Extract Transform Load

Group Members:

- Grace Arhin
- Josh Mills
- Rodney Davermann
- Chithra Sundaram
- Jennifer Gaddie

Question: Create a forecasting model for the occupation that is most likely to win the next Bachelor/Bachelorette.

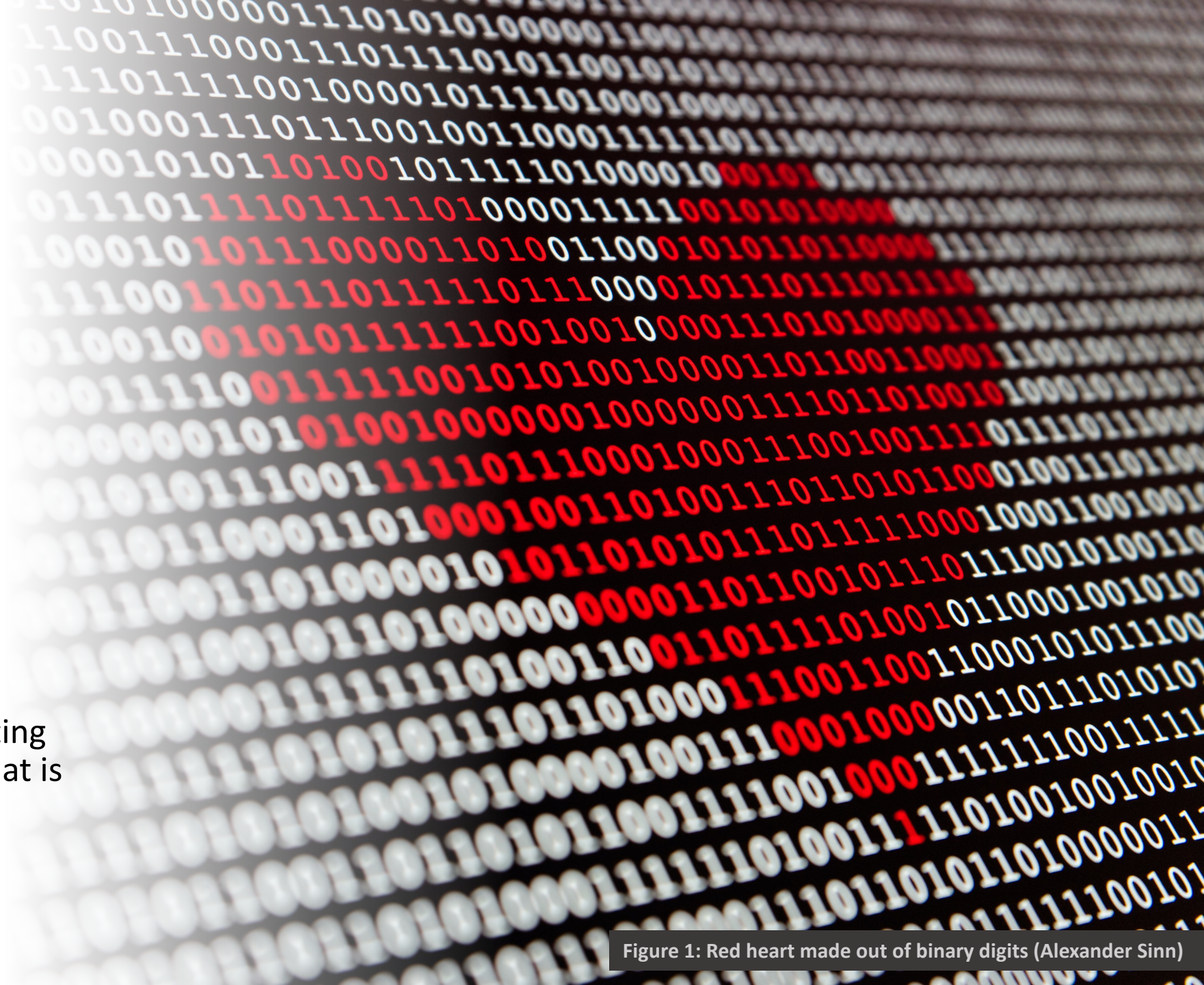


Figure 1: Red heart made out of binary digits (Alexander Sinn)

ETL Process to get the data for the winner

Extract

- Import needed data from two sources: Kaggle and Data World
- Both Files were CSV
- Prior to transform, we examined the data, and became familiar with it

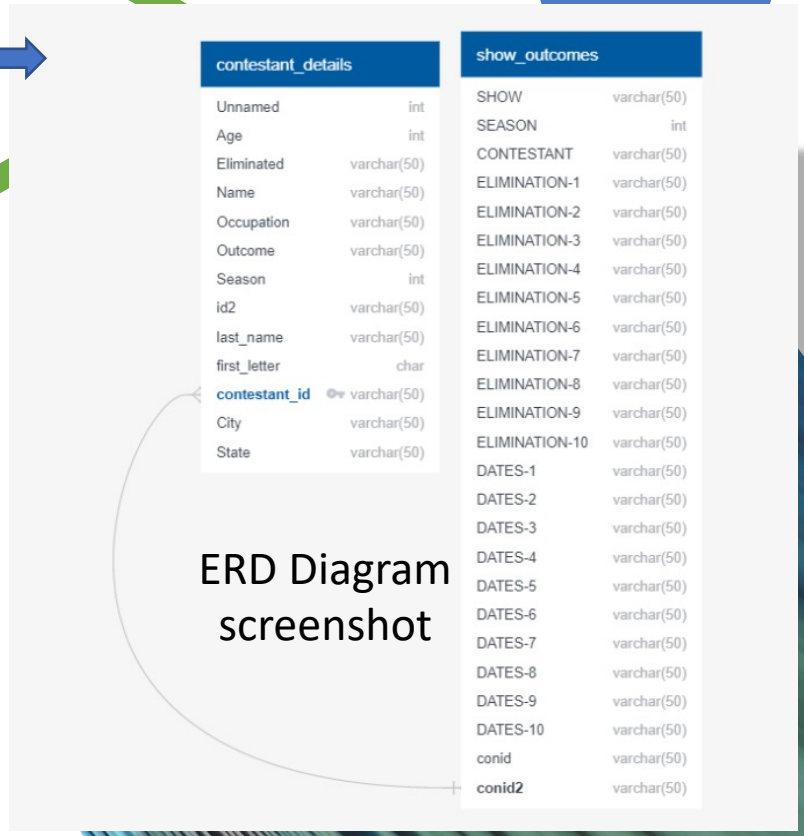
Transform

- Concatenate season with name to create a primary/foreign key
- Make the key all Upper Case
- Separate the id from last name to further make a primary key
- Replace the the last name that says none to X so we could make a primary key
- Took just the first letter of the last name so we could make a primary key
- Concatenate everything as a final key
- Take the first part of contestant id off because one sheet starts with 01 and the other sheet starts with 1
- Made it so both keys had the same format ex. 01_firstname_lastinitial check conid2 at the end
- Drop unneeded duplicated columns
- Split hometown column into cities and state

Load

- Convert files back to csv so we can load into Postgres
- Built EDR to confirm correct SQL and Primary Key
- Created Database, and Tables to load to Postgres
- Use SQL Alchemy to pull from Postrgres
- Check for tables to make sure they are available

- **NOTE:** We chose this path to allow for ease of data review later, and to allow for future test against additional hypothesis testing



ERD Diagram screenshot

[etl_project github link]