DATA607 Project 2

Warner Alexis

2023-10-11

National Science Foundation

This data set contain research information census from Doctorate recipients by historical major fields. This data requires some data manipulation to transform it from wide to long. What are the fields that have the highest numbers of doctorate recipients in 2022. We only selected all the major fields and excluded any field that summarize all the fields. This is how the data look when enter our data platform.

Table 1-3 ◄◄ 1-3 ►►

Research doctorate recipients, by historical major field of doctorate: Selected years, 1992–2022

(Number and percent)

	1992		1997		2002		2007		2012		2017		2022	
Field of doctorate	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percent	Number	Percer
All fields	38,886	100.0	42,539	100.0	40,031	100.0	48,132	100.0	50,943	100.0	54,552	100.0	57,596	100.
Life sciences	7,172	18.4	8,421	19.8	8,478	21.2	10,702	22.2	11,964	23.5	12,554	23.0	13,211	22.
Agricultural sciences and natural resources	1,261	3.2	1,212	2.8	1,129	2.8	1,321	2.7	1,255	2.5	1,493	2.7	1,434	2.
Biological and biomedical sciences	4,799	12.3	5,788	13.6	5,695	14.2	7,238	15.0	8,322	16.3	8,566	15.7	9,218	16.
Health sciences	1,112	2.9	1,421	3.3	1,654	4.1	2,143	4.5	2,387	4.7	2,495	4.6	2,559	4.
Physical sciences and earth sciences	4,517	11.6	4,550	10.7	3,875	9.7	4,956	10.3	5,419	10.6	6,082	11.1	6,649	11.
Chemistry	2,213	5.7	2,148	5.0	1,922	4.8	2,318	4.8	2,416	4.7	2,699	4.9	3,060	5.
Geosciences, atmospheric sciences, and ocean sciences	767	2.0	803	1.9	689	1.7	875	1.8	941	1.8	1,169	2.1	1,181	2.
Physics and astronomy	1,537	4.0	1,599	3.8	1,264	3.2	1,763	3.7	2,062	4.0	2,214	4.1	2,408	4.
Mathematics and computer sciences	1,927	5.0	2,032	4.8	1,729	4.3	3,042	6.3	3,496	6.9	3,842	7.0	4,854	8.
Computer and information sciences	869	2.2	909	2.1	809	2.0	1,654	3.4	1,793	3.5	1,998	3.7	2,606	4.
Mathematics and statistics	1,058	2.7	1,123	2.6	920	2.3	1,388	2.9	1,703	3.3	1,844	3.4	2,248	3.
Psychology and social sciences	6,562	16.9	7,369	17.3	6,925	17.3	7,309	15.2	8,498	16.7	9,034	16.6	9,235	16.
Psychology	3,262	8.4	3,557	8.4	3,207	8.0	3,276	6.8	3,599	7.1	3,925	7.2	3,990	6.
Anthropology	320	0.8	434	1.0	496	1.2	512	1.1	547	1.1	446	0.8	415	0.
Economics	910	2.3	1,030	2.4	908	2.3	1,004	2.1	1,243	2.4	1,239	2.3	1,287	2.
Political science														

screenshot of the file preview

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readxl)
library(hrbrthemes)
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.
##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and
##       if Arial Narrow is not on your system, please see https://bit.ly/arialnarrow
```

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(gt)

raw <- read_csv("https://raw.githubusercontent.com/joewarner89/CUNY-607/main/Project%202/nsf24300-tab001-0033.csv",skip = 3)
```

```
## New names:
## Rows: 45 Columns: 15
## ── Column specification ──────────────────────────────────────────────────────── Delimiter: "," chr
## (15): Field of doctorate, 1992, ...3, 1997, ...5, 2002, ...7, 2007, ...9...
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## • `` -> `...3`
## • `` -> `...5`
## • `` -> `...7`
## • `` -> `...9`
## • `` -> `...11`
## • `` -> `...13`
## • `` -> `...15`
```

```
# read the data set


raw <- data.frame(raw)
head(raw)
```

```
##                          Field.of.doctorate  X1992    ...3  X1997    ...5
## 1                                      <NA> Number Percent Number Percent
## 2                                All fields 38,886   100.0 42,539   100.0
## 3                             Life sciences  7,172    18.4  8,421    19.8
## 4 Agricultural sciences and natural resources  1,261     3.2  1,212     2.8
## 5          Biological and biomedical sciences  4,799    12.3  5,788    13.6
## 6                           Health sciences  1,112     2.9  1,421     3.3
##    X2002    ...7  X2007    ...9  X2012   ...11  X2017   ...13  X2022   ...15
## 1 Number Percent Number Percent Number Percent Number Percent Number Percent
## 2 40,031   100.0 48,132   100.0 50,943   100.0 54,552   100.0 57,596   100.0
## 3  8,478    21.2 10,702    22.2 11,964    23.5 12,554    23.0 13,211    22.9
## 4  1,129     2.8  1,321     2.7  1,255     2.5  1,493     2.7  1,434     2.5
## 5  5,695    14.2  7,238    15.0  8,322    16.3  8,566    15.7  9,218    16.0
## 6  1,654     4.1  2,143     4.5  2,387     4.7  2,495     4.6  2,559     4.4
```

Data Manipulation

This data set comes with numerical columns which R Studio would have difficulty to tidy so We need to rename the variables. R studio put a X in front of all the number-like columns and dots and numbers for empty Column names. For Examples: If a column name starts with 1992, R would represent that column like X1992.

```
# deleted unnecessary rows
raw <- raw[-c(1),]
# Rename the variable so there are no numerical variable in the raw data
raw <- raw %>% rename(Doctorate_Field = 1,
                y1992 = 2,
                y1992_ = 3,
                y1997 = 4,
                y1997_ = 5,
                y2002 = 6,
                y2002_ = 7,
                y2007 = 8,
                y2007_ = 9,
                y2012 = 10,
                y2012_ = 11,
                y2017 = 12,
                y2017_ = 13,
                y2022 = 14,
                y2022_ = 15)



# Transforming the data set
head(raw)
```

```
##                              Doctorate_Field  y1992 y1992_  y1997 y1997_
## 2                                 All fields 38,886  100.0 42,539  100.0
## 3                              Life sciences  7,172   18.4  8,421   19.8
## 4 Agricultural sciences and natural resources  1,261    3.2  1,212    2.8
## 5          Biological and biomedical sciences  4,799   12.3  5,788   13.6
## 6                              Health sciences  1,112    2.9  1,421    3.3
## 7        Physical sciences and earth sciences  4,517   11.6  4,550   10.7
##    y2002 y2002_  y2007 y2007_  y2012 y2012_  y2017 y2017_  y2022 y2022_
## 2 40,031  100.0 48,132  100.0 50,943  100.0 54,552  100.0 57,596  100.0
## 3  8,478   21.2 10,702   22.2 11,964   23.5 12,554   23.0 13,211   22.9
## 4  1,129    2.8  1,321    2.7  1,255    2.5  1,493    2.7  1,434    2.5
## 5  5,695   14.2  7,238   15.0  8,322   16.3  8,566   15.7  9,218   16.0
## 6  1,654    4.1  2,143    4.5  2,387    4.7  2,495    4.6  2,559    4.4
## 7  3,875    9.7  4,956   10.3  5,419   10.6  6,082   11.1  6,649   11.5
```

Tidy package is useful for grouping certain rows and turn them as column. This package simplifies the process of transforming a data set from making them wide or long or spreading the rows into columns. It is essential to creating year, numbers and percent.

```
# Creating pivot pramaters var1 and var2
var1 <- c("y1992","y1997","y2002","y2007","y2012","y2017","y2022")
var2 <- c("y1992_","y1997_","y2002_","y2007_","y2012_","y2017_","y2022_")

# Creating 2 data sets to fully transform the problem
# fist data set transform data for year that match all the numbers

rawdata <- pivot_longer(
  data = raw,
  cols = all_of(var1),
  names_to = "year",
  values_to = "numbers"

)

# 2nd data set transform data for year that match all the percentage
raw_data1 <- pivot_longer(
  data = rawdata,
  cols = all_of(var2),
  names_to = "year_",
  values_to = "percent"
)
raw_data1$year <- str_extract(raw_data1$year,"\\d+")
raw_data1$year_ <- str_extract(raw_data1$year_,"\\d+")

# creating data with the correct alignment
data <- raw_data1 %>% filter(year == year_) %>% select(Doctorate_Field,year,numbers,percent)
data$numbers <- as.numeric(data$numbers)
```

```
## Warning: NAs introduced by coercion
```

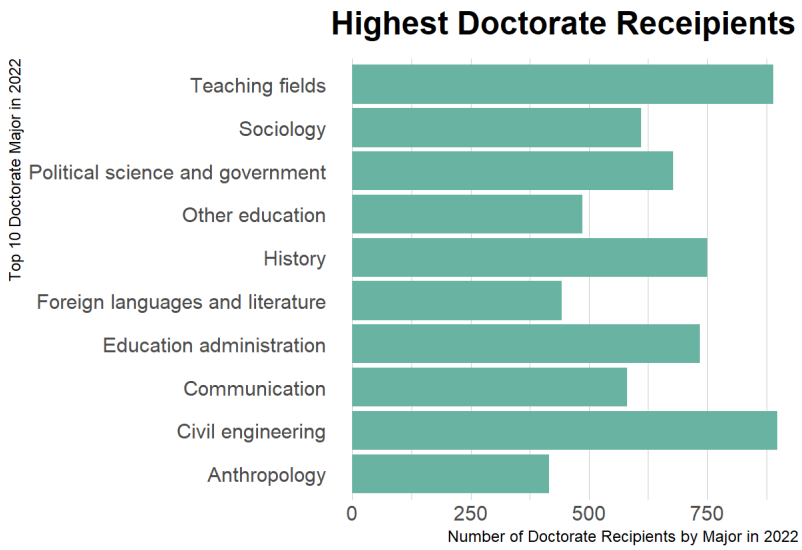
```
data$percent <- as.double(data$percent)

#Subset the data to get all major in 2022

final <- data %>% filter(Doctorate_Field!="All fields",
                year == '2022' ) %>% arrange(desc(numbers) )

head(final)
```

```
## # A tibble: 6 × 4
##   Doctorate_Field                year  numbers percent
##   <chr>                          <chr>   <dbl>   <dbl>
## 1 Civil engineering              2022      898     1.6
## 2 Teaching fields                2022      890     1.5
## 3 History                        2022      750     1.3
## 4 Education administration       2022      734     1.3
## 5 Political science and government 2022      678     1.2
## 6 Sociology                      2022      611     1.1
```

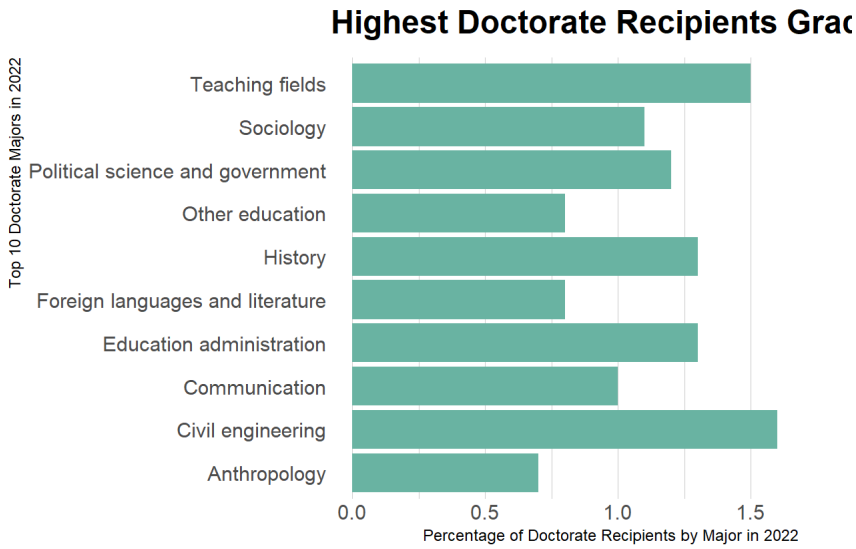
Data Analysis

We want to determine what Doctorate Program have the Highest numbers of recipients in 2022. this questions can help us understand in the near future, what fields that is going to have a higher employment rate.

```
# Number of PhD Graduates
final %>% head(10) %>% filter(Doctorate_Field!="Other") %>%
  ggplot( aes(x=Doctorate_Field, y=numbers) ) +
  geom_bar(stat="identity", fill="#69b3a2") +
  coord_flip() +
  theme_ipsum() +
  theme(
    panel.grid.minor.y = element_blank(),
    panel.grid.major.y = element_blank(),
    legend.position="none"
  ) +
  xlab("Top 10 Doctorate Major in 2022") +   ggtitle("Highest Doctorate Receipients")+
  ylab("Number of Doctorate Recipients by Major in 2022")
```



```
# Increase percentage of Graduate compare to previous year in 2022

final %>% head(10) %>% filter(Doctorate_Field!="Other") %>%
  ggplot( aes(x=Doctorate_Field, y=percent) ) +
  geom_bar(stat="identity", fill="#69b3a2") +
  coord_flip() +
  theme_ipsum() +
  theme(
    panel.grid.minor.y = element_blank(),
    panel.grid.major.y = element_blank(),
    legend.position="none"
  ) +
  xlab("Top 10 Doctorate Majors in 2022") +   ggtitle("Highest Doctorate Recipients Graduate prior previous year")+
  ylab("Percentage of Doctorate Recipients by Major in 2022")
```



What we learn form this data from the National Science Foundation is that Teaching fields major and Civil Engineering have the highest number of graduates in 2022 and seem that the numbers of graduates increase every year. This suggests that these majors would be very important in near future. United of America lays out its plan for national infrastructure. According to New York Times, Biden Details $2 Trillion Plan to Rebuild Infrastructure and Reshape the Economy. The president plan is to fix 20000 miles of roads and 10000 bridges. Civil engineering and Teaching Fields will be essential in the upcoming infrastructure project.

References https://ncses.nsf.gov/pubs/nsf24300/table/1-3 (https://ncses.nsf.gov/pubs/nsf24300/table/1-3)

Domestic Tourism

According to United Nation World Tourism (UNWTO), Tourism is a social, cultural and economic phenomenon which entails the movement of people to countries or places outside their usual environment for personal or business/professional purposes. These people are called visitors (which may be either tourists or excursionists; residents or non-residents) and tourism has to do with their activities, some of which involve tourism expenditure. This data sets has information about how many visitors travel per country.

This is how the data look:

	A	B	C	D	E	F	G	H	I	J	K	L
1	#NAME?	<?xml version="1.0" encoding="utf-16"?><WebTableParameter xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns										
2	Country	Variable		20	20	20	20	20	20	20	20	20
3	Australia	Total domestic trips		210,753,700	215,845,100	225,239,200	233,126,300	248,377,400	240,118,400	260,362,000	269,481,500	280,324,300
4		Overnight visitors (tourists)		72,008,700	67,669,600	69,296,800	71,894,700	74,472,200	75,796,400	84,480,800	87,523,200	90,741,700
5		Same-day visitors (excursionists)		138,745,000	148,175,500	155,942,400	161,231,600	173,905,200	164,322,000	175,881,200	181,958,300	189,582,600
6		Nights in all types of accommodation		277,865,400	262,235,600	265,393,200	270,573,300	281,732,900	282,679,700	310,532,600	317,535,400	334,798,200
7		Hotels and similar establishments		75,646,100	70,740,800	71,929,600	74,367,000	71,529,900	71,977,900	79,604,700	81,120,800	83,092,300
8		Other collective establishments		68,603,000	69,302,700	70,154,400	69,891,400	72,239,000	72,161,500	70,625,100	83,114,200	90,291,300
9		Private accommodation		133,616,300	122,192,000	123,309,300	126,314,900	137,964,100	138,540,300	160,302,700	153,300,400	161,414,600
10	Belgium	Overnight visitors (tourists)		..	..	..	..	..	..	..	4,201,243	3,845,110
11		Same-day visitors (excursionists)		..	..	..	..	..	..	..	13,986,366	..
12		Nights in all types of accommodation		..	..	..	..	..	..	..	14,545,105	15,459,919
13		Hotels and similar establishments		..	..	..	..	..	..	..	8,336,997	10,169,388
14		Other collective establishments		..	..	..	..	..	..	..	1,427,693	1,104,872
15		Private accommodation		..	..	..	..	..	..	..	4,780,416	4,185,659
16	Region of Brussels-Capital	Overnight visitors (tourists)		..	..	..	..	..	..	..	..	..
17		Nights in all types of accommodation		..	..	..	..	..	..	..	..	..
18	Flanders	Overnight visitors (tourists)		..	..	..	..	..	..	..	..	..
19		Nights in all types of accommodation		..	..	..	..	..	..	..	..	..
20	Wallonia	Overnight visitors (tourists)		..	..	..	..	..	..	..	..	..
21		Nights in all types of accommodation		..	..	..	..	..	..	..	..	..
22	Canada	Total domestic trips		..	..	..	..	..	..	..	..	..
23		Overnight visitors (tourists)		87,647,000	91,386,000	93,143,000	105,743,000	108,383,000	108,825,000	108,647,000	108,805,000	113,053,000

screenshot of the file preview

```
# load the data from Github
tourist <- read.csv("https://raw.githubusercontent.com/joewarner89/CUNY-607/main/Project%202/OECD%20-%20Tourism%20data.csv",
stringsAsFactors = F, skip = 1)

# Delete Empty row
tourist$X <- NULL
# Delete unwanted rows at the bottom of the files
tourist <- tourist[-c(208:213),]
head(tourist)
```

```
##     Country                             Variable        X2008        X2009
## 1 Australia                 Total domestic trips 210,753,700  215,845,100
## 2                  Overnight visitors (tourists)  72,008,700   67,669,600
## 3              Same-day visitors (excursionists) 138,745,000  148,175,500
## 4           Nights in all types of accommodation 277,865,400  262,235,600
## 5              Hotels and similar establishments  75,646,100   70,740,800
## 6                Other collective establishments  68,603,000   69,302,700
##          X2010        X2011        X2012        X2013        X2014        X2015
## 1 225,239,200  233,126,300  248,377,400  240,118,400  260,362,000  269,481,500
## 2  69,296,800   71,894,700   74,472,200   75,796,400   84,480,800   87,523,200
## 3 155,942,400  161,231,600  173,905,200  164,322,000  175,881,200  181,958,300
## 4 265,393,200  270,573,300  281,732,900  282,679,700  310,532,600  317,535,400
## 5  71,929,600   74,367,000   71,529,900   71,977,900   79,604,700   81,120,800
## 6  70,154,400   69,891,400   72,239,000   72,161,500   70,625,100   83,114,200
##          X2016        X2017        X2018        X2019        X2020        X2021
## 1 280,324,300  291,796,700  310,166,200  365,796,900  236,705,700  242,543,000
## 2  90,741,700   98,483,900  104,821,900  117,447,700   72,513,800   82,074,100
## 3 189,582,600  193,312,800  205,344,300  248,349,200  164,191,900  160,468,900
## 4 334,798,200  352,085,100  371,527,500  417,906,500  275,403,600  321,108,900
## 5  83,092,300   87,519,400   95,297,500  101,046,800   51,588,600   67,162,100
## 6  90,291,300   94,045,100   94,416,100  111,904,800   79,557,400   90,006,300
```

Data Manipulation

This data set has a lot of empty spaces. We are going to fill out with the country column with the respective country every observation belong to then creating two data set to create the year and visitors column. both newly created columns can be constructed using `gather()` or `pivot_long()` depending on how you select your parameters

```
# fill the values in country so the right country aligned with his population visitors

tourist <- tourist  %>% mutate(Country = as.character(na_if(Country,""))) %>% fill(Country,.direction = 'down')

var_t <- c("X2008","X2009","X2010","X2011","X2012","X2013"
                  ,"X2014","X2015","X2016","X2017","X2018",
                  "X2019","X2020","X2021")

# creating new variables.

prep <- pivot_longer(
  data = tourist,
  cols = all_of(var_t),
  names_to = "year",
  values_to = "visitors"

)

# replace the missing value with o
prep <- prep %>%
  mutate(Country = str_replace(Country, "T<U+FFFD>rkiye","Turkey"))
prep$visitors[prep$visitors == '..'] <- 0
prep$visitors <- as.numeric(gsub(",", "", prep$visitors))

# re-transform the dataset so
prep2 <- spread(prep,
       key = "Variable",
       value = "visitors")
```

```
## Warning: The `x` argument of `as_tibble.matrix()` must have unique column names if
## `.name_repair` is omitted as of tibble 2.0.0.
## ℹ Using compatibility `.name_repair`.
## ℹ The deprecated feature was likely used in the tidyr package.
##   Please report the issue at <https://github.com/tidyverse/tidyr/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
prep2$year <- str_extract(prep2$year,"\\d+")
prep2$V1 <- NULL

### Rename the variables
tourist_final <- prep2 %>% rename(Hotel_Establishments = 3,
                                  Collective_Establishments = 4,
                                  Overnights_visitors = 5,
                                  Private_Accomodation = 6,
                                  SameDay_Accomodation = 7,
                                  Nights_Accomodations = 8,
                                  Total_Domestic_Trips = 9)



tourist_final[is.na(tourist_final)] = 0
tourist_final$Country[tourist_final$Country == 'T<fc>rkiye'] <- 'Turkey'

# look at tourism in 2021
tourist21 <- tourist_final %>% filter(year == 2021)
finaldata <- tourist21 %>% arrange(desc(Hotel_Establishments)) %>% filter(Hotel_Establishments != 0) %>%
  select(Country,year,Hotel_Establishments) %>% head(10)
head(finaldata)
```

```
## # A tibble: 6 × 3
##   Country          year  Hotel_Establishments
##   <chr>            <chr>                <dbl>
## 1 "United States"  2021            1157502833
## 2 "France"         2021             277538829
## 3 "  Saudi Arabia" 2021             204645844
## 4 "  Indonesia"    2021             185971368
## 5 "Germany"        2021             148135171
## 6 "Spain"          2021              67637768
```

Now We are going to look at countries that have a lot visitor booking Hotels.

```
tour_tbl <- gt(finaldata)

tour_tbl <-
  tour_tbl |>
  tab_header(
    title = md("**Top 10 Touristic Country in the World**"),
    subtitle = md("countries with Highest Hotel Accomodation by Tourists")
  )
tour_tbl
```

Top 10 Touristic Country in the World

countries with Highest Hotel Accomodation by Tourists

Country	year	Hotel_Establishments
United States	2021	1157502833
France	2021	277538829
Saudi Arabia	2021	204645844
Indonesia	2021	185971368
Germany	2021	148135171
Spain	2021	67637768
Australia	2021	67162100
Poland	2021	42968711

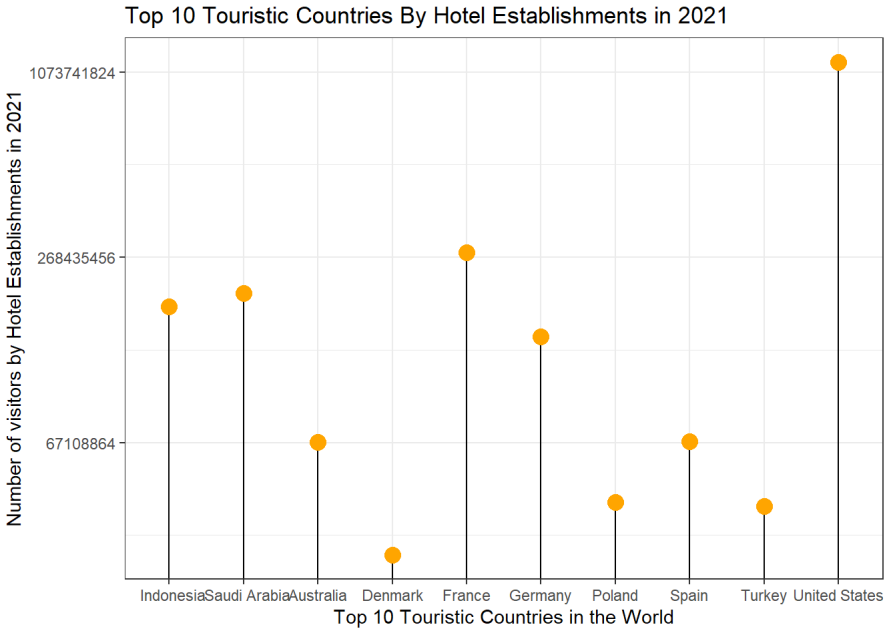
Top 10 Touristic Country in the World

countries with Highest Hotel Accomodation by Tourists

Country	year	Hotel_Establishments
Turkey	2021	41654000
Denmark	2021	28989789

```
finaldata %>% ggplot( aes(x=Country, y=Hotel_Establishments)) +
  geom_segment( aes(xend=Country, yend=0)) +
  geom_point( size=4, color="orange") + scale_y_continuous(trans='log2') +
  theme_bw() +
  xlab("Top 10 Touristic Countries in the World") +   ggtitle("Top 10 Touristic Countries By Hotel Establishments in 2021")+
  ylab("Number of visitors by Hotel Establishments in 2021")
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```



More people have traveled to USA than any other countries. Big Cities in the USA have a lot great establishments such as Hotels, Private Properties and so on. Tourism has a great effect on any country economy. The top 10 countries revealed in the data sets has the most decorative hotels and can the one of the best accommodations in the World.USA and France are considered one of the most popular place to visit.

School Performances

According to the city of Chicago, This data set shows all school level performance data used to create CPS School Report Cards for the 2011-2012 school year. Metrics are described as follows (also available for download at http://bit.ly/uhbzah (http://bit.ly/uhbzah)): NDA indicates “No Data Available.”

We are going to transform the data set so we can look into which high school perform well and how impacted is the teacher involvement in the school year.

First look of the data set

Table Preview

Scho...	Name...	Ele...	Stree...	Phon...	Link	Netw...	Adeq...	Track...	CPS ...	CPS ...	Healt...	Safet...
610040	Henry D L...	ES	2103 N L...	(773) 534...	http://sch...	Fullerton ...	No	Track_E	Probation	Level 3	No	NDA
609852	Eliza Cha...	ES	2135 W F...	(773) 534...	http://sch...	Ravensw...	Yes	Standard	Not on Pr...	Level 1	No	Strong
609818	Luther Bu...	ES	2035 N M...	(773) 534...	http://sch...	Fullerton ...	No	Standard	Not on Pr...	Level 2	No	Weak
610127	Mary Gag...	ES	5510 N C...	(773) 534...	http://sch...	O'Hare El...	No	Standard	Not on Pr...	Level 1	No	Strong
610095	Walter L ...	ES	700 W Wi...	(773) 534...	http://sch...	Fullerton ...	Yes	Standard	Not on Pr...	Level 1	No	Strong
609865	Jordan El...	ES	7414 N W...	(773) 534...	http://sch...	Ravensw...	No	Standard	Not on Pr...	Level 2	No	Strong
610196	George B ...	ES	5900 N W...	(773) 534...	http://sch...	Ravensw...	No	Standard	Not on Pr...	Level 2	No	Strong
609792	Newton B...	ES	4220 N Ri...	(773) 534...	http://sch...	O'Hare El...	No	Standard	Not on Pr...	Level 2	No	Average
610144	Frank W ...	ES	3650 W S...	(773) 534...	http://sch...	Fullerton ...	No	Standard	Probation	Level 2	No	Average
609875	Charles R...	ES	3116 W B...	(773) 534...	http://sch...	Fullerton ...	No	Standard	Not on Pr...	Level 2	No	Average
609901	Edgebroo...	ES	6525 N Hi...	(773) 534...	http://sch...	O'Hare El...	Yes	Standard	Not on Pr...	Level 1	No	Very Str
610145	Peter A R...	ES	3425 N M...	(773) 534...	http://sch...	O'Hare El...	No	Standard	Not on Pr...	Level 1	No	Strong
610101	William B...	ES	24 W Wal...	(773) 534...	http://sch...	Fullerton ...	No	Standard	Not on Pr...	Level 2	No	NDA

```
raw <- read.csv("https://raw.githubusercontent.com/joewarner89/CUNY-607/main/Project%202/schools_by_performance_level.csv",
stringsAsFactors = F,sep = ",")
head(raw)
```

```
##   School.ID                              Name.of.School
## 1    610539              Marvin Camras Elementary School
## 2    609852             Eliza Chappell Elementary School
## 3    609818             Luther Burbank Elementary School
## 4    609680 Walter Payton College Preparatory High School
## 5    610038            Abraham Lincoln Elementary School
## 6    609749     Northside College Preparatory High School
##   Elementary..Middle..or.High.School     Street.Address   Phone.Number
## 1                                ES  3000 N Mango Ave  (773) 534-2960
## 2                                ES 2135 W Foster Ave  (773) 534-2390
## 3                                ES 2035 N Mobile Ave  (773) 534-3000
## 4                                HS   1034 N Wells St  (773) 534-0034
## 5                                ES   615 W Kemper Pl  (773) 534-5720
## 6                                HS 5501 N Kedzie Ave  (773) 534-3954
##                                                                    Link
## 1 http://schoolreports.cps.edu/SchoolProgressReport_Eng/Spring2011Eng_610539.pdf
## 2 http://schoolreports.cps.edu/SchoolProgressReport_Eng/Spring2011Eng_609852.pdf
## 3 http://schoolreports.cps.edu/SchoolProgressReport_Eng/Spring2011Eng_609818.pdf
## 4 http://schoolreports.cps.edu/SchoolProgressReport_Eng/Spring2011Eng_609680.pdf
## 5 http://schoolreports.cps.edu/SchoolProgressReport_Eng/Spring2011Eng_610038.pdf
## 6 http://schoolreports.cps.edu/SchoolProgressReport_Eng/Spring2011Eng_609749.pdf
##                         Network.Manager Adequate.Yearly.Progress.Made.
## 1            Fullerton Elementary Network                           No
## 2     Ravenswood-Ridge Elementary Network                          Yes
## 3            Fullerton Elementary Network                           No
## 4 North-Northwest Side High School Network                          Yes
## 5            Fullerton Elementary Network                           No
## 6 North-Northwest Side High School Network                          Yes
##   Track.Schedule CPS.Performance.Policy.Status CPS.Performance.Policy.Level
## 1       Standard              Not on Probation              Not Enough Data
## 2       Standard              Not on Probation                      Level 1
## 3       Standard              Not on Probation                      Level 2
## 4       Standard              Not on Probation                      Level 1
## 5       Standard              Not on Probation                      Level 1
## 6       Standard              Not on Probation                      Level 1
##   Healthy.Schools.Certified. Safety.Icon Safety.Score Family.Involvement.Icon
## 1                         No     Average           54                 Average
## 2                         No      Strong           70                  Strong
## 3                         No        Weak           37                     NDA
## 4                         No Very Strong           98                     NDA
## 5                        Yes Very Strong           99             Very Strong
## 6                         No Very Strong           99                     NDA
##   Family.Involvement.Score Environment.Icon Environment.Score Instruction.Icon
## 1                       58             Weak                37          Average
## 2                       65          Average                53          Average
## 3                      NDA          Average                42             Weak
## 4                      NDA      Very Strong                80           Strong
## 5                       99           Strong                74           Strong
## 6                      NDA      Very Strong                99      Very Strong
##   Instruction.Score Leaders.Icon Leaders.Score Teachers.Icon Teachers.Score
## 1                41  Very Strong            83   Very Strong              88
## 2                51      Average            56       Average              48
## 3                34          NDA           NDA           NDA             NDA
## 4                77          NDA           NDA           NDA             NDA
## 5                66       Strong            65        Strong              70
## 6                88          NDA           NDA           NDA             NDA
##   Parent.Engagement.Icon Parent.Engagement.Score Parent.Environment.Icon
## 1                Average                      51                 Strong
## 2                Average                      50                Average
## 3                Average                      47                Average
## 4                    NDA                     NDA                    NDA
## 5                 Strong                      56                Average
## 6                 Strong                      57                 Strong
##   Parent.Environment.Score Average.Student.Attendance
## 1                       55                       95.1
## 2                       52                       95.1
## 3                       49                       95.2
## 4                      NDA                       93.4
## 5                       47                       96.0
## 6                       62                       95.7
##   Rate.of.Misconducts..per.100.students. Average.Teacher.Attendance
## 1                                    4.3                       0.0
## 2                                    2.9                      96.7
## 3                                    9.8                      95.5
## 4                                    0.7                      96.1
## 5                                    2.0                      96.4
## 6                                    2.8                      96.8
##   Individualized.Education.Program.Compliance.Rate Pk.2.Literacy.. Pk.2.Math..
```

```
## 1                                        97.7            47.5        39.9
## 2                                        98.9            72.6          57
## 3                                       100.0            60.8        45.2
## 4                                       100.0             NDA         NDA
## 5                                        95.8            80.1        43.3
## 6                                        98.4             NDA         NDA
##   Gr3.5.Grade.Level.Math.. Gr3.5.Grade.Level.Read.. Gr3.5.Keep.Pace.Read..
## 1                     30.8                     27.8                  47.8
## 2                       56                     54.4                  55.2
## 3                     42.8                     34.9                  61.2
## 4                      NDA                      NDA                   NDA
## 5                     89.6                     84.9                  60.7
## 6                      NDA                      NDA                   NDA
##   Gr3.5.Keep.Pace.Math.. Gr6.8.Grade.Level.Math.. Gr6.8.Grade.Level.Read..
## 1                   38.6                    41.1                    38.5
## 2                     69                    53.2                    63.9
## 3                   63.8                      41                    34.4
## 4                    NDA                     NDA                     NDA
## 5                   62.6                    81.9                    85.2
## 6                    NDA                     NDA                     NDA
##   Gr6.8.Keep.Pace.Math. Gr6.8.Keep.Pace.Read.. Gr.8.Explore.Math..
## 1                  49.2                  55.4               15.3
## 2                    60                  59.8               12.8
## 3                    55                  54.5               13.1
## 4                   NDA                   NDA                NDA
## 5                    52                  62.4               66.3
## 6                   NDA                   NDA                NDA
##   Gr.8.Explore.Read.. ISAT.Exceeding.Math.. ISAT.Exceeding.Reading..
## 1                30.5                  11.2                     12.0
## 2                  41                  19.6                     17.6
## 3                17.2                  20.6                     11.3
## 4                 NDA                    NA                       NA
## 5                77.9                  69.7                     64.4
## 6                 NDA                    NA                       NA
##   ISAT.Value.Add.Math ISAT.Value.Add.Read ISAT.Value.Add.Color.Math
## 1                -1.8                -0.3                      Red
## 2                 0.4                 1.0                   Yellow
## 3                -0.2                 0.2                   Yellow
## 4                  NA                  NA                      NDA
## 5                 0.2                 0.9                   Yellow
## 6                  NA                  NA                      NDA
##   ISAT.Value.Add.Color.Read Students.Taking..Algebra..
## 1                    Yellow                       NDA
## 2                     Green                      27.5
## 3                    Yellow                       NDA
## 4                       NDA                       NDA
## 5                     Green                      67.1
## 6                       NDA                       NDA
##   Students.Passing..Algebra.. X9th.Grade.EXPLORE..2009.
## 1                         NDA                      NDA
## 2                        63.6                      NDA
## 3                         NDA                      NDA
## 4                         NDA                     21.2
## 5                        54.5                      NDA
## 6                         NDA                     22.4
##   X9th.Grade.EXPLORE..2010. X10th.Grade.PLAN..2009. X10th.Grade.PLAN..2010.
## 1                       NDA                    NDA                     NDA
## 2                       NDA                    NDA                     NDA
## 3                       NDA                    NDA                     NDA
## 4                      21.8                   23.1                    23.2
## 5                       NDA                    NDA                     NDA
## 6                      22.2                   24.5                    24.7
##   Net.Change.EXPLORE.and.PLAN X11th.Grade.Average.ACT..2011.
## 1                         NDA                            NDA
## 2                         NDA                            NDA
## 3                         NDA                            NDA
## 4                           2                             27
## 5                         NDA                            NDA
## 6                         2.3                           28.8
##   Net.Change.PLAN.and.ACT College.Eligibility.. Graduation.Rate..
## 1                     NDA                  NDA               NDA
## 2                     NDA                  NDA               NDA
## 3                     NDA                  NDA               NDA
## 4                     3.9                 96.4              96.9
## 5                     NDA                  NDA               NDA
## 6                     4.3                   98              97.6
##   College.Enrollment.Rate.. College.Enrollment..number.of.students.
## 1                      NDA                                     826
## 2                      NDA                                     451
```

```
## 3                     NDA                                1139
## 4                    82.4                                 881
## 5                     NDA                                 813
## 6                    90.7                                1053
##   General.Services.Route Freshman.on.Track.Rate.. X_COORDINATE Y_COORDINATE
## 1                     30                     NDA      1137482      1919395
## 2                     31                     NDA      1161017      1934467
## 3                     29                     NDA      1134123      1913042
## 4                     33                    90.7      1174485      1907490
## 5                     33                     NDA      1171699      1915829
## 6                     31                    95.9      1154091      1936414
```

```
#subseting the data set
# We want only high school that their grade benchmark
raw <- raw %>% select(School.ID:Phone.Number,Teachers.Icon,Teachers.Score,
                      ISAT.Exceeding.Math..:X10th.Grade.PLAN..2010.) %>%
  filter(Elementary..Middle..or.High.School == "HS")
# Rename some variable names
raw2 <- raw %>%  select(1:19) %>%    rename(School_id = 1,
                      School = 2,
                      School_Category = 3,
                      Street_Address = 4,
                      Phone_Number = 5,
                      Teachers = 6,
                      Teachers_Score = 7,

                      Algebra = Students.Taking..Algebra..,
                      Algebra_Benchmark = Students.Taking..Algebra..,
                      Explore_9th_Grade_2009 = X9th.Grade.EXPLORE..2009.,
                      Explore_9th_Grade_2010 = X9th.Grade.EXPLORE..2010.,
                      Plan_10th_Grade_2009 = X10th.Grade.PLAN..2009.,
                      Plan_10th_Grade_2010 = X10th.Grade.PLAN..2010.)

# reselect important rows
raw2 <- raw2 %>% select (1:7,16:19)


# Replace String with Another Stirng
raw2$Teachers[raw2$Teachers == 'NDA'] <- 'NA'
raw2$Teachers_Score[raw2$Teachers_Score == 'NDA'] <- '0'
raw2$Explore_9th_Grade_2009[raw2$Explore_9th_Grade_2009 == 'NDA'] <- 0
raw2$Explore_9th_Grade_2010[raw2$Explore_9th_Grade_2010 == 'NDA'] <- 0
raw2$Plan_10th_Grade_2009[raw2$Plan_10th_Grade_2009 == 'NDA'] <- 0
raw2$Plan_10th_Grade_2010[raw2$Plan_10th_Grade_2010 == 'NDA'] <- 0
head(raw2)
```

```
##   School_id                                              School
## 1    609680            Walter Payton College Preparatory High School
## 2    609749                Northside College Preparatory High School
## 3    610394                             Uplift Community High School
## 4    609737 Friedrich W von Steuben Metropolitan Science High School
## 5    609695                               Roald Amundsen High School
## 6    609708                              Edwin G Foreman High School
##   School_Category       Street_Address   Phone_Number Teachers Teachers_Score
## 1              HS     1034 N Wells St  (773) 534-0034       NA              0
## 2              HS   5501 N Kedzie Ave  (773) 534-3954       NA              0
## 3              HS    900 W Wilson Ave  (773) 534-2875  Average             53
## 4              HS  5039 N Kimball Ave  (773) 534-5100     Weak             29
## 5              HS    5110 N Damen Ave  (773) 534-2320  Average             41
## 6              HS 3235 N LeClaire Ave  (773) 534-3400       NA              0
##   Explore_9th_Grade_2009 Explore_9th_Grade_2010 Plan_10th_Grade_2009
## 1                   21.2                   21.8                 23.1
## 2                   22.4                   22.2                 24.5
## 3                   13.6                   13.9                 14.9
## 4                   15.9                   15.7                 17.3
## 5                   13.8                   13.9                 14.9
## 6                     13                   12.7                 14.1
##   Plan_10th_Grade_2010
## 1                 23.2
## 2                 24.7
## 3                 14.7
## 4                 16.7
## 5                 14.8
## 6                 13.7
```

Data Manipulation

we are going transform the data set so we can create new variables for Performance Column, along with survey data fields.

```
# Transforming the data set to Long using the
# performance benchmark for all teachers

temp <- pivot_longer(
  data = raw2,
  cols = Teachers,
  names_to = "Performance",
  values_to = "Teaching_Performance"

)

## Spred the performace acroos different columns
temp_prep <- spread(temp, key = "Teaching_Performance",value = "Teachers_Score")
temp_prep$Not_Recorded <- temp_prep$`NA`
temp_prep$`NA` <- NULL

# Replace All na with 0
temp_prep$Average[temp_prep$Average == NA] <- "0"
temp_prep <-  temp_prep %>% mutate_at(c('Average','Strong','Very Strong','Weak','Not_Recorded',
                                'Explore_9th_Grade_2009','Explore_9th_Grade_2010','Plan_10th_Grade_2009',
                                'Plan_10th_Grade_2010'), as.numeric)

temp_prep$`Very Strong`[temp_prep$`Very Strong` == NA] <- 0
head(temp_prep)
```

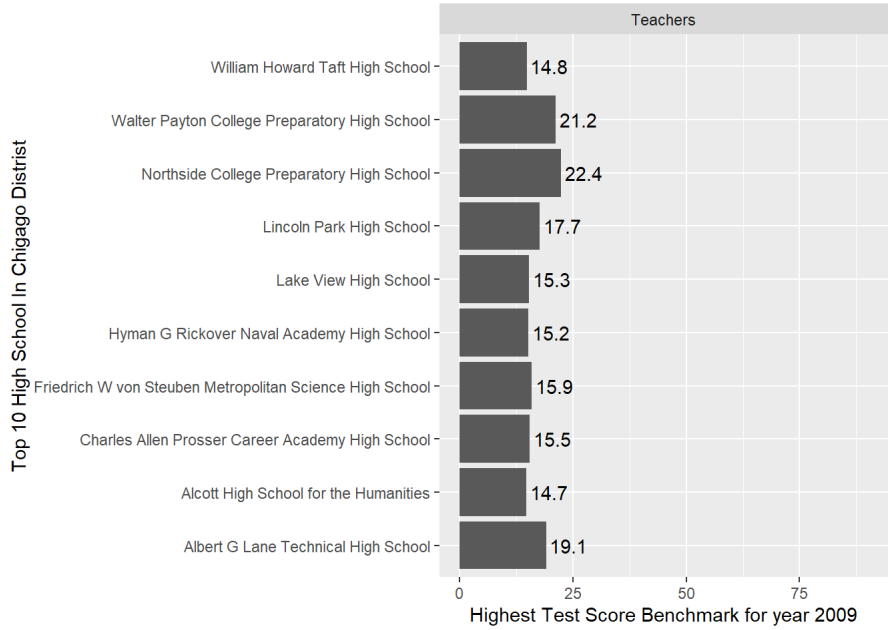
```
## # A tibble: 6 × 15
##   School_id School                   School_Category Street_Address Phone_Number
##       <int> <chr>                    <chr>           <chr>          <chr>
## 1    609680 Walter Payton College P… HS              "1034 N Wells… (773) 534-0…
## 2    609749 Northside College Prepa… HS              "5501 N Kedzi… (773) 534-3…
## 3    610394 Uplift Community High S… HS              "900 W Wilson… (773) 534-2…
## 4    609737 Friedrich W von Steuben… HS              "5039 N Kimba… (773) 534-5…
## 5    609695 Roald Amundsen High Sch… HS              "5110 N Damen… (773) 534-2…
## 6    609708 Edwin G Foreman High Sc… HS              "3235 N LeCla… (773) 534-3…
## # ℹ 10 more variables: Explore_9th_Grade_2009 <dbl>,
## #   Explore_9th_Grade_2010 <dbl>, Plan_10th_Grade_2009 <dbl>,
## #   Plan_10th_Grade_2010 <dbl>, Performance <chr>, Average <dbl>, Strong <dbl>,
## #   `Very Strong` <dbl>, Weak <dbl>, Not_Recorded <dbl>
```

```
# High School that have the highest average  of Student performing well in their test
school <- temp_prep %>% arrange(desc(Explore_9th_Grade_2009)) %>% select(2,3,Performance,Explore_9th_Grade_2009:Plan_10th_Gr
ade_2010,Average:Not_Recorded ) %>%
  head(10)
school <- school %>% replace(is.na(.), 0)
head(school)
```

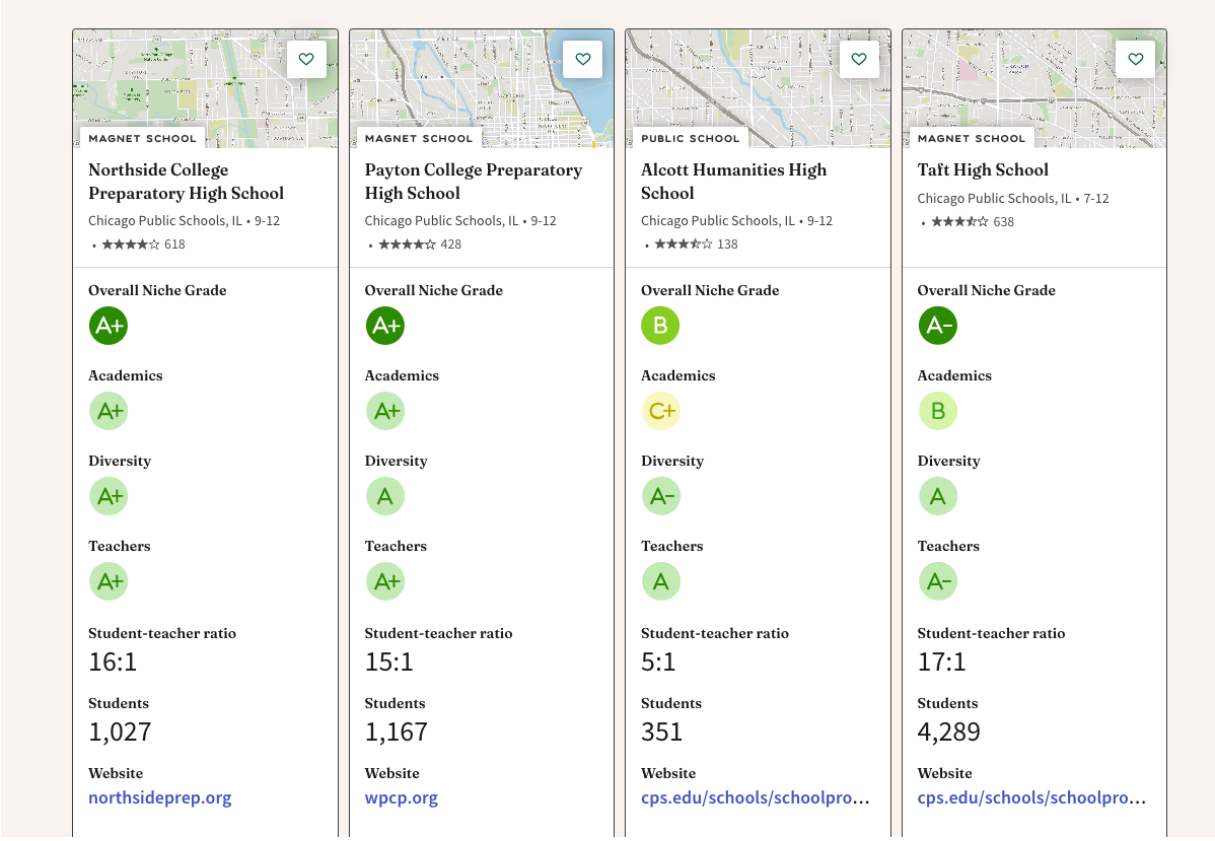
```
## # A tibble: 6 × 12
##   School                      School_Category Performance Explore_9th_Grade_2009
##   <chr>                       <chr>           <chr>                        <dbl>
## 1 Northside College Preparat… HS              Teachers                      22.4
## 2 Walter Payton College Prep… HS              Teachers                      21.2
## 3 Albert G Lane Technical Hi… HS              Teachers                      19.1
## 4 Lincoln Park High School    HS              Teachers                      17.7
## 5 Friedrich W von Steuben Me… HS              Teachers                      15.9
## 6 Charles Allen Prosser Care… HS              Teachers                      15.5
## # ℹ 8 more variables: Explore_9th_Grade_2010 <dbl>, Plan_10th_Grade_2009 <dbl>,
## #   Plan_10th_Grade_2010 <dbl>, Average <dbl>, Strong <dbl>,
## #   `Very Strong` <dbl>, Weak <dbl>, Not_Recorded <dbl>
```

Let look at the schools that performed well in 2009 because of their teacher performances. We are going to check the top school in Niche.com to have an idea why City of Chicago survey range this particular school as one to excel in taking state exam.

```
# the choosen graph
school %>%   ggplot(aes(x = School, y = Explore_9th_Grade_2009)) +
  geom_col() +
  facet_grid(~ Performance) +
  coord_flip() +
  ylim(c(0, 90)) +
  geom_text(aes(label = round(Explore_9th_Grade_2009, 1)), hjust = -.1) +
  labs(x = "Top 10 High School In Chigago District", y = "Highest Test Score Benchmark for year 2009")
```



According to Nice.com, Northside College Preparatory High School is a top rated, public, magnet school located in CHICAGO, IL. It has 1,027 students in grades 9-12 with a student-teacher ratio of 16 to 1. According to state test scores, 94% of students are at least proficient in math and 95% in reading. These Schools has performing since 2009 to present day and the involvements of teachers and parents have been one of the main reasons they do well academically. The picture below shows a comparison of some of the best High School revealed in the top 10 list.



References

Schools, C. P. (2018, September 14). NNW -all schools by Performance Level + AYP: City of chicago: Data Portal. Chicago Data Portal. https://data.cityofchicago.org/Education/NNW-all-schools-by-performance-level-AYP/wefw-pbqz (https://data.cityofchicago.org/Education/NNW-all-schools-by-performance-level-AYP/wefw-pbqz)

Senior, Freshman, & Junior. (2021a, December 31). Northside College Preparatory High School in Chicago, IL. Niche. https://www.niche.com/k12/northside-college-preparatory-high-school-chicago-il/ (https://www.niche.com/k12/northside-college-preparatory-high-school-chicago-il/)

Kelly Kang, J. F. (n.d.). Data tables. NSF. https://ncses.nsf.gov/pubs/nsf24300/table/1-3 (https://ncses.nsf.gov/pubs/nsf24300/table/1-3)

Tankersley, J. (2021, March 31). Biden details $2 trillion plan to rebuild infrastructure and reshape the economy. The New York Times. https://www.nytimes.com/2021/03/31/business/economy/biden-infrastructure-plan.html# (https://www.nytimes.com/2021/03/31/business/economy/biden-infrastructure-plan.html#):~:text=677-,Biden%20Details%20%242%20Trillion%20Plan%20to%20Rebuild%20Infrastructure%20and%20Reshape,inequities%20and%20raising%2(