# *Lecture 2 : Preparing Data for Analysis*

*Sakib Anwar*

AN7914 Data Analytics and Modelling

University of Winchester

2023

# Motivation

- Does immunization of infants against measles save lives in poor countries? To answer that question you can use data on immunization rates in various countries in various years. The World Bank collects such information, and a lot more, on each country for multiple years that is free to download. But how should you store, organize and use the data to have all relevant information in an accessible format that lends itself to meaningful analysis?

- You want to know, who has been the best manager (as coaches are sometimes called in football) in the top English football league. To investigate this question, you have downloaded data on football games played in a professional league, and data on managers including which team they worked at and when. To answer your question you need to combine this data. How should you do that? And are there issues with the data that you need to address?

# Variable types: Qualitative vs quantitative

- Data can be born (collected, generated) in different form, and our variables may capture the quality or the quantity of a phenomenon.
- *Quantitative* variables are born as numbers. Typically take many values.
  - also called numeric variables
  - special case is time (date)
- *Qualitative* variables, also called categorical variables, take on a few values, with each value having a specific interpretation (belonging a category).
  - Another name used is categorical or factor variable.
  - binary variable (YES/NO) is a special case.

# Variable types - binary

- A special case is a *binary variable*, which can take on two values
- *Yes/No* answer to whether the observation belongs to some group.
- Best to represent these as 0 or 1 variables: 0 for no, 1 for yes.
- Sometimes binary variables with 0-1 values are called *dummy variables* or an *indicator*
- For flagging some issue - binary showing the existence of some issue (such as missing value for another variable, presence in another dataset)

# Variable types - formal definition

1. Nominal qualitative
2. Ordinal
3. Interval
4. Ratio

# Variable Types: Detailed Definitions - Part 1

*Nominal (Qualitative) Variables*

- *Characteristics:* Categories without order. Labels attributes.
- *Example:* Types of fruit, Gender.
- *Key Point:* Differences are not mathematically meaningful.

*Ordinal Variables*

- *Characteristics:* Categories with a clear order but inconsistent intervals.
- *Example:* Education level, Customer satisfaction.
- *Key Point:* Enables ranking, but distances between ranks are not quantified.

# Variable Types: Detailed Definitions - Part 2

*Interval Variables*

- *Characteristics:* Numeric values with meaningful distances but no true zero.
- *Example:* Temperature (Celsius/Fahrenheit), Calendar years.
- *Key Point:* Differences are meaningful, but ratios are not.

*Ratio Variables*

- *Characteristics:* Numeric values with meaningful distances and a true zero.
- *Example:* Weight, Age.
- *Key Point:* Allows for full range of mathematical operations, including ratios.

# Data wrangling (data munging)

*Data wrangling* is the process of transforming raw data to a set of data tables that can be used for a variety of downstream purposes such as analytics.

[1] Understanding and storing

- start from raw data
- understand the structure and content
- create tidy data tables
- understand links between tables

[2] Data cleaning

- understand features, variable types
- filter duplicates
- look for and manage missing observations
- understand limitations

# The tidy data approach

A useful concept of organizing and cleaning data is called the *tidy data* approach:

1. Each observation forms a row.
2. Each variable forms a column.
3. Each type of observational unit forms a table.
4. Each observation has a unique identifier (ID)

Advantages:

- standard data tables that turn out to be easy to work with.
- finding errors and issues with data are usually easier with tidy data tables
- transparent, which helps other users to understand
- easy to extend – new observations added as new rows; new variables as new columns.

# Simple tidy data table

Table: A simple tidy table

| | Variables/columns | | |
|---|---|---|---|
| | hotel_id | price | distance |
| Observations/rows | 21897 | 81 | 1.7 |
| | 21901 | 85 | 1.4 |
| | 21902 | 83 | 1.7 |

Source: hotels-vienna data. Vienna, 2017 November weekend.

# Organizing Multi-Dimensional Data: Tidy Data Principles

- *Tidy Data Approach:*
  - Organize data into tables where each row represents a single observation at a specific time point, referred to as '*it*'.
  - This method is known as the *long format*.
  - Each row corresponds to one cross-sectional unit observed in one time period.
  - Successive rows may represent the same unit at subsequent time periods.
  - Challenges include discerning the multi-dimensional structure and constructing tidy data tables.
- *Alternative: Wide Format*
  - In this format, each row corresponds to one cross-sectional unit, with different time periods represented across multiple columns.
  - Preferred for certain types of presentation and analysis, though not ideal for data storage.

*Note:* The choice between long and wide formats depends on the analysis needs and data presentation requirements.

# Displaying immunization rates across countries

- *xt* panel of countries with yearly observations,
- Downloaded from the World Development Indicators data website maintained by the World Bank.
- Illustrate the data structure focusing on the two ID variables (country and year) and two other variables, immunization rate and GDP per capita.

# Displaying immunization rates across countries – WIDE

| Country | imm2015 | imm2016 | imm2017 | gdppc2015 | gdppc2016 | gdppc2017 |
|---|---|---|---|---|---|---|
| India | 87 | 88 | 88 | 5743 | 6145 | 6516 |
| Pakistan | 75 | 75 | 76 | 4459 | 4608 | 4771 |

Wide format of country-year panel data, each row is one country, different years are different variables. imm: rate of immunization against measles among 12–13-month-old infants. gdppc: GDP per capital, PPP, constant 2011 USD. Source: `world-bank-vaccination` data

# Displaying immunization rates across countries – LONG

| Country | Year | imm | gdppc |
|---|---|---|---|
| India | 2015 | 87 | 5743 |
| India | 2016 | 88 | 6145 |
| India | 2017 | 88 | 6516 |
| Pakistan | 2015 | 75 | 4459 |
| Pakistan | 2016 | 75 | 4608 |
| Pakistan | 2017 | 76 | 4771 |

Note: Tidy (long) format of country-year panel data, each row is one country in one year. imm: rate of immunization against measles among 12–13-month-old infants. gdppc: GDP per capital, PPP, constant 2011 USD. Source: `world-bank-vaccination` data.

# Relational Database

- *Concept:* A method to organize data in a structured way, using tables to represent different sets of related information.
- *Structure:*
  - Tables consist of rows and columns.
  - Each row represents a unique record (observation) with a distinct/unique identifier (*ID* or *key*).
  - Columns store attributes related to each record.
- *Relationships:*
  - Records across tables can be interconnected through *foreign IDs*.
  - This allows for the linking of related information stored in separate tables.
- *Operations:*
  - Define and design tables to accurately reflect data relationships.
  - *Merge* or *join* tables based on common IDs to aggregate or compare related data.

# Identifying Successful Football Managers

*Objective:*

- Assess who the best football managers in England have been over the past 11 seasons.

*Data Sources:*

- Analysis based on combined data from two sources: teams/games and managers.
- Spanning 11 seasons of English Premier League (EPL) from 2008/2009 to 2018/2019.
- Sourced from `www.football-data.co.uk`.

*Data Details:*

- Each observation corresponds to a single game.
- *Key Variables:*
  - Date of the game.
  - Names of the home and away teams.
  - Goals scored by the home and away teams.

# Identifying successful football managers

Table: Games data

| Date | HomeTeam | AwayTeam | Home goals | Away goals |
|---|---|---|---|---|
| 2018-08-19 | Brighton | Man United | 3 | 2 |
| 2018-08-19 | Burnley | Watford | 1 | 3 |
| 2018-08-19 | Man City | Huddersfield | 6 | 1 |
| 2018-08-20 | Crystal Palace | Liverpool | 0 | 2 |
| 2018-08-25 | Arsenal | West Ham | 3 | 1 |
| 2018-08-25 | Bournemouth | Everton | 2 | 2 |
| 2018-08-25 | Huddersfield | Cardiff | 0 | 0 |

Source: `football` data.

# Identifying successful football managers

- Is this a tidy data table?

# Identifying successful football managers

- Is this a tidy data table?
- It is!
- Each observation is a game, and each game is a separate row in the data table.
- Three ID variables identify each observation: date, home team, away team. The other variables describe the result of the game.
- From the two scores we know who won, by what margin, how many goals they scored, and how many goals they conceded.

# Identifying successful football managers

- Could we have an alternative tidy table?

# Identifying successful football managers

- Could we have an alternative tidy table?
- There is an alternative way to structure the same data table, which will serve our analysis better
- In this data table, each row is a game played by a team.
- It includes variables from the perspective of that team: when played, who the opponent was, and what the score was.

# Identifying successful football managers

Table: Games data - long table version

| Date | Team | Opponent team | Goals | Opponent goals | Home/away | Points |
|---|---|---|---|---|---|---|
| 2018-08-19 | Brighton | Man United | 3 | 2 | home | 3 |
| 2018-08-19 | Burnley | Watford | 1 | 3 | home | 0 |
| 2018-08-19 | Man City | Huddersfield | 6 | 1 | home | 3 |
| 2018-08-19 | Man United | Brighton | 2 | 3 | away | 0 |
| 2018-08-19 | Watford | Burnley | 3 | 1 | away | 3 |
| 2018-08-19 | Huddersfield | Man City | 1 | 6 | away | 0 |

# Identifying successful football managers

- Also a tidy data table, albeit a different one.
- It has twice as many rows as the original data table: Each game appears twice in this data table, once for each of the playing team's perspectives.
- New variable to denote whether the team at that game was the home team or the away team.
- Now we have two ID variables, one denoting the team, and one denoting the date of the game. The identity of the opponent team is a qualitative variable.

- Tidy data has some key features. But a given multi-dimensional data may be stored as tidy in multiple ways.

# Identifying successful football managers

- Our second data table is on managers.
- One row is one manager-team relationship.
- Each manager may feature more than once in this data table if they worked for multiple teams.
- For each observation, we have the name of the manager, their nationality, the name of the team (club), the start time of the manager's work at the team, and the end time.

# Identifying successful football managers

Table: Managers data

| Name | Nat. | Club | From | Until |
|---|---|---|---|---|
| Arsene Wenger | France | Arsenal | 1 Oct 1996 | 13 May 2018 |
| Unai Emery | Spain | Arsenal | 23 May 2018 | Present* |
| Ron Atkinson | England | Aston Villa | 7 June 1991 | 10 Nov 1994 |
| Brian Little | England | Aston Villa | 25 Nov 1995 | 24 Feb 1998 |
| John Gregory | England | Aston Villa | 25 Feb 1998 | 24 Jan 2002 |
| Dean Smith | England | Aston Villa | 10 Oct 2018 | Present* |
| Alan Pardew | England | Crystal Palace | 2 Jan 2015 | 22 Dec 2016 |
| Alan Pardew | England | Newcastle | 9 Dec 2010 | 2 Jan 2015 |

Source: `football` data. Present = 01 July 2019

# Identifying successful football managers

- This is a relational dataset.
- One data table with team-game observations, and one data table with manager-team observations.
- To work with the data, we need to create a workfile, which is a single data table that is at the team-game level with the additional variable of who the manager was at the time of that game.
- but before we do, need need to merge them...

# Relational data and linking data tables across observations

- Organize and store data in tidy data tables with appropriate ID variables,
- how to combine such table into a workfile to run our analysis?
- The process of pulling different variables from different data tables for well-identified entities to create a new data table is called *linking*, *joining*, *merging*, or *matching*.

# Matching Relational Data

*Matching (Joining) Depends on Data Structure*

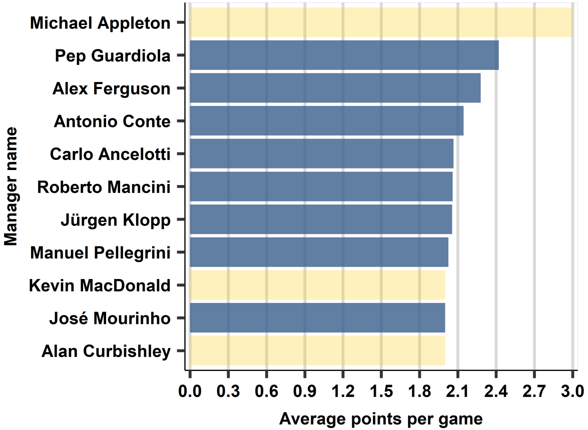
- *One-to-One (1:1) Matching*: Merging tables with the same type of observations.
  - Example: Matching Football Teams with Stadiums.
- *Many-to-One (m:1) Matching* or *One-to-Many (1:m) Matching*: When a value in one table may match to more than one value in the other table.
  - Example: Matching Football Teams with Their Players (Many Players in a Team).
- *Many-to-Many (m:m) Matching*: When values in both tables could match to many others.
  - Example: Matching Football Teams with Their Managers (Some Managers Worked for Multiple Teams).

# Identifying successful football managers

- Started with a relational dataset.
- Merged two data tables and created a work file
- With the workfile at hand, we can describe it.
- The workfile has 8360 team-game observations: in each of the 11 seasons, 20 teams playing 38 games (19 opponent teams twice; $11 \times 20 \times 38 = 8360$).
- There are 137 managers in the data.

# Identifying successful football managers

- Remember: data is 11 seaons, EPL.
- spells at teams: if a manager worked for two teams, we consider it two cases.
- Success: average points per game
- Above 2.0



# Complex data - tidy data- summary

- Creating a tidy data - generating a set of data tables that are easy to understand, combine and extend in the future.
- If relational data, IDs are essential
- Often raw data will not come in a tidy format, and you will need to work understanding the structure, relationships and find the individual ingredients.
- For analysis work, need to combine tidy data tables

# Data wrangling: cleaning

- Entity resolution:
  - Dealing with duplicates
  - ambiguous identification
  - non-entity rows
- Missing values

# Wrangling: Filter out Duplicates

- *duplicates*: some observations appearing more than once in the data.
- Duplicates may be the result of human error or the features of data source
- Often, easy process. Just check and get rid of repeated observations
- Sometimes, same observation is featured number of times.

# Entity identification and resolution

- More generally, you would need to have unique IDs
- It could be that two observations belong to two entities although ID is the same.
  - example: John Smith – there may be many
  - need to figure out, maybe assign unique IDs in raw data
- It could be that two observations have different ID but belong to same entity
  - need to figure out and have a single ID
- Unique IDs crucial. Numerical IDs are better

# Entity resolution - football example

| Team ID | Unified name | Original name |
|---|---|---|
| 19 | Man City | Manchester City |
| 19 | Man City | Man City |
| 19 | Man City | Man. City |
| 19 | Man City | Manchester City F.C. |
| 20 | Man United | Manchester United |
| 20 | Man United | Manchester United F.C. |
| 20 | Man United | Manchester United Football Club |
| 20 | Man United | Man United |

Source: various sources

# Getting rid of non-entity observations

- Rows that do not belong to an entity we want in the data table.
- Find them and drop them
- Such as: a summary row in a table that adds up, or averages, variables across all, or some, entities.
- a data table downloaded from the World Bank on countries often includes observations on larger regions, such as Sub-Saharan Africa

# Missing values

- A frequent and important issue with variables is *missing values*.
- Missing values mean that the value of a variable is not available for some, but not all, observations.

# Missing values

Key issues

1. Look at content of data - related to data quality (esp. coverage)
2. Missing values need to be identified.
3. Most software can have a value=missing.
4. Missing values should be counted. Missing values mean fewer observations with valid information.
5. The third issue is potential *selection bias*. Is data missing at random?

# Missing values - Understanding the selection process

- Random: When missing data really means no information, it may be the result of errors in the data collection process. Rare!
- Often, values are missing systematically. Some survey respondents may not know the answer to a question or refuse to answer it, and such respondents are likely to be different from those who provide valid answers.

# Missing values: what can we do?

Two basic options:

1. Restrict the analysis to observations with non-missing values for all variables used in the analysis.
2. *Imputation* - Fill in some value for the missing values, such as the mean or median value. Be very careful with this!!
3. Be conservative.

# Data wrangling: common steps

1. Write a code - it can be repeated and improved later
2. Understand the structure of the dataset, create data tables, recognize links. Draw a schema.
3. Start by looking into the data table(s) to spot issues
4. Store data in tidy data tables. Make sure one row in the data is one observation and manage duplicates
5. Get each variable in an appropriate format
6. Have a description of variables
7. Make sure values are in meaningful ranges; correct non-admissible values or set them as missing
8. Identify missing values and store them in an appropriate format. Make edits if needed.
9. Document every step of data cleaning