Problem Statement

The "Clustering, EDA - Data Science and STEM Salaries" report provides a clean approach to the classification of data in relation to Data Science and Stem salaries, but its methods for the system of classification could use some improvement. The main objective of this project will be to replicate the results obtained by the original creator, and to test the hypothesis that changing the percentage of explained variance in our model, as well changing the model's algorithm will lead to more clear, and consistent results.

Related Work

By analyzing the project code, I was able to learn the method of classification, and replicate the results. After some data preprocessing, the model accounts for twelve features which are Company, Level, Title, Location,Years of experience, Years at company, Base salary, Stock grant value, Gender, Race, Education, and Total yearly compensation. The creator collects important feature information by conducting Principle Component Analysis with the number of components equal to three. In order to choose the appropriate amount of clusters, the creator uses the Elbow method. The Elbow method plots the explained variance against the number of clusters to provide an optimal solution. Finally the model clusters by Agglomerative Clustering, which is a form of hierarchical clustering. From there, the Agglomerative Clustering model is used to classify the data. The model allows us to then plot any of the features against each other to visualize the clusters.

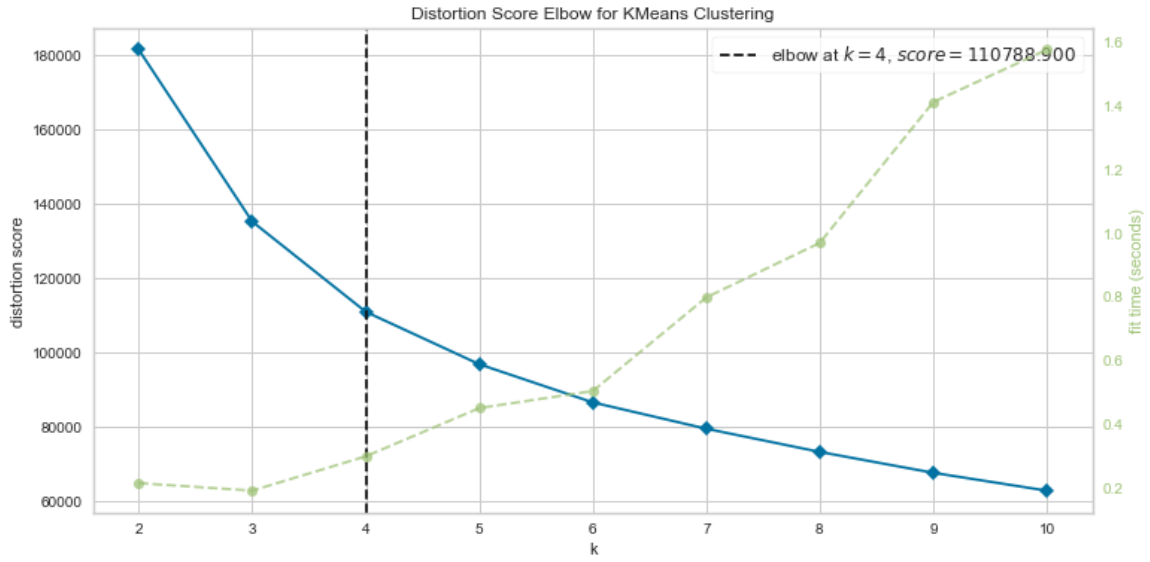

Figure 1: Elbow method for choosing cluster size

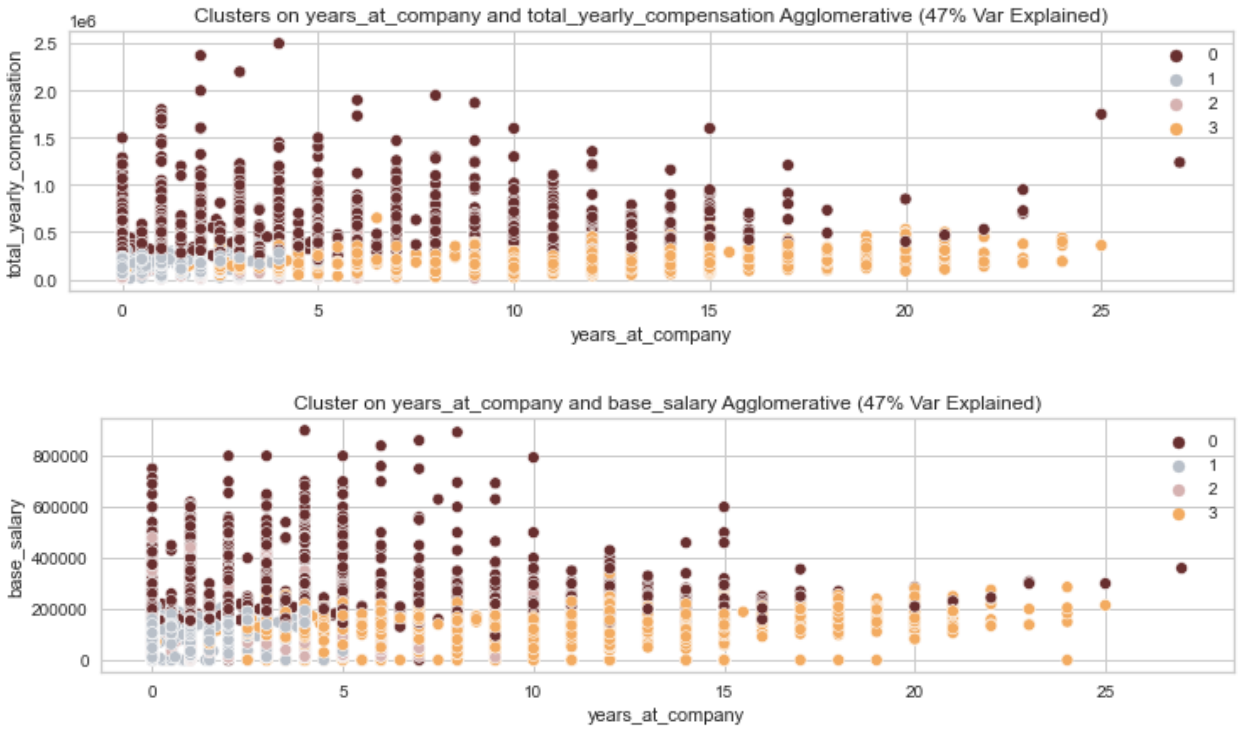

Figure 2 & 3: Clustering coupled with features plotted against each other

Methods

While the model above follows a clean approach to clustering. An improvement I made was in the number of principal components chosen. As shown above, the explained variance ratio was forty-seven percent. Because we are clustering for descriptive purposes. Based on external research, from sources such as UCLA department. The variance accounted for in factor analysis should be at least seventy percent. I believe that in the model's case, it could be losing some important information. I do recognize that the project creator could've had the motive to limit the number of principal components for visualization purposes. But for the purpose of maintaining statistical standards, it is better to choose a model that explains a good portion of the variance. Another improvement I made was by comparing Agglomerative to Kmeans ++, to assess the fit, as well as variance across unseen data.

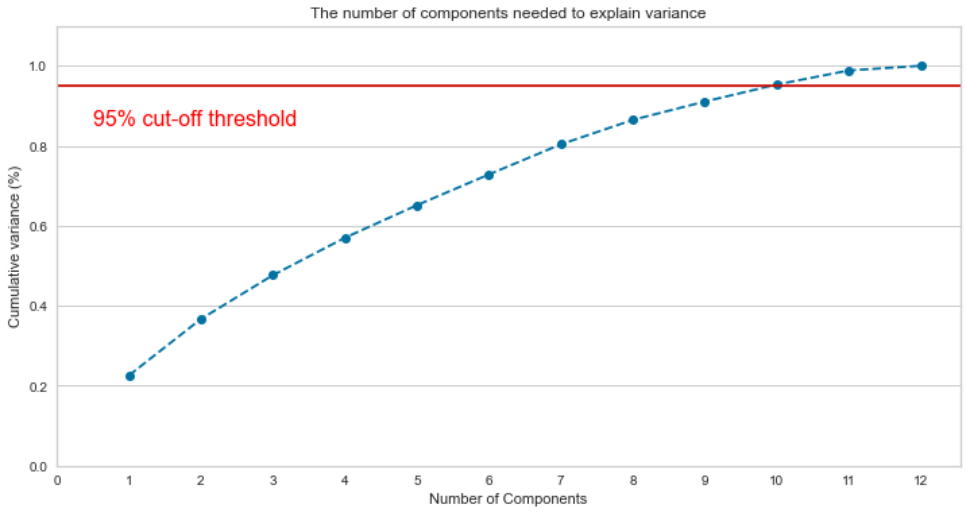

Figure 4: Plot of Number of Components against Cumulative Variance to choose optimal number of components

The hypothesized improvement made was choosing the number of components in order to achieve the explained variance ratio threshold above, which in this case was nine components. As well as clustering using different algorithms. To provide evidence, I ran both models against test data in order to test the consistency of the models under new data.

Results

The hypothesized methods of improvement did not work as expected. My test results are inconclusive as neither model improved or remained consistent under changes to the explained variance ratio,or the algorithm itself. Based on the results of the model with a ninety-five percent explained variance, the model actually seems to be overfitting the data as the clusters begin to overlap more.

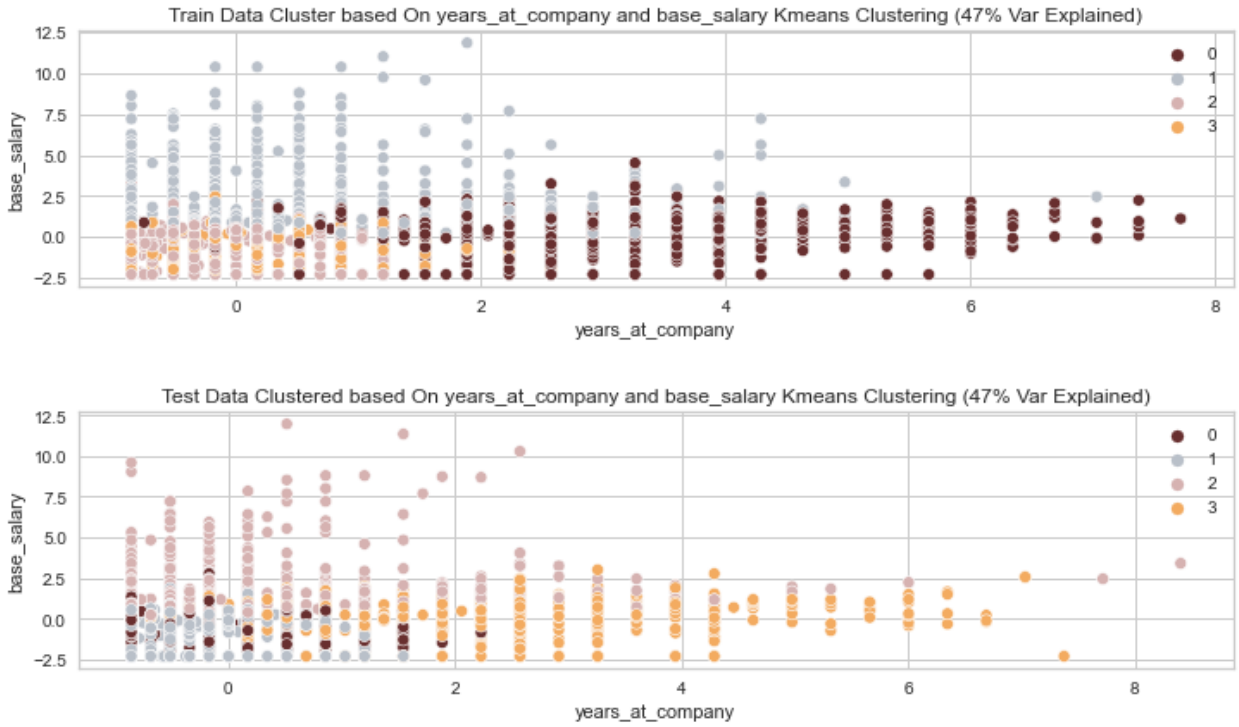

Figure 5 & 6: Train, and test data Kmeans clustering

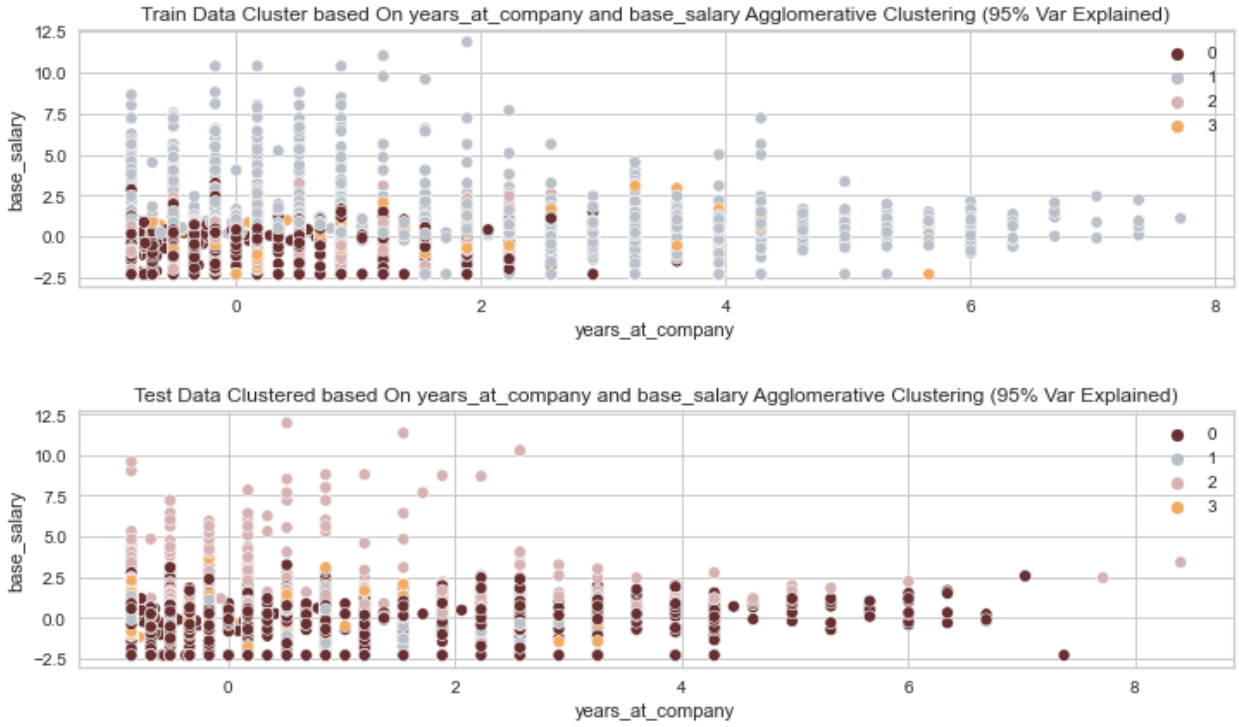

Figure 7 & 8: Train, and test data Agglomerative clustering

Conclusion

This project replicates the "Clustering, EDA - Data Science and STEM Salaries" report, as well as make some improvements in order to control the variability of the model under unseen data. Techniques including PCA, Kmeans ++, and Agglomerative clustering were used to explore and gain insight on the data. Comparing the models, the original project's method limiting the number of principal components in addition to Agglomerative clustering proved to produce the most consistent results. Using a higher number of principal components, in addition with Agglomerative, and Kmeans ++ clustering seemed to over complicate the model, leading to some form of overfitting.

Although these models are hard to gauge from an accuracy standpoint because it is unsupervised. Further testing across different numbers of components, as well as more clustering algorithms could provide a baseline for better understanding, and comparison purposes.

References

UCLA. “A PRACTICAL INTRODUCTION TO FACTOR ANALYSIS: EXPLORATORY FACTOR ANALYSIS.” *IDRE Stats*, https://stats.idre.ucla.edu/spss/seminars/introduction-to-factor-analysis/a-practical-introduction-to-factor-analysis/.

mario931. “Clustering, EDA - Data Science and STEM Salaries.” *Kaggle*, Kaggle, 10 Nov. 2021, https://www.kaggle.com/mario931/clustering-eda-data-science-and-stem-salaries/notebook.

scikit-learn: Machine Learning in Python https://sklearn.org