Data Science in Business:
Data Structures & Analysis

Dr. Peter Molnar

Sarah Zeis

Carly Wieting

Georgia State University | J. Mack Robinson College of Business

Robinson

Course Overview

Class 1: Introduction to Machine Learning and Set-Up Python

Class 2: Data Exploration

Class 3: Machine Learning Models (Decision Tree and KNN)

Class 4: Analyze Celebrity Tweets

Class 5: Forecasting with Facebook Prophet

Goals and Takeaways

- Identifying Data Needs: Case Walkthrough
- Types of Data
- Using data in Python
- Python Notebook Lab

Data Science Begins with Questions!

A question will be asked before any analysis beings!

Asking the right questions can impact:

- The structure of the data you are analyzing
- The feasibility > accuracy > success of a project

Scenario:
You are trying to schedule an advertising campaign for a pizza chain to boost delivery orders!
Your manager asks you: “What is the most popular day for pizza delivery?”

… and Data Science continues with questions

What day of the week do people order delivery pizza the most?

 Sunday 

How can we ask more questions to analyze pizza consumption?

- Is Sunday the most popular day every week?
- Is Sunday the most popular day for every region / state / country?
- Which Sunday has the most pizza deliveries? Is this consistent YoY?
- How do people order delivery pizza on Sundays?

Translating Questions and Logic into Data

What pizza data was needed to create the chart to the right?

- Pizza sales data
- Domino's pizza sales data
- Day pizza of pizza sale
- % of pizza sales
- # of pizzas sold or $ of pizza sold

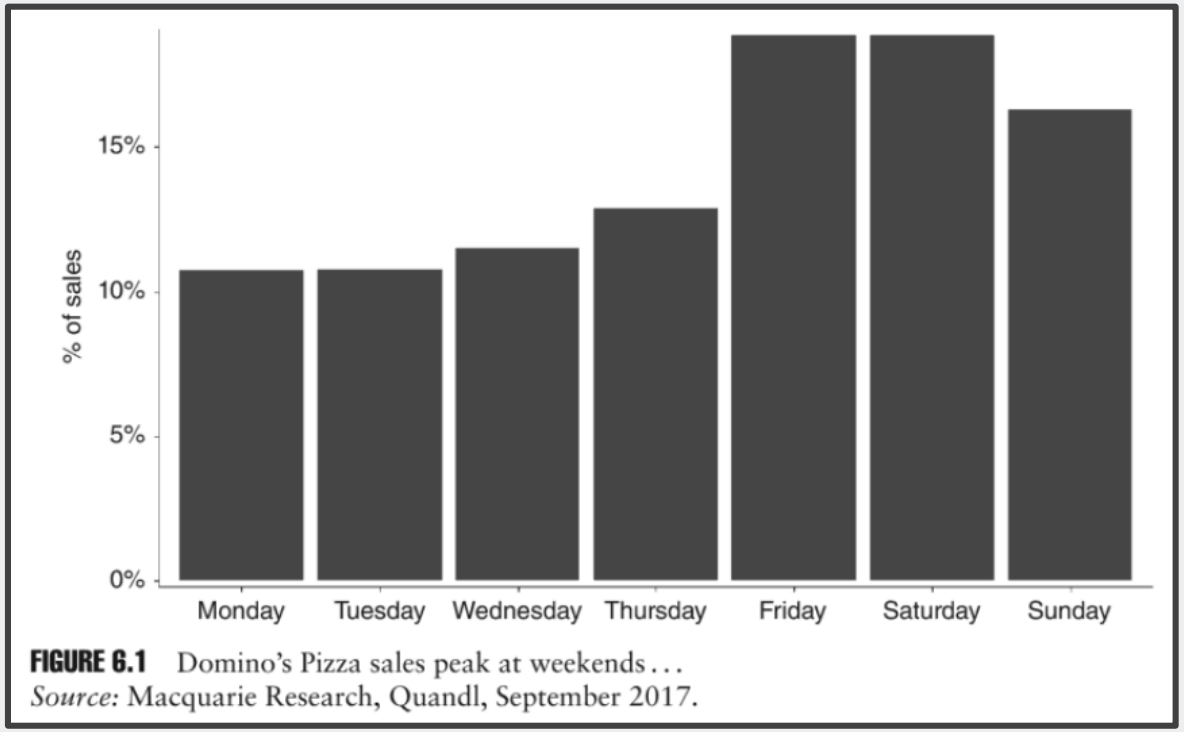


FIGURE 6.1 Domino's Pizza sales peak at weekends . . .
Source: Macquarie Research, Quandl, September 2017.

