INTRODUCTION - STATS Project :

*Joos.K, Jonas R., Ilhem B., Ole B., Assan S.,Trung H.*

Data science cycle control cycle

1. ASK : What the problem(s) we need to solve ?
2. RESEARCH : What data do we need and how do we get it ?
3. MODEL : Which method(s) is appropriate to use ?
4. VALIDATE : Do the model and assumpstions work as expected ?
5. TEST : How does the model generalize to real world data ?
6. INTEPRET : How can we use the conclusion in the real world ?

As the data is already given, what we can do is following :

* 1. Descriptive analysis
     1. Data preparation (wrangling data to right format)

(na, missing values, outliers, data unit, normalization …)

* + 1. Visualization
  1. Inference and (or) prediction statistics
     1. Knowledge discovery inside data
        1. Contribution of each variables toward budget ?
        2. Is there any pattern of buget allocation among the variables ?
     2. Prediction : ?
  2. Modeling : Parameters or non-parameter ?

(Can be non-parameter as we make no assumption about the functional form of predicted variables – in fact, we don’t even know the response variables.)

Supervised or Unsupervised ? As we have no idea about the ‘groups’ or ‘classes’ of variables as in case of classification, we might opt to choose unsupervised method :

* + 1. PCA
    2. Clustering
  1. Validation
     1. Collinearity (Correlation of variables (e.g. over time))

1. **Descriptive analysis :**
   1. Overview by individuals :
      1. Stacked lines
      2. Pies
   2. Overview by variable :
      1. **AN: year**
         1. Table:
            1. Min-Max
            2. Standard deviation
         2. Box-Plot - Histogram
         3. Evolution with time
      2. **PVP: authorities** 
         1. Table:
            1. Min-Max
            2. Standard deviation
         2. Box-Plot - Histogram
         3. Evolution with time
      3. **AGR: agriculture** 
         1. Table:
            1. Min-Max
            2. Standard deviation
         2. Box-Plot - Histogram
         3. Evolution with time
      4. **CMI: trades and companies** 
         1. Table:
            1. Min-Max
            2. Standard deviation
         2. Box-Plot - Histogram
         3. Evolution with time
      5. **TRA: work** 
         1. Table:
            1. Min-Max
            2. Standard deviation
         2. Box-Plot - Histogram
         3. Evolution with time
      6. **LOG: accommodations** 
         1. Table:
            1. Min-Max
            2. Standard deviation
         2. Box-Plot - Histogram
         3. Evolution with time
      7. **EDU: education** 
         1. Table:
            1. Min-Max
            2. Standard deviation
         2. Box-Plot - Histogram
         3. Evolution with time
      8. **ACS: social action** 
         1. Table:
            1. Min-Max
            2. Standard deviation
         2. Box-Plot - Histogram
         3. Evolution with time
      9. **ANC: veterans** 
         1. Table:
            1. Min-Max
            2. Standard deviation
         2. Box-Plot - Histogram
         3. Evolution with time
      10. **DEF: defense** 
          1. Table:
             1. Min-Max
             2. Standard deviation
          2. Box-Plot - Histogram
          3. Evolution with time
      11. **DET: debt refund** 
          1. Table:
             1. Min-Max
             2. Standard deviation
          2. Box-Plot - Histogram
          3. Evolution with time
      12. **DIV: various**
          1. Table:
             1. Min-Max
             2. Standard deviation
          2. Box-Plot - Histogram
          3. Evolution with time
2. **Assumptions :**
   1. Are the individuals (years) independent ?
      1. Correlation Matrix between years
      2. heatMap
   2. Are the variables correlated ?
      1. Correlation Matrix between years
      2. heatMap
3. **PCA Analysis :**
   1. Ascending analysis
   2. PCA groups
   3. Cos2 to check whether data is well represented
   4. Plot (Zi,Zj)
   5. Variables against Zi,Zj
4. **Clustering :**
   1. Kmeans
   2. Fuzzy Kmeans
5. **Issues :**
   1. Data is time dependent
   2. How to take into account the time series aspect