Megan Vaughn

Cai

STA4365

April 18, 2024

Analyzing the Gender Pay Gap

The objective of this report is to determine if a pay gap between genders for the same job title exists. The dataset was found on Kaggle, but is originally from Glassdoor, a website where current and former employees anonymously review companies and are able to post their current position, pay, and other various information. The dataset included features such as Job Title, Gender, Age, PerfEval (Performance Evaluation), Education, Department, Seniority, Base Pay, and Bonus.

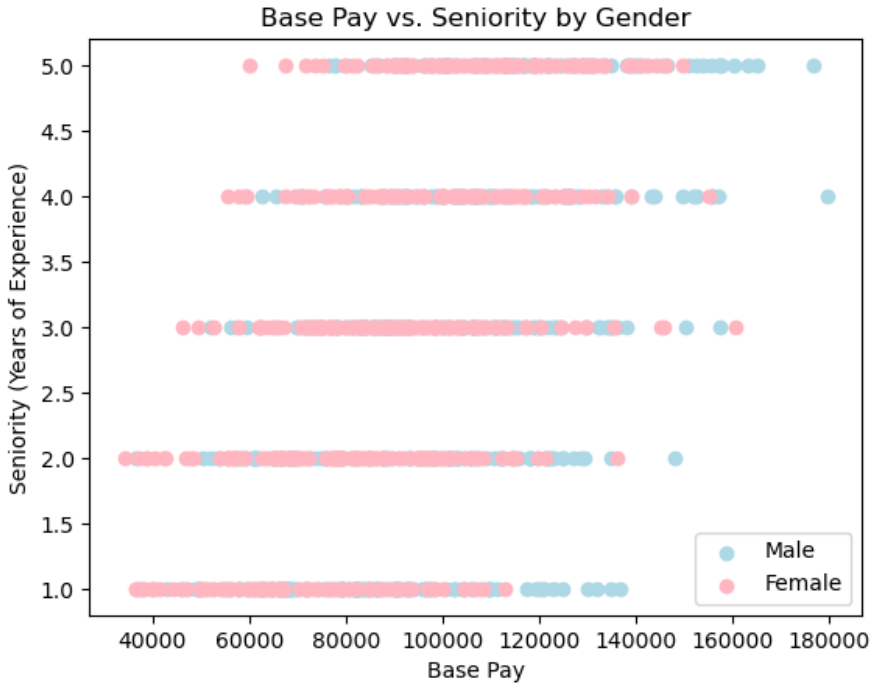
Logistic regression was performed using various python packages such as, pandas, numpy, sklearn, and statsmodels. During the logistic regression process, I made Base Pay the target variable, extracted the non-numeric columns and casted them as categorical variables, and created the features of the model with the previously listed features. I then used preprocessing steps and split the data into training and testing sets and fitted the logistic regression model. The R-Squared of the model was fairly strong at 0.839, and the Adjusted R-Squared was also strong, having a value of 0.834. The R-Squared shows that the logistic regression model is strong and is significant.

OLS Regression Results

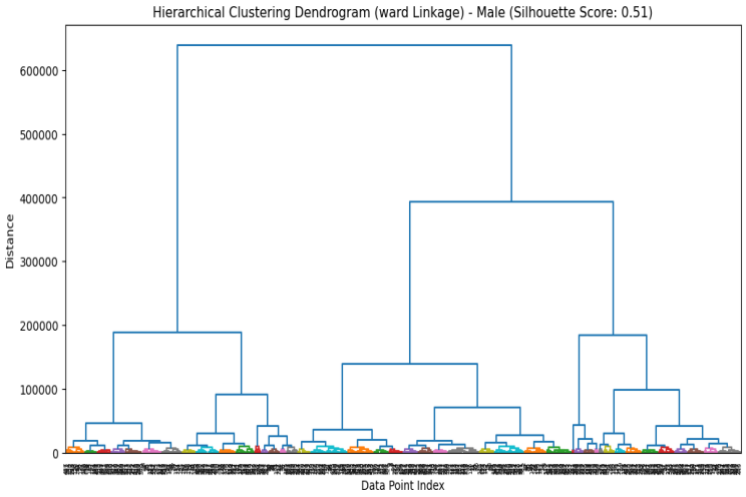
Dep. Variable:	BasePay	R-squared:	0.839
Model:	OLS	Adj. R-squared:	0.834
Method:	Least Squares	F-statistic:	192.8
Date:	Wed, 10 Apr 2024	Prob (F-statistic):	3.49e-291
Time:	18:12:59	Log-Likelihood:	-8518.0
No. Observations:	800	AIC:	1.708e+04
Df Residuals:	778	BIC:	1.718e+04
Df Model:	21		
Covariance Type:	nonrobust		

