Analysis of Skills and Salaries of Careers in Data Science Across the U.S.

Sydney Camenzuli
Department of Computer Science
College of Charleston
Charleston, SC
camenzulisg@g.cofc.edu

Caroline Goodman
Department of Computer Science
College of Charleston
Charleston, SC
goodmancr@g.cofc.edu

Abstract—**The purpose of this project is to determine if particular attributes may lead to a higher data scientist salary. The data analyzed in this research was extracted from Kaggle. It includes around 1,000 different data scientist careers that provide job descriptions, which sectors they work in, city of the job, skills, level of education, etc. The data set is large enough to support this research on its own. We will utilize Python programming to establish if a combination of these attributes, or any one attribute, can contribute to the calculation of a salary. We found that sector, location, and level of education contribute to higher paying jobs. Additionally, less frequent skills such as Mongo and Flink were associated with higher average salaries. There are greater implications that accompany this research. Most students in our data mining course will soon be graduating with majors and minors in Data Science degrees. By extracting information such as locations, skills, and descriptions of the various careers, both our team and our classmates will be offered vital insight to their future specialties.**

I. Introduction

Data science is an interdisciplinary field that incorporates an even mix of programming skills, statistics, and machine learning. This field offers useful information about various sectors extracted from large amounts of complex data. Unlike data engineers and data analysts, data scientists have very comprehensive jobs. Data scientists are responsible for taking on the entire cycle of a data science project. This cycle begins at the conception of the idea, followed by data acquisition, optimization of modeling, and clear communication of results to investors and/or stakeholders.

Businesses can utilize data science to determine which particular demographics they should market to. Doctors can use data analytics to see patterns of health across the country, which is especially useful during global pandemics such as COVID-19. Industries of technology use analytics to strengthen or expedite the running times of the programs they implement. With a rising demand for data scientists, the job interviews are becoming increasingly competitive. Each company is not only looking at your academics or background, but further your skill-set of programming languages and a greater understanding of implementing statistics and machine learning.

Due to the great responsibilities that come with data science careers, it is necessary to have a widely varied skill-set and a strong background in mathematics, particularly statistics. These responsibilities lead data scientists to some of the highest paying salaries across the country, and even across the globe. With salaries ranging anywhere between $90,000 to $200,000, the intensity of interviews and the work of obtaining a sufficient skill-set pays off in the end.

Interviews for data scientists are intense and extensive; therefore, we were interested in determining which skills would place an applicant at an advantage, and ultimately have a higher average salary.

Past research provides greater understanding of our current and future work in this area. O'Reilly Media created a report in 2017 detailing where data scientists make the highest salaries, which tools were most commonly used on the job, which tools contribute to salary the most, how gender can affect salaries, and more [1]. They created advanced figures to accurately visualize these results. These findings prompted us to perform similar actions on our data set and provided a solid basis for what our research would entail.

The data science field involves specific knowledge regarding programming languages that are used on a daily basis in the typical job setting. Our data set provides not only the more common languages as skill sets, but also the less frequently used languages. This provides the opportunity for us to see if having a more unique skill set can lead to a higher salary. We must also take into consideration the contrasting costs of living in different areas of the country, which potentially plays a role in the distribution of the salaries.

II. Data Collection

We received our data set from Kaggle with around 1,000 different data scientist careers [2]. The data set included 42 columns, such as age, job location, sector, size of the company, level of education, various programming languages, and more. For our research, we extracted the columns we believed to lead to impactful outcomes. It was unnecessary to incorporate a separate data set, as the data included in ours was robust in providing sufficient information on each job listing.

```
Rating :  3.8
Company Name :  Tecolote Research
3.8
Location :  Albuquerque, NM
Headquarters :  Goleta, CA
Size :  501 - 1000
Founded :  1973
Type of ownership :  Company - Private
Industry :  Aerospace & Defense
Sector :  Aerospace & Defense
Revenue :  $50 to $100 million (USD)
Competitors :  -1
Hourly :  0
Employer provided :  0
Lower Salary :  53
Upper Salary :  91
Avg Salary(K) :  72.0
company_txt :  Tecolote Research
Job Location :  NM
Age :  48
Python :  1
spark :  0
aws :  0
excel :  1
sql :  0
sas :  1
keras :  0
pytorch :  0
scikit :  0
tensor :  0
hadoop :  0
tableau :  1
bi :  1
flink :  0
mongo :  0
google_an :  0
job_title_sim :  data scientist
seniority_by_title :  na
Degree :  M
```

III. Data Preprocessing

A. Missing Values

Since the data set was extracted from Kaggle, the data was relatively clean, but there were still some factors that needed to be addressed. We began by looking for missing values by summing all of the null values in each column. We decided to approach this problem in this way so that we were able to see which columns had missing values, or more so how many. This would give us the ability to decide which attributes needed to be handled and the correct approach of how to handle them.

B. Creating a New Data Frame

We wanted to look more closely at particular columns, specifically job location. In order to do this, we utilized Python programming to manipulate the data and create dummy variables based on location. Take, for example, our job location column. We found it was more difficult to access information when having one column for all of the different states. Instead, we took the dummy variable approach. We created a new data frame with a separate column for each location that contained either 0 or 1. This indicated if the particular job was in the state specified by that column.

C. Determining Important Attributes

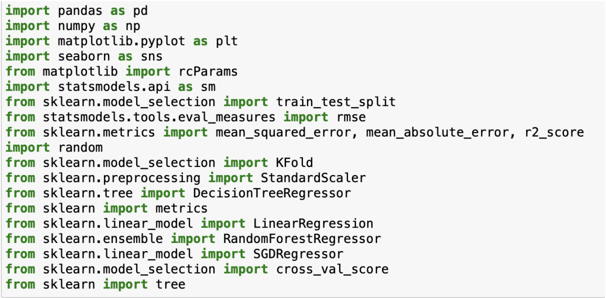
Determining which attributes of the data set will contribute accurate information is vital. After implementing a dummy variable approach, our data set had lots of columns that we deemed unnecessary to the greater implications of our research. Having too many columns/attributes can make it more difficult to visualize or access certain parts of the data.

First, we split our columns into two data frames: quantitative and qualitative. Of the numerical data frame, we decided to drop index, hourly wage, employer provided, lower salary, and upper salary. Hourly wage, lower salary, and upper salary were deemed insufficient, as we already had a column containing average salaries. Employer provided gave no useful information in our particular research.

Of the categorical data frame, we decided to drop job title, salary estimate, job description, company name, location, headquarters, industry, competitors, and company_txt. Job titles were all very similar because the entire data set was about data scientists. We were less interested about the various subsets of careers, and more so interested in the salaries that ranged across the entire field. Salary estimate was unnecessary, as we had the average salary column. Job description was long and unnecessary, but could be useful in further research regarding sentiment analysis. We dropped location because we had made the data frame with all of the dummy variables. Industry was not included because we were more focused on the sectors. Competitors were not a factor of our research.

IV. Machine Learning Models

Supervised learning is a machine learning approach to data analytics can be separated into two types of problems: classification and regression. It is defined by its utilization of labeled data sets that are to be trained to classify or predict accurate outcomes. Some machine learning examples included in our research are decision tress, random forests, and linear regression. Implementing the scikit-learn imports, we were able to use the various methods to run the aforementioned machine learning models, leading us to meaningful conclusions about the data.



```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib import rcParams
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from statsmodels.tools.eval_measures import rmse
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import random
from sklearn.model_selection import KFold
from sklearn.preprocessing import StandardScaler
from sklearn.tree import DecisionTreeRegressor
from sklearn import metrics
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import SGDRegressor
from sklearn.model_selection import cross_val_score
from sklearn import tree
```

A. Decision Trees

The goal of a decision tree is to form a model predicting a target variable's value by learning simple rules inferred from its own features. One advantage of a decision tree is its simplicity. They are easy to understand, interpret, and visualize. Regarding time complexity, the cost of using a decision tree is logarithmic with respect to the data points used in the training model.

On the other hand, decision trees present a disadvantage when the rules of the tree are over-complex and too tightly trained to the particular data set. This makes it difficult to generalize the data to a greater population.

B. Random Forest

Random forest is another supervised machine learning algorithm that implements methods used in decision trees. It builds multiple decision trees based off of different combinations of the attributes and takes the majority vote for the classification and average of the regression. Random forests are sufficient in reducing the variance for unstable classifiers without negatively impacting bias.

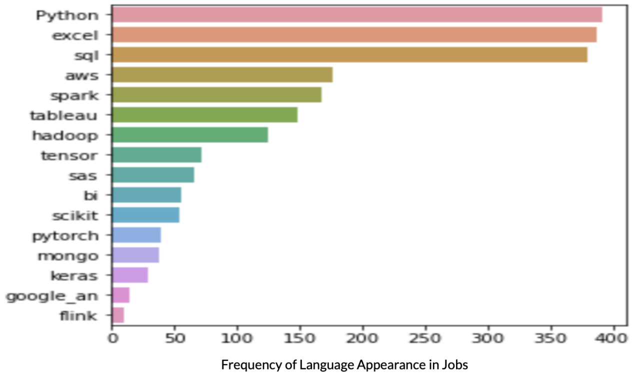
C. Linear Regression

Linear regression works by fitting a linear model that minimizes the residual sum of squares between two continuous variables (the observed targets in the data set and the targets predicted by the linear approximation). In simpler terms, it uses one independent variable to help explain the outcome of the dependent variable.

