Mini projet de base de données:
Comment le sexe et la localisation géographique influencent-ils l'incidence et la manifestation des pathologies ?

Réalisé par:
"LES ALGORYTHMES CEREBRAUX"
Casin Océane
Hakiri Siwar
Samuilava Yelizaveta
Uzan Benjamin-Haroun

Encadré par:

Mme BRINGAY Sandra
Mme BEN-SASSI Imen
Marine DEMANGEOT

Table des matières:

Contents

Préface

Déclaration de non-plagiat

Nous déclarons que ce rapport est le fruit de notre seul travail, à part lorsque cela est indiqué explicitement.

Nous acceptons que la personne évaluant ce rapport puisse, pour les besoins de cette évaluation:

- la reproduire et en fournir une copie à un autre membre de l'université; et/ou,
- en communiquer une copie à un service en ligne de détection de plagiat (qui pourra en retenir une copie pour les besoins d'évaluation future).

Nous certifions que nous avons lu et compris les règles ci-dessus.

En signant cette déclaration, nous acceptons ce qui précède.

Signature: Les Algorythmes cérébraux

Date: _____

Chapitre 1

Sujet

Comment le sexe et la localisation géographique influencent-ils l'incidence et la manifestation des pathologies ?

Introduction

Pour répondre au sujet, nous avons choisi 4 bases de données :

1. Base de données 1 (EFFECTIF) : Effectifs associés aux pathologies concernant la population française en fonction de la région par sexe, concernant l'année 2021, selon la Sécurité Sociale.
   Disponible à https://data.ameli.fr/explore/dataset/effectifs/table/?refine.annee=2021&refine.cla_age_5=tsage&refine.niveau_prioritaire=1.

2. Base de données 2 (DÉPENSES) : Dépenses remboursées par la sécurité sociale par pathologies, selon la Sécurité Sociale.
   Disponible à https://data.ameli.fr/explore/dataset/depenses/table/?refine.annee=2021&refine.niveau_prioritaire=1&exclude.patho_niv1=Total+consommants+tous+r%C3%A9gimes&refine.type_somme=Total&exclude.dep_niv_1=D%C3%A9penses.

3. Base de données 3 (COMORBIDITÉS) : Comorbidités associées aux pathologies.
   Disponible à https://data.ameli.fr/explore/dataset/comorbidites/table/?refine.annee=2021.

4. Base de données 4 (DEPARTEMENT): Département associés à leur région ainsi qu'à la population de celui-ci.
   Disponible à https://www.insee.fr/fr/statistiques/7739582?sommaire=7728826

Objectifs

À l'aide des bases de données, nous allons mettre en place des requêtes SQL afin de mettre en évidence les déterminants de santé ainsi que les tendances pathologiques en fonction du sexe et/ou du département. Nous montrerons ainsi l'importance de prendre en compte certaines informations afin d'élaborer notre politique de soin.

Chapitre 2

Tri des bases de données et création des tables

1. Récupération des données

   La première étape du traitement des données a été de trier les données dans l'espace et le temps. Nous avons donc choisi de nous concentrer sur l'année 2021, sur les pathologies de niveau 1 ainsi que sur les comorbidités de niveau 2.
   Cependant, les 4 données originales ne permettaient ni de répondre à la problématique, ni de correspondre avec notre modèle. Par conséquent, il a été nécessaire de trier et d'adapter ces données.

2. Tri des données

   Afin d'adapter les données à notre problématique, nous avons tout d'abord filtré et supprimé les variables inutiles au projet, telles que le niveau 3 des comorbidités, les niveaux 2 et 3 des pathologies, les variables constantes (telles que l'année toujours égale à 2021), ainsi que les lignes possédant des données manquantes ou incomplètes.
   Ensuite, les données ont été réparties dans différents fichiers afin de les faire correspondre à notre modèle.
   Pour cela, nous avons créé un fichier par table SQL (voir partie...) afin de faciliter l'importation de celles-ci.

3. Importation

   L'importation des données à été effectuée via 6 tables sur phpMyAdmin :

   (a) dep

	#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
☐	1	**id_region**	int			Oui	*NULL*			Modifier Supprimer Plus
☐	2	**region**	varchar(25)	utf8mb3_general_ci		Oui	*NULL*			Modifier Supprimer Plus
☐	3	**id_dep**	varchar(3)	utf8mb3_general_ci		Non	*Aucun(e)*			Modifier Supprimer Plus
☐	4	**dep**	varchar(23)	utf8mb3_general_ci		Oui	*NULL*			Modifier Supprimer Plus
☐	5	**ptot**	int			Oui	*NULL*			Modifier Supprimer Plus

Figure 1: variable de la table dep

A chaque département est associé un identifiant ainsi qu'une région. La région est identifié par un identifiant ainsi qu'un libellé. Ci-dessous la table desciptive des variables de la table.

Nom	Clef	Description
id_dep	Primaire	Identifiant attribuée à un département (ex : 01 = Ains)
dep		Nom du département
id_region		Identifiant attribuée à la une région d'un département
region		Nom de la région attribuée à un département
ptot		Population total d'un département

Il a été nécessaire de rajouter un département 99 non présent initialement afin de gérer les données des pathologies nationales (somme de tous les départements)

(b) sexe

#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
1	**id_sexe**	int			Non	*Aucun(e)*			Modifier Supprimer Plus
2	**libelle_sexe**	varchar(9)	utf8mb3_general_ci		Oui	*NULL*			Modifier Supprimer Plus

Il existe 3 sexe (**Homme, Femme et Non connu**). A chaque sexe est attribué un numéro permettant d'identifier le sexe dans les autres tables. Ci-dessous la table desciptive des variables de la table.

Nom	Clef	Description
id_sexe	Primaire	Numéro Attribuée à un sexe **(1,2 ou 9)**
libelle_sexe	Unique	Description du sexe **(Homme, Femme ou Non connu)**

(c) patho_un

#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
1	**patho_niv1**	varchar(116)	utf8mb3_general_ci		Oui	*NULL*			Modifier Supprimer Plus
2	**top**	varchar(21)	utf8mb3_general_ci		Non	*Aucun(e)*			Modifier Supprimer Plus
3	**montant**	bigint			Non	*Aucun(e)*			Modifier Supprimer Plus

A chaque pathologie est associé un identifiant (**top**) ainsi qu'un nom et un effectif de patient. Ci-dessous la table desciptive des valeurs de la table.

Nom	Clef	Description
top	Primaire	Identifiant de la pathologie
patho_niv1	Unique	Nom de la pathologie
montant		Effectif de patients traité par la pathologie *(0 signifiant pas de renseignement)*

(d) patho_deux

#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
1	**comorbidite**	varchar(16)	utf8mb3_general_ci		Non	*Aucun(e)*			Modifier Supprimer Plus
2	**patho_niv2_comorb**	varchar(73)	utf8mb3_general_ci		Oui	*NULL*			Modifier Supprimer Plus
3	**patho_niv1_comorb**	varchar(71)	utf8mb3_general_ci		Oui	*NULL*			Modifier Supprimer Plus

A chaque comorbidité est associé un identifiant (**comorbidite**), un nom ainsi qu'un groupe de pathologie plus large. Ainsi chaque groupe primaire est sous divisée en 56 pathologies secondaires plus spécifiques *(voir schema 1)*. Ci-dessous la table desciptive des valeurs de la table.

Nom	Clef	Description
comorbidite	Primaire	Identifiant de la comorbidité
patho_niv2_comorb	Unique	Nom de la comorbidité secondaire
patho_niv1_comorb		Nom de la comorbidité primaire

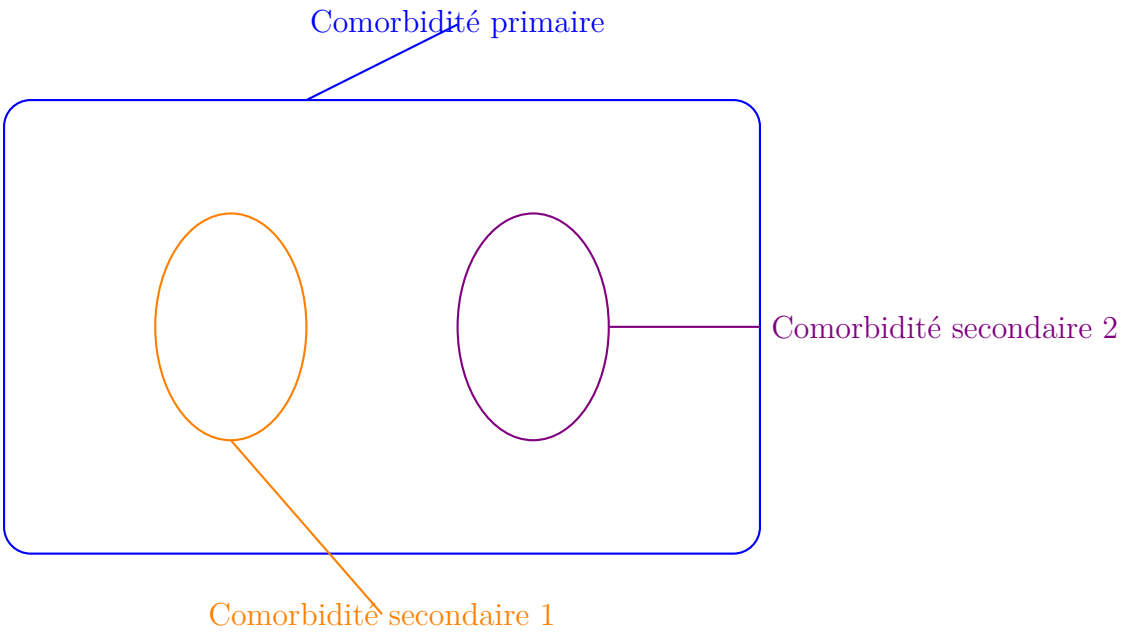


Figure 2: figure descriptive de l'organisation des pathologies

(e) etre_comorbidite

	#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
☐	1	**top**	varchar(17)	utf8mb3_general_ci		Oui	*NULL*			Modifier Supprimer Plus
☐	2	**comorbidite**	varchar(16)	utf8mb3_general_ci		Oui	*NULL*			Modifier Supprimer Plus
☐	3	**Ncomorb**	int			Oui	*NULL*			Modifier Supprimer Plus
☐	4	**Ntop**	int			Oui	*NULL*			Modifier Supprimer Plus
☐	5	**Proportion_comorb**	decimal(7,6)			Oui	*NULL*			Modifier Supprimer Plus
☐	6	**Niveau prioritaire**	int			Oui	*NULL*			Modifier Supprimer Plus

La table *petre_comorbidite* sert de liaison entre les tables c et d. Elle comprend les Effectifs, Proportions ainsi que le niveau de priorité de chaque pathologie associé à une comorbidité. Ci-dessous la table desciptive des valeurs de la table.

Nom	Clef	Description
top	Unique	Liaison vers la table c
comorbidite	Unique	Liaison vers la table d
Ncomorb		Effectif de patients pris en charge pour la comorbidité associée
Ntop		Effectif de patients pris en charge pour la pathologie
Ncomorb		Proportion de patients pris en charge pour la comorbidité par rapport à l'effectif de patients pour la pathologie dont il est question
Niveau prioritaire		Niveau de priorité de soins associé à une comorbidité'*(de 1 à 3 avec 3 le plus élevée)*

(f) exister

	#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra	Action
☐	1	**sexe**	int			Oui	*NULL*			Modifier Supprimer Plus
☐	2	**dep**	varchar(2)	utf8mb3_general_ci		Oui	*NULL*			Modifier Supprimer Plus
☐	3	**top**	varchar(21)	utf8mb3_general_ci		Oui	*NULL*			Modifier Supprimer Plus
☐	4	**prev**	float			Non	*Aucun(e)*			Modifier Supprimer Plus

La table *exister* sert de liaison entre les tables a, b et c. Elle permet d'associer à chaque pathologie une prévalence en fonction du sexe ainsi que du département. Ci-dessous la table desciptive des valeurs de la table.

Nom	Clef	Description
sexe	Index	Liaison vers la table b
dep	Index	Laision vers la table a
top	Index	Laison vers la table c
prev		Prévalence de la pathologie en fonction du sexe et du département

MCD

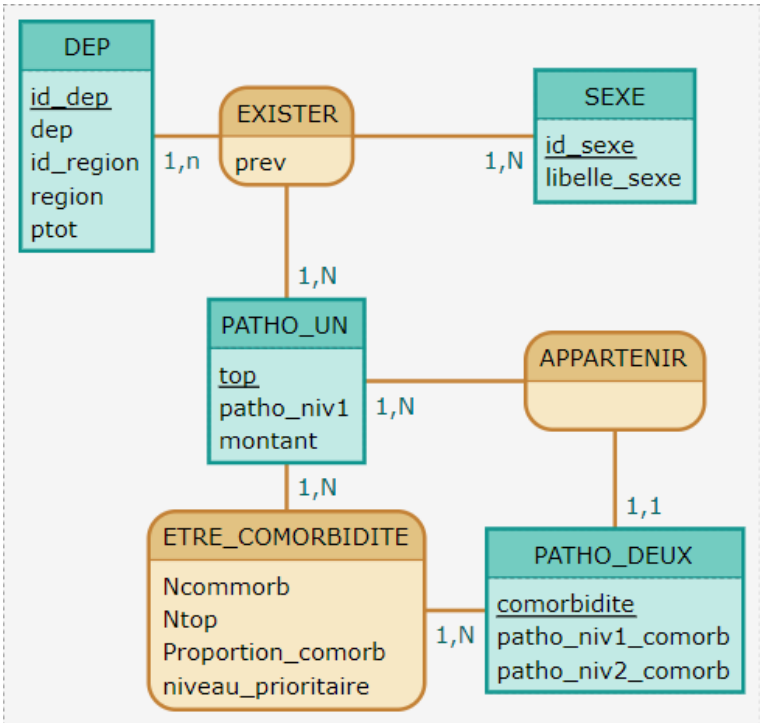


Figure 3: MCD

MOD

DEP (id_dep, dep, id_region, region, ptot)

SEXE (id_sexe, libelle_sexe)

PATHO_UN (top, patho_niv1, montant)

EXISTER (id_dep, id_sexe,top, prev)

ETRE_COMMORBIDITE (top, comorbidite, Ncommorb, Ntop, Proportion_commob, niveau_prioritaire)

PATHO_DEUX (comorbidite, top, patho_niv1_comorb, patho_niv2_comorb)

Chapitre 3

Requêtes et analyse

- Pathologie avec le plus haut montant de remboursement

```
SELECT DISTINCT patho_un.patho_niv1,patho_un.montant,etre_comorbidite.Niveau_prioritaire FROM patho_un,etre_comorbidite,patho_deux where
patho_un.montant in( SELECT max(patho_un.montant) FROM patho_un,etre_comorbidite,patho_deux WHERE patho_un.top=etre_comorbidite.top and
etre_comorbidite.comorbidite=patho_deux.comorbidite);
```

Profilage [Éditer en ligne] [Éditer] [Expliquer SQL] [Créer le code source PHP] [Actualiser]

Tout afficher | Nombre de lignes : 25 Filtrer les lignes: Chercher dans cette table

Options supplémentaires

patho_niv1	montant	Niveau_prioritaire
Cancers	12565883308	1

On peut voir ici que la pathologie ayant le montant de remboursement le plus élevé est le cancer (tous types confondus).

- La prévalence des pathologies par département pour les femmes

```
SELECT patho_un.patho_niv1, dep.dep, sexe.libelle_sexe, exister.prev, AVG(exister.prev) as Moyenne_prev from exister,patho_un,sexe, dep WHERE
patho_un.top=exister.top and sexe.id_sexe=exister.sexe and exister.dep=dep.id_dep and sexe.id_sexe=2 and patho_un.patho_niv1 not like "%pas%" group
by patho_un.patho_niv1, dep.dep ORDER by Moyenne_prev DESC;
```

Profilage [Éditer en ligne] [Éditer] [Expliquer SQL] [Créer le code source PHP] [Actualiser]

1 > >> | Tout afficher | Nombre de lignes : 25 Filtrer les lignes: Chercher dans cette table

Options supplémentaires

patho_niv1	dep	libelle_sexe	prev	Moyenne_prev ▾ 1
Traitement antalgique ou anti-inflammatoire (hors ...	Alpes-de-Haute-Provence	Femme	4.143	4.14300012588501
Traitements psychotropes (hors pathologies)	Eure-et-Loir	Femme	3.581	3.5810000896453857
Traitements psychotropes (hors pathologies)	Eure	Femme	3.193	3.193000078201294
Maladies psychiatriques	Eure-et-Loir	Femme	3.061	3.061000108718872
Maladies psychiatriques	Mayenne	Femme	3.026	3.0260000228881836
Traitements psychotropes (hors pathologies)	Loire-Atlantique	Femme	2.887	2.8870000392334
Traitements psychotropes (hors pathologies)	Gers	Femme	2.886	2.885999917984009
Maladies psychiatriques	Paris	Femme	2.863	2.86299991607666
Traitements psychotropes (hors pathologies)	Paris	Femme	2.849	2.8489999771118164
rs pathologies)	Dordogne	Femme	2.811	2.811000108718872

Console de requêtes SQL

On voit que le département le plus touché par une pathologie pour les femmes sont les Alpes-de-haute-provence avec les traitements antalgiques ou anti-inflammatoires à hauteur de 4,14%.

- Pathologies pour les hommes par département par prévalence décroissante

```
SELECT patho_un.patho_niv1, dep.dep, sexe.libelle_sexe, exister.prev, AVG(exister.prev) as Moyenne_prev from exister,patho_un,sexe, dep WHERE
patho_un.top=exister.top and sexe.id_sexe=exister.sexe and exister.dep=dep.id_dep and sexe.id_sexe=1 and patho_un.patho_niv1 not like "%pas%" group
by patho_un.patho_niv1, dep.dep ORDER by Moyenne_prev DESC;
```

Profilage [Éditer en ligne] [Éditer] [Expliquer SQL] [Créer le code source PHP] [Actualiser]

1 > >> | Tout afficher | Nombre de lignes : 25 | Filtrer les lignes: Chercher dans cette table

Options supplémentaires

patho_niv1	dep	libelle_sexe	prev	Moyenne_prev ▾ 1
Maladies psychiatriques	Mayenne	Homme	3.046	3.0460000038146973
Maladies psychiatriques	Ain	Homme	2.919	2.9189999103546143
Maladies psychiatriques	Eure-et-Loir	Homme	2.911	2.9110000133514404
Maladies psychiatriques	Paris	Homme	2.878	2.878000020980835
Maladies psychiatriques	Seine-Maritime	Homme	2.759	2.759000062942505
Maladies psychiatriques	Alpes-de-Haute-Provence	Homme	2.671	2.6710000038146973
Maladies psychiatriques	Val-de-Marne	Homme	2.576	2.5759999752044678
Maladies psychiatriques	Aisne	Homme	2.576	2.5759999752044678
Maladies psychiatriques	Autre	Homme	2.544	2.5439999103546143
Maladies psychiatriques	Haute-Marne	Homme	2.535	2.5350000858306885
Maladies psychiatriques	Eure	Homme	2.522	2.5220000743865967
Maladies psychiatriques	Vaucluse	Homme	2.497	2.4969999790191650

Console de requêtes SQL

Pour les hommes, le département le plus touché par une pathologie est la Mayenne. On voit également que les maladies psychiatriques sont sur-représentées chez les hommes(pathologie prépondérante).

- Proportion de commorbidités

```
SELECT DISTINCT patho_deux.patho_niv1_comorb,patho_deux.patho_niv2_comorb,etre_comorbidite.Proportion_comorb*100 as proportion_comor FROM
patho_deux,etre_comorbidite WHERE etre_comorbidite.comorbidite=patho_deux.comorbidite group by (patho_deux.patho_niv2_comorb) order by
patho_deux.patho_niv1_comorb ASC;
```

Profilage [Éditer en ligne] [Éditer] [Expliquer SQL] [Créer le code source PHP] [Actualiser]

1 > >> | Tout afficher | Nombre de lignes : 25 | Filtrer les lignes: Chercher dans cette table

Options supplémentaires

patho_niv1_comorb ▲ 1	patho_niv2_comorb	proportion_comor
Cancers	Autres cancers	2.941700
Cancers	Cancer colorectal	0.456700
Cancers	Cancer du poumon	0.357600
Cancers	Cancer de la prostate	1.282500
Cancers	Cancer du sein de la femme	0.426800
Diabète	Diabète	11.274800
Insuffisance rénale chronique terminale	Dialyse chronique	0.418700
Insuffisance rénale chronique terminale	Transplantation rénale	0.024400
Insuffisance rénale chronique terminale	Suivi de transplantation rénale	0.229100
Maladies cardio-neurovasculaires	Artériopathie oblitérante du membre inférieur	3.798100

On voit ici que la maladie impliquant le plus de commorbidités est le diabète impliquant lui-même le diabète. (Par exemple le diabète de type I peut impliquer celui de type II et inversement).

- Département associé à sa région ayant la plus haute population

```
SELECT id_dep,dep,id_region,region,ptot from dep where ptot=(SELECT max(ptot) FROM dep WHERE id_dep<>99);
```

Profilage [Éditer en ligne] [Éditer] [Expliquer SQL] [Créer le code source PHP] [Actualiser]

Tout afficher | Nombre de lignes : 25 Filtrer les lignes: Chercher dans cette table

Options supplémentaires

	id_dep	dep	id_region	region	ptot
Éditer Copier Supprimer	59	Nord	32	Hauts-de-France	2641207

- Pathologie par prévalence décroissante pour les hommes dans le région ayant la plus haute population

```
SELECT patho_un.top,patho_un.patho_niv1,exister.prev, dep.region from exister,dep,patho_un,sexe WHERE exister.dep=dep.id_dep and
exister.top=patho_un.top and exister.sexe=sexe.id_sexe AND dep.region like"%Hauts%" and patho_un.patho_niv1 not like"%pas%" AND sexe.id_sexe=1 ORDER
by exister.prev DESC;
```

Profilage [Éditer en ligne] [Éditer] [Expliquer SQL] [Créer le code source PHP] [Actualiser]

Tout afficher | Nombre de lignes : 25 Filtrer les lignes: Chercher dans cette table

Options supplémentaires

top	patho_niv1	prev 1	region
sup_PsyPat_cat	Maladies psychiatriques	2.576	Hauts-de-France
sup_NeuDeg_cat	Maladies neurologiques ou dégénératives	1.101	Hauts-de-France
sup_InfRarVIH_cat	Maladies inflammatoires ou rares ou VIH ou SIDA	0.919	Hauts-de-France
sup_PsyMed_cat	Traitements psychotropes (hors pathologies)	0.695	Hauts-de-France
sup_Cv_cat	Maladies cardio-neurovasculaires	0.502	Hauts-de-France
sup_FRV_cat	Traitements du risque vasculaire (hors pathologies...	0.246	Hauts-de-France
sup_Arthros_med	Traitement antalgique ou anti-inflammatoire (hors ...	0.214	Hauts-de-France
sup_Can_cat	Cancers	0.171	Hauts-de-France
sup_RIRCT_cat	Insuffisance rénale chronique terminale	0	Hauts-de-France

On voit que dans la région la plus peuplée de France (Hauts-de-France) la pathologie la plus fréquente pour les hommes sont les maladies psychiatriques avec une prévalence de 2,58%.

- Pathologie par prévalence décroissante pour les femmes dans la région ayant la plus haute population

```
SELECT patho_un.top,patho_un.patho_niv1,exister.prev, dep.region from exister,dep,patho_un,sexe WHERE exister.dep=dep.id_dep and
exister.top=patho_un.top and exister.sexe=sexe.id_sexe AND dep.region like"%Hauts%" and patho_un.patho_niv1 not like"%pas%" AND sexe.id_sexe=2 ORDER
by exister.prev DESC;
```

Profilage [Éditer en ligne] [Éditer] [Expliquer SQL] [Créer le code source PHP] [Actualiser]

Tout afficher | Nombre de lignes : 25 | Filtrer les lignes: Chercher dans cette table

Options supplémentaires

top	patho_niv1	prev 1	region
sup_PsyPat_cat	Maladies psychiatriques	1.602	Hauts-de-France
sup_Arthros_med	Traitement antalgique ou anti-inflammatoire (hors ...	0.971	Hauts-de-France
sup_NeuDeg_cat	Maladies neurologiques ou dégénératives	0.911	Hauts-de-France
sup_PsyMed_cat	Traitements psychotropes (hors pathologies)	0.901	Hauts-de-France
sup_InfRarVIH_cat	Maladies inflammatoires ou rares ou VIH ou SIDA	0.671	Hauts-de-France
sup_FRV_cat	Traitements du risque vasculaire (hors pathologies...	0.36	Hauts-de-France
sup_Cv_cat	Maladies cardio-neurovasculaires	0.33	Hauts-de-France
sup_Can_cat	Cancers	0.16	Hauts-de-France
sup_RIRCT_cat	Insuffisance rénale chronique terminale	0	Hauts-de-France

On voit que dans la région la plus peuplée de France (Hauts-de-France) la pathologie la plus fréquente pour les femmes sont les maladies psychiatriques avec une prévalence de 1,60%.

- Pathologie pour chaque département classé par prévalence décroissante

```
SELECT dep.id_dep,dep.dep,patho_un.patho_niv1,exister.prev from exister,dep,patho_un WHERE exister.dep=dep.id_dep and exister.top=patho_un.top a
patho_un.patho_niv1 not like"%pas%" group by dep.id_dep,dep.dep,patho_un.patho_niv1 order by dep.dep ASC,exister.prev DESC;
```

Profilage [Éditer en ligne] [Éditer] [Expliquer SQL] [Créer le code source PHP] [Actualiser]

1 > >> | Tout afficher | Nombre de lignes : 25 | Filtrer les lignes: Chercher dans cette table

Options supplémentaires

id_dep	dep 1	patho_niv1	prev
1	Ain	Maladies psychiatriques	2.919
1	Ain	Maladies neurologiques ou dégénératives	0.821
1	Ain	Traitements psychotropes (hors pathologies)	0.785
1	Ain	Traitement antalgique ou anti-inflammatoire (hors ...	0.758
1	Ain	Maladies inflammatoires ou rares ou VIH ou SIDA	0.539
1	Ain	Maladies cardio-neurovasculaires	0.513
1	Ain	Traitements du risque vasculaire (hors pathologies...	0.286
1	Ain	Cancers	0.173
1	Ain	Insuffisance rénale chronique terminale	0
2	Aisne	Maladies psychiatriques	1.602
2	Aisne	Maladies neurologiques ou dégénératives	1.003
2	Aisne	Traitement antalgique ou anti-inflammatoire (hors ...	0.971
2	Aisne	Traitements psychotropes (hors pathologies)	0.901
2	Aisne	Maladies inflammatoires ou rares ou VIH ou SIDA	0.671

- Prévalence moyenne pour le cancer chez les femmes

```
SELECT patho_un.patho_niv1, AVG(exister.prev) FROM exister,sexe,patho_un WHERE exister.sexe=sexe.id_sexe AND exister.top=patho_un.top AND
patho_un.patho_niv1 like'%cancer%' AND sexe.id_sexe=2;
```

Profilage [Éditer en ligne] [Éditer] [Expliquer SQL] [Créer le code source PHP] [Actualiser]

Tout afficher | Nombre de lignes : 25 | Filtrer les lignes: Chercher dans cette table

Options supplémentaires

	patho_niv1	AVG(exister.prev)
Éditer Copier Supprimer	Cancers	0.2608421060599779

- Prévalence moyenne pour le cancer chez les hommes

```
SELECT patho_un.patho_niv1, AVG(exister.prev) FROM exister,sexe,patho_un WHERE exister.sexe=sexe.id_sexe AND exister.top=patho_un.top AND
patho_un.patho_niv1 like'%cancer%' AND sexe.id_sexe=1;
```

Profilage [Éditer en ligne] [Éditer] [Expliquer SQL] [Créer le code source PHP] [Actualiser]

Tout afficher | Nombre de lignes : 25 | Filtrer les lignes: Chercher dans cette table

Options supplémentaires

	patho_niv1	AVG(exister.prev)
Éditer Copier Supprimer	Cancers	0.2368947385173095

Chapitre 4

Analyse descriptive

Nous allons étudier la nature des variables contenues dans chaque table.
Nous retouvons les variables suivantes:

- **La région** : variable qualitative nominale
- **Le departemennt** : variable qualitative nominale
- **La population totale pour chaque département** : variable quantitative continue
- **Le sexe** : variable qualitative nominale
- **Pathologie niveau 1** : variable qualitative nominale
- **Pathologie niveau 2** : variable qualitative nominale
- **Ntop** : variable quantitative continue
- **Ncommorb** : variable quatitative continue
- **Proportion commorbidité** : variable quantitative continue
- **Niveau priorité** : variable quantitative discrète

- **Prévalence** : variable quantitative continue
- **Montant** : variable quantitative continue

Analyse statistique

Lorsque nous comparons les effectifs des patient atteints du cancer pour les hommes et les femmes nous remarquons qu'ils sont quasiment égaux malgré que certains soient spécifiques à un sexe. Nous allons donc voir si la probabilité d'avoir un cancer est indépendante du sexe grâce à un test du Khi2.

	Homme	Femme	Total
Atteint	1621810	1758530	3380340
Non Atteint	31734830	33597910	65332740
Total	33356640	35356440	68713080

Nous pouvons maintenant à l'aide de R realiser un test Khi2.

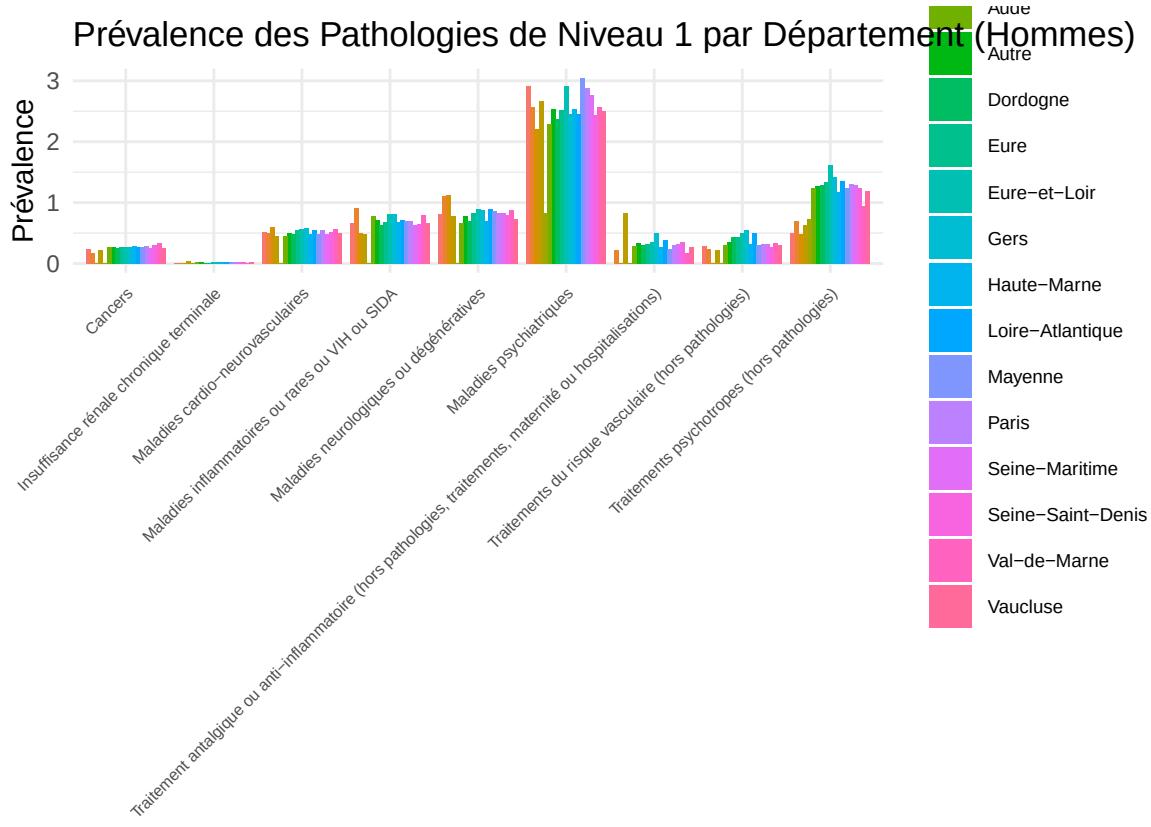
```
val_observ <- matrix(c(1621810, 31734830, 1758530, 33597910), nrow = 2,
                  dimnames = list("Statut" = c("Atteint", "Non Atteint"),
                                  "Genre" = c("Homme", "Femme")))
val_observ
```

```
##              Genre
## Statut           Homme    Femme
##   Atteint      1621810  1758530
##   Non Atteint 31734830 33597910
```

```
test_khi2 <- chisq.test(val_observ)

test_khi2
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  val_observ
## X-squared = 457.71, df = 1, p-value < 2.2e-16
```

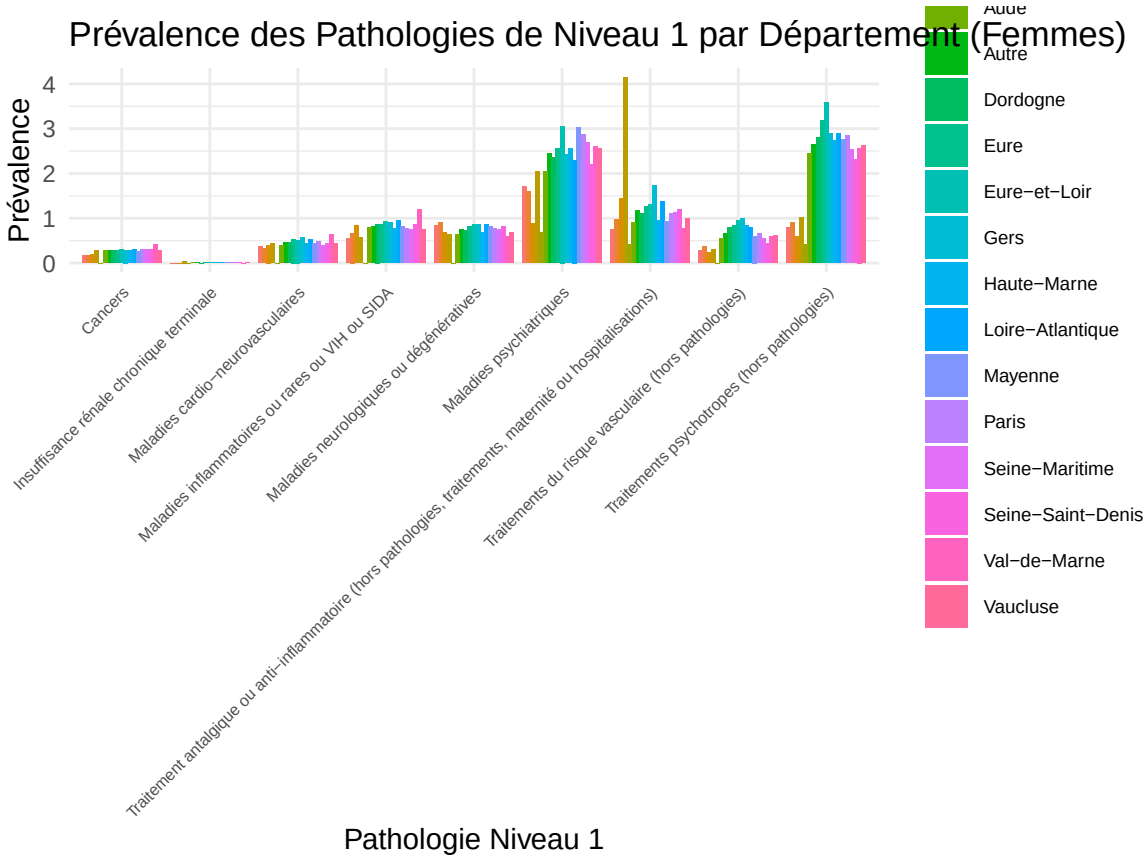
```r
# Exécution de la requête SQL
sql_result <- dbGetQuery(bd, "SELECT patho_un.patho_niv1, dep.dep, exister.prev
                              FROM exister, patho_un, sexe, dep
                              WHERE patho_un.top = exister.top
                              AND sexe.id_sexe = exister.sexe
                              AND exister.dep = dep.id_dep
                              AND sexe.id_sexe = 1
                              AND patho_un.patho_niv1 NOT LIKE '%pas%'
                              ORDER BY exister.prev DESC")

# Chargement du package ggplot2 s'il n'est pas déjà installé
if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}

# Chargement du package ggplot2
library(ggplot2)

# Création du graphique à barres
# Création du graphique à barres avec des barres plus grandes et
# du texte plus petit
ggplot(sql_result, aes(x = patho_niv1, y = prev, fill = dep)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.9) +
  # Ajustement de la largeur des barres
  labs(title = "Prévalence des Pathologies de Niveau 1 par Département (Hommes)",
       x = "Pathologie Niveau 1",
       y = "Prévalence",
       fill = "Département") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 5.5)) +
  # Ajustement de la taille du texte sur l'axe x
  theme(legend.text=element_text(size=7))
```




```
# Ajustement de la taille du texte dans la légende
```

```
# Exécution de la requête SQL
sql_result <- dbGetQuery(bd, "SELECT patho_un.patho_niv1, dep.dep, exister.prev
                              FROM exister, patho_un, sexe, dep
                              WHERE patho_un.top = exister.top
                              AND sexe.id_sexe = exister.sexe
                              AND exister.dep = dep.id_dep
                              AND sexe.id_sexe = 2
                              AND patho_un.patho_niv1 NOT LIKE '%pas%'
                              ORDER BY exister.prev DESC")

# Chargement du package ggplot2 s'il n'est pas déjà installé
if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}

# Chargement du package ggplot2
library(ggplot2)

# Création du graphique à barres
```

```r
# Création du graphique à barres avec des barres
#plus grandes et du texte plus petit
ggplot(sql_result, aes(x = patho_niv1, y = prev, fill = dep)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.9) +
  # Ajustement de la largeur des barres
  labs(title = "Prévalence des Pathologies de Niveau 1 par Département (Femmes)",
       x = "Pathologie Niveau 1",
       y = "Prévalence",
       fill = "Département") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 5.5)) +
  # Ajustement de la taille du texte sur l'axe x
  theme(legend.text=element_text(size=7))
```




```r
# Ajustement de la taille du texte dans la légende
```

Les deux graphiques ci dessous montre la répartition des pathologie en fonction de la zone géographique. Avec respectivement, un pour les hommes et l'autre Pour les femmes. Nous constatons que dans les deux cas, la prévalence des maladies varie significativement d'un département à l'autre, indiquant des facteurs régionaux influents tels que les différences

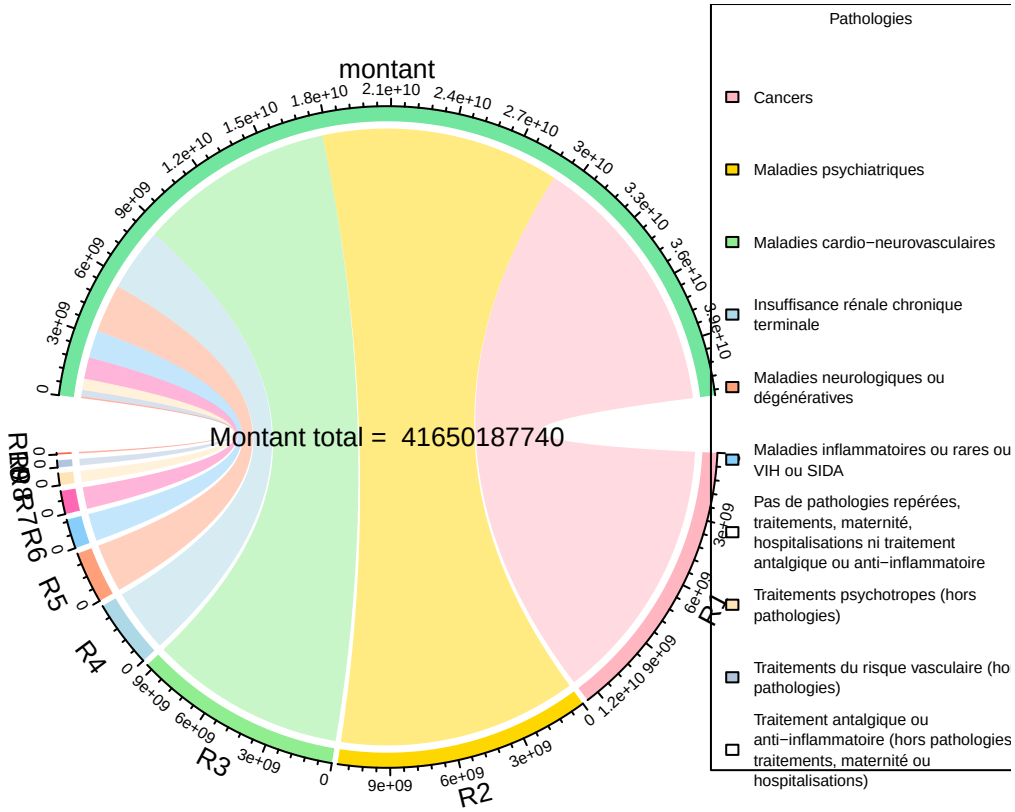
socio-économiques, l'accès aux soins de santé, ou des facteurs environnementaux spécifiques. Par exemple, les maladies cardiovasculaires et certains types de cancer peuvent être plus prévalents chez l'un des sexes en raison de facteurs biologiques et comportementaux. De plus, les maladies psychologiques semblent avoir une prévalence relativement élevée chez les femmes dans plusieurs départements comparé à celle observée chez les hommes, ce qui pourrait refléter des différences dans la manière dont les maladies sont diagnostiquées ou signalées entre les sexes.

```
# Exécution de la requête SQL
sql_result <- dbGetQuery(bd, "SELECT patho_niv1, montant
                              FROM patho_un
                              WHERE montant > 0
                              GROUP BY top
                              ORDER BY montant DESC")

# Calcul du nombre total
total <- sum(sql_result$montant)

# Définition des pathologies
pathologies <- c("Cancers", "Maladies psychiatriques",
                 "Maladies cardio-neurovasculaires",
                 "Insuffisance rénale chronique terminale",
                 "Maladies neurologiques ou dégénératives",
                 "Maladies inflammatoires ou rares ou VIH ou SIDA",
                 "Pas de pathologies repérées, traitements, maternité,
                 hospitalisations ni traitement antalgique ou anti-inflammatoire",
                 "Traitements psychotropes (hors pathologies)",
                 "Traitements du risque vasculaire (hors pathologies)",
                 "Traitement antalgique ou anti-inflammatoire
                 (hors pathologies, traitements, maternité ou hospitalisations)")

# Définition des couleurs pastel
couleurs_pastel <- c("#FFB6C1", "#FFD700", "#90EE90", "#ADD8E6", "#FFA07A", "#87CEFA", "

# Création d'une liste de correspondance avec les pathologies
correspondance_couleurs_pastel <- setNames(couleurs_pastel, pathologies)

# Sélection des couleurs pastel correspondantes à chaque pathologie
patho_colors_pastel <- correspondance_couleurs_pastel[sql_result$patho_niv1]

# Création du diagramme en cordes avec les couleurs pastel
chord_diagram_pastel <- chordDiagram(as.matrix(sql_result[-1]), transparency = 0.5,
                                     grid.col = c(R1 = "#FFB6C1", R2= "#FFD700", R3 = "#
                                                  R4= "#ADD8E6",R5= "#FFA07A", R6="#87CE
                                                  R7="#FF69B4", R8="#FFE4B5", R9= "#B0C4
```

```
                                                    R10="#FF6347"))
 par(xpd=TRUE)
# Définir les valeurs de légende à modifier
legend_labels <- sql_result$patho_niv1

# Appliquer str_wrap pour couper les légendes
wrapped_labels <- str_wrap(legend_labels, width = 38)

text(0, 0, paste("Montant total = ", total), cex = 0.8)

# Création de la légende principale avec les légendes adaptées
legend_title <- "Pathologies"
legend(x = max(sql_result$montant) + 0.4,  # Placer la légende à
       #côté du diagramme
       y = "bottom",  # Aligner la légende en bas
       legend = wrapped_labels,
       fill = patho_colors_pastel,  # Utiliser les couleurs pastel
       title = legend_title


)
```

```
## Warning in xy.coords(x, y, setLab = FALSE): NAs introduits lors de la
## conversion automatique
```

```
# Ajout de la légende de pathologie à côté
#du diagramme avec des couleurs pastel
legend("topright", inset = c(-0.05, -0.05),
       legend = wrapped_labels,
       fill = patho_colors_pastel,  # Utiliser les couleurs pastel
       title = "Pathologies",
       cex = 0.53)
```



Ce diagramme en camembert présente la répartition des coûts associés à différentes pathologies. Les cancers représentent une portion significative des dépenses, ce qui est attendu étant donné le coût élevé des traitements oncologiques, y compris la chimiothérapie, la radiothérapie, et les médicaments spécialisés. L'insufisance rénale malgré sa forte prévalence chez les personnes agées ne représente pas une grosse partie des coût, indiquant surement un faible coût de traitement par patient. Ce graphique permet de nous apprendre l'importance des pathologies dans la gestions du budget des différent instances régionales et nationales liées à la santé.

```
# Charger les packages nécessaires
library(ggplot2)
library(maps)
```

```
## Warning: le package 'maps' a été compilé avec la version R 4.3.3
```

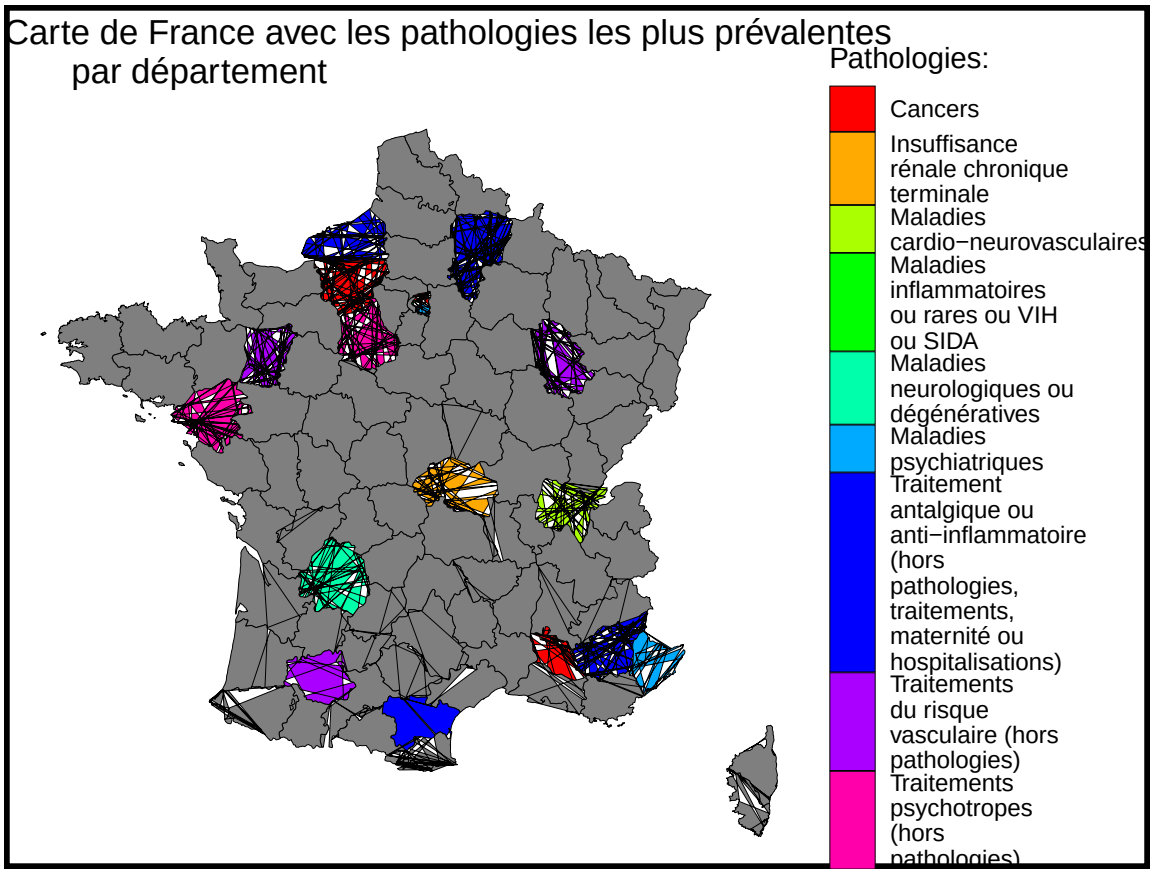
```
library(mapdata)
```

```
## Warning: le package 'mapdata' a été compilé avec la version R 4.3.3
```

```r
# Charger les données de la carte de la France
france <- map_data("france")

# Adapter la requête SQL selon vos besoins pour obtenir les 16 pathologies les
#plus prévalentes par département
patho_data <- dbGetQuery(bd, "SELECT
    dep.dep AS region,
    patho_un.patho_niv1 AS patho,
    MAX(exister.prev) AS max_prev
FROM
    patho_un
JOIN
    exister ON patho_un.top = exister.top
JOIN
    sexe ON sexe.id_sexe = exister.sexe
JOIN
    dep ON dep.id_dep = exister.dep
WHERE
    sexe.id_sexe = 2
AND patho_un.patho_niv1 NOT LIKE '%Pas%'
GROUP BY
    dep.dep,
    patho_un.patho_niv1
ORDER BY
    max_prev DESC
;")

# Définir une palette de couleurs pour chaque niveau de pathologie
palette_couleurs <- rainbow(length(unique(patho_data$patho)))

# Fusionner les données agrégées avec les données de la carte
france_map <- merge(france, patho_data, by.x = "region", by.y = "region",
                    all.x = TRUE)




library(stringr) # Charger le package stringr pour utiliser str_wrap

# Définir les valeurs de légende à modifier
legend_labels <- levels(factor(patho_data$patho))

# Appliquer str_wrap pour raccourcir et mettre en forme les légendes
```

```r
wrapped_labels <- str_wrap(legend_labels, width = 16)

# Créer le graphique avec les données fusionnées et légendes modifiées
ggplot() +
  geom_map(data = france_map, map = france_map,
           aes(map_id = region, fill = patho),
           color = "black", size = 0.1) +
  scale_fill_manual(values = palette_couleurs, name = "Pathologies:",
                    labels = wrapped_labels) +
  expand_limits(x = france_map$long, y = france_map$lat) +
  coord_map() +
  theme_void() +
  theme(plot.background = element_rect(linewidth = 2, fill = "white")) +
  labs(title = "Carte de France avec les pathologies les plus prévalentes
       par département")
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



Cette carte montre la répartition géographique des pathologies les plus prévalentes par département en France. Les cancers (rouge) et les maladies cardio-neurovasculaires (bleu) sont visuellement prédominants, notamment dans les départements avec de grande agglomération présente, indiquant un potentiel facteur liée à la polution ou au stress. Les maladies psychiatriques (violet) apparaissent également de manière notable, ce qui pourrait indiquer une reconnaissance croissante de ces conditions ou une meilleure disponibilité des données. Enfin, dans la diagonale du vide on observe des maladies liées à la viellesse qui sont prédominentes (tel que l'insuffisance rénale, les maladie dégénératives ou nerologique), indiquant probablement une population vieillissante mais avec un plus faible taux de maladie liée à un mauvais environnement.

Chapitre 5

Difficultés rencontrées

La principale difficulté rencontrée a été la création de nos tables. Nos bases de données étant extrêmement denses, le tri et l'organisation étaient donc complexes. Comme décrit précédemment, nous avons dû adapter les données à notre modèle.

La compréhension des jeux de données n'était pas toujours évidente, ce qui nous a confrontés à des défis supplémentaires concernant leur compréhension. Comprendre pleinement les données était crucial, ce qui nous a amenés à modifier notre Modèle Conceptuel de Données (MCD) à de nombreuses reprises (4 à 5 fois) et donc à réorganiser nos tables en conséquence.

En raison du volume important de données, nous avons dû faire des choix stratégiques, comme décrit précédemment, pour déterminer sur quoi nous focaliser et quelles données et quelles requêtes étaient les plus pertinentes pour répondre à notre problématique. Cette sélection minutieuse a été essentielle pour garantir la pertinence et la cohérence de notre analyse.

Nous avons également eu des difficultés concernant l'affichage des légendes sur R, ce qui nous a obligé à supprimer certains graphiques du fait de leur mauvaise lisibilité. R n'acceptant pas les minipage de LateX nous avons du combiner deux pdf pour afficher notre page de garde.

Perspectives

Avec plus de temps, de compétences et de membres dans notre groupe, nous aurions pu étendre notre analyse et préciser nos objectifs. Une direction intéressante aurait été d'explorer le montant remboursé par pathologie et par région. Cette analyse aurait permis de mieux

comprendre les politiques de santé régionales mises en place par les Agences Régionales de Santé.

De plus, l'absence de données sur la prévalence des comorbidités dans notre base de données limite notre analyse. Avec ces données supplémentaires, nous aurions pu évaluer plus précisément les tendances de santé. Leur prise en compte aurait permis une meilleure compréhension des facteurs de risque et des complications associées aux pathologies étudiées.

Conclusion

Dans l'ensemble, cette analyse nous a permis de mieux comprendre la distribution des pathologies, des comorbidités et des remboursements dans différentes régions et départements. Malgré les défis rencontrés, nous avons pu identifier des tendances significatives pour orienter les politiques de santé.

Nos résultats soulignent l'importance de prendre en compte les variations régionales dans la planification des interventions de santé publique. Par exemple, nous avons observé des différences notables dans la prévalence des maladies en fonction des régions, ce qui suggère la nécessité d'adapter les programmes de dépistage et de prévention en fonction des besoins locaux.

De plus, l'identification des pathologies les plus coûteuses en termes de remboursements permet aux décideurs de prioriser les ressources financières là où elles sont le plus nécessaires. Cette information peut contribuer à une allocation plus efficace des budgets de santé.