

TUGAS BESAR: Eksplorasi Data YouTube 2023

IF5100 Pemrograman Basis Data



23525063 - Fadhlan Nazhif Azizy
23525036 - Ivan Hardja
23525037 - Katherine Febrianty Sumartono

Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
2025/2026

Link Github: <https://github.com/Aureus7777/TUBES-PDA-18>

Link Dataset: <https://www.kaggle.com/datasets/nelgiryewithana/global-youtube-statistics-2023>

Pendahuluan

Platform YouTube adalah sebuah sarana yang sering digunakan oleh orang sebagai sumber entertainment baik itu musik, pertunjukkan, drama, dll serta edukasi. Konten yang variasi ini terpelihara dengan ekosistem monetisasi yang bergantung pada jumlah penonton, subscriber, jenis konten, dan lokasi. Sehingga selain sumber entertainment bagi beberapa orang YouTube juga merupakan sarana pendapatan. Oleh karena itu, kelompok kami menentukan untuk menggunakan dataset Global YouTube Statistics 2023 karena data yang disimpan merupakan data yang relevan dengan saat ini. Maka, dapat diketahui estimasi pendapatan channel yang dapat membantu creator, agensi maupun analisis marketing dalam menyusun strategi konten dan investasi.

Tujuan bisnis penggunaan dataset ini adalah untuk membangun model Machine Learning yang mampu untuk memprediksi pendapatan tahunan channel YouTube berdasarkan fitur channel. Lalu, mengidentifikasi faktor-faktor utama yang mempengaruhi pendapatan dan membantu proses pengambilan keputusan yang berbasis data terkait performa channel.

Berhubungan dengan itu, terdapat beberapa pertanyaan yang ingin didapatkan jawabannya berdasarkan dataset:

1. Apakah karakteristik channel dapat memprediksi pendapatan secara akurat?
2. Fitur apa yang paling berpengaruh terhadap pendapatan?
3. Bagaimana pendapatan bervariasi antar negara dan jenis channel?

Model yang dibangun dapat terbilang sukses apabila model mampu memberi prediksi pendapatan dengan error yang rendah dan insight yang dihasilkan relevan serta dapat digunakan secara praktis.

Dataset

Dataset Global YouTube Statistics memiliki sebanyak 28 fitur dan 995 baris data. Berikut adalah tabel dengan fitur dan tipe data yang disimpan.

Fitur	Tipe Data
rank	int
YouTuber	object
subscribers	int
video views	float
category	object (19 nilai unik: Music, Film & Animation,

	Entertainment, Education, Shows, People & Blogs, Gaming, Sports, Howto & Style, News & Politics, Comedy, Trailers, Nonprofits & Activism, Science & Technology, Movies, Pets & Animals, Autos & Vehicles, Travel & Events, nan)
Title	object
uploads	int
Country	object (50 nilai unik untuk nama negara)
Abbreviation	object (50 nilai unik untuk singkatan nama negara)
channel_type	object (15 nilai unik: Music, Games, Entertainment, Education, People, Sports, Film, News, Comedy, Howto, Nonprofit, Autos, Tech, Animals, nan)
video_views_rank	float
video_views_for_the_last_30_days	float
lowest_monthly_earnings	float
highest_monthly_earnings	float
lowest_yearly_earnings	float
highest_yearly_earnings	float
subscribers_for_last_30_days	float
created_year	float
created_month	object (13 nilai unik untuk bulan, dan lebih satu untuk nilai nan)
created_date	float
Gross tertiary education enrollment (%)	float
Population	float
Unemployment rate	float
Urban_population	float
Latitude	float
Longitude	float

Persiapan Data

Berdasarkan tabel terlihat bahwa nama fitur tidak konsisten oleh karena hal yang pertama dilakukan adalah membuat nama fitur menjadi konsisten dimana semua nama fitur di lowercase serta diganti fitur dengan empty space dengan underscore ‘_’. Sesudah itu untuk semua fitur dengan tipe data object dihapus white space next line, untuk merapihkan data.

Selanjutnya data duplikat yang ada juga perlu dihapus, maka pertama dihapus untuk semua nilai duplikat pada fitur YouTuber dan disimpan entitas data yang pertama. Ketika dicek diketahui bahwa baris data sebelum dilakukan drop dan sesudah sama yaitu 995 baris maka tidak ada data duplikat berdasarkan nama YouTuber. Lalu dihapus juga baris data pada kolom subscribers, video views dan uploads yang memiliki nilai negatif, karena tidak mungkin pada kolom tersebut memiliki nilai negatif. Sesudah dilakukan drop sama jumlah baris dengan nilai negatif yang akan dibuang ada sebanyak 0 baris, oleh karena itu data sudah bersih pada kolom ini semua.

Sesudah itu dilakukan pemeriksaan apabila terdapat missing values per kolom, dan didapatkan terdapat beberapa kolom yang memiliki missing value seperti pada kolom ‘category’, ‘channel_type’, ‘country’, dan ‘abbreviation’. Dikarenakan data penting maka untuk missing value nan diisi dengan value bentuk string ‘Unknown’. Namun, masih ada kolom dengan data yang tidak ada seperti ‘Gross tertiary education enrollment (%)’, ‘Population’, ‘Unemployment rate’, dll. Kolom tersebut masuk akal tidak ada valuenya karena kolom ‘country’ mereka tidak ada, oleh karena itu dibiarkan nilainya kosong. Selain itu, didapatkan juga pada kolom ‘created_year’, ‘created_month’ dan ‘created_date’ yang tidak memiliki nilai. Seharusnya sebuah channel pada YouTube disimpan data mereka dibuat, maka data ini dihapus. Ditemukan bahwa setelah dihapus dari 995 baris data menjadi 990 baris data yang berarti terdapat 5 baris data yang memiliki nilai kosong untuk fitur tersebut.

Maka, sebenarnya dataset sudah termasuk bersih namun dipersiapkan karena data yang tidak memiliki value merupakan data yang tergolong sebagai kategori dan banyak data kosong pada fitur lainnya yang berhubungan dengan data tersebut. Selain itu dihapus juga beberapa data yang tidak masuk akal disimpan seperti data ‘created_year’, karena tidak mungkin sebuah channel YouTube tidak disimpan tanggal serta waktu kapan channel itu dibuat.

EDA

Pada tahapan ini dibuat sebuah fitur baru untuk dataset yaitu umur channel aktif di YouTube sudah berapa lama. Isi data ini adalah lama waktu semenjak channel dibuat sampai waktu sekarang ini. Berikut adalah sebuah tabel

	rank	youtuber	subscribers	video_views	category	title	uploads	country
0	1	T-Series	245000000	2.280000e+11	Music	T-Series	20082	India
1	2	YouTube Movies	170000000	0.000000e+00	Film & Animation	youtubemovies	1	United States
2	3	MrBeast	166000000	2.836884e+10	Entertainment	MrBeast	741	United States
3	4	Cocomelon - Nursery Rhymes	162000000	1.640000e+11	Education	Cocomelon - Nursery Rhymes	966	United States
4	5	SET India	159000000	1.480000e+11	Shows	SET India	116536	India

Seperti terlihat pada tabel secara singkat adalah data yang didapatkan dengan menjalankan `df.head()`.

Sesudah itu dijalankan potongan kode untuk mendapatkan statistika dataset pada kolom numerik dataset. Dari ringkasan yang didapatkan terdapat variasi yang luas pada fitur-fitur numerik seperti 'subscribers' dengan nilai rata-rata 22.9 juta dan 'video_views' dengan nilai rata-rata 11 miliar, dengan nilai maksimum yang sangat tinggi mengindikasikan adanya outlier atau channel yang sangat besar.

Selain itu, seperti yang sudah dijelaskan pada tahapan persiapan data, baris data dengan nilai null tentu perlu diketahui secara lebih lanjut. Oleh karena itu dengan menghitung total baris data dengan nilai yang hilang didapatkan seperti pada gambar berikut

