

Problem Statement: The primary objective is

1. Profile the best companies and job positions from Scaler's database.
2. To execute clustering techniques.
3. Evaluate the coherence of your clusters, and provide actionable insights for enhanced learner profiling and course tailoring.
4. To profile learners effectively to tailor course content, ensuring a more personalized and impactful learning experience.

Questionnaire:

1. What percentage of users fall into the largest cluster?

Answer: We've created 20 clusters based on 4 primary features Company, CTC, Appraisal, Year of Experience and in largest cluster we got : 10% (2595/26457) of my complete data.

2. Comment on the characteristics that differentiate the primary clusters from each other.

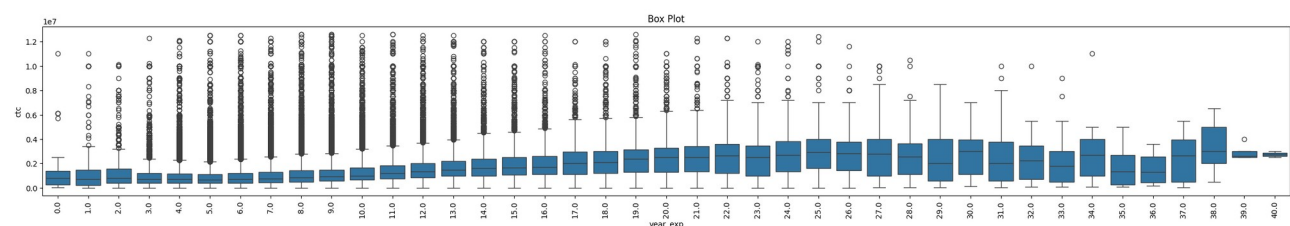
Primary Cluster characteristics includes:

1. Salary bin Medium (5,00,000 to 15,00,000)
2. Exp bin Medium (5 years to 10 years)
3. Last updated CTC was more than 3 years back

	company_hash	salary_bin_Low	salary_bin_Medium	salary_bin_High	salary_bin_Very High	exp_bin_Low	exp_bin_Medium	exp_bin_High	ctc_updated_2020	ctc_updated_2021	ctc_updated_other	label
...	2579	0.762515	0	1	0	0	0	1	0	0	0	1 9
	20848	0.774683	0	1	0	0	0	1	0	0	0	1 9
	17390	0.747821	0	1	0	0	0	1	0	0	0	1 9
	28486	0.777796	0	1	0	0	0	1	0	0	0	1 9
	24131	0.810580	0	1	0	0	0	1	0	0	0	1 9

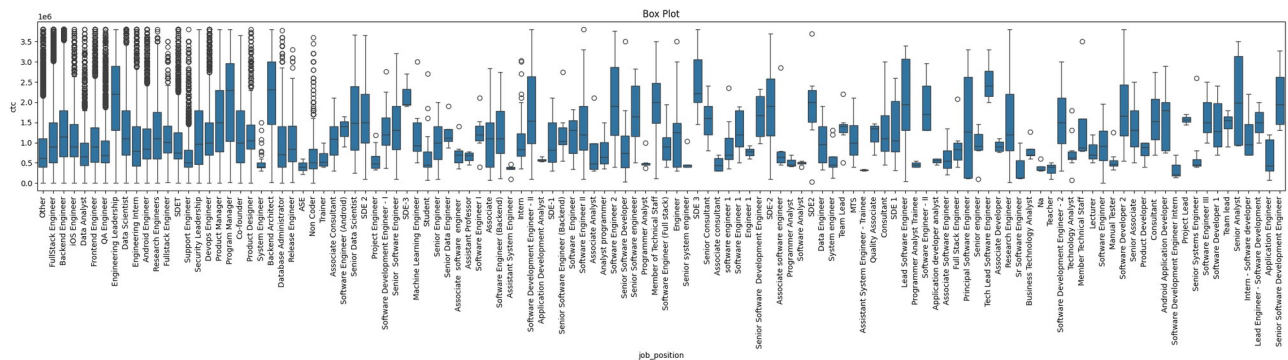
3. Is it always true that with an increase in years of experience, the CTC increases? Provide a case where this isn't true ?

Yes, not CTC but the mean of CTC in general increase.



4. Name a job position that is commonly considered entry-level but has a few learners with unusually high CTCs in the dataset.

From the graph, we can visualize that Data Scientist and Research Engineers may consider as entry-level but has few learners with ususally high CTCs.



