Laporan Tugas Analisis Data

***Abstrak*—Makalah ini bertujuan untuk mengenali persoalan komputasi tentang analisis data, mulai dari menggunakan metode dekomposisi dan abstraksi dalam pemecahan persoalan, menganalisis data, hingga menuliskan solusinya dengan bahasa pemrograman Python.**

Kata Kunci—python, data covid, data cuaca,

I. Deskripsi Data dan File

A. Deskripsi Data

1) Data Covid

Semakin meningkatnya jumlah penderita Covid di dunia, semakin banyak juga penelitian tentang wabah yang merebak semenjak tahun 2020 ini. Data Covid ini berisi tingkat risiko masyarakat terhadap Covid-19. Data yang disajikan berupa tingkat risiko tiap kasur pasien, ruangan ICU pasien, rumah sakit di setiap daerah, dan total risiko secara keseluruhan.

Data Covid yang digunakan untuk tugas besar data analisis ini merupakan data yang dimanfaatkan untuk melihat potensi masyarakat di 50 negara bagian Amerika Serikat, baik daerah micropolitan maupun metropolitan, untuk terjangkit virus Covid-19 berdasarkan persentase risiko yang disajikan. Data Covid ini juga menjadi sumber dari penulisan berbagai artikel, berita, dan jurnal terkait wabah Covid-19 di Amerika Serikat.

Data ini disusun karena ternyata masih banyak rumah sakit yang hanya memiliki jumlah ruang rawat inap dan ICU yang sangat sedikit di beberapa daerah Amerika Serikat. Padahal, jumlah Covid-19 selalu menaik tajam setiap harinya sedangkan fasilitas yang memadai jauh dari kata cukup. Oleh karena itu, data ini dibuat untuk melihat hubungan antara fasilitas ruang rawat inap di rumah sakit dan jumlah penderita Covid-19 sehingga, nantinya, dapat dihasilkan solusi untuk menurunkan angka penderita Covid dan mengalahkan wabah ini.

2) Data Cuaca

Data Cuaca yang digunakan dalam tugas besar ini adalah data cuaca dari negara bagian Philadhelpia. Data ini berisi temperature rata-rata, minimum, dan maksimum sepanjang tahun Philadhelpia dalam kurun waktu 1 tahun, yaitu dari tanggal 1 Juli 2014 sampai dengan 30 Juni 2015.

Data ini disusun tidak hanya sebagai rekaman temperature di Philadhelpia selama satu tahun terakhir, tetapi juga untuk membandingkannya dengan 49 negara bagian Amerika Serikat lainnya. Banyak warga negara Amerika, terutama yang dibagian barat, merasa cuaca dan temperature pada tahun 2014—2015 terasa lebih terik dan panas dibandingkan sebelumnya. Data ini dapat membantu membuktikan asumsi warga negara Amerika dengan membandingkannya dengan negara bagian lainnya.

Dari data ini dapat dibuktikan bahwa ternyata benar asumsi bahwa temperature rata-rata Amerika Serikat pada tahun 2014 menjelang 2015 lebih tinggi dari tahun-tahun sebelumnya. Data ini juga membantu menjelaskan bahwa temperature di semua negara bagian tidak naik semua, tetapi juga ada yang turun. Data ini cukup merepresentasikan temperature di Philadhelpia.

B. Format dan Ukuran File

Data Covid dan data cuaca sama-sama berformat Comma Separated Values (csv). Di setiap data, juga dilampirkan file berformat Markdown (md) yang berisi penjelasan singkat tentang data terkait.

Ukuran file data Covid sebesar 10,461 bytes, sedangkan ruang pada media penyimpanannya sebesar 12,288 bytes. Ukuran file data cuaca sebesar 20,560 bytes, sedangkan ruang pada media penyimpanannya sebesar 24,576 bytes.

C. Sumber Data

1) Data Covid

Data covid ini diambil dari GitHub. Data ini merupakan gabungan dari data yang diambil dari BRFSS (Behavioral Risk Factor Surveillance System) dan data hasil survei yang dilakukan setiap tahunnya oleh Kaiser Family Foundation dengan melibatkan lebih dari 400,000 orang Amerika.

2) Data Cuaca

Data cuaca ini langsung diambil dari Weather Underground. Weather Underground sendiri merupakan perusahan yang bergerak pada bidang informasi terkait cuaca *real-time*

D. Pengolahan Data

Untuk mengolah data Covid dan cuaca yang telah dicari sebelumnya, kami menggunakan bahasa pemrograman Python karena bahasa Python cukup mudah untuk dipelajari karena sintaksnya yang sederhana. Bahasa Python juga dinamis dengan tingkat keterbacaan yang tinggi.

Sedangkan untuk tool yang dipakai untuk mengolah data di Python, kami menggunakan Jupyter Notebook. Tool ini juga cukup mudah penggunaanya. Penggunaannya juga fleksibel serta mudah untuk memodifikasi apabila ada kesalahan pada coding.

Kami juga menggunakan *library* pandas untuk membaca dan mengolah data serta matplotlib untuk menvisualisasikan data.

1) Membaca data

```
import pandas as pd

import matplotlib.pyplot as plt

data_covid=pd.read_csv("mmsa-icu-
beds.csv")

data_cuaca=pd.read_csv("KPHL.csv")
```

2) Mengakses data tertentu

```
data_covid[:10]
```

```
data_cuaca.loc[data_cuaca["date"] ==
"2014-12-18"]
```

3) Menampilkan statistik sederhana

```
data_covid.describe()
```

E. Dimensi Ukuran Data

Dengan menggunakan bahasa pemrograman Python, dapat ditentukan jumlah, baris, kolom, dan sel secara keseluruhan.

```
baris_covid = len(data_covid)

kolom_covid = len(data_covid.columns)

sel_covid = baris_covid * kolom_covid

baris_cuaca = len(data_cuaca)

kolom_cuaca = len(data_cuaca.columns)

sel_cuaca = baris_cuaca * kolom_cuaca
```

Sehingga didapatkan :

1) Data Covid

- Baris = 136
- Kolom = 7
- Total Sel = 952

2) Data Cuaca

- Baris = 365
- Kolom = 13
- Total Sel = 4745

II. KARAKTERISTIK DATA

Untuk menentukan range dari data yang berkarakteristik kategorikal nominal, dapat menggunakan :

```
print(data_covid.min())

print(data_covid.max())

print(data_cuaca.min())

print(data_cuaca.max())
```

A. Data Covid

Pada Data Covid, terdapat 7 kolom yang bertipe dan menjelaskan ;

1) MMSA

Data MMSA bertipe atribut kategoris (kualitatif) nominal dengan nilainya menunjukkan nama-nama daerah metropolitan dan micropolitan yang tersebar di 50 negara bagian Amerika Serikat. MMSA diambil dari *Covid Data Tracker* yang terdapat di BRFF.

2) total_percent_at_risk

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan persentase seseorang di suatu area MMSA yang rentan memiliki risiko tinggi terpapar virus Covid-19. Range dari kolom ini berkisar antara 38,92% -- 80,73%

3) high_risk_per_icu_bed

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan jumlah orang yang memiliki risiko tinggi untuk terpapar COVID-19 dengan membandingkannya dengan jumlah ruang ICU di daerah itu. Misal, *high_risk_per_icu_bed* bernilai 5, berarti ada 5 orang yang kemungkinan terpapar Covid -19 dari total keseluruhan ruang ICU di daerah MMSA itu. Range kolom ini antara 413.668 – 4489.85

4) high_risk_per_hopistal

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan jumlah orang yang memiliki risiko tinggi untuk terpapar COVID-19 dengan membandingkannya dengan jumlah setiap rumah sakit di daerah itu. Misal, *high_risk_per_hopistal* bernilai 10, berate ada 10 orang di setiap rumah sakit di daerah itu yang memiliki risiko tinggi terpapar Covid-19. Range kolom ini berkisar di antara 6670.19 – 91771.3

5) icu_beds

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan jumlah ruangan ICU di daerah MMSA tersebut. Data dari kolom ini digunakan untuk melakukan perbandingan dengan kolom ke-2. Data ini didapatkan dari Kaiser Family Foundation. Range kolom ini berada di antara 8 – 2777.

6) hospitals

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan jumlah rumah sakit di daerah MMSA tersebut. Data dari kolom ini digunakan untuk melakukan perbandingan dengan kolom ke-3. Data ini didapatkan dari Kaiser Family Foundation. Range kolom ini berada di antara 1 – 100.

