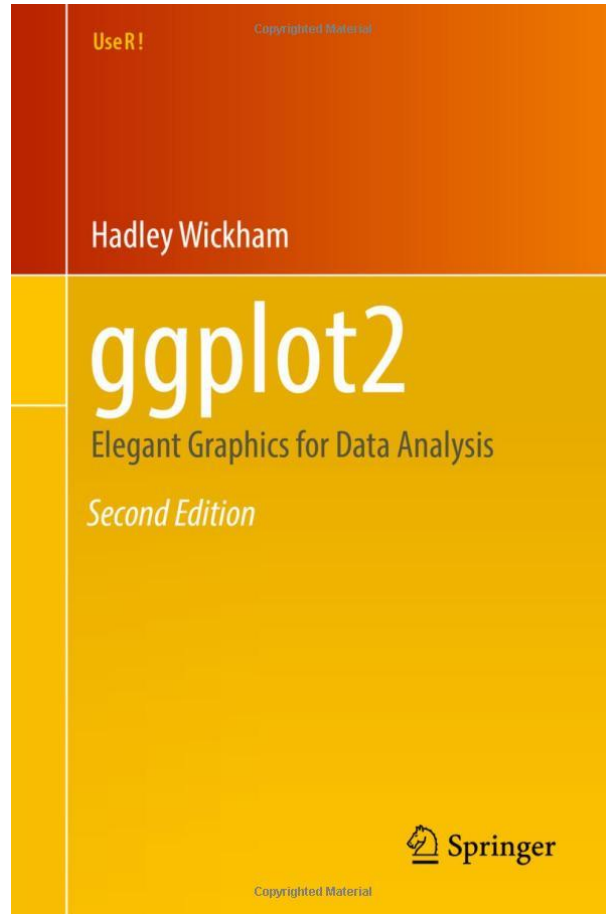# Data visualization

Phuc Loi Luu, PhD
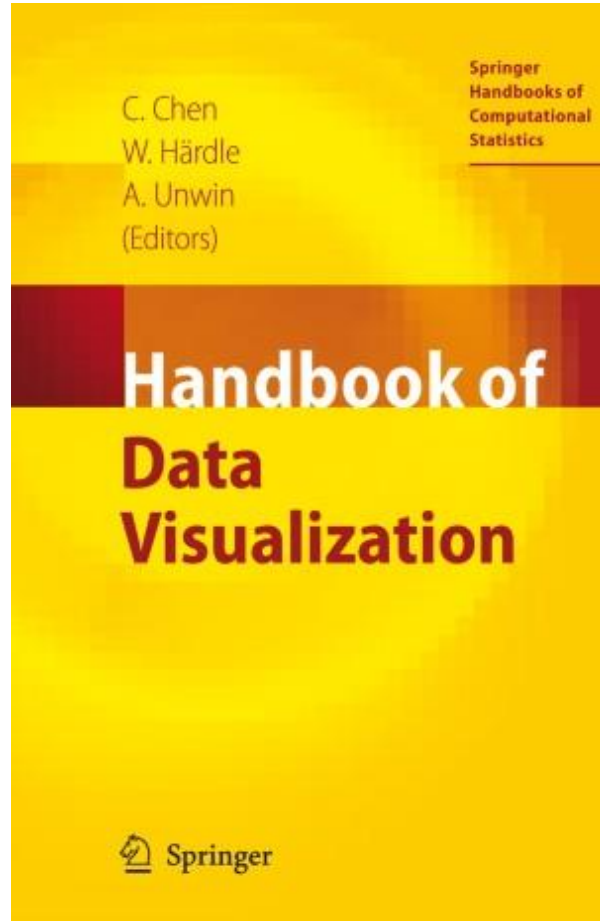
p.luu@garvan.org.au

Organization of the lecture

- Data visualization: 4h30 VN time, 10.01.2022 to 11.04.2022 (12 weeks + 2 weeks)

- Prerequisites:

  - Basic programming skills with R and Python

  - Basic Maths

  - Statistics

- Exercises:

  – weekly exercises, theoretical and practical work (roughly alternating),

  – practical exercises will be in R or Python,

  – 50% of the points in the exercises are needed to take part in the exams.

  - Final Exam:

    - presenting the course project (11.04.2022 at 4h30 pm)

    - script

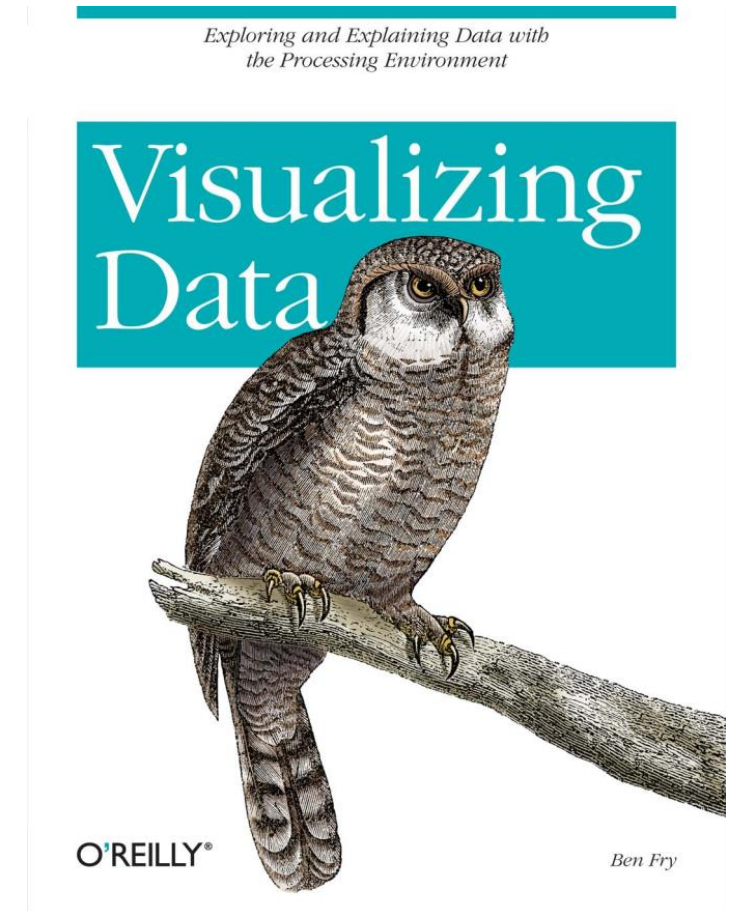 - Grading: passed if you get at least 50% of the points.

Books, slides, exercises and solutions to exercises

https://ggplot2-book.org/index.html
(GGP)

https://link.springer.com/book/10.1007/978-3-540-33037-0
(HDV)

https://www.iaacblog.com/programs/books-about-data-visualization/
(VD)

Roadmap of the lecture

- Introduction to data visualization

- The seven stages of visualizing data

- Design plot and panel

- Sparing and converting input and output data formats

- Maths plots

- Statistics plots

- Machine learning plots

- Map plots

- Time series plots

- Network plots

- Pathology images

- Plot processing with Inkscape

Basic Plots

1.1 Bar/pie/waterfall plot

1.2 Boxplot

1.3 Histogram/density/violin

1.4 Variant of boxplot/histogram/density/violin/cloud

1.5 Parallel plot and its variants

1.6 Heatmap

1.7 Dendogram

1.8 Venn diagram

1.9 Upset plot

1.10 Forest plot

1.11 Survival plot
1.12 Volcano and MA plot
1.13 Scatter plot
1.14 Dot plot
1.15 Scatter + histogram/boxplot/violin plot
1.16 Time series plot
1.17 Spatial/ Map/Location plot
1.18 Hexabin/contour plot
1.19 ROC curve
1.20 RMSE
1.21 Circos plot
1.22 Hexanbin plot
1.23 Maths plot
1.24 Radar plot

