LAPORAN

TUGAS BESAR II ANALISIS DATA

KELOMPOK 12

1. Rahmat Al Fajri - 16520171
2. Rava Naufal Attar - 16520411
3. Ray Clement - 16520371
4. Rio Alexander - 16520181

KU1102

PENGENALAN KOMPUTASI

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA

INSTITUT TEKNOLOGI BANDUNG

DESEMBER 2020

DAFTAR ISI

BAB I

DESKRIPSI DAN ANALISIS DATA

1.1. Deskripsi Data

Data tersebut berisi mengenai besar konsumsi listrik di negara-negara bagian di India dari Januari 2019 hingga Mei 2020, tetapi yang dianalisis dalam laporan ini hanya data pada tahun 2019.

a. Informasi yang dapat diambil dari data tersebut adalah seberapa besar konsumsi listrik di berbagai negara bagian di India.
b. Format data tersebut berupa csv, diambil dari website keggle (https://www.kaggle.com/twinkle0705/state-wise-power-consumption-in-india). Data yang digunakan yaitu file csv yang bernama "long_data_.csv".
c. Dimensi awal dari data tersebut adalah enam kolom dan 16599 baris dengan ukuran file 964 kb lalu setelah di cleansing menjadi 8 kolom 11848 baris dengan ukuran file 760 kb.
d. Pembacaan data dilakukan dengan kode:

```
import pandas as pd
df = pd.read_csv('long_data_.csv')
```

1.2. Karakteristik Data

Dalam data tersebut terdapat delapan atribut sebagai berikut:

a. States
   Atribut States berisi nama negara bagian dan merupakan atribut kategorikal-nominal.
b. Regions
   Atribut Regions berisi zona tempat negara bagian di India dan merupakan atribut kategorikal-nominal.
c. latitude
   Atribut latitude berisi koordinat garis lintang kota tersebut dan merupakan atribut kuantitatif-kontinu.
d. longitude
   Atribut longitude berisi koordinat garis bujur kota tersebut dan merupakan atribut kuantitatif-kontinu.
e. Dates
   Atribut Dates berisi tanggal diambilnya data dan merupakan atribut kuantitatif.
f. Month
   Atribut Month berisi bulan diambilnya data dan merupakan kategorikal-nominal.
g. Year
   Atribut Year berisi tahun diambilnya data dan merupakan atribut kuantitatif-diskrit.
h. Usage
   Atribut Usage berisi besar konsumsi listrik pada tanggal dan state tersebut dalam satuan Mega Watt dan merupakan atribut kuantitatif-kontinu.

Data pada setiap atribut memiliki range atau nilai sebagai berikut:

a. States : Punjab, Haryana, Rajasthan, Delhi, UP, Uttarakhand, HP, J&K, Chandigarh, Chhattisgarh, Gujarat, MP, Maharashtra, Goa, DNH, Andhra Pradesh, Telangana, Karnataka, Kerala, Tamil Nadu, Pondy, Bihar, Jharkhand, Odisha, West Bengal, Sikkim, Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram,

			Nagaland, Tripura
b.	Regions	:	NR, WR, SR, ER, NER
c.	Latitude	:	8.900372741 – 33.45
d.	Longitude	:	71.1924 – 94.21666744
e.	Dates	:	2 Januari 2019 – 23 Mei 2020
f.	Month	:	Januari – Desember
g.	Year	:	2019
h.	Usage	:	0.5 – 522.1

BAB II

PENGOLAHAN DATA

2.1. Statistik

2.1.1. Sampel Data

a. Banyak baris data

```
# Banyaknya data
print('Banyaknya data adalah ' + str(len(df)))
```

```
Banyaknya data adalah 11847
```

b. Sampel 33 baris pertama

```
# Sampel data
df[:33]
```

	States	Regions	latitude	longitude	Dates	Month	Year	Usage
0	Punjab	NR	31.519974	75.980003	02/01/2019	Jan	2019	119.9
1	Haryana	NR	28.450006	77.019991	02/01/2019	Jan	2019	130.3
2	Rajasthan	NR	26.449999	74.639981	02/01/2019	Jan	2019	234.1
3	Delhi	NR	28.669993	77.230004	02/01/2019	Jan	2019	85.8
4	UP	NR	27.599981	78.050006	02/01/2019	Jan	2019	313.9
5	Uttarakhand	NR	30.320409	78.050006	02/01/2019	Jan	2019	40.7
6	HP	NR	31.100025	77.166597	02/01/2019	Jan	2019	30.0
7	J&K	NR	33.450000	76.240000	02/01/2019	Jan	2019	52.5
8	Chandigarh	NR	30.719997	76.780006	02/01/2019	Jan	2019	5.0
9	Chhattisgarh	WR	22.090420	82.159987	02/01/2019	Jan	2019	78.7
10	Gujarat	WR	22.258700	71.192400	02/01/2019	Jan	2019	319.5
11	MP	WR	21.300391	76.130019	02/01/2019	Jan	2019	253.0
12	Maharashtra	WR	19.250232	73.160175	02/01/2019	Jan	2019	428.6
13	Goa	WR	15.491997	73.818001	02/01/2019	Jan	2019	12.8
14	DNH	WR	20.266578	73.016618	02/01/2019	Jan	2019	18.6
15	Andhra Pradesh	SR	14.750429	78.570026	02/01/2019	Jan	2019	164.6
16	Telangana	SR	18.112400	79.019300	02/01/2019	Jan	2019	204.2
17	Karnataka	SR	12.570381	76.919997	02/01/2019	Jan	2019	206.3
18	Kerala	SR	8.900373	76.569993	02/01/2019	Jan	2019	72.7
19	Tamil Nadu	SR	12.920386	79.150042	02/01/2019	Jan	2019	268.3
20	Pondy	SR	11.934994	79.830000	02/01/2019	Jan	2019	6.3
21	Bihar	ER	25.785414	87.479973	02/01/2019	Jan	2019	82.3
22	Jharkhand	ER	23.800393	86.419986	02/01/2019	Jan	2019	24.8
23	Odisha	ER	19.820430	85.900017	02/01/2019	Jan	2019	70.2
24	West Bengal	ER	22.580390	88.329947	02/01/2019	Jan	2019	108.2
25	Sikkim	ER	27.333330	88.616647	02/01/2019	Jan	2019	2.0
26	Arunachal Pradesh	NER	27.100399	93.616601	02/01/2019	Jan	2019	2.1
27	Assam	NER	26.749981	94.216667	02/01/2019	Jan	2019	21.7
28	Manipur	NER	24.799971	93.950017	02/01/2019	Jan	2019	2.7
29	Meghalaya	NER	25.570492	91.880014	02/01/2019	Jan	2019	6.1
30	Mizoram	NER	23.710399	92.720015	02/01/2019	Jan	2019	1.9
31	Nagaland	NER	25.666998	94.116570	02/01/2019	Jan	2019	2.2
32	Tripura	NER	23.835404	91.279999	02/01/2019	Jan	2019	3.4

c. Data usage dari yang terkecil

```
# Sortir data berdasarkan usage
df_u = df.sort_values(['Usage'])
df_u
```

	States	Regions	latitude	longitude	Dates	Month	Year	Usage
9793	Sikkim	ER	27.333330	88.616647	28/10/2019	Oct	2019	0.5
8341	Sikkim	ER	27.333330	88.616647	13/09/2019	Sep	2019	0.5
4381	Sikkim	ER	27.333330	88.616647	18/05/2019	May	2019	0.5
2302	Sikkim	ER	27.333330	88.616647	14/03/2019	Mar	2019	0.6
5998	Sikkim	ER	27.333330	88.616647	07/07/2019	Jul	2019	0.6
...	...	...	...	...	...	...	...	...
5358	Maharashtra	WR	19.250232	73.160175	18/06/2019	Jun	2019	513.6
5391	Maharashtra	WR	19.250232	73.160175	19/06/2019	Jun	2019	513.9
1695	Maharashtra	WR	19.250232	73.160175	23/02/2019	Feb	2019	515.8
1662	Maharashtra	WR	19.250232	73.160175	22/02/2019	Feb	2019	516.4
5424	Maharashtra	WR	19.250232	73.160175	20/06/2019	Jun	2019	522.1

11847 rows × 8 columns

d. Nilai usage terkecil

```
# Mencari usage terkecil
u_min = df_u[:1]
u_min
```

	States	Regions	latitude	longitude	Dates	Month	Year	Usage
9793	Sikkim	ER	27.33333	88.616647	28/10/2019	Oct	2019	0.5

e. Nilai usage terbesar

```
# Mencari usage terbesar
u_max = df_u[11846:11847]
u_max
```

	States	Regions	latitude	longitude	Dates	Month	Year	Usage
5424	Maharashtra	WR	19.250232	73.160175	20/06/2019	Jun	2019	522.1

2.1.2. Statistik Data

a. Jumlah data, mean, standar deviasi, persentil, dan ekstremum

```
# Statistik data
df_s = df[['Usage']].describe(percentiles=[.10, .25, .5, .75, .90])
df_s
```

	Usage
count	11847.000000
mean	102.659348
std	115.637732
min	0.500000
10%	2.200000
25%	6.600000
50%	64.100000
75%	173.450000
90%	273.300000
max	522.100000

b. Distribusi data berdasarkan States

```
# Distribusi tiap states
df[['States']].value_counts()
```

```
States
West Bengal          359
Kerala               359
Arunachal Pradesh    359
Assam                359
Bihar                359
Chandigarh           359
Chhattisgarh         359
DNH                  359
Delhi                359
Goa                  359
Gujarat              359
HP                   359
Haryana              359
J&K                  359
Jharkhand            359
Karnataka            359
MP                   359
Uttarakhand          359
Maharashtra          359
Manipur              359
Meghalaya            359
Mizoram              359
Nagaland             359
Odisha               359
Pondy                359
Punjab               359
Rajasthan            359
Sikkim               359
Tamil Nadu           359
Telangana            359
Tripura              359
UP                   359
Andhra Pradesh       359
```

c. Distribusi data berdasarkan Regions

```
# Distribusi tiap regions
df[['Regions']].value_counts()
```

```
Regions
NR         3231
NER        2513
WR         2154
SR         2154
ER         1795
```

