is positive, negative or neutral about that app. For example, positive words in an app review might include words such as 'amazing', 'friendly', 'good', 'great', and 'love'. Negative words might be words like 'malware', 'hate', 'problem', 'refund', and 'incompetent'.

By plotting sentiment polarity scores of user reviews for paid and free apps, we observe that free apps receive a lot of harsh comments, as indicated by the outliers on the negative y-axis. Reviews for paid apps appear never to be extremely negative. This may indicate something about app quality, i.e., paid apps being of higher quality than free apps on average. The median polarity score for paid apps is a little higher than free apps, thereby syncing with our previous observation.

In this notebook, we analyzed over ten thousand apps from the Google Play Store. We can use our findings to inform our decisions should we ever wish to create an app ourselves.

In [ ]:
```
# Load user_reviews.csv
reviews_df = pd.read_csv("datasets/user_reviews.csv")

# Join the two dataframes
merged_df = apps.merge(reviews_df, on = "App", how ="inner")

# Drop NA values from Sentiment and Review columns
merged_df = merged_df.dropna(subset = ['Sentiment', 'Review'])

sns.set_style('ticks')
fig, ax = plt.subplots()
fig.set_size_inches(11, 8)

# User review sentiment polarity for paid vs. free apps
ax = sns.boxplot(x = merged_df["Type"], y = merged_df["Sentiment_Polarity"], data = merged_df)
ax.set_title('Sentiment Polarity Distribution')
```

Out[ ]:
```
Text(0.5, 1.0, 'Sentiment Polarity Distribution')
```