Types of Data

Structured: data stored within a fixed and defined field

Ex: Database recording each transaction at a retail establishment

Unstructured: information that is not organized in a defined way

Ex: Yelp text reviews of Atlanta based Domino's restaurants

Time-Series: a series of data collected in successive / sequential intervals

Ex: Day-end stock price of GOOGL last month

Real time: information delivered immediately after collection

Ex: # of users on a website right now

Big Data?

Pizza Advertisement Plan Data

	A	B	C	D
1	Day of Week	% of Total Sales	$ Sales	Count of Sales
2	Sunday	15%	$ 750,000.00	21,496
3	Monday	10%	$ 500,000.00	17,825
4	Tuesday	10%	$ 500,000.00	6,546
5	Wednesday	11%	$ 550,000.00	14,339
6	Thursday	12%	$ 600,000.00	21,818
7	Friday	20%	$ 1,000,000.00	33,117
8	Saturday	20%	$ 1,000,000.00	15,000

Structured Unstructured Time-Series Real time

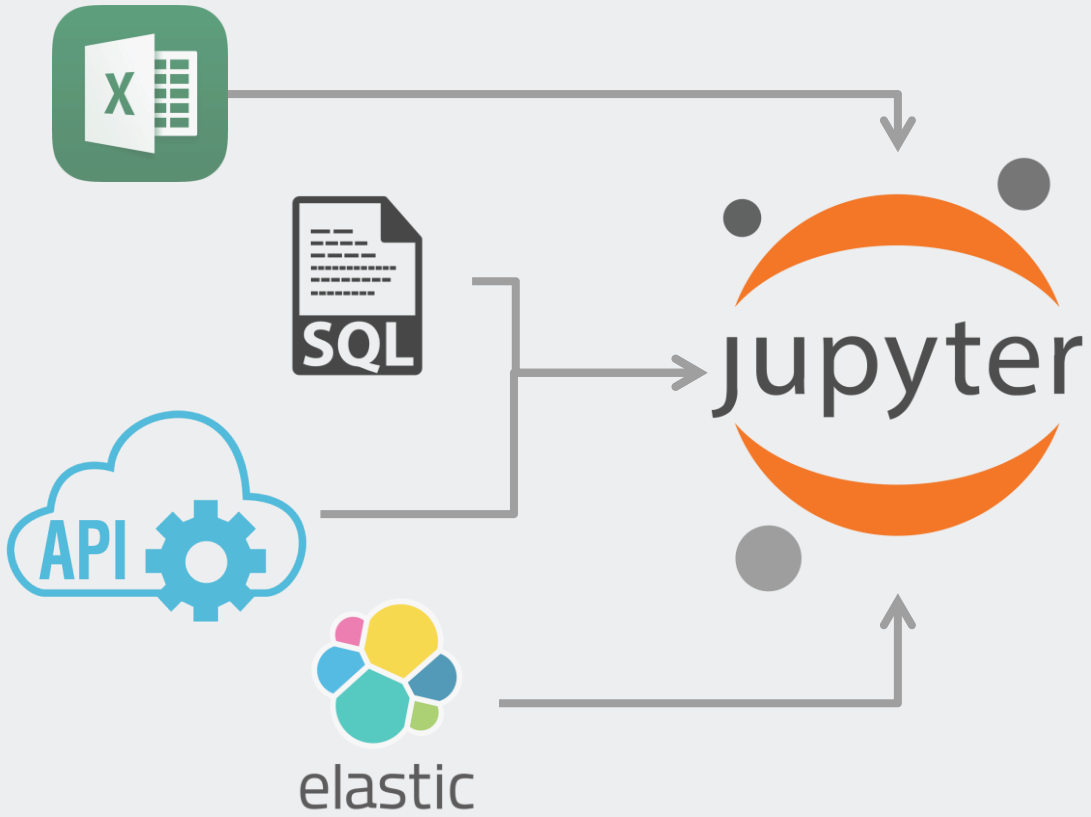
Accessing & Acquiring Data

Files (.txt, .csv, .xlsx)

