1. **Understanding variable types (numeric, string, etc.)**

* Numeric variables

Numeric variables store numbers and can be used for mathematical operations.

**Types:**

* **Integer:** Whole numbers without decimal points.
* **Float:** Numbers with decimal points. These are approximate representations of real numbers.

**\* Clear any existing data in memory**

**clear**

**\* Set seed for reproducibility**

**set seed 12345**

**\* Create a dataset with 100 observations**

**set obs 100**

**\* Generate an integer variable (age)**

**gen int age = round(runiform() \* 52 + 18)**

**\* Generate a float variable (income)**

**gen float income = runiform() \* 80000 + 20000**

* String Variables

String variables store text and are used for non-numeric data.

**\* Generate a short string variable (name)**

**gen str18 name = ""**

**\* Assign values to the string variable**

**replace name = "John Doe" in 1**

**replace name = "Jane Smith" in 2**

* Factor/Categorical variables

Categorical variables take on a limited number of distinct values, representing different categories.

**\* Generate a categorical variable (education)**

**\* if not explicitly defined, by default numerical variables take float type.**

**gen education = 1**

**replace education = 2 in 21/40**

**replace education = 3 in 41/60**

**replace education = 4 in 61/80**

**\* Label the categorical variable**

**label define edu\_labels 1 "High School" 2 "Bachelor's" 3 "Master's" 4 "PhD"**

**label list edu\_labels**

**label values education edu\_labels**

***Exercise:***

Generate a factor/categorical variable named age\_group based on the following rule.

age < 20 🡪 Teen, 20 <= age < 65 🡪 Adult, age >=65 🡪 Senior.

**\* Generate a categorical Age Group variable**

**gen age\_group = .**

**replace age\_group = 1 if age < 20**

**replace age\_group = 2 if age >= 20 & age < 65**

**replace age\_group = 3 if age >= 65**

**\* Label the Age Group variable**

**label define agegrp\_labels 1 "Teen" 2 "Adult" 3 "Senior"**

**label values age\_group agegrp\_labels**

* Date and Time variable

Date and time variables store dates, times, and date-time combinations. They require special formats to perform calculations and manipulations.

**\* Generate a date variable**

**gen date = mdy(12, 25, 2024)**

**\* Format the date variable**

**format date %td**

**\* Format the date in YYYY-MM-DD format**

**format date %tdCCYY-NN-DD**

**\* Format the date in MM/DD/YYYY format**

**format date %tdNN/DD/CCYY**

1. **Converting variable types using destring and tostring**

Converting a string variable to numeric

\* Clear any existing data in memory

**clear**

**set obs 3**

\* Create a string variable with numeric values

**gen** str\_var = "123"

**replace** str\_var = "456" **in** 2

**replace** str\_var = "789" **in** 3

\* Convert the string variable to numeric

**destring** str\_var, **replace** //destring str\_var, gen(num\_var)

Converting a numeric variable to string

\* Clear any existing data in memory

**clear**

**set obs 3**

\* Create a numeric variable

**gen** num\_var = 123

**replace** num\_var = 456 **in** 2

**replace** num\_var = 789 **in** 3

\* Convert the numeric variable to string

tostring num\_var, **replace** //tostring num\_var, gen(str\_var)

1. **Encoding a string variable to a factor/categorical variable**

**\* Clear any existing data in memory**

**clear**

**\* set dataset size to 4 observations**

**set obs 40**

**\* Create a string variable with categorical values**

**gen education\_level = "PhD"**

**replace education\_level = "Bachelor's" in 11/20**

**replace education\_level = "High School" in 21/30**

**replace education\_level = "Master's" in 31/40**

**\* Define a label with a specific order**

**label define edu\_labels 1 "High School" 2 "Bachelor's" 3 "Master's" 4 "PhD"**

**\* Encode the string variable into a numeric variable using the defined label**

**encode education\_level, gen(education\_encoded) label(edu\_labels)**

**\* Encode the string variable without defined labels**

**encode education\_level, gen(education\_encoded1)**

1. **Re-ordering variables**

**\* Load the built-in dataset 'auto'**

**sysuse auto, clear**

**\* Original order of variables**

**describe**

**\* Reorder variables**

**order price mpg rep78 make headroom**

**\* Verify the new order of variables**

**describe**

**\* move a variable to the first position**

**order mpg**

**\* Verify the new order of variables**

**describe**

**\* move a variable to the first position**

**order mpg, last**

**\* Verify the new order of variables**

**describe**