7) total_at_risk

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan total jumlah orang secara keseluruhan yang berpotensi untuk terpapar Covid-19 di area MMSA tersebut. Kolom ini dihitung berdasarkan analisis pakar disana, seperti dari usia dan riwayat penyakit. Range kolom ini berada di antara 17941,5 – 6165100.

B. Data Cuaca

Data cuaca memiliki 13 kolom yang bertipe dan menjelaskan :

1) date

Data pada kolom date bertipe numerik (kuantitatif) interval yang perbedaan diantaranya memiliki makna yang berarti (-). Data pada kolom ini menunjukkan tanggal-tanggal cuaca di Philadhelpia dicatat dengan format YYYY-M-D (tahun-bulan-hari). Data ini berkurun waktu satu tahun antara tanggal 1 Juli 2014 sampai 30 Juni 2015.

2) Actual_mean_temp

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan hasil pengukuran rata-rata temperature selama satu hari di Philadhelphia. Range kolom ini berada di antara 10 – 86.

3) Actual_min_temp

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan temperature minimum pada hari yang sesuai dengan baris yang sama. Range kolom ini berada di antara 2 – 77.

4) Actual_max_temp

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan temperature maksimum pada hari yang sesuai dengan baris yang sama. Range kolom ini berada di antara 17 – 96.

5) Average_min_temp

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan rata-rata temperature terendah (minimum) pada hari itu dengan membandingkannya dengan rata-rata hari-hari sebelumnya sejak tahun 1880. Range kolom ini berada di antara 25 – 70.

6) Average_max_temp

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan rata-rata temperature tertinggi (maksimum) pada hari itu dengan membandingkannya dengan rata-rata hari-hari sebelumnya sejak tahun 1880. Range kolom ini berada di antara 40 – 87.

7) Record_min_temp

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan temperature terendah (minimum) dari tahun 1880 sampai dengan hari itu. Misal, temperature terendah hari itu adalah -7°C, apabila -7°C itu merupakan yang terendah semenjak tahun 1880, maka nilai yang akan ditampilkan akan -7°C. Jika tidak, nilai yang ditampilkan pasti akan lebih rendah. Range kolom ini berada di antara -11 – 59.

8) Record_max_temp

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan temperature tertinggi (maksimum) dari tahun 1880 sampai dengan hari itu. Misal, temperature tertinggi hari itu adalah 38°C, apabila 38°C itu merupakan yang tertinggi semenjak tahun 1880, maka nilai yang akan ditampilkan akan 38°C. Jika tidak, nilai yang ditampilkan pasti akan lebih tinggi. Range kolom ini berada di antara 61 – 106.

9) Record_min_temp_year

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan pada tahun apa data pada kolom *record_min_temp* terjadi. Atau dengan kata lain, kolom ini menunjukkan tahun dengan hari yang memiliki temperature terendah (minimum) dari tahun 1880 hingga hari yang terdapat pada baris yang sama. Range kolom ini berada di antara 1872 – 2014.

10) Record_max_temp

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan pada tahun apa data pada kolom *record_max_temp* terjadi. Atau dengan kata lain, kolom ini menunjukkan tahun dengan hari yang memiliki temperature tertinggi (maksimum) dari tahun 1880 hingga hari yang terdapat pada baris yang sama. Range kolom ini berada di antara 1874 – 2014.

11) Actual_precipitation

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan curah hujan atau salju pada hari yang sesuai dengan baris yang sama. Range kolom ini berada di antara 0 – 2,01.

12) Average_precipitation

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan rata-rata curah hujan atau salju pada hari yang sesuai dengan baris yang sama. Range kolom ini berada di antara 0,08 – 0,15.

13) Record_precipitation

Data di kolom ini bertipe atribut kategorikal nominal yang menunjukkan curah hujan tertinggi dari tahun 1880 sampai dengan hari itu. Range kolom ini berada di antara 0,85 – 8,02.

III. STATISTIK

Pertama-tama, kita akan menginput data dari file "mmsa-icu-beds.csv" dan "KPHL.csv" yang masing-masing berisi data covid di USA dan data cuaca di Philadelphia.

```
import pandas as pd

import matplotlib.pyplot as plt

data_covid=pd.read_csv("mmsa-icu-
beds.csv")

data_cuaca=pd.read_csv("KPHL.csv")
```

Kita akan melihat masing-masing 10 data pertama dari kedua data sebagai sampel data.

```
print("Sampel Data Covid USA: ")

print(data_covid[:10])

print("-----------------------------")

print("Sampel Data Cuaca Philadelphia: ")

print(data_cuaca[:10])
```

Agar kita bisa mendapat gambaran besar data, maka kita akan melihat beberapa statistik data.

