**Neel Acharya**

**STAT355 - Final Project - Happiness in R**

**Introduction**

The World Happiness Report is a landmark survey of the state of global happiness. The report continues to gain global recognition as governments, organizations, and civil society increasingly use happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy, and more – describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness. The happiness scores and rankings use data from the Gallup World Poll. The columns following the happiness score estimate the extent to which each of the six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity. ((Ajaypal Singh, 2021)

The dataset I have chosen to work with for my project is the World happiness report dataset for 2021 found at: <https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021>

Files Attached:

* R file for the code in the project
* Excel file for the dataset (can also found at <https://www.kaggle.com/datasets/ajaypalsinghlo/world-happiness-report-2021> )
* This Google Doc containing the variables analysis and the research questions/hypotheses

**The sample size**: The typical annual sample for each country is 1,000 people. If a typical country had surveys each year, the sample size would be 3,000. A sample size of 2,000 to 3,000 is large enough to give a reasonably good estimate at the national level.

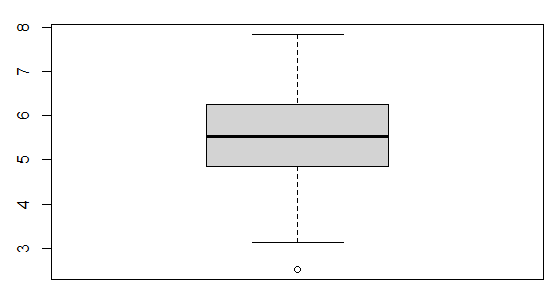
**The dataset itself** (or at least the variables that I'm looking at):

