Data Science project-IT 362

Web scraping project

Maha Almuaythir, Daniyh Aleid, Nahla Aljeraisy, Jomanah Alomar

Introduction

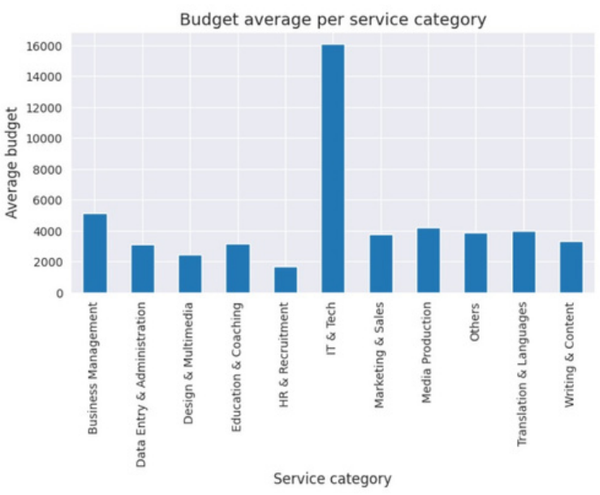
Freelancers domain is an interesting topic, we found a few questions to ask and a suitable website then proceeded to web scrape it creating a dataset for us to analyze and answer our questions. Then we created a classifier prediction model.

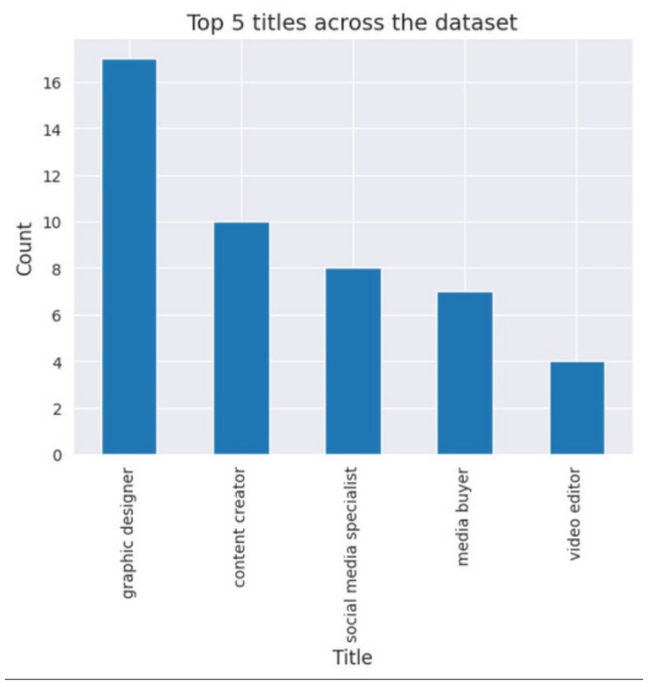
Data Analysis

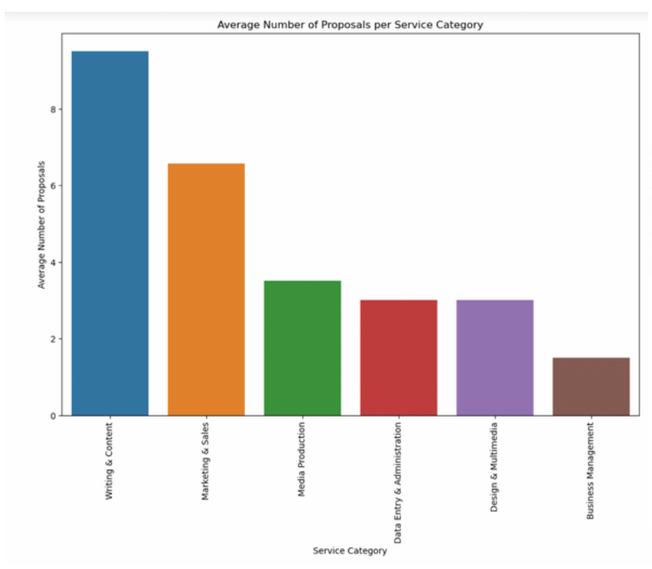
After analyzing our scraped dataset we found that, the budget and experience years are correlated positively, and the average budget across the categories doesn't vary much except for the IT & tech category which had the highest budget average.
We also found that the most popular requested title was 'Graphic designer'.

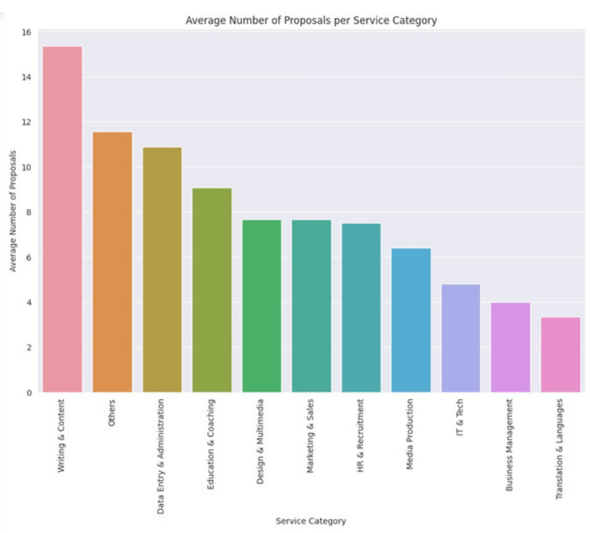
Furthermore, we noticed that after plotting the relationship between the type of company and the budget that the “Social Enterprise: type of company is the highest in the budget among all the rest. While the “individual” type of company is the lowest.

In addition, we also plotted the relationship between category and the number of proposals and found out that "Writing & Content" has the highest number of proposal after that comes "Marketing & Sales" and the category that has the least number of proposals was "Business Management".









Objectives

Our goal of web scraping Freelance yard was to check the Relation between budget and experience level, the relation between budget and category, the most popular job title in all fields, the average budget in each field,
relation between the type of company and the budget, relation between the field and the number of proposals

Data collection

After using the Python's requests library to obtain the site's HTML code.
To scrape the webpage, we used Beautiful Soup, a Python library for processing structured data. Finally, we generated a CSV file using the Pandas library.

Prediction Model and Findings

For classification the label we chose was 'type of company' we used logistic regression and the 5 models that showed the best model was CART with 56% accuracy which may be considered low due to class imbalance

Conclusion

In our project for the Freelance Yard website, we provided valuable insights for freelancers, which may help them in their decision-making process. and through statistical analysis and visualization, we were able to answer specific questions about the data.