Github Scraper
&
Google Playstore Crawler

by
Pratap Timilsina

Objective

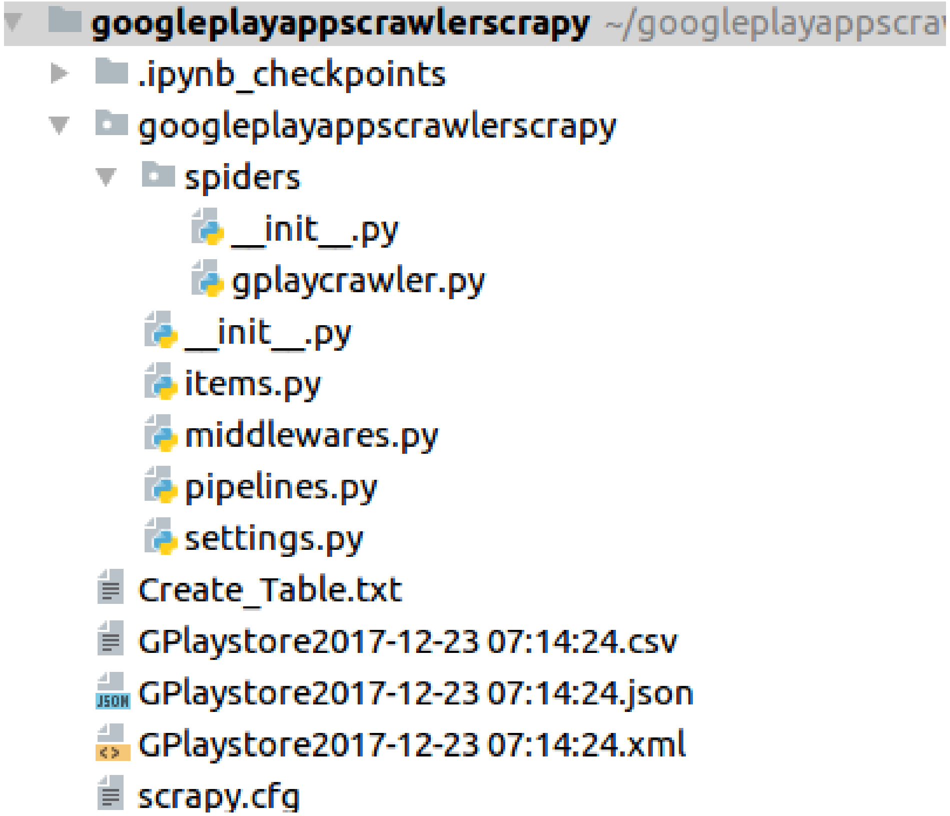
- **Web Scraping using Beautifulsoup**
- **Web Crawling using Scrapy**

Trending Github Repositories Scraping

- Used Beautifulsoup python package
- Daily cron job with email of trends
- Saved file to xml, json, csv, parquet, avro
- Data also pipelined to MySql
- For avro create avsc schema
- Avro doesn’t like trending and trailing spaces

Crawling Google Apps Store

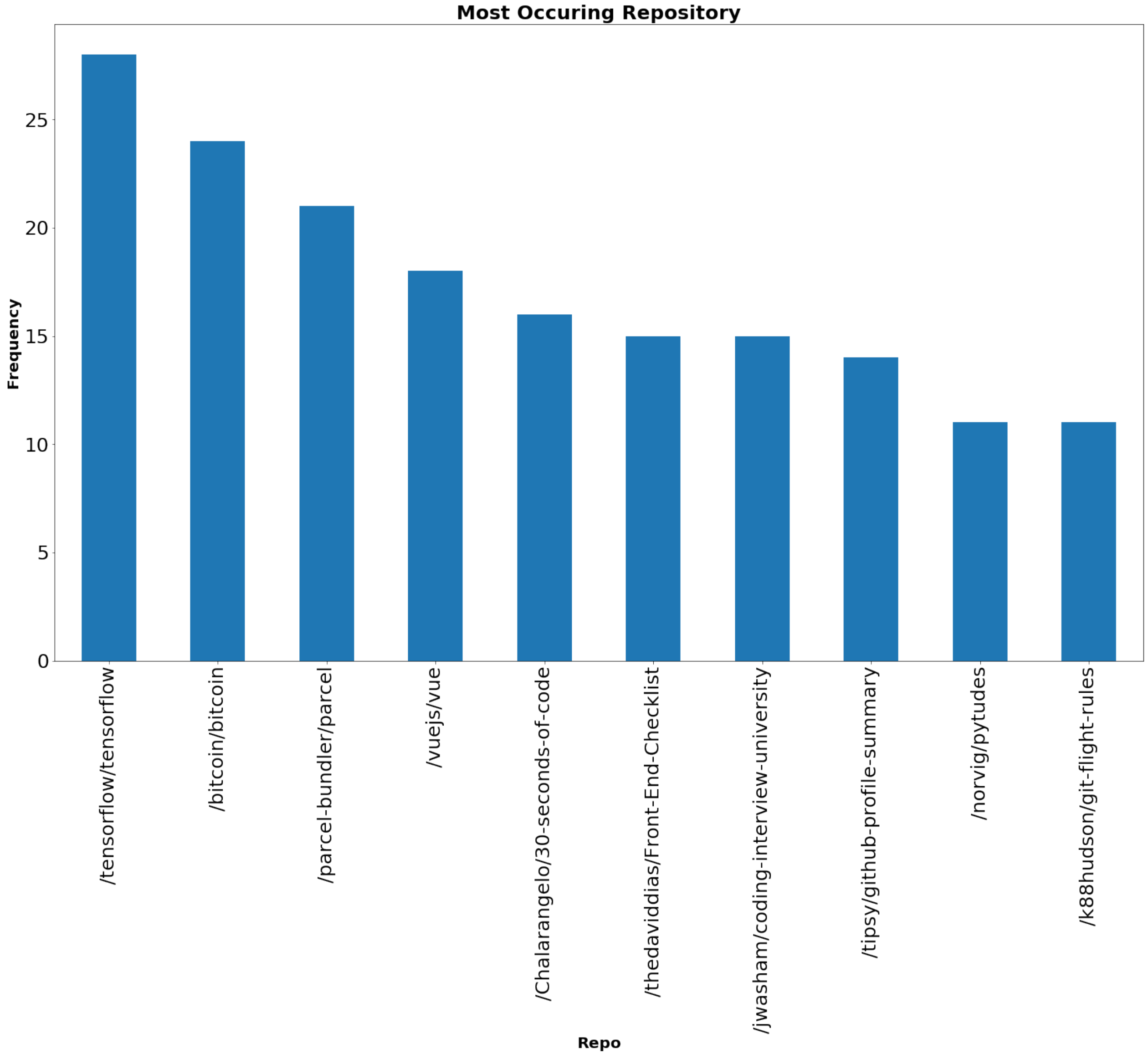
- Used scrapy python package
- Scrapy creates project
- Data saved as csv, xml, json   (well organized)
- Data pipe lined to MySql



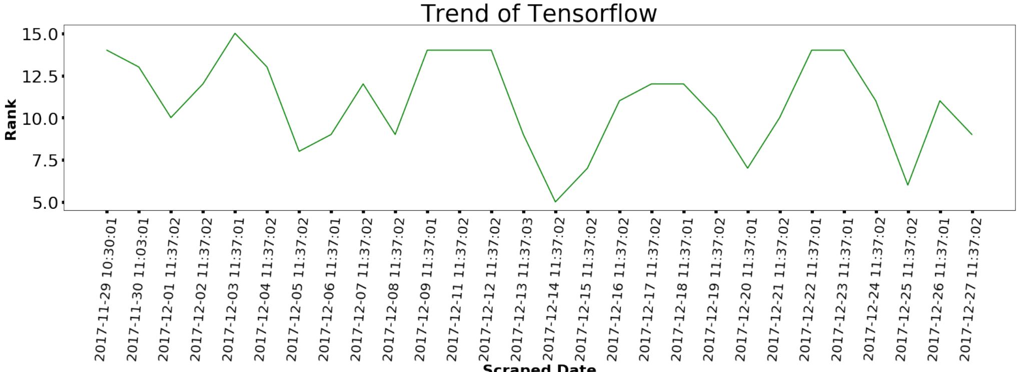
```
googleplayappscrawlerscrapy ~/googleplayappscra
    .ipynb_checkpoints
    googleplayappscrawlerscrapy
        spiders
            __init__.py
            gplaycrawler.py
        __init__.py
        items.py
        middlewares.py
        pipelines.py
        settings.py
    Create_Table.txt
    GPlaystore2017-12-23 07:14:24.csv
    GPlaystore2017-12-23 07:14:24.json
    GPlaystore2017-12-23 07:14:24.xml
    scrapy.cfg
```

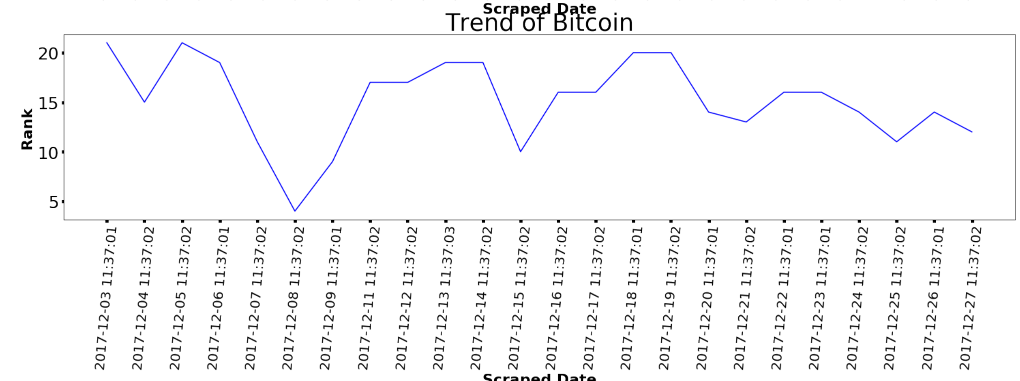
Exploratory Data Analysis
of
Github Trends Data

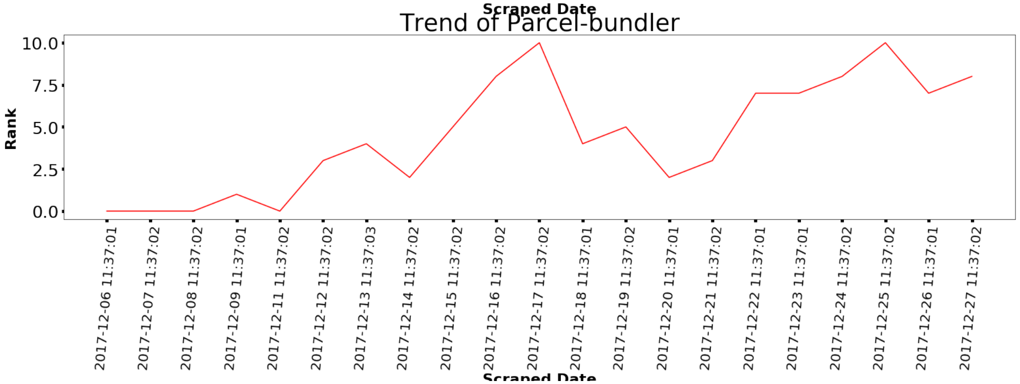
Git Hub Trending Repository

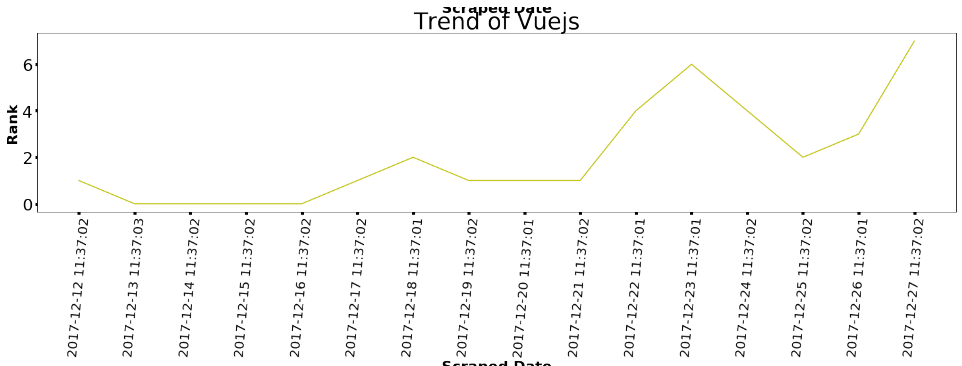


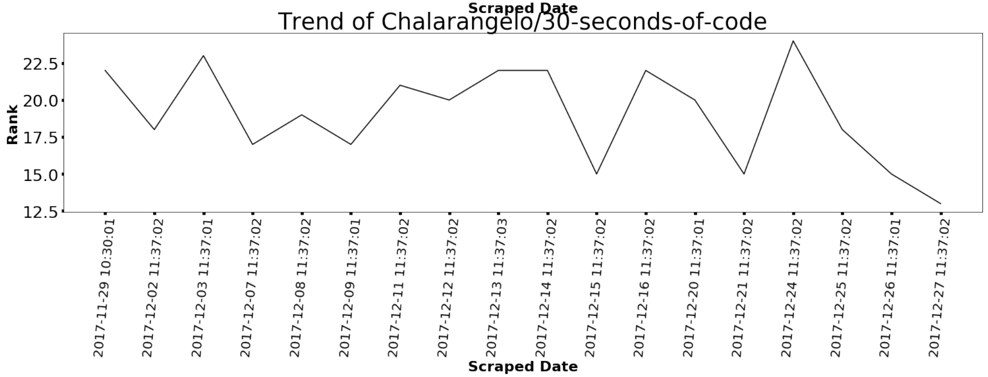
Git Hub Repos Ranking Time Trends



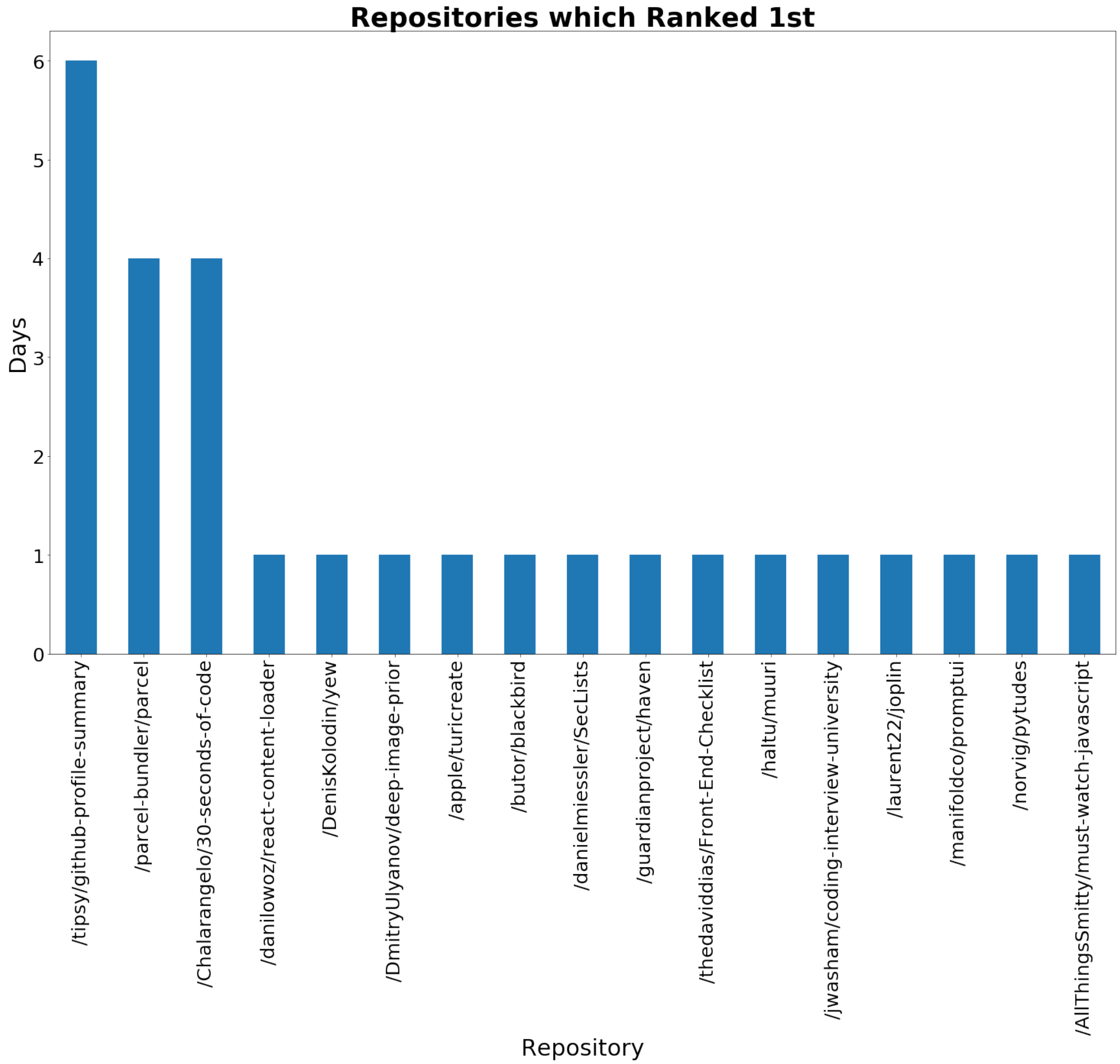




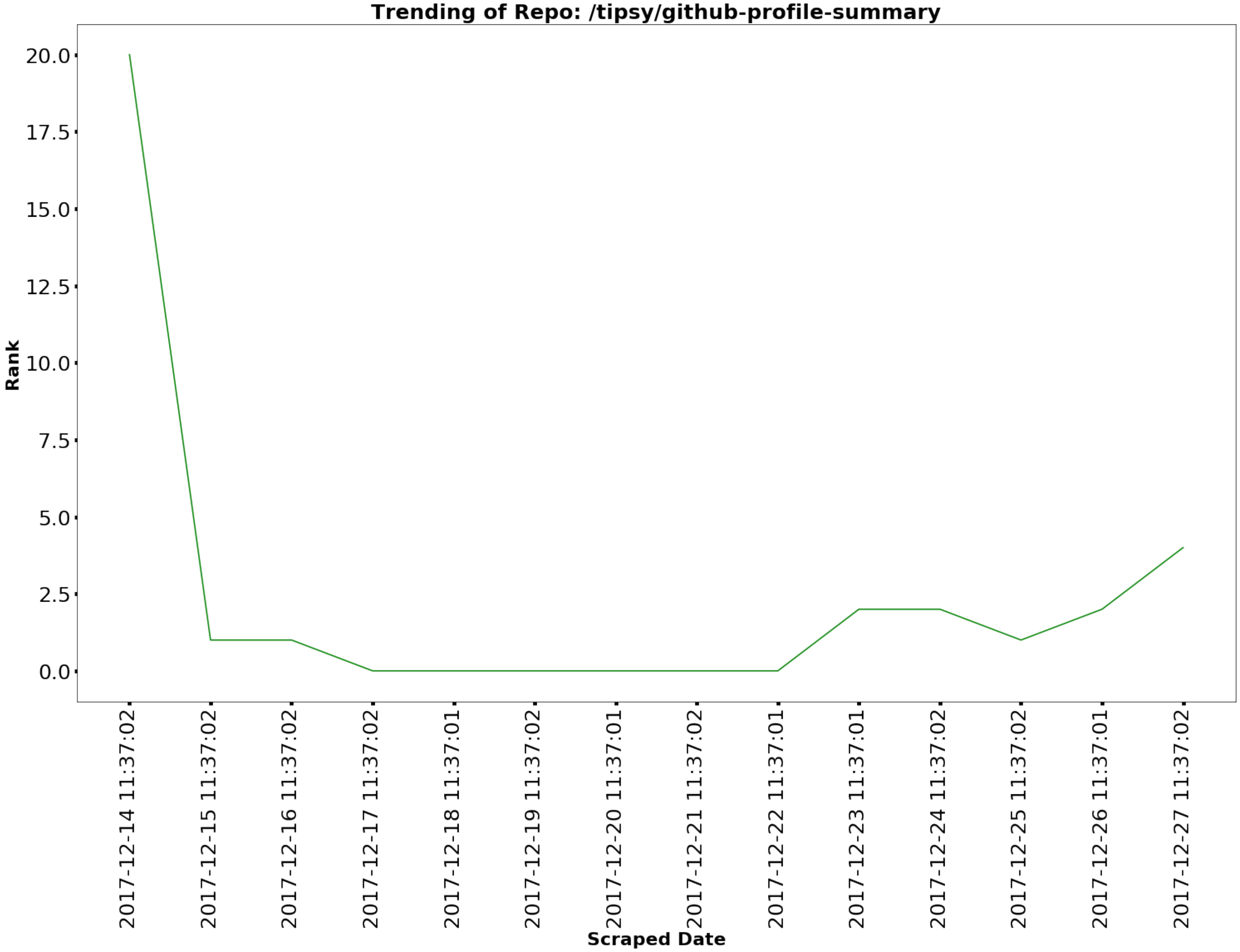




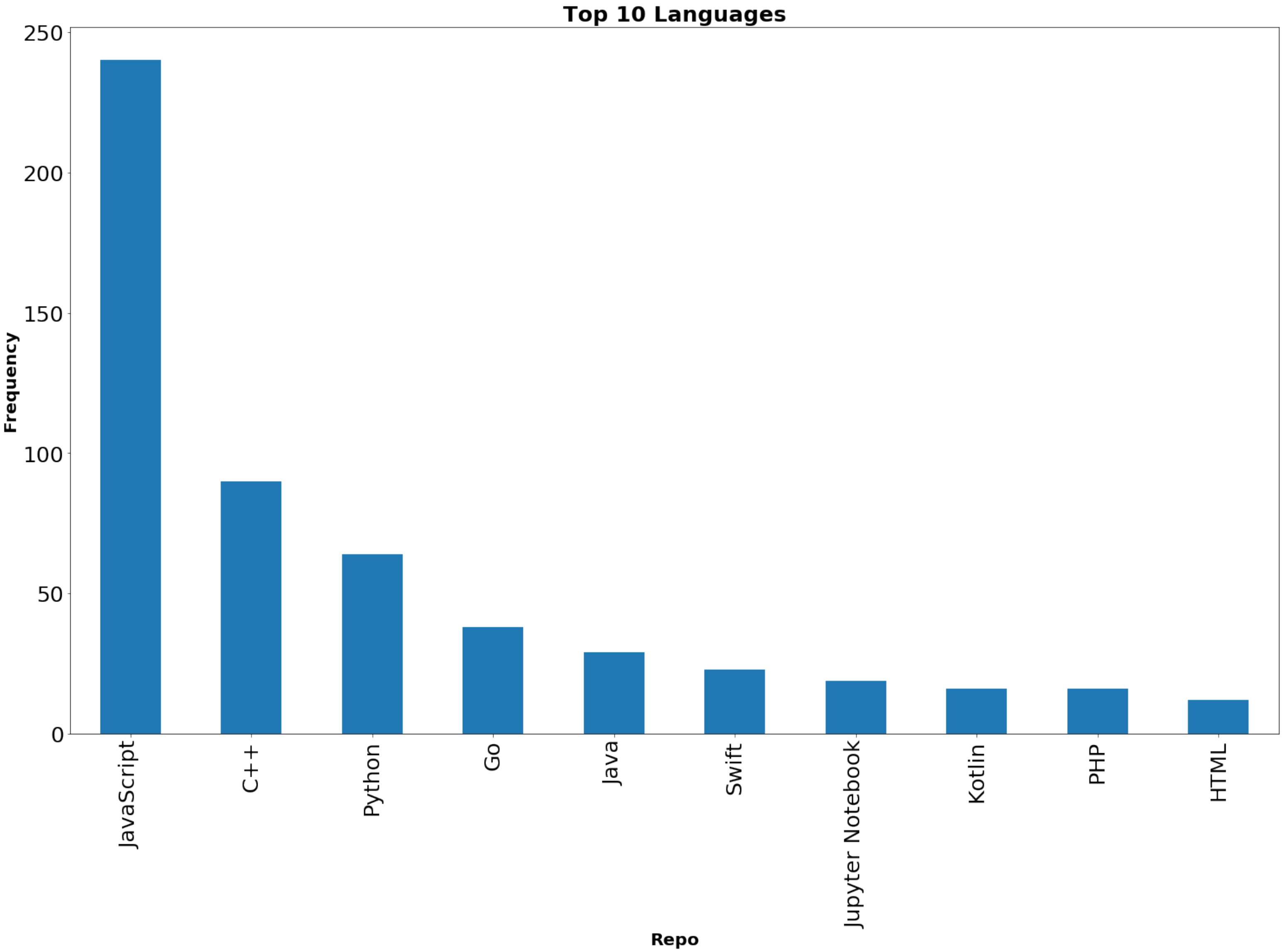
Git Hub Repos Ranked 1st



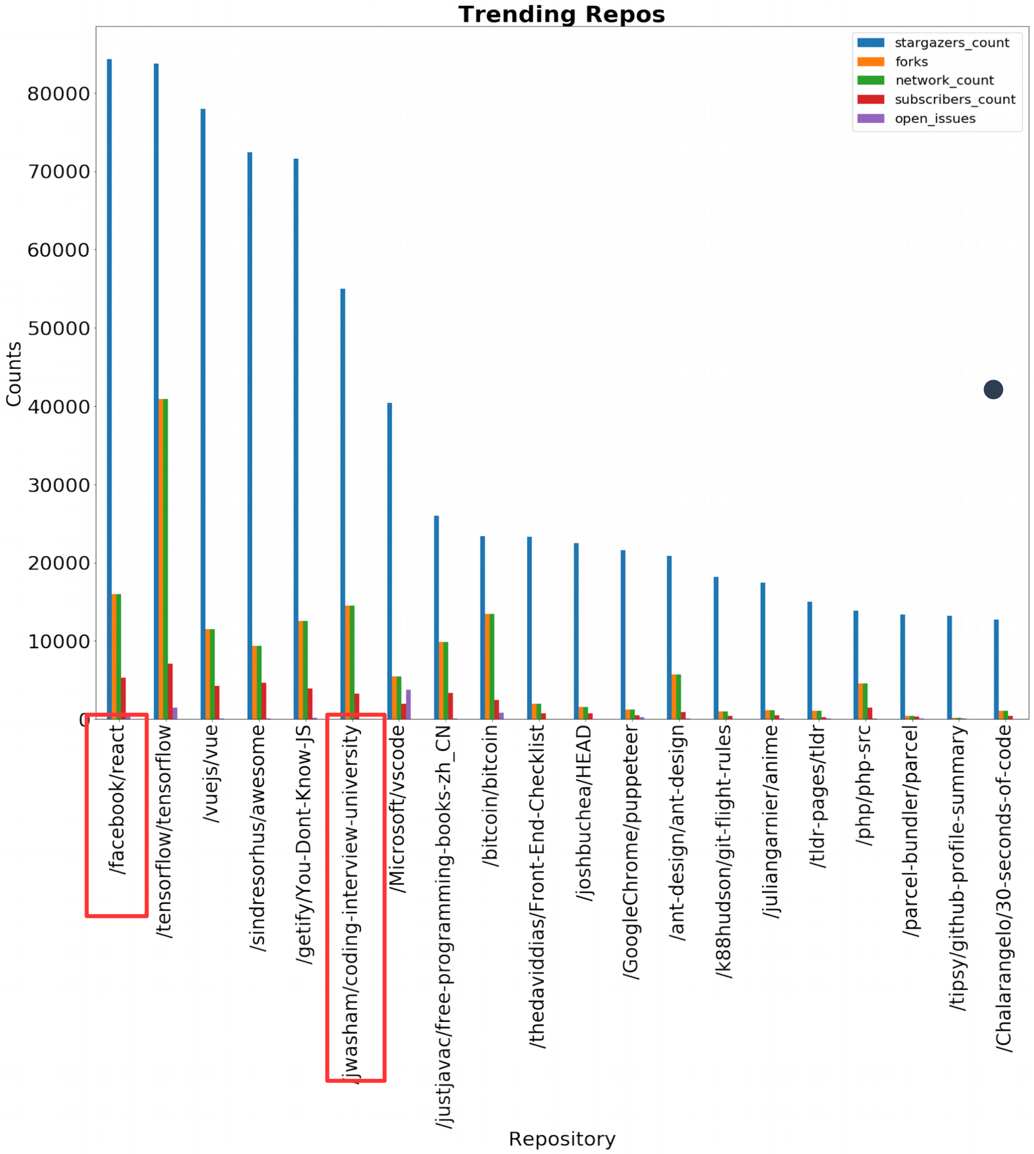
Git Hub Repo Ranking Time Trends



Git Hub Used Languages

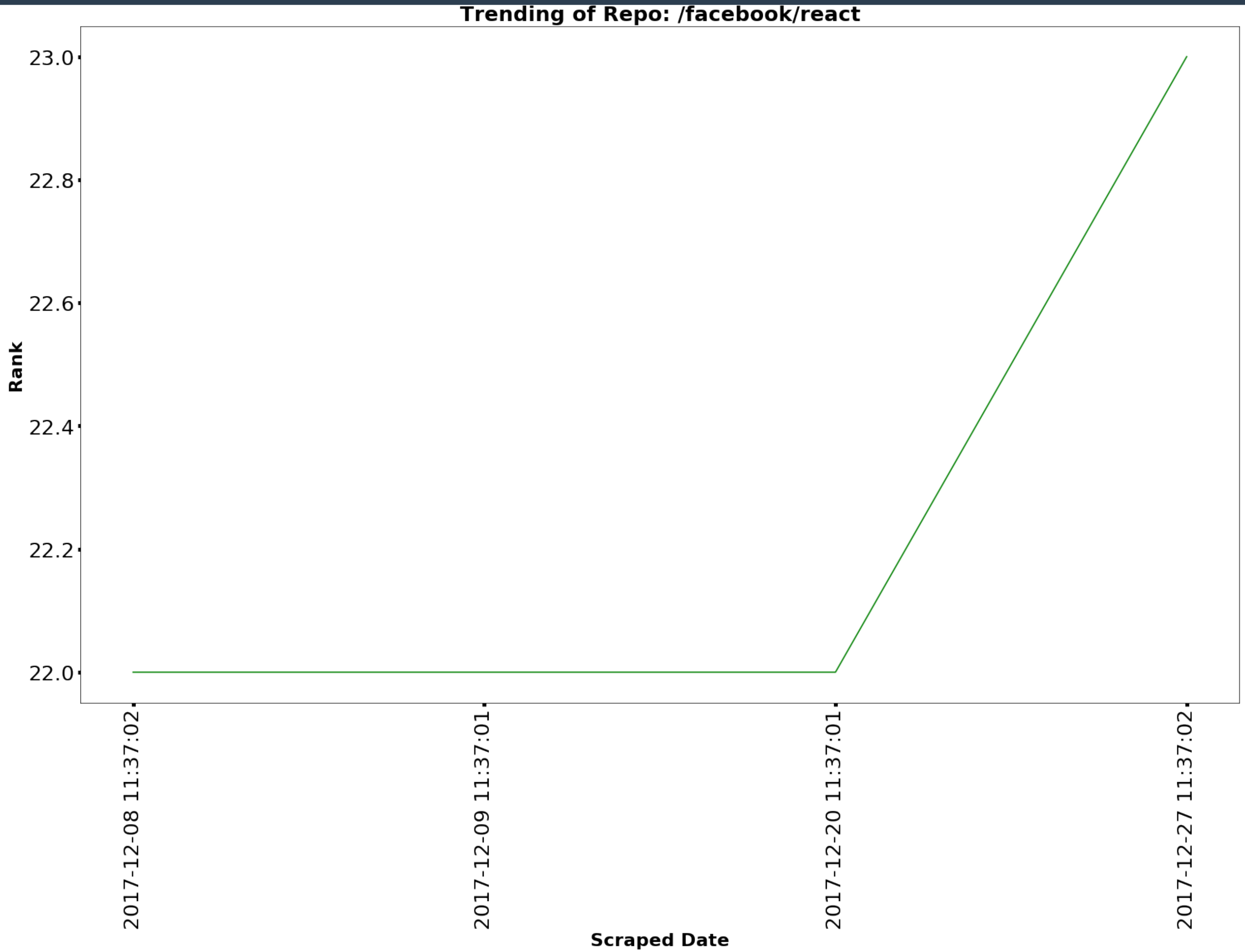


Metrics of Github Repositories

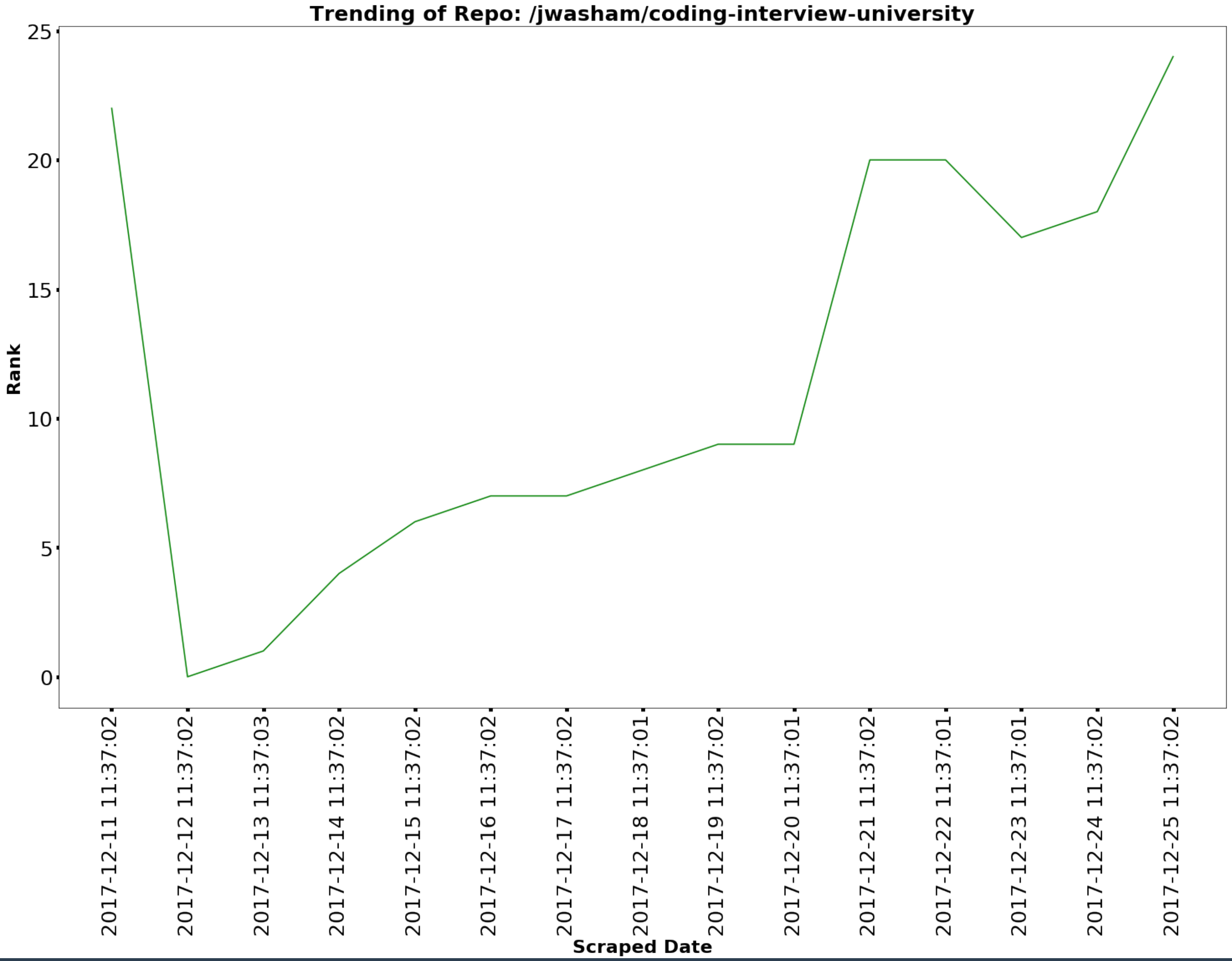


- ***Network Count**

Git Hub Repo Ranking Time Trends



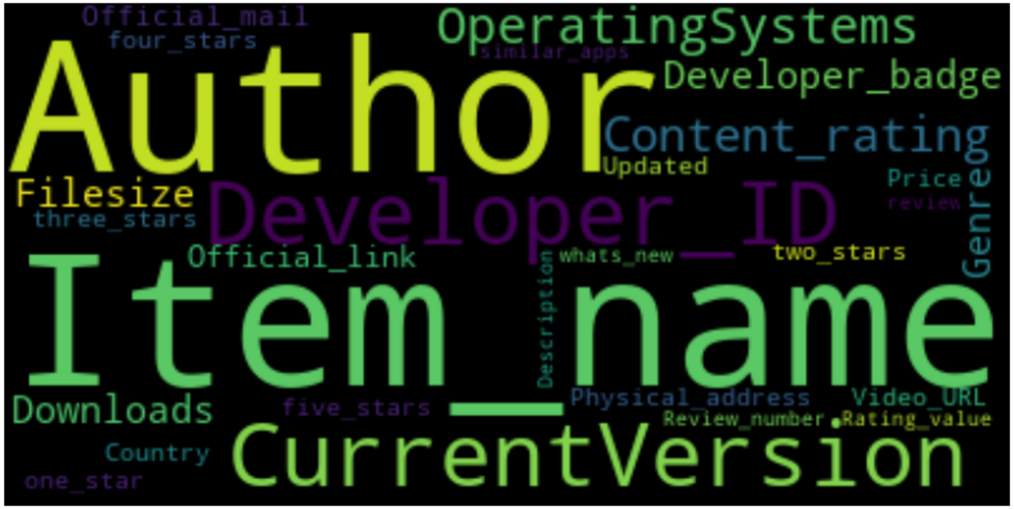
