**What dataset or datasets do you plan to use? What are the features, rows, and data types of each?**

The dataset used in this analysis contains Data Analyst job openings and several attributes related to each job. The dataset is located on Kaggle, and the information was scraped by an individual from the popular job website, Glassdoor. The dataset has 2253 rows and 16 columns. Each row is a record of a unique job opening with the following attributes:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| Unnamed: 0 | This is the record number | int64 |
| Job Title | Job position title | object |
| Salary Estimate | Range of salary estimate (USD) | object |
| Job Description | Description of jobs | object |
| Rating | Company rating on scale from 0 – 5.0 | float64 |
| Company Name | Name of company | object |
| Location | Job location | object |
| Headquarters | Company headquarters location | object |
| Size | Company size | object |
| Founded | Year company was founded | int64 |
| Type of ownership | Nonprofit, private, public, government, etc. | object |
| Industry | Company’s industry | object |
| Sector | Company’s sector within the industry | object |
| Revenue | Annual company revenue | object |
| Competitors | Company’s main competitors | object |
| Easy Apply | Glassdoor’s easy apply available or not | object |

Link to the dataset: <https://www.kaggle.com/andrewmvd/data-analyst-jobs>

**What research or business questions do you want to answer?**

The goal of this analysis is to identify differences in data analyst salaries by job sector. The dataset for this analysis was scraped by an individual from the popular job website, Glassdoor. This analysis can give aspiring and current data analysts (DA) insights on the most popular job sectors hiring and help these professionals identify which sectors pay the most.

The three most popular sectors in the dataset will be analyzed: Information Technology (IT), Business Services, and Finance. Three statistical tests will be performed:

1. T-test of average salaries for IT DA jobs vs Business Services DA jobs
2. T-test of average salaries for IT DA jobs vs Finance DA jobs
3. T-test of average salaries for Business Services DA jobs vs Finance DA jobs

**What are your hypotheses going in?**

I believe that there are significant differences in average Data Analyst salaries from sector to sector. In my opinion, the ranking by average salaries for these three sectors will be: Finance, IT, and Business Services. The following is a summary of the top three sectors prior to data cleaning:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sector** | **Job Count in Dataset** | **Sector Description** | **Popular Companies** |
| Information Technology | 570 | Technology driven companies that offer physical and virtual products; some offer staffing services as well | Staffing Technical Services, LLC, Diverse Lynx, Apple, TikTok, Amazon |
| Business Services | 524 | Business consulting and staffing services | Kforce, RobertHalf, Mondo |
| Finance | 169 | Financial institutions that provide banking and investment products | MUFG, Bank of New York Mellon, Citi, JPMorgan Chase, Goldman Sachs |

Now, having a better understanding of the companies in each sector and based on personal knowledge of salaries from my friends that work for some of these companies, I predict the following hypotheses:

1. A significant difference will be observed for average salaries of IT DA jobs vs Business Services DA jobs
2. A significant difference will be observed for average salaries of IT DA jobs vs Finance DA jobs
3. A significant difference will be observed for average salaries of Business Services DA jobs vs Finance DA jobs

**How will you use your data to test your hypotheses?**

First, I will clean the dataset by dropping all records without a ‘Salary Estimate’ value. Then, for each record, I will extract the upper and lower values from the ‘Salary Estimate’ column and create a column named ‘Average’, that calculates the average salary of the upper and lower ‘Salary Estimate’ columns for that record.

Next, I will create three new DataFrames, one for each of the top three sectors, and use the statistics library function, ttest.ind() with unequal variances, to perform all t-tests as indicated above. The test results will be calculated on a 95% confidence interval.

**Who will find your findings valuable, and how will they use them?**

Individuals interested in data analyst salaries will find this analysis useful. This information will help these individuals identify “in-demand” sectors and provide a quantifiable range of average salary differences from sector to sector.