Capture1

La première chose qu’on remarque en observant les données est que toutes les informations relatives à un individu sont écrites dans une cellule. En particulier, la première cellule de la première ligne contient les en-têtes du tableau. En la sélectionnant, nous pouvons voir que les variables du tableau sont :

* PassengerId
  + Un indice identifiant chaque observation.
* Survived
  + Une variable binaire indiquant si le passager a survécu au naufrage. Un 1 marque que le passager a survécu et un 0 indique qu’il est décédé.
* Pclass
  + Indique si le passage était en première, deuxième ou troisième classe.
* Name
  + Le nom du passager.
* Sex
  + Le sexe du passager.
* Age
  + L’âge du passager.
* SibSp
  + Le nombre de frère, sœur, demi-sœur et demi-frère ou de conjoint à bord du navire avec le passager.
* Parch
  + Le nombre de parent ou d’enfant à bord du Titanic avec le passager.
* Ticket
  + Le numéro de série du billet acheté par le passager.
* Fare
  + Le prix payé par le passage pour son billet.
* Cabin
  + Le numéro de la cabine loué par le passage pour le voyage.
* Embarked
  + Le lieu d’embarquement du passage. C = Cherbourg, Q = Queenstown et S = Southampton

Donc, chaque cellule représente un individu et pour pouvoir faire l’analyse, nous devrons ventiler l’information sur plusieurs colonnes.

Pour ce faire, on sélectionne la colonne contenant toute l’information et on utilise la fonction Text to Columns qui se trouve sous l’onglet DATA du ruban principal.

Capture2

En cliquant sur cette icône, une fenêtre s’ouvre et vous devez faire de différents choix qui influenceront la conversion des cellules en colonnes. En observant les cellules de plus près, on s’aperçoit que les informations dans chaque cellule sont séparées par des virgules. Donc à la première étape, nous devons cocher le bouton delimited, puisqu’un caractère, la virgule, sépare chaque élément de la cellule.

Capture3

Ensuite, « virgule » est sélectionné comme séparateur à la deuxième étape de la conversion des cellules en colonnes, puis nous pouvons cliquer sur terminer.

Capture4

Maintenant, après quelques modifications aux dimensions des cellules pour rendre le tableau plus facile à lire, les données ressemblent à cela :

Capture6

Avant de commencer à manipuler les données, il est important de vérifier si le tableau qui les contient facilite l’analyse en étant clair et bien organisé. La première chose à observer est si chaque observation est représentée dans le tableau par une ligne. Puisque nous nous intéressons aux passagers du Titanic et leur chance de survivre, chacune de nos observations concerne des personnes et chaque ligne doit représenter un individu en particulier. C’est le cas dans ce tableau. Voici un exemple de tableau dont les lignes devraient être ventilées. Comme vous voyez, ces lignes contiennent les informations de deux observations.

Capture7

La deuxième chose à vérifier est si chaque colonne contient la valeur d’une seule variable. Encore une fois, c’est le cas pour notre tableau, car chaque cellule représente la mesure d’une caractéristique pour un individu en particulier. Un exemple de tableau qui ne respecterait pas ce principe est un tableau ou une variable a été codée en dummy variable.

Capture8

La dernière chose à vérifier est que notre tableau contient des informations sur une seule unité d’observation. Notre tableau contient seulement des données sur les passagers du Titanic, alors il contient une seule unité d’observation. En pratique, avoir des données sur le bateau lui-même comme le nombre de canots de sauvetages, leur capacité maximale et leur positionnement, pourrait nous aider à créer un meilleur modèle, mais ajouter ces données à ce tableau nous compliquerais considérablement la tâche. Il serait alors préférable de créer un tableau séparé.

Capture5

En regardant attentivement les données, il est clair que cette banque de données n’est pas complète. Par exemple, la variable Age de la 7eme ligne est manquante. On peut facilement compter combien de cellule vide est contenu dans chaque colonne en utilisant la fonction COUNTBLANK et en sélectionnant la colonne comme champs de la fonction

Capture9

Nous voyons que les trois variables Age, Cabin et Embarked contiennent des cellules vides. À ce point il faut se demander comment?