

The background features three vertical stripes on the left side in shades of pink, blue, and beige. On the right side, there is a pattern of small, light pink dots arranged in a grid-like fashion, with some dots missing to create a sparse effect.

# **EXPLORATORY DATA ANALYSIS**

**Presented By : Synthetica**

# DESCRIPTIVE STATISTIC

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 4410 entries, 0 to 4409
```

```
Data columns (total 29 columns):
```

| #  | Column                  | Non-Null Count | Dtype    |
|----|-------------------------|----------------|----------|
| 0  | EmployeeID              | 4410 non-null  | int64    |
| 1  | EnvironmentSatisfaction | 4385 non-null  | category |
| 2  | JobSatisfaction         | 4390 non-null  | category |
| 3  | WorkLifeBalance         | 4372 non-null  | category |
| 4  | Age                     | 4410 non-null  | int64    |
| 5  | Attrition               | 4410 non-null  | object   |
| 6  | BusinessTravel          | 4410 non-null  | object   |
| 7  | Department              | 4410 non-null  | object   |
| 8  | DistanceFromHome        | 4410 non-null  | int64    |
| 9  | Education               | 4410 non-null  | category |
| 10 | EducationField          | 4410 non-null  | object   |
| 11 | EmployeeCount           | 4410 non-null  | int64    |
| 12 | Gender                  | 4410 non-null  | object   |
| 13 | JobLevel                | 4410 non-null  | category |
| 14 | JobRole                 | 4410 non-null  | object   |
| 15 | MaritalStatus           | 4410 non-null  | object   |
| 16 | MonthlyIncome           | 4410 non-null  | int64    |
| 17 | NumCompaniesWorked      | 4391 non-null  | float64  |
| 18 | Over18                  | 4410 non-null  | object   |
| 19 | PercentSalaryHike       | 4410 non-null  | int64    |
| 20 | StandardHours           | 4410 non-null  | int64    |
| 21 | StockOptionLevel        | 4410 non-null  | category |
| 22 | TotalWorkingYears       | 4401 non-null  | float64  |
| 23 | TrainingTimesLastYear   | 4410 non-null  | int64    |
| 24 | YearsAtCompany          | 4410 non-null  | int64    |
| 25 | YearsSinceLastPromotion | 4410 non-null  | int64    |
| 26 | YearsWithCurrManager    | 4410 non-null  | int64    |
| 27 | JobInvolvement          | 4410 non-null  | category |
| 28 | PerformanceRating       | 4410 non-null  | category |

```
dtypes: category(8), float64(2), int64(11), object(8)
```

# Data Numerikal

|       | Age         | DistanceFromHome | EmployeeCount | MonthlyIncome | NumCompaniesWorked | PercentSalaryHike | StandardHours |
|-------|-------------|------------------|---------------|---------------|--------------------|-------------------|---------------|
| count | 4410.000000 | 4410.000000      | 4410.0        | 4410.000000   | 4391.000000        | 4410.000000       | 4410.0        |
| mean  | 36.923810   | 9.192517         | 1.0           | 65029.312925  | 2.694830           | 15.209524         | 8.0           |
| std   | 9.133301    | 8.105026         | 0.0           | 47068.888559  | 2.498887           | 3.659108          | 0.0           |
| min   | 18.000000   | 1.000000         | 1.0           | 10090.000000  | 0.000000           | 11.000000         | 8.0           |
| 25%   | 30.000000   | 2.000000         | 1.0           | 29110.000000  | 1.000000           | 12.000000         | 8.0           |
| 50%   | 36.000000   | 7.000000         | 1.0           | 49190.000000  | 2.000000           | 14.000000         | 8.0           |
| 75%   | 43.000000   | 14.000000        | 1.0           | 83800.000000  | 4.000000           | 18.000000         | 8.0           |
| max   | 60.000000   | 29.000000        | 1.0           | 199990.000000 | 9.000000           | 25.000000         | 8.0           |

| TotalWorkingYears | TrainingTimesLastYear | YearsAtCompany | YearsSinceLastPromotion | YearsWithCurrManager |
|-------------------|-----------------------|----------------|-------------------------|----------------------|
| 4401.000000       | 4410.000000           | 4410.000000    | 4410.000000             | 4410.000000          |
| 11.279936         | 2.799320              | 7.008163       | 2.187755                | 4.123129             |
| 7.782222          | 1.288978              | 6.125135       | 3.221699                | 3.567327             |
| 0.000000          | 0.000000              | 0.000000       | 0.000000                | 0.000000             |
| 6.000000          | 2.000000              | 3.000000       | 0.000000                | 2.000000             |
| 10.000000         | 3.000000              | 5.000000       | 1.000000                | 3.000000             |
| 15.000000         | 3.000000              | 9.000000       | 3.000000                | 7.000000             |
| 40.000000         | 6.000000              | 40.000000      | 15.000000               | 17.000000            |

# Data Kategorikal dan Target

|        | EnvironmentSatisfaction | JobSatisfaction | WorkLifeBalance | BusinessTravel | Department             | Education | EducationField |
|--------|-------------------------|-----------------|-----------------|----------------|------------------------|-----------|----------------|
| count  | 4385.0                  | 4390.0          | 4372.0          | 4410           | 4410                   | 4410      | 4410           |
| unique | 4.0                     | 4.0             | 4.0             | 3              | 3                      | 5         | 6              |
| top    | 3.0                     | 4.0             | 3.0             | Travel_Rarely  | Research & Development | 3         | Life Sciences  |
| freq   | 1350.0                  | 1367.0          | 2660.0          | 3129           | 2883                   | 1716      | 1818           |

| Education | EducationField | Gender | JobLevel | JobRole         | MaritalStatus | Over18 | StockOptionLevel | JobInvolvement | PerformanceRating |
|-----------|----------------|--------|----------|-----------------|---------------|--------|------------------|----------------|-------------------|
| 4410      | 4410           | 4410   | 4410     | 4410            | 4410          | 4410   | 4410             | 4410           | 4410              |
| 5         | 6              | 2      | 5        | 9               | 3             | 1      | 4                | 4              | 2                 |
| 3         | Life Sciences  | Male   | 1        | Sales Executive | Married       | Y      | 0                | 3              | 3                 |
| 1716      | 1818           | 2646   | 1629     | 978             | 2019          | 4410   | 1893             | 2604           | 3732              |

| Attrition |      |
|-----------|------|
| count     | 4410 |
| unique    | 2    |
| top       | No   |
| freq      | 3699 |



## **Berdasarkan Statistik di atas, observasi yang kami dapatkan :**

A. Kolom 'EnvironmentSatisfaction', 'JobSatisfaction', 'WorkLifeBalance', 'Education', 'JobLevel', 'JobInvolvement', dan 'PerformanceRating' memiliki tipe data float64 dan int64, sehingga diubah tipe datanya menjadi tipe category untuk memudahkan analisis dan pengolahan data.

B. Terdapat beberapa kolom yang memiliki nilai kosong.

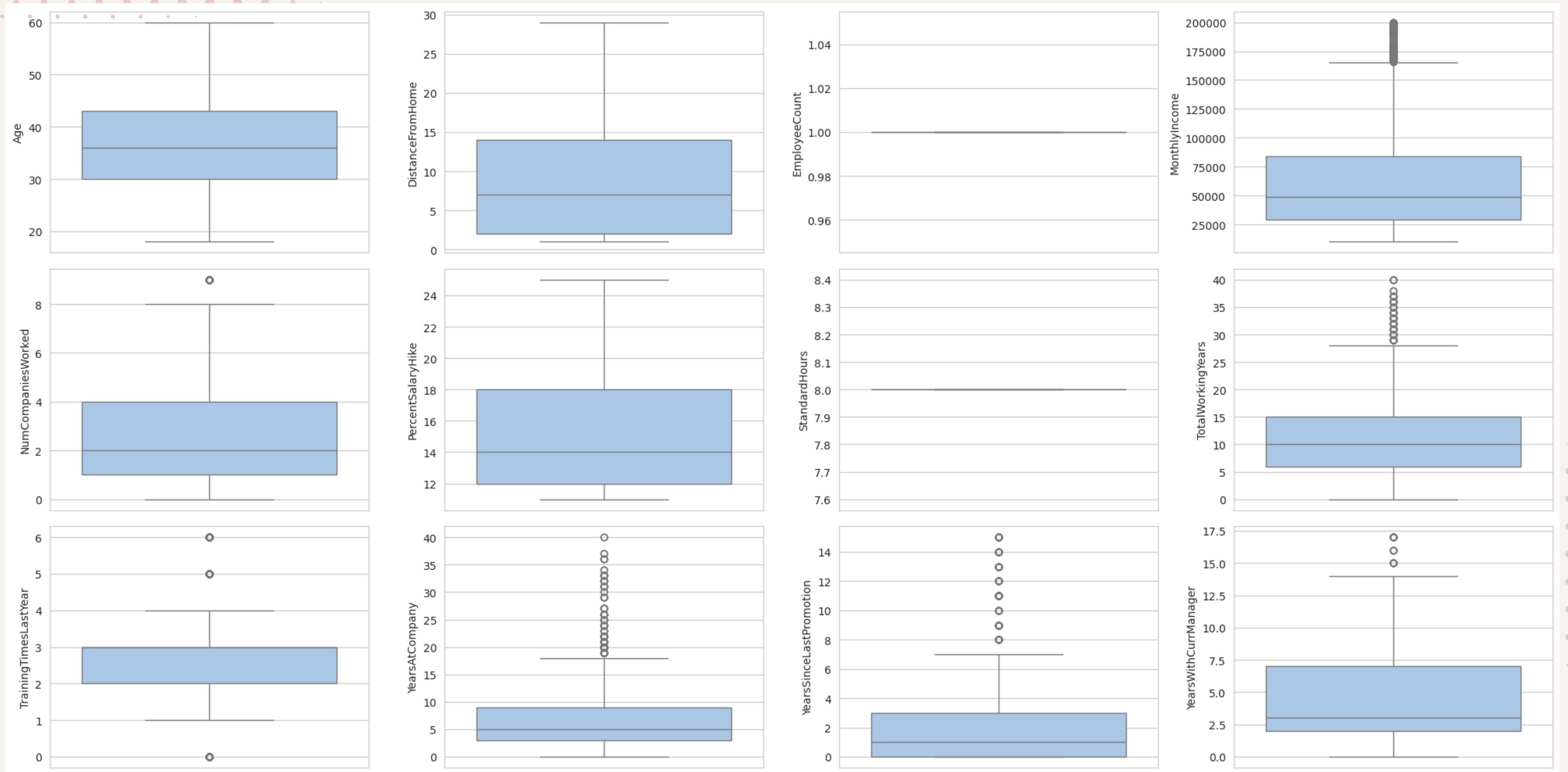
- EnvironmentSatisfaction: 25 nilai kosong, JobSatisfaction: 20 nilai kosong, WorkLifeBalance: 38 nilai kosong, NumCompaniesWorked: 19 nilai kosong, dan TotalWorkingYears: 9 nilai kosong

C. Dari tabel statistik deskriptif yang ditampilkan, berikut adalah beberapa kolom yang perlu diperhatikan:

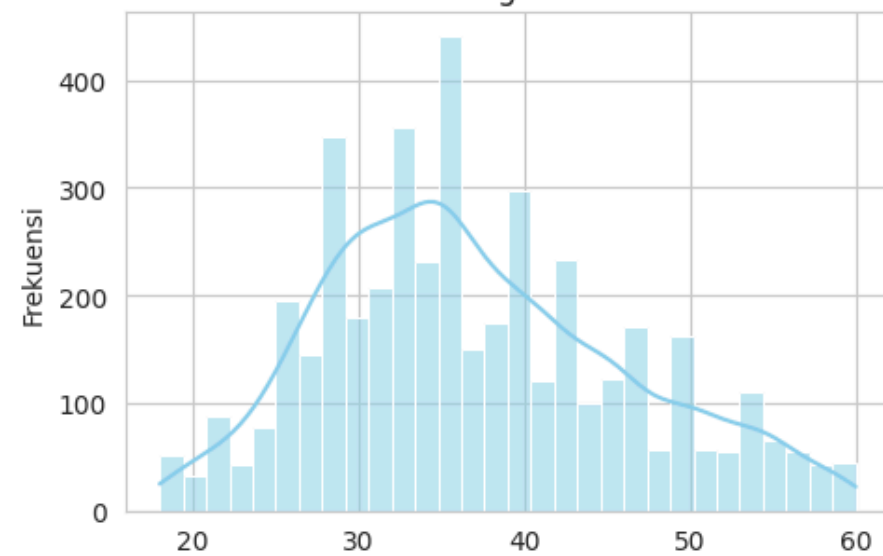
- EmployeeCount: minimum, maksimum, mean, dan nilai kuartilnya semua bernilai 1.0. Hal ini menunjukkan bahwa semua entri memiliki nilai yang sama yaitu 1, sehingga kolom ini tidak bervariasi dan tidak informatif untuk digunakan dalam analisis.
- StandardHours: minimum, maksimum, mean, dan nilai kuartilnya semua bernilai 8.0. Hal ini menunjukkan bahwa semua entri memiliki jam kerja standar yang sama, sehingga kolom ini tidak bervariasi dan tidak informatif untuk digunakan dalam analisis.
- BusinessTravel: total baris 4410 dengan nilai unique paling banyak adalah 'Travel\_Rarely' sejumlah 3129. Hal ini menunjukkan bahwa hampir 71% data kolom berisi 'Travel\_Rarely'.
- Department: total baris 4410 dengan nilai unique paling banyak adalah 'Research & Development' sejumlah 3129. Hal ini menunjukkan bahwa hampir 65% data kolom berisi 'Research & Development'.
- Over18: total baris 4410 dengan total nilai unique 1 yaitu 'Y' sejumlah 4410. Hal ini menunjukkan bahwa semua data kolom berisi 'Y', sehingga kolom ini tidak bervariasi dan tidak informatif untuk digunakan dalam analisis.
- Attrition: total baris 4410 dengan nilai unique paling banyak adalah 'No' sejumlah 3699. Hal ini menunjukkan bahwa hampir 84% data kolom berisi 'No', sehingga terjadi class imbalance pada target.

# UNIVARIATE ANALYSIS

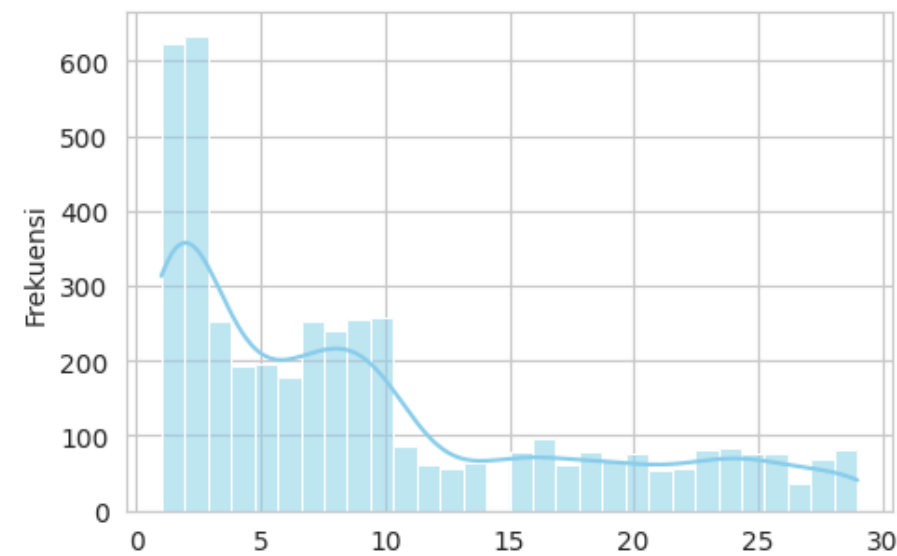
## Data Numerikal



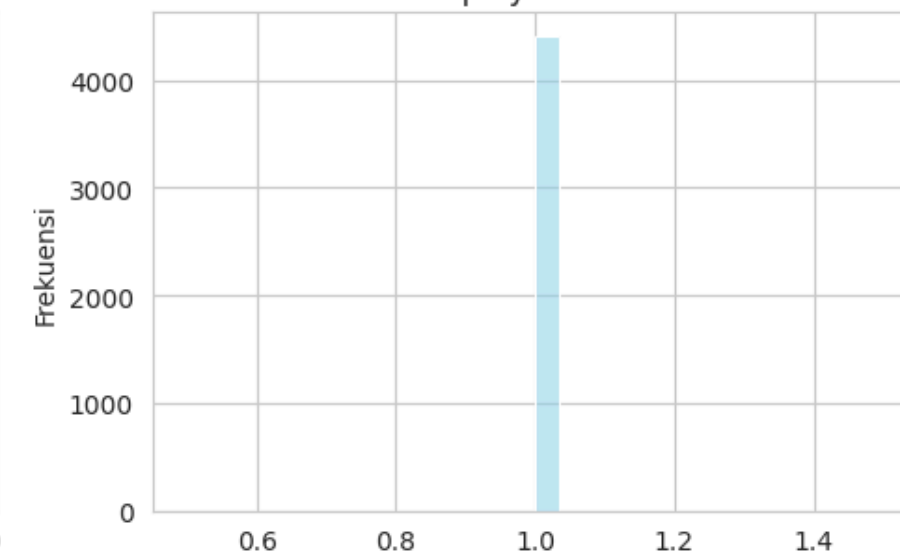
Age



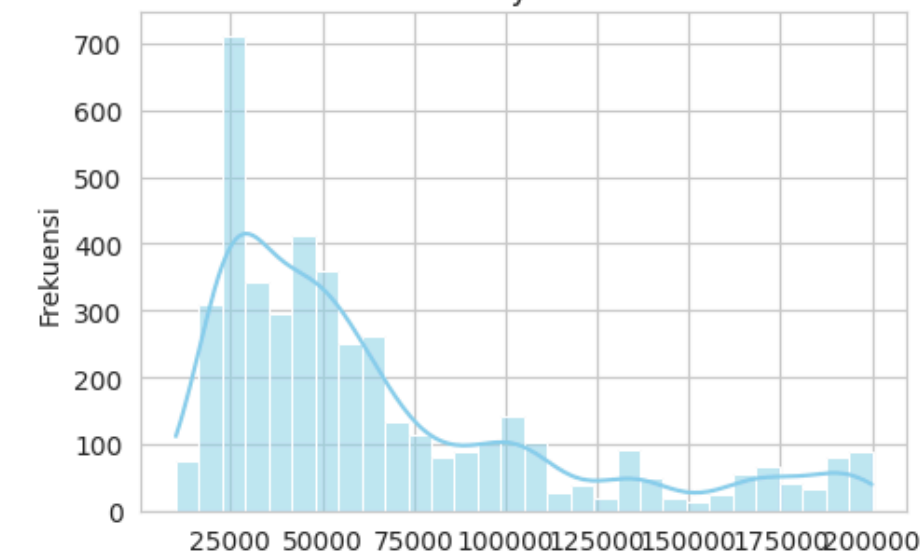
DistanceFromHome



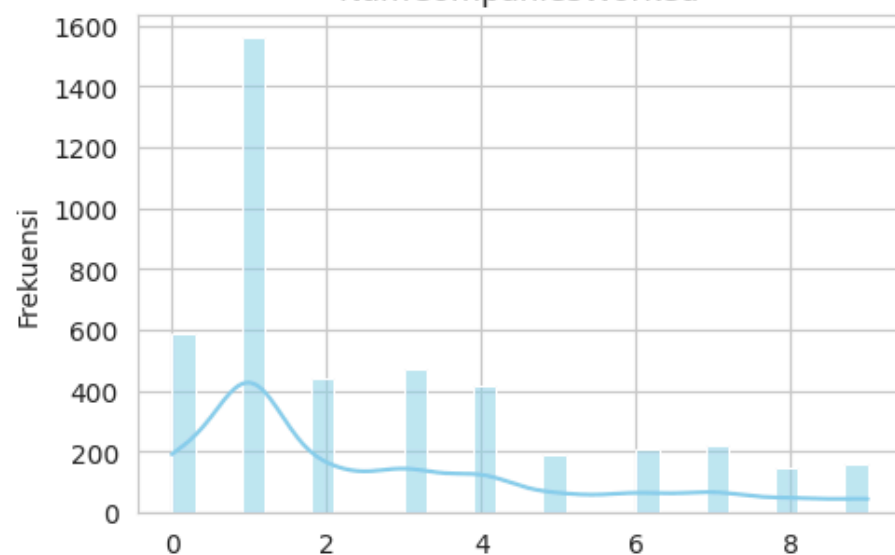
EmployeeCount



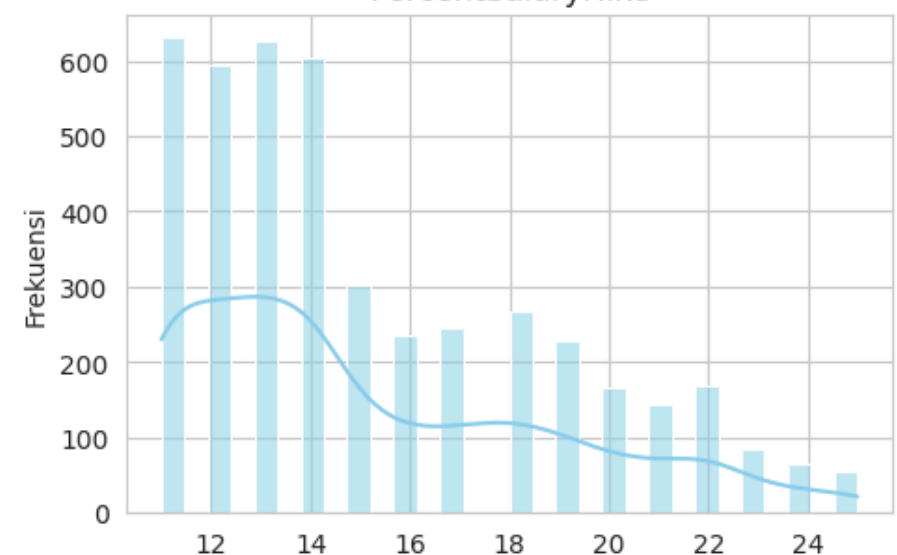
MonthlyIncome



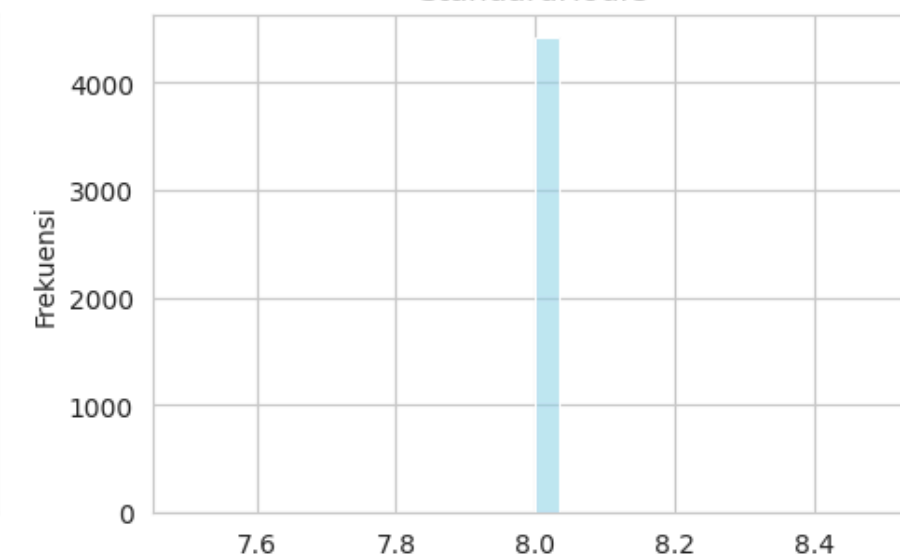
NumCompaniesWorked



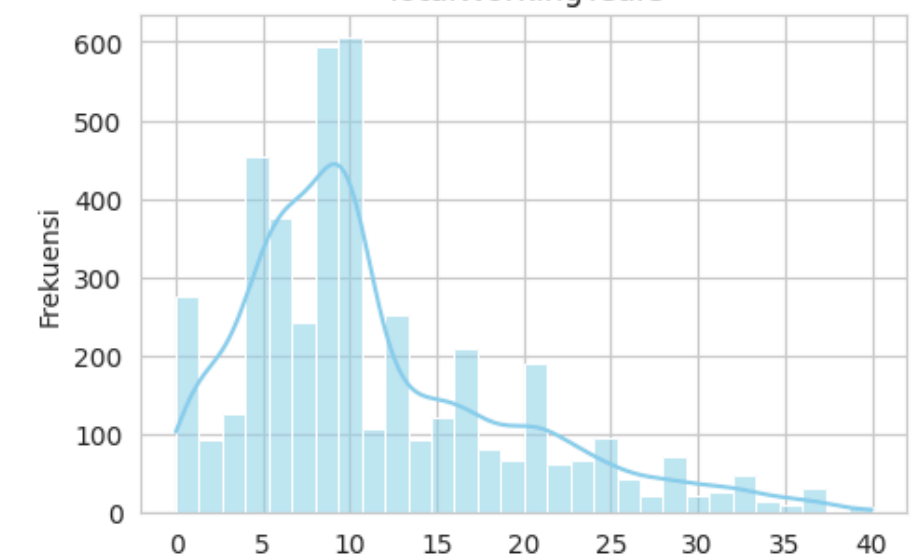
PercentSalaryHike



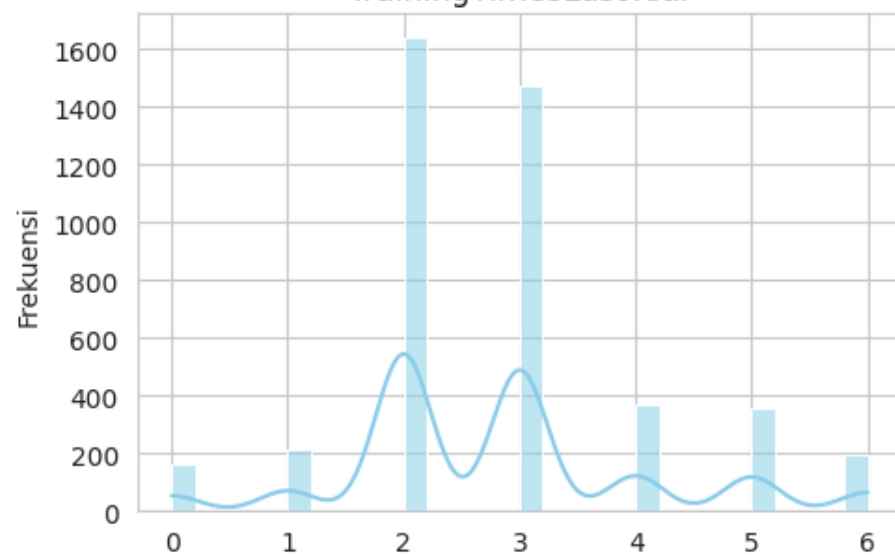
StandardHours



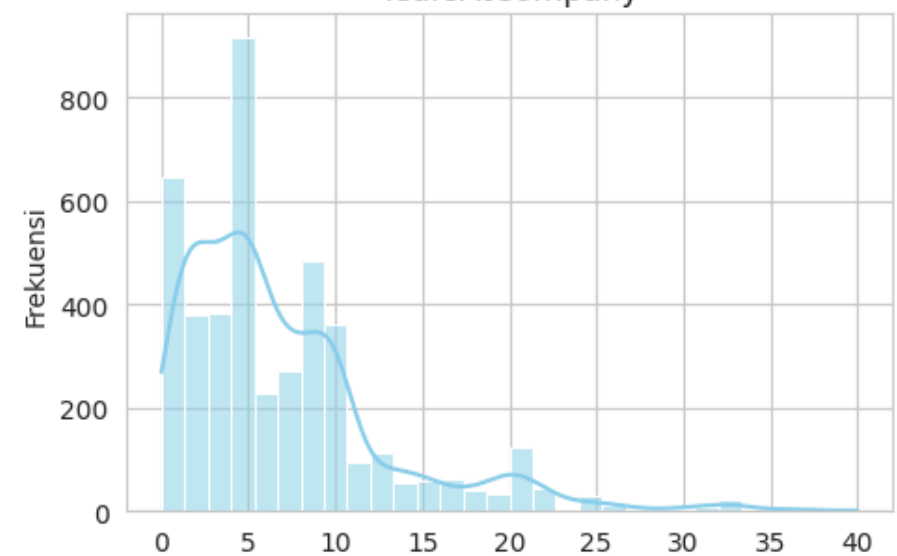
TotalWorkingYears



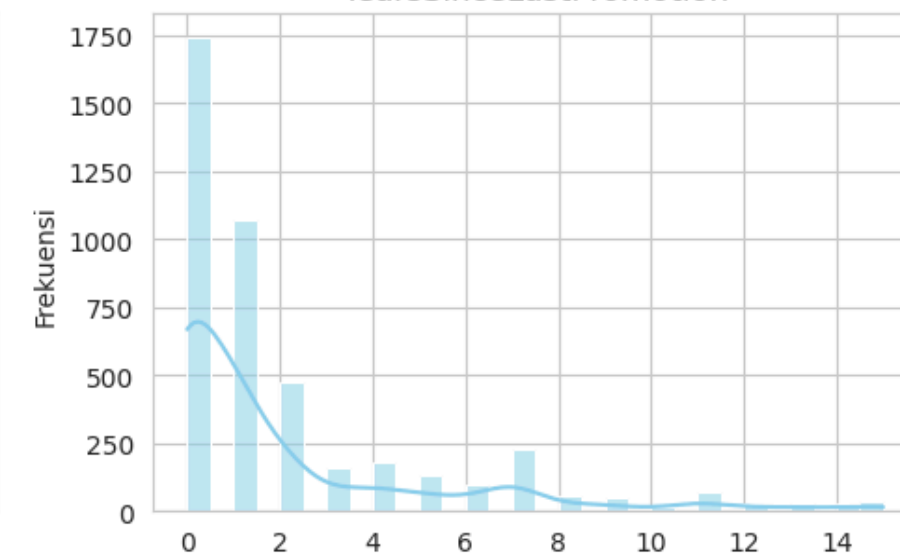
TrainingTimesLastYear



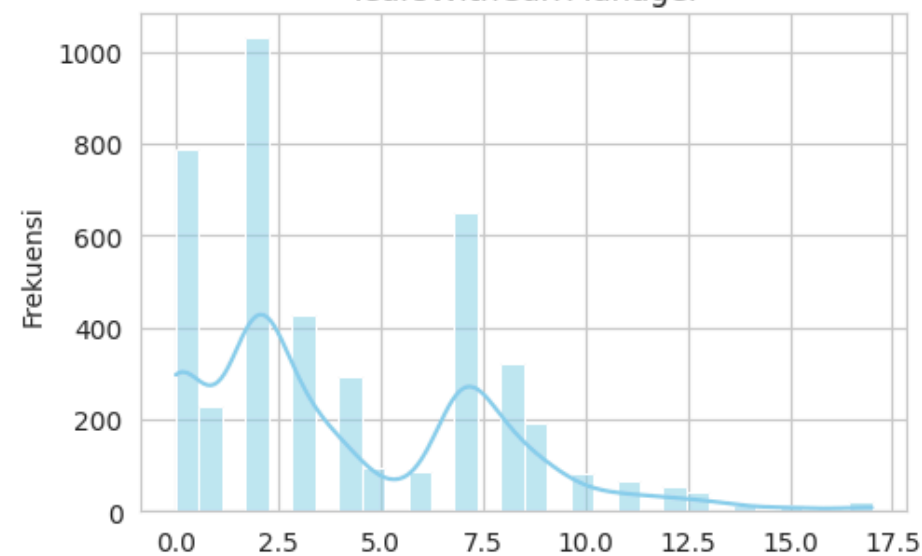
YearsAtCompany



YearsSinceLastPromotion



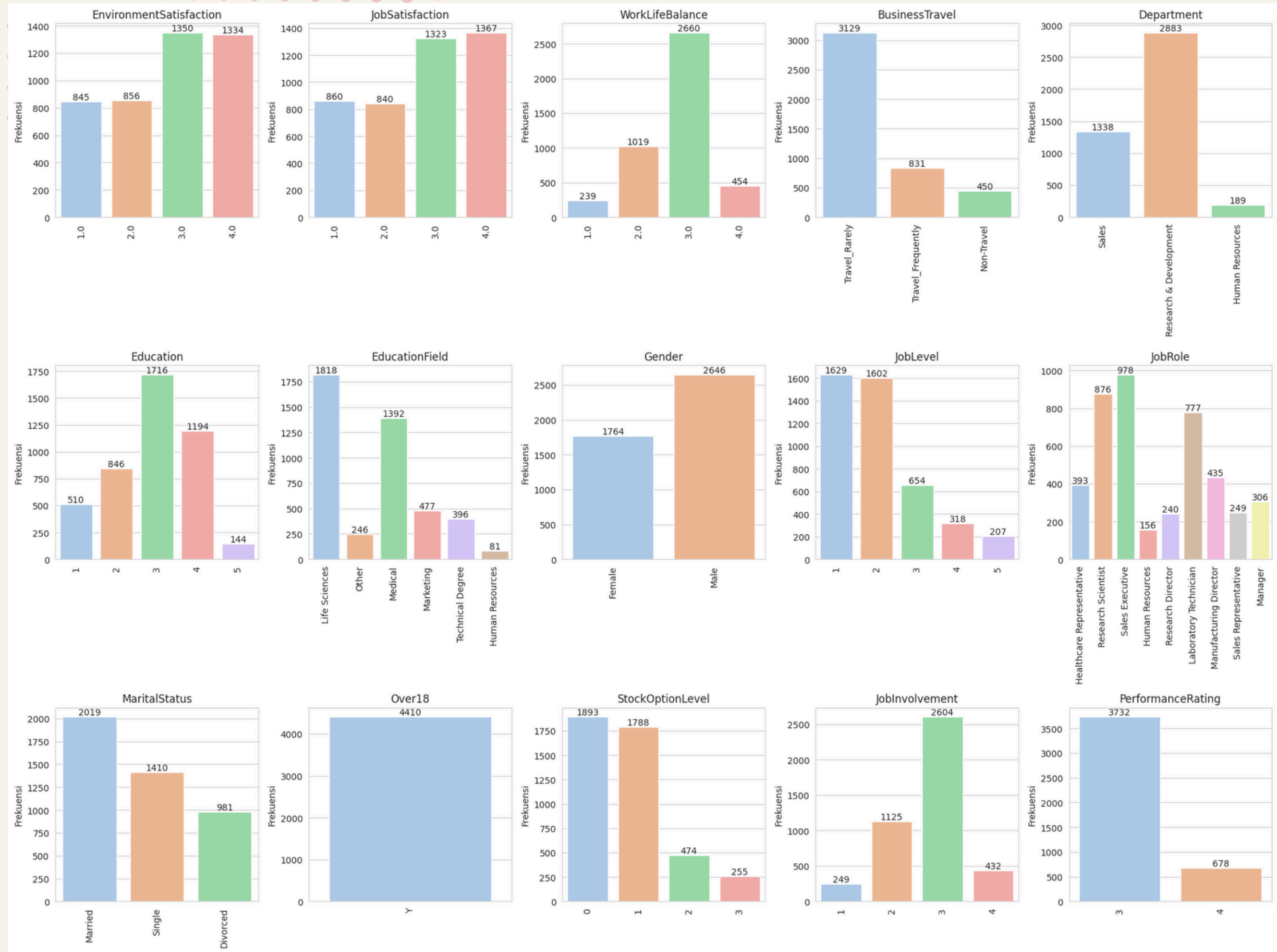
YearsWithCurrManager



## Hasil Observasi Data Numerik

1. Age: berdistribusi positively skewed dengan modus sekitar usia rentang 30-40. Ini menunjukkan mayoritas karyawan berada pada rentang usia muda hingga menengah, dengan sedikit karyawan yang usianya mendekati 60 tahun.
2. DistanceFromHome: berdistribusi positively skewed. Sebagian besar karyawan tinggal relatif dekat dari kantor (0-10 kms).
3. EmployeeCount dan StandardHours: distribusi kolom ini konstan pada nilai 1 dan 8, yang menunjukkan bahwa kedua kolom ini sebenarnya tidak memberikan informasi yang bervariasi.
4. MonthlyIncome: berdistribusi positively skewed dengan beberapa nilai yang sangat tinggi sebagai outlier. Ini menunjukkan ada beberapa karyawan dengan gaji yang jauh di atas rata-rata.
5. NumCompaniesWorked: berdistribusi positively skewed dengan mayoritas karyawan hanya pernah bekerja di 0-4 perusahaan sebelum ini.
6. PercentSalaryHike: berdistribusi positively skewed, menunjukkan kebanyakan karyawan mendapat kenaikan gaji yang kecil hingga sedang, dengan beberapa yang mendapatkan persentase kenaikan lebih besar.
7. TotalWorkingYears: berdistribusi positively skewed dengan beberapa nilai yang tinggi sebagai outlier. Outlier ini mungkin karyawan dengan pengalaman sangat panjang di berbagai perusahaan.
8. TrainingTimesLastYear: berdistribusi bimodal dengan beberapa puncak pada nilai rendah (2 atau 3 kali pelatihan).
9. YearsAtCompany: berdistribusi positively skewed dengan modus di nilai 5 tahun. Banyak karyawan yang baru beberapa tahun bekerja di perusahaan ini, sedangkan hanya sedikit yang memiliki masa kerja sangat panjang (lebih dari 20 tahun).
10. YearsSinceLastPromotion: berdistribusi positively skewed. Sebagian besar karyawan baru saja atau belum lama mendapatkan promosi, dengan hanya beberapa yang lama tidak dipromosikan.
11. YearsWithCurrManager: berdistribusi positively skewed, menunjukkan sebagian besar karyawan belum lama bekerja dengan manajer mereka saat ini.



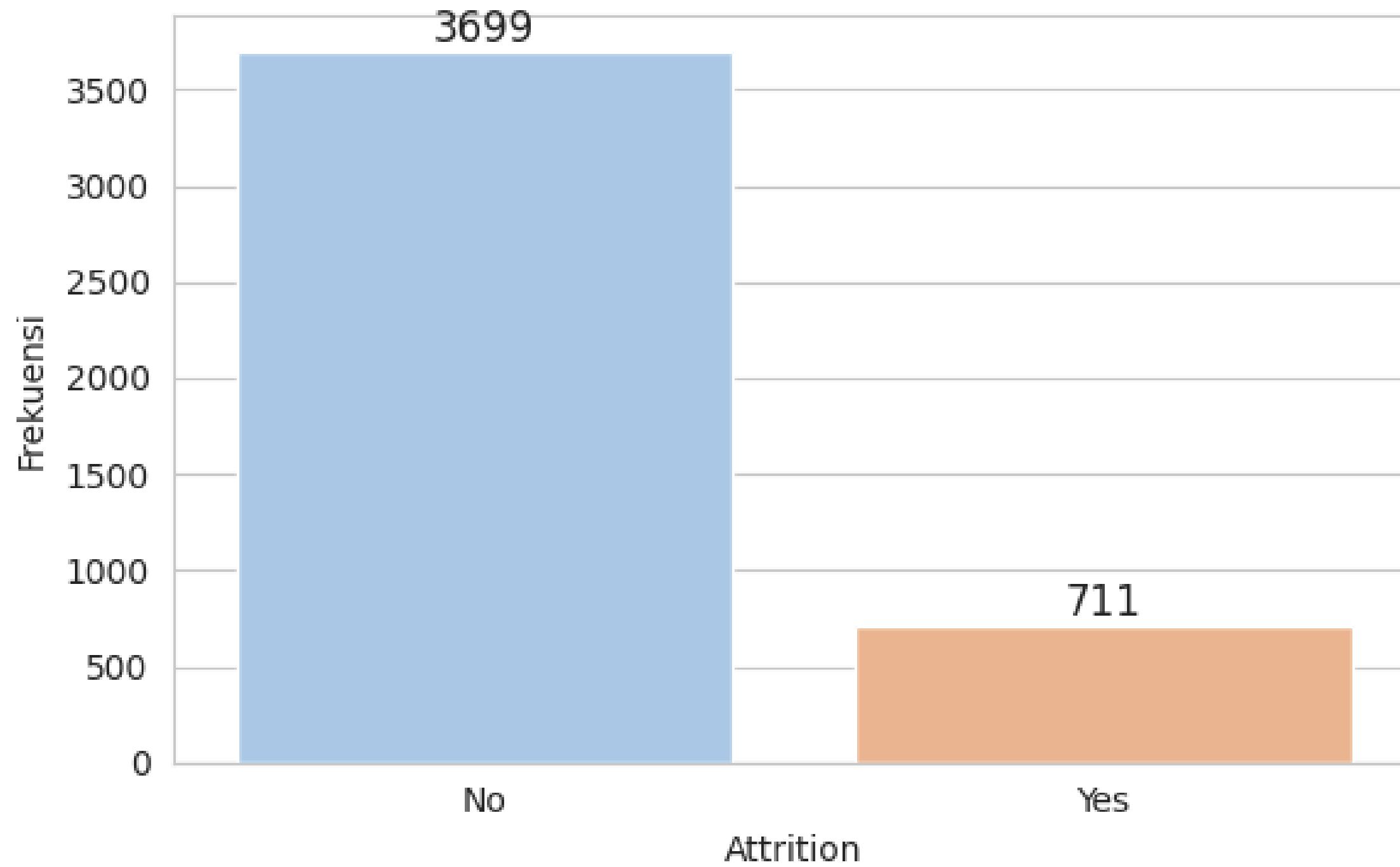


# Data Kategorikal

## Hasil Observasi Data Kategorikal

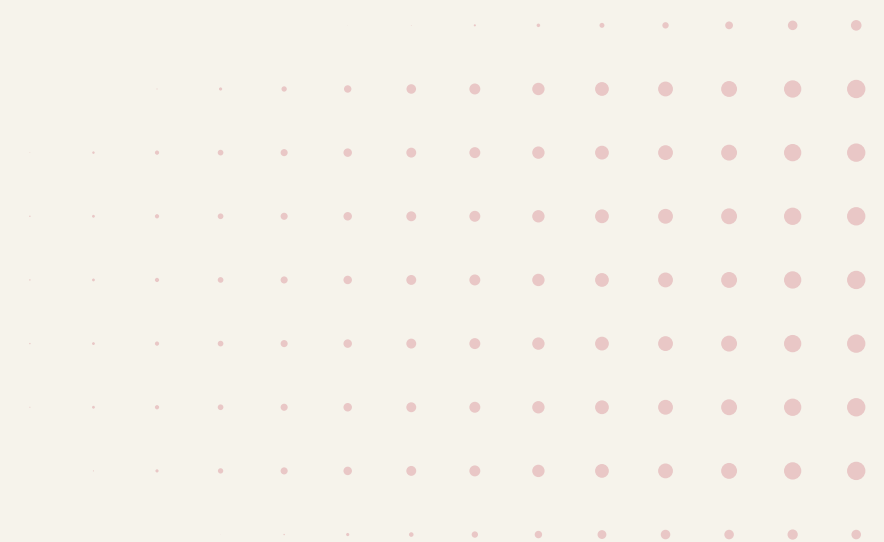
- BusinessTravel: Kategori 'Travel\_Rarely' mendominasi dengan jumlah terbesar, sedangkan 'Non-Travel' memiliki jumlah paling sedikit. Distribusi ini menunjukkan pola mayoritas karyawan tidak sering bepergian.
- Department: Mayoritas karyawan berada di departemen 'Research & Development', sedangkan 'Human Resources' memiliki jumlah terendah.
- EducationField: Mayoritas karyawan berada di bidang 'Life Sciences' dan 'Medical'.
- Gender: Karyawan pria lebih banyak daripada karyawan wanita.
- JobRole: Variasi cukup besar, dengan jumlah karyawan terbesar di peran 'Sales Executive', 'Research Scientist' dan 'Laboratory Technician'.
- MaritalStatus: 'Married' merupakan status pernikahan dominan.
- Over18: Seluruh karyawan berada pada kategori 'Y' yang menunjukkan bahwa variabel ini kurang bermanfaat untuk dianalisis karena tidak memiliki variasi.
- EnvironmentSatisfaction dan JobSatisfaction: Kecenderungan skor yang lebih tinggi (3 dan 4) lebih banyak.
- WorkLifeBalance: Skor '3' memiliki jumlah terbesar, menunjukkan mayoritas merasa cukup puas dengan keseimbangan hidup kerja.
- JobInvolvement: Dominan pada nilai '3' dan skor rendah sangat jarang.
- PerformanceRating: Mayoritas berada pada nilai '3' sedangkan hanya sebagian kecil yang memiliki rating '4'. Distribusi ini menunjukkan skor evaluasi kinerja yang tinggi jarang diberikan.

Distribution of Attrition



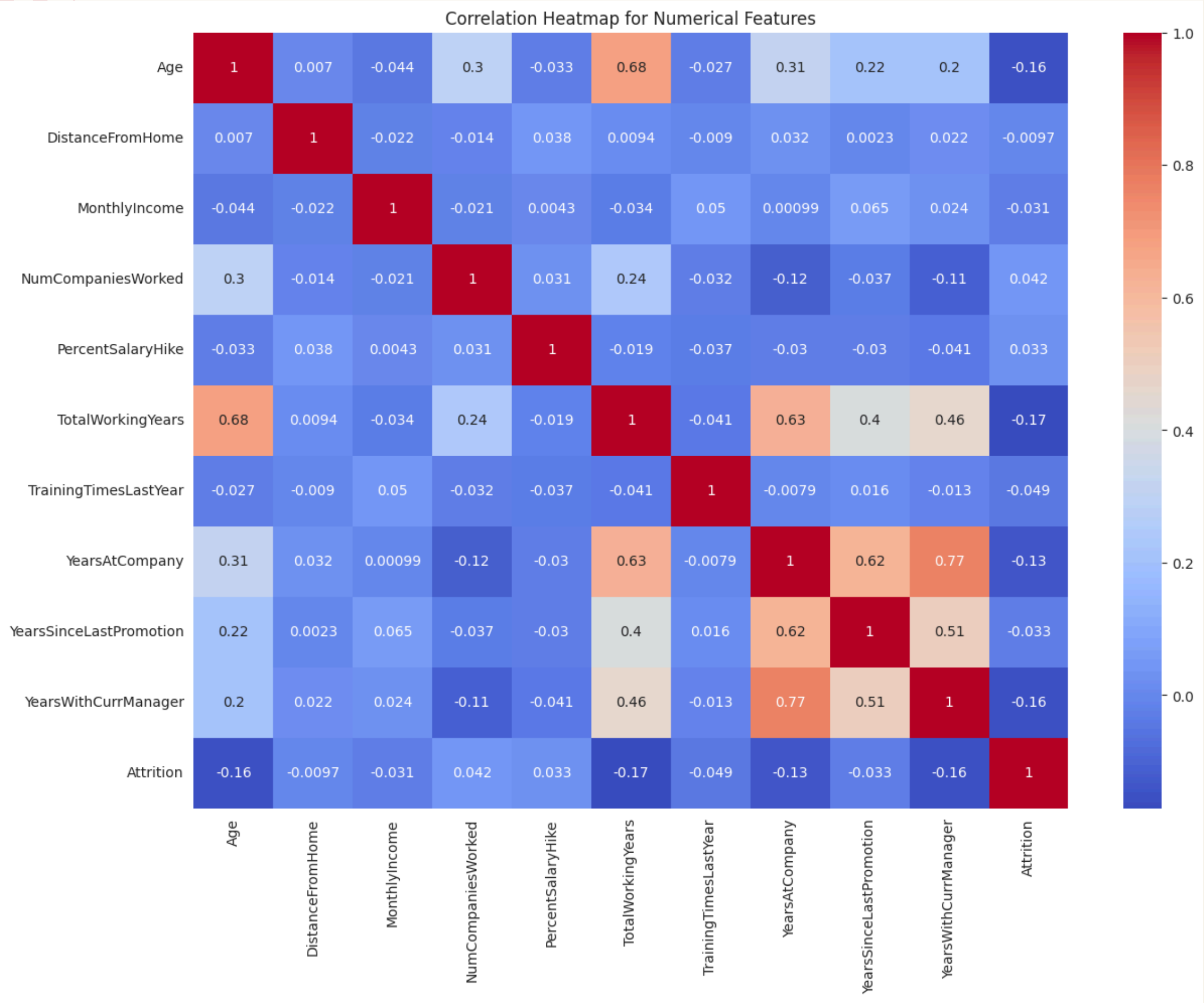


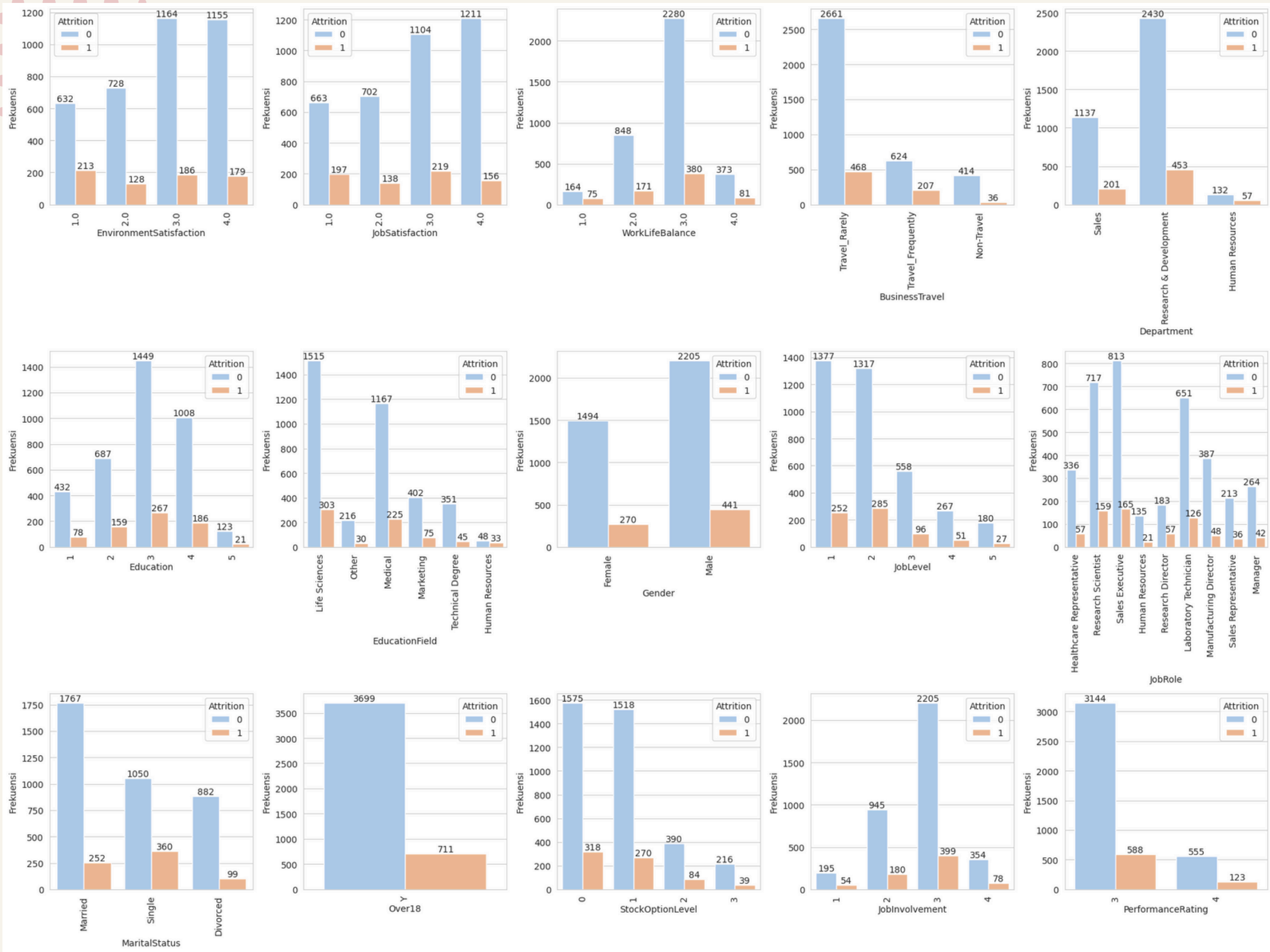
Hal yang harus di follow-up saat data pre-processing:

- Handle missing value: membuang baris-baris dengan data yang hilang jika dirasa data yang dimiliki sudah cukup dan jumlah data yang hilang tidak signifikan atau mengisi kolom yang kosong dengan modus/median
  - Meng-handle outliers
  - Feature Selection: kolom EmployeeCount, StandardHours dan Over18 dapat di-drop karena tidak memiliki variabilitas
  - Melakukan One-Hot encoding terhadap feature kategorik
  - Handle class imbalance dari target
- 



# MULTIVARIATE ANALYSIS





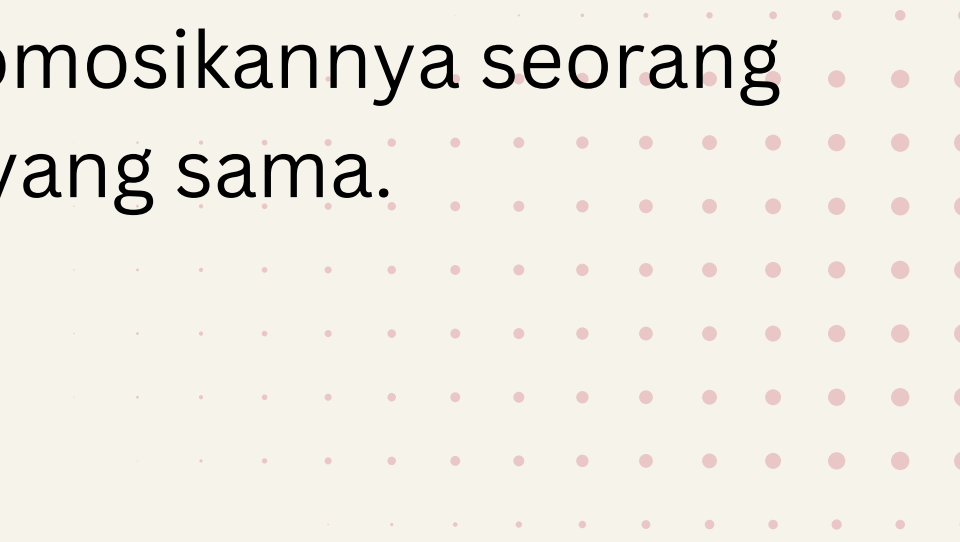
## Hasil Observasi

### A. Korelasi antara Feature dan Target (Attrition):

- Age memiliki korelasi negatif (-0.16) dengan Attrition, yang menunjukkan bahwa semakin tua karyawan, kemungkinan untuk meninggalkan perusahaan cenderung lebih rendah.
- TotalWorkingYears dan YearsWithCurrManager memiliki korelasi negatif dengan Attrition masing-masing sebesar -0.17 dan -0.16, mengindikasikan bahwa karyawan dengan pengalaman lebih lama atau yang telah bekerja lama dengan manajer saat ini cenderung kurang mungkin meninggalkan perusahaan.
- MonthlyIncome memiliki korelasi negatif dengan Attrition sebesar -0.031, sehingga kontribusinya mungkin rendah dalam memprediksi tingkat Attrition.
- NumCompaniesWorked memiliki korelasi positif dengan Attrition sebesar 0.042, hal ini menunjukkan bahwa semakin sering karyawan berganti perusahaan, maka semakin besar kemungkinan karyawan untuk meninggalkan perusahaan.
- EnvironmentSatisfaction rendah akan cenderung melakukan atrisi. Karyawan yang tidak merasa nyaman dengan lingkungan kerjanya berpotensi melakukan atrisi dibandingkan dengan karyawan yang nyaman dengan lingkungan kerjanya.



## B. Korelasi antar Feature

- Age dan TotalWorkingYears memiliki korelasi positif yang tinggi sebesar 0.68. Hal ini wajar karena semakin tua karyawan, biasanya semakin lama pengalaman kerja mereka.
  - YearsAtCompany berkorelasi tinggi dengan TotalWorkingYears (0.63), YearsWithCurrManager (0.77) dan YearsSinceLastPromotion (0.62). Karyawan yang bekerja lebih lama di perusahaan cenderung memiliki lebih banyak pengalaman kerja, bekerja lebih lama dengan manajer saat ini, tetapi cenderung lama tidak dilakukan promosi. Hal ini menunjukkan potensi redundansi antar fitur tersebut.
  - YearsSinceLastPromotion berkorelasi positif dengan TotalWorkingYears (0.4), YearsAtCompany (0.62) dan YearsWithCurrManager (0.51). Yang menunjukkan bahwa dipromosikannya seorang karyawan bukan karena lamanya bekerja pada perusahaan dan manager yang sama.
- 



# BUSINESS INSIGHT AND RECOMENDATION

## **Insight 1: Pengaruh Umur dan Lama Pengalaman Terhadap Attrition**

Berdasarkan korelasi antara Age dan TotalWorkingYears dengan Attrition, terlihat bahwa karyawan yang lebih tua dan memiliki pengalaman kerja yang lebih panjang cenderung lebih loyal dan jarang meninggalkan perusahaan. Korelasi negatif yang signifikan antara ketiga variabel ini menunjukkan bahwa karyawan berpengalaman kurang mungkin untuk keluar.

### **Rekomendasi:**

Fokus pada program retensi untuk karyawan muda atau yang memiliki pengalaman kerja lebih singkat.

Perusahaan dapat mempertimbangkan program mentorship yang menghubungkan karyawan muda dengan yang lebih senior untuk meningkatkan loyalitas dan mengurangi kemungkinan attrition.

Memberikan peluang pengembangan karier dan pelatihan bagi karyawan muda dapat membantu mereka merasa lebih terikat dengan perusahaan.

## **Insight 2: Kepuasan Lingkungan Kerja dan Keseimbangan hidup-kerja Berhubungan dengan Attrition Rendah**

Dari distribusi variabel EnvironmentSatisfaction dan WorkLifeBalance, terlihat bahwa karyawan dengan kepuasan lingkungan kerja dan keseimbangan hidup-kerja yang tinggi cenderung memiliki angka attrition yang lebih rendah.

### **Rekomendasi:**

Investasi dalam program peningkatan lingkungan kerja, seperti fleksibilitas dalam jam kerja, area relaksasi, atau fasilitas yang mendukung kesehatan mental.

Mengembangkan kebijakan keseimbangan hidup-kerja yang lebih baik, misalnya opsi bekerja dari rumah atau jam kerja yang fleksibel, dapat membantu meningkatkan loyalitas karyawan.



### **Insight 3: Tingkat Kenaikan Gaji Berhubungan dengan Attrition**

Variabel PercentSalaryHike menunjukkan sedikit korelasi positif dengan tingkat attrition, yang mungkin menunjukkan bahwa karyawan yang menerima kenaikan gaji yang rendah lebih mungkin untuk meninggalkan perusahaan.

#### **Rekomendasi:**

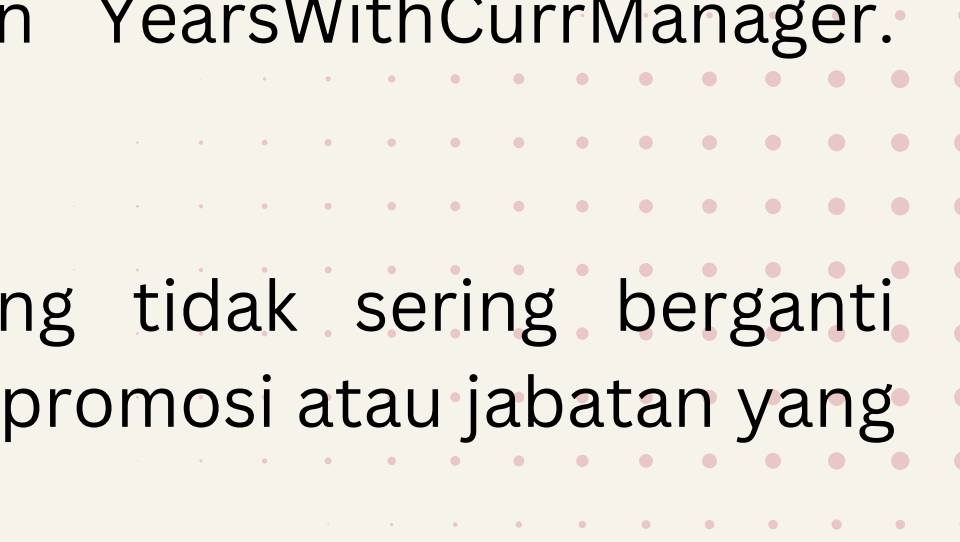
- Perusahaan bisa mempertimbangkan untuk mengkaji kembali kebijakan kenaikan gaji, khususnya bagi karyawan yang memiliki performa tinggi atau masa kerja lama.
- Menyediakan insentif tambahan di luar kenaikan gaji standar, seperti bonus kinerja, penghargaan, atau kesempatan promosi, dapat membantu mempertahankan karyawan yang berpotensi tinggi.

### **Insight 4: Hubungan Loyalitas Karyawan dan Promosi**

Variabel YearsSinceLastPromotion menjelaskan promosi terakhir yang didapat oleh karyawan. Variabel ini berkorelasi positif dengan variabel TotalWorkingYears, YearsAtCompany, dan YearsWithCurrManager. Loyalitas karyawan tidak terlalu diapresiasi oleh perusahaan.

#### **Rekomendasi:**

Mengapresiasi karyawan yang telah menunjukkan loyalitas yaitu mereka yang tidak sering berganti departemen atau perusahaan. Apresiasi tersebut bisa diwujudkan dengan kenaikan promosi atau jabatan yang lebih baik agar ke depannya karyawan merasa lebih diperhatikan oleh perusahaan.



# REPOSITORI GIT SYNTHETICA

Final-Project-SyntheticaPublic

Pin

Unwatch1

Fork0

Star0

develophad recent pushes 20 seconds ago

Compare & pull request

develop4 BranchesTags

Go to file

Add file

Code

This branch is 3 commits ahead of, 2 commits behindmaster.

timothygintingCreate README.md

Final\_Project\_EDA.ipynbMenambahkan File

README.mdCreate README.md

test-git-develop-branch.txtcommit test branch

test-git.txtfirst commit from timothy

README

Insight 1: Pengaruh Umur dan Lama Pengalaman Terhadap Attrition Berdasarkan korelasi antara Age dan TotalWorkingYears dengan Attrition, terlihat bahwa karyawan yang lebih tua dan memiliki pengalaman kerja yang lebih panjang cenderung lebih loyal dan jarang meninggalkan perusahaan. Korelasi negatif yang signifikan antara ketiga variabel ini menunjukkan bahwa karyawan berpengalaman kurang mungkin untuk keluar. Rekomendasi: Fokus pada program retensi untuk karyawan muda atau yang memiliki pengalaman kerja lebih singkat. Perusahaan dapat

Local

Codespaces

Clone

HTTPSSSHGitHub CLI

https://github.com/timothyginting/Final-Project1

Clone using the web URL.

Open with GitHub Desktop

Download ZIP

About

No description, website, or topics provided.

Readme

Activity

0 stars

1 watching

0 forks

Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package

feature/test4 BranchesTags

Go to file

Add file

Code

This branch is 2 commits ahead ofmaster.

Contribute

miaanurfadilahMenambahkan File EDA Final Project

95525ca · yesterday3 Commits

Final\_Project\_EDA.ipynbMenambahkan File EDA Final Project

yesterday

test-git-develop-branch.txtcommit test branch from timothy

yesterday

test-git.txtfirst commit from timothy

yesterday

**SEKIAN DAN TERIMA KASIH**