Rapport de Projet Big DATA

Sommaire:

1. Description des données

- **Quelles sont les informations présentes ?**

Usagers:

- Numéro accident > *int*
- Id usager > *int*
- Code insee > *int*
- Année de naissance > *int*
- Âge > *int*
- **Gravité > indemne, tué, blessé hospitalisé, blessé léger -> passage en int (1 à 4)**
- Motif trajet > *Promenade – loisirs, Domicile – travail, Utilisation professionnelle, Courses – achats, Domicile – école, Autre*

Véhicule:

- Catégorie véhicule >

VL seul Motocyclette > 125 cm3 / Scooter < 50 cm3 / VU seul / 1,5T <= PTAC <= 3,5T avec ou sans remorque / Cyclomoteur <50cm3 / Bicyclette / Scooter > 50 cm3 et <= 125 cm3 / Motocyclette > 50 cm3 et <= 125 cm3 / PL seul > 7,5T / Autobus / Tracteur routier + semi-remorque / Scooter > 125 cm3 / PL > 3,5T + remorque / Autocar / PL seul 3,5T <PTCA <= 7,5T / Voiturette (Quadricycle à moteur carrossé) (anciennement "voiturette ou tricycle à moteur") / Autre véhicule / Tramway / Quad lourd > 50 cm3 (Quadricycle à moteur non carrossé) / Engin spécial / Tracteur agricole / Tracteur routier seul / Quad léger <= 50 cm3 (Quadricycle à moteur non carrossé)

- **Numéro véhicule > exemple : A01 -> passage en int (1 à 56)**
- Places > *int de 1 à 9 ou NULL*
- Dispositif de sécurité >

Utilisation d'une ceinture de sécurité Utilisation d'un casque Présence d'une ceinture de sécurité - Utilisation non déterminable Autre - Non déterminable Présence d'un casque - Utilisation non déterminable Présence de ceinture de sécurité non utilisée Présence d'un casque non utilisé Utilisation d'un dispositif enfant Autre - Utilisé Autre - Non utilisé Utilisation d'un équipement réfléchissant Présence équipement réfléchissant - Utilisation non déterminable Présence d'un équipement réfléchissant non utilisé Présence dispositif enfant - Utilisation non déterminable Présence d'un dispositif enfant non utilisé

Lieux:

- Date > *Datetime*
- Ville > *string*
- Latitude > *float*
- Longitude > *float*
- Agglomération > *binaire : Hors agglo ou En agglo*
- Intersection >

Hors intersection Intersection en X Intersection en T Giratoire Intersection à plus de 4 branches Intersection en Y Autre intersection Place Passage à niveau

Caractéristiques:

- Type de collision >

Deux véhicules – Par le coté Autre collision Deux véhicules – Par l'arrière Deux véhicules - Frontale Sans collision Trois véhicules et plus – En chaîne Trois véhicules et plus – Collisions multiples

- Description atmosphérique >

Normale Pluie légère Temps couvert Pluie forte Temps éblouissant Neige – grêle Autre Brouillard – fumée Vent fort – tempête

- Luminosité >

Plein jour Nuit avec éclairage public allumé Nuit sans éclairage public Crépuscule ou aube Nuit avec éclairage public non allumé

- Etat de la surface >

Normale Mouillée Verglacée Autre Enneigée Corps gras – huile Flaques Boue Inondée

---

1. Préparation

- **Quels traitements appliquer lorsqu'il manque des informations ou que ces informations ne sont pas exploitables ? Présenter quelques exemples et expliquer les traitements effectués.**

Lorsque les données sont inexploitables ou contiennent des valeurs manquantes on supprime toute la donnée. Dans le cas où la latitude vaut 2009,or la latitude est une mesure entre -90 et 90. Afin de déterminer des données aberrantes(outliers) on effectue un boxplot et on détermine la variance. Donc on passe d'une quantité de données de 73 644 à 55 428.

