

# Kelompok Data Lab

1. Oktavian Dwi Putra
2. Muhammad Ilham
3. Maulid Uly Mada
4. Seto Bhanu Adyatma
5. Mega Pangastuti
6. Rasyidah Maulida P. A
7. M. Hasbi Ashshiddiqi
8. Raihan Damar



# **1. Descriptive Statistics**

# 1. Descriptive Statistics

Dilihat menggunakan `df.info` sudah sesuai, karena nama dan type kolom sudah sesuai dengan data yang sudah ada di dalam kolom.

```
Kolom dengan tipe data kurang sesuai atau nama kolom dan isinya kurang sesuai:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                -
0   Age                                  1470 non-null   int64
1   Attrition                           1470 non-null   object
2   BusinessTravel                       1470 non-null   object
3   DailyRate                            1470 non-null   int64
4   Department                           1470 non-null   object
5   DistanceFromHome                     1470 non-null   int64
6   Education                             1470 non-null   int64
7   EducationField                       1470 non-null   object
8   EmployeeCount                         1470 non-null   int64
9   EmployeeNumber                       1470 non-null   int64
10  EnvironmentSatisfaction               1470 non-null   int64
11  Gender                               1470 non-null   object
12  HourlyRate                           1470 non-null   int64
13  JobInvolvement                       1470 non-null   int64
14  JobLevel                             1470 non-null   int64
15  JobRole                              1470 non-null   object
16  JobSatisfaction                       1470 non-null   int64
17  MaritalStatus                        1470 non-null   object
18  MonthlyIncome                        1470 non-null   int64
19  MonthlyRate                           1470 non-null   int64
20  NumCompaniesWorked                   1470 non-null   int64
21  Over18                               1470 non-null   object
22  OverTime                             1470 non-null   object
23  PercentSalaryHike                    1470 non-null   int64
24  PerformanceRating                    1470 non-null   int64
25  RelationshipSatisfaction               1470 non-null   int64
26  StandardHours                         1470 non-null   int64
27  StockOptionLevel                     1470 non-null   int64
28  TotalWorkingYears                    1470 non-null   int64
29  TrainingTimesLastYear                 1470 non-null   int64
30  WorkLifeBalance                       1470 non-null   int64
31  YearsAtCompany                       1470 non-null   int64
32  YearsInCurrentRole                    1470 non-null   int64
33  YearsSinceLastPromotion                1470 non-null   int64
34  YearsWithCurrManager                  1470 non-null   int64
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
```

# 1. Descriptive Statistics

Dilihat menggunakan `df.isnull`, pada kolom kanan menunjukkan jumlah *null* dalam kolom yang ditunjuk ada di sebelah kiri, menunjukkan bahwa tidak ada kolom yang memiliki nilai kosong.

Kolom yang memiliki nilai kosong:

Age	0
Attrition	0
BusinessTravel	0
DailyRate	0
Department	0
DistanceFromHome	0
Education	0
EducationField	0
EmployeeCount	0
EmployeeNumber	0
EnvironmentSatisfaction	0
Gender	0
HourlyRate	0
JobInvolvement	0
JobLevel	0
JobRole	0
JobSatisfaction	0
MaritalStatus	0
MonthlyIncome	0
MonthlyRate	0
NumCompaniesWorked	0
Over18	0
OverTime	0
PercentSalaryHike	0
PerformanceRating	0
RelationshipSatisfaction	0
StandardHours	0
StockOptionLevel	0
TotalWorkingYears	0
TrainingTimesLastYear	0
WorkLifeBalance	0
YearsAtCompany	0
YearsInCurrentRole	0
YearsSinceLastPromotion	0
YearsWithCurrManager	0
dtype:	int64



# 1. Descriptive Statistics

Numerical Describe

	Age	DistanceFromHome	Education	EnvironmentSatisfaction	JobSatisfaction	MonthlyIncome	NumCompaniesWorked	WorkLifeBalance	YearsAtCompany
<b>count</b>	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
<b>mean</b>	36.923810	9.192517	2.912925	2.721769	2.728571	6502.931293	2.693197	2.761224	7.008163
<b>std</b>	9.135373	8.106864	1.024165	1.093082	1.102846	4707.956783	2.498009	0.706476	6.126525
<b>min</b>	18.000000	1.000000	1.000000	1.000000	1.000000	1009.000000	0.000000	1.000000	0.000000
<b>25%</b>	30.000000	2.000000	2.000000	2.000000	2.000000	2911.000000	1.000000	2.000000	3.000000
<b>50%</b>	36.000000	7.000000	3.000000	3.000000	3.000000	4919.000000	2.000000	3.000000	5.000000
<b>75%</b>	43.000000	14.000000	4.000000	4.000000	4.000000	8379.000000	4.000000	3.000000	9.000000
<b>max</b>	60.000000	29.000000	5.000000	4.000000	4.000000	19999.000000	9.000000	4.000000	40.000000

# 1. Descriptive Statistics

Categorical Describe

	Attrition	BusinessTravel	Department	EducationField	Gender	JobRole	MaritalStatus	Over18	OverTime
<b>count</b>	1470	1470	1470	1470	1470	1470	1470	1470	1470
<b>unique</b>	2	3	3	6	2	9	3	1	2
<b>top</b>	No	Travel_Rarely	Research & Development	Life Sciences	Male	Sales Executive	Married	Y	No
<b>freq</b>	1233	1043	961	606	882	326	673	1470	1054

# 1. Descriptive Statistics

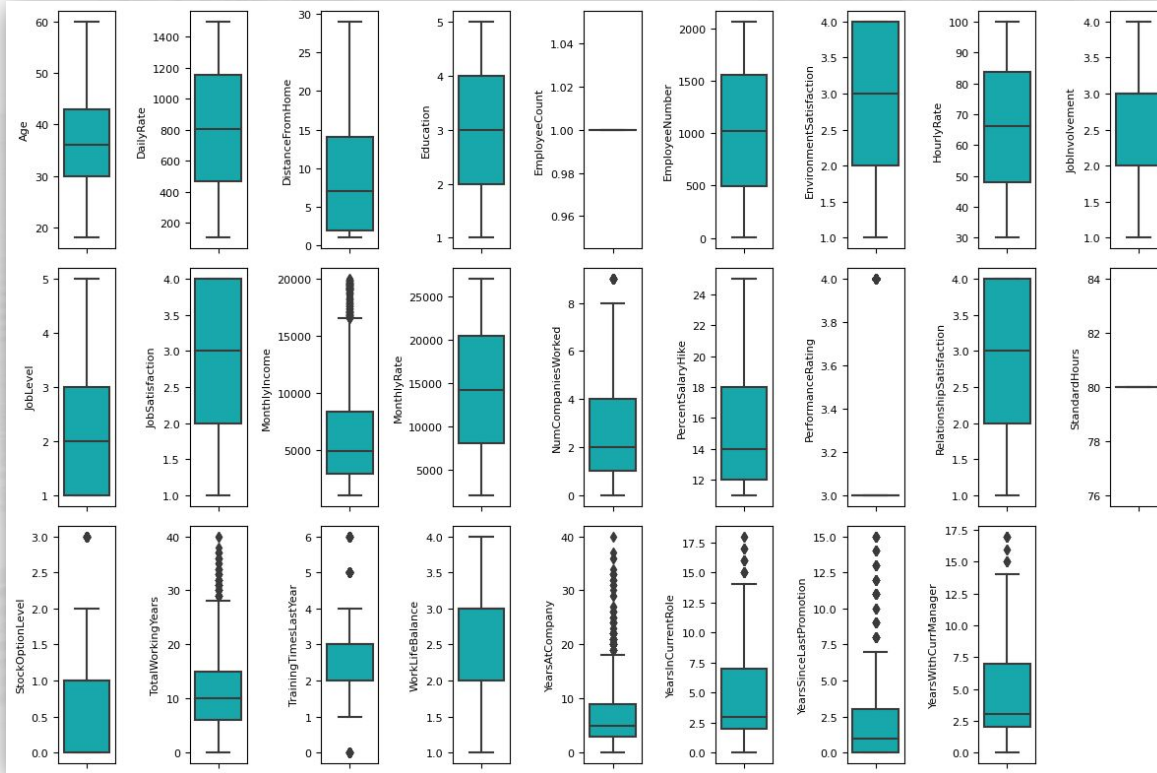
Berdasarkan statistika deskriptif yang sudah dibuat,

- Kolom numerik seperti DistanceFromHome, MonthlyIncome, NumCompaniesWorked, PercentSalaryHike, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, dan YearsWithCurrManager sepertinya memiliki distribusi **skew ke kanan**, jika dilihat dari nilai mean yang jauh lebih besar dari nilai mediannya.
- Pada kolom DailyRate, EmployeeNumber, MonthlyIncome, MonthlyRate, TotalWorkingYears, YearsAtCompany terdapat indikasi adanya **outlier** yang ditunjukkan dari **nilai max**-nya yang berbeda jauh jika dibandingkan dengan nilai lainnya.
- Kolom EmployeeCount, StandardHours, Over18 perlu di-*drop*, karena hanya memiliki **1 nilai unik**. Sedangkan kolom EmployeeNumber, juga perlu di-*drop*, karena memiliki nilai unik yang jumlahnya **sama dengan jumlah baris** yang ada.

