EMPLOYEE ATTRITION ANALYSIS

Exploring attrition risks across employee segments using survival analysis and profiling key factors

Table of Contents

1 Introduction

1.1 Purpose of Study

The purpose of this study is to explore and identify the key factors that contribute to employee attrition within the organization. By employing survival analysis techniques, including Cox Proportional Hazards models and Kaplan-Meier survival estimates, the study aims to uncover the relationships between various demographic, job-related, and compensation factors and their impact on employee turnover.

The analysis also incorporates clustering to segment employees into distinct groups, allowing for a deeper understanding of how attrition risks vary across different employee profiles. This segmentation enables the study to assess the unique attrition drivers within each group, facilitating the development of targeted retention strategies. Ultimately, the goal is to provide actionable insights that will help the organization reduce turnover rates, improve employee satisfaction, and create a more stable and productive workforce.

1.2 Research Objectives

The study specifically aims to

1. **To assess the impact of demographic and job-related factors on employee attrition.**

   Evaluate how variables such as gender, age, department, and overtime contribute to the risk of employee turnover.

2. **To explore the influence of compensation-related factors on attrition rates.**

   Examine the effect of stock options, monthly income, and job satisfaction on employees' likelihood of leaving the organization.

3. **To investigate the role of interactions between job dynamics and compensation in influencing attrition.**

   Analyse how the combination of business travel frequency and overtime, as well as the interaction between age and stock options, affects turnover risk.

4. **To identify distinct employee segments based on satisfaction and job dynamics, and examine their respective attrition behaviours.**

   Use segmentation to uncover clusters of employees and explore how variables like time since last promotion influence attrition within these segments.

5. **To understand the moderating effect of job satisfaction on the relationship between age and attrition.**

Determine if job satisfaction alters how age influences turnover risk, thereby providing insights into the retention challenges across different age groups.

1.3 Methodology Overview:

This study employs a multi-step approach, beginning with k-means clustering to segment employees based on work-related and satisfaction variables, excluding demographic factors. By identifying distinct groups, the analysis aims to capture how various employee characteristics relate to attrition risk. Following segmentation, the study utilizes Cox Proportional Hazards models to explore attrition risk. A baseline model examines primary effects of variables like age, gender, and overtime. Building on these findings, an enhanced model incorporates interaction terms (e.g., business travel and overtime) to understand compounded influences on attrition. Finally, Kaplan-Meier survival analysis will provide a non-parametric validation, visualizing attrition over time across clusters. The dataset lacks specific time information, which limits the precision of survival analysis. "YearsAtCompany" is used as a proxy for time due to the absence of actual dates.

2 Description of the Dataset

The dataset contains comprehensive information about employee demographics, job characteristics, satisfaction levels, and attrition status, with a total of 1,470 observations and various variables converted to numeric form for effective analysis of factors affecting employee retention.

- **Gender (Gender_Num):** Categorical variable representing the gender of employees, converted to Gender_Num, where 0 represents female and 1 represents male.
- **Business Travel (BusinessTravel_Num):** Categorical variable describing the frequency of business travel, converted to BusinessTravel_Num, where 0 represents no travel, 1 represents travel rarely, and 2 represents travel frequently.
- **Overtime (OverTime_Num):** Categorical variable indicating whether an employee works overtime, converted to OverTime_Num, where 0 represents no overtime and 1 represents overtime.
- **Environment Satisfaction (EnvironmentSatisfaction_Num):** Ordinal variable measuring satisfaction with the work environment, converted to EnvironmentSatisfaction_Num, where 1 represents low satisfaction, 2 represents medium satisfaction, 3 represents high satisfaction, and 4 represents very high satisfaction.
- **Job Satisfaction (JobSatisfaction_Num):** Ordinal variable indicating satisfaction with the job, converted to JobSatisfaction_Num, where 1 represents low satisfaction, 2 represents medium satisfaction, 3 represents high satisfaction, and 4 represents very high satisfaction.
- **Work-Life Balance (WorkLifeBalance_Num):** Ordinal variable representing perceived balance between work and personal life, converted to WorkLifeBalance_Num, where 1 represents poor balance, 2 represents fair balance, 3 represents good balance, and 4 represents excellent balance.
- **Relationship Satisfaction (RelationshipSatisfaction_Num):** Ordinal variable measuring satisfaction with relationships at work, converted to RelationshipSatisfaction_Num, where 1 represents low satisfaction, 2 represents medium satisfaction, 3 represents high satisfaction, and 4 represents very high satisfaction.
- **Marital Status (MaritalStatus_Num):** Categorical variable indicating the marital status of employees, converted to MaritalStatus_Num, where 0 represents single, 1 represents married, and 2 represents divorced.
- **Department (Dept_Sales, Dept_RD, Dept_HR):** Categorical variable representing the department of an employee, converted to binary variables:
- **Dept_Sales:** 0 represents not in Sales, and 1 represents in Sales.

- **Dept_RD:** 0 represents not in R&D, and 1 represents in R&D.
- **Dept_HR:** 0 represents not in HR, and 1 represents in HR.
- **Job Role** Categorical variable representing an employee's job role, converted to binary variables:
- **JobRole_SalesExec:** 0 represents not a Sales Executive, and 1 represents Sales Executive.
- **JobRole_ResearchSci:** 0 represents not a Research Scientist, and 1 represents Research Scientist.
- **JobRole_LabTech:** 0 represents not a Lab Technician, and 1 represents Lab Technician.
- **JobRole_ManufDir:** 0 represents not a Manufacturing Director, and 1 represents Manufacturing Director.
- **JobRole_HealthRep:** 0 represents not a Healthcare Representative, and 1 represents Healthcare Representative.
- **JobRole_Manager:** 0 represents not a Manager, and 1 represents Manager.
- **JobRole_SalesRep:** 0 represents not a Sales Representative, and 1 represents Sales Representative.
- **JobRole_ResearchDir:** 0 represents not a Research Director, and 1 represents Research Director.
- **JobRole_HR:** 0 represents not in HR, and 1 represents HR.
- **Age:** Continuous variable representing the age of employees in years.
- **Attrition (Attrition_Num):** Categorical variable indicating whether an employee has left the company, converted to Attrition_Num, where 0 represents retained and 1 represents attrition.
- **Daily Rate:** Continuous variable representing employees' daily wage rate.
- **Distance from Home:** Continuous variable representing the commuting distance from an employee's home to the workplace.
- **Education:** Ordinal variable representing the level of education attained by employees, ranging from 1 (e.g., High School) to 5 (e.g., Doctorate).
- **Education Field:** Categorical variable describing the field of education, such as Life Sciences or Technical Degree.
- **Employee Count:** A constant variable representing the number of employees in the dataset, likely 1 for all observations.
- **Employee Number:** A unique identifier for each employee.
- **Hourly Rate:** Continuous variable indicating the hourly wage of employees.
- **Job Involvement:** Ordinal variable indicating the level of employee involvement in their job, ranging from 1 to 4.
- **Job Level:** Ordinal variable representing the level of the job position within the company hierarchy, ranging from 1 to 5.
- **Monthly Income:** Continuous variable representing the monthly salary of employees.
- **Monthly Rate:** Continuous variable representing the monthly payment rate.
- **Number of Companies Worked:** Discrete variable representing the total number of companies an employee has worked at before joining the current organization.
- **Over18:** A constant variable indicating whether an employee is over 18 years of age.
- **Percent Salary Hike:** Continuous variable representing the percentage increase in salary for employees.
- **Performance Rating:** Ordinal variable indicating employees' performance ratings, typically ranging from 1 to 4.
- **Standard Hours:** A constant variable representing standard work hours, likely the same for all observations.
- **Stock Option Level:** Ordinal variable representing the level of stock options offered to employees, ranging from 0 to 3.
- **Total Working Years:** Continuous variable indicating the total number of years an employee has worked.

- **Training Times Last Year:** Discrete variable representing the number of training sessions attended by employees in the last year.
- **Years at Company:** This variable indicates the number of years an employee has been with the current company.
- **Years in Current Role:** Continuous variable representing the number of years an employee has been in their current role.
- **Years Since Last Promotion:** Continuous variable representing the number of years since an employee's last promotion.
- **Years with Current Manager:** Continuous variable indicating the number of years an employee has worked with their current manager.

```
* Importing data for this assignment;
proc import datafile= "Z:\Downloads\CaseStudy_HR.csv"
   dbms=csv out=work.HRDATA replace;
run;

/* Contents of imported data */
proc contents data=HRDATA; run;
/*Transforming Categorical variables*/
data HRDATA_Transformed;
   set HRDATA;
   /* Binary */
   if Attrition = 'Yes' then Attrition_Num = 1; else Attrition_Num = 0;
   if OverTime = 'Yes' then OverTime_Num = 1; else OverTime_Num = 0;

   /* Ordinal */
   if BusinessTravel = 'Non-Travel' then BusinessTravel_Num = 0;
   else if BusinessTravel = 'Travel_Rarely' then BusinessTravel_Num = 1;
   else if BusinessTravel = 'Travel_Frequently' then BusinessTravel_Num = 2;
   if Gender = 'Male' then Gender_Num = 1; else Gender_Num = 0;
   if MaritalStatus = 'Single' then MaritalStatus_Num = 0;
   else if MaritalStatus = 'Divorced' then MaritalStatus_Num = 1;
   else if MaritalStatus = 'Married' then MaritalStatus_Num = 2;

/* Create Dummy Variables for JobRole */
   if JobRole = 'Sales Executive' then JobRole_SalesExec = 1; else JobRole_SalesExec = 0;
   if JobRole = 'Research Scientist' then JobRole_ResearchSci = 1; else JobRole_ResearchSci = 0;
   if JobRole = 'Laboratory Technician' then JobRole_LabTech = 1; else JobRole_LabTech = 0;
   if JobRole = 'Manufacturing Director' then JobRole_ManufDir = 1; else JobRole_ManufDir = 0;
   if JobRole = 'Healthcare Representative' then JobRole_HealthRep = 1; else JobRole_HealthRep = 0;
   if JobRole = 'Manager' then JobRole_Manager = 1; else JobRole_Manager = 0;
   if JobRole = 'Sales Representative' then JobRole_SalesRep = 1; else JobRole_SalesRep = 0;
   if JobRole = 'Research Director' then JobRole_ResearchDir = 1; else JobRole_ResearchDir = 0;
   if JobRole = 'Human Resources' then JobRole_HR = 1; else JobRole_HR = 0;

/* Create Dummy Variables for Department */
   if Department = 'Sales' then Dept_Sales = 1; else Dept_Sales = 0;
         if Department = 'Research & Development' then Dept_RD = 1; else Dept_RD = 0;
   if Department = 'Human Resources' then Dept_HR = 1; else Dept_HR = 0;
run;
```

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
1	Age	Num	8	BEST12.	BEST32.
2	Attrition	Char	3	$3.	$3.
3	BusinessTravel	Char	17	$17.	$17.
4	DailyRate	Num	8	BEST12.	BEST32.
5	Department	Char	22	$22.	$22.
6	DistanceFromHome	Num	8	BEST12.	BEST32.
7	Education	Num	8	BEST12.	BEST32.
8	EducationField	Char	13	$13.	$13.
9	EmployeeCount	Num	8	BEST12.	BEST32.
10	EmployeeNumber	Num	8	BEST12.	BEST32.
11	EnvironmentSatisfaction	Num	8	BEST12.	BEST32.
12	Gender	Char	6	$6.	$6.
13	HourlyRate	Num	8	BEST12.	BEST32.
14	JobInvolvement	Num	8	BEST12.	BEST32.
15	JobLevel	Num	8	BEST12.	BEST32.
16	JobRole	Char	25	$25.	$25.
17	JobSatisfaction	Num	8	BEST12.	BEST32.
18	MaritalStatus	Char	8	$8.	$8.
19	MonthlyIncome	Num	8	BEST12.	BEST32.
20	MonthlyRate	Num	8	BEST12.	BEST32.
21	NumCompaniesWorked	Num	8	BEST12.	BEST32.
22	Over18	Char	1	$1.	$1.
23	OverTime	Char	3	$3.	$3.
24	PercentSalaryHike	Num	8	BEST12.	BEST32.
25	PerformanceRating	Num	8	BEST12.	BEST32.
26	RelationshipSatisfaction	Num	8	BEST12.	BEST32.
27	StandardHours	Num	8	BEST12.	BEST32.
28	StockOptionLevel	Num	8	BEST12.	BEST32.

29	TotalWorkingYears	Num	8	BEST12.	BEST32.
30	TrainingTimesLastYear	Num	8	BEST12.	BEST32.
31	WorkLifeBalance	Num	8	BEST12.	BEST32.
32	YearsAtCompany	Num	8	BEST12.	BEST32.
33	YearsInCurrentRole	Num	8	BEST12.	BEST32.
34	YearsSinceLastPromotion	Num	8	BEST12.	BEST32.
35	YearsWithCurrManager	Num	8	BEST12.	BEST32.

Figure-1: List of variables in Dataset

2.1 Summary Statistics

The summary statistics provide an overview of key variables related to employee demographics, job characteristics, and satisfaction levels. The average age of employees is around 39 years, indicating a generally mature workforce, with ages ranging from 18 to 60 years. The gender distribution shows a slight skew with an average of 0.6, suggesting one gender is somewhat more prevalent. The average monthly income is approximately 6,503, though there is considerable variation, as seen by the high standard deviation of 4,708, indicating substantial income disparity among employees. The average tenure at the company is 7 years, with a wide range from 0 to 40 years, implying a mix of both new hires and long-term employees. Other variables, like job satisfaction, work-life balance, and environment satisfaction, generally have moderate average scores of around 2.7, reflecting room for improvement in overall employee satisfaction. These insights offer a broad understanding of employee characteristics in the dataset.

Summary Stats

The MEANS Procedure

Variable	N	N Miss	Mean	Std Dev	Minimum	Maximum	Median
Age	1470	0	36.9238095	9.1353735	18.0000000	60.0000000	36.0000000
Gender_Num	1470	0	0.6000000	0.4900647	0	1.0000000	1.0000000
Attrition_Num	1470	0	0.1612245	0.3678630	0	1.0000000	0
MonthlyIncome	1470	0	6502.93	4707.96	1009.00	19999.00	4919.00
YearsAtCompany	1470	0	7.0081633	6.1265252	0	40.0000000	5.0000000
BusinessTravel_Num	1470	0	1.0863946	0.5321699	0	2.0000000	1.0000000
OverTime_Num	1470	0	0.2829932	0.4506065	0	1.0000000	0
StockOptionLevel	1470	0	0.7938776	0.8520767	0	3.0000000	1.0000000
DistanceFromHome	1470	0	9.1925170	8.1068644	1.0000000	29.0000000	7.0000000
JobSatisfaction	1470	0	2.7285714	1.1028461	1.0000000	4.0000000	3.0000000
NumCompaniesWorked	1470	0	2.6931973	2.4980090	0	9.0000000	2.0000000
Dept_Sales	1470	0	0.3034014	0.4598835	0	1.0000000	0
YearsSinceLastPromotion	1470	0	2.1877551	3.2224303	0	15.0000000	1.0000000
TrainingTimesLastYear	1470	0	2.7993197	1.2892706	0	6.0000000	3.0000000
WorkLifeBalance	1470	0	2.7612245	0.7064758	1.0000000	4.0000000	3.0000000
RelationshipSatisfaction	1470	0	2.7122449	1.0812089	1.0000000	4.0000000	3.0000000
EnvironmentSatisfaction	1470	0	2.7217687	1.0930822	1.0000000	4.0000000	3.0000000

Figure-2: Summary Stats

```
/* Summary Statistics */
proc means data=HRDATA_Transformed n nmiss mean std min max median;
var age Gender_Num MonthlyIncome YearsAtCompany
        BusinessTravel_Num OverTime_Num StockOptionLevel
        DistanceFromHome jobsatisfaction NumCompaniesWorked
        Dept_Sales  YearsSinceLastPromotion BusinessTravel_Num
        TrainingTimesLastYear WorkLifeBalance  RelationshipSatisfaction
        EnvironmentSatisfaction
; title "Summary Stats";
run;
```

2.2 Sub Sample Analysis

Summary Statistics for Continuous Variables by Attrition Status

The MEANS Procedure

Attrition_Num	N Obs	Variable	N	N Miss	Mean	Std Dev	Minimum	Maximum	Median
0	1233	Age	1233	0	37.5612328	8.8883600	18.0000000	60.0000000	36.0000000
		Gender_Num	1233	0	0.5936740	0.4913461	0	1.0000000	1.0000000
		MonthlyIncome	1233	0	6832.74	4818.21	1051.00	19999.00	5204.00
		YearsAtCompany	1233	0	7.3690187	6.0962981	0	37.0000000	6.0000000
		BusinessTravel_Num	1233	0	1.0567721	0.5268951	0	2.0000000	1.0000000
		OverTime_Num	1233	0	0.2343877	0.4237874	0	1.0000000	0
		StockOptionLevel	1233	0	0.8450933	0.8419851	0	3.0000000	1.0000000
		DistanceFromHome	1233	0	8.9156529	8.0126335	1.0000000	29.0000000	7.0000000
		JobSatisfaction	1233	0	2.7785888	1.0932774	1.0000000	4.0000000	3.0000000
		NumCompaniesWorked	1233	0	2.6455799	2.4600903	0	9.0000000	2.0000000
		YearsSinceLastPromotion	1233	0	2.2343877	3.2347622	0	15.0000000	1.0000000
		TrainingTimesLastYear	1233	0	2.8329278	1.2935853	0	6.0000000	3.0000000
		WorkLifeBalance	1233	0	2.7810219	0.6819067	1.0000000	4.0000000	3.0000000
		RelationshipSatisfaction	1233	0	2.7339822	1.0716030	1.0000000	4.0000000	3.0000000
		EnvironmentSatisfaction	1233	0	2.7712895	1.0711323	1.0000000	4.0000000	3.0000000
1	237	Age	237	0	33.6075949	9.6893499	18.0000000	58.0000000	32.0000000
		Gender_Num	237	0	0.6329114	0.4830311	0	1.0000000	1.0000000
		MonthlyIncome	237	0	4787.09	3640.21	1009.00	19859.00	3202.00
		YearsAtCompany	237	0	5.1308017	5.9499840	0	40.0000000	3.0000000
		BusinessTravel_Num	237	0	1.2405063	0.5339775	0	2.0000000	1.0000000
		OverTime_Num	237	0	0.5358650	0.4997675	0	1.0000000	1.0000000
		StockOptionLevel	237	0	0.5274262	0.8563614	0	3.0000000	0
		DistanceFromHome	237	0	10.6329114	8.4525253	1.0000000	29.0000000	9.0000000
		JobSatisfaction	237	0	2.4683544	1.1180580	1.0000000	4.0000000	3.0000000
		NumCompaniesWorked	237	0	2.9409283	2.6785186	0	9.0000000	1.0000000
		YearsSinceLastPromotion	237	0	1.9451477	3.1530769	0	15.0000000	1.0000000
		TrainingTimesLastYear	237	0	2.6244726	1.2547842	0	6.0000000	2.0000000
		WorkLifeBalance	237	0	2.6582278	0.8164528	1.0000000	4.0000000	3.0000000
		RelationshipSatisfaction	237	0	2.5991561	1.1254375	1.0000000	4.0000000	3.0000000
		EnvironmentSatisfaction	237	0	2.4641350	1.1697913	1.0000000	4.0000000	3.0000000

Figure-3: Sub Sample Statistics

Employees without Attrition (Attrition_Num = 0):

- **Age:** The average age is 37.6 years, with the maximum age reaching 80, indicating that older employees are more likely to stay.
- **Gender (Gender_Num):** The average value is 0.59, suggesting a relatively balanced gender distribution among retained employees, though slightly skewed.

- **Monthly Income:** The average monthly income is 8,332.74, reflecting higher income levels among those who stay, possibly indicating that better salaries may contribute to retention.
- **Years at Company:** Employees have an average tenure of 7.39 years, implying longer periods of employment and suggesting potential satisfaction with career progression or stability.
- **Business Travel (BusinessTravel_Num):** The average score is 1.05, indicating occasional travel, which might be acceptable for most employees who stay.
- **Overtime (OverTime_Num):** The average score is 0.23, indicating infrequent overtime among retained employees, suggesting better work-life balance.
- **Stock Option Level:** The average score of 0.84 indicates that stock options are available to some extent, potentially contributing to employee loyalty.
- **Distance from Home:** The average distance is 9.81, suggesting that employees are willing to commute longer distances when other retention factors are strong.
- **Job Satisfaction:** With an average score of 2.77, job satisfaction appears to be a key contributor to employee retention.
- **Number of Companies Worked:** The average is 2.46, suggesting that employees who stay might have more previous work experience, which could influence their career stability.
- **Years Since Last Promotion:** The average time since last promotion is 2.23 years, indicating that promotion intervals might be longer for those who stay but do not deter them from remaining employed.
- **Training Times Last Year:** The average score of 2.83 reflects a moderate level of training, suggesting ongoing skill development as a potential retention factor.
- **Work-Life Balance:** The average score is 2.84, indicating relatively good work-life balance, a critical factor for employee satisfaction.
- **Relationship Satisfaction:** The average score is 2.73, indicating positive relationships within the workplace.
- **Environment Satisfaction:** With an average of 2.77, a satisfying work environment appears to play a significant role in retaining employees.

Employees with Attrition (Attrition_Num = 1):

- **Age:** The average age is 33.6 years, indicating that younger employees are more prone to leaving.
- **Gender (Gender_Num):** The average is 0.63, suggesting a slightly different gender distribution compared to those who stayed.
- **Monthly Income:** The average income drops to 4,787.00, highlighting a significant income gap between retained and attrited employees, making salary a likely factor for attrition.
- **Years at Company:** The average tenure is 5.13 years, showing shorter periods of employment before leaving, possibly indicating dissatisfaction with career progression or growth.
- **Business Travel (BusinessTravel_Num):** The average score is 0.99, indicating that business travel frequency is similar to that of non-attrited employees, suggesting it may not be a major factor.
- **Overtime (OverTime_Num):** The average overtime score is 0.53, indicating a higher likelihood of experiencing frequent overtime, which may contribute to burnout and eventual departure.
- **Stock Option Level:** The average is 0.53, indicating that employees who left might have had fewer stock options, possibly impacting long-term financial incentives.
- **Distance from Home:** The average distance is 9.63, similar to those who stayed, suggesting that commute length is not a primary driver of attrition.
- **Job Satisfaction:** The average score of 2.48 is lower than that of retained employees, indicating that dissatisfaction may lead to attrition.

- **Number of Companies Worked:** The average of 2.94 suggests that attrited employees might have a higher tendency to change jobs, indicating a potential preference for exploring new opportunities.
- **Years Since Last Promotion:** The average duration is 1.94 years, slightly shorter than that of retained employees, indicating that recent promotions might not prevent attrition.
- **Training Times Last Year:** The average score is 2.62, indicating slightly less training received, suggesting that better skill development opportunities might help in retention.
- **Work-Life Balance:** The average score is 2.66, lower than that of non-attrited employees, pointing to a poorer perception of work-life balance among those who leave.
- **Relationship Satisfaction:** The average score of 2.59 is lower, hinting that weaker workplace relationships could be a factor in attrition.
- **Environment Satisfaction:** With an average of 2.46, the lower environment satisfaction score suggests that work environment improvement could enhance retention.

The statistics suggest that younger employees, those with lower monthly incomes, and those experiencing more overtime are more likely to leave.

```
/* Sub Sample Analysis of the Dataset */
proc means data=HRDATA_Transformed n nmiss mean std min max median;
class Attrition_Num;
var age Gender_Num MonthlyIncome YearsAtCompany
BusinessTravel_Num OverTime_Num  StockOptionLevel DistanceFromHome
Jobsatisfaction NumCompaniesWorked YearsSinceLastPromotion
BusinessTravel_Num TrainingTimesLastYear WorkLifeBalance
RelationshipSatisfaction EnvironmentSatisfaction
;
title "Summary Statistics for Continuous Variables by Attrition Status";
run;
```

3 Exploratory Data Analysis & Transformation of Variables

3.1 Representation of Attrition by Department.

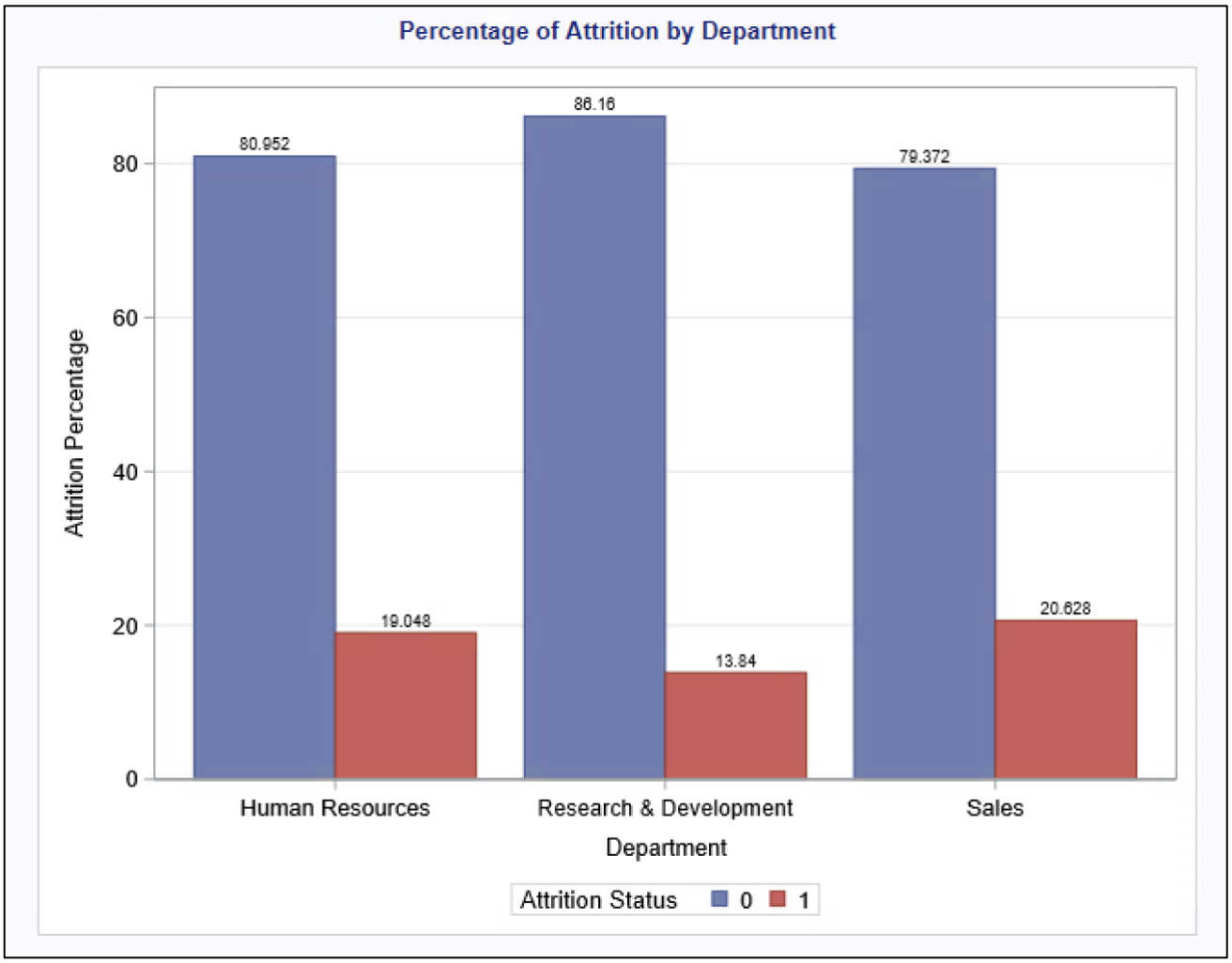


Figure-4: Attrition by Dept.

The bar chart illustrates the percentage of attrition across different departments. It shows the proportion of employees who stayed (Attrition Status = 0) and those who left (Attrition Status = 1) within Human Resources, Research & Development, and Sales departments. Research & Development has the highest retention rate (88.16%), with the lowest attrition percentage (13.84%). The Sales department has the highest attrition rate (20.63%), indicating a higher turnover compared to other departments. Human Resources has a similar pattern to the Sales department, with 19.05% attrition.

3.2 Average Monthly Income by Age and Attrition Status

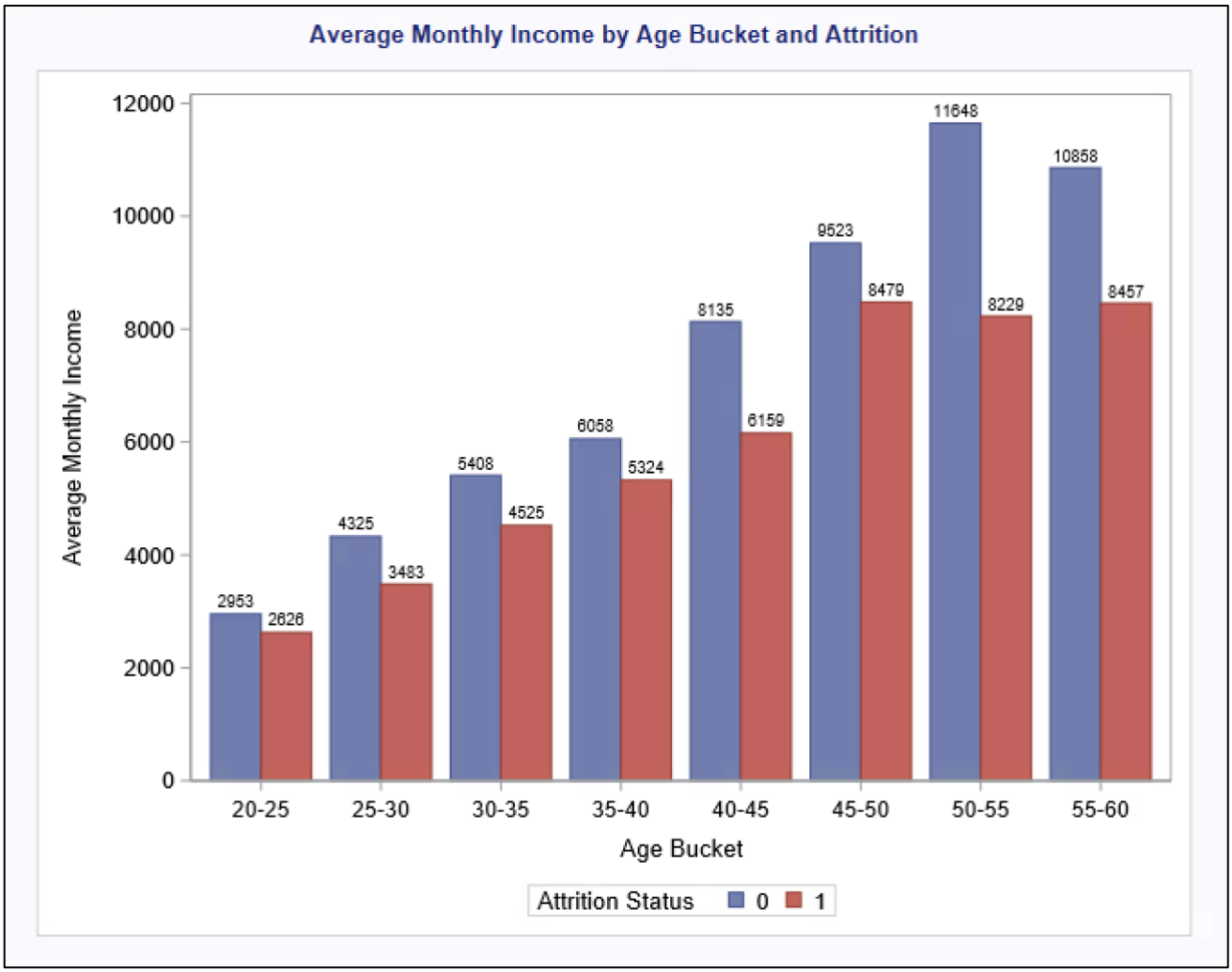


Figure-5: Income by Age/Attrition.

The above bar chart shows the average monthly income across different age buckets for employees who stayed (Attrition Status = 0) and those who left (Attrition Status = 1).

Key observations include:

- Average income generally increases with age across both attrition statuses.
- Employees who stayed tend to have a higher average income than those who left in most age groups, particularly noticeable in the **35-40**, **50-55**, and **55-60** age buckets.
- The income gap between retained and employees who left is more pronounced in older age buckets, suggesting that higher income might be a retention factor among older employees.

This analysis indicates a potential correlation between **age, income**, and **attrition**, with higher incomes contributing to employee retention, especially in older age groups.

```
/* Step 1: Create Age Buckets */
data HRDATA_Bucketed;
   set HRDATA_Transformed;
   length Age_Bucket $10;
   if 20 <= Age < 25 then Age_Bucket = '20-25';
   else if 25 <= Age < 30 then Age_Bucket = '25-30';
   else if 30 <= Age < 35 then Age_Bucket = '30-35';
   else if 35 <= Age < 40 then Age_Bucket = '35-40';
   else if 40 <= Age < 45 then Age_Bucket = '40-45';
   else if 45 <= Age < 50 then Age_Bucket = '45-50';
   else if 50 <= Age < 55 then Age_Bucket = '50-55';
   else if 55 <= Age <= 60 then Age_Bucket = '55-60';
run;

/* Step 2: Calculate Average Monthly Income by Age Bucket and Attrition */
proc means data=HRDATA_Bucketed noprint;
   class Age_Bucket Attrition_Num;
   var MonthlyIncome;
   output out=AvgIncomeByBucket mean=Avg_MonthlyIncome;
run;

/* Step 3: Round Average Monthly Income */
data AvgIncomeByBucket;
   set AvgIncomeByBucket;
   Avg_MonthlyIncome = round(Avg_MonthlyIncome, 1); /* Round to whole number */
run;

/* Step 4: Plot Average Monthly Income by Age Bucket and Attrition */
proc sgplot data=AvgIncomeByBucket;
   vbar Age_Bucket / response=Avg_MonthlyIncome group=Attrition_Num
groupdisplay=cluster datalabel;
   xaxis label="Age Bucket";
   yaxis label="Average Monthly Income";
   title "Average Monthly Income by Age Bucket and Attrition";
   keylegend / title="Attrition Status";
run;
```

3.3 Attrition by Age

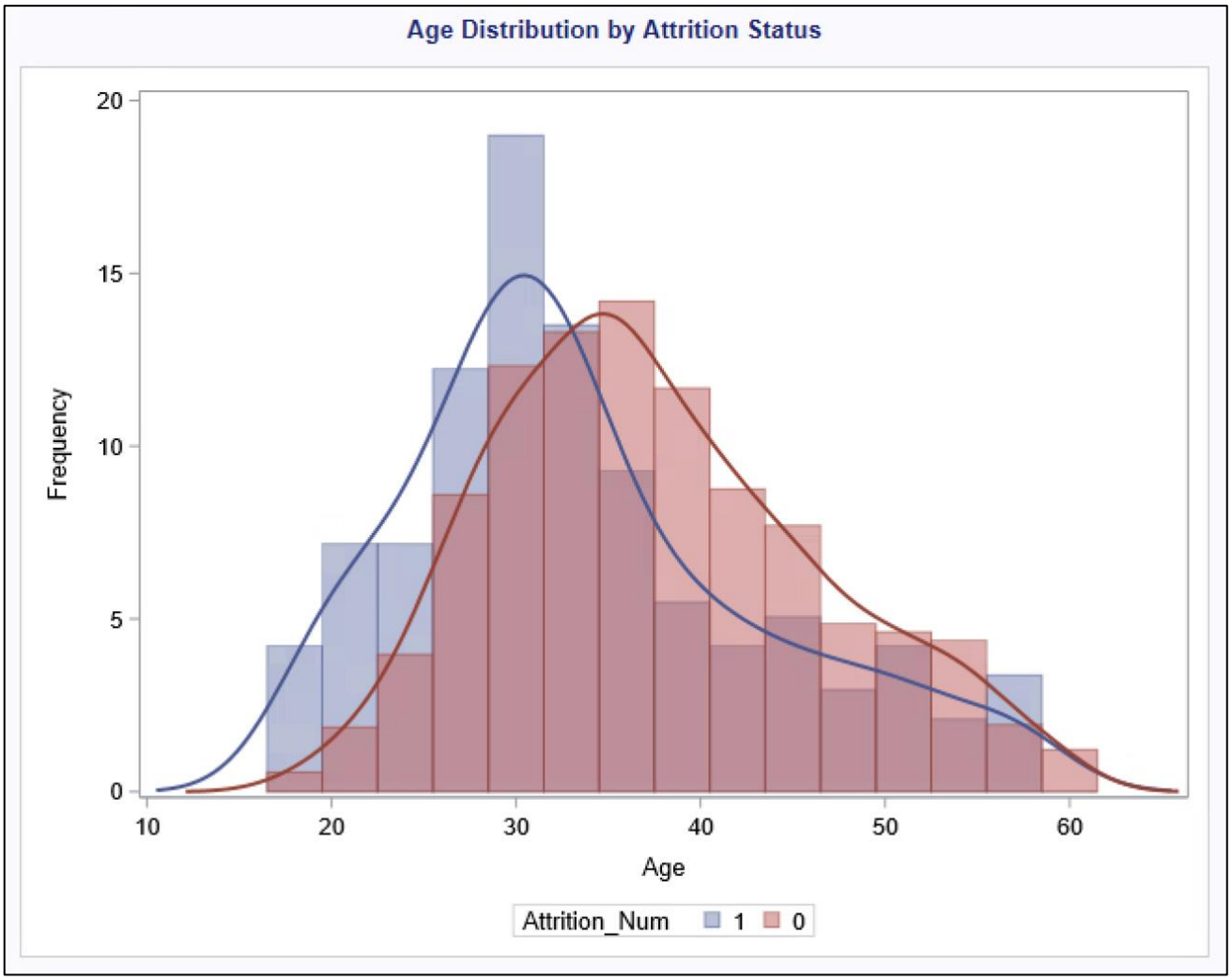


Figure-6: Attrition by Age.

The histogram displays the age distribution of employees categorized by attrition status. Employees who left are more concentrated in the 25-35 age range, indicating higher attrition rates among younger employees. Employees who stayed (blue) are more evenly spread across age groups but peak around 30-40 years. Attrition frequency declines significantly after age 40, suggesting that older employees tend to stay longer. This plot suggests that younger employees are more likely to leave, indicating potential retention challenges within this age group.

```
proc sgplot data=HRDATA_Transformed;
   histogram Age / group=Attrition_Num transparency=0.5;
   density Age / group=Attrition_Num type=kernel;
   title "Age Distribution by Attrition Status";
   xaxis label="Age";
   yaxis label="Frequency";
run;
```

3.4 Impact of Age on Job Satisfaction and Attrition

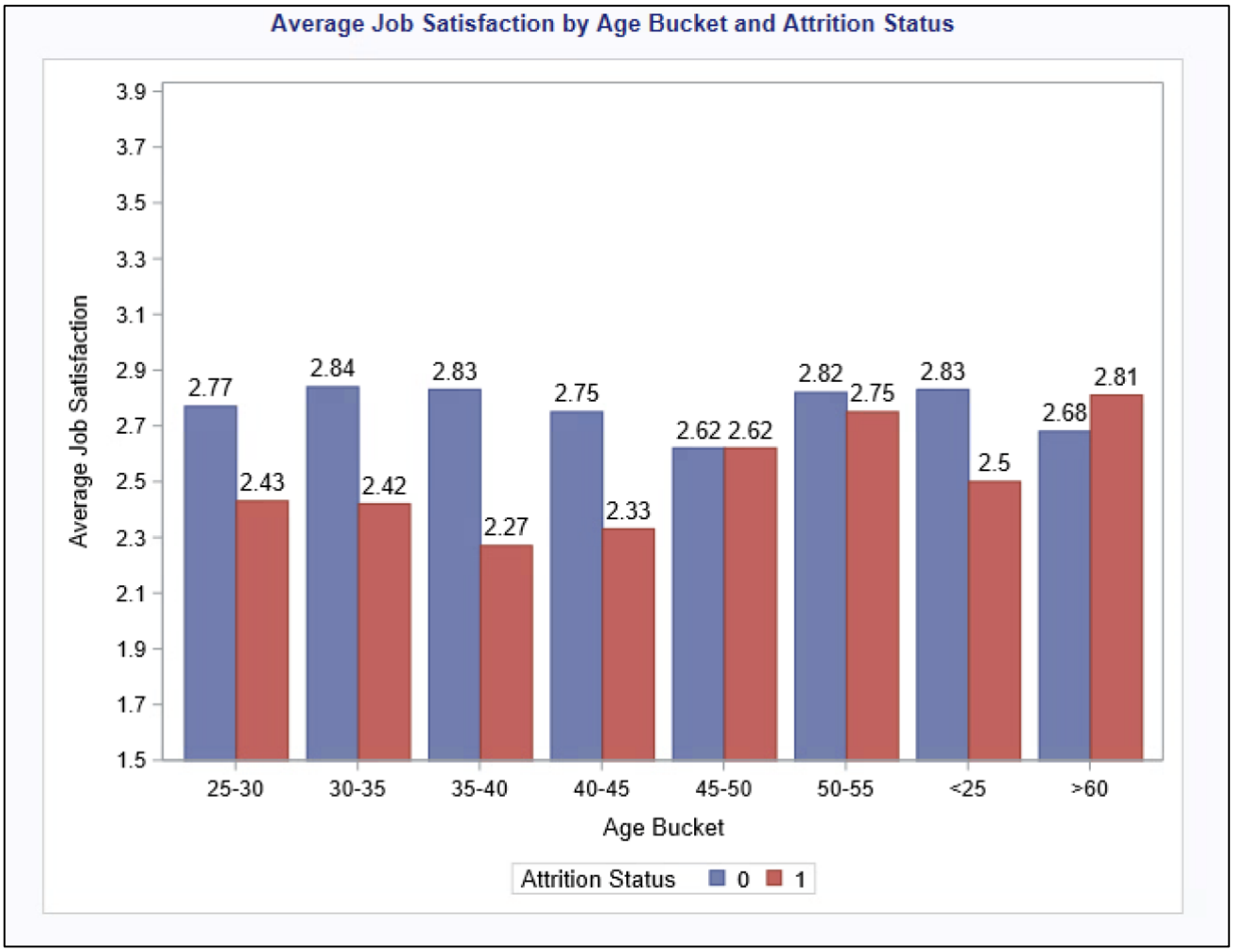


Figure-7: Attrition by Job Satisfaction.

The bar chart shows that **average job satisfaction** is consistently higher among employees who stayed compared to those who left across all age groups. The **30-35** and **<25** age buckets have the highest job satisfaction among retained employees, while the **35-40** age group has the lowest satisfaction among those who left, indicating potential dissatisfaction leading to attrition. The **45-50** age group shows similar job satisfaction levels for both retained and attrited employees, while job satisfaction among those who left slightly improves in older age groups. This suggests that enhancing job satisfaction, particularly among younger employees and those aged 35-40, could be crucial for improving employee retention.

```
/* Step 1: Create Age Buckets */
data HRDATA_Bucketed;
   set HRDATA_Transformed;
   length Age_Bucket $10;
   if 20 <= Age < 25 then Age_Bucket = '<25';
   else if 25 <= Age < 30 then Age_Bucket = '25-30';
   else if 30 <= Age < 35 then Age_Bucket = '30-35';
   else if 35 <= Age < 40 then Age_Bucket = '35-40';
   else if 40 <= Age < 45 then Age_Bucket = '40-45';
   else if 45 <= Age < 50 then Age_Bucket = '45-50';
   else if 50 <= Age < 55 then Age_Bucket = '50-55';
   else  Age_Bucket = '>60';
run;

/* Step 2: Calculate Average Job Satisfaction by Age Bucket and Attrition */
proc means data=HRDATA_Bucketed noprint;
   class Age_Bucket Attrition_Num;
   var JobSatisfaction;
   output out=AvgJobSatisfaction mean=Avg_JobSatisfaction;
run;

/* Step 3: Round Average Job Satisfaction to Two Decimal Places */
data AvgJobSatisfaction;
   set AvgJobSatisfaction;
   Avg_JobSatisfaction = round(Avg_JobSatisfaction, 0.01); /* Round to two decimal
places */
run;

/* Step 4: Plot Average Job Satisfaction by Age Bucket and Attrition */
proc sgplot data=AvgJobSatisfaction;
   vbar Age_Bucket / response=Avg_JobSatisfaction group=Attrition_Num
groupdisplay=cluster datalabel datalabelattrs=(size=10);
   xaxis label="Age Bucket";
   yaxis label="Average Job Satisfaction" values=(1.5 to 4 by 0.2);
   title "Average Job Satisfaction by Age Bucket and Attrition Status";
   keylegend / title="Attrition Status";
run;
```

4 Employee Segmentation

Segmentation was conducted to group employees based on job-related factors that influence their work experience and satisfaction. The selected variables provide insights into different aspects of employees' roles and workplace perceptions. **Monthly Income** differentiates employees by compensation, which can be a significant factor in their job satisfaction and retention decisions. **Business Travel** frequency helps to identify employees with varying work-life demands, as frequent travel can contribute to job stress. **Training Times Last Year** captures the focus on professional development, which can enhance career satisfaction. **Work-Life Balance** assesses how employees perceive their ability to manage work alongside personal responsibilities, a key component of overall well-being. **Relationship Satisfaction** reflects how content employees are with their workplace interactions, impacting their engagement levels. Finally, **Environment Satisfaction** reveals employees' contentment with their physical workspace, which can affect their comfort and productivity. These variables, **focusing on workplace conditions and personal fulfilment**, help in creating meaningful segments that could correlate with employee retention patterns.

Segment Profile by Cluster - Mean and Standard Deviation

	MonthlyIncome		BusinessTravel_Num		TrainingTimesLastYear		WorkLifeBalance		RelationshipSatisfaction		EnvironmentSatisfaction	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Cluster												
1	17280.16	2016.05	1.05	0.49	2.73	1.15	2.82	0.68	2.78	1.06	2.75	1.15
2	8960.97	2041.17	1.08	0.53	2.82	1.39	2.76	0.71	2.71	1.07	2.72	1.06
3	3671.21	1297.25	1.09	0.54	2.81	1.27	2.75	0.71	2.70	1.09	2.72	1.10

Figure-8: Segment Profiles

Based on the segment profiles, we can describe each cluster and suggest names reflecting their characteristics:

Cluster 1: High Earners with Balanced Satisfaction
This segment has the highest Monthly Income (mean of 17,280) and moderate levels across Business Travel (mean of 1.05), Training Times (mean of 2.73), Work-Life Balance, Relationship Satisfaction, and Environment Satisfaction. Characteristics: Employees in this cluster generally have high incomes and balanced satisfaction with their work-life and environment, indicating they may value financial compensation and have moderate job satisfaction.

Cluster 2: Mid-Level Earners with Strong Development Focus
This group has a Monthly Income around 8,901, slightly higher Training Times Last Year (mean of 2.82), and average levels of Business Travel, Work-Life Balance, and Relationship Satisfaction. Characteristics: These employees are mid-level earners with a focus on professional development. They have a balanced view on job satisfaction aspects, indicating a strong interest in career growth and development opportunities.

Cluster 3: Lower Earners with Moderate Satisfaction
This segment has the lowest Monthly Income (mean of 3,671) and similar levels of Business Travel, Training Times, Work-Life Balance, and Relationship Satisfaction to the other segments. Characteristics: Employees in this cluster earn less and maintain average satisfaction in work-life and

relationships, suggesting they might be in entry-level or less compensated positions but have relatively stable satisfaction levels.
Each cluster reflects a distinct group within the organization, differing primarily in income levels and slightly in job satisfaction dimensions.

```
/* Segmentation using k-means clustering */
proc fastclus data=HRDATA_Transformed maxclusters=3 out=SegmentedData;
   var
         MonthlyIncome BusinessTravel_Num
         TrainingTimesLastYear WorkLifeBalance
         RelationshipSatisfaction
         EnvironmentSatisfaction;
run;

proc tabulate data=SegmentedData;
   class Cluster;
   var
         MonthlyIncome BusinessTravel_Num TrainingTimesLastYear
         WorkLifeBalance RelationshipSatisfaction
         EnvironmentSatisfaction
         ;
   table Cluster,
       ( MonthlyIncome
         BusinessTravel_Num TrainingTimesLastYear
         WorkLifeBalance RelationshipSatisfaction
         EnvironmentSatisfaction ) * (mean std);
   title "Segment Profile by Cluster - Mean and Standard Deviation";
run;
```

4.1 Segment Analysis using SAS Viya

- **Monthly Income by Cluster**
  This bar chart displays the average monthly income across different clusters. Cluster 1 has the highest average income, followed by Cluster 2, with Cluster 3 having the lowest. This suggests income variations are significant factors in distinguishing these segments.

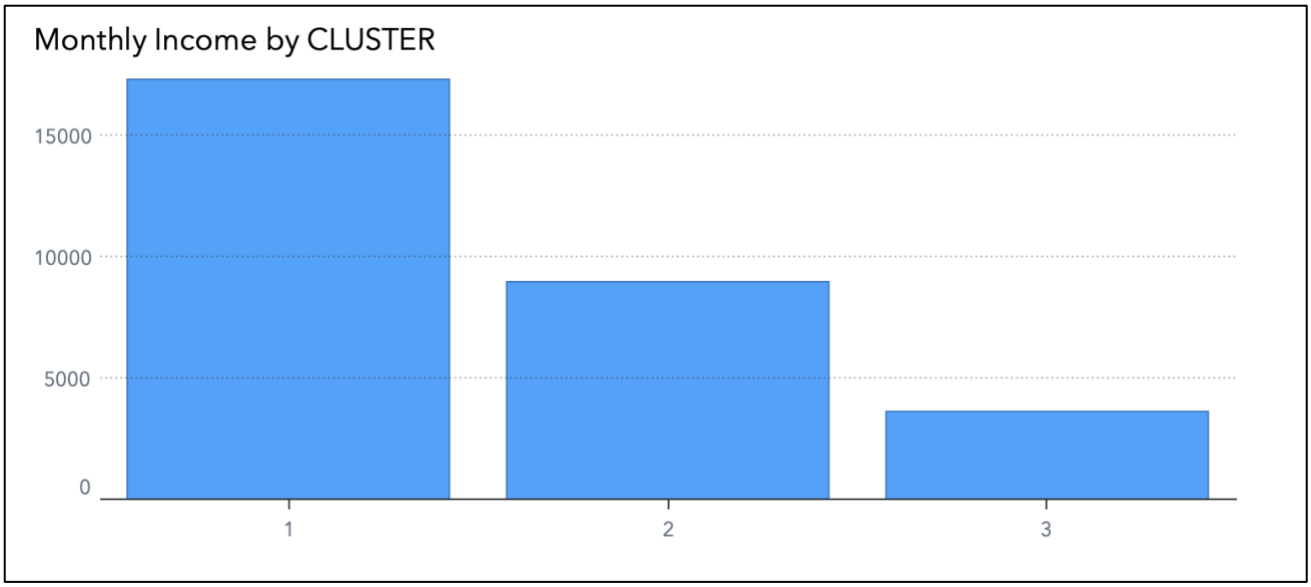


Figure-9: Segment wise Monthly income

- **Age Distribution by Cluster**
  The gauge charts illustrate the age distribution for each cluster, with Cluster 1 having the oldest average age, Cluster 2 a mid-range age, and Cluster 3 the youngest. This age variation may correlate with experience and career stage, impacting attrition risk.

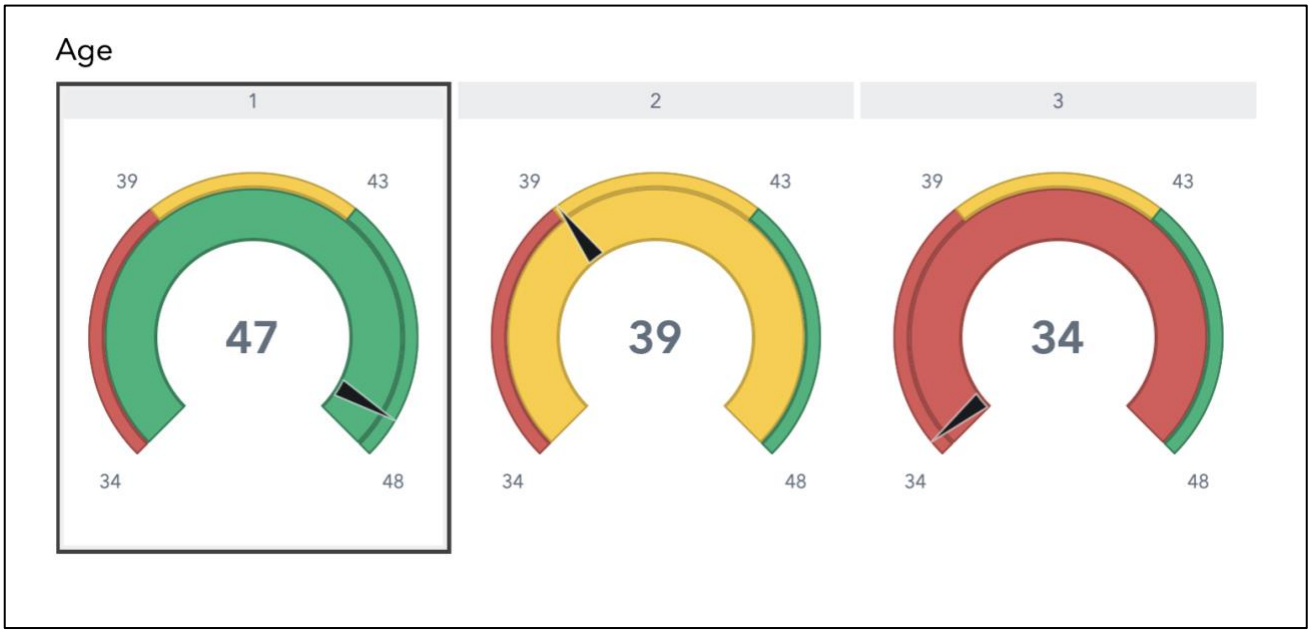


Figure-10: Segment wise Age

- **Monthly Income by Business Travel Frequency and Cluster**
  This chart explores monthly income variations across clusters based on business travel frequency. Cluster 1 employees, who are high earners, exhibit the highest income across all travel categories, with frequent travellers earning more. This indicates that frequent business travel may be associated with higher compensation.

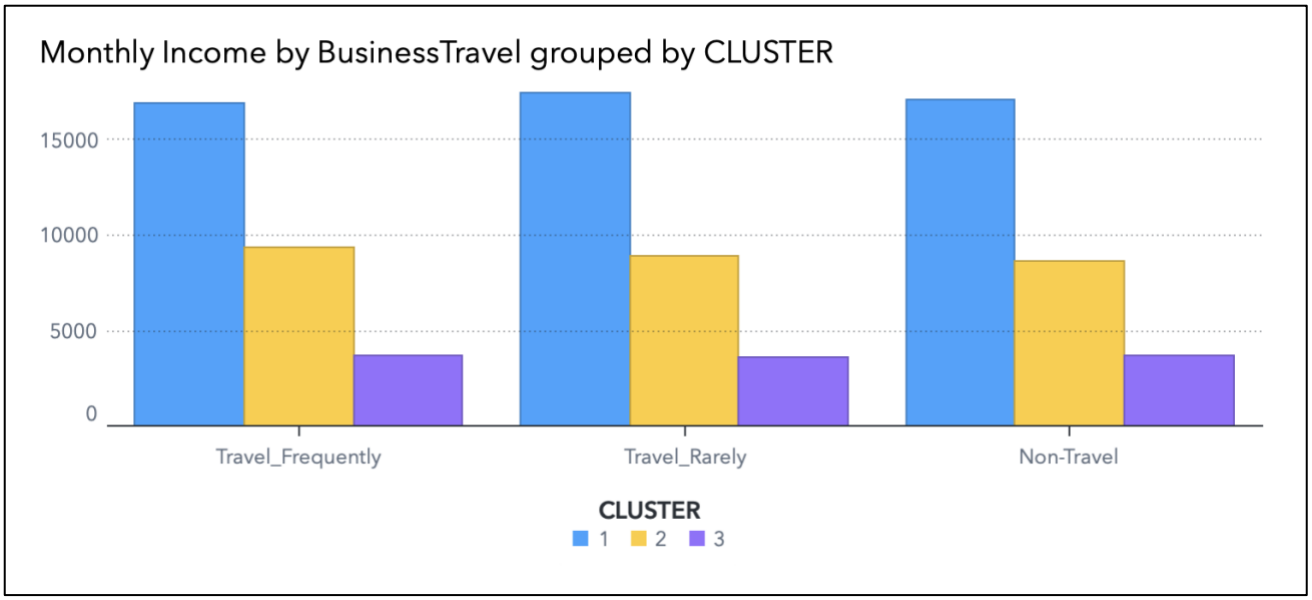


Figure-11: Segment wise Business Travel

- **Distance From Home by Over Time and Attrition**
  This bar chart examines the relationship between distance from home and overtime status, segmented by attrition. Employees who work overtime and experience attrition tend to live farther from work than those who do not, indicating that distance may contribute to turnover when combined with overtime.

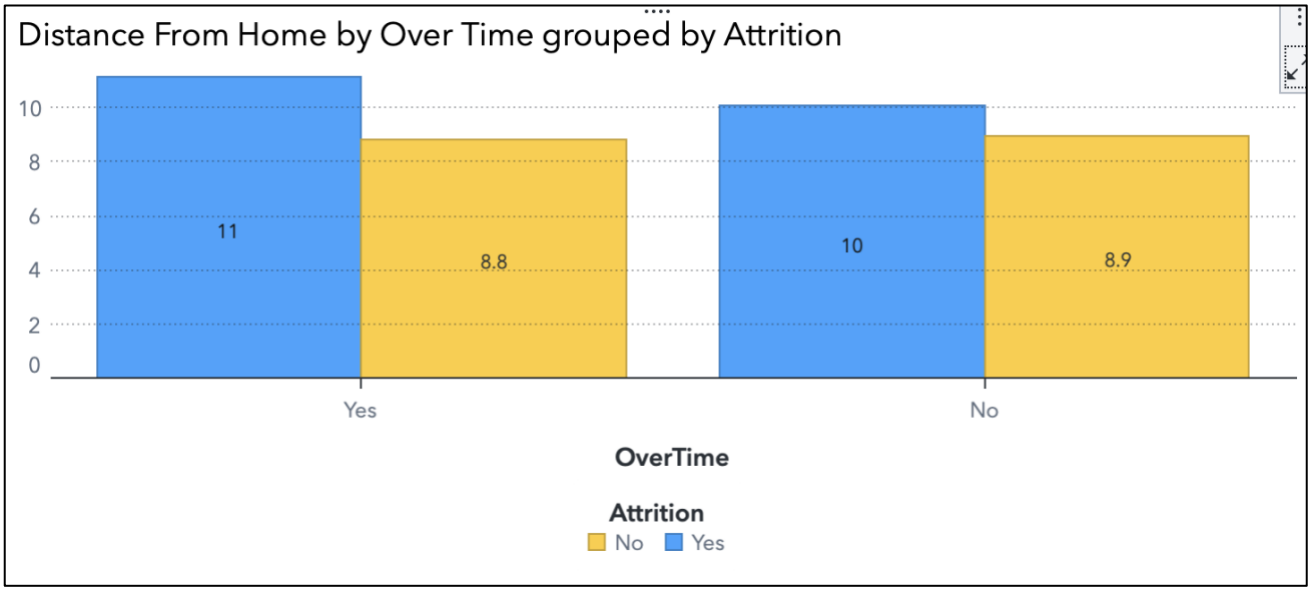


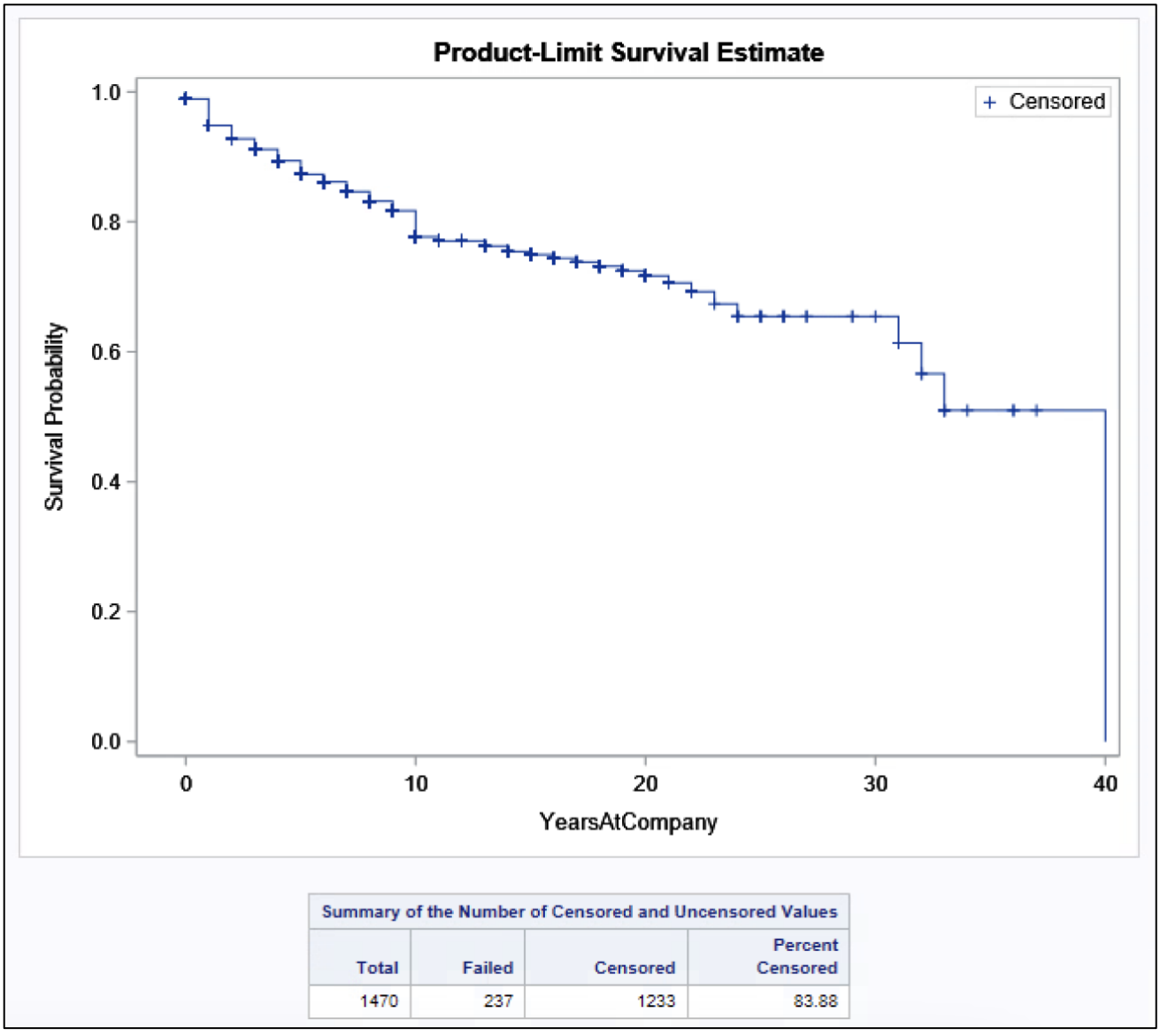
Figure-12: Distance From Home by Over Time and Attrition

5 Survival Analysis & Interpretation

5.1 Rationale for Selecting Variables

The variables included in the baseline Cox Proportional Hazards Model were chosen to align with the research objectives, focusing on factors that influence employee attrition. The study aims to examine the impact of demographic aspects (like age and gender), job-related factors (such as department affiliation and frequency of overtime), and stock options - on attrition rates. This selection was guided by questions about how different departments affect attrition, the influence of stock options on employee retention, and how these levels correlate with leaving intentions. Furthermore, job satisfaction and work-life balance were incorporated to capture the broader impact of workplace experience on employee attrition.

5.2 Overall Survival Curve



Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
1470	237	1233	83.88

Figure-13: Distance From Home by Over Time and Attrition

The Kaplan-Meier survival curve illustrates employee retention over time, showing the probability of employees remaining with the company as their tenure increases. The curve gradually declines, indicating decreasing retention rates as employees spend more years at the company. Each step down represents attrition events, while "+" markers signify censored observations where employees were still with the company at the time of analysis. This curve provides an overall view of retention trends, highlighting key points where attrition may increase as tenure progresses.

The table below the graph shows the total number of observations (1470 employees), with 237 employees having "failed" (left the company) and 1233 remaining censored (still employed). The high percentage of censored data (83.88%) suggests that most employees stayed with the company over the observed period.

5.3 Baseline model

The baseline model examines the influence of various independent variables on employee attrition without including interaction effects. The categorical variables (Gender_Num, BusinessTravel_Num, OverTime_Num, and JobRole) were incorporated with reference categories, allowing for the comparison of hazard ratios across different groups. Continuous variables such as Age, MonthlyIncome, and DistanceFromHome were also included to assess their direct effects on attrition risk.

```
/* Baseline Cox Proportional Hazards Model */
proc phreg data=SegmentedData;
   /* Categorical Variables */
   class cluster(ref='2') Gender_Num(ref='0') BusinessTravel_Num(ref='0')
OverTime_Num(ref='0') JobRole(ref='Sales Executive')
/ param=ref;

/* Model */
   model YearsAtCompany*Attrition_Num(0) =
         Gender_Num Age  NumCompaniesWorked
         OverTime_Num StockOptionLevel Dept_Sales
         DistanceFromHome      JobSatisfaction YearsSinceLastPromotion
                  cluster/ ties=efron;

   title 'Baseline Cox Proportional Hazards Model for Employee Attrition';
run;
```

5.3.1 Interpretation of Baseline Model Output

Each variable in the baseline Cox Proportional Hazards Model contributes uniquely to understanding employee attrition risk. Below is an interpretation of the key variables, including their hazard ratios and implications:

Key Findings from the Output:

- **Gender**: Parameter Estimate = 0.27111 with a Hazard Ratio of 1.311.
  - This suggests that males (Gender_Num = 1) have a 31% higher risk of leaving the company compared to females. The result is statistically significant (p-value = 0.0480).
- **Age**: Parameter Estimate = -0.07925 with a Hazard Ratio of 0.924.
  - As age increases, the risk of attrition decreases (roughly a 7.6% lower risk for every additional year of age). This is a strong, statistically significant result (p-value < 0.0001).
- **Number of Companies Worked**: Parameter Estimate = 0.16983 with a Hazard Ratio of 1.185.
  - Employees who have worked at more companies tend to have a higher risk of leaving. Each additional company worked corresponds to an 18.5% increase in attrition risk (p-value < 0.0001).
- **OverTime**: Parameter Estimate = 1.28909 with a Hazard Ratio of 3.629.
  - Employees who work overtime have a substantially higher risk of attrition (a 262.9% increase in risk). This is one of the strongest predictors in model and is highly significant (p-value < 0.0001).
- **Stock Option Level**: Parameter Estimate = -0.44045 with a Hazard Ratio of 0.644.
  - Higher stock option levels are associated with lower attrition risk. Employees with more stock options are 35.6% less likely to leave the company (p-value < 0.0001).
- **Sales Department**: Parameter Estimate = 0.48477 with a Hazard Ratio of 1.620.
  - Employees in the Sales department have a 62.0% higher risk of leaving the company compared to other departments, indicating this department may have retention issues (p-value = 0.0007).
- **Distance from Home**: Parameter Estimate = 0.02959 with a Hazard Ratio of 1.030.
  - A slight positive association exists between the distance employees live from work and attrition risk (3% higher risk for each unit increase in distance). This is significant (p-value = 0.0002).
- **Job Satisfaction**: Parameter Estimate = -0.25398 with a Hazard Ratio of 0.776.
  - Lower job satisfaction increases the likelihood of attrition. Employees with higher job satisfaction are 22.4% less likely to leave (p-value < 0.0001).
- **Years Since Last Promotion**: Parameter Estimate = -0.10435 with a Hazard Ratio of 0.901.
  - For each additional year since the last promotion, the attrition risk decreases slightly by 9.9% (p-value < 0.0001). This could be because employees who stay longer without a promotion might have reached a level of job stability and satisfaction with their current role, responsibilities.
- **Cluster 1 and Cluster 3**: Cluster 1 has a Hazard Ratio of 0.208, meaning employees in Cluster 1 who are High Earners with Balanced Satisfaction are significantly less likely to leave the company compared to those who are Mid-Level Earners with Strong Development Focus. Cluster 3 who are Lower Earners with Moderate Satisfaction has a Hazard Ratio of 1.882, meaning employees in this cluster have a higher likelihood of leaving compared to Cluster 2 (Mid-Level Earners with Strong Development Focus).

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
Gender_Num	1	1	0.27111	0.13710	3.9104	0.0480	1.311	Gender_Num 1
Age		1	-0.07925	0.01078	54.0748	<.0001	0.924	
NumCompaniesWorked		1	0.16983	0.02468	47.3557	<.0001	1.185	
OverTime_Num	1	1	1.28909	0.13319	93.6703	<.0001	3.629	OverTime_Num 1
StockOptionLevel		1	-0.44045	0.09105	23.4033	<.0001	0.644	
Dept_Sales		1	0.46477	0.13634	11.6205	0.0007	1.592	
DistanceFromHome		1	0.02959	0.00784	14.2460	0.0002	1.030	
JobSatisfaction		1	-0.25396	0.05854	18.8185	<.0001	0.776	
YearsSinceLastPromot		1	-0.10435	0.02557	16.6587	<.0001	0.901	
CLUSTER	1	1	-1.56527	0.51050	9.4013	0.0022	0.209	Cluster 1
CLUSTER	3	1	0.62157	0.17653	12.3980	0.0004	1.862	Cluster 3

Figure-14: Cox Proportional Hazards Model

5.3.2 Testing Global Null Hypothesis

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	418.7337	13	<.0001
Score	386.4248	13	<.0001
Wald	333.3034	13	<.0001

Figure-15: Global Null Hypothesis

Likelihood Ratio Test, Score Test, and Wald Test: All three tests show highly significant p-values ($p < 0.0001$), indicating that the overall model is statistically significant. This suggests that the set of predictors included significantly improves the model compared to a model with no predictors.

5.4 Enhanced Model

The enhanced Cox Proportional Hazards Model delves into employee attrition by incorporating key interactions and demographic adjustments. This model explores how factors like business travel and overtime, as well as stock options in relation to age, influence attrition risks. By including interactions between age and job satisfaction and integrating *Years Since Last Promotion* across different employee clusters, the model captures the nuanced effects of job dynamics on retention.

```
/* Interaction Cox Proportional Hazards Model */
proc phreg data=SegmentedData;
   /* Categorical Variables */
   class MaritalStatus_num(ref='2') cluster(ref='2') Gender_Num(ref='0')
         BusinessTravel_Num(ref='0') OverTime_Num(ref='0')
         JobRole(ref='Sales Executive') / param=ref;

   /* Model Statement */
   model YearsAtCompany*Attrition_Num(0) =
         BusinessTravel_Num*OverTime_Num age StockOptionLevel StockOptionLevel*age
         age*jobsatisfaction NumCompaniesWorked
         Dept_Sales YearsSinceLastPromotion*cluster
         / ties=efron;

title 'Baseline Cox Proportional Hazards Model for Employee Attrition';
run;
```

5.4.1 Interpretation of Model Output

- **Business Travel and OverTime Interaction**: BusinessTravel_Num 1 * OverTime_Num 1: The hazard ratio is high, indicating that employees who travel occasionally and work overtime face significantly increased attrition risk. BusinessTravel_Num 2 * OverTime_Num 1: This has an even higher hazard ratio, suggesting that frequent travellers who work overtime are at the greatest risk of attrition.

- **Age**: **Parameter Estimate = -0.10125**: The negative coefficient indicates that as employees age, their risk of attrition decreases. Older employees may seek stability more than younger ones, thus having lower attrition risk.

- **Stock Option Level**: Parameter Estimate = -1.30623: A negative estimate indicates that employees with higher stock options tend to have lower attrition risk, possibly because stock options act as incentives for retention.

- **Age * Stock Option Level Interaction**: Parameter Estimate = 0.02595: This suggests that the protective effect of stock options on retention weakens slightly with age. Younger employees might perceive stock options as more valuable for future growth compared to older employees.

- **Age * Job Satisfaction Interaction**: Parameter Estimate = -0.00619: The negative coefficient indicates that for older employees, higher job satisfaction correlates with a slightly lower attrition risk, reinforcing the value of a fulfilling job as employees age.

- **Number of Companies Worked For**: Hazard Ratio = 1.182: The positive coefficient and hazard ratio greater than 1 imply that employees with a history of job-hopping are more likely to leave, which aligns with expectations.

- **Department (Sales)**: Hazard Ratio = 1.458**:** Sales employees have a higher risk of attrition than those in other departments, potentially due to high pressure and targets which are typical in sales roles.

- **Years Since Last Promotion * Cluster Interaction**:

  - Cluster 1 (Estimate = -0.13820)**:** Suggests that in this cluster, as the years since the last promotion increase, the risk of attrition decreases within this cluster. This could mean that in Cluster 1, employees who have not received recent promotions are somewhat more likely to stay, which might seem counterintuitive at first. It suggests that these employees might not prioritize promotions as much for their job satisfaction or retention. Instead, they could be motivated by other factors within their cluster, such as work-life balance or relationship satisfaction.
  - Cluster 3 (Estimate = -0.08008)**:** The negative estimate indicates that as the years since the last promotion increase, the risk of attrition decreases, but this effect is weaker compared to Cluster 1. Since Cluster 3 comprises younger individuals, it suggests that these employees may be more tolerant of longer intervals between promotions, or they may wait to receive a promotion and switch jobs.

Analysis of Maximum Likelihood Estimates

Parameter			DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
BusinessT*OverTime_N	1	1	1	1.19032	0.14516	67.2417	<.0001	.	BusinessTravel_Num 1 * OverTime_Num 1
BusinessT*OverTime_N	2	1	1	1.51603	0.19404	61.0423	<.0001	.	BusinessTravel_Num 2 * OverTime_Num 1
Age			1	-0.10125	0.01281	62.4997	<.0001	.	
StockOptionLevel			1	-1.30623	0.41099	10.1014	0.0015	.	
Age*StockOptionLevel			1	0.02595	0.01143	5.1557	0.0232	.	Age * StockOptionLevel
Age*JobSatisfaction			1	-0.00619	0.00170	13.2765	0.0003	.	Age * JobSatisfaction
NumCompaniesWorked			1	0.16713	0.02474	45.6563	<.0001	1.182	
Dept_Sales			1	0.37689	0.13508	7.7846	0.0053	1.458	
YearsSinceLa*CLUSTER	1		1	-0.13820	0.04807	8.2657	0.0040	.	Cluster 1 * YearsSinceLastPromotion
YearsSinceLa*CLUSTER	3		1	-0.08006	0.03316	5.8294	0.0158	.	Cluster 3 * YearsSinceLastPromotion

Figure-16: Cox Proportional Hazards Model

5.5 Kaplan Meir Curve on Segments

The Kaplan-Meier plot illustrates the survival probability over time for employees, stratified by clusters. Each line represents a different cluster, showing how the likelihood of employees staying with the company decreases over the years. The vertical steps represent observed attrition events, while the "+" symbols indicate censored data, where the employee's time at the company ended without attrition.

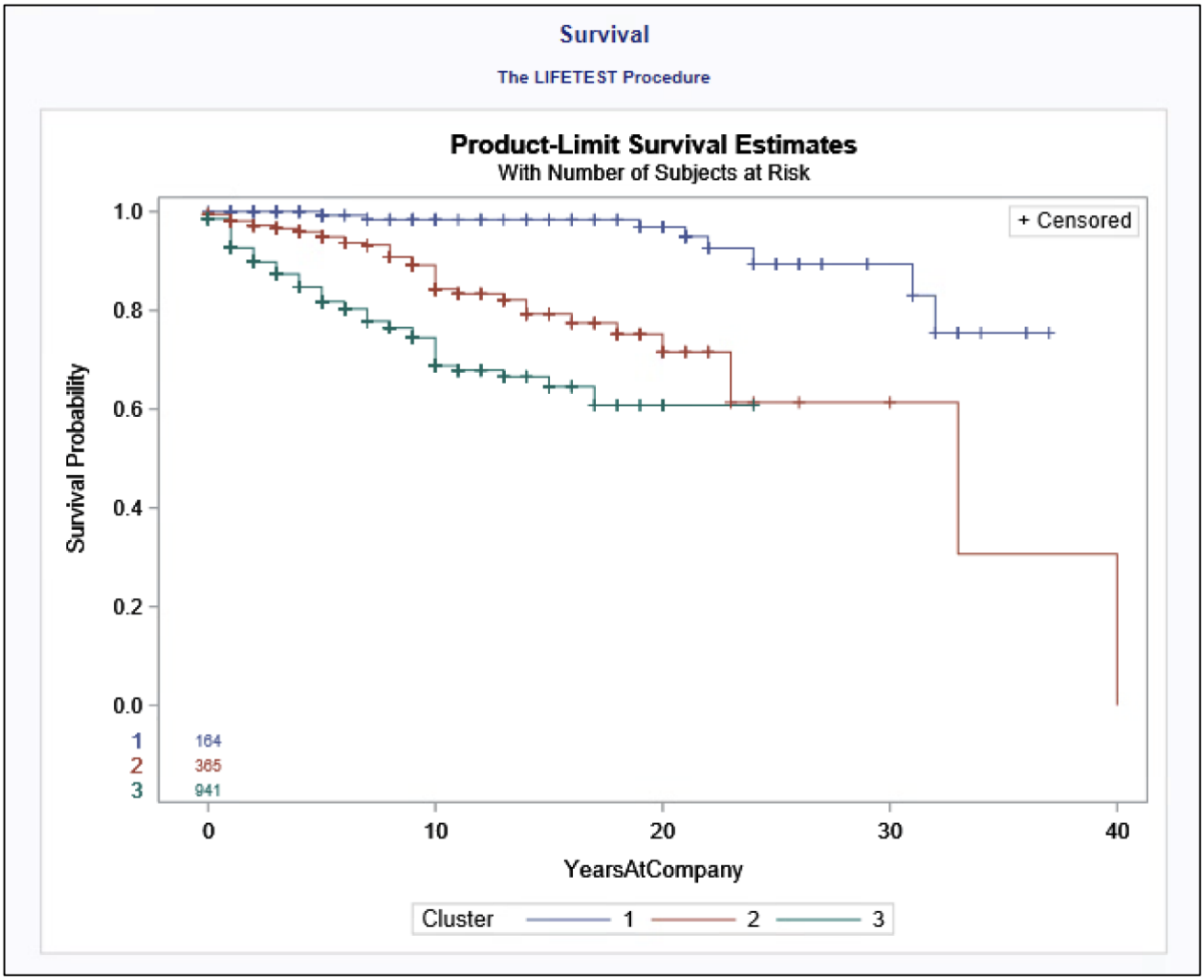


Figure-17: Kaplan-Meier plot

The survival curves reveal differences in attrition risk across clusters:

- **Cluster 3** (in green) has the lowest survival probability, indicating a higher attrition risk over time.
- **Cluster 2** (in red) shows a moderate attrition risk.
- **Cluster 1** (in blue) demonstrates the highest survival probability, suggesting that employees in this cluster are the least likely to leave.

```
/* Kaplan Meier Estimation */
ods       graphics on;
proc lifetest data=SegmentedData plots=survival (atrisk=0);
   time YearsAtCompany*Attrition_Num(0);
   strata cluster; /* Example of stratifying by employed status, you can stratify by other
variables */
           ods select survivalplot;
           title " Survival";
run;
```

6 Summary of Findings

This study aimed to explore factors influencing employee attrition, leveraging Cox Proportional Hazards models and Kaplan-Meier survival analysis. Key variables like age, job satisfaction, stock option levels, and overtime were analysed to understand their impact on attrition risk, with segmentation providing additional insights into distinct employee groups.

1. **Frequent Travel and Overtime:** The combination of frequent business travel and overtime was a significant predictor of higher attrition risk. Employees facing both conditions displayed the steepest decline in survival probability, especially those in Clusters 1 and 2.
2. **Stock Options and Age:** Stock options were a strong retention factor, particularly for younger employees, as indicated by the Age*StockOption interaction. Younger employees appear to be more influenced by stock options in their decision to stay or leave, underscoring the importance of financial incentives for younger demographics.
3. **Departmental Impact:** Employees in the Sales department demonstrated a higher risk of attrition, suggesting that role-specific challenges or expectations may drive turnover rates. Addressing these department-specific factors could potentially improve retention within sales teams.
4. **Job Satisfaction and Promotion Timing:** Higher job satisfaction correlated with a reduced risk of attrition, while more years since the last promotion increased attrition risk, particularly in Cluster 3 (typically younger employees). This trend highlights the need for timely promotions to retain younger talent.
5. **Cluster Analysis:** The segmentation revealed distinct clusters with different attrition risks. Cluster 1, "High Earners with Balanced Satisfaction," exhibited higher attrition risks associated with promotions, while Cluster 3, "Lower Earners with Moderate Satisfaction," demonstrated a unique sensitivity to job role and satisfaction dynamics.

6.1.1 Recommendations

Based on these findings, the following recommendations could help reduce attrition and improve employee retention:

1. **Implement Flexible Work and Travel Policies:** Since employees with frequent travel and overtime obligations face higher attrition risk, introducing flexible work arrangements and limiting mandatory overtime for those who travel frequently could enhance retention.
2. **Enhance Financial Incentives for Younger Employees:** Given the strong impact of stock options on younger employees, consider expanding financial incentives like stock options and bonuses to this demographic to increase their sense of financial security and job satisfaction.

3. **Address Departmental Challenges in Sales:** With sales roles exhibiting higher attrition, it's essential to conduct targeted employee satisfaction surveys and identify specific pain points. Tailored interventions, such as role redesigns or workload adjustments, could help retain talent within this department.
4. **Develop a Transparent Promotion Strategy:** Given that time since the last promotion is a predictor of attrition, implementing a clear, merit-based promotion pathway could reduce turnover, especially for younger employees who may seek faster career progression.
5. **Tailored Retention Programs for Each Cluster:** Customize retention strategies based on the distinct needs and risks associated with each cluster. For example, focusing on financial incentives for Cluster 3 and job satisfaction enhancements for Cluster 1 could be more effective than one-size-fits-all solutions.

By targeting these areas, the organization can take meaningful steps to improve employee retention and reduce the overall attrition rate, ultimately fostering a more stable and motivated workforce.

7 Conclusion

This study identified key factors affecting employee attrition, with overtime, frequent business travel, and stock options playing significant roles. The enhanced model showed that employees who travel often and work overtime face higher attrition risks, while stock options help retain younger employees. Segment analysis revealed that promotion frequency and job role also impact attrition differently across employee groups. Kaplan-Meier analysis further illustrated varying survival rates by cluster. To reduce attrition, recommendations include offering targeted support for high-risk groups, increasing promotion opportunities, and enhancing stock option availability to align with the needs of specific employee segments.

8 Limitations and Future Recommendations

One key limitation of this analysis is the absence of specific time-stamped data in the IBM HR Analytics dataset. The dataset only provides "YearsAtCompany" as a measure of employee tenure, without precise information on hire dates or termination dates. In survival analysis, exact timing of events is crucial for accurately estimating time-to-event metrics. In this case, the model assumes a consistent measurement point for all employees, which may limit the precision of the survival curves and hazard ratios.

To address this limitation, the analysis focuses on understanding the relative influence of different factors, such as overtime, stock options, and job satisfaction, on employee attrition risk, rather than providing precise time-to-event predictions. This approach offers valuable insights into the conditions associated with higher or lower attrition risk.

For future studies, in real world datasets, it is recommended the inclusion of more detailed, time-stamped HR data, such as hire dates, promotion dates, and exit dates. This would allow for a more accurate survival analysis, providing a clearer picture of employee attrition patterns over time.

9 Bibliography

Allen, D. G., Bryant, P. C., & Vardaman, J. M. (2010). Retaining talent: Replacing misconceptions with evidence-based strategies. *Academy of Management Perspectives, 24*(2), 48-64.

Hausknecht, J. P., & Holwerda, J. A. (2013). When does employee turnover matter? Dynamic member configurations, productive capacity, and collective performance. *Academy of Management Journal, 56*(5), 1254-1274.

Shaw, J. D., Gupta, N., & Delery, J. E. (2005). Alternative conceptualizations of the relationship between voluntary turnover and organizational performance. *Academy of Management Journal, 48*(1), 50-68.

Trevor, C. O., & Nyberg, A. J. (2008). Keeping your headcount when all about you are losing theirs: Downsizing, voluntary turnover rates, and the moderating role of HR practices. *Academy of Management Journal, 51*(2), 259-276.

Allison, P. D. (2010). *Survival analysis using SAS: A practical guide*. Cary, NC: SAS Institute.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187-220.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time-to-event data*. John Wiley & Sons.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457-481.

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations*. Kluwer Academic Publishers.

Nyberg, A. J., & Ployhart, R. E. (2013). Context-emergent turnover (CET) theory: A theory of collective turnover. *Academy of Management Review, 38*(1), 109-131.

10 Appendix

10.1 Complete SAS Code

```
* Importing data for this assignment;
proc import datafile= "Z:\Downloads\CaseStudy_HR.csv"
    dbms=csv
    out=work.HRDATA
    replace;
run;

/* Contents of imported data */
proc contents data=HRDATA;
run;


/*Transforming Categorical variables*/
data HRDATA_Transformed;
    set HRDATA;

    /* Binary Coding */
    if Attrition = 'Yes' then Attrition_Num = 1; else Attrition_Num = 0;
    if OverTime = 'Yes' then OverTime_Num = 1; else OverTime_Num = 0;

    /* Ordinal Coding */
    if BusinessTravel = 'Non-Travel' then BusinessTravel_Num = 0;
    else if BusinessTravel = 'Travel_Rarely' then BusinessTravel_Num = 1;
    else if BusinessTravel = 'Travel_Frequently' then BusinessTravel_Num = 2;

    if Gender = 'Male' then Gender_Num = 1; else Gender_Num = 0;

    if MaritalStatus = 'Single' then MaritalStatus_Num = 0;
    else if MaritalStatus = 'Divorced' then MaritalStatus_Num = 1;
    else if MaritalStatus = 'Married' then MaritalStatus_Num = 2;

            /* Create Dummy Variables for JobRole */
    if JobRole = 'Sales Executive' then JobRole_SalesExec = 1; else JobRole_SalesExec = 0;
    if JobRole = 'Research Scientist' then JobRole_ResearchSci = 1; else JobRole_ResearchSci = 0;
    if JobRole = 'Laboratory Technician' then JobRole_LabTech = 1; else JobRole_LabTech = 0;
    if JobRole = 'Manufacturing Director' then JobRole_ManufDir = 1; else JobRole_ManufDir = 0;
    if JobRole = 'Healthcare Representative' then JobRole_HealthRep = 1; else JobRole_HealthRep = 0;
    if JobRole = 'Manager' then JobRole_Manager = 1; else JobRole_Manager = 0;
    if JobRole = 'Sales Representative' then JobRole_SalesRep = 1; else JobRole_SalesRep = 0;
    if JobRole = 'Research Director' then JobRole_ResearchDir = 1; else JobRole_ResearchDir = 0;
    if JobRole = 'Human Resources' then JobRole_HR = 1; else JobRole_HR = 0;

                    /* Create Dummy Variables for Department */
    if Department = 'Sales' then Dept_Sales = 1; else Dept_Sales = 0;
            if Department = 'Research & Development' then Dept_RD = 1; else Dept_RD = 0;
    if Department = 'Human Resources' then Dept_HR = 1; else Dept_HR = 0;


run;
```

```
/* Contents imported data */
PROC CONTENTS data=HRDATA_Transformed;
run;


/*summary Stats

/* Summary Statistics */
proc means data=HRDATA_Transformed n nmiss mean std min max median;
var
        age
        Gender_Num
        MonthlyIncome
        YearsAtCompany
   BusinessTravel_Num
        OverTime_Num
        StockOptionLevel
        DistanceFromHome
        jobsatisfaction
        NumCompaniesWorked
   Dept_Sales
        YearsSinceLastPromotion
        BusinessTravel_Num
        TrainingTimesLastYear
        WorkLifeBalance
        RelationshipSatisfaction
        EnvironmentSatisfaction
; title "Summary Stats";
run;

/*Summary Statistics */



/* Sub Sample Analysis of the Dataset */
proc means data=HRDATA_Transformed n nmiss mean std min max median;
class Attrition_Num;
var
        age
        Gender_Num
        MonthlyIncome
        YearsAtCompany
   BusinessTravel_Num
        OverTime_Num
        StockOptionLevel
        DistanceFromHome
        jobsatisfaction
        NumCompaniesWorked
        YearsSinceLastPromotion
        BusinessTravel_Num
        TrainingTimesLastYear
        WorkLifeBalance
        RelationshipSatisfaction
        EnvironmentSatisfaction
;
```

```
title "Summary Statistics for Continuous Variables by Attrition Status";
run;


/*_____EDA_____________*/
/* Step 1: Calculate Total Count per Department */
proc sql;
   create table DeptTotal as
   select
      Department,
      count(*) as TotalCount
   from HRDATA_Transformed
   group by Department;
quit;

/* Step 2: Calculate Attrition Count and Join with Total Count to Get Percentage */
proc sql;
   create table DeptAttritionPct as
   select
      a.Department,
      a.Attrition_Num,
      count(*) as Count,
      (count(*) * 100.0 / b.TotalCount) as Attrition_Pct
   from HRDATA_Transformed as a
   left join DeptTotal as b
   on a.Department = b.Department
   group by a.Department, a.Attrition_Num;
quit;

proc sort data=DeptAttritionPct nodupkey;
   by Department Attrition_Num;
run;

proc sgplot data=DeptAttritionPct;
   vbar Department / response=Attrition_Pct group=Attrition_Num groupdisplay=cluster datalabel;
   xaxis label="Department";
   yaxis label="Attrition Percentage" grid;
   keylegend / title="Attrition Status";
   title "Percentage of Attrition by Department";
run;




proc sgplot data=HRDATA_Transformed;
   scatter x=Age y=MonthlyIncome / group=Attrition_Num markerattrs=(symbol=circlefilled);
   title "Age vs. Monthly Income by Attrition";
   xaxis label="Age";
   yaxis label="Monthly Income";
run;


/*Avetage Income by Age/Attrition
/* Step 1: Create Age Buckets */
data HRDATA_Bucketed;
```

```
    set HRDATA_Transformed;
    length Age_Bucket $10;
    if 20 <= Age < 25 then Age_Bucket = '20-25';
    else if 25 <= Age < 30 then Age_Bucket = '25-30';
    else if 30 <= Age < 35 then Age_Bucket = '30-35';
    else if 35 <= Age < 40 then Age_Bucket = '35-40';
    else if 40 <= Age < 45 then Age_Bucket = '40-45';
    else if 45 <= Age < 50 then Age_Bucket = '45-50';
    else if 50 <= Age < 55 then Age_Bucket = '50-55';
    else if 55 <= Age <= 60 then Age_Bucket = '55-60';
run;

/* Step 2: Calculate Average Monthly Income by Age Bucket and Attrition */
proc means data=HRDATA_Bucketed noprint;
    class Age_Bucket Attrition_Num;
    var MonthlyIncome;
    output out=AvgIncomeByBucket mean=Avg_MonthlyIncome;
run;

/* Step 3: Round Average Monthly Income */
data AvgIncomeByBucket;
    set AvgIncomeByBucket;
    Avg_MonthlyIncome = round(Avg_MonthlyIncome, 1); /* Round to whole number */
run;

/* Step 4: Plot Average Monthly Income by Age Bucket and Attrition */
proc sgplot data=AvgIncomeByBucket;
    vbar Age_Bucket / response=Avg_MonthlyIncome group=Attrition_Num groupdisplay=cluster datalabel;
    xaxis label="Age Bucket";
    yaxis label="Average Monthly Income";
    title "Average Monthly Income by Age Bucket and Attrition";
    keylegend / title="Attrition Status";
run;




proc sgplot data=HRDATA_Transformed;
    vbox MonthlyIncome / category=Attrition_Num;
    title "Monthly Income by Attrition Status";
    xaxis label="Attrition Status (0 = Stayed, 1 = Left)";
    yaxis label="Monthly Income";
run;



proc sgplot data=HRDATA_Transformed;
    histogram Age / group=Attrition_Num transparency=0.5;
    density Age / group=Attrition_Num type=kernel;
    title "Age Distribution by Attrition Status";
    xaxis label="Age";
    yaxis label="Frequency";
run;
```

```
proc sgplot data=HRDATA_Transformed;
   vbar BusinessTravel_Num / group=Attrition_Num groupdisplay=cluster;
   title "Attrition by Business Travel Frequency";
   xaxis label="Business Travel Frequency";
   yaxis label="Count";
run;

/* Step 1: Create Age Buckets */
data HRDATA_Bucketed;
   set HRDATA_Transformed;
   length Age_Bucket $10;
   if 20 <= Age < 25 then Age_Bucket = '<25';
   else if 25 <= Age < 30 then Age_Bucket = '25-30';
   else if 30 <= Age < 35 then Age_Bucket = '30-35';
   else if 35 <= Age < 40 then Age_Bucket = '35-40';
   else if 40 <= Age < 45 then Age_Bucket = '40-45';
   else if 45 <= Age < 50 then Age_Bucket = '45-50';
   else if 50 <= Age < 55 then Age_Bucket = '50-55';
   else  Age_Bucket = '>60';
run;

/* Step 2: Calculate Average Job Satisfaction by Age Bucket and Attrition */
proc means data=HRDATA_Bucketed noprint;
   class Age_Bucket Attrition_Num;
   var JobSatisfaction;
   output out=AvgJobSatisfaction mean=Avg_JobSatisfaction;
run;

/* Step 3: Round Average Job Satisfaction to Two Decimal Places */
data AvgJobSatisfaction;
   set AvgJobSatisfaction;
   Avg_JobSatisfaction = round(Avg_JobSatisfaction, 0.01); /* Round to two decimal places */
run;

/* Step 4: Plot Average Job Satisfaction by Age Bucket and Attrition */
proc sgplot data=AvgJobSatisfaction;
   vbar Age_Bucket / response=Avg_JobSatisfaction group=Attrition_Num groupdisplay=cluster datalabel
datalabelattrs=(size=10);
   xaxis label="Age Bucket";
   yaxis label="Average Job Satisfaction" values=(1.5 to 4 by 0.2);
   title "Average Job Satisfaction by Age Bucket and Attrition Status";
   keylegend / title="Attrition Status";
run;

/*____Segmentation____*/


/* Segmentation using k-means clustering */
proc fastclus data=HRDATA_Transformed maxclusters=3 out=SegmentedData;
   var
         MonthlyIncome
         BusinessTravel_Num
         TrainingTimesLastYear
         WorkLifeBalance
         RelationshipSatisfaction
```

```
            EnvironmentSatisfaction
            ;
run;

proc tabulate data=SegmentedData;
   class Cluster;
   var
            MonthlyIncome
            BusinessTravel_Num
            TrainingTimesLastYear
            WorkLifeBalance
            RelationshipSatisfaction
            EnvironmentSatisfaction
            ;
   table Cluster,
         (
MonthlyIncome
            BusinessTravel_Num
            TrainingTimesLastYear
            WorkLifeBalance
            RelationshipSatisfaction
            EnvironmentSatisfaction) * (mean std);
   title "Segment Profile by Cluster - Mean and Standard Deviation";
run;

/* Viewing the contents of segmented data */
proc contents data=SegmentedData;
run;


proc export data=SegmentedData
   outfile="Z:\Downloads\SegmentedData.xlsx"
   dbms=xlsx
   replace;
run;


/*_______Survival Analysis________*/



data HRDATA_SURV;
   set HRDATA_Transformed;
   Event=Attrition_Num;
   Time = YearsAtCompany;
run;


proc lifetest data=HRDATA_SURV plots=(s);
   time Time * Event(0);
   strata Department; /* Optional: Stratify by Department or any other categorical variable */
run;
```

```
/* Baseline Cox Proportional Hazards Model */
proc phreg data=SegmentedData;
   /* Categorical Variables */
   class cluster(ref='2') Gender_Num(ref='0') BusinessTravel_Num(ref='0')
                          OverTime_Num(ref='0') JobRole(ref='Sales Executive') / param=ref;

   /* Model Statement */
   model YearsAtCompany*Attrition_Num(0) =
      Gender_Num Age
                  NumCompaniesWorked
                 OverTime_Num
                 StockOptionLevel
      Dept_Sales
      DistanceFromHome
                 JobSatisfaction
                 YearsSinceLastPromotion
                 cluster/ ties=efron;

   title 'Baseline Cox Proportional Hazards Model for Employee Attrition';
run;


/* Kaplan Meier Estimation */
ods      graphics on;
proc lifetest data=SegmentedData plots=survival (atrisk=0);
   time YearsAtCompany*Attrition_Num(0);
   strata cluster; /* Example of stratifying by employed status, you can stratify by other variables */
         ods select survivalplot;
         title " Survival";
run;


/* Kaplan Meier Estimation */
ods graphics on;

proc lifetest data=SegmentedData method=LT intervals=(0 TO 36 by 1) plots=(survival) outsurv=survdata;

         time YearsAtCompany*Attrition_Num(0);
         strata cluster;
run;

ods graphics on;
/*survdata*/
proc sgplot data=survdata;
         title 'Survival Probability Estimates';
         series x=YearsAtCompany y=Survival / group=cluster;
         YAXIS min=0.0;
run;



/* Interaction Cox Proportional Hazards Model */
proc phreg data=SegmentedData;
   /* Categorical Variables */
```

```
    class MaritalStatus_num(ref='2') cluster(ref='2') Gender_Num(ref='0')
            BusinessTravel_Num(ref='0') OverTime_Num(ref='0')
            JobRole(ref='Sales Executive') / param=ref;

    /* Model Statement */
    model YearsAtCompany*Attrition_Num(0) =
            BusinessTravel_Num*OverTime_Num age
            StockOptionLevel
            StockOptionLevel*age
            age*jobsatisfaction
            NumCompaniesWorked
    Dept_Sales
            YearsSinceLastPromotion*cluster
            / ties=efron;

    title 'Baseline Cox Proportional Hazards Model for Employee Attrition';
run;
```