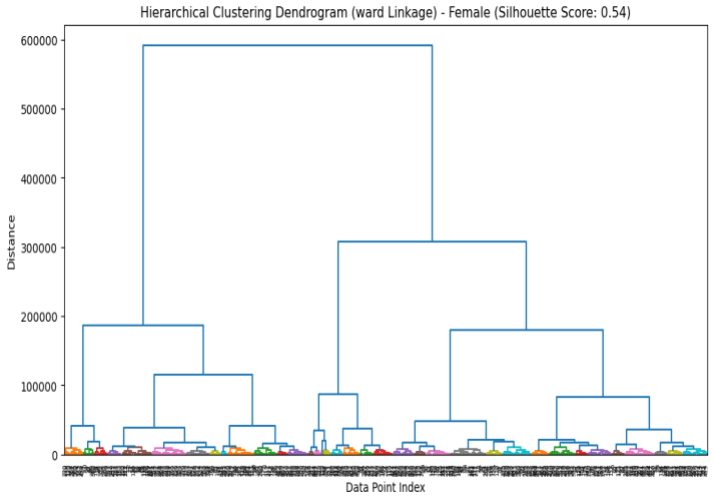
	coef	std err	t	P>|t|	[0.025	0.975]
const	1.282e+04	1669.977	7.675	0.000	9538.437	1.61e+04
x1	-1089.9129	1040.037	-1.048	0.295	-3131.525	951.699
x2	-4541.6003	1218.523	-3.727	0.000	-6933.583	-2149.618
x3	2523.6627	1121.120	2.251	0.025	322.885	4724.441
x4	-3321.9291	1125.612	-2.951	0.003	-5531.526	-1112.332
x5	-3218.8975	1146.226	-2.808	0.005	-5468.960	-968.835
x6	3.074e+04	1229.125	25.011	0.000	2.83e+04	3.32e+04
x7	-1.815e+04	1081.027	-16.794	0.000	-2.03e+04	-1.6e+04
x8	-366.5487	1117.885	-0.328	0.743	-2560.978	1827.880
x9	1.202e+04	1138.132	10.562	0.000	9786.627	1.43e+04
x10	-1775.5685	1210.261	-1.467	0.143	-4151.332	600.195

A scatterplot was made using the matplotlib package. The data was split into male and female subsets and plotted in different colors. In the scatterplot, males have a higher base pay with the seniority level as females. The outliers within the higher base pay range are male. The outliers within the lower base pay range are female.

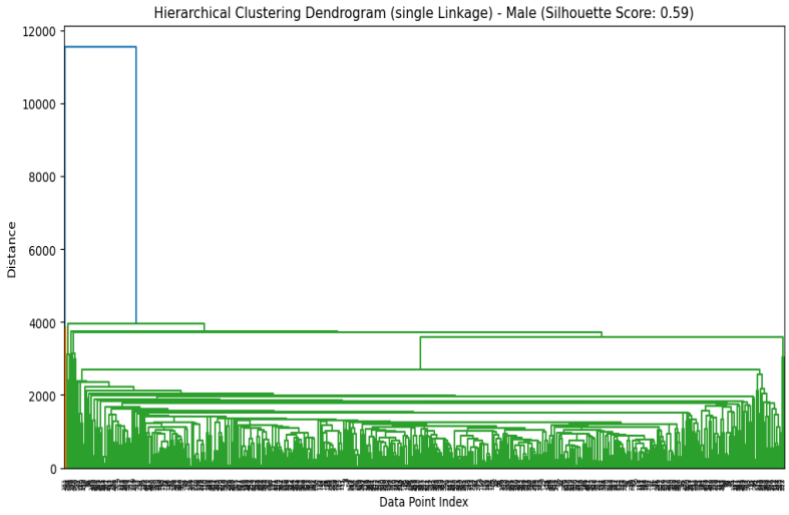


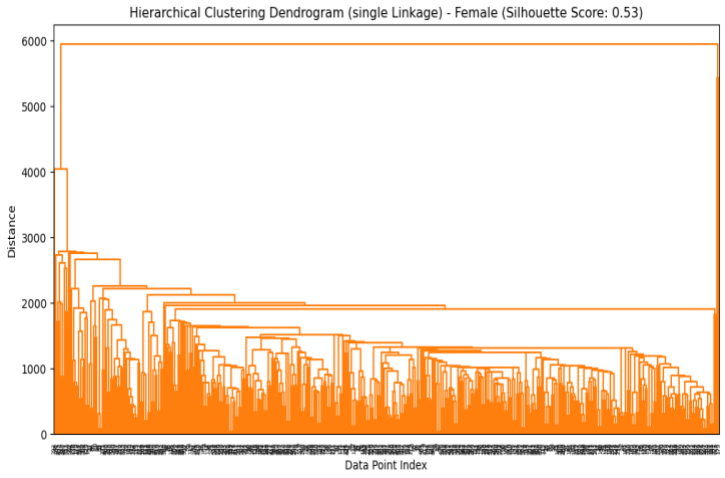
Centroid linkage using hierarchical clustering was used to compare the linkages in the male and female plots. The sklearn, scipy, and matplotlib packages were used to visualize the clusters with dendrograms. Both the male and female centroid linkage dendrograms have a favorable Silhouette Score, with the female score being slightly higher. This means the data points are well matched to their respective clusters. The male score was 0.52 and the female score was 0.54.



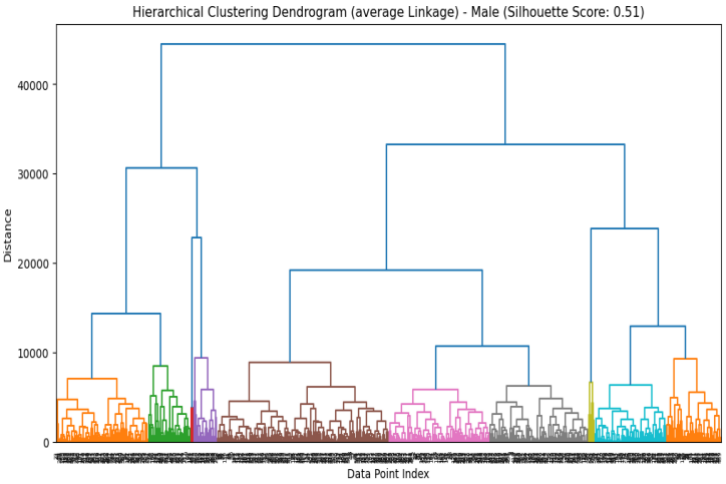


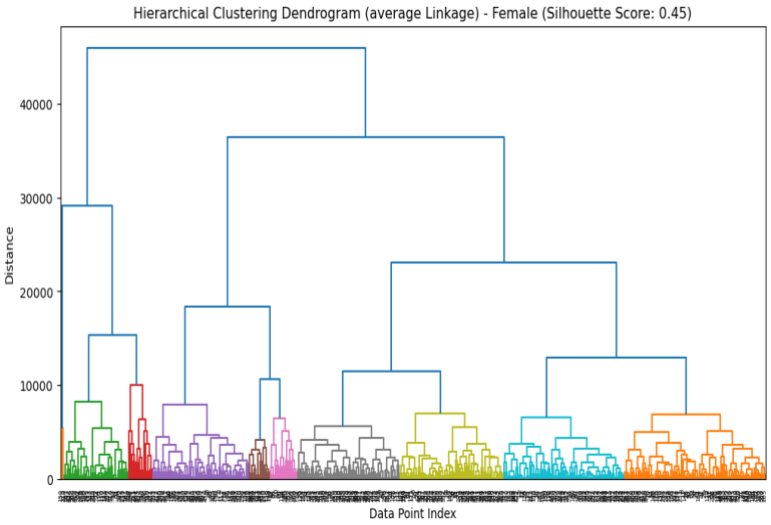
Single linkages using hierarchical clustering was then performed with the sklearn, scipy, and matplotlib packages. Both the male and female dendrograms have favorable Silhouette Scores, with the male score being higher. The data points are well matched within their clusters, however, both dendrograms contain very wide clades, making the dendrogram difficult to interpret. The male score was 0.59 and the female score was 0.53.



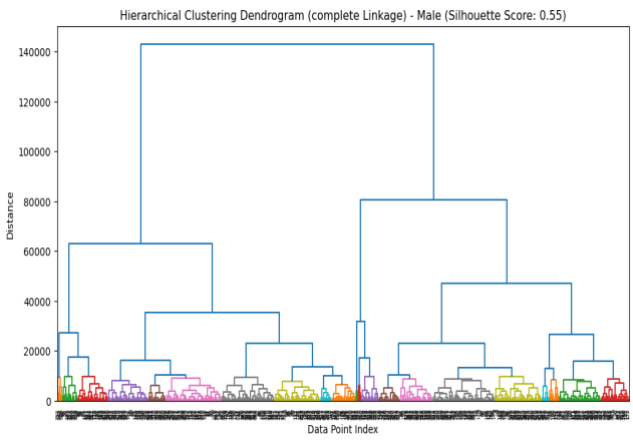


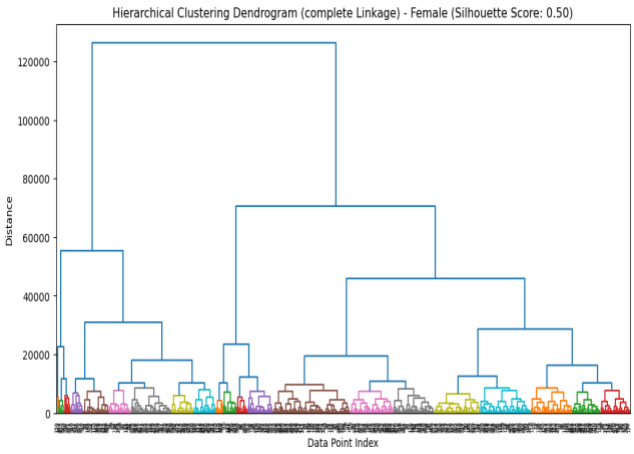
Average linkage was then determined using hierarchical clustering. The python packages sklearn, scipy, and matplotlib were used. The male Silhouette Score was higher than the female, a lower Silhouette Score indicates overlapping cluster or poorly separated data points. The male score was 0.51 and the female score was 0.45.





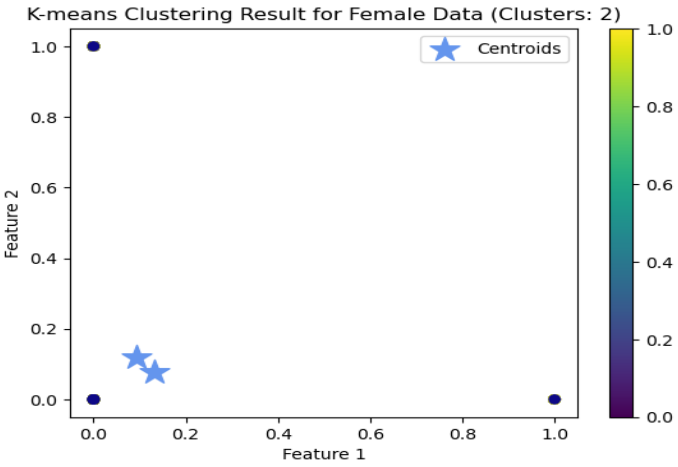
Complete linkage was then performed using hierarchical clustering. python packages sklearn, scipy, and matplotlib were used to visualize the data using dendrograms. The male Silhouette Score is higher indicating that the clusters are more separated and there are fewer overlapping clusters than in the female plot. The male score was 0.55 and the female score was 0.50.





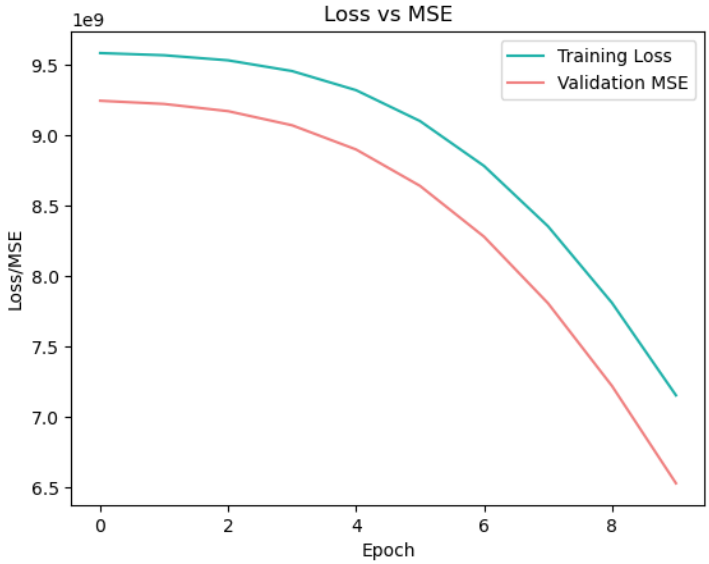
Overall, the single and average linkages had the largest differences in Silhouette Scores between male and female. Single linkage had the highest Silhouette Score for males, 0.59. Centroid linkage had the highest Silhouette Score for females, 0.54.