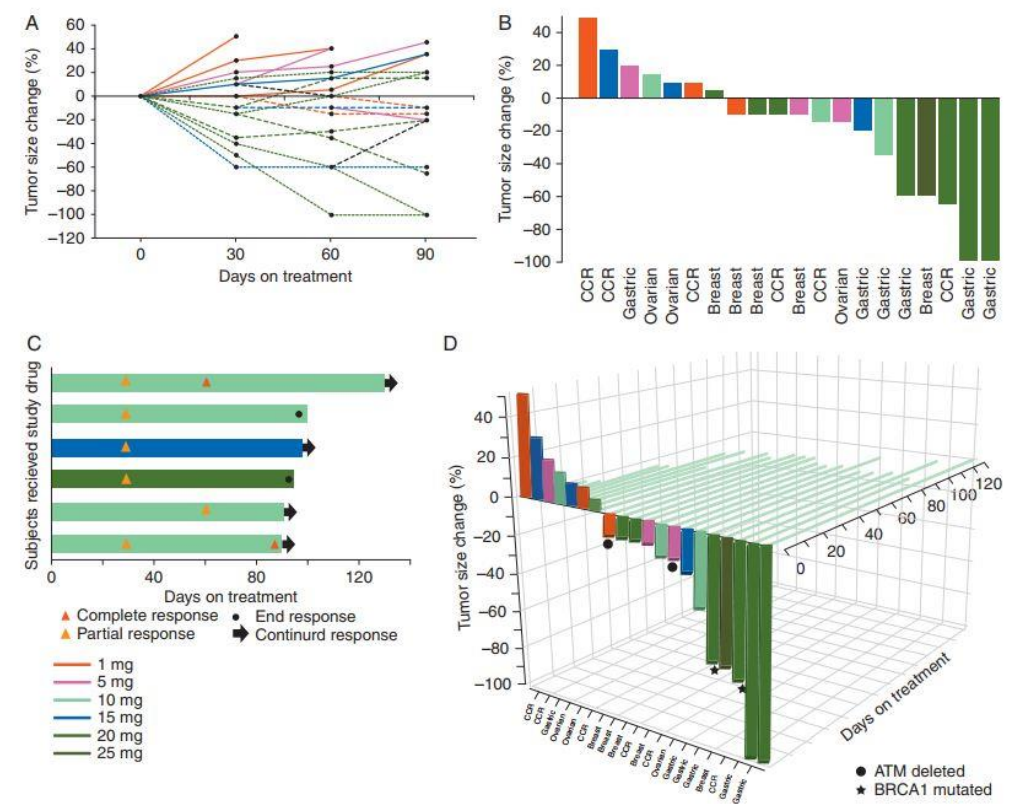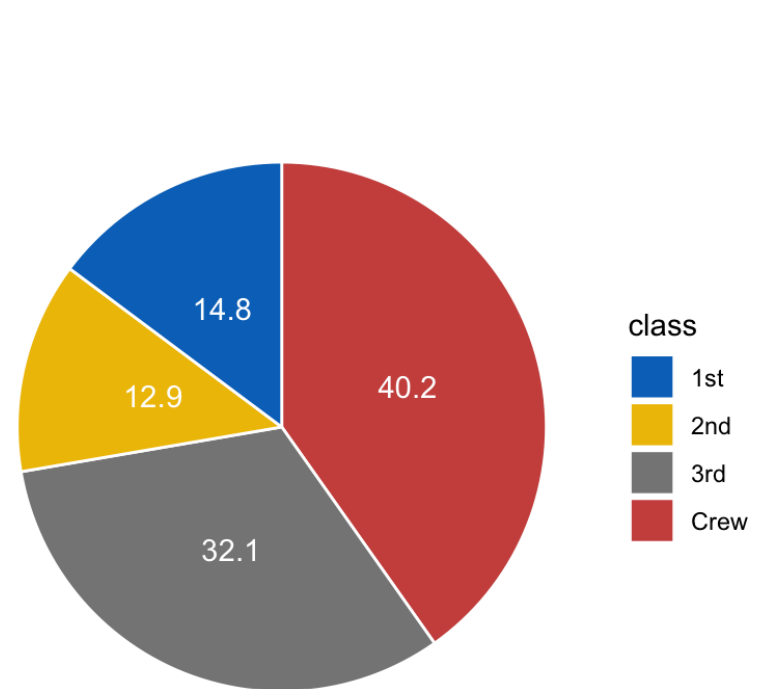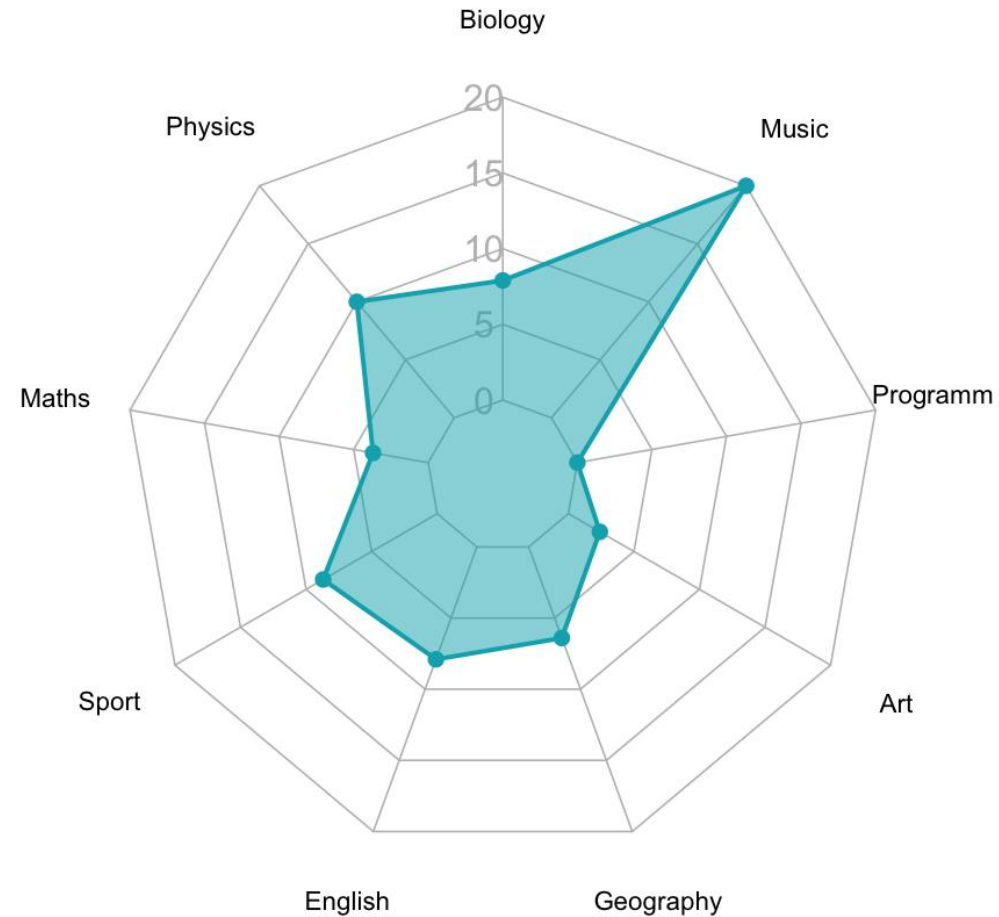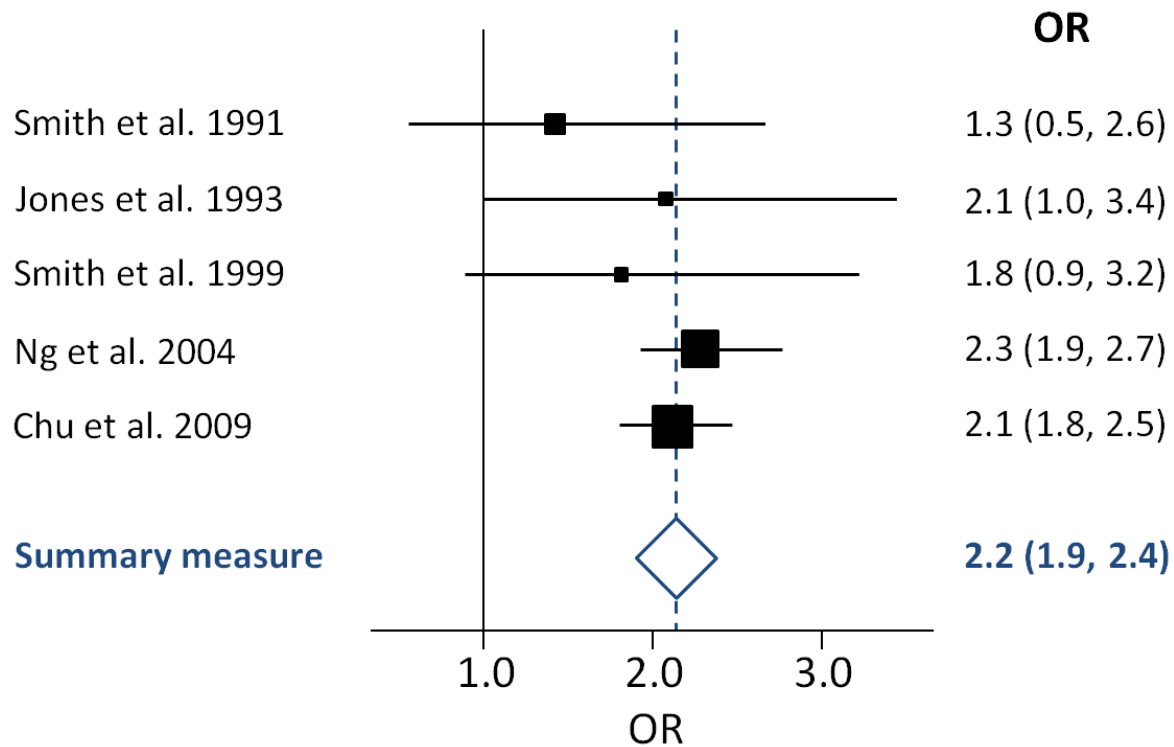# Pie + Bar + water fall + time plot

Figure 1 (A) Spider plot, (B) classical (2D) waterfall plot, (C) swimmer plot and (D) 3D waterfall plot.
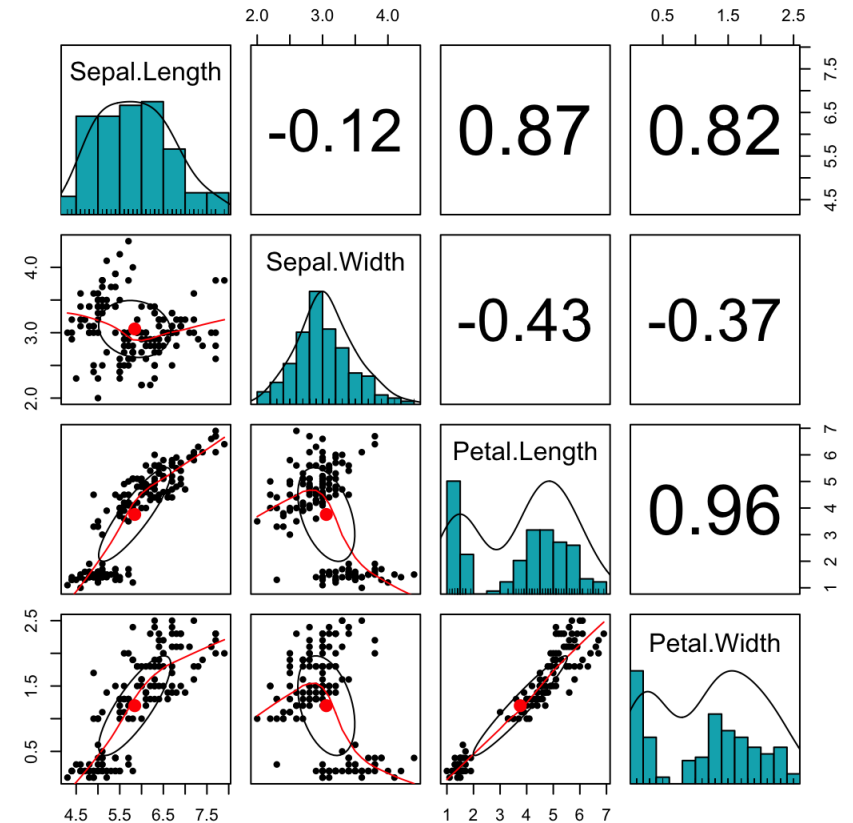
10.1093/annonc/mdw656

Radar plot

https://www.datanovia.com/en/blog/beautiful-radar-chart-in-r-using-fmsb-and-ggplot-packages/
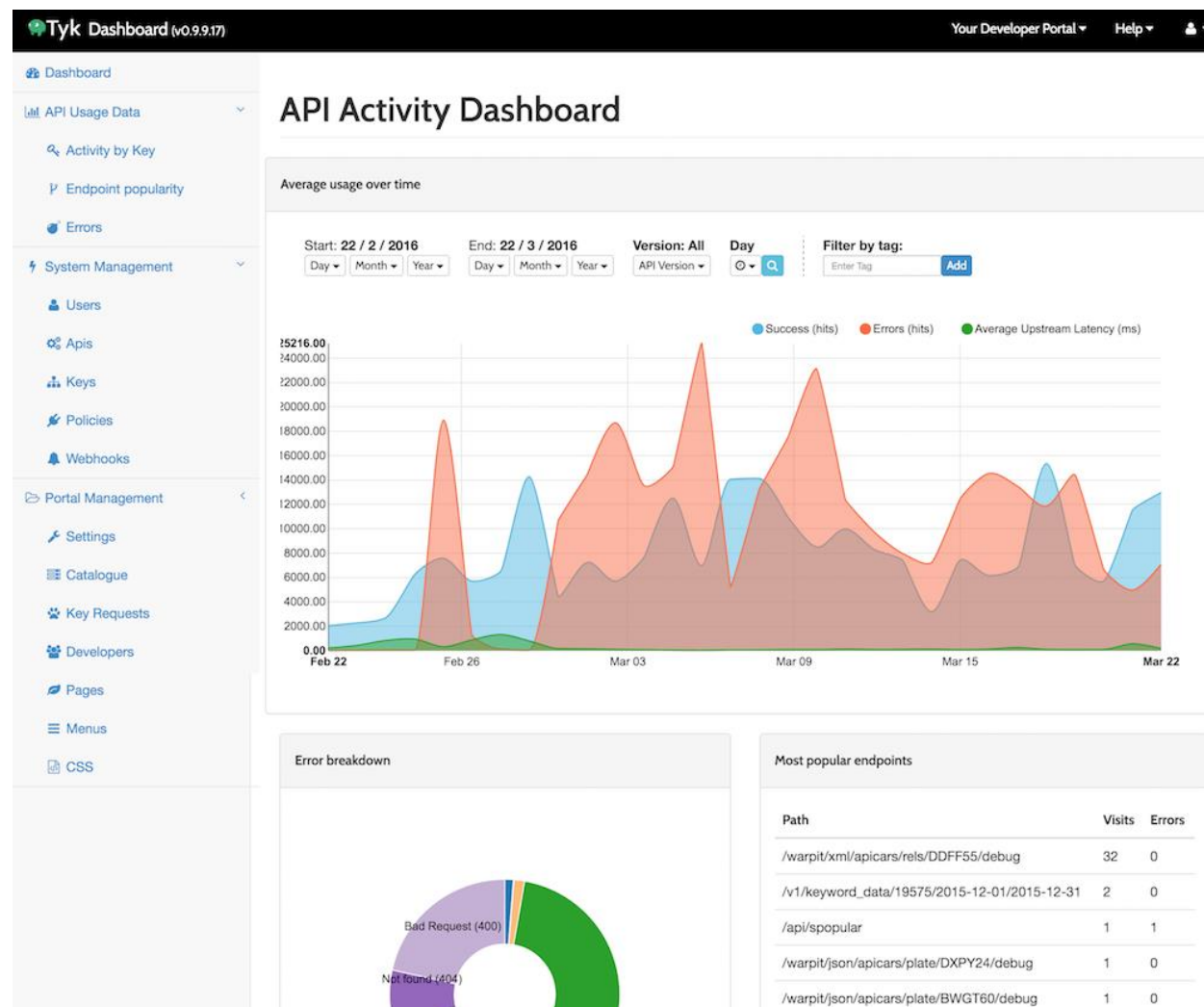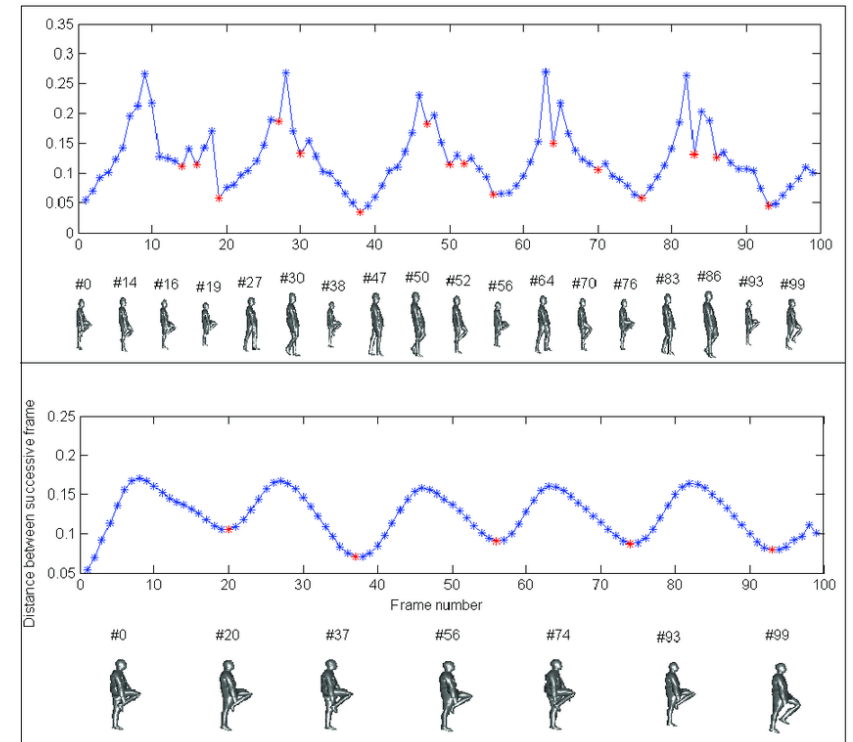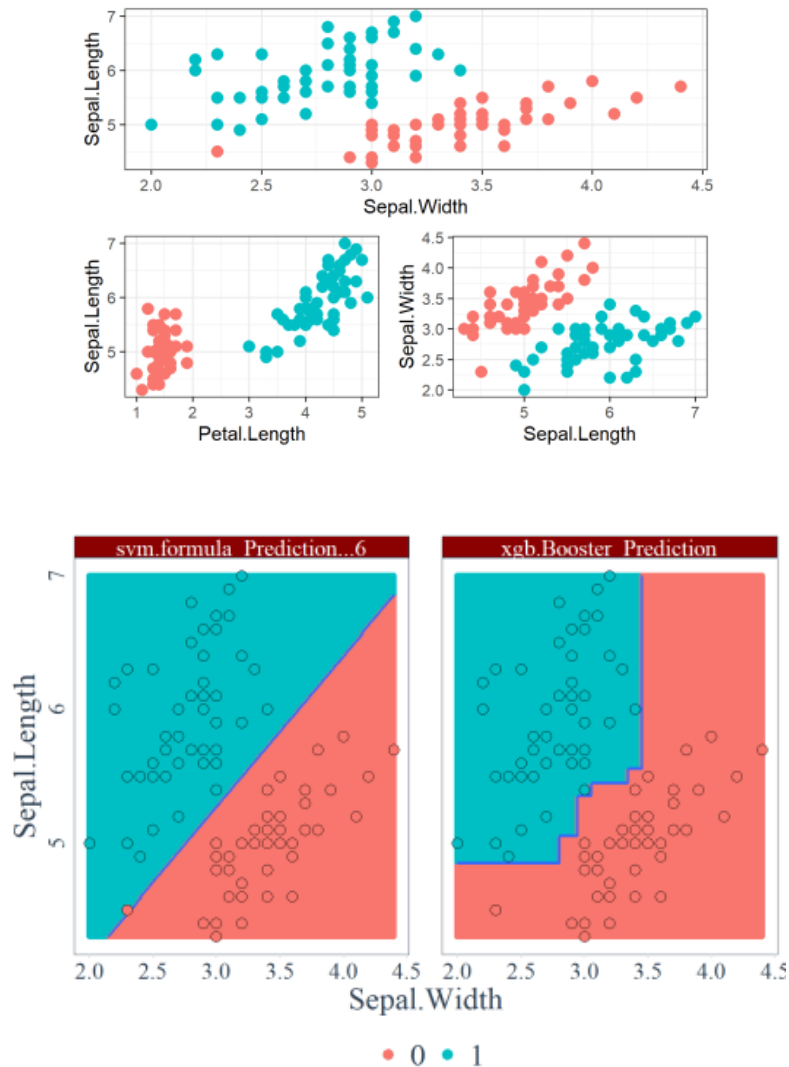
Forest plot vs matrix plot
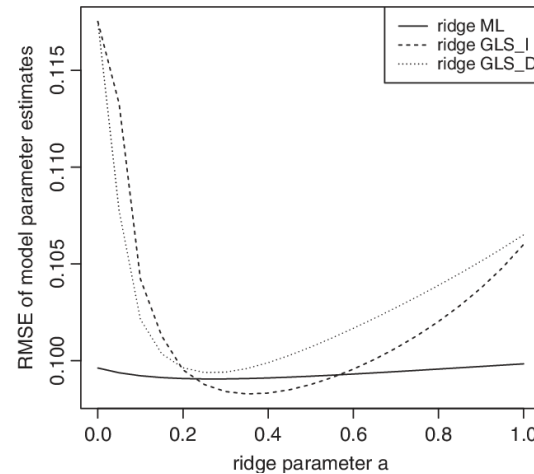
Dashboard with Tyk/Tableau

