EMPLOYEE SEGMENTATION

with Unsupervised Learning

Lim Jun Jie
DAAA/FT/2A/02

TABLE OF CONTENTS

PROBLEM STATEMENT

Tu understand the employees so that appropriate direction can be given to the management to satisfy and retain employees.

Objectives

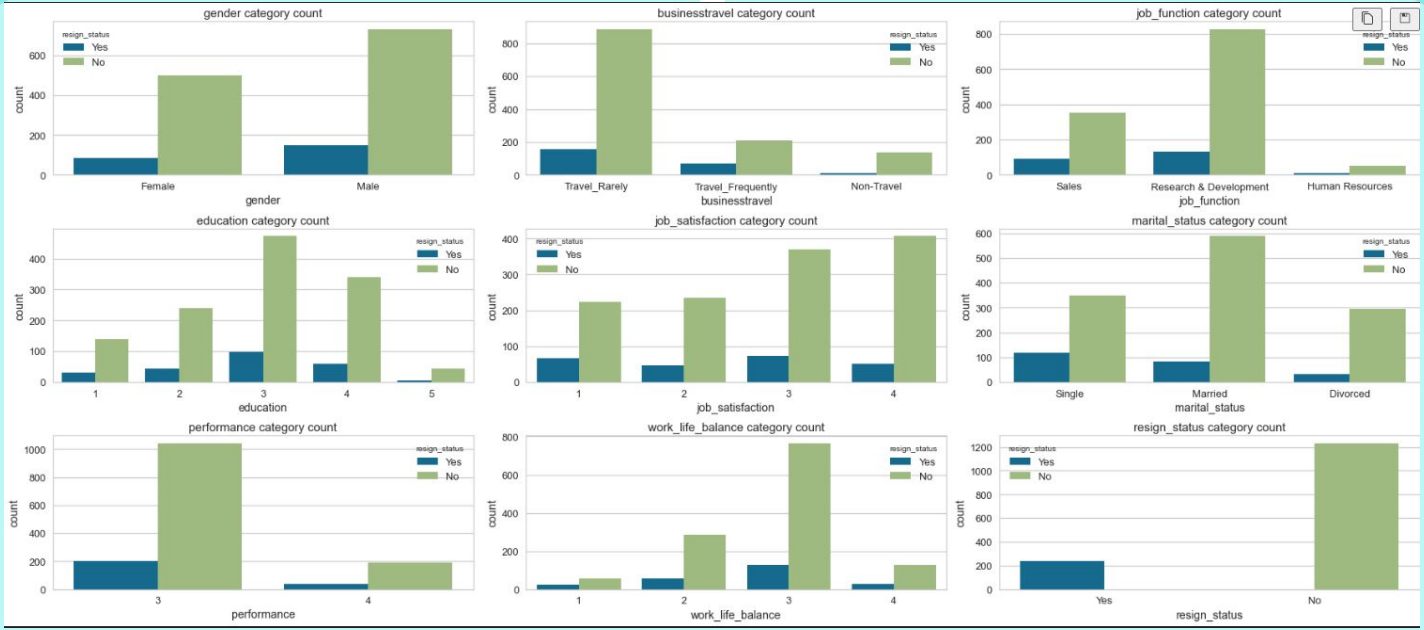
- Achieve employee segmentation using Unsupervised ML
- Describe characteristics of each employee cluster
- Find the most vulnerable group of employees and suggest strategies to retain them

EDA - Dataset Origin

- This dataset seems to be a edited version of the **IBM HR Analytics Attrition and Performance** Dataset
- Fictional dataset created by IBM Data Scientists

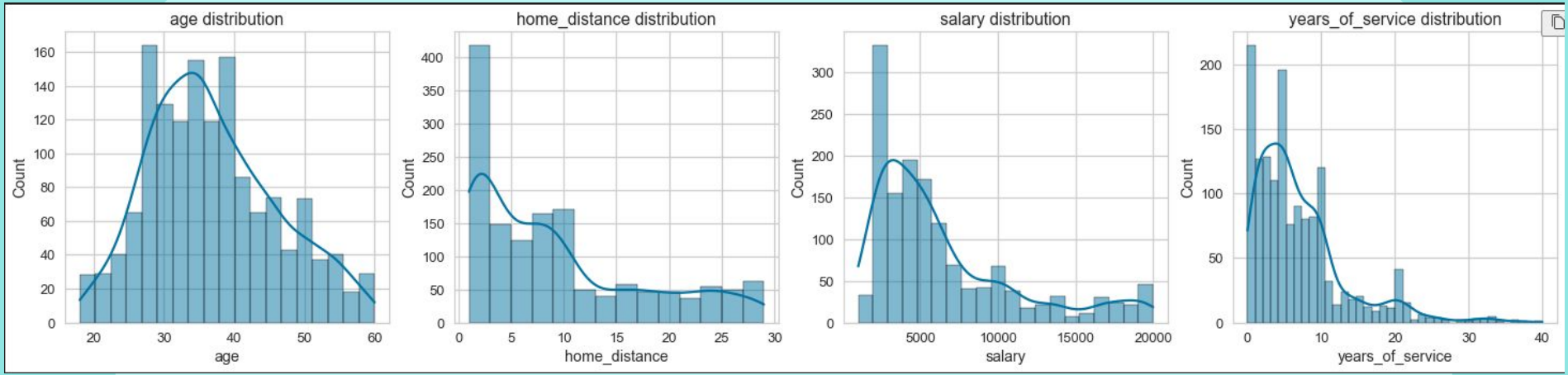
Kaggle Link:
https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

