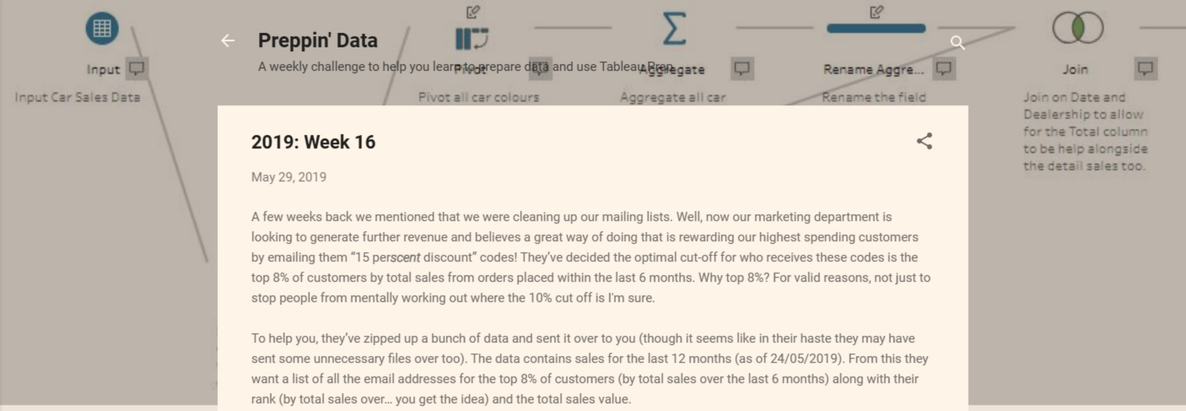
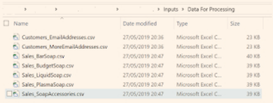


2019: Week 16

May 29, 2019

A few weeks back we mentioned that we were cleaning up our mailing lists. Well, now our marketing department is looking to generate further revenue and believes a great way of doing that is rewarding our highest spending customers by emailing them "15 per*scent* discount" codes! They've decided the optimal cut-off for who receives these codes is the top 8% of customers by total sales from orders placed within the last 6 months. Why top 8%? For valid reasons, not just to stop people from mentally working out where the 10% cut off is I'm sure.

To help you, they've zipped up a bunch of data and sent it over to you (though it seems like in their haste they may have sent some unnecessary files over too). The data contains sales for the last 12 months (as of 24/05/2019). From this they want a list of all the email addresses for the top 8% of customers (by total sales over the last 6 months) along with their rank (by total sales over... you get the idea) and the total sales value.



The files they've provided us. We only wanted the Sales data.

REQUIREMENTS

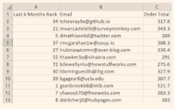
- Download the "Data for Processing" folder containing all the input files.
- Find a way to import & combine all the data from the file without deleting the non-sales data that was accidentally sent to you (they might need it back, they don't seem very tech-savvy). The rest of the detail in the file names isn't important.
- Find a way to rank the customers by total sales across orders placed within the last 6 months.
  - **NB:** *For this challenge, calculate "last 6 months" from 24/05/2019, not today's date.*
- Find a way to filter these down to customers in the top 8%.
- Produce a neatly formatted output matching the structure shown and described below.

The data structure of one of the input files.

Further hints can be found at the end of this post if you get stuck.

OUTPUT

- One output file.
- 3 columns: [Last 6 Months Rank], [Email], [Order Total].
- The number of rows is a potential spoiler so it is contained in the HINTS section..



The data structure of the output.

You can find our full output file here for comparison.

Don't to forget to fill in our participation tracker!

——————— *SPOILERS BELOW* ———————

HINTS

These hints go in order of when they come in handy during our solution:

1. You don't need to import all the "Sales_" files one-by-one.
2. Filtering with a date calculation could come in handy to reduce the date range.
3. You can join two consecutive clean steps to each other even if you don't do anything in one or both steps. Perhaps this, along with the correct join condition, could help figure out who has been generating the most sales?
4. You don't need to do mental math to figure out which ranks are in the top 8% if you aggregate to count how many email addresses there are in total and multiply by 0.08.
5. There should be 72 rows in your final output (not including headers).

Aggregation Join ranking top n percent week16

Popular posts from this blog

2023: Week 1 The Data Source Bank

January 04, 2023

Created by: Carl Allchin Welcome to a New Year of Preppin' Data. These are weekly exercises to help you learn and develop data preparation skills. We publish the challenges on a Wednesday and share a solution the following Tuesday. You can take the challenges whenever you want and we love to see y ...

READ MORE

2023: Week 2 - International Bank Account Numbers

January 11, 2023

Challenge By: Jenny Martin For week 2 of our beginner month, Data Source Bank has a requirement to construct International Bank Account Numbers (IBANs), even for Transactions taking place in the UK. We have all the information in separate fields, we just need to put it altogether in the following order ...

READ MORE

2021: Week 22 - Answer Smash

June 02, 2021

Challenge By: Jenny Martin Recently, my family and I have become quite invested in the TV quiz show Richard Osman's House of Games . The final round is always a round called Answer Smash. In this round you have a picture and question and you have to "smash" the name of the picture with the ans ...

READ MORE