Rapport sur les données météorologiques Australiennes

Table des matières

[1 Introduction 3](#_Toc145076935)

[2 Exploration des données et visualisation 3](#_Toc145076936)

[2.1 Description du dataset 3](#_Toc145076937)

[2.2 Statistiques descriptives : 3](#_Toc145076938)

[2.3 Exploitation des valeurs manquantes 3](#_Toc145076939)

[2.4 Exploitation des valeurs extrèmes 3](#_Toc145076940)

[3 Pre-processing et feature engineering 3](#_Toc145076941)

[3.1 Nettoyage des données 3](#_Toc145076942)

[3.1.1 Traitement des valeurs manquantes 3](#_Toc145076943)

[3.1.2 Traitement des valeurs extrêmes 3](#_Toc145076944)

[3.2 Transformation des données 4](#_Toc145076945)

[3.3 Préparation pour la Modélisation 4](#_Toc145076946)

[3.3.1 Séparation du dataset 4](#_Toc145076947)

[3.3.2 Sélection des variables pour la modélisation 4](#_Toc145076948)

[4 Conclusion 4](#_Toc145076949)

# Introduction

* Présenter la météorologie australienne : une brève description du climat général de l’Australie, ses particularités régionales,…
* Objectif du projet :

# Exploration des données et visualisation

## Description du dataset

* Source de données
* Présenter les variables du data set
* Un Overview du data set: nombre d’observations (ligne), nombre de variables (colonne), période de collection de données, fréquence des observations, …

## Statistiques descriptives :

* Pour les variables catégorielles : fréquence des catégories
* Pour les variables numériques : histogram, boxplot -> permettre de connaitre la distribution de chaque variable.
* Corrélation entre les variables

## Exploitation des valeurs manquantes

* Identification des valeurs manquantes par variable (colonne) et par observation (ligne). Déterminer la corrélation de nullité entre les variables (les 3 graphiques générés par la librairie missingno)
* Identification des problèmes potentiels :
  + Colonnes problématiques (les analyses de Sophie sur certaines variables, certaines villes et certaines périodes)
  + Lignes aves des données manquantes excessives

## Exploitation des valeurs extrèmes

* Identification des valeurs extrêmes/aberrantes des variables numériques.

# Pre-processing et feature engineering

## Nettoyage des données

* Remplacer les valeurs de deux variables booléennes (Yes -> 1 et No -> 0)

### Traitement des valeurs manquantes

* Méthodes utilisées
* Résultats et impact

### Traitement des valeurs extrêmes

* Explication des raisons pour lesquelles qu’on a tendance de ne pas les supprimer.
* Correction (convertir les données des

## Transformation des données

* Encodage des variables catégorielles : décrire la méthode dont les variables catégorielles ont été traitées. (Quyen utilise la méthode get\_dummies de pandas pour encoder les variables catégorielles)
* Suppression des variables ou création de nouvelles variables (Sophie a expliqué ce point dans son rapport).
* Normalisation/Standardisation : techniques choisises et raisonnement (Quyen utilise la méthode sklearn.preprocessing.StandardScaler)

## Préparation pour la Modélisation

### Séparation du dataset

* Comment les données ont été divisées en ensembles d’entraînement, de validation et de test (train-test-split, validation croisée…)
* Proportions

### Sélection des variables pour la modélisation

* Quelles variables seront utilisées pour la modélisation
* Justifications

# Conclusion

Résumer des principales découvertes et des difficultés rencontrées