*Nom de la team : RSA*

*Sprint n°0 - Projet N°6 Criminalité à Chicago*

*Scrum Master : Adrien Golebiewski*

*Teammates : Ryme Youb, Sacha Cymermann*

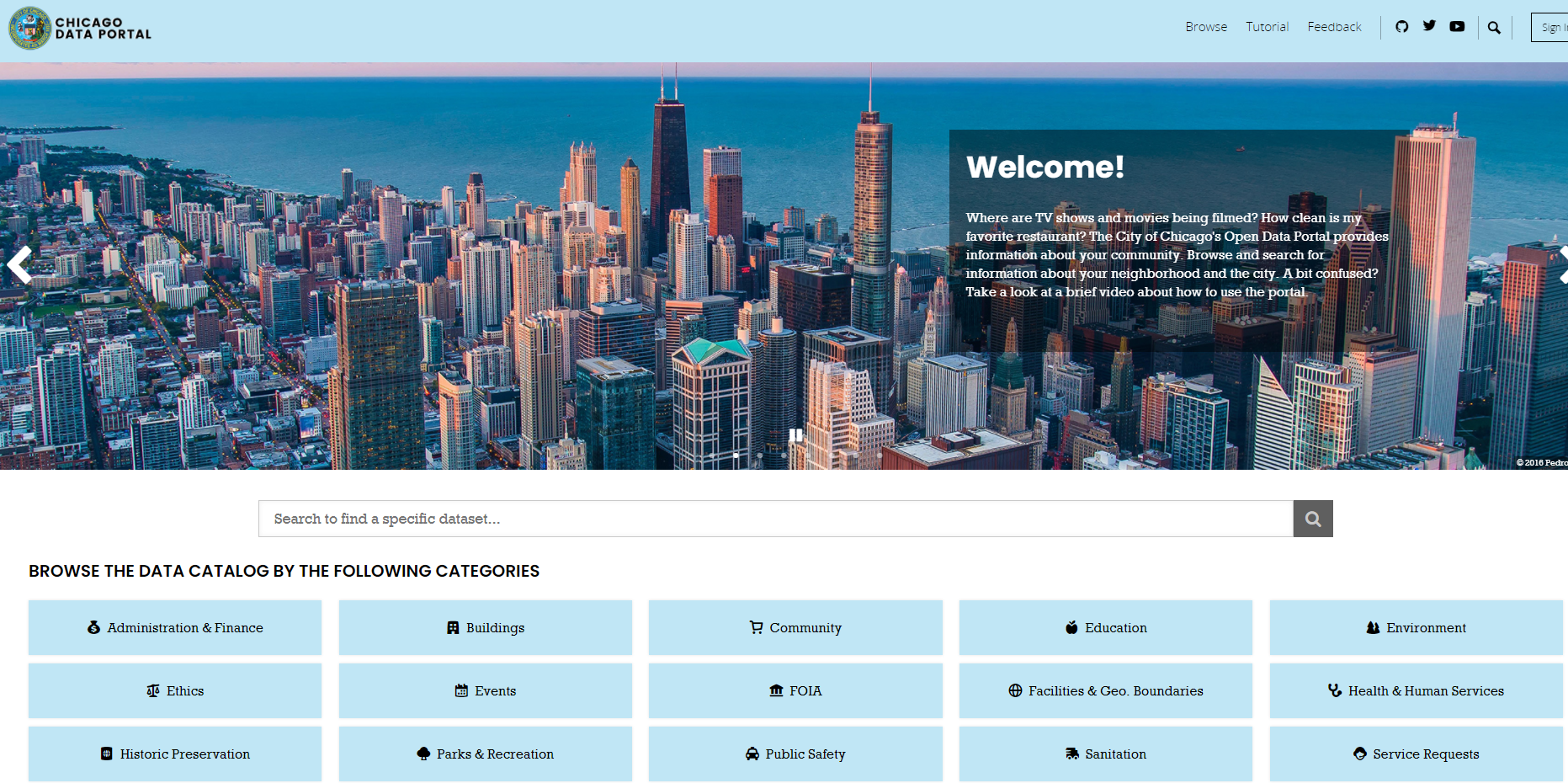
*Début : 06/09/2021*

*Fin : 20/09/2021*

**Sprint 0 - Définition des objectifs et du design de la solution technique**

**Présentation et description de la BDD**

La ville de Chicago est historiquement entachée d’une mauvaise réputation, ce qui lui vaut le surnom de « capitale du crime ». De plus, son urbanisation rapide associée à une hétérogénéité sociale et économique en ont fait un terrain d’étude privilégié des liens entre ville et criminalité.   
  
C’est dans ce contexte que nous avons choisi de traiter cette problématique de criminalité actuelle en étudiant les données fournies par le “Chicago Data Portal”, portail de données en OPEN DATA par la ville de Chicago (<https://data.cityofchicago.org/>).



Le grand volume de données associées à la criminalité à Chicago (environ plus de 1,62 Go) a également été un argument dans le choix de notre base de données dans le cadre des directives de notre projet (projet 6). Il était en effet préférable pour nous de pouvoir exploiter un volume important de données en accord avec le contexte “big data” de l’enseignement.

L’ensemble de données décrit 7,39M de crimes dans la ville de Chicago depuis 2001 selon plusieurs caractéristiques :

* géographiques (community area, distric, coordinate, location),
* temporelles (année du crime, date)
* type de crimes et gravité (arrest, domestic, beat, FBI Code, Primary description)
* Type de fichier

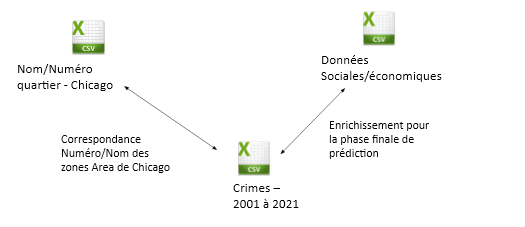
Fichiers csv

* Quelles sont les données ?

→ Fichier csv central “**Crimes - 2001 to Present**“ qui liste les crimes recensés sur Chicago aux USA de 2001 jusqu’à aujourd’hui ainsi que leurs caractéristiques. Taille : 1,62 Go.

→ Fichier csv “**Community Area Name**”. Association des noms des communautés de Chicago et des numéros. Taille : 2 Ko

→ Fichier CVS “**Hardship Index**” (éléments sociaux-économiques par Community Area Name). Taille : 4 Ko. Données complémentaires pour expliquer certaines variables du fichier csv crimes. Données utiles pour l’objectif final de prédiction



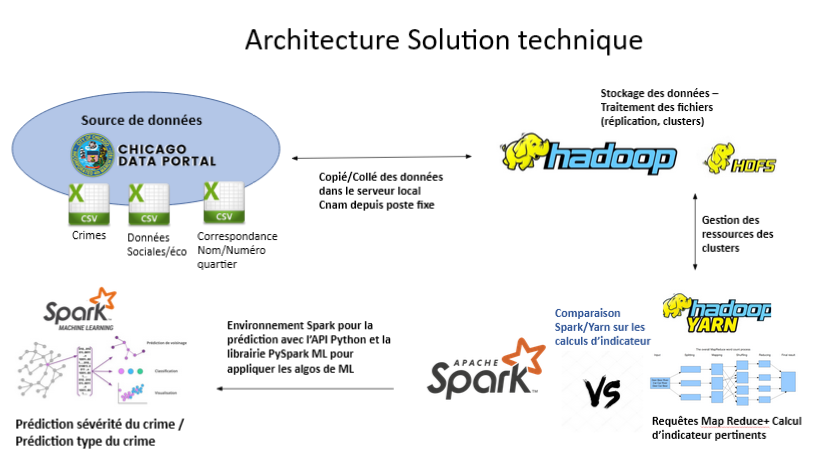
**Définition du design de l’architecture technique de la solution**

La principale caractéristique de ce projet est la variété des outils et des process que nous devons utiliser et ou appliquer.

1. Source de données (Données criminalité Chicago csv - x3 fichiers)
2. Stockage et traitement des fichiers sources (Hadoop/HDFS) - (Petite démo à réaliser pour la présentation plénière du TO)

Environnement de travail depuis les postes fixes du CNAM pour accès au serveur Hadoop + transférer les fichiers de données CSV dans ce serveur

1. Requêtage MapReduce (Hadoop/HDFS - YARN) pour calculs d’indicateurs pertinents résumant la complexité des crimes à Chicago.
2. Comparaison des résultats avec les indicateurs calculés avec Spark : temps de calcul, qualité des résultats, etc ...
3. Prédiction de la sévérité du crime et de son type à l’aide de l’API Python dans Spark et de la librairie Python PySpark pour le Machine learning



**Objectifs finaux :**

* Calcul d’indicateurs pertinents pour expliquer du mieux possible la complexité des crimes à Chicago
* Prédire la sévérité du crime et le type de crime en appliquant les algorithmes de machine learning étudiés durant ces 3 années au CNAM.

**Map Reduce - Définition des indicateurs, portée et objectifs avec Yarn/Spark**

Objectif de résumer la complexité des crimes de Chicago à la fois pour les aspects temporels (évolution du nombre de crimes), géographiques (nombre de crimes par area) et structurels à travers différents indicateurs.

5 exemples d’indicateurs à calculer :

*Nombre total de crimes par an*

*Nombre total d’arrestations par an*

*Nombre total de crimes par “Location Description”*

*Nombre total de crimes par “Primary Type”*

*Nombre total de vols par description et location description*

La définition de ces indicateurs pourra être amenée à évoluer tout au long du projet en fonction de la réalisation technique et des résultats obtenus.

Ces indicateurs seront calculés grâce à la technique Map Reduce à la fois par l’intermédiaire de la couche Yarn d’Hadoop et de Spark.

**Choix des prédictions**

La finalité du projet est l’apprentissage automatique de nos données afin de pouvoir prédire :

* la sévérité du crime (crime sévère / Non sévère)
* le type de crime (vol, infraction …)

Environnement Spark + librairie PySpark pour développer les algorithmes de ML.

En fonction de l’avancement du projet et du temps restant, la phase de prédiction pourra être réduite qu’à une seule prédiction.

Les variables explicatives dans le cadre de la prédiction seront les suivantes :

* 'location\_description'
* 'arrest'
* 'domestic'
* 'beat'
* 'district'
* 'ward'
* 'community\_area'
* 'fbi\_code'
* 'hour'
* 'week\_day'
* 'year\_month'
* 'month\_day'
* 'date\_number'
* les données “sociales/économiques” issues du second fichier csv.
* PERCENT HOUSEHOLDS
* BELOW POVERTY
* PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA
* PERCENT AGED 16+ UNEMPLOYED
* PERCENT AGED UNDER 18 OR OVER 64
* PER CAPITA INCOME

Les réseaux de neurones multi-couches (MlP Classifier), les forêts aléatoires et l’algorithme SVM seront les algorithmes d’apprentissage à faire tester sur nos données d’apprentissage.

L’enjeu est de réaliser les meilleures prédictions possibles mais aussi de retenir le meilleur algorithme en termes de résultats (qualité et robustesse du modèle) en comparaison avec les autres.

**Démonstration technique - Prototype calcul d’indicateurs**

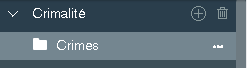
Les différentes problématiques d’accès à Internet ainsi que les retards d’accès au serveur HDFS nous n’ont pas permis en fin de Sprint 0 de faire une démonstration technique sur Hadoop/HDFS (copier/coller des données). Ainsi la réalisation d’indicateurs depuis l’éco-système Hadoop était impossible à réaliser au niveau du timing.

C’est dans ce contexte que le fin du sprint T0 a donné lieu à la réalisation d’un prototype de création d’indicateurs à partir des outils MongoDB et Python. L’enjeu est de pouvoir calculer un certain nombre d’indicateurs à partir de requêtes non pas depuis Hadoop mais depuis d’autres outils Big Data.

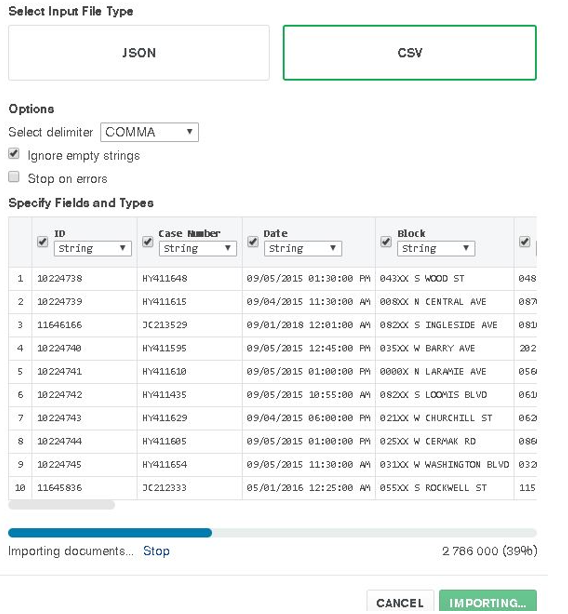
Nous avons décidé d’utiliser MongoDB ainsi que le langage Python car pertinents dans le contexte Big data et correspondent à des outils auxquels nous sommes familiers.

Pour celà, nous sommes passés par plusieurs étapes :

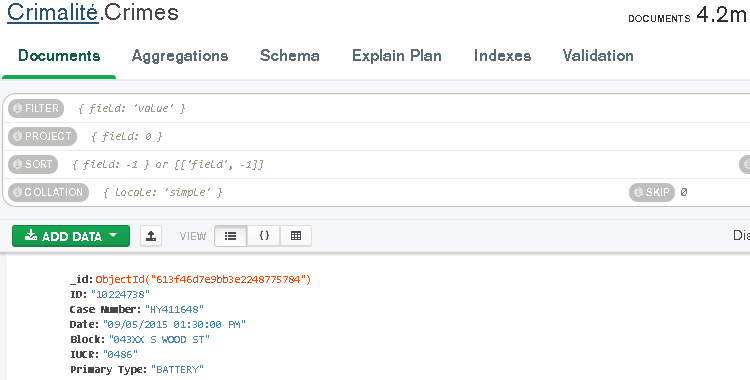
1. Dans le GUI MonGo Compass, création d’une base de données "Criminalité" et d’une première collection “crime”



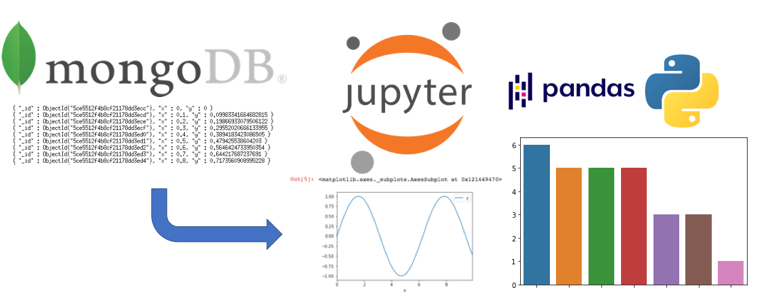
1. Chargement du fichier csv “crimes” dans la collection créée précédemment



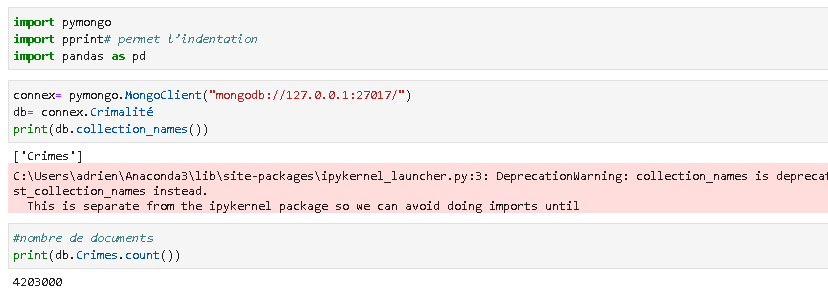
1. Nous retrouvons ensuite l’ensemble des données du fichiers répartis en document, soit au total 4.2 Millions de documents (4.2 millions de crimes)



1. Enfin, les données étant convenablement chargées, l’idée est de pouvoir requêter notre data base et de pouvoir construire un certain nombre d’indicateurs. Pour pouvoir réaliser cette étape, nous avons souhaité appliquer un certain nombre de démarches réalisées l’année dernière lors de l’UE “Python”. En effet, grâce à la librairie Python “Pymongo”, il est possible de requêter et de calculer des indicateurs depuis des bases de données enregistrées en localhost sur MongoDB. Nous avons décidé de choisir cette méthode pour calculer nos indicateurs sur la criminalité à Chicago.



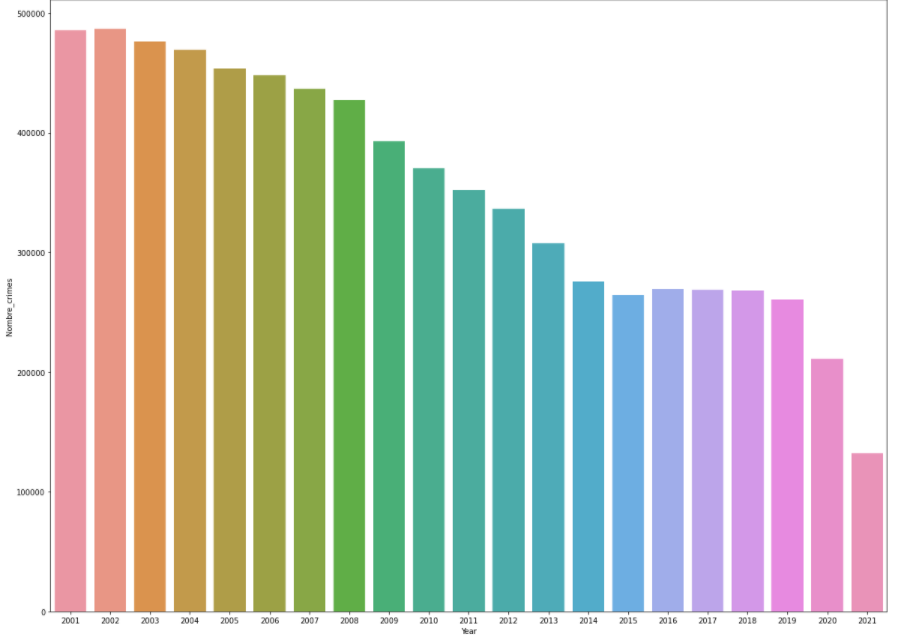
Nous nous connectons à MongoDB à l’aide de la fonction “pymongo.MongoClient” puis récupérons les données de la base “Criminalité” grâce à la fonction “connex”. On peut dès lors recalculer le nombre de crimes (nombre de documents) à l’aide de la fonction “db.Crimes.count()”.



1. Nous avons ainsi calculé l’ensemble des indicateurs définis précédemment à l’aide de la librairie PyMongo de Python. Ces indicateurs sont décrits à travers des résultats sous forme de table et de graphique pour expliquer l’évolution ou le statut d’un phénomène.

Voici les résultats obtenus ci-dessous :

***Nombre total de crimes par an***

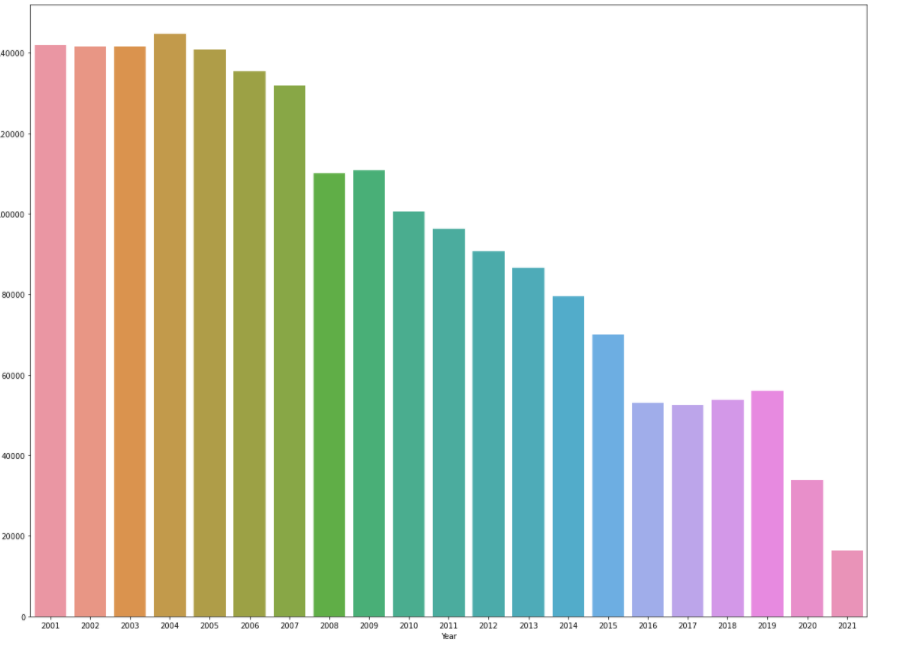
**

On observe une baisse du nombre de crimes par an entre 2001 et 2021. Mais ce nombre de crimes reste élevé et la baisse est ralentie depuis 2015.

Le taux de criminalité a baissé à Chicago progressivement depuis 2003. Et baisse de manière continue depuis 2007 (comme dans d’autres villes de USA).

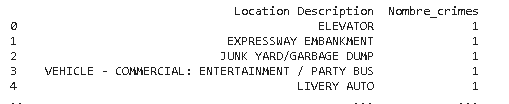
Malgré quelques affaires particulièrement sanglantes, selon des décomptes publiés par les journaux locaux de Chicago, la criminalité est en baisse. Cette baisse peut s’expliquer notamment par des mesures de sécurité prises par les autorités de la ville (mises en place de caméra de surveillance, arrestation ciblées et groupées). Le nombre de crimes a diminué (C'est le chiffre le plus bas depuis la fin de la Deuxième guerre mondiale, en chute spectaculaire par rapport au pic de 2.245 morts enregistrés en 1990) mais les crimes deviennent de plus en plus graves et sanglants.

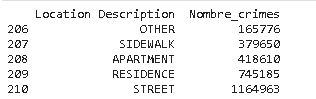
***Nombre total d’arrestations par an :***

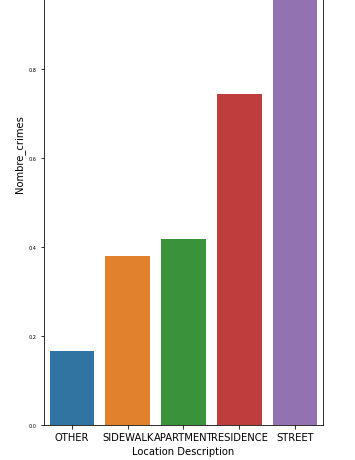
**

Le nombre d’arrestations suit la même évolution que le nombre de crimes.

***Nombre total de crimes par “Location Description”***

**

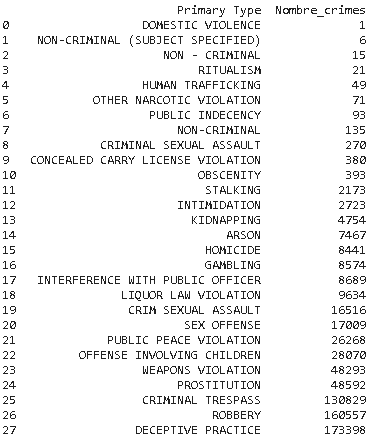
**

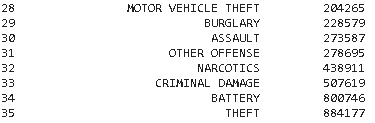
**

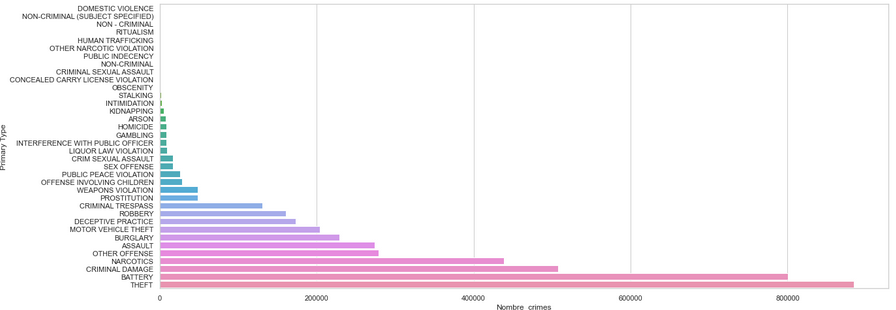
Le graphique ainsi que le tableau récapitulatif dénotent le nombre conséquent de crimes qui a lieu dans les rues de la ville de Chicago. Plus d’1 million de “crimes” depuis 2002 ont eu lieu dans les rues de la ville. Viennent ensuite les crimes dans les résidences, appartement qui sont également à un nombre très élevé.

A l’inverse, très peu de crimes ont été recensés chez les fournisseurs automobiles, ascenseurs, ou dépôts d'ordures.

***Nombre total de crimes par “Primary Type”***

**

**

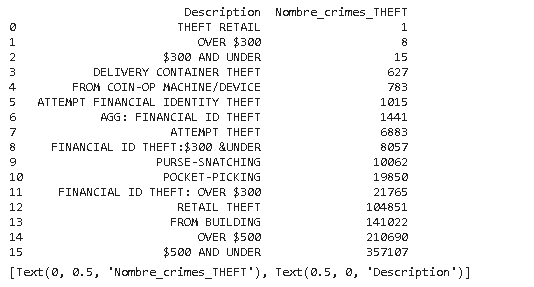
**

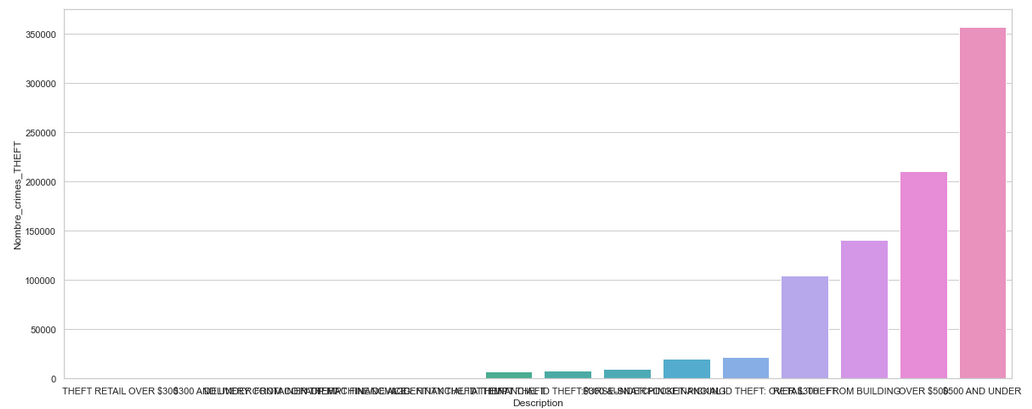
Par le tableau et le graphique, on constate que le Primary Type “Thieft” est le plus courant parmi l’ensemble des crimes. Plus de 884 000 vols depuis 2001 à Chicago sont des vols.

Viennent ensuite “batery” (attaques engendrant des blessures) et “criminal damage”. A l’inverse, les violences domestiques sont quasiment inexistantes.

***Nombre total de vols par description et location description***

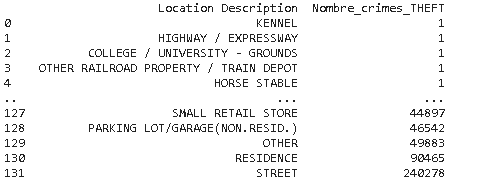
A partir de l’indicateur précédent, nous nous sommes intéressés sur les statistiques de vols qui ont lieu à Chicago. Nous souhaitons notamment récupérer les caractéristiques de chaque vol ainsi que leur nombre :





On remarque ainsi que la majorité des vols effectués sont des vols d’argent avec des gains allant jusqu'à plus de 500 dollars et qu’ils sont souvent réalisés à l’intérieur de bâtiment.

Ensuite, nous souhaitons notamment récupérer les localisations de chaque vol ainsi que leur nombre :



On se rend compte ainsi que la majorité des vols (comme la majorité des crimes, cf 2ème indicateur) s’effectue dans la rue.

Ces indicateurs sont des exemples de mesures que nous pourons calculer depuis l’éco-système d’Hadoop. Ils résument simplement l’état actuel de la criminalité à Chicago selon différents niveaux de granularité.

**Bilan du Sprint TO :**

A l’issue du Sprint TO, plusieurs étapes essentielles ont été réalisées, notamment :

* le choix de notre base de données d’étude
* la définition de notre architecture technique
* la construction de prototype d’indicateurs résumant les données de criminalité de la ville de Chicago

Les problèmes d’accès à internet depuis le lieu de l’école nous a empêché de finaliser notre Sprint T0 avec une démonstration du Stockage des données dans l'écosystème Hadoop (HDFS). Cependant, en stockant les données depuis la base de données MongoDB (Base No SQL) et en les requêtant à l’aide de la librairie “PyMongo” Python, nous avons réussi à prototyper l’ensemble de nos indicateurs définis en début de Sprint.

Nos objectifs fixés en début de Sprint sont donc atteints. De plus, la méthode Agile employée dans le cadre de ce projet nous a permis de rebondir face à une problématique d’accès à internet en définissant notamment un prototype de calcul d’indicateurs.

Le Sprint 1 sera consacré au stockage des données dans HDFS, la gestion des fichiers au sein de l’éco-systèe Hadoop et au lancement de requêtes Map Reduce. Le sprint 2 se concentrera sur la prédiction d’indicateurs et la réalisation d’algorithmes de Machine Learning à travers l’outil Spark et sa librairie Python PySpark.

Ci-dessous, la backlog mise à jour à l’issue de ce Sprint T0 :

