## 

## 

Predicting the Probability of Need for Student Intervention

Team 7

**Amanda**: Project Lead & SME (Business Objectives)

**Dylan:** Data Scientist// Statistician (Technical/Data Understanding)

**Bridge:** Data Engineer/ Statistician (Preparing and Building the Model)

**Brianne**: Business Analyst (Model Accuracy)

QMB6930

Dr. Janice Carrillo

Dr. James Hoover

July 6th, 2020

[**Summary of Student Success Analysis**](#_62b4p976nqz8) **4**

[**Description of the Data Sources**](#_izws1dg6lkkl) **5**

[Description of Key Data Tables](#_rc361q30pdbf) 5

[Data Table Source](#_a7b48bguu4q6) 5

[Data Lake](#_bx8txgq26ogz) 6

[Joining Tables in Data Lake](#_gd3masenlci3) 6

[Data Dictionary Files](#_pngyzdydf9k7) 6

[Important Fields or Variables](#_6ak5ce3gr6lu) 6

[**Data Wrangling**](#_h8023rjjiuap) **9**

[Quality of the Data: Key Variables](#_dqj1y6wd03e5) 9

[Quality of the Data: Visualizations](#_21xvxpk0tplp) 11

[Quality of the Data: Issues Addressed](#_d1dp6dtmw1pj) 14

[Pseudo:Code SQL & Python](#_q4ajxxz2f1n) 15

[Missing Data & Outliers](#_djnlgx8is6yp) 15

[Data Transformations](#_o2yw14qz7pof) 16

[**Data Munging**](#_9jmc7qj6zxz9) **17**

[Additional Data Sources](#_ao1y4n4sy9go) 17

[Identifying New Fields](#_5sd0d4fbyrn0) 18

[Aggregation Methods](#_i8ch1r2ikzwo) 19

[**Development Workflow**](#_7gy0cpe2jf50) **20**

[Data Source Diagram](#_k5j00xwksioi) 20

[Data Source Files](#_fu0dsye79tz6) 22

[**Appendix**](#_fo2gtmhilywa) **23**

[Appendix 1: SQL Historic Indicators for Student Intervention](#_uvzfxpd3v59f) 23

[Appendix 2: SQL Historic Indicators for Auto-Generated Messages](#_exb4blbfudaa) 24

[Appendix 3: SQL Data Tables Utilized](#_you505yo5knj) 25

[Appendix 4: SQL Data Table Creation UF\_R2\_ANALYSIS\_UNDERGRAD](#_4i26uqm6lkz8) 27

[Appendix 5: SQL Data Table Creation UF\_R2\_ANALYSIS\_UNDERGRAD\_0709  
UF\_R2\_ANALYSIS\_UNDERGRAD\_1719](#_v3mzc7u1tgao) 30

[Appendix 6: SQL Data Dictionary](#_fspo6reiloyt) 33

[Appendix 7: Python Quality of the Data Report  
UF\_R1\_SUCCESS\_ANALYSIS\_UGRD](#_izvjhaf7920e) 38

[Appendix 8: Python Visualizations on the Quality of the Data](#_iniy0qjfpqho) 42

[Appendix 9: SQL Summary Counts of New Variables](#_dqrd0lra6gh1) 46

[Appendix 10: SQL UF\_R2\_ANALYSIS\_UGRDDRAFT](#_hwj2rwp5x4eg) 47

[Appendix 11: Student\_term\_enrollment\_model\_v3](#_lf1g2x9031j0) 50

## **Summary of Student Success Analysis**

There are multiple ways for Universities to define student success beyond a single graduation metric. Historically the University of Florida has used six key indicators when determining if a student is on track for graduation and academic success or if there is a need for student intervention.

Historic Indicators for Student Intervention

* Low Term GPA
* Part-Time Academic Load Status
* Not Registered for Classes
* Withdrawn from Classes
* Full-Time Load Status
* Greater Than Full Time Load Status

These historic indicators are displayed on a dashboard for academic coaches to assist them in advising efforts while meeting with students. Additional indicators are utilized to generate automated electronic outreach to students.

Historic Indicators for Auto-Generated Messages

* Audit Plan Status
* Career Status
* Degree Revoked
* Degree Checkout Status
* Eligible for Graduation
* Expected Graduation Term
* Program Hours
* Program Status

The code for how both sets of indicators have been calculated can be found on [Github](https://github.com/ajpiter/StudentSuccess/blob/master/SQL/6.15.20_UF_Historic_Success_Indicatiors) or in Appendix 1 and 2.

Using data gathered from the UF Data Lake provided by UFIT, an updated analytical model will be created to predict the likelihood of undergraduate student success. The model could then be used to provide more information to coaches who work directly with students.

These initial findings will be shared with SMEs, both at the Provost office, UF IT, and the educators that students face. By using the external team, new insight will be gained into the definition, scope, and boundaries of the findings.

## **Description of the Data Sources**

### **Description of Key Data Tables**

Below is a list of key data tables, including a high-level description of how each is anticipated to be used in the analysis. For a full list of all tables that will be utilized in the analysis please reference Appendix 3.

**UF\_B\_Student\_Term:** The main student record information table which includes 27 variables of student record information. Notable variables include GPA, transfer credits, and residency status. Variables will be utilized:

1. To evaluate which category of success a student record has historically been placed in utilizing the UF\_SUCCESS\_INDICATORS table.
2. To determine if any individual variables were a statistically relevant indicator of student success.

Please reference [Github](https://github.com/ajpiter/StudentSuccess/blob/master/SQL/06.29.2020%20UF_R1_SUCCESS_ANALYSIS_UDGD%260709%261719) or Appendix 4 and 5 for SQL queries to create the following tables.

**UF\_R1\_SUCCESS\_ANALYSIS:** Was created by joiningUF\_SUCCESS\_TARGETS to the UF\_B\_STDNT\_TERM table.

**UF\_R1\_SUCCESS\_ANALYSIS\_UNDERGRAD:** This has the same variables as the UF\_R1\_SUCESS\_ANALYSIS table but is filtered for just undergraduates.   
  
**UF\_R1\_SUCCESS\_ANALYSIS\_0709**: The UF\_R1\_SUCCESS\_ANALYSIS TABLE was narrowed to the years of 2007 to 2009. These years were chosen to compare if student success indicators from the Great Recession from 2007 to 2009, were the same indicators for student records in more recent graduating classes.

**UF\_R1\_SUCCESS\_ANALYSIS1719:** The UF\_R1\_SUCCESS\_ANALYSIS TABLE was narrowed to the years of 2007 to 2009 and 2017 to 2019. These years were chosen so it could be compared if student success indicators from the Great Recession from 2007 to 2009, were the same indicators for student records in more recent graduating classes.

### **Data Table Source**

In the building of this model, data were acquired in three different ways. First, available data was utilized in UF’s data lake. Then new data tables were created that combined different data sources within the data lake. Finally, publicly accessible data was utilized and joined with existing student record data for further insights.

### **Data Lake**

Within this report, the term data lake refers to the environment provided by UF IT at the request of the UF Provost office and contains comprehensive anonymized student records dating back to 1977. Before 1977 a small sample of additional records was available as they had been migrated from previous databases. Within the current data lake environment, the data is stored as SQL tables within the SQL Developer (SQL) program. In addition to SQL, Python will be utilized for additional analysis and to build the final model recommendation.

### **Joining Tables in Data Lake**

For ease of use, variables will be joined from multiple tables into new comprehensive tables. First as the UF\_R1\_SUCCESS\_ANALYSIS table. Then tables will be narrowed down to subset the data, based on undergraduate status, as UF\_R1\_SUCCESS\_ANALYSIS\_UNDERGRAD and by year enrolled. The goal will be to compare undergraduates enrolled during the great recession, 2007 to 2009, as the UF\_R1\_SUCCESS\_ANALYSIS\_0709 table to compare it to recent graduates. For this analysis, recent graduates will be defined as any undergraduates who graduated between 2017 and 2019 and will be subsetting their data on the UF\_R1\_SUCCESS\_ANALYSIS1719 table. Based on that comparison a model will be built comparing student record variables against the likelihood for student success.

The SQL code for creating and subsetting the new tables can be found on Github and in Appendix 3 & 4.

### **Data Dictionary Files**

Please Reference Appendix 5, to see the complete Data Dictionary File which describes the contents, format, and structure of a database.

### **Important Fields or Variables**

Below are important variables contained in the data tables and a description of why they are anticipated to be important in the analysis**.**

**ACAD\_CAREER:** This is how data is filtered by undergraduate and graduate-level students.

**ACAD\_PROG\_PRIMARY:** Defines the academic program in which a student is enrolled in.

**ACADEMIC\_LOAD:** Is a student less likely to graduate within the 4-year term if they are taking less than a full academic load?

**AGE\_YEARS:** Does a student's age have an impact on the likelihood of success? If data were bucketed in "Early" "Traditional" "Late20s" "Advanced" age groups does one group have a better likelihood for success than another?

**CUM\_GPA:** Defines what the student’s cumulative GPA is based on all classes completed.

**CUR\_GPA:** What the student’s current GPA is based on all classes completed to the current term.

**ENRL\_CNT:** Do the number of classes a student is enrolled in per semester impact a student's likelihood to succeed?

**ENRL\_FLAG:** Indicates whether a student is currently enrolled.

**ENRL\_SUMMER\_A\_FLAG:** Shows students enrolled for Summer A for their last enrolled module.

**ENRL\_SUMMER\_B\_FLAG:** Shows students enrolled for Summer B for their last enrolled module.

**ENRL\_SUMMER\_C\_FLAG:** Shows students enrolled for Summer C for their last enrolled module.

**JUNIOR\_SENIOR\_FLAG:** Flags students that are juniors or seniors.

**PERSON\_SID:** Other identifications of a student

**RESIDENCY:** This indicates a student’s residency status while enrolled.

**STRM:** This identifies the term a student was enrolled as a string text, for example, ‘Fall 2019’.

**TERM\_BEG\_DT\_SID:** This identifies the beginning term as a student who was enrolled as a DateTime value.

**TERM\_BEG\_DT\_SID\_FALL\_CATGRY:** By comparing other years, this categorizes a start date for classes as “early,” “average,” or “late” for the Fall semesters.

**TERM\_BEG\_DT\_SID\_SPRING\_CATGRY:** By comparing other years, this categorizes a start date for classes as “early,” “average,” or “late” for the Spring semesters.

**TERM\_END\_DT\_SID:** This identifies the end of a term a student was enrolled as a DateTime value.

**TERM\_END\_DT\_SID\_FALL\_EARLY:** By comparing other years, this categorizes an end date for classes as “early,” “average,” or “late” for the Fall semesters.

**TERM\_END\_DT\_SID\_SPRING\_EARLY:** By comparing other years, this categorizes an end date for classes as “early,” “average,” or “late” for the Spring semesters.

**TERM\_LENGTH\_CATEGORY:** Shows the term length by defining as “Short,” “Average,” and “Long” determined by K Means Clustering.

**TERM\_LENGTH\_DAYS:** The number of days of a term length.

**TERM\_SEASON:** The season of which a term is in. I.E. “Summer,” “Spring,” “Fall”

**TERM\_SID:** This is the unique numeric code assigned to each unique term.

**TOT\_GRADE\_POINTS:** The total points of a grade.

**TOT\_PASSD\_PRGRSS:** The total credits a student has received based on the completion of courses.

**TOT\_REQUIRED\_PRGRSS\_PROGRAMNAME:** The total credit hours needed for completion of an individual undergrad program. The programs are listed under “Program Name.”

