GENERAL ASSEMBLY

DATABASES

Abbas Chokor, Ph.D.

Staff Data Scientist, Seagate Technology

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS

- ~~What is Data Science~~ ~~Lesson 1~~
- ~~Research Design and Pandas~~ ~~Lesson 2~~
- ~~Statistics Fundamentals I~~ ~~Lesson 3~~
- ~~Statistics Fundamentals II~~ ~~Lesson 4~~
- ~~Flexible Class Session~~ ~~Lesson 5~~

UNIT 2: FOUNDATIONS OF DATA MODELING

- ~~Introduction to Regression~~ ~~Lesson 6~~
- ~~Evaluating Model Fit~~ ~~Lesson 7~~
- ~~Introduction to Classification~~ ~~Lesson 8~~
- ~~Introduction to Logistic Regression~~ ~~Lesson 9~~
- ~~Communicating Logistic Regression Results~~ ~~Lesson 10~~
- ~~Flexible Class Session~~ ~~Lesson 11~~

UNIT 3: DATA SCIENCE IN THE REAL WORLD

- ~~Decision Trees and Random Forests~~ ~~Lesson 12~~
- ~~Natural Language Processing~~ ~~Lesson 13~~
- ~~Dimensionality Reduction~~ ~~Lesson 14~~
- ~~Time Series Data I~~ ~~Lesson 15~~
- ~~Time Series Data II~~ ~~Lesson 16~~
- Database Technologies Lesson 17 ← Today's Class
- Where to Go Next Lesson 18
- Flexible Class Session Lesson 19
- Final Project Presentations Lesson 20

LAST CLASS

WHAT DID WE LEARN?

- Model and predict from time series data using AR, ARMA, or ARIMA models
- Specifically, coding these models in `statsmodels`

LEARNING OBJECTIVES

- Understanding of the uses and differences of databases
- Accessing databases from Pandas

OPENING

DATABASES

DATABASES

- Today's lesson will be on databases and the SQL query language.

- Databases are the standard solution for data storage. They're far more robust than text and CSV files.

- They come in many flavors, but we'll explore the most common: *relational databases*.

DATABASES

- Relational databases also come in different varieties, but almost all use SQL as a basis for querying (i.e. retrieving) data.

- Most analyses typically involve pulling data from a database.

DATABASES

DATABASES

- Databases are computer systems that manage the storage and querying of datasets.

- They provide a way to organize the data on disk (i.e. hard drive) and efficient methods to retrieve information. Databases allow a user to create rules that ensure proper data management and verification.

- Typically, retrieval is performed using a query language, a mini programming language with a few basic operators for data transformation.

- The most common query language is SQL (Structured Query Language).

DATABASES

- A *relational database* is based on links between data entities or concepts.

- Typically, a relational databases is organized into *tables*.

- Each table should correspond to one entity or concept. Each table is similar to a single CSV file or Pandas dataframe.

- For example, consider an application like Twitter. Our two main entities are Users and Tweets. For each of these, we would have a separate table.

DATABASES

- A table is made up of rows and columns, similar to a Pandas dataframe or Excel spreadsheet.

- Each table has a specific *schema*, a set of rules for what goes in each table. These specify which columns are contained in the table and what *type* of data is in each column (e.g. text, integers, decimals, etc).

Users Table Schema	
user_id	char
user_sign_up_date	date
user_follower_count	int

DATABASES

- This means you can’t add text data to an integer column in that database.
- The additional *type* information make this constraint stronger than the header of a CSV file.
- For this reason and many others, databases allow for stronger consistency of the data and are often a better solution for data storage.

DATABASES

- Each table typically has a *primary* key column. This column has a unique value per row and serves as the identifier for the row.

- A table can have many *foreign keys* as well. A *foreign key* is a column that contains values to link the table to the other tables.

- These keys that link the table together define the relational database.

DATABASES

- For example, the tweets table may have as columns:
  - tweet_id - the primary key tweet identifier
  - tweet_text
  - user_id - a foreign key to the users table

Users Table Schema	
user_id	char
user_sign_up_date	date
user_follower_count	int

Tweets Table Schema	
tweet_id	int
tweet_text	char
user_id	int

DATABASES

- MySQL and Postgres are popular variants of relational databases and are widely used. Both are open-source and available for free.
- Alternatively, many companies use proprietary software such as Oracle or Microsoft SQL databases.
- While these databases offer many of the same features and use the same SQL language, the latter two offer some maintenance features and support that large companies find useful.

NORMALIZED VS DENORMALIZED DATA

- Once we start organizing our data into tables, we start to separate it into *normalized* and *denormalized* setups.

- *Normalized* structures have a single table per entity and use many foreign keys or link tables to connect the entities.

- *Denormalized* structures have fewer tables that combine different entities.

NORMALIZED VS DENORMALIZED DATA

- With our Twitter example, a *normalized* structure would place users and tweets in different tables.

Users Table Schema	
user_id	char
user_sign_up_date	date
user_follower_count	int

Tweets Table Schema	
tweet_id	int
tweet_text	char
user_id	int

NORMALIZED VS DENORMALIZED DATA

- A *denormalized* structure would put them both in one table.

Twitter Table Schema	
tweet_id	int
tweet_text	char
user_id	int
user_sign_up_date	date
user_follower_count	int

NORMALIZED VS DENORMALIZED DATA

Denormalized structures:

- Duplicates a lot of information
- Makes data easy to access since it's all in one table

Normalized structures:

- Save storage space by separating information
- Requires joining of table to access information about two different entities, a slow operation

ALTERNATIVE DATABASES

- While relational databases are the most popular and broadly used, specific applications may require different data organization.

- You don’t need to know every variety, but it’s good to know some overall themes.

KEY-VALUE STORES

- Key-Value databases are nothing more than very large and very fast hashmaps or dictionaries.

- These are useful for storing key based data, e.g. a count of things per user or customer, a last visit per customer.

- Every entry in these databases has two values, a key and a value. We can retrieve any value based upon its key.

KEY-VALUE STORES

- This is exactly like a python dictionary, but it can be larger than your memory (i.e. RAM). So these systems use smart caching algorithms to ensure frequently or recently accessed items are quickly accessible.

- Popular key-value stores include *Cassandra* and *MemcacheDB* (pronounced mem-cash-dee-bee).

NOSQL OR DOCUMENT DATABASES

- "NoSQL" databases are those that don't rely on a traditional relational table setup and more flexible in their data organization.

- Typically they actually **do** have SQL querying abilities but model their data differently.

NOSQL OR DOCUMENT DATABASES

- Relational Structure

user_id	user_name	user_hobby_1	user_hobby_2	user_age
13123	robby_g	guitar	cars	25
18423	jt1235	football		31

- NoSQL Data Structure

```
{
    “user_id”: 13123,
        “user_name”: “robby_g”,
        “user_hobbies”: [“guitar”,
“cars”],
        “user_age”: 25
}
```

```
{
    “user_id”: 19423,
        “user_name”: “jt1235”,
        “user_hobbies”: [“football”],
        “user_age”: 31
}
```

NOSQL OR DOCUMENT DATABASES

- They may organize data on an entity level, but often have denormalized and nested data setups.

- This nested data layout is often similar to that in JSON documents.

- Popular databases include *MongoDB* and *CouchDB*.

NOSQL OR DOCUMENT DATABASES

Relational data model

Highly-structured table organization with rigidly-defined data formats and record structure.

Document data model

Collection of complex documents with arbitrary, nested data formats and varying “record” format.

NOSQL OR DOCUMENT DATABASES

- The following is an example of the storage document for a tweet.


```
{
  "created_at": "Mon Sep 24 03:35:21 +0000 2012",
  "id_str": "250075927172759552",
  "entities": {
    "hashtags": [
      {
        "text": "freebandnames",
        "indices": [
          20,
          34
        ]
      }
    ],
    "user_mentions": [

    ]
  }
}
```

DEMO

ACCESSING DATABASES FROM PANDAS

ACCESSING DATABASES FROM PANDAS

- While databases provide many analytical capabilities, often it’s useful to pull the data back into Python for more flexible programming.

