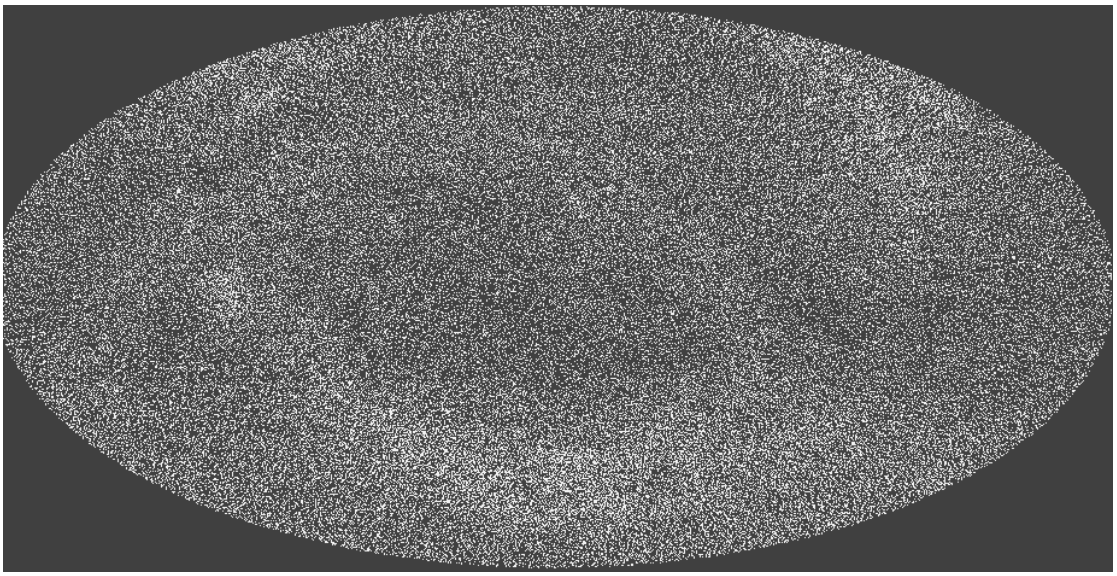


Basic plotting can be fuzzy and difficult to interpret if you have large amounts of data

Visualizing Data Effectively

Published on May 23, 2018 Edit article | View stats

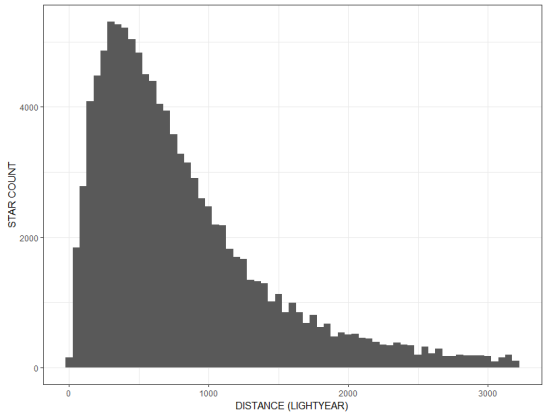
Carter Edie
Crossing the bridge between data analytics and strategy
2 articles

58 12 1 2

A few people asked me about the chart methods I used in my previous post on identifying data trends in large data sets. Different data plotting techniques are used depending on variable statistic (count, value, outcome), variable type (discrete or continuous), visual goal (exploratory, descriptive, or interpretive), and most importantly audience (casual, informed, expert). Data visualization is a wonderful combination of observation, statistics, creativity, analysis, critical thinking, artistic license, and story-telling. We are communicating a message with data visualization and what's wonderful about the medium is that the visual dimensions of size, colour, shape, position, text, and opacity can all be optimized to tell the best stories.

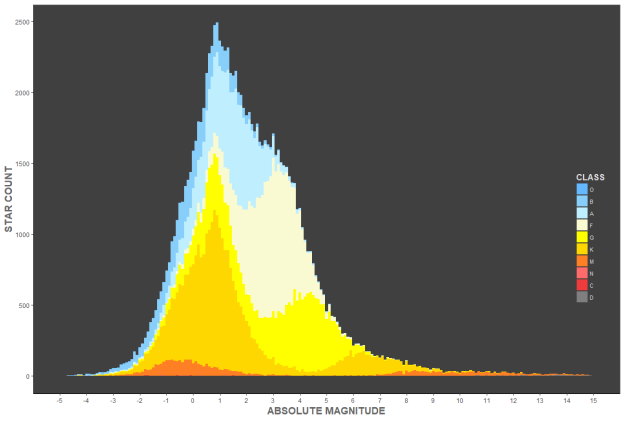
Here I'll take a look at the visualization methods that were used (histogram | scatter | rose | ridgeline density), why the plotting techniques are useful for certain types of data, and the trade-offs that exist in choosing those plots. There will be a couple of new plotting methods included as well (dot | box) as these are closely related to the first four and provide unique interpretation insights.

Histogram



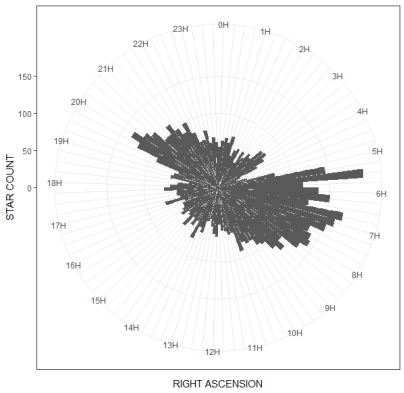
Sometimes simple is best. In large data sets with many, many recorded variables it can be overwhelming on where to even begin. A histogram plot is simple and efficient. It communicates one dimension of a variable, a discrete count. The data is binned according to a user defined width along the continuous variable to be examined and then a count is performed on the number of observations that fall within. Looking at the basic histogram plot, there isn't a lot of value for your plot area money. The interpretation is quite flat, straight forward, and boring.

One way to "punch it up a bit" (a hated phrase by graphic designers) is to include a related discrete variable to increase the visual dimensions of the plot. On the histogram to the right I have included Stellar Classification as the fill to the histogram, making sure to take the colour scheme into account to communicate another dimension of the data, the spectral colours of those types of stars. Now visible in the plot are characteristics of each class that can be further investigated.

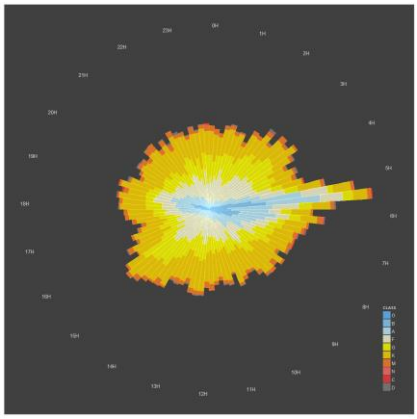


- **Strengths: Quick, simple, effective communication tool | count visualization**
- **Trade-offs: Limited visual dimensionality and difficult to identify precise values | difficult to directly compare multiple categories and discrete variables (such as names, places, gender, class)**
- **When to use it: When you have data sets with large continuous value ranges and you need an exploratory data visualization**

Rose Diagram



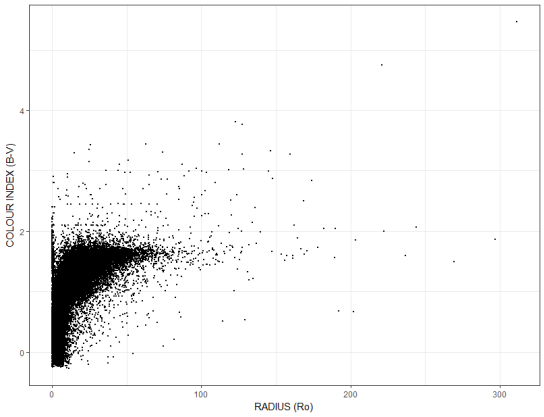
Another way to introduce dimensionality is to alter the direction, position, or coordinate system of the plotting method. A rose diagram is a transformed histogram, where direction is visually included in the mapping of the binned count. This method is best intended when interpretation of a radial direction is needed, such as with sub-surface faulting, reservoir depletion rates, wind-speed, drilling direction trends, or objects around the earth.



Just like the histogram example, we can augment the interpretability of the rose diagram by including a discrete variable as the fill value in the rose segments. In the stellar example on the right we can interpret that the higher counts in the 5 to 6 H range can be attributed to a higher observed O and B class star counts. That trend extends to the opposite side, ~19 H, which would be worth a detailed examination.

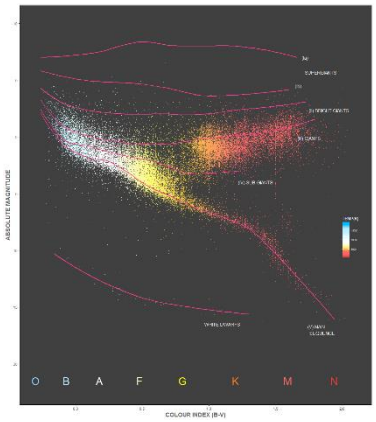
- **Strengths: Ability to show direction as well as volume count for a data set**
- **Trade-offs: Specialized histogram in the radial coordinate system | Directional context required for interpretation**
- **When to use it: When you have large data sets with geographical or radial direction components**

Scatter



Whereas histograms and rose diagrams are one data dimension in axis plotting, scatter plots introduce a second continuous variable axis to reveal trends that exist between them. Each point on the plot represents one observation in the data set with each axis representing the specific measurement recorded. The scatter plot is a work horse in data analytics, with machine learning models and predictive algorithms built on the functions they produce.

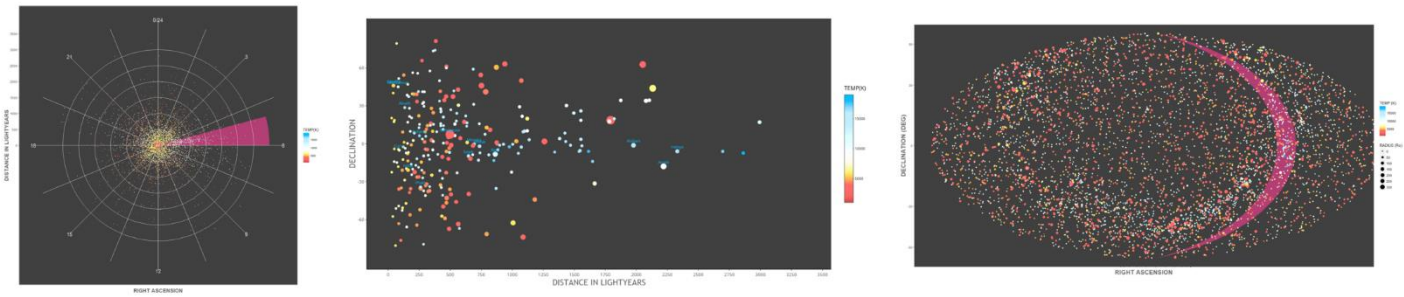
The scatter plot on the right is called a Hertzsprung-Russell diagram. The x-axis is the star's colour-index (a measure of it's blue-white-red spectra) and the y-axis is a star's absolute magnitude (the visible magnitude of a star from a standard distance). The bluer a star, the lower the colour-index. The larger a star, the higher its absolute magnitude. By scatter plotting the observable colour-index and absolute magnitude we can interpret the stellar class (O, B, A, F, G, K, M, N) and sequence (Ia, Ib, II, III, IV, V and WD).



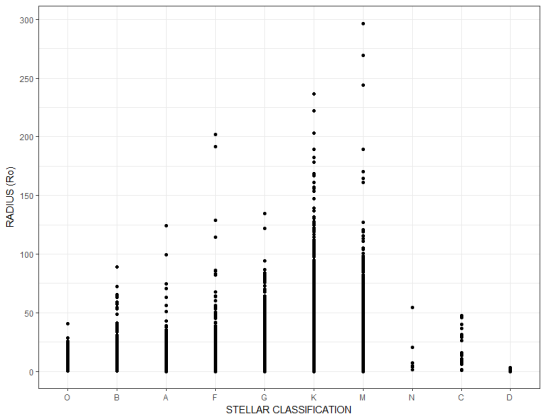
- **Strengths: Relationship mining between continuous variables, categorical cluster mapping, and non-linear trend visualization | Effective use of plotting space**
- **Trade-offs: Isolation of interpretative and actionable insights between scatter plot pairs in wide data sets (many observable continuous variables) is time consuming, domain context is required, and identification of outliers is difficult**
- **When to use it: When you want to isolate continuous variable relationships for model design and identifying correlations in the data**

Examples of Scatter Plotting

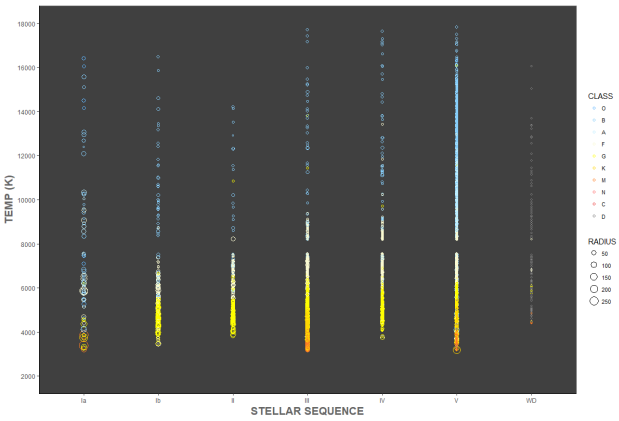
The flexibility of scatter plotting supports many visual transformations. Below is an example of three plots that can be interactively linked to produce a visual mapping of our celestial sphere. The plot on the left is a radial transformation of the x-axis where like the rose diagram from earlier our observation point is at the center. The middle plot represents a distance (in light-years) vs declination plot to parallel the pink wedge mapping in the radial plot. The plot on the right is an elliptical projection of the stars visible to us in the night sky. Colour (temperature), size (Radius), and text (Common named stars) all improve the interpretation of the scatter plot set.



Dot



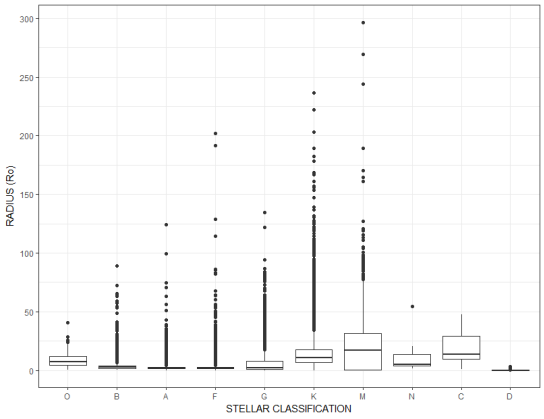
The dot plot is a specific instance of a scatter plot. All the characteristics are the same, except for one. Instead of having two continuous variables plotted against each other (such as IP90 BOE vs lateral length or height vs weight) one of the variables is a category (or discrete). This lends to data points mapped in straight lines, leading to obscure visualizations and undefined groupings.



Dot pots can give more interpretation value when 2nd and 3rd dimension layers are added in. The plot on the right is not only showing the temperature ranges for each of the star sequence groups, but it also has a colour to represent the class (with the class colour scale matching the temperature colours) as well as a size dimension for each point. Each point is plotted as a ring to ensure that there exists visual separation. This type of plot would be useful for plotting production markers (IP90) of operators drilling in the same area segmented by formation or revenue per store per day of the week.

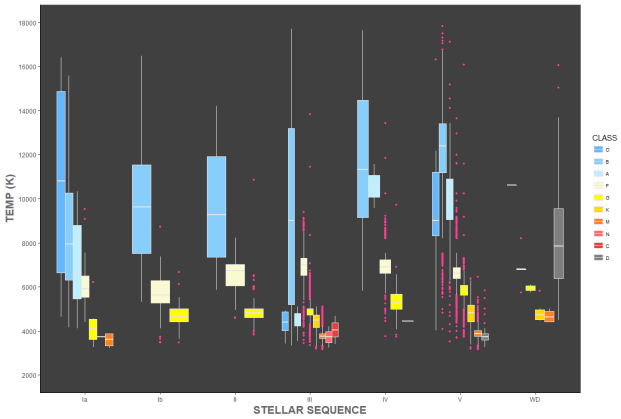
- **Strengths: Trend identification through relationship mining between one continuous variable and one categorical variable | large variable range mapping**
- **Trade-offs: Identification of outliers is difficult | statistical analysis is poor | limited use of plotting space**
- **When to use it: When you have a smaller sized data set and want to isolate continuous variable trends within discrete variables**

Box



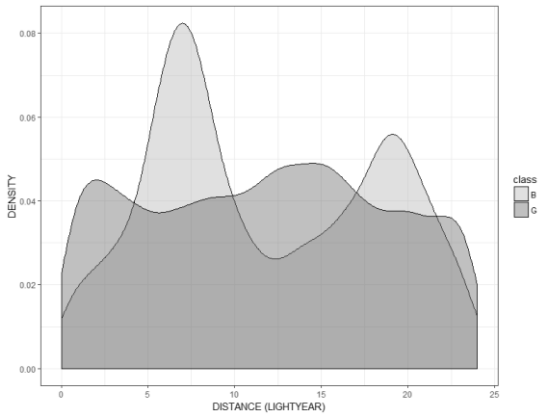
A concern that can be raised is there has been little statistical analysis in the plotting tools to this point. In particular with the scatter and box plots, means, quartiles, and outliers are difficult to observe and interpret. The box plot builds on the structure of the dot plot and introduces some statistical value visuals. Means for the category data points are the horizontal lines inside the boxes, the upper and lower limits of the box are indicative of 25% above and below respectively, and the individual data points that remain are the outliers. This can be a very effective way of visualizing observation points that are performing extremely well or extremely poorly

What can we do to get more interpretive characteristics out of a box plot? Similar to using size in the dot plot we can use other categorical or discrete variables as 3rd dimension mapping tools. Once isolated, categorical means and outliers can be identified, and uncertain data can be investigated (such as the lower temperature boxes for the "O" stars compared to "B" and "A" in the Giant(III) and Main(V) categories).

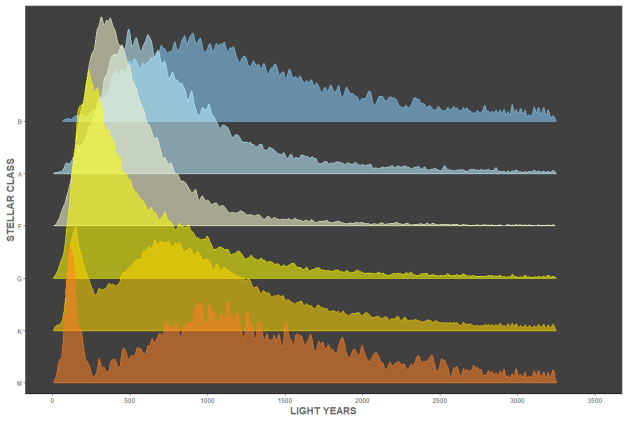


- **Strengths: Statistical analysis through relationship mining between one continuous variable and one categorical variable | large variable range mapping | outlier identification**
- **Trade-offs: High level of interpretation knowledge | high level of domain context required | poor for exploratory visualization**
- **When to use it: When you have a smaller sized data set and want to isolate continuous variable trends within discrete variables**

Density / Ridgeline



The last method I am going to review is the density plot, a relative of the first visual we looked at the histogram. A density plot is a smoothed version of the histogram which is useful for plotting continuous data for comparison on relative scales. Distinct categories can easily be plotted, compared, and interpreted.



Where density plots are particularly useful is when we want to scale categorical data for distribution comparisons. A ridgeline plot is a multi-class single dimension density plot that visually compares distribution profiles. When combined with discrete class colours, the plot on the right is showing observation distance limits in the data set that get shorter as the star class temperature decreases (the profile distribution peaks show B > A > F > G). An interesting observation occurs with the final two classes though. K and M stars have distribution notches taken out of their near-Earth observations, an unexpected result. Further data investigation would be needed to solve this puzzle.

- **Strengths: Relative comparisons of continuous distribution curves | large variable range mapping | distribution quality control**
- **Trade-offs: high level of domain context needed | low observation count can influence visual results**
- **When to use it: When you have multiple categorical variables that you want to compare continuous distribution profiles | you might be concerned about observation artifacts influencing your data insights**

Conclusion

As some might point out, there are many other plotting techniques that I did not investigate here. Line, 2d density, map overlay, bubble, area, tree, chord, marginal, heatmap, lollipop, violin, diverging bar, time series, slope, cluster, and dumbbell all have their time and place to be used (yes, even the dreaded pie chart). Even more important than the charts that you end up deciding to use to tell your data story are the charts that are excluded. Not every technique is effective for the type of data you're working with. Experimentation, flow, aesthetics, context, and audience perspective are all things to consider when choosing what to run with. Speaking from experience, editing is just as important a step in telling your data story.

If you have a data story that needs to be told, or help digging deep into your data mines to find those hidden insight gems message me through LinkedIn or reach out at analytics@carteredie.com