D211_Advanced_Data_Acquisition_SLM1_TASK_1_DATA_ANALYSIS

September 24, 2021

0.1 Part 2 - Item 6 Demonstrate how the databases were created

Import Python data science libraries

[ ]:
```
import pandas as pd
import sqlite3
import sql
```

Establish sqlite3 library and create a connection

[ ]:
```
cnn = sqlite3.connect('jupyter_sql_tutorial.db')
```

Load the SQL module to iPython

[ ]:
```
%load_ext sql
```

Create Jupyter database for the ETL task

[ ]:
```
%sql sqlite:///jupyter_sql_tutorial.db
```

Import tables from churn_clean

[ ]:
```
contract = pd.read_csv('data/contract.csv')
customer = pd.read_csv('data/customer.csv')
job = pd.read_csv('data/job.csv')
location = pd.read_csv('data/location.csv')
payment = pd.read_csv('data/payment.csv')
```

Load churn_clean tables into SQLite Jupyter Database

[ ]:
```
contract.to_sql('contract', cnn)
customer.to_sql('customer', cnn)
location.to_sql('location', cnn)
job.to_sql('job', cnn)
payment.to_sql('payment', cnn)
```

Select initial views of loaded churn_clean tables to get a sense of the data within

Contract table

[ ]:
```
%%sql

SELECT *
FROM contract
LIMIT 10;
```

```
 * sqlite:///jupyter_sql_tutorial.db
Done.
```

[ ]:
```
[(0, 1, 'Month-to-month'), (1, 2, 'One year'), (2, 3, 'Two Year')]
```

Customer table

[ ]:
```
%%sql

SELECT *
FROM customer
LIMIT 10;
```

```
 * sqlite:///jupyter_sql_tutorial.db
Done.
```

[ ]:
```
[(0, 1, 'K409198', 56.251000000000005, -133.37571, 38, 'Urban', 'America/Sitka',
0, 68, 28561.99, 'Widowed', 'No', 'Male', 6.795512947000001, 172.45551899999998,
904.5361101999999, 7.9783229470000006, 10, 0, 1, 'No', 'Yes', 'Yes',
'Environmental health practitioner', 'Credit Card (automatic)', 'One year',
99927),
 (1, 2, 'S120509', 44.32893, -84.2408, 10446, 'Urban', 'America/Detroit', 1, 27,
21704.77, 'Married', 'Yes', 'Female', 1.156680997, 242.632554, 800.9827661,
11.69907956, 12, 0, 1, 'Yes', 'No', 'Yes', 'Programmer, multimedia', 'Bank
Transfer(automatic)', 'Month-to-month', 48661),
 (2, 3, 'K191035', 45.35589, -123.24656999999999, 3735, 'Urban',
'America/Los_Angeles', 4, 50, 9609.57, 'Widowed', 'No', 'Female', 15.75414408,
159.947583, 2054.706961, 10.75280028, 9, 0, 1, 'Yes', 'Yes', 'No', 'Chief
Financial Officer', 'Credit Card (automatic)', 'Two Year', 97148),
 (3, 4, 'D90850', 32.96687, -117.24798, 13863, 'Suburban',
'America/Los_Angeles', 1, 48, 18925.23, 'Married', 'No', 'Male', 17.08722662,
119.95683999999999, 2164.579412, 14.91353964, 15, 2, 0, 'Yes', 'No', 'No',
'Solicitor', 'Mailed Check', 'Two Year', 92014),
 (4, 5, 'K662701', 29.38012, -95.80673, 11352, 'Suburban', 'America/Chicago', 0,
83, 40074.19, 'Separated', 'Yes', 'Male', 1.6709717259999999, 149.948316,
271.49343619999996, 8.147416533, 16, 2, 1, 'No', 'Yes', 'No', 'Medical
illustrator', 'Mailed Check', 'Month-to-month', 77461),
 (5, 6, 'W303516', 32.57032, -83.8904, 17701, 'Urban', 'America/New_York', 3,
```

```
83, 22660.2, 'Never Married', 'No', 'Female', 7.000993555, 185.007692,
1039.357983, 8.420992898, 15, 3, 1, 'No', 'Yes', 'No', 'Chief Technology
Officer', 'Electronic Check', 'One year', 31030),
 (6, 7, 'U335188', 36.4342, -84.27892, 2535, 'Suburban', 'America/New_York', 0,
79, 11467.5, 'Widowed', 'Yes', 'Male', 13.23677381, 200.118516,
1907.2429719999998, 11.18272453, 10, 0, 1, 'Yes', 'No', 'No', 'Surveyor,
hydrographic', 'Electronic Check', 'Month-to-month', 37847),
 (7, 8, 'V538685', 35.43313, -97.52463, 23144, 'Suburban', 'America/Chicago', 2,
30, 26759.64, 'Married', 'Yes', 'Female', 4.26425515, 114.950905, 979.6127078,
7.791632265, 16, 0, 0, 'Yes', 'No', 'No', 'Sales promotion account executive',
'Mailed Check', 'Month-to-month', 73109),
 (8, 9, 'M716771', 28.276459999999997, -81.16273000000001, 17351, 'Suburban',
'America/New_York', 2, 49, 58634.51, 'Separated', 'No', 'Nonbinary',
8.220686373, 117.46859099999999, 1312.874964, 5.739005915, 20, 2, 3, 'No',
'Yes', 'No', 'Teaching laboratory technician', 'Bank Transfer(automatic)',
'Month-to-month', 34771),
 (9, 10, 'I676080', 39.19296, -84.4523, 20193, 'Rural', 'America/New_York', 1,
86, 50231.4, 'Married', 'No', 'Female', 3.4220861389999997, 162.48269399999998,
508.7637913, 8.707823904, 18, 1, 0, 'No', 'Yes', 'No', 'Museum education
officer', 'Mailed Check', 'Two Year', 45237)]
```

Check for missing values in internal dataset table 'customer'

[ ]:
```
%%sql

SELECT COUNT(*) - COUNT(churn) AS missing
FROM customer;
```

```
 * sqlite:///jupyter_sql_tutorial.db
Done.
```

[ ]:
```
[(0,)]
```

Job table

[ ]:
```
%%sql

SELECT *
FROM job
LIMIT 10;
```

```
 * sqlite:///jupyter_sql_tutorial.db
Done.
```

[ ]:
```
[(0, 1, 'Academic librarian'),
 (1, 2, 'Accommodation manager'),
 (2, 3, 'Accountant, chartered'),
```

```
 (3, 4, 'Accountant, chartered certified'),
 (4, 5, 'Accountant, chartered management'),
 (5, 6, 'Accountant, chartered public finance'),
 (6, 7, 'Accounting technician'),
 (7, 8, 'Actor'),
 (8, 9, 'Actuary'),
 (9, 10, 'Acupuncturist')]
```

Location table

[ ]:
```
%%sql

SELECT *
FROM location
LIMIT 10;
```

```
 * sqlite:///jupyter_sql_tutorial.db
Done.
```

[ ]:
```
[(0, 1, 601, 'Adjuntas', 'PR', 'Adjuntas'),
 (1, 2, 610, 'Anasco', 'PR', 'Añasco'),
 (2, 3, 647, 'Ensenada', 'PR', 'Guánica'),
 (3, 4, 652, 'Garrochales', 'PR', 'Arecibo'),
 (4, 5, 662, 'Isabela', 'PR', 'Isabela'),
 (5, 6, 667, 'Lajas', 'PR', 'Lajas'),
 (6, 7, 674, 'Manati', 'PR', 'Manatí'),
 (7, 8, 683, 'San German', 'PR', 'San Germán'),
 (8, 9, 692, 'Vega Alta', 'PR', 'Vega Alta'),
 (9, 10, 694, 'Vega Baja', 'PR', 'Vega Baja')]
```

Payment table

[ ]:
```
%%sql

SELECT *
FROM Contract
LIMIT 10;
```

```
 * sqlite:///jupyter_sql_tutorial.db
Done.
```

[ ]:
```
[(0, 1, 'Month-to-month'), (1, 2, 'One year'), (2, 3, 'Two Year')]
```

Import external telco table from Kaggle and IBM

Links: Telecom Churn Prediction) Telco customer churns (11.1.3+)

[ ]:
```
telco_customer_churns = pd.read_csv('data/WA_Fn-UseC_-Telco-Customer-Churn.csv')
```

Import INTL telecom datasets tables into SQLite Jupyter Database

[ ]:
```
telco_customer_churns.to_sql('telco_customer_churns', cnn)
```

Select initial views of loaded INTL telecom tables

INTL mobile subscriber table

[ ]:
```
%%sql

SELECT *
FROM telco_customer_churns
LIMIT 10;
```

```
 * sqlite:///jupyter_sql_tutorial.db
Done.
```

[ ]:
```
[(0, '7590-VHVEG', 'Female', 0, 'Yes', 'No', 1, 'No', 'No phone service', 'DSL',
'No', 'Yes', 'No', 'No', 'No', 'No', 'Month-to-month', 'Yes', 'Electronic
check', 29.85, '29.85', 'No'),
 (1, '5575-GNVDE', 'Male', 0, 'No', 'No', 34, 'Yes', 'No', 'DSL', 'Yes', 'No',
'Yes', 'No', 'No', 'No', 'One year', 'No', 'Mailed check', 56.95, '1889.5',
'No'),
 (2, '3668-QPYBK', 'Male', 0, 'No', 'No', 2, 'Yes', 'No', 'DSL', 'Yes', 'Yes',
'No', 'No', 'No', 'No', 'Month-to-month', 'Yes', 'Mailed check', 53.85,
'108.15', 'Yes'),
 (3, '7795-CFOCW', 'Male', 0, 'No', 'No', 45, 'No', 'No phone service', 'DSL',
'Yes', 'No', 'Yes', 'Yes', 'No', 'No', 'One year', 'No', 'Bank transfer
(automatic)', 42.3, '1840.75', 'No'),
 (4, '9237-HQITU', 'Female', 0, 'No', 'No', 2, 'Yes', 'No', 'Fiber optic', 'No',
'No', 'No', 'No', 'No', 'No', 'Month-to-month', 'Yes', 'Electronic check', 70.7,
'151.65', 'Yes'),
 (5, '9305-CDSKC', 'Female', 0, 'No', 'No', 8, 'Yes', 'Yes', 'Fiber optic',
'No', 'No', 'Yes', 'No', 'Yes', 'Yes', 'Month-to-month', 'Yes', 'Electronic
check', 99.65, '820.5', 'Yes'),
 (6, '1452-KIOVK', 'Male', 0, 'No', 'Yes', 22, 'Yes', 'Yes', 'Fiber optic',
'No', 'Yes', 'No', 'No', 'Yes', 'No', 'Month-to-month', 'Yes', 'Credit card
(automatic)', 89.1, '1949.4', 'No'),
 (7, '6713-OKOMC', 'Female', 0, 'No', 'No', 10, 'No', 'No phone service', 'DSL',
'Yes', 'No', 'No', 'No', 'No', 'No', 'Month-to-month', 'No', 'Mailed check',
29.75, '301.9', 'No'),
 (8, '7892-POOKP', 'Female', 0, 'Yes', 'No', 28, 'Yes', 'Yes', 'Fiber optic',
'No', 'No', 'Yes', 'Yes', 'Yes', 'Yes', 'Month-to-month', 'Yes', 'Electronic
check', 104.8, '3046.05', 'Yes'),
 (9, '6388-TABGU', 'Male', 0, 'No', 'Yes', 62, 'Yes', 'No', 'DSL', 'Yes', 'Yes',
```

```
'No', 'No', 'No', 'No', 'One year', 'No', 'Bank transfer (automatic)', 56.15,
'3487.95', 'No')]
```

Check for missing values in external dataset table 'telco_customer_churns'

[ ]:
```
%%sql

SELECT COUNT(*) - COUNT(churn) AS missing
FROM telco_customer_churns;
```

```
 * sqlite:///jupyter_sql_tutorial.db
Done.
```

[ ]:
```
[(0,)]
```

Join internal and external tables

[ ]:
```
%%sql

SELECT customer.gender, telco_customer_churns.gender
FROM customer
INNER JOIN telco_customer_churns
ON customer.gender = telco_customer_churns.gender;
```

```
 * sqlite:///jupyter_sql_tutorial.db
Done.
```

[ ]:
```
%%sql

SELECT COUNT(c.customer_id) AS our_company, COUNT(t.customerID) AS competitor
FROM customer AS c
INNER JOIN telco_customer_churns AS t
ON c.churn = t.churn;
```

Queries for exploration and/or comparison

[ ]:
```
%%sql

SELECT
        customer_id,
        job_id
FROM customer
WHERE timezone IN ('America/Los_Angeles', 'America/New_York')
LIMIT 10;
```

[ ]:
```
%%sql
```

```
SELECT
    CASE WHEN income > 100000 THEN 'One Percenters'
    WHEN income > 50000 THEN 'Middle Class'
    ELSE 'Just Over Broke (JOB)'
    END AS social_class,
    COUNT(customer_id) AS totals
FROM customer
GROUP BY social_class
ORDER BY social_class DESC;
```

CASE WHEN . . . AND then some

[ ]:
```
%%sql

SELECT area, age, churn,
    CASE WHEN age = 35 AND tenure > 7
        THEN 'Old and stayed'
    WHEN age = 21 AND tenure <=7
        THEN 'Young and left'
    ELSE 'Outlier' END AS churn_or_stayed,
    COUNT(customer_id) AS totals
FROM customer
WHERE age = 35 OR age = 21
GROUP BY churn;
```

[ ]:
```
%%sql

SELECT area,
    CASE WHEN age >= 40 THEN 'Older'
    WHEN age < 39 THEN 'Thirty something'
    ELSE 'Munchkin'
    END AS age_category,
    COUNT(customer_id) AS totals
FROM customer
GROUP BY age_category
LIMIT 10;
```

[ ]:
```
!wget -nc https://raw.githubusercontent.com/brpy/colab-pdf/master/colab_pdf.py
from colab_pdf import colab_pdf
colab_pdf('D211_Advanced_Data_Acquisition_SLM1_TASK_1_DATA_ANALYSIS_Part_1_SQL_Code.
 ↪ipynb')
```

```
--2021-09-24 19:18:45--  https://raw.githubusercontent.com/brpy/colab-
pdf/master/colab_pdf.py
Resolving raw.githubusercontent.com (raw.githubusercontent.com)...
185.199.111.133, 185.199.108.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com
(raw.githubusercontent.com)|185.199.111.133|:443... connected.
```

```
HTTP request sent, awaiting response... 200 OK
Length: 1864 (1.8K) [text/plain]
Saving to: colab_pdf.py

colab_pdf.py        100%[===================>]   1.82K  --.-KB/s    in 0s

2021-09-24 19:18:46 (27.9 MB/s) - colab_pdf.py saved [1864/1864]

Mounted at /content/drive/

WARNING: apt does not have a stable CLI interface. Use with caution in scripts.


WARNING: apt does not have a stable CLI interface. Use with caution in scripts.

Extracting templates from packages: 100%
```

[ ]: