

THINKFUL

Dimensionality Reduction for Categorical & Mixed Data

Course

Objective

Students can explain dimensionality reduction techniques for categorical (MCA) and mixed (FAMD) data

Students can apply MCA and FAMD to appropriate datasets

Curriculum

Module:

- Checkpoint 1 Title [linked]
- Checkpoint 2 Title [linked]

Agenda

- ◆ Warm Up
- ◆ Categorical Variable Encoding
- ◆ Correspondence Analysis
- ◆ Multiple Correspondence Analysis
- ◆ MCA with Python
- ◆ Visualizing MCA Results
- ◆ Factor Analysis of Mixed Data
- ◆ FAMD with Python
- ◆ Visualizing FAMD Results

Warm Up

We learned previously that we can use PCA to reduce dimensionality when we have a feature set with a large number of continuous variables.

How can we reduce dimensionality when we have a large number of *categorical* variables?

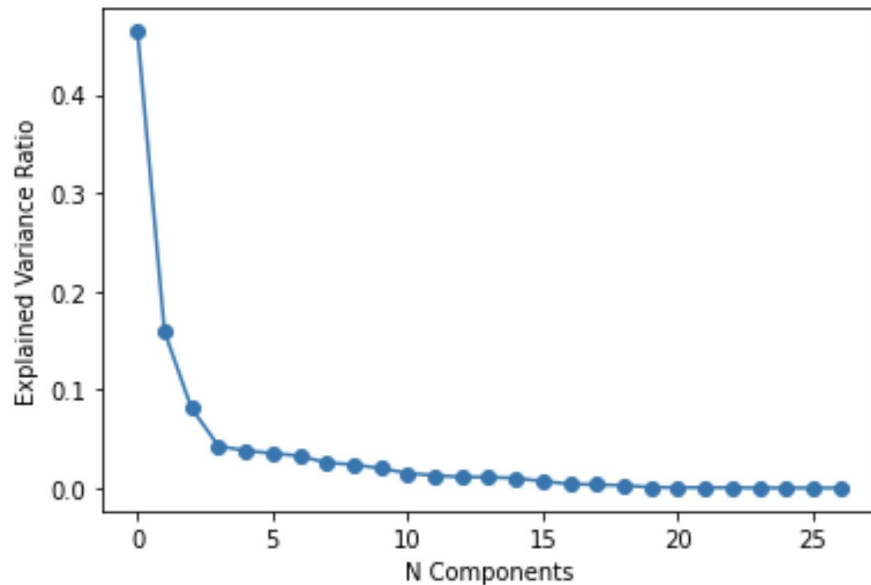
What if our data has a mix of continuous *and* categorical variables?

Warm Up

- What are the goals of PCA?
- In practice, how can we decide the 'right' number of components?

Warm Up

- What is the main goal of PCA?
- In practice, how can we decide the 'right' number of components?
- Based on the plot to the right, what would a reasonable number of principal components be?



Categorical Variable Encoding

- ◆ Categorical variables need to be encoded in order to be interpreted by machine learning algorithms.
- ◆ Common methods for encoding categorical variables include:
 - ◆ One Hot Encoding
 - ◆ Label Encoding
 - ◆ Binary Encoding
 - ◆ Learned Encoding (Distributed Representation)
- ◆ Encoding categorical variables further increases the dimensionality of the data.

Multiple Correspondence Analysis (MCA)

- ◆ MCA is often viewed as the categorical version of PCA.
- ◆ MCA allows us to reduce the dimensionality of categorical variables.
- ◆ Should be used when we have a large number of categorical variables and/or categorical variables with large numbers of values (categories).
- ◆ To understand how MCA works, let's take a step back and discuss simple correspondence analysis.

Correspondence Analysis

- ◆ Correspondence analysis uses contingency tables to analyze relationships or associations between pairs of categorical variables.
- ◆ Contingency tables are crosstabs where the rows represent the values of one categorical variable and the columns represent the values of another categorical variable.
- ◆ Inside the contingency table are counts of how many records fall into each combination of categories.

Contingency Table

- ◆ Below we have a contingency table comparing the frequencies found in the data across two categorical variables - Marital Status and Race.

	Race						
Marital Status	American Indian or Alaska Native	Asian	Black or African American	Hispanic	Two or more races	White	Total
Divorced	2	3	5	0	2	18	30
Married	1	15	21	2	4	80	123
Separated	1	1	4	0	0	6	12
Single	0	13	26	2	12	84	137
Widowed	0	2	1	0	0	5	8
Total	4	34	57	4	18	193	310

- ◆ We can see that there is a high concentration of records that fall into the Married and White variable combination as well as the Single and White combination, signifying a potential relationship between these variables.

Correspondence Analysis

- ◆ The values in the contingency table are normalized by the sum of all values.
- ◆ Pearson's chi-squared test is used to determine whether there is a statistically significant difference between expected frequencies and observed frequencies in one or more categories.
- ◆ When we conclude that an association exists between categorical variables, correspondence analysis provides us with an intuitive display of this association.

Multiple Correspondence Analysis (MCA)

- ◆ MCA can be viewed as an extension of simple correspondence analysis.
- ◆ The idea is simply to compute the one-hot encoded version of a dataset and apply correspondence analysis to it.
- ◆ The result is a reduction of many variables usually into 2-3 dimensions.

MCA with Python

- ◆ We can use the `prince` library to perform dimensionality reduction with MCA. (Note: `prince` will one hot encode your data for you)
- ◆ Similar to `sklearn`, `prince.MCA()` has a `fit` method and a `transform` method to retrieve the coordinates for the resulting components.

```
import prince

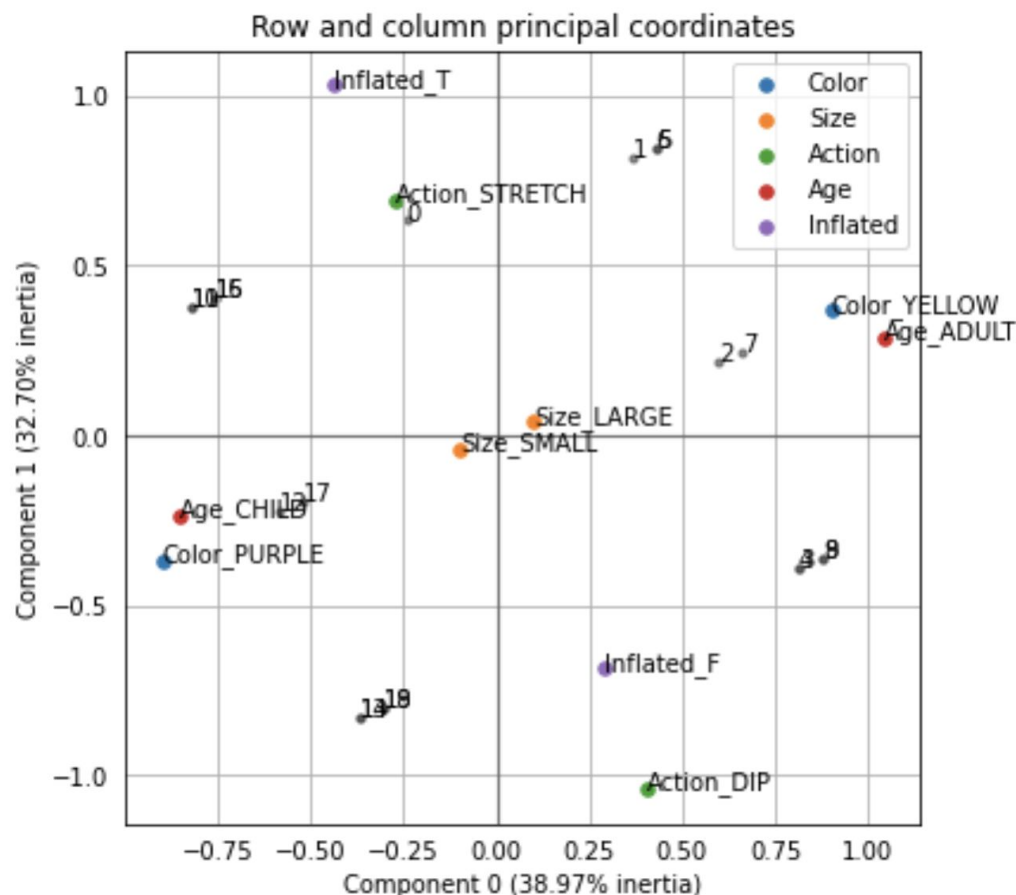
model = prince.MCA()
mca = model.fit(cat_data)
coordinates = mca.transform(cat_data)
```

Visualizing MCA Results

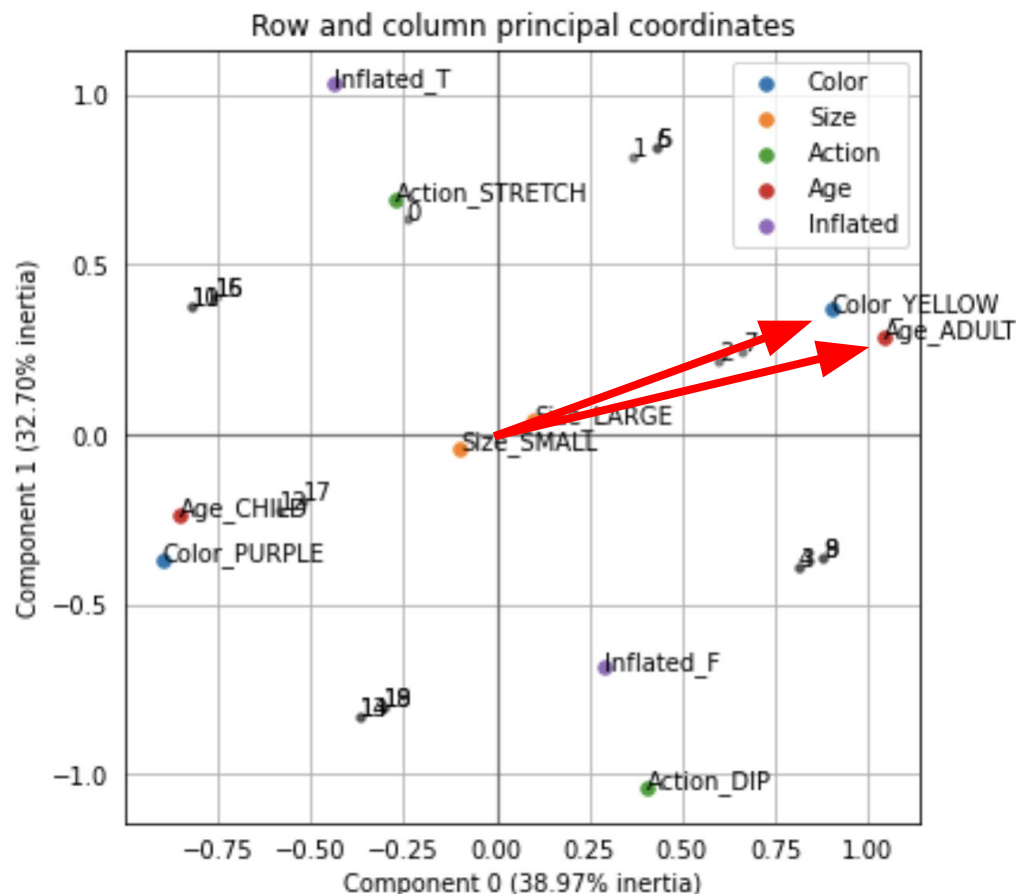
- ◆ We can also visualize the results by calling the `plot_coordinates` method as follows.

```
ax = mca.plot_coordinates(  
    X=cat_data, ← only required parameter  
    ax=None,  
    figsize=(15, 8),  
    show_row_points=True,  
    row_points_size=10,  
    show_row_labels=False,  
    show_column_points=True,  
    column_points_size=30,  
    show_column_labels=False,  
    legend_n_cols=1  
)
```

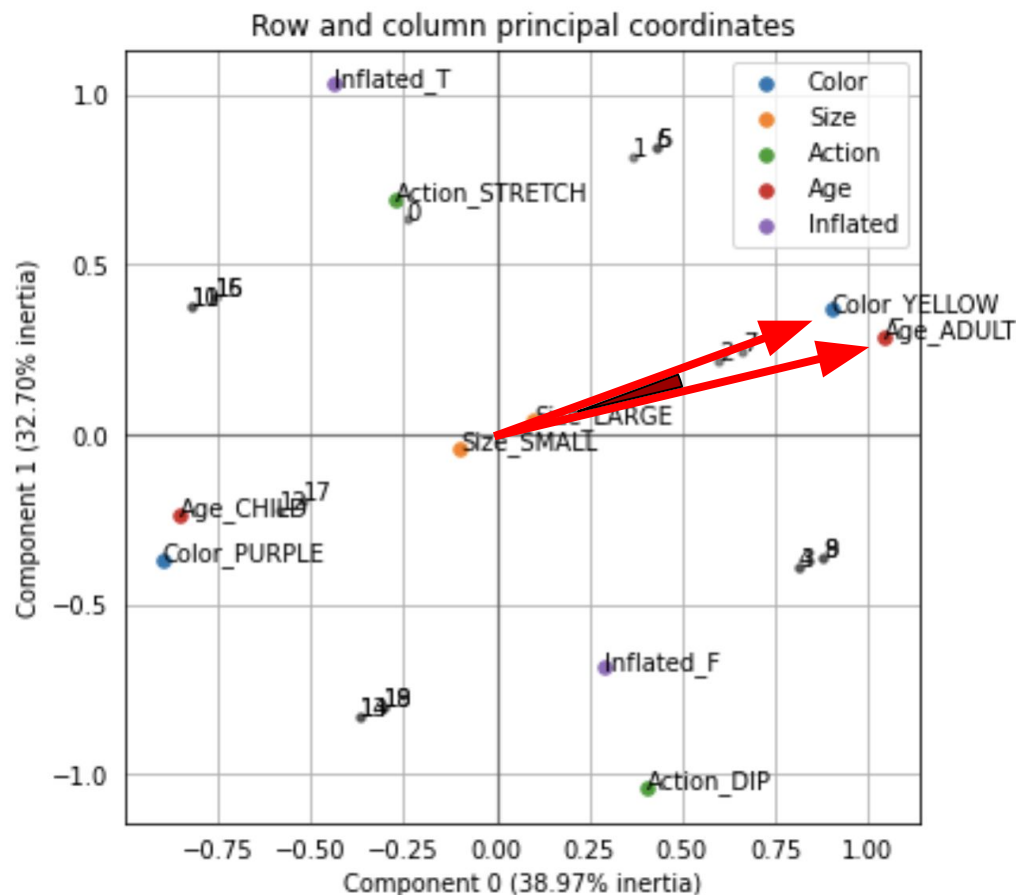
- For interpretation, draw lines from the origin to the 2 points of interest.



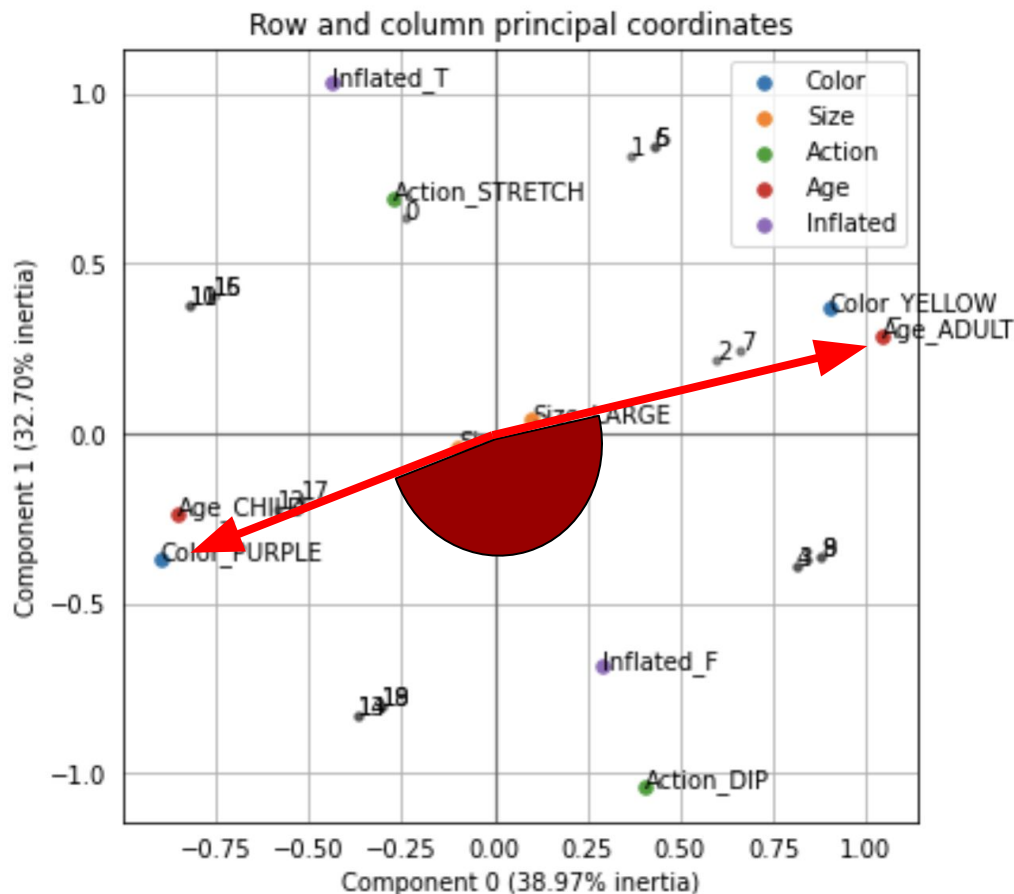
- For interpretation, draw lines from the origin to the 2 points of interest.



- For interpretation, draw lines from the origin to the 2 points of interest.
- If the angle between these lines is acute, these categories occur together often.



- For interpretation, draw lines from the origin to the 2 points of interest.
- If the angle between these lines is acute, the points are closely related.
- If the angle between these lines is obtuse, these categories occur together less often.



Multiple Correspondence Analysis (MCA)

- ◆ The main drawback of MCA is that it only works on categorical data.
- ◆ You can include numeric variables by binning or discretizing them, but only if there are relatively few of them.
- ◆ If the data contains many numeric variables in addition to categorical ones, it would be preferable to use a different approach.
- ◆ Additionally, MCA results can be unstable when there are relatively few rows in the data (ex. < 100).

QUESTION

What if our data contains a mix of many numeric and categorical variables?

Factor Analysis of Mixed Data (FAMD)

- ◆ When we have both numeric and categorical variables for which we want to reduce dimensions, we can use Factor Analysis of Mixed Data (FAMD).
- ◆ “Roughly speaking, we can say that FAMD works as a PCA for quantitative variables and an MCA for qualitative variables.”
- ◆ FAMD works best when there are enough categorical variables that it makes sense to discretize/bin numeric variables.

FAMD with Python

- ◆ We can use the `prince` library to perform dimensionality reduction with FAMD. (Note: `prince` will both one hot encode and scale your data for you)
- ◆ We can use `prince.FAMD()` very similarly to `prince.MCA()`

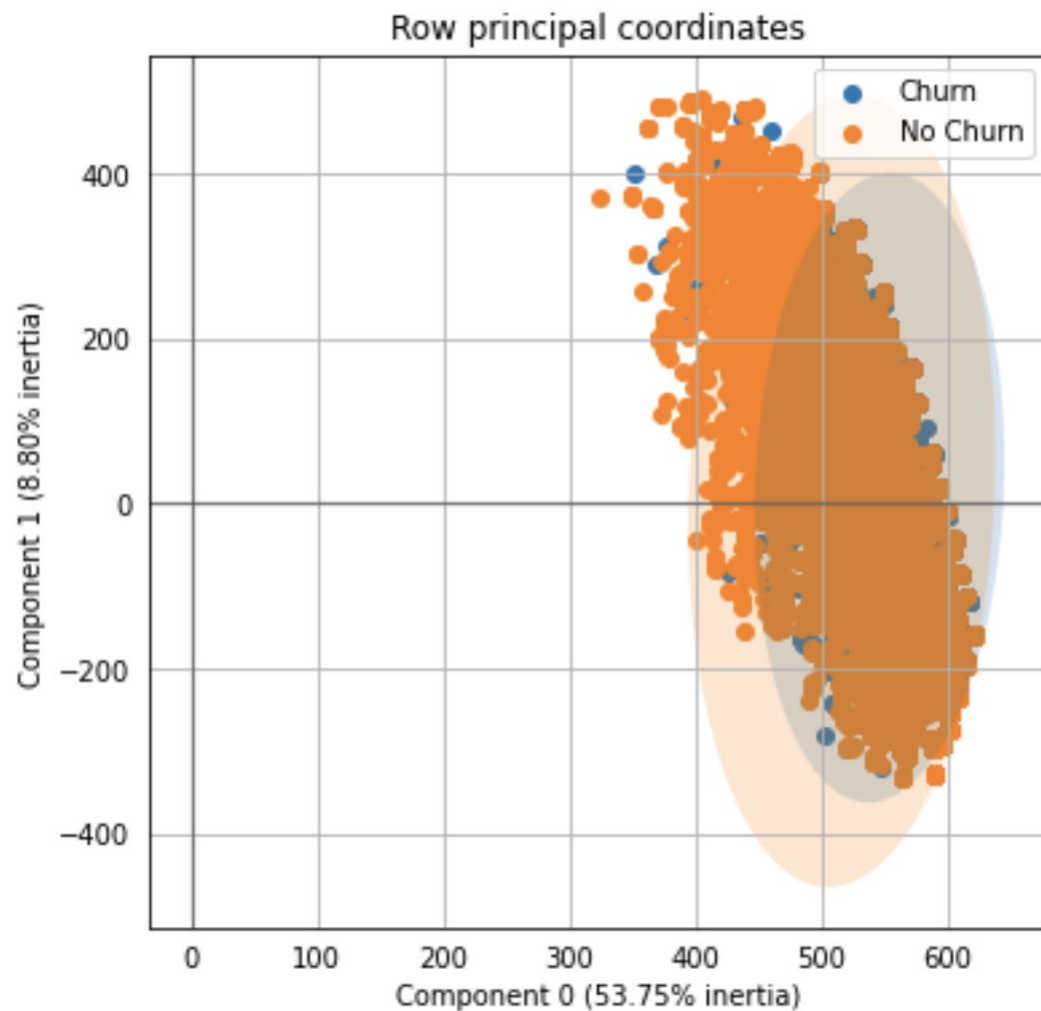
```
1 import prince
2
3 model = prince.FAMD()
4 famd = model.fit(data)
5 coordinates = famd.transform(data)
```

Visualizing FAMD Results

- ◆ We can plot the coordinate values for each component the FAMD model generated by calling the `plot_row_coordinates` method.
- ◆ We can also color code our plot points by passing one of the categorical columns to the `color_labels` argument.

```
1 famd.plot_row_coordinates(data, color_labels=y)  
2 plt.show()
```

THINKFUL



Factor Analysis of Mixed Data (FAMD)

- ◆ FAMD can be very helpful when we have mixed data as we can explore the association between all variables, both numeric and categorical.
- ◆ Like MCA however, FAMD effectiveness can be limited when there are relatively few rows in the data as well as when there are few categorical variables compared to numeric ones.

Summary

In this session, we covered:

- ◆ How encoding categorical variables for machine learning can add dimensionality to our data.
- ◆ What correspondence analysis is and how we can use multiple correspondence analysis (MCA) to reduce the dimensions of categorical variables.
- ◆ What factor analysis of mixed data (FAMD) is and how we can use it for dimensionality reduction on data containing both numeric and categorical data.
- ◆ How to visualize results from MCA and FAMD with Python.

Assignment

Use MCA and FAMD on the HR dataset

[Dataset](#)

[Notebook](#)

THANKFUL

Thank You

THINKFUL

PRESENTATION TITLE

UP NEXT

Subtopic Title



Bulleted List

- ◆ Example 1
 - ◇ Example 2
- ◆ Example 3

Content

When you are beginning a machine learning process, you need to determine if you have labeled data, and if so, whether the labels are categorical or continuous. This will decide what types of modeling algorithms you can use.

Remember that you can always change your label from categorical to continuous using the techniques from feature engineering. It's all about your use case.

QUESTION

What are some considerations
in optimizing k ?

Keyword	/ˈkēˌwərd/)))
---------	------------	-----



Linear SVC is a support vector machine classifier that will use a linear hyperplane to separate between the classes

Later we may look at other types of hyperplanes that can be used to separate between the classes such that we have a maximum distance between the hyperplane and all observations



Our greatest weakness lies in giving up. The most certain way to succeed is always to try just one more time.

Amanda Holden


Diagram

Replace me with a diagram or image
Resize as needed

Caption


resize or move, don't change font size

Double Diagram



Replace me with a diagram or
image
Resize as needed

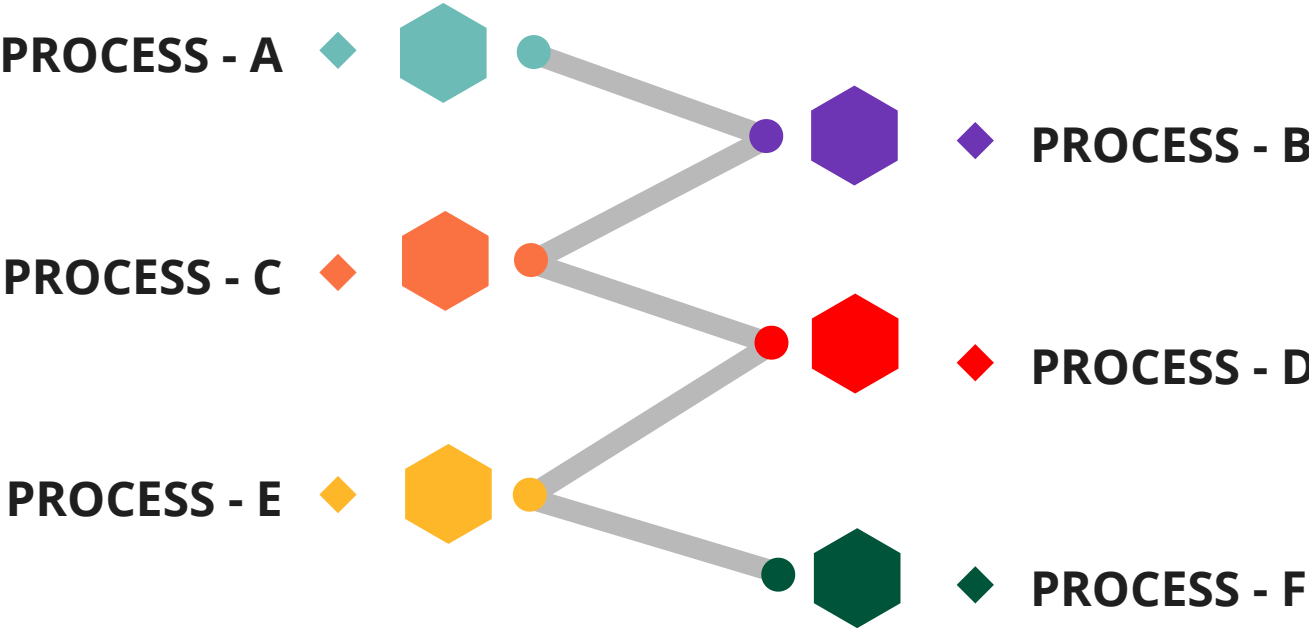
Caption, resize or move,
don't change font size



Replace me with a diagram or
image
Resize as needed

Caption, resize or move,
don't change font size

Diagram



EXERCISE

1. Change the outcome variable in your diabetes data set back from binary into the original 3 classes
2. Run your logistic regression model on it and evaluate the scoring metrics
3. Be careful to mind the accuracy for the least-dominant class

Summary

Brief review, should call back to the objective and make the direct connection for how the objective has now been achieved.

Assignment

1. Instructions...
2. ...
3. ...
4. ...

THANKFUL

Thank You

PRESENTATION GUIDELINES ON COPY

Title

Michroma 27pt

Subtitles

Michroma 18PT

Body copy

Open Sans 18pt

Diagrams/Captions

Open Sans 18pt

COLOR PALETTE

#ffffff

R: 255
G: 255:
B: 255

#000000

R: 0
G: 0
B: 0

#feb729

R: 254
G: 183
B: 41

#fa7242

R: 250
G: 114
B: 66

#00badb

R: 0
G: 186
B: 219

#6d35b3

R: 109
G: 53
B: 179

#00543a

R: 0
G: 84
B: 58

#6cbbb7

R: 108
G: 187
B: 183

#ff0000

R: 255
G: 0
B: 0

#0074bf

R: 0
G: 116
B: 191