# Section 1: Brief

In this technical test we’ll be assessing the way you handle a task given to you. We would like you to complete the tasks below and submit the files outlined. You’ll only be given 5 days to complete and submit these files, to make the test fair across all candidates, so please make sure you submit these files back to your HR contact before the deadline. The final stage of the interview will consist of you presenting a high level breakdown of the first task and the ideas from the second, back to a few members of the Data & Analytics team. Please assume an audience with mixed technical knowledge from none to advanced.

You should have been sent 2 files for this test; this document and SQL code to run in BigQuery. The code references a sample google analytics data source and creates a target variable to model on. The data represents user sessions on a website down to the hit level. You can find the schema for the data and instructions for running the BigQuery code in the sections below. Should you encounter any issues during this test, then please get in contact.

## Tasks

### Predicting Return Visit to our Website

Being able to identify the users most likely to return is vitial and helps us improve the targeting of customers for marketing. For this task you have been supplied with code to create a modelling dataset. We would like you to run the code given, and extract the data in order to build a model that predicts the likelihood a user will return within 31 days of their latest visit. For fair comparison, please first split the dataset with a random seed of 10 and a train to test ratio of 70:30. Once you have built a model, use it to predict over the Test split.

The end result should be an exported csv file which contains two columns; distinct fullvisitorid from the test dataset and your prediction for that user. Please bear in mind the predictions should be at a fullvisitorid level whilst the data you are using is at a hit level. Our focus for this task will be on your process more than the output, which is why we’d like you to submit a single notebook with your code along with your predictions. This can be written in either Python or R.

Projects like this are usually subject to a governance review before they can be implemented for production use. It’s therefore important to demonstrate the ability to present your project at a high level to the department director and heads. We’d would like you to show: an overview and exploratory analysis of the data, insights on key variables, model overview and results, conclusions and recommendations. Your presentation should be kept to 20 minutes to allow time for questions after.

### Innovative Idea

Produce a slide that highlights one or more ideas for a future project that involves session data similar to this. As a member of the Data Science team it is important to be constantly on the look out for new ideas that will help the business move forward. This idea should relate to the insurance industry.

## Submission

### Notebook

(This should include the code used for your first task and we like you to code to be written in either Python or R.)

### Return Visit Test Predictions

(This csv should include a distinct column for the fullvisitorid and a second column for the prediction.)

### Presentation

(This should include a high level break down for your approach to the first task and the ideas from the second task.)

# Section 2: BigQuery

1. In your browser, go to: <https://console.cloud.google.com/bigquery>
2. Either sign in with a google account or create a new one.
3. If you are new to BigQuery then you may be prompted to create a new Project. (This can be deleted at anytime after you have completed the assessment.)
4. You should be taken to the BigQuery editor under the project you have just created or last selected. From here you can paste the code, run it and save it to google drive (so that you can down load the data).

# Section 3: Schema

## Table 1 – Test and Train Datasets

|  |  |  |
| --- | --- | --- |
| Field Name | Description | Type |
| VisitNumber | The session number for this user. If this is the first session, then value is 1. | Int |
| VisitID | An identifier for this session. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId (ID). | String |
| FullVisitorID | The unique visitor ID (also known as client ID). | String |
| VisitStartTime | VisitStartTimeRaw formatted as UNIX timestamp. | TS |
| ChannelGrouping | The Default Channel Group associated with an end user's session for this View. | String |
| Pageviews | Total number of pageviews within the session. | Int |
| Bounces | Total bounces (for convenience). For a bounced session, the value is 1, otherwise it is null. | Int |
| TimeOnSite | Total time of the session expressed in seconds. | Int |
| DeviceCategory | The type of device (Mobile, Tablet, Desktop). | String |
| HitTime | HitTimeRaw formatted as UNIX timestamp. | TS |
| HitType | The type of hit. One of: "PAGE", "TRANSACTION", "ITEM", "EVENT", "SOCIAL", "APPVIEW", "EXCEPTION". | String |
| PageTitle | The page title. | String |
| TransactionID | The transaction ID of the ecommerce transaction. | String |
| TransactionRevenue | Total transaction revenue. | Float |
| ProductSKU | Product Code. | String |
| V2ProductName | Product Name. | String |
| V2ProductCategory | Product Category. | String |
| ProductRevenue | The revenue of the product. | Float |
| ProductQuantity | The quantity of the product purchased. | Int |
| ReturnVisit | Target variable (only on the Train dataset). | Int |