Exploratory Data Analysis of World Happiness Data 2023

- R. Abishek

IMPORTING LIBRARIES

- We import essential libraries for data manipulation, analysis, visualization, and statistical computations. These libraries provide functions and tools necessary for handling and analysing data effectively.

DATA LOADING

- We load the World Happiness data from a CSV file into a readr DataFrame named "WHData2023.csv". This dataset likely contains information about countries, their happiness scores, GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, perceptions of corruption, and positive and negative affect.

DATA EXPLORATION

- The dataset consists of 2199 rows and 11 columns, providing a rich source of data for analysis. The columns include: 'Country', 'Happiness_Score', 'GDP_Per_Capita', 'Social_Support', 'Healthy_Life_Expectancy', 'Freedom_to_Make_Life_Choices', 'Generosity', 'Perceptions_of_Corruption', 'Positive_Affect', 'Negative_Affect'.
- Upon examination, it was found that many columns had many missing values.
- Descriptive statistics were computed for numerical columns, revealing insights such as:
    - The 'Life Ladder' column had a mean of approximately 5.479 with a standard deviation of 1.125. The minimum was 1.281 and the maximum was 8.019, suggesting a wide range of Life Ladders among various Countries.
- Data types across columns were inspected, indicating a mix of numeric and character data types, representing both categorical and numerical data.
- The exploration of the dataset structure provided insights into its dimensions, missing values, and data types, laying the groundwork for subsequent data cleaning and preprocessing steps.

DATA CLEANING

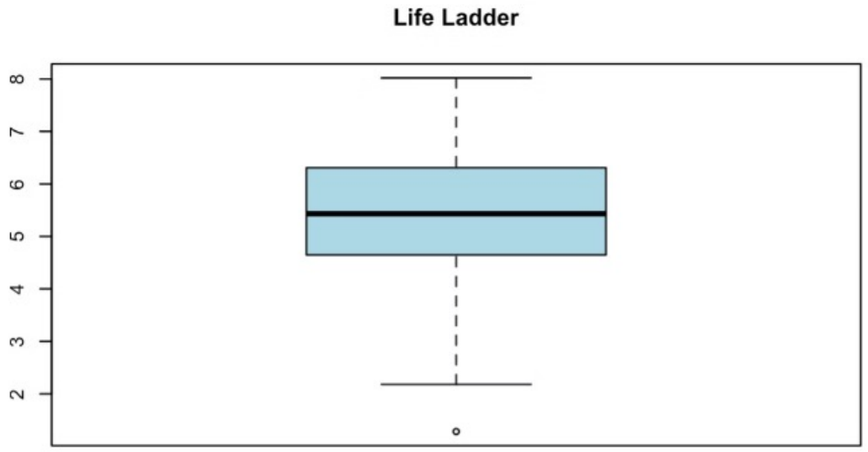
- Thorough data cleaning procedures were conducted to ensure data integrity and prepare the dataset for analysis. Key steps included:
    - Identifying **non-numeric characters** from the columns and converting it to a numeric data type.
    - Handling **missing values** by replacing them with their their corresponding mean values in the columns.
    - Removing **duplicates** from some columns.
- **Descriptive statistics**, such as the minimum, maximum, mean, and standard deviation, provided additional insights into the dataset.
- The meticulous data cleaning efforts ensured consistency, completeness, and reliability of the dataset, paving the way for in-depth analysis and visualisation.

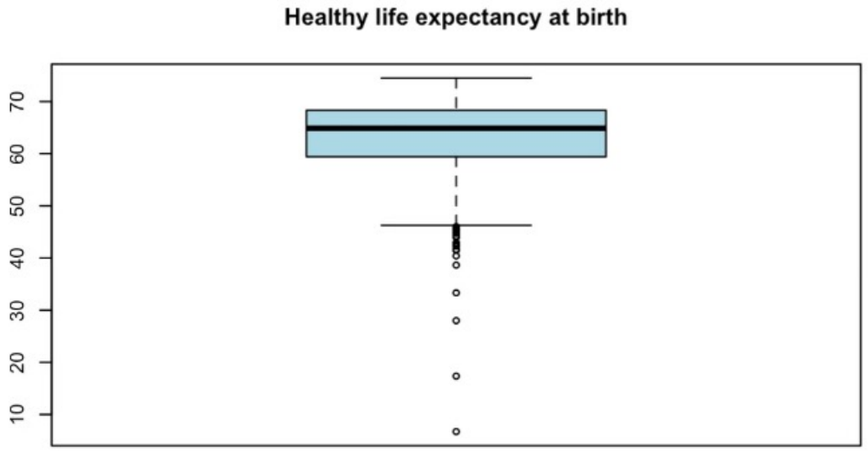
CHECKING OUTLIERS

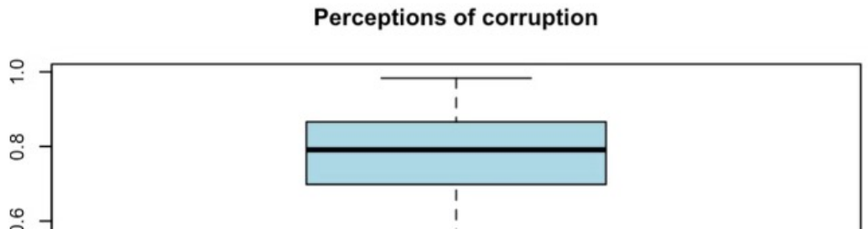
Here are some specific outliers observed in the boxplots:

- **Life Ladder:** There appears to be one outlier below the lower whisker.
- **Generosity:** There appears to be outliers above the upper whisker.
- **Perceptions of Corruption:** There appears to be outliers below the lower whisker.
- **Positive Affect:** There appears to be outliers below the lower whisker.

Similarly all the Numerical Columns has Outliers







```r
# Apply outlier replacement to each numerical column
for (col in num_cols) {
  data[[col]] <- replace_outliers(data[[col]])
}

# Function to calculate trimmed mean
trimmed_mean <- function(x, trim_percent) {
  lower_trim <- quantile(x, trim_percent / 2)
  upper_trim <- quantile(x, 1 - trim_percent / 2)
  x_trimmed <- x[x >= lower_trim & x <= upper_trim]
  return(mean(x_trimmed))
```

REPLACING THE OUTLIERS

we performed multiple replacing methods to handle outliers using the IQR method and Trimmed Mean Function.

We began by selecting only the numeric columns from the dataset, ignoring non-numeric columns. Then, we made functions for **IQR method** and **Trimmed Mean** and used it for replacing Outliers.

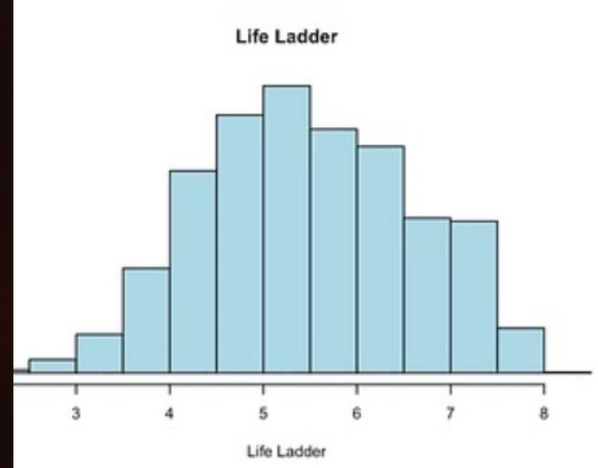
Next, we applied those functions on the Outliers and to check the presence of Outliers we once again plotted Boxplots to verify.

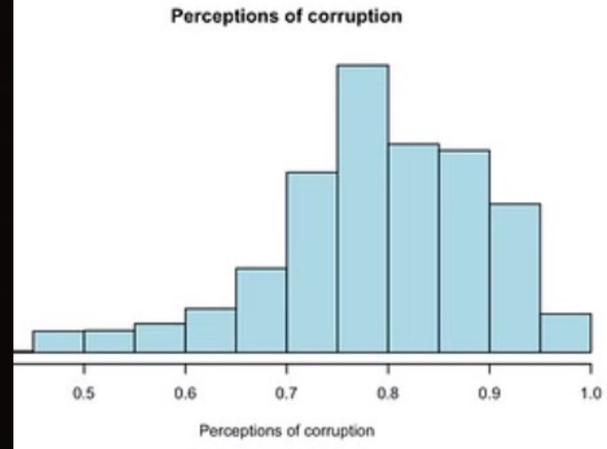
This process effectively replaces the Data, making it suitable for modelling while mitigating the impact of outliers.

