Data Analyst – Work Sample

Code review metrics analytics

"Code review (sometimes referred to as peer review) is a software quality assurance activity in which one or several people check a program mainly by viewing and reading parts of its source code, and they do so after implementation or as an interruption of implementation. At least one of the persons must not be the code's author. The persons performing the checking, excluding the author, are called "reviewers" **– Wikipedia**

BlueOptima gathers data from code reviews which in turn can be used to track the engineering team's performance. There are a range of different metrics which can be used to identify the bottlenecks in the code review process Ex: Number of code reviews being created within a span of time, number of comments made, number of revisions happening within a code review etc.,

In this work sample, we have provided Code review related data extracted from Open source software projects.

Task:

- Understand the provided Code review dataset
- Formulate meaningful questions which can be answered by using the dataset
    - Ex:
        - How long does it take to get a code review merged?
        - What percentage of code reviews gets merged? etc.,
- Find out answers for the questions using statistics
- Come up with charts to visualize the data
- Feel free to formulate your own metrics to indicate code review performance/practices

Datasets:

- Link: https://www.dropbox.com/s/ywgdz03dpvs7n11/gerrithub20161121.rar?dl=0 •
Example queries: https://github.com/kin-y/miningReviewRepo/wiki/Query
- Overview: https://kin-y.github.io/miningReviewRepo/
- Schema: https://github.com/kin-y/miningReviewRepo/wiki/Database-Schema •
Publication: https://sdlab-intra.naist.jp/pman3/pman3.cgi?DOWNLOAD=158

Expected Deliverables:

- The solution must be ready for productionisation. Therefore, we would like to have it written in Python. (***Suggestion****: Candidates should think about using any of the following tools/libraries such as Superset, Kibana or Grafana to create a dashboard with input data stream so that the solution can be productionised.)*
- All source code, along with instructions (document/readme file) to reproduce this work sample. Share the intermediate representations of the data (if any).
- Candidates can use some BI tools (ex: Tableau, Power BI) to make a prototype of the dashboard. In that case, please share the analysis workbook made using one of these tools.
- A brief description of the analysis & findings
- What more would you have done if provided with more time

Please note,

Deadline:
You have 1 week to submit your Work Sample.
If you'd like a quick chat to sense-check your approach don't hesitate to ask as many questions as you need. We look forward to discussing your submission.