Analysing the relationship between Internet Speeds and Happiness Indexes

Mark Curran

ATU Sligo

Abstract

The purpose of this document is to compare the effects of different countries' internet speeds on the happiness score. Hypothesis tests will be performed on the data to validate the given data, and to test the differences of particular variables. Questions of what columns are correlated will then be asked, and a prediction model will be generated for the correlation. The hypothesis test found that there is a difference between Broadband and Mobile speeds globally, with mobile speeds being lower. There is a positive correlation between Broadband speed and happiness score for the countries tested, indicating a rise in happiness in a country with higher broadband speeds. A prediction model was created from this correlation.

Keywords: Happiness score, Broadband, Mbps, p-value, Dystopia

1 Introduction

As of 2022, 5.07 billion people use the internet globally, with the median internet speed of mobile users being 29.9 Megabits per second (Mbps). This is a sufficient speed for streaming a 4K resolution video using mobile data without buffering (Kemp, S., 2022). Internet users not getting instantaneous feedback from their devices can often lead to frustration and distress.

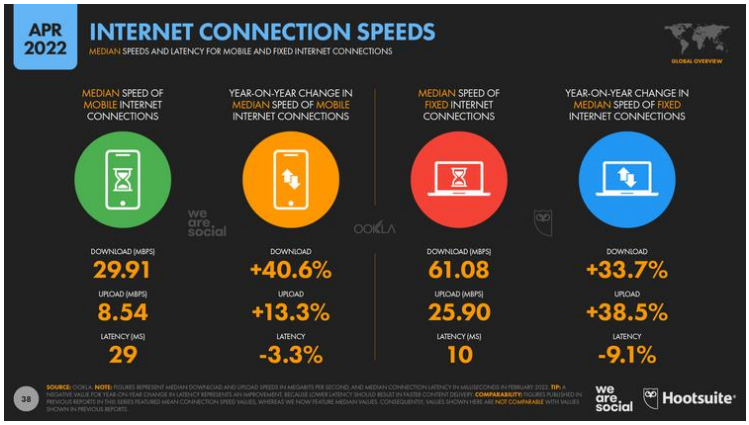

Figure 1 Internet Connection Speeds

The Gallup World Poll World Happiness Report is carried out annually to assess the 'happiness index' of each country worldwide. This document will compare the internet speeds with the happiness indexes of countries around the world to assess their relationship. This document will answer the question of whether a correlation exists between the two.

2 The Data

To answer this question, two datasets were required: to collect data for global internet speeds, and to collect data for the happiness indexes globally. Both datasets were retrieved from Kaggle. It was important to retrieve the two datasets from the same year for a more accurate comparison. The internet speed CSV file contained 179 countries. The columns for the countries included, their broadband and mobile speeds, along with their rank. The happiness report contained 147 countries. Each having their happiness score with confidence intervals, and six variables to explain the final score:

- GDP per capita
- Social Support
- Healthy Life Expectancy
- Freedom to make life choices.
- Generosity
- Corruption

Each happiness score also has the Whisker-high and Whisker-low columns, indicating the upper and lower confidence intervals for the score respectively. Finally, the Dystopia column accounts for the imaginary country of Dystopia which has the world's lowest happiness score.

3 Exploratory Data Analysis

Cleaning the Data Sets

To perform efficient analysis on the datasets, it was necessary to clean the data. This involved removing null columns and correcting data types of columns. The mobile speed rank and speed both had forty null values that had to be removed, while the happiness set had only one null value for most columns. The data types of both sets were already clean when they were validated.

A challenge in combining any data set is matching them by a common column or index. As mentioned above, the internet and happiness sets had 179 and 147 countries respectively. It was then necessary to find the countries that were common to both and remove the others from the final data to be analysed.

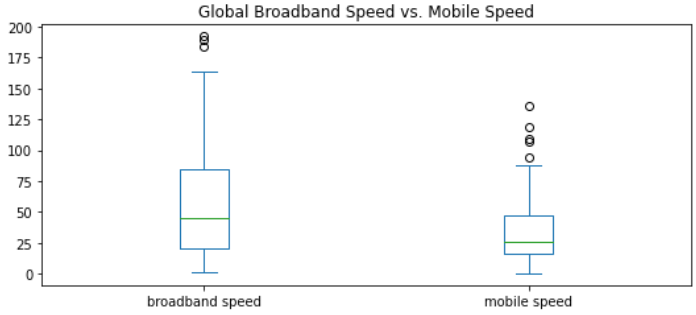

Figure 2 Box plot comparing Internet speeds

The happiness countries set contained some incorrectly formatted country names, and it was necessary to correct this before finding the intersect of the countries from both sets. Once this process was completed, there were 121 countries in both sets, and they were prepared for analysis. Many iterations of these cleaned sets had to be created for different sorting orders for creating different graphical representations of the data.

Results

Figure 2 shows the global broadband and mobile speeds displayed on a box plot. By inspection it is clear that the broadband speeds are generally higher than the mobile speeds. The fastest broadband speed is significantly higher than the highest mobile speed, though the lowest for both are approximately the same. A higher median broadband can be observed for broadband speeds, suggesting a higher distribution for broadband speeds than for mobile. Though it can be seen in Figure 3 that the mobile speeds begin to catch

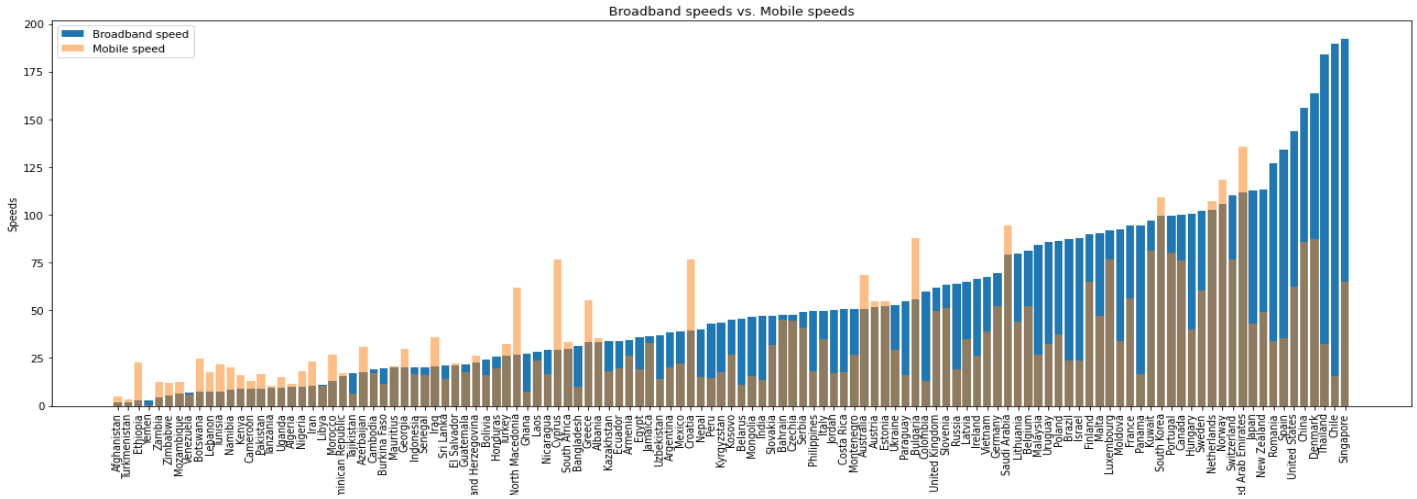

Figure 3 Broadband and Mobile speeds ascending

up to the broadband as broadband decreases, with many having larger mobile speeds than broadband.

A scatter matrix was also created to compare the correlation of broadband speed and mobile speed. Mobile speed is seen to have a positive correlation with broadband speed by inspection. Conversely broadband speed has a slightly wider margin scattered with mobile speed.

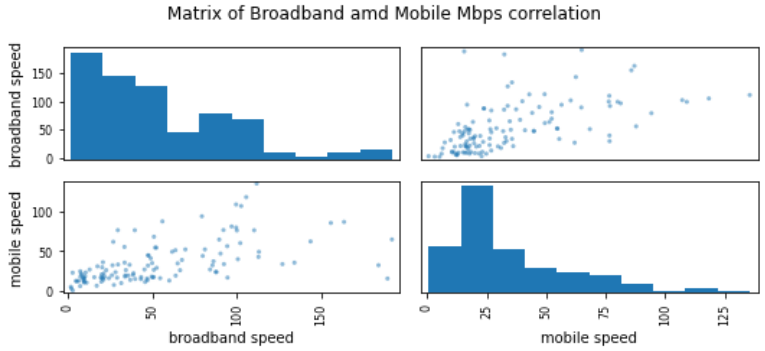

Figure 4 Scatter matrix of broadband and mobile speeds

For the happiness data, a bar chart was created to show the happiness scores in ascending order. A horizontal red line was drawn through it to represent the score of Dystopia (1.83) which no country can be lower than. Afghanistan is closest with a relatively low score of 2.404. This is significantly lower than the mean score for this data of 5.723.

A stacked bar chart was built to represent the six 'Explained by' columns that cumulatively tell the final happiness score. A blue bar representing GDP per capita being primarily the highest proportioned of the factors. The other factors had similar values and tended to fluctuate more with each other.

Figure 5 shows a scatter matrix of broadband speed and happiness score. This figure gives an idea of the correlation between the two variables. They appear to have a positive correlation, with happiness score appearing to increase with higher broadband speeds. A similar plot with mobile speed yielded similar results.

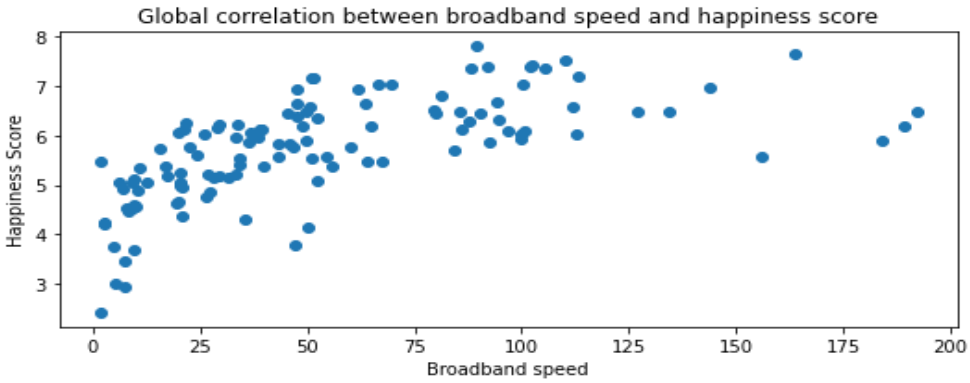

Figure 5 Correlation of broadband speed and Happiness score

4 Questions

Having cleaned the data set and performing exploratory data analysis on them allowed questions to be asked about the data:

1. Is there a significant difference in the overall broadband speed and the overall mobile speed across the countries in question?

2. Are the Whisker-high/low columns accurate confidence intervals for the happiness scores?

3. Are their statistically significant differences between the effects of generosity and perceptions of corruption on happiness scores?

4. Is there a correlation between broadband speed and mobile speed?

5. Is there a correlation between happiness scores and broadband speed and can accurate predictions be made for happiness scores based on given speeds?

These questions will be answered through hypothesis tests, correlation tests, and by applying linear regression concepts.

5 Analysis Results

1. For the first tests of the difference in broadband and mobile speeds, it was necessary to check the variance of the two samples. As they were unequal, it was necessary to perform a one-tailed t-test with unequal variances and two independent samples. It was necessary to form a null hypothesis and an alternative: the null hypothesis stating there is no significant difference in the two variables, and the alternative being that mobile speed is less than broadband. This test was conducted at the 95% significance level, with the null hypothesis being rejected if the calculated p-value was less than 0.05. Having conducted the test, the p-value was less than the significance value. With this there was enough evidence to reject the null hypothesis, suggesting that mobile speeds are less than broadband speeds at a statistically significant level for the samples tested.

2. Being given the confidence intervals for the happiness score, a hypothesis test would be carried out to test the validity of these intervals. Using the mean happiness score and a 95% significance level, the true confidence intervals were calculated. The mean value for the original whisker values was calculated along with the new mean for the true confidence interval. A z-test could then be carried out to test the statistical difference in the two means.

The null hypothesis stated that there was no difference in the two means, while the alternative stated there was a difference. The null hypothesis can be rejected if the resulting p-value was less than 0.05 due to the significance level. After performing a z-test a very high p-value was calculated. Meaning there was not enough evidence to reject the null hypothesis. The whisker confidence intervals can be taken as accurate at a 95% significance level.

3. The third hypothesis test was to test the difference in means between corruption and generosity as factors towards the happiness score. The null hypothesis stated there is no difference between the two variables, with the alternative hypothesis stating there is a difference. With an unequal variance for the variables, a two-sided independent sample t-test was carried out. The null hypothesis could be rejected if the calculated t-statistic was greater than the critical t-value. After performing the test, it was found that the t-statistic was not greater than the critical value and that there was not enough evidence to reject the null hypothesis. This suggests that the effects of generosity and perceptions of corruption were not different at a statistically significant level.

