CS105 Final Report - Salary Prediction

Group 1: Nathanael Shin, Rahul Katwala, Neha Mathews, Michael Bettencourt, Anjali Daryani

Goal

The overall goal of our project is to use the data from salaries.csv to predict the salaries of someone in the AI/Machine field of work. We will use a combination of data cleaning, exploratory data analysis, and machine learning algorithms to explore the data and create a model that can accurately predict salaries based on the given data. To do so, we are taking in the dataset that contains Work Year, Experience Level, Employment Type, Job Title, Salary, Salary Currency, Salary in $USD, Employee Residence, Remote Ratio, Company Location, and Company Size, but focusing on Employment Type, Job Title, Company Size, and Experience Level for this project. We did not use employee residence, remote ratio, and company location, as there were too many categories, with too little data to fill those categories.
Then, by using those categories, we are going to try to predict the salary of a person in the AI/Machine Learning field.

Data Processing/EDA

Data Cleaning

In order to first clean the data, we replaced all of the categorical variables with numerical values. This included the experience_level, employment_type, company_size, and job_title variables, which were replaced with numerical values. The dataset also contained inconsistent data like uppercase and lowercase values for employee residence, so the code converts all the values to uppercase to have a standardized format. We then removed any rows with a company_location that was not "US", as that was outside the scope of our project. Finally, we dropped the company_location column entirely, leaving us with the cleaned data ready for EDA. This cleaning ensures that the data is consistent and accurate for analysis. Then, to create our clean dataset, we used:

```
df_cleaned.to_csv('cleaned_salaries.csv', index=False)
```

Data Visualization

Our code creates visualizations to help identify patterns and trends in the data.

1. We created a **scatter plot** using px.scatter() to show the relationship between different variables such as experience level, employment type, and salary. This was our first graph

made to have a visualization of how the salaries are distributed within the different job titles.

2. We created a **histogram** and **violin plot** using plt.hist() and go.Violin(), respectively, to show the distribution of salaries in the dataset. This helped us identify any outliers within our dataset, and gave us a visual representation to see the range in which most salaries fell in between. It was one part of figuring out why our MSE was so high.

3. We created a **pie chart** to visualize the count of data samples for each experience level. This is useful to clearly see the experience level of individual people in the dataset, which can help us understand more about our KNN prediction model.

4. We created a **bar plot** using sns.barplot() to show the relationship between experience level and salary. This visualization was useful for seeing how the average salary of the different experience levels compared with one another. It was also useful for figuring out why our prediction was not accurate.

5. We created a **heatmap** using sns.heatmap() to show the correlation between different variables in the dataset. This visualization helped us identify if there was any correlation between the salary, and the experience level and company size.

6. We created a **residual plot** created using the scatter() function from the matplotlib library, and displayed using the show() function, identifying any patterns in the residuals and to visualize how well our model predicted the salary.

7. We created an **elbow graph** to determine the optimal value of k for the KNN model. We use the optimal value of k to try and prevent overfitting or underfitting.

8. We created a **3D scatterplot** to visualize the relationship between salary, experience level, and company size for different job titles in the AI/ML field. We can then use this plot to see potential clusters in the data.

9. We created a **dendrogram** to perform hierarchical clustering and was used to visualize the distance between data points and identify potential clusters.

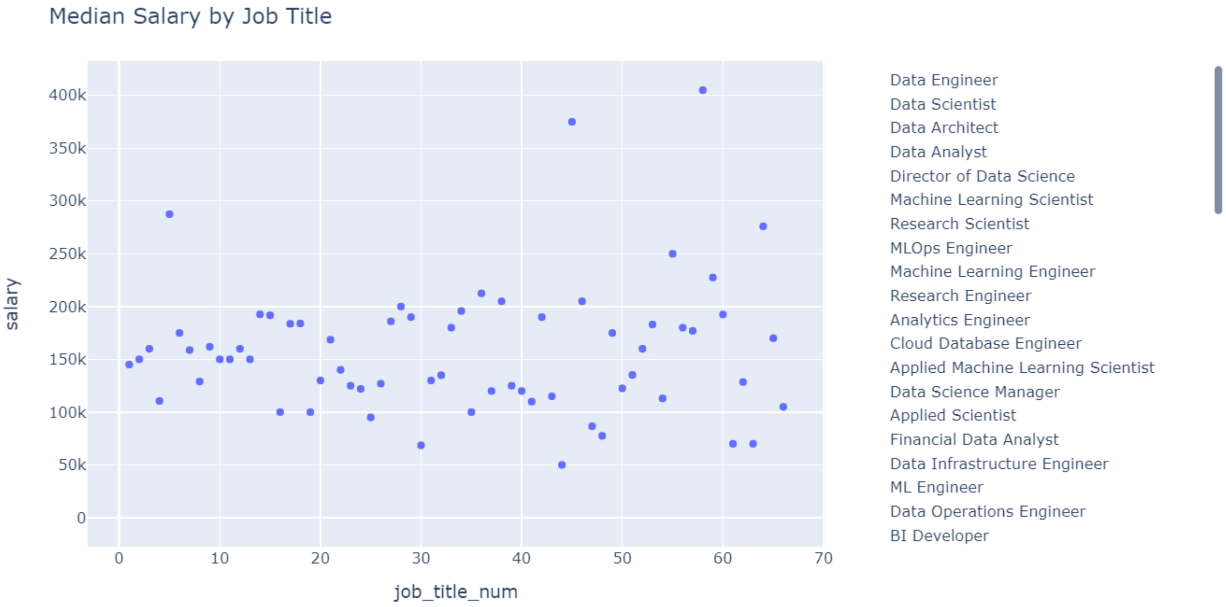
Machine Learning

1. The code uses machine learning techniques such as **K-Nearest Neighbors classifier** to predict salaries based on different parameters. This is helpful in building a predictive model that can accurately predict salaries based on different factors.
2. The **mean_squared_error and the mean_absolute_error** function from scikit-learn library was used to calculate the mean squared error (MSE) and mean absolute error (MAE) between the actual salary values and the predicted salary values. In the context of predicting salaries, a lower MSE indicates that the model's predictions are closer to the actual salaries, and MAE measures the average absolute difference between the actual and predicted values, with higher values again indicating poorer performance.

Analysis

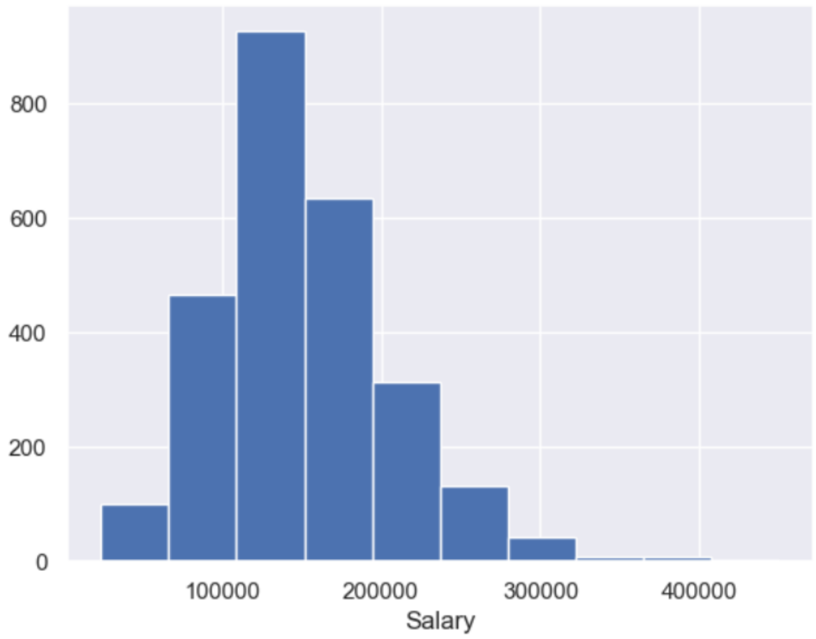
In order to properly analyze our data so we can make predictions, we can inspect our data visualizations to help us identify trends, communicate insights, and confirm our predictions. After explaining which visualizations we chose for our project, how we did it, and why we used it, we will be discussing what we have learned from each visualization in this section.

