Data Scientist Test

Introduction

So that you can get an idea of the type of work Warwick Analytics does, we have devised a couple of questions that test whether you have the basic skills for the role.

Background

One part of dealing with prospective customers is to perform a proof of concept by solving one of their current problems. This involves taking some data from them (usually in some text format) and running it through our algorithms to get a result. The process usually follows these steps

1. Get problem definition and initial data from customer
2. Review data and confirm that there is the right data and enough to run our algorithms
3. If required, interact with customer to get more data and clarification
4. Prepare data for algorithms
5. Run algorithms
6. Interpret results and repeat steps 4 and 5 if necessary.
7. Report back to prospect.

Data from customers comes from disparate sources and can be incomplete or pre-filtered which can be a disadvantage to us. The purpose of our algorithms is to reduce the search space in which to identify a problem with a product. During the manufacturing process of a product, various tolerances, tests and configuration data will be collected for each product. These are known as variables and there can 100s to 10,000s of these. At some point in its lifetime, a product could fail (this could be during manufacture, during quality control or with the consumer). A failure is a Boolean marker for an individual product. Our algorithm reduces the number of the variables down to a manageable size to help engineers resolve an issue.

The ideal input into our algorithms is

ID	Marker	V1	V2	V3	…	V_n
1	0 or 1	*{value}*	*{value}*	*{value}*	*{value}*	*{value}*
2	0 or 1	*{value}*	*{value}*	*{value}*	*{value}*	*{value}*
…	0 or 1	*{value}*	*{value}*	*{value}*	*{value}*	*{value}*
m	0 or 1	*{value}*	*{value}*	*{value}*	*{value}*	*{value}*

The input data has n variables and m individual products. The ID is unique and could be, for example, the serial number. The data must contain be failed and normal products.

We run this data through our algorithms and the output is a series of regions with a more tightly grouped failures. Example results from our algorithms:

Region	Variable	Min	Max
A	V_x	Min$_{xa}$	Max$_{xa}$
A	V_y	Min$_{ya}$	Max$_{ya}$
A	V_z	Min$_{za}$	Max$_{za}$
B	V_a	Min$_{ab}$	Max$_{ab}$
B	V_b	Min$_{bb}$	Max$_{bb}$
C	V_e	Min$_{ec}$	Max$_{ec}$
C	V_f	Min$_{fc}$	Max$_{fc}$
C	V_g	Min$_{gc}$	Max$_{gc}$

In this case, 3 regions have been identified. A region is a subset of the data bounded by the Min and Max of the variables – it will contain some failures and may contain normals (products that have been identified as not failed but could have done or will do in the future!).

Some regions have more relevance than others and can signify different reasons for the same failure or different types of failure.

Questions

This package contains this file and 2 sub-directories – 1 for each question. The sub-directories contain the data necessary to answer the questions.

Question 1

An automotive manufacture has asked to help identify a vibration problem in the steering wheel on a hatchback. They believe that the problem is due to faulty tyre manufacture.

i. It would be difficult to get any meaningful results from this data, why?

Question 2

This is a bit more of a pure data manipulation/processing exercise. We have provided a sample of some real world data (TrendData.csv) of a value over time.

i. We would like you to come up with **the best solution** of segmenting the data by date based on trends in the Value column (an example input and output has been given).

Please write some form of code/procedure or script to do this, the method should be easily repeatable for new data. The output should be a csv file with the following columns **StartDate, EndDate**, StartValue, EndValue. Start and End Date define the bounds of the segment and we ask you to include the Value at the start and end of the trend for convenience.

ii. List your assumptions and explain why you chose them.

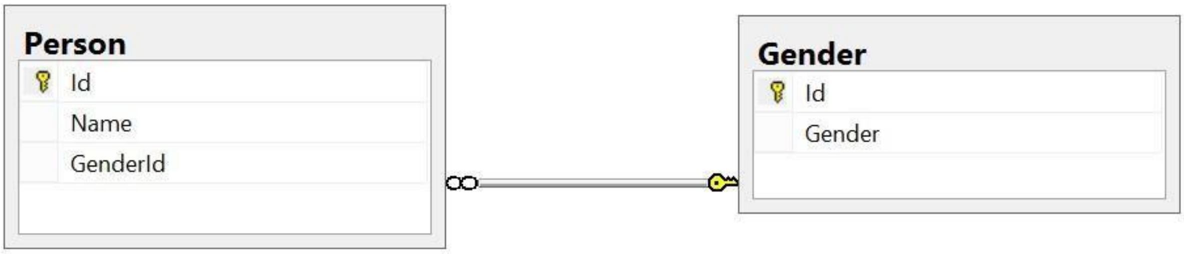
iii. Are there any improvements you would potentially make to your approach / what other information would be helpful in refining/choosing your approach?

Note: Please include your code/script in your submission.

Question 3

We sometimes deal with clients whose data is in an SQL based database. SQL is also a useful tool for data cleaning/manipulation so an understanding is needed.

Given the following data model, write the SQL queries that answer the next questions.



Examples:

i. Who are the people without an assigned gender?

ii. How many people are assigned to each gender? Note: Include non-assigned people in your answer

We are not expecting perfect answers although we would love to get them but you should not need to spend more than a few hours on this.