**University of Connecticut**



**OPIM 5512**

**Data Science using Python**

**Data Science and STEM Salaries**

Provide actionable market intelligence and competitive benchmarking data to Data science students to select the best paying career

**Group 6**

Alpa Puri

Lakshmi Ramya

Rupam Tyagi

Saqib

# Introduction

The idea behind this project was inspired by the most common regular conversations between graduate students in our program who are looking for their golden opportunity to establish a shinning career after doing the MSBAPM program. With multiple opportunities in the job market, companies pitching in the job fairs and recruiters bombarding our handshake messenger it is challenging to focus on what are we looking to achieve in career life. Industry and job titles can only differ so much and that puts the whole game inclined towards the Total yearly compensation. There is a great need for the students to know what are the drivers that result to a high paying data professional job.

We found a few datasets that were apt for our project on data professional salaries. We further explored those data set to find that Levels Fyi salary data was apt for our project scope with more complete and verified data. From here we started exploring the data in excel and tableau to discover different correlation between variables. There was additional variable for total yearly compensation that gave a bifurcation of Base salary, stock grant and bonus earned by different professionals with diverse experience, race and education. We initially wanted to merge this dataset with other data set that has additional information like certificates and computer related educational background but there was a lack of enough data points to converge with our selected dataset.

The markets have several hypotheses on the relationship between job titles and their average earned salaries. Going from there, this project aims to determine if those assumptions are true and what is the true trajectory of the data professionals in terms of the target variable – total yearly compensation. Our visualization also shows what are the driving factors that lead to a higher compensation and what factors contribute to the gender disparities in this industry.

This is where our project can be resourceful to the students looking ahead to join the corporate world as a data professional.

Contents

[Introduction 2](#_gjdgxs)

[Literature 4](#_30j0zll)

[Data 5](#_1fob9te)

[Business Questions 5](#_3znysh7)

[Data Cleaning 6](#_2et92p0)

[Data Exploration 8](#_tyjcwt)

[Data Modeling 17](#_3dy6vkm)

[Models 17](#_1t3h5sf)

[Conclusion 19](#_4d34og8)

[References 20](#_2s8eyo1)

# Literature

**Previous work in this area:**

We found many articles and writeups on this topic of data science salary and data profession demands. For a more precise data study on data professional salaries, we found there is a projectSHERPA which was also a dataset of jobs with the scope of using techniques learned in “Data science with Python: Machine learning” to predict the NYC data science base salaries. But this study was done using the jobs posted on different portals. Predicting salaries from job posts is inconsistent and not reliable. Also, a lot of data was not usable so only a small data sets was utilized. In comparison to our selected data this project lacks the verified salary feature of Levels fyi site. And as we know many times the companies and portals keep old job postings to keep attracting the pool of candidates and hence it is just a marketing gimmick.

**Levels\_fyi data set:**

Our data set is from June 2017 to August 2019 with more than 1000 companies represented in our sample. There are 15 unique data professional job titles along with their verified Total yearly compensation with details of base salary, stock grant and bonus earned. Even location wise our data is rich and vast covering major data of all United states city, state & also some data from other countries.

# Data

For our project, we have used the Data Science and STEM Salaries data from Kaggle website, which has 62,000+ STEM salaries scraped off Level.fyi site. The data set contains useful information such as compensation (base salary, bonus, stock grants), race, education level, company, years of experience and more. Here is a snapshot of the data:

A screenshot of a computer

Description automatically generated with medium confidence

Table

Description automatically generated

# Business Questions

Student are looking for deeper insights about the driving factors that lead to better compensating career in data science field and would like to know the future growth of the field and financial returns. Here are some of the questions we have attempted to answer with our Project study. The data sample from Level.fyi covering more than 1000 different companies can be utilized to create predictive modeling. This will help our students to improve their chances of landing their dream jobs.

# Data Cleaning

Dimension reduction (dropping of insignificant columns):

* There were few insignificant variables that would not aid to our prediction models and hence were removed during data cleaning.

Text

Description automatically generated

Delete rows with base salary is zero

* There were around 2000 rows with missing base value in base salary and hence were removed from the 62000 row data set.

A picture containing chart

Description automatically generated

Replacing blank fields in Race, Education and Gender with N/A

* There were also some variables with missing values and were imputed with N/A for better visualization.

Text

Description automatically generated

Rounding off years of experience and years at company to closest natural number

* The two variables namely years of experience and years at company were rounded to the nearest integer to standardize and group easily.

