IJITCE

ISSN 2347- 3657

International Journal of
Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 12, Issue 1, 2024

A ML-SENTIMENT ANALYSIS ON MONKEYPOX OUTBREAK

Dr. R Sugumar, Professor, Department Of Data Science, SICET, Hyderabad
B.Rishika, P.Samatha, M.Manognya,J .Mallesh
UG Student, Department Of Data Science, SICET, Hyderabad

Abstract:
The sudden and unexpected increase in the number of people suffering from anemia worldwi de has caused increased concern. A zoonotic disease characterized by symptoms similar to s mallpox has spread to nearly two countries and many others and has been labeled as potential ly contagious by experts. There is no specific treatment for monkeypox. However, because s mallpox is very similar to monkeypox, administration of antibiotics and smallpox vaccines ca n be used to prevent and treat scarlet fever. Since the disease has become a global problem, th ere is a need to examine its impact and public health. Number of infections, deaths, hospital v isits, hospitalizations, etc. Analyzing basic results such as can play an important role in preve nting infection. In this study, we analyzed the spread of monkeypox disease in different count ries using machine learning techniques such as linear regression (LR), decision tree (DT), ran dom forest (RF), elastic net regression (EN), Artificial Neural Network (ANN).) and Convol utional Neural Network (CNN). Our research has shown that CNN performs best and uses sta tistics such as mean error (MAE), mean square error (MSE), mean percent error (MAPE) and Rsquared error to measure the effectiveness of this model (R2). This study also presents a ti me series analysis using the Autoregressive Integrated Moving Average (ARIMA) and Seaso nal Autoregressive Integrated Moving Average (SARIMA) models to measure time differenc es. Understanding spread can lead to an understanding of risk that can be used to prevent furt her spread and ensure timely and effective treatment. Machine learning; Neural networks.

Introduction:

Monkeypox (MPX) is a zoonotic disease caused by monkeypox virus. Although it has sympt oms and signs similar to smallpox, it is less contagious than smallpox [1]. This virus is a dou blestranded DNA bag that is part of the genus Orthopoxvirus in the family Poxviridae, which includes variola virus and smallpox vaccine [2]. The virus was first discovered in monkeys in 1958 and has since spread to many parts of Africa and the United States. Most recently, in M ay 2022, a large number of measles cases were reported in some 12 endemic countries, includ ing Australia, Belgium, Spain, Portugal, the United Kingdom and the United States. Spain, P ortugal and England reported the most cases. Recent studies have revealed cases of measles i n Austria, Israel, Switzerland, Taiwan and India. The main symptoms are headache, fever, m uscle pain, respiratory symptoms, cold, etc. Although research to understand epidemiology, tr ansmission, location, and patterns is limited,The recurrence of this disease requires further inf ormation to implement strategies to prevent and treat zoonotic diseases. While scientists and doctors around the world are investigating this disease, sources of information that the diseas e is common in animals in the lake have not yet been confirmed. Clinical studies have shown that the disease can be transmitted between humans and/or animals via MPV [4]. Epidemiolo gical studies, genome sequencing and cross-country linkage have been initiated [5]. Because the number of vaccines is limited, an emerge ncy requires a coordinated response. These smallpox vaccines have been shown to be effectiv e against influenza if given quickly. Due to the rapid spread of the disease and the limited sup ply of vaccines, we need to analyze the burden and impact of the disease on the population, it s epidemiology, patterns, patterns, etc. It is necessary to evaluate. describe and verify the live s of millions of people worldwide[6]. The combination of medical care and global quarantine s stopped the spread of the virus and saved many lives. While experts warn that measles may become an epidemic in the future, it is necessary to take time to monitor its spread and detect the situation [7]. A comprehensive review of the epidemic with scarlet fever is still needed. T his analysis enables epidemiological studies, genome alignment and phylogenetic analysis. It can help understand transmission patterns and significantly inform policymakers in developin g policies and strategies to prevent further transmission. . Traditional machine learning metho ds such as linear regression, decision trees, random forests, elastic net regression, and neural networks such as artificial neural networks and convolutional neural networks were used to p erform the analysis. To measure the performance of these models, we rely on statistics such a s mean error, mean square error, mean percent error, and Rsquared error. To understand trend s and trends over time, we performed a time series analysis using the Autoregressive Integrat ed Moving Average (ARIMA) and the Seasonal Autoregressive Integrated Moving Average (SARIMA). examines patterns and trends and makes observations. To our knowledge, this is t he first paper to investigate influenza using two methods. New and important advantages are: We process and analyze data from a population sample and see how certain factors (character istics) affect performance. We began the data collection process (Figure 1).

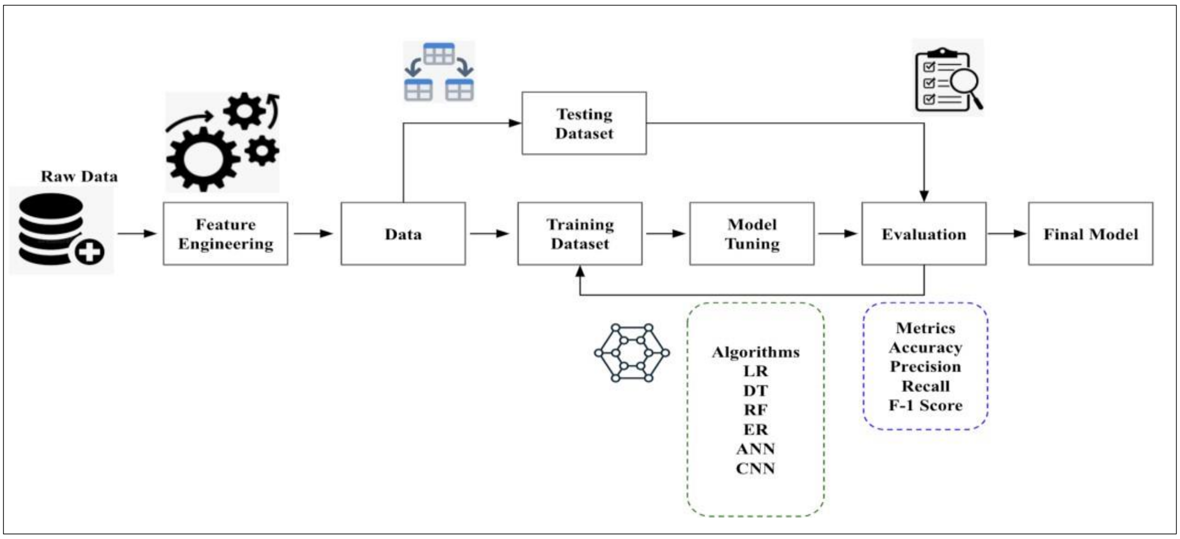


Figure 1. Analysis using machine learning.

Since the quality of the data determines the accuracy of the model, the reliability of the data n eeds to be ensured to ensure that inaccurate data and incorrect data will not lead to misclassifi cation. After collecting data, the next step is to prioritize it; This involves randomizing and cl eaning them to eliminate unwanted results or resolve missing and duplicate data. It is recomm ended to view the data to help understand any patterns and relationships between variables an d groups. After visualization, the data will be divided into training and testing. Here we divid e the data into 80% training data and 20% test data. Create a training program for the model t o learn and a testing program to check the model's performance. Once the data is segmented, the next step is to choose an appropriate machine learning model to run the algorithm on the processed data. In this study, we used a combination of traditional machine learning algorith ms and deep learning algorithms:

Statistical tests and Augmented Dickey Fuller (ADF) test can be used to check stationary dat a.

