Merging and Wrangling Process

#### Executing the following scripts below to achieve the FinalData for our model testing

#### [SQL script of studentVleFeature](#_SQL_Script:)

1. [SQL script of studentCourseRegistrationFULLSTG](#_SQL_Script:_1)
2. SQL script of studentVLE
3. Score\_wranging.ipynb :

<https://github.com/georgetown-analytics/University-Learning-Analytics/blob/master/code/wrangling/Score_wragling.ipynb>

1. [SQL script of analysisFeatures](#_SQL_Script:_2)
2. Data\_wrangling\_2 :

<https://github.com/georgetown-analytics/University-Learning-Analytics/blob/master/code/wrangling/Data_wrangling_2.ipynb>

1. One\_Hot\_Coding\_Regions\_and\_Final\_Wrangling.ipynb :

<https://github.com/georgetown-analytics/University-Learning-Analytics/blob/master/code/wrangling/One_Hot_Encoding_Regions_and_Final_Wrangling.ipynb>

1. **Finalize Data for Features Dataset.ipynb**

<https://github.com/georgetown-analytics/University-Learning-Analytics/blob/master/code/wrangling/Finalize%20Data%20for%20Features%20Dataset.ipynb>

#### TABLE: studentVleFeature

Description: Student interaction with materials

Data Sources: studentVleFULLSTG, courseSTG

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| id\_student | INTEGER | PK, a unique identification number for the student |
| code\_module | CHARACTER VAR | PK, the identification code of the module |
| code\_presentation | CHARACTER VAR | PK, the identification code of the module presentation |
| b4\_sum\_clicks | BIGINT | the number of times a student interacts with the material before the start of the module-presentation. |
| qtr\_sum\_clicks | BIGINT | the number of times in the first quarter that student interacts with the material |
| half\_sum\_clicks | BIGINT | the number of times a student interacts with the material halfway through the module. |
| threeqtr\_sum\_clicks | BIGINT | 75% of the number of times of the student interacts with the material |
| qtr\_half\_sum\_clicks | BIGINT | the number of times in the second quarter that student interacts with the material |
| half\_threeqrt\_sum\_clicks | BIGINT | the number of times in the third quarter that student interacts with the material |
| twothrd\_sum\_clicks | BIGINT | 66.66% of the number of times a student interacts with the material |
| thrd\_twothrd\_sum\_clicks | BIGINT | the number of times between 33.33%- 66.66% that a student interacts with the material |
| all\_clicks | BIGINT | the number of times a student interacts with the material in that day. |

#### SQL Script:

|  |  |
| --- | --- |
| CREATE TABLE public."studentVleFeatures" | |
|  | AS ( |
|  | Select vle.id\_student, vle.code\_module, vle.code\_presentation, |
|  | sum(CASE |
|  | WHEN date\_iact < 0 THEN sum\_click |
|  | ELSE 0 |
|  | END) as b4\_sum\_clicks, |
|  | sum(sum\_click) as allclicks, |
|  | sum(CASE |
|  | WHEN date\_iact between 0 and module\_presentation\_length/4 THEN sum\_click |
|  | ELSE 0 |
|  | END) as qtr\_sum\_clicks, |
|  | sum(CASE |
|  | WHEN date\_iact between 0 and module\_presentation\_length/2 THEN sum\_click |
|  | ELSE 0 |
|  | END) as half\_sum\_clicks, |
|  | sum(CASE |
|  | WHEN date\_iact between 0 and module\_presentation\_length\*3/4 THEN sum\_click |
|  | ELSE 0 |
|  | END) as threeqtr\_sum\_clicks, |
|  | sum(CASE |
|  | WHEN date\_iact between module\_presentation\_length/4 and module\_presentation\_length/2 THEN sum\_click |
|  | ELSE 0 |
|  | END) as qtr\_half\_sum\_clicks, |
|  | sum(CASE |
|  | WHEN date\_iact between module\_presentation\_length/2+1 and module\_presentation\_length\*3/4 THEN sum\_click |
|  | ELSE 0 |
|  | END) as half\_threeqtr\_sum\_clicks, |
|  | sum(CASE |
|  | WHEN date\_iact between 0 and module\_presentation\_length/3 THEN sum\_click |
|  | ELSE 0 |
|  | END) as thrd\_sum\_clicks, |
|  | sum(CASE |
|  | WHEN date\_iact between 0 and module\_presentation\_length\*2/3 THEN sum\_click |
|  | ELSE 0 |
|  | END) as twothrd\_sum\_clicks, |
|  | sum(CASE |
|  | WHEN date\_iact between module\_presentation\_length/3+1 and module\_presentation\_length/2 THEN sum\_click |
|  | ELSE 0 |
|  | END) as thrd\_twothrd\_sum\_clicks |
|  | from public."studentVleFULLSTG" as vle, public."coursesSTG" as crse |
|  | where vle.code\_module = crse.code\_module and vle.code\_presentation = crse.code\_presentation |
|  | group by vle.id\_student, vle.code\_module, vle.code\_presentation |
|  | order by vle.id\_student, vle.code\_module, vle.code\_presentation |
|  | ); |

#### TABLE: studentCourseRegistrationFULLSTG

Description: Student’s demographics and registrations

Data Sources: studentInfoSTG, courseSTG

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| id\_student | INTEGER | PK, a unique identification number for the student |
| code\_module | CHARACTER VAR | PK, identification code of the module |
| code\_presentation | CHARACTER VAR | PK, the identification code of the module presentation |
| module\_presentation\_legth | INTEGER | length of the module-presentation in days. |
| gender | CHARACTER VAR | the student’s gender |
| region | CHARACTER VAR | identifies the geographic region, where the student lived while taking the module-presentation. |
| highest\_education | CHARACTER VAR | highest student education level on entry to the module presentation. |
| imd\_band | CHARACTER VAR | specifies the [Index of Multiple Depravation](https://en.wikipedia.org/wiki/Multiple_deprivation_index) band of the place where the student lived during the module-presentation. |
| age\_band | CHARACTER VAR | band of the student’s age. |
| num\_of\_prev\_attempts | INTEGER | the number times the student has attempted this module. |
| studied\_credits | INTEGER | the total number of credits for the modules the student is currently studying. |
| disability | CHARACTER VAR | indicates whether the student has declared a disability. |
| fianl\_result | CHARACTER VAR | student’s final result in the module-presentation. |
| date\_registration | INTEGER | the date of student’s registration on the module presentation, this is the number of days measured relative to the start of the module-presentation |
| date\_unregistration | INTEGER | date of student unregistration from the module presentation, this is the number of days measured relative to the start of the module-presentation. |
| module\_domain | CHARACTER VAR | Goupr of module, Social Science courses are defined as AAA, BBB, and GGG. STEM courses are defined as CCC, DDD, EEE, FFF |
| term | TEXT | code name of term. “B” for the presentation starting in February and “J” for the presentation starting in October. |
| year | TEXT | the year of module-presentation |

#### 

#### SQL Script:

|  |  |
| --- | --- |
| CREATE TABLE public."studentCourseRegistrationFeatures" | |
|  | AS ( |
|  | SELECT |
|  | crse.module\_presentation\_length, |
|  | SUBSTRING(crse.code\_presentation, 1, 4) as YEAR, |
|  | SUBSTRING(crse.code\_presentation, 5, 1) as TERM, |
|  | stdt.\*, |
|  | stdtreg.date\_registration, stdtreg.date\_unregistration, |
|  | CASE |
|  | WHEN stdt.code\_module in ('AAA','BBB','GGG') THEN 'SocialScience' |
|  | WHEN stdt.code\_module in ('CCC','DDD','EEE','FFF') THEN 'STEM' |
|  | ELSE NULL |
|  | END AS module\_domain |
|  | FROM |
|  | public."studentInfoSTG" as stdt, |
|  | public."coursesSTG" as crse, |
|  | public."studentRegistrationSTG" as stdtreg |
|  | WHERE |
|  | stdt.id\_student = stdtreg.id\_student and stdt.code\_presentation = stdtreg.code\_presentation and stdt.code\_module = stdtreg.code\_module |
|  | AND crse.code\_presentation = stdtreg.code\_presentation and crse.code\_module = stdtreg.code\_module |
|  | ORDER BY stdtreg.id\_student, stdtreg.code\_module, stdtreg.code\_presentation |
|  | ); |

