Data Cleaning With R: Takeaways

Syntax

MANIPULATING COLUMNS USING THE `DPLYR` PACKAGE:

- Converting a single column to numeric:

```
data_frame <- data_frame %>%

mutate(`col name` = as.numeric(`col name`))
```

- Converting multiple columns to numeric with column names:

```
data_frame <- data_frame %>%

mutate_at(vars(`col name 1`: `col name 5`), as.numeric)
```

- Converting multiple columns to numeric with column indexes:

```
data_frame <- data_frame %>%

mutate_at(`beginning index`: `ending index`), as.numeric)
```

- Filtering a data frame:

```
data_frame <- data_frame %>%

filter(`col name` > condition)
```

- Grouping a data frame:

```
data_frame <- data_frame %>%

group_by(`col name 1`, `col name 2`)
```

- Summing up columns:

```
data_frame <- data_frame %>%

mutate(`col name` = `col name 1` + `col name 2`)
```

- Padding character strings:

```
data_frame <- data_frame %>%

str_pad(`col name`, width = 6, side = 'left', pad = "0")
```

- Selecting variables from a data frame:

  ```
  graduation <- graduation %>%
  filter(Cohort == "2006" & Demographic == "Total Cohort") %>%
  select(`col name 1`, `col name 2`, `col name 3`)
  ```

- Removing a column from a data frame:

  ```
  graduation <- graduation %>%
  select(-the_name_of_column_to_remove) #note the presence of the symbol -
  ```

- Renaming a column in a data frame:

  ```
  data_frame %>%
  rename(new_column_name = old_column_name)
  ```

- Identifying duplicated values:

  ```
  duplicated(data_frame)
  ```

- Identifying duplicatated values using purrr and dplyr:

  ```
  list %>%
  map(mutate, is_dup = duplicated(`col name 1`))
  ```

Concepts

- Much of the data you will encounter in the real world requires data cleaning. Data cleaning includes:
  - Removing data you don't need for analysis.
  - Removing duplicate data.
  - Dealing with missing data and outliers.
  - Creating new variables where necessary.
  - Combining separate datasets.
- Metadata refers to any available descriptions of the datasets.
- Tick marks (``) are necessary when referring to variable names with spaces within them.

Resources

- Preparing data analysis
- Duplicated function
- Six steps to data cleaning