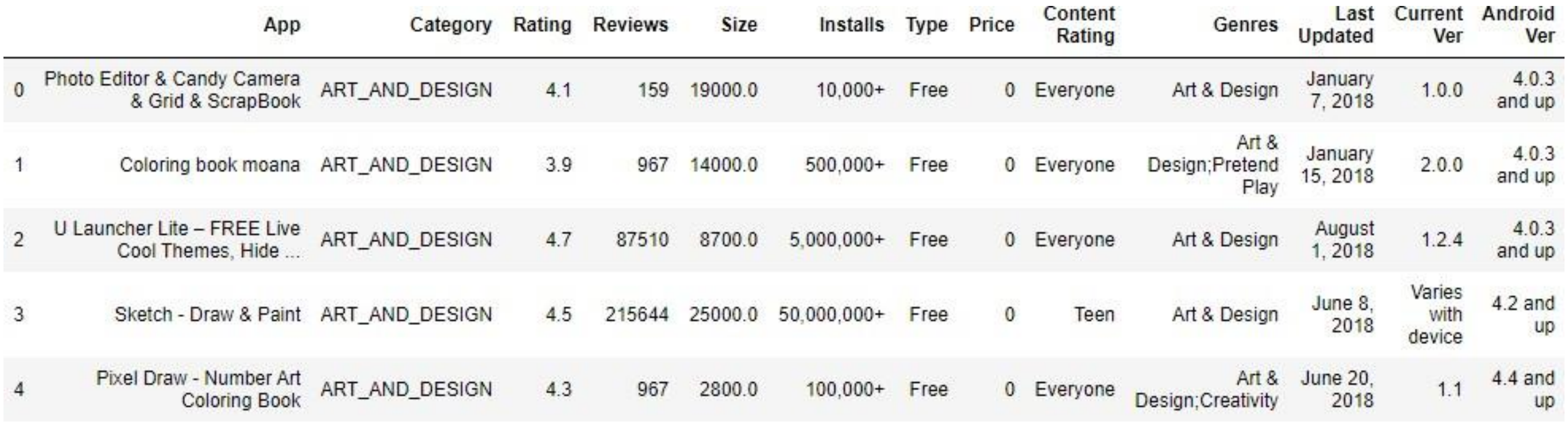
Objective

Google Play Store team is about to launch a new feature where in certain apps that are promising are boosted in visibility. The boost will manifest in multiple ways – higher priority in recommendations sections (“Similar apps”, “You might also like”, “New and updated games”). These will also get a boost in visibility in search results. This feature will help bring more attention to newer apps that have potential.

The task is to understand what makes an app perform well - size? price? category? multiple factors together? Analyze the data and present your insights in a format consumable by business – the final output of the analysis would be presented to business as insights with supporting data/visualizations.

Data



	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

Tasks

1. Data clean up – Missing value treatment
    a. Drop records where rating is missing since rating is our target/study variable
    b. Check the null values for the Android Ver column.
        i. Are all 3 records having the same problem?
        ii. Drop the 3rd record i.e. record for “Life Made WIFI …”
        iii. Replace remaining missing values with the mode
    c. Current ver – replace with most common value
2. Data clean up – correcting the data types
    a. Which all variables need to be brought to numeric types?
    b. Price variable – remove $ sign and convert to float

c. Installs – remove ‘,’ and ‘+’ sign, convert to integer

d. Convert all other identified columns to numeric

3. Sanity checks – check for the following and handle accordingly

a. Avg. rating should be between 1 and 5, as only these values are allowed on the play store.

i. Are there any such records? Drop if so.

b. Reviews should not be more than installs as only those who installed can review the app.

i. Are there any such records? Drop if so.

4. Identify and handle outliers –

a. Price column

i. Make suitable plot to identify outliers in price

ii. Do you expect apps on the play store to cost $200? Check out these cases

iii. After dropping the useless records, make the suitable plot again to identify outliers

iv. Limit data to records with price < $30

b. Reviews column

i. Make suitable plot

ii. Limit data to apps with < 1 Million reviews

c. Installs

i. What is the 95th percentile of the installs?

ii. Drop records having a value more than the 95th percentile

Data analysis to answer business questions

5. What is the distribution of ratings like? (use Seaborn) More skewed towards higher/lower values?

a. How do you explain this?

b. What is the implication of this on your analysis?

6. What are the top Content Rating values?

a. Are there any values with very few records?

b. If yes, drop those as they won’t help in the analysis

7. Effect of size on rating

a. Make a joinplot to understand the effect of size on rating

b. Do you see any patterns?

c. How do you explain the pattern?

8. Effect of price on rating
   a. Make a jointplot (with regression line)
   b. What pattern do you see?
   c. How do you explain the pattern?
   d. Replot the data, this time with only records with price > 0
   e. Does the pattern change?
   f. What is your overall inference on the effect of price on the rating
9. Look at all the numeric interactions together –
   a. Make a pairplort with the colulmns - 'Reviews', 'Size', 'Rating', 'Price'
10. Rating vs. content rating
    a. Make a bar plot displaying the rating for each content rating
    b. Which metric would you use? Mean? Median? Some other quantile?
    c. Choose the right metric and plot
11. Content rating vs. size vs. rating – 3 variables at a time
    a. Create 5 buckets (20% records in each) based on Size
    b. By Content Rating vs. Size buckets, get the rating (20th percentile) for each combination
    c. Make a heatmap of this
       i. Annotated
       ii. Greens color map
    d. What’s your inference? Are lighter apps preferred in all categories? Heavier? Some?