Jurnal Strategi
Volume 2 Nomor 2 November 2020

ANALISIS DATASET GOOGLE PLAYSTORE MENGGUNAKAN METODE EXPLORATORY DATA ANALYSIS

Rafael Stevi Oktavian[#1], Setia Budi[*2]

[#]*Program Studi Sistem Informasi, Universitas Kristen Maranatha*
Jl. Prof. Drg. Surya Sumantri No.65, Sukawarna, Bandung
[1]rafaelstevi@gmail.com
[3]setia.budi@it.maranatha.edu

Abstract — **This journal is the result of a study entitled "Exploratory Data Analysis Datase App Google Playstore". This analysis was made for finding any pattern or indicator that can affect Rating and Installs from mobile application by using theory from John W. Turkey about EDA (Exploratory Data Analysis) and dataset Google Playstore from Kaggle that owned by Lavanya Gupta to uncover pattern or indicator about getting high Rating and Installs for mobile application.**

Keywords— **Exploratory Data Analysis, indicator, Rating and Installs, uncover pattern.**

I. PENDAHULUAN

A. *Identifikasi Masalah*

App store adalah toko aplikasi yang terdapat pada setiap *smartphone*, untuk sistem operasi berbasis IOS terdapat AppleAppStore sebagai toko aplikasi dan untuk sistem opeasi bebasis Android terdapat banyak *app store* seperti GooglePlaystore yang sudah tidak asing ditelinga kita, ada juga *app store* yang ekslusif dibuat oleh perusahaan *smartphone* yaitu Mi Store yang dimiliki oleh Xiaomi, Windows Phone Store oleh Windows, Blackberry App World oleh Blackberry. Google Playstore memiliki karakteristik yang berbeda dengan Apple App Store yaitu lebih terbuka terhadap developer aplikasi mobile sehingga memiliki varian yang lebih beragam dibanding dengan Apple App Store. Setiap aplikasi di dalam *app store* dapat dikelompokan berdasarkan karakteristik yang sama dan dapat disebut sebagai kategori dan *genre*. Kategori lebih mengarah ke karakteristik aplikasi secara umum dan *genre* lebih mengarah ke karakteristik yang spesifik. Untuk setiap aplikasi berbayar dikelompokan sebagai *paid app* dan untuk aplikasi gratis dikelompokan sebagai *free app*. Indikator lain yang terdapat dalam Google Playstore ialah Rating, Review, dan Installs.

Google Playstore memiliki Rating, Review, dan Installs. Rating merupakan nilai yang diberikan pengguna kepada aplikasi yang mereka gunakan, tinggi rendahnya Rating dapat memberikan gambaran terhadap aplikasi tersebut. Review berisikan pendapat pengguna terhadap aplikasi dapat berbentuk kritik maupun saran. Installs merupakan indikasi yang menunjukan keseluruhan dari total aplikasi telah diunduh oleh pengguna. Mendapatkan Rating yang tinggi dan Installs yang banyak merupakan keinginan dari setiap *developer* aplikasi *mobile*, karena dengan begitu mereka tahu bahwa aplikasi yang mereka kembangkan populer dan disukai oleh mayoritas penggunanya, namun masih banyak *developer* aplikasi *mobile* yang tidak tahu indikator apa saja yang mempengaruhi Rating dan Installs dari aplikasi *mobile*.

Ketidaktahuan *developer* aplikasi *mobile* terhadap indikator yang dapat mempengaruhi tingginya Rating dan banyaknya Installs dari aplikasi *mobile* dapat menjadi kendala yang dapat menghambat pengembangan suatu aplikasi. *Developer* menjadi kesulitan untuk menentukan target estimasi yang harus dicapai ketika mengembangkan aplikasi *mobile*. Hal tersebut dapat dihindari bilamana *developer* mengetahui indikator yang dapat mempengaruhi Rating dan Installs, dengan mengetahui itu *developer* mendapatkan gambaran tentang aplikasi *mobile* yang berkemungkinan besar mendapatkan Rating tinggi dan total Installs yang banyak, serta dapat diterima oleh pengguna *smartphone* khususnya pada *app store* Google Playstore. Dengan demikian saya menjadi tertarik untuk mencari pola-pola dan indikator yang dapat mempengaruhi Rating dan Install aplikasi guna membantu *developer* aplikasi *mobile* untuk menentukan target dalam membuat serta mengembangan aplikasi mereka dengan melakukan penelitian berupa eksplorasi *dataset* dari Google Playstore yang berjudul "Exploratory Data Analysis Dataset App Google Playstore".

B. *Tujuan Penelitian*

Tujuan dari penyusunan laporan ini untuk adalah untuk mendapatkan pola-pola dan indikator yang dapat mempengaruhi Rating dan total Installs dari suatu aplikasi dalam Google Playstore serta, data ini juga dapat digunakan oleh *developer* sebagai acuan dalam mengembangakan aplikasinya ataupun, sebagai pengetahuan yang dapat berguna bagi orang lain.

II. KAJIAN TEORI

A. *Exploratory Data Analysis*

1) *Data Analysis*

Data sebenarnya bukanlah informasi, setidaknya jika dilihat dari bentuknya dan data akan sulit dipahami jika tidak terikat dengan angka, kata, atau waktu dilaporkanya data tersebut. Informasi dihasilkan melalui proses pengambilan data mentah dari dataset tertentu, lalu mengekstraknya untuk mendapatkan konklusi yang dapat digunakan dalam berbagai hal, proses ini disebut data analysis.. Tujuan dari Data Analysis adalah untuk mengekstrak informasi yang tidak mudah dijawab atau diputuskan, namun ketika dipahami akan menuntun kepada kemungkinan untuk mempelajari mekanisme sistem yang menghasilkan informasi itu, sehingga dapat memprediksi kemungkinan respons dari sistem dan evolusi kedepannya.

2) *Exploratory Data Analysis :*

Menurut John W. Turkey Exploratory Data Analysis (EDA) didefinisikan sebagai *"detective work – numerical detective work – or counting detective work – or graphical detective work",* dengan menjadi seorang detektif atau melakukan pekerjaan detektif untuk menemukan apa yang dapat diberikan oleh data tanpa didasari prasangka dan bergantung kepada fakta. John W. Turkey menemukan *Confirmatory Data Analysis* (CDA) yang merupakan kontras dari EDA, terfokus pada lingkup analisis data yang berkaitan dengan pengujian hipotesis statistik, *confiendce intevals*, dan estimasi. EDA dan CDA seharusnya digunakan bersamaan secara komplementer untuk dapat menemukan pola dan struktur yang menuju pada hipotesis dan model.

B. *Google Playstore*

App store merupakan "platform penjualan online untuk developer menjual dan menyalurkan produk kepada aktor melalui satu atau lebih platform software ecosystem". App Store Memungkinkan developer untuk menghasilkan keuntungan dari software dan membawa fungsionalitas baru kepada konsumen. App Store ada yang memiliki satu ekosistem seperti Google Play yang hanya melayani platform Android, atau lebih dari satu ekosistem seperti BlinPress app store, dimana kode dapat dibeli untuk lebih dari satu platform dan ekosistem. Google Play ialah app store yang diluncurkan di tahun 2008 dan menjadi app store terbesar di ekosistem Android. Google Play melayani platform Android sebagai open-source sistem operasi untuk perangkat mobile dan komputer tablet.

C. *Python, Dataframe, Pandas*

1. *Python* : Python adalah bahasa pemrograman *interpreted*, yaitu *psuedo-compiled* artinya setelah kode ditulis dibutuhkan *interpreter* untuk menjalankan kode itu. *Interpreter* adalah program yang terinstal dalam setiap perangkat yang memiliki tugas untuk menafsirkan *source code* dan menjalankanya. Tidak seperti bahasa pemrograman C, C++ dan java, tidak terdapat compile time pada python.
2. *Dataframe* : Dataframe adalah data yang berada didalam tabel dengan kumpulan kolom yang tersusun. Struktur dari *dataframe* didesain untuk mengembangkan *series* kedalam berbagai dimensi. Faktanya *dataframe* terdiri dari koleksi kolom yang tersusun dan memiliki nilai yang berbeda dari tiap tipe.
3. *Pandas* : Pandas adalah pustaka *open source* untuk Python untuk spesialisasi analisis data dan timbul dari kebutuhan untuk memiliki pustaka yang terspesifikasi dalam menganalisa data dengan menyediakan cara tersimpel untuk alat *data processing*, *data extraction* dan *data manipulation.*

III. METODOLOGI PENELITIAN

A. *Dataset*

Dataset yang digunakan dalam proyek penelitian ini berjudul "Google Play Store Apps", *dataset* ini diperoleh dari Kaggle. *Dataset* ini bersifat open dan dapat diakses secara publik serta dimanfaatkan untuk keperluan penelitian dan

pengembangan pengetahuan. Terdapat dua data yang tersimpan dalam format CSV (*Comma Separated Values*), yaitu *dataset* googleplaystore.csv dan *dataset* googleplaystore_user_review.csv

1) Karakteristik Dataset

Dataset tersimpan dalam dua format CSV (*Comma Seperated Values*) yaitu googleplaystore.csv dan googleplaystore_user_review. *Dataset* googleplaystore berisikan data dari 10.000 aplikasi yang terdapat dalam Google Playstore dan *dataset* googleplaystore_user_review berisikan data tentang detail review dari tiap aplikasi dalam Google Playstore. *Dataset* googleplaystore terdiri dari 13 kolom.

B. Platform Penelitian

Penelitian EDA ini menggunakan Jupyter Notebook sebagai *platform* penelitian dengan bantuan *library* Pandas dalam mengolah *dataset*. Jupyter Notebook memiliki *library* lainnya seperti Numpy yang mendukung operasi vektor dan matriks ketika mengolah *dataset*, dan Seaborn yang mendukung visualisasi data. Jupyter Notebook digunakan karena cocok untuk menjadi *platform* analisis data dengan bahasa pemrograman Python seperti pada penelitian ini. Jupyter Notebook memiliki kelebihan yaitu dapat menjalankan kode per baris sehingga jika terjadi suatu kesalahan hanya perlu memperbaiki satu baris kode yang dianggap bermasalah.

C. Data Transformation

Dataset terdiri dari dua file CSV yaitu googleplaystore.csv dan googleplaystore_user_review.csv yang akan ditransformasi menjadi dataframe dengan menggunakan library Pandas.

1. *Tujuan Transformasi* Dataset : Data Transformation bertujuan untuk mengubah *dataset* menjadi *dataframe* dengan menggunakan library Pandas, dengan begitu *dataset* dapat dibaca oleh Jupyter Notebook sebagai *platform* penelitian.
2. *Proses Transformasi Dataset* : Transformasi *dataset* menjadi *dataframe* diperlukan agar data dapat dibaca oleh *platform* Jupyter Notebook. *Dataset* dengan format CSV dapat diubah menjadi *dataframe* dengan melalui step seperti berikut :
   - Import *library* Pandas
   - Import data dengan fungsi pd.read.csv("direktori dari *dataset*").
3. *Identifikasi Atribut Dataset* : *Dataset* yang telah ditransformasi menjadi *dataframe* sudah dapat dieksplorasi dalam Jupyter Notebook dan hal pertama yang harus dilakukan adalah mengidentifikasi atribut-atribut yang dimiliki oleh *dataframe* tersebut. Atribut yang dimiliki oleh *dataframe* dapat berupa tipe data dari tiap kolom, jumlah keseluruhan baris dan kolom, serta nilai unik yang dimiliki oleh masing-masing kolom dalam *dataframe.* Untuk melihat tipe data dari keseluruhan kolom dalam *dataframe* menggunkakan kode "nama *dataframe*".dtypes dengan demikian keseluruhan tipe data setiap kolom akan dijabarkan dapat berupa objek, float, int, dan sebagainya. Untuk melihat jumlah keseluruhan baris dan kolom yang dimiliki *dataframe* digunakan kode "nama *dataframe*".shape yang akan menghasilkan jumlah baris dan kolom dari *dataframe* tersebut. Atribut lainnya yang tidak kalah penting adalah nilai unik dari setiap kolom, dengan adanya nilai unik eksplorasi akan menjadi lebih mudah karena dapat menunjukan variasi nilai non-duplikasi yang terdapat dalam setiap kolom. Untuk mengetahui nilai unik digunakan kode "nama *dataframe*".['nama kolom'].unique() dengan demikian seluruh nilai unik dari kolom tersebut akan dijabarkan dalam bentuk array.