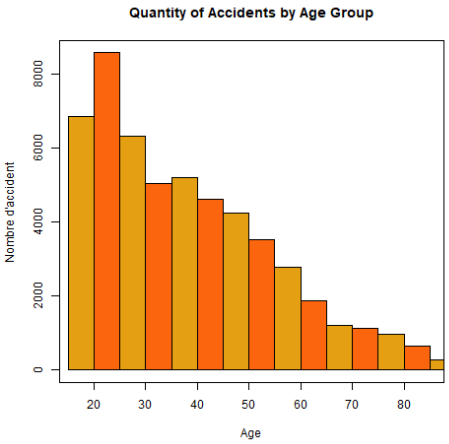
- **Construire des séries chronologiques sur l'évolution du nombre d'accidents par mois et semaines sur l'ensemble de la période. A quel niveau d'agrégation (mois ou semaine) les données collectées permettraient-elles de faire une prévision de bonne qualité avec une régression linéaire ?**

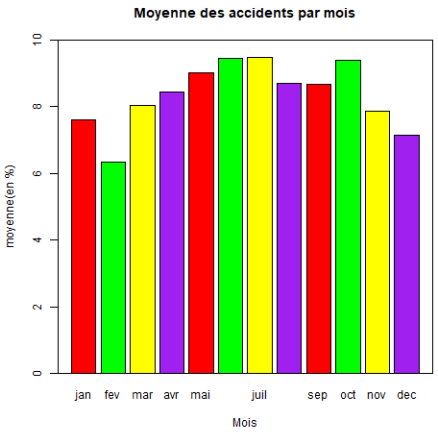
Plus l'agrégation contient de points plus la prévision grâce à une régression linéaire sera précise ; ainsi, il y a 12 points pour les mois et 53 pour les semaines. Donc la régression linéaire sera plus précise dans le cas des semaines que des mois.

---

2. Visualisation

Description et analyse des histogrammes





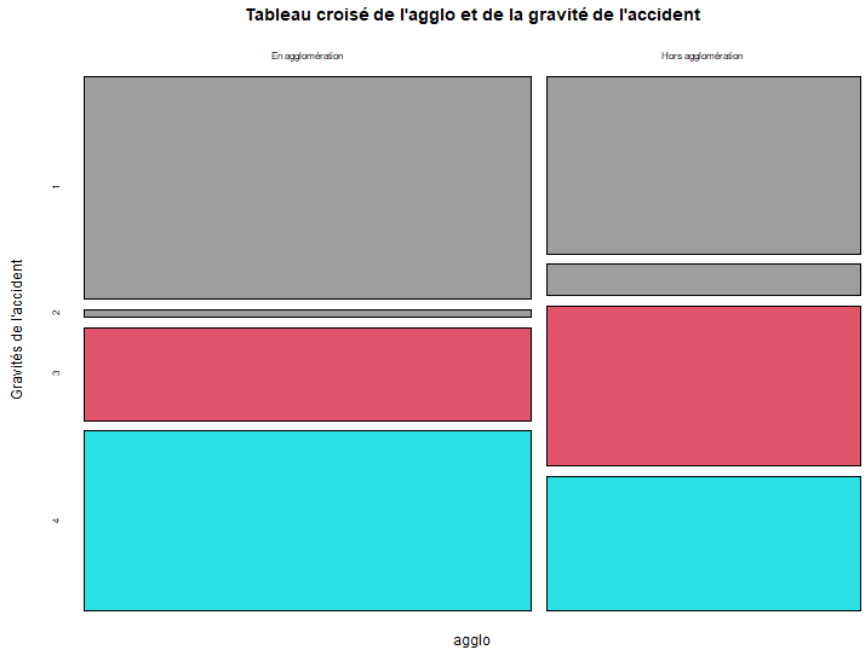
Concernant l'histogramme d'accident par âge, on remarque un très grand nombre d'accidents chez les jeunes conducteurs de la tranche d'âge 18-20 ans. Donc en tant que jeune conducteur on s'expose à plus de risques.

Concernant les mois, la moyenne indique un plus haut taux d'accident au mois de juin-juillet par rapport aux autres. Mais il est de l'ordre de 1% de différence par rapport aux autres. Il est difficile de conclure que le mois d'été augmente le taux d'accident

---

3. Analyse

- **Analyser MosaicPlot. Cas d'étude:**



Pour réaliser le mosaicplot il nous fallait un tableau croisé entre deux variables qualitatives. Après avoir appliqué la fonction mosaicplot() en R nous pouvons analyser le graphique.

La largeur des barres dans un mosaicplot représente l'effectif total des catégories de la variable indépendante (variable sur l'axe horizontal).
La hauteur des barres dans un mosaicplot indique la proportion des catégories de la variable dépendante à l'intérieur de chaque catégorie de la variable indépendante. Ainsi, la hauteur des barres représente la distribution des catégories de la variable dépendante pour chaque catégorie de la variable indépendante.

La surface des tuiles dans un mosaicplot reflète la taille totale de chaque combinaison de catégories. Les tuiles plus grandes indiquent des proportions plus élevées, tandis que les tuiles plus petites indiquent des proportions plus faibles.

Les tuiles d'un mosaicplot peuvent être colorées pour mettre en évidence des associations spécifiques ou des tendances visuelles.

Test du khi-2:

- La statistique du test du khi2 (X-squared) de 2640.4 indique une grande divergence entre les données observées et les données attendues sous l'hypothèse nulle.
- Le degré de liberté (df) de 3 signifie qu'il y avait 3 catégories ou groupes de données qui étaient libres de varier lors du test.

- La valeur p très faible (< 2.2e-16) indique que les résultats du test sont statistiquement significatifs. En d'autres termes, il y a des preuves solides pour rejeter l'hypothèse nulle selon laquelle il n'y a pas d'association entre les variables étudiées.

En résumé, les résultats montrent qu'il existe une association significative entre les variables analysées en fonction des valeurs du test du chi carré, des degrés de liberté et des valeurs de p. Cela signifie qu'il est peu probable que les variables étudiées soient indépendantes. Test chi2 p-value < 2.2e-16, < 0.05 donc les variables (gravité et agglo) sont liées.

- **Analyse des Régressions**

Les deux tableaux montrent une corrélation extrêmement forte entre l'évolution du nombre d'accidents et le temps, que ce soit par mois ou par semaine. Les coefficients de corrélation sont très proches de 1, ce qui suggère une relation linéaire positive entre ces variables.

Les intervalles de confiance à 95% donnent une estimation de la précision de nos mesures. Pour le nombre d'accidents par mois, l'intervalle se situe entre 4708.208 et 4931.3653, tandis que pour le nombre d'accidents par semaine, il se situe entre 1100.359 et 1118.768. Cela signifie que nous sommes très confiants dans la précision de nos estimations.

Les valeurs de R2 mesurent la proportion de la variation observée dans le nombre d'accidents qui peut être expliquée par notre modèle. Dans les deux cas, les valeurs de R2 sont très élevées (0.9989217 et 0.9991479), ce qui suggère que notre modèle est capable d'expliquer la majorité de la variation observée.

Enfin, les valeurs de R2 ajusté prennent en compte le nombre de variables explicatives dans notre modèle. Les valeurs élevées de R2 ajusté (0.9988138 et 0.9991308) indiquent que notre modèle est bien ajusté et qu'il n'y a pas de variables redondantes ou inutiles

Erreur standard des résidus : L'erreur standard des résidus est une mesure de la dispersion des résidus autour de la ligne de régression. Dans ce cas, l'erreur standard des résidus est de 496 pour la semaine .

Dans ce cas, l'erreur standard des résidus est de 598,8 pour la régression d'accident par mois.

En conclusion, les deux analyses montrent des résultats similaires avec une corrélation forte, des coefficients significatifs, des intervalles de confiance étroits et des valeurs élevées de R^2. Cela suggère que le modèle de régression linéaire est capable de bien expliquer la variation du nombre d'accidents en fonction du temps. La différence est minime entre la semaine et le mois, la semaine est très légèrement performante