SQL Project

November 21, 2021

0.0.1 Dependencies

[ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
from config import username, password
%matplotlib inline

from sqlalchemy import create_engine
engine = create_engine(f'postgresql://{username}:{password}@localhost:5432/
 ↪employees2')
connection = engine.connect()
```

0.0.2 Import the Employees table

[2]:
```
employees = pd.read_sql('select * from employees', connection,␣
 ↪parse_dates=['birth_date', 'hire_date'])
```

[3]:
```
employees.head()
```

[3]:

	emp_no	birth_date	first_name	last_name	gender	hire_date
0	10001	1953-09-02	Georgi	Facello	M	1986-06-26
1	10002	1964-06-02	Bezalel	Simmel	F	1985-11-21
2	10003	1959-12-03	Parto	Bamford	M	1986-08-28
3	10004	1954-05-01	Chirstian	Koblick	M	1986-12-01
4	10005	1955-01-21	Kyoichi	Maliniak	M	1989-09-12

0.0.3 Import the Departments table

[4]:
```
departments = pd.read_sql('select * from departments', connection)
```

[5]:
```
departments.head()
```

[5]:

	dept_no	dept_name
0	d001	Marketing
1	d002	Finance
2	d003	Human Resources
3	d004	Production
4	d005	Development

0.0.4 Import the Salaries table

[6]:
```
salaries = pd.read_sql('select * from salaries', connection,␣
↪parse_dates=['from_date', 'to_date'])
```

[7]:
```
salaries.head()
```

[7]:

	emp_no	salary	from_date	to_date
0	10001	60117	1986-06-26	1987-06-26
1	10002	65828	1996-08-03	1997-08-03
2	10003	40006	1995-12-03	1996-12-02
3	10004	40054	1986-12-01	1987-12-01
4	10005	78228	1989-09-12	1990-09-12

0.0.5 Import the Department Manager table

[8]:
```
dept_manager = pd.read_sql('select * from dept_manager', connection,␣
↪parse_dates=['from_date', 'to_date'])
```

[9]:
```
dept_manager.head()
```

[9]:

	dept_no	emp_no	from_date	to_date
0	d001	110022	1985-01-01	1991-10-01
1	d001	110039	1991-10-01	NaT
2	d002	110085	1985-01-01	1989-12-17
3	d002	110114	1989-12-17	NaT
4	d003	110183	1985-01-01	1992-03-21

Replace null dates

[10]:
```
dept_manager.to_date = dept_manager['to_date'].fillna(pd.
↪to_datetime('2050-12-31'))
```

[11]:
```
dept_manager.head()
```

[11]:

	dept_no	emp_no	from_date	to_date
0	d001	110022	1985-01-01	1991-10-01
1	d001	110039	1991-10-01	2050-12-31
2	d002	110085	1985-01-01	1989-12-17
3	d002	110114	1989-12-17	2050-12-31
4	d003	110183	1985-01-01	1992-03-21

0.0.6 Import the Titles table

[12]:
```
titles = pd.read_sql('select * from titles', connection,␣
↪parse_dates=['from_date', 'to_date'])
```

[13]:
```
titles.head()
```

[13]:

	emp_no	title	from_date	to_date
0	10001	Senior Engineer	1986-06-26	NaT
1	10002	Staff	1996-08-03	NaT
2	10003	Senior Engineer	1995-12-03	NaT
3	10004	Engineer	1986-12-01	1995-12-01
4	10004	Senior Engineer	1995-12-01	NaT

Replace null dates

[14]:
```
titles.to_date = titles['to_date'].fillna(pd.to_datetime('2050-12-31'))
```

[15]:
```
titles.head()
```

[15]:

	emp_no	title	from_date	to_date
0	10001	Senior Engineer	1986-06-26	2050-12-31
1	10002	Staff	1996-08-03	2050-12-31
2	10003	Senior Engineer	1995-12-03	2050-12-31
3	10004	Engineer	1986-12-01	1995-12-01
4	10004	Senior Engineer	1995-12-01	2050-12-31

Count by title

[16]:
```
titles.title.value_counts()
```

[16]:
```
Engineer             115003
Staff                107391
Senior Engineer       97750
Senior Staff          92853
Technique Leader      15159
Assistant Engineer    15128
Manager                  24
Name: title, dtype: int64
```

0.0.7 Average salary by title

Create a merged dataframe of titles and salaries

[17]:
```
employee_salaries = titles.merge(salaries, on='emp_no')
```

[18]:
```
employee_salaries.head()
```

[18]:

	emp_no	title	from_date_x	to_date_x	salary	from_date_y	to_date_y
0	10001	Senior Engineer	1986-06-26	2050-12-31	60117	1986-06-26	1987-06-26
1	10002	Staff	1996-08-03	2050-12-31	65828	1996-08-03	
2	10003	Senior Engineer	1995-12-03	2050-12-31	40006	1995-12-03	
3	10004	Engineer	1986-12-01	1995-12-01	40054	1986-12-01	
4	10004	Senior Engineer	1995-12-01	2050-12-31	40054	1986-12-01	

```
1 1997-08-03
2 1996-12-02
3 1987-12-01
4 1987-12-01
```

[19]:
```
employee_salaries_df = employee_salaries[['emp_no', 'title', 'salary']]
```

[20]:
```
employee_salaries_df.head()
```

[20]:

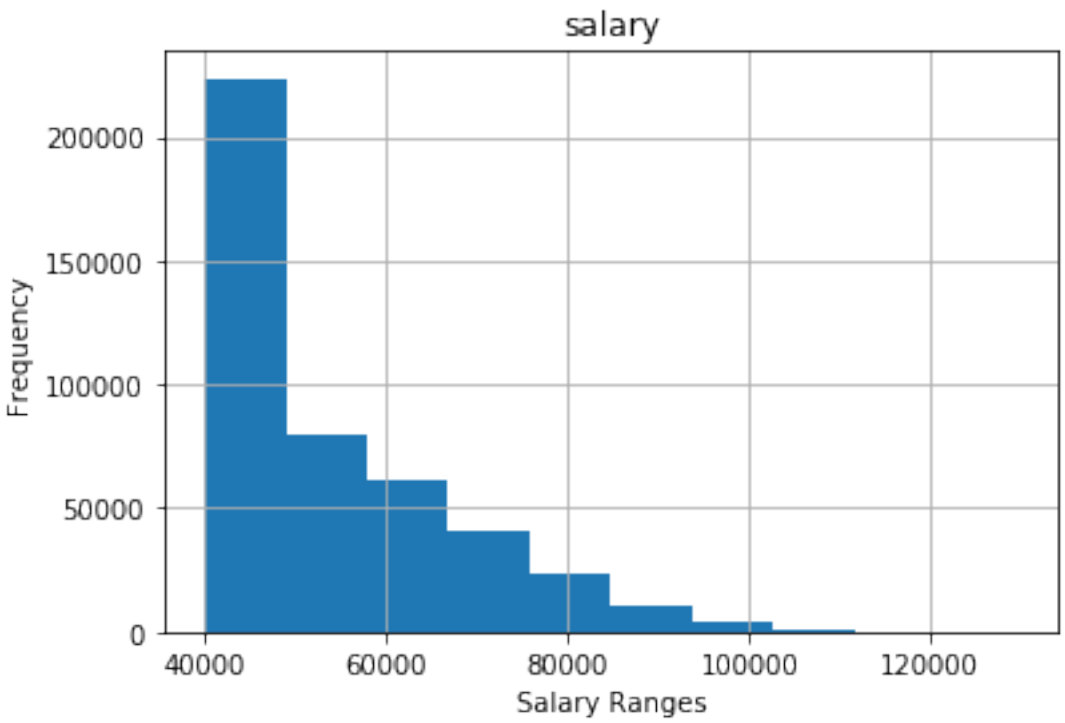
	emp_no	title	salary
0	10001	Senior Engineer	60117
1	10002	Staff	65828
2	10003	Senior Engineer	40006
3	10004	Engineer	40054
4	10004	Senior Engineer	40054

[21]:
```
employee_salaries_df.groupby('title')['salary'].mean().round(2)
```

[21]:
```
title
Assistant Engineer    48493.20