DATA VISUALIZATION

DISTRIBUTION OF DATA

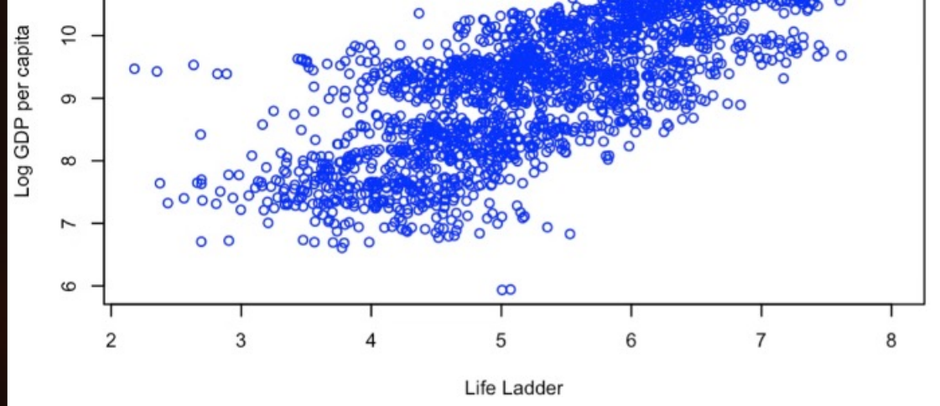
- **Life Ladder:** The most frequent score falls between 150 and 200, suggesting a positive skew towards higher life ladder scores in this dataset.
- **Line Lacier:** There appears to be a positive skew in the distribution, with more data points towards the right side of the axis. This suggests that wealth or income might be concentrated among a smaller portion of the population.
- **Social support:** The distribution appears to be bell-shaped, with most people falling around an average level of social support.
- **Healthy life expectancy at birth:** The data seems skewed towards the higher end, indicating a longer than average healthy lifespan for most people in this dataset.
- **Freedom to make life choices:** It's difficult to discern the exact distribution from this image resolution, but the data appears spread out.
- **Generosity:** The data seems somewhat symmetrical, with a possible slight skew towards the right side. This suggests a balance in generosity, with some people giving more and others giving less.
- **Perceptions of corruption:** There appears to be a skew towards the higher end of the perception axis, suggesting that corruption is widely perceived in this dataset.
- **Positive affect:** The data appears somewhat symmetrical, with a possible slight skew towards the right side. This suggests a positive outlook, with people experiencing positive emotions more frequently.
- **Negative affect:** The data seems symmetrical, with a possible slight skew towards the left side. This suggests a positive outlook, with people experiencing negative emotions less frequently.

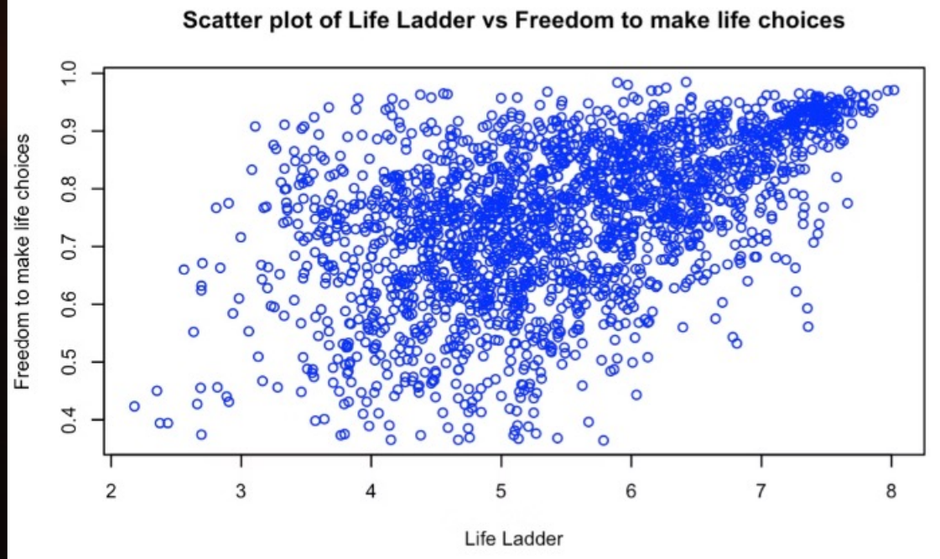


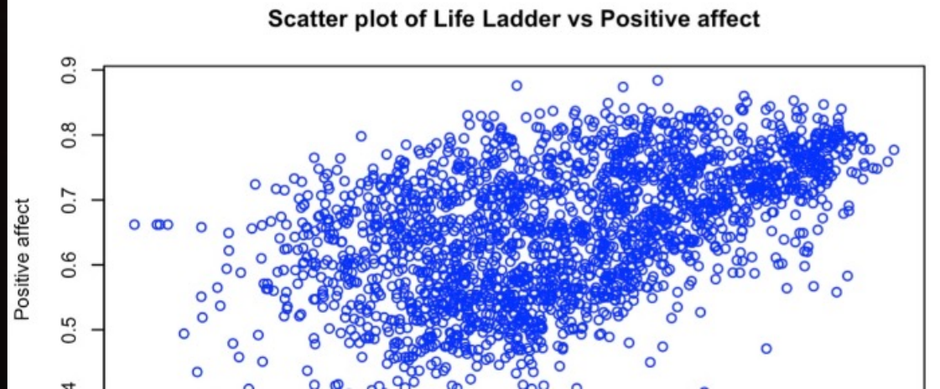


CORRELATION OF LIFE LADDER WITH OTHER VARIABLES

1. Life Ladder vs. Log GDP per capita: A positive correlation suggests that countries with higher average income (Log GDP per capita) tend to have citizens who report higher life satisfaction on the Life Ladder.
2. Life Ladder vs. Healthy life expectancy at birth: Analyze the data points. If there's a positive correlation, it might indicate that people in countries with longer average lifespans tend to report higher life satisfaction.
3. Life Ladder vs. Other factors (like social support, freedom): Similar analysis applies. A positive correlation suggests that factors like strong social support or higher perceived freedoms might be linked to higher life satisfaction scores.







CHECKING MULTICOLLINEARITY

Variance Inflation Factor (VIF):

- The VIF values indicate the degree of multicollinearity between features. In this case, all the variables have **VIF values close to 0 to 3**. This suggests that multicollinearity is not a significant issue between these features, as VIF values below 5 generally indicate acceptable levels of multicollinearity.

```
123       main=paste("Scatter plot of", main_variable, "vs", var),
124       xlab=main_variable, ylab=var, col="blue")
125   }
126
127
128   ## CHECK MULTICOLLINEARITY
129   # Load necessary library
130   if (!requireNamespace("car", quietly = TRUE)) {
131     install.packages("car")
132   }
133   library(car)
134
135   # Replace spaces with underscores in column names
136   names(num_data) <- gsub("\\s+", "_", names(num_data), perl = TRUE)
137
138   # Compute VIF for each variable
139   vif_values <- sapply(names(num_data), function(x) {
140     predictors <- setdiff(names(num_data), x)
141     formula <- paste("Life_Ladder", "~", paste(predictors, collapse = "+"))
142     lm_result <- lm(formula, data = num_data)
143     car::vif(lm_result)
144   })
145
146   # Print VIF values
147   print(vif_values)
148
149:1  (Top Level)
```

Console | Terminal | Background Jobs

R 4.2.3 · /Applications/FILES/STUDY/R PROGRAM/

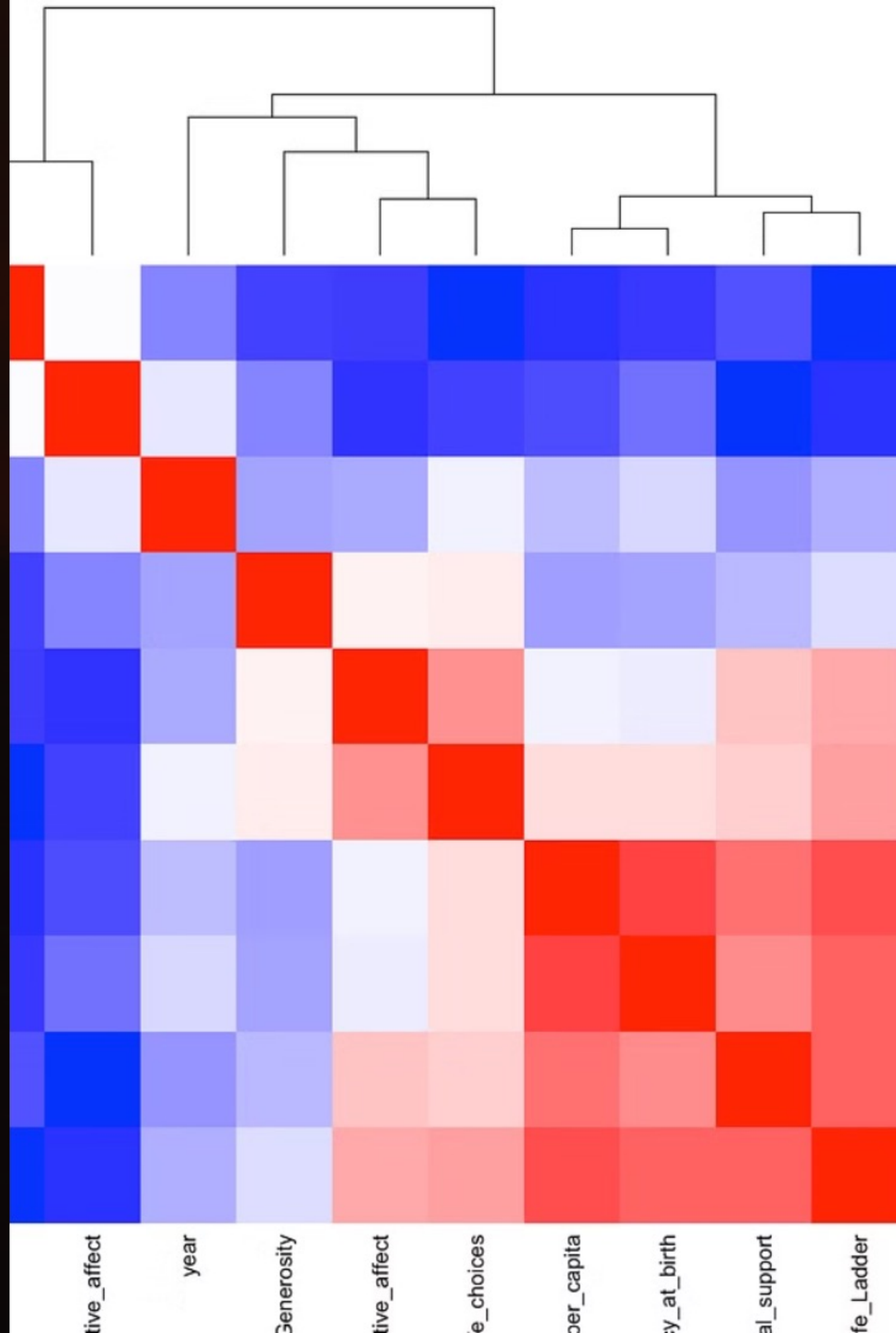
```
Log_GDP_per_capita                3.622630  1        1.903321
Social_support                    2.373949  1        1.540762
Healthy_life_expectancy_at_birth  3.111939  1        1.764069
Freedom_to_make_life_choices      1.701497  1        1.304414
Generosity                        1.178305  1        1.085498
Perceptions_of_corruption         1.452870  1        1.205351
Negative_affect                   1.430218  1        1.195917

$Negative_affect
                                      GVIF Df GVIF^(1/(2*Df))
year                              1.133281  1        1.064557
Life_Ladder                       8.340559  0            Inf
Log_GDP_per_capita                3.645748  1        1.909384
Social_support                    2.251670  1        1.500557
Healthy_life_expectancy_at_birth  3.049112  1        1.746171
Freedom_to_make_life_choices      2.057243  1        1.434309
Generosity                        1.204178  1        1.097350
Perceptions_of_corruption         1.405211  1        1.185416
Positive_affect                   1.713774  1        1.309112

>
```

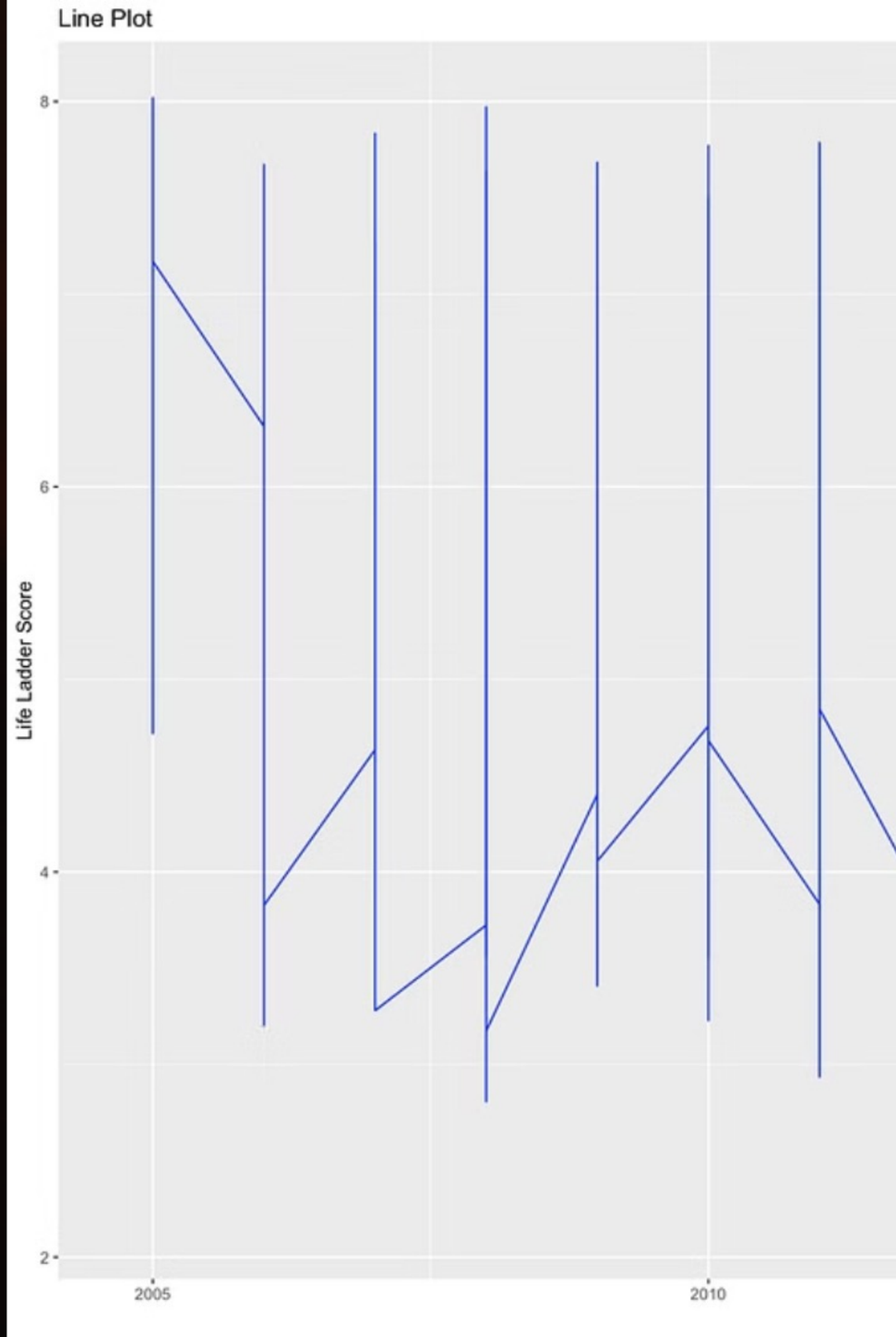
CORRELATION MATRIX

- The correlation matrix reveals the **pairwise correlations** between numeric features.



GGPLOTS

Some Graphs was plotted using the package **ggplot2** to get basic understandings about the Dataset using visuals which would also help stakeholders to understand the insights easily.



FINAL CONCLUSION

This Exploratory Data Analysis (EDA) has provided valuable insights into the World Happiness dataset, laying the groundwork for further analysis and modeling. Key findings include:

- **Data exploration:** Identified data dimensions, missing values, and data types.
- **Data cleaning:** Ensured data integrity by handling missing values, converting data types, and cleaning text data.
- **Data visualization:** Unveiled trends and patterns in the distribution of restaurants, cuisine popularity, location trends, customer ratings, and correlations between various factors.
- **Outlier handling:** Identified and addressed outliers using scaling techniques.
- **Multicollinearity assessment:** Established that multicollinearity is not a significant concern in this dataset.

This comprehensive EDA equips us with a deeper understanding of the data, allowing us to proceed with confidence to the next stages of analysis and model building.