Git Hub Repo Ranking Time Trends



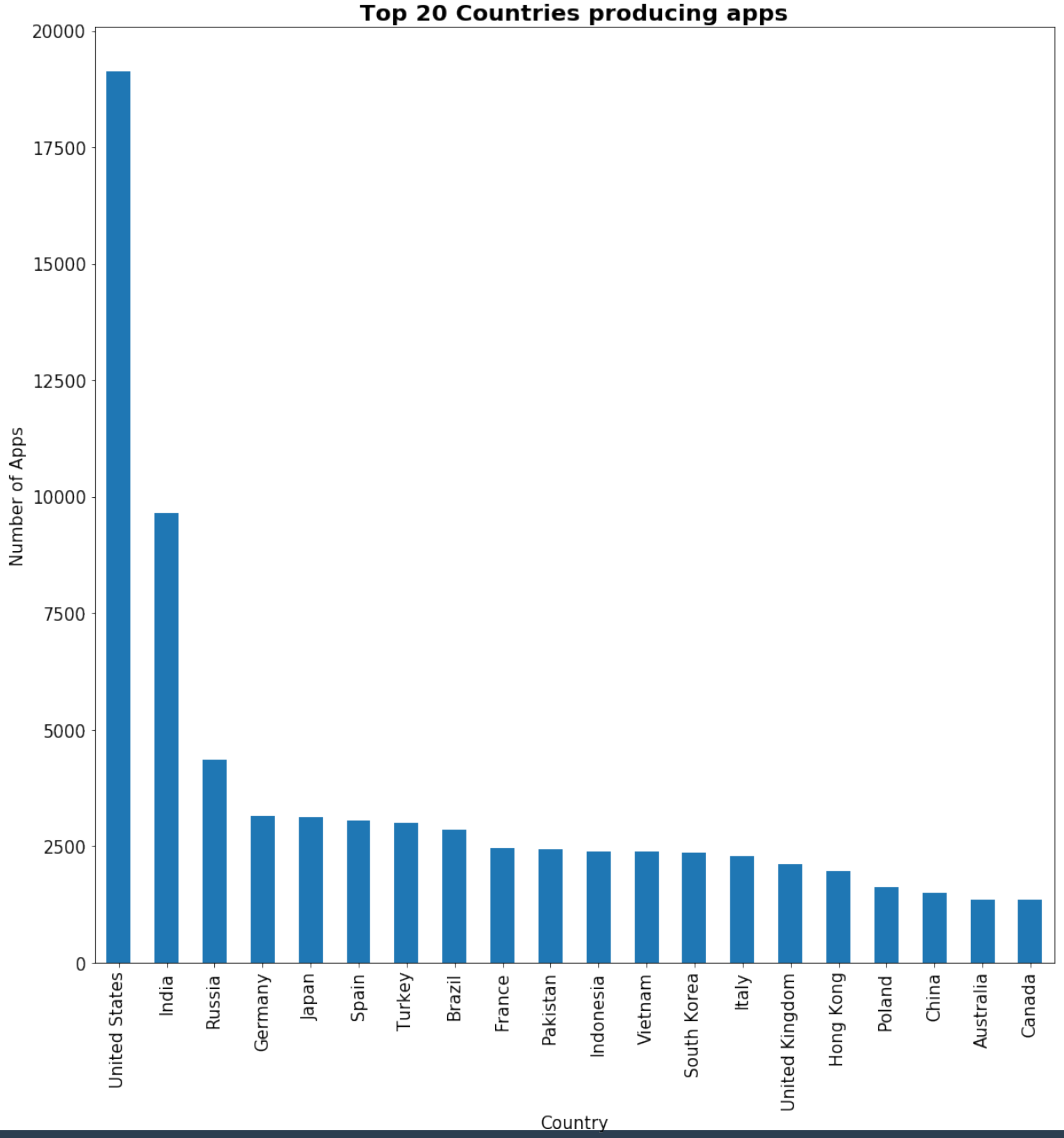
Exploratory Data Analysis of Google Apps Data

Google Apps

- **92919 unique apps**
-
- **15216 unique places**
-
- **146 countries**
-
- **17211 authors**
-
- **0.12% editors choice**
-
- **31% apps were not reviewed.**
-
- **15% apps were using video for promotion**

