Paris Airbnb Analysis

Ruoxian Wu

2024-03-03

Exploratory data analysis (EDA) is used to predict factors that impact a host's likelihood of achieving superhost status. Based on the findings, we can gain insight into characteristics of superhost listings, significance of guest reviews, and pricing strategies.

Introduction

The Parisian Airbnb market is highly popular among tourists from all over the world as a renowned tourist destination. The wide variety of Airbnb listings available in the city offer valuable insights into trends in urban hospitality, pricing dynamics, and the factors that impact guest satisfaction and host performance.

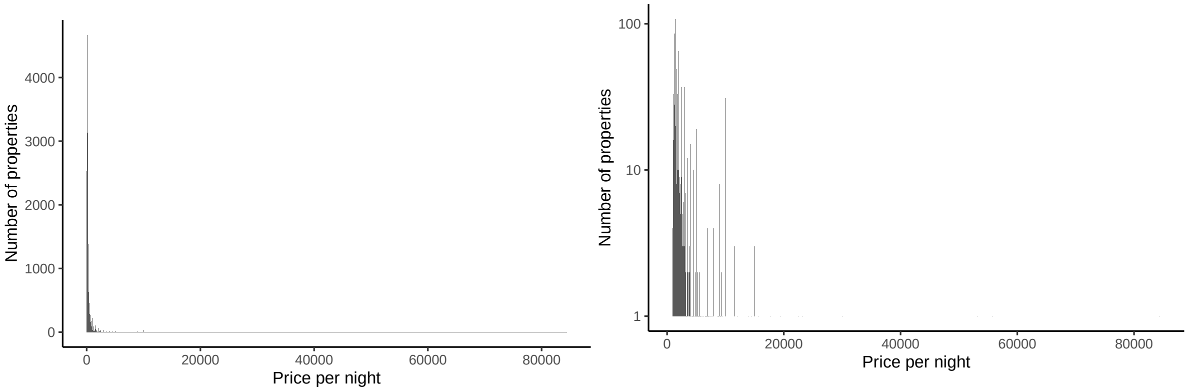
This study uses data from Inside Airbnb as of 12 December 2023, and uses various data processing and analytical techniques to examine the relationship between listing characteristics, host attributes, and market outcomes. The report uses exploratory data analysis (EDA) to explore the distribution of prices across different neighborhoods, identify the traits of highly rated listings, and examine the correlation between superhost status and review scores. Advanced statistical methods, including logistic regression, are used to predict factors that significantly impact a host's likelihood of achieving superhost status.

The report aims to provide valuable recommendations for Airbnb hosts looking to optimize their listings for enhanced profitability and guest experience, while also offering guests a lens through which to make informed accommodation choices in Paris.

Data Analysis

This study involves several stages of data processing. We start the process by acquiring the data, followed by a detailed cleaning and formatting step to ensure that the results are accurate and reliable, including removing non-numeric characters from prices and converting them to integers to make sure they are accurate. The study makes use of various R (R Core Team 2022) packages. Dplyr (Wickham et al. 2023),arrow (Richardson et al. 2024), naniar (Tierney and Cook 2023), janitor (Firke 2021), modelsummary (Arel-Bundock 2022), knitr (Xie 2014), ggplot2 (Wickham 2016), and other tidyverse (Wickham et al. 2019) packages support a wide range of data manipulation, exploration, and visualization tasks for EDA. This EDA focuses on identifying patterns in pricing, the distribution of listings across neighborhoods, and the correlation between review scores and host attributes. The arrow package is utilized for reading and writing data in the Parquet format, enabling efficient data storage and access. The janitor package assists in data cleaning tasks, such as removing duplicate entries and tidying variable names. The knitr package is used for dynamic report generation, allowing for seamless integration of R code and its output into documents. The modelsummary package facilitates the creation of elegant tables summarizing statistical model results. The naniar package offers specialized functions for handling missing data, providing insights into the pattern of missingness.

Result



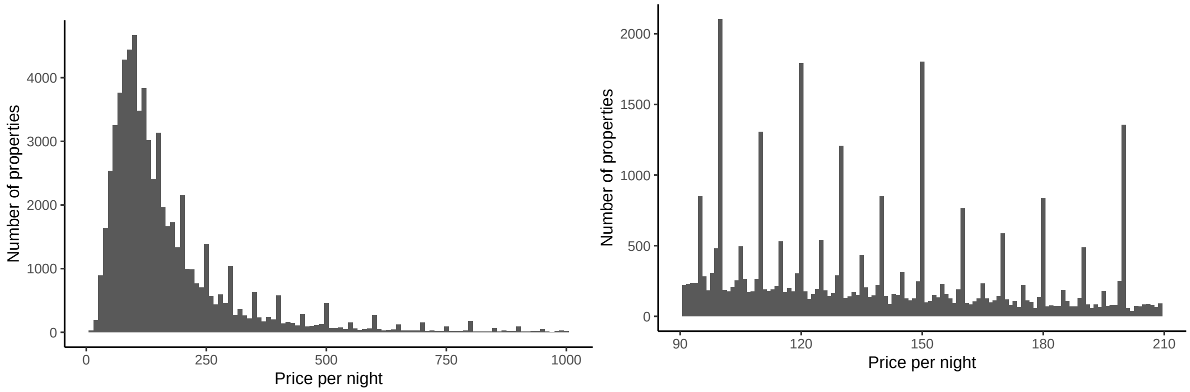
(a) Distribution of price

(b) Loged distribution of prices for prices more than $1,000

Figure 1: Distribution of Airbnb Prices in Paris

Figure 1a illustrates the distribution of nightly rental prices across Airbnb listings in Paris. Figure 1b shows the distribution of prices on the log scale. The histogram highlights the concentration of listings within certain price ranges, providing insights into the affordability and pricing strategies of hosts in the city. The majority of listings appear to cluster around the lower to mid-price range, indicating a competitive market for budget-friendly accommodations.

Figure 2a visualizes the distribution of nightly rental prices for listings priced below 1000 dollars. It reveals the majority of Airbnb listings in Paris are concentrated in the lower to



(a) Distribution of price under $1000

(b) Distribution of prices for prices between $90 and $210

Figure 2: Distribution of Airbnb Prices in Paris under $1000

mid-price ranges, emphasizing the abundance of budget-friendly accommodations available to travelers. Figure 2b narrows the focus to a specific price range, providing a detailed look at how listings are priced within this segment.

Figure 3 shows the frequency of different review scores ratings for Airbnb listings in Paris. Each bar shows the number of properties that received a particular review score rating. This gives a visual representation of the overall quality of accommodations available and helps guests to make informed decisions.

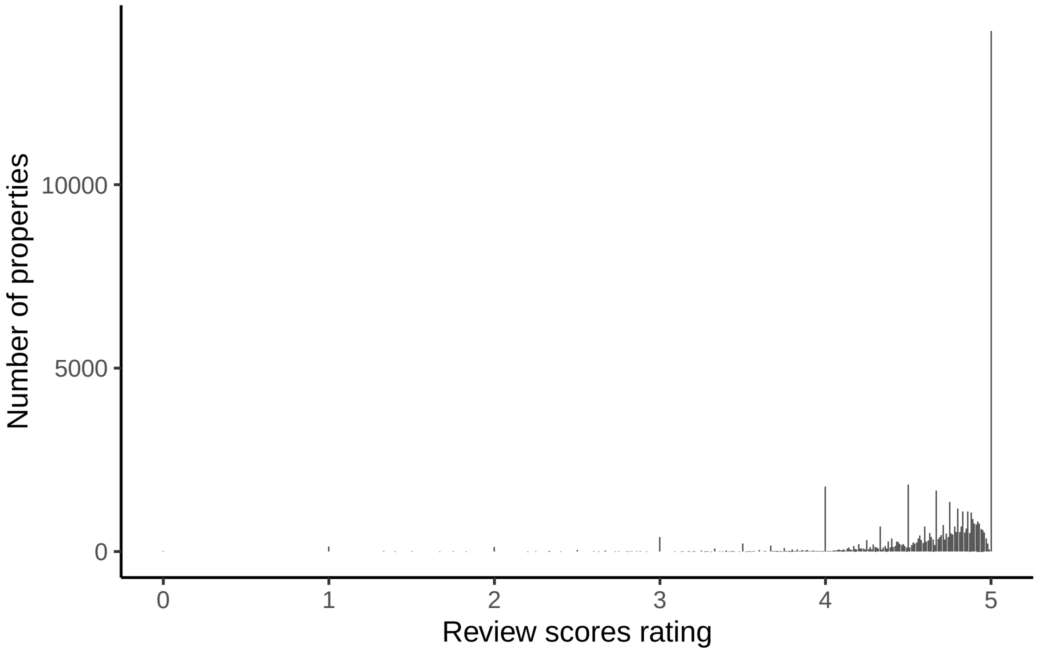


Figure 3: Distribution of review scores for Paris Airbnb listings in December 2023

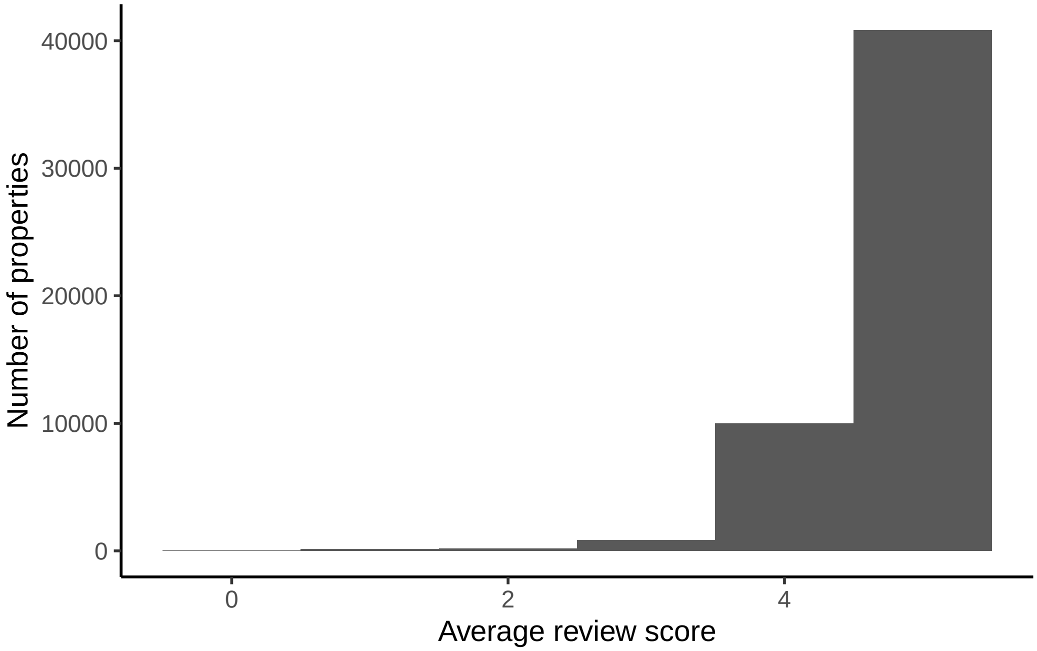


Figure 4: Distribution of review scores for properties with NA response time, for Paris Airbnb rentals in December 2023

Figure 4 details the distribution of average review scores for Airbnb properties in Paris, focusing only on listings with non-missing review ratings. The figure precisely illustrates the range and concentration of review scores, which implies guest satisfaction levels. The figure aims to underline the performance of listings in terms of guest feedback, indicating common ratings.