Text

Description automatically generated

Text

Description automatically generated

Splitting the location data to give both the area and the country

* The location variable was sliced to get the location on the tableau mapping visualization by city state and country. Country was missing for all US city and states and hence was imputed separately.

Graphical user interface, text, application

Description automatically generated

Here was also an erroneous entry in gender which was imputed as N/A.

# Data Exploration

In the data exploration phase, we used a variety of methods to explore the data set. We used correlation analysis to find relationships within the data, linear regression to establish significant variables against a target variable. Lastly, we explored charts and graphs which allowed us to identify trends.

Top 10 job profiles in STEM

Graphical user interface, application

Description automatically generated

The most common title is Software Engineer

Total yearly compensation of job roles

Chart, box and whisker chart

Description automatically generated

From the graph that it can be analyzed that, Total yearly compensation is highest for Software Engineering manager

Total yearly compensation of Software Engineer, Data Scientists, Solution Architects

**Chart, bar chart

Description automatically generated**

Data Scientists with a PhD have average income larger than Data Scientists with only a Master's Degree, whereas the opposite holds for Solution Architects in Master's and Bachelor's.

Total yearly compensation in top 8 companies

**Chart, bar chart

Description automatically generated**

Facebook pays highest compensation in STEM related jobs

Distribution of Top 8 companies

**Chart, pie chart

Description automatically generated**

The biggest amount of STEM jobs are in Amazon

STEM jobs across US cities

**Chart, bar chart

Description automatically generated**

Most common location is Seattle

STEM Jobs in US, UK, India, Canada and Germany

**Chart, line chart

Description automatically generated**

USA seems to be the best location for this kind of jobs.

Total yearly compensation across all education levels among genders

**Chart, bar chart

Description automatically generated**

We can observe that the total\_yearly\_compensation for PhD are the highest in all metrics.

Gender disparity in the industry

**Chart, bar chart

Description automatically generated**

From the data above, gender doesn't have an impact on average salary in STEM. But from the other side - there's more men than women working in this sector.

Gender disparity in STEM jobs across 4 countries

**Chart, bar chart

Description automatically generated**

STEM jobs spread across US

**Map

Description automatically generated**

Heat map of STEM job profiles with their average base salary those who have bachelors or higher-level education degrees

**Chart, treemap chart

Description automatically generated**

Data Exploration: Correlation analysis

We performed correlation analysis on all variables to see how the relationship of each variable can affect another. From the analysis we can see that Total yearly compensation, base salary and years of experience sales have the highest positive correlation.

Correlation with Total Yearly Compensation

**Table

Description automatically generated with medium confidence**

# Data Modeling

We divide the dataset into two parts: training and testing dataset. Training dataset is 60% while the testing dataset is 40%.

Text

Description automatically generated with medium confidence

## Models

We ran 3 different models in evaluating the relationship between the target variable and the other features.

1. **Data Modeling: Linear Regression**

Graphical user interface, text, application, email

Description automatically generated

In the linear regression model, we see that the R-square score is 0.6063 and the estimates for ‘basesalary’ and ‘yearsatcompany’ are 1.85 and 973.121.

1. **Data Modeling: Decision Tree**

Text

Description automatically generated

The decision tree model works well with an R-Square score of 0.463 but it is not as good as the Linear Regression model.

1. **Data Modeling: KNN**

Graphical user interface, text, application

Description automatically generated



The KNN model has an R-Square value of 0.4805. This is better than the decision tree however it is not as good as the Linear Regression model.

**Model Comparison**

We see that the best model is Linear Regression with an R-Square value of 0.6063.

# Conclusion

In this project, we did analysis to identify the most prominent factor that affects the Total Yearly compensation with the help of predictive modeling techniques learnt in class. In this process we did learn the practical application of python and machine learning tools and programs.

When total experience in years and years at a company is used as a regressor, years at company fetches negative coefficient, from this observation we would suggest the candidates look for new opportunities in the market rather than trying to find them in the current organization to find higher compensation. And we can also support this argument with one of our team member’s real-world experiences in Insurance industry where, she has observed that external candidates are likely to get higher pays than any internally hired data professional.

As a female candidate for better opportunities, it is more probable to secure a job in the US than in the rest of the world as, according to an analysis from tableau US has the highest number of STEM job females compared to the rest of the world.

To aim for a job in leading firms, it is easier to get into amazon as the number of opportunities are more but for maximizing yearly compensation Facebook is the best option.

# References

<https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries>

<https://nycdatascience.com/blog/student-works/job-salary-prediction-with-python/>

<https://www.projectpro.io/article/data-scientist-salary-the-ultimate-guide-for-2021/218>