EMPLOYEES ATTRITION OF IBM

group 1

112354015 統碩二 林子涵
112753134 資碩一 管漢程
110405026 廣電三 賴冠儒
110501035 英文三 晏煒翔
110306019 資管三 黃茂勛
110304039 統計三 邱士芳

Outline

一、資料集/目標介紹

二、資料視覺化

三、統計檢定

四、模型配適

五、結論

附錄、海報展演

一、資料集/目標介紹

研究動機與目的

- **IBM員工自願離職**

對一間公司而言，員工是否會穩定地在公司工作是很重要的。
公司會希望能掌握員工是否會繼續留在目前的工作崗位，以減少人資成本。
我們透過IBM HR對內部員工調查的資料集，想以是否離職為因變數，得出可以預測員工是否會離職的模型，並分析不同類型員工的特色，希望能從此得出相關之洞見。若能得出有效的結論，除了人資部門可以安排更穩定的人力而不用擔心人流缺失外，或許某些因素是公司能改善，可以減少離職率。

- **資料來源：Kaggle**

https://www.kaggle.com/code/mragpavank/ibm-hr-analytics-employee-attrition-performance/notebook

資料介紹

資料描述：

這個資料集包含了一家公司的員工資訊，用於分析員工離職（Attrition）情況。每一列代表一名員工，每一欄代表該員工的某些屬性。資料集共包含35個變數，1470筆數據

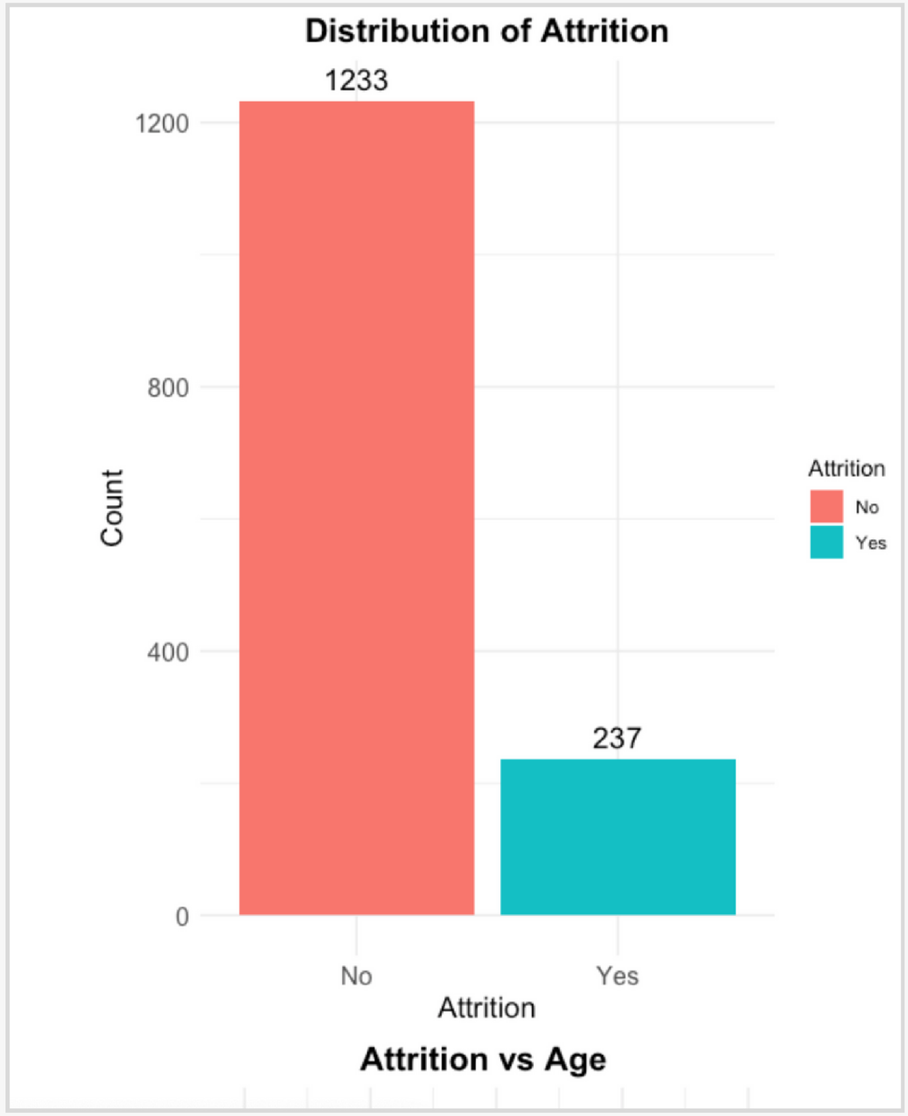
AGE	Attrition	Business Travel	DailyRate	Department
Distance From Home	Education	Education Field	Employee Count	Employee Number
Environment Satisfaction	Gender	HourlyRate	Job Involvement	JobLevel
JobRole	Job Satisfaction	Marital Status	Monthly Income	Monthly Rate
Num Companies Worked	Over18	OverTime	Percent Salary Hike	Performance Rating
Relationship Satisfaction	Standard Hours	StockOptionLevel	Total Working Years	Training Times LastYear
WorkLife Balance	Years At Company	Years In Current Role	Years SinceLast Promotion	Years With Curr Manager

資料概況一覽

# Age Numérica - Discreta	✓ Attrition Categórica	A BusinessTravel Categórica	# EmployeeCount Numérica - Discreta	# EmployeeNumber Numérica - Discreta	# EnvironmentSati... Categórica	A Gender Categórica
 18 60	true 0 0% false 0 0%	Travel_Rarely 71% Travel_Frequently 19% Other (150) 10%	 1 1	 1 2068	 1 4	Male Female
41	Yes	Travel_Rarely	1	1	2	Female
49	No	Travel_Frequently	1	2	3	Male
37	Yes	Travel_Rarely	1	4	4	Male
33	No	Travel_Frequently	1	5	4	Female
27	No	Travel_Rarely	1	7	1	Male

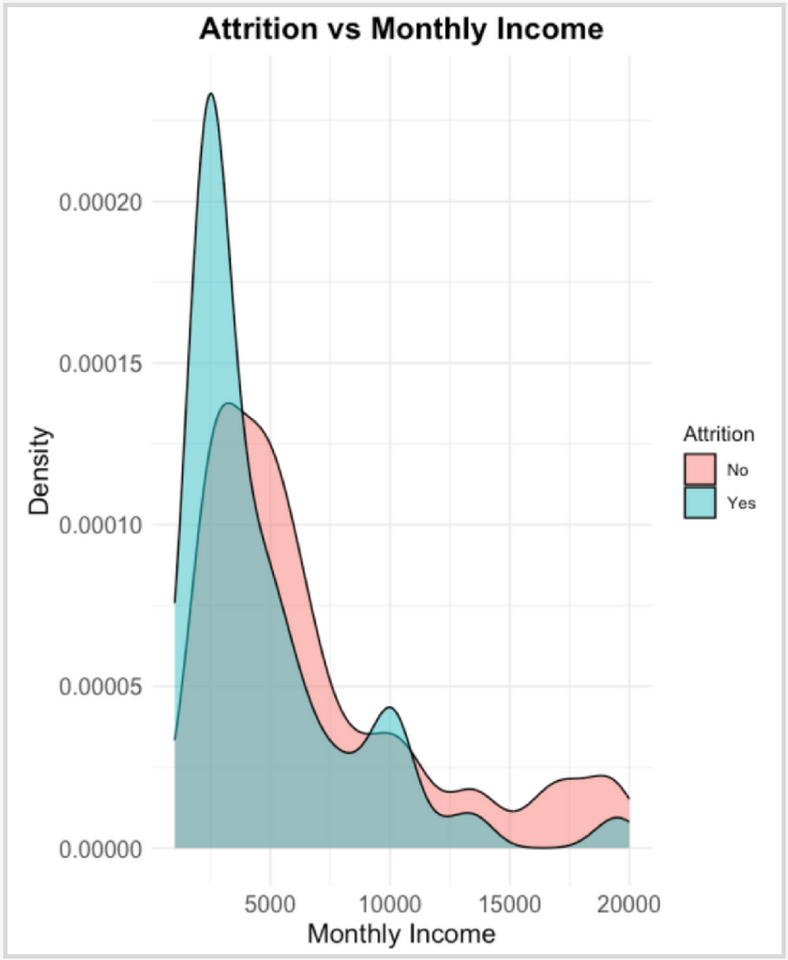
二、資料視覺化

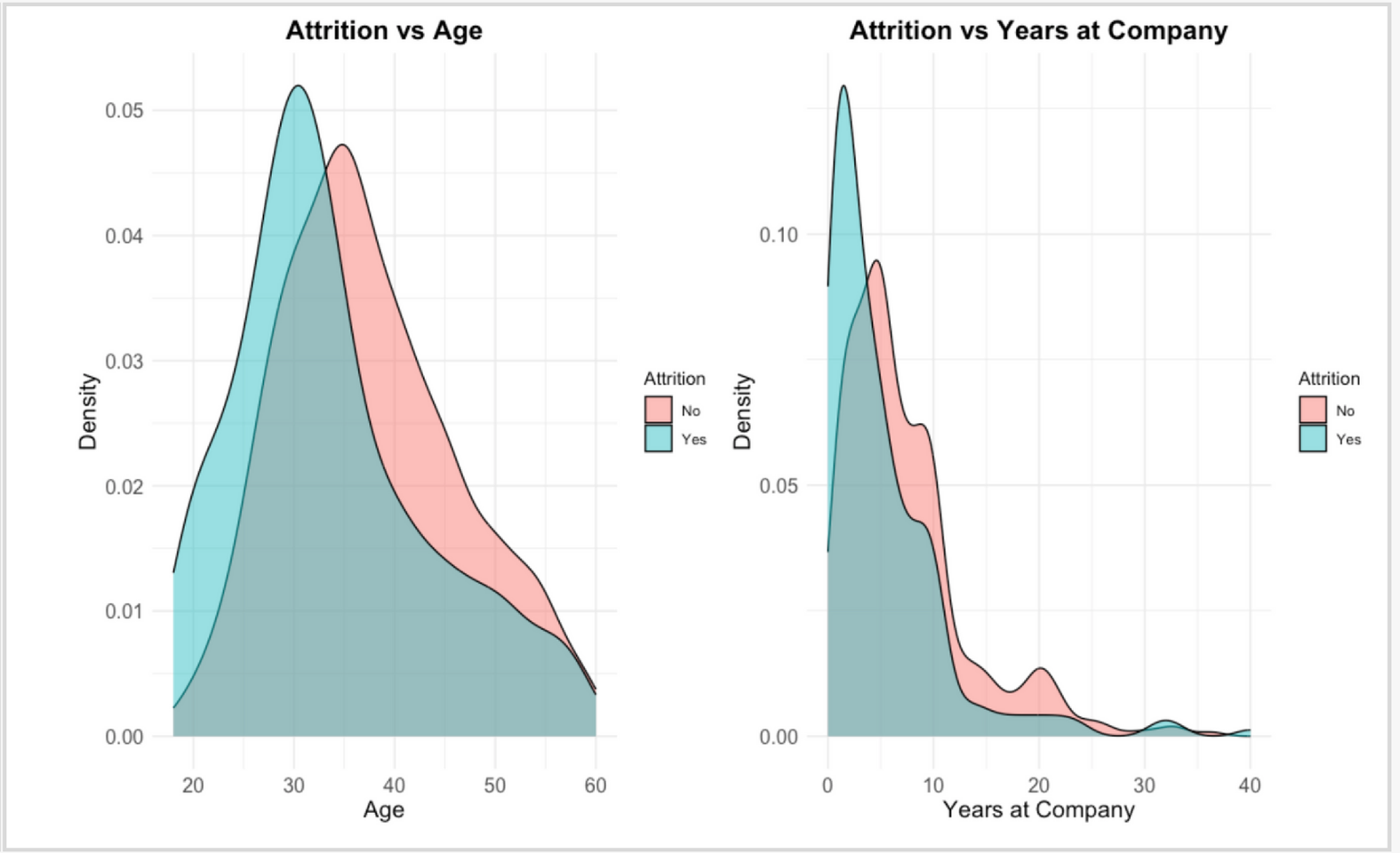
EDA　Histogram



- 大多數員工沒有離職，「否」的人數明顯高於「是」。

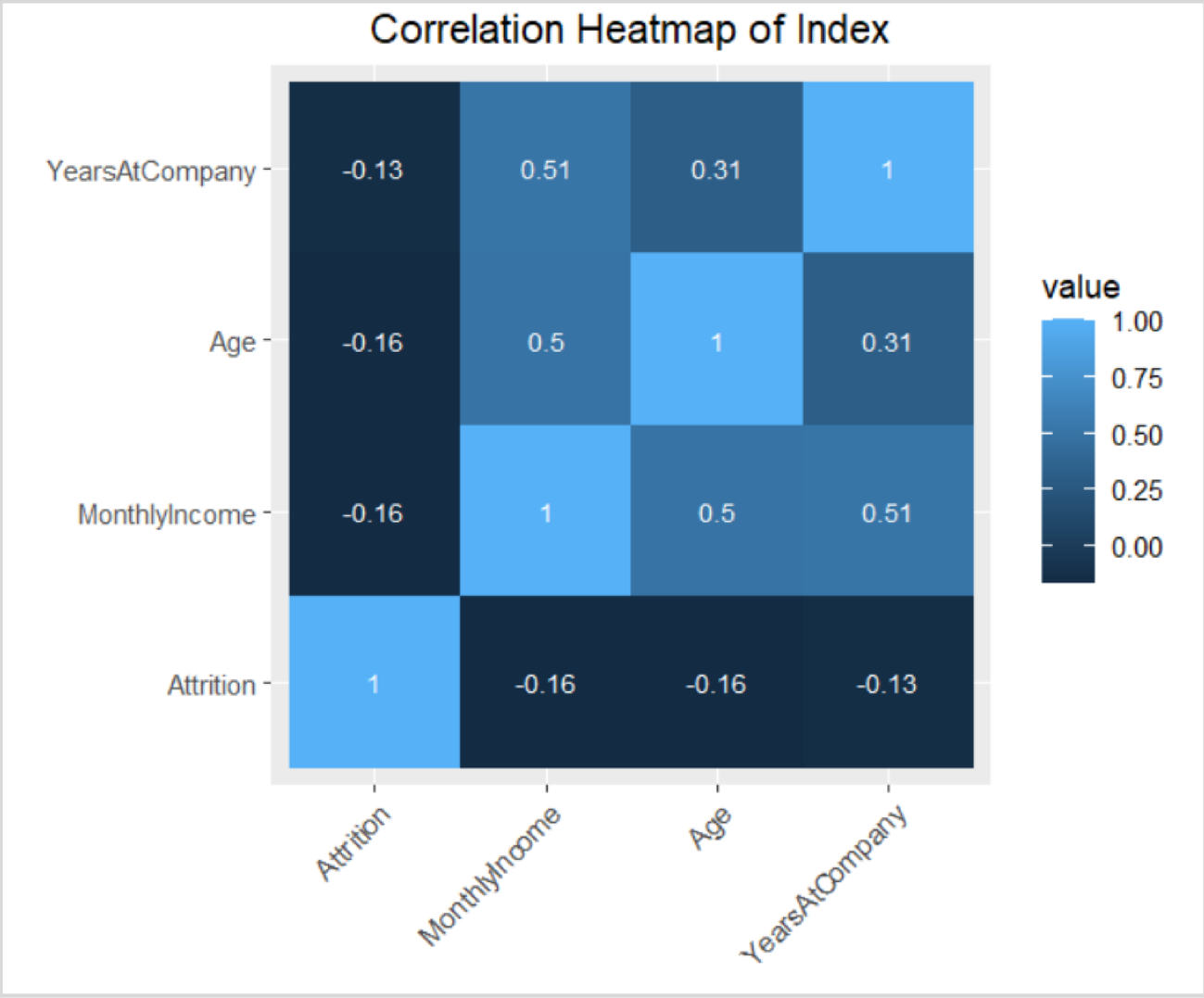
EDA　Density Plot





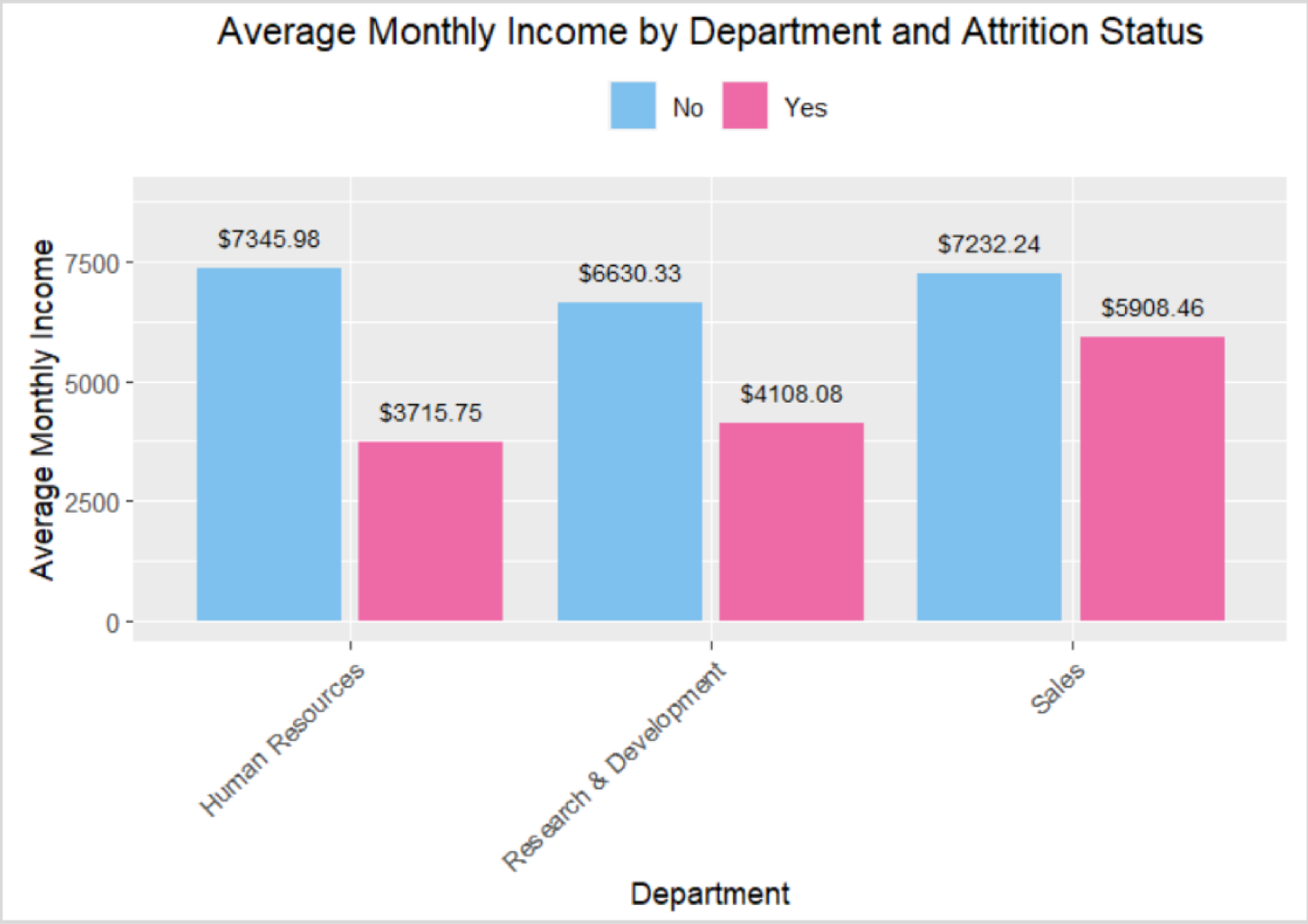
- 離職員工（「是」）通常月收入較低。
- 年輕員工（30 歲左右）的流動率較高。
- 工作年資較短的員工流失率較高。

EDA　Heat Map



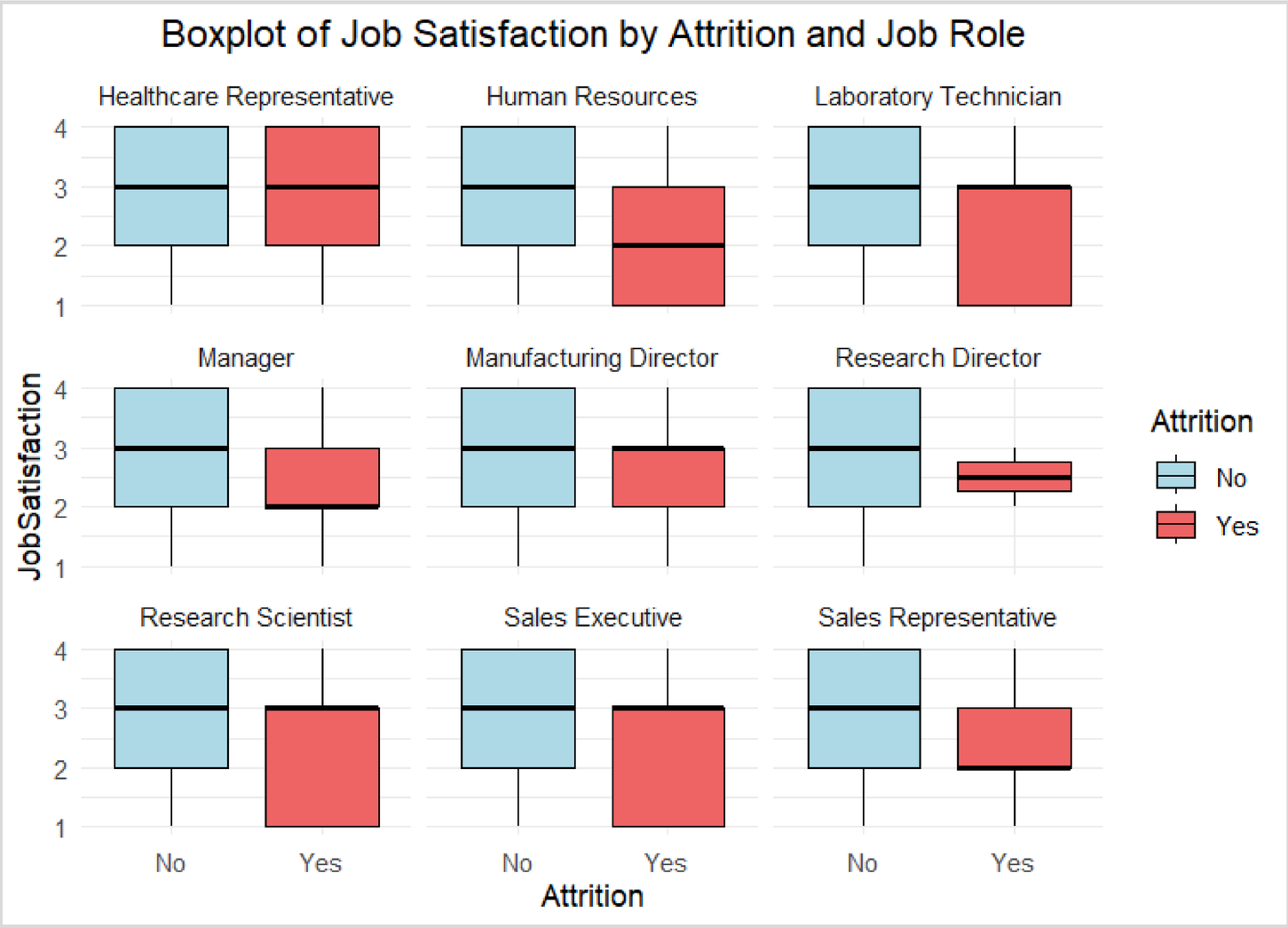
- 員工流失率與月收入、年齡、在公司工作年資皆呈**負相關**：收入越高、年齡越大、任期越長，離職率越低。
- 月收入與年齡、在公司工作年資呈**正相關**：經驗豐富、年齡較大的員工往往收入較高。

EDA　Barchart



- 所有部門中，留下員工（否）之平均月收入通常高於離職員工（是）。
- 人力資源部門中，留下員工和離職員工之間相較其它部門存在著巨大的收入差距。

EDA　Box Plot



- 不同工作角色的工作滿意度水準有所不同。
- 於大部分職位（e.g. 人力資源、實驗室技術員），離職員工（是）的工作滿意度普遍低於留下來的員工（否）。

三、統計檢定

統計檢定　Independent t-Test

H0：mean monthly income在Attirtion(Yes or No)相同 v.s.

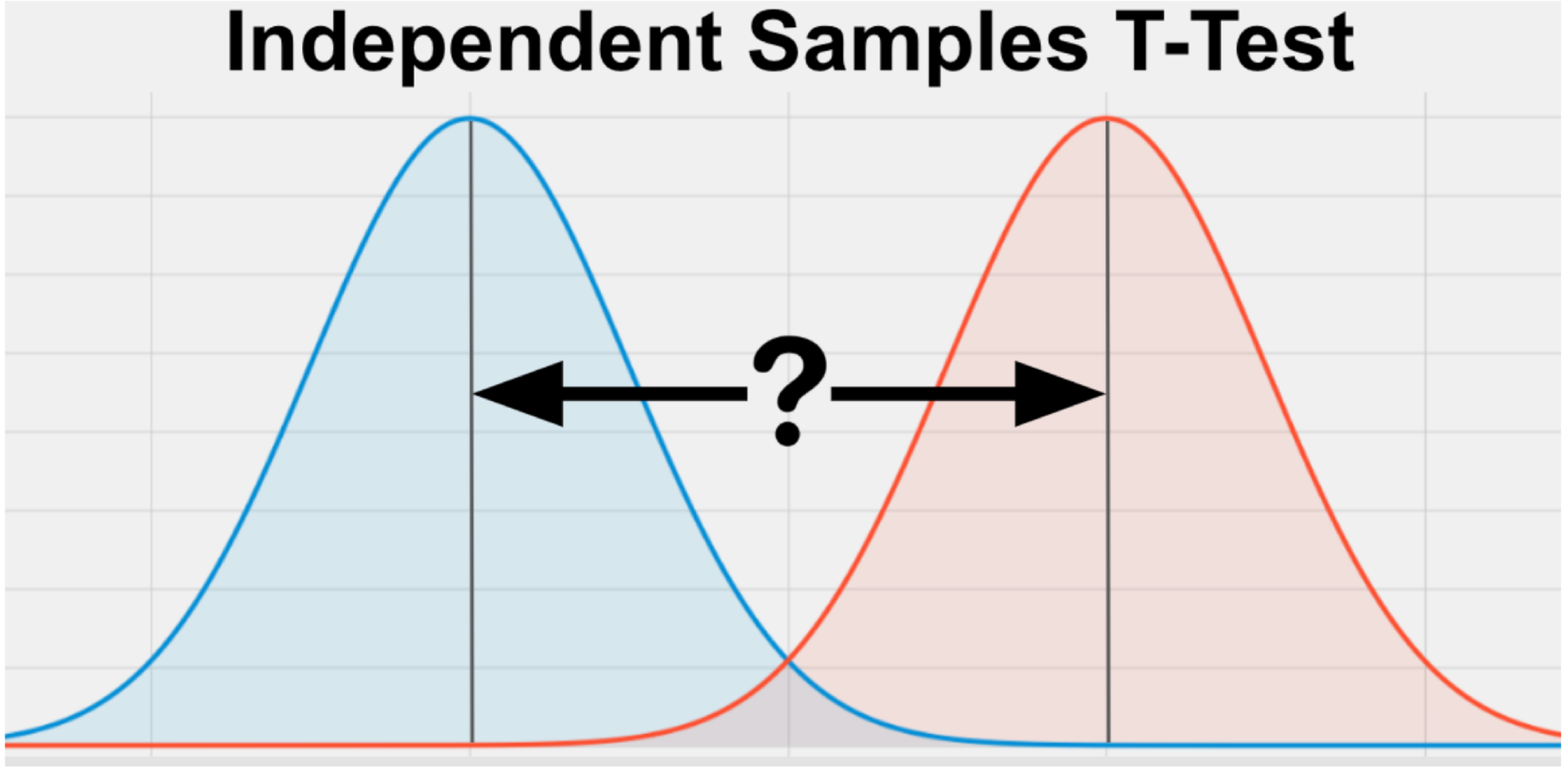
H1：mean monthly income在Attirtion(Yes or No)不同

Data: `attrition_yes_income` and `attrition_no_income`

- **t-statistic**: -7.4826
- **Degrees of freedom**: 412.74
- **p-value**: 4.434×10^{-13}
- **Alternative hypothesis**: True difference in means is not equal to 0
- **95% confidence interval**: -2583.050 to -1508.244
- **Sample estimates**:
  - Mean of `attrition_yes_income`: 4787.093
  - Mean of `attrition_no_income`: 6832.740

顯著差異！

統計検定　Why use Independent t-Test



source: https://www.statstest.com/independent-samples-t-test/

統計檢定　Chi-Square Test

H0：Attrition與Department不相關　v.s. H1：Attrition與Department相關

- **X-squared**: 10.796
- **Degrees of freedom (df)**: 2
- **p-value**: 0.004526

顯著差異！

Statistically Significant Association:

- Since the p-value (0.004526) is less than 0.05, we reject the null hypothesis.
- This means there is a statistically significant association between attrition and department.

Interpretation of Association:

- The association suggests that the attrition rates vary significantly across different departments.

統計檢定　Why use Chi-Square Test

The Chi-Square Test is used to determine if there is a **significant association** between **two categorical variables**. (Independence chi-square test)

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} = \sum \frac{(O-E)^2}{E}, df = k - 1$$

統計檢定　ANOVA Test

H0：不同JobRole間的薪水平均數相等 v.s. H1：至少一類的JobRole間的薪水平均數不相等

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
JobRole	8	2.657e+10	3.321e+09	810.2	<2e-16 ***
Residuals	1461	5.989e+09	4.099e+06		

顯著差異！

統計檢定　ANOVA Test

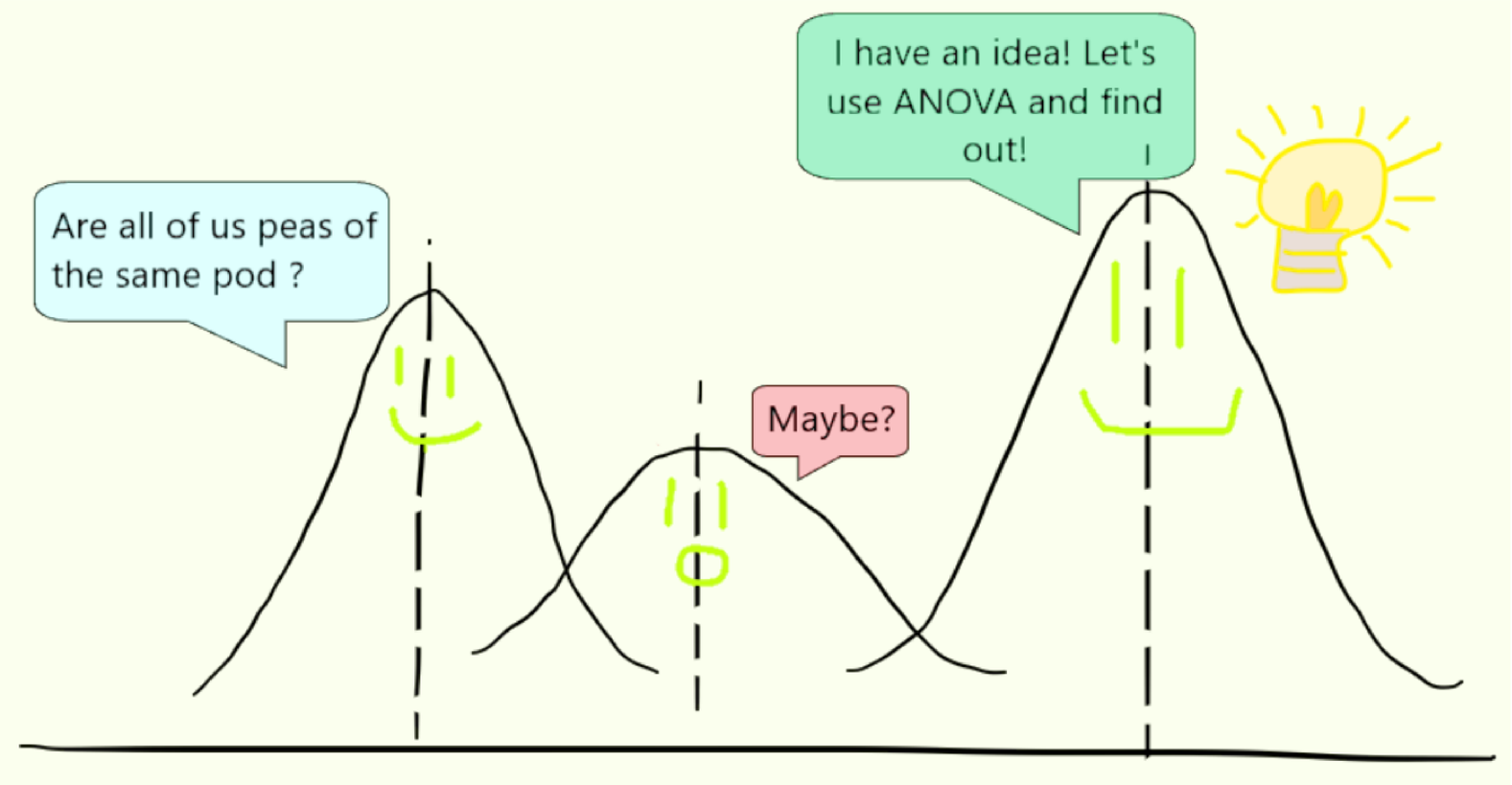
H0：不同Department間的工作滿意度平均數相等 v.s.
H1：至少一類的Department間的工作滿意度平均數不相等

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Department	2	1.2	0.6111	0.502	0.605
Residuals	1467	1785.5	1.2171		

沒有顯著差異！

統計検定　Why use ANOVA Test

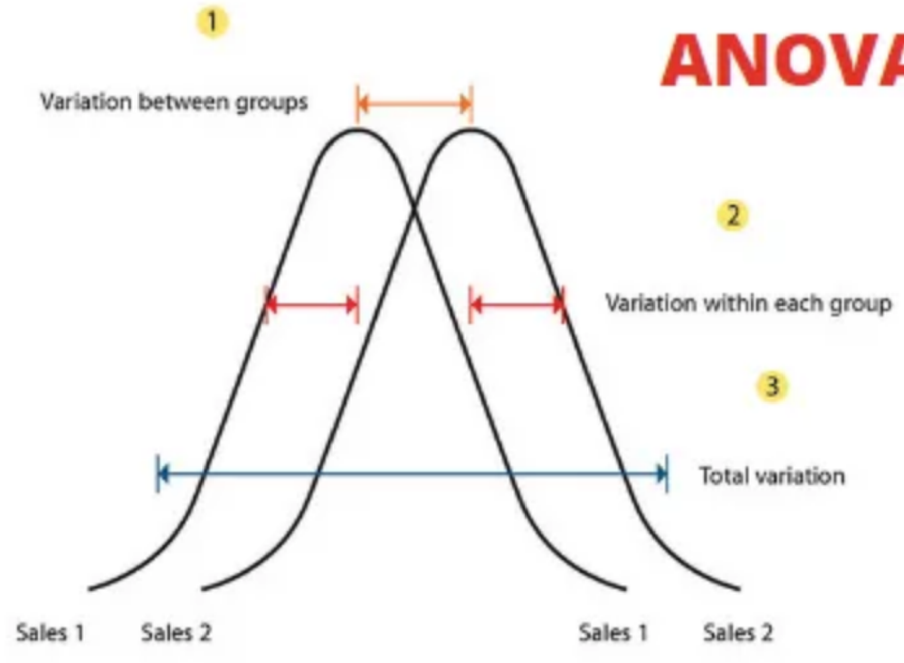
ANOVA is used to compare the means of **three or more groups** to see if at least one group mean is different from the others.



source: https://www.geeksforgeeks.org/one-way-anova/

統計檢定　Why use ANOVA Test

	Group 1	Group 2	Group 3	Group 4
Sample Size	n_1	n_2	n_3	n_4
Sample Mean	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$
Sample Standard Deviation	s_1	s_2	s_3	s_4



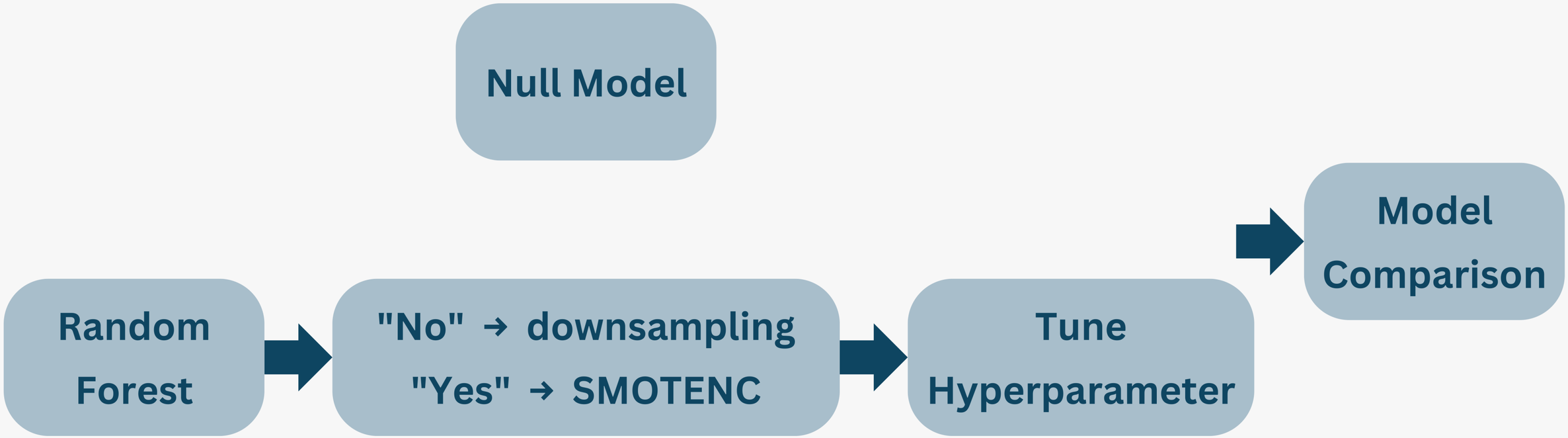
ANOVA testing

F-value= $\dfrac{\text{Mean between the groups}}{\text{Mean Within the groups}}$

Two sample groups of sales data

四、模型配適

建模流程


Null Model
Random Forest
"No" → downsampling
"Yes" → SMOTENC
Tune Hyperparameter
Model Comparison

Null Model　各項指標

Accuracy	0.8386
Kappa	0
Sencitivity	1
Specificity	0

Prediction \ Real	No	Yes
No	369	71
Yes	0	0

Random Forest K-fold

Validation Results by Folds

Fold	Accuracy	Precision	Recall	F1 Score
1	0.8563	0.8561	0.9658	0.9076
2	0.8438	0.8769	0.9268	0.9012
3	0.8500	0.8636	0.9500	0.9048
4	0.8688	0.8846	0.9504	0.9163
5	0.7188	0.2969	1.0000	0.4578

Mean

Metric	Value
Accuracy	0.8275
Precision	0.7556
Recall	0.9586
F1 Score	0.8175

Random Forest　各項指標與Hyper parameter

Accuracy	0.8545
95% CI	(0.8181, 0.8861)
Kappa	0.4161
Sensitivity	0.9322
Specificity	0.4507
Pos Pred Value	0.8982
Neg Pred Value	0.5614
Prevalence	0.8386

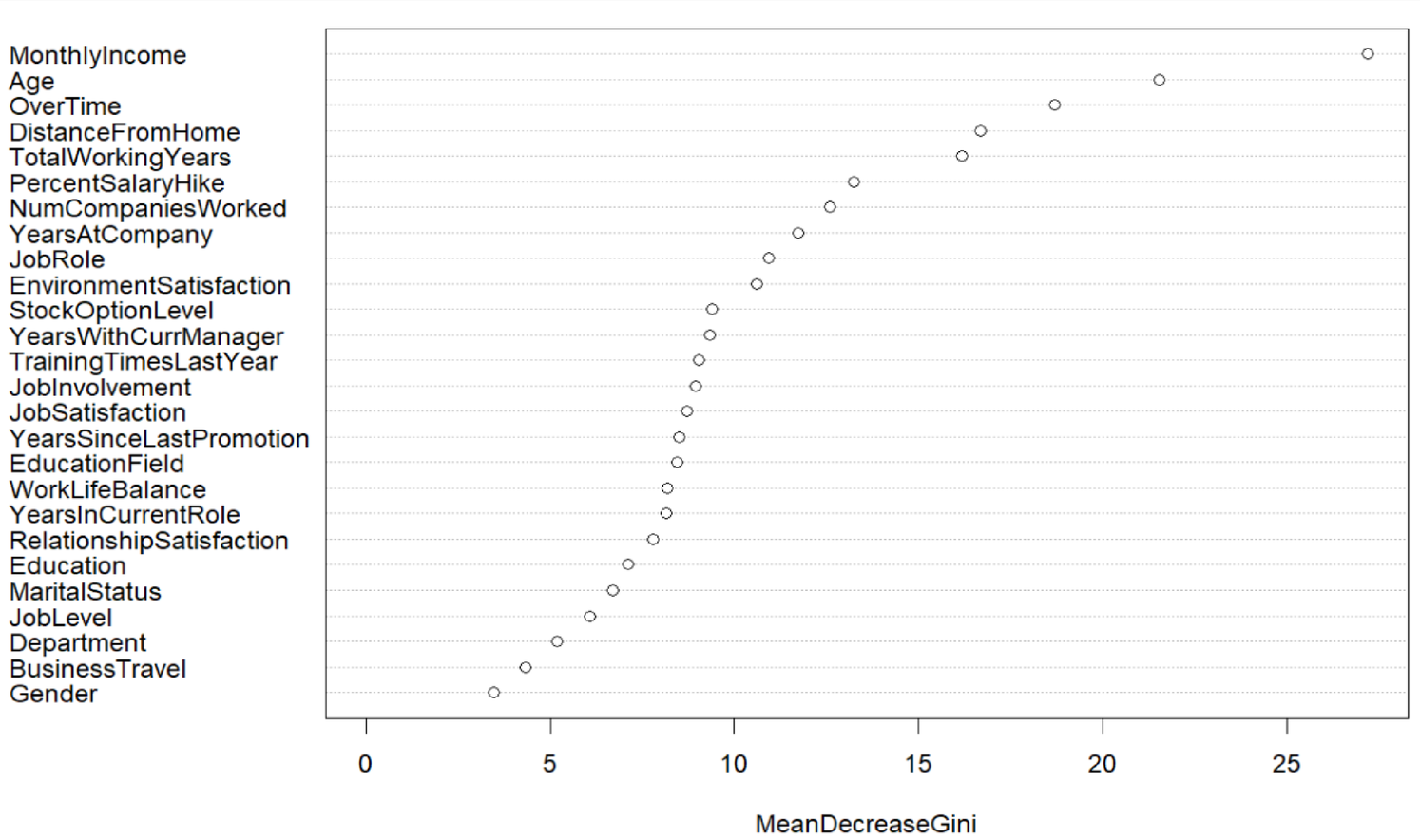
ntree	850
mtry	5

Real

Prediction

	No	**Yes**
No	344	39
Yes	25	32

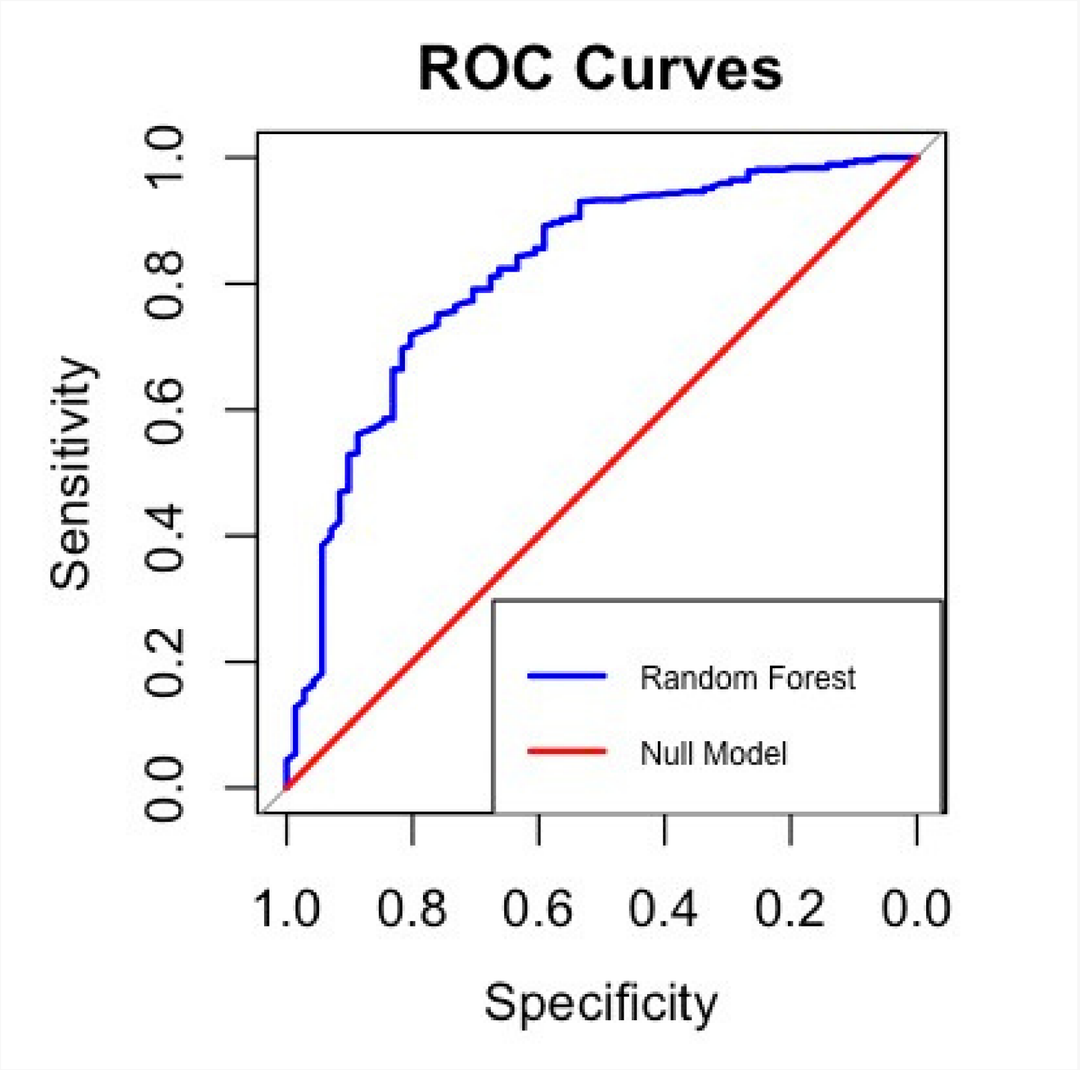
Random Forest　重要變數



由上圖可得重要變數有：MonthlyIncome、Age、OverTime、DistanceFromHome、TotalWorkingYears

ROC curve　Null Model v.s. Random Forest

	AUC
Null Model	0.5
Random Forest	0.8171



五、結論

結論

Null Model

Accuracy	0.8386
Kappa	0
Sencitivity	1
Specificity	0

Random Forest

Accuracy	0.8545
Kappa	0.4161
Sencitivity	0.9322
Specificity	0.4507

使用隨機森林後，模型準確度有顯著提升，其中對結果影響性較大的變數為：
MonthlyIncome、Age、OverTime、DistanceFromHome、TotalWorkingYears

附錄、海報展演

人家給予的回饋

其他人的回饋：

- HR經理：很多feature和檢定都是common sense，我們只是用數學和去驗證，之後可以考慮多加一些其他的因子，例如心理因素、公司文化之類的
- 同學1：對於imbalance的data，可以hybrid做oversampling和downsampling
- 同學2：薪水的Barchart為什麼不要用中位數或眾數而是平均數
  - 都可以，看情形，平均數反映整體水平，但可能被極端值影響、中位數可避免被極端值影響，但可能忽略整體變動的情形
- 同學3：為什麼k-fold CV可以避免overfitting？
  - k-fold 每次穩定切割，驗證模型效能
  - 做多次k-fold再平均結果，減少偏差
  - 如果training loss與validation loss差太多，可能有overfitting產生

看到別組的有趣事物

- 有些組別的專題具有高度互動性
- 結合AI與藝術
- 得獎組應用方面居多

Thank you

Sales Performance Analysis

Current Sales Trends:
Over the past years, we've observed Warner & Spencer's product sales going down because of some reason.

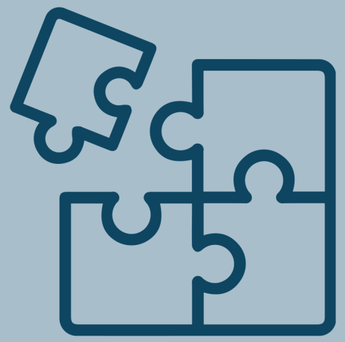
Customer Feedback:
Direct feedback from our customers has highlighted areas for improvement

Market Insights:
Through comprehensive market analysis, we've identified shifts in consumer preferences, competitive landscape changes, and emerging market segments.

Project Objectives

Analysis Phase

- Analyze sales performance, market trends, and consumer behavior.
- Conduct a SWOT analysis of existing sales strategy.



Strategy Development

- Create new strategies using marketing, sales promotions,
- Customer engagement to leverage strengths and opportunities.

Implementation Plan

- Create a timeline with milestones
- And responsibilities. Set KPIs to measure success.

Expected Outcomes

Increased Sales Figures:

- Target a 25% increase in sales over the next years.
- Measure success by tracking sales metrics and revenue growth.

Improved Customer Engagement:

- Foster stronger relationships with customers through personalized engagement strategies.
- Increase customer retention rates and loyalty.

Enhanced Market Reach:

- Expand market reach by tapping into new demographics or geographical regions.
- Strengthen brand presence through effective marketing campaigns.

Methodology

Data Collection:

- Gather sales data, market research, and consumer feedback through surveys and analysis tools.
- Utilize both primary and secondary research methods to gather comprehensive insights.

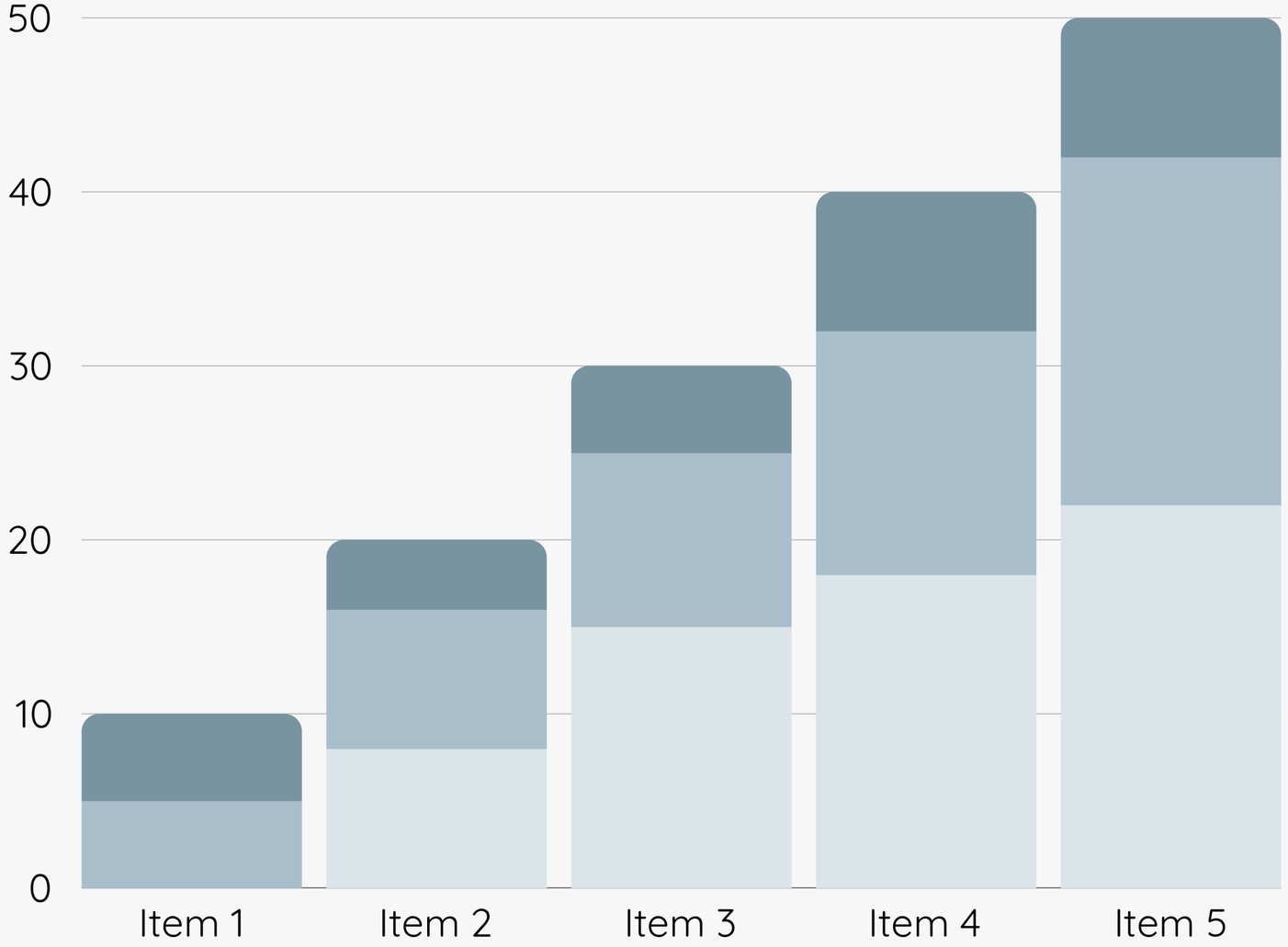
Brainstorming Sessions:

- Collaborate to generate innovative ideas for product positioning, pricing, and promotional campaigns.
- Consider various channels such as online platforms, partnerships, and offline marketing.

Testing and Refinement:

- Pilot the proposed strategies on a smaller scale to assess their effectiveness.
- Gather feedback, iterate, and refine the strategies based on initial results.

Data Analysis



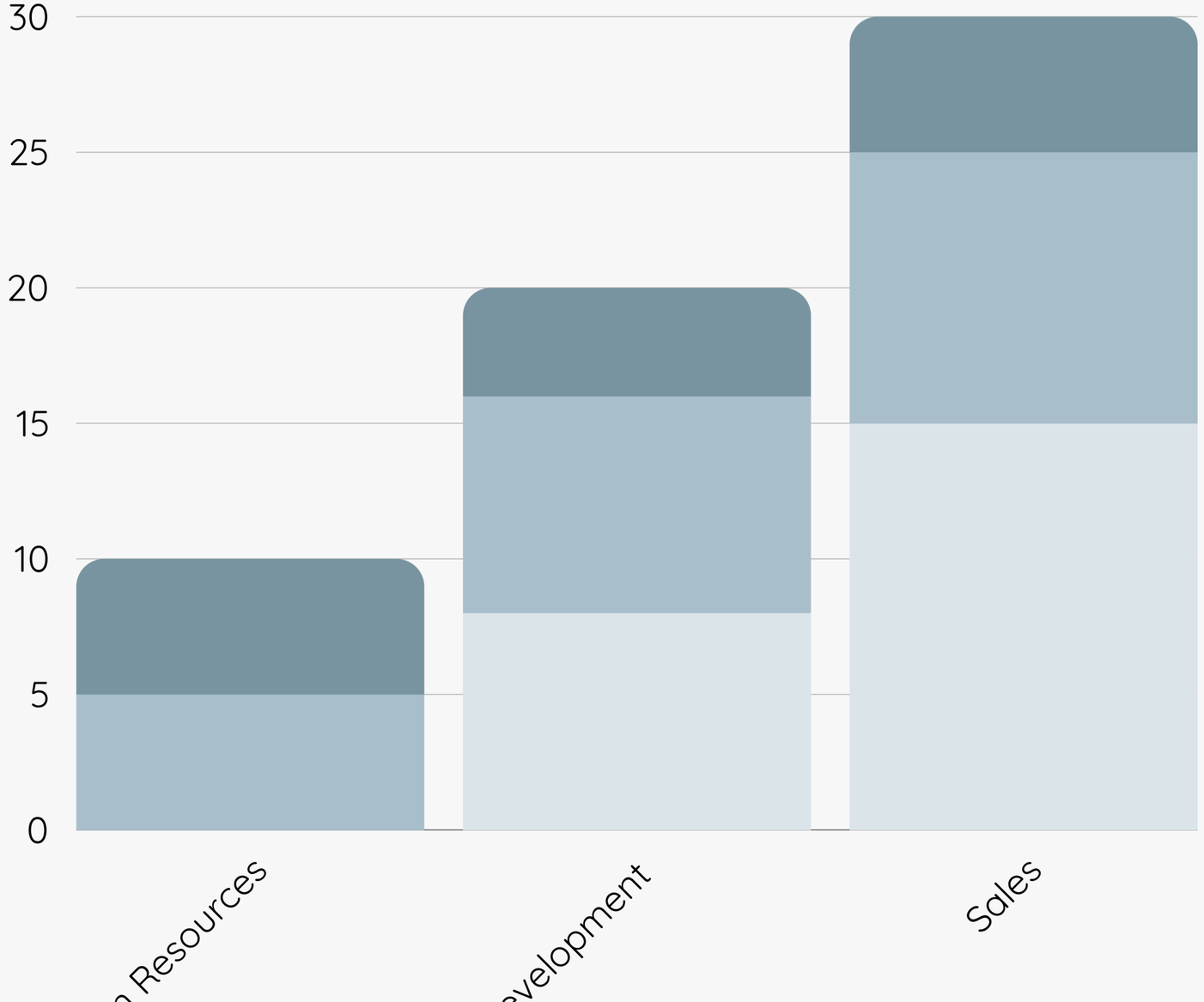
Over the past year, sales for Warner & Spencer have experienced a consistent decline, dropping month-on-month. The graphical representation of sales volumes reveals a noticeable downward trend, especially in the last quarter.

Product 1

Product 2

Product 3

Data Analysis



Over the past year, sales for Warner & Spencer have experienced a consistent decline, dropping month-on-month. The graphical representation of sales volumes reveals a noticeable downward trend, especially in the last quarter.

Product 1

Product 2

Product 3

Team Members

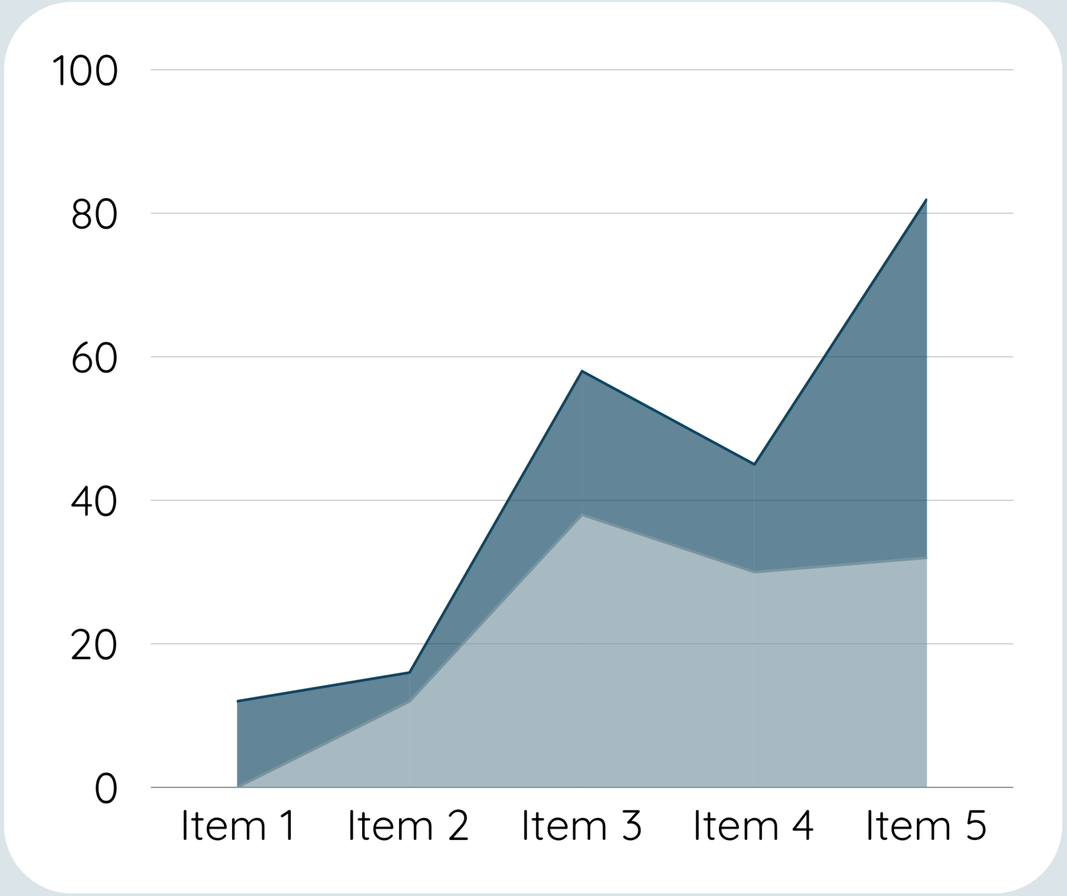
Neil Tran
Member

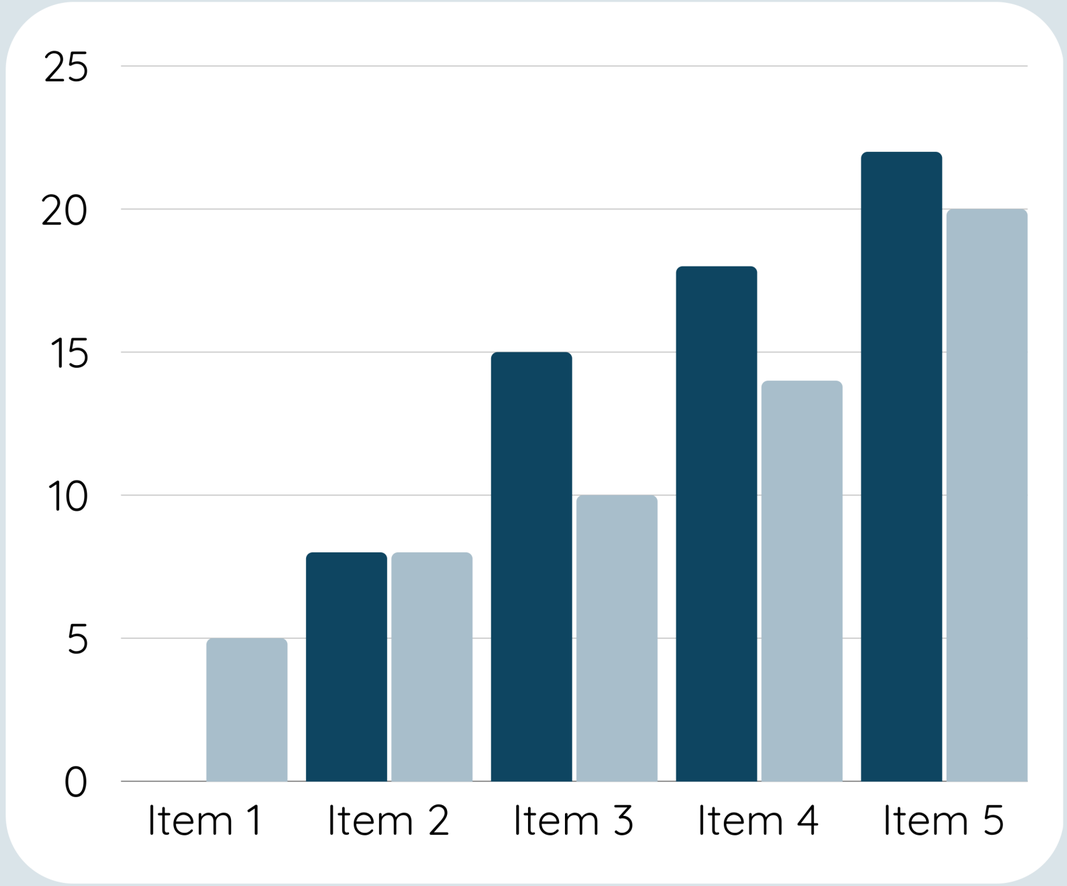
Alexander Aronowitz
Member

Harper Russo
Leader

Data Analysis

Customers are unhappy with Warner & Spencer's new packaging, which may be contributing to a decline in sales. Competitors offer better features and pricing, making it difficult for our product to stand out in the market.





Conclusion

By implementing a well-researched and innovative sales strategy, our goal is not only to boost immediate sales figures but also to establish a sustainable framework for continued growth and success.