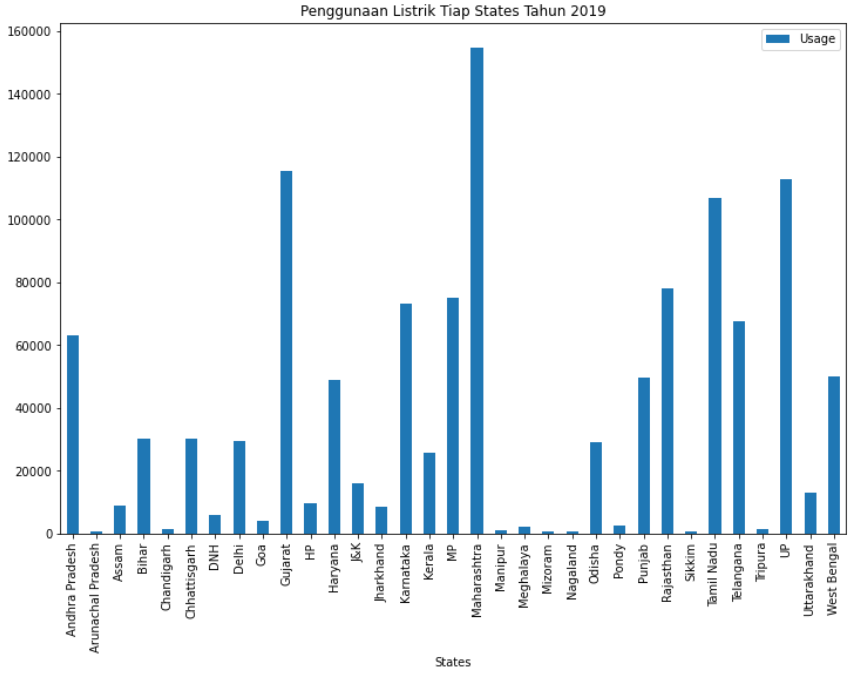
- Informasi/pengetahuan yang didapat

  Dari 33 *States* di India pada data, berdasarkan penggunaan listriknya pada tahun 2019 dapat dilihat masih adanya ketimpangan infrasturuktur dan distribusi sumber energi listrik di antara satu dan lainnya. Penggunaan listrik terkecil ada di Sikkim dengan *usage* 0.5, sedangkan yang terbesar ada di Maharashtra dengan *usage* 522.1. Jika dilihat dari data tersebut, dapat diketahui pula bahwa data tersebut dapat mewakili keseluruhan negara India karena data yang disajikan yaitu 33 dari 36 *States* dan 5 dari 6 *Regions* di India.

2.2. Visualisasi

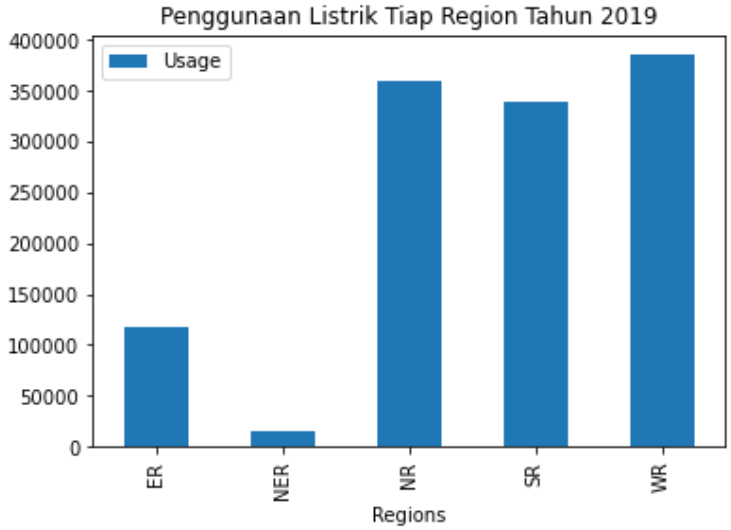
a. Diagram batang penggunaan listrik tiap States pada tahun 2019

```
# Penggunaan listrik tiap states pada tahun 2019

df.groupby('States').sum().plot(kind='bar', y='Usage', figsize = (12,8), title='Penggunaan Listrik Tiap States Tahun 2019')
plt.show()
```



b. Diagram batang penggunaan listrik tiap Regions pada tahun 2019

```
# Penggunaan Listrik Tiap Regions pada tahun 2019

df.groupby('Regions')[['Usage']].sum().plot(kind='bar', title='Penggunaan Listrik Tiap Region Tahun 2019')
plt.show()
# ER  : East
# NER : Northeast
# NR  : North
# SR  : South
# WR  : Western
```

