Wendy Dharmawan - 2206059591

Proyek berikut merupakan proyek yang dibuat untuk course Analisis Data dengan Python, yang membahas data wrangling (mengambil dan membersihkan data), EDA (exploratory data analysis), dan data visualization menggunakan library numpy, pandas, dan matplotlib pada python.

Menentukan Pertanyaan Bisnis

- Seberapa besar persentase korelasi kadar O3 terhadap suhu kota Guanyuan pada tahun 2013-2017?
- Bagaimana perbandingan kadar SO2, NO2, dan CO antara kota Guanyuan dan Gucheng pada tahun 2013-2017?

Import Semua Packages/Library yang Digunakan

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Data Wrangling

Gathering Data

Memuat tabel kota Guanyuan

```python
guanyuan_df =
pd.read_csv("https://raw.githubusercontent.com/marceloreis/HTI/master/PRSA_Data_20130301-20170228/PRSA_Data_Guanyuan_20130301-20170228.csv")
guanyuan_df.head()
```

{"summary":"{\n \"name\": \"guanyuan_df\",\n \"rows\": 35064,\n \"fields\": [\n   {\n   \"column\": \"No\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 10122,\n   \"min\": 1,\n   \"max\": 35064,\n   \"num_unique_values\": 35064,\n   \"samples\": [\n   6071,\n   32601,\n   8847\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"year\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 1,\n   \"min\": 2013,\n   \"max\": 2017,\n   \"num_unique_values\": 5,\n   \"samples\": [\n   2014,\n   2017,\n   2015\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"month\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 3,\n   \"min\": 1,\n   \"max\": 12,\n   \"num_unique_values\": 12,\n   \"samples\": [\n   1,\n   12,\n   3\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\":

\"day\",\n   \"properties\": {\n   \"dtype\": \"number\",\n\"std\": 8,\n   \"min\": 1,\n   \"max\": 31,\n\"num_unique_values\": 31,\n   \"samples\": [\n   28,\n16,\n   24\n   ],\n   \"semantic_type\": \"\",\n\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"hour\",\n   \"properties\": {\n   \"dtype\": \"number\",\n\"std\": 6,\n   \"min\": 0,\n   \"max\": 23,\n\"num_unique_values\": 24,\n   \"samples\": [\n   8,\n16,\n   0\n   ],\n   \"semantic_type\": \"\",\n\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"PM2.5\",\n   \"properties\": {\n   \"dtype\": \"number\",\n\"std\": 80.93349731219654,\n   \"min\": 2.0,\n   \"max\":
680.0,\n   \"num_unique_values\": 557,\n   \"samples\": [\n140.0,\n   14.3,\n   129.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"PM10\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 91.57370871803138,\n
\"min\": 2.0,\n   \"max\": 999.0,\n   \"num_unique_values\":
630,\n   \"samples\": [\n   305.4,\n   45.0,\n
605.0\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"SO2\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 23.600366633044654,\n   \"min\": 1.0,\n   \"max\":
293.0,\n   \"num_unique_values\": 280,\n   \"samples\": [\n
75.0,\n   33.0,\n   142.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"NO2\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 35.15085745143049,\n
\"min\": 2.0,\n   \"max\": 270.0,\n   \"num_unique_values\":
376,\n   \"samples\": [\n   42.7024,\n   70.7,\n
58.0999\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"CO\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 1164.8549453175049,\n   \"min\": 100.0,\n
\"max\": 10000.0,\n   \"num_unique_values\": 117,\n
\"samples\": [\n   4599.0,\n   700.0,\n   200.0\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"O3\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 57.43698266437783,\n
\"min\": 0.2142,\n   \"max\": 415.0,\n
\"num_unique_values\": 767,\n   \"samples\": [\n
20.7774,\n   38.9844,\n   56.5488\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"TEMP\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 11.399096949947065,\n
\"min\": -16.8,\n   \"max\": 40.5,\n
\"num_unique_values\": 967,\n   \"samples\": [\n   18.8,\n
-10.3,\n   -1.575\n   ],\n   \"semantic_type\":
\"\",\n   \"description\": \"\"\n   }\n   },\n   {\n

\"column\": \"PRES\",\n   \"properties\": {\n   \"dtype\":
\"number\",\n   \"std\": 10.404046577854823,\n   \"min\":
985.9,\n   \"max\": 1042.0,\n   \"num_unique_values\": 600,\
n   \"samples\": [\n   998.6,\n   1029.9,\n
1024.25\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"DEWP\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 13.688896009288912,\n   \"min\": -35.3,\n
\"max\": 28.5,\n   \"num_unique_values\": 604,\n
\"samples\": [\n   -2.1,\n   27.2,\n   -30.2\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"RAIN\",\n   \"properties\":
{\n   \"dtype\": \"number\",\n   \"std\":
0.9100559139821632,\n   \"min\": 0.0,\n   \"max\": 72.5,\n
\"num_unique_values\": 127,\n   \"samples\": [\n   2.1,\n
13.8,\n   11.3\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"wd\",\n   \"properties\": {\n   \"dtype\": \"category\",\n
\"num_unique_values\": 16,\n   \"samples\": [\n
\"NNW\",\n   \"N\",\n   \"E\"\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"WSPM\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 1.2040710727929371,\n
\"min\": 0.0,\n   \"max\": 11.2,\n   \"num_unique_values\":
91,\n   \"samples\": [\n   1.9,\n   0.7,\n
3.2\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"station\",\n   \"properties\": {\n   \"dtype\":
\"category\",\n   \"num_unique_values\": 1,\n   \"samples\":
[\n   \"Guanyuan\"\n   ],\n   \"semantic_type\":
\"\",\n   \"description\": \"\"\n   }\n   }\n]\
n}","type":"dataframe","variable_name":"guanyuan_df"}

Memuat tabel kota Gucheng

```
gucheng_df =
pd.read_csv("https://raw.githubusercontent.com/marceloreis/HTI/master/
PRSA_Data_20130301-20170228/PRSA_Data_Gucheng_20130301-20170228.csv")
gucheng_df.head()
```

{"summary":"{\n \"name\": \"gucheng_df\",\n \"rows\": 35064,\n
\"fields\": [\n   {\n   \"column\": \"No\",\n   \"properties\":
{\n   \"dtype\": \"number\",\n   \"std\": 10122,\n
\"min\": 1,\n   \"max\": 35064,\n   \"num_unique_values\":
35064,\n   \"samples\": [\n   6071,\n   32601,\n
8847\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"year\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 1,\n   \"min\": 2013,\n   \"max\": 2017,\n
\"num_unique_values\": 5,\n   \"samples\": [\n   2014,\n

