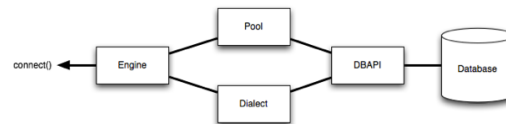**Piruz Alemi SQL & SQLITE Report of Jan 4th, 2020**

This report captures the steps taken to create an SQL data base, and import it in PANDAS, while creating a set of reports in both SQL + PANDAS + Python + SQLITE + MPLOT

Data Modeling

Inspected the CSVs and sketched out an ERD of the tables (Entity Relations Data) by means of a tool like http://www.quickdatabasediagrams.com as noted below:
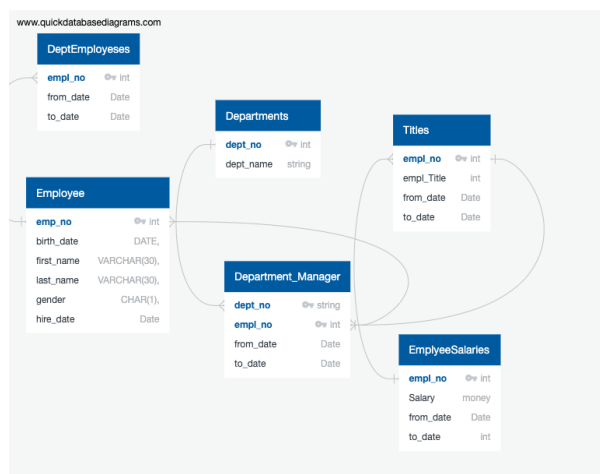
Imported the following files for data analysis:

1. **Deptartments.csv**
   "**dept_no**","dept_name"

2. **Dept_empl.csv**
   "**emp_no**","**dept_no**","from_date","to_date"

3. **Dept_manager.csv**
   "**dept_no**","emp_no","from_date","to_date"

4. **Employee.csv**
   "**emp_no**","birth_date","first_name","last_name","gender","hire_date"

5. **Salaries.csv**
   Emp_no,salary,from_date,to_date

6. **Titles.csv**
   "**emp_no**","title","from_date","to_date"

The above Six data files were linked as a single data base with the **primary** and **foreign keys** highlighted in red.

Data Engineering

1. Using the information I (Piruz Alemi) created a table schema for each of the above six CSV files.

2. In this process I specified:
   1. data types,
   2. primary keys,
   3. foreign keys,
   4. and other constraints.

3. And imported six CSV files into the above defined corresponding SQL tables.

Data Analysis

Once I had a complete SQL database, the following analysis was undertaken:

1. Listed the following details of each employee:
   1. employee number,
   2. last name,
   3. first name,
   4. gender,
   5. and salary.

2. Listed employees who were hired in 1986.

3. Listed the manager of each department with the following information:
   1. department number,
   2. department name,
   3. the manager's employee number,
   4. last name,
   5. first name,
   6. and start employment
   7. end employment dates.

4. Listed the department of each employee with the following information:
   1. employee number,
   2. last name,
   3. first name,
   4. and department name.

5. Listed all employees whose first name is "Hercules" and last names begin with "B."

6. Listed all employees in the Sales department, including:
   1. their employee number,
   2. last name,
   3. first name, and
   4. department name

7. Listed all employees in the Sales and Development departments, including:
   1. Their employee number,
   2. Last name,

3. First name, and
4. Department name.

8. In descending order, listed the frequency count of employee last names, i.e., how many employees share each last name.

Bonus (Optional)

As I examined the data, I was overcome with a creeping suspicion that the dataset is fake. I surmise that my boss handed me spurious data in order to test the data engineering skills of this new employee. To confirm my hunch, I decided to take the following steps to generate a visualization of the data, with which I will never confront my boss:

1. Imported the SQL database into Pandas. (Yes!!!, I can read the CSVs directly in Pandas) This step required some research. I used the code below to get started. *And made sure to make any necessary modifications for my username, password, host, port, and database name:*

2. from sqlalchemy import create_engine

3. engine = create_engine('postgresql://localhost:5432/<your_db_name>')
   connection = engine.connect()

- Consult [SQLAlchemy documentation](#) for more information.

- If using a password, do not upload your password to your GitHub repository. See https://www.youtube.com/watch?v=2uaTPmNvH0I and https://martin-thoma.com/configuration-files-in-python/ for more information.

2. Created a histogram to visualize the most common **last names** for employees. This was done in PANDAS, by importing MPLOT

3. Created a bar chart of average salary by title. (This was done in SQL!)

Epilogue

Evidence in hand, I marched into my boss's office and presented the visualization. With a sly grin, my boss thanked me for my work. On my way out of the office, I heard the words, "Search your ID number." I look down at my badge to see that my employee ID number is 499942. I was very disappointed with the salary, but hey happy to have such a nice boss.

| Data Output | Explain | Messages | Notifications |

emp_no [PK] integer	salary integer	from_date date	to_date date
1	499942	40000	1998-03-28	1999-03-28

Submissions Included:

1. Create an image file of the ERD.
2. Create a .sql/csv file of the table schemata.
3. Create a .sql/csv file of the queries.
4. (Optional) Created a Jupyter Notebook of the bonus analysis.

5. Created and uploaded a repository with the above files to GitHub and posted a link on BootCamp Spot.