

visualizing data ①

The goal of visualizing data

- Know what you have
- Get ideas for how to investigate the data

Knowing What You Have

- Missing data
- Attribute values that are different orders of magnitude
- Non representative / Skewed attributes

These can affect the quality of what you infer from the data

visualizing data ③

How should data be visualized?

There is no recipe.

Here are some guidelines...

visualizing data (4)

1) Get the data into tabular form. Something you can easily manipulate.

- Excel worksheet
- Orange data table

visualizing data ⑤

2) Get a handle on the

structure of the data

- Number of rows
- Number of columns/features
- Types of features
 - Numerical
 - Categorical
- Types of information contained in the features

Visualizing data ⑥

3) Visualize numerical features

What

- Summary Statistics
(Range, Distribution, Percentiles)

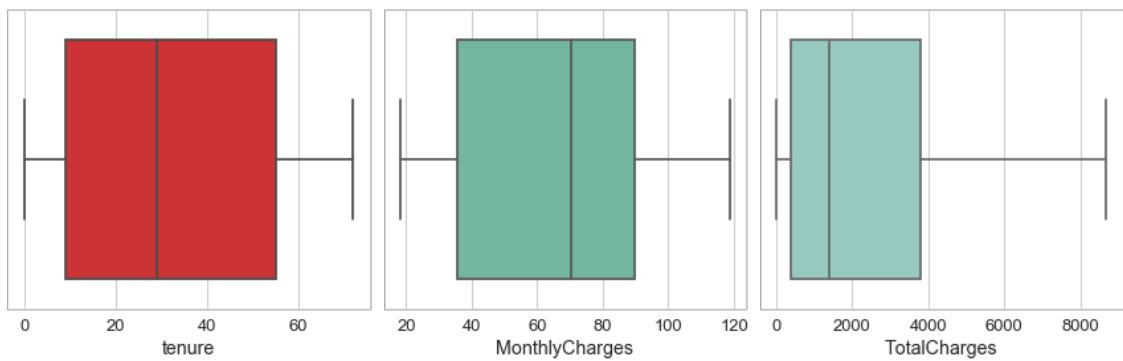
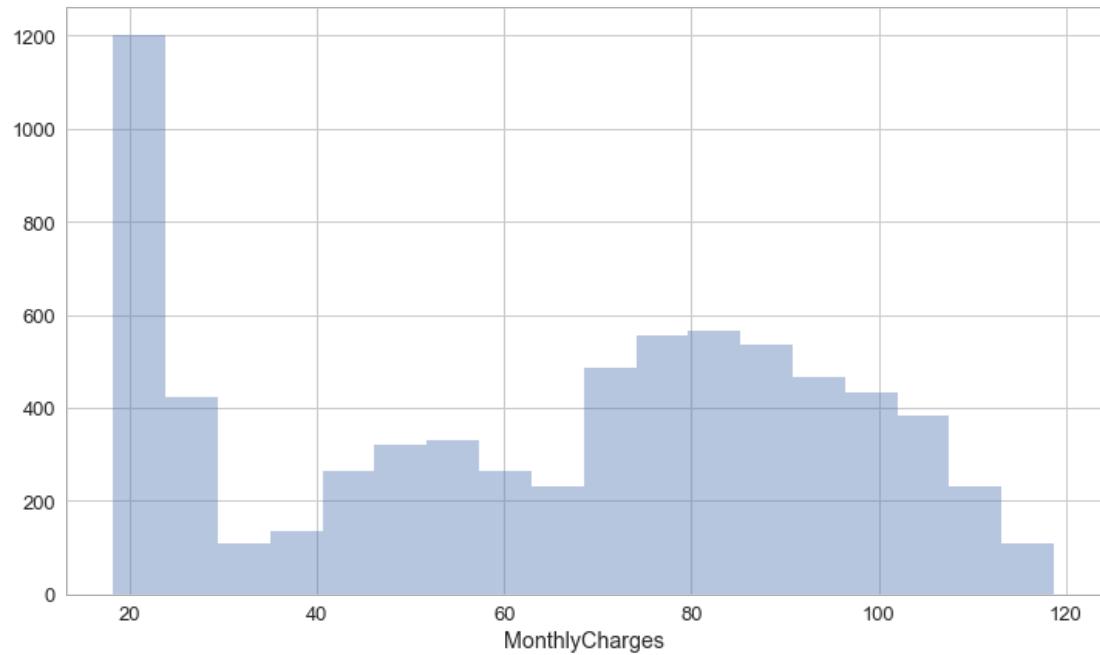
How

- Histogram
- Box Plot

Look For

- Outliers
- skew

visualizing data



visualizing data ⑧

4) Visualize relationships between numerical features.

What

- Comparing quantities

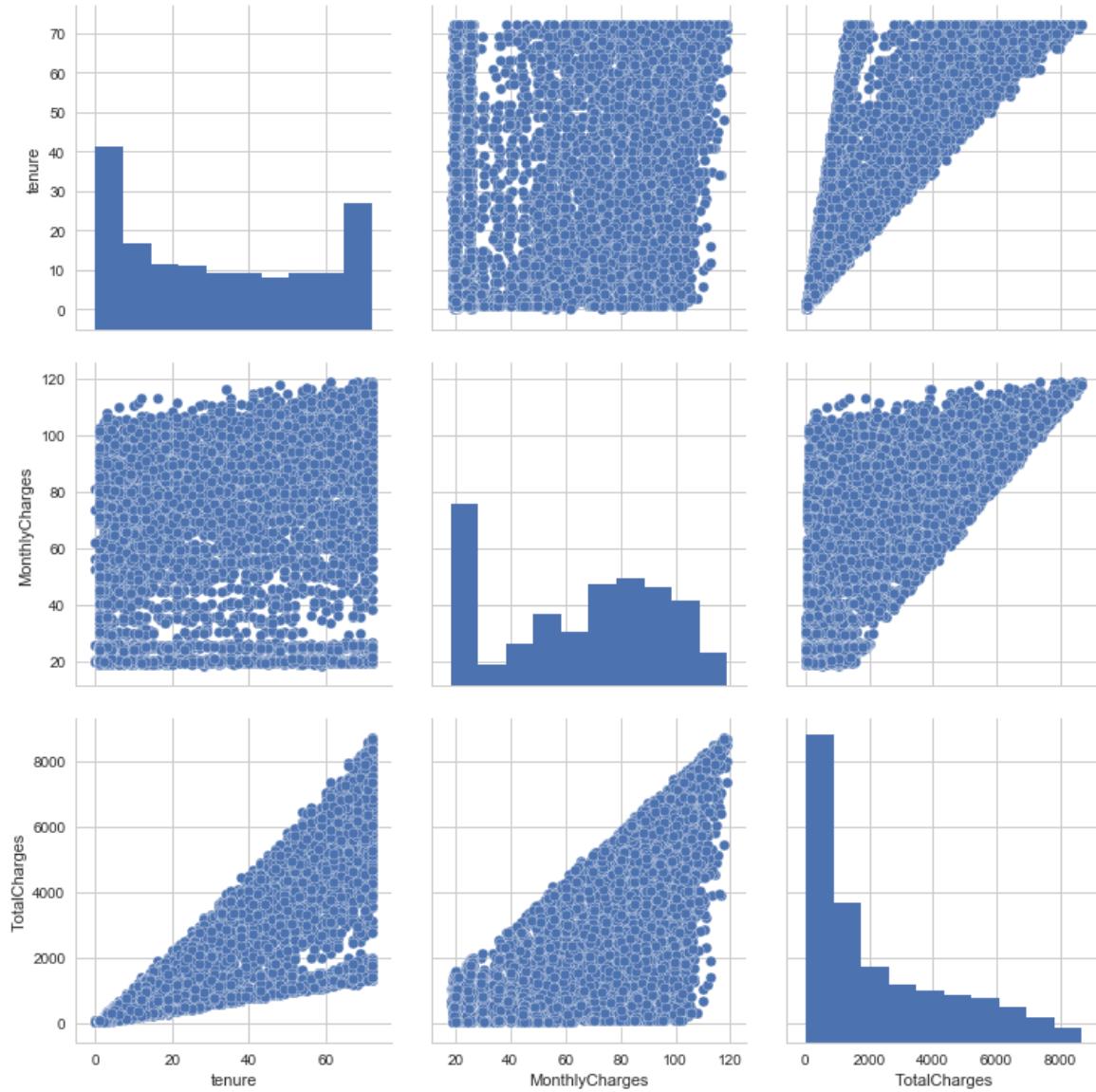
How

- Scatter plot
- Correlation table
- Bar chart
- Overlapping histograms

Look for

- Correlations and anti-correlations
- Differences in order of magnitude

visualizing data ⑨



5) Visualize categorical features.

What

- Unique values within each category
- Distribution of values within each category

How

- Bar graph

Look for

- How values are distributed
- Representativeness of values

visualizing data ⑪

6) Visualize relationships between
Categorical features

What

- Which values occur together
(or not)? How frequently?

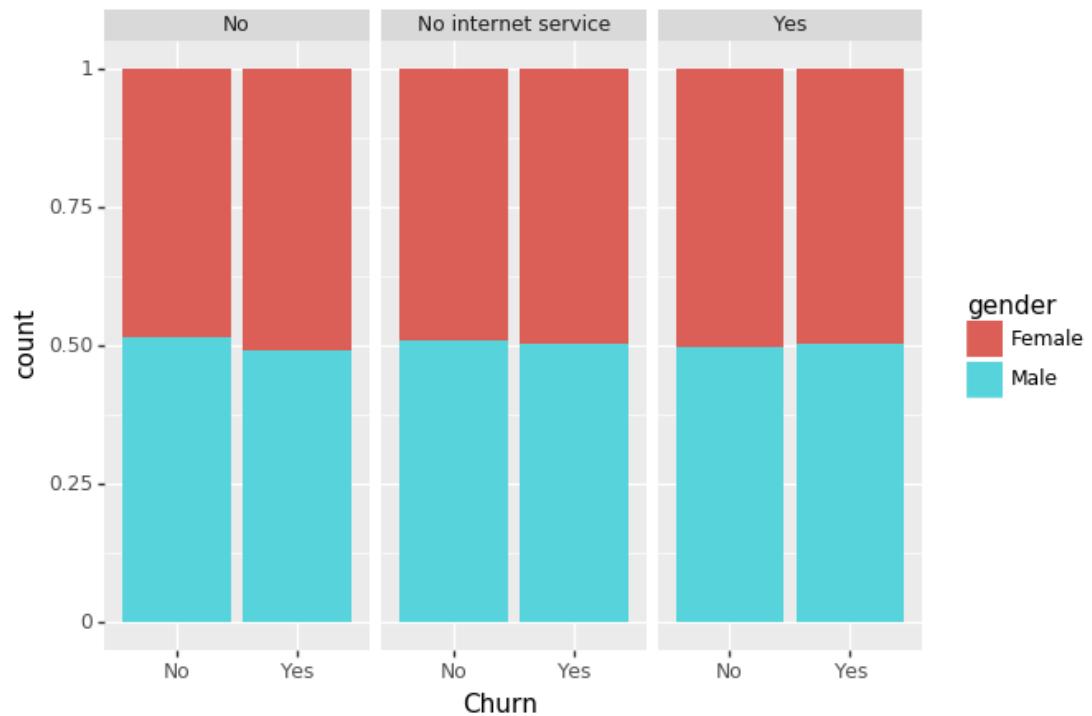
How

- 2×2
- Grouped bar chart
- Facet / Pivot plot

Look for

- Strong and weak relationships
- Confounding attributes

visualizing data (12)



7) Visualize relationships between numerical and categorical attributes

What

- Range of numerical values within each category

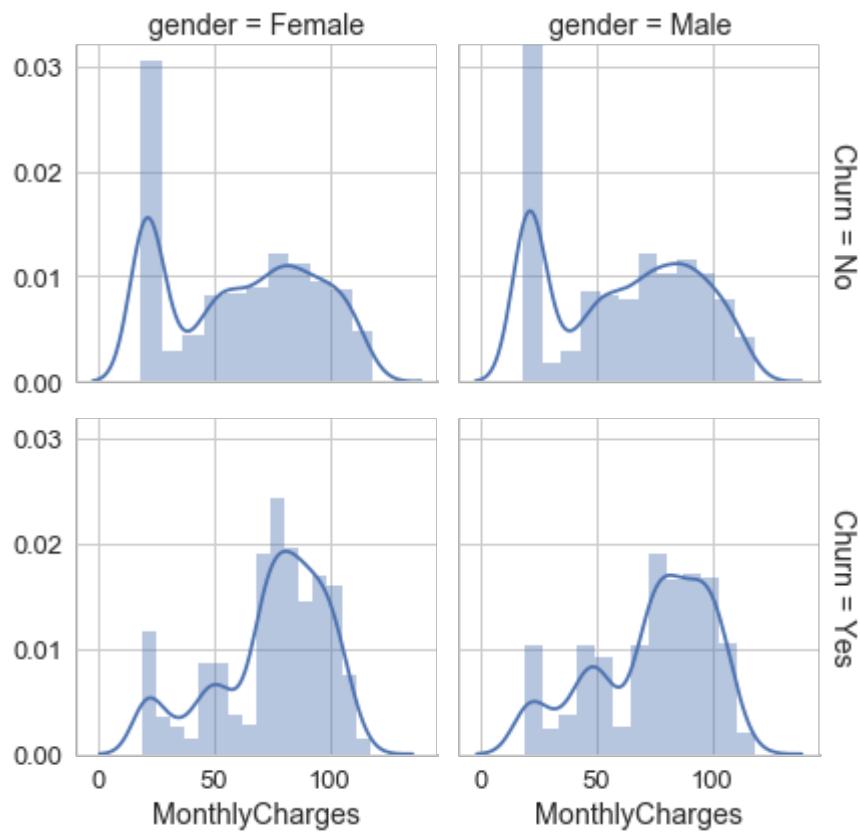
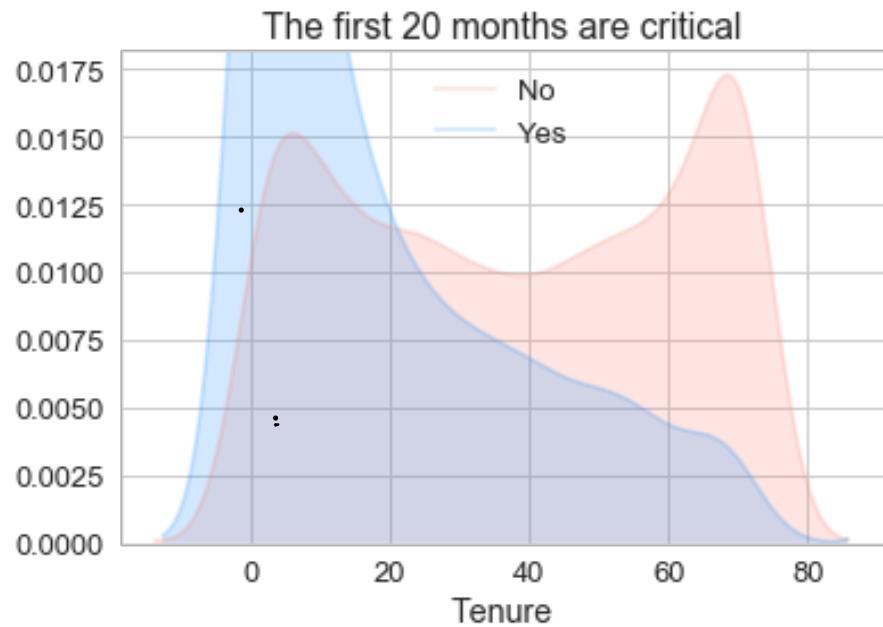
How

- Facet plots
- Density Plots
- Jitter plots

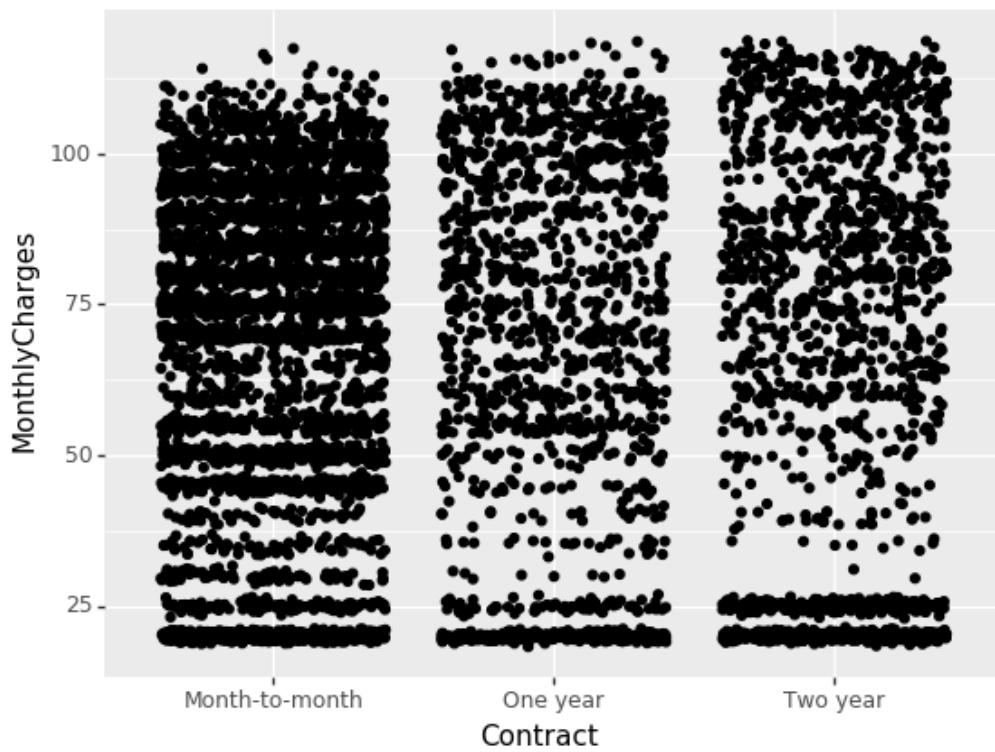
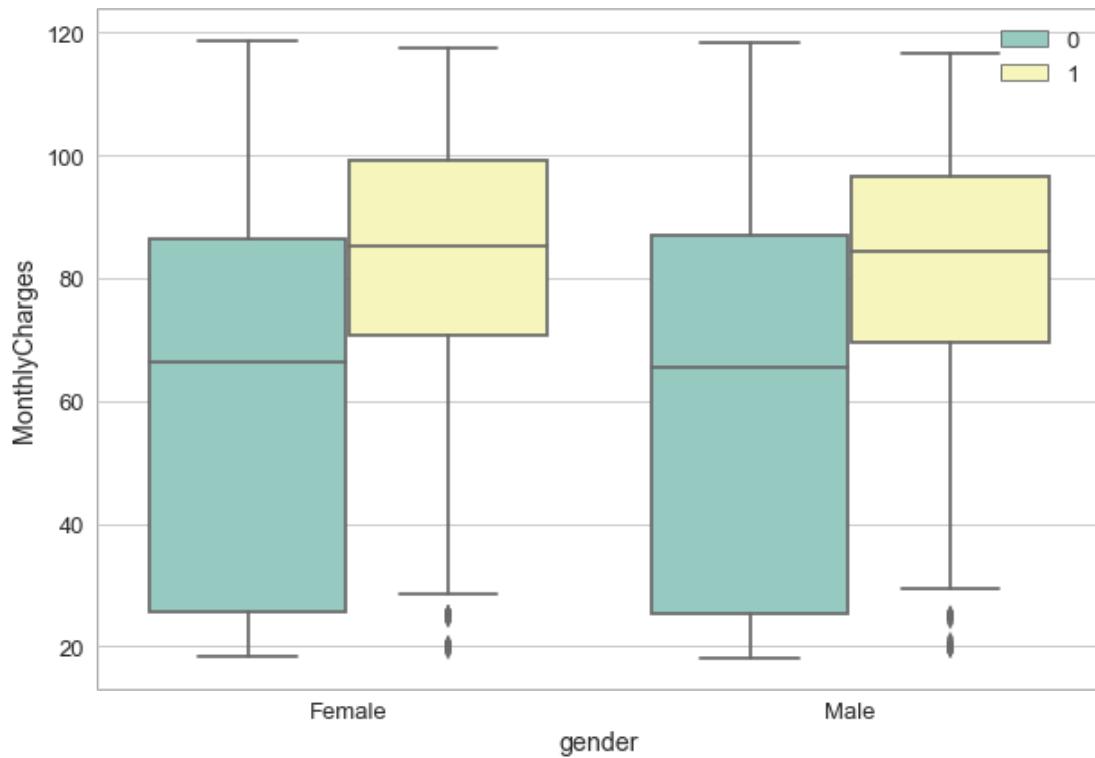
Look for

- How numerical values are distributed as categorical values "switch on" or "switch off"

visualizing data (14)



visualizing data 15



visualizing data (B)

8) For both numerical and categorical data, find and handle missing values.

- Empty cells ''
- Truly empty cells

Figure out how to handle missing values.

- Remove Column(s)
- Remove Row(s)
- Impute values
 - mean
 - median
 - mode

visualizing data (17)

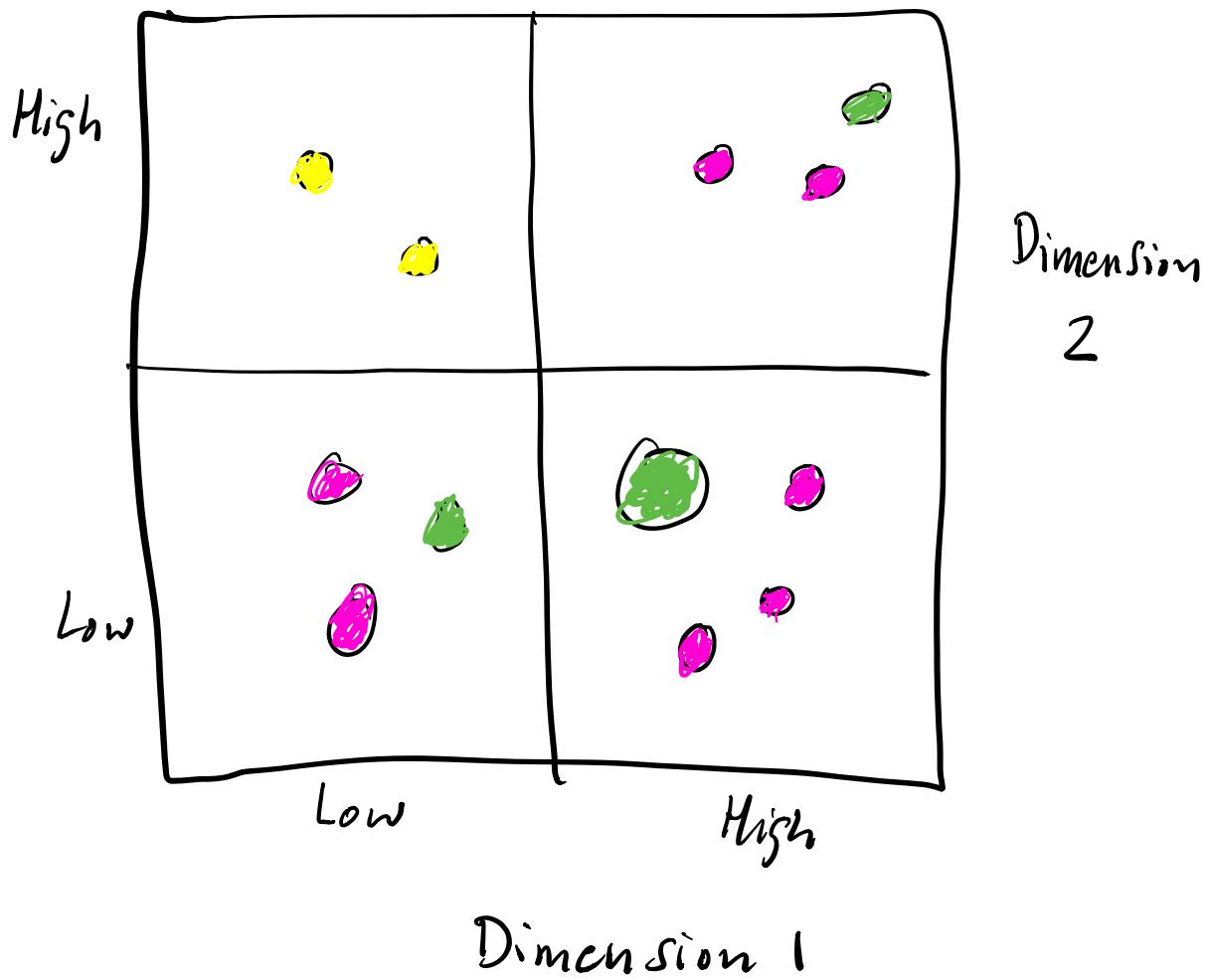
What happens when there are more than 3 features to visualize together?

