# Data Scientist Assessment

## Instructions

Assessment instructions:

* Use whatever programming language/tool you are most comfortable with.
* The assessment is designed to get progressively harder. Do not feel you have to answer all the questions if you get stuck.
* Make it clear in your answers if you have made any assumptions about the data/problems.
* Ensure you use the correct dataset for each question.
* All data in this assessment is simulated but under realistic assumptions.

We expect to receive back:

* Written answers to each question giving answer/describing solution (including any relevant graphs/plots).
* Source code.
* If you answer the final question, a file (in the format described in the question) with your predictions.

## Part I

### Data

transactions.tsv

Dataset of in-store customer transactions. Each transaction would consist of a number of items. The transaction value is the total cost of the items in the basket.

Columns:

* transaction.value – Value of the basket in £’s
* gender – The gender of the customer
* store.type – The type of store the transaction occurred in

**Q1) What is the average transaction value?**

**Q2) Is there significant difference between spend in different store types?**

## Part II

### Data

content\_train.tsv

content\_test.tsv

Each row in content\_train.tsv represents a customer.

The variables content\_1 to content\_9 are different **marketing content cards.** 1 means the customer clicked on the card, 0 means the customer viewed the card but didn’t click, NA means the user was never shown the card.

Our aim is to show the most relevant cards to the customer.

The other variables are:

* customer.id - The unique id of the customer
* gender – Gender of the customer
* affluency – A broad categorisation of how affluent the customer is based on their postcode.
* county – The county the customer lives in
* express.no.transactions – How many transactions the customer has performed in Express stores.
* express.total.spend – The total amount the customer has spent in Express stores (in £’s).
* metro.no.transactions – How many transactions the customer has performed in Metro stores.
* metro.total.spend – The total amount the customer has spent in Metro stores (in £’s).
* superstore.no.transactions – How many transactions the customer has performed in Superstores.
* superstore.total.spend – The total amount the customer has spent in Superstores (in £’s).
* extra.no.transactions – How many transactions the customer has performed in Extra stores.
* extra.total.spend – The total amount the customer has spent in Extra stores (in £’s).
* fandf.no.transactions – How many transactions the customer has performed on the online F&F shop
* fandf.total.spend – The total amount the customer has spent in the online F&F shop (in £’s).
* petrol.no.transactions – How many transactions the customer has performed in petrol stations
* petrol.total.spend – The total amount the customer has spent in petrol stations (in £’s)
* direct.no.transactions – How many transactions the customer has performed in Direct online
* direct.total.spend – The total amount the customer has spent in Direct Online (in £’s).

In content\_test.tsv, you do not have whether or not users clicked on the marketing content. This file will be used in Q10.

**Q3) Create a model to predict if a user will click on 1st marketing content (e.g. content\_1). How well does it perform?**

For this question, you should only be using the data in content\_train.tsv.

It is up to you how to measure performance, but you should be prepared to justify why you choose a particular performance measure.

For the customers in content\_test.tsv we have the truth of whether or not they clicked on the marketing content, but this is not provided to you. You can assume that all of these customers saw all of the marketing content items. Customers can click on more than one marketing content item.

To test the performance of your predictions, we will measure log loss. To see how this is calculated see Scoring Function in: <https://www.kaggle.com/c/tradeshift-text-classification/details/evaluation>

The aim is to minimise log loss.

You should be using content\_train.tsv to build your models and predicting over content\_test.tsv.

**Q4) Please give a high-level description of how you would create a production system to deliver the most relevant content card to the user on the website. Bear in mind:**

* **The spend and website interaction data would be delivered (updated) daily**
* **The decision as to what content to show each customer needs to be delivered in real-time (via an API)**
* **The whole process needs to be automated, with no manual interaction**
* **Please state what tools and technology you would use to achieve this**