TUGAS 1 EXPLORATORY DATA ANALYSIS

Student ID	Student Name	Contribution Description	Contribution (%)
2106724883	Adawia Ananda	Melakukan Analisis dan prepocesing	100%
2106706464	Fernaldy	Melakukan Analisis dan prepocesing	100%
2106700776	Nadia Sukesi Sianipar	Melakukan Analisis dan prepocesing	100%
2106726812	Najwa Salsabila Hakim	Mengerjakan Laporan, membantu diskusi preprocessing	100%
2106726844	Myra Azzahra Putri Syah I.	Mengerjakan PPT, membantu diskusi preprocessing	100%
2105700946	Whitney	Mengerjakan PPT, membantu diskusi preprocessing	100%

Laporan untuk Memenuhi Tugas

Mata Kuliah Eksplorasi dan Visualisasi Data

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS INDONESIA

BAB I

PENDAHULUAN

Energi sangat penting peranannya dalam perekonomian dunia, baik sebagai bahan bakar untuk proses industrialisasi, sebagai bahan baku untuk proses produksi, dan sebagai komoditas ekspor. Sumber energi yang digunakan untuk keperluan domestik meliputi energi fosil (minyak bumi, gas bumi, dan batubara) serta energi terbarukan (tenaga air dan tenaga panas bumi). Secara umum peningkatan kebutuhan energi memiliki keterkaitan yang erat dengan semakin berkembangnya kegiatan ekonomi. Peningkatan penggunaan energi di sektor industri dalam 15 tahun terakhir bukan hanya terjadi karena proses transformasi struktural yang cepat dari sektor pertanian ke sektor industri. Namun hal ini terjadi juga karena ada dugaan **pemborosan penggunaan energi di sektor industri**. Intensitas energi merupakan salah satu ukuran yang sering digunakan untuk melihat tingkat efisiensi energi dalam suatu sektor, dalam laporan ini kami akan membahas tema yaitu "Permasalahan Prediksi Penggunaan Energi Gedung dari PT Ashrae - American Society of Heating, Refrigerating and Air-Conditioning Engineers"

- Kami mendapatkan sumber data dari tautan berikut
  https://www.ashrae.org/about,
  https://www.kaggle.com/c/ashrae-energy-prediction/data

- Jumlah pengukuran variabel awal sebanyak 16

- Tipe data yang dipakai pada proyek kali ini diantaranya :
  1. building_id=int
  2. meter=int
  3. timestamp=object
  4. meter_reading=float64
  5. site_id=int
  6. primary_use=object
  7. square_feet=int
  8. year_built=float
  9. floor_count=float
  10. air_temperature=float
  11. cloud_coverage=float
  12. dew_temperature=float
  13. precip_depth_1_hr=float
  14. sea_level_pressure=float
  15. wind_direction=float
  16. wind_speed=float

- Arti/maksud dari pengukuran-pengukuran tersebut
    1. site_id & building_id: Id lokasi dan gedung
    2. primary_use: Peruntukan Gedung
    3. square_feet: Luas bangunan gedung
    4. year_built: Tahun pembuatan gedung
    5. floor_count: Banyaknya lantai yang ada di gedung.
    6. meter : Jenis meter reading penggunaan energi gedung.
    7. timestamp : Waktu saat pengukuran (per-jam)
    8. meter_reading: Penggunaan energy.
    9. air_temperature: Suhu udara
    10. cloud_coverage: Ukuran berawan
    11. dew_temperature: Suhu embun
    12. precip_depth_1_hr: precipitation (banyaknya air dari langit, karena sebab apapun)
    13. sea_level_pressure: Tekanan permukaan laut.
    14. wind_direction & wind_speed: arah dan kecepatan angin

BAB II PREPROCESSING

Sebelum melakukan preprocessing lebih lanjut ada beberapa step yang harus dilakukan yaitu :

1. *Import* modul
   Pada bagian ini akan di *import* modul-modul yang digunakan antara lain :
   - gc
   - matplotlib.pyplot
   - pandas
   - numpy
   - seaborn
2. *Input* data yang akan digunakan yaitu *input* 'kaggle.json'
3. Pendefinisian tabel dengan membaca tabel csv
   Terdapat 5 tabel yang didefinisikan yaitu :
   - building_metadata
   - train
   - test
   - weather_train
   - weather_test
4. *Memory Reduction*
   Berdasarkan referensi kode dari
   https://www.kaggle.com/kernels/scriptcontent/3684066/download
   akan dilakukan pengurangan memori untuk *reduce memory* tabel saat *preprocessing* dilakukan.

