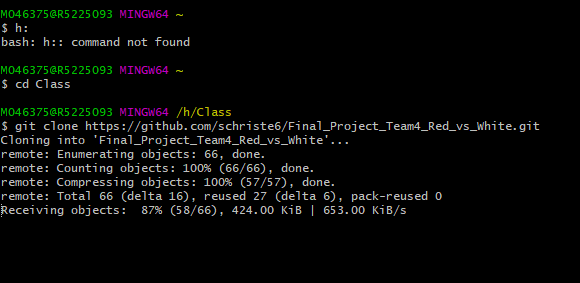
Data sources:

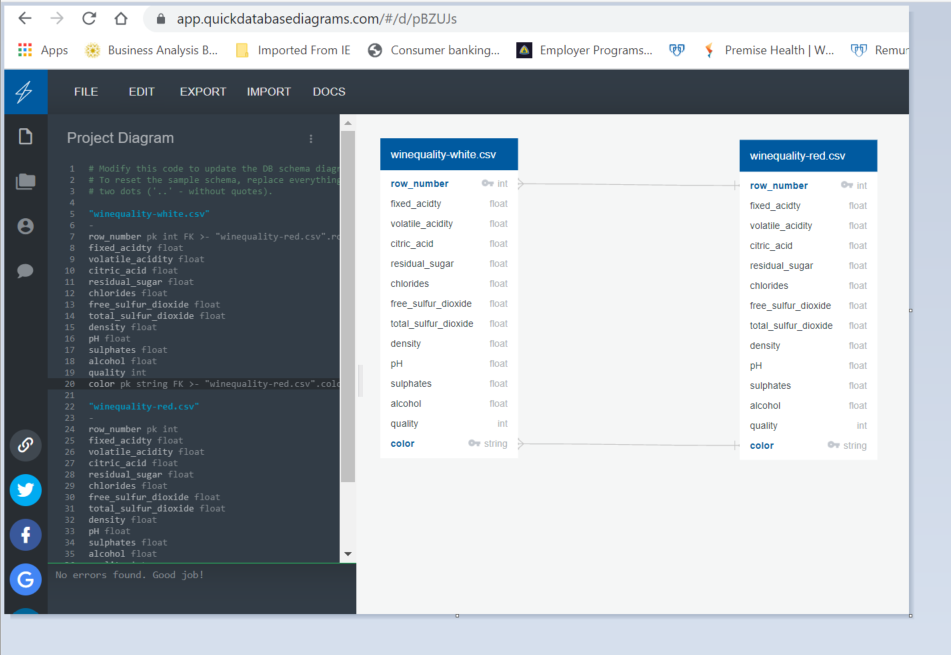
<https://www.kaggle.com/datasets/maitree/wine-quality-selection>

