

Rapport Projet BIG DATA

Du 22/05/2023 au 27/05/2023

**Groupe 1 – trinôme 6 :** LE MOAL Gauthier, LE ROUZIC Emilie, REQUET Hippolyte

**Sommaire**

1. Préparation des données
   1. Nettoyage des données
   2. Recodage des variables multimodales en chiffres
   3. Mettre les variables sous format numérique
   4. Construction de la série chronologique
   5. Construction d’un jeu de données
2. Visualisation des données
   1. Représentation graphiques des données
   2. Implémentation des histogrammes
   3. Représentation des données dans une carte
3. Analyse des données
   1. Relation entre les variables qualitatives
   2. Régressions linéaires du nombre d’accidents
4. **Préparation des données**
5. **Nettoyage des données**

Pour bien débuter ce projet nous avons dû nettoyer le fichier csv donner par les professeurs. Pour cela, nous avons créé une fonction « Nettoyage\_des\_données » nous permettant ainsi de supprimer toutes les lignes contenant une donnée vide ou bien une donnée que l’on considère comme absurde, voici quelques exemples :

* Exemple 1 : Ligne 3683 -> longitude > 90 ;
* Exemple 2 : Ligne 3684 -> latitude > 90 ;
* Exemple 3 : Ligne 52127 -> place = NULL ;

1. **Recodage des variables multimodales en chiffres**

Par la suite nous dû recoder les variables multimodales en chiffres pour nous simplifier l’exploitation des données du tableau.

1. **Mettre les variables sous format numériques**

Lors de la description statistique du tableau de données nous avons remarqué que certaines colonnes du tableau possédaient un problème de type de données (exemple : la colonne « age » était de type character). Afin de régler ce problème nous avons créé la fonction « valeur\_num\_type », cette fonction permet de changer le type des données présentes dans les colonnes qui nous intéressent.

1. **Construction de la série chronologique**

Après avoir nettoyer, recoder et modifier le tableau de données, nous avons construit une série chronologique pour analyser et modéliser l’occurrence des accidents au fil du temps. Ici c’est la fonction « construire\_series\_chronologiques » qui nous permet de réaliser cette série. De plus, elle nous permet d’effectuer une série de prévision en fonction de la période choisie.

Si l'erreur de prédiction pour le niveau d'agrégation par mois est plus faible que celle pour le niveau d'agrégation par semaine, cela signifie que la régression linéaire basée sur les données mensuelles offre une meilleure qualité de prédiction. Dans ce cas, le niveau d'agrégation recommandé serait mensuel.

En revanche, si l'erreur de prédiction pour le niveau d'agrégation par semaine est plus faible, cela indique que la régression linéaire basée sur les données hebdomadaires offre une meilleure qualité de prédiction. Dans ce cas, le niveau d'agrégation recommandé serait hebdomadaire.

1. **Construction du jeu de données**

Pour finir cette partie sur la préparation des données, on finit sur la construction d’un jeu de données avec le nombre d’accidents selon la gravité pour 100.000 habitants par région. Pour ce faire nous avons dû lier plusieurs tableaux de données afin d’obtenir la population par région.

1. **Visualisation des données**
2. **Représentation graphique des données**

Dans cette partie nous avons réaliser plusieurs graphiques avant de visualiser le nombre d’accidents en fonction de différents facteurs, on peut prendre par exemple le nombre d’accidents en fonction des conditions atmosphériques. Cela nous permet de connaitre les facteurs créant le plus d’accident.

1. **Implémentation des histogrammes**

Tout comme la partie précédente, ici on a implémenté deux histogrammes l’un relatant de la quantité d’accidents en fonction des tranches d’âges et l’autre sur la quantité d’accident par mois, voici leur représentation :

Une image contenant capture d’écran, Caractère coloré, texte, Tracé

Description générée automatiquementUne image contenant texte, capture d’écran, Caractère coloré, diagramme

Description générée automatiquement

1. **Représentation des données sur une carte**

Ici on s’est intéressé à la représentation de tous les accidents sur une carte de la France. Premièrement une représentation par région pour basculer par la suite sur une représentation par départements.

Une image contenant texte, carte, atlas

Description générée automatiquement

Une image contenant carte, texte, atlas, capture d’écran

Description générée automatiquement

1. **Analyse des données**
2. **Relations entre les variables qualitatives**

Dans cette partie nous avons réalisé un test du chi2 sur les différentes variables du tableau. Grâce à ce test nous récupérons tous les rapports de corrélation entre toutes les variables du tableau. Par la suite nous avons afficher les résultats de ce test sur une mosaïque, ce qui nous permet d’avoir une meilleure représentation de ces rapports.

1. **Régression linéaire du nombre d’accidents**

Enfin, nous avons réalisé une régression linéaire des données du tableau afin de modéliser la relation entre une variable dépendante continue et une ou plusieurs variables indépendantes. Par la suite nous avons affiché la droite de régression concernant le nombre d’accidents par mois pour afficher ensuite une autre droite mais sur les accidents par semaine.

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquementUne image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement