Lecture 3

Design Plots in Data Visualiztion

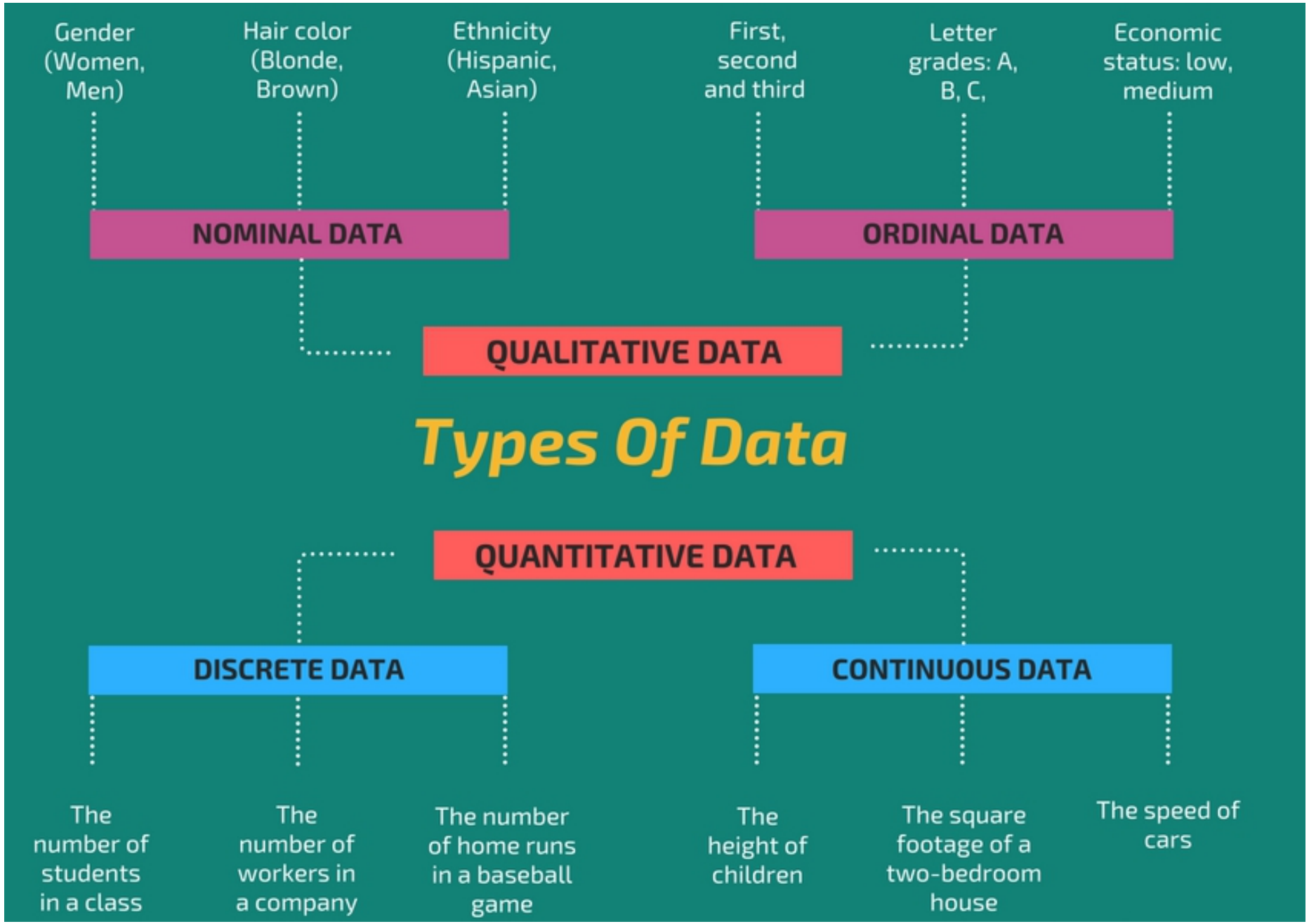
Phuc Loi Luu, PhD

luu.p.loi@googlemail.com

p.luu@garvan.org.au

Content

- Data Types
- Plot Dimension
- Static, Interactive vs Amination
- Plot Types
- Check
- Design a plot and panel/daskboard


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars

Data

Figure 4.1 The process of decoding data into information

```
> head(UScrime)
    M So  Ed Po1 Po2  LF  M.F Pop  NW  U1 U2 GDP Ineq     Prob    Time    y
1 151  1  91  58  56 510  950  33 301 108 41 394  261 0.084602 26.2011  791
2 143  0 113 103  95 583 1012  13 102  96 36 557  194 0.029599 25.2999 1635
3 142  1  89  45  44 533  969  18 219  94 33 318  250 0.083401 24.3006  578
4 136  0 121 149 141 577  994 157  80 102 39 673  167 0.015801 29.9012 1969
5 141  0 121 109 101 591  985  18  30  91 20 578  174 0.041399 21.2998 1234
6 121  0 110 118 115 547  964  25  44  84 29 689  126 0.034201 20.9995  682
```

```
> head(housing)
     Sat   Infl  Type Cont Freq
1    Low    Low Tower  Low   21
2 Medium    Low Tower  Low   21
3   High    Low Tower  Low   28
4    Low Medium Tower  Low   34
5 Medium Medium Tower  Low   22
6   High Medium Tower  Low   36
```

Check 6: MASS package with biopsy

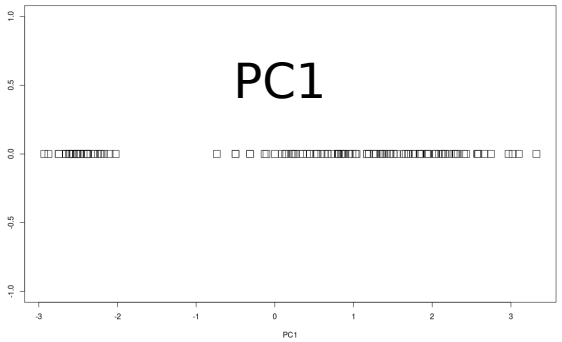
This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. He assessed biopsies of breast tumours for 699 patients up to 15 July 1992; each of nine attributes has been scored on a scale of 1 to 10, and the outcome is also known. There are 699 rows and 11 columns.

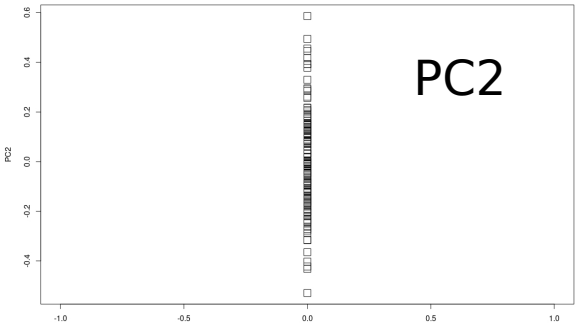
```
> head(biopsy)
        ID V1 V2 V3 V4 V5 V6 V7 V8 V9     class
1 1000025  5  1  1  1  2  1  3  1  1    benign
2 1002945  5  4  4  5  7 10  3  2  1    benign
3 1015425  3  1  1  1  2  2  3  1  1    benign
4 1016277  6  8  8  1  3  4  3  7  1    benign
5 1017023  4  1  1  3  2  1  3  1  1    benign
6 1017122  8 10 10  8  7 10  9  7  1 malignant
```

```
This data frame contains the following columns:

‘ID’ sample code number (not unique).

‘V1’ clump thickness.

‘V2’ uniformity of cell size.

‘V3’ uniformity of cell shape.

‘V4’ marginal adhesion.

‘V5’ single epithelial cell size.

‘V6’ bare nuclei (16 values are missing).

‘V7’ bland chromatin.

‘V8’ normal nucleoli.

‘V9’ mitoses.

‘class’ ‘"benign"’ or ‘"malignant"’.
```

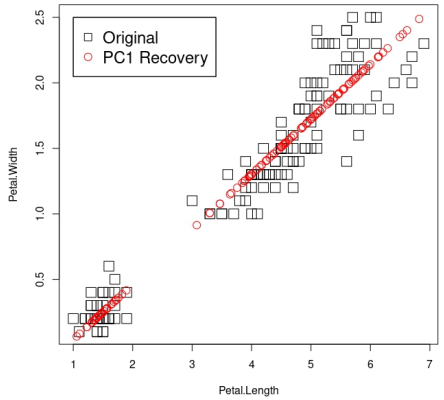
Plot and Data Dimension

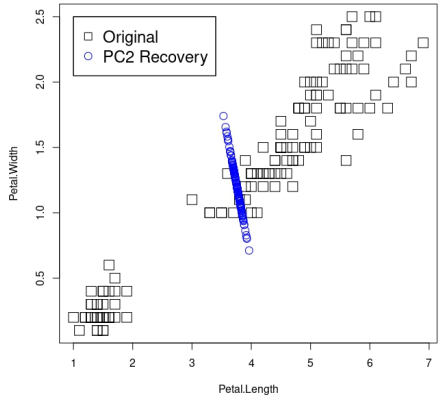
1D



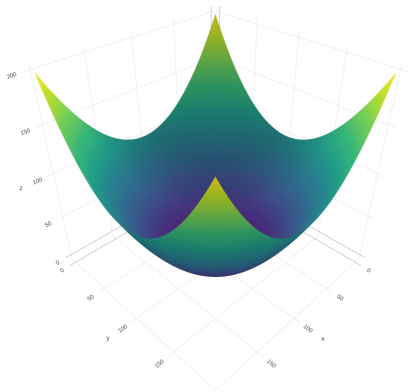


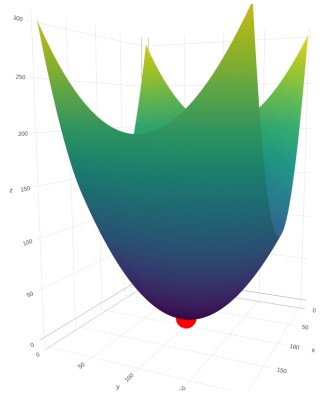
2D



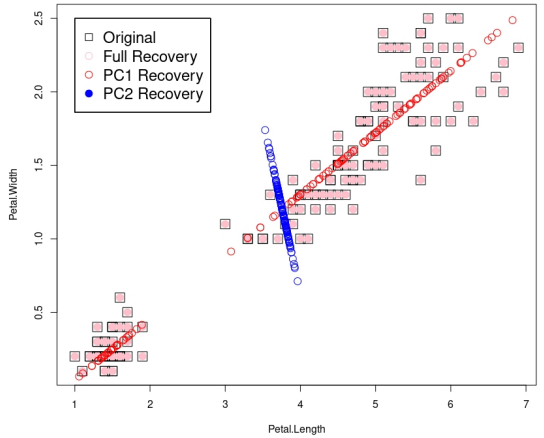


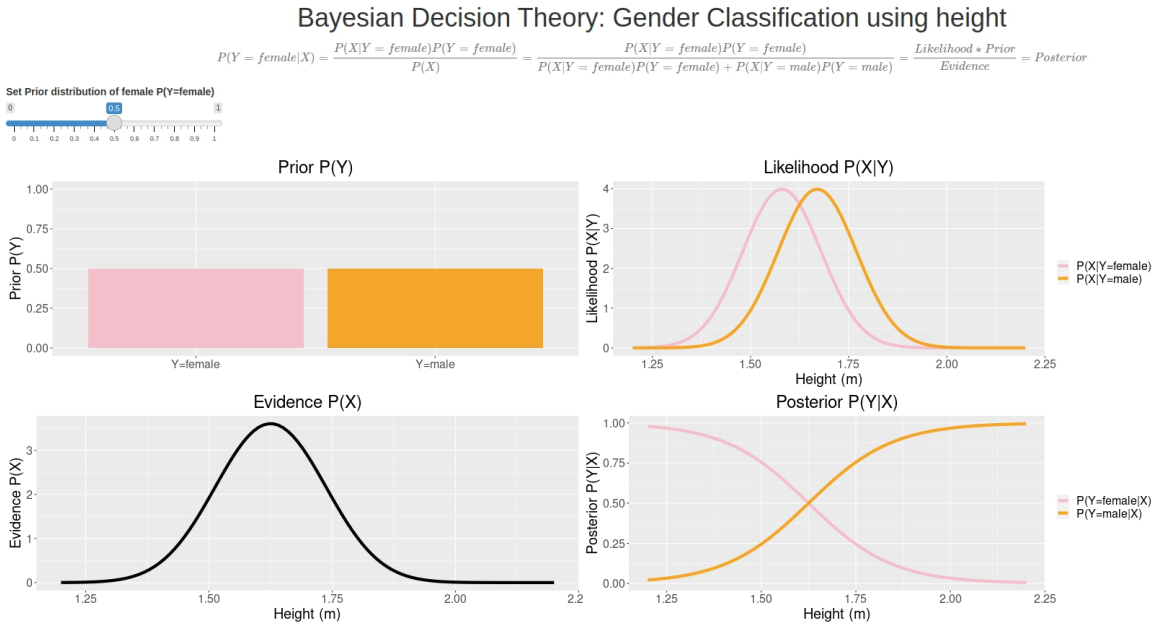
3D

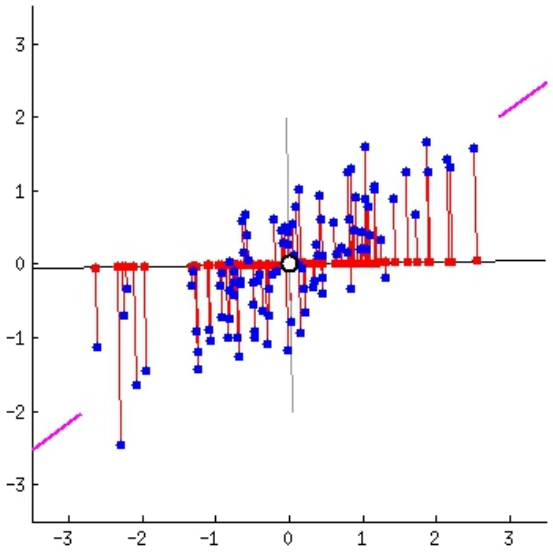




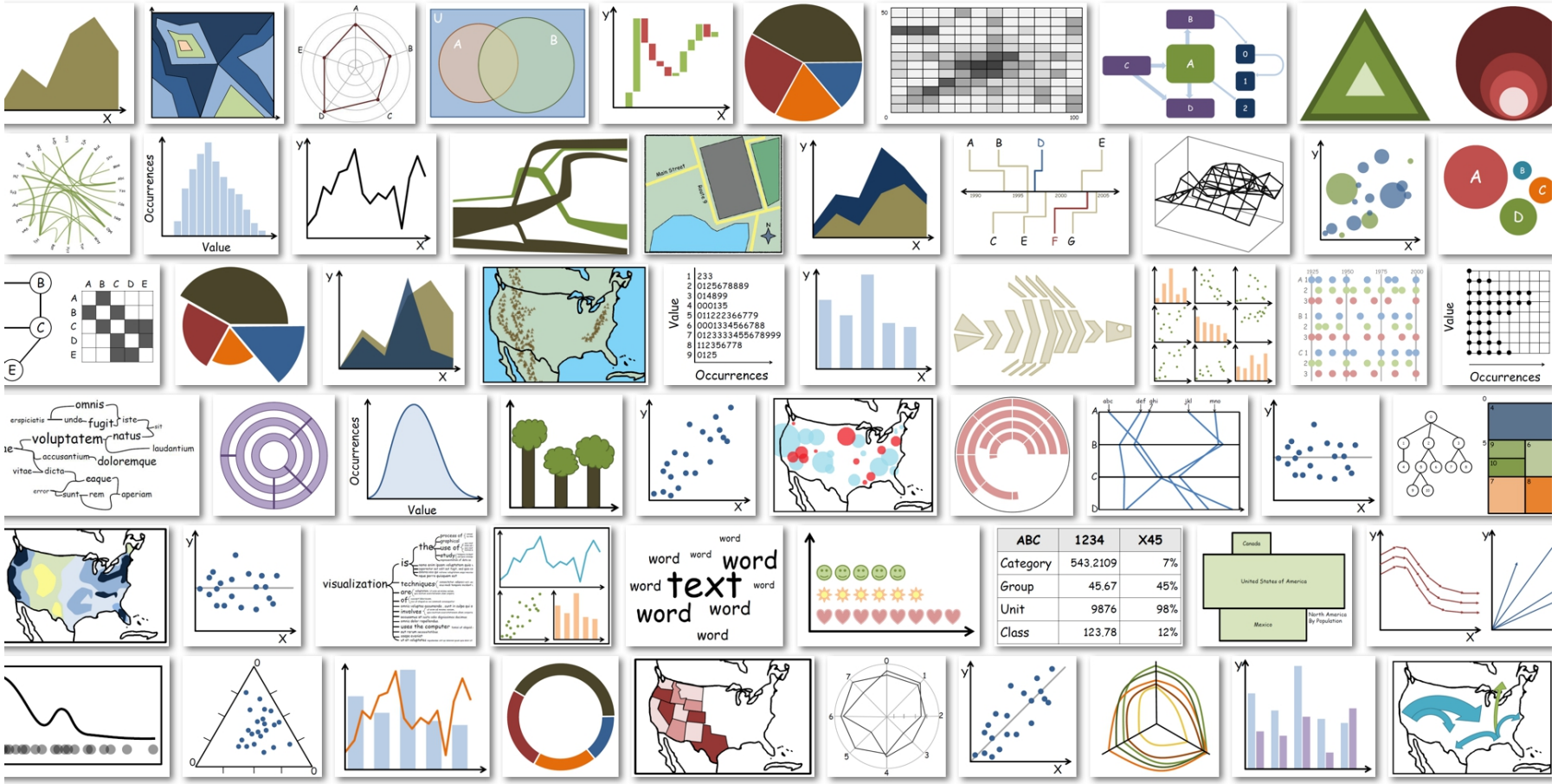
Static, Interactive vs Amination







Plot Types


Occurrences
Value
omnis
fugit
iste
natus
voluptatem
laudantium
accusantium
doloremque
eaque
aperiam
visualization
word
text
ABC
1234
X45
Category
543.2109
7%
Group
45.67
45%
Unit
9876
98%
Class
123.78
12%

Guided Visualizations for Charts and Graphs

What would you like to show?

Relationship

- Two Variables
  - Scatter Chart
  - Bubble Chart
- Three Variables

Distribution

- Single Variables
  - Few Data Points → Column Histogram
  - Many Data Points → Line Histogram
- Two Variables → Scatter Chart
- Three Variables → 3D Area Chart

Comparison

- Among Items
  - Two Variables per Item → Variable Width Column Chart
  - Many Categories → Table or Table with Embedded Charts
  - One Variable per Item
    - Few Categories
      - Many Items → Bar Chart
      - Few Items → Column Chart
- Over Time
  - Many Periods
    - Cyclical Data → Circular Area Chart
    - Non-Cyclical Data → Line Chart
  - Few Periods
    - Single or Few Categories → Column Chart
    - Many Categories → Line Chart

Composition

- Changing Over Time
  - Few Periods
    - Only Relative Differences Matter → Stacked 100% Column Chart
    - Relative and Absolute Differences Matter → Stacked Column Chart
  - Many Periods
    - Only Relative Differences Matter → Stacked 100% Area Chart
    - Relative and Absolute Differences Matter → Stacked Area Chart
- Static
  - Simple Share of Total → Pie Chart
  - Accumulation or Subtraction to Total → Waterfall Chart
  - Components of Components → Stacked 100% Column Chart with Subcomponents

Choose plot type

Table 3.1 The general data insight that corresponds to each data classification

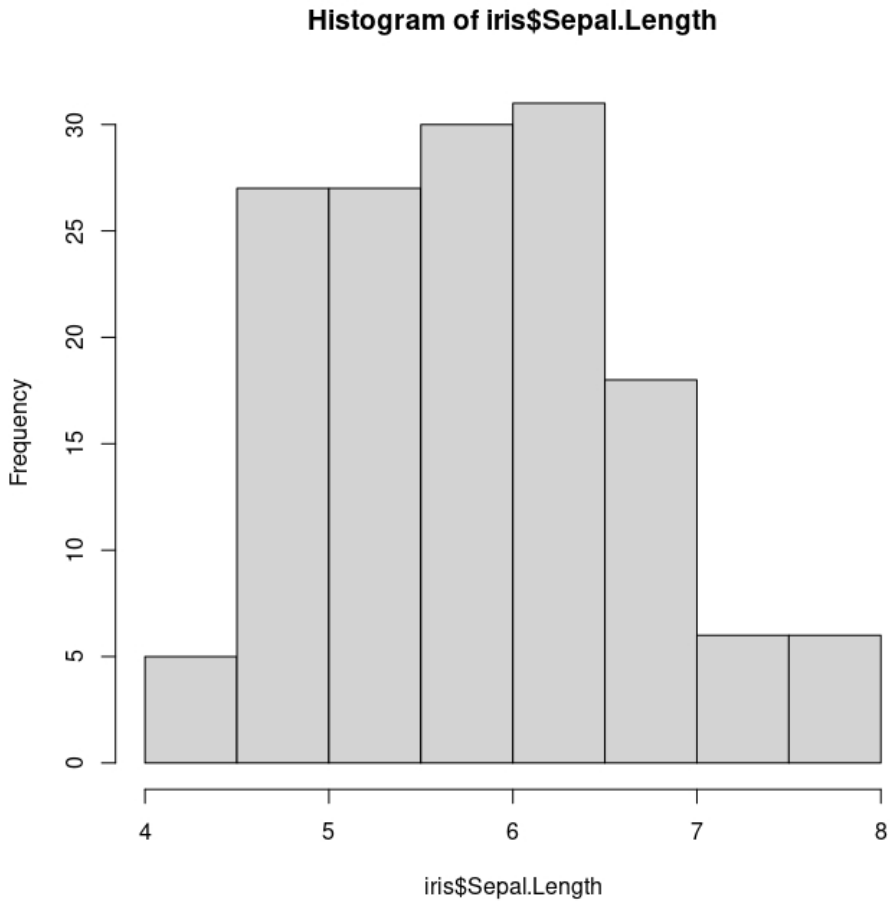
Data	Example	Insight	Chart type
Categorical	Non-numeric data such as types of movies, books, or authors.	Comparisons, proportions	Vertical bar, column bar, horizontal bar, and bullet charts Pie, stacked bar, stacked 100% bar, stacked area, stacked 100% area, and a tree map
Univariate	One numeric variable, such as book price	Distributions, proportions, frequencies	Histogram, density plot, and a boxplot
Geospatial	Specific locations marked by the latitude and longitude, regions coded by zip code, city, state, country, or county boundaries	Locations, comparisons, trends	Choropleth filled-map, bubble map, point map, connection map, and isopleth map
Multivariate	Two or more numeric variables, for example, weight, height, and IQ	Relationships, proportions, comparisons	Scatterplot, scatterplot matrix, bubble, parallel coordinates, radar, bullet, and a heat map.
Time series	Years, months, days, hours, minutes, seconds, or date	Trends, comparisons, cycles	Line chart, sparkline, area, stream graph, as well as bubble, stacked-area, and vertical bar charts.
Text	Single words or phrases, such as keywords from restaurant reviews on Yelp	Sentiment, comparisons, frequency	Word cloud, proportional area chart using size bubbles or squares, histogram, and bar chart
Edge lists or adjacency matrices	Who contacts whom or who knows whom in a network	Connections, relationships, tie strength, centrality, interactions	Undirected network diagram and directed network diagram

Check 1

```
> head(iris$Sepal.Length)
[1] 5.1 4.9 4.7 4.6 5.0 5.4
```

hist(iris$Sepal.Length)

No	Type	For this example
1	Data	Cont.
2	Plot/data Dimension	2D/1D
3	Static, Interactive vs Amination	Static
4	Plot	Histogram

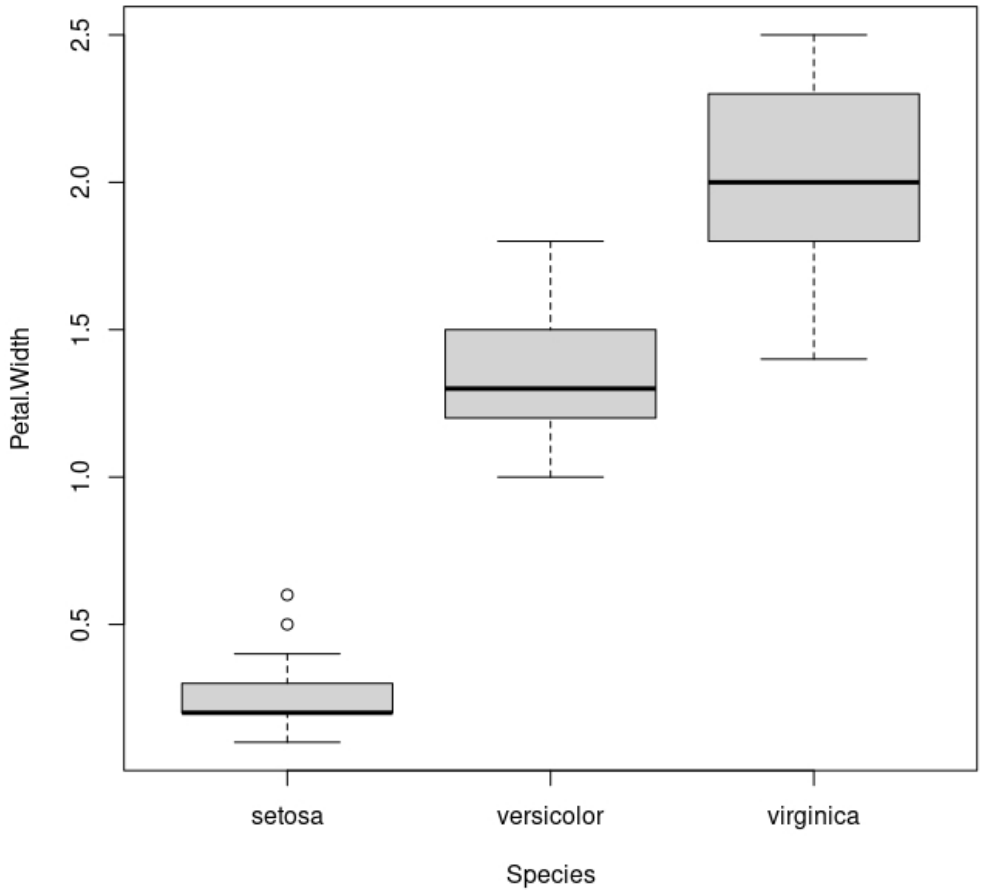


Check 2

```
> head(iris[,c("Petal.Width", "Species")])
  Petal.Width Species
1         0.2  setosa
2         0.2  setosa
3         0.2  setosa
4         0.2  setosa
5         0.2  setosa
6         0.4  setosa
```

No	Type	For this example
1	Data	
2	Plot/data Dimension	
3	Static, Interactive vs Amination	
4	Plot	

plot(Petal.Width~Species, data=iris)



Check 3

```
> head(iris[,c("Petal.Length", "Petal.Width")])
  Petal.Length Petal.Width
1          1.4         0.2
2          1.4         0.2
3          1.3         0.2
4          1.5         0.2
5          1.4         0.2
6          1.7         0.4
```

plot(???)

No	Type	For this example
1	Data	
2	Plot/data Dimension	
3	Static, Interactive vs Amination	
4	Plot	

Check 4

```
> head(iris[,c("Petal.Length", "Petal.Width", "Species")])
  Petal.Length Petal.Width Species
1          1.4         0.2  setosa
2          1.4         0.2  setosa
3          1.3         0.2  setosa
4          1.5         0.2  setosa
5          1.4         0.2  setosa
6          1.7         0.4  setosa
```

plot(???)

No	Type	For this example
1	Data	
2	Plot/data Dimension	
3	Static, Interactive vs Amination	
4	Plot	

Check 5

```
> head(iris[,c("Sepal.Length", "Petal.Length", "Petal.Width", "Species")])
  Sepal.Length Petal.Length Petal.Width Species
1          5.1          1.4         0.2  setosa
2          4.9          1.4         0.2  setosa
3          4.7          1.3         0.2  setosa
4          4.6          1.5         0.2  setosa
5          5.0          1.4         0.2  setosa
6          5.4          1.7         0.4  setosa
```

plot(???)

No	Type	For this example
1	Data	
2	Plot/data Dimension	
3	Static, Interactive vs Amination	
4	Plot	

Check 6: MASS package with biopsy

This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. He assessed biopsies of breast tumours for 699 patients up to 15 July 1992; each of nine attributes has been scored on a scale of 1 to 10, and the outcome is also known. There are 699 rows and 11 columns.

```
> head(biopsy)
        ID V1 V2 V3 V4 V5 V6 V7 V8 V9     class
1 1000025  5  1  1  1  2  1  3  1  1    benign
2 1002945  5  4  4  5  7 10  3  2  1    benign
3 1015425  3  1  1  1  2  2  3  1  1    benign
4 1016277  6  8  8  1  3  4  3  7  1    benign
5 1017023  4  1  1  3  2  1  3  1  1    benign
6 1017122  8 10 10  8  7 10  9  7  1 malignant
```

```
This data frame contains the following columns:

‘ID’ sample code number (not unique).

‘V1’ clump thickness.

‘V2’ uniformity of cell size.

‘V3’ uniformity of cell shape.

‘V4’ marginal adhesion.

‘V5’ single epithelial cell size.

‘V6’ bare nuclei (16 values are missing).

‘V7’ bland chromatin.

‘V8’ normal nucleoli.

‘V9’ mitoses.

‘class’ ‘"benign"’ or ‘"malignant"’.
```

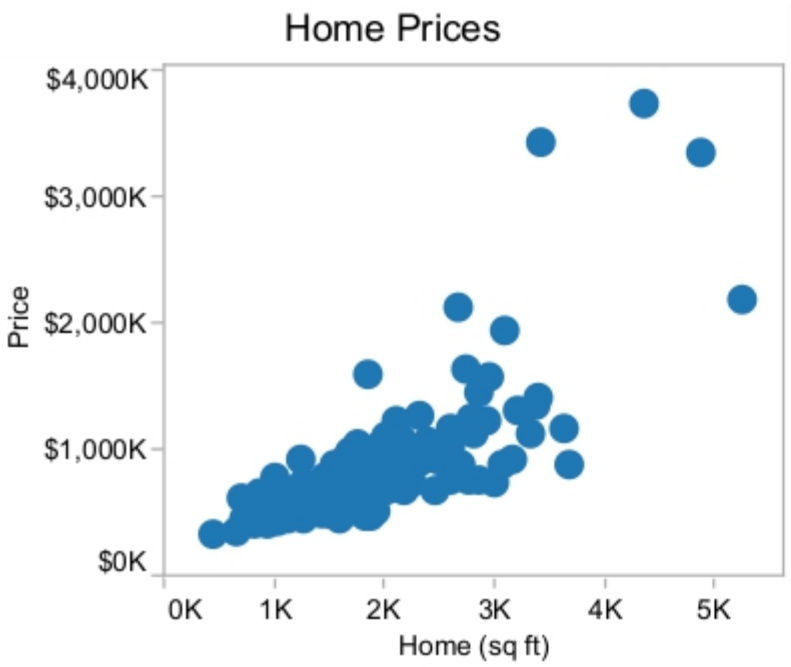
Check 7: MASS package with housing

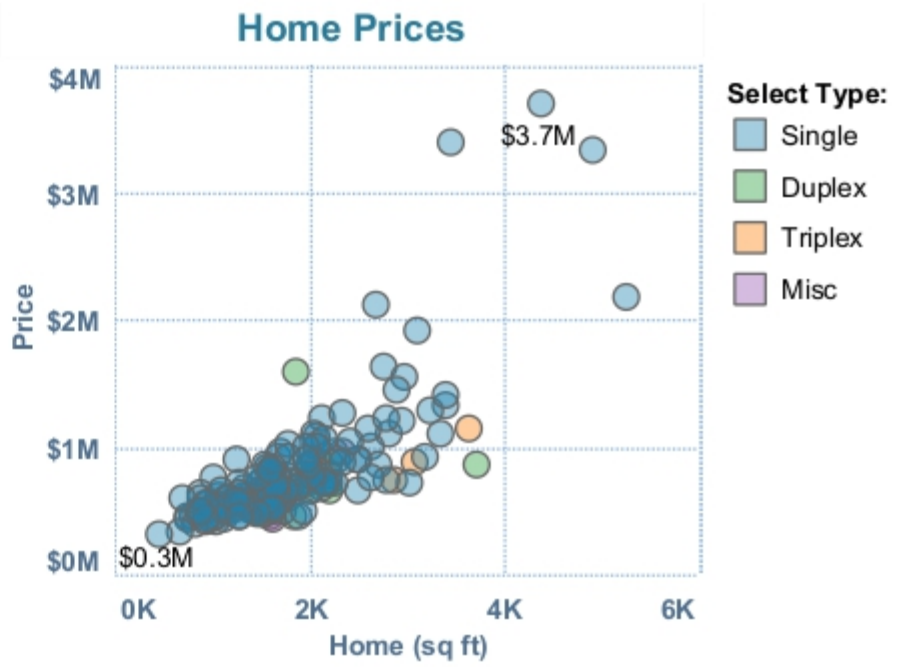
```
> head(housing)
     Sat   Infl  Type Cont Freq
1    Low    Low Tower  Low   21
2 Medium    Low Tower  Low   21
3   High    Low Tower  Low   28
4    Low Medium Tower  Low   34
5 Medium Medium Tower  Low   22
6   High Medium Tower  Low   36
```

Design a plot

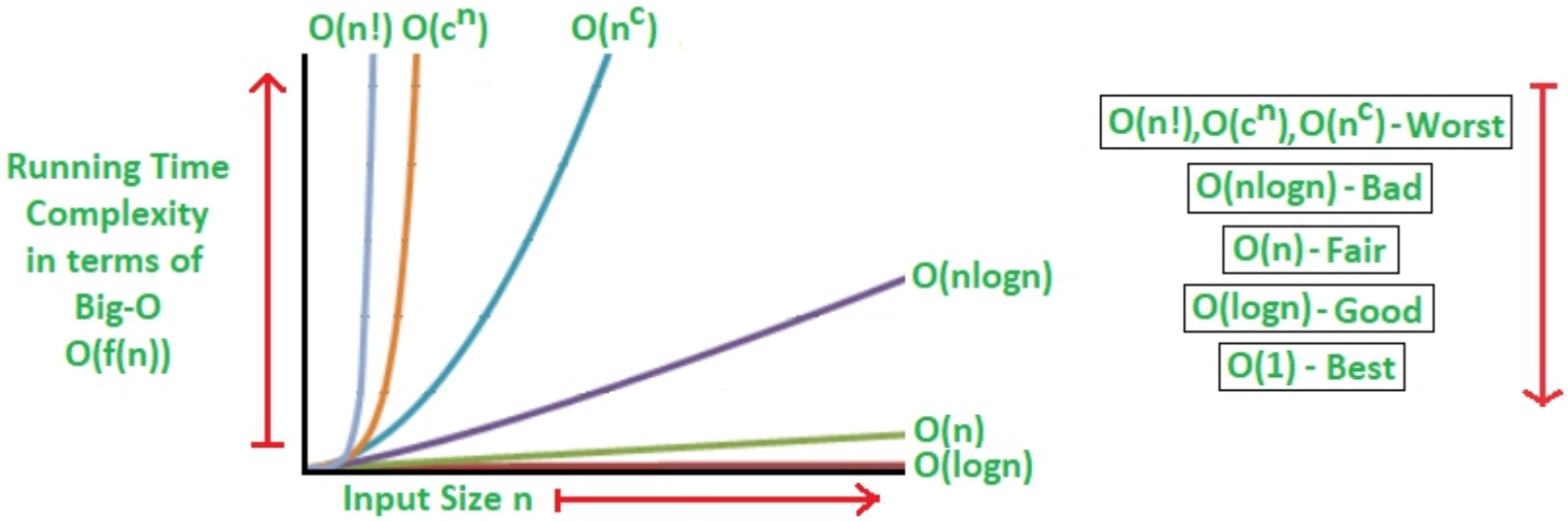
1) Axes
2) Titles
3) Legends
4) Backgrounds & Grid Lines
5) Margins
6) Multi-Panel Plots
7) Colors
8) Themes
9) Lines
10) Text
11) Coordinates
12) Chart Types
13) Ribbons (AUC, CI, etc.)
14) Smoothings

Design a plot

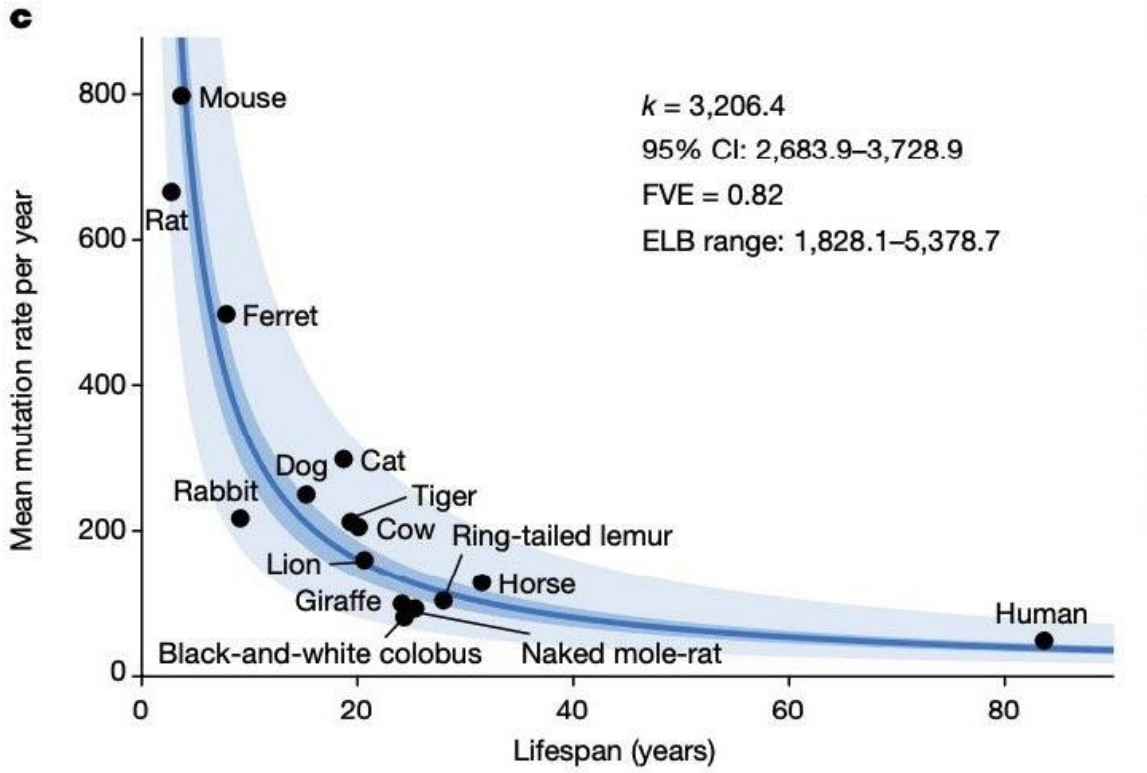


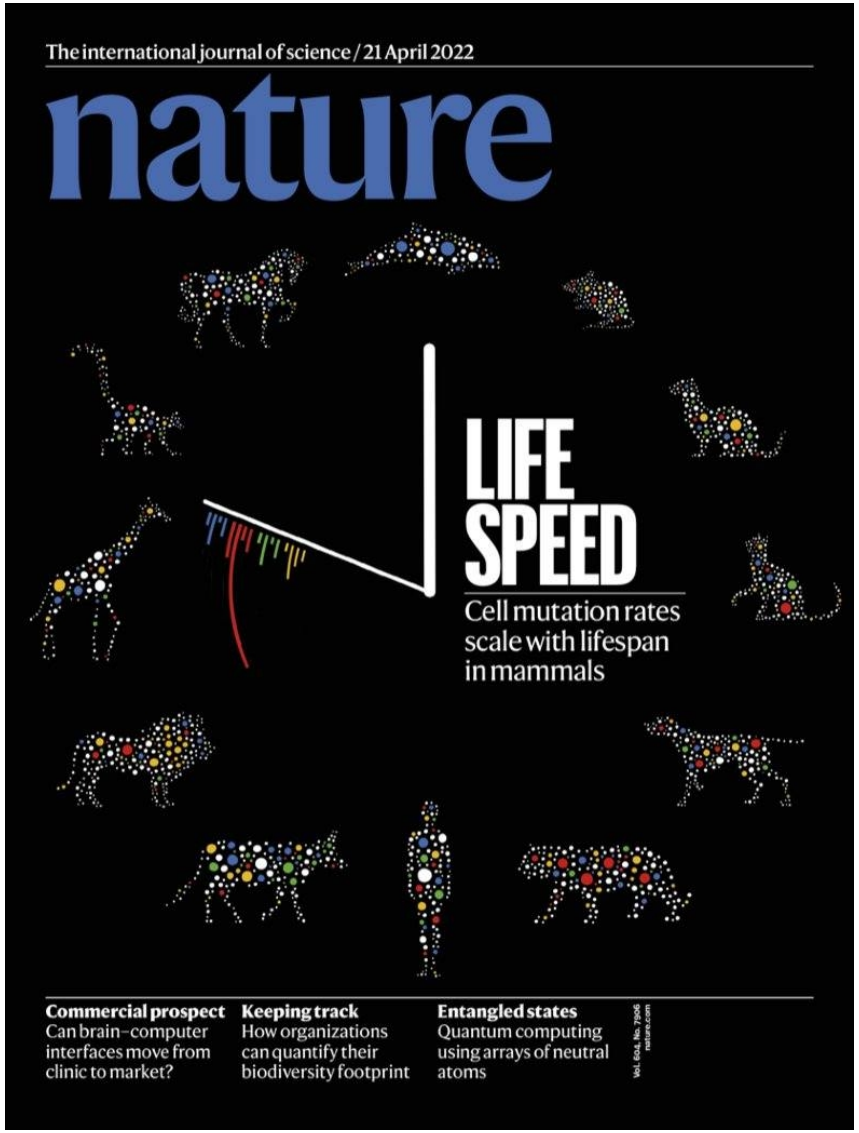


Big-O


O(n!) O(c^n)
O(n^c)
Running Time Complexity in terms of Big-O O(f(n))
O(nlogn)
O(n)
O(logn)
Input Size n
O(n!),O(c^n),O(n^c) - Worst
O(nlogn) - Bad
O(n) - Fair
O(logn) - Good
O(1) - Best

Mean mutation rate





https://www.nature.com/articles/nature12213

Tumor burden in cancer

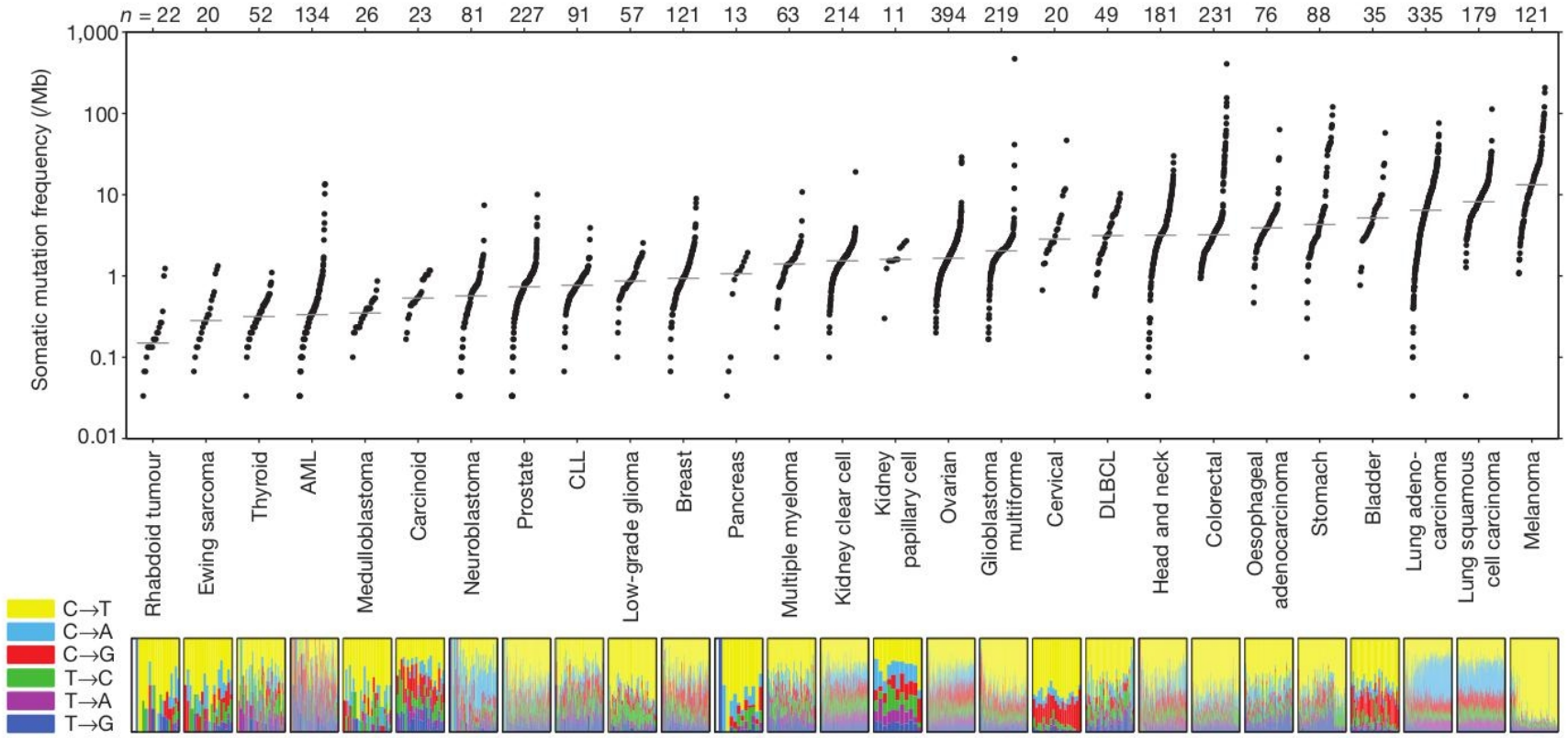
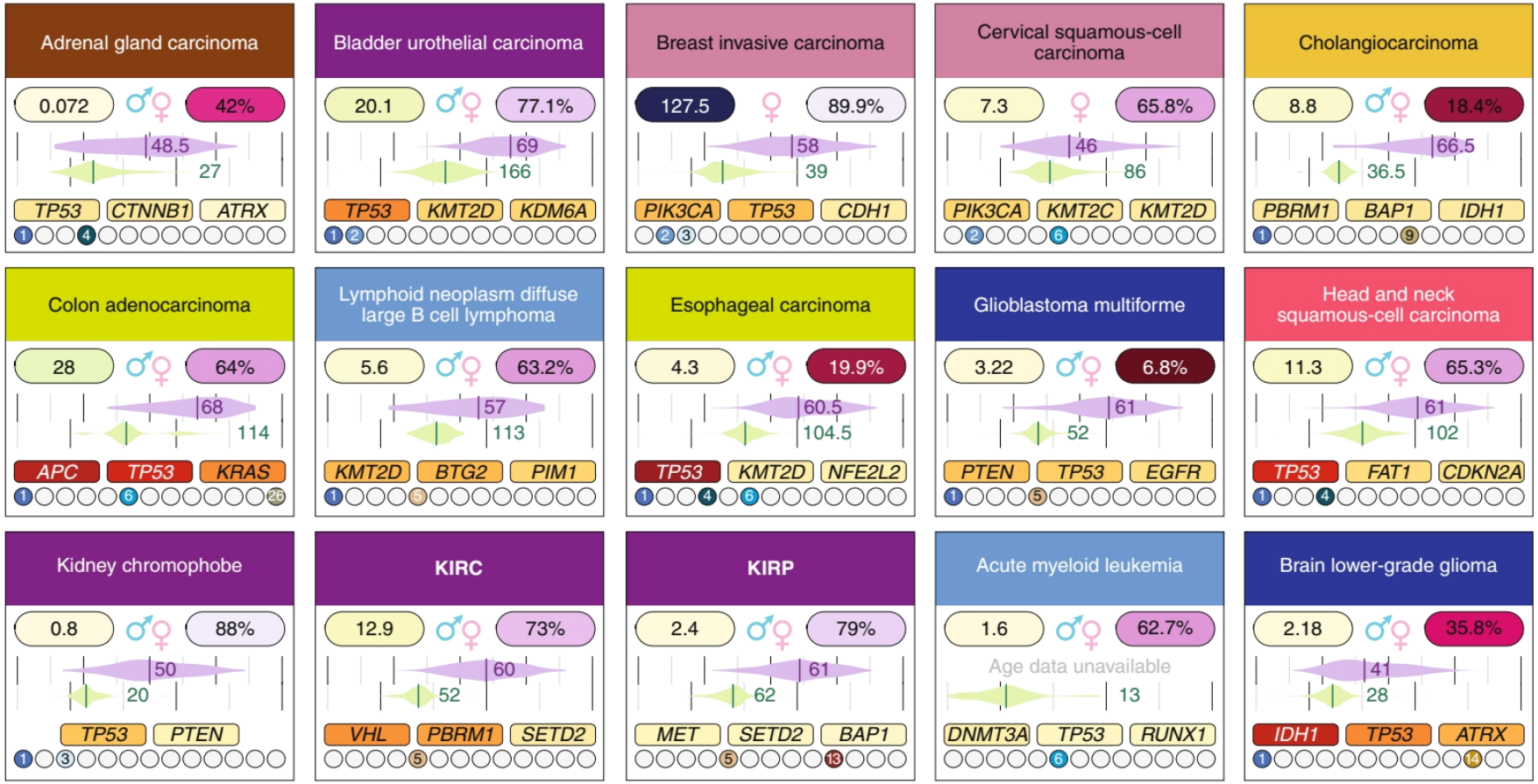


Figure 1 | Somatic mutation frequencies observed in exomes from 3,083 tumour–normal pairs. Each dot corresponds to a tumour–normal pair, with vertical position indicating the total frequency of somatic mutations in the exome. Tumour types are ordered by their median somatic mutation frequency, with the lowest frequencies (left) found in haematological and paediatric tumours, and the highest (right) in tumours induced by carcinogens such as tobacco smoke and ultraviolet light. Mutation frequencies vary more than 1,000-fold between lowest and highest across different cancers and also within several tumour types. The bottom panel shows the relative proportions of the six different possible base-pair substitutions, as indicated in the legend on the left. See also Supplementary Table 2.

- Tumor gene mutation burden (TMB) is the approximate amount of gene mutation that occurs in the genome of a cancer cell.
- Tumors with high TMB are likely to generate more neoantigens, which can be displayed on HLA molecules on the surface of the cancer cells and be recognized by the immune system

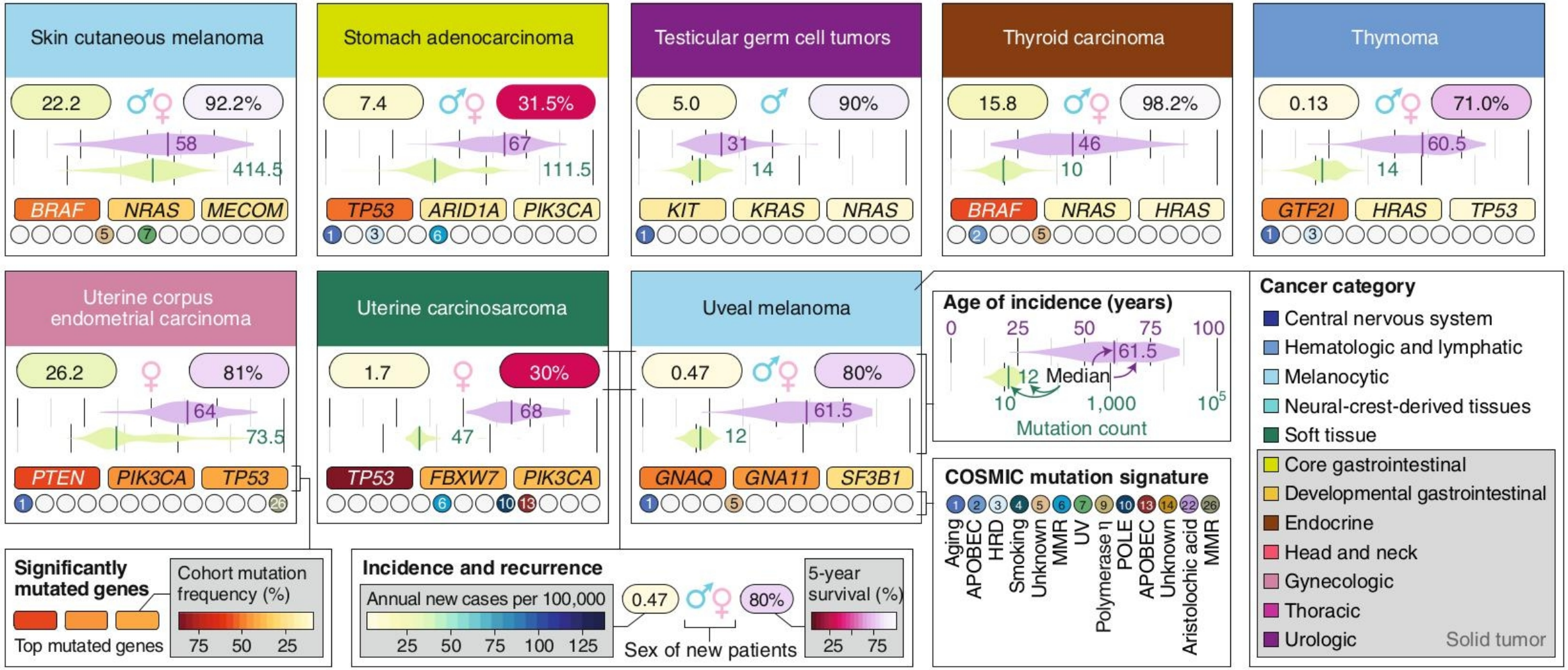
https://www.nature.com/articles/nature12213

Somatic mutations in cancer



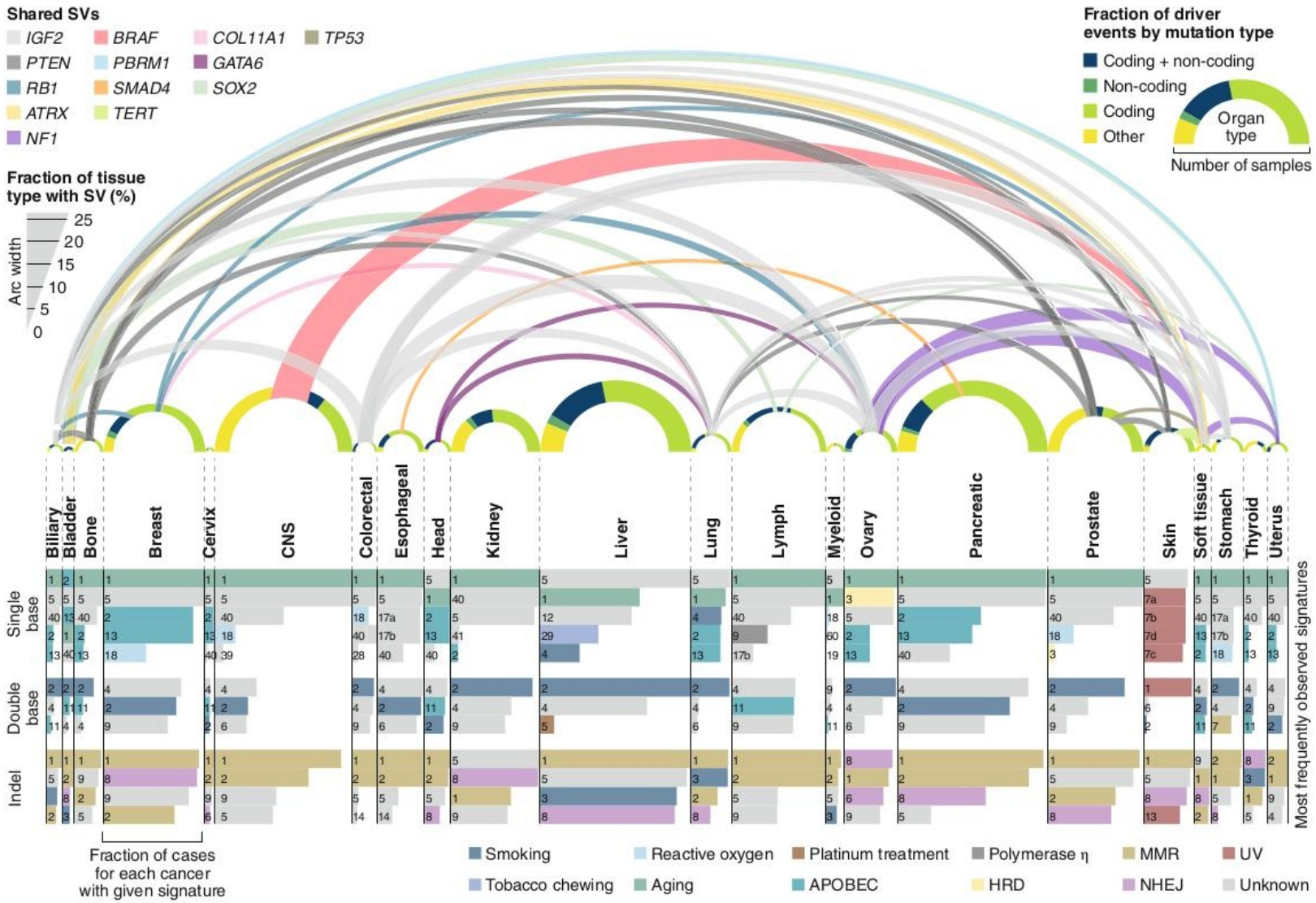
https://doi.org/10.1038/s43018-021-00250-4

Somatic mutations in cancer



https://doi.org/10.1038/s43018-021-00250-4

Somatic mutations in cancer



https://doi.org/10.1038/s43018-021-00250-4

Design a plot: font text and label

Fonts: *Can you read this?* **(Can you read this?)**

- Trebuchet MS or Verdana (especially for tables and numbers)
- Arial
- Georgia
- Tahoma
- Times New Roman
- Lucida sans

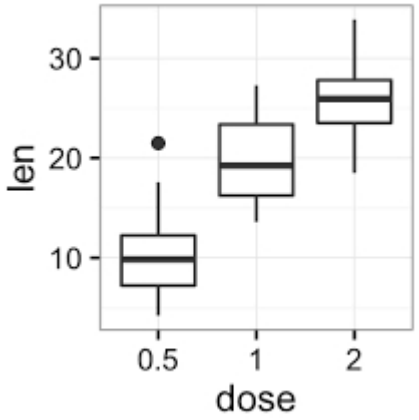
In addition, **Calibri** and **Cambria** are suitable for tooltips (see below), but are not recommended for use in any other part of a visualization.

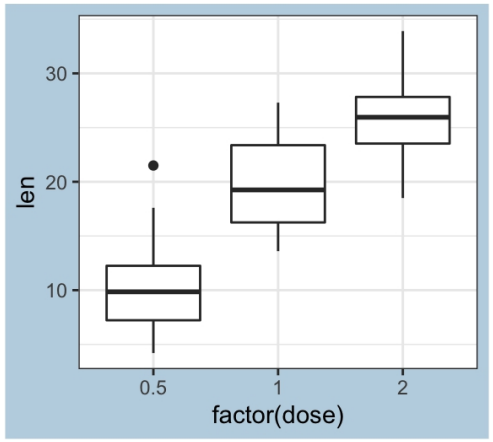
Design a plot: font text and label

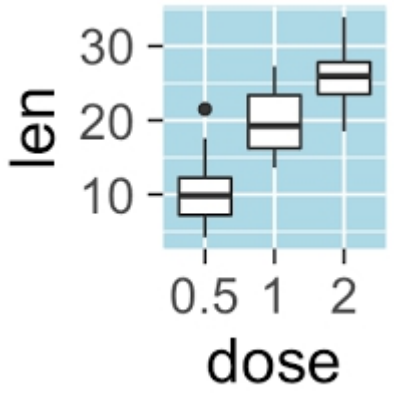
Lastly, never make a change in adjacent text that modifies more than one attribute of a font (such as size, boldness, color, or serif quality).

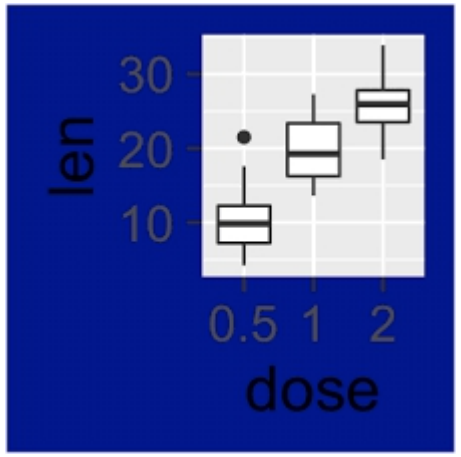
Good Change | ***Bad*** Change

Design a plot: Backgrounds & Grid Lines



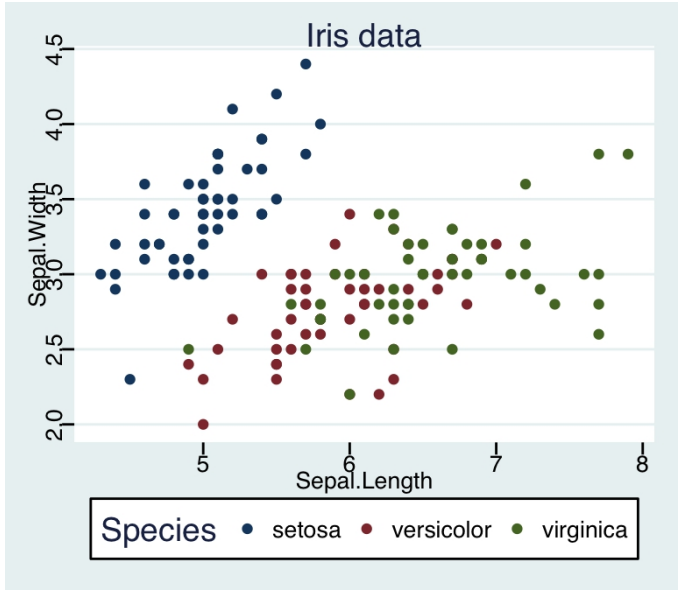


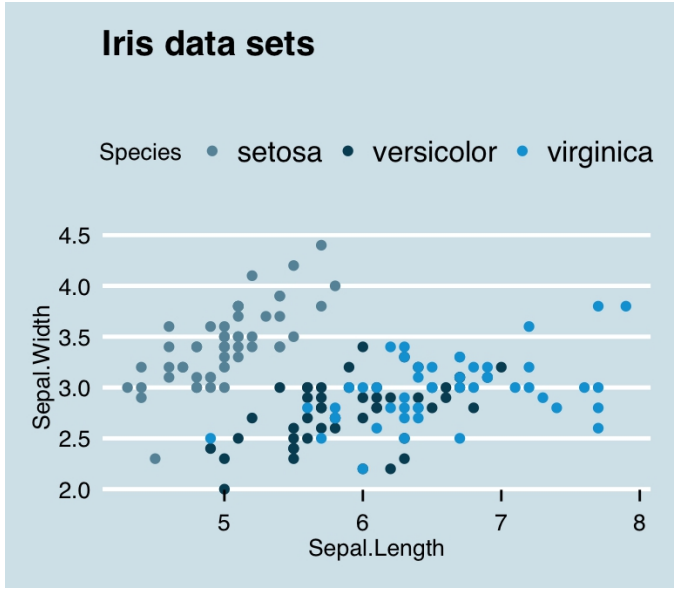


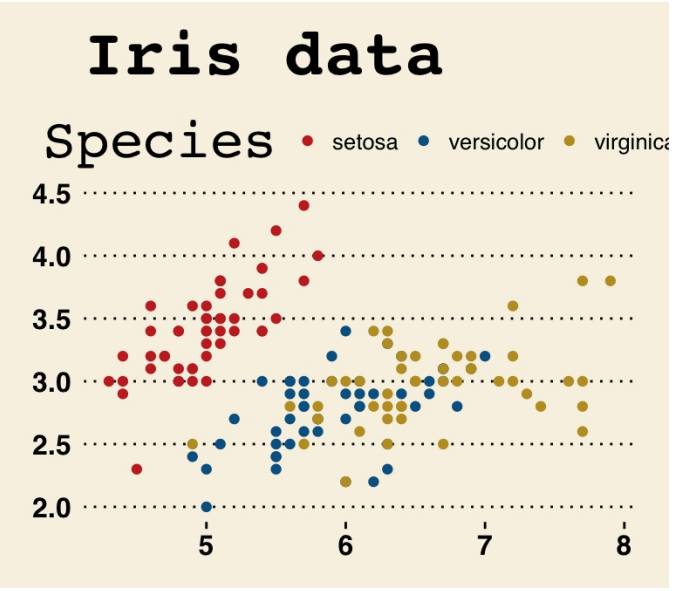


http://www.sthda.com/english/wiki/wiki.php?id_contents=7793

Design a plot: Backgrounds & Grid Lines

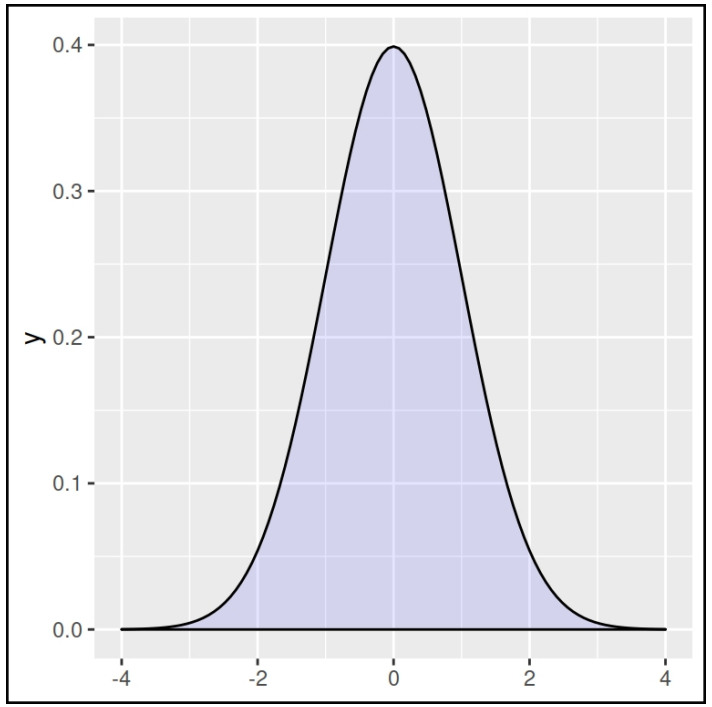


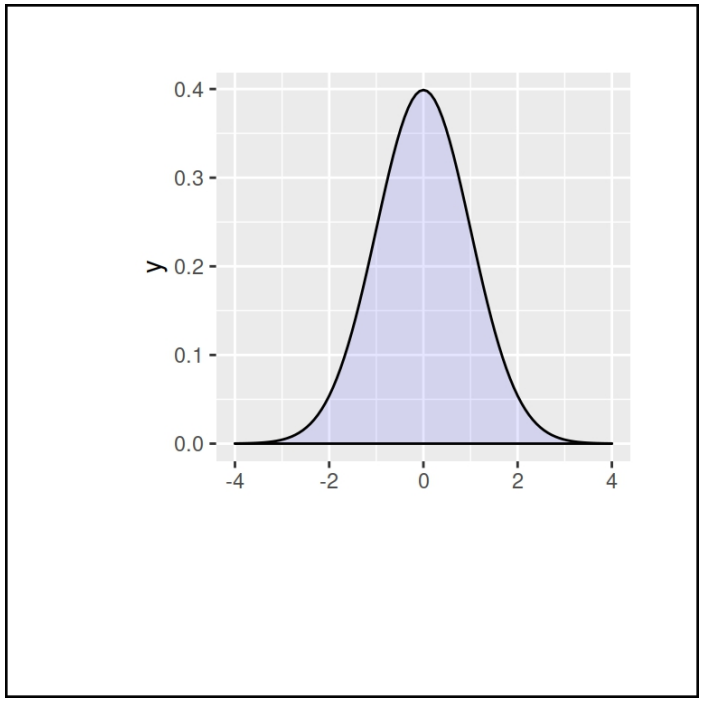




http://www.sthda.com/english/wiki/wiki.php?id_contents=7793

Design a plot: Margin

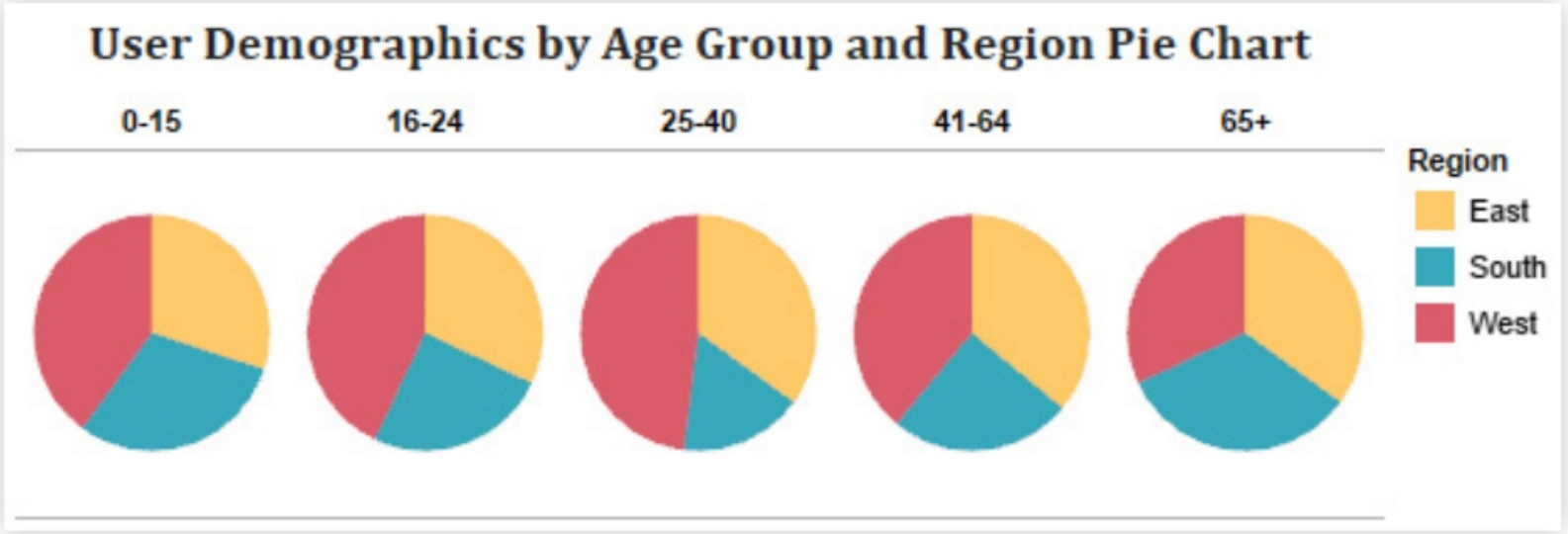




https://r-charts.com/ggplot2/margins/

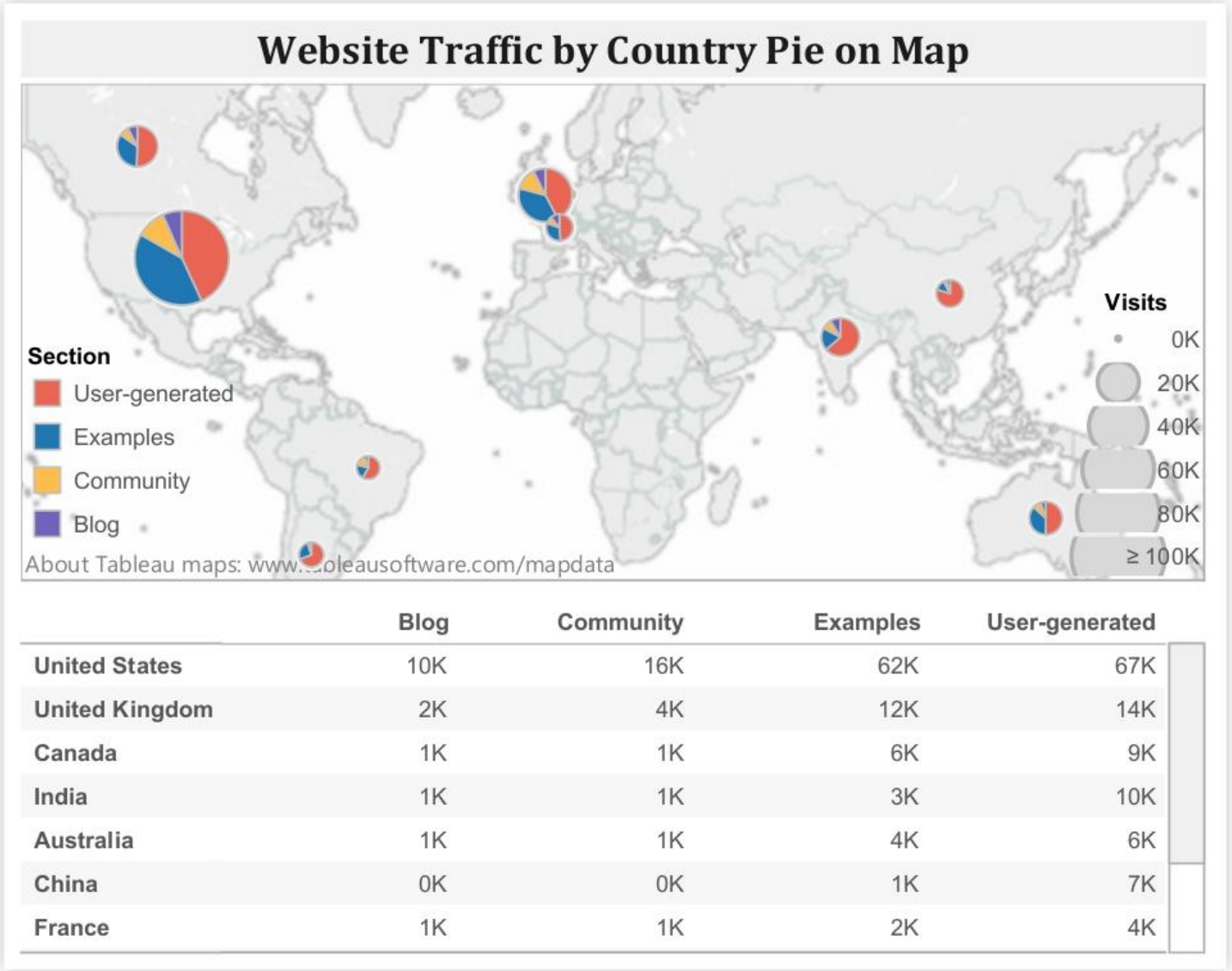
Design a plot: pie

Although pie charts are commonly used in part-to-whole analysis, we suggest avoiding them.



Design a plot: map

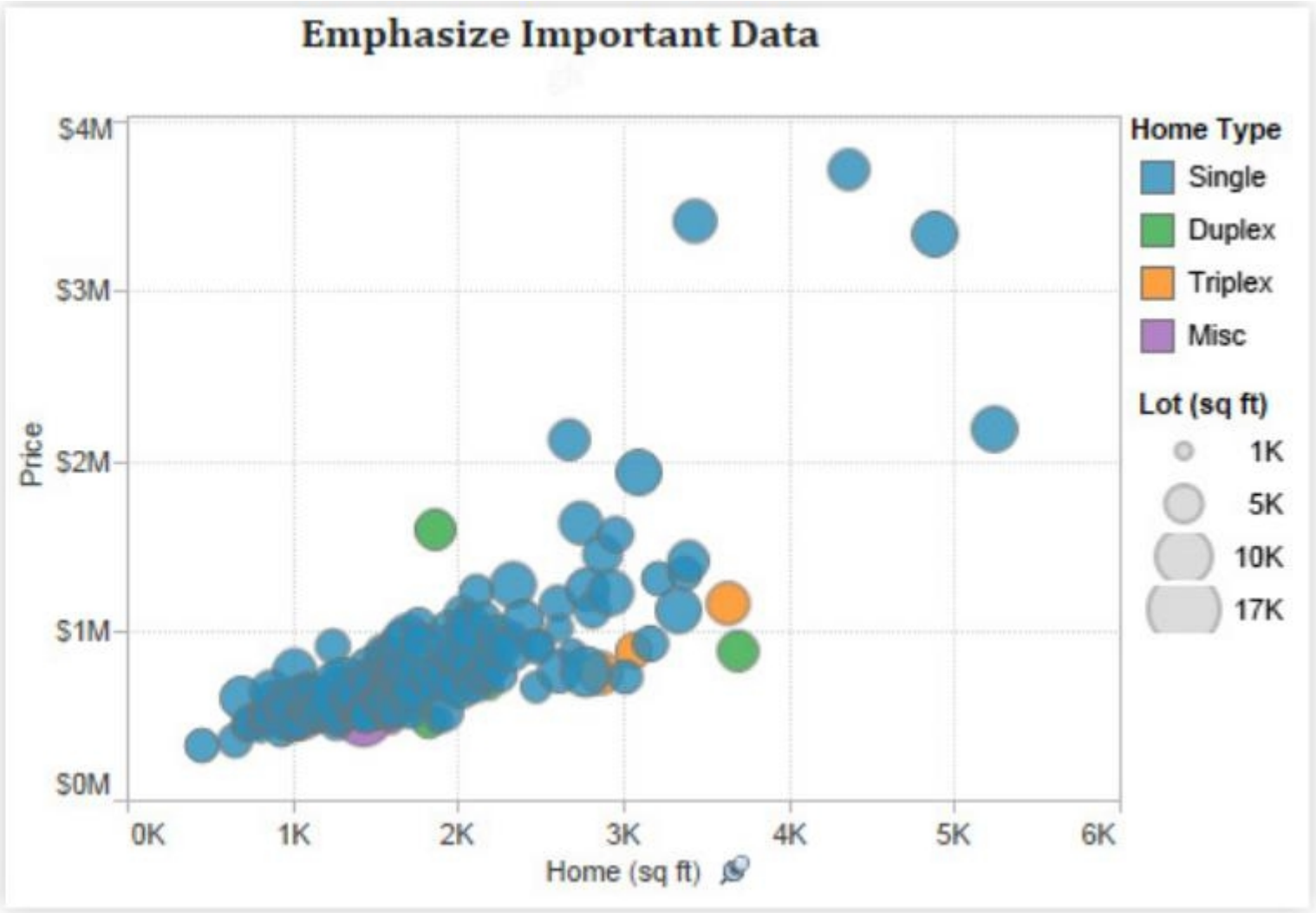
Maps are often best when paired with another chart that details what the map displays.



	Blog	Community	Examples	User-generated
United States	10K	16K	62K	67K
United Kingdom	2K	4K	12K	14K
Canada	1K	1K	6K	9K
India	1K	1K	3K	10K
Australia	1K	1K	4K	6K
China	0K	0K	1K	7K
France	1K	1K	2K	4K

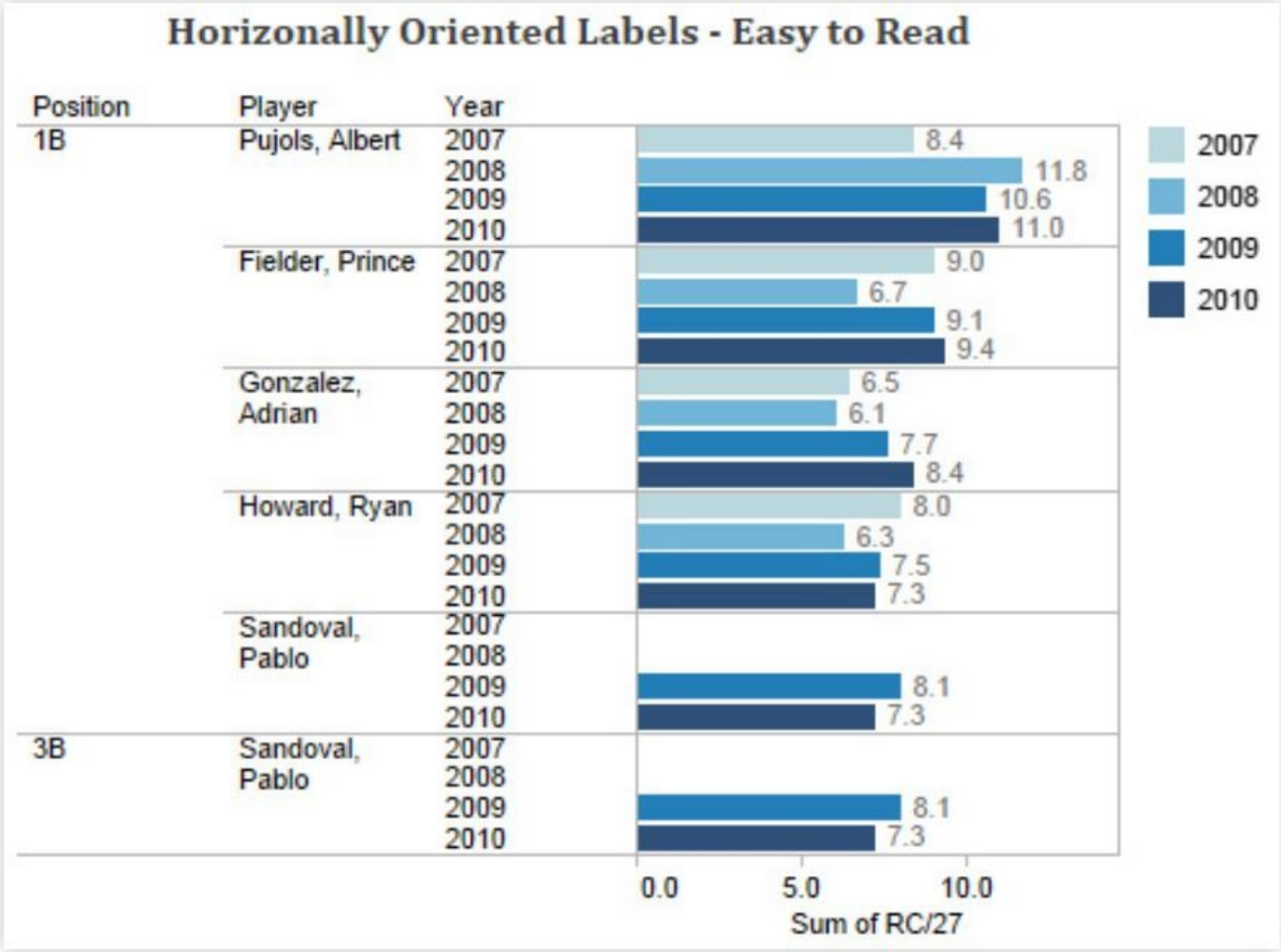
Design a plot: put #variables

A rule of thumb is to put the most important data on the X- or Y- axis and less important data on color, size, or shape.



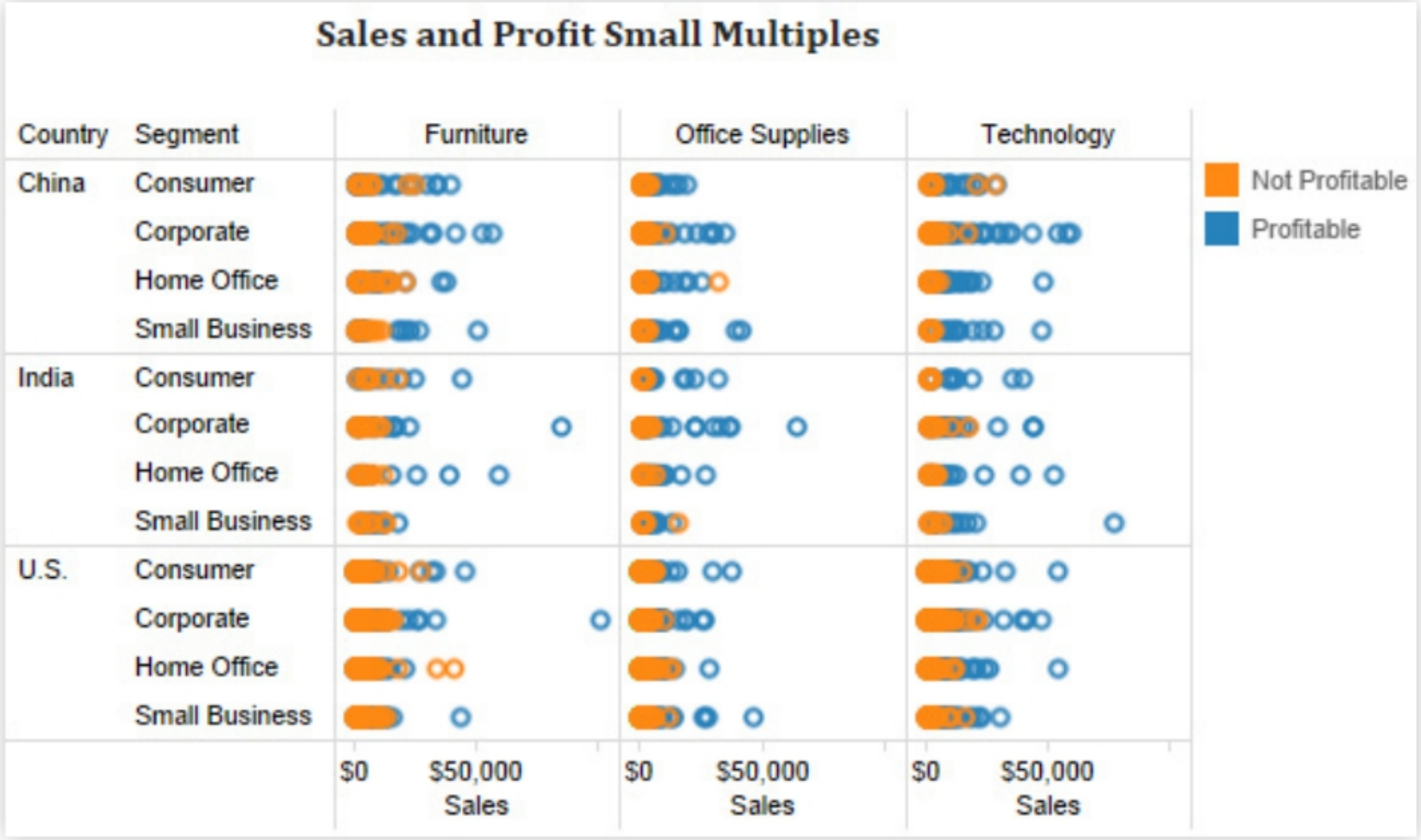
Design a plot: label

If you find yourself with a view that has long labels that only fit vertically, try rotating the view.



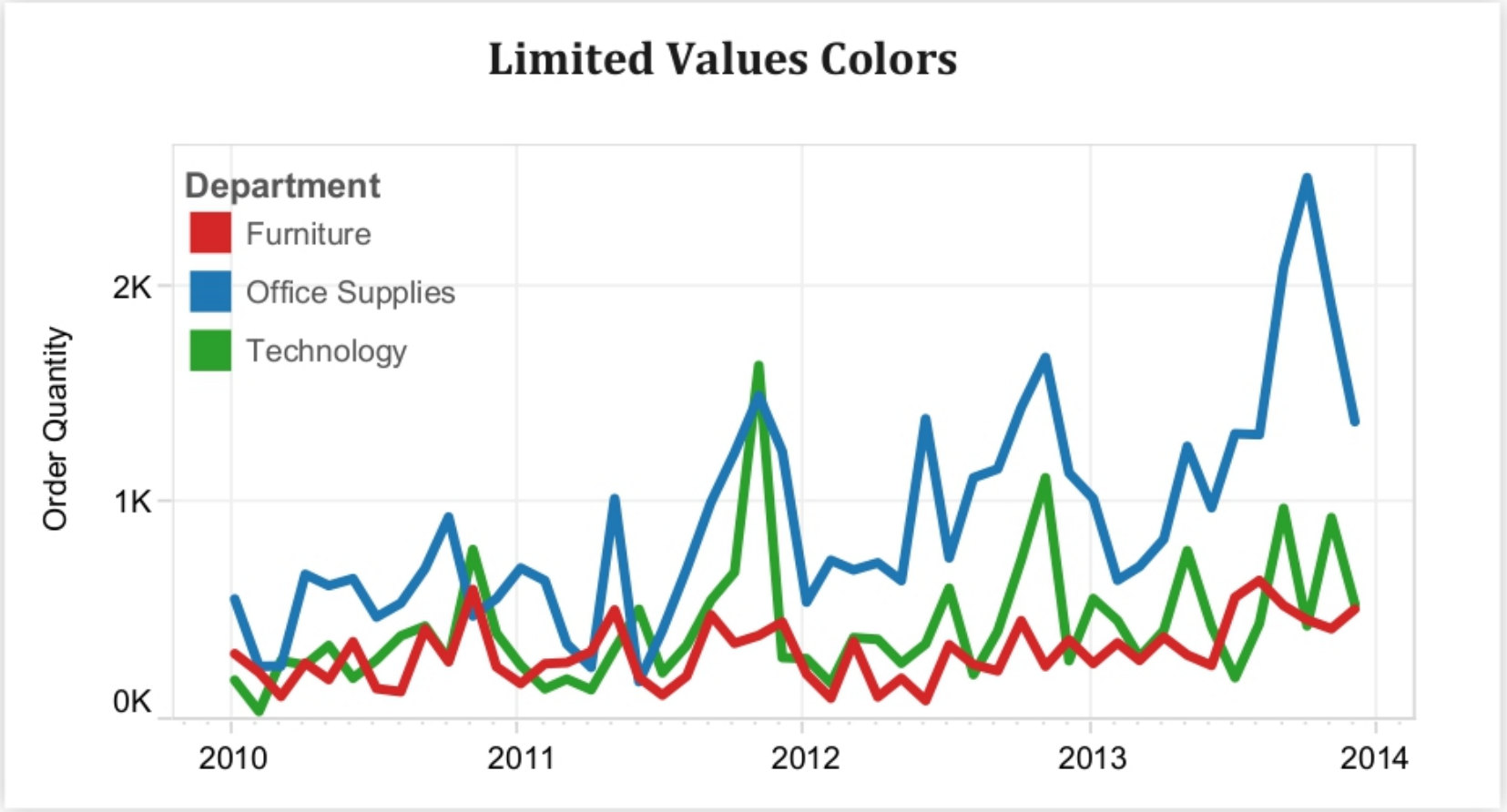
Design a plot: split data

Instead of stacking many measures and dimensions into one condensed view, break them down to small multiples.



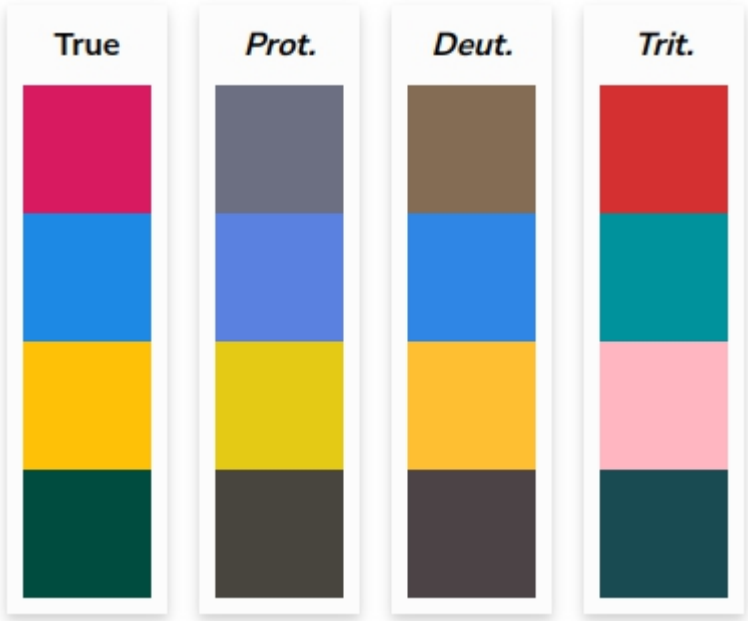
Design a plot: color and shape

Limit the number of colors and shapes in one view to 7-10 so that you can distinguish them and see important patterns.



Design a plot: we see colors differently

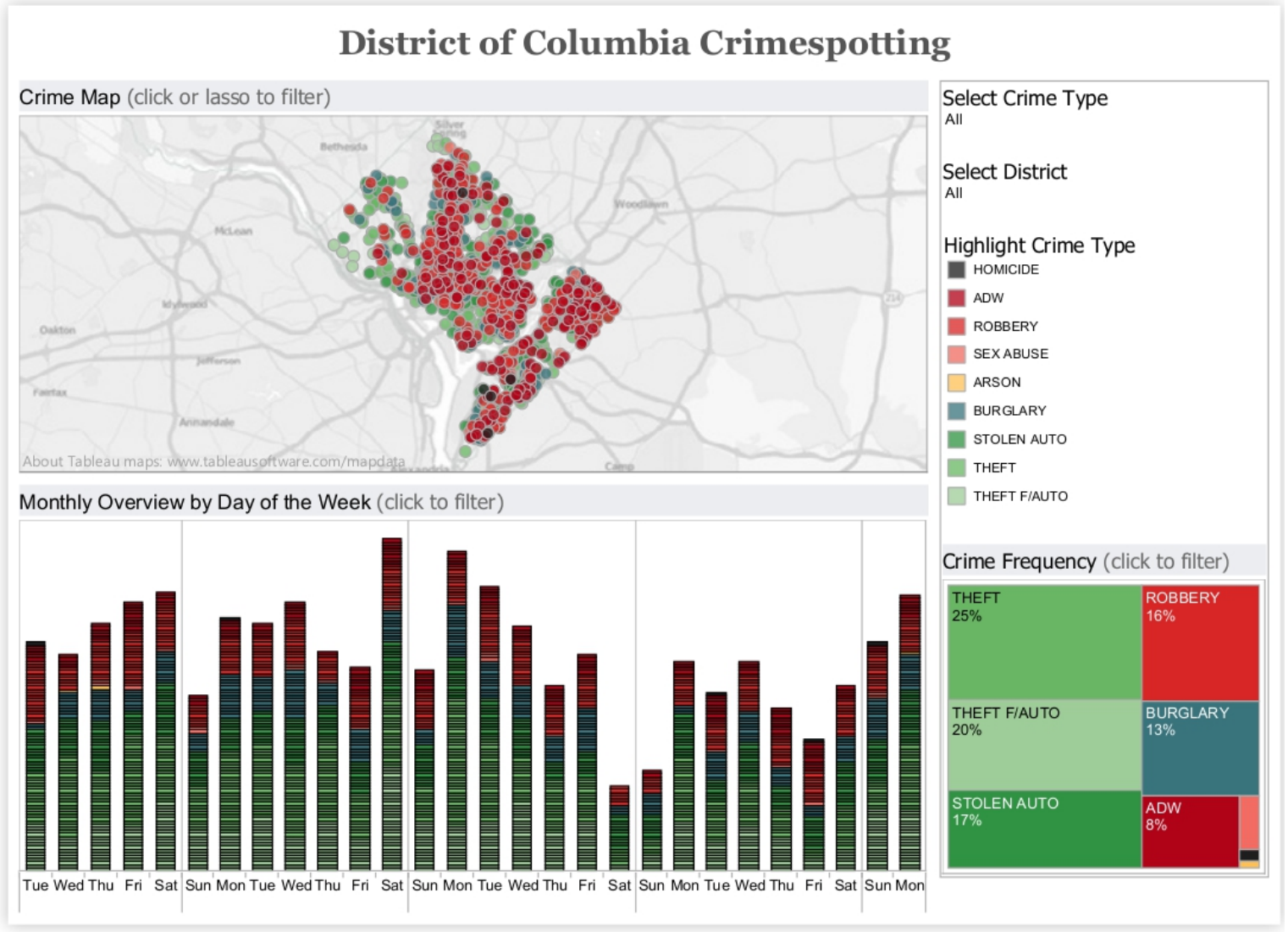
- About 1 in 20 people are colorblind in some way
- The colors in the leftmost column are the "true" colors; these are displayed in the remaining three columns the way that a person with protanopia, deuteranopia, or tritanopia would see them, respectively.
- Deuteranomaly affects 1 in every 16 or 17 men and 1 in every 250 women



https://davidmathlogic.com/colorblind/#%23D81B60-%231E88E5-%23FFC107-%23004D40

Design a dashboard

Use interactive views only when it is necessary: you need to guide a story, encourage user exploration, or there is too much detail to show all at once.



Check 1: What questions are you trying to answer?

- Does this visualization answer all of your questions?
- Is the purpose of the visualization clearly explained in its title or surrounding text?
- Can you understand the visualization in 30 seconds or less, without additional information?
- Does your visualization include a title? Is that title simple, informative, and eye-catching?
- Does your visualization include subtitles to guide your viewers?

Check 2: Do you have the right chart type for your analysis?

- What types of analysis are you performing?
- Have you selected the most suitable chart type(s) for your types of analysis?
- Have you considered alternative chart types that could work better than the ones you have chosen?

Check 3: Are your views effective?

- Are your most important data shown on the X- and Y-axes and your less important data encoded in color or shape attributes?
- Are your views oriented intuitively—do they cater to the way your viewers read and perceive data?
- Have you limited the number of measures or dimensions in a single view so that your users can see your data?
- Have you limited your usage of colors and shapes so that your users can distinguish them and see patterns?

Check 4: Is your dashboard holistic?

- Do all your views fit together to tell a single story?
- Do all your views flow well from one to the next? Are they in a good order?
- Do your most important views appear in the top or top-left corner?
- Are secondary elements in your dashboard placed well so they support the views without interrupting them?
- Are your filters in the right locations?
- Do your filters work correctly? Do views become blank or downright confusing if you apply a filter?
- Do your filters apply to the right scope?

- Are your filter titles informative? Can viewers easily understand how to interactive with your filters?
- Are your legends close to the views they apply to?
- Is your legend highlight button set to “on” or “off” according to your preference?
- Do you have filter, highlight or URL actions? If so, do they work?
- Are your legends and filters grouped and placed intuitively?
- Do you have scrollbars in your views? If so, are they acceptable ones?
- Are your views scrunched?
- Do your views fit consistently well when you apply filters?

Check 5: applying to Traffic data in MASS package

```
> head(Traffic)
  year day limit  y
1 1961   1    no  9
2 1961   2    no 11
3 1961   3    no  9
4 1961   4    no 20
5 1961   5    no 31
6 1961   6    no 26
```

Reference Code:

- data(package = .packages(all.available = TRUE))
- /MASS