Assessment 2: Data modelling report

Timm Rahrt - Revolution Consulting

Table of Contents

1. Introduction ... 1

2. Features overview ... 1

3. Methodology ... 3

3.1 Research goal ... 3

3.2 Process ... 3

3.3 Evaluation strategy ... 3

3.4 Data exploration ... 4

3.5 Data modelling ... 5

4. Results ... 7

5. Discussion ... 10

6. Conclusion ... 12

7. References ... 13

8. Appendix ... 14

1. Introduction

The IT consulting firm Revolution Consulting is currently facing challenges related to high employee turnover, which has led to a decline in project quality and growing concerns from clients. The management suspects that issues such as gender pay gaps, limited career progression, and overall job dissatisfaction are contributing to this problem.

This report focuses on exploring these potential drivers of employee attrition through data-driven analysis. By leveraging machine learning models such as KMeans and DBSCAN, the report intends to uncover patterns and insights within the employee data, enabling the identification of key factors that contribute to turnover. The report is structured to provide a detailed exploratory data analysis (EDA), data modelling to cluster employees into different groups with same characteristics, and recommendations that can help Revolution Consulting to improve overall business performance and employee satisfaction to reduce employee turnover. We will achieve this by analysing the clustered employees based on their characteristics and corresponding resignation rate, so that our management can make data-driven decisions of how to keep the employees risk-groups.

2. Features overview

The dataset contains 1470 observations, each is representing an employee, with 20 features of all data types, such as nominal, oridinal and interval/ratio. Each feature represents a variable that influences the employee to stay or resign. An example are a couple of categorical features which were collected during an employee survey.

Feature name	Number of unique values	Type	Description
EmployeeID	1470	Integer (int64) / Interval	Unique identifier for each employee
Age	43	Integer (int64) / Interval	Employee's age in years
MonthlyIncome	1349	Integer (int64) / Interval	Employee's monthly salary
NumCompaniesWorked	10	Integer (int64) / Interval	Number of different companies employee worked for
PercentSalaryHike	15	Integer (int64) / Interval	Percentage increase in employee's salary
TotalWorkingYears	40	Integer (int64) / Interval	Total number of years employee has been working
TrainingTimesLastYears	7	Integer (int64) / Interval	Number of training sessions the employee had over the last year
YearsAtCompany	37	Integer (int64) / Interval	Number of years employee is employed by our company
YearsInRole	19	Integer (int64) / Interval	Number of years employee has the same position
YearsSinceLastPromotion	16	Integer (int64) / Interval	Number of years since employee's last promotion
YearsWithCurrManager	18	Integer (int64) / Interval	Number of years employee worked under same manager
EducationLevel	5	Integer (int64) / Ordinal	Highest level of education employee helds
JobSatisfaction	4	Integer (int64) / Ordinal	Employee's satisfaction level with their job
PerformanceRating	2	Integer (int64) / Ordinal	Employee's performance rating
WorkLifeBalance	4	Integer (int64) / Ordinal	Employee's perception of their work-life-balance
AverageWeeklyHoursWorked	23	Float (float64) / Ratio	Average number of hours employee works per week
Resigned	2	Object / Nominal	Indicates whether employee left the company
BusinessTravel	3	Object / Nominal	Frequency of employee's business travel
BusinessUnit	3	Object / Nominal	Department in which employee works
Gender	2	Object / Nominal	Gender of employee
MaritalStatus	3	Object / Nominal	Marital status of employee
Overtime	2	Object / Nominal	Indicates whether employee works overtime

3. Methodology

3.1 Research goal

The main goal of this report is to identify what the key factors of employee attrition at Revolution Consulting are, using data-driven analysis to uncover actionable insights. The aim is to provide management with targeted strategies to retain top talent and reduce turnover by clustering the employees into different groups of similar characteristics. Each cluster will then be further investigated in terms of resignation rate and their reason, as well as improvement procedures our management then has to implement.

3.2 Process

This report was separated into two main parts, first we apply Exploratory Data Analysis (EDA) on all features as well as feature pairs to gain a deeper understanding on similar correlations that could lead to resignations. Secondly, we model our preprocessed dataset using KMeans and DBSCAN to further categorise all employees into cluster with similar characteristics. To achieve this:

1. We import the dataset and apply basic summary statistics to display the variables, their type, rows and feature unqiue values
2. Next, we continue with the EDA, create a function that displays various plots of single column to further understand their distribution and concentration
3. We apply standard statistics on each plot to gain further insights
4. The second EDA part concentrates on the visualization of multiple features in one plot to create pairs of attributes using a function that outputs various different plots
5. We visualize different attribute pairs to gain a deeper understanding of each columns correlation to each other
6. Finally, we continue with the Data Modelling part, starting by scaling our datset using MinMaxScaler (Scikit-learn, 2019)
7. Next, we convert all categorical variables to numerical ones by mapping numerical categories to each categorical feature (Hackman, 2024)
8. We define function to display Model Evaluation plots for KMeans (here: WCSS and Silhouette Coefficent) to identify the optimal amount of clusters (k value)
9. We then perform KMeans on our preprocessed dataset to identify patterns in clusters with high resignation rates
10. Lastly, we use and compare the model DBSCAN on our datset by defining a function that plots the optimal eps value and models our dataset using DBSCAN
11. We visualize the same feature pairs to identify which machine learning model performs better

3.3 Evaluation strategy

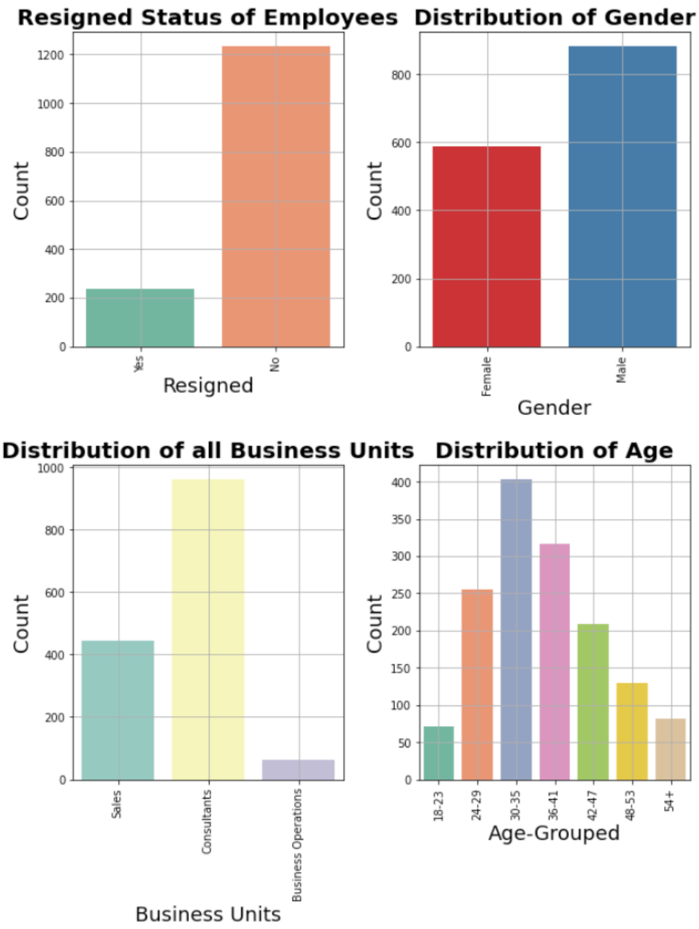
This Assessment focuses on intrinsic evaluation to assess the quality of KMeans clustering as no external labels or ground truth are given. We define a function that will apply two methods to determine the optiomal k value, Elbow Method (WCSS) and Silhouette Coefficient for optiomal k. WCSS measures the compactness of clusters by calculating the sum of squared distances between each data point and its corresponding cluster centroid, where a lower Inertia value indicates that data points are tightly grouped around their centroid. The \`elbow point\` indicates the optiomal number of cluster (k value), all points beyong that value do not reduce the Inertia value significantly. The Silhouette Score measures how well each data point fits within its assigned cluster relative to other clsuters. Its range goes from -1 to 1, where higher values indicate better defined clusters (Gupta, 2019). The combination of both method offer the optimal way of ensuring the number of clusters (k) are chosen correctly, as the identification of the optiomal \`elbow point\` of WCSS can be quite challenging (Scikit-learn, 2024).

For evaluating DBSCAN we use another intrinsic evaluation method k-th Nearest Neighbor as this approach is very effective to the DBSCAN algorithm as it focuses on density-based clustering. The method is used to identify the optimal eps value for DBSCAN by calculating the distance of each data point to its k-th closest neighbor in the dataset. This distance is used to identify the best eps value at its `elbow point`, as this point indicates a cutoff to distinguish between core points, border points and outliers/ noise (Jjunk, 2024).

Lastly, we define a function that will apply an extrinstic evaluation method on KMeans using the Adjusted Rand Index, but as we don’t have any true labels for our modelled dataframe we are unable to use this extrinsic evaluation.

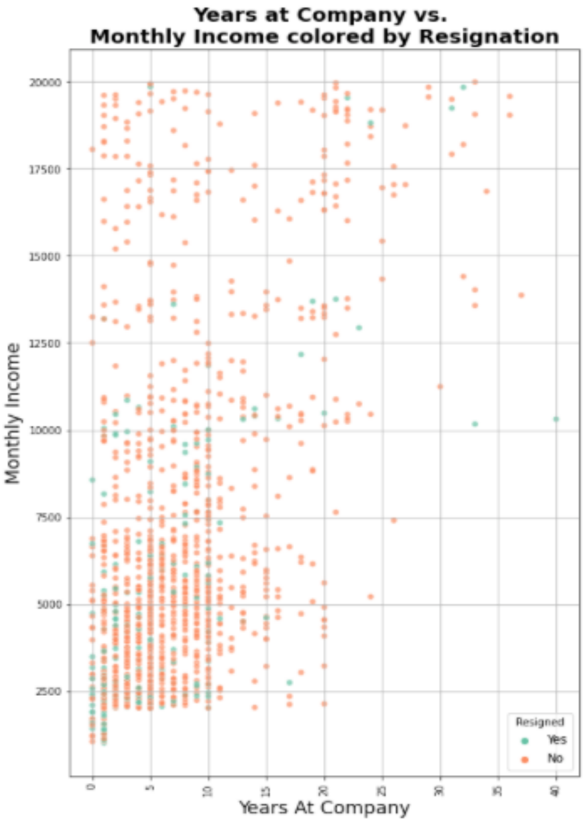
3.4 Data exploration

We begin by applying some exploratory data analysis onto our dataset to gain further insights of our features. Our first ten plots visualise the distribution of some key variables to identify what feature pairs could provide interesting findings to answer our research question. This step is crucial to identify patterns, anomalies, and releationships between variables which are used for the subsequent modelling process.

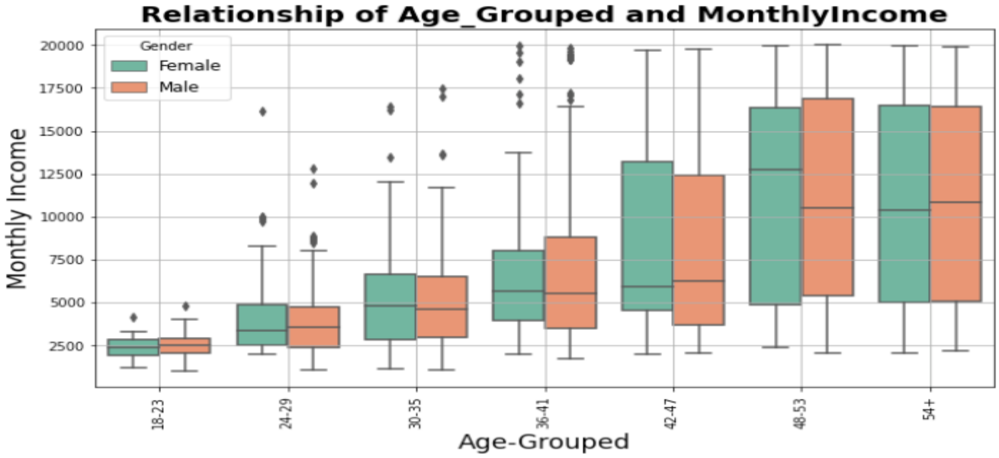


We visualise the distribution of fundamential features of our company to identify the composition and dynamics at Revolution C. The Resigned Status of Employees shows that 16.12% of employees resigned. This is above average and concerning, identifying that employees are unhappy in their current role which leads to higher turnover. We also see that 60% of all employees are male, which is fairly distributed but this plot might help for further analysis whether gender influences resignation rates. Our third plot ‘Distribution of Business Units’ shows that the majority of employees (65.37%) work in Consulting, 30.34% in Sales and just 4.29% in Business Operations. It implies that the low concentration of employees in Business Operations might be a factor for resignation trends. Lastly, our Age Distribution shows that half of our employees are younger than 36 years old, indicating a younger workforce which shapes the culture of the company. This could also correlate with certain trends in resignation.

Next, we explore ten feature pairs to uncover any relationships between variables that could provide insights into employee attrition.



When we compare Monthly Income by Years at Company colored by Resignation, we can identify whether employees who stay longer and earn more are less likely to resigned. The plots identicate that employees who resigned are primarily clustered in the lower income range and new to the company (0-5 years). The violin plot supports this statement, showing a higher concentration of resignations among lower-income groups, while the gender of employees is fairly even distributed. In contrast, employees who stayed tend to have a broader income distribution, implying that higher income plays a role in retaining employees (Violin plot (P1) in Appendix).



The analysis reveals disparities in pay across different age groups and genders. The box plot shows that the median monthly income increases with age, with older age groups earning considerably more than younger groups. In addition to this, within each age group a gender pay gap is also evident, with male employees generally earning more than their femals coutnerparts, especially in higher age brackets. We use a line plot (Plot P2 in the Appendix) to explore the relationship between performance rating and average weekly hours by gender, to see if any performance related variables explain the income difference between the genders. As the plot shows a weak correlation suggests that other factors influence the gender pay gap. This could be demotivating for the female employees as their performance is not being taking into consideration. Revolution C. could implement transparent pay structures to reduce biases and ensure that employees are compensated fairly for their role.

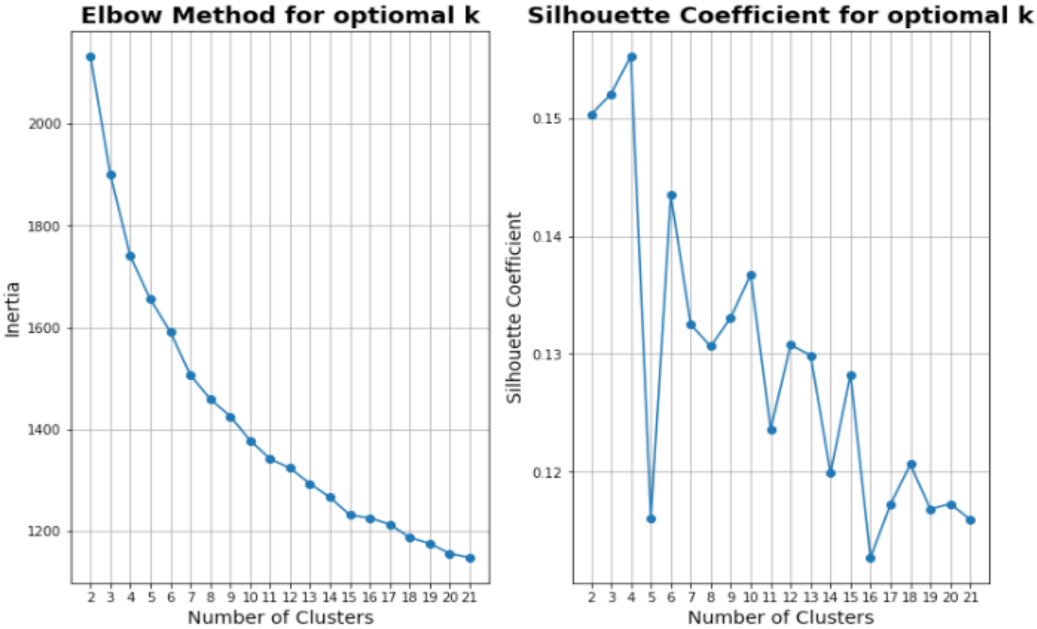
3.5 Data modelling

Next, we proceed by clustering our dataset into various groups representing each with different characteristics which will lead to different resignation rates. After that, we further explore the clusters with the highest resignation rate, what features motivates the employees within this group to resign and what our management could implement or improve to reduce employee turnover and improve the work quality of consultants. This will be done with the help of two unsupervised machine learning techniques, KMeans and DBSCAN.

Model 1 KMeans with the evaluation methods WCSS and Silhouette Coefficient:

Before we apply any model technique onto the dataset, we remove all unnecessary or biased features such as EmployeeID, Age_Groued and Resigned, as our model would be biased. Additionally, we convert all categorical features by mapping numerical representations of their categories, as string values can't be used for our modelling techniques.

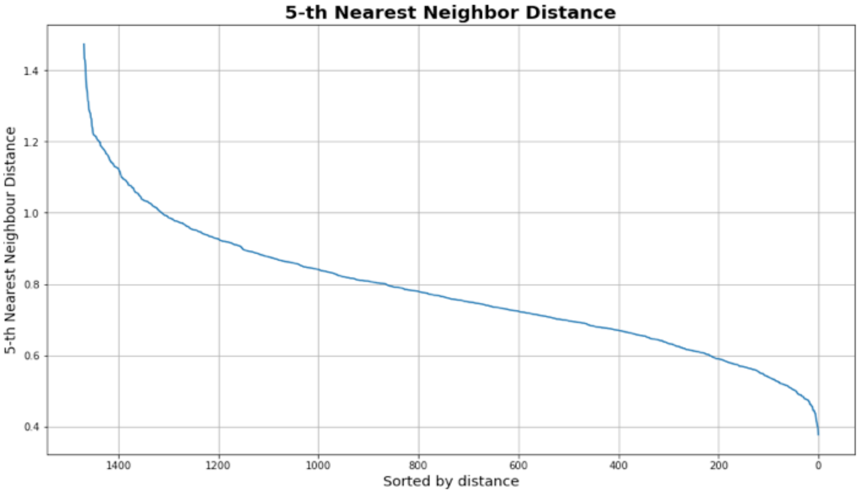
We then apply KMeans clustering to segment the employee population into distinct groups based on their features. The number of clusters are determined by the previously explained KMeans intrinsic model evaluation method WCSS (Elbow Method) and Silhouette score, ensuring the optimal separation between clusters (Hailu, Yu and Fantaye, 2020). The methods evaluate how similar an object is to its own cluster compared to other clusters. Finally, we plot the clusters to visualise each group and further analyse the characteristics of each group to further investigate what characteristics lead to higher resignation rates and which factors were most influential.



As the 'elbow point' of the WCSS graph is difficult to identify, we also use the Silhouette Coefficient. The Silhouette Coefficient has an optiomal number of clusters, within the Elbow Point of WCSS, at the value of 6, as the Silhouette Coefficient is the highest with a value of 0.145.

Model 2 DBSCAN with the evaluation method `kth- Nearest Neighbor Distance`.

We also test a second machine learning method DBSCAN to identify whether a density-based method could reveal better groupings than KMeans. We first use kth Nearest Neighbour Distance to determine the optimal eps value we can use in combination with a reasonable minimum number of samples (min_samples) to handle noise in the dataset. This method is different to KMeans as it does not require a pre-specified number of clusters (k value). Lastly, we will further explore the identified clusters of a high resignation rate for characteristics that lead to a high employee turnover and compare the models performance with KMeans.



The DBSCAN evaluation method shows that the optiomal eps values is between 1 and 1.1, as this is where the curve starts to flatten (Elbow point).