4. To find if a correlation existed between broadband and mobile speed, the covariance of the two columns ‘Broadband Mbps’ and ‘Mobile Mbps’ was calculated as 659.013. This value indicated that the two variables changed together positively. The correlation coefficient was then calculated to be 0.578. This is a moderately strong positive correlation which means the two variables are related to each other. As one speed increases, the other is likely to also increase, based on this data.

5. The covariance and correlation coefficient were calculated for happiness scores against broadband speed in Mbps. The covariance gave a positive value of 2.823, with a moderately strong correlation coefficient of 0.631. A scatter plot showed this positive correlation. It appeared that happiness score increased with faster broadband speeds. A linear regression model could be generated from this data. Once the model was created using the values of both variables, the

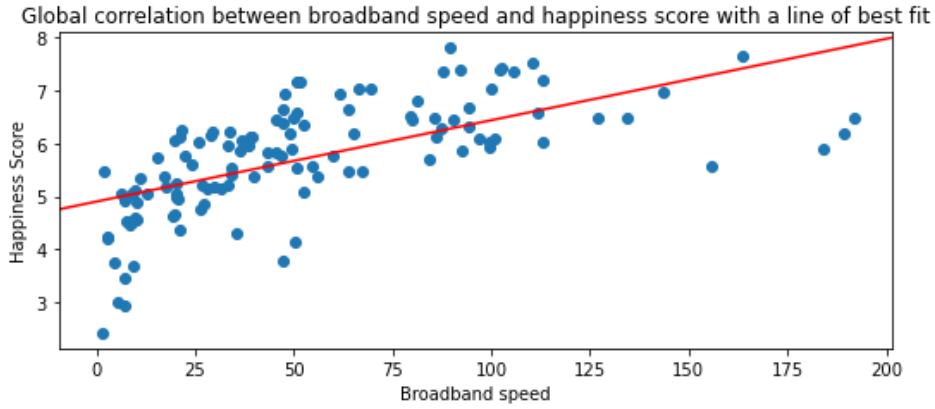

Figure 6 Regression line on scatter plot of broadband and happiness score

model gave an R^2 value of approximately 0.398, a below average score for a regression model. By retrieving the calculated coefficient and intercept of the model, it was possible to plot the regression line as seen in Figure 6.

To try to validate this model, known values of the dataset were put through the model to test the accuracy of its predictions. The model's inputted values proved inaccurate for the expected results as seen in Figure 7.

	Country	Broadband Mbps (input)	Happiness Score	Predicted Happiness Score
0	Afghanistan	1.62	2.404	4.926448
1	Albania	33.50	5.199	5.416082
2	Algeria	9.72	5.122	5.050853

Figure 7 Happiness scores compared with model predicted scores based on their Broadband speeds

6. Conclusion

The analysis of the World Happiness Report and Global Internet Speeds showed a moderately strong positive correlation, though it produced a weak regression model. This could be due to a relatively small sample size of the 121 countries used for this analysis. As all countries were not taken into consideration and all of the data is from one year, it may not be the most optimal data for a strong regression model.

It can be seen from the exploratory data analysis that the affecting factors on happiness score tend to fluctuate outside of GDP primarily. The hypothesis test found that generosity and corruption have a similar effect on the overall happiness score. The given whisker-low/high values in the data set were taken as accurate after testing the hypothesis of the difference of their mean with the calculated confidence intervals' own mean.

A correlation test found a moderately strong positive correlation coefficient for broadband and mobile speed. Though they are likely to increase together, the hypothesis test found that their means' differences are statistically significant.

References

Chauhan, A. (2022) He..he..he... world happiness report 2022, Kaggle. Available at: https://www.kaggle.com/datasets/whenamancodes/world-happiness-report (Accessed: December 20, 2022).

Kanawattanachai, P. (2022) Internet broadband and mobile speeds by country, Kaggle. Available at: https://www.kaggle.com/datasets/prasertk/internet-broadband-and-mobile-speeds-by-country (Accessed: December 20, 2022).

Kemp, S. (2022) Digital 2022: April global statshot report - datareportal – global digital insights, DataReportal. DataReportal – Global Digital Insights. Available at: https://datareportal.com/reports/digital-2022-april-global-statshot (Accessed: December 20, 2022).