2017,\n   2015\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"month\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 3,\n   \"min\": 1,\n   \"max\": 12,\n
\"num_unique_values\": 12,\n   \"samples\": [\n   1,\n
12,\n   3\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"day\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 8,\n   \"min\": 1,\n   \"max\": 31,\n
\"num_unique_values\": 31,\n   \"samples\": [\n   28,\n
16,\n   24\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"hour\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 6,\n   \"min\": 0,\n   \"max\": 23,\n
\"num_unique_values\": 24,\n   \"samples\": [\n   8,\n
16,\n   0\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"PM2.5\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 82.79644460750218,\n   \"min\": 2.0,\n   \"max\":
770.0,\n   \"num_unique_values\": 577,\n   \"samples\": [\n
250.0,\n   444.0,\n   628.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"PM10\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 96.7426255992567,\n
\"min\": 2.0,\n   \"max\": 994.0,\n   \"num_unique_values\":
660,\n   \"samples\": [\n   496.0,\n   502.0,\n
293.0\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"SO2\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 21.204526026319193,\n   \"min\": 0.2856,\n
\"max\": 500.0,\n   \"num_unique_values\": 243,\n
\"samples\": [\n   62.0,\n   12.0,\n   10.8528\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"NO2\",\n   \"properties\": {\
n   \"dtype\": \"number\",\n   \"std\": 36.47386009283308,\n
\"min\": 2.0,\n   \"max\": 276.0,\n   \"num_unique_values\":
367,\n   \"samples\": [\n   56.2522,\n   94.0,\n
30.0\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"CO\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 1208.957772258247,\n   \"min\": 100.0,\n   \"max\":
10000.0,\n   \"num_unique_values\": 116,\n   \"samples\": [\
n   6700.0,\n   1000.0,\n   6900.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"O3\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 57.01958700323256,\n
\"min\": 0.2142,\n   \"max\": 450.0,\n
\"num_unique_values\": 804,\n   \"samples\": [\n
98.1036,\n   20.5632,\n   110.9556\n   ],\n

\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"TEMP\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 11.292857008072325,\n   \"min\": -15.6,\n   \"max\": 41.6,\n   \"num_unique_values\": 995,\n   \"samples\": [\n   -8.02,\n   -15.5,\n   17.75\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"PRES\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 10.103255921223534,\n   \"min\": 984.0,\n   \"max\": 1038.1,\n   \"num_unique_values\": 592,\n   \"samples\": [\n   1034.6,\n   1015.33333333333,\n   1013.0\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"DEWP\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 13.782990710571664,\n   \"min\": -34.6,\n   \"max\": 27.4,\n   \"num_unique_values\": 594,\n   \"samples\": [\n   -16.0,\n   15.1,\n   -1.2\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"RAIN\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 0.838653860679221,\n   \"min\": 0.0,\n   \"max\": 41.9,\n   \"num_unique_values\": 125,\n   \"samples\": [\n   9.3,\n   3.4,\n   12.8\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"wd\",\n   \"properties\": {\n   \"dtype\": \"category\",\n   \"num_unique_values\": 16,\n   \"samples\": [\n   \"NW\",\n   \"WNW\",\n   \"NE\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"WSPM\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 1.1510643966097016,\n   \"min\": 0.0,\n   \"max\": 12.0,\n   \"num_unique_values\": 101,\n   \"samples\": [\n   8.1,\n   5.4,\n   8.2\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"station\",\n   \"properties\": {\n   \"dtype\": \"category\",\n   \"num_unique_values\": 1,\n   \"samples\": [\n   \"Gucheng\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   }\n]\n}","type":"dataframe","variable_name":"gucheng_df"}

Assessing Data

Menilai data Guanyuan

Menghitung jumlah data yang tidak lengkap

```python
guanyuan_df.info()
guanyuan_df.isna().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35064 entries, 0 to 35063
Data columns (total 18 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   No       35064 non-null  int64
 1   year     35064 non-null  int64
 2   month    35064 non-null  int64
 3   day      35064 non-null  int64
 4   hour     35064 non-null  int64
 5   PM2.5    34448 non-null  float64
 6   PM10     34635 non-null  float64
 7   SO2      34590 non-null  float64
 8   NO2      34405 non-null  float64
 9   CO       33311 non-null  float64
 10  O3       33891 non-null  float64
 11  TEMP     35044 non-null  float64
 12  PRES     35044 non-null  float64
 13  DEWP     35044 non-null  float64
 14  RAIN     35044 non-null  float64
 15  wd       34983 non-null  object
 16  WSPM     35050 non-null  float64
 17  station  35064 non-null  object
dtypes: float64(11), int64(5), object(2)
memory usage: 4.8+ MB

No            0
year          0
month         0
day           0
hour          0
PM2.5       616
PM10        429
SO2         474
NO2         659
CO         1753
O3         1173
TEMP         20
PRES         20
DEWP         20
RAIN         20
wd           81
WSPM         14
station       0
dtype: int64
```

Terlihat bahwa terdapat beberapa data yang kosong pada column PM2.5, PM10, SO22, NO2, CO, O3, TEMP, PRES, DEWP, RAIN, wd dan WSPM

Menampilkan parameter statistik tabel Guanyuan

```
guanyuan_df.describe()
```

{"summary":"{\n \"name\": \"guanyuan_df\",\n \"rows\": 8,\n
\"fields\": [\n   {\n   \"column\": \"No\",\n   \"properties\":
{\n   \"dtype\": \"number\",\n   \"std\":
12636.930448670224,\n   \"min\": 1.0,\n   \"max\": 35064.0,\
n   \"num_unique_values\": 6,\n   \"samples\": [\n
35064.0,\n   17532.5,\n   26298.25\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"year\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 11807.362916902399,\n
\"min\": 1.1772134318198135,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
2014.662559890486,\n   2015.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"month\",\n   \"properties\": {\
n   \"dtype\": \"number\",\n   \"std\": 12394.775690880602,\
n   \"min\": 1.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
6.522929500342231,\n   7.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"day\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12391.770525834587,\n
\"min\": 1.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
15.729637234770705,\n   16.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"hour\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12393.163442376766,\n
\"min\": 0.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 7,\n   \"samples\": [\n   35064.0,\
n   11.5,\n   17.25\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"PM2.5\",\n   \"properties\": {\
n   \"dtype\": \"number\",\n   \"std\": 12128.523448483706,\
n   \"min\": 2.0,\n   \"max\": 34448.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
82.93337203901532,\n   59.0,\n   34448.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"PM10\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12174.91689123717,\n
\"min\": 2.0,\n   \"max\": 34635.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
109.02330301717916,\n   89.0,\n   34635.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"SO2\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12211.212534718983,\n
\"min\": 1.0,\n   \"max\": 34590.0,\n

\"num_unique_values\": 8,\n   \"samples\": [\n 17.59094149754264,\n   8.0,\n   34590.0\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\ n   },\n   {\n   \"column\": \"NO2\",\n   \"properties\": {\n \"dtype\": \"number\",\n   \"std\": 12137.76628206752,\n \"min\": 2.0,\n   \"max\": 34405.0,\n \"num_unique_values\": 8,\n   \"samples\": [\n 57.901642517076006,\n   51.0,\n   34405.0\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\ n   },\n   {\n   \"column\": \"CO\",\n   \"properties\": {\n \"dtype\": \"number\",\n   \"std\": 11451.576920854763,\n \"min\": 100.0,\n   \"max\": 33311.0,\n \"num_unique_values\": 8,\n   \"samples\": [\n 1271.294377232746,\n   900.0,\n   33311.0\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\ n   },\n   {\n   \"column\": \"O3\",\n   \"properties\": {\n \"dtype\": \"number\",\n   \"std\": 11949.820184258542,\n \"min\": 0.2142,\n   \"max\": 33891.0,\n \"num_unique_values\": 8,\n   \"samples\": [\n 55.79504428314302,\n   41.0,\n   33891.0\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\ n   },\n   {\n   \"column\": \"TEMP\",\n   \"properties\": {\n \"dtype\": \"number\",\n   \"std\": 12385.411047634085,\n \"min\": -16.8,\n   \"max\": 35044.0,\n \"num_unique_values\": 8,\n   \"samples\": [\n 13.584607417568478,\n   14.5,\n   35044.0\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\ n   },\n   {\n   \"column\": \"PRES\",\n   \"properties\": {\n \"dtype\": \"number\",\n   \"std\": 12087.68535182901,\n \"min\": 10.404046577854823,\n   \"max\": 35044.0,\n \"num_unique_values\": 8,\n   \"samples\": [\n 1011.8469198246108,\n   1011.4,\n   35044.0\n   ],\ n   \"semantic_type\": \"\",\n   \"description\": \"\"\n }\n   },\n   {\n   \"column\": \"DEWP\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12388.863098113037,\n   \"min\": -35.3,\n   \"max\": 35044.0,\n   \"num_unique_values\": 8,\n   \"samples\": [\n 3.1230624357950005,\n   3.8,\n   35044.0\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\ n   },\n   {\n   \"column\": \"RAIN\",\n   \"properties\": {\n \"dtype\": \"number\",\n   \"std\": 12386.239715567051,\n \"min\": 0.0,\n   \"max\": 35044.0,\n \"num_unique_values\": 5,\n   \"samples\": [\n 0.06742095651181372,\n   72.5,\n   0.9100559139821632\n],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n }\n   },\n   {\n   \"column\": \"WSPM\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12391.106770540331,\n   \"min\": 0.0,\n   \"max\": 35050.0,\ n   \"num_unique_values\": 8,\n   \"samples\": [\n

```
1.708496433666191,\n               1.4,\n               35050.0\n        ],\n
\"semantic_type\": \"\",\n          \"description\": \"\"\n        }\
n     }\n  ]\n}","type":"dataframe"}
```

Dapat dilihat terdapat banyak data kosong pada kolom rain yang ditandai dengan median dan quartil bawah / atasnya bernilai 0

Menampilkan jumlah data duplikat pada tabel Guanyuan

```
print("Jumlah duplikasi: ", guanyuan_df.duplicated().sum())
```

```
Jumlah duplikasi:  0
```

Menilai data Gucheng

Menghitung jumlah data yang tidak lengkap

```
gucheng_df.info()
gucheng_df.isna().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35064 entries, 0 to 35063
Data columns (total 18 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   No       35064 non-null  int64
 1   year     35064 non-null  int64
 2   month    35064 non-null  int64
 3   day      35064 non-null  int64
 4   hour     35064 non-null  int64
 5   PM2.5    34418 non-null  float64
 6   PM10     34683 non-null  float64
 7   SO2      34557 non-null  float64
 8   NO2      34396 non-null  float64
 9   CO       33663 non-null  float64
 10  O3       34335 non-null  float64
 11  TEMP     35013 non-null  float64
 12  PRES     35014 non-null  float64
 13  DEWP     35013 non-null  float64
 14  RAIN     35021 non-null  float64
 15  wd       34905 non-null  object
 16  WSPM     35022 non-null  float64
 17  station  35064 non-null  object
dtypes: float64(11), int64(5), object(2)
memory usage: 4.8+ MB

No            0
year          0
month         0
day           0
```

```
hour            0
PM2.5         646
PM10          381
SO2           507
NO2           668
CO           1401
O3            729
TEMP           51
PRES           50
DEWP           51
RAIN           43
wd            159
WSPM           42
station         0
dtype: int64
```

Terlihat bahwa terdapat beberapa data yang kosong pada column PM2.5, PM10, SO22, NO2, CO, O3, TEMP, PRES, DEWP, RAIN, wd dan WSPM

