

TDU.

Introduction à R pour les enquêtes de santé publique

Cahier d'exercices pour le Jour 3



Public Health
Agency of Canada

Agence de la santé
publique du Canada

Canada

Contenu

Exercice de mise en pratique	2
Consignes.....	2
Introduction	3
Organisation de votre espace de travail	5
Configurez votre session R Studio	10
Fonctions vedettes	39

EXERCICE DE MISE EN PRATIQUE

Consignes

Durée : 4 heures et 10 minutes

Les apprenants et apprenantes ont un scénario, des questions et des tâches avec le code associé pour effectuer chaque tâche. Nous recommandons aux :

Utilisateurs et utilisatrices novices (maîtres d'équipage) : Utilisez ce cahier d'exercices ainsi que le ou les scripts R fournis par le biais de GitHub. En vous servant de ce cahier d'exercices comme guide, exécutez le code que nous vous avons fourni étape par étape pour comprendre ce à quoi chaque partie du code sert et ce que font les différentes fonctions. À ce stade-ci, ne vous préoccupez pas de rédiger ou de corriger le code.

Utilisateurs et utilisatrices de niveau débutant/intermédiaire (seconds capitaines et secondes capitaines) : Le code R est fourni sous forme de capture d'écran dans ce cahier d'exercices. Vous devriez avoir une assez bonne compréhension du codage pour avoir une idée générale du code en le lisant (avec la documentation d'aide, quelques recherches Google au besoin et les commentaires se trouvant dans les scripts fournis par l'entremise de GitHub). Notre but est de vous faire écrire le code à partir du guide à mesure que vous avancez dans le scénario. Consultez les scripts fournis dans GitHub si vous rencontrez des problèmes.

Utilisateurs et utilisatrices avancés (capitaines au long cours) : Nous vous encourageons à essayer d'écrire votre propre code où vous voulez ainsi qu'à le contraster avec le code utilisé pour l'exercice et à aider vos pairs si des questions surviennent. Consultez les scripts fournis dans GitHub si vous rencontrez des problèmes.

Avancez autant que vous le pouvez dans cet exercice en quatre heures et dix minutes (maximum). Ne vous en faites pas si vous avez besoin de plus de temps. La courbe d'apprentissage pour R est abrupte et les apprenants ainsi que les apprenantes bénéficieront encore plus du temps consacré à la mise en pratique. Nous planifierons un webinaire de suivi d'une durée d'une heure après le cours pour faire le point sur les exercices (date à venir). Avant cela, si vous avez besoin d'aide avec le matériel de cours, contactez les personnes responsables de l'animation de votre cours par le biais de Slack ou par courriel.

Introduction

Au cours de cette session, vous mettrez en pratique vos compétences en lien avec R et la gestion des données afin de préparer et de recueillir les données provenant de plusieurs ensembles de données, d'appliquer un algorithme d'identification des cas et d'effectuer une analyse épidémiologique descriptive de base. On vous a fourni cinq ensembles de données qui sont des extraits fictifs d'ensembles de données administratifs (factures de médecins, autorisations de sortie de l'hôpital et liste de clients de l'assurance maladie provinciale) et qui ont été simplifiés à des fins de formation. Ces ensembles de données doivent être combinés afin d'appliquer efficacement les définitions de cas administratives (algorithmes) pour que vous puissiez décrire la fréquence de l'utilisation des services de santé pour les maladies chroniques parmi les résidents d'une petite province. Les ensembles de données et la définition de cas suivants vous ont été fournis :

Factures de médecins : Les ensembles de données *mSP1213* et *mSP1314* comprennent des informations provenant de factures de la rémunération à l'acte des médecins (médecins généralistes et spécialistes) pour les années financières 2012-2013 ainsi que 2013-2014 respectivement. Chaque dossier correspond à une visite chez un médecin, à la date de la visite et à un seul code de diagnostic. Bien que des personnes puissent avoir besoin de discuter de plusieurs problèmes de santé lors de leur rendez-vous avec leur médecin, ce dernier sélectionne le problème de santé « le plus responsable » pour la facturation.

Autorisations de sortie de l'hôpital : Les ensembles de données *dad1213* et *dad1314* comprennent des informations à propos du résumé d'autorisation produit à la sortie de l'hôpital. Les données comprennent seulement les hospitalisations de soins actifs. Les autres hospitalisations, comme les services hospitaliers ambulatoires ou de soins de longue durée, sont exclues. Les diagnostics multiples sont inclus dans le résumé d'autorisation de sortie de l'hôpital, avec les dates d'admission et de sortie de l'individu.

Liste de clients de l'assurance maladie provinciale : L'ensemble de données *reg1314* comprend une liste de toutes les personnes inscrites à l'assurance maladie provinciale (p. ex. Régime de service médical, Assurance-santé de l'Ontario, régime d'assurance maladie, etc.). L'inscription est automatique lors de la délivrance d'un certificat de naissance pour les résidents de la province (dans le cas de nouvelles naissances) ou la demande est effectuée lors d'un déménagement dans la province. L'inscription est annulée en cas de déménagement, de décès ou d'inscription à l'assurance maladie provinciale dans une nouvelle province. Les services de soins de santé financés par l'État sont facturés à la province de résidence ou au programme

fédéral offrant une couverture d'assurance maladie. Pour les personnes qui ne sont pas inscrites, ceux-ci doivent être payés de leur poche ou par une assurance privée. Quoi qu'il en soit, les personnes inscrites qui ne résident pas dans la province ainsi que celles ayant accès aux services de soins de santé et n'étant pas inscrites au Régime de service médical provincial ont été exclues de cet ensemble de données.

Tableau 1 : Contenu des ensembles de données fournis pour le Jour 3 du cours *Introduction à R*

Ensemble de données	Champs	Type	Définition
<i>msp1213.csv</i>	<i>uniqueid</i>	Chaîne	Identifiant unique (personne)
	<i>servdat</i>	Date	Date de la visite chez le médecin
	<i>diag1</i>	Chaîne	Code de diagnostic (ICD9) apparaissant sur la facture
<i>msp1314.csv</i>	<i>uniqueid</i>	Chaîne	Identifiant unique (personne)
	<i>servdat</i>	Date	Date de la visite chez le médecin
	<i>diag1</i>	Chaîne	Code de diagnostic (ICD9) apparaissant sur la facture
<i>dad1213.csv</i>	<i>uniqueid</i>	Chaîne	Identifiant unique (personne)
	<i>admitdat</i>	Date	Date d'admission à l'hôpital
	<i>sepdat</i>	Date	Date d'autorisation de sortie de l'hôpital
	<i>diag1</i>	Chaîne	Code de diagnostic (ICD9) apparaissant sur l'autorisation de sortie
	<i>diag2</i>	Chaîne	Code de diagnostic (ICD9) apparaissant sur l'autorisation de sortie
	<i>diag3</i>	Chaîne	Code de diagnostic (ICD9) apparaissant sur l'autorisation de sortie
<i>dad1314.csv</i>	<i>uniqueid</i>	Chaîne	Identifiant unique (personne)

