**HAP 780-001**

**MIMIC III DATASET EXPLORATION IN SQL AND WEKA**

Abhishek Shambhu

**Prediction of Mortality for Patients admitted in first 24 hours to hospital based on top 25 Lab Events**

**Source Code:**

---1. Preprocessing the Admissions table

--- so as to have only one random HADM\_ID for one SUBJECT\_ID.

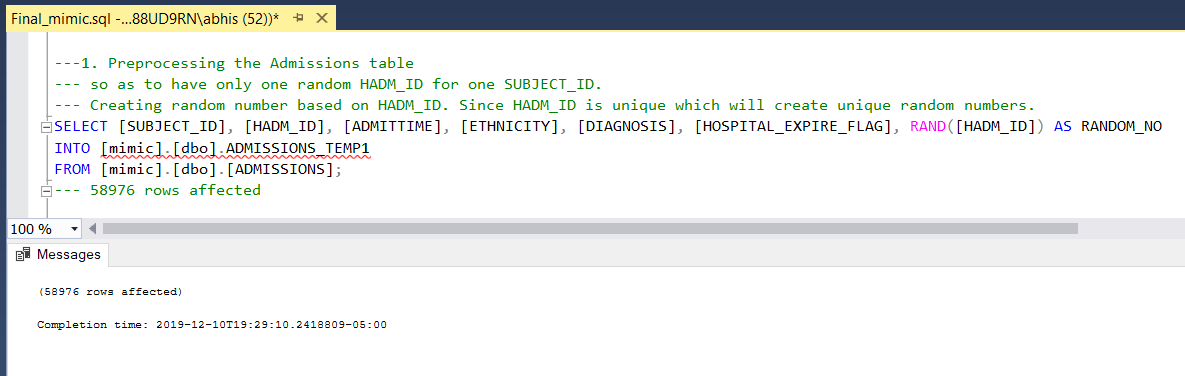
--- Creating random number based on HADM\_ID. Since HADM\_ID is unique which will create unique random numbers.

SELECT [SUBJECT\_ID], [HADM\_ID], [ADMITTIME], [ETHNICITY], [DIAGNOSIS], [HOSPITAL\_EXPIRE\_FLAG], RAND([HADM\_ID]) AS RANDOM\_NO

INTO [mimic].[dbo].ADMISSIONS\_TEMP1

FROM [mimic].[dbo].[ADMISSIONS];

--- 58976 rows affected



--- Assigning Min random number to each subject\_id

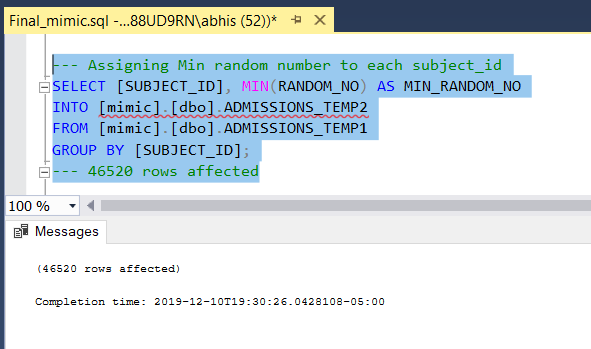
SELECT [SUBJECT\_ID], MIN(RANDOM\_NO) AS MIN\_RANDOM\_NO

INTO [mimic].[dbo].ADMISSIONS\_TEMP2

FROM [mimic].[dbo].ADMISSIONS\_TEMP1

GROUP BY [SUBJECT\_ID];

--- 46520 rows affected



--- Combining the two tables based on Min Random Number and performing cleaning of [HOSPITAL\_EXPIRE\_FLAG] and [DIAGNOSIS]

SELECT T1.[SUBJECT\_ID], [HADM\_ID], [ADMITTIME], [ETHNICITY], [DIAGNOSIS], [HOSPITAL\_EXPIRE\_FLAG]

