Assessment 2 Template: Data modelling report

Michael Teixeira S4133975

Revolution Consulting Data Modelling Report

Table of contents

Introduction

In the context of Revolution Consulting's challenges with declining work quality and increased employee turnover, our mission is to:

1. Analyze employee data to uncover patterns related to attrition.
2. Segment employees into distinct groups based on their characteristics and risk of leaving.
3. Provide actionable insights and recommendations to improve employee retention and satisfaction.

This analysis will help Revolution Consulting address their concerns about losing valuable consultants and ultimately improve the quality of work delivered to clients.

Features overview

Feature name	Number of unique values	Type	Description
EmployeeID	1470	Numerical	Unique identifier for each employee
Age	43	Numerical	Age of the employee
Resigned	2	Categorical	Indicates whether the employee has resigned (Yes/No)
BusinessTravel	3	Categorical	The frequency of business travel for the employee (Rarely, Frequently, Non-Travel).
BusinessUnit	3	Categorical	The department or division within the company where the employee works.
EducationLevel	5	Categorical (Ordinal)	Level of education of the employee
Gender	2	Categorical	The gender of the employee (Male, Female).
JobSatisfaction	4	Categorical (Ordinal)	The employee's rating of their overall job satisfaction (Very Dissatisfied, Somewhat Dissatisfied, Somewhat Satisfied, Very Satisfied).
MaritalStatus	3	Categorical	The marital status of the employee (Single, Married, Divorced).
MonthlyIncome	1349	Numerical	Monthly income of the employee
NumCompaniesWorked	10	Numerical	Number of companies the employee has worked for
OverTime	2	Categorical	Indicates if the employee works overtime

PercentSalaryHike	15	Numerical	Percentage of salary increase
PerformanceRating	2	Categorical (Ordinal)	Performance rating of the employee
AverageWeeklyHoursWorked	23	Numerical	Average number of hours worked per week
TotalWorkingYears	40	Numerical	Total number of years the employee has worked
TrainingTimesLastYear	7	Numerical	Number of training sessions attended last year
WorkLifeBalance	4	Categorical (Ordinal)	Rating of employee's work-life balance
YearsAtCompany	37	Numerical	Number of years the employee has been with the current company
YearsInRole	19	Numerical	Number of years in the current role
YearsSinceLastPromotion	16	Numerical	Years since the employee's last promotion
YearsWithCurrManager	18	Numerical	Years working under the current manager

Methodology

Overview

The primary objective of this analysis is to pinpoint the primary factors contributing to employee turnover within Revolution Consulting. To achieve this, we will employ machine learning models (K-means and DBSCAN) to segment employees into distinct clusters based on shared characteristics. Subsequently, we will evaluate the likelihood of resignation for each cluster.

Clusters exhibiting a higher propensity for turnover will be subjected to in-depth examination to identify distinguishing characteristics. These characteristics can then be inferred as influential factors driving employee resignations.

Process

1. **Data Exploration:** We examined the dataset's structure, identified key variables, and visualized distributions and relationships.
2. **Feature Selection:** We chose relevant features for clustering based on their potential impact on employee attrition and another group with all numerical variables for comparison.
3. **Data Preprocessing:** We standardized the selected features to ensure equal weight in the clustering algorithms.
4. **Clustering:** We applied K-means and DBSCAN algorithms to segment employees into distinct groups.
5. **Visualization:** We plotted different charts and visualize the clustering results for different models and groups.
6. **Evaluation:** We assessed the quality of the clusters using silhouette scores and analysed the characteristics and statistics of each cluster.

Evaluation strategy

Intrinsic Measures

- **Within-Cluster Sum of Squares (WCSS):** For K-Means clustering, WCSS measures the average squared distance of each data point to its respective cluster center. Lower WCSS values indicate better clustering.
- **Silhouette Coefficient**: This metric evaluates how similar each data point is to its own cluster compared to other clusters. A higher silhouette coefficient suggests better clustering.

DBSCAN Evaluation

- **k-Distance Graph:** To determine the appropriate value for the eps parameter in DBSCAN, we can analyse the k-distance graph. This plot helps identify the knee point, which suggests a suitable value for eps.

Overall Evaluation

Given the research goal of identifying meaningful employee clusters and their propensity to resign, the primary focus of evaluation will be on the usefulness of the clustering output rather than solely relying on statistical measures.

Cluster Analysis

- **Resignation Percentage Differences:** We will examine the clusters to identify those with significant differences in the percentage of employees who resign.
- **Actionable Insights:** We will seek clusters that offer valuable insights into the characteristics of employees who are more likely to resign. These insights should be actionable, allowing for targeted interventions to improve retention.

Data modelling and exploration

Data Preparation

Before modelling, we prepared the data by addressing text features and scaling numerical values to ensure consistent feature distributions for clustering algorithms.

Data Transformations

- **Feature Removal:** The *EmployeeID* column was removed from the dataset as it is not considered a relevant feature for the analysis.
- **Categorical Data Encoding:** Categorical variables were encoded using the *LabelEncoder* technique, assigning unique numerical labels to each category.
- **Feature Scaling:** The *MinMaxScaler* was applied to standardize numeric features, ensuring they have values within a specific range (typically 0 to 1). This helps prevent features with larger scales from dominating the clustering process.
- **Relevant variables:** Created a group of only relevant variables and then compare against all numeric variables results.

Clustering Algorithms

For the analysis, two clustering algorithms will be used:

1. **K-Means:** A centroid-based algorithm that's efficient and works well for finding globular clusters. It's suitable for our dataset as we expect to find distinct groups of employees based on their characteristics.
2. **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise): A density-based algorithm that can find clusters of arbitrary shape and is robust to outliers. This can help us identify any non-globular clusters that K-means might miss.

Using both algorithms allows us to compare results and potentially uncover different patterns in the data.

K-Means Clustering Steps

1. **K Value Determination:** Based on the dataset's size and complexity, a range of potential cluster numbers (k) was established. In this case, the range was set from 2 to 10.

2. **KMeans Model Initialization:** A KMeans object was created in Python using the Scikit-learn library. This object represents the K-Means clustering algorithm.

3. **Iterative Clustering and Evaluation:**
    a. For each value of k within the defined range:
        i. The K-Means algorithm was applied to the data to partition it into k clusters.

        ii. The Within-Cluster Sum of Squares (WCSS) was calculated to measure the overall distance of data points to their respective cluster centers.
        iii. The Silhouette Coefficient was computed to evaluate the quality of the clustering by comparing the similarity of each data point to its own cluster versus other clusters.

4. **Visualization and Analysis:**
   The WCSS was plotted to visualize the relationship between these metrics and the number of clusters. This analysis helped identify the optimal value of k.
   b. The "elbow" method was used to determine the optimal k by observing the point where the WCSS curve starts to flatten, indicating diminishing returns with additional clusters.
   c. A high Silhouette Coefficient and a relatively high standard deviation of the Resignation percentage were also considered desirable factors for selecting the optimal k.

5. **Clustering and Interpretation:**
   Once the optimal k was selected, the K-Means algorithm was applied to the dataset to assign each data point to a cluster. The resulting cluster statistics were then analyzed to identify meaningful patterns and insights related to employee turnover.

DBSCAN Clustering Steps

1. **Choosing Epsilon (eps):**
   a. The "knee" point in the k_distance_graph suggests a suitable value for eps, the minimum distance between points to be considered neighbours.

2. **DBSCAN Object Creation:**
   A DBSCAN object is instantiated in Python, specifying the chosen eps and minimum number of neighbors (min_samples) required to form a dense cluster.

3. **DBSCAN Clustering:**
   The DBSCAN model is fitted to the data using the defined eps and min_samples.

4. **Cluster Label Assignment:**
   Each data point receives a cluster label:
   a. Core points belong to dense clusters.
   b. Border points are on the fringe of clusters.
   c. Noise points are considered outliers.

5. **Analysis:**
   The cluster labels are analysed to understand the distribution of data points into dense regions (clusters) and outlying areas (noise). Cluster characteristics, such as size and core point composition, may be explored for further insights.

Data Exploration

Resignation rates

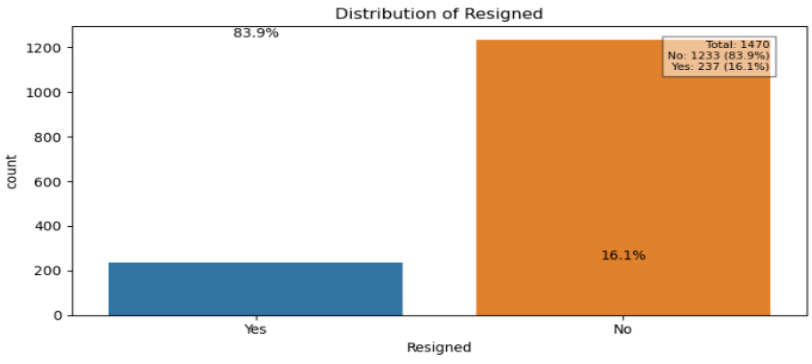


Fig.1. 16.1% of employees have resigned, which is a significant attrition rate. This directly relates to the company's concern about losing valuable consultants.

Overtime work and employee resignation

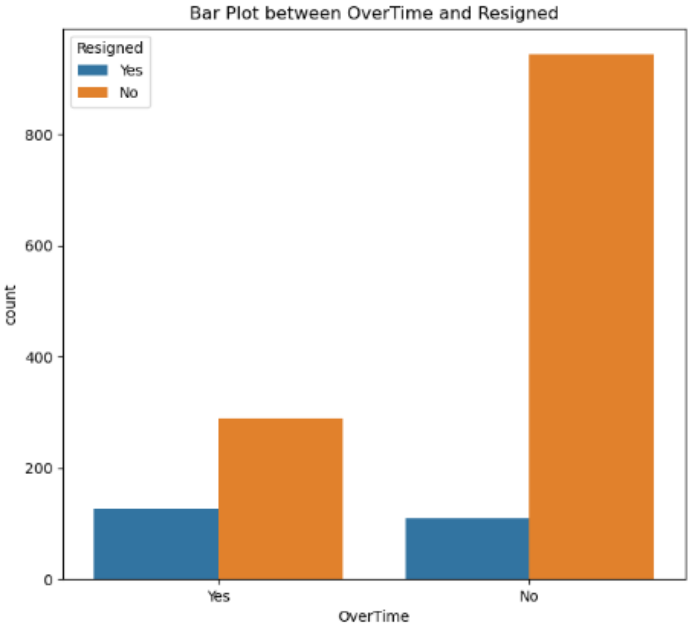


Fig.2. The bar plot shows the relationship between overtime work and employee resignation.

- Employees who work overtime are more likely to resign than those who do not.
- The count of employees who resigned and worked overtime is significantly higher than the count of those who resigned and did not work overtime.
- The plot suggests that there is a strong association between overtime work and employee resignation.

It is important to note that correlation does not necessarily imply causation. Further analysis would be needed to determine whether overtime work is actually causing employees to resign, or if there are other factors at play.

Monthly Income and Years at Company

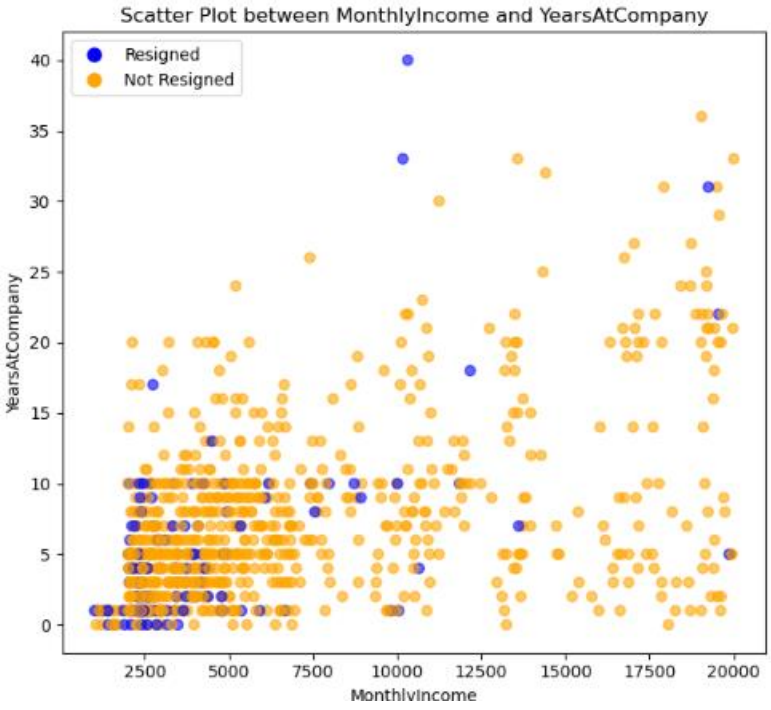
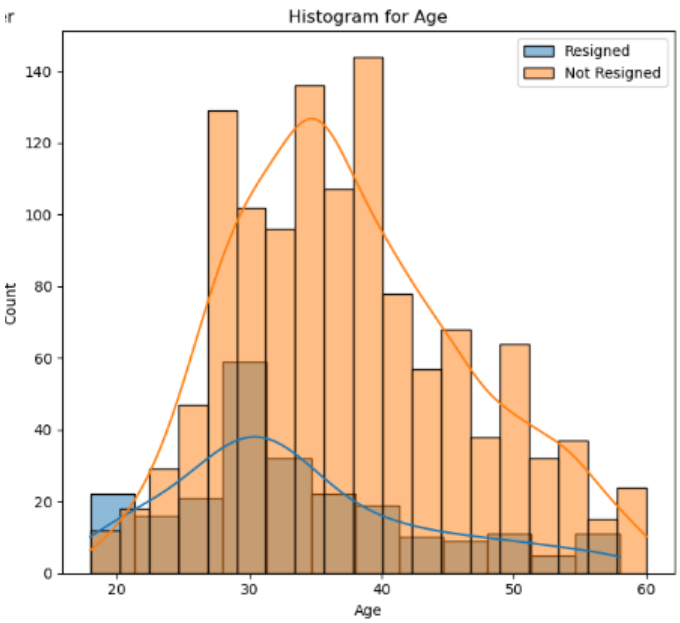


Fig.3. The scatter plot shows the relationship between monthly income and years at company, with the color of the points indicating whether the employee resigned.

- There seems to be a slight negative correlation between monthly income and years at company. This suggests that employees who earn higher salaries tend to stay with the company for fewer years.
- However, the correlation is not very strong, and there is a lot of overlap between the points representing employees who resigned and those who did not.
- The plot also shows that there are a few outliers, such as the employee who earned a high salary but stayed with the company for many years.

Overall, the scatter plot suggests that while there may be a slight negative relationship between monthly income and years at company, it is not a strong one and there are other factors that likely influence employee turnover.

Fig.4. The histogram shows the distribution of employees' ages, with the colour representing whether the employee resigned.



- There appears to be a concentration of resignations among younger employees, particularly those in their late 20s and early 30s. This suggests that younger employees are more likely to leave the company.
- On the other hand, the employees who did not resign are more evenly distributed across age groups, with a peak in the mid-30s to early 40s. This indicates that older employees tend to stay longer with the company.
- The plot also shows that resignations are much less common in employees over the age of 40, with very few resignations seen in the age group of 45 and above.

Overall, the histogram suggests that younger employees are more prone to resign, while employees in their mid-career (ages 30-45) and beyond are more likely to remain with the company.

Data Modelling

Model 1: KMeans - All Numerical
Model 2: DBSCAN - All Numerical
Model 3: KMeans - Relevant Variables
Model 4: DBSCAN - Relevant Variables

Clustering with All and Relevant Variables:

- I used both all numerical variables and a subset of relevant variables to perform KMeans and DBSCAN clustering. The goal was to observe how the clustering results varied when different input variables were used.

Data Filtering for Consultants:

- To gain more targeted insights, I filtered the dataset to include only employees from the Consultants business unit. This allowed me to compare the clustering results for the entire dataset with the results specifically for consultant employees.

Visualization of Clusters:

- I visualized the distribution of clusters by plotting Age vs. Monthly Income for both KMeans and DBSCAN models, comparing the results for the entire dataset against those for consultants only.
- I also created box plots to display the distribution of Job Satisfaction and Monthly Income across clusters for the general dataset and consultants specifically. This visual comparison helps identify potential differences in satisfaction or income levels among clusters.

Cluster Analysis:

- I analysed the cluster summaries, calculating the average values of important variables like job satisfaction and monthly income within each cluster.
- Additionally, I evaluated the resignation rate for each cluster to identify which group had the highest turnover rate, focusing on both the overall dataset and the consultants group.

Gender Distribution Comparison:

- I further explored the gender distribution within each cluster, comparing the proportion of males and females across clusters. This analysis was done for both the overall dataset and consultants, allowing for gender-related insights within clusters.

Results

Cluster Evaluation

The quality of each cluster was evaluated based on the percentage of employees who resigned within each cluster. A "good" cluster was defined as having a resignation rate below 10%, while a "concerning" cluster had a rate above 16%

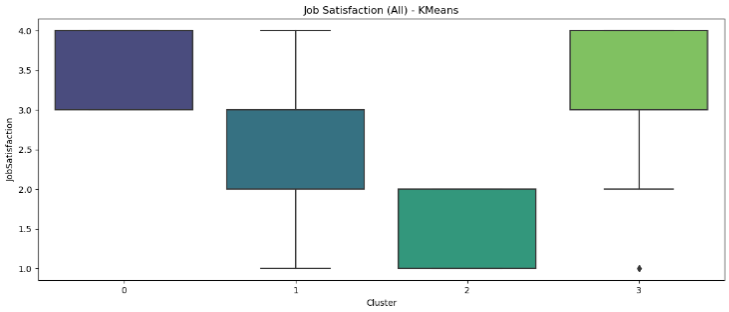
Resignation Rates:

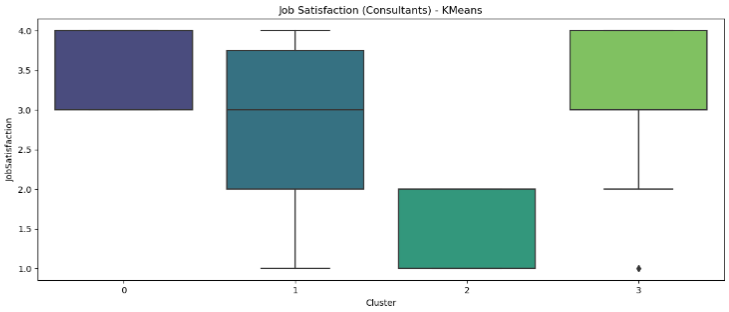
All Employees	Consultants
M1 - K-Means (All Numerical Features): • Cluster 0: 8.91% • Cluster 1: 16.37% • Cluster 2: 16.23% • Cluster 3: 20.74%	**K-Means (All Numerical Features):** • Cluster 0: 06.47% • Cluster 1: 16.66% • Cluster 2: 13.75% • Cluster 3: 17.44%
M2 - DBSCAN (All Numerical Features): • Cluster -1: 14.58% • Cluster 0: 16.25% • Cluster 1: 18.06% • Cluster 2: 0%	**DBSCAN (All Numerical Features):** • Cluster -1: 13.33% • Cluster 0: 13.28% • Cluster 1: 18.51% • Cluster 2: 0%
M3 - K-Means (Relevant Variables): • Cluster 0: 17.01% • Cluster 1: 06.67% • Cluster 2: 22.39% • Cluster 3: 11.72%	**K-Means (Relevant Variables):** • Cluster 0: 14.92% • Cluster 1: 04.92% • Cluster 2: 19.23% • Cluster 3: 10.11%
M4 - DBSCAN (Relevant Variables): • Cluster -1: 15.85% • Cluster 0: 16.11% • Cluster 1: 25.00%	**DBSCAN (Relevant Variables):** • Cluster -1: 17.64% • Cluster 0: 13.65% • Cluster 1: 0%

Model 3 - K-Means (Relevant Variables)

This model was particularly interesting for gaining insights as it revealed distinct clusters with varying resignation rates. Among the clusters, Cluster 2 and 0 stood out with the highest resignation rate, suggesting that employees in this group may have specific characteristics or conditions that make them more likely to leave the company.

Job Satisfaction per cluster:



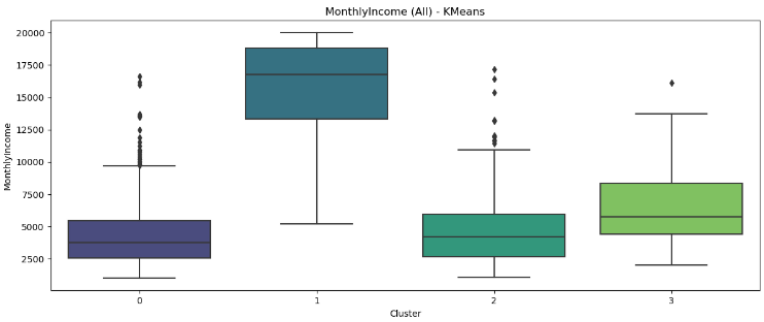


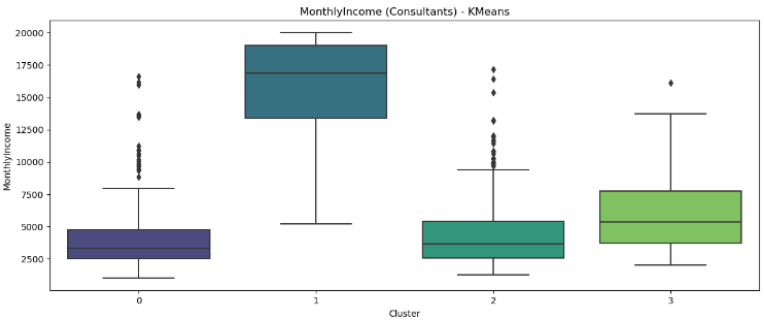
Cluster 2: Low job satisfaction combined with unfavourable work conditions might drive resignations.

Cluster 0: High job satisfaction doesn't guarantee retention. External factors like better opportunities or personal reasons could lead employees to leave.

Consultants follow similar trends to the general population

Monthly Income per cluster:





Most population of **Cluster 0 and 2** stands out with low income overall, that could be contributing to the high turnover.

Cluster 1, with the highest income and the lowest resignation rate (6.67%), suggests that higher-paying positions are associated with better retention.

Consultants follow similar trends to the general population, but the data highlights specific clusters where resignation risk is particularly high, such as Cluster 2.

Cluster Summaries (KMeans - Relevant Variables - All Data):

Cluster	Age	MonthlyIncome	JobSatisfaction	EducationLevel	WorkLifeBalance
0	34.60	4,406.76	3.51	2.83	2.73
1	48.61	15,623.10	2.58	3.06	2.76
2	34.46	4,796.90	1.45	2.86	2.75

3	36.48	6,354.05	3.36	3.04	2.84

Cluster	YearsAtCompany	YearsSinceLastPromotion	PercentSalaryHike
0	3.62	0.71	15.16
1	14.53	4.87	14.97
2	4.86	1.28	15.31
3	11.05	4.33	15.32

Cluster	AverageWeeklyHoursWorked	TotalWorkingYears	YearsInRole	YearsWithCurrManager
0	43.54	7.29	2.09	1.90
1	42.79	25.62	6.92	6.57
2	42.89	8.34	3.19	3.18
3	42.55	12.70	7.78	7.86

Key Observations:

1. **Age:**
   - **Cluster 1:** Has the highest average age, suggesting it may represent more experienced employees.
   - **Cluster 0 and 2:** Have lower average ages, indicating they might consist of younger or less tenured employees.
2. **Monthly Income:**
   - **Cluster 1:** Exhibits the highest average monthly income, possibly indicating higher-paying roles or longer tenure.
   - **Cluster 0 and 2:** Have lower average monthly incomes, suggesting lower-paying positions or earlier career stages.
3. **Job Satisfaction:**
   - **Cluster 0:** Shows the highest average job satisfaction, potentially indicating a more positive work environment or greater employee engagement.
   - **Cluster 2:** Has the lowest average job satisfaction, suggesting potential areas for improvement in employee satisfaction.
4. **Education Level:**
   - **Cluster 1 and 3:** Demonstrate similar average education levels, suggesting a comparable educational background among these clusters.
   - **Cluster 0 and 2:** Have slightly lower average education levels, indicating a slightly different educational profile.
5. **Work-Life Balance:**
   - **All clusters:** Show relatively similar average work-life balance scores, suggesting a consistent level of balance across different employee groups.
6. **Tenure and Experience:**
   - **Cluster 1:** Has the highest average years at the company, years since last promotion, total working years, and years in role, indicating longer tenure and more experience.
   - **Cluster 0 and 2:** Have lower averages in these metrics, suggesting shorter tenure and less experience.

Cluster Profiles:

- **Cluster 0:** Likely represents younger, less experienced employees with higher job satisfaction and a relatively good work-life balance.
- **Cluster 1:** May represent older, more experienced employees with higher incomes and potentially longer tenures.
- **Cluster 2:** Might represent employees with lower job satisfaction and a mix of experience levels.
- **Cluster 3:** Appears to be a cluster with moderate levels of age, experience, and income, potentially representing a diverse group of employees.

Discussion

In this analysis, the comparison between K-Means and DBSCAN yielded interesting results. While K-Means initially appeared more effective in achieving the research objectives by generating clusters that provided valuable insights into resignation rates, the Silhouette Analysis revealed a more nuanced picture.
The Silhouette Analysis, which measures how similar an object is to its own cluster compared to other clusters, surprisingly showed better results for DBSCAN. This suggests that DBSCAN may have created more cohesive and well-separated clusters from a mathematical standpoint. However, there was a disconnect between these quantitative results and the practical interpretability of the clusters. Despite DBSCAN's superior performance in the Silhouette Analysis, it was more challenging to verify and interpret these results through the selected clusters. This difficulty in interpretation limits the practical applicability of DBSCAN's results in this context.
K-Means, while potentially scoring lower on the Silhouette Analysis, provided more readily interpretable clusters, particularly in identifying groups with higher resignation rates. This interpretability made K-Means more immediately useful for deriving actionable insights.
This situation highlights an important consideration in cluster analysis: the balance between mathematical optimality and practical interpretability. While DBSCAN may have produced more mathematically robust clusters according to the Silhouette Analysis, K-Means delivered results that were more aligned with the research goals and easier to translate into actionable insights.
This outcome emphasizes the importance of considering multiple evaluation metrics and the ultimate research objectives when selecting and interpreting clustering algorithms.
(Graphs and further statics are available on the Jupiter notebook).

Conclusion

Based on our analysis, we recommend the following actions for Revolution Consulting:

1. Implement targeted retention strategies for high-risk employee segments, focusing on Cluster 0 and 2.
2. Review and adjust compensation structures, particularly for Cluster 0 and 2.
3. Enhance career development programs to address stagnation in roles, especially for employees in Cluster 0 and 2.
4. Improve work-life balance initiatives, which our analysis showed to be a significant factor in job satisfaction.
5. Conduct regular employee satisfaction surveys to monitor the effectiveness of these interventions.
6. Consider personalized retention plans for high-value consultants based on their cluster characteristics.

By implementing these recommendations, Revolution Consulting can address the root causes of employee attrition, improve job satisfaction, and ultimately enhance the quality of work delivered to clients.

.

References

Scikit-learn: Machine Learning in Python, Pedregosa *et al*, JMLR 12, pp. 2825-2830,2011

Appendix