pyladies.yk

Modul Analisis Data Dasar

1. Pendahuluan

Modul ini dirancang untuk memberikan pemahaman fundamental mengenai proses analisis data menggunakan bahasa pemrograman Python dalam lingkungan Google collaboratory. Sebagai contoh praktis, kita akan menggunakan dataset dari Netflix, yang awalnya bernama netflix.csv dan diperoleh dari Kaggle. Penting untuk dicatat bahwa metode dan teknik yang diajarkan dalam modul ini bersifat umum dan dapat diaplikasikan pada berbagai jenis dataset di masa mendatang. Kita akan melakukan pembersihan data langsung di Google collab, yang kemudian akan menghasilkan *file* cleaned_data_netflix.csv. Proses analisis selanjutnya akan dilaksanakan dalam lingkungan interaktif Google collab, yang sangat ideal untuk eksperimen dan visualisasi data.

2. Tujuan Pembelajaran

Setelah menyelesaikan modul ini, peserta diharapkan mampu:

- Memahami struktur dan karakteristik dataset contoh (Netflix).
- Melakukan pra-pemrosesan data untuk menghasilkan dataset yang bersih dan siap analisis.
- Melakukan eksplorasi data awal dan menerapkan teknik visualisasi data sederhana.
- Menganalisis tren temporal pada data contoh.
- Memahami relevansi dan aplikasi analisis data dalam konteks industri, serta termotivasi untuk menjelajahi dataset lain.

3. Lingkungan Pengembangan dan Pustaka

Lingkungan Pengembangan

- Google collaboratory (collab): Platform komputasi awan yang memungkinkan penulisan dan eksekusi kode Python secara interaktif tanpa memerlukan konfigurasi lingkungan lokal.
- Kaggle: Platform komunitas data science yang menyediakan beragam dataset publik dan lingkungan *notebook* berbasis *cloud* untuk analisis data.
- Looker Studio (Google Data Studio): Alat visualisasi data dan *dashboarding* berbasis *cloud* yang memungkinkan pembuatan laporan interaktif dari berbagai sumber data.

Pustaka Python

- pandas: Pustaka esensial untuk manipulasi dan analisis data, menyediakan struktur data DataFrame yang fleksibel.
- matplotlib.pyplot: Pustaka dasar untuk pembuatan visualisasi statis dan interaktif dalam Python.
- seaborn: Pustaka visualisasi data tingkat tinggi yang dibangun di atas Matplotlib, dirancang untuk membuat grafik statistik yang menarik dan informatif.

4. Persiapan Data

Dataset awal yang akan digunakan sebagai contoh adalah netflix.csv, yang dapat diunduh dari Kaggle. Sebelum analisis, dataset ini akan melalui proses pembersihan di Google collab untuk menangani nilai yang hilang dan informasi tidak relevan. Hasil pembersihan akan disimpan sebagai cleaned_data_example.csv. Kolom date_added dalam dataset ini akan menjadi fokus utama untuk analisis tren temporal pada data contoh ini.

5. Tahapan Analisis Data

Bagian ini menguraikan langkah-langkah sistematis untuk melakukan analisis data dasar.

a. Impor Pustaka dan Pemuatan Dataset Awal

Langkah awal dalam analisis data adalah mengimpor pustaka yang diperlukan dan memuat dataset awal netflix.csv ke dalam DataFrame Pandas. Pastikan Anda telah mengunduh netflix.csv dari Kaggle dan mengunggahnya ke sesi Google Collab Anda (lihat bagian Panduan Platform Pendukung untuk cara mengunggah *file*).

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Memuat dataset awal dari file CSV
# Pastikan file 'netflix.csv' sudah diunggah ke sesi collab Anda
try:
    df = pd.read_csv("netflix.csv")
    print("Dataset 'netflix.csv' berhasil dimuat.")
except FileNotFoundError:
     print("Error: File 'netflix.csv' tidak ditemukan. Pastikan Anda telah mengunggahnya ke Google collab.")
    # Kode untuk mengunggah file bisa ditambahkan di sini atau instruksikan user
    from google.collab import files
    uploaded = files.upload()
    df = pd.read_csv("netflix.csv") # Coba lagi setelah upload
```

```
# Menampilkan 5 baris pertama untuk inspeksi awal
print("\n5 baris pertama dataset awal:")
print(df.head())
print("\nInformasi dataset awal:")
df.info()
```

b. Pra-pemrosesan Data dan Penyimpanan Hasil

Tahap ini melibatkan pembersihan data, termasuk penanganan nilai yang hilang, dan kemudian menyimpan dataset yang telah dibersihkan sebagai cleaned_data_example.csv.

```
# Membuat salinan DataFrame untuk operasi pembersihan
df_cleaned = df.copy()

# Menangani nilai yang hilang di kolom 'date_added' dengan menghapus baris terkait
# Kolom ini krusial untuk analisis tren temporal pada contoh ini
df_cleaned.dropna(subset=['date_added'], inplace=True)

# Mengonversi kolom 'date_added' ke tipe datetime
# 'errors='coerce'' akan mengubah nilai yang tidak dapat dikonversi menjadi NaT (Not a Time)
df_cleaned['date_added'] = pd.to_datetime(df_cleaned['date_added'], errors='coerce')

# Mengekstrak tahun dari kolom 'date_added' dan menyimpannya di kolom baru 'year_added'
# Baris dengan NaT (Not a Time) setelah konversi akan diabaikan oleh .dt.year
df_cleaned['year_added'] = df_cleaned['date_added'].dt.year

# Menghapus baris yang mungkin memiliki NaT di 'date_added' setelah konversi
# (jika ada data tanggal yang sangat tidak valid sehingga tidak bisa diconvert)
df_cleaned.dropna(subset=['year_added'], inplace=True)

# Mengonversi 'year_added' ke integer yang dapat menangani nilai NA (Int64Dtype)
# atau langsung ke int biasa jika sudah yakin tidak ada NaN
df_cleaned['year_added'] = df_cleaned['year_added'].astype(int)

# Menghapus duplikat jika ada (opsional, tergantung dataset)
df_cleaned.drop_duplicates(inplace=True)


# Menyimpan dataset yang telah dibersihkan ke file CSV baru
df_cleaned.to_csv("cleaned_data_example.csv", index=False)
print("\nDataset telah dibersihkan dan disimpan sebagai 'cleaned_data_example.csv'.")
print("Informasi dataset setelah pembersihan:")
df_cleaned.info()
print("\n5 baris pertama dataset yang telah dibersihkan:")
print(df_cleaned.head())
```

c. Visualisasi Tren Penambahan Konten per Tahun

Visualisasi ini bertujuan untuk mengidentifikasi pola dan tren dalam penambahan konten Netflix dari tahun ke tahun menggunakan dataset yang telah dibersihkan.

```
# Menghitung jumlah konten yang ditambahkan setiap tahun dari dataset yang sudah bersih
content_per_year = df_cleaned['year_added'].value_counts().sort_index()

