Statistical modelling

#1.c Exploratory Data Analysis

Dr. Léo Belzile
HEC Montréal

Type of data

Data base typically comprise many different *variable types*.

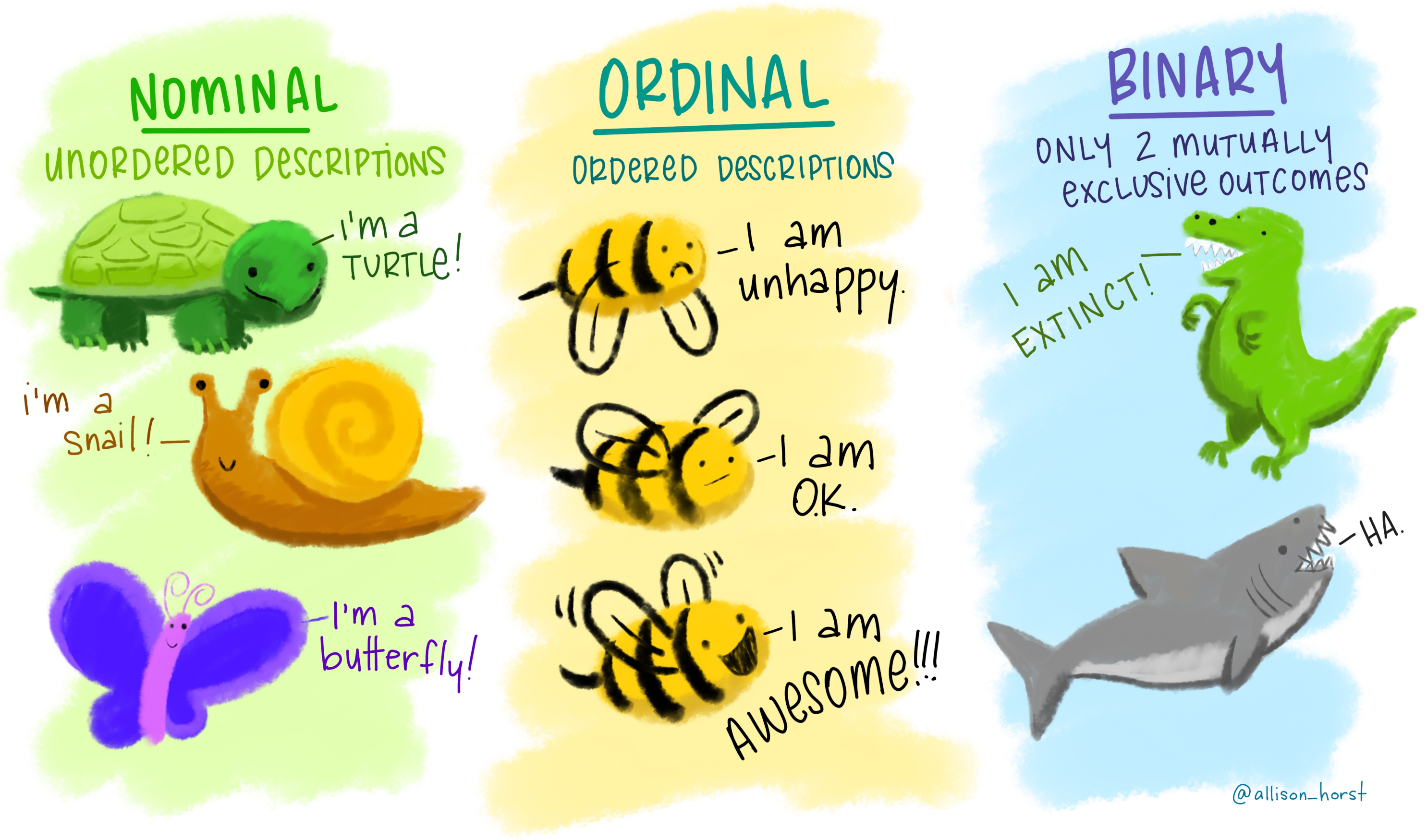
Distinguishing between the later is needed for

- correct model choice,
- proper graphical representation,
- interpretation of effects.

Numerical variables

Drawing by Allison Horst of continuous (left) and discrete (right) numerical variables.

Categorical variables



Drawing by Allison Horst of nominal (left), ordinal (middle) and binary (right) categorical variables.

Graphics and data

> *The simple graph has brought more information to the data analyst's mind than any other device.*
>
> — John Tukey

What is a good graph?

> *Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency … Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.*
>
> — Tufte, 1983

Grammar of graphics

> Wilkinson, L. (2005), *The Grammar of Graphics(2nd ed.)* Statistics and Computing, New York: Springer.

- Elements (layers):
  - data
  - mapping (variables -> aesthetics)
  - geometric objects
  - transformations
  - positioning
- Scale / guide
- Coordinates (facets, coordinate system)

Rule 1: the choice of graphic depends on the variable type

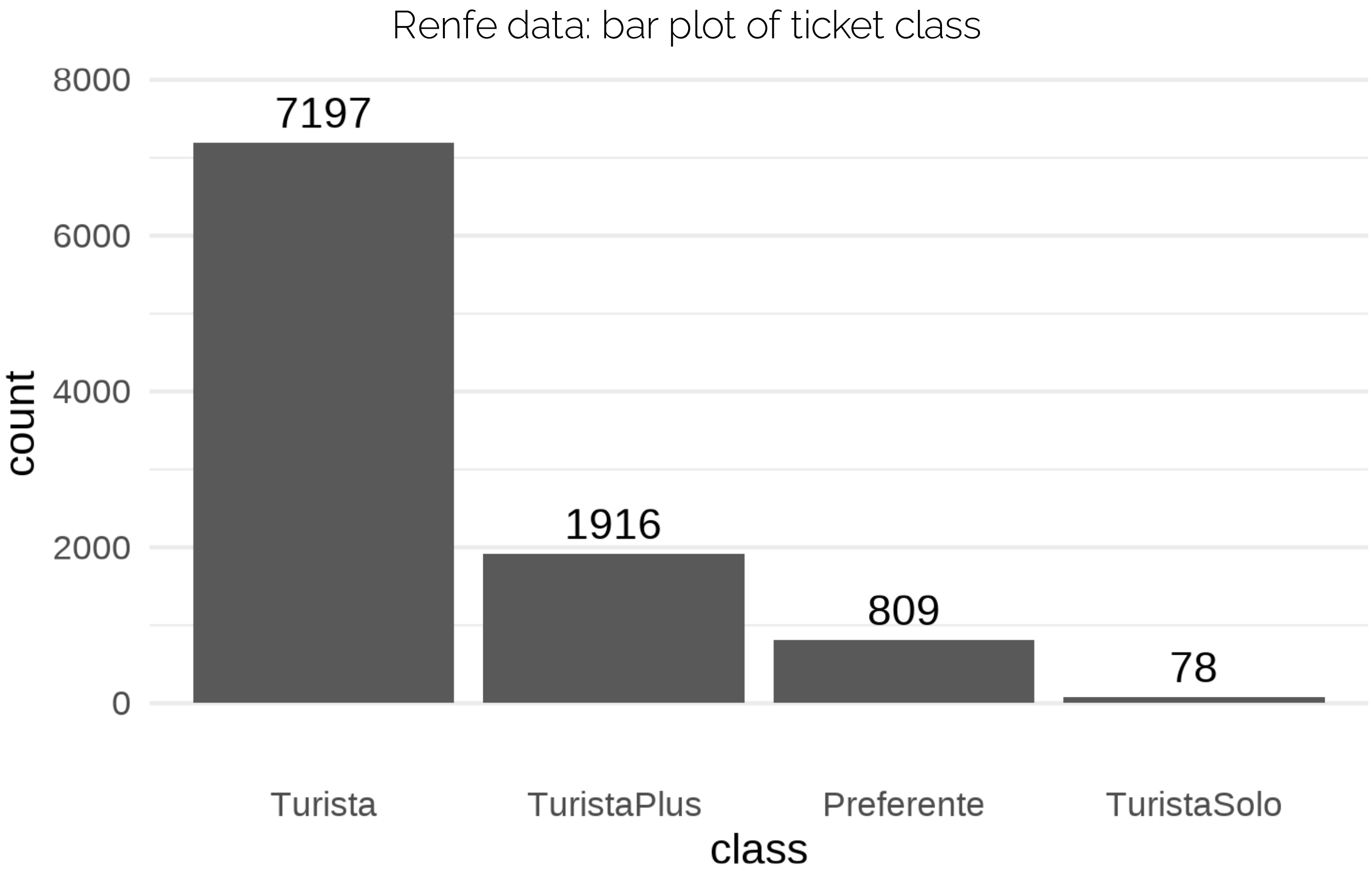
One variable

- continuous: histogram/density plot
- discrete: bar plot
- categorical: bar plot (frequency or percentage)

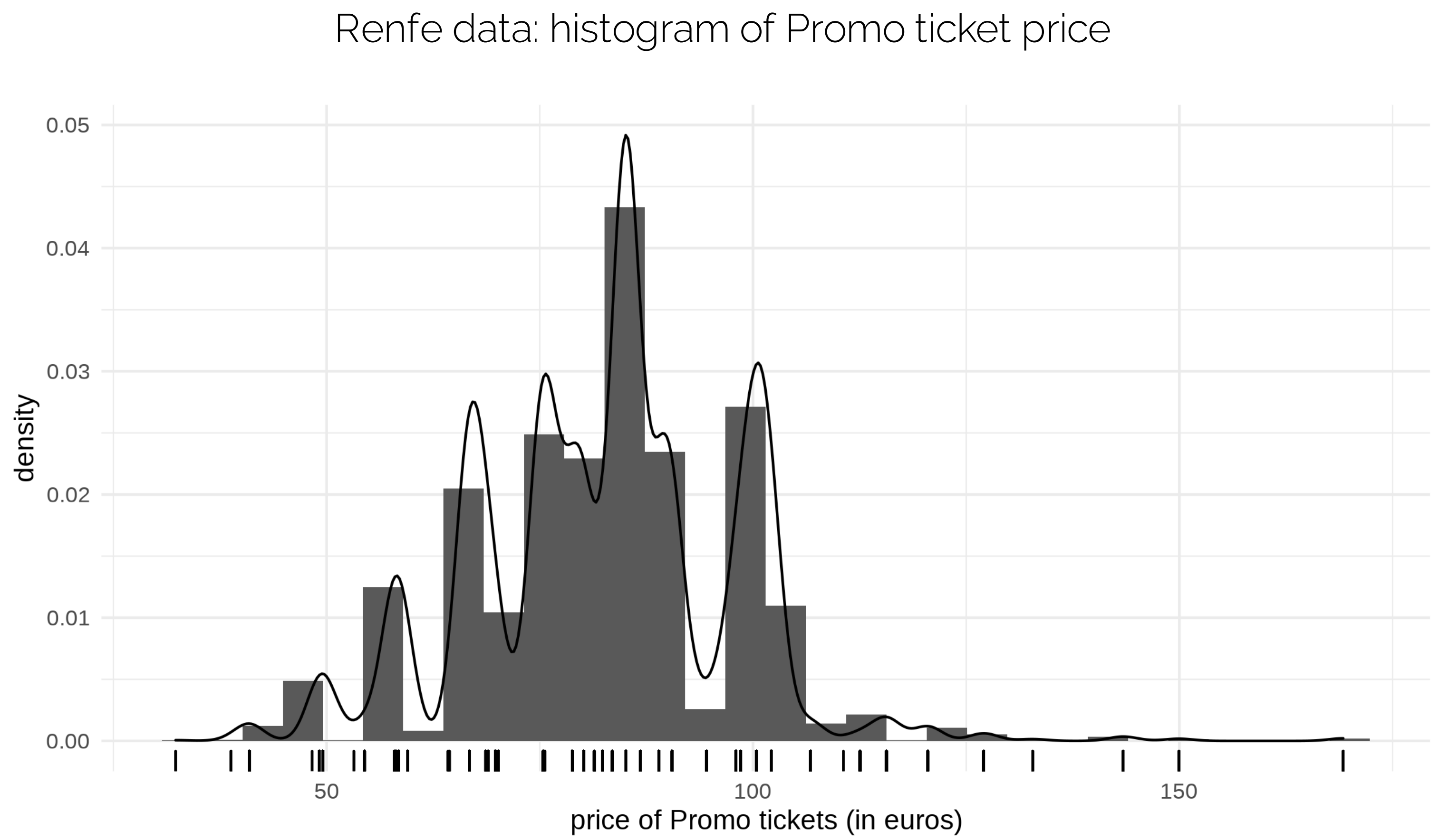
Two variables

- continuous: scatterplot
- categorical: bar plots (one group via color), heatmap
- continuous × categorical: box-and-whisker plot, violin plot

+ R graph + R code + SAS graph + SAS code

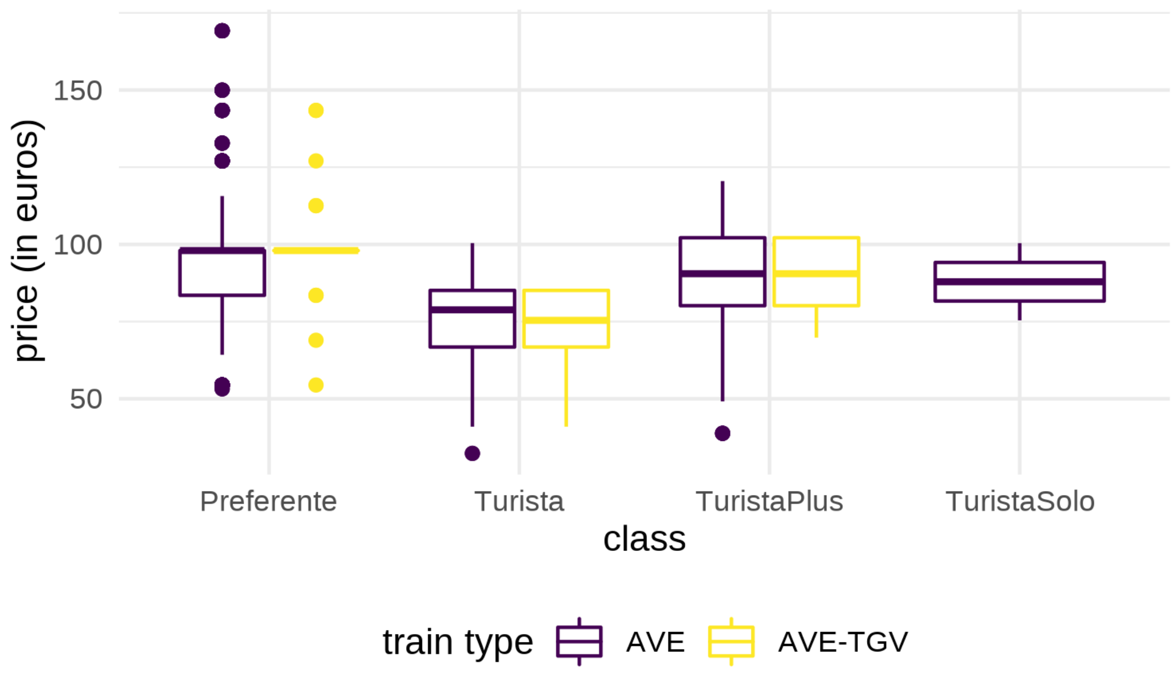


+ R graph + R code + SAS graph + SAS code



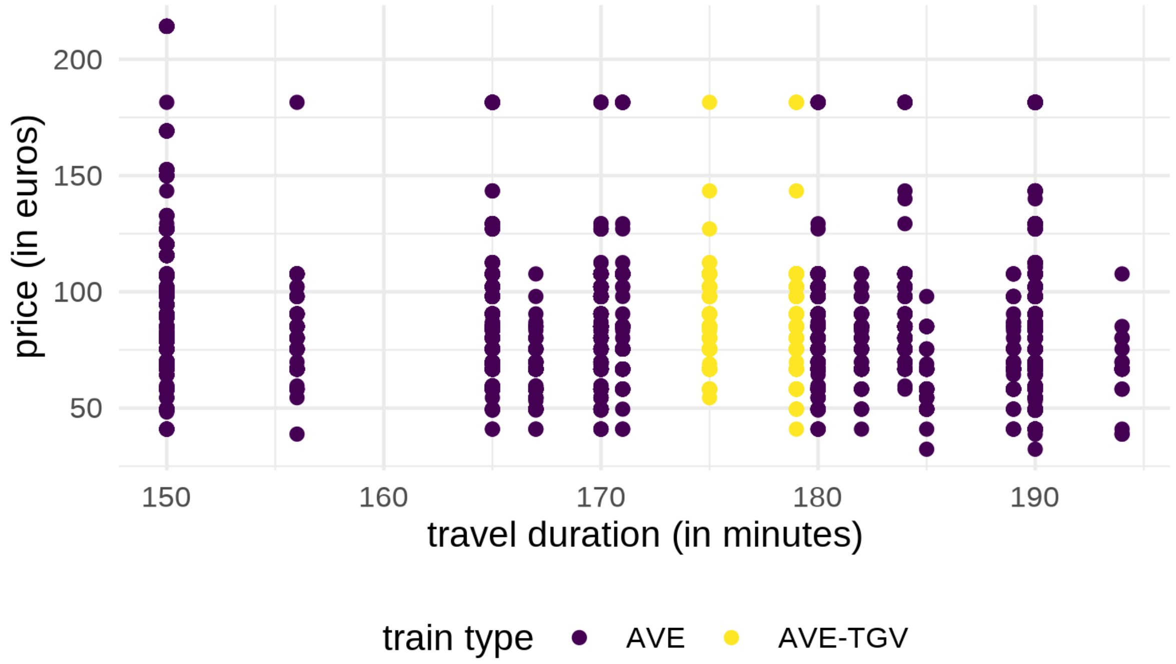
+ R graph + R code + SAS graph + SAS code

Renfe data: box-and-whiskers plot of Promo tickets price as a function of class



+ R graph + R code + SAS graph + SAS code

Renfe data: scatterplot of ticket price as a function of travel time for high speed trains



Rule 2: A graph tells a story by itself

Your graphic must be standalone with the legend.

- some visualization choices are more effective/adequate than others
- include both variable name **and** units if ambiguous
- add a description in the text and cross-reference
- pay attention to scale (adequate font size for legibility)

Rule 3: pay attention to human visual perception

Avoid junk chart

- ratio length/width
- spacing between bands
- axis limits (with or without zero)
- choice of color (grayscale, color-blind friendly palettes)
- comparing areas is difficult
- avoid 3D graphs / rotation

Graphical exploratory data analysis

> *Numerical quantities focus on expected values, graphical summaries on unexpected values.*
>
> — John Tukey

- Ask questions related to the data
- Look for answers using graphs
- Infirm or confirm your intuitions
- Refine questions based on preliminary findings
- Rince and repeat
- Write a summary of key findings

Discussion

Perform and exploratory data analysis of the `diamonds` dataset in small groups.

Questions

-

Summary

-

References

- *Fundamentals of Data Visualization* par Claus O. Wilke
- Chapter 3 of ***R*** *for Data Science* by Garrett Grolemund and Hadley Wickham
- Chapter 1 of *Data Visualization: A practical introduction* by Kieran Healy