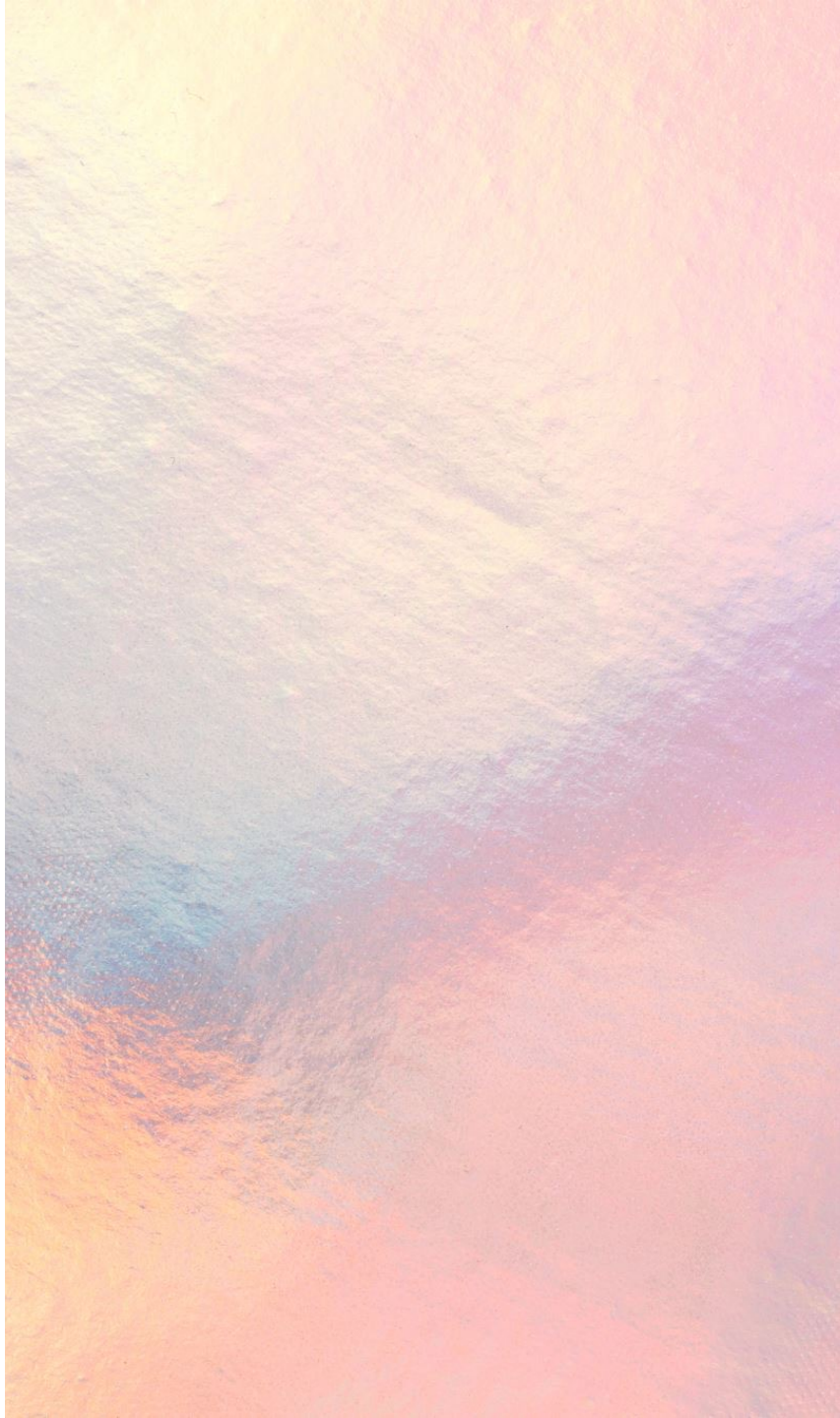
ANALYZING THE GENDER PAY GAP

Megan Vaughn

STA4365



DATA INFORMATION

	JobTitle	Gender	Age	PerfEval	Education	Dept	Seniority	BasePay	Bonus
0	Graphic Designer	Female	18	5	College	Operations	2	42363	9938
1	Software Engineer	Male	21	5	College	Management	5	108476	11128
2	Warehouse Associate	Female	19	4	PhD	Administration	5	90208	9268
3	Software Engineer	Male	20	5	Masters	Sales	4	108080	10154
4	Graphic Designer	Male	26	5	Masters	Engineering	5	99464	9319
...	...	...	...	...	...	...	...	...	...
995	Marketing Associate	Female	61	1	High School	Administration	1	62644	3270
996	Data Scientist	Male	57	1	Masters	Sales	2	108977	3567
997	Financial Analyst	Male	48	1	High School	Operations	1	92347	2724
998	Financial Analyst	Male	65	2	High School	Administration	1	97376	2225
