SYSTEMATIC APPROACH

TO

VISUALIZING DATA

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

How should data be visualized?

There is no recipe.

Here are some guidelines...

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

- Excel worksheet
- Orange data table

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

3) Visualize numerical features

What

- Summary Statistics
  (Range, Distribution, Percentiles)

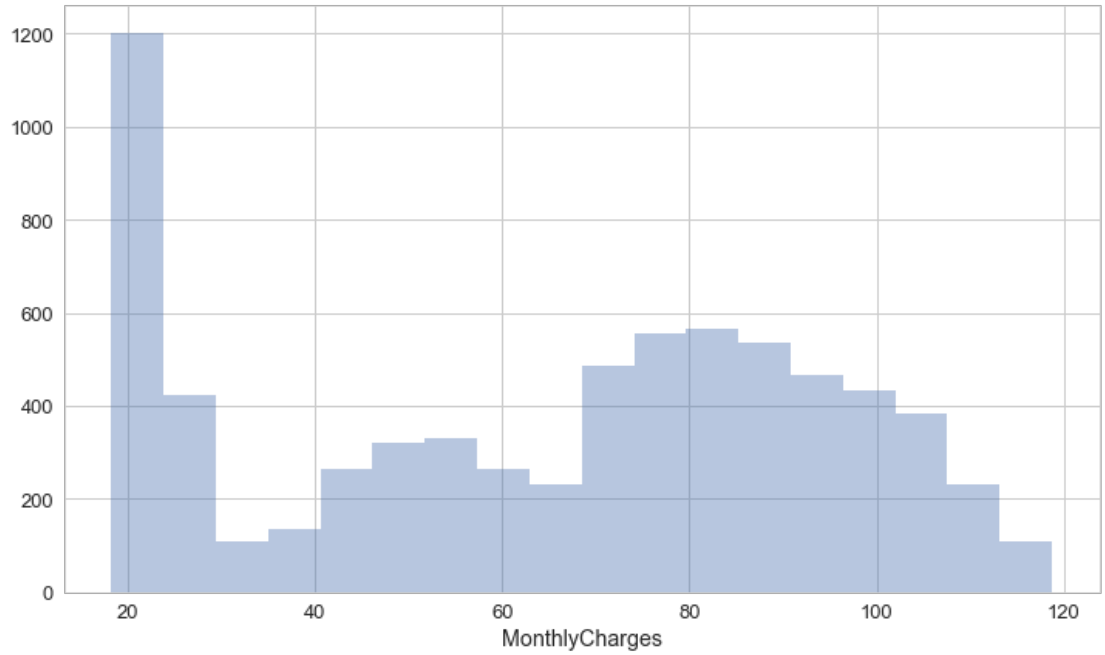
How

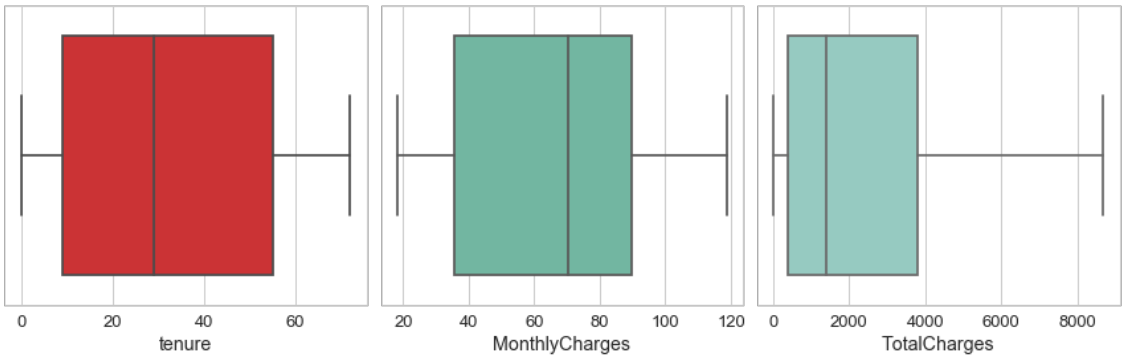
- Histogram
- Box Plot

Look For

- Outliers
- Skew

visualizing data 7





4) Visualize relationships between numerical features.