Engineer              48539.78
Manager               51531.04
Senior Engineer       48506.75
Senior Staff          58503.29
Staff                 58465.27
Technique Leader      48580.51
Name: salary, dtype: float64
```

[22]:
```
employee_salaries_df.hist(column='salary')
plt.xlabel("Salary Ranges")
plt.ylabel("Frequency")
```

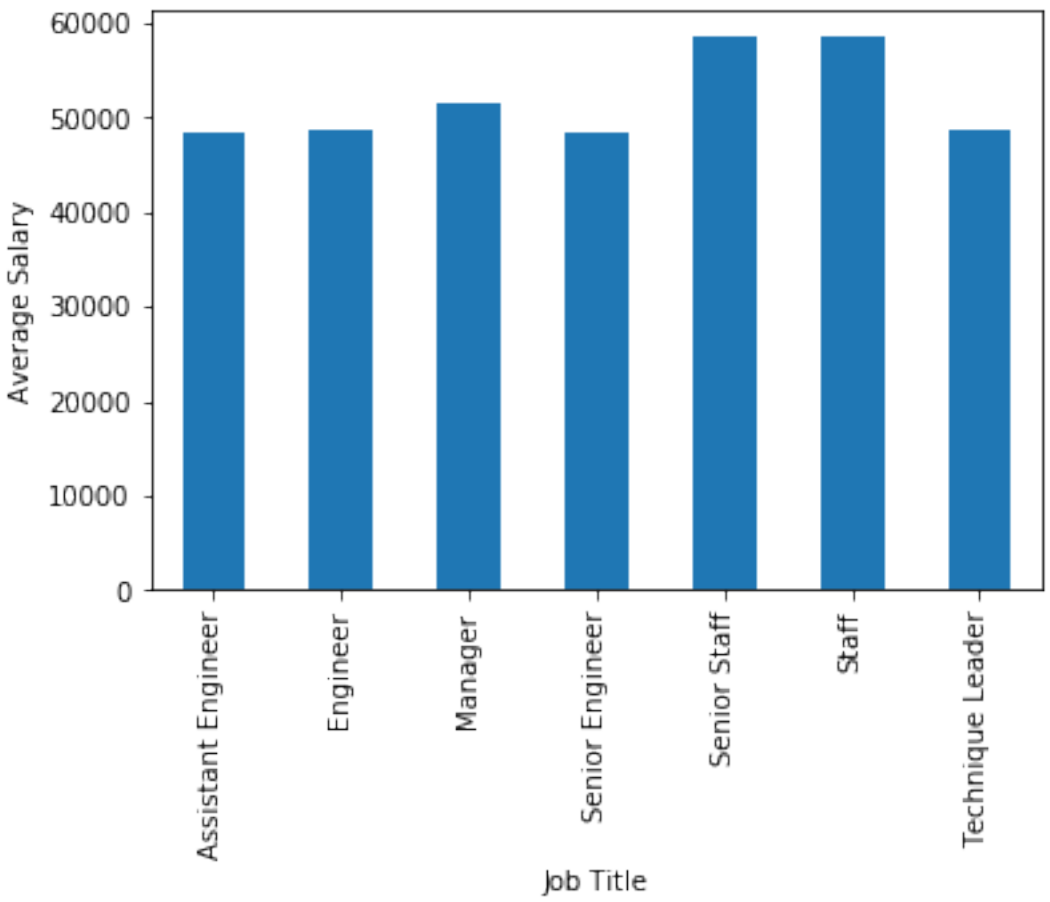
[22]:
```
Text(0, 0.5, 'Frequency')
```



[23]:
```
employees_grouped_by_title = employee_salaries_df.groupby(['title'])['salary'].
↪mean()

employees_grouped_by_title.plot.bar()
plt.xlabel("Job Title")
plt.ylabel("Average Salary")
```

[23]: Text(0, 0.5, 'Average Salary')



0.1 Schema

```
CREATE TABLE employees (     emp_no INT   NOT NULL,     birth_date DATE   NOT
NULL,     first_name VARCHAR   NOT NULL,     last_name VARCHAR   NOT NULL,
gender VARCHAR   NOT NULL,     hire_date DATE   NOT NULL,     PRIMARY KEY
(emp_no) );

CREATE TABLE departments (     dept_no VARCHAR   NOT NULL,     dept_name VARCHAR
NOT NULL,     PRIMARY KEY (dept_no) );

CREATE TABLE dept_manager (     dept_no VARCHAR   NOT NULL,     emp_no INT   NOT
NULL,     from_date DATE   NOT NULL,     to_date DATE   NOT NULL,     FOREIGN
KEY (emp_no) REFERENCES employees (emp_no),     FOREIGN KEY (dept_no) REFERENCES
departments (dept_no) );

CREATE TABLE dept_emp (     emp_no INT   NOT NULL,     dept_no VARCHAR   NOT
NULL,     from_date DATE   NOT NULL,     to_date DATE   NOT NULL,     FOREIGN
KEY (emp_no) REFERENCES employees (emp_no),     FOREIGN KEY (dept_no) REFERENCES
departments (dept_no),     PRIMARY KEY (emp_no,dept_no) );

CREATE TABLE titles (     emp_no INT   NOT NULL,     title VARCHAR   NOT NULL,
```

```
from_date DATE   NOT NULL,     to_date DATE   NOT NULL,     FOREIGN KEY (emp_no)
REFERENCES employees (emp_no),     PRIMARY KEY (emp_no, title, from_date) );

CREATE TABLE salaries (     emp_no INT   NOT NULL,     salary INT   NOT NULL,
from_date DATE   NOT NULL,     to_date DATE   NOT NULL,     FOREIGN KEY (emp_no)
REFERENCES employees (emp_no),     PRIMARY KEY (emp_no, from_date) );
```

0.2 Queries

– Salary by employee

```
SELECT  emp.emp_no,
        emp.last_name,
        emp.first_name,
        emp.gender,
        sal.salary
FROM employees as emp
    LEFT JOIN salaries as sal
    ON (emp.emp_no = sal.emp_no)
ORDER BY emp.emp_no;
```

– Employees hired in 1986

```
SELECT first_name, last_name
FROM employees
WHERE hire_date BETWEEN '1986-01-01' AND '1986-12-31';
```

– Manager of each department `SELECT  dm.dept_no,         d.dept_name, dm.emp_no,         e.last_name,         e.first_name,         dm.from_date, dm.to_date FROM dept_manager AS dm     INNER JOIN departments AS d         ON (dm.dept_no = d.dept_no)     INNER JOIN employees AS e         ON (dm.emp_no = e.emp_no);`

– Department of each employee `SELECT  e.emp_no,         e.last_name, e.first_name,         d.dept_name FROM employees AS e     INNER JOIN dept_emp AS de         ON (e.emp_no = de.emp_no)     INNER JOIN departments AS d         ON (de.dept_no = d.dept_no) ORDER BY e.emp_no;`

– Employees whose first name is “Hercules” and last name begins with “B” `SELECT * FROM employees WHERE first_name = 'Hercules' AND last_name LIKE 'B%';`

– Employees in the Sales department `SELECT  e.emp_no,         e.last_name, e.first_name,         d.dept_name FROM employees AS e     INNER JOIN dept_emp AS de         ON (e.emp_no = de.emp_no)     INNER JOIN departments AS d         ON (de.dept_no = d.dept_no) WHERE d.dept_name = 'Sales' ORDER BY e.emp_no;`

– Employees in Sales and Development departments `SELECT  e.emp_no,         e.last_name, e.first_name,         d.dept_name FROM employees AS e     INNER JOIN dept_emp AS de         ON (e.emp_no = de.emp_no)     INNER JOIN departments AS d         ON (de.dept_no = d.dept_no) WHERE d.dept_name IN ('Sales', 'Development') ORDER BY e.emp_no;`

– The frequency of employee last names `SELECT last_name, COUNT(last_name) FROM employees GROUP BY last_name ORDER BY COUNT(last_name) DESC;`