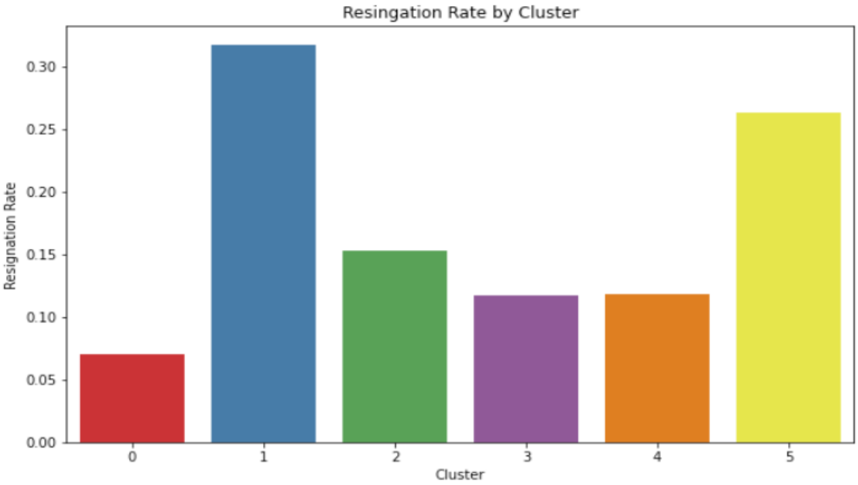
4. Results

This section will present the insights of both unsupervised machine learning models KMeans and DBSCAN on our preprocessed dataframe. We will further investigate which cluster is at risk of leaving the company and what characteristics they share. We will also compare both models.

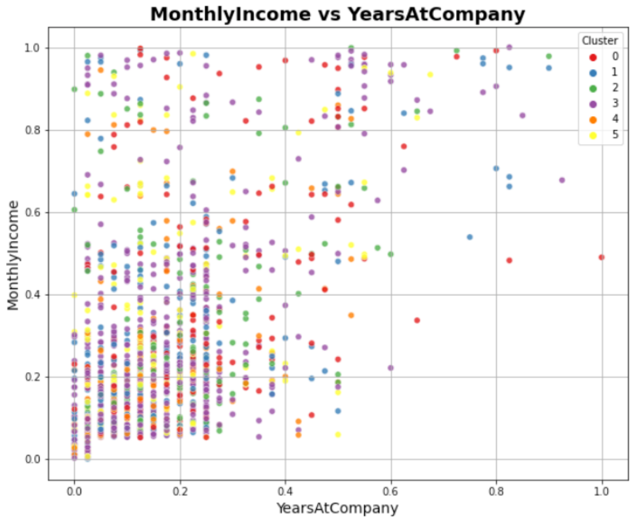
Model	KMeans on all features incl. categorical features

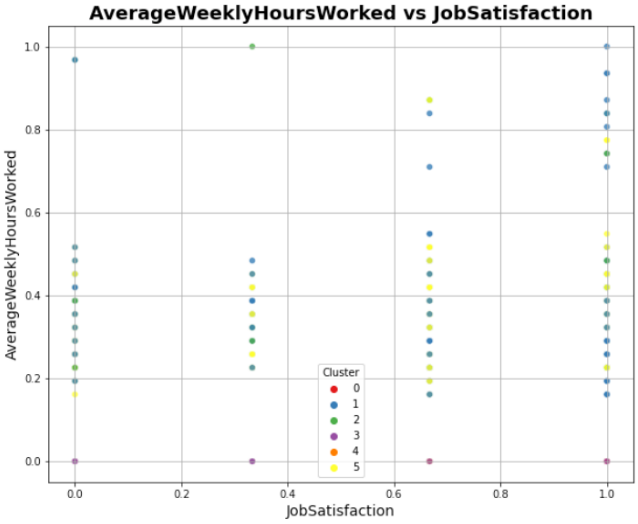
KMeans with **6 clusters**	# KMeans Model using the optimal k = 6
	km_1 = KMeans(n_clusters=6, random_state=30)
	df_1['Cluster'] = km_1.fit_predict(df_1)
	# Add the Resigned Column
	df_1['Resigned'] = df['Resigned']

Model 1: KMeans with 6 clusters (optiomal cluster as per W model evaluation methods):



Our model clusters our population dataset into six clusters, the above barplot visualizes the resignation rate of each cluster. Cluster 1 (blue) has the highest resignation rate of >30%, followed by cluster 5 (yellow) with >25%. This visualization is key as we can now further explore what characteristics lead to such a high resignation rate and what strategies management has to follow to reduce the turnover rate for employees identified in this group.



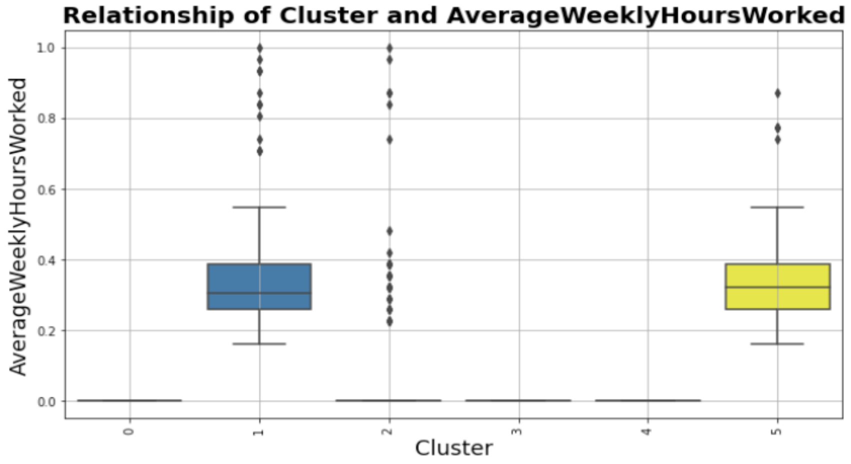


To identify the characteristics of cluster 1 and 5 (blue and yellow) we proceed with two strategies. **First, we identify feature pairs and color our values as per identified cluster** to gain a deeper understanding of where which group is sitting within the scatterplot. Secondly, we visualize the newly created column 'Cluster' in a boxplot to understand the distribution of this column and it's correlation to other variables.

Our first visualization MonthlyIncome vs YearsAtCompany confirms our previous assumption we made in the EDA, the majority of our risk clusters are employees within the lower income group and with lower years at the company. This suggests that early-career employees with lower pay are at higher risk of leaving.

Furthermore, the second relationship between AverageWeeklyHoursWorked and JobSatisfaction shows almost all employees who work more than 40 hours per week are part of Cluster blue and yellow. Both clusters are represented in all levels of JobSatisfaction (from low to high), therefore we can see that long work hours impact the happiness of some of the risk employees and some are still happy with working overtime (Appendix graph P3). This plot also emphazises that another cluster group is strongly represented in the area of lower JobSatisfaction, it is the green group which dominates that sector together with blue and yellow. Other clusters are working the average 40 hours

Strategie Two – Pairing our Cluster feature with other features

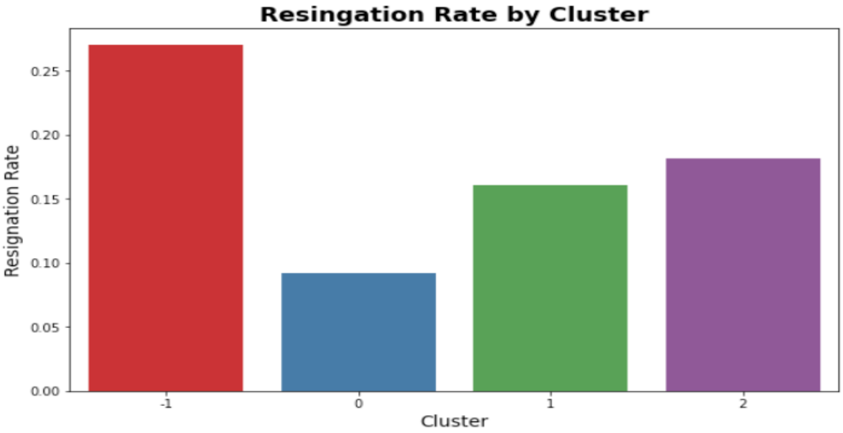


Next, we further explore if our previous insights are correct by plotting our cluster column with the AverageWeeklyHoursWorked feature in a boxplot to identify the distribution of all clusters. The above boxplot confirms our hypothesis that the risk clusters of employees with high turnover rate work way longer than the rest of the employee groups. Our management has to further investigate why this is the case and ensure that everyone is looking after themselves. One explanation could be that both clusters are (as previously expolored) fairly new employees who want to proof that they were a good hire or due to a high workload and a lot of internal pressure. Revolution C. has to ensure that every new person is not feeling overwhelmed by the work and must guarantee an easy start for new employees, as they the internal pressure can lead to higher turnover rates.

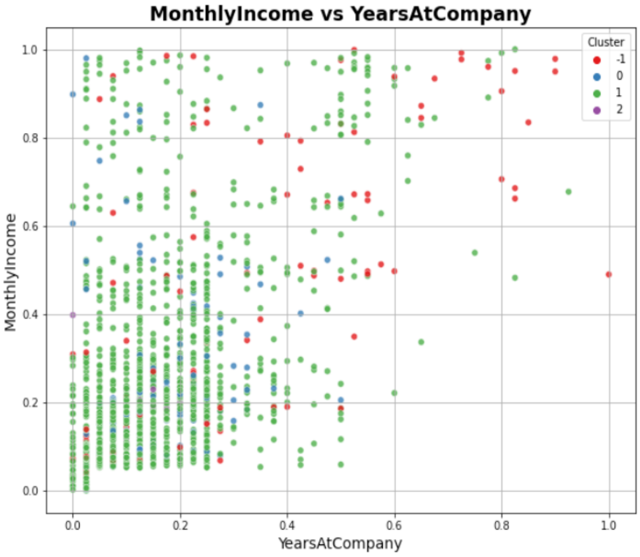
Model	**DBSCAN on all features incl. categorical features**

DBSCAN with **1.07 eps, min_samples=10**	# DBSCAN on our scaled df_2 dataset with the optimal eps value
	dbscan_df = DBSCAN(eps=1.07, min_samples=10)
	df_2['Cluster'] = dbscan_df.fit_predict(df_2) # Add the Resigned Column
	df_1['Resigned'] = df['Resigned']

Model 2 DBSCAN with an optimal eps of 1.08 and min_sample = 10



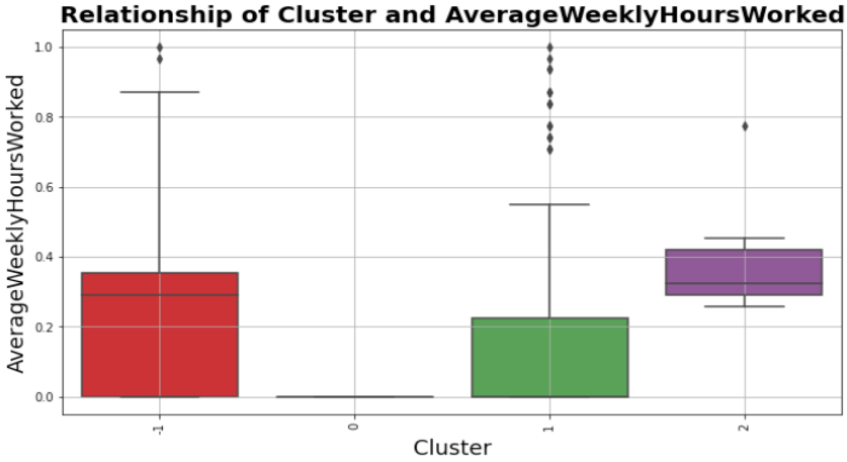
DBSCAN clusters our dataset into four groups, Cluster red (-1) has the highest resignation rate of >25%, cluster purple (2) comes second with >17.5% and cluster green (1) is right after that with a resignation rate of >15%. We will plot the same pair features as before to identify the groupings of our DBSCAN model and also to compare both model performance.



