**Histograms**

Histograms are a type of plot that takes one continuous variable as its input.

When we can use ?

1. If you have a single continuous variable

2. You want to ask questions about the shape of its distribution. For ex: you might want to know lowest and highest values or which values are

most common

Chart, histogram

Description automatically generatedIt’s important to choose **the bin size** to get good insights, but before plotting you never know which bin size is good. So you’ll have to play with it. The right one is 5 years size. Above one is 1 year.

So, Size matters ! ☺

Chart, bar chart

Description automatically generated

Interpreting **Histograms**

1. **Modality: How many peaks?**

When interpereting histograms first thing to look at it is the modality of the distrubition. That is modality, how many peaks there are ?

There are 3 types of modality

1.unimodal: 1 peak

2. bimodal: 2 peaks

3. trimodal : 3 peaks

Chart, histogram

Description automatically generated

1. **Skewness: is it symmetric?**

Second thing to look at it is skewness of distrubition.

1. A picture containing histogram

   Description automatically generatedLeft-skewed : Has outliers on the left
2. Right skewed: has outliers on the right
3. Symettric: It’s a normal distrubiton

**Kurtosis: how many extreme values ?**

1. Histogram

   Description automatically generatedLeptokurtic : has a narrow peak and a lot of extreme values. Leptokurtic distrubitions are important in finance, because weird stuffs are happening in stock market.
2. A platykurtic: has a broad peak and few extreme values.

X: axis time of the day

Y axis : The agouti caughted by camera

When an animal passed the camera, a photo was taken with a timestamp, so the histogram shows the distribution of the time of day when the agouti were most active.

Chart, histogram

Description automatically generated

***Interpretation: The agouti were most active for a couple of hours after sunrise (6:30am to 8:30am), and before sunset (4pm to 6m).***

**Box Plots**

Box plot split a continuous variable (like age) by a categorical variable (like a royal house).

And allow us to compare the resulting distributions in a space-efficient way.

***Chart, box and whisker chart

Description automatically generated***

Comparisons of Monarchs by house.

Royal houses are ordered from oldest at the top to newest at the bottom. ***Chart, box and whisker chart

Description automatically generated***

**Trend** is visible: Since the Plantagenets in the 14th century,the boxes are moving to right, shows the ages have been increasing.

Goldwing and Blois are appear as a single line, because they had one king from each house. The Anjou had only 3 kings, so there is onlt one whisker not 2.

Notice, House of Denmark shows 1 point. Points are for extreme values, that is, values that are outside of the range of the whiskers. He become king at 53. This was extremely old at 11th century.

***Diagram

Description automatically generated with medium confidence***

Some insights about this box plot and dataset for cigarette consumption per person in the USA

* In 1990, three states were considered to have extreme values in the number of packets of cigarettes smoked per capita
* The median number of packets of cigarettes smoked per capita was below 100 from 1991 onwards
* The lower quartile number of packets of cigarettes smoked per capita decreased every year from 1985 to 1995.

**Scatter Plot**

When you should use?

1. You have 2 continuous variable
2. You want to answer questions about the relationship between 2 variables. If one is increasing does other is increasing as well ?

One important concept is Correlation:

How close are you being able to fit a straight line throught the points?

Chart, scatter chart

Description automatically generated

Red line show perfect negative correlation.X values increases, Y values are decreases. Green line shows perfect positive correlation. X increases Y increases as well.

If your plot doesn’t have clear relationship, you’ll need to be a bit creative for ex: X and Y have a slight negative correlation. Instead of saying Plot looks like a dinosaur.

**Adding trend lines :** The trend line has a close fit to the points . Sometimes it could be terrible fit, for example pic in the below, line completely misses the more expensive homes.

When a straight trend line is a poor fit, one alternative is to use a curve.

**Here you can say, as a area increases, the price increases faster than linearly. Chart, line chart, scatter chart

Description automatically generated**

**Line Plot**

When you should use a line plot?

1. You have **2 continuous** variables
2. You want to answer questions about their relationship
3. Consecutive observations are connected somehow

A screenshot of a graph

Description automatically generated with low confidence

Some insights about this line plot. This describes adoption of different machines during 1920-1980

1. After 1940, adoption of refrigerators was always higher than adoption of stoves
2. In 1930, adoption of automobiles was greater than 50%
3. In 1945,two out of the four technologies had lower adoption than in 1940.

If your data spans many orders of magnitude,choosing a logarithmic y-axis can make it easier to distinguish between the lines.

A picture containing text, screenshot, plot, line

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generated

**Bar Plots**

Bar plots are close relative of box plots.

When should you use a bar plot?

Most common cases:

1. You have a categorical variable
2. You want counts or percentages for each category

Occasionally:

1. You want another numeric score for each category, and need to include zero in the plot.

Example:

Here’s a dataset from the Health Survey for England in 2018.

A screenshot of a computer

Description automatically generated with low confidence

The metric is a percentage rather than a count. Since the percentages of children for each year always add up to 100%, it’s helpful to stack the bars on top of each other. In 2001, for example, you can see that the bottom two blocks reach 25 %, meaning that 25 % children ate at least four pieces of fruit and veg per day in that year.

In 2003, The UK government started a campaign to encourage to eat five portions of fruit and veggies per day. Look at the bottom blocks, and notice that the percentage of children eating five portions increased each year from 2003 to 2006, and stayed roughly constant until 2014. Similarly, the pale block at the top of the plot show that the percentage of children eating zero portions per day decreased from 2003 to 2006 then stayed constant.

It looks like campaign was successful.

A picture containing text, screenshot, line, rectangle

Description automatically generated

**Dot Plots**

When should you use dot plots?

1. You have a categorical variable
2. You want to display numeric scores for each category on a log scale, or
3. You want to display multiple numeric scores for each category

A picture containing text, screenshot, font, diagram

Description automatically generatedA screenshot of a graph

Description automatically generated with low confidence

**Higher Dimensions**

X and Y are not the only dimensions.

Points also have these dimensions

1. Color
2. Size
3. Transparency
4. Shape

Mostly Best practice is to distinguish with color.

Now let’s learn how to choose color for our vizualizations.

A picture containing text, screenshot, font

Description automatically generated

Let’s look at wrong color choice here.

1. Yellow points are harder to see than reds
2. Blue and purple are perceptually quite close

The colorscpace designed for data visualization is called hue-chroma- luminance, or HCL.

It’s designed to deal with issues of color perception.

1.Hue is like the rainbow, from red through orange, green and blue, to purple and back to red.

2. Chroma is the intensity of the color, from grey to a bright color

3. Luminance is the brightness of the color, from black to white. (For example, you can go from black through cyan to white, or black through red to white)

A picture containing text, screenshot, font, colorfulness

Description automatically generated

**Types of color scale:**

1. **Qualitative palette example**

When choosing the color for your plot, you can pick one of three types of color scale. Qualitativee color scales are used to distinguish unordered categories, A picture containing text, plot, screenshot, line

Description automatically generated

Easy to distinguish, because they different hue, but constant chroma and luminance .

1. **Sequential – when you need to show ordering (Chroma and Luminance should vary while keeping hue fixed)**

To empahasize ordering in your data, that is, to show that values are greater than or less than each other, you need to sequential color scale

A picture containing text, screenshot, line, font

Description automatically generated

1. **Diverging – show above or below a midpoint ( change chroma or luminance with 2 hues )**

It’s similar sequential color palette but in the middle we have white color or gray, and have increasingly bold colors with different hues on either edge.

A picture containing text, screenshot, font, design

Description automatically generated

**Plotting many variables at once**

**Pair plots**

When should you use pair plots?

1. You have up to 10 variables (either continuous, categorical or a mix)
2. You want to see the distribution for each variable
3. You want to see the relationship between each pair of variables

A screenshot of a graph

Description automatically generated with medium confidence

**Correlation Heatmap**

When should you use pair plots?

1. You have lots of continuous variables
2. You want to a simple overview of how each pair of variables is related

A picture containing text, screenshot, diagram, font

Description automatically generated

When 2 feautures consistently received similar scores by customers, they are positively correlated and colored in a more vibrant red.

**Parallel coordinates plot**

When should you use ?

1. You have lots of continuous variables
2. You want to find patterns across these variables, or
3. You want to visualize clusters of observations

A picture containing screenshot, line, design

Description automatically generated

Each line represents one country, and each continuos variable appears on the x-axis, just like a bar plot. To make each variable comparable, the y-axis simply ranges from the lowest to the highest value. This plot is a mess.

**Pie Plot**

*Pie plots* (sometimes called pie charts) are extremely popular, but often difficult to interpret. They are just bar plots converted into polar coordinates, and humans are generally worse at perceiving angles accurately compared to lengths.

Measures of a good visualization

1. How many interesting insights can your reader get from the plot?
2. How quickly can they get those insights?

Chartjunk is anything in a plot that distracts from getting insight

Chartjunk refers to anything gin the plot that makes it harder for the reader to get insight into the data.

1. Pictures
2. Skueomorphisim: reflections, shadows and etc
3. Extra dimensions
4. Ostentatious colors or lines

**Summary:** *We learned that higstograms are excellent for showing the distribution of a continuous variable and that box plots can compactly show the distibutions of lots of continuous variables.*

*Scatter plots can show the relationship between two continuous variables, and line plots are great for showing trends over time.*

*Bar plots show counts or proportions split by categories, and dot plots will do the same, but allow for logarithmic scales and showing multiple metrics at once.*

*We saw that using colors or multiple panels are ofthen the best way to add a third dimension to you plot, since 3D plots are hard to interpret.*

*We also learned that there are 3 types of color scale: qualitative, sequential and diverging. For cases where you need to analyze many varuables at once, we saw three types of plots.*

*Pair plots show relationships between each pair of variables, correlation heatmaps show related variables, and parallel coordinates plots show patterns across many variables.*

*We learned that plots with polar coordinates are usually a bad idea, but they have niche uses when data is cyclical, like a time of day. We also learned that using dual axes is almost always misleading, and minimalism is a good idea. We should eliminate anything from the plot that distract from interpretations.*