Data Science and STEM Salaries

Karen Xu, Ryan Huang, Ruiming Li, and Michelle Lo

Northeastern University: Khoury College of Computer Sciences

ABSTRACT

As STEM majors we would like to know our future approximant salaries for after we graduate. To do this we would like to investigate the public dataset "Data Science and STEM Salaries." [4] This data set, as its name suggests, contains information regarding the different salaries of people working in the STEM fields. The data contains information such as company names, role title, years of employee experience, and their total yearly compensation. We will be using the support vector machine, random forest, and K-nearest neighbor algorithms to explore our data.

INTRODUCTION

Problem:

The primary goal of the project is to find common factors related to career levels and compensation packages across multiple tech-oriented companies. As levels.fyi catalogs thousands of employee companies, levels, salaries, locations, and many other attributes relevant to total compensation. This information is invaluable to job-seeking engineers and project managers as wanting to maximize compensation for a given amount of experience is something all workers have felt at some point in time [2]. The desire to earn more is valid for all job-seeking workers. Especially in a time where wages are generally agreed upon as having not increased relative to inflation, this information has never been more important [3]. The ability to share salaries and how location and years of experience can optimize it is critical to short-term and long-term career success.

Motivation:

In the dataset, we can see that there are product managers, software engineers, data scientists, and more of whom are all close to our STEM-related field jobs. In the dataset, the job positions are all from well-known companies such as Microsoft, eBay, Amazon, and Apple. We were motivated to choose this topic because in this dataset we can see our potential salary after we graduated or if there are any jobs with high salaries that we can work for in the remainder of our years.

Goals and Objectives:

Our goal for this project is to be able to discover the correlation between career levels and compensation packages across technology-oriented companies. After analyzing the dataset, we hope to be able to explain which companies offer better stock options, which companies promote individuals to high-level positions faster, and which companies offer better salary options. Additionally, we hope to be able to discover the trends between low entry-level paid jobs and experience along with which positions' salaries increase significantly throughout the different experience levels.

As for our individual team objectives, we hope to be able to effectively communicate our ideas with one another using the available resources such as Microsoft team channels, email, and text. Despite the challenges posed by remote learning, we hope that throughout the different milestones, each team member will be able to hone leadership skills, make decisive calls, and guide the team towards a successful final project.

RELATED WORK

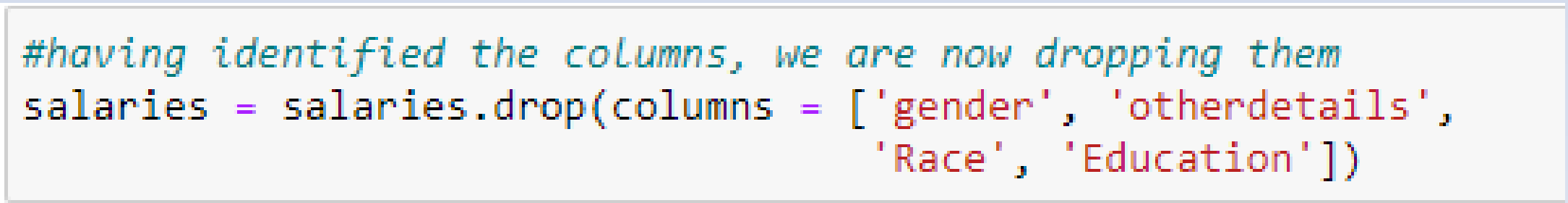
The New York Times Data Scientist Salary[6] studies Data Scientist's salaries, which is part of what we are studying. In their studies, they concluded with different graphs that represent compensations by department, average equity, Compensations by gender, and compensations by ethnicity. Lastly, something that is different than us is they ask their data scientist to rate their compensation which was not included with in our data.

Fastest-Growing Tech Occupations Include Data Scientists Engineers[5] in this paper its es how each tech occupation has the biggest increases in salary between 2018 and 2019. This paper was analyzed by 12,837 technologists.

2021 Salary Guide to Careers in Data Science[1] is also very similar to what we are studying it analyzes which states have the highest pay alongside the highest-paid company. It also studies how the level of education and year of experience would affect your pay. Lastly it studies which city has the highest paid.

