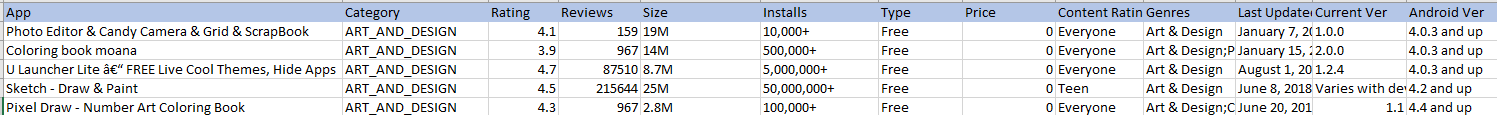
**Google Play Store Summary**

Data was uploaded by user Lava18 via Kaggle (Link to the data found below):

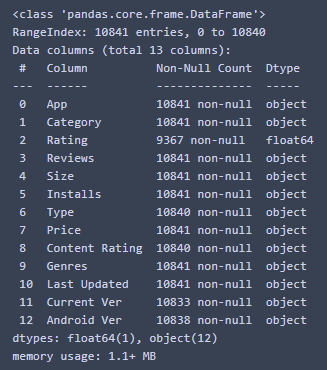
<https://www.kaggle.com/lava18/google-play-store-apps>

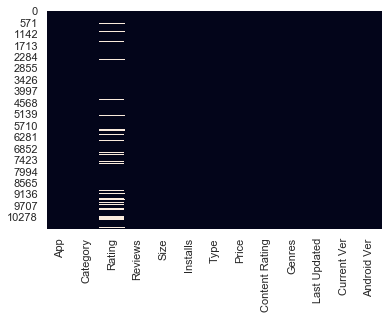
Column explanation:

* App – Name of the app within the Google Play store.
* Category – The genre of the app.
* Rating – The average rating of the app given by its users.
* Reviews – Number of reviews of the app by users.
* Size – The amount of storage the app uses.
* Installs – The number of people who have installed the app, values given as an exponential increment to the nearest whole number.
* Type – Whether the app costs or not.
* Price – The price of the app.
* Content Rating – The target audience of the app.
* Genres – The same as Category
* Last Update – The last time the app was updated.
* Current Ver – Current Version of the app.
* Android Ver – Current Android version.

Above the column explanation, we can see the first five rows of the dataset. To begin this summary, I will be walking through the steps I took to cleaning this dataset.

**Data Cleaning:**

There are almost eleven thousand apps in this list, therefore to find the missing values effectively, I used these two methods:

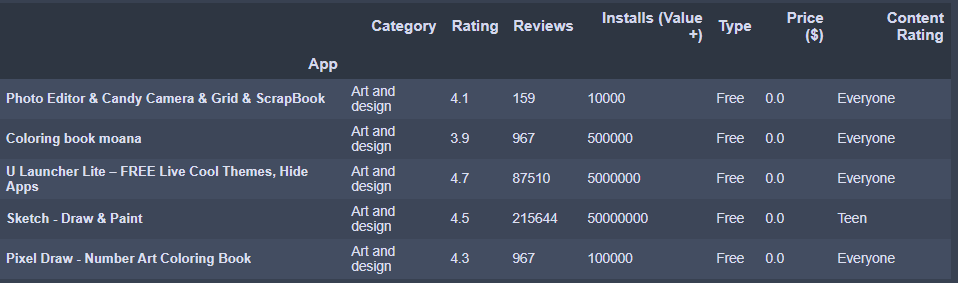


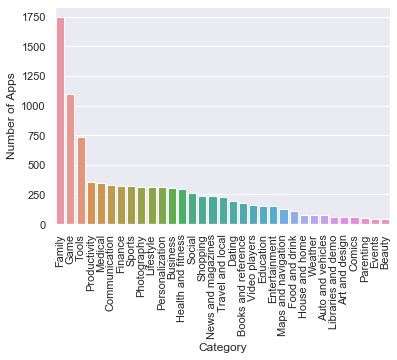
The image on the left is a heatmap showing us the missing values, spaces in white are null, which tells us that a vast majority of the missing data is in the rating column. I decided to drop the rows with missing values rather than the entire column as the rating column is an important part of this dataset.

The other columns don’t seem to have any missing values shown in the heatmap, as we have thousands of values, it's unlikely we would see a few missing values from those columns. The table on the right tells us the exact number of non-null values in the dataset, here I decided to also drop those rows and it shouldn’t affect results later.

Here are the fixes I have made to the messy values in this dataset:

* Category – Removed hyphens, turned characters into lowercase and capitalised the starting letter.
* Rating – Removed null values.
* Reviews – changed data type from objects to integers.
* Installs – Removed “+” and “,” from the values, converted the cells from objects into integers and renamed the column to indicate that the values aren’t exact.
* Price – Removed the “$” sign from the values and converted the values from strings to floats.
* Content Rating – remained the same
* Dropped the following columns as they’re irrelevant to the rest of the data (Size, Genres, Last Updated, Current Ver, Android Ver).

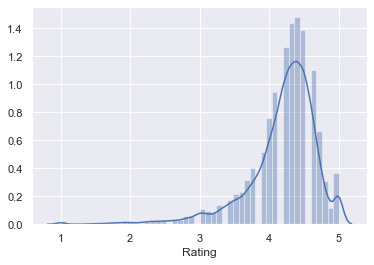


Above is the resultant table after the data cleaning, the data is now more usable and looks much more orderly compared to the CSV file seen on the first page.

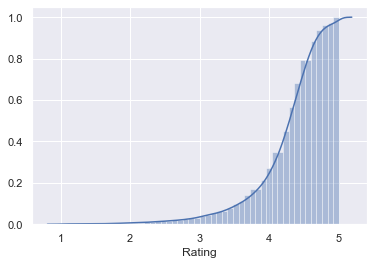
**Data Visualisations:**

What category is the majority among the apps on the Google Play Store?

Family apps take an overwhelming majority. This may be a surprise for some as games have a whole section of its own on the Play Store.



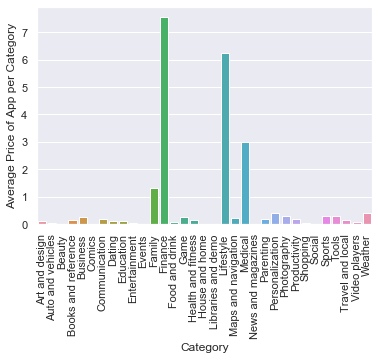
Here on the left, is a probability density function that tells us the probability density of a given rating of an app on the Play Store. This means that most of the apps in the Play Store average around a 4.4 rating.



Here is a different representation of the probability density function called the cumulative density function. 80% of the data is between 0 – 4.5. This tells us that it’s a 1 in 5 chance that an app you randomly find in this dataset will have a rating between 4.5 and 5.

What percentage of apps in the Google Play Store cost money?

Only 6.9% of apps have a price tag.

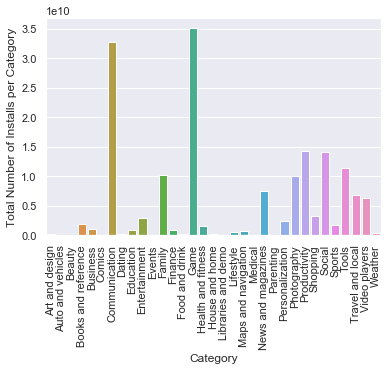
What Category has the most expensive apps?

On average, categories finance, lifestyle and medical apps cost the most.

You can imagine these apps are averagely more expensive because they’re being outweighed by a few extremely expensive ones.

When we look for prices that are above $100, the following dataset is returned:



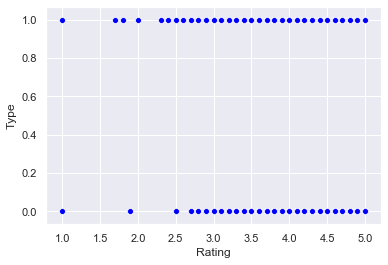
They are all mostly finance and lifestyle and all named the same thing. It looks to me like some apps that you have just to claim that you have the most expensive app. This explains why finance and lifestyle apps averagely cost the most amount of money.

Which category has the greatest number of users installing their apps?

As you would expect, we have games in the lead closely followed by communications. Communication apps are used for calling and messaging, social media comes under the social category.

**Logistical Regression:**

For this logistical regression, I went in to try and find a relationship between ratings and type, could I predict what rating an app is more likely to be based on whether it was free or paid. As the number of free apps largely outweighs that of the paid apps, I took a slice of the free apps to match the number of paid apps and merged the two tables.

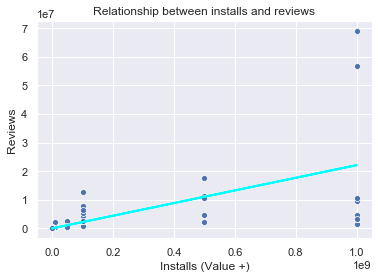
When we plot a scatter of this, the following returns:

It doesn’t seem to have much correlation given this graph. You would expect apps that cost is more likely to have higher ratings.

I put the data through a train and test split and the model does predict higher ratings to be paid apps. The model output a 60% reliance score, which is slightly below acceptable. This concludes that paying money for an app may be better quality, but it’s not a certainty.

**Linear Regression:**

For this linear regression, I attempted to predict the number of reviews you’d expect to get given on four new apps as per installations by users. This relationship outputs an R-Squared value of 0.450, this means it doesn’t explain the data too well, but we have a high F-statistic of 1056 which means the relationship is relevant. Let’s see how the data plots below:

The regression line on the right shows the relationship between the two.

Below is a table of four new apps and the number of users who have installed them, that has been joined with the predicted number of reviews each of these apps are likely to receive.