Human Development Analysis

Jenn Griffiths's Data Science Portfolio

Install Libraries

These libraries allow us to use the various functions within this document

```
#install.packages("tidyverse")
library(tidyverse)
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
```

Load the Data

Loading the original data into the analysis environment. The original data will not be altered. Versions of the data are created in this environment which can be freely manipulated. If an alteration, such as a removal of a column, was a mistake - one can always reload the original data.

```
Indicators <- read_csv("Data/Indicators.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   dimension = col_character(),
##   indicator_name = col_character(),
##   iso3 = col_character(),
##   country_name = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

View the Dataframes

This function will show us the top 6 rows of the data frame in order to give us an idea of the structure of table.

```
head(Indicators)
```

dimension <chr>	indicator_id <dbl>	indicator_name <chr>	is… <chr>	country_name <chr>	1… <dbl>	1… <dbl>	1… <dbl>	1… <dbl>	1.. <dl
Composite indices	146206	HDI rank	AFG	Afghanistan	*NA*	*NA*	*NA*	*NA*	*N*
Composite indices	146206	HDI rank	ALB	Albania	*NA*	*NA*	*NA*	*NA*	*N*
Composite indices	146206	HDI rank	DZA	Algeria	*NA*	*NA*	*NA*	*NA*	*N*
Composite indices	146206	HDI rank	AND	Andorra	*NA*	*NA*	*NA*	*NA*	*N*

dimension <chr>	indicator_id <dbl>	indicator_name <chr>	is…	country_name <chr>	1… <dbl>	1… <dbl>	1… <dbl>	1… <dbl>	1.. <dl
Composite indices	146206	HDI rank	AGO	Angola	*NA*	*NA*	*NA*	*NA*	*N*
Composite indices	146206	HDI rank	ATG	Antigua and Barbuda	*NA*	*NA*	*NA*	*NA*	*N*

6 rows | 1-10 of 34 columns

Inspecting the Data

```
#Count of rows
nrow(Indicators)
```

```
## [1] 25636
```

```
#Count of columns
ncol(Indicators)
```

```
## [1] 34
```

List of the variables in the dataframe

```
names(Indicators)
```

```
##  [1] "dimension"      "indicator_id"   "indicator_name" "iso3"
##  [5] "country_name"   "1990"           "1991"           "1992"
##  [9] "1993"           "1994"           "1995"           "1996"
## [13] "1997"           "1998"           "1999"           "2000"
## [17] "2001"           "2002"           "2003"           "2004"
## [21] "2005"           "2006"           "2007"           "2008"
## [25] "2009"           "2010"           "2011"           "2012"
## [29] "2013"           "2014"           "2015"           "2016"
## [33] "2017"           "9999"
```

Tidy The Data

Tidying the data is an important part of preparing your data for analysis. Making your data tidy can be time consuming but it makes the format of your data consistent and easier to manage when you're analyzing the data.

Categorical Variables

Change variable 'dimension' to a categorical variable, meaning there are only a certain number of values it could possibly be. This will help with analysis later.

```
Indicators$dimension <-as.factor(Indicators$dimension)

#Display number of levels in 'dimension' variable
nlevels(Indicators$dimension)
```

```
## [1] 14
```

Change variable ‘indicator_name’ to a categorical variable

```
Indicators$indicator_name <-as.factor(Indicators$indicator_name)

#Display levels in 'indicator_name' variable
nlevels(Indicators$indicator_name)
```

```
## [1] 157
```

Change variable ‘country_name’ to a categorical variable

```
Indicators$country_name <-as.factor(Indicators$country_name)

#Display levels in 'country_name' variable
nlevels(Indicators$country_name)
```

```
## [1] 195
```

Change variable ‘indicator_id’ to a categorical variable

```
Indicators$indicator_id <-as.factor(Indicators$indicator_id)
```

Change variable ‘iso3’ to a categorical variable

```
Indicators$iso3 <-as.factor(Indicators$iso3)
```

Gather year columns

An important feature of tidy data is that “each variable must have its own column” As in the variable years should be in a single column. This is done because it makes transforming the data a much smoother process.

```
Indicators <-gather(Indicators, key = Years, "1990", "1991", "1992", "1993", "1994", "1995", "19
96", "1997", "1998","1999", "2000", "2001", "2002","2003","2004","2005","2006", "2007","2008","2
009","2010","2011","2012","2013","2014","2015", "2016", "2017","9999")

#rename last column to rating
colnames(Indicators)[colnames(Indicators) == "1990" ] <- "HDI_value"
```

Subset the data based on ‘dimensions’ category

To break up the data by ‘dimensions’, we must take subsets of the original table. It is good practice to have each table be one thing, rather than a collection of everything. Dimensions in this data set are the category under which the Human

Development Index is scored. For example dimensions include education, health, demographics, and so on.

```
Composite_indices <- subset(Indicators, dimension == "Composite indices" )
Demography <- subset(Indicators, dimension == "Demography" )
Education <- subset(Indicators, dimension == "Education" )
Environmental_sustainability <- subset(Indicators, dimension == "Environmental sustainability" )
Gender <- subset(Indicators, dimension == "Gender" )
Health <- subset(Indicators, dimension == "Health" )
Human_Security <- subset(Indicators, dimension == "Human Security" )
Income_resources <- subset(Indicators, dimension == "Income/composition of resources" )
Inequality <- subset(Indicators, dimension == "Inequality" )
Mobility_communication <- subset(Indicators, dimension == "Mobility and communication" )
Poverty <- subset(Indicators, dimension == "Poverty" )
Sustainability <- subset(Indicators, dimension == "Socio-economic sustainability" )
Financial_flows <- subset(Indicators, dimension == "Trade and financial flows" )
Work_vulnerability <- subset(Indicators, dimension == "Work, employment and vulnerability" )
```

Creating Tibbles

A tibble is a subset table from a larger table (Rows and columns of data). Creating a tibble for each 'dimension, or category for which the Human Development Index was calculated. Here we are creating a tibble from each of the subset tables.

```
  #Composite Indices Tibble
Tib_CompositeIndices <- tibble(indicator = Composite_indices$indicator_name, country = Composite
_indices$country_name, year = Composite_indices$Years, HDI_value = Composite_indices$HDI_value)
  #Demography Tibble
Tib_Demography <- tibble(indicator = Demography$indicator_name, country = Demography$country_nam
e, year = Demography$Years, HDI_value = Demography$HDI_value)
  #Education Tibble
Tib_Education <- tibble(indicator = Education$indicator_name, country = Education$country_name,
year = Education$Years, HDI_value = Education$HDI_value)
  #Environmental Sustainability Tibble
Tib_EnvironmentalSustainability <- tibble(indicator = Environmental_sustainability$indicator_nam
e, country = Environmental_sustainability$country_name, year = Environmental_sustainability$Year
s, HDI_value = Environmental_sustainability$HDI_value)
  #Gender Tibble
Tib_Gender <- tibble(indicator = Gender$indicator_name, country = Gender$country_name, year = Ge
nder$Years, HDI_value = Gender$HDI_value)
  #Health Tibble
Tib_Health <- tibble(indicator = Health$indicator_name, country = Health$country_name, year = He
alth$Years, HDI_value = Health$HDI_value)
  #Human Security Tibble
Tib_HumanSecurity <- tibble(indicator = Human_Security$indicator_name, country = Human_Securit
y$country_name, year = Human_Security$Years, HDI_value = Human_Security$HDI_value)
  #Income Resources Tibble
Tib_IncomeResources <- tibble(indicator = Income_resources$indicator_name, country = Income_reso
urces$country_name, year = Income_resources$Years, HDI_value = Income_resources$HDI_value)
  #Inequality Tibble
Tib_Inequality <- tibble(indicator = Inequality$indicator_name, country = Inequality$country_nam
e, year = Inequality$Years, HDI_value = Inequality$HDI_value)
  #Mobility Communication Tibble
Tib_MobilityCommunication <- tibble(indicator = Mobility_communication$indicator_name, country =
Mobility_communication$country_name, year = Mobility_communication$Years, HDI_value = Mobility_c
ommunication$HDI_value)
  #Poverty Tibble
Tib_Poverty <- tibble(indicator = Poverty$indicator_name, country = Poverty$country_name, year =
Poverty$Years, HDI_value = Poverty$HDI_value)
  #Sustainability Tibble
Tib_Sustainability <- tibble(indicator = Sustainability$indicator_name, country = Sustainabilit
y$country_name, year = Sustainability$Years, HDI_value = Sustainability$HDI_value)
  #Financial Flows Tibble
Tib_FinancialFlows <- tibble(indicator = Financial_flows$indicator_name, country = Financial_flo
ws$country_name, year = Financial_flows$Years, HDI_value = Financial_flows$HDI_value)
  #Work Vulnerability Tibble
Tib_WorkVulnerability <- tibble(indicator = Work_vulnerability$indicator_name, country = Work_vu
lnerability$country_name, year = Work_vulnerability$Years, HDI_value = Work_vulnerability$HDI_va
lue)
```

There are still a lot of missing values (NA in a cell). For now we will leave them in, but an important tool to use when analyzing the data is the function to remove NA values from the calculations:

na.rm = TRUE

Now It’s Time For STEP ZERO: Plot The Data

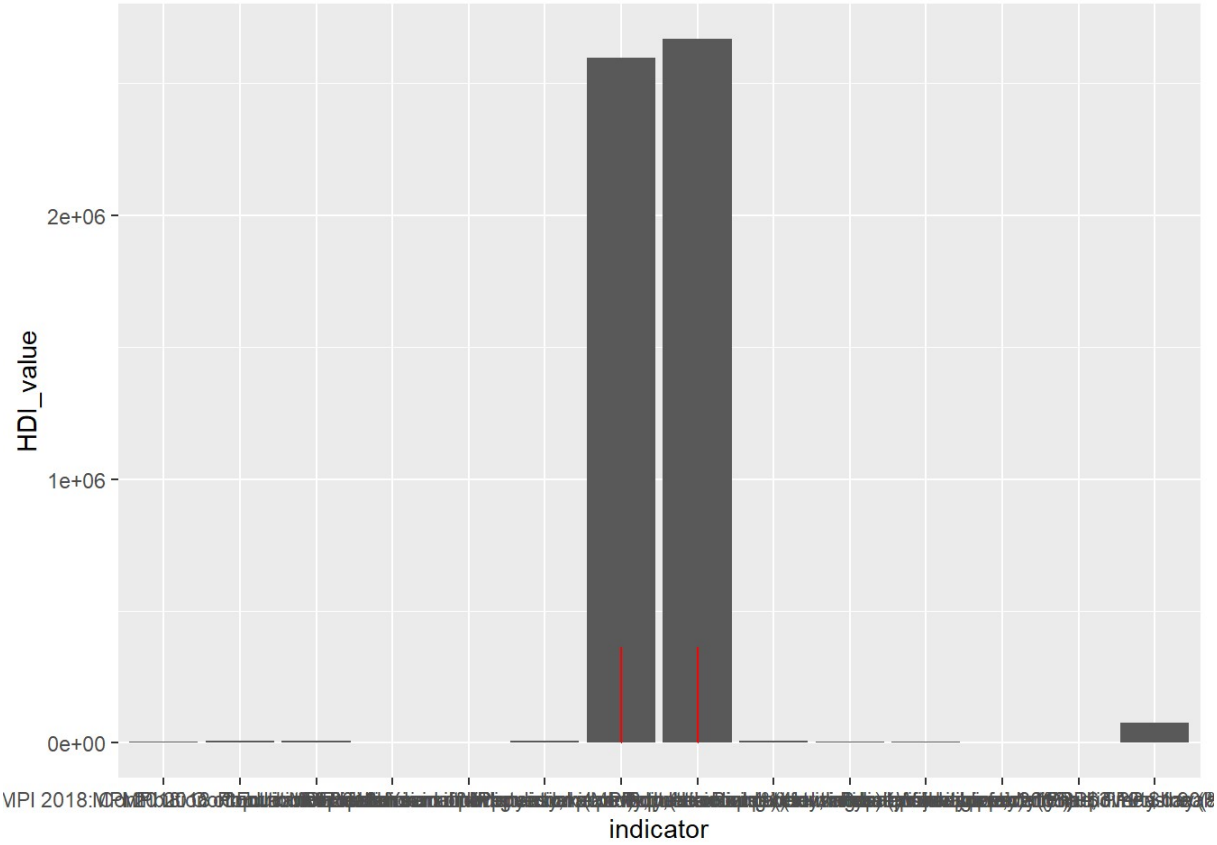
Nothing gives you a better perspective on your data than plotting your data.

Bar Chart across 2 variable

At first glance the two tall bars in the center look as though there is something weird going on here. In looking into the tibble, it is evident that the HDI_value is scoring each indicator with different units. Most scores are in context of percentages, whereas the two tall bars are in terms of headcounts of certain population groups. This leads to the conclusion that these tibbles need to be sliced again based on the indicators in order for the HDI_values in a tibble to be in the same units.

```
ggplot(Tib_Poverty, aes(indicator, HDI_value) ) + geom_col() + geom_path(colour = "red", na.rm = TRUE)
```

```
## Warning: Removed 36655 rows containing missing values (position_stack).
```



Bar Chart

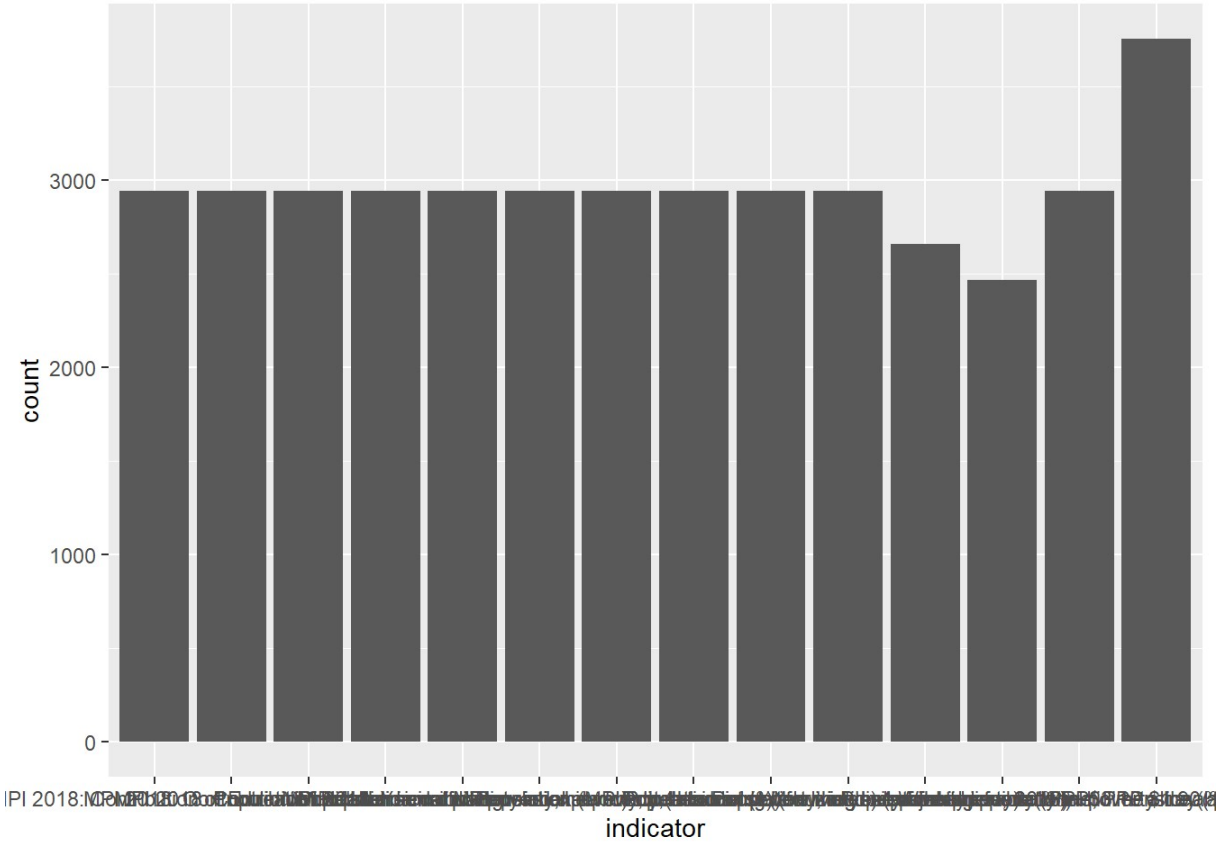
Looking at distributions of a single variable The assumption would be that there is an even spread of the amount of indicators across the data set but here we see there a subtle dip then a rise at the far right of the bar graph.

Is this due to the missing data being removed? Or possibly some other reason for the uneven distribution?

```
names(Tib_Poverty)
```

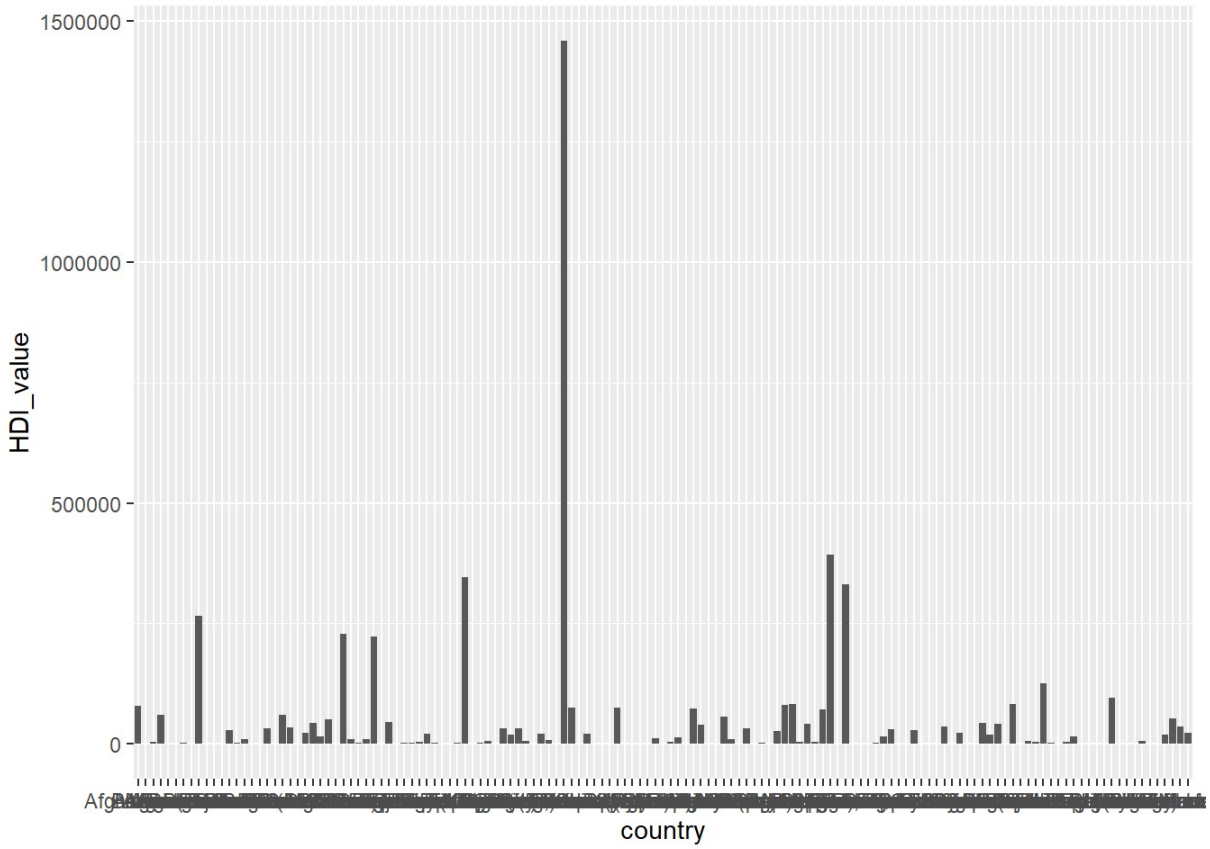
```
## [1] "indicator" "country"   "year"      "HDI_value"
```

```
chart<- ggplot(Tib_Poverty, aes(indicator), na.rm = TRUE)
chart + geom_bar()
```




```
ggplot(Tib_Poverty , aes(x = country, y = HDI_value)) + geom_col()
```

```
## Warning: Removed 36655 rows containing missing values (position_stack).
```


1500000
1000000
500000
0
HDI_value
country