data-preperation

October 7, 2021

1 Data preperation

[32]:
```
import pandas
from IPython.display import display

def read_data(file_name):
    csv_file = pandas.read_csv(f'../data/{file_name}.csv')
    return csv_file

student_course_identifier = ["code_module", "code_presentation", "id_student"]

student_info = read_data('studentInfo')
student_info.loc[0]
```

[32]:
```
code_module                          AAA
code_presentation                  2013J
id_student                         11391
gender                                 M
region                East Anglian Region
highest_education        HE Qualification
imd_band                         90-100%
age_band                            55<=
num_of_prev_attempts                   0
studied_credits                      240
disability                             N
final_result                        Pass
Name: 0, dtype: object
```

1.1 Merge 1

merge: studentinfo
with: studentRegistration

[33]:
```
# merge studentinfo
# with studentRegistration
student_registration = read_data("studentRegistration")
students_merged_step_1 = pandas.merge(student_info, student_registration,␣
 ↪on=student_course_identifier)
students_merged_step_1.loc[0]
```

[33]:
```
code_module                          AAA
code_presentation                  2013J
id_student                         11391
gender                                 M
region                East Anglian Region
highest_education        HE Qualification
imd_band                         90-100%
age_band                           55<=
num_of_prev_attempts                  0
studied_credits                     240
disability                            N
final_result                       Pass
date_registration                -159.0
date_unregistration                 NaN
Name: 0, dtype: object
```

1.2 Merge 2

merge: studentinfo and studentRegistration
with: courses

[34]:
```
courses = read_data("courses")
students_merged_step_2 = pandas.merge(students_merged_step_1, courses,␣
↪on=['code_module', 'code_presentation'])
students_merged_step_2.loc[0]
```

[34]:
```
code_module                              AAA
code_presentation                      2013J
id_student                             11391
gender                                     M
region                    East Anglian Region
highest_education            HE Qualification
imd_band                             90-100%
age_band                               55<=
num_of_prev_attempts                      0
studied_credits                         240
disability                                N
final_result                           Pass
date_registration                    -159.0
date_unregistration                     NaN
module_presentation_length              268
Name: 0, dtype: object
```

1.3 Merge 3

merge: studentinfo, studentRegistration and courses
with: vles and studentvles

[35]:
```
# prepare vles for merging later
vles = pandas.merge(read_data("vle"), read_data("studentVle"),␣
 ↪on=['code_module', 'code_presentation', 'id_site'])

# group vle clicks per day
grouped_vles_per_day = vles.groupby(["code_module", "code_presentation",␣
 ↪"id_student", "id_site", "date", "activity_type"]).agg({
    "sum_click": "sum"
}).reset_index()

# combine vle data as a single column value
grouped_vles_per_day["vles"] = grouped_vles_per_day[["id_site", "date",␣
 ↪"activity_type", "sum_click"]].values.tolist()
grouped_vles_per_day.head()

# combine all seperate from rows to a single row with a list
grouped_vles_per_student = grouped_vles_per_day.
 ↪groupby(student_course_identifier, as_index=False).agg({
    "vles": lambda x: list(x)
})
```

[36]:
```
students_merged_step_3 = pandas.merge(grouped_vles_per_student,␣
 ↪students_merged_step_2, on=student_course_identifier)
students_merged_step_3.loc[0]
```

[36]:
```
code_module                                                          AAA
code_presentation                                                  2013J
id_student                                                         11391
vles                    [[546614, -5, homepage, 7], [546614, 0, homepa…
gender                                                               M
region                                                East Anglian Region
highest_education                                        HE Qualification
imd_band                                                        90-100%
age_band                                                          55<=
num_of_prev_attempts                                                 0
studied_credits                                                    240
disability                                                           N
final_result                                                      Pass
date_registration                                                -159.0
date_unregistration                                                NaN
module_presentation_length                                         268
Name: 0, dtype: object
```