INTO [mimic].[dbo].ADMISSIONS\_PROCESSED

FROM [mimic].[dbo].ADMISSIONS\_TEMP1 T1

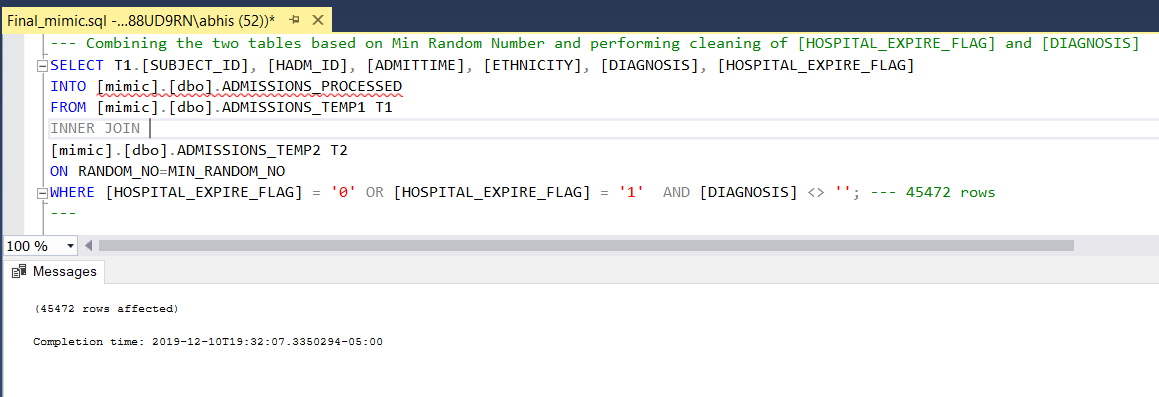
INNER JOIN

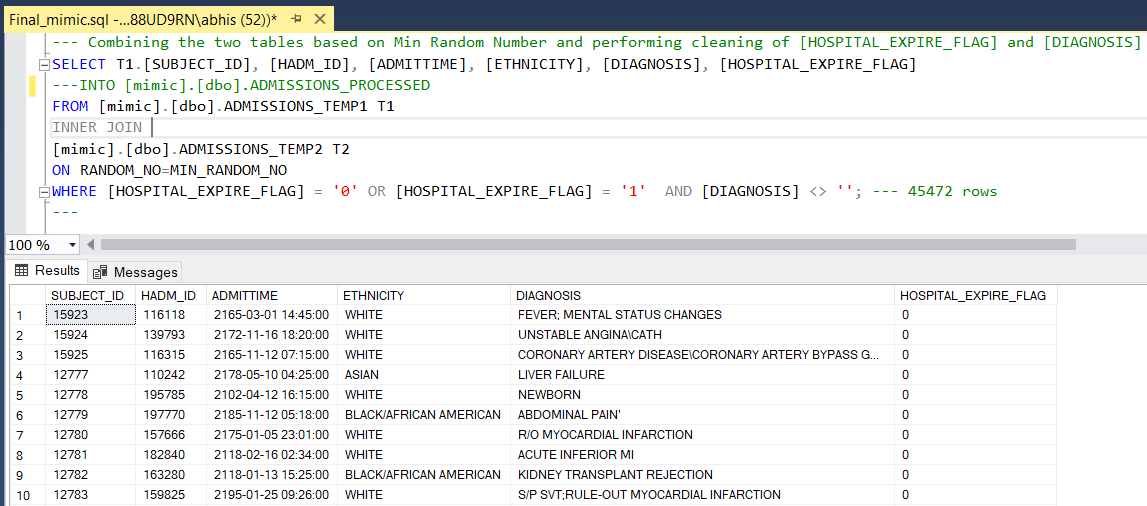
[mimic].[dbo].ADMISSIONS\_TEMP2 T2

ON RANDOM\_NO=MIN\_RANDOM\_NO

WHERE [HOSPITAL\_EXPIRE\_FLAG] = '0' OR [HOSPITAL\_EXPIRE\_FLAG] = '1' AND [DIAGNOSIS] <> ''; --- 45472 rows

---



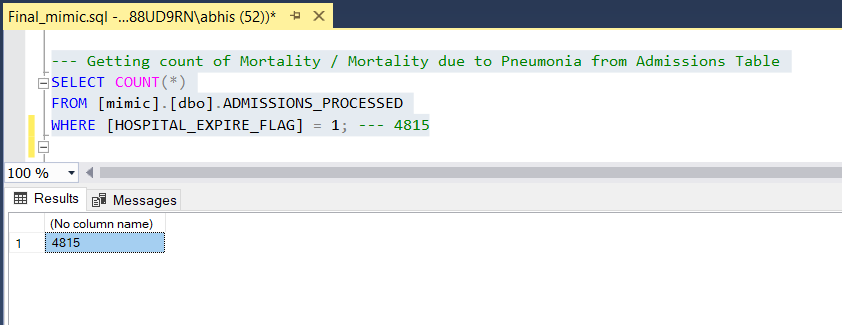


--- Getting count of Mortality from Admissions Table

SELECT COUNT(\*)

FROM [mimic].[dbo].ADMISSIONS\_PROCESSED

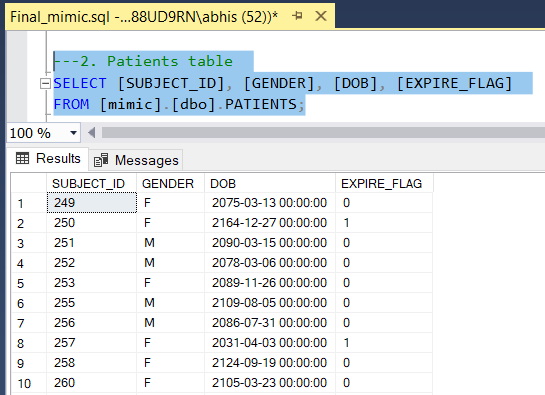
WHERE [HOSPITAL\_EXPIRE\_FLAG] = 1; --- 4815



---2. Patients table

SELECT [SUBJECT\_ID], [GENDER], [DOB], [EXPIRE\_FLAG]

FROM [mimic].[dbo].PATIENTS;



---3. Joining Processed Admissions Table with the Patients Table

SELECT ap.[SUBJECT\_ID], [HADM\_ID], [ADMITTIME], [ETHNICITY], [DIAGNOSIS], [HOSPITAL\_EXPIRE\_FLAG], [GENDER], [DOB], [EXPIRE\_FLAG]

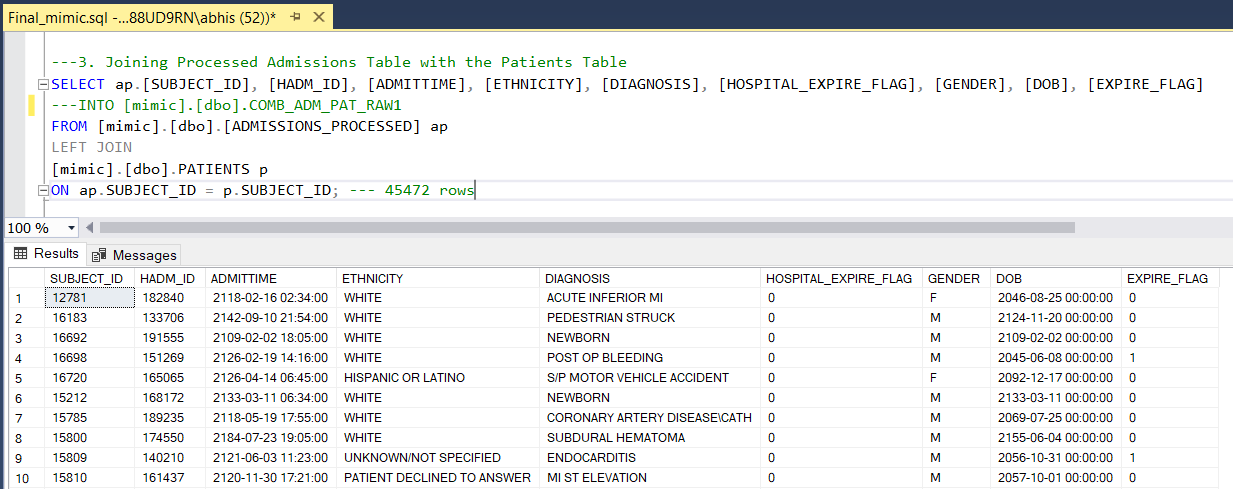
INTO [mimic].[dbo].COMB\_ADM\_PAT\_RAW1

FROM [mimic].[dbo].[ADMISSIONS\_PROCESSED] ap

LEFT JOIN

[mimic].[dbo].PATIENTS p

ON ap.SUBJECT\_ID = p.SUBJECT\_ID; --- 45472 rows



---4. Since we have DOB and ADMITTIME, we calculate the Age

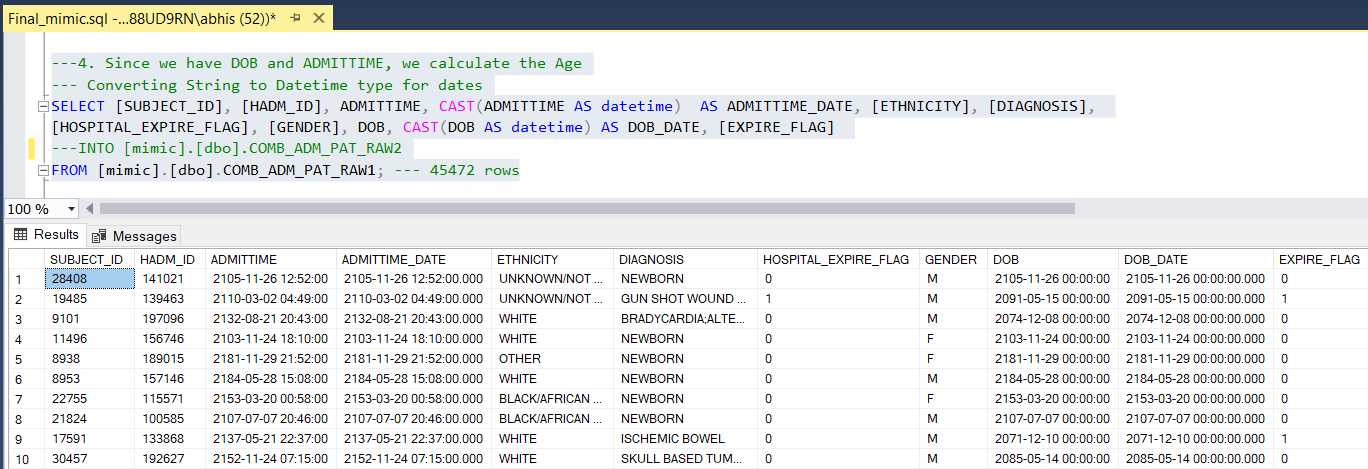
--- Converting String to Datetime type for dates

SELECT [SUBJECT\_ID], [HADM\_ID], ADMITTIME, CAST(ADMITTIME AS datetime) AS ADMITTIME\_DATE, [ETHNICITY], [DIAGNOSIS],

[HOSPITAL\_EXPIRE\_FLAG], [GENDER], DOB, CAST(DOB AS datetime) AS DOB\_DATE, [EXPIRE\_FLAG]

INTO [mimic].[dbo].COMB\_ADM\_PAT\_RAW2

FROM [mimic].[dbo].COMB\_ADM\_PAT\_RAW1; --- 45472 rows



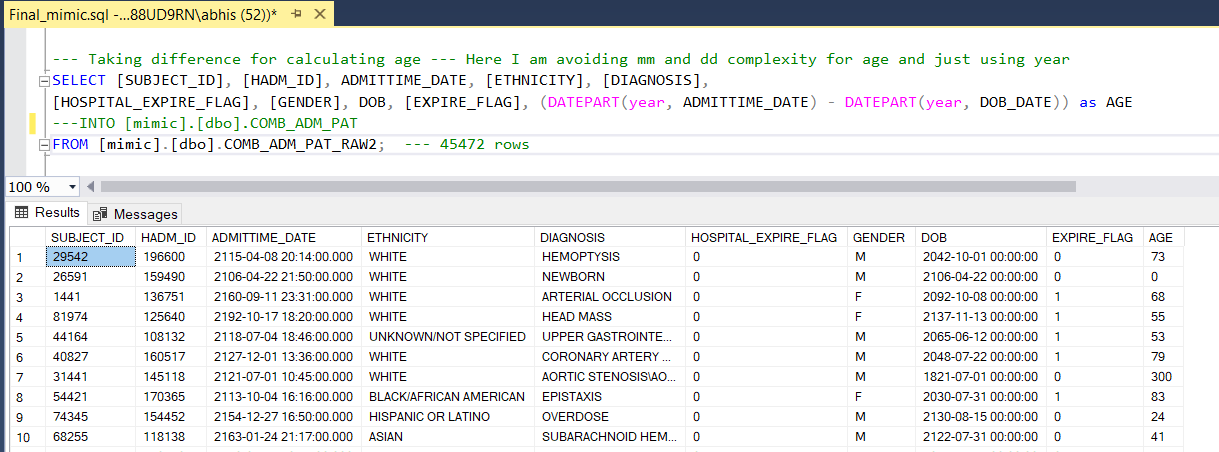
--- Taking difference for calculating age --- Here I am avoiding mm and dd complexity for age and just using year

SELECT [SUBJECT\_ID], [HADM\_ID], ADMITTIME\_DATE, [ETHNICITY], [DIAGNOSIS],

[HOSPITAL\_EXPIRE\_FLAG], [GENDER], DOB, [EXPIRE\_FLAG], (DATEPART(year, ADMITTIME\_DATE) - DATEPART(year, DOB\_DATE)) as AGE

INTO [mimic].[dbo].COMB\_ADM\_PAT

FROM [mimic].[dbo].COMB\_ADM\_PAT\_RAW2; --- 45472 rows



--- Now since we have Age we go forward to assign classes to Age (Age category)

--- 0-1 --> Neonate 15-59 --> Adult 60-89 --> Senior Adult 89+ --> >89

SELECT [SUBJECT\_ID], [HADM\_ID], ADMITTIME\_DATE, [ETHNICITY], [DIAGNOSIS], [HOSPITAL\_EXPIRE\_FLAG], [GENDER], DOB, [EXPIRE\_FLAG], [AGE]

