CHENNAI MATHEMATICAL INSTITUTE

Big Data Project on

Visualizing Play Store App Data

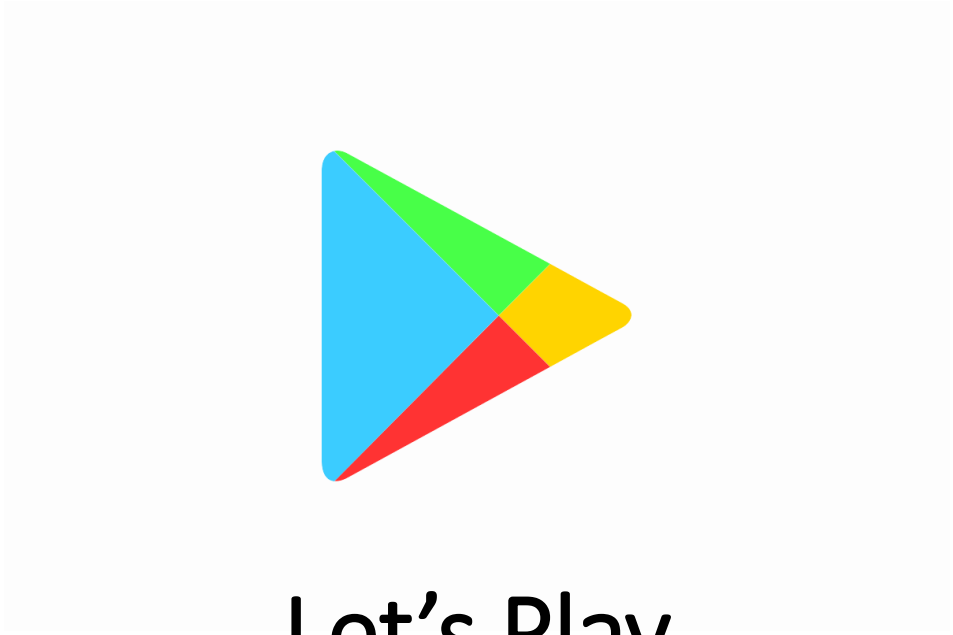
by

Avirup Das
Prosenjit Dey
Shiuli Subhra Ghosh
Suman Roy

“Let everything happen to you.
Beauty and terror.
Just keep going. No Feeling is Final.”

Rainer Maria Rilke

- From a Visualization's point of view



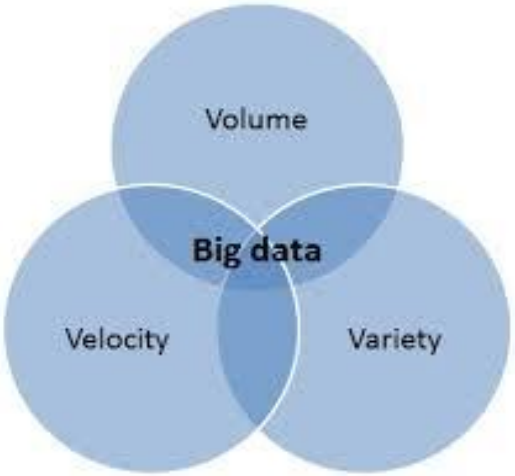
Let’s Play

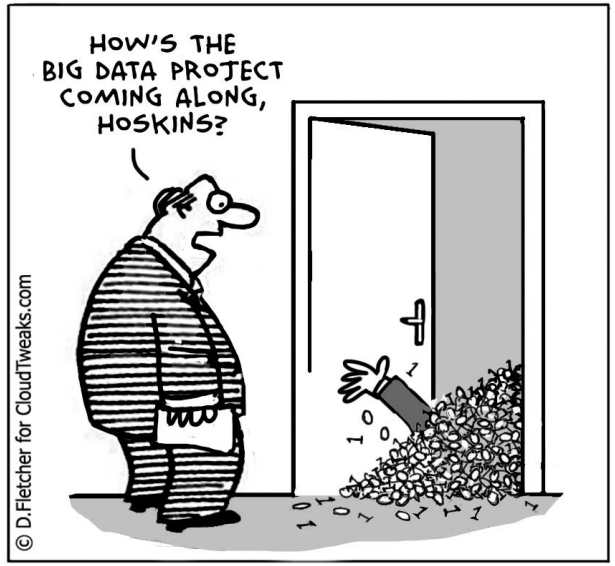
Why do we need to visualise Google Play store App data?

- Earlier till 2020, the total number of apps on the Google Play Store was 2.87 million which has now **risen to nearly 5 million**.
- In 2020, there were **108.5 billion apps downloaded** by the users on Google Play Store.
- If we analyse and visualise this huge data, we can gain insights on what parameters an app outperforms others.
- Developers might find these insights very useful to before creating an app or making their apps better.

Why is this a Big Data Project?

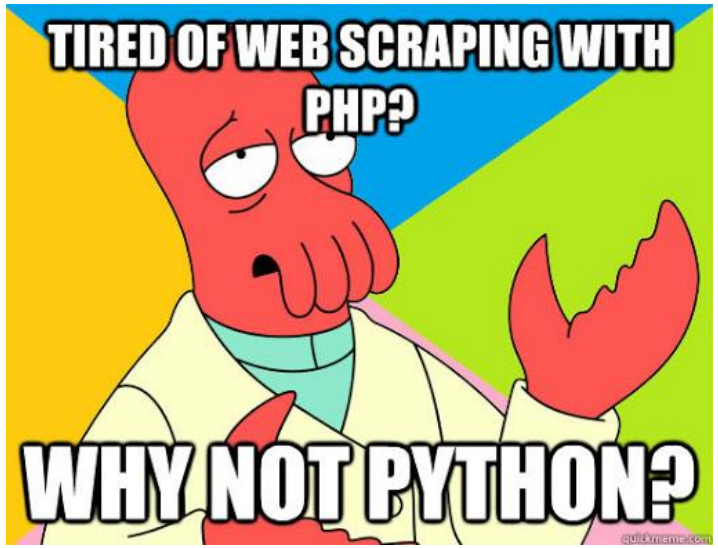
- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time. Google Play Store data is also of huge volume and growing every day.
- Big data is often characterized by the three V's. **Volume, Variety and Velocity**.
- Volume: The number of apps in Google Play Store is almost **5 Million**.
- Variety: For each app there are various types of data available like number of installs, score, size, reviews, content rating, price etc.
- Velocity: These data are generated in an enormous speed. New apps, updates of apps, reviews and scores given by users and many other things contribute in generation of data. For effective analysis and visualisation these are to be processed in similar speed.

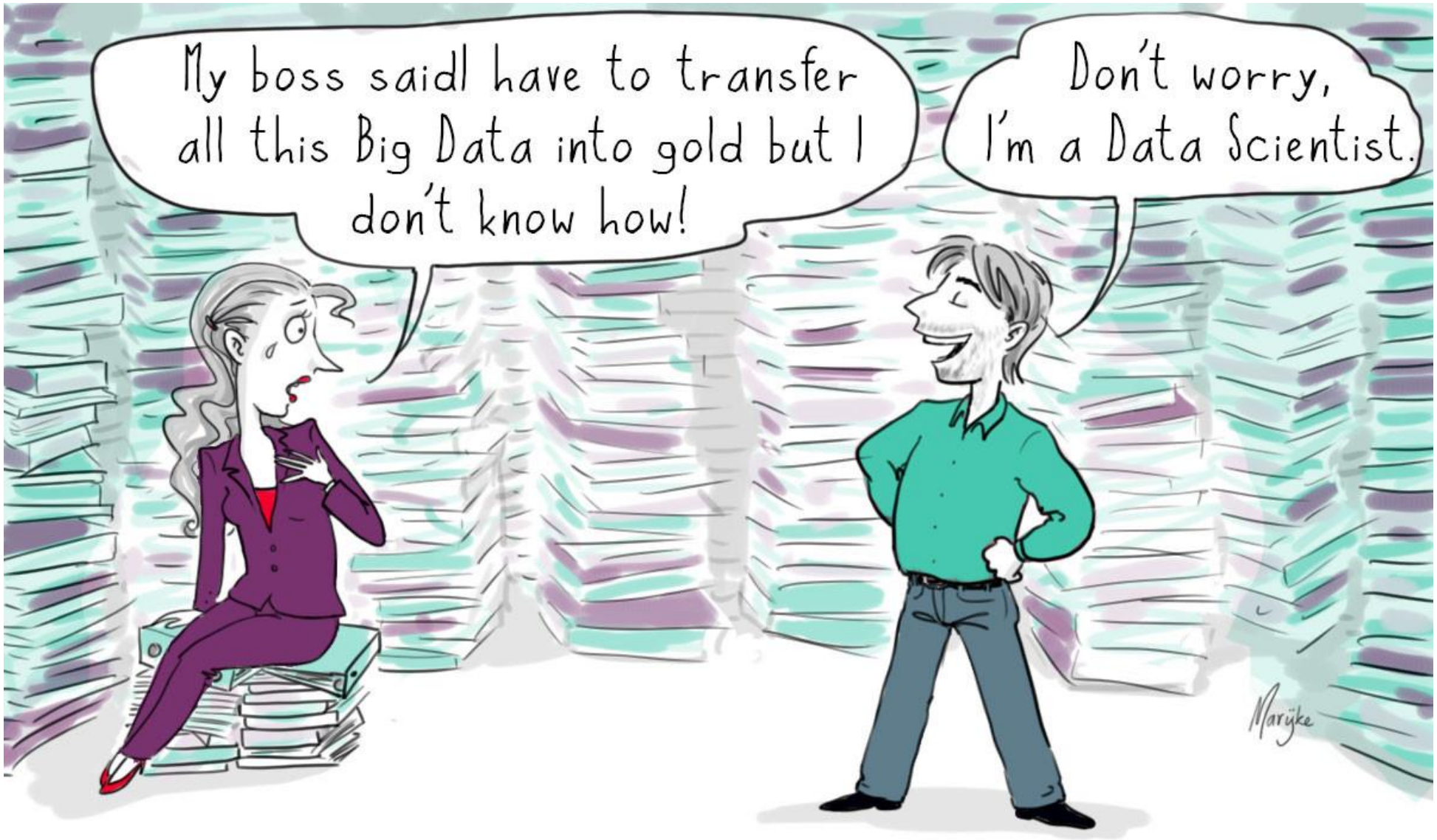




Web Scraping

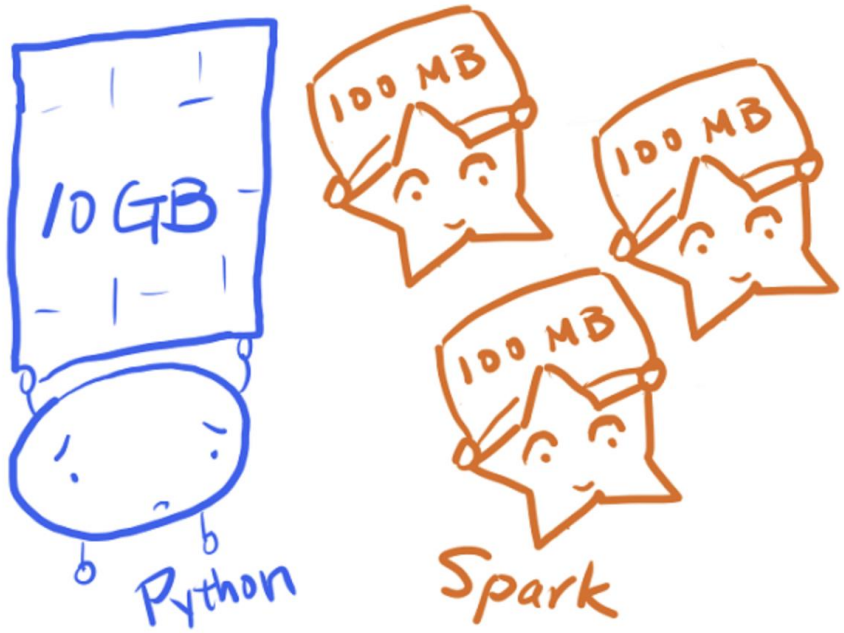
- We designed a code which will crawl over the Play-store using the links provided and collect all relevant data.
- The crawler visits each of the links of different categories of app, waits for 5 seconds for the page to load and scrolls down to the bottom of the page.
- Our bot waits for 3 seconds for the new apps to show up and will continue doing this until there are no more apps to show.
- At this stage, it has the app-id's of all the apps it has come across and will visit the app-page to collect the relevant data.
- All this data is then transferred to a CSV file.

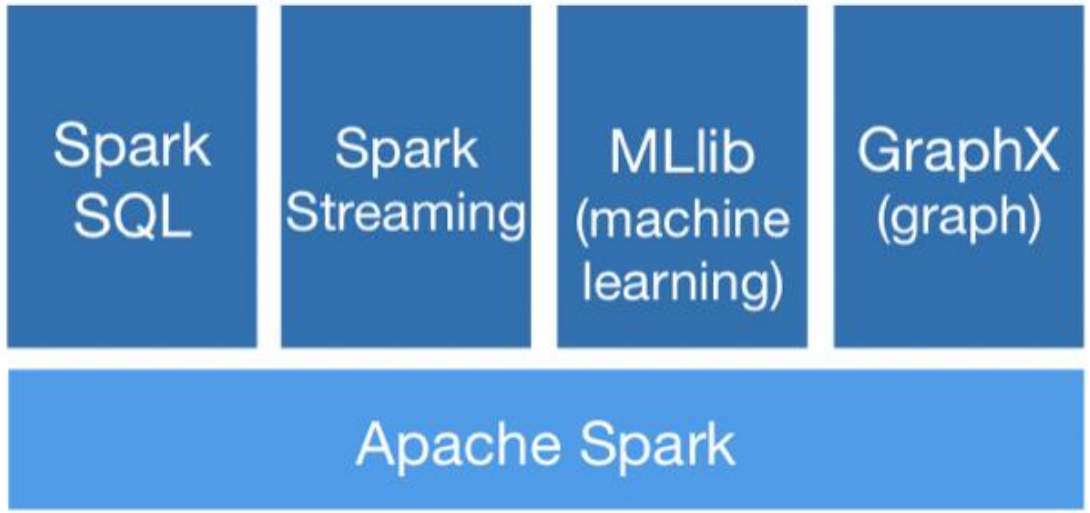


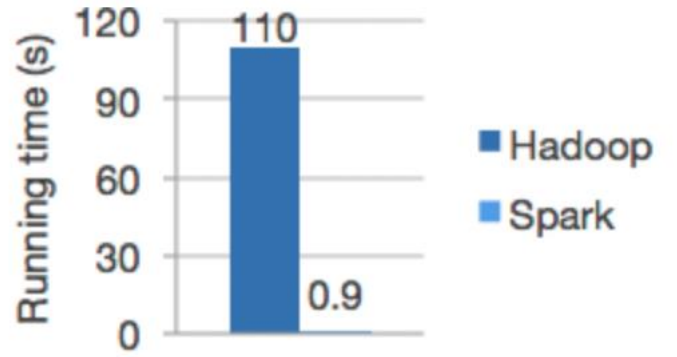

My boss saidI have to transfer all this Big Data into gold but I don't know how!
Don't worry, I'm a Data Scientist.
Marijke

Why spark?

- Spark has useful features such as Spark SQL, DataFrame, Streaming, MLlib and Spark Core.
- When the dataset is very large, Python will load the data all at once and will perform all operations on the whole data. Most of the time this becomes not possible due to our primary memory constraint.
- Spark re-partitions the data and picks it up one at a time. This way we can handle huge data even with comparatively less resources.
- Apache Spark achieves high performance for both batch and streaming data resulting in running workloads 100 times faster than Hadoop.



