

Assessment

Applied Statistics and Data Visualisation

MSc Data Science



University of  
**Salford**  
MANCHESTER

Research Title

**DETERMINANTS OF LIFE SPAN ( LIFE EXPECTANCY): A COMPREHENSIVE ANALYSIS ACROSS  
GLOBAL REGIONS**

AMADI EMMANUEL OTUDE

@00705726

6/12/2023

## TABLE OF CONTENTS

### Part One: Statistical Analysis

- 1.1 Introduction
- 1.2 Background research
- 2. Exploration of data set
- 2.1 Results and analysis
- 2.2 Comprehensive descriptive statistical analysis
- 2.3 Correlation analysis
- 2.4 Regression analysis
- 3. Time series analysis
- 3.1 Hypothesis testing
- 3.2 Checking normality of data
- 3.3 Transform for normality
- 4. Hypothesis 1: Kruskal Walls Test
- 4.2 Hypothesis 2: Mann-Witney: Wilcoxon test of Hypothesis
- 4.3 Discussion
- 4.4 Conclusion

### Part Two: interactive dashboard design

- 2.1 Background research
- 2.2 Exploration of data set
- 2.3 Investigation of data workflow and proposal for design of dashboard
- 2.4 Discussion
- 2.5 Conclusion

### Part Three: Reference and Appendices

## PART ONE: STATISTICAL ANALYSIS

### 1. Introduction

Life tables are used to statistically measure the Average Life Expectancy (Life Expectancy and Their Different Users, 2018).

Life expectancy at birth is "the lifespan or duration an individual is expected to live starting from birth" (Databank, n.d).

Medical advancement and better hygiene Have increased Life globally over time. However, Life span Expectancy is visibly affected by genetics, lifestyle, gender, nutrition and external factors that determine Life Expectancy, including economic and environmental growth (Longevity: Extending Life Span Expectancy, 2022). For this study, the data was sourced from 2011 to 2022, countries were selected from two regions, and indicators were picked based on background research, personal knowledge, and reasoning.

The research aims to identify and highlight the different longevity levels in selected regions worldwide and discuss factors that impact it. Further hypothesis testing will be used to discover these differences in life expectancy, and forecasts will show future trends.

Research question:

Regarding socioeconomic factors and world development indicators, what are the influential attributes of Life Expectancy at birth in different regions? Considering the above question, the following objectives will be met.

The objectives of this research are:

1. Explore the life expectancy difference between Europe and Africa.
2. Assess the relationship between Life Expectancy and other indicators.
3. Identify the relationship of Life Expectancy with selected indicators using regression analysis
4. Trend Analysis and forecast of average Life Expectancy (2011-2022)
5. Using hypothesis testing to identify regional groups' average Life Expectancy rate.

### 1.2 Background Research

Life expectancy has been a subject of interest amongst researchers observing indicators like the ones selected in this analysis. Linear regression predicts Average Life Expectancy across nations by using Labor force population amongst other factors, and Eberhart shows it, M. Also, there have been several approaches to this research depending on the indicator of interest.

Regression analysis on Life Expectancy (2022) found that life expectancy is related to health expenditure. However, this research includes basic sanitation and drinking water along with undernourishment. These variables substantially impact life equality and, hence, the Longevity of Life.

### 1.3 Correlation Analysis

The coefficient correlation value ranges from 0, meaning no correlation,  $-1$  to  $+1$ ,  $+1$  means positively correlated, and  $-1$  means negatively correlated. Correlation thus shows how variables are related to each other.

Similarities between variables, the dependent variable and independent variables, are also known as Regression Analysis as described by Handyman, R. and Athanasopoulos, G. (Hyndman & Athanasopoulos, 2018)

Time series Analysis is the monitoring of items harnessed via measurements of a homogenous time frame. The observed series has three components, including.

1. Seasonal systematic calendar movements
2. Long term direction
3. Irregular unsystematic fluctuations (Time series Analysis, n.d.).

Hypothesis testing is an act in statistics where an analyst tests the assumption regarding the population parameter (Hypothesis testing, n.d.). P. value is the determinant of rejecting or accepting the null hypothesis. Usually, the significance level is 0.05 or 0.01.

### 1. Exploration of Data Set

The statistical analysis uses R. for data preparation; firstly, unwanted columns are deleted, and the names of series are changed. Data is sourced from the World Bank (World Development Indicators, n.d.) 12 countries have been stratified into two regions, namely Europe and Africa, see Table 3. Selection is made depending on the country's geographical location. The time range selected is from 2011 to 2022.

Columns with missing values are deleted. The indicators used for this research and their meaning are in the table below. Definitions were taken from metadata (Databank, n.d.).

Countries and Regions		
Region	Western Europe	Sub-Saharan Africa
Countries	1. Austria	1. Ethiopia
	2. Ireland	2. Liberia
	3. Hungary	3. Rwanda
	4. Denmark	4. Mozambique
	5. France	5. Sierra Leone
	5. United Kingdom	6. Senegal

**R studio** is used for loading and viewing of Dataset

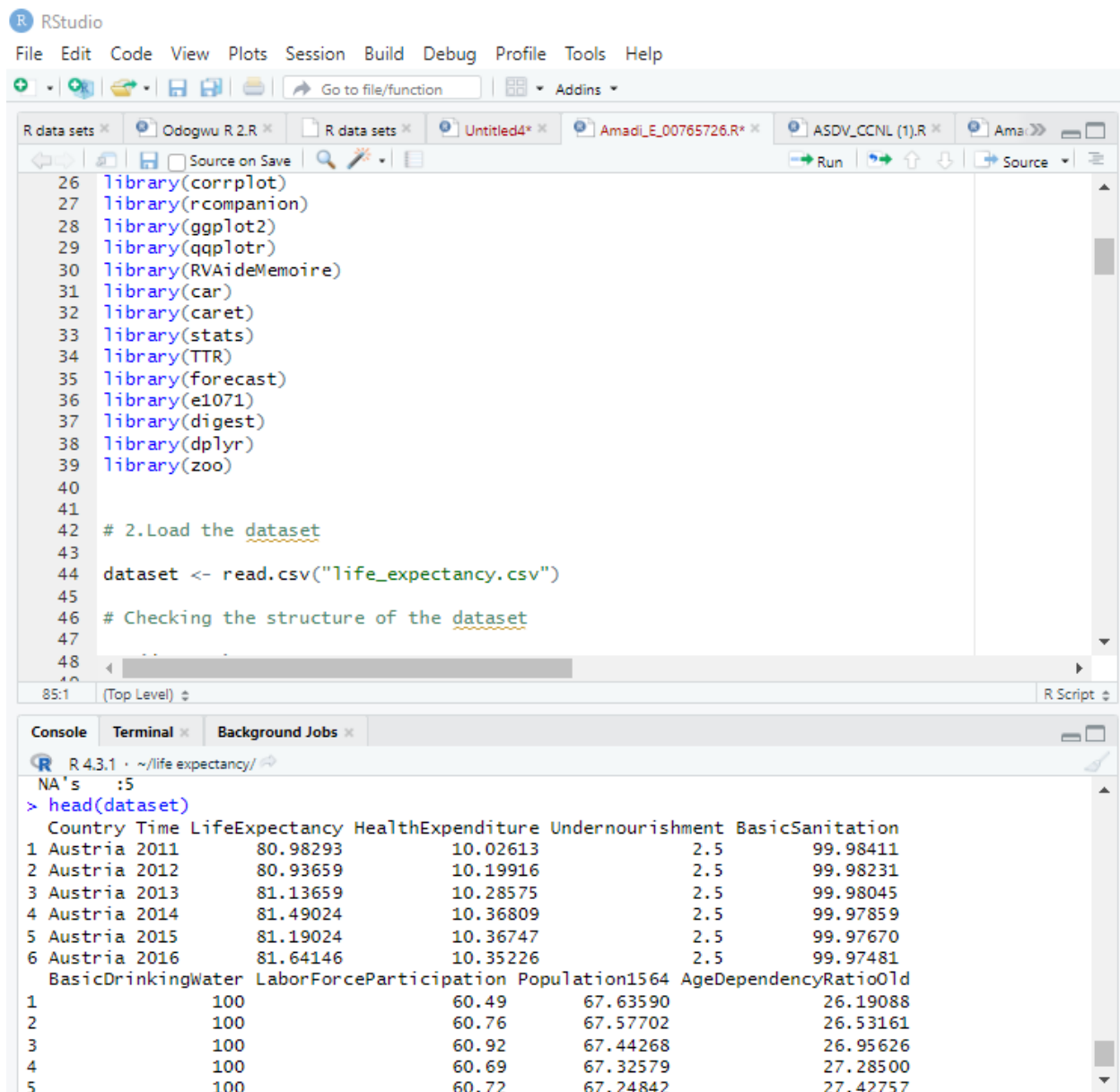


Figure 1 loading of dataset and dropping of column

Dropping unnecessary columns such as year code and country code

**Changing Type:** Columns except year and country are numeric because they are all continuous statistically measured values.

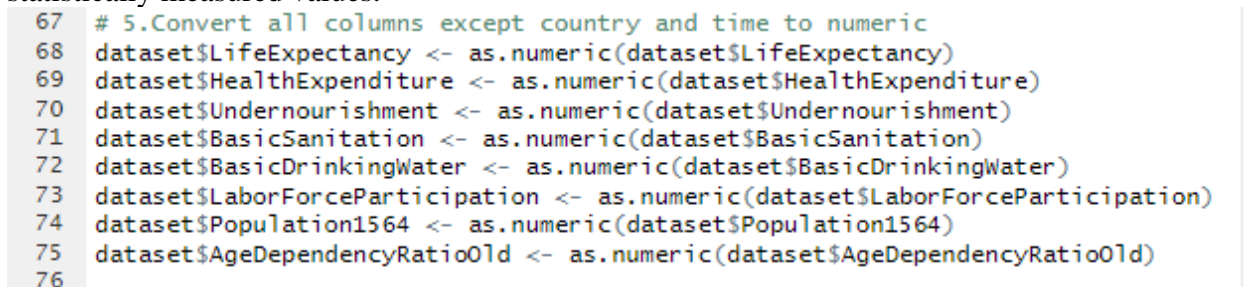


Figure 2

Checking for missing values for this dataset is not an acceptable option since countries differ in their demographics, size, population, GDP and Political situation, so they are removed.

```

88 # Checking for Missing Values
89 any(is.na(dataset))
90
91 # dropping of missing values
92 dataset <- na.omit(dataset)
93
94
95
96
99:1 (Top Level)
R Script

```

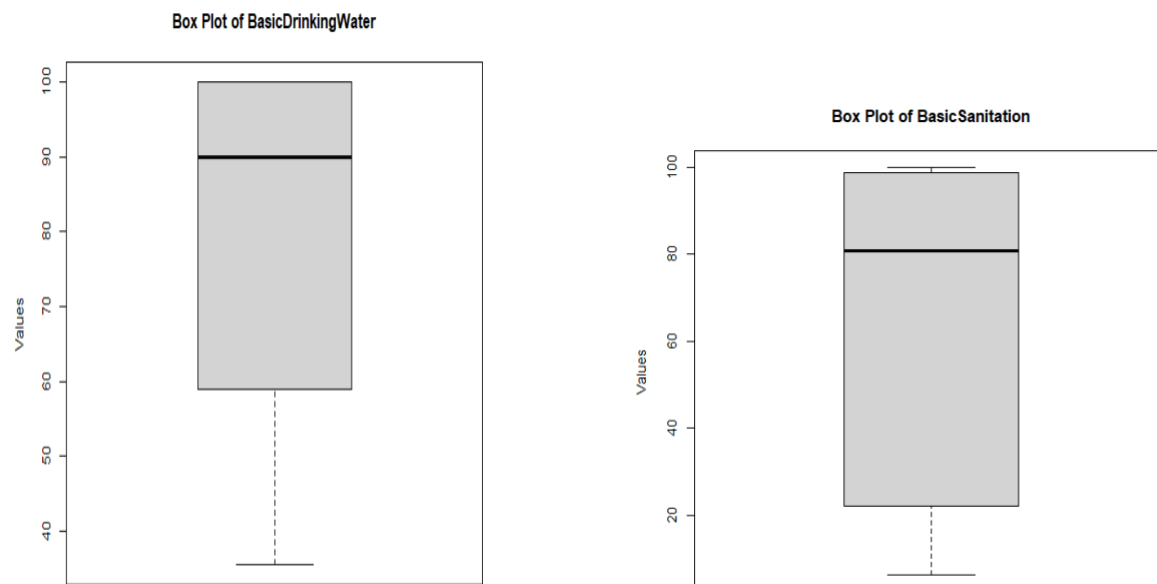
```

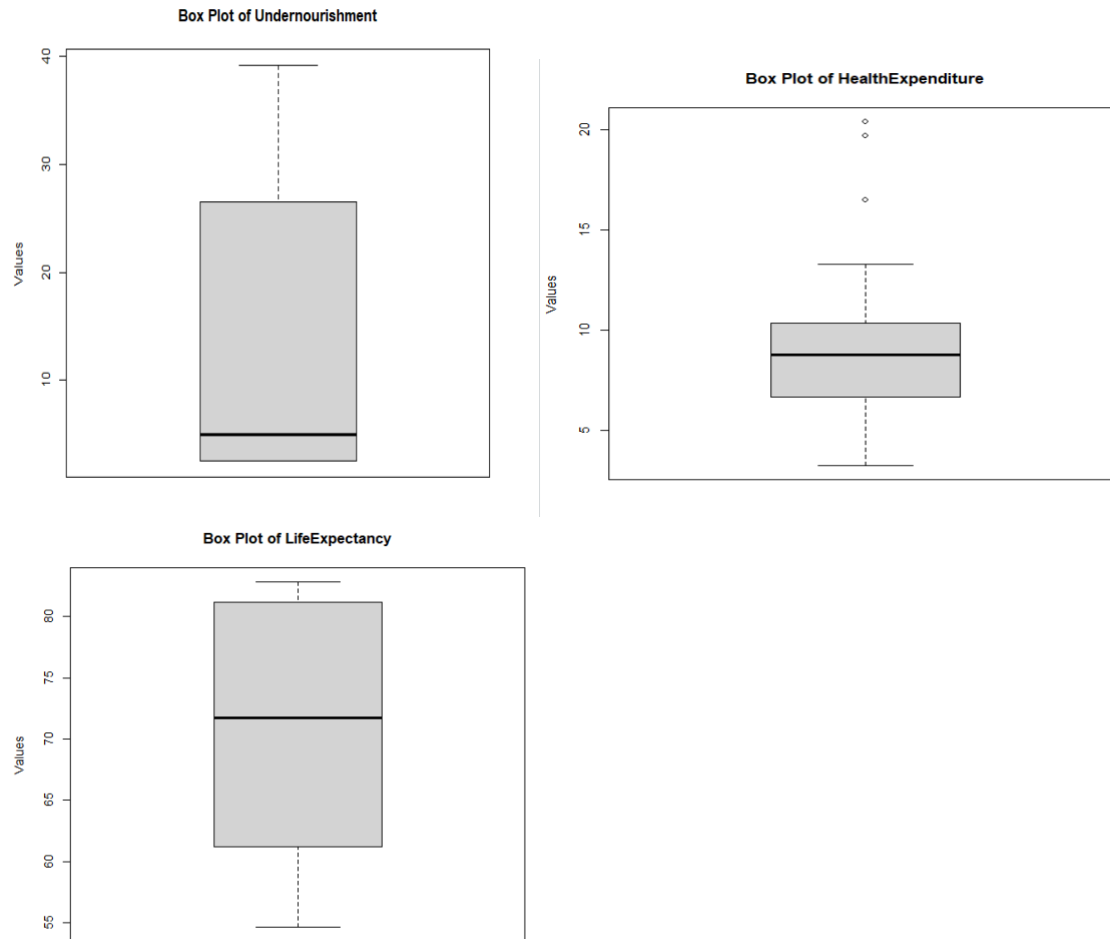
R 4.3.1 ~ /life expectancy/
6 Austria 2016 81.64146 10.35226 2.5 99.97481
BasicDrinkingWater LaborForceParticipation Population1564 AgeDependencyRatioOld
1 100 60.49 67.63590 26.19088
2 100 60.76 67.57702 26.53161
3 100 60.92 67.44268 26.95626
4 100 60.69 67.32579 27.28500
5 100 60.72 67.24842 27.42757
6 100 61.21 67.14602 27.52965
> # Checking for Missing Values
> any(is.na(dataset))
[1] TRUE
> # dropping of missing values
> dataset <- na.omit(dataset)
> mat<- matrix(c(1,2,3,4,5,6,7,8,9), nrow = 3, byrow = TRUE)

```

Figure 3

**Outliers' detection:** Outlier removal will interfere with the integrity of the objectives of this project hence no outliers are removed.





*Figure4Boxplot*

## Life expectancy

Africa shows the highest variance followed by Africa in primary drinking water availability to population. The same is valid for health expenditure undernourishment. Distribution for Europe has low variance and standard deviation, indicating that most countries have good living conditions. This is statistically determined in the next section.

### 2.1 Comprehensive Descriptive Statistical Analysis

#### Histograms for exploration of the dataset:

Figure 5 & 6 describes the distribution using Histograms. Life Expectancy distribution indicates that many countries have higher Life Expectancy. Basic sanitation and primary drinking water distribution have similar shapes, with the highest count towards the graph's right. Undernourishment has increased in a few countries; hence, the graph's tail is towards the right. This will be later explained when skewness and kurtosis are discussed below. Gvt\_health\_exp distribution is built.

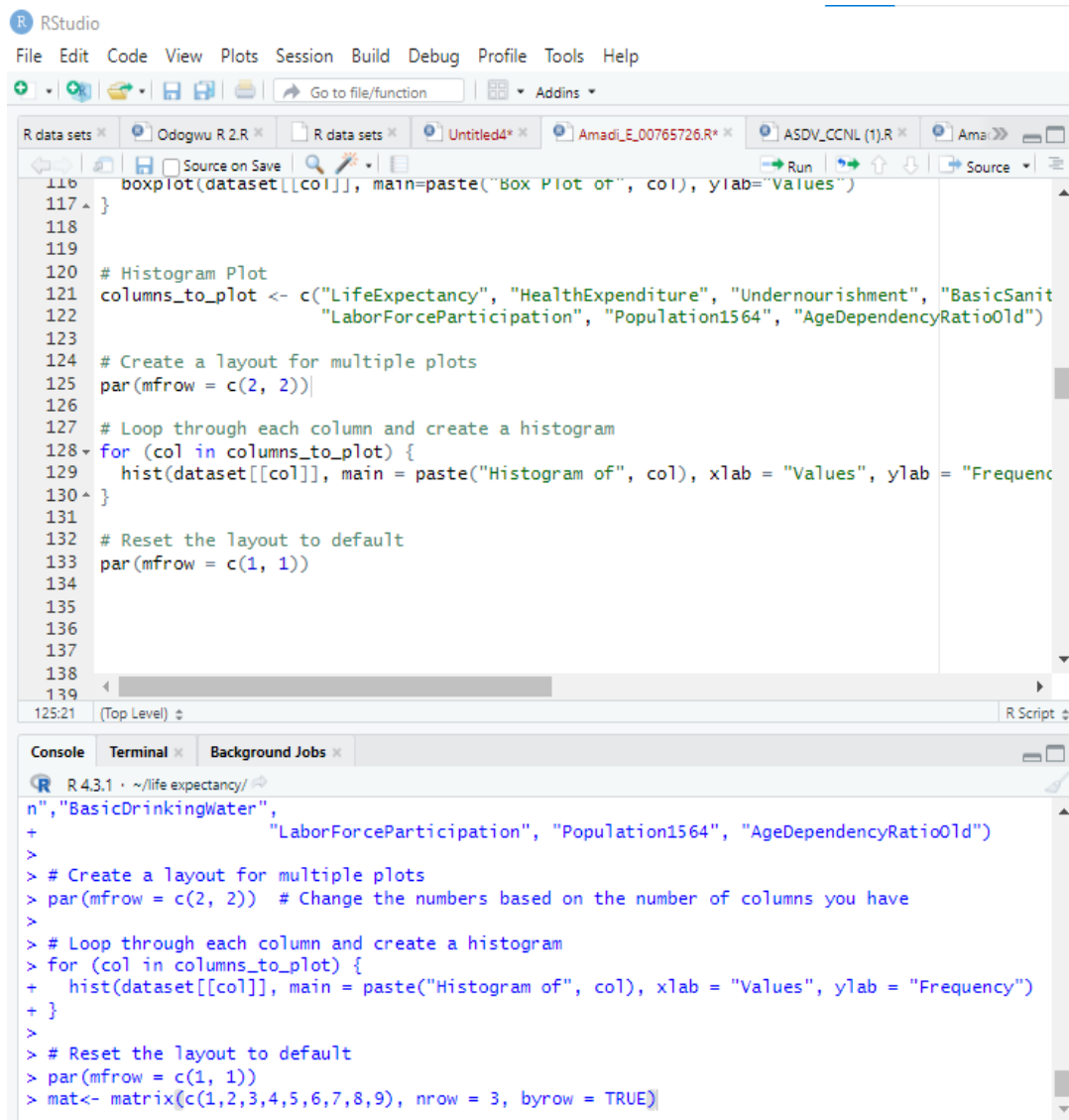
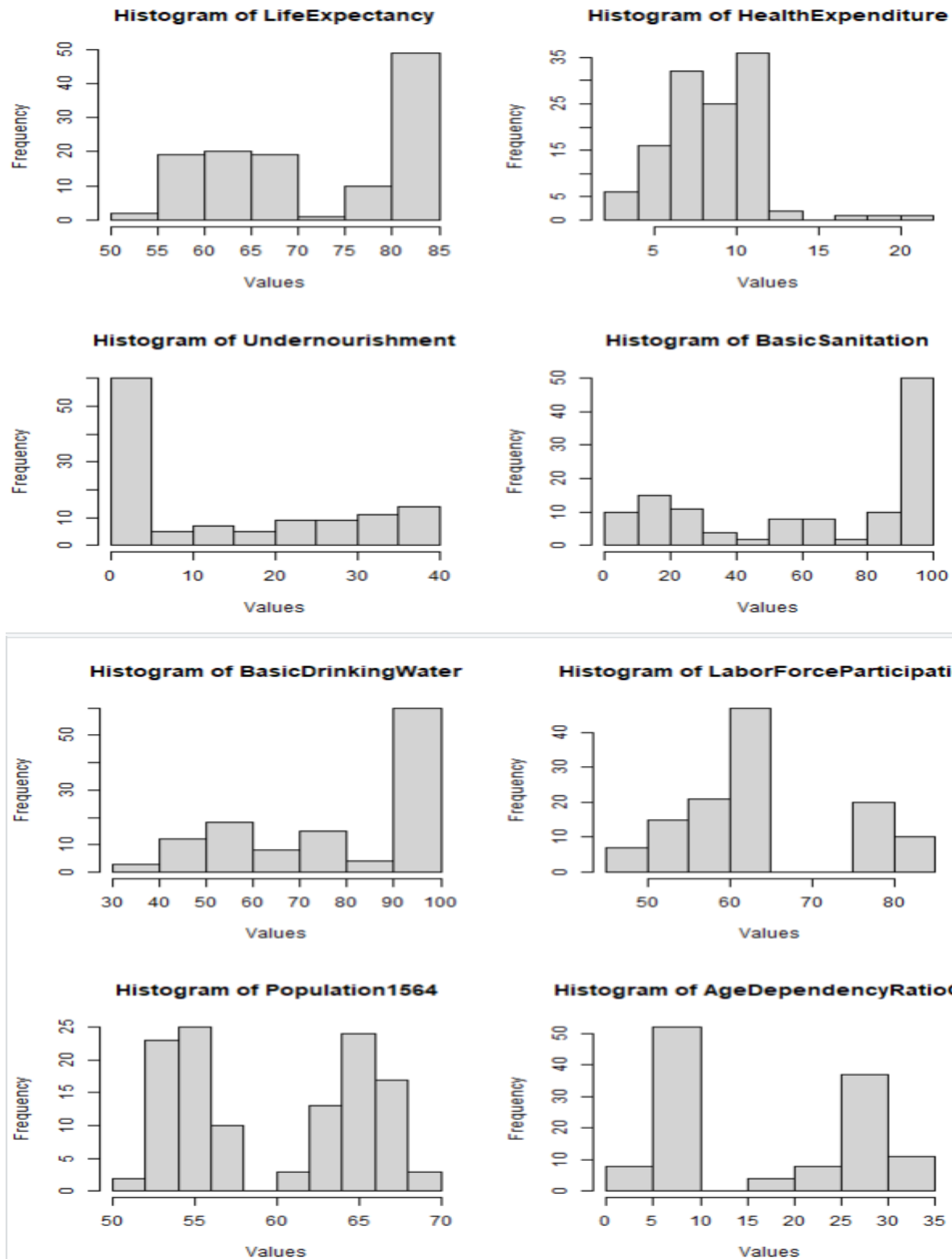


Figure 5





*Figure 6*

Mean and median: below is the mean and median per country Code and Output:

The Mean LE of Europe is the highest, undernourishment is the lowest at 2.5, and basic sanitation is the highest. Africa has a long way to go with the highest undernourishment and most insufficient basic sanitation access. It is interesting to note that life expectancy in this region is also the lowest. Skewness and Kurtosis values and interpretation

Skewness value interpretation,

- Symmetric: Value between  $-0.5$  to  $0.5$
- Moderated Skewed data:  $-1$  and  $-0.5$  or between  $0.5$  n.d  $1$
- Highly Skewed data: Values less than  $-1$  or greater than  $1$

Life expectancy distributions for all regions are almost symmetric, with the skewness of Africa being the highest negative and Africa being the lowest negative. Kurtosis values for Africa are the highest, indicating a longer tail.

```
171 # Columns for which you want to calculate statistics
172 selected_columns <- c("LifeExpectancy", "HealthExpenditure", "Undernourishment", "BasicSani
173                       "LaborForceParticipation", "Population1564", "AgeDependencyRatioOld"
174
175 # Function to calculate skewness (using e1071 package)
176 calculate_skewness <- function(x) {
177   skewness <- e1071::skewness(x, na.rm = TRUE)
178   return(skewness)
179 }
180
181 # Function to calculate kurtosis (using e1071 package)
182 calculate_kurtosis <- function(x) {
183   kurtosis <- e1071::kurtosis(x, na.rm = TRUE)
184   return(kurtosis)
185 }
186
187 # Calculate statistics For 12 countries
188 selected_countries <- c("Austria", "Denmark", "Ireland", "Hungary", "France", "United Kingdo
189                       "Ethiopia", "Liberia", "Rwanda", "Mozambique", "Sierra Leone", "Sene
190
191 # Create new columns for mean, median, skewness, and kurtosis for each selected country
192 statistics_by_country <- data.frame(Country = character(), stringsAsFactors = FALSE)
193
194 for (country in selected_countries) {
195   subset_data <- dataset[dataset$Country == country, ]
196   statistics <- sapply(dataset[, selected_columns], function(x) {
197     mean_val <- mean(x, na.rm = TRUE)
198     median_val <- median(x, na.rm = TRUE)
199     skewness_val <- calculate_skewness(x)
200     kurtosis_val <- calculate_kurtosis(x)
201     return(c(Mean = mean_val, Median = median_val, Skewness = skewness_val, Kurtosis = kurt
202   })
203   statistics_by_country <- rbind(statistics_by_country, cbind(Country = country, statistics
204 }
205
206 # Display the updated dataset
207 print(statistics_by_country)
```

Output:

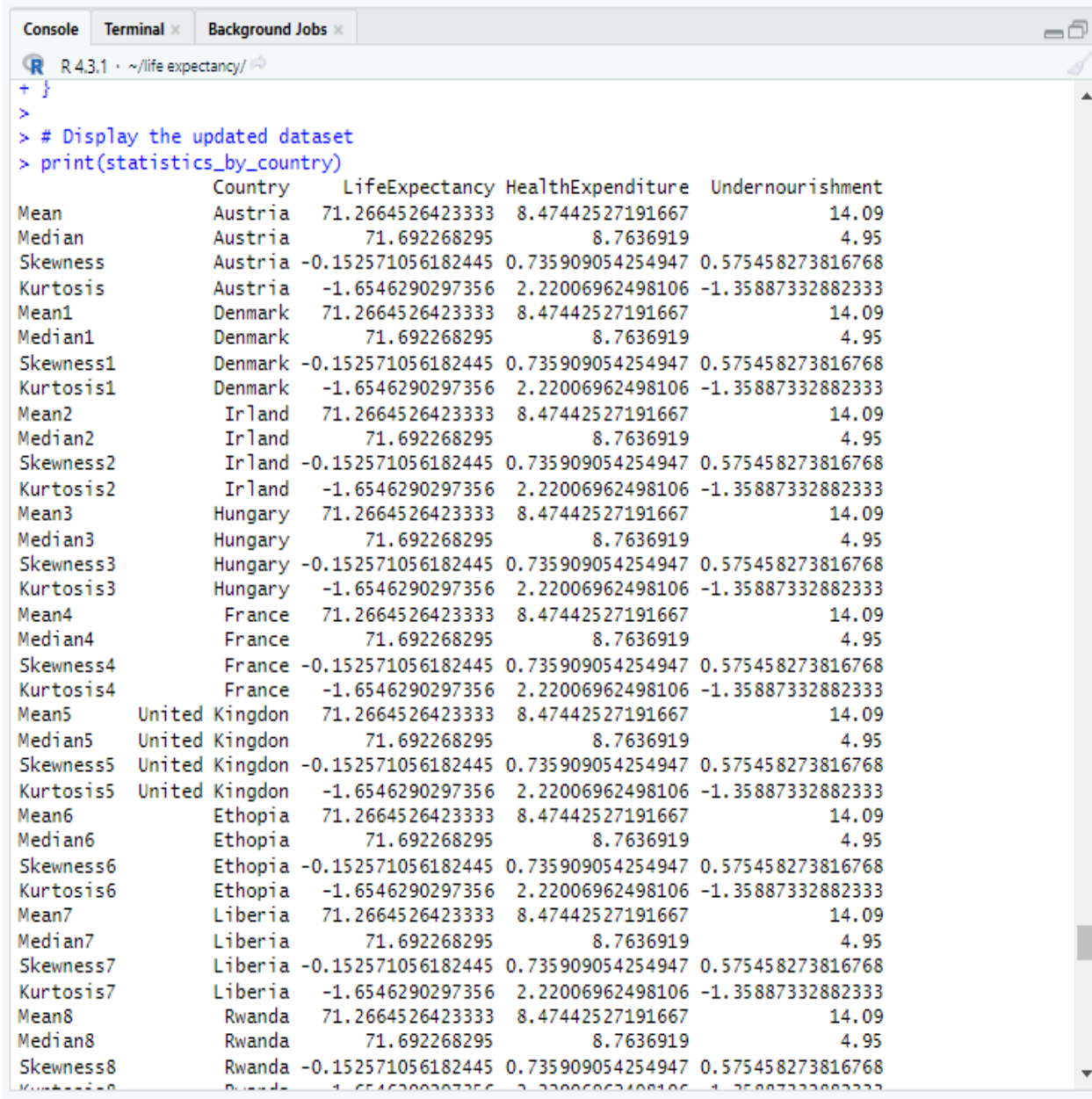


Figure 7

## 2.1 Correlation Analysis

Unlike the Pearson coefficient, the Spearman Ranks method replaces the observation values with their rank; hence, it is more robust to outliers and non-gaussian, skewed distribution (Mukaka, 2012). Since there are outliers and non-normal distribution. Spearman ranks are used in the figure below. The data frame is processed by removing categorical features such as country (step 1). Use core data; the method is specified as Spearman (Step 2).

```

178 # Correlation Analysis
179
180 # Columns for which you want to calculate correlations
181 selected_columns <- c("LifeExpectancy", "HealthExpenditure", "Undernourishment", "BasicSani
182                       "LaborForceParticipation", "Population1564", "AgeDependencyRatioOld")
183
184
185 # Calculate correlations
186 correlation_matrix <- cor(dataset[, selected_columns], use = "complete.obs")
187
188 # Display the correlation matrix
189 print(correlation_matrix)
190 # Plot the correlation matrix using corrrplot
191 corrrplot(correlation_matrix, method = "color")
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

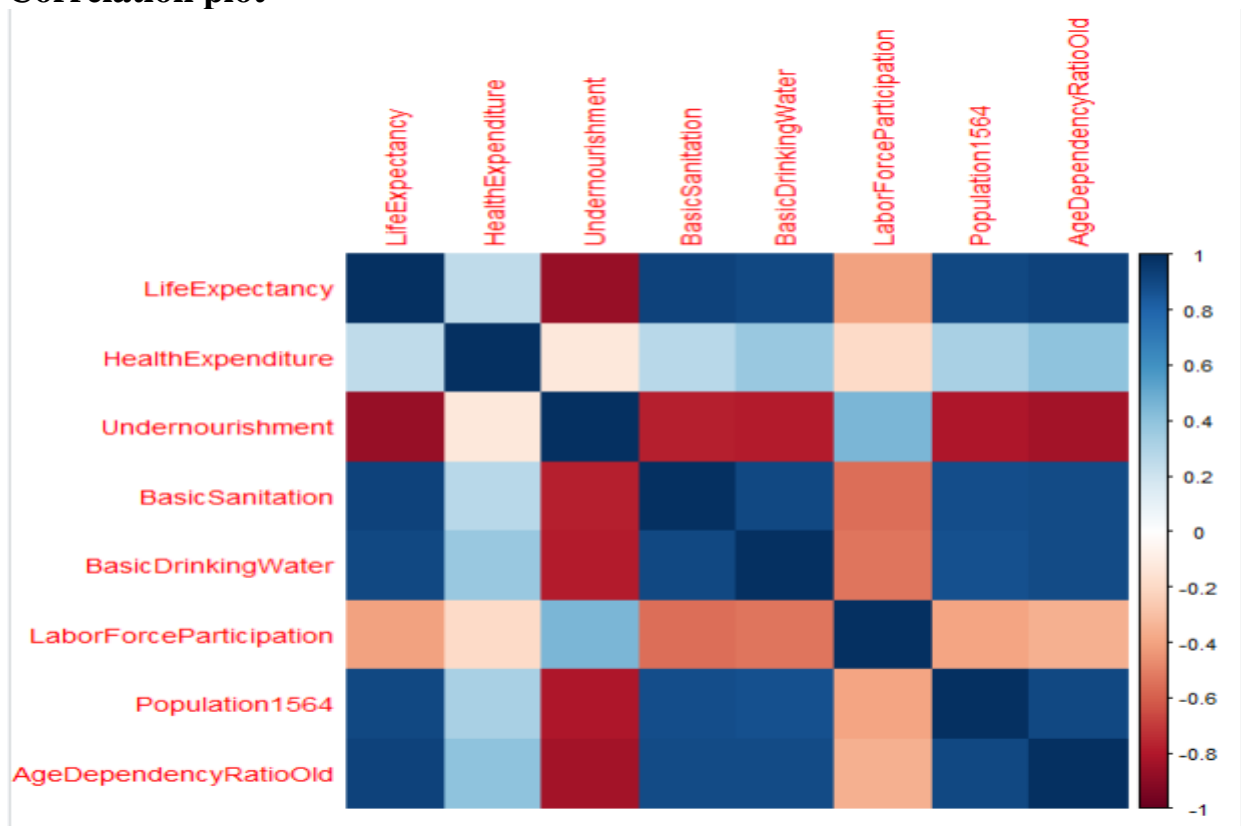
```

> # Display the correlation matrix
> print(correlation_matrix)
      LifeExpectancy HealthExpenditure Undernourishment BasicSanitation
LifeExpectancy      1.0000000      0.2517817      -0.8673051      0.9287870
HealthExpenditure    0.2517817      1.0000000      -0.1251030      0.2759039
Undernourishment     -0.8673051     -0.1251030      1.0000000     -0.7794806
BasicSanitation       0.9287870      0.2759039     -0.7794806      1.0000000
BasicDrinkingWater    0.9006805      0.3772831     -0.7889652      0.9055227
LaborForceParticipation -0.4092650     -0.1963274      0.4528002     -0.5576778
Population1564        0.9077090      0.3256494     -0.8054175      0.8885698
AgeDependencyRatioOld 0.9207605      0.4044849     -0.8358336      0.8927431
      BasicDrinkingWater LaborForceParticipation Population1564
LifeExpectancy          0.9006805          -0.4092650          0.9077090
HealthExpenditure        0.3772831          -0.1963274          0.3256494
Undernourishment        -0.7889652           0.4528002         -0.8054175

```

Figure 8

## Correlation plot



## Figure 9

### Evaluation:

According to the results of the Spearman correlation, life expectancy is adversely connected with undernourishment for 149 observations,  $p < 0.01$ , and highly correlated with drinking water and sanitation for 149 observations. This demonstrates that undernourishment and shorter life expectancy are associated with increased malnutrition rates (The impact of malnutrition on infant mortality and life expectancy in Africa, 2022).

Attributes highly correlated to Life expectancy (all p-values for below table  $< 0.01$ )

Attributes	Correlation coefficient
Basic Sanitation	0.89
Age Dependency Ratio Old	0.84
Basic Drinking Water	0.79
Undernourishment	-0.63
Health expenditure	0.57

*Table 10 Summary of correlation*

## 2.2 Regression Analysis

All assumptions must be met to build a successful model for regression. Assumptions include:

1. Linearity,
  1. Normality:
  2. Residuals independence.
  3. No homoscedasticity
  4. No multicollinearity

Regression Model 1: The objective of this task is to observe for the year 2022 if average Life Expectancy (regress) can be predicted as a function of attributes that affect it. From the scatter plot, it is evident that there is a linear relationship between Life Expectancy – Basic Sanitation and Life Expectancy - gvt\_health\_exp. The next step is to build a forward stepwise mode. The following text includes details on how this is performed and the findings.

Pr ( $>|t|$ ) coefficients are significant for the intercept and basic Sanitation. The residual error is small, 4.5 with 16 degrees of freedom, and R squared is 0.75, meaning that basic Sanitation can explain 75% variance in LE. Below is the regression line. Now, assumptions will be checked.

The homoscedasticity assumption is met, as seen below. The red line between residuals is almost flat, signifying no apparent pattern. Residuals are randomly scattered.

Figure 46

All assumptions have been met. This is a good model to predict LE with the following factors. Life expectancy equation is:  $LE = 56.26 + 0.24 \times (\text{basic\_sanitation})$

The model is tested against data, and the results are based on the equation below.

Basic Sanitation	Life Exp. Predicted	Life Exp. Actual
44.7	66.9	66.6
99.2	80	78.4

*Table 11 Testing*

Multiple Regression Model 2 and Comparison with Model 1:

The multiple regression model is done stepwise, and gvt\_health expense is added. As a result, the R square increased from 0.75 to 0.78. Residual standard errors have reduced from 4.58 to 4.26, and the p-value for RSE is below the 0.05 significance level.

```
# Linear Regression model with multiple independent variables
my_multiple_linear_model <- lm(LifeExpectancy ~ HealthExpenditure + Undernourishment + Basi

# Summary of the multiple regression model
summary(my_multiple_linear_model)

# Diagnostic plots
residuals_multiple <- residuals(my_multiple_linear_model)

# Diagnostic plots

# Linearity Check
plot(my_multiple_linear_model)

# Normality Check
hist(residuals_multiple)
qqnorm(residuals_multiple)
qqline(residuals_multiple)

# Residuals Independence Check
plot(residuals_multiple ~ predict(my_multiple_linear_model))

# Homoscedasticity Check
plot(predict(my_multiple_linear_model), residuals_multiple)

# Multicollinearity Check
vif(my_multiple_linear_model)
```

Output:

```
Call:
lm(formula = LifeExpectancy ~ HealthExpenditure + Undernourishment +
    BasicSanitation + BasicDrinkingWater + LaborForceParticipation +
    Population1564 + AgeDependencyRatioOld, data = dataset)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-6.8732	-0.7533	0.1506	1.6008	5.0371

Coefficients:

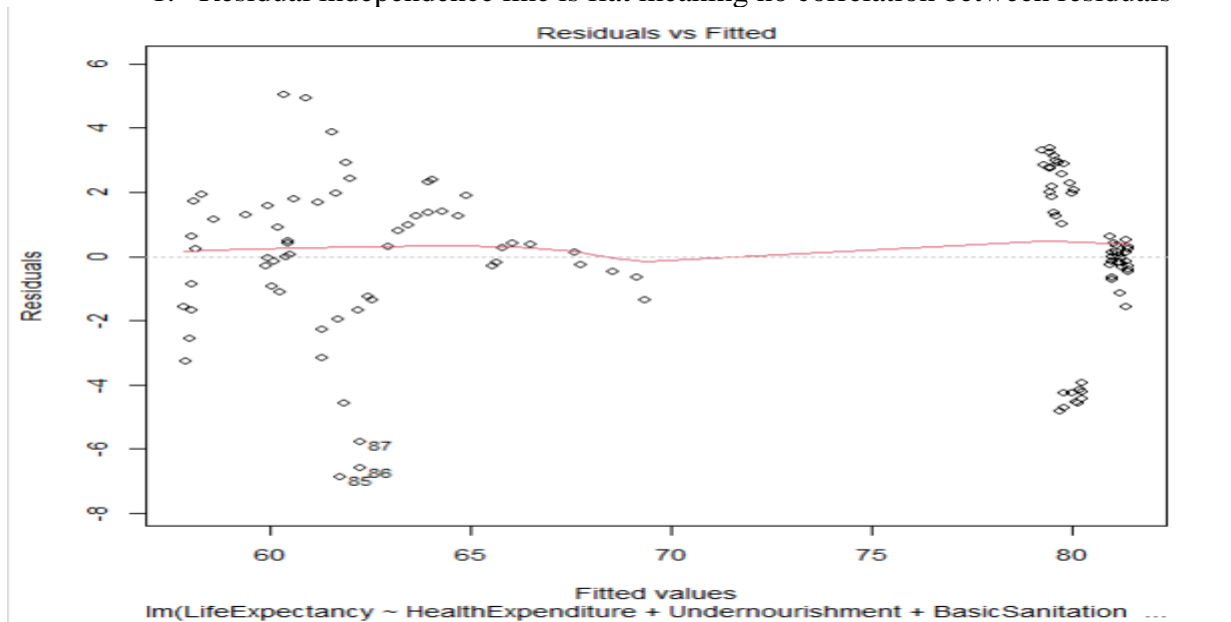
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	39.6730543	5.7603321	6.887	3.49e-10 ***
HealthExpenditure	0.0004066	0.0989353	0.004	0.9967
Undernourishment	-0.2244089	0.0372483	-6.025	2.21e-08 ***
BasicSanitation	0.1447664	0.0200731	7.212	6.93e-11 ***
BasicDrinkingWater	0.0830695	0.0294053	2.825	0.0056 **
LaborForceParticipation	0.1571716	0.0335582	4.684	7.97e-06 ***
Population1564	0.1480774	0.1065306	1.390	0.1673
AgeDependencyRatioOld	-0.0014659	0.0764454	-0.019	0.9847

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.414 on 112 degrees of freedom  
Multiple R-squared: 0.9427, Adjusted R-squared: 0.9391  
F-statistic: 263.1 on 7 and 112 DF, p-value: < 2.2e-16

Checking Assumptions:

1. Residual independence line is flat meaning no correlation between residuals



2. Normal Q-Q –Normality of residuals: most datapoints fall on line indicating normality

Normal Q-Q plot:

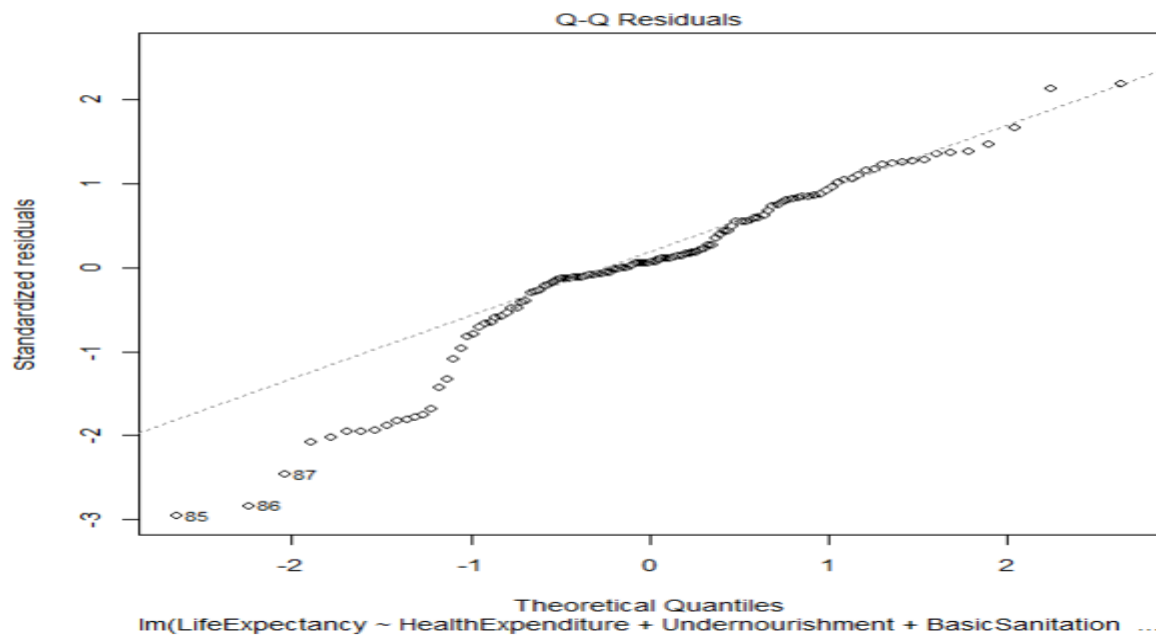


Figure 8

3. **Homoscedalstcity:** Equal variability between fitted values; since fewer observations are used, the below graph is acceptable.

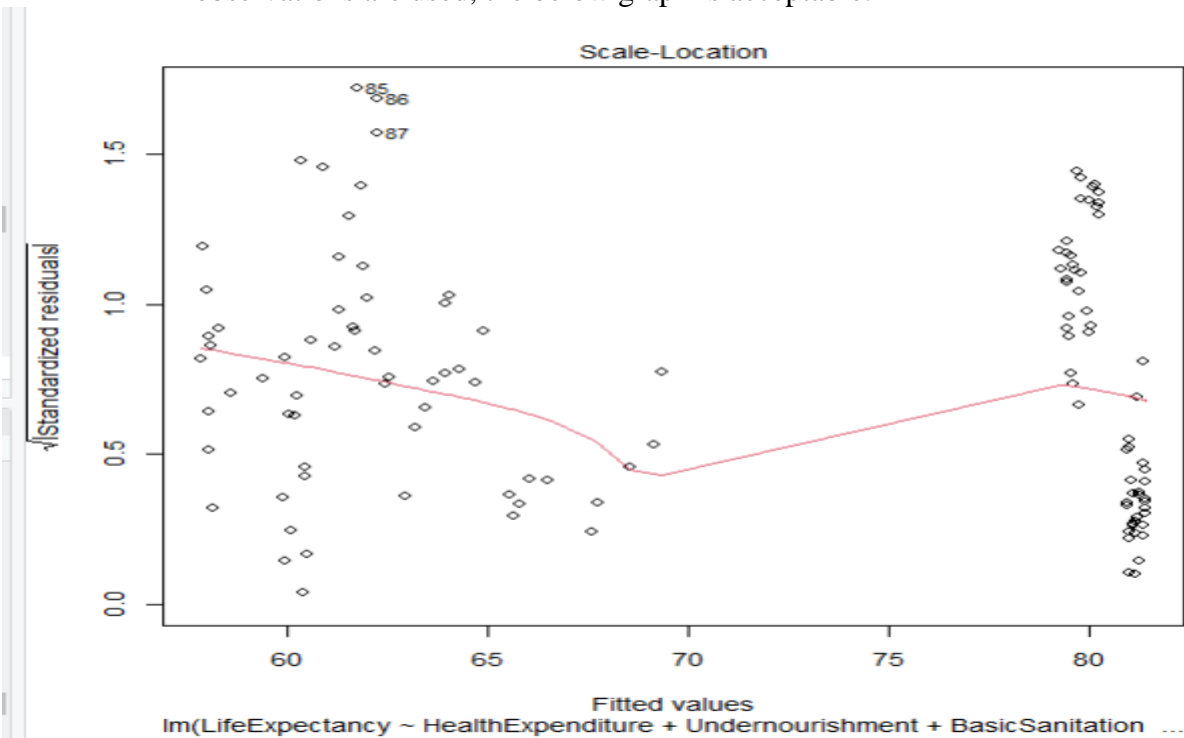


Figure 9

4. **Checking Multicollinearity:**



```

222
223 # Multicollinearity Check
224 vif(my_multiple_linear_model)

```

Output:

```

>
> # Multicollinearity Check
> vif(my_multiple_linear_model)
      HealthExpenditure      Undernourishment      BasicSanitation
      1.800861           5.161989           10.807136
      BasicDrinkingWater LaborForceParticipation      Population1564
      8.384524           2.368477           7.367068
      AgeDependencyRatio0ld
      14.674900

```

Figure 9

Since the variance inflation factor value is  $1.667 < 5$ , as seen in Figure 47, there is a small collinearity between the two metrics used in the model.

All four assumptions are approved n.d. The fitted regression equation is:  $LE = 53.9 + 0.18 * (\text{basic\_sanitation}) + 0.48 * (\text{gvt\_health\_exp})$

### 3.0 Time Series Analysis

In this section, two widely used time series models will be developed using Holt-Winters and ARIMA to forecast LE across two regions, and model performance will be compared. The below shows that the mean L is gradually increasing.

Time series code: Right: Time Series Data frame Created

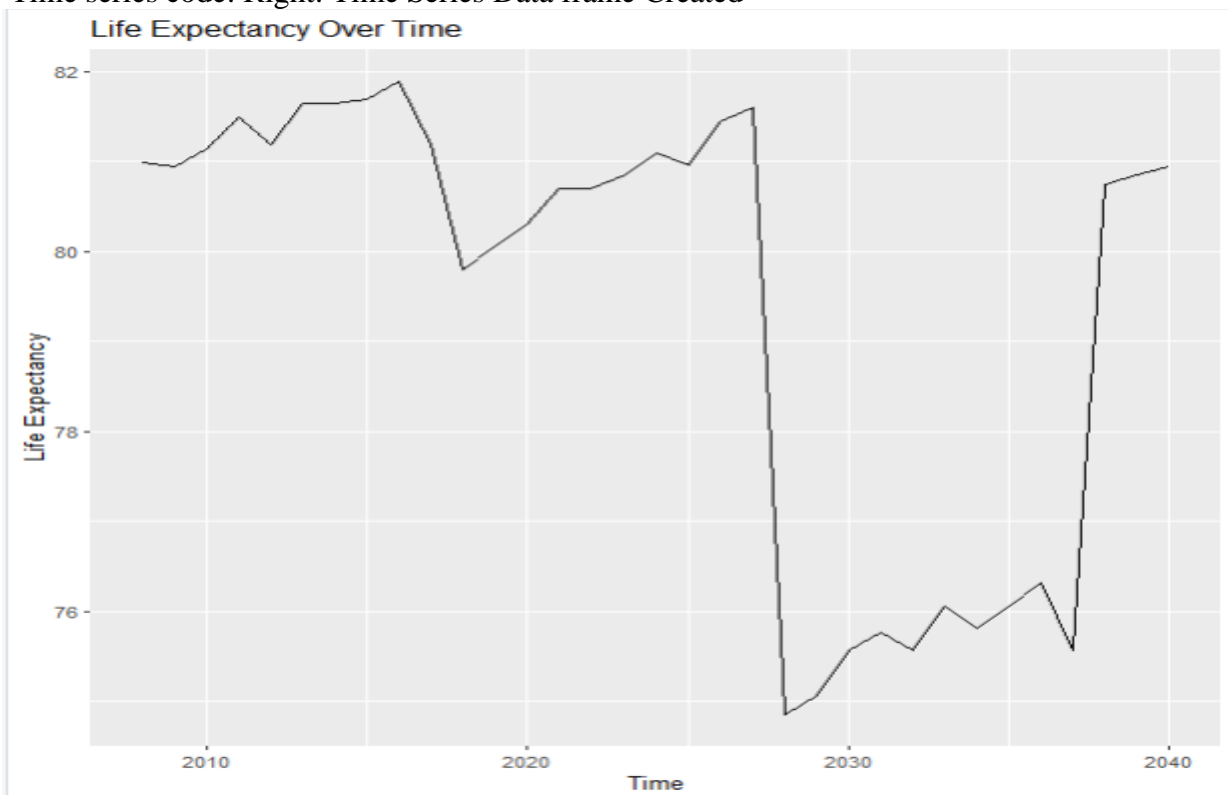


Figure 11

The Adf test below has  $p > 0.05$ ; the Alternative hypothesis is rejected at a 5% significance level, meaning the series is non-stationary.

### Time Series Model

Holt-Winter has three parameters, namely alpha, beta and gamma, that specify the coefficients for level smoothing, trend smoothing, and seasonal smoothing, respectively (Holt-Winters (Times Series); n.d). Gamma is set to FALSE because this series is stationary.

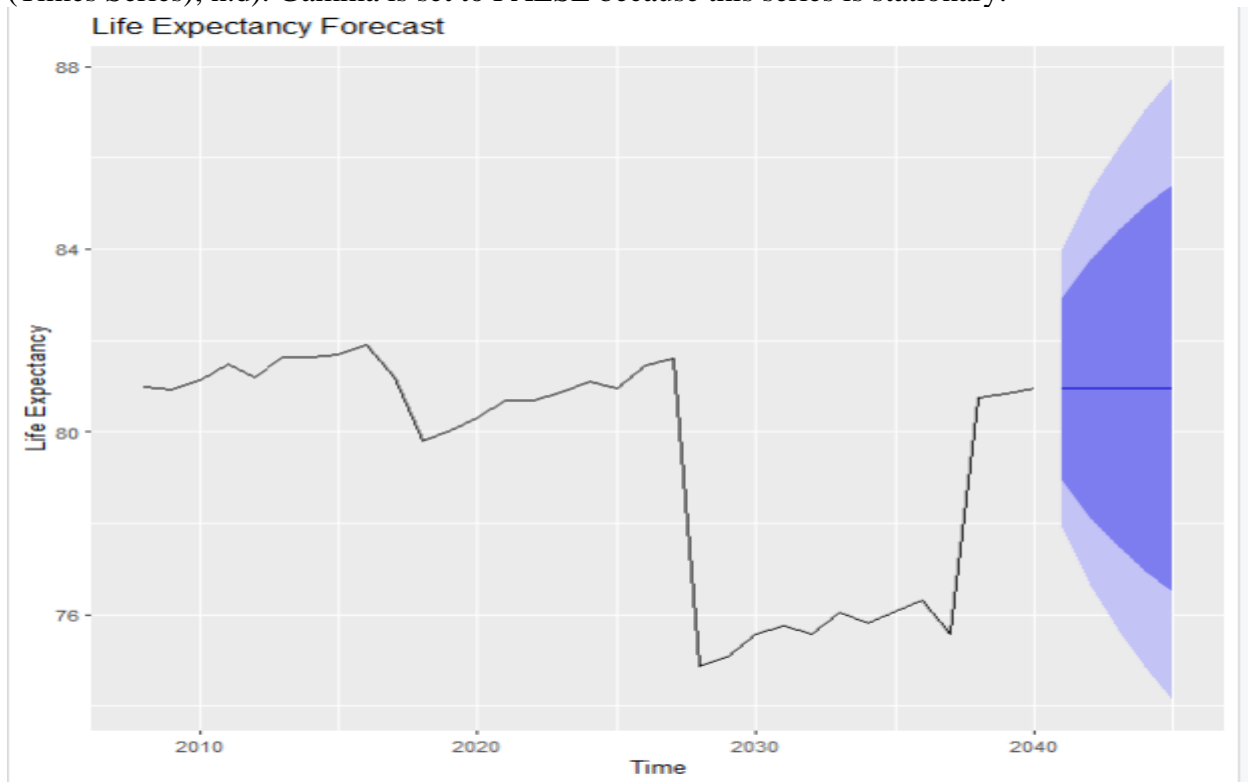


Figure 12

The forecast was plotted for 10 years. It shows that LE gradually increases globally; however, the increasing gradient decreases over time. The forecast predicts an LE of 75.56

The figure below shows that the autocorrelations for the forecast errors always do not exceed the significance bound for lags 0 to 10. The blue lines in the ACF plot represent the significance threshold. Lines that cross the level show a correlation between forecast errors, signifying that the model can be improved. This model has no autocorrelation for any lags.

```
> # Fit ARIMA model
> arima_model <- auto.arima(ts_life.expectancy)
> summary(arima_model)
Series: ts_life.expectancy
ARIMA(0,1,0)

sigma^2 = 2.415: log likelihood = -59.51
AIC=121.03 AICc=121.16 BIC=122.49

Training set error measures:
      ME      RMSE      MAE      MPE      MAPE      MASE      ACF1
Training set 0.001419289 1.530368 0.6394089 -0.01756697 0.8192202 0.973433 -0.07152622
```

Ljung-Box test (figure 54) shows test statistics as 1.39 and  $p > 0.05$  (p-value 0.84), indicating little evidence of autocorrelations for forecast errors for lags 1-4

**Plotting Residuals:** Forecast errors plotted on the time graph show constant variance over time. A slight skew is to the right, but most errors are typically distributed across mean 0.

### Time Series Model 2 ARIMA:

Arima assumes the series is stationary, as seen in the ACF test; the series is non-stationary; we can proceed with ARIMA by differencing the series' first. Exploring Health Expenditure Indicator.

```
> # Fit ARIMA model
> arima_model1 <- auto.arima(ts_HealthExpenditure)
> summary(arima_model1)
Series: ts_HealthExpenditure
ARIMA(0,1,0)

sigma^2 = 0.7652: log likelihood = -41.12
AIC=84.25   AICc=84.38   BIC=85.71

Training set error measures:
              ME      RMSE      MAE      MPE      MAPE      MASE      ACF1
Training set 0.008867413 0.8614018 0.3888246 -0.3795731 4.384871 0.9704553 0.04011005
```

Figure 13

Auto Arima option has already computed p.d.q values as highlighted below.

Comparison of Holt Winter and ARIMA:

A minor difference exists between the mean residuals of both models from the Box-Ljung test: Holt-Winters' p-value is higher, indicating better prediction. On the other hand, ACF is lower for the ARIMA model than Holt-Winters. Holt-Winters predicts a value of 2040 LE 75.56, and Arima indicates 75.14. Given this information, Arima fits better into the dataset and is the better choice for the time forecast of LE.

### 4.5 Hypothesis Testing

Observing the distribution of life expectancy amongst three regions:

Density plots show the difference in means of the two populations: Africa and Europe. There is some overlap between the density plots of the three areas. We carry out the hypothesis test to determine whether the difference in LE is statistically significant between areas.

**Forming a hypothesis:**

**H<sub>0</sub>**= there is no visible significant difference in the means of life expectancy between different regions

**H<sub>0</sub>:**

**...Europe=...Africa**

**H<sub>1</sub>**= At least one region out of three has a different mean life expectancy

**Confidence level, Alpha=0.05**

```

> ### Hypothesis test 1 #####
>
> ### Pearson corr Life Expectancy vs Current.health.expenditure #####
> cor_result <- cor.test(dataset$LifeExpectancy, dataset$HealthExpenditure, method = "pearson")
>
> # Print the correlation coefficient and p-value
> cat("Correlation Coefficient:", cor_result$estimate, "\n")
Correlation Coefficient: 0.2517817
> cat("P-value:", cor_result$p.value, "\n")
P-value: 0.005535466
>
> # Set significance level (e.g., 0.05)
> alpha<-0.05
>
> # Compare p-value with significance level
> if (cor_result$p.value < alpha) {
+   cat("The hypothesis test is statistically significant. We reject the null hypothesis.\n")
+ } else {
+   cat("The hypothesis test is not statistically significant. We fail to reject the null hypothesis.\n")
+ }
The hypothesis test is statistically significant. We reject the null hypothesis.
.

```

*Figure 14*

The objective of the hypothesis is to assess mean LE for more than two regions; therefore, it is compulsory to use a hypothesis test that allows comparison for two or more groups. Hypothesis tests such as .... test and ANOVA assume normality of data before proceeding onto the test data needs to ... for normality. This can be checked in various ways:

Quantile-Quantile Plot:

Shapiro Wik tests statistically check if the data is normally distributed; it assumes null hypothesis = data is not normally distributed. A P-value more significant than 0.05 means that the data is usually spread.

The image below shows the result and analysis of the LE mean normality test; Europe has a p-value of 0.44, indicating I usually distributed.

Transform for Normality

To conduct a parametric hypothesis test, it is essential to satisfy the normality assumption for accurate results. This task has multiple ways of transforming data to a normal distribution. Three will be tested.

1. Log Transform
2. Square Root
3. Cube Root

## **HYPOTHESIS 1: Kruskal Wallis Test**

Kruskal Wallis test is also known as non-parametric ANOV. That means the means are tested between groups whose distribution is not necessarily typical. Kruskal Wallis test is robust to assumptions of variance. The next step is to find out if assumptions are met.

```

> ### Hypothesis test 1 #####
>
> #### Pearson corr Life Expectancy vs HealthExpenditure #####
> cor_result <- cor.test(dataset$LifeExpectancy, dataset$HealthExpenditure, method = "pearson")
>
> # Print the correlation coefficient and p-value
> cat("Correlation Coefficient:", cor_result$estimate, "\n")
Correlation Coefficient: 0.2517817
> cat("P-value:", cor_result$p.value, "\n")
P-value: 0.005535466
>
> # Set significance level (e.g., 0.05)
> alpha<-0.05
>
> # Compare p-value with significance level
> if (cor_result$p.value < alpha) {
+   cat("The hypothesis test is statistically significant. We reject the null hypothesis.\n")
+ } else {
+   cat("The hypothesis test is not statistically significant. We fail to reject the null hypothesis.\n")
+ }
The hypothesis test is statistically significant. We reject the null hypothesis.

```

*Figure 15*

Assumptions of the test:

1. All samples are independent: Each observation was randomly selected independently. No two samples or observations were repeated
2. More than two categorical groups: (Europe, Africa)
3. Groups are factors: Groups are not numeric; they are factors, as seen in the report's exploratory data analysis section.
4. The dependent variable is continuous: The dependent variable is Average life expectancy, a constant variable.

In this task, mean life expectancy is the dependent variable; when the medians are being considered, it is essential to ensure that all the groups have the same variance; however, for this task, we do not need to have the same variances.

#### Kruskal Willis & Wilcoxon

P-value < 0.05, at a 5% significance level, the null hypothesis (mean life expectancy for all regions is the same) is rejected. The Kruskal test tells us that the null hypothesis is rejected but does not tell which areas have a different mean life expectancy; hence, another test, the Wilcoxon test, is performed to gain more insight.

**Interpretation:** As per the table above, all p-values are below 0.05; at a 5% significance level, we reject the null hypothesis; this means that the life expectancy in all regions is statistically significantly different.

**Kruskal Effect Size:** To measure how the value of mean LE is dependent upon the region, the Kruskal effect size function is used from the library (static); the confidence level at 95% shows that the effect size is 0.72 which means mean LE is highly dependent upon the Group.

#### Kruskal Effect size

Hypothesis 2: Mann-Witney\_Wilcoxon Test of Hypothesis

Although the meaning of LE in Europe and Asia appears different in Density plots of LE, there is a significant overlap of values given this information. Hypothesis 2 is formed to consider whether the difference in mean LE between Africa and Europe is significantly different using the Mann Whitney Wilcoxon Test; this is a non-parametric alternative to the parametric Student t-test.

Assumptions: Samples are independent

$H_0$  = there is no significant difference in the means of life expectancy between Africa and Europe.

$H_0$ : ...Europe = ... Africa

$H_1$  = The mean life expectancy in the three regions is significantly different.

A New Data frame is created by removing Africa...

Since  $p < 0.01$ , there is a visible statistically significant difference between the life expectancies in Europe and Africa. Hence, the null hypothesis is rejected.

## DISCUSSION

Analyzing statistical and exploratory data provided an intriguing viewpoint on the socioeconomic development of many nations. The log, square root, or cube root transforms could not be used to establish normalcy because of the nature of the data belonging to specific countries. Regression normalization would also make it more challenging to understand the equation. Time constraints resulted in non-gauss and non-parametric data. The Wilcoxon and Kruskal Wallis tests were applied. Non-parametric tests are more likely to result in type 2 errors than parametric tests because the former are more likely to reject the null hypothesis in cases where the distribution is not Gaussian. When data is normally distributed, parametric tests are more dependable than non-parametric tests (Soetewey, 2020).

Even though independent variables were included to create the regression models, some might not accurately reflect life expectancy. One regression line cannot be fitted to forecast all regions with the same level of accuracy due to outliers and extreme values. Because these measures naturally rise as the years go by, it was challenging to discover clear linear correlations for both the dependent and independent variables per country or region due to the nature of the dataset. In order to determine whether repeating this study yields better results, a time-series regression model will preferably be fitted, and more countries and areas will be chosen to provide a more accurate picture of the link between LE and other parameters.

It is crucial to remember that while there is a substantial correlation between life expectancy and government health spending, undernourishment, and basic sanitation, this correlation should not be confused with a causal relationship that may exist but is outside the purview of this study (Madhavan, 2019).

It would be fascinating to predict each country's average LE about the regional mean LE; unfortunately, time constraints prevented that from being done. A far more thorough investigation would be needed. The ARIMA model performed well and was approved. Ljung-Box and ACF tests, but if the same test were carried out after dividing the data into test and train

sets, it would be more appropriate to state that ARIMA is the best model for forecasting life expectancy.

## **6. CONCLUSION**

Although there has been an overall increase in average life expectancy between 2011 and 2022, it should be emphasized that average life expectancy varies significantly between locations. A statistical study determined that Europe has the longest life expectancy while Africa has the lowest. The Mann-Whitney Wilcoxon test and Kruskal Wallis, two non-parametric techniques, demonstrated that this difference was statistically significant.

Since there is low collinearity and a strong connection between LE and these indicators, life expectancy may best be described as a consequence of basic sanitation and government health expenditure. Because of its smaller sum of squares, the time series RIMA model has proven to be an effective predictor of LE for short-term projections.

Because the time series RIMA model has lower sum squared errors and ACF values, it effectively predicts LE for short-term future forecasts. In order to assess how the wealth and standard of living of a nation affect life expectancy and to compare findings with those of other studies, it could be a good idea for future research to incorporate additional variables like the PPP, Gini index, and political stability metrics.

## Part Two Introduction: **Interactive Dashboard Design**

### **1. Introduction**

Longevity has increased in the last few decades with the advent of antibiotics, biomedical advances and other factors such as access to healthcare, education, nourishment, genetics and lifestyle (Zhaurova, 2008); however, there is still significant inequality in life expectancy amongst different regions of the world.

This task aims to investigate this inequality using an interactive dashboard that effectively illustrates the disparities between the average life expectancy at birth in Europe and Africa. It also explores the underlying relationships between the target variable—life expectancy at birth—and other socioeconomic factors, such as the percentage of GDP per capita that goes toward health spending, the frequency of undernourished individuals using essential sanitation services, etc. The dashboard's objectives include forecasting the value of LE in the upcoming years and presenting these relationships and disparities in the most aesthetically pleasing style possible. The goal of the Sige page dashboard, which was made with Power BI and clever application of gestalt concepts and pre-attentive attribute information, is to tell a narrative.

### **2.1 Background Research**

Today, a vast amount of data is available for free. According to Stephen Few, an expert in dashboard design, it is incumbent upon us to find ways to represent this data in a way that would make sense to the audience.

"A Power BI dashboard is succinctly a visual display of the essential information that needs to achieve several objectives, consolidated and organized on a single screen so the information can be visualized at a glance." (Few, Dashboard Confusion, 2004)

The human brain has a working memory that can hold limited information in one instance. Working memory is where information perceived through vision is transferred for processing; therefore, whilst creating a display, it is crucial to consider pre-attentive attributes to keep the dashboard easily understood by the public. According to Baskett et al., dashboard design methodology should be in the following order: identifying the objectives of the dashboard, creating a hierarchy, determining the look and feel and developing the dashboard (Bskett et al., 2008).

A study focused on understanding how people scan visuals summarized that F and Z patterns are amongst common patterns in checking pages from top left through to right and then going down as in the letter "F". Therefore, any visual hierarchy of dashboard design should be based on eye movement when exploring a page (Moga, 2020). It is recommended to display critical information on the top left corner of the dashboard since that is the first focal point. Based on an article by Tableau, some of the best practices to make this dashboard include understanding your audience, eliminating clutter, including interactivity for more engagement and reducing the number of colours (10 best practices for Building Effective Dashboards, 2022).

The purpose of graphs is to take work from our brain to our eyes (Few, Dashboard design for at-a-glance monitoring, 2010). Pre-attentiveness means the ability of a visual to be understood



without requiring much effort from the audience (The Principles of Visual Design for Dashboards, 2021). Pre-attentive attributes are essential as they do not require special attention from the user and are easily absorbed in working memory. Gestalt principles, in conjunction with pre-attentive attributes such as hue and size, have been used in making the dashboard for this report.

Information about differences in LE trends for different groups is easy to follow and best presented as a line chart, a good example of which is visual by Our World in Data (Life expectancy at birth including the UN Projections, 2022), another eye-catching visual by Stephanie included the LE vs Income graph where graph titles are named to stir the audience's emotion and increase engagement according to this paper, titles of graph can impact the interest of the user (Arevalo & Devan, 2017).

Using the to differentiate a value helps a user search it far quicker than any other attribute (Few et al. for a glance monitoring, 2010). To avoid incorrect interpretations, bar charts are used since they allow users with elementary graph reading knowledge to easily spot the most significant or highest value on top of the left value at the bottom, as seen in the visuals of this dashboard.

According to the Big Book of Dashboard design, using strategies such as keeping white space and highlighters, keeping colour blind-suited displays such as green and blue instead of red and green and applying filters/slicers allows for easy exploration of the dashboard elements. This sparks the user's interest (Wexler et al., 2017).

As per the World Bank, the countries in regions selected for this project fall into the categories of high-income and Low-income countries, respectively (Databank, n.d). Since access to healthcare, education and nourishment is directly linked to income group, this may explain the stark difference in LE amongst these regions (Kpolovie et al., 2016). Section 4 of this report explains how the abovementioned research is exercised in making the dashboard.

### **3. Exploration of Data Set**

Downloaded from the World Bank website (DataBank, n.d.), the dataset spans 12 countries and 12 series between 2011 and 2022. Because there was not enough data available for the years chosen, the selection process was based on geography. The research aims to investigate the various global indicators that affect life expectancy; series are chosen in large quantities for most European and African nations.

The grouping of countries is shown in Table 1.

Region	Countries
Europe	<ul style="list-style-type: none"> <li>• France</li> <li>• Austria</li> <li>• Ireland</li> <li>• United Kingdom</li> <li>• Denmark</li> <li>• Hungary</li> </ul>
Africa	<ul style="list-style-type: none"> <li>• Rwanda</li> <li>• Ethiopia</li> <li>• Liberia</li> <li>• Mozambique</li> <li>• Sierra Leone</li> <li>• Senegal</li> </ul>

*Table 1 Regions and Countries*

Attributes named in Dashboard	Definition
<b>Life Expectancy</b> Life expectancy at birth, total (years)	Lifespan of an individual
<b>Undernourishment</b> Prevalence of undernourishment (% of population)	Populations unable to sustain healthy life.
<b>Health Expenditure (% GDP)</b> Domestic general government health expenditure (% of GDP)	Govt. expenditure on health from domestic sources as a share of the economy as measured by GDP.
<b>Basic sanitation (%population)</b> People using basic sanitation services (% of population)	Percentage of people using basic sanitation services unshared with other households.
<b>Basic Drinking Water (%population)</b> People using basic water services (% of population)	Percentage population using basic water services.

*Table 2 Attributes and Definitions*

### Nature Of Data Attributes:

Years and nations are the categorical attributes in this dataset; the remaining values are continuous numerical values. The aim variable is life expectancy, which is related to basic sanitation, healthcare, and nutrition, among other things (Allel et al., 2022). The association this study examines should be distinct from the topic of possible causes, which is outside the purview of this investigation.

## 2.3 Investigation of Data Workflows & Proposal for the Design of Dashboard

Data Workflow: Steps in data workflow follow as illustrated in the pattern below:

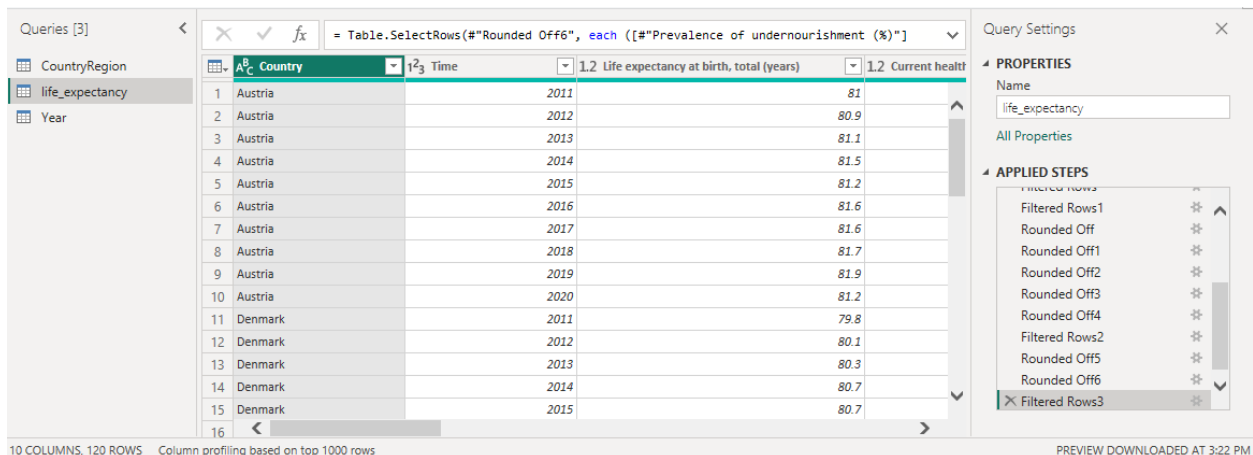
Background Research on Topic= Attribute selection =Data acquisition =Data cleansing=Planning of layout= Dashboard development and visualization

### Workflow Summary

Steps showing Workflow for Dashboard Design:

1. Background studies on LE and pertinent global inequities and indicators
2. Attribute selection, grounded in the study
3. Data acquisition: columns representing nations and time (years) as rows, with data downloaded from Databank.org.
4. Data purification and analysis: Power Query Editor verifies and removes missing values.
5. Planning Layout according to industry standards. It took a lot of trial and error to find the best layout based on the research; several layout changes were made.
6. Visualization: Visuals are made with Power BI.

**Pre-processing data in Power Query:** Following import, data is cleaned for appropriate usage, which entails several procedures, including but not limited to.

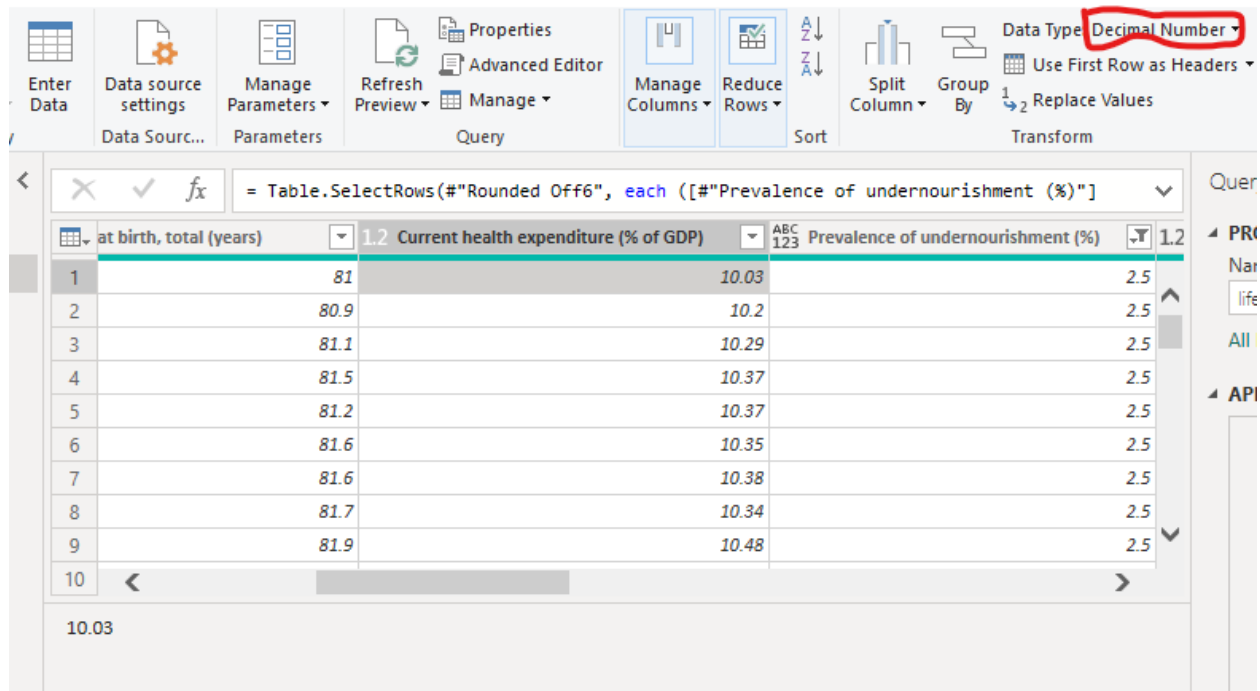


	Country	Time	1.2 Life expectancy at birth, total (years)	1.2 Current health
1	Austria	2011	81	
2	Austria	2012	80.9	
3	Austria	2013	81.1	
4	Austria	2014	81.5	
5	Austria	2015	81.2	
6	Austria	2016	81.6	
7	Austria	2017	81.6	
8	Austria	2018	81.7	
9	Austria	2019	81.9	
10	Austria	2020	81.2	
11	Denmark	2011	79.8	
12	Denmark	2012	80.1	
13	Denmark	2013	80.3	
14	Denmark	2014	80.7	
15	Denmark	2015	80.7	

### Diagram showing applied steps in Power Query Editor

**Correct Data Type:** The data type for the year was reviewed from the whole number to date, the categorical variable 'Country' was kept as text, and all the other continuous variables were changed to decimal number types.

**Change Decimal Places:** LE was rounded off to 1 decimal place. Other attributes to 2 decimal places for simplicity. This decision is justified since the numbers being compared already have significant differences.



**Figure 3 Change decimal**

**Grouping:** Power BI ‘grouping’ feature groups the column countries into Europe and Africa. Countries were manually selected and grouped based on location.

Grouping

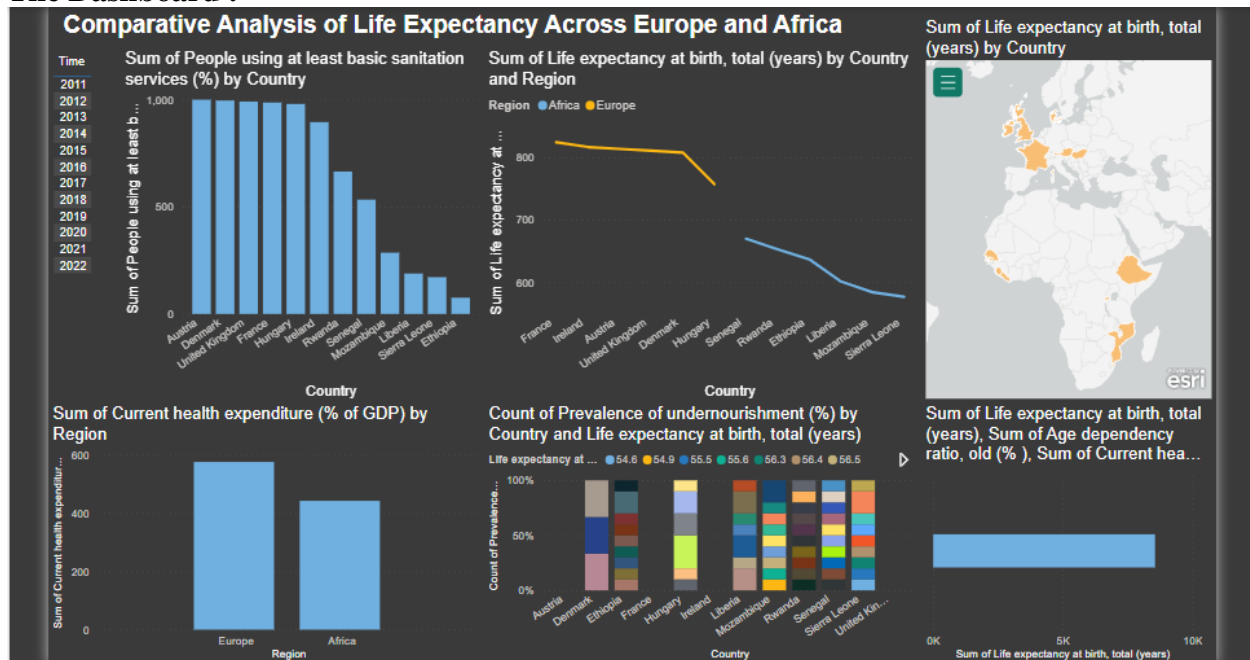
Country	Region
Austria	Europe
Denmark	Europe
Hungary	Europe
Ireland	Europe
France	Europe
United Kingdom	Europe
Sierra Leone	Africa
Ethiopia	Africa
Mozambique	Africa
Liberia	Africa
Rwanda	Africa
Senegal	Africa

*figure 5*

**Tools tips:**

Figure 6 shows tooltips for visual 2, which include all the important measures that impact LE. Tools tips help analyze information while keeping everything in an eye span whilst reducing clutter. (Wexler, Shaffer, & Cotgreave, 2017)

## The Dashboard :



## Comparative Analysis of Life Expectancy in Europe and Africa

From the top, left underneath the title is the yearly slicer visual that filters all the other visuals in the dashboard for detailed exploration.

## A brief overview of the visual paradigm:

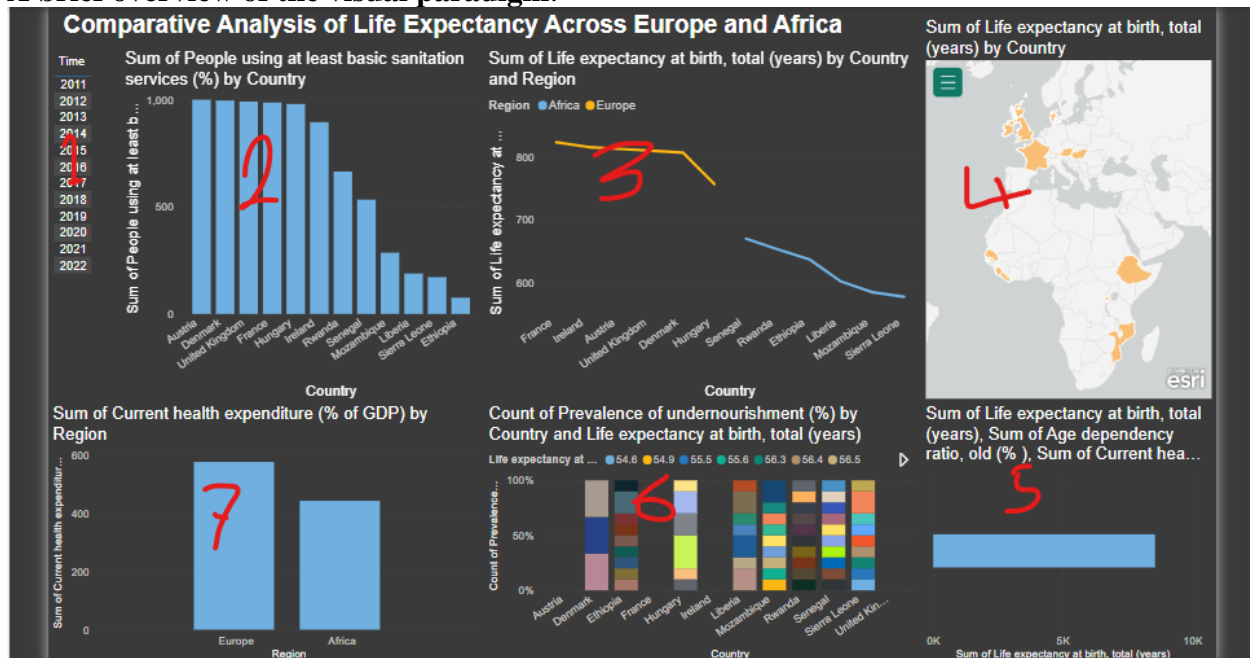


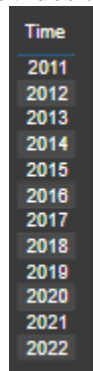
Figure 8 Labels for Dashboard Visuals

For ease of reference, Figure 8 has been created, and it will be used to describe each visual throughout the remainder of this report. There is a cyclical structure to the layout. A slicer called Visuals1 makes it possible to choose data specific to that year. The average life expectancy at birth is summarized in Visual 3, which also briefly introduces the story by placing all of the selected countries in the upper left.

In contrast to cyan and black (used for other bars), which highlight countries on a shallow spectrum of developmental indicators, the orange hue of visual encoding highlights the nation with the lowest indicator value. According to Stephen Few's books, colour is the first unconsciously seen thing by the human eye.

### **Year Slicer**

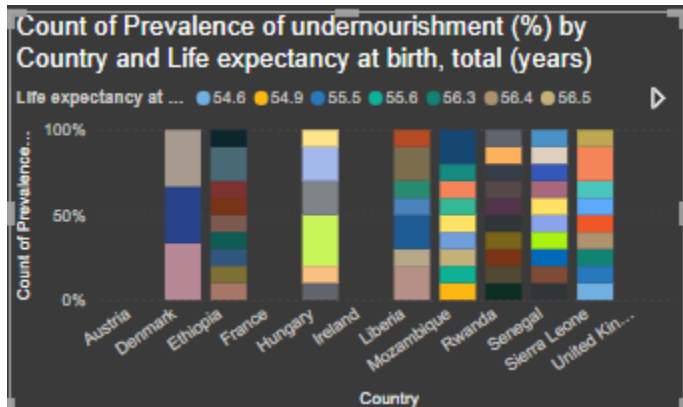
Year Slicer is right underneath the title to display which time frame and range the data refers to. On clicking the relevant year, it only shows data for that year and highlights the other visuals. It also gives an option to 'Select all ', which provides average values for all-inclusive years.



*Figure: Year Slicer*

### **Undernourishment Impact on Life Expectancy**

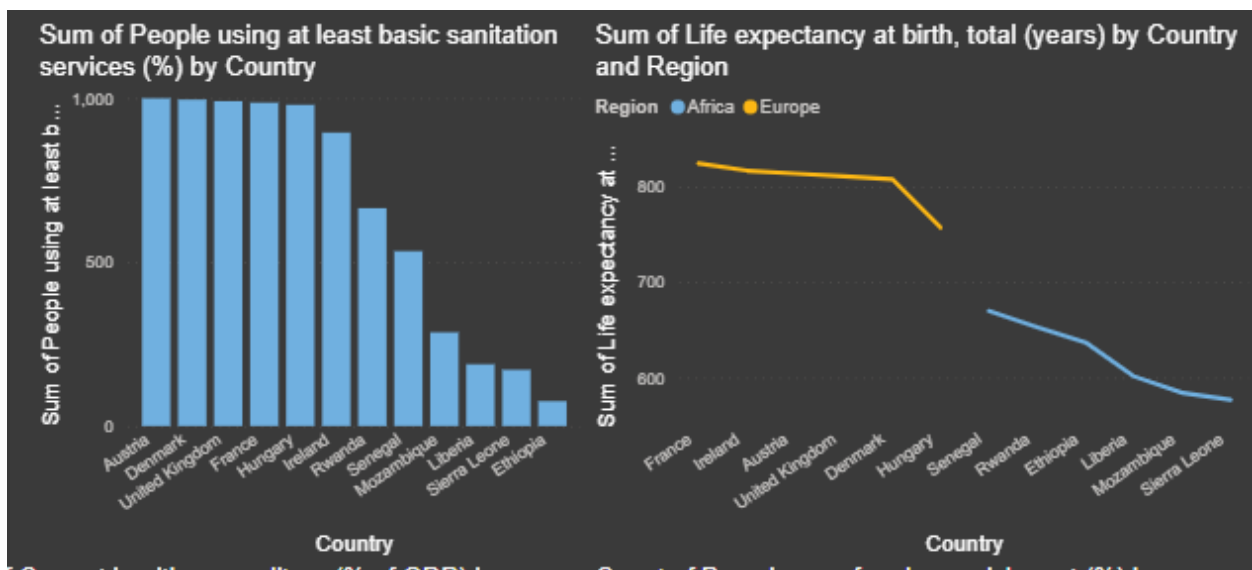
The combination chart is used in this image to illustrate the connection between LE and the frequency of undernourishment. The nation is a categorical attribute with an x-axis. Undernourishment is less common in continuous variables. This graphic efficiently conveys a lot of information; according to the analysis, Liberia has the highest percentage of undernourishment. Hue is used to indicate this. In this graphical representation, Sierra Leone stands out because, while its frequency of undernourishment is marginally lower than that of Liberia, its LE is noticeably lower. This is because of the 2014 Ebola outbreak, which lasted for two years until 2006.



### Sanitation Services used by the percentage of the total population

One of the pre-attentive characteristics of images is size. The spread shows the proportion of the population using essential sanitation services around a particular region. Maps provide a spatial or geographical location and combine measurable or numeric variables/metrics (in this case, the percentage of the population using basic sanitation) to create a more visually appealing visual (Visualization types in Power BI, 2022).

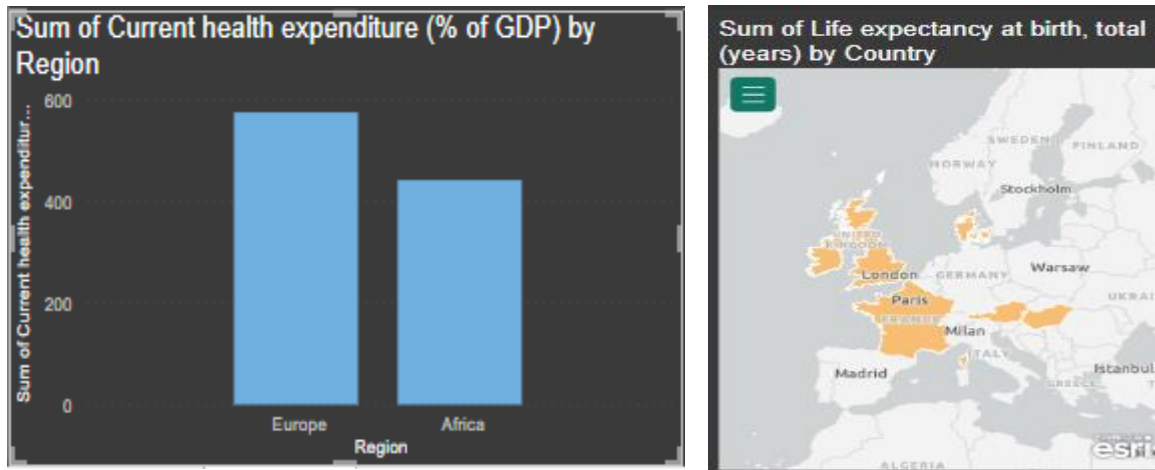
The bubble chart is utilized because there is a significant disparity between the percentages of people in Europe and Africa who use basic sanitation, as seen by the bubble sizes; the graph thus makes these discrepancies quite evident.



**Visual 7: What is the influence of Health Expenditure on Life Expectancy?**

The question posed in the headline of this image is intended to pique viewers' interest and maintain audience engagement. For the same reason as in graphic 5, a combination chart was chosen to illustrate the relationship between LE and government spending on healthcare as a % of GDP. The health expense percentage is shown on the left chart, and the life expectancy of various nations is shown on the correct chart. This graphic demonstrates that LE is higher in countries where the government spends more of the GDP on the healthcare system. LE and healthcare have a close association, but it is not casual (Allel, Hajizadeh, & Kiadaliri, 2022). African nations spend the least amount of money on healthcare.

**Figure 19 Health expenditure and life Expectancy:**



#### 4. Discussion

The dashboard style and content decisions can benefit from feedback-based research for information comprehension and design. A rough dashboard sketch was created following an in-depth study on birth weight and its relationship to undernourishment, medical costs, access to basic sanitation and drinking water, and best practices in dashboard design. With the additional literature reading, this design changed until it took on its ultimate shape. Visual two was filtered using Power BI's grouping feature to colour-code regions for easy comparison. Since bar and line charts are the most readable and appropriate for displaying data properties in this way, they were primarily utilized.

#### 5. Conclusion

There are notable disparities in life expectancy between Europe and Africa. This narrative needs to be told to draw attention to significant disparities in socioeconomic indicators and to urge action to encourage socioeconomic growth. Dashboards with just one-page help give viewers crucial information. The graphics on the dashboard are arranged according to priority. The



dashboard created by applying Gestalt principles was clear, concise, and had a minimalistic theme, which satisfied all of the project's goals.

## REFERENCES

10 Best Practices for building Dashboards. (2022).

<https://www.tableau.com/learn/whitepapers/10-best-practices-building-effective-dashboards>

Allel, K., Hajizadeh, M., & Kiadaliri, A. (2022, June 8). The gap in life expectancy and lifespan inequality between Iran and neighbor countries; the contributions of avoidable causes of death. *International journal for Equity in Health*. <https://doi.org/10.1186/s12939.022.01683.8>.

Arevalo, S., & Devan, A (2017). Unveiling storytelling and visualization of data. 14<sup>th</sup> Student Colloquium University of Groningen, Groningen.

<https://www.researchgate.net/publication/317236927> Unveiling storytelling and visualization\_of\_data

Baskett, L., Lerouge, C., & Tremblay, M.C. (2008, September). Using the dashboard technology properly Health Progress, 89(5), 16-23.

DataBank, (n.d.). The World Bank: <https://databank.worldbank.org/metadataglossary/world-development-indicators/series/>

Eberhart, M. (2021). Average Life Expectancy Across Nations. Proceedings of the Jepson Undergraduate Conference on International Economics, 3(3)

<https://scholarworks.com.uni.edu/jucie/vol3/iss1/3>

Few, S. (2004, March 20). Dashboard Confusion. Intelligent Enterprise

Few, S. (2010). Dashboard Design for at-a-glance monitoring. Perceptual Edge

Zhaurova, K. (2008). Genetic causes of adult-onset disorders. Nature Education, 1(1):49.

<https://www.nature.com/scitable/topicpage/genetic-causes-of-adult-onset-disorders-34609/>

Regression Analysis on Life Expectancy. (2019, December 13). Hackernoon.

<https://hackernoon.com/regression-analysis-on-on-life-expectancy-4wp34rf>

Maity, A., Rheman, E., & Sanders, E. (2017). Factors Explaining Average Life Expectancy: An Explanation Across Nations.