2_2 Data description report

Alexander J Malt

28/11/2019

Data description report

This report contains a description of (i) the data sets provided and (ii) a description of the structure of each type of data.

N.B. This report does not contain a description of data sources - this data was provided directly by the University rather than being accessed via/extracted from a system (see 2_1 Initial data collection report).

User activity data (activityData)

The User activity datasets consist of seven files which - I assume - contain logs when users begin and complete steps on the course. The files are entitled:

- cyber-security-1_step-activity.csv
- cyber-security-2_step-activity.csv
- cyber-security-3_step-activity.csv
- cyber-security-4_step-activity.csv
- cyber-security-5_step-activity.csv
- cyber-security-6_step-activity.csv
- cyber-security-7_step-activity.csv

Each file contained the following fields:

- *learner_id*, containing a hash uniquely identifing an individual student. I assume that each hash/learner_id is unique to the individual student but that the hash will be common to all datasets (permitting merging/joining of the data frames).
- *step*, containing a value for the 'step' of the course, e.g. '1.2'. I have assumed that this value is essentially a concatenation of the next two fields: that the number to the left of the decimal refers to the week of the course and the number to the right of the decimal refers to the step number.
- *week_number*, giving the week on which the user's activity is logged.
- *step_number*, giving the step of the course which the user's activity relates to.
- *first_visited_at*, giving a date and time. I assume this timestamp denotes when the user first accessed this step of the course.
- *last_completed_at*, giving a date and time. I assume this timestamp denotes when the user completed this step of the course.

Because these columns are common to all files in the dataset, I added a new *courseID* column to each dataset - to denote which run of the course the data refers to - and then merged the data from each individual file into a single data frame (**activityData**) using `rbind()`. Once merged, I added two additional fields derived from *first_visited_at* and *last_completed_at*, containing only the date values (ommitting time) and formated as dates rather than factors (anticipating possible time series analysis).

The activityData data frame is therefore structured as follows:

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    423072 obs. of  7 variables:
##  $ learner_id       : Factor w/ 20285 levels "000a49d0-39c8-4cef-848c-a37d3574d179",..: 2942 4730 405
##  $ step             : Factor w/ 63 levels "1.1","1.10","1.11",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ week_number      : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ step_number      : Factor w/ 23 levels "1","10","11",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ first_visited_at : POSIXct, format: "2016-08-02 13:45:37" "2016-08-02 15:40:48" ...
##  $ last_completed_at: POSIXct, format: NA NA ...
##  $ courseID         : Factor w/ 7 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
```

User personality archetypes data

The User personality archetypes data consists of seven files, however the files for the first two years of the course were empty, i.e. contained no data beyond column headings:

- cyber-security-1_archetype-survey-responses.csv - *empty*
- cyber-security-2_archetype-survey-responses.csv - *empty*
- cyber-security-3_archetype-survey-responses.csv
- cyber-security-4_archetype-survey-responses.csv
- cyber-security-5_archetype-survey-responses.csv
- cyber-security-6_archetype-survey-responses.csv
- cyber-security-7_archetype-survey-responses.csv

Each file contained the following fields:

- *id*, containing a numerical figure identifying each row of data in the archetype survey responses.
- *learner_id*, containing a hash uniquely identifing an individual student. I assume that each hash/learner_id is unique to the individual student but that the hash will be common to all datasets (permitting merging/joining of the data frames).
- *responded_at*, giving a date and time. I assume this timestamp denotes when a behavioural archetype was assigned to the user (following the user's answering the relevant questions).
- *archetype*, categorising each student in the dataset as one of a set of behavioural archetypes. The list of behavioural archetypes and their respective meanings are viewable on the Future Learn website.

Because these columns are common to all files in the dataset, I added a new *courseID* column to each dataset - to denote which run of the course the data refers to - and then merged the data from each individual file into a single data frame (**archetypeData**) using `rbind()`.

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1074 obs. of  3 variables:
##  $ learner_id: Factor w/ 1059 levels "0100d15a-f66a-4615-91aa-efe5e2ad9ece",..: 2 29 11 33 47 16 18
##  $ archetype : Factor w/ 8 levels "Advancers","Explorers",..: 3 3 7 2 1 7 7 7 7 2 ...
##  $ courseID  : Factor w/ 5 levels "3","4","5","6",..: 1 1 1 1 1 1 1 1 1 1 ...
```

User enrollment data

The User enrollment data consists of seven files:

- cyber-security-1_enrolments.csv
- cyber-security-2_enrolments.csv
- cyber-security-3_enrolments.csv
- cyber-security-4_enrolments.csv
- cyber-security-5_enrolments.csv
- cyber-security-6_enrolments.csv
- cyber-security-7_enrolments.csv

These files contain the following fields:

- *learner_id*, containing a hash uniquely identifing an individual student. I assume that each hash/learner_id is unique to the individual student but that the hash will be common to all datasets (permitting merging/joining of the data frames).
- *enrolled_at*, giving a date and time. I assume this timestamp denotes when the user enrolled as a student on the course.
- *unenrolled_at*, giving a date and time. I assume this timestamp denotes when a user ceased to be a student on the course, i.e. withdrew from study.

- *role*, denoting whether the user is a student (in which case they are assigned the value ‘learner’) or an organisational administrator (in which case they are assiged the value ‘organisation_admin’).

```
## [1] "learner"            "organisation_admin"
```

- *fully_participated_at*, giving a date and time. I assume this timestamp denotes when the student fully completed all steps of the course.
- *purchased_statement_at*, giving a date and time. I assume this timestamp denotes when the student purchased the course content.
- *gender*, denoting the gender of the student with one of the following values:

```
## [1] "female"    "male"      "nonbinary" "other"
```

- *country*, giving a two letter country code for each row (or ‘Unknown’ if the value is unknown). I assume this information originated from students on the course, and is therefore differentiated from *detected_country* (see below). The possible values in the data are:

```
##   [1] "AD" "AE" "AF" "AL" "AM" "AO" "AR" "AT" "AU" "AZ" "BA" "BB" "BD" "BE" "BG"
##  [16] "BH" "BN" "BO" "BR" "BS" "BW" "BY" "CA" "CD" "CG" "CH" "CI" "CL" "CM" "CN"
##  [31] "CO" "CR" "CW" "CY" "CZ" "DE" "DK" "DO" "DZ" "EC" "EE" "EG" "ES" "ET" "FI"
##  [46] "FJ" "FR" "GA" "GB" "GE" "GG" "GH" "GI" "GM" "GN" "GR" "GY" "HK" "HN" "HR"
##  [61] "HT" "HU" "ID" "IE" "IL" "IN" "IQ" "IR" "IT" "JE" "JM" "JO" "JP" "KE" "KH"
##  [76] "KR" "KW" "KZ" "LB" "LC" "LK" "LR" "LS" "LT" "LV" "LY" "MA" "MD" "ME" "MG"
##  [91] "MK" "MM" "MN" "MT" "MU" "MW" "MX" "MY" "MZ" "NE" "NG" "NL" "NO" "NP" "NZ"
## [106] "OM" "PE" "PG" "PH" "PK" "PL" "PR" "PS" "PT" "QA" "RO" "RS" "RU" "RW" "SA"
## [121] "SD" "SE" "SG" "SK" "SL" "SN" "SO" "SS" "SV" "SY" "SZ" "TG" "TH" "TJ" "TL"
## [136] "TN" "TR" "TT" "TW" "TZ" "UA" "UG" "US" "UY" "UZ" "VE" "VN" "WF" "XK" "YE"
## [151] "ZA" "ZM" "ZW"
```

- *age_range*, giving - for each row in the data - an age range within which the user’s age falls (or ‘Unknown’ if the value is unknown). The following ranges are in the data:

```
## [1] "<18"   ">65"   "18-25" "26-35" "36-45" "46-55" "56-65"
```

- *highest_education_level*, giving - for each row in the data - the highest educaton level attained by the user at time of taking the course (or ‘Unknown’ if the value is unknown). I assume this has been inputted by the user. The following levels are in the data:

```
## [1] "apprenticeship"       "less_than_secondary"  "professional"
## [4] "secondary"            "tertiary"             "university_degree"
## [7] "university_doctorate" "university_masters"
```

- *employment_status*, giving - for each row in the data - the user’s employment status at time of taking the course (or ‘Unknown’ if the value is unknown). I assume this has been inputted by the user. The following levels are in the data:

```
## [1] "full_time_student" "looking_for_work"  "not_working"
## [4] "retired"           "self_employed"     "unemployed"
## [7] "working_full_time" "working_part_time"
```

- *employment_area*, giving - for each row in the data - the sector in which the user was employed at time of taking the course (or ‘Unknown’ if the value is unknown). I assume this has been inputted by the user. The following levels are in the data:

```
##  [1] "accountancy_banking_and_finance"     "armed_forces_and_emergency_services"
##  [3] "business_consulting_and_management"  "charities_and_voluntary_work"
##  [5] "creative_arts_and_culture"           "energy_and_utilities"
##  [7] "engineering_and_manufacturing"       "environment_and_agriculture"
##  [9] "health_and_social_care"              "hospitality_tourism_and_sport"
## [11] "it_and_information_services"         "law"
## [13] "marketing_advertising_and_pr"        "media_and_publishing"
## [15] "property_and_construction"           "public_sector"
## [17] "recruitment_and_pr"                  "retail_and_sales"
## [19] "science_and_pharmaceuticals"         "teaching_and_education"
## [21] "transport_and_logistics"
```

- *detected_country*, giving (I assume) the country the system detects the user in when taking the course (or “–” for unknown).

```
##   [1] "AD" "AE" "AF" "AG" "AI" "AL" "AM" "AO" "AR" "AS" "AT" "AU" "AW" "AZ" "BA"
##  [16] "BB" "BD" "BE" "BF" "BG" "BH" "BI" "BJ" "BM" "BN" "BO" "BR" "BS" "BT" "BW"
##  [31] "BY" "BZ" "CA" "CD" "CG" "CH" "CI" "CL" "CM" "CN" "CO" "CR" "CU" "CV" "CW"
##  [46] "CY" "CZ" "DE" "DJ" "DK" "DM" "DO" "DZ" "EC" "EE" "EG" "ER" "ES" "ET" "FI"
##  [61] "FJ" "FR" "GA" "GB" "GD" "GE" "GG" "GH" "GI" "GM" "GN" "GR" "GT" "GU" "GY"
##  [76] "HK" "HN" "HR" "HT" "HU" "ID" "IE" "IL" "IM" "IN" "IQ" "IR" "IS" "IT" "JE"
##  [91] "JM" "JO" "JP" "KE" "KG" "KH" "KI" "KN" "KR" "KW" "KY" "KZ" "LA" "LB" "LC"
## [106] "LK" "LR" "LS" "LT" "LU" "LV" "LY" "MA" "MC" "MD" "ME" "MG" "MK" "ML" "MM"
## [121] "MN" "MO" "MQ" "MT" "MU" "MV" "MW" "MX" "MY" "MZ" "NE" "NG" "NI" "NL" "NO"
## [136] "NP" "NR" "NZ" "OM" "PA" "PE" "PG" "PH" "PK" "PL" "PR" "PS" "PT" "PY" "QA"
## [151] "RE" "RO" "RS" "RU" "RW" "SA" "SB" "SC" "SD" "SE" "SG" "SI" "SK" "SL" "SN"
## [166] "SO" "SR" "SS" "SV" "SX" "SY" "SZ" "TC" "TD" "TG" "TH" "TJ" "TL" "TN" "TR"
## [181] "TT" "TW" "TZ" "UA" "UG" "US" "UY" "UZ" "VC" "VE" "VG" "VN" "VU" "WS" "XK"
## [196] "YE" "ZA" "ZM" "ZW"
```

Merged with archetype data.