Our first visualization of MonthlyIncome vs YearsAtCompany colored by Clusters of employees with similar characteristics, groups almost all values to our cluster green (-1) regardless if this employee is new or old to the company or if this employee earns a high or low income per month. DBSCAN seem to struggle when creating clearly defined groups, suggesting that the displayed visualisation is not very insightful and misleading.

We can't gain any deeper understanding which group of employee is at higher risk of leaving the company, therefore our management can't make any decision what factors play a part in their high employee turnovers.

Strategie Two – Pairing our Cluster feature with other features



When comparing the DBSCAN cluster directly with the feature `AverageWeeklyHoursWorked`, we are also getting conflicting results of why employees quit their job. As per our previous barplot visualization, Cluster red (-1) represents the cluster with the highest resignation rate but our boxplot indicates that cluster purple (2) works longer average hours than all other clusters. Therefore, the feature AverageWeeklyHoursWorked or Overtime wouldn't have any relation with our modelled dataframe. The densities in our dataframe might not be optiomal for DBSCAN to create valuable insights and usable cluster.

Model Performance of KMeans vs. DBSCAN:

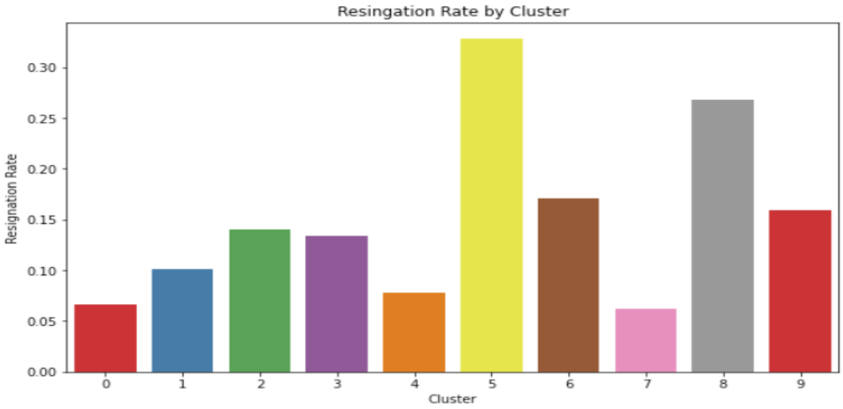
The analysis showed that KMeans outperformed DBSCAN in clustering employees into distinct groups with varying resignation rates. KMeans was more effective in identifying key factors contributing to attrition, as we saw for example in our scatterplot of MonthlyIncome vs YearsAtCompany. DBSCAN grouped almost all values into one cluster as this model is very sensitive to the choice of parameters like eps, which represents the distance between points in a cluster. The model is less effective on high-dimensional datasets with a large set of features as this causes the model to struggle when clustering the data into meaningful portions.

5. Discussion

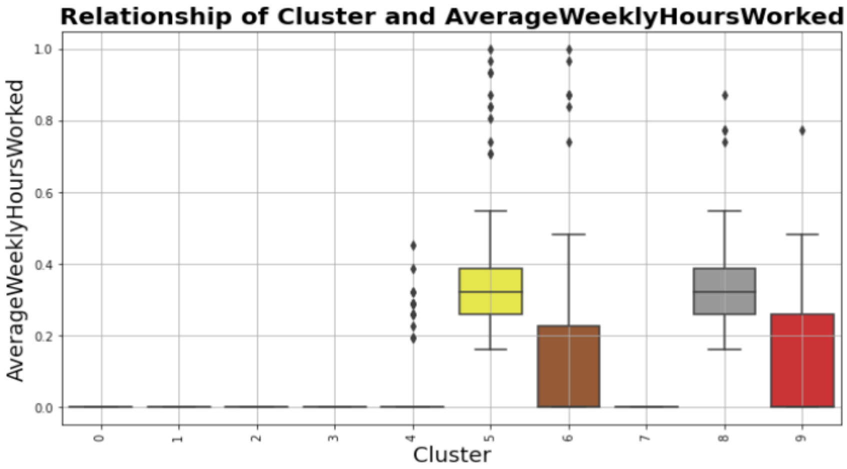
As we saw before, we discussed how KMeans is better suitable to cluster features in distinct groups when working with high-dimensional datasets with a large set of features. To confirm that our used parameters weren't the main reason for incorrect results, we perform one more model for KMeans with a higher k-value (number of clusters) and an additional DBSCAN model with adjusted eps and min_samples parameter to generate more cluster than the previous four. The below will show if those adjustments had any impact on the model performance and it will also disclose if our management can win further insights of how to reduce staff turnover within the company.

Model	**KMeans on all features incl. categorical features**
KMeans with **10 clusters**	km_1_higher_k = KMeans(n_clusters=10, random_state=30) df_1_higher_k['Cluster'] = km_1_higher_k.fit_predict(df_1_higher_k)

Model 3: KMeans with 10 clusters



The above visualization shows that even if we adjust our k value, our model still produces distinct clusters with different resignation rates. We can check if this newly created model changes anything on our previously identified issues Revolution C. is facing – that employees with an above average work hours are more likely to quit. All clusters with an alarming rate of resignation rate should be reprented when plotting our boxplot of AverageWeeklyHours by Clusters.



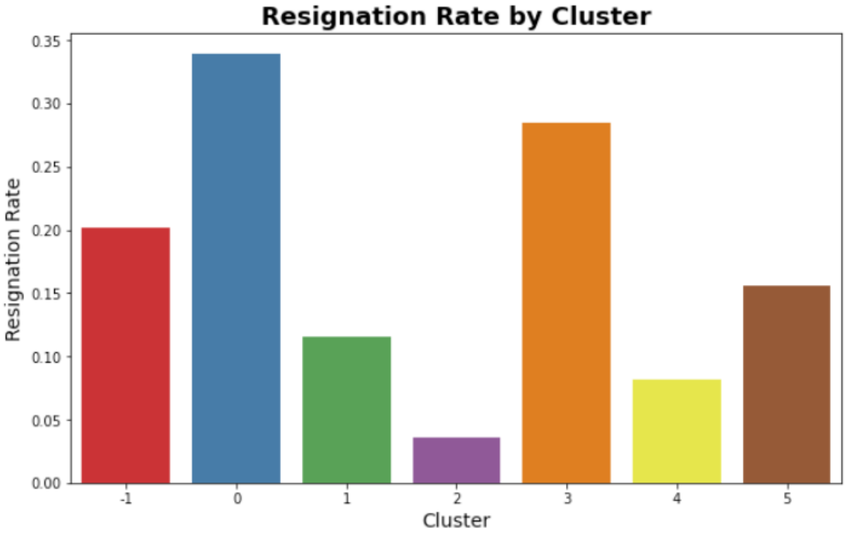
As expected, all clusters with an above average resignation rate are represented in our boxplot. This confirms that the KMeans model performed great even with an higher number of clusters. Our management has to focus on improving the work environment for people who are forced to work Overtime, or they will loose their motivated staff.

Let’s explore if an adjusted DBSCAN model improved the models performance.

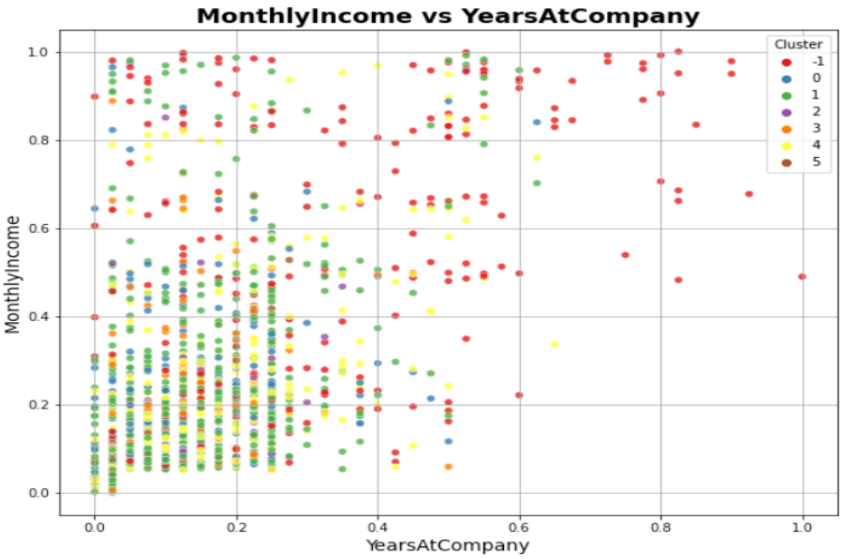
Model	**DBSCAN on all features incl. categorical features**
DBSCAN with **0.95 eps, min_samples=15**	dbscan_df = DBSCAN(eps=0.95, min_samples=15) df_2_lower_eps['Cluster'] = dbscan_df_loweps.fit_predict(df_2_lower_eps)

Model 4: DBSCAN with an lower eps of 0.95 and min_sample = 15

This section checks if the DBSCAN model improved its performance after we lowered the eps value and increased the min_sample. First, we plot the resignation rate against all clusters again.



The new model created more clusters this time, where cluster blue (0), orange (3) and red (-1) represent the groups with the highest resignation rate. We can plot the scatterplot vs YearsAtCompany again to identify if one cluster still dominates.