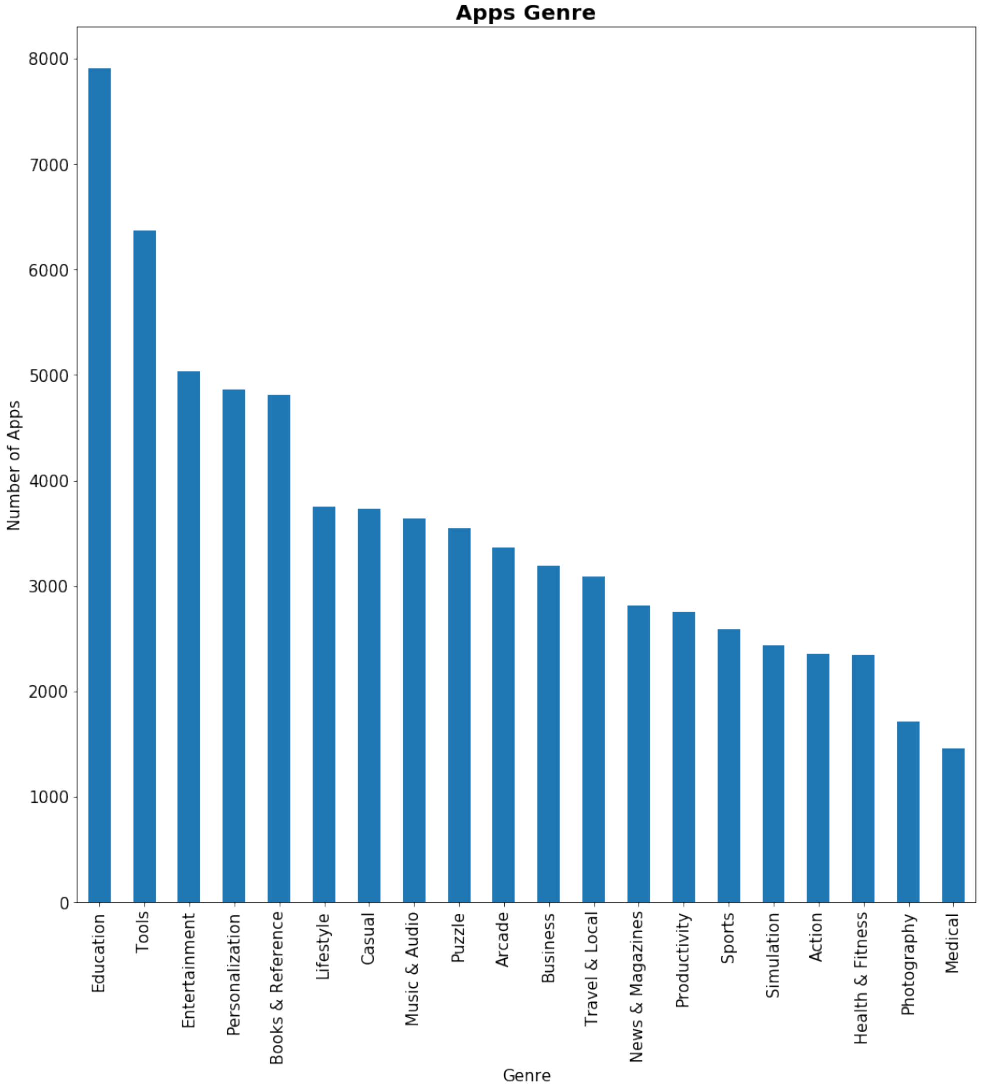


Top 20 Countries Producing Google Apps



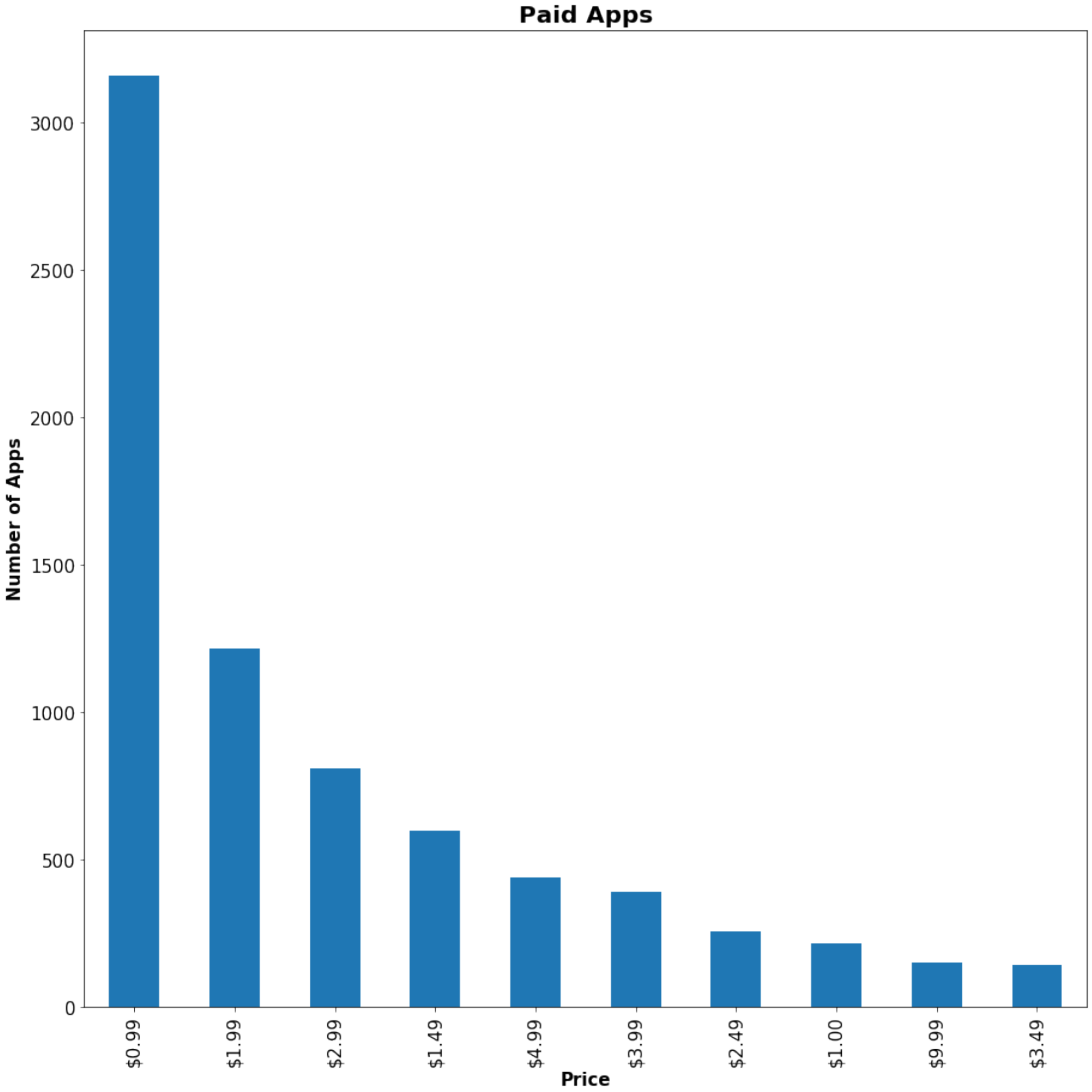
- **USA produces most apps followed by India**
-

Apps Genre



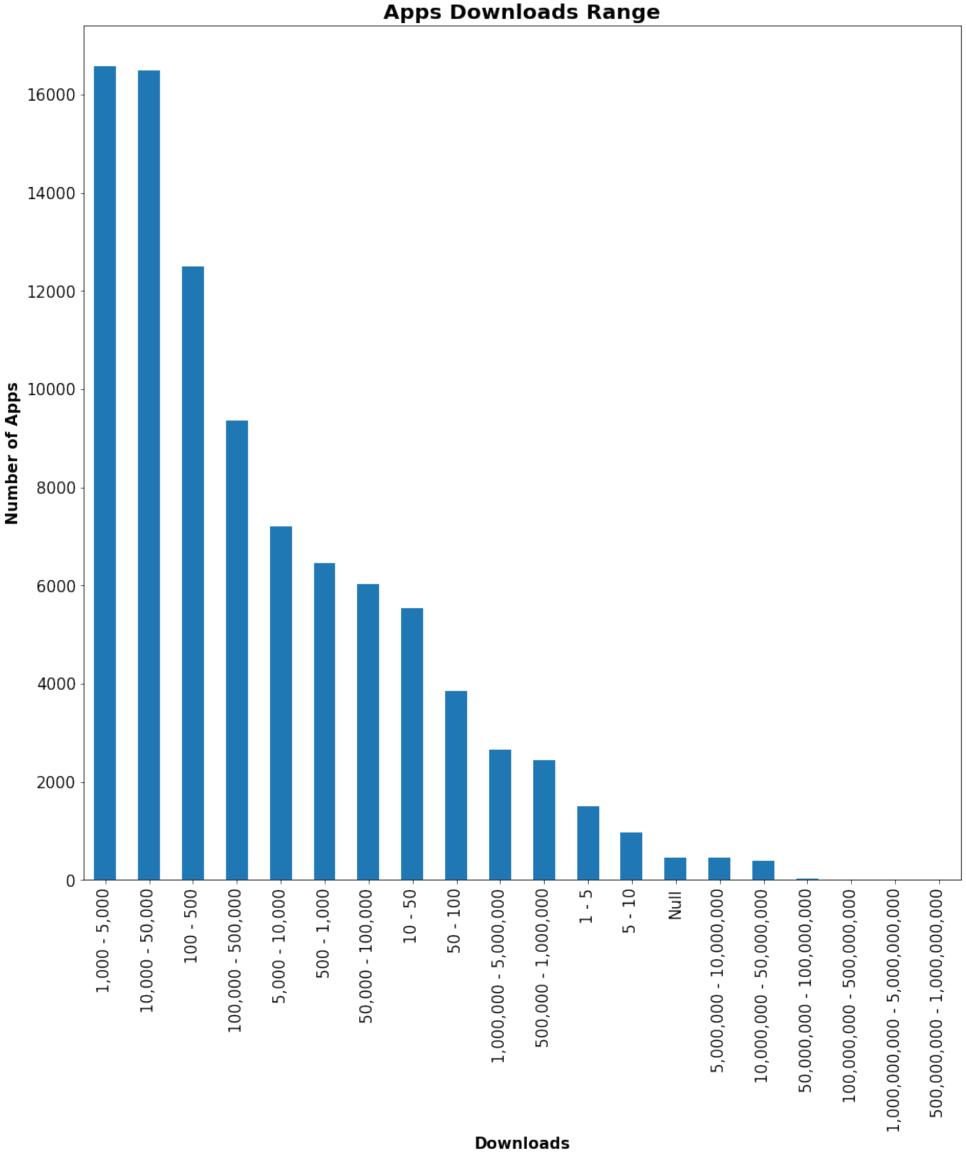
- **Most Apps Genre is Education**

Apps Pricing

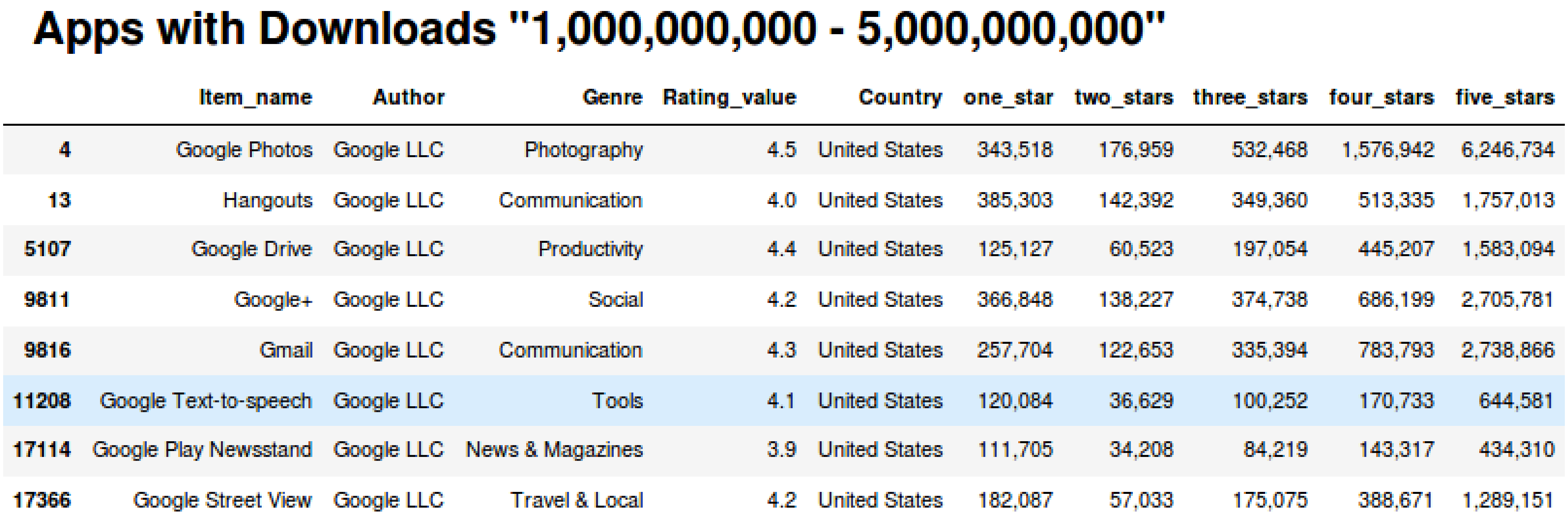


- About 89% apps are free where as 11% are paid apps.
- In paid apps price is mostly $0.99