- Large, fixed operations would be more efficient in a database, but Pandas allows for interactive processing.

- For example, if you just want to aggregate login or sales data to present a report or dashboard, this operation is operating on a large dataset and not often changing.

- This would run very efficiently in a database vs connecting to Python.

ACCESSING DATABASES FROM PANDAS

- However, if we want to investigate the login or sales data further and ask more interactive questions, then using Python would come in very handy.

```
import pandas as pd
from pandas.io import sql
```

- Pandas can be used to connect to most relational databases.

ACCESSING DATABASES FROM PANDAS

- In this demonstration, we will create and connect to a SQLite database. SQLite creates portable relational databases saved in a single file.

- These databases are stored in a very efficient manner and allow fast querying, making them ideal for small databases or databases that need to be moved across machines.

- Additionally, SQLite databases can be created with the setup of MySQL or Postgres databases.

ACCESSING DATABASES FROM PANDAS

- We can create a SQLite databases as follows.

```
import sqlite3
```

```
conn = sqlite3.connect('dat-test.db')
```

- This creates a file, `dat-test.db`, which will act as a relational/SQL database.

WRITING DATA INTO A DATABASE

- Data in Pandas can be loaded into a relational database. For the most part, Pandas can use the databases column information to infer the schema for the table it creates.

- Let’s return to the Rossmann sales data and load it into our database.

```
import pandas as pd

data = pd.read_csv('../../datasets/rossmann.csv', low_memory=False)
data.head()
```

WRITING DATA INTO A DATABASE

- Data is moved to the database with the `to_sql` command, similar to the `to_csv` command.

- `to_sql` takes several arguments.
  - `name` - the table name to create
  - `con` - a connection to a database
  - `index` - whether to input the index column
  - `schema` - if we want to write a custom schema for the new table
  - `if_exists` - what to do if the table already exists. We can overwrite it, add to it, or fail

WRITING DATA INTO A DATABASE

- The following code loads the Rossmann sales data to our database.

```
data.to_sql('rossmann_sales',
            con=conn,
            if_exists='replace',
            index=False)
```

READING FROM A DATABASE

- If we already have data in the database, we can use Pandas to query our database.

- Querying is done through the `read_sql` command in the `sql` module.

```
import pandas as pd
from pandas.io import sql

sql.read_sql('select * from rossmann_sales limit 10', con=conn)
```

- This runs the query passed in and returns a dataframe with the results.

DEMO

SQL SYNTAX: SELECT, WHERE, GROUP BY, JOIN

SQL OPERATORS: SELECT

- Every query should start with `SELECT`. `SELECT` is followed by the names of the columns in the output.

- `SELECT` is always paired with `FROM`, which identifies the table to retrieve data from.

```
SELECT <columns>
FROM <table>
```

- `SELECT *` denotes returning *all* of the columns.

SQL OPERATORS: SELECT

- Rossmann Stores example:

```
SELECT Store, Sales
FROM rossmann_sales;
```

SQL OPERATORS: WHERE

- `WHERE` is used to filter a table using a specific criteria. The `WHERE` clause follows the `FROM` clause.

```
SELECT <columns>
FROM <table>
WHERE <condition>
```

- The condition is some filter applied to the rows, where rows that match the condition will be output.

SQL OPERATORS: WHERE

- Rossmann Stores example:

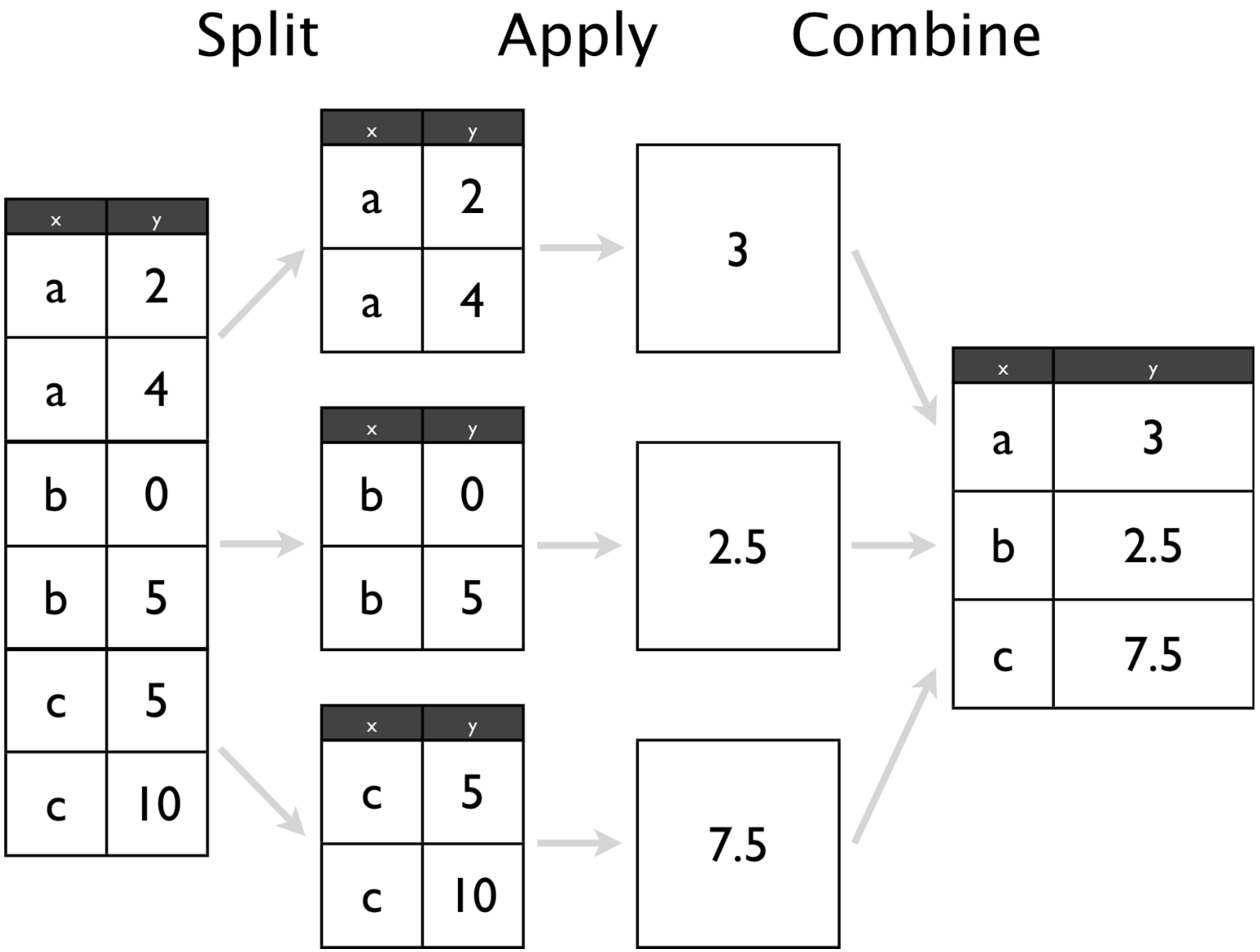
```
SELECT Store, Sales
FROM rossmann_sales
WHERE Store = 1;
```

```
SELECT Store, Sales
FROM rossmann_sales
WHERE Store = 1 and Open = 1;
```

SQL OPERATORS: GROUP BY

- `GROUP BY` allows us to aggregate over any field in the table by applying the concept of Split Apply Combine.

- We identify some key with which we want to segment the rows. Then, we roll up or compute some statistics over all of the rows that match that key.



SQL OPERATORS: GROUP BY

- `GROUP BY` *must* be paired with an aggregate function, the statistic we want to compute in the rows, in the `SELECT` statement.

- `COUNT(*)` denotes counting up all of the rows. Other aggregate functions commonly available are `AVG` (average), `MAX`, `MIN`, and `SUM`.