Next, K-Means Clustering was performed using the sklearn and matplotlib packages. A range of clusters was determined to find the optimal number of clusters for each gender. Both the male and female plot have centroids in the bottom left corner of the plot. The centroids are closer together in the female plot, meaning their Silhouette Scores would be lower due to overlapping clusters. Both the male and female plots' points are in the same cluster which leads me to believe there was an error during the preprocessing stage.



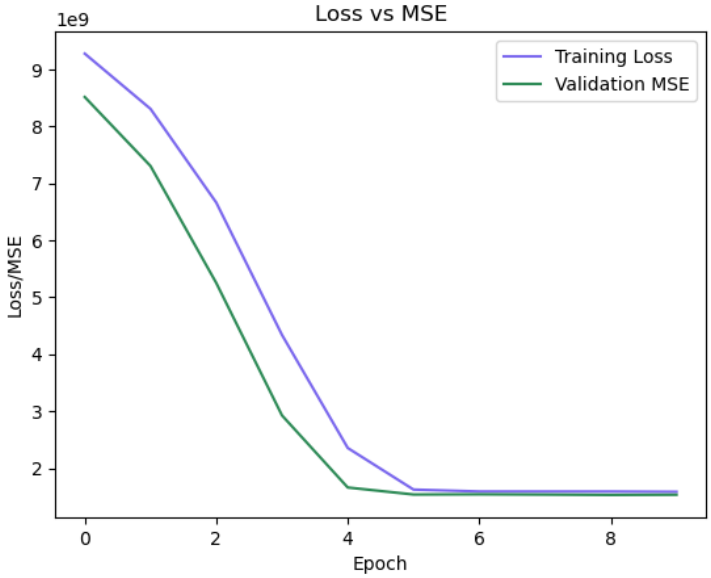
Convolutional Neural Networks (CNN) were then created using the sklearn, tensorflow, and matplotlib packages. The data was reshaped for CNN input, and the model was compiled and trained with the data to create a plot of Loss v. MSE. Training Loss and Validation MSE

both have negative slope. As the neural network is trained, the cycle increases, and the Loss and MSE decrease exponentially.




```
Epoch 1/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 2s 18ms/step - loss: 9517377536.0000 - val_loss: 9245394944.0000
Epoch 2/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 9643018240.0000 - val_loss: 9222652928.0000
Epoch 3/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - loss: 9527646208.0000 - val_loss: 9171393536.0000
Epoch 4/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 9216134144.0000 - val_loss: 9071098880.0000
Epoch 5/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 9431699456.0000 - val_loss: 8899126272.0000
Epoch 6/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 9272471552.0000 - val_loss: 8640323584.0000
Epoch 7/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 8837728256.0000 - val_loss: 8279440384.0000
Epoch 8/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 8438584832.0000 - val_loss: 7806589440.0000
Epoch 9/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 7983575552.0000 - val_loss: 7218782208.0000
Epoch 10/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 7470205952.0000 - val_loss: 6527037952.0000
```

Feed Forward Neural Network (FNN) was created using the sklearn, tensorflow, and matplotlib packages. The FNN model was defined, complied, and trained to create a plot of Loss v. MSE. Training Loss and Validation MSE both have negative slope. As the neural network is trained, and the cycle increases, the Loss and MSE decrease exponentially.




```
25/25 ━━━━━━━━━━━━━━━━━━━━ 2s 11ms/step - loss: 9537742848.0000 - val_loss: 8513298432.0000
Epoch 2/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 8566194688.0000 - val_loss: 7302814720.0000
Epoch 3/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 7463686656.0000 - val_loss: 5250487808.0000
Epoch 4/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 4865720320.0000 - val_loss: 2926100992.0000
Epoch 5/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 2780562944.0000 - val_loss: 1661103232.0000
Epoch 6/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step - loss: 1608840320.0000 - val_loss: 1537989504.0000
Epoch 7/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 1577142528.0000 - val_loss: 1541173376.0000
Epoch 8/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 1729822720.0000 - val_loss: 1536525568.0000
Epoch 9/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 1527928704.0000 - val_loss: 1530213760.0000
Epoch 10/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 8ms/step - loss: 1559807872.0000 - val_loss: 1533918080.0000
```

In conclusion, there are some discrepancies in the salaries of males and females with the same levels of experience (seniority). More analysis could be done with the dataset using regression models. More emphasis could be placed on education level and performance evaluation scores. Bonuses were not included in the analysis because they can fluctuate based on performance evaluation, financial stability of the employer, and outside factors, such as the stock market. For example, employees would be less likely to get a bonus during a recession, regardless of their performance.