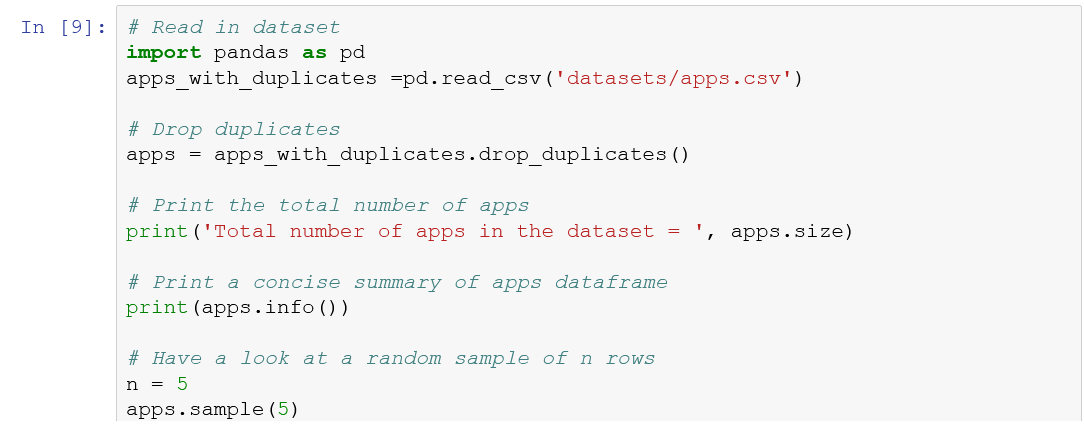
**1. Google Play Store apps and reviews**

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](https://www.androidpolice.com/2019/01/21/google-play-stores-redesigned-ratings-and-reviews-section-lets-you-easily-filter-by-star-rating/). 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.



## 2. Data cleaning

The four features that we will be working with most frequently henceforth are Installs, Size, Rating and Price. The info() function (from the previous task) told us that Installs and Price columns are of type object and not int64 or float64 as we would expect. This is because the column contains some characters more than just [0,9] digits. Ideally, we would want these columns to be numeric as their name suggests.   
Hence, we now proceed to data cleaning and prepare our data to be consumed in our analyis later. Specifically, the presence of special characters (, $ +) in the Installs and Price columns make their conversion to a numerical data type difficult.



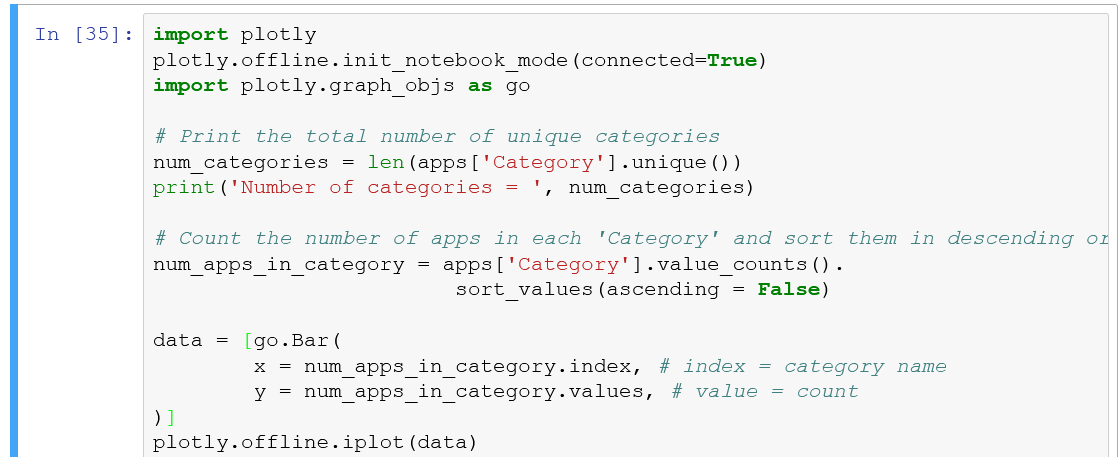
**3. Exploring app categories**

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?

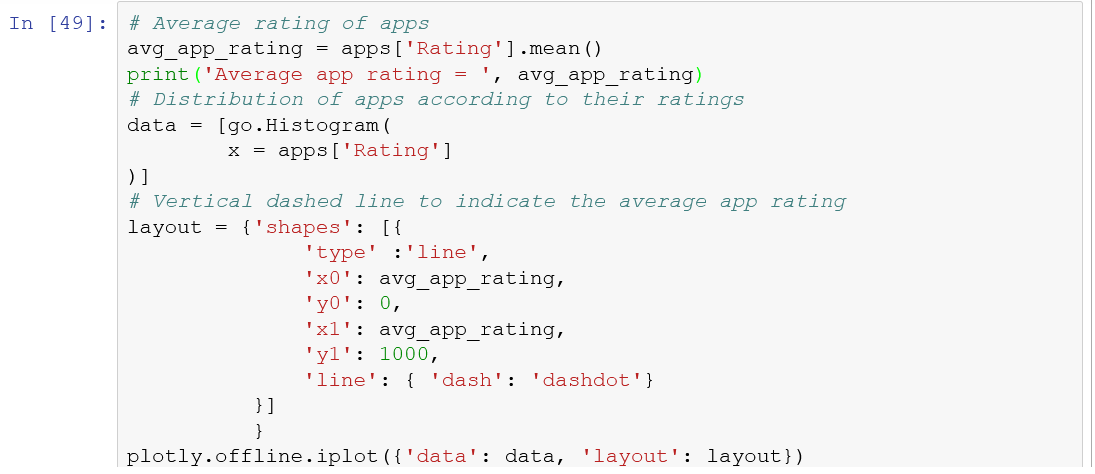
We will see that there are 33 unique app categories present in our dataset. *Family* and *Game* apps have the highest market prevalence. Interestingly, *Tools*, *Business* and *Medical* apps are also at the top.



## 4. 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.



**5. 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.

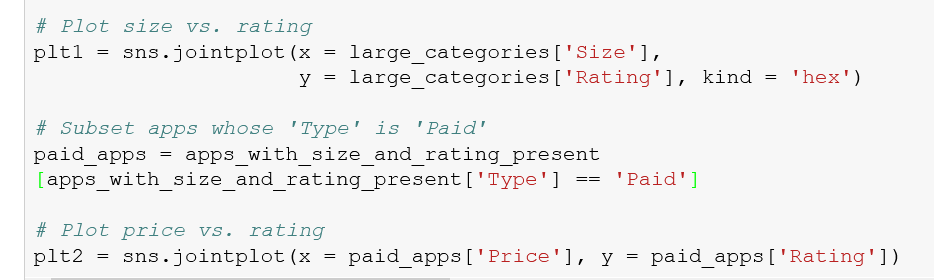
How can we effectively come up with strategies to size and price our app?

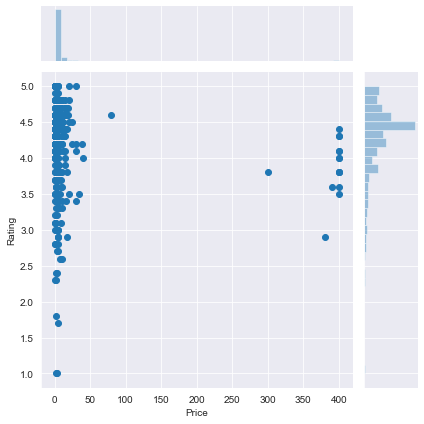
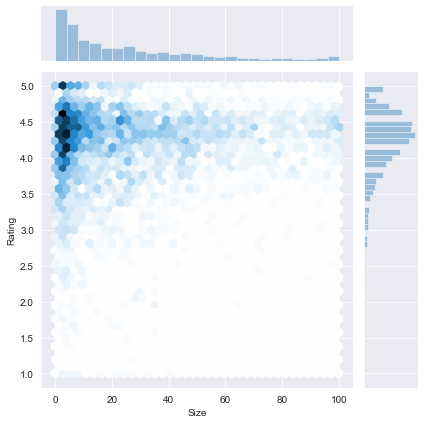
* Does the size of an app affect its rating?
* Do users really care about system-heavy apps or do they prefer light-weighted apps?
* Does the price of an app affect its rating?
* Do users always prefer free apps over paid apps?

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.

* Select rows from apps where Rating is not null and Size is not null. Notice the use of ~ for NOT operation and the bitwise operator & to AND the two conditions. Store the result in the apps\_with\_size\_and\_rating\_present dataframe.
* Apply the groupby function on apps\_with\_size\_and\_rating\_present['Category']. Using filter() function, select only those groups or categories having greater than or equal to 250 apps. Assign the result to large\_categories dataframe.
* Fill out x and y to create a joint plot of Rating as a function of Size.
* Subset apps\_with\_size\_and\_rating\_present dataframe to select apps of type Paid. Assign the result to paid\_apps dataframe.
* Fill out x and y to create a joint plot of Rating as a function of Price.







## 6. 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.

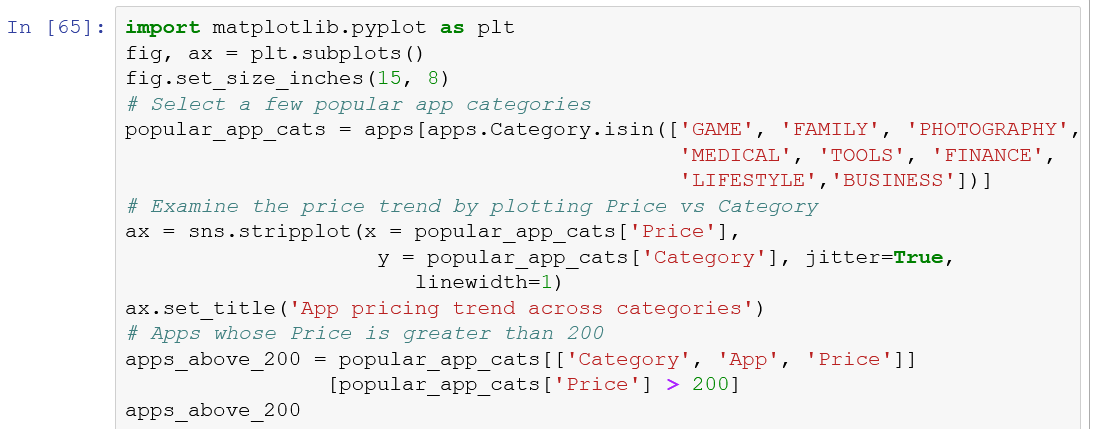
Different categories demand different price ranges. Some apps that are simple and used daily, like the calculator app, should probably be kept free. However, it would make sense to charge for a highly-specialized medical app that diagnoses diabetic patients. Below, we see that Medical and Family apps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

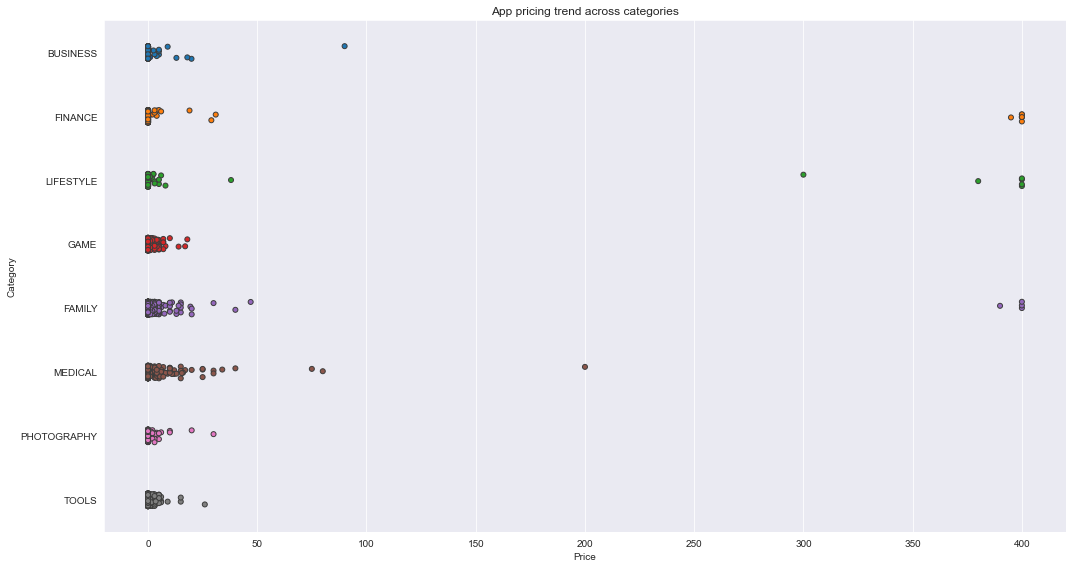
Use a strip plot to visualize the distribution of paid apps across different categories.

* Plot a strip plot with x-axis extending along the Price range and y-axis depicting the Category.
* Print Category, App and Price for apps that are priced above 200.

Here are some interesting websites that can estimate app price:

* [Estimate my app](https://estimatemyapp.com/)
* [How much to make an app](http://howmuchtomakeanapp.com/)

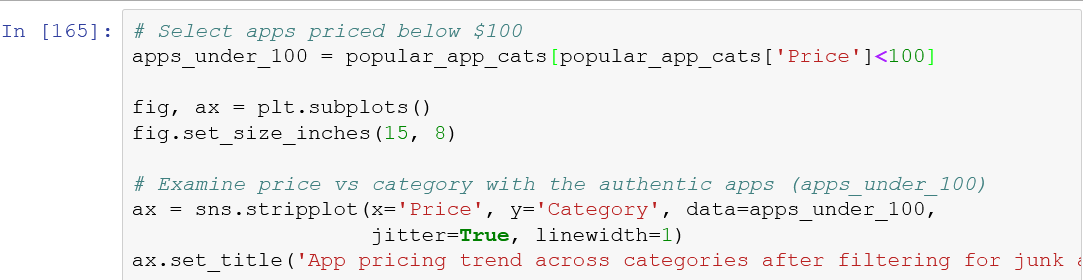


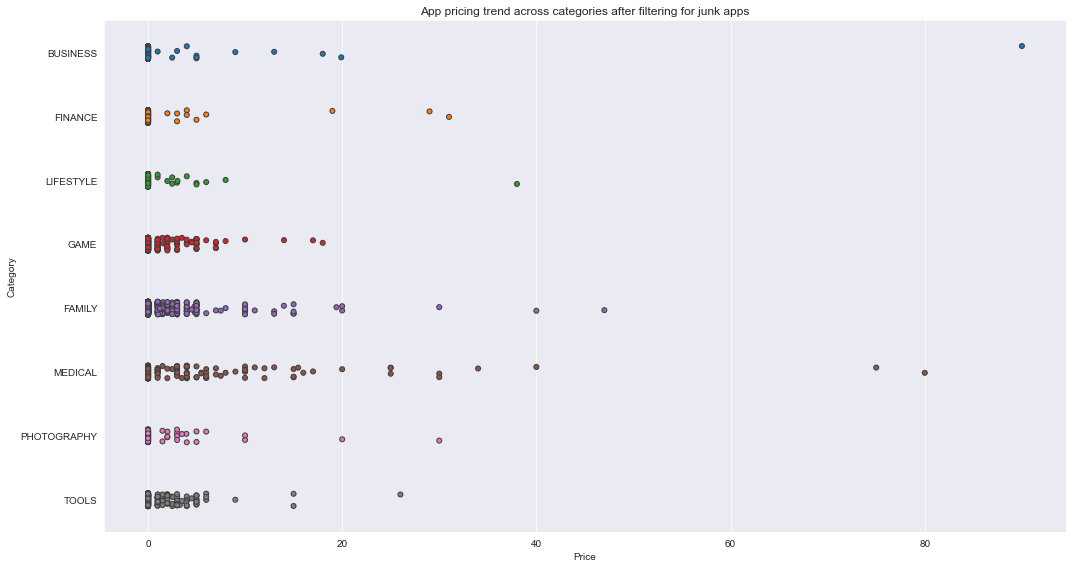


## 7. Filter out "junk" apps

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.





**8. 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:

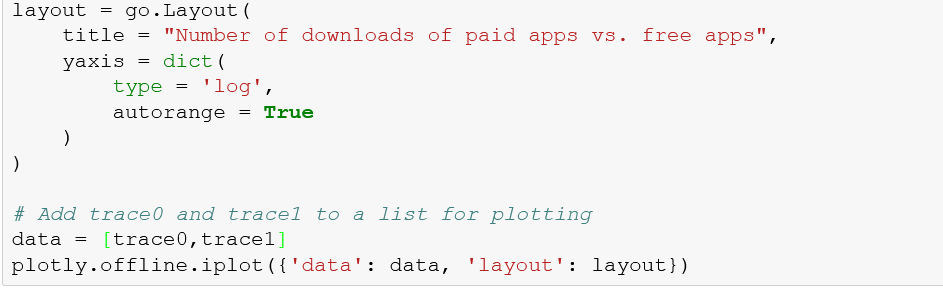
* Users are asked to pay once for the app to download and use it.
* The user can't really get a feel for the app before buying it.

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!

Prep the data for a box plot that compares the number of installs of paid apps vs. number of installs of free apps.

* From apps, filter rows where for Type == Paid, and select the Installs column and assign it to y of trace0.
* From apps, filter rows where for Type == Free, and select the Installs column and assign it to y of trace1.
* Create a Python list containing variables trace0 and trace1.





## 9. 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.

Load the user review data and plot it to visualize sentiment of paid vs. free apps.

* Read datasets/user\_reviews.csv into the reviews\_df DataFrame.
* Merge apps and reviews\_df DataFrames on the common column App and assign the result to merged\_df.
* Create a box plot with Type on the x-axis and Sentiment\_Polarity on the y-axis.