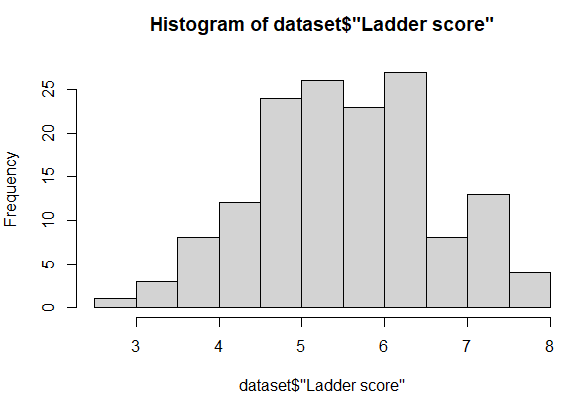
| Country name | Ladder score | Logged GDP per capita | Social support | Healthy life expectancy | Freedom to make life choices | Generosity | Perceptions of corruption |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Finland | 7.842 | 10.775 | 0.954 | 72 | 0.949 | -0.098 | 0.186 |
| Denmark | 7.62 | 10.933 | 0.954 | 72.7 | 0.946 | 0.03 | 0.179 |
| Switzerland | 7.571 | 11.117 | 0.942 | 74.4 | 0.919 | 0.025 | 0.292 |
| Iceland | 7.554 | 10.878 | 0.983 | 73 | 0.955 | 0.16 | 0.673 |
| Netherlands | 7.464 | 10.932 | 0.942 | 72.4 | 0.913 | 0.175 | 0.338 |
| Norway | 7.392 | 11.053 | 0.954 | 73.3 | 0.96 | 0.093 | 0.27 |
| Sweden | 7.363 | 10.867 | 0.934 | 72.7 | 0.945 | 0.086 | 0.237 |
| Luxembourg | 7.324 | 11.647 | 0.908 | 72.6 | 0.907 | -0.034 | 0.386 |
| New Zealand | 7.277 | 10.643 | 0.948 | 73.4 | 0.929 | 0.134 | 0.242 |
| Austria | 7.268 | 10.906 | 0.934 | 73.3 | 0.908 | 0.042 | 0.481 |
| Australia | 7.183 | 10.796 | 0.94 | 73.9 | 0.914 | 0.159 | 0.442 |
| Israel | 7.157 | 10.575 | 0.939 | 73.503 | 0.8 | 0.031 | 0.753 |
| Germany | 7.155 | 10.873 | 0.903 | 72.5 | 0.875 | 0.011 | 0.46 |
| Canada | 7.103 | 10.776 | 0.926 | 73.8 | 0.915 | 0.089 | 0.415 |
| Ireland | 7.085 | 11.342 | 0.947 | 72.4 | 0.879 | 0.077 | 0.363 |
| Costa Rica | 7.069 | 9.88 | 0.891 | 71.4 | 0.934 | -0.126 | 0.809 |
| United Kingdom | 7.064 | 10.707 | 0.934 | 72.5 | 0.859 | 0.233 | 0.459 |
| Czech Republic | 6.965 | 10.556 | 0.947 | 70.807 | 0.858 | -0.208 | 0.868 |
| United States | 6.951 | 11.023 | 0.92 | 68.2 | 0.837 | 0.098 | 0.698 |
| Belgium | 6.834 | 10.823 | 0.906 | 72.199 | 0.783 | -0.153 | 0.646 |
| France | 6.69 | 10.704 | 0.942 | 74 | 0.822 | -0.147 | 0.571 |
| Bahrain | 6.647 | 10.669 | 0.862 | 69.495 | 0.925 | 0.089 | 0.722 |
| Malta | 6.602 | 10.674 | 0.931 | 72.2 | 0.927 | 0.133 | 0.653 |
| Taiwan Province of China | 6.584 | 10.871 | 0.898 | 69.6 | 0.784 | -0.07 | 0.721 |
| United Arab Emirates | 6.561 | 11.085 | 0.844 | 67.333 | 0.932 | 0.074 | 0.589 |
| Saudi Arabia | 6.494 | 10.743 | 0.891 | 66.603 | 0.877 | -0.149 | 0.684 |
| Spain | 6.491 | 10.571 | 0.932 | 74.7 | 0.761 | -0.081 | 0.745 |
| Italy | 6.483 | 10.623 | 0.88 | 73.8 | 0.693 | -0.084 | 0.866 |
| Slovenia | 6.461 | 10.529 | 0.948 | 71.4 | 0.949 | -0.101 | 0.806 |
| Guatemala | 6.435 | 9.053 | 0.813 | 64.958 | 0.906 | -0.038 | 0.775 |
| Uruguay | 6.431 | 9.966 | 0.925 | 69.1 | 0.896 | -0.092 | 0.59 |
| Singapore | 6.377 | 11.488 | 0.915 | 76.953 | 0.927 | -0.018 | 0.082 |
| Kosovo | 6.372 | 9.318 | 0.821 | 63.813 | 0.869 | 0.257 | 0.917 |
| Slovakia | 6.331 | 10.369 | 0.936 | 69.201 | 0.766 | -0.124 | 0.911 |
| Brazil | 6.33 | 9.577 | 0.882 | 66.601 | 0.804 | -0.071 | 0.756 |
| Mexico | 6.317 | 9.859 | 0.831 | 68.597 | 0.862 | -0.147 | 0.799 |
| Jamaica | 6.309 | 9.186 | 0.877 | 67.5 | 0.89 | -0.137 | 0.884 |
| Lithuania | 6.255 | 10.499 | 0.935 | 67.906 | 0.773 | -0.203 | 0.826 |
| Cyprus | 6.223 | 10.576 | 0.802 | 73.898 | 0.763 | -0.015 | 0.844 |
| Estonia | 6.189 | 10.481 | 0.941 | 68.8 | 0.909 | -0.106 | 0.527 |
| Panama | 6.18 | 10.35 | 0.896 | 69.652 | 0.872 | -0.166 | 0.856 |
| Uzbekistan | 6.179 | 8.836 | 0.918 | 65.255 | 0.97 | 0.311 | 0.515 |
| Chile | 6.172 | 10.071 | 0.882 | 70 | 0.742 | -0.044 | 0.83 |
| Poland | 6.166 | 10.382 | 0.898 | 69.702 | 0.841 | -0.165 | 0.735 |
| Kazakhstan | 6.152 | 10.155 | 0.952 | 65.2 | 0.853 | -0.069 | 0.733 |
| Romania | 6.14 | 10.284 | 0.832 | 67.355 | 0.845 | -0.219 | 0.938 |
| Kuwait | 6.106 | 10.817 | 0.843 | 66.9 | 0.867 | -0.104 | 0.736 |
| Serbia | 6.078 | 9.787 | 0.873 | 68.6 | 0.778 | 0.002 | 0.835 |
| El Salvador | 6.061 | 9.054 | 0.762 | 66.402 | 0.888 | -0.11 | 0.688 |
| Mauritius | 6.049 | 10.008 | 0.905 | 66.701 | 0.867 | -0.054 | 0.789 |
| Latvia | 6.032 | 10.315 | 0.927 | 67.1 | 0.715 | -0.162 | 0.8 |
| Colombia | 6.012 | 9.557 | 0.847 | 68.001 | 0.837 | -0.135 | 0.841 |
| Hungary | 5.992 | 10.358 | 0.943 | 68 | 0.755 | -0.186 | 0.876 |
| Thailand | 5.985 | 9.805 | 0.888 | 67.401 | 0.884 | 0.287 | 0.895 |
| Nicaragua | 5.972 | 8.62 | 0.864 | 67.657 | 0.836 | 0.02 | 0.664 |
| Japan | 5.94 | 10.611 | 0.884 | 75.1 | 0.796 | -0.258 | 0.638 |
| Argentina | 5.929 | 9.962 | 0.898 | 69 | 0.828 | -0.182 | 0.834 |
| Portugal | 5.929 | 10.421 | 0.879 | 72.6 | 0.892 | -0.244 | 0.887 |
| Honduras | 5.919 | 8.648 | 0.812 | 67.3 | 0.857 | 0.081 | 0.809 |
| Croatia | 5.882 | 10.217 | 0.924 | 70.799 | 0.754 | -0.118 | 0.939 |
| Philippines | 5.88 | 9.076 | 0.83 | 62 | 0.917 | -0.097 | 0.742 |
| South Korea | 5.845 | 10.651 | 0.799 | 73.9 | 0.672 | -0.083 | 0.727 |
| Peru | 5.84 | 9.458 | 0.832 | 68.25 | 0.822 | -0.154 | 0.891 |
| Bosnia and Herzegovina | 5.813 | 9.59 | 0.87 | 68.098 | 0.706 | 0.113 | 0.931 |
| Moldova | 5.766 | 9.454 | 0.857 | 65.699 | 0.822 | -0.079 | 0.918 |
| Ecuador | 5.764 | 9.313 | 0.821 | 68.8 | 0.842 | -0.124 | 0.843 |
| Kyrgyzstan | 5.744 | 8.538 | 0.893 | 64.401 | 0.935 | 0.119 | 0.908 |
| Greece | 5.723 | 10.279 | 0.823 | 72.6 | 0.582 | -0.288 | 0.823 |
| Bolivia | 5.716 | 9.046 | 0.81 | 63.901 | 0.875 | -0.077 | 0.839 |
| Mongolia | 5.677 | 9.4 | 0.935 | 62.5 | 0.708 | 0.116 | 0.856 |
| Paraguay | 5.653 | 9.448 | 0.893 | 65.9 | 0.876 | 0.028 | 0.882 |
| Montenegro | 5.581 | 9.94 | 0.858 | 68.699 | 0.708 | -0.034 | 0.812 |
