ANALISIS DATA KUALITAS UDARA DI STASIUN JAKARTA TAHUN 2020

KAPITA SELEKTA STATISTIKA

KELAS A

KELOMPOK MURID PROF

KENNETH HOLIVIANTO	**6162001018**
JASON HINARDI	**6162001022**
BRYAN ERNESTIN	**6162001097**
LEONARDO ALINDRA	**6162001111**

UNIVERSITAS KATOLIK PARAHYANGAN

BANDUNG

2023

BAB I

PENDAHULUAN

1.1 Latar Belakang

Kualitas udara yang kita hirup akan memengaruhi kesehatan orang-orang yang berada di daerah tersebut. Namun, polusi udara akan selalu ada sehingga menurunkan kualitas udara yang dihirup. Polutan udara meliputi berbagai senyawa kimia yang berbahaya bagi manusia, diantaranya adalah sulfur dioksida, karbon monoksida, dan nitrogen dioksida. Karena polutan ini berbahaya, sebaiknya konsentrasi polutan ini dipantau dan dijaga agar konsentrasinya di udara rendah.

Dalam penelitian ini, akan dilakukan analisis kualitas udara pada 5 stasiun di Jakarta. Dengan mencari tahu ISPU DKI Jakarta, bisa dilihat seberapa baik atau buruknya pencemaran udara yang telah terjadi di DKI Jakarta, sehingga bisa menunjukkan adanya urgensi bagi pemerintah untuk mengeluarkan kebijakan demi mengurangi polusi udara di DKI Jakarta bila dinilai kurang baik atau mempertahankan kebersihan udara bila sudah cukup baik.

1.2 Rumusan Masalah

Adapun rumusan masalah untuk karya tulis ini adalah sebagai berikut:

- Bagaimana kualitas udara pada setiap stasiun DKI Jakarta tahun 2020?
- Polutan apa yang memiliki ISPU tertinggi di stasiun DKI Jakarta tahun 2020?
- Bagaimana tren nilai rata-rata ISPU polutan berdasarkan tanggal yang sama pada tahun 2020?
- Bagaimana rata-rata ISPU untuk masing-masing jenis polutan berdasarkan kategori?
- Polutan apa yang paling kritis setiap bulannya di stasiun DKI Jakarta tahun 2020?

1.3 Tujuan Penelitian

Adapun tujuan penelitian untuk karya tulis ini adalah sebagai berikut:

- Mengetahui kualitas udara pada setiap stasiun DKI Jakarta tahun 2020.
- Mengetahui polutan yang memiliki ISPU tertinggi di stasiun DKI Jakarta tahun 2020.
- Mengetahui tren nilai rata-rata ISPU polutan berdasarkan tanggal yang sama pada tahun 2020.
- Mengetahui rata-rata ISPU untuk masing-masing jenis polutan berdasarkan kategori.
- Mengetahui polutan yang paling kritis setiap bulannya di stasiun DKI Jakarta tahun 2020.

BAB II
PEMBAHASAN

2.1 Dataset

Tabel di bawah merupakan sepuluh baris teratas dari dataset yang digunakan untuk analisis ISPU di 5 stasiun DKI Jakarta dan bersumber dari Jakarta Open Data.

tanggal	stasiun	pm10	so2	co	o3	no2	max	critical	categori
1/1/2020	DKI1 (Bunderan HI)	30	20	10	32	9	32	O3	BAIK
2/1/2020	DKI1 (Bunderan HI)	27	22	12	29	8	29	O3	BAIK
3/1/2020	DKI1 (Bunderan HI)	39	22	14	32	10	39	PM10	BAIK
4/1/2020	DKI1 (Bunderan HI)	34	22	14	38	10	38	O3	BAIK
5/1/2020	DKI1 (Bunderan HI)	35	22	12	31	9	35	PM10	BAIK
6/1/2020	DKI1 (Bunderan HI)	46	23	16	32	9	46	PM10	BAIK
7/1/2020	DKI1 (Bunderan HI)	37	23	26	33	11	37	PM10	BAIK
8/1/2020	DKI1 (Bunderan HI)	41	26	20	30	11	41	PM10	BAIK
9/1/2020	DKI1 (Bunderan HI)	52	23	29	24	12	52	PM10	SEDANG
10/1/2020	DKI1 (Bunderan HI)	24	24	18	25	8	25	O3	BAIK

Dataset di atas berisikan Indeks Standar Pencemaran Udara (ISPU) dari lima stasiun DKI Jakarta selama tahun 2020, yaitu Bunderan HI, Kelapa Gading, Jagakarsa, Lubang Buaya, dan Kebon Jeruk. Dataset tersebut terdiri dari 1830 baris dan 10 kolom, yaitu:

1. tanggal : tanggal pengukuran kualitas udara
2. stasiun : nama stasiun
3. pm10 : nilai indeks dari partikel udara berukuran lebih kecil dari 10 mikron
4. so2 : nilai indeks dari sulfur dioksida
5. co : nilai indeks dari karbon monoksida
6. o3 : nilai indeks dari ozon
7. no2 : nilai indeks dari nitrogen dioksida
8. max : nilai indeks terbesar dalam suatu baris data
9. critical : jenis polutan yang memiliki nilai indeks terbesar dalam suatu baris data
10. categori : kualitas udara suatu baris data, bernilai “baik”, “sedang”, “tidak sehat”

2.2 Pemrosesan Data

Pertama-tama, install pyspark dan import library yang dibutuhkan termasuk SparkSession dan pyplot.

```
# Instalasi pyspark
!pip install pyspark
```

```
Collecting pyspark
  Downloading pyspark-3.4.1.tar.gz (310.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 310.8/310.8 MB 4.6 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.4.1-py2.py3-none-any.whl size=311285398 sha256=71b6767c37cebbc983bc
  Stored in directory: /root/.cache/pip/wheels/0d/77/a3/ff2f74cc9ab41f8f594dabf0579c2a7c6de920d584206e0834
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.4.1
```

```
# Import library yang digunakan
import requests
from pyspark.sql import SparkSession
from pyspark.sql.functions import avg, min, max, stddev, desc
import matplotlib.pyplot as plt
```

1. Membuat SparkSession untuk mendaftarkan tabel sebagai dataframe.

```
# Membuat spark session
spark = SparkSession.builder.appName("Analisis ISPU").getOrCreate()
```

2. Mengambil data csv, menyimpan, dan menampilkannya untuk 10 baris teratas.

```
# Membaca dan mengekstrak data
df = spark.read.csv("/content/Data_ISPU.csv", header=True)

df.show(10)
```

```
+----------+------------------+----+---+---+---+---+---+--------+--------+
|   tanggal|           stasiun|pm10|so2| co| o3|no2|max|critical|categori|
+----------+------------------+----+---+---+---+---+---+--------+--------+
|2020-01-01|DKI1 (Bunderan HI)|  30| 20| 10| 32|  9| 32|      O3|    BAIK|
|2020-01-02|DKI1 (Bunderan HI)|  27| 22| 12| 29|  8| 29|      O3|    BAIK|
|2020-01-03|DKI1 (Bunderan HI)|  39| 22| 14| 32| 10| 39|    PM10|    BAIK|
|2020-01-04|DKI1 (Bunderan HI)|  34| 22| 14| 38| 10| 38|      O3|    BAIK|
|2020-01-05|DKI1 (Bunderan HI)|  35| 22| 12| 31|  9| 35|    PM10|    BAIK|
|2020-01-06|DKI1 (Bunderan HI)|  46| 23| 16| 32|  9| 46|    PM10|    BAIK|
|2020-01-07|DKI1 (Bunderan HI)|  37| 23| 26| 33| 11| 37|    PM10|    BAIK|
|2020-01-08|DKI1 (Bunderan HI)|  41| 26| 20| 30| 11| 41|    PM10|    BAIK|
|2020-01-09|DKI1 (Bunderan HI)|  52| 23| 29| 24| 12| 52|    PM10|  SEDANG|
|2020-01-10|DKI1 (Bunderan HI)|  24| 24| 18| 25|  8| 25|      O3|    BAIK|
+----------+------------------+----+---+---+---+---+---+--------+--------+
only showing top 10 rows
```

3. Menghitung jumlah baris pada data. Dapat dilihat bahwa terdapat 1830 baris data.

```
# Mengetahui banyaknya baris pada data
df.count()
```

```
1830
```

4. Menghapus baris yang memiliki informasi tidak lengkap.

```
df = df.dropna()
```

5. Mengetahui apakah ada baris yang dihapus. Dari 1830 baris data, terdapat 157 baris data yang dihapus sehingga tersisa 1673 baris data.

```
# Melihat kembali banyaknya baris pada data
df.count()
```

```
1673
```

6. Melihat tipe data dari masing-masing kolom.

```
# Melihat tipe data dari masing-masing kolom
df.dtypes
```

```
[('tanggal', 'string'),
 ('stasiun', 'string'),
 ('pm10', 'string'),
 ('so2', 'string'),
 ('co', 'string'),
 ('o3', 'string'),
 ('no2', 'string'),
 ('max', 'string'),
 ('critical', 'string'),
 ('categori', 'string')]
```

7. Mengganti tipe data menjadi tipe yang sesuai, yaitu integer untuk semua kolom jenis polutan dan kolom “max”.

```
# Mengganti tipe data menjadi tipe yang sesuai
df = df.withColumn("pm10", df["pm10"].cast('int'))
df = df.withColumn("so2", df["so2"].cast('int'))
df = df.withColumn("co", df["co"].cast('int'))
df = df.withColumn("o3", df["o3"].cast('int'))
df = df.withColumn("no2", df["no2"].cast('int'))
df = df.withColumn("max", df["max"].cast('int'))
```

8. Mengubah dataframe pyspark ke dalam bentuk pandas, kemudian ekstrak data yang sudah dibersihkan ke dalam bentuk csv untuk divisualisasikan menggunakan Looker Studio.

```
# Mengubah pyspark df ke pandas df
# Lalu, ekstrak kembali data yang sudah bersih ke csv untuk proses visualisasi di looker studio
df.toPandas().to_csv('Data_ISPU_clean.csv')
```

2.3 Analisis Data

Akan dilihat statistik dari kandungan polutan di udara sekitar stasiun DKI Jakarta tahun 2020. Tabel di bawah menunjukkan nilai dari rata-rata, standar deviasi, minimum, dan maksimum dari kelima jenis polutan.

```
# Melihat statistik dari masing-masing senyawa yang terkandung di udara
df.describe("pm10", "so2", "co", "o3", "no2", "max").show()
```

```
+-------+-----------------+-----------------+-----------------+-----------------+-----------------+---------------+
|summary|             pm10|              so2|               co|               o3|              no2|            max|
+-------+-----------------+-----------------+-----------------+-----------------+-----------------+---------------+
|  count|             1673|             1673|             1672|             1673|             1673|           1673|
|   mean| 49.09683203825463|23.160789001793187|21.455143540669855| 53.38971906754333|18.702331141661684|63.0442319187089|
| stddev|17.623017944980827| 16.21409813478605| 22.83718863562032|30.036778723048727| 25.86780620138816|27.7918046553943|
|    min|                3|                1|                3|                2|                1|              1|
|    max|              111|              112|              197|              191|              197|            197|
+-------+-----------------+-----------------+-----------------+-----------------+-----------------+---------------+
```

Pembuatan sebuah tabel sementara.

```
# Analisis data dengan spark sql
# Membuat temporary tabel
df.createOrReplaceTempView("data_ISPU")
```

2.3.1 Analisis Kualitas Udara pada Setiap Stasiun DKI Jakarta Tahun 2020

Analisis pertama yang dilakukan adalah menghitung seberapa seringnya muncul hasil kualitas udara “baik”, “sedang”, dan “tidak baik” di setiap stasiun.

```
# (Analisis 1) Kualitas Udara Masing-masing Stasiun
a1 = spark.sql("""
  SELECT
    stasiun,
    categori,
    COUNT(categori)
  FROM data_ISPU
  GROUP BY stasiun, categori
  ORDER BY stasiun, categori
""")

# Menampilkan tabel: Kualitas Udara Masing-masing Stasiun
a1.show()
```

stasiun	categori	count(categori)
DKI1 (Bunderan HI)	BAIK	120
DKI1 (Bunderan HI)	SEDANG	230
DKI1 (Bunderan HI)	TIDAK SEHAT	4
DKI2 (Kelapa Gading)	BAIK	75
DKI2 (Kelapa Gading)	SEDANG	249
DKI2 (Kelapa Gading)	TIDAK SEHAT	31
DKI3 (Jagakarsa)	BAIK	40
DKI3 (Jagakarsa)	SEDANG	254
DKI3 (Jagakarsa)	TIDAK SEHAT	31
DKI4 (Lubang Buaya)	BAIK	46
DKI4 (Lubang Buaya)	SEDANG	252
DKI4 (Lubang Buaya)	TIDAK SEHAT	21
DKI5 (Kebon Jeruk)	BAIK	48
DKI5 (Kebon Jeruk)	SEDANG	220
DKI5 (Kebon Jeruk)	TIDAK SEHAT	52

Program sebelumnya akan menghasilkan tabel di atas. Dapat dilihat bahwa stasiun Bunderan HI secara umum merupakan stasiun yang memiliki udara paling bersih karena memiliki frekuensi kualitas udara baik paling tinggi dan kualitas tidak sehat paling rendah. Kemudian, stasiun Kebon Jeruk memiliki frekuensi tidak sehat paling tinggi, sehingga udara di sekitar stasiun tersebut dinilai paling tercemar dibandingkan yang lain.

2.3.2 Analisis Jenis Polutan dengan ISPU Tertinggi di Stasiun DKI Jakarta 2020

Untuk mencari nilai jenis polutan dengan ISPU tertinggi, akan dijumlahkan setiap nilai ISPU masing-masing jenis polutan yang tercatat dalam setiap stasiun.

```
# (Analisis 2) Jumlah Nilai Indeks Polutan pada Masing-masing Stasiun
a2 = spark.sql("""
  SELECT
    stasiun,
    SUM(pm10) AS jumlah_pm10,
    SUM(so2) AS jumlah_so2,
    SUM(co) AS jumlah_co,
    SUM(o3) AS jumlah_o3,
    SUM(no2) AS jumlah_no2
  FROM data_ISPU
  GROUP BY stasiun
  ORDER BY jumlah_pm10 ASC
""")

# Menampilkan tabel: Jumlah Nilai Indeks Polutan pada Masing-masing Stasiun
a2.show()
```

Dengan menggunakan kode diatas, didapatkan hasil tabel sebagai berikut:

```
+--------------------+-----------+----------+---------+---------+----------+
|             stasiun|jumlah_pm10|jumlah_so2|jumlah_co|jumlah_o3|jumlah_no2|
+--------------------+-----------+----------+---------+---------+----------+
|  DKI5 (Kebon Jeruk)|      14724|      5924|     8509|    20784|      6200|
|    DKI3 (Jagakarsa)|      15492|      8519|     8602|    17959|      5769|
|  DKI1 (Bunderan HI)|      15897|      7189|     4669|    14856|      6422|
|DKI2 (Kelapa Gading)|      17662|      8265|     5975|    19112|      7082|
| DKI4 (Lubang Buaya)|      18364|      8851|     8118|    16610|      5816|
+--------------------+-----------+----------+---------+---------+----------+
```

Bisa dilihat bahwa pada umumnya, PM_{10} dan ozon merupakan polutan dengan nilai ISPU terbesar. Di stasiun Bunderan HI dan Lubang Buaya, PM_{10} merupakan polutan dengan nilai ISPU tertinggi, sedangkan di stasiun Kebon Jeruk, Jagakarsa, dan Kelapa Gading, ozon merupakan polutan dengan nilai ISPU tertinggi.

2.3.3 Analisis Tren Nilai Rata-rata ISPU Polutan Berdasarkan Tanggal

Analisis ini dilakukan untuk melihat tren nilai rata-rata indeks masing-masing polutan berdasarkan tanggal yang sama dan tanggal berapa suatu jenis polutan memiliki nilai indeks tertinggi sepanjang tahun 2020. Pertama-tama, akan dibuat tabel dengan kolom data yang akan digunakan.

```
# (Analisis 3) Rata-rata Nilai Indeks Polutan Berdasarkan Tanggal

# Membuat tabel yang hanya terdiri dari kolom "tanggal", "pm10", "so2", "co", "o3", "no2"
a3 = spark.sql("""
  SELECT
    tanggal,
    pm10,
    so2,
    co,
    o3,
    no2
  FROM data_ISPU
""")
```

Kemudian, data akan diubah ke dalam bentuk pandas lalu menambahkan kolom untuk menunjukkan hari ke berapa dalam setiap bulan data tersebut diambil.

```
import pandas as pd
a3_pd = a3.toPandas() # Mengubah hasil menjadi pandas dataframe
a3_pd["tanggal"] = pd.to_datetime(a3_pd["tanggal"]) # Mengubah tipe kolom "tanggal" dari object ke datetime

a3_pd["hari"] = a3_pd["tanggal"].dt.day # Menambahkan kolom hari
print(a3_pd)
```

```
        tanggal  pm10  so2    co  o3  no2  hari
0    2020-01-01    30   20  10.0  32    9     1
1    2020-01-02    27   22  12.0  29    8     2
2    2020-01-03    39   22  14.0  32   10     3
3    2020-01-04    34   22  14.0  38   10     4
4    2020-01-05    35   22  12.0  31    9     5
...         ...   ...  ...   ...  ..  ...   ...
1668 2020-12-24    29   31   9.0  28    2    24
1669 2020-12-25    24   27   7.0  18    3    25
1670 2020-12-28    22   33   5.0  35    3    28
1671 2020-12-30    16    7   3.0  21    2    30
1672 2020-12-31    18   13   6.0  24    3    31

[1673 rows x 7 columns]
```

Lalu, akan dibentuk tabel menggunakan data di atas

```
a3_py = spark.createDataFrame(a3_pd)
a3_py.show()

# Membuat temporary tabel
a3_py.createOrReplaceTempView("data_analisis3")
```

```
+-------------------+----+---+----+---+---+----+
|            tanggal|pm10|so2|  co| o3|no2|hari|
+-------------------+----+---+----+---+---+----+
|2020-01-01 00:00:00|  30| 20|10.0| 32|  9|   1|
|2020-01-02 00:00:00|  27| 22|12.0| 29|  8|   2|
|2020-01-03 00:00:00|  39| 22|14.0| 32| 10|   3|
|2020-01-04 00:00:00|  34| 22|14.0| 38| 10|   4|
|2020-01-05 00:00:00|  35| 22|12.0| 31|  9|   5|
|2020-01-06 00:00:00|  46| 23|16.0| 32|  9|   6|
|2020-01-07 00:00:00|  37| 23|26.0| 33| 11|   7|
|2020-01-08 00:00:00|  41| 26|20.0| 30| 11|   8|
|2020-01-09 00:00:00|  52| 23|29.0| 24| 12|   9|
|2020-01-10 00:00:00|  24| 24|18.0| 25|  8|  10|
|2020-01-11 00:00:00|  34| 31| NaN| 23|  8|  11|
|2020-01-12 00:00:00|  27| 23| 9.0| 33|  4|  12|
|2020-01-13 00:00:00|  33| 26|12.0| 36|  8|  13|
|2020-01-14 00:00:00|  34| 28|13.0| 27|  7|  14|
|2020-01-15 00:00:00|  29| 22|13.0| 36|  8|  15|
|2020-01-16 00:00:00|  52| 60|19.0| 30|  8|  16|
|2020-01-17 00:00:00|  51| 34|21.0| 74| 20|  17|
|2020-01-18 00:00:00|  37| 29|14.0| 31|  6|  18|
|2020-01-19 00:00:00|  61| 34|36.0| 58| 15|  19|
|2020-01-20 00:00:00|  47| 30|15.0| 33|  9|  20|
```

Dengan menggunakan tabel di atas, akan dijumlahkan setiap baris data dengan hari yang sama, lalu dihitung nilai rata-ratanya

```
# Membuat tabel: Rata-rata Nilai Indeks Polutan Berdasarkan Tanggal
a_3a = spark.sql("""
  SELECT
    hari,
    CAST(AVG(pm10) AS DECIMAL(5,2)) AS avg_pm10,
    CAST(AVG(so2) AS DECIMAL(5,2)) AS avg_so2,
    CAST(AVG(co) AS DECIMAL(5,2)) AS avg_co,
    CAST(AVG(o3) AS DECIMAL(5,2)) AS avg_o3,
    CAST(AVG(no2) AS DECIMAL(5,2)) AS avg_no2
  FROM data_analisis3
  GROUP BY hari
  ORDER BY hari ASC
""")

# Menampilkan tabel: Rata-rata Nilai Indeks Polutan Berdasarkan Tanggal
a_3a.show(31)
```

Program di atas akan menghasilkan tabel yang menunjukkan rata-rata ISPU polutan pada setiap tanggal sebagai berikut

hari	avg_pm10	avg_so2	avg_co	avg_o3	avg_no2
1	43.10	20.44	20.40	45.28	7.42
2	46.65	15.75	25.24	47.80	8.65
3	13.53	21.98	54.63	8.37	58.35
4	48.12	14.00	13.80	61.55	6.24
5	51.94	13.75	13.04	56.50	7.42
6	54.67	15.36	13.16	60.00	7.81
7	61.49	19.59	12.31	65.81	9.59
8	67.92	20.34	13.73	74.41	9.39
9	61.79	24.00	9.86	83.40	8.79
10	52.32	19.14	41.56	52.90	42.00
11	56.17	61.81	23.59	38.14	41.31
12	31.33	33.02	7.53	33.61	6.98
13	48.92	23.38	19.17	53.43	17.53
14	48.90	25.88	29.34	51.44	30.52
15	49.96	22.65	25.13	51.00	21.04
16	50.49	24.27	24.56	49.24	21.16
17	48.98	24.07	24.12	59.16	22.37
18	47.95	23.10	22.50	56.17	19.64
19	53.36	24.39	22.39	60.04	18.11
20	48.85	23.29	18.40	46.94	16.08
21	43.79	23.26	20.72	48.92	17.21
22	52.14	22.77	19.89	48.11	14.93
23	47.98	23.79	19.23	55.13	16.11
24	47.53	22.36	20.09	53.47	18.62
25	46.96	22.16	21.07	51.11	16.53
26	50.70	23.19	21.24	58.39	17.72
27	44.70	21.83	17.69	57.80	15.56
28	50.09	23.53	27.23	59.91	26.64
29	49.85	21.09	23.42	57.45	19.07
30	48.17	20.08	21.50	49.90	18.15
31	42.94	18.15	23.61	50.27	22.70

Berdasarkan hasil di atas, PM_{10} paling banyak pada tanggal 19, sulfur dioksida, karbon monoksida, dan nitrogen dioksida paling banyak pada tanggal 14, dan ozon paling banyak pada tanggal 12. Maka itu, bisa dilihat bahwa pada tanggal 14 tingkat polusi udara itu pada umumnya paling tinggi.

2.3.4 Analisis Rata-rata ISPU untuk Masing-masing Jenis Polutan Berdasarkan Kategori

Analisis ini bertujuan untuk melihat bagaimana penilaian sebuah baris data untuk dikategorikan ke dalam setiap kategori. Yang akan dilakukan adalah mencari nilai rata-rata setiap polutan serta hasil penjumlahannya untuk setiap kategori. Didapatkan tabel sebagai berikut:

```
# (Analisis 4) Rata-rata Nilai Indeks Polutan Setiap Kategori

a4 = spark.sql("""
  SELECT
    categori,
    CAST(AVG(pm10+so2+co+o3+no2) AS DECIMAL(5,2)) AS avg_jmlh,
    CAST(AVG(pm10) AS DECIMAL(5,2)) AS avg_pm10,
    CAST(AVG(so2) AS DECIMAL(5,2)) AS avg_so2,
    CAST(AVG(co) AS DECIMAL(5,2)) AS avg_co,
    CAST(AVG(o3) AS DECIMAL(5,2)) AS avg_o3,
    CAST(AVG(no2) AS DECIMAL(5,2)) AS avg_no2
  FROM data_ISPU
  GROUP BY categori
  ORDER BY avg_jmlh ASC
""")

a4.show()
```

```
+-----------+--------+--------+-------+------+------+-------+
|   categori|avg_jmlh|avg_pm10|avg_so2|avg_co|avg_o3|avg_no2|
+-----------+--------+--------+-------+------+------+-------+
|       BAIK|  108.14|   30.96|  19.52| 15.12| 31.13|  11.41|
|     SEDANG|  172.19|   53.53|  23.80| 20.71| 56.00|  18.15|
|TIDAK SEHAT|  246.94|   53.57|  26.19| 42.96| 83.43|  40.79|
+-----------+--------+--------+-------+------+------+-------+
```

Didapatkan bahwa jumlah ISPU dari kelima polutan yang baik adalah sekitar 108.14, jumlah ISPU yang sedang adalah sekitar 172.19, dan jumlah ISPU yang tidak sehat adalah 246.94. Semakin rendah nilai ISPU, maka kualitas udara akan semakin membaik.

2.3.5 Analisis Jenis Polutan yang Paling Kritis Setiap Bulan

Analisis ini dilakukan agar bisa diketahui jenis polutan apa yang paling banyak terdapat di udara setiap harinya sehingga bisa diutamakan upaya untuk mereduksi polutan tersebut. Pertama-tama akan diambil kolom data yang dibutuhkan.

```
# (Analisis 5) Jenis Polutan dengan ISPU Tertinggi Setiap Bulan

a5 = spark.sql("""
  SELECT
    tanggal,
    pm10,
    so2,
    co,
    o3,
    no2,
    critical
  FROM data_ISPU
""")

a5.show(5)
```

```
+--------+----+---+---+---+---+--------+
| tanggal|pm10|so2| co| o3|no2|critical|
+--------+----+---+---+---+---+--------+
|1/1/2020|  30| 20| 10| 32|  9|      O3|
|2/1/2020|  27| 22| 12| 29|  8|      O3|
|3/1/2020|  39| 22| 14| 32| 10|    PM10|
|4/1/2020|  34| 22| 14| 38| 10|      O3|
|5/1/2020|  35| 22| 12| 31|  9|    PM10|
+--------+----+---+---+---+---+--------+
only showing top 5 rows
```

Kemudian, ditambahkan kolom bulan, yaitu bulan data diambil lalu diubah ke dalam bentuk tabel

```
a5_py = spark.createDataFrame(a5_pd)
a5_py.show()

# Membuat temporary tabel
a5_py.createOrReplaceTempView("data_analisis5")
```

```
+-------------------+----+---+---+---+---+--------+-----+
|            tanggal|pm10|so2| co| o3|no2|critical|bulan|
+-------------------+----+---+---+---+---+--------+-----+
|2020-01-01 00:00:00|  30| 20| 10| 32|  9|      O3|    1|
|2020-02-01 00:00:00|  27| 22| 12| 29|  8|      O3|    2|
|2020-03-01 00:00:00|  39| 22| 14| 32| 10|    PM10|    3|
|2020-04-01 00:00:00|  34| 22| 14| 38| 10|      O3|    4|
|2020-05-01 00:00:00|  35| 22| 12| 31|  9|    PM10|    5|
|2020-06-01 00:00:00|  46| 23| 16| 32|  9|    PM10|    6|
|2020-07-01 00:00:00|  37| 23| 26| 33| 11|    PM10|    7|
|2020-08-01 00:00:00|  41| 26| 20| 30| 11|    PM10|    8|
|2020-09-01 00:00:00|  52| 23| 29| 24| 12|    PM10|    9|
|2020-10-01 00:00:00|  24| 24| 18| 25|  8|      O3|   10|
|2020-11-01 00:00:00|  34| 31| 25| 23|  8|    PM10|   11|
|2020-12-01 00:00:00|  27| 23|  9| 33|  4|      O3|   12|
|2020-01-13 00:00:00|  33| 26| 12| 36|  8|      O3|    1|
|2020-01-14 00:00:00|  34| 28| 13| 27|  7|    PM10|    1|
|2020-01-15 00:00:00|  29| 22| 13| 36|  8|      O3|    1|
|2020-01-16 00:00:00|  52| 60| 19| 30|  8|     SO2|    1|
|2020-01-17 00:00:00|  51| 34| 21| 74| 20|      O3|    1|
|2020-01-18 00:00:00|  37| 29| 14| 31|  6|    PM10|    1|
|2020-01-19 00:00:00|  61| 34| 36| 58| 15|    PM10|    1|
|2020-01-20 00:00:00|  47| 30| 15| 33|  9|    PM10|    1|
+-------------------+----+---+---+---+---+--------+-----+
```

Lalu, akan dikelompokkan ke dalam bulan, lalu lihat nilai kritikal maksimum di setiap bulannya dan tambahkan jenis polutan dengan nilai kritikal maksimum tersebut. Didapatkan hasil sebagai berikut:

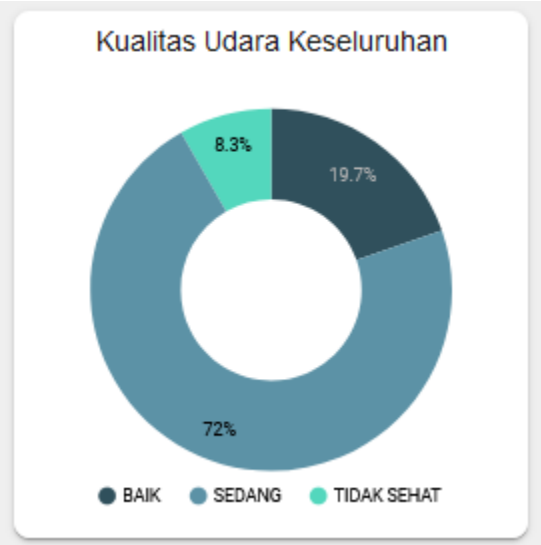
bulan	critical	max_jumlah_critical
1	O3	70
2	O3	75
3	CO	69
4	O3	78
5	O3	92
6	O3	70
7	O3	76
8	O3	75
9	O3	96
10	O3	75
11	SO2	48
11	O3	48
12	O3	73

Bisa dilihat bahwa pada umumnya polutan ozon adalah polutan yang paling sering merupakan polutan dengan nilai ISPU maksimal, sehingga polutan ozon sangat besar di lingkungan stasiun DKI Jakarta.

2.4 Visualisasi Data

2.4.1 Kualitas Udara Secara Keseluruhan

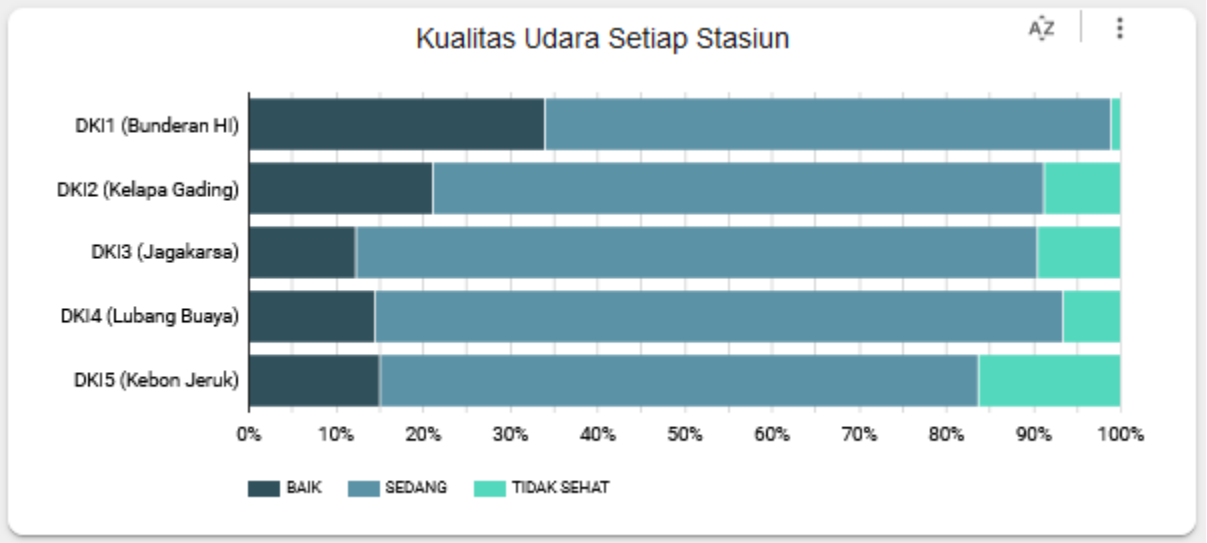
Berikut adalah visualisasi kualitas udara secara keseluruhan dengan cara menjumlahkan frekuensi setiap kategori.



Bisa dilihat bahwa kualitas udara dominan di kategori “sedang”. Selain itu, kategori baik lebih sering muncul daripada kategori tidak sehat, dengan frekuensi lebih dari dua kali lipat. Maka dari itu, bisa dinilai bahwa secara umum kualitas udara di Jakarta cukup baik, namun masih memiliki banyak potensi peningkatan.

2.4.2 Kualitas Udara Setiap Stasiun

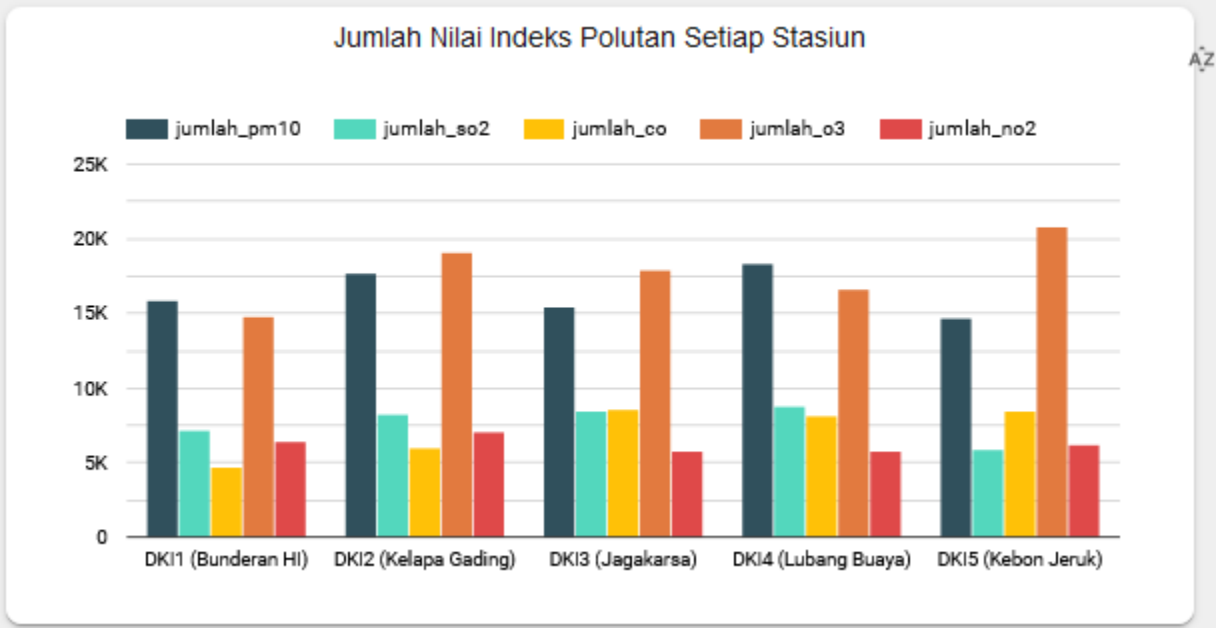
Berikut adalah visualisasi kualitas udara di setiap stasiun Jakarta:



Bisa dilihat bahwa stasiun Bunderan HI sangat baik karena frekuensi kategori “baik” sangat tinggi dibandignkan kategori tidak sehat. Namun, frekuensi kategori “tidak sehat” di stasiun Kebon Jeruk sangat tinggi sehingga udara di daerah tersebut perlu ditingkatkan.

2.4.3 ISPU Setiap Jenis Polutan di Setiap Stasiun

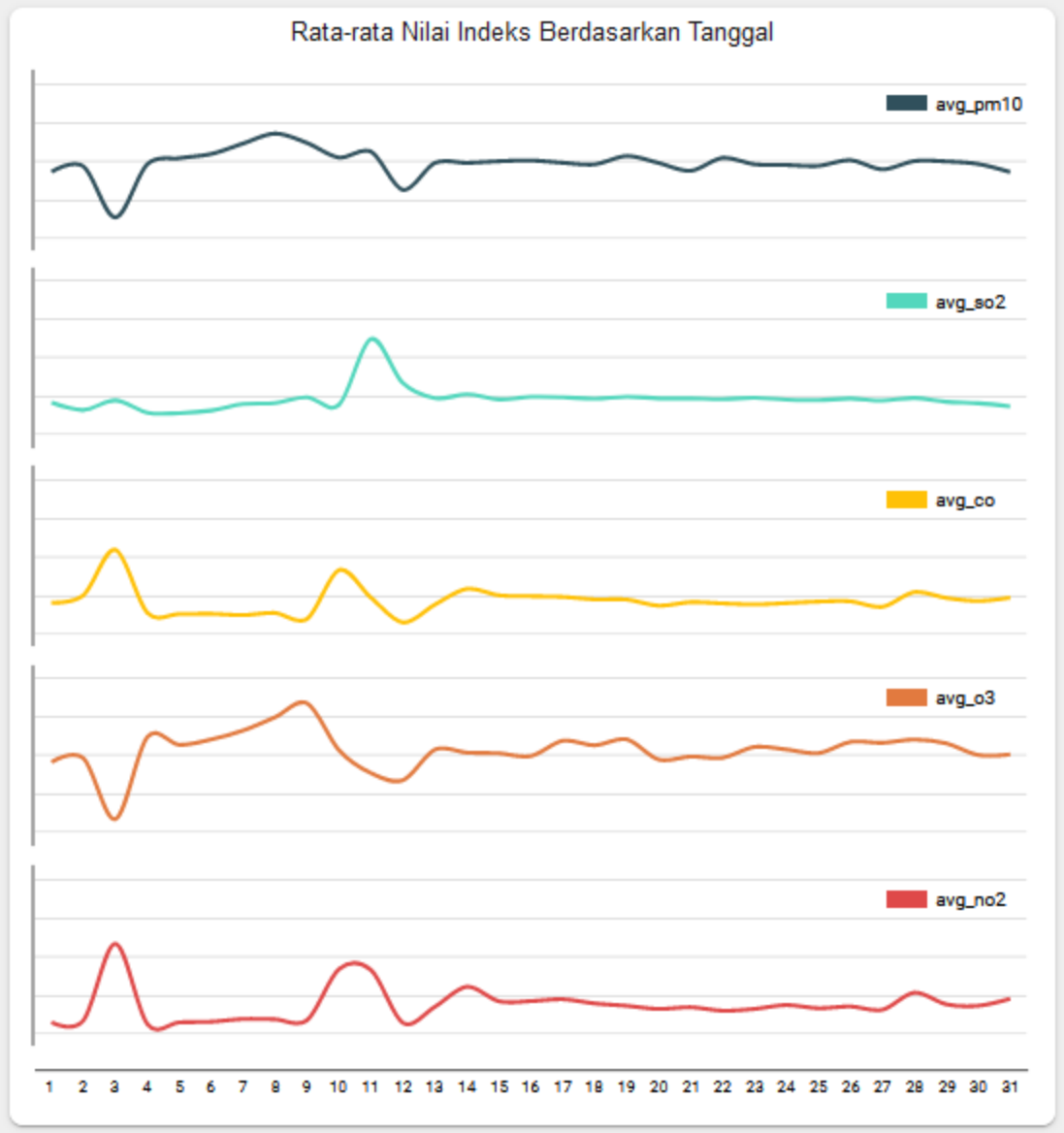
Berikut adalah visualisasi setiap jenis polutan di setiap stasiun:



Bisa dilihat bahwa secara umum jenis polutan yang paling banyak adalah ozon dan PM_{10}. kemudian, yang mencolok adalah total ISPU polutan ozon di stasiun Kebon Jeruk sangat banyak dibandingkan dengan lain, dengan nilai lebih besar dari dua puluh ribu, sehingga perlu dikurangi kandungan ozon di Jakarta, terutama di sekitar stasiun Kebon Jeruk.

2.4.4 Tren Rata-Rata ISPU Setiap Jenis Polutan Terhadap Waktu

Berikut adalah tren rata-rata ISPU setiap jenis polutan terhadap waktu dalam hari:



Bisa dilihat bahwa tren mengalami perubahan besar pada tanggal 3. Dua jenis polutan naik drastis, dua jenis polutan turun drastis, dan satu mengalami peningkatan kecil, namun semuanya kembali mendekati nilai ISPU di hari sebelumnya. Kemudian, pada tanggal 5-8, ISPU ozon dan PM_{10} meningkat secara perlahan. Lalu, di tanggal 9-12, tiga jenis polutan meningkat cukup signifikan dan dua polutan lainnya menurun signifikan, namun semuanya juga kembali mendekati rata-ratanya.

2.4.5 Rata-Rata ISPU Setiap Kategori Kualitas Udara

Berikut adalah nilai rata-rata dari penjumlahan ISPU setiap jenis polutan dalam setiap kategori:

categori	avg_jmlh
BAIK	108.14
SEDANG	172.19
TIDAK SEHAT	246.94

Kemudian, berikut adalah visualisasi rata-rata ISPU setiap kategori kualitas udara:

Berdasarkan grafik di atas, kategori kualitas udara yang semakin baik akan diikuti oleh rata-rata ISPU untuk setiap jenis polutan yang berkurang.

2.4.6 Jenis Polutan Paling Kritis Setiap Bulan

Berikut adalah visualisasi polutan paling kritis, yaitu polutan dengan ISPU tertinggi, di setiap bulan:

Jenis Polutan Paling Kritis Setiap Bulan

bulan	critical	max_jumlah_critical
6	O3	70
7	O3	76
8	O3	75
9	O3	96
10	O3	75
11	SO2	48
11	O3	48
12	O3	73

Jenis Polutan Paling Kritis Setiap Bulan

bulan	critical	max_jumlah_critical
6	O3	70
7	O3	76
8	O3	75
9	O3	96
10	O3	75
11	SO2	48
11	O3	48
12	O3	73

Bisa dilihat bahwa nilai ISPU Ozon sangat sering menjadi polutan kritis, sehingga jenis polutan ini cukup urgen untuk dikurangi, sebab nilai ISPU ozon relatif tinggi.

Visualisasi data sebelumnya dibuat pada link Looker Studio sebagai berikut:
https://lookerstudio.google.com/reporting/cad6ff97-fed1-43e4-a5e9-1c38199d1d09

BAB III
KESIMPULAN DAN SARAN

3.1 Kesimpulan

Kualitas udara di 5 stasiun di Jakarta mayoritas berada pada kategori “sedang”. Kualitas udara terbaik berada pada stasiun DK1 (Bunderan HI), sedangkan terburuk berada pada stasiun DKI5 (Kebon Jeruk). Polutan ozon (o3) dan pm10 adalah jenis yang paling banyak ditemukan pada setiap stasiun di mana ozon merupakan yang paling kritis. Selanjutnya, rata-rata nilai ISPU polutan fluktuatif pada setengah bulan awal dan lebih stabil pada setengah bulan akhir.

Secara umum, kualitas udara di Jakarta masih memiliki banyak potensi peningkatan. Untuk melakukan hal tersebut, nilai ISPU setiap jenis polutan harus dikurangi, terutama polutan ozon dan PM_{10} yang memiliki indeks yang tinggi. Kemudian, kualitas udara di sekitar stasiun Kebon Jeruk harus ditingkatkan juga karena kualitas udara di stasiun tersebut kurang baik. Dengan mengurangi polutan di udara, kualitas udara pun akan meningkat sehingga diperlukan upaya untuk mengurangi polutan di udara.

3.2 Saran

Adapun saran bagi penelitian berikutnya adalah:

- Menggunakan data ISPU yang lebih terkini sehingga bisa lebih menggambarkan masa sekarang dan prediksi kedepannya dengan lebih akurat.
- Membuat model yang mampu memprediksi ISPU polutan di tahun-tahun kedepan.

Adapun saran bagi pemerintah adalah:

- Melakukan upaya untuk mengurangi pencemaran udara di Jakarta, terutama di stasiun Kebon Jeruk karena 16,25% dari 366 hari termasuk dalam kategori tidak sehat.
- Melakukan upaya untuk mengurangi emisi gas O3 karena gas O3 merupakan polutan terbanyak di mayoritas stasiun dan yang memiliki indeks pencemaran terbesar di hampir setiap bulan.