===== Kolom yang mengandung nilai null beserta jumlahnya =====

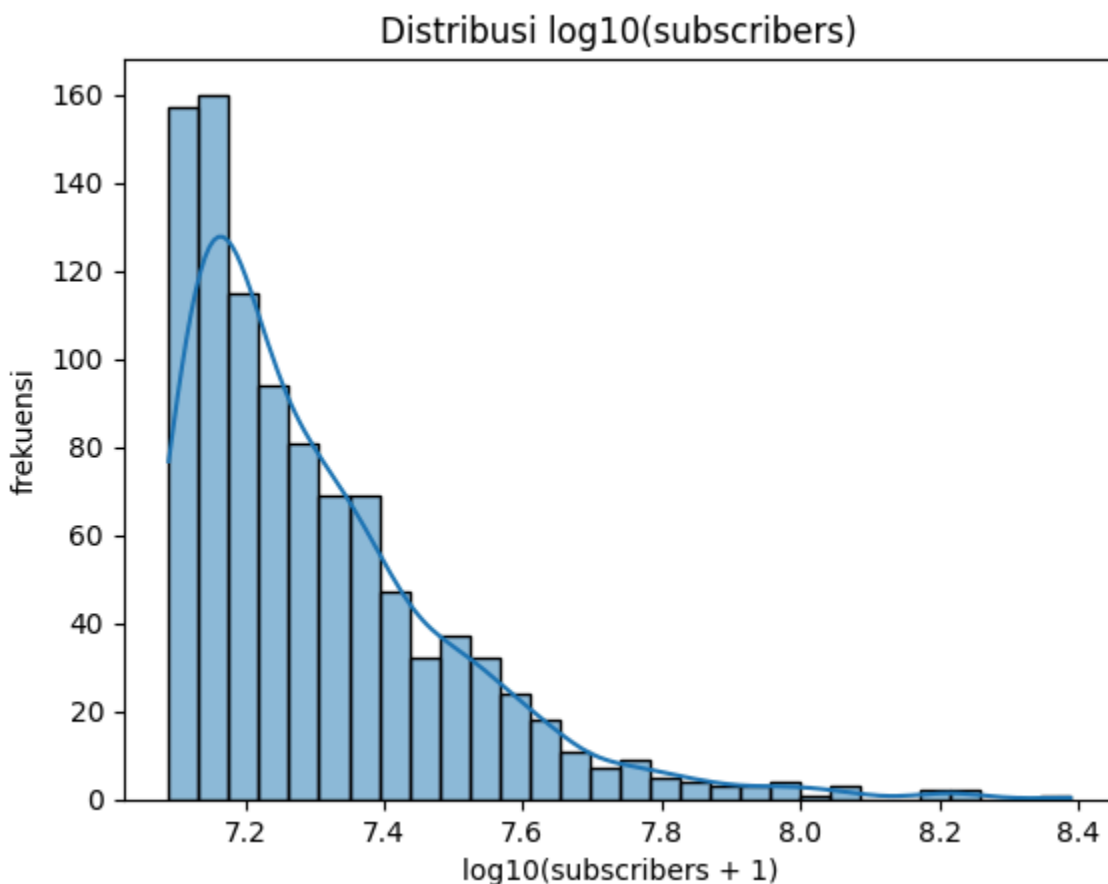
	null_count	null_pct
country	120	12.121212
abbreviation	120	12.121212
country_rank	114	11.515152
channel_type_rank	29	2.929293
video_views_for_the_last_30_days	51	5.151515
subscribers_for_last_30_days	332	33.535354
gross_tertiary_education_enrollment_(%)	121	12.222222
population	121	12.222222
unemployment_rate	121	12.222222
urban_population	121	12.222222
latitude	121	12.222222
longitude	121	12.222222

Berdasarkan gambar tersebut diketahui bahwa sebagian besar nilai yang hilang ditemukan pada fitur terkait informasi negara. Serta fitur 'subscribers_for_last_30_days' yang memiliki persentase kehilangan tertinggi sebesar 33.87%.

Didapatkan juga kalau negara dengan jumlah channel terbanyak dalam dataset adalah Amerika Serikat dengan 312 channel dan India dengan 168 channel sementara channel ketiga terbanyak yaitu Brazil hanya sebanyak 61 channel. Hal tersebut mencerminkan pasar YouTube terbesar. Juga didapatkan untuk kategori dengan jumlah subscribers terbanyak didominasi oleh kategori 'Shows', 'Trailers' dan 'Film & Animation' yang memiliki rata-rata subscribers tertinggi, menunjukkan bahwa jenis konten tersebut menarik banyak subscribers.

