Project Portfolio

1. Stroke Prediction

2. Visualization Dashboard on Attrition and Retention Level of Labour Market

Stroke Prediction

Stroke, according to the World Health Organization (WHO) is the second leading cause of death globally since 2000. Unless controlled, stroke can lead to several health complications.

- Analysis to identify relationship between individuals' lifestyle and physical measurements to the likelihood of Stroke

Problem

Every 40 seconds, someone in the US suffers a stroke

Long-term disability - cognitive and motor dysfunction

What factors increase the risk of Stroke?

Analysis to draw conclusions about the likelihood of Stroke among several physical parameters.

Process

1 **Understanding the Data**
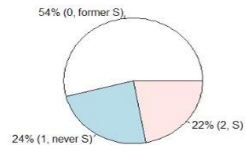
2 **Data Cleaning**

3 **Data Modeling**

4 **Data Analysis**

5 **Insights**

EDA
(categorical variables)

Piechart of smoking_status

54% (0, former S)

24% (1, never S)

22% (2, S)

Smoking status

0	1	2
1852	837	737

0: Formerly smokes
1: Never smoked
2: Smokes

Piechart of Stroke

95% (0, if patient never had stroke)

5% (1, if patient has had stroke)
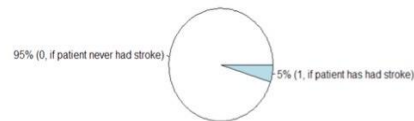
Stroke count

0	1
3246	180

0: Never had a stroke
1: Has had a stroke

Piechart of Hypertension

88% (0, no HT)

12% (1, HT)
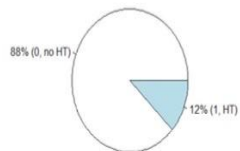
Hypertension

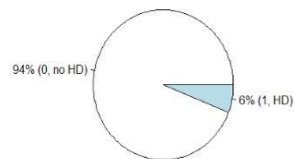| 0	1
3018	408

0: Never had hypertension
1: Has hypertension

Piechart of Heart_disease

94% (0, no HD)

6% (1, HD)

Heart disease

0	1
3220	206

0: Never had heart disease
1: Has heart disease

EDA
(numerical variables)

Histogram of log(y$avg_glucose_level) | Boxplot of avg_glucose_level without outliers

- 465 outlying values are removed using function in R (data points between 175.29 and 271.74)

- The log-transformation (base *e*) is applied.

- After removing outliers, *avg_glucose_level* follows a log-normal distribution.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
55.12	74.80	88.10	91.76	103.89	174.54

Histogram of clean$age | Boxplot of clean$age

No outliers detected

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10.00	34.00	50.00	48.65	63.00	82.00

Histogram of bmi | Boxplot of bmi without outliers

- 90 outlying values (ranging from 11.5 to 92.0) are removed using a function in R

- After removing outliers, *bmi* follows a normal distribution.

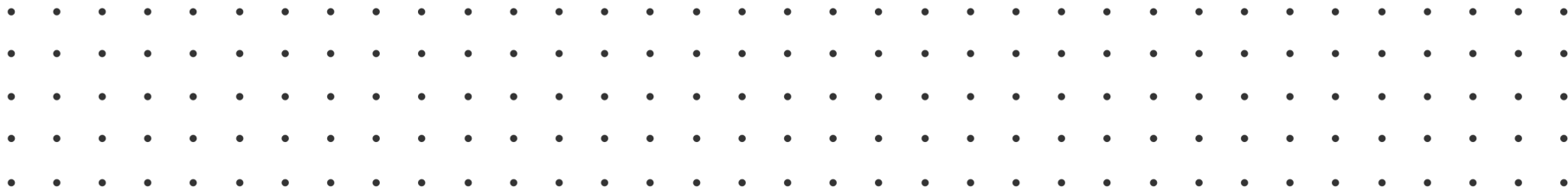Min. 1st Qu. Median Mean 3rd Qu. Max.

Insights

LIKELIHOOD OF
STROKE IS
INDEPENDENT
OF THEIR BMI

LIKELIHOOD OF
STROKE IS
DEPENDENT ON
AVERAGE
GLUCOSE LEVEL

AGE, HYPERTENSION,
HEART DISEASE AND
AVERAGE GLUCOSE
LEVEL ARE
SIGNIFICANT
FACTORS

Association of BMI and Stroke

Null Hypothesis H_0: The mean BMI of people with stroke is equal to those without stroke.
$(\mu_{stroke} = \mu_{no\,stroke})$

Alternative Hypothesis H_1: The mean BMI of people with stroke is less than those without stroke. $(\mu_{stroke} < \mu_{no\,stroke})$

```
> var.test(bmi ~ stroke)

        F test to compare two variances

data:  bmi by stroke
F = 1.1737, num df = 3158, denom df = 176, p-value = 0.1631
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9365309 1.4404280
sample estimates:
ratio of variances
        1.173699
```

Association of BMI and Stroke

```
> t.test(bmi~stroke, var.equal=T, alt="less")

        Two Sample t-test

data:  bmi by stroke
t = -1.3921, df = 3334, p-value = 0.08199
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf 0.1213788
sample estimates:
mean in group 0 mean in group 1
        29.63941        30.30678
```

Association of Hypertension and Stroke

	No HT	Yes HT
No Stroke	2895	351
Yes Stroke	123	57

Proportion table

	No HT	Yes HT
No Stroke	0.84500876	0.10245184
Yes Stroke	0.03590193	0.01663748

Null Hypothesis H_0: There is no association between stroke & hypertension.
Alternative Hypothesis H_1: There is some association between stroke & hypertension.

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  table
X-squared = 68.72, df = 1, p-value < 2.2e-16
```

Association of Smoking status and Stroke

Null Hypothesis H_0: There is no association between stroke & smoking status.
Alternative Hypothesis H_1: There is some association between stroke & smoking status.

	Former S	Never S	Current S
No Stroke	1768	780	698
Yes Stroke	84	57	39

```
          Pearson's Chi-squared test

data:  table
X-squared = 5.9932, df = 2, p-value = 0.04996
```

Key Factors of Stroke

```
> mod = glm(clean2$stroke~clean2$age+clean2$hypertension+clean2$heart_disease+clean2$a
vg_glucose_level+clean2$bmi+clean2$smoking_status)
> summary(mod)

Call:
glm(formula = clean2$stroke ~ clean2$age + clean2$hypertension +
    clean2$heart_disease + clean2$avg_glucose_level + clean2$bmi +
    clean2$smoking_status)

Deviance Residuals:
    Min       1Q    Median        3Q       Max
-0.20830  -0.06365  -0.02989   0.00054   0.99850

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               8.509e-01  2.551e-02  33.356  < 2e-16 ***
clean2$age                1.710e-03  2.036e-04   8.397  < 2e-16 ***
clean2$hypertension       3.613e-02  1.169e-02   3.090 0.002022 **
clean2$heart_disease      5.558e-02  1.557e-02   3.571 0.000362 ***
clean2$avg_glucose_level  2.134e-04  7.976e-05   2.676 0.007490 **
clean2$bmi               -7.143e-04  5.842e-04  -1.223 0.221525
clean2$smoking_status     3.084e-03  4.415e-03   0.698 0.484968
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.03629292)

    Null deviance: 110.41  on 2881  degrees of freedom
Residual deviance: 104.34  on 2875  degrees of freedom
AIC: -1369.3

Number of Fisher Scoring iterations: 2
```

Summary

ANALYSIS

Analyze the relationship between a patient's lifestyle and physical measurements to determine significant factors that lead to Stroke

INSIGHTS

Age, hypertension, heart disease and glucose level are significant predictor variables to determine the likelihood of Stroke. Precautionary measures such as observing glucose levels and being mindful to their exposure to hypertension and smoking should be considered to lower risk of Stroke.

NEXT STEPS

This analysis is insightful but future research could look towards a deeper and wider analysis of data from official health records or partnering with health organisations.

Attrition and Retention

Attrition is the reduction in the number of employees through retirement, resignation or death.

- Reduce attrition rates of the organisation
- Control employee churn through predictive modelling which can be used for developing retention strategies
- Visuals grouped into categories - demographics, job roles, etc and a summary of the analysis

Insights

Largest percentage
of attrition
Age 26 to 35

Education, department
and role that accounts
for the most attrition
**Degree, R&D
department, lab
technician role**

Salary with highest
attrition rates
$2,000 to $5,000

Attrition - Education

Marital Status
- [ ] Divorced
- [ ] Married
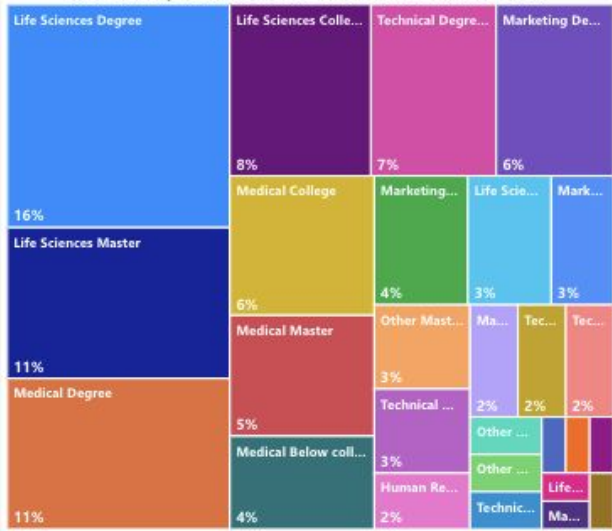- [ ] Single

Gender
- [ ] Female
- [ ] Male

Attrition by Education and Educational Qualification

Life Sciences Degree **16%**	Life Sciences Colle... **8%**	Technical Degre... **7%**	Marketing De... **6%**
Life Sciences Master **11%**	Medical College **6%**	Marketing... **4%**	Life Scie... **3%** / Mark... **3%**
Medical Degree **11%**	Medical Master **5%**	Other Mast... **3%**	Ma... **2%** / Tec... **2%** / Tec... **2%**
	Medical Below coll... **4%**	Technical ... **3%**	Other ...
		Human Re... **2%**	Other ... / Life... / Technic... / Ma...

Attrition by Educational Qualification
100%

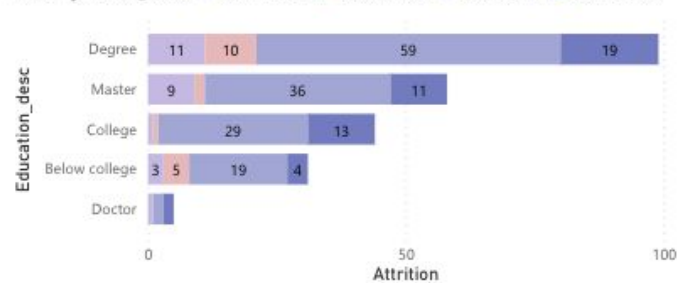| Qualification	Attrition
Degree	41.8%
Master	24.5%
College	18.6%
Below college	13.1%
Doctor	2.1%

5.1%
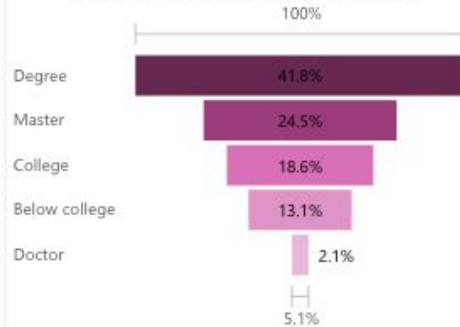
Insights

- Employees with a Degree Qualification has the highest attrition rate (41.8%) as compared the other qualification levels

- employees in the Life Sciences field have the highest attrition rate (37.6%), while those in Human Resource field have the lowest rate (2.95%)

- Single employees with a Degree qualification contribute the most to attrition, while Single and Divorced employees with Doctor qualification contribute the least.

- Degree holders whose monthly income ranges between $2,000 to $5,000 have the highest attrition rates, while doctors across the different monthly income range, have the lowest attrition rates
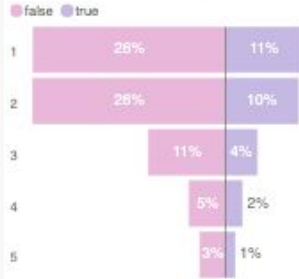
Attrition by Educational Qualification

MonthlyIncome_Gro... ● 10,001-20,000 ● 2,000 or less ● 2,001-5,000 ● 5,001-10,000
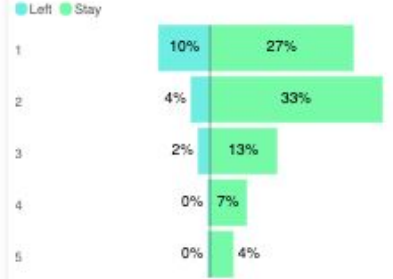
Education_desc	Attrition
Degree	11 / 10 / 59 / 19
Master	9 / 36 / 11
College	29 / 13
Below college	3 / 5 / 19 / 4
Doctor	

0   50   100
Attrition

Attrition - Department

Employee Count by Job Level and Overtime
● false ● true

Level	false	true
1	26%	11%
2	26%	10%
3	11%	4%
4	5%	2%
5	3%	1%

Attrition by Job Level
● Left ● Stay

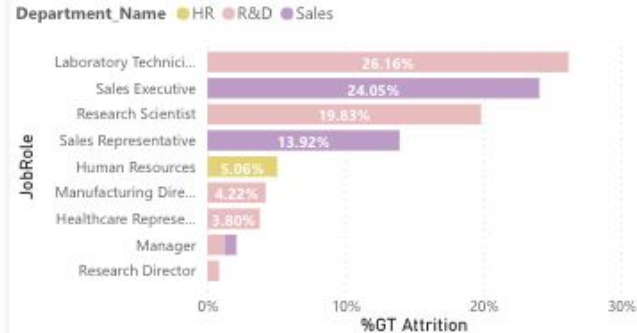| Level	Left	Stay
1	10%	27%
2	4%	33%
3	2%	13%
4	0%	7%
5	0%	4%

Insights

- Employees in the R&D department have the most ratings for high job satisfaction, but also have the highest attrition rate. Employees in HR have the lowest ratings for job satisfaction

- Laboratory Technician in particular have the highest attrition rate, followed by Sales Executive.

- Employees at the lower levels clock in the most over-time, as compared to those of higher levels

- Employees of lower levels contribute more significantly than the other levels to attrition.
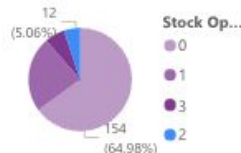
Attrition by Department
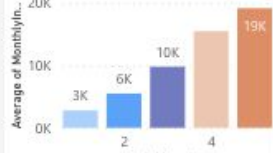job_satisfaction_... ● high ● low ● medium ● very high

Department_Name	high	low	medium	very high
R&D	18.14%	16.03%	10.13%	11.81%
Sales	11.39%	9.70%	8.44%	9.28%
HR				

%GT Attrition

Attrition by Job Role
Department_Name ● HR ● R&D ● Sales

JobRole	%GT Attrition
Laboratory Technici...	26.16%
Sales Executive	24.05%
Research Scientist	19.83%
Sales Representative	13.92%
Human Resources	5.06%
Manufacturing Dire...	4.22%
Healthcare Represe...	3.80%
Manager	
Research Director	

Attrition - Financial

Attrition by StockOptionLevel

12
(5.06%)

Stock Op...
- 0
- 1
- 3
- 2

154
(64.98%)

Average of MonthlyIncome by JobLevel

Average of MonthlyIn...

20K

19K

10K
10K

6K
3K

0K

JobLevel: 2, 4

Attrition by Monthly Income and Age Group

AgeGroup: ●18-25 ●26-35 ●36-45 ●46-55 ●56 and Above

%GT Attrition

30% — 33%

20% —

13% 11%
10% — 12%
3% 2% 4% 4% 5%
4% 1% 4% 4% 2%
0% — 0% 0% 1% 1% 0%

2,001-5,000   5,001-10,000   10,001-20,000   2,000 or less

MonthlyIncome_Group
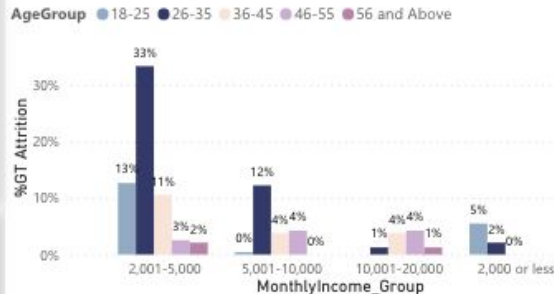
Attrition by Monthly Income Group and Department

Department_Name: ●HR ●R&D ●Sales

145

Sales

R&D

49

Sales

25   18
Sales   R&D   Sales   R&D   HR

10,001-20,000   2,000 or less   2,001-5,000   5,001-10,000
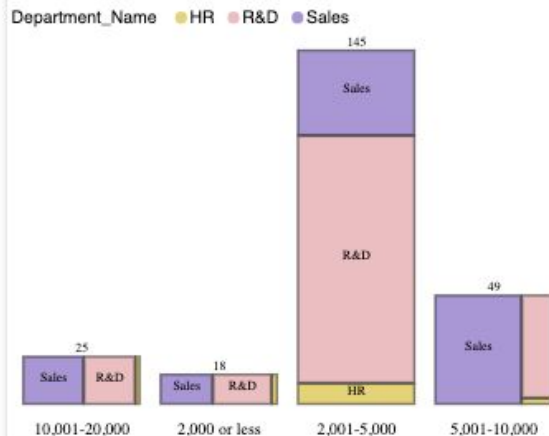
Attrition by Monthly Income and Job Level

JobLevel: ●1 ●2 ●3 ●4 ●5

MonthlyIncome_Group

2,001-5,000 | 52.74% | 8.44%

5,001-10,000 | 13.50% | 7.17%

10,001-20,000 |

2,000 or less | 7.59%

0%   20%   40%   60%
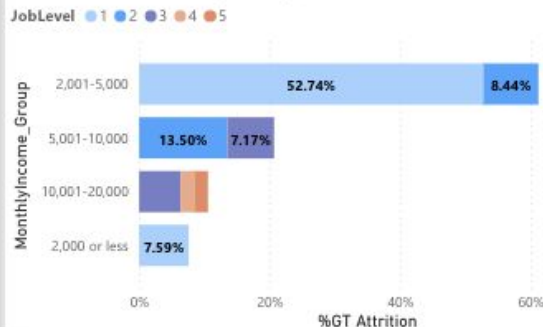
%GT Attrition

Insights

- Employees earning between $2,000 to $5,000 have the highest attrition rate, while those earning less than $2,000 have the lowest rate.

- Average salary is in direct proportion to job level. Employees working in higher levels earning higher wages.

- Highest movement is evident in employees of ages 26 to 35

- 72% of employees earning less than $2,000 are of the age group 18 to 25.

- Individuals' earning between $2,000 to $5,000 are the greatest contributors to attrition levels compared to other income groups. Of that income group, the department with the highest attrition is R&D.

- Employees with no stock option account for 65% of attrition levels, as opposed to those with some stock option holdings.
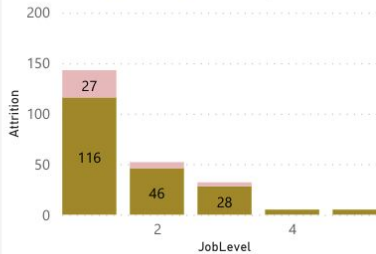
Attrition - Job Performance

Attrition by YearsSinceLastPromotion and years_at_company_grp

years_at_company_grp ● 1-10 Years ● 11-20 Years ● 21-30 Years ● 31-40 Years
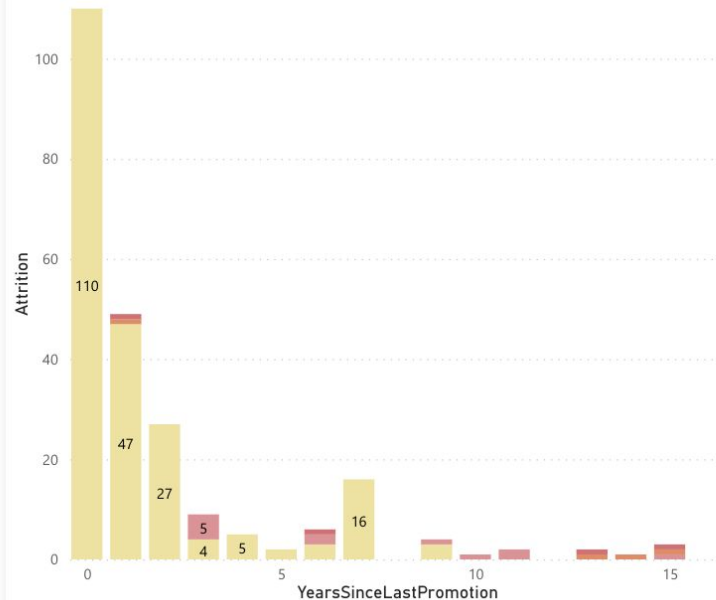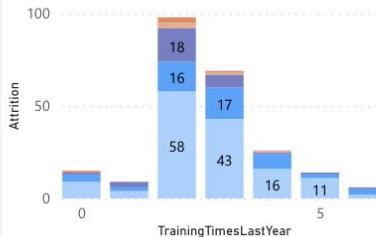

Attrition by Job Level and Performance Rating

Performan... ● Excellent ● outstanding

Attrition by Training Times Last Year and Job Level
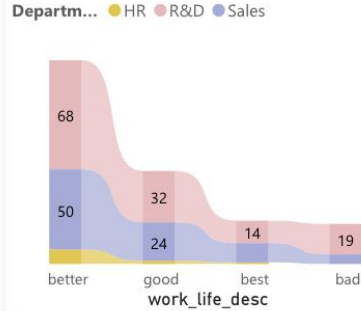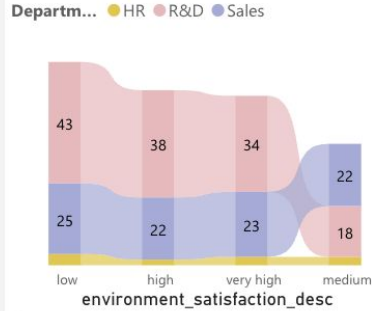
JobLevel ● 1 ● 2 ● 3 ● 4 ● 5

Insights

- Employees who were trained 2 to 3 times last year contributed the most to attrition, as opposed to those with more than 5 training times

- Most employees who were trained 2 to 3 times last year were from the lower levels

- Employees with an excellent job rating have the highest attrition rates, of which most are those working at lower levels

- Employees working in the company for 1 to 10 years have the shortest time frame of years since their last promotion. However, these employees are the ones with the highest attrition value

Summary

Key influencers | Top segments

AgeGroup ✕
26-35

Department_Na... ✕
R&D

JobRole ✕
Laboratory Technician

overtim...

What influences Attrition_desc to be [Left ▾] ?

Attrition
237

26-35
116

18-25
44

36-45
43

46-55
26

R&D
67

Sales
41

HR
8

Laboratory Technician
31

Research Scientist
29

Healthcare Representa...
3

Manufacturing Director
3

True
16

False
15

When...

....the likelihood of Attrition_desc being Left increases by

MonthlyIncome_Group is 2,000 or less — 3.58x

TotalWorkingYears is 2 or less — 3.23x

overtime_desc is True — 2.93x

JobRole is Sales Representative — 2.70x

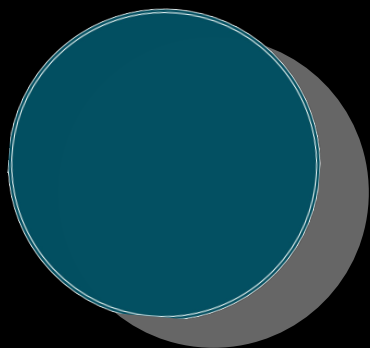years_at_cp2_grp is 1-10 Years — 2.60x

YearsWithCurrManager is 0 or less — 2.57x

AgeGroup is 18-25 — 2.50x

job_involvement_new is Low — 2.24x

Summary:
- An attrition rate of 16% for the company is not ideal
- Age groups 26 to 35 account for a large percentage of the attrition
- Majority of those who left were single males with a degree
- the R&D department, in particular Lab Technician Role, accounts the most for attrition
- Employees earning between $2,000 to $5,000, in particular, those from R&D department account the most for high attrition rates
- Employees from lower job levels with excellent ratings hve the highest attrition rates
- The following are a few contributors to high attrition rates:
Long over-time hours clocked in
Low environment satisfaction
No stock option level
Less training times in the last year

Recommendations:
- overtime is a key factor - look into reviewing workload, increase resource capacity
- Greater equity on stock option levels amongst employees of different job levels
- Higher managerial levels should look into building a more inclusive culture to increase retention of new employees, through team-building activities
- Look into promoting good work-life balance culture, especially for those in lower job levels who clock in the most over-time, through policies like flexible working hours
- Review possibility of raising monthly income based on job performance

Thank you!

ANY QUESTIONS?