FIT3152 Assignment 1

SHIN MINSEO 35865377

2025-04-17

*All detailed codes are attached in the appendix.

Question 1a.

Overall Dataset

The dataframe consists 50,000 rows and 40 columns. This represents a sample of 50,000 respondents randomly selected from the survey dataset, and 40 variables that span a range of social, political, economic concepts. The first column `Country` is a character variable, and remaining 39 variables are integers that match with responses to survey questions.

Missing Data

When tested with `is.na()`, the dataset appears to have no missing values, but after checking the codebook provided by WVS, it was revealed that the response uses special negative values to represent missing or invalid answers. For example, input '-1' means "Don't Know", '-2' means 'No Answer', and so on. There are total 41,676 missing values.

Distribution of Numerical attributes

Using summary(`VC`), I could observe minimum, maximum, median, mean values. The summary reveals that most variables use Likert-type ordinal scales. Some variables range from 1 to 4, 1 to 10, or 0 to 10.

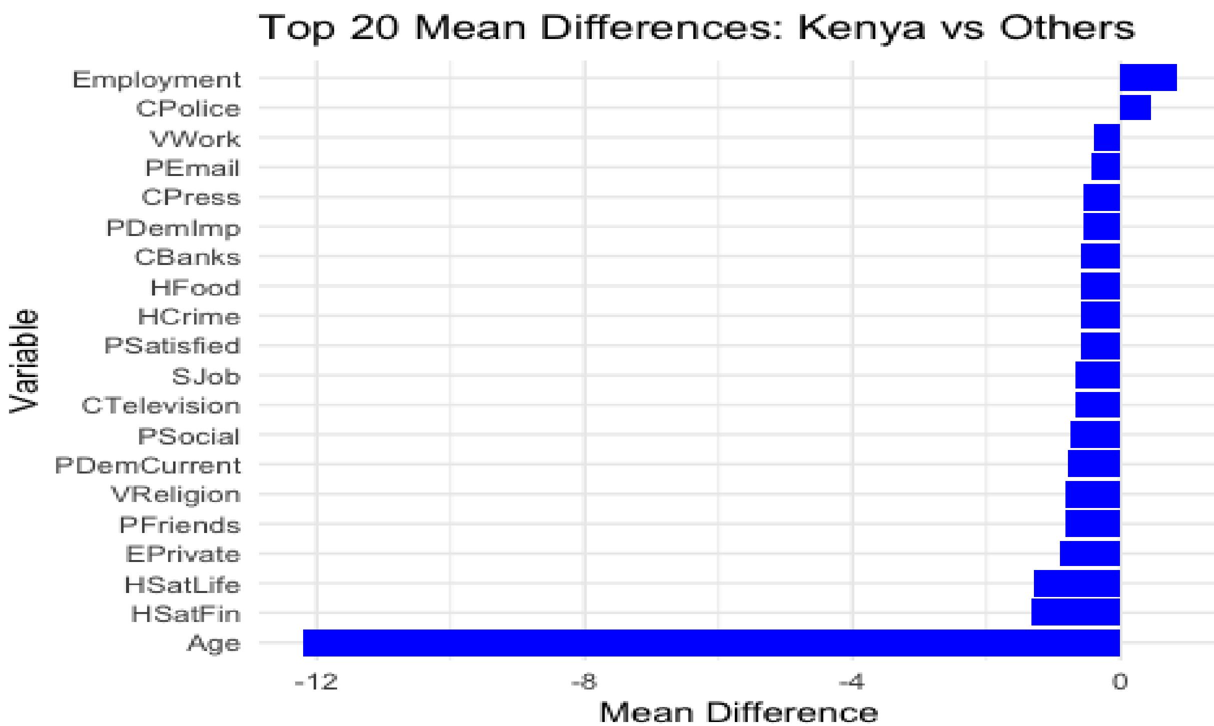
Variety of Non-numerical attributes

The only non-numerical attribute is Country, a character variable that represents country of birth using three-letter codes. Using `table`(`VC$Country`), we can observe that 650 responses are from Kenya.

Question 2a.

Comparing Mean Differences: Kenya vs Others

To explore how participant responses in Kenya differ from other countries, the first approach was to compute the average scores across all variables and plot the top 20 variables with the largest gaps. This analysis helps us to identify areas where Kenya respondents' attitudes or experiences differ from global average.



As seen in the bar plot, one of the most striking differences is in Age. In average, Kenya participants are 12.4 years younger than those from other countries. This large demographic gap likely influences many other attitude differences observed in the data. Also, it is observed that Kenya respondents have lower satisfaction with life `HSatLife` and with finances `HSatFin`, scoring mean differences -0.8 and 0.9 respectively. This shows a lower well-being among Kenyan sample compared to the global sample. Additionally, Kenya has lower engagement with digital communication, reflected through variables like `PEmail` and `PSocial`. Similarly, there is lower average values for `VFriends`, `PFriends`, `EPrivate`, which all indicate a relatively low importance on personal and social connections.

However, there are also few areas where Kenyans scored higher average than other countries. Employment status and Confidence in the police `Cpolice`. A high value for confidence in Police may reflect a strong value of formal institutions. However, it is important to note that mean difference for Employment is positive, so Kenya has a higher average employment value. The employment states falls into 7 scores (1 = Full-time, 7 = unemployed). It is likely that Kenya has a higher proportion of people who are unemployed, students, or not in-full time work, compared to global sample.

```
#run t-tests

ttest_df

##                          Variable p_value
## TPeople                   TPeople  0.0000
## TFamily                   TFamily  0.2953
## TNeighbourhood TNeighbourhood  0.0000
## TKnow                       TKnow  0.0007
```

```
## TMeet                       TMeet  0.0027
## VFamily                   VFamily  0.0000
## VFriends                 VFriends  0.0000
## VWork                       VWork  0.0000
## VReligion               VReligion  0.0000
## HHealth                   HHealth  0.0000
## HSatLife                 HSatLife  0.0000
## HSatFin                   HSatFin  0.0000
## HFood                       HFood  0.0000
## HCrime                     HCrime  0.0000
## EPrivate                 EPrivate  0.0000
## SJob                         SJob  0.0000
## PIA                           PIA  0.0000
## PIAB                         PIAB  0.0083
## STBetter                 STBetter  0.6109
## PEmail                     PEmail  0.0000
## PSocial                   PSocial  0.0000
## PFriends                 PFriends  0.0000
## PDemImp                   PDemImp  0.0000
## PDemCurrent           PDemCurrent  0.0000
## PSatisfied             PSatisfied  0.0000
## MF                             MF  0.4727
## Age                           Age  0.0000
## Edu                           Edu  0.0021
## Employment             Employment  0.0000
## CArmedForces         CArmedForces  0.0000
## CPress                     CPress  0.0000
## CTelevision           CTelevision  0.0000
## CUnions                   CUnions  0.0107
## CPolice                   CPolice  0.0000
## CGovernment           CGovernment  0.0593
## CParliament           CParliament  0.0357
## CUniversities       CUniversities  0.0073
## CMajCompanies       CMajCompanies  0.0000
## CBanks                     CBanks  0.0000
```

Welch's T-test Results

T-tests were conducted to determine whether the differences in attribute scores for Kenya vs. Others were statistically significant. Since this task compare the means between two independent groups, and the dataset includes numeric values, the t-test is an appropriate approach. Welch's t-test is applied because it doesn't assume equal variance between groups and is better for unequal sample sizes.

Out of tested variables, majority showed statistically significant difference ($p<0.05$). Most significant differences ($p<0.001$) include variables like `TPeople`, `TNeighborhood`. There are a few that was not significantly different, such as `TFamily` which has $p=0.2953$. This suggest that views about family are relatively consistent across Kenya and global average.

Small -values should be interpreted alongside effect size because very large samples can generate statistically significant difference that are not actually meaningful.

Question 2b.

```
sorted_results
```

```
##    Confidence_Var R_squared                          Top_Predictors
## 6     CGovernment     0.192 TPeople, VReligion, TNeighbourhood
## 5         CPolice     0.183     TPeople, TKnow, TNeighbourhood
## 4         CUnions     0.170              VFamily, TMeet, TKnow
## 7     CParliament     0.168     VFamily, TNeighbourhood, TMeet
## 1    CArmedForces     0.149        MF, VFamily, TNeighbourhood
## 10         CBanks     0.135         PFriends, VFamily, TFamily
## 9   CMajCompanies     0.131                 TKnow, VFamily, MF
## 8   CUniversities     0.123                 VFamily, TKnow, MF
## 2          CPress     0.089               VFamily, TFamily, MF
## 3     CTelevision     0.052                 MF, TFamily, VWork
```

To assess predicting confidence in social organisms, I fit multiple linear regression models for each confidence variables (filtered by prefix "C"). The predictors include demographic variables, trust-related variables and value measures.

How well do particiapnt responses predict confidence?

Using R-squared value which indicates the proportion of variance in the confidence variable explained by the model, the predictive performance of this survey is tested. As shown in results, the highest R-squared value is 0.192 for Government. It suggests a meaningful ability to predict confidence in government using participant responses. Others like Police, Unions, and Parliament also show a reasonable score for R-squared value. In contrast, variables like Television and Press show a very low R-squared value, meaning that they were poorly predicted by participant attributes. This likely suggest that ideas of media are influenced by some other factors. Overall, the R-squared values range from 0.05 to 0.19, showing some predictive power.

Which attributes are the best predictors?

`VFamily` appeared the most (7 out of 10 top models) among all attributes. It is the most influential and consistent predictor. `TNeighborhood` and `TMeet` also appeared commonly, showing how social connection influence confidence.

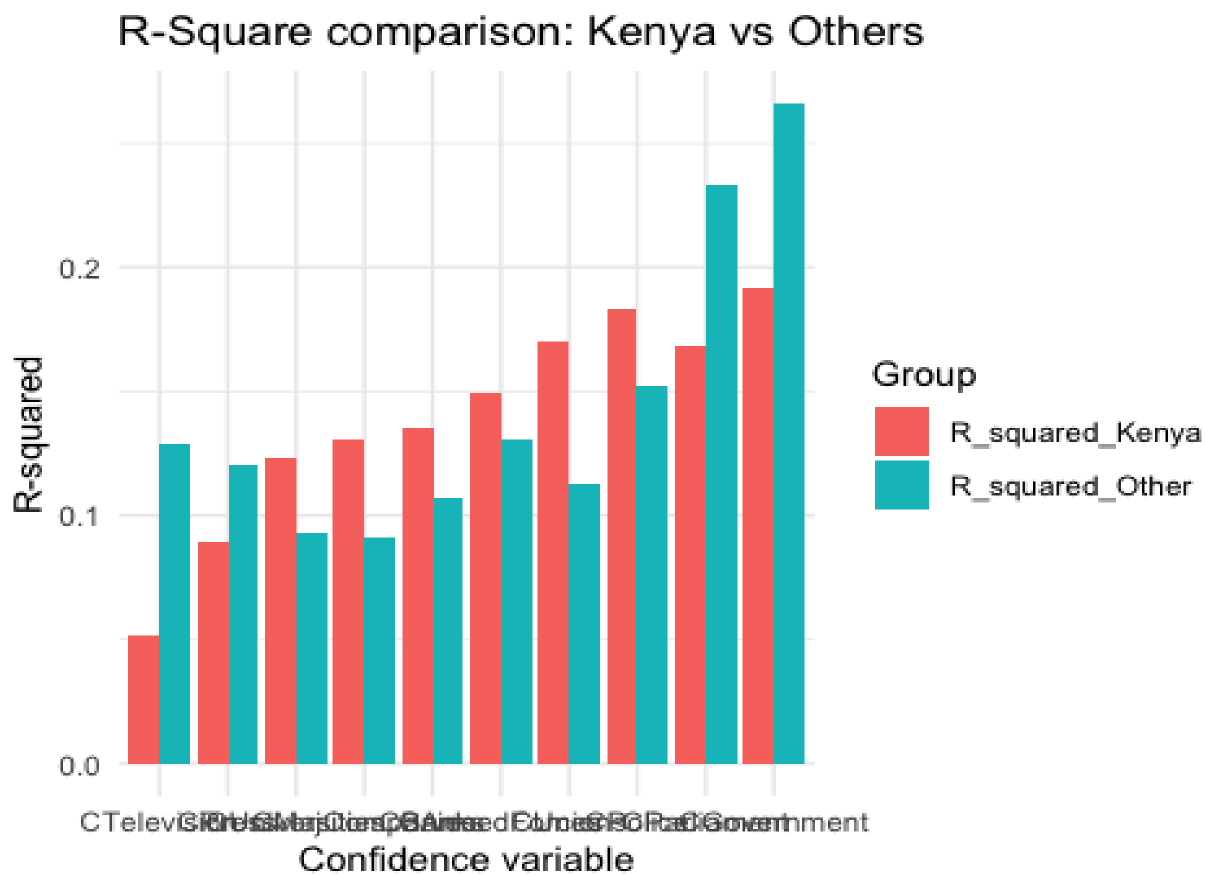
Which confidence variables are more reliably predicted?

In the sorted results, the top rows including `CGovernment`, `CPolice`, `CUnions`, `CParliament` are the most reliable predicted social organizations.

Question 2c.

```
sorted_results_other

##    Confidence_Var R_squared                            Top_Predictors
## 6     CGovernment     0.266 PSatisfied, TNeighbourhood, TPeople
## 7     CParliament     0.233 TPeople, TNeighbourhood, PSatisfied
## 5         CPolice     0.152    TFamily, TPeople, TNeighbourhood
## 1     CArmedForces    0.131  TFamily, TNeighbourhood, VReligion
## 3     CTelevision     0.129    TNeighbourhood, TFamily, TPeople
## 2          CPress     0.120      TNeighbourhood, TPeople, TMeet
## 4         CUnions     0.113      TNeighbourhood, TPeople, TMeet
## 10         CBanks     0.107      TNeighbourhood, TKnow, TPeople
## 8   CUniversities     0.093               TFamily, TKnow, VWork
## 9   CMajCompanies     0.091      TMeet, TNeighbourhood, TPeople
```



When the same method is applied on all other countries, we can observe the results as shown above. The predictive power of participant responses is slightly stronger globally when compared to that of Kenya. The top model is Government, with R^2 value of 0.266, whereas it is 0.192 for Kenya. R^2 for Parliament is 0.233 globally, whereas Kenya is 0.168. `TPeople` and `TNeighborhood` were most frequent predictors. Political Satisfaction `PSatisfied` is one top predictor for other countries, but was not one of top predictors in Kenya. `VFamiy` and `TMeet` also appeared commonly, similar to Kenya. There is some overlap in important predictors like trust and social values, but Kenya is more influenced by family and religion.

Question 3a.

Data Set Reference

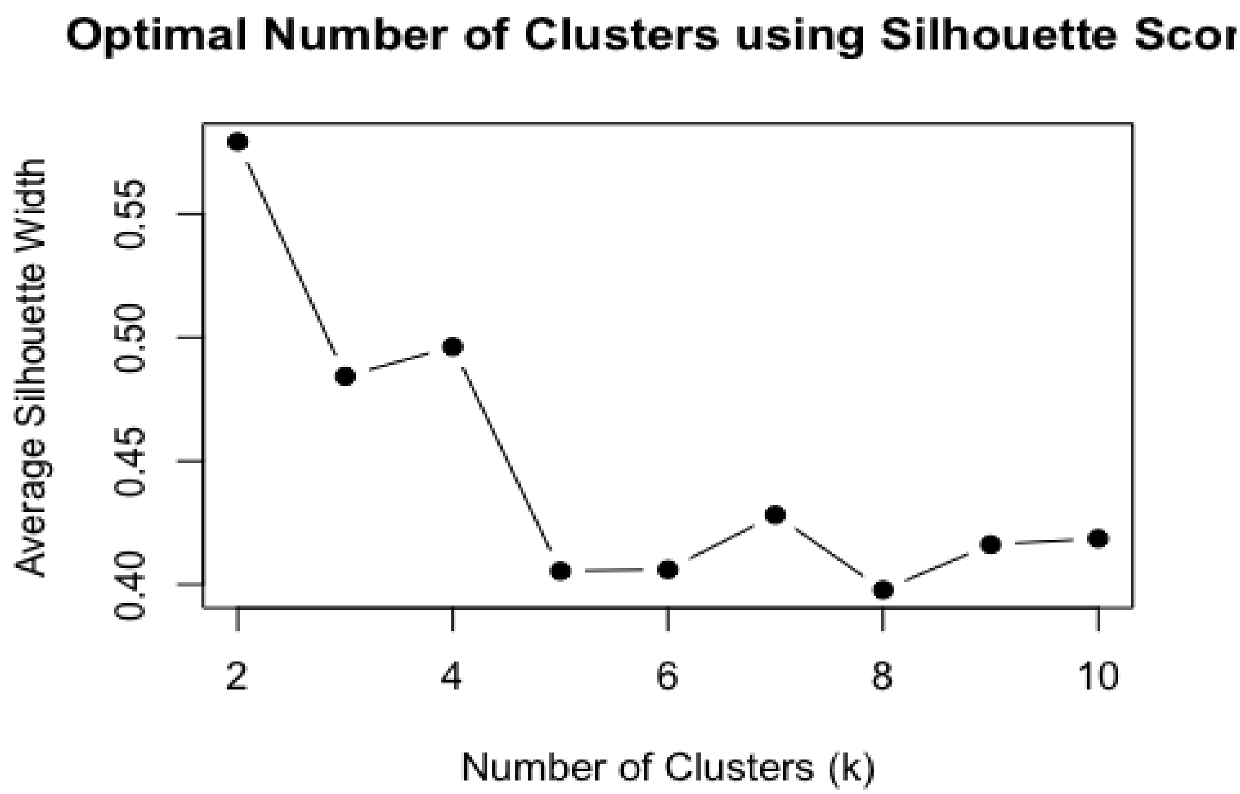
Political Stability https://data.worldbank.org/indicator/PV.PER.RNK

GDP per Capita https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

Life Expectancy https://data.worldbank.org/indicator/SP.DYN.LE00.IN

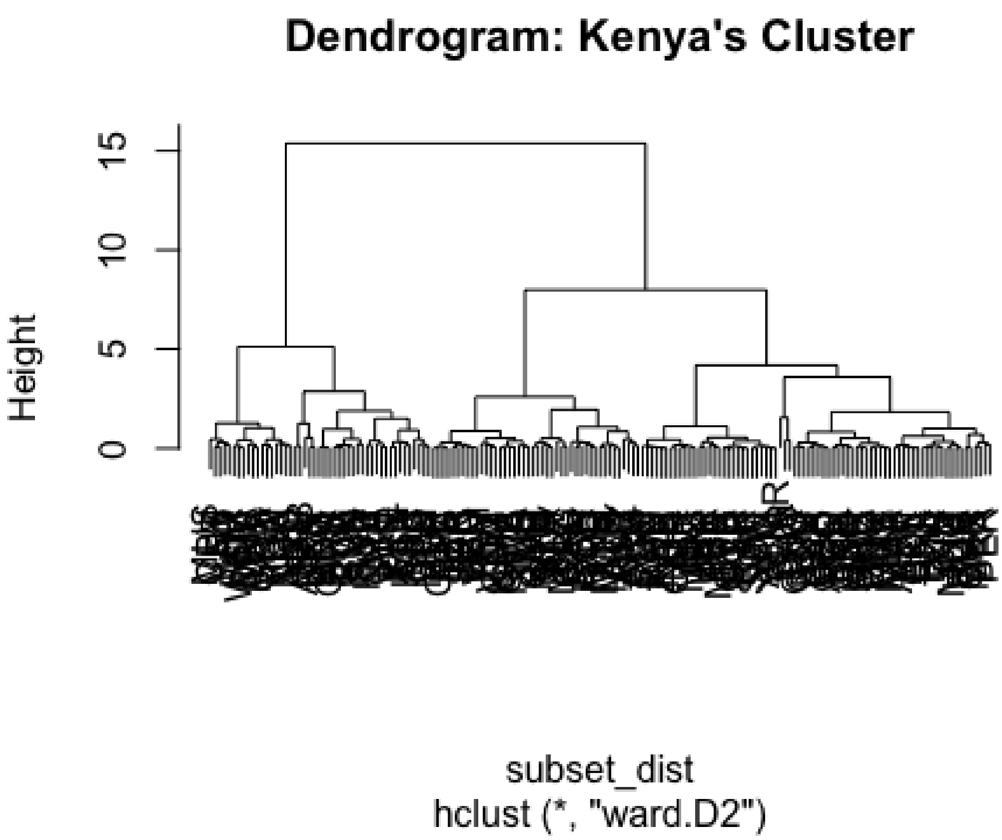
Three country-level datasets are chosen from World Bank (year 2023 filtered). To identify countries most similar to Kenya, hierarchial clustering analysis based on a combination of social, economic, political indicators will be performed. GDP per capita measures economic development. Life expectancy reflects public health. Political stability represents governance quality.

The datasets were merged by country code and missing values are cleaned.



Since the silhouette score peaks at k=2, it is the best clustering solution to divide countries into 2 broad clusters. Then, Ward's hierarchical clustering is used to separate countries into different groups.

```
##    Country.Code    X2023.x    X2023.y     X2023 Cluster
## 1           ABW 33984.7906 33984.7906 97.630333       1
## 3           AFG   415.7074   415.7074  1.421801       1
## 5           AGO  2308.1598  2308.1598 32.227489       1
## 6           ALB  8575.1711  8575.1711 51.658768       1
## 7           AND 46812.4484 46812.4484 98.578201       1
## 10          ARG 14187.4827 14187.4827 41.706161       1
```



Kenya was grouped into Cluster 1, which includes countries that tend to have lower GDP per capita, moderate life expectancy, and less political stability. Aruba, Afghanistan, Angola, Albania, Angola are countries that are similar to Kenya.

Question 3b.

```
results_cluster
```

```
##    Confidence_Var R_squared                        Top_Predictors
## 6     CGovernment     0.255 TPeople, TNeighbourhood, PSatisfied
## 7     CParliament     0.226 TPeople, TNeighbourhood, PSatisfied
## 5         CPolice     0.145    TPeople, TFamily, TNeighbourhood
## 1    CArmedForces     0.135    TFamily, TPeople, TNeighbourhood
## 3     CTelevision     0.127    TNeighbourhood, TPeople, TFamily
## 2          CPress     0.123      TNeighbourhood, TPeople, TMeet
```

```
## 4        CUnions     0.121      TPeople, TNeighbourhood, TMeet
## 10        CBanks     0.109      TPeople, TNeighbourhood, TKnow
## 9   CMajCompanies     0.088               TPeople, TMeet, TKnow
## 8   CUniversities     0.083               TKnow, TFamily, VWork
```

To observe how well participant-level attributes predict confidence in social organizations within the cluster of Kenya, the same regression modeling method done in Question 2 is repeated.

For each confidence variable, a multiple linear regression model was fir using different predictors.

Overall Predictability

`CGovernment` is most predictable variable, having R-squared value of 0.255. Next is `CParliament` has 0.226, and `CPolice` has 0.145.

These values are higher than those for Kenya alone, where `CGovernment` had R^2= 0.192 (Q2b), and slightly lower than the global results reaching $R^2 = 0.266$ (Q2c).

Strongest Predictors

The most influential predictors across models in this cluster include trust in others, trust is neighborhood. Also, political satisfaction plays important roles. This is very similar to that of the global patterns (Q2c). While for Kenya, Family and Religion values had stronger influence.

Comparison: Cluster vs Kenya vs Other Countries

Comparison group	Best R^2	Top Predictors	Overall Fit
Kenya (Question2b)	0.192	`VFamily, TPEople, VReligion`	Moderate
Others (Question2c)	0.266	`TPeople, TNeighborhood, PSatisfied`	Strong
Cluster (Question3b)	0.255	`TPeople, TNeighborhood, PSatisfied`	Strong

For overall predictive strength, the cluster group outperforms the Kenya model. It is also shown that the cluster group shows high similarity to Kenya's predictor pattern. Thus, the cluster model is likely a better match for Kenya than the other-countries model because it is based on countries with more comparable socio-political and economic perceptions.

Appendix

```
#Data Setup
rm(list = ls())
set.seed(35865377)
VCData = read.csv("WVSExtract.csv")
VC = VCData[sample(1:nrow(VCData),50000, replace=FALSE),] #sample 50,000
respondents
VC = VC[,c(1:6, sort(sample(7:46,17, replace = FALSE)), 47:53,
sort(sample(54:69,10, replace = FALSE)))]
```

Question 1

```
#Q1 observe data
dim(VC)

## [1] 50000    40

str(VC)

## 'data.frame':    50000 obs. of  40 variables:
##  $ Country       : chr  "UKR" "MNG" "ROU" "MDV" ...
##  $ TPeople       : int  2 2 2 2 2 1 2 2 2 2 ...
##  $ TFamily       : int  1 1 3 2 1 1 1 1 1 1 ...
##  $ TNeighbourhood: int  1 1 4 3 3 1 1 2 3 2 ...
##  $ TKnow         : int  1 2 3 2 2 1 2 2 3 3 ...
##  $ TMeet         : int  1 2 4 4 4 2 4 2 3 4 ...
##  $ VFamily       : int  2 1 2 1 1 1 1 1 1 1 ...
##  $ VFriends      : int  3 1 3 1 2 2 1 1 4 1 ...
##  $ VWork         : int  2 3 2 2 1 1 1 1 1 1 ...
##  $ VReligion     : int  2 4 4 1 1 1 2 1 2 1 ...
##  $ HHealth       : int  3 2 3 3 2 3 1 3 1 1 ...
##  $ HSatLife      : int  -1 6 4 4 1 7 5 -1 3 6 ...
##  $ HSatFin       : int  2 5 4 4 1 7 8 -1 7 4 ...
##  $ HFood         : int  3 4 2 3 4 4 4 1 1 4 ...
##  $ HCrime        : int  4 4 4 4 3 4 4 2 2 4 ...
##  $ EPrivate      : int  -2 5 8 4 7 5 2 -1 5 7 ...
##  $ SJob          : int  1 1 1 3 1 4 4 1 2 3 ...
##  $ PIA           : int  4 1 1 2 2 2 2 1 2 1 ...
##  $ PIAB          : int  1 3 4 3 1 4 4 2 1 3 ...
##  $ STBetter      : int  10 8 8 7 8 6 1 6 4 10 ...
##  $ PEmail        : int  5 5 5 1 5 1 5 1 5 5 ...
##  $ PSocial       : int  1 2 5 1 5 1 1 1 2 2 ...
##  $ PFriends      : int  1 2 4 2 3 1 1 1 4 4 ...
##  $ PDemImp       : int  10 8 9 4 10 8 7 5 6 7 ...
##  $ PDemCurrent   : int  3 5 4 4 5 7 5 6 6 7 ...
##  $ PSatisfied    : int  1 3 7 2 3 8 5 -1 6 5 ...
##  $ MF            : int  1 1 1 1 1 1 2 1 2 2 ...
##  $ Age           : int  33 32 54 28 55 18 38 33 24 22 ...
##  $ Edu           : int  4 3 3 2 4 3 2 3 6 3 ...
```

```
##  $ Employment    : int  3 1 1 7 7 7 5 1 6 6 ...
##  $ CArmedForces  : int  1 2 2 3 2 2 3 1 1 1 ...
##  $ CPress        : int  3 2 4 4 4 3 2 1 2 3 ...
##  $ CTelevision   : int  3 2 3 4 4 3 2 2 2 3 ...
##  $ CUnions       : int  4 2 3 4 4 3 2 3 1 3 ...
##  $ CPolice       : int  2 2 2 3 2 2 2 1 1 2 ...
##  $ CGovernment   : int  -1 4 2 4 3 2 1 1 1 3 ...
##  $ CParliament   : int  -1 4 2 4 2 3 1 2 1 3 ...
##  $ CUniversities : int  -1 2 -1 3 2 1 1 1 4 2 ...
##  $ CMajCompanies : int  -1 2 -1 4 3 2 1 2 4 3 ...
##  $ CBanks        : int  3 2 4 2 3 2 1 1 1 3 ...
```

```
summary(VC)
```

```
##    Country             TPeople          TFamily        TNeighbourhood
##  Length:50000       Min.   :-5.000   Min.   :-5.000   Min.   :-5.000
##  Class :character   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 2.000
##  Mode  :character   Median : 2.000   Median : 1.000   Median : 2.000
##                     Mean   : 1.708   Mean   : 1.263   Mean   : 2.155
##                     3rd Qu.: 2.000   3rd Qu.: 1.000   3rd Qu.: 3.000
##                     Max.   : 2.000   Max.   : 4.000   Max.   : 4.000
##      TKnow            TMeet           VFamily          VFriends
##  Min.   :-5.000   Min.   :-5.000   Min.   :-5.000   Min.   :-5.000
##  1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 2.000   Median : 3.000   Median : 1.000   Median : 2.000
##  Mean   : 2.043   Mean   : 2.953   Mean   : 1.108   Mean   : 1.697
##  3rd Qu.: 2.000   3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.: 2.000
##  Max.   : 4.000   Max.   : 4.000   Max.   : 4.000   Max.   : 4.000
##      VWork          VReligion         HHealth          HSatLife
##  Min.   :-5.000   Min.   :-5.000   Min.   :-5.000   Min.   :-5.000
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 6.000
##  Median : 1.000   Median : 2.000   Median : 2.000   Median : 7.000
##  Mean   : 1.508   Mean   : 1.945   Mean   : 2.175   Mean   : 7.013
##  3rd Qu.: 2.000   3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 9.000
##  Max.   : 4.000   Max.   : 4.000   Max.   : 5.000   Max.   :10.000
##     HSatFin           HFood            HCrime          EPrivate
##  Min.   :-5.000   Min.   :-5.000   Min.   :-5.000   Min.   :-5.000
##  1st Qu.: 5.000   1st Qu.: 3.000   1st Qu.: 3.000   1st Qu.: 3.000
##  Median : 6.000   Median : 4.000   Median : 4.000   Median : 5.000
##  Mean   : 6.148   Mean   : 3.458   Mean   : 3.417   Mean   : 5.408
##  3rd Qu.: 8.000   3rd Qu.: 4.000   3rd Qu.: 4.000   3rd Qu.: 8.000
##  Max.   :10.000   Max.   : 4.000   Max.   : 4.000   Max.   :10.000
##       SJob             PIA              PIAB           STBetter
##  Min.   :-5.000   Min.   :-5.000   Min.   :-5.000   Min.   :-5.000
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 6.000
##  Median : 2.000   Median : 2.000   Median : 2.000   Median : 8.000
##  Mean   : 1.946   Mean   : 1.809   Mean   : 2.173   Mean   : 7.182
##  3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.:10.000
##  Max.   : 4.000   Max.   : 4.000   Max.   : 4.000   Max.   :10.000
##      PEmail          PSocial          PFriends         PDemImp
```

```
##  Min.   :-5.000   Min.   :-5.000   Min.   :-5.00   Min.   :-5.000  
##  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.00   1st Qu.: 7.000  
##  Median : 4.000   Median : 2.000   Median : 2.00   Median : 9.000  
##  Mean   : 3.386   Mean   : 2.446   Mean   : 2.28   Mean   : 8.195  
##  3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 3.00   3rd Qu.:10.000  
##  Max.   : 5.000   Max.   : 5.000   Max.   : 5.00   Max.   :10.000  
##   PDemCurrent       PSatisfied           MF              Age        
##  Min.   :-5.000   Min.   :-5.000   Min.   :-5.000   Min.   : -5.00  
##  1st Qu.: 4.000   1st Qu.: 3.000   1st Qu.: 1.000   1st Qu.: 29.00  
##  Median : 6.000   Median : 5.000   Median : 2.000   Median : 41.00  
##  Mean   : 5.955   Mean   : 5.049   Mean   : 1.519   Mean   : 42.84  
##  3rd Qu.: 8.000   3rd Qu.: 7.000   3rd Qu.: 2.000   3rd Qu.: 55.00  
##  Max.   :10.000   Max.   :10.000   Max.   : 2.000   Max.   :103.00  
##       Edu          Employment     CArmedForces        CPress      
##  Min.   :-5.000   Min.   :-5.00   Min.   :-5.000   Min.   :-5.000  
##  1st Qu.: 2.000   1st Qu.: 1.00   1st Qu.: 1.000   1st Qu.: 2.000  
##  Median : 3.000   Median : 3.00   Median : 2.000   Median : 3.000  
##  Mean   : 3.505   Mean   : 3.05   Mean   : 1.896   Mean   : 2.611  
##  3rd Qu.: 6.000   3rd Qu.: 5.00   3rd Qu.: 3.000   3rd Qu.: 3.000  
##  Max.   : 8.000   Max.   : 8.00   Max.   : 4.000   Max.   : 4.000  
##   CTelevision        CUnions          CPolice        CGovernment    
##  Min.   :-5.000   Min.   :-5.000   Min.   :-5.000   Min.   :-5.000  
##  1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000  
##  Median : 3.000   Median : 3.000   Median : 2.000   Median : 3.000  
##  Mean   : 2.569   Mean   : 2.418   Mean   : 2.221   Mean   : 2.454  
##  3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 3.000  
##  Max.   : 4.000   Max.   : 4.000   Max.   : 4.000   Max.   : 4.000  
##   CParliament     CUniversities    CMajCompanies        CBanks     
##  Min.   :-5.000   Min.   :-5.000   Min.   :-5.000   Min.   :-5.00  
##  1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.00  
##  Median : 3.000   Median : 2.000   Median : 3.000   Median : 2.00  
##  Mean   : 2.652   Mean   : 2.011   Mean   : 2.401   Mean   : 2.31  
##  3rd Qu.: 4.000   3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.: 3.00  
##  Max.   : 4.000   Max.   : 4.000   Max.   : 4.000   Max.   : 4.00
```

```r
#check missing values
VC[VC<0] <- NA
colSums(is.na(VC))
```

```
##        Country        TPeople        TFamily TNeighbourhood          TKnow 
##              0            668            144            399            277 
##          TMeet        VFamily       VFriends          VWork      VReligion 
##            697             83            173            591            481 
##        HHealth       HSatLife        HSatFin          HFood         HCrime 
##            132            266            326            273            297 
##       EPrivate           SJob            PIA           PIAB       STBetter 
##           1783           2264           1343           2526           1301 
##         PEmail        PSocial       PFriends        PDemImp    PDemCurrent 
##           1021           1881            556            925           1402 
##     PSatisfied             MF            Age            Edu     Employment
```

```
##          1898          52         244         515         598
##   CArmedForces      CPress  CTelevision     CUnions     CPolice
##          2201        1115         749        3667        1237
##    CGovernment CParliament CUniversities CMajCompanies    CBanks
##          1618        1793        2011        2722        1447
```

```
sum(is.na(VC))
```

```
## [1] 41676
```

```
table(VC$Country)
```

```
##
##  AND  ARG  ARM  AUS  BGD  BOL  BRA  CAN  CHL  CHN  COL  CYP  CZE  DEU  ECU
EGY
##  527  517  641  916  627 1050  890 2055  514 1573  782  537  605  754  623
636
##  ETH  GBR  GRC  GTM  HKG  IDN  IND  IRN  IRQ  JOR  JPN  KAZ  KEN  KGZ  KOR
LBN
##  646 1384  603  617 1085 1665  903  787  605  588  722  672  650  622  645
597
##  LBY  MAC  MAR  MDV  MEX  MMR  MNG  MYS  NGA  NIC  NLD  NZL  PAK  PER  PHL
PRI
##  626  508  625  532  929  617  843  673  636  642 1091  550 1075  727  608
610
##  ROU  RUS  SGP  SRB  SVK  THA  TJK  TUN  TUR  TWN  UKR  URY  USA  UZB  VEN
VNM
##  635  912 1040  524  639  778  620  616 1282  624  667  505 1320  653  617
620
##  ZWE
##  618
```

Question 2a.

```
#create new variable to label kenya or other
VC$group <- ifelse(VC$Country == "KEN", "Kenya", "Other")
#create new data frame and filter columns (exclude "Country" and "group")
predictors <- VC[, !(names(VC) %in% c("Country", "group"))]
#compute means
group_means <- VC |>
  group_by(group) |>
  summarise(across(where(is.numeric), ~mean(.x, na.rm = TRUE)))
#compte difference
kenya_means <- group_means[group_means$group == "Kenya", -1]
other_means <- group_means[group_means$group == "Other", -1]
mean_diff <- as.numeric(kenya_means - other_means)
diff_df <- data.frame(
  Variable = colnames(kenya_means),
  Kenya_Mean = as.numeric(kenya_means),
  Other_Mean = as.numeric(other_means),
  Mean_Difference = round(mean_diff,3)
```

```
)
diff_df <- diff_df[order(abs(diff_df$Mean_Difference), decreasing = TRUE),]
#order

#barplot difference in means
top_diffs <- head(diff_df, 20)

ggplot(top_diffs, aes(x=reorder(Variable, Mean_Difference),
y=Mean_Difference))+
  geom_bar(stat="identity", fill="blue") +
  coord_flip() +
  labs(title = "Top 20 Mean Differences: Kenya vs Others",
       y = "Mean Difference",
       x = "Variable") +
  theme_minimal()
```

```
#run t-tests
ttest_results <- sapply(names(predictors), function(var) {
  t.test(VC[[var]] ~ VC$group)$p.value
})
ttest_results

##        TPeople        TFamily TNeighbourhood          TKnow          TMeet
##   3.373614e-29   2.953042e-01   3.692575e-06   7.005891e-04   2.705435e-03
##        VFamily       VFriends          VWork      VReligion        HHealth
##   2.105460e-23   1.094339e-06   3.038567e-82  6.163552e-173   2.549262e-15
##       HSatLife        HSatFin          HFood         HCrime       EPrivate
##   7.221659e-31   5.614963e-31   6.231559e-42   1.980443e-46   2.481736e-13
##           SJob            PIA           PIAB        STBetter         PEmail
##   1.526503e-60   5.685741e-20   8.295969e-03   6.109033e-01   2.960523e-12
##        PSocial       PFriends        PDemImp    PDemCurrent     PSatisfied
##   6.890368e-33   1.152085e-79   1.569966e-07   4.374608e-11   7.408817e-08
##             MF            Age            Edu     Employment   CArmedForces
##   4.727282e-01  2.814376e-127   2.050486e-03   5.232745e-21   2.743558e-07
##         CPress    CTelevision        CUnions        CPolice    CGovernment
##   5.447180e-43   1.529636e-64   1.067179e-02   1.645045e-27   5.926406e-02
##    CParliament  CUniversities  CMajCompanies         CBanks
##   3.570012e-02   7.318697e-03   1.726640e-07   1.637073e-48

#convert result to dataframe
ttest_df <- data.frame(
  Variable = names(ttest_results),
  p_value = round(ttest_results, 4)
)
```

Question 2b.

```r
#filter kenya only
VC_Kenya <- VC[VC$Country == "KEN", ]
#confidence in social organizations (starting with C)
conf_vars <- names(VC_Kenya)[grepl("^C", names(VC_Kenya))]
#predictor variables exclude country, group conf_var
predictors <- names (VC_Kenya)[!(names(VC_Kenya) %in% c("Country", "group",
conf_vars))]

results <- data.frame(Confidence_Var = character(),
                      R_squared = numeric(),
                      Top_Predictors = character(),
                      stringsAsFactors = FALSE)

for (conf_var in conf_vars) {

  formula <- as.formula(paste(conf_var, "~", paste(predictors, collapse =
"+")))

  model_data <- VC_Kenya[, c(conf_var, predictors)]

  # Make sure response is numeric
  model_data[[conf_var]] <- as.numeric(model_data[[conf_var]])

  # Remove rows with any NAs
  model_data <- na.omit(model_data)

#in case data is too small
  if (nrow(model_data) < 30) next

  #clean data, remove NA
  y <- model_data[[conf_var]]
  if (any(is.na(y)) || any(is.nan(y)) || any(is.infinite(y))) next

  # Fit the model
  model <- lm(formula, data = model_data)

  #find R^2
  r2 <- summary(model)$r.squared

  # Get top 3 predictors
  coefs <- summary(model)$coefficients[-1, "Estimate"]
  top_preds <- names(sort(abs(coefs), decreasing = TRUE))[1:3]

  results <- rbind(results, data.frame(
    Confidence_Var = conf_var,
    R_squared = round(r2, 3),
    Top_Predictors = paste(top_preds, collapse = ", ")
```

```
  ))
}

## Warning: NAs introduced by coercion

head(results)

##   Confidence_Var R_squared                         Top_Predictors
## 1   CArmedForces     0.149        MF, VFamily, TNeighbourhood
## 2         CPress     0.089               VFamily, TFamily, MF
## 3    CTelevision     0.052                 MF, TFamily, VWork
## 4        CUnions     0.170              VFamily, TMeet, TKnow
## 5        CPolice     0.183     TPeople, TKnow, TNeighbourhood
## 6    CGovernment     0.192 TPeople, VReligion, TNeighbourhood

sorted_results <- results[order(-results$R_squared),]
sorted_results

##    Confidence_Var R_squared                         Top_Predictors
## 6     CGovernment     0.192 TPeople, VReligion, TNeighbourhood
## 5         CPolice     0.183     TPeople, TKnow, TNeighbourhood
## 4         CUnions     0.170              VFamily, TMeet, TKnow
## 7     CParliament     0.168     VFamily, TNeighbourhood, TMeet
## 1    CArmedForces     0.149        MF, VFamily, TNeighbourhood
## 10         CBanks     0.135          PFriends, VFamily, TFamily
## 9   CMajCompanies     0.131                 TKnow, VFamily, MF
## 8   CUniversities     0.123                 VFamily, TKnow, MF
## 2          CPress     0.089               VFamily, TFamily, MF
## 3     CTelevision     0.052                 MF, TFamily, VWork
```

Question 2c.

```
#filter non-kenya
VC_Other <- VC[VC$Country != "KEN", ]
#confidence in social organizations (starting with C)
conf_vars <- names(VC_Other)[grepl("^C", names(VC_Other))]
#predictor variables exclude country, group conf_var
predictors <- names (VC_Other)[!(names(VC_Other) %in% c("Country", "group",
conf_vars))]

results_other <- data.frame(Confidence_Var = character(),
                            R_squared = numeric(),
                            Top_Predictors = character(),
                            stringsAsFactors = FALSE)

#repeat 2b
for (conf_var in conf_vars) {
```

```r
  formula <- as.formula(paste(conf_var, "~", paste(predictors, collapse =
"+")))

  model_data <- VC_Other[, c(conf_var, predictors)]

  # Make sure response is numeric
  model_data[[conf_var]] <- as.numeric(model_data[[conf_var]])

  # Remove rows with any NAs
  model_data <- na.omit(model_data)

#in case data is too small
  if (nrow(model_data) < 30) next

  #clean data, remove NA
  y <- model_data[[conf_var]]
  if (any(is.na(y)) || any(is.nan(y)) || any(is.infinite(y))) next

  # Fit the model
  model <- lm(formula, data = model_data)

  #find R^2
  r2 <- summary(model)$r.squared

  # Get top 3 predictors
  coefs <- summary(model)$coefficients[-1, "Estimate"]
  top_preds <- names(sort(abs(coefs), decreasing = TRUE))[1:3]

  results_other <- rbind(results_other, data.frame(
    Confidence_Var = conf_var,
    R_squared = round(r2, 3),
    Top_Predictors = paste(top_preds, collapse = ", ")
  ))
}

## Warning: NAs introduced by coercion

head(results_other)

##   Confidence_Var R_squared                     Top_Predictors
## 1   CArmedForces     0.131  TFamily, TNeighbourhood, VReligion
## 2         CPress     0.120      TNeighbourhood, TPeople, TMeet
## 3    CTelevision     0.129    TNeighbourhood, TFamily, TPeople
## 4        CUnions     0.113      TNeighbourhood, TPeople, TMeet
## 5        CPolice     0.152    TFamily, TPeople, TNeighbourhood
## 6    CGovernment     0.266 PSatisfied, TNeighbourhood, TPeople

sorted_results_other <- results_other[order(-results_other$R_squared),]
sorted_results_other
```

```
##    Confidence_Var R_squared                          Top_Predictors
## 6     CGovernment     0.266 PSatisfied, TNeighbourhood, TPeople
## 7     CParliament     0.233 TPeople, TNeighbourhood, PSatisfied
## 5         CPolice     0.152    TFamily, TPeople, TNeighbourhood
## 1    CArmedForces     0.131  TFamily, TNeighbourhood, VReligion
## 3     CTelevision     0.129    TNeighbourhood, TFamily, TPeople
## 2          CPress     0.120      TNeighbourhood, TPeople, TMeet
## 4         CUnions     0.113      TNeighbourhood, TPeople, TMeet
## 10         CBanks     0.107      TNeighbourhood, TKnow, TPeople
## 8   CUniversities     0.093               TFamily, TKnow, VWork
## 9   CMajCompanies     0.091      TMeet, TNeighbourhood, TPeople
```

```
#visualize 2c

merge_results <- merge(results, results_other, by = "Confidence_Var",
suffixes = c("_Kenya", "_Other"))

#make into long format
long <- merge_results |>
  select(Confidence_Var, R_squared_Kenya, R_squared_Other) |>
  pivot_longer(cols = starts_with("R_squared"),
               names_to = "Group",
               values_to = "R_squared")

#plot
ggplot(long, aes(x=reorder(Confidence_Var, R_squared), y = R_squared, fill =
Group)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "R-Square comparison: Kenya vs Others",
       x = "Confidence variable",
       y = "R-squared") + theme_minimal()
```

Question 3a.

```
#Extractpolitical stability
politics = read.csv("PoliticalStability.csv", skip = 4)

pol_2023 <- politics[,c("Country.Code","X2023")]

#Extract GDP per capita
GDP = read.csv("GDPperCapita.csv", skip = 4)

gdp_2023 <- GDP[,c("Country.Code","X2023")]


#Extract Life expectancy
Life = read.csv("LifeExpectancy.csv", skip=4)
```

```r
life_2023 <- Life[,c("Country.Code","X2023")]


#Merge
indicators <- merge(gdp_2023, life_2023, by = "Country.Code")
indicators <- merge(indicators, pol_2023, by = "Country.Code")
indicators <- na.omit(indicators)

scaled <- scale(indicators[,-1])
rownames(scaled) <- indicators$Country.Code
dist_matrix <- dist(scaled)

#silhouette for k=2 to 10
avg_sil_width <- numeric(9)
for (k in 2:10) {
  clusters <- cutree(hclust(dist_matrix, method = "ward.D2"), k)
  sil <- silhouette(clusters, dist_matrix)
  avg_sil_width[k - 1] <- mean(sil[, 3])
}

plot(2:10, avg_sil_width, type = "b", pch = 19,
     xlab = "Number of Clusters (k)",
     ylab = "Average Silhouette Width",
     main = "Optimal Number of Clusters using Silhouette Score")

#hierarchical clustering
hier <- hclust(dist_matrix, method = "ward.D2")
clusters<-cutree(hier, k=2)
indicators$Cluster <- clusters

# find just for countries in Kenya's cluster
kenya_cluster <- indicators$Cluster[indicators$Country.Code == "KEN"]
similar <- indicators[indicators$Cluster == kenya_cluster, ]
head(similar)

##    Country.Code    X2023.x    X2023.y     X2023 Cluster
## 1           ABW 33984.7906 33984.7906 97.630333       1
## 3           AFG   415.7074   415.7074  1.421801       1
## 5           AGO  2308.1598  2308.1598 32.227489       1
## 6           ALB  8575.1711  8575.1711 51.658768       1
## 7           AND 46812.4484 46812.4484 98.578201       1
## 10          ARG 14187.4827 14187.4827 41.706161       1

# Redo clustering only on that subset
subset_scaled <- scaled[indicators$Cluster == kenya_cluster, ]
subset_dist <- dist(subset_scaled)
subset_hc <- hclust(subset_dist, method = "ward.D2")
plot(subset_hc, labels = rownames(subset_scaled), main = "Dendrogram: Kenya's
Cluster")
```

Question 3b.

```r
cluster_countries <- similar$Country.Code

VC_cluster <- VC[VC$Country %in% cluster_countries, ]

conf_vars <- names(VC_cluster)[grepl("^C", names(VC_cluster))]
predictors <- names(VC_cluster)[!(names(VC_cluster) %in% c("Country",
"group", conf_vars))]

#run regression
results_cluster <- data.frame(Confidence_Var = character(),
                              R_squared = numeric(),
                              Top_Predictors = character(),
                              stringsAsFactors = FALSE)

for (conf_var in conf_vars) {

  formula <- as.formula(paste(conf_var, "~", paste(predictors, collapse =
"+")))
  model_data <- VC_cluster[, c(conf_var, predictors)]
  model_data[[conf_var]] <- as.numeric(model_data[[conf_var]])
  model_data <- na.omit(model_data)

  if (nrow(model_data) < 30) next
  y <- model_data[[conf_var]]
  if (any(is.na(y)) || any(is.nan(y)) || any(is.infinite(y))) next

  model <- lm(formula, data = model_data)
  r2 <- summary(model)$r.squared
  coefs <- summary(model)$coefficients[-1, "Estimate"]
  top_preds <- names(sort(abs(coefs), decreasing = TRUE))[1:3]

  results_cluster <- rbind(results_cluster, data.frame(
    Confidence_Var = conf_var,
    R_squared = round(r2, 3),
    Top_Predictors = paste(top_preds, collapse = ", ")
  ))
}

## Warning: NAs introduced by coercion

results_cluster <- results_cluster[order(-results_cluster$R_squared), ]
results_cluster

##    Confidence_Var R_squared                        Top_Predictors
## 6     CGovernment     0.255 TPeople, TNeighbourhood, PSatisfied
## 7     CParliament     0.226 TPeople, TNeighbourhood, PSatisfied
```

```
## 5          CPolice     0.145    TPeople, TFamily, TNeighbourhood
## 1     CArmedForces     0.135    TFamily, TPeople, TNeighbourhood
## 3      CTelevision     0.127    TNeighbourhood, TPeople, TFamily
## 2           CPress     0.123      TNeighbourhood, TPeople, TMeet
## 4          CUnions     0.121      TPeople, TNeighbourhood, TMeet
## 10          CBanks     0.109      TPeople, TNeighbourhood, TKnow
## 9   CMajCompanies     0.088               TPeople, TMeet, TKnow
## 8   CUniversities     0.083               TKnow, TFamily, VWork
```