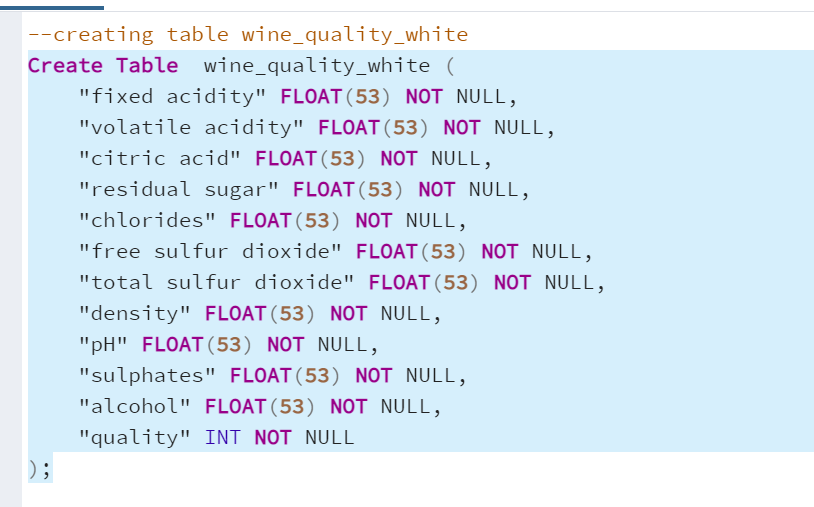
1. Clone Github

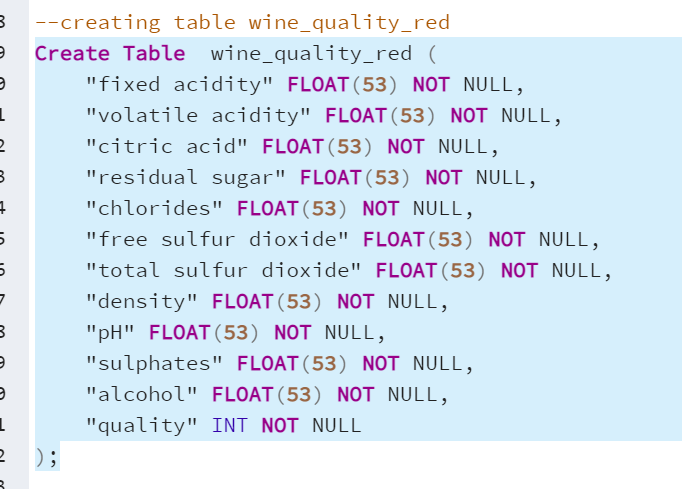


https://github.com/schriste6/Final\_Project\_Team4\_Red\_vs\_White.git

1. Create branch: in github and gitbash--done
2. Gets CSVs and place in github: done
3. Make join plan
4. Rough ERD
5. Make ERD

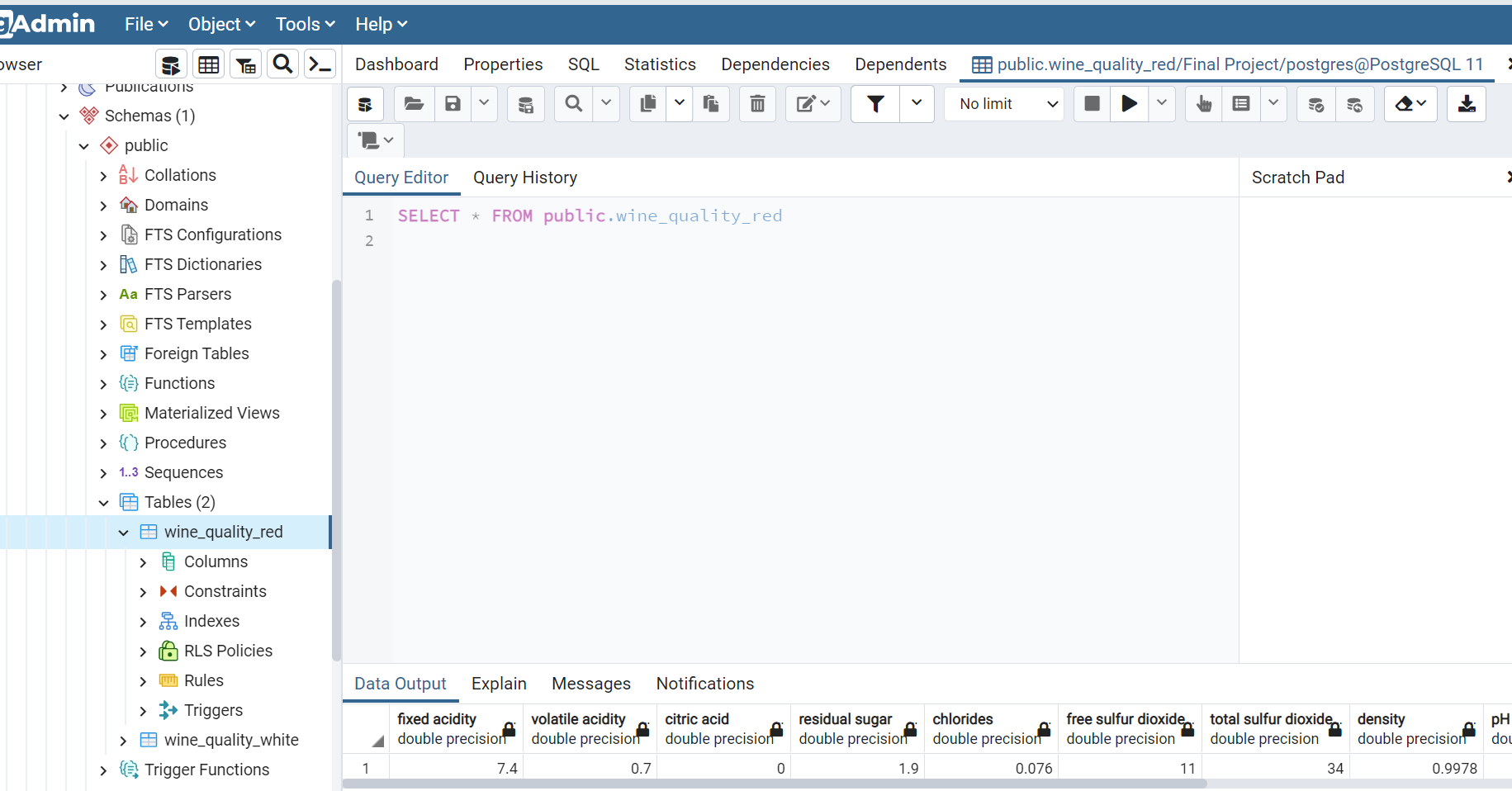


1. Use SQL to make new file
   1. Create DB
   2. Create tables
   3. done
2. 

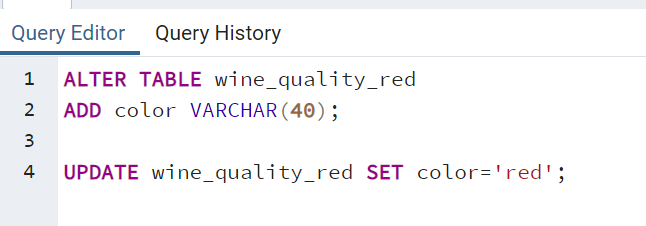


* 1. Import data

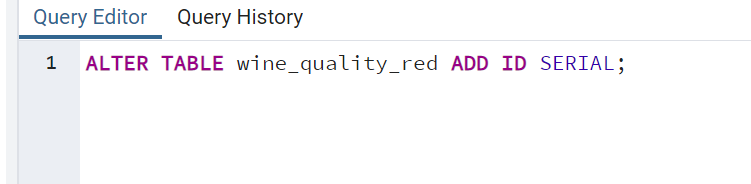
Done



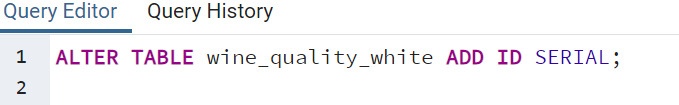
* 1. Troubleshoot
  2. Add Color columns:

1. 
   1. Add row number columns:

Red:

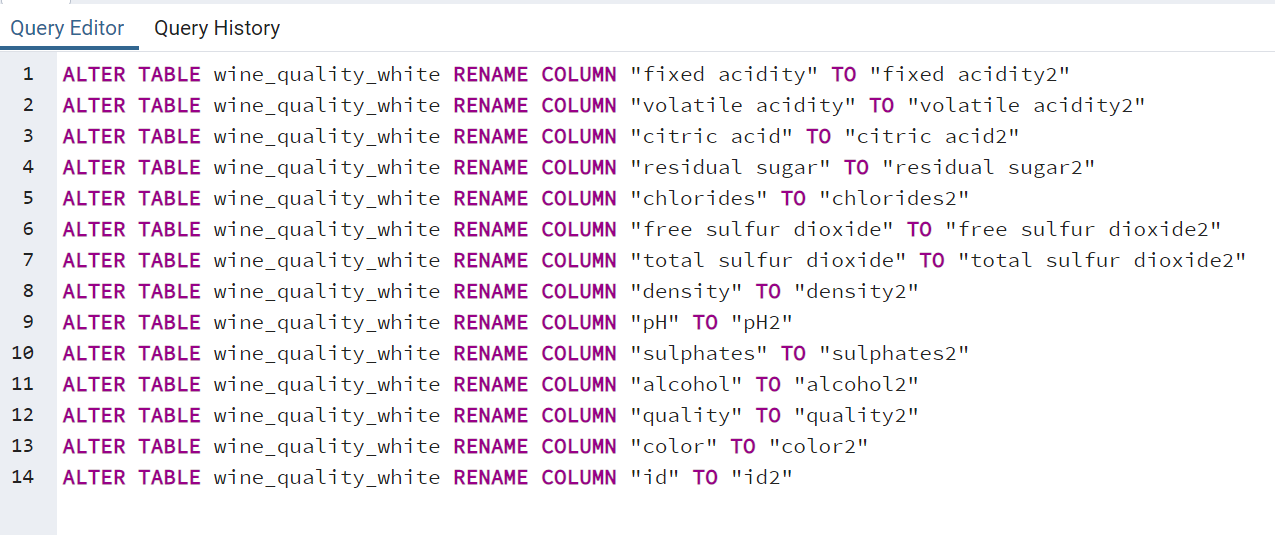


**White:**



* 1. Remove blanks/clean data

1. Rename White columns as are duplicate from red:



1. Join Tables

--joining wine\_quality\_white and wine\_quality\_red

SELECT wine\_quality\_red."fixed acidity",

wine\_quality\_red."volatile acidity",

wine\_quality\_red."citric acid",

wine\_quality\_red."residual sugar",

wine\_quality\_red."chlorides",

wine\_quality\_red."free sulfur dioxide",

wine\_quality\_red."total sulfur dioxide",

wine\_quality\_red."density",

wine\_quality\_red."pH",

wine\_quality\_red."sulphates",

wine\_quality\_red."alcohol",

wine\_quality\_red."quality",

wine\_quality\_red."color",

wine\_quality\_red."id",

wine\_quality\_white."fixed acidity2",

wine\_quality\_white."volatile acidity2",

wine\_quality\_white."citric acid2",

wine\_quality\_white."residual sugar2",

wine\_quality\_white."chlorides2",

wine\_quality\_white."free sulfur dioxide2",

wine\_quality\_white."total sulfur dioxide2",

wine\_quality\_white."density2",

wine\_quality\_white."pH2",

wine\_quality\_white."sulphates2",

wine\_quality\_white."alcohol2",

wine\_quality\_white."quality2",

wine\_quality\_white."color2",

wine\_quality\_white."id2",

INTO combined\_wine\_quality

FROM wine\_quality\_red.id

FULL OUTER JOIN wine\_quality\_white

ON wine\_quality\_red."fixed acidity" = wine\_quality\_white."fixed acidity2"

AND wine\_quality\_red."volatile acidity" = wine\_quality\_white."volatile acidity2"

AND wine\_quality\_red."citric acid" = wine\_quality\_white."citric acid2"

AND wine\_quality\_red."residual sugar" = wine\_quality\_white."residual sugar2"

AND wine\_quality\_red."chlorides" = wine\_quality\_white."chlorides2"

AND wine\_quality\_red."free sulfur dioxide" = wine\_quality\_white."free sulfur dioxide2"

AND wine\_quality\_red."total sulfur dioxide" = wine\_quality\_white."total sulfur dioxide2"

AND wine\_quality\_red."density" = wine\_quality\_white."density2"

AND wine\_quality\_red."pH" = wine\_quality\_white."pH2"

AND wine\_quality\_red."sulphates" = wine\_quality\_white."sulphates2"

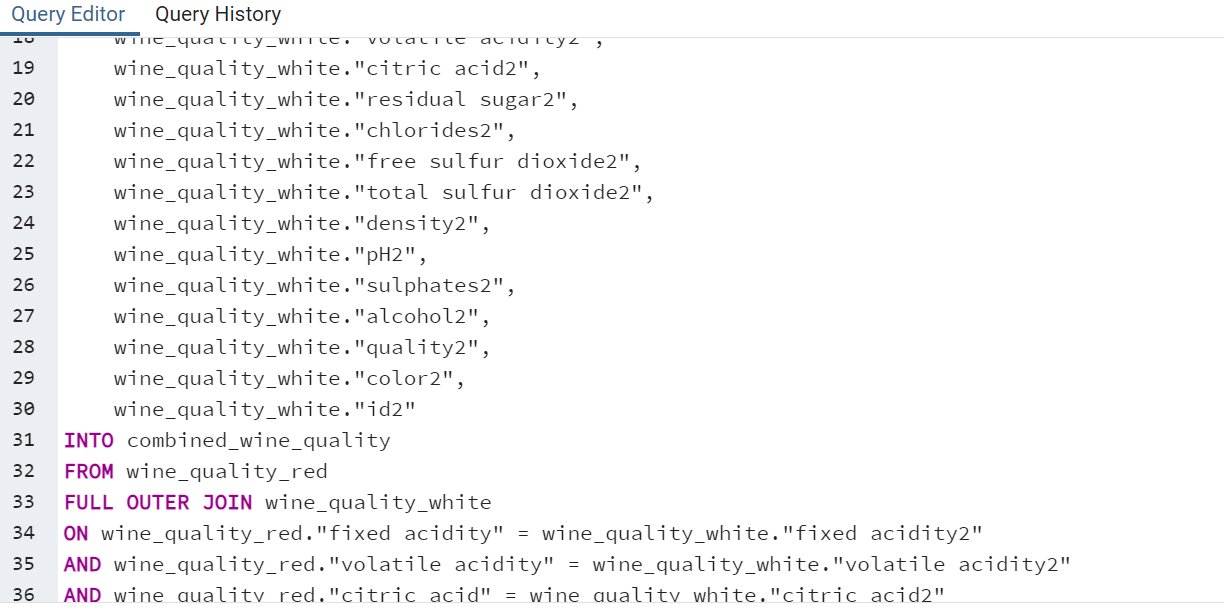
AND wine\_quality\_red."alcohol" = wine\_quality\_white."alcohol2"

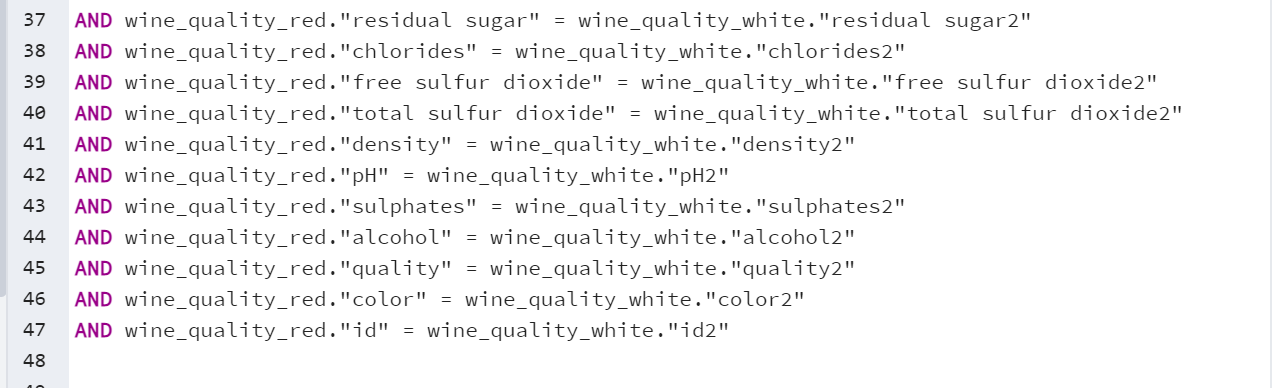
AND wine\_quality\_red."quality" = wine\_quality\_white."quality2"

AND wine\_quality\_red."color" = wine\_quality\_white."color2"

AND wine\_quality\_red."id" = wine\_quality\_white."id2"







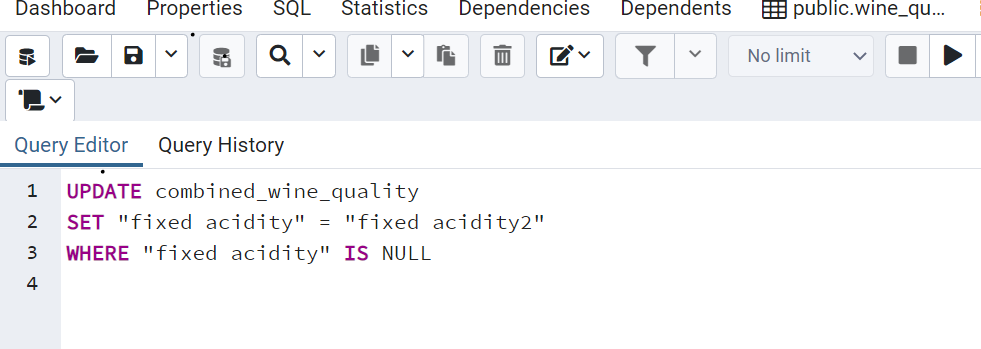
1. COMBINE COLUMNS AND Delete/drop unnecessary columns

Did the following for each of the columns

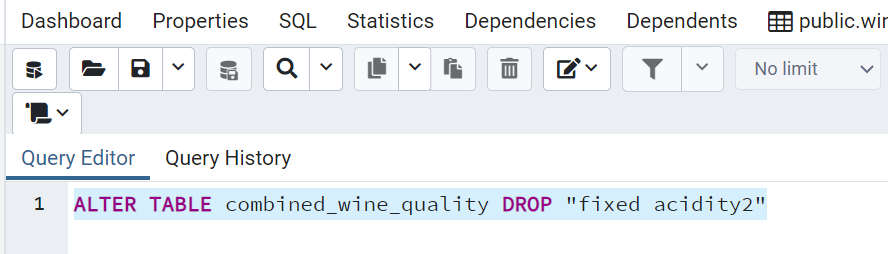
UPDATE combined\_wine\_quality

SET "fixed acidity" = "fixed acidity2"

WHERE "fixed acidity" IS NULL



Then dropped each of the columns with a “2” appended:



1. Load data into python or sQL alchemy, etc.
   1. Connect pandas and sql
2. Create branch
3. Upload to github Jessica branch
4. Upload to github—commit to main branch (4 commits)

