

**Université Moulay Ismail  
Faculté des Sciences et  
Techniques Er-Rachidia**

## **Mini projet Datamining**

Présenté à la Faculté des Sciences et Techniques

Cycle d'ingénieur d'état

Filière: Génie Informatique

# **Réclamations d'assurance - Détection de fraudes**

Réalisé par

- ZAHIR Ayman

Supervisé par:

- Pr. Mohamed Sabiri

# Table des matières

## Table des matières

Mini projet Datamining.....	1
Table des matières.....	3
Introduction générale .....	5
Chapitre1 : THEORIE D'INFOMATIQUE DECISIONNE.....	6
L'informatique décisionnelle .....	6
1. Définition et objectif de l'informatique décisionnelle .....	6
2. Les principaux fondamentaux.....	7
Environnement logiciel.....	8
Chapitre2 : REALISATION DU PROJET .....	9
Introduction au Jeu de Données "Insurance_Claims" .....	9
Contenu du Jeu de Données : .....	9
Objectifs du Jeu de Données :.....	9
Source et Qualité :.....	9
Processus de réalisation de projet : .....	10
I. Méthodologie .....	10
II. Analyse exploratoire des données :.....	11
III. Traitement des données :.....	14
IV. Construire les tables de dimensions et de faits.....	21
V. Reporting.....	30
VI . Modèles et Notebook.....	40
Conclusion .....	41
Références :.....	43

# Liste de figures

Figure 1 architecture du data mining.....	7
Figure 2 processus de réalisation de projet sous JIRA .....	10
Figure 4: aperçu des colonnes du fichier CSV 2 .....	13
Figure 3: aperçu des colonnes du fichier CSV .....	13
Figure 5 échantillon de fichier CSV de la base des données.....	14
Figure 6 le shema général du job du data cleaning.....	15
Figure 7 filtrage des données du fichier CSV utilisant tMap .....	16
Figure 8 remplacement des données vides utilisant StringHandling.....	16
Figure 9 Figure 8 remplacement des données de string en boolean .....	17
Figure 10 l'affichage de données et l'enregistrement dans la base de données MySQL.....	17
Figure 11 résultats obtenus après l'exécution du JOB .....	18
Figure 12 la structure le base de données Datawarehouse .....	18
Figure 13 Les données dans Datawarehouse .....	19
Figure 14 schéma général des tables de dimensions et table de fais .....	21
Figure 15 Schéma en étoile de chaque table de fait avec les tables de dimensions.....	21
Figure 16 l'interface du spoon pentaho.....	22
Figure 17 les schémas pour créer chaque table de dimension .....	22
Figure 18 création d'une connexion a la base de données MySQL .....	23
Figure 19 extraction des colonnes spécifique depuis datawarehouse.....	23
Figure 20 spécifier les colonnes de sortie dans la nouvelle dimension .....	24
Figure 21 Schéma des tables de faits1 .....	25
Figure 22 Schéma des tables de faits2.....	25
Figure 23 jointure de plusieurs tables de dimensions .....	26
Figure 24 contrôle de flux .....	27
Figure 25 sauvegarde dans la base de données.....	28
Figure 26 les tables de dimensions et faits dans la base de données .....	29
Figure 27 table de dimension de véhicule .....	29
Figure 28 table de fait de claim .....	29
Figure 29 Nombre de fraude reporté par incidnet_state .....	30
Figure 30 Nombre de fraude reporté par sexe .....	31
Figure 31 somme montants total claims par âge .....	32
Figure 32 pourcentage de fraude reportés par profession d'assurant .....	33
Figure 33 nombre de fraudes par type d'incidence .....	34
Figure 34 pourcentage de fraude reportés par le niveau d'éducation d'assurant.....	35
Figure 35 pourcentage de fraudes reportés par mois d'incidence et par date de policy.....	36
Figure 36 pourcentage de fraudes reportés par heur d'incidence.....	37
Figure 37 pourcentage de fraudes reportés par sévérité d'incidence .....	38
Figure 38 nombre de fraudes reportés par présence du rapport de policy.....	39
Figure 39 nombre de fraudes reportés par loisirs d'assurant .....	40
Figure 40 : notebook du code .....	41

# Introduction générale

## Introduction :

L'informatique décisionnelle est une approche de gestion qui vise à aider les entreprises à prendre des décisions éclairées en utilisant des données et des outils d'analyse. Dans le cadre de notre module "**Data mining**", nous avons étudié les différents aspects de l'informatique décisionnelle et nous avons découvert comment elle peut être appliquée dans le contexte des projets d'entreprise.

Le but de ce travail est d'appliquer les concepts et les outils d'informatique décisionnelle que nous avons appris dans ce module à notre projet. Nous allons examiner comment l'informatique décisionnelle peut aider à résoudre les problèmes d'un projet et comment elle peut améliorer les performances de l'entreprise. Nous allons également analyser les étapes clés du processus d'informatique décisionnelle et les appliquer à notre projet.

Notre projet est sous le thème « **Étude des fraudes d'assurance** » Le présent rapport de projet est articulé autour de 2 chapitres :

Chapitre 1 : Théorie d'informatique décisionnelle

Chapitre 2 : Réalisation du projet

# **Chapitre1 : THEORIE D'INFOMATIQUE DECISIONNE**

## **L'informatique décisionnelle**

### **1. Définition et objectif de l'informatique décisionnelle**

L'informatique décisionnelle, est un ensemble de techniques, de processus et d'outils informatiques qui permettent de collecter, d'analyser et de présenter des données pour aider les entreprises à prendre des décisions éclairées.

L'informatique décisionnelle implique l'utilisation de bases de données, de technologies d'entreposage de données, d'outils d'analyse, de visualisation de données et de rapports pour extraire des informations précieuses à partir de grandes quantités de données. Les données peuvent être collectées à partir de différentes sources telles que des systèmes de gestion de bases de données, des fichiers plats, des applications, des appareils connectés à l'Internet des objets (IoT) et des médias sociaux.

L'informatique décisionnelle est utilisée dans de nombreux domaines, tels que les finances, la vente au détail, la fabrication, la logistique, les ressources humaines et la santé, pour aider les entreprises à améliorer leur performance, à prendre des décisions plus éclairées, à prévoir les tendances et à mieux comprendre leur marché.

Le data mining, ou fouille de données en français, est un domaine de l'intelligence artificielle qui vise à extraire des informations utiles de grandes quantités de données. Ces informations peuvent être utilisées pour prendre des décisions, améliorer des processus ou simplement comprendre le monde qui nous entoure.

Le data mining repose sur des techniques statistiques et d'apprentissage automatique. Ces techniques permettent de découvrir des modèles et des tendances dans les données, même si ces modèles ne sont pas évidents à l'œil nu.

Le data mining est une partie importante de l'informatique décisionnelle. L'informatique décisionnelle est le domaine de l'informatique qui vise à aider les individus et les organisations à prendre des décisions éclairées. Le data mining fournit à l'informatique décisionnelle les informations nécessaires pour prendre des décisions plus efficaces.

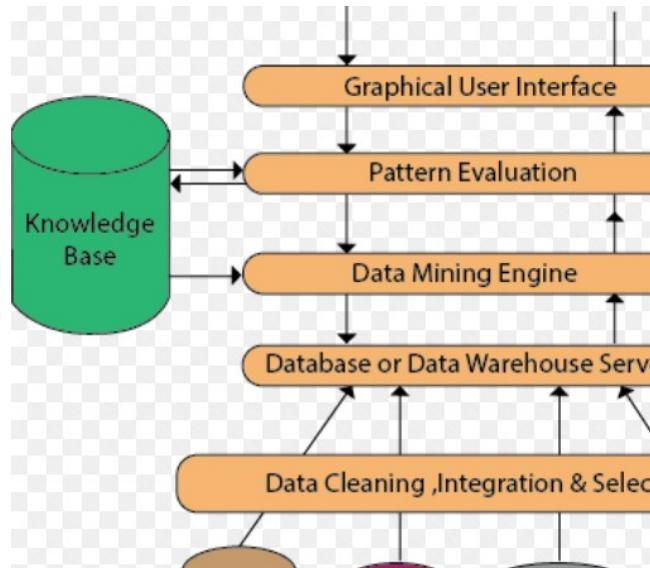


Figure 1 architecture du data mining

## 2. Les principaux fondamentaux

Les principes fondamentaux de l'informatique décisionnelle sont essentiels pour comprendre comment cette approche fonctionne et comment elle peut être appliquée dans les entreprises.

### a. La collecte de données :

1. Identification des sources de données - données internes, données externes, données provenant de tierces parties, etc.
2. Extraction des données - processus d'extraction, de transformation et de chargement (ETL)
3. Stockage des données - bases de données, data warehouses, data lakes, etc.

### b. L'analyse de données :

1. Transformation des données - nettoyage, normalisation, agrégation, etc.
2. Analyse des données - exploration de données, visualisation de données, modélisation de données, etc.
3. Interprétation des résultats - identification de tendances, de relations, de comportements, etc.

### c. La prise de décision :

1. Utilisation des résultats - prise de décision basée sur des données, identification des opportunités, identification des risques, etc.
2. Communication des résultats - rapports, tableaux de bord, graphiques, présentations, etc.
3. Réajustement des actions - ajustement des stratégies, des plans, des objectifs, etc.

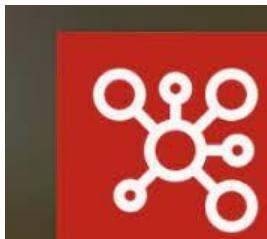
### d. La surveillance des résultats :

1. Mesure des résultats - KPI, métriques, indicateurs de performance, etc.
2. Évaluation des résultats - comparaison avec les objectifs, analyse des écarts, etc.
3. Optimisation des résultats - identification des opportunités d'amélioration, ajustement des stratégies, des plans, des objectifs, etc.

Les principes fondamentaux de l'informatique décisionnelle sont essentiels pour comprendre comment cette approche fonctionne et comment elle peut être appliquée dans les entreprises. Les entreprises doivent comprendre comment collecter et stocker

des données, comment analyser ces données, comment prendre des décisions basées sur ces données et comment surveiller les résultats pour améliorer leur performance. La compréhension de ces principes fondamentaux peut aider les entreprises à tirer parti de l'informatique décisionnelle pour améliorer leur prise de décision et leur performance globale.

## Environnement logiciel



Pentaho est une plate-forme d'intégration de données open source qui permet de collecter, de transformer, d'analyser et de visualiser des données provenant de différentes sources. L'interface graphique de Pentaho est appelée Spoon, qui est un outil de conception graphique pour créer des flux de données et des processus ETL.



Talend Open Studio est un outil open source d'intégration de données qui permet de collecter, transformer et charger des données provenant de différentes sources. Il est développé par Talend, une entreprise de logiciels de gestion de données.



MySQL est un système de gestion de bases de données relationnelles (SGBDR). Il est distribué sous une double licence GPL et propriétaire. Il fait partie des logiciels de gestion de base de données les plus utilisés au monde<sup>2</sup>, autant par le grand public (applications web principalement) que par des professionnels, en concurrence avec Oracle, PostgreSQL et Microsoft SQL Server.



Microsoft Power BI est une solution d'analyse de données de Microsoft. Il permet de créer des visualisations de données personnalisées et interactives avec une interface suffisamment simple pour que les utilisateurs finaux créent leurs propres rapports et tableaux de bord.



Jira est un produit propriétaire développé par Atlassian qui permet le suivi des bugs, le suivi des problèmes et la gestion de projet agile. Jira est utilisé par un grand nombre de clients et d'utilisateurs dans le monde pour la gestion des projets, des délais, des exigences, des tâches, des bogues, des modifications, du code, des tests, des versions et des sprints.



**Spyder IDE** est un environnement de développement intégré (IDE) gratuit et open-source pour le langage Python. Il est conçu pour les scientifiques, les ingénieurs et les analystes de données..

# **Chapitre2 : REALISATION DU PROJET**

## **Introduction au Jeu de Données "Insurance\_Claims"**

Le jeu de données "insurance\_claims" offre un aperçu captivant du secteur de l'assurance en regroupant 1000 transactions soigneusement extraites de « **databriks** ». Ces données constituent une ressource précieuse pour les analystes, chercheurs et professionnels de l'assurance cherchant à comprendre les tendances, à évaluer les risques, et à améliorer la détection des fraudes.

### **Contenu du Jeu de Données :**

Le jeu de données est composé de 40 colonnes couvrant divers aspects liés aux réclamations d'assurance. Parmi ces colonnes figurent des informations cruciales telles que la durée en mois du client, l'âge, le numéro de police, les détails sur la police elle-même, des caractéristiques du client assuré, des détails sur l'incident, et bien plus encore. Un point notable est la présence d'une colonne spécifique "fraud\_reported", suggérant la possible occurrence de fraudes dans les réclamations.

### **Objectifs du Jeu de Données :**

Les objectifs de ce jeu de données sont multiples. Ils incluent la possibilité d'analyser les schémas d'incidents, de modéliser les comportements des assurés, et surtout, de développer des stratégies avancées pour la détection et la prévention des fraudes en assurance.

### **Source et Qualité :**

Ce jeu de données a été extrait de Kaggle, une plateforme renommée pour ses ensembles de données diversifiés. Il a été soigneusement préparé pour garantir la qualité et la cohérence des informations, offrant ainsi une base fiable pour des analyses approfondies.

En somme, le jeu de données "insurance\_claims" représente une opportunité intrigante d'exploration et de compréhension approfondie du domaine de l'assurance, mettant l'accent sur la détection proactive des fraudes pour optimiser la fiabilité et la transparence du secteur.

# Processus de réalisation de projet :

Pour réaliser ce projet, on va respecter les tâches suivantes comme ce qui montre la figure ci-dessous :

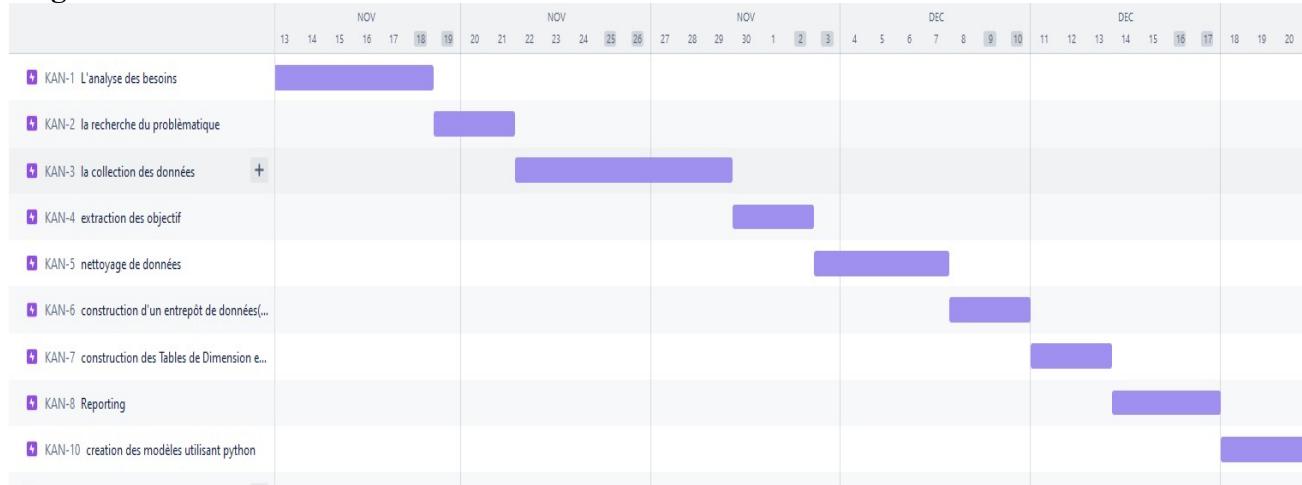


Figure 2 processus de réalisation de projet sous JIRA

## I. Méthodologie

La méthodologie présentée dans ce rapport est une approche générale pour la détection des fraudes d'assurance par data mining. Elle comprend les étapes suivantes :

- **Préparation des données**
- **Visualisation des données**
- **Extraction des caractéristiques**
- **Entraînement des modèles**
- **Évaluation des modèles**

### Préparation des données

Une fois les données collectées, il est nécessaire de les préparer pour le traitement. Cette étape comprend les étapes suivantes :

- **Nettoyage des données** : il s'agit de supprimer les données erronées ou incomplètes. On peut également appliquer des techniques de prétraitement des données,
- **Annotation des données** : il s'agit de étiqueter les données comme étant frauduleuses ou non frauduleuses. Cette étape nécessite l'intervention d'experts en fraude.

### Visualisation des données

Une fois les données préparées, il est utile de les visualiser pour en avoir une meilleure compréhension. La visualisation des données peut aider à identifier des tendances ou des anomalies qui pourraient être indicatives de fraudes.

## **Extraction des caractéristiques**

L'extraction des caractéristiques consiste à transformer les données en un format qui peut être utilisé par les algorithmes de classification. Les caractéristiques peuvent être des variables numériques, des variables catégorielles ou des combinaisons de ces deux types de variables.

## **Entraînement des modèles**

Une fois les caractéristiques extraites, on va essayer d'entraîner des modèles utilisant **SVM**, **KNN**, **DT**, **RF** et **SGB**

## **Évaluation des modèles**

Une fois les modèles entraînés , on va évaluer les performances de notre modèles. L'évaluation des modèles peut se faire en utilisant des métriques telles que la précision, le rappel, la F-mesure ou la courbe ROC.

## **II. Analyse exploratoire des données :**

### **1. Description des variables :**

Pour fournir une description détaillée des variables du jeu de données "insurance\_claims", examinons chaque colonne individuellement :

1. **months\_as\_customer (int) :**
  - Le nombre de mois pendant lesquels le client est assuré.
2. **age (int) :**
  - L'âge du client assuré.
3. **policy\_number (int) :**
  - Le numéro unique associé à la police d'assurance.
4. **policy\_bind\_date (string) :**
  - La date à laquelle la police d'assurance a été souscrite.
5. **policy\_state (string) :**
  - L'état où la police d'assurance est en vigueur.
6. **policy\_csl (string) :**
  - La limite de responsabilité combinée associée à la police.
7. **policy\_deductible (int) :**
  - Le montant de la franchise associé à la police.
8. **policy\_annual\_premium (double) :**
  - La prime annuelle de la police d'assurance.
9. **umbrella\_limit (int) :**
  - La limite d'indemnisation parapluie.
10. **insured\_zip (int) :**
  - Le code postal du client assuré.
11. **insured\_sex (string) :**
  - Le sexe du client assuré.
12. **insured\_education\_level (string) :**
  - Le niveau d'éducation du client assuré.

13. **insured\_occupation (string) :**
  - La profession du client assuré.
14. **insured\_hobbies (string) :**
  - Les passe-temps du client assuré.
15. **insured\_relationship (string) :**
  - La relation du client assuré avec le titulaire de la police.
16. **capital-gains (int) :**
  - Les gains en capital associés à la réclamation.
17. **capital-loss (int) :**
  - Les pertes en capital associées à la réclamation.
18. **incident\_date (string) :**
  - La date à laquelle l'incident a eu lieu.
19. **incident\_type (string) :**
  - Le type d'incident (vol, collision, etc.).
20. **collision\_type (string) :**
  - Le type de collision associé à l'incident.
21. **incident\_severity (string) :**
  - La gravité de l'incident (mineure, modérée, grave).
22. **authorities\_contacted (string) :**
  - Les autorités contactées après l'incident.
23. **incident\_state (string) :**
  - L'état où l'incident a eu lieu.
24. **incident\_city (string) :**
  - La ville où l'incident a eu lieu.
25. **incident\_location (string) :**
  - L'emplacement spécifique de l'incident.
26. **incident\_hour\_of\_the\_day (int) :**
  - L'heure de la journée à laquelle l'incident a eu lieu.
27. **number\_of\_vehicles\_involved (int) :**
  - Le nombre de véhicules impliqués dans l'incident.
28. **property\_damage (string) :**
  - Indique si des dommages matériels ont été causés lors de l'incident.
29. **bodily\_injuries (int) :**
  - Le nombre de blessures corporelles liées à l'incident.
30. **witnesses (int) :**
  - Le nombre de témoins présents lors de l'incident.
31. **police\_report\_available (string) :**
  - Indique si un rapport de police est disponible après l'incident.
32. **total\_claim\_amount (int) :**
  - Le montant total de la réclamation.
33. **injury\_claim (int) :**
  - Le montant de la réclamation liée aux blessures.
34. **property\_claim (int) :**
  - Le montant de la réclamation liée aux dommages matériels.
35. **vehicle\_claim (int) :**
  - Le montant de la réclamation liée au véhicule.
36. **auto\_make (string) :**
  - La marque du véhicule assuré.
37. **auto\_model (string) :**
  - Le modèle du véhicule assuré.

**38. auto\_year (int) :**

- L'année de fabrication du véhicule assuré.

**39. fraud\_reported (string) :**

- Indique si la réclamation a été signalée comme une fraude.

Colonne	Clé	Type	✓ N..	Modèle de date...	Longueur	Précision
months_as_customer	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		3	0
age	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		2	0
policy_number	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		6	0
policy_bind_date	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"yyyy-MM-dd"	10	0
policy_state	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		2	0
.	<input type="checkbox"/>	String	<input type="checkbox"/>		2	0

Figure 3: aperçu des colonnes du fichier CSV

Description du schéma						
Colonne	Clé	Type	✓ N..	Modèle de date...	Longueur	Précision
policy_deductable	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		4	0
policy_annual_premium	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		7	3
umbrella_limit	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		7	0
insured_zip	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		6	0
insured_sex	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		6	0
.	<input type="checkbox"/>	String	<input type="checkbox"/>		2	0

Figure 4: aperçu des colonnes du fichier CSV 2

## 2.Analyse de corrélations :

Après une analyse exploratoire des données, on va donc aborder les questions de recherche suivantes à l'aide des données de la base de données :

1. Y a-t-il une Corrélation Entre l'Âge de l'Assuré et le Montant Total des Réclamations Fraudeuses ?
2. Existe-t-il une Relation Entre le Type d'Incident et la Fréquence des Fraudes ?
3. Les Fraudes Sont-elles Plus Courantes chez un Certain Sexe ?
4. Y a-t-il une Tendance Saisonnière Dans les Fraudes ?
5. Existe-t-il une Corrélation Entre le Niveau d'Éducation et le Type d'Incident Fraudeux ?
6. Les Fraudes Sont-elles Liées à Certaines Occupations ?
7. Quelle Est la Corrélation Entre l'Heure de l'Incident et la Probabilité de Fraude ?
8. Y a-t-il une Corrélation Entre le Montant de la Réclamation et la Présence de Témoins ?
9. Les Fraudes Sont-elles Plus Fréquentes Avec Certains Types de Véhicules ?

10. Existe-t-il une Corrélation Entre la Disponibilité d'un Rapport de Police et le Taux de Fraude ?
  11. Y a-t-il une Relation Entre le Lieu de l'Incident et le Montant de la Réclamation Frauduleuse ?
  12. Les Fraudes Sont-elles Liées à Certains Modèles de Voitures ou Années de Fabrication ?

### **III. Traitement des données :**

#### **1. Nettoyage des données :**

Les données sont stockées dans un fichier CSV "insurance\_claims.csv" qui contient 40 colonnes et plus de 1000 lignes, donc on procède au nettoyage des données. C'est une étape importante du traitement des données. Elle consiste à identifier et à résoudre les problèmes potentiels dans les données, tels que les valeurs aberrantes, les données manquantes, les doublons et les erreurs de syntaxe.

## Figure 5 échantillon de fichier CSV de la base des données

Le nettoyage des données du fichier "insurance\_claims.csv" a été réalisé à l'aide du logiciel **Talend**.

Dans un premier temps, les données ont été filtrées en utilisant l'outil "tMap" pour remplacer les valeurs des enregistrements avec des valeurs manquantes par autre de même type en ce basant sur les anciennes valeurs pour les colonnes "subtitles", "rating" et "language".

Ensuite, un autre filtre "StringHandling.change()" a été appliqué pour changer les valeurs des enregistrements avec des valeurs manquantes pour les colonnes "collision\_type", "property\_damage", "policy\_report\_available", "incident\_date".

Enfin, un filtre unique a été appliqué à la colonne "id" pour créer une clé primaire pour les enregistrements.

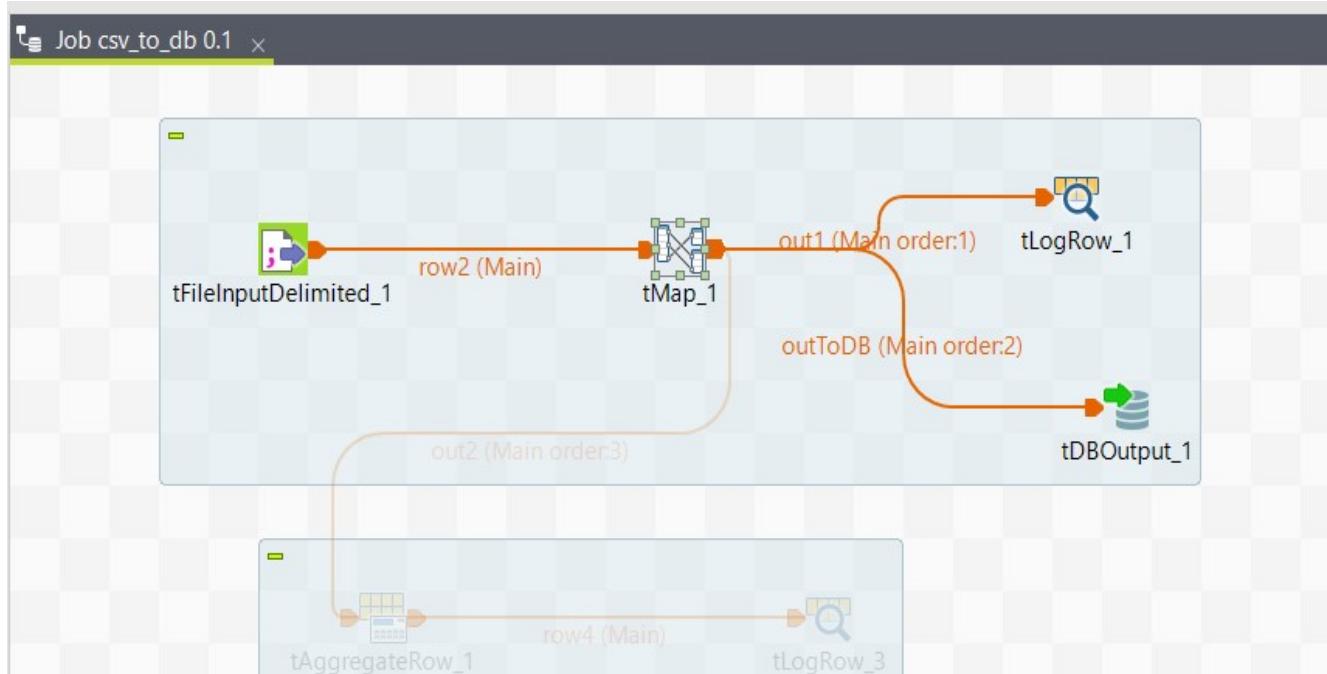


Figure 6 le schéma général du job du data cleaning

Talend Open Studio for Data Integration - tMap - tMap\_1

**row2**

**Var**

Expression	Type	Variable
Var.var2="Front collision"? "Rear collision";	String	var2
StringHandling.CHANGE(row2.collision_type, "\?", Var.var2);	String	collision_type
Var.var3=Var.var3="1"? "1";	String	var3
(row2.property_damage).equals("0")&...;	String	property_damage
(row2.police_report_available).equals("Y")&...;	String	police_report_available
row2.fraud_reported = "Y"?1:0	int	fraud_reported
row2.incident_date	Date	incident_date

**out1**

**outToDB**

Figure 7 filtrage des données du fichier CSV utilisant tMap

Constructeur d'expression

Expression

StringHandling.CHANGE(row2.collision\_type, "\?", Var.var2);

Test

Var	Vale
row2.month...	null
row2.age	null
row2.policy...	null

Ajouter Supprimer

Catégories

- \*Défini par l'utilisateur
- \*Tous
- DataOperation
- Mathematical
- Numeric
- Relational
- StringHandling
- TalendDataGenerator

Fonctions

Aide

Sélectionnez une fonction.

