Tugas 1: Exploratory Data Analysis

Student ID	Student name	Contribution description	Contribution (%)
2006614771	Suridian Pratama	coding, perumusan masalah, penulisan laporan, diskusi kelompok	100%
2106779472	Gabriella Aileen Mendrofa	coding, perumusan masalah, penulisan laporan, diskusi kelompok	100%
2106779516	Nadhilah Farhana	coding, perumusan masalah, penulisan laporan, diskusi kelompok	100%

Bagian 1. Pendahuluan

Dalam projek ini, kelompok kami mengangkat masalah penggunaan energi oleh beberapa gedung di suatu daerah. Data yang digunakan dalam projek ini diperoleh dari *kaggle.com* pada section competition ASHRAE - Great Energy Predictor III, yang meliputi tiga jenis dataset. Ketiga jenis dataset tersebut adalah building_metadata, train, dan weather_train. Berikut adalah deskripsi untuk masing-masing dataset.

1. **Dataset building_metadata**

   Ukuran (baris x kolom): 1449 x 6

   Kolom/Variabel:

   - site_id: foreign key untuk **weather_train**
   - building_id: foreign key untuk dataset **train**
   - primary_use: indikator untuk kategori primer aktivitas yang dilakukan di gedung tersebut berdasarkan *EnergyStar property type definitions*.
   - square_feet: luas lantai kotor bangunan
   - year_built: tahun gedung dibuka
   - floor_count: jumlah lantai bangunan

2. **Dataset train**

   Ukuran (baris x kolom): 20216100 x 4

   Kolom/Variabel:

- building_id: foreign key untuk dataset **building_metadata**
- meter: kode untuk jenis fasilitas yang dimiliki oleh suatu gedung (0: electricity, 1: chilled water, 2: steam, 3: hot water). Tidak semua gedung memiliki semua jenis fasilitas.
- timestamp: waktu pengukuran penggunaan energi dilakukan
- meter_reading: konsumsi energi (kWh, kecuali site 0 dalam satuan kBTU). Untuk mengubah kBTU menjadi kWh, maka perlu dikalikan dengan 0.2931.

3. **Dataset weather_train**

   Ukuran (baris x kolom): 139773 x 9

   Kolom/Variabel:

   - site_id
   - air_temperature: temperatur udara (°C)
   - cloud_coverage: bagian dari langit yang tertutup awan (okta)
   - dew_temperature: temperatur embun (°C)
   - precip_depth_1_hr (mm)
   - sea_level_pressure (millibar/hectopascal)
   - wind_direction: arah kompas (0 - 360)
   - wind_speed (meters/second)

Note: Dataset test tidak digunakan dalam projek ini karena ukurannya terlalu besar sehingga menyebabkan *crash* saat proses merging.

Rumusan masalah yang diangkat dalam projek ini adalah apakah terdapat tren penggunaan energi untuk masing-masing meter pada tahun 2016?

Bagian 2. Pre-processing

Sebelum melakukan tahap pre-processing data, pada penelitian ini dilakukan proses loading data langsung dari kaggle. Hal yang perlu dilakukan untuk menghubungkan google colab dengan kaggle adalah sebagai berikut:

1. Mendownload file API dari akun kaggle milik pribadi (kaggle.json).
2. Mengupload file tersebut ke dalam file repository google colab.
3. Melakukan mount drive atau izin akses penggunaan google drive (bersifat *temporary*).

4. Mendownload file yang diinginkan dari kaggle dengan cara memasukkan link repository data yang tersedia di kaggle.
5. Mengecek file apa saja yang telah terupload pada google colab (untuk memastikan data yang kita inginkan sudah tersimpan dalam file repository google colab).
6. Membuka zip file (file yang terdownload berbentuk zip).
7. Melakukan loading dataset sesuai dengan keperluan dari hasil extract zip sebelumnya.

Setelah berhasil melakukan loading dataset, tahap berikutnya adalah melakukan pre-processing data. Pada tahap pre-processing data, tahapan yang dilakukan adalah:

1. "mengintip" setiap dataset (building_metadata, train, weather_train) dengan menampilkan sebagian dari data untuk mengetahui variabel dari data dan tipe data yang sesuai untuk masing-masing variabel.
2. Mengecek tipe data dari setiap variabel dan mencatat variabel mana saja yang tipe datanya belum sesuai.
3. Mengecek persentase missing value setiap variabel dalam setiap dataset dan menghapus kolom/variabel dengan missing value yang relatif besar (>30%) karena kolom tersebut menjadi kurang informatif.
4. Mengubah tipe data dari variabel yang tipe datanya belum sesuai.
5. Mengecek foreign key dari masing-masing dataset (building_metadata, train, weather_train), dilanjutkan dengan proses merging data.
6. Menghapus dataset yang sudah tidak terpakai setelah merging (optional, untuk menghemat memori).
7. Mengecek korelasi antar variabel dalam data yang sudah dimerge dan menghapus kolom yang memiliki korelasi kuat (>= 0.75) dengan kolom yang lain, dengan catatan kolom tersebut memiliki missing value lebih banyak dari kolom yang lain (untuk memaksimalkan pengambilan informasi). Penghapusan kolom ini dilakukan untuk menghemat memori karena data dalam kolom tersebut telah diwalikan oleh kolom yang lain.
8. Mengecek adanya outlier dengan membandingkan nilai mean dan median dari setiap variabel. Apabila nilai mean dan median dari suatu variabel berbeda jauh, maka dapat diduga terdapat outlier dalam data.
9. Membuat boxplot untuk variabel yang diduga mengandung outlier. Boxplot dipilih karena dapat mendeteksi keberadaan outlier ditandai dengan adanya titik-titik di luar box.

10. Handling outlier menggunakan transformasi log. Cara ini dipilih karena kita tidak perlu melakukan penghapusan outlier, hanya mempersempit interval antardata. Hal ini baik, karena bisa saja outlier tersebut mengandung informasi yang penting (terutama apabila akan dilakukan pemodelan pada data).
11. Mengecek persentase missing value untuk setiap variabel pada dataset yang sudah dimerge.
12. Membuat plot variabel dengan missing value terhadap timestamp untuk mengecek adanya tren/seasonality (ditandai dengan naik-turunnya grafik atau adanya pola berulang).
13. Melakukan imputasi menggunakan metode interpolasi linier, karena terdapat tren dan seasonality pada variabel dengan missing value.

Bagian 3. Analisis Dasar Statistika

Setelah melakukan tahap loading dan pre-processing data, maka diperoleh informasi data tentang penggunaan energi electricity, chilled water, steam, hot water pada gedung education, lodging/residential, office, entertainment/public assembly, retail, parking, public services, warehouse/storage, food sales and services, religious worship, healthcare, utility, technology/science, manufacturing/industrial, dan other sebagai berikut:

1. Gedung mana sajakah yang paling banyak dan paling sedikit untuk pengeluaran pada semua energi ?

   Penjelasan: Untuk pengeluaran dari semua energi terdapat gedung education yang menggunakan energi paling banyak, kemudian untuk gedung religious worship yang menggunakan energi paling sedikit. Untuk urutan penggunaan energi secara lengkap bisa dilihat di laporan jupyter notebook.

2. Berapa banyak gedung untuk masing-masing sumber keluaran energi?

   Penjelasan: Pada pengeluaran energi electricity paling banyak digunakan yaitu sebanyak 12060910 gedung, untuk urutan kedua adalah energi chilled water sebanyak 4182440 gedung, untuk urutan ketiga adalah energi steam sebanyak 2708713 gedung, kemudian energi yang paling sedikit digunakan adalah hot water sebanyak 1264037 gedung.

3. Bagaimana hubungan pengeluaran masing-masing energi untuk setiap gedung?

Penjelasan: Pada pengeluaran energi electricity terdapat pada semua gedung, dengan pengeluaran yang paling banyak terdapat di gedung education dan office, sedangkan pengeluaran yang paling sedikit terdapat di gedung religious worship. Gedung yang tidak mengeluarkan energi chilled water terdapat pada gedung service dan warehouse/storage, dengan pengeluaran yang paling banyak terdapat di gedung education dan office, sedangkan pengeluaran yang paling sedikit terdapat di gedung manufacturing/industrial. Gedung yang tidak mengeluarkan energi steam terdapat pada gedung religious worship dan retail, dengan pengeluaran yang paling banyak terdapat di gedung education dan office, sedangkan pengeluaran yang paling sedikit terdapat di gedung technology/science. Pada pengeluaran energi hot water terdapat pada gedung education, office, entertainment/public assembly, lodging/residential, public services, healthcare, food sales and service, dan technology/science, dengan pengeluaran yang paling banyak terdapat di gedung education dan office, sedangkan pengeluaran yang paling sedikit terdapat di gedung technology/science

4. Bagaimana persentase masing-masing gedung pada pengeluaran setiap energi?

   Penjelasan: Persentase tertinggi pada pengeluaran energi electricity terdapat pada gedung warehouse/storage yaitu sebesar 92.15% dari jumlah gedung itu sendiri, dan persentase terendah terdapat pada gedung food sales and service yaitu sebesar 38.41% dari jumlah gedung itu sendiri. Persentase tertinggi pada pengeluaran energi chilled water terdapat pada gedung food sales and services yaitu sebesar 30.79% dari jumlah gedung itu sendiri, dan dan persentase terendah terdapat pada gedung parking yaitu sebesar 4.11% dari jumlah gedung itu sendiri. Persentase tertinggi pada pengeluaran energi steam terdapat pada gedung utility yaitu sebesar 26.59% dari jumlah gedung itu sendiri, dan dan persentase terendah terdapat pada gedung public services yaitu sebesar 5.04% dari jumlah gedung itu sendiri. Persentase tertinggi pada pengeluaran energi hot water terdapat pada gedung food sales and services yaitu sebesar 15.39% dari jumlah gedung itu sendiri, dan dan persentase terendah terdapat pada gedung lodging/residential yaitu sebesar 4.47% dari jumlah gedung itu sendiri.

5. Apakah terdapat hubungan antara musim dengan penggunaan masing-masing energi?

   Penjelasan: Jenis musim dan waktunya terjadinya:

   - Musim semi yang berlangsung pada bulan Maret sampai dengan bulan Mei.
   - Musim panas berlangsung pada bulan Juni sampai dengan bulan Agustus.
   - Musim gugur berlangsung pada bulan September sampai dengan November

- Musim dingin berlangsung pada bulan Desember sampai dengan Februari.

Berikut pola data yang didapat:

1. Pada penggunaan energi electricity paling melonjak terjadi pada dari bulan Juni dan puncak pemakaian tertinggi pada bulan Agustus, kemudian turun lagi pada bulan Oktober, dan pemakaian paling kecil pada bulan Maret sampai April.
2. Pada penggunaan energi chilled water bertahap naik dari bulan Februari dan puncak pemakaian tertinggi terjadi pada bulan September, kemudian turun lagi pada bulan November, dan pemakaian paling kecil pada Januari.
3. Pada penggunaan energi steam paling melonjak terjadi pada dari bulan Maret dan puncak pemakaian tertinggi pada bulan April, kemudian turun drastis untuk pemakaian paling sedikit terjadi pada bulan Juli sampai Oktober.
4. Pada penggunaan energi hot water paling melonjak terjadi pada dari bulan Desember dan puncak pemakaian tertinggi pada bulan Desember sampai Januari, kemudian turun drastis untuk pemakaian paling sedikit terjadi pada bulan Juni.

Pada keterangan diatas dapat diduga bahwa:

1. Penggunaan energi electricity sering digunakan pada musim panas, dan sedikit digunakan pada musim semi.
2. Penggunaan energi chilled water sering digunakan pada musim gugur, dan sedikit digunakan pada musim dingin
3. Penggunaan energi steam sering digunakan pada musim semi, dan sedikit digunakan pada musim panas sampai awal musim gugur
4. Penggunaan energi hot water sering digunakan pada musim dingin, dan sedikit digunakan pada musim panas.

Bagian 4. Penutup

Kesimpulan

Terdapat tren penggunaan energi untuk masing-masing jenis keluaran energi (electricity, chilled water, steam, dan hot water) selama satu tahun. Hal ini diduga disebabkan oleh adanya perubahan musim dan suhu udara.

a. https://colab.research.google.com/drive/1r7AbyDfsNMgftQJ59656z8R2N8svRh2t#scrollTo=19iI46XcsQlN (Akses langsung google colab)

b. https://drive.google.com/drive/folders/1UrQjhONbl7YMpsScbClT2lCD478ND2Sk
(Lampiran kaggle.json, code .ipynb, video presentasi, dan data hasil preprocessing)