What

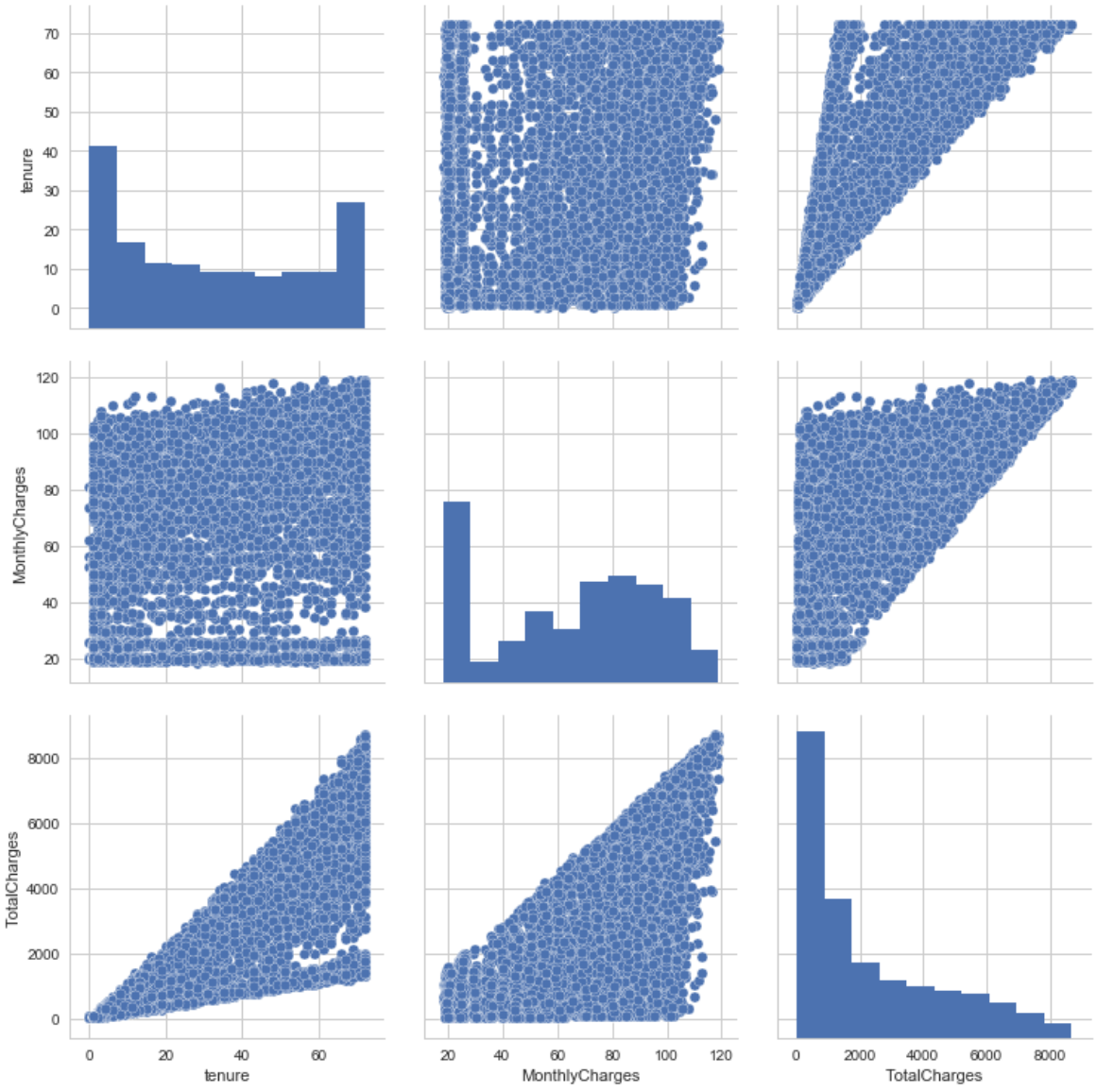
- Comparing quantities

How

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

Look for

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


tenure
MonthlyCharges
TotalCharges
0
10
20
30
40
50
60
70
20
40
60
80
100
120
0
2000
4000
6000
8000
0
20
40
60
tenure
20
40
60
80
100
120
MonthlyCharges
0
2000
4000
6000
8000
TotalCharges

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

6) Visualize relationships between Categorical features

What

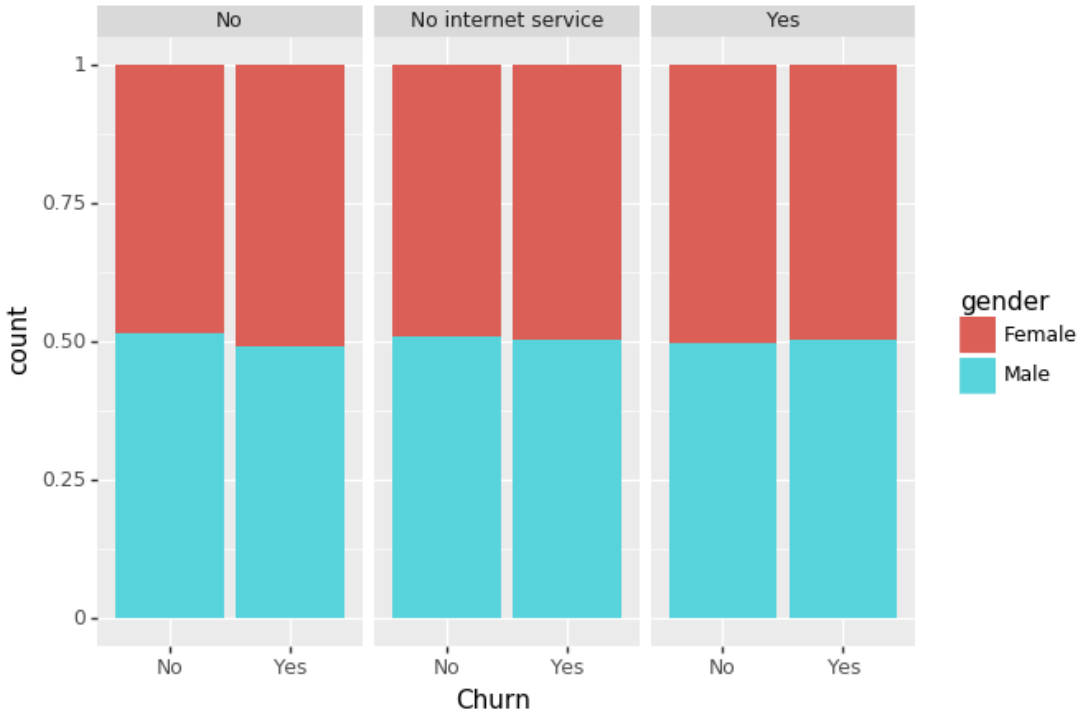
- which values occur together (or not)? How frequently?

How

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

Look for

- strong and weak relationships
- confounding attributes


No
No internet service
Yes
1
0.75
0.50
0.25
0
count
No
Yes
No
Yes
No
Yes
Churn
gender
Female
Male

7) Visualize relationships between numerical and categorical attributes

What

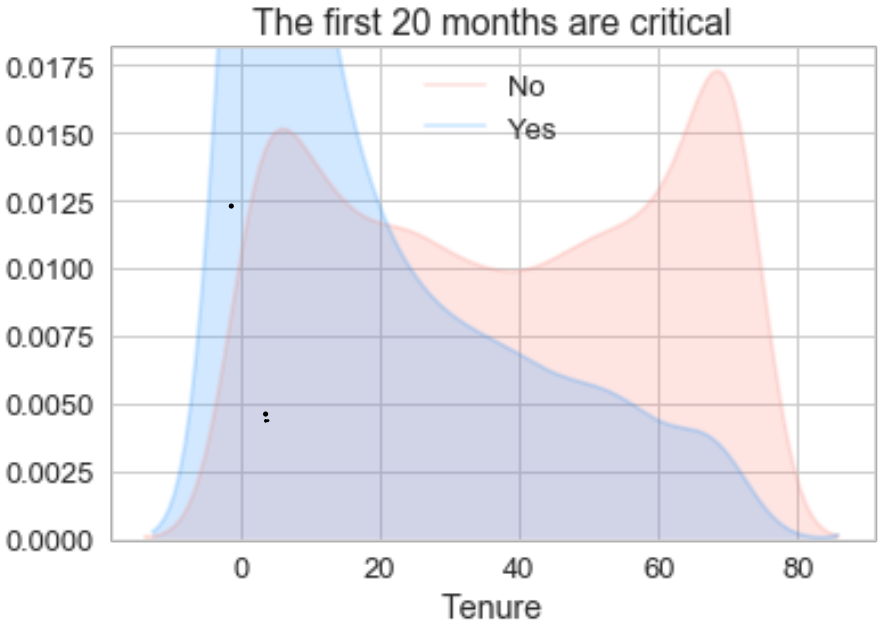
- Range of numerical values within each category

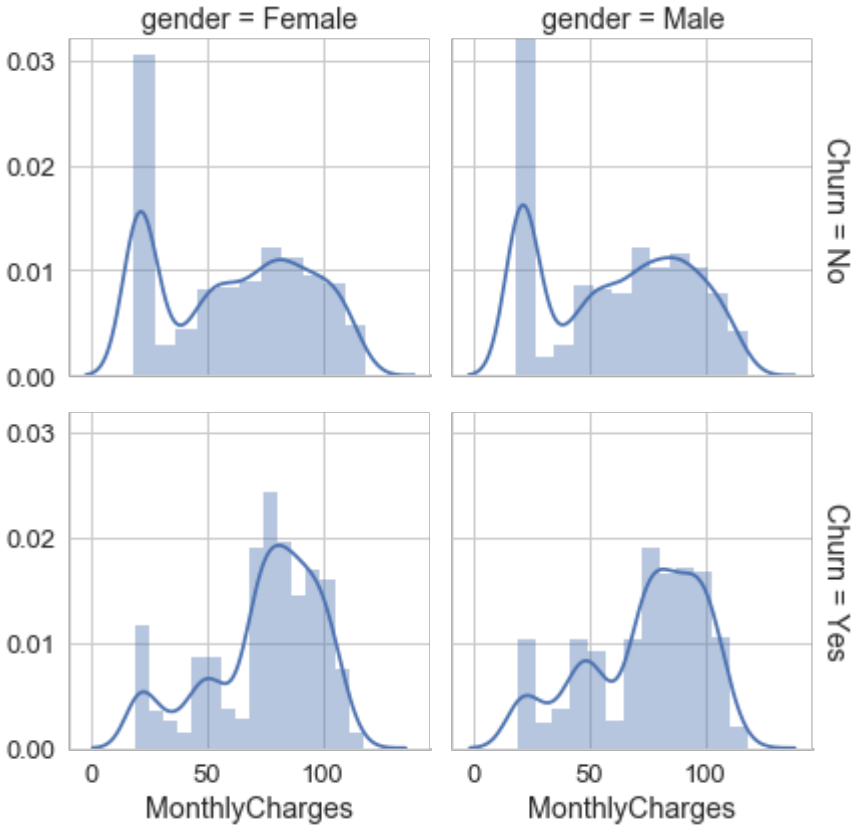
How

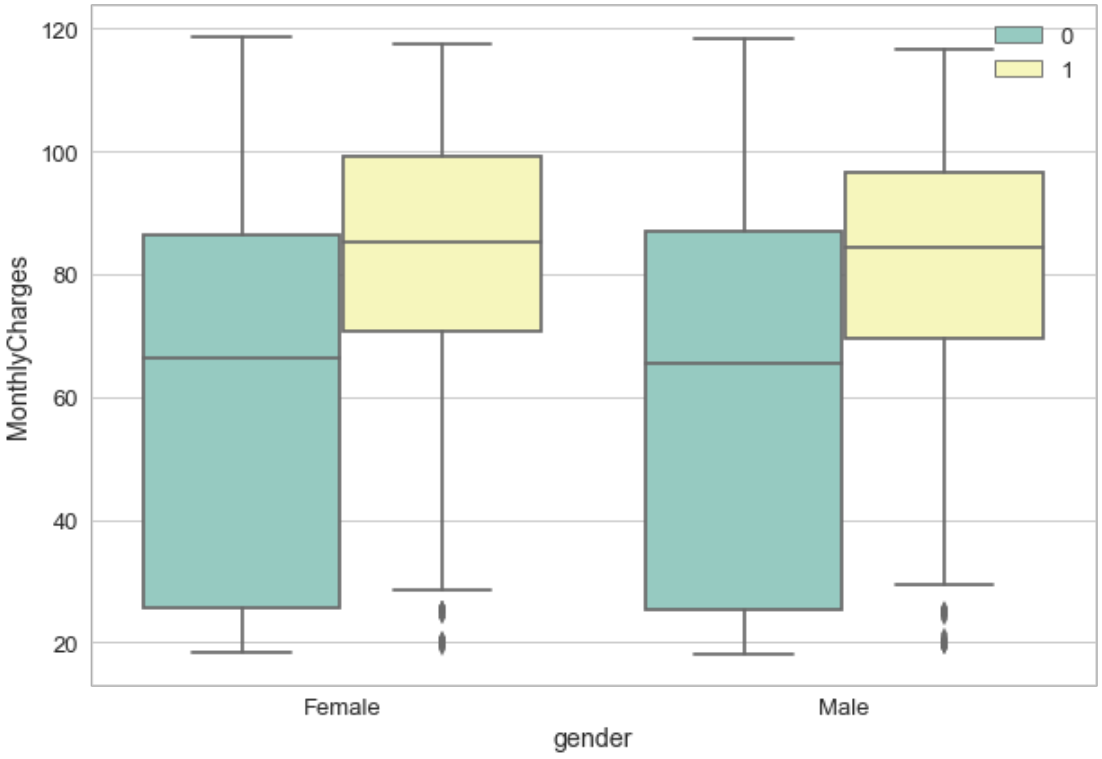
- Facet plots
- Density Plots
- Jitter plots

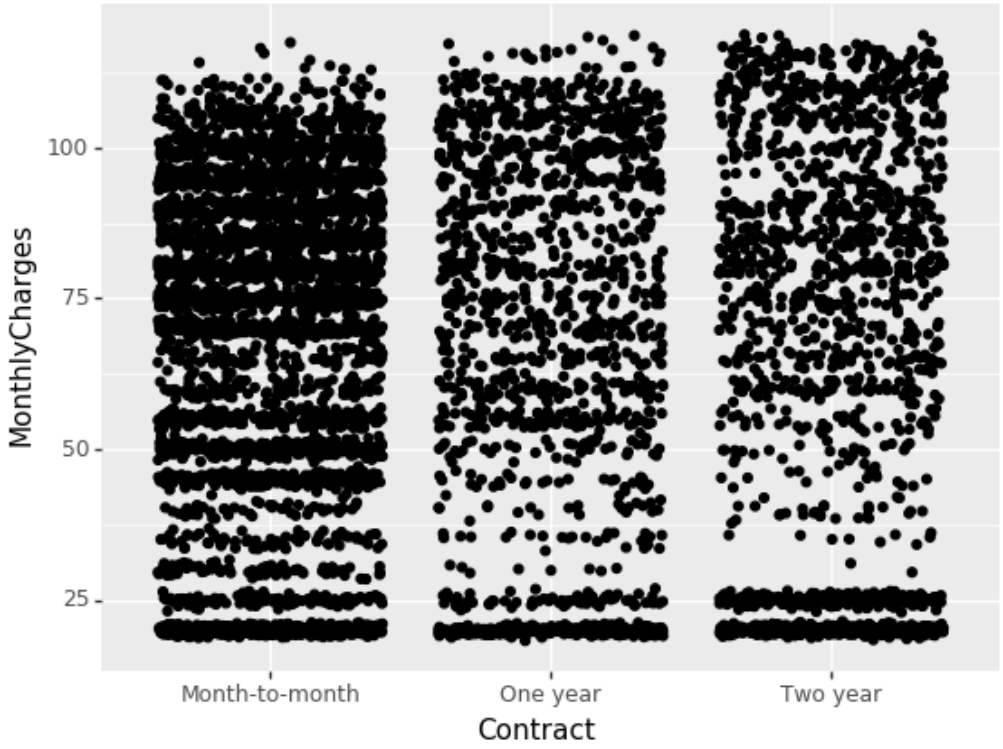
Look for

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


The first 20 months are critical
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
No
Yes
0
20
40
60
80
Tenure


gender = Female
gender = Male
Churn = No
Churn = Yes
0.03
0.02
0.01
0.00
0
50
100
MonthlyCharges


120
100
80
60
40
20
MonthlyCharges
0
1
Female
Male
gender


100
75
50
25
MonthlyCharges
Month-to-month
One year
Two year
Contract

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

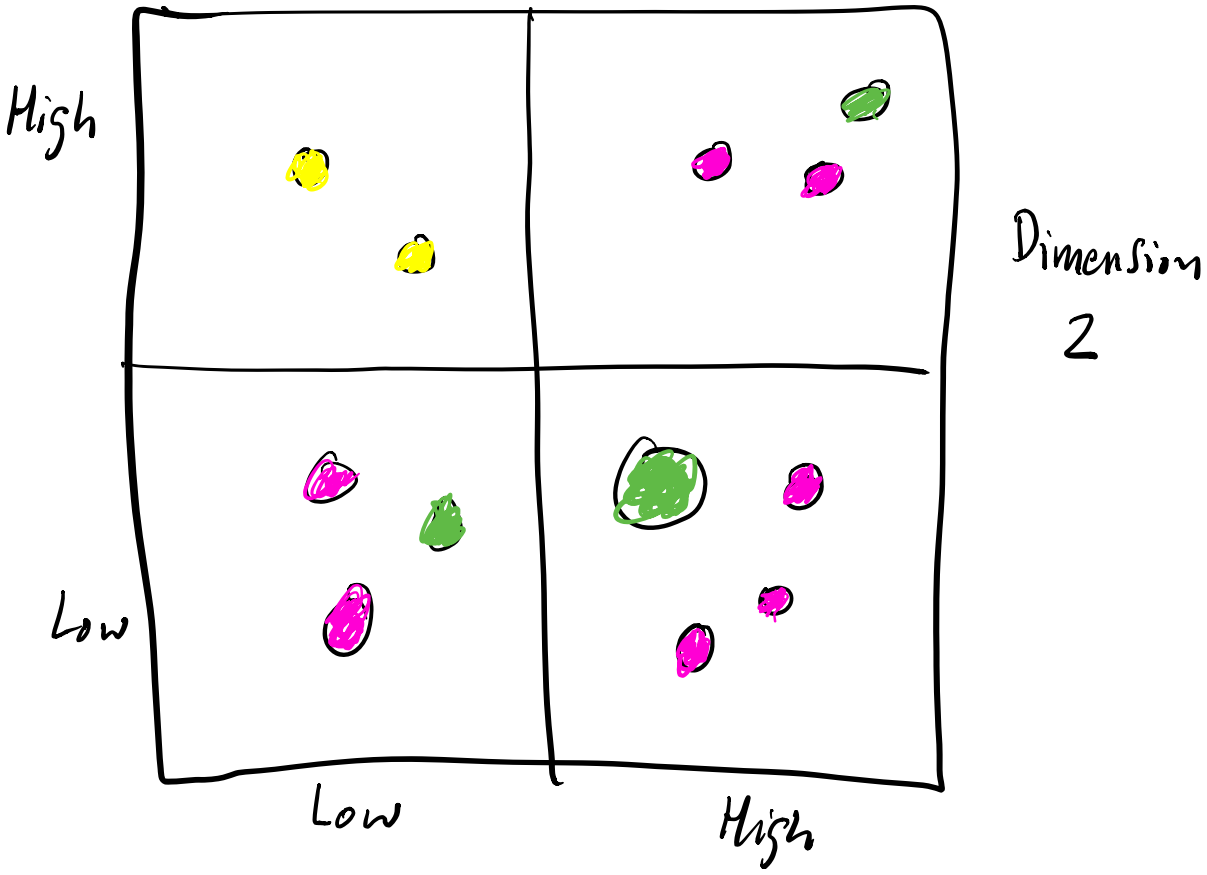
What happens when there are more than 3 features to visualize together?

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

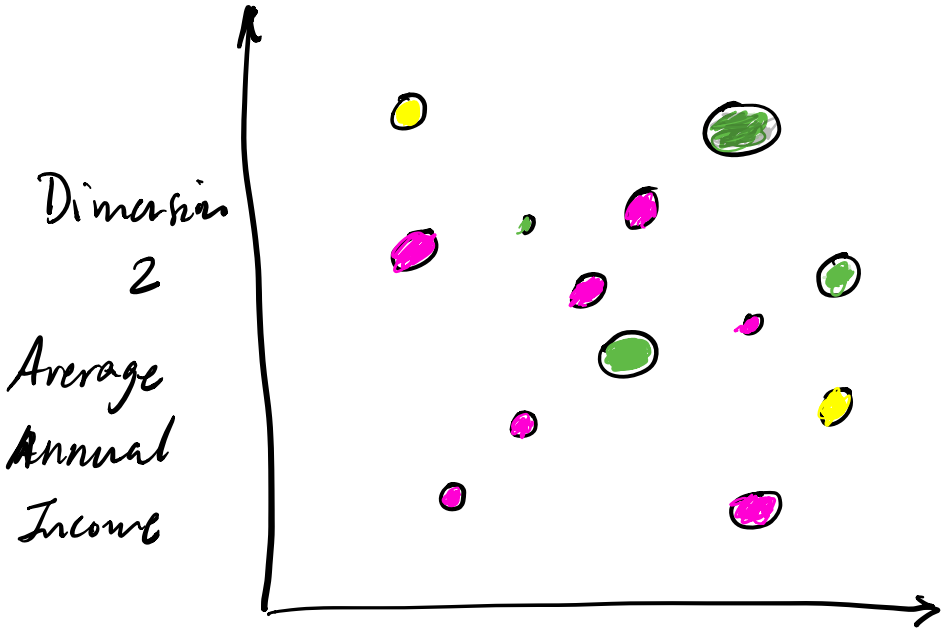
4 Dimensions



Dimension 3 - Bubble Size ○ ○ ○

Dimension 4 - Bubble Color

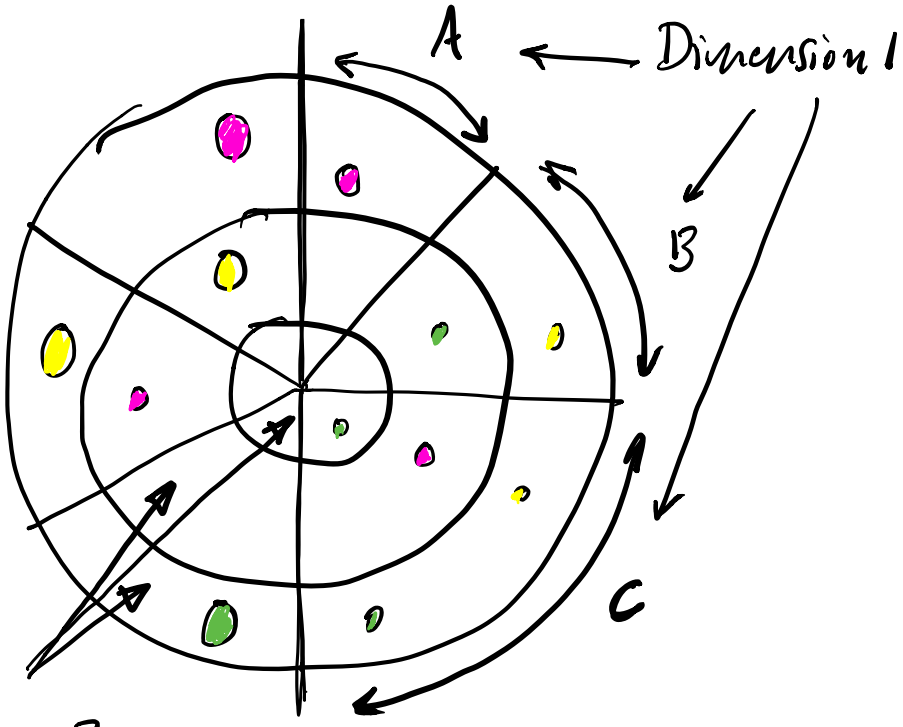
Bubble Chart - 4 Dimensions



Dimension 3 - Bubble Size

Dimension 4 - Bubble Color

Bullseye Chart - 4 Dimensions



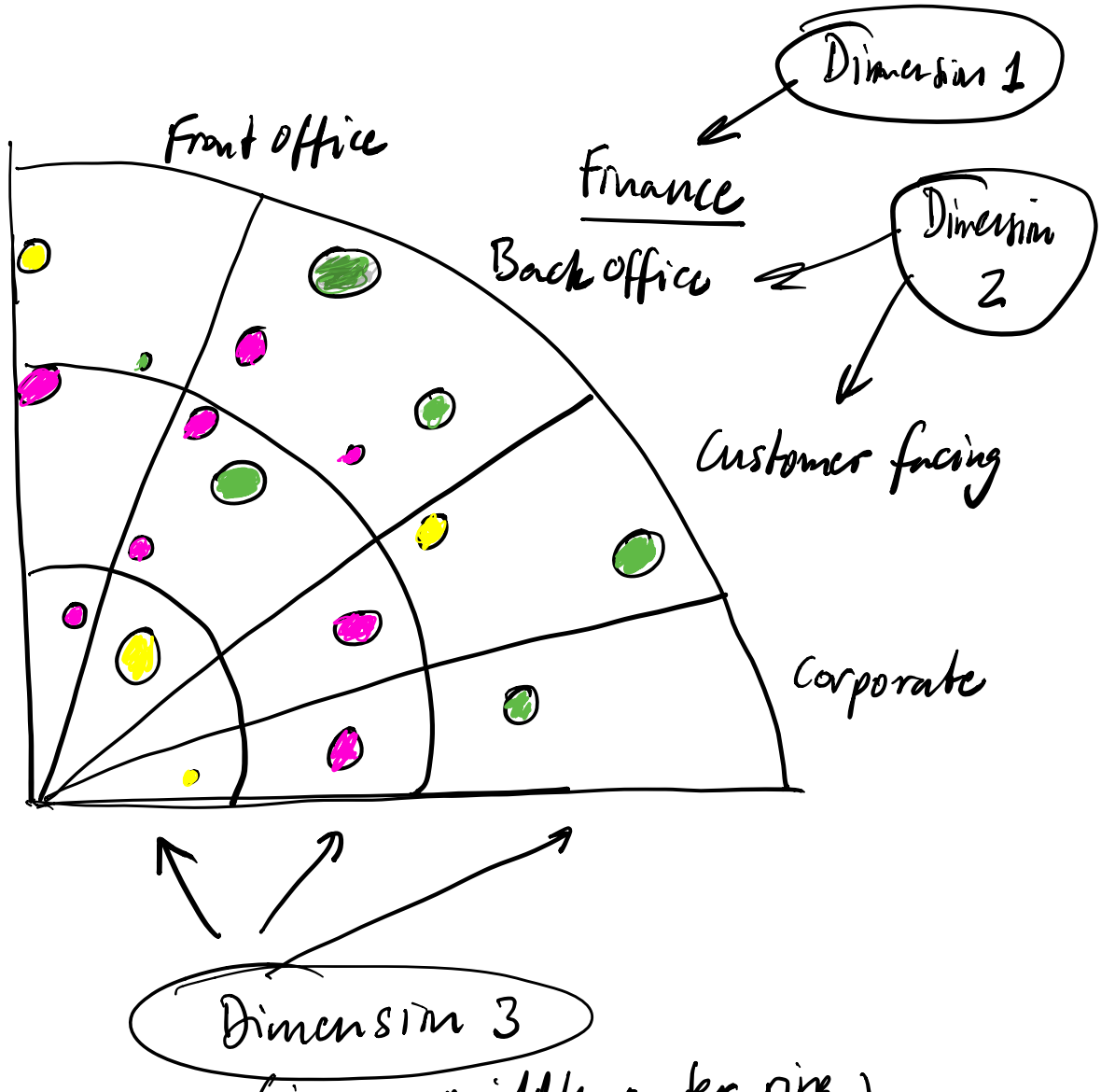
Dimension 2

(Outer, middle, inner ring)

Dimension 3 - Bubble Size o O O

Dimension 4 - Bubble Color

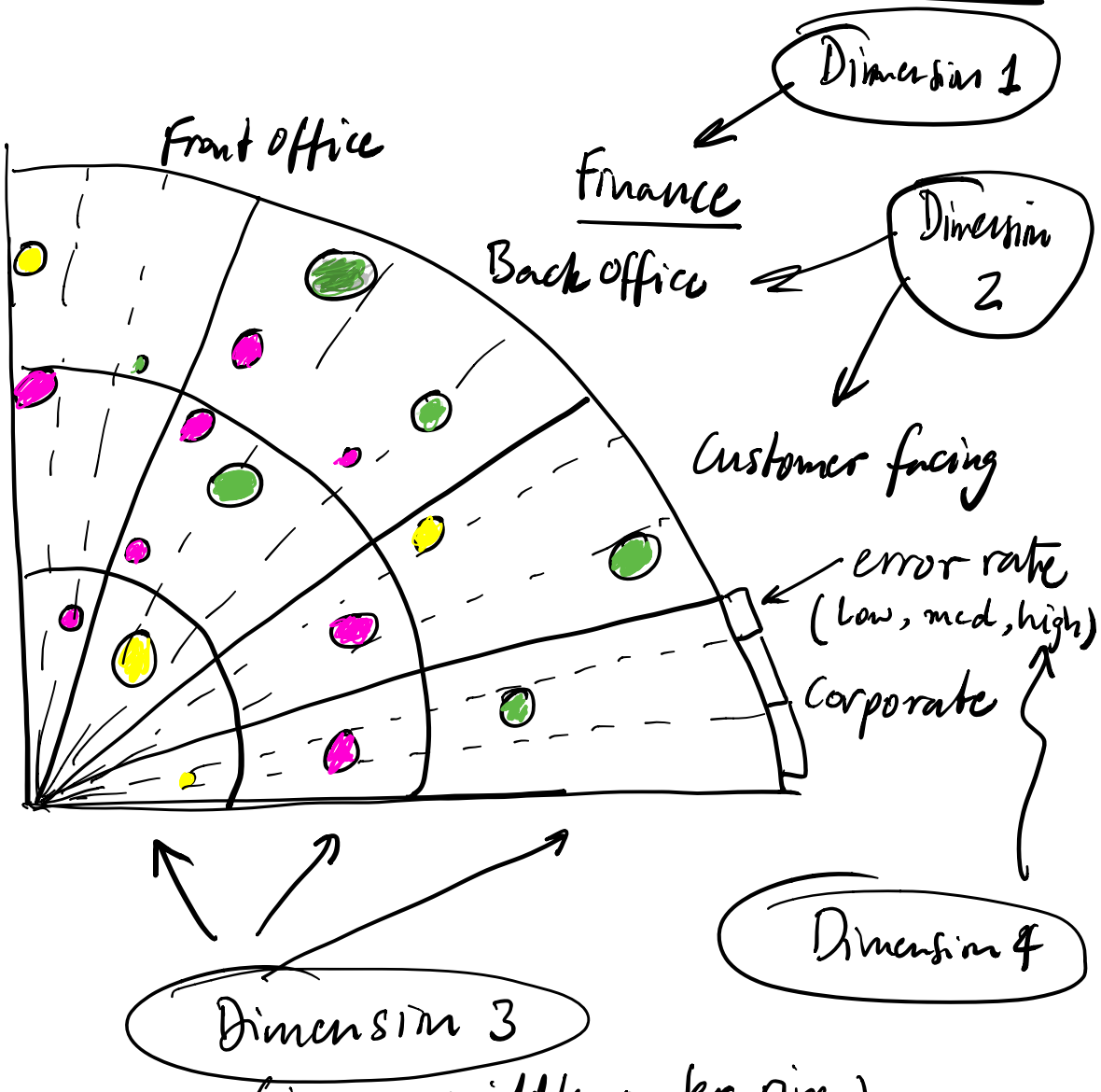
Bullseye Chart - 5 Dimensions



Dimension 4 - Bubble size

Dimension 5 - Bubble color

Bullseye Chart – 6 Dimensions



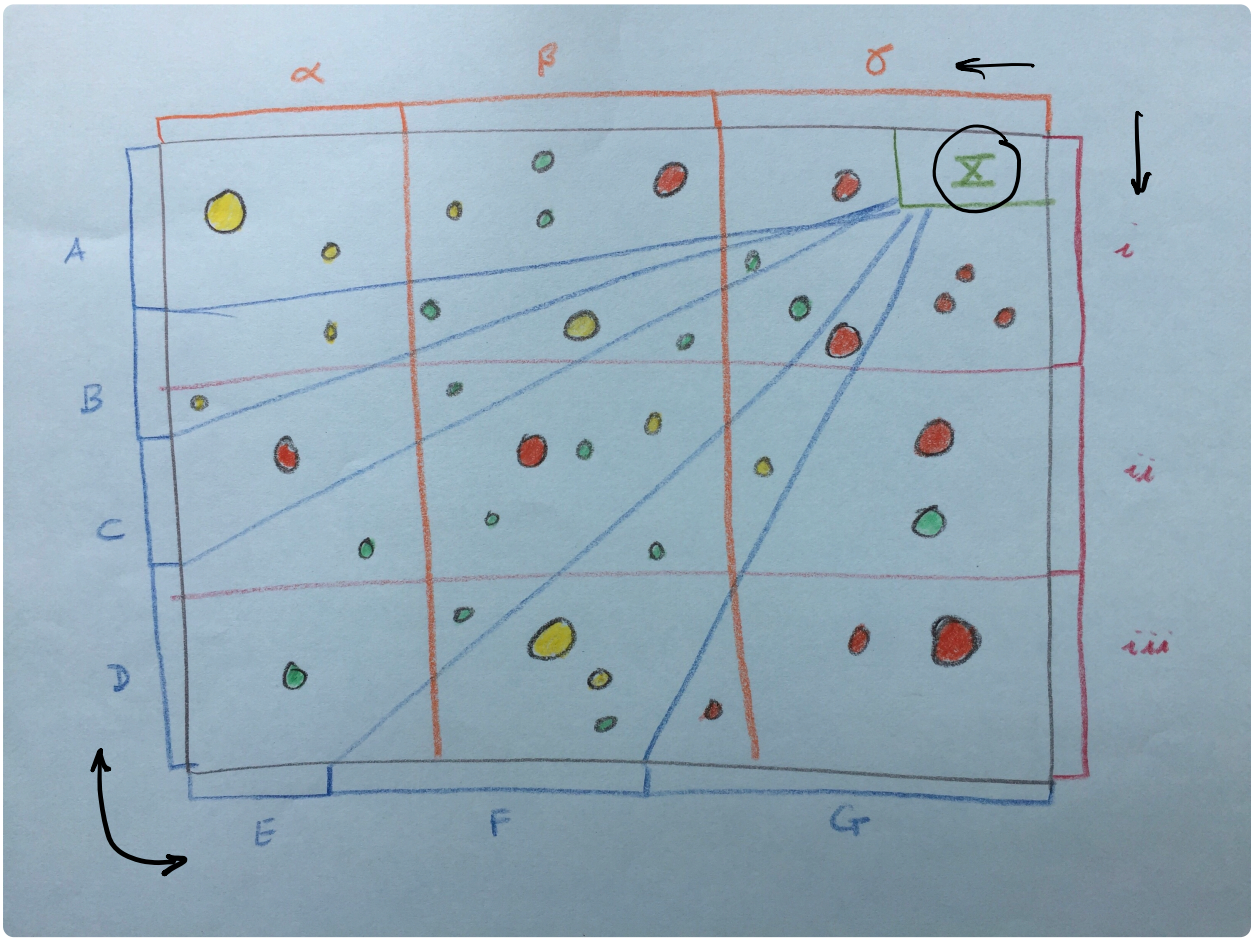
Dimension 5 – Bubble Size ○ ○ ○

Dimension 6 – Bubble Color

visualizing data (23)

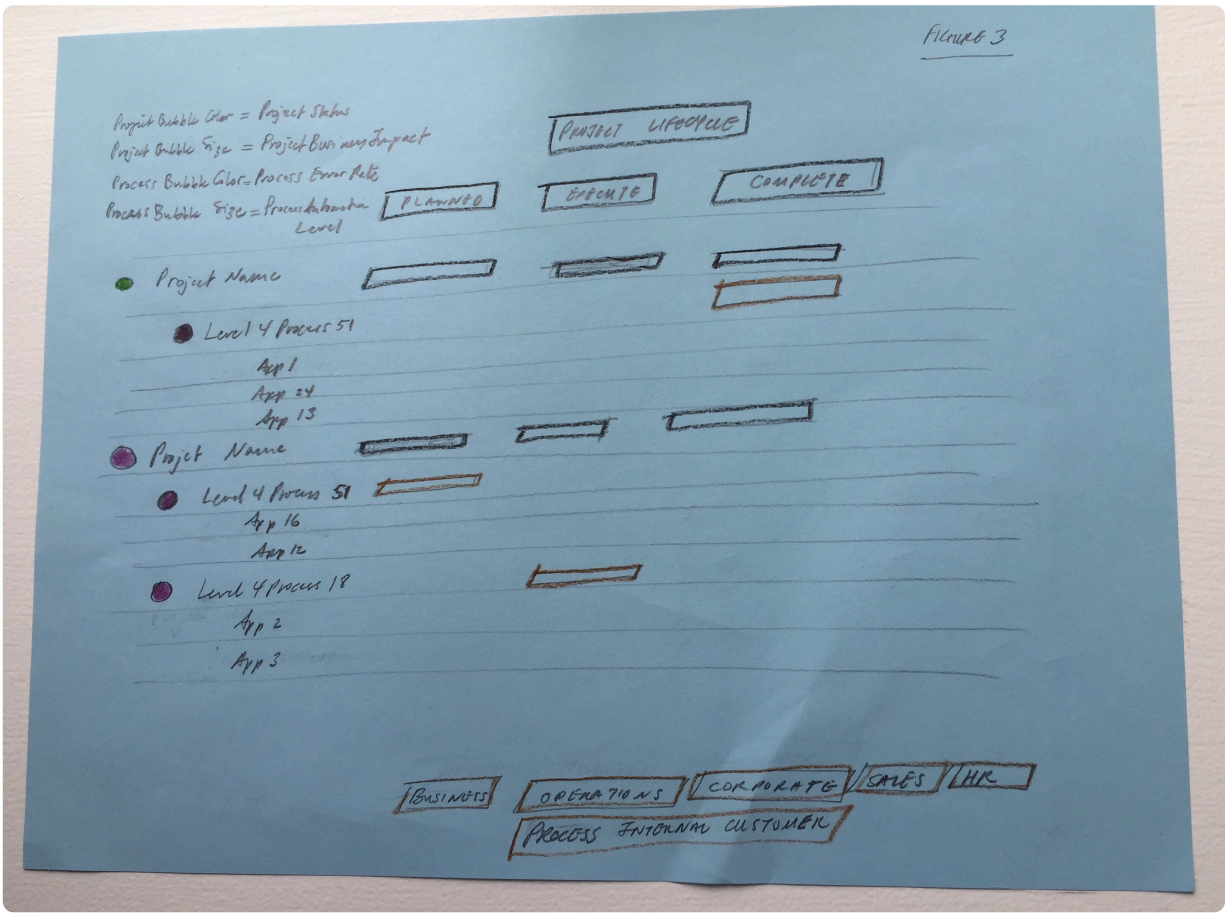
Roadmap Diagrams

6 + Dimensions



visualizing data (24)

Gantt Chart – 7 Dimensions



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