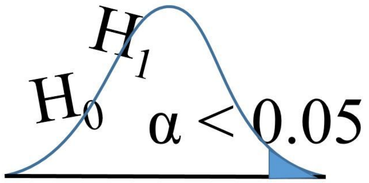
Research Question:
What factors will affect an employee's decision to leave the company?

Name: Tan Wei Rong (21224533), Tan Siao Shuen (20229564)

Hypothesis:

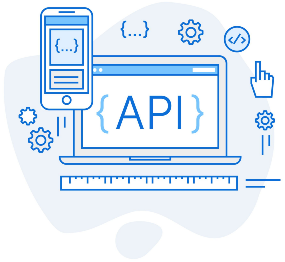
H1: **Age** is a factor in influencing employee's decision to leave the company.
H2: **Department** which employees belong to is a factor in influencing the employee's decision to leave the company.
H3: **Gender** is a factor in influencing employee's decision to leave the company.
H4: **Monthly Income** of employee's is a factor in influencing employee's decision to leave the company.

Variables:



Dependent variable: Attrition

Independent variables: Age, Department, Gender, Monthly Income



Importing libraries and API

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
%matplotlib inline
```

```
import kaggle
from kaggle.api.kaggle_api_extended import KaggleApi
api = KaggleApi()
api.authenticate()
```

```
x = api.dataset_list(search="attrition rate of company")
print(*x)
```

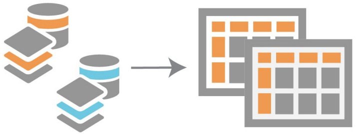
```
anujachintyabiswas/attrition-rate-of-a-company thedevastator/employee-attrition-and-factors iamsouravbanerjee/software-professi
onal-salaries-2022 colearninglounge/employee-attrition muhammadshahidazeem/customer-churn-dataset saadharoon27/hr-analytics-dat
aset sanjanchaudhari/employees-performance-for-hr-analytics dhawan123/attrition-rate-data-set prachi13/employeeattritionrate ri
schan/workshop-shopee-machine-learning-ann-scikitlearn kadirduran/hr-dataset bhavikapuri2811/hr-analytics-dashboard-using-excel
rohan0301/hackerearth-ml-will-your-employees-leave-you adityaghuse/employee-attrition-and-engagement synful/churn-classificatio
n mohammadkaiftahir/hr-analytics danielggak/summit-biotech-attrition-2021 sunilhit120/ibm-hr-analytics-and-visualization chaudh
urimdtausif/company-employee-attrition wrucebaynebot/hr-dataset
```

```
api.dataset_list_files('anujachintyabiswas/attrition-rate-of-a-company').files
```

```
[general_data.csv]
```

```
api.dataset_download_file('anujachintyabiswas/attrition-rate-of-a-company','general_data.csv')
```

```
True
```



Loading dataset

```
df=pd.read_csv('general_data.csv')
df
```

	EmployeeID	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Gender	...	TotalWorkingYears
0	1	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	Female	...	1.0
1	2	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	Female	...	6.0
2	3	32	No	Travel_Frequently	Research & Development	17	4	Other	1	Male	...	5.0
3	4	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	Male	...	13.0
4	5	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	Male	...	9.0
...	...	...	...	...	...	...	...	...	...	...	...	...
4405	4406	42	No	Travel_Rarely	Research & Development	5	4	Medical	1	Female	...	10.0
4406	4407	29	No	Travel_Rarely	Research & Development	2	4	Medical	1	Male	...	10.0
4407	4408	25	No	Travel_Rarely	Research & Development	25	2	Life Sciences	1	Male	...	5.0
4408	4409	42	No	Travel_Rarely	Sales	18	2	Medical	1	Male	...	10.0
4409	4410	40	No	Travel_Rarely	Research & Development	28	3	Medical	1	Male	...	NaN

4410 rows × 29 columns

Extracting related columns and building a subdf (df1)

```
df1=df[['Age','Attrition','Department','Gender','MonthlyIncome']]
df1
```

	Age	Attrition	Department	Gender	MonthlyIncome
0	51	No	Sales	Female	131160
1	31	Yes	Research & Development	Female	41890
2	32	No	Research & Development	Male	193280
3	38	No	Research & Development	Male	83210
4	32	No	Research & Development	Male	23420
...	...	...	...	...	...
4405	42	No	Research & Development	Female	60290
4406	29	No	Research & Development	Male	26790
4407	25	No	Research & Development	Male	37020
4408	42	No	Sales	Male	23980
4409	40	No	Research & Development	Male	54680

4410 rows × 5 columns

Understanding the data

```
df1.head()
```

	Age	Attrition	Department	Gender	MonthlyIncome
0	51	No	Sales	Female	131160
1	31	Yes	Research & Development	Female	41890
2	32	No	Research & Development	Male	193280
3	38	No	Research & Development	Male	83210
4	32	No	Research & Development	Male	23420

```
df1.tail()
```

	Age	Attrition	Department	Gender	MonthlyIncome
4405	42	No	Research & Development	Female	60290
4406	29	No	Research & Development	Male	26790
4407	25	No	Research & Development	Male	37020
4408	42	No	Sales	Male	23980
4409	40	No	Research & Development	Male	54680

```
df1.shape
```

```
(4410, 5)
```

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4410 entries, 0 to 4409
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Age            4410 non-null   int64 
 1   Attrition      4410 non-null   object
 2   Department     4410 non-null   object
 3   Gender         4410 non-null   object
 4   MonthlyIncome  4410 non-null   int64 
dtypes: int64(2), object(3)
memory usage: 172.4+ KB
```

```
df1.describe()
```

	Age	MonthlyIncome
count	4410.000000	4410.000000
mean	36.923810	65029.312925
std	9.133301	47068.888559
min	18.000000	10090.000000
25%	30.000000	29110.000000
50%	36.000000	49190.000000
75%	43.000000	83800.000000
max	60.000000	199990.000000

Cleaning the data
(unnecessary in our case)

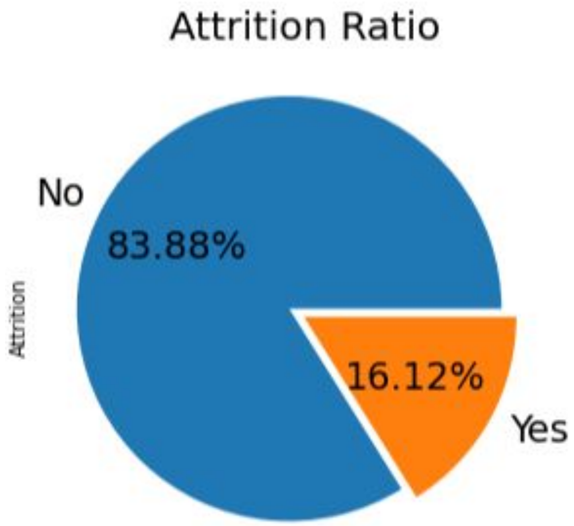
```
df1.isnull().any()
```

```
Age              False
Attrition        False
Department       False
Gender           False
MonthlyIncome    False
dtype: bool
```

```
df1.isnull().sum()
```

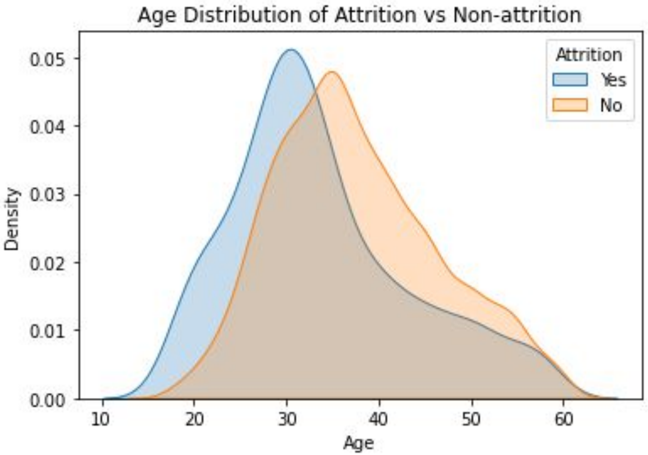
```
Age              0
Attrition        0
Department       0
Gender           0
MonthlyIncome    0
dtype: int64
```

Attrition
Univariate Analysis - Pie chart



83.88% of employees are still working for the company; while 16.12% of the employees have already resigned from the company.

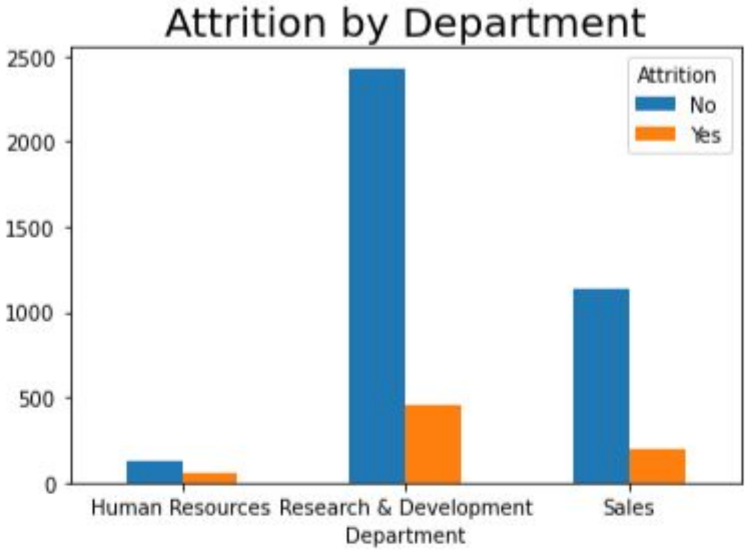
Attrition by Age
Bivariate Analysis - Multi distribution plot
SIGNIFICANT



We can observed that the distribution of the non-attrition employees follows an approximate normal distribution. The age distribution for attrition is positively skewed to the right. The age of attrited employees mostly fall under the range 29-35 (approx).

The millennial generation (1981 to 1996), in particular, has been stigmatized as job-hoppers. Seventy-five percent of millennials believe job hopping can be good for their careers. (Gregory, 2018)

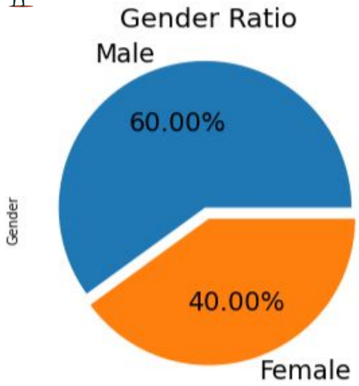
Attrition by Department
Bivariate Analysis - Crosstab
SIGNIFICANT

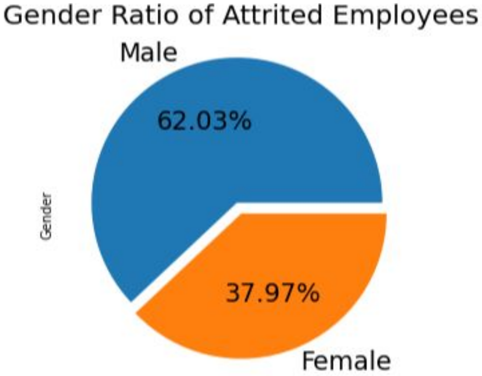


R&D department has the highest attrition rate among all departments.

Over the 7-year period after their organizational entry, R&D professionals were more likely to leave the organization.
(Chang et al., 2010)

Attrition by Gender
Univariate Analysis - Pie chart
INSIGNIFICANT

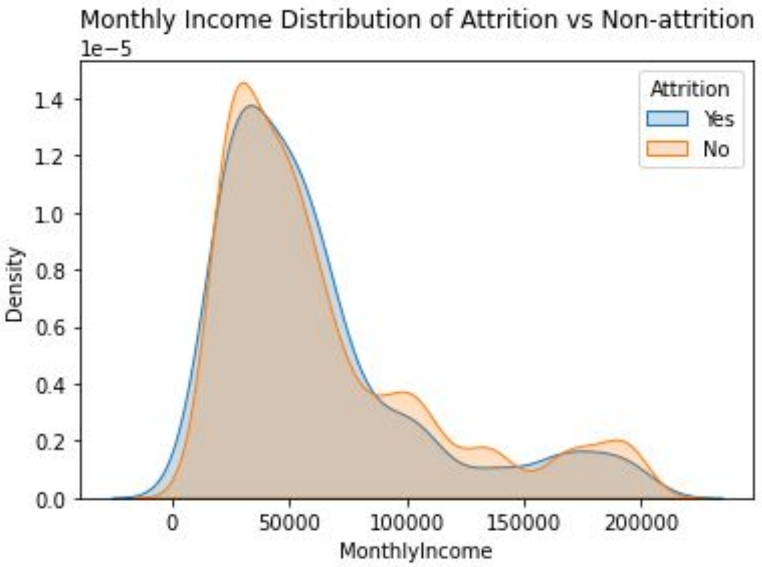




Based on the graphs, we can see that the Gender ratio of the company as well as the Gender ratio for the Attrition are fairly similar suggesting that Gender might not be a significant factor in affecting employee's decision to leave the company.

Gender does not moderate the effect of job satisfaction on employees' intention to quit.
(Emmanuel, N. & Agaha, A.N.L., 2021)

Attrition by Monthly Income
Bivariate Analysis - Multi distribution plot
INSIGNIFICANT



We observe that the distribution of both attrition and non-attrition are similar and skewed to the right. This suggests that Monthly Income might not be a significant impact in affecting employee's decision to leave the company.

Salaries of employees might not be an important factor in employees' attrition compared to other factors such as toxic work culture.
(Rubenstein et al., 2017)

Dealing with Categorical Data

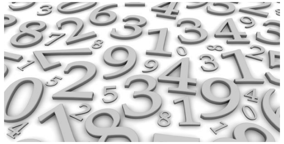
	Age	Attrition	Gender	MonthlyIncome	Research & Development	Sales
0	51	0	0	131160	0.0	0.0
1	31	1	0	41890	0.0	1.0
2	32	0	1	193280	0.0	1.0
3	38	0	1	83210	0.0	1.0
4	32	0	1	23420	0.0	1.0
...	...	...	...	...	...	...
4405	42	0	0	60290	0.0	1.0
4406	29	0	1	26790	0.0	1.0
4407	25	0	1	37020	0.0	1.0
4408	42	0	1	23980	0.0	0.0
4409	40	0	1	54680	0.0	1.0

4410 rows × 6 columns

Attrition:
LabelEncoder.fit_transform()
assign ‘1’ to ‘Yes’;
assign ‘0’ to ‘No’

Gender:
LabelEncoder.fit_transform()
assign ‘1’ to ‘Male’;
assign ‘0’ to ‘Female

Department:
pd.getdummies()
create dummy variables



Dealing with Numerical Data

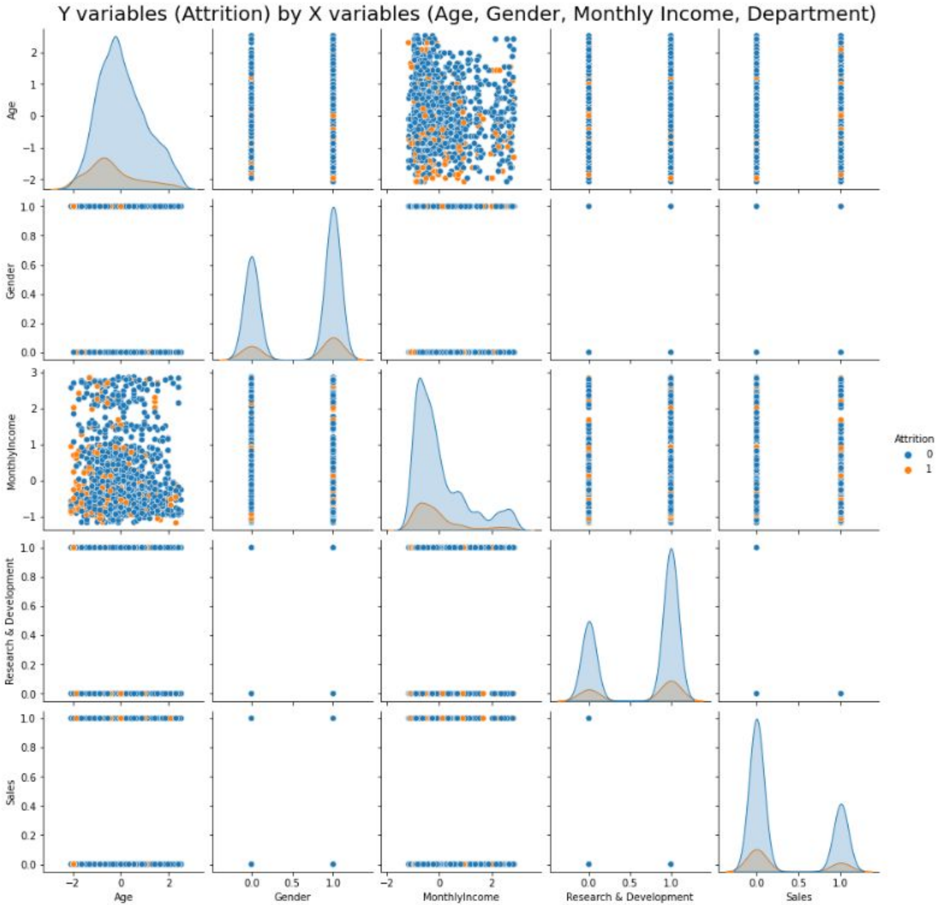
	Age	Attrition	Gender	MonthlyIncome	Research & Development	Sales
0	1.541369	0	0	1.405136	0.0	0.0
1	-0.648668	1	0	-0.491661	0.0	1.0
2	-0.539166	0	1	2.725053	0.0	1.0
3	0.117845	0	1	0.386301	0.0	1.0
4	-0.539166	0	1	-0.884109	0.0	1.0
...	...	...	...	...	...	...
4405	0.555852	0	0	-0.100700	0.0	1.0
4406	-0.867672	0	1	-0.812504	0.0	1.0
4407	-1.305679	0	1	-0.595138	0.0	1.0
4408	0.555852	0	1	-0.872210	0.0	0.0
4409	0.336849	0	1	-0.219901	0.0	1.0

4410 rows × 6 columns

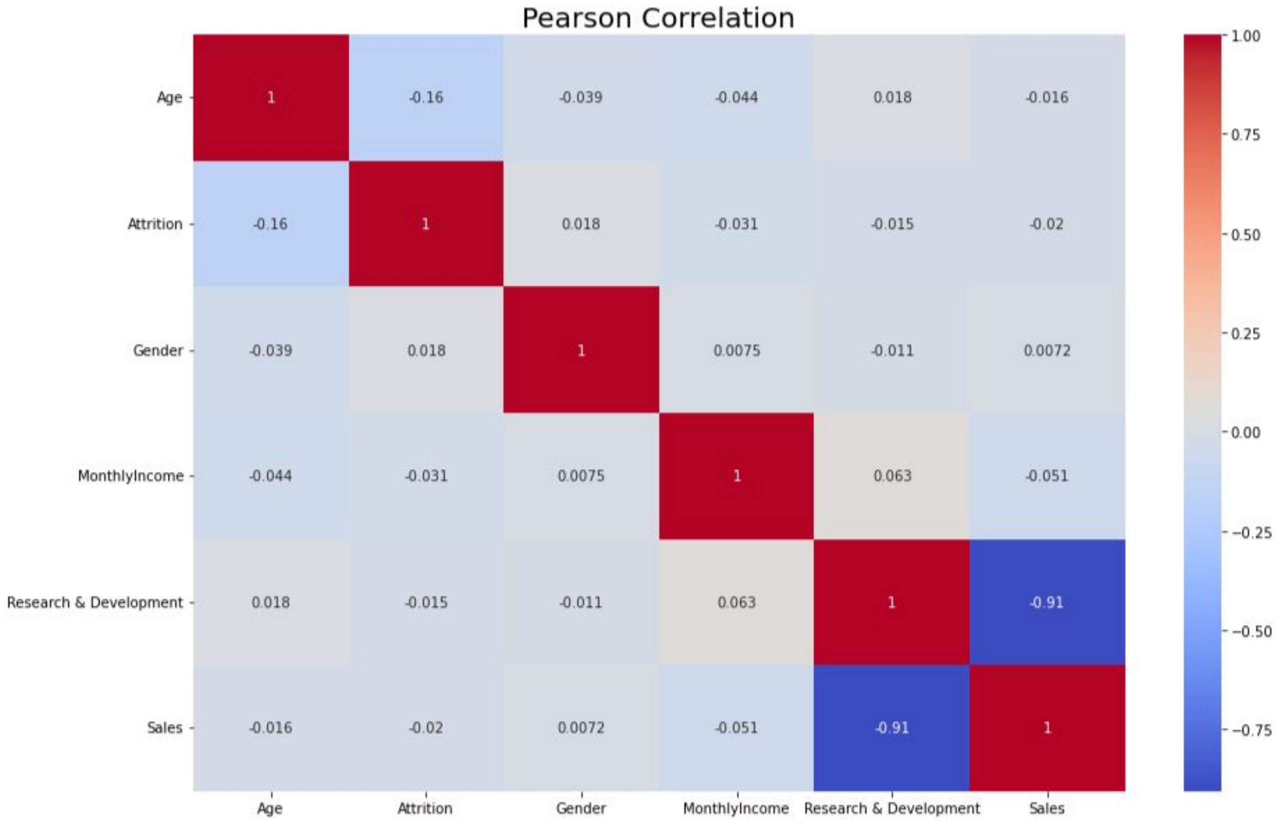
Age & **Monthly Income**:

StandardScaler.fit_transform()
to normalize the data

Attrition by Age, Gender, Monthly Income & Department
Advanced Visualization - Pairplot



Age, Attrition, Gender, Monthly Income & Department Multivariate Analysis - Heatmap



Logistic Regression Model

```python
y=df1['Attrition']
x=df1[['Age','Gender','MonthlyIncome','Research & Development','Sales']]
```

```python
import statsmodels.api as sm
model=sm.Logit(y,x)
result=model.fit()
print(result.summary())
```

```
==============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Age                         -0.4748      0.046    -10.410      0.000      -0.564      -0.385
Gender                      -0.0980      0.080     -1.230      0.219      -0.254       0.058
MonthlyIncome               -0.0940      0.043     -2.179      0.029      -0.178      -0.009
Research & Development      -1.6887      0.071    -23.881      0.000      -1.827      -1.550
Sales                       -1.7644      0.092    -19.258      0.000      -1.944      -1.585
==============================================================================================
```

Conclusion

Based on our findings from the logistic regression, we can conclude that our initial analysis from the EDA and the test for significance in the Logistic Regression differs slightly in the table below.

Variables	EDA	Logistic Regression
Age	Significant	Significant
Department	Significant	Significant
Gender	Insignificant	Insignificant
Monthly Income	Insignificant	Significant

Thank You!

References

Chang, J. Y., Choi, J. N. & Kim, M. U. (2010). Turnover of highly educated R&D professionals: The role of pre-entry cognitive style, work values and career orientation. *Journal of Occupational & Organizational Psychology, 81*(2), 299-317. https://doi.org/10.1348/096317907X204453

Gregory, E. (August, 2018). The Age of Employment Turnover. *ADP*. https://www.adp.com/spark/articles/2018/08/the-age-of-employment-turnover-take-two.aspx

Nwahanye E. & Atabong N.L.A. (2021). Testing the Moderating Effect of Gender on Job Satisfaction and Employees' Behaviours Relationship: Evidence from Mobile Telecommunication Network (MTN) Buea, Cameroon. *European Scientific Journal*, *17*(1), 236. https://doi.org/10.19044/esj.2021.v17n1p236

Rubenstein, A., Lee, T., Mitchell, T. & Eberly, M. (2017). Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. *Personnel Psychology, 71*(1), https://doi.org/10.1111/peps.12226