Visualization

Max Turgeon

STAT 4690–Applied Multivariate Analysis

Tidyverse

- For graphics, I personally prefer using `ggplot2` than base `R` functions.
  - Of course, you're free to use whatever you prefer!
- Therefore, I often use the `tidyverse` packages to prepare data for visualization
- Great resources:
  - The book *R for Data Science*
  - RStudio's cheatsheets

Pipe operator

- One of the important features of the `tidyverse` is the pipe operator %>%
- It takes the output of a function (or of an expression) and uses it as input for the next function (or expression)

```r
library(tidyverse)

count(mtcars, cyl)
# Or with the pipe
mtcars %>% count(cyl)
```

Pipe operator

- Note that the LHS (`mtcars`) becomes the first argument of the function appearing on the RHS (`count`)
- In more complex examples, where multiple transformations are applied one after another, the pipe operator improves readibility and avoids creating too many intermediate variables.

Main tidyverse functions

- `mutate`: Create a new variable as a function of the other variables

```
mutate(mtcars, liters_per_100km = mpg/235.215)
```

- `filter`: Keep only rows for which some condition is `TRUE`

```
filter(mtcars, cyl %in% c(6, 8))
```

- `summarise`: Apply summary function to some variables. Often used with `group_by`.

```
mtcars %>% group_by(cyl) %>%
  summarise(avg_mpg = mean(mpg))
```

Data Visualization

Main principles

Why would we want to visualize data?

- Quality control
- Identify outliers
- Find patterns of interest (EDA)

Visualizing multivariate data

- To start, you can visualize multivariate data one variable at a time.
- Therefore, you can use the same visualizing tools you're likely familiar with.

Histogram i

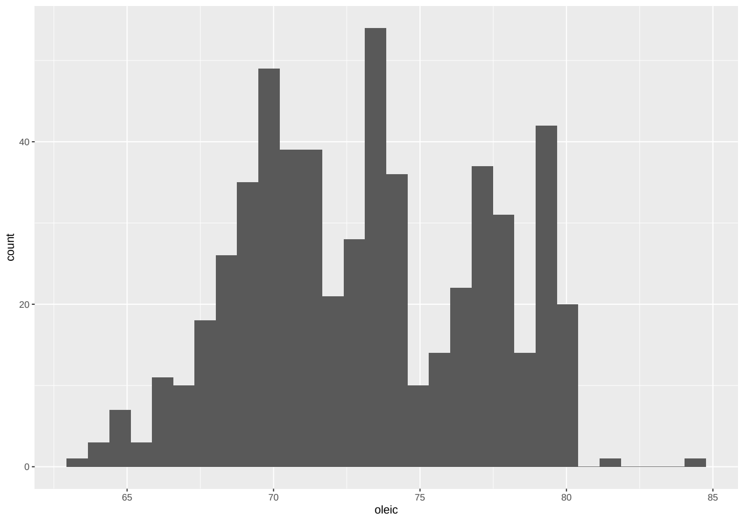
```r
library(tidyverse)
library(dslabs)

dim(olive)
```

```
## [1] 572  10
```

```r
olive %>%
  ggplot(aes(oleic)) +
  geom_histogram()
```

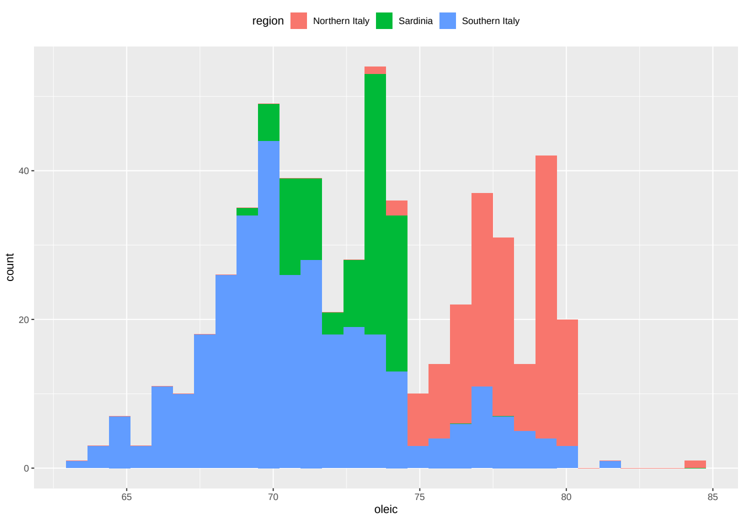
Histogram ii



Histogram iii

```r
olive %>%
  ggplot(aes(oleic, fill = region)) +
  geom_histogram() +
  theme(legend.position = 'top')
```

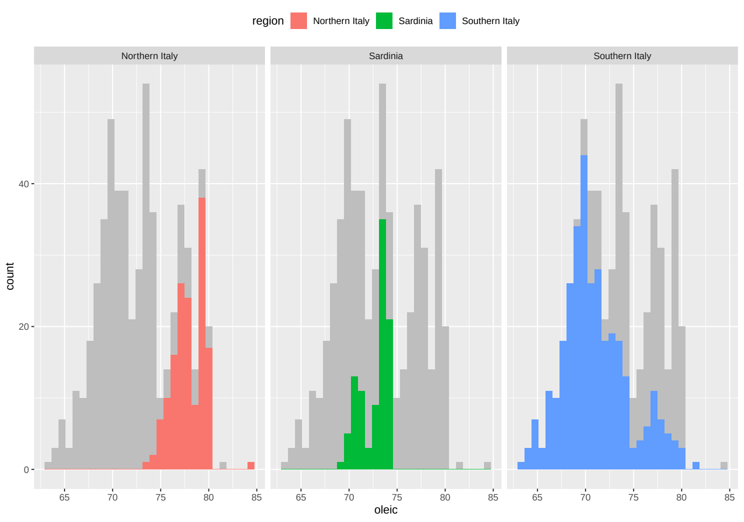
Histogram iv



Histogram v

```r
# Or with facets
olive_bg <- olive %>% dplyr::select(-region)
olive %>%
  ggplot(aes(oleic, fill = region)) +
  geom_histogram(data = olive_bg,
                 fill = 'grey') +
  geom_histogram() +
  facet_grid(. ~ region) +
  theme(legend.position = 'top')
```

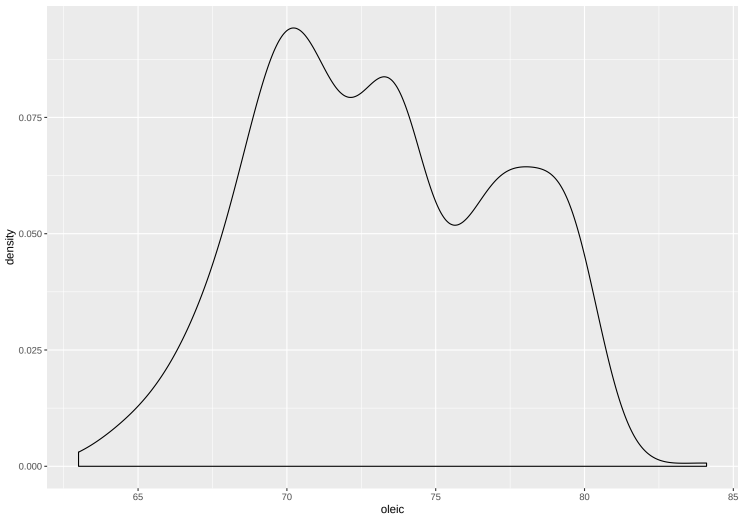
Histogram vi



Density plots i

```
olive %>%
  ggplot(aes(oleic)) +
  geom_density()
```

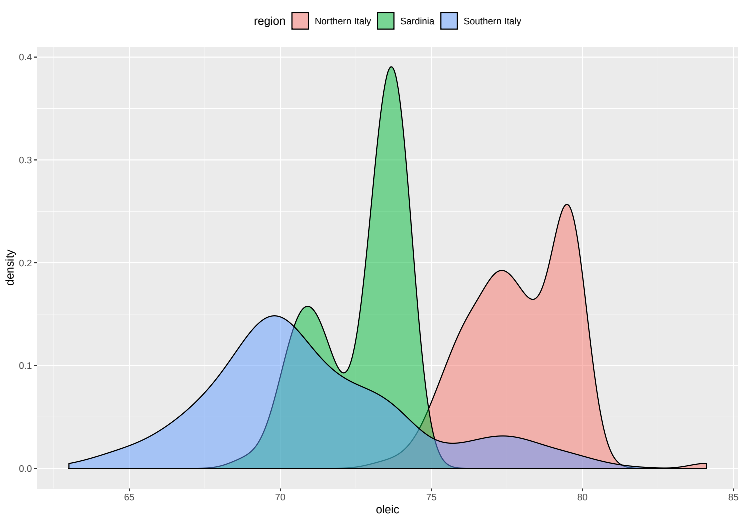
Density plots ii



Density plots iii

```r
olive %>%
  ggplot(aes(oleic, fill = region)) +
  geom_density(alpha = 0.5) +
  theme(legend.position = 'top')
```

