**Data Science for Industry Project 1**

**Due date: 5September 2017**

The aim of this project is to build a recommender system for Getwine, an online wine retailer (http://getwine.co.za/). The function of the recommender system is to recommend to a user wines that, on the basis of his or her past purchases, they might like to buy in the future.

The data for the project consists of all purchase orders made by 1000 randomly selected Getwine users in 2017. A purchase order is a single “delivery” of wine to a user. A single order can consist of any number and selection of bottles of wine, but usually consists of 12 or more bottles.

Please start by signing the attached confidentiality agreement on the next page of this document and emailing the signed copy to [ian.durbach@uct.ac.za](mailto:ian.durbach@uct.ac.za) or DM it to me on Slack. Customer data is valuable and it was nice of Getwine to share theirs with us, so please respect the agreement. Please remember not to push the data to GitHub (add to .gitignore if need be)!

The data has been provided in the following two files:

**orders.csv**: each row in this file constitutes a single customer order. A single order may consist of any number of bottles of wine. The order is identified by a unique key *orders\_id* that can be used to identify the wines that were purchased in the accompanying file **order-items.csv**. Other variables are:

* Customer’s gender
* Customer’s date of birth
* Customer’s location
* Date of purchase
* Payment method
* Monetary value of order

**order-items.csv**: contains the wines bought in each order, together with the quantity bought and the price per bottle.

Please note that the data is provided to you “as is” from the Getwine database. It needs to be checked and cleaned (see in particular the *products\_name* variable in **order-items.csv.**

Write up your work in the form of a report **written in R Markdown**. The report should contain a description of the problem, the approach you took, and your results. Your code should be integrated into the document, and this code should be clearly described and commented (see the class .Rmd notebooks for examples).

Anyone else should be able to run the code in the R markdown document to completion – if necessary use *set.seed()* to set a random seed so that your final results don't change.

The submission deadline is on or before 5 September 2017. Details on how/where to submit to follow.

**CONFIDENTIALITY AGREEMENT FOR DATA PROVIDED FOR DATA SCIENCE FOR INDUSTRY ASSIGNMENT 1**

In exchange for the release by Getwine of the data attached to Assignment 1 of the Data Science for Industry module, I hereby agree to the following conditions:

1. I will only use these data for the purposes of the assignment.

2. I will not release, furnish, disclose, publish or otherwise disseminate these data without the written consent of a representative of Getwine.

3. I understand that Getwine shall retain all ownership rights to the data. I shall have no right, title, or interest to any of the data. I agree to reference Getwine as the source of the data in all reports, publications, tables, graphs, or other products produced from the data.

4. Until such time as I have either destroyed or returned to Getwine the data subject to this Agreement, I will maintain in effect the policies and procedures on confidentiality of information described above.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
(Signature)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
Date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
(Print Name)