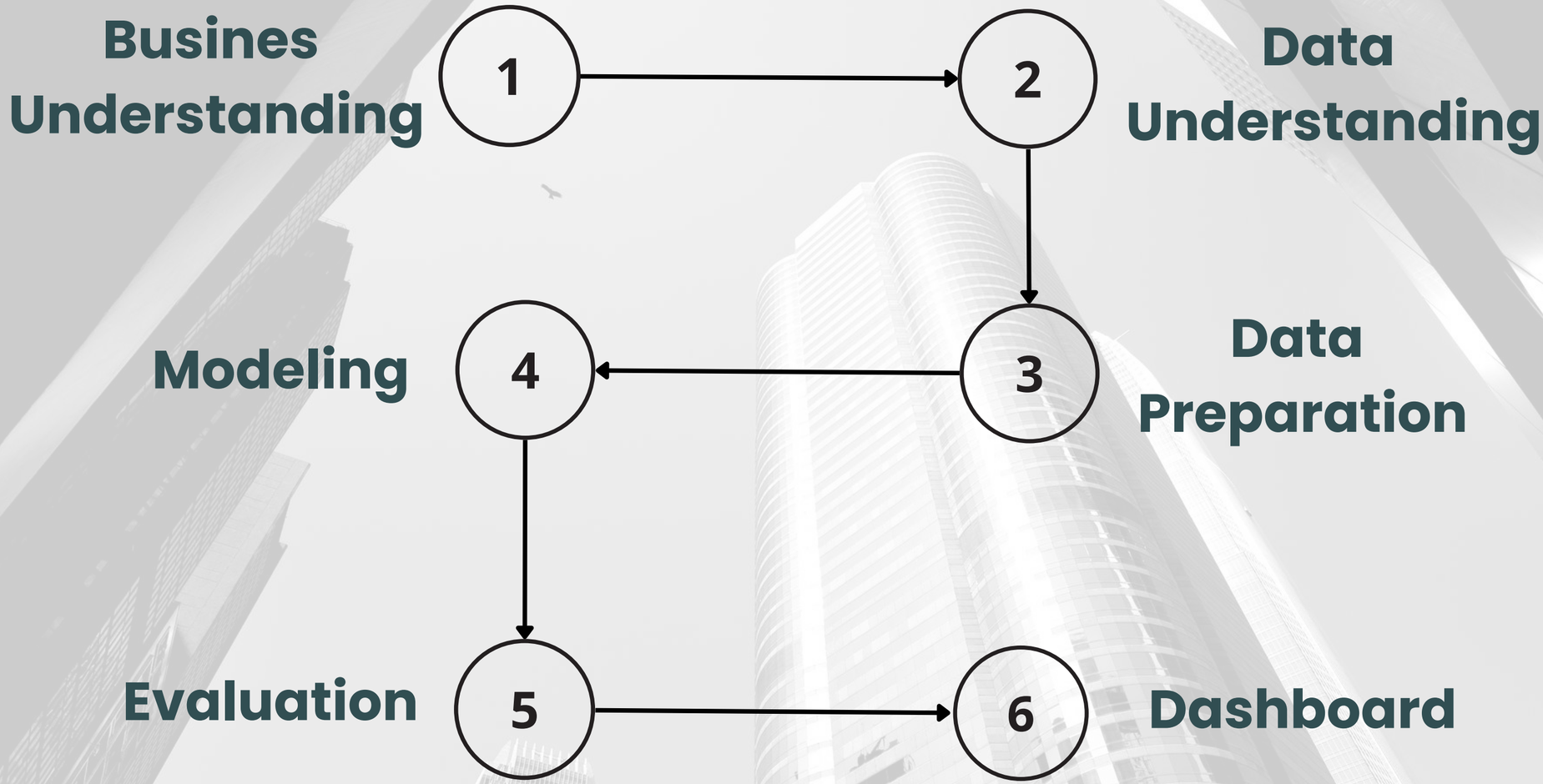
ruang guru x Kampus Merdeka INDONESIA JAYA

# FINAL PROJECT

## HR DEPARTMENT

Data, Bussines Analytics
and Operations

# DAFTAR ISI


Busines Understanding
1
2
Data Understanding
Modeling
4
3
Data Preparation
Evaluation
5
6
Dashboard

# Business Understanding

# HR DEPARTMENT

## Goals

*Mengurangi rata-rata absensi staff perusahaan selama setahun maksimal 16 jam saja supaya dapat meningkatkan kontribusi , produktifitas, kedisiplinan staff perusahaan , serta meminimalisirkan kerugian pada perusahaan*

- **WHAT**

  *Mengurangi waktu absensi staff perusahaan*

- **WHO**

  *Staff perusahaan*

- **WHERE**

  *The Look E-commerce*

- **WHEN**

  *Data periode 2019 -2022*

- **WHY**

  *Untuk meningkatkan kontribusi , produktifitas, kedisiplinan staff, serta meminimalisirkan kerugian pada perusahaan*

- **HOW**

  *Dengan menganalisis data rata-rata absensi staff serta menggunakan Logistic Regression*

# Data Understanding

**TOOLS YANG DIGUNAKAN :**

PostgreSQL

Google Collab

# HUMAN RESOURCES DEPARTMENT

- **COLLECT INITIAL DATA (DATA EMPLOYEES - PRIMARY)**

| Field name | Description |
|---|---|
| Fisrt_Name | Nama depan pegawai |
| Last_Name | Nama belakang pegawai |
| Gender | Jenis kelamin pegawai |
| Age | Usia pegawai |
| Length_Service | Lama pegawai bekerja diperusahaan The looker |
| Absent_Hours | Lama absen dalam satu tahun (jam) |
| distribution_centers_id | Distribution centers id (tabel) |

*Link Dataset From Kaggle :*
*https://www.kaggle.com/datasets/HRAnalyticRepository/absenteeism-dataset*

# HUMAN RESOURCES DEPARTMENT

- **COLLECT INITIAL DATA (DATA DISTRIBUTION CENTERS - SECONDARY)**

| Field name | Description |
|---|---|
| id | Id distribution centers |
| name | Nama Distribution Centers |
| latitude | Titik koordinat latitude |
| longitude | Titik koordinat longitude |

# HUMAN RESOURCES DEPARTMENT

- **COLLECT INITIAL DATA - DATA PRODUCTS (SECONDARY)**

| Field name | Description |
|---|---|
| id | ID Products |
| cost | Harga produk saat produksi |
| category | Kategori dari produk |
| name | Nama dari produk |
| brand | Brand dari produk |

| Field Name | Description |
|---|---|
| retail_price | Harga retail |
| department | Departemen produk (untuk laki-laki/perempuan) |
| sku | SKU produk |
| distribution_center_id | Id distribution center (tabel) |

# HUMAN RESOURCES DEPARTMENT

- **DESCRIBE DATA ( WITH SQL )**
  - **DATA EMPLOYEES**

*SELECT * FROM Employees*
*LIMIT 10;*

| | fisrt_name character varying | last_name character varying | gender character varying (1) | age double precision | length_service double precision | absent_hour double precision | distribution_centers_id integer |
|---|---|---|---|---|---|---|---|
| 1 | Gutierrez | Molly | F | 32.02881569 | 6.018478474 | 36.57730606 | 5 |
| 2 | Hardwick | Stephen | M | 40.32090167 | 5.532444578 | 30.16507231 | 9 |
| 3 | Delgado | Chester | M | 48.82204661 | 4.389973118 | 83.80779766 | 10 |
| 4 | Simon | Irene | F | 44.59935722 | 3.081735738 | 70.02016505 | 2 |
| 5 | Delvalle | Edward | M | 35.69787561 | 3.619091448 | 0 | 4 |
| 6 | Jones | Ernie | M | 48.44031059 | 2.717692452 | 81.83007916 | 6 |
| 7 | Buford | Ralph | M | 50.75273 | 10.157918 | 60.49507152 | 6 |
| 8 | Lee | Gregory | M | 36.2160312 | 4.432122862 | 30.07290192 | 10 |
| 9 | Smith | Jerry | M | 58.42738025 | 6.940120524 | 181.630819 | 4 |
| 10 | Beard | Robert | M | 39.85398 | 13.848321 | 30.66440832 | 10 |

# HUMAN RESOURCES DEPARTMENT

- **DESCRIBE DATA ( WITH SQL )**
  - **DATA DISTRIBUTION CENTERS**

*SELECT * FROM Distribution_centers*
*LIMIT 10;*

| | id integer | name character varying | latitude double precision | longitude double precision |
|---|---|---|---|---|
| 1 | 1 | Memphis TN | 35.1174 | -89.9711 |
| 2 | 2 | Chicago IL | 41.8369 | -87.6847 |
| 3 | 3 | Houston TX | 29.7604 | -95.3698 |
| 4 | 4 | Los Angeles CA | 34.05 | -118.25 |
| 5 | 5 | New Orleans LA | 29.95 | -90.0667 |
| 6 | 6 | Port Authority of N... | 40.634 | -73.7834 |
| 7 | 7 | Philadelphia PA | 39.95 | -75.1667 |
| 8 | 8 | Mobile AL | 30.6944 | -88.0431 |
| 9 | 9 | Charleston SC | 32.7833 | -79.9333 |
| 10 | 10 | Savannah GA | 32.0167 | -81.1167 |

# HUMAN RESOURCES DEPARTMENT

- **DESCRIBE DATA ( WITH SQL )**
  - **DATA PRODUCTS**

*SELECT * FROM Products*
*LIMIT 10;*

| | id integer | cost double precision | category character varying | name character varying | brand character v | retail_price double precision | department character var | sku character varying | distribution_center_id integer |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 27569 | 92.6525625942932 | Swim | 2XU Men's Swimm... | 2XU | 150.41000366210938 | Men | B23C5765E165D83AA924FA8F... | 1 |
| 2 | 27445 | 24.71966119429112 | Swim | TYR Sport Men's Sq... | TYR | 38.9900016784668 | Men | 2AB7D3B23574C3DEA2BD278... | 1 |
| 3 | 27457 | 15.89760025509596 | Swim | TYR Sport Men's So... | TYR | 27.600000381469727 | Men | 8F831227B0EB6C6D09A05555... | 1 |
| 4 | 27466 | 17.850000048056245 | Swim | TYR Sport Men's S... | TYR | 30 | Men | 67317D6DCC4CB778AEB92195... | 1 |
| 5 | 27481 | 29.408000515669585 | Swim | TYR Alliance Team ... | TYR | 45.950000762939446 | Men | 213C888198806EF1A0E2BBF2... | 1 |
| 6 | 27487 | 15.655589914467036 | Swim | TYR Sport Men's 4-I... | TYR | 26.489999771118164 | Men | 978F39314267ADC0E1C50DB2... | 1 |
| 7 | 27510 | 22.57175048463792 | Swim | TYR Sport Men's So... | TYR | 39.950000762939446 | Men | 4ECBB790F241666326D31F79... | 1 |
| 8 | 27529 | 22.824000120162964 | Swim | TYR Sport Men's Po... | TYR | 36 | Men | C386CBA5332D11385672EE52... | 1 |
| 9 | 27537 | 24.353911066282688 | Swim | TYR Sport Men's All... | TYR | 39.9900016784668 | Men | D012C46243D7E2391B64B221... | 1 |
| 10 | 27552 | 19.317550501902403 | Swim | TYR Sport Men's So... | TYR | 33.95000076293945 | Men | 2AF9B1A840B4ECD522FE1CD... | 1 |

# HUMAN RESOURCES DEPARTMENT

- **EXPLORE DATA ( WITH SQL )**
  - **DATA EMPLOYEES**

*SELECT Gender, AVG(age) as Rata_rata_usia, AVG(absent_hour) as Rata_rata_absent, AVG(length_service) as Rata_rata_Length_service FROM Employees GROUP BY Gender*

***Menampilkan rata2 usia, rata2 absent dan rata2 masa kerja berdasarkan gender***

| | gender character varying (1) | rata_rata_usia double precision | rata_rata_absent double precision | rata_rata_length_service double precision |
|---|---|---|---|---|
| 1 | M | 42.696780467084956 | 56.0026664661769 | 4.78616725392837 |
| 2 | F | 41.30132035237216 | 66.68834857788447 | 4.779575968330343 |

*SELECT CASE WHEN FLOOR (length_service) <= 1 THEN 'Baru (<=1 Tahun)' WHEN FLOOR (length_service) BETWEEN 1 AND 4 THEN 'Sedang (1-4 Tahun)' WHEN FLOOR (length_service) > 4 THEN 'Lama (>4 Tahun)' END AS Kategori_lama_kerja,COUNT (*) AS total_karyawan FROM Employees GROUP BY Kategori_lama_kerja*

| | kategori_lama_kerja text | total_karyawan bigint |
|---|---|---|
| 1 | Sedang (1-4 Tahun) | 4682 |
| 2 | Baru (<=1 Tahun) | 357 |
| 3 | Lama (>4 Tahun) | 3297 |

***Menjumlahakn Staff berdasarkan kategori lama kerja yaitu :***

- ***BARU (<=1 Tahun) artinya pengalaman kerja masih sedikit/pemula***
- ***SEDANG (1-4 Tahun) artinya pengalaman kerja sudah cukup banyak***
- ***LAMA (>=4 Tahun) artinya pengalaman kerja sudah banyak dapat di katakan sebagai senior***

# HUMAN RESOURCES DEPARTMENT

- **EXPLORE DATA ( WITH SQL )**
  - **DATA EMPLOYEES**

```
SELECT CONCAT (fisrt_name, ' ',last_name) AS Nama_Lengkap,
FLOOR(age) AS Usia, length_service AS Lama_Kerja,
absent_hour AS Lama_Absen FROM employees WHERE absent_hour = 0
GROUP BY fisrt_name,last_name,age,absent_hour,length_service
ORDER BY usia
```

- ***Menampilkan nama karyawan yang tidak pernah absent (lama_absent = 0) diurutkan berdasarkan usia yang paling muda.***
- ***Terdapat hal yang tidak wajar didalam data tersebut. dapat dilihat ada staff-staff yang berusia <15 tahun dengan masa kerja antara 3 - 6 tahun. Berarti bahwa staf-staf tersebut telah bekerja dengan usia <12 tahun***

| | nama_lengkap text | usia double precision | lama_kerja double precision | lama_absen double precision |
|---|---|---|---|---|
| 1 | Wilson Marci | 3 | 4.67873141 | 0 |
| 2 | Smith Filomena | 6 | 3.290380056 | 0 |
| 3 | Lynam Dorothy | 7 | 6.816120798 | 0 |
| 4 | Stearns Gary | 8 | 0.527552377 | 0 |
| 5 | Moore Tamara | 8 | 5.454766407 | 0 |
| 6 | Miles Josefina | 9 | 5.90351999 | 0 |
| 7 | Shank William | 11 | 3.138604758 | 0 |
| 8 | Russo Lisa | 11 | 4.873719892 | 0 |
| 9 | Harrison Viva | 11 | 5.763168102 | 0 |
| 10 | Kaiser Noreen | 11 | 6.588401386 | 0 |
| 11 | Montgomery Marguerite | 11 | 5.244620706 | 0 |
| 12 | Parks Richard | 11 | 5.546536998 | 0 |
| 13 | Broadwater Donna | 12 | 5.894723724 | 0 |
| 14 | Maze David | 12 | 3.690920973 | 0 |
| 15 | Bryant Minerva | 12 | 3.032072975 | 0 |
| 16 | Bell Melanie | 13 | 5.756020591 | 0 |
| 17 | Black Faith | 13 | 4.69560243 | 0 |

# HUMAN RESOURCES DEPARTMENT

- **EXPLORE DATA ( WITH SQL )**
  - **INNER JOIN DATA EMPLOYEES AND DISTRIBUTION_CENTERS**

*SELECT c.id, c.name, FLOOR(AVG(e.age)) AS Rata_rata_usia, AVG(e.length_service) AS Rata_rata_masa_kerja,AVG(e.absent_hour) AS Rata_rata_lama_absen FROM employees e INNER JOIN distribution_centers c ON e.distribution_centers_id = c.id GROUP BY c.name, c.id ORDER BY c.id*

| | id integer | name character varying | rata_rata_usia double precision | rata_rata_masa_kerja double precision | rata_rata_lama_absen double precision |
|---|---|---|---|---|---|
| 1 | 1 | Memphis TN | 42 | 4.758048723935715 | 66.07615928833455 |
| 2 | 2 | Chicago IL | 42 | 4.880384673436602 | 61.281467003440184 |
| 3 | 3 | Houston TX | 41 | 4.729099971281587 | 58.50572144477607 |
| 4 | 4 | Los Angeles CA | 41 | 4.806741322385436 | 58.08348038951548 |
| 5 | 5 | New Orleans LA | 42 | 4.76575044756779 | 62.88907391036073 |
| 6 | 6 | Port Authority of New York/New Jersey NY/NJ | 42 | 4.90343505390779 | 61.42654621974135 |
| 7 | 7 | Philadelphia PA | 42 | 4.701137736878384 | 61.62005487656261 |
| 8 | 8 | Mobile AL | 41 | 4.757417332399304 | 61.85231196461815 |
| 9 | 9 | Charleston SC | 41 | 4.859489417395957 | 60.575304890275916 |
| 10 | 10 | Savannah GA | 41 | 4.662480351232845 | 60.495727273150706 |

***Menampilkan rata2 usia, rata2 masa kerja, dan rata2 lama absen berdasarkan distribution centers***

# HUMAN RESOURCES DEPARTMENT

- **EXPLORE DATA ( WITH SQL )**
  - **INNER JOIN DATA EMPLOYEES AND DISTRIBUTION_CENTERS**

*SELECT c.id, c.name, FLOOR(AVG(e.age)) AS Rata_rata_usia, AVG(e.length_service) AS Rata_rata_masa_kerja,AVG(e.absent_hour) AS Rata_rata_lama_absen FROM employees e INNER JOIN distribution_centers c ON e.distribution_centers_id = c.id GROUP BY c.name, c.id ORDER BY c.id*

| | id integer | name character varying | gender character varying (1) | rata_rata_usia double precision | rata_rata_masa_kerja double precision | rata_rata_lama_absen double precision |
|---|---|---|---|---|---|---|
| 1 | 1 | Memphis TN | F | 41 | 4.806191381366432 | 71.22370109197401 |
| 2 | 1 | Memphis TN | M | 43 | 4.7092133663980835 | 60.85455213500241 |
| 3 | 2 | Chicago IL | M | 43 | 4.968709141092235 | 56.04139091772327 |
| 4 | 2 | Chicago IL | F | 41 | 4.794559954865568 | 66.37323904899523 |
| 5 | 3 | Houston TX | M | 42 | 4.722445499744677 | 54.68364486612055 |
| 6 | 3 | Houston TX | F | 40 | 4.735999092507352 | 62.46831554470592 |
| 7 | 4 | Los Angeles CA | F | 40 | 4.693406112171642 | 61.70120293587315 |
| 8 | 4 | Los Angeles CA | M | 42 | 4.911238465747708 | 54.74787382154352 |
| 9 | 5 | New Orleans LA | F | 41 | 4.7530105512214185 | 67.32010694659361 |
| 10 | 5 | New Orleans LA | M | 43 | 4.778367549732533 | 58.50074962628433 |
| 11 | 6 | Port Authority of New York/New Jersey NY/NJ | F | 41 | 4.80562706619802 | 67.72099968518562 |
| 12 | 6 | Port Authority of New York/New Jersey NY/NJ | M | 42 | 4.995115859092806 | 55.526408864661235 |

***Menampilkan rata2 usia, rata2 masa kerja, dan rata2 lama absen berdasarkan gender pada setiap distribution centers***

# HUMAN RESOURCES DEPARTMENT

- **EXPLORE DATA ( WITH SQL )**
  - **INNER JOIN (DATA EMPLOYEES AND PRODUCT ) DAN (DISTRIBUTION_CENTERS)**

*SELECT p.distribution_center_id, c.name, COUNT(DISTINCT (CONCAT (e.fisrt_name, ' ',e.last_name))) AS Jumlah_Karyawan, AVG(e.age) as rata_rata_usia, AVG(e.length_service) as Rata_rata_lama_kerja, AVG(e.absent_hour) as Rata_Rata_Absent, COUNT(DISTINCT p.id) as Jumlah_produk FROM employees e INNER JOIN products p ON p.distribution_center_id = e.distribution_centers_id INNER JOIN distribution_centers c ON e.distribution_centers_id = c.id GROUP BY c.name, p.distribution_center_id ORDER BY p.distribution_center_id*

| | distribution_center_i integer | name character varying | jumlah_karyawan bigint | rata_rata_usia double precision | rata_rata_lama_kerja double precision | rata_rata_absent double precision | jumlah_produk bigint |
|---|---|---|---|---|---|---|---|
| 1 | 1 | Memphis TN | 839 | 42.75407426028886 | 4.758048723942426 | 66.07615928845647 | 3891 |
| 2 | 2 | Chicago IL | 831 | 42.28834599443617 | 4.88038467344369 | 61.28146700349462 | 3929 |
| 3 | 3 | Houston TX | 831 | 41.478908588015216 | 4.729099971269418 | 58.50572144470026 | 3667 |
| 4 | 4 | Los Angeles CA | 835 | 41.47713094503124 | 4.806741322387947 | 58.08348038963622 | 2761 |
| 5 | 5 | New Orleans LA | 824 | 42.260925500365325 | 4.7657504475572114 | 62.88907391042467 | 2112 |
| 6 | 6 | Port Authority of New York/N... | 835 | 42.33876935972821 | 4.9034350539133 | 61.4265462196542 | 2572 |
| 7 | 7 | Philadelphia PA | 811 | 42.22331890486783 | 4.701137736874813 | 61.62005487643439 | 2669 |
| 8 | 8 | Mobile AL | 858 | 41.68506360304681 | 4.757417332392147 | 61.85231196459547 | 2919 |
| 9 | 9 | Charleston SC | 839 | 41.74812854238941 | 4.859489417396058 | 60.57530489021188 | 2719 |
| 10 | 10 | Savannah GA | 815 | 41.82592364700737 | 4.662480351235694 | 60.4957272731642 | 1881 |

***Menampilkan, jumlah karyawan, rata2 usia, rata2 masa kerja, dan rata2 lama absen, serta jumlah produk yang dihandle pada setiap distribution centers***

# HUMAN RESOURCES DEPARTMENT

- **EXPLORE DATA ( WITH SQL )**
  - **INNER JOIN (DATA EMPLOYEES DAN DISTRIBUTION_CENTERS) DAN (PRODUCTS)**

```
SELECT CONCAT (e.fisrt_name, ' ',e.last_name) AS Nama_Lengkap, e.gender,  e.age, e.length_service  e.absent_hour, e.distribution_centers_id, c.name AS
Distribution_centers_name  c.latitude, c.longitude, COUNT(DISTINCT p.id) as Jumlah_produk FROM employees e INNER JOIN distribution_centers c
ON  e.distribution_centers_id = c.id INNER JOIN products p ON  p.distribution_center_id = e.distribution_centers_id GROUP BY Nama_Lengkap, e.gender, e.age,
e.length_service, e.absent_hour, e.distribution_centers_id, Distribution_centers_name,  c.latitude, c.longitude ORDER B e.distribution_centers_id
```

| | nama_lengkap text | gender characte | age double precision | length_service double precision | absent_hour double precision | distribution_centers_id integer | distribution_centers_name character varying | latitude double prec | longitude double precision | jumlah_produk bigint |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Abbate Laurie | F | 39.89188323 | 1.766600113 | 51.5270002 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 2 | Abbott Edward | M | 54.24806682 | 3.159910592 | 131.7836282 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 3 | Abbott George | M | 38.10977405 | 6.466156286 | 0 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 4 | Abbott Monroe | M | 62.25049073 | 4.210838124 | 163.8682021 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 5 | Adamek Adam | M | 51.09250968 | 0.330490734 | 113.7179294 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 6 | Adams Allegra | F | 53.01589958 | 5.742959344 | 174.0971093 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 7 | Adams Gregory | M | 32.12486046 | 4.158933534 | 0 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 8 | Adan Joseph | M | 36.36691048 | 2.218496065 | 0.303980077 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 9 | Adolph Roger | M | 42.52130644 | 3.9796129 | 42.14236855 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 10 | Adorno Robert | M | 30.61769644 | 5.439549939 | 16.81946984 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 11 | Aldridge Arlene | F | 30.5012917 | 3.563761327 | 27.08492361 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 12 | Alexander Lorraine | F | 32.69666964 | 4.168768011 | 0.78204329 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 13 | Alexander Sarah | F | 53.8894244 | 5.409040791 | 111.6737578 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 14 | Allen Devon | M | 49.40034753 | 4.809700994 | 106.2001379 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 15 | Alvarado Erica | F | 28.22209548 | 3.640241404 | 0 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 16 | Anaya Raymond | M | 65.08544718 | 6.608769352 | 168.2968241 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |

***Tabel ini diexport ke format csv dan akan dipakai untuk melakukan anaisis lebih jauh menggunakan Python***

***Menampilkan, nama staff, gender, usia, masa kerja, lama absen, id dan nama distribution center serta latitude dan longitudenya juga jumlah produk yang dihandle pada setiap distribution centers***