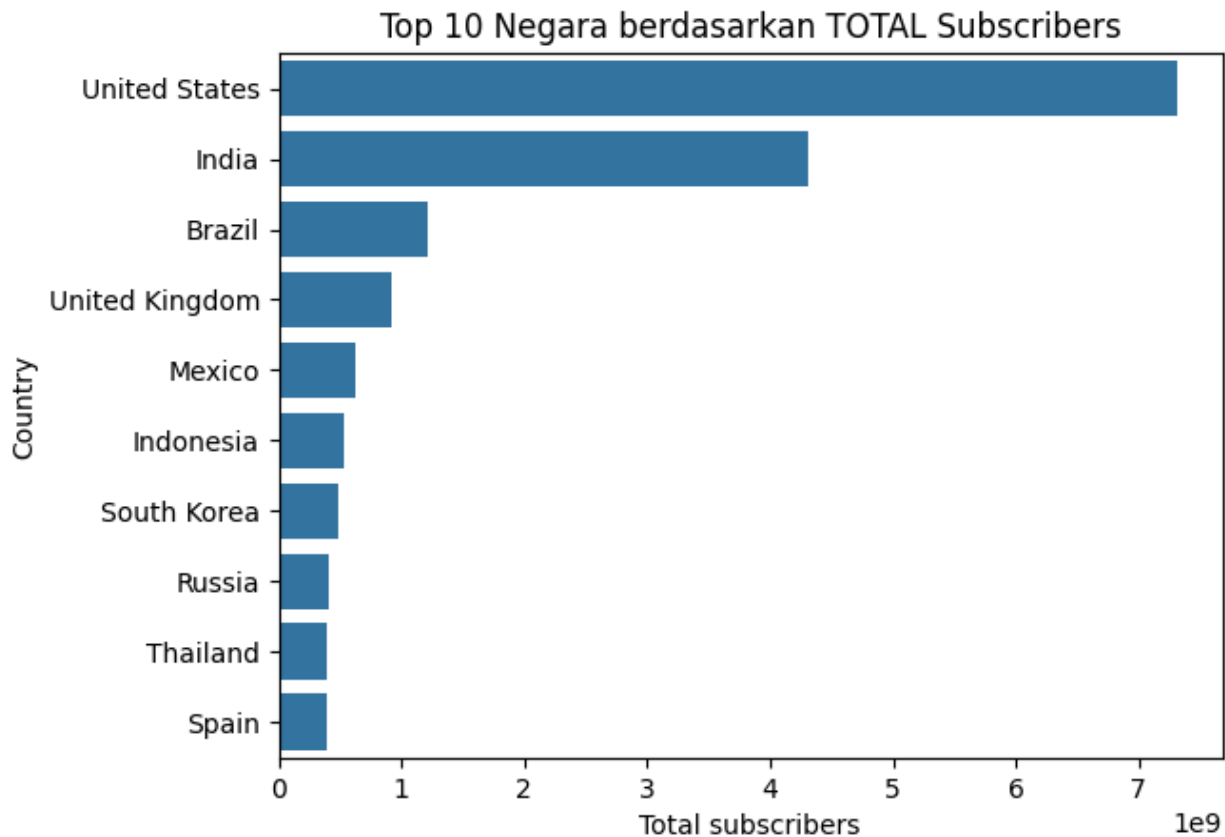
Visualisasi Data

Distribusi Subscribers



Distribusi 'log-transformed subscribers' menunjukkan mayoritas channel memiliki jumlah subscribers yang rendah, dengan tail yang memanjang ke kanan menjadi semakin kecil menunjukkan keberadaan beberapa channel yang memiliki jumlah subscribers yang sangat tinggi.

Distribusi Negara



Pada horizontal bar chart diatas terlihat kalau Amerika Serikat dan India secara signifikan memimpin dalam total 'subscribers', hal ini konsisten dengan jumlah channel terbanyak yang juga berbasis di Amerika Serikat dan India.