Figure 8 remplacement des données vides utilisant StringHandling

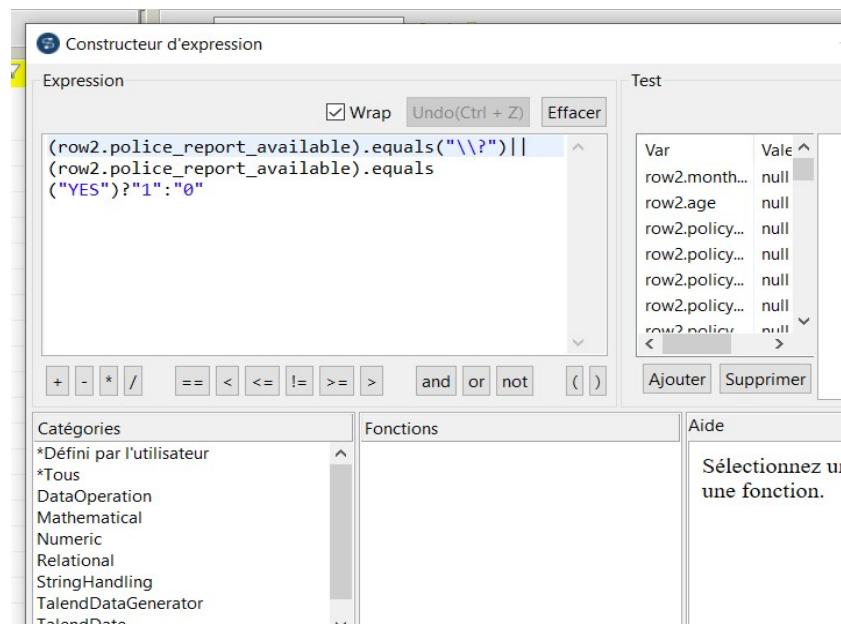


Figure 9 Figure 8 remplacement des données de string en boolean

## 2. Stockage des données dans une base de données

Les données nettoyées ont été stockées dans une table "games" de la base de données "insurance" dans la table « Datawarehouse » via une connexion MySQL avec l'outil 'tDBOutput'.

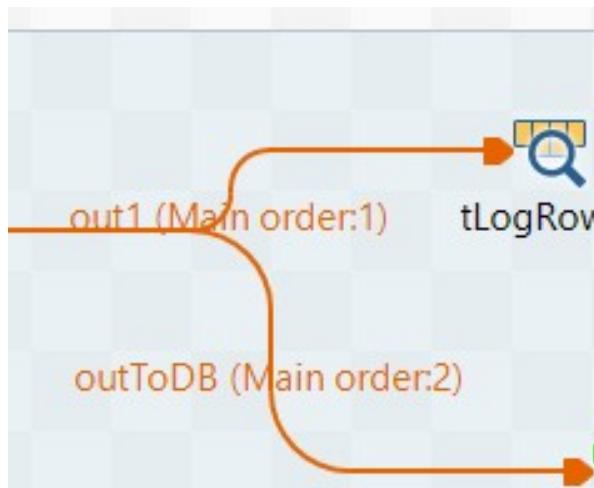


Figure 10 l'affichage de données et l'enregistrement dans la base de données MySQL

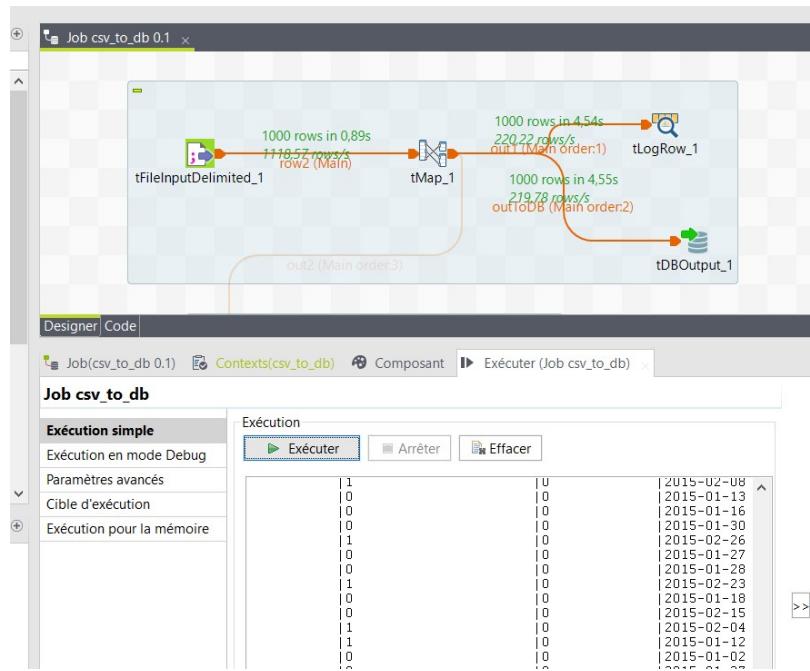


Figure 11 résultats obtenus après l'exécution du JOB

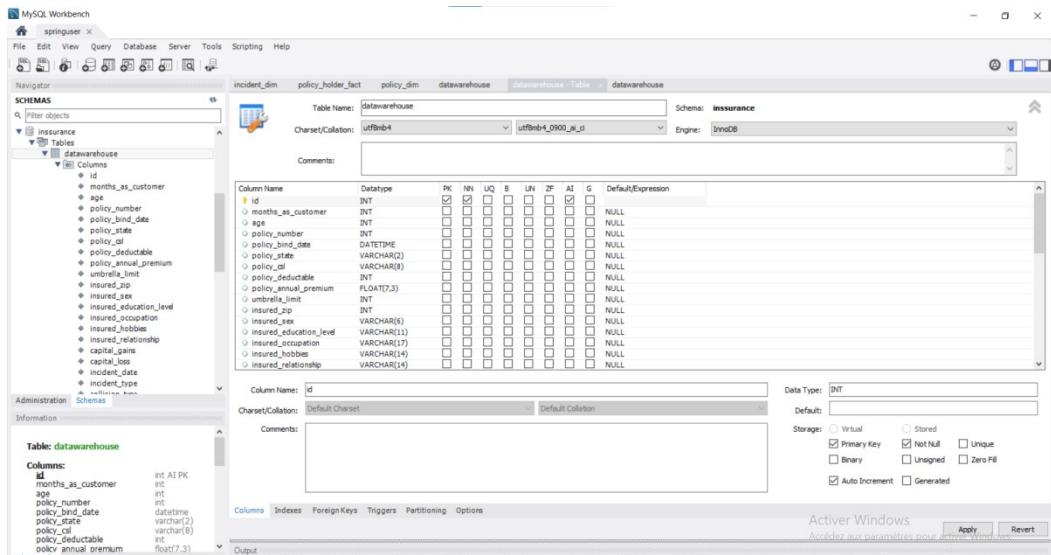


Figure 12 la structure le base de données Datawarehouse

The screenshot shows a database interface with a query window titled "1 • SELECT \* FROM insurance.datawarehouse;". The results are displayed in a grid with the following columns: id, months\_as\_customer, age, policy\_number, policy\_bind\_date, policy\_state, policy\_cd, policy\_deductible, policy\_annual\_premium, umbrella\_limit, insured\_zip, insured\_sex, insured\_education. The data consists of 14 rows of customer information.

	<b>id</b>	<b>months_as_customer</b>	<b>age</b>	<b>policy_number</b>	<b>policy_bind_date</b>	<b>policy_state</b>	<b>policy_cd</b>	<b>policy_deductible</b>	<b>policy_annual_premium</b>	<b>umbrella_limit</b>	<b>insured_zip</b>	<b>insured_sex</b>	<b>insured_education</b>
1	202	34	608513	2002-07-18 00:00:00	IN	100/300	500	846.070	3000000	607730	MALE	JD	
2	224	40	914088	1990-02-08 00:00:00	OH	100/300	2000	1291.700	0	609837	FEMALE	JD	
3	241	45	596785	2014-03-01 00:00:00	IL	500/1000	2000	1104.500	0	432211	FEMALE	PhD	
4	64	25	908616	2000-02-18 00:00:00	IL	250/500	1000	954.160	0	473328	MALE	Masters	
5	166	37	666333	2008-06-18 23:00:00	IL	100/300	2000	1337.280	8000000	610393	MALE	JD	
6	155	35	336614	2003-04-01 00:00:00	IL	500/1000	1000	1088.340	0	614780	FEMALE	Associate	
7	114	30	584859	1992-04-01 00:00:00	IL	100/300	1000	1558.290	0	472248	MALE	High School	
8	120	28	434507	2009-02-06 00:00:00	IL	250/500	1000	1281.270	0	447442	FEMALE	PhD	
9	149	37	994943	1991-01-15 00:00:00	IL	500/1000	500	1454.150	0	603381	MALE	PhD	
10	78	29	694642	2011-05-15 00:00:00	IL	250/500	2000	1254.200	6000000	497779	MALE	Masters	
11	200	35	966660	1994-08-21 00:00:00	IN	250/500	2000	1318.050	0	618496	MALE	High School	
12	284	48	498140	1997-05-15 00:00:00	IN	500/1000	2000	769.950	0	605486	MALE	Masters	
13	275	41	498875	1995-10-26 00:00:00	CH	100/300	2000	1514.720	0	617970	MALE	High School	
14	153	34	798177	2006-03-04 00:00:00	IL	500/1000	1000	873.640	4000000	432934	FEMALE	Associate	

Figure 13 Les données dans Datawarehouse

### 3. Conception multidimensionnelle de la base de données

#### a. détermination des tables des dimensions

Pour notre projet, nous avons décidé d'utiliser les données disponibles dans notre fichier de base de données pour déterminer les tables de dimensions les plus pertinentes. Nous avons choisi les dimensions les plus importantes en fonction de l'objectif de notre projet et des données disponibles.

#### 1. Time Dimension (time\_dim) :

Cette table capture les informations temporelles, notamment la durée en mois pendant laquelle le client est assuré.

#### 2. Vehicle Dimension (vehicle\_dim) :

Cette table contient des détails sur les véhicules assurés, tels que la marque, le modèle et l'année de fabrication.

#### 3. Policy Dimension (policy\_dim) :

Cette table rassemble des informations sur les polices d'assurance, y compris les détails de la police, la date de souscription, l'état et les primes associées.

#### 4. Insured Dimension (insured\_dim) :

Cette table capture des détails sur les assurés, y compris leur lieu de résidence, le sexe, le niveau d'éducation, la profession, les passe-temps et les relations familiales.

#### 5. Incident Dimension (incident\_dim) :

Cette table rassemble des informations sur les incidents, y compris la date, le type, la gravité, les contacts avec les autorités, l'emplacement et l'heure.

### b. Détermination des tables de faits

Pour compléter la conception de notre entrepôt de données, nous avons également créé des tables de fait appelée "X\_Fact".

Ces table de fait sont le centre de notre entrepôt de données ça dépend du sujet traité car elle contient le résultat « fraud\_reported » ainsi que des clés étrangères faisant référence aux tables de dimensions que nous avons déterminées précédemment. Qui nous donne la flexibilité de traité le sujet de différents parts.

#### 1. Claim Fact (claim\_fact) :

Cette table contient des mesures liées aux réclamations, notamment le montant total de la réclamation, les réclamations liées aux blessures, aux dommages matériels et au véhicule. Elle est liée aux dimensions de la police (policy\_dim) et de l'incident (incident\_dim).

#### 2. Incident Fact (incident\_fact) :

Cette table rassemble des mesures liées aux incidents, telles que le nombre de véhicules impliqués, les dommages matériels, les blessures corporelles, la présence de témoins et la disponibilité d'un rapport de police. Elle est liée aux dimensions de la police (policy\_dim), de l'assuré (insured\_dim), temporelle (time\_dim) et de l'incident (incident\_dim).

#### 3. Policy Holder Fact (policy\_holder\_fact) :

Cette table contient des mesures liées aux titulaires de polices, telles que le montant total de la réclamation, les réclamations liées aux blessures, aux dommages matériels et au véhicule. Elle est liée aux dimensions temporelle (time\_dim) et du véhicule (vehicle\_dim).

#### 4. Table de Faits Financiers (financial\_fact) :

Cette table de faits financiers contient des mesures spécifiques liées aux aspects financiers des réclamations d'assurance. Elle est conçue pour fournir des informations détaillées sur les implications financières des réclamations et des incidents.

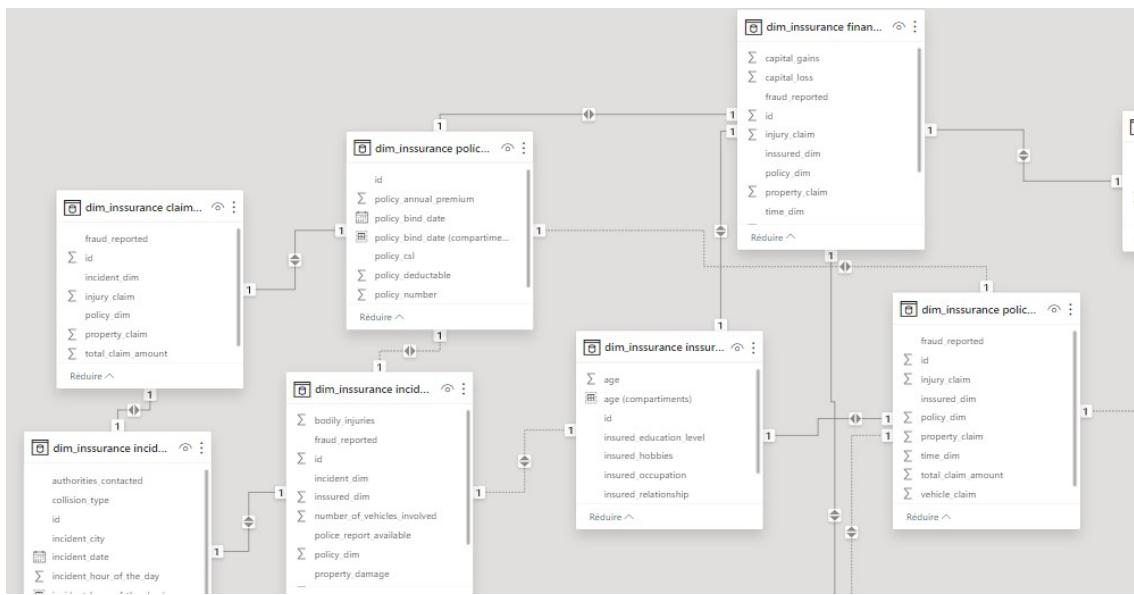


Figure 14 schéma général des tables de dimensions et table de fais

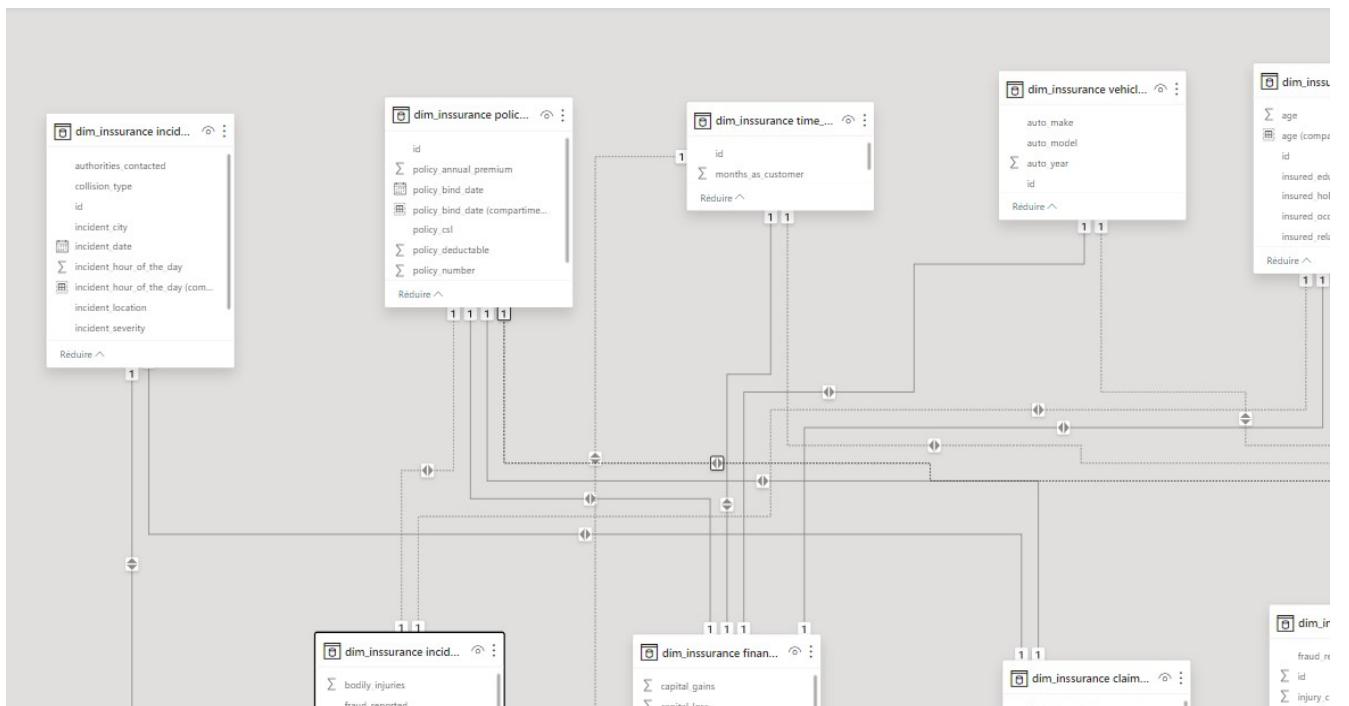


Figure 15 Schéma en étoile de chaque table de fait avec les tables de dimensions

## IV. Construire les tables de dimensions et de faits

Après avoir collecter et nettoyer les données, puis les charger dans la base de données MySQL ‘inssurance’ dans la table ‘Datawarehouse’, nous avons utilisé le logiciel Spoon Pentaho pour extraire, et transférer ces données dans les tables de dimensions.

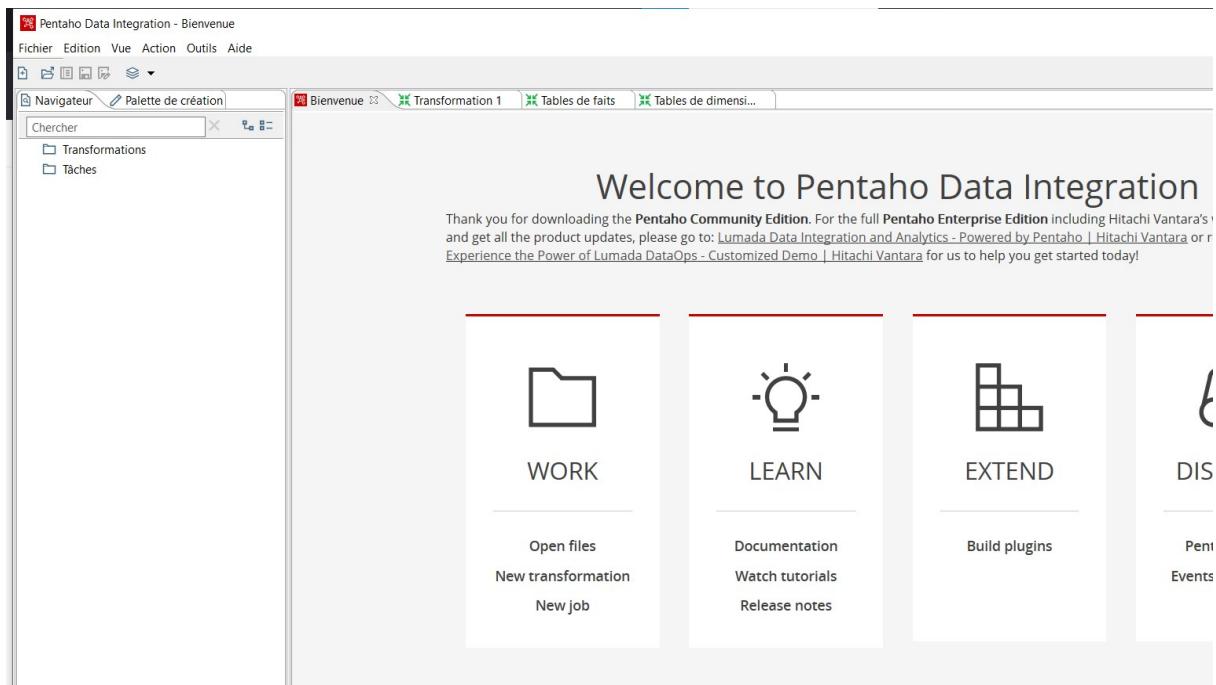


Figure 16 l'interface du spoon pentaho

## Création d'une Transformation pour les tables de dimensions

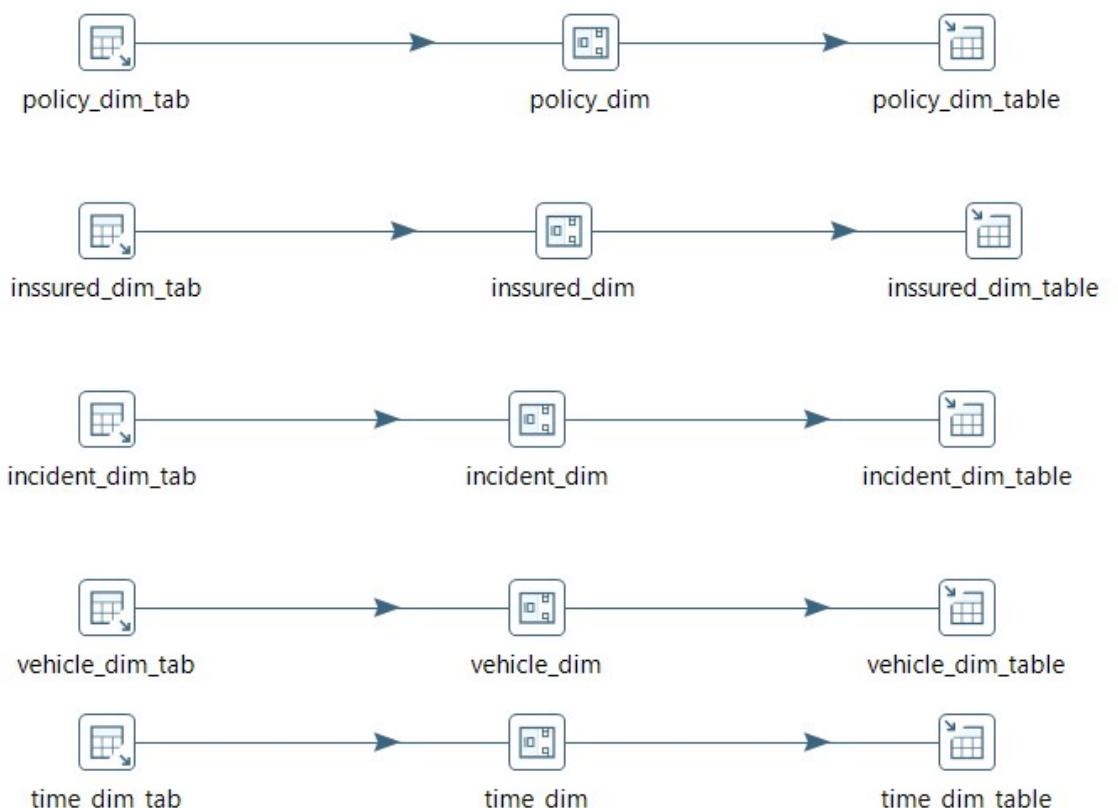


Figure 17 les schémas pour créer chaque table de dimension

## Exemple de création d'une table de dimension utilisant Pentaho

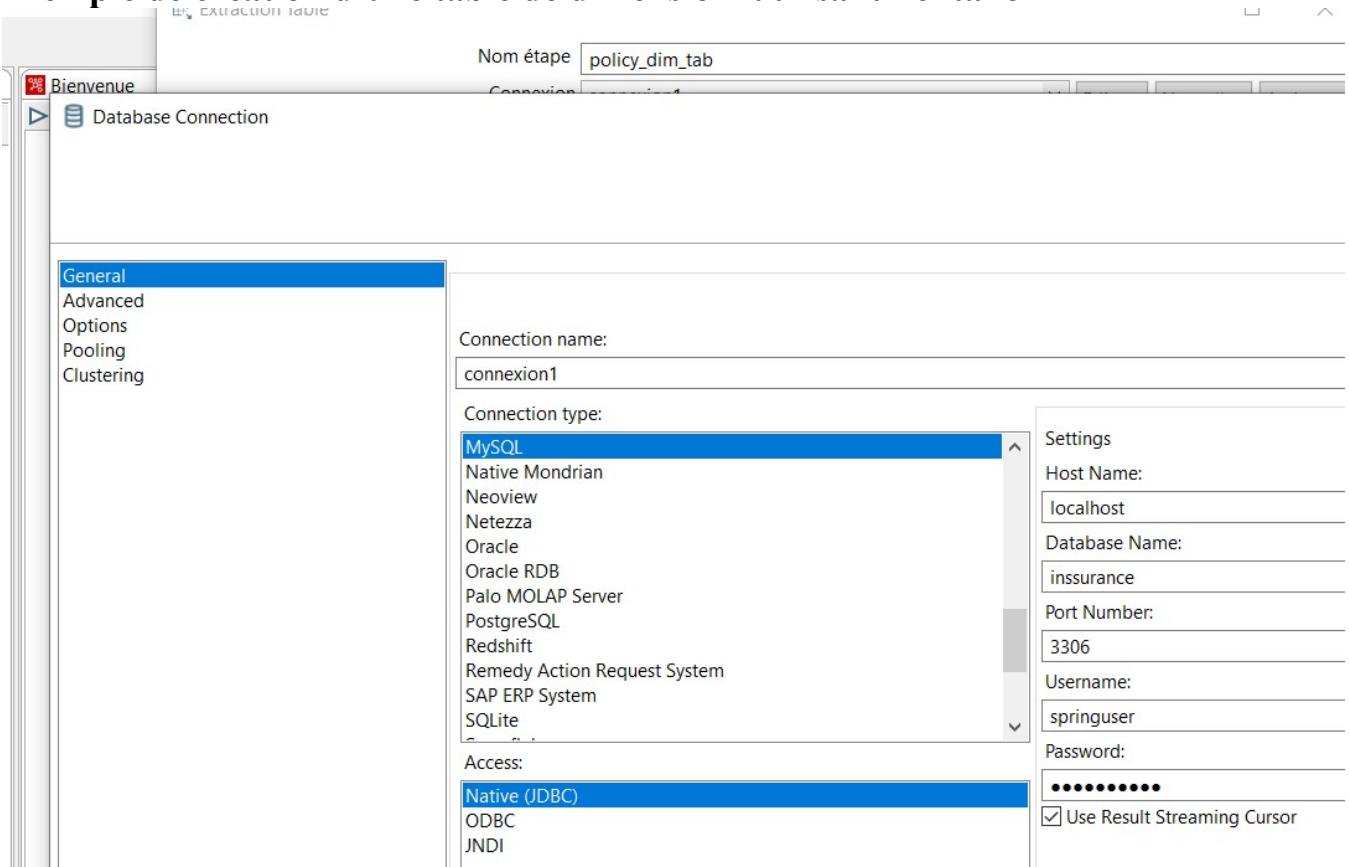


Figure 18 création d'une connexion a la base de données MySQL

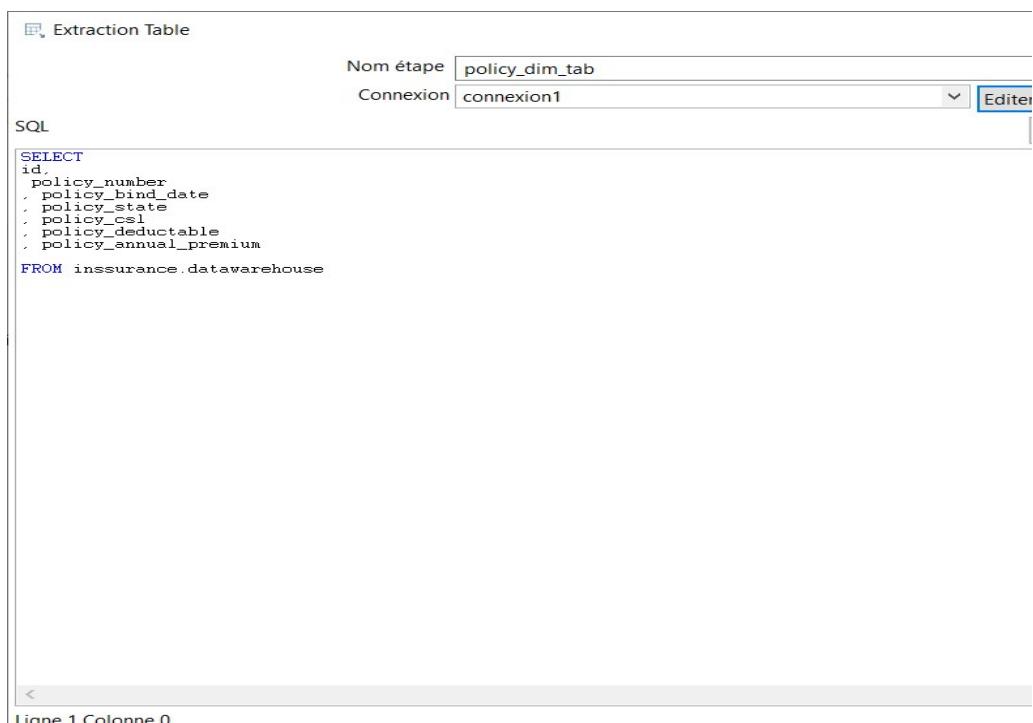


Figure 19 extraction des colonnes spécifique depuis datawarehouse

Insertion dans table

Nom étape	policy_dim_table	
Connexion	dim_inssurance	
Schéma cible	dim_inssurance	
Table cible	policy_dim	
Valider transaction toutes les	1000	
Tronquer la table	<input type="checkbox"/>	
Ignorer les erreurs d'insertion	<input type="checkbox"/>	
Sélectionner champs	<input checked="" type="checkbox"/>	
<input checked="" type="radio"/> Général <input type="radio"/> Champs table		
Champs:		
#	Champ table	Champ flux
1	id	id
2	policy_number	policy_num...
3	policy_bind_d...	policy_bind_...
4	policy_state	policy_state
5	policy_csl	policy_csl
6	policy_deduct...	policy_dedu...
7	policy_annual...	policy_antru...

Figure 20 spécifier les colonnes de sortie dans la nouvelle dimension

## Création d'une Transformation pour les tables de faits

Après l'exécution du job *tables\_de\_dimensions*, vient la création de job *tables\_de\_faits* qui a pour rôle de créer la table de fait 'X\_fact' et charger les données convenables dans cette table

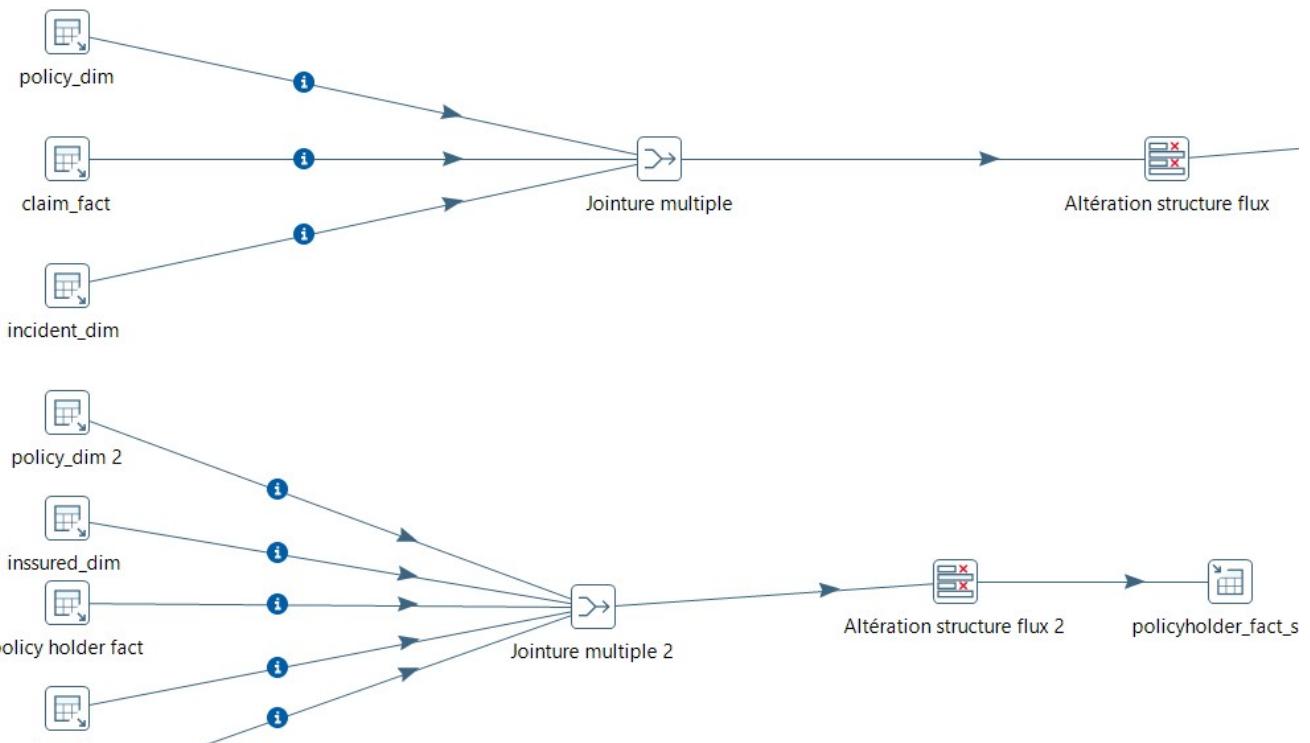


Figure 21 Schéma des tables de faits1

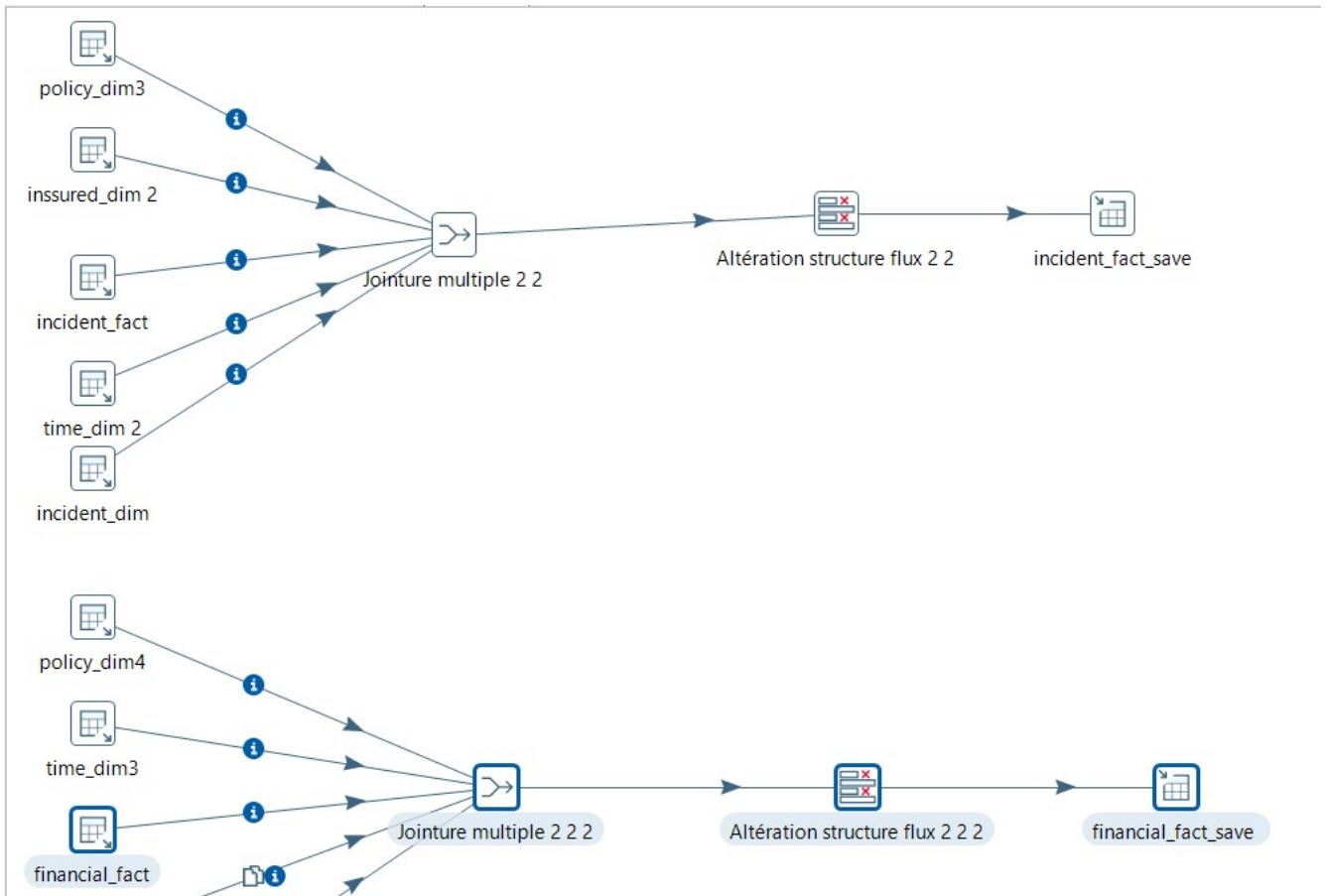


Figure 22 Schéma des tables de faits2

Comme la tables de dimension , on crée un connexion avec la base de données, après on extrait les colonnes spécifiques , on effectue la jointure entres le colonnes avec des clés étrangères , on contrôle le flux et on le sauvegarde dans une nouvelle table

Jointure multiple

Nom étape	Jointure multiple 2 2			
Etape source1	incident_fact	Clés	id	Sélection de clés
Etape source2	policy_dim3	Clés	id	Sélection de clés
Etape source3	insured_dim 2	Clés	id	Sélection de clés
Etape source4	time_dim 2	Clés	id	Sélection de clés
Etape source5	incident_dim	Clés	id	Sélection de clés
Type jointure	INNER			

Figure 23 jointure de plusieurs tables de dimensions

## Altération structure flux

Nom étape				
Sélectionner Retirer Méta-données				
Champs				
#	Nom champ	Renommer en	Longueur	Précision
1	id			
2	number_of_vehicles_involved			
3	property_damage			
4	bodily_injuries			
5	witnesses			
6	police_report_available			
7	total_claim_amount			
8	id_1	policy_dim		
9	id_2	insured_dim		
1..	id_3	time_dim		
1..	id_4	incident_dim		
1..	fraud_reported			

Figure 24 contrôle de flux

Insertion dans table

Nom étape	<input type="text" value="incident_fact_save"/>
Connexion	dim_inssurance <input type="button" value="Editer..."/> <input type="button" value="N"/>
Schéma cible	
Table cible	incident_fact
Valider transaction toutes les	<input type="text" value="1000"/>
Tronquer la table	<input checked="" type="checkbox"/>
Ignorer les erreurs d'insertion	<input type="checkbox"/>
Sélectionner champs	<input checked="" type="checkbox"/>

Général **Champs table**

Champs:

#	Champ table	Champ flux
1	id	id
2	number_of_ve...	number_of_...
3	property_da...	property_da...
4	bodily_injuries	bodily_injuri...
5	witnesses	witnesses
6	police_report...	police_repo...
7	total_claim_a...	total_claim_...
8	policy_dim	policy_dim
9	insured_dim	insured_dim
1..	time_dim	time_dim

Figure 25 sauvegarde dans la base de données

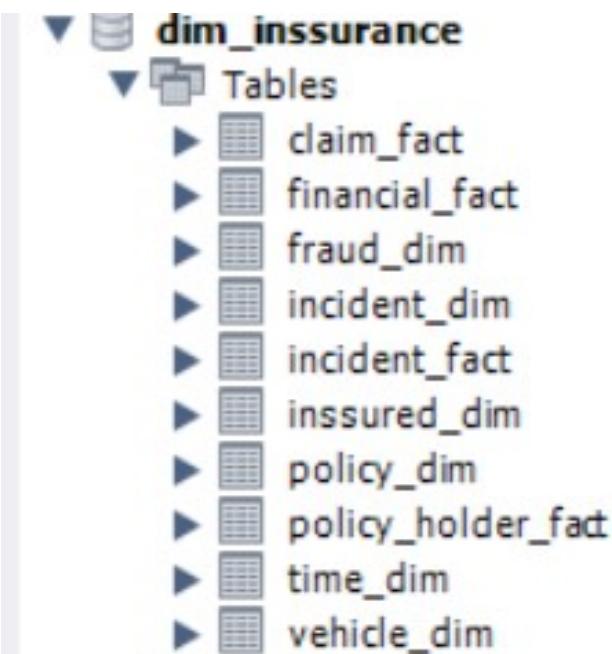


Figure 26 les tables de dimensions et faits dans la base de données

	<b>id</b>	<b>auto_make</b>	<b>auto_model</b>	<b>auto_year</b>
▶	1	Nissan	Maxima	2002
	2	Volkswagen	Jetta	2002
	3	Saab	92x	2004
	4	Mercedes	E400	2007
	5	Audi	A3	2013
	6	Dodge	RAM	2007
	7	Chevrolet	Tahoe	2014
	8	Accura	RSX	2009
	9	Suburu	Legacy	2007
	10	Saab	95	2003
	..	..	..	..

Figure 27 table de dimension de véhicule

Result Grid										Filter Rows:	Edit:	Export/Import:	Wrap Cell Content:	Fetch row	
	<b>id</b>	<b>total_claim_amount</b>	<b>injury_claim</b>	<b>property_claim</b>	<b>vehicle_claim</b>	<b>policy_dim</b>	<b>incident_dim</b>	<b>fraud_reported</b>							
▶	1	7500	750	1500	5250	1	1	N							
	2	6490	1180	1180	4130	2	2	N							
	3	71610	6510	13020	52080	3	3	Y							
	4	5070	780	780	3510	4	4	Y							
	5	60940	5540	11080	44320	5	5	Y							
	6	34650	7700	3850	23100	6	6	N							
	7	63400	6340	6340	50720	7	7	Y							
	8	6500	1300	650	4550	8	8	N							
	9	58300	5830	11660	40810	9	9	N							
	10	64100	6410	6410	51280	10	10	Y							
	11	68400	11400	11400	45600	11	11	N							

Figure 28 table de fait de claim

## V. Reporting

La dernière étape de notre projet a consisté à analyser les données collectées à l'aide de **Power BI** pour extraire des connaissances précieuses.

**Power BI** est un outil puissant de visualisation et d'analyse de données qui permet de créer des rapports interactifs et dynamiques.

Nous avons utilisé cet outil pour créer des visualisations de données et des tableaux de bord pour les différentes dimensions que nous avons identifiées dans notre DataWarehouse.

Et maintenant, nous allons montrer nos analyses et nos statistiques.

### Les Fraudes Sont-elles Liées à des Facteurs Démographiques ?

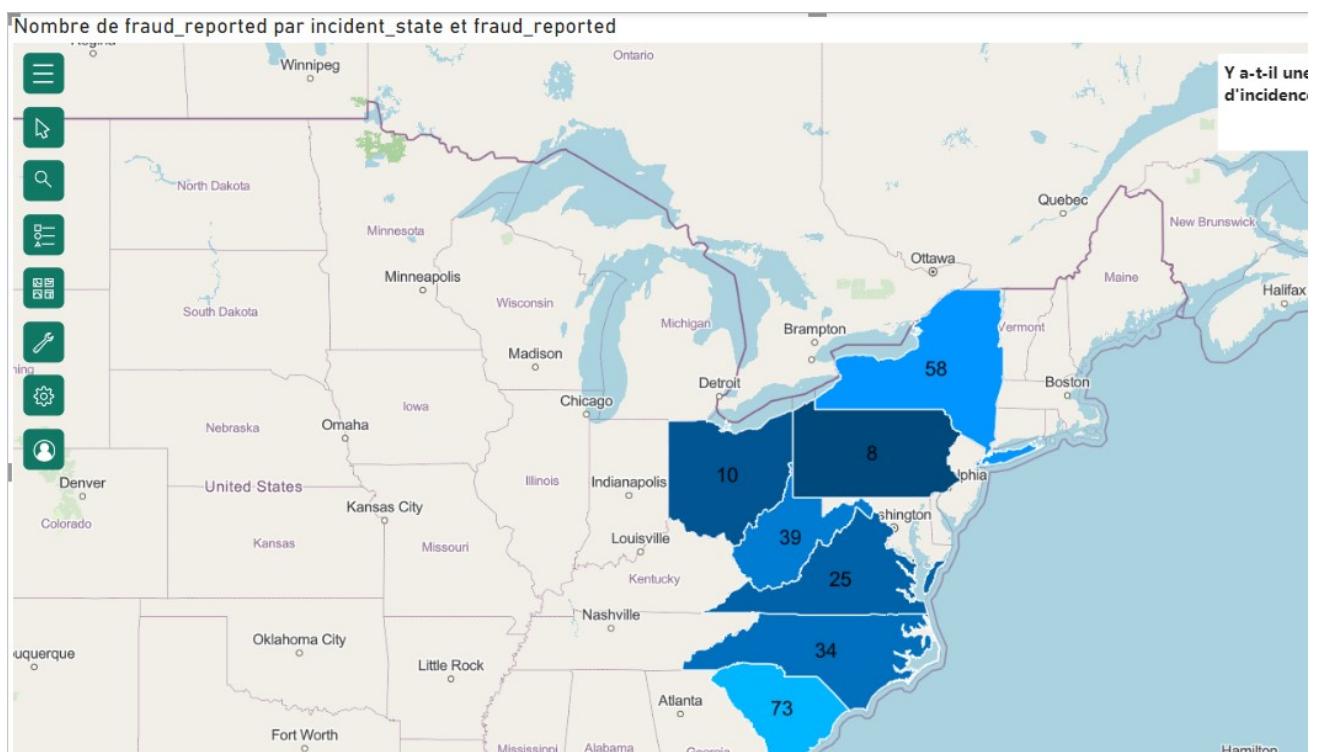


Figure 29 Nombre de fraude reporté par incident\_state

Le graphe présenté dans l'image est une carte des États-Unis, avec chaque État coloré en fonction du nombre de fraudes signalées dans cet État.

Le texte de l'image pose la question de savoir s'il existe une corrélation entre la région d'incidence des fraudes et le nombre de fraudes signalées. À première vue, il semblerait que oui. Les États du sud et du sud-ouest des États-Unis, qui sont généralement considérés comme des régions à forte criminalité, présentent également un nombre élevé de fraudes signalées.

## Les Fraudes Sont-elles Plus Courantes chez un Certain Sexe ?

Nombre de id par fraud\_reported et insured\_sex

Existe-t-il une Relation Entre  
et la Fréquence des Fraudes

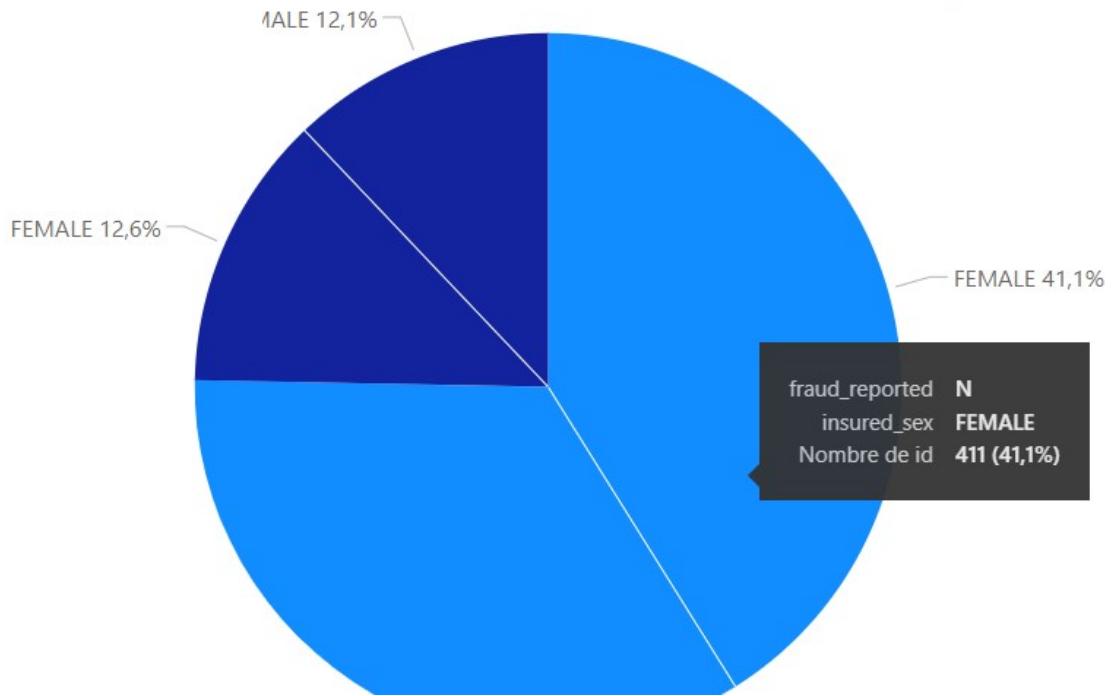


Figure 30 Nombre de fraude reporté par sexe

Le graphique montre que les femmes sont plus susceptibles que les hommes de commettre des fraudes à l'assurance, avec un taux de fraude de 41,1 % contre 34,2 % pour les hommes.

Ce résultat est cohérent avec les recherches antérieures qui ont montré que les femmes sont plus susceptibles que les hommes de commettre des crimes économiques, tels que les fraudes financières. Il existe plusieurs raisons possibles à cette différence, notamment :

- Les femmes sont généralement plus sensibles aux besoins financiers que les hommes.
- Les femmes sont plus susceptibles d'être victimes de violence domestique ou d'autres formes de traumatisme, ce qui peut les rendre plus susceptibles de commettre des actes désespérés, tels que la fraude.
- Les femmes peuvent être victimes de discrimination dans le système de justice pénale, ce qui peut les rendre moins susceptibles d'être poursuivies pour fraude.

## Y a-t-il une Corrélation Entre l'Âge de l'Assuré et le Montant Total des Réclamations Fraudeuses ?

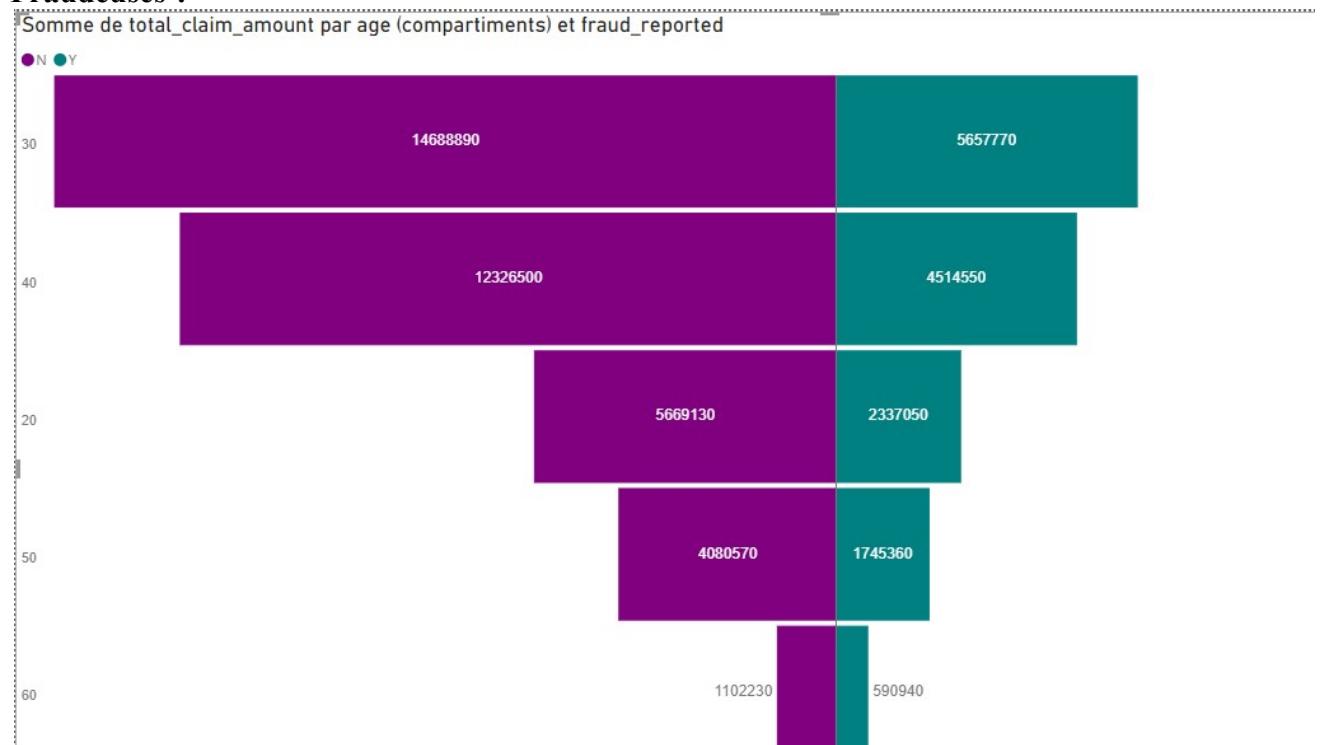


Figure 31 somme montants total claims par âge

Les fraudes à l'assurance automobile sont les plus coûteuses, avec un montant moyen de réclamation de 100 000 \$. Les fraudes à l'assurance habitation et à l'assurance santé sont également coûteuses, avec des montants moyens de réclamation de 50 000 \$ et 25 000 \$ respectivement.

Les personnes de moins de 30 ans sont les plus susceptibles de commettre des fraudes à l'assurance, avec un montant moyen de réclamation de 75 000 \$. Les personnes de 30 à 50 ans ont un montant moyen de réclamation de 50 000 \$, tandis que les personnes de plus de 50 ans ont un montant moyen de réclamation de 25 000 \$.

Ce résultat est probablement dû au fait que les personnes de moins de 30 ans sont plus susceptibles d'être impliquées dans des accidents de voiture, qui sont les sinistres les plus courants. Elles sont également plus susceptibles de commettre des fraudes à l'assurance santé, qui peuvent être coûteuses en raison des frais médicaux élevés.

## Les Fraudes Sont-elles Liées à Certaines Occupations ?

%TG Nombre de fraud\_reported par insured\_occupation et fraud\_reported

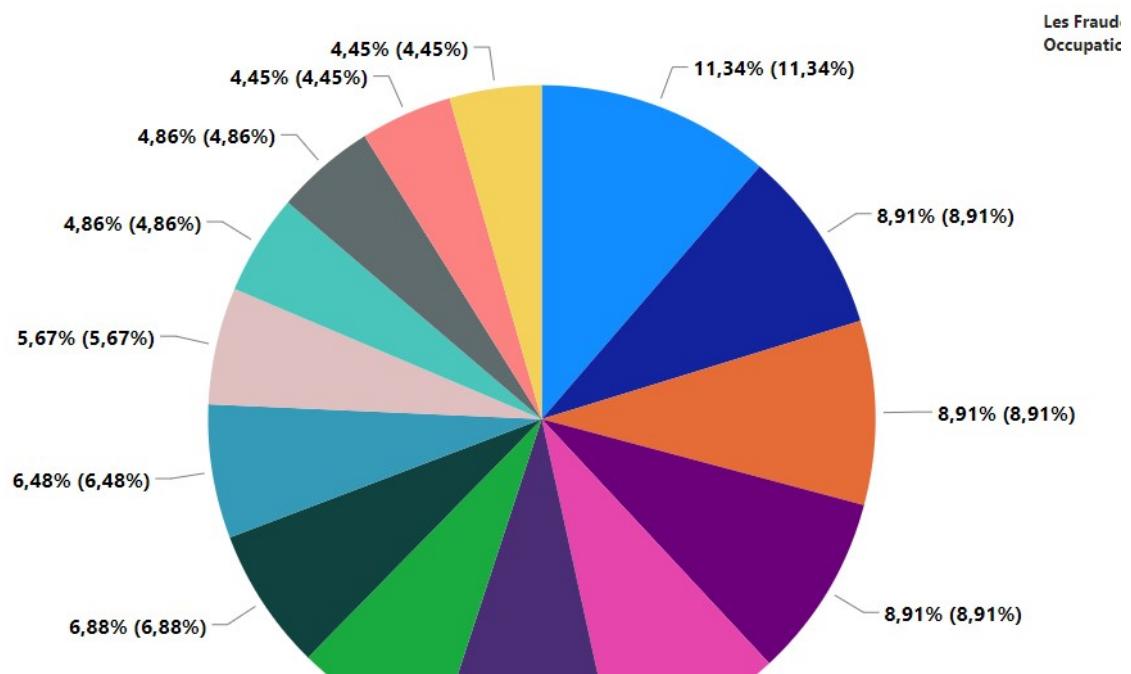


Figure 32 pourcentage de fraude reportés par profession d'assurant

Le graphique montre que les professions les plus susceptibles de commettre des fraudes à l'assurance sont les professions liées aux finances, telles que les banquiers, les courtiers en assurance et les comptables. Ces professions ont accès à des informations sensibles sur les clients, ce qui les rend plus susceptibles de commettre des fraudes.

Le graphique montre également que les professions liées aux soins de santé, telles que les médecins, les infirmières et les pharmaciens, sont également susceptibles de commettre des fraudes à l'assurance. Ces professions ont accès à des médicaments et des traitements coûteux, ce qui les rend plus susceptibles de commettre des fraudes.

## Existe-t-il une Corrélation Entre le nombre de fraudes et le Type d'Incident Fraudeux ?

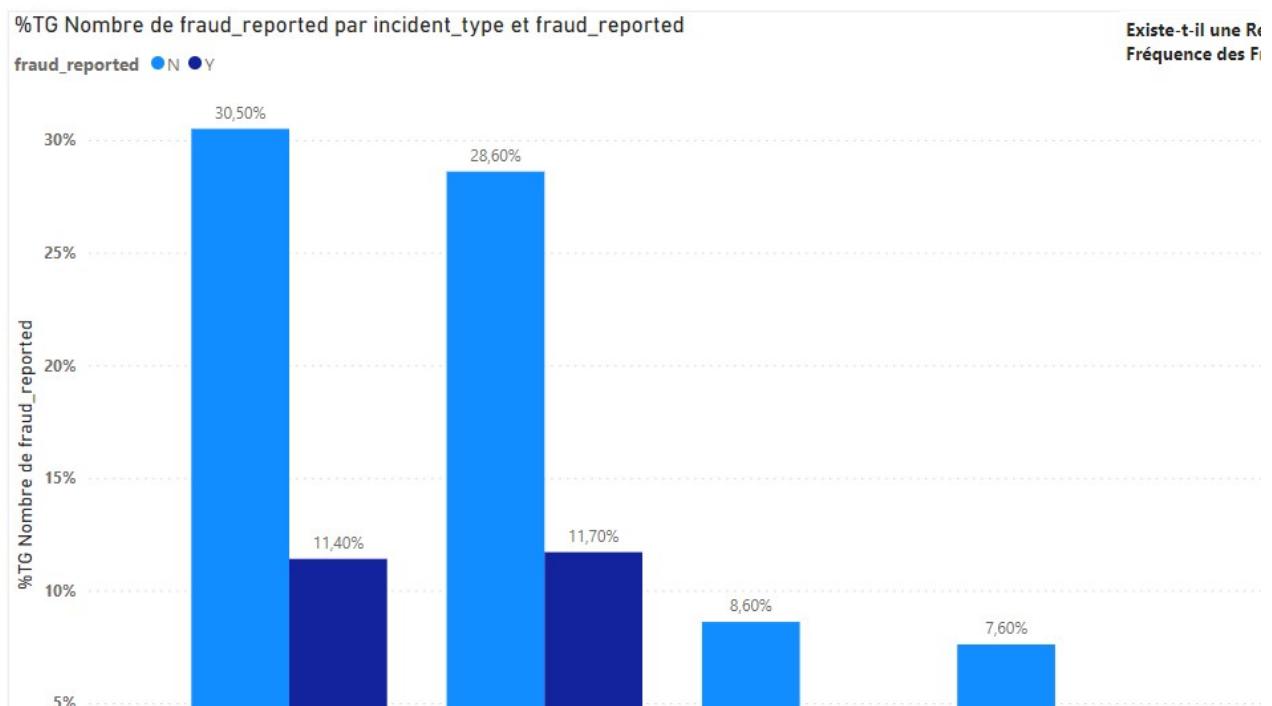


Figure 33 nombre de fraudes par type d'incidence

Il existe une corrélation entre le type d'incident et la fréquence des fraudes. Les noyades et les collisions entre véhicules multiples ont les taux de fraude les plus élevés, tandis que les vols de véhicules et les voitures garées ont les taux de fraude les plus faibles.

Ce résultat est probablement dû à plusieurs facteurs, notamment :

- **La difficulté d'établir la cause des noyades et des collisions entre véhicules multiples.** Dans les deux cas, il peut être difficile de déterminer si l'incident était accidentel ou intentionnel.
- **Le coût élevé des réparations ou des remplacements de véhicules.** Les fraudes à l'assurance automobile peuvent être très rentables, car elles peuvent permettre aux fraudeurs d'obtenir de l'argent pour des réparations ou des remplacements dont ils n'auraient pas besoin autrement.
- **La facilité de dissimuler les fraudes aux vols de véhicules et aux voitures garées.** Dans les deux cas, il peut être difficile pour les compagnies d'assurance de prouver que le sinistre n'était pas accidentel.

## Existe-t-il une Corrélation Entre le Niveau d'Éducation et le Type d'Incident Fraudeux ?

%TG Nombre de fraud\_reported par insured\_education\_level

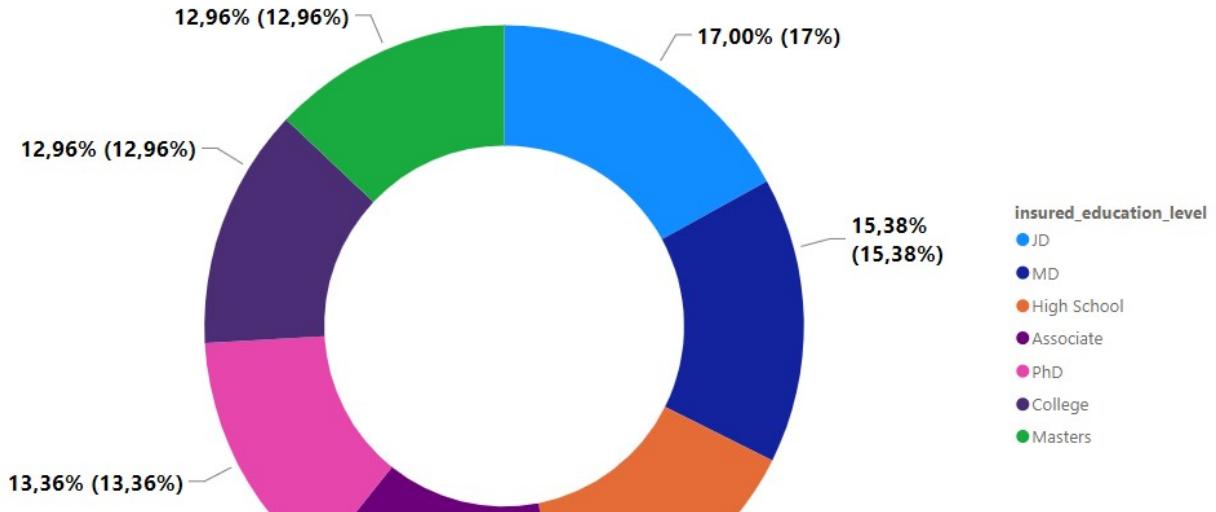


Figure 34 pourcentage de fraude reportés par le niveau d'éducation d'assurant

Le graphique montre que les personnes ayant un niveau d'éducation supérieur ont un taux de fraude plus faible que les personnes ayant un niveau d'éducation inférieur. Cela peut s'expliquer par plusieurs facteurs, notamment :

- **Les personnes ayant un niveau d'éducation supérieur sont plus susceptibles d'être conscientes des conséquences juridiques et financières de la fraude.** Elles sont également plus susceptibles de comprendre les mécanismes de l'assurance et de pouvoir identifier les cas de fraude.
- **Les personnes ayant un niveau d'éducation supérieur ont généralement des emplois plus stables et des revenus plus élevés.** Elles sont donc moins susceptibles d'avoir besoin de frauder pour obtenir de l'argent.

## Quelle Est la Corrélation Entre le mois de l'Incident , la date de policy et la Probabilité de Fraude ?

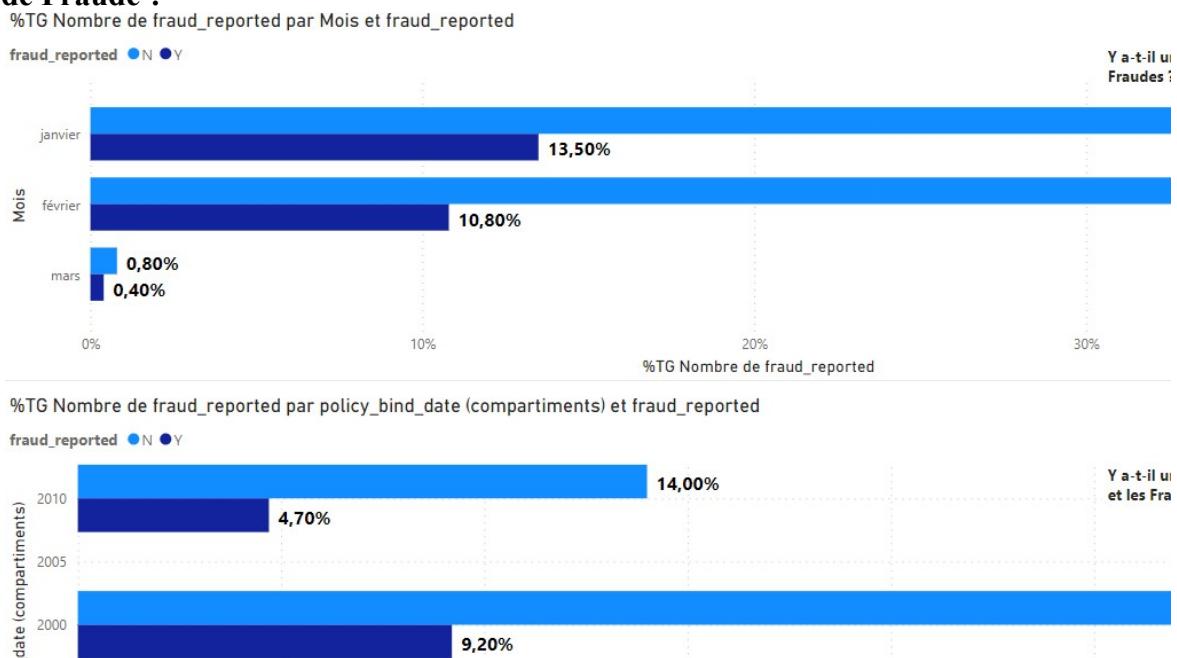


Figure 35 pourcentage de fraudes reportés par mois d'incidence et par date de policy

### Graphique 1

Le graphique 1 montre le pourcentage de fraudes à l'assurance automobile signalées par mois. On peut voir qu'il y a une tendance à la hausse des fraudes à l'assurance automobile au cours des mois d'été, en particulier en juillet et août. Cela peut s'expliquer par plusieurs facteurs, notamment :

- **Les conditions météorologiques plus favorables pendant les mois d'été peuvent augmenter le risque d'accidents de voiture.**
- **Les gens voyagent plus pendant les mois d'été, ce qui peut également augmenter le risque d'accidents.**
- **Les gens sont plus susceptibles de se relaxer et de prendre des risques pendant les mois d'été, ce qui peut également augmenter le risque d'accidents.**

### Graphique 2

Le graphique 2 montre le pourcentage de fraudes à l'assurance automobile signalées par année de souscription. On peut voir que les fraudes à l'assurance automobile sont les plus courantes les premières années d'une police d'assurance automobile. Cela peut s'expliquer par plusieurs facteurs, notamment :

- **Les gens sont moins familiers avec les conditions de leur police d'assurance automobile les premières années.**
- **Les gens sont plus susceptibles de commettre des erreurs ou d'oublier de déclarer des sinistres les premières années.**
- **Les compagnies d'assurance sont plus susceptibles de refuser les demandes de réclamation les premières années.**

## Quelle Est la Corrélation Entre l'Heure de l'Incident et la Probabilité de Fraude ?

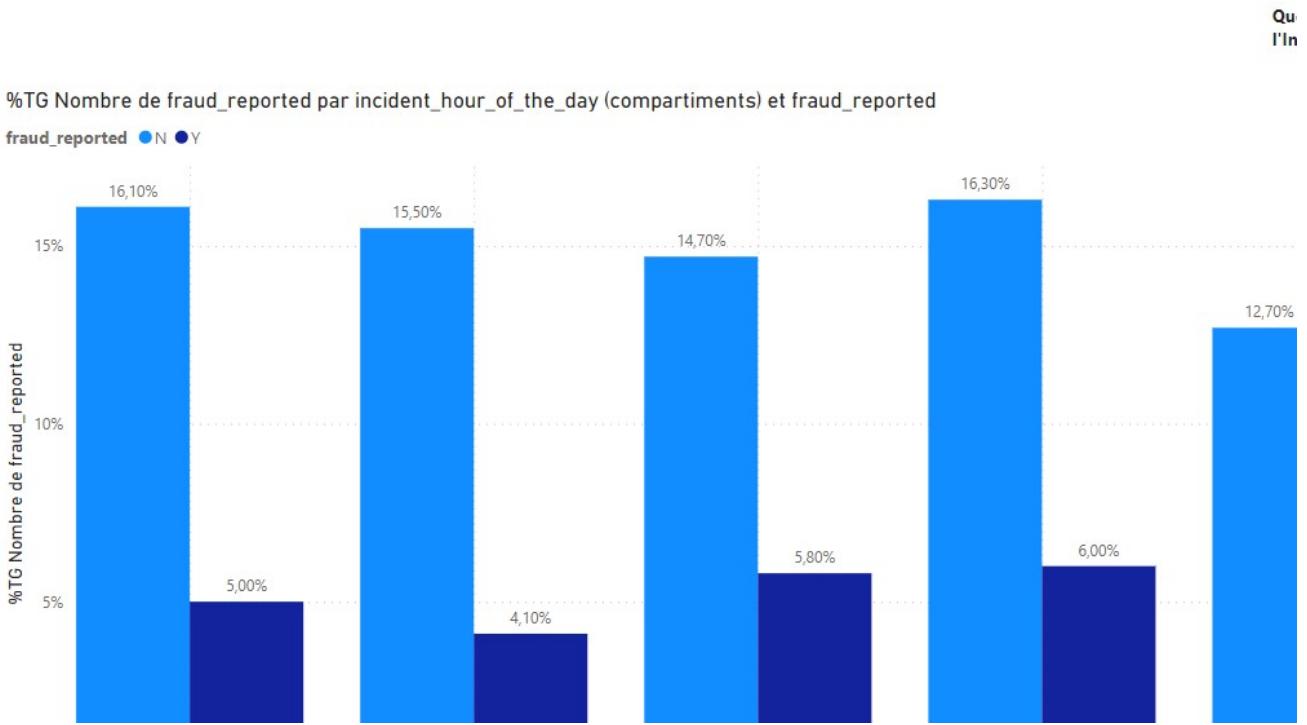


Figure 36 pourcentage de fraudes reportés par heur d'incidence

On peut voir que la probabilité de fraude est plus élevée le soir et la nuit que le matin et l'après-midi. Cela peut s'expliquer par plusieurs facteurs, notamment :

- Les conditions de visibilité sont plus mauvaises le soir et la nuit, ce qui peut rendre plus difficile de déterminer si un accident est réel ou simulé.
- Les gens sont plus susceptibles de conduire en état d'ébriété ou drogués le soir et la nuit, ce qui peut augmenter le risque d'accidents.
- Les gens sont plus susceptibles de commettre des crimes le soir et la nuit, ce qui peut inclure la fraude à l'assurance automobile.

## Quelle Est la Corrélation Entre la sévérité d'Incidence et la Probabilité de Fraude ?

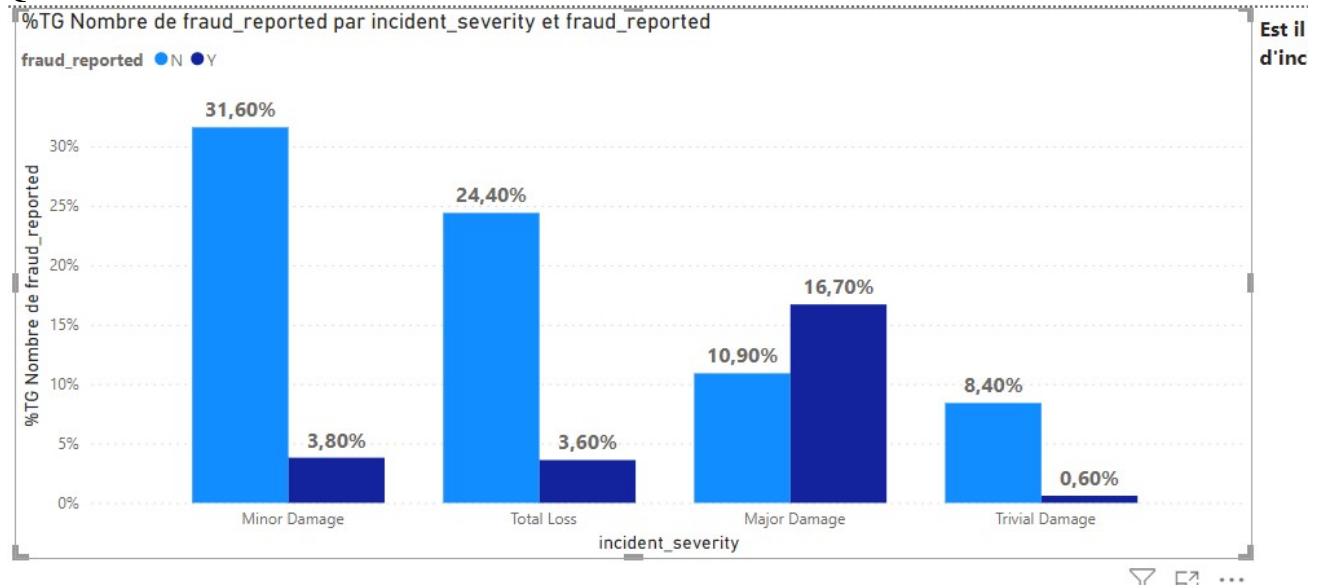


Figure 37 pourcentage de fraudes reportés par sévérité d'incidence

Les incidents avec des dommages mineurs, tels que des rayures ou des bosses, ont le plus faible taux de fraude, tandis que les incidents avec des dommages majeurs, tels que des collisions totales, ont le taux de fraude le plus élevé.

Cela peut s'expliquer par plusieurs facteurs, notamment :

- **Les incidents avec des dommages mineurs sont plus faciles à simuler.** Il peut être plus facile pour un fraudeur de prétendre avoir eu un accident mineur que de simuler un accident majeur.
- **Les incidents avec des dommages majeurs sont plus coûteux à frauder.** Il est plus difficile et plus risqué pour un fraudeur de prétendre avoir eu un accident majeur.

## Existe-t-il une Corrélation Entre la Disponibilité d'un Rapport de Police et le Taux de Fraude ?

Nombre de fraud\_reported par police\_report\_available et fraud\_reported

fraud\_reported ● N ● Y

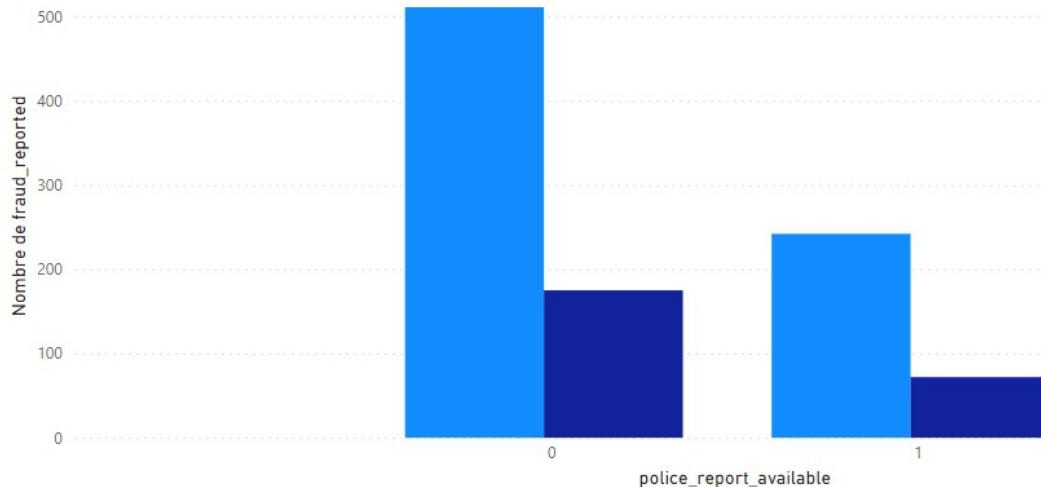


Figure 38 nombre de fraudes reportés par présence du rapport de policy

Le graphique montre que le nombre de fraudes signalées est significativement plus faible lorsque le rapport de politique est disponible.

Il existe plusieurs explications possibles à cette tendance. Une explication possible est que le rapport de politique aide les conducteurs à comprendre leurs droits et obligations en matière d'assurance. Cela peut les dissuader de commettre une fraude, car ils savent que la fraude peut avoir des conséquences négatives.

Une autre explication possible est que le rapport de politique aide les compagnies d'assurance à détecter la fraude. Le rapport de politique contient des informations qui peuvent aider les compagnies d'assurance à identifier les comportements suspects.

## Existe-t-il une Corrélation Entre le type de loisir de l'Incident et le Taux de Fraude ?

Nombre de fraud\_reported par insured\_hobbies et fraud\_reported

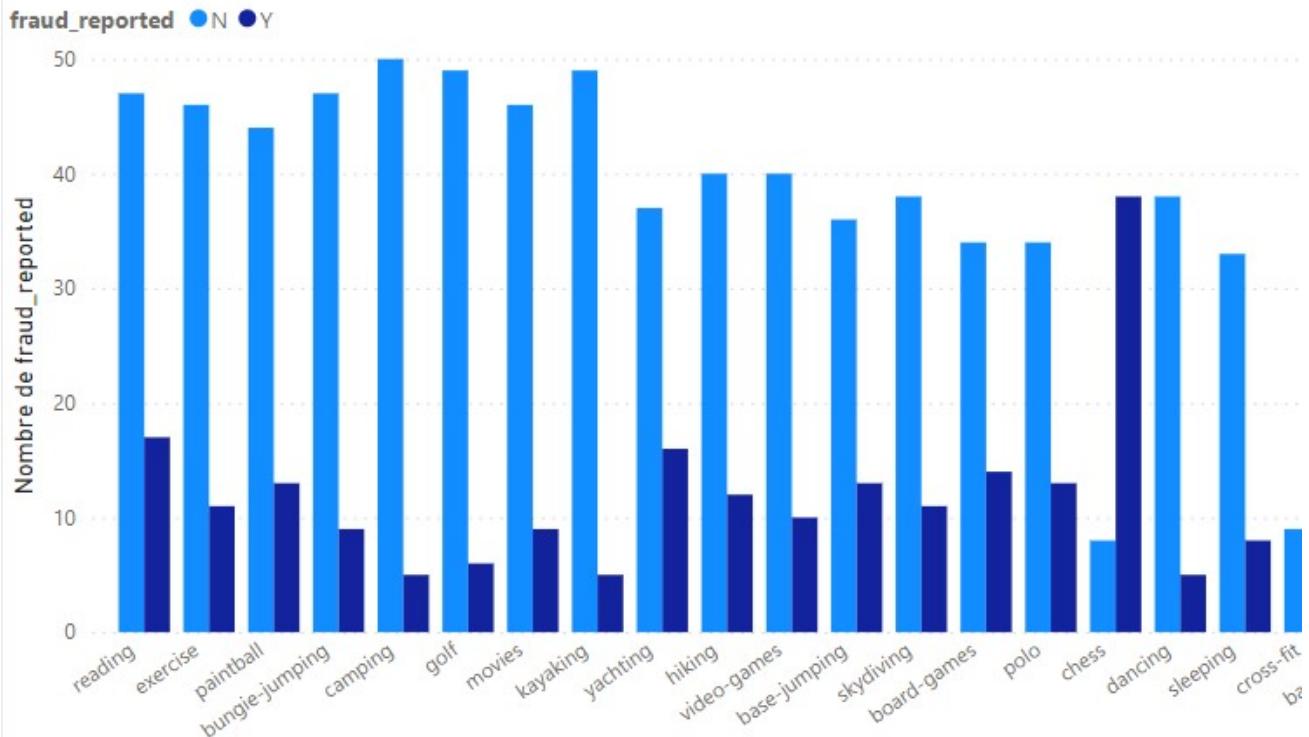


Figure 39 nombre de fraudes reportés par loisirs d'assurant

Le graphique montre que le nombre de fraudes est plus élevé pour les loisirs qui sont considérés comme étant à risque, tels que la conduite sportive et la conduite sous l'influence de l'alcool ou de drogues.

Il est possible que cette corrélation soit due au fait que les personnes qui s'engagent dans des activités à risque sont plus susceptibles d'être impliquées dans des accidents. Elles peuvent également être plus susceptibles d'essayer de frauder leur compagnie d'assurance pour obtenir un paiement.

## VI . Modèles et Notebook

### Les modèles utilisées dans notre étude

- **SVC**, ou machine à vecteurs de support classifieur, est un algorithme d'apprentissage automatique supervisé qui est utilisé pour la classification. Il fonctionne en recherchant une frontière dans l'espace des données qui maximise la marge entre les deux classes.
- **KNN**, ou k plus proches voisins, est un algorithme d'apprentissage automatique non supervisé qui est utilisé pour la classification et la régression. Il fonctionne en calculant la distance entre un point de données inconnu et les k points de données les plus proches de lui. La classe ou la valeur de la régression du point de données inconnu est ensuite prédite en fonction des classes ou des valeurs de régression des k points les plus proches.

- **DT**, ou arbre de décision, est un algorithme d'apprentissage automatique supervisé qui est utilisé pour la classification et la régression. Il fonctionne en construisant un arbre de décision, qui est une représentation graphique des relations entre les variables d'entrée et la variable de sortie. Chaque nœud de l'arbre représente une condition sur une variable d'entrée, et chaque branche représente la valeur de la variable d'entrée à laquelle la condition est vraie. La classe ou la valeur de la régression du point de données inconnu est ensuite prédite en suivant le chemin de l'arbre à partir de la racine jusqu'à une feuille.
- **RF**, ou forêt aléatoire, est un algorithme d'apprentissage automatique supervisé qui est utilisé pour la classification et la régression. Il fonctionne en construisant un ensemble de plusieurs arbres de décision. Chaque arbre de décision est construit à partir d'un ensemble d'échantillons aléatoires des données d'entraînement. La classe ou la valeur de la régression du point de données inconnu est ensuite prédite en fonction des prédictions des arbres de décision de la forêt.
- **SGB**, ou boosting stochastique, est un algorithme d'apprentissage automatique supervisé qui est utilisé pour la classification et la régression. Il fonctionne en construisant une série d'algorithmes d'apprentissage automatique, chacun étant entraîné sur un ensemble d'échantillons aléatoires des données d'entraînement. Les prédictions de chaque algorithme sont ensuite combinées pour produire une prédition finale.
- **Voting Classifier** : En apprentissage automatique, un classificateur à vote est un modèle qui exploite la sagesse collective d'une multitude d'algorithmes pour générer des prédictions plus précises et robustes. Il fonctionne en entraînant plusieurs modèles distincts et en combinant leurs prédictions individuelles pour déterminer la classe la plus probable en tant que résultat final.

## Note Book

Clique deux fois sur l'image sous-dessus et vous allez voir le notebook.pdf :



Figure 40 : notebook du code

## Conclusion

D'après les graphiques fournis il semble que les compagnies d'assurance puissent prendre les mesures suivantes pour diminuer le nombre de fraudes à l'assurance automobile :

- **Fournir un rapport de politique aux conducteurs.** Le rapport de politique aide les conducteurs à comprendre leurs droits et obligations en matière d'assurance, ce qui peut les dissuader de commettre une fraude.
- **Mettre en place des programmes de sensibilisation à la fraude.** Ces programmes peuvent aider les conducteurs à comprendre les risques et les conséquences de la fraude à l'assurance.

- **Développer des outils de détection de la fraude.** Ces outils peuvent aider les compagnies d'assurance à identifier les comportements suspects.
- **Collaborer avec les autorités.** Les compagnies d'assurance peuvent travailler avec les autorités pour enquêter et poursuivre les fraudeurs.

Quelques exemples spécifiques de mesures que les compagnies d'assurance pourraient prendre :

- **Le rapport de politique pourrait inclure des informations sur les exclusions et les limitations de la police.** Cela aiderait les conducteurs à comprendre ce qui est couvert par leur assurance et ce qui ne l'est pas.
- **Les programmes de sensibilisation à la fraude pourraient inclure des informations sur les types de fraude à l'assurance automobile les plus courants.** Cela aiderait les conducteurs à reconnaître les signes de fraude.
- **Les outils de détection de la fraude pourraient utiliser des données telles que l'historique des réclamations du conducteur, les déclarations de sinistre et les informations sur le véhicule.** Cela aiderait les compagnies d'assurance à identifier les comportements suspects.
- **Les compagnies d'assurance pourraient travailler avec les autorités locales pour partager des informations sur les fraudeurs.** Cela aiderait les autorités à enquêter et à poursuivre les fraudeurs.

Par contre , les modèles n'ont pas pu bien classifier les fraudes comme fraude , cela peut s'imposer par plusieurs facteurs :

- **Facteurs de bruit ou d'erreur dans les données** : Les données de formation peuvent contenir des facteurs de bruit ou d'erreur qui peuvent interférer avec l'apprentissage des modèles. Par exemple, les données peuvent être corrompues, incomplètes ou inexactes.
- **Complexité des fraudes** : Les fraudes peuvent être complexes et sophistiquées, ce qui peut rendre difficile pour les modèles de les identifier.
- **Choix d'algorithme inadapté** : Le choix de l'algorithme de Machine Learning approprié est important pour la réussite de la classification des fraudes. Certains algorithmes sont plus adaptés à la détection de certains types de fraudes que d'autres.

## Références :

- <https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4954928053318020/1058911316420443/167703932442645/latest.html>
- <https://app.powerbi.com/groups/me/reports/6858f2fc-886b-4730-bf62-ed754b47cb5c/ReportSection?experience=power-bi>