The model does indeed have multiple cluster within the scatterplot now, without one cluster being too dominant. Since we know that the turnover rate is most significant for employees who are new at the company and part of a lower monthly income group, we need to investigate the lower left corner of our visualisation. All clusters are represented equally without one indicating that there is a relationship between new underpaid employees and high turnover rate, which indicates that even after the eps and min_samples adjustments DBSCAN is still struggling to create meaningful clusters.

6. Conclusion

Our primary goal of this report was to identify key factors of high employee turnover at Revolution Consulting and how the management can take steps to actively reduce those visualised resignment rates within clusters of employees with similar characteristics.

We started by displaying various graphs and feature pairs in the EDA of this analysis, where we gained a deeper understanding of the features we are working with. It was evident that a substantial proportion of resignations occurred among employees with lower income and shorter tenures. These insights were crucial for the model building process where we clustered our dataset into groups with different resignation rates and similar character traits.

We proceed with two modelling techniques, KMeans and DBSCAN and highlighted what clusters of employees management has to focus on and what groups are not impacted by the high turnover rate. After we plotted different feature pairs with the newly identified column ‘Clusters’, we e.g. identified the significant impact above average working hours and overtime contributed towards a high resignation rate.

Conversely to that, we also showcased how the DBSCAN model did not perform well when handling a high-dimensional dataset with large feature sets, as this model approach is clustering its data based on value density. The scatterplots of DBSCAN failed to distinctly separate the data into meaningful groups.

Based on our findings, the management of Revolution C. should consider the following key recommendations that can decrease the employee resignation rate:

1. Enhancing Compensation: Our analysis highlighted that lower-income employees are more prone to resign. Revolution C. needs to consider revising its salary structure to be more competitive, especially for new employees and those in lower-paying roles.

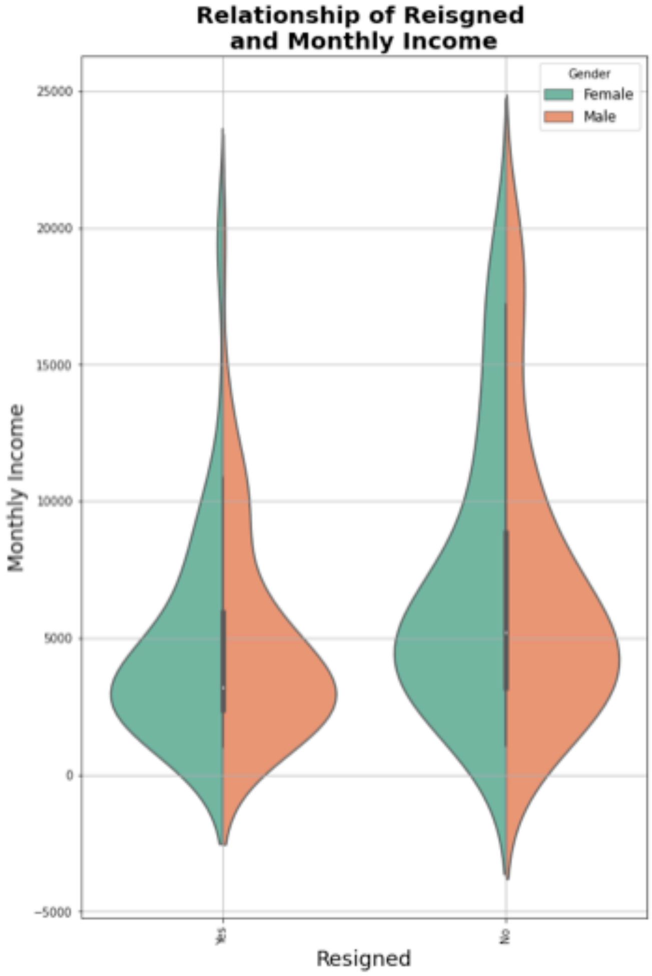
2. Promoting Work-Life Balance: Our ‘Relationship of Cluster and AverageWeeklyHoursWorked’ graph showed that employees with above average working hours face a higher risk of resigning. The management should prioritze programms that promote a healthier work-life balance such as flexible work schedule, working from home etc.
3. Supporting younger employees: Our ‘Distribution of Age’ showed that the employees of the company are farily young with its mean of 36 years. As younger people are more likely to resign if no career growth is achievable, our management should provide clear mentorship programs or career development opportunities and reward attendants.
4. Improving Job Satisfaction: Our management should regulary assess job satisfaction levels through surveys and face-to-face meetings to proactively take action if any issues that occur that lower the employees job satisfaction.

7. References

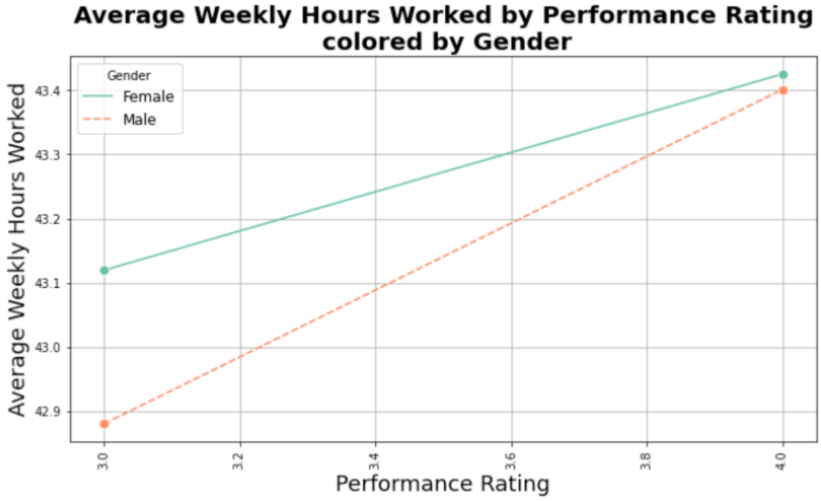
- Scikit-learn (2019). *sklearn.preprocessing.MinMaxScaler — scikit-learn 0.22.1 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html [Accessed 16. Sep. 2024].
- Hackman, M. (2024). *Mapping string categories to numbers using pandas and numpy*. [online] Stack Overflow. Available at: https://stackoverflow.com/questions/43882652/mapping-string-categories-to-numbers-using-pandas-and-numpy [Accessed 19 Sep. 2024].
- Gupta, A. (2019). *Elbow Method for optimal value of k in KMeans*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/ [Accessed 16. Sep. 2024].
- scikit-learn. (2024). *Selecting the number of clusters with silhouette analysis on KMeans clustering*. [online] Available at: https://scikit-learn.org/1.5/auto_examples/cluster/plot_kmeans_silhouette_analysis.html [Accessed 16 Sep. 2024].
- Jjunk (2024). *How can I find just the kth-nearest neighbor?* [online] Stack Overflow. Available at: https://stackoverflow.com/questions/64573623/how-can-i-find-just-the-kth-nearest-neighbor [Accessed 17 Sep. 2024].
- Hailu, T.T., Yu, J. and Fantaye, T.G. (2020). Intrinsic and Extrinsic Automatic Evaluation Strategies for Paraphrase Generation Systems. *Journal of Computer and Communications*, 08(02), pp.1–16. doi:https://doi.org/10.4236/jcc.2020.82001 [Accessed 19 Sep. 2024].