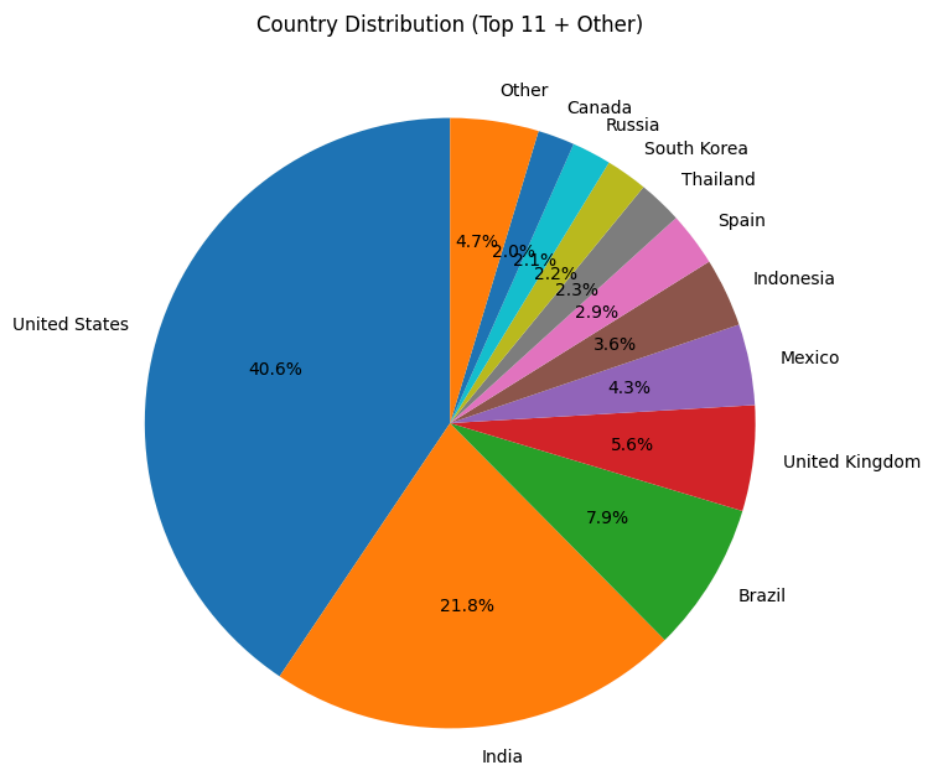
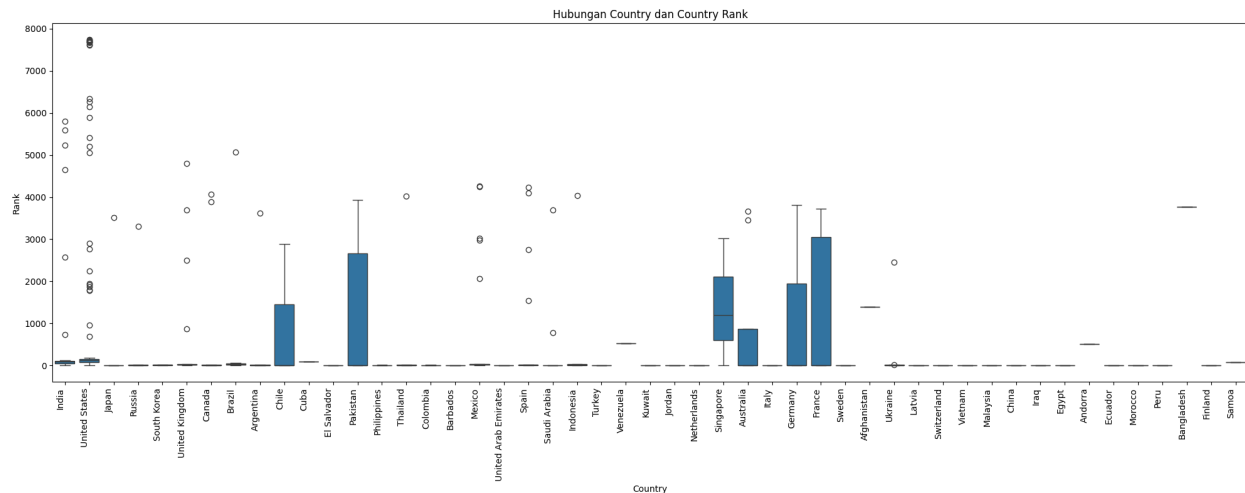


Diagram lingkaran dibuat untuk memvisualisasikan distribusi kategori channel dengan negara, dan didapatkan diagram yang menunjukkan dominasi channel dari Amerika Serikat (40.6%) dan India (21.8%) dalam dataset, yang selaras dengan temuan sebelumnya mengenai jumlah channel dan total subscribers.

Hubungan country dan country_rank

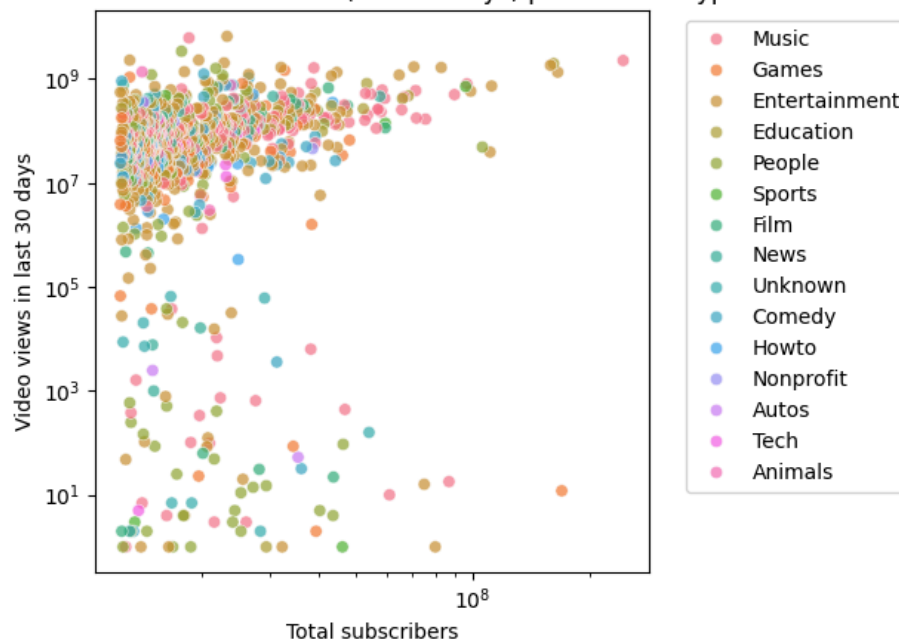


Visualisasi ini menunjukkan distribusi 'country rank' untuk setiap negara. Terdapat banyak outlier (titik-titik di luar batas kotak) untuk beberapa negara, yang mengindikasikan variasi 'rank' yang signifikan dalam suatu negara atau keberadaan channel dengan rank yang sangat berbeda