Menampilkan parameter statistik tabel Gucheng

```
gucheng_df.describe()
```

{"summary":"{\n \"name\": \"gucheng_df\",\n \"rows\": 8,\n \"fields\": [\n   {\n   \"column\": \"No\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12636.930448670224,\n   \"min\": 1.0,\n   \"max\": 35064.0,\n   \"num_unique_values\": 6,\n   \"samples\": [\n 35064.0,\n   17532.5,\n   26298.25\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"year\",\n   \"properties\": {\n \"dtype\": \"number\",\n   \"std\": 11807.362916902399,\n \"min\": 1.1772134318198135,\n   \"max\": 35064.0,\n \"num_unique_values\": 8,\n   \"samples\": [\n 2014.662559890486,\n   2015.0,\n   35064.0\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"month\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12394.775690880602,\n   \"min\": 1.0,\n   \"max\": 35064.0,\n \"num_unique_values\": 8,\n   \"samples\": [\n 6.522929500342231,\n   7.0,\n   35064.0\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"day\",\n   \"properties\": {\n \"dtype\": \"number\",\n   \"std\": 12391.770525834587,\n \"min\": 1.0,\n   \"max\": 35064.0,\n \"num_unique_values\": 8,\n   \"samples\": [\n 15.729637234770705,\n   16.0,\n   35064.0\n   ],\n \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"hour\",\n   \"properties\": {\n \"dtype\": \"number\",\n   \"std\": 12393.163442376766,\n
```

\"min\": 0.0,\n          \"max\": 35064.0,\n
\"num_unique_values\": 7,\n          \"samples\": [\n          35064.0,\
n          11.5,\n          17.25\n          ],\n
\"semantic_type\": \"\",\n          \"description\": \"\"\n     }\
n     },\n     {\n       \"column\": \"PM2.5\",\n       \"properties\": {\
n        \"dtype\": \"number\",\n        \"std\": 12113.69298356956,\n
\"min\": 2.0,\n          \"max\": 34418.0,\n
\"num_unique_values\": 8,\n          \"samples\": [\n
83.85208902318554,\n          60.0,\n          34418.0\n          ],\n
\"semantic_type\": \"\",\n          \"description\": \"\"\n      }\
n     },\n     {\n       \"column\": \"PM10\",\n       \"properties\": {\n
\"dtype\": \"number\",\n         \"std\": 12189.609191626367,\n
\"min\": 2.0,\n          \"max\": 34683.0,\n
\"num_unique_values\": 8,\n          \"samples\": [\n
118.86197849090333,\n          99.0,\n          34683.0\n          ],\n
\"semantic_type\": \"\",\n          \"description\": \"\"\n      }\
n     },\n     {\n       \"column\": \"SO2\",\n       \"properties\": {\n
\"dtype\": \"number\",\n         \"std\": 12190.368028002727,\n
\"min\": 0.2856,\n       \"max\": 34557.0,\n
\"num_unique_values\": 8,\n          \"samples\": [\n
15.366161622826054,\n          7.0,\n          34557.0\n          ],\n
\"semantic_type\": \"\",\n          \"description\": \"\"\n      }\
n     },\n     {\n       \"column\": \"NO2\",\n       \"properties\": {\n
\"dtype\": \"number\",\n         \"std\": 12134.586480743063,\n
\"min\": 2.0,\n          \"max\": 34396.0,\n
\"num_unique_values\": 8,\n          \"samples\": [\n
55.87107495348296,\n          50.0,\n          34396.0\n          ],\n
\"semantic_type\": \"\",\n          \"description\": \"\"\n      }\
n     },\n     {\n       \"column\": \"CO\",\n       \"properties\": {\n
\"dtype\": \"number\",\n         \"std\": 11558.282691821732,\n
\"min\": 100.0,\n        \"max\": 33663.0,\n
\"num_unique_values\": 8,\n          \"samples\": [\n
1323.9744229569558,\n          900.0,\n          33663.0\n        ],\n
\"semantic_type\": \"\",\n          \"description\": \"\"\n      }\
n     },\n     {\n       \"column\": \"O3\",\n       \"properties\": {\n
\"dtype\": \"number\",\n         \"std\": 12104.622867372267,\n
\"min\": 0.2142,\n       \"max\": 34335.0,\n
\"num_unique_values\": 8,\n          \"samples\": [\n
57.694878835008005,\n          45.0,\n          34335.0\n        ],\n
\"semantic_type\": \"\",\n          \"description\": \"\"\n      }\
n     },\n     {\n       \"column\": \"TEMP\",\n       \"properties\": {\
n \"dtype\": \"number\",\n        \"std\": 12374.275383198543,\n
\"min\": -15.6,\n        \"max\": 35013.0,\n
\"num_unique_values\": 8,\n          \"samples\": [\n
13.864524478028411,\n          14.8,\n          35013.0\n          ],\n
\"semantic_type\": \"\",\n          \"description\": \"\"\n      }\
n     },\n     {\n       \"column\": \"PRES\",\n       \"properties\": {\n
\"dtype\": \"number\",\n         \"std\": 12077.960428885428,\n
\"min\": 10.103255921223534,\n          \"max\": 35014.0,\n

```
\"num_unique_values\": 8,\n          \"samples\": [\n
1008.8295917481486,\n          1008.5,\n          35014.0\n          ],\
n       \"semantic_type\": \"\",\n          \"description\": \"\"\n
}\n     },\n    {\n       \"column\": \"DEWP\",\n      \"properties\":
{\n         \"dtype\": \"number\",\n          \"std\":
12378.039705764257,\n          \"min\": -34.6,\n          \"max\":
35013.0,\n          \"num_unique_values\": 8,\n          \"samples\": [\n
2.6104418358895276,\n          3.0,\n          35013.0\n          ],\n
\"semantic_type\": \"\",\n          \"description\": \"\"\n       }\
n     },\n    {\n       \"column\": \"RAIN\",\n      \"properties\": {\n
\"dtype\": \"number\",\n          \"std\": 12379.640032154402,\n
\"min\": 0.0,\n          \"max\": 35021.0,\n
\"num_unique_values\": 5,\n          \"samples\": [\n
0.06445275691727821,\n          41.9,\n          0.838653860679221\n
],\n       \"semantic_type\": \"\",\n          \"description\": \"\"\n
}\n     },\n    {\n       \"column\": \"WSPM\",\n      \"properties\":
{\n         \"dtype\": \"number\",\n          \"std\":
12381.243654409724,\n          \"min\": 0.0,\n          \"max\": 35022.0,\
n          \"num_unique_values\": 8,\n          \"samples\": [\n
1.3433099194791847,\n          1.0,\n          35022.0\n          ],\n
\"semantic_type\": \"\",\n          \"description\": \"\"\n       }\
n     }\n   ]\n}","type":"dataframe"}
```

Menampilkan jumlah data duplikat pada tabel Gucheng

```python
print("Jumlah duplikasi: ", gucheng_df.duplicated().sum())
```

```
Jumlah duplikasi:  0
```

Tidak terdapat data duplikat pada tabel tersebut

Dapat dilihat terdapat banyak data kosong pada kolom rain yang ditandai dengan median dan quartil bawah / atasnya bernilai 0

Cleaning Data

Karena kedua data berupa time series, maka data yang dianalisa harus utuh sehingga tidak dapat menggunakan drop. Maka untuk mengatasi data yang hilang digunakan metode interpolation