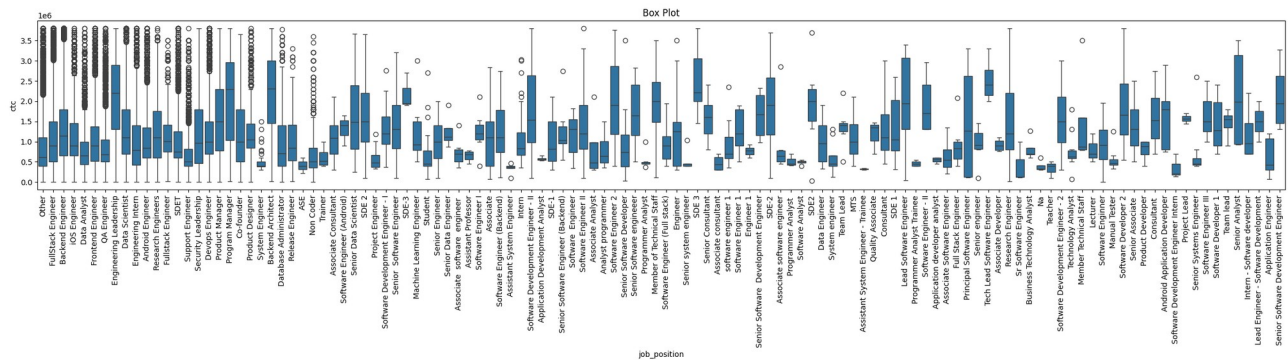
5. What is the average CTC of learners across different job positions?

The meadian seems to be 9.5 lakhs and mean seems to be 22 lakhs across all the learners but mean is not the right metric here we believe.

	orgyear	ctc	ctc_updated_year
count	205757.000000	2.058430e+05	205843.000000
mean	2014.882750	2.271685e+06	2019.628231
std	63.571115	1.180091e+07	1.325104
min	0.000000	2.000000e+00	2015.000000
25%	2013.000000	5.300000e+05	2019.000000
50%	2016.000000	9.500000e+05	2020.000000
75%	2018.000000	1.700000e+06	2021.000000
max	20165.000000	1.000150e+09	2021.000000

6. For a given company, how does the average CTC of a Data Scientist compare with other roles?

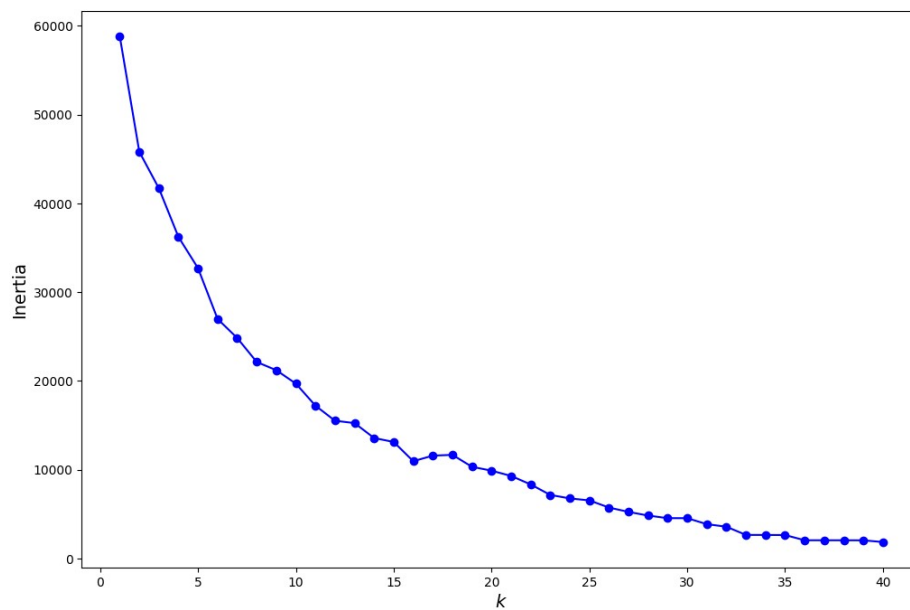
It seems in comparison to most of the other roles the average CTC is higher in terms of median and mean both.



8. After performing unsupervised clustering:

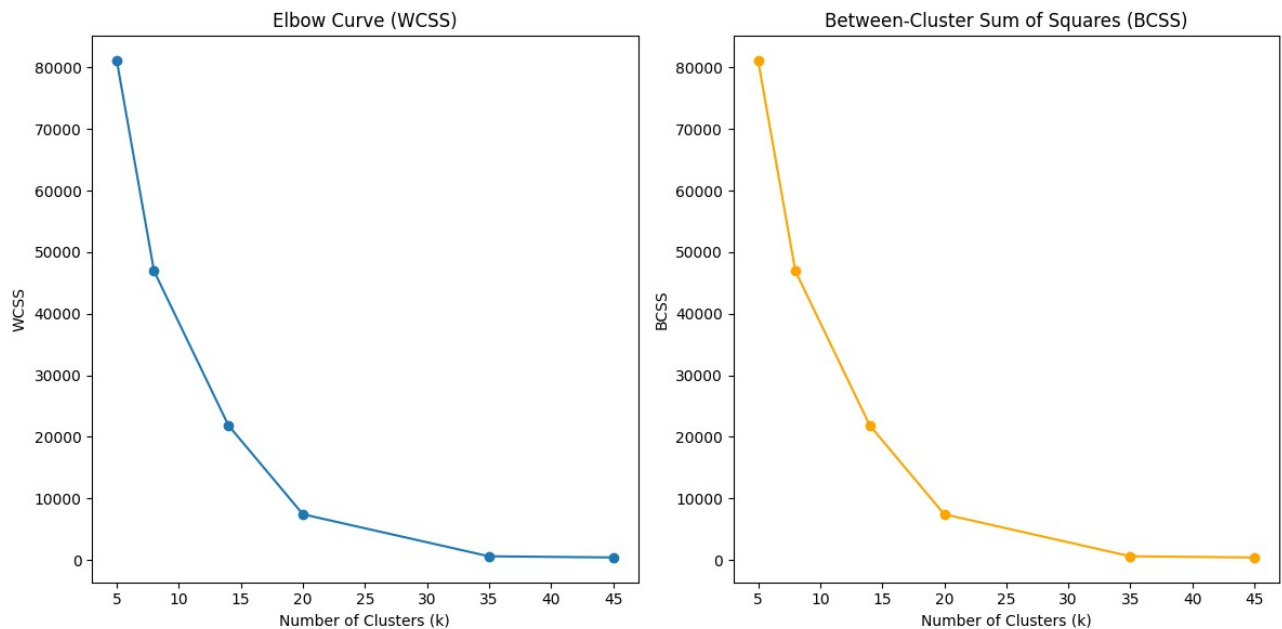
1. How many clusters have been identified using the Elbow method?

It seems 8 clusters looks good, but 20 also seems to be a good number for deciding the number of clusters to go with.



2. Do the clusters formed align or differ significantly from the manual clustering efforts? If so, in what way?

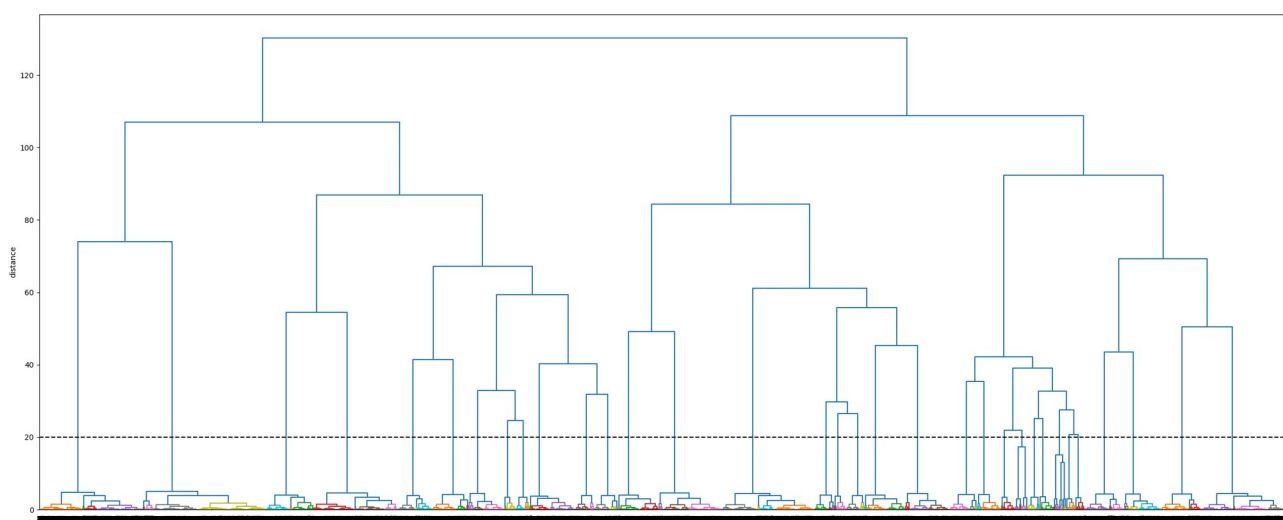
On 8-D data can't say much but the WCSS values and BCSS values seems promising and clusters seems in good shape.



9. From the Hierarchical Clustering results:

1. Are there any clear hierarchies or patterns formed that could suggest the different levels of seniority or roles within a company?

Yes, there is clear hierarchical/pattern observed while performing hierarchical clustering.



2. How does the dendrogram representation correlate with the 'Years of Experience' feature?

From the attached visualization, different clusters seems to represent different segment of data clearly:

	company_hash	salary_bin_Low	salary_bin_Medium	salary_bin_High	salary_bin_Very High	exp_bin_Low	exp_bin_Medium	exp_bin_High	ctc_updated_2020	ctc_updated_2021	ctc_updated_other	Cluster
0	0.691859	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0
1	0.831993	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0
2	0.617750	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	2.0
3	0.762515	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	3.0
4	0.821460	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	4.0
5	0.718640	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	5.0
6	0.736795	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	6.0
7	0.864317	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	7.0
8	0.957344	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	8.0
9	0.716022	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	9.0
10	0.770263	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	10.0
11	0.818774	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	11.0
12	0.743082	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	12.0
13	0.770320	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	13.0
14	0.778515	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	14.0
15	0.730386	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	15.0
16	0.664193	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	16.0
17	0.549593	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	17.0
18	0.867860	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	18.0
19	0.918840	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	19.0

<END>