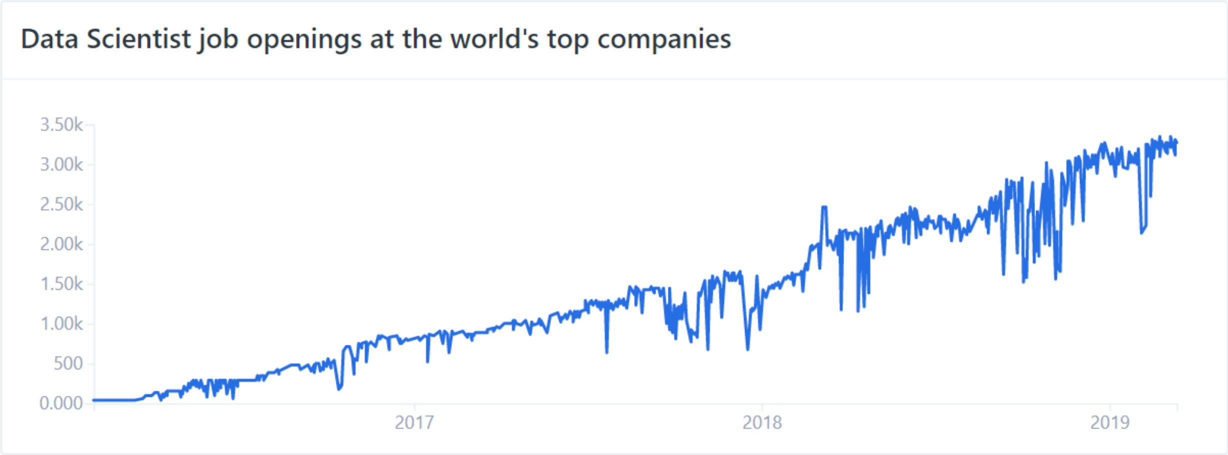
PROJECT REPORT
ANALYSIS OF SALARIES IN DATA-RELATED JOBS

CSIS 3360 – Fundamentals of Data Analytics - (Summer 2023)

by
AMIRHOSSEIN KARIMI- 300360903
SUDARSAN HARIDAS – 300353099

DISCOVERY

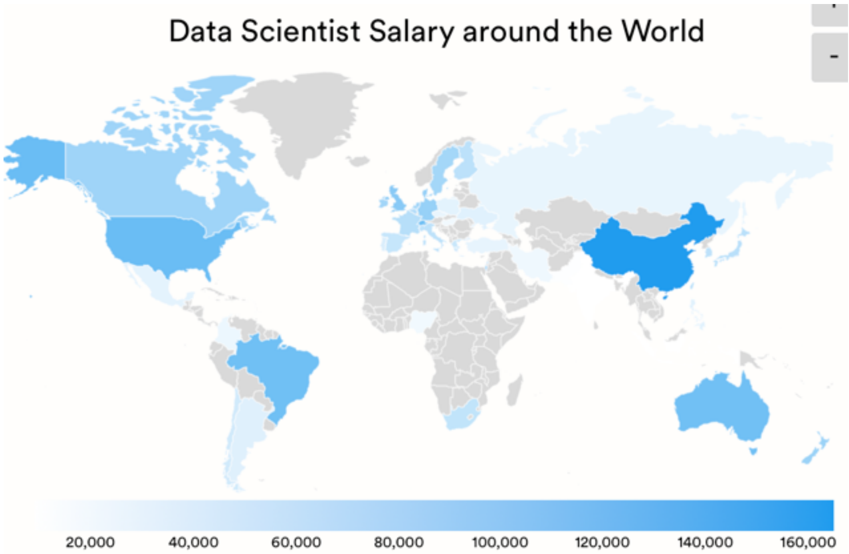
The growth in the amount of data being generated globally over the years has proportionately increased the number of jobs related to data. The upwards trend, though seeming to be stabilized is expected to grow exponentially in the next decade. The world needs people with knowledge in handling data.



The surge in demand over the years has also increased the number of candidates interested in those jobs. And as any organization's goal is to minimize costs, the expanding supply of skilled and trained personnel is a gold mine they can buy for cheap. There has been a long-standing conversation if employees are being fairly compensated. The extreme variation in pay scale in this field has drastic effects on how over or under-paid a beginner enters this field. The factors that could affect the salary of an individual would be his/her educational background, experience, the size of the organization, etc. The data was collected during Stack Overflow developer summits over the years 2017 - 2020. It was then filtered to have only entries from data-related jobs. The analysis of salaries in the field is currently the need of the hour for many young and aspiring individuals.

DATA PREPARATION

The data collected in the survey, except the columns where programming languages in use, database in use, and database that the developer wishes to learn were collected, are used for this analysis. The categorical columns containing a long list of unique entries were grouped down to a maximum of five unique entries. For example, the education level column had multiple entries but was grouped down to have if the developer completed a Bachelor, Master, or Doctoral degree or finished certificate-level studies, or had completed no formal education at all. Also, since the converted compensation was in US Dollars, and also because countries other than Canada and USA had very low salaries comparatively, Canada and USA



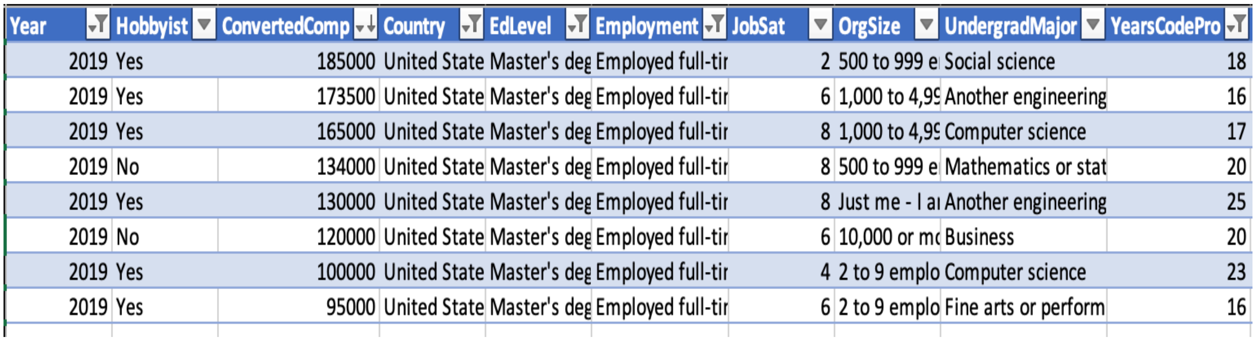
were grouped together as North America, and every other entry was grouped as Other. These are some examples of the data processing carried out.

MODEL PLANNING & IMPLEMENTATION

The use of a Multiple Linear Regression model to predict the salary of an individual and a Logistic Regression model to classify if an individual is satisfied with his/her job or not is proposed. The Linear Regression model in the first case is due to the continuous nature of the target variable and the Logistic Regression model in the case of classification is due to the discrete and binary nature of the target variable. The prediction process can be implemented well with just a simple Linear model. The classification, however, can be carried out with multiple techniques but in this case, the classification is binary and also because it is not the main objective. The chosen techniques provide results and coefficients to verify those results and to gain insights on how and by how much a factor affects the outcome.

RESULTS INTERPRETATION

The Linear Regression model performed moderately with an R-squared score of 0.4262, where the provided factors explain 42.62% variability in the outcome variable. Although the score was low, the coefficients produced by the model happen to have statistical and literal significance. The inaccurate nature of the model, despite the valid nature of the coefficients produced, is due to the high variability in salary in the dataset. As shown below, even when all the factors are matched, the salary varies by a lot.



Year	Hobbyist	ConvertedComp	Country	EdLevel	Employment	JobSat	OrgSize	UndergradMajor	YearsCodePro
2019	Yes	185000	United State	Master's deg	Employed full-tir	2	500 to 999 e	Social science	18
2019	Yes	173500	United State	Master's deg	Employed full-tir	6	1,000 to 4,9	Another engineering	16
2019	Yes	165000	United State	Master's deg	Employed full-tir	8	1,000 to 4,9	Computer science	17
2019	No	134000	United State	Master's deg	Employed full-tir	8	500 to 999 e	Mathematics or stat	20
2019	Yes	130000	United State	Master's deg	Employed full-tir	8	Just me - I a	Another engineering	25
2019	No	120000	United State	Master's deg	Employed full-tir	6	10,000 or m	Business	20
2019	Yes	100000	United State	Master's deg	Employed full-tir	4	2 to 9 emplo	Computer science	23
2019	Yes	95000	United State	Master's deg	Employed full-tir	6	2 to 9 emplo	Fine arts or perform	16

The addition of data or a more comprehensive data processing of each column might solve the low accuracy problem. And as expected every single factor proves to be statistically significant in determining the salary of an individual. The model tends to overcompensate to fit for values from the United States, but even when running the model with data only from the United States the accuracy problem exists, due to high variability in salary as shown in the above list of values. The same can be said for the classification model, as job satisfaction is very subjective, regardless of external factors.

PREDICTIVE ANALYSIS

The predictive Linear Regression model was used with generated data for a beginner Data Analyst, looking for a job in the United States after completing his/her Master’s Degree. Initially, the expected salary is around $90,000 when looking to work for a company. The same individual if looking for a job outside of the United States, is expected to earn only around $41,000, which is less than 50% of the expected North American beginner salary. If the same individual chooses to be a Data Analysis consultant and be self-employed, the expected salary increases by almost 10% to around $98,500.

Over the years, after gaining almost 5 years of experience, the salary is expected to jump to around $108,000. These values give an approximate insight into what an aspiring Data Analyst can expect to earn. These predictions prove that the model has valid assumptions on the provided real-world factors.

CONCLUSION

The above analysis gives an individual looking to enter the data field or already in the data field a brief insight into what to expect as a salary. The topic of salary transparency is an ongoing need of the hour to tackle disparity in pay. With regards to the prediction of salary, all the provided factors are shown to be statistically significant in determining the salary of an individual. But, out of the given factors, an individual not having any formal education does not have as much significance as the other factors. This might be because individuals without a formal education, tend to have niche skills that apply to a particular job. And surprisingly an individual having a major relevant to the job negatively affects his/her salary. This does not make sense in the real world but might be because a person with an irrelevant degree is hired into a technical job because of his/her domain expertise, which is valued more than technical knowledge in the data field. Domain expertise is required to be able to interpret the results provided by technically equipped individuals. Thus, an individual looking to enter the field also already existing and looking to get a hike should look to improve his/her domain expertise, negotiate a fair salary after gathering necessary input, and also explore the options of being self-employed as it provides more control in terms of compensation and work-life balance.