Density plots iv



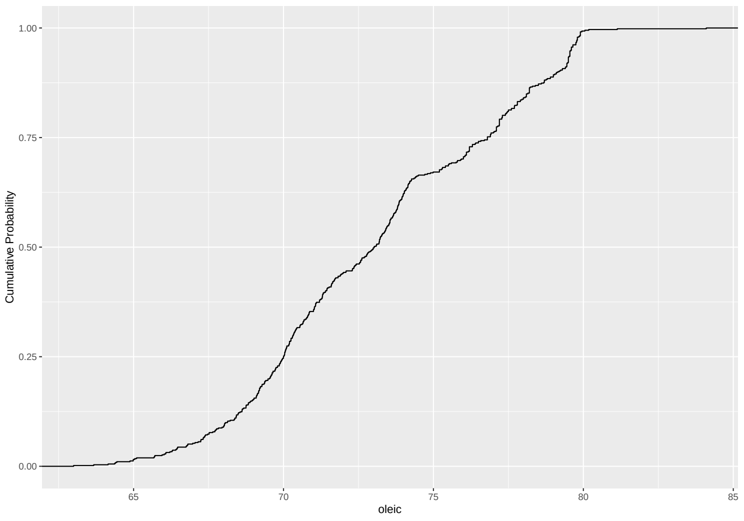
ECDF plots i

- Density plots are “smoothed histograms”
- The smoothing can hide important details, or even create artifacts
- Another way of looking at the distribution: **Empirical CDFs**
  - Easily compute/compare quantiles
  - Steepness corresponds to variance

ECDF plots ii

```r
olive %>%
  ggplot(aes(oleic)) +
  stat_ecdf() +
  ylab("Cumulative Probability")
```

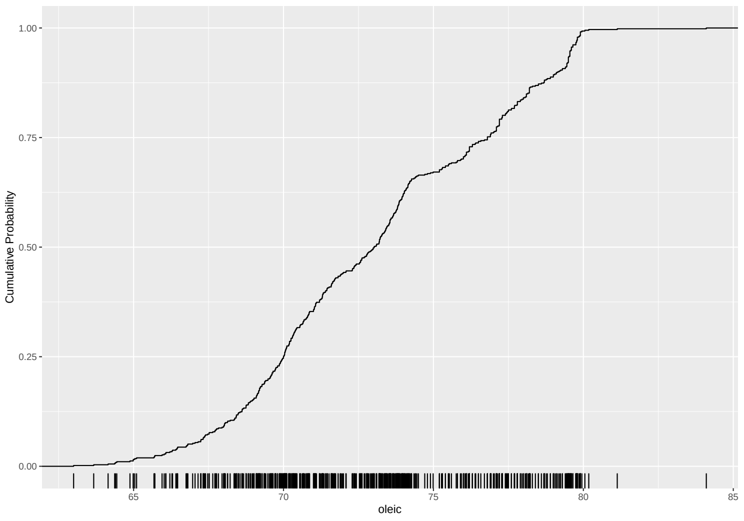
ECDF plots iii



ECDF plots iv

```
# You can add a "rug"
olive %>%
  ggplot(aes(oleic)) +
  stat_ecdf() +
  geom_rug(sides = "b") +
  ylab("Cumulative Probability")
```

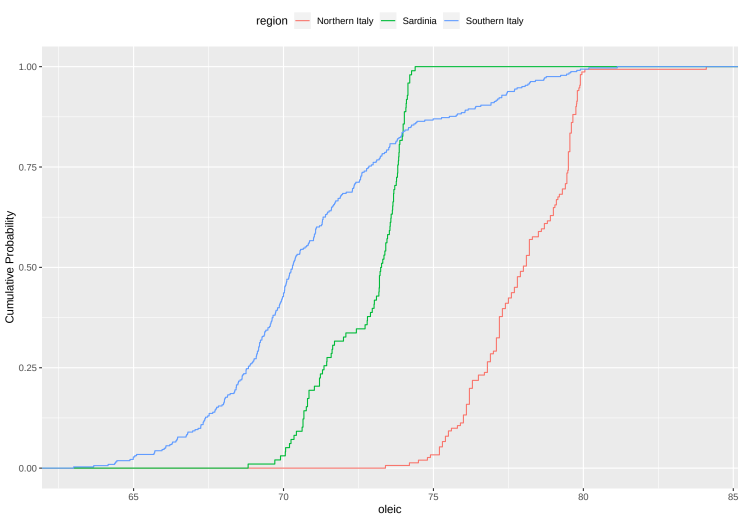
ECDF plots v


1.00
0.75
0.50
0.25
0.00
Cumulative Probability
65
70
75
80
85
oleic

ECDF plots vi

```r
olive %>%
  ggplot(aes(oleic, colour = region)) +
  stat_ecdf() +
  ylab("Cumulative Probability") +
  theme(legend.position = 'top')
```

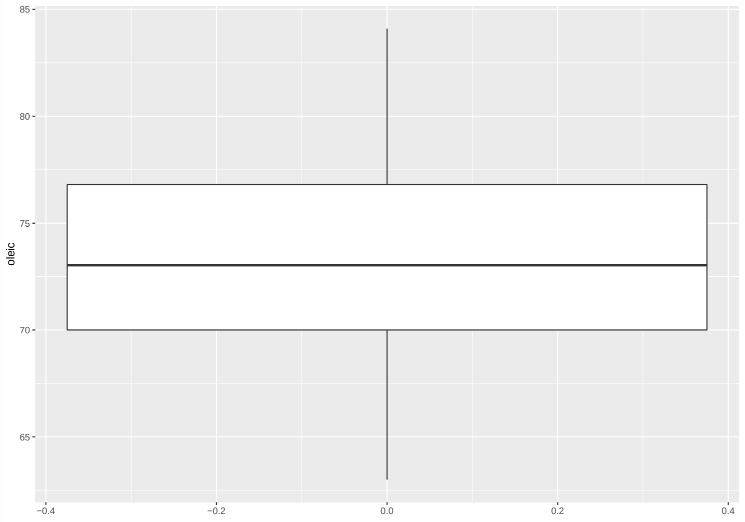
ECDF plots vii



Boxplots i

```
olive %>%
  ggplot(aes(y = oleic)) +
  geom_boxplot(x = 0)
```

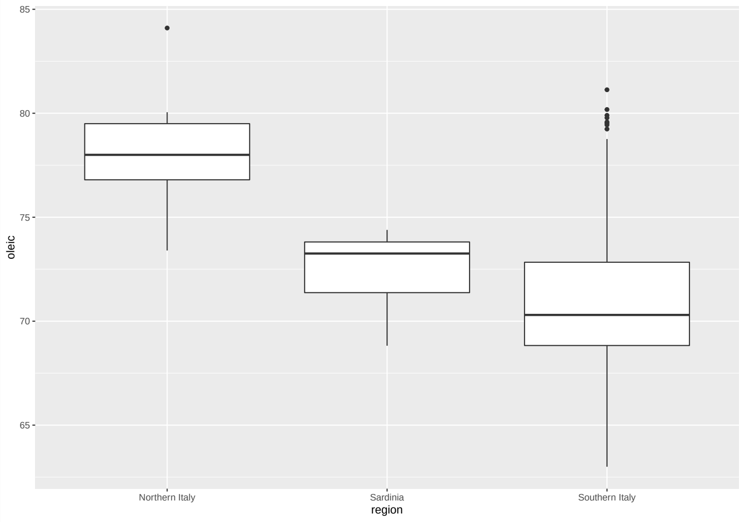
Boxplots ii



Boxplots iii

```
olive %>%
  ggplot(aes(x = region, y = oleic)) +
  geom_boxplot()
```

Boxplots iv

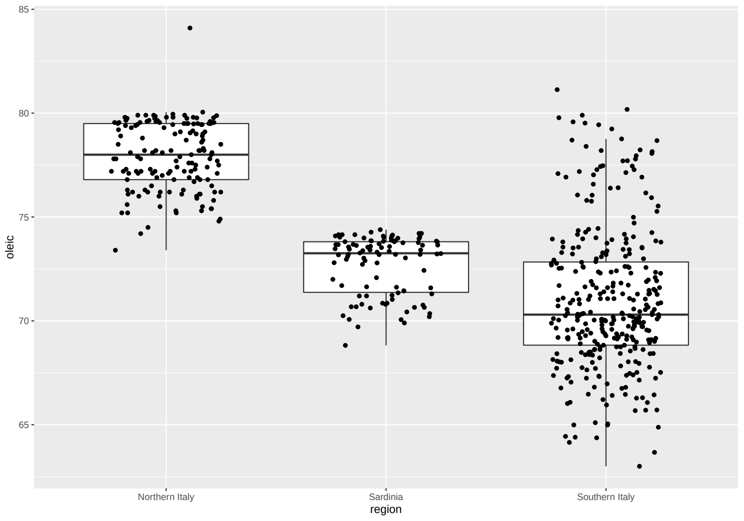


85
80
75
70
65
oleic
Northern Italy
Sardinia
Southern Italy
region

Boxplots v

```
# Add all points on top of boxplots
# Note: need to remove outliers or you will get
#       duplicates
olive %>%
  ggplot(aes(x = region, y = oleic)) +
  geom_boxplot(outlier.colour = NA) +
  geom_jitter(width = 0.25, height = 0)
```

Boxplots vi



85
80
75
70
65
oleic
Northern Italy
Sardinia
Southern Italy
region