#### 

#### TABLE: studentAssessmentFullSTG

Description: Student Assessments scores

Data Sources: studentAssessmentSTG, courseSTG

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| id\_assessment | INTEGER | PK, a unique identification number of assessments |
| assessment\_type | CHARACTER VAR | type of assessment. Three types of assessments exis: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam). |
| id\_student | INTEGER | FK, a unique identification number for the student |
| code\_module | CHARACTER VAR | FK, the identification code of the module |
| code\_presentation | CHARACTER VAR | FK, the identification code of the module presentation |
| date\_submitted | INTEGER | the date of student submission, measured as the number of days since the start of the module presentation. |
| final\_sub\_date | INTEGER | information about the final submission date of the assessment calculated as the number of days since the start of the module-presentation. |
| half\_weight | NUMERIC | The number of module assessments weights halfway of module-presentation length |
| half\_module\_length | INTEGER | halfway of module-presentation length |
| is\_banked | INTEGER | a status flag indicating that the assessment result has been transferred from a previous presentation. |
| score | INTEGER | the student’s score in this assessment. The range is from 0 to 100. |
| weight | NUMERIC | weight of the assessment in %. Typically, Exams are treated separately and have the weight 100%; the sum of all other assessments is 100%. |

SQL Script: ??????

#### 

### FEATURE: std\_half\_score

Description: the score% of student assessments before halfway of the module-presentation length

Source: studentAssessmentFULLSTG

Python Script: <https://github.com/georgetown-analytics/University-Learning-Analytics/blob/master/code/wrangling/Score_wragling.ipynb>

|  |
| --- |
| studentDf = df.copy()  studentDf['stdn\_half\_weight\_score'] = 0  studentDf['score'] = studentDf['score'].fillna(0)  *# Assessments score percentage*  *studentDf['assmt\_score'] = studentDf['score']\*studentDf['weight']/100*  *# the number of student assessment scores group by student, module, term and year*  *# groupby() as\_index = False to keep column*  *studentScoreDf = studentDf.loc[studentDf['final\_sub\_date'] < studentDf['hlf\_module\_length'] ].groupby(['id\_student','code\_module','code\_presentation','hlf\_weight'], as\_index = False)['assmt\_score'].sum()*  *# the number of student assessment scores on half of module-presentation weights*  *studentScoreDf['stdn\_half\_score'] = studentScoreDf['assmt\_score']/studentScoreDf['hlf\_weight']*  *# write dataframe to database*  *from sqlalchemy import create\_engine*  *engine=create\_engine('postgresql://postgres:Georgetown16@database-1.c5vispb5ezxg.us-east-1.rds.amazonaws.com:5432/Dataset')*  *studentScoreDf.to\_sql('studentAssessmentHalfScore', engine, if\_exists='replace')* |

# TABLE: AnalysisFeatures

Description: Student Info, activities and performance

Data Sources: studentVleFullSTG, studentAssessmentFullSTG, studentCourseRegistrationFullSTG

