Exploring Data Science Field Salaries

Joe Trotti, Alexis Rivera, and Christina Rodriguez

State University of New York at New Paltz

Abstract:

Data Science is an ever-growing field over the last decade with 650% growth since 2012. It continues to flourish with a predicted 36% growth over the next ten years, according to the Bureau of Labor Statistics. This is much faster than the average occupations expected growth. With that, Data Science is a considerably well-paying field with salaries above average. We will explore data science field salaries and various conditions influencing them, as well as perform a linear regression to predict salaries in years to come.

Introduction:

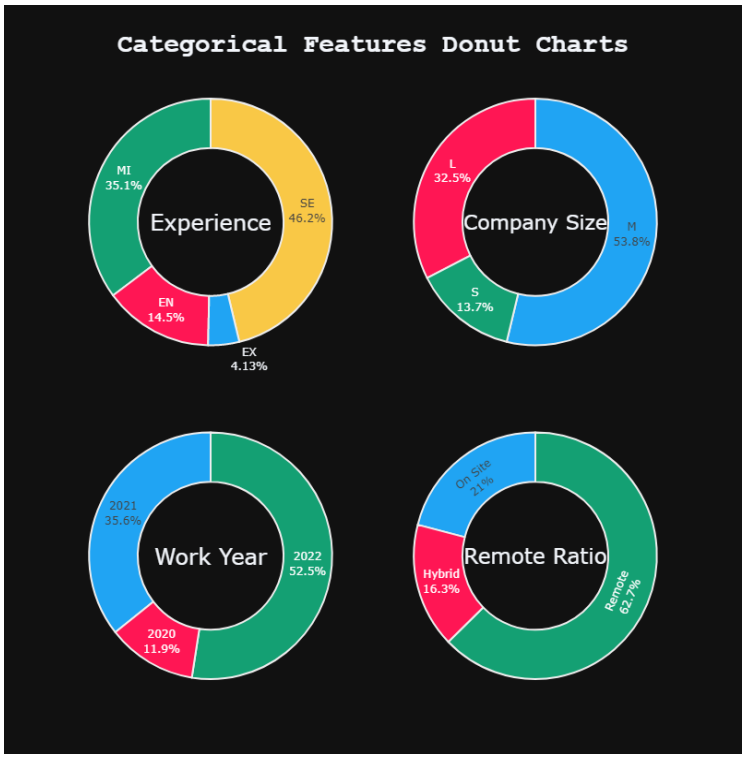
We are interested in looking at the effect on salaries based on various conditions, as well as making future predictions about salaries in the data science field. We will look at all salaries in US dollars. The conditions we will look at include job title, year, location, level of experience, company size, and remote work status. There are a wide variety of different types of jobs in our data set which include research scientist, machine learning engineer, and data architects. The years in our data set include 2020, 2021, and 2022. The locations include 49 different countries from across the world. The level of experience is broken down into four categories which are entry-level, mid-level, senior, and expert. Additionally, we have company size broken down into small, medium, and large, where we define a small company as one with less than 50 employees, a medium company as one with 50-250 employees, and a large company as one with greater than 250 employees. Lastly, the remote work status is broken down into three categories labeled on-site, hybrid, and remote, where on-site represents less than 20% remote work, hybrid represents 20-80% remote work, and remote represents greater than 80% remote work. We will first look at each condition individually and their relationship with salary.

Approach:

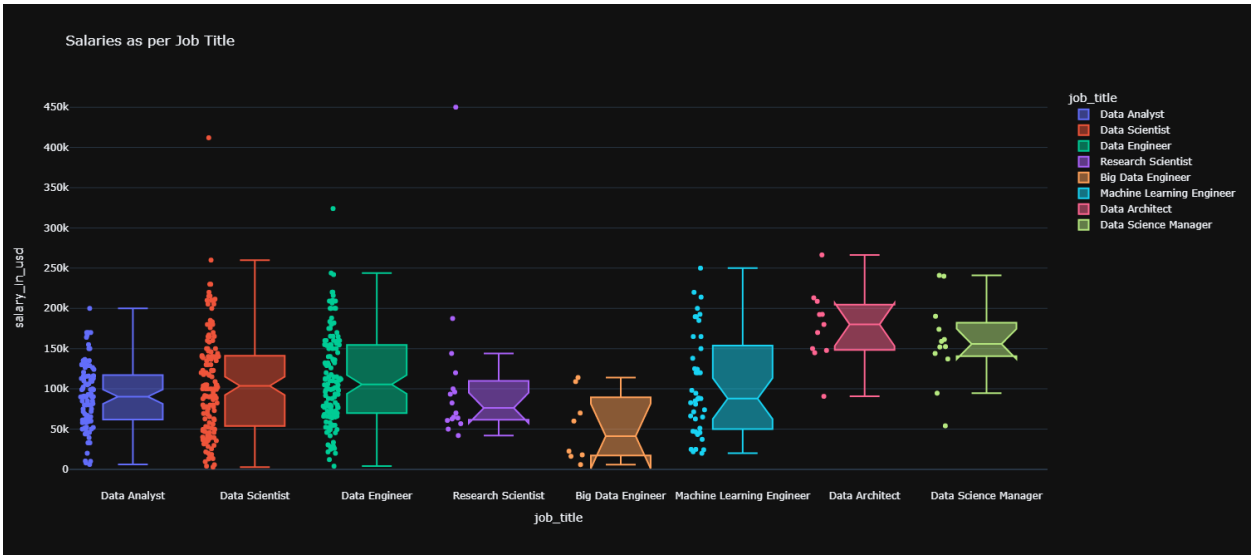
Before we could start working with our data set, our first step was to clean the data. To do so, we first checked for and replaced any missing data value. We also checked for and removed any duplicated data points or outliers. Next, we ensured consistent formatting and capitalization for better readability. We then used Hadoop to count the number of data points that meet specific criteria to help us organize the data. Hadoop counted 588 full time employees out

of 607. This meant that 16 were either part-time, contract or free-lance employees. Since these only made up 3% of our data, we omitted all except full-time employees.

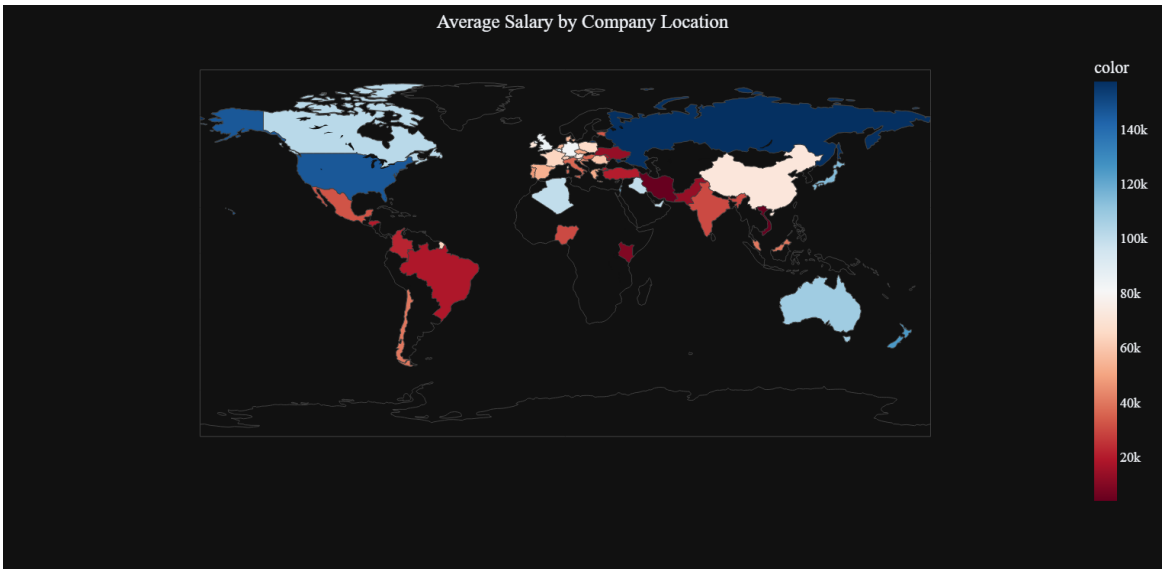
We looked at a breakdown of our data to see what type of data was the most abundant. For level of experience, our data included 46.2% senior, 35.1% mid-level, 14.5% entry-level, and 4.13% expert. For company size, our data included 53.8% Medium, 32.5% Large, and 14.7% small. For work year, our data included 57.5% for 2022, 35.6% for 2021, and 11.9% for 2020. Lastly, for remote work status, our data included 62.7% remote, 21% on-site, and 16.3% hybrid. The donut chart below shows the results.



We used Hadoop to find the eight most abundant job types in our data set. These jobs were research scientist, big data engineer, machine learning engineer, data architect and data science manager. The box and whisker plot below shows these results. We found that data architects had the highest median salary of ~180k with an upper extreme of ~250k, while big data engineers had the lowest median salary of ~40k, with a lower extreme of ~5k. We noted that data scientists had the most widespread data, meaning that there was a wide variation between data points. Meanwhile, research scientists had a much smaller spread compared to other jobs, meaning that the data points stayed within a much shorter range.

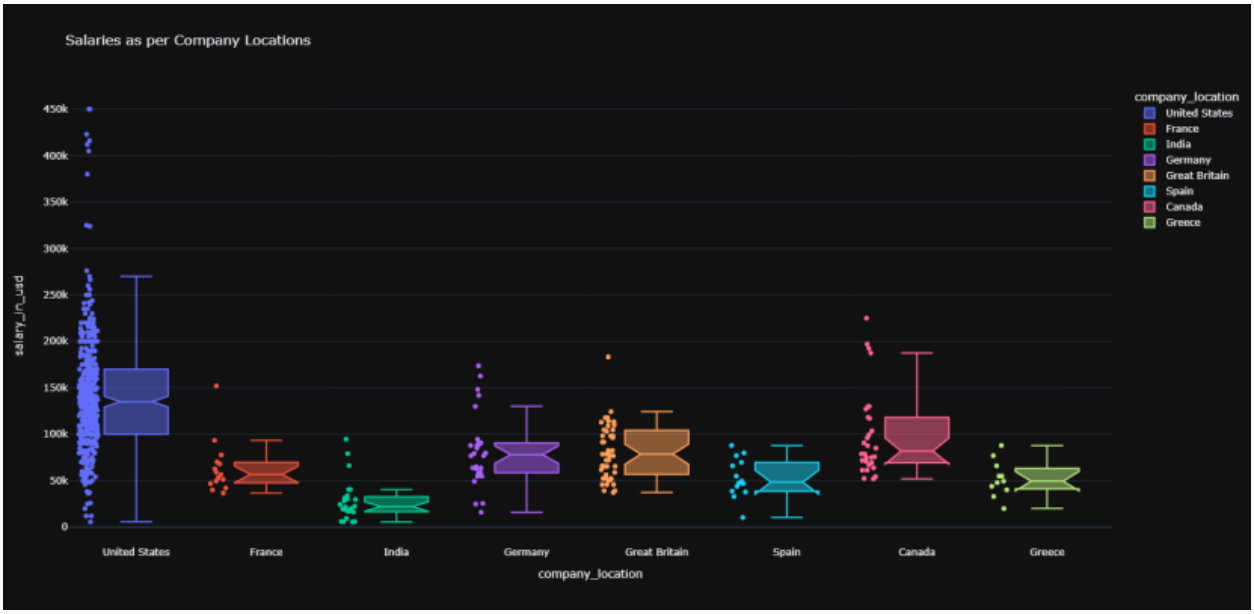


We also looked at the salaries by company location. The graph below is a visual representation of average salaries by country. Cooler colors represent higher salaries, while warmer colors represent lower salaries. As you can see, Russia and the US have the highest average salary, while Iran has the lowest average salary. This was immediately surprising to us, which prompted us to take a deeper look at the salaries in Russia. We realized that we had very little data for Russia and that one of the few data points was a major outlier of ~230k skewing the data so that Russia had the highest average.



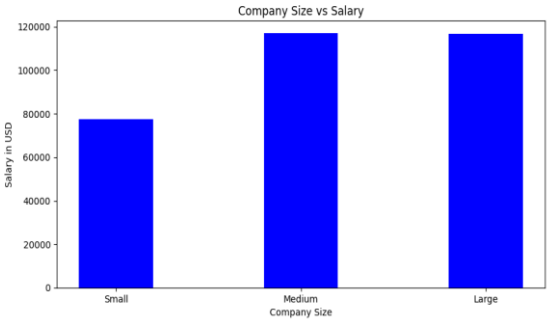
We realized the above graph was not an effective way to look at the relationship between salary and location. Therefore, we decided to reevaluate the way we viewed this data. We first use Hadoop to find the top eight most abundant countries. That is, the United States, France, India, Germany, United Kingdom, Spain, Canada, and Greece. We decided to create a box and whisker plot to look at the top eight countries. Using the box and whisker plot

eliminated skewing of the data due to outliers. This is because this form of graph uses the median instead of the average. This graph allowed us to clearly see how the data was spread as well as easily identify outliers. We found that the US had the highest median salary of ~140k as well as the highest upper extreme of ~275k, while India had the lowest median salary of ~20k and lowest lower extreme of ~5k.



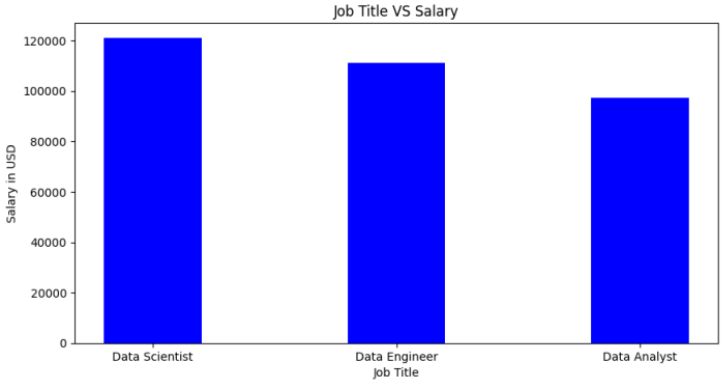
The following bar graph represents the relationship between salary and remote work status. We noticed that the average salary for remote and on-site work did not have a substantial difference, with remote averaging at ~115k and on-site averaging at ~105k. This meant that remote work status had little influence on salary and therefore was not a good predictor of salary.

This next graph looks at salary versus company size. You can see that there is a drastic difference between average salaries in small companies versus medium and large companies. Small companies averaged at ~78k, while medium and large companies averaged at ~115k. We noted that company size turned out to be a good predictor of salary as the disparities between small versus medium or large companies was substantial.

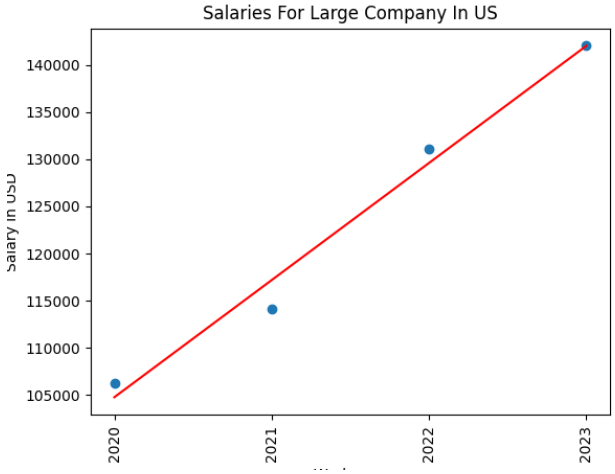


Lastly, we looked at the relationship between salary and level of experience. As expected, there was a positive correlation. That is, entry-level employees made the lowest salary and experts made the highest salary.

Another step we took was using Hadoop to categorize jobs into three broader categories: data analyst, data scientist, and data engineers. This is because we needed a way to group jobs that were extremely similar or the same. For example, a “big data engineer” and “data analytics engineer” were both grouped into the data engineer category. Performing a wordcount in Hadoop, we counted 143 data scientists, 218 data analysts, and 246 data engineers in our data. We used this to look at salary versus job title. As shown in the graph below, we found that data scientists had the highest salaries at ~120k, with data engineers following behind at ~110k and data analysts at ~100k.



Our final step was to perform a linear regression to predict salaries for future years. Because our data had many factors influencing salary, we first filtered the data by looking at data points from the US only. This was in order to eliminate the effect that company location had on salary. In addition, we decided to look at large companies only to eliminate the effect that company size had on salary. Below shows the results of our linear regression. Our line of best fit was the line y = 12412.86x + 92369.96. We used this to predict the average salary for 2023. Our prediction was ~142k USD. This is a 9.6% increase from 2022 where the average was $129,600.



We realized that we would have had a more accurate prediction if we had performed a multi-regression instead of a linear regression. Unfortunately, we did not realize this until late in the project and struggled with time constraints and related issues. For this reason, we performed a linear regression. Another aspect that could have been better would be finding a data set with more much more data and a greater range of years. These changes would have drastically improved our project and made for more accurate results.

Conclusion:

The Data Science field has grown rapidly over the years and continues to grow. As expected, our data has shown that salaries for data science fields are projected to increase in future years. Our linear regression has shown that the projected salary for 2023 is ~$142,000. We concluded that the best predictors of salary were company size, location, and level of experience. Overall, the field of data science offers a wide range of potential salaries. Those interested in pursuing a career in data science can expect to earn a competitive salary, with opportunities for growth and advancement.

References:

Chauhan, Aman. “Data Science Fields Salary Categorization.” *Kaggle*, 10 Sept. 2022, https://www.kaggle.com/datasets/whenamancodes/data-science-fields-salary-categorization.

Leekahwin. “Mastering Tuning 10 Regressors (>0.5 r2).” *Kaggle*, Kaggle, 23 Sept. 2022, https://www.kaggle.com/code/leekahwin/mastering-tuning-10-regressors-0-5-r2.

“Data Scientists: Occupational Outlook Handbook.” *U.S. Bureau of Labor Statistics*, U.S. Bureau of Labor Statistics, 8 Sept. 2022, https://www.bls.gov/ooh/math/data-scientists.htm.