Stata Training Camp

Lecture 1 Software Basics and Data Processing

Juncheng Jiang

jc_jiang@stu.zuel.edu.cn

School of Public Finance and Taxation, Zhongnan University of Economics & Law
Associate Editor, Student Academic Research Center

2024.11.17

Pre-class Survey

STATA: STATistics Aided by Technology, Applications, and Software.

Camp Design

In today's economic academic research, theoretical models are inseparable from empirical testing. With the advent of the big data era, processing economic data is an essential skill for economics students.
In previous coursework, more emphasis was placed on introducing and explaining econometric methods, but the practical aspect of using these methods to handle data was relatively lacking.
Based on feedback and suggestions from faculty and students in our school, it seems necessary to organize some seminars and training to provide an introductory overview of basic data processing and Stata operations. Thus, the Stata Basic Training Camp came into being.

Camp Design

This training camp is positioned for absolute Stata beginners. Therefore, it will not involve overly complex operational techniques.

We will start from the most basic installation and understanding of Stata, gradually explaining how to clean, organize, visualize data, perform regressions, and output results. Finally, using publicly available code from top journals as examples, we will demonstrate the practical process of replicating a paper.

Trainee Composition

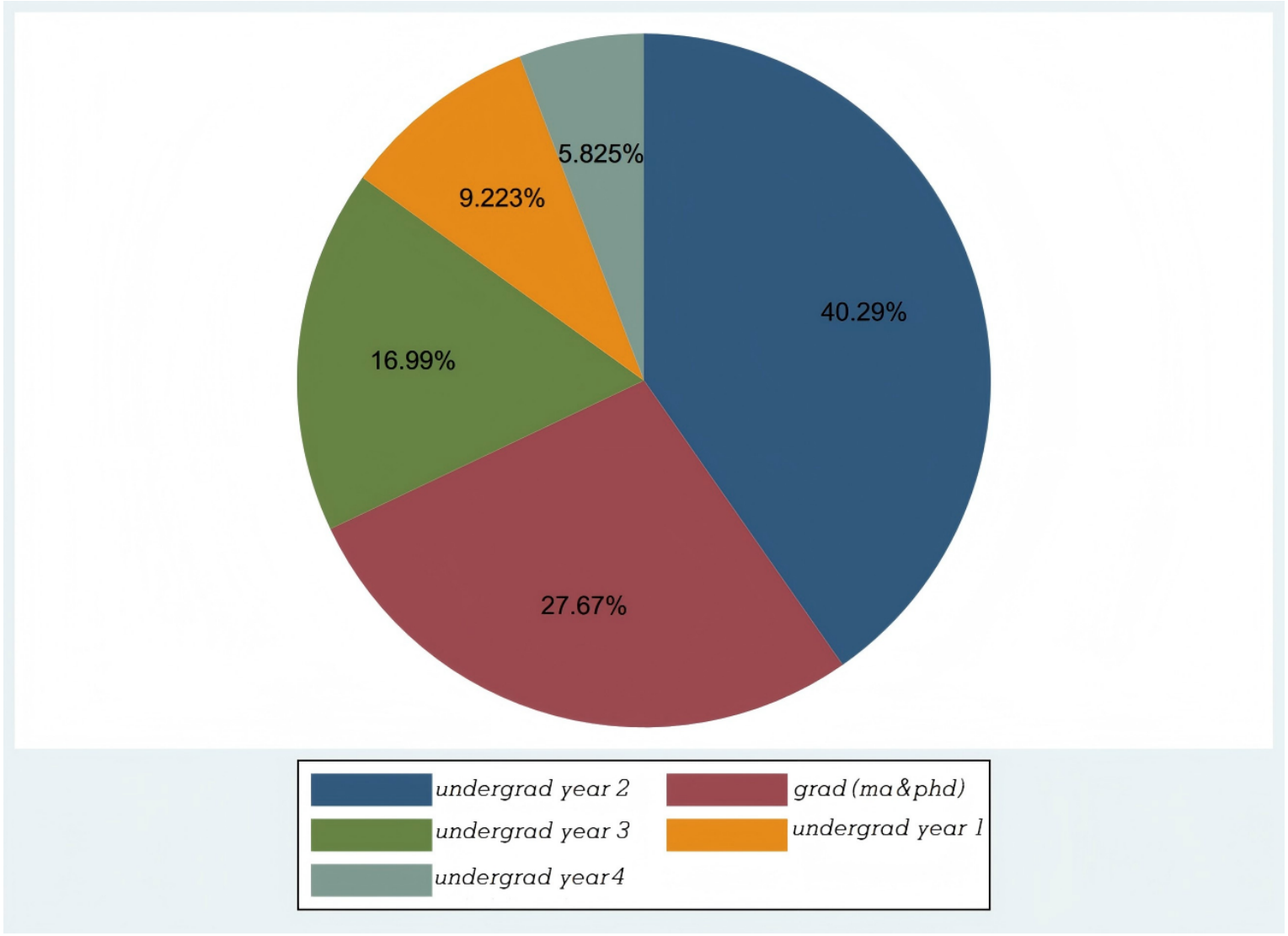


Figure: Trainee Composition

Stata Software Version Introduction

Stata typically has three common versions for each release: MP, SE, BE

- Stata/MP: The fastest Stata version (for quad-core, dual-core, and multi-core/multiprocessor computers), can analyze the largest datasets
- Stata/SE: Standard Edition; for larger datasets
- Stata/BE: Basic Edition; for medium-sized datasets

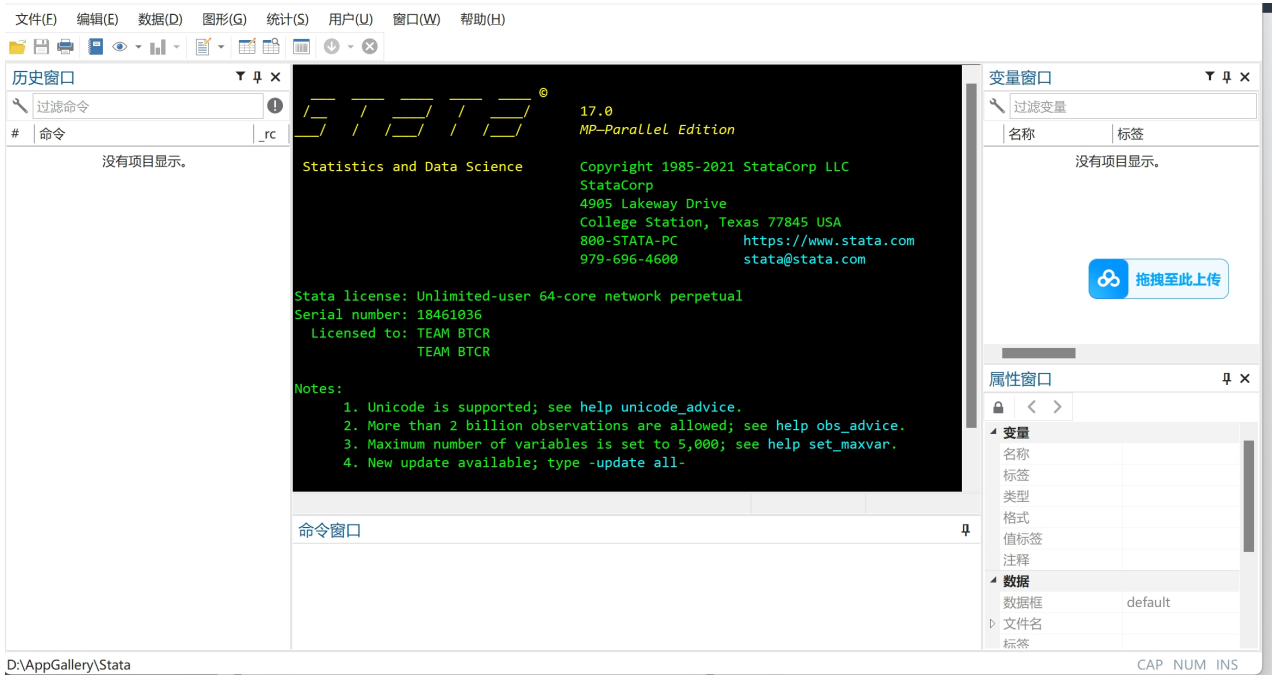
Simply put, the functional comparison of the three versions is MP $>$ SE $>$ BE. The version we are sharing in this session is the MP version, the highest-spec Stata version.

Currently, Stata has been updated to version 18. Installation steps refer to the installation tutorial in the camp preparation guide.

The first version of Stata was released in January 1985.

Stata Windows and Menu Bar

- Command Window
- Results Window
- History Window (Review)
- Variables Window
- Properties Window



File Formats in Stata

- Data files recognizable by Stata have the suffix .dta; You can initially store collected data as .xlsx or .csv, then import it into Stata and save it. The import methods will be introduced shortly.

File Formats in Stata

- Data files recognizable by Stata have the suffix .dta; You can initially store collected data as .xlsx or .csv, then import it into Stata and save it. The import methods will be introduced shortly.
- Similar to other programming languages, you can write all commands into a do-file, and execute the file instead of choosing interactive, step-by-step command entry. The storage suffix for do-files is .do.

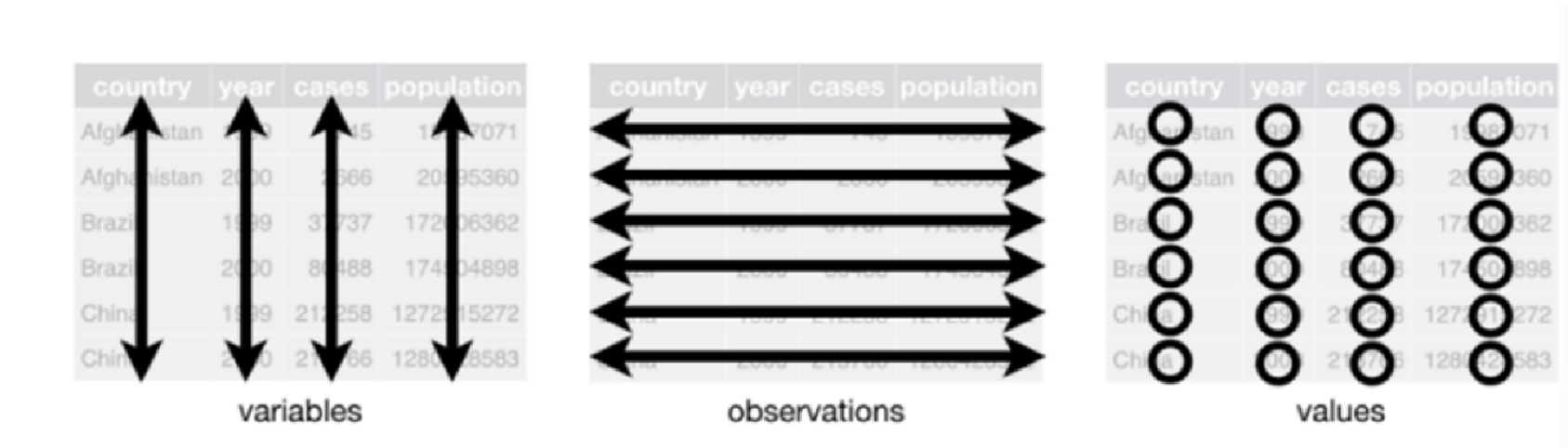
File Formats in Stata

- Data files recognizable by Stata have the suffix .dta; You can initially store collected data as .xlsx or .csv, then import it into Stata and save it. The import methods will be introduced shortly.
- Similar to other programming languages, you can write all commands into a do-file, and execute the file instead of choosing interactive, step-by-step command entry. The storage suffix for do-files is .do.
- Furthermore, when you run a do-file, you can choose to record your process. Stata has a specific log command to help you record it. Its suffix is .smcl (can only be opened with Stata), but you can translate it to .log or .pdf format using the ‘translate‘ command.

Data Organization FormTidy Data

We generally call data where each column is a variable, each row is an observation tidy data;
There are three interrelated rules that make a dataset tidy (Wickham, H. 2014):

1. Each variable forms a column; Each column is a variable.
2. Each observation forms a row; Each row is an observation.
3. Each value forms a cell; Each cell is a single value.



Other Organizational Forms

```
table1
#> # A tibble: 6 × 4
#>   country      year  cases population
#>   <chr>       <dbl>  <dbl>      <dbl>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3 Brazil       1999  37737  172006362
#> 4 Brazil       2000  80488  174504898
#> 5 China        1999 212258 1272915272
#> 6 China        2000 213766 1280428583
```

```
table3
#> # A tibble: 6 × 3
#>   country      year rate
#>   <chr>       <dbl> <chr>
#> 1 Afghanistan  1999 745/19987071
#> 2 Afghanistan  2000 2666/20595360
#> 3 Brazil       1999 37737/172006362
#> 4 Brazil       2000 80488/174504898
#> 5 China        1999 212258/1272915272
#> 6 China        2000 213766/1280428583
```

```
table2
#> # A tibble: 12 × 4
#>   country      year type           count
#>   <chr>       <dbl> <chr>          <dbl>
#> 1 Afghanistan  1999 cases            745
#> 2 Afghanistan  1999 population   19987071
#> 3 Afghanistan  2000 cases           2666
#> 4 Afghanistan  2000 population   20595360
#> 5 Brazil       1999 cases           37737
#> 6 Brazil       1999 population  172006362
#> # ℹ 6 more rows
# Spread across two tibbles
table4a  # cases
#> # A tibble: 3 × 3
#>   country     `1999` `2000`
#>   <chr>        <dbl>  <dbl>
#> 1 Afghanistan    745   2666
#> 2 Brazil       37737  80488
#> 3 China       212258 213766
table4b  # population
#> # A tibble: 3 × 3
#>   country         `1999`      `2000`
#>   <chr>            <dbl>       <dbl>
#> 1 Afghanistan   19987071    20595360
#> 2 Brazil       172006362   174504898
#> 3 China       1272915272  1280428583
```

Common Problems

Five common problems with messy datasets (remedial measures refer to the original text):

- Column headers are values, not variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple types of observational units are stored in the same table.
- A single observational unit is stored in multiple tables.

Structures of Economic Data

Generally speaking, economic data can be divided into three/four categories:

- Cross-section data
  - Consists of a sample of individuals, households, firms, cities, countries, or various other units taken at a given point in time.
  - Minor time differences are ignored, e.g., surveying different people in different months within a year.
- Time series data
  - Consists of observations on one or more variables over time.
  - It is rarely assumed that economic data are independent across time.
  - Need to pay attention to the data frequency at the time of collection (Daily/Weekly/Monthly/Quarterly/Annually?).
- Pooled cross section
  - Has both cross-sectional and time-series characteristics. E.g., A group of people was surveyed in 2022, and a new group of people was surveyed in 2024.
- Panel data
  - Consists of a time series for each cross-sectional unit in the dataset.

Identifying Structure Type

	电影名	电影id	年代	海报	时光评分	评论人数	导演	编剧	类型	国家地区
4	丛林大冒险	274737	2019	http://img5.mtime.c…	0.0	1人评分	咪咪·梅纳德	Martin BehnkeMarina…	动画/冒险/喜剧	比利时
5	艾尔米塔什博物馆：艺术的力…	274325	2019	http://img5.mtime.c…	0.0	0人评分			纪录片	意大利
6	京门烽火	268322	2020	http://movie.mtime.…	0.0	0人评分	刘子铭		2D	中国
7	野蛮人入侵	270412	2021	http://img5.mtime.c…	0.0	7人评分	陈翠梅	陈翠梅	剧情	中国香港
8	拉斐尔：年轻的画圣	274660	2021	http://img5.mtime.c…	0.0	0人评分			纪录片	意大利
9	湄公河巨兽	274736	2022	http://img5.mtime.c…	0.0	0人评分	李·通坎	李·通坎	惊悚/科幻/恐怖	澳大利亚
10	热烈	272254	2022	http://img5.mtime.c…	7.7	201人评分	大鹏	大鹏苏彪	剧情/喜剧	中国
11	女王的肖像	274659	2022	http://img5.mtime.c…	0.0	0人评分			纪录片	意大利
12	消失的她	273294	2022	http://img5.mtime.c…	6.8	327人评分	崔睿	陈思诚顾舒怡殷一雄	中国	中国
13	火山挚恋	272736	2022	http://img5.mtime.c…	8.4	73人评分	Sara Dosa	Shane BorisErin Cas…	中国	加拿大
14	爱犬奇缘	270236	2022	http://img5.mtime.c…	0.0	0人评分	陈国辉			中国
15	欢乐丛林之外星人大冒险	274180	2023	http://img5.mtime.c…	0.0	1人评分	邢海顺		中国	中国
16	最遗憾是错过你	274094	2023	http://img5.mtime.c…	0.0	0人评分			爱情/剧情	中国台湾
17	学爸	270250	2023	http://img5.mtime.c…	0.0	34人评分	苏亮	苏亮	剧情/喜剧	中国
18	忠爱无言2	262838	2023	http://img5.mtime.c…	0.0	0人评分	谈宜之	谈宜之	剧情	中国
19	潜艇总动员：环游地球80天	274622	2023	http://img5.mtime.c…	0.0	1人评分			中国	中国
20	北京2022	273815	2023	http://img5.mtime.c…	0.0	4人评分	陆川		纪录片/剧情/运动	中国
21	雪域少年	272275	2023	http://img5.mtime.c…	0.0	0人评分	路奇	王运生路武楠	中国	中国
22	巨齿鲨2：深渊	270273	2023	http://img5.mtime.c…	6.5	209人评分	本·维特利	乔·霍贝尔埃里希·霍…	动作/科幻/冒险	中国
23	长安三万里	272782	2023	http://img5.mtime.c…	8.3	423人评分	谢君伟	红泥小火炉	动画/历史	中国
24	力量密码	274483	2023	http://img5.mtime.c…	0.0	2人评分			剧情/历史/传记	中国
25	看不见的他们	274797	2023	http://img5.mtime.c…	0.0	0人评分	李飞		家庭/惊悚/恐怖	中国
26	川流不“熄”	270506	2023	http://img5.mtime.c…	0.0	7人评分	冯钶予		剧情	中国
27	暗杀风暴	255768	2023	http://img5.mtime.c…	0.0	37人评分	邱礼涛	沈锡然	悬疑/犯罪	中国
28	GT赛车：极速狂飙	273470	2023	http://img5.mtime.c…	0.0	6人评分	尼尔·布洛姆坎普	杰森·霍尔扎克·贝林…	动作/剧情/冒险	美国
29	我经过风暴	270490	2023	http://img5.mtime.c…	0.0	34人评分	秦海燕	秦海燕	家庭/剧情	中国
30	百变马丁：怪兽大作战	274470	2023	http://img5.mtime.c…	0.0	3人评分	张天晓		动画/冒险	中国

Figure: Pooled Cross-section Data

Most Commonly Used Type – Panel Data

Balanced panel data means that each observation unit has complete observations over a specific period. That is, for every individual, data is available for every time point.

Data Editor (Browse) — wc14.dta

var620[11]

	pid	year	age	male	drink_lastyr
1	010104102001	2011	57	0	0.None
2	010104102001	2013	59	0	0.None
3	010104102001	2015	61	0	0.None
4	010104102001	2018	64	0	0.None
5	010104102002	2011	60	1	1.Yes
6	010104102002	2013	62	1	1.Yes
7	010104102002	2015	64	1	1.Yes
8	010104102002	2018	67	1	1.Yes
9	010104104001	2011	61	1	0.None
10	010104104001	2013	63	1	0.None
11	010104104001	2015	64	1	0.None
12	010104104001	2018	67	1	0.None
13	010104104002	2011	53	0	.m:missing
14	010104104002	2013	55	0	0.None
15	010104104002	2015	57	0	0.None
16	010104104002	2018	60	0	1.Yes
17	010104105001	2011	66	0	0.None
18	010104105001	2013	69	0	0.None
19	010104105001	2015	70	0	0.None
20	010104105001	2018	73	0	0.None
21	010104105002	2011	68	1	1.Yes
22	010104105002	2013	70	1	1.Yes
23	010104105002	2015	72	1	1.Yes
24	010104105002	2018	75	1	0.None
25	010104107001	2011	79	1	0.None
26	010104107001	2013	81	1	0.None
27	010104107001	2015	83	1	0.None

Vars: 5 of 617 Order: Dataset Obs: 36,080

Figure: Balanced Panel Data

Data Editor (Browse) — wc14.dta

var626[9]

	pid	year	age	male	drink_lastyr
1	010104101001	2015	50	0	0.None
2	010104101001	2018	53	0	0.None
3	010104101002	2013	50	1	1.Yes
4	010104101002	2015	52	1	1.Yes
5	010104101002	2018	55	1	0.None
6	010104102001	2011	57	0	0.None
7	010104102001	2013	59	0	0.None
8	010104102001	2015	61	0	0.None
9	010104102001	2018	64	0	0.None
10	010104102002	2011	60	1	1.Yes
11	010104102002	2013	62	1	1.Yes
12	010104102002	2015	64	1	1.Yes
13	010104102002	2018	67	1	1.Yes
14	010104103001	2013	50	1	1.Yes
15	010104103001	2015	52	1	1.Yes
16	010104103001	2018	55	1	1.Yes
17	010104103002	2015	50	0	0.None
18	010104103002	2018	53	0	0.None
19	010104104001	2011	61	1	0.None
20	010104104001	2013	63	1	0.None
21	010104104001	2015	64	1	0.None
22	010104104001	2018	67	1	0.None
23	010104104002	2011	53	0	.m:missing
24	010104104002	2013	55	0	0.None
25	010104104002	2015	57	0	0.None
26	010104104002	2018	60	0	1.Yes
27	010104105001	2011	66	0	0.None

Vars: 5 of 617 Order: Dataset Obs: 63,135

Figure: Unbalanced Panel Data

Converting Unbalanced Data

In econometric analysis, balanced panel data is often easier to handle than unbalanced panel data. However, this does not mean you must use balanced panels to run regressions.

If you want to keep only individuals that exist in all given years within an unbalanced panel dataset, thereby converting it to a balanced panel, the specific command used is: ‘xtbalance‘

To download and install the external command: ‘ssc install xtbalance‘
Syntax format: ‘xtbalance, range(numlist) [miss(varlist)]‘
Note: You need to declare the dataset as panel data before using it.

Specific conversion operations can be referred to in Lianxh’s article, which will not be elaborated here:
Stata Data Processing: xtbalance - Converting Unbalanced Panels
https://blog.csdn.net/arlionn/article/details/118815391

Data Types

Three types of data (distinguished by color in Stata)

- Red: String (can be sorted alphabetically) type: string(str)
  - One type: pure numbers (e.g., stock code stkcd)
  - Another type: pure text (e.g., province name, industry name)
- Black: Numeric (can be used in regression analysis)
- Blue: Numeric - text correspondence table (categorical/dummy variables); essentially also numbers

Common Data Processing Steps & Stata Commands

- Check working directory (‘pwd‘); Change working directory (‘cd‘)
- Import Excel file (‘import excel‘), CSV file (‘import delimited‘), dta file (‘use‘)
- Add variable or value labels (‘label‘)
- Generate new variables (‘gen‘) or statistical variables (‘egen‘)
- Sort observations based on variable values (‘sort‘, ‘gsort‘)
- Change the order of variables (‘order‘)
- Reshape datasets between wide and long formats (‘reshape‘)
- Merge datasets (‘merge‘)
- Delete (‘drop‘) or keep (‘keep‘) observations or variables
- Export Excel, CSV files (‘export‘) or dta files (‘save‘)

STATA Environment Command Examples

```
use xxx.dta, clear
keep a b c
gen d
gen dummy=1 if year>=2022 // Generate dummy variable

.......

outreg2 using statistic, word replace sum(log)  /*
*/ sortvar(price income interest) /*
*/ keep(price income interest)
* Basic regression
```

Basic Command List

```
1  use "file.dta", clear // Read .dta data
2  sort air_time dest // Sort columns (observations)
3  order month day // Order variables (columns)
4  rename (old) (new) // Rename variable
5  keep in 1/200 // Keep first 200 observations
6  keep if day>5 & day<10 // Keep observations meeting c
7  keep if in range(day,5,10) // Alternative condition
8  keep month day carrier // Keep subset of variables
9  gen dist_sq=distance^2 // Create new variable
10 replace a=dep_delay+arr_delay // Modify existing varia
11 drop varlist [if] [in]  // Drop variables/observations
12 .......
```

What You Will Learn in This Lecture:

```
use xxx.dta, clear
log // logging commands
global local // global and local macros
rename // rename variables
label // label variables/values
encode // encode string to numeric with label
destring // convert string to numeric
keep // keep variables/observations
gen // generate new variable
replace // replace values of a variable
sort // sort data
.......
```

Launch Stata

Let's DO it!