, CASE

-- all ages > 89 in the database were replaced with 300

WHEN [AGE] <= 1

THEN 'NEONATE' --- 7874 rows

WHEN [AGE] > 14 AND [AGE] <=59

THEN 'ADULT' --- 14017 rows

WHEN [AGE] > 59 AND [AGE] <=89

THEN 'SENIOR ADULT' --- 21645 rows

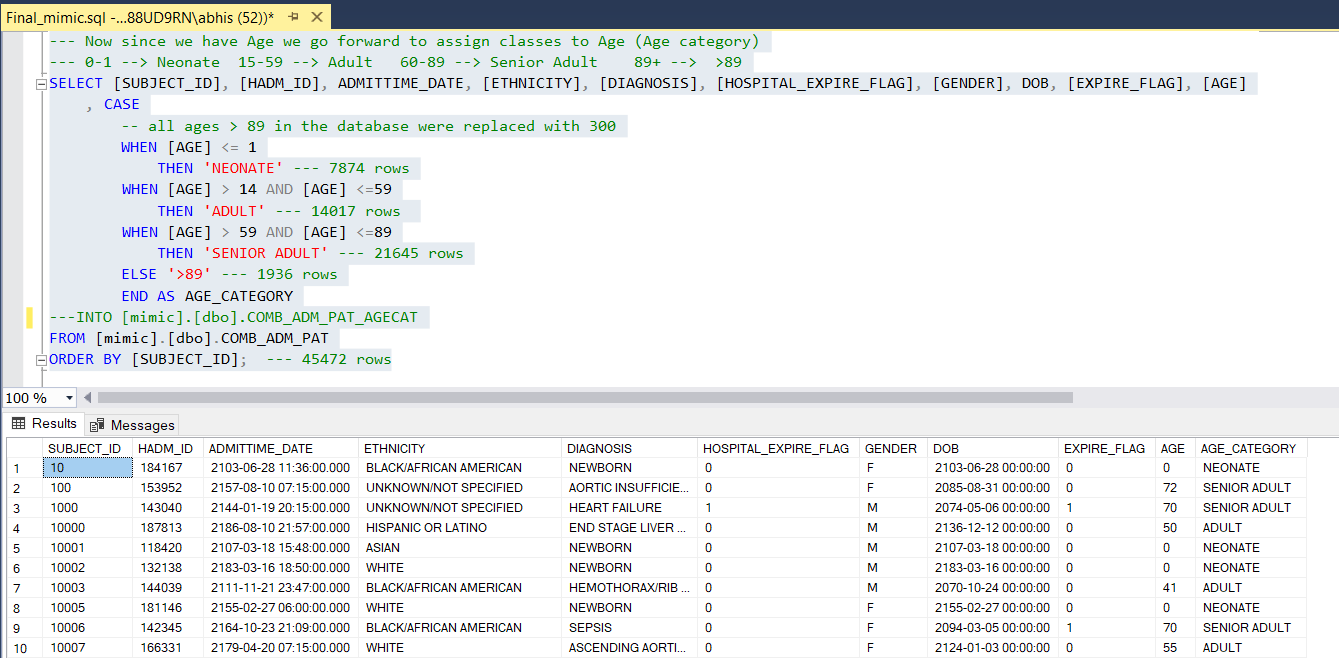
ELSE '>89' --- 1936 rows

END AS AGE\_CATEGORY

INTO [mimic].[dbo].COMB\_ADM\_PAT\_AGECAT

FROM [mimic].[dbo].COMB\_ADM\_PAT

ORDER BY [SUBJECT\_ID]; --- 45472 rows



--- Reducing 41 Ethnicity classes to top 4 classes

SELECT [SUBJECT\_ID], [HADM\_ID], [GENDER], [DOB], [ADMITTIME\_DATE], [AGE], [AGE\_CATEGORY],

CASE WHEN lower(ethnicity) like '%white%'

THEN 'WHITE'

WHEN lower(ethnicity) like '%black%'

THEN 'BLACK'

WHEN lower(ethnicity) like '%asian%'

THEN 'ASIAN'

WHEN lower(ethnicity) like '%hispanic%'

THEN 'HISPANIC/LATINO'

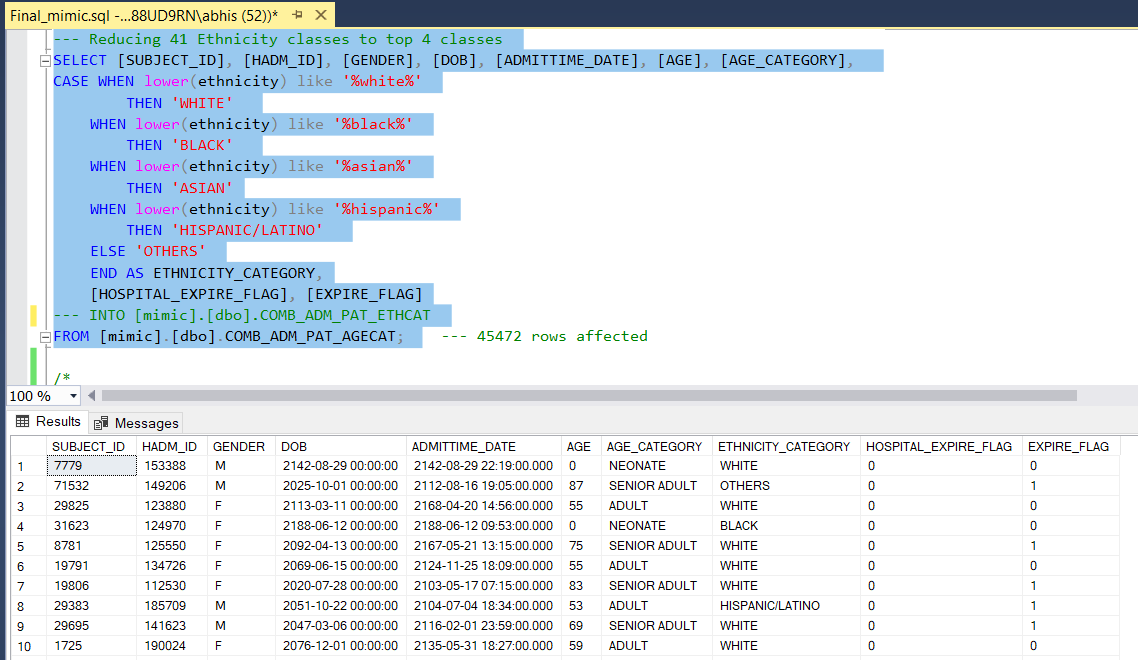
ELSE 'OTHERS'

END AS ETHNICITY\_CATEGORY,

[HOSPITAL\_EXPIRE\_FLAG], [EXPIRE\_FLAG]