1. Scatter Plot

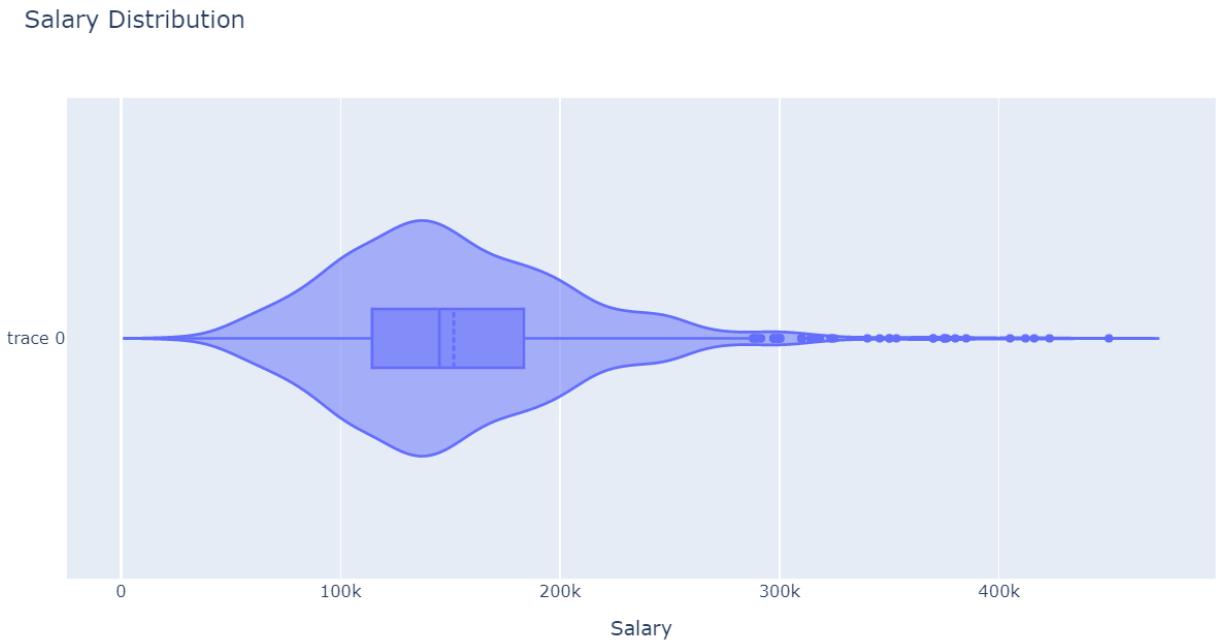


This scatter plot shows the distribution of salaries based on the job title. From this graph, we can see that there are only 2 job titles that make above 300k. This indicates that there may be outliers and under or over representation of certain salaries in the dataset. At this point, we are not sure if this will cause issues with predicting later, so we will continue exploring.

2. Histogram and Violin Plot

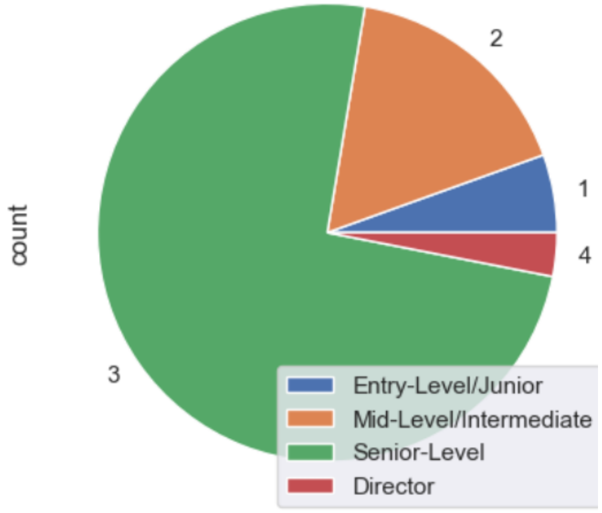


```
Median Salary: 145000.0
```



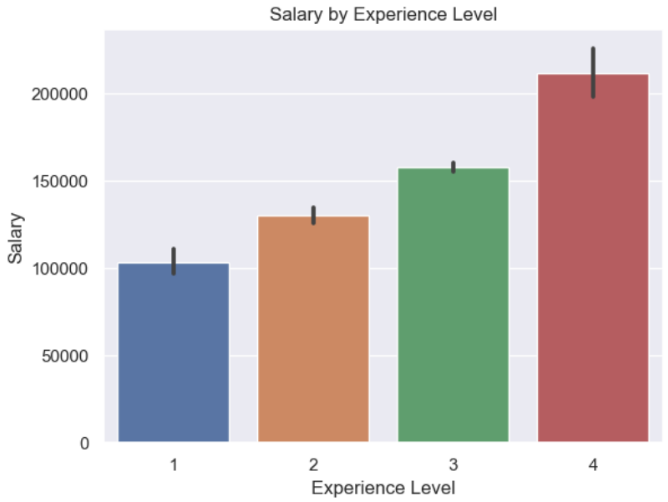
The violin and histogram show the frequency of salaries between individual people. From the visuals, we can clearly see that there are outliers in this dataset. We will explore more about this dataset when we talk about the MSE.

3. Pie Chart



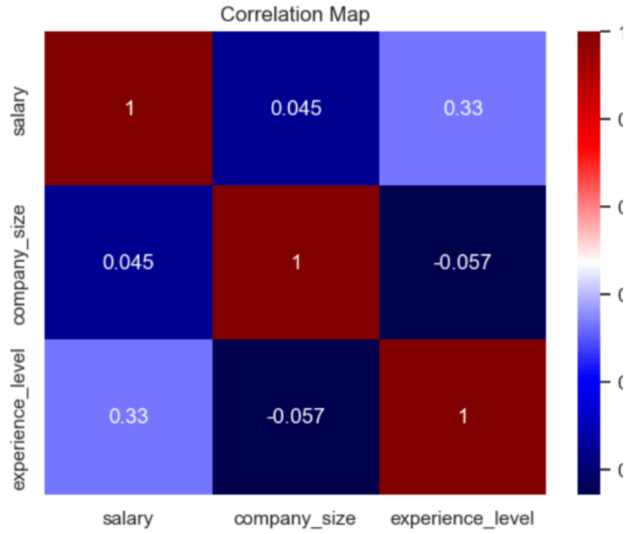
This pie chart shows the experience level of individual people in the dataset. From this pie chart, we can clearly see that the majority of people in the dataset are senior-level engineers. This indicates that there may be problems with our KNN prediction model, since there is an underrepresentation of the other 3 experience levels, which may not give us enough data to make accurate predictions. Just from the pie chart, we see that senior-level makes up 3/4 of the dataset, and the other 3 categories combined only make up 1/3.

4. Bar Plot



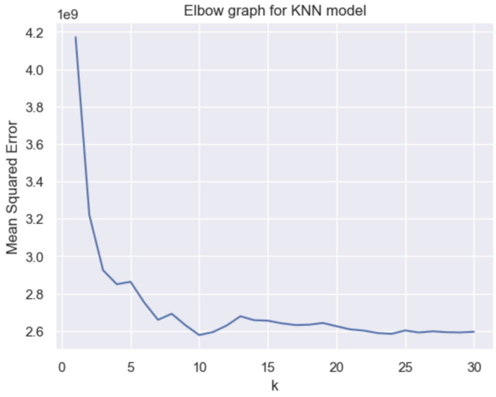
This chart shows that senior-level engineers make around 150k. Looking at the two previous visuals, the violin/histogram and the pie chart, we know that there are outliers, and that there is an overrepresentation of senior-level engineers. As such, this can cause predicted salary values for people in lower experience levels to be over the actual, and can cause predicted salary values for people with higher experience level to be lower than the actual. When grouping with KNN, since there are so many points within the senior-level category, it can cause data to be miscategorized, and therefore, wrongly predict the value.

5. Heatmap



This heatmap shows the correlation between experience level, salary, and the company size. We expected there to be correlations between these three variables, as we thought company size and experience level would be the biggest predictor of salary amount. However, from the heatmap, we see that this is not the case, which can give us a sense of how our model is going to work. The only notable correlation is between salary and experience level, but even then, the correlation is only .33, which indicates only a small positive correlation, but the only correlation between our three categories.

6. Elbow Graph



```
Best k value: 10
```

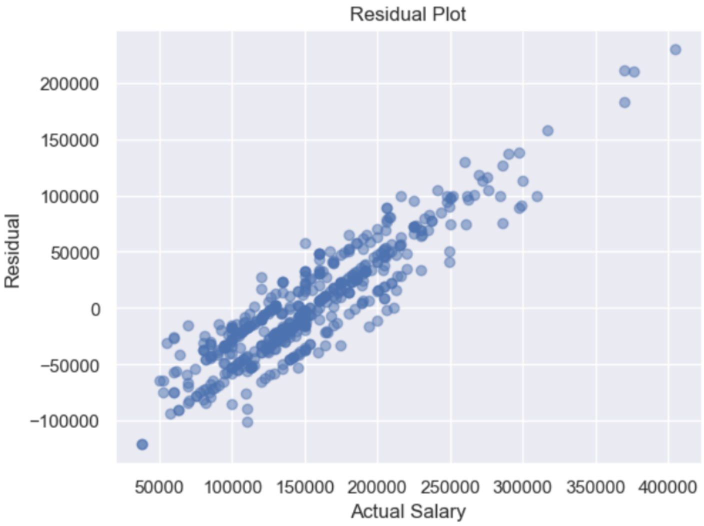
```
This shows why we used k=10 in our KNN model.
```

The elbow graph gives an output that specifies what value of k is the best. In this case, the best k value is 10, which is what we will use in our KNN Prediction model.

7. **MSE and MAE**
    - MSE: 2578060865.731207
    - MAE: 40276.842720306515

The MAE and MSE values show that our model is not accurate at predicting the salaries of people in the AI/Machine Learning field.

8. **KNN and Residual Plot**

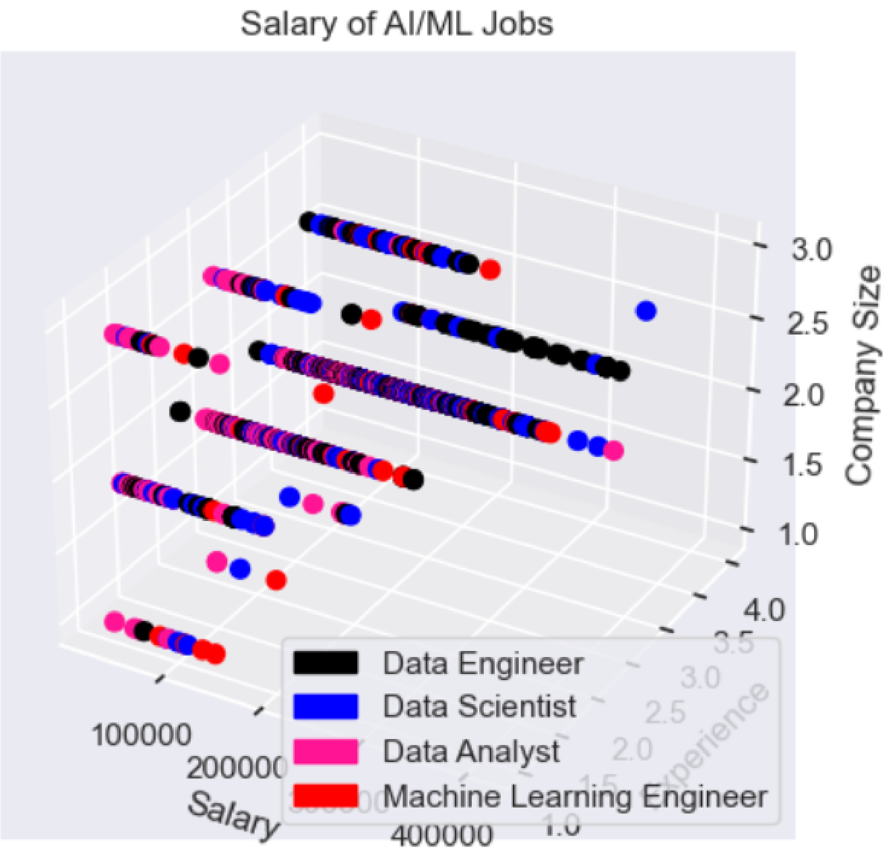


We used KNN to predict the salary of the people in the AI/Machine Learning industry. Our model calculates the distance between each data point (the four features for each individual) and the k = 10 number of nearest neighbors. Then, the model predicts the salary for a new data point based on the average salary of its k nearest neighbors.

The model is trained on a training set and tested on a test set. We first normalize the data for each feature so that we can have them on the same scale. We then fit the training data to the KNN model and then use our test data to predict the salaries. We then made a residual plot to visualize the accuracy of our predictions.

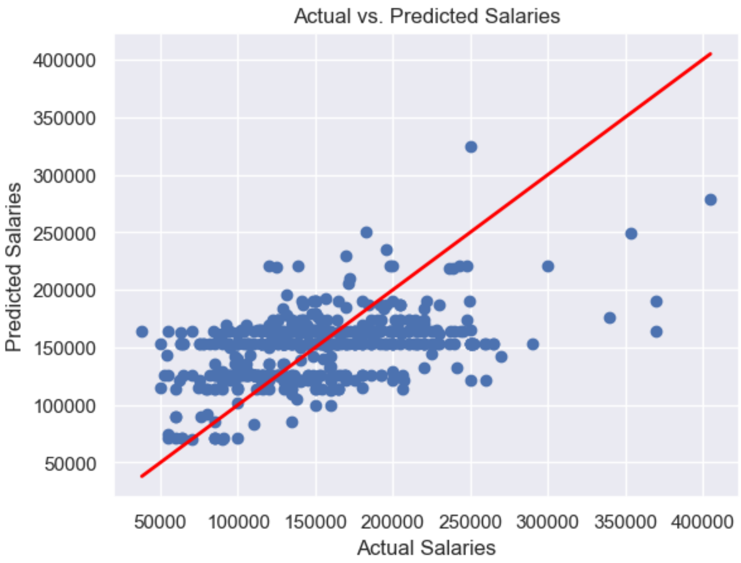
Since our residual plot is showing a positive trend, this indicates that our model is performing poorly, as it is underestimating the predicted values compared to the actual values. This result is likely because of a combination of the issues mentioned above. Because there are so many senior-level engineers in the dataset, with around 2/3, and the rest combined making up only 1/3, there most likely was not enough data to make accurate predictions about the other 3 job experiences. People with lower actual salaries were predicted to have higher salaries, and people with higher actual salaries were predicted to have lower salaries. This is close to what we have explained with our pie chart, histogram/violin plot, and the bar graph.

9. 3D Scatterplot



This 3D plot shows the relationship between salary, experience, and company size of data engineers, scientists, and analysts as well as machine learning engineers, the 4 most common jobs in the AI/ML field. We can use this plot to see some potential clusters in the data. For example, the employees with 1 experience tend to be closer together in salary or that employees at small companies tend to have smaller salaries and less experience.

10. Decision Tree



We tried to verify that our previous conclusion with the KNN was incorrect with a decision tree regression mode. The decision tree works as follows:

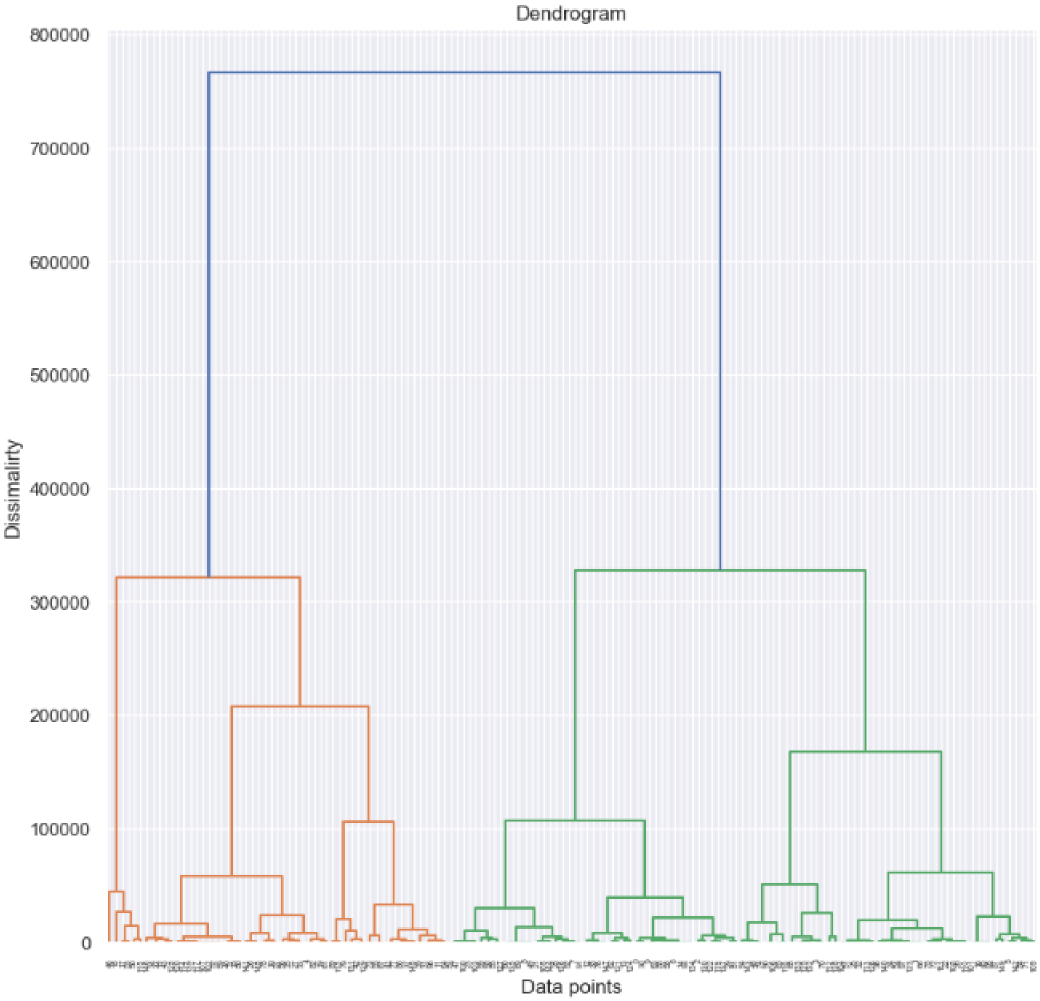
1. Recursively splits data by examining the salary (target) and the rest of the categories based on category features. The goal of splitting is to reduce the variance of the salary values for each subset.
2. Chooses the best split according to MSE, which is done to reduce the variance within the subsets upon splitting
3. Steps 1 and 2 repeat until the tree is done building.

Upon building our decision tree, we found that our MSE was around 220 million. This told us that our model was not accurate at predicting the salaries. We further concluded this by making a graph that plotted the actual vs predicted salaries. The red line is a representation of the actual salaires, and points on the red line indicate a perfect prediction. However, as seen from the model, it performed extremely poorly at predicting the value.

11. Hierarchical Clustering

The second method that we used was hierarchical clustering, specifically ward's linkage clustering. This was used to create the dendrogram below. We removed some data such as the label encoding of the job titles which would reduce the effectiveness of the hierarchical clustering. We removed the label encoding of job titles because there were too many to have any meaningful effect on the clusters created in the dendrogram. We then performed agglomerative clustering using Ward's method. This method minimizes the total variance of each data point in each cluster. We used this because it tends to perform better when there is quantitative data, which our data is. We used a sample of 150 data points to create the dendrogram (shown in the next section). If we had used the entire dataset it would have created a dendrogram that was much harder to read and draw meaningful conclusions from. In the dendrograms that were created using this method it showed that the two largest clusters represented the experience level of employees. Roughly 75% of the data points were in one of the clusters, which matches some of the earlier visualizations that showed that 75% of employees had the highest level of experience. If we split the dendrogram into 4 clusters the sub-clusters represented the salaries of these employees. For example, the sub-cluster that was by far the smallest showed employees with higher salaries for employees without the highest level of experience.

12. Dendrogram



This dendrogram shows hierarchical clustering. It uses a random sample of 150 employees to make it more readable but it tends to show 2 main clusters which typically is followed by 4 sub clusters of similar heights.

Conclusion

In conclusion, we were not able to successfully develop a model within our dataset that can predict the salary of a person in the AI/Machine Learning field based on their employment type, job title, company size, and experience level.

Although our analysis revealed that the salary of a person in the AI/Machine Learning field is slightly correlated with their experience level, we cannot deduce a solid analysis or prediction of one's salary in the AI/ML field is because in our dataset, there was a high overrepresentation of people with senior experience levels, and a high underrepresentation of those in any other experience level.

This will cause much more points with senior-level experience level to be on the graph, resulting in an increased likelihood of new predicted data having inappropriate neighbors. Since we predict the salary by taking the average salary of the nearest neighbors, this will cause issues. The predicted salary for people with lower salaries than senior level experience will be higher than their actual, and the predicted salary for people with higher salaries than senior level experience will be lower than their actual. Therefore, our model cannot accurately predict the salary of people in the AI/Machine Learning field with our dataset.

Contributions

Anjali Daryani:

- Final Report
- Project Proposal and Final Project Slides Presentations
- Presenting introduction, goal of project, cleaning of data and scatter plot analysis
- Found dataset and formed project question
- Made three questions

Michael Bettencourt:

- Final project slides
- Found data set
- Created 3D chart and dendrogram (hierarchical clustering)
- Description for hierarchical clustering

Nathanael Shin:

- Final Report
- Descriptions for graphs made from EDA
- Made residual graph for KNN model
- calculated MAE for KNN model
- Made violin plot more readable
- fixed misc. bugs

Neha Mathews:

- Created pie chart, heatmap, and violin chart and added the descriptions
- Normalized data and created KNN model
- Calculated MSE for model
- Added to final project slides

Rahul Katwala:

- Cleaned dataset
- Created histogram, scatterplot, and bar chart and added the descriptions
- Created elbow plot and found optimal K
- Added to final project slides