2021-2022

RAPPORT DE PROJET 1A

Développement d'un outil de traitement de données réelles par les statistiques

Tuteur : Dominique Benmouffek

Auteurs : Rim M'hamdi - Elie Thellier - Hugo Cocher- Léon Neis - Youssef Mabrouk

Introduction

Le traitement de données est une série de processus permettant d'extraire de l'information ou de produire du savoir à partir de données brutes. Il est composé d'une étape de collecte, puis de la préparation, de l'importation et du traitement des données. Enfin, il sert à interpréter un jeu de données par l'utilisation de statistiques ou de graphiques.

Même s'il existe déjà des outils permettant de traiter les données par les statistiques, notre projet « Développement d'un outil de traitement de données par les statistiques » vise à créer une interface similaire à celles existantes en nous faisant réfléchir sur les limites du nettoyage des données et sur le problème de représentation de données inconnues.

Nous souhaitons remercier Mme. Benmouffek pour son aide quant aux décisions relatives au projet.

Table des matières

Résumé

Après 9 mois, notre travail sur le projet « Plateforme d'analyse statistique de données réelles » touche à sa fin.

Nous avons commencé notre travail lors du premier semestre en suivant deux formations à distance. La première est le MOOC gestion de projet imposé à tous les projets et qui nous a permis entre autres de facilité la distribution des tâches en nommant un chef de projet qu'est Youssef et en formant deux équipes ; une pour la création de l'interface qui était composée de Leon et Rim et une autre, composée d'Elie et Hugo, chargée de trouver les dataset à utiliser et de les nettoyer afin de pouvoir les utiliser à bon escient. Nous avons également suivi une formation sur open class room qui avait pour objectif de nous donner les bases du langage R et qui a été très utile pour la seconde partie du projet.

Cette formation à distance nous a permis, lors du second semestre de nous lancer sur la partie plus informatique du projet qu'était celle du codage de la plateforme ainsi que de la sélection des dataset que nous allions utiliser. La partie codage était la plus importante puisque c'est le squelette de notre interface. Le but était de pouvoir avoir des histogrammes, des nuages de points et de pouvoir configurer ça en entrant le dataset choisi, cette partie a été assez compliqué car assez technique du point de vue informatique parlant et nous a pris beaucoup de temps car nous étions tous les cinq très exigeants sur ce sur quoi nous voulions aboutir. Il a également fallu choisir les bases de données que nous voulions utiliser ce qui n'a pas été tâche facile puisque nous avons décidé de changer d'avis au milieu du projet. En effet, en début de projet nous avions comme idée de se concentrer sur des bases de données proche du sport et notamment concernant le football. Mais nous avons changé d'avis afin d'avoir un choix plus large de bases de données à étudier. Le problème est que cela nous limitait dans ce les résultats que nous voulions avoir en sorti.

Pour conclure, ce projet fut très enrichissant pour tous les cinq. Il nous a permis d'apprendre un nouveau langage informatique et de progresser dans notre capacité à s'organiser et à communiquer.

Mots clefs :

Traitement de données réelles, Développement en langage R, Interface graphique avec Shiny, Analyse statistique, Nettoyage des données.

Summary

After 9 months, our work on the project "Platform for statistical analysis of real data" is coming to an end.

We started our work during the first semester by following two distance learning courses. The first one is the MOOC project management imposed to all projects and which allowed us, among other things, to facilitate the distribution of tasks by appointing a project leader, Youssef, and by forming two teams; one for the creation of the interface, which was composed of Leon and Rim, and another one, composed of Elie and Hugo, in charge of finding the datasets to be used and of cleaning them in order to be able to use them for good purposes. We also followed a training course from Openclassroom which aimed to give us the basics of the R language and which was very useful for the second part of the project.

This distance learning course allowed us, during the second semester, to start the more computerized part of the project which was the coding of the platform as well as the selection of the datasets we were going to use. The coding part was the most important since it is the core of our interface. The aim was to be able to have histograms, scattered graphs and to be able to configure this by entering the chosen dataset, this part was quite complicated because it was quite technical from a computer point of view, and it took us a lot of time because all five of us were very demanding on what we wanted to achieve. We also had to choose the databases we wanted to use, which was not an easy task as we decided to change our minds in the middle of the project. Indeed, at the beginning of the project we had the idea to focus on databases related to sports and especially football. But we changed our mind in order to have a wider choice of databases to study. The problem was that this limited the results we wanted to get out.

To conclude, this project was very rewarding for all five of us. It allowed us to learn a new computer language and to progress in our ability to organize and communicate.

Key words:

Real data processing, Development in R language, Graphical interface with Shiny, Statistical analysis, Data cleaning.

Partie I : Enjeux et appropriation du projet

a - Intitulé du projet et enjeux :

Le projet nous a été présenté en début d'année comme suit :

16. Développement d'un outil de traitement de données par les statistiques

Objectifs généraux

- Formation au Langage R.
- Développement d'une application qui prend des données issues d'Internet, calcule des données statistiques caractéristiques et affiche des critères pertinents

Descriptif du projet

L'outil doit proposer une interface graphique qui permet de sélectionner des données massives issues d'Internet, avec une structure connue. L'utilisateur pourra alors choisir de calculer des résultats statistiques fondés sur ces données, et ensuite d'afficher graphiquement des critères pertinents qui pourront mettre en évidence des caractéristiques intéressantes sur ces données.

Ce projet permettra à l'équipe qui le choisira d'apprendre un nouveau langage (R), de travailler sur des données réelles et de les exploiter avec des outils statistiques pour afficher finalement des graphiques explicitant les données de départ.

Le début de l'année sera dédié à l'auto-formation en langage R. Puis, l'application sera imaginée pour exploiter des données par les statistiques.

Un expert scientifique pourra être consulté en la personne de Rémi Peyre, enseignant d'Inférence Statistique en S6.

Figure 1 : Extrait du document de présentation des projets 1A

b - Prise en main du projet et déroulement du premier semestre :

1- Organisation de l'équipe :

Pour mener à bien ce projet, nous avons organisé l'équipe de 5 étudiants en deux pôles principaux. Cette division du travail s'est révélée très efficace. Ponctuellement, des réorganisations ont eu lieu.

L'équipe est la suivante :

Youssef - Chef de projet
Elie et Hugo - Pôle traitement de données
Rim et Léon - Pôle développement de l'interface graphique

2- Objectifs du premier semestre :

Le premier semestre avait pour but de mettre en marche le projet, de se l'approprier et de monter en compétence afin de le mener à bien.

Nous avons donc suivi deux MOOC de formation d'une part à la gestion de projet et d'autre part au codage en langage R.
Nous avons cependant été confronté à une première difficulté : le sujet étant proposé pour la première fois, ses enjeux étaient encore vagues et un long travail d'analyse était nécessaire afin de débroussailler le sujet et d'en éclaircir les enjeux.

Figure 2 : Logo du MOOC de Gestion de Projet

Figure 3 : Aperçu du Mooc d'introduction au Langage R

Le suivi d'un MOOC et de vidéos puis la réalisation d'exercices d'application et de mini-projets nous a permis de valider ces compétences pour le premier jalon du projet : la soutenance de mi-parcours.

3- Conclusions de la soutenance de mi-parcours :

L'année se déroulant en deux semestres, une soutenance de projet de mi-parcours a eu lieu mi-janvier.

Nos faiblesses établies dans la matrice SWOT (voir figure X) du projet au premier semestre nous ont effectivement porter préjudice. Effectivement, l'orientation qu'avait pris notre projet a été remis en question par le jury. L'ambition d'en faire un outil de traitement de données avec un enjeu social permettant de prendre du recul par rapport aux informations que l'on trouve dans les médias, ne l'a pas convaincu.
Nous avons donc du remettre en perspective notre travail d'analyse qui était resté assez scolaire car nous voulions à tout prix répondre aux objectifs pré-définis qui nous avaient été présenté. En prenant conscience que le travail de projet nous offrait de plus grandes libertés quant à la manière de l'aborder, nous avons à l'issue du premier semestre réorienter le projet.

Partie II : Réorientation du projet :

a - Etat des lieux du projet:

L'objectif du projet est donc de développer un objet de traitement des données par la statistique en langage R, langage de programmation destiné aux statistiques et à la science des données.
Nos compétences nous ont imposés plusieurs options :

- Traiter un grand nombre de data set pour en faire un outil assez universel quitte à afficher un nombre restreint de données statistiques intéressantes.

Figure 4: Matrice SWOT du projet

- Réduire le nombre de data set possiblement analysable mais l'analyser sous tous les aspects possibles et faire apparaitre des corrélations intéressantes.

Nous avons finalement décidé après concertation avec notre responsable de projet de choisir la première option. En effet, nous avons l'ambition qu'à terme l'outil conçu soit utilisable par un maximum d'utilisateurs extérieurs. Nous avons donc donné la priorité à cet aspect, quitte à developer les capacités techniques de notre interface au fur et à mesure le but étant de bâtir notre interface sur de bonnes fondations.

Après avoir fait un état des lieux et un brainstorming de ce qu'il nous reste à faire et des idées à prendre en compte dans la réorientation du projet, nous avons réalisé la matrice SWOT suivante :

	Facteurs Positifs	**Facteurs Négatifs**
Diagnostic Interne	**FORCES** • Compétence en informatique • Répartition des tâches • Facilité d'utilisation de la plateforme • Nombre de possibilités pour la fonctionnalité de la plateforme	**FAIBLESSES** • Communication • Manque de temps • Complexité des codes
Diagnostic Externe	**OPPORTUNITÉS** • MOOC gestion de projet • MOOC initiation au langage R • Diversité des data set à utiliser	**MENACES** • Propreté des dataset • Problèmes avec Rstudio • Nombre important de plateforme qui ont un cahier des charges similaire

b - Objectifs du semestre :

Diagramme de Gantt

Nom	Date de début	Date de fin
Semestre 1	15/09/2021	23/02/2022
Appropriation et définition du projet	15/09/2021	29/09/2021
Répartition des tâches	15/09/2021	22/09/2021
Choix du module pour les interfaces	23/09/2021	29/09/2021
Choix du langage de programmation : R + Shiny	23/09/2021	29/09/2021
MOOC R + exercices R	30/09/2021	15/12/2021
Spécialisation en Shiny	16/12/2021	23/02/2022
Spécialisation traitement de données	16/12/2021	23/02/2022
Préparation soutenance mi-parcours	03/01/2022	20/01/2022
Soutenance mi-parcours	21/01/2022	21/01/2022
Semetre 2	24/02/2022	17/06/2022
Choix des données définitives	24/02/2022	09/03/2022
Traitement des données et questionnements su...	10/03/2022	22/04/2022
Création de l'interface graphique = du script	24/03/2022	18/05/2022
Finalisations et rapport	19/05/2022	07/06/2022
Préparation soutenance finale	08/06/2022	17/06/2022
Soutenance finale	20/06/2022	20/06/2022

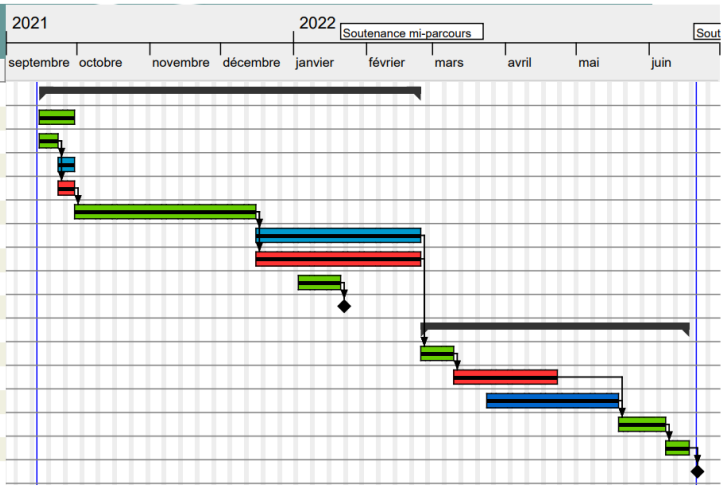


Figure 4 : Diagramme de Gantt

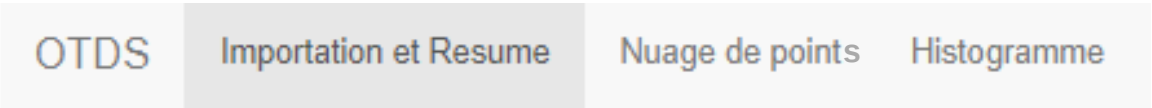
La soutenance de mi-parcours nous a permis de redéfinir les objectifs. Après des questionnements sur la représentation des données et sur le nettoyage de celles-ci, nous avons codé un script R permettant le traitement de données réelles présentées sous forme de dataset (format txt ou csv). Il affiche le résumé statistique des variables, permet de représenter une variable par rapport à une autre dans un nuage de points et d'afficher l'histogramme d'une variable. L'outil de traitement de données est ainsi fonctionnel pour les derniers jalons du projet : la remise du rapport et la soutenance finale.

Partie III : Réalisation technique du projet :

a - Fonctionnement du script (R + Shiny) :

L'objectif principal du projet était le développement d'un outil permettant d'afficher des critères pertinents par le calcul de statistiques caractéristiques sur tout jeu de données réelles provenant d'internet. Pour réaliser cet objectif, nous avons tout de suite pensé à utiliser le langage R, reconnu pour le traitement statistique. De plus, le module Shiny de R s'est révélé être aussi très important car il permet de créer une interface graphique interactive.

Ainsi, l'application comporte trois onglets (tabPanel sur R), le premier permet d'importer un dataset et d'afficher le résumé des données. Le deuxième permet la création d'un nuage de points entre deux variables et le troisième représente l'histogramme d'une variable.



Après l'importation des modules nécessaires, la partie générale de l'interface utilisateur est définie grâce aux tabPanel.

```r
library(shiny)
library(datasets)
library(ggplot2)


ui= fluidPage(
  navbarPage("OTDS",tabPanel("Importation et Resume", value=1, uiOutput('page1')),
        tabPanel("Nuage de point", value=2, uiOutput('page2')),
        tabPanel("Histogramme", value=3, uiOutput('page3'))
  )
)
```

Figure 5 : Extrait 1 du script de l'outil de traitement

Il reste ensuite à définir la partie serveur qui est la partie de commande de l'interface. Pour une meilleure clarté du code, nous divisons les parties interface utilisateur, affichage et commande de chaque page les unes à la suite des autres.

Dans la première page, nous définissons sa partie interface qui interagit avec l'utilisateur (importation du fichier, case à cocher, bouton de choix) et la partie d'affichage (texte et table de données). La suite du code permet de contrôler les affichages en fonction du choix de l'utilisateur dans la partie interface. Ainsi, après l'importation d'un fichier txt ou csv, le résumé est calculé et affiché dans la partie affichage. De même pour le tableau de données. Aussi, les options d'importation (existence d'entête, séparateur de variable, séparateur décimal, indicateur de citation et le nombre de lignes à afficher dans les observations) de ce dataset sont gérées par l'utilisateur dans la partie interface.

```r
server = function(input, output, session) {
# page 1 ----
    output$page1 = renderUI({
      sidebarLayout(
        sidebarPanel(
          fileInput("file1", "Choisir un fichier CSV :", multiple = TRUE, accept = c("text/csv", "text/comma-separated-values,text/plain", ".csv")),
          tags$hr(),
          checkboxInput("header", "Cocher si votre fichier contient une entete", TRUE),
          radioButtons("sep", "Separateur de variable", choices = c("Virgule" = ",", "Point-virgule" = ";", "Tabulation" = "\t"), selected = ","),
          radioButtons("dec", "Separateur decimal", choices = c("Virgule" = ",", "Point" = "."), selected = "."),
          radioButtons("quote", "Indicateur de citation", choices = c("Aucun" = "", "Double Quote" = '"', "Single Quote" = "'"), selected = '"'),
          tags$hr(),
          radioButtons("disp", "Nombre de lignes a afficher", choices = c("Juste le debut" = "head", Toutes = "all"), selected = "head")
        ),

        mainPanel(
          h4("Resume"),
          verbatimTextOutput("summary"),
          tags$hr(),
          h4("Observations"),
          tableOutput("view")
        )
      )
    })

    output$summary = renderPrint({
      req(input$file1)
      df = read.csv(input$file1$datapath, header = input$header, sep = input$sep, dec = input$dec, quote = input$quote)
      return(summary(df))
    })
    output$view = renderTable({
      req(input$file1)
      df = read.csv(input$file1$datapath, header = input$header, sep = input$sep, dec = input$dec, quote = input$quote)
      if(input$disp == "head") {
        return(head(df))
      }
      else {
        return(df)
      }
    })
```

Figure 6 : Extrait 2 du script de l'outil de traitement

De la même façon que la première page, la page « Nuage de points » comporte d'abord le code de la partie interface, puis celui de la partie affichage et enfin celui de la partie commande. Dans la partie

interface, l'utilisateur peut choisir les variables à afficher sur l'axe des ordonnées et sur l'axe des abscisses. Il peut aussi ajouter une courbe de régression (qui décrit la relation entre deux variables) ainsi qu'un intervalle de confiance autour de cette courbe. La partie affichage décrit le graphe et ajoute un curseur permettant de lire graphiquement les valeurs de l'abscisse et de l'ordonnée d'un point. Enfin, la partie commande créer ce graphe à partir du dataset importé à la page précédente en fonction du choix (de x, de y et de l'ajout de courbe de régression) de l'utilisateur. Elle permet aussi de contrôler le curseur.

```
# page 2 ----
    output$page2 = renderUI({

      sidebarLayout(
        sidebarPanel(
          selectInput(inputId = "x_axis",
              label = "variable de l'axe des abscisses",
              choices = colnames(read.csv(input$file1$datapath, header = input$header, sep = input$sep, dec = input$dec, quote = input$quote))),
          selectInput(inputId = "y_axis",
              label = "variable de l'axe des ordonnees",
              choices = colnames(read.csv(input$file1$datapath, header = input$header, sep = input$sep, dec = input$dec, quote = input$quote))),
          checkboxInput("regress", "Cocher pour ajouter une courbe de regression", TRUE)
        ),

        mainPanel(
          plotOutput("myplot", click = "plot_click1"),
          verbatimTextOutput("info1")
        )
      )
    })

    output$myplot = renderPlot({
        df = read.csv(input$file1$datapath, header = input$header, sep = input$sep, dec = input$dec, quote = input$quote)
        ggplot(data=df, mapping = aes(x=unlist(df[input$x_axis]),y=unlist(df[input$y_axis]))) + geom_point() + labs(x=input$x_axis,y=input$y_axis) + if(input$regress) {geom_smooth(method="auto", se=TRUE, fullrange=FALSE, leve
    })
    output$info1 = renderText({
        paste0("x=", input$plot_click1$x, "\ny=", input$plot_click1$y)
      })

```

Figure 7 : Extrait 3 du script de l'outil de traitement

La dernière page implémentée dans notre script de traitement de données est une page d'histogramme, il a pour but de représenter la fréquence d'apparition des données d'une variable dans le dataset. La structure de page reste la même, l'utilisateur peut choisir la variable à afficher, l'histogramme et le curseur sont créés. Puis le graphe est affiché avec la moyenne de la variable.
Enfin, la fonction shinyApp permet de créer une page sur un navigateur web, contenant les informations d'interface utilisateur et de serveur de commande de l'affichage.

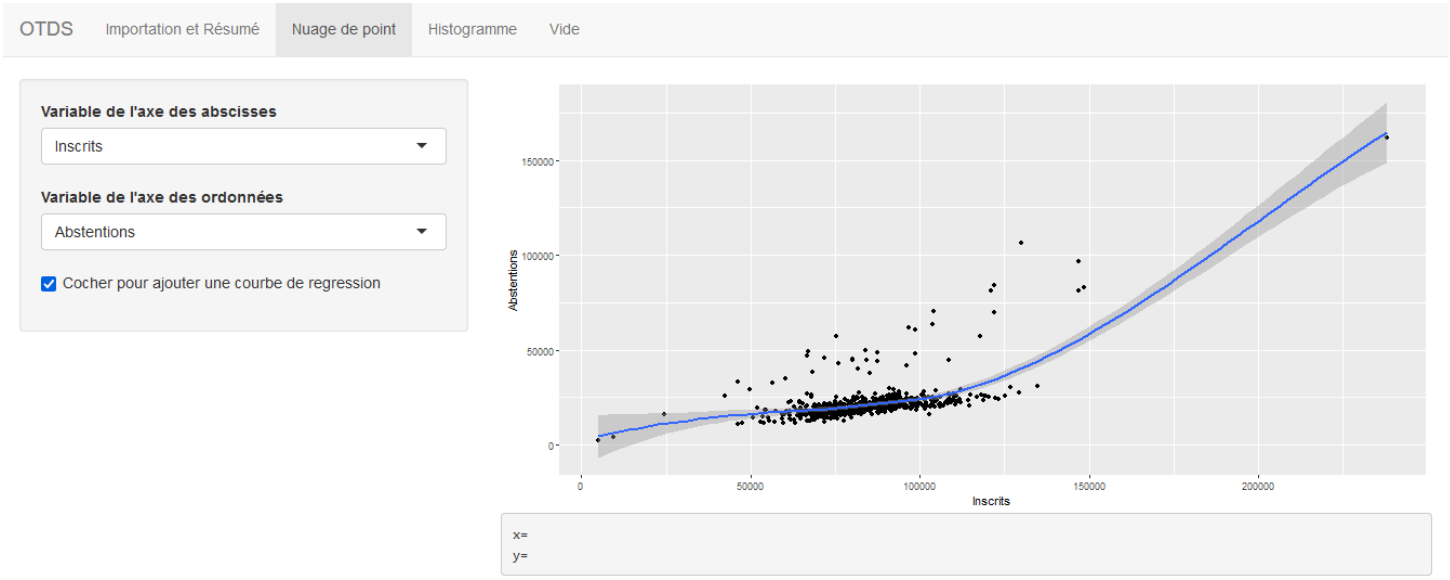
```
# page 3 ----
    output$page3 = renderUI({

      sidebarLayout(
        sidebarPanel(
          selectInput(inputId = "varia", label = "Variable a afficher", choices = colnames(read.csv(input$file1$datapath, header = input$header, sep = input$sep, dec = input$dec, quote = input$quote))),
        ),

        mainPanel(
          plotOutput("histogram", click = "plot_click2"),
          verbatimTextOutput("info2")
        )
      )
    })

    output$histogram = renderPlot({
        df = read.csv(input$file1$datapath, header = input$header, sep = input$sep, dec = input$dec, quote = input$quote)
        ggplot(df, aes(x=unlist(df[input$varia]))) + stat_count() + geom_vline(aes(xintercept=mean(unlist(df[input$varia]))),
            color="blue", linetype="dashed", size=1) + labs(x=input$varia, y="Frequence")
    })
    output$info2 = renderText({
        paste0("x=", input$plot_click2$x, "\ny=", input$plot_click2$y)
      })
}

shinyApp(ui, server, options = list(launch.browser = TRUE))
```

Figure 8 : Extrait 4 du script de l'outil de traitement

b - Développement de l’interface graphique :

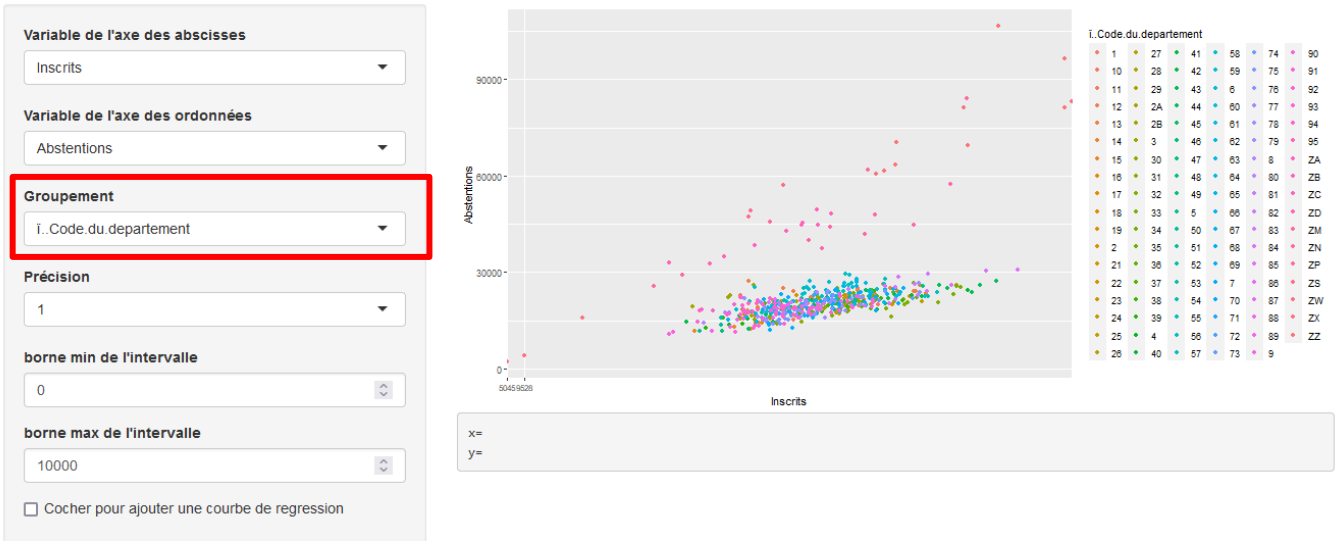
Comment exprimer les données le plus clairement possible ? Cela reste le but fondamental du projet : modularité et clarté. On a procédé par étapes :
Pour commencer, on a cherché à développer un graphe à nué de points et un histogramme, c’est les deux manières les plus classiques de mettre en exergue un ensemble de données. Puis au fur et à mesure, on a voulu développer des fonctionnalités pour répondre à des problèmes (par exemple problème d’affichage) ou bien juste améliorer le code pour offrir à l’utilisateur une meilleure expérience.

Exemple avec une base de données sur les résultats du 1er tour de l’élection présidentielle 2022 :

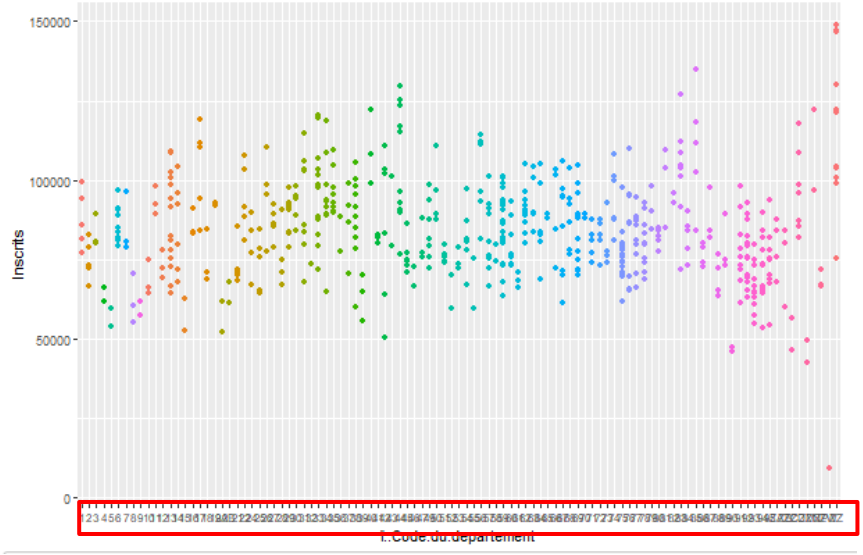


[On observe le nombre d’abstentions en fonction du nombre d’inscrits]

Pour continuer d’améliorer l’interface, l’idée a été de donner la possibilité à l’utilisateur de classer les différents points du graphe en fonction d’un critère. Cela permet de mettre en corrélation 3 attributs de la table des données.

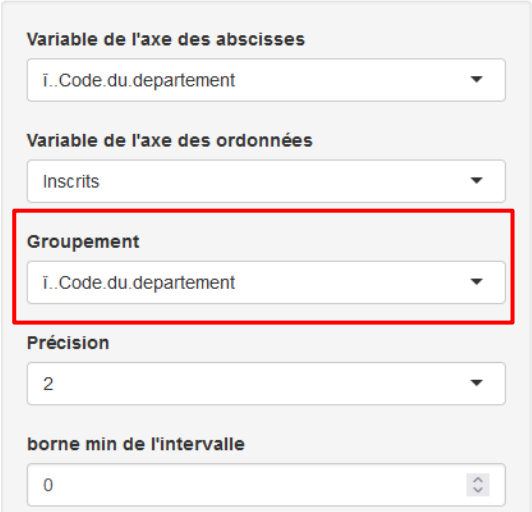


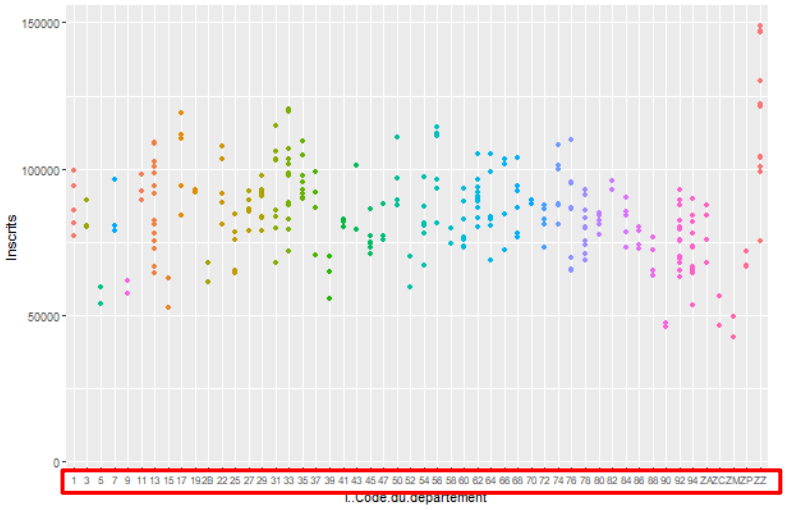
La sélection de l'attribut de regroupement se fait par l'entrée « Groupement ». Ainsi dans le cas ci-dessus, on affiche les abstentions en fonction des inscrits regroupé par départements
On s'est ensuite heurté à un problème. Il est arrivé que dans le cas où la classe de l'attribut est une chaine de caractère, l'affiche soit illisible :



On remarque qu'il est impossible de déchiffrer la graduation des abscisses. Après des recherches, on a eu l'idée de proposer à l'utilisateur de pouvoir choisir entre montrer toutes les valeurs sur l'axe des abscisses, 1 valeur sur 2, 1 valeur sur 3. Cela offre plus de visibilité mais on permet malheureusement des informations, vu qu'à chaque fois, on prend une valeur sur 2 ou une valeur sur 3.