INTO [mimic].[dbo].COMB\_ADM\_PAT\_ETHCAT

FROM [mimic].[dbo].COMB\_ADM\_PAT\_AGECAT; --- 45472 rows affected



/\*

select [SUBJECT\_ID], [HADM\_ID], [GENDER], [DOB], [ADMITTIME\_DATE], [AGE],

CASE WHEN [AGE\_CATEGORY]='NEONATE' THEN 1 ELSE 0 END as 'NEONATE',

CASE WHEN [AGE\_CATEGORY]='ADULT' THEN 1 ELSE 0 END as 'ADULT',

CASE WHEN [AGE\_CATEGORY]='SENIOR ADULT' THEN 1 ELSE 0 END as 'SENIOR ADULT',

CASE WHEN [AGE\_CATEGORY]='>89' THEN 1 ELSE 0 END as '>89',

CASE WHEN [ETHNICITY\_CATEGORY]='WHITE' THEN 1 ELSE 0 END as 'WHITE',

CASE WHEN [ETHNICITY\_CATEGORY]='BLACK' THEN 1 ELSE 0 END as 'BLACK',

CASE WHEN [ETHNICITY\_CATEGORY]='ASIAN' THEN 1 ELSE 0 END as 'ASIAN',

CASE WHEN [ETHNICITY\_CATEGORY]='HISPANIC/LATINO' THEN 1 ELSE 0 END as 'HISPANIC/LATINO',

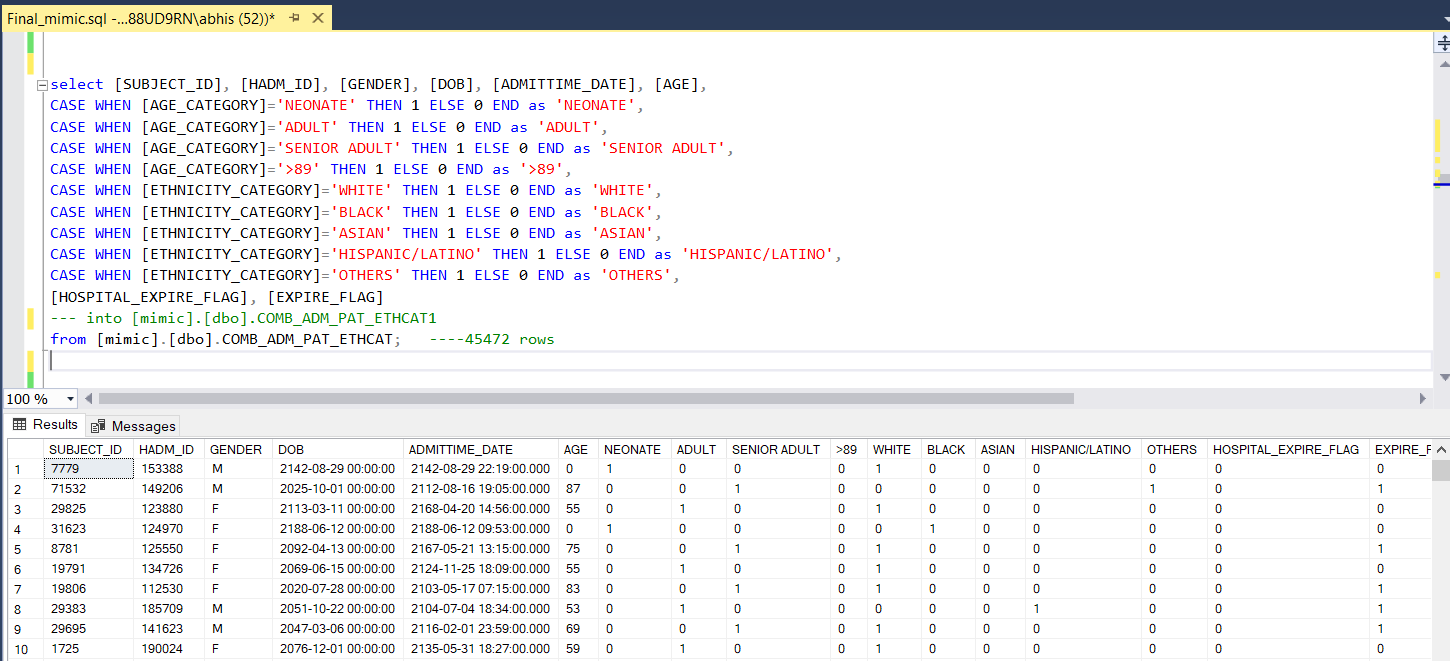
CASE WHEN [ETHNICITY\_CATEGORY]='OTHERS' THEN 1 ELSE 0 END as 'OTHERS',

[HOSPITAL\_EXPIRE\_FLAG], [EXPIRE\_FLAG]

into [mimic].[dbo].COMB\_ADM\_PAT\_ETHCAT1

from [mimic].[dbo].COMB\_ADM\_PAT\_ETHCAT; ----45472 rows

\*/



Select \* from [mimic].[dbo].COMB\_ADM\_PAT\_ETHCAT; --- 45472 rows

-----------------------------------------------------------------------------------------------------------

--- Taking Lab events of patients

SELECT \* FROM [mimic].[dbo].[LABEVENTS]

/\*

--- 1% of the Lab events data

select \* from [mimic].[dbo].[LABEVENTS] tablesample(1 percent) --- Not as random as newid() but will give fast results --- 275701 rows

--- select top 1 percent \* from [mimic].[dbo].[LABEVENTS] order by newid()

--- Creating LABEVENTS\_SAMPLE\_RAW(1%) file

select \* INTO [mimic].[dbo].LABEVENTS\_SAMPLE\_RAW from [mimic].[dbo].[LABEVENTS] tablesample(1 percent) --- 274089 rows affected

--- select top 1 percent \* INTO LABEVENTS\_SAMPLE\_RAW from [mimic].[dbo].[LABEVENTS] order by newid()

select \* FROM [mimic].[dbo].LABEVENTS\_SAMPLE\_RAW;

\*/

--- Preprocessing LabEvents data

--- To remove quotes

SET QUOTED\_IDENTIFIER OFF

UPDATE [mimic].[dbo].[LABEVENTS]

SET [VALUEUOM] = REPLACE([VALUEUOM],"""","")

SET QUOTED\_IDENTIFIER ON --- 27854055 rows affected

SET QUOTED\_IDENTIFIER OFF

UPDATE [mimic].[dbo].[LABEVENTS]

SET VALUENUM = REPLACE(VALUENUM,"'","")

SET QUOTED\_IDENTIFIER ON --- 27854055 rows affected

--- 1. Considering only in patient data(Removing all nulls from HADM\_ID & nulls from Valuenum column)

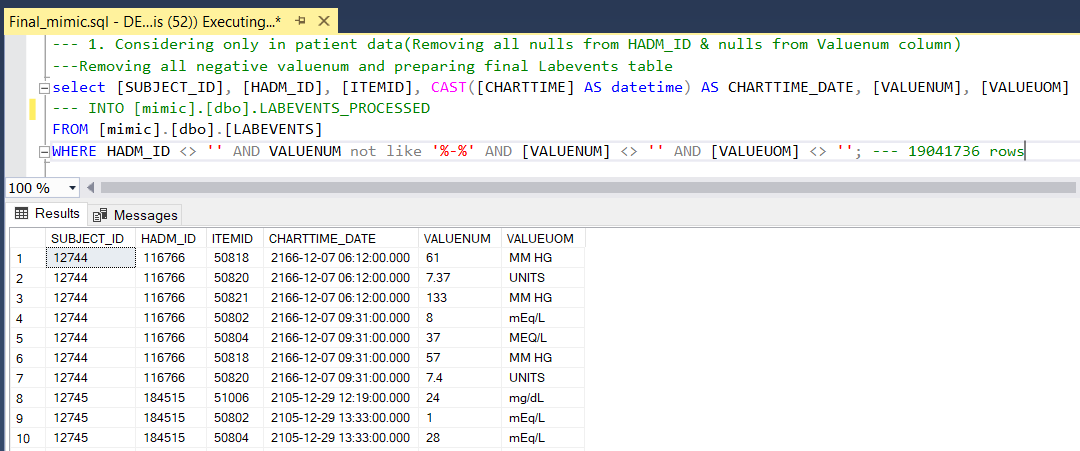
---Removing all negative valuenum and preparing final Labevents table

select [SUBJECT\_ID], [HADM\_ID], [ITEMID], CAST([CHARTTIME] AS datetime) AS CHARTTIME\_DATE, [VALUENUM], [VALUEUOM]

INTO [mimic].[dbo].LABEVENTS\_PROCESSED

FROM [mimic].[dbo].[LABEVENTS]

WHERE HADM\_ID <> '' AND VALUENUM not like '%-%' AND [VALUENUM] <> '' AND [VALUEUOM] <> ''; --- 19041736 rows



