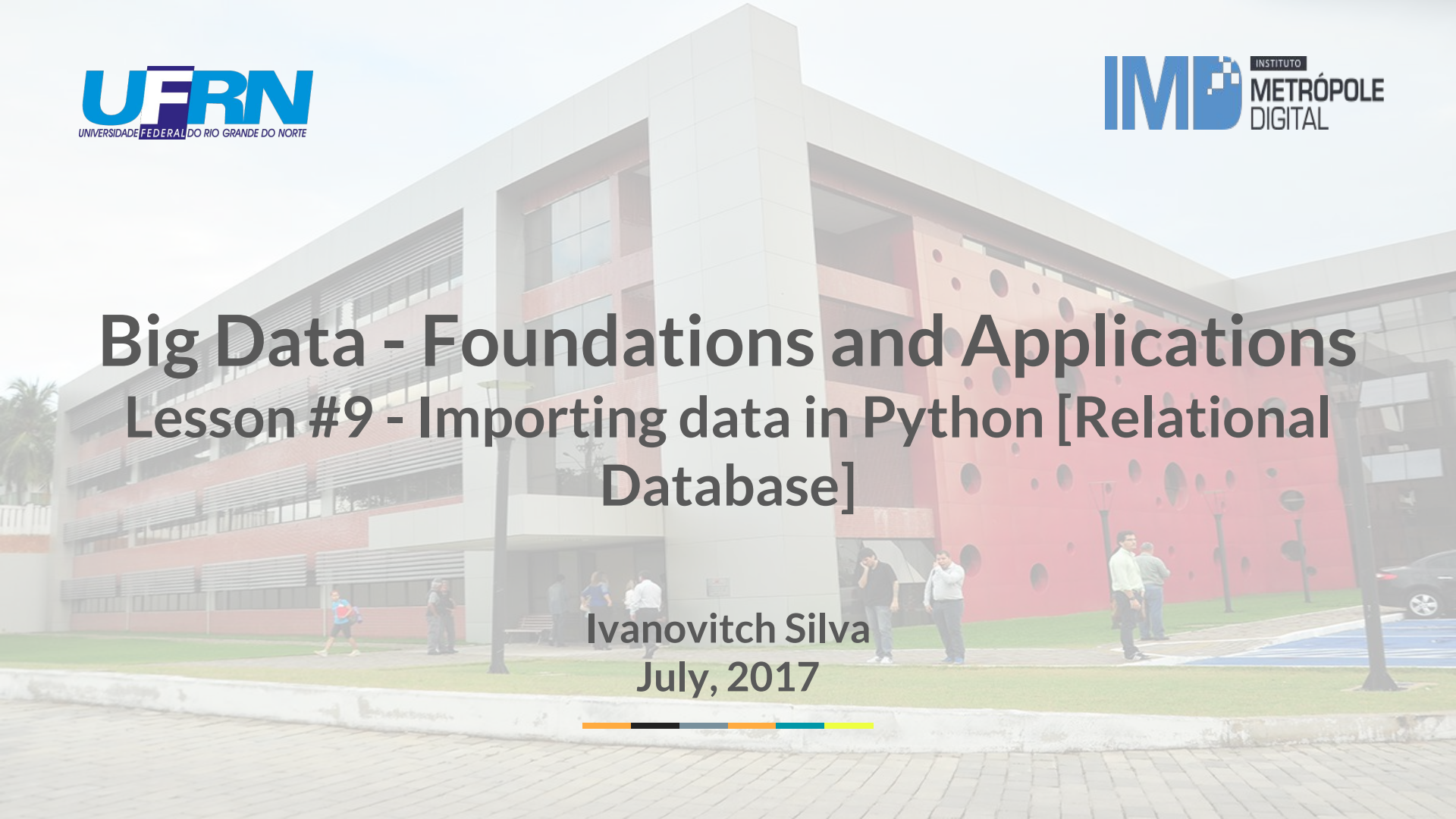


Big Data - Foundations and Applications

Lesson #9 - Importing data in Python [Relational Database]

Ivanovitch Silva

July, 2017

Agenda

- Import data
- Relational database
- Creating a engine
- Querying data
- Pandas way
- Exploiting tables relationship
- Exercises

Previously on last class (...)

Importing Data

As a Data Scientist, on a daily basis you will need:

- clean data
- wrangle and munge
- visualize
- build predictive models and interpret these models.

Before doing any of these, however, you will need to know how to get data. IMPORTING DATA!!!

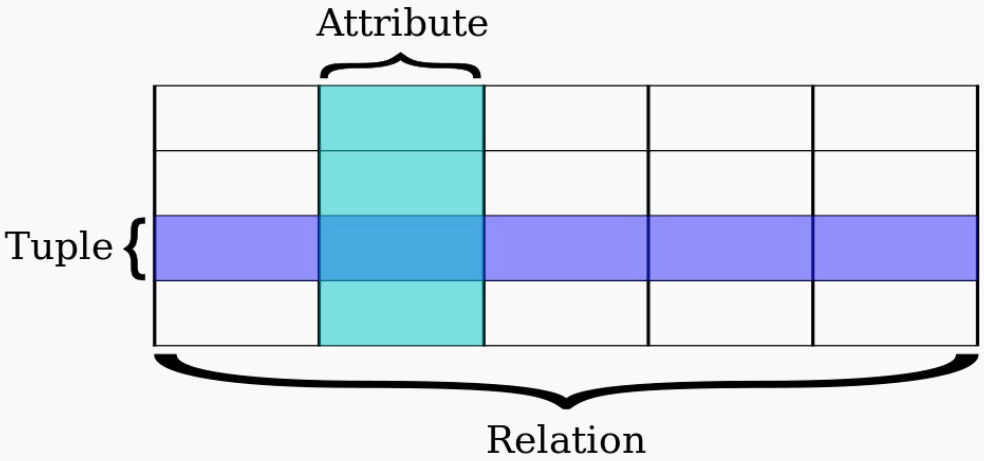
Working with relational database

We will learn how to extract meaningful data from relational databases

- creating SQL queries
- filtering and ordering your SQL records
- advanced querying by JOINing database tables

What is a relational database?

- Based on relational model of data
- First described by Edgar “Ted” Codd (**1970**)



Tables are linked

- Orders table

OrderID	CustomerID	EmployeeID	OrderDate	RequiredDate	ShippedDate	ShipVia	Freight	ShipName	ShipAddress
10248	VINET	5	7/4/1996 12:00:00 AM	8/1/1996 12:00:00 AM	7/16/1996 12:00:00 AM	3	32.38	Vins et alcools Chevalier	59 rue de l'Abbaye
10251	VICTE	3	7/8/1996 12:00:00 AM	8/5/1996 12:00:00 AM	7/15/1996 12:00:00 AM	1	41.34	Victuailles en stock	2, rue du Commerce
10254	CHOPS	5	7/11/1996 12:00:00 AM	8/8/1996 12:00:00 AM	7/23/1996 12:00:00 AM	2	22.98	Chop-suey Chinese	Hauptstr. 31

- Customers table

CustomerID	CompanyName	ContactName	ContactTitle	Address	City	Region	PostalCode	Country
ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Obere Str. 57	Berlin	None	12209	Germany
AROUT	Around the Horn	Thomas Hardy	Sales Representative	120 Hanover Sq.	London	None	WA1 1DP	UK
BLAUS	Blauer See Delikatessen	Hanna Moos	Sales Representative	Forsterstr. 57	Mannheim	None	68306	Germany
BONAP	Bon app'	Laurence Lebihan	Owner	12, rue des Bouchers	Marseille	None	13008	France

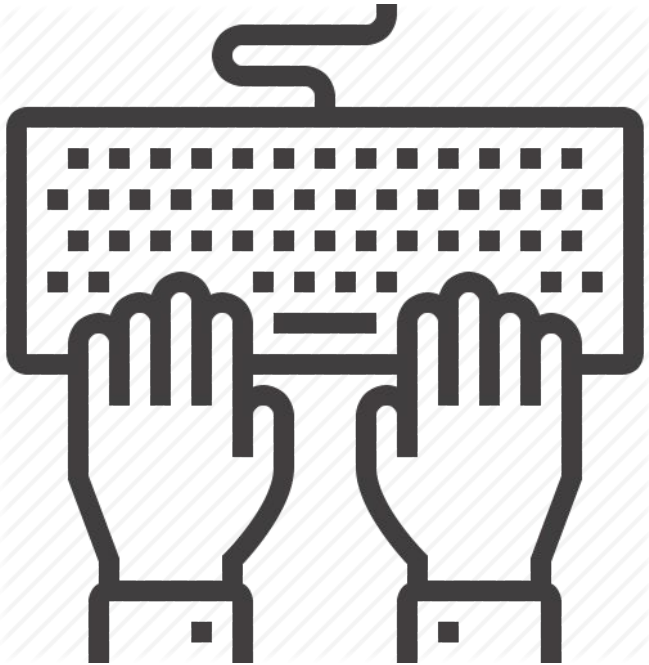
- Employees table

EmployeeID	LastName	FirstName	Title	TitleOfCourtesy	BirthDate	HireDate	Address	City	Region
1	Davolio	Nancy	Sales Representative	Ms.	12/8/1948 12:00:00 AM	5/1/1992 12:00:00 AM	507 - 20th Ave. E.\r\nApt. 2A	Seattle	WA
2	Fuller	Andrew	Vice President, Sales	Dr.	2/19/1952 12:00:00 AM	8/14/1992 12:00:00 AM	908 W. Capital Way	Tacoma	WA
3	Leverling	Janet	Sales Representative	Ms.	8/30/1963 12:00:00 AM	4/1/1992 12:00:00 AM	722 Moss Bay Blvd.	Kirkland	WA

Relational Database Management System

- PostgreSQL
- MySQL
- SQLite
- SQL = Structured Query Language

Hands On



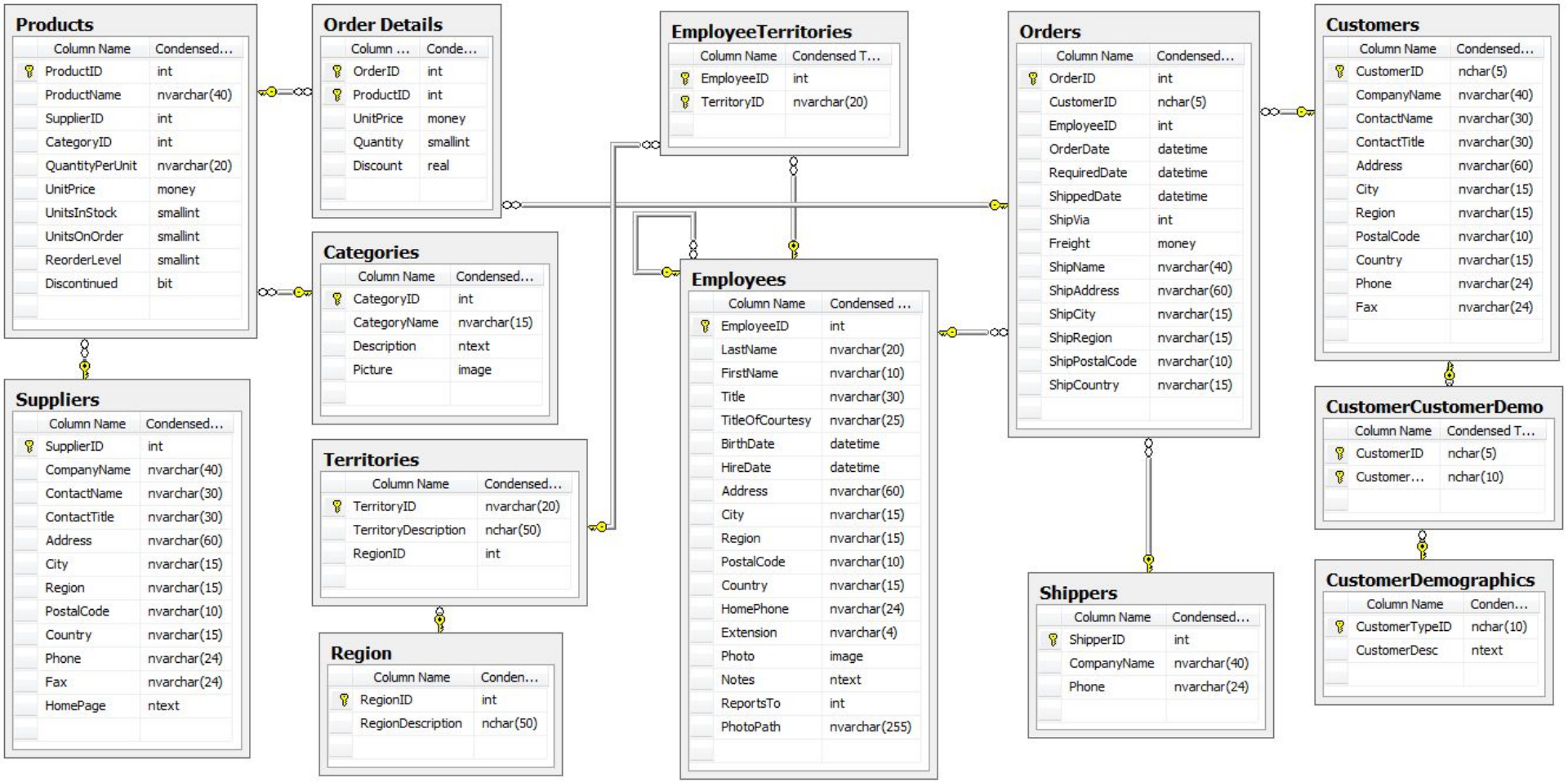
Creating a Database Engine

- SQLite database
  - Fast and simple
- SQLAlchemy
  - Works with many Relational Database Management Systems

```
In [1]: from sqlalchemy import create_engine

In [2]: engine = create_engine('sqlite:///Northwind.sqlite')
```

https://northwinddatabase.codeplex.com/


Products
Column Name Condensed...
ProductID int
ProductName nvarchar(40)
SupplierID int
CategoryID int
QuantityPerUnit nvarchar(20)
UnitPrice money
UnitsInStock smallint
UnitsOnOrder smallint
ReorderLevel smallint
Discontinued bit
Order Details
Column ... Conde...
OrderID int
ProductID int
UnitPrice money
Quantity smallint
Discount real
EmployeeTerritories
Column Name Condensed T...
EmployeeID int
TerritoryID nvarchar(20)
Orders
Column Name Condensed...
OrderID int
CustomerID nchar(5)
EmployeeID int
OrderDate datetime
RequiredDate datetime
ShippedDate datetime
ShipVia int
Freight money
ShipName nvarchar(40)
ShipAddress nvarchar(60)
ShipCity nvarchar(15)
ShipRegion nvarchar(15)
ShipPostalCode nvarchar(10)
ShipCountry nvarchar(15)
Customers
Column Name Condensed...
CustomerID nchar(5)
CompanyName nvarchar(40)
ContactName nvarchar(30)
ContactTitle nvarchar(30)
Address nvarchar(60)
City nvarchar(15)
Region nvarchar(15)
PostalCode nvarchar(10)
Country nvarchar(15)
Phone nvarchar(24)
Fax nvarchar(24)
Categories
Column Name Condensed...
CategoryID int
CategoryName nvarchar(15)
Description ntext
Picture image
Employees
Column Name Condensed ...
EmployeeID int
LastName nvarchar(20)
FirstName nvarchar(10)
Title nvarchar(30)
TitleOfCourtesy nvarchar(25)
BirthDate datetime
HireDate datetime
Address nvarchar(60)
City nvarchar(15)
Region nvarchar(15)
PostalCode nvarchar(10)
Country nvarchar(15)
HomePhone nvarchar(24)
Extension nvarchar(4)
Photo image
Notes ntext
ReportsTo int
PhotoPath nvarchar(255)
CustomerCustomerDemo
Column Name Condensed T...
CustomerID nchar(5)
Customer... nchar(10)
Suppliers
Column Name Condensed...
SupplierID int
CompanyName nvarchar(40)
ContactName nvarchar(30)
ContactTitle nvarchar(30)
Address nvarchar(60)
City nvarchar(15)
Region nvarchar(15)
PostalCode nvarchar(10)
Country nvarchar(15)
Phone nvarchar(24)
Fax nvarchar(24)
HomePage ntext
Territories
Column Name Condensed...
TerritoryID nvarchar(20)
TerritoryDescription nchar(50)
RegionID int
Shippers
Column Name Condensed...
ShipperID int
CompanyName nvarchar(40)
Phone nvarchar(24)
CustomerDemographics
Column Name Conden...
CustomerTypeID nchar(10)
CustomerDesc ntext
Region
Column Name Conden...
RegionID int
RegionDescription nchar(50)

Getting Table Names

```
In [1]: from sqlalchemy import create_engine

In [2]: engine = create_engine('sqlite:///Northwind.sqlite')
```

```
In [3]: table_names = engine.table_names()

In [4]: print(table_names)
['Categories', 'Customers', 'EmployeeTerritories',
'Employees', 'Order Details', 'Orders', 'Products',
'Region', 'Shippers', 'Suppliers', 'Territories']
```

Querying Relational Database in Python

```
SELECT * FROM Table_Name
```

- Returns all columns of all rows of the table
- Example:

```
SELECT * FROM Orders
```

- We'll use SQLAlchemy and pandas

Workflow of SQL Querying

- Import packages and functions
- Create the database engine
- Connect to the engine
- Query the database
- Save query results to a DataFrame
- Close the connection

A First SQL Query

```
In [1]: from sqlalchemy import create_engine

In [2]: import pandas as pd

In [3]: engine = create_engine('sqlite:///Northwind.sqlite')

In [4]: con = engine.connect()

In [5]: rs = con.execute("SELECT * FROM Orders")

In [6]: df = pd.DataFrame(rs.fetchall())

In [7]: con.close()
```

Print the Results

```
In [8]: print(df.head())
      0       1   2                          3                          4
0  10248   VINET   5   7/4/1996 12:00:00 AM   8/1/1996 12:00:00 AM
1  10251   VICTE   3   7/8/1996 12:00:00 AM   8/5/1996 12:00:00 AM
2  10254   CHOPS   5  7/11/1996 12:00:00 AM   8/8/1996 12:00:00 AM
3  10256   WELLI   3  7/15/1996 12:00:00 AM  8/12/1996 12:00:00 AM
4  10258   ERNSH   1  7/17/1996 12:00:00 AM  8/14/1996 12:00:00 AM
```

Set the DataFrame Column Names

```
In [1]: from sqlalchemy import create_engine

In [2]: import pandas as pd

In [3]: engine = create_engine('sqlite:///Northwind.sqlite')

In [4]: con = engine.connect()

In [5]: rs = con.execute("SELECT * FROM Orders")

In [6]: df = pd.DataFrame(rs.fetchall())

In [7]: df.columns = rs.keys()

In [8]: con.close()
```

Print the new results

```
In [9]: print(df.head())
   OrderID CustomerID  EmployeeID              OrderDate
0    10248      VINET           5   7/4/1996 12:00:00 AM
1    10251      VICTE           3   7/8/1996 12:00:00 AM
2    10254      CHOPS           5  7/11/1996 12:00:00 AM
3    10256      WELLI           3  7/15/1996 12:00:00 AM
4    10258      ERNSH           1  7/17/1996 12:00:00 AM
```

The Pandas way to query

```
In [1]: from sqlalchemy import create_engine

In [2]: import pandas as pd

In [3]: engine = create_engine('sqlite:///Northwind.sqlite')

In [4]: with engine.connect() as con:
   ...:     rs = con.execute("SELECT * FROM Orders")
   ...:     df = pd.DataFrame(rs.fetchall())
   ...:     df.columns = rs.keys()
```

```
In [5]: df = pd.read_sql_query("SELECT * FROM Orders", engine)
```

Exploiting Table Relationships

- Orders table

OrderID	CustomerID	EmployeeID	OrderDate	RequiredDate	ShippedDate	ShipVia	Freight	ShipName	ShipAddress
10248	VINET	5	7/4/1996 12:00:00 AM	8/1/1996 12:00:00 AM	7/16/1996 12:00:00 AM	3	32.38	Vins et alcools Chevalier	59 rue de l'Abbaye
10251	VICTE	3	7/8/1996 12:00:00 AM	8/5/1996 12:00:00 AM	7/15/1996 12:00:00 AM	1	41.34	Victuailles en stock	2, rue du Commerce
10254	CHOPS	5	7/11/1996 12:00:00 AM	8/8/1996 12:00:00 AM	7/23/1996 12:00:00 AM	2	22.98	Chop-suey Chinese	Hauptstr. 31

- Customers table

CustomerID	CompanyName	ContactName	ContactTitle	Address	City	Region	PostalCode	Country
ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Obere Str. 57	Berlin	None	12209	Germany
AROUT	Around the Horn	Thomas Hardy	Sales Representative	120 Hanover Sq.	London	None	WA1 1DP	UK
BLAUS	Blauer See Delikatessen	Hanna Moos	Sales Representative	Forsterstr. 57	Mannheim	None	68306	Germany
BONAP	Bon app'	Laurence Lebihan	Owner	12, rue des Bouchers	Marseille	None	13008	France

- Employees table

EmployeeID	LastName	FirstName	Title	TitleOfCourtesy	BirthDate	HireDate	Address	City	Region
1	Davolio	Nancy	Sales Representative	Ms.	12/8/1948 12:00:00 AM	5/1/1992 12:00:00 AM	507 - 20th Ave. E.\r\nApt. 2A	Seattle	WA
2	Fuller	Andrew	Vice President, Sales	Dr.	2/19/1952 12:00:00 AM	8/14/1992 12:00:00 AM	908 W. Capital Way	Tacoma	WA
3	Leverling	Janet	Sales Representative	Ms.	8/30/1963 12:00:00 AM	4/1/1992 12:00:00 AM	722 Moss Bay Blvd.	Kirkland	WA

JOINing Tables

- Orders table

OrderID	CustomerID	EmployeeID	OrderDate	RequiredDate	ShippedDate	ShipVia	Freight	ShipName	ShipAddress
10248	VINET	5	7/4/1996 12:00:00 AM	8/1/1996 12:00:00 AM	7/16/1996 12:00:00 AM	3	32.38	Vins et alcools Chevalier	59 rue de l'Abbaye
10251	VICTE	3	7/8/1996 12:00:00 AM	8/5/1996 12:00:00 AM	7/15/1996 12:00:00 AM	1	41.34	Victuailles en stock	2, rue du Commerce
10254	CHOPS	5	7/11/1996 12:00:00 AM	8/8/1996 12:00:00 AM	7/23/1996 12:00:00 AM	2	22.98	Chop-suey Chinese	Hauptstr. 31

- Customers table

CustomerID	CompanyName	ContactName	ContactTitle	Address	City	Region	PostalCode	Country
ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Obere Str. 57	Berlin	None	12209	Germany
AROUT	Around the Horn	Thomas Hardy	Sales Representative	120 Hanover Sq.	London	None	WA1 1DP	UK
BLAUS	Blauer See Delikatessen	Hanna Moos	Sales Representative	Forsterstr. 57	Mannheim	None	68306	Germany
BONAP	Bon app'	Laurence Lebihan	Owner	12, rue des Bouchers	Marseille	None	13008	France

INNER JOIN in Python (Pandas)

```
In [1]: from sqlalchemy import create_engine

In [2]: import pandas as pd

In [3]: engine = create_engine('sqlite:///Northwind.sqlite')

In [4]: df = pd.read_sql_query("SELECT OrderID, CompanyName FROM
Orders INNER JOIN Customers on Orders.CustomerID =
Customers.CustomerID", engine)

In [5]: print(df.head())
   OrderID                CompanyName
0    10248  Vins et alcools Chevalier
1    10251       Victuailles en stock
2    10254          Chop-suey Chinese
3    10256     Wellington Importadora
4    10258               Ernst Handel
```

Now, it is your turn!!!!

Notebook

https://goo.gl/SxYQfh

Analyze the performance of IMD's students

Grades related to "programming" disciplines

Notebook: https://goo.gl/oofIPX
Database: https://goo.gl/xRon1c