A. Standar Deviasi

1) Data Covid

```
print(data_covid.std())
```

Dari data tersebut, kita dapat melihat bahwa deviasi pasien "high risk" baik secara total maupun per rumah sakit sangat besar. Ini menunjukkan bahwa sebaran virus corona tidak menentu. Selain itu, deviasi kasur ICU dan pasien per kasur ICU relatif lebih kecil karena kasur ICU merupakan faktor yang bisa dikendalikan dan tiap distribusi ICU ini diregulasi oleh pemerintah. Kemudian, deviasi banyak rumah sakit jauh lebih kecil karena rumah sakit jumlahnya

memang tidak terlalu banyak. Seperti halnya dengan kasur ICU juga, pembangunan rumah sakit diregulasi oleh pemerintah sehingga distribusinya lebih merata.

2) Data Cuaca

```
print(data_cuaca.std())
```

Seperti yang diharapkan, deviasi data cuaca yang merupakan fenomena teratur memiliki deviasi yang relatif rendah. Deviasi yang lebih besar disebabkan nilai data yang lebih besar (seperti tahun yang nilainya ribuan) dan deviasi kecil disebabkan data kecil (seperti nilai presipitasi yang berkisar di 1).

B. Rata-rata

1) Data Covid

```
print(data_covid.mean())
```

Nilai rata-rata diatas menunjukkan tingkat keganasan covid yang mencapai 667 ribu orang per daerah dengan 43 ribu beresiko tinggi, tetapi hanya ada sekitar 14 rumah sakit per daerah. Kasur ICU yang tersedia hanya 360 dan tiap kasur secara ideal harus digunakan 1900 orang secara rata-rata.

2) Data Cuaca

```
print(data_cuaca.mean())
```

Nilai rata-rata tersebut menunjukkan gambaran besar suhu dan presipitasi selama satu tahun di USA. Namun, tentu saja nilai rata-rata tahun rekor maksimum dan minimum suhu tidak berarti apa-apa secara cuaca.

C. Kuartil

1) Data Covid

```
print(data_covid.quantile([0.25,0.5,0.75])
)
```

Kita dapat melihat bahwa perbedaan rumah sakit dan kasur ICU di tiap kuartil cukup signifikan. Hal ini disebabkan lebih banyak infrastruktur yang dibangun di daerah yang lebih padat dan sebaliknya. Selain itu, sebaran covid yang mencapai 900 ribu penduduk pada kuartil tiga juga menunjukkan bahwa covid menyebar lebih mudah di daerah padat tersebut.

2) Data Cuaca

```
print(data_cuaca.quantile([0.25,0.5,0.75])
)
```

Seperti yang diharapkan, data temperatur relatif berbeda pada tiap kuartil. Ini wajar karena sepanjang tahun, cuaca berubah drastis mulai dari musim panas hingga musim dingin. Lain halnya, data kuartil presipitasi menunjukkan bahwa kejadian hujan relatif sama sepanjang tahun dan sangat jarang terjadi.

D. Ekstremum

1) Data Covid

```
print("Maksimum")

print(data_covid.max())

print("-----------------------------")

print("Minimum")

print(data_covid.min())
```

Kita dapat melihat bahwa ada daerah yang hanya memiliki satu rumah sakit, sedangkan ada yang memiliki 100. Selain itu, ada daerah yang memiliki 81% pasien beresiko dan ada yang hanya memiliki 39%. Ini menunjukkan sebaran covid dan kapasitas perawatan tiap daerah keduanya memiliki rentang yang sangat besar.

2) Data Cuaca

```
print("Maksimum")

print(data_cuaca.max())

print("-----------------------------")

print("Minimum")

print(data_cuaca.min())
```

Data tersebut menunjukkan fluktuasi cuaca yang relatif besar. Ada hari dimana temperatur mencapai 96 derajat farenheit dan ada hari dimana temperatur mencapai 2 derajat farenheit.