```python
guanyuan_df.interpolate(method='linear',limit_direction='forward',inplace = True)
gucheng_df.interpolate(method='linear',limit_direction='forward',inplace = True)
```

Tidak ada data yang duplikat pada kedua tabel. Data yang tidak lengkap hanya terdapat pada kolom rain yang tidak digunakan dalam analisis data.

Exploratory Data Analysis (EDA)

Explore ...

Melihat rangkuman parameter statistik menggunakan describe

```
guanyuan_df.describe(include="all")
```

{"summary":"{\n \"name\": \"guanyuan_df\",\n \"rows\": 11,\n \"fields\": [\n   {\n   \"column\": \"No\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12636.930448670224,\n   \"min\": 1.0,\n   \"max\": 35064.0,\n   \"num_unique_values\": 6,\n   \"samples\": [\n   35064.0,\n   17532.5,\n   26298.25\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"year\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 11807.362916902399,\n   \"min\": 1.1772134318198135,\n   \"max\": 35064.0,\n   \"num_unique_values\": 8,\n   \"samples\": [\n   2014.662559890486,\n   2015.0,\n   35064.0\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"month\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12394.775690880602,\n   \"min\": 1.0,\n   \"max\": 35064.0,\n   \"num_unique_values\": 8,\n   \"samples\": [\n   6.522929500342231,\n   7.0,\n   35064.0\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"day\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12391.770525834587,\n   \"min\": 1.0,\n   \"max\": 35064.0,\n   \"num_unique_values\": 8,\n   \"samples\": [\n   15.729637234770705,\n   16.0,\n   35064.0\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"hour\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12393.163442376766,\n   \"min\": 0.0,\n   \"max\": 35064.0,\n   \"num_unique_values\": 7,\n   \"samples\": [\n   35064.0,\n   11.5,\n   17.25\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"PM2.5\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12346.322972172586,\n   \"min\": 2.0,\n   \"max\": 35064.0,\n   \"num_unique_values\": 8,\n   \"samples\": [\n   82.89752167465207,\n   59.0,\n   35064.0\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"PM10\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12326.480225532672,\n   \"min\": 2.0,\n   \"max\": 35064.0,\n   \"num_unique_values\": 8,\n   \"samples\": [\n

109.37234913301391,\n   89.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"SO2\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12378.788944684795,\n
\"min\": 1.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
17.609312194843717,\n   8.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"NO2\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12370.723959287741,\n
\"min\": 2.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
58.139297945185945,\n   51.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"CO\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12054.999033976908,\n
\"min\": 100.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
1258.3270448323067,\n   900.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"O3\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12364.728431222053,\n
\"min\": 0.2142,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
54.817142160050196,\n   40.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"TEMP\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12392.482203450325,\n
\"min\": -16.8,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
13.581414016976662,\n   14.5,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"PRES\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12094.753180472728,\n
\"min\": 10.404517471480327,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
1011.8516500779623,\n   1011.4,\n   35064.0\n   ],\
n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"DEWP\",\n   \"properties\":
{\n   \"dtype\": \"number\",\n   \"std\":
12395.934225953106,\n   \"min\": -35.3,\n   \"max\":
35064.0,\n   \"num_unique_values\": 8,\n   \"samples\": [\n
3.120295744923568,\n   3.8,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"RAIN\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12393.310783634011,\n
\"min\": 0.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 5,\n   \"samples\": [\n
0.06738250057038557,\n   72.5,\n   0.909797752298308\n

],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"wd\",\n   \"properties\": {\n   \"dtype\": \"category\",\n   \"num_unique_values\": 4,\n   \"samples\": [\n   16,\n   \"5140\",\n   \"34983\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\":\n   \"WSPM\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12396.0565081987,\n   \"min\": 0.0,\n   \"max\":\n   35064.0,\n   \"num_unique_values\": 8,\n   \"samples\": [\n   1.7085771731690622,\n   1.4,\n   35064.0\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"station\",\n   \"properties\":\n   {\n   \"dtype\": \"category\",\n   \"num_unique_values\":\n   3,\n   \"samples\": [\n   \"35064\",\n   1,\n   \"Guanyuan\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   }\n   ]\n}","type":"dataframe"}

```
gucheng_df.describe(include="all")
```

{"summary":"{\n \"name\": \"gucheng_df\",\n \"rows\": 11,\n
\"fields\": [\n   {\n   \"column\": \"No\",\n   \"properties\":
{\n   \"dtype\": \"number\",\n   \"std\":
12636.930448670224,\n   \"min\": 1.0,\n   \"max\": 35064.0,\
n   \"num_unique_values\": 6,\n   \"samples\": [\n
35064.0,\n   17532.5,\n   26298.25\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"year\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 11807.362916902399,\n
\"min\": 1.1772134318198135,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
2014.662559890486,\n   2015.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"month\",\n   \"properties\": {\
n   \"dtype\": \"number\",\n   \"std\": 12394.775690880602,\
n   \"min\": 1.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
6.522929500342231,\n   7.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"day\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12391.770525834587,\n
\"min\": 1.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
15.729637234770705,\n   16.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"hour\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12393.163442376766,\n
\"min\": 0.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 7,\n   \"samples\": [\n   35064.0,\
n   11.5,\n   17.25\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\

n   },\n   {\n   \"column\": \"PM2.5\",\n   \"properties\": {\
n   \"dtype\": \"number\",\n   \"std\": 12342.006141102263,\
n   \"min\": 2.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
84.07480179101073,\n   60.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"PM10\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12324.153392354208,\n
\"min\": 2.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
119.26156741957564,\n   100.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"SO2\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12369.478387183306,\n
\"min\": 0.2856,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
15.705820887519966,\n   7.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"NO2\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12363.682565745099,\n
\"min\": 2.0,\n   \"max\": 35044.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
55.85131964958338,\n   50.0,\n   35044.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"CO\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12023.861261682056,\n
\"min\": 100.0,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
1338.100059890486,\n   985.7142857142858,\n   35064.0\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"O3\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12362.27111519826,\n
\"min\": 0.2142,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
58.000000866986085,\n   45.0,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"TEMP\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12392.306924014538,\n
\"min\": -15.6,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
13.855824022278949,\n   14.8,\n   35064.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"PRES\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 12095.630322460285,\n
\"min\": 10.104301856861493,\n   \"max\": 35064.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
1008.8356926040858,\n   1008.5,\n   35064.0\n   ],\
n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"DEWP\",\n   \"properties\":

{\n   \"dtype\": \"number\",\n   \"std\": 12396.07612358495,\n   \"min\": -34.6,\n   \"max\": 35064.0,\n   \"num_unique_values\": 8,\n   \"samples\": [\n   2.6003507871321023,\n   3.0,\n   35064.0\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"RAIN\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12394.842847229575,\n   \"min\": 0.0,\n   \"max\": 35064.0,\n   \"num_unique_values\": 5,\n   \"samples\": [\n   0.06437371663244354,\n   41.9,\n   0.838142490923598\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"wd\",\n   \"properties\": {\n   \"dtype\": \"category\",\n   \"num_unique_values\": 4,\n   \"samples\": [\n   16,\n   \"3949\",\n   \"34905\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"WSPM\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 12396.092878501477,\n   \"min\": 0.0,\n   \"max\": 35064.0,\n   \"num_unique_values\": 8,\n   \"samples\": [\n   1.3436672940908054,\n   1.0,\n   35064.0\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"station\",\n   \"properties\": {\n   \"dtype\": \"category\",\n   \"num_unique_values\": 3,\n   \"samples\": [\n   \"35064\",\n   1,\n   \"Gucheng\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   }\n]\n}","type":"dataframe"}

Mengelompokkan kadar rata-rata gas O3 kota Gucheng per tahun

