Syracuse University

Memo

To: Dr. Landowski

From: Nora Lin and Michael Morrey

Date: May 8th,2022

Re: Project Proposal

Topic: Analyze data scientist salaries

Data Description:

The main data set is from Kaggle, and consists of job postings related to the position of “Data Scientist” in the USA, scrapped from Glassdoor.com. There are 742 observations and 42 columns. Here is the link to our dataset: https://www.kaggle.com/datasets/nikhilbhathi/data-scientist-salary-us-glassdoor

Fields	Description	Example
index	Index	1
Job Title	The title of the job, e.g. Data Scientist, Junior Data Scientist, etc.	Data Scientist
Salary Estimate	Range of Salary and the source	$53k - $91k (Glassdoor est.)
Job Description	Indicates what qualities that company wants and what is expected out of the job title	Data Scientist Location: Albuquerque, NM Education Required: Bachelor’s degree required, preferably in

		math, engineering, business, or the sciences. Skills Required: Bachelor's Degree in relevant field, e.g
Rating	Rating of the company	3.8
Company Name	Name of the company	Tecolote Research 3.8
Location	Location of the job (city, state)	Albuquerque, NM
Headquarters	Location of the headquarters of the company (city, state)	Goleta, CA
Size	Range of the number of employees working in the company	501-1000
Founded	Company founded in year	1973
Type of ownership	Type of ownership	Company - Private
Industry	Industry of the company	Aerospace & Defense
Sector	Sector of the company	Aerospace & Defense
Revenue	Revenue of the company	$50to $100 million (USD)
Competitors	Company competitors. If not successfully scrubbed, then -1	-1
Hourly	If the job is paid hourly (0=no, 1=yes)	0
Employer Provided	If the employer name is provided (0=no, 1=yes)	0
Lower salary	Lower end of the salary range	53
Upper Salary	Higher end of the salary range	91
Avg Salary(K)	Average salary of the job posting	72
company_txt	Name of the company	Tecolote Research
Job Location	State abbreviation of the job location	NM

Age	Age of the company in years	48
Python	Python programming required (0=no, 1=yes)	1
Spark	Spark programming required (0=no, 1=yes)	0
AWS	AWS programming required (0=no, 1=yes)	0
Excel	Excel required (0=no, 1=yes)	1
SQL	SQL required (0=no, 1=yes)	0
SAS	SAS required (0=no, 1=yes)	1
Pytorch	Pytorch required (0=no, 1=yes)	0
Scikit	Scikit required (0=no, 1=yes)	0
Tensor	Tensor required (0=no, 1=yes)	0
Hadoop	Hadoop required (0=no, 1=yes)	0
Tableau	Tableau required (0=no, 1=yes)	1
Bi	BI required (0=no, 1=yes)	0
Flink	Flink required (0=no, 1=yes)	0
Mongo	Mongo required (0=no, 1=yes)	0
Google_An	Google analytics required (0=no, 1=yes)	0
Job_title_sim	Simplification of the job title (ex. Data scientist, analyst)	Data scientist
seniority_by_title	Seniority by title (sr, jr, na)	na
Degree	Degree level required (M=Masters, P=PhD, na=not specified)	M

We may find another dataset to merge this one with. However, we currently do not have a secondary data set.

Research Questions:

1. Which companies have the highest rating?
2. Do company ratings have an impact on Data Scientist salaries?
3. Which geographic locations offer the highest salary for Data Scientists?
4. In which states are data scientists the most in demand?
5. Which sectors rank the highest based on the average salary?
6. Which technical skills have the highest demand for Data Scientist related positions?
7. What is the relationship between salary and education level?
8. Which job titles have the highest salaries, and which are the most in demand?

Data Preparation Plan:

1. Load the data into Python
2. Exclude columns that we are not interested in analyzing
3. Rename any column names that have spaces so they no longer have white space
4. Decide how to handle the various -1 and NA values that are in the data, whether it be deleting them or replacing the values with something else
5. Explore the data for additional anomalies or missing data, and decide on how to handle those issues if any are found
6. Clean columns that have string values for numeric ranges, such as Size, so they can be converted to a numeric data type