	<i>admitdat</i>	Date	Date d'admission à l'hôpital
	<i>sepdat</i>	Date	Date d'autorisation de sortie de l'hôpital
	<i>diag1</i>	Chaîne	Code de diagnostic (ICD9) apparaissant sur l'autorisation de sortie
	<i>diag2</i>	Chaîne	Code de diagnostic (ICD9) apparaissant sur l'autorisation de sortie
	<i>diag3</i>	Chaîne	Code de diagnostic (ICD9) apparaissant sur l'autorisation de sortie
<i>reg1314.csv</i>	<i>uniqueid</i>	Chaîne	Identifiant unique (personne)
	<i>dob</i>	Date	Date de naissance
	<i>sex</i>	Numérique	F = 1, M = 2

Définition de l'asthme selon le SCSMC¹: La définition de cas de l'asthme diagnostiqué est la suivante : une personne âgée d'un an ou plus ayant au moins deux réclamations de facturation de médecin sur une période de deux ans avec un diagnostic d'asthme dans le premier champ de diagnostic, ou ayant au moins une sortie de l'hôpital avec un diagnostic d'asthme dans n'importe quel champ de diagnostic, et dont le code diagnostique était 493 selon la Classification internationale des maladies (CIM), neuvième révision ou CIM-9-CM, ou J45 ou J46 selon la CIM-10-CA. Veuillez noter que la mise en application de la définition de cas dans cet exercice est très simplifiée.

Organisation de votre espace de travail

1. Créez de nouveaux dossiers sur votre ordinateur pour organiser les fichiers de l'exercice d'aujourd'hui :
 - a. À l'intérieur du dossier *IntoToR*, créez un sous-dossier pour le Jour 3 et nommez-le *Exercise_Day3*.
 - b. À l'intérieur du dossier *Exercise_Day3*, créez les nouveaux dossiers suivants:
 - i. *data*

¹ <https://www.canada.ca/fr/sante-publique/services/publications/maladies-et-affections/asthme-maladie-pulmonaire-obstructive-chronique-canada-2018.html>

ii. *scripts*

iii. *output*

- c. Déplacez les fichiers pour le Jour 3 vers leurs dossiers respectifs.

Remarque : Pour le Jour 3, plutôt que d'écrire plusieurs scripts individuels, nous allons travailler à partir d'un seul fichier .Rmd dans votre dossier de scripts pour cet exercice.

2. Maintenant que vous avez organisé vos dossiers de projet :

Ouvrez les ensembles de données dans Excel (msp1213.csv, msp1314.csv, dad1213.csv, dad1314.csv, reg1314.csv).

- a. Est-ce qu'*uniqueid* est réellement unique? Que représente chaque dossier selon vous?
- b. Selon vous, quels champs pourraient être utiles pour relier les données entre les tableaux?
- c. Reportez-vous à la définition de cas administrative. Selon vous, quels sont les ensembles de données et les champs qui pourraient être nécessaires pour identifier les cas d'asthme et leurs dates?
- d. Compte tenu de la définition de cas administrative et des données que vous avez, quels calculs épidémiologiques descriptifs aimeriez-vous effectuer pour décrire la présence de l'asthme dans cette population?
- e. Considérant le fait que *reg1314* est un registre de clients du Régime de service médical pour l'année financière 2013-2014, pensez-vous qu'il est acceptable de l'utiliser comme registre de population? Pourquoi ou pourquoi pas? Quelles pourraient être les limites s'il devait être utilisé comme registre de population?

- f. Quelles sont les informations supplémentaires dont vous avez besoin de la part des quatre autres tableaux afin de les ajouter à l'ensemble de données *reg1314* dans le but d'effectuer votre analyse descriptive de l'asthme?

- g. Comment ces informations doivent-elles être présentées sous forme de champs de données dans l'ensemble de données *reg1314* pour que vous puissiez effectuer votre analyse descriptive? Quelles sont les étapes à suivre pendant le traitement des données pour inclure ces champs de données dans l'ensemble de données *reg1314*?

Indice : dessinez le fichier de résultat. En ayant clairement en tête votre point de départ (c.-à-d. les ensembles de données fournis) et votre point final (c.-à-d. le fichier plat final), vous pourrez identifier les étapes précises que vous devez suivre pendant le traitement des données.

Dans le cadre de cet exercice, les étapes suivantes seront suivies pour créer ce fichier plat :

uniqueid	dob	age	sex	asthma_case	asthma_casedate	popn_1plus
1	1936-09-21	76	1	0	NA	1
2	1916-04-08	97	2	1	2012-09-24	1
3	1924-10-17	88	1	0	NA	1
4	1995-04-01	18	2	0	NA	1
5	1997-06-28	16	1	0	NA	1

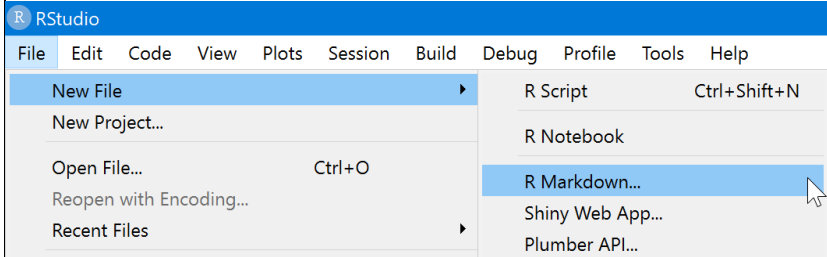
Aperçu des étapes de cet exercice :

- Annexe *DAD*.
- Indiquez le tri effectué à l'hôpital en lien avec les diagnostics d'asthme
 - Gardez seulement les hospitalisations liées à l'asthme
 - Identifiez les doublons
 - Identifiez et comptez les dossiers en double pour chaque *uniqueid*, conservez uniquement le plus ancien *sepdad*.
- Sauvegardez les données *DAD* en tant que nouveau fichier annexé.
- Annexe *MSP*.
- Indiquez les diagnostics d'asthme découlant de visites chez le médecin
 - Gardez seulement les visites chez le médecin liées à l'asthme
 - Identifiez les doublons
 - Identifiez et comptez les dossiers en doubles pour chaque *uniqueid*, conservez uniquement le plus ancien *servdat*.
- Sauvegarder les données *MSP* en tant que nouveau fichier annexé
- Fusionnez le registre aux données *DAD* et *MSP*
 - Fusionnez *DAD* et *Registry* (*full join, uniqueid*)
 - Fusionnez *MSP* et *Registry* (*full join, uniqueid*)
 - Fusionnez les deux fichiers de résultat (*full join, uniqueid dob sex*).
- Sauvegardez-le en tant que nouveau fichier de registre avec des informations sur l'état de la maladie chronique.
- Appliquez la définition de cas : indiquez les personnes qui répondent à la définition de cas.
- Calculez la date du cas (plus ancien service de santé codé pour l'asthme)
- Préparez le registre pour les calculs d'épidémiologie descriptive
 - Calculez l'âge au milieu de l'année 2013.
 - Indiquez les personnes faisant partie du dénominateur.
 - Appliquez les catégories de groupes d'âge.
- Analysez les données
 - Additionnez l'indicateur de définition de cas (numérateur) et l'indicateur de dénominateur, calculez le taux.

- Additionnez l'indicateur de définition de cas (numérateur) et l'indicateur de dénominateur (par groupe d'âge et/ou sexe), calculez les taux.
- Créez les figures.

Configurez votre session R Studio

Remarque : Essayez d'ajouter vos propres commentaires au code pour garder des traces des choses essentielles telles que le nombre de dossiers et de variables (afin de vous assurer que les opérations se déroulent comme prévu) et pour clarifier le code pour votre compréhension (afin de vous assurer que, dans plusieurs semaines, vous pourrez comprendre ce que vous avez produit dans le cadre de cet exercice). Veuillez noter qu'un fichier .rmd principal est fourni pour cet exercice (avec commentaires) afin de vous aider si vous rencontrez des problèmes. Comme cet exercice est compliqué, n'hésitez surtout pas à consulter le fichier .rmd fourni pour réaliser l'exercice. Vous pouvez suivre le fichier .rmd principal si vous cherchez à développer votre compréhension en effectuant un copier-coller du code dans votre propre fichier .rmd ou vous pouvez taper votre propre code à partir des images fournies dans votre nouveau fichier .rmd. Quelle que soit votre approche dans le cadre de cet exercice, il est essentiel que vous portiez une attention aux changements dans vos fichiers (par exemple, le nombre de variables et d'observations) afin de vous assurer que les liens s'effectuent comme prévu.

Tâche	Code
Ouvrez un nouveau fichier R Markdown, sélectionnez HTML comme format d'extrait par défaut et nommez-le : <i>Day3_Exercise_script.Rmd</i> .	
Notez que pour l'exercice du Jour 3, nous allons travailler à partir d'un seul fichier R Markdown plutôt que de travailler avec plusieurs fichiers en même temps et d'utiliser notre fichier R Markdown comme un carnet de notes en	<pre> --- title: "Day3_Exercise_Script" output: html_document ---</pre>

<p>format HTML. D'abord, vous allez devoir préciser ce format dans l'en-tête YAML située dans le haut.</p>	
<p>Ensuite, vous aurez besoin d'un bout de code pour configurer votre document pour les analyses. Ici, nous chargerons directement les <i>librairies</i> que nous prévoyons d'utiliser, nous initialiserons l'emplacement du script .Rmd actuel et nous définirons l'emplacement du chemin d'accès au dossier <i>output</i> pour une utilisation ultérieure.</p>	<pre> ### Setup ```{r setup, include=FALSE} require("knitr") library(tidyverse) library(here) knitr::opts_chunk\$set(echo = TRUE) here::i_am("Exercise_Day3/scripts/Exercise_Day3.Rmd") output_folder <- here::here("Exercise_Day3", "output") ``` </pre>

3. Préparez les données :



Chargez les ensembles de données d'hospitalisation (*dad1213.csv* et *dad1314.csv*) et conservez-les en tant qu'objets en mémoire. Notez le nombre d'observations et de variables pour chacun d'eux.

Remarque : Si vous avez besoin d'ajouter des bouts supplémentaires à votre fichier R Markdown, vous pouvez cliquer sur *Insert -> R* au haut de la barre d'outils Markdown ou utiliser le raccourci clavier Ctrl + Alt + I (Cmd + Option + I pour macOS).

```
```{r}
dad1213 <- read_csv(here("Exercise_Day3", "data", "dad1213.csv"))
dad1314 <- read_csv(here("Exercise_Day3", "data", "dad1314.csv"))
```
```

Annexez les ensembles de données d'hospitalisation (en utilisant *bind_rows*). Combien d'hospitalisations y a-t-il dans chaque fichier?

Confirmez les dates de sorties d'hôpital tombant dans la

Append DAD datasets:

```
```{r}
dad_appnd <- bind_rows(dad1213, dad1314) %>%
 mutate(sepdat = as.Date(sepdat, "%d-%b-%Y")) %>%
 filter(sepdat > "2012-04-01" & sepdat < "2014-03-31")
```
```

bonne année financière.
Combien de dossiers doivent
être supprimés?

Remarque : Si vous voulez
seulement exécuter une partie
du code, utilisez le menu
déroulant *Run* au haut de la
barre d'outils Markdown.

Indice : essayez de mettre `print(unique(dad_appnd$sepdat))` dans la fenêtre de commandes

Ou, consultez et triez le fichier en double-cliquant sur le fichier dans le panneau
d'environnement et en cliquant sur l'ensemble de triangles à côté du nom de la variable :



| uniqueid | admitdat | sepdat | diag1 | diag2 | diag3 |
|----------|-------------|------------|-------|-------|-------|
| 1779 | 29-Mar-2012 | 2012-04-02 | 1011 | 9500 | 2790 |
| 3016 | 02-Apr-2012 | 2012-04-02 | 1080 | 6230 | 8850 |
| 1034 | 02-Apr-2012 | 2012-04-02 | 1100 | 1100 | N/A |

Indiquez et gardez les hospitalisations pour lesquelles un diagnostic d'asthme a été confirmé. Comme nous nous intéressons à l'ensemble du bloc de code du diagnostic 493, nous prendrons une sous-chaîne de code pour filtrer. Combien d'hospitalisations liées à l'asthme y a-t-il eu en 2013-2014 et 2012-2013?

Pour s'exercer : combien y a-t-il de personnes uniques (indice : taille du fichier à la fin de ce code)? Quel est le nombre maximal d'hospitalisations par personne durant la période d'intérêt (indice : fonctions *view* et *sort*, *table* ou *count*)?

Process DAD Datasets:

```
```{r}
```

```
#create flag variable.
```

```
#####Step 1: Extract first three digits of diag code.
```

```
dad_appnd <- dad_appnd %>% mutate(diag1_sbstr = str_sub(diag1, end=3)) %>%
```

```
 mutate(diag2_sbstr = str_sub(diag2, end=3)) %>%
```

```
 mutate(diag3_sbstr = str_sub(diag3, end=3))
```

```
#####Step 2: create flag where 0 = no asthma, 1= asthma if any diag variables equal to 493.
```

```
dad_appnd <- dad_appnd %>% mutate(flag = case_when(diag1_sbstr == "493" | diag2_sbstr == "493" | diag3_sbstr == "493" ~ 1, TRUE ~ 0))
```

Comme certains individus ont été hospitalisés plus d'une fois au cours de cette période, conservez uniquement la plus ancienne date de sortie. Comment la fonction *distinct* supprime-t-elle les doublons? Les données doivent-elles être triées avant d'utiliser cette fonction? Les données devraient-elles être triées par ordre croissant ou décroissant, et selon quelles variables?

Gardez et renommez seulement les variables dont vous aurez besoin plus tard (c.-à-d. *uniqueid*, *sepdatt*, *count\_duplicates*).

Organisez les données et l'environnement de travail.

```
#keep unique asthma-related hosps
#####Step 1: Keep only asthma related events.
#####Step 2: Group by unique ID (a) and count how many uniqueIDs appear in each group (b)
#####Step 3: Keep only earliest sepdatt where unique persons are duplicated.
#####Step 4: Keep (a) and rename (b) only the variables you need for your analysis.
dad_appnd_asthma <- dad_appnd %>%
 filter(dad_appnd$flag == 1) %>% #Step 1
 group_by(uniqueid) %>% #Step 2 (a)
 add_tally(name = "count_duplicates") %>% #Step 2 (b)
 arrange(uniqueid, sepdatt) %>% #Step 3 (a)
 distinct(uniqueid, .keep_all = TRUE) %>% #Step 3 (b)
 select(uniqueid, sepdatt, count_duplicates) %>% #Step 4 (a)
 rename(dad_sepdatt_asthma = sepdatt, dad_count_asthma=count_duplicates) #Step 4 (b)

#tidy workspace
rm(dad_appnd, dad1213,dad1314)

...
```

*Remarque : Le nombre de dossiers restants devrait être égal au nombre d'individus uniques obtenus ci-dessus.*

<p>Annexer les ensembles de données des médecins (<i>msp1213.csv</i> et <i>msp1314.csv</i>). Combien de visites de médecins y a-t-il dans chaque fichier? Combien y en a-t-il dans le fichier annexé?</p>	<pre>#### Prepare MSP datasets ```{r} ##load files msp1213 &lt;- read_csv(here("Exercise_Day3", "data", "msp1213.csv")) msp1314 &lt;- read_csv(here("Exercise_Day3", "data", "msp1314.csv"))  ##append datasets msp_appnd &lt;- bind_rows(msp1213, msp1314)</pre> <p><i>Conseil : Si vous partez de zéro, pensez à réutiliser le code DAD.</i></p>
<p>Comme nous l'avons fait pour les ensembles de données <i>DAD</i>, nous devons maintenant procéder au traitement des ensembles de données <i>MSP</i>.</p> <p>Pour ce faire, nous voulons formater à nouveau <i>servdat</i> en une variable date-classe à utiliser pour le filtrage; gardez les dates entre le 1<sup>er</sup> avril 2012 et le 13 mars 2014, puis utilisez les codes des médecins en commençant par</p>	<pre>##### Traitez les ensembles de données MSP :  ### process MSP. #####Step 1 - Reformat servdat to date-class variable. #####Step 2 - Filter by servdat: keep if servdat between Apr 1 2012 &amp; Mar 13 2014 #####Step 3 - Extract first three digits from diag1. #####Step 4 - Create new flag variable, 1=asthma, 0=no asthma; code block 493 (incl 4390...4939)  msp_appnd &lt;- msp_appnd %&gt;%   mutate(servdat = as.Date(servdat, "%d%b%Y")) %&gt;% #Step 1   filter(servdat &gt;= "2012-04-01" &amp; servdat &lt;= "2014-03-31") %&gt;% #Step 2   mutate(diag1_sbstr = as.numeric(str_sub(diag1, end=3))) %&gt;% #Step 3   mutate(flag = case_when(diag1_sbstr == "493" ~ 1, TRUE ~ 0)) #Step 4</pre>



<p>493 afin de créer un indicateur pour l'asthme.</p>	
<p>Laissez tomber toutes les visites chez les médecins qui ne sont pas liées à l'asthme. Combien y a-t-il d'observations restantes (indice : taille du fichier)? Quel est le nombre maximal de visites chez les médecins codées pour l'asthme par personne (indice : fonctions <i>view</i> et <i>sort</i> ou <i>table</i> et <i>count</i>)?</p> <p>Comme certains individus ont consulté leur médecin plus d'une fois au cours de cette période, gardez seulement la plus ancienne date de visite. Parmi les individus uniques qui ont consulté leur médecin, combien de visites ont été codées pour l'asthme (indice : fonctions <i>view</i> et <i>sort</i> ou <i>table</i> et <i>count</i>)?</p>	<pre>#Step 1 - Keep &amp; count phys visits for asthma #Step 2 - Identify and count how many observations per <u>uniqueid</u> #Step 3 - keep only <u>earliest servdat</u> where unique persons are duplicated msp_appnd_asthma &lt;- msp_appnd %&gt;% filter(msp_appnd\$flag == 1) %&gt;% #Step 1   group_by(uniqueid) %&gt;% add_tally(name="count_duplicates") %&gt;% #Step 2   arrange(uniqueid, servdat) %&gt;% distinct(uniqueid, .keep_all = TRUE) #Step 3</pre>



Organisez les données et l'environnement de travail.

Gardez et renommez seulement les variables dont vous aurez besoin plus tard (c.-à-d. *uniqueid*, *servdat*, *count\_duplicates*).

```
#Tidy the data.
msp_appnd_asthma <- msp_appnd_asthma %>%
 select(uniqueid, servdat, count_duplicates) %>%
 rename(msp_servdat_asthma = servdat, msp_count_asthma = count_duplicates)

#clear work environment.
rm(msp_appnd, msp1213, msp1314)
```

4. Rassemblez les données :

Tâche	Code
Chargez <i>reg1314</i> . Combien de déclarants y a-t-il?	<pre>##### Registry Dataset ```{r} #####Load registry file reg1314 &lt;- read_csv(here("Exercise_Day3", "data", "reg1314.csv"))</pre>
Afin de rassembler les données de MSP, de DAD et du registre, de quels	

<p>types de fusion aurez-vous besoin? Comment fusionnerez-vous les trois fichiers (c.-à-d. quelles variables clés) et quelles seront les étapes de cette action? Justifiez.</p> <p>Conseil : Passez en revue les types de jointures dans votre Guide du participant ou de la participante.</p>	
<p>Fusionnez les données d'hospitalisation avec le registre. Combien de dossiers y a-t-il dans le fichier de données fusionné? Combien de déclarants ont été hospitalisés en</p>	<pre>##### Merge the dad dataset with the registry: ```{r} #--- Merge dad to reg asthma_dad_reg1314 &lt;- full_join(reg1314, dad_appnd_asthma, by=c("uniqueid"="uniqueid"))  #--- Frequency of each result in dad_count_asthma asthma_dad_reg1314 %&gt;%   count(dad_count_asthma) ```</pre> <p><i>Indice : Vous pouvez trouver le nombre d'observations dans le fichier fusionné de résultat dans la fenêtre d'environnement. Consultez le fichier fusionné et voyez le nombre d'individus qui ont des données relatives aux hospitalisations pour l'asthme. Le nombre de personnes devrait être égal au nombre de personnes</i></p>

<p>2013-2014 avec un diagnostic d'asthme au cours des deux années précédentes?</p>	<p><i>dans le fichier dad_appnd_asthma. Le fichier fusionné contient également des dossiers relatifs à toutes les autres personnes du registre.</i></p>
<p>Fusionnez les données de visite chez les médecins avec le registre. Combien de dossiers y a-t-il dans le fichier de données fusionné? Combien de déclarants ont consulté leur médecin en 2013-14 et ont reçu une facture codée pour l'asthme au cours des deux années précédentes?</p>	<pre>##### Merge the MSP dataset with the registry: ```{r} #--- Merge msp to reg asthma_msp_reg1314 &lt;- full_join(reg1314, msp_appnd_asthma, by=c("uniqueid"="uniqueid"))  asthma_msp_reg1314 %&gt;%   count(msp_count_asthma) ```</pre> <p><i>Indice : Vous pouvez trouver le nombre d'observations dans le fichier fusionné de résultat dans la fenêtre d'environnement. Consultez le fichier fusionné et voyez le nombre d'individus ayant des données relatives aux visites pour l'asthme. Le nombre de personnes devrait être égal au nombre de personnes dans le fichier msp_appnd_asthma. Le fichier fusionné contient également des dossiers relatifs à toutes les autres personnes du registre.</i></p>



Combinez les ensembles de données d'asthme. Combien d'observations y a-t-il dans le fichier de résultat? Combien de personnes se trouvent dans les deux ensembles de données?

```
Merge Asthma-Registry datasets
library(r)
#create combined asthma datasets
asthma_reg1314 <- full_join(asthma_dad_reg1314, asthma_msp_reg1314, by=c("uniqueid"="uniqueid", "sex"="sex", "dob"="dob"))

table(asthma_reg1314$msp_count_asthma, asthma_dad_reg1314$dad_count_asthma)
count(asthma_reg1314, msp_count_asthma)
count(asthma_dad_reg1314, dad_count_asthma)
...
```

*Remarque : Comme sex et dob proviennent du même fichier, de mauvaises correspondances indiqueraient que des erreurs se sont produites – nous ne voulons pas de colonnes qui se répètent.*

*Indice :*

- *Vous pouvez trouver le nombre d'observations dans le fichier fusionné de résultat dans la fenêtre d'environnement.*
- *En utilisant la commande table, nous pouvons voir un tableau croisé du nombre d'hospitalisations par visite chez le médecin. Le tableau illustrera les personnes qui sont apparues dans les deux fichiers et leurs habitudes d'utilisation des services de santé pour l'asthme dans la période donnée.*
- *En utilisant la commande count, nous pouvons voir le nombre de personnes dans l'ensemble de données divisé par le nombre de visites chez le médecin ou d'hospitalisations pour l'asthme.*



<p>Combien d'individus répondent à la composante d'hospitalisation et de consultation d'un médecin de la définition de cas pour l'asthme? Quelles sont les autres composantes de cette définition?</p>	<pre>##### Process Asthma-Registry dataset ```{r Flag individuals who meet hospitalisation and physician vist portion of the case definition}  # 1H Ever or 2P in 2Y asthma_reg1314 &lt;- asthma_reg1314 %&gt;%   mutate(PH_flag = case_when(dad_count_asthma &gt;= 1 ~ 1,                              msp_count_asthma &gt;= 2 ~ 1,                              TRUE~0))  asthma_reg1314 %&gt;%   count(PH_flag) # table(asthma_reg1314\$PH_flag) ```</pre>
<p>Calculez l'âge du cas en 2013-2014. Quelle date de référence utiliseriez-vous pour ce calcul?</p>	<pre>```{r Calculate age}  asthma_reg1314 &lt;- asthma_reg1314 %&gt;%   mutate(dob = as.Date(dob, "%d%b%Y")) %&gt;% # format as date   mutate(age = as.numeric(difftime(time1 = as.Date("01jul2013", "%d%b%Y"), time2 = dob, units="days"))/365.25)</pre>



Comment identifieriez-vous les personnes qui devraient être incluses dans la définition de cas en fonction de l'âge? N'oubliez pas que vous aurez besoin d'obtenir des données de dénominateur du registre pour vos calculs d'épidémiologie descriptive.

```
#--- Identify asthma cases based on age and case definition
#143 cases
asthma_reg1314 <- asthma_reg1314 %>%
 mutate(popn_1plus = case_when(age >= 1 ~ 1,
 TRUE ~ 0)) %>%
 mutate(asthma_case = case_when(popn_1plus == 1 & PH_flag == 1 ~ 1,
 TRUE~0))
```

Combien d'individus ont répondu à la définition de cas pour l'asthme en 2013-2014?

```
asthma_reg1314 %>%
 count(asthma_case)
````
```



Précisez la date du cas, ou la date du premier service de santé codé pour l'asthme. Pour ce faire, nous créerons une série d'indicateurs en fonction de la date de sortie de l'hôpital et de la visite chez un ou une médecin. Nous pouvons utiliser ces indicateurs pour identifier puis recoder afin de préciser la date à utiliser comme date du cas ou le premier moment où une personne a correspondu à la définition de cas.

Vérifiez vos

```

####{r Process when case met definition}

#identify case date, or the earliest point at which a person met the case definition
asthma_reg1314 <- asthma_reg1314 %>%
  mutate(date_flag = case_when(
    asthma_case ==1 & dad_sepdat_asthma < msp_servdat_asthma ~ 1,
    asthma_case ==1 & dad_sepdat_asthma > msp_servdat_asthma ~ 2,
    asthma_case ==1 & dad_sepdat_asthma == msp_servdat_asthma ~ 3,
    asthma_case ==1 & is.na(dad_sepdat_asthma) == TRUE ~ 4,
    asthma_case ==1 & is.na(msp_servdat_asthma) == TRUE ~ 5,
    TRUE~0))

asthma_reg1314 <- asthma_reg1314 %>%
  mutate(asthma_casedate = case_when(
    date_flag == 1 ~ dad_sepdat_asthma,
    date_flag == 2 ~ msp_servdat_asthma,
    date_flag == 3 ~ msp_servdat_asthma,
    date_flag == 4 ~ msp_servdat_asthma,
    date_flag == 5 ~ dad_sepdat_asthma))

#count asthma case flags
asthma_reg1314 %>%
  count(asthma_case)

#count date flags
asthma_reg1314 %>%
  count(date_flag)

#compare case and date flags
asthma_reg1314 %>%
  count(asthma_case, date_flag)

#compare casedate and case flags
asthma_reg1314 %>%
  count(asthma_casedate, date_flag)

```



| | |
|---|---|
| énoncés logiques au moyen de tableaux et de tableaux croisés (c.-à-d. voir les fonctions <i>count</i> à droite et dans le fichier rmd pour cet exercice). | |
| Le nombre de cas d'asthme reflète-t-il les cas incidents?
Pourquoi ou pourquoi pas? | |
| <p>Rangez le produit final et organisez l'espace de travail au besoin.</p> <p>Gardez seulement les variables dont vous aurez besoin plus tard :
<i>uniqueid, dob,</i></p> | <pre> #---Tidy data str(asthma_reg1314) #review structure of the data #select variables to keep asthma_reg1314 <- asthma_reg1314 %>% select(uniqueid, dob, age, sex, asthma_case, asthma_casedate, popn_1plus) #clear out work environment rm(asthma_dad_reg1314, asthma_msp_reg1314, dad_appnd_asthma, msp_appnd_asthma, reg1314) </pre> |



age, sex,
asthma_case,
asthma_casedate
 , et *popn_1plus*.

5. Analysez les données

| Tâche | Code |
|---|--|
| Quel était le taux brut de prévalence de l'asthme en 2013-2014? | <pre>#### Data Analysis '''{r Calculate crude prevalence} #crude prevalence: #add up n cases and n population asthma_overall <- asthma_reg1314 %>% summarize(cases = sum(asthma_case), popn = sum(popn_1plus)) asthma_overall <- asthma_reg1314 %>% summarize(cases = sum(asthma_case), popn = sum(popn_1plus)) %>% mutate(crude_rate=(cases/popn)*100000) print(asthma_overall) '''</pre> |



Calculez le nombre de cas d'asthme par âge en créant tout d'abord des catégories d'âge et en classant les cas selon ces catégories. Quels groupes d'âge utiliserez-vous?

Nous allons calculer les taux et créer une figure dans la prochaine étape.

```

```{r Calculate age-specific prevalence}
#age-specific prevalence
asthma_reg1314 <- asthma_reg1314 %>%
 mutate(agegrp = case_when(
 age >= 1 & age <= 9 ~ "<10",
 age >= 10 & age <= 19 ~ "10-19",
 age >= 20 & age <= 29 ~ "20-29",
 age >= 30 & age <= 39 ~ "30-39",
 age >= 40 & age <= 49 ~ "40-49",
 age >= 50 & age <= 59 ~ "50-59",
 age >= 60 & age <= 69 ~ "60-69",
 age >= 70 & age <= 79 ~ "70-79",
 age >= 80 & age <= 89 ~ "80-89",
 age >= 90 ~ "90+"))

#tabulate age groupings
asthma_reg1314 %>%
 count(agegrp)

#tabulate cases per age grouping
asthma_reg1314 %>%
 count(agegrp, asthma_case) %>%
 pivot_wider(names_from = asthma_case, values_from = n)
```

```



Calculez et tracez les taux d'asthme par âge en utilisant les mêmes groupes d'âge de 10 ans que vous avez créés à l'étape précédente.

```
```{r Age-specific asthma case rate}

Summarize cases and population per age range
asthma_agespecific <- asthma_reg1314 %>%
 group_by(agegrp) %>%
 summarize(cases = sum(asthma_case), popn = sum(popn_1plus), .groups = "drop") %>%
 filter(!is.na(agegrp))

Calculate rates
asthma_agespecific <- asthma_agespecific %>%
 mutate(agespecific_rate=(cases/popn)*100000)

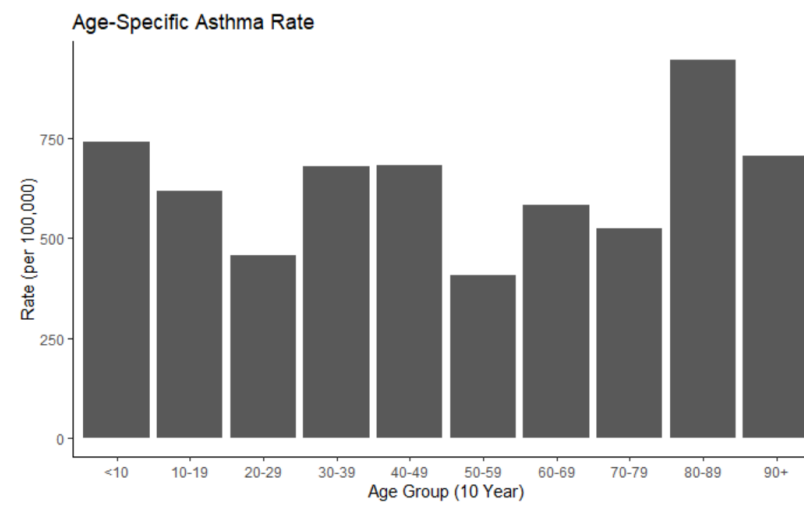
Basic Plot
agespec_barchart <- ggplot(data=asthma_agespecific, aes(x=agegrp, y=agespecific_rate)) +
 geom_bar(stat="identity")

Add simple options to the plot
agespec_barchart <- agespec_barchart +
 labs(title="Age-Specific Asthma Rate", x="Age Group (10 Year)", y = "Rate (per 100,000)") +
 theme_classic()

#Call plot
agespec_barchart

```
```





Calculez et tracez les taux d'asthme par âge et par sexe.

```

```{r Age and sex-specific asthma rates}
Age and sex-specific prevalence

#add sex variable to the dataset
asthma_reg1314 <- asthma_reg1314 %>%
 mutate(sexchar = case_when(asthma_reg1314$sex == 1 ~ "Female",
 asthma_reg1314$sex == 2 ~ "Male",
 TRUE~"Unknown"))

#summarize based on age and sex
asthma_as_specific <- asthma_reg1314 %>%
 group_by(agegrp, sexchar) %>%
 summarize(cases = sum(asthma_case), popn = sum(popn_1plus), .groups = "drop") %>%
 filter(!is.na(agegrp))


#calculate rates
asthma_as_specific <- asthma_as_specific %>%
 mutate(as_specific_rate=(cases/popn)*100000)

Plot age and sex-specific results
agesexspec_barchart <- ggplot(data=asthma_as_specific, aes(fill=sexchar, x=agegrp, y=as_specific_rate)) +
 geom_bar(position="dodge", stat="identity") +
 labs(title="Age- and Sex-Specific Asthma Rate",
 x="Age Group (10 Year)",
 y = "Rate (per 100,000)",
 fill="Sex") +
 theme_classic()

#Call plot
agesexspec_barchart
```

```

| | <div><div>Age- and Sex-Specific Asthma Rate</div><table><thead><tr><th>Age Group (10 Year)</th><th>Female (per 100,000)</th><th>Male (per 100,000)</th></tr></thead><tbody><tr><td><10</td><td>380</td><td>1050</td></tr><tr><td>10-19</td><td>320</td><td>1150</td></tr><tr><td>20-29</td><td>220</td><td>650</td></tr><tr><td>30-39</td><td>500</td><td>850</td></tr><tr><td>40-49</td><td>320</td><td>1050</td></tr><tr><td>50-59</td><td>480</td><td>450</td></tr><tr><td>60-69</td><td>580</td><td>950</td></tr><tr><td>70-79</td><td>750</td><td>550</td></tr><tr><td>80-89</td><td>950</td><td>1150</td></tr><tr><td>90+</td><td>680</td><td>750</td></tr></tbody></table></div> | Age Group (10 Year) | Female (per 100,000) | Male (per 100,000) | <10 | 380 | 1050 | 10-19 | 320 | 1150 | 20-29 | 220 | 650 | 30-39 | 500 | 850 | 40-49 | 320 | 1050 | 50-59 | 480 | 450 | 60-69 | 580 | 950 | 70-79 | 750 | 550 | 80-89 | 950 | 1150 | 90+ | 680 | 750 |
|---|--|---------------------|----------------------|--------------------|-----|-----|------|-------|-----|------|-------|-----|-----|-------|-----|-----|-------|-----|------|-------|-----|-----|-------|-----|-----|-------|-----|-----|-------|-----|------|-----|-----|-----|
| Age Group (10 Year) | Female (per 100,000) | Male (per 100,000) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| <10 | 380 | 1050 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 10-19 | 320 | 1150 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 20-29 | 220 | 650 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 30-39 | 500 | 850 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 40-49 | 320 | 1050 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 50-59 | 480 | 450 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 60-69 | 580 | 950 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 70-79 | 750 | 550 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 80-89 | 950 | 1150 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 90+ | 680 | 750 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Assurez-vous que les détails pertinents sont inclus dans le titre du script. | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Tricotez le document (<i>Knit</i>).

Remarque : Le document sera enregistré dans votre dossier de scripts ou dans le dossier que vous avez précisé au départ. | <div> Knit</div> | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Voici un exemple de ce à quoi ressemblera votre extrant lorsque vous l'ouvrirez dans votre navigateur Web :

R for Public Health Investigations: Day 3

Script Information:

Purpose: Exercise in advanced data management training for epidemiologists in R by using mock admin datasets to identify cases of Asthma

Author: J. Stares

Date: 26-Oct-2020

Last modified by: B. Hetman

Date last modified: 2020-11-15

Notes:

- Mock admin SAS datasets processed in R and Stata in 2017
- Training module location: Dropbox
- Case definition - Asthma: 1 year and older - 1H Ever or 2P in 2Y <https://www.canada.ca/en/public-health/services/publications/canadian-chronic-disease-surveillance-system-factsheet.html>

Setup:

```
require("knitr")
```

```
## Loading required package: knitr
```

```
## setting working directory
opts_knit$set(root.dir = "C:/IntroToR/Exercise_Day3/data/")
opts_chunk$set(fig.path = "C:/IntroToR/Exercise_Day3/data/")
```

```
#Load libraries
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## -- Attaching packages ----- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2    v purrr   0.3.4
## v tibble  3.0.4    v dplyr  1.0.2
```



6. Pourquoi avons-nous construit le produit final sous la forme d'un tableau de registre indiquant les numérateurs et dénominateurs pour l'asthme? Y a-t-il un meilleur moyen?

7. Quelles sont les possibles limites de l'approche consistant à utiliser les dossiers médicaux administratifs pour identifier les cas de maladie? Quels sont les avantages de cette approche?

8. Que pensez-vous de la façon dont cet exercice a été structuré en comparaison avec les exercices des deux premiers jours (c.-à-d. plusieurs courts scripts c. un long script, des bouts de code c. un code divisé en lignes et une structure de répertoire de travail)?

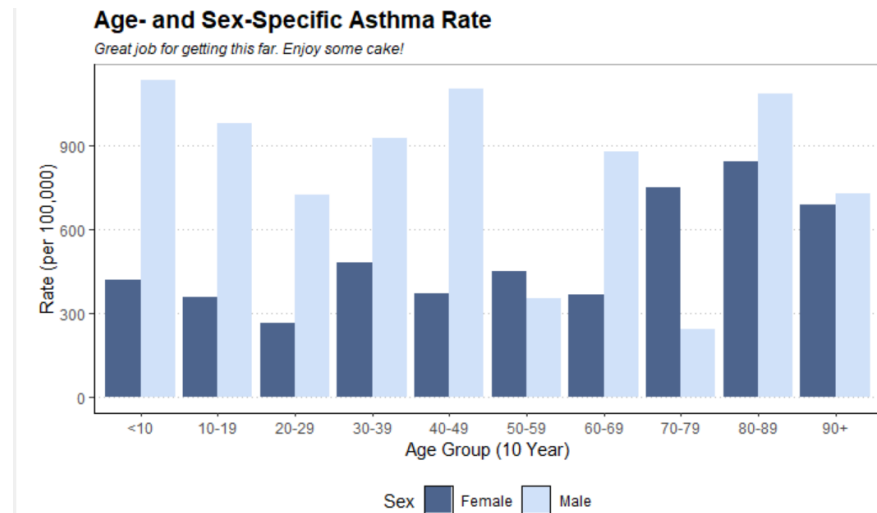


9. Contenu avancé en prime! Démonstration des fonctions définies par l'utilisateur :

| Tâche | Code |
|--|---|
| <p>Réviser le matériel suivant sur les fonctions dans R :</p> <p>https://www.tutorialspoint.com/r/r_functions.htm (EN)</p> <p>https://towardsdatascience.com/your-first-user-defined-function-in-r-1eedc634ead4 (EN)</p> <p>https://swcarpentry.github.io/r-novice-inflammation/02-func-R/ (EN)</p> <p>https://datasciencebeginners.com/2018/11/02/10-user-defined-functions-in-r/ (EN)</p> <p>Examinez et exécutez le code de démonstration fourni ici dans votre fichier .rmd.</p> <p>Quelles sont les composantes de base d'une fonction définie par l'utilisateur? Quelle partie du code définit le nom de la fonction? Y a-t-il des arguments intrants à cette fonction? Que fait cette fonction?</p> | <pre>##### So many plotting options! ### {r Demonstrating a custom function} # Create your own function to format plot styles! theme_awesome <- function(){ theme_classic() + theme(line = element_line(color="black"), panel.background = element_rect(fill = "white", colour = "darkgrey"), panel.grid.major.y = element_line(colour = "grey", linetype = 3, size = 0.5), plot.background = element_rect(fill="white"), strip.background = element_rect(fill = "white"), legend.background = element_rect(fill = "white"), legend.key = element_rect(fill = "white"), legend.position = "bottom", plot.title = element_text(size = 14, family = "Helvetica Neue", face = "bold"), plot.subtitle = element_text(size = 9, family = "Helvetica Neue", face = "italic"), axis.title.x = element_text(), text = element_text(size = 11),) }</pre> |

Appliquez votre fonction personnalisée à l'objet `agesexspec_barchart` créé plus tôt.

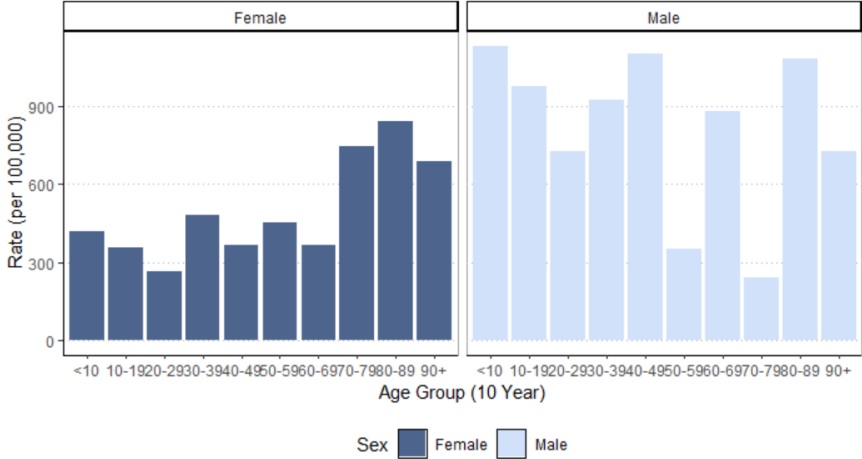
```
agesexspec_barchart +  
  theme_awesome() +  
  scale_fill_manual(values = c("#4D648D", "#D0E1F9")) +  
  scale_color_manual(values = "darkgrey") +  
  labs(subtitle = "Great job for getting this far. Enjoy some cake!")  
...
```



Tricotez le document (*Knit*).



10. Contenu avancé en prime! Mieux vivre grâce aux technologies modernes - démonstration des facettes et des boucles :

| Tâche | Code |
|--|--|
| <p>Passez en revue le tutoriel suivant sur les facettes pour <i>ggplot</i> dans R.</p> <p>http://www.sthda.com/english/wiki/ggplot2-facet-split-a-plot-into-a-matrix-of-panels
(EN)</p> <p>Examinez et ajoutez le code de démonstration fourni ici dans votre fichier rmd.</p> <p>Qu'est-ce qu'une facette? Que réalise-t-elle plus précisément? En quoi les facettes sont-elles applicables à votre travail?</p> | <pre>##### Demonstrating facets and a simple loop: {r Facetting on sex} # What if we wanted to compare sexes by themselves? agesexspec_barchart + theme_awesome() + scale_fill_manual(values = c("#4D648D", "#D0E1F9")) + scale_color_manual(values = "darkgrey") + labs(subtitle = "Great job for getting this far. Enjoy some cake!") + facet_wrap(~sexchar) }</pre> <p>Age- and Sex-Specific Asthma Rate
<i>Great job for getting this far. Enjoy some cake!</i></p>  <p>Rate (per 100,000)</p> <p>Age Group (10 Year)</p> <p>Sex ■ Female ■ Male</p> |

Passez en revue le tutoriel suivant sur les boucles dans R.

<https://www.r-bloggers.com/2015/12/how-to-write-the-first-for-loop-in-r/> (EN)

Examinez et exécutez le code de démonstration fourni ici.

Quelles sont les composantes de base pour une boucle? Quel est le but de l'énoncé « i in unique »? À quoi sert le bloc de code? Quelles fonctionnalités supplémentaires la boucle fournit-elle? En quoi les boucles sont-elles applicables à votre travail?

```
```{r A simple loop}
What if we wanted a separate chart for each age group?
The following code will iterate through each age group created earlier,
and produce a unique chart for that age category only; then save that chart
in your output directory.

Check where your working directory is currently set to:
getwd()

for (i in unique(asthma_as_specific$agegrp)){
 # Create titles for the individual plots based on the age group
 title <- paste0("Asthma rate among age group ", i)
 # Special note: R will fail if the file name includes non-alpha-numeric characters
 # We use str_replace_all to remove these
 title <- str_replace_all(title, "[^[:alnum:]]", " ")

 # subset the data we need to plot
 loop_plot <- asthma_as_specific %>% filter(agegrp==i)

 # Now we plot and save the individual files to our working directory
 ggplot(data=loop_plot, aes(fill=sexchar, x=agegrp, y=as_specific_rate))+
 geom_bar(position="dodge", stat="identity")+
 labs(title="Age- and Sex-Specific Asthma Rate",
 x=paste0("Age (years)"),
 y="Rate (per 100,000)",
 fill="Sex") +
 theme_awesome() +
 scale_fill_manual(values = c("#4D648D", "#D0E1F9")) +
 scale_color_manual(values = "darkgrey") +
 labs(title = title,
 subtitle = "Great job for getting this far. Enjoy some cake!")

 ggsave(path = output_folder, filename = paste0(title, ".png"), device = "png")
}
Check your output directory for the exported plots!

...`
```

Tricotez le document (*Knit*).



**FONCTIONS VEDETTES**

En guise de référence, nous vous avons fourni une liste des principales fonctions utilisées dans le cadre de cet exercice. Les fonctions en gras sont les plus grandes vedettes de la journée.

<b>Fonctions vedettes</b>	
<b><i>bind_rows()</i></b>	<b><i>full_join()</i></b>
<b><i>as.numeric()</i></b>	<b><i>trunc()</i></b>
<b><i>str_sub()</i></b>	<b><i>difftime()</i></b>
<b><i>add_tally</i></b>	<b><i>is.na()</i></b>
<b><i>table()</i></b>	