SQL Powered EDA

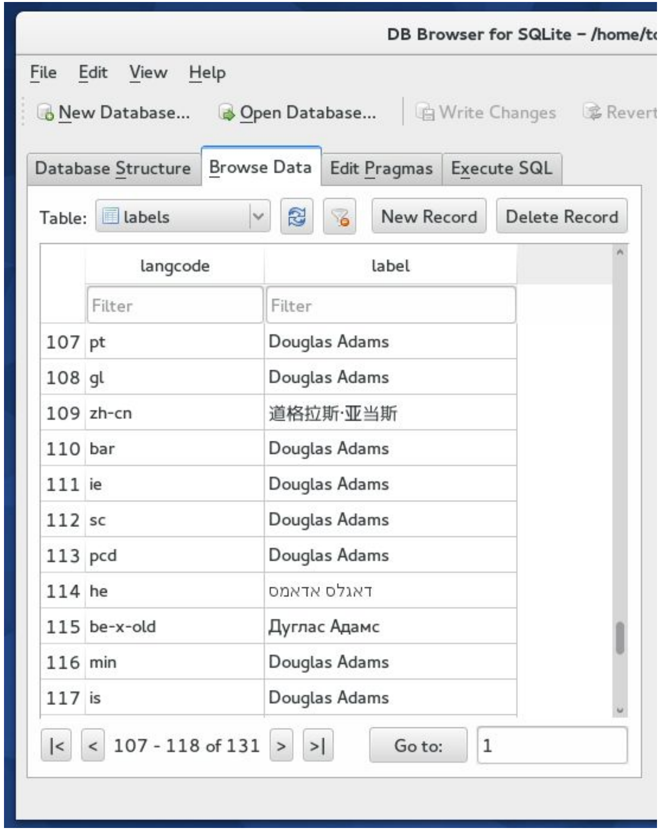
What You Need Before We Start

Please go to the following GitHub Repository:
https://github.com/thenileriver/DS3_Workshop_4

- Click “Fork”
- Open up your command prompt/terminal
- Go the the folder you want to upload repository to
- Ex: cd Desktop if you want to upload repo to Desktop
- Type git clone and copy+paste the URL of your forked repository
- To see if it worked, use the ls command
- Change directories to DS3_Workshop_4
- Type: cd DS3_Workshop_4

Intro to SQL

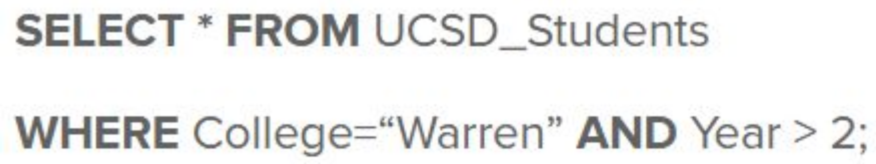
- SQL (Structured Query Language) is a programming language for storing/retrieving information in databases
    - Databases appear in form of SQL Table, which is like a spreadsheet
    - Similar to Pandas DataFrame
- Popular for data scientists because it can be integrated with other languages
    - Allows for creation/retrieval/storage of data for analysis



	langcode	label
	Filter	Filter
107	pt	Douglas Adams
108	gl	Douglas Adams
109	zh-cn	道格拉斯·亚当斯
110	bar	Douglas Adams
111	ie	Douglas Adams
112	sc	Douglas Adams
113	pcd	Douglas Adams
114	he	דאגלס אדמס
115	be-x-old	Дуглас Адамс
116	min	Douglas Adams
117	is	Douglas Adams

Basic SQL Commands

- **CREATE DATABASE** ***[database name]*** creates the database
- **SELECT * FROM** ***[database]*** is the basic query that allows you to select rows from a given database
  - Can be modified to be more specific
  - In example, College and Year are rows in UCSD_Students
  - For a list of full commands visit: https://www.w3schools.com/sql/default.asp



```
SELECT * FROM UCSD_Students

WHERE College="Warren" AND Year > 2;
```

What is EDA?

- As data scientists, we want the ability to understand the data we’re working with
- Exploratory Data Analysis (EDA) is used to understand a dataset’s underlying structure, properties, and patterns
    - EDA is done on **exploratory variables** to see changes in **response variables**
    - Think of an experiment: independent (exploratory) variables are changed to see what happens to dependent (response) variables
- This allows us to better model the data and explore areas for further investigation

EDA Summarized

- Cleaning: removing irregularities and unnecessary features from the data
- Visualization: start to gain insight from the data through charts/graphs
  - calculating summary statistics and analyzing distributions of variables
- Analysis: explore relationships between different variables in the data
  - finding correlations between variables, graphing relationships
- Pandas libraries we use:
  - Pandas: data manipulation
  - Numpy: perform mathematical operations on data
  - Seaborn: data visualization

SQLAlchemy and EDA

- SQL is commonly used for data science
- In this workshop, we’ll use SQLAlchemy (a Python SQL toolkit) to perform EDA
- Download SQLAlchemy here: https://www.sqlalchemy.org/download.html
  - Installation guide: https://docs.sqlalchemy.org/en/20/intro.html#installation

https://tinyurl.com/WorkshopEQSurvey

Please fill out the survey so we can improve our workshops next quarter :)