Assessment 2 Template: Data modelling report

Title page

Data Modelling Report, Thaddeus Lee, Revolution Consulting.

Table of contents

Introduction

Revolution Consulting, an IT Consulting firm, has been facing issues with declining work quality produced by its consultants. The company is facing higher turnover in consultants and newer consultants are not as knowledgeable or skilled. Following the initial data preparation report, management speculate that resignations are driven by factors related to employee renumeration, gender pay gap, age groups and employees likely to leave due to a lack of career progression.

This report documents the process for data exploration, data modelling and model evaluation with the goal of identifying the underlying factors driving employee resignations, and identifying employees that are likely to resign. Solutions will then be recommended to management based on the report's findings.

Features overview

Briefly outline the features from your dataset using the table below.

Feature name	Number of unique values	Type	Description
AGE	43	Interval/Ratio	Age of employee in years
RESIGNED	2	Nominal	Denotes if employee has left job.
BUSINESSTRAVEL	3	Nominal	Notes if employee travels, doesn’t travel or travels rarely for work
BUSINESSUNIT	3	Nominal	Role in company as either: Sales, Consulting and Business Operation

EDUCATIONLEVEL	5	Ordinal	Tiers labelled 1, 2, 3, 4, 5 in ascending order
GENDER	2	Nominal	Male and Female
JOBSATISFACTION	4	Ordinal	Tiers labelled 1, 2, 3, 4 in ascending order
MARITALSTATUS	3	Nominal	Nominal. Responses are: Single, Divorce, Married
MONTHLYINCOME	1349	Interval/Ratio	Salary earned per month
NUMCOMPANIESWORKED	10	Interval/Ratio	Number of different companies worked with.
OVERTIME	2	Nominal	Notes if employee works overtime or not 'Yes', 'No', binary response.
PERCENTSALARYHIKE	15	Interval/Ratio	Percent salary increase
PERFORMANCERATING	2	Ordinal	Ordinal tiered rating in ascending order.
AVERAGEWEEKLYHOURSWORKED	23	Interval/Ratio	Average hours worked in a week.
TOTALWORKINGYEARS	40	Interval/Ratio	Number of years employee has been working (even outside of company)
TRAININGTIMESLASTYEAR	7	Interval/Ratio	Number of times employee trained in the last year.
WORKLIFEBALANCE	4	Ordinal	Tiers labelled 1, 2, 3, 4 in ascending order
YEARSATCOMPANY	37	Interval/Ratio	No of years worked in company
YEARSINROLE	19	Interval/Ratio	No. of years spent in job
YEARSSINCELASTPROMOTION	16	Interval/Ratio	Years since last promotion
YEARSWITHCURRMANAGER	18	Interval/Ratio	Years under same manager

Methodology

Overview

The goal of this report is to provide further analysis into the data collected by Revolution Consulting on its employees and utilise data exploration and data modelling to provide further analysis into the factors behind their staff retention issues.

Management also speculate that resignations are driven by employee renumeration, gender pay gap, age groups and employees likely to leave due to a lack of career progression.

This report aims to identify the factors driving resignations including the possible factors above, as well as identifying satisfied employees that don't require focus, new employees that may leave due to concerns about pay and Work Life Balance, and possible cultural issues that might be driving resignations.

This report will document the process of data exploration, identifying relationships among the data and how it correlates with the research goal. We will then select the features that best correlate to resignations and use clustering data modelling methods to group employees by resignation status to examine features of resigning employees and identify potential resignations.

The usefulness of the models will be evaluated and the data modelling results will inform recommendations to management.

Process

We start with data exploration to gain a better understanding of the data and the relationships between all data features. As we are trying to understand the reasons behind Revolution Consulting's retention issues, we use statistical analysis to explore the relationship of each data feature for correlations to our target feature which is the 'RESIGNED' column in the data set.

Feature Selection is then used, selecting the features that have a strong correlation to our target feature, while avoiding any potential overlap. We then utilize these selected features in clustering data modelling methods KMeans and DBSCAN and evaluate each model for optimal selection of hyper parameters for training and model usefulness.

Evaluation strategy

For KMeans and DBSCAN clustering methods, the models are evaluated for the optimum parameters for our dataset using the evaluation methods below. As we are conducting our data modelling with the view of examining the features for employees with a resignation status, we expect K to be equal to 2, or 2 clusters. We will utilise Intrinsic and Extrinsic evaluations to evaluate our models which are suited for unsupervised learning.

Intrinsic Evaluations:

Intrinsic evaluations evaluate quality of clusterings based on an internal criterion without labels. The following will be utilised in our report.

Within Cluster Sum of Squares (WCSS): WCSS is the sum of squared distance between each point and the centroid in a cluster. This evaluation determines the optimal amount of clusters (k) given a WCSS score. Generally, a relatively small number of clusters and a low WCSS score is a good result.

Silhouette Coefficient Method: A measure of the tightness and separation of the clusters. Average Silhouette width provides an evaluation of clustering validity and could be used to select an appropriate number of clusters. The silhouette score is between -1 to 1 and a higher score is better.

Extrinsic Evaluations:

Evaluates clustering result based on the degree the result will assist us in solving a subsequent problem. Extrinsic evaluation used in this report is the Adjusted Rand Index.

Adjusted Rand Index: Measures the agreement of two clusterings or partitions of the same dataset by calculating the agreement ratio of the pairs that fall in the same or different groups. The Rand index is bound between the values of 0 and 1, where 1 indicates full agreement. Adjusted Rand Index (ARI), adjusts the Rand Index for chance groupings , that the value for a random clustering will be close to 0.

Data modelling and exploration

Data Exploration and Feature Selection:

Data exploration is used to find the existence of strong relationships among the features in a dataset that can be driving resignations.

We begin our data exploration by producing a Scattermatrix with hue = RESIGNED column (Appendix 1) with the following features: AGE, EDUCATION LEVEL, JOBSATISFACTION, MONTHLYINCOME, NUMCOMPANIESWORKED, OVERTIME, PERCENTSALARYHIKE, PERFORMANCERATING, AVERAGEWEEKLYHOURSWORKED, TOTALWORKINGYEARS, TRAININGTIMESLASTYEAR, WORKLIFEBALANCE, YEARSATCOMPANY, YEARSINROLE, YEARSSINCELASTPROMOTION, YEARSWITHCURRMANAGER.

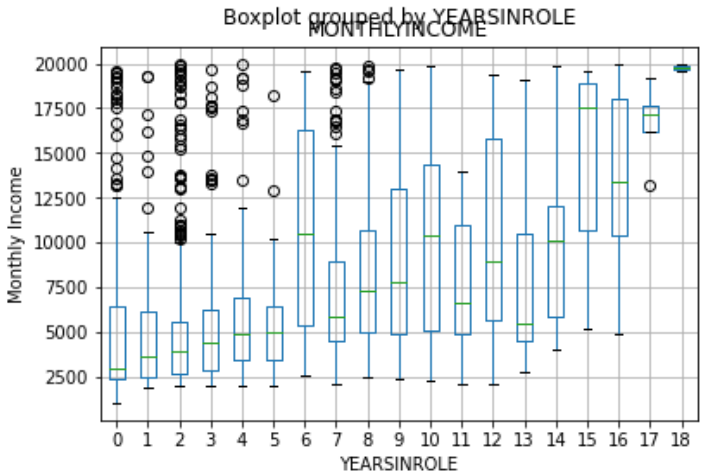
We can observe the following:

Under features YEARSWITHCURRMANAGER, YEARSSINCELASTPROMOTION, YEARSINROLE, TOTALWORKINGYEARS and YEARSATCOMPANY. We can observe that newer employees, within the first 5 years, work longer hours compared to more tenured employees. We can also observe that employees that resign tend to not have been with the company for a long time.

WORKLIFEBALANCE compared to MONTHLYINCOME, we can observe that there is a concentration of employees with lower incomes that have resigned, regardless of rating. Compared with AVERAGEWEEKLYHOURSWORKED, employees who have given a rating of 1 and 2 work the most hours compared to those who have given a rating of 3 or 4.

PERCENTSALARYHIKE and PERFORMANCERATING: Relationship exists where highest performers, get higher salary increases.

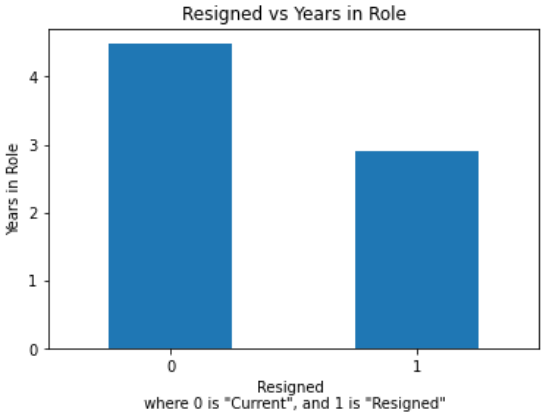
Compared to YEARSATCOMPANY, Increases to MONTHLYINCOME are staggered, with no increases to new employees for a number of years. Highest MONTHLYINCOME is given to employees who've been with the company longest. From the graph below, we can see there's very little change in monthly income for more than 5 years.



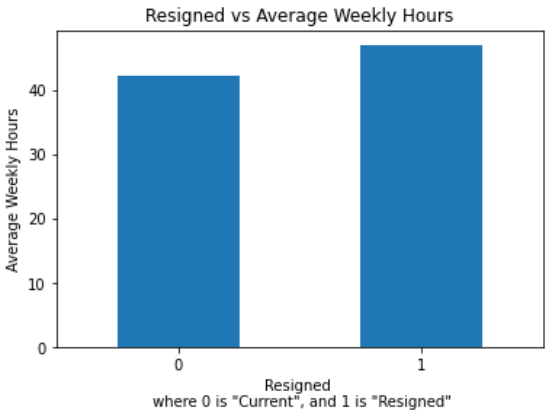
YEARSWITHCURRMANAGER, YEARSSINCELASTPROMOTION and YEARSINROLE: We can see similar observations among these columns with the other data features.

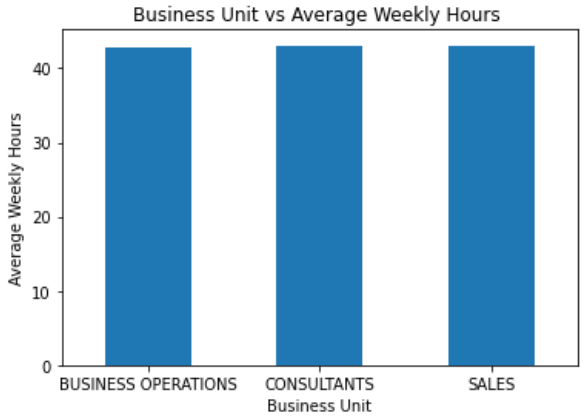
TOTALWORKINGYEARS and YEARSATCOMPANY: we can see very similar observations across the other features in the dataset.

YEARSWITHCURRMANAGER, YEARSSINCELASTPROMOTION, YEARSINROLE, TOTALWORKINGYEARS and YEARSATCOMPANY: We can observe that newer employees, within the first 5 years, work longer hours compared to more tenured employees. We can also observe that employees that resign tend to not have been with the company for a long time. From below, we can see on average employees that resign are with the company for 5.13 years.



Average Weekly hours among resigned employees is 46.9 hours, compared to 42.3 for current employees. We can also see that the people who tend to do overtime are likely to resign. Business Unit vs Average Weekly Hours, Consultants work more hours on average. Business Operations = 42.84, Consultants = 42.06, Sales 43.02.





When looking at WORKLIFEBALANCE and JOBSATISFACTION, compared to resigned the median is 3 for both. Mean for JOBSATISFACTION (Appendix 2) is 2.47 for resigned and 2.78 for current employees. Mean for WORKLIFEBALANCE (Appendix 3) for resigned is 2.66. and 2.78 for current

When looking at the AGE data feature compared to resign status, the mean of resigned is 37.56 compared to 33.6, median of resigned is 32 compared to 36 for current employees (Appendix 4)

Looking at Gender we see that on average, men are paid lower than women (Appendix 5). FEMALE mean income = 6686.566327, MALE mean income = 6380.507937.

Furthermore, looking at the break-up of resignations by Gender (Appendix 6), men have a higher rate of resigning.

Feature Selection:

For the purposes of data modelling, our analysis into other features such as BUSINESSTRAVEL, BUSINESSUNIT, EDUCATIONLEVEL, MARITALSTATUS, NUMCOMPANIESWORKED, OVERTIME, PERCENTSALARYHIKE,PERFORMANCERATING, TRAININGTIMESLASTYEAR did not reveal any significant correlation to an employees' RESIGNED status(Appendix 6). Features such as OVERTIME and BUSINESSTRAVEL overlapped with AVERAGEWEEKLYHOURSWORKED.

Given the correlations we can observe in our data exploration, and the considerations asked of management. The best features to be incorporated into our data set for modelling are: AGE, JOBSATISFACTION, MONTHLYINCOME, AVERAGEWEEKLYHOURSWORKED, TOTALWORKINGYEARS, WORKLIFEBALANCE and YEARSATCOMPANY are selected.

Data Modelling (Clustering):

Clustering, an unsupervised method of data modelling where objects are grouped together in a certain way, is used. KMeans and DBSCAN methods were used and are briefly described below.

