Projet Météo : Rapport n° 1 proposition de plan

# Description dataset

## Introduction :

* Description générale, objectif : prédire la pluie le lendemain à partir de données du jour.
* Volumétrie : lignes, colonnes, taille fichier, % de nan  : 01. Exploration des données
* Description des variables : template Excel Variables .xlsx

## Autour des NaN :

* Graph striplot : certaines grandeurs ne sont pas du tout mesurées par certaines stations.

02. Description stat

* Exploration : heatmap des nan par station pour confirmer 03. Stations
* Recherche des grandeurs non mesurées : dict, dict\_reverse, modalités = 9 03. Stations
* D’où une première idée de gestion des nan : faire 9 df
* Analyse de la distribution temporelle des nan : fichier etude\_nan.ipynb .Conclusion : choix du niveau de granularité. Voir aussi 15. Stat des moyennes par station pour une justification de la nécéssité de faire des moyennes par stations vs la moyenne tout court : il ya une grande variablitié des moyennes d’une grandeur selon les stations.

# Description statistique :

* Intro via un simple describe : rien à signaler, la distribution des valeurs semble cohérente.

dataviz\_quanti

* Analyse plus poussée : boxplot des var num, puis des var num scalées , puis des boxplot par station

dataviz\_quanti

* Il faudrait connaitre la quantité de valeurs extrêmes pour savoir quoi en faire. D’où le plot et le calcul en % -> les valeurs sont normales, on choisit la moyenne/médiane comme gestion des nan.

dataviz\_quanti

* Var cat : count plot. Tout a l’air normal aussi

dataviz\_quali

* Reste un pb : la distribution de rainfall. dataviz\_quanti

## Analyse des corrélation entre variables : dataviz\_correlations

* Var num : bilan des corrélation
* Var cat : idem

## Trous Chronologiques : 07. Séries temporelles

Mise en évidence, détermination des dates associées, commentaires

# Nettoyage des données :

## Quels choix ?

* + Var num : moyenne par station ou drop
  + Var cat : mode[0] par station ou drop
  + Suppression pour Rainfall, RainToday et RainTomorrow
  + Bilan en termes de pertes de données : % de perdu ?

## Pre processing :

* + Scale des var num,
  + dichotomisation des var cat avec un problème posé : les 3 var cat WindDir (WindGustDir, WindDir9am, WindDir3pm) qui contiennent chacune 16 modalités vont ajouter 3 x 16 = 48 colonnes au tableau, alors que les corrélations vis-à-vis de ces variables sont faibles. Je préconise donc de les supprimer les colonnes.

Feature engineering : Création des colonnes Year, Month, Day. 10. Création dates

* D’autres idées?

Insérer les élements fournis par les contacts de Fred ?