fas1001\_Plan\_Rousseau

#### Librairies & Packages:

# Librairies & Packages:   
#install.packages("tidyverse")   
suppressMessages(library(tidyverse))   
#install.packages("ggplot2")   
library(ggplot2)  
#install.packages("stats")  
library(stats)  
#install.packages("viridis")  
library(viridis)

Warning: le package 'viridis' a été compilé avec la version R 4.3.3

Le chargement a nécessité le package : viridisLite

## 1 - Question de recherche et raisons motivant cette dernière.

Quel est l’impact de l’activité physique le niveau d’éducation dans la population des États-Unis?

Voici plusieurs raisons qui motivent se choix de sujet:

1. **Intérêt personnel** : En tant qu’étudiant athlète passionné par l’activité physique, cette question permet d’explorer un sujet qui me concerne directement, ce qui peut rendre la recherche plus motivante et significative. De plus, étant un étudiant en science politique, il s’agit d’une oppotunité de joindre des loisirs et passions a mon sujet d’étude, quelque chose que je n’ai pas vraiment eu la chance de faire a cette ampleure auparavant dans mon cheminement académique.
2. **Impact potentiel** : Comprendre la relation entre le niveau d’éducation, l’activité physique et la santé chez les étudiants athlètes peut avoir un impact important sur la promotion de modes de vie sains et l’amélioration des performances académiques et sportives.
3. **Portée large** : En examinant l’impact de l’activité physique sur la santé et le niveau d’éducation dans la population des États-Unis, cette étude pourrait avoir des implications pour la santé publique à l’échelle nationale.
4. **Compréhension globale** : Cette question permet d’explorer comment l’activité physique peut influencer à la fois la santé et le niveau d’éducation, offrant ainsi une vision globale de son importance pour le bien-être global des individus.
5. **Politiques publiques** : Les résultats de cette recherche pourraient être utilisés pour informer les politiques publiques visant à promouvoir l’activité physique et à améliorer la santé et le niveau d’éducation de la population.
6. **Comparaison entre groupes** : Vous pourriez également envisager de comparer les effets de l’activité physique sur la santé et le niveau d’éducation entre différents groupes démographiques, ce qui pourrait mettre en lumière des disparités et des facteurs de risque potentiels

## 2 - Données et Méthodes d’analyse:

### 2.1 Données:

Une partie des données qui sera utilisée provient du Behavioral Risk Factor Surveillance System (BRFSS) un système de surveillance des facteurs de risque comportementaux. Il s’agit d’une enquête téléphonique menée chaque année aux États-Unis par les Centers for Disease Control and Prevention (CDC) pour collecter des informations sur les comportements liés à la santé de la population, tels que le tabagisme, la consommation d’alcool, l’activité physique, l’alimentation, l’utilisation de la ceinture de sécurité, etc. Les données recueillies par le BRFSS sont utilisées pour surveiller les tendances des comportements à risque, planifier et évaluer les programmes de santé publique, et orienter les politiques de santé.

les variables qui nous intréssent vont comme suit: L’année de début de l’étude, représentée par la variable YearStart, permet d’identifier quand les données ont été collectées. La variable Data\_Value indique les valeurs des données, exprimées en pourcentage, qui représentent les niveaux d’activité physique, de santé ou de niveau d’éducation. La variable LocationDesc fournit une description de l’emplacement, permettant d’identifier les lieux où les données ont été collectées. La variable Question décrit la question posée dans l’enquête, définissant ainsi le sujet de l’étude. Les variables StratificationCategory1 et Stratification1 sont utilisées pour regrouper les données selon différentes caractéristiques, telles que le groupe d’âge ou le sexe. La variable Sample\_Size indique la taille de l’échantillon, permettant d’évaluer la fiabilité des estimations basées sur les données. En outre, les variables Income, Race/Ethnicity et Gender sont utilisées pour évaluer l’impact du revenu, de l’origine ethnique et du genre sur les niveaux d’activité physique, de santé et de niveau d’éducation, respectivement. Ces variables sont essentielles pour comprendre les résultats de l’étude et évaluer les différences entre les groupes.

Une autre partie des données provient du département d’éducation du gouvernement des États-Unis et donné en open source. Les données utilisées comprennent plusieurs variables importantes. La variable “state” représente l’état dans lequel les données ont été collectées, fournissant ainsi des informations sur la région étudiée. La variable “region” fournit des détails supplémentaires sur la région géographique à laquelle appartient l’état. La variable “state\_median\_income” indique le revenu médian de l’état, offrant ainsi un aperçu de la situation économique générale de la région. La variable “state\_income\_percentile” indique le percentile du revenu de l’état par rapport aux autres états, ce qui permet de comparer le niveau de revenu de l’état avec celui des autres régions. La variable “state\_proportion\_with\_bachelors\_or\_higher” représente la proportion de personnes dans l’état ayant un diplôme de bachelier ou un niveau d’éducation plus élevé, ce qui peut être un indicateur de la qualité de l’éducation dans la région. Les variables “total\_colleges\_within\_25km\_of\_state\_counties”, “total\_colleges\_within\_50km\_of\_state\_counties”, “total\_colleges\_within\_75km\_of\_state\_counties” et “total\_colleges\_within\_100km\_of\_state\_counties” indiquent le nombre total de collèges situés à différentes distances des comtés de l’état, ce qui peut être utile pour évaluer l’accessibilité à l’éducation supérieure dans la région. Ces variables sont cruciales pour comprendre la relation entre le niveau de revenu, le niveau d’éducation et l’accessibilité à l’éducation supérieure dans chaque état.

Une troisième base de données provenant de la même source que la deuxième nous offre des données sur la démographie de la population américaine, selon les États ainsi que sur trois années (2017, 2018, 2019), ce qui nous indique comment filtrer ou nettoyer nos donner selon les dates, questions que nos analyses futures soient pertinentes et précises.

# importation des données sur l'activité physique:  
Physical\_act <- read.csv("C:/Users/rouss/OneDrive - Universite de Montreal/fas\_1001\_rousseau/Nutrition\_\_Physical\_Activity\_\_and\_Obesity\_-\_Behavioral\_Risk\_Factor\_Surveillance\_System.csv")   
  
View(Physical\_act)   
  
# importation des données sur l'éducation:  
Education <- readxl::read\_xlsx("C:/Users/rouss/OneDrive - Universite de Montreal/fas\_1001\_rousseau/State - Income, Education Level by State.xlsx")   
  
View(Education)

### 2.2 Nettoyage des données

# Nettoyage de données de la banque de pour l'activité physique:   
  
questions\_activite\_physique <- c(  
 "Percent of adults who engage in no leisure-time physical activity",  
 "Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)",  
 "Percent of adults who achieve at least 300 minutes a week of moderate-intensity aerobic physical activity or 150 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)",  
 "Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic physical activity and engage in muscle-strengthening activities on 2 or more days a week"  
)  
  
noms\_questions <- data.frame(  
 Question = questions\_activite\_physique,  
 niveau\_activite = c(  
 "Aucune activité physique",  
 "150+ min/sem",  
 "300+ min/sem",  
 "renforcement musculaire"  
 ),  
 ordre\_activite = c(1, 2, 3, 4)  
)  
  
physical\_act\_clean <- Physical\_act %>%   
 filter(YearStart %in% c(2017)) %>%   
 select(LocationDesc, Question, Data\_Value, Stratification1, Sample\_Size) %>%  
 filter(Question %in% questions\_activite\_physique) %>%  
 left\_join(noms\_questions, by = "Question") %>%  
 select(-Question) %>%   
 rename(state = LocationDesc) %>%   
 rename(proportion\_activite = Data\_Value) %>%   
 drop\_na()  
  
  
  
  
# Nettoyage de la banque de données d'éducation:   
education\_clean <- Education %>%   
 select(-region, -state\_income\_percentile) %>%   
 rename(diplomes\_universitaires = state\_proportion\_with\_bachelors\_or\_higher,  
 uni\_au\_25km = total\_colleges\_within\_25km\_of\_state\_counties,  
 uni\_au\_50km = total\_colleges\_within\_50km\_of\_state\_counties,  
 uni\_au\_75km = total\_colleges\_within\_75km\_of\_state\_counties,  
 uni\_au\_100km = total\_colleges\_within\_100km\_of\_state\_counties)  
  
  
# Groupement des bases de données:  
# Joindre la base de données sur l'éducation et les universités :  
pa\_education <- merge(physical\_act\_clean, education\_clean, by = "state", all.x = TRUE)%>%  
 mutate(  
 ordre\_activite\_1 = as.integer(niveau\_activite == "Aucune activité physique"),  
 ordre\_activite\_2 = as.integer(niveau\_activite == "150+ min/sem"),  
 ordre\_activite\_3 = as.integer(niveau\_activite == "300+ min/sem"),  
 ordre\_activite\_4 = as.integer(niveau\_activite == "renforcement musculaire"),  
 replace(diplomes\_universitaires, is.na(diplomes\_universitaires), 0)  
 )

### 2.3 Visualisation des différentes données:

#### 2.3.1 Activité physique:

# Calculer les moyennes par État  
mean\_data <- pa\_education %>%  
 group\_by(state, ordre\_activite) %>%  
 summarize(mean\_proportion\_diplomes = mean(diplomes\_universitaires),  
 proportion\_activite = first(proportion\_activite))

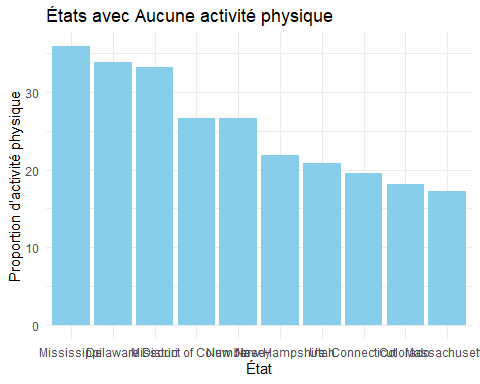
`summarise()` has grouped output by 'state'. You can override using the  
`.groups` argument.

# Identifier les États avec les taux les plus élevés dans chaque catégorie  
top\_states <- mean\_data %>%  
 group\_by(ordre\_activite) %>%  
 top\_n(10, mean\_proportion\_diplomes) %>%  
 arrange(ordre\_activite, desc(mean\_proportion\_diplomes))  
  
#View(top\_states)  
  
# Convertir ordre\_activite en facteur pour maintenir l'ordre des niveaux  
top\_states\_sorted <- top\_states %>%  
 arrange(ordre\_activite, mean\_proportion\_diplomes)

##### Aucune activité physique:

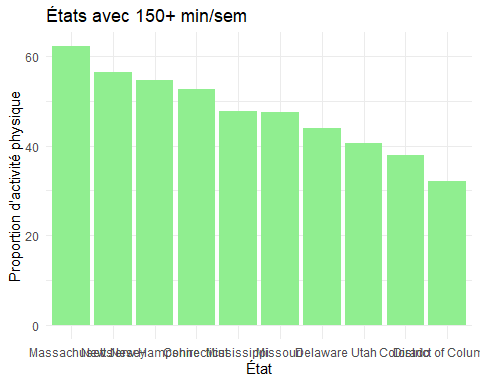
!!!!!!refaire les titres:

graph\_0pa <- ggplot(subset(top\_states\_sorted, ordre\_activite == 1),   
 aes(x = reorder(state, -proportion\_activite),   
 y = proportion\_activite)) +  
 geom\_bar(stat = "identity", fill = "skyblue") +  
 labs(title = "États avec Aucune activité physique",  
 x = "État", y = "Proportion d'activité physique") +  
 theme(panel.grid.major.x = element\_blank(),  
 axis.text.x = element\_text(angle = 45, hjust = 1),  
 axis.line = element\_line(colour = "black")) +  
 theme\_minimal()  
graph\_0pa



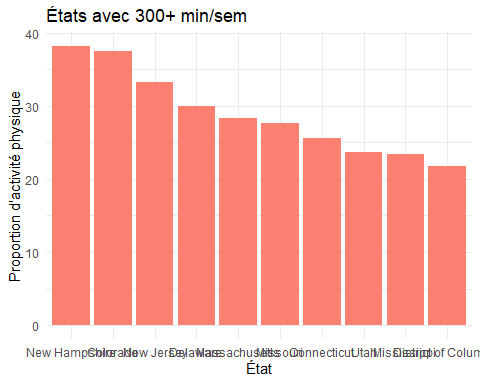
##### 150 min/sem

# Graphique pour "150+ min/sem"  
graph\_150min\_pa <- ggplot(subset(top\_states\_sorted, ordre\_activite == 2),   
 aes(x = reorder(state, -proportion\_activite),   
 y = proportion\_activite)) +  
 geom\_bar(stat = "identity", fill = "lightgreen") +  
 labs(title = "États avec 150+ min/sem",  
 x = "État", y = "Proportion d'activité physique") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1),  
 axis.line = element\_line(colour = "black"))+  
 theme\_minimal()  
  
graph\_150min\_pa



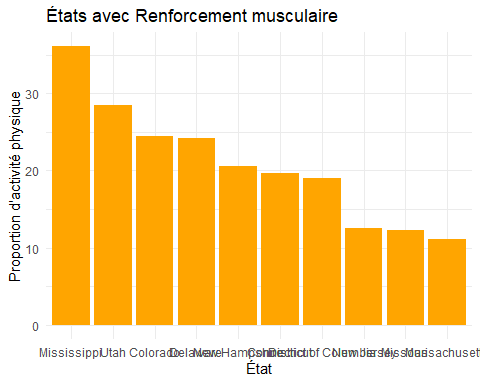
##### 300 + min/sem

# Graphique pour "300+ min/sem"  
graph\_300\_min <- ggplot(subset(top\_states\_sorted, ordre\_activite == 3),   
 aes(x = reorder(state, -proportion\_activite),   
 y = proportion\_activite)) +  
 geom\_bar(stat = "identity", fill = "salmon") +  
 labs(title = "États avec 300+ min/sem",  
 x = "État", y = "Proportion d'activité physique") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1),  
 axis.line = element\_line(colour = "black"))+  
 theme\_minimal()  
  
graph\_300\_min



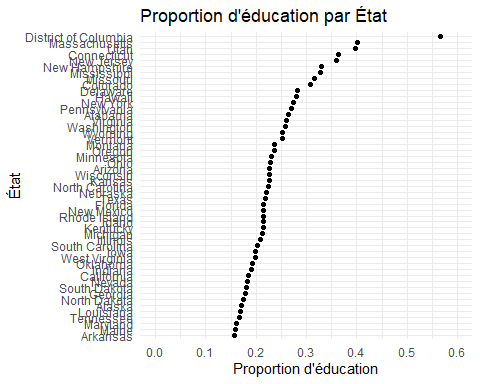
##### renforcement musculaire

# Graphique pour "Renforcement musculaire"  
graph\_muscu <- ggplot(subset(top\_states\_sorted, ordre\_activite == 4),   
 aes(x = reorder(state, -proportion\_activite),   
 y = proportion\_activite)) +  
 geom\_bar(stat = "identity", fill = "orange") +  
 labs(title = "États avec Renforcement musculaire",  
 x = "État", y = "Proportion d'activité physique") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1),  
 axis.line = element\_line(colour = "black"))+  
 theme\_minimal()  
  
graph\_muscu



#### 2.3.2 Visualisation des proportion d’éducation:

mean\_education <- pa\_education %>%  
 filter(!is.na(diplomes\_universitaires)) %>%  
 group\_by(state) %>%  
 summarize(mean\_diplomes\_universitaires = mean(diplomes\_universitaires))  
  
# Trier les États en fonction de la moyenne de diplomes\_universitaires  
pa\_education\_sorted <- pa\_education %>%  
 left\_join(mean\_education, by = "state") %>%  
 filter(!is.na(mean\_diplomes\_universitaires)) %>%  
 arrange(desc(mean\_diplomes\_universitaires))  
  
# Créer un graphique ggplot pour la proportion d'éducation par État  
ggplot(pa\_education\_sorted, aes(x = reorder(state, mean\_diplomes\_universitaires), y = diplomes\_universitaires)) +  
 geom\_point(stat = "identity", fill = "skyblue") +  
 scale\_y\_continuous(limits = c(0, 0.6),  
 breaks = seq(0, 0.6, by = 0.1))+  
 labs(title = "Proportion d'éducation par État",  
 x = "État", y = "Proportion d'éducation") +  
 theme( panel.grid.minor.y = element\_blank(),  
 axis.text.x = element\_text(angle = 45, hjust = 1),  
 axis.line = element\_line(colour = "black")) +  
 theme\_minimal() +  
 coord\_flip()

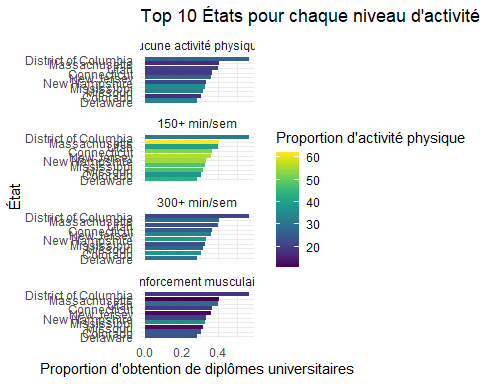


## 3- Analyse par État:

# Créer un graphique ggplot avec des facettes pour chaque niveau d'ordre\_activite  
super\_graph1 <- ggplot(top\_states\_sorted, aes(x = reorder(state, mean\_proportion\_diplomes), y = mean\_proportion\_diplomes, fill = proportion\_activite)) +  
 geom\_bar(stat = "identity") +  
 labs(title = "Top 10 États pour chaque niveau d'activité physique",  
 x = "État", y = "Proportion d'obtention de diplômes universitaires",  
 fill = "Proportion d'activité physique") +  
 facet\_wrap(~ ordre\_activite, scales = "free\_y", ncol = 1,labeller = labeller  
 (ordre\_activite = c("1" = "Aucune activité physique",  
 "2" = "150+ min/sem",  
 "3" = "300+ min/sem",   
 "4" = "renforcement musculaire"))) +  
 scale\_fill\_viridis(discrete = TRUE) + # Utiliser la palette de couleurs viridis) +  
 scale\_fill\_viridis() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1),  
 legend.position = "top") +  
 coord\_flip() +  
 theme\_minimal()

Scale for fill is already present.  
Adding another scale for fill, which will replace the existing scale.

super\_graph1



#### Méthodes d’analyse

Le but de l’analyse des données et d’essayer de voir s’il y a une relation entre le niveau d’activité physique et de santé et le niveau d’éducation de la population des États-Unis. Pour se faire nous allons procéder a des analyse statistiques afin de démontrer que notre hypothèse selon laquelle, les États américains où le niveau d’activité physique est le plus élevé auront également le plus haut niveau d’éducation en pourcentage.

1. **Analyse descriptives** : Pour décrire les caractéristiques de la population étudiée en termes d’activité physique, de santé et de niveau d’éducation, en utilisant des statistiques descriptives telles que les moyennes, les médianes et les écarts-types.
2. **Analyse de corrélation** : Pour examiner les relations entre l’activité physique, la santé et le niveau d’éducation, en utilisant des techniques telles que la corrélation de Pearson ou de Spearman pour mesurer les associations entre les variables.
3. **Régression linéaire** : Pour étudier comment l’activité physique influence la santé et le niveau d’éducation, en contrôlant les variables confondantes potentielles.

Vos conseils sont le bienvenue sur quels types d’analyse je devrais faire!Analyse:

summary\_data <- pa\_education %>%  
 drop\_na(diplomes\_universitaires) %>%  
 group\_by(niveau\_activite) %>%  
 summarize(mean\_proportion\_diplomes = mean(diplomes\_universitaires),  
 median\_proportion\_diplomes = median(diplomes\_universitaires),  
 sd\_proportion\_diplomes = sd(diplomes\_universitaires))  
summary\_data

# A tibble: 4 × 4  
 niveau\_activite mean\_proportion\_diplomes median\_proportion\_diplomes  
 <chr> <dbl> <dbl>  
1 150+ min/sem 0.243 0.225  
2 300+ min/sem 0.243 0.225  
3 Aucune activité physique 0.243 0.225  
4 renforcement musculaire 0.243 0.225  
# ℹ 1 more variable: sd\_proportion\_diplomes <dbl>

# Analyse de régression du niveau d'activité physique en fonction du niveau d'éducation  
model <- lm(diplomes\_universitaires ~ ordre\_activite + state\_median\_income + uni\_au\_25km + uni\_au\_50km + uni\_au\_75km + uni\_au\_100km, data = pa\_education)  
  
# Résumé de la régression  
summary(model)

Call:  
lm(formula = diplomes\_universitaires ~ ordre\_activite + state\_median\_income +   
 uni\_au\_25km + uni\_au\_50km + uni\_au\_75km + uni\_au\_100km, data = pa\_education)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.132119 -0.029491 -0.008042 0.039953 0.184580   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 7.006e-02 5.672e-03 12.351 < 2e-16 \*\*\*  
ordre\_activite -2.737e-05 7.510e-04 -0.036 0.9709   
state\_median\_income 3.502e-06 9.286e-08 37.709 < 2e-16 \*\*\*  
uni\_au\_25km 7.087e-04 3.707e-05 19.120 < 2e-16 \*\*\*  
uni\_au\_50km -8.921e-05 5.213e-05 -1.711 0.0871 .   
uni\_au\_75km -3.592e-04 4.506e-05 -7.971 1.92e-15 \*\*\*  
uni\_au\_100km 1.725e-04 1.534e-05 11.243 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.06045 on 5174 degrees of freedom  
 (296 observations effacées parce que manquantes)  
Multiple R-squared: 0.3594, Adjusted R-squared: 0.3587   
F-statistic: 483.9 on 6 and 5174 DF, p-value: < 2.2e-16

Coefficients :

Intercept : L’intercept est de 0.2666. Cela signifie que lorsque toutes les autres variables sont nulles, la proportion de diplômés universitaires est estimée à 0.2666. ordre\_activite : Le coefficient est -2.357e-05, mais il n’est pas significatif (p-value = 0.978). Cela suggère qu’il n’y a pas de relation significative entre le niveau d’activité physique et la proportion de diplômés universitaires. uni\_au\_25km, uni\_au\_50km, uni\_au\_75km, uni\_au\_100km : Ces variables représentent le nombre d’universités à différentes distances de l’État. Par exemple, pour uni\_au\_25km, chaque augmentation d’une unité dans le nombre d’universités dans un rayon de 25 km est associée à une augmentation de 0.0004545 dans la proportion de diplômés universitaires. Residuals :

Les résidus (erreurs de prédiction) semblent être distribués symétriquement autour de zéro, ce qui indique que le modèle de régression est approprié.