8. Appendix

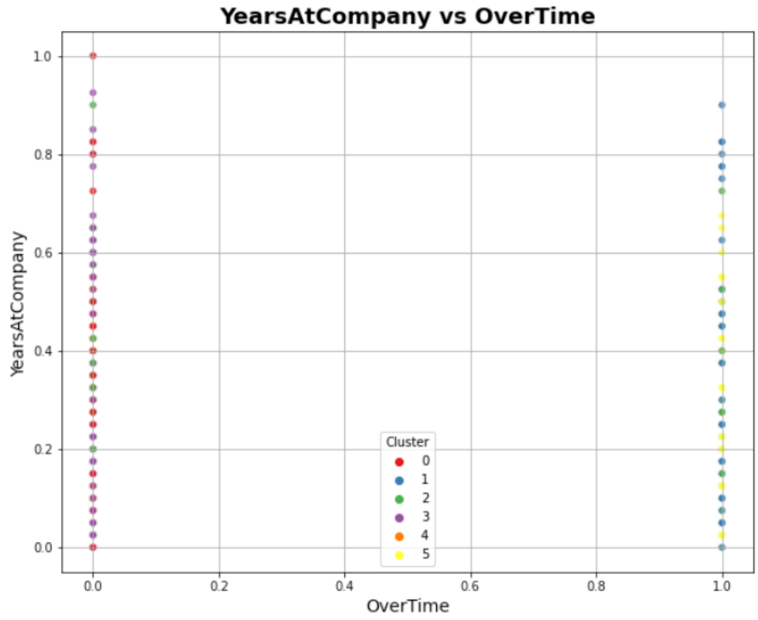
P1: Violin Plot EDA



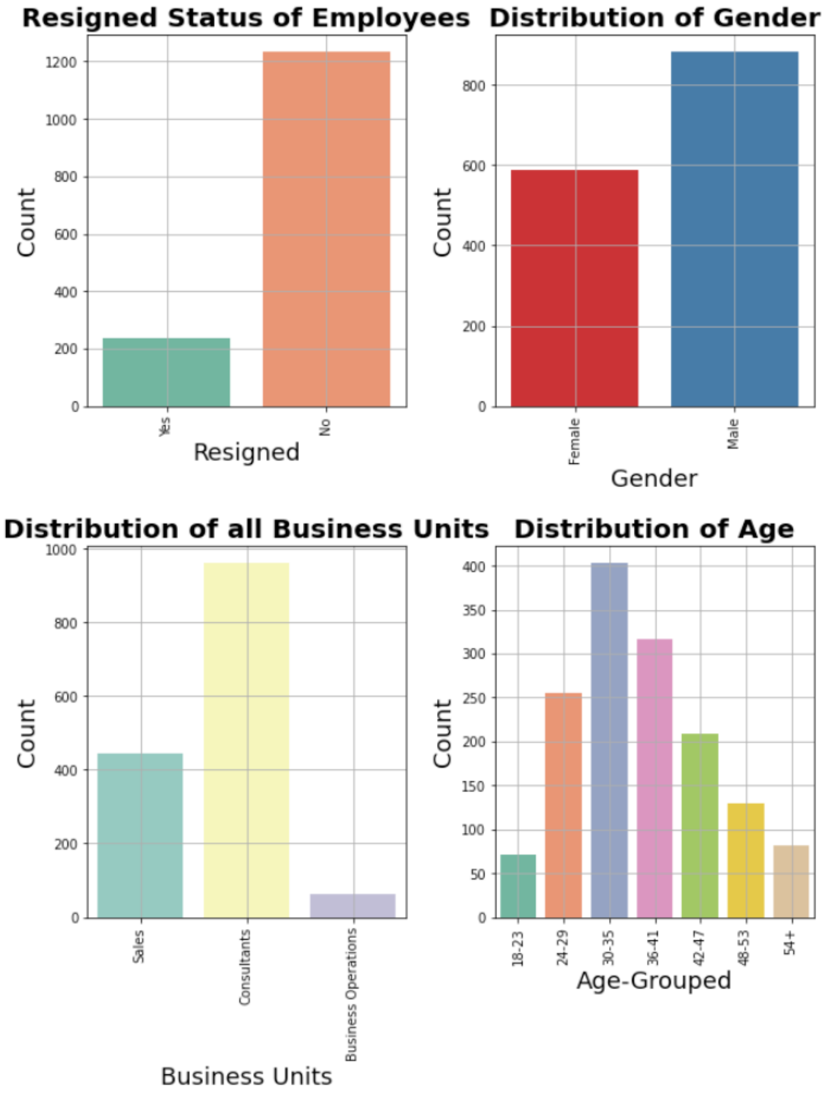
P2 Lineplot EDA



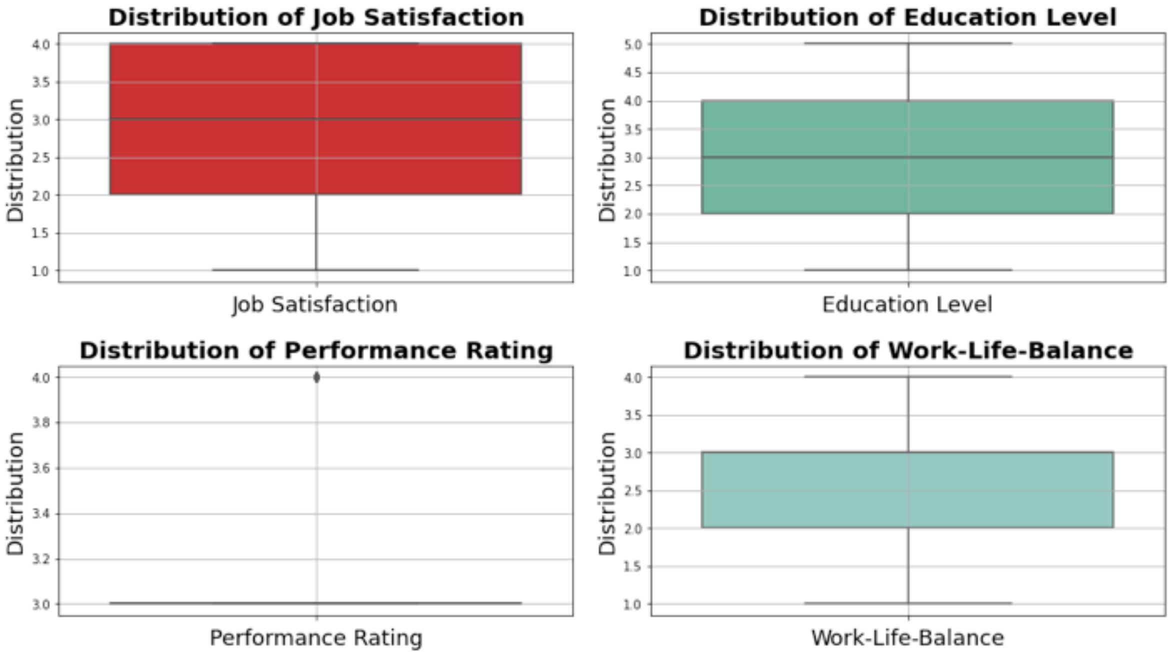
P3 OverTime KMeans on optimal k = 6



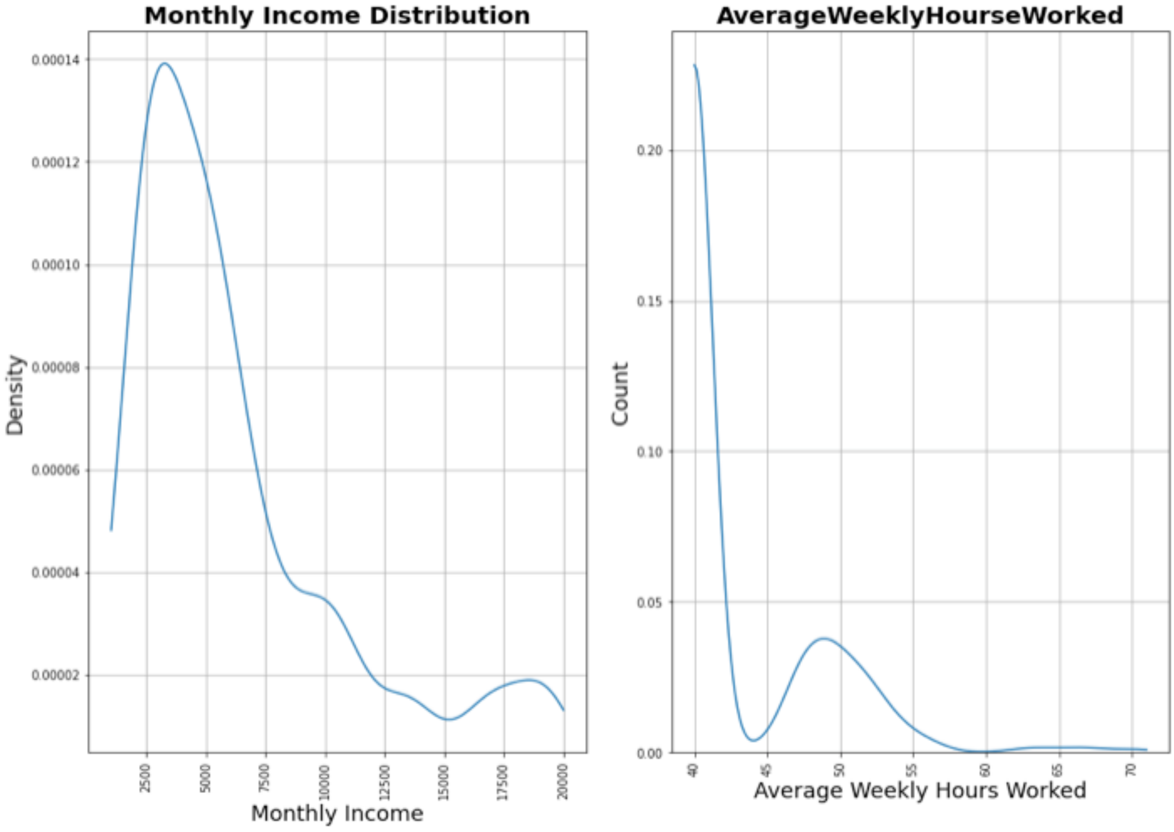
P4 Plots of all column specific EDA plots - Countplots



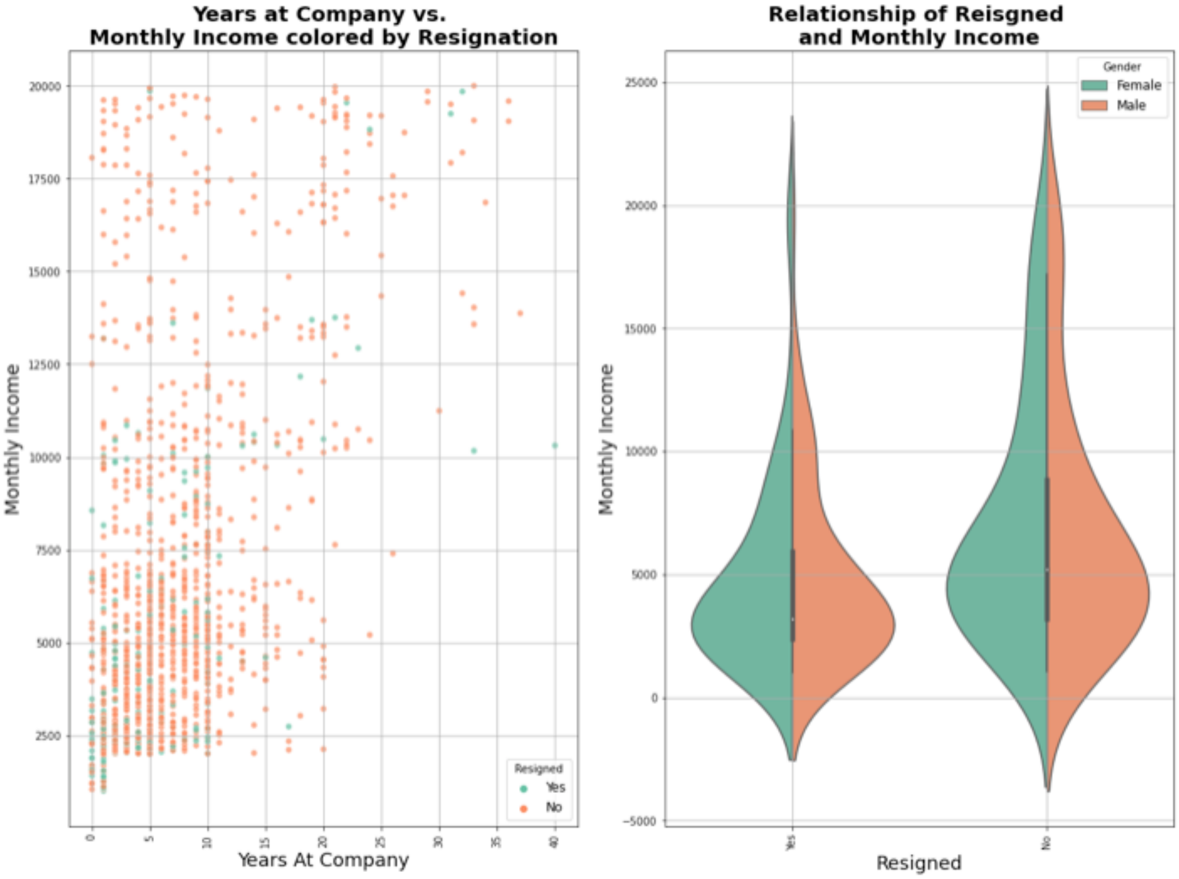
P5 Plots of all column specific EDA plots – Boxplots



