MS&E 125: Intro to Applied Statistics

Data Munging

Professor Udell

Management Science and Engineering
Stanford

April 11, 2023

Announcements

- hw 1 due Friday
- quiz 1 Friday
- convert colab to pdf
- complete participation before subsequent class
- section today

Outline

Messy data

SQL

Data types

- continuous values (*e.g.*, 4.2, π)
- discrete values (*e.g.*, 0, 4, 994)
- nominal values (*e.g.*, apple, banana, pear)
- ordinal values (*e.g.*, rarely, sometimes, often)
- graphs or networks (*e.g.*, person 1 is friends with person 2)
- text (*e.g.*, doctor's note describing symptoms)
- sets (*e.g.*, items purchased)

Messy data

- heterogeneous: values of many different types
- missing: some values are missing, inconsistent, not recorded, or lost
- noise: some (or all) values suffer errors, inaccuracies, or malicious corruption
- duplicated values

Data cleaning

- remove duplicates
- remove missing values
- remove noise
- convert to a single type (usually, numeric)

how? by taking a careful look...

Demo

https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/fires.ipynb

Outline

Messy data

SQL

SQL

- most data is stored in relational databases
- Structured Query Language (SQL) is a language for querying relational databases
- we will use pandas in python, not SQL
- but if you know the ideas, you can easily write SQL queries

SQL: example

team_name	player_name	player_height
Los Angeles Lakers	LeBron James	6'9"
Boston Celtics	Jaylen Brown	6'6"
Chicago Bulls	Zach LaVine	6'7"
Miami Heat	Jimmy Butler	6'6"
San Antonio Spurs	DeMar DeRozan	6'6"
Golden State Warriors	Stephen Curry	6'3"
Houston Rockets	Christian Wood	6'10"
Dallas Mavericks	Luka Doncic	6'7"

"ChatGPT, write an sql query to find the NBA team with the highest average height."

SQL: example

team_name	player_name	player_height
Los Angeles Lakers	LeBron James	6'9"
Boston Celtics	Jaylen Brown	6'6"
Chicago Bulls	Zach LaVine	6'7"
Miami Heat	Jimmy Butler	6'6"
San Antonio Spurs	DeMar DeRozan	6'6"
Golden State Warriors	Stephen Curry	6'3"
Houston Rockets	Christian Wood	6'10"
Dallas Mavericks	Luka Doncic	6'7"

"ChatGPT, write an sql query to find the NBA team with the highest average height."

```
SELECT team_name, AVG(player_height) AS avg_height
FROM nba_teams
GROUP BY team_name
ORDER BY avg_height DESC
LIMIT 1;
```

SQL poll

match the query to the question:

A. What is the average sales per month for each product line?
B. Which product line has the highest sales?
C. Which country is most profitable?

1.
```
SELECT product_line, SUM(sales) AS total_sales
FROM orders
GROUP BY product_line
ORDER BY total_sales DESC
LIMIT 1;
```

2.
```
SELECT country, SUM(sales) AS total_profit
FROM orders
GROUP BY country
ORDER BY total_profit DESC
LIMIT 1;
```

3.
```
SELECT product_line, AVG(sales) AS avg_monthly_sales
FROM orders
GROUP BY product_line, YEAR(orderdate), MONTH(orderdate)
ORDER BY product_line;
```

SQL-style munging

- select rows
- select columns
- on condition
- sort
- group (aggregate using a function)
- join (combine tables)

Demo

https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/join.ipynb