TP 2 - Étude sur la santé mentale et le taux de suicide aux États-Unis (Option 1)

Yun Shen (Yoshi) Li

Introduction

Ces dernières années, avec l'impact de l'isolation durant la pandémie de la COVID-19, il semble que la population partout dans le monde souffre de taux alarmant de santé mentale catatrophique ainsi qu'un taux de suicide qui est de plus en plus élevé. Personellement, la santé mentale a toujours été un sujet touchant pour moi, mais c'est exactement la raison pour laquelle j'aimerais faire cette étude: pour voir les données pour moi-même au lieu de me référer à des articles sur l'Internet qui me raconte à quel point le monde souffre.

Parfois, prendre soins de sa santé mentale peut simplement dire de se détendre avec un livre, se reposer après une longue journée, s'exercer physiquement bref, prendre soins de soi à travers la relaxation. Cela dit, parfois les gens sont malheureusement affectés par des troubles mentaux qui nécessitent de l'aide professionel, que ce soit avec un conseiller/psychiatre et/ou des médicaments pour essayer de réguler ses symptômes.

Dans ma recherche pour des données disponibles sur le taux de suicide et des données de questionnaire sur la santé mentale des gens et comment ils se sentent, j'ai trouvé des données plus récentes pour le taux de suicide de 1999 à 2020 aux États-Unis (par état) et un questionnaire sur comment les gens décrivent leur santé mentale dans le moment par le CDC (conduit de 2017 à 2020).

Bref, ma question de recherche est l'effet de la santé mentale sur la possibilité de suicide d'un individu aux États-Unis. La dépression est un trouble mental qui affecte les émotions de manière variée de personne en personne, mais généralement cette maladie la rend souvent incapable d'exprimer de la joie véritable et de se sentir très mal émotionellement.

Données et méthodes

Comme mentionné ci-dessus, j'aimerais étudier l'effet de la santé mentale sur la possibilité d'un individu à se suicider. Je vais utiliser la base de donnée 'suicide_rate' qui donne le taux

de décès par suicide de chacun des 51 états pour obtenir donc le taux nationale. Ensuite, je vais additioner les rangées pour obtenir le score sur 27 de chaque répondant du questionnaire, et donc obtenir un pourcentage pour signifier le "pourcentage" d'être déprimé d'une personne. Finalment, je vais combiner les données pour

Pour cette recherche, j'ai choisi l'option 1 en croisant:

- Des données d'un questionnaire/sondge par le Center for Disease Control and Prevention (CDC) des États-Unis. Les questions incluent "Having little interest in doing things", "Feeling down, depressed, hopeless", bref des questions pour voir si cette personne peut être atteint par la depression ou non. Les individus répondent avec comment ils se sentent d'après ce code: 0 (not at all), 1 (several days), 2 (more than half the days), 3 (nearly everyday), 7 (refused) et 9(don't know). À part l'état d'origine du répondant et leurs réponses, il n'y a aucun d'autre variable considérée.
- Le taux de décès par suicide catégorisé par état, avec des données recueillis de 2009 à 2020 (nommés suicide_rate). Les données sur le site du CDC sont catégorisés sous "intentional self-harm" dans la banque de données des décès, que je considère dans cette analyse comme décès par suicide. Il y a un score total de 27, avec 0 décrivant une personne complètement non-déprimé, et 27 décrivant une personne à risque d'être atteinte par la dépression.

Transformation des variables

J'ai importé 'tidyverse', 'dplyr' et 'haven' pour m'aider dans cette analyse.

Depression screener CDC

Pour les données 'Depression screener CDC' que j'ai obtenu du site officiel du CDC, j'ai d'abord commandé R à omettre les variables NA puisqu'il y en avait beaucoup et ils servaient à rien, et j'ai aussi enlevé la variable SEQN qui correspondait à un code d'identification des répondants.

Ensuite, il y avait beaucoup d'observations pour ce sondage, alors j'ai commandé R de me choisir 50 individus de façon aléatoire, que j'ai enregistré sous "random_screener". Pour chacun de ces 50 individus, j'ai additionné leurs réponses pour ensuite diviser par 27 et obtenir un pourcentage que j'utiliserai pour mesurer leur "percentage of depression" (sous la variable "precent_depressed").

- Ceci a été fait en disant à R d'additionner les rangées débutant par "DPQ" (catégorie de réponse)

Finalement, j'ai supprimé les colonnes dont j'avais plus besoin et renommé une variable pour faire plus beau

Suicide Rate

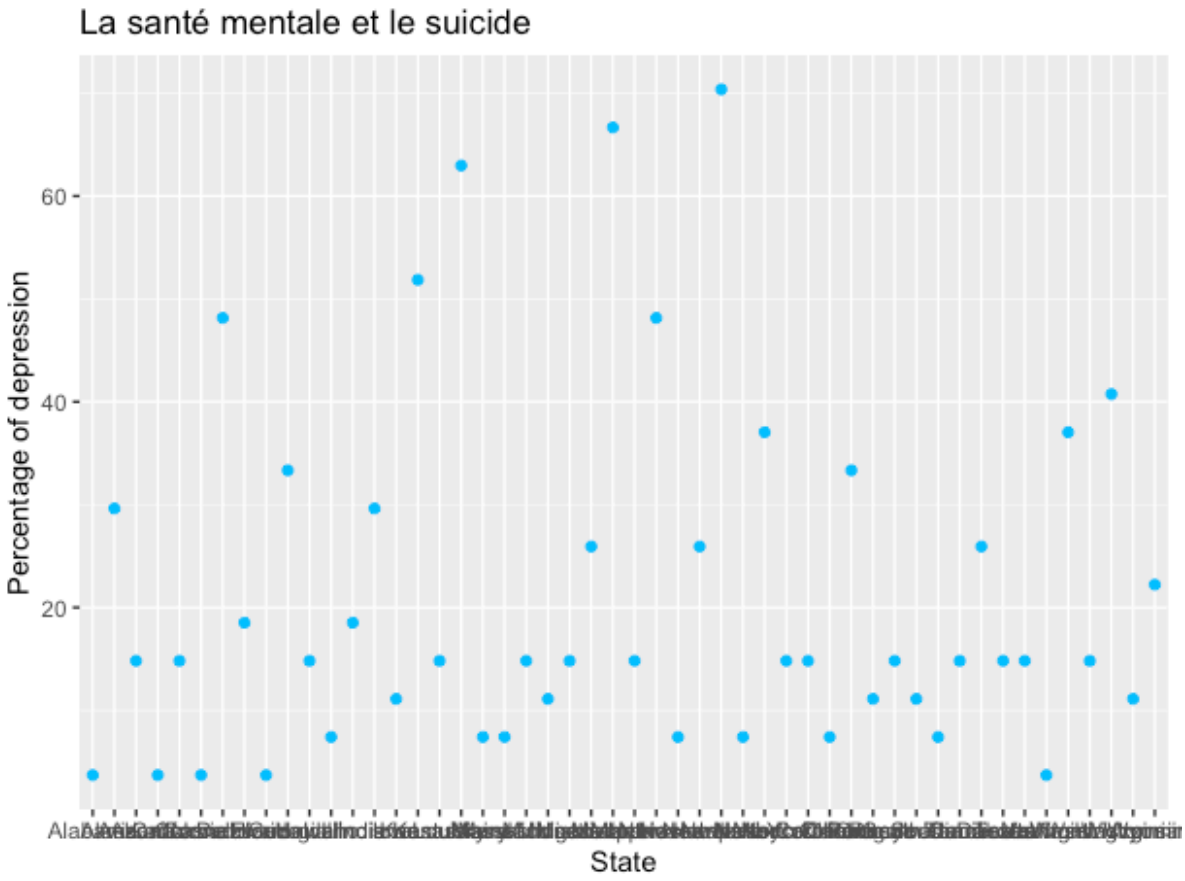
J'ai d'abord importé le fichier 'Multiple.Cause.of.Death, .1999. 2020' et renommé 'suicide_rate'. Il n'y avait pas grand chose à nettoyer, alors j'ai commandé R à omettre les valeurs NA comme d'habitude, ainsi que variables 'Notes' des chercheurs, le 'State.Code' et 'Crude.Rate'.

Puisque les données sont par états, j'ai additionné la somme totale des décès par suicide et la population totale. Ces données nettoyées sont enregistrées sous 'suicide_rate_clean'.

Croisement de données

En me basant sur la variable 'State', le seul truc en commun de mes deux bases de données, j'ai fait un croisement 'left_join' de 'random_screener' et 'suicide_rate_clean'. La variable 'State' a été mise en première et j'ai supprimer les colonnes commençants par 'DPQ' car j'en ai plus besoin. J'ai ensuite renommé 'Deaths' à 'deaths_by_suicide' pour faire plus claire. Finalement, j'ai ajouté une variable 'Rate_of_suicide' en divisant 'deaths_by_suicide' par 'Population'. Le croisement de ces deux bases de données ont été enregistré sous 'effect'.

Pour le graphique, j'ai appeler 'ggplot2' avec 'library()' en nommant ce graphique 's' pour faire simple. Les données viennent de 'effect' et j'ai dit à R ce que je voulais comme variable indépendante et dépendante. Le graphique est un graphique à point avec variable indépendant 'State' et dépendante 'Percentage of depression'. Finalement, j'ai renommé les axes et donné le titre "La santé mentale et le suicide" à mon graphique.



Honnêtement, je ne sais pas trop il représente quoi le graphique et mes données, et si je peux même répondre à ma question de recherche. J'ai passé beaucoup de temps à chercher des données mais peut-être elles sont pas les meilleurs pour ma question. Au moins j'ai pu utiliser des commandes que j'ai appris en cours FAS 1001 et 1003, mais bon.

Je vais quand même tenter de faire une analyse: d'après le graphique, les individus dans certains états semblent êtres plus vulnérables à être atteint par la dépression, et donc aussi plus vulnérables à se suicider. Cela peut-être à cause d'une multitude de raisons, économique, physique, financière, médicale, etc etc. Bref, une personne peut sembler complètement normale, mais on ne sait jamais ce que les gens sont en train de vivre. Mieux vaut être plus gentil/gentille :)

Annexe

`lrl`{r eval=FALSE} #FAS 1001 - research TP 2

#libraries library(tidyverse) library(dplyr) library(haven)

#US survey for mental health status based on individual's emotions (ranking out of 27) file_path <- "/Users/yoshili/fas_1001_Li/depression screener CDC.xpt" depression_screener <- read_xpt(file_path) #open xpt file in R

#cleaning depression_screener_clean <- na.omit(depression_screener) |> select(-SEQN) #there's too many observations so I'm going to have R randomly select 51 of them, one per state random_screener <- depression_screener_clean |> sample_n(50, replace = FALSE) #addition to obtain score out of 27 screener_sums <- rowSums(random_screener, na.rm = TRUE) print(screener_sums)

#displaying individuals' scores random_screener <- random_screener %>% mutate(score = rowSums(select(., starts_with("DPQ")))) random_screener <- random_screener %>% mutate(percent_depressed = score / 27) random_screener <- random_screener %>% mutate(percentage = percent_depressed * 100) #after performing all operations, delete the unneeded columns random_screener <- random_screener |> select(-score, -percent_depressed)

random_screener <- random_screener |> rename(percent_depressed = percentage)

#probably could've just used the pipe over and over again #but i forgot and got too lazy to change it so eh :P

#rate of death by suicide in US suicide_rate <- `Multiple.Cause.of.Death,.1999.2020` #file imported directly from CDC data bank rm(`Multiple.Cause.of.Death,.1999.2020`)

#cleaning suicide_rate_clean <- na.omit(suicide_rate) |> #omitting cells with NA select(-Notes, -State.Code, -Crude.Rate) #omitting irrelevant variables

#obtaining rate of death by suicide column_sums <- colSums(suicide_rate_clean[, -which(names(suicide_rate_clean) == "State")], na.rm = TRUE) print(column_sums) #sum deaths by suicide: 840204, US population sum: 6746356647

#joining the two cleaned datasets effect <- left_join(random_screener, suicide_rate_clean, by = "State")

effect <- effect |> select(State, everything()) #to put State in the front effect <- effect |> select(-DPQ010, -DPQ020, -DPQ030, -DPQ040, -DPQ050, -DPQ060, -DPQ070, -DPQ080, -DPQ090, -DPQ100)

effect <- effect |> rename(Deaths_by_suicide = Deaths)

effect <- effect |> mutate(Rate_of_suicide = Deaths_by_suicide/Population)

"'