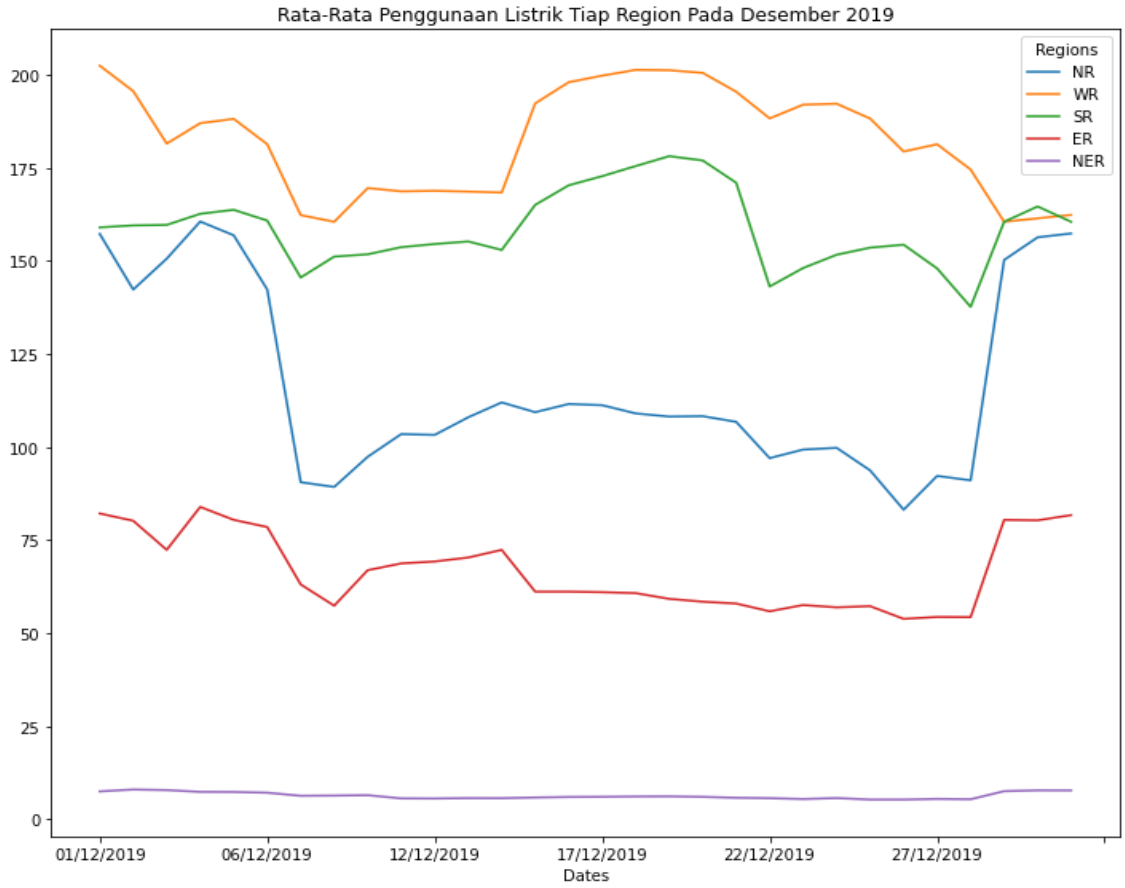


- *Insight* yang didapat

  Persebaran sumber energi listrik dan infrastruktur di India belum merata, terutama di India bagian timur.

c. Diagram garis rata-rata penggunaan listrik tiap region pada Desember 2019

```
# Rata-rata penggunaan listrik tiap region pada desember 2019

u_reg = df.loc[df['Month'] == 'Dec'].groupby(['Dates', 'Regions'], sort=False)['Usage'].mean().unstack()
u_reg.plot(kind="line", y=["NR","WR","SR","ER","NER"]
           ,figsize=(12,10)
           ,title='Rata-Rata Penggunaan Listrik Tiap Region Pada Desember 2019')
plt.show()
```

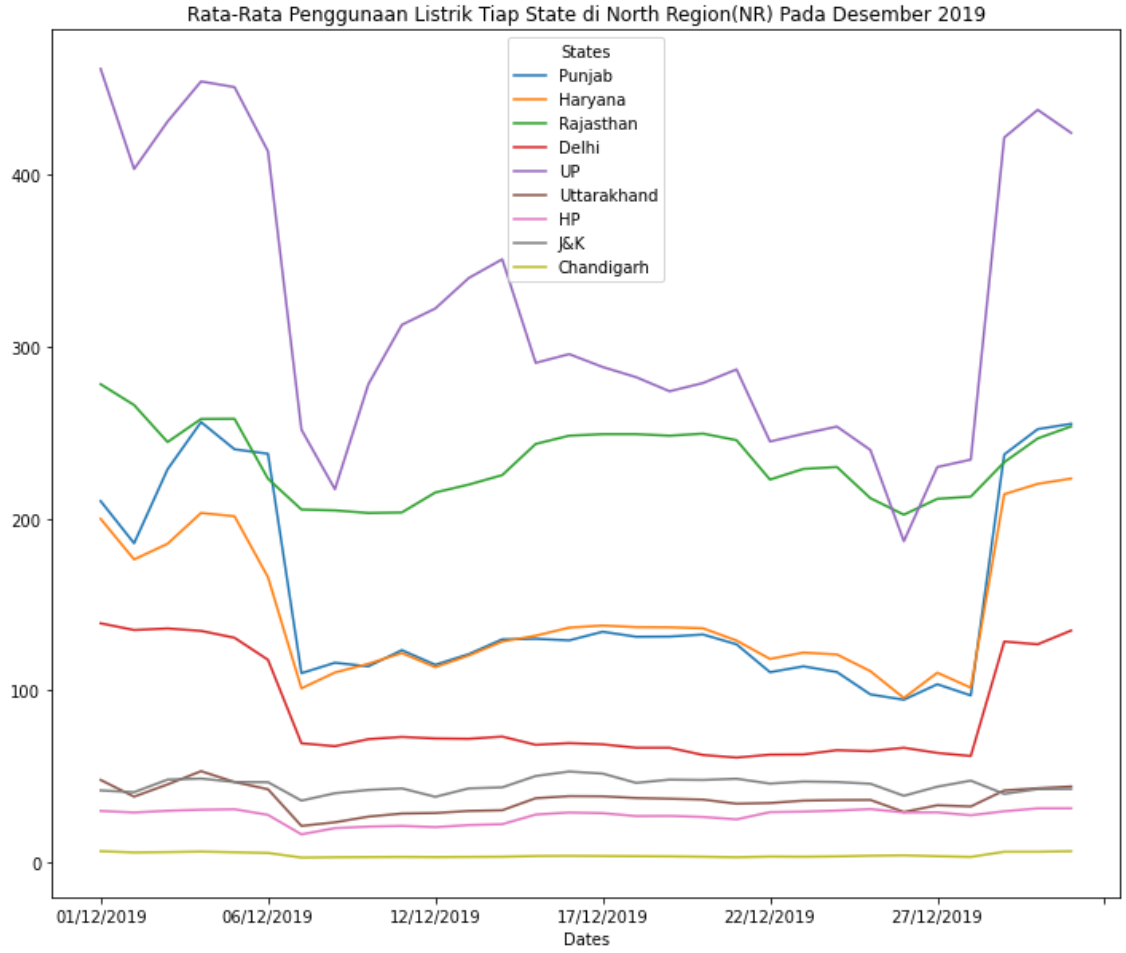


- *Insight* yang didapat

  Tim penulis mengambil data pada Desember 2019 karena pertimbangan banyak hari libur dan menjelang perayaan tahun baru, untuk mengetahui penggunaan listrik tiap *region*. Didapati grafiknya cukup fluktuatif dan penggunaan listrik di North-East Region (NER) sangat rendah dibandingkan dengan *region* lain. Di sisi lain, Tim Penulis tertarik pada North Region (NR), karena memiliki penurunan dan kenaikan yang sangat signifikan pada waktu tertentu, oleh karena itu kami menganalisa lebih lanjut pada grafik berikutnya.

d. Diagram garis rata-rata penggunaan listrik tiap states di NR pada Desember 2019

```
# Rata-rata penggunaan listrik tiap states di North Region pada Desember 2019

u_nr = df.loc[ (df['Regions'] == 'NR')&(df['Month'] == 'Dec')].groupby(['Dates','States'], sort=False)['Usage'].mean().unstack()
u_nr.plot(kind="line", y=["Punjab","Haryana","Rajasthan","Delhi","UP","Uttarakhand", "HP", "J&K", "Chandigarh"]
         ,figsize=(12,10),
         title='Rata-Rata Penggunaan Listrik Tiap State di North Region(NR) Pada Desember 2019')
plt.show()

# ambil bln dec karena akhir tahun, pake NR karena based on grafik atas
# karena di NR ada kenaikan signifikan pada akhir des 2019
# 4 dec AL INDIA
```

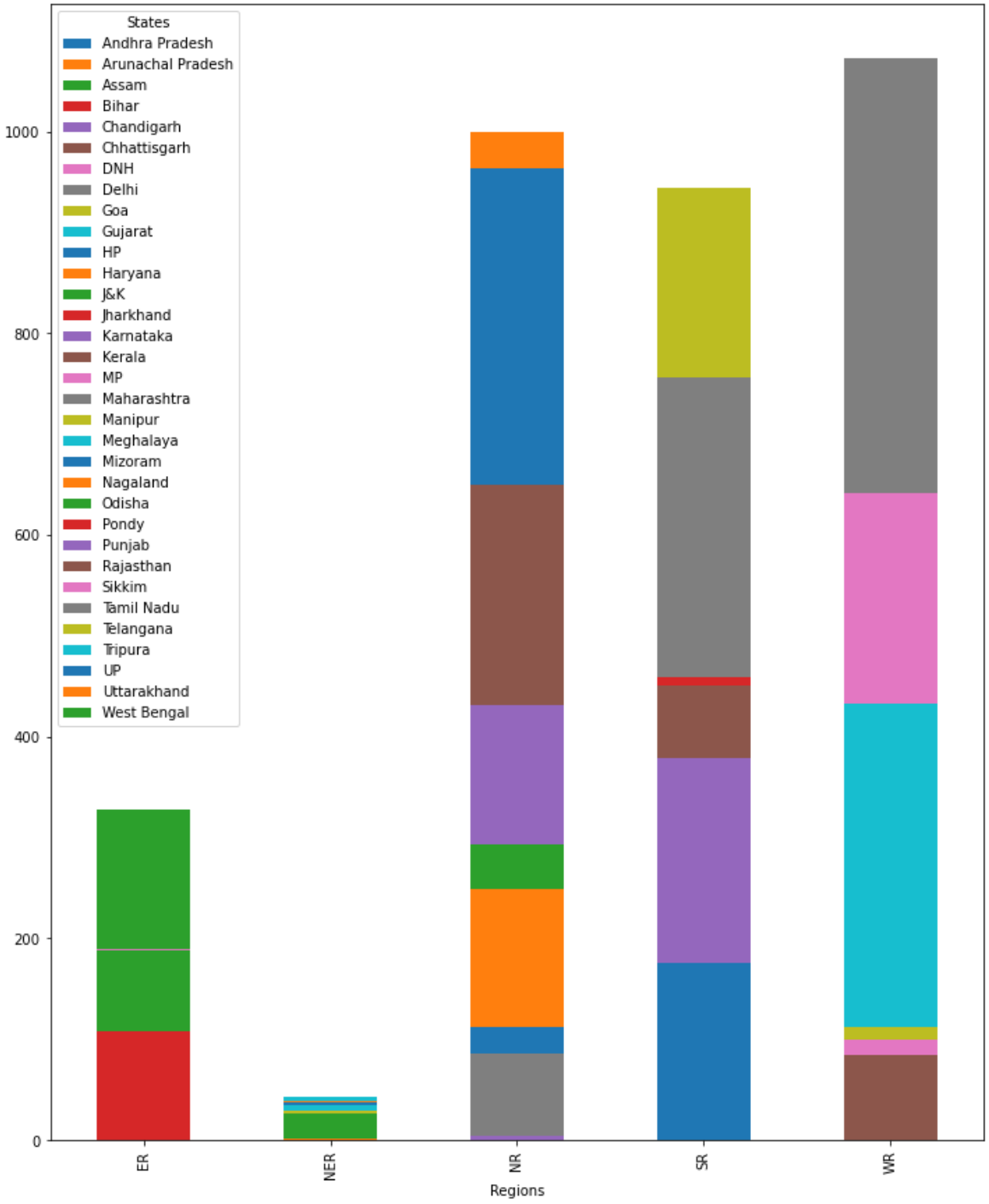


- *Insight* yang didapat

Tim penulis pengadakan riset sederhana dan didapati terdapat penggunaan listrik yang cukup tinggi pada tanggal 4 Desember disebabkan oleh adanya perayaan *Navy Day to commemorate Operation Trident* di India yang dirayakan setiap tahunnya. Dapat dilihat pula dari grafik, bahwa penyumbang usage terbesar dari North Region (NR) adalah Uttar Pradesh (UP).

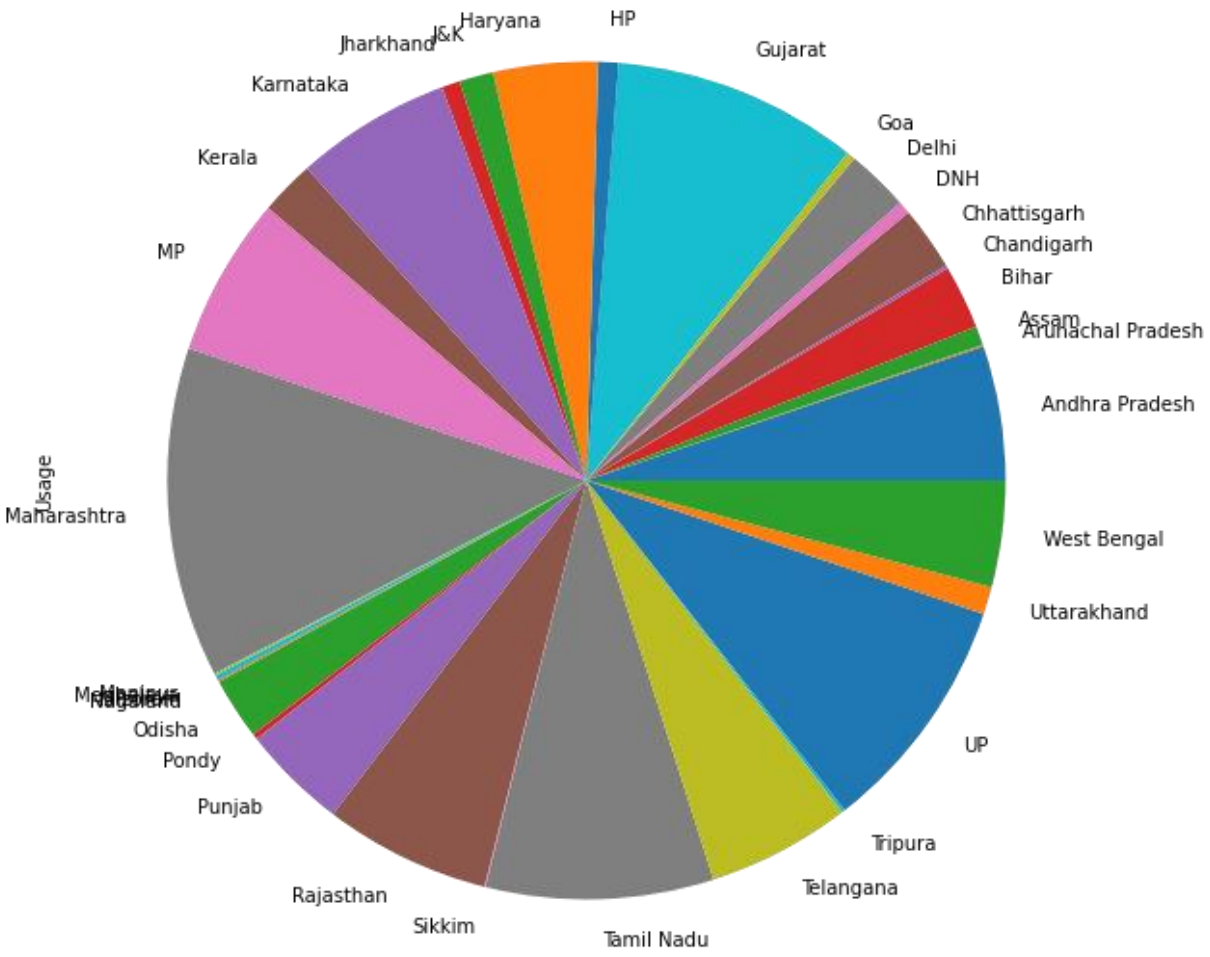
e. Diagram batang penampilan hierarki dan hubungan keseluruhan bagian

```
#Penampilan hierarki dan hubungan keseluruhan bagian
df_pt = df.pivot_table(index='Regions', columns='States', values='Usage')
df_pt.plot.bar(stacked=True, figsize=(12, 15))
plt.show()
```



f. Pie chart penampilan hierarki dan hubungan keseluruhan bagian

```
#Penampilan hierarki dan hubungan keseluruhan bagian
df.groupby('States').sum()['Usage'].plot(kind = "pie", figsize = (14,10))
plt.show()
```

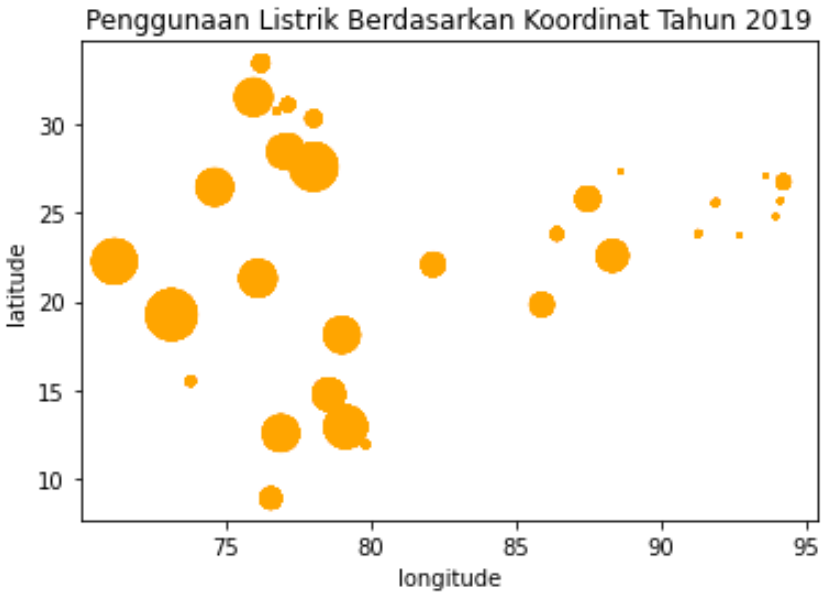


- *Insight* yang didapat

  Dalam visualisasi hierarki data lebih proper digunakan dengan *stacked-bar*, karena pembagian *region*, *states*, dan *usage* dapat dilihat lebih jelas dibandingkan dengan *pie-chart*.

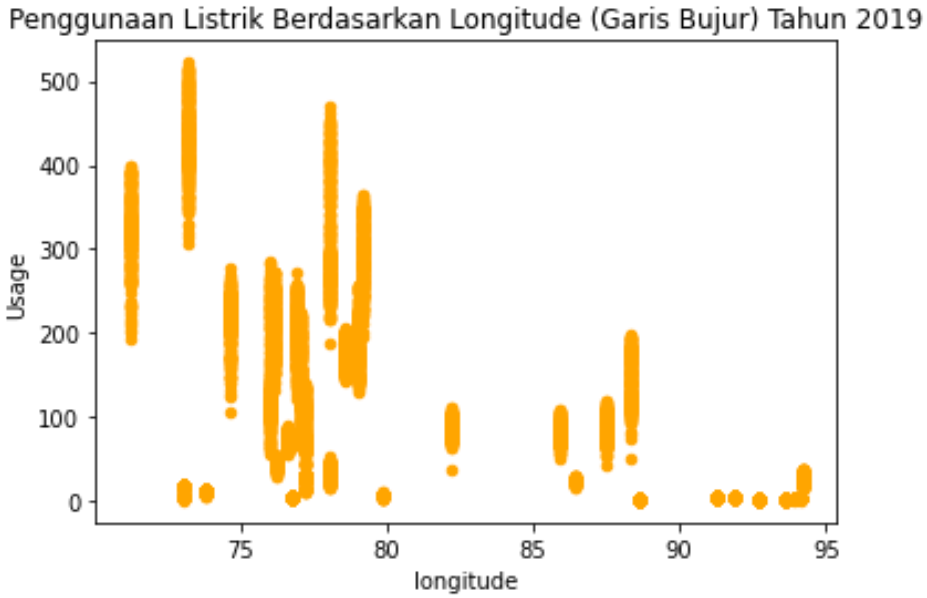
f. Diagram scatter bubble penggunaan listrik berdasarkan koordinat pada tahun 2019

```
# Bubble plot ; Penggunaan Listrik Berdasarkan Koordinat Tahun 2019
u_long_lat = df[['latitude','longitude','Usage']]
u_long_lat.plot(kind="scatter",x="longitude", y="latitude", sizes=a['Usage'] ,color="orange"
       , title='Penggunaan Listrik Berdasarkan Koordinat Tahun 2019')
plt.show()
```



g. Diagram scatter penggunaan listrik berdasarkan garis bujur pada tahun 2019

```
# Scatter plot ; Penggunaan Listrik Berdasarkan Longitude (Garis Bujur) Tahun 2019
u_long = df[['latitude','longitude','Usage']]
u_long.plot(kind="scatter",x="longitude", y="Usage", color="orange"
       ,title='Penggunaan Listrik Berdasarkan Longitude (Garis Bujur) Tahun 2019')
plt.show()
```



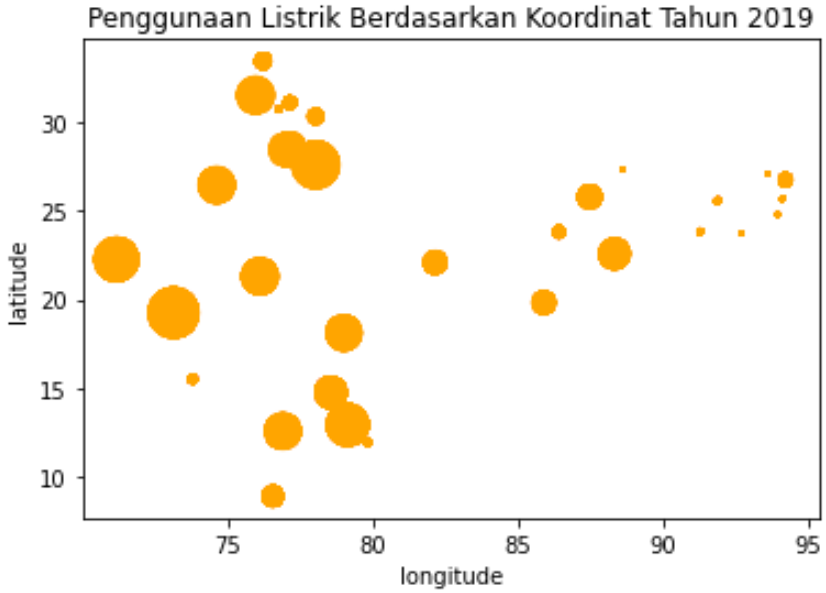
- *Insight* yang didapat

  Dapat diilihat dari bubble dan scatter plot yang ada, India bagian barat lebih maju dalam hal infrastruktur dan distribusi sumber energi listrik.

2.3. Korelasi

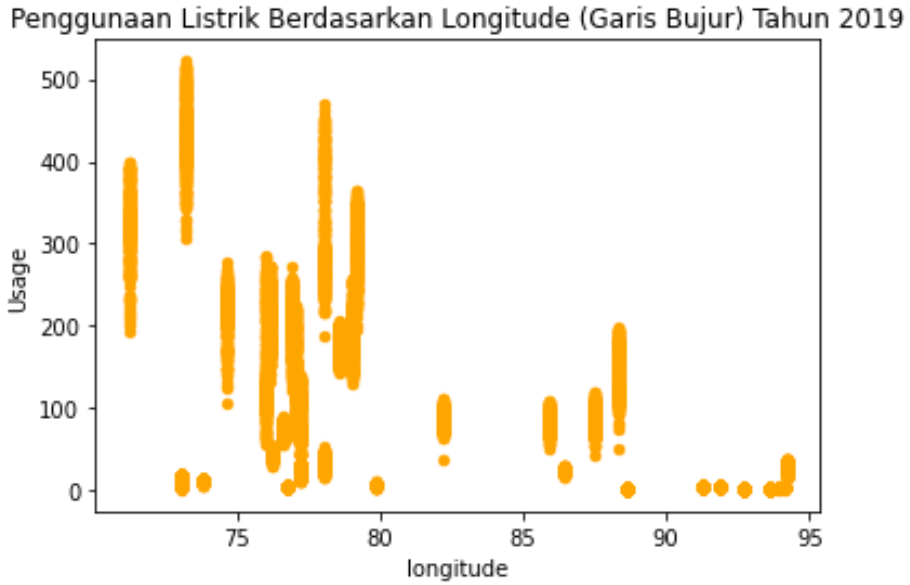
a. Diagram scatter bubble penggunaan listrik berdasarkan koordinat pada tahun 2019

```
# Bubble plot ; Penggunaan Listrik Berdasarkan Koordinat Tahun 2019
u_long_lat = df[['latitude','longitude','Usage']]
u_long_lat.plot(kind="scatter",x="longitude", y="latitude", sizes=a['Usage'] ,color="orange"
       , title='Penggunaan Listrik Berdasarkan Koordinat Tahun 2019')
plt.show()
```



b. Diagram scatter penggunaan listrik berdasarkan garis bujur pada tahun 2019

```
# Scatter plot ; Penggunaan Listrik Berdasarkan Longitude (Garis Bujur) Tahun 2019
u_long = df[['latitude','longitude','Usage']]
u_long.plot(kind="scatter",x="longitude", y="Usage", color="orange"
       ,title='Penggunaan Listrik Berdasarkan Longitude (Garis Bujur) Tahun 2019')
plt.show()
```



c. Nilai korelasi antara longitude dan Usage

```
# Nilai korelasi antara longitude dan usage
nilai_corr = u_long['longitude'].corr(u_long['Usage'])
print('Nilai korelasi antara longitude dan usage adalah ' + str(nilai_corr))
print()
print('Dapat dikatakan pula low negative relationships')
```

```
Nilai korelasi antara longitude dan usage adalah -0.5242078967797845

Dapat dikatakan pula low negative relationships
```

- *Insight* yang didapat

Nilai korelasi yang didapat antara garis bujur dan penggunaan listrik pada tahun 2019 yaitu berkisar pada -0.52 yang berarti memiliki *low negative relationships*.

2.4. Data Cleansing

- Deskripsi kekotoran data

Pada tahun 2020 data hanya mencakup lima hari di setiap bulannya. Juga ditemukan pula format atribut ‘Dates’ yang berbeda-beda, ada yang mengikuti US format, adapula yang tidak, serta terdapat pula beberapa data *overleap*.

- Cara mengatasi

Kami menulis ulang data baru dan hanya mengambil data tahun 2019, karena dengan tidak lengkapnya data tahun 2020 akan membuat data yang diolah dan kesimpulan yang didapatkan akan dipertanyakan kebenarannya. Mengenai format atribut ‘Dates’ yang berbeda-beda kami atasi dengan cara membuat atribut baru yaitu ‘Month’ dan ‘Year’, agar persortiran dan pencarian data terhadap waktu dapat dilakukan tanpa adanya kesalahan. Untuk data yang overleap kami hilangkan agar analisis perhitungan lebih akurat.

Dalam mengatasi hal tersebut kami melakukan dengan Microsoft Excel, dengan munggunakan tool ‘find & replace’, fungsi perubahan format tanggal, dan menulis ulang data tersebut pada file baru untuk ditetapkan.

BAB III

KESIMPULAN

3.1. Informasi yang Didapat

1. Pada grafik rata-rata penggunaan listrik tiap state di North Region (NR) pada bulan Desember 2019, terjadi lonjakan penggunaan listrik pada tanggal 4 Desember 2019. Setelah diteliti, tanggal tersebut adalah tanggal perayaan Indian Navy Day. Sehingga dapat disimpulkan pada perayaan tersebut terjadi lonjakan penggunaan listrik
2. Pada grafik scatter plot penggunaan listrik berdasarkan koordinat tahun 2019 dapat dilihat bahwa penggunaan listrik India bagian barat lebih besar dibandingkan dengan India bagian timur. Hal ini dapat disebabkan karena konsentrasi perkotaan dan aktivitas di negara India terjadi di daerah barat
3. Nilai korelasi antara longitude dan usage adalah -0.52 yang berarti korelasinya yaitu *low negative relationship*. Hal ini artinya semakin ke timur, penggunaan listrik semakin kecil

3.2. Kesimpulan

1. Pandas merupakan library yang cukup fleksibel dan mudah dipelajari untuk digunakan dalam data analisis bagi pemula
2. Konsentrasi perkotaan dan aktivitas berpengaruh pada penggunaan listrik pada daerah tersebut, semakin banyak perkotaan dan aktivitas, maka penggunaan listrik semakin tinggi
3. Pada hari perayaan tertentu, konsumsi listrik cenderung naik

PEMBAGIAN TUGAS

Pada tugas besar kali ini, pembagian tugas sebagai berikut:

Tugas 3 & 4	: Ray Clement, Rahmat Al Fajri
Tugas 5, 6, 7, & 8	: Rio Alexander , Rava Naufal A
Laporan, dan PPT	: Ray Clement, Rahmat Al Fajri

REFERENSI

https://www.kaggle.com/twinkle0705/state-wise-power-consumption-in-india

https://indianexpress.com/article/explained/navy-day-2020-why-india-remembers-operation-trident-every-year-on-december-4-7092063/#:~:text=Every%20year%2C%20India%20celebrates%20December,%2C%20too%2C%20is%20celebrated%20annually.

https://stackoverflow.com/questions/47861085/using-recently-created-attributes-in-pandas-dataframe-to-create-new-attribute