# Membuat plot garis untuk memvisualisasikan tren
plt.figure(figsize=(12, 7)) # Mengatur ukuran figure untuk tampilan yang lebih baik
sns.lineplot(x=content_per_year.index, y=content_per_year.values, marker='o', linewidth=2.5, color='#e50914') # Menggunakan warna merah Netflix
plt.title('Tren Penambahan Konten ke Netflix per Tahun (Data Contoh)', fontsize=16, fontweight='bold')
plt.xlabel('Tahun', fontsize=12)
plt.ylabel('Jumlah Konten', fontsize=12)
plt.grid(True, linestyle='--', alpha=0.7)
plt.xticks(rotation=45) # Memutar label sumbu x agar tidak tumpang tindih
plt.tight_layout() # Menyesuaikan tata letak agar semua elemen terlihat
plt.show()
```

6. Interpretasi Hasil

Visualisasi tren penambahan konten menunjukkan dinamika yang menarik pada data contoh Netflix ini. Terlihat adanya peningkatan signifikan dalam penambahan konten pada periode tertentu, khususnya antara tahun 2018 hingga 2020. Puncak ini kemungkinan besar mencerminkan strategi ekspansi konten Netflix yang agresif untuk menarik dan mempertahankan pelanggan. Namun, setelah periode tersebut, dapat diamati adanya penurunan dalam jumlah konten yang ditambahkan. Penurunan ini mungkin disebabkan oleh berbagai faktor eksternal, termasuk dampak pandemi COVID-19 yang memengaruhi produksi film dan serial, serta perubahan strategi investasi konten perusahaan. Data ini dapat menjadi masukan penting bagi tim strategi konten Netflix dalam pengambilan keputusan di masa mendatang. Penting untuk diingat bahwa interpretasi ini spesifik untuk dataset contoh Netflix; pada dataset lain, tren dan implikasinya bisa sangat berbeda.

7. Dimensi dan Metrik (Integrasi dengan Looker Studio)

Untuk analisis dan pelaporan lebih lanjut menggunakan platform seperti Looker Studio (sebelumnya Google Data Studio), berikut adalah dimensi dan metrik yang relevan dari dataset contoh ini:

Dimensi

- type: Kategori konten (misalnya, film, acara TV).

- country: Negara asal produksi konten.
- release_year: Tahun rilis konten asli.
- year_added: Tahun penambahan konten ke platform (hasil ekstraksi).
- rating: Peringkat usia konten.

Metrik

- count: Jumlah total konten (dapat digunakan untuk menghitung jumlah konten berdasarkan dimensi tertentu).
- duration: Durasi konten (perlu dibersihkan dan dikonversi ke format numerik jika akan digunakan sebagai metrik kuantitatif).

8. Panduan Platform Pendukung

Bagian ini menyediakan panduan singkat untuk platform-platform yang dapat mendukung proses analisis data ini, dari pengembangan kode hingga visualisasi interaktif.

8.1 Google collaboratory (collab)

Google collab adalah lingkungan pengembangan Python berbasis *cloud* yang dihosting oleh Google. collab memungkinkan peserta untuk menulis dan mengeksekusi kode Python langsung dari *browser*, tanpa memerlukan konfigurasi lingkungan lokal. collab sangat cocok untuk eksperimen cepat, pembelajaran, dan proyek data science yang tidak memerlukan *hardware* khusus.

Keunggulan collab:

- Akses Gratis ke GPU/TPU: Menyediakan akselerator perangkat keras untuk komputasi berat.
- Tidak Perlu Instalasi: Lingkungan Python sudah siap pakai.
- Integrasi Google Drive: Memudahkan penyimpanan dan akses *file*.
- Kolaborasi Real-time: Memungkinkan beberapa pengguna bekerja pada *notebook* yang sama.

Cara Mengakses dan Login:

1. Kunjungi portal web collab di: https://collab.research.google.com
2. Lakukan *login* menggunakan akun Google Anda.
3. Setelah login, Anda memiliki beberapa opsi:
    - Pilih "New Notebook" untuk membuat berkas .ipynb baru.
    - Gunakan opsi "Upload" untuk mengunggah berkas .ipynb dari komputer lokal Anda.

4. Pekerjaan Anda akan secara otomatis tersimpan di Google Drive, memudahkan pengelolaan dan kolaborasi.

Pengaturan Lingkungan:

- Runtime > Change runtime type: Pilih Python 3 sebagai *runtime*. Anda juga dapat memilih akselerator perangkat keras seperti GPU atau TPU jika analisis Anda membutuhkan daya komputasi tinggi (misalnya untuk *machine learning*).
- Mengunggah Berkas CSV: Untuk mengunggah dataset langsung ke sesi collab, gunakan kode berikut dalam sel kode:

```
from google.collab import file
uploaded = files.upload()
```

- Setelah menjalankan kode ini, jendela akan muncul untuk memilih dan mengunggah *file* dari komputer Anda.

Tips Manajemen Proyek:

- Simpan Pekerjaan Secara Berkala: Meskipun collab memiliki fitur simpan otomatis, disarankan untuk secara manual menyimpan perubahan penting melalui File > Save atau Ctrl + S.
- Eksekusi Sel: Gunakan Ctrl + Enter (atau Cmd + Enter di macOS) untuk mengeksekusi sel kode saat ini. Shift + Enter akan mengeksekusi sel dan pindah ke sel berikutnya.
- Dokumentasi Kode: Manfaatkan sel *Markdown* (# untuk judul, ''' ''' untuk blok teks) atau komentar dalam kode (#) untuk mendokumentasikan langkah-langkah dan interpretasi Anda.

8.2 Kaggle

Kaggle adalah platform komunitas daring terkemuka untuk ilmu data dan pembelajaran mesin. Kaggle tidak hanya menyediakan beragam dataset publik (termasuk dataset Netflix yang kita gunakan sebagai contoh), tetapi juga menawarkan lingkungan *notebook* berbasis *cloud* (Kaggle Kernels) untuk analisis data, serta menyelenggarakan kompetisi ilmu data global yang menantang.

Keunggulan Kaggle:

- Sumber Dataset Melimpah: Akses ke ribuan dataset yang diunggah oleh komunitas dan perusahaan.
- Lingkungan Komputasi Gratis: Menawarkan *notebook* dengan akses ke GPU/TPU.

- Komunitas Aktif: Forum diskusi, *kernel* yang dibagikan, dan *challenge* yang mengasah *skill*.
- Kompetisi Data Science: Kesempatan untuk menguji kemampuan dan memenangkan hadiah.

Cara Mengakses dan Login:

1. Buka situs web Kaggle di: https://www.kaggle.com
2. *Login* menggunakan akun Google Anda atau buat akun Kaggle baru.
3. Untuk mencari dataset publik, navigasikan ke tab "Datasets". Cari "Netflix" atau topik lain yang menarik untuk menemukan dataset yang relevan.
4. Untuk memulai analisis data langsung di Kaggle, pilih tab "Code" kemudian klik "New Notebook".

Pengaturan Lingkungan:

- Akselerator Perangkat Keras: Anda dapat memilih antara GPU atau TPU dari menu di pojok kanan atas *notebook* Kaggle.
- Instalasi Pustaka Tambahan: Gunakan sintaks %pip install nama_pustaka di dalam sel *notebook* untuk menginstal pustaka Python yang tidak tersedia secara *default*.

Tips Efisien:

- Mengunduh Dataset: Untuk mengunduh dataset netflix.csv dari Kaggle ke komputer Anda, cari datasetnya di Kaggle, lalu klik tombol "Download".
- Salinan Dataset ke Notebook Kaggle: Jika Anda bekerja langsung di Kaggle Notebook, dataset yang relevan biasanya sudah tersedia di direktori ../input/.
- Kaggle CLI: Untuk pengunduhan dataset yang lebih cepat ke lingkungan lokal atau collab, pertimbangkan untuk menginstal *Kaggle Command Line Interface (CLI)* dan gunakan perintah: kaggle datasets download [nama_pengguna]/[nama_dataset].

8.3 Looker Studio (Google Data Studio)

Looker Studio (sebelumnya dikenal sebagai Google Data Studio) adalah alat visualisasi data dan *dashboarding* berbasis *cloud* dari Google. Alat ini memungkinkan peserta untuk mengubah data mentah dari berbagai sumber menjadi laporan interaktif dan *dashboard* yang mudah dipahami, menjadikannya jembatan penting antara analisis data teknis dan presentasi bisnis.

Keunggulan Looker Studio:

- Antarmuka Drag-and-Drop: Sangat intuitif untuk membuat visualisasi tanpa perlu *coding*.
- Konektivitas Data Beragam: Terhubung ke puluhan sumber data, mulai dari Google Sheets, BigQuery, Google Analytics, hingga database SQL.

- Dashboard Interaktif: Filter dan kontrol yang dapat disesuaikan untuk eksplorasi data oleh pengguna.
- Kolaborasi Mudah: Berbagi laporan dengan tim atau klien.

Cara Mengakses dan Login:

1. Akses Looker Studio melalui: https://lookerstudio.google.com
2. *Login* menggunakan akun Google Anda.
3. Untuk memulai laporan baru, klik "Blank Report" atau pilih dari berbagai "Template" yang tersedia.

Pengaturan Sumber Data dan Visualisasi:

- Koneksi Sumber Data: Looker Studio mendukung koneksi ke berbagai sumber data. Untuk dataset yang telah dibersihkan (`cleaned_data_example.csv`), Anda dapat mengunggahnya ke Google Drive, lalu menghubungkannya sebagai sumber data "Google Sheets" atau "CSV upload" di Looker Studio.
- Pemilihan Dimensi dan Metrik: Pada panel sebelah kanan, Anda dapat memilih "Dimension" (variabel kategori yang digunakan untuk mengiris data, seperti `year_added` atau `type`) dan "Metric" (variabel kuantitatif yang diukur atau dihitung, seperti `count` atau `duration`) yang akan digunakan dalam visualisasi Anda.
- Kontrol Interaktif: Atur filter dan kontrol interaktif (seperti *dropdown* atau *slider*) untuk memungkinkan peserta *dashboard* menjelajahi data secara dinamis.

Tips Manajemen Laporan:

- Tema dan Tata Letak: Manfaatkan opsi "Theme and Layout" untuk menjaga konsistensi tampilan dan *branding* laporan Anda.
- Judul Visualisasi: Berikan judul yang jelas dan deskriptif untuk setiap visualisasi dan halaman laporan.
- Hindari Judul Ganda: Pastikan hanya satu judul yang aktif per komponen grafik atau per halaman untuk menjaga kejelasan.

9. Pentingnya Analisis Data bagi Dunia Industri

Analisis data telah menjadi tulang punggung bagi pengambilan keputusan strategis di berbagai sektor industri. Dalam era digital saat ini, di mana data dihasilkan dalam volume besar setiap detik, kemampuan untuk mengumpulkan, memproses, menganalisis, dan menginterpretasikan data adalah aset yang tak ternilai. Menguasai analisis data berarti memiliki kekuatan untuk mengubah data mentah menjadi wawasan yang dapat ditindaklanjuti, memberikan keunggulan kompetitif bagi organisasi.

Manfaat Utama Analisis Data dalam Industri:

1. Pengambilan Keputusan Berbasis Data: Industri dapat beralih dari intuisi atau spekulasi menjadi keputusan yang didukung oleh bukti empiris. Ini mengurangi risiko dan meningkatkan akurasi.
2. Pemahaman Pelanggan yang Lebih Mendalam: Dengan menganalisis data perilaku, preferensi, dan umpan balik pelanggan, perusahaan dapat memahami kebutuhan mereka, mempersonalisasi produk/layanan, dan meningkatkan kepuasan pelanggan. Contoh: Netflix menggunakan data tontonan untuk merekomendasikan konten dan mengembangkan serial orisinal.
3. Optimalisasi Proses Bisnis: Analisis data dapat mengidentifikasi *bottleneck*, inefisiensi, dan area yang perlu perbaikan dalam operasional perusahaan, mulai dari rantai pasok hingga manajemen inventaris.
4. Identifikasi Tren dan Prediksi Masa Depan: Dengan menganalisis data historis, perusahaan dapat mengidentifikasi tren pasar, memprediksi permintaan di masa depan, dan mengantisipasi perubahan kompetitif. Ini memungkinkan perencanaan yang lebih proaktif.
5. Pengembangan Produk dan Layanan Baru: Wawasan dari analisis data dapat mengungkapkan celah pasar, kebutuhan yang belum terpenuhi, atau peluang untuk inovasi produk/layanan yang lebih relevan.
6. Peningkatan Efisiensi Pemasaran: Data memungkinkan penargetan kampanye pemasaran yang lebih tepat sasaran, pengukuran efektivitas kampanye, dan pengoptimalan pengeluaran iklan.
7. Manajemen Risiko dan Deteksi Penipuan: Analisis data dapat digunakan untuk mendeteksi anomali dan pola yang mencurigakan, membantu perusahaan mengidentifikasi potensi risiko atau aktivitas penipuan.
8. Peningkatan Profitabilitas: Seluruh manfaat di atas pada akhirnya berkontribusi pada peningkatan profitabilitas melalui pengurangan biaya, peningkatan pendapatan, dan efisiensi operasional.

Dalam konteks contoh Netflix kita, analisis data tentang tren penambahan konten, preferensi *genre*, atau durasi tontonan adalah kunci untuk mempertahankan dominasi pasar, mengoptimalkan investasi konten, dan memastikan relevansi platform bagi jutaan pelanggannya di seluruh dunia. Oleh karena itu, kemampuan analisis data bukan lagi sekadar keahlian tambahan, melainkan kompetensi inti yang sangat dicari di berbagai industri.

10. Penutup

Modul ini telah memperkenalkan tahapan dasar analisis data menggunakan Python dan dataset Netflix sebagai contoh, dari pengunduhan dan pembersihan data awal di Google collab hingga interpretasi hasil visualisasi. Dengan memanfaatkan *tools open-source* seperti Python, Google collab, dan Kaggle, serta kemampuan visualisasi interaktif dari Looker Studio, proses analisis data dapat dilakukan secara efektif dan efisien, bahkan bagi peserta yang baru memulai di bidang ilmu data.

Peserta didorong untuk menjelajahi berbagai dataset lain yang tersedia di Kaggle atau sumber lainnya. Setiap dataset memiliki cerita uniknya sendiri dan menawarkan peluang baru untuk menerapkan teknik analisis data yang telah Anda pelajari di modul ini. Kunci untuk menjadi analis data yang mahir adalah dengan berlatih secara konsisten dan selalu ingin tahu tentang wawasan apa yang dapat diungkap dari data.

Dokumentasi:
DATAVERSE