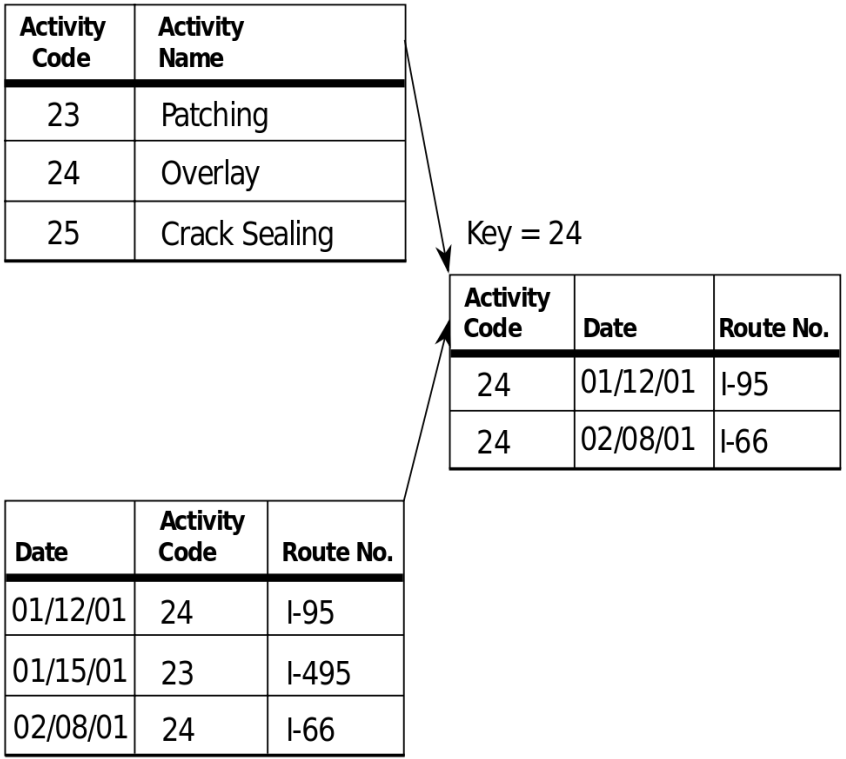
What are relational databases and why use them?

You can think of a relational database as a collection of (sometimes very, very big) spreadsheet files.

Relational databases are generally used for “tidy” data (Wickham 2014):

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

Relational Model



You can think of a relational database as a collection of (sometimes very, very big) spreadsheet files.

Relational databases are generally used for “tidy” data (Wickham 2014):

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

Spreadsheets	**Databases**
Spreadsheet	Table
Rows	Records
Columns	Attributes

When should you use a relational database?

Do use a relational database when...

- The data you want is already in a database
- You have a large amount of data (too big to fit in local memory)
- Many people need to be able to access/change the data
- You're learning to use relational databases ;)

Don't use a relational database when...

- You have a very small amount of data and don't expect to get more in the future
- You don't have tabular data (images, text, hierarchical data structures, etc.). These are better stored and accessed using other tools (like noSQL or various ways of storing "blobs")

BigQuery!

- We'll be learn SQL using datasets hosted on BigQuery (don't worry, the SQL syntax will be pretty much the same whatever service you use!)
- Relational database service from Google, optimized for fast performance on very big datasets
- We've built access to the public BigQuery datasets into Kaggle
- FYI:
  - To use a public BigQuery dataset in your kernel, you need to add it as a dataset
  - Each Kaggle user can scan 5TB every 30 days for free. If you go over this limit, you'll need to wait for it to reset.
- If you want to work locally, you'll need to set up a Google Cloud Platform account, install the client library and link your account