STA130H1S TUT0109
W3: Boxplots, Data Wrangling

Jan 25, 2019

Email: sonia.chhay@mail.utoronto.ca
Website: soniachhay.github.io

Announcements

- Be sure to answer all parts of the question, including the explanation if it asks for it
- You’re allowed to discuss assignment with other students but you cannot share your answers, anything you submit must be your own work
- If there are things about your assignment that you're confused about, you can go to OH or ask on Piazza
- Solutions are also posted on Quercus (after tutorial)

Overview

1st half of tutorial:

- Vocabulary for this week (1/2)
  - Brief overview of boxplots
- Vocabulary for this week (2/2)
  - Data wrangling
- Group discussion on Question (1d)
- Short written evaluation

2nd half of tutorial:

- Meet your mentors!

Vocabulary for this week (1/2):

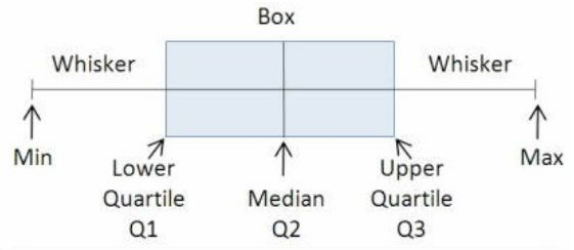
General terms:

- Boxplot
- Interquartile range
- Proportion
- Outlier

Boxplots Overview

When should boxplots be used?
What types of information does it show?

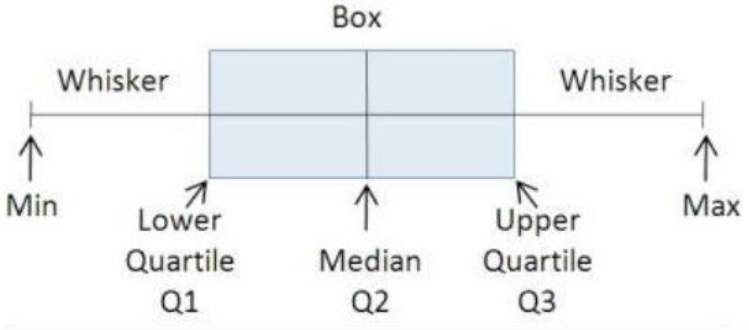
Boxplots



(1) When should boxplots be used?
- To summarize distribution of quantitative (numerical) variable
  - E.g. Babies' birth weight

(2) What does a boxplot do?
- Visualizes five statistics
  - Minimum, maximum, median, 1st quartile and 3rd quartile
- Plots unusual observations (outliers)

Boxplots

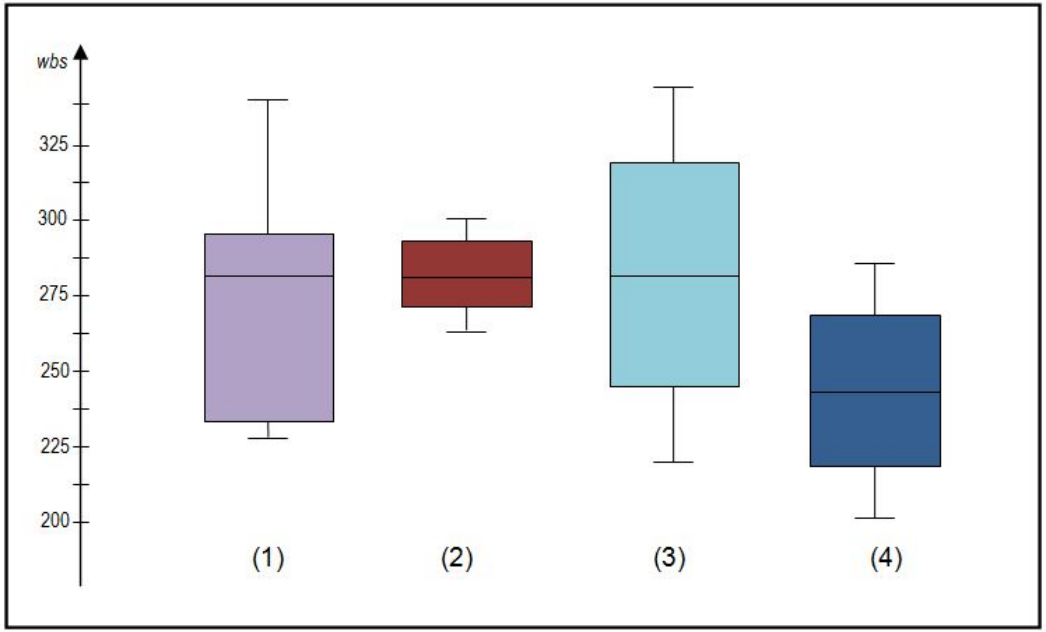


- Interquartile Range (IQR): Q3 - Q1
  - Gives indication of how spread out the data is
- Whiskers on box extend to the most extreme value that is outside of the box (but still within 1.5 x IQR)
- Points beyond the whiskers (outliers) are farther than 1.5 x IQR from the box
  - I.e. Lower than (Q1 - 1.5 x IQR) or higher than (Q3 + 1.5 x IQR)*

* 1.5xIQR rule

Boxplots

- Look at centers and spreads



https://www.wellbeingatschool.org.nz/information-sheet/understanding-and-interpreting-box-plots

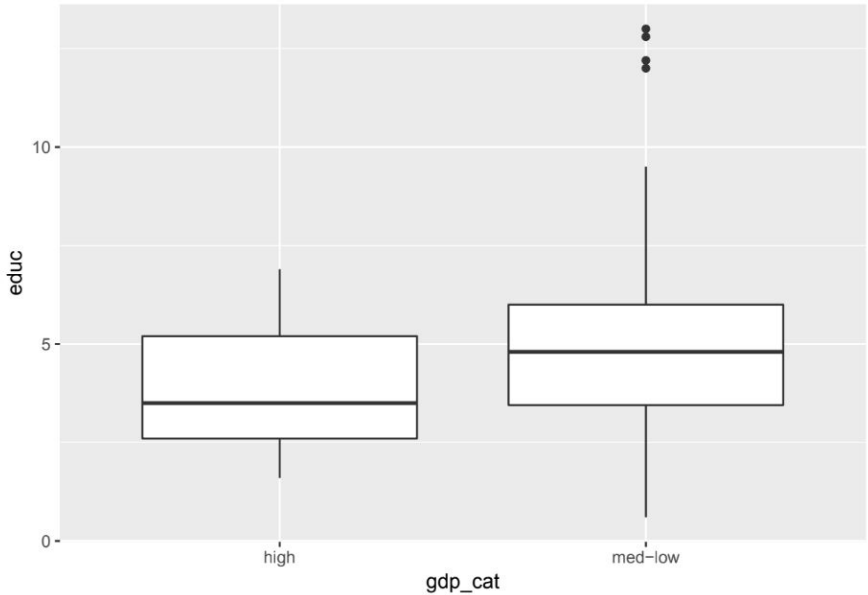
Question (2a)

Used **CIACountries** dataset which contains **gdp** variable

Question (2a)

Wanted to compare the distributions of the proportions of GDP spent on education for countries with a GDP of at least $50,000 compared to countries with a GDP of less than $50,000

```
mutate(gdp_cat = ifelse(gdp >= 50000, "high","med-low")) %>%
filter(is.na(gdp_cat) == F) %>%
ggplot(aes(x = gdp_cat, y= educ)) + geom_boxplot()
```



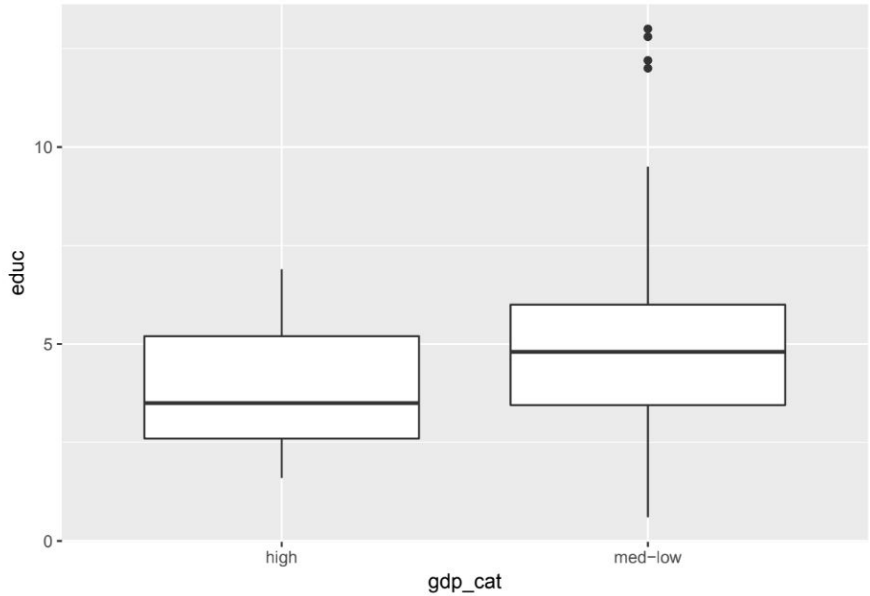
Question (2a)

Interpret the boxplots:

- Center
- Spread

Conclusion: ?

```
mutate(gdp_cat = ifelse(gdp >= 50000, "high","med-low")) %>%
filter(is.na(gdp_cat) == F) %>%
ggplot(aes(x = gdp_cat, y= educ)) + geom_boxplot()
```



Vocabulary for this week (2/2):

Terms for describing data wrangling:

- Cleaning data
- Tidy data
- Create a dummy variable
- Removing the column
- Replace values above/below a certain threshold
- Taking the subset of variables
- Filtering the data frame based on a condition (e.g. based on one of the variables/columns)
- Renaming the variables
- Grouping the categories

Cleaning Data and Tidy Data

Cleaning Data:

- Data tidying structures datasets, making it easier to model, manipulate, and visualize the data
  - I.e. Standard way of structuring dataset

Characteristics of Tidy Data:

(1) Each variable has its own column
(2) Each observation has its own row
(3) Each value has its own cell

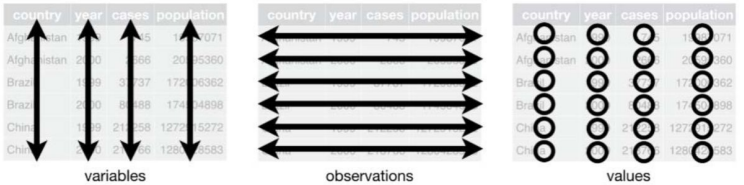


Image from: https://r4ds.had.co.nz/tidy-data.html

If you want to read more: https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html

Quick Summary of Functions from Week 3

select(), filter(), mutate(), rename(), arrange(), summarise()

Summary of Functions (Week 3)

FUNCTION	DESCRIPTION	WHEN USED TOGETHER
select()	Select variables/columns	
filter()	Choose observations/rows	
mutate()	Create new variables from existing variables	Logical test on vector; Populates cells of new variable depending on test results
ifelse()	Evaluates test to get a logical vector	
rename()	Rename variables	
arrange()	Sort data frame	
summarise()	Summarise data frame	Data frame groups that contains summary information for each group
group_by()	Group one or more variables	

mutate() and ifelse()

- Data frame:
  - Gradebook
- Vectors:
  - Section, grade, student

```
section <- c("MATH111", "MATH111", "ENG111")
grade <- c(78, 93, 56)
student <- c("David", "Kristina", "Mycroft")
gradebook <- data.frame(section, grade, student)
mutate(gradebook, Pass.Fail = ifelse(grade > 60,
"Pass", "Fail"))
```

```
  section grade  student Pass.Fail
1 MATH111    78    David      Pass
2 MATH111    93 Kristina      Pass
3  ENG111    56  Mycroft      Fail
```

https://rstudio-pubs-static.s3.amazonaws.com/116317_e6922e81e72e4e3f83995485ce686c14.html#/5

group_by() and summarise()

In this dataset,

- Grouped by types of cells
- Summary information:
  - Number of cells and preferred colour

```
# A tibble: 8 × 3
      sampleGroup     n   color
            <chr> <int>   <chr>
1           E3.25    36 #CAB2D6
2 E3.25 (FGF4-KO)    17 #FDBF6F
3      E3.5 (EPI)    11 #A6CEE3
4  E3.5 (FGF4-KO)     8 #FF7F00
5       E3.5 (PE)    11 #B2DF8A
6      E4.5 (EPI)     4 #1F78B4
7  E4.5 (FGF4-KO)    10 #E31A1C
8       E4.5 (PE)     4 #33A02C
```

https://www.huber.embl.de/users/klaus/Teaching/slides_graphics_and_data_handling.html#(8)