R for Data Science

AITI Talk Summary

Talk by Dr Haziq Jamil

What is tidy data?

A dataset is **tidy** if it follows these three rules:

1. Each variable forms a column.
   e.g. height, weight, age
2. Each observation forms a row.
   e.g. data for one person, one transaction, one event
3. Each type of observational unit forms a table.
   e.g. people in one table, hospitals in another

How to install & load packages?

Only need to install ones

```
install.packages("<package name>")
```

But you need to load everytime!

```
library(tidyverse)  # data wrangling tool
library(tidytext)   # bigrams
library(tm)         # text mining
library(wordcloud)  # word clouds
library(gtsummary)  # pretty summary tables
library(bruneimap)  # for mapping
```

What are the data types?

Type	Subtype	Example
Logical	-	TRUE,FALSE
Numeric	Integer	1L, 314L
Numeric	Double	1.23, 3.141
Complex	-	1+2i, 3+4i
Character	-	"cat", "blue"
Character	Factor	"MOE", "MTIC"
Character	Ordered	"Disagree", "Agree"

What's in our data set?

```
dat <- read_csv("fake_survey.csv")
```

```
glimpse(dat)
Rows: 2,000
Columns: 13
$ id          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9
$ kampong     <chr> "Kg. Lorong Tiga Selatan"
$ mukim       <chr> "Mukim Seria", "Mukim Ber
$ district    <chr> "Belait", "Brunei-Muara",
$ gender      <chr> "Female", "Male", "Female
$ age         <dbl> 47, 38, 42, 47, 50, 33, 5
$ education   <chr> "O Level", "O Level", "O
$ q_fbspeed   <dbl> 54, 58, 711, 56, 187, 58,
$ q_fbqual    <chr> "Fair", "Very Good", "Poo
$ q_mbqual    <chr> "Poor", "Good", "Good", "
$ q_fbexpend <dbl> 782, 53, 744, 53, 635, 61
$ q_fbusage   <dbl> 620, 260, 750, 320, 290,
$ q_limiting <chr> "I'll almost never downlo
```

Some Important Functions

Data Wrangling

- `mutate()` – Add or modify columns
- `pivot_longer()` – Reshape wide to long format

Data Type Conversion

- `factor()` – Convert to categorical variable
- `ordered()` – Create an ordered factor
- `as.numeric()` – Convert to numeric type

Data Exploration

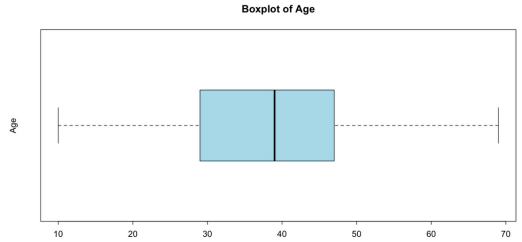
- `head()` – View first few rows
- `unique()` – Show unique values
- `table()` – Frequency count of values

Summary Statistics

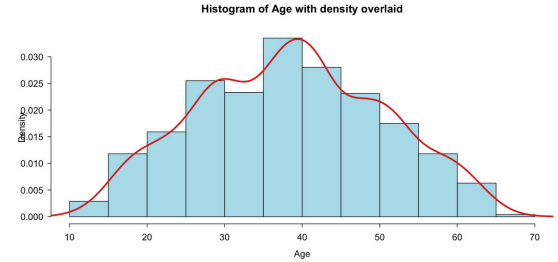
- `mean()` – Calculate average
- `sd()` – Standard deviation
- `summary()` – Summary of each column

Simple Visualization

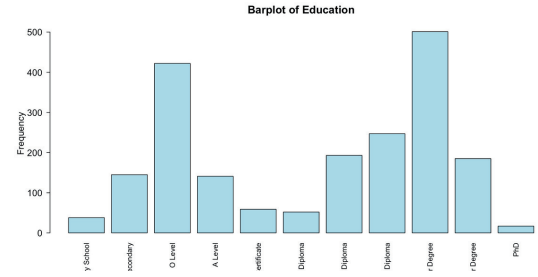
```
x <- dat$age
boxplot(x, horizontal = TRUE, main = "Boxplot of
        Age", ylab = "Age",col = "lightblue")
```




```
hist(x, main = "Histogram of Age with density
  overlaid", xlab = "Age",ylab = "Density",
  col = "lightblue", breaks = 10, prob = TRUE)
lines(density(x), lwd = 3, col = "red3")
```




```
barplot(table(dat$education), las = 2, cex.names
       = 0.8, main = "Barplot of Education",
       ylab = "Frequency", col = "lightblue")
```



Tables

```
x <- dat$gender
table(x)
```

```
-->  x
   Female   Male
     1010    990
```

```
tab1 <- table(dat$gender, dat$q_fbqual)
print(tab1)
```

```
       Very Poor Poor Fair Good Very Good Excellent
Male          26   68  204  333       271        88
Female        16   62  236  333       254       109
```