**TOT\_TAKEN\_GPA:** This is the total number of GPA points per student.

**TOT\_TAKEN\_PRGRSS:** This is the total number of classes that a student has taken.

**TOT\_TEST\_CREDIT:** Some students have tested out of classes. This shows the total amount of credits they have received from testing out.

**TOT\_TRNSFR:** The total amount of credits transfer students have transferred with them.

## 

## **Data Wrangling**

### **Quality of the Data: Key Variables**

Multiple reports were created to determine the quality of the data. You can view the outcome of these reports on [Github](https://github.com/ajpiter/StudentSuccess/blob/master/Python/7.2.20%20UF_R1_SUCCESS_UNDERGRAND%20Counts%20on%20Outliers%20Report.docx) or Appendix 7.

**ACAD\_CAREER:** It appears that every student is coded as either a GRAD, MED, PROF, UGRD, or VEM and the data is reasonably distributed.

**ACAD\_PROG\_PRIMARY:** This is a reasonable distributed categorical variable with students categorized as either ‘GRAGL’, ‘GRENG’, ‘GRLAS’, ‘PRPBH’, ‘UGENG’, ‘UGLAS’, ‘UGPBH’, ‘UNVEM’ and ‘VMVEM’.

**ACADEMIC\_LOAD:** This is a reasonable distributed categorical variable with students categorized as either ‘F’, ‘H’, ‘L’, ‘N’, or ‘T’ indicating a student’s academic load. The distribution ins F: 512780, H: 369640, L: 344164, N: 493853, T: 239887

**AGE\_YEARS:** The age is reasonably distributed between ages 17 and 26. After 26 years of age, there is not a single student that is between 26 and 31 years old. Then there are 33698 students listed as 31 years old. The issue repeats for students aged 33 years of age with 25214 of students listed.

**CUR\_GPA:** There are 462,094 students records indicating their current GPA was between 1.9 and 4.0. However, there are 2,459,4888 student records with a 0.0 GPA. Of those 2.4 million records 501,873 student records were for students with a 0.0 current GPA listed even when their TOT\_PASSD\_PRGRSS is greater than 1. Please see Appendix 7 which runs the counts on CUR\_GPA in further detail.

**DATE\_OF\_BIRTH:** There are 27,108 unique dates of births for student records. For a database whose main records cover 43 years of student records from 1977 to 2020, this number seems plausible.

**ENRL\_CNT:** The number of credits a student is enrolled in seems reasonably accurate and all data falls between 0 and 13 credits.

**ENRL\_FLAG:** This is a reasonable distributed binary variable with students categorized as either ‘Y’ for yes or ‘N’ for no indicating where they are an upperclassman.

**ENRL\_SUMMER\_A\_FLAG:** Every student is listed an ‘N’ for No indicating that not a single student was enrolled in Summer A classes, which seems unlikely.

**ENRL\_SUMMER\_B\_FLAG:** This is a reasonable distributed binary variable with students categorized as either ‘Y’ for yes or ‘N’ for no indicating whether they were enrolled in Summer B classes.

**ENRL\_SUMMER\_C\_FLAG:** This is a reasonable distributed binary variable with students categorized as either ‘Y’ for yes or ‘N’ for no indicating whether they were enrolled in Summer C classes.

**JUNIOR\_SENIOR\_FLAG:** This is a reasonable distributed binary variable with students categorized as either ‘Y’ for yes or ‘N’ for no indicating where they are an upperclassman.

**PERSON\_SID:** This is one of the main join ids used to connect the tables and represents a unique numeric identifier assigned to each student.

**RESIDENCY:** This is a reasonable distributed categorical variable with students categorized as either ‘F’, ‘A’ or ‘N’ indicating a student’s residency status.

**STRM:** There appear to be 61,056 inputs for the years 1727, 1729, 1730, 1732, and 1734 which are likely dummy variables since this database is only supposed to represent students to 1977.

**TERM\_BEG\_DT\_SID:** The data appears evenly distributed with no clear outliers and formatted in the YEARMONTHDAY date format.

**TERM\_END\_DT\_SID:**  The data appears evenly distributed with no clear outliers and formatted in the YEARMONTHDAY date format.

**TERM\_SID:** There are dozens of terms identified, and the data appears clean with an equal distribution between terms.

**TOT\_GRADE\_POINTS:** The data is reasonably distributed with total grade points ranging from 0 to 448.69.

**TOT\_PASSD\_PRGRSS:** The majority of the data is reasonable. However, four students report their total credits taken in unreasonable decimals (28.68, 40.01, 50.68, and 61.35). Additionally, there are two students whose total taken credits are 146, and 161 which is unreasonably high considering students only need 120 total credits for graduation. Additionally, there are 154,960 student records that report a 0 passed progress. Please reference Appendix 8 for the Quality of Data Report for additional counts.

**TOT\_TAKEN\_PRGRSS:** The majority of the data is reasonable, however, four students report their total credits taken in unreasonable decimals (28.68, 40.01, 50.68, and 61.35). Additionally, there are two students whose total taken credits are 146, and 161 which is unreasonably high considering students only need 120 total credits for graduation.

**TOT\_TEST\_CREDIT:** As expected most students have tested out of 0 credits (389,720 students) and 36337 students have tested out of between 22 and 30 credits. It is unusual that there are 0 students who have tested out of less than 22 credits.

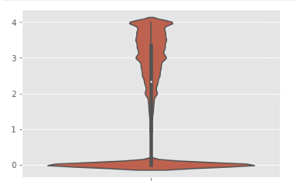
**TOT\_TRNSFR**: As expected most students have 0 transfer credits (309,985 students) and 122248 students transferred between 3 and 15 credits from other institutions.

**UF\_CLASS\_EOT:** The data is a reasonably distributed variable ranging from 0 to 9.

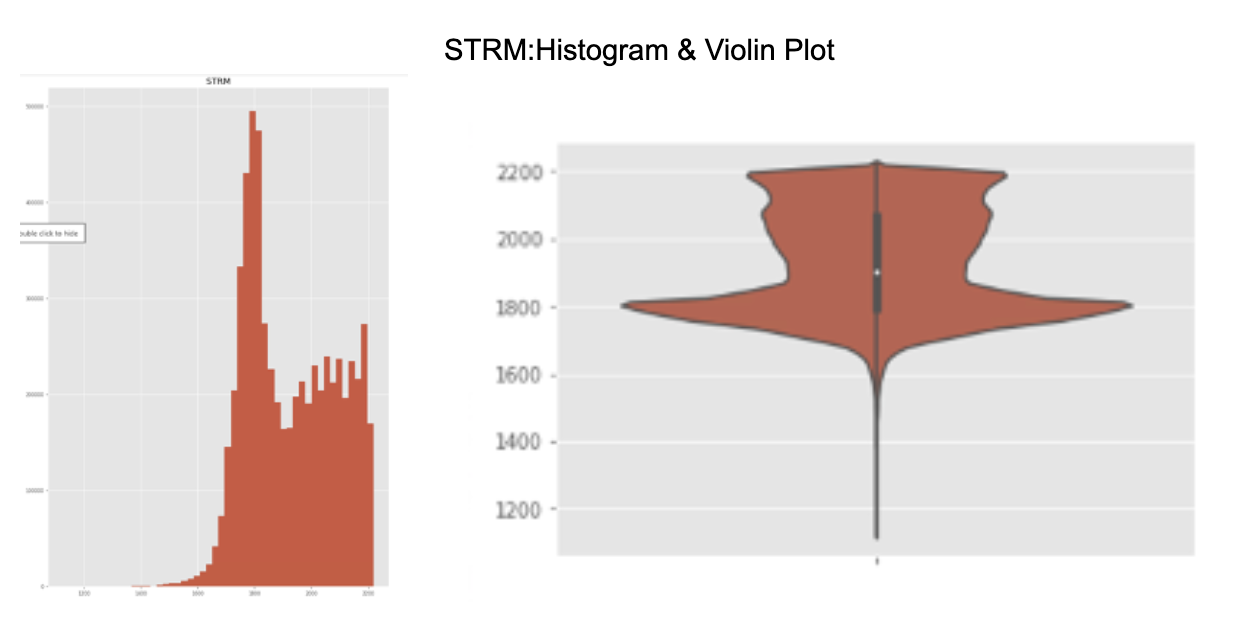
### **Quality of the Data: Visualizations**

Below are data visualizations and further discussions on the quality of the data for variables where there are unexpected values. For additional Visualizations please reference the project [Github](https://github.com/ajpiter/StudentSuccess/blob/master/Python/Data%20Visualization%20for%20Data%20Quality) page or Appendix 8.

