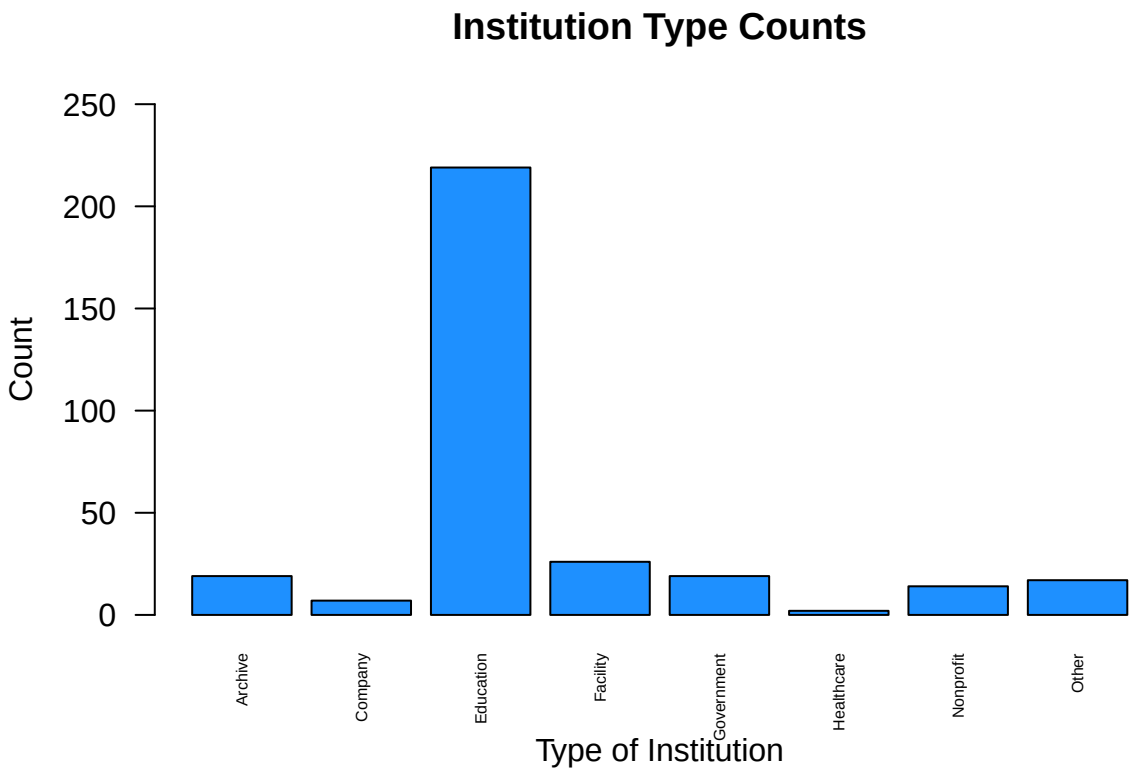
final_institution_doi_assessment_report

Read in the data

```
library(readr)
Institution_doi <- read_csv("final_institution_doi.csv")
```

Visualize the counts of institution types

```
freq = table(Institution_doi$`Institution Type`)
barplot(freq, las=2, cex.names=.5, col=c("dodgerblue"),
        xlab="Type of Institution", ylab = "Count", main="Institution Type Counts", ylim=c(0,250))
```



Number of NA Values

```
institution_type = Institution_doi$`Institution Type`
na_values = sum(is.na(institution_type))
na_values
```

```
## [1] 223
```

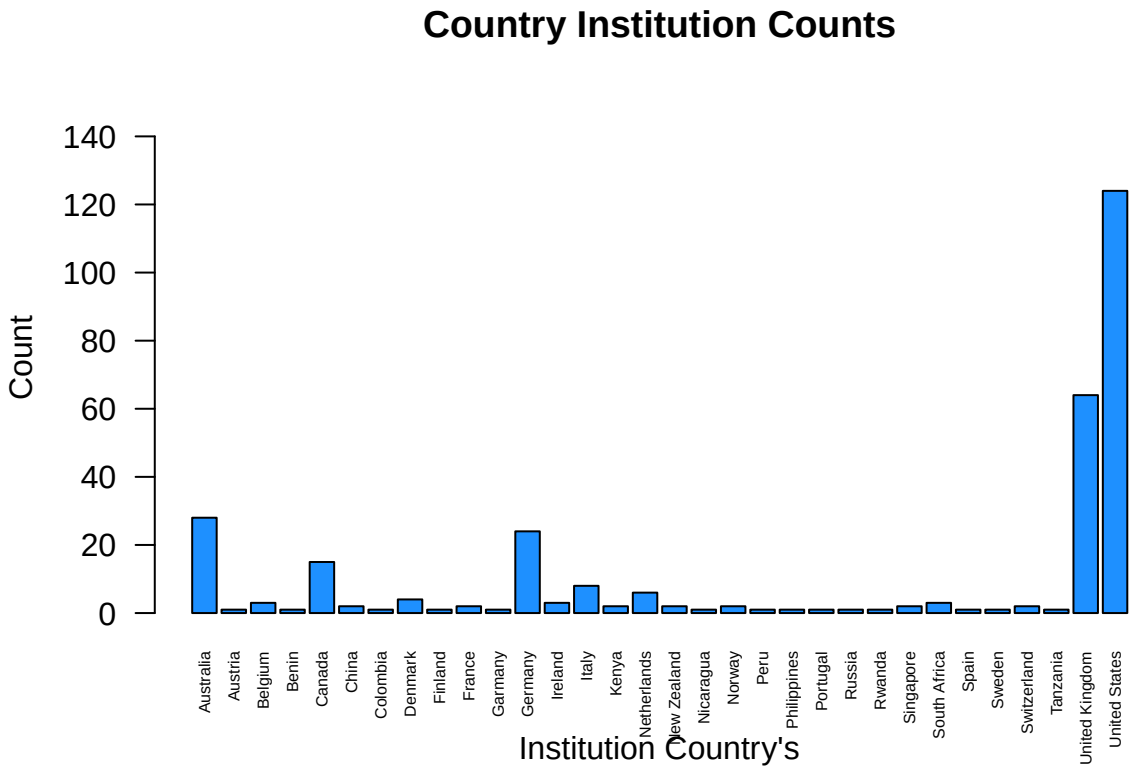
Number of Education Institution types in my dataset

```
education_count = Institution_doi$`Institution Type` == "Education"
sum(education_count, na.rm = TRUE)
```

```
## [1] 219
```

Visualize the counts by country

```
freq = table(Institution_doi$`Institution's Country`)
barplot(freq, las=2, cex.names=.5, col=c("dodgerblue"),
        xlab="Institution Country's", ylab = "Count", main="Country Institution Counts", ylim=c(0,150))
```



In this visualization, we can see that more education institutions share their research data than any other institution type. However, I should note that from the GRID database, I had less missing institution type data from education institutions than I did from these other types. For example, there are 223 institutions in my dataset that did not have an institution type, which amounts to 40% of my data. However, since 219 of the institution typs are education, at the very least Education institutions represent at least 50% of the my data, if not more, this suggests that education institutions are sharing their data more so than others. However, this does not get into the total quantity of data that each institution is sharing. With this information, we can then ask if being a public or private sector institution plays a role. Another thought that should be considered is if reproducibility efforts are being more prevalent in certain places around the world? Here we can see that the US and the UK are more representative in my dataset. But I then need to ask myself, if this is a symptom of my data sources or sharing policies? I also need to note that I deleted my CC licenses columns becuase I had incorrectly portrayed that data. Anyways, does this suggest that the US and UK simply have better means of sharing data? This is where doing analysis on regional data aggregator DOI counts might be insightful. Nevertheless, the data was worth pursuing because this suggests that either more institutions in a couple countries and certain types of institutions are producing more datasets than others, or simply that they are making their datasets more reproducible by virtue of more being accessible in open data repositories.