SUMMARY OF CLEANING MONSTER JOBS 2018 DATA SET

GOAL:

-Extract all jobs in the Administrative/Clerical category, clean and export them to their own table

-Prove SQL data cleaning skills

THE DATA SET:

This data set is an extensive CSV Excel sheet of job listings on monster.com in 2018.

I downloaded it from [Kaggle](https://www.kaggle.com/datasets/PromptCloudHQ/us-jobs-on-monstercom) and it is credited to a company named PromptCloud.

The dataset has 12 columns and 22000 rows.

The columns explained:

-country - string data - country the job is located in

-country-code - string data - an abbreviation of the country name where the job is located

-date\_added - date data - the day the listing was made

-has\_expired - string data - if the job listing has expired at the time of the CSV file’s creation

-job\_board - string data - the job board the job was posted too

-job\_desprction - string data - description of job as written by the employer

-job\_title - string data - title of job on offer

-location - string data -where the job is located

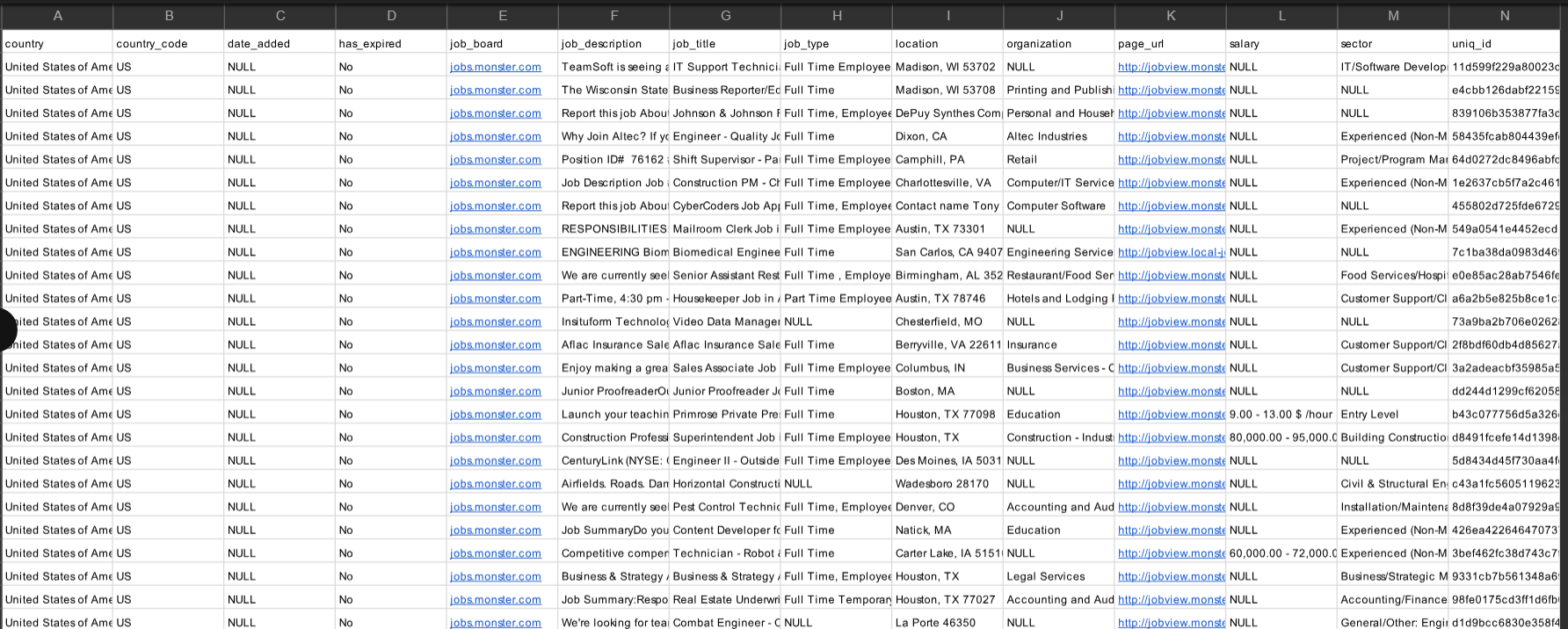
-Organization - string data -type of company offering the job

