Tugas 1: Exploratory Data Analysis

Student ID	Student name	Contribution description	Contribution (%)
2106724832	Justin Martinus	Mengerjakan proses pre-processing, Mengikuti diskusi via zoom, Mengerjakan laporan, dan Membuat video presentasi.	100%
2106725053	Kalisha Rahma Firza	Mengerjakan proses Insight, Mengikuti diskusi via zoom, Mengerjakan laporan, dan Membuat video presentasi.	100%
2106725034	Kamal Muftie Yafi	Mengerjakan proses pre-processing, Mengikuti diskusi via zoom, Mengerjakan laporan, dan Membuat video presentasi.	100%
2106653035	Laras Kirana Anindita	Mengerjakan proses Insight, Mengikuti diskusi via zoom, Mengerjakan laporan, dan Membuat video presentasi.	100%
2106652562	Raqi Akbar Robbani	Mengerjakan proses pre-processing, Mengikuti diskusi via zoom, Mengerjakan laporan, dan Membuat video presentasi.	100%

Bagian 1. Pendahuluan

Pada laporan ini, kami akan membahas mengenai preprocessing serta insight yang didapat dari dataset pada kompetisi di Kaggle bernama "ASHRAE - Great Energy Predictor III". Kompetisi tersebut menantang para peserta untuk membuat model dari empat jenis energi berdasarkan tingkat penggunaan historis dan cuaca yang diamati. Dataset mencakup pembacaan meter per jam selama tiga tahun dari lebih dari seribu bangunan di beberapa lokasi berbeda di seluruh dunia. Berikut adalah file-file yang diberikan beserta nama kolomnya dengan sedikit penjelasan:

- train.csv (20216100 x 4)
  - building_id (int64) - Kunci asing untuk metadata.
  - meter (int64) - Kode ID meter. Baca sebagai {0: electricity, 1: *chilledwater*, 2: steam, 3: hotwater}. Tidak setiap bangunan memiliki semua jenis meteran.
  - timestamp (object) - Kapan pengukuran dilakukan
  - meter_reading (float64) - Variabel target. Konsumsi energi dalam kWh (atau setara). Catat bahwa ini data asli dengan error pengukuran
- building_meta.csv (1449 x 6)
  - site_id (int64) - Kunci asing untuk data cuaca.
  - building_id (int64) - Kunci asing untuk training.csv
  - primary_use (object) - Indikator kategori utama aktivitas untuk bangunan berdasarkan definisi tipe properti EnergyStar
  - square_feet (int64) - Luas kotor lantai bangunan
  - year_built (float64) - Tahun bangunan tersebut dibuka
  - floor_count (float64) - Jumlah lantai pada bangunan
- weather_[train/test].csv (139773 x 9 dan 277243 x 9)
  - site_id (int64)
  - air_temperature (float64) - Derajat Celsius

    - cloud_coverage (float64) - Bagian dari langit yang tertutup awan, dalam oktas
    - dew_temperature (float64) - Derajat Celsius
    - precip_depth_1_hr (float64) - Millimeters
    - sea_level_pressure (float64) - Milibar/hektopaskal
    - wind_direction (float64) - Arah kompas (0-360)
    - wind_speed (float64) - Meter per detik
- test.csv (41697600 x 4)
    - row_id (int64) - ID baris untuk file submisi
    - building_id (int64) - Kode ID bangunan
    - meter (int64) - Kode ID meter
    - timestamp (object) - Timestamps untuk periode data pengujian
- sample_submission.csv (data ini tidak kami gunakan)

Bagian 2. Pre-processing

Sebelum melakukan preprocessing, kami akan melakukan **pengurangan memori** dengan mengambil referensi code di Kaggle. Pengurangan memori dilakukan agar pada program-program selanjutnya menjadi lebih ringan untuk dilakukan bagi Google Colab karena RAM yang diberikan terbatas.

Pertama-tama, kami akan melihat datanya terlebih dahulu. Dengan melihat dataset dan memahami kolom dari setiap variabel, kami mengetahui apa saja isi dari tiap dataset dan apa saja yang harus diolah. Kami juga mengganti variabel pada kolom ‘meter’ sesuai dengan informasi pada dataset.

Kemudian, pada data ‘building_meta’ didapat kolom ‘building_id’ & ‘site_id’, sehingga dapat dilakukan penggabungan dengan data lainnya. Dataset ‘train’ digabungkan dengan dataset ‘weather_train’, dan dataset ‘test’ dengan dataset ‘weather_test‘. Akhirnya, didapat dua tabel besar yang kemudian akan dilakukan proses preprocessing. Variabel yang mengandung data-data yang sudah tidak terpakai setelah penggabungan kemudian dihapus untuk menghemat memori.

Lalu, kami akan coba untuk mengurai waktu pada variabel ‘timestamp’ agar dapat dilakukan analisis terhadap waktu pada proses pencarian insight.

Selanjutnya, kami akan melakukan identifikasi outlier. Pertama, kami akan coba untuk mencoba visualisasi dataset ‘train’ yang sudah digabung. Terlihat bahwa terdapat satu bar pada ‘site_id’ yang memiliki nilai sangat tinggi dibandingkan bar yang lain. Kemudian, kami perhatikan lebih dalam menggunakan visualisasi kembali dan menemukan bahwa ternyata outlier tersebut terdapat pada ‘building_id’ 1099. Akhirnya, kami drop kolom tersebut sehingga didapatkan distribusi data yang lebih baik.

Terakhir, kami akan menangani missing value pada data. Pertama, kami cari terlebih dahulu berapa persentase data yang hilang dari tiap kolom. Didapatkan bahwa terdapat persentase yang sangat besar pada variabel ‘year_built’ dan ‘floor_count’, yakni masing-masing sekitar 59.98% dan 82.65%. Akhirnya, kami drop kolom tersebut karena terlalu banyak missing valuenya sehingga tidak cukup untuk mendapatkan insight yang baik dan data yang diberikan tidak signifikan untuk arah insight yang akan kami cari dan dapatkan.

Pada missing values lainnya, akan dilakukan proses pengisian dengan mempertimbangkan variabelnya. Ada yang diisi dengan mean dan ada juga yang diisi dengan median.

Bagian 3. Analisis Dasar Statistika

[Tuliskan analisis apa saja yang dilakukan, kenapa/hasil apa yang diharapkan (dan yang diperoleh) dari melakukan analisis tersebut (**jika sudah dituliskan di Jupyter Notebook/Google Colab tidak perlu dituliskan lagi di sini**. Apakah ada kesinambungan antara hasil yang diperoleh? Apakah ada konstruksi cerita yang "utuh" dari berbagai proses yang dilakukan dan hasil yang diperoleh? Informasi bermakna apa saja yang dihasilkan dari proses ini?**]**
Dilakukan analisis sebagai berikut:

1. Dilakukan visualisasi rata-rata penggunaan energi per-jam, per-hari, dan per-bulan.
2. Dilakukan visualisasi heatmap. Jika angka yang muncul mendekati angka 1, maka korelasi antar variabel bisa dibilang KUAT. Sebaliknya, jika angka yang didapat menjauhi angka 1, maka korelasi antar variabelnya LEMAH.
3. Mencari nilai maksimum dan minimum dari 'meter_reading' dengan tujuan untuk bisa menentukan dan membandingkan penggunaan energi paling banyak dan paling sedikit untuk setiap variabel.
    - Langkah selanjutnya, dilihat berdasarkan nilai maksimum 'meter_reading' untuk 'primary_use' dan membandingkannya dengan nilai maksimum 'primary_use' untuk mengetahui korelasi antara keduanya.
    - Selanjutnya, dilihat berdasarkan nilai maksimum 'meter_reading' untuk 'square_feet' dan membandingkannya dengan nilai maksimum 'square_feet' untuk mengetahui korelasi antara keduanya.
    - Untuk langkah selanjutnya, dilihat berdasarkan nilai minimum 'meter_reading' untuk 'meter' dan membandingkannya dengan nilai minimum 'meter' untuk mengetahui korelasinya antara kedua variabel tersebut.
4. Membuat visualisasi untuk 'square_feet' dan visualisasi untuk menggambarkan korelasi antara 'meter_reading' terhadap 'primary_use', dan 'meter'.

Bagian 4. Penutup
Link Google Drive:
https://drive.google.com/drive/folders/1hPe6Whf_fq5xeo0ZfGCoqbzcoJuWa1yV?usp=sharing
Kesimpulan :

1. Berdasarkan visualisasi untuk rata-rata penggunaan energi diperoleh:
    - Per-jam: bahwa penggunaan energi pada interval antara pukul 10 hingga 15 meningkat. Ini menunjukkan aktivitas yang menggunakan banyak energi meningkat di siang hari.
    - Per-hari: bahwa aktivitas yang menggunakan banyak energi meningkat di pertengahan bulan, yaitu sekitar tanggal 10 sampai 15 dan 16 sampai 20 .
    - Per-bulan: bahwa penggunaan energi pada interval bulan ke 5 sampai bulan ke 7 meningkat. Ini menunjukkan aktivitas yang menggunakan banyak energi meningkat di pertengahan tahun.
2. Berdasarkan hasil program, pada tabel Statistika Deskriptif diperoleh energi maksimum adalah 880374 dan energi minimumnya adalah 0.
    - Primary_use: Sektor yang paling banyak menggunakan energi adalah *Entertainment,* sementara sektor yang paling banyak menghabiskan untuk penggunaan gedung adalah *Education*. Jadi dapat ditunjukkan bahwa sektor paling banyak digunakan tidak menjamin bahwa penggunaan energinya paling besar.
    - Square_Feet: Diperoleh luas bangunan gedung yang menghabiskan energi terbanyak adalah 108339. Namun, luas bangunan terbesar adalah 875000. Hal ini menunjukkan bahwa luas bangunan gedung terbesar tidak menjamin penggunaan energi paling besar.

- Meter_reading: Diperoleh energi terbesar pada saat menggunakan *Chilled Water* sedangkan jenis meter reading yang paling banyak digunakan adalah Electricity. Jadi dapat ditunjukkan bahwa jenis meter reading yang terbanyak digunakan tidak menjamin penggunaan energi terbesar.

3. Penggunaan visualisasi diperuntuk melihat data secara keseluruhan, sehingga kesimpulan yang diperoleh sedikit berbeda dengan kesimpulan tabel.
    - Primary_use: Dari hasil visualisasi, diperoleh secara keseluruhan sektor *Healthcare* merupakan pengguna energi terbanyak. Disusul yang kedua adalah *Utility.*
    - Meter_Reading: Melalui visualisasi secara menyeluruh, diperoleh jenis meter reading yang mengonsumsi energi terbanyak adalah *steam* sesuai dengan yang paling banyak digunakan.
    - Square_feet: Tidak digunakan visualisasi secara menyeluruh untuk ‘square_feet’. Visualisasi hanya dilakukan pada tabel saja dan dapat diinterpretasikan sebagai luasnya paling banyak untuk yang kurang dari 200.000.