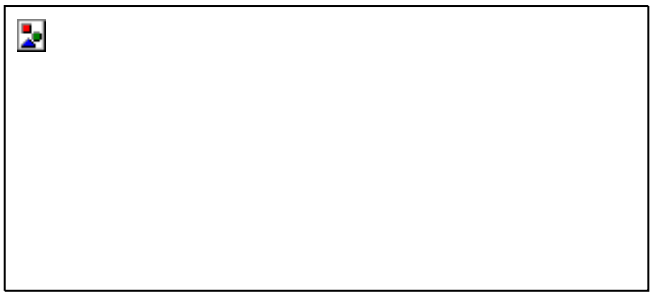
IV. VISUALISASI DATA

Pertama-tama, kita akan menginput data dari file "mmsa-icu-beds.csv" dan "KPHL.csv" yang masing-masing berisi data covid di USA dan data cuaca di Philadelphia.

```
import pandas as pd

import matplotlib.pyplot as plt

data_covid=pd.read_csv("mmsa-icu-
beds.csv")

data_cuaca=pd.read_csv("KPHL.csv")
```

Selanjutnya, akan ditampilkan visualisasi data. Pertama, kita akan menampilkan perbandingan data per kategori.

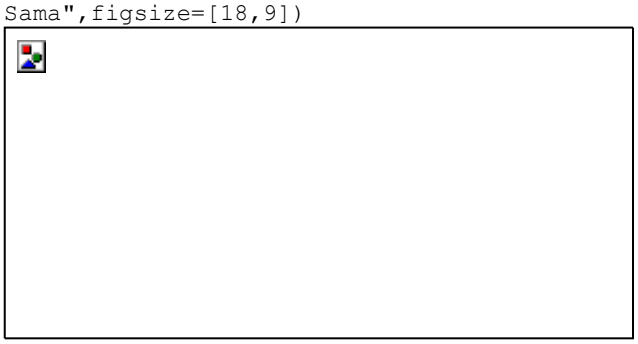
```
jmlh_rs=data_covid["hospitals"]

jmlh_rs.plot(kind="hist",bins=[0,5,10,15,20,2
5,30,35,40], rwidth=0.5,title="Banyaknya
Rumah Sakit per Daerah",figsize=[10,5])
```



Hal pertama yang ditampilkan adalah Frekuensi Banyaknya Rumah Sakit per Daerah. Pada tabel, sumbu x merupakan jumlah rumah sakit dalam suatu daerah dan sumbu y merupakan frekuensi. Dari tabel, dapat disimpulkan bahwa banyak daerah yang memiliki jumlah rumah sakit yang sedikit.

Tabel per kategori yang kedua adalah Frekensi Hari dengan Temperatur yang sama.

```
temperature=data_cuaca["actual_mean_temp"]

temperature.plot(kind="hist",bins=[0,20,40
,60,80,100], rwidth=0.5,title="Banyak Hari
dengan Temperatur Rata-Rata
Sama",figsize=[18,9])
```

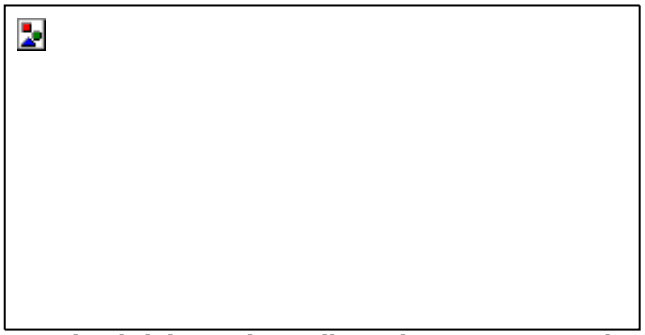


Pada tabel diatas, sumbu x merupakan temperature rata-rata per hari dan sumbu y merupakan frekuensi. Dari tabel dapat ditarik kesimpulan bahwa suhu rata-rata yang sebenarnya sebagian besar terletak di kisaran 60-80 Fahrenheit.

Selanjutnya, akan ditampilkan tabel penampilan perubahan terhadap waktu.

Hal pertama yang ditampilkan adalah perubahan temperature rata-rata tiap hari.