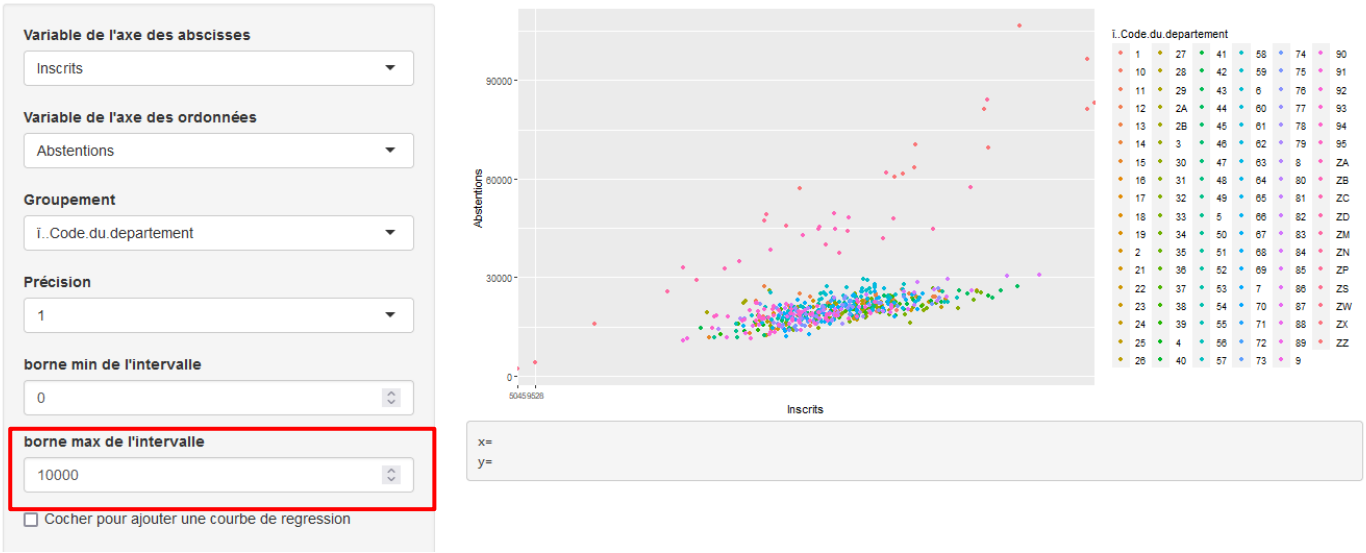
Voici ce qu'on obtient avec une précision de 2. On choisit une donc de prendre une valeur sur 2.





Enfin, on offre la possibilité à l'utilisateur, dans le cas où il est manipule une entrée constituée d'entier ou de réel, de d'étudier les données sur un intervalle qu'il peut choisir.

Ici l'utilisateur observe donc le nb d'abstentions en fonction du nombre d'inscrits sur l'intervalle [0, 100000]



c - Difficultés rencontrés:

1- Limites du nettoyage des données :

Le problème principal auquel s'est confronté l'équipe « Traitement de données » est le nettoyage de celles-ci. C'est un processus qui vise à identifier et corriger les données altérées, inexactes ou non pertinentes. Cela permet d'améliorer la cohérence, la fiabilité et la valeur des données. Cependant, comme nous ne connaissons pas le dataset importé par l'utilisateur en entrée, comment pouvons-nous nettoyer efficacement ce jeu de données ? Comment détecter des données inexactes ou non pertinentes ?

Les données inexactes sont généralement des valeurs manquantes (notées NA par R) ou des fautes de frappe. Ainsi, le mieux serait de corriger ces valeurs mais ne connaissant pas les données à priori, nous allons juste supprimer les données inexactes. Par chance, R a la fonctionnalité d'ignorer les valeurs manquantes mais pas de corriger les fautes de frappe. Il reste donc des erreurs potentielles dans le jeu de données.

Ensuite, les données non pertinentes dépendent fortement de ce que le statisticien veut mettre en avant. Pour une étude générale, ce sont les données les plus extrêmes qui semblent moins représentatives, alors que pour une étude aux limites, celles-ci sont tout aussi importantes. Ainsi, pour éliminer ces données non pertinentes, nous avons pensé (mais pas encore implémenté dans le script) un algorithme de nettoyage des données dont la première étape est l'identification des données essentielles en supprimant par exemple les variables d'indiçage des observations. La deuxième étape réalisera un tri des données permettant le calcul de statistiques. Les données étant collectées, nous pouvons ensuite résoudre les incohérences et les erreurs, en commençant par la suppression des observations qui sont en doubles. Il faut ensuite s'occuper des valeurs manquantes. Selon la proportion de ces valeurs dans une variable, il semble plus intéressant de soit supprimer intégralement la variable qui ne serait alors plus représentative, soit d'ignorer les valeurs manquantes. Enfin, nous laissons une option à l'utilisateur pour lisser son jeu de données en éliminant les valeurs qui sont hors de l'intervalle de confiance à n% (avec n entre 0,5 et 1 choisit par l'utilisateur) ou même d'éliminer des données trop éloignées de la moyenne calculée précédemment. Néanmoins, cet algorithme n'est pas standard à chaque dataset et il est donc préférable que l'utilisateur statisticien nettoie au préalable son jeu de données en fonction de ce qu'il souhaite faire.

2- Problème de représentation de données inconnues :

L'équipe « Affichage des données » s'est, quant à elle, confrontée au problème de la représentation d'un jeu de données inconnu. Ils ont réfléchi à quels graphes et quelles statistiques permettent au mieux d'analyser et de représenter des données que l'on ne connait pas au départ. En effet, les données provenant de l'utilisateur peuvent être de tout type (chaîne de caractère, bouléen, nombre (entier, décimaux, date, pourcentage), …) et même pour les variables de nombre, celles-ci peuvent être affichées dans des unités différentes (m/pouce, kg/livre, °C/°F, …). Un premier résumé des variables permet de déterminer ce type alors que l'unité, définie généralement de manière affine par rapport à une autre, n'est pas un problème majeur dans la représentation.

Ainsi, la représentation optimale d'un jeu de données correspond à leur type. Par exemple, on préférera utiliser des graphiques en courbe pour des variables possédant un changement dans la durée, des nuages de point pour relever une corrélation entre deux variables ou des histogrammes pour montrer la répartition des événements par leur fréquence. Il serait aussi intéressant d'afficher des données correspondantes à des facteurs géographiques sur des cartes, même si ce cas est parfaitement réalisable avec R + Shiny, il est beaucoup trop spécifique à un type de données et n'est alors pas caractéristique dans le cas général.

Pour réaliser l'objectif de représenter graphiquement tout dataset inconnu, nous avons choisi deux graphes et un ensemble de statistique descriptive. Les données statistiques caractéristiques calculées sont les suivantes. Pour des chaînes de caractère, le nombre d'observation. Pour les valeurs bouléennes, le nombre de FAUX, le nombre de VRAI et le nombre de NA (valeur manquante). Et pour les nombres, le minimum et le maximum, les 1er et 3ème quartile, la médiane, la moyenne et le nombre de NA. Ainsi, le script ne comprend pas lui-même si un nombre représente un pourcentage ou non. L'utilisateur doit en être conscient pour utiliser l'outil avec cohérence. Le premier graphe choisi est le nuage de points car il permet de représenter au mieux les corrélations entre deux variables, que se soit l'identification de relations ou de répartitions plus ou moins homogènes. Nous y ajoutons une courbe de régression calculée par R. Le second graphe choisi pour la représentation de critères pertinents dans le jeu de données est l'histogramme car il permet de montrer la distribution des données de façon pratique. Cette distribution permet aussi d'approximer les lois des variables aléatoires à des lois connues (uniforme, normale, exponentielle, …).

Une autre statistique pertinente aurait pu être les observations extrêmes d'une variable qui permettent de communiquer des informations sur les limites.

Partie IV : Mise en application du script sur des exemples :

a- Exemple d'utilisation du tableau de bord:

On a ici pris pour exemple un fichier excel fournissant des données sur les matchs de football de premières divisions européennes entre 2012 et 2017.

Ce fichier enregistre les événements réalisés lors des matchs dans la colonne event_type qui couvrent tous les types d'actions possibles dans un match ; du carton jaune, aux penaltys concédés en passant par les coups francs gagnants ou encore les tentatives de tir.

Pour les tentatives de tirs, on trouve dans la colonne shot_place l'information portant sur la localisation du tir.

Pour les tirs encore, on retrouve aussi le résultat du tir en question ; s'il est cadré ou non, s'il a été arrêté par le gardien ou si le ballon a touché le poteau ou la barre transversale.

On retrouve également des informations sur la localisation -sur le terrain- de l'action, la partie du corps avec laquelle le joueur a tiré, le contexte de l'action (coups de pieds arrêtés ou non), etc.

Toutes ces informations sont compilées dans les colonnes du tableau par des chiffres.

Un dictionnaire, sous format texte est fourni avec les tables pour lire le tableau. (Voir en annexe ce qu'on y retrouve)

L'onglet Importation et Résumé donne un ensemble de statistiques élémentaires sur chaque colonne du dataset (premier quartile, moyenne, médiane, troisième quartile, minimum, maximum) :

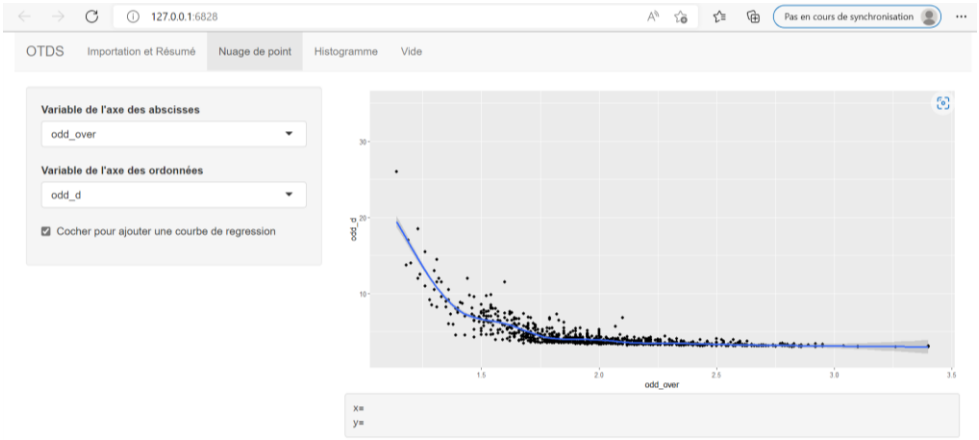
```
     odd_d            odd_a
Min.   : 1.910   Min.   : 1.110
1st Qu.: 3.430   1st Qu.: 2.740
Median : 3.680   Median : 3.860
Mean   : 4.278   Mean   : 5.538
3rd Qu.: 4.300   3rd Qu.: 6.000
Max.   :35.000   Max.   :81.000

   odd_over       odd_under
Min.   :1.140   Min.   :1.420
1st Qu.:1.790   1st Qu.:1.780
Median :2.030   Median :1.970
Mean   :2.047   Mean   :2.106
3rd Qu.:2.280   3rd Qu.:2.270
Max.   :3.400   Max.   :7.500
NA's   :9135    NA's   :9135
```

Ce tableau de bord dispose également d'un onglet histogramme. On donne ici par exemple l'histogramme de la colonne shot_place qui indique la position du tir, repartie dans les catégories suivantes (le chiffre indique le code correspondant à la catégorie en question) - Voir annexe.

Dans une table annexe qui donne des informations sur les cotes de pari liées aux matchs analysés, on peut tenter de chercher des corrélations. Pour rappel, une cote de pari traduit la probabilité que l'évènement en question se produise. Elle est calculée en tenant compte de l'historique des rencontres entre les deux équipes et de toutes autres informations jugées intéressantes par les bookmakers. Plus la cote est élevée, plus la probabilité que l'évènement en question se produise est jugée faible par les bookmakers. A titre d'échelle, une cote supérieure à 2 est relativement élevée.

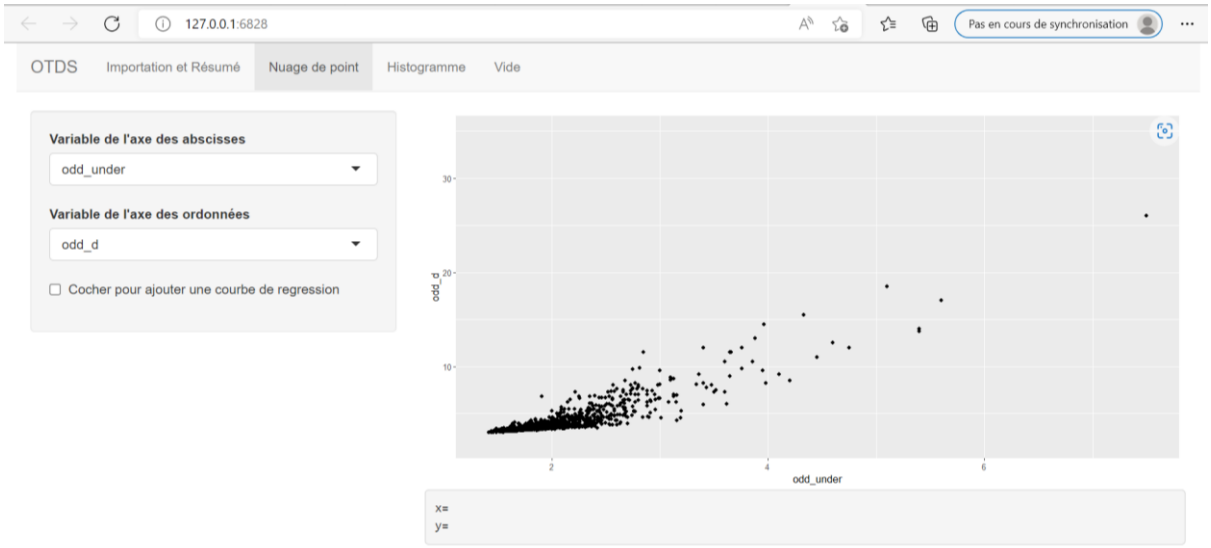
Pour en revenir à la table, on peut tracer par exemple l'évolution de la cote liée à l'événement : « le nombre de buts marqués est supérieur à 2,5 » en fonction de la cote liée au fait que le match se solde en un match nul.

On obtient le nuage de points suivant (avec une courbe de régression) :



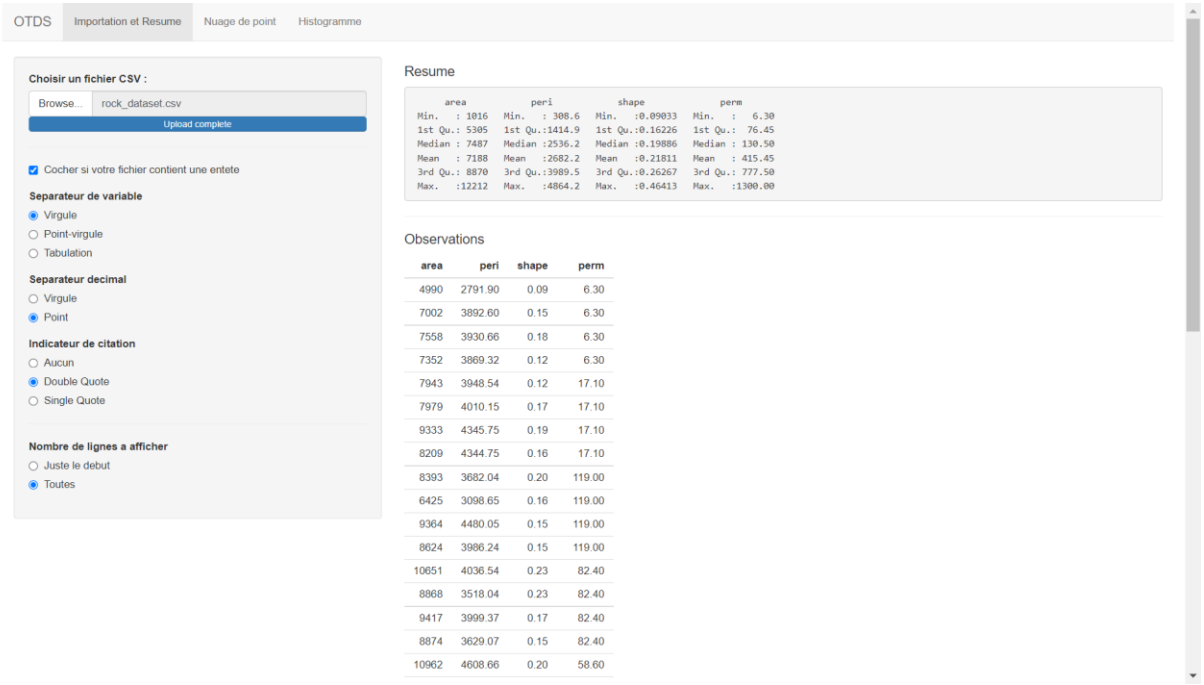
Ainsi, les bookmakers considèrent que les matches nuls les plus probables sont des matchs où le nombre de buts inscrits est susceptible d'être faible.

On peut vérifier la tendance inverse, en traçant l'évolution de la cote liée à l'événement : « le nombre de buts inscrits est inférieur à 2.5 » en fonction de la cote liée au fait que le match se solde en un match nul. On obtient le graphique suivant, qui nous mène à la même conclusion que précédemment :

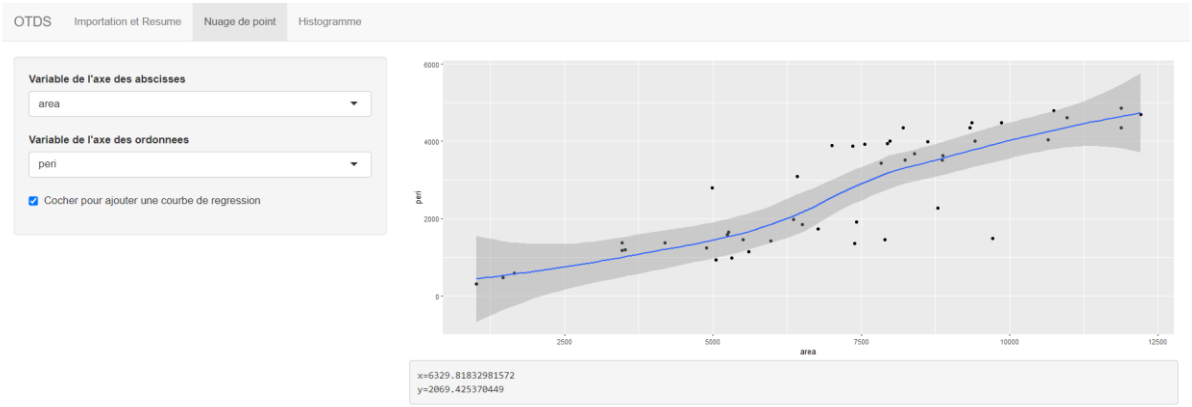


b- Dataset « Rock »:

Le premier jeu de données est le dataset « Rock » implémenté de base dans R qui contient la mesure de 48 échantillons de roche d'un réservoir pétrolier. Les variables sont area : la surface de l'espace des pores (en pixels sur 256x256), peri : le périmètre (en pixels), shape : défini comme le périmètre divisé par la racine carré de la surface et perm : la perméabilité des roches.
La première page permet d'importer le dataset, donne le résumé des variables et affiche les observations.

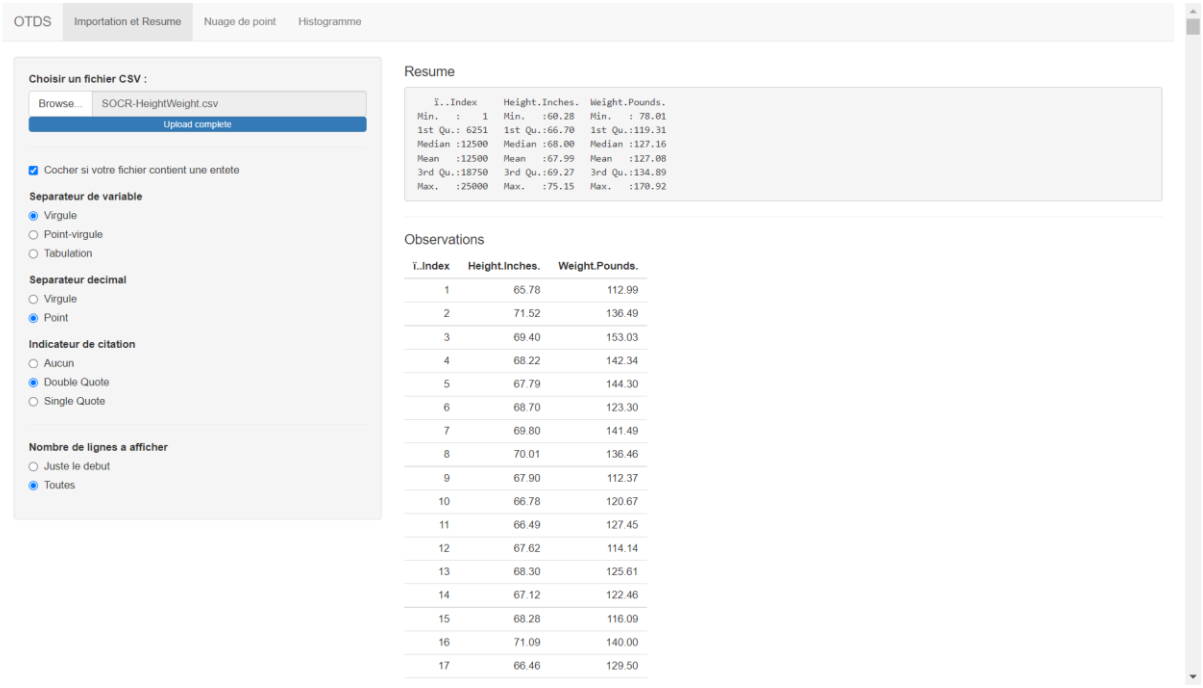


La page nuage de points permet d'observer une variable en fonction d'une autre, par exemple ici nous avons affiché le périmètre des roches en fonction de la surface des pores. On aperçoit une relation presque linéaire entre ces variables. Ainsi, ce sont les roches les plus grandes qui possèdent les pores les plus grands ce qui semble cohérent physiquement. On pourrait aussi calculer la relation par approximation linéaire entre ces deux variables.



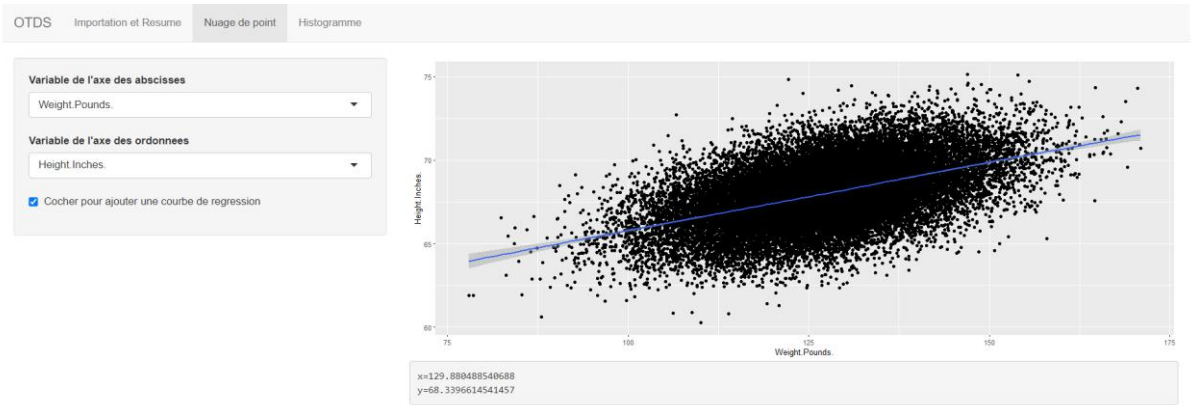
c- Dataset « Tailles et Poids »

Un second exemple d'application de notre script provient d'un dataset nommé « Heights and Weights Dataset » téléchargé sur kaggle.com. Il contient la taille (en pouces) et le poids (en livres) de 25 000 personnes de 18 ans.



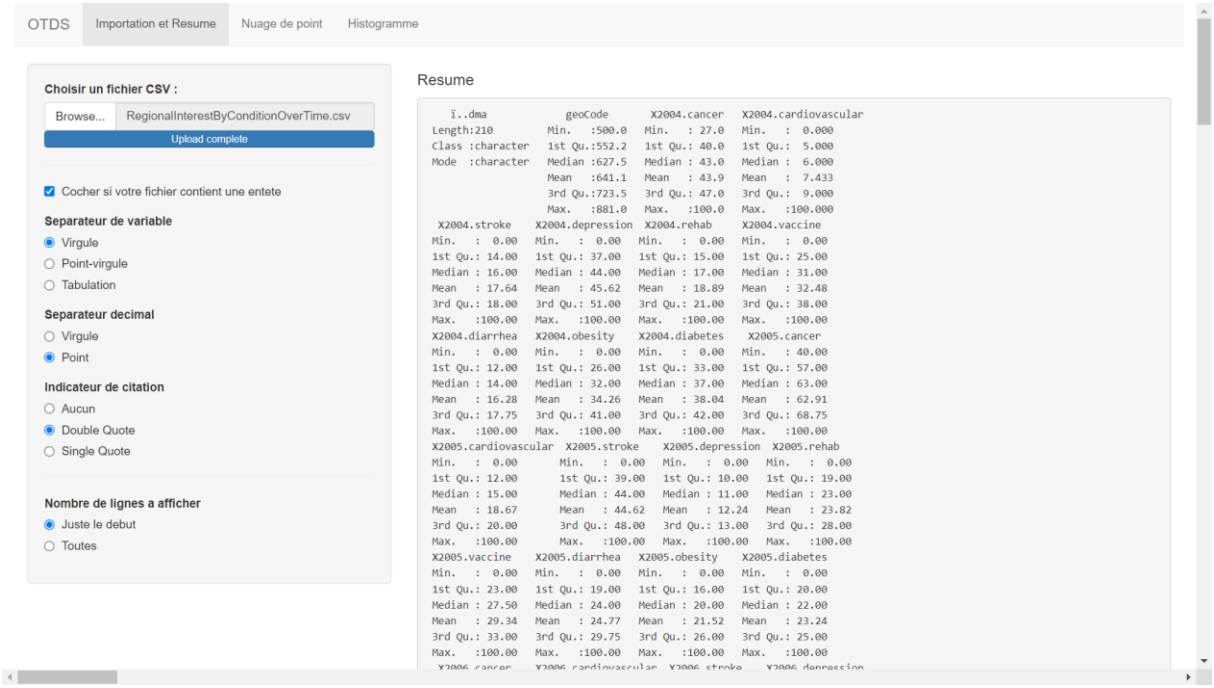
ï..Index	Height.Inches.	Weight.Pounds.
1	65.78	112.99
2	71.52	136.49
3	69.40	153.03
4	68.22	142.34
5	67.79	144.30
6	68.70	123.30
7	69.80	141.49
8	70.01	136.46
9	67.90	112.37
10	66.78	120.67
11	66.49	127.45
12	67.62	114.14
13	68.30	125.61
14	67.12	122.46
15	68.28	116.09
16	71.09	140.00
17	66.46	129.50

En représentant le taille en fonction du poids, on retrouve une relation parfaitement affine entre ces variables ce qui est assez connu scientifiquement. Le fait qu'il est beaucoup de données permet de créer un intervalle de confiance à 95% de la droite de régression très étroit et donc de valider le modèle affine. On trouve ainsi la relation Hauteur=0,1*Poids+55 (avec Hauteur en pouces et Poids en livre) pour des humains de 18 ans. Cette relation permet alors d'approximer la taille moyenne ou le poids moyen de quelqu'un de 18 ans en connaissant l'une des deux variables.

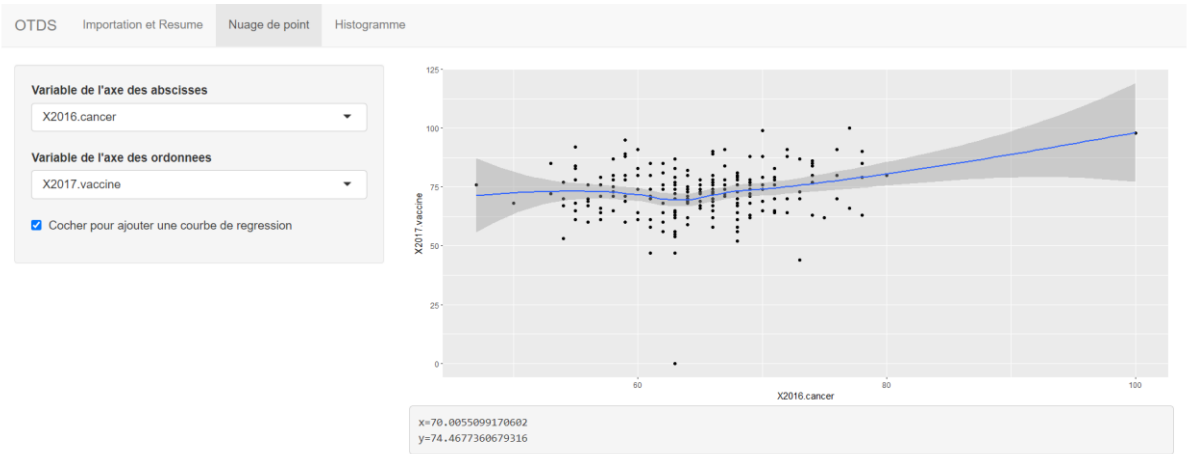


d- Dataset « Intérêts dans la recherche sur différentes maladies »

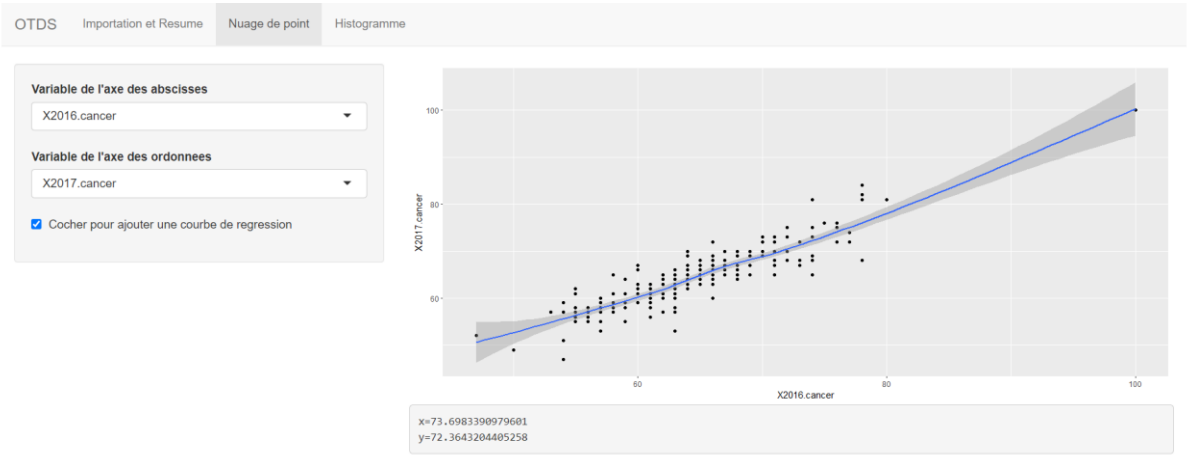
Nous avons aussi étudié un dataset présentant l'intérêt des Américains à propos de la recherche sur différentes maladies de 2004 à 2017. Ainsi, nous pouvons décrire l'évolution de l'intérêt des Américains sur ces maladies.



Dans cet exemple, nous voyons que l'intérêt pour les vaccins en 2017 ne dépend pas de l'intérêt pour les cancers en 2016. Ces variables sont indépendantes.

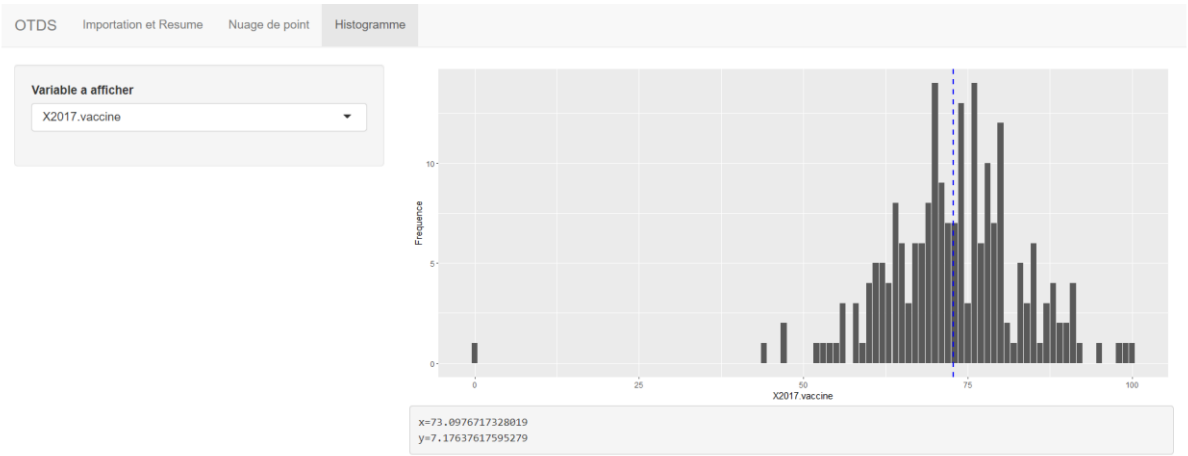


Dans l'exemple suivant, nous voyons que l'intérêt pour les cancers est le même en 2016 et en 2017 (droite y=x).



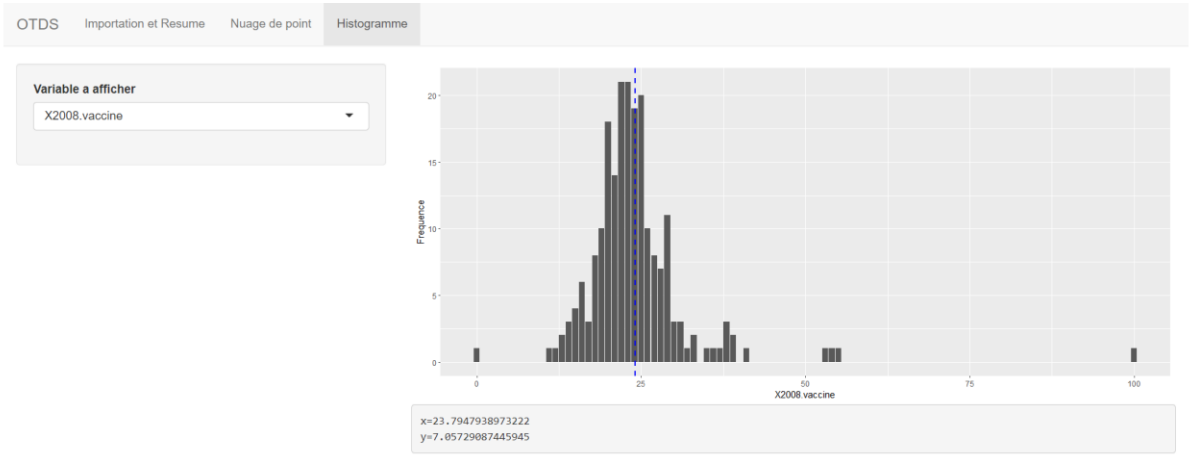
En utilisant ces données, nous pouvons montrer la relation entre l'obésité et le diabète qui ressort le plus.

En utilisant l'histogramme, on voit que la distribution de la variable « Intérêt pour la recherche sur les vaccins en 2017 » suit une loi normale de moyenne égale à 73% avec un assez grand écart-type. Donc en moyenne, 73% des Américains exprimaient un intérêt pour la recherche sur les vaccins en 2017.



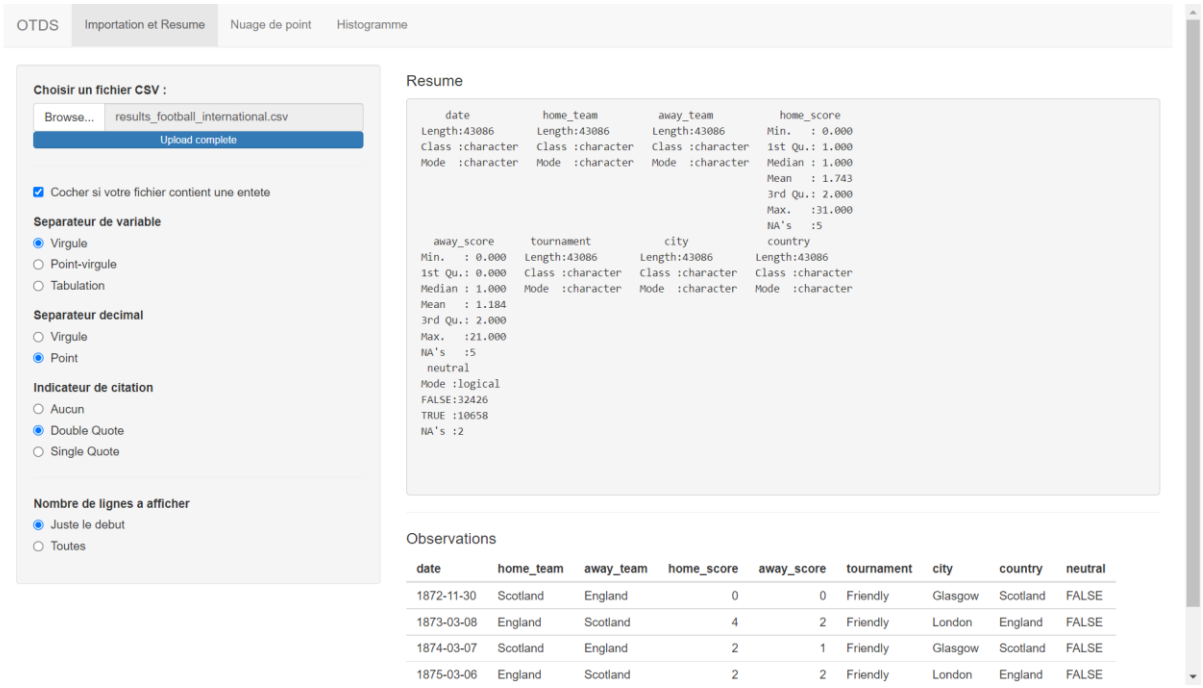
Alors que la variable « Intérêt pour la recherche sur les vaccins en 2008 » suit une loi normale de moyenne 24% avec un écart-type faible. Donc en moyenne, 24% des Américains exprimaient un intérêt pour la recherche sur les vaccins en 2008.

Cette différence avec 2017 montre une évolution croissante de l'intérêt pour les vaccins aux Etats-Unis dans les années 2010.



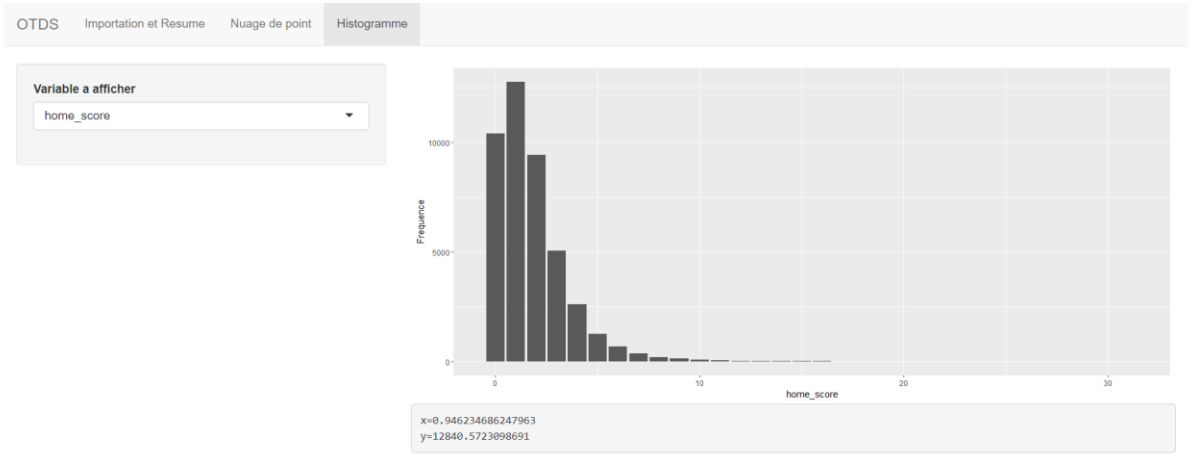
e- Dataset « Résultats internationaux de Football »

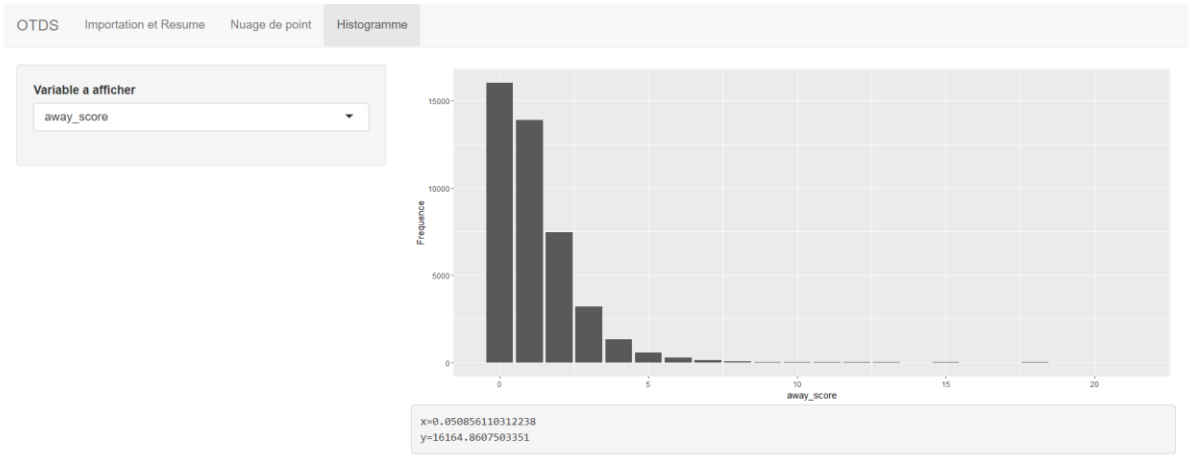
Dans un tout autre sujet, nous nous sommes posé la question de l'avantage de jouer à domicile par rapport à jouer à l'extérieur au Football. Ainsi, nous avons récupéré un dataset donnant les résultats internationaux de football dans les 150 dernières années.



date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland	FALSE
1873-03-08	England	Scotland	4	2	Friendly	London	England	FALSE
1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland	FALSE
1875-03-06	England	Scotland	2	2	Friendly	London	England	FALSE

Puis nous avons tracé les histogrammes des nombres de but mis à domicile ou en extérieur.

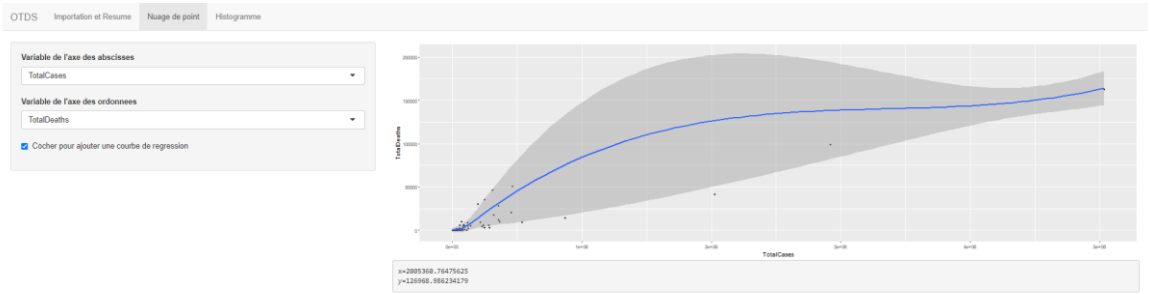




Sur ces graphiques, nous voyons qu'en moyenne, l'équipe à domicile marque 0,559 but de plus que l'équipe extérieure. Il est plus probable que l'équipe à domicile marque 1 but que 0 but alors qu'il est plus probable que l'équipe extérieure marque 0 but que 1 but. Ainsi, nous avons prouvé par analyse statistique de données qu'il y a un avantage incontestable de jouer à domicile même si cette analyse ne donne pas les raisons de cet avantage. Le statisticien qui utilise le programme doit alors prendre du recul sur ces données et appliquer une méthode d'ingénieur pour comprendre les résultats du traitement effectué. Le script que nous avons créé n'est donc pas un analyseur magique de données mais un outil permettant à l'utilisateur de traiter ses données simplement avec les statistiques.

f- Dataset « covid-19 »

Un dernier exemple parlant provient de données sur le covid-19. Nous représentons le nombre total de morts par rapport au nombre total de cas. Nous voyons sur ce graphe que lorsqu'il y a très peu de cas, il y a, logiquement, très peu de mort. Puis, nous voyons que le nombre de mort augmente linéairement par rapport au nombre de cas mais que, plus il y a de cas, plus la pente de cette droite est faible. Ainsi, à partir d'un certain nombre de cas, cette variable n'a plus d'influence sur le nombre de mort.



Partie V- Clôture du projet :

a- Améliorations possibles de l'outil développé - Ce qu'il reste à faire:

Même si nous avons validé les objectifs généraux proposés par le projet, le travail que nous avons fourni peut servir de base lors de la continuation de ce projet. En effet, les questions de représentation et de traitement des données inconnues peuvent être poussées plus loin. Le script en lui-même peut aussi être amélioré par l'ajout de fonctionnalité comme des autres graphiques. Un algorithme de nettoyage des données en fonction du choix de l'utilisateur peut être mis en place. Nous avons aussi réfléchi à l'ajout de fonctionnalités pour traiter les données par groupe (group by) pour qu'on puisse traiter certains dataset d'une meilleure façon. Aussi, il est possible d'améliorer les graphismes en ajoutant un style graphique moins brut à l'application.

b- Conclusions personnelles:

Rim : Ce projet a été très formateur pour moi. Tout d'abord, il m'a permis de monter en compétence technique. L'apprentissage du langage R sera un atout indéniable pour le reste de ma scolarité et l'entrée sur le marché du travail. D'un autre côté, ce projet m'a réellement formé au travail en équipe et j'y ai compris ce qu'impliquait de mener à bien un projet. La communication mais aussi la rigueur et l'échange de critiques constructives sont nécessaires. L'expérience de ce projet nous a permis de concrétiser les apprentissages du mooc de Gestion de projet et de manière plus générale, toute nos connaissances théoriques. En cela, ce projet m'a été très profitable et j'en tire des leçons qui me seront très utile pour la suite.

Elie : Malgré des incompréhensions sur l'objectif et l'intérêt du projet dans un premier temps, ce dernier s'est révélé être très formateur au niveau technique avec une formation poussée au langage R, à un module très utilisé : Shiny et beaucoup de questionnements sur le traitement et la représentation de données par les statistiques via le développement d'un outil de traitement de données. J'ai apprécié le fait d'avoir beaucoup de libertés dans l'interprétation du sujet en ayant en même temps un encadrement disponible. Enfin, ce projet m'a permis de découvrir ce que la gestion de projet impliquait en pratique avec la répartition du travail dans l'équipe, les idées et compétences de chacun et le rendu des livrables.

Youssef : Ce projet a été riche d'enseignements pour moi. En premier lieu, j'ai eu l'occasion de mettre en application les connaissances acquises grâce au MOOC gestion de projet et également par l'intermédiaire d'un cours d'apprentissage du langage R sur Openclassroom. A vrai dire, je crois que c'est la première fois dans l'histoire de mon cursus académique que je suis amené à mettre en œuvre concrètement des connaissances récemment intégrées. J'ai progressé tant en termes de « soft skills » que sur un plan plus

technique en informatique par l'apprentissage d'un nouveau langage de programmation. De plus, ce projet m'a offert l'opportunité de comprendre, de manière empirique j'entends, l'importance de la communication dans un groupe de travail ainsi que la nécessité de s'adapter, de faire évoluer ses objectifs et ses méthodes de travail au fil des problèmes rencontrés. Ce projet a donc été une expérience qui m'a été profitable à tous les niveaux.

Hugo : Ce projet fut pour moi très enrichissant. En effet, ce dernier alliait compétences techniques approfondies comme l'apprentissage du langage R, comment manier des data set ou le traitement des données statistiques. Et compétences plus propres à la gestion de projet comme la planification des réunions et la répartition des tâches de chacun. J'ai aimé travailler en équipe car cela oblige une rigueur commune très différente de ce à quoi j'étais habitué dans le passé. Finalement, ce projet fut pour moi la possibilité de m'initier au monde des statistiques qui m'intriguait beaucoup avant le début de ce projet. Le projet m'a donc, sans surprise, beaucoup intéressé.

Léon : Après une année à chercher, réfléchir, se tromper et finalement réussir, ce projet est et a été pour moi une expérience très enrichissante que ce soit sur le plan technique ou sur le plan organisationnel. En effet, j'ai appris un nouveau langage de programmation et pu développer un programme avec elle, chercher et trouver les moyens de présenter les informations de manières claires, offrant à l'utilisateur le plus de possibilités de traiter les données. D'autant plus que chacun dans l'équipe effectuait son travail, cela change des expériences passées et m'a permis de découvrir le travail à plusieurs !

Conclusion

Ce projet nous a apporté plusieurs compétences :

Dans un premier temps, ce sont les compétences techniques que nous avons dû développer, on a donc décider de suivre un cours en ligne afin d'apprendre à coder en langage R et de comprendre comment fonctionne Rstudio. On a aussi dû comprendre la structure des bases de données notamment savoir comment les « nettoyer » afin qu'elles deviennent exploitables.
De plus, il a fallu créer des programmes afin d'avoir les histogrammes et les nuages de points en couleur sur l'interface. Enfin il a fallu apprendre à rédiger des rapports afin de tenir informer de l'avancée du projet Madame Benmouffek.

Il a fallu dans un second temps acquérir des compétences non techniques mais tout aussi importantes pour mener à bien le projet. En effet, nous avons eu des problèmes de communications en début de projet qu'on dû régler et c'est à ce moment-là que nous avons décidé d'attribuer à chacun un rôle bien précis. On a également dû faire preuve d'adaptation et de pouvoir gérer une situation assez critique puisqu'on a changé au début de second semestre l'orientation de projet et notamment savoir quels dataset nous intéressait et qu'est-ce qu'on voulait en faire ressortir. Finalement, on a également appris à s'organiser en équipe afin que les réunions entre nous ne soit pas vide de sens et soit efficace. Ainsi, nous devions avant chaque réunion savoir ce que nous allions dire et qu'elles étaient nos perspectives.

Bibliographie

- Cours « Inférence Statistique » par Rémi PEYRE
- MOOC Gestion de projet
- Cours Openclassroom « Initiez-vous au langage R pour analyser vos données »

Annexe 1 : Partie IV - a) Exemple d’utilisation du tableau de bord

Dictionnaire fourni avec les tables pour lire le tableau :

Location :

1-Attacking half

2-Defensive half

3-Centre of the box

4-Left wing

5-Right wing

6-Difficult angle and long range

7-Difficult angle on the left

8-Difficult angle on the right

9-Left side of the box

10-Left side of the six yard box

11-Right side of the box

12-Right side of the six yard box

13-Very close range

14-Penalty spot

15-Outside the box

16-Long range

17-More than 35 yards

18-More than 40 yards

19-Not recorded

Annexe 2 : Partie IV - a) Exemple d’utilisation du tableau de bord

Histogramme de la colonne shot_place indiquant la position du tir, repartie dans les catégories suivantes (le chiffre indique le code correspondant à la catégorie en question) :

shot_place

1-Bit too high

2-Blocked

3-Bottom left corner

4-Bottom right corner

5-Centre of the goal

6-High and wide

7-Hits the bar

8-Misses to the left

9-Misses to the right

10-Too high

11-Top centre of the goal

12-Top left corner

13-Top right corner