SAS project - 3rd year BFA

Jerome Lepagnol[*] Pierre Lepagnol[†]

Abstract

This year is a presidential election year in France. Behind the estimates of the major polling institutes, there are modelling choices and data choices. Can we do better ? Can we beat their predictions ? Will we be able to discover, even before the first round, who will be the next president of France?

Preliminaries

Deadline for submission: ***9 April 2022 - EOB (19h30-GMT+1 - MAX)***

Expected:

- SAS codes with comments
- Input file (if modified)
- Output file (if any)
- Professional report document in PDF format (max 10 useful pages)

  If you **imperatively** need to modify the input data files (in XLS for example) :
  Please provide these modified files with a note explaining your changes to these files.
  All option consisting in scripting the modification of data is prefered

  During all your project you should check for outliers, remove row if needed (In case of `NULL/UNDETERMINED` values, etc).
  Please document those removals and include them in SAS script

  Please Have a look at `meta_xxx.csv` files, thoses are meta-data documentation, written in french. They may help you to interpret variables present in your charts, models and SAS tables. No need to load those tables in SAS.

We prepared many datasets. All the files can be found under the public git repository here : https://github.com/LepagnolCorp/SAS2022-PROJECT

[*] jerome.lepagnol@dauphine.fr
[†] pierre@lepagnol.net

PART 1 – Import and qualificaiton of the data

- List of towns in France : `commune2021.csv`
- Results of 2017 presidential elections by town : `election_2017.csv`
- Political classification of 2017's candidates : `candidates2017.txt`
- Demographics data by town for years `2015`, `2017` and `2020` : `birth20xx.csv` and `death20xx.csv` (where xx is either 15/17/20)
- Wealth and Revenues data by town for years `2014` and `2019` : `revenues_declared_2014.csv` and `revenues_declared_2019.csv`
- Equipment and jobs by town : `equipement_codes.csv`, `selected_equip.txt`

1. Import in SAS all data mentioned above :

   a. Import `election_2017.csv`

   b. For the years `2015`,`2017` and `2020` :

   - Import `birth20xx.csv` and `death20xx.csv` + sanitize those datasets (check for outliers, remove row if needed, etc)

   c. For the years `2014` and `2019` :

   - Import `revenues_declared_20xx.csv` (you should probably skip[1] rows or reformat csv file to get a proper import)

2. With as many step `Data` steps as necessary :

   a. Merge demographics variables in a single table for all years. The final table will be keyed by town.

   b. Merge revenue variables in a single table for all years. The final table will be keyed by town.

   c. Group by departments and produce agregated tables for both revenues and demographics at this level.
   The final table will be keyed by departments .

   d. At the departments level and for each year :
   (**Pay attention to scales, orders of magnitude: it will be necessary to normalise the data according to the population density.**)

   - Compute `count` of different equipments
   - Compute the increase rate of equipments between the two years, by equipment types.
   - Compute the correlation between of each equipment type increase rates (the observation being the departments).

   $$\forall i, j \in \text{list of equipments}, CORR(\text{Dep\_EquipTypeIncreaseRate}_i, \text{Dep\_EquipTypeIncreaseRate}_j)$$

   - Conclude in 5 lines maximum.

   e. Merge all data from revenues, demographics and equipments with the table of electoral results of 2017. The final table will be keyed by town. Pay attention to the years you include in your table.

   f. Given the classification of candidates in `candidates2017.txt` : Group candidates and compute the `score` of each political categories at town level.
   Whats is the `score` ? It is the number of votes casted in elections. (El famoso `voix exprimées` in french), to be more accurate in our analysis, let consider abstention as real candidate. Therefore to compute proportions we must look at the number of registered elector. The computed score will be based on the `SUM` of the number of voices of each candidates, by political categories.

[1] Here is a way to do the skip : see Scenario 2: Variable names and data begin "later" than row : https://blogs.sas.com/content/sgf/2017/10/20/tips-for-using-the-import-procedure-to-read-files-that-contain-delimiters/

PART 2 – Statistical Analysis

3. Describe with **appropriate statistical procedures** data created in previous steps, notably answering the following questions :

   a. For each of the 6 political categories (from FAR-LEFT to FAR-RIGHT + ABSTENTION) in which departments the 3 top best score had been performed.

   Example :

   - FAR-RIGHT : Dep_1 with `Number of voices in Dep_1`, Dep_2 with `Number of voices in Dep_2`, etc.
   - FAR-LEFT : Dep_xx with `Number of voices in Dep_xx`, etc.
   - CENTER : Dep_xx with `Number of voices in Dep_xx`, etc.
   - etc.

   With observation at the department level :

   b. How the score of Macron is positionned and spread in terms of standard deviation, kurtosis, skewness. You can display a scatter plot or any appropriate graph to illustrate this data.

   With observation at the town level :

   c. Same question and explain wich level you would choose and why.

   Pick a political category :

   With observation at the department level :

   d. How the score of your category is positionned and spread in terms of standard deviation, kurtosis, skewness. You can display a scatter plot or any appropriate graph to illustrate this data.

   With observation at the town level :

   e. Same question and explain wich level you would choose and why.
   Is there a difference in granularity between Macron and your political category ? If there is a difference, explain why ? 5 lines max.

   With observation at the departement level :

   f. Look in deeper details to the number of births and deaths :

   - Is there any variable that is normally distributed? Perform the necessary test.
   - Regarding the evolution of those variables between 2015-2017, is the evolution rate normally distributed ?
   - Can something similar be done regarding the 2020 data ?
     (Evolution 2017-2020, please take into account the duration of the evolution 2015-2017 : 2 years // 2017-2020 : 3 years)

   g. Compute the correlation between revenues 2014 (resp. 2019) and equipments 2015 (resp. 2020) ?[2]
   Analyze the given correlations and illustrate them with scatter plots, boxplots, histogramms or any appropriate graphs.

   h. Is there a correlation between demographics data 2015 (resp 2020) and equipments of the same year ? As you made it before illustrate your analyze with any appropriate graphs.

[2] i.e revenue is lagged by one year to equipment

PART 3 – Development of linear models

4. Produce correlation matrices between all the variables of demographics, revenues and scores of the political categories at town level.
5. Then, for the candidate Emmanuel Macron (the only one in political category "center") :

a. Produce a model predicting the score of the candidat within a departement depending on the various caracteristics of the department :
   - Revenues 2014 (i.e 3 years lag to the 2017 election)
   - Demographics 2015 (i.e 2 years lag to the 2017 election)
   - Equipment 2015 (i.e 2 years lag to the 2017 election)

b. BONUS : Produce a model predicting whether the candidat will be majority candidate within a choosen department based on the department caracteristics. (binary discrimination[3])

6. Pick a political category (other than "center")

a. Produce a model predicting the score of the category within a departement depending on the various caracteristics of the department
   - Revenues 2014 (i.e 3 years lag to the 2017 election)
   - Demographics 2015 (i.e 2 years lag to the 2017 election)
   - Equipment 2015 (i.e 2 years lag to the 2017 election)

b. BONUS : Produce a model predicting whether the candidat will be majority candidate within a choosen department based on the department caracteristics. (binary discrimination[4])

7. Elaborate on the validities of model 5.a and 6.a assumptions (normality of residuals).

Check the site **https://www.theanalysisfactor.com/assumptions-of-linear-models/** are there any other assumptions you want to check ? What is the outcome of this checking ?

8. Elaborate on the model 5.a and 6.a performances (measure of the quality of the prediction : R square + RMSE).

[3] Please look into the following exemple : https://communities.sas.com/t5/SAS-Communities-Library/A-Guide-to-Logistic-Regression-in-SAS/ta-p/564323

[4] Please look into the following exemple : https://communities.sas.com/t5/SAS-Communities-Library/A-Guide-to-Logistic-Regression-in-SAS/ta-p/564323

PART 4 - Determination of the influence of categorial variables on models

7. Quantitative variables discretization :

   a. You will make the 2014 revenue variable discrete in 5 modalities (4 tranches of values having the same *number* of observations and one with the missing values, if any). The new variable is called `REV2014_disc`.

   b. You will perform a complete analysis on a GLM/ANCOVA linear model explaining the Macron score by the newly created variable in 7.a and by region. (Region variable is located in `REG` form `communes2021.csv`)
   Please provide any relevant details on the obtained resultats and discuss the limits of this modelisation.

   c. Elaborate on the model performance

PARTIE 5 – Conclusion

8. For the center (E. Macron) and the alternate choosen category (in PART3/Q.6):
   - Please compute prediction of the 2022 scores of those 2 categories based on the better model (as of 2017 model performances).
   - We want a national score at the end : make it happen.

   Please note that the result expected is a % of the votes for a category. Therefore, you will need to aggregate the departmental score a single one. This aggregation shall be weighted by the population of the departement.

Bonus Calculation

A bonus will be computed the following way :

- S_o is the score obtained of your category
- S_e is the score estimated by you model

$$BONUS = 8 \times \max\left(0, \frac{1}{2} - \frac{abs(S_o - S_e)}{(\frac{S_o+S_e}{2})}\right)$$

EG: if $S_e = 10\%$ and $S_o = 7.5\%$ will lead to a bonus of 1.7 points.

$$\begin{aligned} BONUS &= 8 \times \max\left(0, \frac{1}{2} - \frac{abs(S_o - S_e)}{(\frac{S_o+S_e}{2})}\right) & (1) \\ &= 8 \times \max\left(0, \frac{1}{2} - \frac{abs(7.5 - 10)}{(\frac{7.5+10}{2})}\right) & (2) \\ &= 8 \times \max\left(0, \frac{1}{2} - \frac{abs(-2.5)}{8,75}\right) & (3) \\ &= 8 \times \max(0, 0, 214285714) & (4) \\ &= 8 \times 0, 214285714 & (5) \\ &= 1, 714285712 & (6) \end{aligned}$$

Classification in 2022

Candidate Name	Policial Side
J-L. Mélenchon	FAR-LEFT
F. Roussel	FAR-LEFT
N. Arthaud	FAR-LEFT
A. Kazib	FAR-LEFT
P. Poutou	FAR-LEFT
A. Hidalgo	LEFT
C. Taubira	LEFT
Y. Jadot	LEFT
J-M. Governatori	LEFT
A. Waechter	LEFT
E. Macron	CENTER
V. Pécresse	RIGHT

Candidate Name	Policial Side
M. Le Pen	FAR-RIGHT
N. Dupont-Aignan	FAR-RIGHT
F. Philippot	FAR-RIGHT
A. Martinez	FAR-RIGHT
E. Zemmour	FAR-RIGHT
F. Asselineau	OTHER
G. Koenig	OTHER
J. Lassalle	OTHER
M. Cau	OTHER
C. Egger	OTHER
A. Langlois	OTHER
H. Thouy	OTHER
G. Kuzmanovic	OTHER

Appendix

USEFUL DATASETS :

Name	URL - webpage	URL - download data	Description
Presidential election results - 2017	https://www.data.gouv.fr/fr/datasets/election-presidentielle-des-23-avril-et-7-mai-2017-resultats-du-1er-tour-1/	https://www.data.gouv.fr/fr/datasets/r/06fde31f-a399-489a-8176-2d8a3e11d97c	Presidential election results 2017
deces2014-2020	https://www.insee.fr/fr/statistiques/1893253	https://www.insee.fr/fr/statistiques/fichier/1893253/base_deces_2020_csv.zip	
naissances2014-2020	https://www.insee.fr/fr/statistiques/1893255	https://www.insee.fr/fr/statistiques/fichier/1893255/base_naissances_2020_csv.zip	

Other sources (non used, yet)

- deces all https://www.insee.fr/fr/statistiques/1893253
    - Décès de 2014 à 2019 : https://www.insee.fr/fr/statistiques/fichier/1893253/base_deces_2019M_csv.zip
    - Décès de 2010 à 2019 : https://www.insee.fr/fr/statistiques/fichier/1893253/base-deces-2019_CSV.zip
    - Décès de 2009 à 2018 : https://www.insee.fr/fr/statistiques/fichier/1893253/base_deces_2018.zip
    - Décès de 2008 à 2017 : https://www.insee.fr/fr/statistiques/fichier/1893253/base_deces_2017.zip
- naissances all
    - Naissances de 2014 à 2020 : https://www.insee.fr/fr/statistiques/fichier/1893255/base_naissances_2020_csv.zip
    - Naissances de 2014 à 2019 : https://www.insee.fr/fr/statistiques/fichier/1893255/base_naissances_2019M_csv.zip
    - Naissances de 2010 à 2019 : https://www.insee.fr/fr/statistiques/fichier/1893255/base-naissances-2019_CSV.zip
    - Naissances de 2009 à 2018 : https://www.insee.fr/fr/statistiques/fichier/1893255/base_naissances_2018.zip
    - Naissances de 2008 à 2017 : https://www.insee.fr/fr/statistiques/fichier/1893255/base_naissances_2017.zip
- Description : https://www.data.gouv.fr/fr/datasets/r/6e235020-f141-4427-ad57-370efcab8f74 https://www.data.gouv.fr/fr/datasets/election-presidentielle-des-23-avril-et-7-mai-2017-resultats-du-1er-tour-1/ data : https://www.data.gouv.fr/fr/datasets/r/06fde31f-a399-489a-8176-2d8a3e11d97c
- Base de comp / com https://www.insee.fr/fr/statistiques/2521169 https://www.insee.fr/fr/statistiques/fichier/2521169/base_cc_comparateur_csv.zip
- tourisme https://www.insee.fr/fr/statistiques/2021703
    - Capacité des communes en hébergement touristique en 2020 data : https://www.insee.fr/fr/statistiques/fichier/2021703/base-cc-tourisme-2020.zip

  - Capacité des communes en hébergement touristique en 2019 (en géographie au 01/01/2019) data : https://www.insee.fr/fr/statistiques/fichier/2021703/base-cc-tourisme-2019-geo2019.zip
  - Capacité des communes en hébergement touristique en 2019 (en géographie au 01/01/2018) data : https://www.insee.fr/fr/statistiques/fichier/2021703/base-cc-tourisme-2019.zip
  - Capacité des communes en hébergement touristique en 2018 data : https://www.insee.fr/fr/statistiques/fichier/2021703/base-cc-tourisme-2018.zip
  - Capacité des communes en hébergement touristique en 2017 data : https://www.insee.fr/fr/statistiques/fichier/2021703/base-cc-tourisme-2017.zip
  - Capacité des communes en hébergement touristique en 2016 data : https://www.insee.fr/fr/statistiques/fichier/2021703/base-cc-tourisme-2016.zip
  - Capacité des communes en hébergement touristique en 2015 data : https://www.insee.fr/fr/statistiques/fichier/2021703/base-cc-tourisme-2015.zip
  - Capacité des communes en hébergement touristique en 2014 data : https://www.insee.fr/fr/statistiques/fichier/2021703/base-cc-tourisme-2014.zip

- rev 2020 + sas https://www.data.gouv.fr/fr/datasets/revenus-et-pauvrete-des-menages-aux-niveaux-national-et-local-revenus-localises-sociaux-et-fiscaux/#resources https://www.insee.fr/fr/statistiques/5393560 data : https://www.insee.fr/fr/statistiques/fichier/5393560/fd_eec20_csv.zip SAS SCRIPT : https://www.insee.fr/fr/statistiques/fichier/5393560/Pgm_CSV_SAS_EEC2020.zip
- nais / deces mar 2015 https://www.insee.fr/fr/statistiques/2406457
- rev ++ 2015 *https://www.insee.fr/fr/statistiques/3560118* https://www.insee.fr/fr/statistiques/3560121 https://www.insee.fr/fr/statistiques/5055909 data : https://www.insee.fr/fr/statistiques/fichier/3560121/filo-revenu-pauvrete-menage-2015.zip
- Évolution du nombre d'équipements et de services entre 2015 et 2020 https://www.insee.fr/fr/statistiques/3606476?sommaire=3568656
- Structure et distribution des revenus, inégalité des niveaux de vie en 2017 https://www.insee.fr/fr/statistiques/4291712
- Structure et distribution des revenus, inégalité des niveaux de vie en 2019 *https://www.insee.fr/fr/statistiques/fichier/6036907/indic-struct-distrib-revenu-2019-COMMUNES_csv.zip*

https://www.data.gouv.fr/fr/datasets/niveau-de-vie-des-francais-par-commune/ https://www.data.gouv.fr/fr/datasets/fichier-des-personnes-decedees/ https://www.data.gouv.fr/fr/datasets/revenus-et-pauvrete-des-menages-aux-niveaux-national-et-local-revenus-localises-sociaux-et-fiscaux/ https://www.data.gouv.fr/fr/datasets/activite-emploi-et-chomage-enquete-emploi-en-continu-fichiers-detail/ https://www.data.gouv.fr/fr/datasets/comptes-des-communes-2012-2020/ https://www.data.gouv.fr/fr/datasets/balances-comptables-des-communes/ https://www.data.gouv.fr/fr/datasets/code-officiel-geographique-cog/ https://www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncieres/ https://www.data.gouv.fr/fr/datasets/data-insee-sur-les-communes/ https://www.data.gouv.fr/fr/datasets/donnees-geographiques-des-communes-par-code-insee-nd/