DS 3001: Project - Proposal

Group Details	**Group 3** **Tharindu Fernando : 15522** **Navod Madhuwantha : 15541** **Janith Ravindu : 15553**

Title of the Project

Influential factors of the remunerations of Data Science related jobs and how it's been evolved

Introduction:

* Data Science is a multidisciplinary field where computer science and statistics meet each other and blends up with business acumen with sheer intention of extracting hidden insights from the data and realize them into impactful decisions .

* The origin of data science can be traced back to the early days of statistics , when scientists and engineers began to use statistical methods to analyse data . However, it was not until the 1990s that the field of data science really began to stemmed out . Because of the overwhelming growth of data and their potential utilities it urges to widen up the capabilities of data analysing techniques . So it got collide with AI and that kind of cause for the role of data scientist to be branched out in several aspects and a lot of new roles emerges with that such as ;

-> Data scientist / Data engineer (ETL) / Machine Learning Engineer / Big Data Engineer / Computer Vision Engineer / Finance Data Analyst / Business Intelligence Analyst / NLP Engineer

* Even though the initial role of data scientist has been split up into a couple of different roles as stated above we are curious to know that still that demand is outstripping the supply , or else the job opportunities of the lower tier (freshers) have been already saturated . Since we don't have any direct measure to gauge that straight away , we going to observe how the salaries have been evolved over the course of most recent years and

What it reflects .

* So this will help to thousands of data science aspirants who are out there looking forward to take career transitions towards data science let alone who are willing to make their path forward in data science or AI related field .

Objectives:

* To understand the current market value of data science skills & To identify factors that influence the salaries of data science related jobs :

(There are quite a lot of factors are playing around determining the incentives for any designation that related to data science . So our primary concern is to pinpoint some of the key factors and analyse their unique traits as well as the impact them having on the wages using variety of makers . And we looking forward to put a lot of strain on visual representations (plots, charts , graphs , etc,....) to acquire those makers.

* To track the trends in the salaries of data science related jobs over time.

*To compare salaries across different industries and locations at the same time how the salary varies with the scale of the organisation .

Data:

work_year : The year which the observation was taken (Numerical)
executive_level : Wether the individual is holding an executive position or not(Categorical)
employment_type : Wether it is full time or part time (Categorical)
job_title : Designation (Categorical)
salary : Annual salary in USD (Numerical)
employee_residence : Location of the employee (Categorical)
remote_ratio : Wether the employee is working fully remotely or not (Categorical)
company_location : Location (country) of the company (Categorical)
company_size : The scale of the firm (Categorical)

Proposed Analysis Plan

- First we going to go ahead with the pre ritual of data analytics which is data cleaning . This consists with;

  (Remove outliers , Impute missing values , Scale down the necessary numerical variables , Encode categorical features , Reduce the cardinality of categorical features , In here we going to use suitable plots both before and after each technique being applied to make sure that , the method that we used didn't distort the original structure of the data)

- **Step 02** - , we going to analyse each an every variable individually . So we can observe and more importantly capture subtle insights that each individual feature is holding . In here we going to split it up into 2 parts , the first part is for numerical variables and the second part is for the categorical variables
- **Step 03** - Now we going to conditionally render the data frame based on the designation , and carry out separate analysis for each distinguish category .
- **Step 04** - We going to analyse the key factors which above plots convey and after that we going to be making comparisons between each different designation
- **Step 05** – Analyse the relationship between different numerical features and the relationship between each variable with our dependent variable salary . Hence we try to come with new meaningful features which are more correlated with our dependent variable
- **Step 06** - We going to summarize all of the key conclusions that were drawn throughout the analysis .

References: Kaggle