Apps Downloads



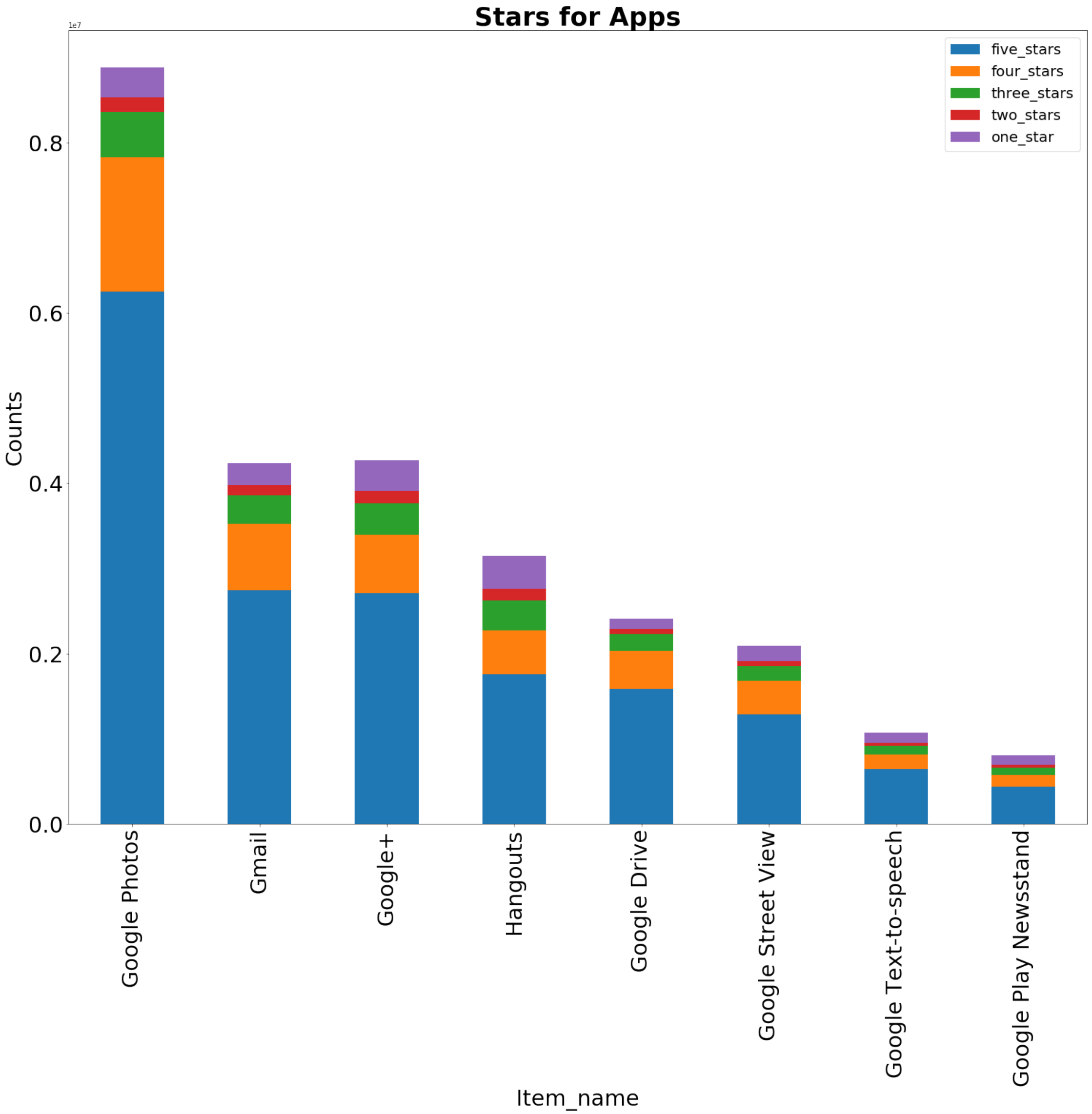
Apps Dowloaded Most



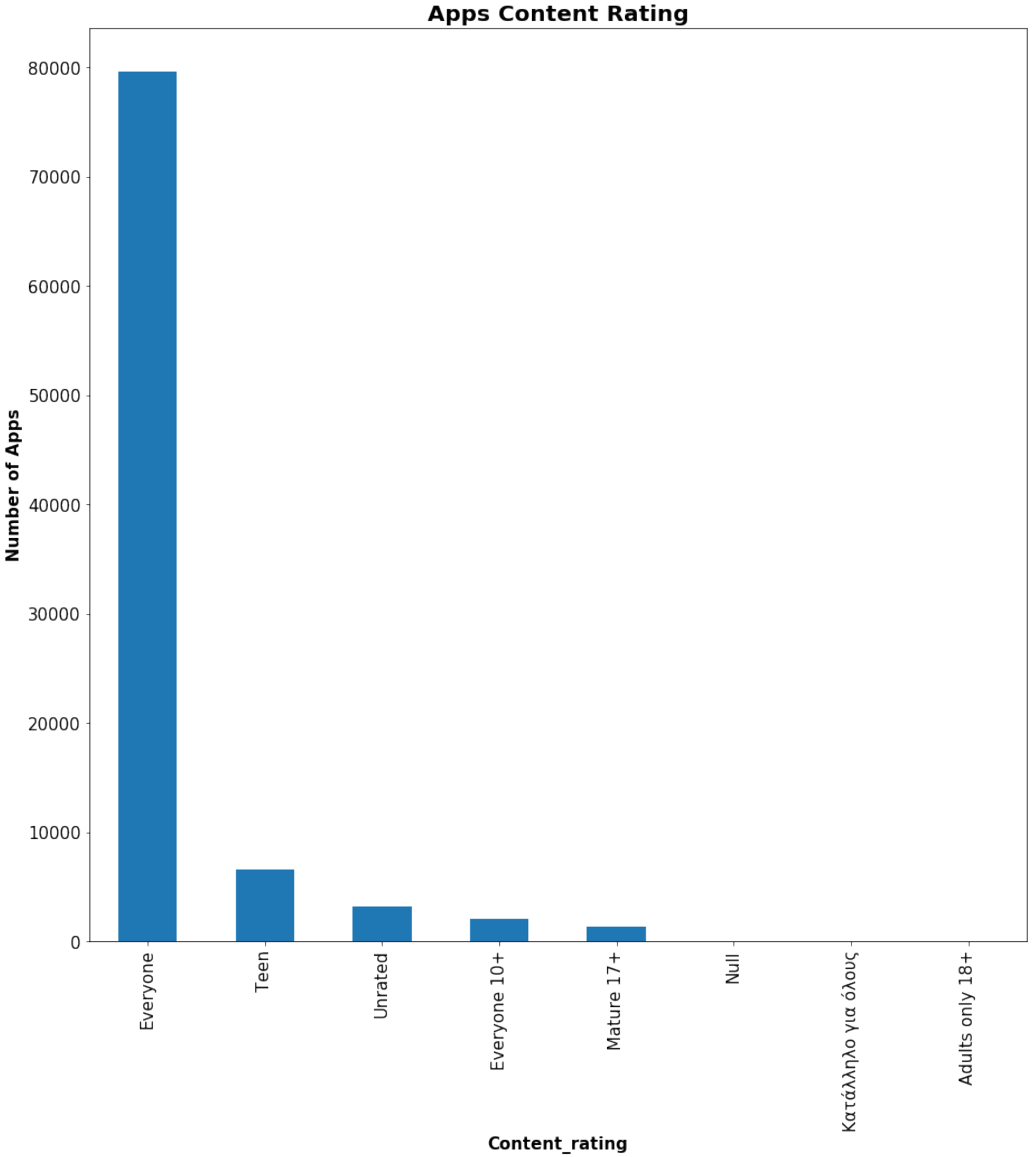
Apps with Downloads "1,000,000,000 - 5,000,000,000"

	Item_name	Author	Genre	Rating_value	Country	one_star	two_stars	three_stars	four_stars	five_stars
4	Google Photos	Google LLC	Photography	4.5	United States	343,518	176,959	532,468	1,576,942	6,246,734
13	Hangouts	Google LLC	Communication	4.0	United States	385,303	142,392	349,360	513,335	1,757,013
5107	Google Drive	Google LLC	Productivity	4.4	United States	125,127	60,523	197,054	445,207	1,583,094
9811	Google+	Google LLC	Social	4.2	United States	366,848	138,227	374,738	686,199	2,705,781
9816	Gmail	Google LLC	Communication	4.3	United States	257,704	122,653	335,394	783,793	2,738,866
11208	Google Text-to-speech	Google LLC	Tools	4.1	United States	120,084	36,629	100,252	170,733	644,581
17114	Google Play Newsstand	Google LLC	News & Magazines	3.9	United States	111,705	34,208	84,219	143,317	434,310
17366	Google Street View	Google LLC	Travel & Local	4.2	United States	182,087	57,033	175,075	388,671	1,289,151

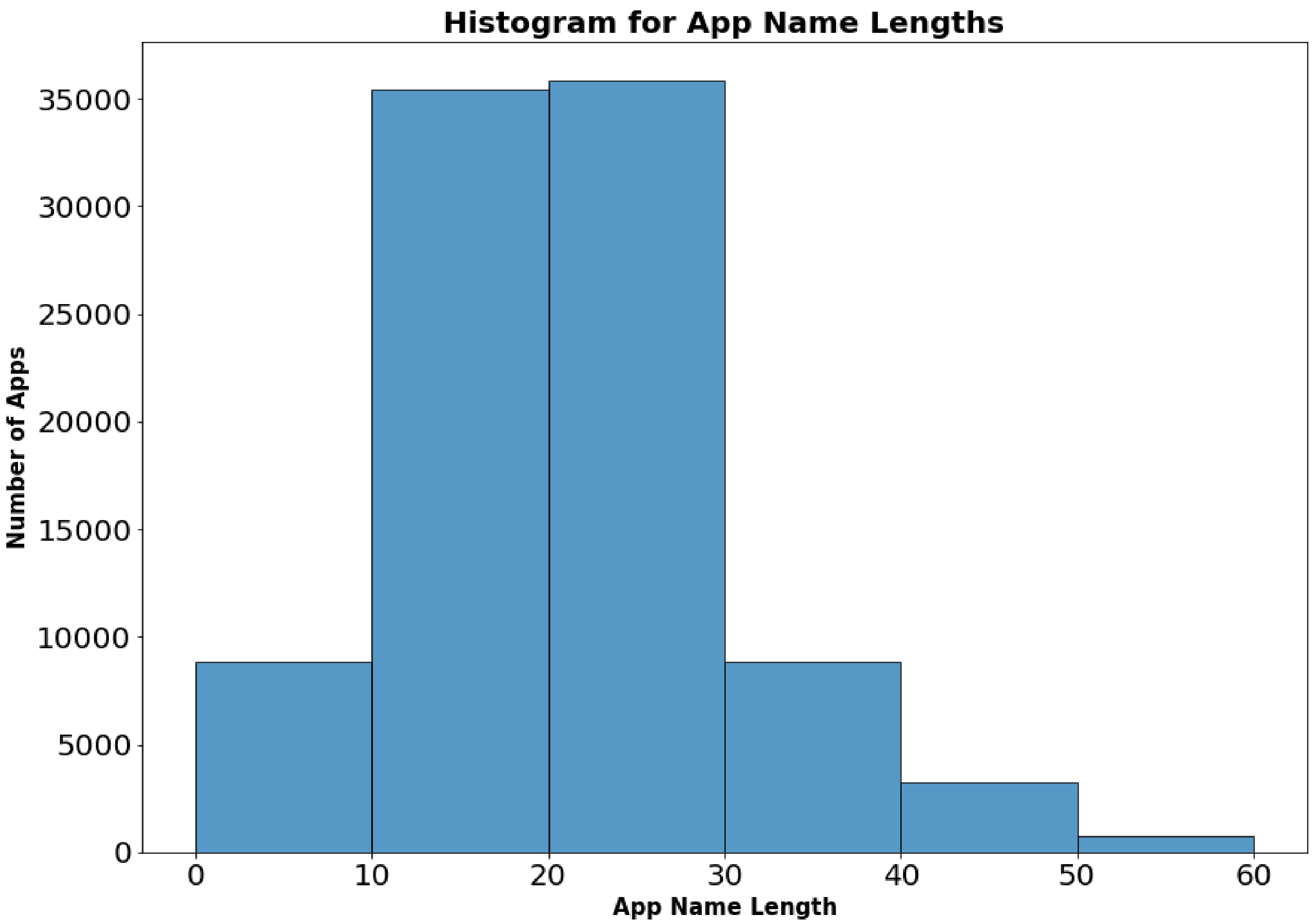
Most Downloaded Apps’ Stars



Apps Content Rating

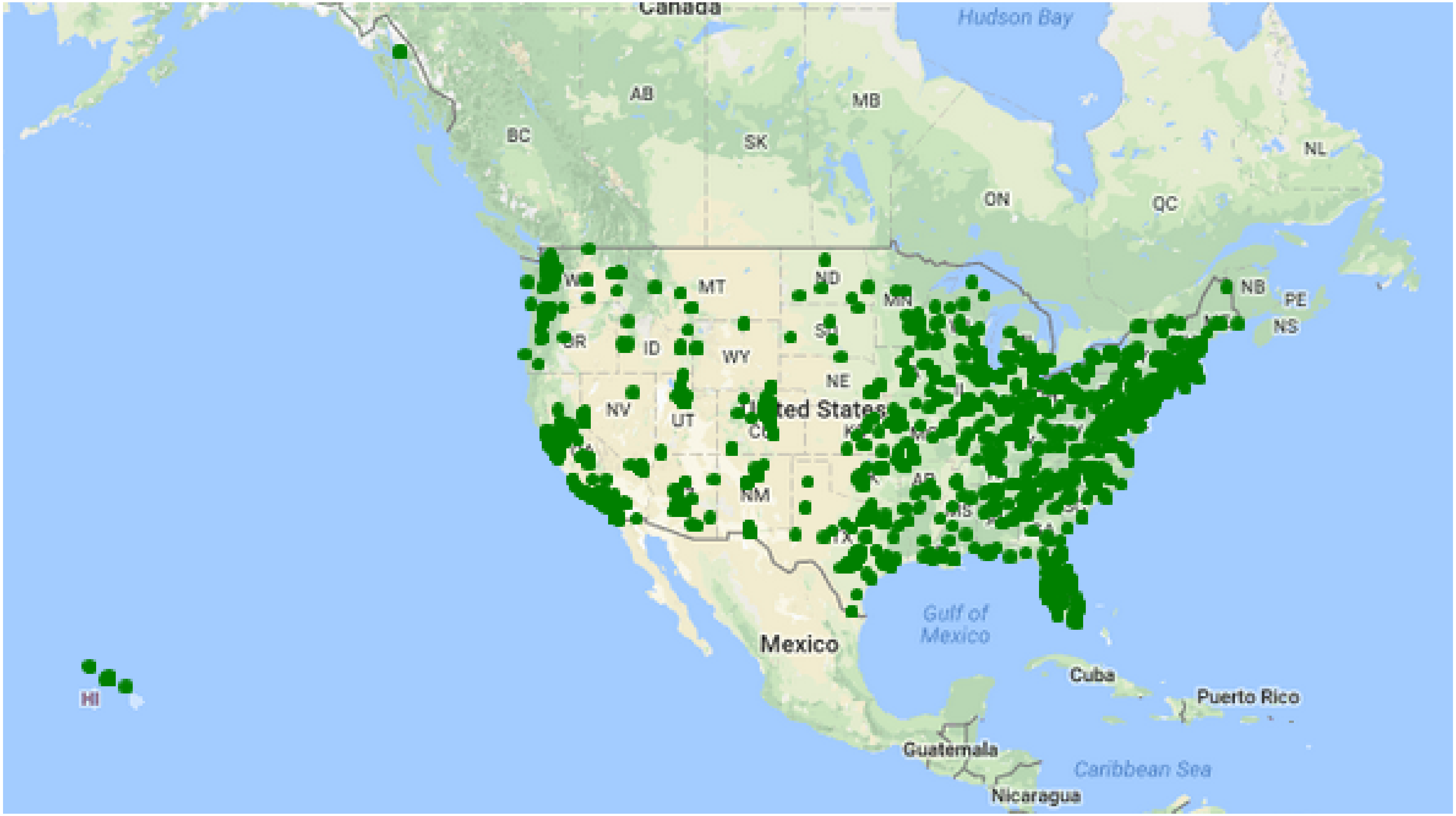


Apps Naming Scheme

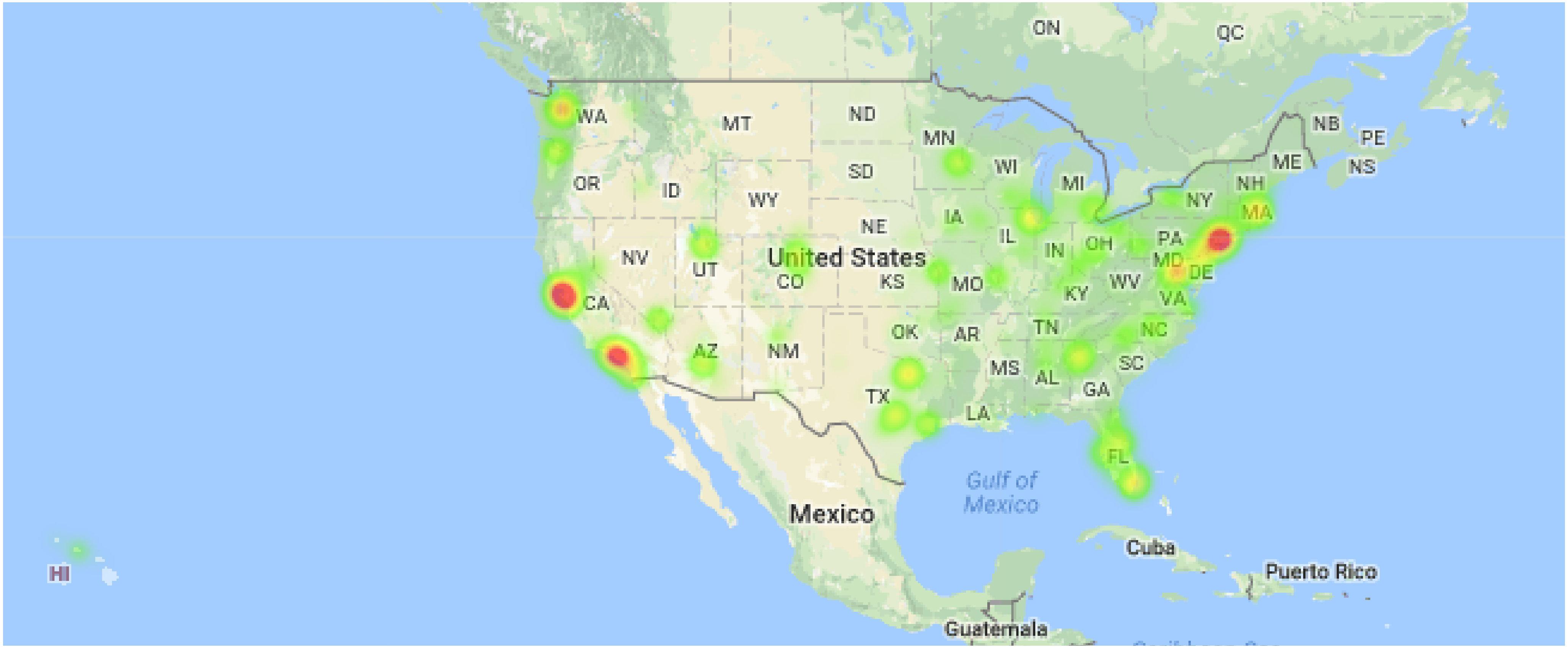


Google used to have apps name with max charecter limit of 30, later limit was increased to 50. Let see the apps naming convention

USA Location for Apps Production



USA Heat Map for Apps Production



Conclusion

- Implemented Beautifulsoup Scraper
- Implemented Scrapy Crawler
- Saved data to serial and columnar files
- Data pipelined daily to Mysql
- Daily email alerts of trending repositories
- Analyzed the scraped and crawled data

Thank You!!