-page\_url - string data -the url address of the job posting

-sector - string data- the industry of the job on offer

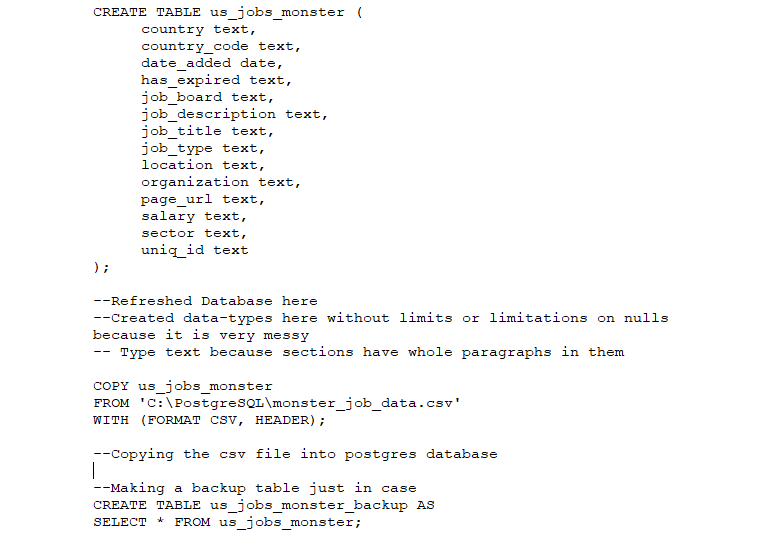
-uniq\_id - string data -unique identifier for each job listing

The data set before being cleaned :

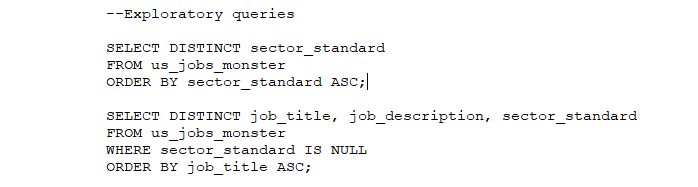


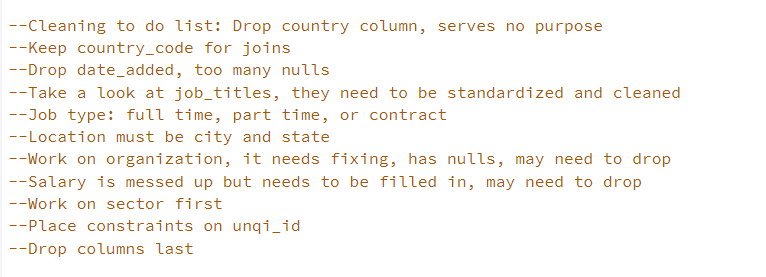
SUMMARY OF CLEANING:

1. I used the query editor in Postgresql to create a table. I then import the data from the monster\_job\_data csv file into the table I created. After importing the table, I copied the contents of the table to another table. I did this to ensure the data would be preserved if mistakes were made during data cleaning.

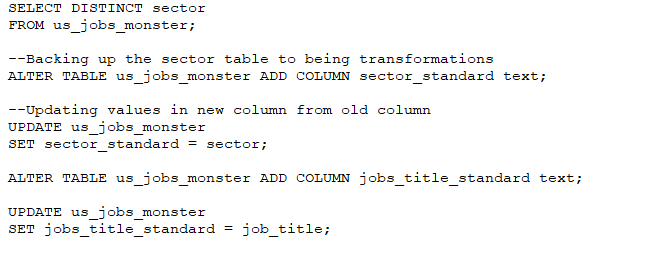


1. I began using queries to explore the data and see what information it held as well as what sort of cleaning needed to be done.