```
gucheng_df.groupby(by="year")["O3"].mean()

year
2013    55.781492
2014    60.253139
2015    60.779538
2016    59.522883
2017    28.924788
Name: O3, dtype: float64
```

Mengelompokkan rata-rata suhu kota Gucheng per tahun

```
gucheng_df.groupby(by="year")["TEMP"].mean()

year
2013    15.932823
2014    14.431661
2015    14.027483
2016    13.547228
2017     0.373600
Name: TEMP, dtype: float64
```

Terlihat bahwa kadar O3 tertinggi ada pada tahun 2015, sedangkan, sedangkan suhu rata-rata tertinggi berada pada tahun 2013. Kadar O3 dan suhu terrendah ada pada tahun 2017

Mengelompokkan produksi rata-rata Gas SO2, NO2, dan CO Guanyuan berdasarkan tahun

```
guanyuan_df.groupby(by="year")["SO2"].mean()

year
2013    21.859392
2014    23.133562
2015    14.073579
2016    11.484517
2017    21.259181
Name: SO2, dtype: float64

guanyuan_df.groupby(by="year")["NO2"].mean()

year
2013    58.866182
2014    64.343179
2015    54.921958
2016    53.459984
2017    64.920904
Name: NO2, dtype: float64

guanyuan_df.groupby(by="year")["CO"].mean()

year
2013    1132.745030
2014    1362.357306
2015    1313.704338
2016    1168.169399
2017    1482.768362
Name: CO, dtype: float64
```

Mengelompokkan produksi rata-rata Gas SO2, NO2, dan CO Gucheng berdasarkan tahun

```
gucheng_df.groupby(by="year")["SO2"].mean()

year
2013    17.508021
2014    21.290103
2015    13.338265
2016    10.204121
2017    20.587924
Name: SO2, dtype: float64

gucheng_df.groupby(by="year")["NO2"].mean()
```

```
year
2013    61.514691
2014    62.424463
2015    49.588271
2016    49.777265
2017    62.319915
Name: NO2, dtype: float64

gucheng_df.groupby(by="year")["CO"].mean()

year
2013    1341.740264
2014    1442.163242
2015    1323.144977
2016    1170.952869
2017    1804.837571
Name: CO, dtype: float64
```

Visualization & Explanatory Analysis

Pertanyaan 1:

```
sampled_data = gucheng_df.sample(n=1000)
plt.hexbin(sampled_data['O3'],sampled_data['TEMP'],  gridsize=50,
cmap='Blues')
plt.colorbar(label='count')
plt.title('Hexbin Plot of Ozone Levels and Temperature')
plt.xlabel('Ozone Levels')
plt.ylabel('Temperature')
plt.show()
```

Hexbin Plot of Ozone Levels and Temperature
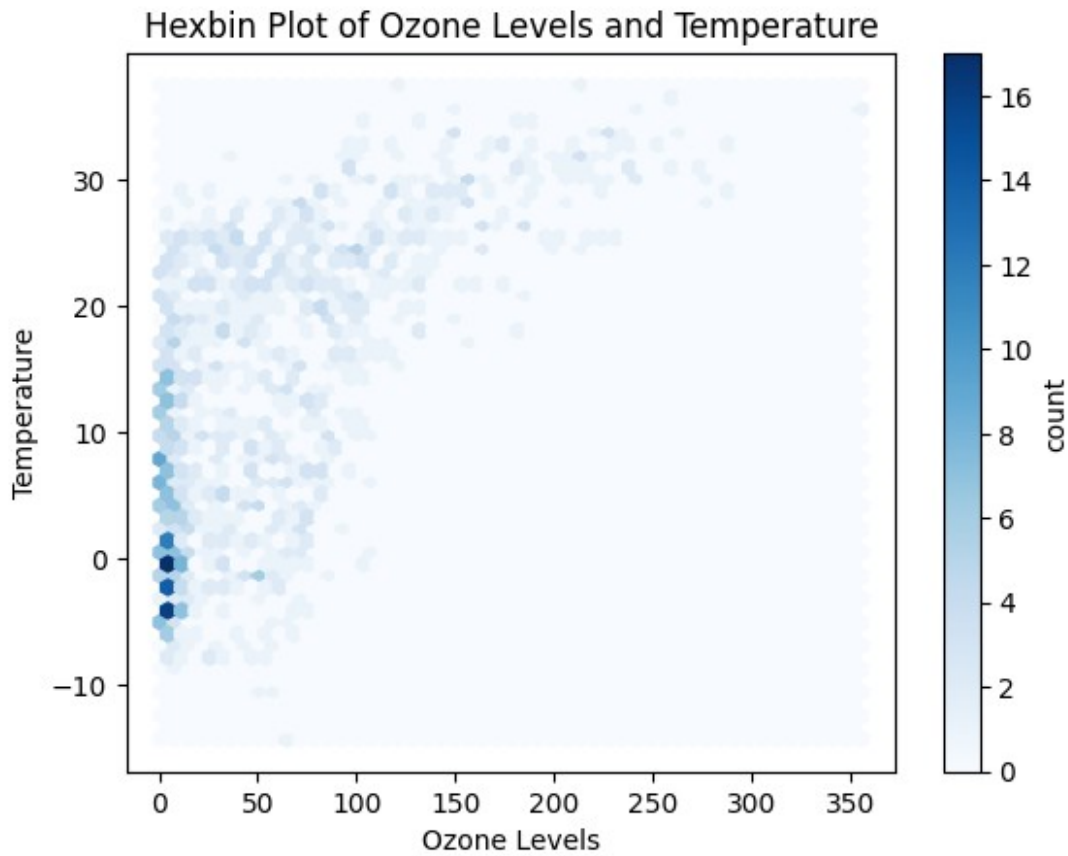

```
temperature = sampled_data['TEMP']
ozone_levels = sampled_data['O3']

correlation_coefficient = np.corrcoef( ozone_levels,temperature)[0, 1]

print("Correlation Coefficient:", correlation_coefficient)

Correlation Coefficient: 0.6300191856439685
```

Dari scatterplot dan nilai koefisien, maka bisa ditarik kesimpulan bahwa terdapat moderate positive correlation antara kadar Ozon dan Temperatur di kota tersebut. Hal ini berarti dengan adanya peningkatan kadar ozon, maka juga diikuti dengan peningkatan suhu udara dan sebaliknya.

Pertanyaan 2:

```
guanyuan_so2_mean = guanyuan_df.groupby(by="year")["SO2"].mean()
gucheng_so2_mean = gucheng_df.groupby(by="year")["SO2"].mean()

plt.plot(guanyuan_so2_mean.index, guanyuan_so2_mean.values,
label='Guanyuan', marker='o')
plt.plot(gucheng_so2_mean.index, gucheng_so2_mean.values,
label='Gucheng', marker='o')
```

```python
plt.xlabel('Year')
plt.ylabel('Average SO2 Production')
plt.title('Comparison of Average SO2 Production Between Guanyuan and
Gucheng')
plt.legend()

plt.show()
```


Comparison of Average SO2 Production Between Guanyuan and Gucheng
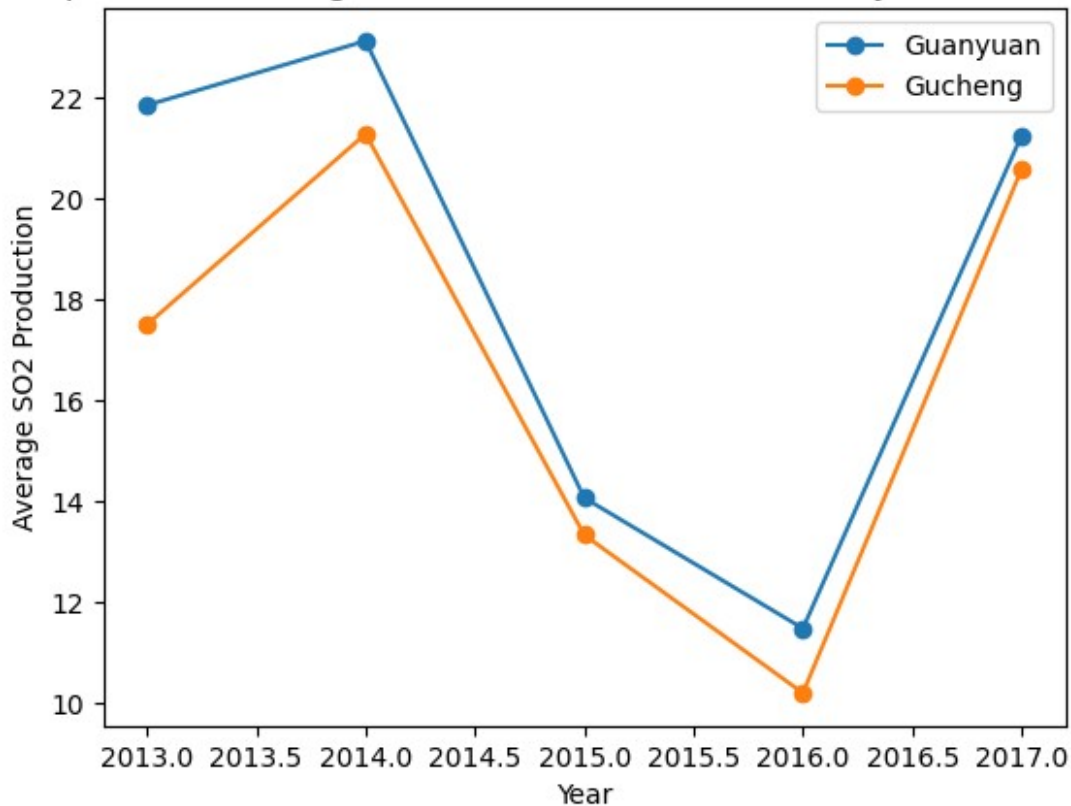
```python
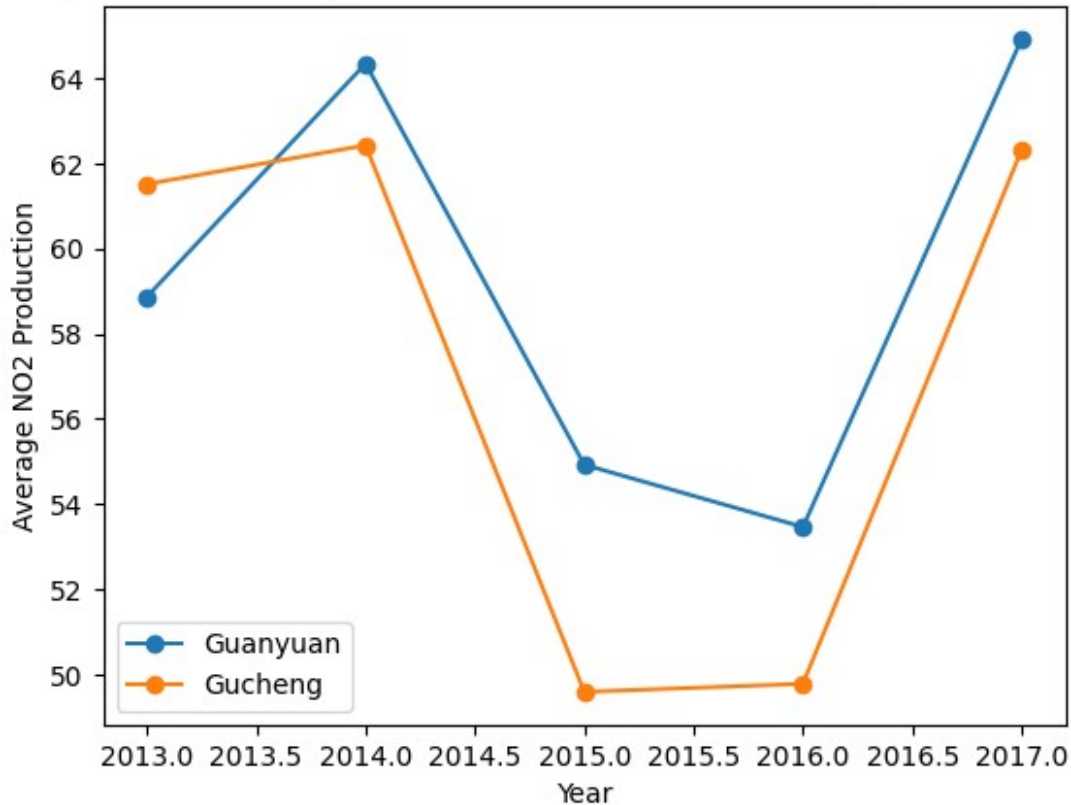guanyuan_no2_mean = guanyuan_df.groupby(by="year")["NO2"].mean()
gucheng_no2_mean = gucheng_df.groupby(by="year")["NO2"].mean()

plt.plot(guanyuan_no2_mean.index, guanyuan_no2_mean.values,
label='Guanyuan', marker='o')
plt.plot(gucheng_no2_mean.index, gucheng_no2_mean.values,
label='Gucheng', marker='o')

plt.xlabel('Year')
plt.ylabel('Average NO2 Production')
plt.title('Comparison of Average NO2 Production Between Guanyuan and
Gucheng')
plt.legend()
```

```
plt.show()
```


Comparison of Average NO2 Production Between Guanyuan and Gucheng

```
guanyuan_co_mean = guanyuan_df.groupby(by="year")["CO"].mean()
gucheng_co_mean = gucheng_df.groupby(by="year")["CO"].mean()

plt.plot(guanyuan_co_mean.index, guanyuan_co_mean.values,
label='Guanyuan', marker='o')
plt.plot(gucheng_co_mean.index, gucheng_co_mean.values,
label='Gucheng', marker='o')

plt.xlabel('Year')
plt.ylabel('Average CO Production')
plt.title('Comparison of Average CO Production Between Guanyuan and
Gucheng')
plt.legend()

plt.show()
```

Comparison of Average CO Production Between Guanyuan and Gucheng
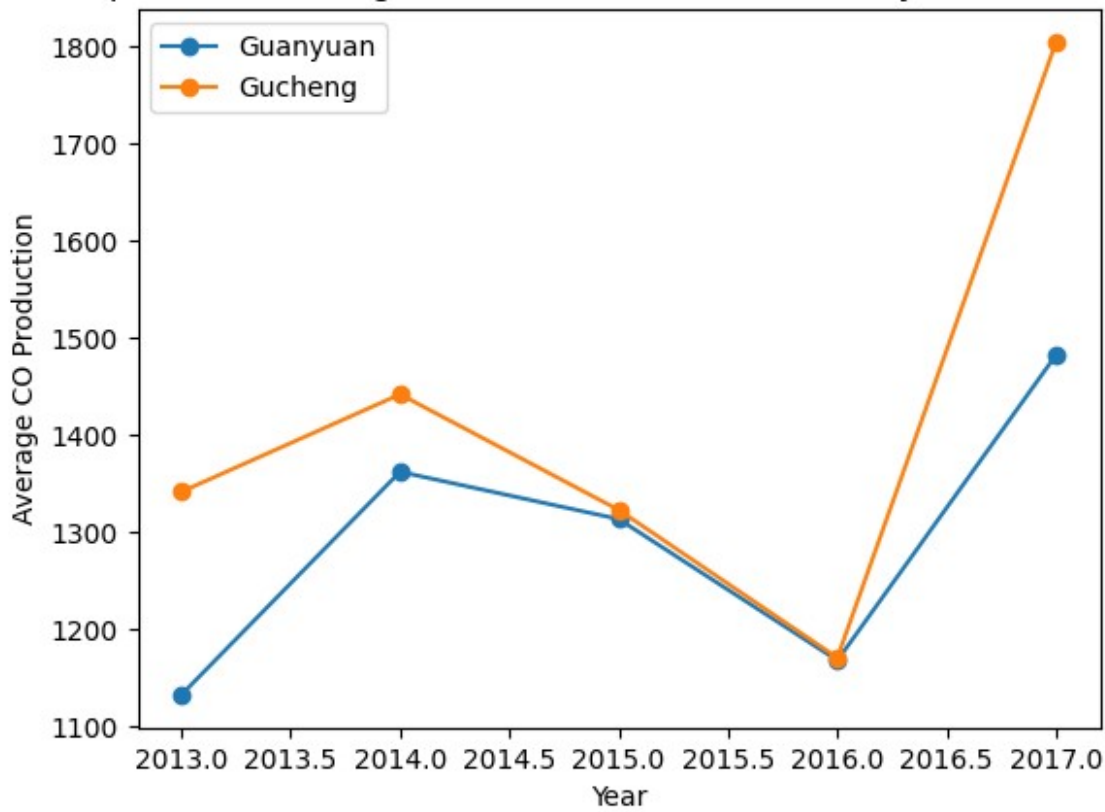
Dari ketiga grafik, dapat dilihat perbandingan produksi gas polutan rata-rata per tahun dari kedua kota. Kesimpulan yang bisa ditarik dari ketiga grafik adalah sebagai berikut:

- Untuk gas SO2, kota Guanyuan memiliki produksi rata-rata yang lebih banyak dibanding Gucheng.
- Untuk Gas NO2, kota Gucheng memiliki produksi rata-rata yang lebih banyak dibanding Guanyuan pada 2013, namun menurun dan lebih rendah dibanding Guanyuan untuk tahun-tahun selanjutnya

- Untuk Gas CO, kota Gucheng memiliki produksi yang lebih banyak dibanding Guanyuan pada tahun 2013-2015, dan menurun pada tahun 2015-2016 hingga mendekat dengan produksi rata-rata Guanyuan, sebelum menjulang tinggi pada tahun 2016

Perbandingan kualitas udara pada kota tersebut dapat dihubungkan pada kandungan gas polutan yang diukur, sehingga dapat menjadi referensi dalam menjaga kualitas udara kedua kota tersebut.

Conclusion
- Terdapat korelasi sebesar 63% diantara kadar ozon dan suhu udara kota Guanyuan
- Rata-rata kadar gas SO2 dan CO pada 2016 di Guanyuan lebih besar dibanding Gucheng, sedangkan rata-rata kadar NO2 dan CO pada 2013-2015 lebih banyak dibanding Guanyuan