EDA - Bar Charts



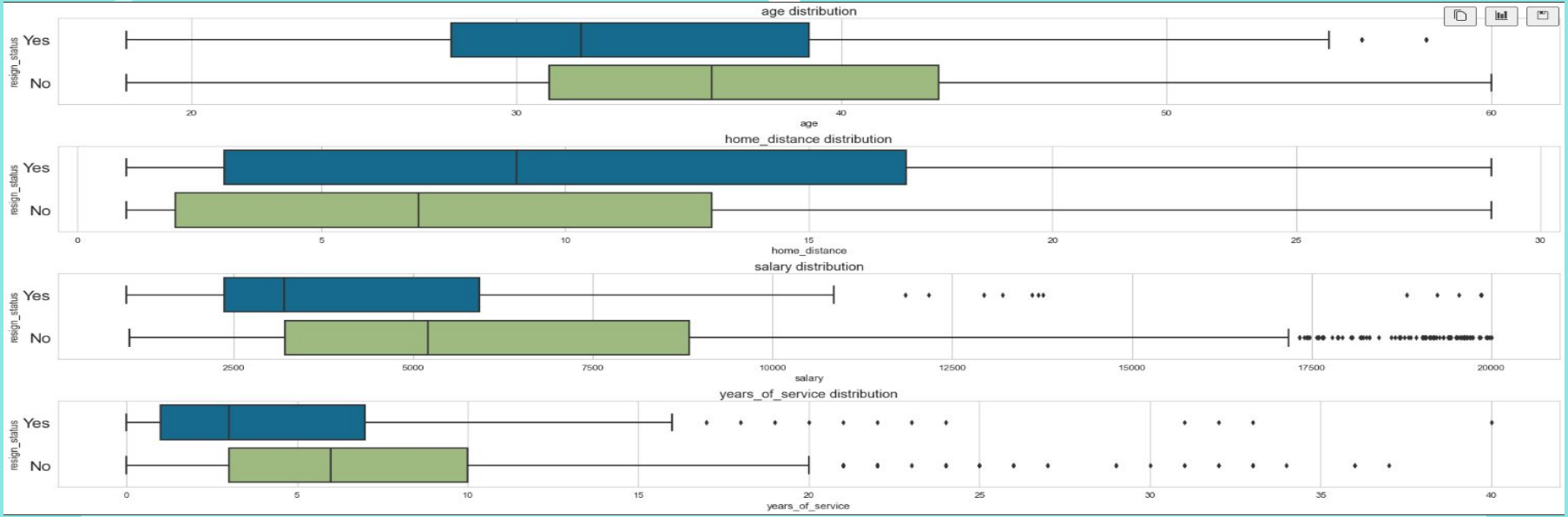
- No low performers in the company - The performance ratings are 3 or 4
- Employees are fairly satisfied with work-life balance
- Splitting by resignation status, no interesting or abnormal trends are spotted