Data visualization for machine learning

Speed curve smoothing: (top) speed curve before smoothing, (bottom) speed curve after smoothing.
https://hal.archives-ouvertes.fr/tel-01094740/file/RimSlamaThesis2014.pdf

10.1080/10705511.2015.1077335

Amination plot

Survival plot (Kaplan-Meier Curve)

http://www.sthda.com/english/wiki/survival-analysis-basics
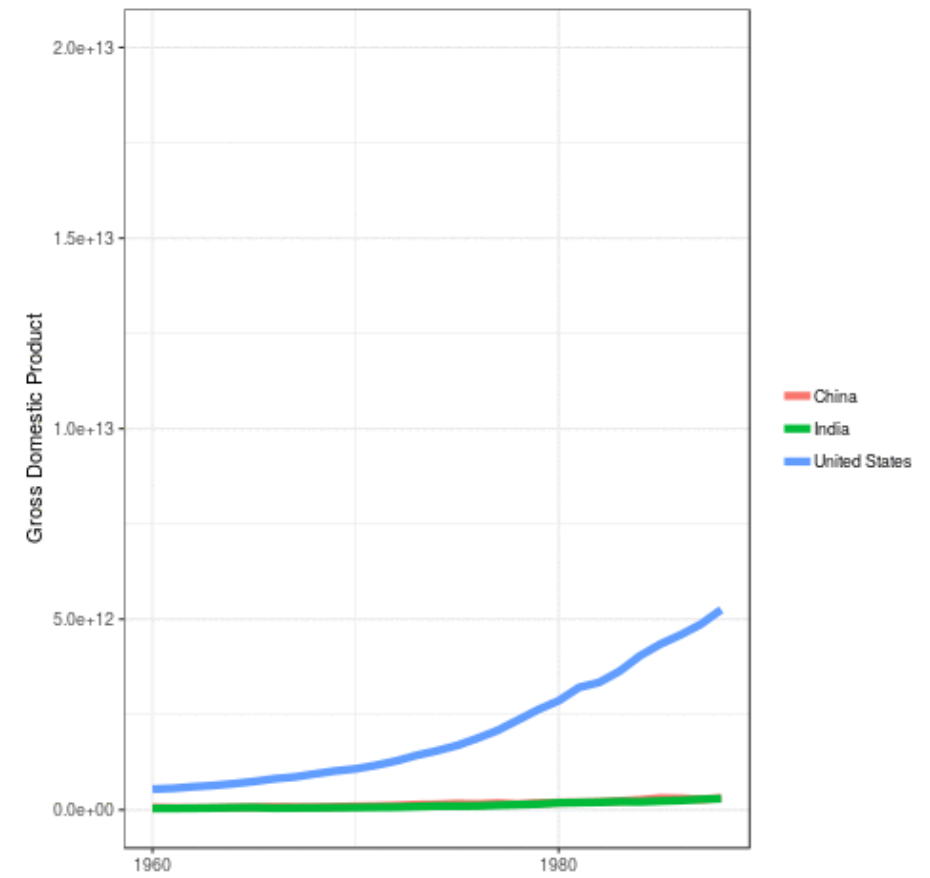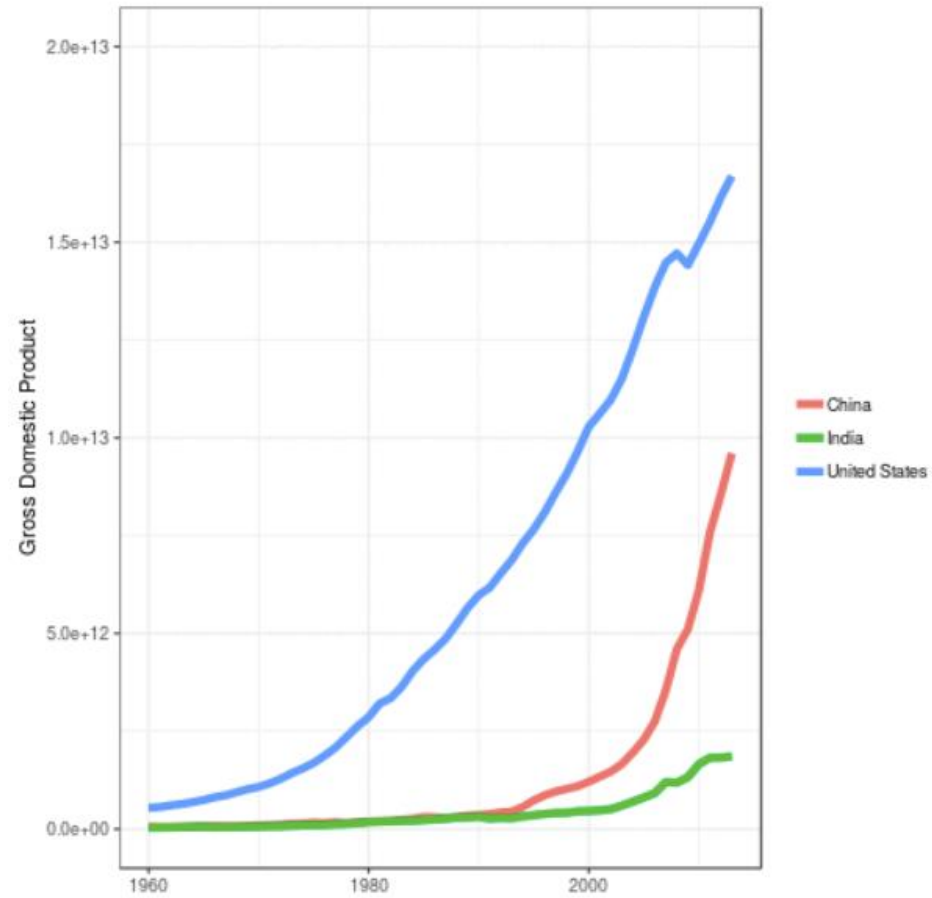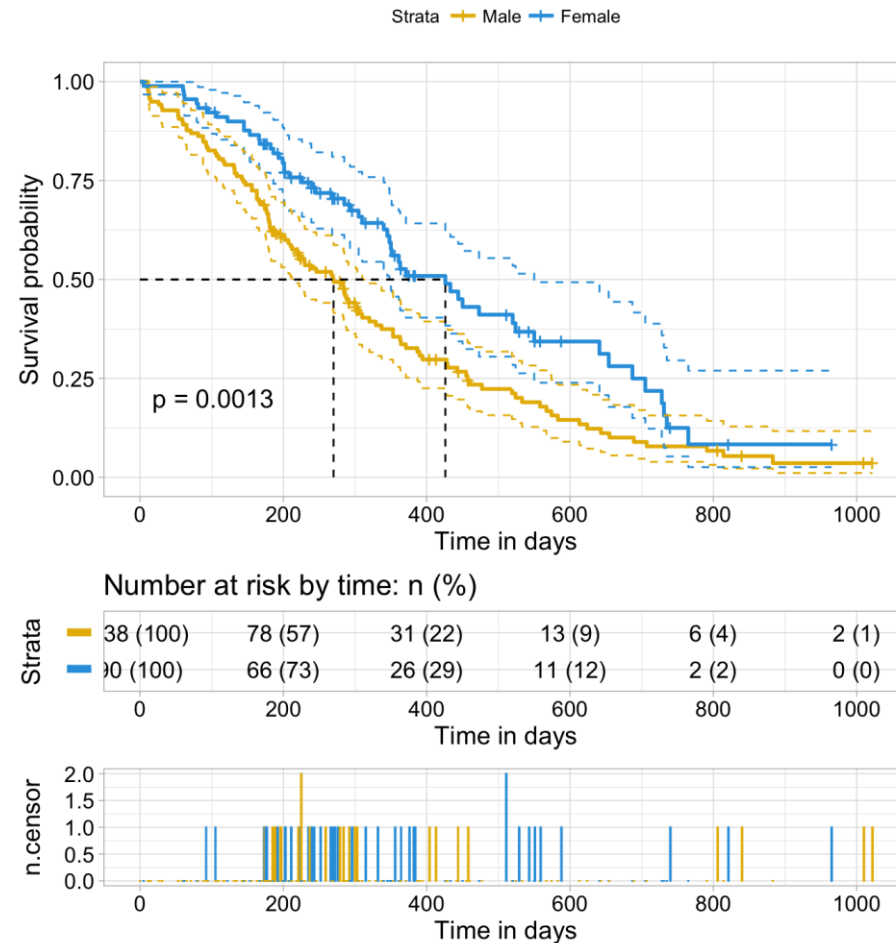
Linked Views for Visual Exploration
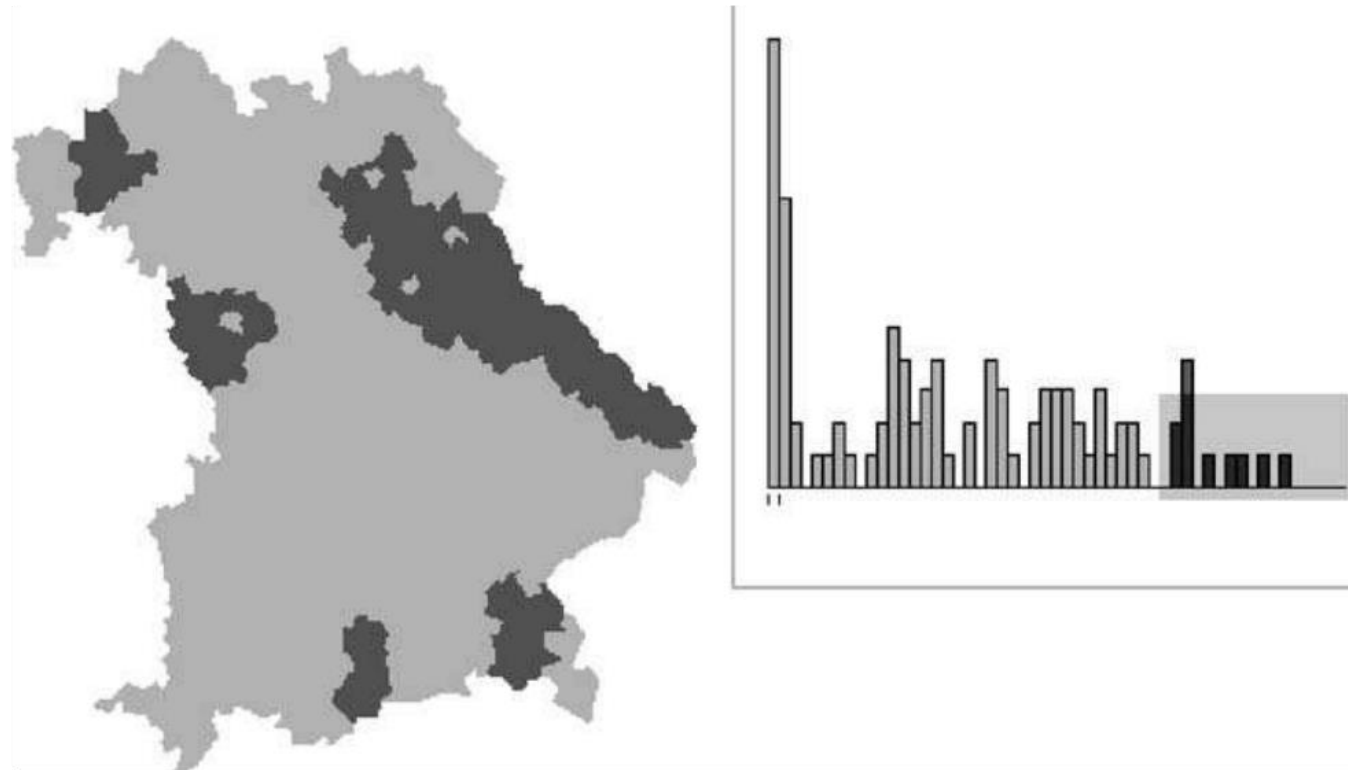
Figure 8.3. Land usage on the county level in Bavaria. Highlighted are those counties with a high percentage of forestry

