Dans le but de faciliter le travail d'extraction, on utilise les fonctions R *setwd* et *getwd* qui permettent respectivement de pointer sur un répertoire et de définir un répertoire. L'utilisation du double point ('..') dans l'argument de la fonction setwd permet de pointer un répertoire au-dessus du répertoire actuel. Ce nouveau répertoire est ainsi affecté à la variable path, qui seras réutilisé plus tard.

Cette section est certainement la plus importante de tous, elle vise à faire un traitement adéquat et pertinent des données afin de pouvoir les réutiliser facilement dans les sections suivantes. Une mauvaise application des concepts d'extraction, traitement et visualisation des données peut entrainer une interprétation inadéquate de la situation.

Le fichier source des données étant disponible en ligne on effectue l'extraction directement via leurs adresses à l'aide de la fonction R *read.csv*.On utilise trois bases de données, la première sur les aéroports mondiaux [**?**], la seconde sur les trajets aériens mondiaux [**?**] et la dernière sur les compagnies aériennes mondiales [**?**]. La présentation des données ne comporte pas de titre, l'argument **header** a donc été mis à FALSE, et l'attribuons manuellement des titres des colonnes est effectuées plus loin.

Afin de facilité l'analyse des données qualitatives, il est beaucoup plus facile de représentées celle-ci à l'aide de variable *factors* [**?**]. La transformation des données en factors permet l'association d'une chaîne de caractères à un nombre entier. Autrement dit, chaque valeur qualitative est *entreposé* à l'aide d'un entier.

De plus, on remarque que dans la base de données les valeurs absentes sont représentées par \\n ou ' '. Cette donnée n'a pas un sens significatif. On lui attribut plutôt la valeur *NA* à l'aide du dernier argument de la fonction.

L'attribution manuelle des noms des différentes colonnes des données s'éffectue à l'aide de la fonction R *colnames*. Il devient ainsi très facile de filtrer les aéroports canadien à l'aide de l'objet *country* dans *aiports*. Autrement dit, on copie dans une variable toutes les entrées (lignes) dont le pays dans la colonne *country* correspond à *Canada*.

La visualisation des données est une étape cruciale dans l'interprétation de celle-ci. De nombreuses fonctions R permettent de sortir plusieurs informations pertinentes sur les données. Tel que

- *View*[**?**] qui permet de visualiser le *data frame* R dans un onglet à part ;
- *head* [**?**] qui permet de visualiser dans la console les premières entrées ;
- *summary*[**?**] qui permet de visualiser différentes informations sur les données quantitatives et qualitatives telles que les quartiles, la moyenne, le maximum et le minimum pour les données quantitatives et la fréquence des observations de chacune des données qualitatives.

Par la suite, il devient possible de sortir différentes informations spécifiques de chaque variable. Par exemple, le nombre d'aéroports par ville a été extrait et présenter à l'aide de la fonction R *table* [**?**] L'utilisation de la fonction R *as.character*[**?**] permet de convertir les *factors* en chaîne de caractères.

La prochaine étape consiste à nettoyer la pertinence des données ainsi que

le remplissage des données absentes. On observe que les relations (colonne) *typeAirport*, *country* et *Source* ne sont pas pertinentes à notre situation puisque nous observons uniquement les aéroports canadiens. Elles seront retirées à l'aide du code suivant,

On cherche maintenant les données absentes, on observe à l'aide de la fonction *summary* que 27 aéroports ne comportent pas leur indicatif IATA, que l'on peut visualiser ainsi

Deux solutions sont possibles concernant cette situation, étant donné que seulement 18% des IATA sont manquants, il pourrait être possible d'ignorer et de retirer les données. Par contre, à l'aide de l'indicatif ICAO il est possible de déterminer l'indicatif IATA. En effet, le ICAO correspond à un caractère unique par pays concaténer avec le IATA. À l'aide de cette information, il est possible de retrouver les informations manquantes facilement

La fonction R *substr* [**?**] permet de faire un découpage de la chaîne de caractère.

Finalement, on peut aussi observer que 52 aéroports ne comportent pas de fuseau horaire, deux options sont envisageables pour résoudre la problématique.

1. Étant donné que les fuseaux horaires sont déterminés par des positions géographiques, il est possible de déduire les informations manquantes à partir des aéroports à proximité
2. Utiliser des outils de cartographie pour retrouver les vrais fuseaux horaires.

La seconde solution a été adoptée, elle peut sembler complexe, mais avec les bons outils elle s'avère beaucoup plus simple et efficace. Cette partie nécessite l'installation de deux paquetages R soit *sp* [**?**] et *rgdal* [**?**].

On constate qu'il y a absence de la province de chacun des aéroports, alors pour le calcul des taxes cette information est nécessaire. Il est donc possible à l'aide des méthodes de cartographie vue précédemment ainsi qu'avec les données sur les frontières des provinces [**?**] de déterminer la province de l'aéroport. L'installation du paquetage R *sqldf* [**?**] est nécessaire pour l'exécution de cette partie. En effet, on applique les mêmes concepts de cartographie au territoire canadien afin de *quadriller* les provinces. Par la suite, à l'aide du langage *sql* [**?**] on [...]

Ainsi, on obtient des données complètes pour l'ensemble des aéroports canadiens. Certaines informations sont toutefois devenues obsolètes pour la suite de l'étude de cas, en effet les relations (colonnes) *timezone*, *DST* et *city* ne sont plus pertinentes. De plus, la relation *tzformat* doit être retirée car elle sera remplacé par la relation *tzmerge* créée par la requête *sql* précédente. Afin de renommer les nouvelles colonnes ajouter par la requête, on utilise la fonction R *rename* [**?**, **?**] du paquetage *plyr* [**?**].

On s'intéresse maintenant aux deux dernières bases données pertinentes pour l'étude de cas, les voies aériennes et les compagnies aériennes. Tout d'abord, on sélectionne à l'aide d'une requête *sql* les voies aériennes canadiennes. On analyse les informations présentes pour les voies aériennes et on observe que seulement 2 trajets ne sont pas des vols directs. Pour des fins de simplifications, tous les

vols seront considérés comme des vols directs. Ainsi, les colonnes *codeshare* et *stops* sont inutiles et elles sont retirées.