V. Results

A. Frequency Distribution of Programming Languages

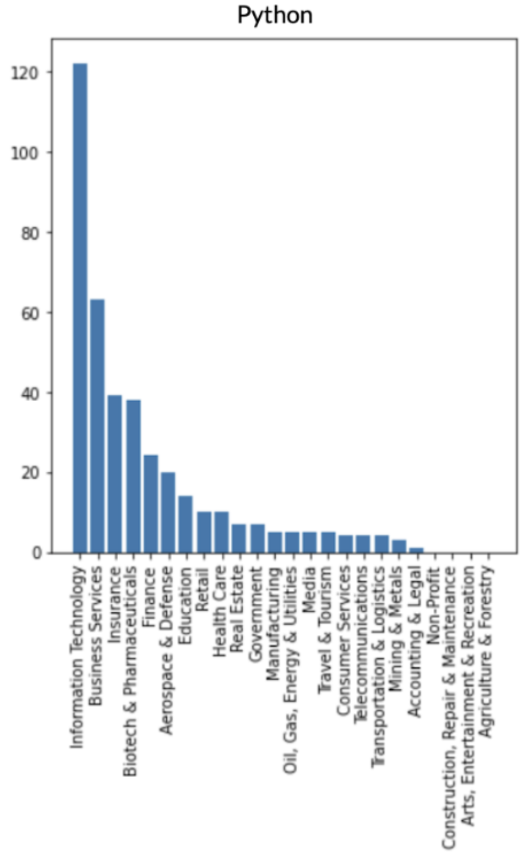
For our first visualization, we wanted to start by looking at the frequencies of each programming language in each of the job listings. The chart below displays the top three languages among our data set: Python, Excel, and SQL. Out of all the data, these three languages had similar frequencies. The next most common language, AWS, appeared in less than half of the job listings that contained the top three.



The figure above brought us to the conclusion that, if an applicant is looking to get a job in this field, it is almost necessary to be proficient in Python, Excel, and SQL because it appears so frequently in the data.

B. Python

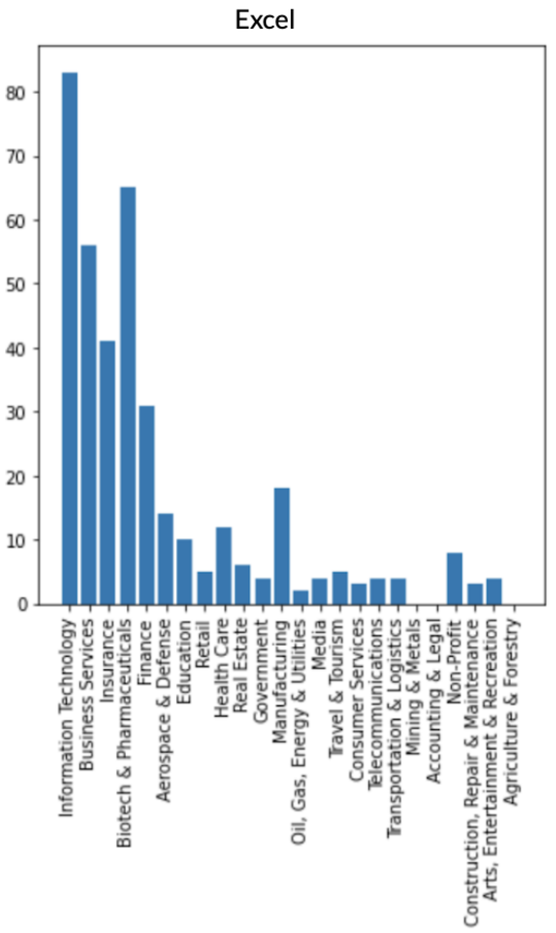
After determining that Python, Excel, and SQL were the top three programming languages necessary for a data science career. We wanted to further visualize their impact based on sector. To accomplish this, we created individual graphs to see how the most frequent languages translate to how common they are in each sector.



In the depiction of Python above, we found that sectors such as information technology, business services, biotechnology & pharmaceuticals, and finance are the top sectors for using Python. This makes sense because these are the sectors that would utilize Python in their every day job responsibilities. However, sectors such as mining & metals, media, and travel & tourism have less of a demand for this type of application.

C. Excel

Our next visualization depicts the same figure as Python's above, but instead for Excel. By comparing the two graphs against each other, we were able to observe some distinct differences between the distributions of Python and Excel, when analyzed based on sector.



It is evident that Excel has a greater variety of uses regarding different sectors. Excel was significantly more popular in biotechnology & pharmaceuticals than Python. There was also a jump in frequency for Excel's manufacturing sector.

Regarding the statistical analysis of our machine learning models, we wanted to see not just predictions of average salary for various data science careers, but also how accurate those predictions are based on the data we extracted. The random forest proved to have the best R-Squared value. For example, the random forest had an R-squared value of 0.759 whereas the decision tree model and linear regression model had R-squared values of 0.596 and 0.663, respectively. Therefore, they were all relatively close, but random forest proved the highest correlation among the data. In terms of adjusted R-squared values, the machine learning models showed similar behavior as the R-squared value outcomes. Additionally, the random forest model had the lowest root mean squared error (RMSE), 18.291. This was 5-10 units lower than the root mean squared error of the linear regression model and the decision tree model.

Consequently, we found that the leading attributes that revealed a correlation to higher salaries were: location, education level, and sector. With that, most companies mentioned having a masters degree when looking for candidates. For companies that mentioned having a PhD, they offered augmented salaries compared to those without. This proves that having a higher education level is important when discussing salary distribution. Additionally, the companies that were greater in size can afford to pay their employees more versus smaller, start-up companies.

Skills that were less common such as Flink, Mongo, and Keras had higher average salaries for a number of reasons. Firstly, not many people are familiar with these languages, so it is more difficult to find a skilled candidate in these areas. Further, these skills are more difficult to learn. Having a more complex skill set may lead to a higher salary. On the other hand, more common programming skills such as Python, SQL and Excel did not impact the salary distribution because it is expected for nearly every candidate to be more than proficient skills.

VI. Discussion

A. What Makes Our Research Unique?

There was an aspect of uniqueness that accompanied our research. Our data and findings were focused on real-life research that would impact not only us, but also anyone looking to enter the data science field. It is not only important to see which attributes lead to the highest salary, but also the skills that are most useful when joining the job force. As opposed to the other findings on Kaggle, we employed multiple machine learning algorithms to see the full extent of our data and provide visualizations that help communicate our results.

B. Limitations

Although our data set was extremely thorough and information dense, there were limitations to our data.

It surprisingly did not include R, a very well-known and useful programming language. R is more commonly associated with statisticians. But because statistics plays such a vital role in data science jobs, it still holds its value in the data science field.

Another limitation would be the size of our data set. Even though our data included a lot of information about each job description, it did not have as many data science job listing records as we would have liked. For instance, we had about 1,000 records of different data scientists, but it would be helpful to have more records if we wanted to look for stronger patterns in correlations. It may even bring about new conclusions regarding average salary, programming languages, and sectors.

VII. Conclusion

A. Concluding Statements

In conclusion, there are a number of reasons that can lead to having a higher average salary in the data science field. For example, where an applicant lives, how large the company they apply to is, skill set, and the education level of the candidate can all vary the distribution of the predicted average salary.

Out of all the machine learning algorithms employed in our research, we found the random forest model to be more robust than the decision tree model and the linear regression model. This is because the random forest revealed the best RMSE, R-squared, and adjusted R-squared values.

B. Where Can We Go From Here?

There are many directions the future of this research can take. In particular, targeting the gender wage gap and the lack of female employment in the data science field. Studies conducted by the Center for Global Development reveal that women currently represent just 18 percent of the entire data science field across the country [3]. This representation is further lowered when considering the accessibility of education and technology in underdeveloped countries. To battle this using future research, we would be able to gather data on women already in the data science field and which skills were the greatest factors in securing her employment.

References

[1] J. King and R. Magoulas, "O'Reilly Media - Technology and Business Training," 2016 Data Science Salary Survey, 2016. [Online]. [Accessed: 26-Apr-2022].

[2] N. Bhathi, "Data scientist salary," Kaggle, 29-Dec-2021. [Online]. [Accessed: 26-Apr-2022].

[3] U. Nwankwo and Michael Pisa and M. Pisa, "Why the World Needs More Women Data Scientists," Center for Global Development, 08-Mar-2021. [Online]. [Accessed: 26-Apr-2022]