# HUMAN RESOURCES DEPARTMENT

- **DESCRIBE DATA (WITH PYTHON)**

*df.info()*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8336 entries, 0 to 8335
Data columns (total 10 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   nama_lengkap             8336 non-null   object
 1   gender                   8336 non-null   object
 2   age                      8336 non-null   float64
 3   length_service           8336 non-null   float64
 4   absent_hour              8336 non-null   float64
 5   distribution_centers_id  8336 non-null   int64
 6   distribution_centers_name  8336 non-null   object
 7   latitude                 8336 non-null   float64
 8   longitude                8336 non-null   float64
 9   jumlah_produk            8336 non-null   int64
dtypes: float64(5), int64(2), object(3)
memory usage: 651.4+ KB
```

*Pada data tersebut terdapat 8336 Baris dan 10 Kolom*

# HUMAN RESOURCES DEPARTMENT

- **EXPLORE DATA (WITH PYTHON)**
  - **Descriptive Statistics**

*df.describe().T*

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| age | 8336.0 | 41.502039 | 9.948626 | 3.000000 | 35.000000 | 42.000000 | 48.000000 | 77.000000 |
| length_service | 8336.0 | 4.782910 | 2.462990 | 0.012098 | 3.575892 | 4.600248 | 5.623922 | 43.735239 |
| absent_hour | 8336.0 | 61.283978 | 49.038365 | 0.000000 | 19.127590 | 56.005808 | 94.284692 | 272.530123 |
| distribution_centers_id | 8336.0 | 5.493762 | 2.872447 | 1.000000 | 3.000000 | 5.000000 | 8.000000 | 10.000000 |
| latitude | 8336.0 | 34.667625 | 4.334152 | 29.760400 | 30.694400 | 32.783300 | 39.950000 | 41.836900 |
| longitude | 8336.0 | -87.987046 | 12.073775 | -118.250000 | -90.066700 | -88.043100 | -79.933300 | -73.783400 |
| jumlah_produk | 8336.0 | 2915.999640 | 669.356455 | 1881.000000 | 2572.000000 | 2761.000000 | 3667.000000 | 3929.000000 |

- ***Rata2 usia staff yaitu 41 tahun dengan minimum usia yaitu 3 tahun dan maksimum usia yaitu 77 tahun.***
- ***Rata2 masa kerja staff yaitu 4 Tahun dengan minimum masa kerja yaitu 0.0 Tahun dan maksimum masa kerja yaitu 43 tahun.***
- ***Rata2 absen hours yaitu 61 jam dengan minimum absen yaitu 0 jam dan maksimum yaitu 272 jam***
- ***Dari Hasil-hasil tersebut terdapat beberapa data yang tidak wajar yang perlu untuk ditangani lebih lanjut seperti minimum usia staff serta masa kerja yaitu 0.0 Tahun***

# HUMAN RESOURCES DEPARTMENT

- **VERIFY DATA QUALITY (WITH PYTHON)**
  - **Check Duplicate Row**
    - *df[df.duplicated(keep=False)]*
    - *df.duplicated()*

***Tidak ada duplikat row***

| nama_lengkap | gender | age | length_service | absent_hour | distribution_centers_id | distribution_centers_name | latitude | longitude | jumlah_produk |
|---|---|---|---|---|---|---|---|---|---|

  - **Check Missing Value**

*df.isnull().sum()*

```
nama_lengkap              0
gender                    0
age                       0
length_service            0
absent_hour               0
distribution_centers_id   0
distribution_centers_name 0
latitude                  0
longitude                 0
jumlah_produk             0
dtype: int64
```

***Tidak ada missing value***

  - **Check Unique Values**

*df.nunique()*

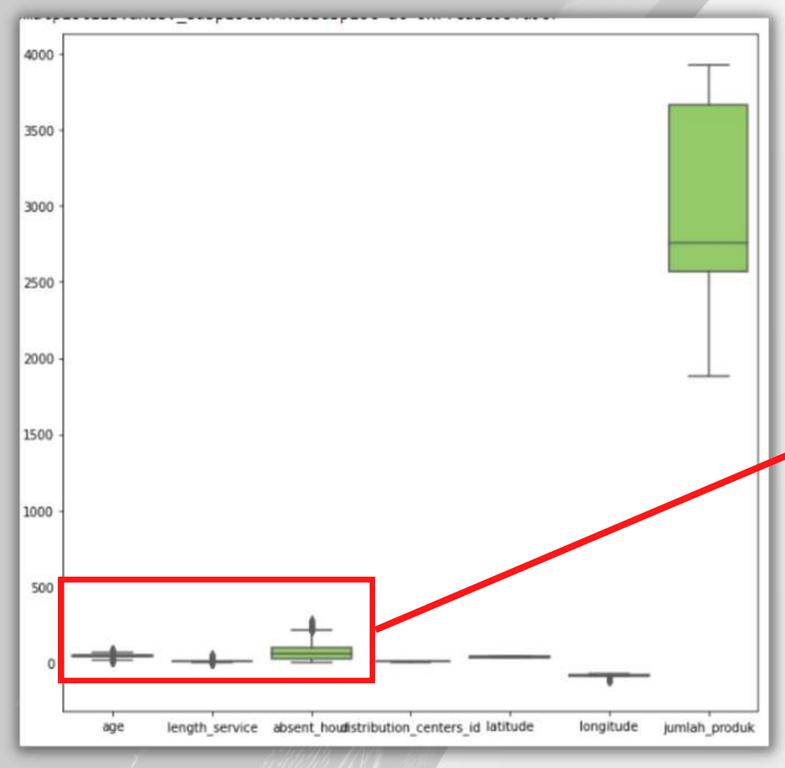
```
nama_lengkap              8209
gender                       2
age                         68
length_service            8301
absent_hour               7016
distribution_centers_id     10
distribution_centers_name   10
latitude                    10
longitude                   10
jumlah_produk               10
dtype: int64
```

***Jumlah Unique Values pada masing-masing kolom***

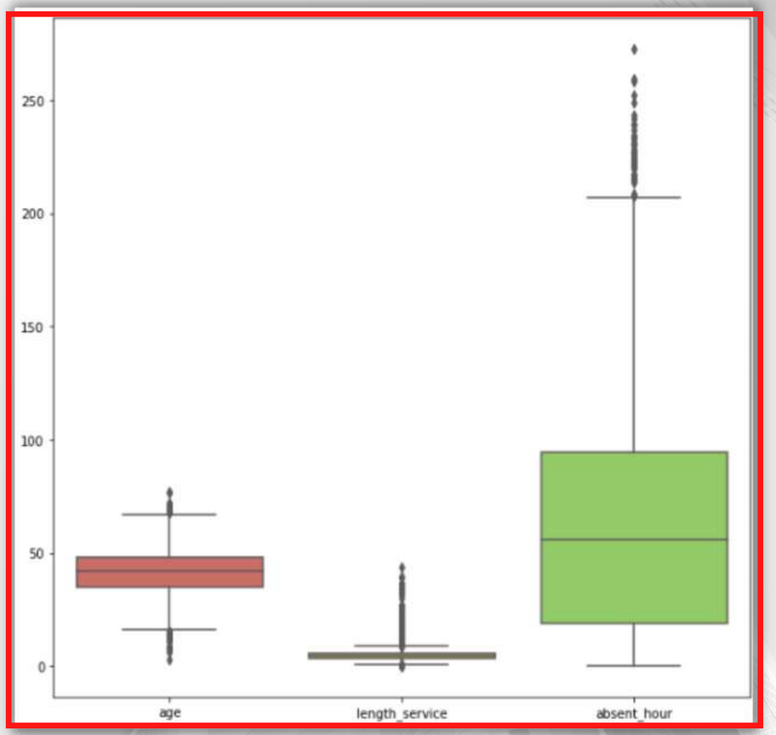
# HUMAN RESOURCES DEPARTMENT

- **VERIFY DATA QUALITY (WITH PYTHON)**
  - **Check Outliers**

*fig, ax = plt.subplots(figsize=(10,10))*
*sns.boxplot(data=df, palette=('#db5f57', '#dbc257', '#91db57', '#57d3db'))*




*fig, ax = plt.subplots(figsize=(10,10))*
*sns.boxplot(data=df[['age','length_service','absent_hour']], palette=('#db5f57', '#dbc257', '#91db57', '#57d3db'))*




***Pada kolom age,length_service, dan absent_hour terdapat banyak outlier yang harus dianalisis dan ditangani***

# HUMAN RESOURCES DEPARTMENT

- **VERIFY DATA QUALITY (WITH PYTHON)**
  - **Mencari Nilai Ambang Batas atas dan Bawah**

```
print("Highest allowed : \n",df[['age','length_service','absent_hour']].mean() + 3*df[['age','length_service','absent_hour']].std())
print("Lowest allowed : \n",df[['age','length_service','absent_hour']].mean() - 3*df[['age','length_service','absent_hour']].std())
```

```
Highest allowed :
 age                 71.347918
length_service       12.171878
absent_hour         208.399071
dtype: float64
Lowest allowed :
 age                 11.656161
length_service      -2.606059
absent_hour        -85.831116
dtype: float64
```

***Nilai ambang batas atas dan bawah pada masing-masing kolom yang terdeteksi memiliki outliers. Nilai-nilai ini akan di pakai untuk mengetahui rows/data yang terdeteksi sebagai outlier***

# HUMAN RESOURCES DEPARTMENT

- **VERIFY DATA QUALITY (WITH PYTHON)**
  - **Data Outliers Kolom "age"**

*df[(df['age'] > 71.34) | (df['age'] < 11.65)]*

| | nama_lengkap | gender | age | length_service | absent_hour | distribution_centers_id | distribution_centers_name | latitude | longitude | jumlah_produk |
|---|---|---|---|---|---|---|---|---|---|---|
| 240 | Fleury Susanna | F | 72 | 2.708078 | 215.694986 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 1535 | Smith Filomena | F | 6 | 3.290380 | 0.000000 | 2 | Chicago IL | 41.8369 | -87.6847 | 3929 |
| 2198 | Moore Tamara | F | 8 | 5.454766 | 0.000000 | 3 | Houston TX | 29.7604 | -95.3698 | 3667 |
| 2962 | Lynam Dorothy | F | 7 | 6.816121 | 0.000000 | 4 | Los Angeles CA | 34.0500 | -118.2500 | 2761 |
| 3006 | Miles Josefina | F | 9 | 5.903520 | 0.000000 | 4 | Los Angeles CA | 34.0500 | -118.2500 | 2761 |
| 3073 | Parks Richard | M | 11 | 5.546537 | 0.000000 | 4 | Los Angeles CA | 34.0500 | -118.2500 | 2761 |
| 3148 | Robinson Carmen | F | 72 | 5.814468 | 189.948165 | 4 | Los Angeles CA | 34.0500 | -118.2500 | 2761 |
| 3165 | Russo Lisa | F | 11 | 4.873720 | 0.000000 | 4 | Los Angeles CA | 34.0500 | -118.2500 | 2761 |
| 3705 | Higbee Connie | F | 72 | 3.657865 | 198.723435 | 5 | New Orleans LA | 29.9500 | -90.0667 | 2112 |
| 3765 | Kaiser Noreen | F | 11 | 6.588401 | 0.000000 | 5 | New Orleans LA | 29.9500 | -90.0667 | 2112 |
| 3867 | Montgomery Marguerite | F | 11 | 5.244621 | 0.000000 | 5 | New Orleans LA | 29.9500 | -90.0667 | 2112 |
| 4046 | Stearns Gary | M | 8 | 0.527552 | 0.000000 | 5 | New Orleans LA | 29.9500 | -90.0667 | 2112 |
| 4158 | Wilson Marci | F | 3 | 4.678731 | 0.000000 | 5 | New Orleans LA | 29.9500 | -90.0667 | 2112 |
| 4832 | Shank William | M | 11 | 3.138605 | 0.000000 | 6 | Port Authority of New York/New Jersey NY/NJ | 40.6340 | -73.7834 | 2572 |
| 5429 | Livingston Lester | M | 77 | 3.203850 | 44.273109 | 7 | Philadelphia PA | 39.9500 | -75.1667 | 2669 |
| 6754 | Bonneau Archie | M | 77 | 7.517812 | 164.882358 | 9 | Charleston SC | 32.7833 | -79.9333 | 2719 |
| 6886 | Detrick Nannie | F | 77 | 3.056848 | 257.924958 | 9 | Charleston SC | 32.7833 | -79.9333 | 2719 |
| 7008 | Harrison Viva | F | 11 | 5.763168 | 0.000000 | 9 | Charleston SC | 32.7833 | -79.9333 | 2719 |
| 7123 | Lee Mary | F | 72 | 4.802511 | 236.692995 | 9 | Charleston SC | 32.7833 | -79.9333 | 2719 |

***Data pada kolom age yang dideteksi sebagai outlier***

*df[(df['age'] > 71.34) | (df['age'] < 11.65)].count()*

```
nama_lengkap              19
gender                    19
age                       19
length_service            19
absent_hour               19
distribution_centers_id   19
distribution_centers_name 19
latitude                  19
longitude                 19
jumlah_produk             19
dtype: int64
```

***Jumlah Outliers pada kolom age***

# HUMAN RESOURCES DEPARTMENT

- **VERIFY DATA QUALITY (WITH PYTHON)**
  - **Data Outliers Kolom "length_service"**

*df[(df['length_service'] > 12.17) | (df['length_service'] < -2.60)]*

| | nama_lengkap | gender | age | length_service | absent_hour | distribution_centers_id | distribution_centers_name | latitude | longitude | jumlah_produk |
|---|---|---|---|---|---|---|---|---|---|---|
| 20 | Archer Renee | F | 39 | 21.147491 | 88.627496 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 46 | Baum Stephanie | F | 45 | 14.410268 | 69.478210 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 80 | Bryan Larry | M | 45 | 12.688114 | 47.511210 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 89 | Busch Alice | F | 35 | 13.707758 | 0.000000 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 108 | Carpentier William | M | 53 | 12.259574 | 0.000000 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 7654 | Clark Ozella | F | 41 | 17.692410 | 52.098354 | 10 | Savannah GA | 32.0167 | -81.1167 | 1881 |
| 7794 | Gomez Irene | F | 39 | 13.412318 | 0.000000 | 10 | Savannah GA | 32.0167 | -81.1167 | 1881 |
| 7895 | Johnson Paul | M | 36 | 18.480086 | 37.115712 | 10 | Savannah GA | 32.0167 | -81.1167 | 1881 |
| 8225 | Tandy Genoveva | F | 37 | 17.692462 | 75.158456 | 10 | Savannah GA | 32.0167 | -81.1167 | 1881 |
| 8268 | Wagner Robert | M | 54 | 17.149135 | 30.458020 | 10 | Savannah GA | 32.0167 | -81.1167 | 1881 |

***Data pada kolom length_service yang dideteksi sebagai outlier***

*df[(df['length_service'] > 12.17) | (df['length_service'] < -2.60)].count()*

```
nama_lengkap               100
gender                     100
age                        100
length_service             100
absent_hour                100
distribution_centers_id    100
distribution_centers_name  100
latitude                   100
longitude                  100
jumlah_produk              100
dtype: int64
```

***Jumlah Outliers pada kolom length_service***

# HUMAN RESOURCES DEPARTMENT

- **VERIFY DATA QUALITY (WITH PYTHON)**
  - **Data Outliers Kolom "absent_hours"**

*df[(df['age'] > 71.34) | (df['age'] < 11.65)]*

| | nama_lengkap | gender | age | length_service | absent_hour | distribution_centers_id | distribution_centers_name | latitude | longitude | jumlah_produk |
|---|---|---|---|---|---|---|---|---|---|---|
| 240 | Fleury Susanna | F | 72 | 2.708078 | 215.694986 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 270 | Glenn Jillian | F | 65 | 1.637244 | 224.351575 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 295 | Groves Martha | F | 69 | 4.564354 | 220.594185 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 397 | Jones Elia | F | 65 | 5.384517 | 233.102158 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 629 | Rieke Melinda | F | 66 | 5.464500 | 226.921085 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 704 | Springer James | M | 69 | 3.629952 | 224.426736 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 724 | Sullivan Cynthia | F | 65 | 3.033414 | 223.350398 | 1 | Memphis TN | 35.1174 | -89.9711 | 3891 |
| 874 | Bailey Helen | F | 62 | 2.703653 | 228.565495 | 2 | Chicago IL | 41.8369 | -87.6847 | 3929 |
| 1372 | Nelson Latonya | F | 67 | 3.377847 | 219.561445 | 2 | Chicago IL | 41.8369 | -87.6847 | 3929 |
| 1506 | Schmaltz Peggy | F | 62 | 5.326187 | 249.055872 | 2 | Chicago IL | 41.8369 | -87.6847 | 3929 |
| 1970 | Green Carmen | F | 69 | 4.073706 | 243.116431 | 3 | Houston TX | 29.7604 | -95.3698 | 3667 |
| 2042 | Huston Bonnie | F | 62 | 6.773482 | 230.973976 | 3 | Houston TX | 29.7604 | -95.3698 | 3667 |
| 3098 | Peterson Shannon | F | 57 | 6.345608 | 221.317630 | 4 | Los Angeles CA | 34.0500 | -118.2500 | 2761 |
| 3413 | Bigelow Caroline | F | 59 | 4.811746 | 230.167571 | 5 | New Orleans LA | 29.9500 | -90.0667 | 2112 |
| 3426 | Borrero Stephanie | F | 68 | 5.655202 | 259.532225 | 5 | New Orleans LA | 29.9500 | -90.0667 | 2112 |
| 3529 | Curry Jamie | F | 62 | 5.669483 | 238.909816 | 5 | New Orleans LA | 29.9500 | -90.0667 | 2112 |
| 3591 | Feeney Jennifer | F | 67 | 2.868066 | 223.110038 | 5 | New Orleans LA | 29.9500 | -90.0667 | 2112 |
| 4059 | Strickland Kimberly | F | 69 | 6.677262 | 272.530123 | 5 | New Orleans LA | 29.9500 | -90.0667 | 2112 |
| 4130 | Wayman Olive | F | 65 | 5.781213 | 222.679777 | 5 | New Orleans LA | 29.9500 | -90.0667 | 2112 |
| 4236 | Bloomfield Ellen | F | 63 | 5.943978 | 214.565795 | 6 | Port Authority of New York/New Jersey NY/NJ | 40.6340 | -73.7834 | 2572 |
| 4293 | Carrero Kristi | F | 67 | 6.647194 | 213.664592 | 6 | Port Authority of New York/New Jersey NY/NJ | 40.6340 | -73.7834 | 2572 |
| 4362 | Derossett Janice | F | 67 | 6.128350 | 226.195155 | 6 | Port Authority of New York/New Jersey NY/NJ | 40.6340 | -73.7834 | 2572 |

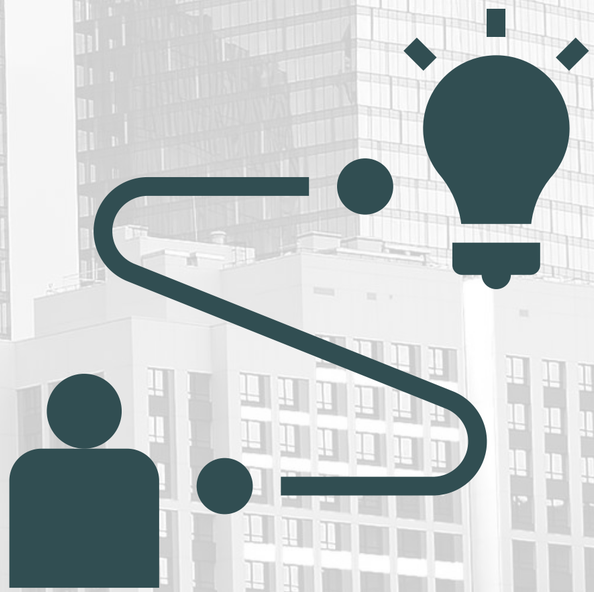
***Data pada kolom absent_hour yang dideteksi sebagai outlier***

*df[(df['age'] > 71.34) | (df['age'] < 11.65)].count()*

```
nama_lengkap              42
gender                    42
age                       42
length_service            42
absent_hour               42
distribution_centers_id   42
distribution_centers_name 42
latitude                  42
longitude                 42
jumlah_produk             42
dtype: int64
```

***Jumlah Outliers pada kolom absent_hour***

# Data Preparation

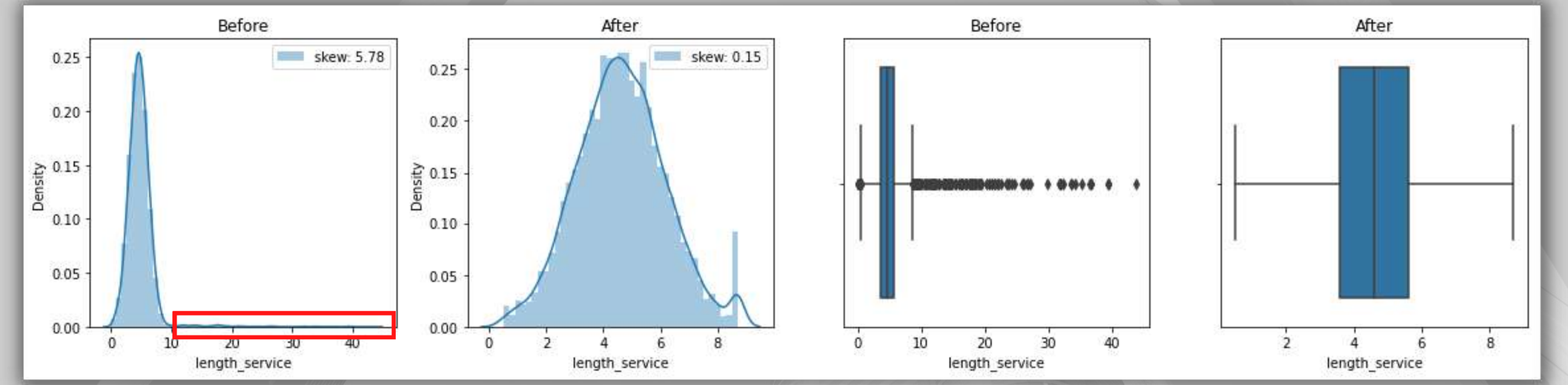


**TOOLS YANG DIGUNAKAN :**

Google Collab

# HUMAN RESOURCES DEPARTMENT

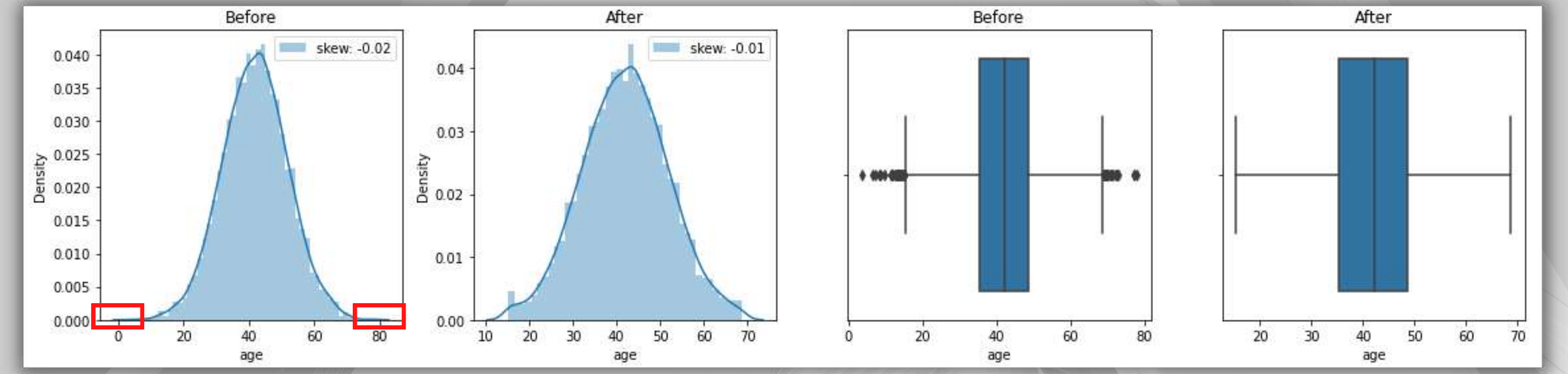
- **DATA CLEANING**
  - **Outlier Removal (IQR Method)**




- ***Pada 'before' di length service terlihat plot distribusinya terdapat outlier sehingga condong mendekati sumbu y***
- ***Pada after di length_service terlihat plot distribusinya setelah outlier dibersihkan dapat dikatakan adalah distribusi normal.***

# HUMAN RESOURCES DEPARTMENT

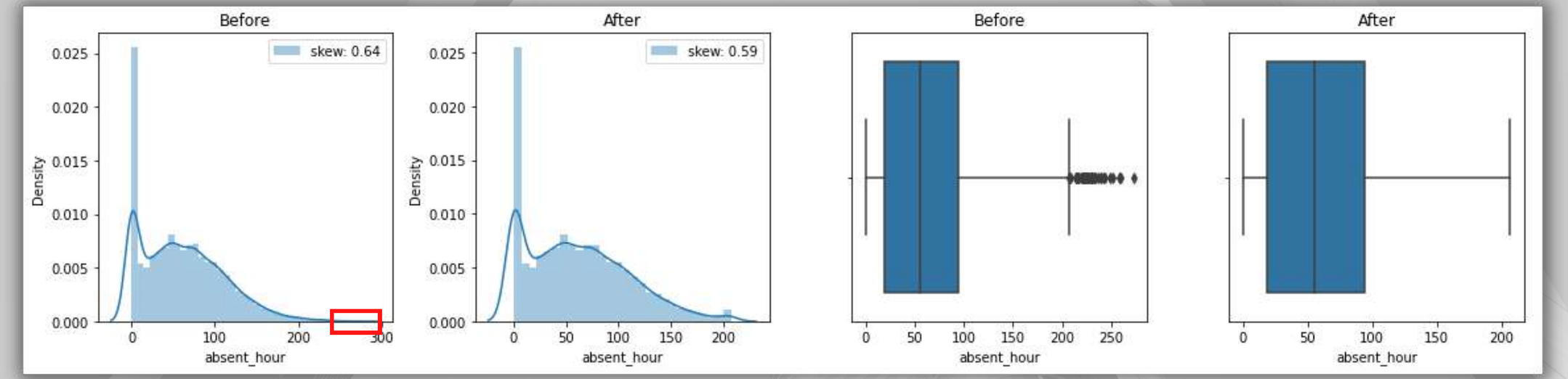
- **DATA CLEANING**
  - **Outlier Removal (IQR Method)**




- ***Pada 'before' dikolom 'age' terlihat plot distribusinya terdapat outlier***
- ***Pada after dikolom 'age terlihat plot distribusinya setelah outlier dibersihkan dapat dikatakan sebagai distribusi normal.***

# HUMAN RESOURCES DEPARTMENT

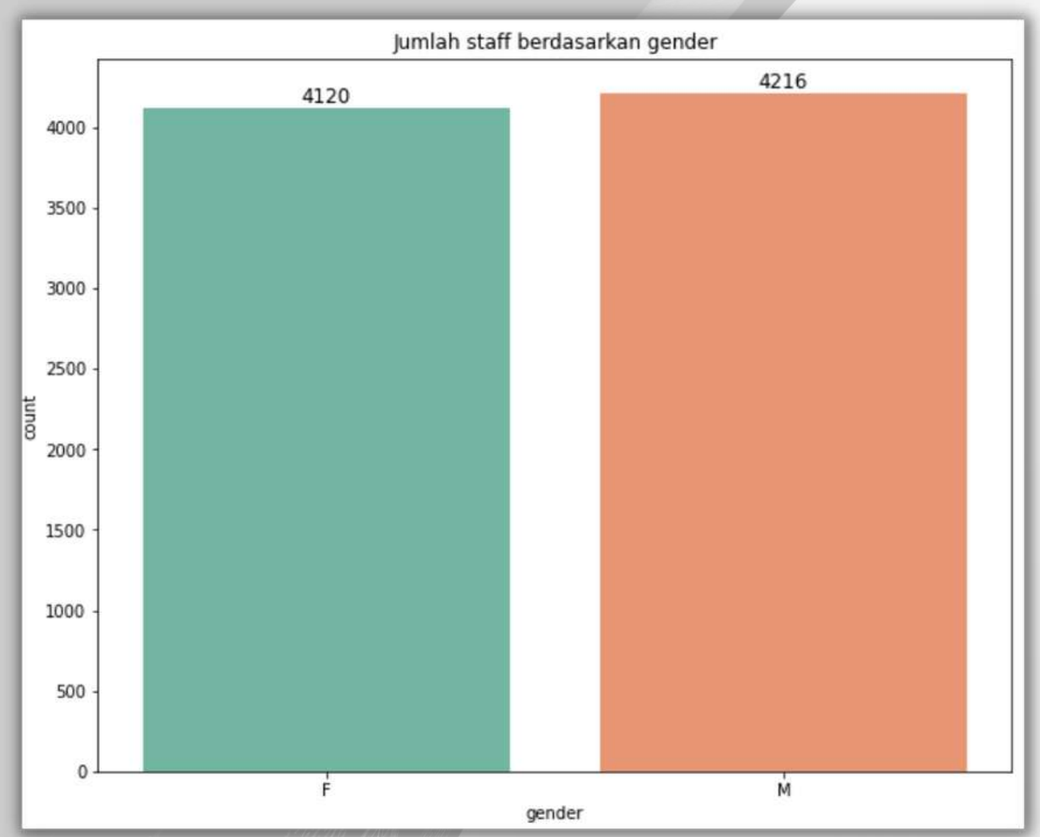
- **DATA CLEANING**
  - **Outlier Removal (IQR Method)**




- ***Pada 'before' dikolom 'absent_hour' terlihat plot distribusinya terdapat outlier***
- ***Pada after dikolom 'absent_hour' terlhat ada sedikit perbuahan pada plot distribusinya setelah outlier dibersihkan***

# HUMAN RESOURCES DEPARTMENT

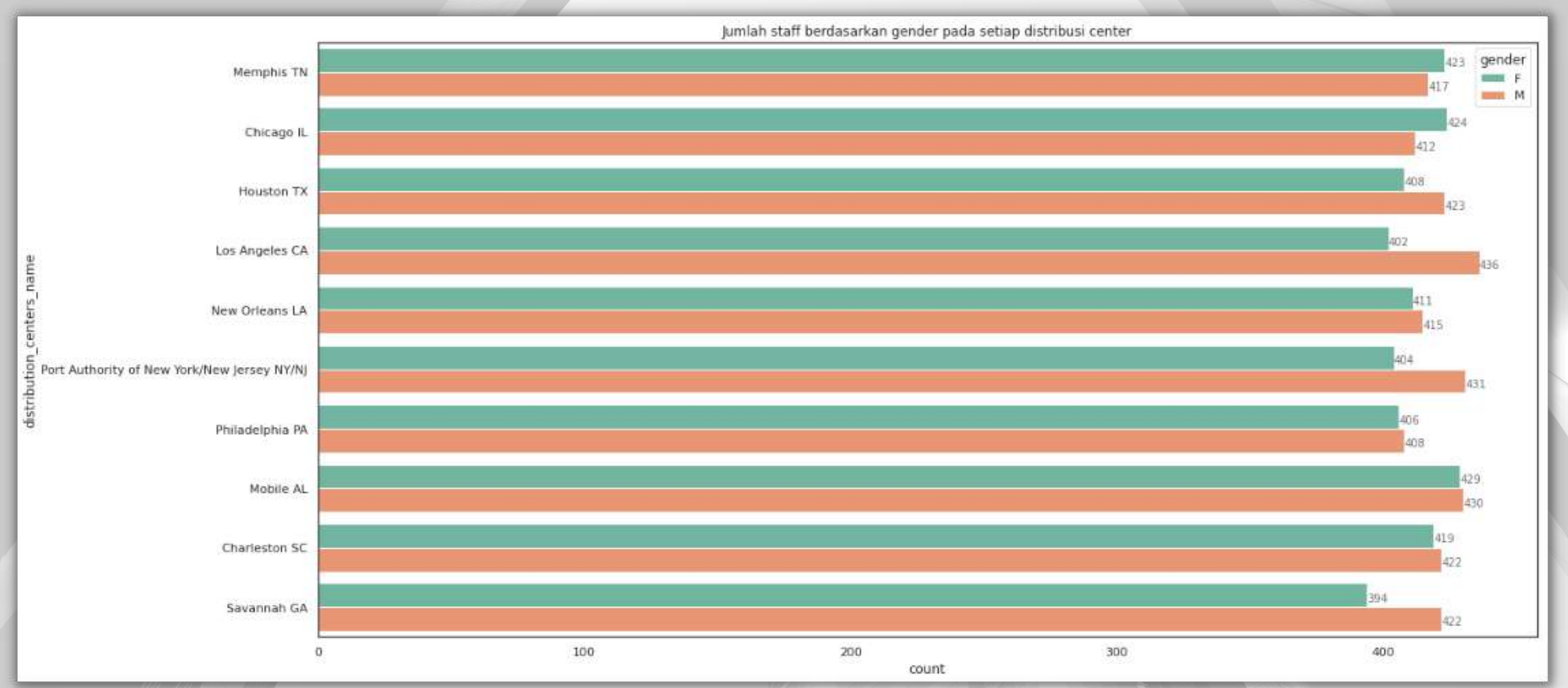
- **Expolatory  Data Analysis (EDA)**




- ***Pada Perusahaan The Look terdapat***
  - ***4120 Staff  berjenis kelamin Perempuan dan***
  - ***4216 Staff berjenis kelamin Laki-laki***

# HUMAN RESOURCES DEPARTMENT

- **Expolatory  Data Analysis (EDA)**




***Jumlah staff (berdasarkan gender) pada setiap tempat distribusi  di perusahaan The Look***

# HUMAN RESOURCES DEPARTMENT

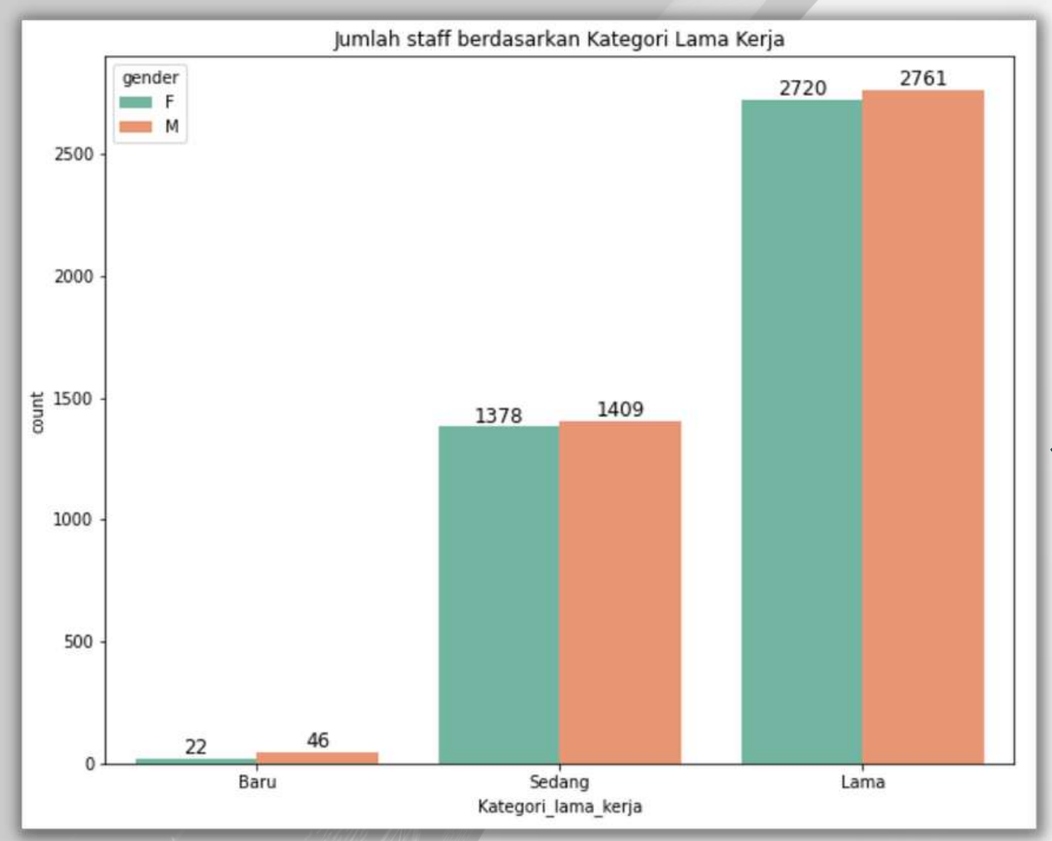
- **Expolatory  Data Analysis (EDA)**

| | distribution_centers_name | Jumlah Karyawan | age | length_service | absent_hour | Absen_rate | jumlah_produk |
|---|---|---|---|---|---|---|---|
| 0 | Charleston SC | 841 | 41.727011 | 4.672856 | 60.426914 | 2.905140 | 2719.0 |
| 1 | Chicago IL | 836 | 42.298647 | 4.647374 | 61.190412 | 2.941847 | 3929.0 |
| 2 | Houston TX | 831 | 41.492887 | 4.593761 | 58.433459 | 2.809301 | 3667.0 |
| 3 | Los Angeles CA | 838 | 41.500818 | 4.664838 | 58.066419 | 2.791655 | 2761.0 |
| 4 | Memphis TN | 840 | 42.748152 | 4.603493 | 65.933738 | 3.169891 | 3891.0 |
| 5 | Mobile AL | 859 | 41.584640 | 4.657351 | 61.672084 | 2.965004 | 2919.0 |
| 6 | New Orleans LA | 826 | 42.292633 | 4.604481 | 62.641123 | 3.011592 | 2112.0 |
| 7 | Philadelphia PA | 814 | 42.207047 | 4.633006 | 61.519165 | 2.957652 | 2669.0 |
| 8 | Port Authority of New York/New Jersey NY/NJ | 835 | 42.346885 | 4.674082 | 61.348799 | 2.949461 | 2572.0 |
| 9 | Savannah GA | 816 | 41.329809 | 4.570904 | 60.447284 | 2.906119 | 1881.0 |

***Jumlah Karywana, rata-rata (usia,length service,absent hour,absen rate) serta jumlah produk yang di handle pada setiap tempat distribusi diperusahaan The Look***

# HUMAN RESOURCES DEPARTMENT

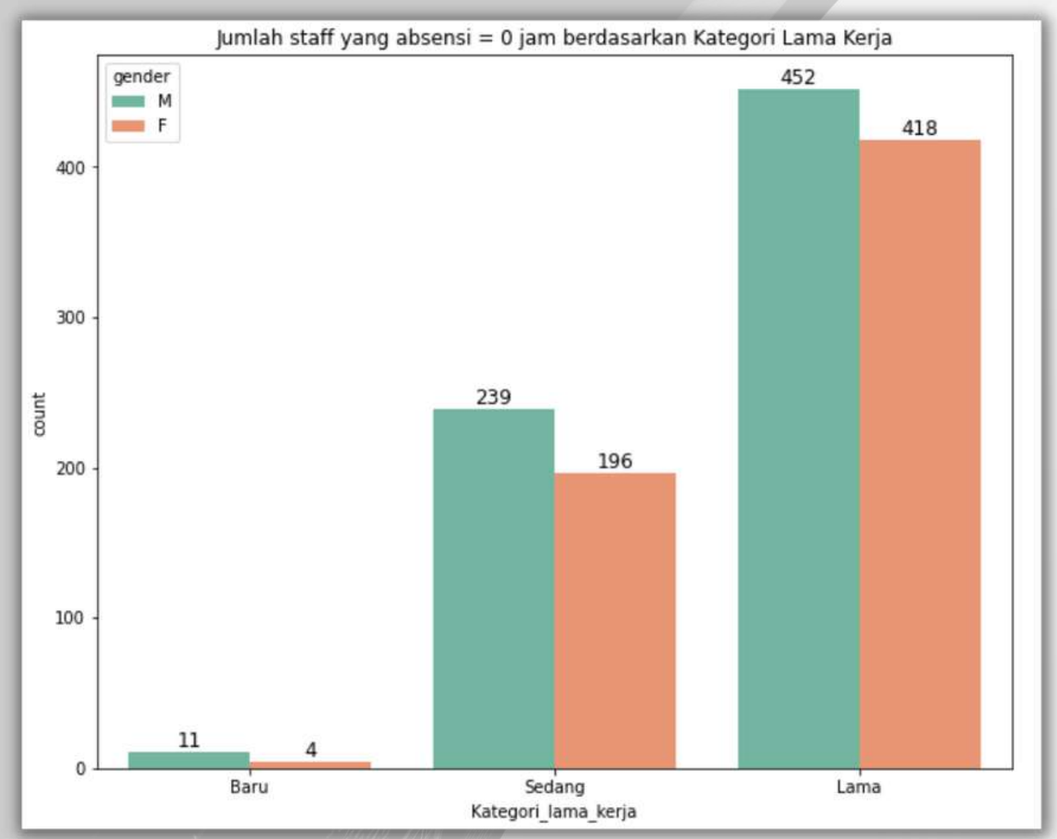
- **Expolatory  Data Analysis (EDA)**




- ***Staff dengan kategori kerja baru ( sudah bekerja 0-1 Tahun ) sebanyak 22 orang adalah perempuan dan 46 adalah laki-laki***
- ***Staff dengan kategori kerja sedang ( sudah bekerja 1-4 Tahun ) sebanyak 1378 adalah perempuan dan 1409 adalah laki-laki***
- ***Staff dengan kategori kerja lama ( sudah bekerja > 4 Tahun ) sebanyak 2720 adalah perempuan dan 2761 laki-laki***

# HUMAN RESOURCES DEPARTMENT

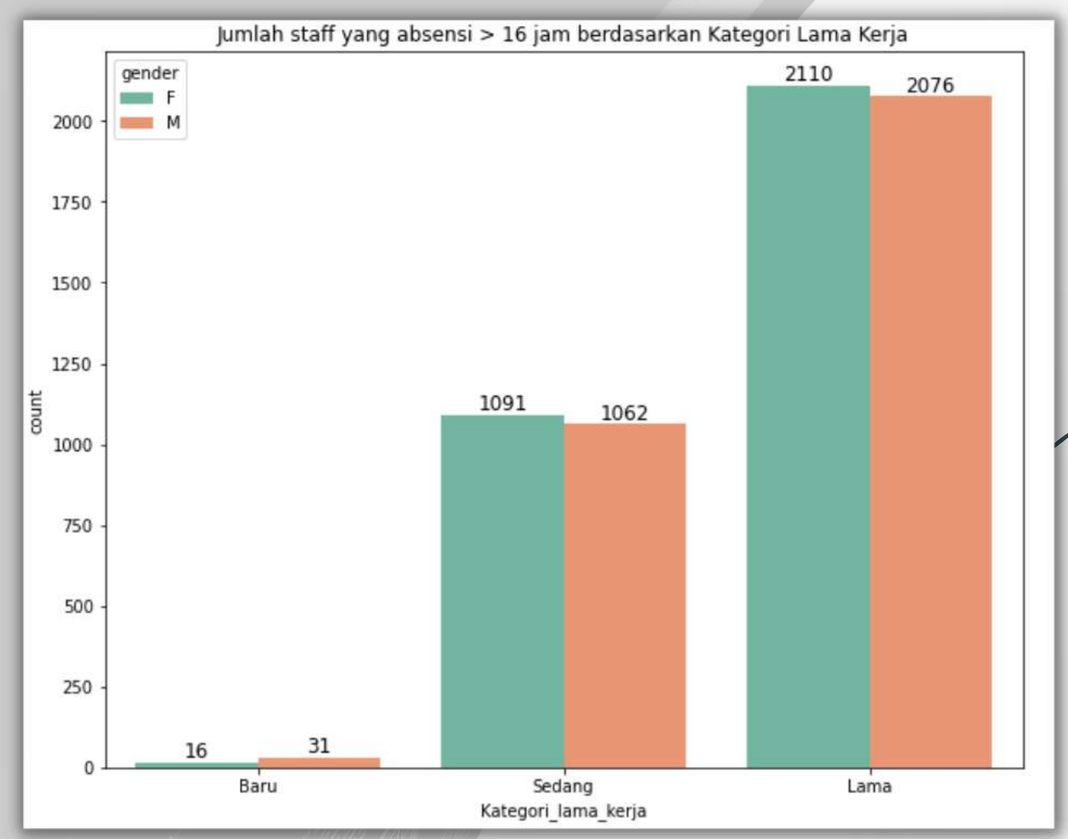
- **Expolatory  Data Analysis (EDA)**




- ***Staff dengan kategori kerja baru ( sudah bekerja 0-1 Tahun ) dan Absent Hour (total jam absen/tahun) = 0 (tidak pernah absen) sebanyak 11 Pria dan 4 Wanita***
- ***Staff dengan kategori kerja sedang ( sudah bekerja 1-4 Tahun ) dan Absent Hour (total jam absen/tahun) = 0 (tidak pernah absen) sebanyak 239 Pria dan 196 Wanita***
- ***Staff dengan kategori kerja lama ( sudah bekerja > 4 Tahun ) dan Absent Hour (total jam absen/tahun) = 0 (tidak pernah absen) sebanyak 452 Pria dan 418 Wanita***

# HUMAN RESOURCES DEPARTMENT

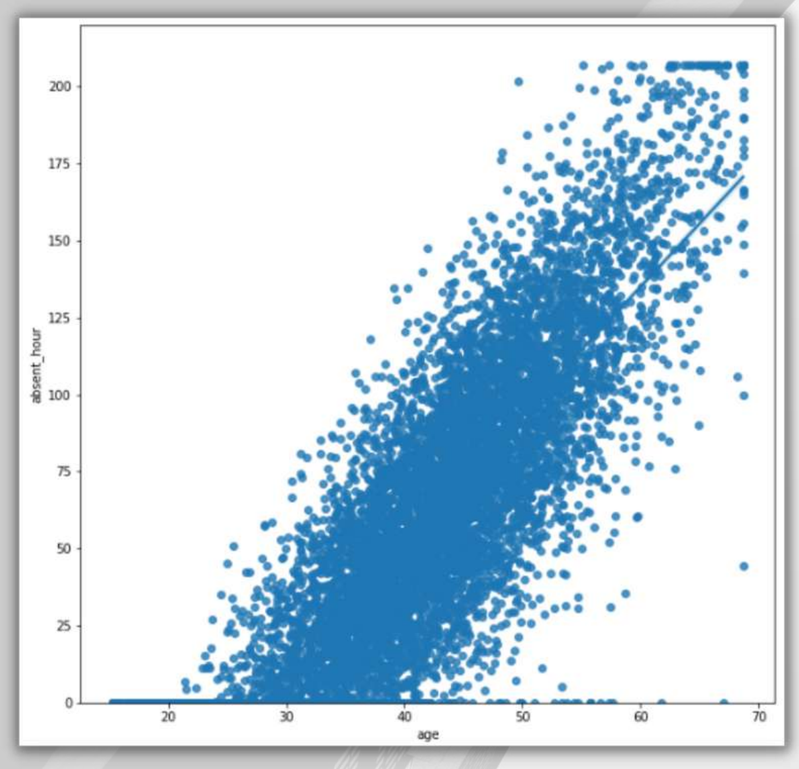
- **Expolatory Data Analysis (EDA)**



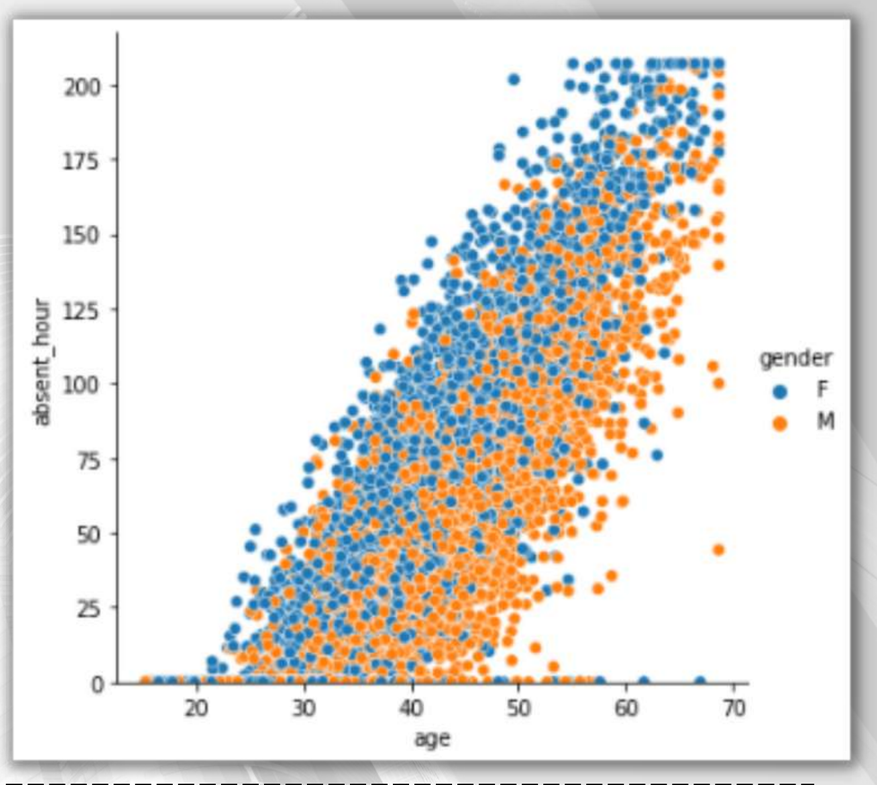

- ***Staff dengan kategori kerja baru ( sudah bekerja 0-1 Tahun ) dengan Absent Hour (total jam absen/tahun) > 16 jam sebanyak 16 Pria dan 31 Wanita***
- ***Staff dengan kategori kerja sedang ( sudah bekerja 1-4 Tahun ) dengan Absent Hour (total jam absen/tahun) > 16 jam sebanyak 1091 Pria dan 1062 Wanita***
- ***Staff dengan kategori kerja lama ( sudah bekerja > 4 Tahun ) dengan Absent Hour (total jam absen/tahun) >16 jam sebanyak 2110 Pria dan 2076 Wanita***

# HUMAN RESOURCES DEPARTMENT

- **Expolatory Data Analysis (EDA)**
  - **Korelasi**



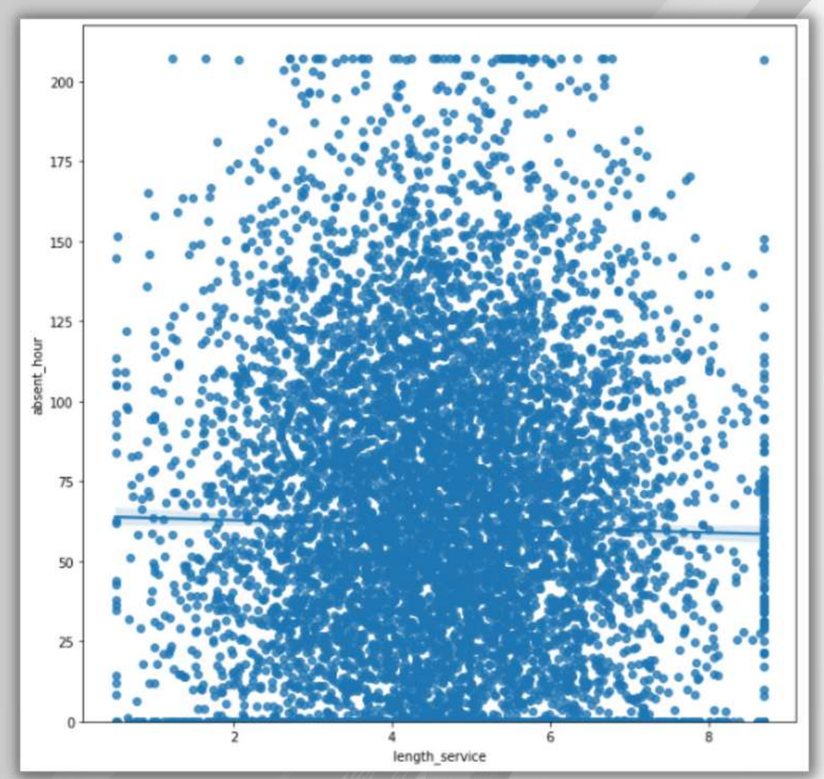

*Fitur absent_hour dan age memiliki korelasi yang positif*



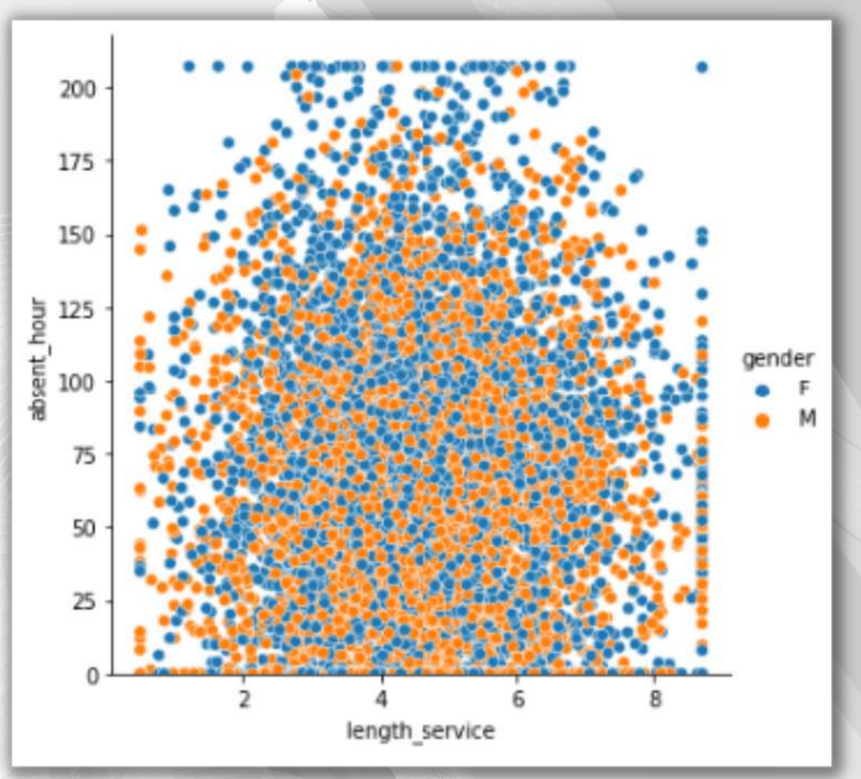

*Jumlah Absent_hour pada pria rata-rata memiliki umur lebih tua dibandingkan wanita*

# HUMAN RESOURCES DEPARTMENT

- **Expolatory Data Analysis (EDA)**
  - **Korelasi**




***Fitur absent_hour dan length_service tidak memiliki korelasi***




***Banyak staff yang absent memiliki masa kerja antara 2-7 Tahun***

# HUMAN RESOURCES DEPARTMENT

- **Expolatory Data Analysis (EDA)**
  - **Korelasi**


- ***Dapat dilihat korelasi antara age dan absent_hour sangat kuat (korelasi positif) yaitu 0.83***
- ***sedangkan korelasi antara length_service dan absent_hour bernilai negatif yaitu -0.044***

# HUMAN RESOURCES DEPARTMENT

- **Formatting and Transforming Data**
  - Label Encoding Jenis Kelamin

```
mapping_jenis_kelamin = {
     'F' : 0,
     'M' : 1
 }
df_new['gender'] = df_new['gender'].map(mapping_jenis_kelamin)
```

- ***Melakukan Label Encoding menggunakan Map Untuk jenis kelamin dengan 0 adalah Female/Wanita dan 1 adalah Male/Pria***

| | gender |
|---|---|
| 0 | 0 |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 0 |
| 6 | 1 |
| 7 | 1 |
| 8 | 1 |
| 9 | 1 |

- ***Hasil label Encoding***

# HUMAN RESOURCES DEPARTMENT

- **Formatting and Transforming Data**
  - Label Encoding Target

```
targets = np.where(df_new['absent_hour'] >=16, 1, 0)
```

- ***Target awalanya adalah ingin mengurangi jam absensi staff maksimal 16 jam/tahun,***
- ***Staff yang tidak hadir lebih dari 16 jam dianggap berlebihan (1)***
- ***Staff yang tidak hadir < 16 jam dianggap normal (0)***

| | absent_hour |
|---|---|
| 0 | 1 |
| 1 | 1 |
| 2 | 0 |
| 3 | 1 |
| 4 | 1 |
| 5 | 1 |
| 6 | 0 |
| 7 | 0 |
| 8 | 1 |
| 9 | 1 |

- ***Hasil label Encoding***

# HUMAN RESOURCES DEPARTMENT

- **Formatting and Transforming Data**

- ***Variabel Independen***

```
features=df_new[['gender','age','length_service']]
```

- ***Variabel Dependen***

```
y = df_new['absent_hour']
```

- ***Melakukan standardisasi semua fitur numerik yang sudah terbentuk***

```
scaler = StandardScaler()
scaler.fit(features)
X = pd.DataFrame(scaler.transform(features),columns=features.columns)

X.head()
```

- ***Hasil Standarisasi***

| | gender | age | length_service |
|---|---|---|---|
| **0** | -1.011583 | -0.214589 | -1.838671 |
| **1** | 0.988549 | 1.238571 | -0.944767 |
| **2** | 0.988549 | -0.394977 | 1.176418 |
| **3** | 0.988549 | 2.048591 | -0.270524 |
| **4** | 0.988549 | 0.919160 | -2.648814 |

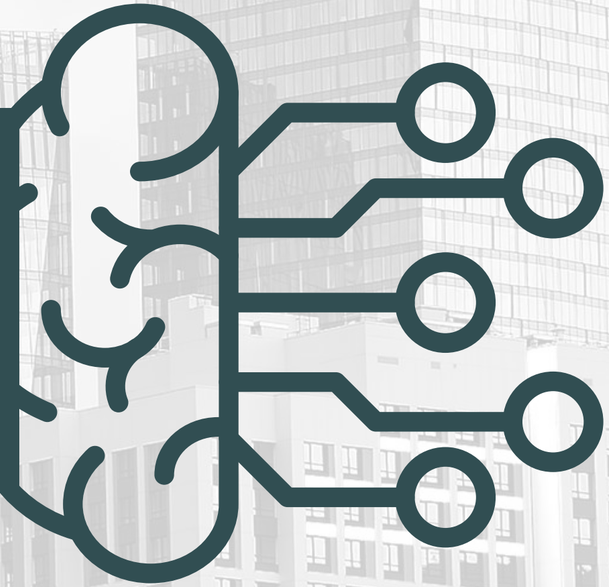
# HUMAN RESOURCES DEPARTMENT

- **Splitting Data**

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.10)
```

| | TRAINING | TESTING |
|---|---|---|
| *PERBANDINGAN* | 90 | 10 |
| *JUMLAH DATA* | 7502 | 834 |

# Modeling



**TOOLS YANG DIGUNAKAN :**

Google
Collab

# HUMAN RESOURCES DEPARTMENT

- **Modeling and Prediction**

```
from sklearn.linear_model import LogisticRegression
#Define Model
classifier = LogisticRegression(random_state = 0)
#Fit the Model
classifier.fit(X_train, y_train)
```

- *Mendefenisikan model Logistic Regression dan melakukan fit model dengan data train*

```
predicted_probability = classifier.predict_proba(X_test)
# let's check that out
predicted_probability
```

- *Prediksi Probabilitas staff yang berpotensi memiliki absnet_hour >16 jam pada dataset test*

```
model_outputs = classifier.predict(X_test)
df1= pd.DataFrame(data=model_outputs)
df1.head(10)
```

- *Melakukan Prediksi pada data testing*

- *Hasil Prediksi*
  - *0 = Absent_hour <=16 Jam*
  - *1 = Absent_hour > 16 jam*

| | 0 |
|---|---|
| 0 | 1 |
| 1 | 0 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 0 |
| 6 | 1 |
| 7 | 1 |
| 8 | 0 |
| 9 | 1 |

# HUMAN RESOURCES DEPARTMENT

- **Modeling and Prediction**

| | age | length_service | gender_F | gender_M | Probability | Prediction | Actual |
|---|---|---|---|---|---|---|---|
| **0** | 39.891883 | 1.766600 | 1.0 | 0.0 | 0.931055 | 1 | 1 |
| **1** | 54.248067 | 3.159911 | 0.0 | 1.0 | 0.997786 | 1 | 1 |
| **2** | 38.109774 | 6.466156 | 0.0 | 1.0 | 0.691260 | 1 | 0 |
| **3** | 62.250491 | 4.210838 | 0.0 | 1.0 | 0.999800 | 1 | 1 |
| **4** | 51.092510 | 0.503847 | 0.0 | 1.0 | 0.995274 | 1 | 1 |
| **5** | 53.015900 | 5.742959 | 1.0 | 0.0 | 0.998271 | 1 | 1 |
| **6** | 32.124860 | 4.158934 | 0.0 | 1.0 | 0.296256 | 0 | 0 |
| **7** | 36.366910 | 2.218496 | 0.0 | 1.0 | 0.649967 | 1 | 0 |
| **8** | 42.521306 | 3.979613 | 0.0 | 1.0 | 0.915958 | 1 | 1 |
| **9** | 30.617696 | 5.439550 | 0.0 | 1.0 | 0.191236 | 0 | 1 |

- ***Dengan 10 data disamping, dapat dilihat probabilitas untuk staff yang memiliki absent_hour > 16 jam (1) serta hasil prediksi model dan nilai aktualnya***

# HUMAN RESOURCES DEPARTMENT

- **Modeling and Prediction**
  - **P-Value**

```
import statsmodels.api as sm
# building the model and fitting the data
log_reg = sm.Logit(y_train, X_train).fit()
# printing the summary table
print(log_reg.summary())
```

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | absent_hour | No. Observations: | 7502 |
| Model: | Logit | Df Residuals: | 7499 |
| Method: | MLE | Df Model: | 2 |
| Date: | Tue, 21 Jun 2022 | Pseudo R-squ.: | 0.02869 |
| Time: | 18:24:08 | Log-Likelihood: | -3946.6 |
| converged: | False | LL-Null: | -4063.1 |
| Covariance Type: | nonrobust | LLR p-value: | 2.394e-51 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| age | 1.5057 | 0.038 | 39.659 | 0.000 | 1.431 | 1.580 |
| length_service | -0.0609 | 0.028 | -2.208 | 0.027 | -0.115 | -0.007 |
| gender_F | 0.0993 | 2.19e+06 | 4.53e-08 | 1.000 | -4.3e+06 | 4.3e+06 |
| gender_M | -0.0993 | 2.19e+06 | -4.53e-08 | 1.000 | -4.3e+06 | 4.3e+06 |

- ***P-Value pada variabel age sebesar 0.000. karena p-value < a = 0.05 Maka, dapat disimpulkan bahwa variabel age berpengaruh terhadap variabel absent_hour***
- ***P-Value pada variable length_service sebesar 0.027. karena p-value< a = 0.05 Maka dapat disimpulkan bawah variabel length_service berpengaruh terhadap variabel absent hour***
- ***P-Value pada variable gender sebesar 1.000. karena p-value > a =0.05 Maka dapat disimpulkan bahwa variable gender tidak berpengaruh terhadap varaible absent_hour***

# HUMAN RESOURCES DEPARTMENT

- **Modeling and Prediction**
  - **Odds Ratio**

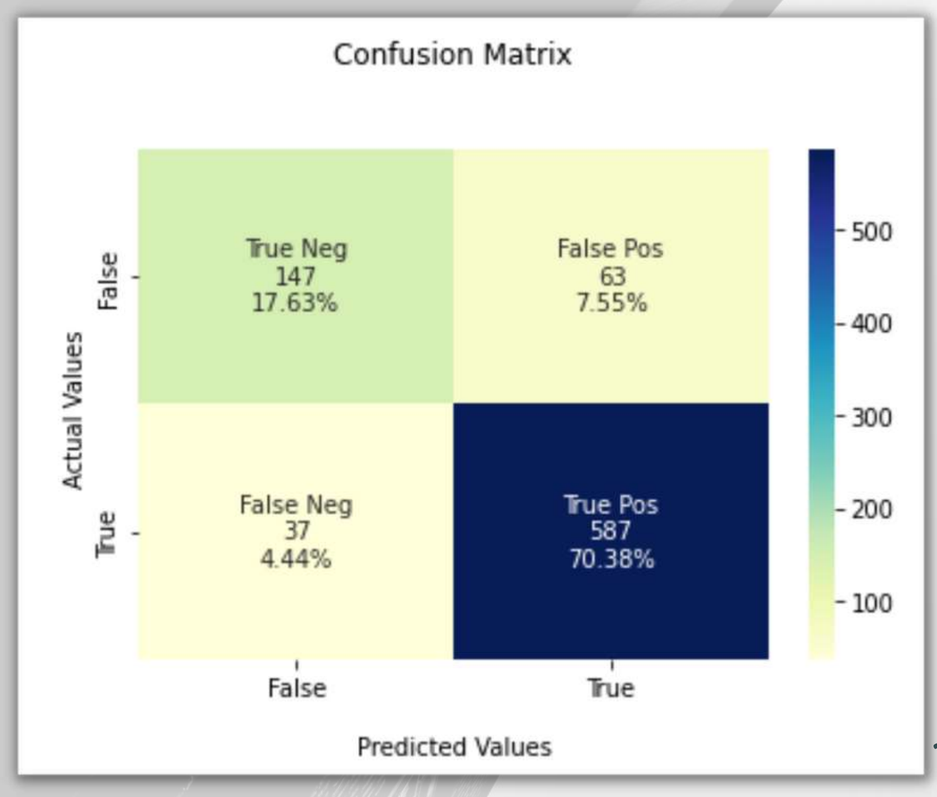
```
summary_table['Odds_ratio'] = np.exp(summary_table.Coefficient)
```

| | Feature name | Coefficient | Odds_ratio |
|---|---|---|---|
| 1 | age | 3.077707 | 21.708578 |
| 0 | Intercept | 2.594424 | 13.388877 |
| 3 | gender_F | 0.212015 | 1.236166 |
| 2 | length_service | -0.130573 | 0.877593 |
| 4 | gender_M | -0.212015 | 0.808953 |

- ***Intercept : Ketika variabel X = 0: Peluang staff memiliki absent_hours >16 jam sebesar 13.38 kali***
- ***Staff yang memiliki umur tergolong tua berpeluang memiliki jam absen yang lebih tinggi (>16 jam) sebesar 21.7 kali lebih mungkin dibandingkan staff dengan usia yang tergolong muda***
- ***Staff yang berjenis kelamin Perempuan (gender_F) berpeluang memiliki jam absensi yang tinggi (>16 jam) sebesar 1.23 kali lebih mungkin dibandingakn staff yang bukan perempuan***
- ***Staff yang berjenis kelamin Laki-laki (gender_M) berpeluang memiliki jam absensi yang tinggi (>16 jam) sebesar 0.808 kali lebih mungkin dibandingakn staff yang bukan perempuan***
- ***Staff yang memiliki length_service lama berpeluang memiliki jam absen yang lebih tinggi (>16 jam) sebesar 0.877 kali lebih mungkin dibandingan staff yang tidak memiliki length_service yang lama***

# HUMAN RESOURCES DEPARTMENT

- **Evaluation Metrics**
  - **Confusion Metrics**




***Absent berlebihan = True (1)***

- ***TRUE POSITIF = Memprediksi Staff yang absent berlebihan (>16 jam) dengan benar sebanyak 586***
- ***TRUE NEGATIF = Memprediksi Staff yang absent tidak berlebihan (<16 jam) dengan benar sebanyak 150***
- ***FALSE POSITIF = Memprediksi dengan salah bahwa data staff yang seharusnya absent <16 jam diprediksi absent > 16 jam sebanyak 55***
- ***FALSE NEGATIF = Memprediksi dengan salah bahwa data staff yang seharusnya absent > 16 jam diprediksi <16 jam sebanyak 43***

# HUMAN RESOURCES DEPARTMENT

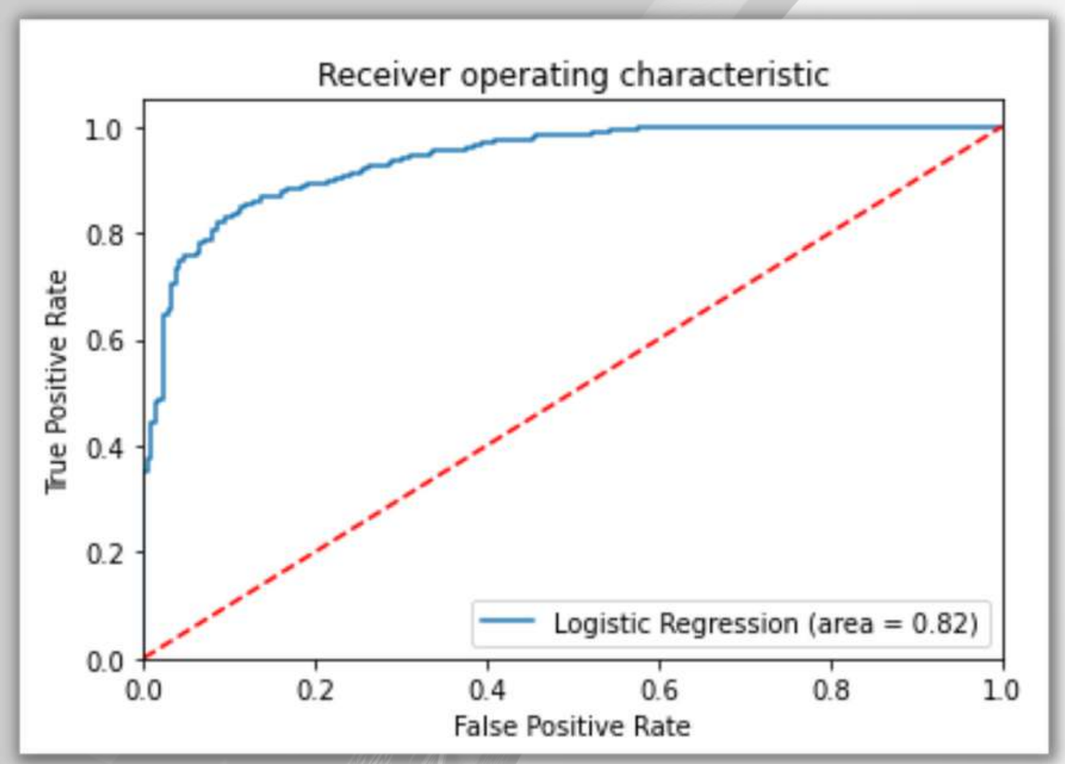
- **Evaluation Metrics**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.80 | 0.70 | 0.75 | 210 |
| 1 | 0.90 | 0.94 | 0.92 | 624 |
| accuracy | | | 0.88 | 834 |
| macro avg | 0.85 | 0.82 | 0.83 | 834 |
| weighted avg | 0.88 | 0.88 | 0.88 | 834 |

- ***Nilai Akurasi yang diperoleh sebesar 0.88 atau 88%***
- ***Nilai Precision dimana merupakan prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif sebesar 0.90 atau 90% untuk jumlah absent_hour staff yang > 16 jam***
- ***Nilai Recall dimana merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif sebesar 0.94 atau 94%***
- ***Nilai F1 Score dimana merupakan merupakan perbandingan rata-rata presisi dan recall yang dibobotkan sebesar 0.92 atau 92%***

# HUMAN RESOURCES DEPARTMENT

- **Evaluation Metrics**
  - **ROC Curve**

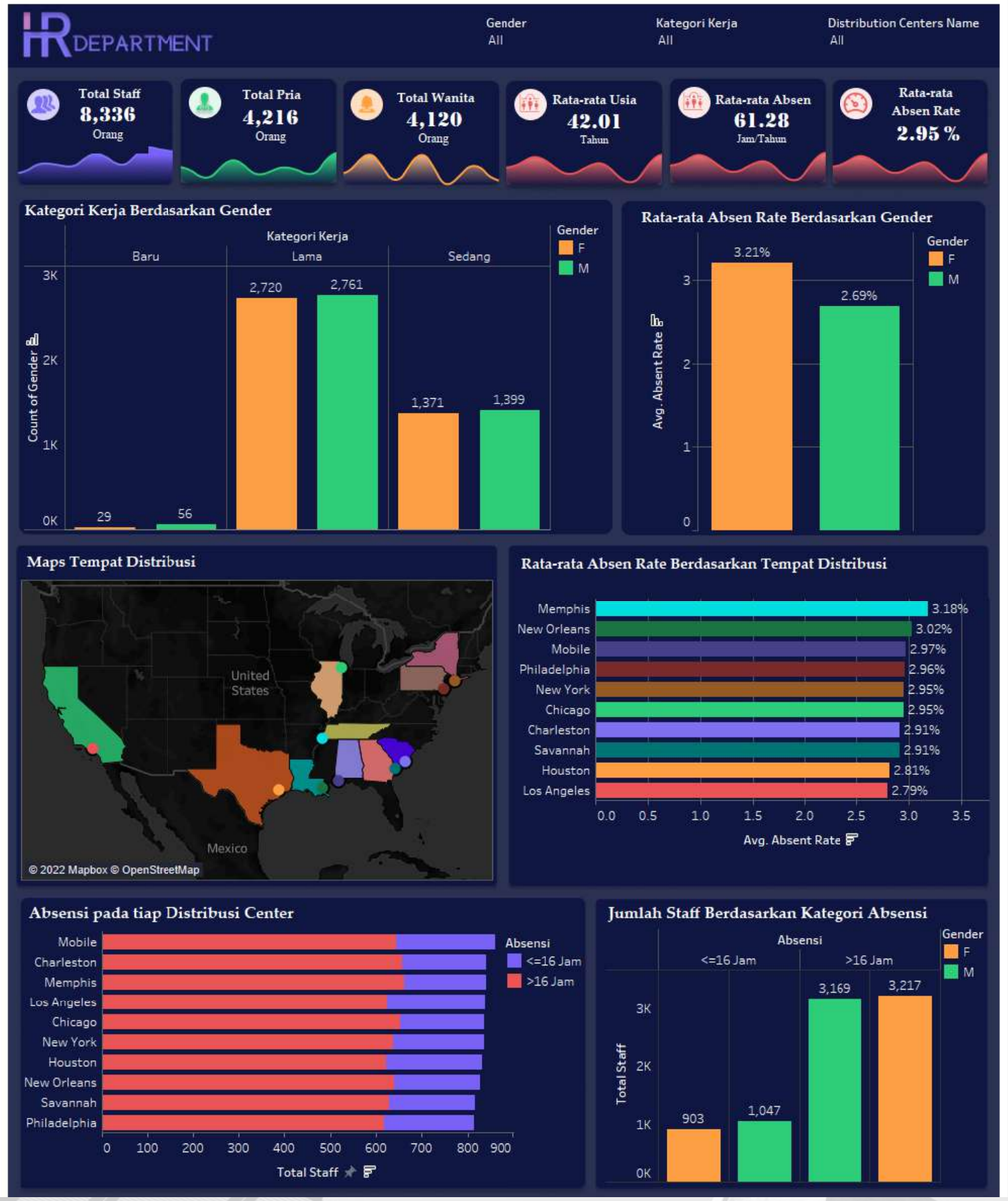



- ***Kurva ROC cenderung keatas dan berada di area True Positif. hal ini mengambbarkan bawah daerah trus positif sangat dominan pada model kita***

# DASHBOARD

tableau

# HUMAN RESOURCES DEPARTMENT





***Link Dashboard :***

*https://public.tableau.com/app/profile/richard7534/viz/FinalProject-DBA/Dashboard*

# FOR THE FULL CODE, PLEASE VISIT MY GITHUB HERE :

CLICK HERE

# THANK YOU

**Author**


**Richardo Z. Damarjanan**
Teknik Informatika
Universitas 17 Agustus 1945 SUrabaya

linkedin.com/in/richardo-damarjanan/
github.com/richardzefan
richarddamarjanan@gmail.com