P6 Plots of all column specific EDA plots – Density plots

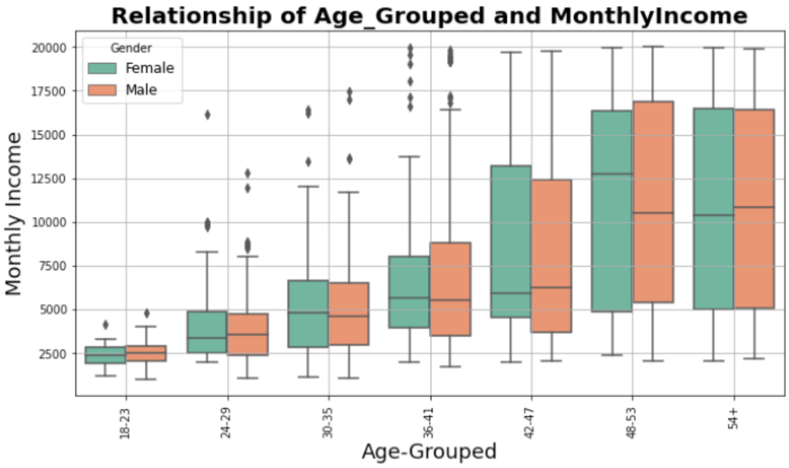


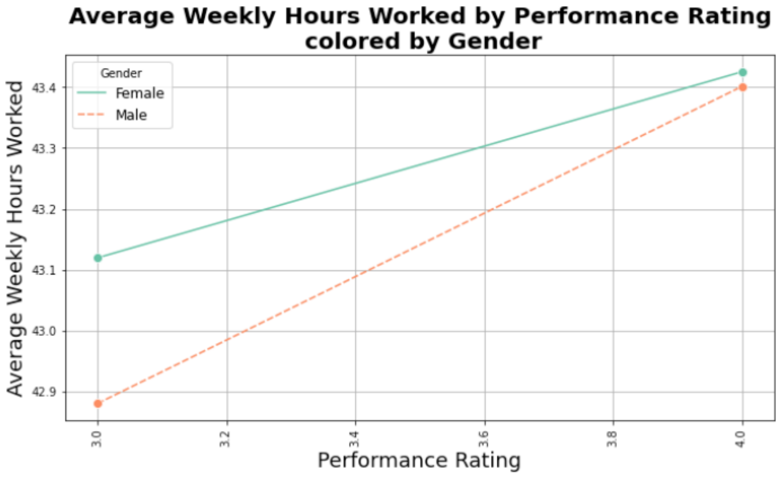
P7 Feature pairs – YearsAtCompany vs MonthlyIncome by Resignation and Gender


Years at Company vs. Monthly Income colored by Resignation

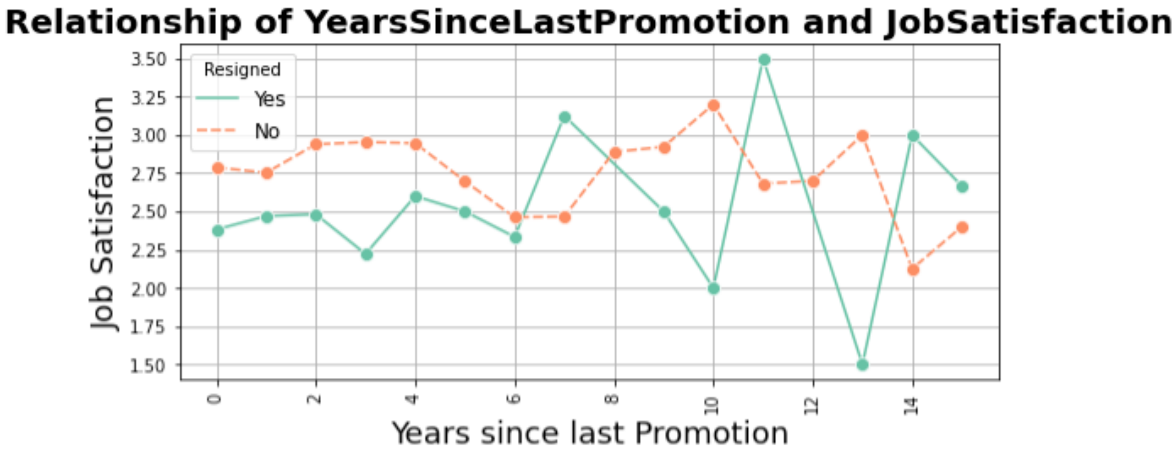
Relationship of Reisgned and Monthly Income

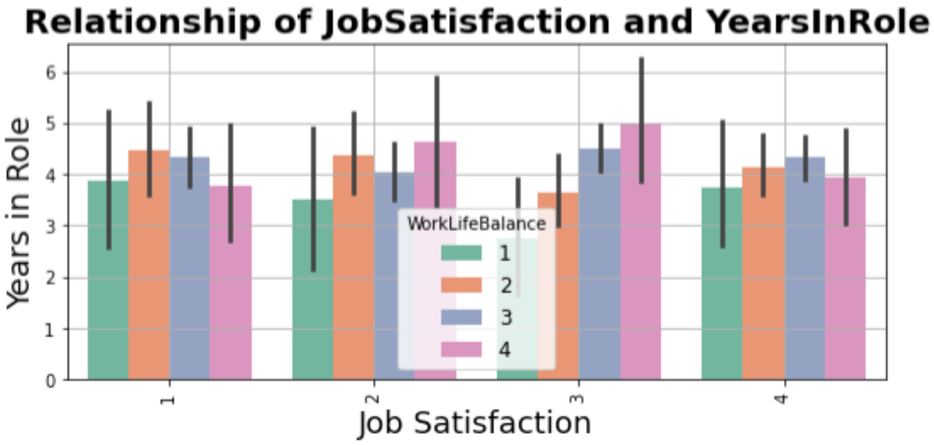
P8 Feature pairs – Relationship of Age_group and MonthlyIncome


Relationship of Age_Grouped and MonthlyIncome

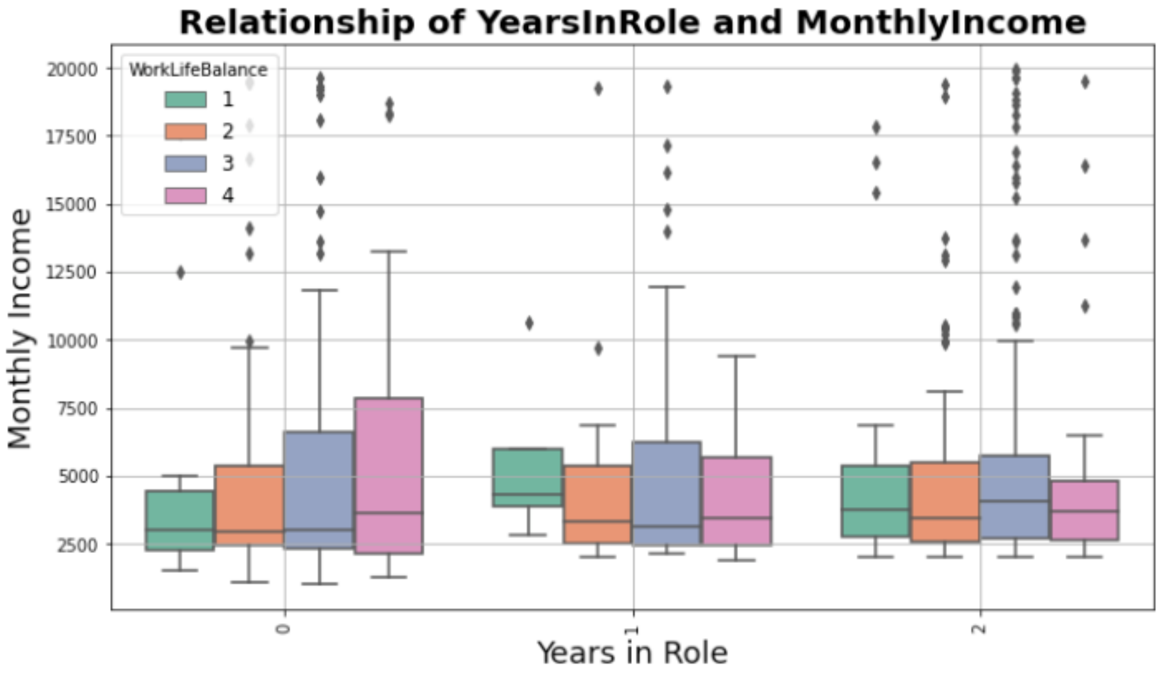

Average Weekly Hours Worked by Performance Rating colored by Gender

P9 Feature pairs – Employees who are likely to churn vs Satisfied employees

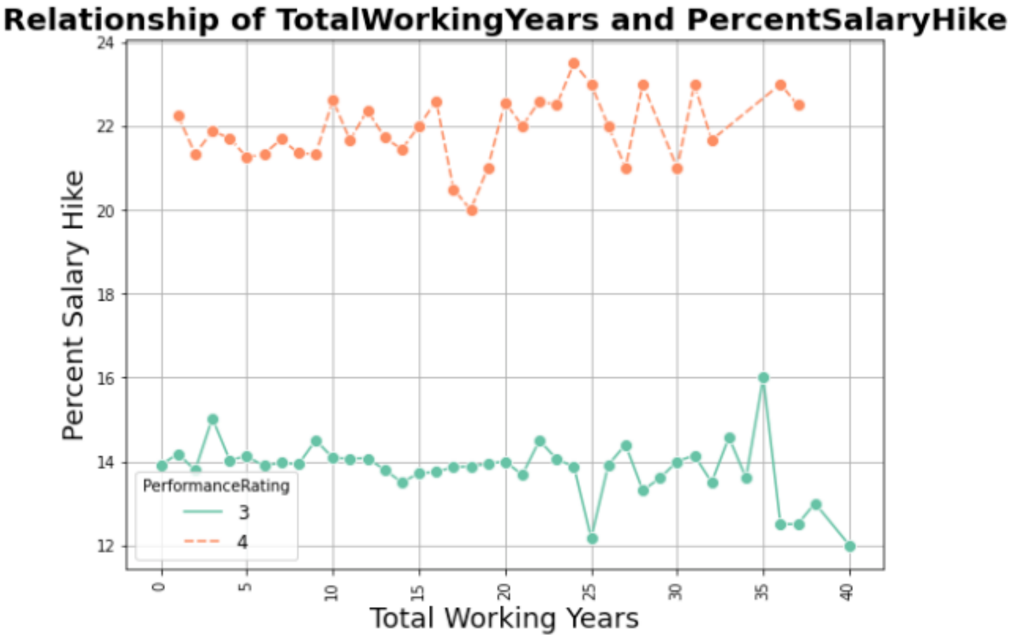


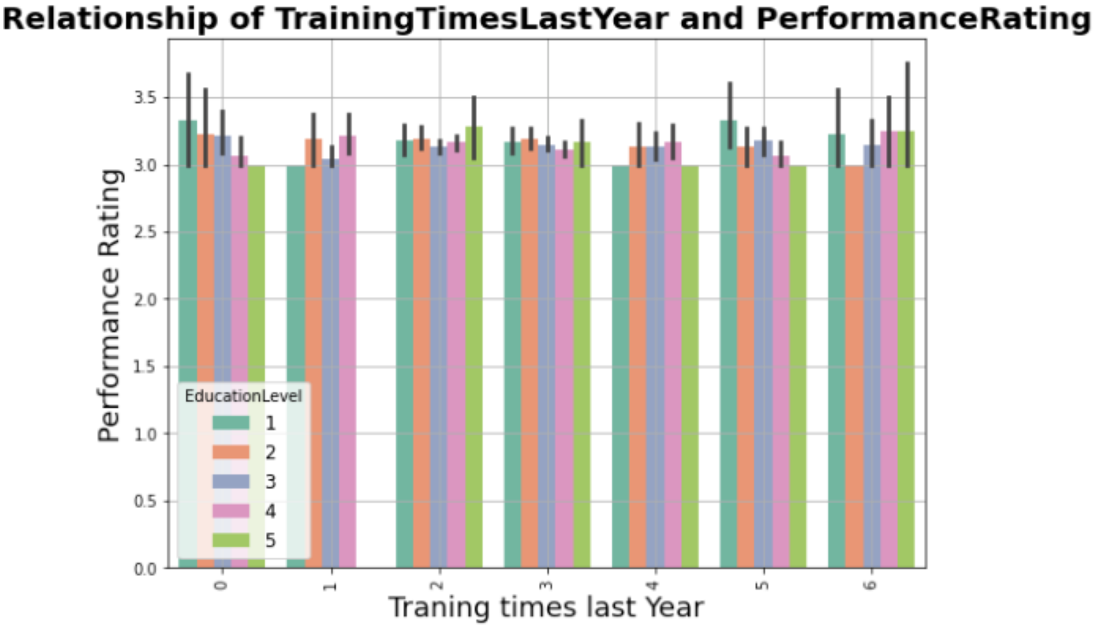


P10 Feature pairs – Relationship of YearsInRole and MonthlyIncome

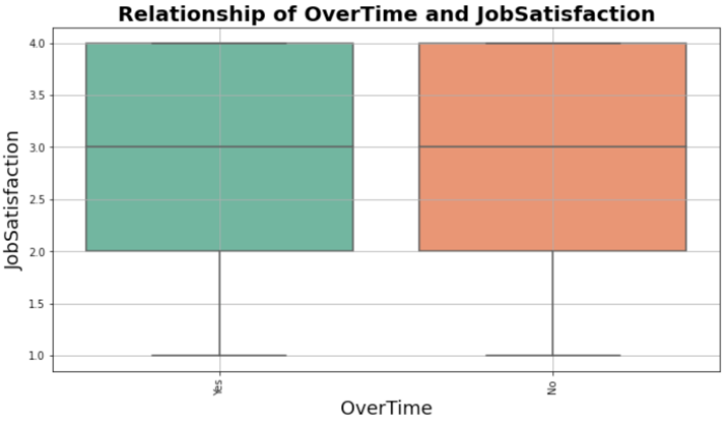


P11 Feature pairs – Engagement within the company

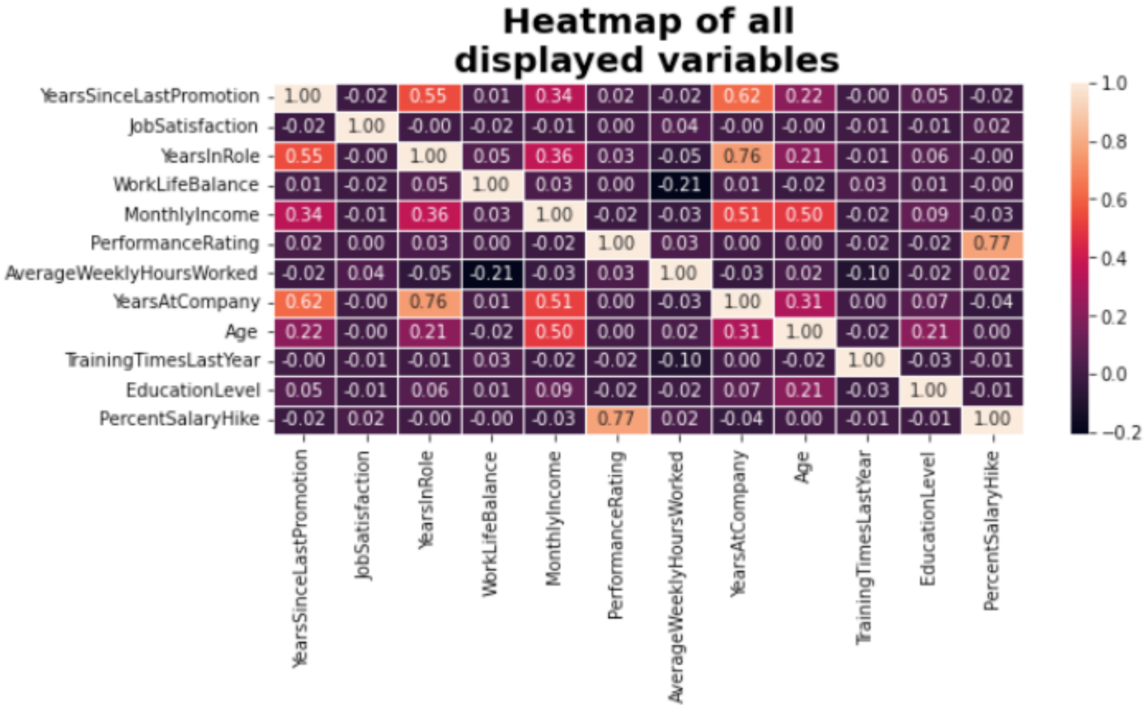




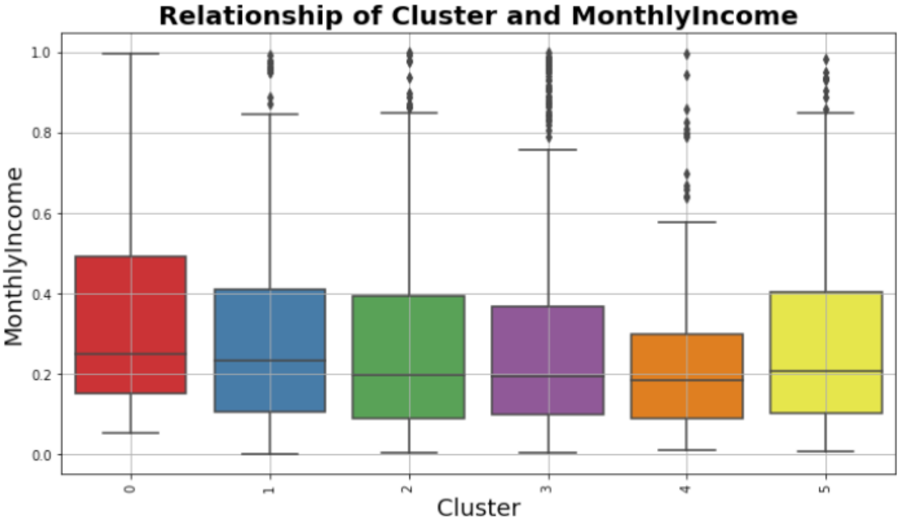
P12 Feature pairs – Correlation between JobSatisfaction and OverTime



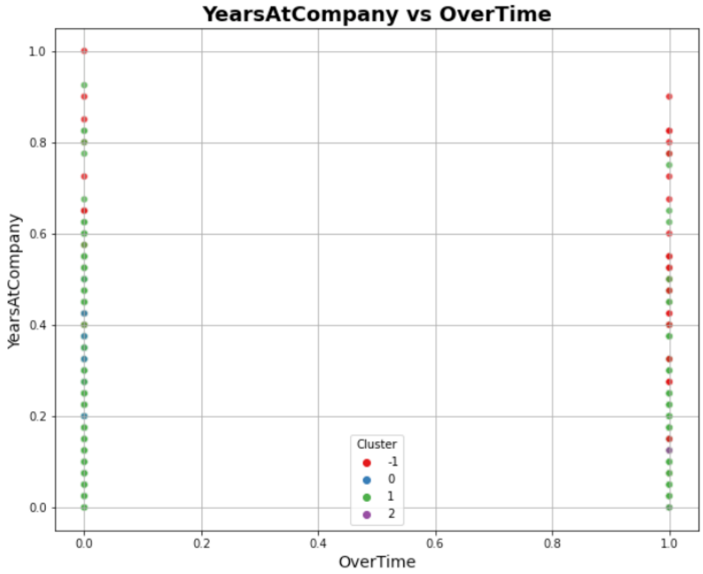
P13 Correlation of all analysed features



P14 KMeans – Relationship of Cluster and MonthlyIncome



P15 DBSCAN– YearsAtCompany vs OverTime


YearsAtCompany vs OverTime
YearsAtCompany
OverTime
Cluster