| Dominican Republic | 5.545 | 9.802 | 0.853 | 66.102 | 0.86 | -0.133 | 0.714 |
| North Cyprus | 5.536 | 10.576 | 0.82 | 73.898 | 0.795 | 0.012 | 0.626 |
| Belarus | 5.534 | 9.853 | 0.91 | 66.253 | 0.65 | -0.18 | 0.627 |
| Russia | 5.477 | 10.189 | 0.903 | 64.703 | 0.718 | -0.111 | 0.845 |
| Hong Kong S.A.R. of China | 5.477 | 11 | 0.836 | 76.82 | 0.717 | 0.067 | 0.403 |
| Tajikistan | 5.466 | 8.091 | 0.86 | 64.281 | 0.832 | -0.056 | 0.553 |
| Vietnam | 5.411 | 8.973 | 0.85 | 68.034 | 0.94 | -0.098 | 0.796 |
| Libya | 5.41 | 9.622 | 0.827 | 62.3 | 0.771 | -0.087 | 0.667 |
| Malaysia | 5.384 | 10.238 | 0.817 | 67.102 | 0.895 | 0.125 | 0.839 |
| Indonesia | 5.345 | 9.365 | 0.811 | 62.236 | 0.873 | 0.542 | 0.867 |
| Congo (Brazzaville) | 5.342 | 8.117 | 0.636 | 58.221 | 0.695 | -0.068 | 0.745 |
| China | 5.339 | 9.673 | 0.811 | 69.593 | 0.904 | -0.146 | 0.755 |
| Ivory Coast | 5.306 | 8.551 | 0.644 | 50.114 | 0.741 | -0.016 | 0.794 |
| Armenia | 5.283 | 9.487 | 0.799 | 67.055 | 0.825 | -0.168 | 0.629 |
| Nepal | 5.269 | 8.12 | 0.774 | 64.233 | 0.782 | 0.152 | 0.727 |
| Bulgaria | 5.266 | 10.016 | 0.931 | 67 | 0.788 | -0.096 | 0.932 |
| Maldives | 5.198 | 9.826 | 0.913 | 70.6 | 0.854 | 0.024 | 0.825 |
| Azerbaijan | 5.171 | 9.569 | 0.836 | 65.656 | 0.814 | -0.223 | 0.506 |
| Cameroon | 5.142 | 8.189 | 0.71 | 53.515 | 0.731 | 0.026 | 0.848 |
| Senegal | 5.132 | 8.118 | 0.71 | 59.802 | 0.695 | -0.046 | 0.801 |
| Albania | 5.117 | 9.52 | 0.697 | 68.999 | 0.785 | -0.03 | 0.901 |
| North Macedonia | 5.101 | 9.693 | 0.805 | 65.474 | 0.751 | 0.038 | 0.905 |
| Ghana | 5.088 | 8.58 | 0.727 | 57.586 | 0.807 | 0.123 | 0.848 |
| Niger | 5.074 | 7.098 | 0.641 | 53.78 | 0.806 | 0.018 | 0.693 |
| Turkmenistan | 5.066 | 9.629 | 0.983 | 62.409 | 0.877 | 0.273 | 0.888 |
| Gambia | 5.051 | 7.686 | 0.69 | 55.16 | 0.697 | 0.424 | 0.746 |
| Benin | 5.045 | 8.087 | 0.489 | 54.713 | 0.757 | -0.034 | 0.661 |
| Laos | 5.03 | 8.947 | 0.728 | 58.968 | 0.91 | 0.123 | 0.658 |
| Bangladesh | 5.025 | 8.454 | 0.693 | 64.8 | 0.877 | -0.041 | 0.682 |
| Guinea | 4.984 | 7.838 | 0.639 | 55.008 | 0.697 | 0.095 | 0.766 |
| South Africa | 4.956 | 9.403 | 0.86 | 56.904 | 0.749 | -0.067 | 0.86 |
| Turkey | 4.948 | 10.24 | 0.822 | 67.199 | 0.576 | -0.139 | 0.776 |
| Pakistan | 4.934 | 8.458 | 0.651 | 58.709 | 0.726 | 0.098 | 0.787 |
| Morocco | 4.918 | 8.903 | 0.56 | 66.208 | 0.774 | -0.236 | 0.801 |
| Venezuela | 4.892 | 9.073 | 0.861 | 66.7 | 0.615 | -0.169 | 0.827 |
| Georgia | 4.891 | 9.585 | 0.671 | 64.3 | 0.783 | -0.238 | 0.655 |
| Algeria | 4.887 | 9.342 | 0.802 | 66.005 | 0.48 | -0.067 | 0.752 |
| Ukraine | 4.875 | 9.436 | 0.888 | 64.902 | 0.724 | -0.011 | 0.924 |
| Iraq | 4.854 | 9.24 | 0.746 | 60.583 | 0.63 | -0.053 | 0.875 |
| Gabon | 4.852 | 9.603 | 0.776 | 59.962 | 0.731 | -0.2 | 0.84 |
| Burkina Faso | 4.834 | 7.678 | 0.672 | 54.151 | 0.695 | -0.009 | 0.748 |
| Cambodia | 4.83 | 8.36 | 0.765 | 62 | 0.959 | 0.034 | 0.843 |
| Mozambique | 4.794 | 7.158 | 0.744 | 54.706 | 0.882 | 0.061 | 0.684 |
| Nigeria | 4.759 | 8.533 | 0.74 | 50.102 | 0.737 | 0.037 | 0.878 |
| Mali | 4.723 | 7.744 | 0.724 | 51.969 | 0.697 | -0.036 | 0.827 |
| Iran | 4.721 | 9.584 | 0.71 | 66.3 | 0.608 | 0.218 | 0.714 |
| Uganda | 4.636 | 7.677 | 0.781 | 56.101 | 0.709 | 0.122 | 0.855 |
| Liberia | 4.625 | 7.288 | 0.72 | 56.498 | 0.735 | 0.05 | 0.85 |
| Kenya | 4.607 | 8.361 | 0.688 | 60.704 | 0.779 | 0.287 | 0.825 |
| Tunisia | 4.596 | 9.266 | 0.691 | 67.201 | 0.656 | -0.201 | 0.87 |
| Lebanon | 4.584 | 9.626 | 0.848 | 67.355 | 0.525 | -0.073 | 0.898 |
| Namibia | 4.574 | 9.161 | 0.818 | 56.799 | 0.719 | -0.149 | 0.847 |
| Palestinian Territories | 4.517 | 8.485 | 0.826 | 62.25 | 0.653 | -0.163 | 0.821 |
| Myanmar | 4.426 | 8.541 | 0.779 | 59.302 | 0.876 | 0.509 | 0.66 |
| Jordan | 4.395 | 9.182 | 0.767 | 67 | 0.755 | -0.167 | 0.705 |
| Chad | 4.355 | 7.364 | 0.619 | 48.478 | 0.579 | 0.041 | 0.807 |
| Sri Lanka | 4.325 | 9.47 | 0.827 | 67.299 | 0.841 | 0.079 | 0.863 |
| Swaziland | 4.308 | 9.065 | 0.77 | 50.833 | 0.647 | -0.185 | 0.708 |
| Comoros | 4.289 | 8.031 | 0.626 | 57.349 | 0.548 | 0.082 | 0.781 |
| Egypt | 4.283 | 9.367 | 0.75 | 61.998 | 0.749 | -0.182 | 0.795 |
| Ethiopia | 4.275 | 7.694 | 0.764 | 59 | 0.752 | 0.082 | 0.761 |
| Mauritania | 4.227 | 8.542 | 0.795 | 57.161 | 0.561 | -0.106 | 0.731 |
| Madagascar | 4.208 | 7.396 | 0.686 | 59.305 | 0.552 | -0.005 | 0.803 |
| Togo | 4.107 | 7.362 | 0.569 | 54.914 | 0.619 | 0.032 | 0.772 |
| Zambia | 4.073 | 8.145 | 0.708 | 55.809 | 0.782 | 0.061 | 0.823 |
| Sierra Leone | 3.849 | 7.434 | 0.63 | 51.651 | 0.717 | 0.084 | 0.866 |
| India | 3.819 | 8.755 | 0.603 | 60.633 | 0.893 | 0.089 | 0.774 |
| Burundi | 3.775 | 6.635 | 0.49 | 53.4 | 0.626 | -0.024 | 0.607 |
| Yemen | 3.658 | 7.578 | 0.832 | 57.122 | 0.602 | -0.147 | 0.8 |
| Tanzania | 3.623 | 7.876 | 0.702 | 57.999 | 0.833 | 0.183 | 0.577 |
| Haiti | 3.615 | 7.477 | 0.54 | 55.7 | 0.593 | 0.422 | 0.721 |
| Malawi | 3.6 | 6.958 | 0.537 | 57.948 | 0.78 | 0.038 | 0.729 |
| Lesotho | 3.512 | 7.926 | 0.787 | 48.7 | 0.715 | -0.131 | 0.915 |
| Botswana | 3.467 | 9.782 | 0.784 | 59.269 | 0.824 | -0.246 | 0.801 |
| Rwanda | 3.415 | 7.676 | 0.552 | 61.4 | 0.897 | 0.061 | 0.167 |
| Zimbabwe | 3.145 | 7.943 | 0.75 | 56.201 | 0.677 | -0.047 | 0.821 |
| Afghanistan | 2.523 | 7.695 | 0.463 | 52.493 | 0.382 | -0.102 | 0.924 |

**Analysis of Variables** (not ANOVA:( )

The variables in question for this dataset are

1. ‘Country name: The name of the country for which the data is reported. This is a categorical variable and has 150 different values, one for each country.

1. Ladder score: The happiness score, was measured by asking the sampled people the question: "How would you rate your happiness on a scale from 0 to 10 where 10 is the happiest?" The average is 5.532838926, it follows a central tendency like any other ‘nice’ data. This is a continuous numerical variable between 1 and 10 The boxplot is:

The histogram:

The other summary statistics for the dataset are:

mean(dataset$"Ladder score")

[1] 5.532839

> median(dataset$"Ladder score")

[1] 5.534

> range(dataset$"Ladder score")

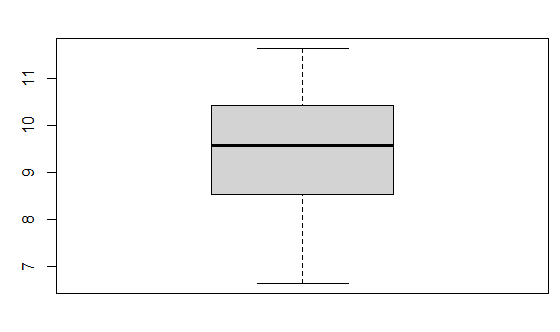
[1] 2.523 7.842

> quantile(dataset$"Ladder score")

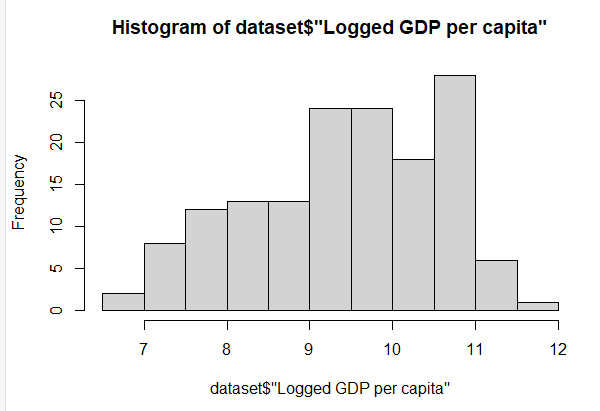
0% 25% 50% 75% 100%

2.523 4.852 5.534 6.255 7.842

1. Logged GDP per capita: The natural logarithm of the GDP per capita, which is a measure of a country's economic output per person. Higher values indicate greater economic prosperity. The average is 9.432208054. This is a continuous numerical variable.

The boxplot: 

The histogram:



As we can see, this is slightly skewed to the left. The other statistics are as follow:

mean(dataset$"Logged GDP per capita")

[1] 9.432208

> median(dataset$"Logged GDP per capita")

[1] 9.569

> range(dataset$"Logged GDP per capita")

[1] 6.635 11.647

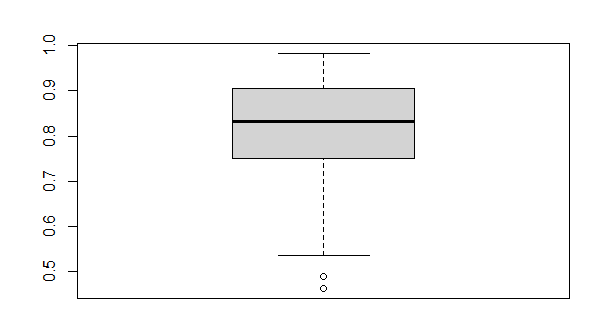
> quantile(dataset$"Logged GDP per capita")

0% 25% 50% 75% 100%

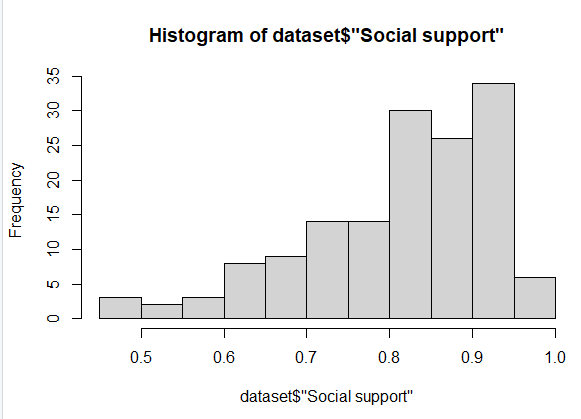
6.635 8.541 9.569 10.421 11.647

1. Social support: The national average of the binary responses (either 0 or 1) to the question: "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?", Here, the average is quite high - more people rely on family than half - with an average of 0.814744966/1. It is a continuous numerical variable between the range of 0 and 1

The boxplot for the data is as follows:



The histogram for the plot is as follows:



As we expect, it is skewed to the left, which means there is more data on the right hadn side of the median.

The other statististics for Social support are:

mean(dataset$"Social support")

[1] 0.814745

> median(dataset$"Social support")

[1] 0.832

> range(dataset$"Social support")

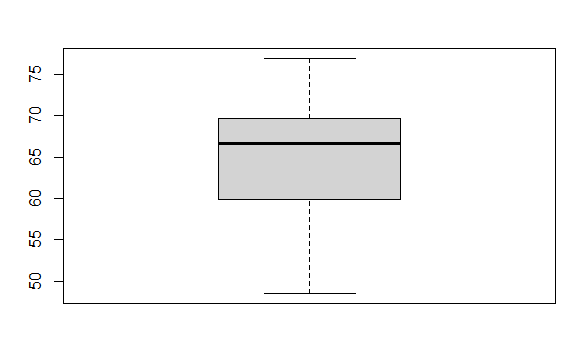
[1] 0.463 0.983

> quantile(dataset$"Social support")

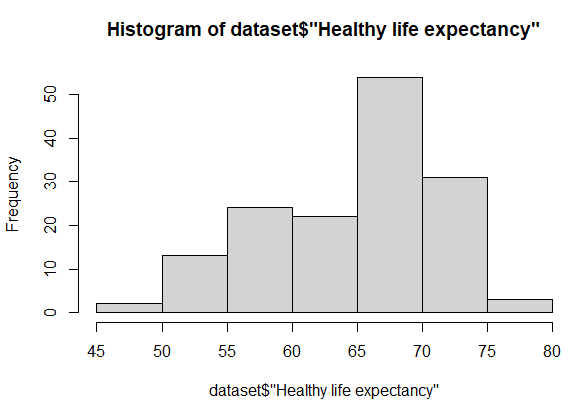
0% 25% 50% 75% 100%

0.463 0.750 0.832 0.905 0.983

1. Healthy life expectancy: The average number of years a person can expect to live in good health, considering the current mortality rates and prevalence of diseases in the country. This number varies a lot with countries and it is a continuous numerical variable. The average is 64.99279866 and the boxplot is as follows:



The histogram is:



The other statistics for this are:

mean(dataset$"Healthy life expectancy")

[1] 64.9928

> median(dataset$"Healthy life expectancy")

[1] 66.603

> range(dataset$"Healthy life expectancy")

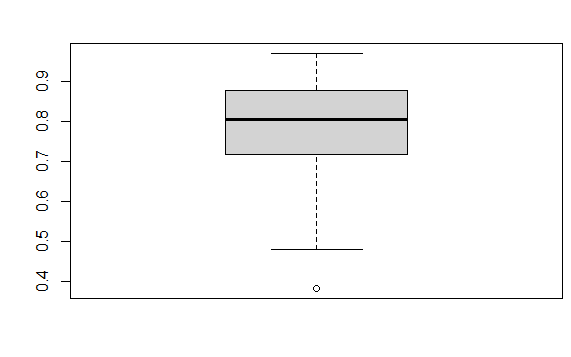
[1] 48.478 76.953

> quantile(dataset$"Healthy life expectancy")

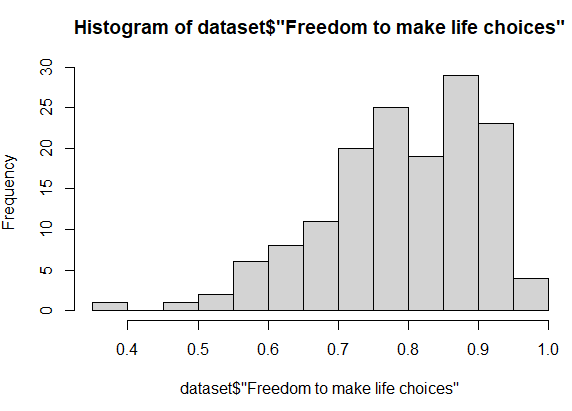
0% 25% 50% 75% 100%

48.478 59.802 66.603 69.600 76.953

1. Freedom to make life choices: The national average of responses to the question: "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?" Higher values indicate greater satisfaction with the freedom to make life choices. This is a continuous numerical variable with an average eof 0.791. The boxplot is:



The histogram is:



The other data for this variable is:

mean(dataset$"Freedom to make life choices")

[1] 0.7915973

> median(dataset$"Freedom to make life choices")

[1] 0.804

> range(dataset$"Freedom to make life choices")

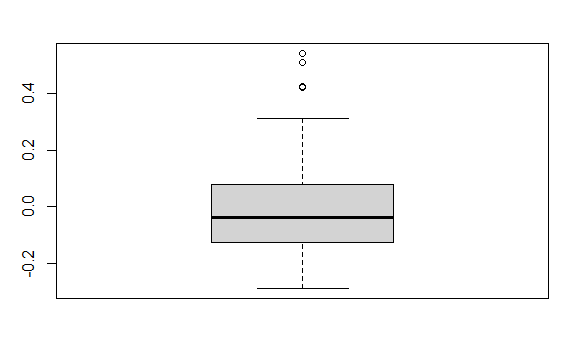
[1] 0.382 0.970

> quantile(dataset$"Freedom to make life choices")

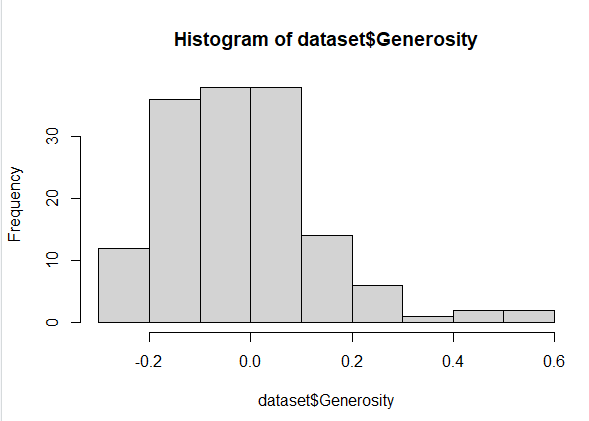
0% 25% 50% 75% 100%

0.382 0.718 0.804 0.877 0.970

1. Generosity: The residual of regressing national average of responses to the question: "Have you donated money to a charity in the past month?" on GDP per capita. The generosity variable is measured using a survey question that asks people to rate how often they have donated money to charity or volunteered their time to help others in the past month. The scores are then normalized to a scale ranging from -1 to 1, where a score of 0 represents the global average level of generosity.



The histogram looks like:



And the 5 number summary looks like:

mean(dataset$"Generosity")

[1] -0.01513423

> median(dataset$"Generosity")

[1] -0.036

> range(dataset$"Generosity")

[1] -0.288 0.542

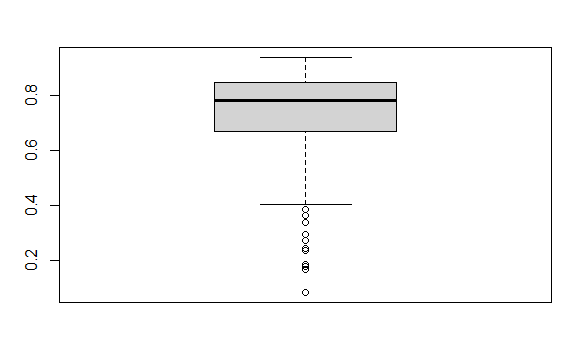
> quantile(dataset$"Generosity")

0% 25% 50% 75% 100%

-0.288 -0.126 -0.036 0.079 0.542

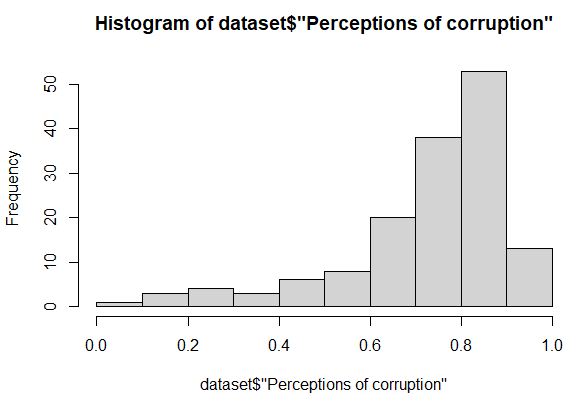
In context of this report, we will look at this variable since it is a residual, as if higher the generosity residual, higher the generosity since it has increased by that much in recent years, so we can use recent years correlation residual as an accurate representation of generosity.

1. Perceptions of corruption: The national average of the survey responses to two questions: "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?" Higher values indicate higher perceived levels of corruption. This is a continuous numerical variable. the values of the perception of corruption variable are being rescaled or transformed to a 0 to 1 scale for comparative or analytical purposes. Rescaling the variable to a 0 to 1 scale can make it easier to compare with other variables in the dataset that are also on a 0 to 1 scale, such as the generosity variable. However, it is important to note that the perception of corruption variable is originally measured on a scale from 0 to 100 and that any rescaling or transformation should be clearly stated and justified in the analysis, as I am doing since I plan to use the variable like this. The boxplot looks like this:



There seems to be more corruption, or a grater perception of more corruption, in recent years.

The histogram looks like this:



And the other summary looks like this:

mean(dataset$"Perceptions of corruption")

[1] 0.7274497

> median(dataset$"Perceptions of corruption")

[1] 0.781

> range(dataset$"Perceptions of corruption")

[1] 0.082 0.939

> quantile(dataset$"Perceptions of corruption")

0% 25% 50% 75% 100%

0.082 0.667 0.781 0.845 0.939

**RESEARCH QUESTIONS**

These variables represent various aspects of a country's well-being, including economic, social, health, and freedom-related factors. The World Happiness Report uses these variables to analyze and rank countries based on their overall happiness levels. Based in these variables, I have developed a few research questions which are the following:

**Research Question 1:** Is there a correlation between GDP per capita and happiness scores?

* Method: Pearson correlation
* Justification: Both variables are continuous, and we want to test their linear association.
* Assumptions: The relationship between the variables is linear, and both variables are normally distributed.

The code to create the correlation, make a plot of it, and then add a regression line to the correlation is the following:

library(tidyverse)

dataset <- read\_csv("D:/world-happiness-report-2021.csv")

correlation <- cor(dataset$"Logged GDP per capita", dataset$"Ladder score", use = "complete.obs", method = "pearson")

cat("Pearson Correlation Coefficient:", correlation, "\n")

# Create a scatter plot

plot(dataset$"Logged GDP per capita", dataset$"Ladder score",

xlab = "Log GDP per capita", ylab = "Happiness Score",

main = "Relationship between GDP per capita and Happiness Scores",

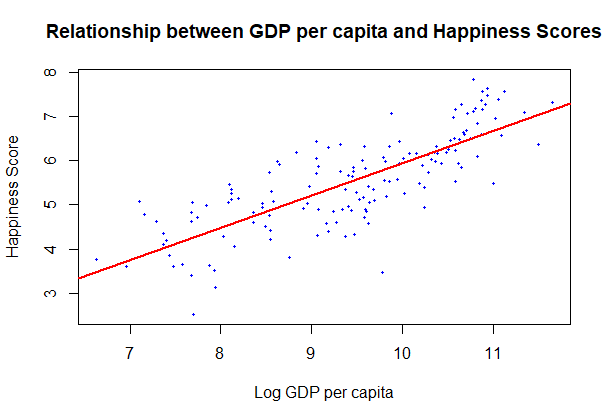
col = "blue", pch = 20, cex = 0.6)

# Add a linear regression line

fit <- lm(dataset$"Ladder score" ~ dataset$"Logged GDP per capita", data = dataset)

abline(fit, col = "red", lwd = 2)

The output of the above code is the following:

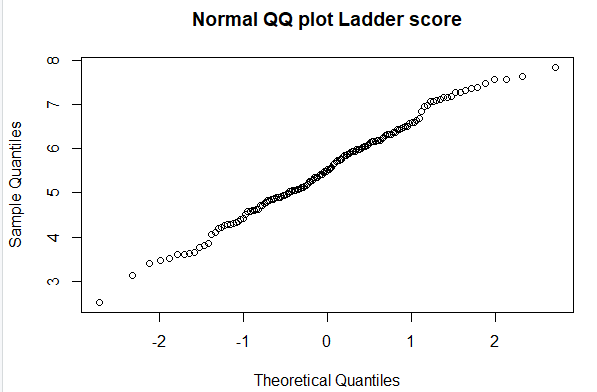
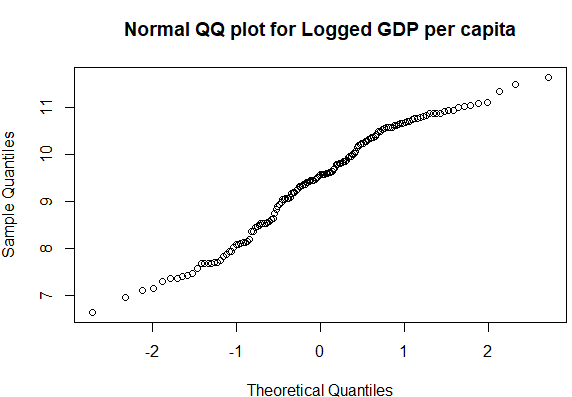


From the output of the correlation, we can see that the value of the correlation index is the following:

correlation

[1] 0.7897597

This means that about 79% of the data is explained by the correlation, This means that in 79% of the values of the ladder score will increase with an increase in the Logged GDP per capita. We can test the normality assumptions by running qqnorms on each of the variables. We get the following outputs:



Since the lines look pretty much straight at the places where there are excessive points, we can say that the data is pretty much normal.

From this, we can conclude that the strong positive correlation between logged GDP per capita and ladder score in the World Happiness Report highlights the importance of economic prosperity for happiness and well-being, but also underscores the need for a holistic and inclusive approach to promoting happiness and well-being around the world.The strong positive correlation between the logged GDP per capita and ladder score in the World Happiness Report suggests that economic prosperity is an important determinant of happiness and well-being. When individuals have access to higher levels of economic resources, they are more likely to have better living conditions, access to healthcare, education, and other opportunities that can contribute to happiness and well-being.

**Research Question 2**: Is there a significant difference in the average generosity scores between countries with high and low levels of perceived corruption?

* Method: Independent two-sample t-test
* Justification: We have two independent groups (countries with high and low levels of perceived corruption), and we want to compare their average generosity scores.
* Assumptions: The two groups are independent, and the data is normally distributed with equal variances.

H0 (Null Hypothesis): There is no difference between the average generosity scores between low and high levels of perceived corruption.

Ha (Alternative Hypothesis): There is a significant difference between the average generosity scores between low and high levels of perceived corruption.

The code to run the t-test in R is the following:

median\_corruption <- median(dataset$"Perceptions of corruption", na.rm = TRUE)

low\_corruption <- dataset[dataset$"Perceptions of corruption" <= median\_corruption,]

high\_corruption <- dataset[dataset$"Perceptions of corruption" > median\_corruption,]

t\_test <- t.test(low\_corruption$"Generosity", high\_corruption$"Generosity", var.equal = TRUE)

print(t\_test)

> print(t\_test)

Two Sample t-test

data: low\_corruption$Generosity and high\_corruption$Generosity

t = 2.0434, df = 147, p-value = 0.04279

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.001641656 0.098178524

sample estimates:

mean of x mean of y

0.009653333 -0.040256757

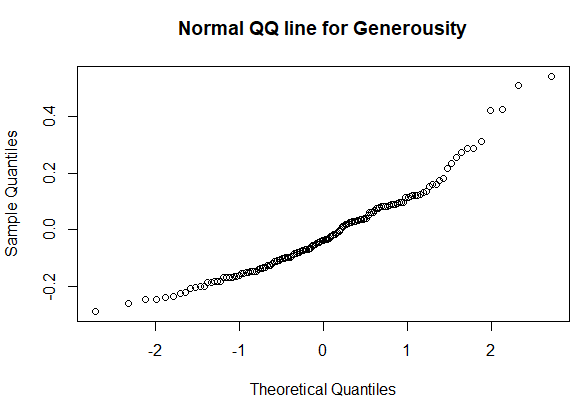
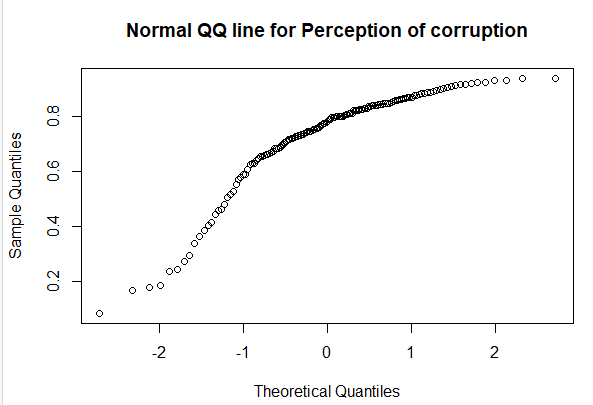
The p value in this t test is 0.0429. When we take a significance level of 0.05, which is the usual significance level taken, the null hypothesis can be rejected. This means that there is significant evidence to say that there is a significant difference between the average generosity scores between low and high levels of perceived corruption.

The t-test calculates a test statistic (t) of 2.0434 and a p-value of 0.04279. The degrees of freedom (df) are 147. The 95 percent confidence interval (CI) for the difference in means ranges from 0.0016 to 0.0982. This means that we can be 95 percent confident that the true difference in means between the two groups falls within this range.

The sample estimates of the mean generosity score for the low corruption group is 0.0097, while the mean generosity score for the high corruption group is -0.0403.

In conclusion, based on the p-value of 0.04279, we can reject the null hypothesis that there is no difference in generosity scores between the low and high corruption groups at a significance level of 0.05. Therefore, we can conclude that there is evidence to suggest that there is a statistically significant difference in the generosity scores between the two groups. Additionally, based on the 95% CI, we can be confident that the true difference in means is likely positive and falls between 0.0016 and 0.0982.

We can test the assumptions that we made about the data being normal by using the normal qq plots. They are:



As we can see, at the parts where the points are concentrated seem to be in a string line so the assumption holds that the datasets are normal.

Thus, we can say that there is a significant difference between the average generosity scores between low and high levels of perceived corruption.

The significant difference in generosity scores between low and high levels of perceived corruption has important implications for social and economic well-being. Higher generosity scores in countries with low corruption may indicate a greater level of trust in the government and in others, contributing to a more positive social and economic environment. On the other hand, lower generosity scores in countries with high levels of corruption may indicate greater inequality and weaker social cohesion. The relationship between generosity and corruption suggests that good governance and trust play an important role in shaping positive social behavior. Addressing corruption could therefore be an important step in promoting positive social behavior, such as generosity. By promoting transparency and reducing corruption, countries may be able to create a more generous and cohesive society, which may lead to a more positive social and economic environment.

**Research Question 3:** How do GDP per capita, social support, life expectancy, freedom to make life choices, generosity, and perceptions of corruption contribute to predicting happiness scores?

* Method: Multiple linear regression
* Justification: We want to model the relationship between the happiness ladder score (continuous dependent variable) and multiple independent variables.
* Assumptions: Linearity, independence of variables, normality of residuals, and absence of multicollinearity.

model <- lm(dataset$"Ladder score" ~ dataset$"Logged GDP per capita" + dataset$"Social support" + dataset$"Healthy life expectancy" + dataset$"Freedom to make life choices" + dataset$"Generosity" + dataset$"Perceptions of corruption", data = dataset)

summary(model)

summary(model)

Call:

lm(formula = dataset$"Ladder score" ~ dataset$"Logged GDP per capita" +

dataset$"Social support" + dataset$"Healthy life expectancy" +

dataset$"Freedom to make life choices" + dataset$Generosity +

dataset$"Perceptions of corruption", data = dataset)

Residuals:

Min 1Q Median 3Q Max

-1.85049 -0.30026 0.05735 0.33368 1.04878

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.23722 0.63049 -3.548 0.000526 \*\*\*

dataset$"Logged GDP per capita" 0.27953 0.08684 3.219 0.001595 \*\*

dataset$"Social support" 2.47621 0.66822 3.706 0.000301 \*\*\*

dataset$"Healthy life expectancy" 0.03031 0.01333 2.274 0.024494 \*

dataset$"Freedom to make life choices" 2.01046 0.49480 4.063 7.98e-05 \*\*\*

dataset$Generosity 0.36438 0.32121 1.134 0.258541

dataset$"Perceptions of corruption" -0.60509 0.29051 -2.083 0.039058 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5417 on 142 degrees of freedom

Multiple R-squared: 0.7558, Adjusted R-squared: 0.7455

F-statistic: 73.27 on 6 and 142 DF, p-value: < 2.2e-16

As we can see, the regression model is very accurate. It has a p value that is much less than the significance value, which means that they seem to highly affect the ladder score, which is what the dataset is supposed to be indicating. In the model, we can also see that almost all of the variables have a p value less than the significance value, which means that all of them affect the regression model, and we have significant evidence to say that. However, the variable generosity doesnt seem to be that effective in the multiple regression model. This doesnt mean that it doesnt affect the value, it just means that it seems to be the case that it affects it less. We can test this model and make it as effective as possible by using the step function. However, the value seems pretty accurate, with a multiple R squared of 75.58%, which means that 75.58% of the variance can be explained by the variables in the model.

The equation for this model is:

Ladder score = -2.23722 + 0.27953 \* Logged GDP per capita + 2.47621 \* Social support + 0.03031 \* Healthy life expectancy + 2.01046 \* Freedom to make life choices + 0.36438 \* Generosity - 0.60509 \* Perceptions of corruption.

Using the step function, we get the following model

summary(reduced\_model)

Call:

lm(formula = dataset$"Ladder score" ~ dataset$"Logged GDP per capita" +

dataset$"Social support" + dataset$"Healthy life expectancy" +

dataset$"Freedom to make life choices" + dataset$"Perceptions of corruption",

data = dataset)

Residuals:

Min 1Q Median 3Q Max

-1.93303 -0.29768 0.06863 0.33924 1.02304

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.11039 0.62112 -3.398 0.000880 \*\*\*

dataset$"Logged GDP per capita" 0.26400 0.08584 3.075 0.002518 \*\*

dataset$"Social support" 2.50670 0.66835 3.751 0.000256 \*\*\*

dataset$"Healthy life expectancy" 0.02936 0.01332 2.204 0.029095 \*

dataset$"Freedom to make life choices" 2.13266 0.48342 4.412 2.01e-05 \*\*\*

dataset$"Perceptions of corruption" -0.66778 0.28549 -2.339 0.020718 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5423 on 143 degrees of freedom

Multiple R-squared: 0.7536, Adjusted R-squared: 0.745

F-statistic: 87.49 on 5 and 143 DF, p-value: < 2.2e-16

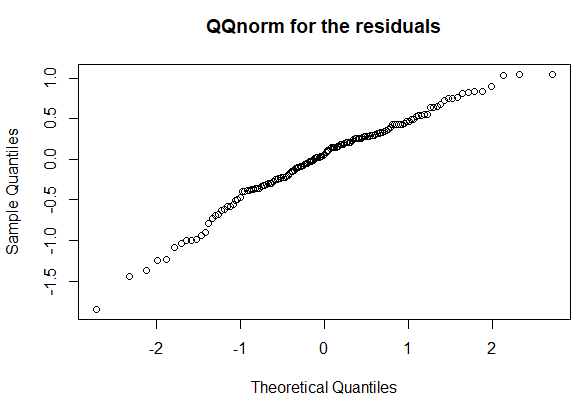
Here, we have the following regression equation:

Ladder score = -2.11039 + 0.26400 \* Logged GDP per capita + 2.50670 \* Social support + 0.02936 \* Healthy life expectancy + 2.13266 \* Freedom to make life choices - 0.66778 \* Perceptions of corruption

Here, the independent variable "Generosity" has been removed from the model, resulting in a simpler model with five independent variables. The coefficients (i.e., the numbers multiplied by each independent variable) represent the estimated effect of each independent variable on the dependent variable, while holding all other variables constant.

As we can see as per the analysis that the generosity variable was not that much of an indicator, we have eliminated it to get a model which has almost the same R square and adjusted R square, but we have to deal with less variables for a model which is just as accurate.

Let us now test the assumptions. First we make a qqplot for the residuals. This outputs the following:



Here, there seems to be straight line without much variation which means that the residuals are mostly normal.

We have to also account for the absence of multicollinearity. We can do this by using a vif test, which stands for variance inflation factor. On running this , we get the following output:

> vif(model)

dataset$"Logged GDP per capita" dataset$"Social support"

5.104890 2.972200

dataset$"Healthy life expectancy" dataset$"Freedom to make life choices"

4.099348 1.585807

dataset$Generosity dataset$"Perceptions of corruption"

1.180982 1.367122

Low values means better in this test, and the 5.1 value doesnt work, which indicates there may be a little amount of multicollinearity, but an amount that is small enough that it can be disregarded. Therefore, based on these VIF values, it does not appear that there is a complete absence of multicollinearity in the model. However, the presence of some multicollinearity may not necessarily invalidate the results of the analysis, as it depends on the extent and nature of the collinearity and the specific research question being investigated. Thus, we can say that the assumptions are mostly true.

Thus, from the above model, we can say that:

1. Logged GDP per capita has a positive and statistically significant effect on happiness scores. This suggests that countries with higher levels of economic development tend to have higher levels of happiness.
2. Social support has a positive and statistically significant effect on happiness scores. This suggests that having strong social connections and support networks is an important factor for happiness.
3. Healthy life expectancy has a positive and statistically significant effect on happiness scores. This suggests that good health and well-being are important for happiness.
4. Freedom to make life choices has a positive and statistically significant effect on happiness scores. This suggests that having the ability to make important life choices and have control over one's own life is important for happiness.
5. Generosity has a positive effect on happiness scores, but the effect is not statistically significant. This suggests that while being generous may contribute to happiness, it may not be as important as other factors.
6. Perceptions of corruption have a negative and statistically significant effect on happiness scores. This suggests that countries with high levels of corruption tend to have lower levels of happiness.

Overall, these findings suggest that a combination of economic, social, and health factors, as well as individual freedoms and low levels of corruption, are important for predicting happiness levels across countries.

**Conclusion**

From the above tests, we have determined quite a few statistical relationships between the variables. Even some variables which do not seem to be useful, such as generosity, we see it doesn't affect the ladder score as much as the other variables in the third and final research question, we can also see that it is somewhat related and useful from the second question which says that the generosity is based on the perception of corruption in the country. This goes to show that even though the backward iteration process of the multiple regression model excludes it, it is still an important variable in determining the relationship since it is related to the other variables in greater ways that cannot be statistically ignored. Thus, by looking into the world happiness index dataset, we have gained some useful insights as to how some of the variables are related, and by looking more into those variables, we will be able to get more useful insights as to how they are related and how we can gradually increase the happiness in different countries by examining different variables and their relationships to generate plausible solutions to some problems in some countries.

By conducting research using the World Happiness Index dataset, we can gain a better understanding of the factors that contribute to happiness and well-being across different populations, which can inform the development of policies and interventions aimed at improving quality of life and promoting well-being around the world.