There are two sets of exercises in this evaluation: The first block tests your ability to handle and manipulate data stored in a relational database. The second block aims to help us understand your analytical strengths.

Although the test was designed without a fixed time limit, completing this exercise should not take you longer than 3 hours.

**Data dictionary**

All data is stored in 2 CSV files which have been provided to you. The data contains the following tables:

|  |  |  |  |
| --- | --- | --- | --- |
| **Table** | **Column** | **Type** | **Content** |
| users | user\_id | int | User ID |
| dob | date | Date of birth |
| country | character | 2 letter country code of user |
| registration\_date | date | Date of registration |
| gender | character | “M” (male) or “F” (female) |
| activity | user\_id | int | User ID |
| date | date | Date of activity |
| revenue | float | Revenue generated by that user on that date |
| days\_since\_reg | int | number of days since registration |

**Output Format**

Please provide a GitHub repository which provides the answers to the questions on the next page. In addition to whether the answers are correct, coding style and readability of the code will also form part of the assessment.

Please send a link to the repository to [henry.gilks@systemiq.earth](mailto:henry.gilks@systemiq.earth) and [shajeeshan.lingeswaran@systemiq.earth](mailto:shajeeshan.lingeswaran@systemiq.earth).

**Question Block A: Data cleaning & merging**

1. Please load the users and activity tables into Python.
   1. How many variables are in the datasets?
   2. How many observations are in the datasets?
2. How many
   1. Male users are in the dataset?
   2. Female users are in the dataset?
   3. For how many users is no gender information available?

If there is no gender information available, please assume that the gender is male (“M”) for all further questions.

1. Merge the two datasets and calculate the days since registration for each activity. What is the total mean and median revenue generated per user (ie across all activities)? How do you interpret the result?
2. Please visualise the distribution of the revenues by user.
3. What is the average *week 1 revenue*, ie the revenue generated by the user in their first week (ie within the first 7 days since registration)?

**Question Block B: Analytics**

1. In this dataset men generated a higher week 1 revenue than women, on average. Is this difference in revenue between men and women statistically significant? What is an appropriate statistical test to determine this and what is it’s p-value?
2. In which country is this difference biggest?
3. Please visualise the relationship between country, gender and week 1 revenue with an appropriate chart.
4. Build a linear regression model to predict the week 1 revenue of a user based on the following variables: Gender, Age, Country and *day 1 revenue* (ie the revenue generated by the user on their registration date). Which of these variables have a statistically significant predictive power according to this model?
5. What revenues do you expect women from France, Germany and the UK to generate, assuming they are all aged 40 and all generated £20 on their registration day?