Pre-processing terbagi menjadi 3 bagian yaitu :

A. Data Integration

Pada step ini, pada tabel building_meta akan ditemukan variabel 'building_id' & 'site_id', sehingga tabel tabel ini bisa digabung/*join* dengan 'train'/'test' dan 'weather train'/'weather test'. Dilakukan *join* pada tabel-tabel ini, sehingga didapatkan 2 tabel yang lebih besar untuk 'train' dan 'test'.

B. Cleaning Data

Data yang akan difokuskan pada step ini adalah tabel 'train'.

Proses ini memiliki beberapa tahap yaitu:

1. Mengoreksi Tipe Variabel

Tahap ini untuk melihat apakah tipe variabel sudah sesuai. Didapatkan hasil bahwa tipe data sudah sesuai.

2. Menelusuri Statistika Deskriptif

Tahap ini untuk melihat hasil statistika deskriptif seperti mean, median, modus, standar deviasi, nilai max, nilai min, dll.

3. *Fixing Variable Type*

Memangkas kolom menjadi kolom yang kita butuhkan saja untuk dilakukan analisis lebih lanjut.

4. Removing Duplicate Data

Pada tahap ini data yang duplikat akan di-*remove* karena tidak dibutuhkan pada analisis. Data duplikat membuat data kurang akurat. Terdapat 24589426 data duplikat pada tabel hasil join untuk tabel train lalu di-*remove*.

5. Menentukan *Outlier*

Ditampilkan beberapa visualisasi untuk melihat apakah terdapat *outlier* atau tidak pada data. Selain itu, akan dihitung nilai maksimum, minimum, dan mean untuk melihat apakah ada outlier dari perbandingan nilai maksimum dan minimum terhadap mean. Ternyata, terlihat terdapat outlier pada 'square_feet' dan 'meter_reading' karena nilai minimum dan maksimum sangat jauh dari nilai rata-rata.

6. Remove *Outlier* ke variabel baru

Setelah *outlier* ditemukan, pada tahap ini *outlier* tidak akan langsung dimusnahkan tetapi di*remove* atau dipindahkan ke variabel baru.

7. Menangani *Missing Value*

Missing value akan dideteksi, lalu dibuat keputusan apakah *missing value* akan diimputasi dengan mean, modus, atau median atau justru membiarkan *missing value* tersebut.

Pada tahap ini, walaupun *missing values* yang terdapat dalam variabel-variabel tersebut lumayan banyak, kelompok kami memutuskan untuk tidak *drop* variable tersebut karena penting untuk analisis kami. Selain itu, kami juga memutuskan untuk tidak melakukan imputasi karena persentase *missing values* yang cukup besar (di atas 10%). Dikhawatirkan apabila dilakukan imputasi, data yang dihasilkan tidak akurat

C. Saving (preprocessed) Data

Pada tahap ini data hasil *preprocessing* akan disimpan menjadi bentuk csv kembali untuk analisis yang ingin dilakukan kemudian hari**.**

Visualisasi Data

Beberapa visualisasi data yang dihasilkan beserta

1. Dari visualisasi boxplot dan countplot, dapat dilihat bahwa sektor *Healthcare* dan *Utility* menggunakan energi terbanyak, sementara sektor *Religious Worship* menggunakan energi paling sedikit.
2. Dari visualisasi square_feet, dapat dilihat bahwa mayoritas gedung memiliki luas kurang dari 200000 kaki.
3. Dari visualisasi year_built, dapat dilihat bahwa bangunan paling banyak dibangun pada *range* tahun 1960-1970.
4. Dari visualisasi air_temperature, dapat dilihat bahwa mayoritas gedung memiliki suhu udara 0 sampai 30 derajat celsius.
5. Dari visualisasi floor_count, dapat dilihat bahwa mayoritas gedung memiliki jumlah lantai 1 sampai 5
6. Dari visualisasi heatmap, dapat dilihat bahwa korelasi antara energi yang digunakan (meter_reading) dengan variabel-variabel lainnya cenderung rendah, yaitu $-0.2<r<0.2$

Keterkatikan antar proses pre-processing

Menurut kelompok kami, setiap langkah pada *preprocessing* saling terkait dan sangat membantu langkah lain untuk menghasilkan data yang lebih terstruktur untuk dianalisis.

Permasalahan yang dialami saat pre-processing

- Google Colab sering *crash* akibat ukuran data yang diproses sangat besar. Namun, kelompok kami memiliki solusi yaitu dengan menggunakan Google Colab dengan RAM 25 GB.
- Data ‘train’ yang sangat besar membutuhkan waktu yang sangat lama untuk ter-*upload* secara sempurna apabila menggunakan metode *upload* berkas manual. Namun, kelompok kami memiliki solusi yaitu dengan meng-*input* langsung data yang digunakan melalui Kaggle.

BAB III
ANALISIS DASAR STATISTIKA

Analisis yang akan dilakukan yaitu:

1. Mencari nilai minimum ‘year_built’ untuk mengetahui bangunan dengan ‘building_id’ mana dan ‘site_id’ mana yang paling lama berdiri dan menghabiskan energi dalam jangka waktu yang lama. Selanjutnya, mencari ‘meter_reading’ dari ‘building_id’ yang didapatkan dan melihat ‘primary_use’ untuk mengetahui kegunaan dari penggunaan energi yang banyak di PT tersebut. Analisis ini digunakan untuk mengetahui korelasi waktu dengan besar penggunaan energi.
2. Mencari nilai maksimum ‘square_feet’ untuk mengetahui bangunan dengan ‘building_id’ mana dan ‘site_id’ mana yang paling luas. Selanjutnya, mencari ‘meter_reading’ dari ‘building_id’ yang didapatkan dan melihat ‘primary_use’ untuk mengetahui kegunaan dari penggunaan energi yang banyak di PT tersebut. Analisis ini digunakan untuk mengetahui korelasi luas bangunan dengan besar penggunaan energi.
3. Mencari nilai maksimum ‘floor_count’ untuk mengetahui bangunan dengan ‘building_id’ mana dan ‘site_id’ mana yang paling banyak jumlah lantai nya. Selanjutnya mencari ‘meter_reading’ dari ‘building_id’ yang didapatkan dan melihat ‘primary_use’ untuk mengetahui kegunaan dari penggunaan energi yang banyak di PT tersebut. Analisis ini digunakan untuk mengetahui korelasi jumlah lantai dengan besar penggunaan energi.
4. Mencari nilai minimum dan maksimum ‘air_temperature’ untuk mengetahui bangunan dengan ‘building_id’ mana dan ‘site_id’ mana yang memiliki suhu udara terendah dan tertinggi. Selanjutnya mencari ‘meter_reading’ dari ‘building_id’ yang didapatkan dan melihat ‘primary_use’ untuk mengetahui kegunaan dari penggunaan energi yang banyak di PT tersebut. Analisis ini digunakan untuk mengetahui korelasi suhu udara dengan besar penggunaan energi.
5. Mencari nilai maksimum ‘meter_reading’ untuk mengetahui bangunan dengan ‘building_id’ mana dan ‘site_id’ mana yang sebenarnya (pada faktanya) menghabiskan energi paling banyak.
6. Membandingkan semua hasil pada poin 1,2,3,4 terhadap poin 5.
7. Tarik kesimpulan.

Hasil yang diperoleh yaitu :

1. Dari hasil tabel penggunaan energi maksimum yang sebenarnya, terlihat bahwa pada faktanya gedung yang menghabiskan energi paling banyak sebesar **2190470** adalah gedung dengan '**building_id' 1099 dan 'site_id' 13**. Hasil ini berbeda dengan gedung yang menghabiskan energi terbanyak berdasarkan 'year_built' yaitu '**building_id' 124 dan 'site_id' 1**. Hal ini menunjukkan bangunan paling lama tidak menjamin bangunan itu akan menggunakan energi paling besar. **Tidak ada korelasi penggunaan energi terhadap waktu secara signifikan**. Namun dapat dilihat, 'primary_use' kedua hasil itu sama yaitu sama-sama digunakan untuk keperluan pendidikan.
2. Dari hasil tabel penggunaan energi maksimum yang sebenarnya terlihat bahwa pada faktanya gedung yang menghabiskan energi paling banyak sebesar **2190470** yaitu gedung dengan '**building_id' 1099 dan 'site_id' 13**. Hasil ini berbeda dengan gedung yang menghabiskan energi terbanyak berdasarkan 'square_feet' yaitu '**building_id' 869 dan 'site_id' 8**. Hal ini menunjukkan bangunan dengan luas paling besar tidak menjamin bangunan itu akan menggunakan energi paling besar. **Tidak ada korelasi penggunaan energi terhadap luas bangunan secara signifikan**. Selain itu dapat dilihat, 'primary_use' kedua hasil itu berbeda, dimana salah satu gedung digunakan untuk keperluan pendidikan, sementara yang lain digunakan untuk keperluan *entertainment/public assembly*
3. Dari hasil tabel penggunaan energi maksimum yang sebenarnya, terlihat bahwa pada faktanya gedung yang menghabiskan energi paling banyak sebesar **2190470** yaitu gedung dengan '**building_id' 1099 dan 'site_id' 13** .Hasil ini berbeda dengan gedung yang menghabiskan energi terbanyak berdasarkan 'floor_count' yaitu '**building_id' 799 dan 'site_id' 7**. Hal ini menunjukkan bangunan dengan jumlah lantai paling banyak tidak menjamin bangunan itu akan menggunakan energi paling besar. **Tidak ada korelasi penggunaan energi terhadap jumlah lantai secara signifikan**. Namun dapat dilihat, 'primary_use' kedua hasil itu sama, yaitu sama-sama digunakan untuk keperluan pendidikan.
4. Dari hasil tabel penggunaan energi maksimum yang sebenarnya, terlihat bahwa pada faktanya gedung yang menghabiskan energi paling banyak sebesar **2190470** yaitu gedung dengan '**building_id' 1099 dan 'site_id' 13** . Hasil ini berbeda dengan gedung yang menghabiskan energi terbanyak berdasarkan suhu maksimum yaitu '**building_id' 156 dan site_id 2** dan suhu minimum yaitu '**building_id' 1069 dan 'site_id' 13**. Hal ini menunjukkan suhu udara paling maksimum ataupun minimum tidak membuat suatu bangunan akan menggunakan energi paling besar. **Tidak ada**

korelasi terhadap penggunaan energi suhu udara secara signifikan. Selain itu dapat dilihat, ‘primary_use’ kedua hasil itu berbeda yaitu satu digunakan untuk keperluan pendidikan dan dua nya lagi untuk keperluan *public services* dan *office*.

BAB 4
PENUTUP

Kesimpulan yang dapat ditarik :

Tidak ada korelasi yang cukup kuat antara 'meter_reading' (banyak penggunaan energi) dengan variabel-variabel lainnya yang kelompok kami uji pada kesempatan kali ini, yaitu 'year_built' (tahun berdirinya gedung), 'square_feet' (luas bangunan), 'floor_count' (jumlah lantai gedung), dan 'air_temperature' (suhu udara gedung). Dari visualisasi boxplot, dapat dilihat bahwa korelasi antara energi yang digunakan (meter_reading) dengan variabel-variabel lainnya cenderung rendah, yaitu -0.2<r<0.2

Link Google Drive :

https://drive.google.com/drive/folders/1_QCtHWO36kldERFPuz52fjsLGHSLEp6w?usp=sharing

DAFTAR PUSTAKA

Pambudi, Harry Gustara .2009. “Analisis faktor-faktor yang mempengaruhi intensitas energi industri menengah-besar indonesia” diakses pada https://repository.ipb.ac.id/handle/123456789/12744 pada 3-Maret-22 pukul 23.00

https://www.kaggle.com/kernels/scriptcontent/3684066/download pada 5-April-22 pukul 22.38