EDA - Histograms



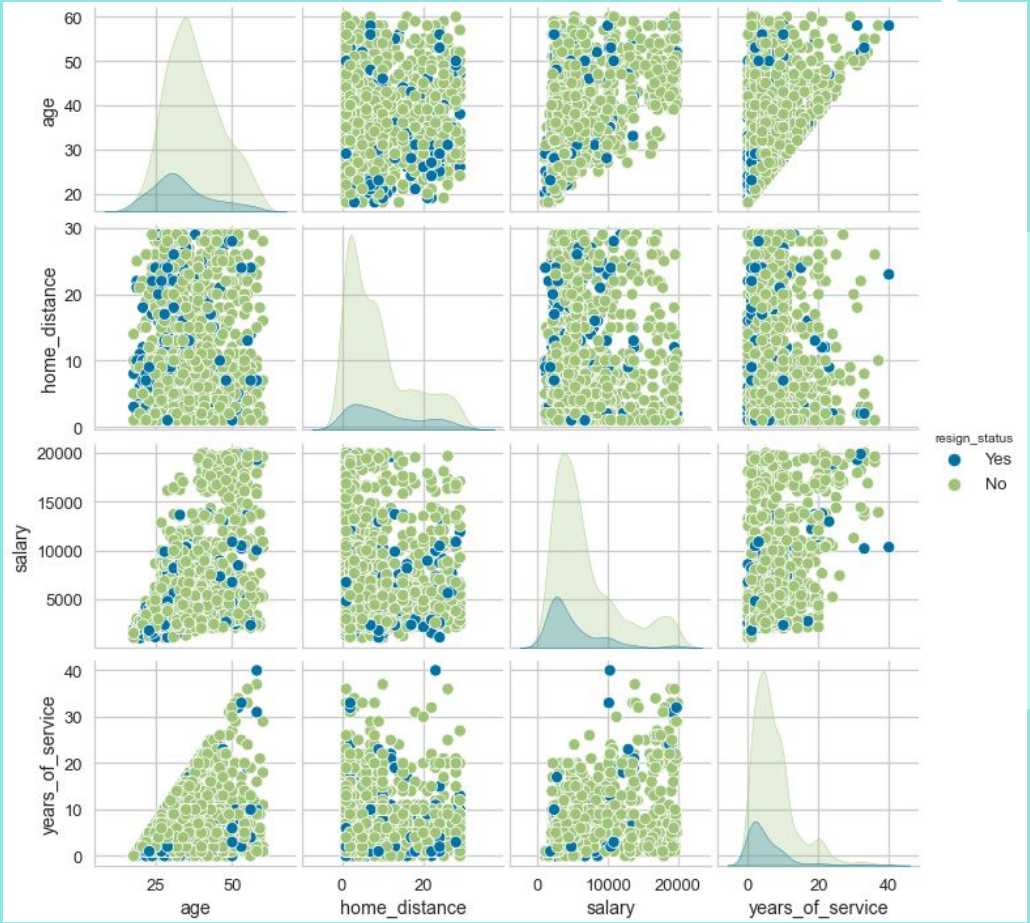
- Home distance, salary and years of service seems to be positively skewed, and might contain outliers
- Age distribution is near normal
- Box plots have to be plotted to check for outliers

EDA - Box Plots



- There are differing medians when each feature is plotted by resignation status
- Outliers in years of service are due to employees who stayed with the company for a long time, causing the distribution to be skewed
- Outliers in salary can be explained as in a company there will be employees who earn way above the median wage.

EDA - Pair Plot



- Pair plot is plotted with resignation status set as the color
- There seems to be no meaningful clusters formed for each continuous feature when using resign status as color on a 2D scatter plot

EDA - Heat Map

- Heatmap of continuous and ordinal features
- Used Spearman Correlation
- There is a weak relationship between salary and age
- There are no strong relationship between features

DATA PREPROCESSING

Nominal Encoding

OneHotEncoder was used to encode nominal features:

gender, resign_status, job_function, marital_status

Ordinal Encoding

OrdinalEncoder was used to encode ordinal features:

business_travel

Feature Selection

As there are 15 features after encoding, feature selection is needed to prevent the 'Curse of Dimensionality' . Feature is selection is done only on categorical features.

Chi Square test is used to select features, setting the target feature as resign_status.

Features that are independent of the resign_status feature are removed.

DATA PREPROCESSING

Feature Scaling

As ordinal and continuous features have different magnitudes, feature scaling has to be done.

Feature scaling is done using StandardScaler on continuous and ordinal features

Hopkins Test

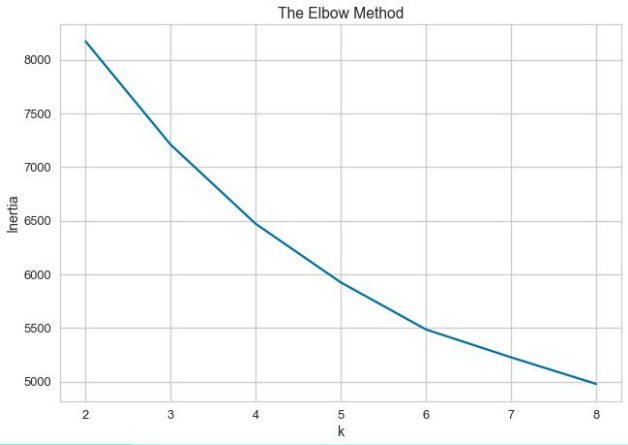
The Hopkins Statistic is used to assess the clustering tendency of a dataset.

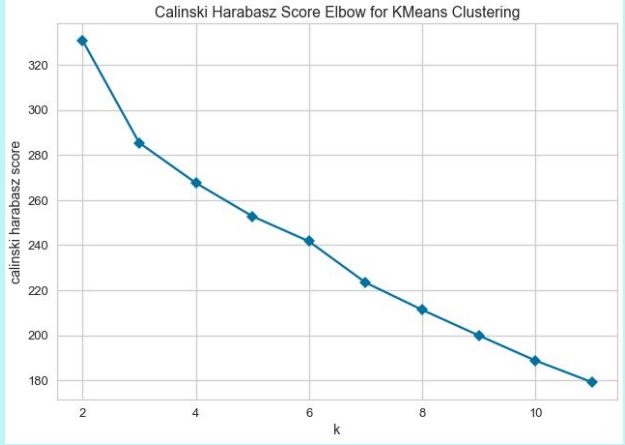
If Hopkins Statistic < 0, then it is unlikely that dataset has statistically significant clusters.

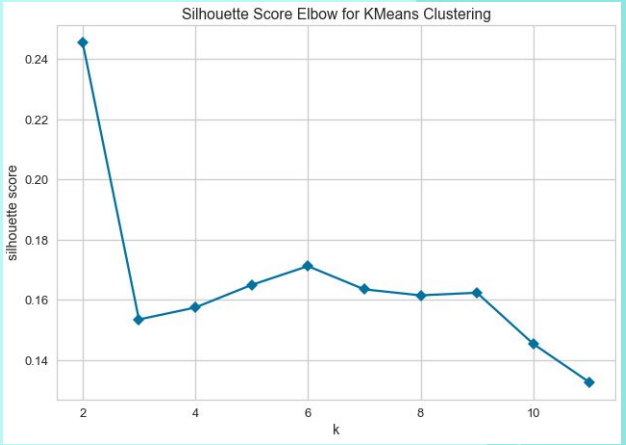
If it is close to 1, null hypothesis can be rejected and dataset has a significantly clusterable data

Modeling: Clustering Algorithms

K-Means Clustering







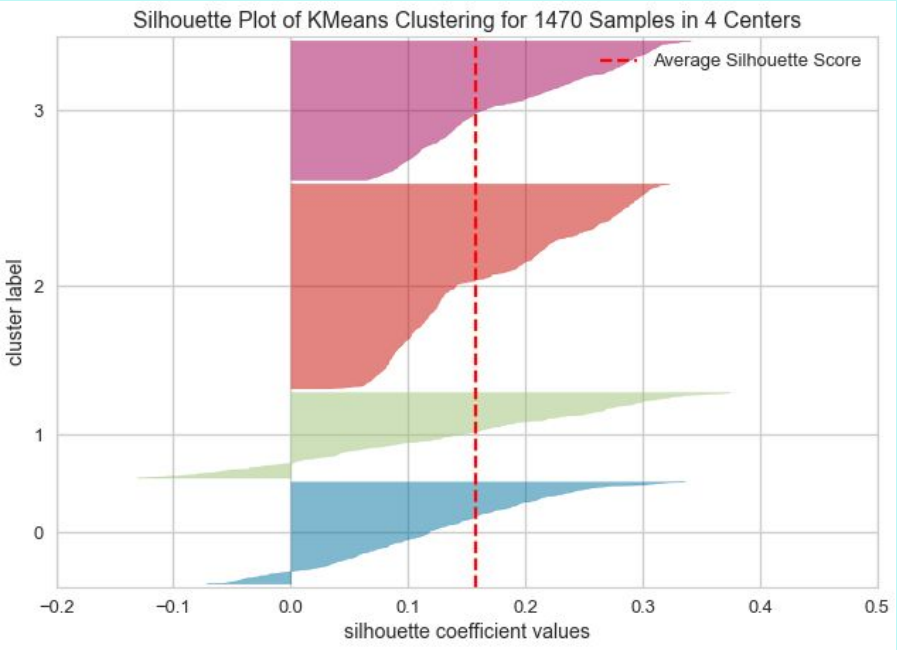
Number of k

- Elbow Method using Inertia does not have a obvious elbow, inertia decreases uniformly
- Silhouette Score and Calinski Harabasz Score are highest at k=2
- Silhouette score also spikes up at k=4 onwards, while Calinski Harabasz score decreases uniformly

K-Means Clustering

Silhouette Analysis

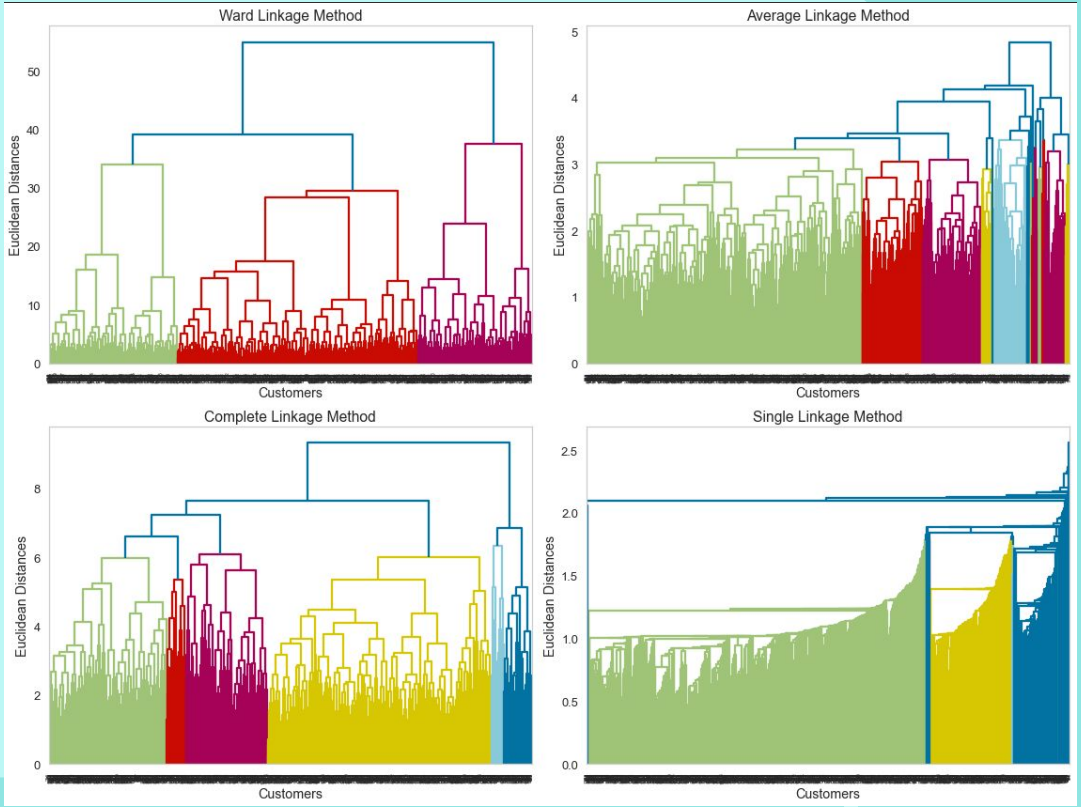
- k=4 is used, with 0.159 silhouette score and Calinski-Harabasz Index of 265
- Negative values indicate that that some data points may be clustered wrongly



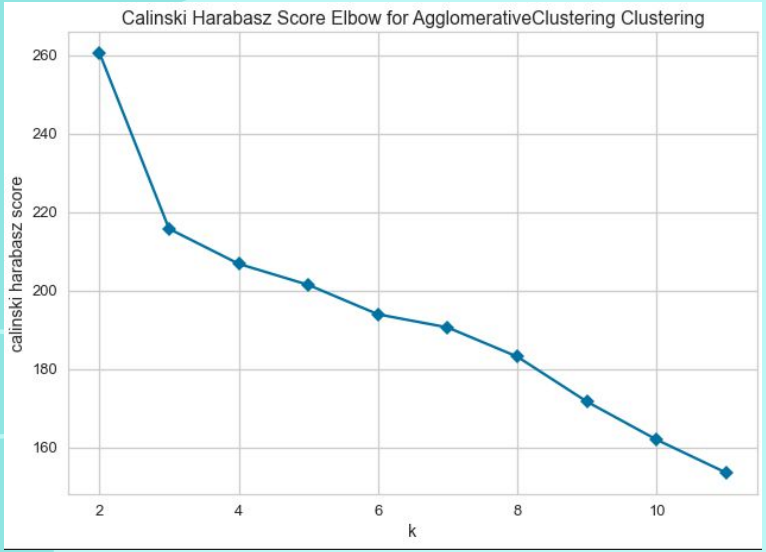
Agglomerative Clustering

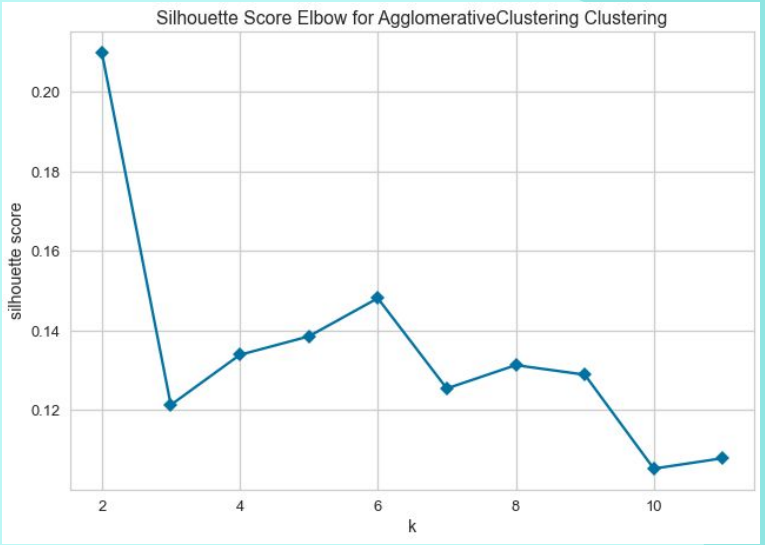
Linkage Parameter

- Ward Linkage gives the most even cluster sizes.
- Thus Ward Linkage will be used to test out Hierarchical Clustering



Agglomerative Clustering





Elbow Method

- Calinski-Harabasz Score decreases as number of clusters increases
- Silhouette Score is decreases at k=3, then spikes up at k=4 onwards.
- The Silhouette Score and Calinski-Harabasz Scores are generally lower compared to K-Means Clustering

Mean Shift Clustering

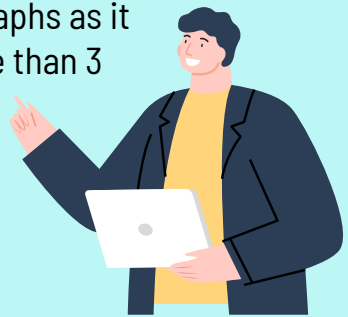
```
bandwidth: 2.0, clusters: 18, silhouette_score: 0.08783377847630124, calinski_harabasz: 34.12223226356367
bandwidth: 2.1, clusters: 8, silhouette_score: 0.16091627393696473, calinski_harabasz: 52.70976985360581
bandwidth: 2.2, clusters: 5, silhouette_score: 0.1958010925094073, calinski_harabasz: 80.18435768693384
bandwidth: 2.3, clusters: 4, silhouette_score: 0.20941615211744227, calinski_harabasz: 62.71134071932303
bandwidth: 2.4, clusters: 2, silhouette_score: 0.35523820320078364, calinski_harabasz: 92.31803957384875
```

Testing with different bandwidth numbers

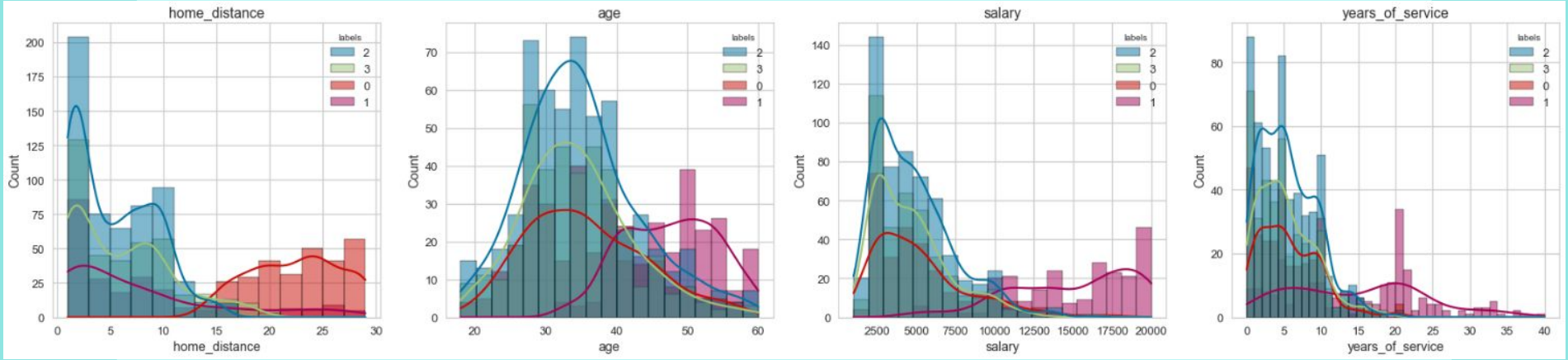
- After iteratively testing with different bandwidths, bandwidth 2-2.4 produced more than 1 cluster as a result
- When testing for bandwidth 2 to 2.4, both silhouette score and Calinski-Harabasz score increases and bandwidth increases
- Both scores are generally lower compared to scores from K-Means and Agglomerative Clustering

Final Model: K-Means

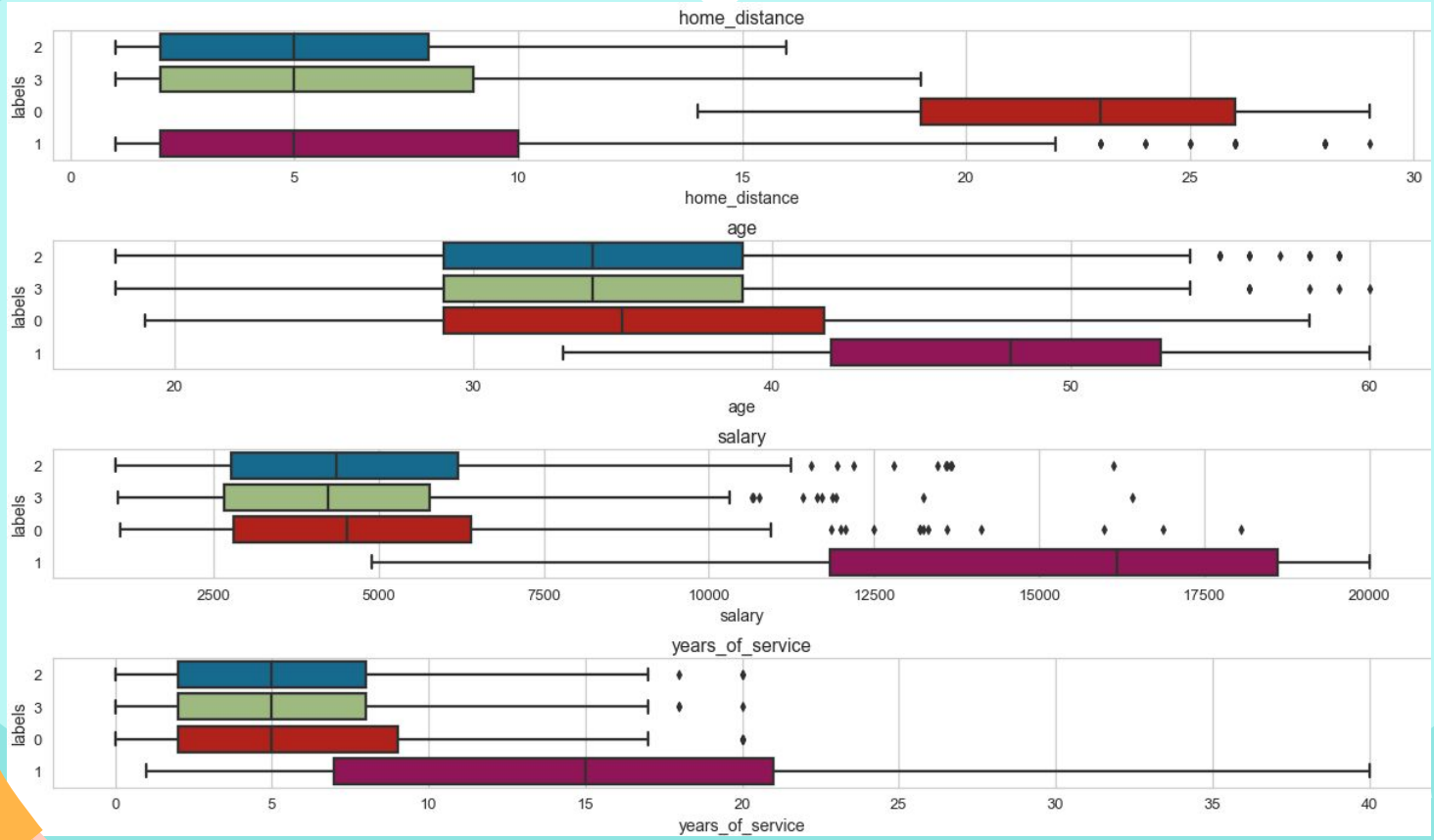
- K-means with 4 clusters is chosen as the final model as both Silhouette and Calinski-Harabasz scores are generally higher than 2 other models
- Choosing 2 and 3 clusters would make no sense as Inertia is at the highest
- Thus, 4 clusters are chosen as they would yield the best results
- Interpretations of the clusters will be made using various graphs as it is not possible to visualize clusters on scatter plot with more than 3 features

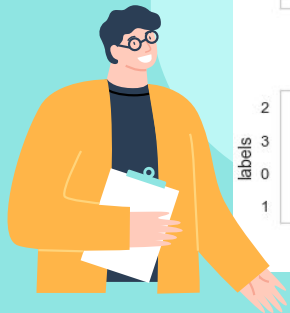


Cluster Interpretations: Graphs (1)

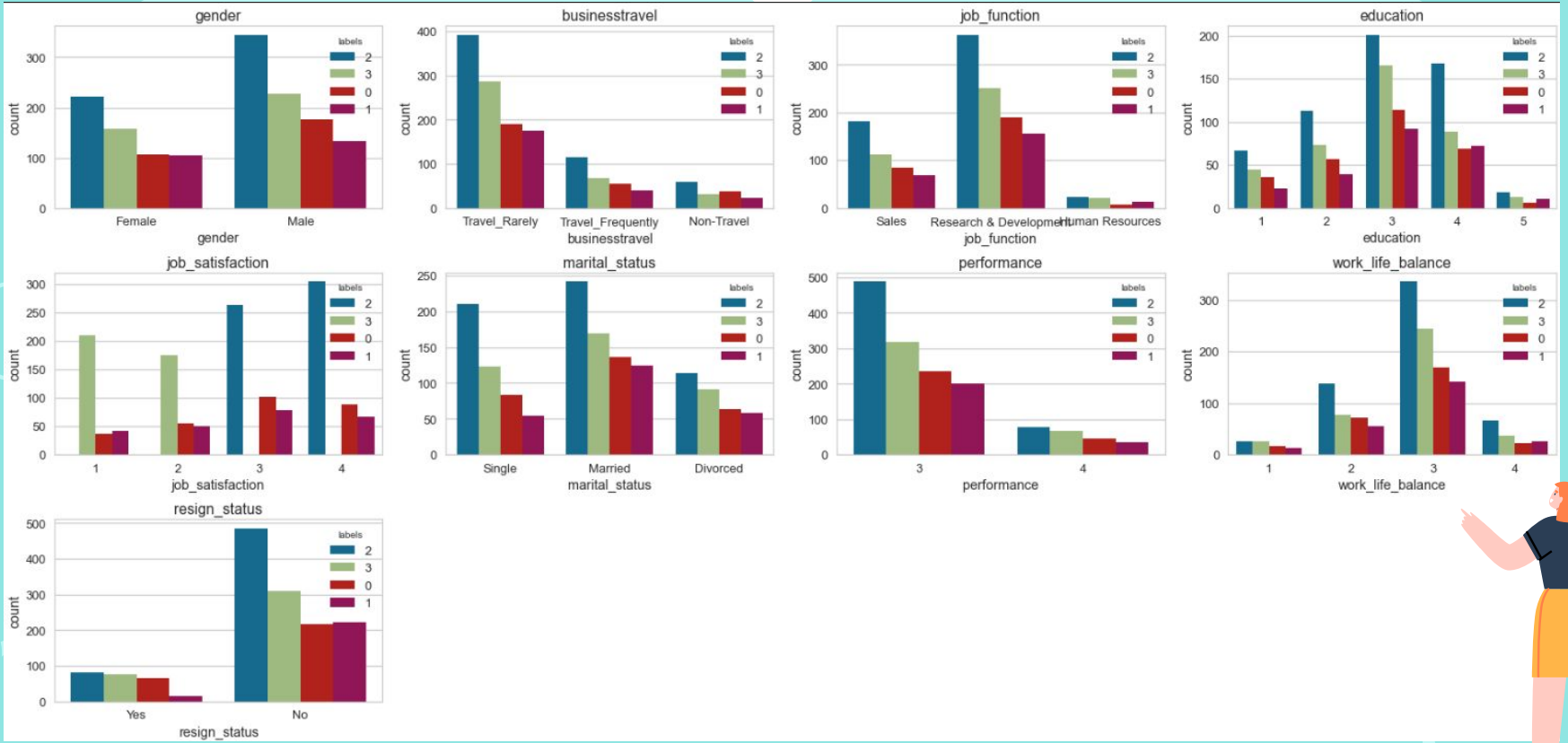


Cluster Interpretations: Graphs (2)





Cluster Interpretations: Graphs (3)



Cluster Interpretations

Cluster 0: **The Travellers**

- Living further from the office than everyone else, this cluster has the highest resignation rate at 23.4%

Cluster 1: **The Seniors**

- They are older and generally have higher salaries than everyone else, the population is the smallest in the company. They also have the lowest resignation rate at 6.75%

Cluster 2: **The Employees who love their Job**

- This cluster is 38.5% of the total population of employees in this dataset. Most, if not all of them voted 3 or 4 for the job satisfaction rating

Cluster 3: **The Unsatisfied Employees**

- Opposite of Cluster 2, most, if not all of them voted 1 or 2 for the job satisfaction rating. This cluster has the lowest median salary at $4000 and also has the 2nd highest resignation rate at 19.48%

Conclusion: Back to Problem Statement

Answering the Question: **Find the most vulnerable group of employees and suggest strategies to retain them**

The Travellers are the most vulnerable group, having a resignation rate of 23.4%. This can be due to them wanting to find a workplace closer to their home

Why:

- Save money on petrol/public transport/private hire fees
- Save time commuting to the workplace

Possible Strategies

- Public Transport/Petrol/Private Hire Fees Subsidy Program
- A company car program where employees can loan cars from the company
- Contract a bus operator to transport employees from home to the office

THANK YOU