

Census case study

INTRODUCTION TO DATABASES IN PYTHON



Jason Myers

Co-Author of Essential SQLAlchemy and
Software Engineer

Census case study

- Preparing SQLAlchemy and the database
- Loading data into the database
- Solving data science problems with queries

Part 1: preparing SQLAlchemy and the database

- Create an engine and `MetaData` object

```
from sqlalchemy import create_engine, MetaData
engine = create_engine('sqlite:///census_nyc.sqlite')
metadata = MetaData()
```

Part 1: preparing SQLAlchemy and the database

- Create and save the census table

```
from sqlalchemy import (Table, Column, String,
                        Integer, Decimal, Boolean)

employees = Table('employees', metadata,
                  Column('id', Integer()),
                  Column('name', String(255)),
                  Column('salary', Decimal()),
                  Column('active', Boolean()))

metadata.create_all(engine)
```

Let's practice!

INTRODUCTION TO DATABASES IN PYTHON

Populating the database

INTRODUCTION TO DATABASES IN PYTHON



Jason Myers

Co-Author of Essential SQLAlchemy and
Software Engineer

Part 2: populating the database

- Load a CSV file into a values list

```
values_list = []  
for row in csv_reader:  
    data = {'state': row[0], 'sex': row[1], 'age': row[2],  
           'pop2000': row[3], 'pop2008': row[4]}  
    values_list.append(data)
```

Part 2: Populating the Database

- Insert the values list into the census table

```
from sqlalchemy import insert  
  
stmt = insert(employees)  
  
result_proxy = connection.execute(stmt, values_list)  
print(result_proxy.rowcount)
```

2

Let's practice!

INTRODUCTION TO DATABASES IN PYTHON

Querying the database

INTRODUCTION TO DATABASES IN PYTHON



Jason Myers

Co-Author of Essential SQLAlchemy and
Software Engineer

Part 3: answering data science questions with queries

- Determine average age for males and females

```
from sqlalchemy import select

stmt = select([census.columns.sex,
               (func.sum(census.columns.pop2008 *
                           census.columns.age) /
                func.sum(census.columns.pop2008)
               ).label('average_age')])

stmt = stmt.group_by('census.columns.sex')

results = connection.execute(stmt).fetchall()
```

Part 3: answering data science questions with queries

- Determine the percentage of Females for each state

```
from sqlalchemy import case, cast, Float

stmt = select([
    (func.sum(
        case([
            (census.columns.state == 'New York',
             census.columns.pop2008)
        ], else_=0)) /
    cast(func.sum(census.columns.pop2008),
        Float) * 100).label('ny_percent')])
```

Part 3: answering data science questions with queries

- Determine the top 5 states by population change from 2000 to 2008

```
stmt = select([census.columns.age,  
               (census.columns.pop2008-  
                census.columns.pop2000).label('pop_change')  
            ])  
  
stmt = stmt.order_by('pop_change')  
stmt = stmt.limit(5)
```

Let's practice!

INTRODUCTION TO DATABASES IN PYTHON

Congratulations!

INTRODUCTION TO DATABASES IN PYTHON



Jason Myers

Co-Author of Essential SQLAlchemy and
Software Engineer

Congratulations!

INTRODUCTION TO DATABASES IN PYTHON