Lecture 2

Data in Real Estate

Haoyu Yue / yohaoyu@washington.edu
Ph.D. Student, Interdisciplinary Urban Design and Planning
University of Washington

RE 519 Real Estate Data Analytics and Visualization
Course Website: www.yuehaoyu.com/data-analytics-visualization/
Autumn 2025

What is Data

Quantitative vs. Qualitative

This Class is Focusing on Quantitative Data

Data is a collection of facts, numbers, words, observations, or other useful information.

Quantitative or qualitative refers more to the way we study and analyze data, rather than to the data alone.

Quantitative Data: values that can be measured numerically, such as population, housing prices.

Qualitative Data: descriptive and non-numerical, capturing characteristics, concepts, or experiences that numbers cannot measure, such as an interview transcript or a survey response.

Methods

→ Statistics, Modeling, Simulations, etc.

→ Case Study, Text Analysis, Grounded Theory, etc.

What is Data

Levels of Measurement

- **Nominal**
  - Numbers or other symbols are assigned to a set of categories for naming, labeling, or classifying the observations.
  - Nominal categories cannot be rank-ordered, such as religion, gender, etc.
- **Ordinal**
  - Nominal levels that can be ranked from low to high, such as very satisfactory, satisfactory, and not satisfactory.
  - Be careful when you process ordinal data
- **Interval-Ratio**
  - All cases are expressed in the same units, such as age, income, and GRE scores.
  - Cumulative property: Variables that can be measured at the interval-ratio level of measurement can also be measured at the ordinal and nominal level.

Group	Level of satisfaction (from respondents)	Satisfaction score (encoding from left column)	Average score
A	Very satisfactory	3	1.667
A	Not satisfactory	1	
A	Not satisfactory	1	
B	Not satisfactory	1	1.667
B	Satisfactory	2	
B	Satisfactory	2	

If we compare group A and B using average scores, we assume they are equally spaced!

Some Data Formats

Tabular Data / Dataframe

Data that is presented in columns or rows. Most commonly used file type: *CSV (comma-separated values)*.

Common Data Types in Columns

- Numeric
  - Integer - 67
  - Float – 12.76486
- Character/String – “Real Estate”
- Data and Time – 2025/09/15
- Binary – TRUE/FALSE
- Others
  - Geography
  - Image
  - …

Unique ID (Also called primary key in a database)

Each column represents an attribute of this record. We call each column a **variable or feature.**

Note: When processing data in R/Python, we should avoid using spaces between words. We can use underline or no space instead, such as: AllSectors / all_sector.

ID	Entity	Type of provider	All sectors	Residential	Commercial	Industrial	Transportation
1	Consolidated Edison Co-NY Inc	Investor-owned	20,710,482	11,344,135	9,313,179	51,643	1,525
2	Long Island Power Authority	Public	18,291,720	9,535,372	8,494,216	0	262,132
3	New York Power Authority	Public	17,611,535	0	7,536,115	7,606,846	2,468,574
4	Niagara Mohawk Power Corp.	Investor-owned	15,172,227	10,510,294	3,703,277	958,656	0
5	Constellation NewEnergy, Inc	Retail power marketer	14,871,737	1,578,041	9,492,345	3,781,969	19,382

Top five retailers of electricity in the State of New York, with end-use sectors, 2021.

Each row represents a record we observe. We call each row an **observation**.

Some Data Formats

Time Series Data

A time series is a series of data points indexed (or listed or graphed) in time order.

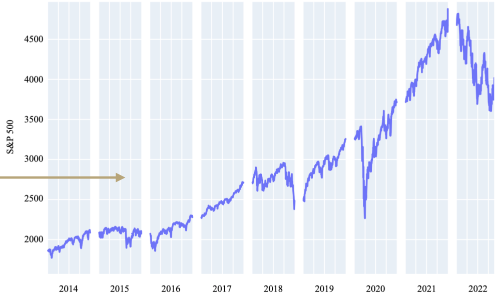
Temporal order with consistent intervals, a month in this case

Region	1/31/00	2/29/00	3/31/00	4/30/00	5/31/00
New York, NY	220834.76256321764	221773.1857048372	222720.30033586253	224639.52106624088	226626.94059518576
Los Angeles, CA	222015.51137538892	222841.691	223942.1538408446	226131.70568265315	228526.38896950847
Chicago, IL	156057.927	156202.39142660523	156477.5260008363	157161.99244626865	157985.32325153003
Dallas, TX	128589.52523097747	128646.7994546991	128712.64203812725	128883.62831123867	129109.06903064142
Houston, TX	124446.06013047668	124469.24061036404	124382.35154965024	124434.28868529193	124482.1990285256

Zillow Housing Price Index in a few US Cities. Source: Zillow Research

Time dependency: it is clear that the latter data depends on the previous data.

Sometimes, it is possible to find some **trends or repeating patterns** directly from the visualization.



S&P from 2014-2022. Source: Visualize ML: https://github.com/Visualize-ML

Time series data is important for financial and business applications. We can do forecasting, pattern recognition, and trend analysis.

Some courses at UW:
STAT 519 Time Series Analysis
CSSS 512 Time Series and Panel Data for the Social Sciences

Some Data Formats

Cross-sectional Data

Data collected by observing many subjects at a single point or period of time.

Region	1/31/00	2/29/00	3/31/00	4/30/00	5/31/00
New York, NY	220834.76256321764	221773.1857048372	222720.30033586253	224639.52106624088	226626.94059518576
Los Angeles, CA	222015.51137538892	222841.691	223942.1538408446	226131.70568265315	228526.38896950847
Chicago, IL	156057.927	156202.39142660523	156477.5260008363	157161.99244626865	157985.32325153003
Dallas, TX	128589.52523097747	128646.7994546991	128712.64203812725	128883.62831123867	129109.06903064142
Houston, TX	124446.06013047668	124469.24061036404	124382.35154965024	124434.28868529193	124482.1990285256

Single time for multiple subjects

Some Data Formats

Panel/Longitudinal Data

Each individual or entity is observed at multiple points in time.

Temporal order with consistent intervals, a month in this case

Region	1/31/00	2/29/00	3/31/00	4/30/00	5/31/00	Popu2024
New York, NY	220834.76256321764	221773.1857048372	222720.30033586253	224639.52106624088	226626.94059518576	19,940,274
Los Angeles, CA	222015.51137538892	222841.691	223942.1538408446	226131.70568265315	228526.38896950847	12,927,614
Chicago, IL	156057.927	156202.39142660523	156477.5260008363	157161.99244626865	157985.32325153003	9,408,576
Dallas, TX	128589.52523097747	128646.7994546991	128712.64203812725	128883.62831123867	129109.06903064142	8,344,032
Houston, TX	124446.06013047668	124469.24061036404	124382.35154965024	124434.28868529193	124482.1990285256	7,796,182

Multiple subjects are included.

More attributes can be added to the panel data, such as population, as **cross-sectional attributes.**

Panel data is the combination of time series and cross-sectional data.

Some Data Formats

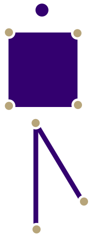
Spatial Data

Data about physical locations and shapes of objects, including both their geographical position and attributes. Broadly speaking, there are two types of spatial data:

- **Vector data**
  - Vectors are composed of discrete geometric locations.
  - Data format: Shapefile, GeoJSON, etc.

- **Raster data**
  - Pixelated data where each pixel is associated with a specific geographical location. The value of a pixel can be continuous (population) or categorical (land use types).
  - Data format: GeoTIFF, etc.

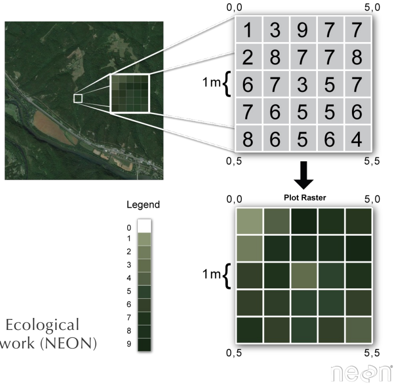
We will go back to spatial data later this quarter.



Point-shape
Individual location with (x,y), typically (longitude, latitude)

Polygon-shape
3 or more locations are connected and closed

Line-shape
2 or more locations are connected



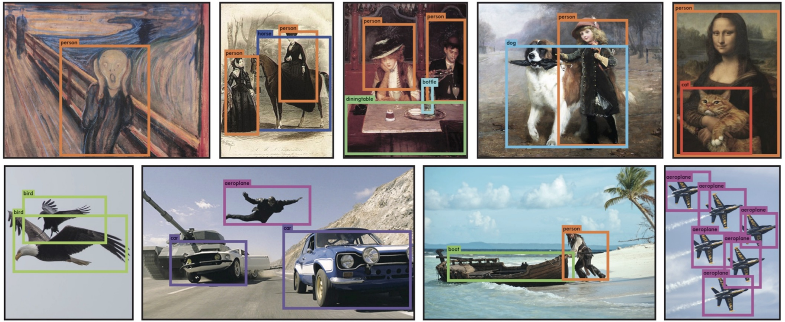
1	3	9	7	7
2	8	7	7	8
6	7	3	5	7
7	6	5	5	6
8	6	5	6	4

Source: National Ecological Observatory Network (NEON)

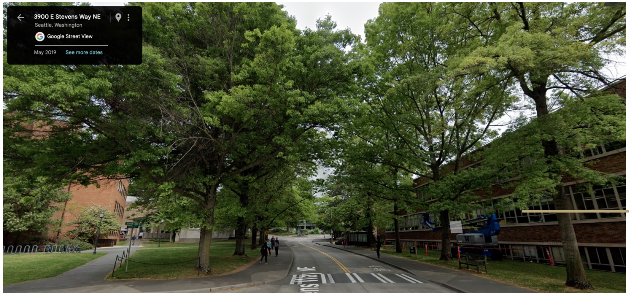
Some Data Formats

Image/Video/Street View

Source: ImageNet; https://www.image-net.org



YOLO. Source: Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection (2015) – developed by UW Huskies



Source: Google Street View of UW Campus

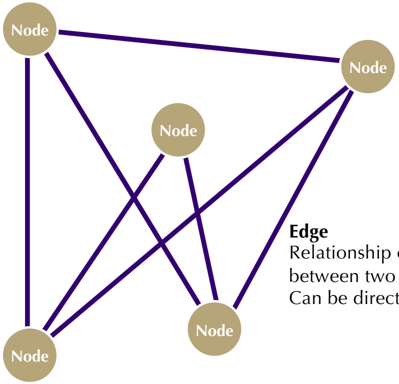
Object Detection + Feature Extraction
Identify the objects in images and extract features, such as the building color, building types, and quality

More serves as a new approach for us to extra real-world information

Some Data Formats

Network Data

Data about the entities within a network (called **nodes**) and the connections or interactions between them (called **edges**).

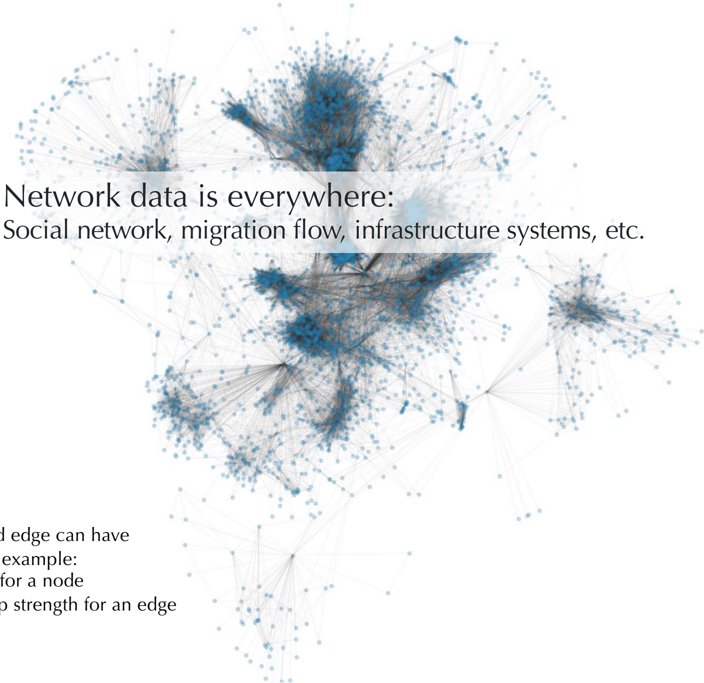


Node
People, location, posts, or any entity.

Edge
Relationship or interactions between two nodes.
Can be directed or undirected.

Attributes
Both node and edge can have attributes. For example:

- Population for a node
- Relationship strength for an edge



Network data is everywhere:
Social network, migration flow, infrastructure systems, etc.

A large network. Source: Visualize ML: https://github.com/Visualize-ML

Real Estate Data by Scopes

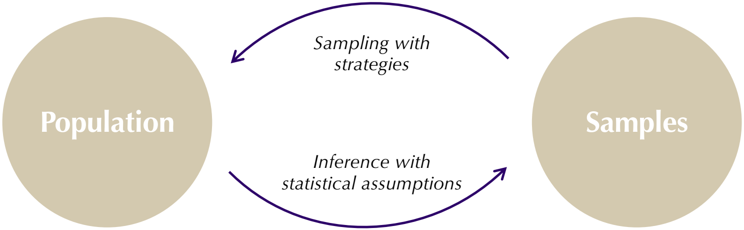
The Purpose of Data and Usage in Property Industry

Property-based	**Extra-locational**	
Specific/Core	***Intentional / Static Spatial***	***Collateral / Peripheral***
Property (Physical), transaction, and financial data	*Layers of related data beyond property-based data*	*By-product of other processes (especially about human behavior)*
Sales Transactions	Census Bureau Data	Internet Searches
Lease Transactions	Road Network Data	Transit Ridership Data
Mortgage Data	Aggregate (Core) Data	Live Traffic Data
Tax Assessment Values	Urban/Spatial (Core) Data	Point-of-Sale (POS) Data
Property Level Data (PLD)	Planning Forecasts	Geo-coded Tweets
REIT/Real Estate Stock Data	Spatial Economic Indicators	Pedestrian Traffic Counts

Source: Real Estate Analysis in the Information Age, by Kimberly Winson-Geideman, Andy Krause, Clifford A. Lipscomb, Nick Evangelopoulos (2018).

We will work with many types of datasets during this quarter.

Population and Sample



- **Population**
  - A population is the entire group of people or things about which we want information.
  - *Example*: All properties in Seattle.

- **Sample**
  - It is more practical to gain information about the whole population by only examining a part.
  - A smaller subset of a larger set of data to draw inferences about the large set.
  - *Example*: 1,000 selected properties in Seattle.

- It is important to first identify whether your data is “population” or “sample”.
  - If it is the population, no **inference** is needed. Just descriptive analysis.
  - If it is the sample, we need to consider sample errors and uncertainty.
    - The sampling strategies (review RE 506 if you need): random, stratified, cluster, etc.

Data and Source Quality

Make sure the source(s) tell a story

If you're not "telling a story" with your infographics (read: explaining a narrative or allowing a narrative to be explored), then you're doing it wrong. Essentially, that story will be derived from the sources that you decide to use.

- **Always use data sets from as unbiased a producer as possible.**
- Good sources include data collected or produced by government agencies, such as the statistics compiled by the U.S. Census Bureau or the Department of Labor.
- Other top-tier data sources can include industry white papers, surveys conducted by reputable research organizations, or findings published in academic publications.
    - Brookings, Urban Institute, Puget Sound Regional Council...
- Note that surveys conducted by polling agencies or think tanks, while usable, often have a political agenda, so always use discretion.
- About non-public access data - find out more about the data, such as how it was gathered, how old it is, and how many people were surveyed.

Data and Source Quality

Make Sure Your Sources are Relevant

- Use the most recently published version of the data you've decided to use if possible.
- As a rule of thumb, try not to use data that is more than a year old. Two years is acceptable in some cases, if that's the best you can get. Beyond this, use discretion. In all cases, be up front about the age of the data set you are using.
- If you are using multiple sources to craft a narrative, make sure they are complementary. Even if you only use two data sources, they can still create a lot of **variance**. Using two data sets that clash, such as data collected by think tanks on opposite sides of the political spectrum, makes crafting a narrative difficult.

Extra Credits

Dataset Sharing (up to 2%)

- From now to week 6, share high-quality datasets related to real estate and housing on this thread.
- The dataset cannot be repeated with the previous datasets shared by other students or the datasets already listed on the Resources page.
- Up to 2% extra credits: 5 unique datasets to receive 2%, 3 unique datasets to receive 1%.
- I will verify the datasets, and any unique datasets will be listed on the Resources page for others to refer to.
- It is a great time to collect data and start to think about the topics of your final project!

Census Data

U.S. Constitution

The United States Constitution mandates in Article I, Sections 2 and 9, that a complete enumeration of the US population be taken every 10 years.

"The actual enumeration shall be made within three years after the first meeting of the Congress of the United States, and within every subsequent term of ten years, in such manner as they shall by law direct."

AKA the **decennial census** (https://www.census.gov/programs-surveys/decennial-census.html)

However, 10 years is a long time to wait for new information…

Census Data

American Community Survey (ACS)

The American Community Survey (ACS) is a survey of 3.5 million US households annually (~3% of the US population).

Far more detailed than the decennial census and more timely.

"Through the ACS, we know more about jobs and occupations, educational attainment, veterans, whether people own or rent their homes, and other topics. Public officials, planners, and entrepreneurs use this information to assess the past and plan the future. When you respond to the ACS, you are doing your part to help your community plan for hospitals and schools, support school lunch programs, improve emergency services, build bridges, and inform businesses looking to add jobs and expand to new markets, and more."

-- U.S. Census Bureau

https://www.census.gov/programs-surveys/acs/about.html

Census Data

American Community Survey (ACS)

- **1-year estimates - “acs1”**
  - areas of population 65,000 and greater
- **5-year estimates - “acs5”**
  - Pooling of 1-year estimates into a 5-year moving average, which allows access to more granular data, such as population subgroups and geographies
- **These are estimates, so note the margin of error!**
  - Be cautious when working with very small subgroups (ex., American Indian or Alaska Native) or geographies with low populations. The estimates may lose much of their meaning.

Geography	Geographic Area Name	Estimate!!Total:	Margin of Error!!Total:
1400000US06001400100	Census Tract 4001, Alameda County, California	3035	402
1400000US06001400200	Census Tract 4002, Alameda County, California	1983	209

GEOID

90% confidence interval

What are Confidence Intervals?
https://seeing-theory.brown.edu/frequentist-inference/index.html#section2

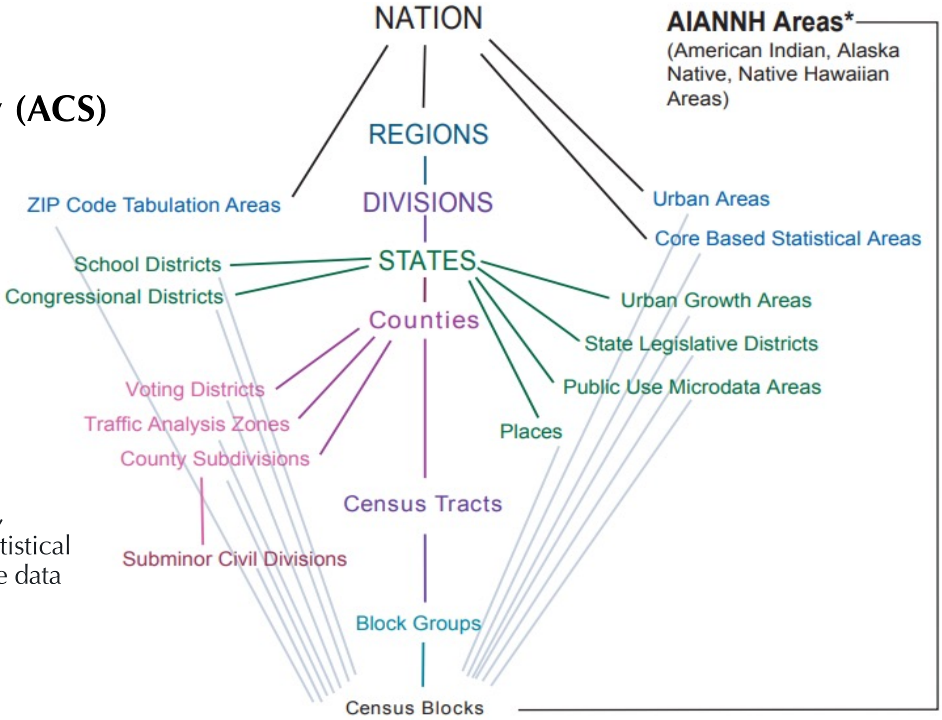
Census Data

American Community Survey (ACS)

Levels of geographic aggregation (nested hierarchy)

- Nation
- State
- County
- Census tract
- Census block group
- Census block

Sometimes, we may also use school districts, municipal boundaries, zip code areas, or statistical areas. Need to be careful when you combine data with different **geographic boundaries.**



Census Data

GEOIDs

Notes: GEOIDs may be slightly different across years. Try to stick with one version of GEOIDs and their boundaries.

53 033 005303 2 034

- The first two digits, 53, represent the FIPS code, also known as the **state**.
- Digits 3 through 5, 033, are the **county**.
- The next six digits, 005303, represent the block's **Census tract**.
- The twelfth digit, 2, represents the parent **block group** of the Census block.
- The last three digits, 034, represent the individual **Census block.**

- Seattle Census Block Map
- Seattle Census Block Group Map
- Seattle Census Tract Map

Census Data

A Note on Race/Ethnicity in the Census

The census codes race differently from how we typically think about it.

- **Racial groups include:**
  - White
  - Black or African American
  - American Indian or Alaska Native
  - Asian
  - Native Hawaiian or Other Pacific Islander

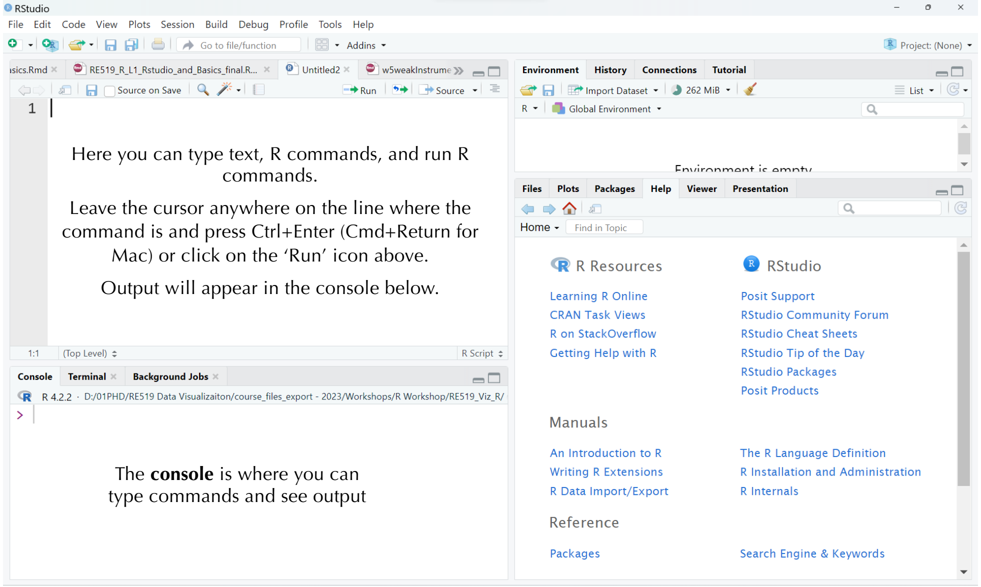
- **Ethnicity includes:**
  - Hispanic or Latino

People who identify their origin as Hispanic, Latino, or Spanish may be of any race.

Lab Session 1-B

Introduction to R & RStudio

RStudio Interface



```
R 4.2.2 · D:/01PHD/RE519 Data Visualizaiton/course_files_export - 2023/Workshops/R Workshop/RE519_Viz_R/
>
```

Here you can type text, R commands, and run R commands.

Leave the cursor anywhere on the line where the command is and press Ctrl+Enter (Cmd+Return for Mac) or click on the 'Run' icon above.

Output will appear in the console below.

The **console** is where you can type commands and see output

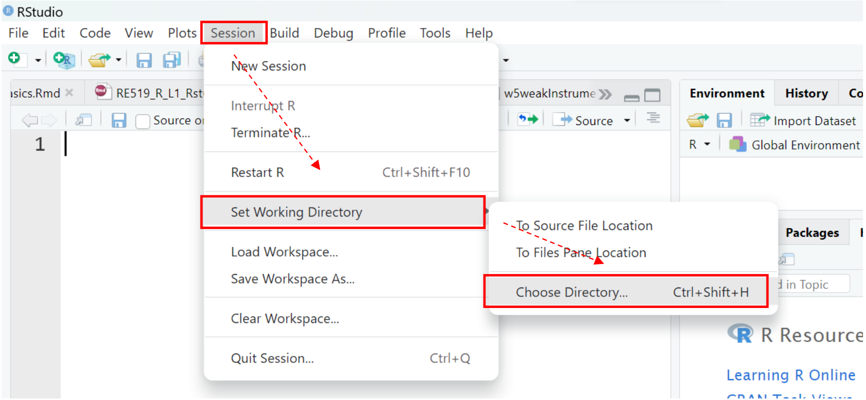
The **environment** tab shows all the active objects (see next slide). The **history** tab shows a list of commands used in the session

The **files** tab shows all the files and folders in your default workspace as if you were on a PC/Mac window. The **plots** tab will show all your graphs. The **packages** tab list the packages available, can install packages.

Introduction to R & RStudio

Set working directory

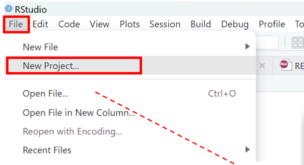
- The working directory is the folder where your files are saved and where RStudio will export files. You can check the current working directory by typing the following:
  - `getwd()`
- You can change the working directory by typing (if using Windows):
  - `setwd("C:/myfolder/data")`
- Or you can use the menu:

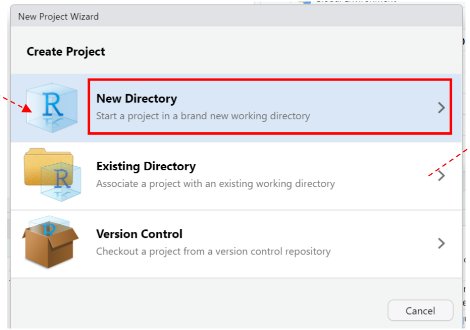


Introduction to R & RStudio

Create A New Project

An R project enables your work to be bundled in a **portable, self-contained folder**. Within the project, all the relevant scripts, data files, figures/outputs, and history are stored in sub-folders, and importantly, the *working directory* is the project's root folder.



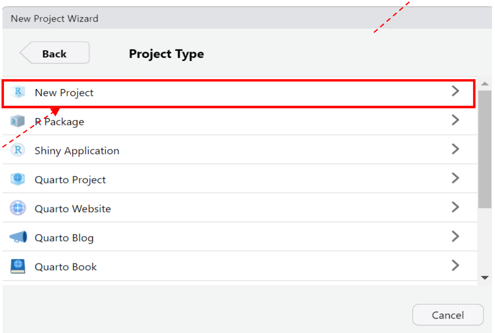

New Project Wizard

Create Project

New Directory
Start a project in a brand new working directory

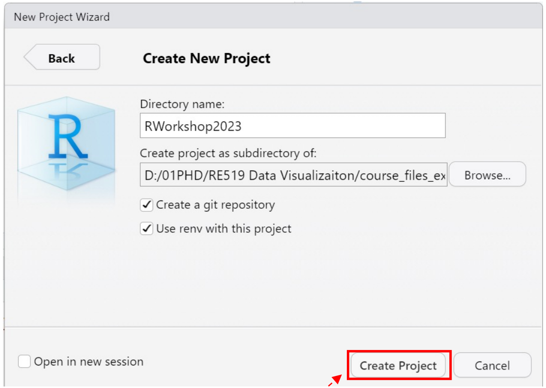
Existing Directory
Associate a project with an existing working directory

Version Control
Checkout a project from a version control repository


New Project Wizard

Project Type

- New Project
- R Package
- Shiny Application
- Quarto Project
- Quarto Website
- Quarto Blog
- Quarto Book


New Project Wizard

Create New Project

Directory name: RWorkshop2023

Create project as subdirectory of: D:/01PHD/RE519 Data Visualizaiton/course_files_ex

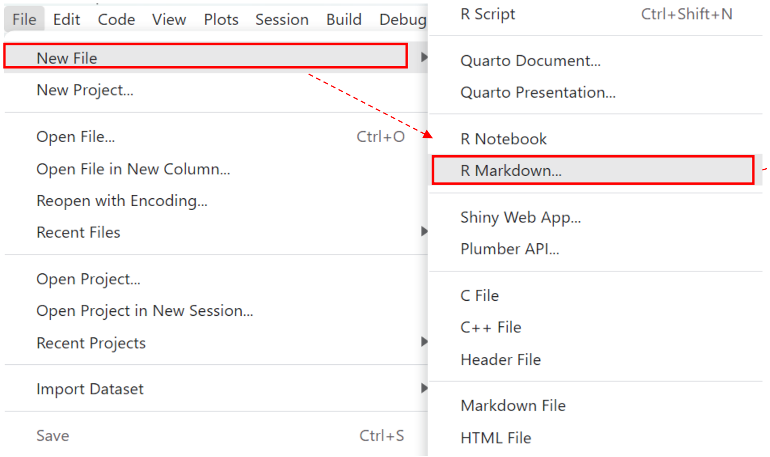
☑ Create a git repository

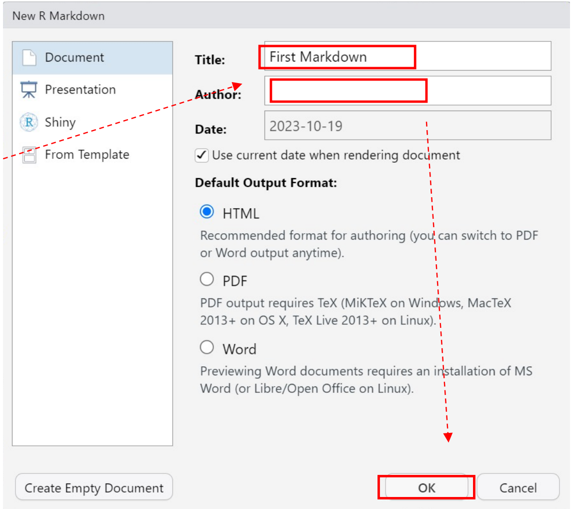
☑ Use renv with this project

□ Open in new session

Introduction to R & RStudio

Create A New R Markdown File





Title: First Markdown

Author:

Date: 2023-10-19

☑ Use current date when rendering document

Default Output Format:

● HTML

Recommended format for authoring (you can switch to PDF or Word output anytime).

○ PDF

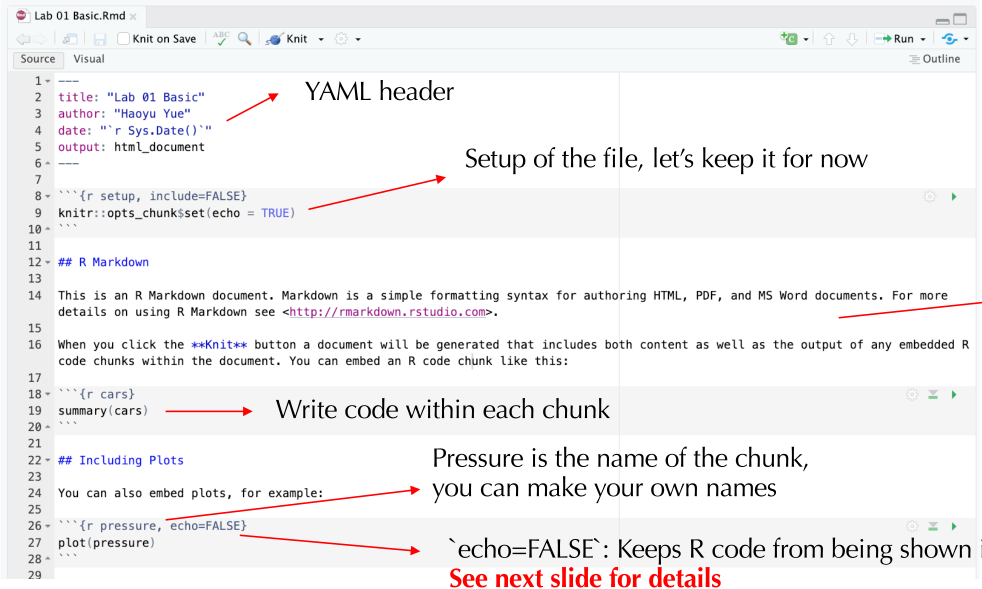
PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

○ Word

Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).

Introduction to R & RStudio

R Markdown Interface



~~~
---
title: "Lab 01 Basic"
author: "Haoyu Yue"
date: "`r Sys.Date()`"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

Including Plots

You can also embed plots, for example:

```{r pressure, echo=FALSE}
plot(pressure)
```

```

YAML header

Setup of the file, let's keep it for now

Write texts using Markdown

Write code within each chunk

Pressure is the name of the chunk, you can make your own names

`echo=FALSE`: Keeps R code from being shown in the document*

**See next slide for details**
```

Introduction to R & RStudio

R Markdown Interface-Other Chunk Settings

- `**echo=FALSE**`: Keeps R code from being shown in the document*
- `**eval=FALSE**`: Shows R code in the document without running it*
- `**include=FALSE**`: Hides all output but still runs code (good for `setup` chunks where you load packages!)*
- `**results='hide'**`: Hides R's (non-plot) output from the document*
- `**cache=TRUE**`: Saves results of running that chunk so if it takes a while, you won't have to re-run it each time you re-knit the document

Reminders

- Let me know if you’d like to use Python for this course.

Thank you!

Haoyu Yue / yohaoyu@washington.edu
Ph.D. Student, Interdisciplinary Urban Design and Planning
University of Washington

RE 519 Real Estate Data Analysis and Visualization
Course Website: www.yuehaoyu.com/data-analytics-visualization/
Autumn 2025

The course was developed based on previous instructors: Christian Phillips, Siman Ning, Feiyang Sun
Cover page credits: Visax