---- LAB events within 24 hours of admission

select a.[SUBJECT\_ID], a.[HADM\_ID], [GENDER], [DOB], [ADMITTIME\_DATE], [AGE], [AGE\_CATEGORY], [ETHNICITY\_CATEGORY],

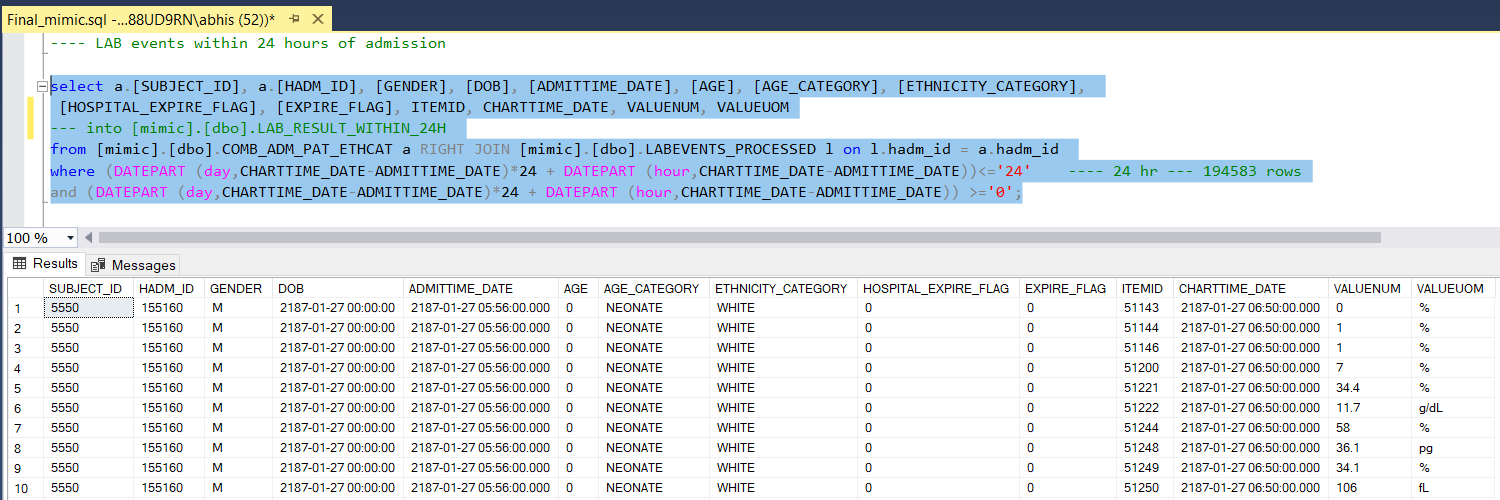
[DIAGNOSIS], [HOSPITAL\_EXPIRE\_FLAG], [EXPIRE\_FLAG], ITEMID, CHARTTIME\_DATE, VALUENUM, VALUEUOM

into [mimic].[dbo].LAB\_RESULT\_WITHIN\_24H

from [mimic].[dbo].COMB\_ADM\_PAT\_ETHCAT a RIGHT JOIN [mimic].[dbo].LABEVENTS\_PROCESSED l on l.hadm\_id = a.hadm\_id

where (DATEPART (day,CHARTTIME\_DATE-ADMITTIME\_DATE)\*24 + DATEPART (hour,CHARTTIME\_DATE-ADMITTIME\_DATE))<='24' ---- 24 hr --- 194583 rows

and (DATEPART (day,CHARTTIME\_DATE-ADMITTIME\_DATE)\*24 + DATEPART (hour,CHARTTIME\_DATE-ADMITTIME\_DATE)) >='0';



/\* select a.[SUBJECT\_ID], a.[HADM\_ID], [GENDER], [DOB], [ADMITTIME\_DATE], [AGE], [NEONATE], [ADULT], [SENIOR ADULT], [>89], [WHITE], [BLACK], [ASIAN], [HISPANIC/LATINO]

[DIAGNOSIS], [HOSPITAL\_EXPIRE\_FLAG], [EXPIRE\_FLAG], ITEMID, CHARTTIME\_DATE, VALUENUM, VALUEUOM

into [mimic].[dbo].LAB\_RESULT\_WITHIN\_24H1

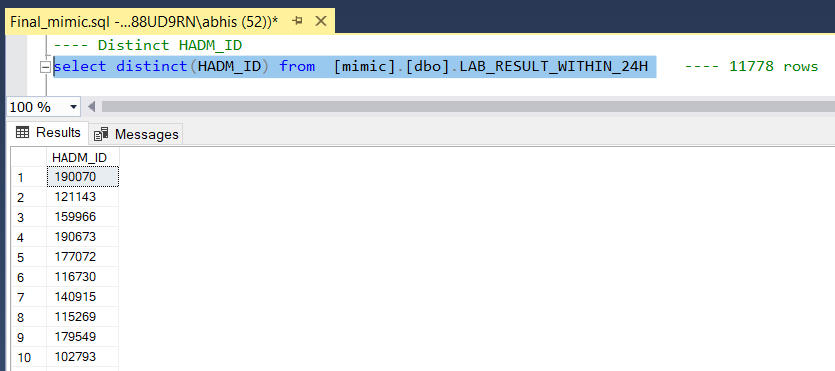
from [mimic].[dbo].COMB\_ADM\_PAT\_ETHCAT1 a RIGHT JOIN [mimic].[dbo].LABEVENTS\_PROCESSED l on l.hadm\_id = a.hadm\_id

where (DATEPART (day,CHARTTIME\_DATE-ADMITTIME\_DATE)\*24 + DATEPART (hour,CHARTTIME\_DATE-ADMITTIME\_DATE))<='24' ---- 24 hr --- 194583 rows

and (DATEPART (day,CHARTTIME\_DATE-ADMITTIME\_DATE)\*24 + DATEPART (hour,CHARTTIME\_DATE-ADMITTIME\_DATE)) >='0'; \*/

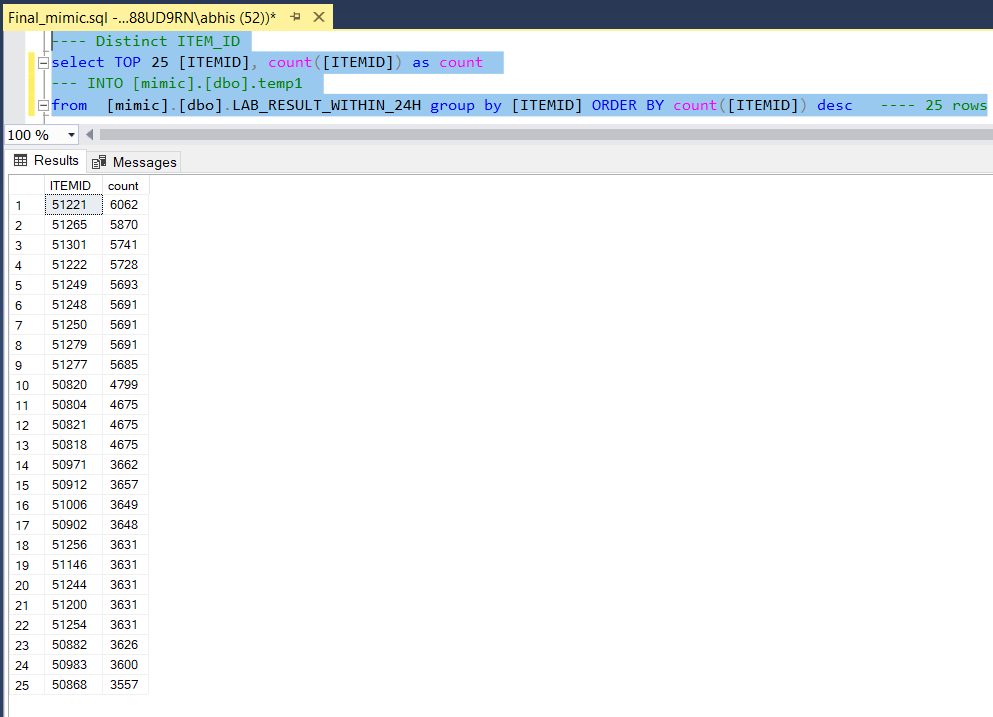
---- Distinct HADM\_ID

select distinct(HADM\_ID) from [mimic].[dbo].LAB\_RESULT\_WITHIN\_24H ---- 11778 rows