Databases (SQL, NOSQL)

Web Sourcing (API)

Datastreams (Elasticsearch)



In this course, we will provide CSV Files containing data

Data Types: Tabular Data

SQL, csv, xlsx

Easiest to work with

	A	B	C	D
1	Day of Week	% of Total Sales	$ Sales	Count of Sales
2	Sunday	15%	$ 750,000.00	21,496
3	Monday	10%	$ 500,000.00	17,825
4	Tuesday	10%	$ 500,000.00	6,546
5	Wednesday	11%	$ 550,000.00	14,339
6	Thursday	12%	$ 600,000.00	21,818
7	Friday	20%	$ 1,000,000.00	33,117
8	Saturday	20%	$ 1,000,000.00	15,000
9				

Data Types: Hierarchical Data

JSON

Common in web / software development

There are packages to translate this information into tabular data


```
{
  "business_id": "PK6aSizckHFWk8i0oxt5DA",
  "full_address": "400 Waterfront Dr E\nHomestead\nHomestead, PA 15120",
  "hours": {},
  "open": true,
  "categories": [
    "Burgers",
    "Fast Food",
    "Restaurants"
  ],
  "city": "Homestead",
  "review_count": 5,
  "name": "McDonald's",
  "neighborhoods": [
    "Homestead"
  ],
  "longitude": -79.910032,
  "state": "PA",
  "stars": 2,
```

Questioning your Data

After defining the question of your analysis and retrieving your data, review the quality of your data.

Examples:

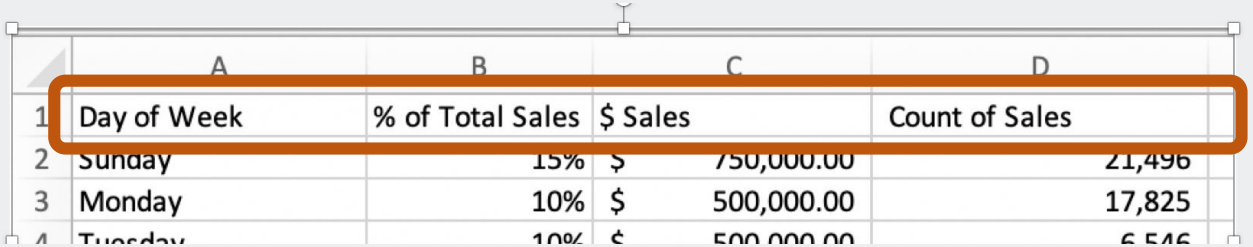
- Where did this come from? Is it accurate?
- Is all the information you need present in the current dataset?
- Are all fields populated as expected?
- Do I know enough about the context of the problem?
- **How accurate is our pizza data?**

Exploratory Data Analysis

Exploratory data analysis (EDA) is the use of statistical tools and methods to understand the **features** of a dataset.

In the pizza sales dataset, there are four features:

- Day of Week
- % of Total Sales
- $ Sales
- Count of Sales



	A	B	C	D
1	Day of Week	% of Total Sales	$ Sales	Count of Sales
2	Sunday	15%	$ 750,000.00	21,496
3	Monday	10%	$ 500,000.00	17,825
4	Tuesday	10%	$ 500,000.00	6,546

With a new dataset, you should start EDA by understanding the **observations** for each feature. Then move on to understanding the relationships between the different features.

Types of Features in a Dataset

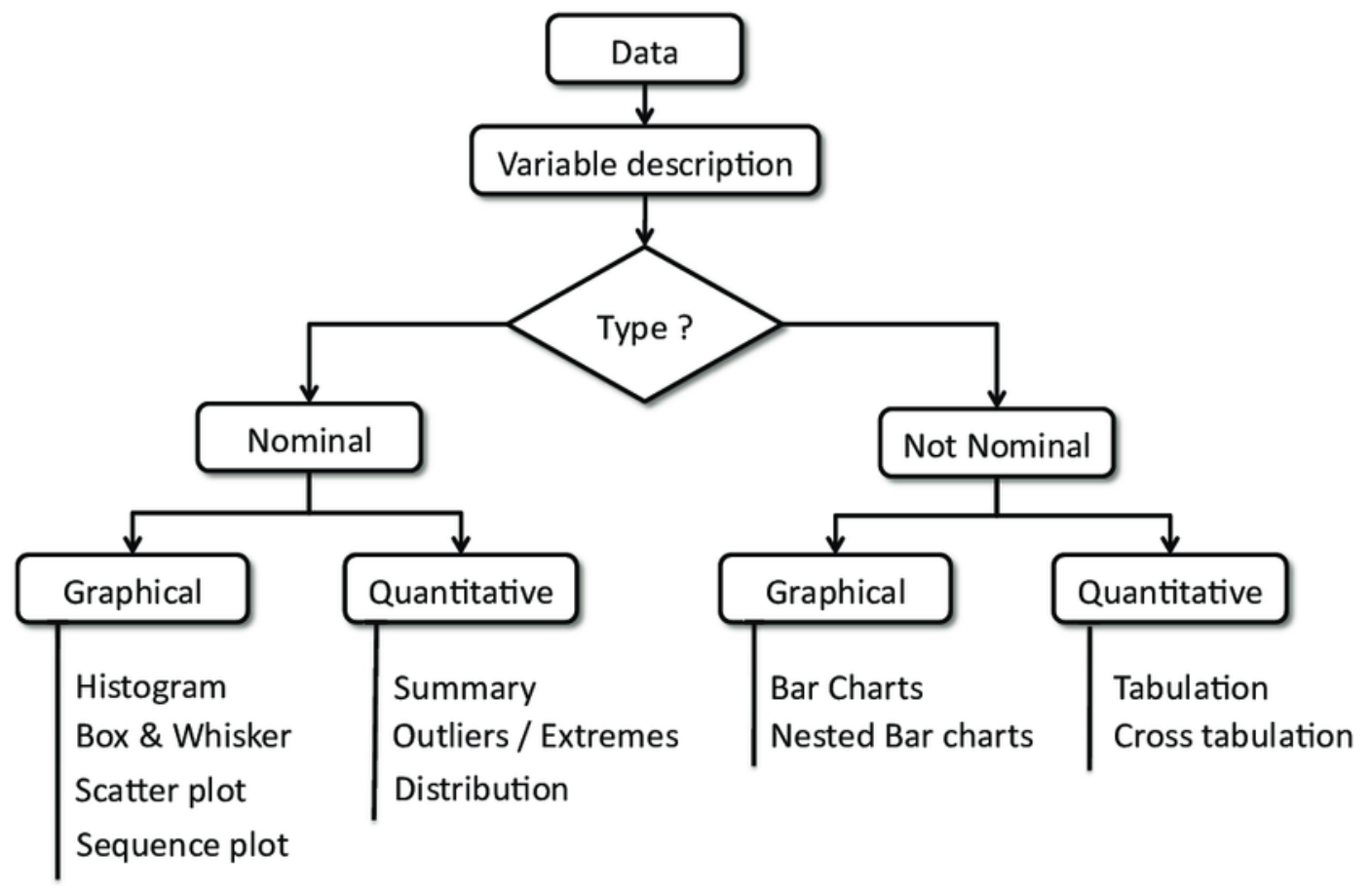
Categorical Data: Represents characteristics of data like color and grade. Can be represented with numbers or strings.

- Nominal: Used to label variables – they do not have an order or hierarchy (ie. Gender, Language)
- Ordinal: Used to represent variables that are ordered and discrete (ie. School Level)

Numerical Data: Represents numbers

- Discrete: Numerical data that can be classified and cannot be measured (ie. 6 cats)
- Continuous: Represents measurements (ie. 6.5254 lbs)
- Interval: Discrete numerical data that has a specific and same difference (ie. 5, 10, 15, 20)
- Ratio: Same as interval data but includes 0 (ie. 0, 5, 10, 15, 20)

Techniques for Exploratory Data Analysis



Understanding Data using Python

There are many packages and approaches that can assist in analyzing data in Python. In this course, we will use the Pandas package.

- Note: Pandas is built on another popular data package, Numpy
- Numpy slices and indexes data using arrays (lists)

	Day of Week	% of Total Sales	$ Sales	Count of Sales
0	Sunday	0.15	750000	28117.946350
1	Monday	0.10	500000	7955.062466
2	Tuesday	0.10	500000	14191.926220
3	Wednesday	0.11	550000	9194.807966
4	Thursday	0.12	600000	15746.016670
5	Friday	0.20	1000000	13297.475780
6	Saturday	0.20	1000000	27584.681920

Pandas uses a **Dataframe** data structure

Dataframes look like tables in a csv or excel file

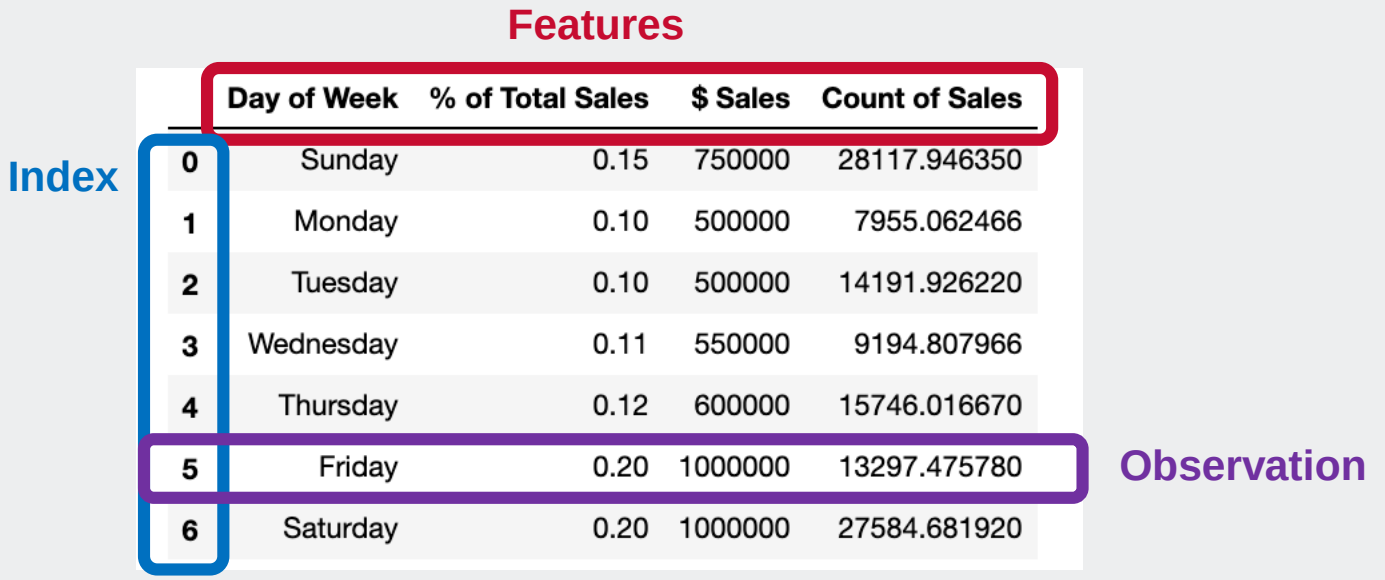
Reviewing your Pandas Dataframe

Features: 4

Observations: 6

Index: Position within an ordered list

Index values: [0:6]



	Day of Week	% of Total Sales	$ Sales	Count of Sales
0	Sunday	0.15	750000	28117.946350
1	Monday	0.10	500000	7955.062466
2	Tuesday	0.10	500000	14191.926220
3	Wednesday	0.11	550000	9194.807966
4	Thursday	0.12	600000	15746.016670
5	Friday	0.20	1000000	13297.475780
6	Saturday	0.20	1000000	27584.681920

Python Notebook Review

Open the Class 2 Python Notebook!