## **2. Univariate Analysis**



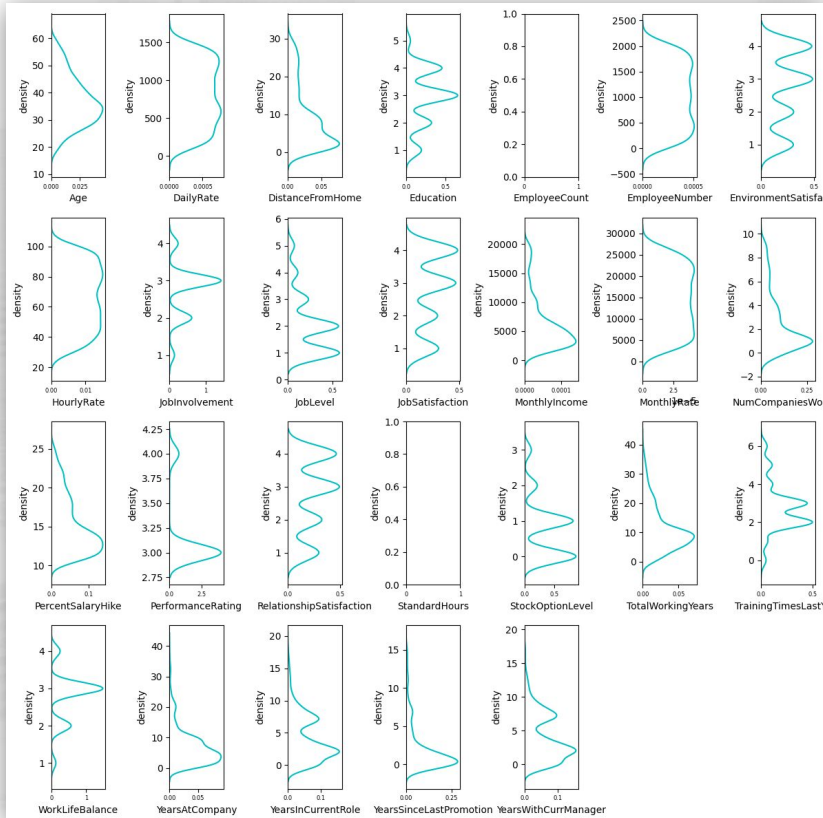
## 2. Univariate Analysis



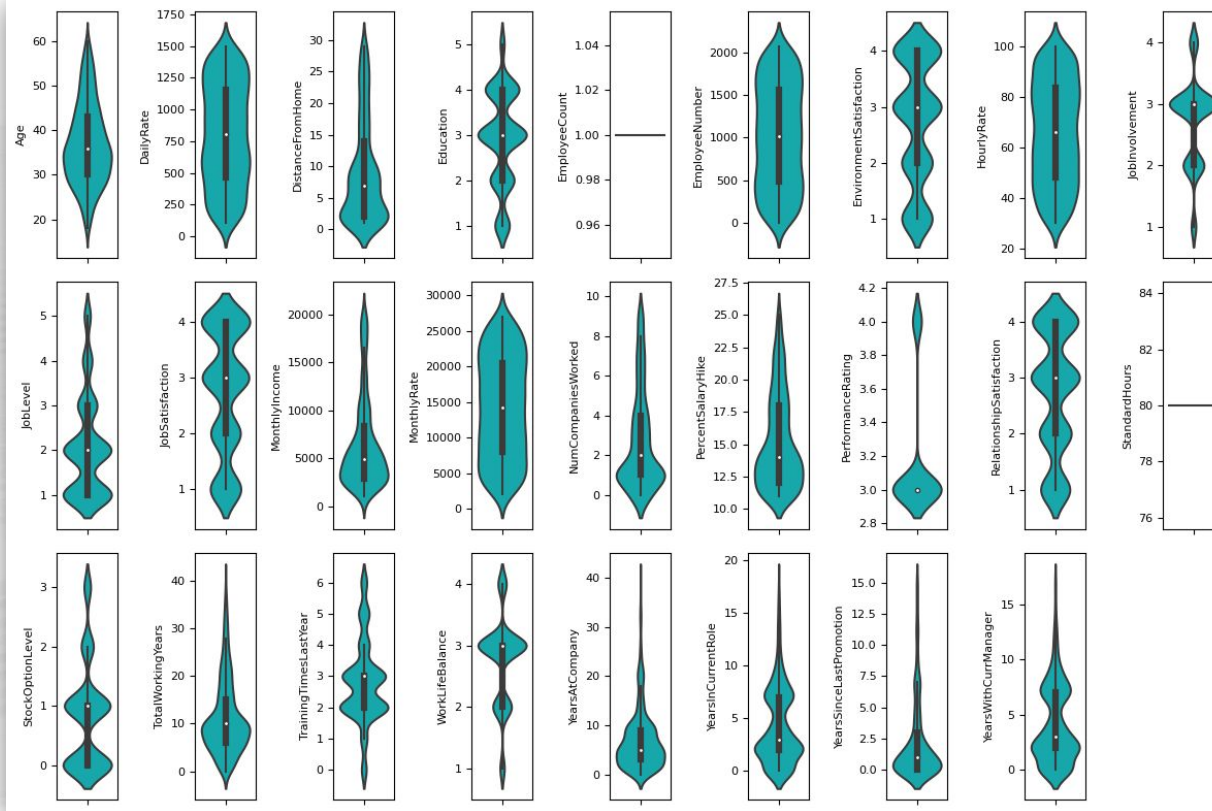
Dari grafik distribusi di samping terlihat bahwa ada beberapa **outlier** yang mana dapat dilihat dari kesenjangan dari skor yang ada hingga skor yang extreme

## 2. Univariate Analysis

Pada grafik distribusi di samping terdapat banyak kolom yang berjenis **positive skewed** dan **multimodal**, serta beberapa kolom **bimodal** dan berdistribusi **normal**.

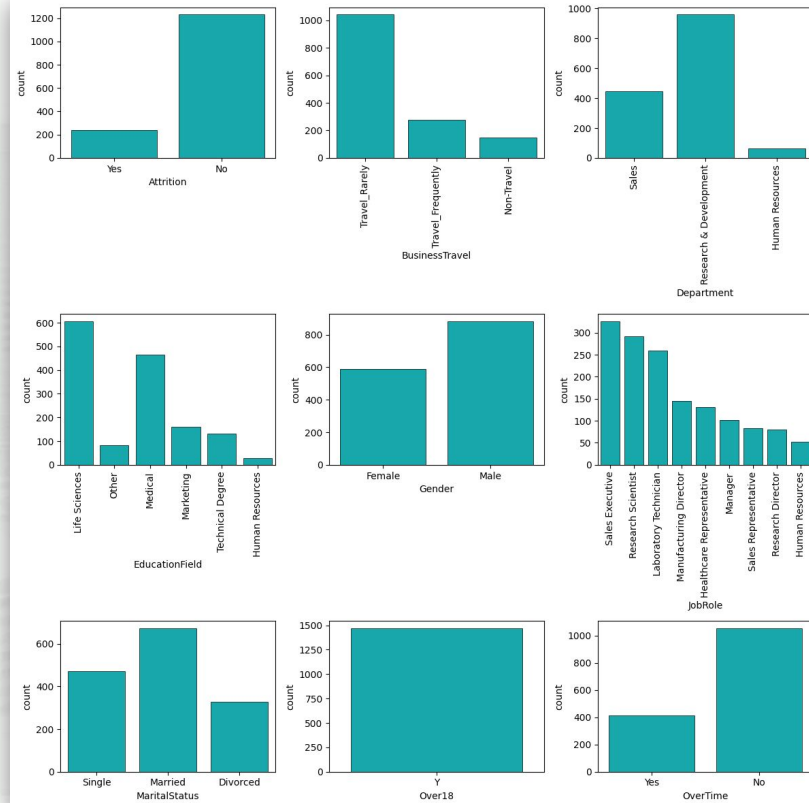


## 2. Univariate Analysis



Pada grafik distribusi di samping, terlihat bahwa distribusi nilai yang ada terdapat **outlier** yang mana garis tipis berwarna hitam yang menjulang panjang ketimbang dengan garis tebal sebagai rentang yang ada.

## 2. Univariate Analysis



### Insight

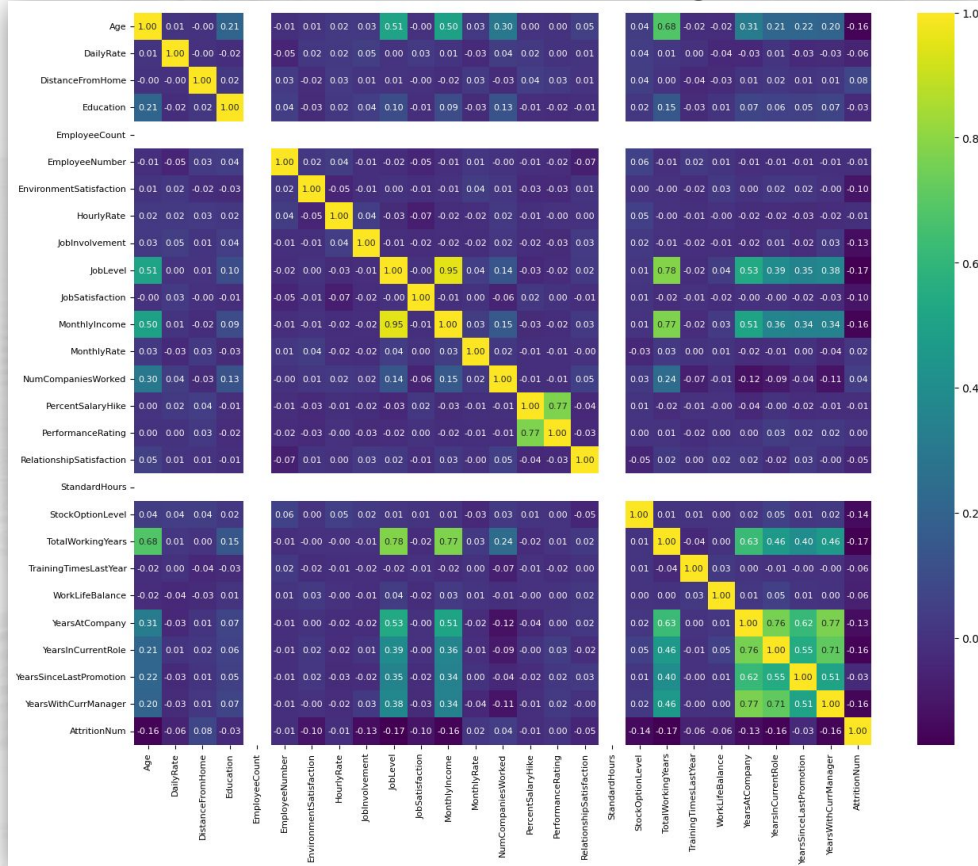
Beberapa insight yang dapat kita tarik dari *univariate analysis* di atas diantaranya:

- Beberapa kolom numerik yang memiliki **outlier** adalah kolom `MonthlyIncome`, `NumCompaniesWorked`, `PerformanceRating`, `StockOptionLevel`, `TotalWorkingYears`, `TrainingTimesLastYear` dan `YearsAtCompany`, `YearsInCurrentRole`, `YearsSinceLastPromotion`, dan `YearsWithCurrentManager`
- Terdapat kolom numerik yang memiliki distribusi **skew ke kanan** seperti kolom `DistanceFromHome`, `MonthlyIncome`, `NumCompaniesWorked`, `PercentSalaryHike`, `TotalWorkingYears`, `YearsAtCompany` dan `YearsSinceLastPromotion` sehingga bisa dilakukan **log transformation** pada data *pre-processing*.
- Kolom numerik yang memiliki distribusi **bi-modal** adalah `PerformanceRating`, `YearsInCurrentRole` dan `YearsWithCurrManager`. Sedangkan kolom yang memiliki distribusi **multi-modal** adalah `Education`, `EnvironmentSatisfaction`, `JobInvolvement`, `JobLevel`, `JobSatisfaction`, `RelationshipSatisfaction`, `StockOptionLevel`, `TrainingTimesLastYear` dan `WorkLifeBalance`.
- Kolom numerik yang berdistribusi normal adalah `Age`, `DailyRate`, `EmployeeNumber`, `HourlyRate`, dan `MonthlyRate`
- Terjadi ketimpangan data **mild** pada distribusi kolom kategorik `BusinessTravel`, `Department`, `Gender` dan `OverTime`. Sedangkan ketimpangan data **moderate** pada distribusi kolom kategorik `Attrition`.

# **3. Multivariate Analysis**



# 3. Multivariate Analysis

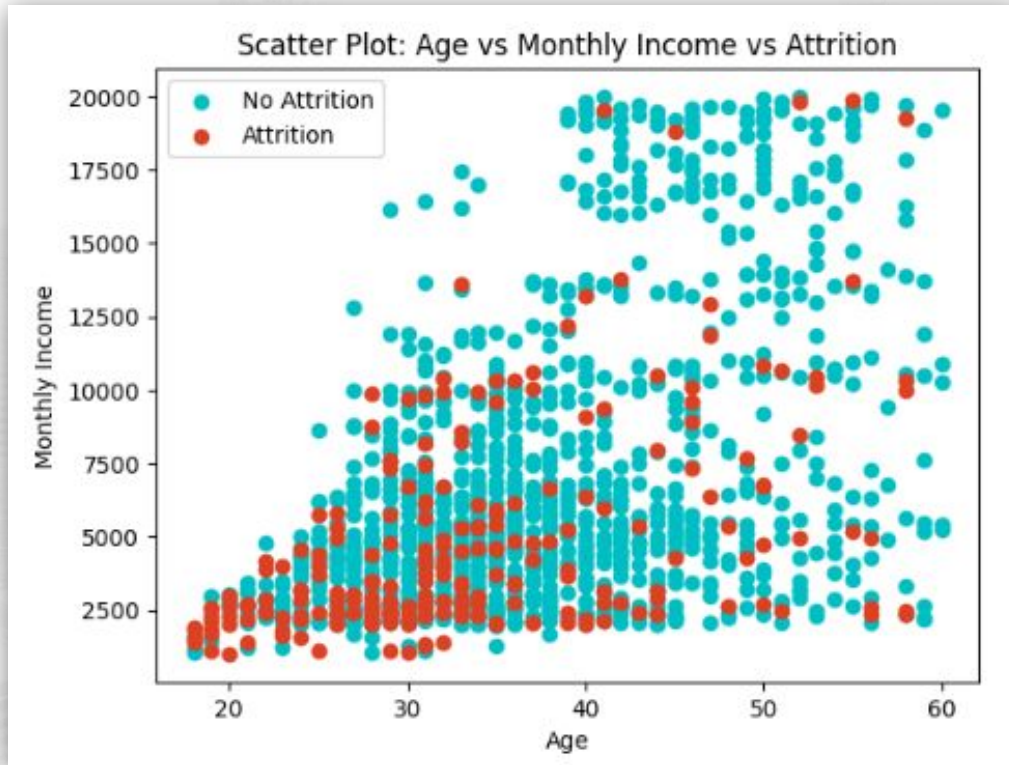


Beberapa insight yang dapat kita tarik dari *multivariate analysis* di atas diantaranya:

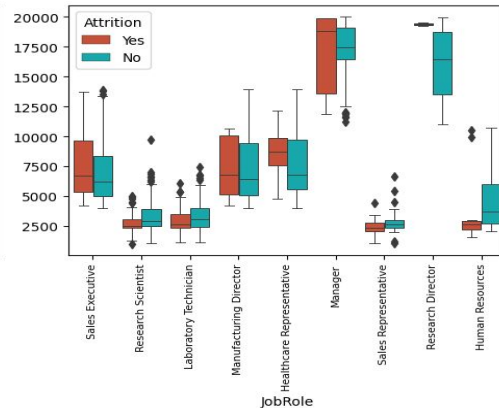
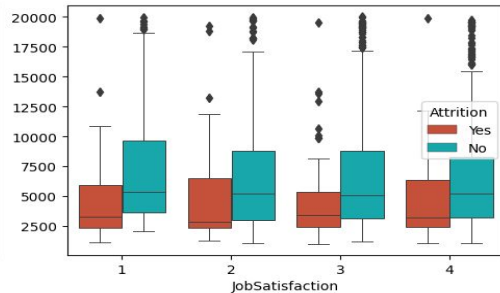
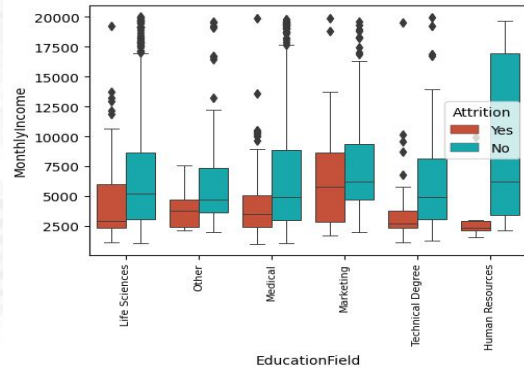
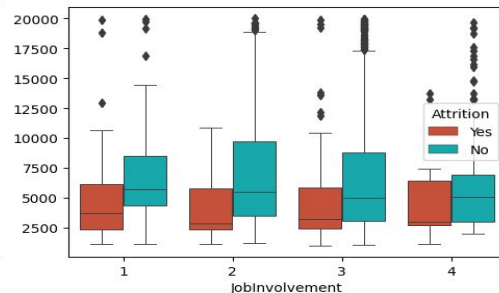
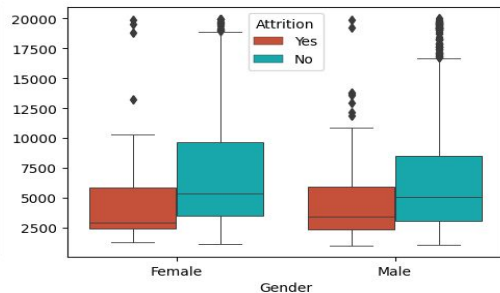
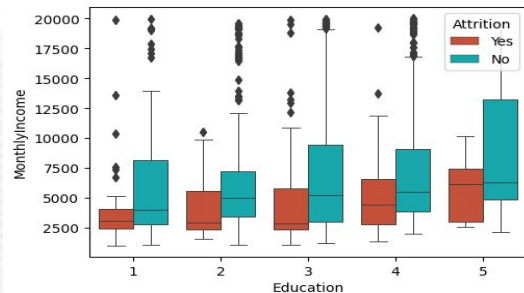
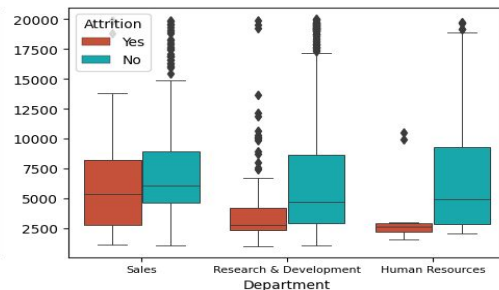
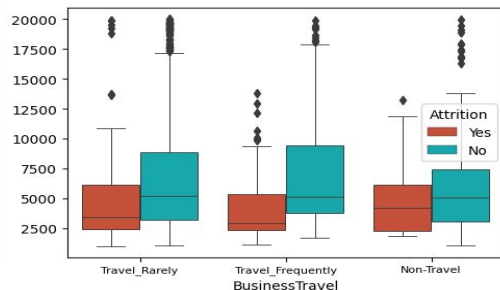
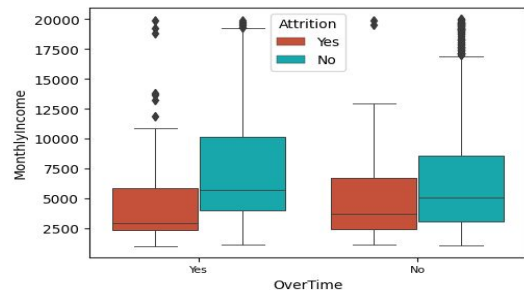
- Terdapat **multicollinearity** atau korelasi yang memiliki nilai diatas 0.7 antara kolom **JobLevel**, **MonthlyIncome** dan **TotalWorkingYears** serta kolom **YearsAtCompany**, **YearsInCurrentRole** dan **YearsWithCurrManager** sehingga kita bisa memilih salah satu saja.
- Tidak ada kolom numerik yang berkorelasi kuat dengan target (**Attrition**)

A. Dari heatmap di atas, maka dapat dicari kriteria fitur yang tidak dipakai/dibuang karena kurang relevan seperti `EmployeeCount` dan `StandardHours` serta redundan ( $> 0.7$ ), yaitu : `JobLevel & TotalWorkingYears`, `JobLevel & MonthlyIncome`, `MonthlyIncome & TotalWorkingHours`, `PercentSalaryHike & PerformanceRating`, `YearsAtCompany & YearsWithCurrManager`, `YearsAtCompany & YearsInCurrentRole`, `YearsInCurrentRole & YearsWithCurrManager`. Jadi, dapat disimpulkan bahwa terjadi multikolinearitas sehingga salah satu fitur yang mempunyai korelasi  $> 0.7$  tersebut harus dihapus dan digunakan hanya salah satu fitur saja.

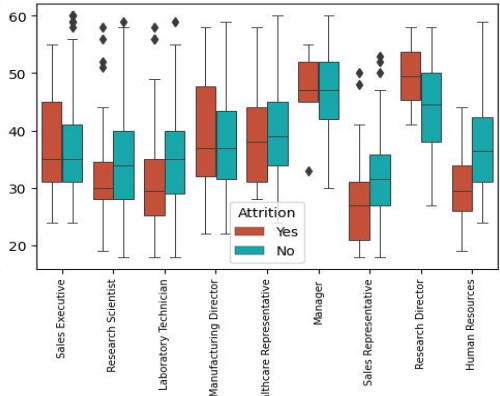
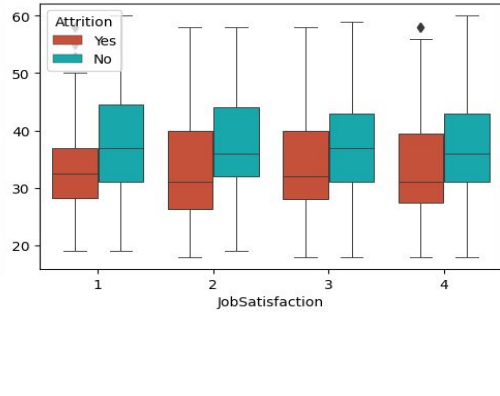
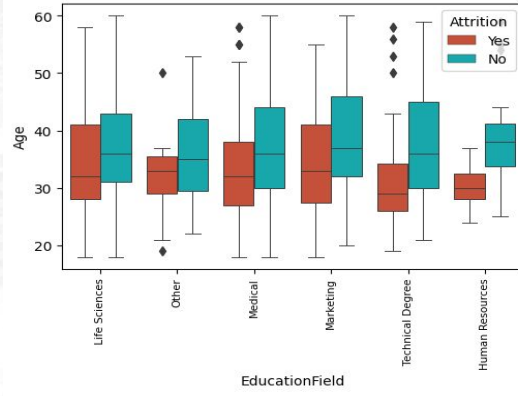
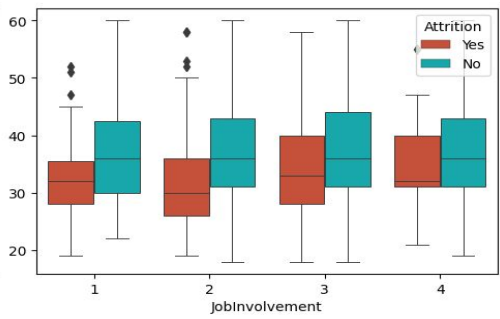
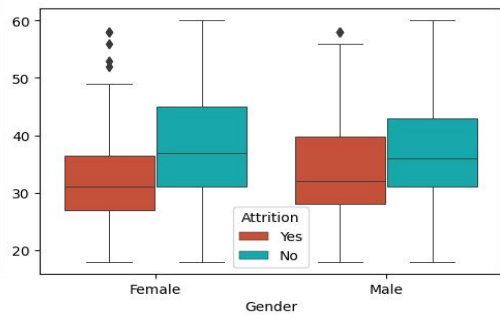
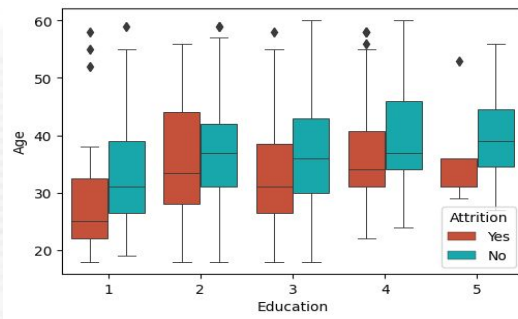
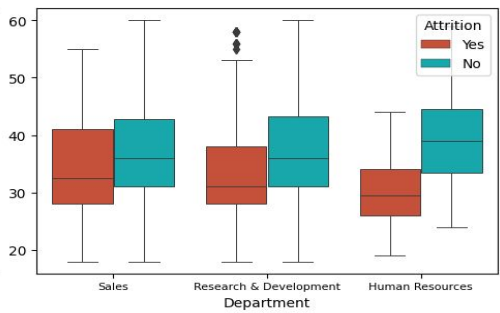
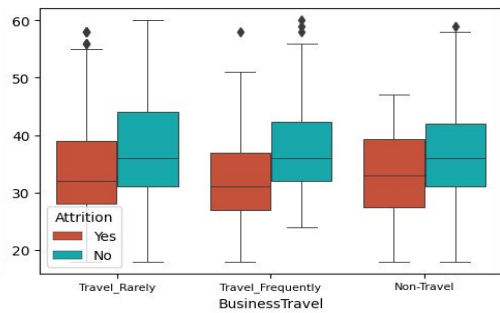
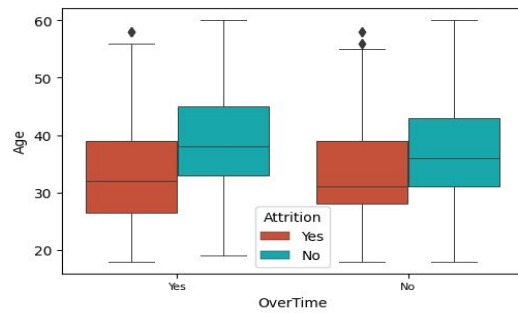
B. Tidak ada fitur yang mempunyai korelasi yang kuat terhadap target karena nilai korelasi paling tinggi adalah hanya 0.17, fitur dengan nilai korelasi terhadap target 0.17 diantaranya `TotalWorkingYears` dan `JobLevel`. Sedangkan, hal menarik dapat dilihat dari korelasi antara `TotalWorkingYears` dan `Age` yang memiliki korelasi positif yang wajar, karena semakin lama pengalaman kerja maka semakin tinggi usia karyawan. Korelasi antara `YearsAtCompany` dan `TotalWorkingYears` memiliki korelasi positif, yang berarti semakin lama seseorang bekerja di perusahaan, semakin lama pula tahun mereka bekerja. Hal yang menarik lain adalah korelasi antara `YearsSinceLastPromotion` dan `YearsAtCompany` memiliki korelasi positif, menunjukkan bahwa semakin lama seseorang berada di perusahaan, maka semakin lama pula sejak mereka dipromosikan terakhir kali.



Grafik **scatterplot** di samping adalah hubungan antara usia (``Age``), pendapatan bulanan (``Monthly Income``), dan ``Attrition``. Dari grafik tersebut terlihat bahwa karyawan dengan rentang **usia 25 - 35** dan pendapatan bulanan **dibawah 5000** lebih banyak untuk meninggalkan perusahaan.









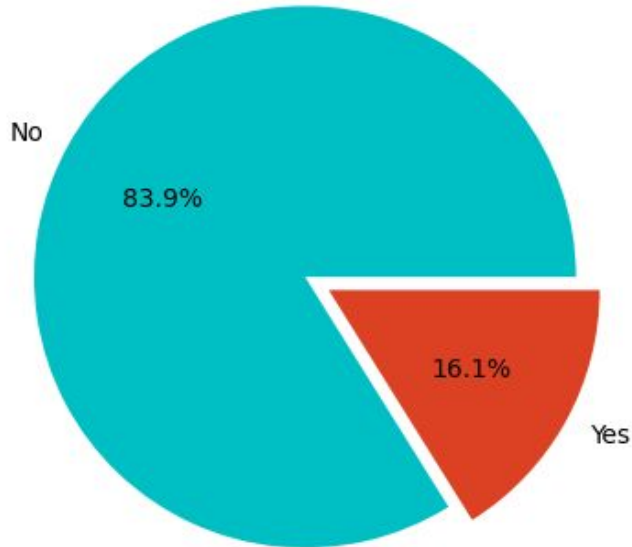
Dari grafik di atas yang menunjukkan hubungan `MonthlyIncome` dengan fitur lain (`OverTime, BusinessTravel, Department, Education, Gender, JobInvolvement, EducationField, JobSatisfaction, dan JobRole`) dapat disimpulkan bahwa karyawan dengan pendapatan bulanan (`Monthly Income`) **di bawah 5000** memiliki *Attrition Rate* yang lebih tinggi atau kecenderungan untuk meninggalkan perusahaan lebih banyak.

Dari grafik di atas yang menunjukkan hubungan `Age` dengan fitur-fitur lain (`OverTime, BusinessTravel, Department, Education, Gender, JobInvolvement, EducationField, JobSatisfaction, dan JobRole`) dapat disimpulkan bahwa karyawan dengan range **usia antara 25 - 35** juga memiliki *Attrition Rate* yang lebih tinggi atau kecenderungan untuk meninggalkan perusahaan lebih banyak.

# **4. Business Insight**

## 4. Insight

Ratio of Employee Attritions



Berdasarkan ilustrasi di atas, dapat disimpulkan bahwa **16.1%** dari total karyawan memilih untuk **keluar** dari perusahaan sedangkan **83.9%** dari total karyawan memilih untuk **bertahan**.

## 4. Insight

Dengan menggunakan framework **5W + 1H**, permasalahan *employee attrition* tersebut akan di-breakdown menjadi:

**What : Apa yang membuat karyawan meninggalkan perusahaan?**

feature : OverTime, JobInvolvement, BusinessTravel

**Who : Siapa saja yang karyawan yang meninggalkan perusahaan?**

feature : Age, Gender, JobRole, JobLevel, MaritalStatus, Education, EducationField

**Why : Alasan apa yang membuat karyawan yang meninggalkan perusahaan?**

feature : Monthly Income, EnvironmentSatisfaction, JobSatisfaction

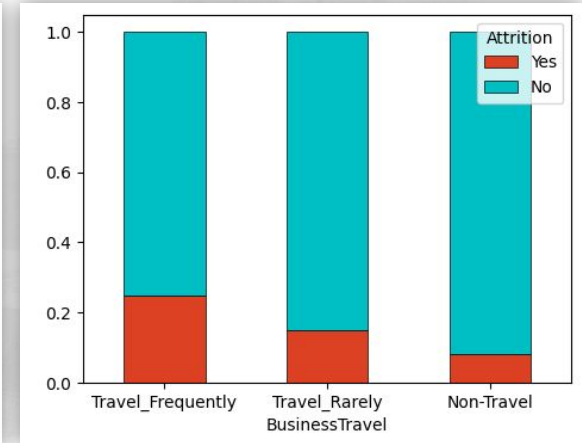
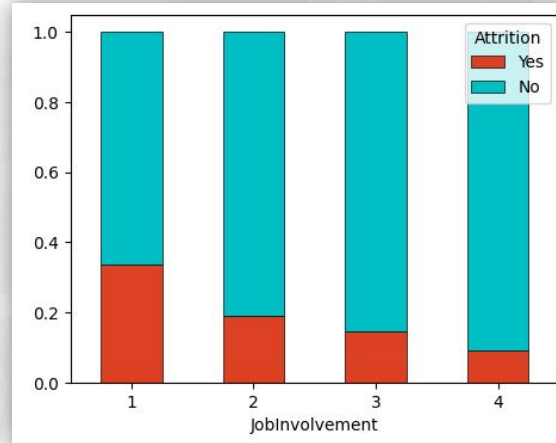
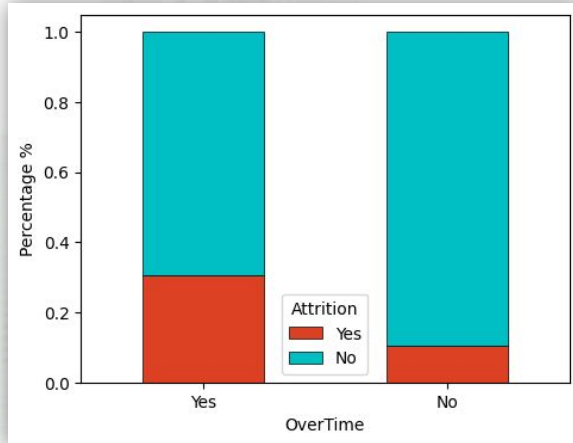
**Where : Dari mana sajakah karyawan yang meninggalkan perusahaan?**

feature : Department

**When : Kapan biasanya karyawan meninggalkan perusahaan?**

feature : TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager

## 4. Insight (What)

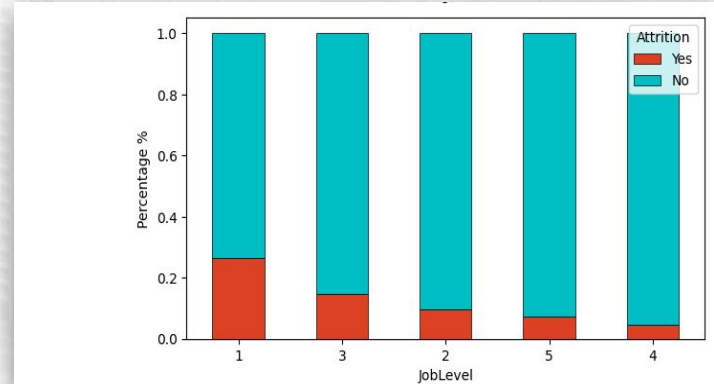
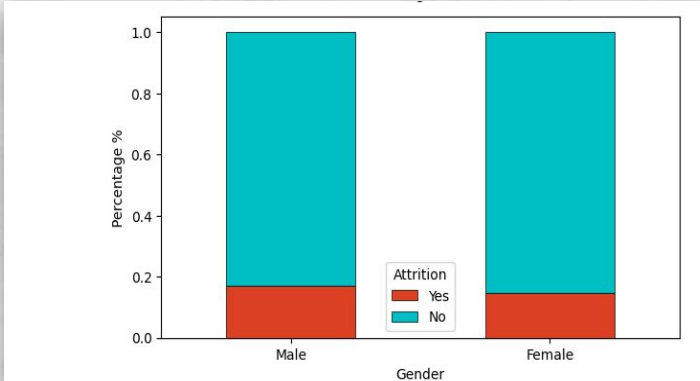
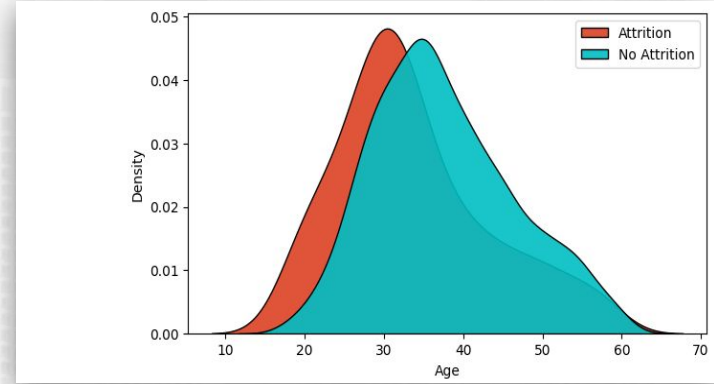
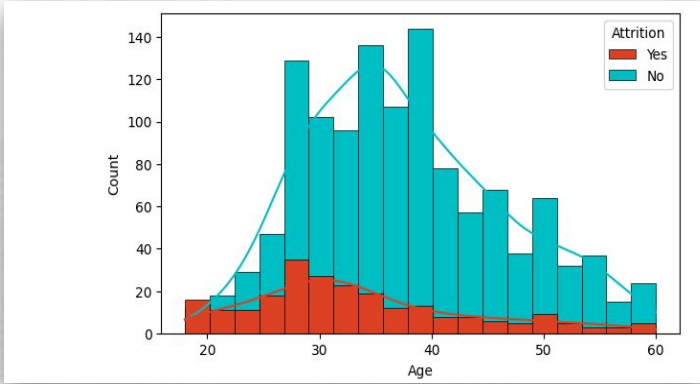


Setelah dilakukan proses analisa, didapatkan beberapa kesimpulan berupa :

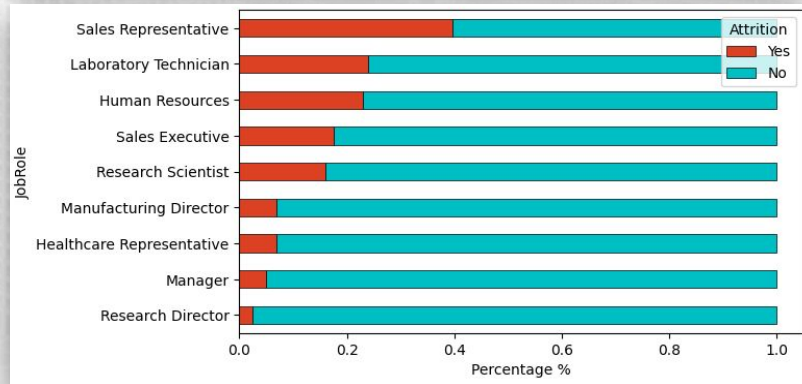
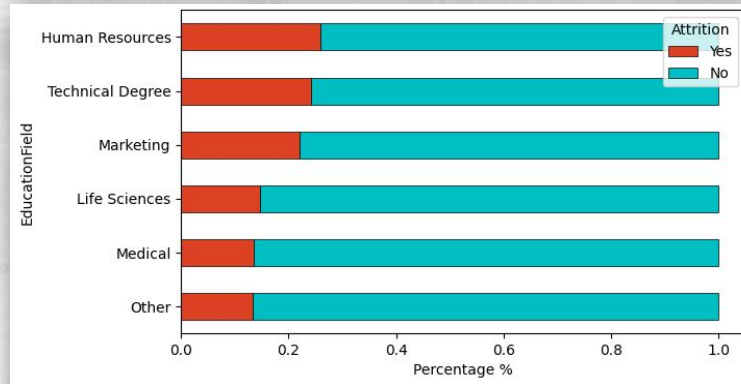
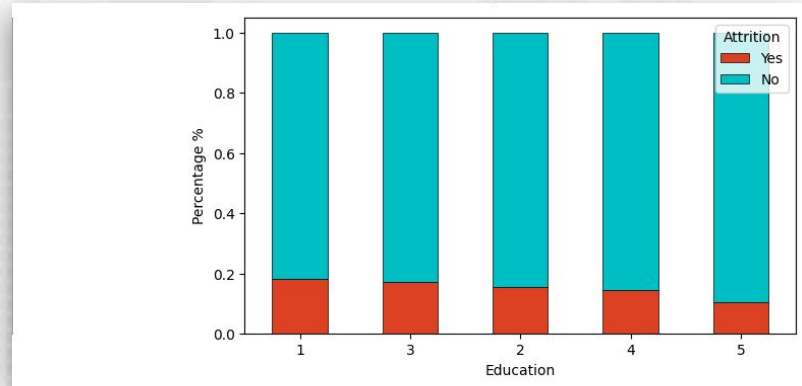
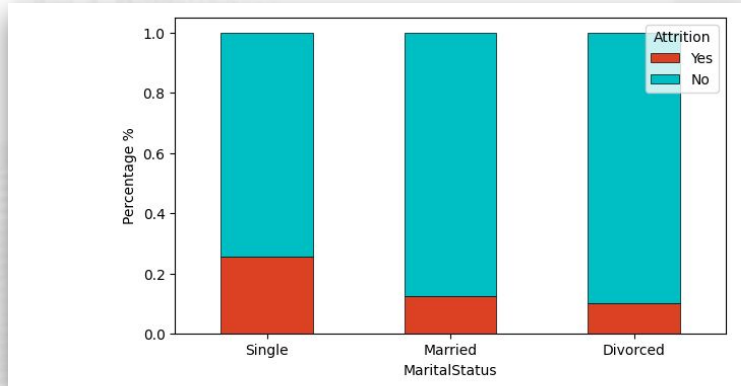
1. Karyawan yang bekerja melebihi jam kerja reguler (*over time*) memiliki *Attrition Rate* yang lebih tinggi dibandingkan dengan karyawan yang tidak bekerja lembur.
2. Karyawan yang memiliki keterlibatan kerja yang lebih rendah memiliki *Attrition Rate* yang lebih tinggi.
3. Karyawan yang lebih sering melakukan *business travel* memiliki *Attrition Rate* yang lebih tinggi.



## 4. Insight (Who)



## 4. Insight (Who)

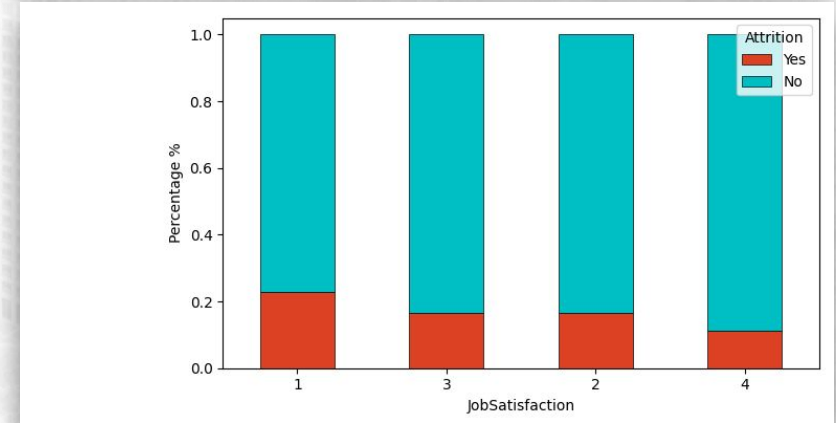
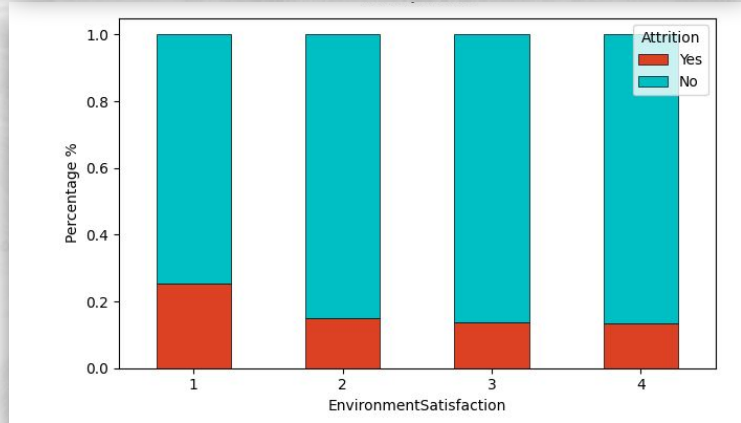
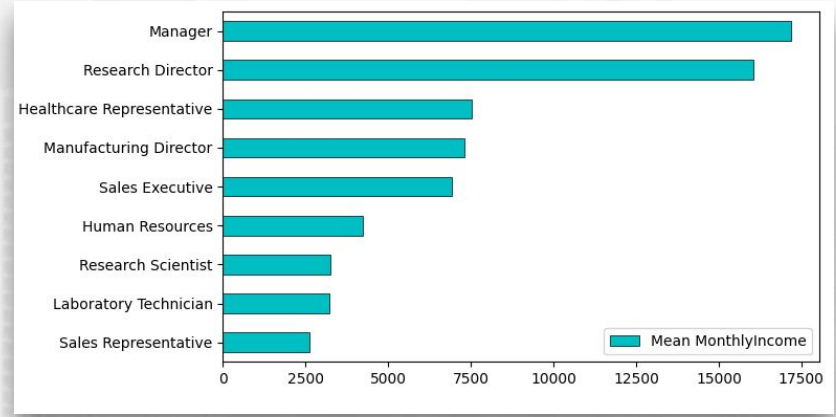
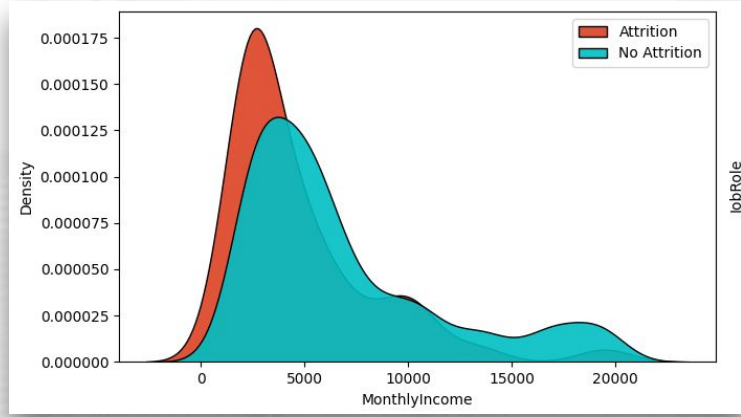


## 4. Insight (Who)

Berdasarkan grafik di atas dapat disimpulkan bahwa :

1. Karyawan dengan rentang umur di bawah 40 tahun cenderung memiliki *Attrition Rate* yang lebih tinggi. Kemudian, karyawan dengan rentang umur sekitar 25 - 35 tahun memiliki *Attrition Rate* yang paling tinggi.
2. Karyawan pria cenderung memiliki *Attrition Rate* yang sama dengan karyawan wanita.
3. Karyawan yang memiliki Job Level yang lebih rendah memiliki *Attrition Rate* yang lebih tinggi. Kemudian, karyawan yang memiliki Job Level 1 merupakan tipe karyawan dengan *Attrition Rate* paling tinggi.
4. Karyawan dengan marital status belum menikah (*single*) cenderung memiliki *Attrition Rate* yang lebih tinggi dibandingkan karyawan yang sudah menikah dan yang berstatus sudah bercerai. Selain itu, karyawan dengan marital status *divorced* memiliki *Attrition Rate* paling rendah atau merupakan tipe karyawan yang paling banyak bertahan di perusahaan.
5. Karyawan yang memiliki *educational level* yang lebih rendah cenderung memiliki *Attrition Rate* yang lebih tinggi. Kemudian, karyawan dengan yang memiliki *educational level* 1 merupakan tipe karyawan yang memiliki *Attrition Rate* paling tinggi.
6. Karyawan yang memiliki latar belakang pendidikan di bidang *Human Resources*, *Technical Degree*, dan *Marketing* merupakan 3 tipe karyawan yang memiliki *Attrition Rate* paling tinggi.
7. Karyawan yang menjabat sebagai *Sales Representative*, *Laboratory Technician*, dan *Human Resources* merupakan 3 tipe karyawan yang memiliki *Attrition Rate* paling tinggi.

## 4. Insight (Why)



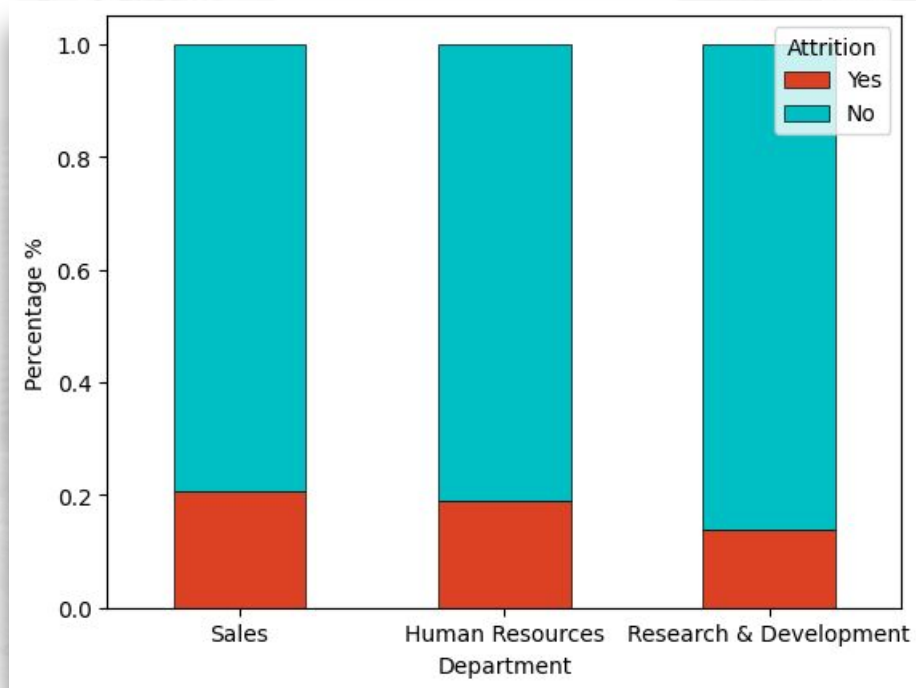
## 4. Insight (Why)

Berdasarkan grafik di atas dapat disimpulkan bahwa :

1. Karyawan dengan *Monthly Income* sekitar 1800 - 3200 merupakan tipe karyawan yang memiliki *Attrition Rate* yang paling tinggi. Kemudian karyawan dengan *Monthly Income* sekitar 13000 - 19000 merupakan tipe karyawan yang memiliki *Attrition Rate* yang paling rendah.
2. Karyawan yang memiliki jabatan sebagai *Manager* memiliki *Monthly Income* paling tinggi sedangkan karyawan dengan jabatan sebagai *Sales Representative* memiliki *Monthly Income* paling rendah.
3. Karyawan yang memiliki tingkat *Environment Satisfaction* yang lebih rendah cenderung memiliki *Attrition Rate* yang lebih tinggi. Kemudian, didapatkan bahwa karyawan yang memiliki *Environment Satisfaction* bernilai 1 merupakan tipe karyawan yang memiliki *Attrition Rate* yang paling tinggi.
4. Karyawan yang memiliki tingkat *Job Satisfaction* yang lebih rendah cenderung memiliki *Attrition Rate* yang lebih tinggi. Kemudian, didapatkan bahwa karyawan yang memiliki *Job Satisfaction* bernilai 1 merupakan tipe karyawan yang memiliki *Attrition Rate* yang paling tinggi.

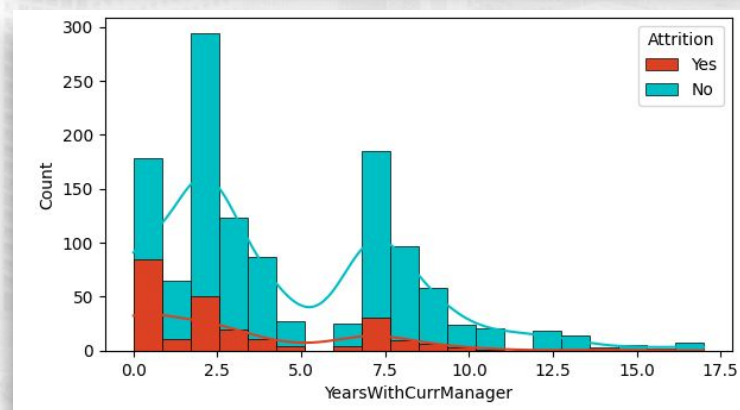
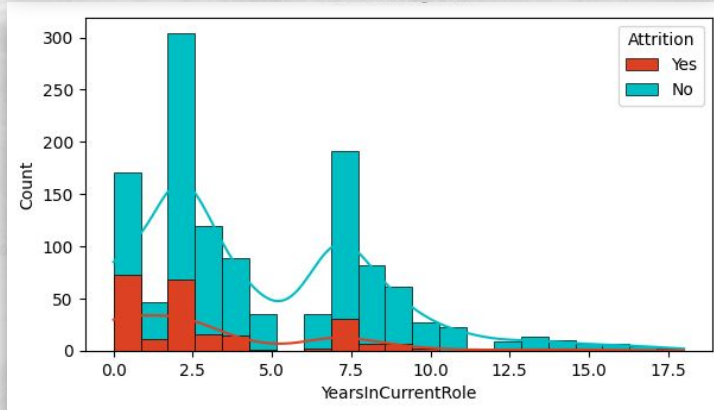
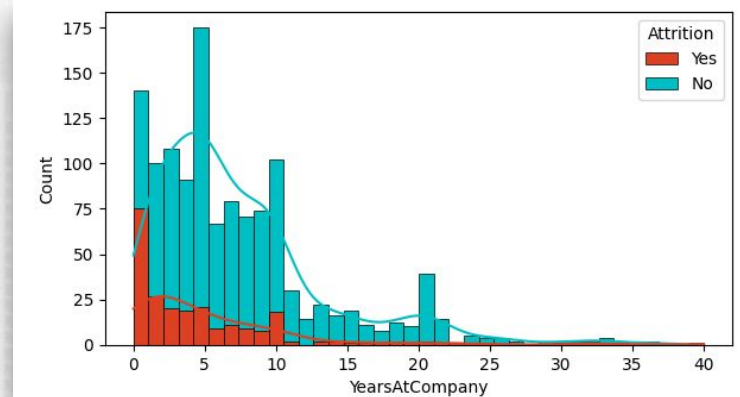
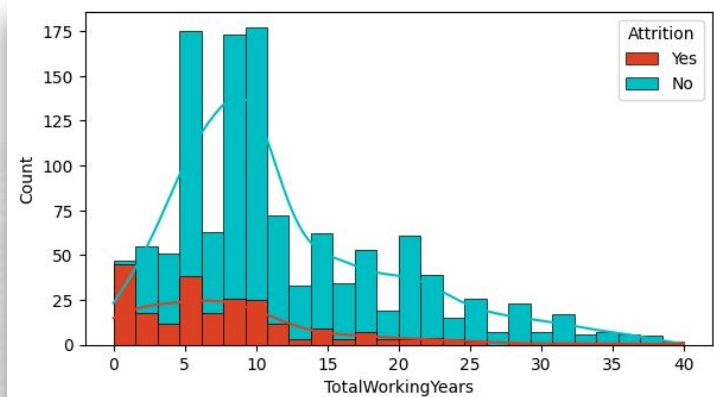


## 4. Insight (Where)



Berdasarkan grafik di samping dapat disimpulkan bahwa karyawan yang bekerja pada Departemen Sales memiliki *Attrition Rate* yang paling tinggi dibandingkan dengan departemen lainnya.

## 4. Insight (When)



## 4. Insight (When)

Berdasarkan grafik di atas dapat disimpulkan bahwa :

1. Karyawan dengan *Total Working Years* sekitar 0 - 2 tahun serta 4,5 - 6 tahun merupakan dua tipe karyawan yang memiliki *Attrition Rate* yang paling tinggi.
2. Karyawan dengan nilai *Years at Company* sekitar 0 - 1 tahun tipe karyawan yang memiliki *Attrition Rate* yang paling tinggi.
3. Karyawan dengan nilai *Years in Current Role* sekitar 0 - 1 tahun serta 1,75 - 2,6 tahun merupakan dua tipe karyawan yang memiliki *Attrition Rate* yang paling tinggi.
4. Karyawan dengan *Years with Current Manager* dibawah 1 tahun memiliki nilai *Attrition Rate* yang lebih tinggi dibanding yang lainnya.

## 4. Business Recommendation

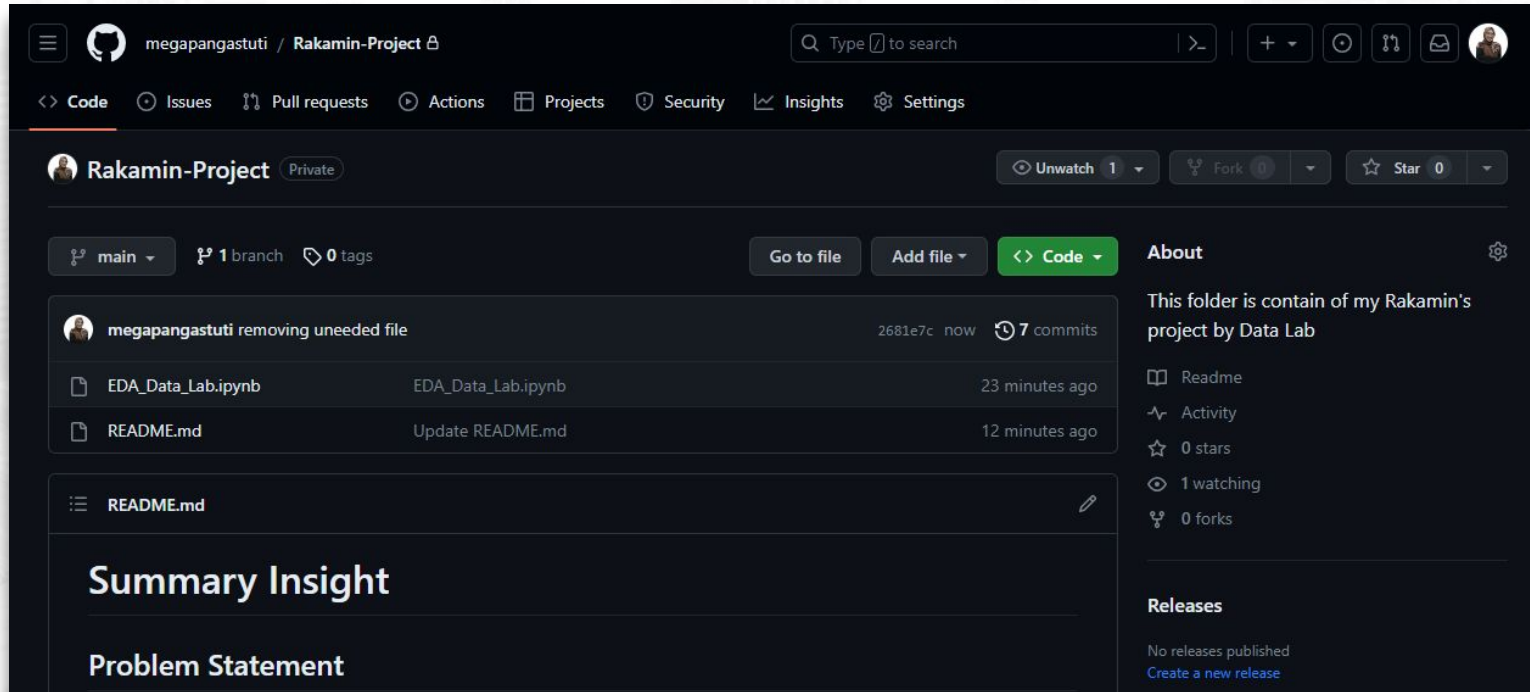
1. Kami merekomendasikan untuk melakukan analisa beban kerja masing-masing karyawan untuk meminimalisir jumlah karyawan yang bekerja diluar jam kerja reguler (*over time*) dengan memperhitungkan jumlah karyawan yang berada pada masing-masing departemen.
2. Kami merekomendasikan untuk mendistribusikan kesempatan *business travel* secara merata.
3. Kami merekomendasikan untuk memberikan perlakuan khusus seperti perusahaan memberikan fasilitas bagi karyawan untuk melakukan *self development* serta pemberian jenjang karir yang jelas kepada karyawan yang berstatus *Job Level 1*.
4. Kami merekomendasikan untuk melakukan peninjauan terhadap karyawan dengan jabatan sebagai *Sales Executive*, *Research Scientist*, dan *Laboratory Technician* untuk mendapatkan faktor-faktor penyebab karyawan pada jabatan tersebut tidak memilih untuk bertahan pada perusahaan. Kemudian, kami menyarankan untuk memberikan *treatment* khusus berdasarkan hasil peninjauan tersebut .
5. Kami merekomendasikan untuk melakukan peninjauan kembali terhadap fasilitas yang sudah atau akan diberikan kepada masing-masing karyawan seperti gaji, insentif, dan sebagainya.
6. Kami merekomendasikan untuk melakukan penerapan *positive culture* sebagai budaya perusahaan serta membuat sebuah kegiatan untuk meningkatkan *engagement* antar karyawan untuk meningkatkan kepuasan para karyawan.
7. Kami merekomendasikan untuk melakukan peninjauan terhadap Departemen *Sales* untuk mendapatkan faktor-faktor penyebab karyawan pada departemen tersebut tidak memilih untuk bertahan pada perusahaan. Kemudian, kami menyarankan untuk memberikan *treatment* khusus berdasarkan hasil peninjauan tersebut .

# 5. Git



# 5. Git

Berikut merupakan [Git Repository](#) untuk menyimpan file Home Work Stage 1 :



The screenshot shows a GitHub repository interface for 'megapangastuti / Rakamin-Project'. The repository is private and has 1 branch (main) and 0 tags. It contains two files: 'EDA\_Data\_Lab.ipynb' (updated 23 minutes ago) and 'README.md' (updated 12 minutes ago). The repository description states: 'This folder is contain of my Rakamin's project by Data Lab'. The 'About' section lists 0 stars, 1 watching, and 0 forks. The 'Releases' section indicates no releases are published. The 'Summary Insight' section shows the 'Problem Statement'.