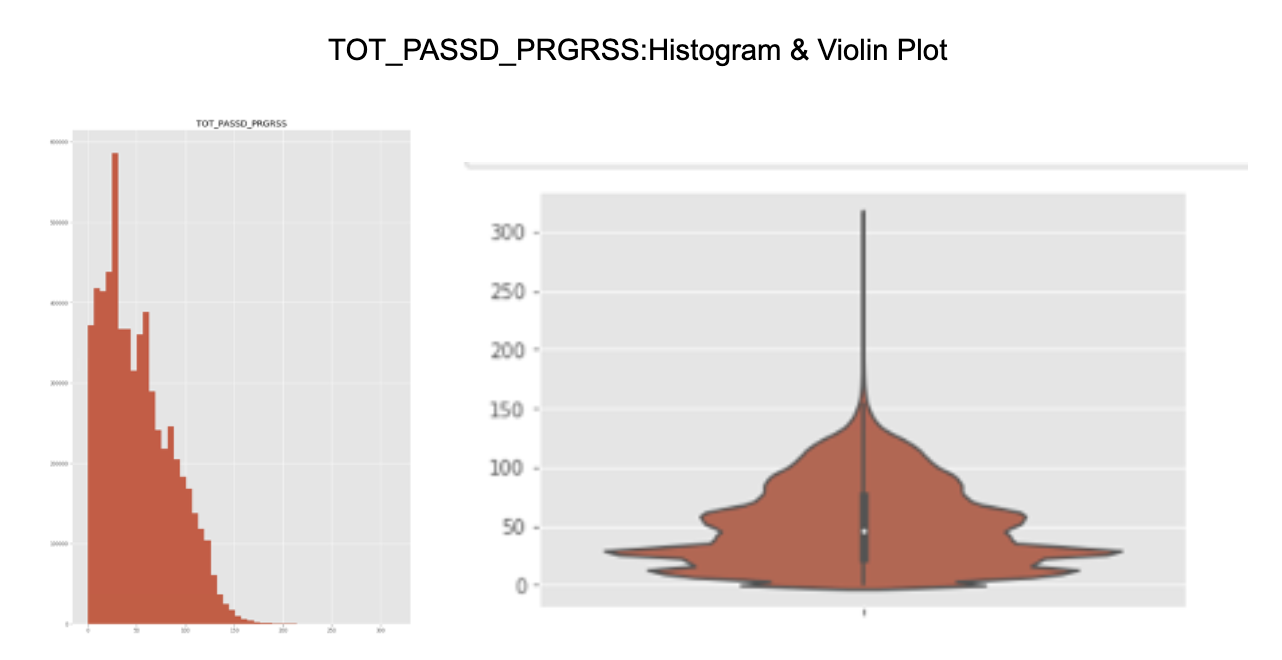
**CUR\_GPA: Violin Plot**



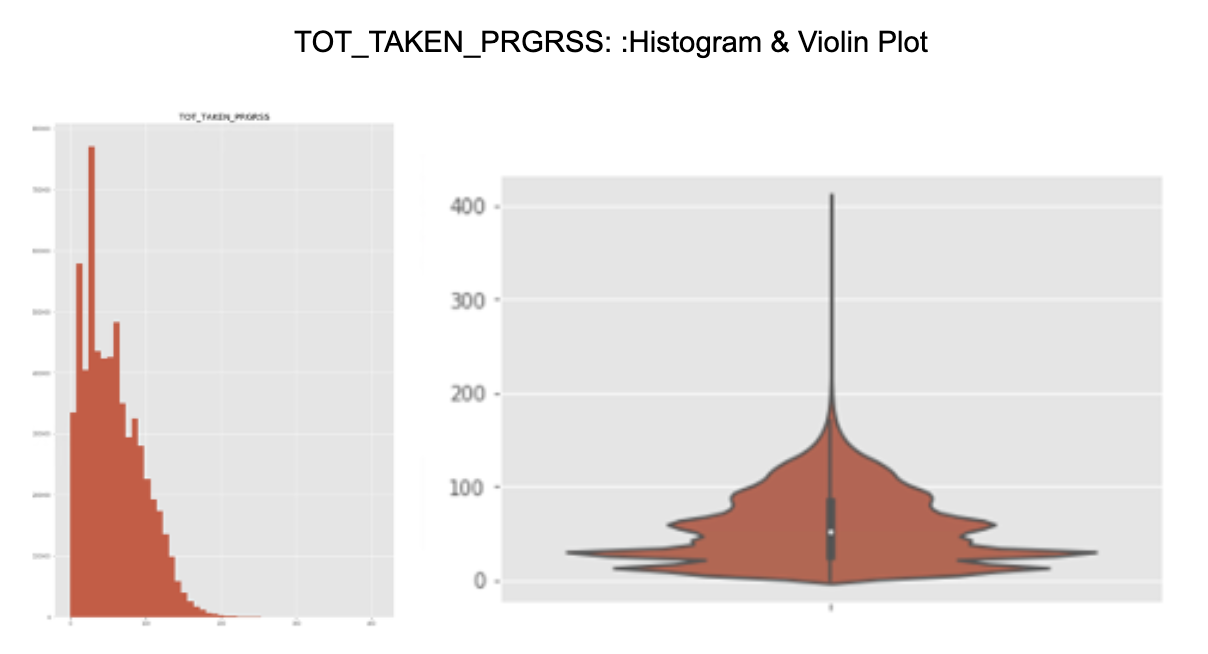
**CUR\_GPA: Violin Plot:** For less than half, 462,094, of students the current GPA is between 1.9 and 4.0. However, there are 501,873 student records with a 0.0 current GPA listed even when their TOT\_PASSD\_PRGRSS is greater than 1.



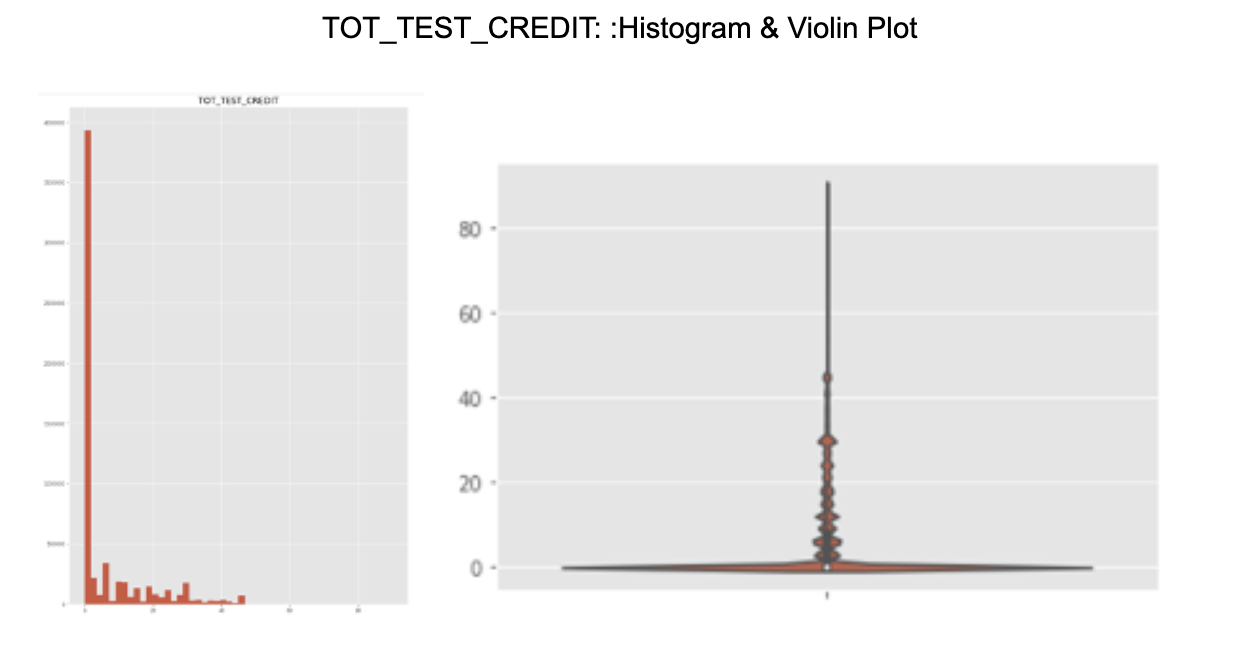
**STRM:** There appear to be 61,056 inputs for the years 1727, 1729, 1730, 1732, and 1734 which are likely dummy variables since this database is only supposed to represent student records to 1977.



**TOT\_PASSD\_PRGRSS:** The majority of the data is reasonable. However, four student records report their total credits taken in unreasonable decimals (28.68, 40.01, 50.68, and 61.35). Additionally, there are two student records whose total taken credits are 146, and 161 which is unreasonably high considering students only need 120 total credits for graduation.



**TOT\_TAKEN\_PRGRSS:** The majority of the data is reasonable. However, four students report their total credits taken in unreasonable decimals (28.68, 40.01, 50.68, and 61.35). Additionally, there are two students whose total taken credits are 146, and 161 which is unreasonably high considering students only need 120 total credits for graduation.



**TOT\_TEST\_CREDIT:** As expected most students have tested out of credits (252609 students) and 33556 students have tested out of between 22 and 30 credits. It is unusual that there are 108419 students who have tested out of less than 22 credits.

### 

### **Quality of the Data: Issues Addressed**

Below is a short summary explaining how it was decided to address quality issues with the data.

After address the data quality issues below there are #### records on the UF\_R1\_SUCCESS\_ANALYSIS\_UNDERGRAD table #### records on the UF\_R1\_SUCCESS\_ANALYSIS\_0709 table and ### records on theUF\_R1\_SUCCESS\_ANALYSIS1719 table.

**CUR\_GPA:** For less than half, 462,094, of students the current GPA is between 1.9 and 4.0. However, there are 501,873 students with a 0.0 current GPA listed even when their TOT\_PASSD\_PRGRSS is greater than 1. Any records where a GPA is 0 has been eliminated.

**TOT\_PASSD\_PRGRSS:** The majority of the data is reasonable. However, four students report their total credits taken in unreasonable decimals (28.68, 40.01, 50.68, and 61.35). Additionally, there are two students whose total taken credits are 146, and 161 which is unreasonably high considering students only need 120 total credits for graduation. Any records where 28.68, 40.01, 50.68, and 61.35 are the results have been eliminated.

**TOT\_TAKEN\_PRGRSS:** The majority of the data is reasonable. However, four students report their total credits taken in unreasonable decimals (28.68, 40.01, 50.68, and 61.35). Additionally, there are two students whose total taken credits are 146, and 161 which is unreasonably high considering students only need 120 total credits for graduation. Any records where 28.68, 40.01, 50.68, and 61.35 are the results have been eliminated.

**TOT\_TEST\_CREDIT:** As expected most students have tested out of credits (252609 students) and 33556 students have tested out of between 22 and 30 credits. It is unusual that there are 108419 students who have tested out of less than 22 credits.More research is needed to be done to determine with the University of Florida project stakeholders before a decision can be made on how to address the data.

### 

### **Pseudo:Code SQL & Python**

Code used throughout this report can be found in the Appendixes to this document or on the Github page at:

[**Appendix**](#_fo2gtmhilywa)[**23**](#_fo2gtmhilywa)

[Appendix 1: SQL Historic Indicators for Student Intervention](#_uvzfxpd3v59f) [23](#_uvzfxpd3v59f)

[Appendix 2: SQL Historic Indicators for Auto-Generated Messages](#_exb4blbfudaa) [24](#_exb4blbfudaa)

[Appendix 3: SQL Data Tables Utilized](#_you505yo5knj) [25](#_you505yo5knj)

[Appendix 4: SQL Data Table Creation UF\_R2\_ANALYSIS\_UNDERGRAD](#_4i26uqm6lkz8) [27](#_4i26uqm6lkz8)

[Appendix 5: SQL Data Table Creation UF\_R2\_ANALYSIS\_UNDERGRAD\_0709  
UF\_R2\_ANALYSIS\_UNDERGRAD\_1719](#_v3mzc7u1tgao) [30](#_v3mzc7u1tgao)

[Appendix 6: SQL Data Dictionary](#_fspo6reiloyt) [33](#_fspo6reiloyt)

[Appendix 7: Python Quality of the Data Report  
UF\_R1\_SUCCESS\_ANALYSIS\_UGRD](#_izvjhaf7920e) [38](#_izvjhaf7920e)

[Appendix 8: Python Visualizations on the Quality of the Data](#_iniy0qjfpqho) [42](#_iniy0qjfpqho)

[Appendix 9: SQL Summary Counts of New Variables 46](#_dqrd0lra6gh1)

[Appendix 10: SQL UF\_R2\_ANALYSIS\_UGRDDRAFT](#_hwj2rwp5x4eg) [47](#_hwj2rwp5x4eg)

[Appendix 11: Student\_term\_enrollment\_model\_v3](#_lf1g2x9031j0) [50](#_lf1g2x9031j0)

### **Missing Data & Outliers**

**ENRL\_SUMMER\_A\_FLAG:** Every student record is listed an ‘N’ for No indicating that not a single student was enrolled in Summer A classes, which seems unlikely.

**STRM:** There appear to be 61,056 inputs for the years 1727, 1729, 1730, 1732, and 1734 which are likely dummy variables since this database is only supposed to represent student rcords to 1977.

**TOT\_PASSD\_PRGRSS:** The majority of the data is reasonable. However, four students report their total credits taken in unreasonable decimals (28.68, 40.01, 50.68, and 61.35). Additionally, there are two students whose total taken credits are 146, and 161 which is unreasonably high considering students only need 120 total credits for graduation.

**TOT\_TAKEN\_PRGRSS:** The majority of the data is reasonable. However, four students report their total credits taken in unreasonable decimals (28.68, 40.01, 50.68, and 61.35). Additionally, there are two students whose total taken credits are 146, and 161 which is unreasonably high considering students only need 120 total credits for graduation.

### 

### **Data Transformations**

Please reference the Identifying New Fields section on page 17 for information on the data transformation techniques we utilized.

## **Data Munging**

### **Additional Data Sources**

From researching the topic, it was gathered that outside forces could play a factor in student success. Currently, outside data sources that are being considered are as follows:

* [**http://zipatlas.com/us/fl/zip:code:comparison/unemployment:rate.htm**](http://zipatlas.com/us/fl/zip-code-comparison/unemployment-rate.htm)**.**
  + This data shows unemployment rates for Florida based on zip code. It is hypothesized that income available might have an effect on student records in different ways but overall will affect their abilities to succeed in college. By finding unemployment rates in Florida, this could be compared to the student record to see if there is a correlation.
* [**https://archive.catalog.ufl.edu/ugrad/1718//liberalarts/majors/home.html**](https://archive.catalog.ufl.edu/ugrad/1718//liberalarts/majors/home.html)
  + This data shows the different majors and credit requirements. It is hypothesized this might mean student success might need to be met differently depending on the major.
* [**https://floridagators.com/sports/2015/12/10/\_overview\_.aspx**](https://floridagators.com/sports/2015/12/10/_overview_.aspx)
  + This data shows when the University of Florida won championships. It is hypothesized this may show an increase in admissions depending on the years. This might affect the student success percentage of graduation as the University was accepting students who would not have otherwise applied and got in. Additionally, it is believed that championships were occurring that might have affected schoolwork for students. This data might show a correlation between grade drops and championships that could affect student success.
* [**https://www.mapsofworld.com/hurricane/dates.html**](https://www.mapsofworld.com/hurricane/dates.html)**.**
  + This data shows major hurricanes that hit different states during the years. From calls with the sponsor, it was discussed that hurricanes can cause students to leave the school to help their families. It is hypothesized this might affect student success as grades might suffer from such hardships or might need to take breaks to help families in other states.

**Merging Data Sources**

The following tables were merged UF\_R1\_ANALYSIS\_UNDERGRAD, UF\_B\_Person\_Student\_GRP, & UF\_B\_TERM\_SPLAN into a new table called UF\_R2\_ANALYSIS\_UNDERGRAD.

The tables were joined as part of the Round 2 Analysis plan, the code to create the tables can be found on [Github](https://github.com/ajpiter/StudentSuccess/blob/master/SQL/07.01.2020%20UF_R2_ANALYSIS_UNDERGRAD) or Appendix 10.

For the Round 3 analysis, it will be determined if one outside source will need to merge with the newly created UF\_R2\_ANALYSIS\_UNDERGRAD table and save as UF\_R3\_ANALYSIS\_UNDERGRAD.

### I**dentifying New Fields**

**TERM\_BEG\_DT\_SID\_FALL\_CATGRY:** Categorize fall start dates as Early, Average, or Late compared to other years. By using K Means Clustering on Fall Start Dates in TERM\_BEG\_DT\_SID

**TERM\_BEG\_DT\_SID\_SPRING\_CATGRY:** Categorize spring start dates as Early, Average, or Late compared to other years. By using K Means Clustering on Spring Start Dates in TERM\_BEG\_DT\_SID

**TERM\_END\_DT\_SID\_FALL\_EARLY:** Categorize fall ends dates as Early, Average, or Late compared to other years. By using K Means Clustering on Fall Start Dates in TERM\_END\_DT\_SID

**TERM\_END\_DT\_SID\_SPRING\_EARLY:** Categorize spring end dates as Early, Average, or Late compared to other years. By using K Means Clustering on Spring Start Dates in TERM\_END\_DT\_SID

**TERM\_LENGTH\_CATEGORY:** Term Length in categorical buckets of Short, Average, and Long determined by K Means Clustering

**TERM\_LENGTH\_DAYS:** Term Length in the number of days

**TERM\_SEASON:** =IF(OR(TERM\_SID, "Summer...", Summer, ("Spring...", Spring), ("Fall...", Fall)

**TOT\_REQUIRED\_PRGRSS\_PROGRAMNAME:** The number of credit hours required for the individual undergrad program listed in "Program Name" in the variable: will likely be 50+ new variables.

### **Aggregation Methods**

Aggregating multiple variables we considered by taking variables that exist and combining their outputs to form new variables that could be analyzed against the target definition of student success.

Historically the University of Florida created six variables, as indicators for student success. Two of those new variables LOW\_TERM\_GPA\_IND and WITHDRWL\_TERM\_IND were created by aggregating information from the Academic\_LOAD, UNT\_TAKEN\_GPA, and UNT\_PASSD\_NOGPA.

SUM(CASE WHEN "CUR\_GPA" <= 2.2 AND "ACADEMIC\_LOAD" <> 'N' AND UNT\_TAKEN\_GPA<> 0 THEN 1 ELSE 0 END) AS "LOW\_TERM\_GPA\_IND",

SUM(CASE WHEN "ACADEMIC\_LOAD" <> 'N' AND UNT\_TAKEN\_GPA=0 AND UNT\_PASSD\_NOGPA=0 THEN 1 ELSE 0 END) AS "WITHDRWL\_TERM\_IND",

Those aggregated indicators were then utilized to determine which students were flagged for intervention and auto-generated electronic follow up. For the complete historic code please see the [Github](https://github.com/ajpiter/StudentSuccess/blob/master/SQL/6.15.20_UF_Historic_Success_Indicatiors) page, or reference Appendix 1 and 2.

WHEN ((IND.LOW\_TERM\_GPA\_IND + IND.PARTTIME\_TERM\_IND + IND.NOT\_REG\_TERM\_IND + IND.WITHDRWL\_TERM\_IND)>0) THEN ('5')

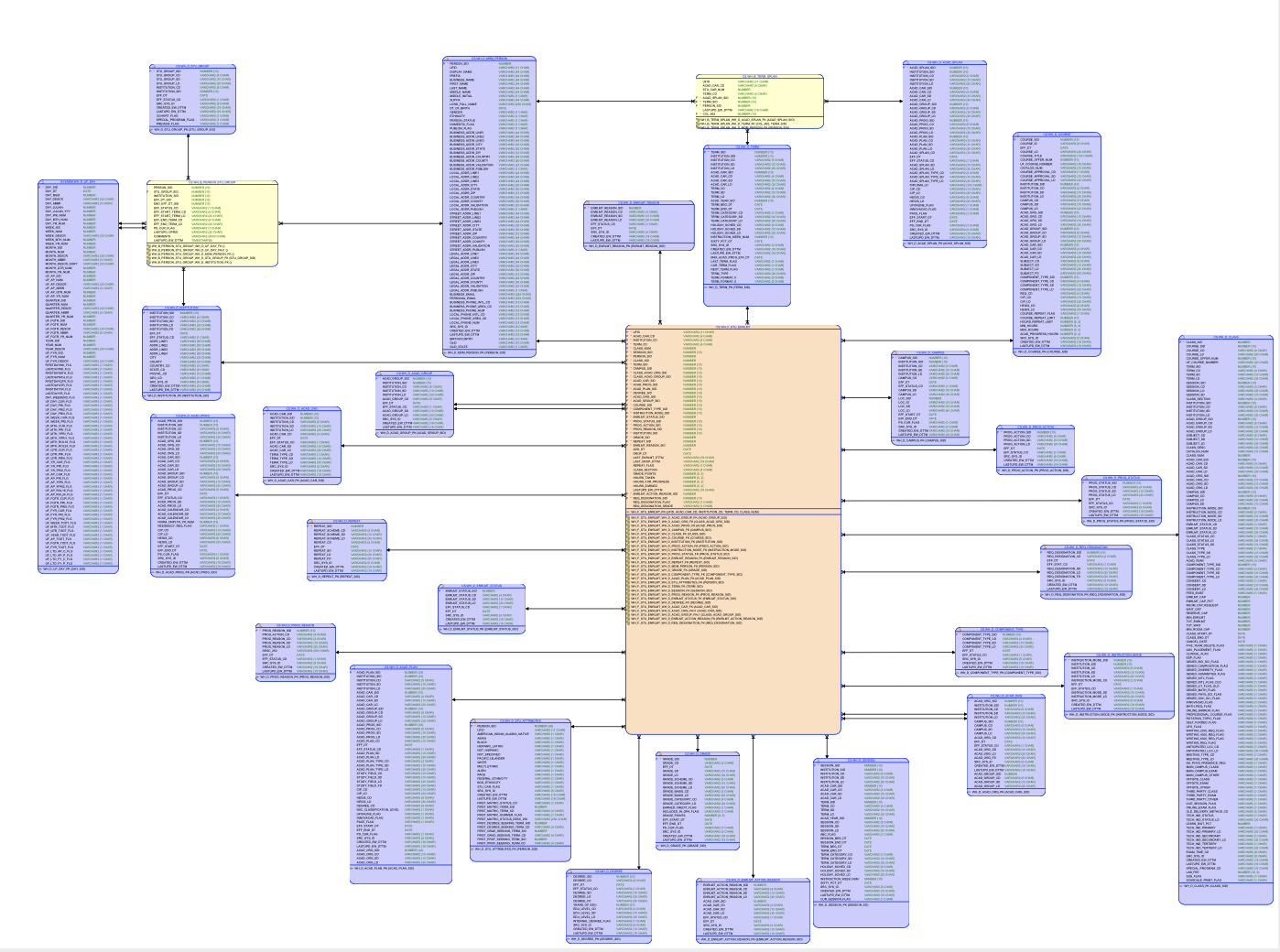
Moving forward aggregated LOW\_TERM\_GPA\_IND and WITHDRWL\_TERM\_IND will be utilized and discussion, where further aggregation could be beneficial, will be continued.

## **Development Workflow**

### **Data Source Diagram**

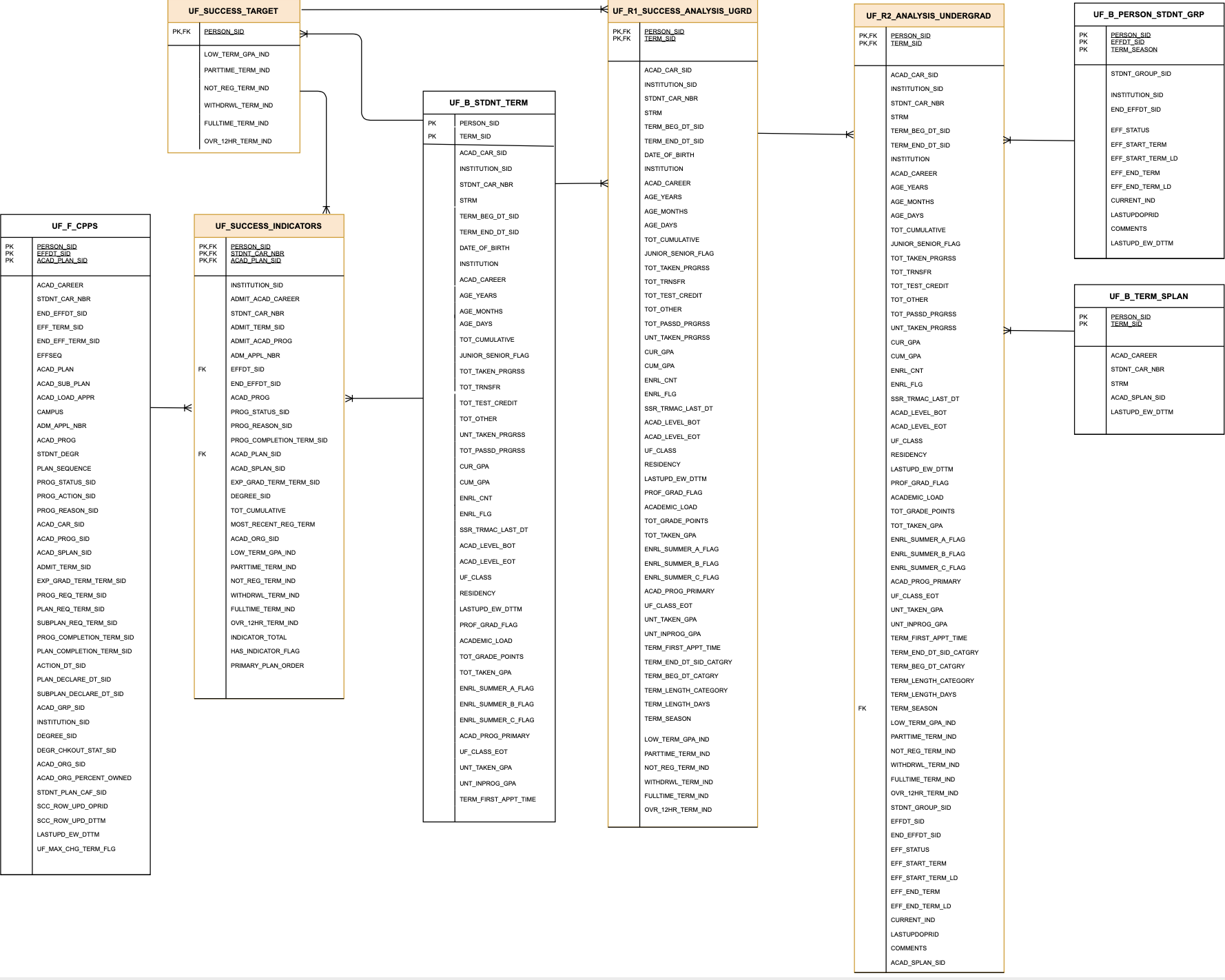
While the University of Florida data lake is home to numerous tables, the tables that are most relevant to modeling student success are in the term workgroup.