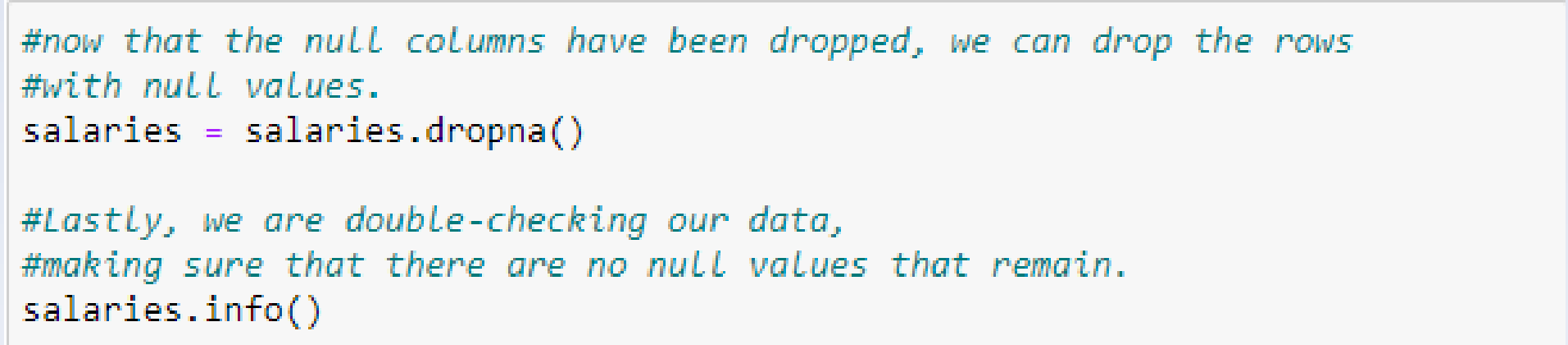
METHODOLOGY

Initially, we began by analyzing the "Data Science and STEM Salaries." As we examined the data, we realized that the data required cleaning. Thus, we began by examining null values via heatmap and dropping the columns that were mostly nulls. We then dropped the columns that were mostly null values.



```
#having identified the columns, we are now dropping them
salaries = salaries.drop(columns = ['gender', 'otherdetails',
                                    'Race', 'Education'])
```

After the null values were dropped, we then examined the rows that contained null values and dropped those as well.



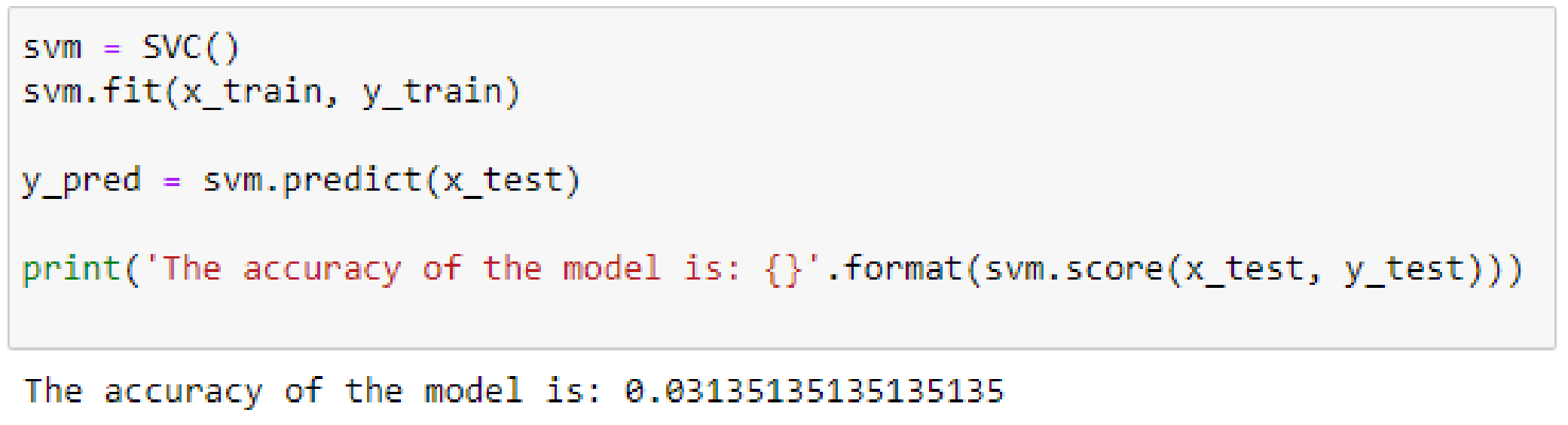
```
#now that the null columns have been dropped, we can drop the rows
#with null values.
salaries = salaries.dropna()

#Lastly, we are double-checking our data,
#making sure that there are no null values that remain.
salaries.info()
```

Finally, now that our data was cleaned and ready to use, we exported it as a new csv. When running algorithms, we determined that our data (60,000+ entries) was too large for some of the algorithms to reasonable run. Therefore, we use pandas.DataFrame.sample() to take a random sample of the entries, thus shrinking our dataset from 60 thousand entries to about two thousand.

Three algorithms we used to analyze our data are:

Support Vector Machine (SVM):

After encoding our data, we passed it into the support vector machine.

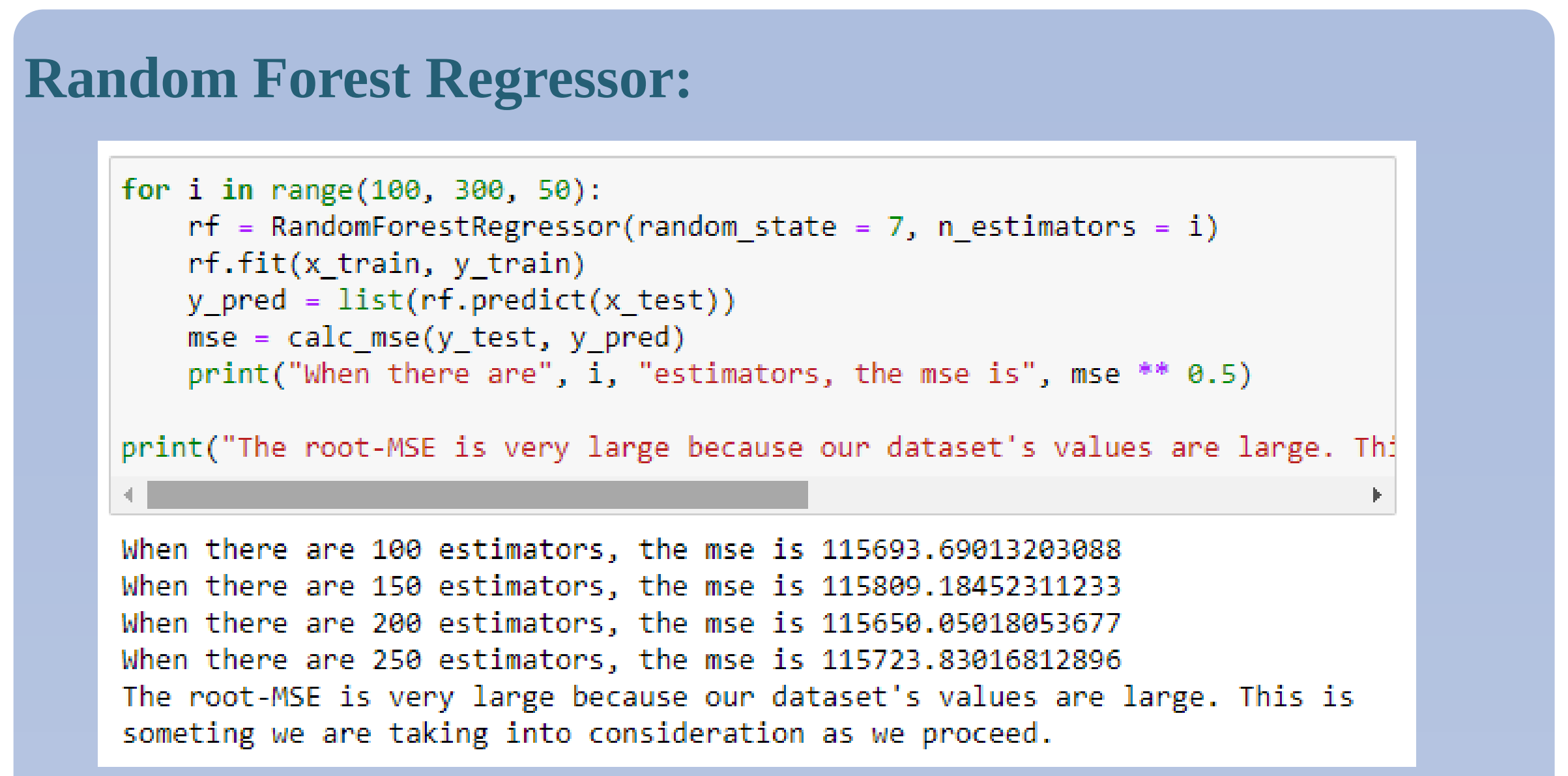


```
svm = SVC()
svm.fit(x_train, y_train)

y_pred = svm.predict(x_test)

print('The accuracy of the model is: {}'.format(svm.score(x_test, y_test)))
```

```
The accuracy of the model is: 0.03135135135135135
```

However, since the accuracy of this model is poor, we determined that it was not a good predictor.

Random Forest Regressor:

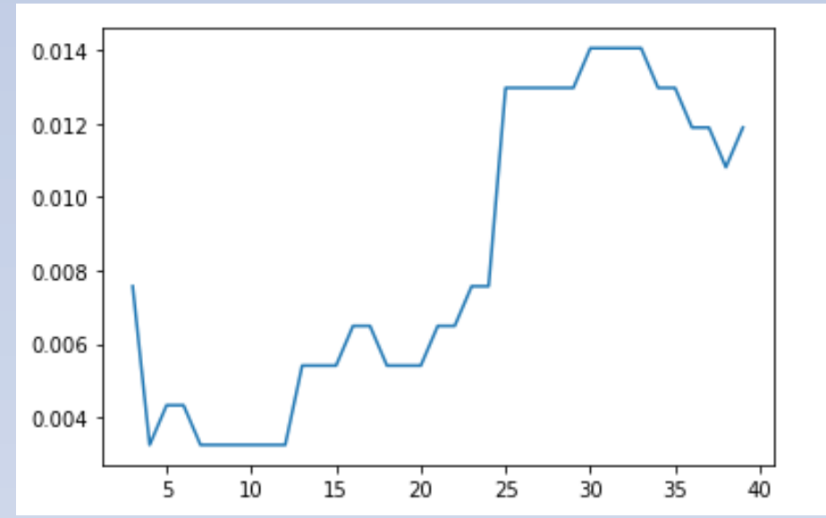


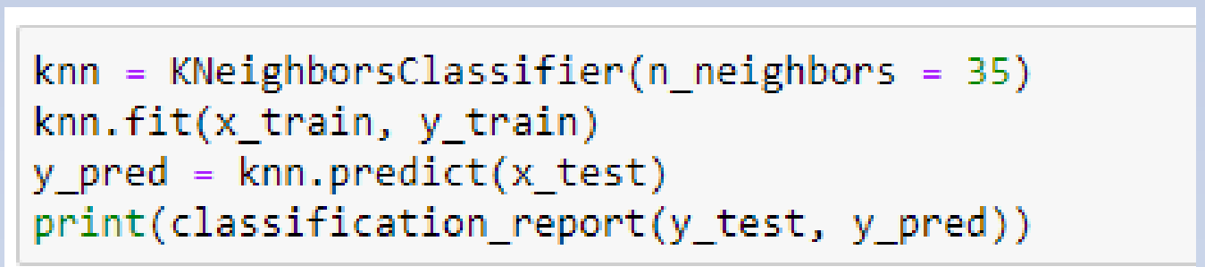
```
for i in range(100, 300, 50):
    rf = RandomForestRegressor(random_state = 7, n_estimators = i)
    rf.fit(x_train, y_train)
    y_pred = list(rf.predict(x_test))
    mse = calc_mse(y_test, y_pred)
    print("When there are", i, "estimators, the mse is", mse ** 0.5)

print("The root-MSE is very large because our dataset's values are large. Thi
```

```
When there are 100 estimators, the mse is 115693.69013203088
When there are 150 estimators, the mse is 115809.18452311233
When there are 200 estimators, the mse is 115650.05018053677
When there are 250 estimators, the mse is 115723.83016812896
The root-MSE is very large because our dataset's values are large. This is
someting we are taking into consideration as we proceed.
```

The random forest regressor was able to predict the model with a root-mean squared error of about 115 thousand. Since the error is also very large, we determined that this is not a good model.

K-nearest Neighbor:





```
knn = KNeighborsClassifier(n_neighbors = 35)
knn.fit(x_train, y_train)
y_pred = knn.predict(x_test)
print(classification_report(y_test, y_pred))
```

Finally, we also ran a k-nearest-neighbor clustering model. We used the graph on the left to determine the best value of k. We then used that k-value to then run the KNeighborsClassifier algorithm using our value of k = 35. However, as the graph shows, our maximum accuracy is only at 0.014.

RESULTS AND EVALUATION

After running our algorithms, we determined that none of the algorithms were satisfactorily accurate. Since the errors and the accuracy rates were out of an acceptable range, we do not believe that this data has an extremely strong correlation. A few factors may contribute to this randomness.

First, we believe that the diverseness of tech jobs is likely a contributing factor. Since this dataset is global, we are not accounting for location in a lot of our models. We think one of the main factors of salary may be the location's cost of living. A higher cost of living may correspond to a higher salary, which is not accounted for in our models.

Second, we believe that, by not accounting for the other categories, such as stock grants, bonuses, or other benefits, we are being too specific in our examination. Some jobs may have only paid in the form of stock grants, rather than an annual compensation, so those entries would have created problems in our dataset.

Lastly, one cause of poor modelling is because of our sample. If we had taken the original, cleaned dataset of over sixty thousand entries, perhaps our model would have been better at predicting. However, given our computer limitations, we were forced to take a sample form the original data.

Thus, given these factors, we believe that the lack of correlation in our data is expected. However, we would have considered examining our data using some other methods, such as using other algorithms that may yield a better outcome.

THE IMPACTS

The information obtained by the dataset has the most relevance to university STEM students or early career employees. could directly influence when and where students decide to focus their career path as salary/compensation are significantly relevant factors to seeking work. As certain companies have specific trends to promotion and base salary levels for a certain number of years of experience, having this information ahead of time better informs workers as to how they can maximize their pay and whether they are being treated unfairly. While our results do not reveal much surrounding STEM related salaries/compensation, the dataset contains valuable information for any job seeking workers.

CONCLUSIONS

For our project, we analyzed the Kaggle dataset "Data Science and STEM Salaries", with the intention of providing young people with accurate information to aid them in deciding their career paths. We started the analysis by examining our data, cleaning it, and then running three different algorithms on it: SVM, random forest regression, and k-nearest neighbors. However, all three of our algorithms found little correlation in the dataset. A lack of correlation could be many factors, such as location, which can influence salary. We also did not count other monetary gains, such as stock grants or bonuses, which may affect our results. Lastly, if we had used our original, larger dataset instead of taking a sample (due to computer and runtime limitations), perhaps our models would have stronger correlations.

REFERENCES

[1] "2021 Salary Guide to Careers in Data Science." *DiscoverDataScience.org*, 19 Apr. 2022, https://www.discoverdatascience.org/a-salary-guide-to-careers-in-data-science/#:~:text=The%20median%20salary%20for%20all,making%20well%20into%20six%20figures.

[2] Casselman, Ben. "Only 17% of Workers Say Their Pay Has Kept Pace with Inflation." *The New York Times*, The New York Times, 4 Jan. 2022, https://www.nytimes.com/2022/01/04/business/economy/worker-pay-inflation.html.

[3] DeSilver, Drew. "For Most Americans, Real Wages Have Barely Budged for Decades." *Pew Research Center*, Pew Research Center, 30 May 2020, https://www.pewresearch.org/fact-tank/2018/08/07/for-most-us-workers-real-wages-have-barely-budged-for-decades/.

[4] Ogozaly, Jack. "Data Science and STEM Salaries." *Kaggle*, 10 Oct. 2021, https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries.

[5] Nick KolakowskiFebruary 13, 202013 min read. "Fastest-Growing Tech Occupations Include Data Scientists, Engineers." *Dice Insights*, 4 May 2021, https://insights.dice.com/2020/02/13/fastest-growing-tech-occupations-data-scientist-engineer/.

[6] *The New York Times Data Scientist Salary | Comparably*. https://www.comparably.com/companies/the-new-york-times/salaries/data-scientist.