BANA 212_Project report

STEM Salary Compensation Analysis and Prediction

Team 19

Harinishri Srikanth, Pei-Ling Wu, Praveen Aravindar, Xinfa Pan

Table of Contents

1. Introduction

Topic the team choose to focus on

The team chose to focus on the topic of STEM-related salaries in top companies. We found a database regarding this topic that has the variable of education level, compensation (base salary, bonus, stock grants), race, years of experience and more. And we are interested in finding relations between these variables and salary.

Brief description of why we proposed this topic

No one doesn't care about salary. Salary is always an interesting and sensitive issue on the job market. In one year, we will face job hunting and salary negotiation as a STEM graduate student, and this raises our curiosity to learn more about what is the range on the market and if it is driven by any certain variables which we could start to prepare now.

We are generally interested in salary in STEM related position to see the correlation between salary and qualifications; since salary, companies, and company cultures are big parts of what the job seekers look at when they are searching for positions. Also, while doing this project, it also allows us to understand what we expect in terms of qualifications and companies for STEM positions.

What questions you attempt to answer from the data

"STEM Salary Compensation Analysis and Prediction"
The ultimate question we are trying to answer is what attributes play an important role in STEM salaries. It could also be extended to qualifications for companies when we are hunting for jobs. For instance, we would like to understand what the minimal requirements or qualifications for education for a certain range of salary are. If everyone in these companies has a college degree, we can assume a college degree is the minimum requirement in terms of education for the companies.

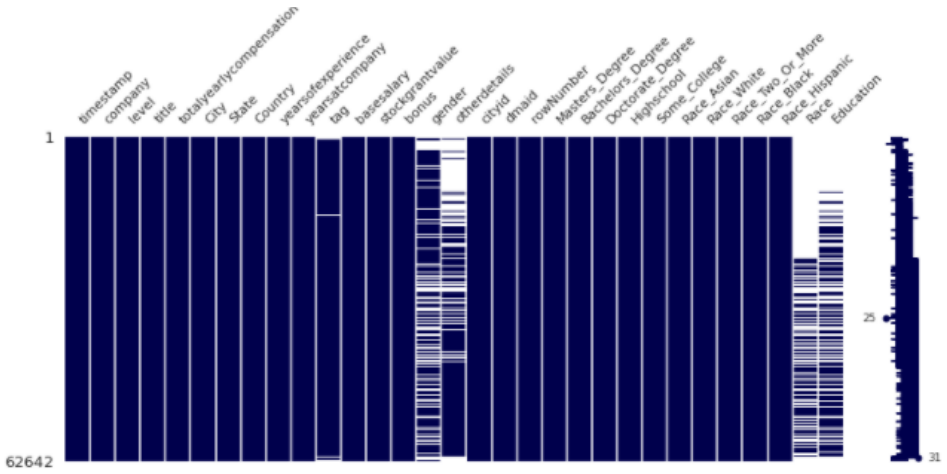
Data Source and Analysis Method

"Data Science and STEM Salaries- 62,000+ STEM salaries scraped from levels.fyi" from Kaggle is the dataset we will research on. Levels.fyi is a website for any anonymous to report their salary and could be used to compare career levels and compensation packages across different

companies. This dataset includes 29 attributes and 62,000 records dated ranging from Jun. 7th, 2017, to Aug. 17th, 2021. Key attributes including company name, company level, title, location, total yearly compensation, race, education level etc.

First, we will start with exploratory data analysis in order to have some general idea about the dataset. Second, clustering is one of the analysis methods we plan to use in order to find key drivers of salary. Regression is also considered to be used since we have chronological data. Third, we might want to segment the data by our interests no matter by company, race, education level or location for in-depth comparison.

2. Data Preprocessing



We dropped rows with NA values in company and level. We also noticed that in gender column, other than female and male and NA, so we dropped rows with unapplicable values in gender, and changed NA values to unknown. Also the categorical variables like Company, title including gender were label encoded to convert it to numerical variable.

Then we dropped columns such as 'timestamp', 'rowNumber', 'cityid', 'dmaid', 'otherdetails', 'tag', 'Race', 'Education', because some of them contain useless information, and Race and education are redundant columns in the dataset. We dropped dmaid because the dataset did not specify what dmaid values are. Other details have lots of NA values and unorganized information, so we dropped it.

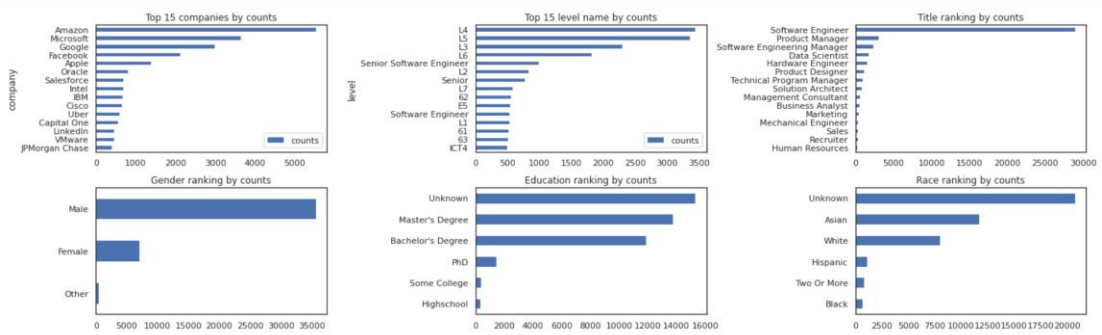
Race and education have couple binary value columns already, so we deleted Race and Education.

We also filtered the dataset so the total annual compensation is within 2,500,000, which is an abnormal range of compensation for regular roles and removed companies that had less than 5 frequency count in the dataset.

For the classification and regression analysis, only the top 50 companies were chosen from 20 countries around the world.

3. Exploratory Data Analysis

Data Overview



Top companies and levels

The majority of the data comes from companies such as Amazon, Microsoft, Google, Facebook, Apple. L4, L5, L3, L6, senior software engineer, are the most common levels of positions in the dataset. Senior software engineer in each company has different level names and salary range: Apple (ICT4, estimated annual compensation $328,000), Amazon (SDE 3 or L6, estimated annual compensation $330,000), Google (L5, estimated annual compensation $357,000), Facebook (E5, estimated annual compensation $395,000), Microsoft (63, estimated annual compensation $223,000). Lx level in commonly used in Google and Amazon, since Amazon has the most sample size in this dataset, we will use Amazon level to compare to other companies. Amazon (L4, $164,000) is corresponding to Apple (ICT2, $164,000), Google (L3, $192,000), Facebook (E3, $182,000), Microsoft (59-60, $158,000- $166,000). Amazon (L5, $ 232,000) is corresponding to Apple (ICT3, $221,000), Google (L4, $271,000), Facebook (E4, $271,000), Microsoft (61-62, $183,000 - $190,000). Below is the comparison table from levels.fyi of top companies and their levels:

Title

The most common title in the dataset is software engineer, outnumbered the rest of the titles combined. The number of software engineers is around 35,000, the second common title is product manager only around 5,000.

Gender

In terms of gender, we only looked at male and female since it is the majority of the gender. There are a lot more male than female in the dataset, male (appr. 35000) outnumbered female (appr. 5000) by around 30,000 people.

Education

When it comes to the education section, a large proportion of people did not enter the information, meaning that they do not wish to answer the question. In the rest of the population, majority of the population have a master degree or a bachelor degree, around 12000-14000. The number of PHD, some college, and high school is significantly less than master degree or bachelor degree.

Race

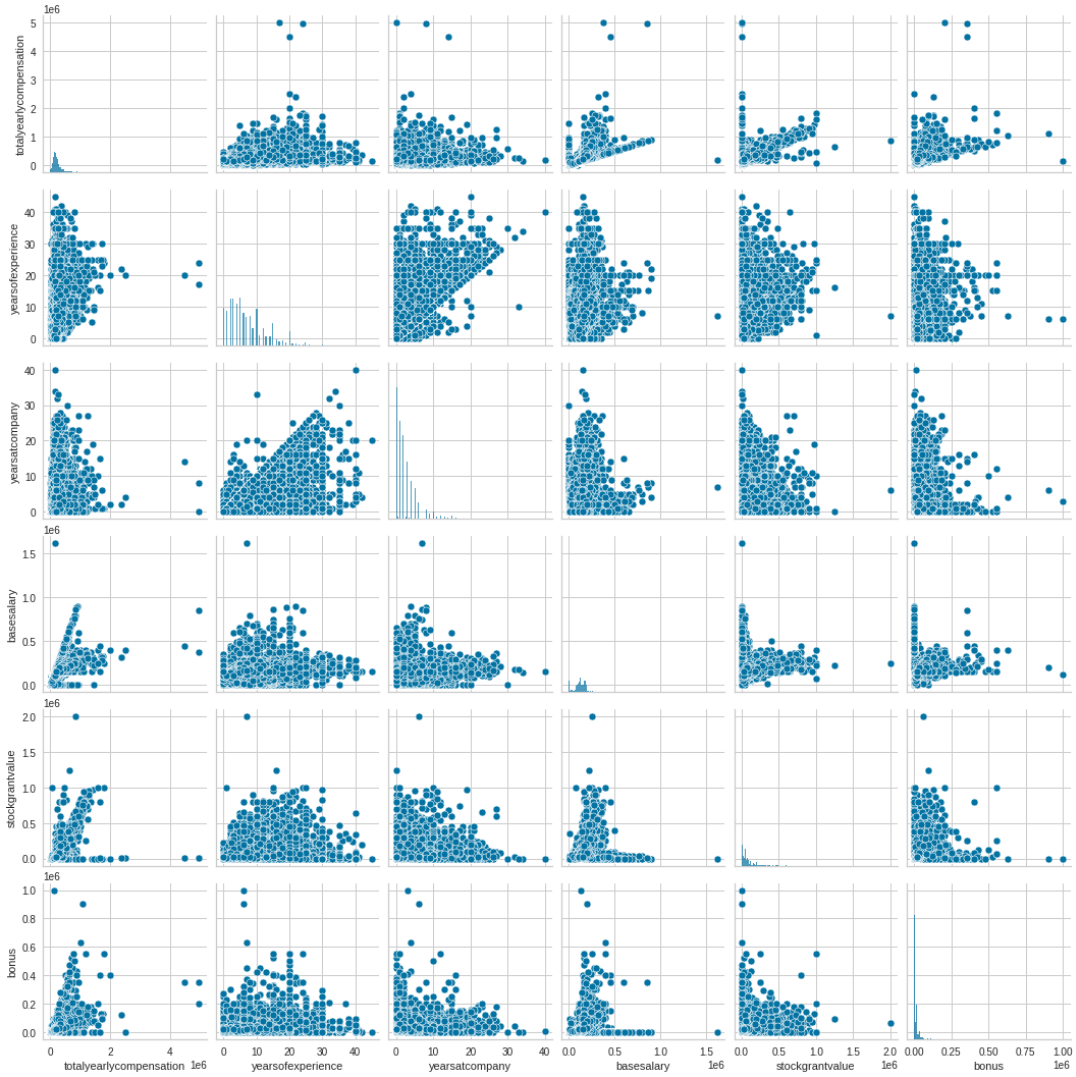
Ignoring the population that did not fill the race section, the most common races in the dataset are Asian and White.

Numerical attributes distribution and correlation

Our most interesting question regarding the dataset was what numerical attributes contributed the most to the total yearly compensation. We selected some of the attributes that we believe have high correlation with total annual compensation such as years of experience, years at the company,

base salary, stock grant value, and bonus. Without a surprise, the total annual compensation is positively correlated with the base salary, the stock grant value, and the bonus, because they are parts of the annual compensation. It is surprise to see that year at company does not really matter for the annual compensation, there were a lot more people with higher annual compensation that joined the company for a brief period of time than people with lower annual compensation that joined the company for a lengthy period. We analyzed this result as the faster way for employees to increase their annual compensation is to jump from one company to another. We also found that the year of experience is a significant factor in annual compensation; but once we have the median years of experience to 6 years, we will almost have our maximum compensation (without considering other factors). Additional years of experience won't increase the annual compensation by a lot, we call it the marginal utility return of the job experience in years.

	totalyearlycompensation	yearsofexperience	yearsatcompany	basesalary	stockgrantvalue	bonus
count	43024.0000	43024.0000	43024.0000	43024.0000	43024.0000	43024.0000
mean	209576.1902	7.0967	2.6760	135951.3729	49507.2817	19044.8195
std	129210.1335	5.8494	3.2482	58764.8924	77878.9218	26301.5168
min	10000.0000	0.0000	0.0000	0.0000	0.0000	0.0000
25%	131000.0000	3.0000	0.0000	107000.0000	0.0000	2000.0000
50%	184000.0000	6.0000	2.0000	139000.0000	24000.0000	14000.0000
75%	255000.0000	10.0000	4.0000	165000.0000	62000.0000	26000.0000
max	2500000.0000	45.0000	40.0000	1620000.0000	2000000.0000	1000000.0000

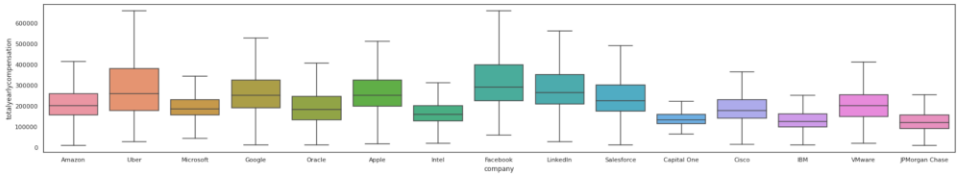


Categorical attributes to total yearly compensation

Company

The top five high pay companies of the total average compensation are Facebook ($339,804) > LinkedIn ($296,441) > Uber ($289,872) > Google ($278,451) > Apple ($274,888).

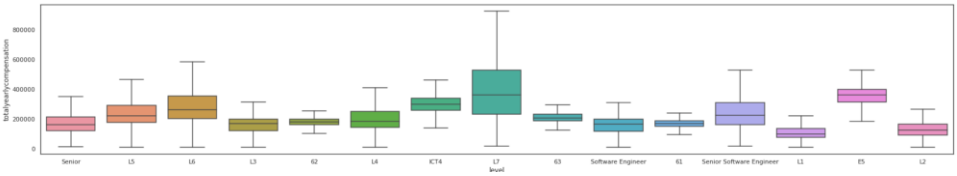
The distribution of total average compensation for Uber and Facebook is the most spread across the companies. Then it comes down to Linkedin, after that it comes down to Google and Apple. This spread represents the difference between the highest salary and the lowest salary, the colored bars are the median 50 percent of total average compensation. It is unexpected that the top five highest pay companies also have the most spread across companies, which means there are lots of flexibilities in annual compensation.



Level

The top five highest pay level of the total average compensation are L7 ($39,2334)> E5 ($35,3368)> ICT4 ($29,9629)> L6 ($28,8667)> Senior Software Engineer ($24,9973).

The distribution of total average compensation for L7 is the most spread across the levels. Then it comes down to L6, after that it comes down to L5 and senior software engineer. The other levels look significantly smaller than these 4 levels mentioned above.

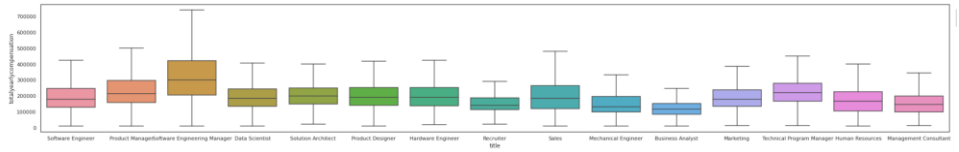


Title

The top five highest pay title of the total average compensation are Software Engineering Manager ($34,4552) > Product Manager ($25,0215) >Technical Program Manager ($23,2659) > Solution Architect ($21,0384) > Product Designer ($20,9489).

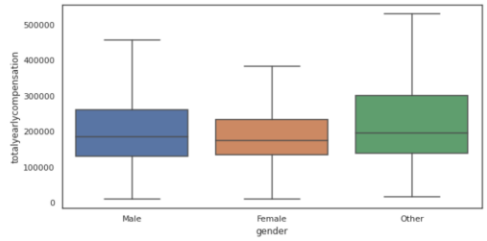
The bottom five lowest pay title of the total average compensation in the dataset are Human Resources ($17,7898) > Management Consultant ($15,7732) > Mechanical Engineer ($15,7069) > Recruiter ($15,3523) > Business Analyst ($12,5344).

The distribution of total average compensation for software engineering manager is the most spread across the title.



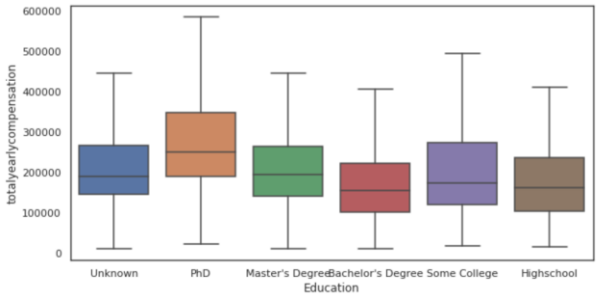
Gender

Male has the highest total average compensation compared to female. Male also have the largest spread between highest compensation and lowest compensation. This might be caused by the large difference between male population and female population.



Education

By looking at the average total compensation: PhD ($286,567) > Master($218505) > Some College($211,293) > Highschool ($183,967) > Bachelor ($177,424) Education earns PhD more, but not for Bachelor degree from the dataset. However, it is also possible that the population with only education of some college and high school must have special skillsets that outstand them in technology companies.



4. Regression and Classification

4.1 Linear regression:

Linear regression is used to model and predict continuous variable. It identifies relation between independent variables (the predictors) and the dependent variable, in this case the total yearly compensation.

$$Y= mx + b$$

It was found that attributes Country, years of experience, gender and Education had positive coefficients and lower P values which reflects its importance in estimating the total yearly compensation. Below is the summary of the linear regression model.

OLS Regression Results

Dep. Variable:	totalyearlycompensation	R-squared:	0.818
Model:	OLS	Adj. R-squared:	0.818
Method:	Least Squares	F-statistic:	7565.
Date:	Thu, 02 Dec 2021	Prob (F-statistic):	0.00
Time:	08:32:40	Log-Likelihood:	-3.5327e+05
No. Observations:	28564	AIC:	7.066e+05
Df Residuals:	28546	BIC:	7.067e+05
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.051e+04	2654.874	-3.958	0.000	-1.57e+04	-5303.318
company	-27.9632	27.651	-1.011	0.312	-82.161	26.234
Country	3051.3424	137.277	22.228	0.000	2782.273	3320.412
yearsofexperience	2731.3924	72.923	37.456	0.000	2588.460	2874.325
yearsatcompany	-211.1961	117.969	-1.790	0.073	-442.421	20.029
basesalary	0.8097	0.007	110.657	0.000	0.795	0.824
stockgrantvalue	0.9541	0.005	205.952	0.000	0.945	0.963
gender	2572.6541	875.527	2.938	0.003	856.580	4288.728
Masters_Degree	-1.082e+04	995.590	-10.872	0.000	-1.28e+04	-8873.122
Bachelors_Degree	-1.383e+04	1246.979	-11.091	0.000	-1.63e+04	-1.14e+04
Doctorate_Degree	4290.7317	1876.718	2.286	0.022	612.276	7969.187
Highschool	-1.218e+04	4535.663	-2.685	0.007	-2.11e+04	-3287.511
Some_College	-1.857e+04	4360.717	-4.258	0.000	-2.71e+04	-1e+04
Race_Asian	-3035.7041	1069.310	-2.839	0.005	-5131.602	-939.807
Race_White	-1772.5127	1251.663	-1.416	0.157	-4225.831	680.805
Race_Two_Or_More	2090.1977	2781.650	0.751	0.452	-3361.968	7542.363
Race_Black	1893.7154	2852.343	0.664	0.507	-3697.012	7484.443
Race_Hispanic	-4029.3093	2333.726	-1.727	0.084	-8603.523	544.904

4.2 Decision Tree, Random Forest and XG Boost Model:

Prediction models was built to accurately predict the total yearly compensation and the dataset was split into training and validation set using sklearn package to avoid over fitting. In order to measure the model's performance, Mean Absolute error metric was used. Below are the results from Decision Tree, Random Forest and XG Boost models. As expected, XG Boost model performed better than other models and reduced the mean absolute error.

Decision Tree Regressor:

Mean Absolute Error = 61740.57

```
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
train_X, val_X, train_y, val_y = train_test_split(x, y,random_state = 0)
stem_model=DecisionTreeRegressor(max_leaf_nodes=100,random_state=1)
stem_model.fit(train_X,train_y)
y_pred=stem_model.predict(val_X)
print(mean_absolute_error(y_pred,val_y))
```

```
61740.57468565379
```

Random Forest Regressor:

Mean Absolute Error=53192.67

```
from sklearn.ensemble import RandomForestRegressor

ran_model = RandomForestRegressor(random_state=1)
ran_model.fit(train_X,train_y)
y_pred=ran_model.predict(val_X)
print(mean_absolute_error(y_pred,val_y))
```

```
53192.67925724376
```

XG Boost:

Mean Absolute Error = 50634.30

```
from xgboost import XGBRegressor

xg_model = XGBRegressor(n_estimators=700,learning_rate=0.08, n_jobs=4)
xg_model.fit(train_X, train_y)
y_pred=xg_model.predict(val_X)
print(mean_absolute_error(y_pred,val_y))
```

Regression - Mean Absolute Error			
Decision Tree		**Random Forest**	**XGBoost**
Max_leaf node = 100	61740.57	53192.67	50634.3
Max_leaf node = 500	56824.77		

4.3 Classification through bins:

In order to get better predictions, The target variable total yearly compensation was classified into three bins assigning labels to each bin from 1 to 3.

Salary range:

9999 to 75000 $ - 1

75000 to 150000 – 2

150000 to 2500000 – 3

Decision Tree and Random Forest models were built without oversampling to classify the target variable and the confusion matrix was used to measure the accuracy of the model. Though the overall accuracy was in the range of 83 to 85%, the stratified accuracy for individual classes were low due to the imbalance in target class. Oversampling the target class using random sampler method resulted in better overall and stratified accuracy.

Features
=['company','title','Country','gender','yearsofexperience','yearsatcompany','Masters_Degree',

'Bachelors_Degree', 'Doctorate_Degree', 'Highschool', 'Some_College',

'Race_Asian', 'Race_White', 'Race_Two_Or_More', 'Race_Black',

'Race_Hispanic']

Results:

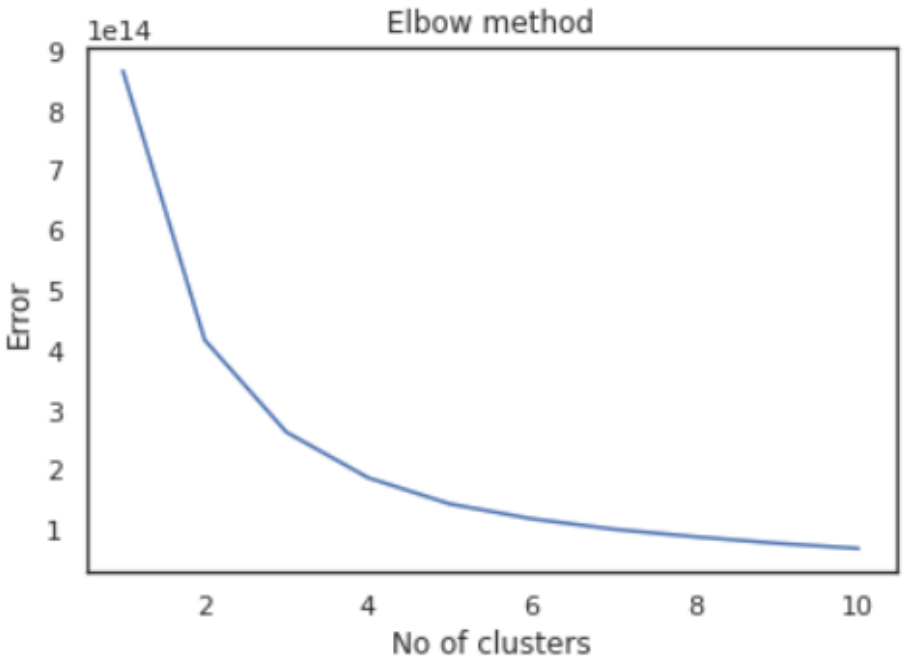
Classification - Accuracy		
	Decision Tree	**Random Forest**

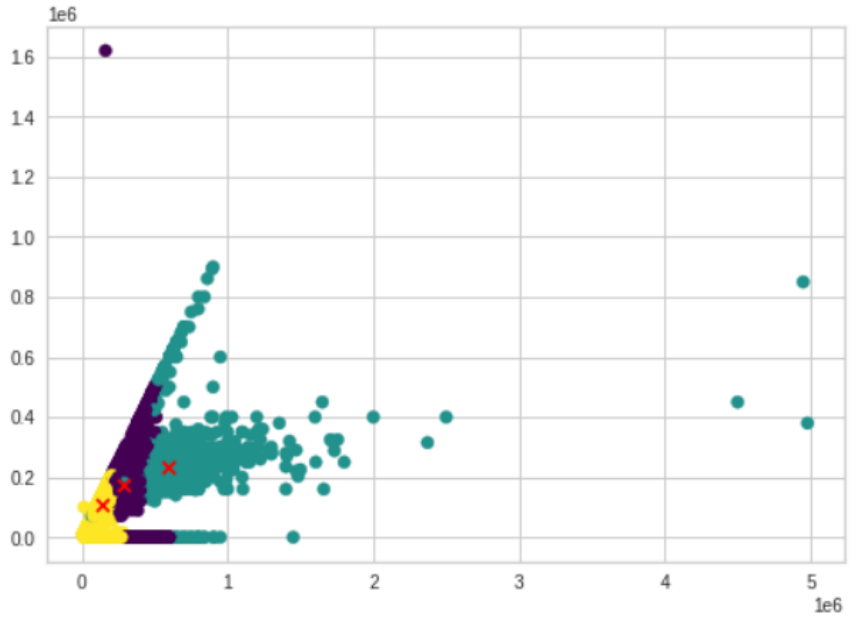
Without Oversampling	**83.19 %**	**85.03 %**
With Oversampling	**93 %**	**93.43%**

