Analysis of Global Data on Internet Usage, Education and Religious Predictions, and Life Satisfaction

Ana Espinosa (ake524) 2021-05-10

Introduction

Three of the four datasets used for this project were gathered from the online publication Our World in Data. The fourth dataset, which is eventually renamed as “world_religion_projections”, was the only one acquired from the website Datahub.

The “happiness” sheet contains data on life satisfaction. The World Happiness Report does a Gallup World Poll that asks participants to rank their life on a scale from one to ten (this is reffered to as the “Cantril Ladder”). The responses from every country are averaged and recorded from the year 2005 to 2018. The variable “satisfaction” is the cantril ladder score. The “internet_usage” dataset contains information about the percentage of countries’ population that use the internet from the year 1990 to 2017. The variables are simply the countries, country code, year, and the percentage of internet users.

The “education_projections” data contains predictions of the rates of no education in countries from the year 1970 all the way to the year 2050. This data was published in 2015 by the International Institute of Applied Systems Analysis. Fiinally, the dataset “world_religion_projections” contains the estimated religious composition of 198 countries for 2010 to 2050. This dataset has many variables including country, region, year, and the percentage of people that are Christians, Muslims, Jews, Hindus, practice folk religions, are unaffiliated or answered “other.” The projections for the education and religion data were based on the current trends seen in each country.

I found these datasets interesting because I tend to be curious about global trends and how life style varies from country to country. I am also really interested to know how these differences in life style can impact people’s quality of life. I know education, religion, and (increasingly so) internet usage are major factors that can impact people’s perception of the world as well as their general satisfaction with life. I wanted to see how these factors/trends related to one another. My intuition tells me that countries with higher predicted education *rates and lower religious affiliation percentages will also see higher satisfaction scores. I am not sure how internet usage will relate to satisfaction.*

```
#opening datasets
library(readxl)

#happiness_cantril_ladder <- read_excel("C:/Users/anaes/Desktop/SDS/happiness-cantril

#share_of_individuals_using_the_internet <- #read_excel("C:/Users/anaes/Desktop/SDS/s

#projections_of_the_rate_of_no_education_based_on_current_global_education_trends <-

#by_rounded_percentage_share_csv <- read_excel("C:/Users/anaes/Desktop/SDS/by_rounded
```

```
#downloading necessary libraries
library(dplyr)
library(tidyr)
library(tidyverse)
library(cluster)
library(factoextra)
library(kableExtra)
```

Tidying Data

```
#changing the datasets names to simpler and more concise names

#happiness <- happiness_cantril_ladder

#internet_usage <- share_of_individuals_using_the_internet

#world_religion_projections <- by_rounded_percentage_share_csv%>%
# filter(year== "2050")

#education_projections <- projections_of_the_rate_of_no_education_based_on_current_gl
```

*Next, the four datasets were joined by using the inner_join function. I chose this join because I only wanted to keep the rows that had matches for each dataset and not have a bunch of NA values. Of the several rows in each dataset, only 87 rows had matches in all of them. To simplify my final dataset, I also decided to only use data from the year 2016 and predictions for the year 2050. The year 2016 is the most recent year that allows the dataset to have over 50 observations. I also felt using education and religious projections of 2050 would better highlight differences between countries. My final dataset did not include any information from any original sheet that was not from these years, so many rows were removed. The size of the final joined, clean dataset was 17 columns wide and 87 rows long. *

contains revision

```
#joining datasets with inner join
#x_data<- world_religion_projections%>%
#  inner_join(happiness, by = c("country"="entity"))

#y_data<- x_data%>%
#  inner_join(internet_usage, by = c("country"="entity"))

#z_data<- y_data%>%
#  inner_join(education_projections, by = c("country"= "entity"))

#only keeping cases of happiness and internet_usage datasets  of the latest year most
#only keeping predictions for 2050
#final_data<- z_data%>%
#  filter(year.y== '2016')%>%
#  filter(year.x.x== '2016')%>%
#  filter(year.y.y == '2050')%>%
#  select(-region, -code.x, -code.y)

#creating an excel sheet with all of my data to access it more easily
#write_csv(final_data,"Project1_data.csv")

#ncol(Project_One)
#nrow(Project_One)
```

Summary Statistics

The next part of the process was to create summary statistics for each of the main variables. But first, I created a new numeric variable by manipulating the existing "unaffiliated" variable in my dataset. Now I could see how many people had any kind of religious association. Aside for looking at the number of rows and columns of my dataset, I found the interquartile values, mean, standard deviation, and min/max for each variable, displaying them in tables by using the kableEstra package. I also created a correlation matrix for my numeric variables. The two variables that had the strongest correlation were percent of internet usage and satisfaction score, with a strong correlation coefficient of 0.78.

I then created a new categorical variable, “top_religion” that shows what is the dominant religion of each country. I did this by counting the religion with anything above 44 percent as the top religion. Since some percentages were more spread out and some countries’ top religion had a smaller percentage than 44, some observations had NAs in this category and were removed. These countries were Singapore, Togo, and Veitnam. This categorical variable was used to create more summary statistics. The mean satisfaction, the top satisfaction score, and the country who this top score belonged to were recorded for every religion. The category for unaffiliated had the highest mean satisfaction score while the christianity category had the country with the highest individual satisfaction score- Norway.

```
#opening my project data
Project_One <- read.csv("C:/Users/anaes/Desktop/SDS/Project1_data.csv")

#creating new variable with mutate that gives percentage of people affiliated with an
Project_One <- Project_One %>%
  #using mutate function
  mutate(affiliated = 100- unaffiliated)

#summarizes basic information about dataset
Project_One %>%
  summarize('number of rows' = n(),
            'number of colums' = ncol(Project_One),
            'number of countries' = n_distinct(country))%>%
  kbl() %>%
  kable_paper("hover", full_width = F) #used to make tables
```

number of rows	number of colums	number of countries
87	18	87

*A total of 87 rows- one for each country with 18 variables. Three of those variables are just there to specify the year the data point took place or the year for the projections. *

```
#summarizes the quantile values of each main variable
Project_One %>%
  summarize('Internet Usage' =quantile(int_per_pop),
    'Satisfaction'= quantile(satisfaction),
       'Affiliated'= quantile(affiliated),
          'No education Rates'=  quantile(no_education_rates)) %>%
  data.frame(Stats = c("min", "Q1", "median", "Q3", "max"))%>%
kbl() %>%
  kable_paper("hover", full_width = F)%>%
  add_header_above(c(" ", "Quantile Values" = 4))
```

	Quantile Values			
Internet.Usage	**Satisfaction**	**Affiliated**	**No.education.Rates**	**Stats**
4.00000	2.693061	40.9	0.0043700	min
22.49808	4.493280	88.6	0.2474282	Q1
53.40000	5.416875	96.4	1.1051812	median
71.83314	6.013322	99.0	4.8657473	Q3
97.99998	7.596332	99.0	40.1670285	max

```
#summarizes the mean values of each main variable
Project_One %>%
  summarize('Internet Usage(%)' = mean(int_per_pop), 'Satisfaction'= mean(satisfactio
            'Affiliated(%)' = mean(affiliated),
            'No education Rates(%)' = mean(no_education_rates))%>%
kbl() %>%
  kable_paper("hover", full_width = F)%>%
  add_header_above(c(" ", "Mean Values" = 3))
```

	Mean Values		
Internet Usage(%)	**Satisfaction**	**Affiliated(%)**	**No education Rates(%)**
48.8431	5.293748	90.66207	4.789264

```
Project_One %>%
  summarize('Internet Usage' = sd(int_per_pop), 'Satisfaction'= sd(satisfaction),
            'Affiliated' = sd(affiliated),
            'No education Rates' = sd(no_education_rates))%>%
kbl() %>%
  kable_paper("hover", full_width = F)%>%
  add_header_above(c(" ", "Standard Deviation" = 3))
```

	Standard Deviation		
Internet Usage	**Satisfaction**	**Affiliated**	**No education Rates**
26.99001	1.046043	12.87419	8.507755

```
#summarizes the min and max values of each main variable
Project_One %>%
  summarize(min_int=min(int_per_pop),
            max_int=max(int_per_pop),
    min_sat=min(satisfaction),
     max_sat=max(satisfaction),
        min_aff=min(affiliated),
    max_aff=max(affiliated),
          min_edu=min(no_education_rates),
          max_edu=max(no_education_rates))%>%
kbl() %>%
  kable_styling(bootstrap_options = c("striped", "hover"))%>%
  add_header_above(c("Internet Usage (%)" = 2, "Satisfaction" = 2, "Affiliated (%)" =
```

Internet Usage (%)		Satisfaction		Affiliated (%)		No Educati
min_int	max_int	min_sat	max_sat	min_aff	max_aff	min_edu
4	97.99998	2.693061	7.596332	40.9	99	0.00437

Note that the "no education rates" variable is counting the percentage of people /without/ an education. So, it makes sense that it has a negative correlation with satisfaction and internet usage.

```
#creates object with only the main numeric varaibles
project_num <- Project_One%>%
  select(satisfaction, int_per_pop, no_education_rates, affiliated)

#makes correlation matrix of numeric variables
cor(project_num, use = "pairwise.complete.obs")%>%
   kbl() %>%
  kable_classic_2(full_width = F)%>%
   add_header_above(c(" ", "Correlation Matrix" = 4))
```

	Correlation Matrix			
	satisfaction	int_per_pop	no_education_rates	af
satisfaction	1.0000000	0.7829212	-0.4772952	-0.3
int_per_pop	0.7829212	1.0000000	-0.5953827	-0.5
no_education_rates	-0.4772952	-0.5953827	1.0000000	0.2

Correlation Matrix

no_education_rates	[illegible]	[illegible]	[illegible]	[illegible]
affiliated	-0.3442495	-0.5156783	0.2880909	1.0

```
#creating a new categorical variable besides country
Project_One_2 <- Project_One%>%
  mutate(top_religion = case_when(christians>44 ~ "christianity",
                                  buddhists>44 ~ "buddhism",
                                  muslims>44 ~ "islam",
                                  jews>44~ "judaism",
                                  hindus>44~ "hinduism",
                                  unaffiliated>44~ "unaffiliated",
                                  folk_religions >44 ~ "folk")) %>%
  mutate(education_rates = 100 - no_education_rates)
```

```
#removing incomplete cases which resulted in the removal of three rows
Project_One_3 <- Project_One_2%>%
  filter(complete.cases(top_religion))
```

```
#The following paragraphs of code finds the mean satisfaction score of each religion
Project_One_3%>%
  group_by(top_religion)%>%
  filter(top_religion == "christianity")%>%
  summarise(mean_sat= mean(satisfaction))%>%
  arrange(desc(mean_sat))
```

```
## # A tibble: 1 x 2
##   top_religion mean_sat
##   <chr>           <dbl>
## 1 christianity     5.41
```

```
Project_One_3%>%
  group_by(top_religion)%>%
  filter(top_religion == "islam")%>%
  summarise(mean_sat= mean(satisfaction))%>%
  arrange(desc(mean_sat))
```

```
## # A tibble: 1 x 2
##   top_religion mean_sat
##   <chr>           <dbl>
## 1 islam            5.01
```

```
Project_One_3%>%
  group_by(top_religion)%>%
  filter(top_religion =="hinduism")%>%
  summarise(mean_sat= mean(satisfaction))%>%
  arrange(desc(mean_sat))
```

```
## # A tibble: 1 x 2
##   top_religion mean_sat
##   <chr>           <dbl>
## 1 hinduism         4.96
```

```
Project_One_3%>%
  group_by(top_religion)%>%
  filter(top_religion == "buddhism")%>%
  summarise(mean_sat= mean(satisfaction))%>%
  arrange(desc(mean_sat))
```

```
## # A tibble: 1 x 2
##   top_religion mean_sat
##   <chr>           <dbl>
## 1 buddhism         5.27
```

```
Project_One_3%>%
  group_by(top_religion)%>%
  filter(top_religion =="unaffiliated")%>%
  summarise(mean_sat= mean(satisfaction))%>%
  arrange(desc(mean_sat))
```

```
## # A tibble: 1 x 2
##   top_religion mean_sat
##   <chr>           <dbl>
## 1 unaffiliated     5.78
```

```
#creating table for mean values
summary_tbl <- data.frame(
  `Mean` = c(5.405919   ,5.005976       ,4.962907,5.26745       ,5.783765   ),
  Religion= c("christianity","islam","hinduism","buddhism","unaffiliated"))

  kbl(summary_tbl) %>%
kable_classic_2(full_width = F)%>%
  add_header_above(c("Mean Satisfaction Score for Each Religion" = 2))
```

Mean Satisfaction Score for Each Religion	
Mean	**Religion**
5.405919	christianity
5.005976	islam
4.962907	hinduism
5.267450	buddhism
5.783765	unaffiliated

The unaffiliated category has the highest satisfaction mean, with christianity coming in second.

```
#The following paragraphs of code finds the maximum satisfaction score of each religi
Project_One_3%>%
  group_by(country)%>%
  filter(top_religion == "christianity")%>%
  summarise( satisfaction)%>%
  arrange(desc(satisfaction))
```

```
## # A tibble: 52 x 2
##    country       satisfaction
##    <chr>                <dbl>
##  1 Norway                7.60
##  2 Canada                7.24
##  3 Costa Rica            7.14
##  4 Austria               7.05
##  5 Ireland               7.04
##  6 Mexico                6.82
##  7 United States         6.80
##  8 Chile                 6.58
##  9 Argentina             6.43
## 10 Brazil                6.37
## # ... with 42 more rows
```

```
Project_One_3%>%
  group_by(country)%>%
  filter(top_religion == "islam" )%>%
  summarise(satisfaction)%>%
  arrange(desc(satisfaction))
```

```
## # A tibble: 22 x 2
##    country      satisfaction
##    <chr>               <dbl>
##  1 Saudi Arabia         6.47
##  2 Bahrain              6.17
##  3 Uzbekistan           5.89
##  4 Turkmenistan         5.89
##  5 Pakistan             5.55
##  6 Kazakhstan           5.53
##  7 Morocco              5.39
##  8 Turkey               5.33
##  9 Jordan               5.27
## 10 Nigeria              5.22
## # ... with 12 more rows
```

```
Project_One_3%>%
  group_by(country)%>%
  filter(top_religion =="hinduism")%>%
  summarise(satisfaction)%>%
  arrange(desc(satisfaction))
```

```
## # A tibble: 3 x 2
##   country   satisfaction
##   <chr>            <dbl>
## 1 Mauritius         5.61
## 2 Nepal             5.10
## 3 India             4.18
```

```
Project_One_3%>%
  group_by(country)%>%
  filter(top_religion == "buddhism")%>%
  summarise(satisfaction)%>%
  arrange(desc(satisfaction))
```

```
## # A tibble: 2 x 2
##   country  satisfaction
##   <chr>           <dbl>
## 1 Thailand         6.07
## 2 Cambodia         4.46
```

```
Project_One_3%>%
  group_by(country)%>%
  filter(top_religion =="unaffiliated")%>%
  summarise(satisfaction)%>%
  arrange(desc(satisfaction))
```

```
## # A tibble: 5 x 2
##   country     satisfaction
##   <chr>              <dbl>
## 1 France              6.48
## 2 South Korea         5.97
## 3 Estonia             5.65
## 4 Hong Kong           5.50
## 5 China               5.32
```

```
#Creates table displaying the top scores for each religion and the country they belon
summary_tbl <- data.frame(
  Country= c("Norway", "Saudi Arabia", "Mauritius",'Thailand',"France" ),
  `Max Satisfaction Score` = c(7.596332,6.473921    ,5.610003,6.073640  ,6.475209),
  Religion= c("christianity","islam","hinduism","buddhism","unaffiliated"))

  kbl(summary_tbl) %>%
kable_classic_2(full_width = F)%>%
  add_header_above(c("Top Satisfaction Score for Each Religion" = 3))
```

Top Satisfaction Score for Each Religion		
Country	**Max.Satisfaction.Score**	**Religion**
Norway	7.596332	christianity
Saudi Arabia	6.473921	islam
Mauritius	5.610003	hinduism
Thailand	6.073640	buddhism
France	6.475209	unaffiliated

Norway is the country with the highest satisfaction score and is also christian

Visualizations

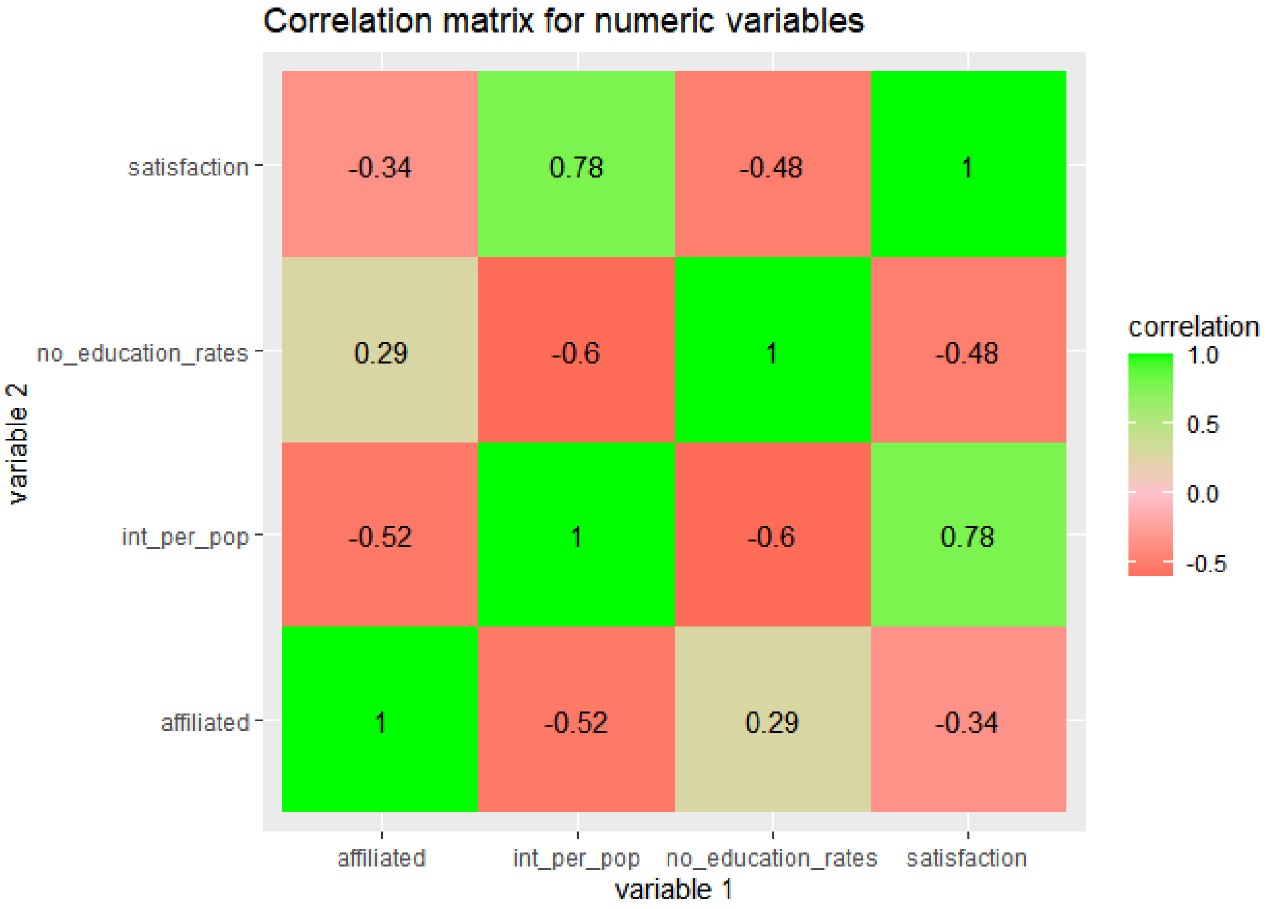
After the summary statistics, I then created four different visualizations. I first made a heatmap for the previous correlation matrix. It makes it easier to see that education rates and percent of affiliation are the least correlated variables while satisfaction and internet usage are the most correlated. The next visualization is a scatterplot of percent of internet usage vs satisfaction. I made a new variable “education rates” which is just the inverse of “no education rates” to make it easier to interplet the graph. The color of the data points shows the level of education rates. There is a clear positive correlation between satisfaction and internet usage. As satisfaction and internet usage grows, so do the education rates. This is consistent with the correlation heatmap information.

The next plot shows percent of religious affiliation against satisfaction levels. This graph does not have a relationship as clear as the first one. Though there does seem to be a slight dip in satisfaction at high affiliation. Education rates increase as satisfaction increases and affiliation decreases.

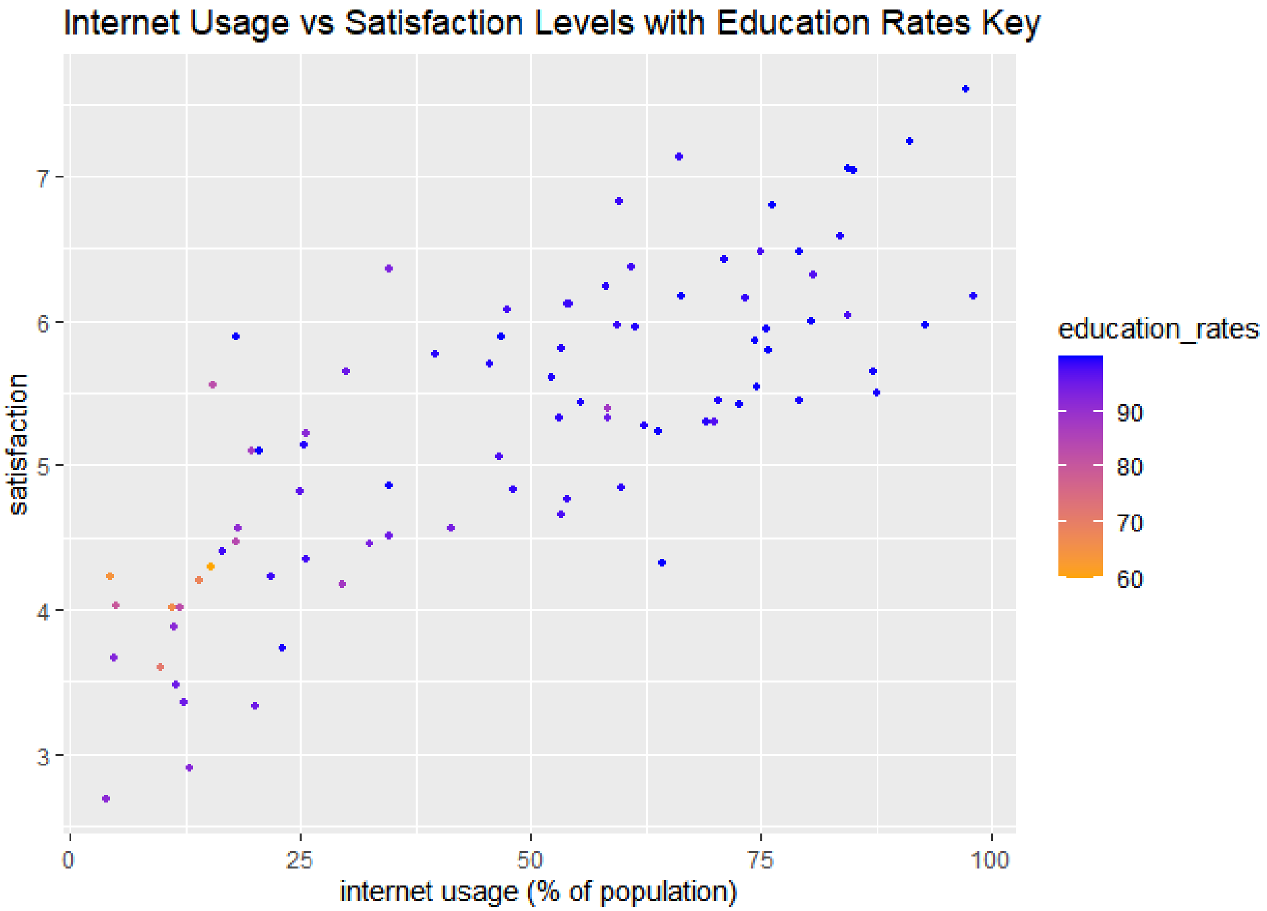
The final visualization is a bar chart which integrates a newly created categorical variable. This new variable groups each country into a “top” and “bottom” satisfaction type. The top countries have satisfaction scores higher than 5 (out of 10) and bottom countries have scores lower than 5. It seems like both halfs had similar ratios of each major religion but the top half contained all the countries in the unaffiliated categories. This bar chart also helps to easily visualize the different of internet usage for each satisfaction type.

-

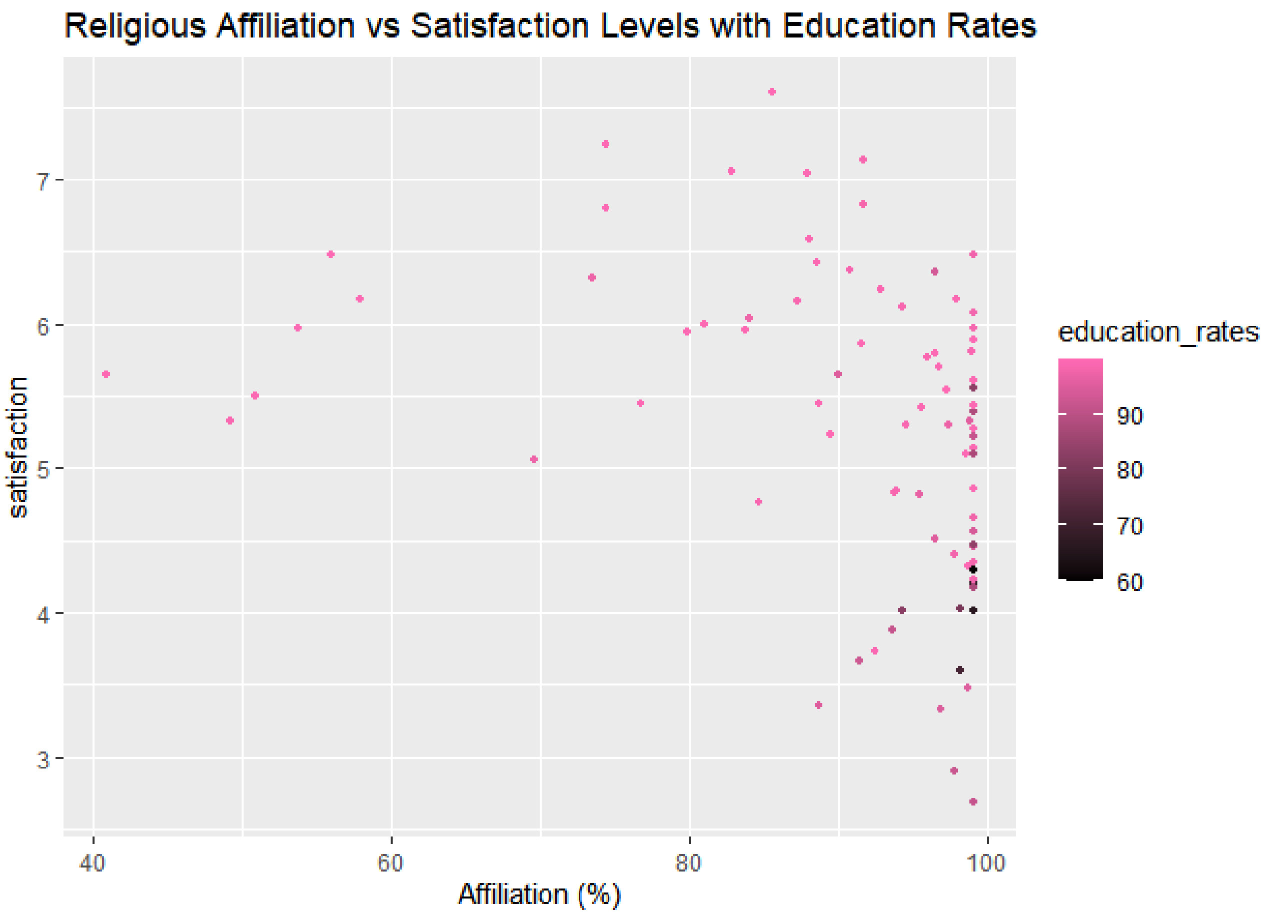
```
#creates heatmap

cor(project_num, use = "pairwise.complete.obs") %>%
  # Save as a data frame
  as.data.frame %>%
  # Convert row names to an explicit variable
  rownames_to_column %>%
  # Pivot so that all correlations appear in the same column
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  ggplot(aes(rowname, other_var, fill=correlation)) +
  # Heatmap with geom_tile
  geom_tile() +
  # Change the scale to make the middle appear neutral
  scale_fill_gradient2(low="red",mid="pink",high="green") +
  # Overlay values
  geom_text(aes(label = round(correlation,2)), color = "black", size = 4) +
  # Give title and labels
  labs(title = "Correlation matrix for numeric variables", x = "variable 1", y = "var
```



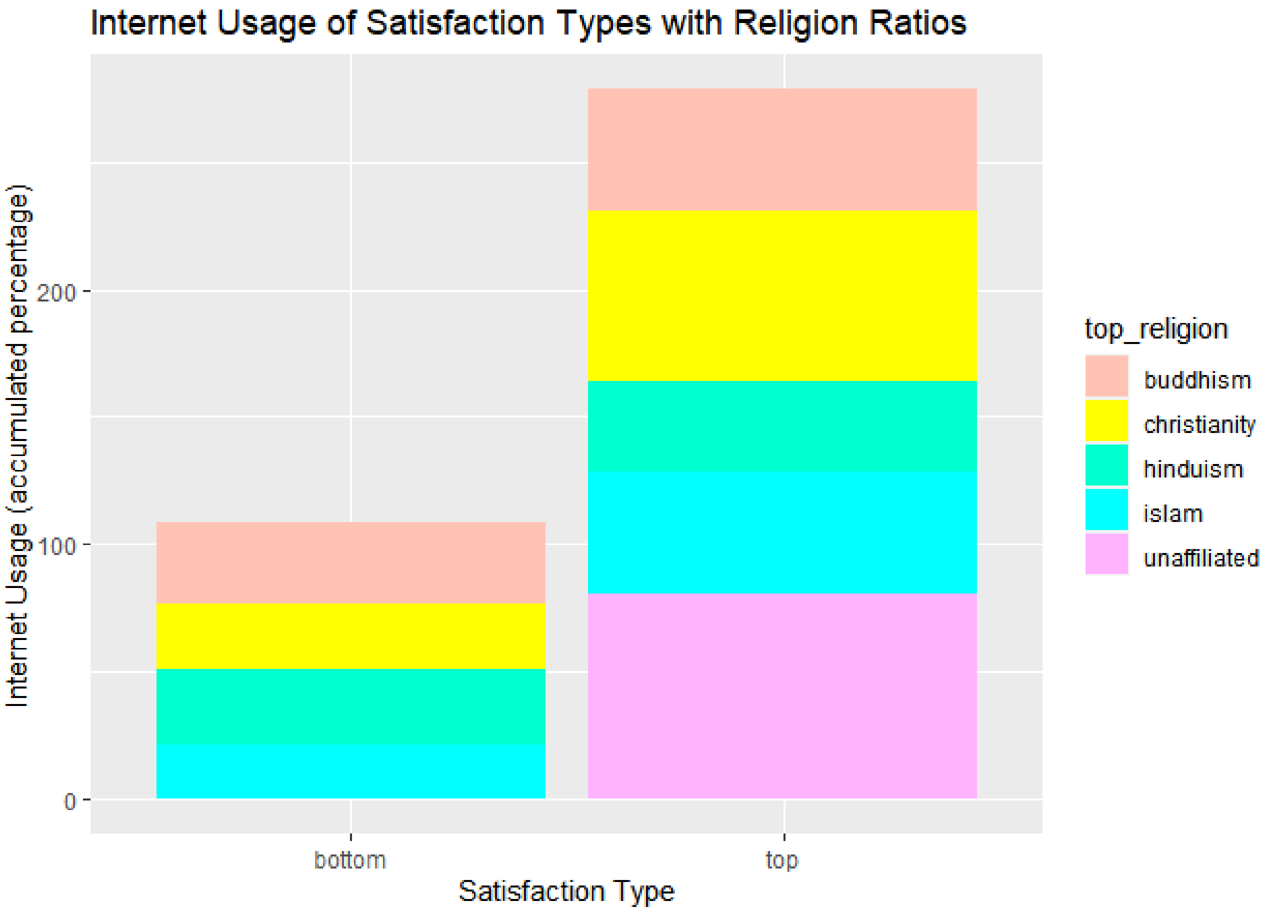
```
 # creates scatterplot
ggplot(data = Project_One_2, aes(x = int_per_pop, y = satisfaction, color = education
  geom_point(size = 1)   + ggtitle("Internet Usage vs Satisfaction Levels with Educat
  # choose colors of the scale
  scale_color_gradient(low="orange", high="blue") + xlab("internet usage (% of popula
```



```
#creates scatterplot
ggplot(data = Project_One_2, aes(x = affiliated, y = satisfaction, color = education_
  geom_point(size = 1)   + ggtitle("Religious Affiliation vs Satisfaction Levels with
  # choose colors of the scale
  scale_color_gradient(low="black", high="hot pink")+ xlab("Affiliation (%)")
```



```
#creating new categorical variable "satisfaction_half"
Project_One_3 <- Project_One_3%>%
  mutate(satisfaction_half = case_when(satisfaction>5  ~ "top",
                                       satisfaction <5 ~ "bottom"))

ggplot(Project_One_3, aes( x= satisfaction_half, fill=top_religion,)) +
  geom_bar(aes(y = int_per_pop),stat="summary", fun="mean") +  scale_fill_hue(l = 100
```



Performing PCA

*The final part of this project was to perform a Principal Component Analysis (PCA). The first step in the PCA was to scale the numeric variables and then use the ‘prcomp’ function to perform the PCA. The eigenvalue for each PC was calculated. Only PC1 had a eigenvalue higher than one. I then created a scree plot to see how much variance each PC contributes. There is a big jump from PC1 and PC2 and both of these contribute to over 80% of the variance. Two principal components were chosen after this scree plot analysis. A correlation circle plot was used to visually demonstrate the contribution of each variable to PC1 and PC2.From the circle, one can see that internet usage contributes most to dimension one while affiliation contributes the most to dimension 2. *

*Finally, the PAM function was used to cluster the observations into two separate clusters. The number of clusters were chosen after using the fviz_nbclust() function. The two clusters are most distinct along dimension one. Cluster one seems to group all of the countries with higher education rates, satisfaction scores, and internet usage away from countries with higher religious affiliation in cluster 2. *

```
#switching "no_education_rates" to "education_rates" to simplify comprehension
projectnum2 <- project_num %>%
  mutate(education_rates = Project_One_2$education_rates)%>%
  select(-no_education_rates)

#scaling the dataset
project_scaled <- projectnum2 %>%
    scale

#Performing PCA!
project_pca <- project_scaled %>%
  prcomp
names(project_pca)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

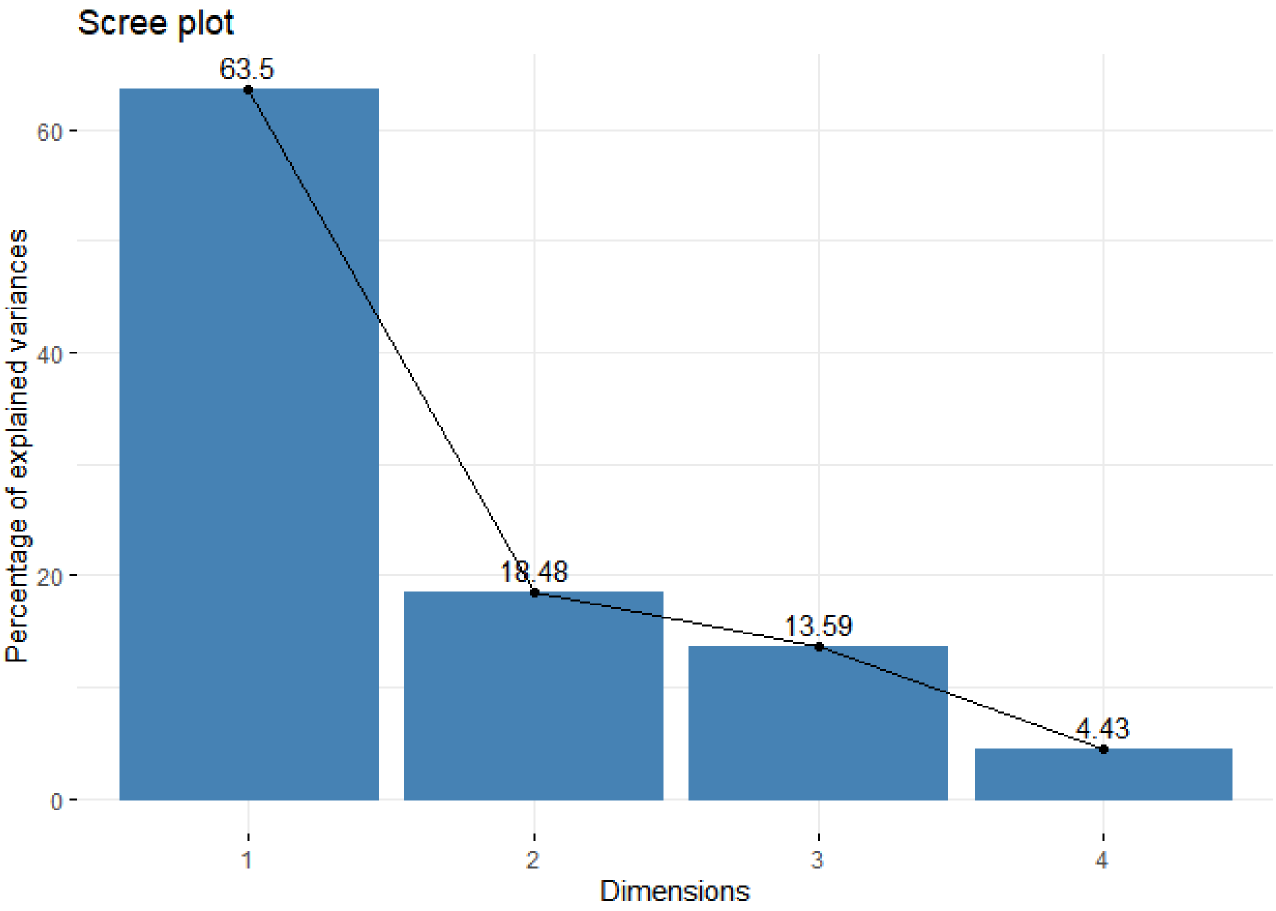
```
percent <- 100* (project_pca$sdev^2 / sum(project_pca$sdev^2))
percent[1] + percent[2]
```

```
## [1] 81.98174
```

```
#gives eigen values
get_eigenvalue(project_pca)
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1  2.5401428        63.503570                    63.50357
## Dim.2  0.7391267        18.478168                    81.98174
## Dim.3  0.5434092        13.585230                    95.56697
## Dim.4  0.1773213         4.433033                   100.00000
```

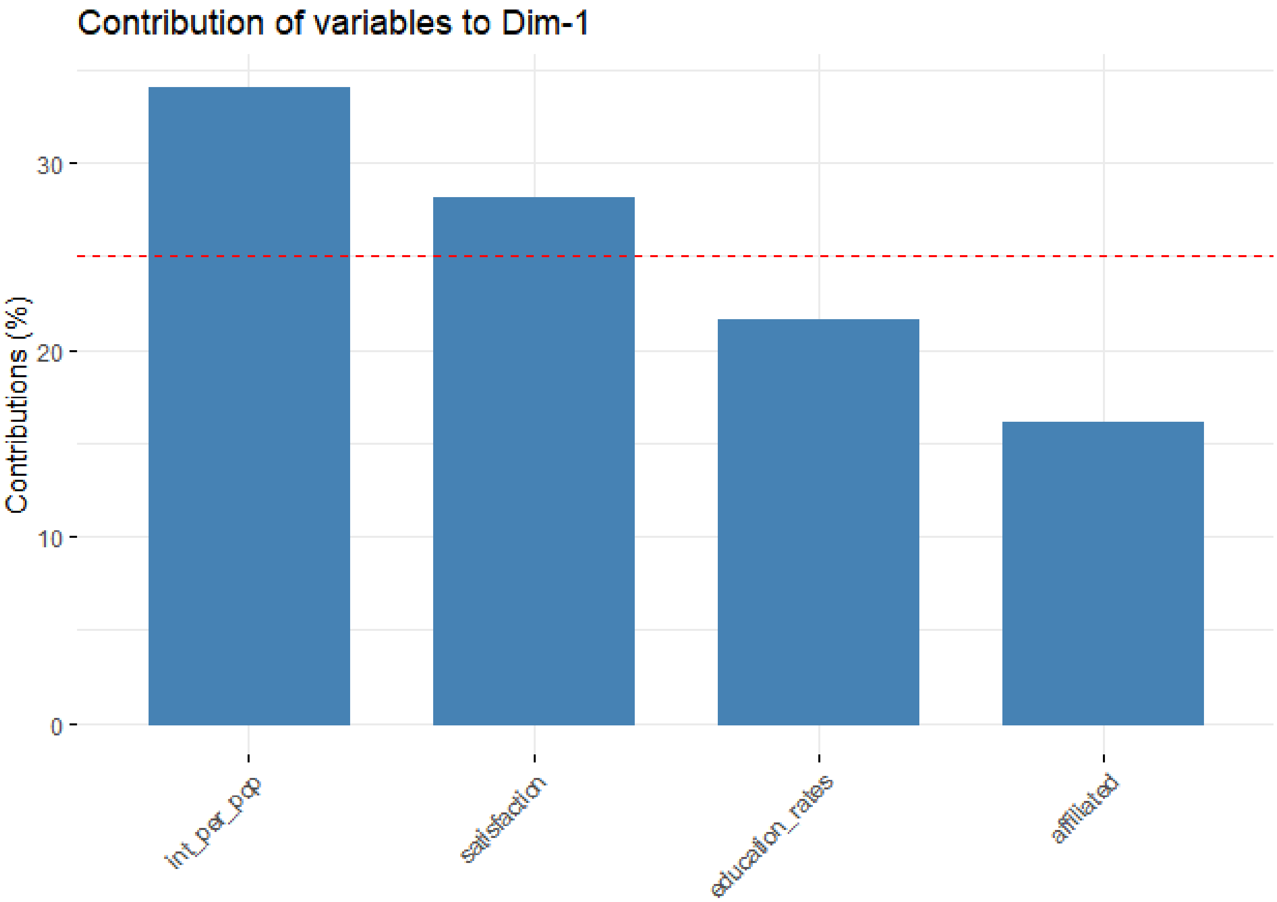
```
# Scree plot with percentages as text
fviz_screeplot(project_pca) +  geom_text(aes(label = round(percent, 2)), size = 4, vj
```



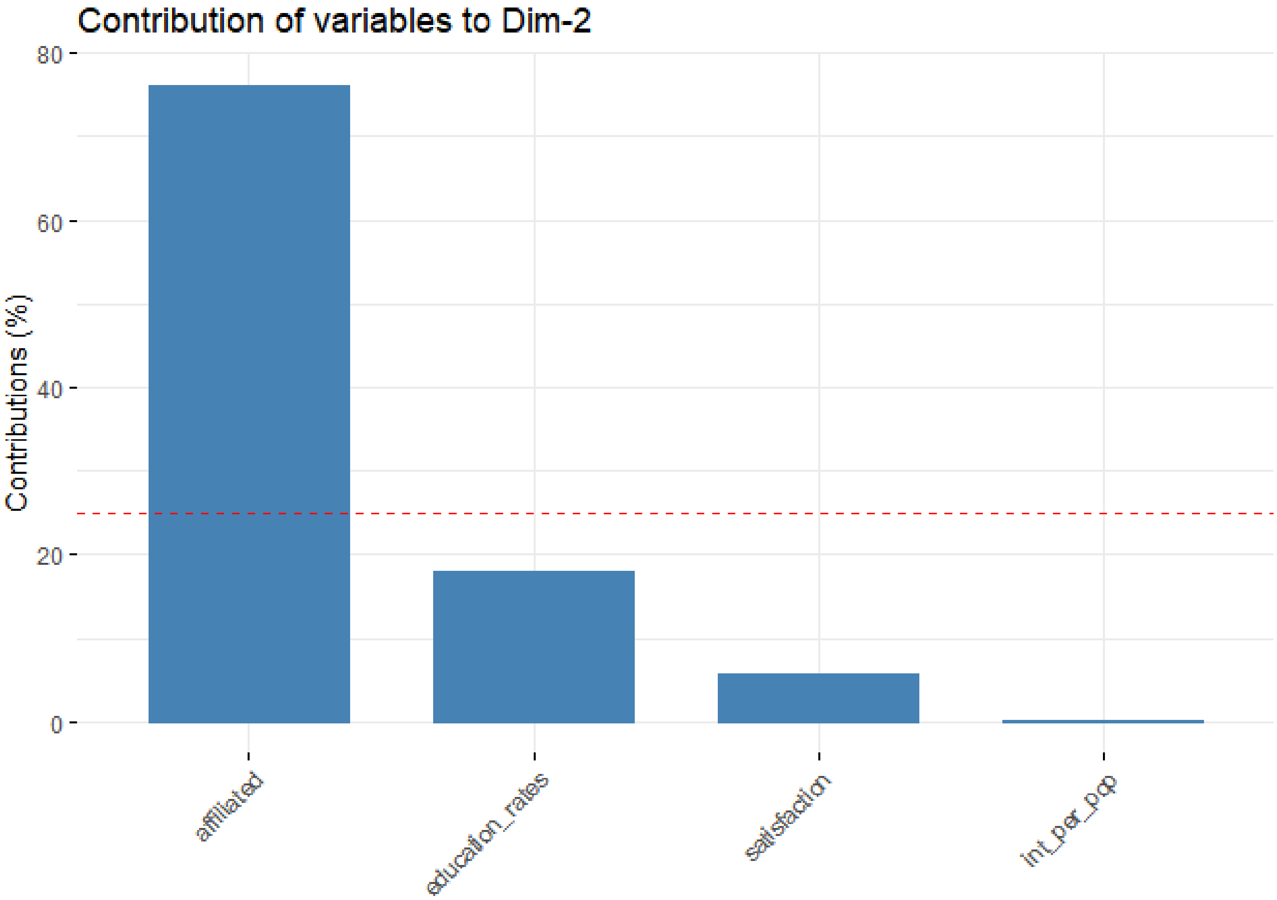
```
#displays the variation each variable contributes to each PC
get_pca_var(project_pca)$contrib
```

```
##                   Dim.1      Dim.2     Dim.3     Dim.4
## satisfaction    28.16315  5.6344249 34.303037 31.899390
## int_per_pop     34.09888  0.2164187  4.306909 61.377788
## affiliated      16.09313 76.2071181  3.903693  3.796054
## education_rates 21.64483 17.9420383 57.486362  2.926768
```

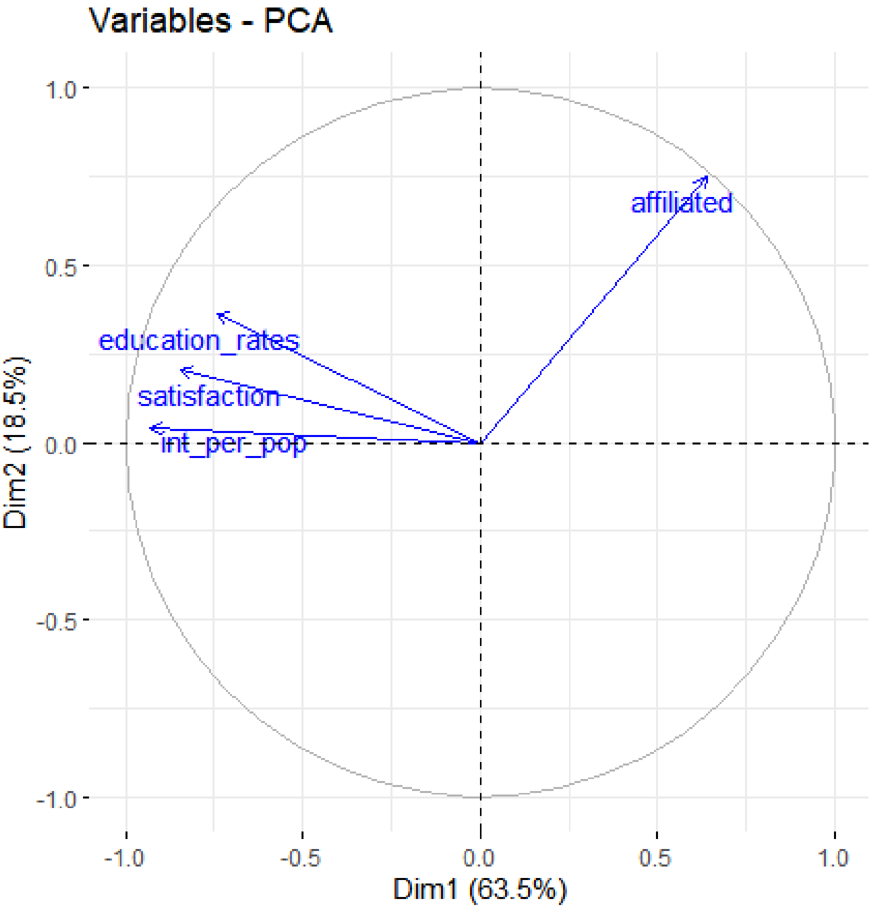
```
fviz_contrib(project_pca, choice = "var", axes = 1, top = 4)
```



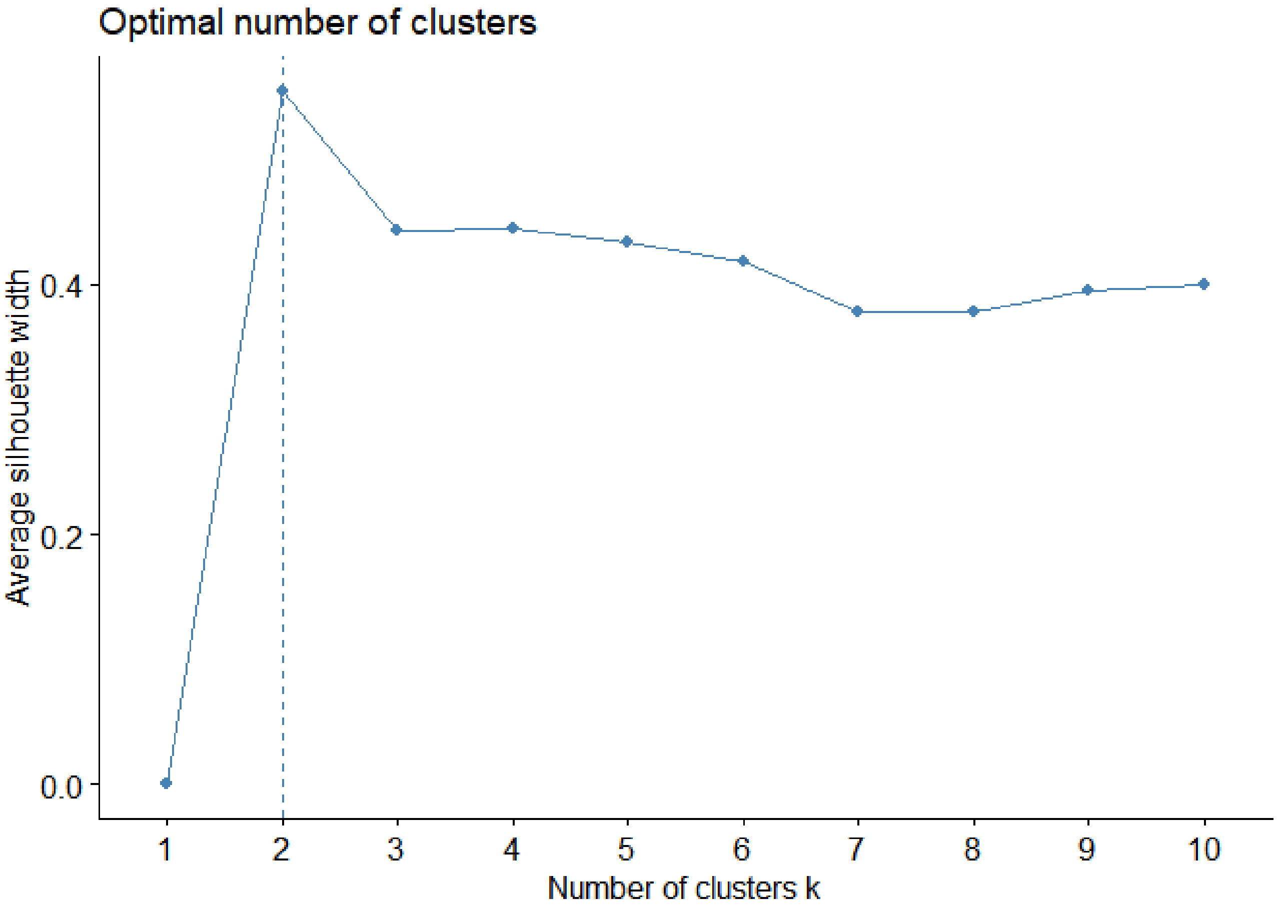
```
fviz_contrib(project_pca, choice = "var", axes = 2, top = 4)
```



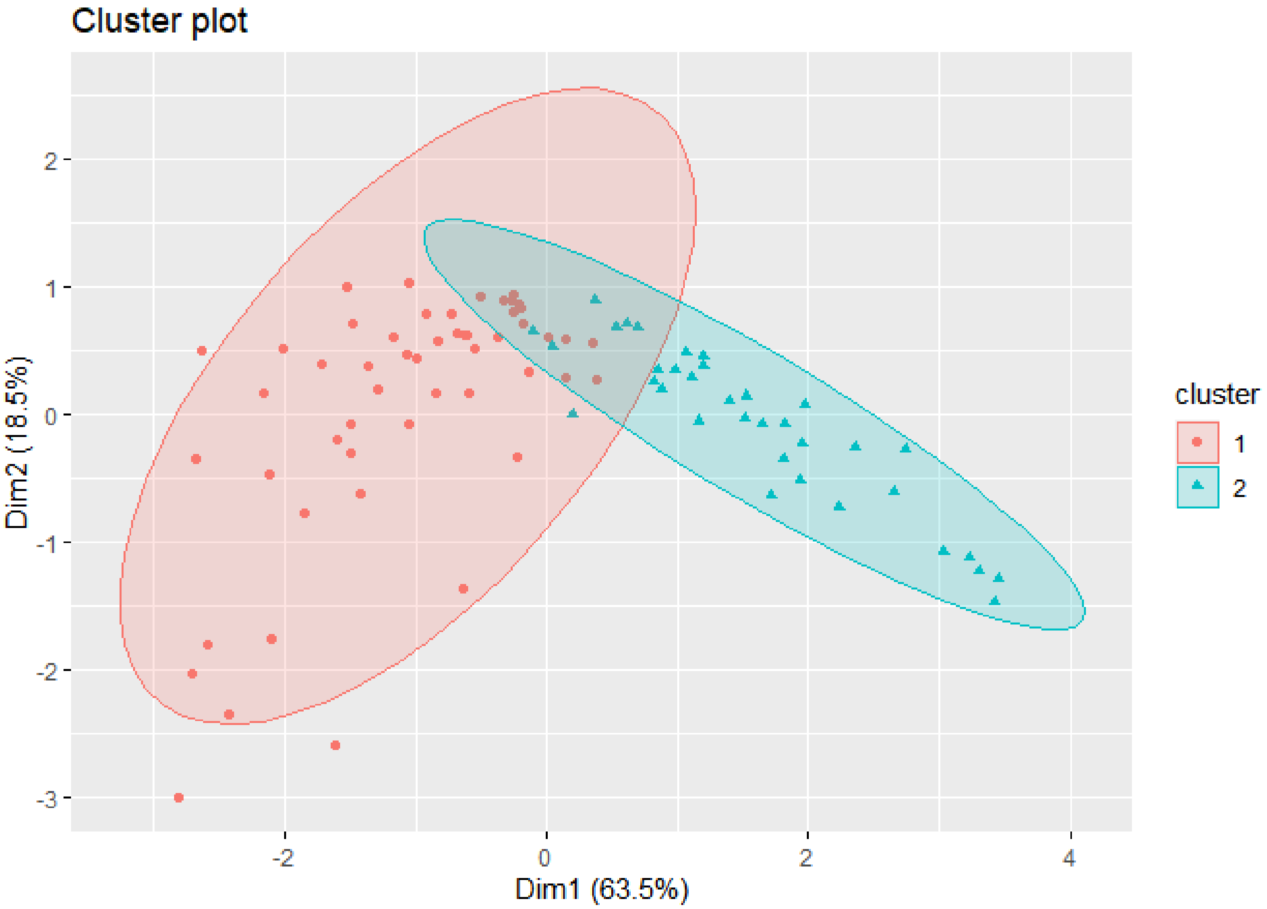
```
#visualizing contribution
fviz_pca_var(project_pca, col.var = "blue",axes = 1:2, repel = TRUE)
```



```
# selects appropriate amount of clusters
fviz_nbclust(projectnum2, FUNcluster = pam, method = "s")
```



```
#using PAM functions with two clusters
projectpam <- project_num %>%
  pam(k=2)

#Visualizing the cluster groups according to the variables PC1 and PC2
fviz_cluster(projectpam,
             geom = "point",  # show points only (not "text")
             ellipse.type = "norm")
```


Cluster plot
Dim2 (18.5%)
2
1
0
-1
-2
-3
Dim1 (63.5%)
-2
0
2
4
cluster
1
2