# Data Tidying Activity

The data for this project is set at the grain level of a single student for a single term. I am aggregating information about prior registrations, charges, aid and loans into single row of data for that student to create a profile of what their current prior behavior up to that point is. The associated excel file has a set of data definitions, but this document should serve as a supplement to that.

To make the data more manageable, I have limited the selection to Master’s level students enrolled in classes from Fall 2010 through Spring 2013 in the Arts and Humanities Department. Additionally to remove records that could throw off the analysis significantly, I have eliminated any students who prior to the term have completed a degree at the college. The database is not easily configured to handle multiple degrees, so if a student had gotten one Master’s degree from TC five years ago and is now returning for a second, the Possible Terms Registered would record over 10 terms of missing registration.

Lastly, the CBSA field in the excel file is not reflected in the SQL listed below, but instead was created from a VLOOKUP function against the CBSA Data tab in the excel file.

## SQL

SELECT stvterm\_code,

substr(stvterm\_code,6,1) termtype,

sgbstdn\_majr\_code\_1,

sgbstdn\_degc\_code\_1,

(SELECT sum(rpratrm\_paid\_amt)

FROM rpratrm

INNER JOIN rfrbase ON rpratrm\_fund\_code = rfrbase\_fund\_code

WHERE rpratrm\_pidm = sgbstdn\_pidm

AND rpratrm\_period <= stvterm\_code

AND rfrbase\_ftyp\_code = 'LOAN') tc\_loans,

(SELECT sum(rpratrm\_paid\_amt)

FROM rpratrm

INNER JOIN rfrbase ON rpratrm\_fund\_code = rfrbase\_fund\_code

WHERE rpratrm\_pidm = sgbstdn\_pidm

AND rpratrm\_period <= stvterm\_code

AND rfrbase\_fsrc\_code IN ('INST','ENDW')

AND rfrbase\_ftyp\_code NOT IN ('LOAN','WORK')) tc\_aid,

(SELECT sum(tbraccd\_amount)

FROM tbraccd

INNER JOIN tbbdetc ON tbraccd\_detail\_code= tbbdetc\_detail\_code

WHERE tbraccd\_pidm = sgbstdn\_pidm

AND tbraccd\_term\_code <= stvterm\_code

AND tbbdetc\_type\_ind = 'C') total\_charges,

smbpgen\_req\_credits\_overall,

(SELECT sum(sfrstcr\_bill\_hr)

FROM sfrstcr

INNER JOIN ssbsect ON sfrstcr\_term\_code = ssbsect\_term\_code

AND sfrstcr\_crn = ssbsect\_crn

WHERE sfrstcr\_pidm = sgbstdn\_pidm

AND sfrstcr\_term\_code <= stvterm\_code

AND sfrstcr\_rsts\_code LIKE 'R%') credits\_so\_far,

(SELECT count(term.stvterm\_code)

FROM stvterm term

CROSS JOIN sgbstdn stdn

WHERE substr(term.stvterm\_code,5,2) IN ('01','09')

AND stdn.sgbstdn\_pidm = sgbstdn.sgbstdn\_pidm

AND stdn.sgbstdn\_term\_code\_eff = (SELECT MIN(dn.sgbstdn\_term\_code\_eff)

FROM sgbstdn dn

WHERE dn.sgbstdn\_pidm = stdn.sgbstdn\_pidm

)

AND term.stvterm\_code BETWEEN stdn.sgbstdn\_term\_code\_eff AND stvterm.stvterm\_code) possible\_reg\_terms,

(SELECT count(DISTINCT sfrstcr\_term\_code)

FROM sfrstcr

INNER JOIN sgbstdn stdn ON sfrstcr\_pidm = stdn.sgbstdn\_pidm

WHERE sfrstcr\_pidm = sgbstdn.sgbstdn\_pidm

AND stdn.sgbstdn\_term\_code\_eff = (SELECT MIN(dn.sgbstdn\_term\_code\_eff)

FROM sgbstdn dn

WHERE dn.sgbstdn\_pidm = stdn.sgbstdn\_pidm

)

AND substr(sfrstcr\_term\_code,5,2) IN ('01','09')

AND sfrstcr\_rsts\_code LIKE 'R%'

AND sfrstcr\_term\_code BETWEEN stdn.sgbstdn\_term\_code\_eff AND stvterm.stvterm\_code) registered\_terms,

NULL reqclassoffnextterm,

spbpers\_citz\_code,

round(months\_between(stvterm\_start\_date, spbpers\_birth\_date)/12,0) ageattermstart,

nvl((SELECT DISTINCT 'Y'

FROM shrdgmr

WHERE shrdgmr\_pidm = sgbstdn\_pidm

AND shrdgmr\_term\_code\_grad =stvterm\_code

AND shrdgmr\_appl\_date <= stvterm\_start\_date + 20

AND shrdgmr\_degs\_code IN ('DA','AF','IE','CP')),'N') appliedtograd,

nvl("AmerIndianAlaskaNative",0) amerindianalaskanative,

nvl("Asian",0) asian,

nvl("BlackAfrAmerican",0) blackaframerican,

nvl("NativeHawaOthPI",0) nativehawaothpi,

nvl("White",0) white,

spbpers\_sex,

(SELECT MIN(spraddr\_cnty\_code)

FROM spraddr

WHERE spraddr.ROWID = f\_get\_address\_rowid(sgbstdn\_pidm, 'STDNADDR', 'A', stvterm\_start\_date + 10 ,1,'S','')

) county,

nvl((SELECT DISTINCT 'Y'

FROM sfrstcr

WHERE sfrstcr\_pidm = sgbstdn\_pidm

AND sfrstcr\_term\_code = CASE

WHEN substr(stvterm\_code,5,2) IN ('01','05')

THEN substr(stvterm\_code,1,4) || '09'

ELSE to\_char(to\_number(substr(stvterm\_code,1,4))+1) ||'01'

END

AND sfrstcr\_rsts\_code LIKE 'R%'

),'N') reg\_next\_fall\_spring

FROM stvterm

CROSS JOIN (sgbstdn

INNER JOIN spbpers ON sgbstdn\_pidm = spbpers\_pidm

INNER JOIN smbpgen ON sgbstdn\_program\_1 = smbpgen\_program

LEFT OUTER JOIN (SELECT \*

FROM (SELECT gorprac\_pidm,

gorprac\_race\_cde,

1 counter

FROM gorprac

)

pivot

( MAX(counter)

FOR gorprac\_race\_cde IN (1 AS "AmerIndianAlaskaNative",

2 AS "Asian",

3 AS "BlackAfrAmerican",

4 AS "NativeHawaOthPI",

5 AS "White")

)

) races ON sgbstdn\_pidm = races.gorprac\_pidm

)

WHERE 1=1

AND sgbstdn\_degc\_code\_1 IN ('ME','MA','MS')

AND stvterm\_code BETWEEN '200909' AND '201301'

AND sgbstdn\_dept\_code = 'A&H'

AND sgbstdn\_majr\_code\_conc\_1 IS NULL

AND substr(stvterm\_code,5,2) IN ('01','09')

AND sgbstdn\_stst\_code IN ('AS','LD','LP','LM')

AND sgbstdn\_term\_code\_eff = (SELECT MAX(stdn.sgbstdn\_term\_code\_eff)

FROM sgbstdn stdn

WHERE stdn.sgbstdn\_pidm = sgbstdn.sgbstdn\_pidm

AND stdn.sgbstdn\_term\_code\_eff <= stvterm\_code)

AND smbpgen\_term\_code\_eff = (SELECT MAX(pgen.smbpgen\_term\_code\_eff)

FROM smbpgen pgen

WHERE pgen.smbpgen\_program = smbpgen.smbpgen\_program

AND pgen.smbpgen\_term\_code\_eff<= '201601')

AND nvl((SELECT DISTINCT 'Y'

FROM sfrstcr

WHERE sfrstcr\_pidm = sgbstdn\_pidm

AND sfrstcr\_rsts\_code LIKE 'R%'

AND sfrstcr\_term\_code = stvterm\_code),'N') = 'Y'

AND NOT EXISTS (SELECT 'X'

FROM shrdgmr

WHERE shrdgmr\_pidm = sgbstdn\_pidm

AND shrdgmr\_term\_code\_grad < stvterm\_code

AND shrdgmr\_degs\_code = 'DA')

ORDER BY sgbstdn\_pidm, stvterm\_code;