There are a few tricks we can use to display as many as 7 dimensions (without causing eye strain) on a piece of 2-D paper.

Let's see how...

2 x 2 chart

visualising data (18)
4 Dimensions

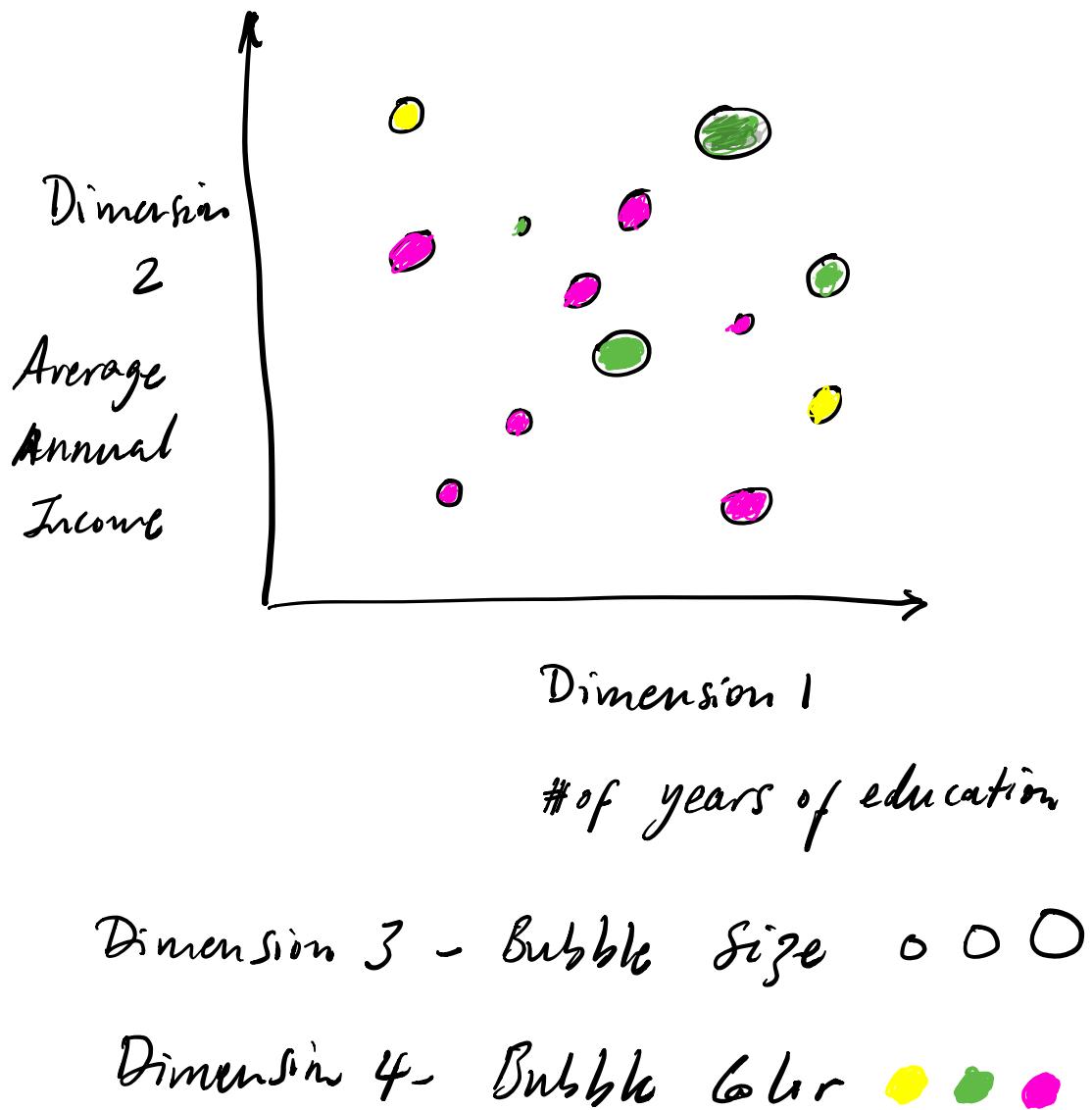


Dimension 3 - Bubble size ○○○

Dimension 4 - Bubble color ● ● ●

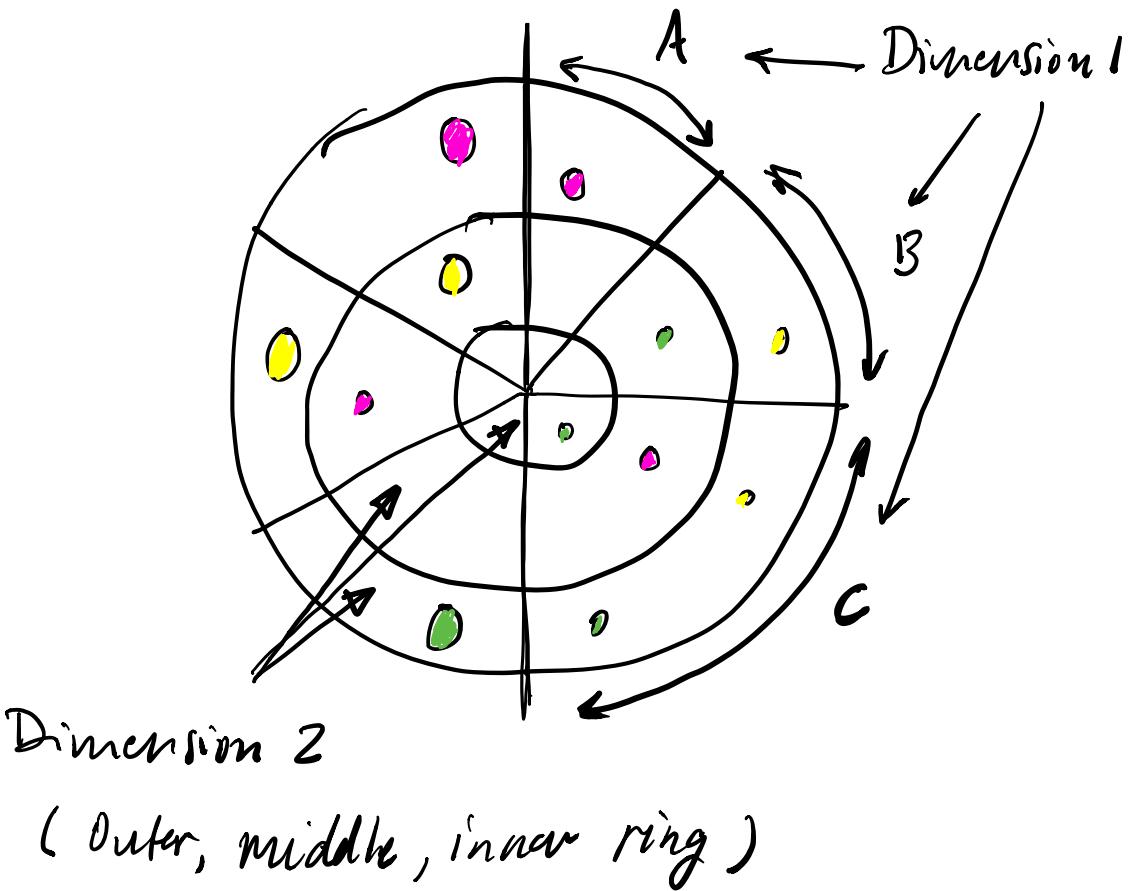
visualizing data (19)

Bubble Chart - 4 Dimensions



Bulls eye chart - 4 Dimensions

visualizing data (20)

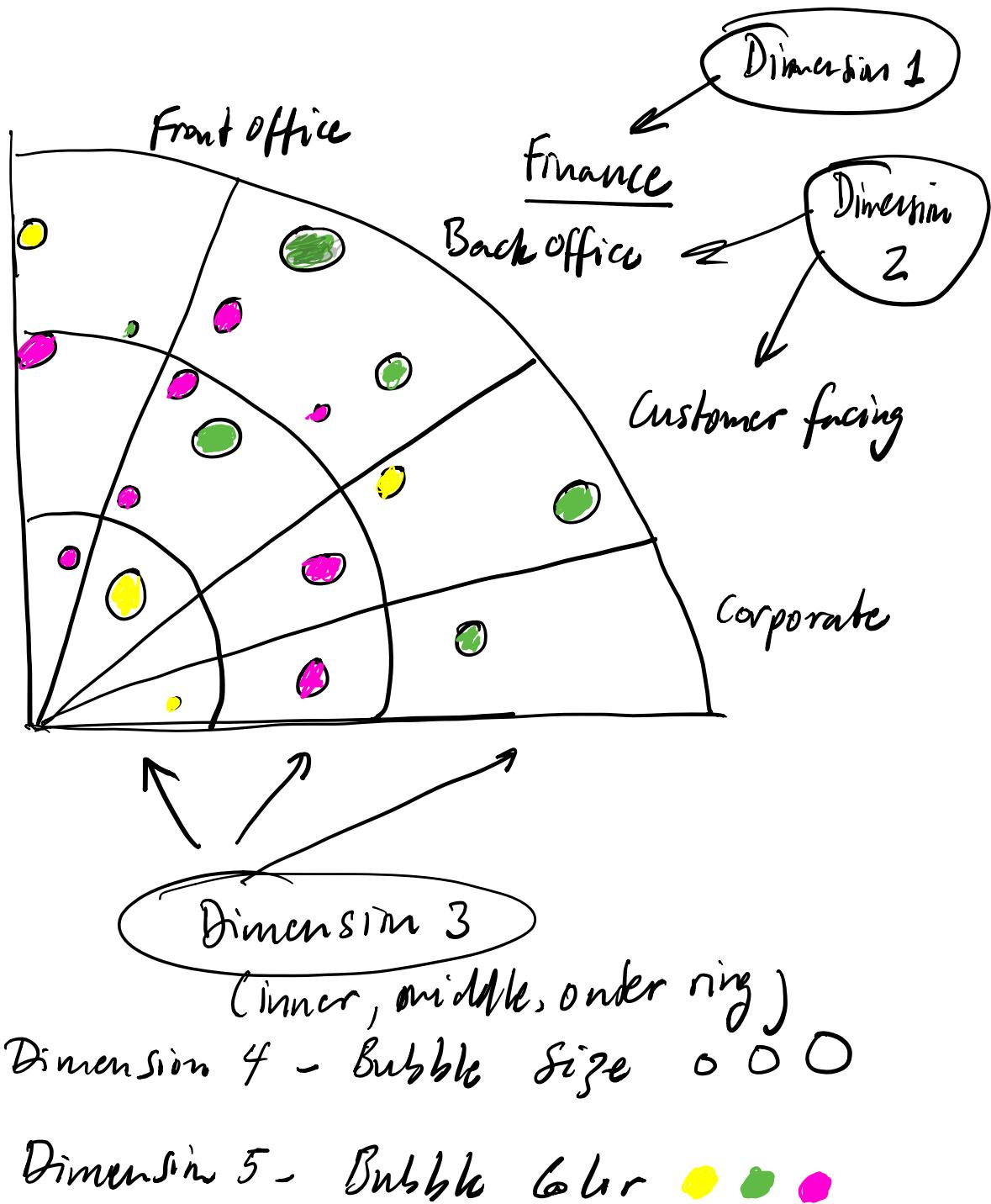


Dimension 3 - Bubble size ○○○

Dimension 4 - Bubble Color ●●●

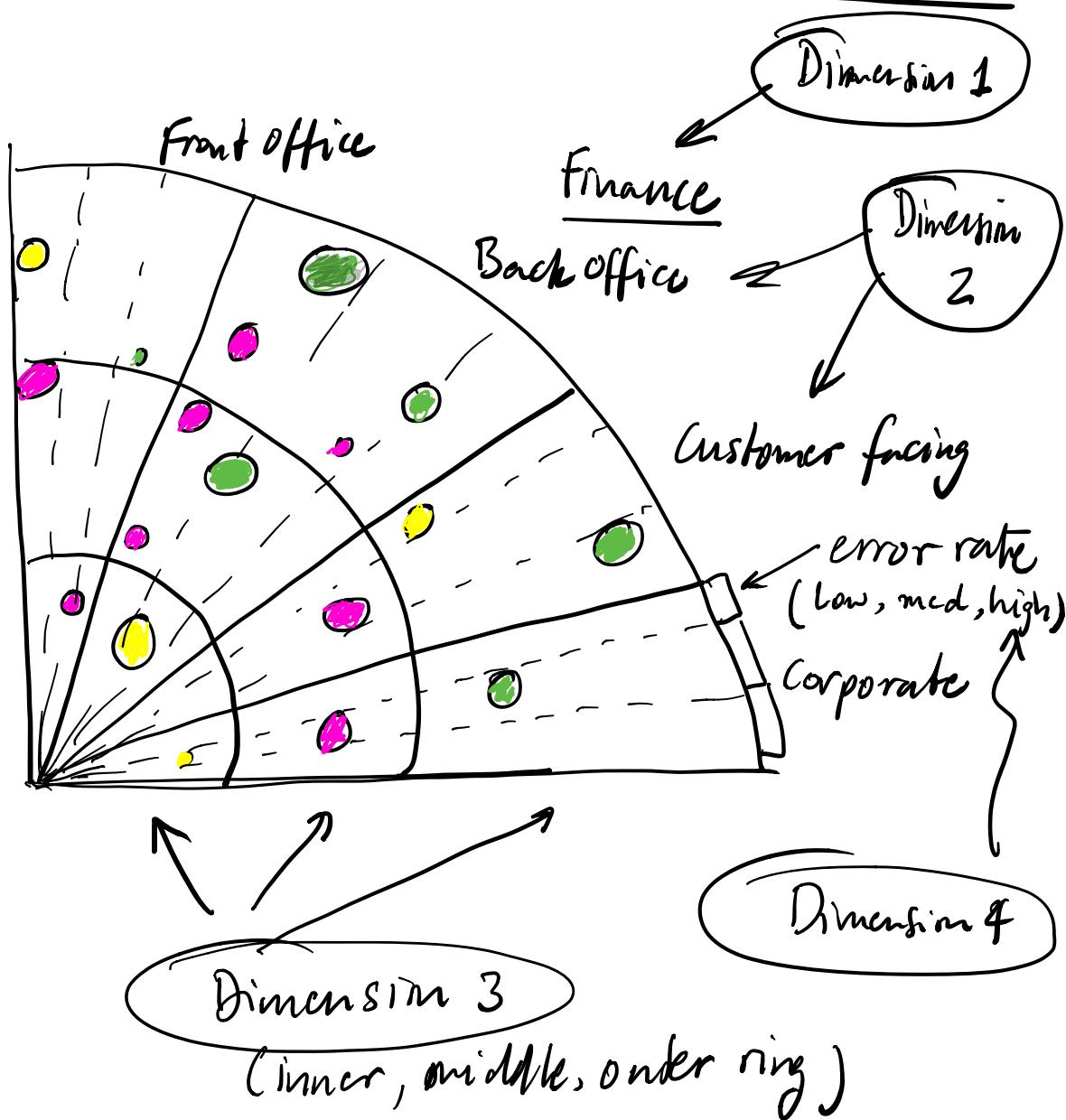
visualizing data (2)

Bullseye Chart - 5 Dimensions



Visualizing data (22)

Bulls eye Chart - 6 Dimensions



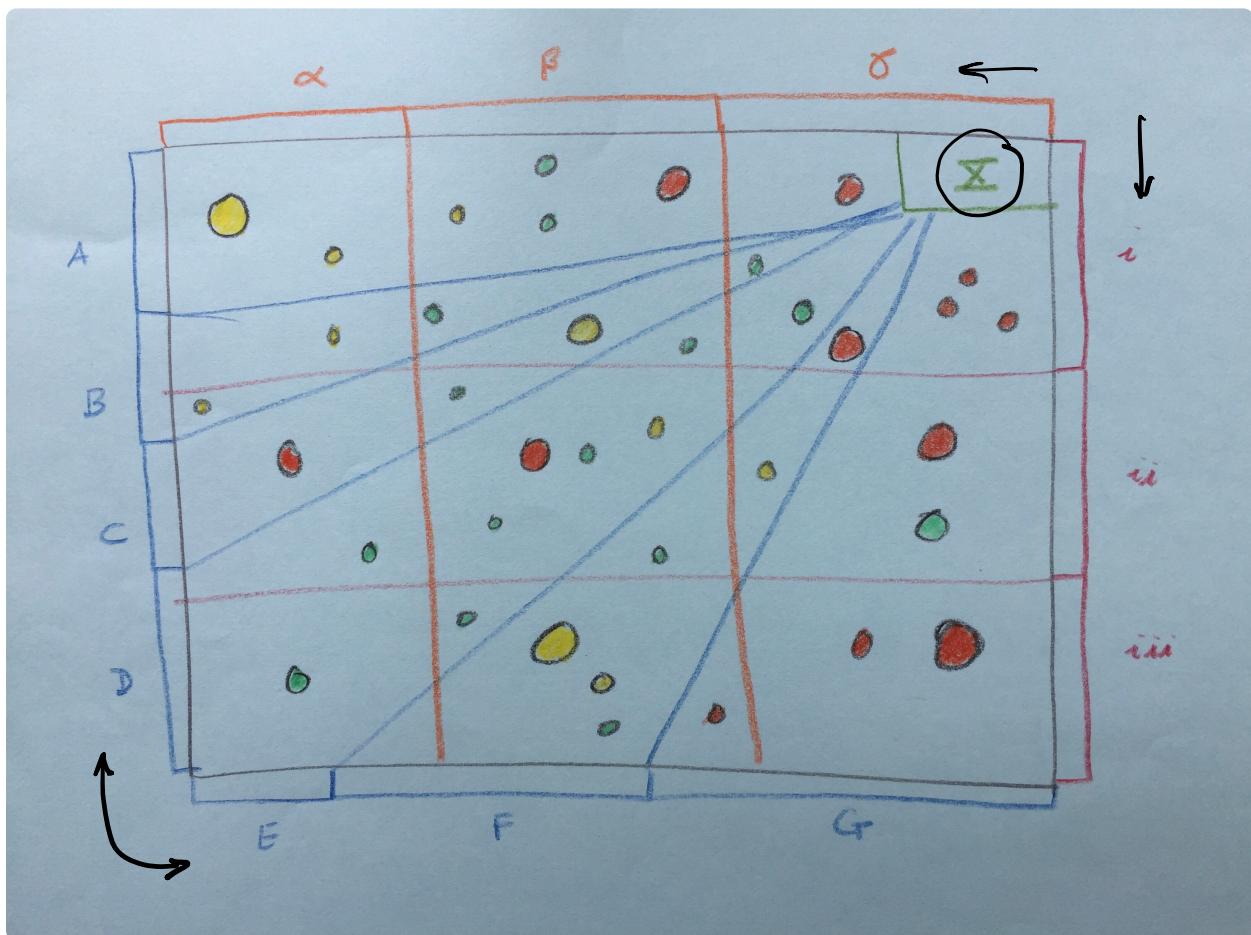
Dimension 5 - Bubble size ○ ○ ○

Dimension 6 - Bubble color ● ● ●

Visualizing data (23)

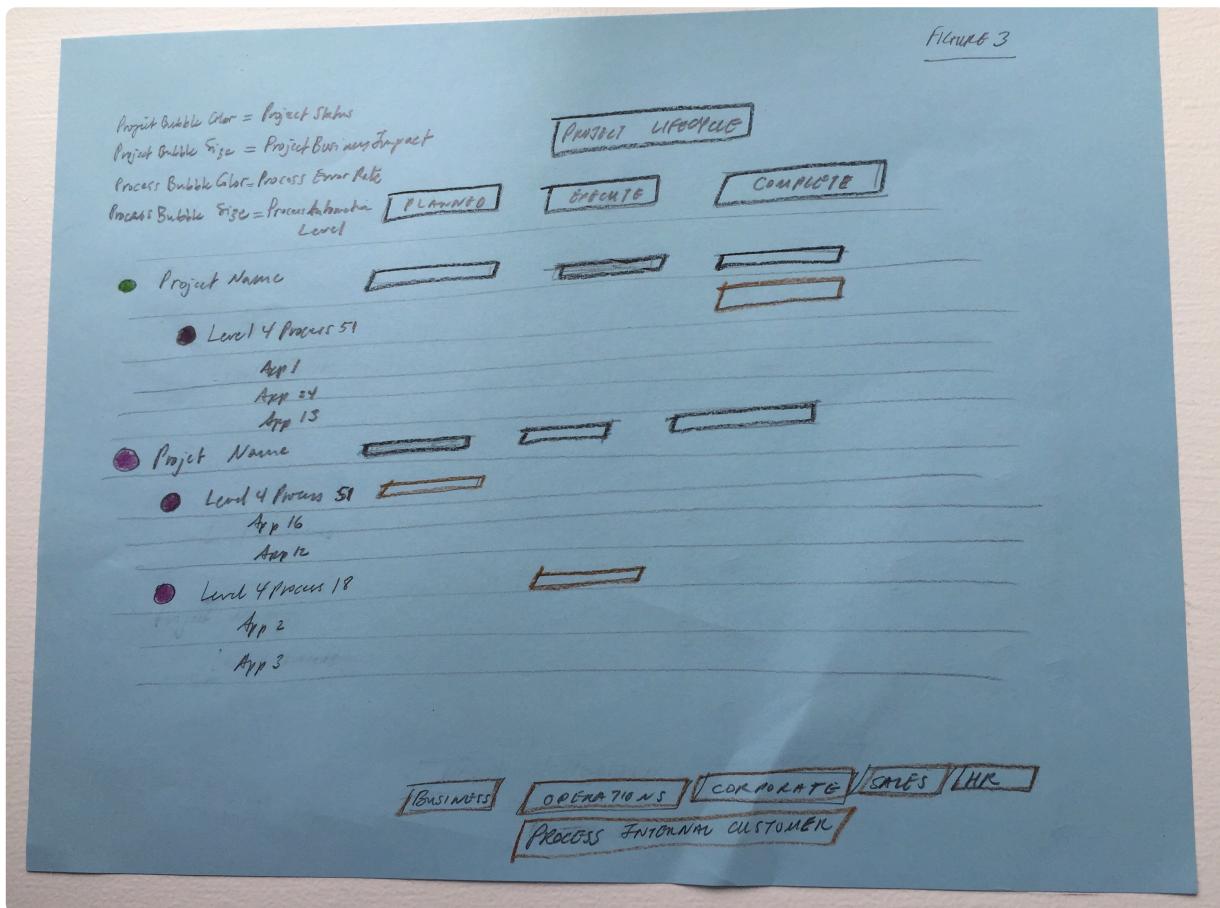
Roadmap Diagrams

6 + Dimensions



visualizing data (24)

Gantt Chart - 7 Dimensions



visualizing data (25)

But what if you have 10, or 20,
or 100 dimensions?

(In machine learning problems it's
not unusual to have 100,000
dimensions!)

Answer: Find out if a select
handful of features matter more than
the rest and use these features to
visualize the dataset.

Some common techniques for finding relevant features (if they exist).

- Domain knowledge
- Convert / transform certain features and drop the rest
- Combine / transform one or more features
- Reduce the dimensionality of the dataset using compression algorithms.

We'll cover these techniques in the session on Feature Engineering