SQL OPERATORS: GROUP BY

- Rossmann Stores example:

```
SELECT Store, SUM(Sales), AVG(Customers)
FROM rossmann_sales
WHERE Open = 1
GROUP BY Store;
```

SQL OPERATORS: ORDER BY

- `ORDER BY` is used to sort the results of a query.

```
SELECT <columns>
FROM <table>
WHERE <condition>
ORDER BY <columns>
```

- You can order by multiple columns in ascending (`ASC`) or descending (`DESC`) order.

SQL OPERATORS: ORDER BY

- Rossmann Stores example:

```
SELECT Store, SUM(Sales) as total_sales, AVG(Customers)
FROM rossmann_sales
GROUP BY Store
WHERE Open = 1;
ORDER BY total_sales desc;
```

- `COUNT(*) AS cnt` renames the `COUNT(*)` value to `cnt` so we can refer to it later in the `ORDER BY` clause.

SQL OPERATORS: JOIN

- `JOIN` allows us to access data across many tables. We specify how a row in one table links to another.

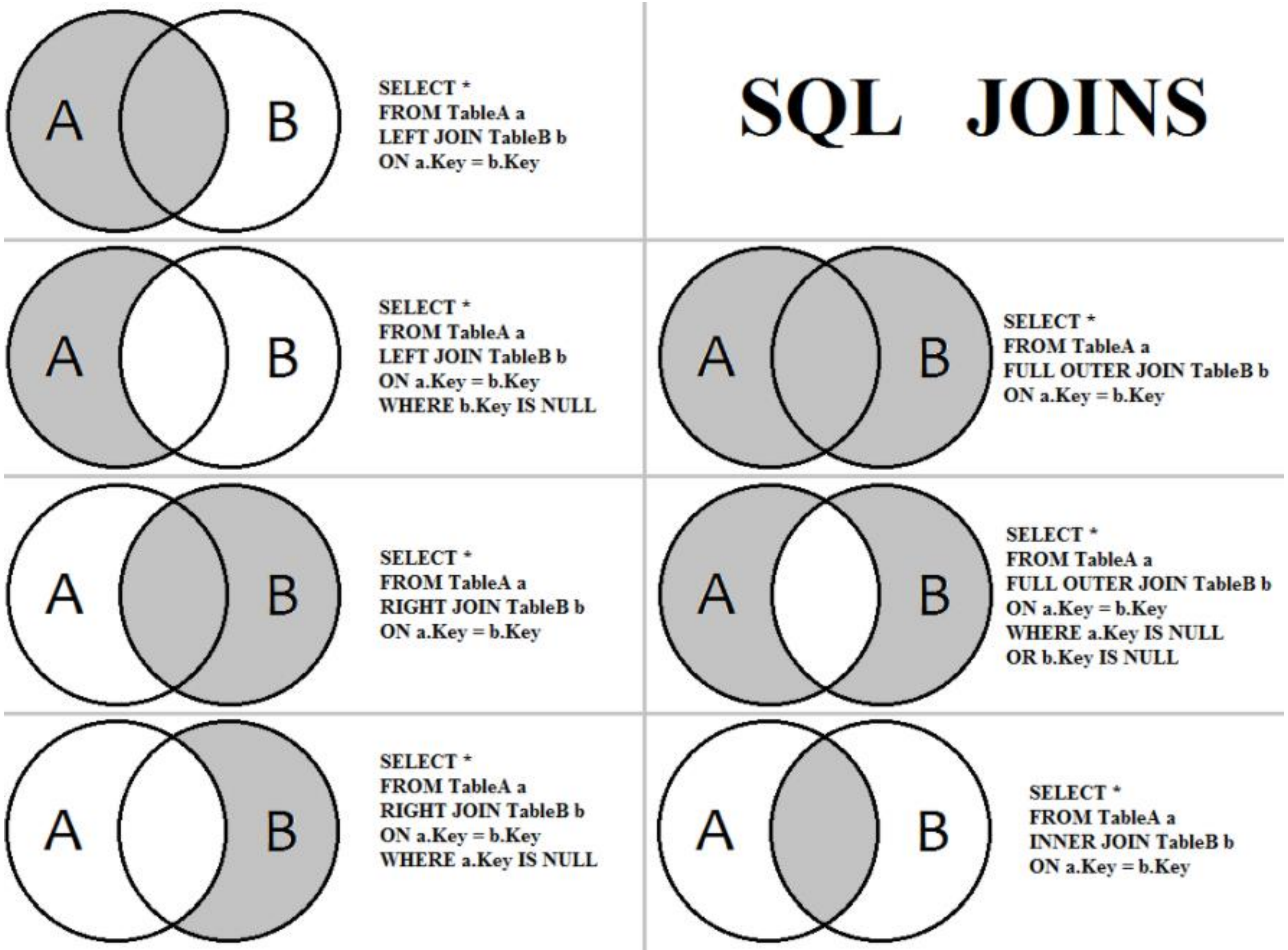
```sql
SELECT a.Store, a.Sales, s.CompetitionDistance
FROM rossmann_sales a
JOIN rossmann_stores s
ON a.Store = s.Store
```

- Here, ON denotes an *inner* join.

SQL OPERATORS: JOIN

- By default, most joins are an *Inner Join*, which means only when there is a match in both tables does a row appear in the results.
- If we want to keep the rows of one table *even if there is no matching counterpart*, we can perform an *Outer Join*.
- Outer joins can be `LEFT`, `RIGHT`, or `FULL`, meaning keep all of the left rows, all the right rows, or all the rows, respectively.

SQL OPERATORS: JOIN



INDEPENDENT PRACTICE

PANDAS AND SQL

ACTIVITY: PANDAS AND SQL

EXERCISE

DIRECTIONS (40 minutes)

1. Load the Walmart sales and store features data.
2. Create a table for each of those datasets.
3. Select the store, date and fuel price on days it was over 90 degrees.
4. Select the store, date and weekly sales and temperature.
5. What were average sales on holiday vs. non-holiday sales?
6. What were average sales on holiday vs. non-holiday sales when the temperature was below 32 degrees?

DELIVERABLE

Answers to the above questions

CONCLUSION

TOPIC REVIEW

CONCLUSION

- While this was a brief introduction, databases are often at the core of any data analysis. Most analysis starts with retrieving data from a database.

- SQL is a key language that any data scientist should understand.

  - `SELECT`: Used in every query to define the resulting columns

  - `WHERE`: Filters rows based on a given condition

  - `GROUP BY`: Groups rows for aggregation

  - `JOIN`: Combines two tables based upon a given condition

CONCLUSION

- Pandas can be used to access data from databases as well. The result of the queries will end up in a Pandas dataframe.

- There is much more to learn about query optimization if one dives further!

COURSE

BEFORE NEXT CLASS

DUE DATE

- **Project**: Final Project, Part 4

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS	~~What is Data Science~~	~~Lesson 1~~
	~~Research Design and Pandas~~	~~Lesson 2~~
	~~Statistics Fundamentals I~~	~~Lesson 3~~
	~~Statistics Fundamentals II~~	~~Lesson 4~~
	~~Flexible Class Session~~	~~Lesson 5~~
UNIT 2: FOUNDATIONS OF DATA MODELING	~~Introduction to Regression~~	~~Lesson 6~~
	~~Evaluating Model Fit~~	~~Lesson 7~~
	~~Introduction to Classification~~	~~Lesson 8~~
	~~Introduction to Logistic Regression~~	~~Lesson 9~~
	~~Communicating Logistic Regression Results~~	~~Lesson 10~~
	~~Flexible Class Session~~	~~Lesson 11~~
UNIT 3: DATA SCIENCE IN THE REAL WORLD	~~Decision Trees and Random Forests~~	~~Lesson 12~~
	~~Natural Language Processing~~	~~Lesson 13~~
	~~Dimensionality Reduction~~	~~Lesson 14~~
	~~Time Series Data I~~	~~Lesson 15~~
	~~Time Series Data II~~	~~Lesson 16~~
	~~Database Technologies~~	~~Lesson 17~~
	Where to Go Next	Lesson 18
	Flexible Class Session	Lesson 19
	Final Project Presentations	Lesson 20

Next Class

LESSON

Q & A

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET