Life Expectancy Data Research

Justin Rzepko

```
data <- read.csv("1a-lifeExpect.csv")
data_missing <- read.csv("1a-lifeExpectMissing.csv", na.strings = c("", ".", "*"))

str(data_missing)
```

```
## 'data.frame':    40 obs. of  6 variables:
##  $ Country               : chr  "Argentina" "Bangladesh" "Brazil" "Canada" ...
##  $ Life.Expectancy       : num  70.5 53.5 65 76.5 70 71 60.5 51.5 78 76 ...
##  $ People.TV             : num  4 315 4 1.7 8 5.6 15 503 2.6 2.6 ...
##  $ People.Dr             : int  370 6166 684 449 643 1551 616 36660 403 346 ...
##  $ Female.Life.Expectancy: int  74 NA 68 80 72 74 61 NA 82 79 ...
##  $ Male.Life.Expectancy  : int  NA 54 62 73 68 68 60 50 NA NA ...
```

```
data$People.TV <- as.numeric(data_missing$People.TV)
```

```
model <- lm(Life.Expectancy ~ People.TV + People.Dr, data = data_missing)

if (!require(car)) install.packages("car")
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
library(car)

# Calculate VIF values
vif_values <- vif(model)
print(vif_values)
```

```
## People.TV People.Dr
##  1.623494  1.623494
```

Both predictors (People.TV and People.Dr) have VIF values around 1.24, which are well below the commonly used thresholds for concern (5 or 10).

The results confirm that People.TV and People.Dr are sufficiently independent in this dataset to be valid predictors in the regression model. This strengthens the model's reliability and interpretability.

```
# Finding best method to use to estimate the missing values
colSums(is.na(data_missing))
```

```
##                Country        Life.Expectancy              People.TV
##                      0                      0                      2
##             People.Dr Female.Life.Expectancy   Male.Life.Expectancy
##                      0                      7                      6
```

```
model_life_exp <- lm(Life.Expectancy ~ People.TV + People.Dr, data = data_missing, na.action = na.exclu
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ People.TV + People.Dr, data = data_missing,
##     na.action = na.exclude)
##
## Coefficients:
## (Intercept)    People.TV    People.Dr
##   70.251957    -0.023495    -0.000432
```

```
summary(model_life_exp)
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ People.TV + People.Dr, data = data_missing,
##     na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2894  -4.6266   0.3977   5.0872   9.0535
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.2519573  1.0877047  64.587   <2e-16 ***
## People.TV   -0.0234954  0.0096469  -2.436   0.0201 *
## People.Dr   -0.0004320  0.0002023  -2.136   0.0398 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.003 on 35 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.44,  Adjusted R-squared:  0.408
## F-statistic: 13.75 on 2 and 35 DF,  p-value: 3.916e-05
```

The model predicts Life.Expectancy using People.TV and People.DR as predictors:

The model predicts `Life.Expectancy` using `People.TV` and `People.DR` as predictors:

$$Life.Expectancy = \beta_0 + (\beta_1 \times People.TV) + (\beta_2 \times People.DR)$$

We are assessing the suitability of Regression Imputation as a method for predicting the missing values in our dataset by evaluating the strength of the correlation between Life Expectancy and the People Per TV and People Per Doctor ratios. In this analysis, we focus on two key metrics: the Residual Standard Error and the R-squared value.

The Residual Standard Error of 6 indicates that, on average, the predicted values deviate from the observed values by 6 years. Meanwhile, the R-squared value of 0.44 suggests that the model explains 44% of the variability in Life Expectancy. However, this leaves 56% of the variability unexplained, implying that additional

factors influencing life expectancy are not captured by the current model. This highlights the potential need for incorporating other relevant variables to improve predictive accuracy.

Given that regression imputation is not suitable, we will replace the missing values in Male Life Expectancy and Female Life Expectancy with the mean Life Expectancy for each country. This approach is simple, fast, and ensures the dataset is complete without introducing significant biases. However, it does have limitations, as it ignores variability in the data and may distort the distribution of the imputed variables.

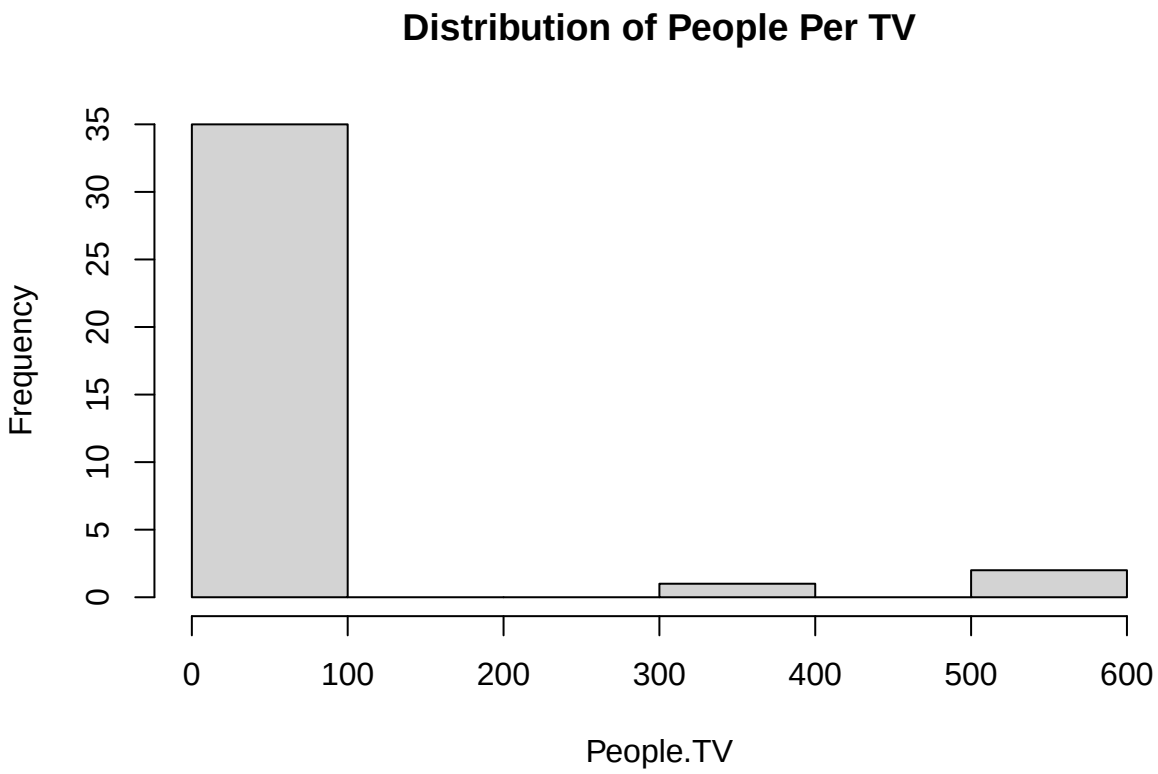
```
# Loop through each row to replace missing values
for (i in 1:nrow(data_missing)) {
  # Check for missing Male Life Expectancy
  if (is.na(data_missing$Male.Life.Expectancy[i])) {
    # Replace with the average Life Expectancy for that country
    data_missing$Male.Life.Expectancy[i] <- round(data_missing$Life.Expectancy[i])
  }

  # Check for missing Female Life Expectancy
  if (is.na(data_missing$Female.Life.Expectancy[i])) {
    # Replace with the average Life Expectancy for that country
    data_missing$Female.Life.Expectancy[i] <- round(data_missing$Life.Expectancy[i])
  }
}
```

```
summary(data_missing$People.TV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.30    3.35    6.30   51.98   23.00  592.00       2
```

```
hist(data_missing$People.TV, main = "Distribution of People Per TV", xlab = "People.TV")
```



Given that there are only two missing values in the People per TV column, we are evaluating whether to use the mean or median to fill in these missing values. The mean would be a better choice if the distribution is symmetric, as it uses all the data points and provides a more accurate estimate of the central tendency in such cases. However, based on the histogram of the People per TV column, we observe that the distribution is highly skewed. In this situation, the median is the more appropriate option, as it is robust to outliers and better represents the central tendency for skewed data. By using the median, we ensure that the imputed values do not disproportionately influence the data set or distort its overall structure.

```
median_people_tv <- median(data_missing$People.TV, na.rm = TRUE)

data_missing$People.TV[is.na(data_missing$People.TV)] <- median_people_tv
```

```
numerical_data <- data_missing[, sapply(data, is.numeric)]

summary(numerical_data)
```

```
##  Life.Expectancy   People.TV        People.Dr       Female.Life.Expectancy
##  Min.   :51.50   Min.   :  1.30   Min.   :  226.0   Min.   :52.00
##  1st Qu.:61.00   1st Qu.:  3.65   1st Qu.:  472.2   1st Qu.:63.00
##  Median :69.50   Median :  6.30   Median :  990.5   Median :72.00
##  Mean   :67.04   Mean   : 49.70   Mean   : 3997.7   Mean   :69.35
##  3rd Qu.:73.38   3rd Qu.: 23.00   3rd Qu.: 3193.2   3rd Qu.:77.25
##  Max.   :79.00   Max.   :592.00   Max.   :36660.0   Max.   :82.00
##  Male.Life.Expectancy
##  Min.   :50.00
```

```
##  1st Qu.:59.75
##  Median :66.00
##  Mean   :65.00
##  3rd Qu.:70.25
##  Max.   :79.00
```

```
categorical_data <- data_missing[, sapply(data, is.character) | sapply(data, is.factor)]
summary(categorical_data)
```

```
##    Length     Class      Mode
##        40 character character
```

```
mean(data_missing$Life.Expectancy, na.rm = TRUE)
```

```
## [1] 67.0375
```

On average, individuals in the dataset have a life expectancy of approximately 67 years.

```
median(data_missing$Life.Expectancy, na.rm = TRUE)
```

```
## [1] 69.5
```

The median life expectancy is slightly higher than the mean(69.5 years), suggesting that the distribution of life expectancy may be left-skewed. This central value is a strong measure of the dataset's typical life expectancy because it it less influenced by outliers compared to the mean.

```
sd(data_missing$Life.Expectancy, na.rm = TRUE) # Mean
```

```
## [1] 8.248844
```

The standard deviation of life expectancy in the data set is approximately 8.25 years. In this data set, most life expectancy values are within 8.25 years of the mean (67 years). A standard deviation of 8.25 indicates that life expectancy in this data set show a moderate spread around the mean. Some regions or groups might have significantly lower or higher life expectancy, contributing to this variability.

```
range(data_missing$Life.Expectancy, na.rm = TRUE) # Mean
```

```
## [1] 51.5 79.0
```

The range of life expectancy is from 51 years to 79 years, covering a span of 28 years. This indicates that there is a wide varitaion in life expectancy across the data set, possibly reflecting differences in healthcare, living conditions, or other socio-economic factors within the populations represented.

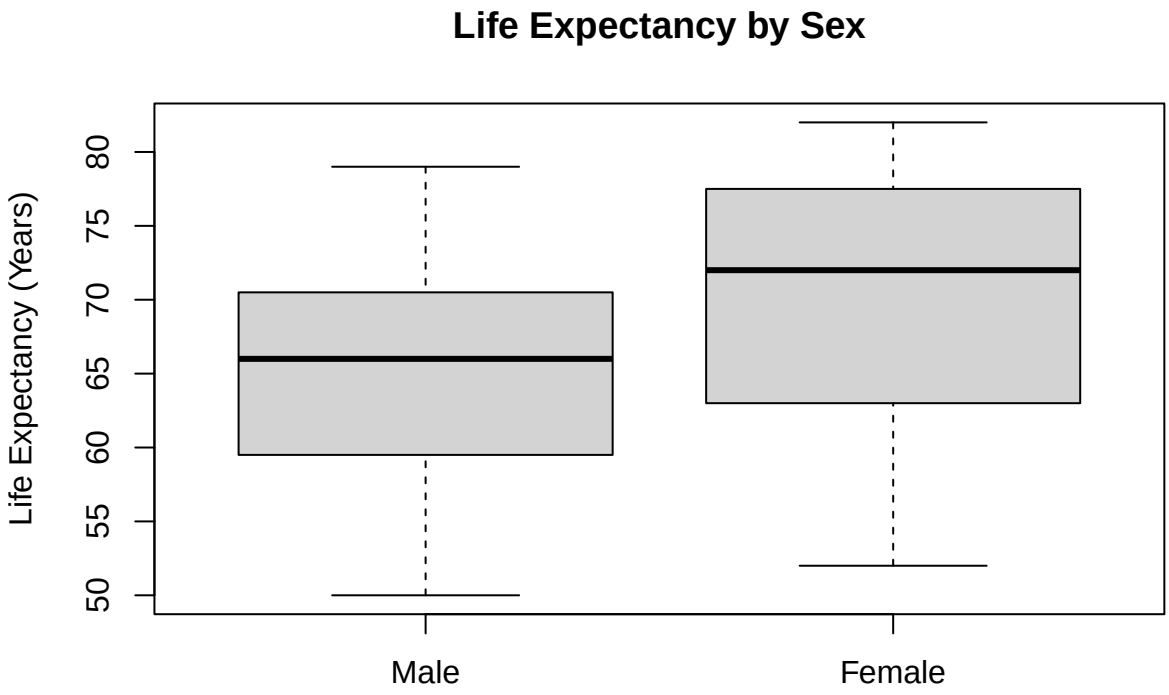
Research Question 1: Do life expectancies differ by sex?

```
data_missing$People.TV <- as.numeric(as.character(data_missing$People.TV))
data_missing$People.DR <- as.numeric(as.character(data_missing$People.Dr))

data_missing$Life.Expectancy[is.na(data_missing$Life.Expectancy)] <- mean(data_missing$Life.Expectancy,
data_missing$People.TV[is.na(data_missing$People.TV)] <- mean(data_missing$People.TV, na.rm = TRUE)
```

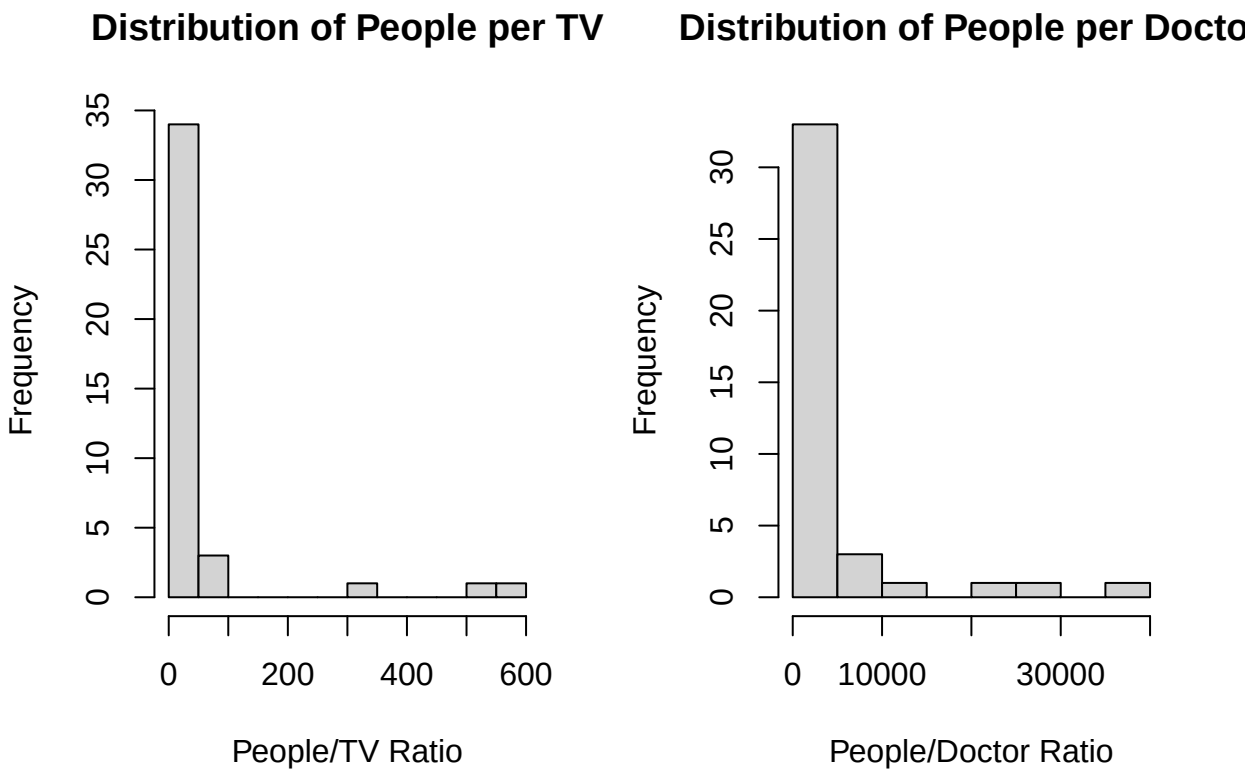
```r
data_missing$People.DR[is.na(data_missing$People.Dr)] <- mean(data_missing$People.Dr, na.rm = TRUE)

# Boxplot for life expectancy by sex
boxplot(data_missing$Male.Life.Expectancy, data_missing$Female.Life.Expectancy,
        names = c("Male", "Female"),
        main = "Life Expectancy by Sex",
        ylab = "Life Expectancy (Years)")
```



The median life expectancy for females is higher than for males, as seen by the central line within each box. The box represents the middle 50% of the data(IQR). Males have a slightly smaller IQR(width of the box) compared to females, suggesting less variability in male life expectancy within the middle 50% of the data compared to females. The whiskers suggest that both males and females have a similar lower bound for life expectancy, but the upper bound is higher for females. Overall, females not only have a higher median life expectancy than males, but also show more variability in the middle range of the data.

```r
par(mfrow = c(1, 2))
hist(data_missing$People.TV, main = "Distribution of People per TV", xlab = "People/TV Ratio", breaks =
hist(data_missing$People.Dr, main = "Distribution of People per Doctor", xlab = "People/Doctor Ratio", b
```




```
par(mfrow = c(1, 1))
```

The histogram on the left(distribution of people per TV) represents the frequency distribution of the People/TV ratio (number of people per TV). The distribution is heavily right-skewed, with most data points concentrated near the lower end of the ratio(close to 0). This indicates that the majority of countries in the data set have a relatively low number of people per TV, suggesting high access to televisions. A few bars on the right represent countries with very high People/TV ratios, indicating the limited access to televisions in these regions.

The histogram on the right(distribution of people per doctor) shows the frequency distribution of the People/Doctor ratio(number of people per doctor). Similar to the People/TV ratio, the distribution is heavily right-skewed, with most countries having low People/Doctor ratios which indicates good access to doctors. There are notable outliers on the far right, where some countries have extremely high People/Doctor ratios, reflecting very limited access to healthcare.

Comparison between the two histograms:

Both distributions are right-skewed, indicating that most countries have good access to resources while a minority have poor access. The People/Doctor ratio shows a wider spread compared to the People/TV ratio. This suggests that disparities in access to doctors are more pronounced than disparities in access to TVs.

Implications:

The skewed distribution suggest that while most countries enjoy relatively good access to both TVs and doctors, there are significant inequalities for a small number of countries. The outliers(countries with high ratios) warrant further investigation to identify potential barriers to healthcare or socioeconomic development. Efforts to reduce inequality in healthcare access should focus on countries represented by these outliers.

```r
# Paired t-test
t_test_result <- t.test(data_missing$Male.Life.Expectancy, data_missing$Female.Life.Expectancy, paired =
print(t_test_result)
```

```
##
##  Paired t-test
##
## data:  data_missing$Male.Life.Expectancy and data_missing$Female.Life.Expectancy
## t = -11.019, df = 39, p-value = 1.526e-13
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -5.148472 -3.551528
## sample estimates:
## mean difference
##           -4.35
```

$H_0 : \mu_0 = 0$ (The mean difference between male and female life expectancy is 0. (Null Hypothesis))
$H_a : \mu_a \neq 0$ (The mean difference is not equal to 0. (Alternate Hypothesis))

The test statistic t = -11.02 is a measure of how far the observed mean difference is from 0(the null hypothesis) in terms of the standard error. A large value of t(such as -11.02) indicates a strong deviation from the null hypothesis. We calculated this test statistic using a degrees of freedom(df) value of 39 which reflects the number of pairs in the data set minus one. Since there are 40 pairs of observations, df = 39. The p-value(.00000000000015), which is the probability of observing data, assuming the null hypothesis is true, is extremely small, essentially less than 0.0000001, which is far below the typical significance level of 0.05. This means the result is highly statistically significant, and we can reject the null hypothesis with confidence. The alternate hypothesis tested is that the true mean difference between male and female life expectancies is NOT equal to 0. Based on the p-value, we accept this alternative hypothesis. The 95% confidence interval for the mean difference is between -5.15 and -3.55 years. This means we are 95% confident that the true average difference in life expectancy(Female - Male) lies within this range. Since the interval does not include 0, it provides further evidence that there is a significant difference between male and female life expectancies. With the sample mean difference being -4.35 years, this indicates that, on average, females live 4.35 years longer than males in the data set. With all that being said, we can conclude that there is strong evidence that male and female life expectancies are significantly different. Specifically, females live significantly longer than males, with an average of approximately 4.35 years.

Research Question 2: Does TV or Doctor Ratio associate with life expectancies?

```r
cor_tv <- cor(data_missing$People.TV, data_missing$Life.Expectancy, use = "complete.obs")
print(cor_tv)
```

```
## [1] -0.5256639
```

The correlation coefficient ranges from -1 to 1. A value of -0.5256 indicates a moderate negative correlation between the number of people per TV and life expectancy. In other words, as the number of people per TV increases(indication fewer TVs per person or less access), life expectancy tends to decrease. This could suggest that greater access to television(a possible proxy for wealth or access to information) might be associated with higher life expectancy.

```r
cor_dr <- cor(data_missing$People.Dr, data_missing$Life.Expectancy, use = "complete.obs")
print(cor_dr)
```

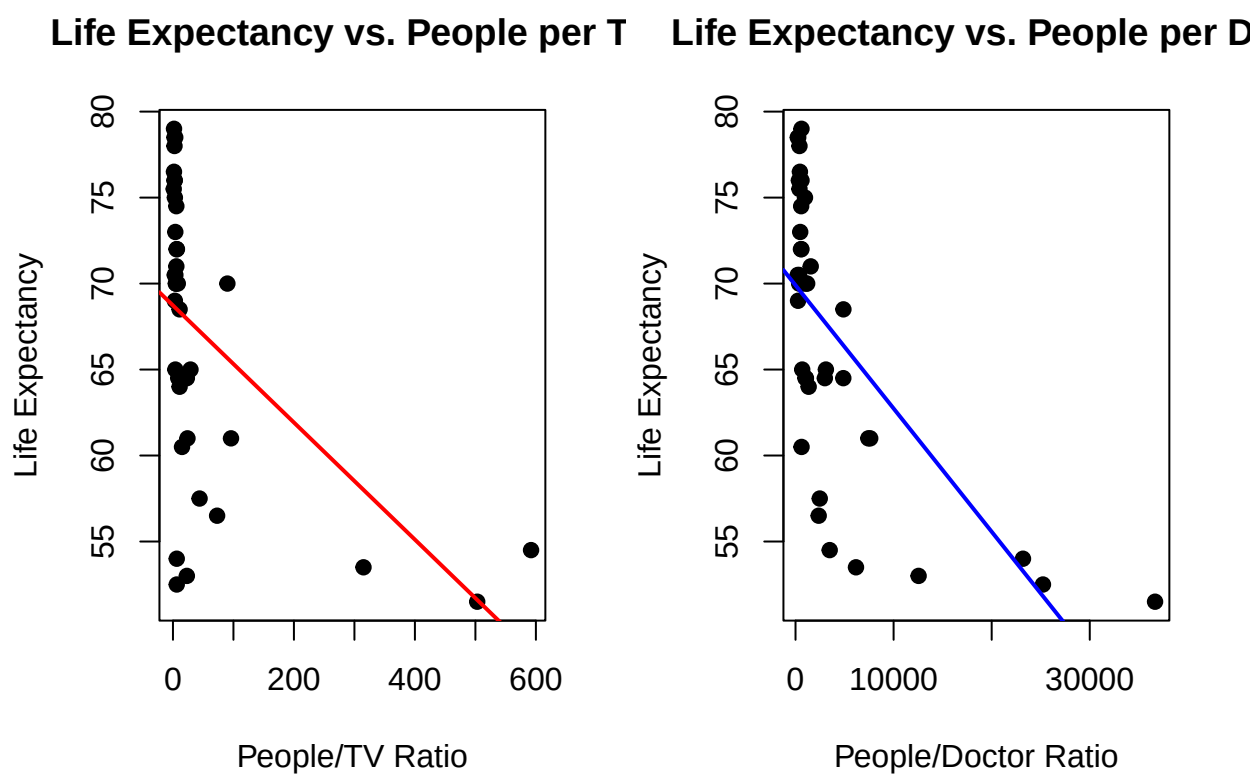
```
## [1] -0.6659967
```

The correlation coefficient between people per doctor and life expectancy has a value of -0.666. This indicates a stronger negative correlation between the number of people per doctor and life expectancy. In other words, as the number of people per doctor increases(indicating fewer doctors available per person), life expectancy tends to decrease significantly. This relationship highlights the importance of access to healthcare in influencing life expectancy.

From both, the correlation between people per doctor and people per TV with life expectancy, suggests that improving healthcare access and potentially socioeconomic conditions(as represented by TV ownership) could positively influence life expectancy. However, correlation does not imply causation so further analysis is needed to determine whether these variables directly impact life expectancy or if other factors are involved.

```
par(mfrow = c(1, 2))
# Scatterplots
plot(data_missing$People.TV, data_missing$Life.Expectancy, main = "Life Expectancy vs. People per TV",
     xlab = "People/TV Ratio", ylab = "Life Expectancy", pch = 19)
abline(lm(data_missing$Life.Expectancy ~ data_missing$People.TV), col = "red", lwd = 2)

plot(data_missing$People.DR, data_missing$Life.Expectancy, main = "Life Expectancy vs. People per DR",
     xlab = "People/Doctor Ratio", ylab = "Life Expectancy", pch = 19)
abline(lm(data_missing$Life.Expectancy ~ data_missing$People.Dr), col = "blue", lwd = 2)
```




```
par(mfrow = c(1, 1))
```

The scatterplot on the left shows the relationship between the People/TV ratio (x-axis) and Life Expectancy (y-axis). The red regression line indicates a negative relationship between People/TV ratio and Life Expectancy. As the People/TV ratio increases (fewer TVs per person), life expectancy tends to decrease. The

data points are moderately scattered around the regression line, reflecting a moderate negative correlation (r = -0.526). Higher access to TVs may serve as a proxy for improved socioeconomic conditions, which are positively associated with higher life expectancy.

The scatterplot on the right shows the relationship between the People/Doctor ratio (x-axis) and Life Expectancy (y-axis). The blue regression line shows a stronger negative relationship between People/Doctor ratio and life expectancy compared to the left plot. As the People/Doctor ratio increases (fewer doctors available per person), life expectancy sharply decreases. The data points are more tightly clustered around the regression line compared to the left plot, indicating a stronger correlation (r = -0.666). Limited access to healthcare has a significant impact on reducing life expectancy, highlighting the importance of medical infrastructure in improving longevity.

Comparison between the Two plots:

The People/Doctor ratio shows a stronger negative relationship with Life Expectancy compared to the People/TV ratio, as evident from both the tighter clustering of points and the steeper slope. While both access to TVs (indicating socioeconomic conditions) and access to doctors (indicating healthcare availability) impact life expectancy, healthcare access appears to have a larger and more direct effect. We can observe from these two plots that policies focusing on improving access to healthcare (e.g., reducing the People/Doctor ratio) would likely yield a more significant improvement in life expectancy than those targeting socioeconomic proxies like People/TV ratios. However, both factors are important to consider.

```
model <- lm(Life.Expectancy ~ People.TV + People.Dr, data = data_missing)
summary(model)
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ People.TV + People.Dr, data = data_missing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0493  -3.8473   0.4127   4.8232  11.8874
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.2858970  1.0726711  65.524  < 2e-16 ***
## People.TV   -0.0187027  0.0082903  -2.256 0.030072 *
## People.Dr   -0.0005801  0.0001377  -4.212 0.000156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.923 on 37 degrees of freedom
## Multiple R-squared:  0.5108, Adjusted R-squared:  0.4844
## F-statistic: 19.32 on 2 and 37 DF,  p-value: 1.798e-06
```

The model predicts Life.Expectancy using People.TV and People.DR as predictors:

The model predicts `Life.Expectancy` using `People.TV` and `People.DR` as predictors:

$$Life.Expectancy = \beta_0 + (\beta_1 \times People.TV) + (\beta_2 \times People.DR)$$

Response Variable (Dependent): Life.Expectancy
Predictor Variables (Independent): People.TV, People.Dr

The coefficients of the model are as follows:

The intercept is 70.29. This indicates that, when both People Per TV and People Per Doctor are 0, the predicted value of Life Expectancy is 70.29 years. However, in the context of this model, this scenario does not make practical sense. A People.TV value of 0 would imply that there are no people per TV (unrealistic in this context), and a People.Dr value of 0 would suggest infinite access to doctors, which is also unrealistic. Therefore, we will not be interpreting this into our findings, as it is essentially a mathematical artifact.

The coefficient for People Per TV is -0.0187. This mean that for every 1-unit increase in People Per TV(indicating fewer TVs per person), life expectancy decreases by 0.0187 years, assuming People Per Doctor remains constant. This relationship is statistically significant, with a p-value of 0.030

The coefficient for People Per Doctor is -0.000508. This indicates that for every 1-unit increase in People Per Doctor(indicating fewer doctors per person), Life Expectancy decreases by 0.000508 years, assuming People.TV remains constant. This relationship is highly statistically significant, with a p-value of less than 0.001.

The model's performance metrics provide further insights:

The Residual Standard Error is 5.923, which means that, on average, the predicted values of Life Expectancy deviate from the observed values by approximately 5.92 years.

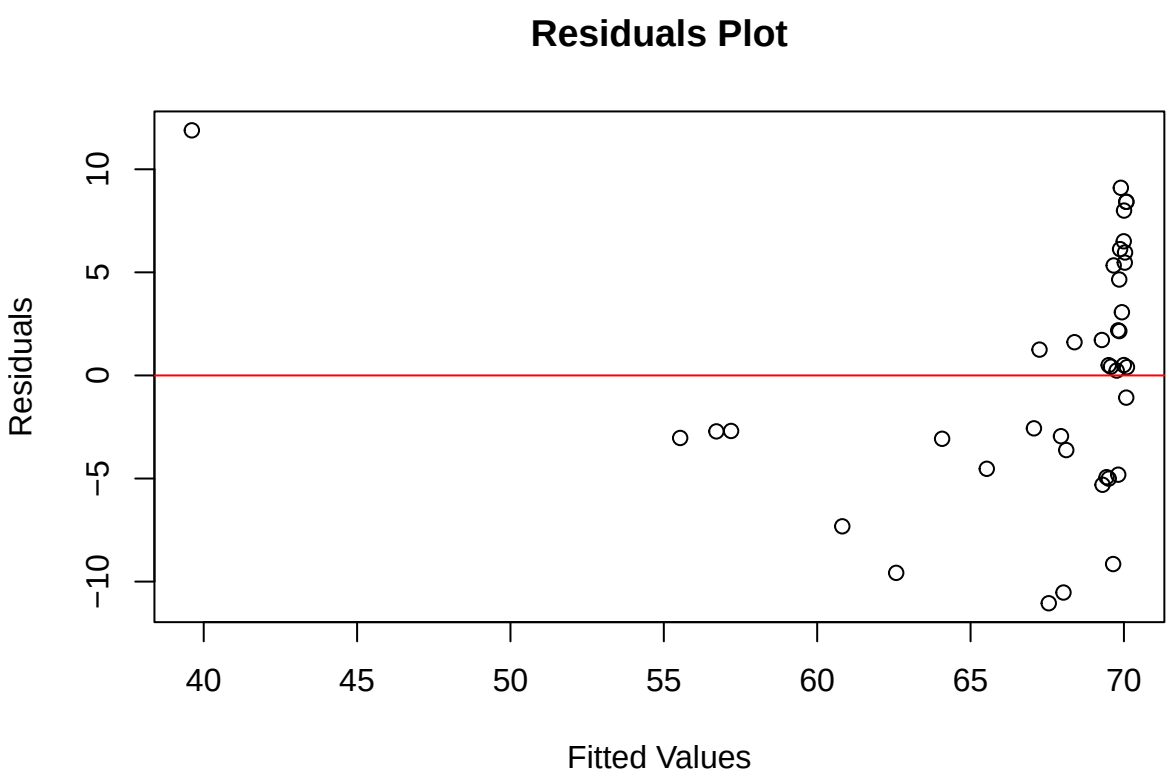
The R-squared value is 0.5108, indicating that the model explains about 51.08% of the variability in Life Expectancy.

The Adjusted R-squared value is 0.4844, accounting for the number of predictors in the model.

The p-value is 1.798e-6, showing that the overall model is statistically significant.

In summary, both predictors (People per TV and People per Doctor) have a significant negative association with Life Expectancy. The model moderately explains the variability in Life.Expectancy (51.08%), but additional predictors may be needed to capture the remaining unexplained variability (48.92%).

```
plot(fitted(model), model$residuals, main = "Residuals Plot", xlab = "Fitted Values", ylab = "Residuals
abline(h = 0, col = "red")
```



The Residuals vs. Fitted Values Plot in this context is used to evaluate how well the regression model predicts Life Expectancy based on the predictors People per TV and People per Doctor ratios. **Response Variable (Dependent):** Life.Expectancy **Predictors (Independent):** People.TV (people per TV) and People.Dr (people per doctor). **Purpose:** This plot assesses the assumptions of the regression model, such as linearity, constant variance, and the presence of outliers, and provides insights into its performance.

The residuals (y-axis) are the differences between the observed and predicted values of Life Expectancy ranging from approximately -11 to 11. Residuals indicate how far the model's predictions deviate from the actual data. The fitted values (x-axis) are the predicted Life Expectancy based on the regression model, ranging from 40 to 70, which align with the model's predictors People.TV and People.Dr.

The plot shows that residuals are not even scattered around the red horizontal line at 0 across the range of fitted values. At the lower fitted values (40-50), there is one noticeable outlier with a large positive residual (~10), indicating the model significantly under predicts Life Expectancy for this observation.
At the higher fitted values (65-70), the residuals appear to fan out, with greater variability above and below the horizontal line, suggesting the model's predictions are less consistent at this range.

While the residuals do not show a clear non-linear trend (e.g., curvature), the increasing spread at higher fitted values suggests that the model struggles to predict accurately for larger Life Expectancy values.

Observations to Highlight
At lower fitted values (40-50): The model has one prominent outlier with a large positive residual (~10), suggesting a significant under prediction for this case.

At higher fitted values (65-70): The residuals show increasing spread, indicating the model's predictions are less reliable for higher Life Expectancy values.

No Curvature: The residuals do not show a curved pattern, meaning the assumption of linearity is reasonable.

Conclusion
The residuals plot reveals some systematic bias and increasing spread, which indicate the model's predictions are not equally accurate across the range of fitted values. This plot suggests potential improvements using additional predictors.

Hypothesis Testing For Research Question 2

$H_0 : r = 0$ (No linear correlation (Null Hypothesis))
$H_a : r \neq 0$ (Linear correlation exists (Alternate Hypothesis))

We are using the t-statistic derived from the correlation coefficient:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where:
r = correlation coefficient
n = sample size

We will be using the standard significance level ($\alpha = 0.05$)

```
people_tv <- data_missing$People.TV
people_doctor <- data_missing$People.Dr
life_expectancy <- data_missing$Life.Expectancy


cor_tv <- cor.test(people_tv, life_expectancy)
print(cor_tv)
```

```
##
##  Pearson's product-moment correlation
##
## data:  people_tv and life_expectancy
## t = -3.8091, df = 38, p-value = 0.0004953
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.719376 -0.256089
## sample estimates:
##       cor
## -0.5256639
```

```
cor_doctor <- cor.test(people_doctor, life_expectancy)
print(cor_doctor)
```

```
##
##  Pearson's product-moment correlation
##
## data:  people_doctor and life_expectancy
## t = -5.5037, df = 38, p-value = 2.729e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8095528 -0.4472824
## sample estimates:
##       cor
## -0.6659967
```

```r
summary_table <- data.frame(
  Predictor = c("People/TV", "People/Doctor"),
  Correlation_r = c(cor_tv$estimate, cor_doctor$estimate),
  t_statistic = c(cor_tv$statistic, cor_doctor$statistic),
  p_value = c(cor_tv$p.value, cor_doctor$p.value),
  Significant = c(cor_tv$p.value < 0.05, cor_doctor$p.value < 0.05)
)
print(summary_table)
```

```
##       Predictor Correlation_r t_statistic      p_value Significant
## 1     People/TV    -0.5256639   -3.809145 4.952976e-04        TRUE
## 2 People/Doctor    -0.6659967   -5.503657 2.729360e-06        TRUE
```

Key Findings from the Correlation Analysis and Summary Table

The correlation analysis shows that both the People/TV and People/Doctor ratios are significantly associated with life expectancy:

People/TV Ratio:

The correlation coefficient (r = -0.5266) indicates a moderate negative relationship between the number of people per TV and life expectancy. This suggests that as the People/TV ratio increases (fewer TVs per person), life expectancy tends to decrease.
The p-value ($p = 4.953 \times 10^{-6}$) confirms that this relationship is statistically significant at the 0.05 level. Thus, we reject the null hypothesis that there is no correlation between the People/TV ratio and life expectancy.

People/Doctor Ratio:

The correlation coefficient (r = -0.666) indicates a stronger negative relationship between the number of people per doctor and life expectancy. As the People/Doctor ratio increases (fewer doctors available per person), life expectancy decreases significantly.
The p-value ($p = 2.729 \times 10^{-6}$) is extremely small, providing strong statistical evidence for rejecting the null hypothesis.

The People/Doctor ratio has a stronger negative association with life expectancy compared to the People/TV ratio, suggesting that healthcare access plays a more direct and substantial role in determining life expectancy.

Conclusion

Based on the correlation analysis, scatter plots, residual plot, and summary table, both the People/TV and People/Doctor ratios are clearly linked to life expectancy. The People/Doctor ratio shows a stronger connection, emphasizing how important healthcare access is for life expectancy. The People/TV ratio also reflects the impact of broader socioeconomic factors. While these relationships are statistically significant, we cannot assume they directly cause changes in life expectancy. Future research should look deeper into these connections, considering other factors that might influence the results and using methods to better understand cause and effect.