This is a graded discussion: 15 points possible

due -

13 21

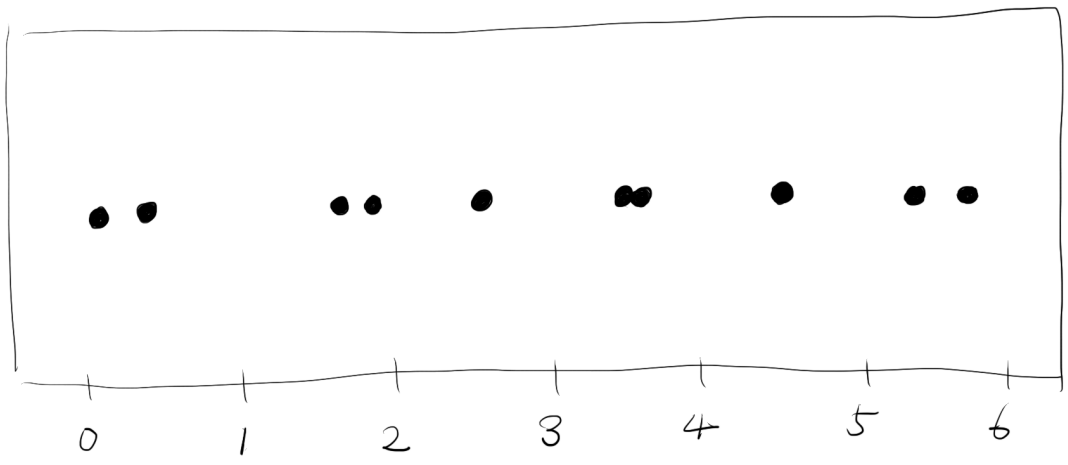
Directly visualizing 1-D data

One of the simplest forms of data you can think of is a **list of 'something'**. Something can be categorical values (nominal) or numbers. When it's about categories, you can probably want to count each category and plot a bar chart. When they are numbers, it may require a bit more thoughts.

So imagine a bunch of numbers (i.e. 1D quantitative data). Examples can be: a list of incomes, a list of 100m dash records for all professional athletic events in 2017, a list of heights of all adults living in South Korea, the amount of time each of you has spent on Candy Crush Saga ever, etc. etc. They can be stored with just one column.

What would be the most *direct way* to visualize your dataset?

I think it'll be the **scatterplot**, or **strip chart**, where you simply draw a symbol for each data point like this:



So here are my questions for you.

1. What are the **benefits** of this visualization method?

2. What are the **drawbacks** of this visualization method? (If you're not sure, think about having more data points)
3. What would be **alternative methods** to overcome the drawbacks? There are numerous ways and I'd like to ask you to think of two directions. First, how can we improve the method *while maintaining the basic concept—showing every data point*? Second, how can we visualize the data in *different ways*? Try to come up with as many solutions you can think of!

Search entries or author | Unread | Subscribed

Reply

Prem Amal (https://iu.instructure.com/courses/2165942/users/6684842)

Monday

What are the **benefits** of this visualization method?

It's a simple and straightforward way to visualize a univariate dataset. When we have a single dimension of data, a 1D scatterplot provides a clear and uncluttered representation of data points along that axis. We can observe the density of data points along the axis. Areas with a higher concentration of points indicate regions of greater data density. Outliers, if present, are easily identifiable as data points that deviate significantly from the majority along the axis.It allows for a quick visual assessment of the data's distribution, central tendency, and spread.

What are the **drawbacks** of this visualization method?

This method provides a simplified view of data, often focusing on a single variable's distribution. It may not capture complex relationships or interactions between multiple variables. When dealing with a large number of data points, they may overlap, leading to overplotting. This can obscure patterns, density, and individual data points, reducing clarity. For example, if the dataset contains a cluster of the same value, we cannot represent it that effectively using a scatter plot.This method is not suitable for visualizing relationships between multiple variables or exploring multivariate datasets, where understanding correlations is essential.

What would be **alternative methods** to overcome the drawbacks? There are numerous ways and I'd like to ask you to think of two directions. First, how can we improve the method *while maintaining the basic concept—showing every data point*? Second, how can we visualize the data in *different ways*? Try to come up with as many solutions you can think of!

Instead of individual points, we can create a scatter plot overlay on top of the density plot to show the distribution of data along the axis. This provides a smoother representation of data density and helps overcome issues of overplotting and also shows the points. Or we can plot every point on x axis at different heights to show unique points. This may create issues with very large datasets having higher frequency.

Different ways to visualize the data :

Histogram: Create a histogram to visualize data distribution within bins or intervals along the axis.

Box Plot: Use a box plot to visualize the median, quartiles, and potential outliers in the data.

Line Plot: When dealing with time series or sequential data, a line plot can show trends and changes over time

Violin Plot: A violin plot combines a box plot and a kernel density plot to show both summary statistics and data density

Edited by **Prem Amal** (https://iu.instructure.com/courses/2165942/users/6684842) on Sep 18 at 2:43am

Reply

Dustin Cole (https://iu.instructure.com/courses/2165942/users/6701715)

Monday

1. The plot above is beneficial because you can see every datapoint (as long as there aren't duplicate values). It's also easy to be able to visualize the difference in relational distance between points.

2. The drawbacks would be the difficulty of understanding the chart if there were thousands of data points. Also there is no way to be able to see the number of identical data points. You can only see that there is at least one at each point.

3. It would be helpful to put the data in bins and make a bar chart with the count of data points in each bin. Or a box plot would be another option. Currently it is difficult to see distribution. You could also keep the chart above and enlarge the data points based on quantity.

Reply

Thomas Jablenski (https://iu.instructure.com/courses/2165942/users/6701599)

Tuesday

The strip chart does a good job at showing all of the data points in a precise way and to the point. The data-ink ratio is very good here while there a few data points. Once we reach a higher number of data points it becomes much harder to read the chart as the points will be overlapping with each other and in some cases if all the points are closely grouped the result would just be a line with no definition between data points.

Some ways to improve upon the strip chart would be to make the points transparent to an extent so you could see other points underneath and if there are many points at one section the area would be darker. To solve this problem even more we could add some noise to the strip chart so the points aren't all on a single axis but have some depth showing giving more room for data points. Another way to see the data is translating the strip chart into a bar chart. All you need for bar charts are the measure (which is depicted already) and then a category alongside it. For example, the example a list of 100m dash records for all professional athletic events in 2017 could have a name of the athlete accompanying the data point.

Reply

Sneha Satish (https://iu.instructure.com/courses/2165942/users/6679606)

Tuesday

Benefits of these plots include:

- You can see every data point individually, so it is easy to see trends or outliers.
- It is easy to assess correlation between variables

Drawbacks include:

- If there are too many data points, overplotting can occur making it difficult to distinguish between individual points. It becomes messy.
- It's not very good for categorical data

Alternative methods:

- Box plots: Shows median, quartiles and potential outliers.

- Beeswarm plots: It is a variation of scatterplots and arrange points within a narrow band to reduce overplotting.
- Histograms: To address overplotting, histograms can be created by dividing the data into bins and showing frequency of data points in each bin.

Reply

Gary Croke (https://iu.instructure.com/courses/2165942/users/6706306)

Tuesday

1) It's easy to draw / produce. You can see all the data. It's easy to interpret.

2) The density of the data may become a problem. With a lot of data points it becomes difficult to distinguish them.

3) Without changing the basic format, selecting a minimal footprint marker may help. Binning data points together in some fashion might be possible, e.g. consolidate multiple close points together into a marker of greater size. Adding a false second dimension (allow points to scatter vertically, understanding they are still part of the same line). Actual binning (convert to a histogram) might be appropriate for some data. If the data covers a large range (several orders of magnitude), a log scale may help. Breaking the line up into several sections may help overcome limited horizontal presentation space. For dynamic presentations, allowing the viewer to zoom into the data might be a useful feature.

Reply

Erik Gonzalez (https://iu.instructure.com/courses/2165942/users/6352173)

Tuesday

1. The benefits of a scatterplot or strip chart in this 1-D scenario are that these methods help to display ordinality relatively well.

2. The main drawback I see is how to deal with overlapping data points - if there are two numbers with a value of 3, would the audience be able to easily grasp that from this visualization? I don't believe they would.

3. My first thought would be to display something similar to a histogram, where overlapping data points are stacked on top of each other to display frequency. If there was no need to

display each individual data point, I would likely lean toward displaying this data as a histogram.

Reply

Sangzun Park (https://iu.instructure.com/courses/2165942/users/6703376)

Thursday

1. Benefits of visualizing with a scatterplot or strip chart

First of all, this type of visualization is simple to present, which means it is easy to create. Additionally, the material is easy to read. Nonetheless, it also represents the data accurately.

2. Drawbacks of this visualization

This is usually beneficial only in non-linear relationships. As data grows, analysis becomes confusing. Especially in the case of scatter plots, it is difficult to check the relationship between three or more variables.

3. Alternatives to overcome drawbacks

In order to compensate for the drawback of difficulty in analysis due to large amounts of data, preliminary work to extract only specific characteristics may be useful. This means reducing the number of data, but retaining data that is key to analysis. If there are many overlapping points, it may be a good idea to randomly add or subtract very small values from the values. It would also be a good idea to use color, size, and shape. I think the Alternative graphs that can complement them are line graphs, cluster representations, or multidimensional (usually 3D) graphs.

Reply

Shruthi Senthilmani (https://iu.instructure.com/courses/2165942/users/6688148)

Thursday

1. Generally, 1D visualizations are easy and clear to visualize and understand due to their simplicity and reduced dimensions. It also helps to analyze the single dimension efficiently to identify trends, and outliers with ease. Uses minimal space and temporal aspects as a time series can represent the trends and patterns in 1D.
2. The drawbacks can include- cluttering of too many data points along a single dimension which may hinder the purpose of visualization. Gives minimal information which sometimes

does not help in analyzing the purpose of the visualization. Another drawback would be the case where there are two values coinciding with each other, it is hard to determine and plot them in 1D.

3. The distribution of the data can be plotted as box plots or violin plots to analyze the mean, median, and how well the data points and outliers are distributed and overcome the problem of coinciding values by showing the median of those distributions(use of histograms can be useful too). We can convert the data points into two dimensions by using kernel techniques and this will help to overcome the drawbacks of 1D separation of data points.

Reply

Akash Patil (https://iu.instructure.com/courses/2165942/users/6699404)

Thursday

1. This 1D visualization is easy to understand as we just have 1 dimension and no complex relationships with other variables.
2. Too many data points can make it cluttered and it might be difficult to tell them apart, essentially it would just become a line.
3. We can use kernel techniques to introduce more dimensions which can create separations in the points. Another way to visualize these points could be distribution plots or boxplots, which will help us understand the points and provide information about their distributions.

Reply

Onur Tekiner (https://iu.instructure.com/courses/2165942/users/6758180)

Thursday

What are the benefits of this visualization method?

It is easy to see the data point's density in the axis value. Even looking at it, we can estimate the dataset's mean without calculating anything.

What are the drawbacks of this visualization method? (If you're not sure, think about having more data points)

If we have many data points, because of data points overwrite on each other, it is hard to see which part's density is higher. For example, if there are two data points and both are "4", if we

draw a period on four in two times. From an outside look, we can see only one period for data points because they were written down exactly on the same point.

What would be alternative methods to overcome the drawbacks? There are numerous ways and I'd like to ask you to think of two directions. First, how can we improve the method *while maintaining the basic concept—showing every data point*? Second, how can we visualize the data in *different ways*? Try to come up with as many solutions you can think of!

If too many values exist in the dataset, it is hard to clearly see data points in the axis.

If dataset discrete values;

I would write down the same values on top of each other on the axis and create histograms manually. It is the best way to see a distribution of the dataset.

if dataset continuous values;

Box plots could also be one of the best data visualizations. You can see data distribution and outliers easily in that way.

Reply

Ao Zhang (https://iu.instructure.com/courses/2165942/users/6703098)

Yesterday

This visualization method is easily to check the distributions of different groups. Like the figure above, we can find see the differences between different groups.

However, there are also some drawbacks of this method. First of all, we are not able to know the specific numbers of different groups. What we are aware of based on the figure is the approximate number. Secondly, if the values of two groups are same, the dots are overlapped in this method. Therefore, it could lead to misunderstanding of datasets. And it is also difficult to show the trend by this chart. We can find that all groups have similar heights of bar.

We can do some improvements of this chart to make it more accurate. Firstly, we can label out specific numbers on the point to show the exact values of them. And we can also use different colors to show the same numbers of two different groups. Alternatively, we can set the y-axis as the number so that we could see the differences among several groups.

Reply

Mukul Gharpure (https://iu.instructure.com/courses/2165942/users/6678592)

Yesterday

Benefits of Strip Chart:

- Clearly displays individual data points.
- Good for identifying outliers.
- Compact for smaller datasets.

Drawbacks:

- Overplotting with larger datasets.
- Doesn't show summary statistics.
- Cluttered with multiple categories.

Improvements:

- Jittering: Reduce overlap by adding slight noise to data points.
- Transparency: Adjust to visualize density.
- Adding Summary Statistics: Include mean, median, or quartiles for more information.

Alternative Methods:

- Histogram: Useful for showing distribution, especially for large datasets.
- Box Plot: Displays summary of distribution, including outliers.
- Violin or KDE Plot: Combines features of box plots and density plots, good for comparing distributions.

These alternatives and improvements can help address the limitations of strip charts, especially when dealing with larger datasets or multiple categories.

Reply

Andrea Chung (https://iu.instructure.com/courses/2165942/users/6443321)

Yesterday

The benefit of drawing scatterplot or strip chart is that we can see trend lines easily. Also, outliers can be easily pointed. However, one of the drawbacks would be that if we have more data points, we would not be easily visualize them. As we have more data points, the graph would only show big chunks of data points and hard to understand the plot. To overcome these drawbacks, one of the alternative methods is bee swarm plot or strip plots. Also, violin plot can be useful as it shows the distribution of the data using summary statistics.

Reply

Aditya Sanjay Mhaske (https://iu.instructure.com/courses/2165942/users/6692144)

Yesterday

Benefits:

- Shows Every Data Point: The most significant advantage is that it displays every individual data point, providing a complete view of the dataset.
- Visual Clarity: Each data point is visually distinct, which aids in pattern recognition.

Drawbacks:

- Overplotting: As the dataset grows larger, the points may overlap (overplot), making it difficult to distinguish individual data points. This can lead to a loss of detail and precision.
- Limited Information: While it shows the distribution of data, it doesn't provide summary statistics or insights into central tendency, variability, or other statistical characteristics.

Alternative Methods:

- To address overplotting, you can introduce slight random variations to the data points' positions along the axis. This spreads the points out horizontally, reducing overlap.
- Making points slightly transparent, overlapping points become visually distinguishable.
- Visualizing the Data in Different Ways is Histogram or Box Plot

Reply

Yumeng Liang (https://iu.instructure.com/courses/2165942/users/6587577)

9:41am

the drawbacks would be when you are dealing with a large amount of data, scatterplots can suffer from overplotting, where points overlap and obscure each other, making it challenging to see individual data points. However, the scatterplots are excellent for visualizing the relationship between two continuous variables. They help identify patterns, trends, clusters, and correlations in the data.

To overcome the drawbacks, we can use different colors to categorize the different data. Also, we can try to use other visualization methods to have a better-visualized graph.

Reply

Renu Jaiswal (https://iu.instructure.com/courses/2165942/users/6704404)

12:33pm

Benefits can be:

- This plot shows the individual points.
- Useful for identifying patterns in data.

Drawbacks can be:

- Not very useful for large dataset.
- Not able to show distribution of whole data.

The scatterplot can be improved by reducing the opacity of points to make overlapping points partially visible. The second thing we can do i aggregate the points and present only summary statistics which be average, median or mode depending on the context of the data.

I believe when dealing with a 1D quantitative dataset the most direct way to visualize it is by creating a histogram. It is good for showing the distribution of data.

Boxplots are also a very good option for the representation of 1D data as they show quartiles and are very useful in showing outliers too.

Reply

Simon Driver (https://iu.instructure.com/courses/2165942/users/6818242)

12:55pm

1. I think that the benefit of this visualization method is that it is relatively clean, sparse, and it is easy to see what value each data point has. It is essentially like have a list of data.

2. I think that the major drawback of this is that as there are more data points in a line, the line will eventually just become a straight line with enough points and we could not even distinguish any individual points any more.

3. I think that to maintain this method and still keep working with the basic concept, we would need to offset the data points. They can't all be in the same line or plane; if they were scatted along the y axis, we could fit more points in.

To visualize the data in a different way, I think we might consider using a histogram to show the frequency of values that fall inside a certain range. This was would lose the concept of seeing

every data point, but it would help us out as we continue to add more data to the visualization.

We could also think about representing this data as a bubble chart, with each bubble corresponding to a range, and the bubble sizes showing the relative frequency of values within that range.

Reply

Madhuri Patibandla (*she/her/hers*) (https://iu.instructure.com/courses/2165942/users/6760559)

1:35pm

1) Benefits of visualization method is simplicity and explicit nature

2) If we have millions of data , Data points will overlap in ID scatter plot and we can visualize the data correctly.

3) Alternative methods are, perform summarization and Aggregation and define it in box or histogram plots to visualize the data and we can use Beeswarm plot to avoid the overlaps.

Reply

Shreedeep Sadasivan Nair (*he/him/his*) (https://iu.instructure.com/courses/2165942/users/6813278)

2:29pm

1.) The benefit of using scatterplot or strip chart for 1D visualization is that you can plot each and every data point in a single plot hence information about the entire dataset is visible in a single plot also you can clearly spot if there are any outliers or not.

2.) The drawback would be that if the dataset was a large one then a lot of points would overlap with each other hence understanding any meaningfull information will be difficult.

3.)We could use the alpha function so that even if points overlap they are still somewhat visible or use histograms to plot the frequencies, we could also use something like a boxplot to visualize the range of the data

Edited by **Shreedeep Sadasivan Nair** (https://iu.instructure.com/courses/2165942/users/6813278) on Sep 23 at 2:32pm

Reply

Jash Shah (https://iu.instructure.com/courses/2165942/users/6684840)

4:00pm

A strip chart has certain benefits as follows:

A scatterplot can identify individual data points and also we can get a preliminary view of the data.

Also we can fit a reg plot and understand the trend between two variables.

we can easily identify patterns, clusters, or outliers in the data. This makes it suitable for detecting trends or anomalies.

Drawbacks of ScatterPlot Visualization:

With a large number of data points or limited space, points can overlap, making it difficult to distinguish individual data points.

High-density scatterplots can become cluttered and hard to interpret, especially when there are many data points in a small area.

Scatterplots are best suited for 1-dimensional data. They become less effective when trying to visualize multi-dimensional data.

Fixing the drawbacks:

Reduce the opacity of data points to handle overlapping. This allows you to see regions of high data density more clearly.

Add a small amount of random noise to the data points along the x-axis (horizontal) to prevent overlap.

Use different colors or shades for data points to indicate categories or groups within the dataset.

Alternative visualizations:

Use histograms and boxplots

Histogram with frequency to understand the distribution of the data and box plots for outliers detection

Reply

Robert Perez (*he/him/his*) (https://iu.instructure.com/courses/2165942/users/6701521)

4:20pm

1. The benefit of the strip chart is that it is simple to read and understand and the data is easy to compare to other data points.

2. The drawbacks of the strip chart are that with a large data set, you could have a lot of closely grouped or bunched data points that will be difficult to individually discern. Some points will overlap with others and the simple readability the chart offers for a smaller dataset would be lost.

3. Binning the data would overcome the drawbacks. In fact, I would probably use a histogram with n number of bins rather than a strip chart. The histogram would make it easier to divide the data into groups; however, you would lose visibility of individual data points.

Reply