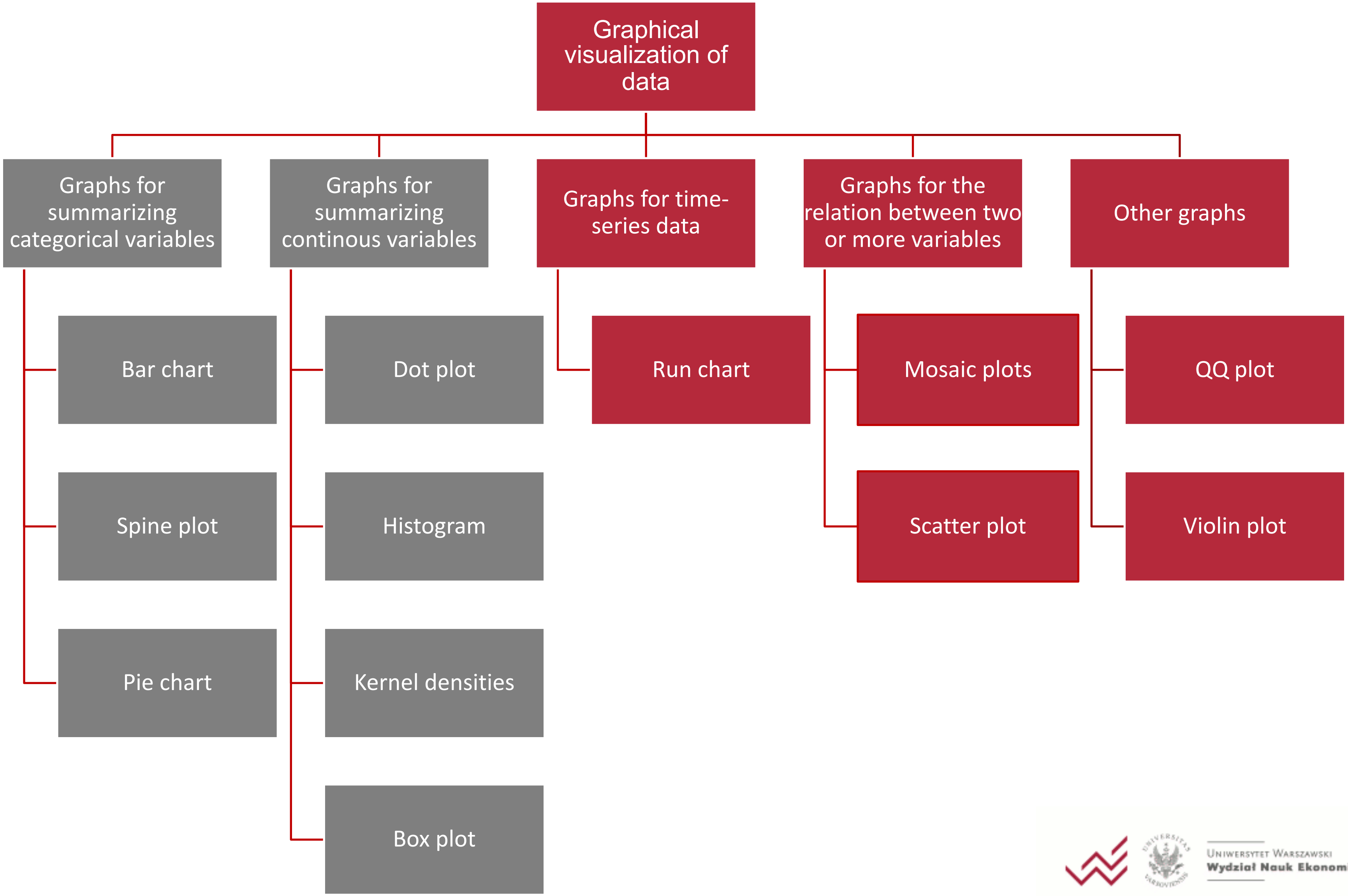
# Graphical analysis of data (I)

**Marcin Chlebus, Ewa Cukrowska-Torzewska**
**Faculty of Economic Sciences**
**University of Warsaw**

**Lecture 4: 24-25.10.2017**

Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

# Types of graphs



Graphical visualization of data

- Graphs for summarizing categorical variables
  - Bar chart
  - Spine plot
  - Pie chart
- Graphs for summarizing continous variables
  - Dot plot
  - Histogram
  - Kernel densities
  - Box plot
- Graphs for time-series data
  - Run chart
- Graphs for the relation between two or more variables
  - Mosaic plots
  - Scatter plot
- Other graphs
  - QQ plot
  - Violin plot

# Categorical data

- Categorical data include:
  - Nominal data → the values cannot be ordered (e.g. male-female)
  - Ordinal data → the values can be ordered but the differences between the values are not informative (e.g. education level)
  - Discrete data → the values are counted (e.g. the number of children)

**There is a certain number of values of the categorical variable that we want to plot!**

# Bar chart

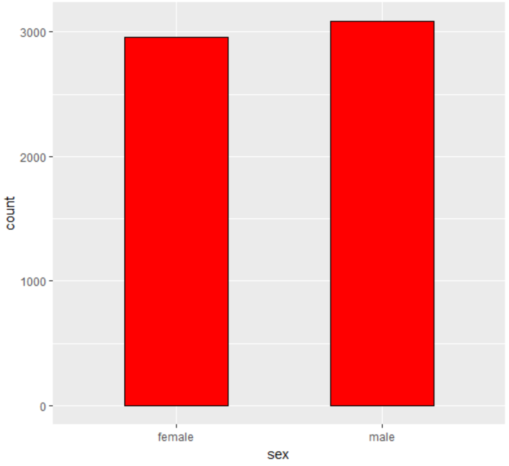
- Any categorical variable may be summarized by one-dimensional table

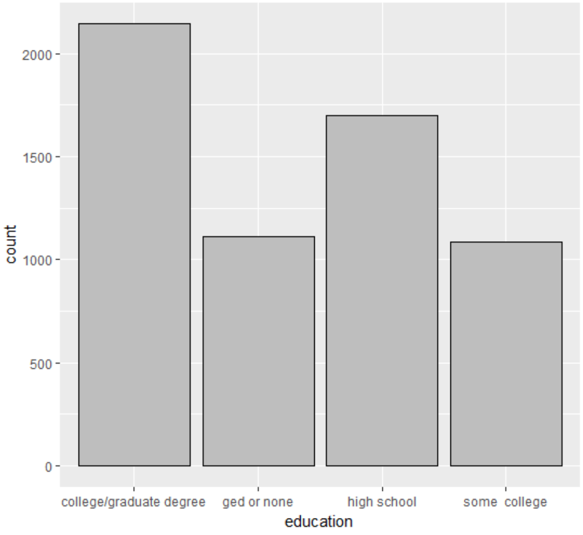
- The easiest way to summarize the data is a bar chart, where the area of the bar represents the count for its category

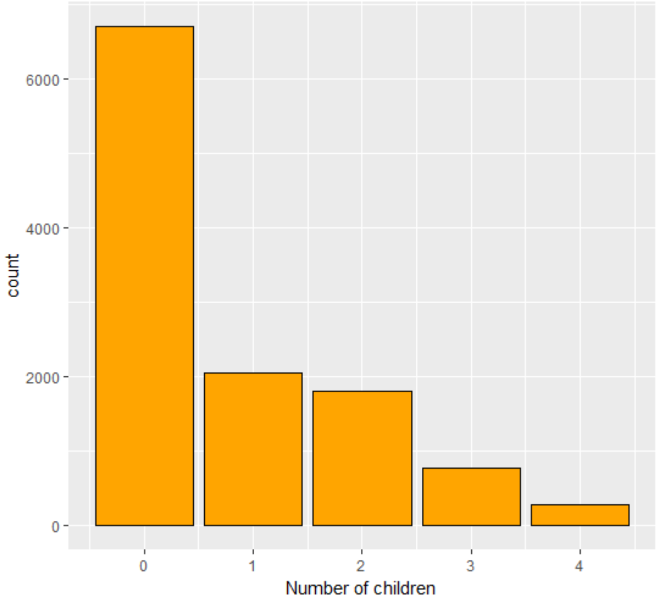
| Female | Male |
|---|---|
| 2959 | 3085 |

| college/ graduate degree | ged or none | high school | some college |
|---|---|---|---|
| 2144 | 1114 | 1699 | 1087 |

| 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 6703 | 2051 | 1801 | 762 | 275 |

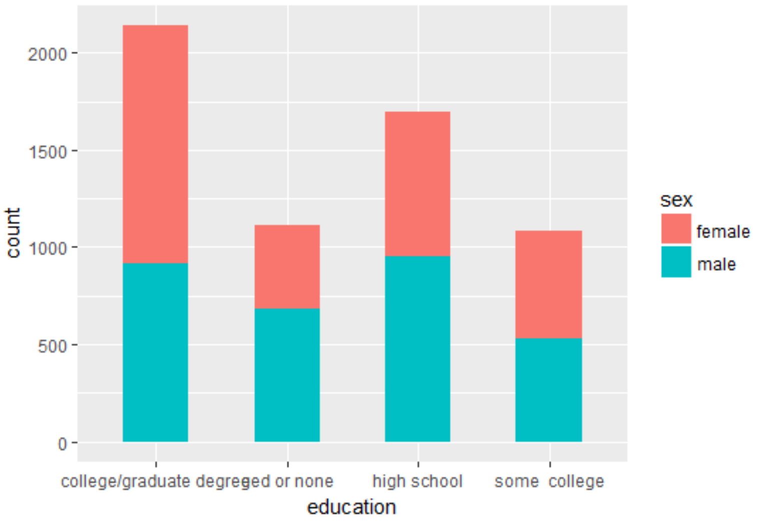


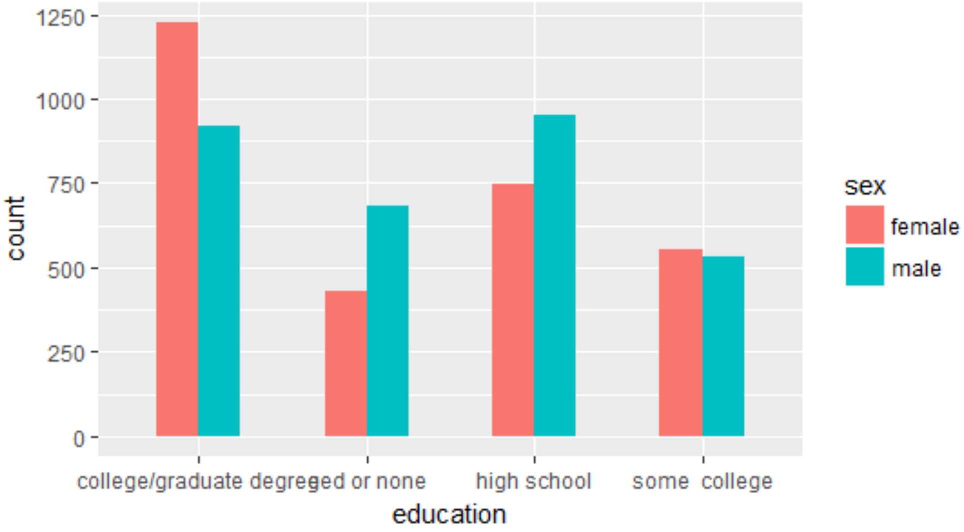




# Bar chart

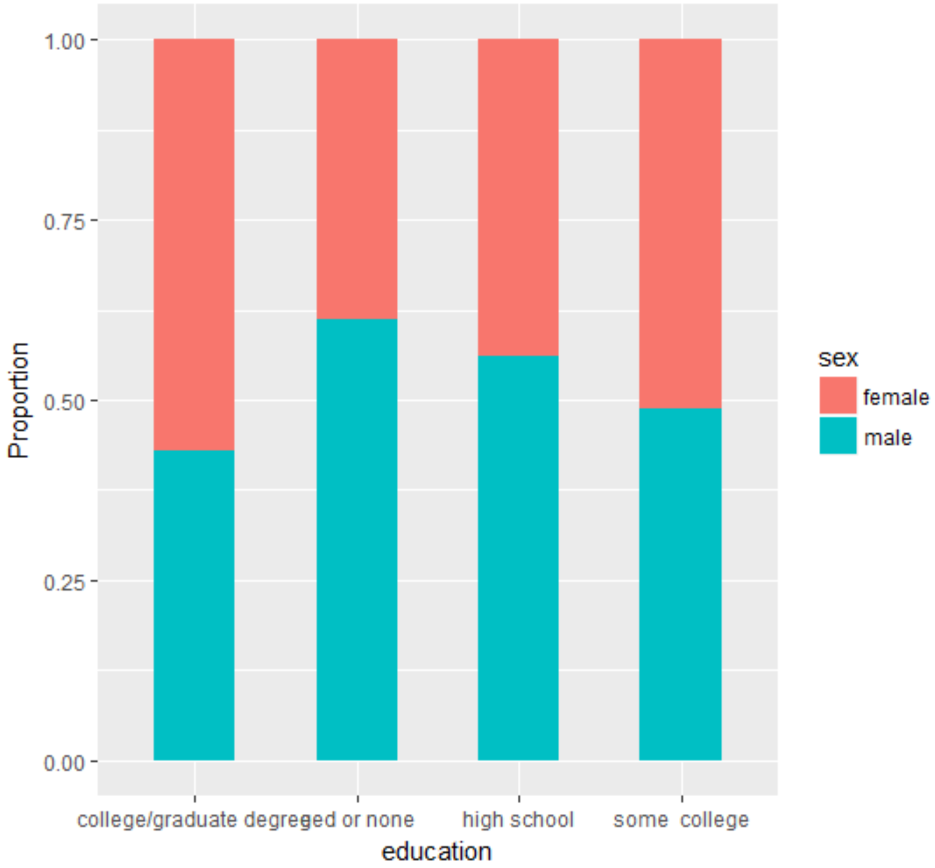
- In many situations it is desirable to look at the distribution of a subgroup of a categorical variable → **e.g. education by gender**





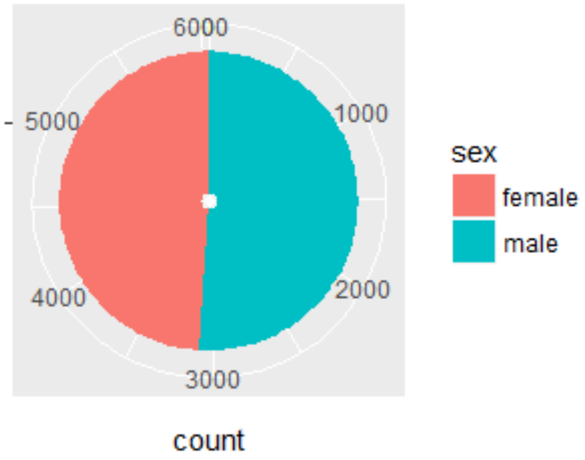
# Spine plot

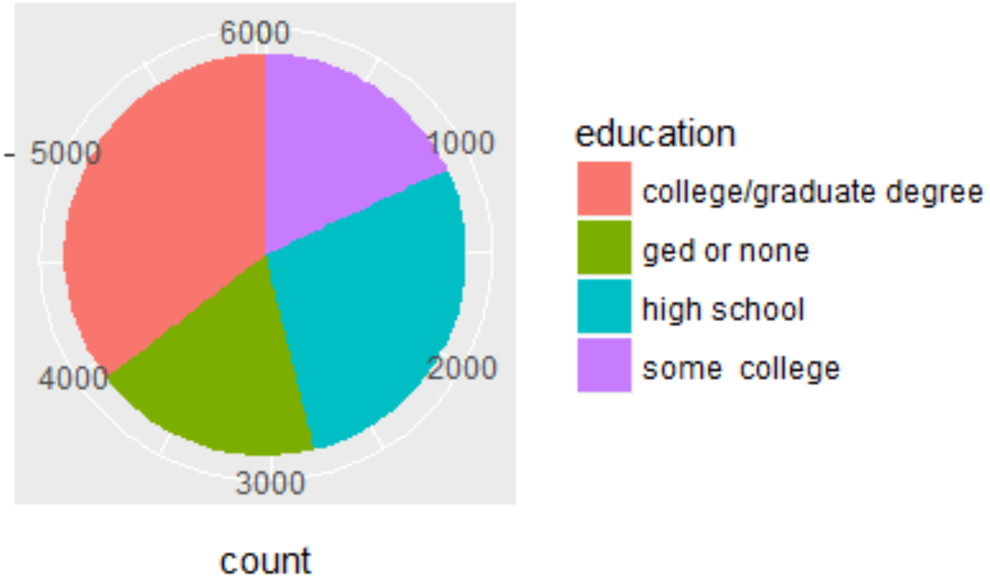
- While the bar chart shows the absolute counts of the subgroup, the spine plot shows the proportions



# Pie chart

- Pie chart is also a useful chart to summarize proportions
- Each „slice” of the pie represents the relative size of a given category in the data.





# Graphs for categorical variables in R

**Exercise 1:**

Install package „car” and use the dataset „Salaries” from that package. The dataset contains salaries of the nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. Using the dataset:

- identify categorical variables
- visualize the data using the graphs for categorical variables.

# Continous variables

- As opposed to discrete and categorical variables, which are counted, **continous variables are measured.**

- Examples: lenght, height, weight,
  wages, hours of work, prices, etc.

# Dot plot

- The simpliest way to plot the values of one continous variable is the dot plot.

- Because the points are distributed along only one axis, overplotting is a serious problem.

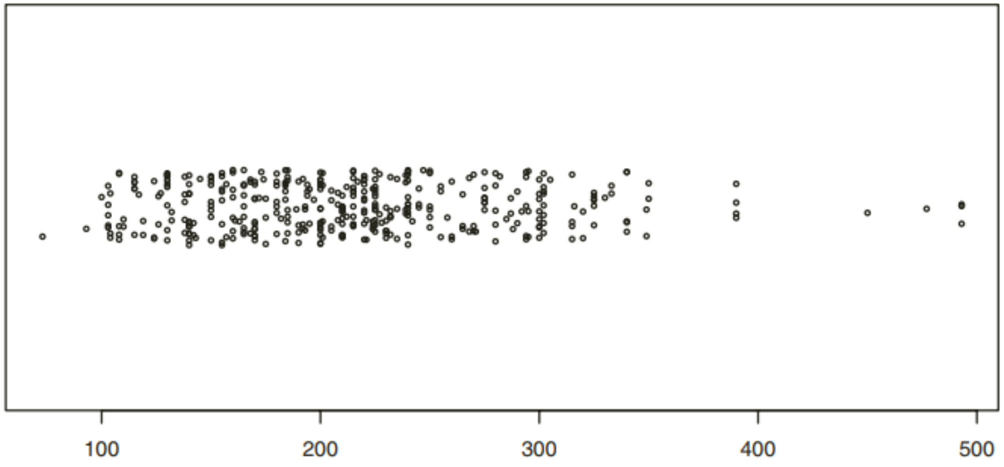
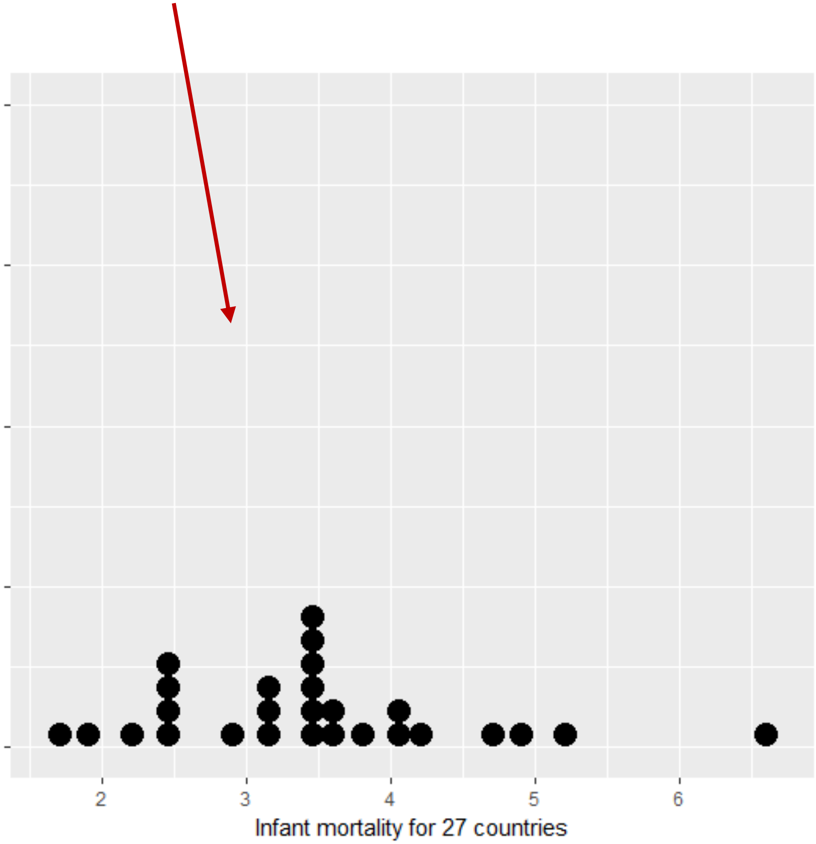


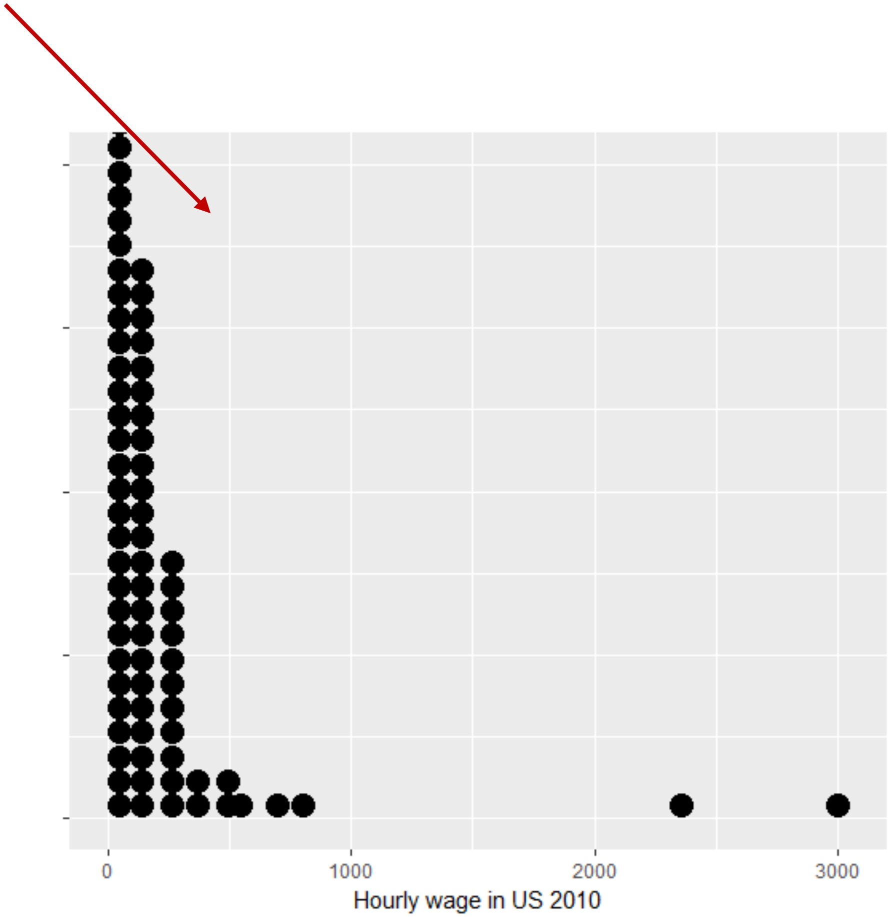
**Fig. 2.7.** *A jittered dotplot of* Horsepower *for the Cars2004 data.*

Source: Graphic of large datasets – Visualizing a Milion; Antony Unwin et al. (2006)

# Wilkinson dot plot

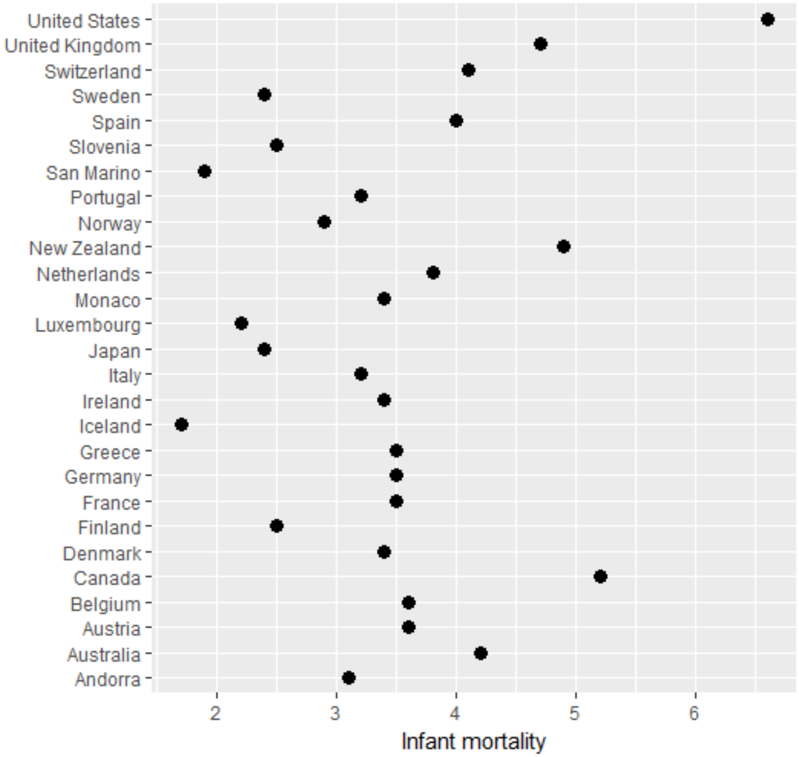
- In this plot each dot represents one observation of a given value of x variable
- It may not be efficient way of visualizing **large datasets**, but it performs well in the case of **small samples**

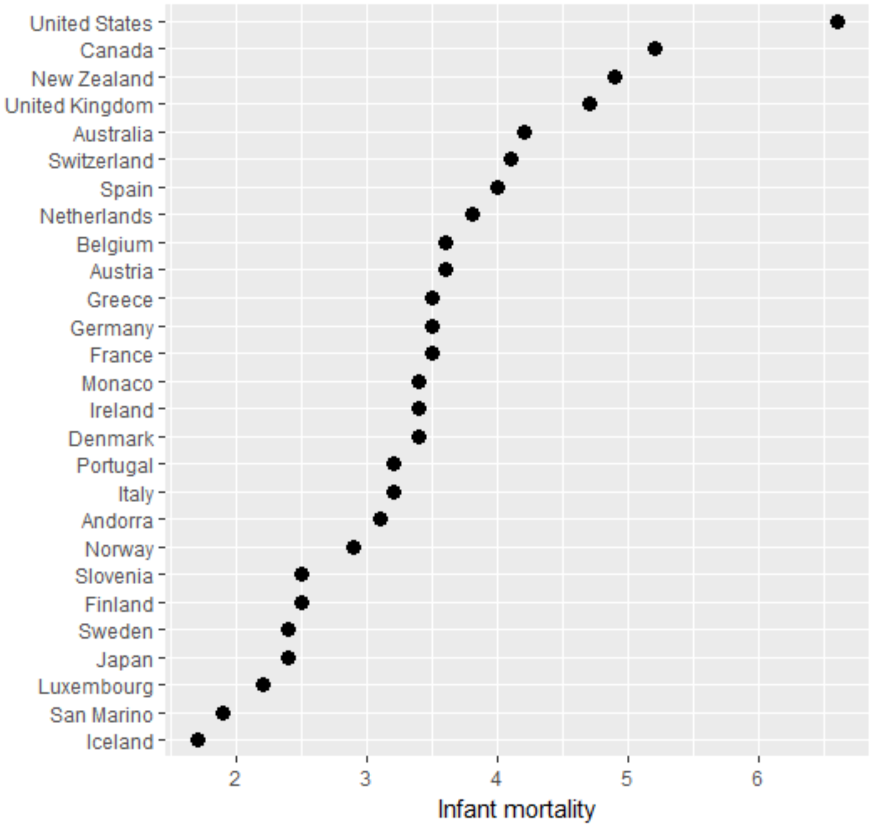




# Cleveland dot plot

- Cleveland dot plots are sometimes used instead of bar graphs.
- The main advantage of these plots over bar plots is that they are easier to read
- They usually pertain to two variables





# Dot plot

**Exercise 2:**

Use the dataset „Salaries”.

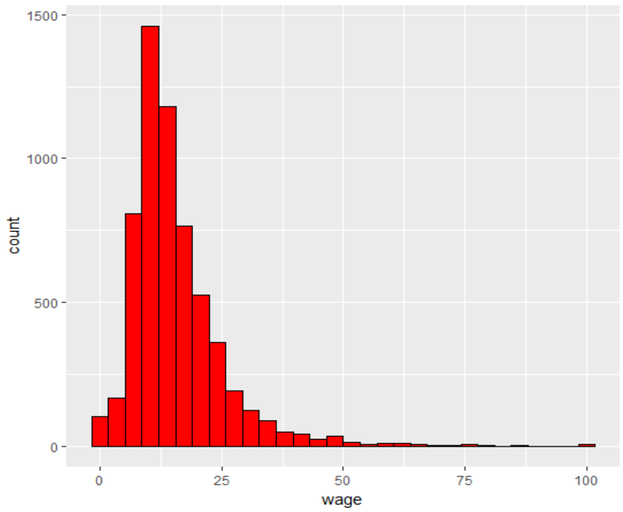
- Use simple dot plot to visualize data on Professors’ salaries.
  What can you say about the distribution of the salaries based on this plot?

- Use Wilkinson dot plot to visualize data on Professors' salaries.
  What can you say about the distrubtion of the salaries based on this plot?

- Try to use Cleveland dot plot to visualize data on Professor’s salaries.

# Histogram

- Histogram is very often used to represent counts of the distribution of the continous variable.

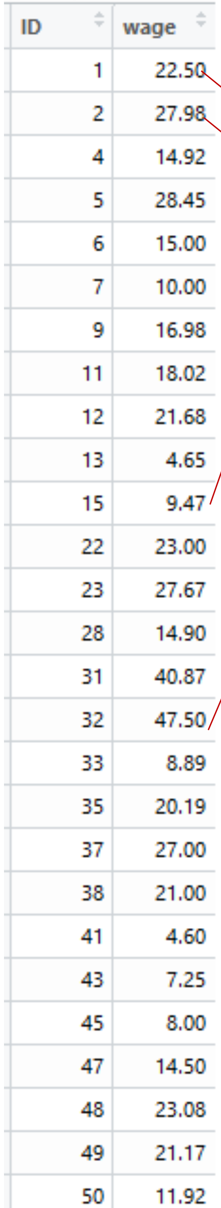


- The interpretation of the histograms makes them somewhat comparable to bar charts for categorical variables nd Wilkinson dot plots for small samples.

- The difference between the histogram and the bar chart is that **the number of the bins of a histogram is not determined a priori and the bins are set to represent the continous scale of the data** (i.e. we need to create „categories” for the continous scale)

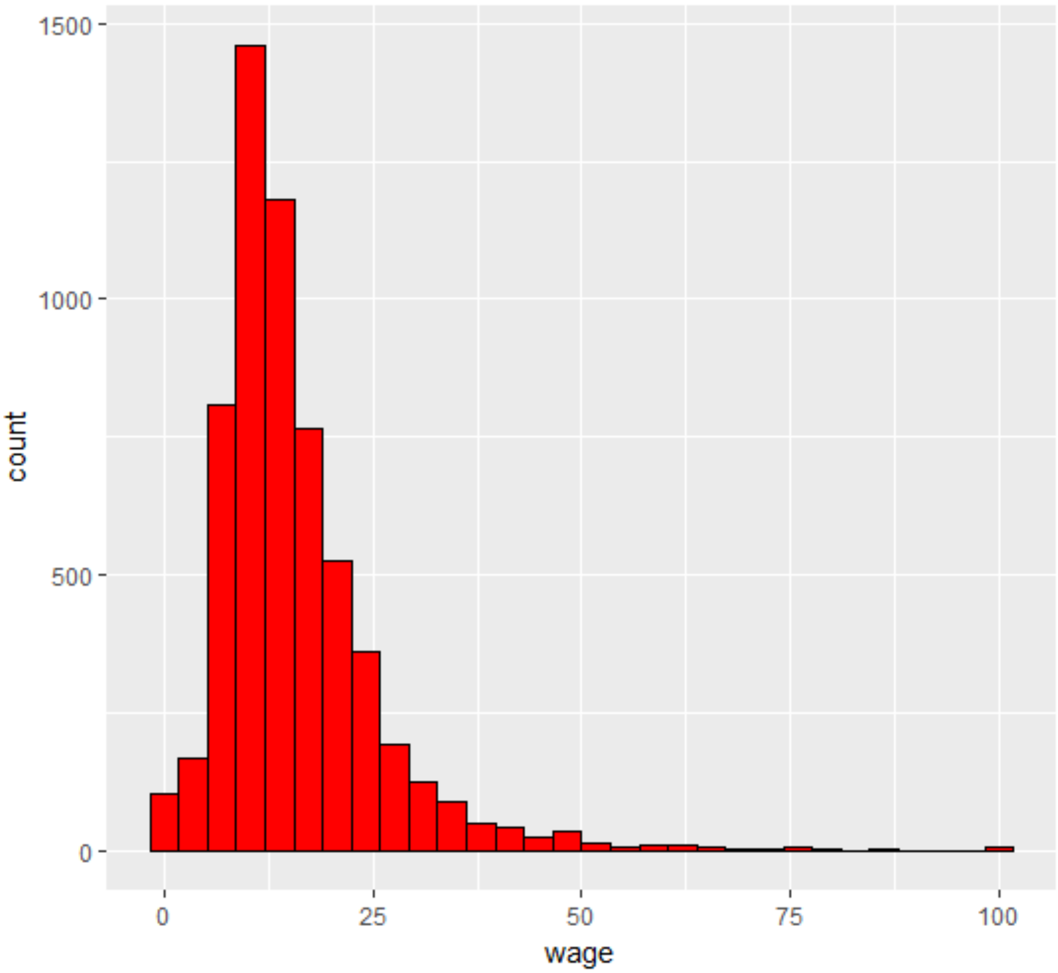
- To determine histogram we need to specify bins width and the starting position of the first bin.

# Histogram

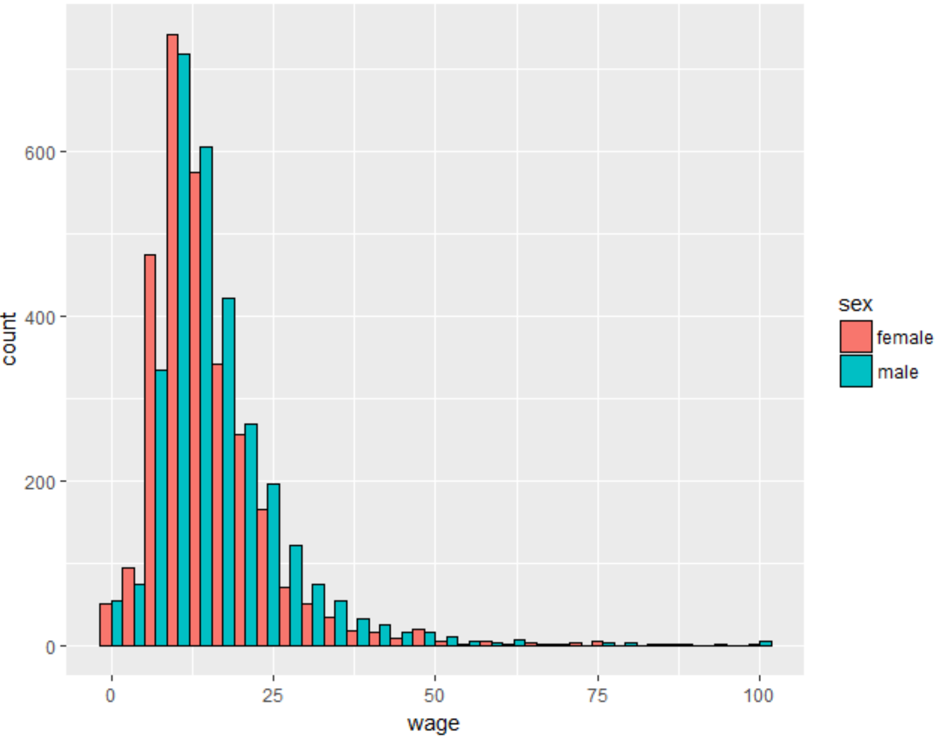


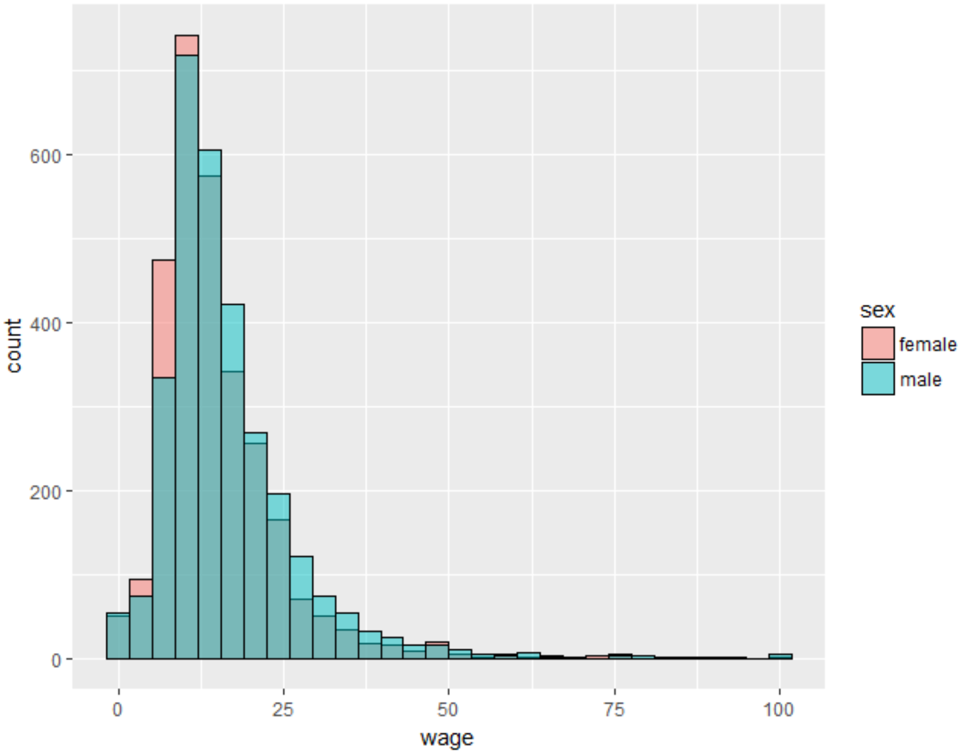
| ID | wage |
|---|---|
| 1 | 22.50 |
| 2 | 27.98 |
| 4 | 14.92 |
| 5 | 28.45 |
| 6 | 15.00 |
| 7 | 10.00 |
| 9 | 16.98 |
| 11 | 18.02 |
| 12 | 21.68 |
| 13 | 4.65 |
| 15 | 9.47 |
| 22 | 23.00 |
| 23 | 27.67 |
| 28 | 14.90 |
| 31 | 40.87 |
| 32 | 47.50 |
| 33 | 8.89 |
| 35 | 20.19 |
| 37 | 27.00 |
| 38 | 21.00 |
| 41 | 4.60 |
| 43 | 7.25 |
| 45 | 8.00 |
| 47 | 14.50 |
| 48 | 23.08 |
| 49 | 21.17 |
| 50 | 11.92 |

| "Category" = bin | Count |
|---|---|
| 0-5 | 105 |
| 5-10 | 169 |
| 10-15 | 809 |
| 15-20 | 1459 |
| 20-25 | 1179 |
| 25-30 | 764 |
| 30-35 | 525 |
| 35-40 | 361 |
| 40-45 | 192 |
| 45-50 | 126 |
| 50-55 | 90 |
| 55-60 | 50 |
| 60-65 | 41 |
| 65-70 | 26 |
| 70-75 | 36 |
| 75-80 | 15 |
| 80-85 | 6 |
| 85-90 | 9 |
| 90-95 | 10 |
| 95-100 | 6 |
| 100-105 | 2 |
| 105-110 | 4 |
| 110-115 | 8 |
| 115-120 | 4 |
| 120-125 | 1 |
| 125-130 | 2 |
| 130-135 | 1 |
| 135-140 | 1 |
| 140-145 | 0 |
| 145-150 | 7 |



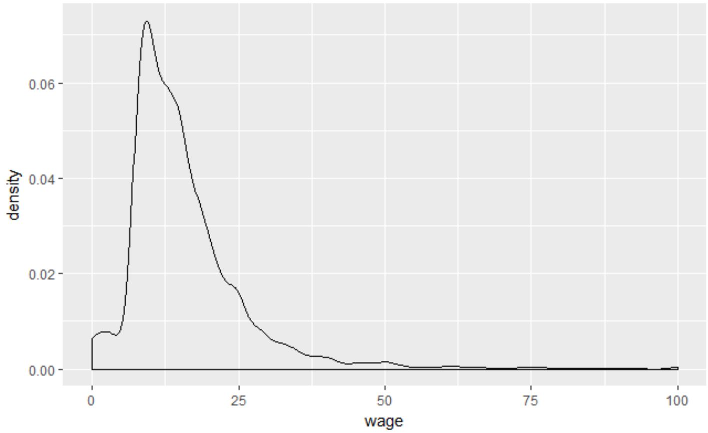
# Histogram


count
600
400
200
0
0
25
50
75
100
wage
sex
female
male


count
600
400
200
0
0
25
50
75
100
wage
sex
female
male

# Density plot (kernel density)

- A density plot is a variation of histogram that uses kernel smoothing function to smooth the distribution.



- Compared to histograms, density plots are better at determining the shape of the distribution because they are not affected by the number of bins

- More formally, kernel smoothing functions smooth out the contribution of each observed data point (x) over a local neighborhood of that data point (x(i)):

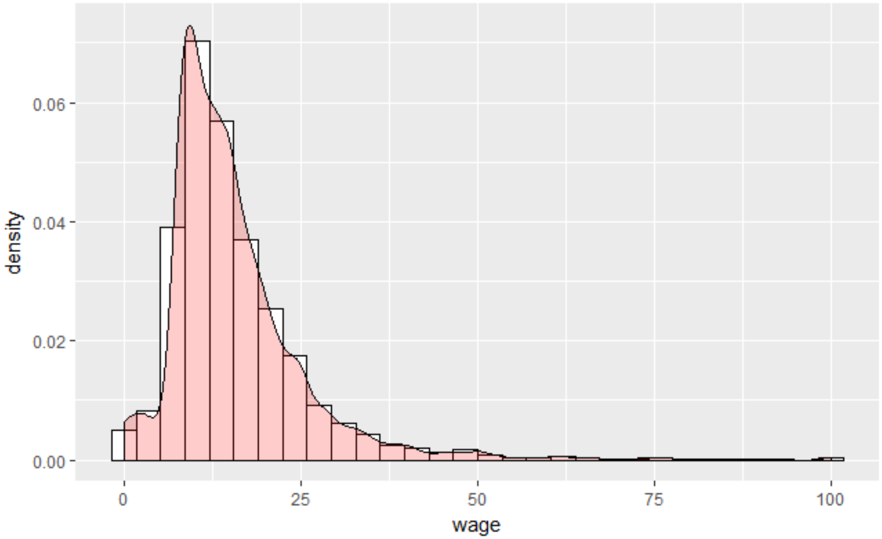
$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left( \frac{x - x(i)}{h} \right)$$

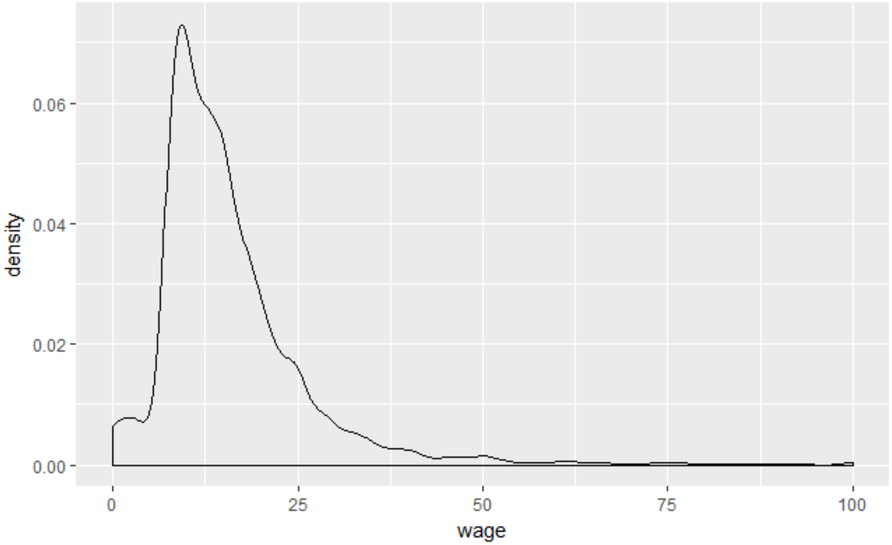
The estimated density at any point x

Kernel function – it determines the weight given to each x at point x(i) based on their proximity

Bandwidth = smoothing parameter = the size of the neighborhood around x(i):

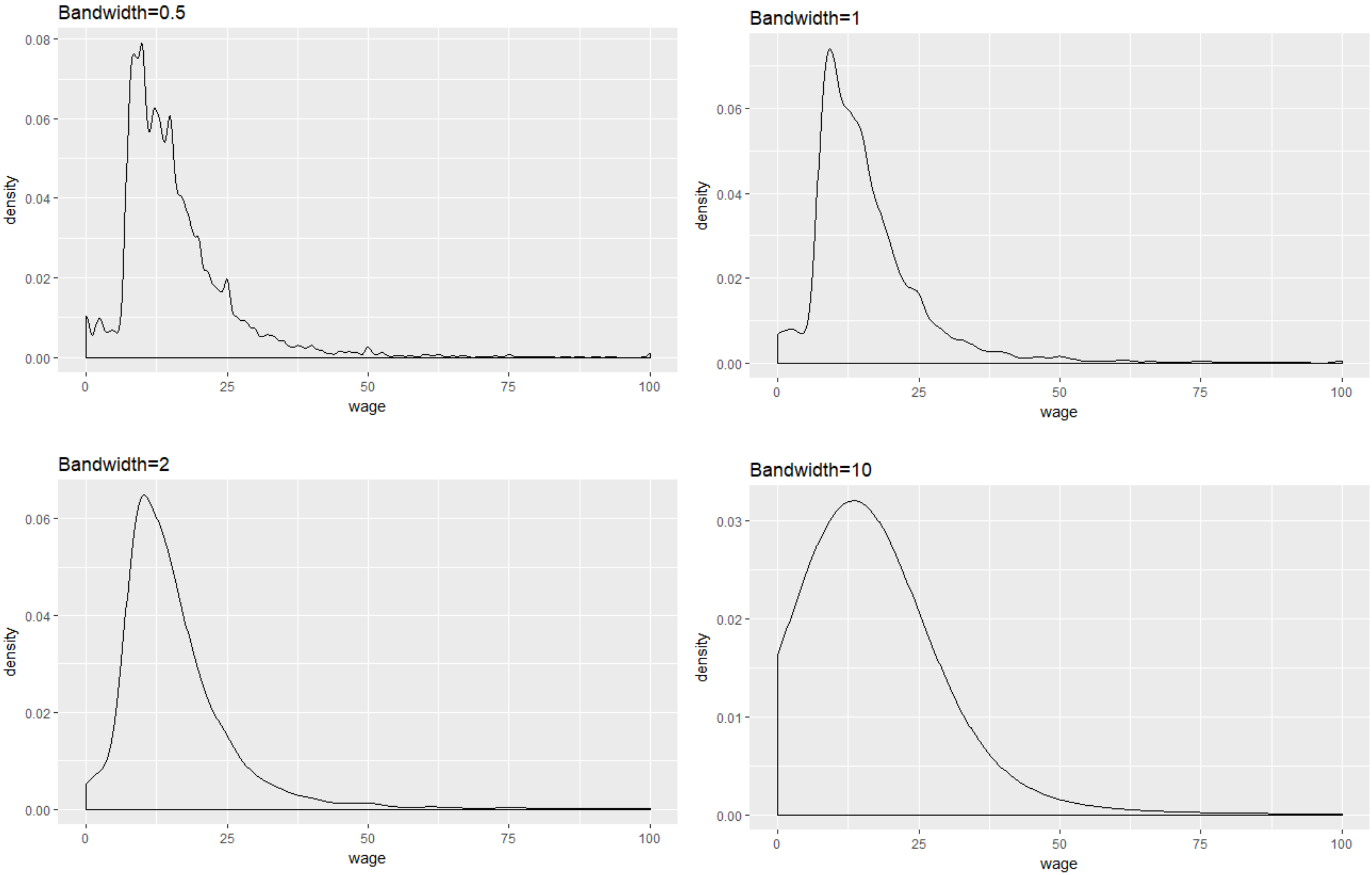
# Density plot (kernel density)





# Density plot (kernel density)

Too large h will oversmooth the data
Too small h will undersmooth the data



# Histogram and density plot

**Exercise 3:**

Use the dataset „Salaries”.

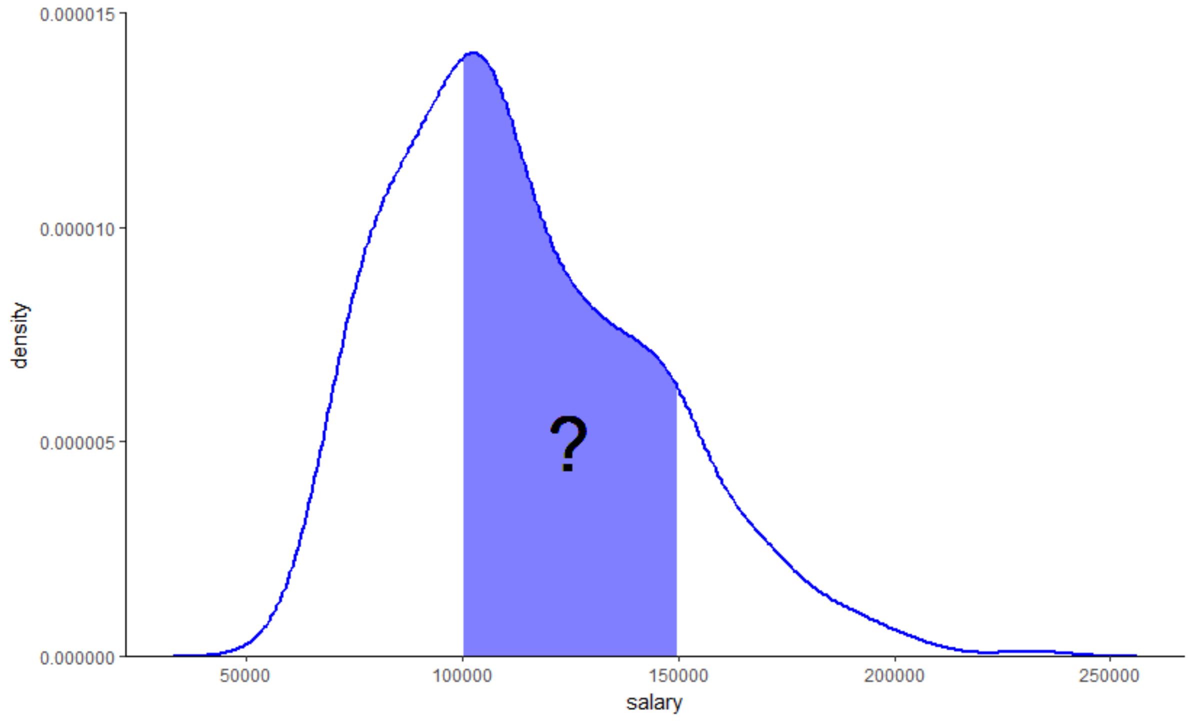
- Create histograms and density plots for Professors’ salaries
- Create histograms and density plots for Professors’ salaries by sex/rank/discipline

Intepret the graphs.

# Histogram and density plot

**Exercise 4:**

How will you interpret the area filled with blue on the following density graph?

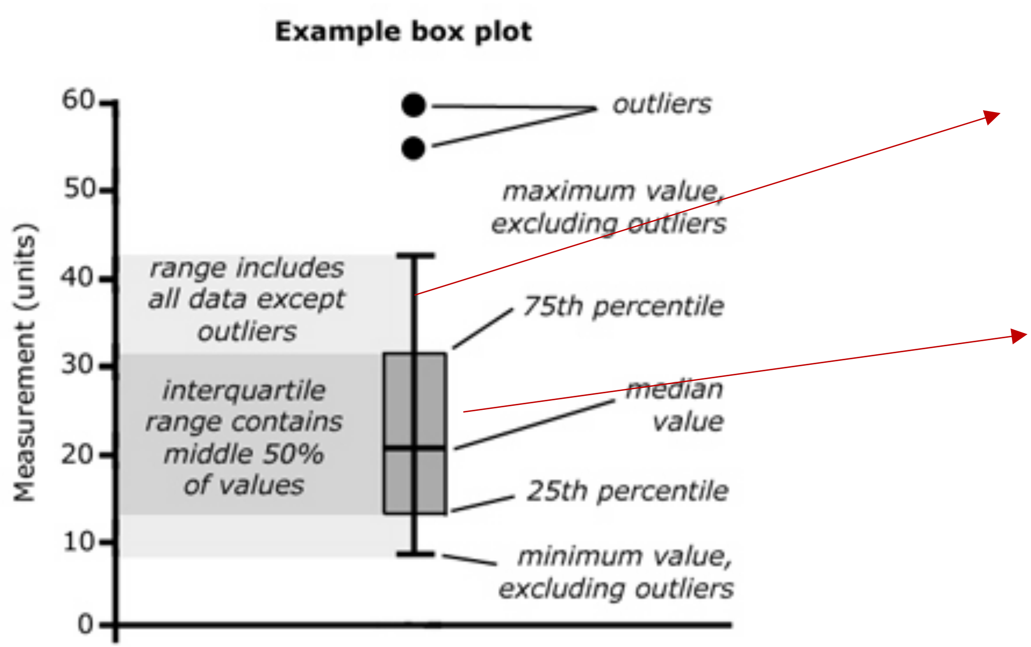


# Box plot

- Boxplots are a mixture of summary information (like histograms) and information on individual points (like dot plots).

- It summarizes the distribution by plotting quartiles.

- Outliers are often plotted as individual points → boxplot is often used to detect outliers


Example box plot
60
50
40
30
20
10
0
Measurement (units)
outliers
maximum value, excluding outliers
range includes all data except outliers
75th percentile
interquartile range contains middle 50% of values
median value
25th percentile
minimum value, excluding outliers

Whisker – it is set using the IQR rule (1.5+/- IQR)

Box

Source: https://lon03.wordpress.com/

# Box plot

**Exercise 5:**

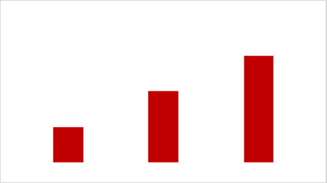
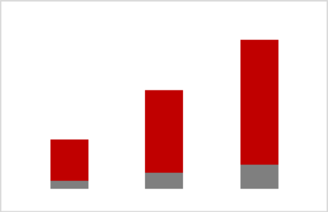
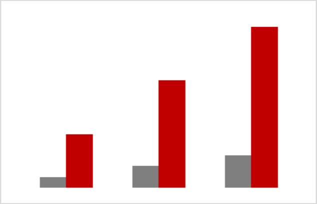
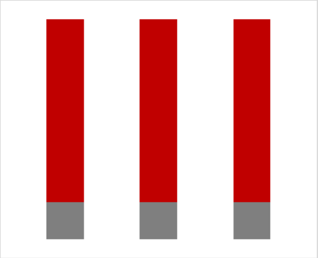
Create the boxplot for salaries data.
Based on your graph give answers to the following questions:

- What are the range, the three quartiles and the interquartile range? Check your answers with calculating the relevant quartiles

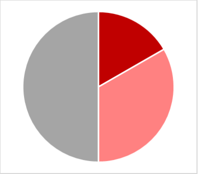
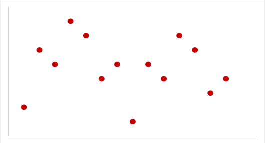
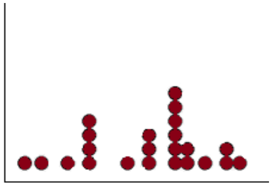
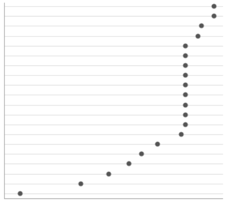
- If the 40th percentile is equal to 100 000, about how much of the sample earns between 91 000 and 100 000?

- About how many outliers can you identify in the data?

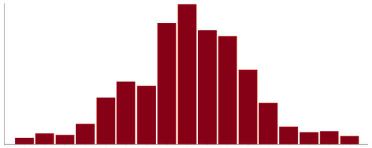
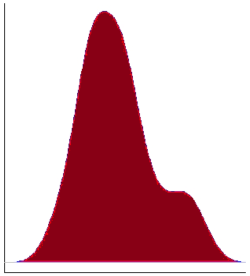
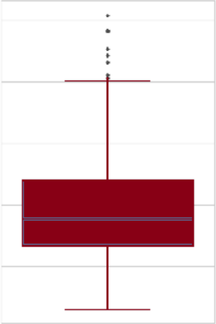
# Data visualization in R

| | Built in functions | Ggplot package |
|---|---|---|
| Simple bar chart  | bars <- table(var1)<br>barplot(bars) | ggplot(data=., aes(x=var1)+<br>geom_bar(stat="count") |
| Stacked bar chart  | bars <- table(var1,var2)<br>barplot(bars) | ggplot(data=., aes(x=var1, fill=var2))+<br>geom_bar(stat="count") |
| Grouped bar chart  | bars <- table(var1,var2)<br>barplot(bars, beside=TRUE) | ggplot(data=., aes(x=var1, fill=var2))+<br>geom_bar(stat="count", position="dodge") |
| Spine plot  | spineplot(var1 ~ var2) | ggplot(data=., aes(x=var1, fill=var2))+<br>geom_bar(stat="count", position="fill") |

# Data visualization in R

| | Built in functions | Ggplot package |
|---|---|---|
| Pie chart<br> | slices <- table(var1)<br>pie(slices) | g <- ggplot(data=., aes(x="", fill=var1))+geom_bar()<br>g + coord_polar(theta="y", start=0) |
| Dot plot<br> | plot(var1, type='p') | NA |
| Wilkinson dot plot<br> | Not available with built in funcitons.<br>Can use:<br>install.packages("BHH2")<br>library(BHH2)<br>dotPlot(var1) | ggplot(data=., aes(x=var1)) + geom_dotplot() |
| Cleveland dot plot<br> | dotchart(var1) | ggplot(data=., aes(x=var1, y=reorder(var2, var1))) + geom_point() |

# Data visualization in R

| | **Built in functions** | **Ggplot package** |
|---|---|---|
| Histogram  | hist(var1) | ggplot(data=., aes(x=var1)) +<br>geom_histogram() |
| Kernel density  | d <- density(var1)<br>plot(d) | ggplot(data=., aes(x=var1)) +<br>geom_density() |
| Boxplot  | boxplot(var1) | ggplot(data=., aes(x="", y=var1)) +<br>geom_boxplot() |

# Bibliography

Antony Unwin, Martin Theus, Heike Hofmann, Graphics of Large Datasets Visualizing a Million. Springer 2006.

Winston Chang. Practical Recipes for Visualizing Data. R Graphics Cookbook. O’Reilly 2012.

For data visualization using ggplot:

http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

http://www.cookbook-r.com/Graphs/Plotting_distributions_(ggplot2)

Thank you for your attention

Time for practice!