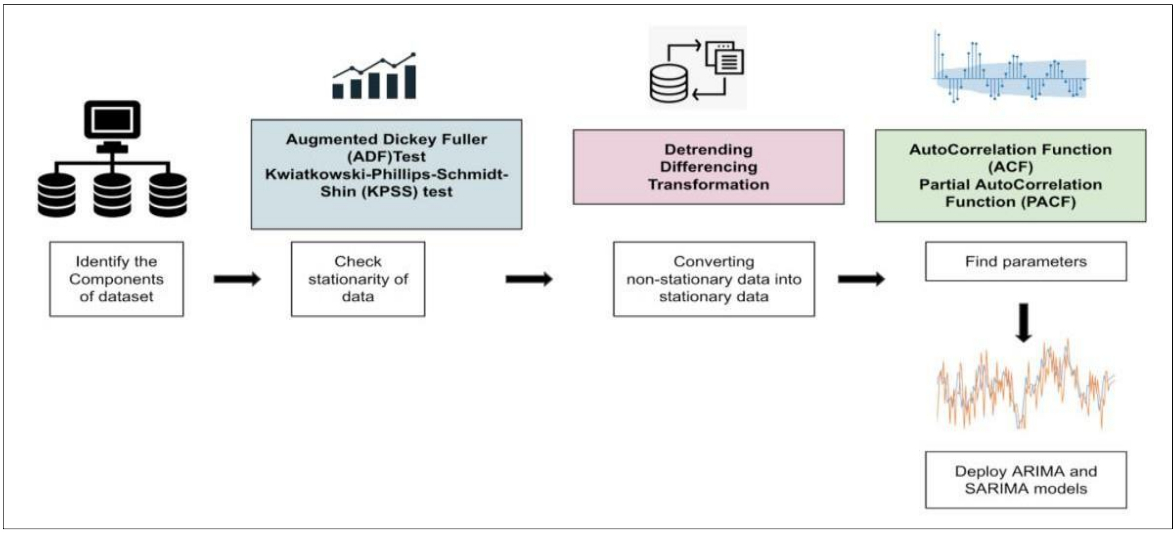


Figure 2. Time series analysis of data.

In this study, ADF test was performed to confirm that the series is not stationary under the null hypothesis. For the other hypothesis, the series is stationary, for example, if the p value is greater than 0.05, the null hypothesis will be rejected. However, if the p value is less than or equal to 0.05, the hypothesis is accepted. If time series data are not stationary, special methods can be used to convert them into stationary data. In this study, we use a different method that simply replaces existing systems with new ones. Here we eliminate the time dependence of the series and fix the mean. This results in reduced variance and seasonality during the transition period. Once the data becomes stable, we can use ARIMA and SARIMA models to analyze.

The most confirmed cases are in the United States, followed by Spain, Germany, the United Kingdom and Brazil, which have almost the same number of cases, while Switzerland, Austria and Israel have different numbers.

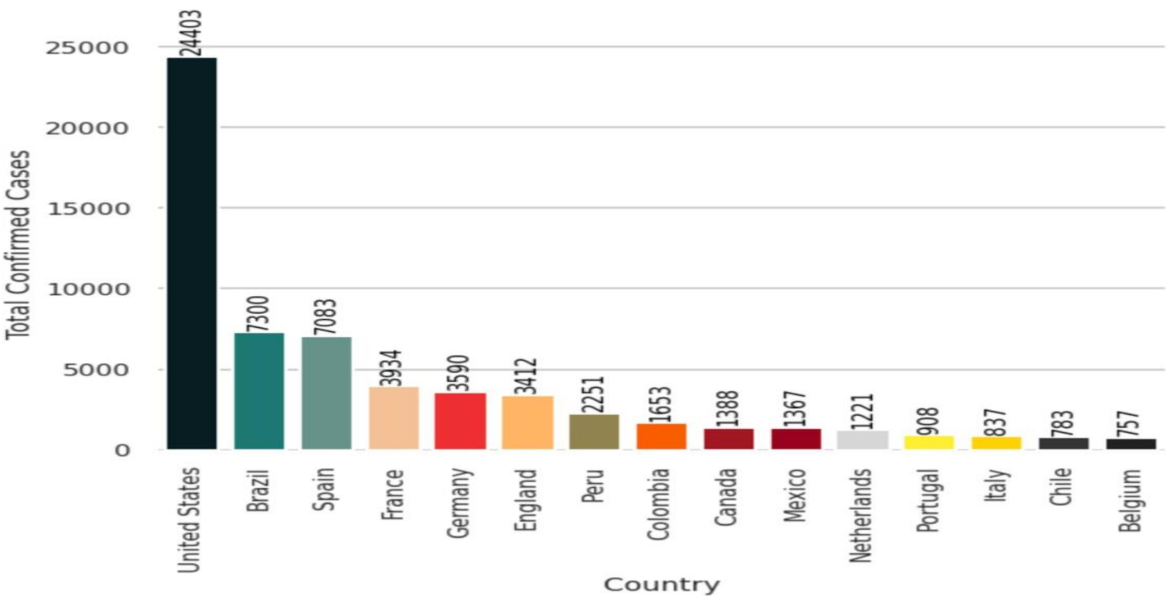


Figure 3. Total confirmed cases based on country.