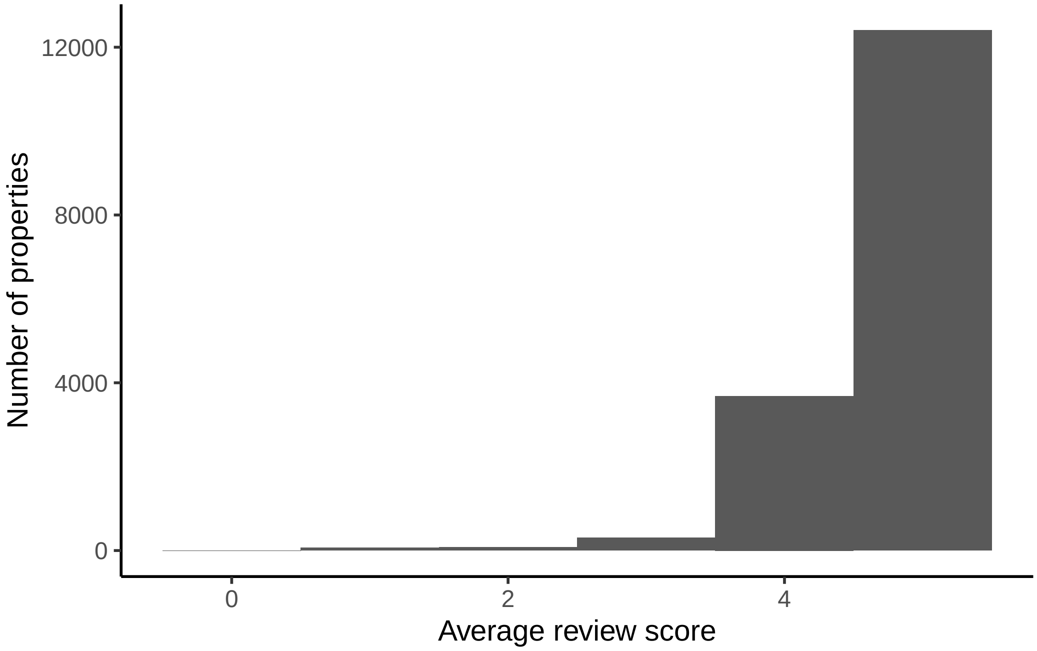


Figure 5: Impact of missing host response time on review scores

Figure 5 examines the distribution of average review scores for properties where the host's response time is missing. This presents a focused analysis of how the absence of responsive communication might correlate with guest satisfaction. The review scores reveal a relationship between host engagement and guest experience. Hosts can learn how to maintain high review scores by communicating promptly.

Figure 6 explores the connection between the accuracy of review scores given by guests and the time it takes for hosts to respond. The scatter plot enhanced with missing data points is used to provide a more complete view of the relationship between these two variables. The x-axis of the plot categorizes listings based on the host's response time, while the y-axis shows the review scores related to accuracy. This allows us to see how these two variables interact and how the absence of data on host responsiveness might affect the accuracy of reviews.

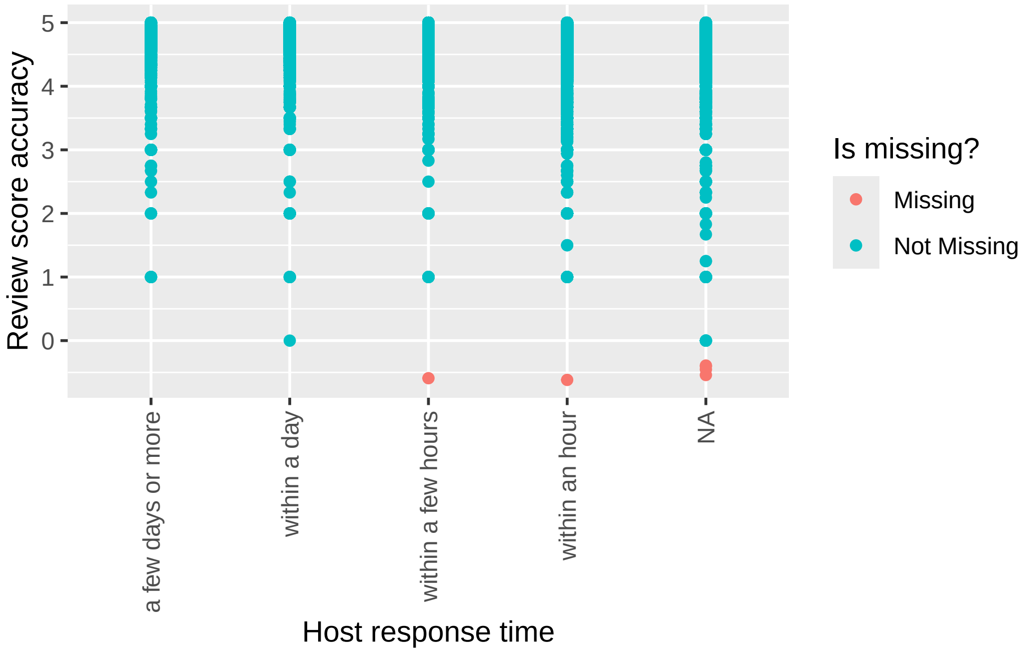


Figure 6: Relationship between host response time and review score accuracy

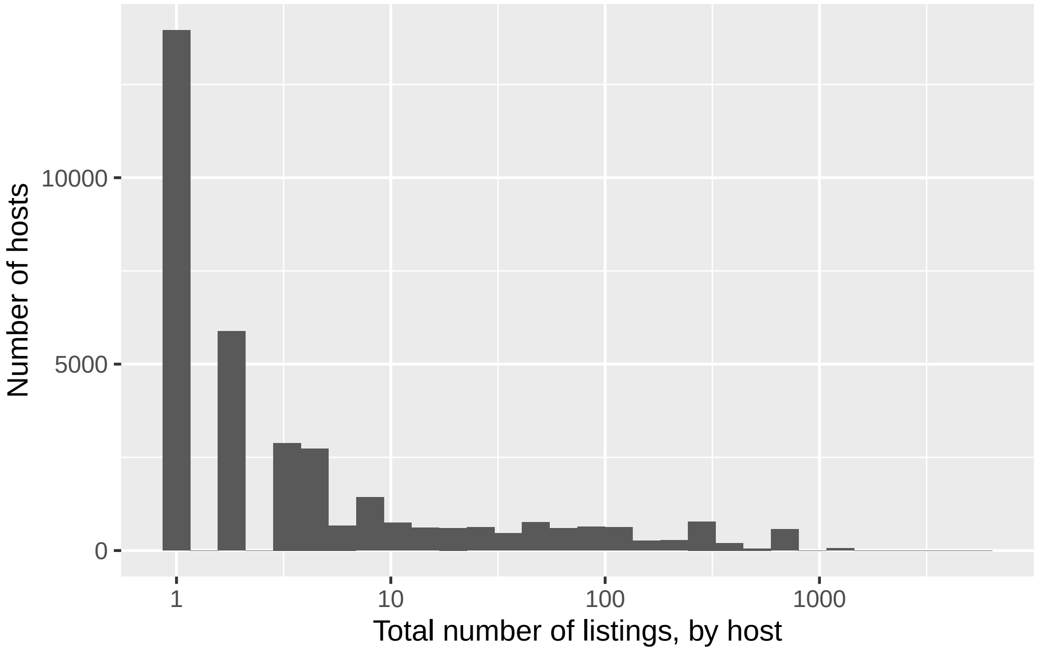


Figure 7: Distribution of host listings count in Paris Airbnb market

Figure 7 focuses on the distribution of the total number of listings managed by each host in the Paris Airbnb market. A logarithmic scale has been applied to the x-axis. This examines the number of properties managed by individual hosts, from those with a single property to those managing hundreds of listings. We can understand the diversity in hosting strategies within the market. Some hosts specialize in personalized, single-listing experiences, while others are professional managers overseeing a large portfolio of properties. The pattern suggests that the market is largely made up of individual hosts or small-scale operators, with a smaller segment of hosts managing larger numbers of properties.

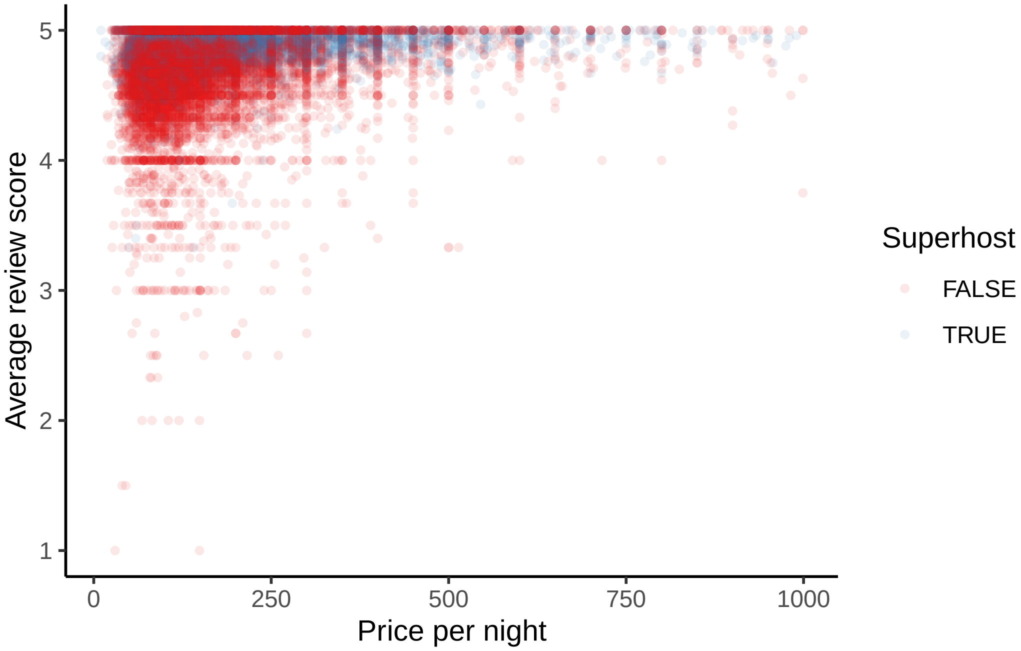


Figure 8: Influence of price on review scores by superhost status

Figure 8 shows the relationship between nightly prices and average review scores of Airbnb listings in Paris. The graph distinguishes between superhosts (in red) and non-superhosts (in blue). The graph shows that superhosts tend to have higher review scores regardless of the listing price, indicating that guests have a better experience with superhosts.

We use the equation below to make the estimate model.

$$\text{Prob(Is superhost} = 1) = \text{logit}^{-1}\left(\beta_0 + \beta_1 \text{Response time} + \beta_2 \text{Reviews} + \epsilon\right)$$

The logistic regression model depicted in Table 1 evaluates the probability of an Airbnb host in Paris being deemed a superhost. The model uses host response time and review scores as predictive variables. Key takeaways include a negative intercept, indicating a generally low probability of being a superhost, but with notable increases associated with faster response times and higher review scores. The positive coefficients for various response time categories

Table 1: Logistic regression analysis of superhost status relative to response time and review scores

	(1)
(Intercept)	−16.262
	(0.481)
host_response_timewithin a day	2.019
	(0.211)
host_response_timewithin a few hours	2.695
	(0.210)
host_response_timewithin an hour	2.972
	(0.209)
review_scores_rating	2.624
	(0.089)
Num.Obs.	22 047
AIC	24 165.0
BIC	24 205.0
Log.Lik.	−12 077.507
RMSE	0.43

affirm that responsiveness is a critical factor in achieving superhost status. Similarly, the model underscores the role of guest satisfaction, as seen in the positive relationship between review scores and superhost probability. Model fit and predictive accuracy are assessed through AIC, BIC, log-likelihood, and RMSE metrics.

Discussion

This analysis of Paris Airbnb listings has provided insights into the factors that influence a host's success on the platform. The findings of the logistic regression model emphasize the significant role of responsiveness and guest satisfaction in achieving superhost status. Hosts who respond promptly and maintain high review scores are more likely to be classified as superhosts. This indicates that attentiveness and service quality are crucial factors in the competitive Paris market. This study highlights the importance of these factors for hosts who want to improve their Airbnb presence and for guests who are looking for quality accommodations. As the Airbnb market continues to evolve, such data-driven insights are necessary for hosts to adapt and thrive in a dynamic hospitality environment.

Reference

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Tierney, Nicholas, and Dianne Cook. 2023. "Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations." *Journal of Statistical Software* 105 (7): 1–31. https://doi.org/10.18637/jss.v105.i07.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jenny Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.