Task 6.1: Sourcing Open Data

As an aspiring data analyst who will soon be looking for a job, I have decided to do some research on the job market for data analysts. I am having a hard time deciding between two datasets which are both about salaries in data-related fields. I'll call the first dataset "Analyst Salaries" and the second dataset "STEM Salaries." What I like about "Analyst Salaries" is that it is more specific to the career that I'm pursuing, and it has information on whether the work is remote or on-site and whether the work is part time, full time, contract, or free-lance. What I like about "STEM Salaries" is that it has more specific work locations ("Analyst Salaries" only gives the country, while "STEM Salaries" gives the city). It's also more specific on the experience level, giving the years of experience rather than a category. It also has years at the job as a variable, as well as other variables not present in "Analyst Salaries," such as gender, race, and education level. I think that both datasets could be useful, though I'm not sure that I'd be able to merge them in any useful way. I will go ahead and profile both datasets below.

Analyst Salaries

Data Source

- Source: This data was downloaded directly from ai-jobs.net on June 20, 2023.
- Data Collection: ai-jobs.net is a platform where employers can post jobs in the fields of artificial intelligence, machine learning, and data science/analysis. The website allows users to answer a survey anonymously about their salary, experience level, job title, etc. The salary data is then regularly updated with results from these surveys, and the dataset is available to the public for download.
- Data Contents: This dataset contains information about the salaries of individuals (from 2020 to the present) who work in the fields of data science, data analytics, data engineering, machine learning, and artificial intelligence. There are 5,293 rows, each representing an individual person. The columns are work_year, experience_level, employment_type, job_title, salary, employee_residence, remote_ratio, company_location, and company_size. For more information, see the Data Profile below.
- Data Relevance: This dataset is very relevant to my aim to understand the job market for data analysts. In particular, I am interested in the information in the employment_type and remote_ratio columns. The dataset is as current as possible, given that it's updated on an ongoing basis, was accessed this June, and contains data going back only to 2020.

Data Profile

- The data originally had 5293 rows and 11 columns. After cleaning the data, it has 5293 rows and 9 columns.

Variable	Description	time-variant?	structured?	variable type	subtype
work_year	The year the information was collected	time-invariant	structured	quantitative	discrete
experience_level	EN = entry level / junior MI = mid-level / Intermediate SE = senior-level / expert EX = executive-level / director	time-variant	structured	qualitative	ordinal
employment_type	PT = part-time FT = full-time CT = contract FL = freelance	time-variant	structured	qualitative	nominal
job_title	The job title of the employee	time-variant	unstructured	qualitative	nominal
salary_in_usd	The salary of the employee	time-variant	unstructured	quantitative	continuous
employee_country	The employee's primary country of residence as an ISO 3166 country code	time-variant	structured	qualitative	nominal
remote_ratio	The amount of work done remotely. Possible values: 0 = no remote work (less than 20%) 50 = partially remote/hybrid 100 = fully remote (more than 80%)	time-variant	structured	quantitative	discrete
company_country	The country of the employer's main office or contracting branch as an ISO 3166 country code	time-variant	structured	qualitative	nominal
company_size	The number of employees at the company. S = small (less than 50) M = medium (50 to 250) L = large (more than 250)	time-variant	structured	qualitative	ordinal

- Limitations and ethics
  - Ethics: All data here were supplied voluntarily, and they contain no PII.

- Limitations: Because the data were obtained by voluntary response, there is a possibility of sampling bias. In particular, individuals not looking for work (perhaps because they are happy with their current salary, for example) would not have much occasion to visit ai-jobs.net, where the voluntary survey is housed.

STEM Salaries

Data Source

- Source: I accessed this data on Kaggle. It was scraped from levels.fyi by Jack Ogozaly using the technique outlined in this article. These data were scraped in 2021.
- Data Collection: The data collection here is similar to that of the Analyst Salaries dataset. levels.fyi is another job posting and searching platform that allows users to anonymously submit information about their salary, experience level, job title, etc. The data is not directly available for download, and thus, was scraped from the website.
- Data Contents: This dataset contains information about the salaries of individuals who work in STEM and data fields. These data were collected between 2017 and 2021. There are over 62,000 rows, each representing an individual. The columns include information on salary, company, experience level, years at company, job location, education, gender, and race. For more information, see the data profile below.
- Data Relevance: Because this dataset was scraped in 2021, it may be a little outdated. It is also broader than what I need in order to understand the job market for my particular field. However, it does contain information about jobs in data analytics, and understanding those within the broader context of STEM jobs could be useful. Moreover, this dataset has good information about companies, location, experience level, education, race, and gender that are not present in the "Analyst Salaries" dataset, which I think could be very useful.

Data Profile

- The data originally had 62,642 rows and 29 columns. After cleaning, it has 62,598 rows and 16 columns.

Variable	Description	time-variant?	structured?	variable type	subtype
timestamp	Date and time the information was collected	time-invariant	structured	qualitative	ordinal
company	Company the employee worked for	time-variant	unstructured	qualitative	nominal
level	The employee's level in the company's management structure	time-variant	unstructured	qualitative	ordinal

title	The employee's job title	time-variant	unstructured	qualitative	nominal
totalyearlycompensation	The employee's total compensation per year	time-variant	unstructured	quantitative	continuous
location	The city, state, and country of the company's main office or contracting branch	time-variant	structured	qualitative	nominal
yearsofexperience	The number of years the employee has worked in this field	time-variant	unstructured	quantitative	continuous
yearsatcompany	The number of years the employee has worked for this company	time-variant	unstructured	quantitative	continuous
tag	An additional column used to place job titles into categories	time-variant	not sure	qualitative	nominal
basesalary	The employee's salary	time-variant	unstructured	quantitative	continuous
stockgrantvalue	The contribution (in usd) of stock to totalyearlycompensation	time-variant	unstructured	quantitative	continuous
bonus	The contribution of bonuses to totalyearlycompensation	time-variant	unstructured	quantitative	continuous
gender	The employee's gender	time-invariant	structured	qualitative	nominal
otherdetails	Other information about the job	time-variant	unstructured	qualitative	nominal
race	The employee's race	time-invariant	structured	qualitative	nominal
education	The highest level of education completed by the employee	time-variant	structured	qualitative	ordinal

- Limitations and ethics
  - Ethics: All data here were supplied voluntarily, and they contain no PII.
  - Limitations: Like the previous dataset, these data were obtained by voluntary response, so there is a possibility of sampling bias. Again, individuals not looking for work (perhaps because they are happy with their current salary, for example) would not have much occasion to visit ai-jobs.net, where the voluntary survey is housed. However, this dataset has a couple additional issues. One, it is two years old, so the information may be a little dated. Two, it has a lot of missing values. Unfortunately, some of the data that I was really excited about in this dataset (like race, gender, and education) are missing in almost half the rows.

Questions to Explore

1. What are my career options as a data analyst?
    a. What is the likelihood of finding a remote job? How has the availability of remote work changed in recent years?
    b. What is the likelihood of finding a part-time job?
    c. What are the top companies hiring data analysts? What size are those companies?
2. What might my salary be like as a data analyst?
    a. How do salaries compare among the various job titles in the datasets?
    b. How is salary affected by years of experience? How is it affected by years at the company?
    c. How have salaries changed over time?
    d. Is salary affected by size of the company? Amount of work done remotely? Employment type?
3. What kind of quality of life do data analysts have?
    a. Where do they tend to live?
    b. How many of the survey respondents are currently at their first job?
4. What demographic factors (race, gender, education level) affect salary?