

Visualizing Data / Perception, Visualizations & Plot Settings

Data Visualization and Graphical Perception

Data visualization is a powerful means for the human brain to visually translate and understand complex data to make more accessible interpretations. In any given data science team, data analysts, scientists, and engineers focus on efficient ways for cleaner data sets and building and deploying accurate predictive models delivering optimal business outcomes. Additionally, teams need to communicate data patterns, trends, and outliers through data visualization to business stakeholders for decision making.

But quite often, the simple semantics of graphical perception in data visualization are overlooked in different phases of a data science pipeline, whether exploratory data analysis, data preparation/transformation, modeling, evaluation, or deployment. For example, while interpreting the customer churn data, your business stakeholders might be better at reading bar charts or histograms than pie charts to study the frequency distribution of data points!

While conveying this information to stakeholders, it is vital to keep in mind that the human eye perception is powerful at some things; it is not good in others. This article briefly introduces the idea of graphical perception, the best practices and the general hierarchy of graphs, and, finally, how to choose charts and visualizations based on the former two.

Graphical Perception and creating cleaner charts

Graphical perception is the human potential to understand and interpret visual information in graphs/charts, visualizations, and maps. William Cleveland is perhaps the most notable figure known for data visualization and its graphical perception. The cleaner and minimalistic charts in practice today are accredited to Cleveland's best practices and the general hierarchy of data visualization.

Graphs are like predictive models - they should be simpler and interpretable when possible. The figure below provides a hierarchy of best practices. It ranks the perception parameters on a scale of "Simple" to "Complex" with a graph having a "position along a common scale" the easiest to interpret and "Color hue" the most difficult.

| Simple ↑ ↓ Complex | Rank | Aspect |
|-----------------------------|------|---|
| | 1 | Position Along a common Scale |
| | 2 | Position on identical but nonaligned scales |
| | 3 | Length |
| | 4 | Angle Slope |
| | 5 | Area |
| | 6 | Volume Density Color saturation |
| | 7 | Color hue |

Approaches to learning about data

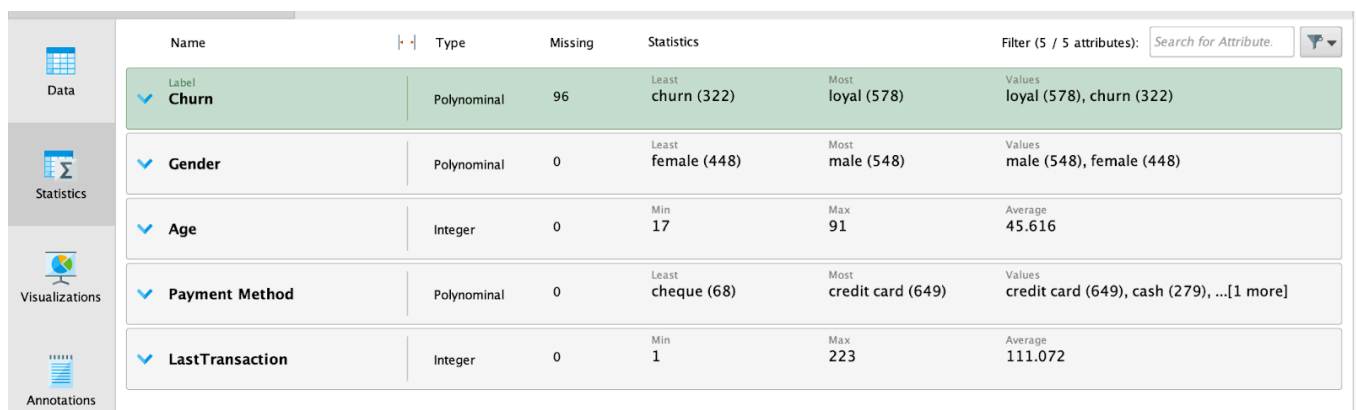
To better understand the approaches to learn about the data through data visualization, here we use the Customer Churn dataset in RapidMiner studio Training repository that consists of the following five attributes:

- Gender (Binominal - Male or Female)
- Age (Integer)
- Payment method (Polynominal - cheque, credit card, cash)
- Last transaction (Integer)
- Churn (Binominal - Churn, Loyal)

Descriptive Analytics

Descriptive analysis constructively summarizes a sample or the entire dataset in a meaningful way such that patterns might emerge in the data. The techniques often include calculating descriptive statistics - like mean, standard deviation, variance, or min/max - and univariate and multivariate analysis using different visualizations.

Within the RapidMiner studio, we will retrieve the customer churn data. Then, we will join the output to the results and run the process. In the Data section, you will see the entire dataset is displayed. In the Statistics section, we can see the descriptive statistics (like min, max, average) for each attribute and the number of missing values as well - in this case, Churn.



The screenshot shows the 'Statistics' panel in RapidMiner. On the left is a sidebar with icons for Data, Statistics, Visualizations, and Annotations. The main area displays a table of statistics for five attributes: Churn, Gender, Age, Payment Method, and LastTransaction. The table has columns for Name, Type, Missing, and Statistics. The 'Statistics' column is expanded for each attribute, showing 'Least' and 'Most' values for categorical variables and 'Min', 'Max', and 'Average' for numerical variables. A search bar and a filter icon are at the top right of the table.

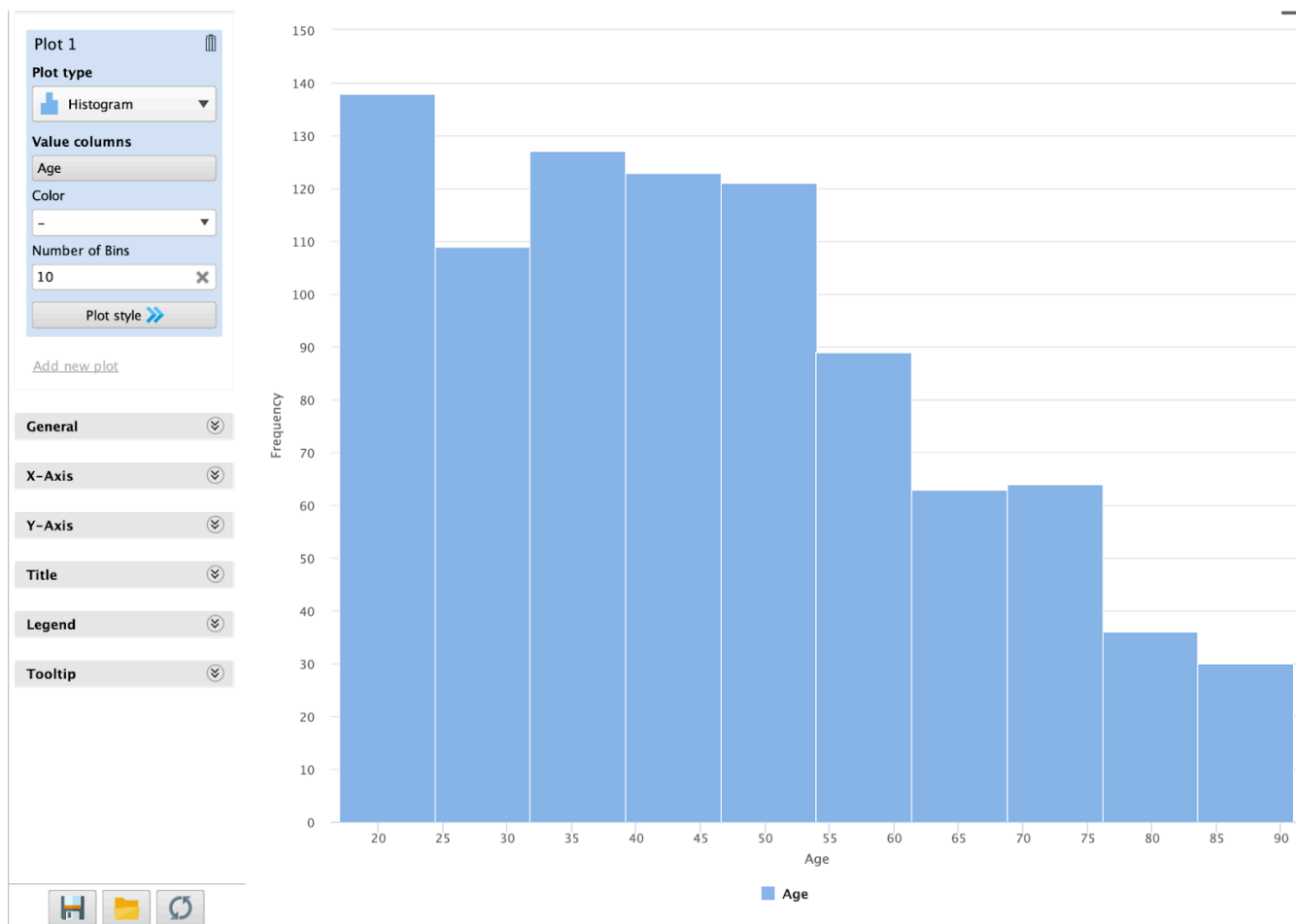
| Name | Type | Missing | Statistics |
|-----------------|-------------|---------|--|
| Label Churn | Polynominal | 96 | Least churn (322) Most loyal (578) Values loyal (578), churn (322) |
| Gender | Polynominal | 0 | Least female (448) Most male (548) Values male (548), female (448) |
| Age | Integer | 0 | Min 17 Max 91 Average 45.616 |
| Payment Method | Polynominal | 0 | Least cheque (68) Most credit card (649) Values credit card (649), cash (279), ...[1 more] |
| LastTransaction | Integer | 0 | Min 1 Max 223 Average 111.072 |

Univariate Analysis

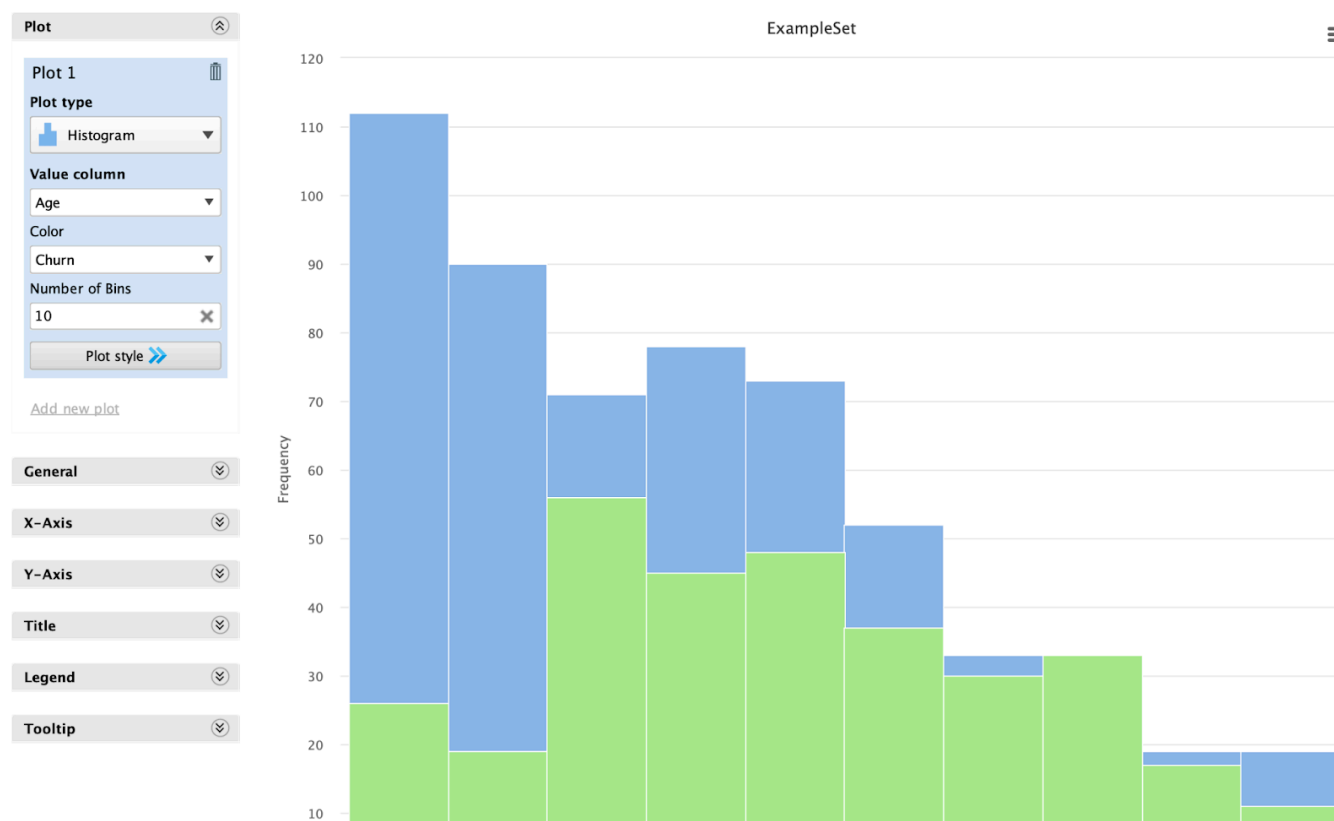
Univariate analysis means the analysis of one variable or one feature in a given data set. It tells us how data in each feature (or attribute) is distributed and also tells us about central tendencies like mean, median, and mode.

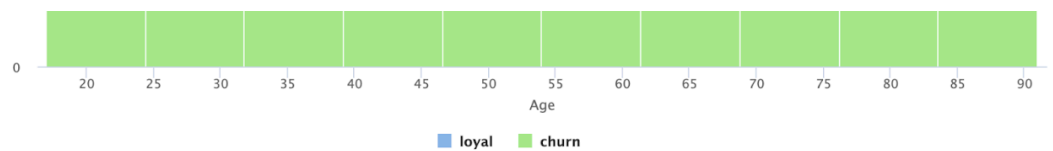
Histograms

One way to visualize the frequency distributions of an integer type feature is by using Histogram charts. In our customer churn dataset, below we can see the distribution of Age. Here it indicates the occurrences of each age bin -- in this case 10 bins.



Now let's say we want to visualize the above frequency distribution with the number of people that are loyal or churned in each age group/bin. In that case, we can set the "Color" setting as Churn. Below is the resulted visualization.

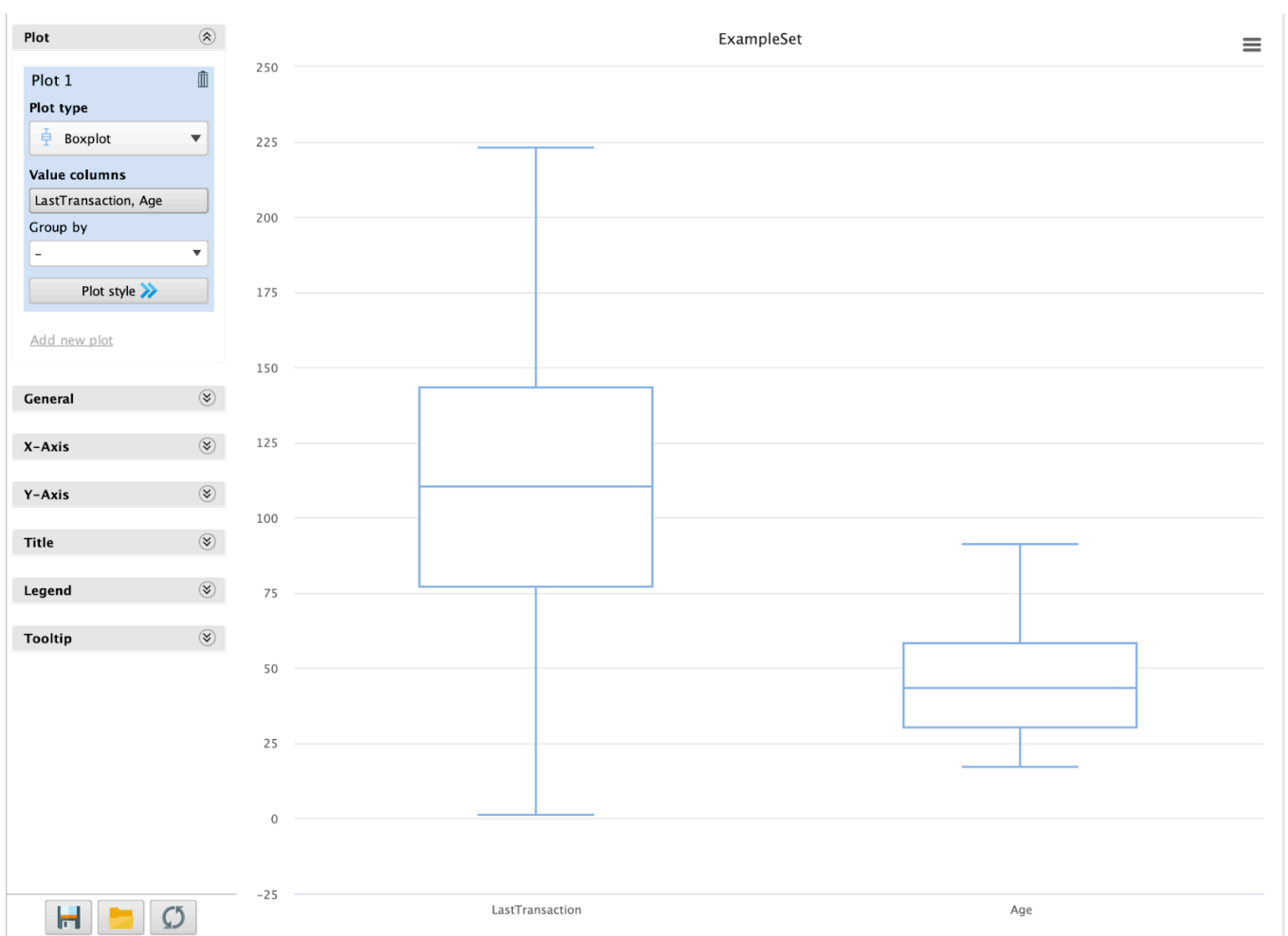




Box plots

Histograms and box plots are quite similar as the purpose of both is to help visualize and describe numeric/integer data attributes. While histograms are better in determining the underlying distribution of any given attribute, box plots allow us to compare multiple attributes better than histograms as they are less detailed and take up less space on a common scale.

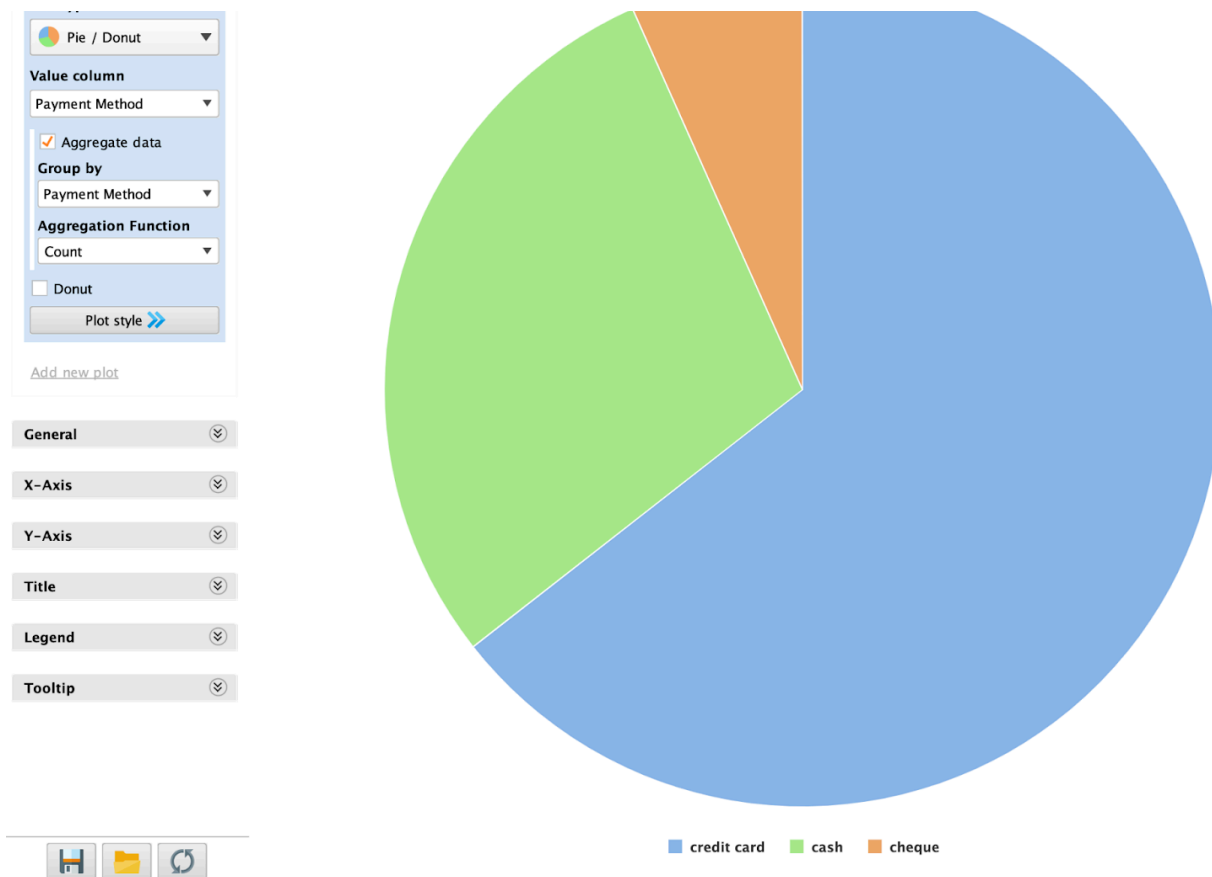
Below is a box plot visualization of the Last Transaction and Age attributes in the customer churn dataset.



Pie charts

Pie charts come to use when we want our viewers to have a general understanding of the part-of-whole relationship of a given attribute, and comparing the slices to each other is not so much important. As we mentioned previously, pie charts may not be the best visualization when we want to learn about the distribution of a dataset as compared to histograms or box plots, hence their rare applicability.





For Continued Learning

Apart from descriptive and univariate analysis, bivariate and multivariate analysis comes to use when we want to visualize the relationship between two or more attributes or features. There are many visualization techniques in the RapidMiner Studio to do such level analytics within a few clicks. Below are some types of charts to explore for continued future learning.

Bivariate Analysis

- Line chart for trends
- Correlation matrix
- Scatter plot and Scatter matrix (example: high correlation between gender and churn)
- Parallel coordinates

Multivariate Analysis

- Scatter plot - three dimensions (Example: Churn, Gender, Age)