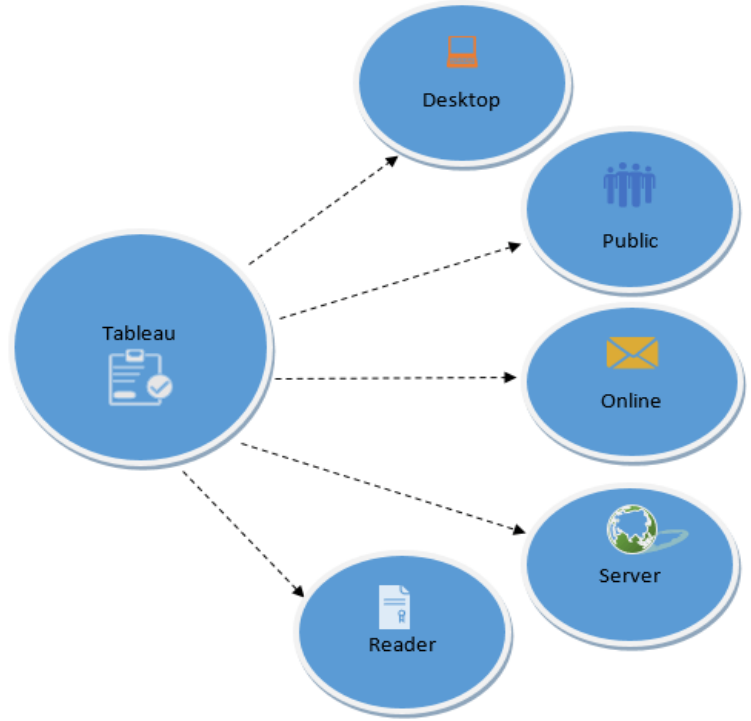


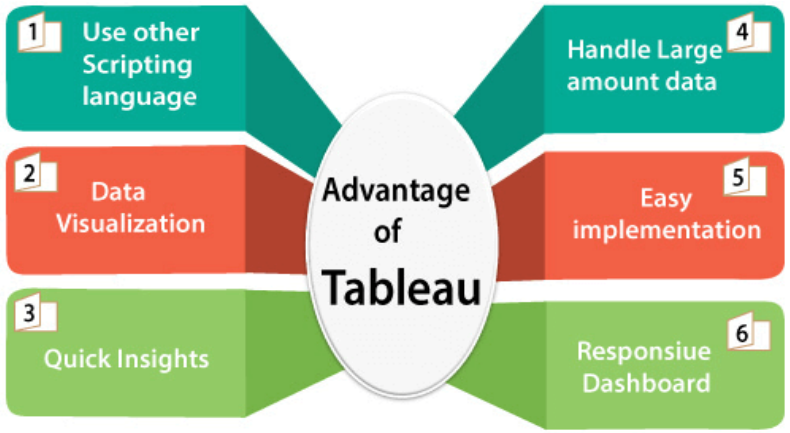


Logistic regression in Hadoop and Spark

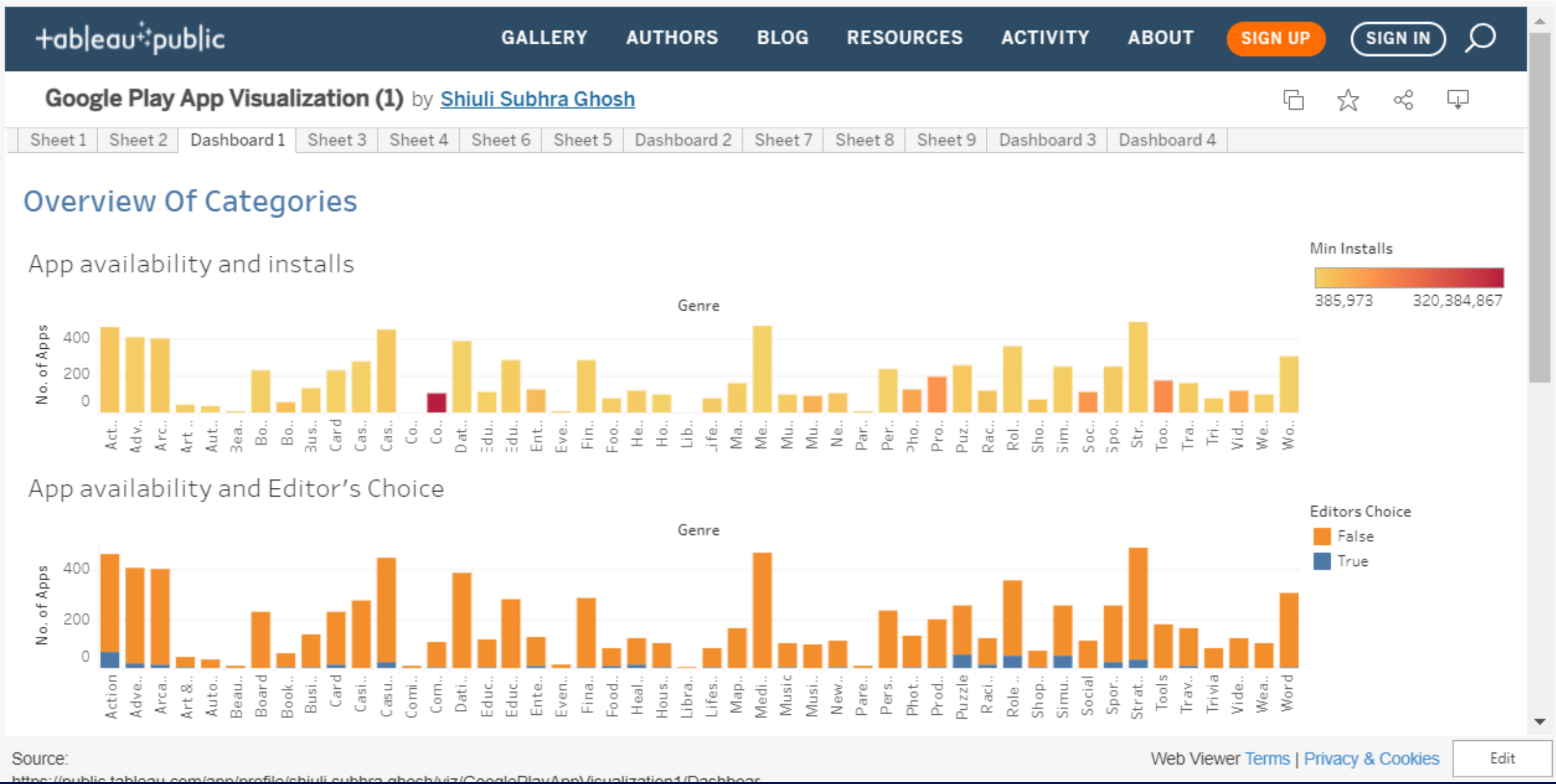
Why Tableau?

- Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions.
- Tableau Software is a tool that helps make Big Data small, and small data insightful and actionable.
- The main use of tableau software is to help people see and understand their data.
- One of the coolest Tableau uses is performing basic ETL operations fast!
- Different users can collaborate on the same project seamlessly in Tableau.
- Tableau comes with beautiful dashboards for better visualization.

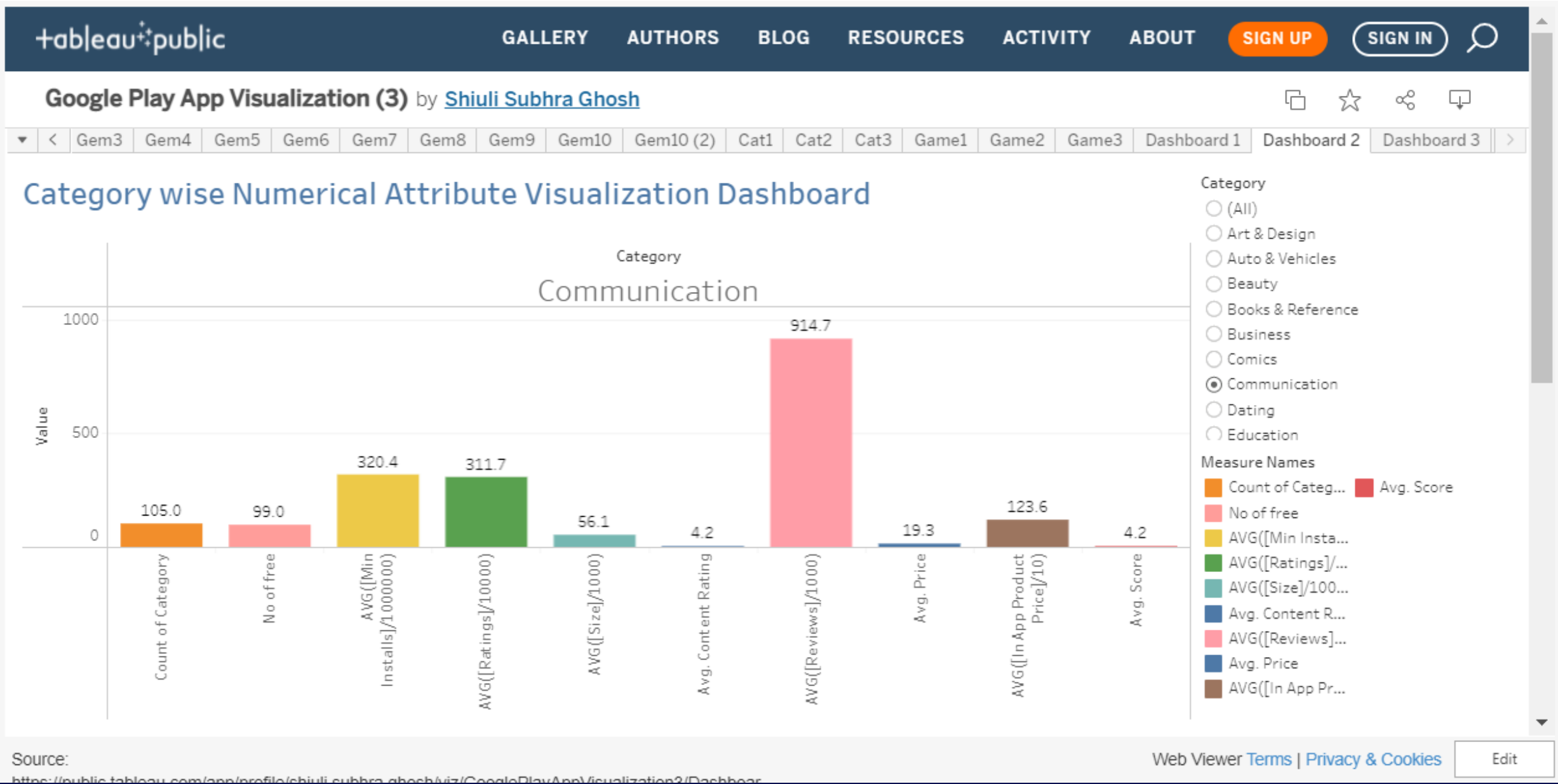




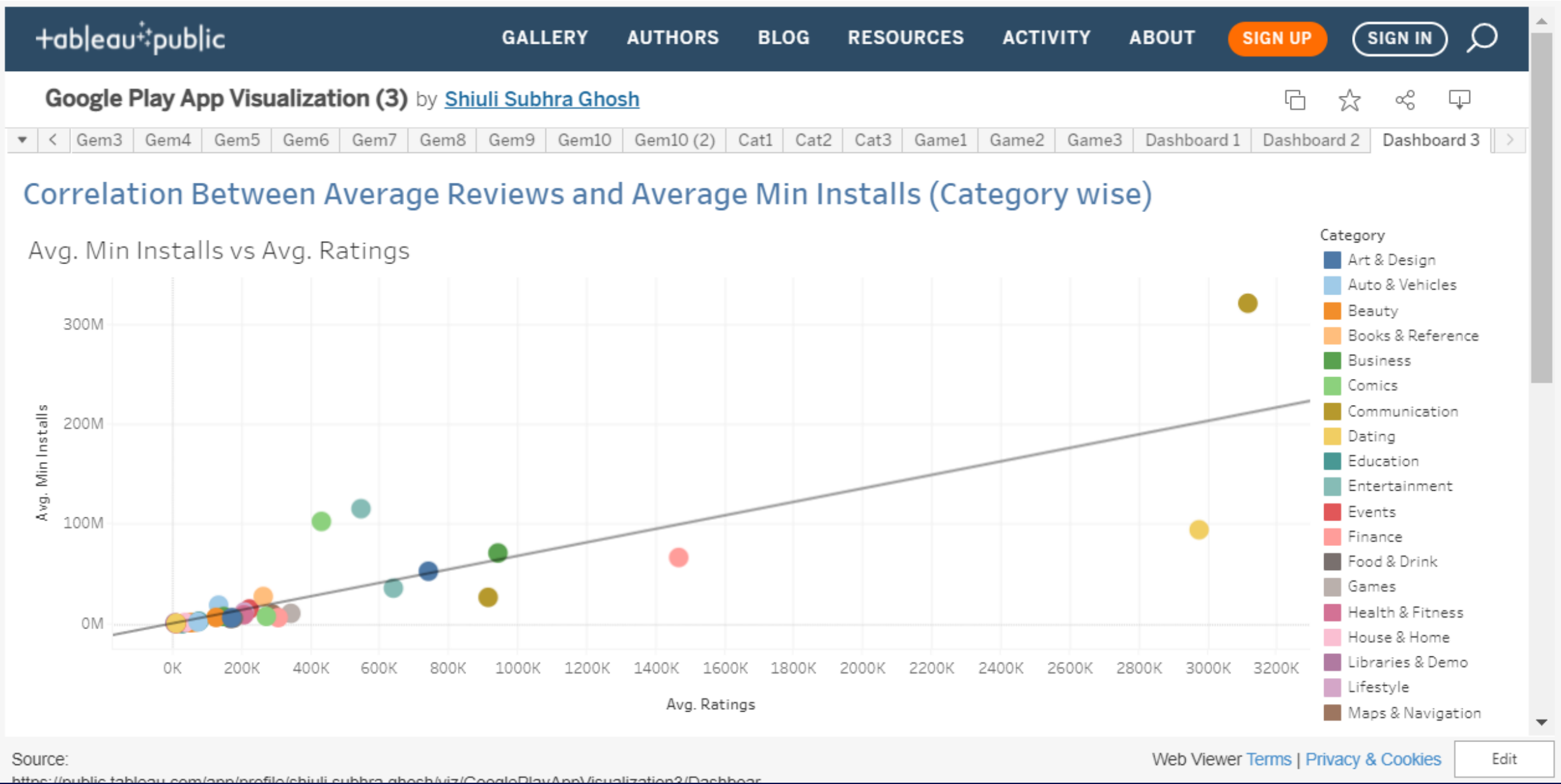
Overview of Data (Category Wise)



Overview of Data (Category wise)



Overview of Data (Category wise)



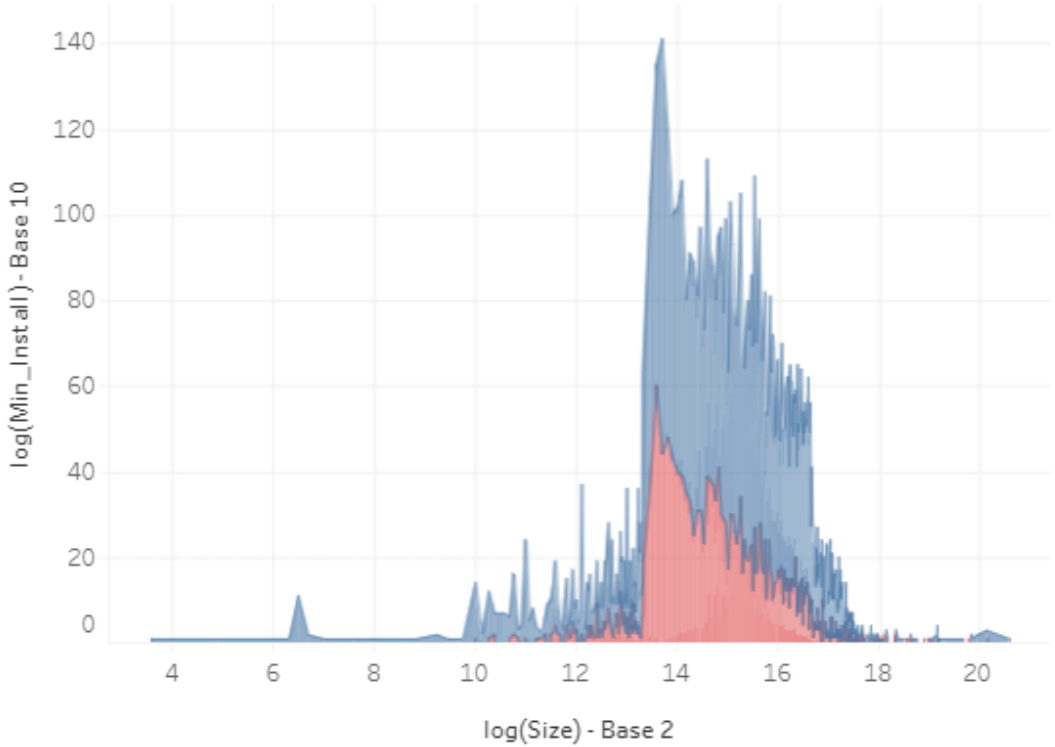
+ableau+public GALLERY AUTHORS BLOG RESOURCES ACTIVITY ABOUT SIGN UP SIGN IN

Google Play App Visualization (1) by Shiuli Subhra Ghosh

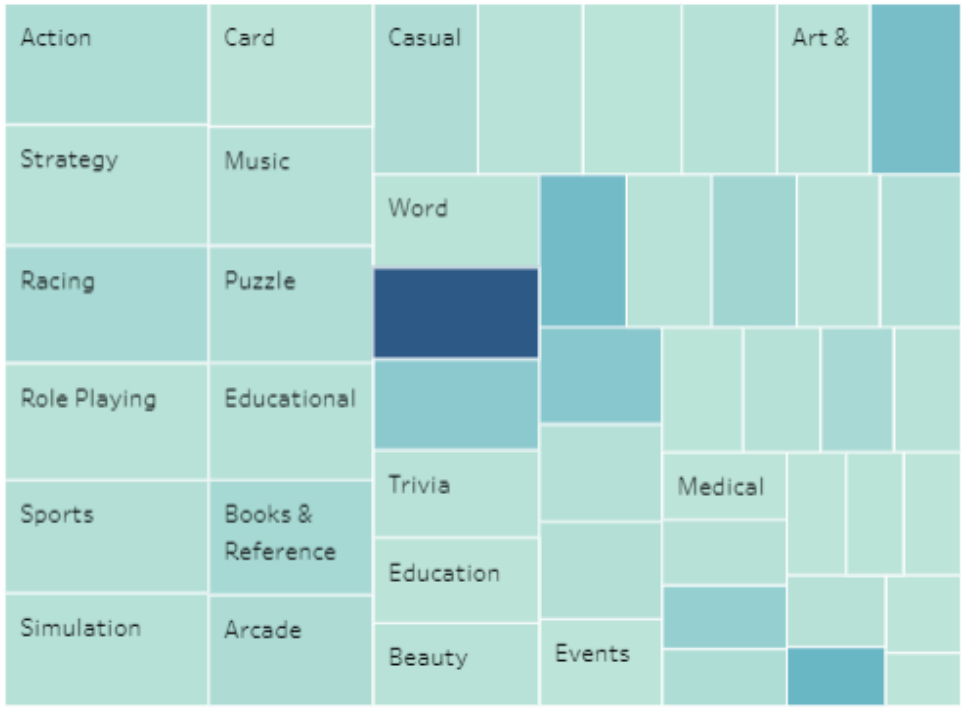
Sheet 1 | Sheet 2 | Dashboard 1 | Sheet 3 | Sheet 4 | Sheet 6 | Sheet 5 | Dashboard 2 | Sheet 7 | Sheet 8 | Sheet 9 | Dashboard 3 | Dashboard 4

Diving Deep Into Continuous Variables

Comparison on Downloads and Size



Category-wise Size



App-Permission: Camera

Numerical Parameters: Min Installs

Gradient: 385,973 – 320M

Genre: (All)

Permissions: False, True

Source:

Web Viewer Terms | Privacy & Cookies Edit

Visualization of Google Apps based on Minimum Android Version Supported

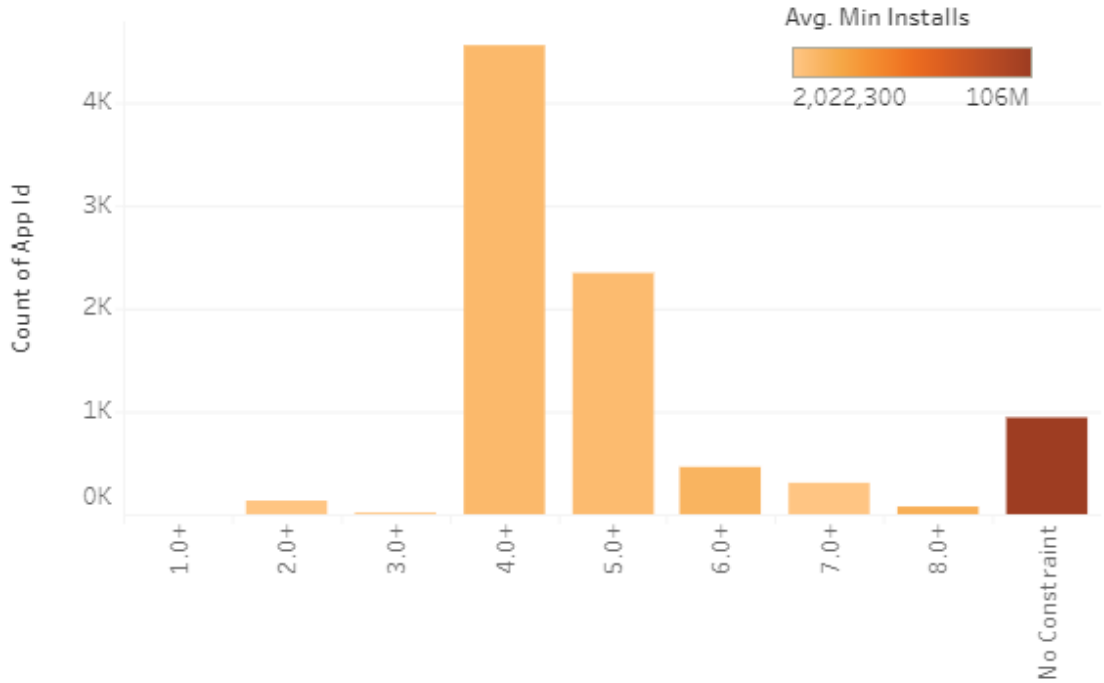
tableau public GALLERY AUTHORS BLOG RESOURCES ACTIVITY ABOUT SIGN UP SIGN IN

Google Play App Visualization (2) by Shiuli Subhra Ghosh

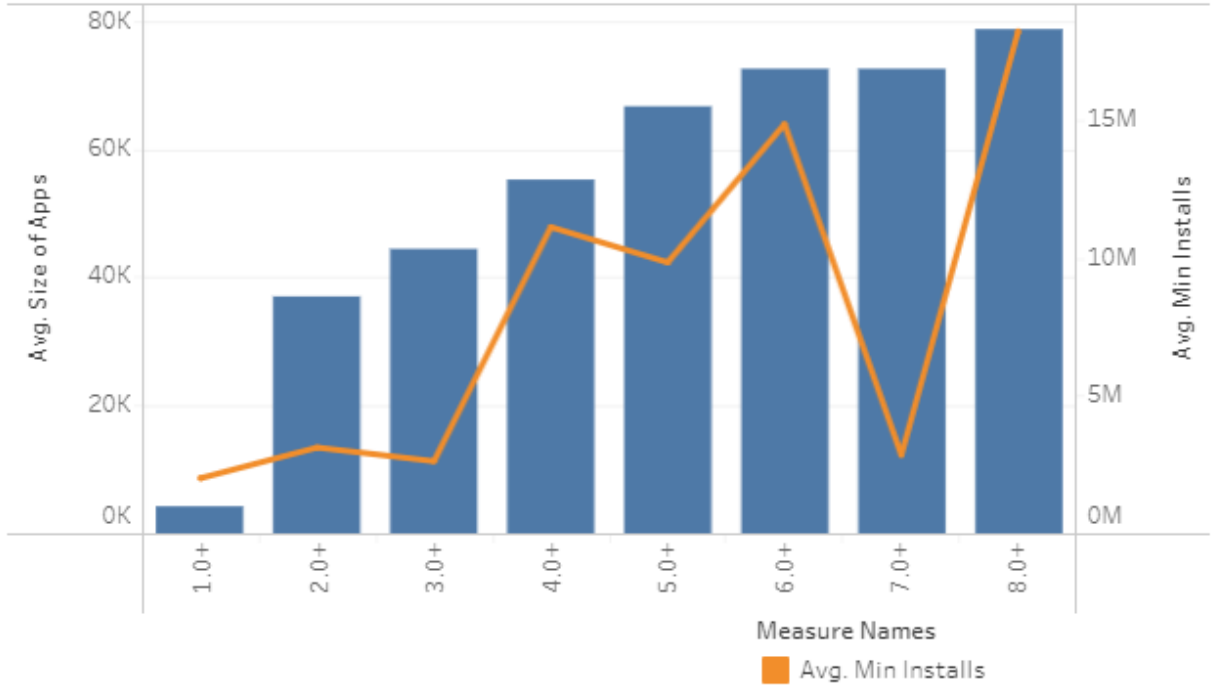
Sheet 1 | Sheet 2 | Sheet 3 | Sheet 4 | Sheet 5 | Sheet 5 (2) | Sheet 6 | Sheet 8 | Visualtisation of Google Apps ... | Dashboard 4 | Sheet 9 | Sheet 9 (2) | Sheet 9 (3) | Das

Visualisation of Google Apps based on Minimum Android Version Supported

Count of Apps and Average Minimum Install with respect to **Android Version**



Average Size and Average Minimum Install of Apps with respect to **Android Version**



Source:

Web Viewer Terms | Privacy & Cookies Edit

Distribution of Apps based on Content rating (Everybody/Adult)

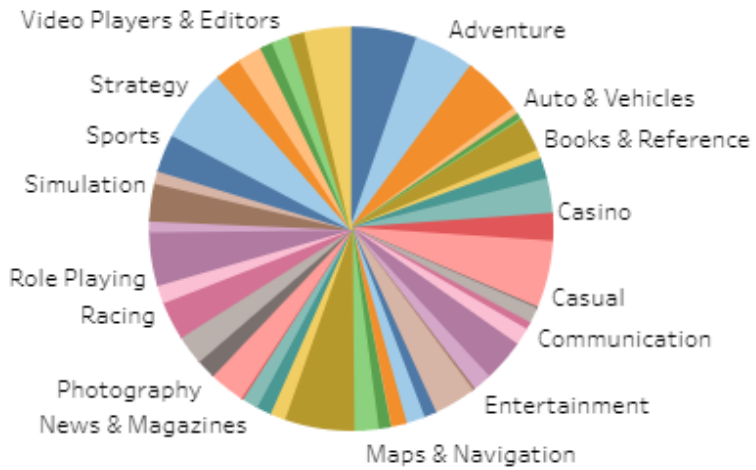
tableau public GALLERY AUTHORS BLOG RESOURCES ACTIVITY ABOUT SIGN UP SIGN IN

Google Play App Visualization (2) by Shiuli Subhra Ghosh

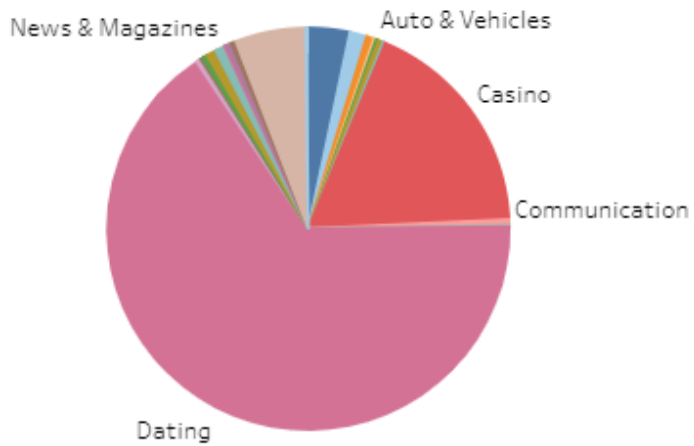
Sheet 1 | Sheet 2 | Sheet 3 | Sheet 4 | Sheet 5 | Sheet 5 (2) | Sheet 6 | Sheet 8 | Visualtisation of Google Apps ... | Dashboard 4 | Sheet 9 | Sheet 9 (2) | Sheet 9 (3) | Das

Distribution of Apps based on Content Rating(Everybody/Adult)

For Everybody



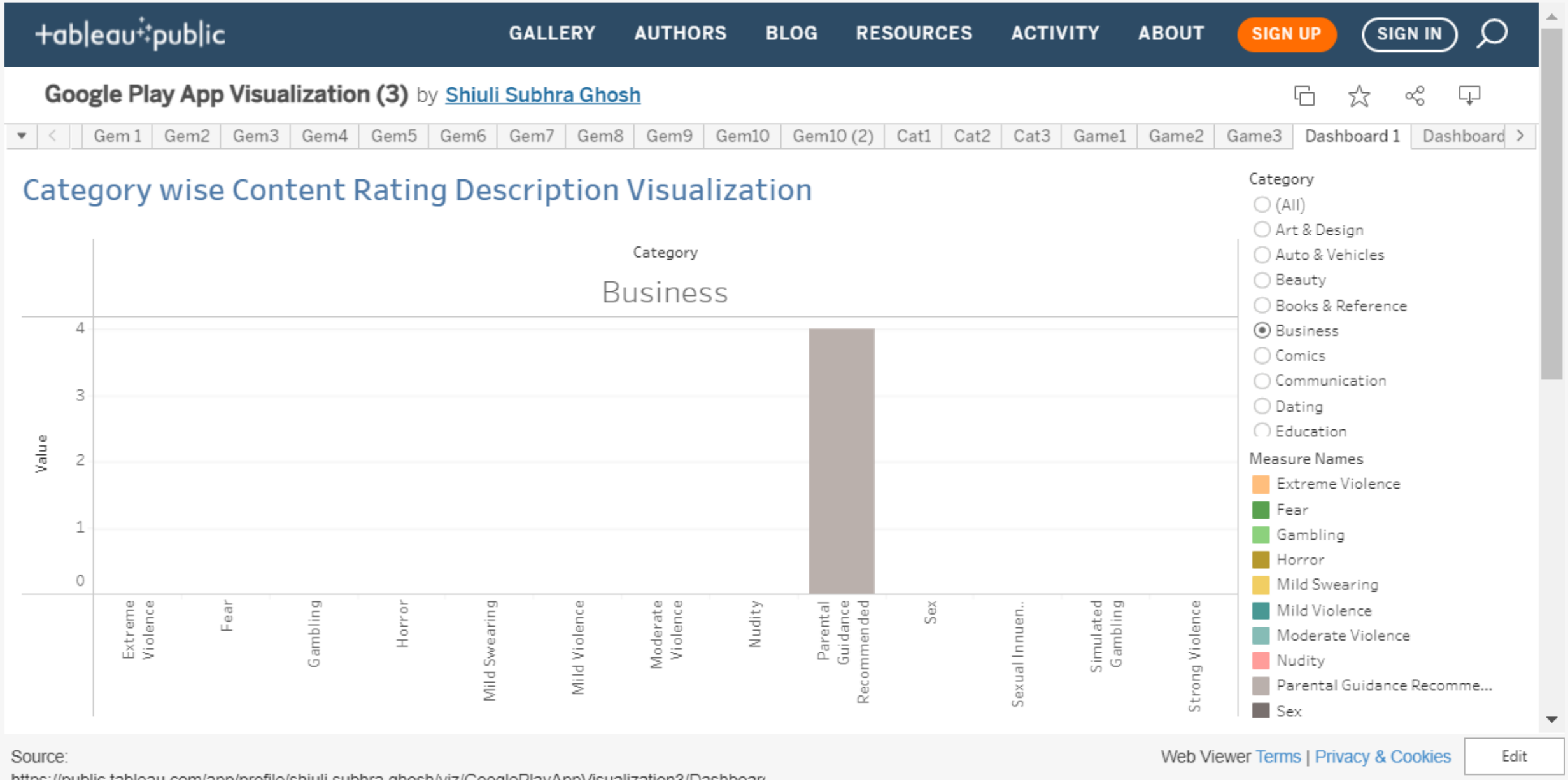
For Adults



Source:

Web Viewer Terms | Privacy & Cookies Edit

Category wise Content Rating Description Visualization



Visualizing Editor's Choice

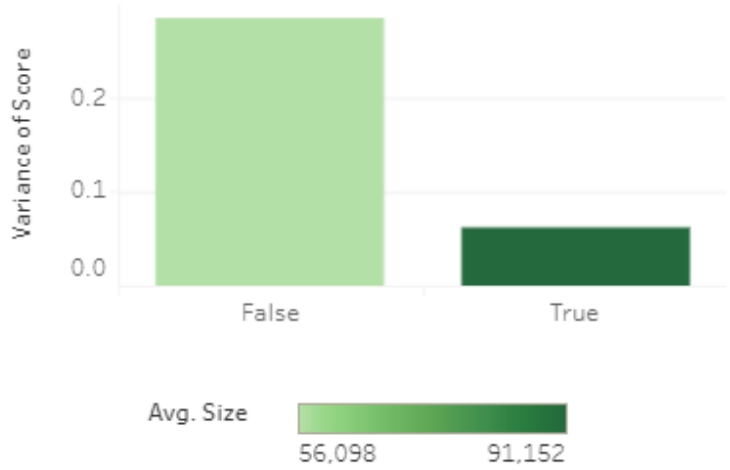
tableau public GALLERY AUTHORS BLOG RESOURCES ACTIVITY ABOUT SIGN UP SIGN IN

Google Play App Visualization (2) by Shiuli Subhra Ghosh

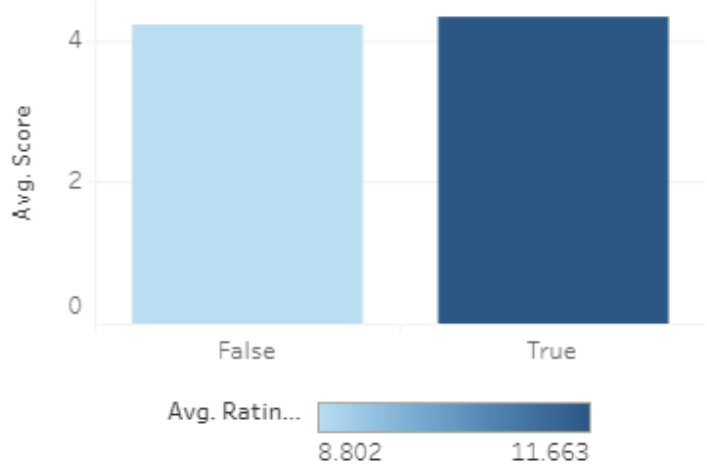
ıaltisation of Google Apps ... | Dashboard 4 | Sheet 9 | Sheet 9 (2) | Sheet 9 (3) | Dashboard 5 | Dashboard 5_1 | Sheet 10 | Sheet 13 | Sheet 14 | Sheet 15 | Dashboard 6 | She

Editor's Choice

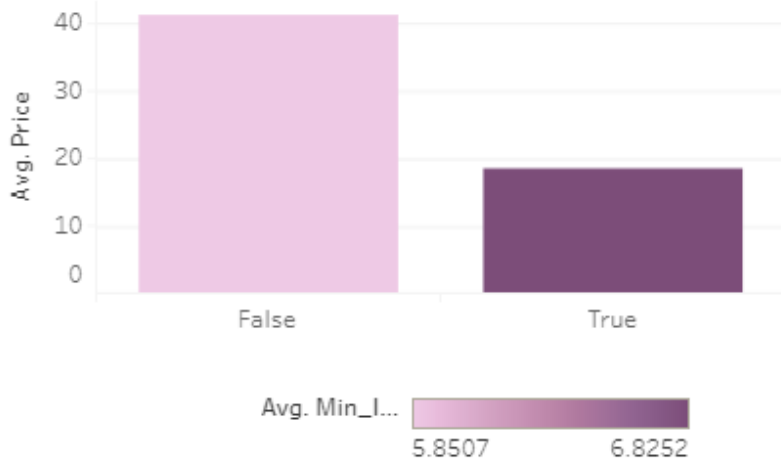
Variance of Score with respect to Editor's choice selection



Average Score of Apps with Respect to Editor's choice



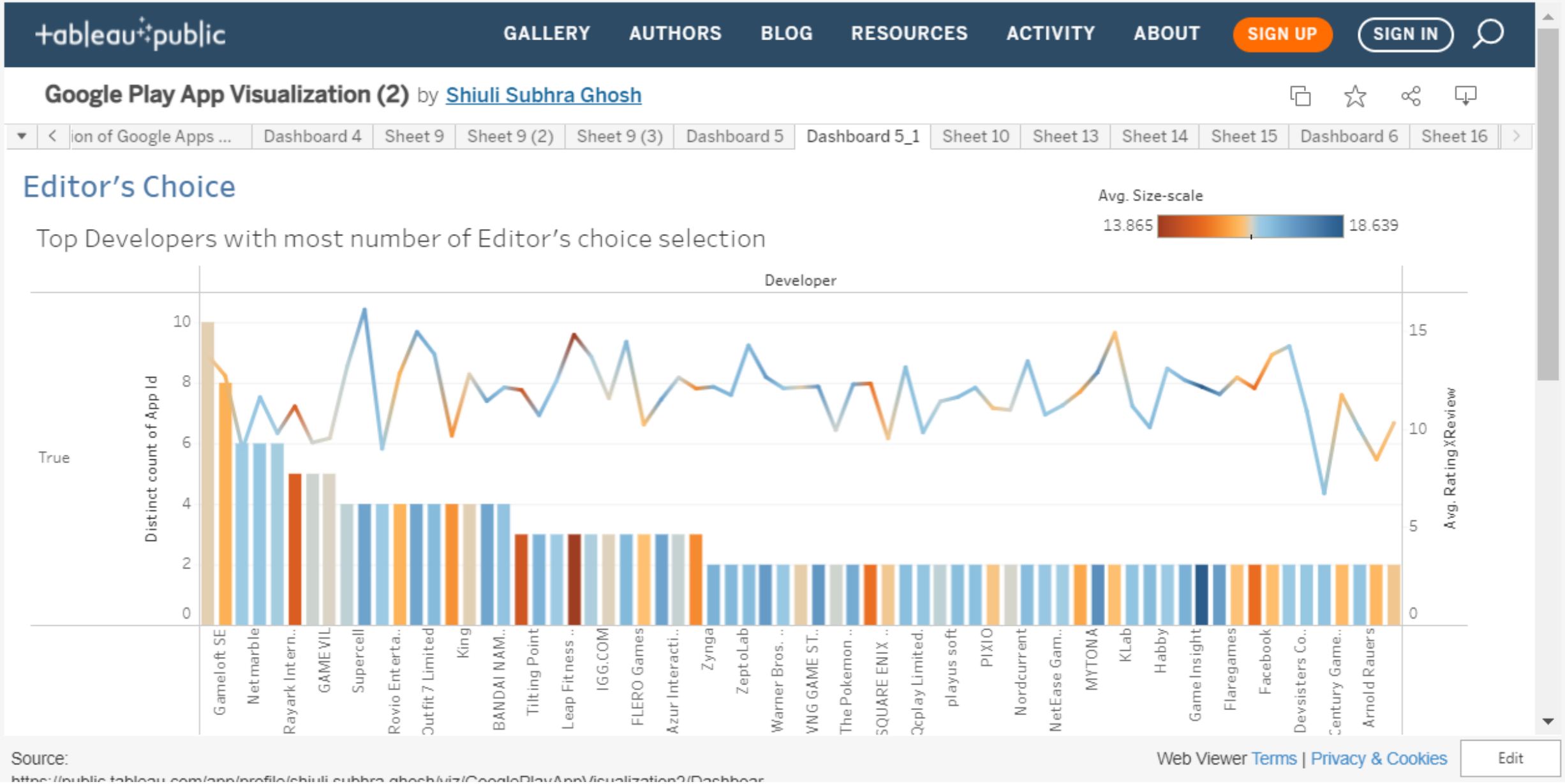
Average Price with respect to Editor's choice selection



Source:

Web Viewer Terms | Privacy & Cookies Edit

Visualizing Editor's Choice



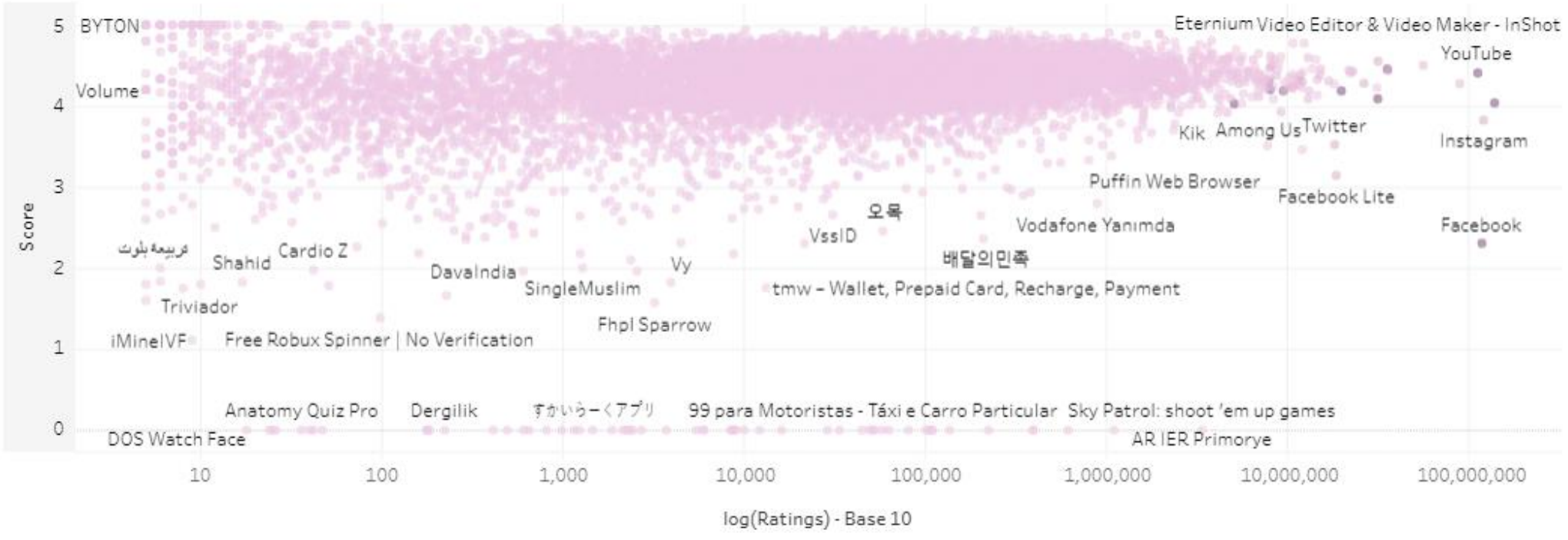
+ableau+public GALLERY AUTHORS BLOG RESOURCES ACTIVITY ABOUT SIGN UP SIGN IN

Google Play App Visualization (1) by Shiuli Subhra Ghosh

Sheet 1 | Sheet 2 | Dashboard 1 | Sheet 3 | Sheet 4 | Sheet 6 | Sheet 5 | Dashboard 2 | Sheet 7 | Sheet 8 | Sheet 9 | Dashboard 3 | Dashboard 4

The Industry

Score and Ratings Of Apps



Min Installs
10 5B

Genre
- (All)
- Action
- Adventure
- Arcade
- Art & Design
- Auto & Vehicles
- Beauty
- Board
- Books & Reference
- Business
- Card
- Casino
- Casual
- Comics
- Communication

Source:

Web Viewer Terms | Privacy & Cookies Edit

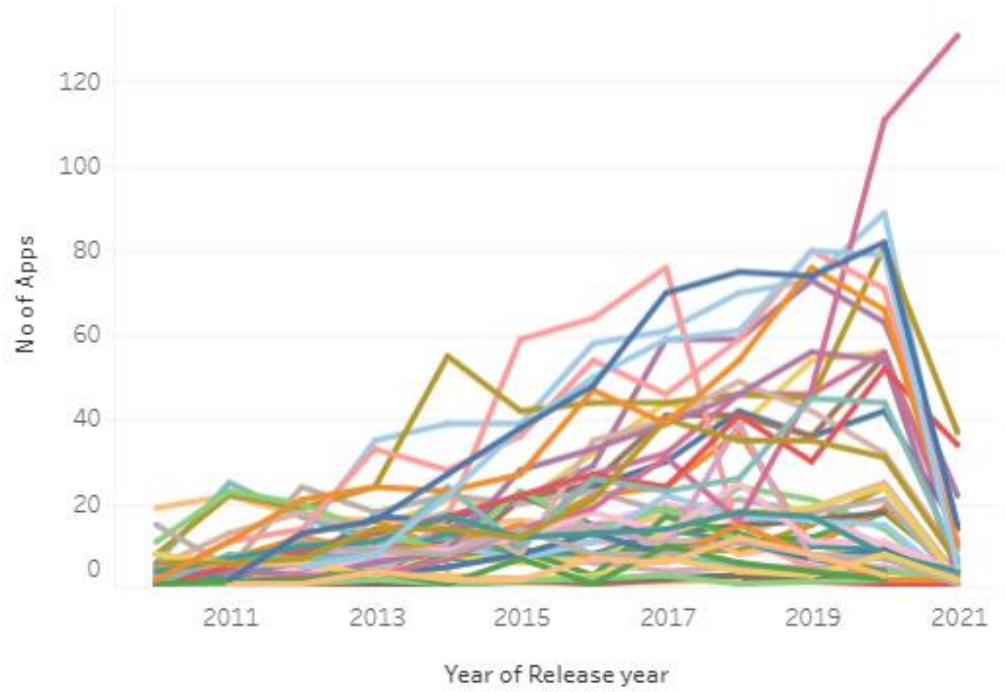
+ableau+public GALLERY AUTHORS BLOG RESOURCES ACTIVITY ABOUT SIGN UP SIGN IN

Google Play App Visualization (1) by Shiuli Subhra Ghosh

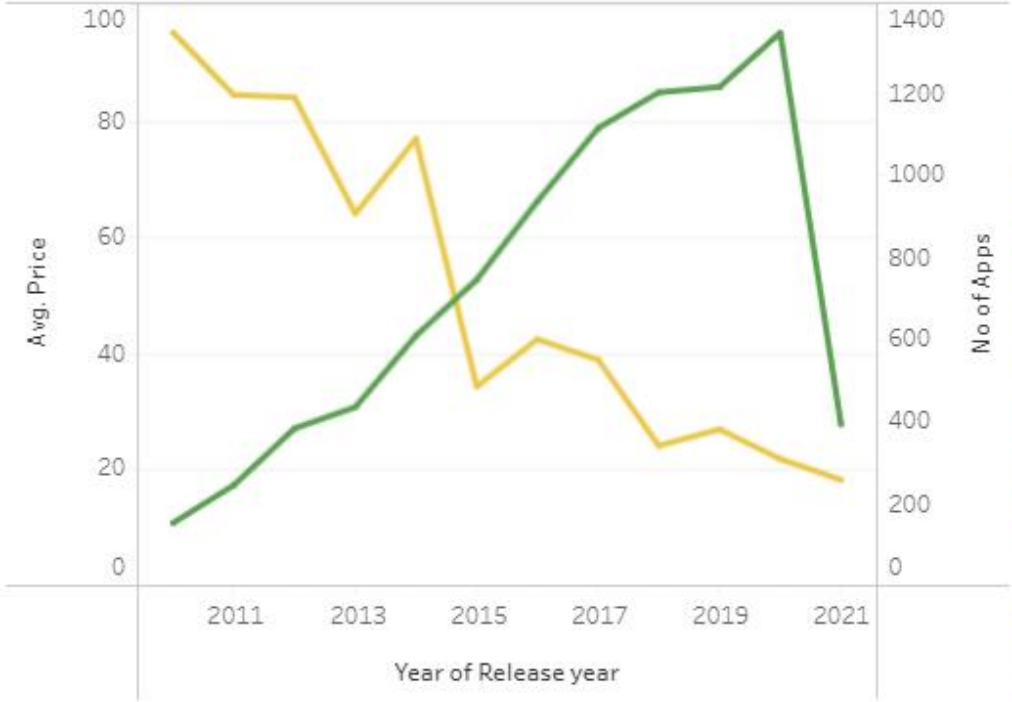
Sheet 1 | Sheet 2 | Dashboard 1 | Sheet 3 | Sheet 4 | Sheet 6 | Sheet 5 | Dashboard 2 | Sheet 7 | Sheet 8 | Sheet 9 | Dashboard 3 | Dashboard 4

The Industry

App Availability Over The Years



Growth Of The Industry and Pricing Scheme



Genre
(All)

Genre
Action
Adventure
Arcade
Art & Design
Auto & Vehicles
Beauty
Board
Books & Reference
Business
Card
Casino
Casual
Comics

Measure Names
Avg. Price
Distinct count of App...

Source:
https://public.tableau.com/app/profile/shiuli.subhra.ghosh/viz/GooglePlayAppVisualization1/Dashboar

Web Viewer Terms | Privacy & Cookies Edit

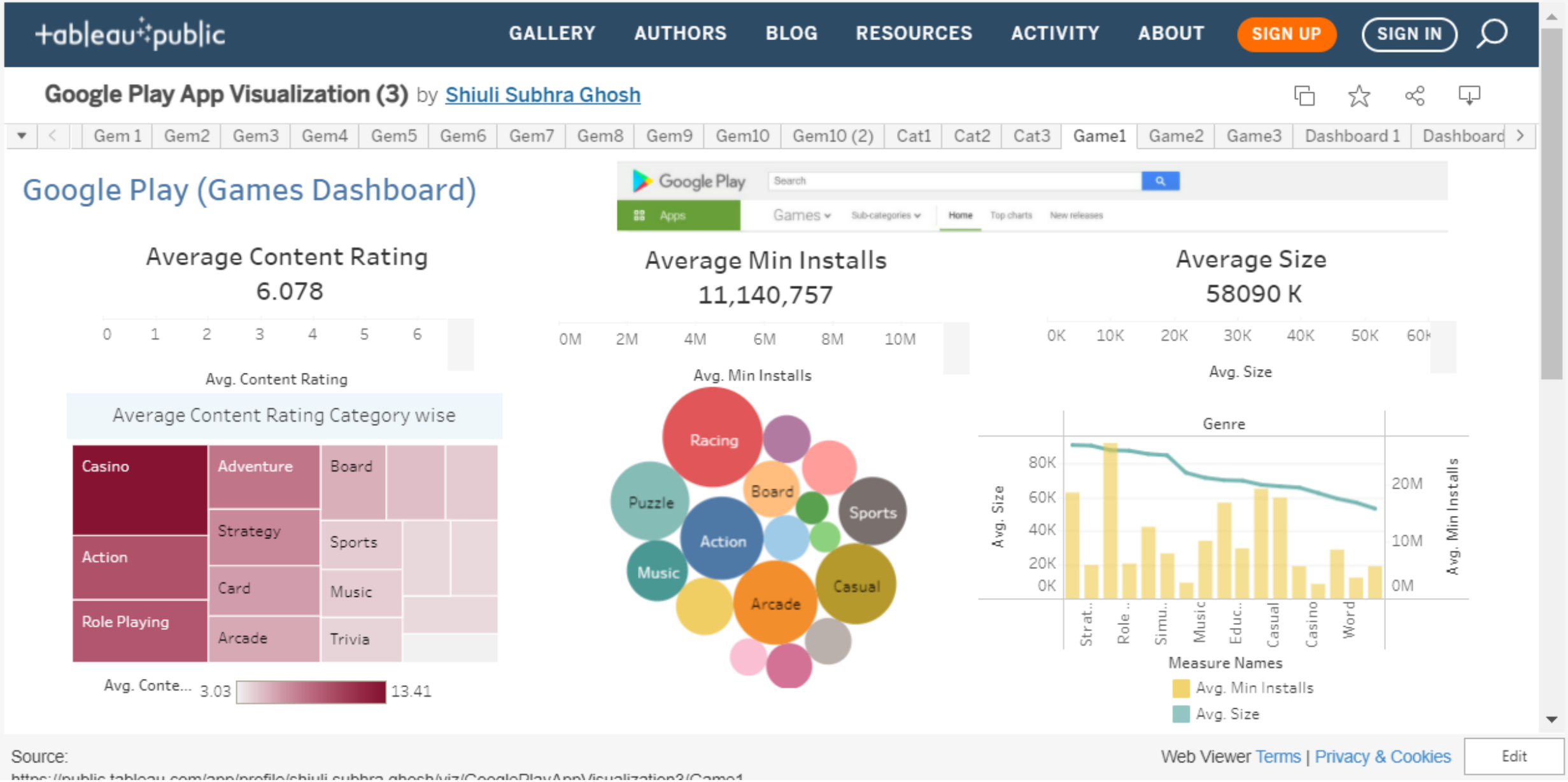


tableau public
GALLERY
AUTHORS
BLOG
RESOURCES
ACTIVITY
ABOUT
SIGN UP
SIGN IN
Google Play App Visualization (3) by Shiuli Subhra Ghosh
Gem 1
Gem2
Gem3
Gem4
Gem5
Gem6
Gem7
Gem8
Gem9
Gem10
Gem10 (2)
Cat1
Cat2
Cat3
Game1
Game2
Game3
Dashboard 1
Dashboard
Google Play (Games Dashboard)
Average Content Rating
6.078
0
1
2
3
4
5
6
Avg. Content Rating
Average Content Rating Category wise
Casino
Adventure
Board
Strategy
Sports
Action
Card
Music
Role Playing
Arcade
Trivia
Avg. Conte...
3.03
13.41
Google Play
Search
Apps
Games
Sub-categories
Home
Top charts
New releases
Average Min Installs
11,140,757
0M
2M
4M
6M
8M
10M
Avg. Min Installs
Racing
Board
Puzzle
Sports
Action
Music
Arcade
Casual
Average Size
58090 K
0K
10K
20K
30K
40K
50K
60K
Avg. Size
Genre
80K
60K
40K
20K
0K
Avg. Size
20M
10M
0M
Avg. Min Installs
Strat..
Role ..
Simu..
Music
Educ..
Casual
Casino
Word
Measure Names
Avg. Min Installs
Avg. Size
Source:
Web Viewer Terms | Privacy & Cookies
Edit

Game Dashboard Category wise Visualization

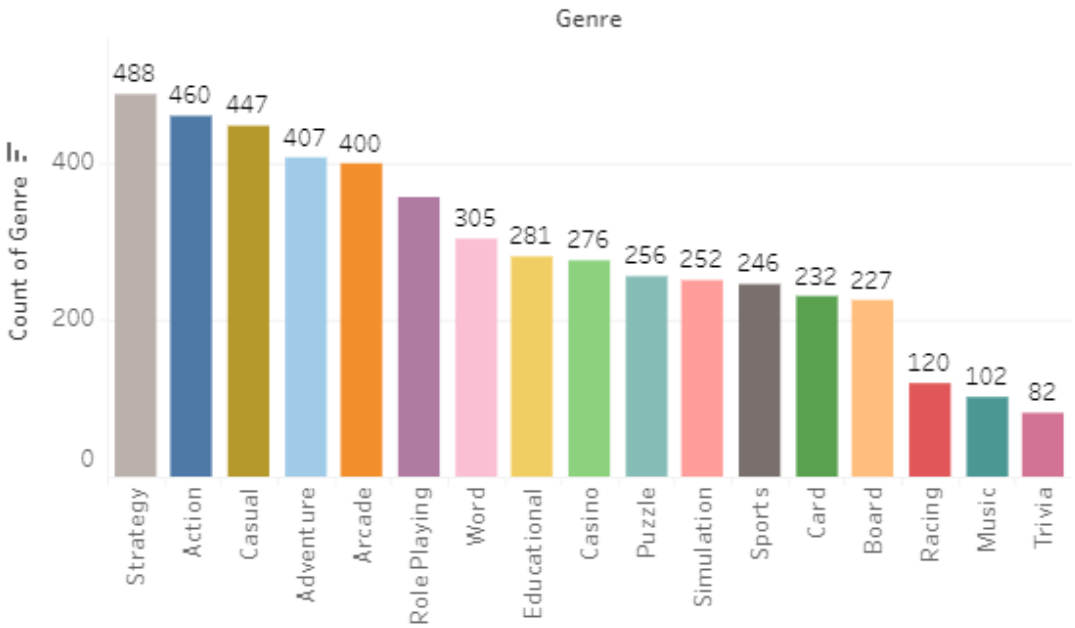
tableau public GALLERY AUTHORS BLOG RESOURCES ACTIVITY ABOUT SIGN UP SIGN IN

Google Play App Visualization (3) by Shiuli Subhra Ghosh

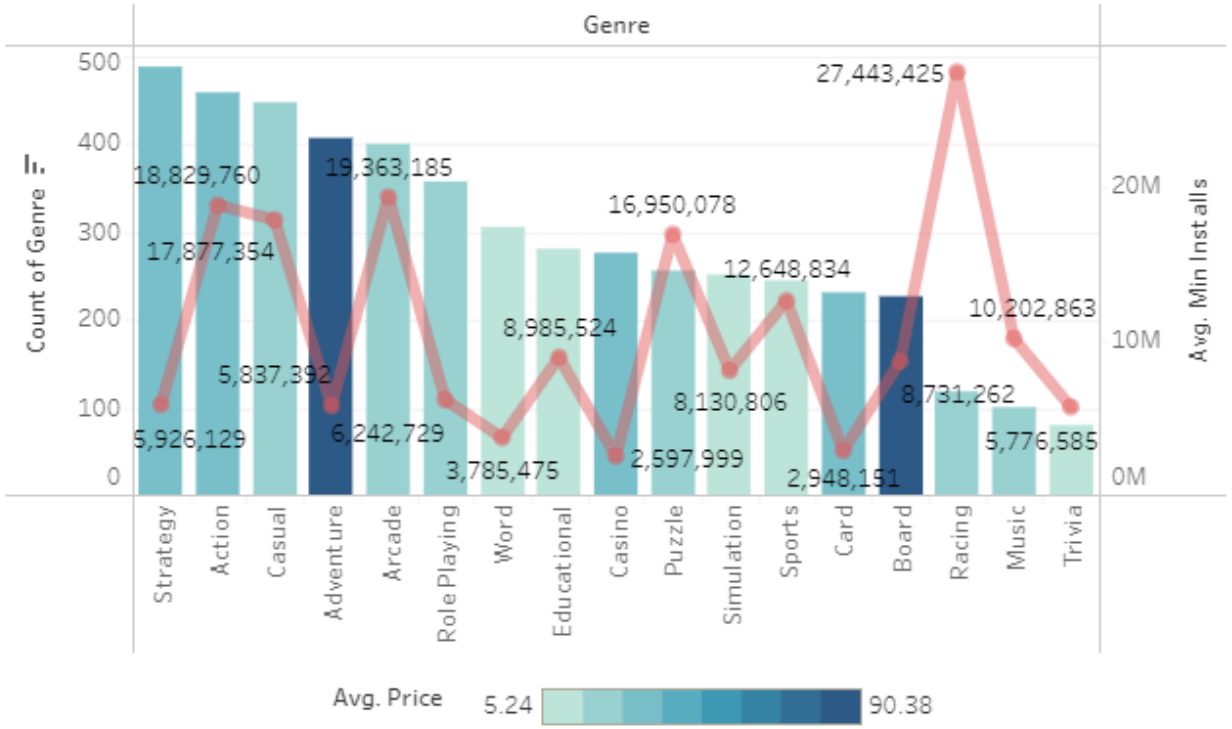
Gem 1 | Gem2 | Gem3 | Gem4 | Gem5 | Gem6 | Gem7 | Gem8 | Gem9 | Gem10 | Gem10 (2) | Cat1 | Cat2 | Cat3 | Game1 | Game2 | Game3 | Dashboard 1 | Dashboard

Google Play (Games Dashboard)

Games Category



Average price vs min installs



Source:

Web Viewer Terms | Privacy & Cookies Edit

Famous games

Famous Games











Most Famous Games Statistics

Genre	Title	Min Installs	Content Rating	Price	Score	Size
Action	Temple Run 2	500,000,000	7	0	4	133,120
	Sniper 3D: Fun Free Online FPS Shooting G..	500,000,000	18	0	4	139,264
	Garena Free Fire - Rampage	500,000,000	12	0	4	47,104
Arcade	Subway Surfers	1,000,000,000	7	0	4	191,488
	Temple Run	500,000,000	3	0	4	47,104
	Fruit Ninja®	500,000,000	3	0	4	128,000
Board	Ludo King™	500,000,000	3	0	4	50,176
Casual	Candy Crush Saga	1,000,000,000	3	0	5	46,080
	Pou	500,000,000	3	0	4	22,528
	My Talking Tom	500,000,000	3	0	4	116,736
	My Talking Angela	500,000,000	3	0	4	117,760
Racing	Hill Climb Racing	500,000,000	3	0	4	60,416
Sports	8 Ball Pool	500,000,000	12	0	4	68,608
Strategy	Clash of Clans	500,000,000	7	0	4	178,176

Conclusions and Inferences

- What are the top categories on Google Play Store?
- Percentage of free and paid apps.
- Distribution of the content ratings of the apps.
- How does the size of an app influence installation of an app?
- Genre wise analysis of different numerical factors.
- How is an app selected in "Editor's Choice" different from the rest?
- How did the count of apps progress yearly for different genres?

Thank You