999	Financial Analyst	Male	60	1	PhD	Sales	2	123108	2244

- Objective: to find the pay gap between the gender for the same job title.
- Features: Job Title, Gender, Age, PerfEval (Performance Evaluation), Education, Dept , Seniority, Base Pay, and Bonus
- Originally from Glassdoor but the dataset itself is from Kaggle
- Link: https://www.kaggle.com/datasets/nilimajauhari/glassdoor-analyze-gender-pay-gap/data

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

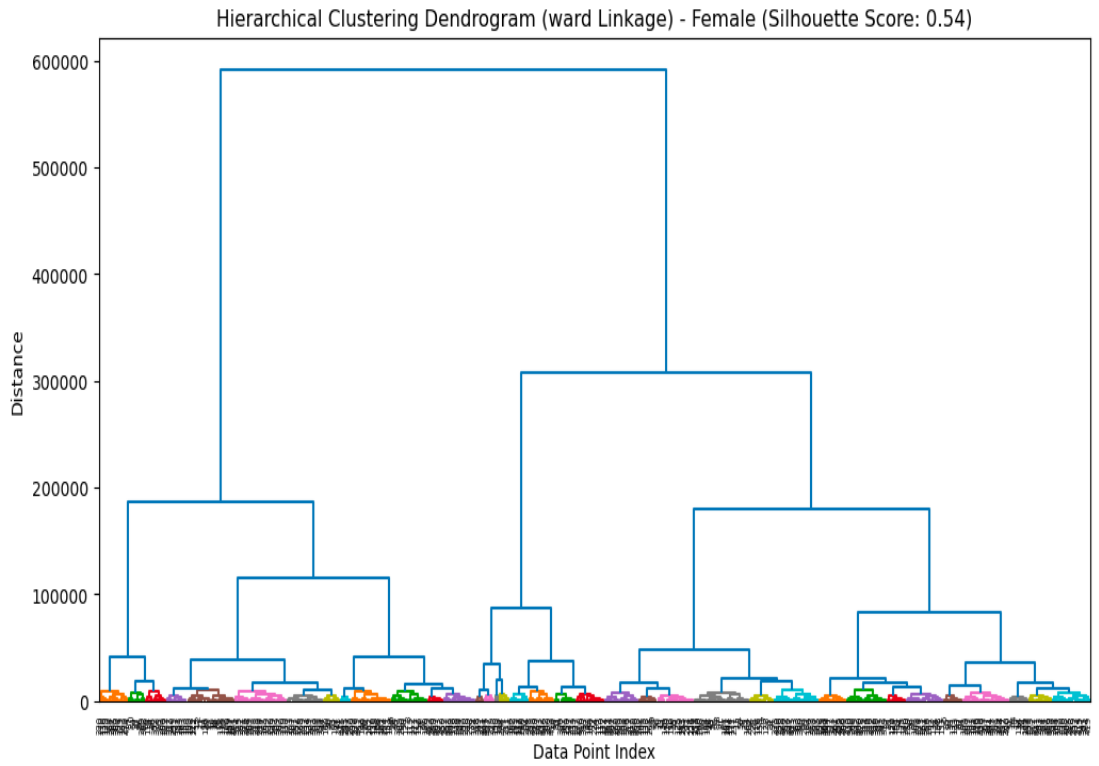
x1: JobTitle
x2: Gender
x3: Age
x4: PerfEval
x5: Education
x6: Dept
x7: Seniority
x8: Base Pay
x9: Bonus
x10: Constant

LOGISTIC REGRESSION

- R-Squared: 0.839
- The model using the features listed to predict Base Pay is strong.
- Gender, Department, Seniority, and Bonus are the most significant features.

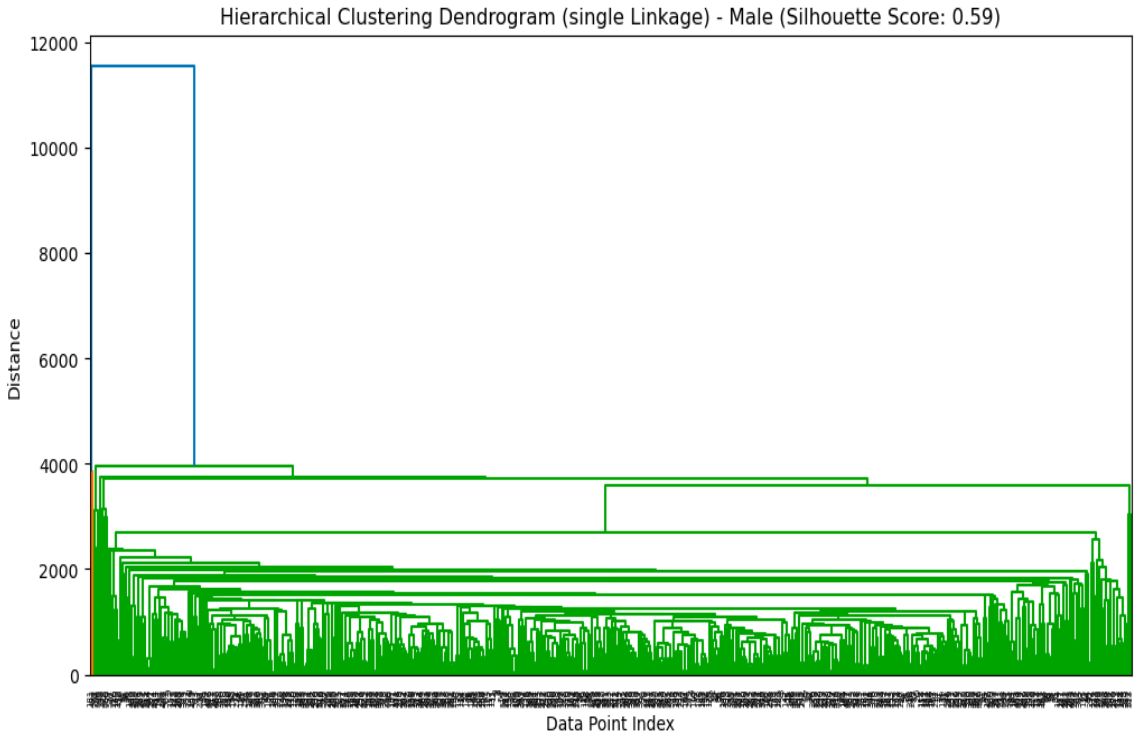
PLOT OF BASE PAY V. SENIORITY

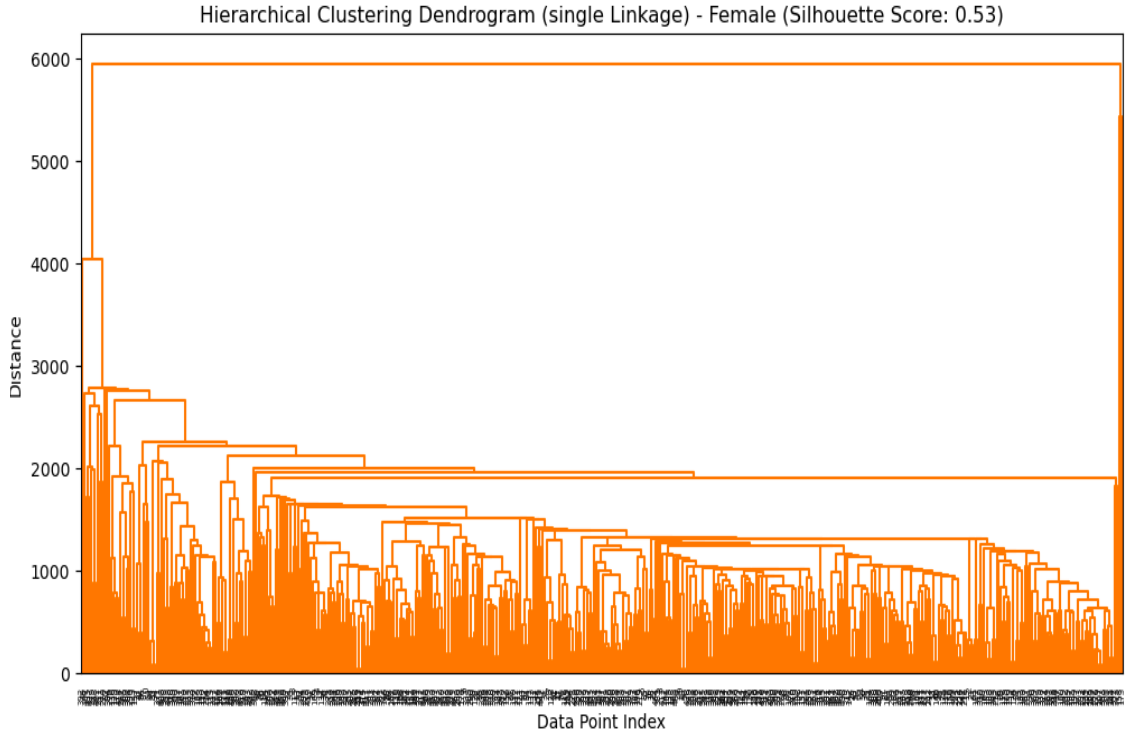
- Males have a higher base pay with the same seniority as females.
- Males have a higher base pay with the same seniority as females.
- The outliers within the higher base pay range are male.
- The outliers within the lower base pay range are female.



CENTROID LINKAGE MALE V. FEMALE

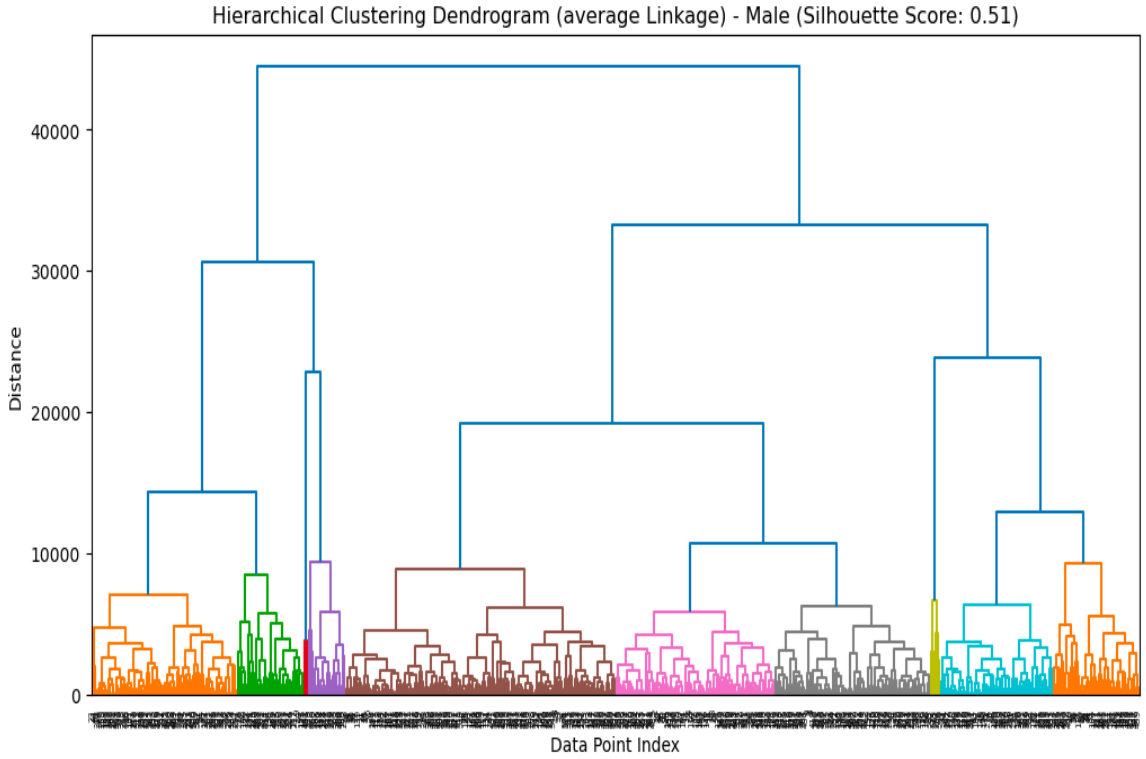
Both the male and female centroid linkage dendrograms have a favorable Silhouette Score, with the female score being slightly higher, this means the points are well matched to their respective clusters.