D. Data Cleaning

Data Cleaning dilakukan untuk kedua *dataset* yang akan digunakan dalam penelitian ini yaitu googleplaystore dan googleplaystore_user_review. Tujuan dilakukannya data cleaning adalah untuk mengidentifikasi *missing value* yang ada dalam *dataset* dan melakukan tindakan penanggulangan untuk data tersebut.

1. *Penanganan Missing Value* : *Missing value* atau dapat disebut sebagai *missing data* merupakan data yang hilang atau bernilai Nan dalam suatu *dataset*. Dalam *dataset* googleplaystore dan googleplaystore_user_review terdapat beberapa data yang berkarakteristik seperti *missing value* sehingga harus dilakukan identifikasi lebih lanjut. Untuk *dataset* googleplaystore terdapat kejanggalan data pada nilai unik dari kolom Category yaitu terdapat nilai '1.9' dan setelah ditelusuri ditemukan bahwa data mengalami pergeseran dan menghilangkan kategori dari aplikasi yang bernama 'Life Made Wi-Fi Touchscreen Photo Frame'. Dan ditemukan aplikasi yang bersangkutan seharusnya memiliki kategori 'Lifestyle' dan untuk menanggulanginya maka data akan dishift satu kolom ke arah kanan, serta kolom kategori diisi dengan 'Lifestyle'. Untuk memastikan bahwa tidak ada duplikasi data dan nilai Null/Nan dilakukan metode drop duplicate dan metode mensubstitusi nilai Nan dengan *mean* setiap kolom serta drop data bilamana masih tersisa data yang bernilai Null/Nan. Pada *dataset* googleplaystore_user_review terdapat *missing value* dengan nilai Nan pada baris yang sama di kolom Translated Review, Sentiment dan Sentiment_Polarity, Sentiment_Subjectivity, dengan ini hanya kolom App saja yang diisikan oleh nama Aplikasi. Penemuan *missing value* ini dapat mempengaruhi keseluruhan data yang ada dalam *dataset*. Penanganan yang dilakukan diantaranya ialah dengan mengidentifikasi terlebih dahulu setiap data yang mengandung *missing value*, lalu menghilangkan data yang berupa missing value atau Nan didalamnya dengan melakukan *drop* data.

2. *Penyesuaian Tipe Data di Setiap Kolom* : Dalam *dataset* googleplaystore terdapat beberapa kolom yang harus disesuaikan terlebih dahulu tipe datanya untuk mempermudah proses eksplorasi, kolom-kolom yang harus disesuaikan diantaranya ialah Size, Installs, Price, Last Updated, Rating dan Reviews. Untuk kolom Size, Installs, Price akan diubah tipe datanya menjadi numerik dengan kode pd.to_numeric("nama *dataframe*"['nama kolom']) dan hasil konversinya akan dijadikan kolom baru yaitu SizeC, InstallsC dan PriceC. Dengan tujuan agar nantinya data tersebut dapat divisualisasikan dengan menggunakan matplotlib dan seaborn. Untuk Kolom Last Update dikonversi menjadi tipe datetime object dengan kode pd.to_datetime("nama *dataframe*"['nama kolom']). Untuk kolom Rating dan Reviews akan diubah tipe datanya menjadi numerik dengan pd.to_numeric("nama *dataframe*"['nama kolom']).

E. Exploratory Data Analysis Google Playstore

Setelah melakukan Data Cleaning maka *dataframe* sudah siap ditelusuri untuk menemukan hal-hal menarik apa saja yang ada dalam kedua *dataframe* dan indikator yang dapat mempengaruhi Rating dan total Installs dari suatu aplikasi *mobile*. Dan setelah dilihat secara keseluruhan *dataframe* googleplaystore timbul beberapa pertanyaan menarik mengenai dataframe tersebut, yaitu :

1. *Category Aplikasi Terpopiler dalam Google* Playstore : Dalam *dataframe* googleplaystore terdapat 33 Category aplikasi dan diantaranya terdapat Category yang memiliki jumlah aplikasi terbanyak dibandingkan Category lainya. Dengan mencari Category aplikasi terpopuler diantaranya dapat ditemukan Category mana saja yang paling banyak diambil oleh *developer* aplikasi *mobile*. Dengan cara menghitung jumlah aplikasi dari setiap Category dapat dilihat gambaran jelas mengenai distribusi aplikasi dari setiap Category. Untuk menemukan Category terpopuler maka harus mencari Category yang memiliki Rating tinggi dengan jumlah aplikasi yang banyak dan akan diurutkan berdasarkan jumlah aplikasi dari yang tertinggi hingga ke yang terendah dan nantinya akan dikelompokan menjadi top 5 Category aplikasi yang memiliki jumlah aplikasi tertinggi. Hasil dari pengelompokan dapat divisualisasikan dalam bentuk pie chart untuk mempermudah pembacaan data..

2. *Distribusi Rating Disetiap Category* : Tinggi rendahnya Rating aplikasi pada Google Playstore ditentukan oleh pengalaman dari setiap pengguna ketika memakai aplikasi tersebut. Rating juga dapat menjadi salah satu faktor ketika pengguna baru ingin mengunduh aplikasi pada Google Playstore, hal tersebut menimbulkan rasa ingin tahu mengenai distribusi Rating di setiap Category Google Playstore. Dengan menggunakan Jupyter Notebook disertai library pandas dan seaborn dapat ditelusuri bagaimana distribusi Rating di setiap Category yang ada, dengan menggunakan bentuk visualisasi histogram dapat dilihat secara keseluruhan Rating aplikasi yang dalam Google Playstore. Untuk distribusi Rating setiap Category dapat dicoba dengan menggunakan box plot .untuk melihat apakah terdapat perbedaan signifikan terhadap Rating di setiap Category.

3. *Tipe Aplikasi yang Mendominasi Google Playstore* : Aplikasi dalam Google Playstore memiliki dua tipe yang berbeda yaitu aplikasi gratis yang disebut *free app* dan aplikasi berbayar yang disebut *paid app*. Dengan menggunakan Jupyter Notebook diharapkan dapat menemukan tipe aplikasi apa yang paling mendominasi. Langkah yang akan dilakukan ialah mengelompokkan terlebih dahulu tipe aplikasi kedalam satu tampungan dan menghitung

total keseluruhan aplikasi yang ada di dalamnya, setelah itu dibuatlah gambaran visualisasi dari data tersebut dapat berupa bar chart. Untuk mengetahui distribusi tipe aplikasi dari setiap Category dapat dilihat dengan cara yang sama hanya Category dan tipe harus dikelompokkan terlebih dahulu, dengan demikian dapat dilihat tipe aplikasi yang paling mendominasi secara keseluruhan .maupun per Category.

4. *Hubungan antara Rating dan Review* : Rating dapat menggambarkan pendapat pengguna terhadap suatu aplikasi *mobile*, namun apakan Rating dapat mempengaruhi jumlah total dari Reviews?. Untuk menjawabnya maka dilakukan penelusuran *dataframe* googleplaystore menggunakan Jupyter Notebook. Rating dan Reviews bertipe data numerik setelah dilakukan penyesuaian data dalam proses data cleaning, sehingga kedua data dapat langsung dibandingkan utnuk memastikan apakah keduanya saling berkorelasi. Setelah dilakukan penelusuran dengan mengunakan metode Pearson Correlation Coeficient ditemukan kolom Reviews dan Rating memiliki angka 0.51 dan jika angka dari Pearson Correlation Coeficient melebihi 0 dapat disimpulkan bahwa kedua data memiliki korelasi yang positif. Dan untuk visualisasi hubungan Rating dengan Reviews.
5. *Hubungan antara Rating dan Installs* : Installs merupakan total jumlah aplikasi diunduh oleh pengguna Google Playstore dengan bentuk seperti *milestone* yang terus bertambah. Developer aplikasi *mobile* pasti menginginkan aplikasi yang mereka buat mendapatkan angka Installs yang banyak disertai dengan Rating yang tinggi, sehingga muncul pertanyaan apakah Rating mempengaruhi jumlah total dari Installs ?. Untuk menjawabnya maka dilakukan penelusuran dengan menggunakan Jupyter Notebook. Setelah dilakukan data cleaning ditemukan bahwa tipe data Rating dan Installs ialah numerik, namun sebelum keduanya dibandingkan harus dilakukan penyesuaian terlebih dahulu terhadap kolom Installs dengan mengurutkan *milestone* dari terendah ke tertinggi dan menjadikannya angka integer mulai dari 0. Hal ini dilakukan untuk mengurangi panjang data Installs sehingga ketika dibuatkan bentuk visualisasinya data dapat terbaca dengan baik. Setelah itu barulah dilakukan penelusuran dengan mengunakan metode Pearson Correlation Coeficient dan ditemukan angka 0.21 dan jika angka dari Pearson Correlation Coeficient melebihi 0 kedua data memiliki korelasi yang positif. Untuk bentuk visualisasi dari hubungan antara Rating dengan Installs dapat terlihat..

IV. HASIL PENELITIAN

A. *Dataset*

Berikut ini adalah gambar dari *dataset* yang digunakan pada penelitian ini yaitu Gambar 1 *dataset* Google Playstore dan Gambar 2 *dataset* User Review.

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
1	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
2	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
3	U Launcher Lite – FREE Live Cool Themes, Hide Apps	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
4	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
5	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
6	Paper flowers instructions	ART_AND_DESIGN	4.4	167	5.6M	50,000+	Free	0	Everyone	Art & Design	March 26, 2017	1.0	2.3 and up
7	Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN	3.8	178	19M	50,000+	Free	0	Everyone	Art & Design	April 26, 2018	1.1	4.0.3 and up
8	Infinite Painter	ART_AND_DESIGN	4.1	36815	29M	1,000,000+	Free	0	Everyone	Art & Design	June 14, 2018	6.1.61.1	4.2 and up
9	Garden Coloring Book	ART_AND_DESIGN	4.4	13791	33M	1,000,000+	Free	0	Everyone	Art & Design	September 20, 2017	2.9.2	3.0 and up
10	Kids Paint Free - Drawing Fun	ART_AND_DESIGN	4.7	121	3.1M	10,000+	Free	0	Everyone	Art & Design;Creativity	July 3, 2018	2.8	4.0.3 and up

Gambar 1 Dataset Google Playstore

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
1	10 Best Foods for You	I like eat delicious food. That's I'm cooking food myself, case "10 Best Foods" helps lot, also "Best Before (Shelf Life)"	Positive	1.0	0.5333333333333333
2	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.28846153846153844
3	10 Best Foods for You	nan	nan	nan	nan
4	10 Best Foods for You	Works great especially going grocery store	Positive	0.4	0.875
5	10 Best Foods for You	Best idea us	Positive	1.0	0.3
6	10 Best Foods for You	Best way	Positive	1.0	0.3
7	10 Best Foods for You	Amazing	Positive	0.6000000000000001	0.9
8	10 Best Foods for You	nan	nan	nan	nan
9	10 Best Foods for You	Looking forward app,	Neutral	0.0	0.0
10	10 Best Foods for You	It helpful site ! It help foods get !	Neutral	0.0	0.0

Gambar 2 Dataset User Review

B. Platform Penelitian

Berikut ini adalah *platform* yang digunakan pada Exploratory Data Analysis *dataset* App Google Playstore adalah Jupyter Notebook disertai dengan beberapa *library*-nya yaitu Pandas mengolah *dataset* , Numpy yang mendukung operasi vektor dan matriks ketika mengolah *dataset*, Seaborn dan matplotlib yang mendukung visualisasi data. Berikut ini adalah logo dari *platform* Jupyter Notebook beserta *library*-nya.

Gambar 3 Jupyter Notebook Logo

Gambar 4 Python Library

Gambar 5 matplotlib

C. Data Transformation

1) Proses Transformasi Dataset

Transformasi dataset googleplaystore dan googleplaystore_user_review menggunakan metode import dataset dengan menggunakan pandas seperti pada Gambar 6 dan Gambar 7.

```
# Importing data
dfGapps = pd.read_csv('dataset/googleplaystore.csv')
```

Gambar 6 Import Dataset Google Playstore

```
# Importing data
dfGUserReview = pd.read_csv('dataset/googleplaystore_user_reviews.csv')
```

Gambar 7 Import Dataset User Review

Dan hasil transformasi *dataset* googleplaystore mejadi *dataframe* dapat dilihat pada Gambar 9 dan *dataset* googleplaystore_user_review pada Gambar 8.

	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	19M	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	14M	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	8.7M	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	25M	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	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

Gambar 8 Dataframe Google Playstore

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
1	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
2	10 Best Foods for You	NaN	NaN	NaN	NaN
3	10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000
4	10 Best Foods for You	Best idea us	Positive	1.00	0.300000

Gambar 9 Dataframe User Review

2) Identifikasi Atribut Dataset

Dataset googleplaystore dan googleplaystore_user_review memiliki beberapa atribut yang dapat diidentifikasi seperti tipe data dari setiap kolom, jumlah baris dan kolom, dan nilai unik yang dimiliki oleh setiap kolom. Berikut ini code yang digunakan untuk melihat tipe data dari kolom-kolom yang dimiliki googleplaystore seperti pada Gambar 10 dan kolom-kolom googleplaystore_user_review seperti pada Gambar 11.

```
In [2]: dfGapps.dtypes

Out[2]: App              object
        Category         object
        Rating          float64
        Reviews          object
        Size             object
        Installs         object
        Type             object
        Price            object
        Content Rating   object
        Genres           object
        Last Updated     object
        Current Ver      object
        Android Ver      object
        dtype: object
```

Gambar 10 Tipe Data dataset Google Playstore

```
In [2]: # Baris dan tipe data
        dfGUserReview.dtypes

Out[2]: App                      object
        Translated_Review        object
        Sentiment                object
        Sentiment_Polarity      float64
        Sentiment_Subjectivity  float64
        dtype: object
```

Gambar 11 Tipe Data dataset User Review

Atribut lainnya yaitu total baris dapat dilihat dengan menggunakan kode pada Gambar 12 untuk dataset googleplaystore dan kode pada Gambar 13 untuk dataset googleplaystore_user_review.

```
In [3]: # Total baris dan kolom dalam Dataset
        dfGapps.shape

Out[3]: (10841, 13)
```

Gambar 12 Baris dan Kolom Dataset GooglePlaystore

```
In [3]: # Jumlah baris dan kolom
        dfGUserReview.shape

Out[3]: (64295, 5)
```

Gambar 13 Baris dan Kolom Dataset User Review

Atribut terakhir yang tidak kalah penting adalah nilai unik dari setiap kolom dan untuk dataset googleplaystore berikut ini adalah nilai unik yang dimiliki sebagian kolom :

```
In [5]: # nilai unik dari kolom App
        dfGapps['App'].unique()

Out[5]: array(['Photo Editor & Candy Camera & Grid & ScrapBook',
               'Coloring book moana',
               'U Launcher Lite – FREE Live Cool Themes, Hide Apps', ...,
               'Parkinson Exercices FR', 'The SCP Foundation DB fr nn5n',
               'iHoroscope - 2018 Daily Horoscope & Astrology'], dtype=object)
```

Gambar 14 Nilai Unik Kolom App

```
In [6]: # nilai unik dari kolom Category
        dfGapps['Category'].unique()

Out[6]: array(['ART_AND_DESIGN', 'AUTO_AND_VEHICLES', 'BEAUTY',
               'BOOKS_AND_REFERENCE', 'BUSINESS', 'COMICS', 'COMMUNICATION',
               'DATING', 'EDUCATION', 'ENTERTAINMENT', 'EVENTS', 'FINANCE',
               'FOOD_AND_DRINK', 'HEALTH_AND_FITNESS', 'HOUSE_AND_HOME',
               'LIBRARIES_AND_DEMO', 'LIFESTYLE', 'GAME', 'FAMILY', 'MEDICAL',
               'SOCIAL', 'SHOPPING', 'PHOTOGRAPHY', 'SPORTS', 'TRAVEL_AND_LOCAL',
               'TOOLS', 'PERSONALIZATION', 'PRODUCTIVITY', 'PARENTING', 'WEATHER',
               'VIDEO_PLAYERS', 'NEWS_AND_MAGAZINES', 'MAPS_AND_NAVIGATION',
               '1.9'], dtype=object)
```

Gambar 15 Nilai Unik Kolom Category

```
In [7]: # nilai unik dari kolom Rating
        dfGapps['Rating'].unique()

Out[7]: array([ 4.1,  3.9,  4.7,  4.5,  4.3,  4.4,  3.8,  4.2,  4.6,  3.2,  4. ,
                nan,  4.8,  4.9,  3.6,  3.7,  3.3,  3.4,  3.5,  3.1,  5. ,  2.6,
                3. ,  1.9,  2.5,  2.8,  2.7,  1. ,  2.9,  2.3,  2.2,  1.7,  2. ,
                1.8,  2.4,  1.6,  2.1,  1.4,  1.5,  1.2, 19. ])
```

Gambar 16 Nilai Unik Rating

```
In [10]: # nilai unik dari kolom Installs
         dfGapps['Installs'].unique()

Out[10]: array(['10,000+', '500,000+', '5,000,000+', '50,000,000+', '100,000+',
                '50,000+', '1,000,000+', '10,000,000+', '5,000+', '100,000,000+',
                '1,000,000,000+', '1,000+', '500,000,000+', '50+', '100+', '500+',
                '10+', '1+', '5+', '0+', '0', 'Free'], dtype=object)
```

Gambar 17 Nilai Unik Kolom Installs

Dan untuk *dataset* googleplaystore_user_review berikut ini nilai unik dari sebagian kolom :

```
# nilai unik dari kolom Sentiment
dfGUserReview['Sentiment'].unique()

array(['Positive', nan, 'Neutral', 'Negative'], dtype=object)
```

Gambar 18 Nilai Unik Kolom Sentiment

```
# nilai unik dari kolom Sentiment Polarity
dfGUserReview['Sentiment_Polarity'].unique()

array([ 1.        ,  0.25      ,         nan, ..., -0.52857143,
       -0.37777778,  0.17333333])
```

Gambar 19 Nilai Unik Kolom Sentiment Polarity

```
# nilai unik dari kolom Sentiment Subjectivity
dfGUserReview['Sentiment_Subjectivity'].unique()

array([0.53333333, 0.28846154,        nan, ..., 0.51145833, 0.7172619 ,
       0.2594697 ])
```

Gambar 20 Nilai Unik Kolom Sentiment Subjectivity

D. Data Cleaning

1) Penanggulangan Missing Value

Pada kolom Category terdapat nilai ‘1.9’ yang sangatlah berbeda dengan keseluruhan data pada kolom Category dan setelah di telusuri dengan menggunakan Jupyter Notebook Ditemukan bahwa data terdapat pada baris ke-10472 dan keseluruhan data pada baris tersebut mengalami pergeseran satu kolom ke arah kiri seperti pada Gambar 21.

```
# Mencari nilai dalam kolom Category
dfGapps.loc[dfGapps['Category'] == '1.9']
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10472	Life Made WI-Fi Touchscreen Photo Frame	1.9	19.0	3.0M	1,000+	Free	0	Everyone	NaN	February 11, 2018	1.0.19	4.0 and up	NaN

Gambar 21 Penelusuran nilai '1.9' pada kolom Category

Dan Setelah dilakukkan upaya untuk memperbaiki data yang tergeser dengan melakukan metode shifting dan setelah menelusuri ditemukan kategori yang sebenarnya adalah ‘Lifestyle’ maka kolom Category pada baris tersebut diisi dengan ‘Lifestyle’ dan menghasilkan data seperti Gambar 22

```
# Menambahkan Category yang hilaing yoitu 'LIFESTYLE' ke Aplikasi "Life Made WI-Fi Touchscreen Photo Frame"
dfGapps.iloc[10472,1] = 'LIFESTYLE'
dfGapps[10470:10475]
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10470	Jazz Wi-Fi	COMMUNICATION	3.4	49	4.0M	10,000+	Free	0	Everyone	Communication	February 10, 2017	0.1	2.3 and up
10471	Xposed Wi-Fi-Pwd	PERSONALIZATION	3.5	1042	404k	100,000+	Free	0	Everyone	Personalization	August 5, 2014	3.0.0	4.0.3 and up
10472	Life Made WI-Fi Touchscreen Photo Frame	LIFESTYLE	1.9	19	3.0M	1,000+	Free	0	Everyone	NaN	February 11, 2018	1.0.19	4.0 and up
10473	osmino Wi-Fi: free WiFi	TOOLS	4.2	134203	4.1M	10,000,000+	Free	0	Everyone	Tools	August 7, 2018	6.06.14	4.4 and up
10474	Sat-Fi Voice	COMMUNICATION	3.4	37	14M	1,000+	Free	0	Everyone	Communication	November 21, 2014	2.2.1.5	2.2 and up

Gambar 22 Hasil dari Proses Penanganan Missing Data pada Kolom Category

Terdapat missing data lain pada *dataset* Google Playstore dan User Review yang telah di temukan juga seperti pada Gambar 23 dan Gambar 24.

	Total	Percent
Rating	1458	0.151104
SizeC	1226	0.127060
Genres	1	0.000104
Android Ver	0	0.000000
Current Ver	0	0.000000
Last Updated	0	0.000000
Content Rating	0	0.000000
PriceC	0	0.000000
Price	0	0.000000
Type	0	0.000000
InstallsC	0	0.000000
Installs	0	0.000000
Size	0	0.000000
Reviews	0	0.000000
Category	0	0.000000

Gambar 23 Hasil Pencarian data yang bernilai Nan pada dataset Google Playstore

	Total	Percent
Translated_Review	26868	0.417886
Sentiment_Subjectivity	26863	0.417809
Sentiment_Polarity	26863	0.417809
Sentiment	26863	0.417809
App	0	0.000000

Gambar 24 Hasil Identifikasi data yang bernilai Nan pada dataset User Review

Dan missing data pada kedua *dataset* akan ditanggulangi dengan menggunakan metode drop data dan substitusi menggunakan mean dari tiap kolom seperti pada

```
# Mengisi data yang bernilai Null dengan nilai rata-rata tiap kolom
column_means = dfGapps.mean()
dfGapps = dfGapps.fillna(column_means)

# Untuk nilai mean pada Kolom Rating nilai akan dibulatkan menjadi 1 desimal
dfGapps['Rating'] = round(dfGapps['Rating'], 1)

# Menghapus sisa data yang masih bernilai Nan
dfGapps.dropna(inplace=True)

# Memastikan sudah tidak ada missing value
print("Data yang Bernilai NaN : ")
dfGapps.isnull().sum().max()
```

```
Data yang Bernilai NaN :
0
```

Gambar 25 Penanganan nilai Nan pada dataset Google Playstore

```
In [9]: dfGUserReview = dfGUserReview.drop(dfGUserReview.loc[dfGUserReview['Translated_Review'].isnull()].index)
        dfGUserReview = dfGUserReview.drop(dfGUserReview.loc[dfGUserReview['Sentiment_Subjectivity'].isnull()].index)
        dfGUserReview = dfGUserReview.drop(dfGUserReview.loc[dfGUserReview['Sentiment_Polarity'].isnull()].index)
        dfGUserReview = dfGUserReview.drop(dfGUserReview.loc[dfGUserReview['Sentiment'].isnull()].index)
        dfGUserReview.isnull().sum().max()

Out[9]: 0
```

Gambar 26 Drop data yang bernilai Nan pada dataset User Review

2) Penyesuaian Tipe Data

Berikut ini adalah kode-kode yang digunakan untuk menyesuaikan beberapa kolom pada *dataset* Google Playstore :

```
#2 Mengkonversi Kolom Size ke SizeC
dfGapps.insert(5, "SizeC", dfGapps['Size'].apply(lambda x: x.replace(',', '') if ',' in str(x) else x), True)
```

```
# Menyamakan Ukuran data dari Kolom SizeC
def change_size(size):
    if 'M' in size:
        x = size[:-1]
        x = float(x)*1000000
        return(x)
    elif 'k' == size[-1:]:
        x = size[:-1]
        x = float(x)*1000
        return(x)
    else:
        return None

dfGapps["SizeC"] = dfGapps["SizeC"].map(change_size)
```

Gambar 27 Konversi kolom Size menjadi SizeC

```
#3 Mengkonversi Kolom Installs ke InstallsC
dfGapps.insert(7, "InstallsC", dfGapps['Installs'].apply(lambda x: x.replace(',', '') if ',' in str(x) else x), True)
dfGapps.InstallsC = dfGapps['InstallsC'].apply(lambda x: x.replace('+',''))
dfGapps.InstallsC = dfGapps.InstallsC.apply(lambda x: int(x))
```

Gambar 28 Proses Konversi kolom Installs menjadi InstallsC

```
#4 Mengkonversi Kolom Price ke PriceC
dfGapps.insert(10, "PriceC", dfGapps['Price'])
dfGapps["PriceC"] = dfGapps.PriceC.apply(lambda x: x.strip('$'))
dfGapps['PriceC'] = pd.to_numeric(dfGapps['PriceC'])
```

Gambar 29 Proses Konversi kolom Price mnejadi PriceC

```
#6 mengkonversi tipe data dari kolom Rating & kolom Reviews
dfGapps['Rating'] = pd.to_numeric(dfGapps['Rating'])
dfGapps['Reviews'] = pd.to_numeric(dfGapps['Reviews'])
```

Gambar 30 Proses Konversi kolom Rating dan Reviews ke tipe numerik

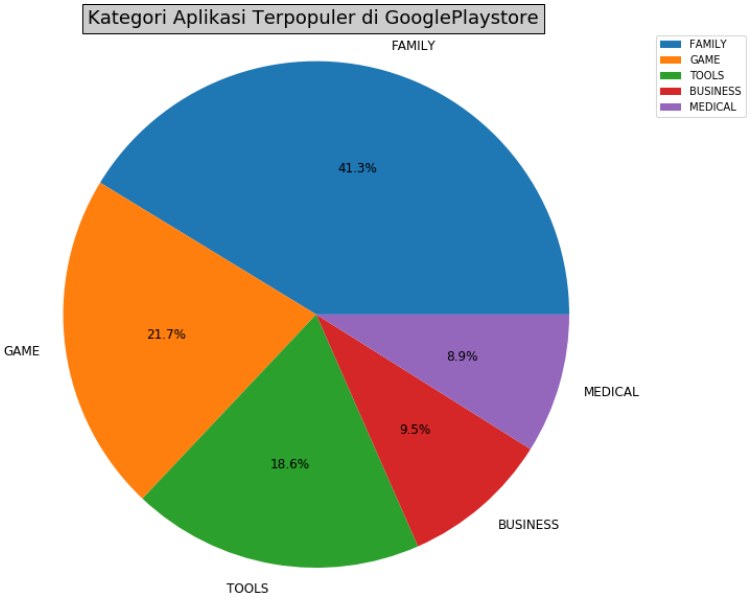
```
#5 Mengkonversi kolom "Last Updated" menjadi Datetime objects
dfGapps['Last Updated'] = pd.to_datetime(dfGapps['Last Updated'])
```

Gambar 31 Proses Konversi kolom Last Update ke tipe datetime object

E. Exploratory Data Analysis Google Playstore

1) Category Apikasi Terpopuler dalam Google Playstore

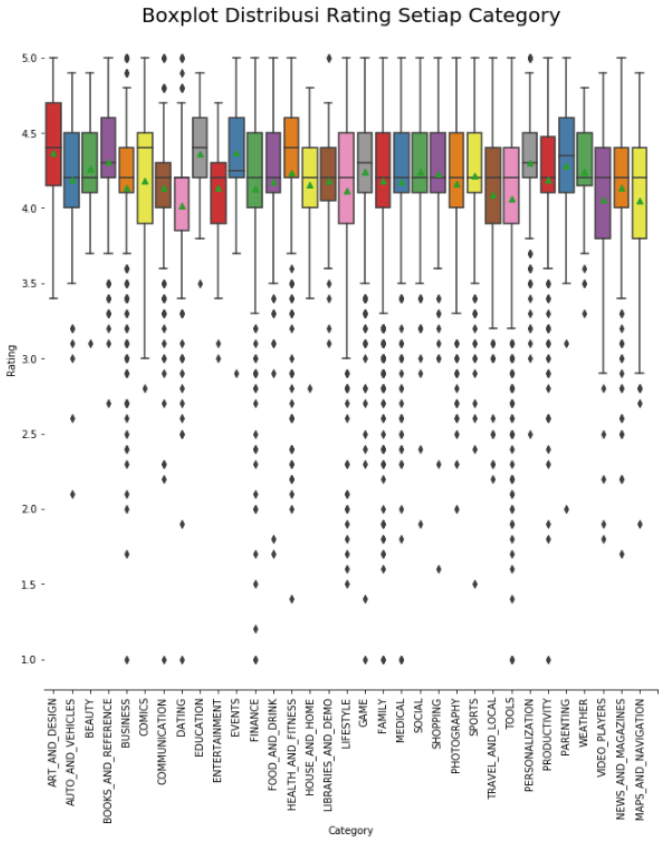
Berikut ini adalah hasil visualisasi dari eksplorasi mengenai Category Aplikasi Terpopiler dalam Google Playstore :



Gambar 32 Hasil Visualisasi Category Aplikasi Terpopuler

2) Distribusi Rating di Setiap Category

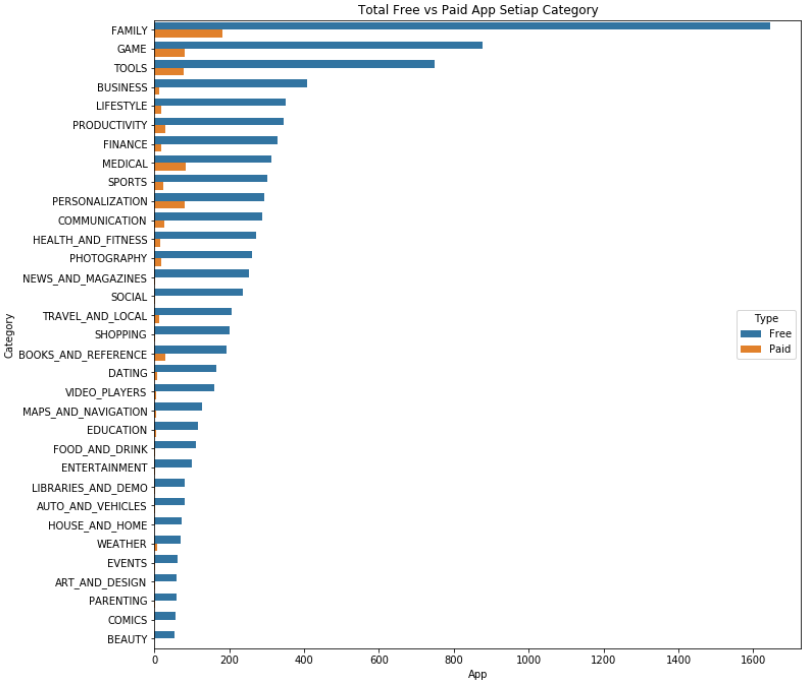
Berikut ini adalah hasil visualisasi dari eksplorasi mengenai Distribusi Rating di Setiap Category :



Gambar 33 Hasil Visualisasi Distribusi Rating di Setiap Category

3) Tipe Aplikasi yang Mendominasi Google Playstore

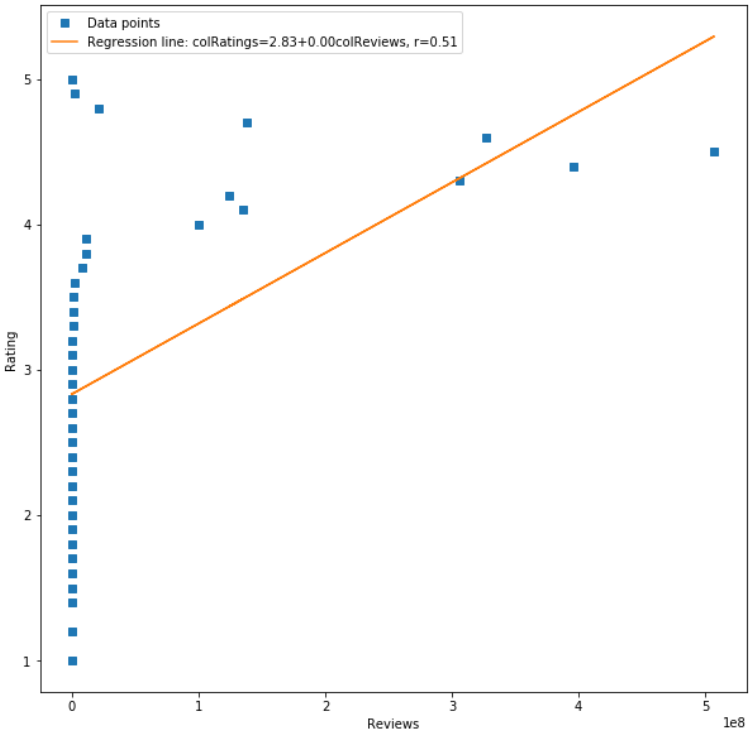
Berikut ini adalah hasil visualisasi dari eksplorasi mengenai Tipe Aplikasi yang Mendominasi Google Playstore :



Gambar 34 Hasil Visualisasi Tipe Aplikasi yang Mendominasi Google Playstore

4) Hubungan antara Rating dan Review

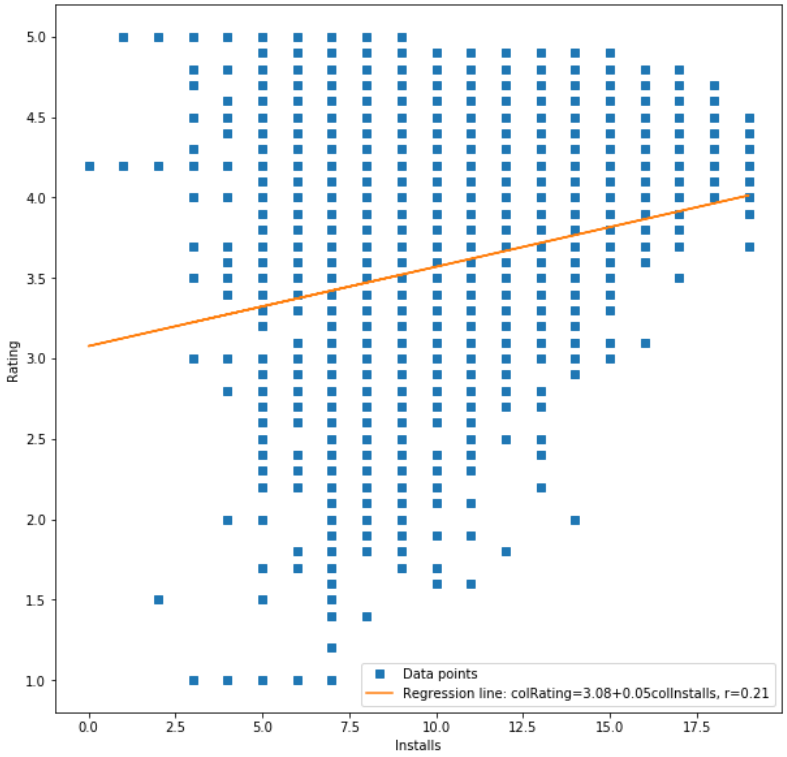
Berikut ini adalah hasil visualisasi dari eksplorasi mengenai Hubungan antara Rating dan Review :



Gambar 35 Hasil Visualisasi Hubungan antara Rating dan Review

5) Hubungan antara Rating dan Installs

Berikut ini adalah hasil visualisasi dari eksplorasi mengenai Hubungan antara Rating dan Review :



Gambar 36 Hasil Visualisasi Hubungan antara Rating dan Installs

V. KESIMPULAN

Setelah dilakukan penelitian berupa Exploratory Data Analysis pada *dataset* App Google Playstore ditemukan bahwa :

1. Top 5 Category aplikasi terpopuler pada *dataset* Google Playstore ialah Family, Game, Tools, Business, dan Medical.
2. Distribusi rata-rata Rating di setiap Category berada pada angka 4.0 hingga 4.5
3. Top 5 Category dengan aplikasi *free* terbanyak adalah Family, Game, Tools, Business, dan Lifestyle, sedangkan top 5 Category dengan aplikasi *paid* terbanyak adalah Family, Medical, Game, Personalization, dan Tools.
4. Rating memiliki korelasi positif dengan Reviews dari aplikasi Google Playstore.
5. Rating memiliki korelasi positif dengan Installs dari aplikasi Google Playstore.

DAFTAR PUSTAKA

[1] F. Nelli, Python Data Analytics With Pandas, NumPy, and Matplotlib, Rome, Italy: Apress Media LLC, 2018.
[2] W. L. Martinez, A. R. Martinez dan L. J. Solka, Exploratory Data Analysis with MATLAB®, Third Edition, Anbingdon, UK: Taylor and Francis Group, 2017.
[3] S. Jansen dan E. Bloemendal, Defining App Stores: The Role of Curated Marketplaces in Software Ecosystems, Utrecht, Netherlands, 2013.
[4] W. McKinney, Python for Data Analysis, United States of America, 2018.