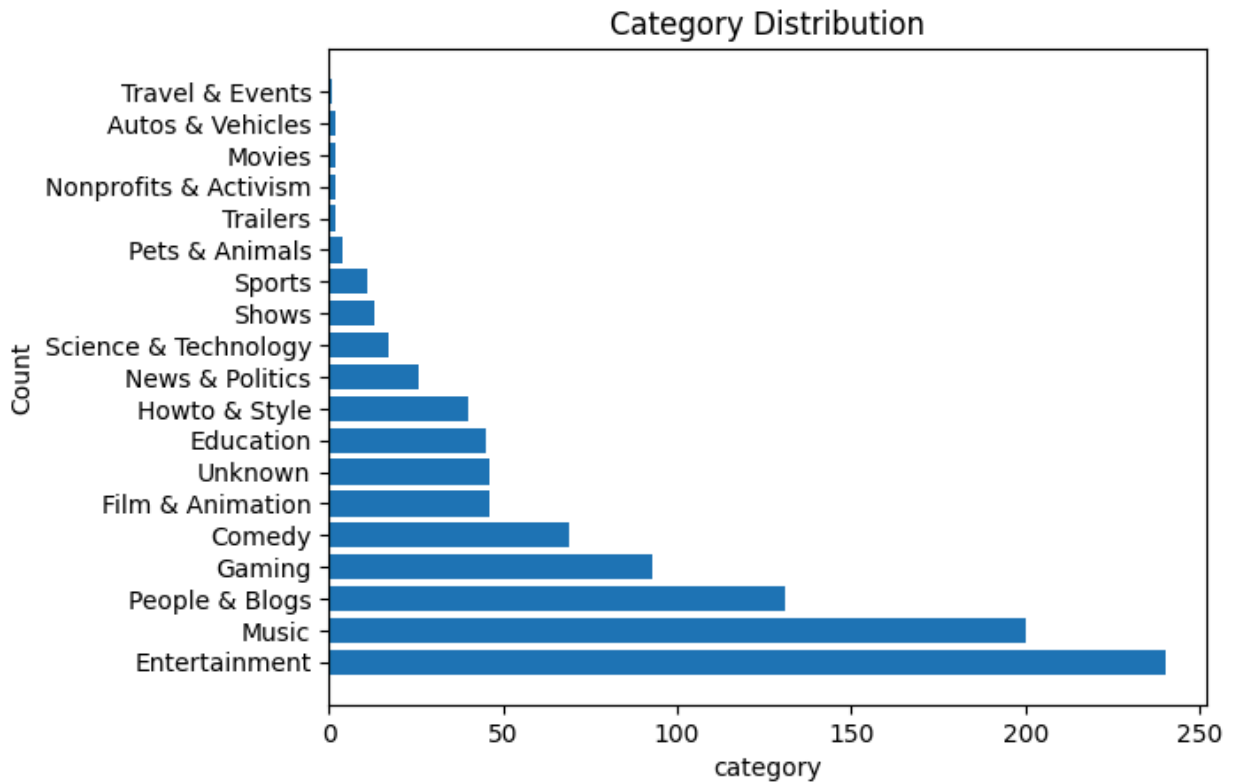
Subscribers vs Video Views

Plot menunjukkan hubungan antara jumlah subscribers dan video views dalam 30 hari terakhir, yang bervariasi antara jenis channel. Sebagian besar channel menunjukkan hubungan positif, di mana subscriber dan views cenderung meningkat bersama.

Subscribers vs Video Views (Last 30 days) per Channel Type

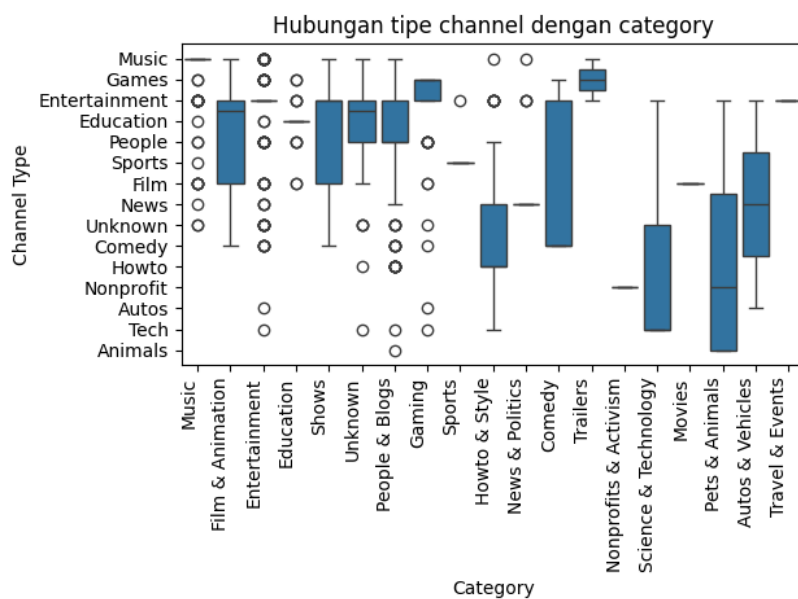


Distribusi Kategori



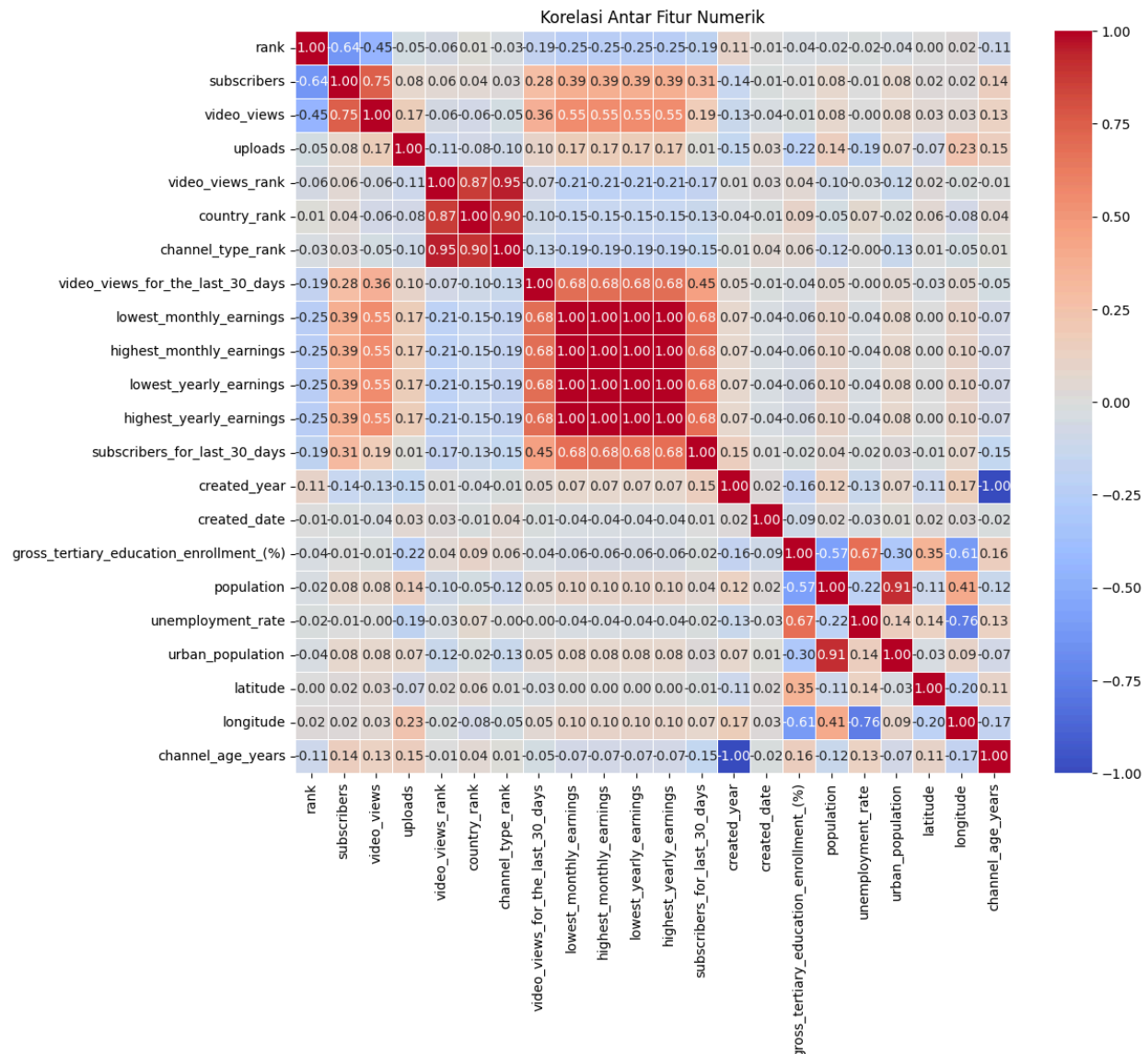
Pada diagram terlihat kalau kategori 'Entertainment', 'Music' dan 'People & Blogs' adalah kategori yang paling umum dalam dataset ini, menunjukkan popularitas jenis konten tersebut

