**Data Science Assignment**

**Introduction**

This assignment has been created with the purpose of giving you an opportunity to demonstrate your strengths and skills. The key factors that we will be observing are your programming, creativity and analytical reasoning.

The assignment is split into two sections. The first section will be an exploratory analysis of the dataset we have provided. The second section will be a more fundamental prediction problem. These sections are not necessarily mutually exclusive, and we hope that you can use some of the insight gained in section 1 to inform your decisions in section 2.

For section 1, you are free to use whichever approach you feel most comfortable with or find most suitable. For example, you could treat this as an exploratory task and provide an annotated notebook; treat it as a customer relations task and provide a visual dashboard; or even provide a program that is tailored towards a technical end-user, such as an object/library that performs functions on this dataset, fully documented. We do ask, however, that what you provide is reproducible on our end.

**Section 1**

**The data**

You will only need to use the data we have provided in the data directory. The dataset we have provided are the public results of Stack Overflow’s developer survey:

• **survey results public.csv** - The survey results from the Stack Overflow Developer Survey. Each row is one respondent’s answers.

• **survey results schema.csv** - A schema providing a brief description of each column of the above dataset.

**Exploratory Data Analysis**

Please provide an exploratory analysis of this dataset. We would usually treat this section as a narrative, providing an explanation for each piece of analysis we produce, with any observation we can glean from this. We expect part of this exploration to be a basis for the next section.

**Basic examples**

• What is the distribution of home countries for each respondent?

• What do respondents rank as the most important aspect of a job when assessing a potential job opportunity?

**Advanced Examples**

Can you visualise any clear partitions in a low-dimensional representation of the data when the target variable is, for example, Gender or Age?

**Section 2**

**Predictive Model**

Finally, we would like you to create a model that can predict the number of years' experience a respondent has had programming.

For this section, you will need to partition the data into a training and testing set.

We would like to see a model that works best on this prediction task, with a clear representation of the training, validation, and testing results they have achieved.

Please provide a clear explanation of your choices and results in the walk through.

**Walkthrough (30 min)**

Lastly, we would like you to walk us through your work. This will be a 30-minute chat on how you approached the problem, the solution you arrived at and any problems you may have encountered. Please provide a walk through and reflection on the analysis you have provided.

We suggest about a 2–3-hour guideline for section 1 and section 2.

Good Luck!