Linked Views for Visual Exploration
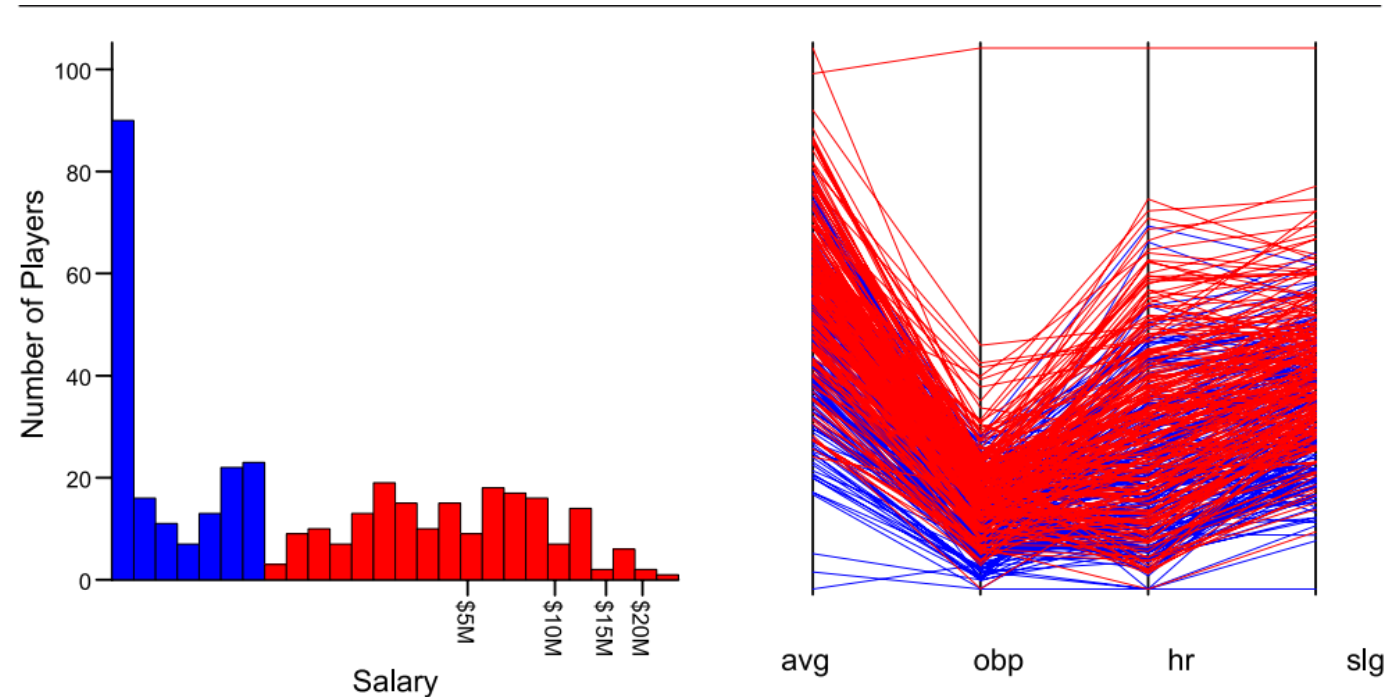
Figure 9.5. Faceting and parallel coordinates. For the 2004 season, the *left view* shows a histogram of salary on a log scale. Players earning over $1,000,000 were selected and are shown in the linked *parallel coordinates plot* as *dashed line*. The parallel coordinates view shows four major measures of performance: batting average, on-base percentage, home-run rate, and slugging average

Binning and contour plot
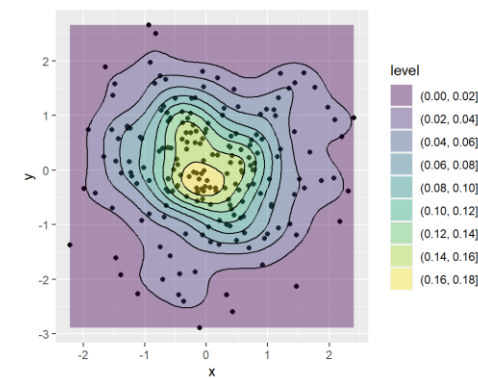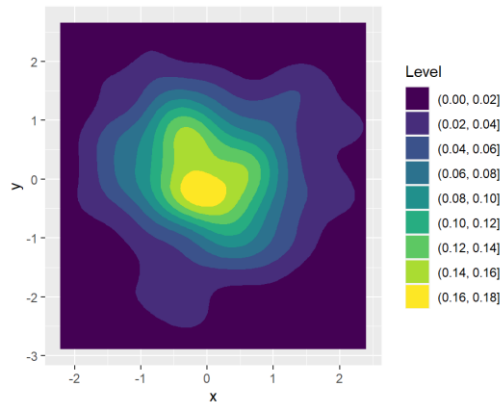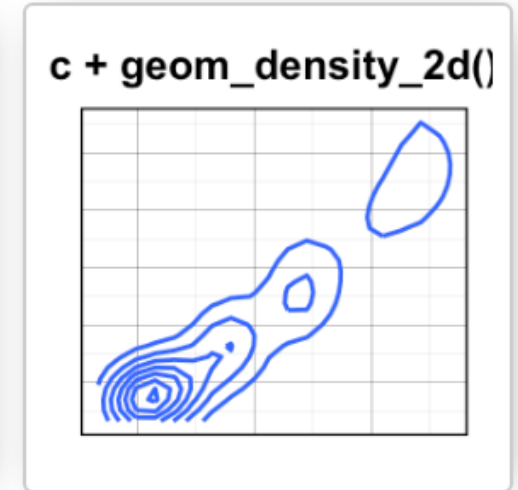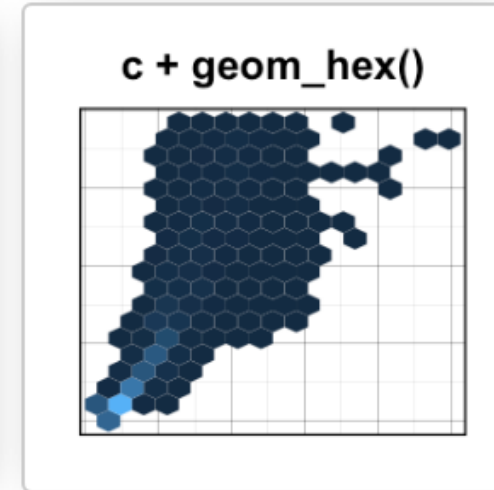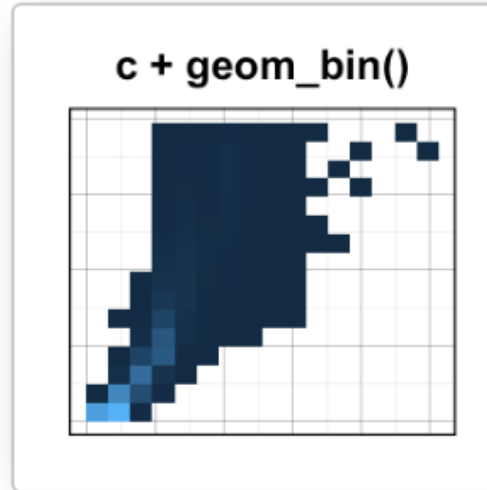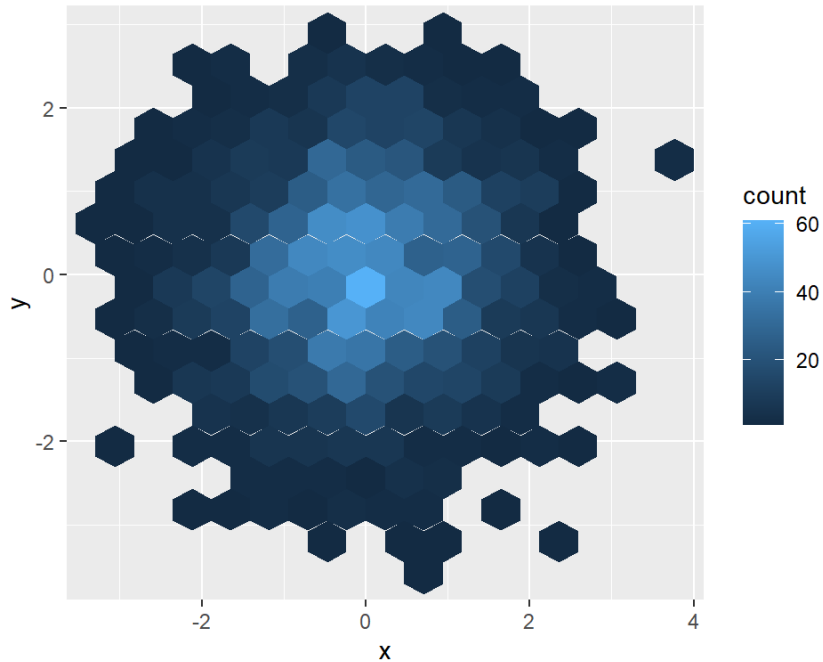
3D contour plot for Gradient Descent
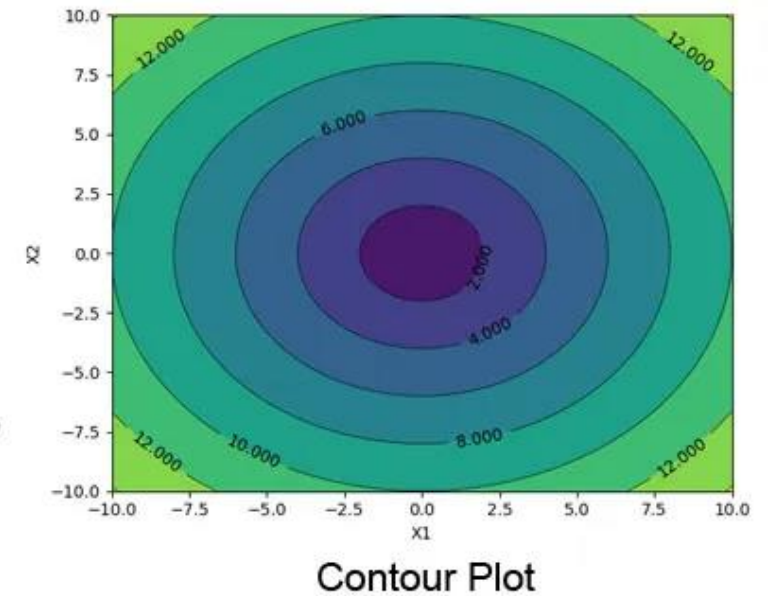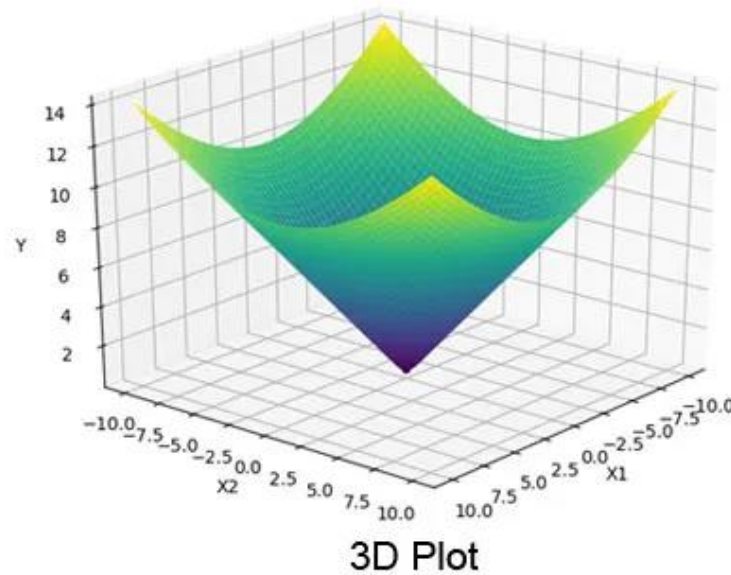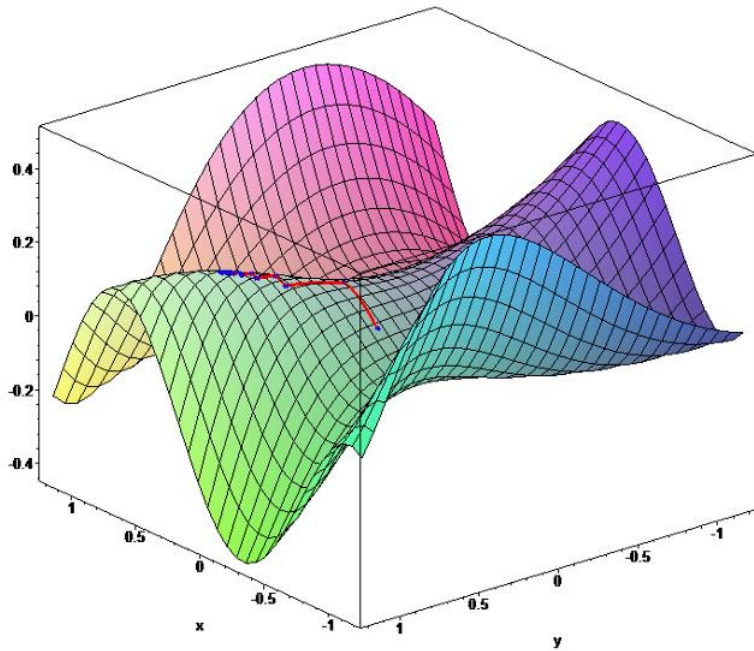
Dotplot and jitter plot
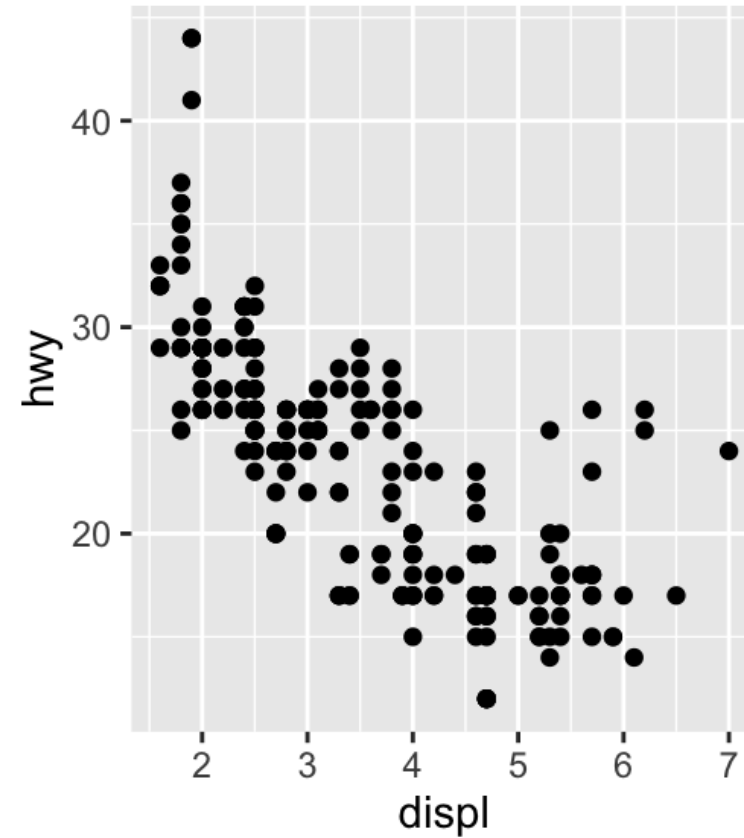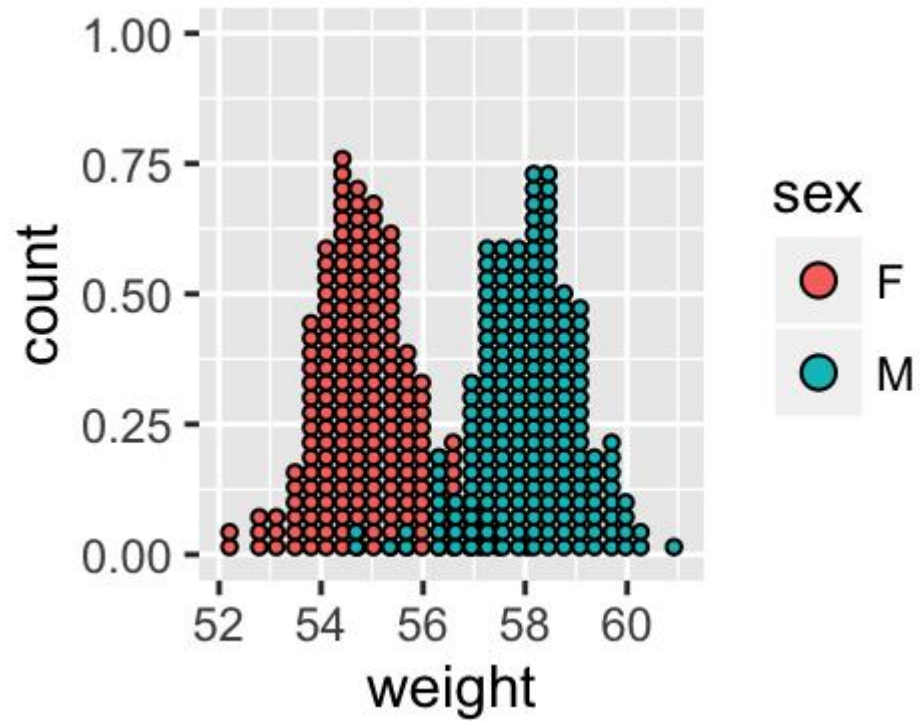
Volcano vs MA (Bland–Altman) plot

http://onetipperday.sterding.com/2012/11/t-test-volcano-plot.html

https://biocorecrg.github.io/CRG_RIntroduction/volcano-plots.html
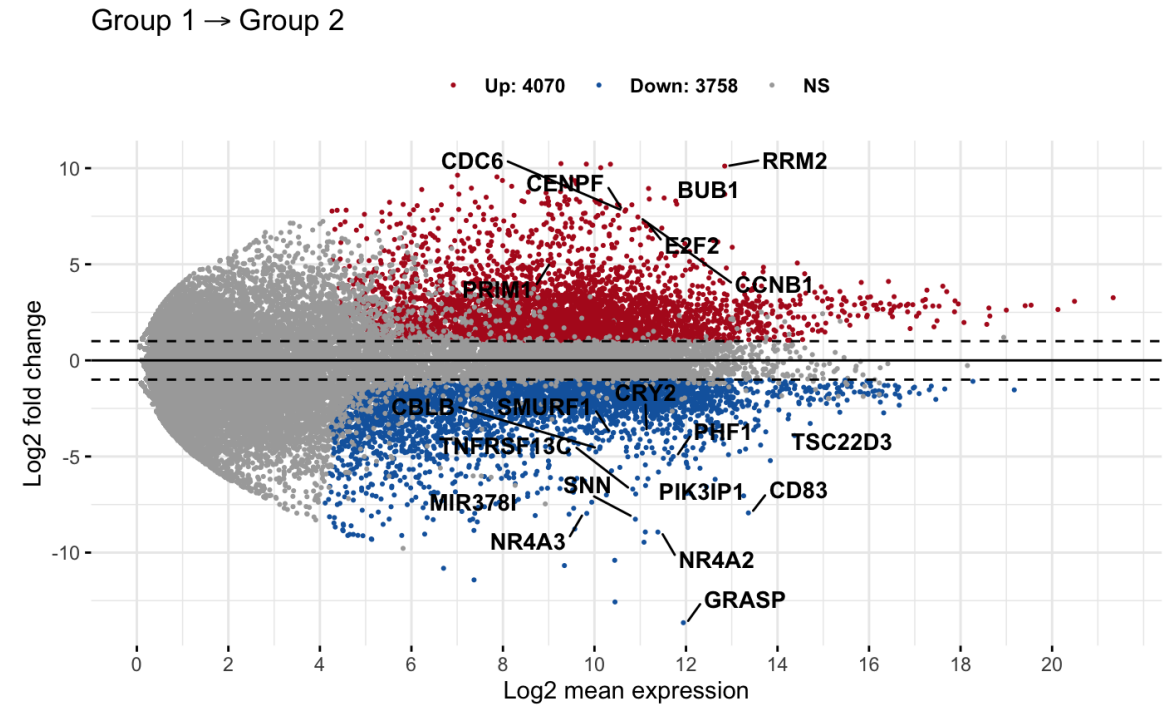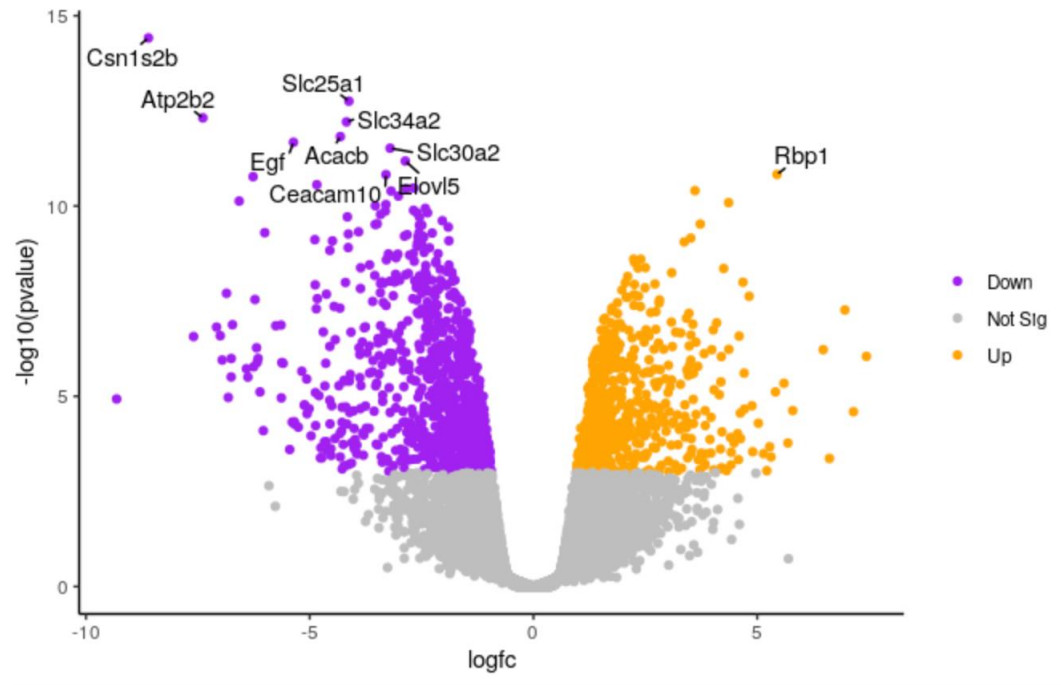
https://rpkgs.datanovia.com/ggpubr/reference/ggmaplot.html

RCircos 2D

RCircos Demo: Mouse and Rat Gene Expression

Right: GSE42081 Mouse
Left: GSE42081 Rat
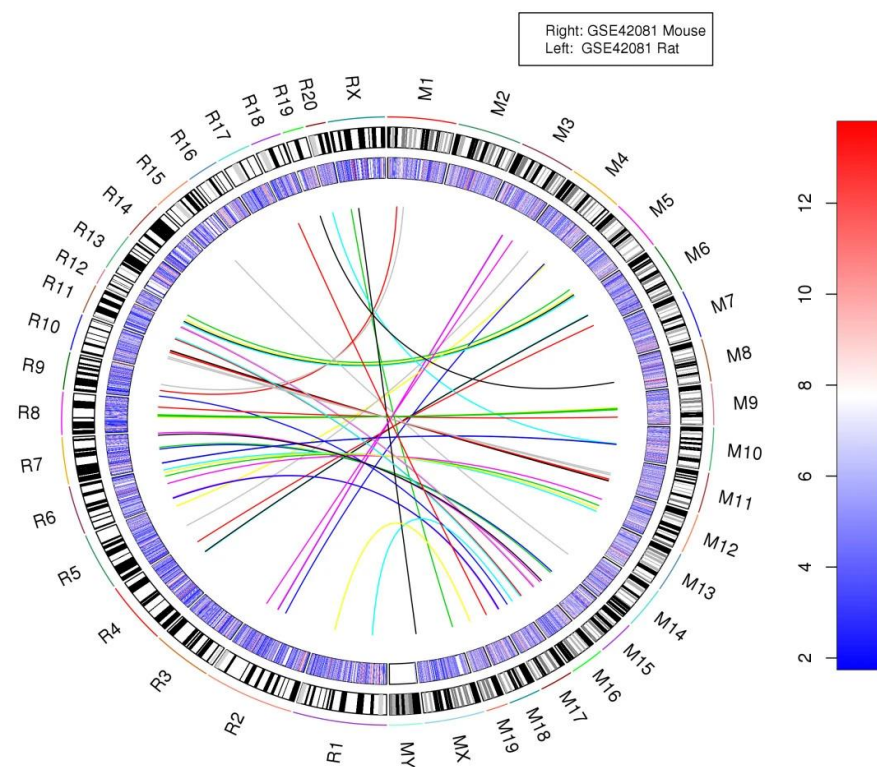
RCircos 2D Track Plot with Human Genome

3D plot

USA arrests

http://www.sthda.com/english/wiki/impressive-package-for-3d-and-4d-graph-r-software-and-data-visualization
https://stackoverflow.com/questions/43330727/how-to-plot-three-dimension-data-or-four-dimension-data-to-filled-contour-plot-o
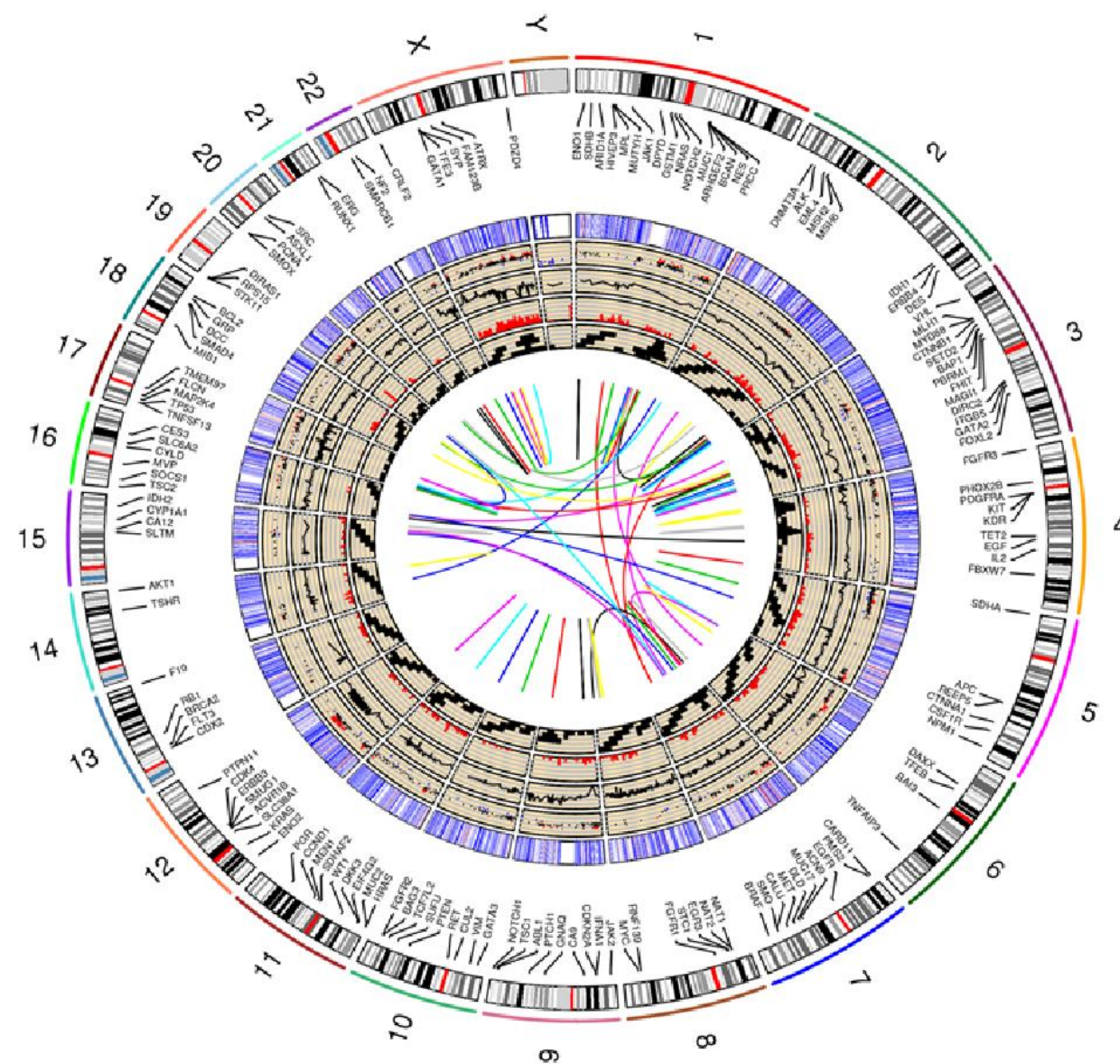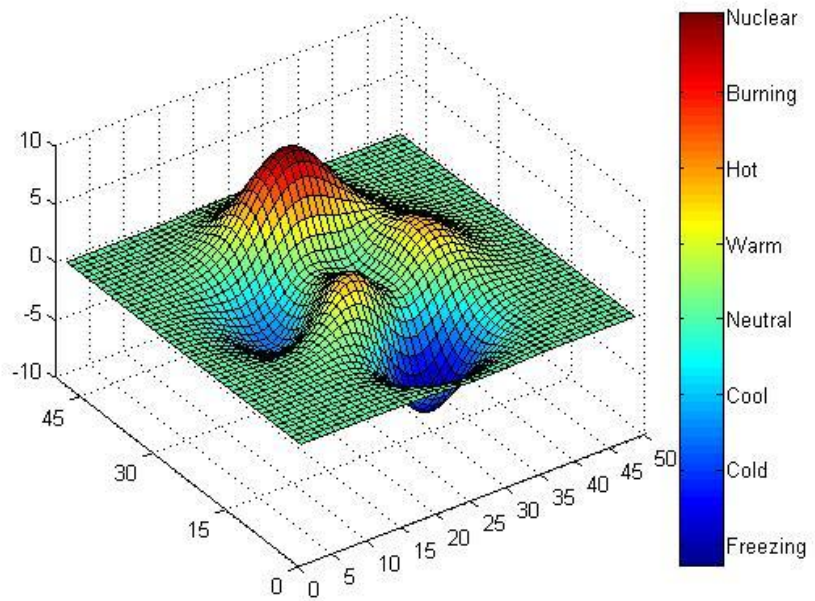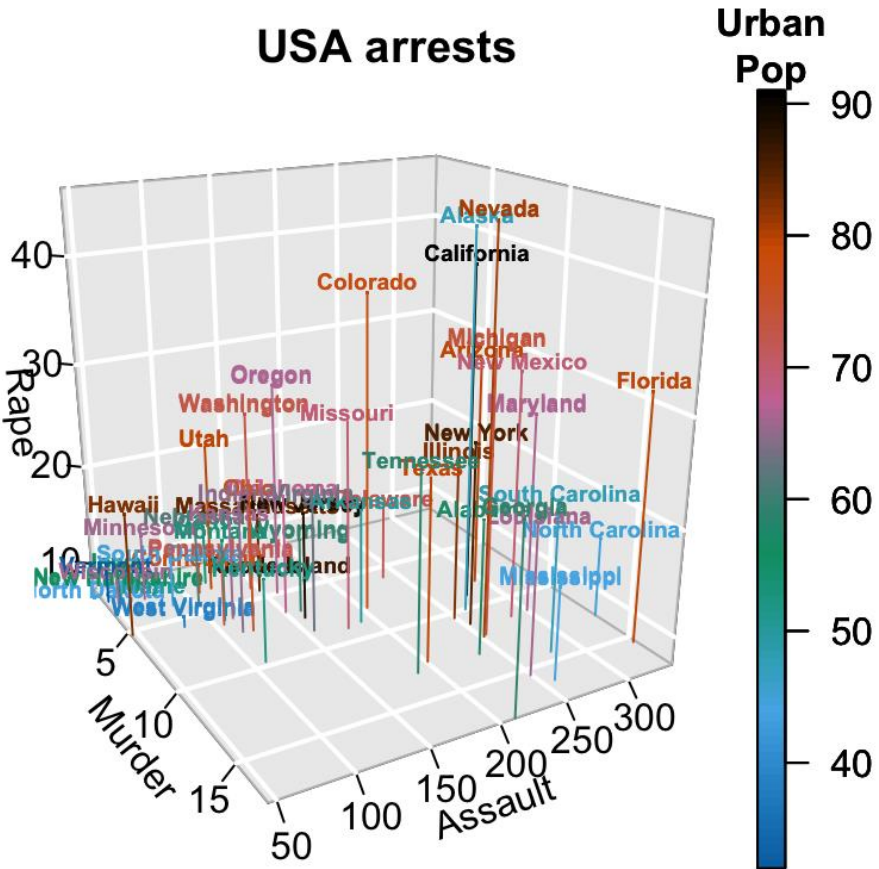
Draw Cell
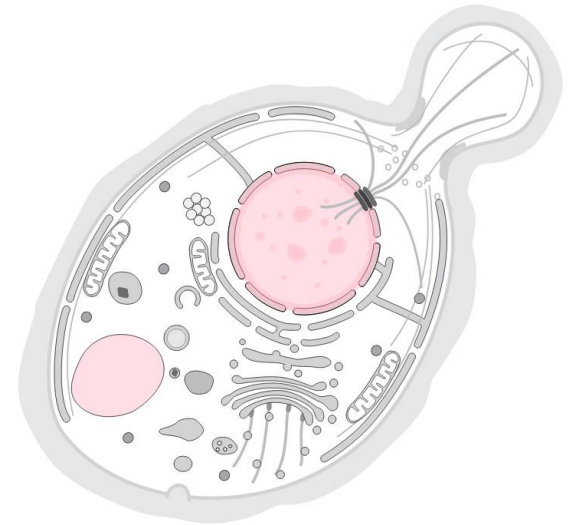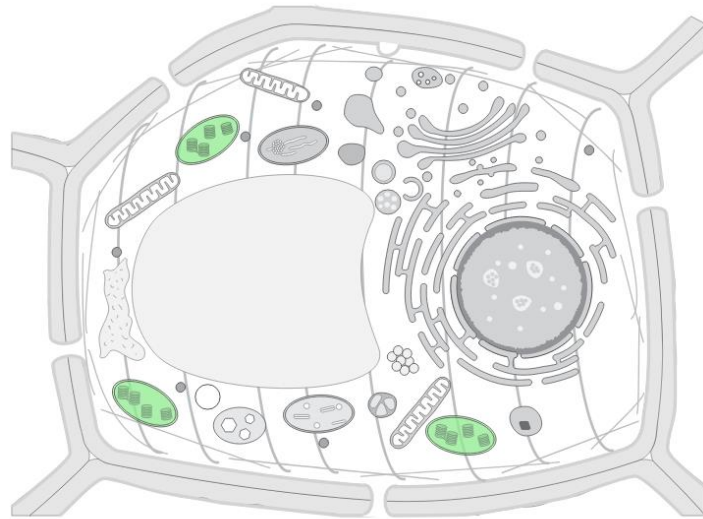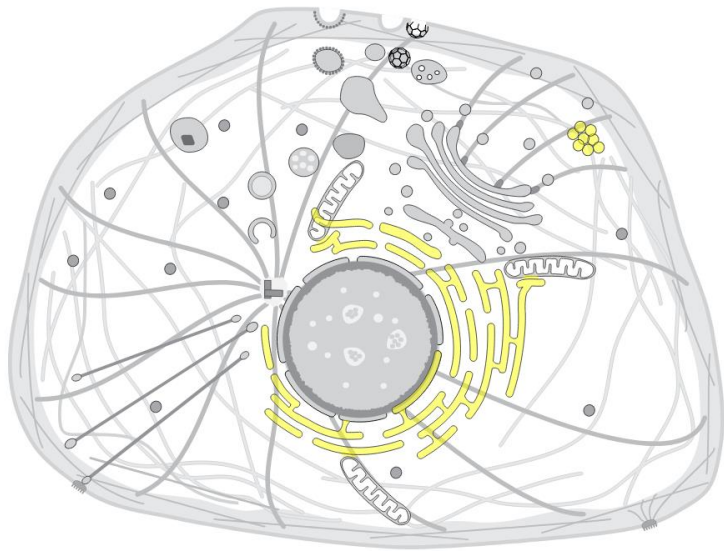
Inkscape for scientists: Making a data figure for a publication
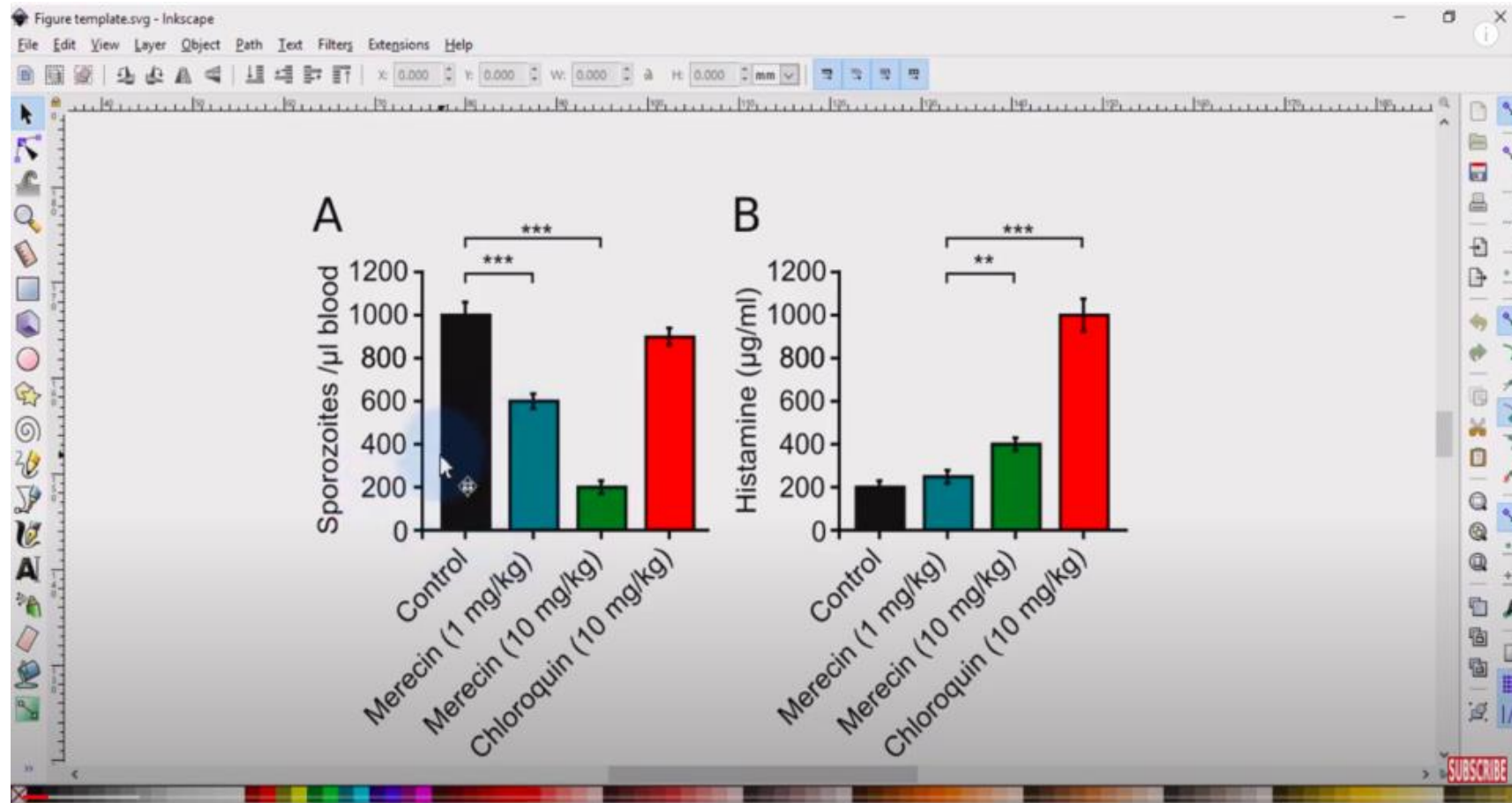
Why Do We Visualize Data?

TABLE 1.1 Table with four groups of numbers: What do they tell you?

Group A		Group B		Group C		Group D	
x	y	x	y	x	y	x	y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

Why Do We Visualize Data?
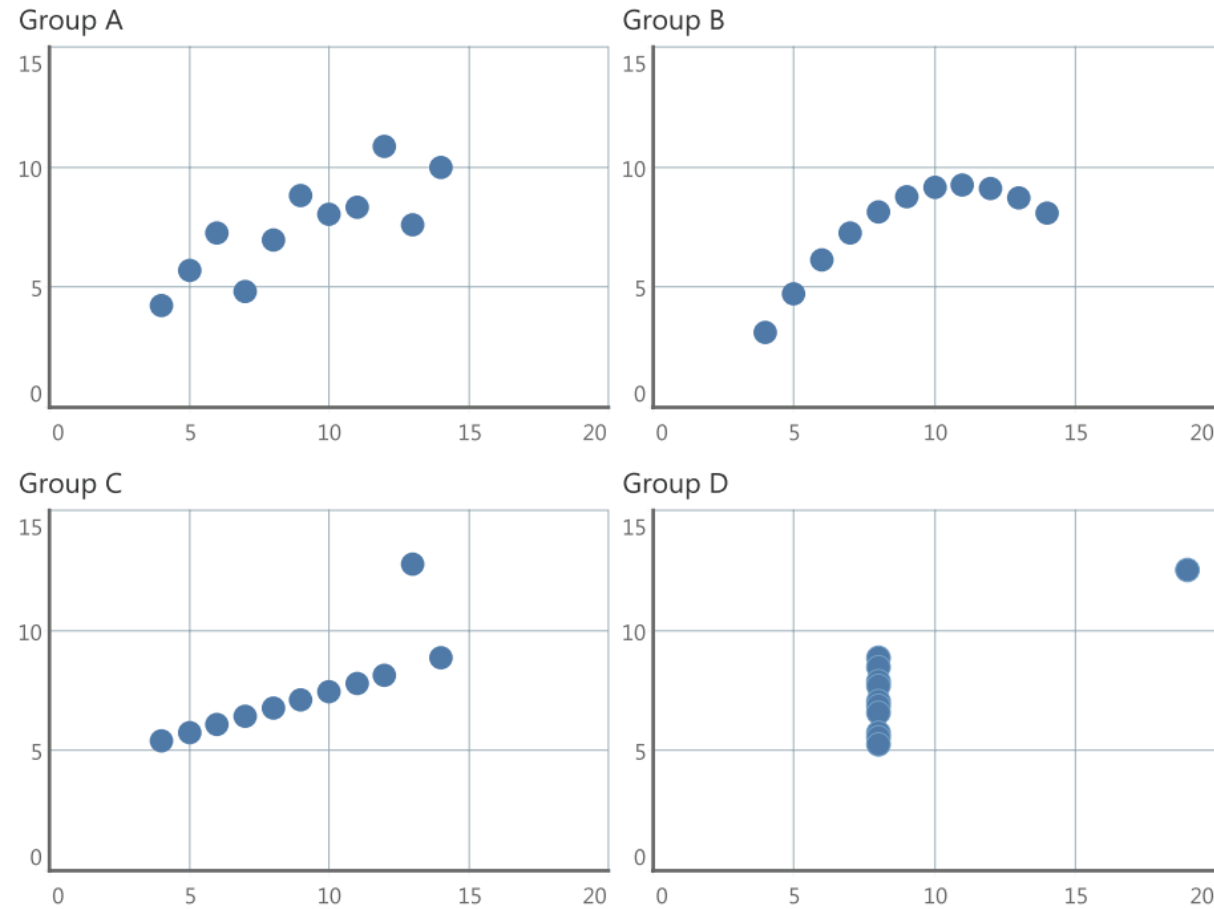
FIGURE 1.1   Now can you see a difference in the four groups?

"Anscombe's Quartet"—in the paper "Graphs in Statistical Analysis" in 1973.

Why Do We Visualize Data?

TABLE 1.2   What are the trends in sales?

Category	2013 Q1	2013 Q2	2013 Q3	2013 Q4	2014 Q1	2014 Q2	2014 Q3	2014 Q4
Furniture	$463,988	$352,779	$338,169	$317,735	$320,875	$287,934	$319,537	$324,319
Office Supplies	$232,558	$290,055	$265,083	$246,946	$219,514	$202,412	$198,268	$279,679
Technology	$563,866	$244,045	$432,299	$461,616	$285,527	$353,237	$338,360	$420,018
Category	2015 Q1	2015 Q2	2015 Q3	2015 Q4	2016 Q1	2016 Q2	2016 Q3	2016 Q4
Furniture	$307,028	$273,836	$290,886	$397,912	$337,299	$245,445	$286,972	$313,878
Office Supplies	$207,363	$183,631	$191,405	$217,950	$241,281	$286,548	$217,198	$272,870
Technology	$333,002	$291,116	$356,243	$386,445	$386,387	$397,201	$359,656	$375,229

Sales by category

FIGURE 1.2   Now can you see the trends?

Neural network Interpretable

Neural network Interpretable

Figure 5: The multi-class experiment result. (A,B,C) show the decision boundary results on a

Projects 1

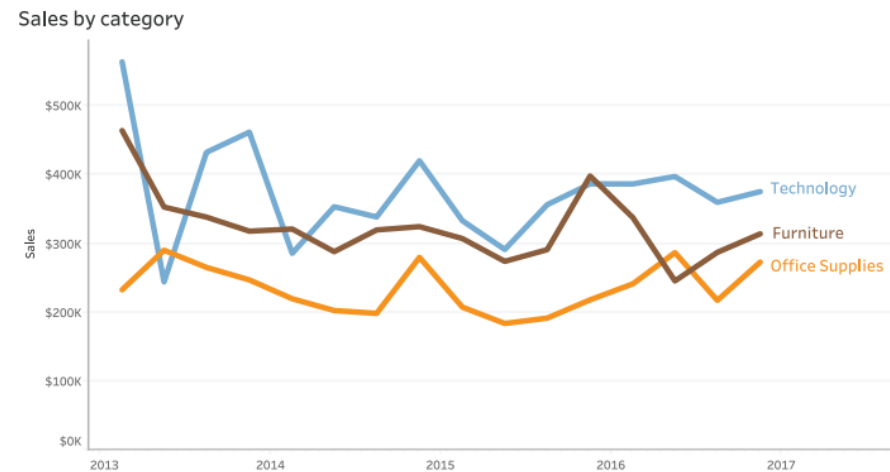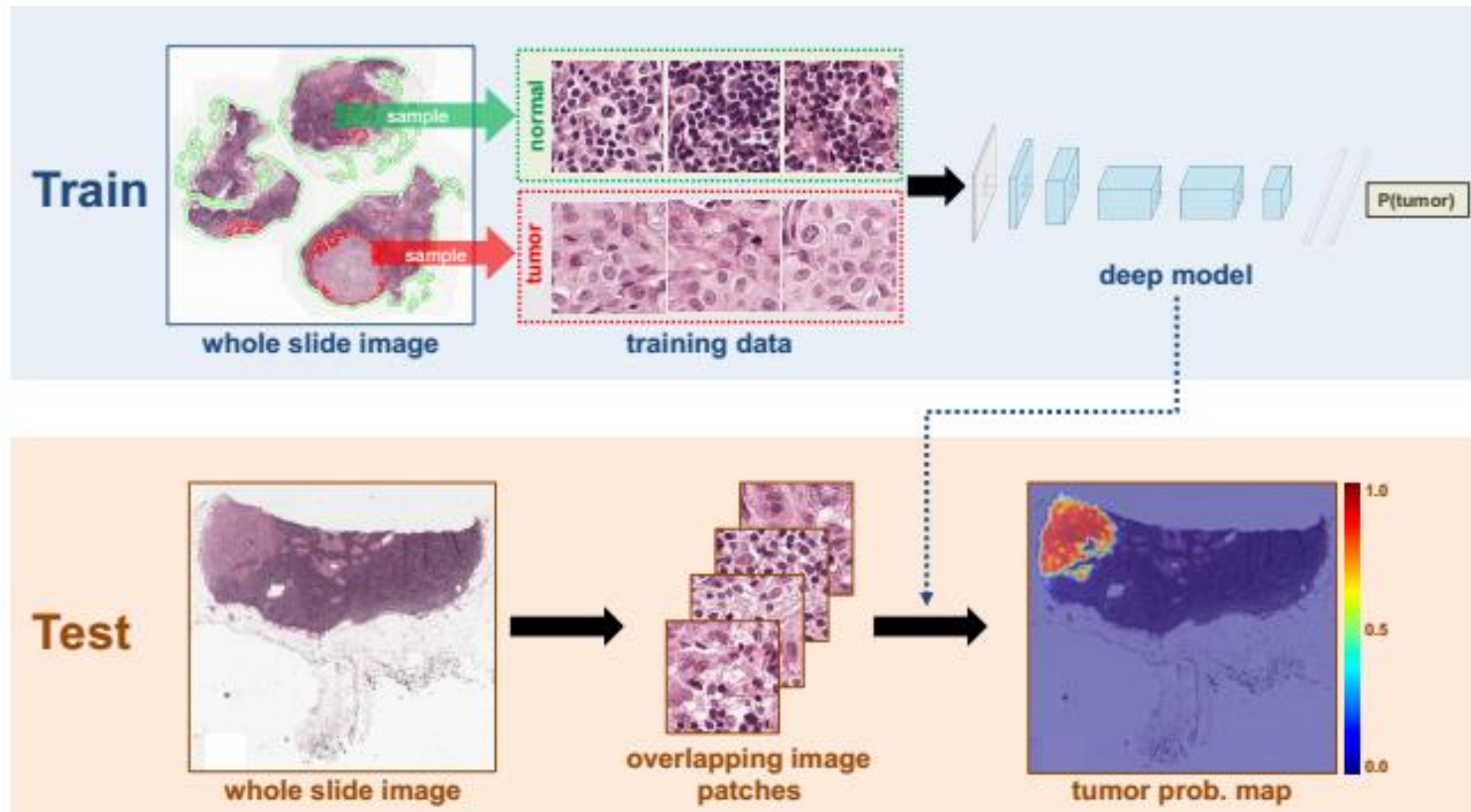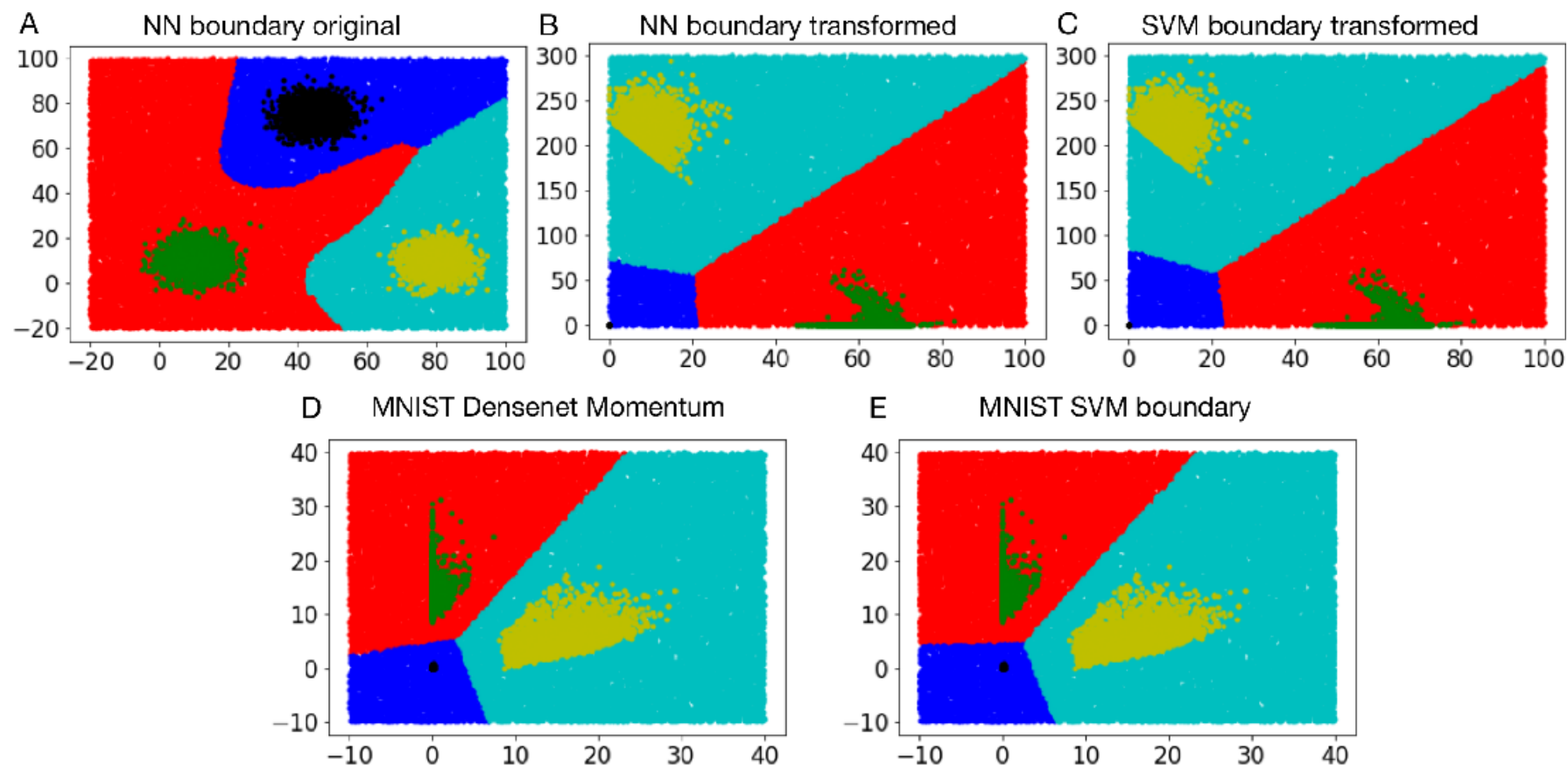| No	Project	Students	Slide/Book	Video
1	Multiple/Simple Regression by Parts RMSE Logistic regression ROCcurve and AUC	Thanh Giang Thien	HDV: chapter III.7 https://towardsdatascience.com/logistic-regression-from-scratch-in-r-b5b122fd8e83 https://www.analyticssteps.com/blogs/introduction-logistic-regression-sigmoid-function-code-explanation http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/	
2	Animation and Interactive math function animate, gganimate, googlevis	Thong	https://towardsdatascience.com/animating-your-data-visualizations-like-a-boss-using-r-f94ae20843e3	https://www.youtube.com/watch?v=PBTrvPewJ_s
3	Map and Location	Hoang, Kim Cuong	https://ggplot2-book.org/maps.html Chapter 3: Mapping, VD	
4	Time Series	Nhuong	Chapter 4: Time Series, VD	
5	Connections and Correlations	Thien	Chapter 5: Connections and Correlations, VD	
6	Boxplot + Histogram + Density + Violin + Scatterplot + dotplot + jitter + Variant of boxplot/histogram/density/violin/cloud bar/pie/waterfall plot	Kim Cuong	https://ggplot2-book.org/maps.html http://www.sthda.com/english/wiki/be-awesome-in-ggplot2-a-practical-guide-to-be-highly-effective-r-software-and-data-visualization HDV: chapter III.6	
7	Venn diagram vs Upset plot Heatmap vs Parallele plot Forest plot vs matrix plot	Hoang	http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs https://jokergoo.github.io/ComplexHeatmap-reference/book/	
8	Binning and contour plot 3D contour plot for Gradient Descent	Giang	https://r-charts.com/correlation/hexbin-chart-ggplot2/ https://r-charts.com/correlation/contour-plot-ggplot2/ https://vincenzocoia.com/post/contour_plots/	https://medium.com/@jackyfeng530/machine-learning-gradient-descent-8681c40d27c3 http://www.adeveloperdiary.com/data-science/how-to-visualize-gradient-descent-using-contour-plot-in-python/

Projects 2

No	Project	Students	Slide/Book	Video
9	Volcano vs MA (Bland–Altman) plot	Minh	https://rpkgs.datanovia.com/ggpubr/reference/ggmaplot.html https://biocorecrg.github.io/CRG_RIntroduction/volcano-plots.html	
10	Trees, Hierarchies, and Recursion Visualizing Trees and Forests	Truong, Phu	HDV: chapter II.10 VD: chapter 7	
11	Visualizing Cluster Analysis and Finite Mixture Models Multidimensional Scaling	Xuan Minh	HDV: chapter III.11 and III.3	
12	Smoothing Techniques for Visualisation	Thinh	HDV: chapter III.9	
13	Parallel Coordinates: Visualization, Exploration and Classification of High-Dimensional Data	Xuan	HDV: chapter III.14	
14	III.15 Matrix Visualization III.12 Visualizing Contingency Tables	Bac	HDV: chapter III.12 and III.15	
15	Survival plot (Kaplan-Meier Curve)	Tam	http://www.sthda.com/english/wiki/survival-analysis-basics	
16	Networks and Graphs	Giang	VD: chapter 8	
17	Visualization in Bayesian Data Analysis	Hanh	HDV: chapter III.16	

Projects 3

No	Project	Students	Slide/Book	Video
18	Dashboard	Hieu	https://community.tyk.io/t/setting-up-tib-and-dashboard/662/4	https://www.youtube.com/watch?v=DjYStQQTje0
19	Quantitative Pathology & BioImage Analysis: QuPath HE staining image visualization and NDPview2 tutorial	Hiep	http://www.uwhistologyandimaging.org/wp-content/uploads/2014/11/NDP_view2_procedure.pdf https://www.hamamatsu.com/sp/sys/en/manual/NDPview2_manual_en.pdf	https://www.youtube.com/watch?v=4An5n6Y_rRI&t=1832s https://www.youtube.com/watch?v=NUH-0xJImRQ
20	QuPath tutorial 1 Convert image to matrix of number	Hieu		https://www.youtube.com/watch?v=J-47tzXAFdE&list=PL4ta8RxZklWkPB_pwW-ZDVAGPGktAlE5Y
21	QuPath tutorial 2 for IHC analysis	Q Thanh		https://www.youtube.com/watch?v=aTVfJk6yNKs&list=PL4ta8RxZklWk_O_Z7K0bZlhmHtaH73vlh
22	Generating a CAMELYON16 training set using CVEDIA Plot training, testing accuracy and loss	Trung		https://www.youtube.com/watch?v=rH6Mfjdck0I https://www.youtube.com/watch?v=eHEgV-XqKnw
23	Plotly and ShinyApp	Thinh	https://towardsdatascience.com/animating-your-data-visualizations-like-a-boss-using-r-f94ae20843e3	
24	Inkscape	Hiep	https://www.youtube.com/watch?v=8f011wdiW7g&list=PLqazFFzUAPc5lOQwDoZ4Dw2YSXtO7lWNv	https://www.youtube.com/watch?v=eyqH0IrzYLc&list=PLxtauMB7RON_2tg-mRQTuieFUr29IOKzW
25	Circos 2D	Hanh	https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-244	https://jokergoo.github.io/circlize_book/book/
26	Draw cell	Truong	https://github.com/svalvaro/drawCell	