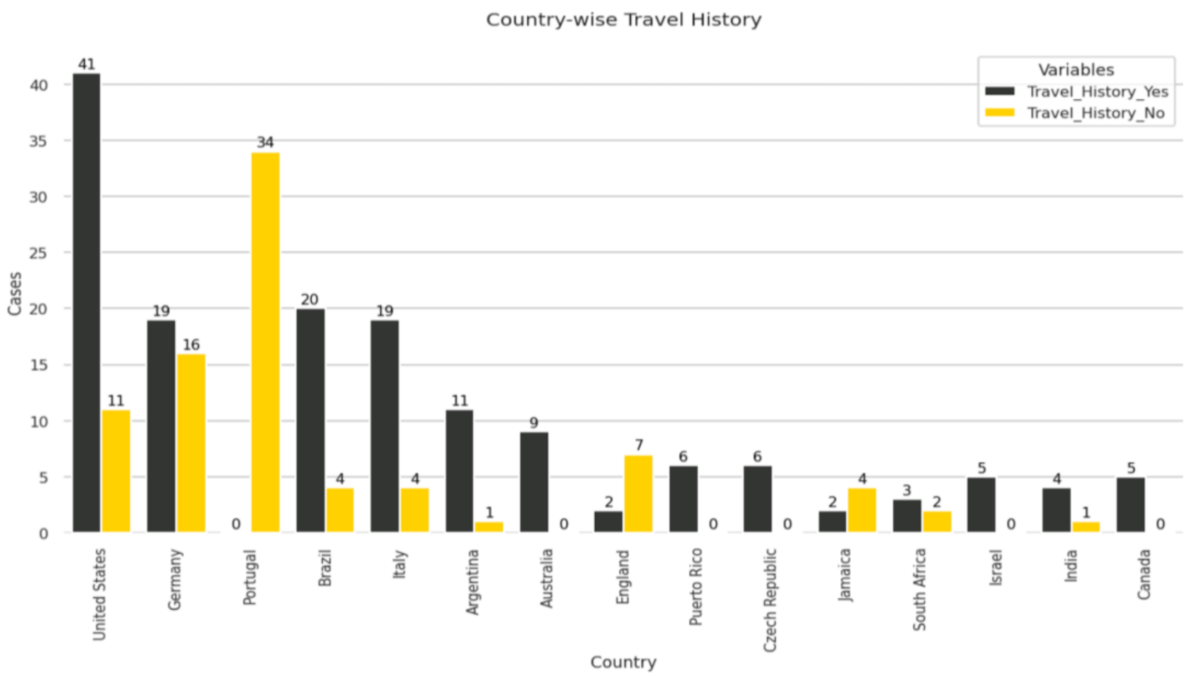


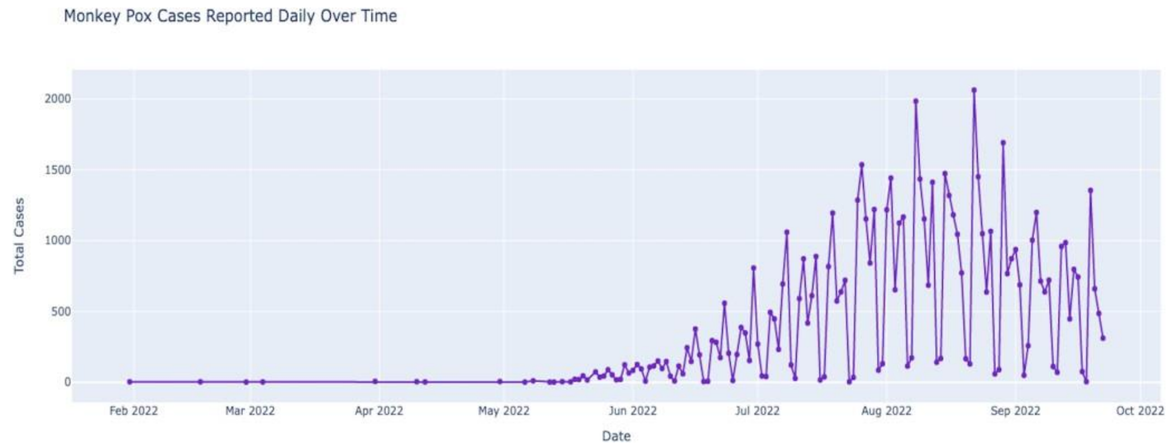
Figure 4. Travel history based on country.

4.2.2. Monkeypox Outbreak Analysis Using Machine Learning Models

In the first stage of data analysis, we determined that our data was out of date. Skewed data of ten leads to model errors in analysis and forecasting. Therefore, we use the minimummaximu m normalization technique to remove skewness. In minmax normalization, the minimum valu e for each feature is changed to 0 and the maximum value is changed to 1. After normalizatio n and data separation (80% training data and 20% testing data), we used some machine learni ng algorithms (linear regression, decision tree, random forest, elastic net regression) and neur al networks (Convolutional Neural Network or CNN, Artificial Neural). network or ANN). P erformance was evaluated using the MAE, MSE, MAPE, and R2 results shown in Table 2.

Table 2. Results from applying ML models to monkeypox data.

ML Model	MAE	MSE	MAPE	R2
Linear regression	526.666	61,776	29,448	0.287
Decision tree	519.233	13,215	36,004	0.423
Random forest	321.944	33,666	23,356	0.656
Elastic net	474.006	45,992	24,433	0.449
ANN	389.664	58,502	18,096	0.736
ANN with grid search	394.262	58,588	18,244	0.894
CNN	284.90	29,112	17,003	0.792
CNN with grid search	290.12	32,066	19,884	0.912



Based on the results shown in Table 2, we see that statistical data has many benefits. MAE, MSE, and MAPE values can vary from zero to infinity.We found that the search network improves the overall p erformance of ANNs and CNNs; so we use grid search for each model to check the performance. We also found that gridsearch CNN performed the best, followed by gridsearch random forest and gridsea rch ANN. Linear regression using grid search seems to be the most effective method. Time Series An alysisFigure 9 shows the smallpox epidemic in the last few months (February 2022 to October 2022). In the first month of 2022, we saw that the number of patients remained almost constant until the first few cases appeared after May 2022. Increase. The number of patients in July was higher than in June, and the number of patients in August was higher than in July. Now that we have data on red blood cell s over time (time records), we can create data to better understand the nature of the data. Data can be decomposed into components to identify hidden patterns and clusters. After construction, we retrieve t he model, season, and remaining data (see Figure 10). Therefore, we accept that the time series data ar e not stationary by chance. Nonstationary data should be confirmed by statistical analysis. We rely on the Augmented Dickey Fuller Test to achieve this. To validate the test, we need to evaluate whether it is a null hypothesis (nonstationary) or another hypothesis (stationary). In order for the data to be statio nary, the p value must be less than or equal to 0.05. The p value measures the probability of obtaining an observation if the null hypothesis is true. Lower p values indicate greater statistical significance. T herefore, a p value of 0.05 or less indicates statistical significance.

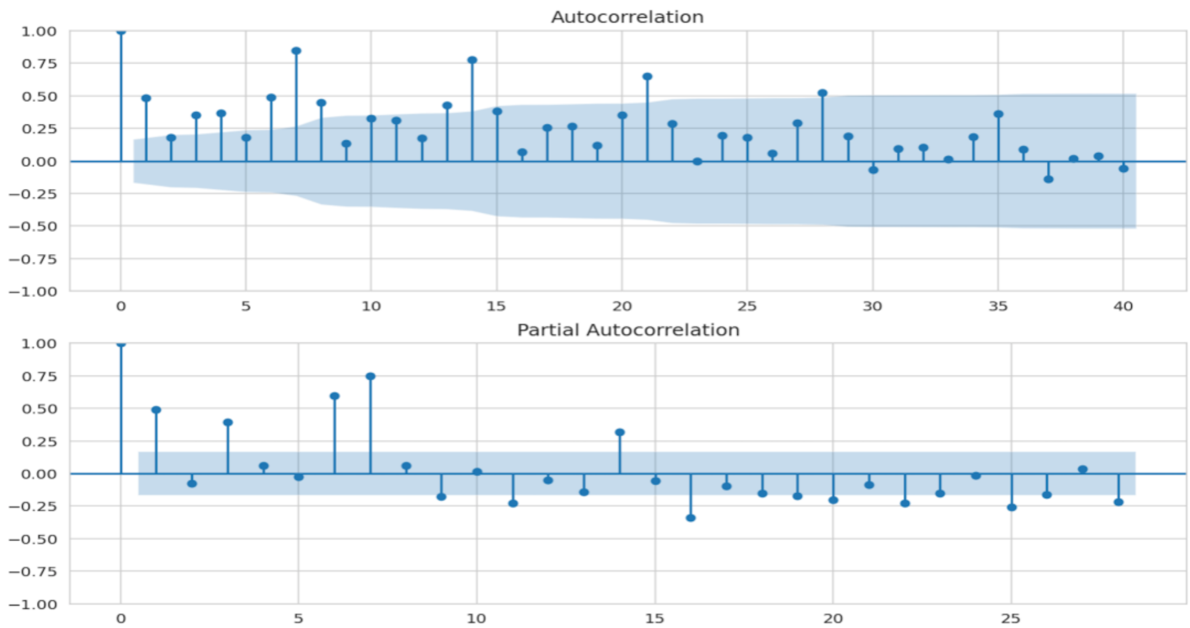


Figure 12. ACF and PACF plots.

We must evaluate a few criteria to assert that the model is a good fit. The residuals must not have any patterns; therefore, the mean must be zero, and the variance must be uniform. The kernel density estimate (KDE) plot is used to visualize the data distribution and should be similar to the normal distribution. The points fall on a 45-degree reference line if the data are normally distributed. The normal Q-Q plot (see Figure 13) indicates univariate normality. Hence, the data points must be in a straight line. In the ACF plot, if data points lie outside the confidence band, they are statistically significant. Our study shows only a few data points lie outside the band, which shows that the model may require additional parameters for better accuracy.Figure 14 depicts the deployed ARIMA model on the test set.

4.4. Discussion

In this section, we will discuss three aspects of this study that form the main part of the study, the main findings regarding all evaluations and the limitations of this study. The main results of this study are as follows:

1. Compared with previous studies, this study demonstrates the machine learning-induced spread of smallpox virus;
2. Neural networks such as random forest and elastic net regression and artificial neural netw ork and convolutional neural network;
3. . This study conducted extensive data collection to identify patterns in the data to draw con clusions. The United States has the most patients with travel history, while Portugal has the m ost patients without travel history. This means that the source of the epidemic may be Portuga l. The most common symptoms reported by patients are fever, rash, and ulcerative lesions in t he genital area. Myalgia was detected in a small number of patients. According to the correlat ion matrix, people with travel history were positively associated with hospitalization, and hos pitalization was positively associated with admitted patients. Most of the patients are in the 4 050 age group. Evaluation is based on MAE, MSE, MAPE and R2. Time series analysis sho ws that the performance of ARIMA and SARIMA models is satisfactory. As seen in time seri es analysis, if the data points in the ACF chart fall outside the confidence interval, it is statisti cally significant. Our study shows only a few points outside the cluster; This suggests that the model may need additional parameters to achieve better accuracy.

Conclusion

In the last few months, scarlet fever has spread all over the world and the increasing number of patients has become a global problem. From clinical studies to mode of transmission and tr avel history, scientists are trying to determine the significance of a potential outbreak before i t spreads further. In this study, we used machine learning techniques to identify infectious dis eases in terms of data, machine learning algorithms, and time analysis. Our research draws co nclusions based on data found on country, age, symptoms, travel history and more. We used f our machine learning algorithms and two neural networks to analyze the data and found that CNN performed best. In addition, time series analysis is performed with ARIMA and SARIM A models. Due to the limitations we discussed in this study, we hope to use deep learning met hods for more data to analyze data transfer in the future. It would be interesting to add additio nal features and use them for machine learning. Also transfer learning, Transformer etc. Adva nced machine learning techniques such as can also be used for analysis.

References

Abelson, H., Sussman, G. J., & Sussman, J. (1996). ***Structure and interpretation of computer programs*** (2nd ed.). The MIT Press. https://mitpress.mit.edu/sites/default/files/sicp/

Allaire, J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2024). *rmarkdown****: Dynamic documents for R***. https://github.com/rstudio/rmarkdown

Anderson, J. R. (1990). ***The adaptive character of thought***. Lawrence Erlbaum.

Anscombe, F. J. (1973). Graphs in statistical analysis. ***The American Statistician***, ***27***(1), 17–21. https://doi.org/10.2307/2682899

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. ***Science***, ***211***(4489), 1390–1396. https://doi.org/10.1126/science.7466396

Bache, S. M., & Wickham, H. (2022). *magrittr****: A forward-pipe operator for R***. https://magrittr.tidyverse.org

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. ***Acta Psychologica***, ***44***(3), 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Bar-Hillel, M., & Falk, R. (1982). Some teasers concerning conditional probabilities. ***Cognition***, ***11***(2), 109–122. https://doi.org/10.1016/0010-0277(82)90021-X

Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., McDine, D., & Brooks, C. (2010). Useful junk? The effects of visual embellishment on comprehension and memorability of charts. ***Proceedings of the SIGCHI Conference on Human Factors in Computing Systems***, 2573–2582. https://doi.org/10.1145/1753326.1753716

Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2021). *Modern Data Science with R* (2nd ed.). Chapman; Hall/CRC. https://mdsr-book.github.io/mdsr2e/

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. ***Psychological Methods***, ***2***(2), 131–160. https://doi.org/10.1037/1082-989X.2.2.131

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? ***Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency***, 610–623. https://doi.org/10.1145/3442188.3445922

Bergstrom, C. T., & West, J. D. (2021). ***Calling bullshit: The art of skepticism in a data-driven world***. Random House Trade Paperbacks. https://www.callingbullshit.org/

Bertin, J. (2011). *Semiology of graphics: Diagrams, networks, maps* (Vol. 1). ESRI Press.

Billings, Z. (2021). *bardr****: Complete works of William Shakespeare in tidy format***. https://CRAN.R-project.org/package=bardr

Binder, K., Krauss, S., & Wiesner, P. (2020). A new visualization for probabilistic situations containing two binary events: The frequency net. ***Frontiers in Psychology***, ***11***,

750. https://doi.org/10.3389/fpsyg.2020.00750

Box, G. E. (1979). Robustness in the strategy of scientific model building. In ***Robustness in statistics*** (pp. 201–236). Elsevier. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Box, G. E. P. (1976). Science and statistics. ***Journal of the American Statistical Association***, ***71***(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Cairo, A. (2012). *The functional art: An introduction to information graphics and visualization*. New Riders.

Cairo, A. (2016). *The truthful art: Data, charts, and maps for communication*. New Riders.

Chang, W. (2012). ***R graphics cookbook: Practical recipes for visualizing data*** (2nd ed.). O'Reilly Media. https://r-graphics.org/

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. ***Science***, ***229***(4716), 828–833. https://doi.org/10.1126/science.229.4716.828

Craik, K. J. W. (1943). ***The nature of explanation***. Cambridge University Press.

Crameri, F., Shephard, G. E., & Heron, P. J. (2020). The misuse of colour in science communication. ***Nature Communications***, ***11***(1), 1–10. https://doi.org/10.1038/s41467-020-19160-7

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. ***Harvard Business Review***, ***90***(5), 70–76. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

De Cruz, H., Neth, H., & Schlimm, D. (2010). The cognitive basis of arithmetic. In B. Löwe & T. Müller (Eds.), *PhiMSAMP. Philosophy of mathematics: Sociological aspects and mathematical practice* (pp. 59–106). College Publications. http://www.lib.uni-bonn.de/PhiMSAMP/Data/Book/PhiMSAMP-bk_DeCruzNethSchlimm.pdf

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., et al. (2017). Curriculum guidelines for undergraduate programs in data science. ***Annual Review of Statistics and Its Application***, ***4***, 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930