Please note that many table names and variable names have been updated, after the data was scrubbed from the original tables of students identifying information and then populated into a more anonymized format. Therefore the below historic ER diagram should be seen as a reference to the quantity of tables that could be included in the analysis from the term ecosystem. For a clearer picture of ‘Student\_term\_enrollment\_model\_v3’ please refer to the project [Github](https://github.com/ajpiter/StudentSuccess/blob/master/ERDiagrams/ER%20Student_term_enrollment_model_v3.pdf) page or Appendix 10.



For purposes in the initial analysis, UF\_B\_STDNT\_TERM table will be utilized and joined with the UF\_SUCCESS\_TARGET and UF\_SUCCESS\_INDICATORS tables. The combination of variables was then filtered to include only undergraduate students on the UF\_R1\_SUCCESS\_ANALYSIS\_UGRD table.

Below is a diagram depicting how tables were joined for our Round 1 Analysis. ,



As analysis moved to round 2 tables within the UF\_Term ecosystem will continue to be joined to the UF\_R1\_SUCCESS\_ANALYSIS\_UGRD table utilizing the PERSON\_SID.

Moving forward there are 8 table ecosystems within the UF data lake that could be relevant when modeling student success. The following tables have been identified:

* UF\_D\_ACAD\_CAR
* UF\_D\_ACAD\_GRP
* UF\_D\_ACAD\_ORG
* UF\_D\_ACAD\_PLAN
* UF\_D\_ACAD\_PROG
* UF\_D\_ACAD\_SPLAN
* UF\_D\_CAMPUS
* UF\_D\_CLASS
* UF\_D\_CRSE
* UF\_D\_INSTITUTION
* UF\_D\_INSTRCTN\_MODE

### **Data Source Files**

When examining the data, and deciding which of the numerous tables to join for the analysis of student success below were divided:

**Round 1 Data Sources**

Historic ER Student\_term\_enrollment\_model\_v3: [Github](https://github.com/ajpiter/StudentSuccess/blob/master/ERDiagrams/ER%20Student_term_enrollment_model_v3.pdf)

UF\_B\_STDNT\_TERM: Output available in the data lake F: Drive

**Round 2 Data Sources**

Round 2 ER Diagram UF\_R2\_ANALYSIS\_UNDERGRAD: Github

UF\_B\_TERM\_SPLAN: Output available in the data lake F: Drive

**Round 3 Data Sources**

Historic ER CPP\_model\_v10: [Github](https://github.com/ajpiter/StudentSuccess/blob/master/ERDiagrams/ER%20CPP_model_v10%20(2).pdf)  
 Historic ER Class\_Meeting\_Patterns\_and\_Instructors\_model\_v1: [Github](https://github.com/ajpiter/StudentSuccess/blob/master/ERDiagrams/ER%20Class_Meeting_Patterns_and_Instructors_model_v1.pdf)

Historic ER External\_test\_score\_model\_v2: [Github](https://github.com/ajpiter/StudentSuccess/blob/master/ERDiagrams/ER%20External_test_score_model_v2.pdf)

Historic ER Milestone\_model\_v1: [Github](https://github.com/ajpiter/StudentSuccess/blob/master/ERDiagrams/ER%20Milestone_model_v1.pdf)  
Historic ER Service\_indicators\_flags\_holds\_model\_v2: [Github](https://github.com/ajpiter/StudentSuccess/blob/master/ERDiagrams/ER%20Service_indicators_flags_holds_model_v2.pdf)  
Historic ER Student\_term\_enrollment\_model\_v3: [Github](https://github.com/ajpiter/StudentSuccess/blob/master/ERDiagrams/ER%20Student_term_enrollment_model_v3.pdf)   
Historic ER Student\_term\_progress\_model\_v3: [Github](https://github.com/ajpiter/StudentSuccess/blob/master/ERDiagrams/ER%20Student_term_progress_model_v3.pdf)  
Historic ER Student\_transfer\_enrollment\_model\_v1: [Github](https://github.com/ajpiter/StudentSuccess/blob/master/ERDiagrams/ER%20Student_transfer_enrollment_model_v1.pdf)

## **Appendix**

### **Appendix 1: SQL Historic Indicators for Student Intervention**

IND AS (SELECT

"EMPLID" EMPLID,

SUM(CASE WHEN "CUR\_GPA" <= 2.2 AND "ACADEMIC\_LOAD" <> 'N' AND UNT\_TAKEN\_GPA<> 0 THEN 1 ELSE 0 END) AS "LOW\_TERM\_GPA\_IND",

SUM(CASE WHEN "ACADEMIC\_LOAD" in ('P','H','L','T') THEN 1 ELSE 0 END) AS "PARTTIME\_TERM\_IND",

SUM(CASE WHEN "ACADEMIC\_LOAD" = 'N' THEN 1 ELSE 0 END) AS "NOT\_REG\_TERM\_IND",

SUM(CASE WHEN "ACADEMIC\_LOAD" <> 'N' AND UNT\_TAKEN\_GPA=0 AND UNT\_PASSD\_NOGPA=0 THEN 1 ELSE 0 END) AS "WITHDRWL\_TERM\_IND",

SUM(CASE WHEN "ACADEMIC\_LOAD" = 'F' THEN 1 ELSE 0 END) AS "FULLTIME\_TERM\_IND",

SUM(CASE WHEN "UNT\_PASSD\_GPA" > 12 THEN 1 ELSE 0 END) AS "OVR\_12HR\_TERM\_IND"

FROM CS.UF\_IPR\_COHORTS CHT INNER JOIN CS.PS\_STDNT\_CAR\_TERM CAR on CHT.UNIV\_ROW\_ID=CAR.EMPLID and CHT.STRM in ('2165','2168')

where CAR.STRM < (Select DISTINCT(TERM\_CD) from CS.UF\_D\_TERM where UF\_D\_TERM.UF\_CURR\_TERM\_FLG = 'Y')

AND ACAD\_CAREER='UGRD'

GROUP BY EMPLID),

### 

### **Appendix 2: SQL Historic Indicators for Auto-Generated Messages**

WHEN ((D.UF\_COLLVL\_APPRV ='Y' AND D.UF\_DGC\_REVOKED<> 'Y' ) or B.DEGR\_CHKOUT\_STAT='AW') THEN ('1')

WHEN (D.UF\_COLLVL\_FINAL='Y' AND D.UF\_COLLVL\_APPRV='N' AND D.EXP\_GRAD\_TERM='2201') THEN ('2')

WHEN (B.PROG\_STATUS = 'DM') THEN ('No Message to Be Sent')

WHEN (B.PROG\_STATUS = 'DC') THEN ('7')

WHEN ((D.UF\_DGC\_REVOKED = 'Y' OR D.UF\_COLLVL\_APPRV\_P='N') AND D.EXP\_GRAD\_TERM='2201') THEN ('1b')

WHEN (D.ACTIVE\_FLAG='Y' AND D.EXP\_GRAD\_TERM='2201' AND ( AUD.AUD\_PLAN\_STATUS='FAIL' OR AUD.CAREER\_STATUS='FAIL')) THEN ('1b')

WHEN (D.UF\_DGC\_REVOKED = 'N' AND D.UF\_COLLVL\_APPRV\_P='Y' AND D.EXP\_GRAD\_TERM='2201') THEN ('1a')

WHEN (D.UF\_DGC\_REVOKED = 'N' AND D.ACTIVE\_FLAG='Y' AND D.EXP\_GRAD\_TERM='2201') THEN ('1a')

WHEN (B.DEGR\_CHKOUT\_STAT in ('EG','WD') AND AUD.AUD\_PLAN\_STATUS<>'FAIL' AND AUD.CAREER\_STATUS<>'FAIL' ) THEN ('3') --EG-Eligible for Graduation, DN-Denied, WD-Withdran removed DN

WHEN (E.UF\_PRGM\_HRS>=125) THEN ('4') --2014 and 2015 this bucket will need to be removed WHEN ((IND.LOW\_TERM\_GPA\_IND + IND.PARTTIME\_TERM\_IND + IND.NOT\_REG\_TERM\_IND + IND.WITHDRWL\_TERM\_IND)>0) THEN ('5')

### 

### **Appendix 3: SQL Data Tables Utilized**

**UF\_B\_Student\_Term:** The main student information table which includes 27 variables of student information. Notable variables include GPA, transfer credits, and residency status. Variables will be utilized:

1. To evaluate which category of success a student is placed in on the UF\_SUCCESS\_INDICATORS table.
2. To determine if any individual variables were a statistically relevant indicator of student success.

**UF\_R1\_SUCCESS\_ANALYSIS:** Was created by joiningUF\_SUCCESS\_TARGETS to the UF\_B\_STDNT\_TERM table. Below is the code utilized for this process.

**UF\_R1\_SUCCESS\_ANALYSIS\_UNDERGRAD:** This has the exact same variables as the UF\_R1\_SUCESS\_ANALYSIS table but is filtered for just undergraduate student records.   
  
**UF\_R1\_SUCCESS\_ANALYSIS\_0709**: The UF\_R1\_SUCCESS\_ANALYSIS TABLE was narrowed to the years of 2007 to 2009. These years were chosen to compare if student success indicators from the Great Recession from 2007 to 2009, were the same indicators for students in more recent graduating classes.

**UF\_R1\_SUCCESS\_ANALYSIS1719:** The UF\_R1\_SUCCESS\_ANALYSIS TABLE was narrowed to the years of 2007 to 2009 and 2017 to 2019. These years were chosen to compare if student success indicators from the Great Recession from 2007 to 2009, were the same indicators for students in more recent graduating classes.

**UF\_SUCCESS\_INDICATORS:**

Other Data Tables Utilized

UF\_B\_TERM\_SPLAN:

UF\_D\_ACAD\_CAR

UF\_D\_ACAD\_GRP

UF\_D\_ACAD\_ORG

UF\_D\_ACAD\_PLAN

UF\_D\_ACAD\_PROG

UF\_D\_ACAD\_SPLAN

UF\_D\_CAMPUS

UF\_D\_CLASS

UF\_D\_CRSE

UF\_D\_INSTITUTION

UF\_D\_REQUIREMENT

UF\_D\_SRVC\_IMPACT

**UF\_F\_CPSS:**

**UF\_F\_CPSS\_Terms:**

### 

### 

### **Appendix 4: SQL Data Table Creation UF\_R2\_ANALYSIS\_UNDERGRAD**

CREATE TABLE UF\_R2\_ANALYSIS\_UGRDDRAFT as (SELECT

A.PERSON\_SID,

A.ACAD\_CAR\_SID,

A.INSTITUTION\_SID,

A.STDNT\_CAR\_NBR,

A.TERM\_SID,

A.STRM,

A.TERM\_BEG\_DT\_SID,

A.TERM\_END\_DT\_SID,

A.INSTITUTION,

A.ACAD\_CAREER,

A.AGE\_YEARS,

A.AGE\_MONTHS,

A.AGE\_DAYS,

A.TOT\_CUMULATIVE,

A.JUNIOR\_SENIOR\_FLAG,

A.TOT\_TAKEN\_PRGRSS,

A.TOT\_TRNSFR,

A.TOT\_TEST\_CREDIT,

A.TOT\_OTHER,

A.TOT\_PASSD\_PRGRSS,

A.UNT\_TAKEN\_PRGRSS,

A.CUR\_GPA,

A.CUM\_GPA,

A.ENRL\_CNT,

A.ENRL\_FLG,

A.SSR\_TRMAC\_LAST\_DT,

A.ACAD\_LEVEL\_BOT,

A.ACAD\_LEVEL\_EOT,

A.UF\_CLASS,

A.RESIDENCY,

A.LASTUPD\_EW\_DTTM,

A.PROF\_GRAD\_FLAG,

A.ACADEMIC\_LOAD,

A.TOT\_GRADE\_POINTS,

A.TOT\_TAKEN\_GPA,

A.ENRL\_SUMMER\_A\_FLAG,

A.ENRL\_SUMMER\_B\_FLAG,

A.ENRL\_SUMMER\_C\_FLAG,

A.ACAD\_PROG\_PRIMARY,

A.UF\_CLASS\_EOT,

A.UNT\_TAKEN\_GPA,

A.UNT\_INPROG\_GPA,

A.TERM\_FIRST\_APPT\_TIME,

A.TERM\_END\_DT\_SID\_CATGRY,

A.TERM\_BEG\_DT\_CATGRY,

A.TERM\_LENGTH\_CATEGORY,

A.TERM\_LENGTH\_DAYS,

A.TERM\_SEASON,

A.LOW\_TERM\_GPA\_IND,

A.PARTTIME\_TERM\_IND,

A.NOT\_REG\_TERM\_IND,

A.WITHDRWL\_TERM\_IND,

A.FULLTIME\_TERM\_IND,

A.OVR\_12HR\_TERM\_IND,

B.STDNT\_GROUP\_SID,

B.EFFDT\_SID,

B.END\_EFFDT\_SID,

B.EFF\_STATUS,

B.EFF\_START\_TERM,

B.EFF\_START\_TERM\_LD,

B.EFF\_END\_TERM,

B.EFF\_END\_TERM\_LD,

B.CURRENT\_IND,

B.LASTUPDOPRID,

B.COMMENTS,

C.ACAD\_SPLAN\_SID

FROM UF\_R1\_SUCCESS\_ANLS\_UDINTSMALL A INNER JOIN UF\_B\_PERSON\_STDNT\_GRP B

ON A.PERSON\_SID = B.PERSON\_SID

AND ((A.TERM\_SEASON = substr(B.EFF\_START\_TERM\_LD,0,4) AND substr(A.TERM\_BEG\_DT\_SID,0,4) = substr(B.EFF\_START\_TERM\_LD,-4))

OR (A.TERM\_SEASON = substr(B.EFF\_END\_TERM\_LD,0,4)AND substr(A.TERM\_BEG\_DT\_SID,0,4) = substr(B.EFF\_END\_TERM\_LD,-4)))

INNER JOIN UF\_B\_TERM\_SPLAN C

ON A.PERSON\_SID = C.PERSON\_SID

AND A.TERM\_SID = C.TERM\_SID

)

CREATE TABLE UF\_R2\_ANALYSIS\_UNDERGRAD as (SELECT

\*

FROM UF\_R2\_ANALYSIS\_UNDERGRAD A

WHERE (

A.EFFDT\_SID = (SELECT MAX(B.EFFDT\_SID)

FROM UF\_R2\_ANALYSIS\_UNDERGRAD B

WHERE A.PERSON\_SID = B.PERSON\_SID

AND A.STDNT\_GROUP\_SID = B.STDNT\_GROUP\_SID

AND A.STDNT\_CAR\_NBR = B.STDNT\_CAR\_NBR

)

)

)

### 

### **Appendix 5: SQL Data Table Creation UF\_R2\_ANALYSIS\_UNDERGRAD\_0709 UF\_R2\_ANALYSIS\_UNDERGRAD\_1719**

CREATE TABLE UF\_R1\_SUCCESS\_ANALYSIS\_UGRD as (SELECT

A.ACAD\_CAR\_SID,

A.INSTITUTION\_SID,

A.STDNT\_CAR\_NBR,

A.TERM\_SID,

A.STRM,

A.TERM\_BEG\_DT\_SID,

A.TERM\_END\_DT\_SID,

A.INSTITUTION,

A.ACAD\_CAREER,

A.AGE\_YEARS,

A.AGE\_MONTHS,

A.AGE\_DAYS,

A.TOT\_CUMULATIVE,

A.JUNIOR\_SENIOR\_FLAG,

A.TOT\_TAKEN\_PRGRSS,

A.TOT\_TRNSFR,

A.TOT\_TEST\_CREDIT,

A.TOT\_OTHER,

A.TOT\_PASSD\_PRGRSS,

A.UNT\_TAKEN\_PRGRSS,

A.CUR\_GPA,

A.CUM\_GPA,

A.ENRL\_CNT,

A.ENRL\_FLG,

A.SSR\_TRMAC\_LAST\_DT,

A.ACAD\_LEVEL\_BOT,

A.ACAD\_LEVEL\_EOT,

A.UF\_CLASS,

A.RESIDENCY,

A.LASTUPD\_EW\_DTTM,

A.PROF\_GRAD\_FLAG,

A.ACADEMIC\_LOAD,

A.TOT\_GRADE\_POINTS,

A.TOT\_TAKEN\_GPA,

A.ENRL\_SUMMER\_A\_FLAG,

A.ENRL\_SUMMER\_B\_FLAG,

A.ENRL\_SUMMER\_C\_FLAG,

A.ACAD\_PROG\_PRIMARY,

A.UF\_CLASS\_EOT,

A.UNT\_TAKEN\_GPA,

A.UNT\_INPROG\_GPA,

A.TERM\_FIRST\_APPT\_TIME,

(CASE WHEN substr(A.TERM\_BEG\_DT\_SID,-4)<=0230 AND substr(A.TERM\_END\_DT\_SID,-4)<=0430 THEN( 'EARLY')

WHEN substr(A.TERM\_BEG\_DT\_SID,-4)<=0230 AND substr(A.TERM\_END\_DT\_SID,-4)>0430 AND substr(A.TERM\_END\_DT\_SID,-4)<=0504 THEN( 'AVERAGE')

WHEN substr(A.TERM\_BEG\_DT\_SID,-4)<=0230 AND substr(A.TERM\_END\_DT\_SID,-4)>0504 THEN( 'LATE')

WHEN substr(A.TERM\_BEG\_DT\_SID,-4)>=0800 AND substr(A.TERM\_END\_DT\_SID,-4)<=1216 THEN( 'EARLY')

WHEN substr(A.TERM\_BEG\_DT\_SID,-4)>=0800 AND substr(A.TERM\_END\_DT\_SID,-4)>1216 AND substr(A.TERM\_END\_DT\_SID,-4)<=1220 THEN( 'AVERAGE')

WHEN substr(A.TERM\_BEG\_DT\_SID,-4)>=0800 AND substr(A.TERM\_END\_DT\_SID,-4)>1220 THEN( 'LATE')

ELSE NULL

END )AS "TERM\_END\_DT\_SID\_CATGRY",

(CASE WHEN substr(A.TERM\_BEG\_DT\_SID,-4)<=0103 THEN ('EARLY')

WHEN substr(A.TERM\_BEG\_DT\_SID,-4)>0103 AND substr(A.TERM\_BEG\_DT\_SID,-4)<=0105 THEN ('AVERAGE')

WHEN substr(A.TERM\_BEG\_DT\_SID,-4)>0105 AND substr(A.TERM\_BEG\_DT\_SID,-4)<=0230 THEN ( 'LATE')

WHEN substr(A.TERM\_BEG\_DT\_SID,-4)>0800 AND substr(A.TERM\_BEG\_DT\_SID,-4)<=0821 THEN ('EARLY')

WHEN substr(A.TERM\_BEG\_DT\_SID,-4)>0821 AND substr(A.TERM\_BEG\_DT\_SID,-4)<=0825 THEN ('AVERAGE')

WHEN substr(A.TERM\_BEG\_DT\_SID,-4)>0825 THEN ( 'LATE')

ELSE NULL

END) AS "TERM\_BEG\_DT\_CATGRY",

(CASE WHEN (TO\_DATE(A.TERM\_END\_DT\_SID,'YYYYMMDD')-TO\_DATE(A.TERM\_BEG\_DT\_SID,'YYYYMMDD')) <=107 THEN('SHORT')

WHEN (TO\_DATE(A.TERM\_END\_DT\_SID,'YYYYMMDD')-TO\_DATE(A.TERM\_BEG\_DT\_SID,'YYYYMMDD')) >107 AND (TO\_DATE(A.TERM\_END\_DT\_SID,'YYYYMMDD')-TO\_DATE(A.TERM\_BEG\_DT\_SID,'YYYYMMDD')) <=119 THEN ('AVERAGE')

WHEN (TO\_DATE(A.TERM\_END\_DT\_SID,'YYYYMMDD')-TO\_DATE(A.TERM\_BEG\_DT\_SID,'YYYYMMDD')) >119 THEN('LONG')

END) AS "TERM\_LENGTH\_CATEGORY",

(TO\_DATE(A.TERM\_END\_DT\_SID,'YYYYMMDD')-TO\_DATE(A.TERM\_BEG\_DT\_SID,'YYYYMMDD')) AS "TERM\_LENGTH\_DAYS",

(CASE WHEN substr(A.TERM\_BEG\_DT\_SID,-4)>=0800 THEN ('FALL')

WHEN substr(A.TERM\_BEG\_DT\_SID,-4)<=0230 THEN ('SPRING')

ELSE ('SUMMER')

END) AS "TERM\_SEASON",

B.PERSON\_SID,

B.LOW\_TERM\_GPA\_IND,

B.PARTTIME\_TERM\_IND,

B.NOT\_REG\_TERM\_IND,

B.WITHDRWL\_TERM\_IND,

B.FULLTIME\_TERM\_IND,

B.OVR\_12HR\_TERM\_IND

FROM UF\_SUCCESS\_TARGET B INNER JOIN UF\_B\_STDNT\_TERM A

ON A.PERSON\_SID = B.PERSON\_SID

WHERE A.ACAD\_CAR\_SID = 8

AND A.CUR\_GPA<>0

AND (A.TOT\_PASSD\_PRGRSS-trunc(A.TOT\_PASSD\_PRGRSS/1,0))=0

AND (A.TOT\_TAKEN\_PRGRSS-trunc(A.TOT\_TAKEN\_PRGRSS/1,0))=0

)

CREATE TABLE UF\_R1\_SUCCESS\_ANALYSIS\_0709 AS(SELECT

\* from UF\_R1\_SUCCESS\_ANALYSIS\_UGRD

WHERE TERM\_BEG\_DT\_SID>20070000 AND TERM\_BEG\_DT\_SID<20091231

)

CREATE TABLE UF\_R1\_SUCCESS\_ANALYSIS\_1719 AS(SELECT

\* from UF\_R1\_SUCCESS\_ANALYSIS\_UGRD

WHERE TERM\_BEG\_DT\_SID>20170000 AND TERM\_BEG\_DT\_SID<20191231

### 

### **Appendix 6: SQL Data Dictionary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Table** | **Variable** | **Description** | **Data Type** |
| UF\_D\_ACAD\_CAR | ACAD\_CAR\_SID | Academic Career surrogate identification | NUMBER(10,0) |
| UF\_D\_ACAD\_PLAN | ACAD\_CAR\_SID | Academic Career surrogate identification | NUMBER(10,0) |
| UF\_D\_ACAD\_PROG | ACAD\_CAR\_SID | Academic Career surrogate identification | NUMBER(10,0) |
| UF\_D\_ACAD\_SPLAN | ACAD\_CAR\_SID | Academic Career surrogate identification | NUMBER(10,0) |
| UF\_D\_CLASS | ACAD\_CAR\_SID | Academic Career surrogate identification | NUMBER(10,0) |
| UF\_D\_CRSE | ACAD\_CAR\_SID | Academic Career surrogate identification | NUMBER(10,0) |
| UF\_B\_STDNT\_TERM | ACAD\_CAREER | Describes whether the student is in the undergrad or graduate program. | VARCHAR2(4 CHAR) |
| UF\_B\_TERM\_SPLAN | ACAD\_CAREER |  | VARCHAR2(4 CHAR) |
| UF\_B\_STDNT\_TERM | ACAD\_LEVEL\_BOT |  | VARCHAR2(3 CHAR) |
| UF\_B\_STDNT\_TERM | ACAD\_LEVEL\_EOT |  | VARCHAR2(3 CHAR) |
| UF\_B\_STDNT\_TERM | ACAD\_PROG\_PRIMARY | Which Academic program a student is enrolled in ie UGENG = Undergraduate Engineering | VARCHAR2(5 CHAR) |
| UF\_B\_TERM\_SPLAN | ACAD\_SPLAN\_SID | Academic Sub Plan Surrogate Identification | NUMBER(10,0) |
| UF\_B\_STDNT\_TERM | ACADEMIC\_LOAD | The number of classes a student takes per term, also known as the academic load. | VARCHAR2(1 CHAR) |
| UF\_B\_STDNT\_TERM | AGE\_YEARS | Age in Years of the person for the term | NUMBER(10,0) |
| UF\_B\_STDNT\_TERM | CUM\_GPA | Cumulative GPA for all activity up to the term | NUMBER(8,3) |
| UF\_B\_STDNT\_TERM | CUR\_GPA | Current GPA for the term activity | NUMBER(8,3) |
| UF\_B\_Person\_Student\_GRP | EFF\_END\_TERM |  | VARCHAR2(4 CHAR) |
| UF\_B\_Person\_Student\_GRP | EFF\_END\_TERM\_LD | Effective End PeopleSoft Term Long Description | VARCHAR2(30 CHAR) |
| Create\_New | EFF\_LENGTH\_TERM | =EFF\_END\_TERM : EFF\_START\_TERM | VARCHAR2(4 CHAR) |
| UF\_B\_Person\_Student\_GRP | EFF\_START\_TERM |  | VARCHAR2(4 CHAR) |
| UF\_B\_Person\_Student\_GRP | EFF\_START\_TERM\_LD | Effective Start PeopleSoft Term Long Description | VARCHAR2(30 CHAR) |
| UF\_B\_Person\_Student\_GRP | END\_EFFDT\_SID | unsure what these numbers stand for. | NUMBER(10,0) |
| UF\_B\_STDNT\_TERM | ENRL\_SUMMER\_A\_FLAG | If a student had enrolled in summer A classes for their last reported Mod | VARCHAR2(1 CHAR) |
| UF\_B\_STDNT\_TERM | ENRL\_SUMMER\_B\_FLAG | If a student had enrolled in summer B classes for their last reported Mod | VARCHAR2(1 CHAR) |
| UF\_B\_STDNT\_TERM | ENRL\_SUMMER\_C\_FLAG | If a student had enrolled in summer C classes for their last reported Mod | VARCHAR2(1 CHAR) |
| UF\_D\_ACAD\_CAR | INSTITUTION\_SID | Institution surrogate identification | NUMBER(10,0) |
| UF\_D\_ACAD\_GRP | INSTITUTION\_SID | Institution surrogate identification | NUMBER(10,0) |
| UF\_D\_ACAD\_ORG | INSTITUTION\_SID | Institution surrogate identification | NUMBER(10,0) |
| UF\_D\_ACAD\_PLAN | INSTITUTION\_SID | Institution surrogate identification | NUMBER(10,0) |
| UF\_D\_ACAD\_PROG | INSTITUTION\_SID | Institution surrogate identification | NUMBER(10,0) |
| UF\_D\_ACAD\_SPLAN | INSTITUTION\_SID | Institution surrogate identification | NUMBER(10,0) |
| UF\_D\_CAMPUS | INSTITUTION\_SID | Institution surrogate identification | NUMBER(10,0) |
| UF\_D\_CLASS | INSTITUTION\_SID | Institution surrogate identification | NUMBER(10,0) |
| UF\_D\_CRSE | INSTITUTION\_SID | Institution surrogate identification | NUMBER(10,0) |
| UF\_D\_INSTITUTION | INSTITUTION\_SID | Institution surrogate identification | NUMBER(10,0) |
| UF\_D\_INSTRCTN\_MODE | INSTITUTION\_SID | Institution surrogate identification | NUMBER(10,0) |
| UF\_D\_REQUIREMENT | INSTITUTION\_SID | Institution surrogate identification |  |
| UF\_D\_SRVC\_IMPACT | INSTITUTION\_SID | Institution surrogate identification | NUMBER(38,0) |
| UF\_B\_STDNT\_TERM | JUNIOR\_SENIOR\_FLAG | Flag (Y/N) indicating if Student is a Junior or Senior | VARCHAR2(1 CHAR) |
| UF\_B\_Person\_Student\_GRP | PERSON\_SID | Person surrogate identification | NUMBER(10,0) |
| UF\_B\_STDNT\_TERM | PERSON\_SID | Person surrogate identification | NUMBER(38,0) |
| UF\_B\_TERM\_SPLAN | PERSON\_SID | Person surrogate identification | NUMBER(10,0) |
| UF\_B\_STDNT\_TERM | RESIDENCY |  | VARCHAR2(5 CHAR) |
| UF\_D\_SRVC\_IMPACT | SRVC\_IMPACT\_LD |  | VARCHAR2(30 CHAR) |
| UF\_D\_SRVC\_IMPACT | SRVC\_IMPACT\_SD |  | VARCHAR2(10 CHAR) |
| UF\_B\_STDNT\_TERM | SSR\_TRMAC\_LAST\_DT |  | DATE |
| UF\_B\_Person\_Student\_GRP | STDNT\_GROUP\_SID |  | NUMBER(10,0) |
| UF\_B\_STDNT\_TERM | STRM |  | VARCHAR2(4 CHAR) |
| UF\_B\_STDNT\_TERM | TERM\_BEG\_DT\_SID |  | NUMBER(38,0) |
| Create\_New\_UF\_B\_STDNT\_TERM | TERM\_BEG\_DT\_SID\_FALL\_CATGRY | Categorize fall start dates as Early, Average, or Late compared to other years. By using K Means Clustering on Fall Start Dates in TERM\_BEG\_DT\_SID | Categorical |
| Create\_New\_UF\_B\_STDNT\_TERM | TERM\_BEG\_DT\_SID\_SPRING\_CATGRY | Categorize spring start dates as Early, Average, or Late compared to other years. By using K Means Clustering on Spring Start Dates in TERM\_BEG\_DT\_SID | Categorical |
| UF\_B\_STDNT\_TERM | TERM\_END\_DT\_SID |  | NUMBER(38,0) |
| Create\_New\_UF\_B\_STDNT\_TERM | TERM\_END\_DT\_SID\_FALL\_EARLY | Categorize fall ends dates as Early, Average, or Late compared to other years. By using K Means Clustering on Fall Start Dates in TERM\_END\_DT\_SID | Categorical |
| Create\_New\_UF\_B\_STDNT\_TERM | TERM\_END\_DT\_SID\_SPRING\_EARLY | Categorize spring end dates as Early, Average, or Late compared to other years. By using K Means Clustering on Spring Start Dates in TERM\_END\_DT\_SID | Categorical |
| Create\_New\_UF\_B\_STDNT\_TERM | TERM\_LENGTH\_CATEGORY | Term Length in categorical buckets of Short, Average, and Long determined by K Means Clustering | Number |
| Create\_New\_UF\_B\_STDNT\_TERM | TERM\_LENGTH\_DAYS | Term Length in the number of days | Number |
| Create\_New\_UF\_B\_STDNT\_TERM | TERM\_SEASON | =IF(OR(TERM\_SID, "Summer...", Summer, ("Spring...", Spring), ("Fall...", Fall) | Categorical |
| UF\_B\_STDNT\_TERM | TERM\_SID | Term surrogate identification | NUMBER(38,0) |
| UF\_B\_STDNT\_TERM | TOT\_GRADE\_POINTS | Total Grade Points | NUMBER(9,3) |
| UF\_B\_STDNT\_TERM | TOT\_PASSD\_PRGRSS | Total number of credits a student has passed | NUMBER(8,3) |
| Create\_New\_UF\_B\_STDNT\_TERM | TOT\_REQUIRED\_PRGRSS\_*PROGRAMNAME* | The number of credit hours required for the individual undergrad program listed in "Program Name" in the variable: will likely be 50+ new variables | Number |
| UF\_B\_STDNT\_TERM | TOT\_TAKEN\_GPA |  | NUMBER(8,3) |
| UF\_B\_STDNT\_TERM | TOT\_TAKEN\_PRGRSS | The total number of credit hours a student has taken? Amanda asking Andrew to clarify. | NUMBER(22,8) |
| UF\_B\_STDNT\_TERM | TOT\_TEST\_CREDIT | The total number of credits a student tested out of. | NUMBER(22,8) |
| UF\_B\_STDNT\_TERM | TOT\_TRNSFR | The total number of transfer credits | NUMBER(22,8) |
| UF\_B\_STDNT\_TERM | UF\_CLASS |  | VARCHAR2(1 CHAR) |
| UF\_B\_STDNT\_TERM | UF\_CLASS\_EOT |  | VARCHAR2(1 CHAR) |
| **UF R1 SUCCESS ANALYSIS:** |  |  |  |
| UF\_R1\_SUCCESS\_ANALYSIS\_UNDERGRAD: |  |  |  |
| UF\_R1\_SUCCESS\_ANALYSIS\_0709: |  |  |  |
| **UF\_R1\_SUCCESS\_ANALYSIS1719:** |  |  |  |
| UF\_SUCCESS\_INDICATORS: |  |  |  |

## 

### **Appendix 7: Python Quality of the Data Report UF\_R1\_SUCCESS\_ANALYSIS\_UGRD**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

plt.style.use('ggplot')

import datetime as dt

%matplotlib inline

uf = pd.read\_csv('UF\_R1\_SUCCESS\_ANALYSIS.csv')

df2=uf.loc[:,['CUR\_GPA']]

df3=uf.loc[:,['DATE\_OF\_BIRTH']]

df4=uf.loc[:,['STRM']]

df5=uf.loc[:,['TOT\_PASSD\_PRGRSS']]

df6=uf.loc[:,['TOT\_TAKEN\_PRGRSS']]

df7=uf.loc[:,['TOT\_TEST\_CREDIT']]

sns.violinplot(data=uf,y=df2)

plt.show()

df2.hist(figsize=(10, 15), bins=50, xlabelsize=8, ylabelsize=8)

plt.show()

f, ax = plt.subplots(figsize=(10,15))

sns.countplot(x="DATE\_OF\_BIRTH", data=uf, palette="Greens\_d")

DOB=uf['DATE\_OF\_BIRTH'].value\_counts()

sns.violinplot(data=uf,y=df4)

plt.show()

df4.hist(figsize=(10, 15), bins=50, xlabelsize=8, ylabelsize=8)

plt.show()

sns.violinplot(data=uf,y=df5)

plt.show()

df5.hist(figsize=(10, 15), bins=50, xlabelsize=8, ylabelsize=8)

plt.show()

sns.violinplot(data=uf,y=df6)

plt.show()

df6.hist(figsize=(10, 15), bins=50, xlabelsize=8, ylabelsize=8)

plt.show()

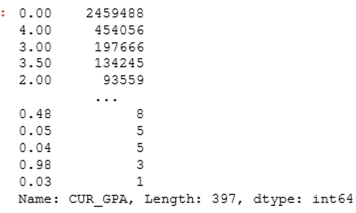
sns.violinplot(data=uf,y=df7)

plt.show()

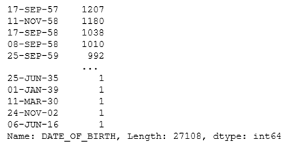
df7.hist(figsize=(10, 15), bins=50, xlabelsize=8, ylabelsize=8)

plt.show()

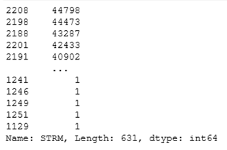
CUR\_GPA：

****

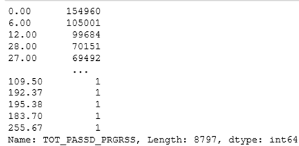
DATE\_OF\_BIRTH:

****

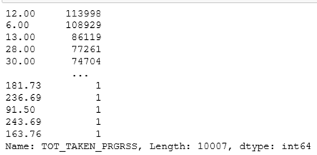
STRM:

****

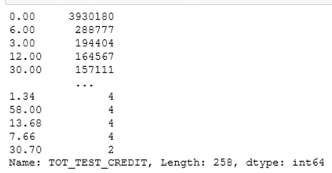
TOT\_PASSD\_PRGRSS:

****

TOT\_TAKEN\_PRGRSS:

****

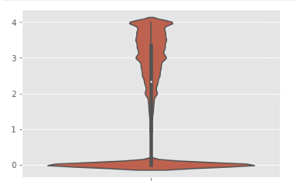
TOT\_TEST\_CREDIT:



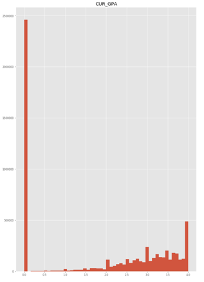
## 

### **Appendix 8: Python Visualizations on the Quality of the Data**

CUR\_GPA Violin Plot:   
Y-axis: All values in this variable, X-axis: Wilder places mean the larger amounts.



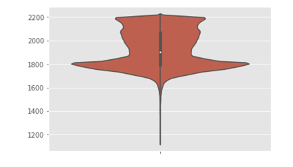
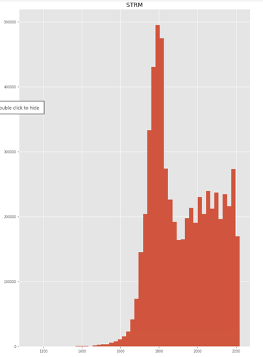
CUR\_GPA Histogram:   
Y-axis: count of one value in this variable, X-axis: All values in this variable.



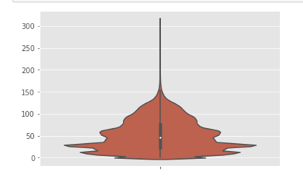
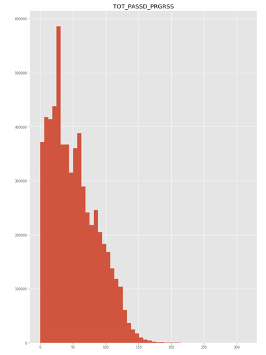
DATE\_OF\_BIRTH: CountPlot

## 

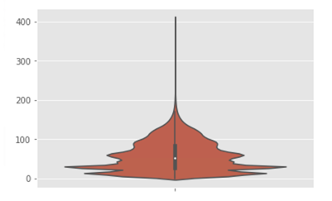
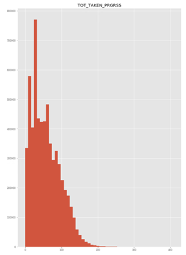
STRM:Histogram & Violin Plot



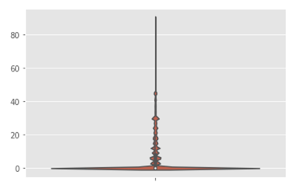
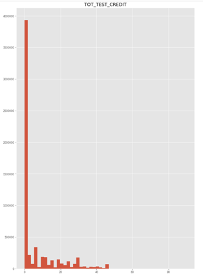
TOT\_PASSD\_PRGRSS:Histogram & Violin Plot



TOT\_TAKEN\_PRGRSS: :Histogram & Violin Plot

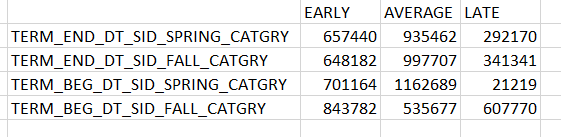


TOT\_TEST\_CREDIT: :Histogram & Violin Plot



### **Appendix 9: SQL Summary Counts of New Variables Created Using KMeans Term\_End & Term\_BEG**

By creating new variables that categorized the start and end of the Fall and Spring terms as “Early”, “Average” or “Late” using a K Means methodology it is able to be determined if the term calendar had an impact on the likelihood for student success.



As you can see from the above chart:

The Fall Semester was well distributed between early, average, and late start dates while leaning slightly on the early side. In contrast, the Spring was much more consistent and the majority of start dates fell between either the early or average category with a very small minority of years having a late Spring term start date.

In comparison regardless of start or end dates the end of the term tended to fall around the same average time period. When there were variations the term-end dates tended to lean on the early side.

### 

### **Appendix 10: SQL UF\_R2\_ANALYSIS\_UGRDDRAFT**

CREATE TABLE UF\_R2\_ANALYSIS\_UGRDDRAFT as (SELECT

A.PERSON\_SID,

A.ACAD\_CAR\_SID,

A.INSTITUTION\_SID,

A.STDNT\_CAR\_NBR,

A.TERM\_SID,

A.STRM,

A.TERM\_BEG\_DT\_SID,

A.TERM\_END\_DT\_SID,

A.INSTITUTION,

A.ACAD\_CAREER,

A.AGE\_YEARS,

A.AGE\_MONTHS,

A.AGE\_DAYS,

A.TOT\_CUMULATIVE,

A.JUNIOR\_SENIOR\_FLAG,

A.TOT\_TAKEN\_PRGRSS,

A.TOT\_TRNSFR,

A.TOT\_TEST\_CREDIT,

A.TOT\_OTHER,

A.TOT\_PASSD\_PRGRSS,

A.UNT\_TAKEN\_PRGRSS,

A.CUR\_GPA,

A.CUM\_GPA,

A.ENRL\_CNT,

A.ENRL\_FLG,

A.SSR\_TRMAC\_LAST\_DT,

A.ACAD\_LEVEL\_BOT,

A.ACAD\_LEVEL\_EOT,

A.UF\_CLASS,

A.RESIDENCY,

A.LASTUPD\_EW\_DTTM,

A.PROF\_GRAD\_FLAG,

A.ACADEMIC\_LOAD,

A.TOT\_GRADE\_POINTS,

A.TOT\_TAKEN\_GPA,

A.ENRL\_SUMMER\_A\_FLAG,

A.ENRL\_SUMMER\_B\_FLAG,

A.ENRL\_SUMMER\_C\_FLAG,

A.ACAD\_PROG\_PRIMARY,

A.UF\_CLASS\_EOT,

A.UNT\_TAKEN\_GPA,

A.UNT\_INPROG\_GPA,

A.TERM\_FIRST\_APPT\_TIME,

A.TERM\_END\_DT\_SID\_CATGRY,

A.TERM\_BEG\_DT\_CATGRY,

A.TERM\_LENGTH\_CATEGORY,

A.TERM\_LENGTH\_DAYS,

A.TERM\_SEASON,

A.LOW\_TERM\_GPA\_IND,

A.PARTTIME\_TERM\_IND,

A.NOT\_REG\_TERM\_IND,

A.WITHDRWL\_TERM\_IND,

A.FULLTIME\_TERM\_IND,

A.OVR\_12HR\_TERM\_IND,

B.STDNT\_GROUP\_SID,

B.EFFDT\_SID,

B.END\_EFFDT\_SID,

B.EFF\_STATUS,

B.EFF\_START\_TERM,

B.EFF\_START\_TERM\_LD,

B.EFF\_END\_TERM,

B.EFF\_END\_TERM\_LD,

B.CURRENT\_IND,

B.LASTUPDOPRID,

B.COMMENTS,

C.ACAD\_SPLAN\_SID

FROM UF\_R1\_SUCCESS\_ANLS\_UDINTSMALL A INNER JOIN UF\_B\_PERSON\_STDNT\_GRP B

ON A.PERSON\_SID = B.PERSON\_SID

AND ((A.TERM\_SEASON = substr(B.EFF\_START\_TERM\_LD,0,4) AND substr(A.TERM\_BEG\_DT\_SID,0,4) = substr(B.EFF\_START\_TERM\_LD,-4))

OR (A.TERM\_SEASON = substr(B.EFF\_END\_TERM\_LD,0,4)AND substr(A.TERM\_BEG\_DT\_SID,0,4) = substr(B.EFF\_END\_TERM\_LD,-4)))

INNER JOIN UF\_B\_TERM\_SPLAN C

ON A.PERSON\_SID = C.PERSON\_SID

AND A.TERM\_SID = C.TERM\_SID

)

CREATE TABLE UF\_R2\_ANALYSIS\_UNDERGRAD as (SELECT

\*

FROM UF\_R2\_ANALYSIS\_UNDERGRADDRAFT A

WHERE (

A.EFFDT\_SID = (SELECT MAX(B.EFFDT\_SID)

FROM UF\_R2\_ANALYSIS\_UNDERGRADDRAFT B

WHERE A.PERSON\_SID = B.PERSON\_SID

AND A.STDNT\_GROUP\_SID = B.STDNT\_GROUP\_SID

AND A.STDNT\_CAR\_NBR = B.STDNT\_CAR\_NBR

)

)

)

### 

### **Appendix 11: Student\_term\_enrollment\_model\_v3**