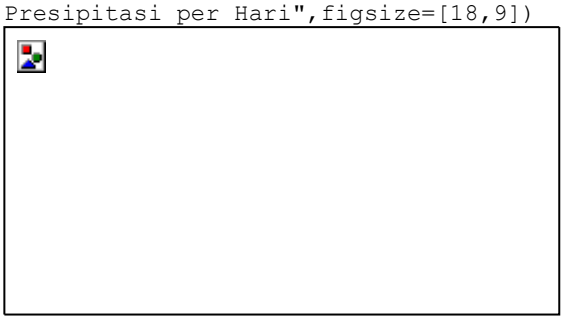
```
suhu_rerata=data_cuaca["actual_mean_temp"]

suhu_rerata.plot(kind="line",x="date",
y="actual_mean_temp",title="Perubahan
Temperatur Rata-Rata tiap
Hari",figsize=[18,9])
```



Pada tabel di atas ditampilkan suhu rata-rata setiap hari, sumbu x merupakan hari ke-n sedangkan sumbu y merupakan temperatur. Dapat disimpulkan bahwa perubahan suhu bersifat siklik, sebab dari grafik dapat dilihat bahwa setelah hampir 365 hari (1 tahun), temperatur kembali ke keadaan yang sama.

Tabel perubahan terhadap waktu yang kedua adalah Data Presipitasi per Hari.

```
presipitasi=data_cuaca["average_precipitat
ion"]

presipitasi.plot(kind="line",x="date",y="a
verage_precipitation",title="Data
Presipitasi per Hari",figsize=[18,9])
```

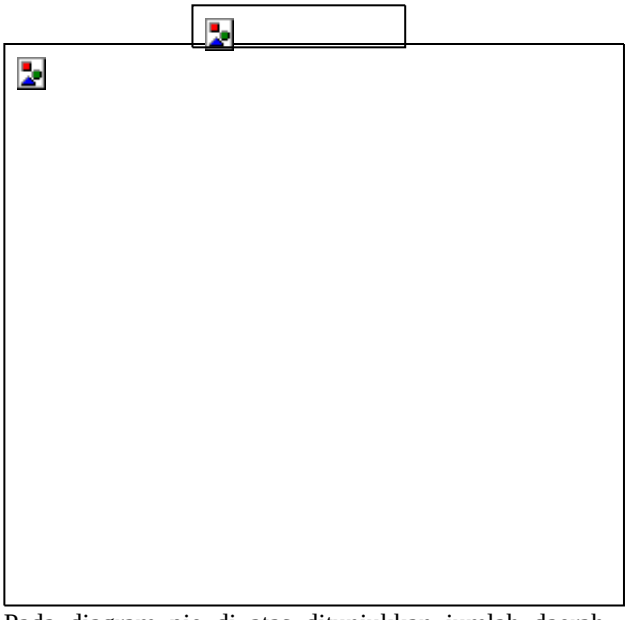


Pada tabel di atas ditampilkan data presipitasi setiap hari, sumbu x adalah tanggal dan sumbu y adalah presipitasi. Dari grafik dapat diambil kesimpulan bahwa presipitasi memang tak menentu, namun dapat berlangsung konstan selama beberapa hari yang berurutan.

Selanjutnya akan ditampilkan tabel penampilan hirearki dan hubungan keseluruhan bagian.

Hal pertama yang ditampilkan adalah komposisi banyaknya rumah sakit.

```
jmlh_rs.value_counts().plot(kind="pie",tit
le="Data Rumah Sakit",figsize=[28,15])
```

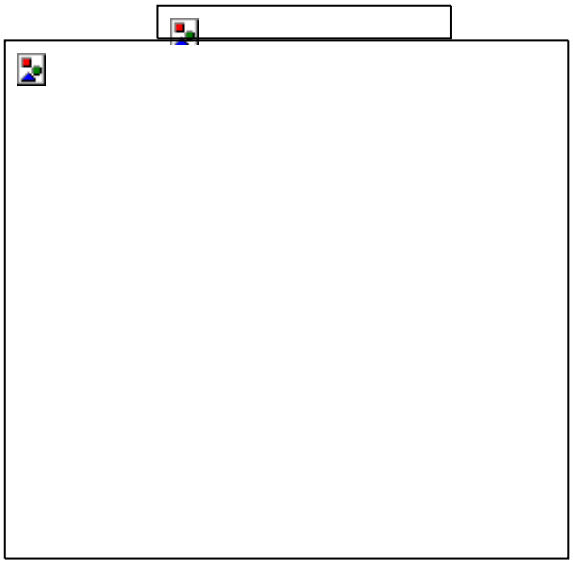


Pada diagram pie di atas ditunjukkan jumlah daerah yang memiliki rumah sakit dalam jumlah tertentu. Dapat ditarik kesimpulan bahwa daerah yang memiliki jumlah rumah sakit sedikit lebih banyak dibanding daerah yang memiliki jumlah rumah sakit yang banyak.

Yang kedua adalah komposisi banyak daerah dengan suhu maksimum rata-rata.

```
suhumaxrerata=data_cuaca["average_max_tem
p"]
```

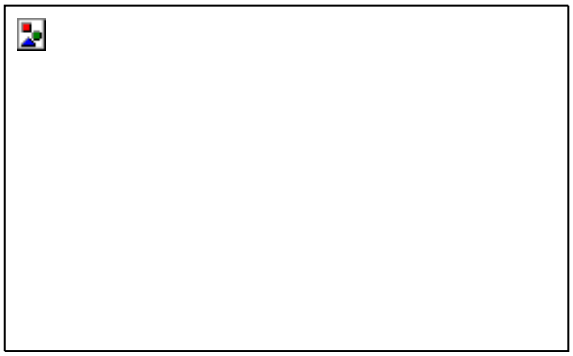
```
suhumaxrerata.value_counts().plot(kind="pie",title="Data Suhu Maksimum Rata-Rata",figsize=[28,15])
```



Pada diagram pie di atas ditunjukan banyaknya jumlah daerah yang memiliki nilai suhu maksimum rata-rata yang sama. Dapat dilihat bahwa daerah dengan suhu maksimum 87 Fahrenheit adalah yang terbanyak.

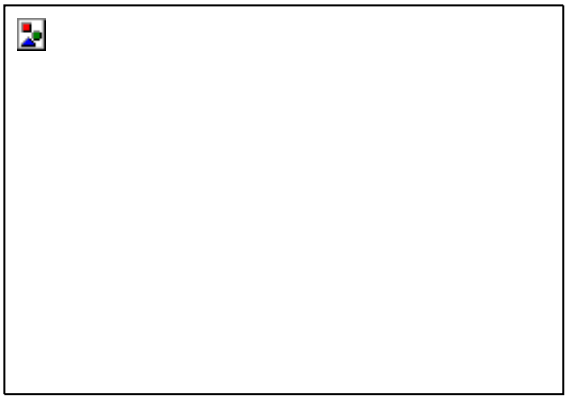
Selanjutnya akan dibuat Plotting Relationship.

Hal pertama yang akan ditampilkan adalah *scatter plot* antara suhu rata-rata minimum dengan suhu rata-rata maksimum.



Pada gambar diatas ditampilkan hubungan suhu rata-rata maks dan min pada suatu daerah, dengan sumbu x adalah suhu rata-rata minimum dan sumbu y adalah suhu rata-rata maksimum, dapat disimpulkan bahwa kenaikan suhu rata-rata minimum sebanding dengan suhu rata-rata maksimum.

Selanjutnya akan ditampilkan *scatter plot* antara ketersedian ICU dan Jumlah Rumah Sakit.



Pada gambar di atas ditampilkan hubungan antara ketersediaan ICU dengan jumlah rumah sakit pada suatu daerah tertentu. Sumbu x merupakan jumlah kamar ICU dan sumbu y merupakan jumlah rumah sakit. Dari grafik dapat dilihat bahwa ketersediaan kamar ICU sangat minim, meskipun berbanding lurus dengan jumlah rumah sakit, namun perbedaan konstanta sangat sedikit, sehingga dapat ditarik kesimpulan jika kamar ICU pada rumah sakit sudah banyak yang penuh.

```
In [9]:
```

TABLE I. TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a] Sample of a Table footnote. (*Table footnote*)

Fig. 1. Example of a figure caption. (*figure caption*)

PEMBAGIAN TUGAS

Adelline Kania : Laporan bagian 1 dan 2, penyusunan laporan, tugas 3 dan 4
Frederik Imanuel : Laporan bagian 3, tugas 1, 2, dan 5
Andreana Hartadi : Laporan bagian 4, tugas 6

DAFTAR PUSAKA

https://repository.unikom.ac.id/49055/1/Pertemuan%203%20-%20Materi%20%5BDM%20-%202016%5D.pdf . Diakses pada 15 Desember 2020