---- Distinct ITEM\_ID

select TOP 25 [ITEMID], count([ITEMID]) as count INTO [mimic].[dbo].temp1 from [mimic].[dbo].LAB\_RESULT\_WITHIN\_24H group by [ITEMID] ORDER BY count([ITEMID]) desc ---- 25 rows



--- Combining with D\_LabItems Table to get Labtest labels.

Select d.ITEMID, t.count, d.[LABEL]

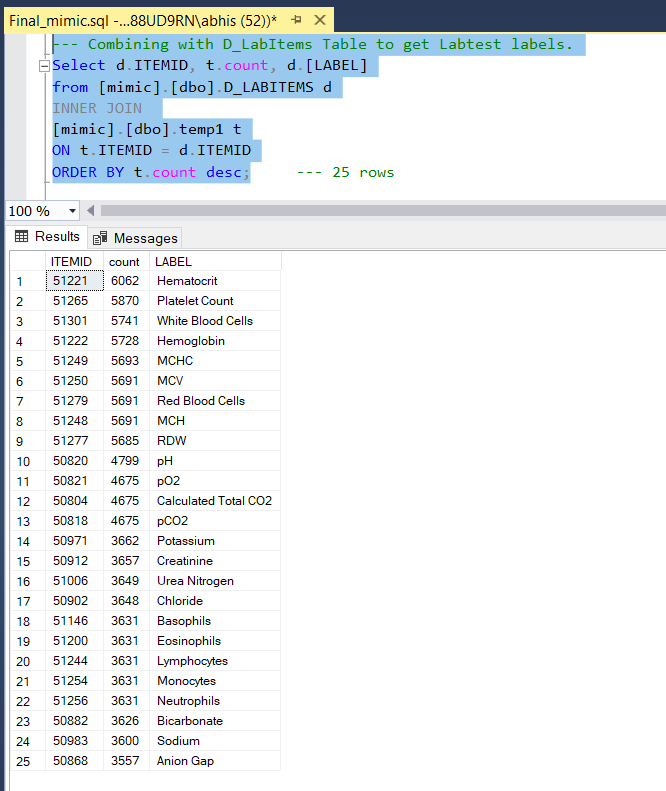
from [mimic].[dbo].D\_LABITEMS d

INNER JOIN

[mimic].[dbo].temp1 t

ON t.ITEMID = d.ITEMID

ORDER BY t.count desc; --- 25 rows

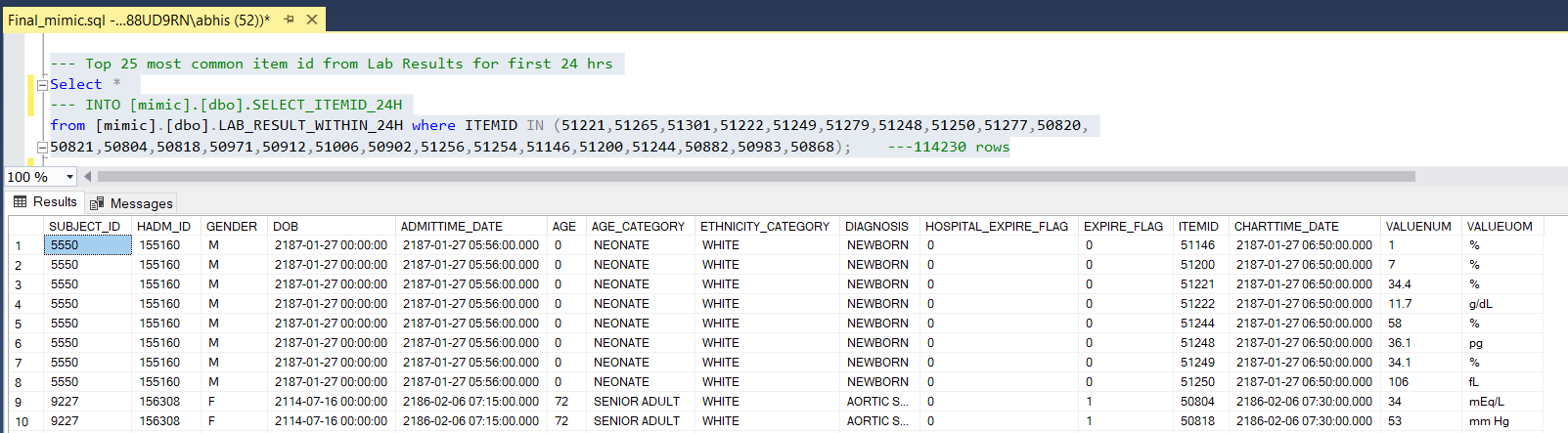


--- Top 25 most common item id from Lab Results for first 24 hrs

Select \* INTO [mimic].[dbo].SELECT\_ITEMID\_24H

from [mimic].[dbo].LAB\_RESULT\_WITHIN\_24H where ITEMID IN (51221,51265,51301,51222,51249,51279,51248,51250,51277,50820,

50821,50804,50818,50971,50912,51006,50902,51256,51254,51146,51200,51244,50882,50983,50868); ---114230 rows



Select \* INTO [mimic].[dbo].SELECT\_ITEMID\_24H

from [mimic].[dbo].LAB\_RESULT\_WITHIN\_24H where ITEMID IN (51221,51265,51301,51222,51249,51279,51248,51250,51277,50820,

50821,50804,50818,50971,50912,51006,50902,51256,51254,51146,51200,51244,50882,50983,50868);

---- Ignoring all other HADM\_ID not associated with any of the above item ids

--- Select top 25 most common item ids, getting label from D\_LABITEMS and generating pivot table

select [HADM\_ID],

avg(CASE WHEN [ITEMID]=51221 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as HEMATOCRIT,

avg(CASE WHEN [ITEMID]=51265 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as PLATELET\_COUNT,

avg(CASE WHEN [ITEMID]=51301 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as WHITE\_BLOOD\_CELLS,

avg(CASE WHEN [ITEMID]=51222 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as HEMOGLOBIN,

avg(CASE WHEN [ITEMID]=51249 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as MCHC,

avg(CASE WHEN [ITEMID]=51279 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as RED\_BLOOD\_CELLS,

avg(CASE WHEN [ITEMID]=51248 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as MCH,

avg(CASE WHEN [ITEMID]=51250 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as MCV,

avg(CASE WHEN [ITEMID]=51277 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as RDW,

avg(CASE WHEN [ITEMID]=50820 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as pH,

avg(CASE WHEN [ITEMID]=50821 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as pO2,

avg(CASE WHEN [ITEMID]=50804 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as CALCULATED\_TOTAL\_CO2,

avg(CASE WHEN [ITEMID]=50818 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as pCO2,

avg(CASE WHEN [ITEMID]=50971 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as POTASSIUM,

avg(CASE WHEN [ITEMID]=50912 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as CREATININE,

avg(CASE WHEN [ITEMID]=51006 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as UREA\_NITROGEN,

avg(CASE WHEN [ITEMID]=50902 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as CHLORIDE,

avg(CASE WHEN [ITEMID]=51256 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as NEUTROPHILS,

avg(CASE WHEN [ITEMID]=51254 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as MONOCYTES,

avg(CASE WHEN [ITEMID]=51146 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as BASOPHILS,

avg(CASE WHEN [ITEMID]=51200 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as EOSINOPHILS,

avg(CASE WHEN [ITEMID]=51244 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as LYMPHOCYTES,

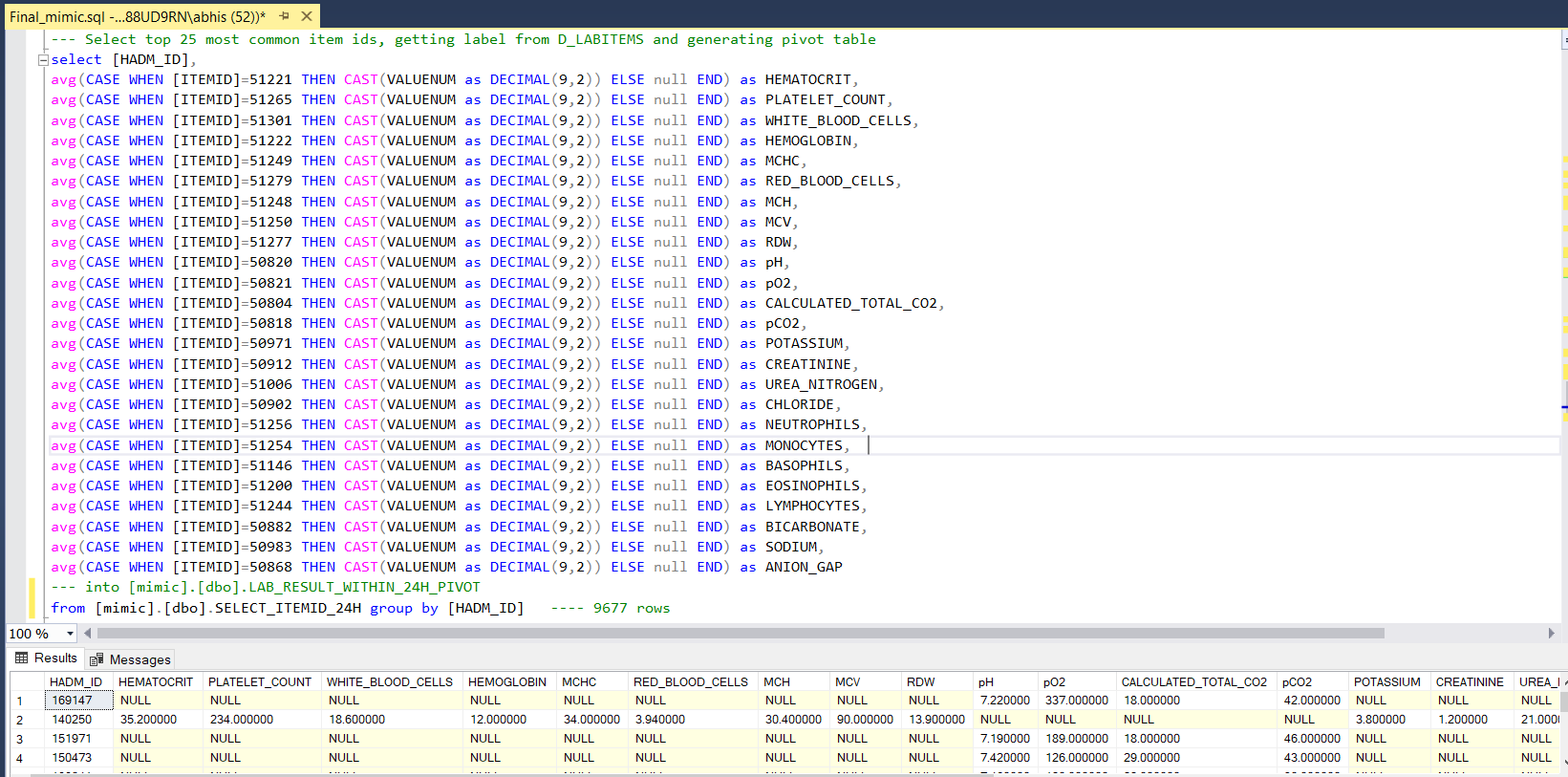
avg(CASE WHEN [ITEMID]=50882 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as BICARBONATE,

avg(CASE WHEN [ITEMID]=50983 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as SODIUM,

avg(CASE WHEN [ITEMID]=50868 THEN CAST(VALUENUM as DECIMAL(9,2)) ELSE null END) as ANION\_GAP

into [mimic].[dbo].LAB\_RESULT\_WITHIN\_24H\_PIVOT

from [mimic].[dbo].SELECT\_ITEMID\_24H group by [HADM\_ID] ---- 9677 rows



Select \* from [mimic].[dbo].COMB\_ADM\_PAT\_ETHCAT; ----- 45472 rows

SELECT \* FROM [mimic].[dbo].LAB\_RESULT\_WITHIN\_24H\_PIVOT; ---- 9677 rows

---- Performing join on two files to make a single final csv file that can be dumped in Weka

select c.\* , HEMATOCRIT, PLATELET\_COUNT, WHITE\_BLOOD\_CELLS, HEMOGLOBIN, MCHC, RED\_BLOOD\_CELLS,

MCH, MCV, RDW, pH, pO2, CALCULATED\_TOTAL\_CO2, pCO2, POTASSIUM, CREATININE, UREA\_NITROGEN, CHLORIDE,

NEUTROPHILS, MONOCYTES, BASOPHILS, EOSINOPHILS, LYMPHOCYTES, BICARBONATE, SODIUM, ANION\_GAP

---INTO [mimic].[dbo].FINAL\_TO\_WEKA

from [mimic].[dbo].COMB\_ADM\_PAT\_ETHCAT c

INNER JOIN

[mimic].[dbo].LAB\_RESULT\_WITHIN\_24H\_PIVOT l on c.hadm\_id=l.hadm\_id; ---- 9677 rows

---- View of the final file

Select \* from [mimic].[dbo].FINAL\_TO\_WEKA;

