

Outlier detection using R language

Soumitra Kar (ZSUK)

Senior Statistician

Novo Nordisk

Disclaimer

Views expressed in the presentation by the presenter are the views, thoughts, and opinions belonging solely to the presenter, and not necessarily to the presenter's employer, organization, committee or other group or individual.

This presentation is intended to exhibit the ways of detecting outliers and beyond that using R language through an open source package ["trelliscopejs"](#) available in cran, hence no data used from Novo Nordisk. The data is simulated for the purpose of exploration.

Agenda

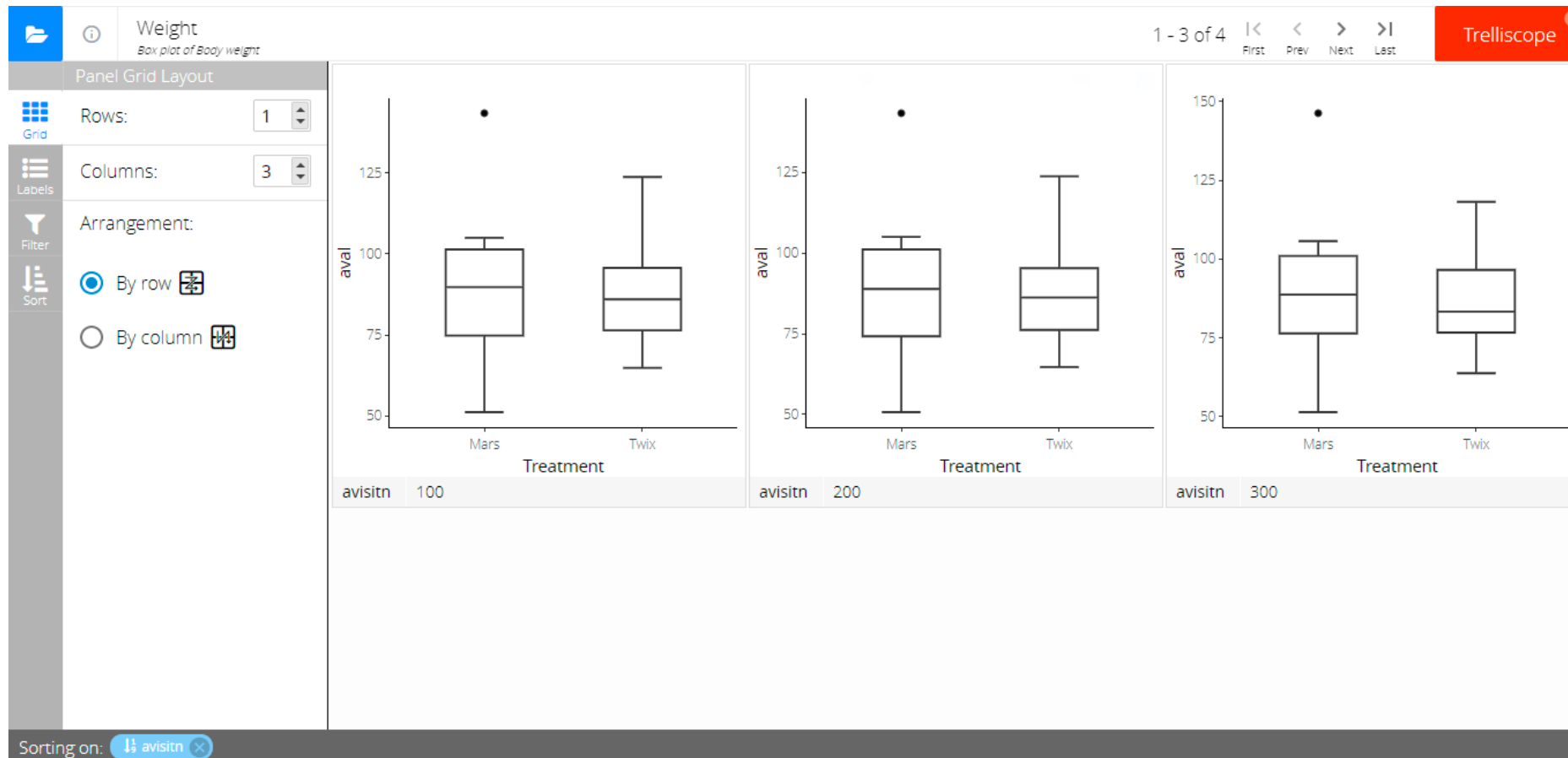
- Cause and importance of Outliers
- Outlier detection using R
- Controls in trelliscopejs & further use
- plotly with trelliscopejs
- Scope of trelliscopejs
- Reference

Cause and importance of outliers

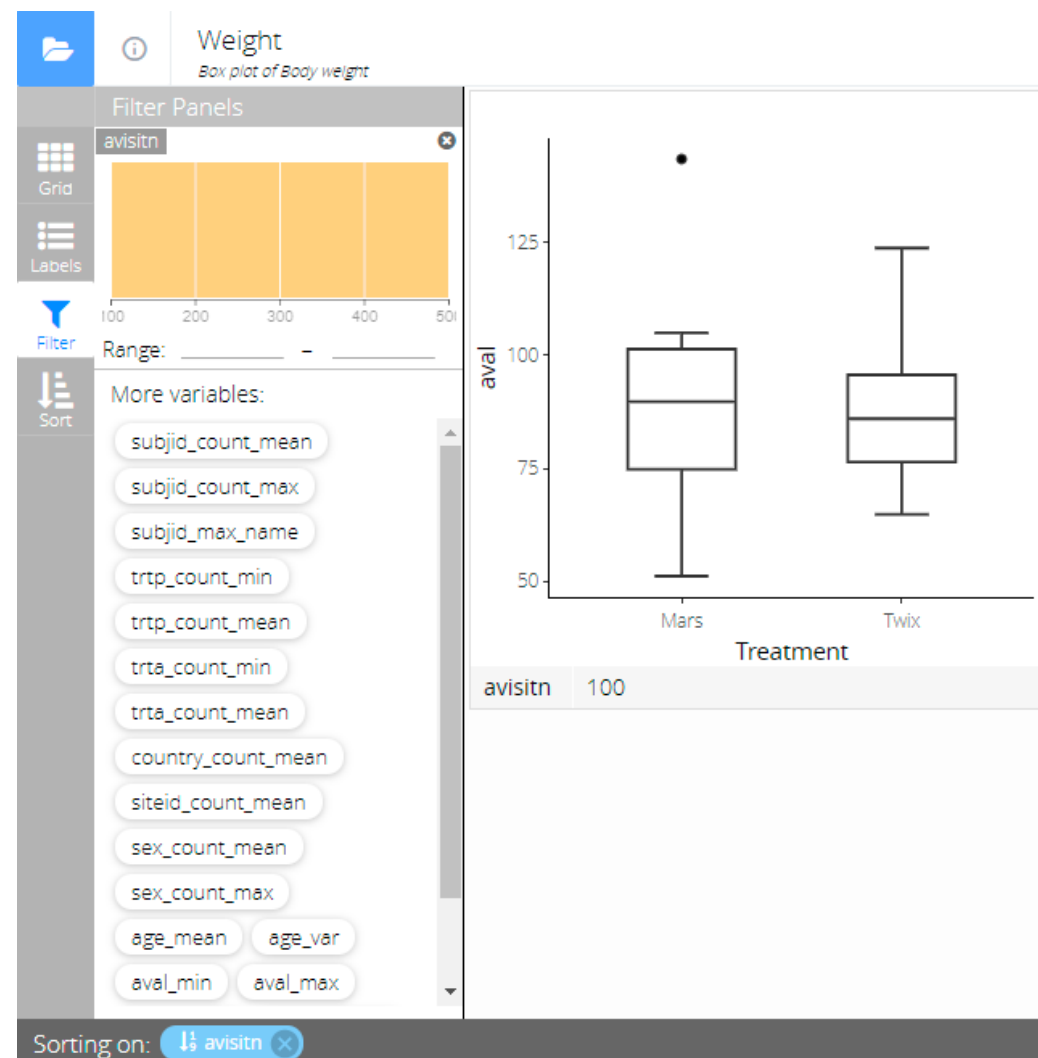
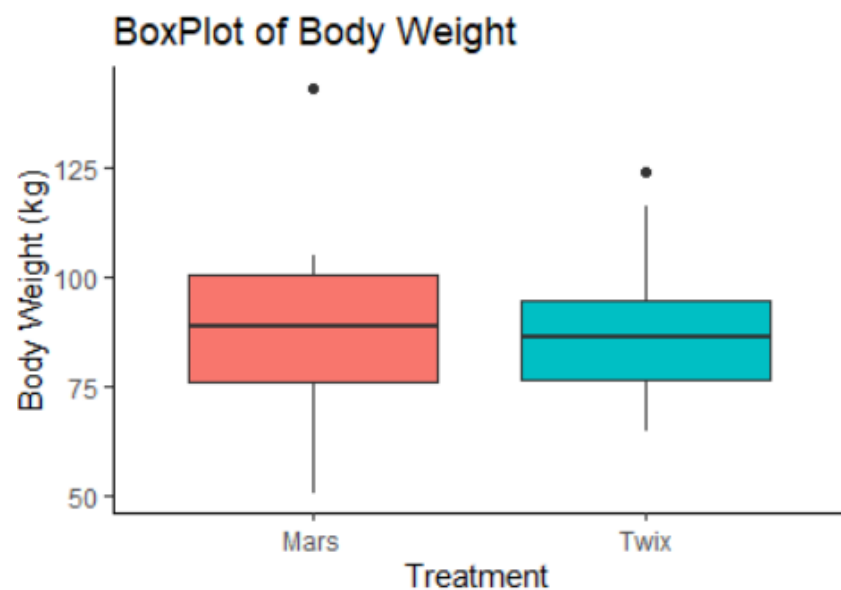
- Suppose we run a multivariate experiment in a business setting and we end up excluding top 5% spenders being outliers. But if we tactfully consider it, we may get some insights about their spending behavior and boost bottom line of a result.
- We cannot remove outliers from clinical studies directly. We should listen to what the outliers are telling us. Without understanding the cause we can never delete the outliers.
- As we are working in biosciences where processes of multiplicative nature are common we can never assure the data will absolutely follow a symmetric distribution and hence we should not automatically delete to symmetrize it. We may throw away the essence of the data.
- We should investigate the data thoroughly to find underlying cause of outliers, and its important in such cases to look at different parameters subject wise. Hence, graphical method of investigating the data is of great use in such cases.
- R brings an advantage to us to easily visualize the data in a way that it helps to understand the trend in the data for different parameters interactively.

Outlier detection in R

- Multiple packages available for visualizations in R but focus for this presentation will be on trelliscopejs (Available in cran)
- Brings **faceted visualizations** with the help of 'ggplot2' or 'tidyverse'
- Advantages : Cost effective, interactive interface and flexible to adapt to different purposes.



Outlier detection in R



Outlier detection in R

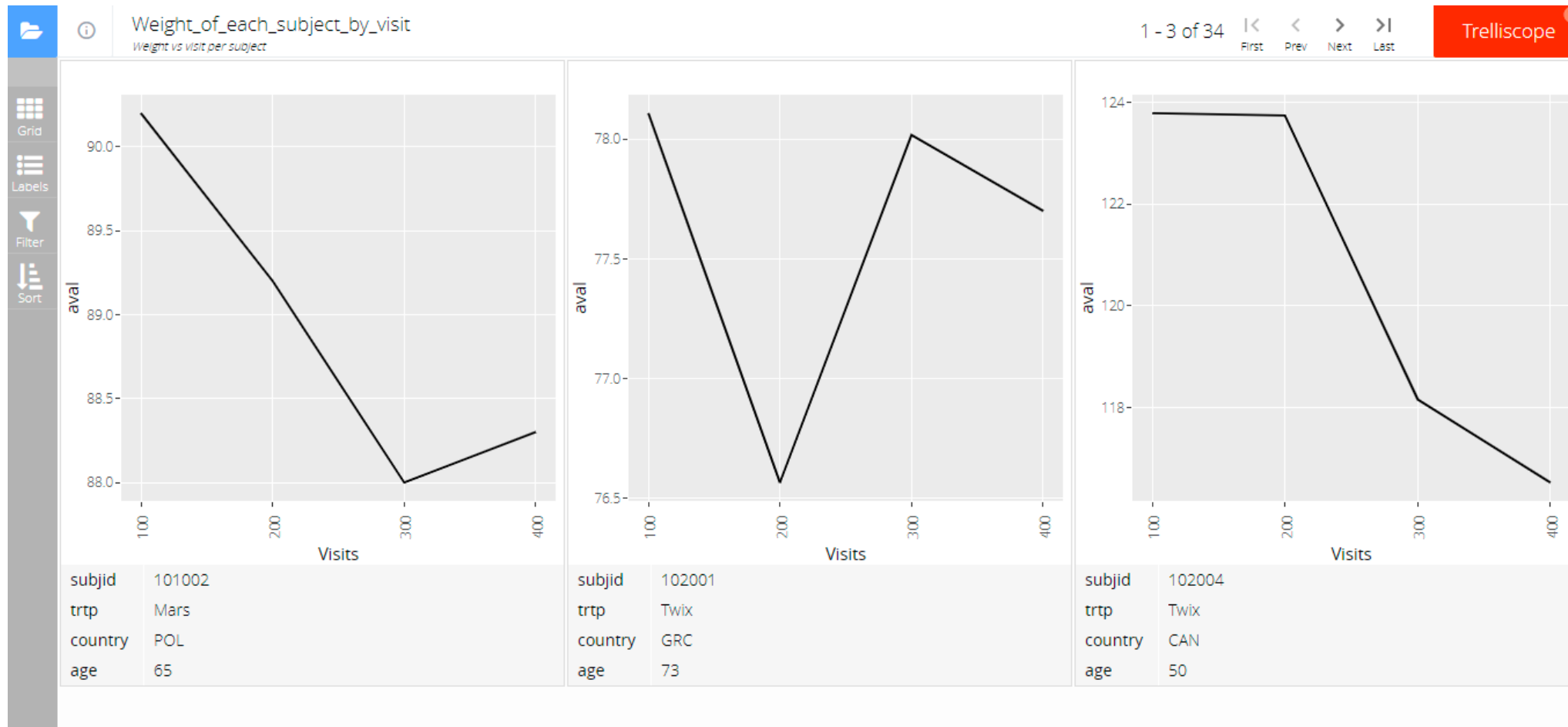
Code

```
# ggplot2 + trelliscope + plotly
ggplot(data=advs,aes(trtp,aval))+
  geom_boxplot()+
  theme(axis.text.x = element_text(angle=90,vjust=0.5))+
  scale_x_discrete("Treatment",labels=as.character(advs$trtp),breaks=advs$trtp)+
  facet_trelliscope(~avisitn,scales = "free",
                    name="Weight",
                    desc="Box plot of Body weight",
                    nrow=1,
                    ncol=3,path = "~/trelliscope",
                    as_plotly = TRUE
  )
```

```
# ggplot2 + trelliscope + plotly
ggplot(data=advs,aes(avisitn,aval))+
  geom_line()+
  theme(axis.text.x = element_text(angle=90,vjust=0.5))+
  scale_x_continuous("Visits",labels=as.character(advs$avisitn),breaks=advs$avisitn)+
  facet_trelliscope(~subjid,scales = "free",
                    name="Weight of each subject by visit",
                    desc="Weight vs visit per subject",
                    nrow=1,
                    ncol=3,path = "~/trelliscope",
                    as_plotly = TRUE
  )
```

Controls in trelliscopejs & further use

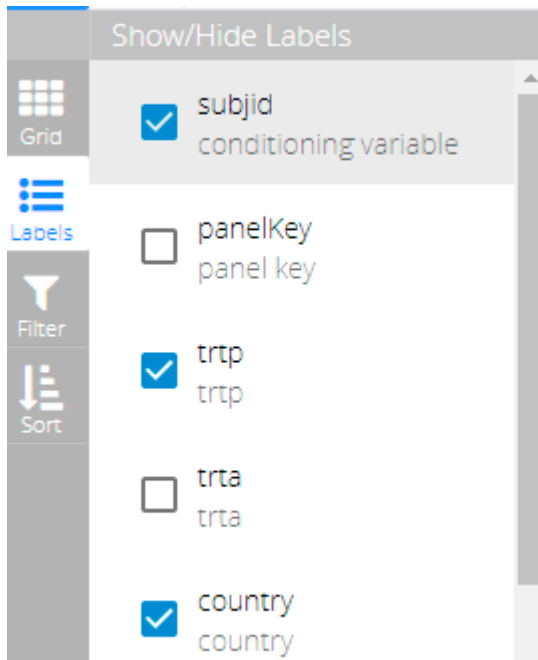
- Helps in further exploration to detect irregularities in some characteristics of the subjects such as country, age, actual treatment given, sex etc.



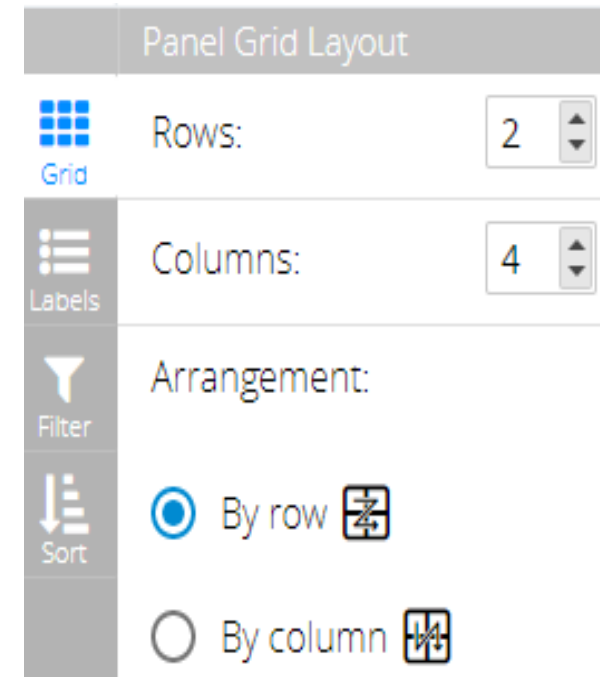
Controls in trelliscopejs & further use

Customization of Panel Grid

How many rows and columns



Adding labels



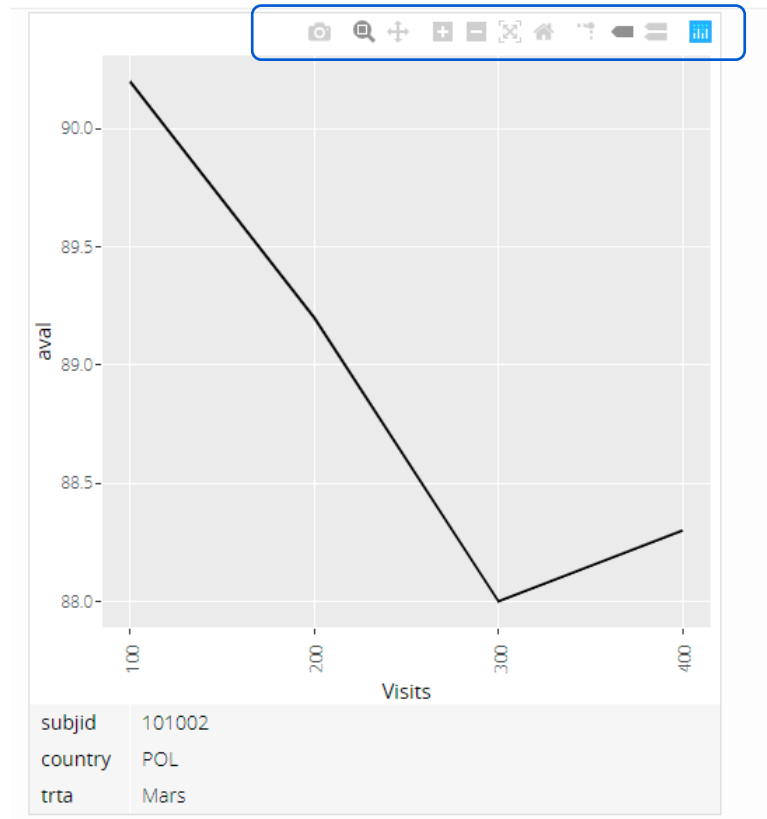
Controls in trelliscopejs & further use

- Sort & filter on any variable. For e.g., to view the data of a specific country or a specific site.



plotly with trelliscopejs

- Makes the visualization procedure interactive.



- Download the image
- Autoscaling
- Zooming in at a particular region
- Individual characteristics

Scope of trelliscopejs

- Detecting irregularities in different parameters on basis of laboratory parameters.
- Chaotic system behavior was discovered by focusing on outliers, not by discarding them.
- For phase 1 clinical studies we can use this, for PK-PD review where we need to investigate the data subject wise.

Reference

- Hafen R, Schloerke B (2020). trelliscopejs: Create Interactive Trelliscope Displays. R package version 0.2.4, URL [CRAN - Package trelliscopejs \(r-project.org\)](https://cran.r-project.org/web/packages/trelliscopejs/index.html)
- Outlier Analysis 2nd Edition, Charu C Agarwal

