Modélisation statistique

#1.c Analyse exploratoire des données

Dr. Léo Belzile
HEC Montréal

Types de données

Vos base de données incluent plusieurs *types de variable*.

Il faut faire la distinction entre ces dernières

- pour la modélisation,
- pour la représentation graphique,
- pour l'interprétation adéquate des effets.

Types de données numériques

Illustration par Allison Horst de variables numériques continues (gauche) et discrètes (droite).

Types de données catégorielles

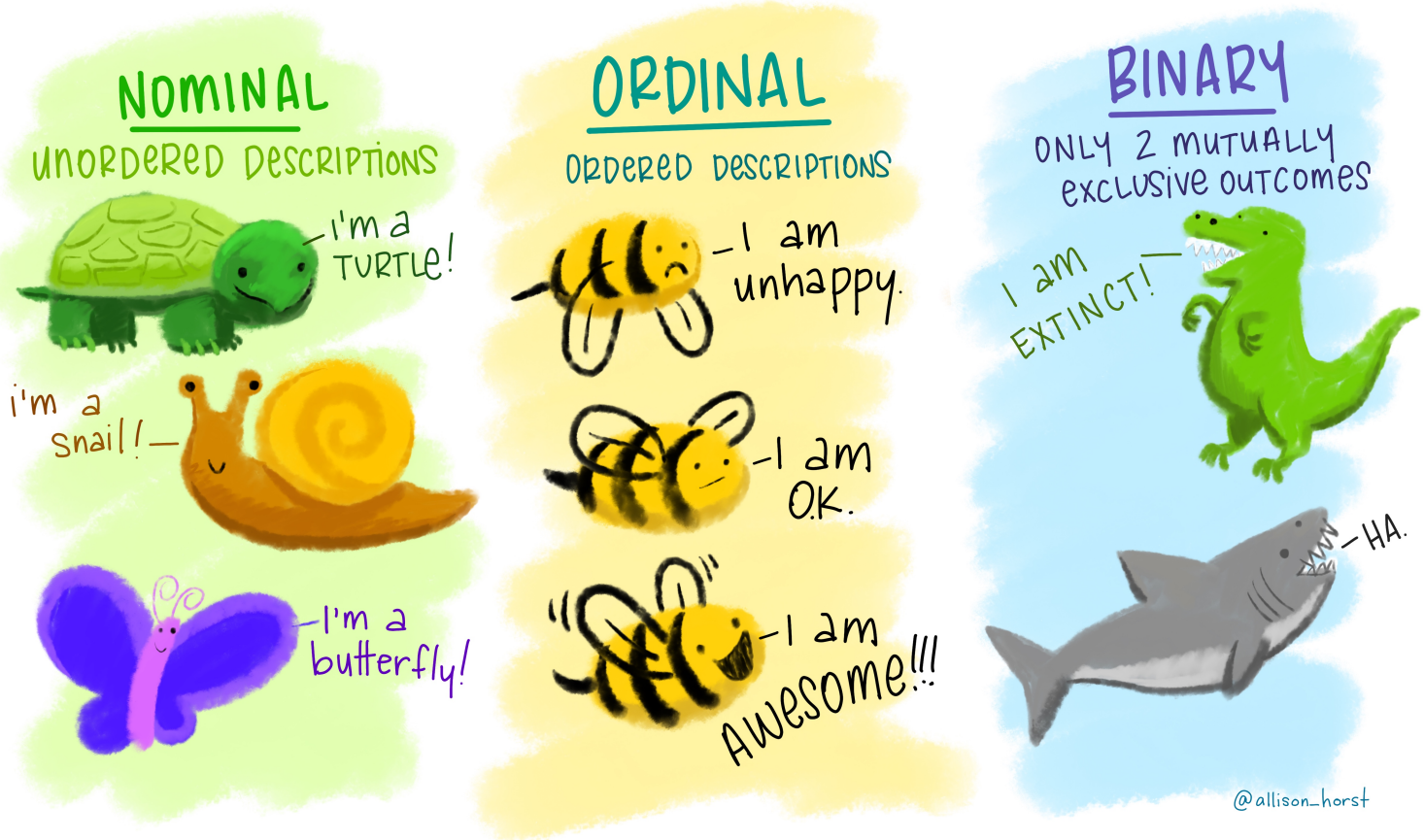


Illustration par Allison Horst de variables catégorielles nominales (gauche), ordinales (centre) et binaires (droite).

Graphiques et données

> *Un simple graphique transmet plus d'information à l'analyste que n'importe quel autre option*
>
> — John Tukey

Qu'est ce qu'un bon graphique?

> *communique des idées complexes avec clarté, précision et efficacité ... [le graphique] qui offre au lecteur le plus grand nombre d'idées le plus rapidement possible avec le moins d'encre et le plus petit espace possible*
>
> — Tufte, 1983

Grammaire des graphiques

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

- Éléments (couches):
  - données
  - application (variable -> esthétique)
  - objets géométriques
  - transformations
  - positionnement
- Échelle / guide
- Coordonnées (facettes, système de coordonnés)

Voici quelques règles d'or pour une visualisation effective

Règle 1: le choix du graphique dépend du type de variable

Une seule variable

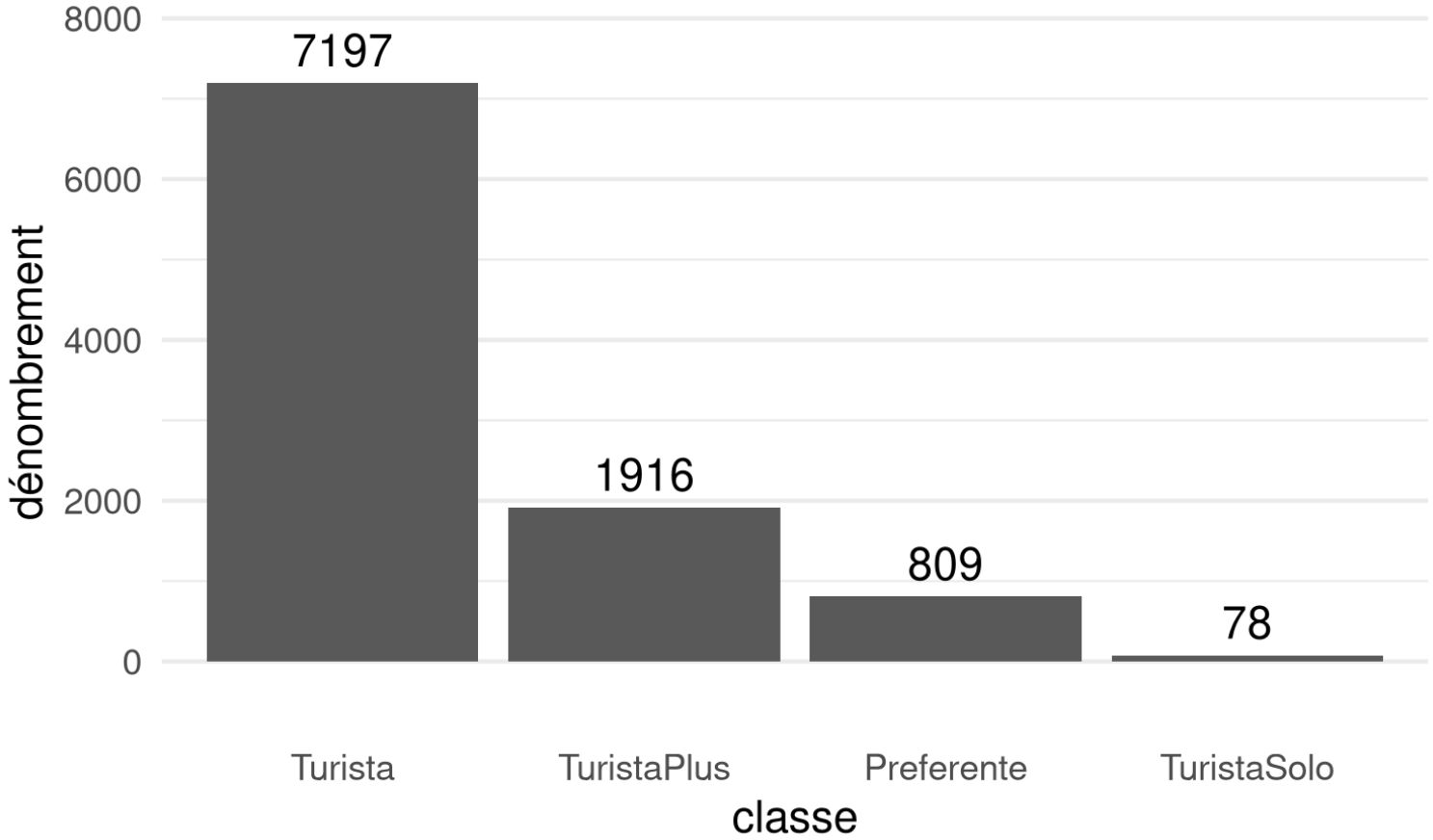
- continue: histogramme/densité
- discrète: diagramme en bâton
- catégorielle: diagramme en bâton (fréquence ou pourcentage)

Deux variables

- continues: nuage de points
- catégorielles: diagramme à bande (avec couleurs), carte thermique
- continue $\times$ catégorielle: boîte à moustache, graphique violon

Graphiques R ✚ Code R ✚ Graphique SAS ✚ Code SAS

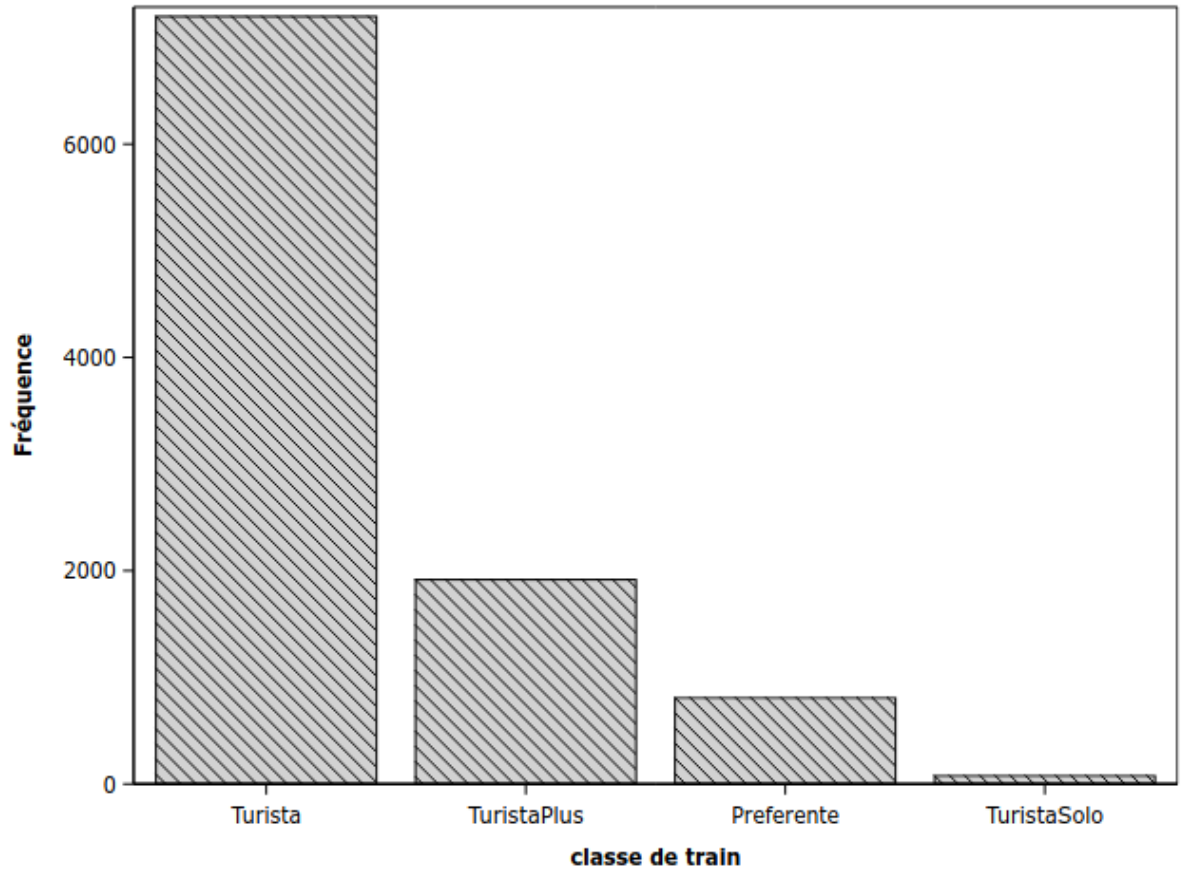
Diagramme en bâtons pour la classe des billets de trains du jeu de données Renfe



Une seule variable catégorielle: diagramme en bâton

```
ggplot(data = renfe,
       aes(x = forcats::fct_infreq(classe))) +
  geom_bar() +
  geom_text(stat='count', aes(label=..count..), vjust=-0.5) +
  labs(x = "classe",
       y = "dénombrement")  +
  scale_y_continuous(expand = c(.125, 0)) +
  theme(panel.grid.major.x = element_blank())
```

- On ordonne les valeurs selon la fréquence.
- Si les étiquettes sont trop longues, faites une rotation via `+ coord_flip()`.

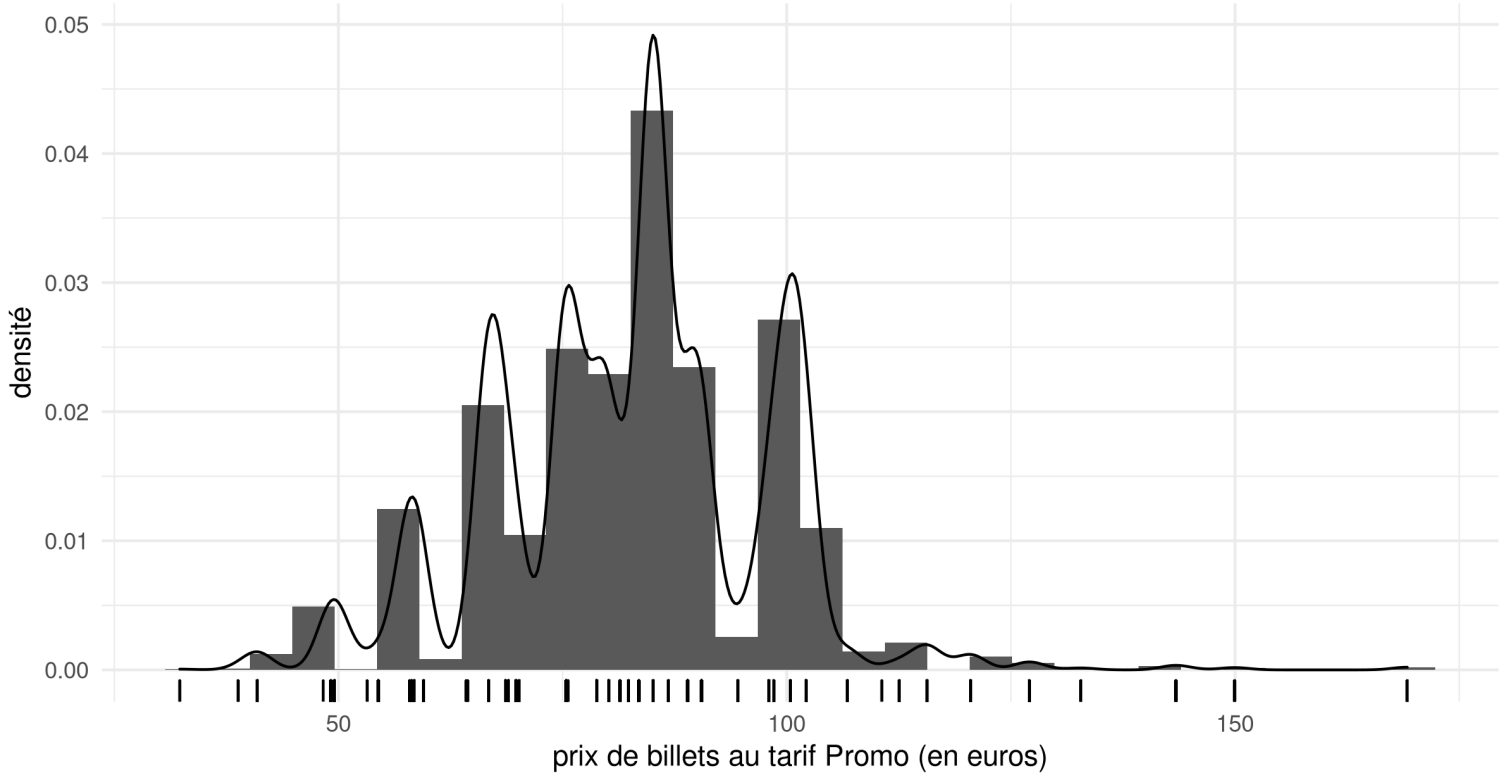

6000
4000
2000
0
Fréquence
Turista
TuristaPlus
Preferente
TuristaSolo
classe de train

- Graphiques R ✚ Code R ✚ Graphique SAS ✚ Code SAS

```
proc sgplot data=modstat.renfe;
vbar classe / categoryorder=respdesc;
xaxis label="classe de train";
run;
```

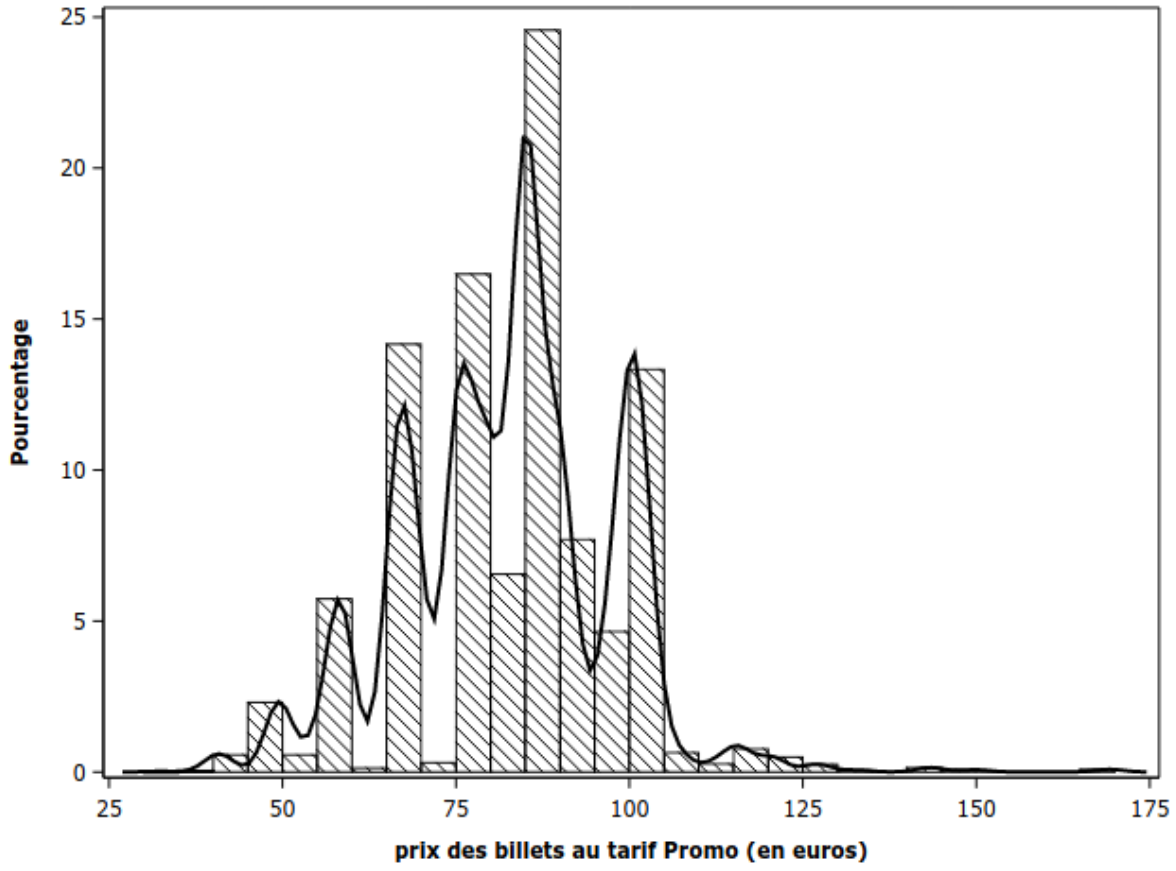
Graphiques R ✚ Code R ✚ Graphique SAS ✚ Code SAS

Histogramme du prix des billets au tarif Promo de trains du jeu de données Renfe



Une seule variable continue: histogramme et/ou densité

```
renfe %>% subset(tarif == "Promo") %>%
  ggplot(aes(x = prix)) +
    geom_histogram(aes(y = ..density..), bins = 30) +
    geom_density() +
    geom_rug(sides = "b") +
    labs(x = "prix de billets au tarif Promo (en euros)",
         y = "densité")
```


Pourcentage
25
20
15
10
5
0
25
50
75
100
125
150
175
prix des billets au tarif Promo (en euros)

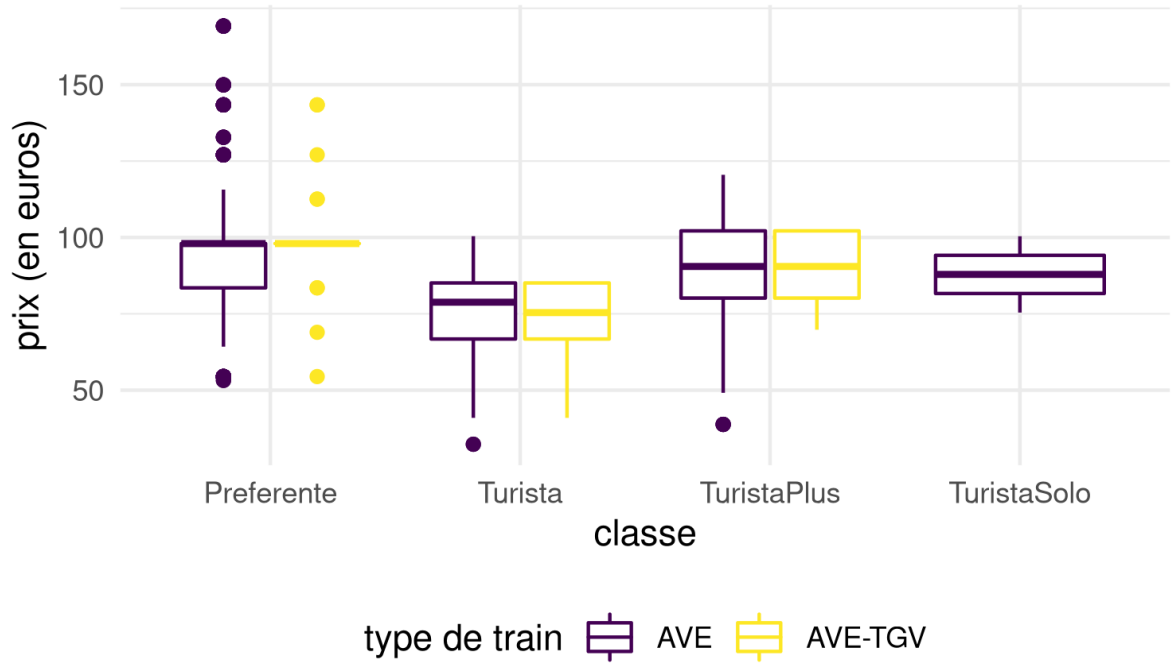
- Graphiques R ✚ Code R ✚ Graphique SAS ✚ Code SAS

```
data renfe_promo;
set modstat.renfe;
where tarif ="Promo";
run;

proc sgplot data=renfe_promo noautolegend;
histogram prix;
density prix / type=kernel;
xaxis label = "prix des billets au tarif Promo (en euros)";
run;
```

Graphiques R ✚ Code R ✚ Graphique SAS ✚ Code SAS

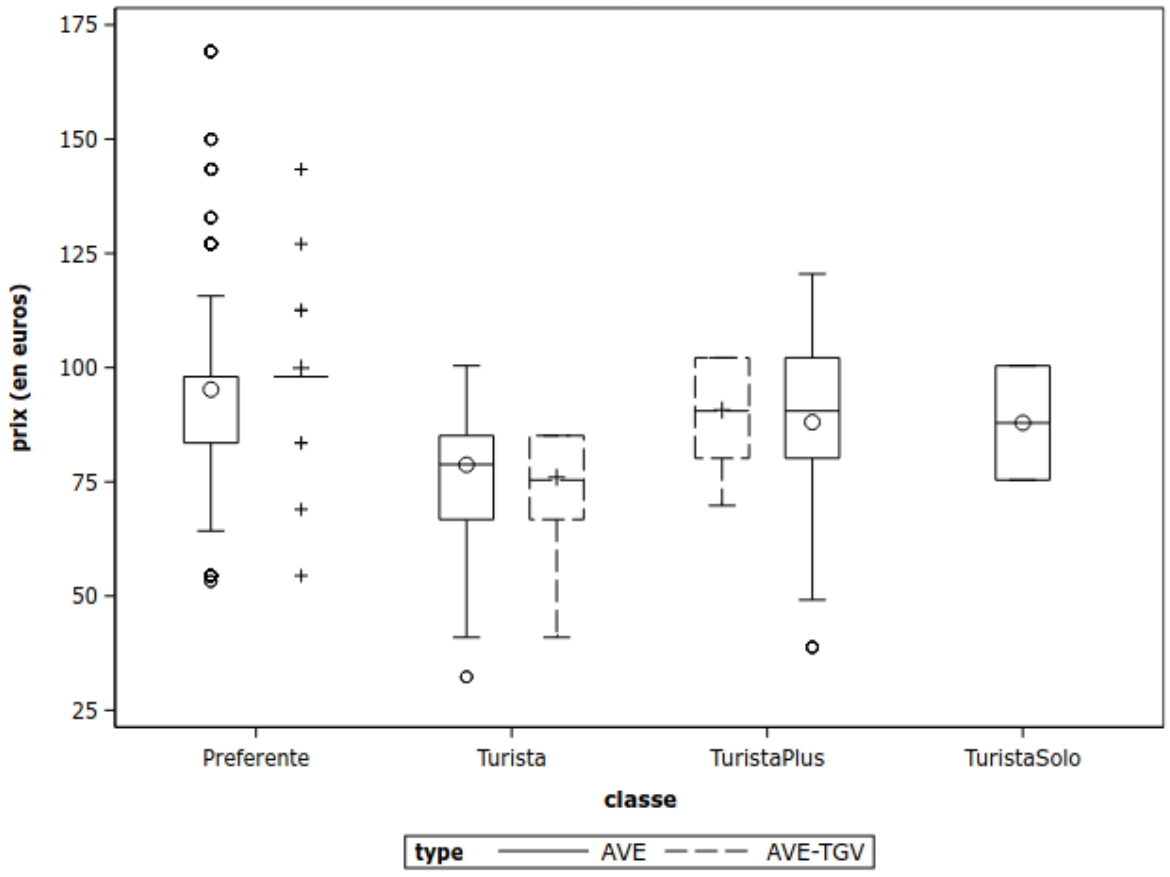
Boîte à moustache du prix des billets au tarif Promo en fonction de la classe pour le jeu de données Renfe



Deux variables (continue et catégorielle): boîte à moustache

```
renfe %>% subset(tarif == "Promo") %>%
    ggplot(aes(y = prix, x = classe, col = type)) +
    geom_boxplot() +
    labs(y = "prix (en euros)", col = "type de train") +
    theme(legend.position = "bottom") +
    scale_colour_viridis_d()
```

- On ajoute une autre variable catégorielle (`type`) à l'aide de la couleur.
- On utilise une palette de couleurs adéquate (daltonisme, impression noir et blanc).

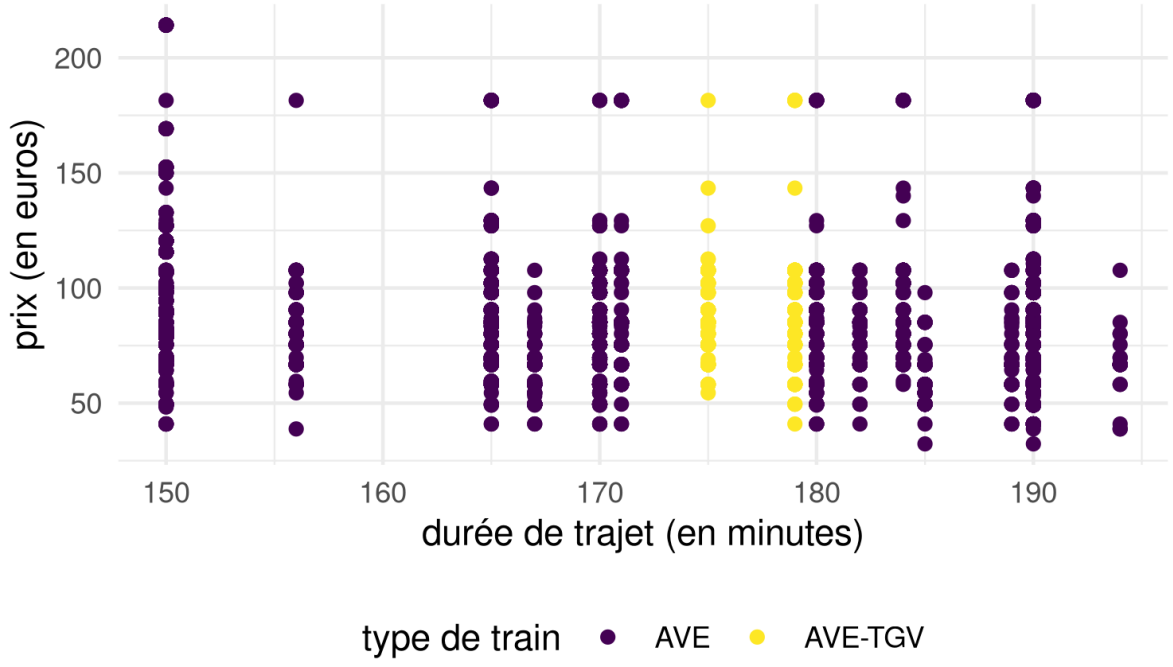

175
150
125
100
75
50
25
prix (en euros)
Preferente
Turista
TuristaPlus
TuristaSolo
classe
type
AVE
AVE-TGV

- Graphiques R ✚ Code R ✚ Graphique SAS ✚ Code SAS

```
proc sgplot data=renfe_promo;
vbox prix / category=classe group=type;
yaxis label = "prix (en euros)";
run;
```

Graphiques R + Code R + Graphique SAS + Code SAS

Nuage de points du prix en fonction du temps de trajet annoncé pour les billets de train à grande vitesse du jeu de données Renfe

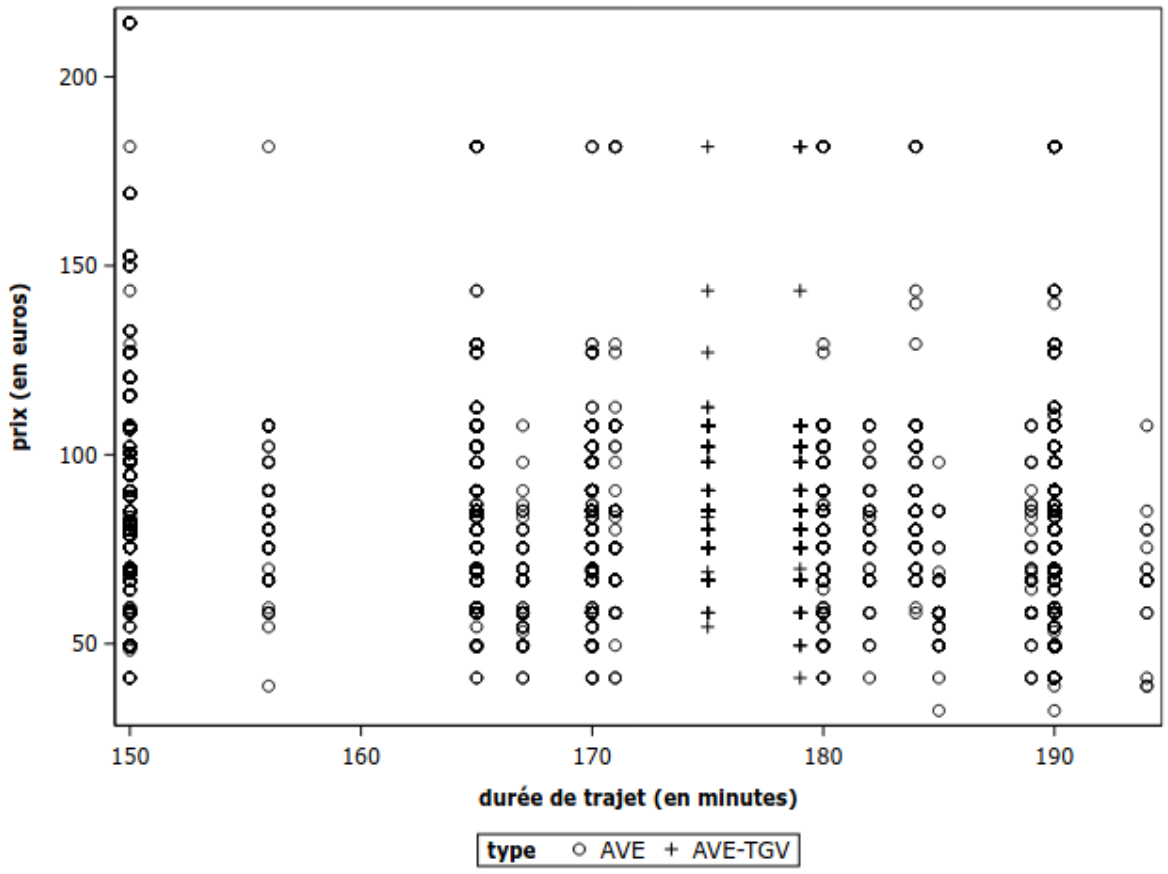


Deux variables (continues) et une variable catégorielle: nuage de points

```
renfe %>% subset(type != "REXPRESS") %>%
    ggplot(aes(x = duree, y = prix, col = type)) +
    geom_point() +
    labs(y = "prix (en euros)",
         x = "durée de trajet (en minutes)",
         col = "type de train") +
    theme(legend.position = "bottom") +
    scale_colour_viridis_d()
```

Qu'est-ce qui cloche dans la représentation graphique précédente?

Comment pourrait-on remédier aux problèmes soulevés?


prix (en euros)
200
150
100
50
150
160
170
180
190
durée de trajet (en minutes)
type
AVE
AVE-TGV

- Graphiques R ✚ Code R ✚ Graphique SAS ✚ Code SAS

```
data renfe_ave;
set modstat.renfe;
where type NE "REXPRESS";
run;

proc sgplot data=renfe_ave;
scatter y=prix x=duree / group=type;
xaxis label="durée de trajet (en minutes)";
yaxis label="prix (en euros)";
run;
```

Règle 2: soignez les apparences

Votre graphique doit être interprétable uniquement avec la légende.

- certaines visualisations sont plus effectives/adéquates que d'autres
- inclure les noms de variables **et** les unités
- ajouter une description dans le texte et faire une référence croisée
- attention à la lisibilité (taille de police adéquate)

Règle 3: Portez une attention particulière à la perception visuelle humaine

- ratio longueur/largeur
- espace entre bandes
- étendu des axes (incluant ou pas zéro)
- choix de couleurs (noir/blanc avec contraste, palette pour daltoniens)
- comparaison d'aires/superficies (difficile)
- graphiques 3D / avec rotation: à éviter

Analyse exploratoire graphique des données

> *Les résumés numériques focalisent l'attention sur les valeurs attendues, les résumés graphiques sur les valeurs inattendues.*
>
> — John Tukey

- Poser des questions en lien avec les données
- Chercher les réponses à l'aide de graphiques
- Infirmer/confirmer nos intuitions
- Raffiner les questions suite aux observations
- Répéter le processus
- Écrire un résumé des trouvailles et aspects importants

Références

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