**Data task**

In the dropbox folder you will find two types of datasets with BBC data:

* verbatim data from a daily survey
* web usage data from our signed-in users

Select the type of data you find more interesting or that you think allows you to best showcase your skills. **There is no specific problem to solve**, we would like to leave it as open as possible but we provide you some context for each dataset and some examples of what you could do with them.

We expect that you will use SQL, Python, R, or any combination of the three to analyse the data.

**Deliverables:**

1. A document covering the following requirement: “Present your findings with visualisations and written insights/observations that you would present to the stakeholders."
2. Any code you have written, ideally as a text file or “notebook”, and commented to explain what each step in the code does.

**Pulse verbatim datasets**

These 2 datasets have been extracted from our audience appreciation survey called Pulse This survey is filled in each day by a panel of respondents. The dataset contains comments from viewers about a show, specifically the answer to the question: “what did you like or did not like about the show”.

Respondents are also asked to rate each episode on a scale of 1 to 10. This rating is called the appreciation index (AI score) so each comment is associated with a rating.

The datasets also include some basic demographics information and the date of the episode.

Datasets:

* Eastenders: Contains all comments from 2017 (28,563) from January to December.
* The apprentice: Contains all comments of 2017 (3,029) from October to December.

Analysis example:

Audience researchers have been reading these comments manually to understand the reaction of the audience to the show. They pay special attention to the AI score to see how the viewers are rating the show. One line of investigation could be to find out if there is any valuable information in the comments that would help identify drivers of AI score. This would give insights to executive producers and scriptwriters or could give them warnings about an upcoming drop in AI and to help understand how the audience is engaging with the programmes.

**Web usage data**

These datasets represent traffic to the BBC’s websites from users during June and July of 2016. It contains information such as BBC product visited (News, Sport, TV & iplayer, Weather, etc…), platform, date/time of the visit, type of app and region. We encourage you to use this data to find insights, patterns of behaviour of our users or use it to build a predictive model.

The small dataset is simply a sample from the large dataset based on 10k users in case you are not able to load the large file. We would expect you to attempt using the large file if possible.

**Data Dictionary:**

|  |  |
| --- | --- |
| **Field** | **Definition** |
| User\_id | User identifier |
| Date\_time | A timestamp to indicate when the event occurred |
| search\_term | If a search term was entered by the user on the BBC website, it will appear here |
| platform | How the content was accessed: Mobile, Computer, Tablet, Big screen |
| app\_type | How the content is delivered to the browser: web, mobile-web or responsive |
| product | Which BBC product that the content is part of |
| Name\_page | The content identifier for the page viewed (e.g. home.page, news.page) |
| Page\_url | The web address of the page visited |
| region | Geographical region where the browser appears to have arrived from |

Analysis example:

Are visitors to the BBC sites only interested in big events? Can we identify patterns of behaviours? For instance you could build a user level segmentation to help the product team better influence user behaviour. Could you use the data to define and predicit future use of a product or lack thereof (ie lapsing).

**In the interview please be prepared to explain your decision making process, any statistical techniques you use, and your reasons for doing so.**