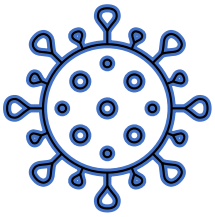
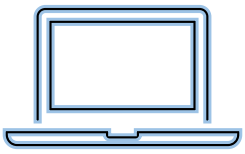
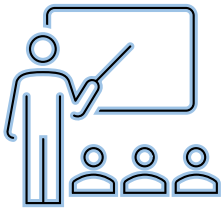
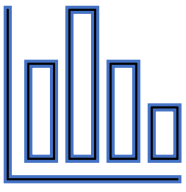
IMPACT DE LA PANDEMIE COVID-19 SUR L'APPRENTISSAGE AUX USA

4BIGF – Big Data Fundamentals









BENYOUCEF Lynda
HULLOUX Guillaume
LEPLAE Louan

2022
Campus de Lille

Table des matières

Introduction

La COVID-19, une pandémie apparue en novembre 2019 à Wuhan en Chine, a entraîné, par sa virulence, une multitude de défis sans précédent pour l'apprentissage et l'éducation des élèves à travers le monde. La flambée mondiale de cas de COVID-19 a contribué à la fermeture de plusieurs écoles, collèges et universités en 2020 dans presque toutes les régions du monde et le passage à l'apprentissage en ligne, ce qui a eu un impact sur l'apprentissage des élèves de différentes manières.

Cela a conduit les éducateurs et les étudiants à passer plus que jamais du temps sur Internet, et par conséquent s'adapter à l'apprentissage à distance en apprenant et se familiarisant avec les informations, les outils, les applications mis à leur disposition.

L'étude de ces comportements web, sous la forme d'exploration et d'analyse de Big Data, provenant de différents pays du monde offre la possibilité d'identifier, d'enquêter et de quantifier les besoins, les intérêts et les défis liés à l'apprentissage en ligne dans différents pays du monde imposés par la COVID-19.

A travers ce projet, nous utiliserons des données concernant les étudiants américains ainsi que des outils d'analyse de données pour déterminer les tendances de l'apprentissage numérique et son efficacité envers les communautés étudiées. Nous comparerons les districts et les États sur des facteurs tels que la démographie, l'accès à Internet, l'accès aux produits d'apprentissage et les finances. Enfin, nous résumerons notre rapport et indiquerons les limites et recommandations qui découlent de notre analyse des données disponibles.

Objectifs

L'objectif de ce projet est d'explorer l'état de l'apprentissage numérique en 2020 et comment l'engagement de l'apprentissage numérique est lié à des facteurs tels que la démographie du district ; mais aussi, d'analyser les ensembles de données, trouver des informations significatives et créer des visualisations simples et pertinentes.

Nous étudierons l'engament des étudiants à l'apprentissage à distance selon cinq axes :

- L'état de la connectivité et engagement numérique en 2020
- L'effet de la COVID-19 sur l'apprentissage à distance
- La relation entre l'engagement étudiant et les technologies éducatives
- Le potentiel lien entre l'engagement étudiant et les facteurs sociogéographiques
- Le rôle des politiques d'Etat dans l'engagement étudiant

Données disponibles

Le jeu de données fourni est un ensemble de données sur l'engagement « edtech » de plus de 200 districts scolaires des Etats-Unis en 2020. Celui-ci inclut trois ensembles de fichiers de base.

Engagement étudiant

Il s'agit d'un dossier comportant un ensemble de fichiers .csv basé sur l'extension Student Chrome de la plateforme LearnPlatform. L'extension collecte les événements de chargement de page de plus de 10 000 produits de technologie éducative dans la bibliothèque de produits, y compris les sites Web, les applications, les applications Web, les programmes logiciels, les extensions, les livres électroniques, les matériels et les services utilisés dans les établissements d'enseignement. Les données sur l'engagement ont été agrégées au niveau du district scolaire, et chaque fichier représente les données d'un district scolaire.

Les données d'engagement sont agrégées au niveau du district scolaire, et chaque fichier du dossier **engagement_data** représente les données d'un district scolaire. Le nom de fichier à 4 chiffres représente **district_id** qui peut être utilisé pour établir un lien vers les informations du district, **lp_id** peut être utilisé pour créer un lien vers des informations sur le produit de technologie éducative.

Nom	***Description***
temps	Date en "AAAA-MM-JJ"
lp_id	Identifiant unique du produit
pct_access	Pourcentage d'élèves du district ayant au moins un événement de chargement de page d'un produit donné et un jour donné
engagement_index	Nombre total d'événements de chargement de page pour mille étudiants d'un produit donné et un jour donné

Technologies éducatives

Ces données se présente sous la forme d'un fichier .csv comprenant des informations sur les caractéristiques des 372 meilleurs produits avec le plus d'utilisateurs en 2020. Les catégories répertoriées dans le fichier **products_info.csv** font partie de la taxonomie des produits de LearnPlatform. Certains produits peuvent ne pas avoir d'étiquettes en raison d'un doublon, d'un manque d'URL précise ou d'autres raisons.

Nom	***Description***
LP ID	Identifiant unique du produit
URL	Lien Web vers le produit spécifique
Product Name	Nom du produit spécifique
Provider/Company Name	Nom du fournisseur du produit
Sector(s)	Secteur de l'éducation où le produit est utilisé

Nom	***Description***
Primary Essential Function	Fonction de base du produit. Il y a deux couches d'étiquettes ici. Les produits sont d'abord étiquetés dans l'une de ces trois catégories : LC = Apprentissage et programme d'études, CM = Gestion de classe et SDO = Opérations de l'école et du district. Chacune de ces catégories comporte plusieurs sous-catégories avec lesquelles les produits ont été étiquetés

Districts scolaires

Il s'agit d'un fichier .csv comprenant des informations sur les caractéristiques des districts scolaires, y compris des données du NCES (pour la connexion IP) et de la FCC (pour les impayés).

Dans cet ensemble de données, **district districts_info.csv**, les informations identifiables sur les districts scolaires ont été anonymisés. À des fins de généralisation des données, certains points de données sont publiés avec une plage dans laquelle se situe la valeur réelle. De plus, il existe de nombreuses données manquantes marquées comme "NaN", indiquant que les données ont été supprimées pour maximiser l'anonymisation de l'ensemble de données.

Nom	***Description***
distict_id	Identifiant unique du district scolaire
state	Etat où le district se trouve
local	Classification locale NCES qui classe le territoire américain en quatre types de zones : ville, banlieue, ville et campagne
pct_black/hispanic	Pourcentage d'élèves dans les districts identifiés comme noirs ou hispaniques sur la base des données du NCES 2018-19
pct_free/reduced	Pourcentage d'élèves dans les districts éligibles à un déjeuner gratuit ou à prix réduit sur la base des données du NCES 2018-19
county_connections_ratio	Ratio (connexions haut débit fixes résidentielles supérieures à 200 kbps dans au moins une direction/foyers) sur la base des données au niveau du comté de FCC From 477 (version de décembre 2018)
pp_total_raw	Dépenses totales par élève (somme des dépenses locales et fédérales) d'après le projet National Education Resource Database on Schools (NERD$) d'Edunomics Lab

Préparation des données

Pour commencer nous avons importé les librairies utilisées dans le scipt afin de réaliser les analyses :

```
library(ggplot2)
library(skimr)
library(readr)
library(plyr)
library(dplyr)
```

- Nous utilisons gplot2 afin de réaliser des graphiques
- Nous utilisons skimr pour avoir des informations complémentaires sur les données provenant des csv
- La librairie readr nous sert pour lire les fichiers csv.
- Plyr et dplyr nous servent pour la réalisation d'opérations ce certaines opérations pour la création de nos graphiques.

Avant de passer à l'analyse, nous modifions ensuite le chemin afin de ne pas avoir de problème en changeant d'environnement.

```
if (substr(getwd(),3,14) != "/4BIGF/data") {
  setwd(paste(getwd(),"data", sep = "/"))
}
```

Tout d'abord, pour pouvoir analyser les données provenant d'une source externe, nous devons savoir quel type de données nous allons traiter. Pour cela, nous avons réalisé des rapports sur les fichiers. Nous avons utilisé la fonction spec() afin d'avoir les listes des colonnes et des informations sur le type de données contenues :

```
>   spec(districts)
cols(
  district_id = col_double(),
  state = col_character(),
  locale = col_character(),
  `pct_black/hispanic` = col_character(),
  `pct_free/reduced` = col_character(),
  county_connections_ratio = col_character(),
  pp_total_raw = col_character()
)
>   spec(products)
cols(
  `LP ID` = col_double(),
  URL = col_character(),
  `Product Name` = col_character(),
  `Provider/Company Name` = col_character(),
  `Sector(s)` = col_character(),
  `Primary Essential Function` = col_character()
)
```

Figure 1 - Détails des colonnes

Ensuite, nous avons utilisé la fonction skim(), provenant de la librairie skimr, afin d'avoir un rapport sur le contenu des colonnes. Comparé à la fonction summary(), la fonction skim() nous permet de nous fournir plus d'informations et de manières plus simple.

```
>   skim(districts)
-- Data Summary ------------------------
                           Values
Name                       districts
Number of rows             233
Number of columns          7
_________________________
Column type frequency:
  character                6
  numeric                  1
_________________________
Group variables            None

-- Variable type: character ------------------------------------------------------------------------
# A tibble: 6 x 8
  skim_variable            n_missing complete_rate   min   max empty n_unique whitespace
* <chr>                         <int>         <dbl> <int> <int> <int>    <int>      <int>
1 state                             0             1     3    20     0       24          0
2 locale                            0             1     3     6     0        5          0
3 pct_black/hispanic                0             1     3    10     0        6          0
4 pct_free/reduced                 28             0.880 3    10     0        6          0
5 county_connections_ratio         14             0.940 3     9     0        3          0
6 pp_total_raw                     58             0.751 3    14     0       12          0

-- Variable type: numeric --------------------------------------------------------------------------
# A tibble: 1 x 11
  skim_variable n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
* <chr>            <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 district_id          0             1 5220. 2596.  1000  2991  4937  7660  9927 ▇▇▇▇▇
```

Figure 2 - Skim fichier districts

```
>   skim(products)
-- Data Summary ------------------------
                           Values
Name                       products
Number of rows             372
Number of columns          6
_________________________
Column type frequency:
  character                5
  numeric                  1
_________________________
Group variables            None

-- Variable type: character ------------------------------------------------------------------------
# A tibble: 5 x 8
  skim_variable             n_missing complete_rate   min   max empty n_unique whitespace
* <chr>                         <int>         <dbl> <int> <int> <int>    <int>      <int>
1 URL                               0             1    14   101     0      372          0
2 Product Name                      0             1     2    45     0      372          0
3 Provider/Company Name             1             0.997 3    55     0      290          0
4 Sector(s)                        20             0.946 7    29     0        5          0
5 Primary Essential Function       20             0.946 11   73     0       35          0

-- Variable type: numeric --------------------------------------------------------------------------
# A tibble: 1 x 11
  skim_variable n_missing complete_rate   mean    sd    p0   p25   p50   p75  p100 hist
* <chr>            <int>         <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 LP ID                0             1 54566. 26248. 10533 30451 53942. 77497 99916 ▇▇▇▇▇
```

Figure 3 - Skim fichier products

Nous chargeons ensuite les données en créant les data frame products et districts :

```
# Districts
districts <<- read_csv("districts_info.csv")
```

```
# Products
products <<- read_csv("products_info.csv")
```

Nous avons remarqué que tous les fichiers contenus dans le dossier « engagement_data » contiennent les mêmes colonnes. Nous avons donc créé un dataframe appelé « tabl » il contient toutes les données dans les fichiers de « engagement_data ». Cela nous permet d'accéder plus simplement aux données.

```
# tabl
id_list <- districts$district_id
for (i in id_list){
  file <- c(i, ".csv$")
  list_id_file <- list.files(path = paste(getwd(),"engagement_data", sep = "/"),
                             recursive = TRUE,
                             pattern = file,
                             full.names = TRUE)
  new_tabl <- read_csv(list_id_file)
  new_tabl$district_id <- i
  if (exists("tabl") && is.data.frame(get("tabl"))) {
    tabl <- rbind(tabl, new_tabl)
  } else {
    tabl <- new_tabl
  }
}
id_list <- products$`LP ID`
product_name_list <- products$`Product Name`
for(i in id_list){
  tabl$productName[tabl$lp_id == i] <- product_name_list[match(i, id_list)]
                                       }
```

Nous avons ajouté à ce dataframe le product_name et le district_id pour des traitements.

Dans le but de faciliter les analyses, nous commençons par regrouper les trois sources de données, concernant l'engagement étudiant, les technologies éducatives, et les districts scolaires, en une unique source.

Pour constituer la base commune, on utilise les noms à 4 chiffres des fichiers présents dans le dossier **engagement_data**, qui représente la valeur **district_id** pour établir un lien vers les informations du district du fichier **district_info.csv** ; et on utilise le champ **lp_id** de ces fichiers pour créer un lien vers des informations sur le produit de technologie éducative du fichier **product_info.csv**.

Code R :

```
complete_base <- merge(x=tabl,y=districts,by="district_id",all.x=TRUE)
complete_base   <-   merge(x=complete_base,y=products,by.x="lp_id",   by.y="LP
ID",all.x=TRUE)
```

Une fois la source de données définie, nous allons mettre en lumière les proportions de valeurs inconnues ou non définies pour avoir un aperçu sur la précision de nos analyses.

Valeurs manquantes

Fichiers dans Engagement_data

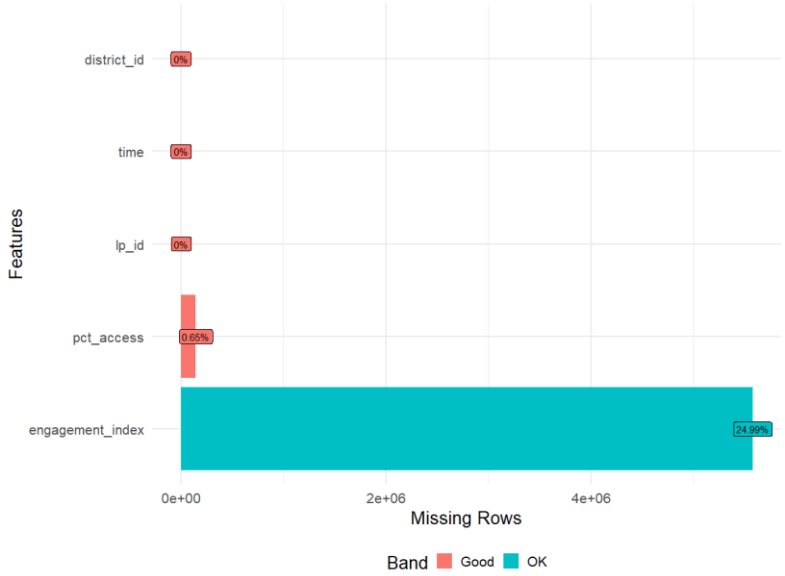


Figure 4 - Valeurs manquantes engagement_data

Ce diagramme en bar horizontal nous montre les valeurs pours lesquels certaines données sont NA (non disponible) dans l'ensemble des fichiers contenus dans le dossier engagement_data.

Code R :

```
# analyse missing values
  plot_missing(tabl, ggtheme = theme_minimal(base_size = 20))
```

Fichier districts_info

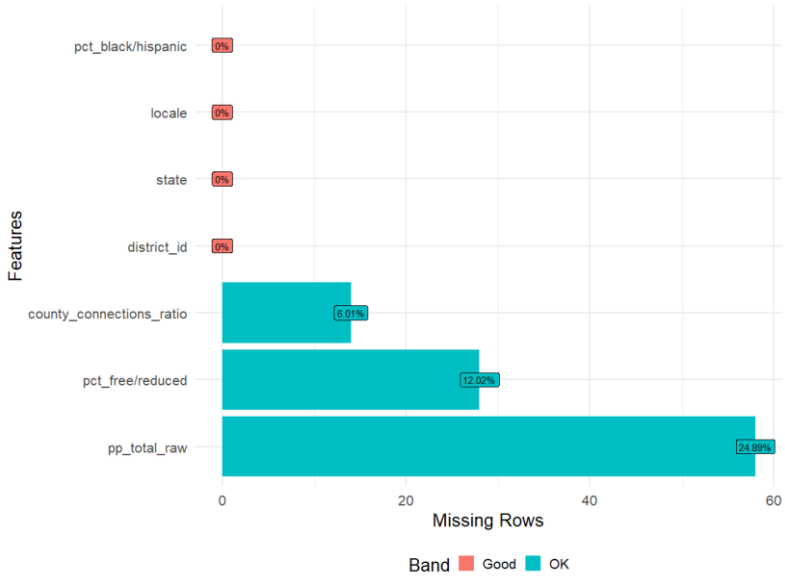


Figure 5 - Valeurs manquantes districts_info

Ce diagramme en bar horizontal nous montre les valeurs pours lesquels certaines données sont NA (non disponible) dans le fichier districts_info.csv.

Code R :

```
# analyse missing values
plot_missing(districts, ggtheme = theme_minimal(base_size = 20))
```

Fichier products_info

Figure 6 - Valeurs manquantes products_info

Ce diagramme en bar horizontal nous montre les valeurs pours lesquels certaines données sont NA (non disponible) dans le fichier products_info.csv.

Code R

```
# analyse missing values
plot_missing(products, ggtheme = theme_minimal(base_size = 20))
```

A la vue de ces graphiques, la proportion de valeurs manquantes reste raisonnable pour pouvoir effectuer des analyses pertinentes.

Connectivité et engagement numérique

Au plus fort de la pandémie, la connectivité numérique ne se réduit plus aux seuls moyens de communication traditionnels et à la recherche d'informations. Elle joue désormais un rôle clé pour les particuliers et les entreprises qui souhaitent utiliser les données, contenus et applications numériques pour assurer la continuité de l'activité économique et sociale malgré les règles de distanciation sociale et le confinement strict imposé presque partout.

D'autre part, cette situation a obligé tous les étudiants à s'engager davantage avec Internet, à tout faire en utilisant Internet. Les plateformes d'apprentissage numérique connaissent une croissance exponentielle plus que jamais. L'engagement des utilisateurs sur Internet atteint le plus haut de l'histoire et ne cesse de croître.

Informations générales

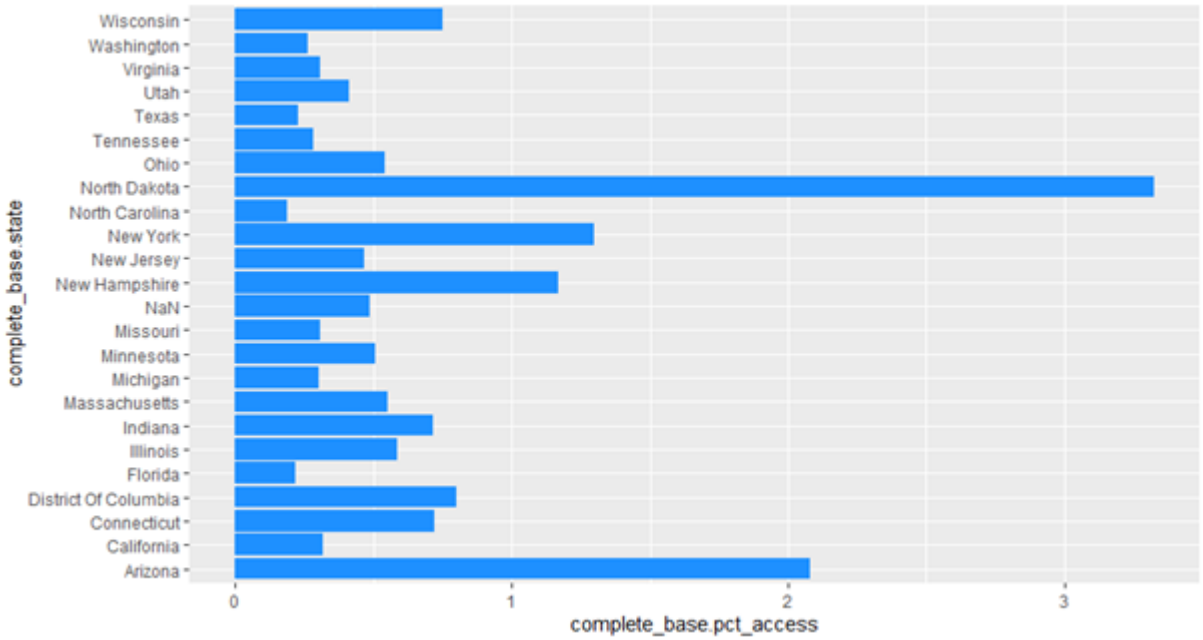


Figure 7 - Chargement de page par Etat

Ce graphique représente le pourcentage d'élèves du district ayant au moins un événement de chargement de page d'un produit en fonction de leur Etat.

On observe que le Dakota du Nord et l'Arizona ont les pourcentages les plus élevés.

La Floride et la Caroline du Nord ont les pourcentages les plus faibles.

Code R :

```
graph <- data.frame(complete_base$pct_access, complete_base$state)
ggplot(data=graph, aes(x=complete_base.pct_access, y=complete_base.state)) +
geom_bar(stat = "summary", fill="#1E90FF")
```

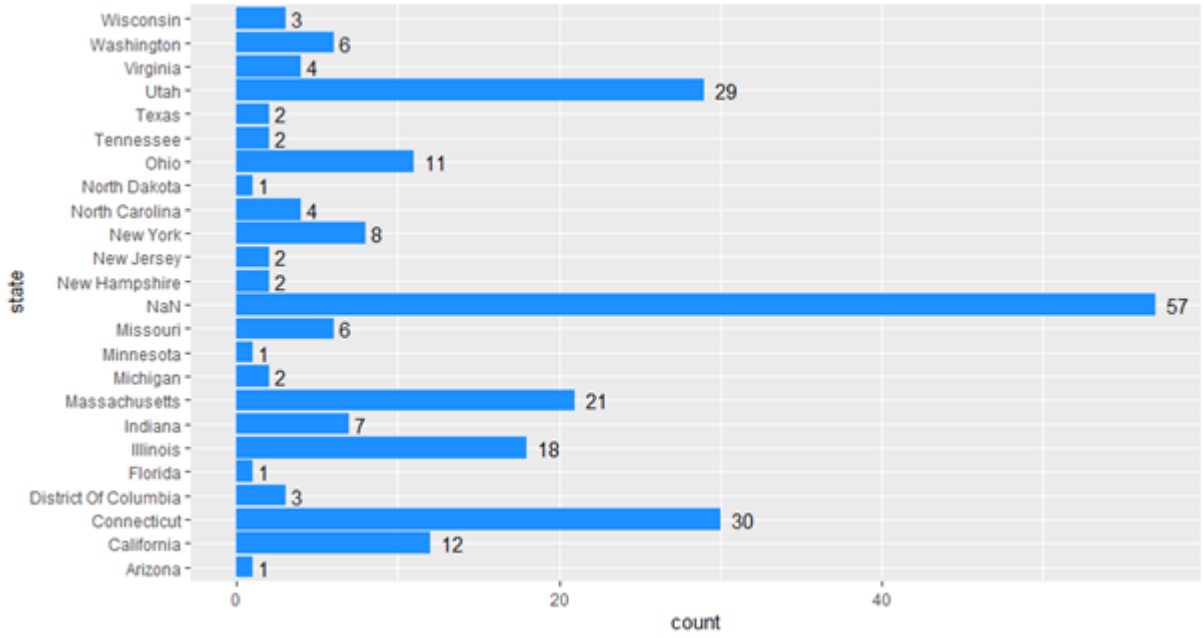


Figure 8 - Nombre de district scolaire par Etat

Ce diagramme en bar horizontales montre le nombre de districts scolaires dans chaque Etats.

On observe que :

- De nombreuses données sont inconnues
- Les Etats ayant le plus de districts sont le Connecticut et l'Utah
- Parmi les Etats ayant le moins de district, on retrouve l'Arizona et le Dakota du Nord qui avaient le plus grand pourcentage de page chargées (cf Figure7), ce qui explique les moyennes supérieurs aux autres Etats.

Code R :

```
ggplot(data=districts, aes(y=state)) + geom_bar(stat = "count",
fill="#1E90FF") + geom_text(aes(label=..count..), stat="count", hjust=-0.5)
```

Etat de la connectivité au niveau des comtés

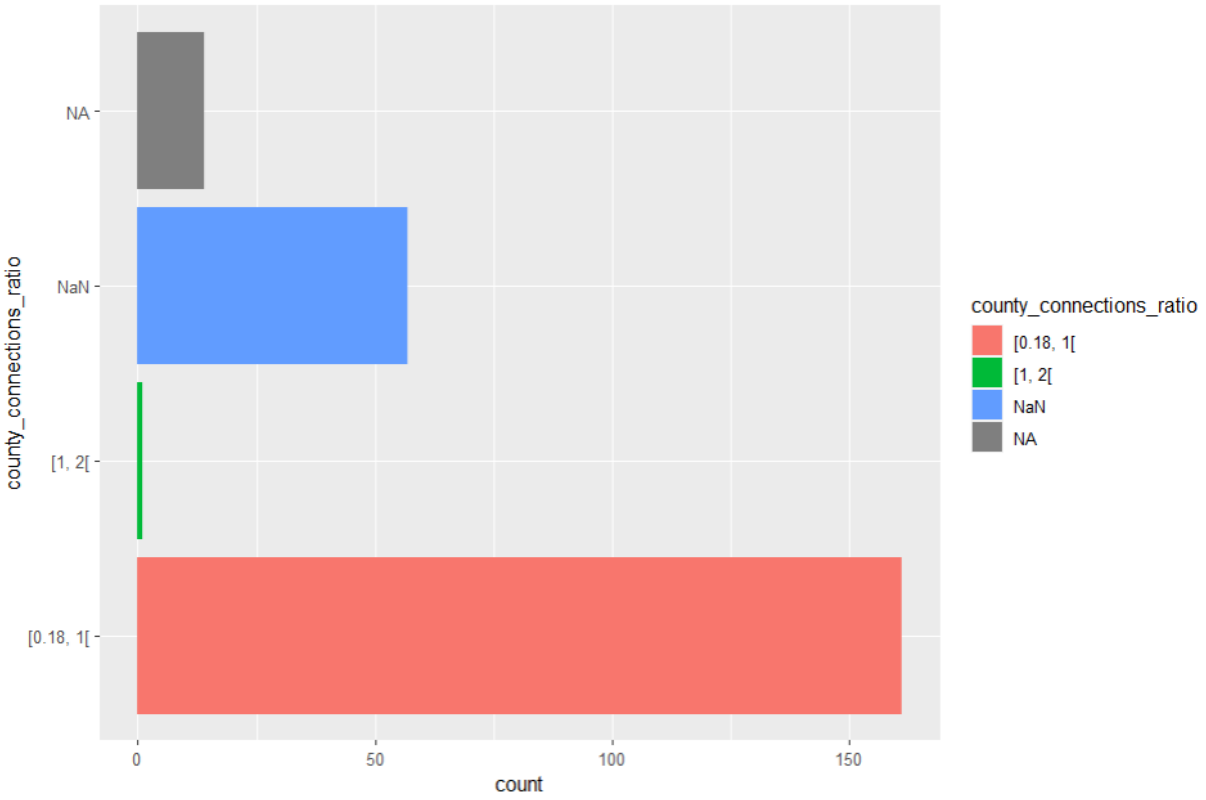


Figure 9 - Compte de county connections ratio

Ce graphique représente le nombre de districts scolaires par taux de connexion du comté, qui correspond aux connexions fixes résidentielles à haut débit de plus de 200 kbps dans au moins une direction/ménages.

On remarque que la plupart des états ont un taux de connexion entre 18% et 100%. Cela montre que tous les Etats ayant communiqués cette donnée possèdent une bonne connexion internet à au moins 20%.

Code R :

```
ggplot(data = districts, aes(y = county_connections_ratio, fill =
county_connections_ratio)) + geom_bar(stat = "count")
```

COVID-19 et apprentissage à distance

La pandémie de COVID-19 a perturbé l'apprentissage de plus de 90% des élèves qui ont été touchés par la fermeture de leurs écoles, ce qui a mené de nombreux établissements scolaires à s'adapter et à trouver des moyens alternatifs pour mener à bien leurs processus. De nombreuses écoles ont eu recours à l'apprentissage en ligne pour faire face aux restrictions imposées en raison de la pandémie.

Engagement en fonction du temps en 2020

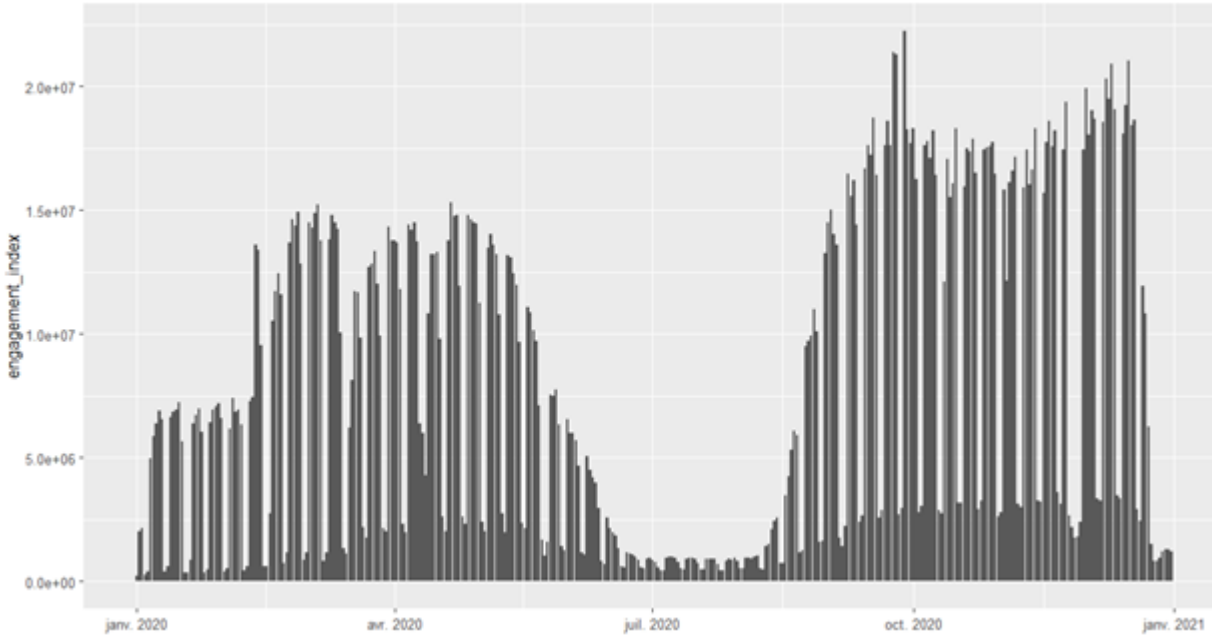


Figure 10 - Engagement des étudiants en 2020

Ce graphique représente l'engagement des étudiant, par le nombre total d'événements de chargement de page d'une solution éducative, sur la période du 1er janvier au 31 décembre 2020.

On remarque que :

- L'épidémie a provoqué une fermeture temporaire des écoles à partir de mi-février avant l'annonce par l'OMS officielle du 11 mars 2020 qui définit le COVID-19 comme pandémie. Et c'est précisément à partir de cette période que le nombre de pages consultées augmente, ce qui signifie que l'apprentissage numérique est utilisé plus fréquemment qu'auparavant.
- Le très faible nombre de pages consultées chaque week-end transparaît distinctement sur le graphique, ce qui dépeint la plage d'apprentissage privilégiée des étudiants est du lundi au vendredi.
- Tout comme les activités les week-ends, le nombre de pages consultées est également très faible pendant les vacances d'été (fin juin à début septembre), ce qui signifie également que les étudiants utilisent peu les produits en dehors des plages réservées à l'apprentissage.

Code R :

```
ggplot(data = tabl, aes(x = time, y = mean(engagement_index, na.rm = TRUE))) +
geom_bar(stat = "identity")
```

Engagement étudiant et technologies éducatives

Avant la pandémie, la technologie était un outil utilisé pour améliorer l'apprentissage dans n'importe quelle matière. Mais avec cette pandémie, elle est devenue le canal et la substance de l'éducation elle-même. Elle a changé la façon dont les enseignants et les élèves collectent, accèdent, analysent, présentent et transmettent les informations. La technologie a rendu l'apprentissage plus interactif et collaboratif, a aidé les étudiants à s'impliquer davantage dans le matériel qu'ils apprennent et avec lequel ils rencontrent des difficultés.

Catégories et secteurs des produits d'apprentissage

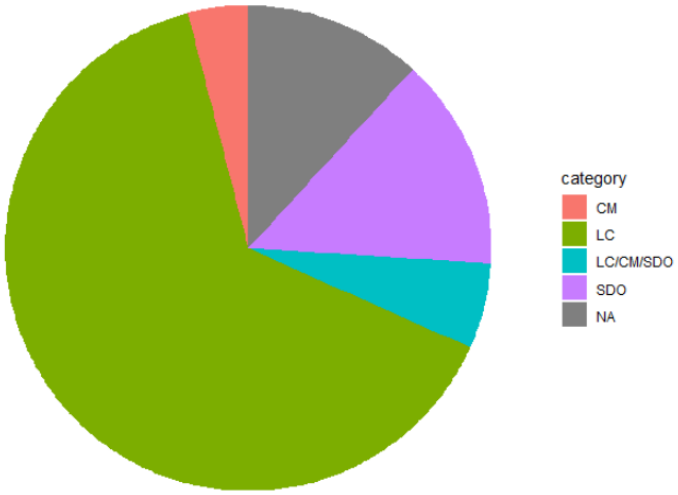


Figure 11 – Répartition des fonctionnalités primaires des 372 produits

LC = Apprentissage et programme d'études
CM = Gestion de classe
SDO = Opérations de l'école et du district

Ce graphique montre les fonctions primaires les plus présentes.

On observe que :

- LC est la fonction primaire la plus présente et représente les eux tiers du graphique
- Une quantité non négligeable de fonctions primaires sont dans la catégorie NA

Code R :

```
  #engagement etudiants et technologies éducatives
  new_category <- substr(products$`Primary Essential Function`, 1, 3)
  new_category[new_category == "LC/"] <- "LC/CM/SDO"
  products$category <- new_category
  ggplot(data = products, aes(x = "", y = category, fill = category)) +
geom_bar(stat = "identity") + coord_polar("y") +theme_void()
```

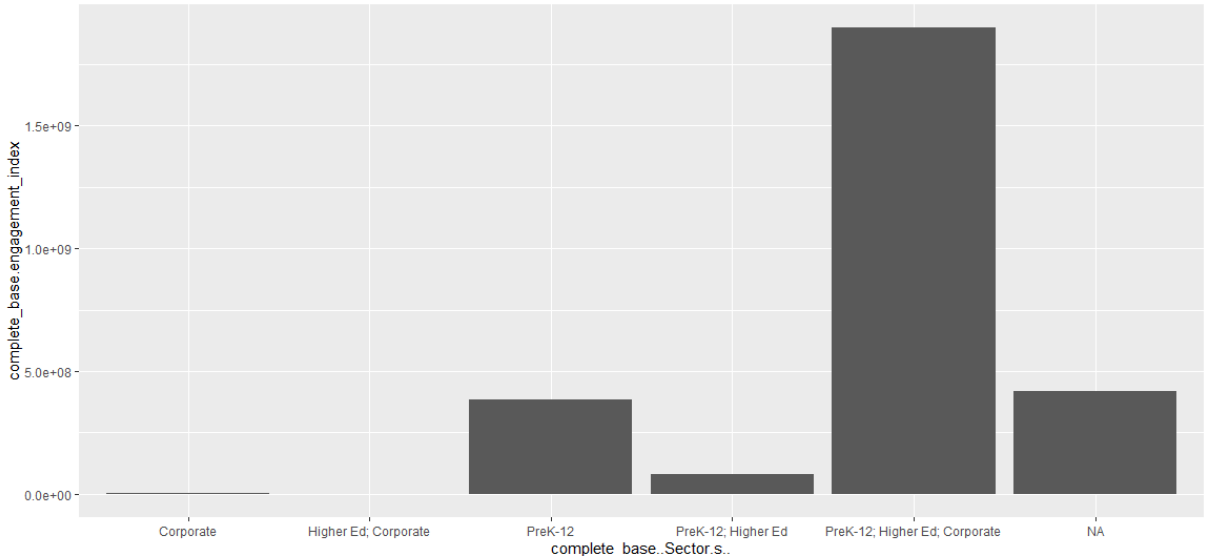


Figure 12 - Pages chargées par secteurs

Ce graphique représente le nombre des pages chargées par secteurs.

On observe que :

- Les secteurs de l'éducation sont divisés en trois catégories, PreK-12, Higher Education et Corporate.
- L'éducation préscolaire (PreK-12) est le secteur ou le produit est plus utilisé, ce qui signifie que l'apprentissage en ligne est plus souvent utilisé pour Prek-12

Code R :

```
graph <- data.frame(complete_base$Sector(s), complete_base$engagement_index)
ggplot(data=graph, aes(x=complete_base..Sector.s..,
y=complete_base.engagement_index)) + geom_bar(stat = "identity")
```

Produits prédominants

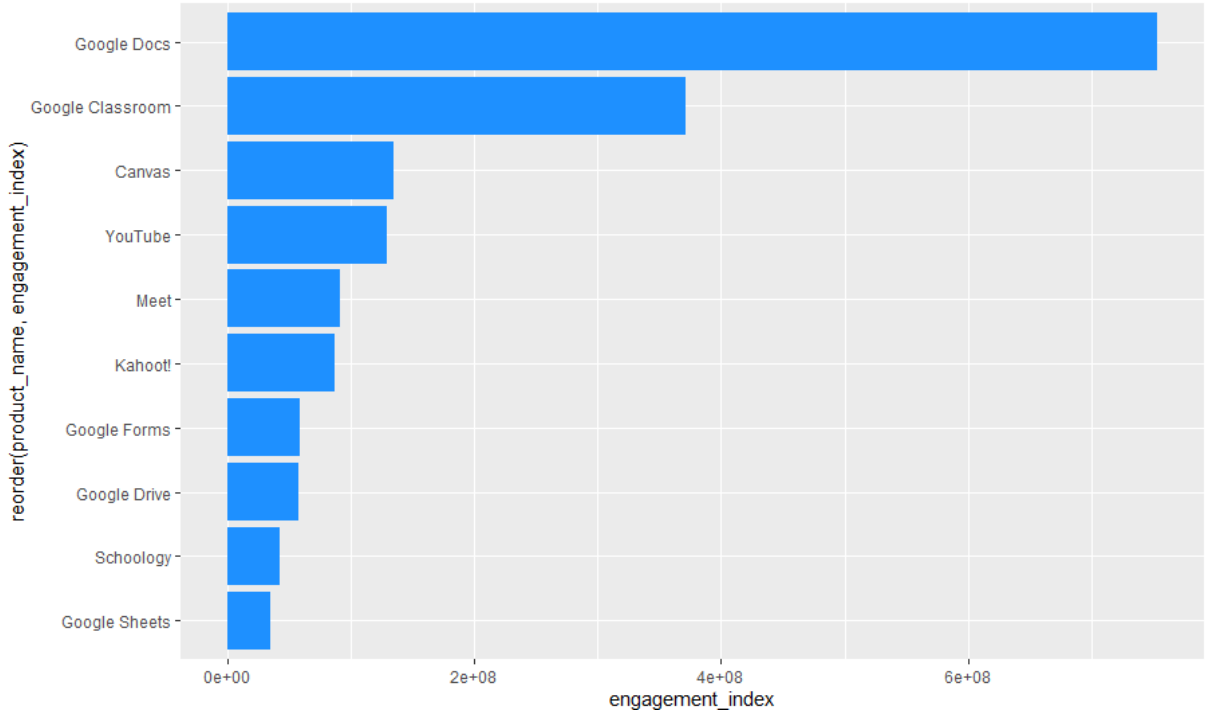


Figure 13 - Produits les plus utilisés

Ce graphique montre les 10 produits les plus utilisés pour l'apprentissage en ligne.

On observe que

- Google docs est le produit le plus utilisé avec plus de 700 millions de pages chargées.
- Le groupe Google LC est omniprésent dans le top 10 des produits utilisés (Google Docs, Google Classroom, Youtube, Meet, Google Forms, Google Drive, Google Sheet)

Les valeurs manquantes n'ont pas été traitées pour plus de pertinence.

Code R :

```
graph <- data.frame(complete_base$`Product Name`,
complete_base$engagement_index)
df <-
aggregate(graph$complete_base.engagement_index~graph$complete_base..Product.Name., FUN=sum, na.rm=TRUE)
names(df)[1]<-"product_name"
names(df)[2]<-"engagement_index"
df %>%
  arrange(desc(engagement_index)) %>%
  slice(1:10) %>%
  ggplot(., aes(x=engagement_index, y=reorder(product_name, engagement_index)))+
  geom_bar(stat='identity', fill="#1E90FF")
```

Engagement étudiant et facteurs sociogéographiques

Selon une étude menée par MYMOVE, des inégalités importantes entre les pays, les régions, les milieux urbains et ruraux augmentent avec la pandémie de COVID19. Par exemple, seuls 5 % des enfants et des jeunes d'Afrique de l'Ouest et du Centre ont accès à Internet à domicile.

Les zones rurales semblaient générer moins ou parfois même pas de trafic sur les plateformes d'apprentissage numérique dans certains pays comme (Arizona, Utah) ; cela peut s'expliquer par le paysage désertique de ces états qui déterminent moins de population dans les zones rurales et qui ont une couverture plus faible d'Internet, qui est nécessaire pour soutenir et faciliter les activités d'apprentissage en ligne.

En conséquent, dans les états où les zones sont couvertes par une bonne connectivité, l'engagement de l'étudiant augmente et produit des résultats académiques favorables.

Répartition géographique des engagements étudiants

Nous allons étudier la possible corrélation entre les données géographiques et le nombre de pages chargées, et par conséquent des informations sur les activités des élèves dans l'apprentissage numérique. Nos données nous informent que la source de données comporte 233 districts dans 23 États et 4 localités (si nous excluons les données non renseignées).

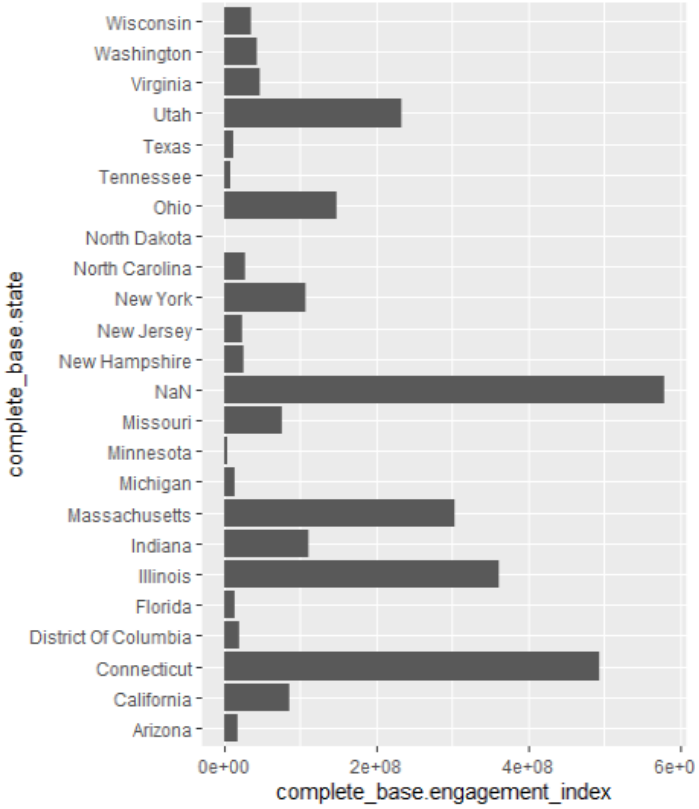


Figure 14 - Engagement des étudiants par Etats

Ce graphique représente l'engagement des étudiant, par le nombre total d'événements de chargement de page d'une solution éducative, selon les différents Etats disponibles.

On remarque que :

- La catégorie comportant le plus grand nombre de pages chargées, avec près de 600 millions, est celle représentant les données non connues. Il est donc important de le prendre en considération
- Il y a 3 États qui ont plus de 300 millions de pages chargées en 2020 : le Connecticut, l'Illinois et le Massachusetts. Sur 2,84 milliards d'enregistrements, ces 3 États représentent environ 40% (environ 1,1 milliard) du nombre total des pages chargées en 2020, si nous incluons les données non définies, la proportion s'élève à plus de 60% (environ 1 ,7 milliards).

Code R :

```
graph <- data.frame(complete_base$state, complete_base$engagement_index)
ggplot(data=graph, aes(x=complete_base.engagement_index,
y=complete_base.state)) + geom_bar(stat = "identity")
```

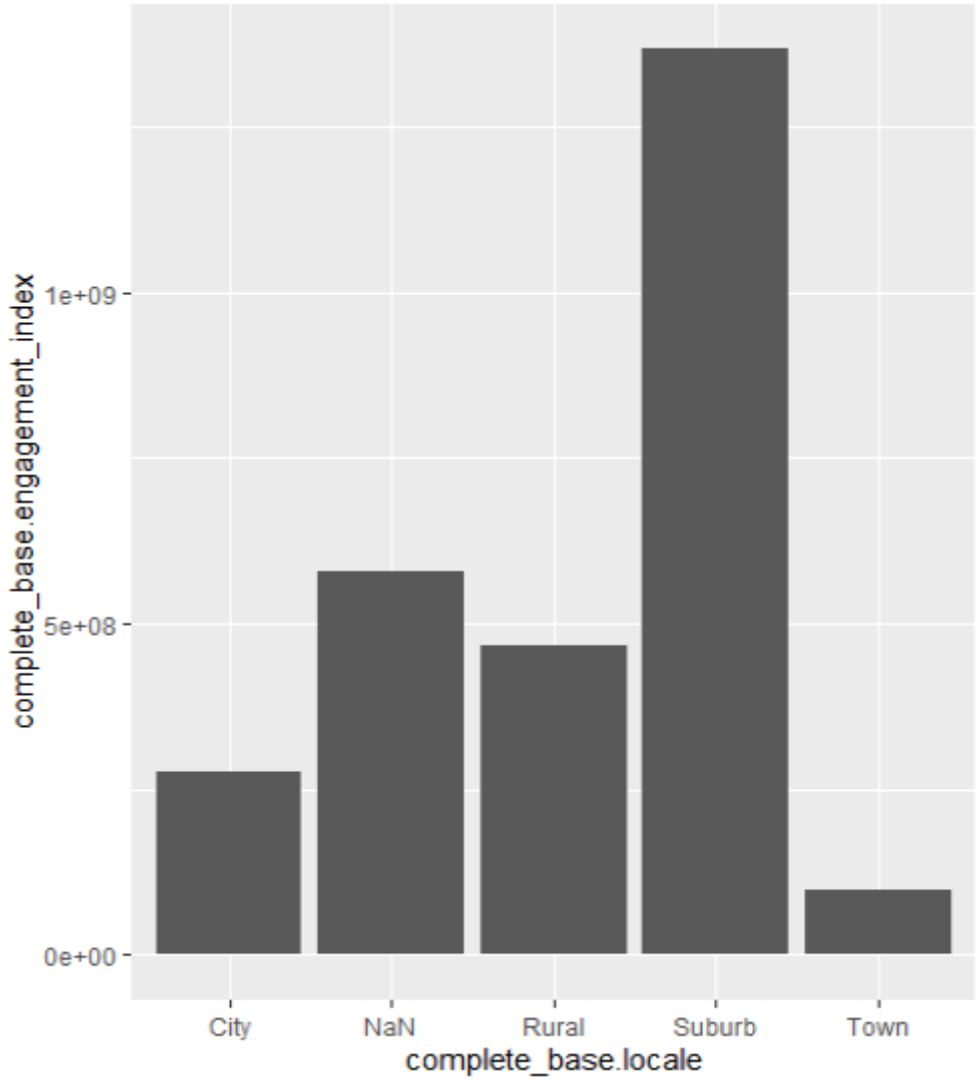


Figure 15 - Engagement des étudiants selon leurs localités

Ce graphique représente l'engagement des étudiant, par le nombre total d'événements de chargement de page d'une solution éducative, selon les différentes localités disponibles.

On observe que :

- Sur 2,84 milliards de pages téléchargées en 2020, la plupart provenant de la banlieue, qui représente près de la moitié du total.
- La catégorie représentant les données inconnues est la deuxième valeur la plus élevée dans le chargement des pages, soit près de 600 millions, avec une proportion de 21%.
- Les grandes villes (City) et les villes (Town) sont les plus faibles avec environ 300 millions et 100 millions de pages chargées.

Code R

```
graph <- data.frame(complete_base$locale, complete_base$engagement_index)
ggplot(data=graph, aes(x=complete_base.locale,
y=complete_base.engagement_index)) + geom_bar(stat = "identity")
```

Engagements des communautés noire et hispanique

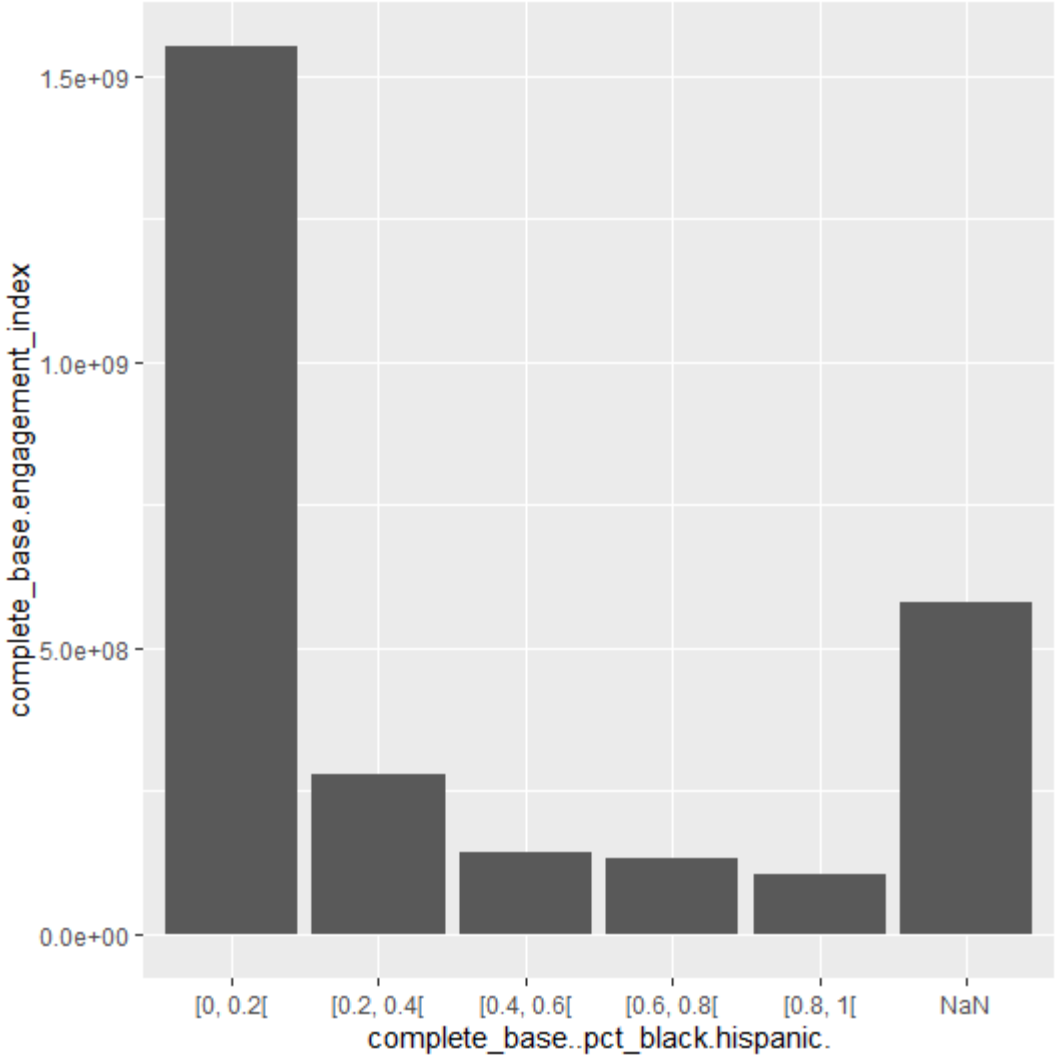


Figure 16 - Pages chargées par les communautés

Ce graphique représente, le nombre de pages chargées celons la proportion de personnes appartenant aux communautés black/hispanic.

On observe que :

- La densité de population entre 0 et 20% représente la tranche la plus significative des données par rapport au nombre de pages chargée.
- On note également que le nombre de valeurs non communiquées ou inconnue reste une partie des données importante.

Code R :

```
graph <- data.frame(complete_base$`pct_black/hispanic`,
complete_base$engagement_index)
ggplot(data=graph, aes(x=complete_base..pct_black.hispanic.,
y=complete_base.engagement_index)) + geom_bar(stat = "identity")
```

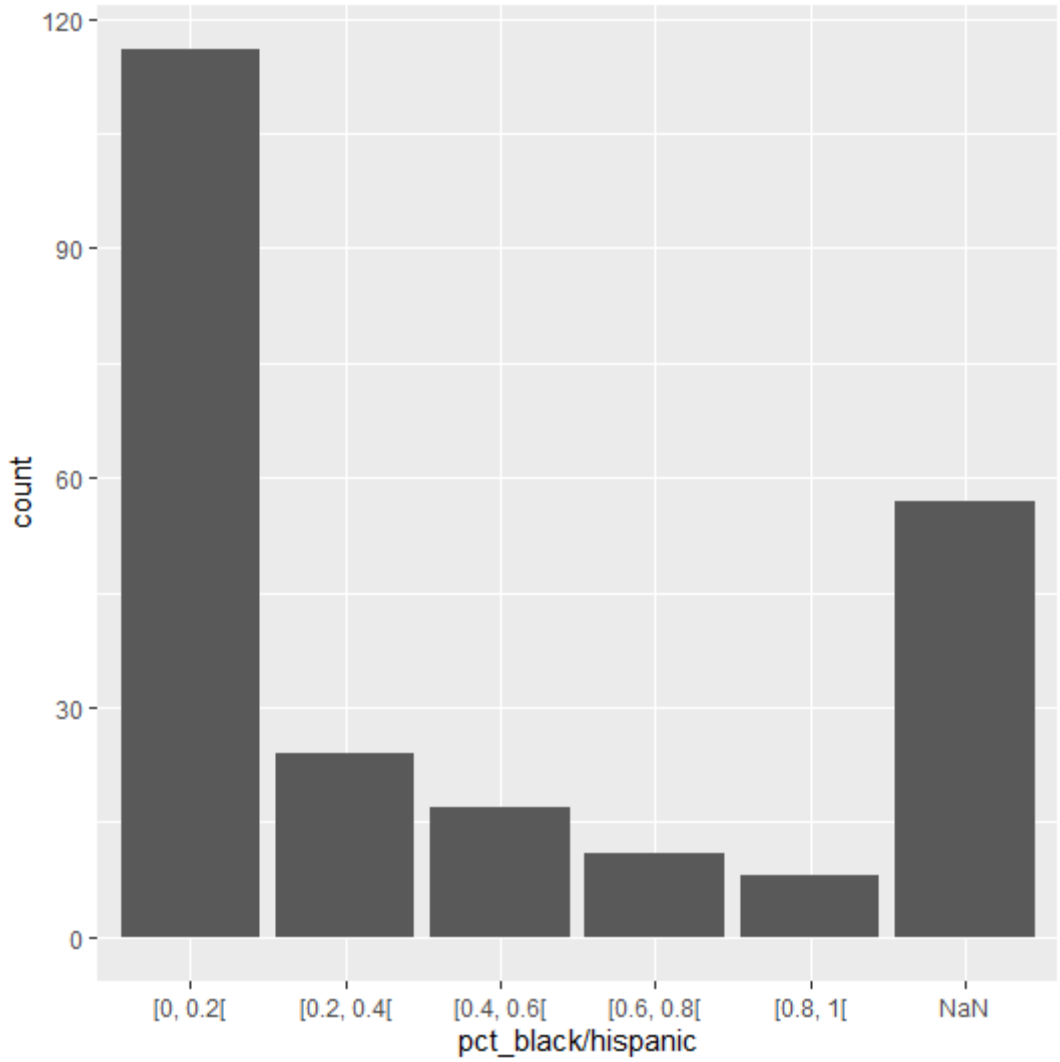


Figure 17 - Communautés par districts scolaires

Ce graphique représente le le nombre de district scolaires en fonction de la proportion des communautés black/hispanic dans leur district.

On observe que :

- Le nombre de district dans la densité de population de 0 à 20% est encore une fois le pourcentage le plus élevé.
- Il existe une corrélation claire entre les 2 graphiques précédents, entre le nombre de pages chargées par les communautés et les communautés par districts (cf Figure 16 et 17).

Code R :

```
ggplot(data=districts, aes(x=`pct_black/hispanic`)) + geom_bar(stat = "count")
```

Engagement étudiant et politiques d'Etat

L'égalité d'accès à une éducation efficace est depuis longtemps un problème dans les établissements scolaires. Le matériel d'apprentissage de qualité n'est pas nécessairement disponible pour tous les élèves ; ceux qui vivent dans des zones à faible revenu et ayant accès à des écoles mal financées sont souvent les plus touchés. Même avant la pandémie de COVID-19, de nombreux étudiants n'avaient pas un accès adéquat à la technologie nécessaire à leurs études.

Les facteurs financiers font toujours des différences cruciales dans tous les aspects d'une société. L'éducation n'est pas une exception. Ces différences sont encore plus marquées entre les pays riches et les pays pauvres. Seulement 6 % des enfants et des jeunes dans les pays à faible revenu ont accès à Internet, contre 87 % dans les pays à revenu élevé.

Ce facteur affecte négativement l'engagement des étudiants qui ont un accès inégal aux différentes ressources technologiques nécessaires à l'apprentissage à distance.

Dépenses totales locales et fédérales par élève

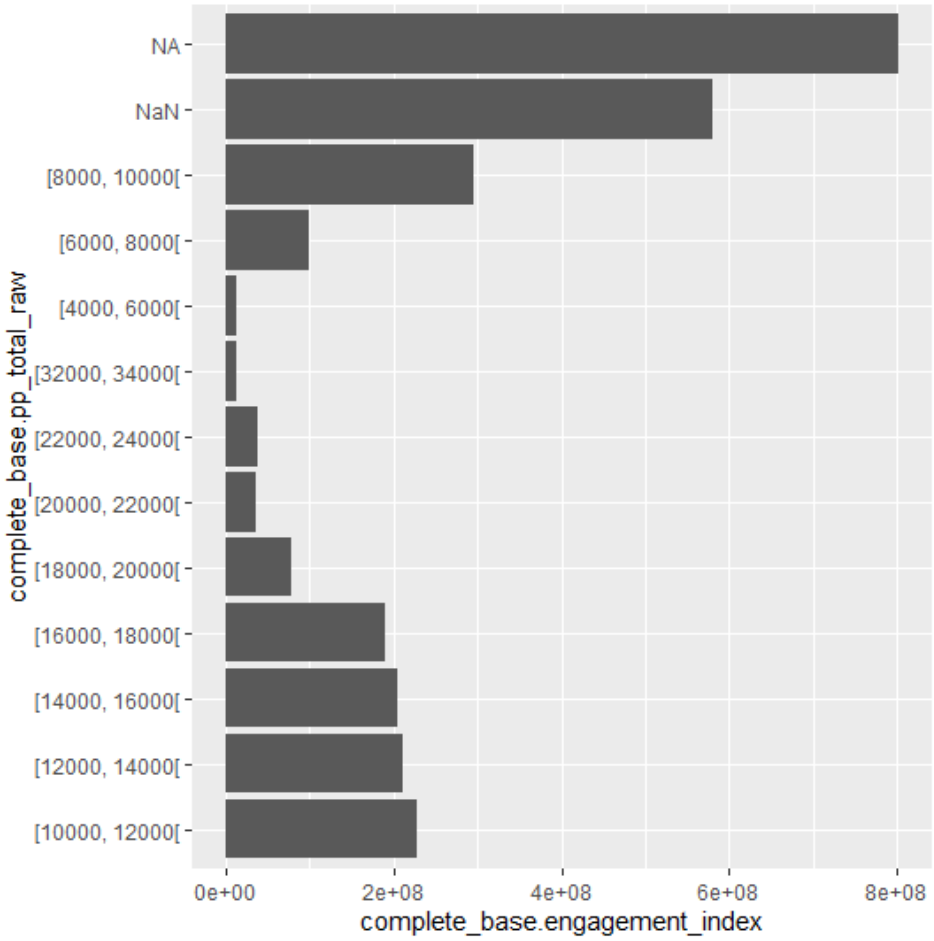


Figure 18 - Nombre de pages chargées et budget

Ce graphique montre le nombre de pages chargées en fonction du montant des budgets alloués aux écoles provenant du projet National Education Resource Database on Schools (NERD$) du Edunomics Lab.

On observe que :

- Près de la moitié des informations sur les dépenses totales par élève sont marquées comme inconnues.
- À partir des données restantes, nous constatons que la plupart des pages chargées proviennent de districts scolaires dont les dépenses sont comprises entre 8 000 et 10 000.
- Nous pouvons remarquer également que les districts scolaires dont les dépenses sont comprises entre 10 000 et 16 000 ont plus de 200 millions de pages chargées en 2020.
- Les deux résultats les plus faibles sont celles des districts dont les dépenses sont comprises entre 4 000 et 6 000 et entre 32 000 et 34 000, soient les dépenses les plus faibles et les plus élevées de l'ensemble de données. Cela signifie que le montant des dépenses n'est pas forcément lié à l'engagement des étudiants

Code R :

```
graph <- data.frame(complete_base$pp_total_raw,
complete_base$engagement_index)
ggplot(data=graph, aes(x=complete_base.engagement_index,
y=complete_base.pp_total_raw)) + geom_bar(stat = "identity")
```

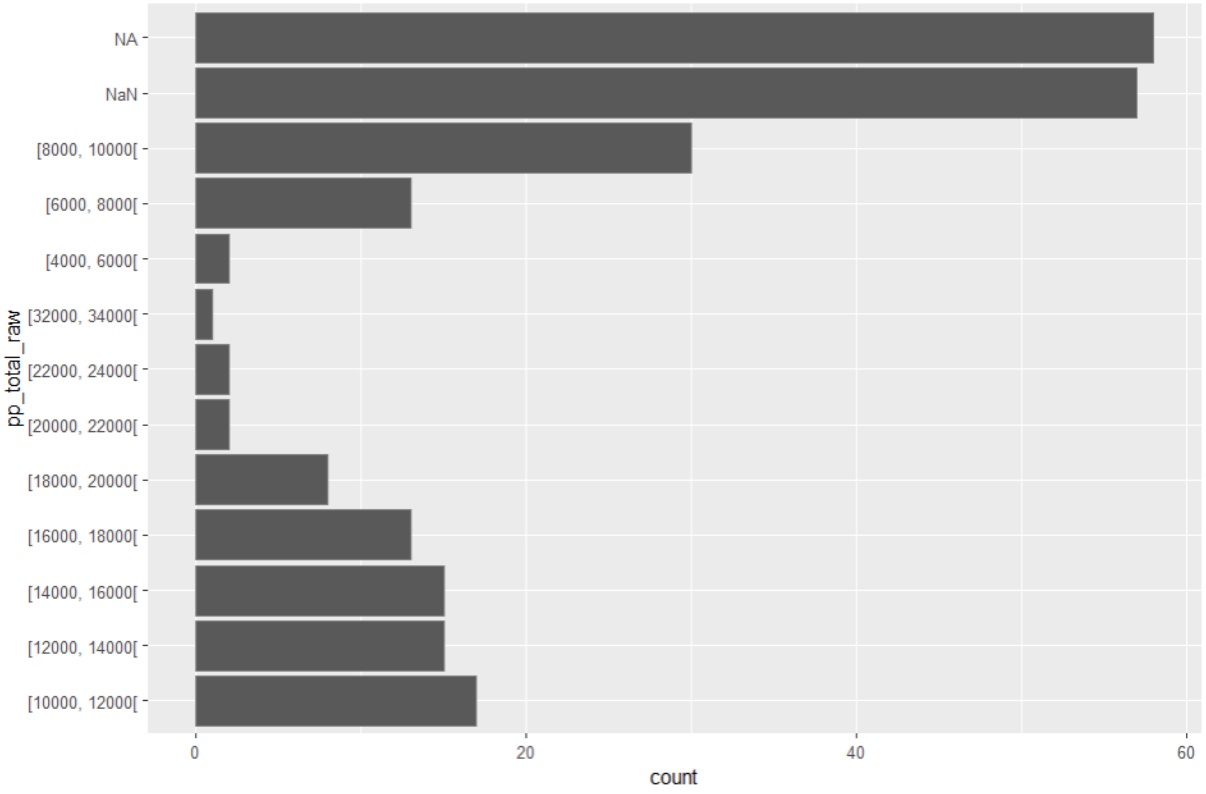


Figure 19 - Budgets alloués par rapport au nombre de district

Ce graphique montre le nombre de districts scolaires par rapport au montant des budgets alloués aux écoles provenant du projet National Education Resource Database on Schools (NERD$) du Edunomics Lab.

On observe que :

- Ce graphique avec le nombre de districts scolaires dans une catégorie de budget semble être en accord avec le graphe concernant le nombre de pages chargées pour un même budget ; et préciser les observations qui en découlent.
- Une fois encore, la catégorie ayant le plus grand nombre de districts scolaires est celle regroupant les valeurs non renseignées.
- Il y a 30 districts scolaires dont les dépenses totales sont comprises entre 8 000 et 10 000, ce qui explique également le nombre de pages chargés plus élevé dans le graphique précédent.

Code R :

```
ggplot(data = districts, aes(y = pp_total_raw)) + geom_bar(stat = "count")
```

Engagement des élèves de foyers modestes

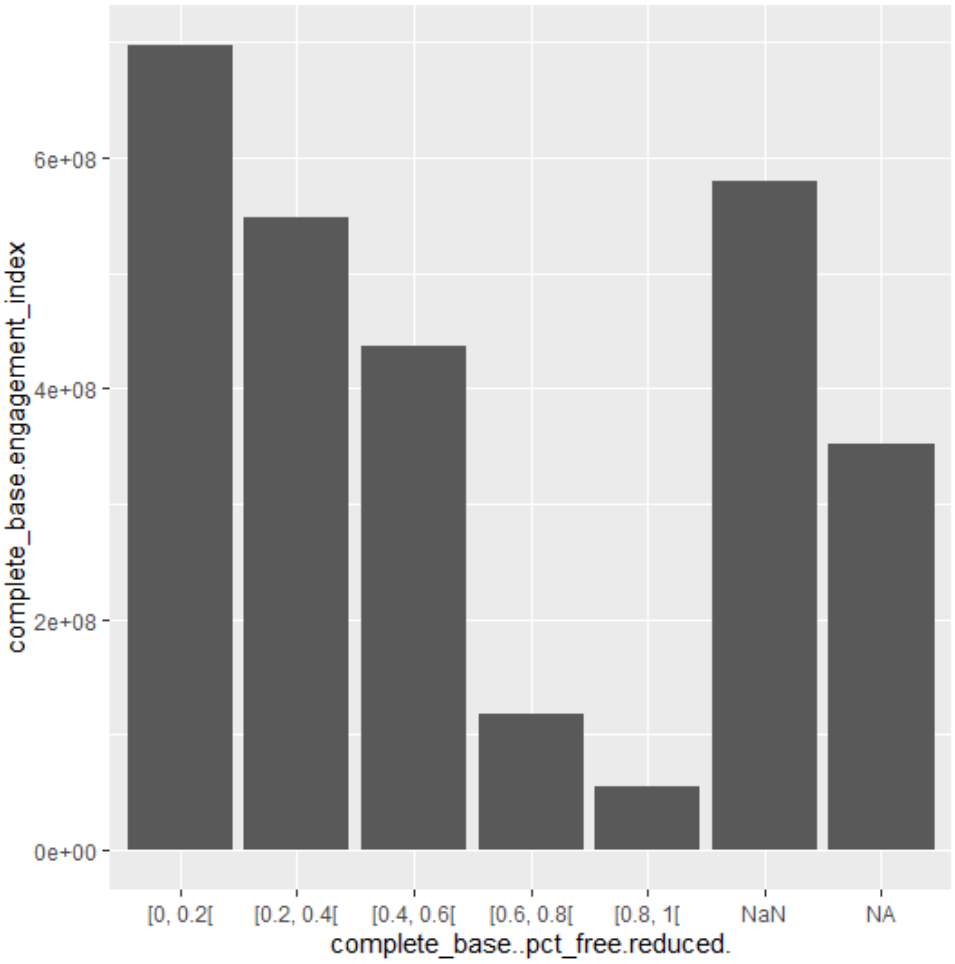


Figure 20 - Pages chargées par les personnes bénéficiaires d'aides

Ce graphique représente, le nombre de pages chargées celons la proportion de personnes bénéficiant d'une aide.

On constate que :

- De nombreuses données sont inconnues
- Entre 0 et 20% des personnes bénéficiant d'aides ont chargées plus de 600 millions de pages

Code R :

```
graph <- data.frame(complete_base$`pct_free/reduced`,
complete_base$engagement_index)
ggplot(data=graph, aes(x=complete_base..pct_free.reduced.,
y=complete_base.engagement_index)) + geom_bar(stat = "identity")
```

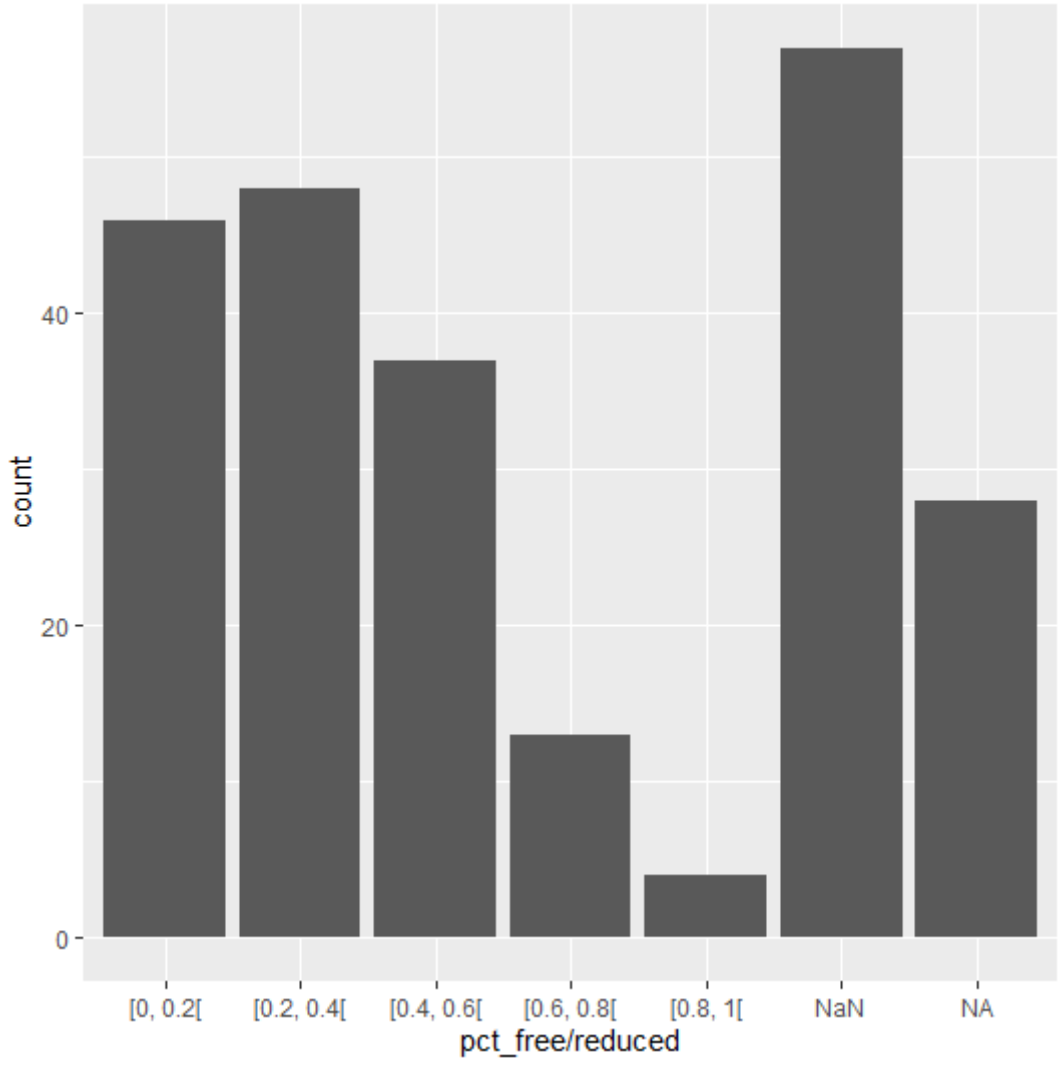


Figure 21 - Nombre de personnes bénéficiant d'aide dans les districs

Ce graphique représente le nombre de district scolaires en fonction de la proportion des personnes bénéficiant d'aides par districts.

On observe que :

- La majorité des données sont manquantes
- Les districts composés de 20 à 40 % de personnes avec des aides représente la valeur la plus élevée
- Contrairement au nombre de pages chargées (cf Figure 19) les districts possédants entre 0 et 20 % de bénéficiaires d'aide ne sont pas les majoritaires.

Code R :

```
ggplot(data=districts, aes(x=`pct_free/reduced`)) + geom_bar(stat = "count")
```

Conclusion

L'apprentissage à distance a le potentiel d'être la méthode d'apprentissage la plus flexible et la plus axée pour les étudiants. Cependant, pour que ce potentiel soit atteint, les problèmes d'accessibilité et la méthodologie d'enseignement adaptée doivent être résolus pour mieux répondre aux besoins éducatifs des élèves.

L'éducation hors ligne a ses propres limites qui ne peuvent être surmontées et qui sont battues avec succès par l'éducation en ligne :

- Coût de l'éducation (principalement pour l'enseignement supérieur) ;
- Des règles de discipline strictes que les jeunes élèves et leurs parents peuvent ne pas approuver ;
- Programmes éducatifs définis par l'État.

L'analyse de cet ensemble de données montre qu'aux États-Unis les étudiants ont un accès suffisant aux ressources, indépendamment des mesures de financement scolaire ou de l'accès au haut débit.

Les données suggèrent également que ni le statut socio-économique ni la démographie n'affectent la capacité d'un élève à s'engager dans l'apprentissage numérique. Les banlieues ont le plus grand nombre d'étudiants accédant à ces produits et on peut voir une corrélation claire entre l'accès aux produits et l'indice d'engagement. Il existe également une forte corrélation entre l'éducation noire/hispanique et l'éducation gratuite/tarif réduit, ce qui signifie que le gouvernement fait de son mieux pour aider la communauté pauvre en offrant une éducation gratuite.

Limites et recommandations

Durant cette analyse, nous avons pu être bloqués par certaines limites. Tout d'abord de nombreuses données sont manquantes au sein des données étudiées et ne peuvent pas être utilisées dans tous les calculs. Ensuite d'autres données manquent, comme des données concernant l'ensemble des Etats.

Il serait également intéressant d'avoir des données sur la réussite des élèves ayant expérimenté l'apprentissage à distance.

Les écoles et les gouvernements doivent prendre des mesures pour s'assurer que chaque élève a accès à Internet et aux appareils nécessaires à son éducation. En outre, si les écoles n'ont pas les moyens d'assumer le coût des supports d'apprentissage, elles doivent clarifier ce problème afin que les gouvernements, et les autres organisations axées sur l'éducation puissent trouver la meilleure façon de rendre ces supports accessibles afin de garantir la réussite des élèves.