5. Clustering Analysis

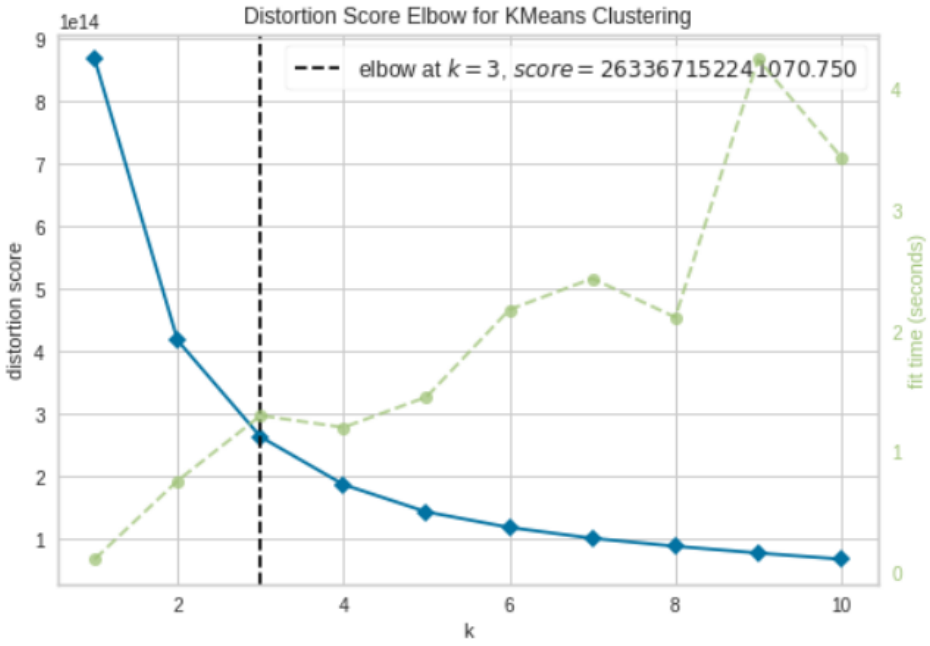
Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data analysis, and a common technique for statistical data analysis, used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including parameters such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

We have used K-Means Clustering in our Analysis. K-Means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances.



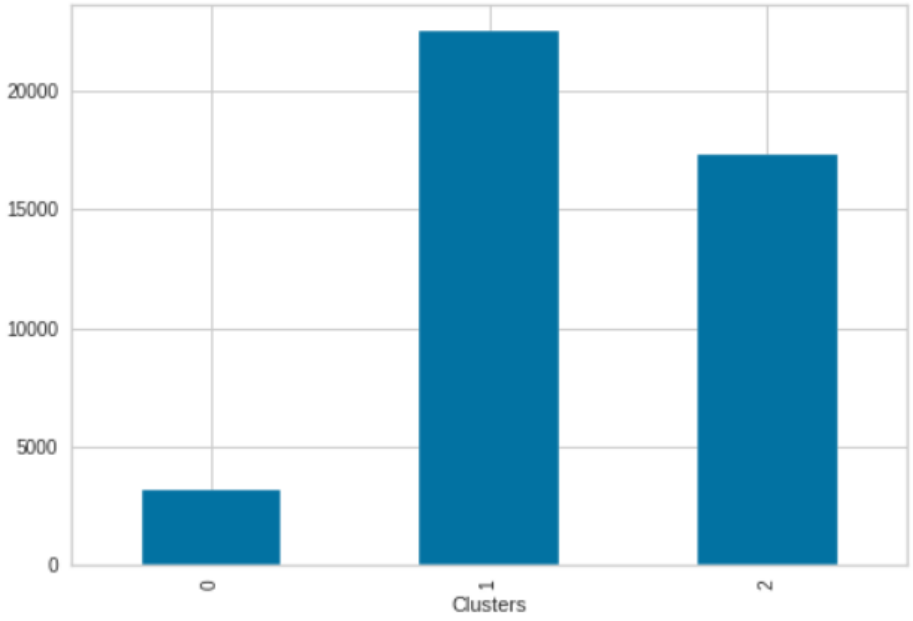


In this Graph, we can see how widespread the Cluster values are and the X marks depict the cluster centers.



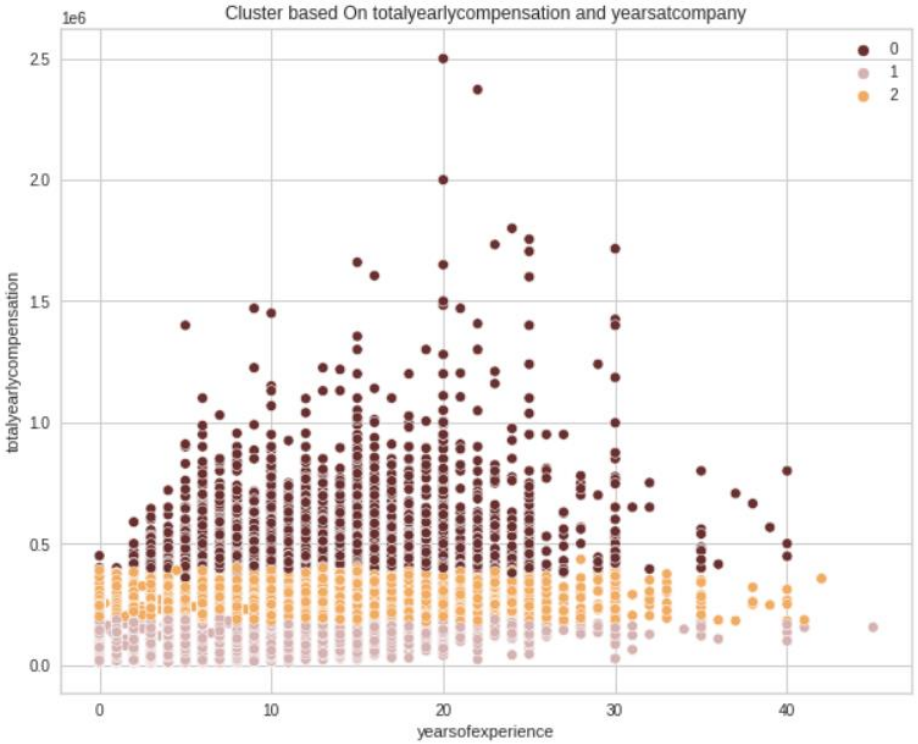
There are in total 3 clusters in our Analysis. The cluster centers are

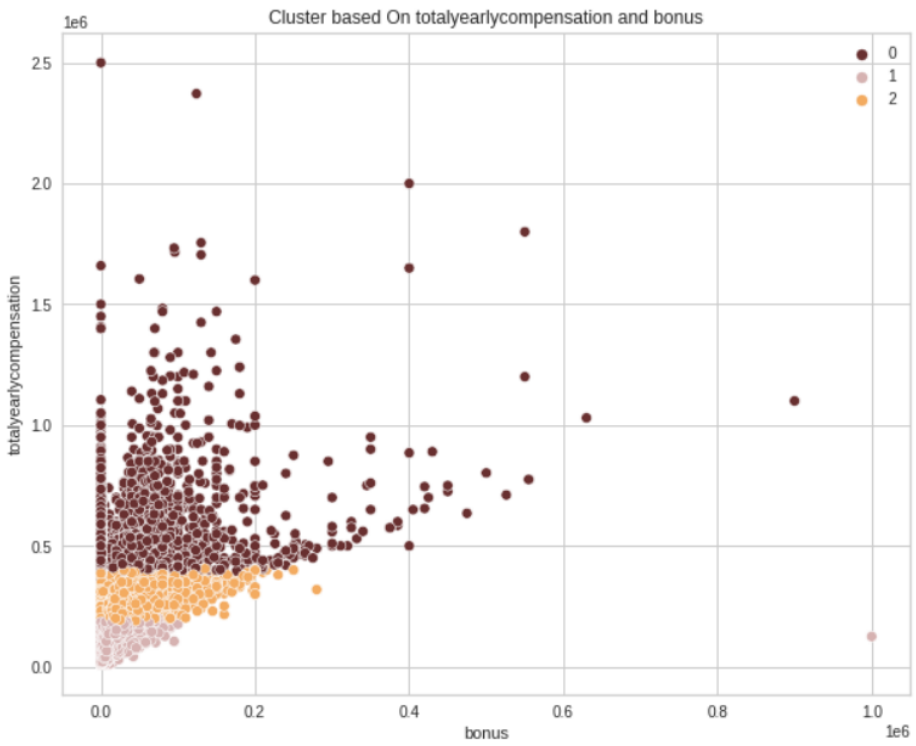
```
array([[2.56086980e+05, 8.66784036e+00, 3.06466550e+00, 1.64881723e+05],
       [1.25424267e+05, 4.97448931e+00, 2.18659583e+00, 9.87547199e+04],
       [5.30234732e+05, 1.30135719e+01, 3.89651450e+00, 2.32072205e+05]])
```



Weightage Cluster wise

- Count of values present in each cluster
- Cluster 0 has the Maximum number of values (26000 approx.) among the three.
- Cluster 2 has the minimum number of Values which is less than 5000 of all the clusters.





- Cluster 0 (low income): 26979 （62.7%) people with average total compensation of $138,727, which is below the average total compensation of $184,000
- Cluster 1 (high income): 2084 (4.8%) people with average total compensation of $596,116, which is below the average total compensation of $184,00
- Cluster 2 (middle income): 13964 (32.4%) people with average total compensation of $289,758, which is above the average total compensation of $184,00

Clusters	counts	average total yearly compensation	percentage
0	3159	533793.2890	7.3424
1	22514	126491.2724	52.3289
2	17351	258355.6279	40.3287

6. Inference and Key Takeaways:

This project was useful in analyzing the job market and salaries earned by STEM professionals. Prediction models like Decision tree, Random Forest regressor and XGBoost were also built to understand key attributes and predict the total yearly compensation.
Below are our key takeaways and learnings from this analysis.

- Average yearly compensation of Male is greater than Female.
- Employees with a PHD or Master's degree have higher salary compensation.
- Software engineer and Product manager roles are the highest paying jobs.
- Best classification model - Random Forest with oversampling - Accuracy - 93.43% Significant attributes based on Regression- Years of Experience, Education, gender, Country