Hubungan tipe channel dengan category



Plot ini menggambarkan bagaimana tipe channel tersebar di berbagai kategori. Beberapa kategori memiliki distribusi tipe channel yang lebih bervariasi daripada yang lain, sementara ada juga kategori yang didominasi oleh satu jenis channel tertentu.

Korelasi Antar Fitur Numerik



Heatmap korelasi menunjukkan hubungan yang kuat antara beberapa fitur numerik. Kolom pendapatan memiliki korelasi positif yang tinggi satu sama lain. 'Subscribers', 'video_views' dan 'video_views_for_lat_30_days' juga memiliki korelasi positif. 'Rank' berkorelasi negatif dengan metrik performa channel yang berarti peringkat yang lebih rendah (angka kecil) menunjukkan performa yang lebih baik.

Model Inferensi

Dalam membangun model inferensi dilakukan beberapa tahapan persiapan lagi. Pertama dipastikan kembali kalau data NaN sudah dihapus dari dataset. Setelah itu, diputuskan bahwa kolom 'highest_yearly_earnings' digunakan sebagai target prediksi dengan tujuan untuk memprediksi pendapatan terbesar tiap tahunnya. Data kemudian dibagi dua untuk train dan test dengan pembagian sebesar 60:40 sehingga didapatkan data training sebesar 594 sampel dan data testing sebesar 398 sampel. Lalu digunakan model *RandomForestRegressor* dengan *preprocessing* dan *feature selection* menunjukkan performa cross-validation yang sangat baik dengan nilai rata-rata R^2 sebesar 0.9682 dan standar deviasi yang relatif kecil (0.0468). Ini menunjukkan bahwa model cukup stabil dan generalisasi yang baik pada data yang belum pernah dilihat sebelumnya.

```
=== Performance Test ===
Train R2: 0.9948
Test  R2: 0.9442
-----
MAE   : 396796.9578
MSE   : 14088758576693.3340
RMSE  : 3753499.5107

=== Sample Predictions ===
Actual: 828600.00   Predicted: 846929.64
Actual: 76700000.00 Predicted: 79077536.91
Actual: 5700000.00  Predicted: 5652060.85
Actual: 56300000.00 Predicted: 58089270.11
Actual: 1000000.00  Predicted: 1029098.12
```

Dari hasil yang didapatkan, model telah menunjukkan performa yang sangat baik pada data training (R^2 : 0.9919) dan cukup baik pada data testing (R^2 : 0.943), mengindikasikan kemampuan prediksi yang kuat. MAE, MSE dan RMSE menunjukkan bahwa model memiliki rata-rata kesalahan prediksi sekitar 442.8 ribu untuk pendapatan tahunan tertinggi. Contoh prediksi menunjukkan kalau perbedaan nilai aktual dan nilai prediksi yang sangat dekat.

Model ini dibuat untuk membantu orang yang ingin membuat channel YouTube baru untuk merencanakan tipe kategori konten yang ingin dibuat dan berdasarkan dari data tahun 2023 ini yang masih relevan seberapa besar penghasilan tahunan yang bisa didapatkan.

Kesimpulan

Berdasarkan percobaan ini didapatkan model yang mampu untuk memprediksi pendapatan tahunan dari YouTube yang baik karena nilai error yang kecil. Pendapatan tersebut paling dipengaruhi dengan jumlah subscribers dan views. Selain itu, country dan category juga berpengaruh pada performa channel itu sendiri. Model ini berguna untuk strategi bagi para content creator atau yang mau menjadi content creator mengenai konten yang hendak mereka buat jika ingin memaksimalkan keuntungan yang bisa mereka dapatkan dari YouTube.

Kontribusi Anggota

NIM	Nama	Kontribusi
23525063	Fadhlan Nazhif Azizy	Merapikan + Insight EDA, Menambahkan Feature Selection, Menambahkan K-Cross Validation
23525036	Ivan Hardja	Mencari dataset yang digunakan, menambahkan bagian data preprocessing , membuat laporan, mengedit video presentasi
23525037	Katherine Febrianty Sumartono	Data Preprocessing, EDA, Visualization, membuat slide powerpoint untuk video