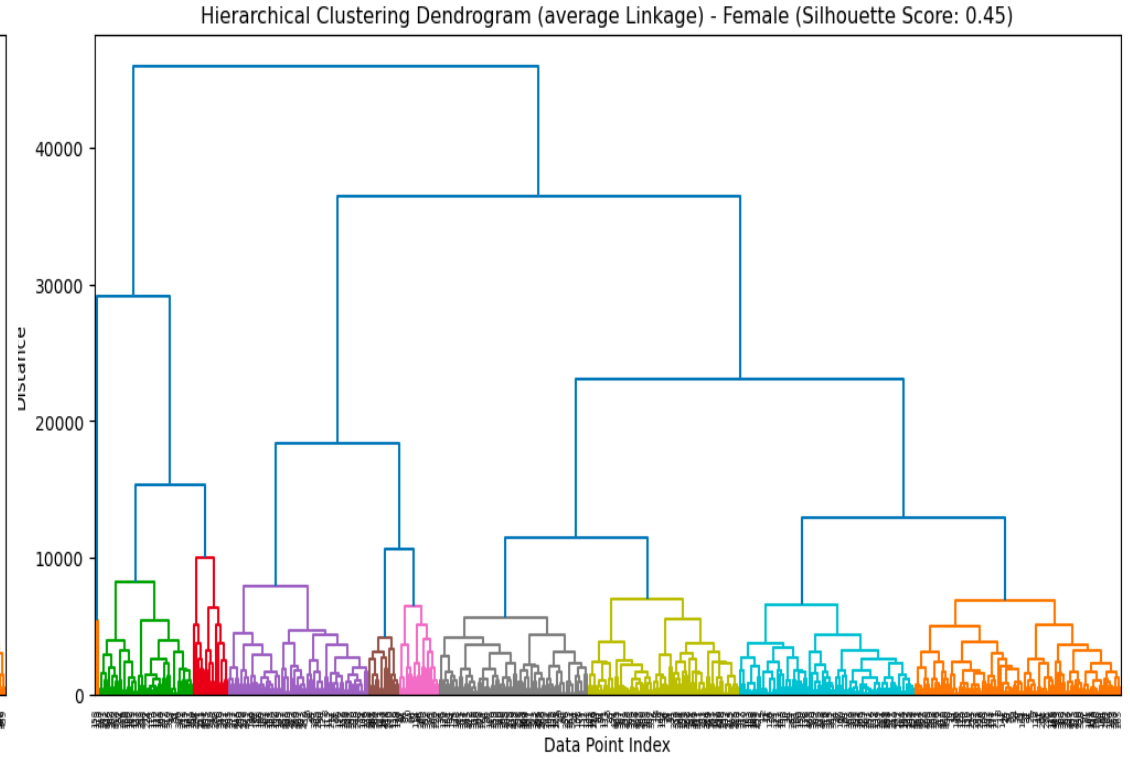




SINGLE LINKAGE MALE V. FEMALE

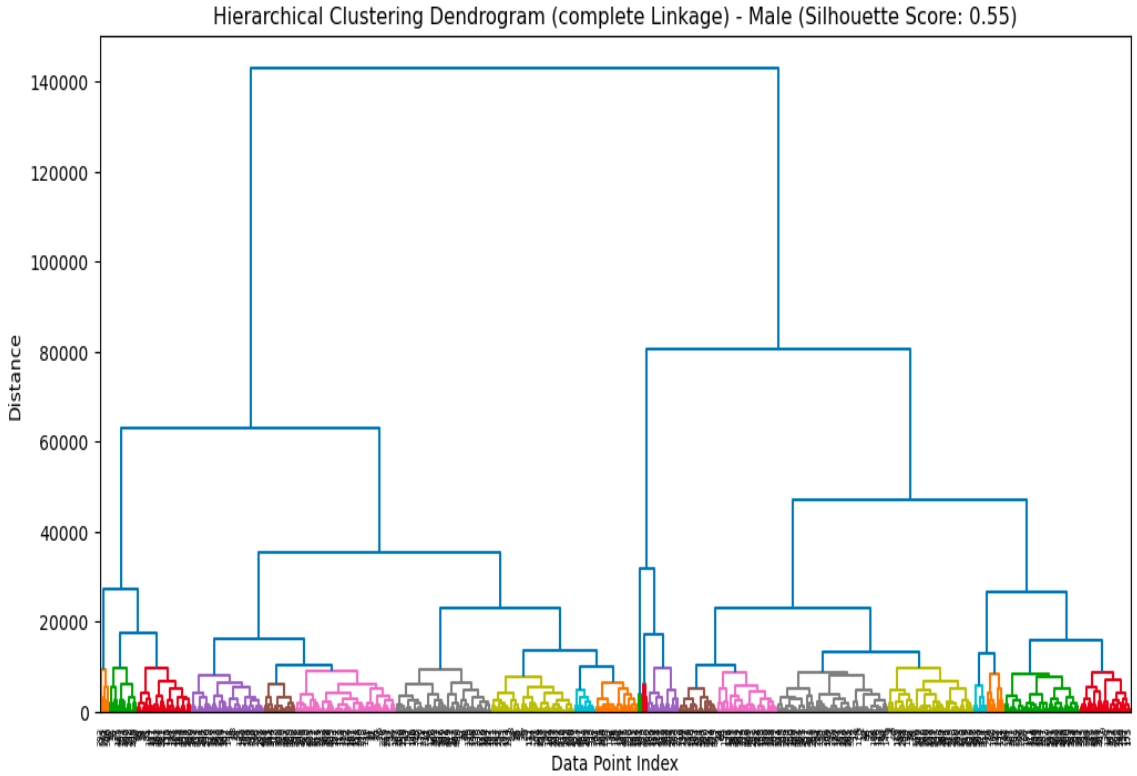
Both dendrograms have a favorable Silhouette Score, with the male score being higher, the points are well matched within their clusters, both dendrograms contain very wide clades.

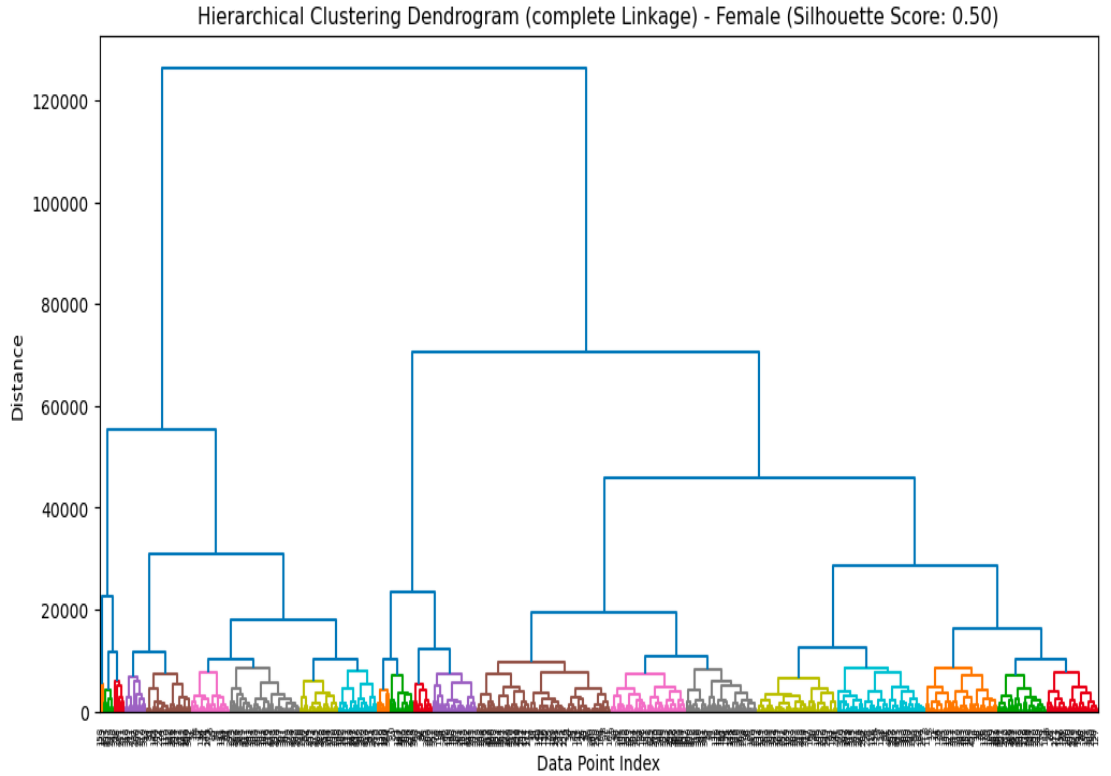




AVERAGE LINKAGE MALE V. FEMALE

The male Silhouette Score was higher than the female, a lower Silhouette Score indicates overlapping clusters or poorly separated data points.



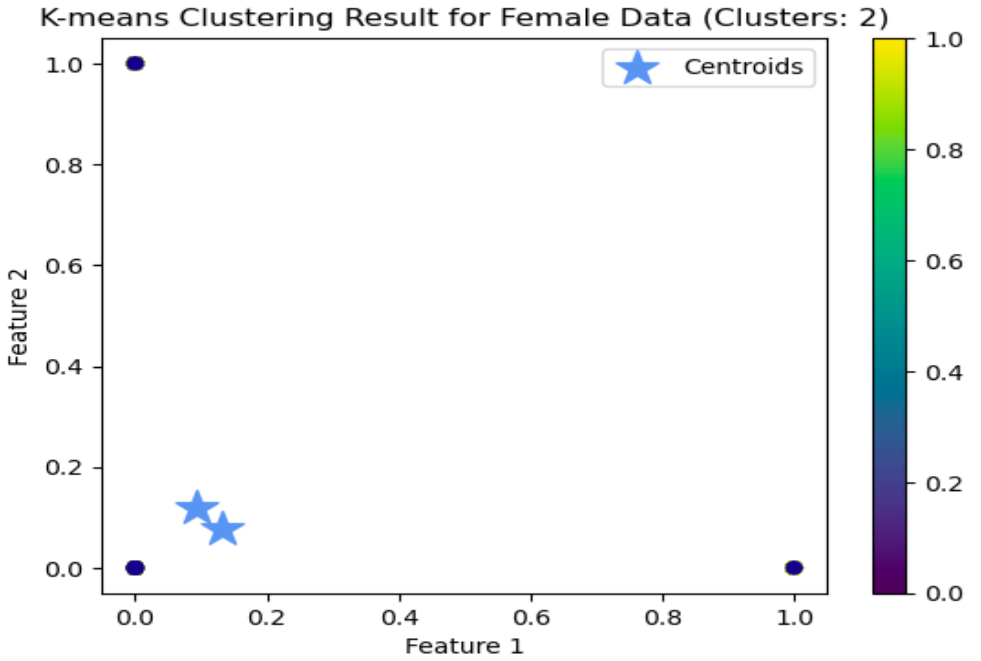


COMPLETE LINKAGE MALE V. FEMALE

The male Silhouette Score is higher indicating that the clusters are more separated and there are fewer overlapping clusters than in the female plot.

HIERARCHICAL CLUSTERING WITH LINKAGES

- Single and average linkages had the largest difference 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.



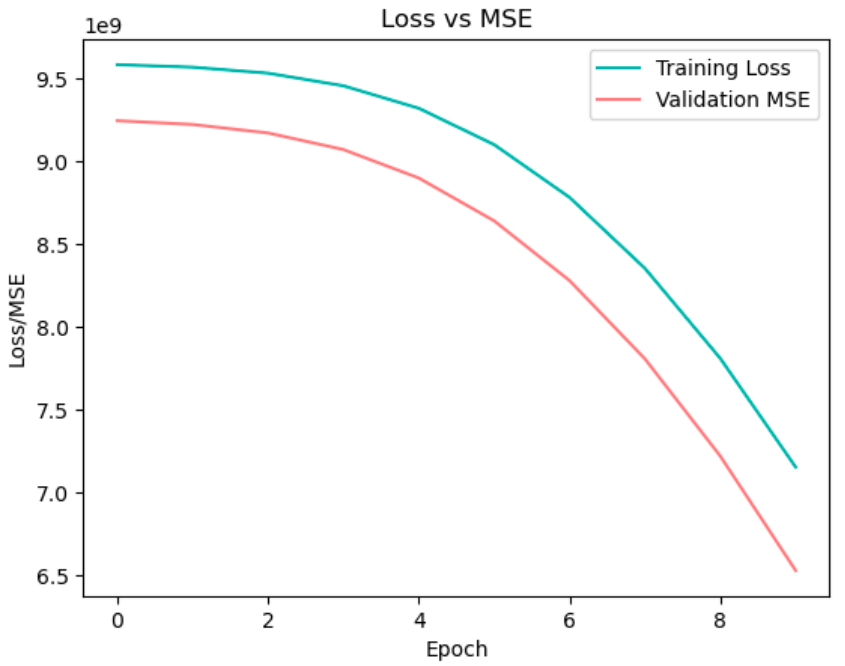
K-MEANS CLUSTERING MALE V. FEMALE

- Both the male and female plot have centroids in the bottom left 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 in the preprocessing stage.

CONVOLUTIONAL NEURAL NETWORK (CNN)

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

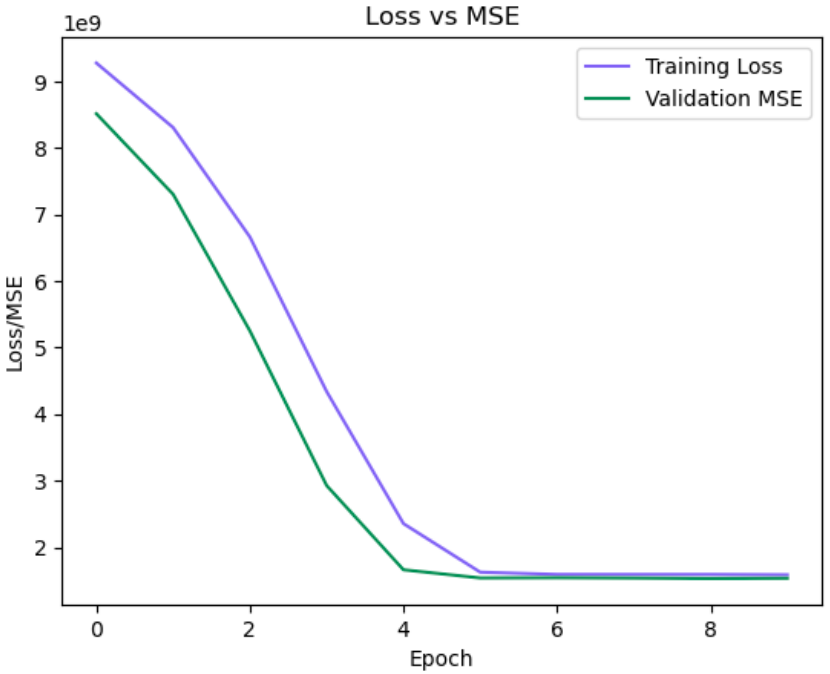
```
Epoch 1/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 2s 18ms/step - loss: 9517377536.0000 - val_loss: 9245394944.0000
Epoch 2/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 9643018240.0000 - val_loss: 9222652928.0000
Epoch 3/10
25/25 ━━━━━━━━━━━━━━━━━━━━ 0s 5ms/step - loss: 9527646208.0000 - val_loss: 9171393536.0000
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)

Training Loss and Validation MSE both have negative slope. As the neural network is trained, the cycle increases, the Loss and Validation 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
```



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 regression methods.
- 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 i.e., the stock market.