EDA project report

Ai Yukino

Contents

Data collection

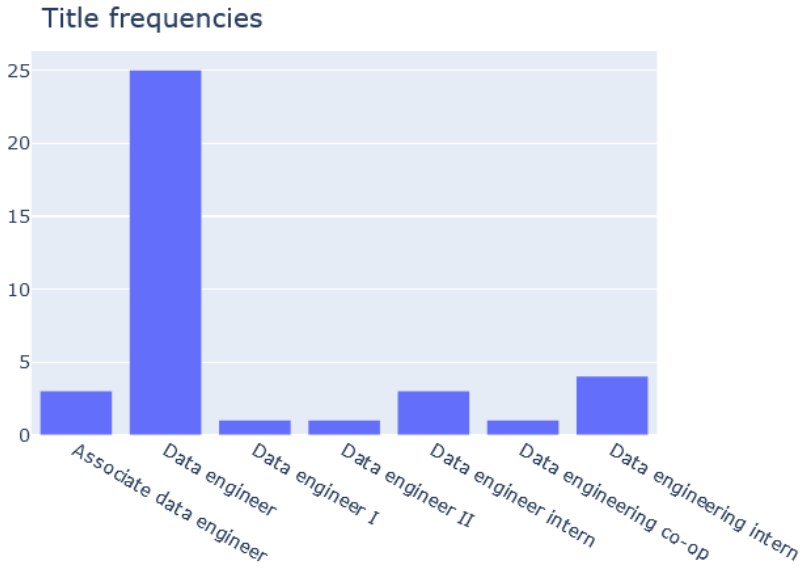
I searched LinkedIn for people who had "data engineer" job titles. I excluded listings that

- did not clarify exact responsibilities/tasks that a person worked on or
- were too senior (e.g. "senior data engineer" or "principal data engineer")

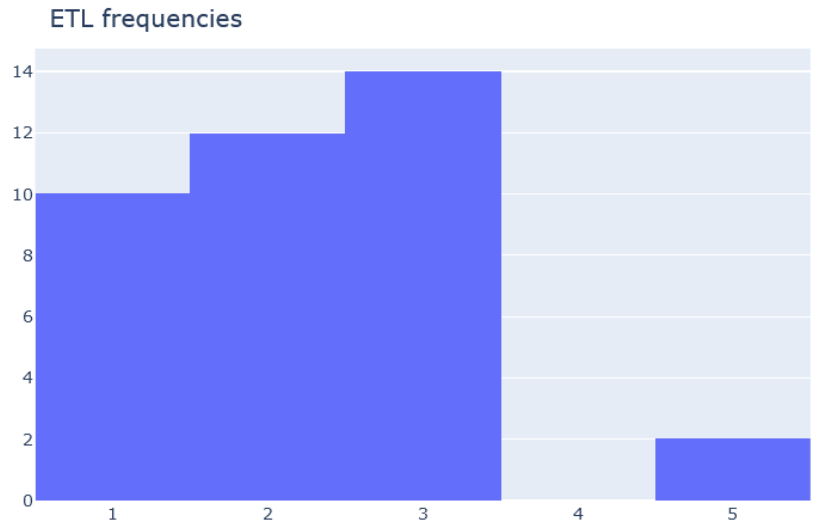
Each row was a role, and each column was either metadata (e.g. an ID, a person's name) or a natural number count of what I subjectively thought of as relevant task categories for a data engineering role. For example, I did not make a column for "analytics" or "statistical inference" as I feel those should be the responsibility of people who formally have the role of data scientist or even data analyst.

Plots

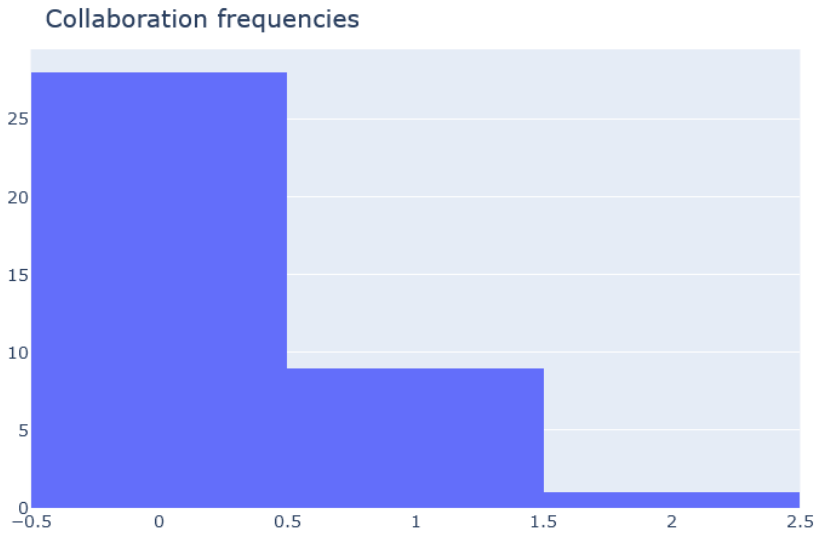
The title bar chart is not particular interesting besides the fact that we see "data engineer" and "data engineer**ing**" are separate titles.



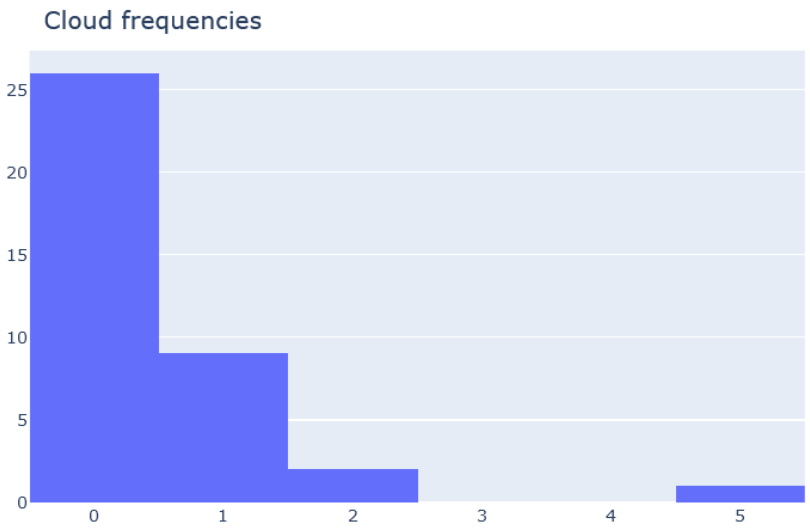
All the single-variable (i.e. column) histograms have single-value bins. The ETL histogram is the only histogram whose first bin is not the 0 bin.



The other histograms are also decreasing instead of increasing. For example, see the collaboration histogram.



The cloud histogram does have a gap like the ETL histogram, though.



Box plots are also possible, but I don't think they would explain much more due to how limited each variable's range is.

Conclusion

Due to the limitations of the data I collected (e.g. small sample size, weak statistical design), I can't make any strong conclusions. However for future work, it might be interesting to subdivide the ETL variable into small divisions, e.g. testing, automation of previous tasks, "standard" SQL queries instead of "optimizing SQL queries". Moreover, I doubt that variables like cloud, DB migration, and APIs are as irrelevant in the job market as implied by their histograms.