Data Visualization in JASP: Implementation of Customizable Graphs

Introduction

This report demonstrates the application of the graphics implemented in the statistical software JASP. The general idea of the internship involved building complex graphs from the ground up while solely relying on the R package ggplot2 (Wickham, 2016a) and its respective dependencies. By introducing new descriptive graphs, the project aimed to offer novel solutions to data visualization problems of JASP users. In total, five conceptually different graphs were introduced, partly inspired and adapted from existing R packages, partly designed and built from the ground by the intern. The initial series of graphs proposed in the first project description was revised and adapted to the learning outcomes. Precisely, the implementation of standard bar plots in all t-Test and ANOVA modules was prioritized to provide a highly requested and academically indispensable feature to JASP users.

Before looking at the individual plots and their potential use cases, we look at the general procedure of graphical implementations within JASP: The first step, after finding an appropriate sample picture from published papers or existing R packages, addresses the generalization of the graphic. This involves the creation of a test build that can be applied to various shapes of data. Therefore, a dynamical implementation is needed, requiring a flexible axis and label creation. After testing the graph with simulated data sets, including ones with missing data and variable variances, R functions capturing the plotting process and required statistical procedures are constructed. The final step involves the implementation within JASP. As graphs rely on different types of data and data structures, it is important to consider how the data reaches the appropriate functions, thus, certain data cleaning operations are undertaken. Due to the high degree of individuality of the visualizations, methods of data piping and cleaning varied significantly. All graphics displayed in this report followed this procedure and were generated with JASP. Additional features of the implementation included the creation of descriptive help files to guide users and unittests to test the core functionality of each graphic. Following, we will look at the features and use cases of the constructed graphics.

Visualizations and Use Case

Likert Graph

The first graph in the series concerns a diverged stacked bar chart. Its primary function is to visualize responses to likert scales, a rating system often encountered in survey research (Heiberger & Robbins, 2014). Participants are asked to rank their agreement towards statements by selecting options ranging from, for example, “strongly agree” to “strongly disagree”. The number of levels differ between questionnaires but typically covers a range of five to nine levels (Robbins & Heiberger, 2011).

Multiple features offered by the final graph enable a straightforward understanding of the data. A legend at the bottom indicates the different levels used in the survey. Each layer of the horizontally flipped bar chart represents the percent contribution of a level to the overall proportion of levels for the specified item. Each flipped bar chart illustrates the different level contributions for one item, which is specified on the y-axis. The levels can be arranged in the desired sequence. The color palette uses automatically adjusted gradations of yellow and cyan tones to distinguish between the two extremes of the scale. Accordingly, the x-axis has a bipolar alignment of percentages. Percentages indicating the combined contribution of all levels belonging to one extreme (e.g., “agree” vs “disagree”) are displayed on the respective side of the graph with a percentage in the middle of the x-axis illustrating the contribution of the neutral level, provided the questionnaire used one. After displaying the graph in JASP, users are offered two options for customization: First, an adjustable font size for variable descriptions on the y-axis. Entire questions may be displayed on the y-axis to directly compare the likert visualization to the question or statement at hand, thus, it is sensible to provide some adaptability in this regard. Furthermore, users can decide to visualize all items within the same graphic by checking “Assume all variables share the same levels”, instead of being displayed in separate plots as it is per default. Note, however, that this graph only works for ordinal and nominal variables as factor levels are investigated.

A special characteristic of this graphic is the fact that all selected items can be visualized in the same graph, given they have the same number of levels. This allows users to simultaneously compare different items which might belong to the same test battery. Additionally, the noticeable split of both extremes enables the user to compare the two sides more intuitively and shows readers clear distinctions in item selection by participants.

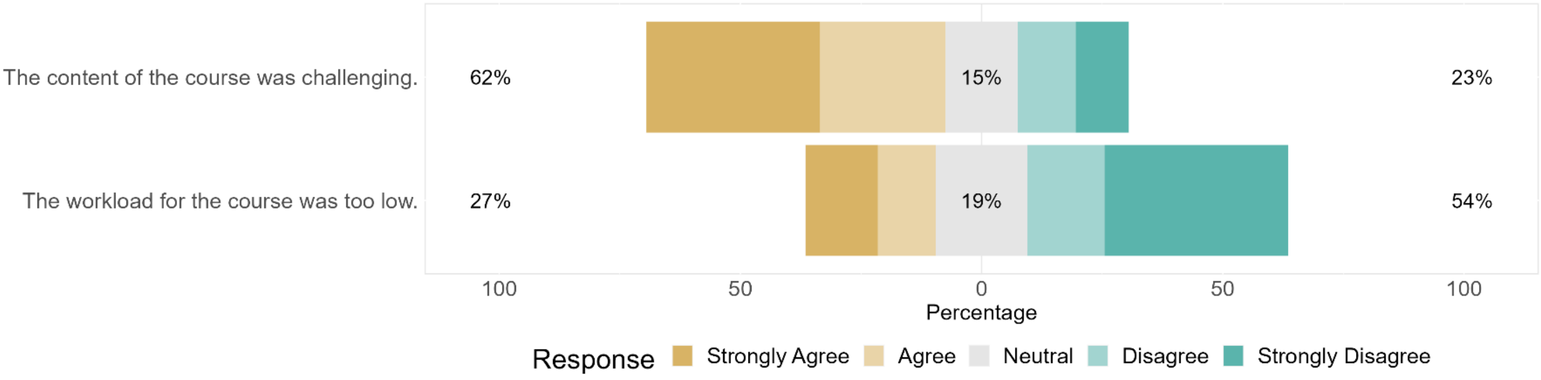

Figure 1: Likert graph visualizing the results of items from a course evaluation.

The above figure illustrates the functionality of the discussed graph. Immediately, we can see that for the first item, concerning the difficulty of a university course, nearly two-thirds of the respondents agreed with the statement while approximately a quarter opposed it. This gives a quick and intuitive indication regarding the students’ perception of course difficulty. On the contrary, most students perceived the workload for the class in question to be not too low, with more than half of respondents expressing disagreement with the item concerning the workload. Even more so, a considerable proportion of answers indicated “Strongly Disagree” in response to this item, suggesting a higher workload. Lastly, we observe that around a fifth of responses showed a neutral stance toward the statements. All the above-mentioned benefits are apparent in this illustration, including a clear legend and representation of the extreme proportions.

Pareto Graph

Next, the Pareto chart is discussed. Its purpose is to visualize the frequency of different factors to identify the most important ones. Decisions about their respective importance are made by looking at a cumulative summary regarding their percent contributions (Tague, 2005, p. 376 - 377). Although mostly applied in settings surrounding quality control (Tague, 2005, p. 376), this graph also offers researchers opportunities to visualize the contributions of analysis components, such as from principal component analyses (e.g., Sun et al., 2022).

The visualization represents a bar chart of the factor frequencies, sorted in descending order. The x-axis shows the different factor levels whereas the y-axis depicts their frequency within the specified variable as counts. A second y-axis on the right side of the graph illustrates percentages and scales with the height of the original y-axis. A default feature of this graph is a cumulative line stretching over all the investigated factors. It illustrates the cumulative proportional contributions of the factors, that is, each point on this line includes the percentage contribution of the current and previous factors to the overall count. The values of this additional plot line correspond to the second y-axis whereas the bars, representing the counts, align with the original y-axis. An option named “Pareto rule” offers users to manually input percentages to receive an additional visual indication of how many factors contribute to the specified percentage. As the process concerns discrete factors, only ordinal and nominal variables can be visualized. A special characteristic of this graph is the second y-axis which, combined with the cumulative line, provides an additional analysis of factor contribution.

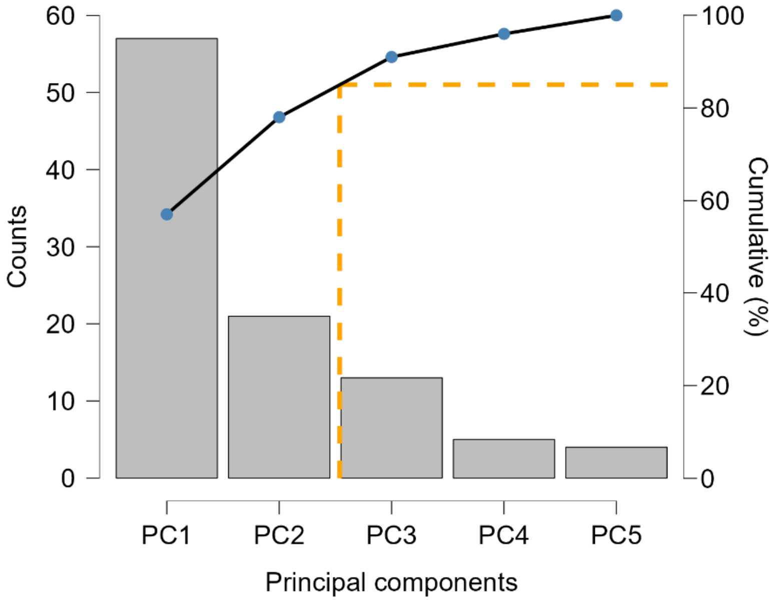

Figure 2: Pareto graph visualizing the (cumulative) explained variance of principal components.

Often multiple methods are used to determine the relevant number of components resulting from principal component analysis. One important method considers the explained variance. It looks at how much of the total variance can be explained by a certain number of principal components. Thereby, the cumulative explained variance is important, reflecting the combined contribution of components, with those contributing the most being added first. A Pareto graph can visualize the cumulative explained variance well, showing how many factors should be considered. The above example follows the convention of successively adding components until an explained variance of 85% is reached (conventions differ, this is just for illustration purposes). It demonstrates that choosing two components is not sufficient to meet the requirement of 85% and that at least three components are needed to explain the defined variance. Although not essential for principal component analysis, this chart provides a convenient visualization of the established method, facilitating readers' understanding of discussed procedures. Other usage areas include quality control analyses trying to determine the significance of problems based on their frequency (Tague, 2005, p. 376).

Bland-Altman Graph

Bland-Altman charts are used to visualize the level of agreement between two methods of measurement, often needed in clinical measurement comparisons. Hereby, a new technique is compared to an established one. If both methods agree sufficiently, the new one might replace the old one (Bland & Altman, 1986). Bland and Altman (1986) argued that evaluating the agreement between two measures by simply looking at the correlation coefficient might be misleading. They proposed the following visualization to assess measurement agreement:

To investigate the level of agreement, a scatterplot displaying the difference of the measurements (y-axis) against the mean of the measurements (x-axis) is generated. A horizontal line illustrating the mean difference of both measures is drawn. The level of agreement is further operationalized by computing the bias, thus, the limits of agreement (95%) around the mean difference. Subsequently, two horizontal lines representing these limits are drawn as well. Assuming differences in this interval are not clinically important, methods could be used interchangeably (Bland & Altman, 1986). Additionally, users can display the confidence intervals of the mean difference and its limits, enabling greater accuracy in the method comparison. Selecting the option "Shading" will highlight these confidence bounds. The above-mentioned values (e.g., mean difference, limits) can be shown separately in a JASP table. The JASP visualization only considers ordinal and continuous variables as both forms work with certain interval ranges. Additionally, plot generation only occurs when a pair of variables is provided, therefore, a maximum of two measures is compared at a time.

A benefit of the visualization of differences against means is that it emphasizes the relationship between measurement errors (differences) and the true value (estimated by the mean of both measurements). Note, that the plot only quantifies the level of agreement and does not tell if the agreement is sufficient. Therefore, an appropriate distance between measurements should be defined in advance to facilitate the method comparison (Bland & Altman, 1986).

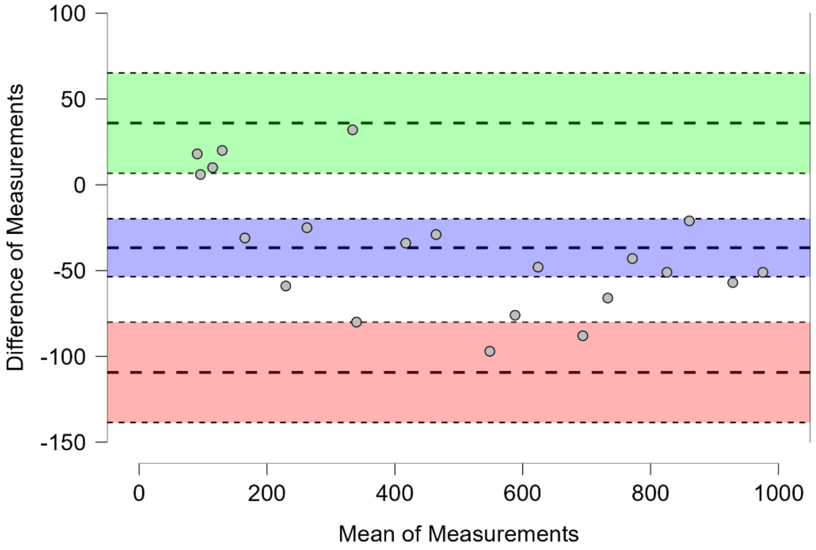

Figure 3: Bland-Altman graph visualizing the level of agreement between method A and method B.

The interpretation of Bland-Altman graphs follows an informal approach, also involving visual investigations, and depends on a priori defined discrepancies. In our example, we investigate reaction time data. For both methods to be exchangeable, the mean difference should be comparatively small and the limits of agreement in a reasonable range considering the concept of reaction times. Furthermore, we will look at observable trends and the overall variability of the data.

The above visualization compares two reaction time measurements, namely method A and method B. The mean difference is -36.7ms, which can be considered a small average discrepancy in terms of reaction times. Further, the limits of agreement amounted to 35.9ms and -109.3ms, respectively, indicating a moderate range, which is reasonable and expected between similar measurements. When investigating the plotted data, we see a consistent variability across the data points. Although some values are close to the determined limits of agreement, they are still within the acceptable bounds. Lastly, a negative trend within the data can be observed. The higher the average reaction times, the more computed differences increase. Further measurements of higher magnitudes might reveal systematic differences. In our case, however, we only observe a mild trend within the acceptable boundaries. Therefore, considering all evaluated aspects, the methods of measurement are not producing systematically different results, therefore, may be used interchangeably.

Density Graph

This graph was requested by the intern himself due to its usefulness in visualizing distributions of variables on a continuous interval. By drawing a reasonable continuous curve, it shows the probability distribution of our data (Wilke, 2019, p. 61), making it easier to detect the shapes of distributions.

The implemented graphic visualizes the probability distribution of the selected variable. The y-axis represents the probability density from the kernel density estimation, a common technique to estimate the continuous curve from the data. By default, our implementation uses the Gaussian kernel to smoothen out the noise (Wilke, 2019, p. 61). Peaks in the curve reveal where values are concentrated and the colored area below the curve always equals 1 (Wilke, 2019, p. 62). Dimensions of the variable in question are represented on the x-axis. The Density graph section offers two options for customization: Another variable can be inserted in the section “Separate densities”, allowing the user to receive multiple overlapping distributions corresponding to the different levels of the variable. The distributions will be displayed in the same plot and the colors used by each distribution can be adjusted using the color palette menu. Accordingly, the transparency of the colors can be changed, offering a range between 0 to 100. This graph only displays the distribution of continuous variables. The additional variable to split up the densities only considers ordinal and nominal variables to display the different distributions for different levels.

Although histograms have been the first choice in visualizing distributions for a long time, since they are relatively easy to create, increasing computational power allowed us to find more elegant solutions, like Density plots (Wilke, 2019, p. 61). The Density graph is simply a smoother version of the histogram, allowing for an easier interpretation of the distribution shape. It should be noted, however, that the interpretation of Density charts is subject to certain limitations: Firstly, the selected kernel affects the displayed shape. In our case, this might result in a more Gaussian-like estimate. This is true, especially for smaller data sets (Wilke, 2019, p. 62). Secondly, these plots tend to show non-existing data, most noticeable in the tails (Wilke, 2019, p. 63).

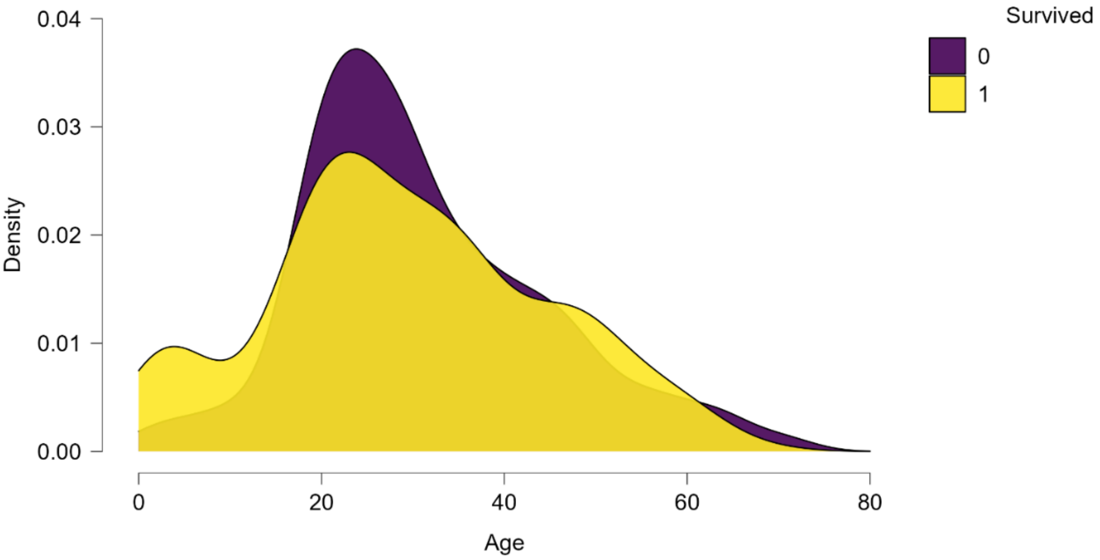

Figure 4: Density graph visualizing the age distribution of survivors (1) vs. non-survivors (0).

A data set often used in teaching contexts concerns the passenger records of the sunken cruise ship titanic. The above graph compares the age distribution of survivors and non-survivors of the fatal encounter, with “1” indicating survivorship. Investigating the overall age distribution reveals that a significant portion of passengers was between the ages of 20 to 30, a tendency also reflected in the above figure. Although both curves show a similar shape, differences in their peaks are evident. A noticeable difference can be observed for passengers in their twenties. Considerably fewer of the survivors were in this age range, compared to the non-survivor proportions. Further, survivors show a small peak at younger ages, indicating that proportionally more children were among the survivors compared to the non-survivors. The non-survivor distribution in this age range is deeper and shallower, suggesting that fewer children were among the deceased, compared to the other age ranges. Lastly, survivors seemed to also slightly favor elderlies compared to non-survivors, evident by a small bump at ages 45 to 55. In summary, both distributions suggest that, even though many passengers were in their twenties and thirties, children and certain elderly ages might have been prioritized for evacuation. However, the numbers of survivors and non-survivors differ, and interpretations are only sensible within their respective distributions.

Bar Graph

Lastly, a highly requested feature by JASP users concerned the implementation of Bar charts to visualize t-Test and ANOVA results. The graph serves the purpose of visually comparing the differences between group means, enabling users to showcase whether differences between groups are truly distinct and align with the computed test statistics.

As the name suggests, the Bar chart illustrates bar diagrams of the respective groups’ means. While the t-Test version can only showcase the means of a maximum of two groups within the same graph, the ANOVA version can include two or more groups in its visualization. Additionally, the ANOVA version enables users to further split the investigation of differences in group means into the levels of another specified factor variable, allowing a more sophisticated visual analysis and simultaneous view of different factor levels. In all Bar graphs, error bars can be displayed, being the default option in the t-Test version. Users can switch between error bars representing the standard error and confidence intervals for the Frequentist version or credible intervals for the Bayesian version. By default, the Bar charts zoom to the relevant position on the y-axis displaying cut-off bar diagrams at the means’ respective positions. This distorted view, however, might mislead users by making differences look bigger than they are. Therefore, although users can adjust the axis of graphs with JASP’s plot editing feature, Bar charts provide the option to fixate the x-axis at the y value of 0, offering an unbiased view with minimal effort. All grouping variables, including the additional factor variable available for the ANOVA versions, must be of ordinal or nominal nature.

While the bar charts of the different ANOVA versions work approximately the same, the t-Test versions differ noticeably in their appearance. The graph for the independent samples t-Test visualizes the means of a grouping variable on the specified dependent variable whereas the paired samples t-Test compares the means of the two selected variables. The one-sample t-Test simply shows one bar diagram of the specified variable, visualizing the adjustable “Test value” in the process. For illustrative reasons, the following use case will only showcase the Bar Charts of an ANOVA procedure.

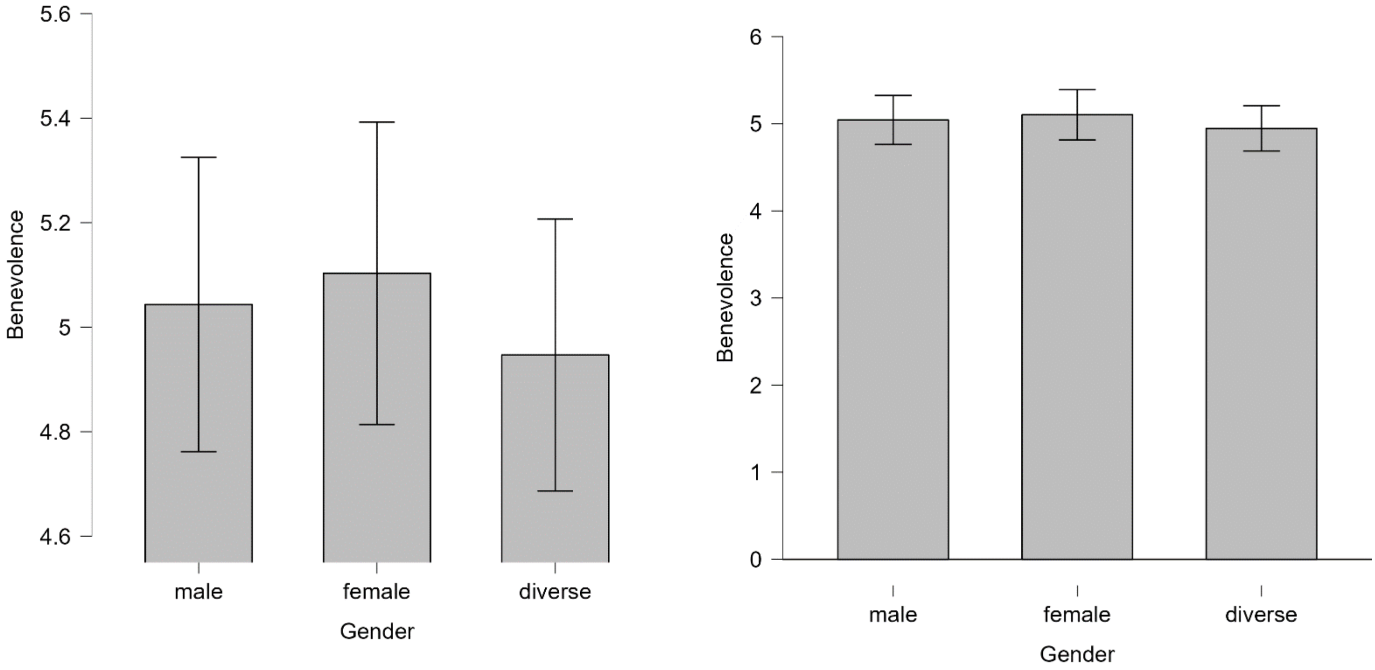

Figure 5: Bar graphs visualizing the means and CIs of three groups regarding their benevolence. Left chart illustrates the zoomed-in version, the chart on the right the version with a fixed x-axis.

A researcher is interested in comparing the benevolence of different genders, including a third option for people identifying with neither male nor female categories, namely diverse. In an elaborate study, the researcher collects ratings of the test subjects ranging from zero (low benevolence) to six (high benevolence). To get an overview and compare the three groups regarding the dependent variable benevolence, the researcher displays the Bar graphs of their respective mean values. The initial Bar chart (fig. 5, left) shows the means and their 95% confidence interval. It may seem like females have on average higher benevolence compared to their counterparts, with the diverse group performing the lowest out of the groups. However, investigating the values with proper scaling of the y-axis (fig. 5, right) reveals that groups only differ marginally in their benevolence, with the means’ confidence intervals displaying a lot of overlap. Our intuition matches the results of the overall significance test from the ANOVA, with an associated p-value of 0.725. The example summarizes the functionality of the Bar graphs and emphasizes the axis-fixation feature which allows users to avoid a distorted representation of the results.

Challenges and Outlook

During the implementation of the above graphics, several difficulties in the creation of the graphs and integration into JASP arose. Both required a rethinking of the problem at hand and fostered the intern's problem-solving skills. First, challenges appeared when simply creating the visualization in the designated testing directories. The adaptation of existing R packages led to the need of reducing the number of dependencies on external R packages. Therefore, several convenience functions outside of the JASP environment had to be rewritten in base R (R Core Team, 2021) and adjusted for the context of solely using ggplot2. This approach, unfortunately, narrowed down the selection of established plotting packages and, consequently, led to the manual creation of components using existing methods as templates. Furthermore, visualization functions required a dynamical axis generation which, in specific cases, posed a challenge due to multiple factors and options determining the axis range. Therefore, axis creation had to consider generated confidence intervals (e.g., Bar graphs) or measurement biases (e.g., Bland-Altman graphs) to ensure interpretable and complete graphs. A rather specific problem appeared when constructing the Pareto chart as it makes use of a secondary y-axis and a cumulative plot line scaling directly with this second axis. In the beginning, ggplot2 had limited capabilities for its construction, just recently adding supportive structures for additional axes (Wickham, 2016b). Subsequently, additional computations for proper scaling of the secondary y-axis were incorporated and combined with base R's interpolation functions to linearly approximate required data points.

Besides challenges in creating the visualizations, the general integration within JASP caused some unforeseen troubles. An early problem arose when working on the visualization of survey data. Because the Likert graph can incorporate all variables in one graph, the entire dataframe is processed within the plot generating function. This function, however, only works for nominal or ordinal data types, requiring continuous variables to be filtered out of the dataframe. Therefore, screening and cleaning of the entire dataframe were implemented, ensuring the final function accesses viable data types and does not break. Another challenge involved the implementation of the Density Chart. The beginning of the R script for the "Descriptives" module shows a unique characteristic: Missing values are only removed when a split variable is defined. Precisely, rows in the variable data set that contain missing values within the column of the split variable are removed, however, when no split variable is defined no removal of missing values takes place. Although most functions solve this by simply omitting the missing values from the data set before any plotting or computation happens, the functions concerning the Density charts could not do that. This is because the chart introduces a potential third variable enabling the user to split the displayed density into the number of factors provided by this third variable. The field showing the usable third variables accesses these from the list containing all variables rather than from the list of specified variables as it greatly simplifies the usage of the feature and avoids confusion for users. However, this leads us into a dilemma as this third variable is never processed by the responsible script and has to be read in separately after the missing values of the split variable have already been removed. Therefore, it was decided to read in all involved variables manually and separately and to start the data cleaning process afterward. Lastly, graphics were constructed with JASP's *plot editing* function in mind, ensuring the displayed information is not distorted or lost after user inputs.

Visualizing information appropriately enables us to display hidden patterns within the data and communicate those in a meaningful way to a broader audience (Tableau, n.d.). Good visualizations can improve our decision-making processes and might aid in ad hoc analysis planning (Ali et al., 2016; Vellido et al., 2011). Consequently, making good graphics benefits the scientific workflow and the communication of complex findings. Hopefully, these new JASP structures can help with that.

References

Ali, S. M., Gupta, N., Nayak, G. K., & Lenka, R. K. (2016, December). Big data visualization: Tools and challenges. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)* (pp. 656-660). IEEE. https://doi.org/10.1109/IC3I.2016.7918044

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet, 327*(8476), 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8

Heiberger, R., & Robbins, N. (2014). Design of diverging stacked bar charts for Likert scales and other applications. *Journal of Statistical Software, 57*, 1-32. https://doi.org/10.18637/jss.v057.i05

Robbins, N. B., & Heiberger, R. M. (2011, July). Plotting Likert and other rating scales. In *Proceedings of the 2011 joint statistical meeting* (Vol. 1). American Statistical Association.

R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. https://www.R-project.org/

Sun, S., Huang, T., Zhang, B., He, P., Yan, L., Fan, D., ... & Chen, J. (2022). A novel intelligent system based on adjustable classifier models for diagnosing heart sounds. *Scientific Reports, 12*(1), 1-17. https://doi.org/10.1038/s41598-021-04136-4

Tableau. (n.d.). *What Is Data Visualization? Definition, Examples, And Learning Resources*. Tableau. Retrieved June 17, 2022, from https://www.tableau.com/learn/articles/data-visualization#:%7E:text=Data%20visualization%20is%20the%20graphical,outliers%2C%20and%20patterns%20in%20data

Tague, N. R. (2005). *The quality toolbox* (Vol. 600). Milwaukee, WI: ASQ Quality Press.

Vellido, A., Martín, J. D., Rossi, F., & Lisboa, P. J. (2011). Seeing is believing: The importance of visualization in real-world machine learning applications. In *Proceedings: 19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2011: Bruges, Belgium, April 27-29, 2011* (pp. 219-226). https://upcommons.upc.edu/handle/2117/20273

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. NY: Springer. https://CRAN.R-project.org/package=ggplot2

Wickham, H. (2016). *ggplot2 2.2.0 coming soon!* RStudio. Retrieved June 24, 2022, from https://www.rstudio.com/blog/ggplot2-2-2-0-coming-soon/

Wilke, C. O. (2019). *Fundamentals of data visualization: a primer on making informative and compelling figures*. O'Reilly Media.