KMeans:

A simple clustering algorithm that defines clusters by partitioning all obersvations into groups. Clusters are formed in a way where each observation is grouped by their nearest mean. KMeans only requires that the number of clusters be set as an input. The model runs itself multiple times until the sum of squares from points to the assigned cluster centres is minimised. It works by the following procedures:

1. Initialise Cluster Centroids.
2. Assign points to clusters whose centroid is nearest.
3. Re-compute means centroids
4. Repeat step 2 and 3 until convergence (where points don't move between clusters).

KMeans advantages are due to it's robust and efficient nature; guarantees convergence and provides the best results when the data sets are distinct and well separated. It does come with disadvantages being that: Number of clusters need to be specified in advance and an incorrectly chosen one will give poor results; it is unable to identify non-spherical or overlapping data; is unsuitable for clustering data with differing sizes and density; is unable to handle outliers; Does not scale well with large numbers of dimensions.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN):

A density-based clustering algorithm that finds areas in data where there is high density of observations versus areas with low observations. Unlike KMeans algorithm, it can sort data into clusters of varying shapes. Density based clusters are formed by measuring distance between points and centroid of each cluster.

Advantages of DBSCAN are: Number of clusters do not need to be predefined by user, can find arbitrarily shaped cluster, is robust to outliers, insensitive to ordering of points in a database,

designed for use with databases that can accelerate region queries and parameters ‘Eps’ and ‘minPts’ can be set by domain expert.

Disadvantages are: not entirely deterministic. The quality of DBSCAN relist on distance measure and with high-dimensional data, there is little difference I the distances between different pairs of samples. DBSCAN is not well suited to data sets with large differences in densities because minPts-Eps combination cannot then be chosen appropriately for all clusters.

DBSCAN requires that 2 parameters, Eps and MinPts be specified by the user. ‘Eps’ specifies how close points should be to each other to be part of a cluster. MinPts specifies the min number of points to form a dense region.

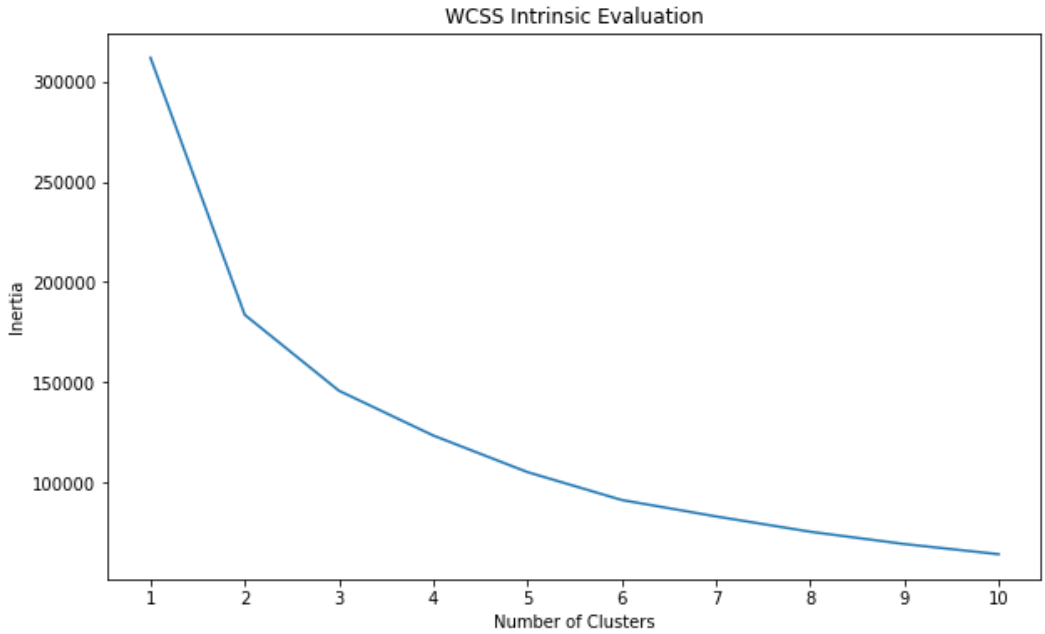
In setting Eps, the value can be chosen by using a K-distance graph, plotting the distance to the k = minPts nearest neighbour. In setting Minpts, this can generally be set to the number of dimensions in a data set plus 1. However, data sets with noise can benefit from higher values. Given the amount of data features and data involved, it is suggested to use Minpts of the number of features multiplied by 2. In our case we have selected 7 features, so MinPts = 14 is used for our DBSCAN.

Results

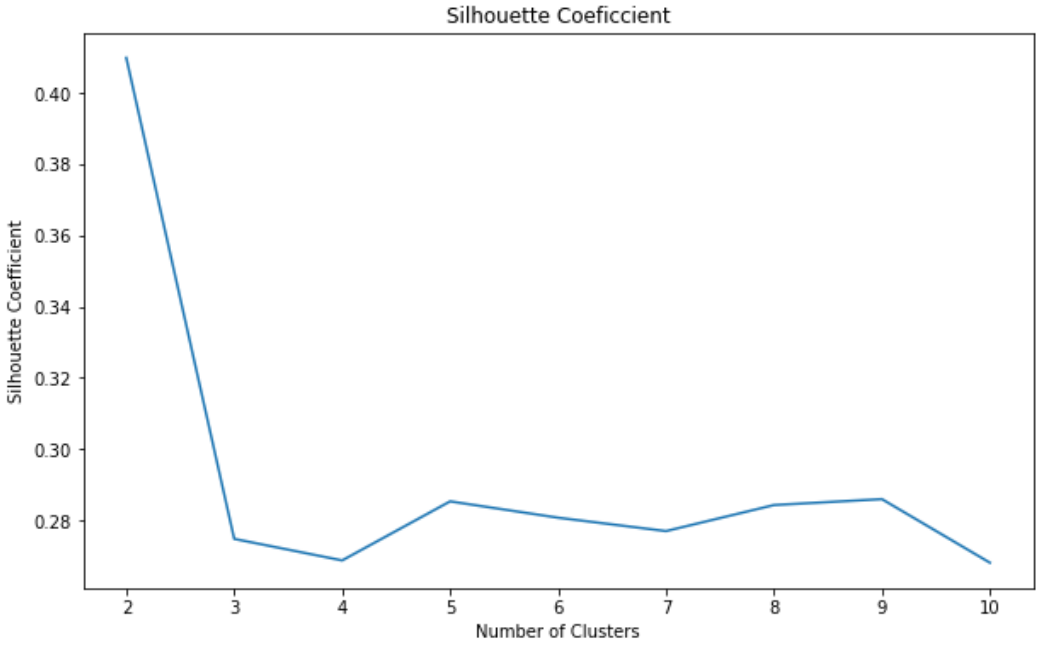
Model	**Within Cluster Sum of Squares (WCSS)**	**Silhouette Coefficient**	**Adjusted Rand Index**
KMeans **K= 2**	Optimum result with: Plot bend at k = 2, WCSS = 180000	Max point at k=2, Silhouette coefficient = 0.41	For K = 6, ARI = 0.0295
DBSCAN **MinPts = 14** **Eps = 7.5**		Silhouette Coefficient = 0.4566111184563737	
DBSCAN **MinPts = 14** **Eps = 8.24** **(Most Optimum)**		Silhouette Coefficient = 0.43396087944129746	
DBSCAN **MinPts = 14** **Eps = 79**		Silhouette Coefficient = 0.4345867733022825	

KMeans:

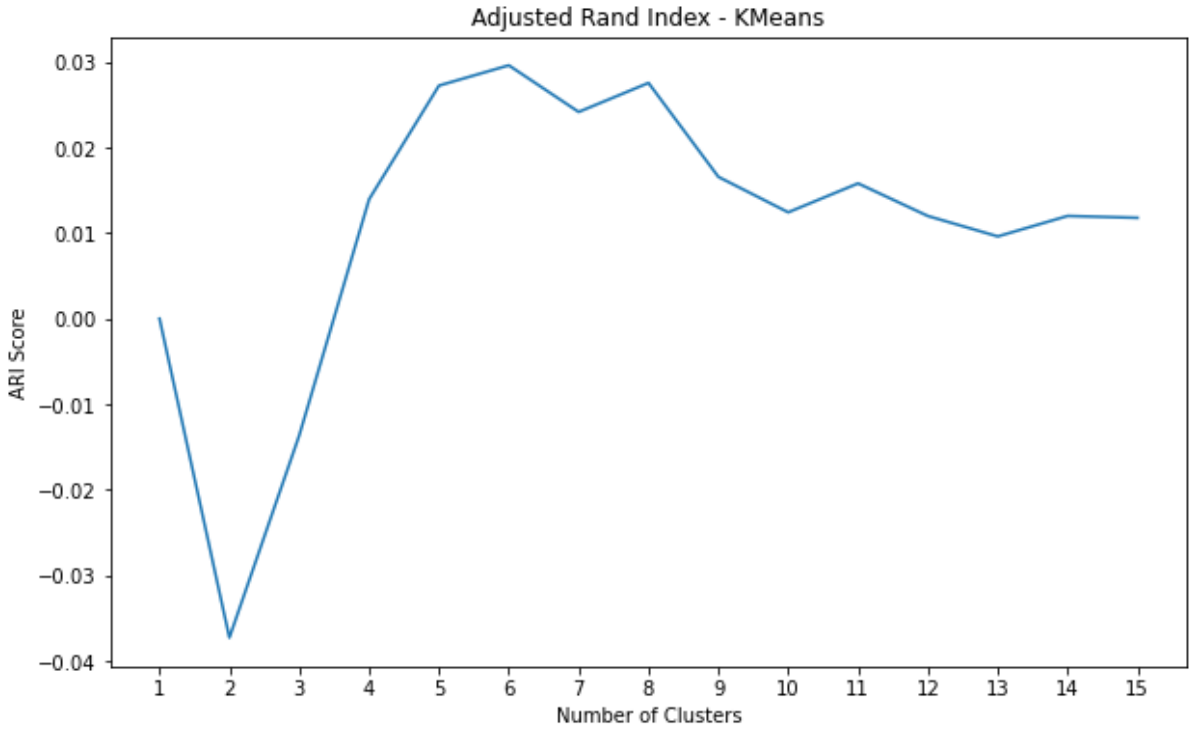
Selected value of k =2. As discussed above and as shown in the WCSS graph below, the most optimum result is to select k at the point where WCSS stops decreasing significantly. In our case this is where k = 2.



Using the Silhouette Coefficient returned a result of k =2 as being the most optimal



Using the Adjusted Rand Index provided a result suggesting that k = 6 was optimal for our data set. This would however yield 6 clusters and is illogical given that there can only be 2 possible choices or clusters. Resigned or Not Resigned .



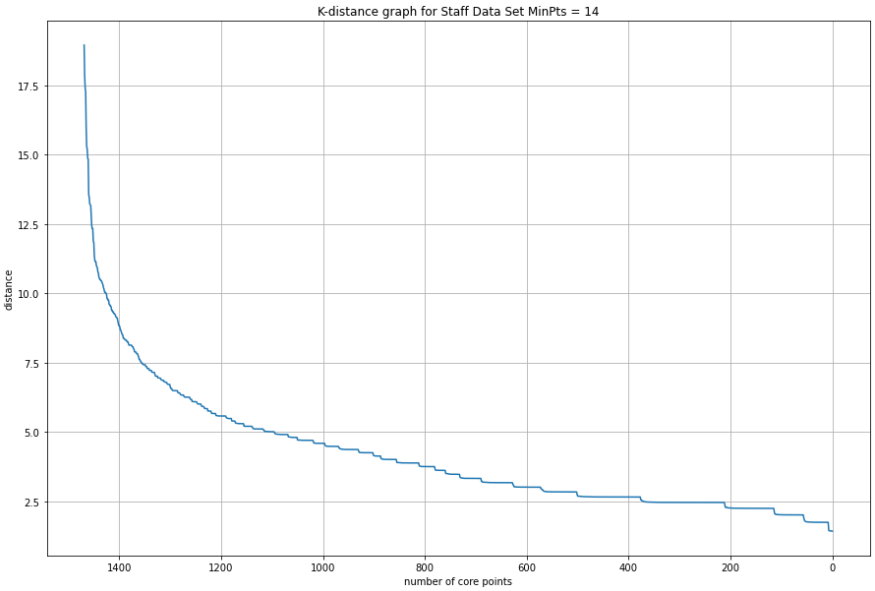
Results for KMeans where k = 2 are as follows:

KMeans Cluster (0 is Not Resigned, 1 is Resigned)	Resigned Data
0 = 1040 1 = 430	0 = 1233 1 = 237

DBSCAN:

For the DBSCAN model, MinPts was set with a suggested value of 14 (7 features part of data set, multiplied by 2 as this was suggested for data sets with many data features and a lot of data. Lower Minpts of 8 for example did not provide us with useful clustering.

K distance graph is then plotted in determining the optimum value for Eps. The graph below suggests the optimal values for Eps are 7.5, 8.25 and 9.



In evaluating the performance of our DBSCAN, we use the Silhouette Coefficient. The Silhouette Coefficients calculated using the values of Eps graphed shows suggests that these are solid scores. For the Silhouette coefficient, the values are bound between -1 and 1, with a value closer to 1 indicating that the model has better defined and condensed clusters.

Eps = 7.5	**Eps = 8.25**	**Eps = 9**
0.4566111184563737	0.43396087944129746	0.4345867733022825

The results of our DBSCAN clusters:

0 Not Resigned, 1 represents Resigned, and -1 represents noise/ non-cluster.

DBSCAN Clusters (MinPts = 14, Eps = 7.5)	DBSCAN Clusters (MinPts = 14, Eps = 8.25)	DBSCAN Clusters (MinPts = 14, Eps = 9)
0 = 1419 -1 = 51	0 = 1432 -1 = 26 1 = 12	0 = 1440 -1 = 20 1 = 10

Discussion

Using the KMeans model, with k= 2, the model identified a higher amount of resigned employees than recorded in the original dataset. 430 compared to the 237 in the original RESIGNED data.

In comparison, the DBSCAN did not provide clusters representing anything close to original RESIGNED Data.

Where Eps = 7.5, the model identified 1419 as Not Resigned, and 51 that could not be grouped into a cluster.

Where Eps = 8.25, the model identified 12 as the amount of designed employees, 26 could not be grouped into a cluster.

Where Eps = 9, 1440 employees were identified as Not Resigned, 10 people as RESIGNED, and 20 people as could not be classified into a group.

The DBSCAN results may be due to the high-dimensional nature of our data set. The high number of features and data, there would be little differences in the pairs of samples which would make it difficult in finding an appropriate value for Eps.

Though the KMeans model has identified a higher amount of resigned employees versus the original data results, it has stronger support from the evaluation methods used, and the results are more representative to the ground truth. Thus it seems more useful and suitable in answering our research goal.

Conclusion

The data exploration and modelling was conducted with the goal of explaining the underlying issues for employee retention affecting Revolution Consulting and informing management of appropriate and actionable remedies. We have identified data features within the staff data set which can be used to create data models that can be used to predict staff resignations.

Our data exploration has enabled us to identify a number of factors contributing to the company's retention woes including:

- *Remuneration*
- *Gender pay*
- *Lack of Career Progression*
- *Concerns with Work Life Balance among New Employees*

One of the several issues identified is one of progression, the average time for employees stay at the company is 5.13 years, but our data analysis shows that the monthly income for new employees takes 6 years for them to see any considerable increase, hi-lighting issues where employees are likely to resign due to a lack of career progression.

Among new employees, there exists a significant correlation where they work more over time compared with more tenured and senior employees. Senior employees look to be satisfied given they are less likely to work overtime and receive significantly more pay.

Data analysis conducted among genders reveal that there exists a gap in pay favouring women. Mean female incomes = 6686.566327, Male mean income = 6380.507937. This disparity in income is reflected by higher resignation rates among men.

Ratings for Job Satisfaction and Work Life Balance don't appear to be accurately capturing the reality of employees. Medians for Job Satisfaction and Work Life Balance tend to be the same, with only slight changes in means measured across those that have resigned and across business units. This Suggests there's may be a cultural issue within the company where employees, particularly newer employees, are may doing a lot of overtime in order to reach quality and, or performance metrics but may be afraid of speaking out about their dissatisfaction.

Actionable recommendations that will assist with Revolution Consulting's retention issues, would be to provide equal pay for men and women. Remuneration and incentives should also be reviewed for newer employees either through promotion or other incentives such as bonuses and work benefits.

Revolution Consulting should also take measures to change their culture so that employees will feel more comfortable in speaking out about their concernts.

References

stats.stackexchange

Nabriya, P 2020, A routine to choose eps and minPts for DBSCAN

<https://stats.stackexchange.com/questions/88872/a-routine-to-choose-eps-and-minpts-for-dbscan/>

Medium

Mullin, T 2020, DBSCAN — Overview, Example, & Evaluation

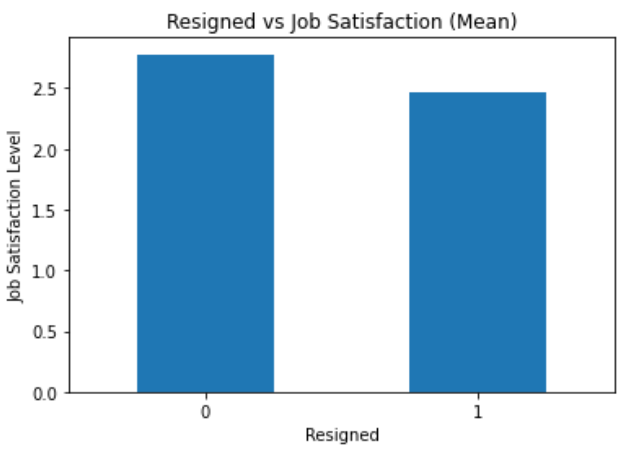
<https://medium.com/@tarammullin/dbscan-2788cfce9389/>

Appendix

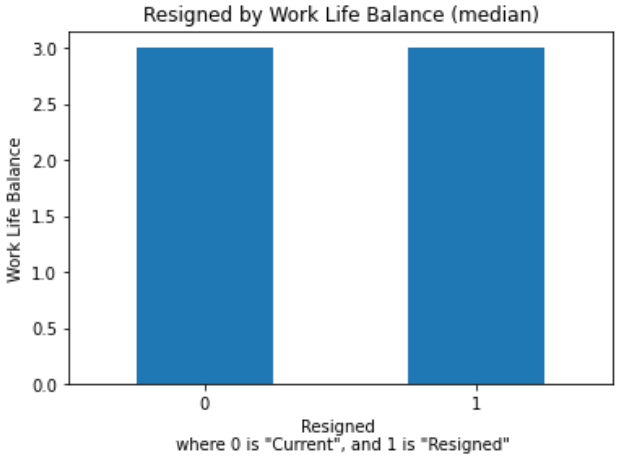
If required, include larger graphs in a numbered appendix.

Appendix 1

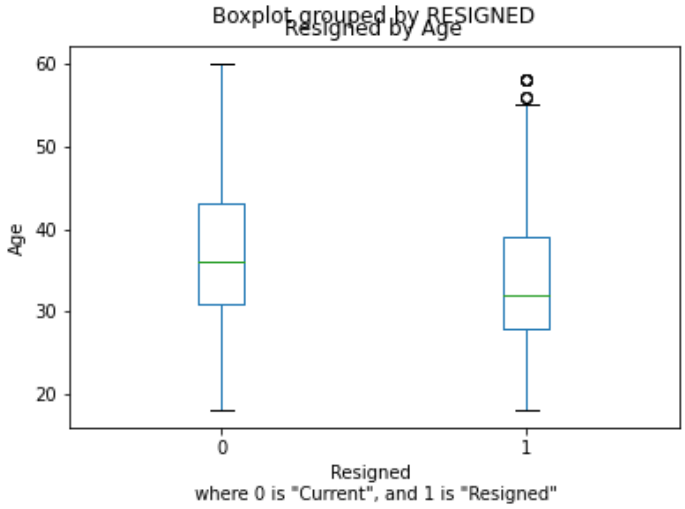
Appendix 2



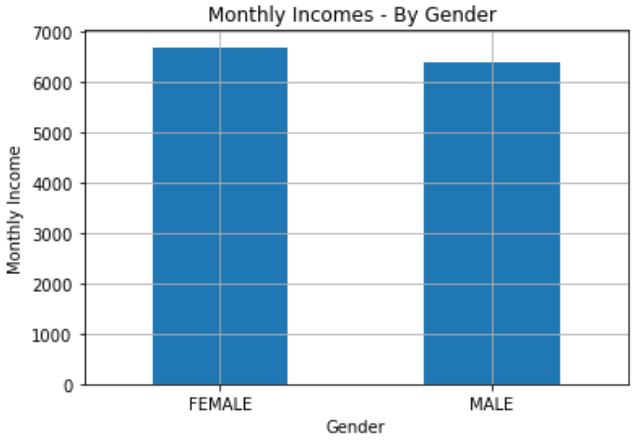
Appendix 3



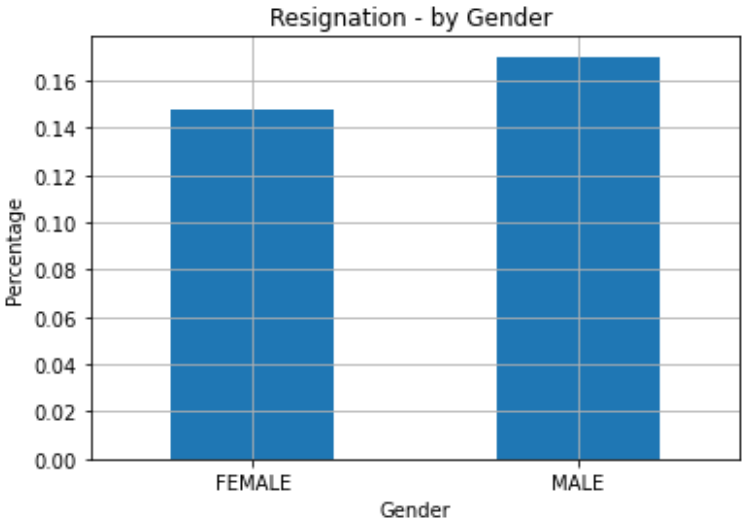
Appendix 4



Appendix 5



Appendix 6


Resignation - by Gender
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Percentage
FEMALE
MALE
Gender