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| id\_student | INTEGER | PK, a unique identification number for the student |
| code\_module | CHARACTER VAR | PK, the identification code of the module |
| code\_presentation | CHARACTER VAR | PK, the identification code of the module presentation |
| module\_domain | TEXT | Group of the module, Social Science courses are defined as AAA, BBB, and GGG. STEM courses are defined as CCC, DDD, EEE, FFF |
| module\_presentation\_length | INTEGER | length of the module-presentation in days. |
| term | TEXT | code name of the term. “B” for the presentation starting in February and “J” for the presentation starting in October. |
| year | TEXT | the year of module-presentation |
| final\_result | CHARACTER VAR | student’s final result in the module-presentation |
| gender | CHARACTER VAR | the student’s gender |
| region | CHARACTER VAR | identifies the geographic region, where the student lived while taking the module-presentation. |
| highest\_education | CHARACTER VAR | highest student education level on entry to the module presentation. |
| imd\_band | CHARACTER VAR | specifies the [Index of Multiple Depravation](https://en.wikipedia.org/wiki/Multiple_deprivation_index) band of the place where the student lived during the module-presentation. |
| age\_band | CHARACTER VAR | band of the student’s age. |
| num\_of\_prev\_attempts | INTEGER | the number times the student has attempted this module. |
| disability | CHARACTER VAR | indicates whether the student has declared a disability. |
| date\_registration | INTEGER | the date of student’s registration on the module presentation, this is the number of days measured relative to the start of the module-presentation |
| date\_unregistration | INTEGER | date of student unregistration from the module presentation, this is the number of days measured relative to the start of the module-presentation. |
| b4\_sum\_clicks | BIGINT | the number of times a student interacts with the material before the start of the module-presentation. |
| qtr\_sum\_clicks | BIGINT | the number of times in the first quarter that student interacts with the material |
| half\_sum\_clicks | BIGINT | the number of times a student interacts with the material halfway though the module. |
| threeqtr\_sum\_clicks | BIGINT | 75% of the number of times of the ta student interacts with the material |
| qtr\_half\_sum\_clicks | BIGINT | the number of times in the second quarter that student interacts with the material |
| half\_threeqrt\_sum\_clicks | BIGINT | the number of times in the third quarter that student interacts with the material |
| twothrd\_sum\_clicks | BIGINT | 66.66% of the number of times a student interacts with the material |
| thrd\_twothrd\_sum\_clicks | BIGINT | the number of times between 33.33%- 66.66% that a student interacts with the material |
| all\_clicks | BIGINT | the number of times a student interacts with the material on that day. |
| std\_half\_score | DOUBLE PRECISION | the score% of student assessments before halfway of the module-presentation length |
| std\_total\_weight | NUMERIC | the number of weight of student assessments before halfway of the module-presentation length |

# SQL Script:

|  |  |
| --- | --- |
| Create table public."analysisFeatures" | |
|  | as |
|  | (select stdtreg.id\_student, stdtreg.code\_module, stdtreg.code\_presentation, stdtreg.module\_domain, stdtreg.module\_presentation\_length, |
|  | stdtreg.term, stdtreg.year, stdtreg.num\_of\_prev\_attempts, |
|  | stdtreg.final\_result, stdtreg.pass\_fail\_ind, |
|  | stdtreg.reg\_period, stdtreg.date\_registration, stdtreg.date\_unregistration, |
|  | stdtreg.disability, stdtreg.gender, stdtreg.age\_band, stdtreg.region, stdtreg.highest\_education, |
|  | stdtreg.imd\_band, stdtreg.studied\_credits,asmtVle.qtr\_sum\_clicks,asmtVle.half\_sum\_clicks,asmtVle.threeqtr\_sum\_clicks,asmtVle.qtr\_half\_sum\_clicks,asmtVle.half\_threeqtr\_sum\_clicks,asmtVle.thrd\_sum\_clicks,asmtVle.twothrd\_sum\_clicks,asmtVle.thrd\_twothrd\_sum\_clicks, |
|  | asmtVle.cma\_assmt\_score, asmtVle.tma\_assmt\_score, asmtVle.tma\_cma\_assmt\_score, asmtVle.final\_exam, asmtVle.total\_weight,asmtVle.is\_reenrolled,asmtVle.final\_exam\_score |
|  | from |
|  | public."studentCourseRegistrationFeatures" as stdtreg LEFT join |
|  |  |
|  | (select vle.qtr\_sum\_clicks,vle.half\_sum\_clicks,vle.threeqtr\_sum\_clicks,vle.qtr\_half\_sum\_clicks,vle.half\_threeqtr\_sum\_clicks,vle.thrd\_sum\_clicks,vle.twothrd\_sum\_clicks,vle.thrd\_twothrd\_sum\_clicks, |
|  | asmt.\* |
|  | from |
|  | public."studentAssessmentFeaturesSTG" asmt RIGHT JOIN |
|  | public."studentVleFeatures" vle |
|  | on asmt.id\_student = vle.id\_student AND asmt.code\_module = vle.code\_module AND asmt.code\_presentation = vle.code\_presentation) as asmtVle |
|  | on stdtreg.id\_student = asmtVle.id\_student AND stdtreg.code\_module = asmtVle.code\_module AND stdtreg.code\_presentation = asmtVle.code\_presentation |
|  | ); |

#### TABLE: FinalData

Description: Student Info, activities and performance for data modeling and EDA

Data Sources: analysisFeature

Python Script:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| id\_student | INTEGER | PK, a unique identification number for the student |
| code\_module | INTEGER | PK, the identification code of the module |
| code\_presentation | INTEGER | PK, the identification code of the module presentation |
| module\_domain | INTEGER | Group of the module, Social Science courses are defined as AAA, BBB, and GGG. STEM courses are defined as CCC, DDD, EEE, FFF |
| module\_presentation\_length | INTEGER | length of the module-presentation in days. |
| term | INTEGER | code name of the term. “B” for the presentation starting in February and “J” for the presentation starting in October. |
| year | INTEGER | the year of module-presentation |
| final\_result | INTEGER | student’s final result in the module-presentation (fail = 0, pass=1) |
| gender | INTEGER | the student’s gender (male =0, female =1) |
| region | INTEGER | identifies the geographic region, where the student lived while taking the module-presentation. |
| highest\_education | INTEGER | highest student education level on entry to the module presentation. |
| imd\_band | INTEGER | specifies the [Index of Multiple Deprivation](https://en.wikipedia.org/wiki/Multiple_deprivation_index) band of the place where the student lived during the module-presentation. |
| age\_band | INTEGER | band of the student’s age. |
| num\_of\_prev\_attempts | INTEGER | the number times the student has attempted this module. |
| disability | INTEGER | indicates whether the student has declared a disability. |
| date\_registration | INTEGER | the date of student’s registration on the module presentation, this is the number of days measured relative to the start of the module-presentation |
| date\_unregistration | INTEGER | date of student unregistration from the module presentation; this is the number of days measured relative to the start of the module-presentation. |
| scotland | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| east\_anglian\_region | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| london\_region | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| south\_region | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| north\_western\_region | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| west\_midlands\_region | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| south\_west\_region | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| east\_midlands\_region | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| south\_west\_region | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| wales | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| yorkshire\_region | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| north\_region | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| ireland | BOOLEAN | if a student lived in this region while taking the module-presentation. (no=0,yes=1) |
| b4\_sum\_clicks | DOUBLE PRECISION | the number of times a student interacts with the material before the start of the module-presentation. |
| qtr\_sum\_clicks | BIGINT | the number of times in the first quarter that student interacts with the material |
| half\_sum\_clicks | BIGINT | the number of times a student interacts with the material halfway through the module. |
| threeqtr\_sum\_clicks | BIGINT | 75% of the number of times of the student interacts with the material |
| qtr\_half\_sum\_clicks | BIGINT | the number of times in the second quarter that student interacts with the material |
| half\_threeqrt\_sum\_clicks | BIGINT | the number of times in the third quarter that student interacts with the material |
| twothrd\_sum\_clicks | BIGINT | 66.66% of the number of times a student interacts with the material |
| thrd\_twothrd\_sum\_clicks | BIGINT | the number of times between 33.33%- 66.66% that a student interacts with the material |
| all\_clicks | DOUBLE PRECISION | the number of times a student interacts with the material on that day. |
| std\_half\_score | DOUBLE PRECISION | the score% of student assessments before halfway of the module-presentation length |