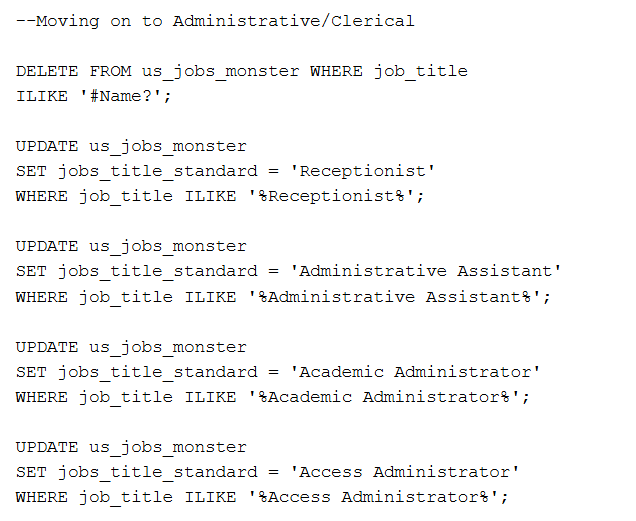




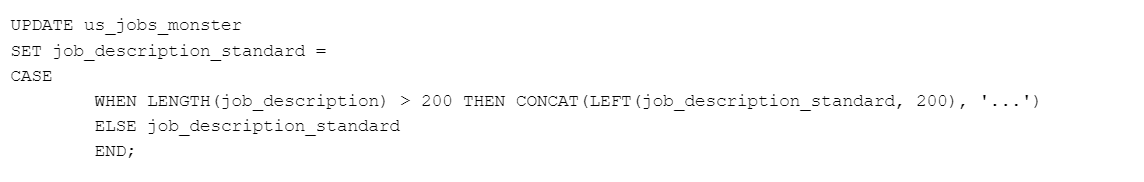
1. I made another column to place the cleaned data in. I did this because it would ensure I have a backup of the original data on the same table.



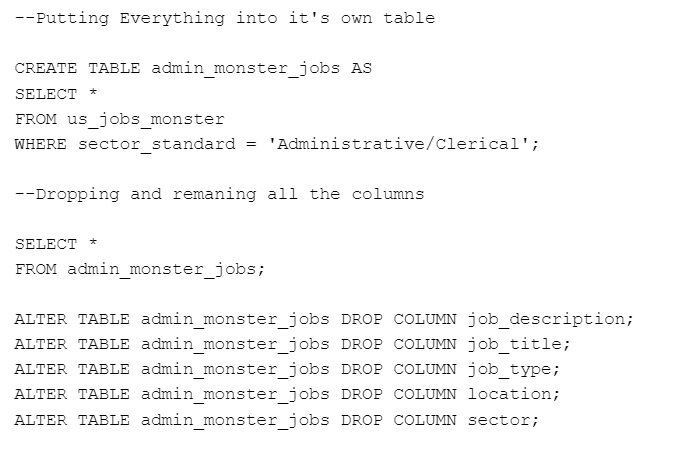
1. I came to the conclusion that cleaning the sector column first would make cleaning the rest of the data easier, as it would allow me to break the data by category. I used a combination of UPDATE and WHERE to change the values in sector\_standard. With WHERE I could pick out words similar jobs had in common and convert sector types in batches. Using UPDATE allowed me to take stock of what effect the queries had on the date. I was able to fix any mistakes I made because I took it slowly.
2. After cleaning each job’s sector, I called all the jobs in the Administrative/Clerical sectors and began to clean job\_titles. I did this with UPDATE and WHERE statements. As well as deleting rows with no useful information was provided.



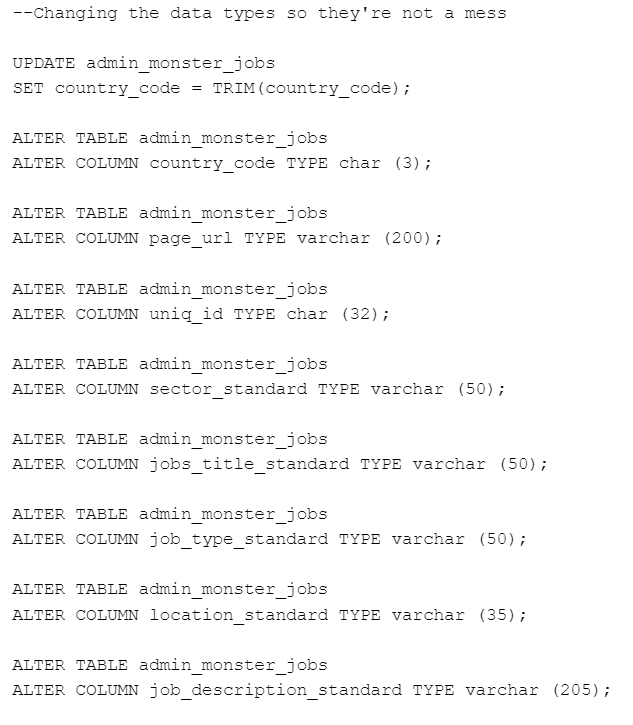
1. Job\_type and location were cleaned in a similar way. I made new columns for the standardized information. And using a combination of UPDATE and WHERE I cleaned these columns as well.
2. Next, I dropped columns that were littered with junk information or contained an overabundance of nulls. I dropped date\_added because it had too many nulls and the information couldn’t be gathered via other columns. I dropped has\_expired as well, the data is from 2018 and it is safe to assume the jobs listed are no longer on offer. I dropped job\_board because it is redundant information. The entire table is from monster.com job listings. It did not need to be listed in every row. Organization was dropped for being redundant to the Sector column. Salary was dropped for having too many nulls and other columns did not have the information to fill the nulls in with.
3. I standardized the job\_description column with INTICAP and TRIM. I then used a CASE WHEN statement along with CONCAT to shorten the length of the job descriptions. The first 200 characters of each description were more than enough to understand what each job would entail while ensuring the data storage requirements of the column to a minimum. I added ‘...’ at the end to show there was previously more to the description.



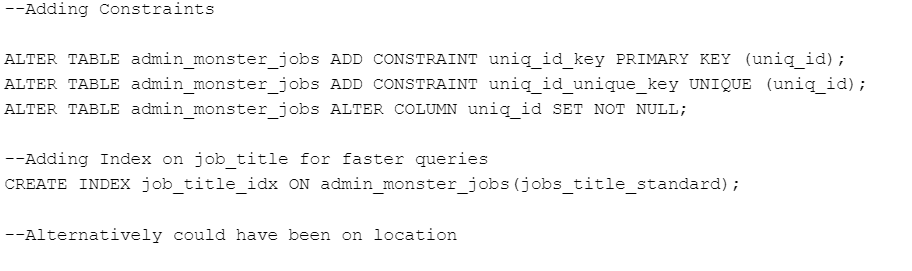
1. After this was complete I created a table of jobs categorized as Administrative/Clerical, and dropped the columns containing the original data.



1. With everything clean it was time to fix the data types in each column. I assigned most columns the varchar data type with varying length limits. Country\_code and uniq\_id were given the data type char, as neither of these columns should have data longer than a certain length.



11) I gave the uniq\_id column the primary key constraint. I added the unique constraint to the column as well to ensure all data in the column would truly be unique. Finally, I added an index to the job\_title column for faster queries. I could have added an index elsewhere depending on what sort of queries the table could be used for.



12) After this was complete the table was ready to be exported as a CSV file. 

After cleaning the data looks like this: