Happiness Defined by Country Development Indicators

Benjamin Merrill and Nicholas Dunbar

1. Motivation

Every country in the world has a unique mix of cultures, economies, and people. Many reports provide interesting measurements on the current status of a country's population. Our proposed project is to use the World Development Indicators dataset from the World Bank to better understand the happiness as measured by the World Happiness Reports. We want to know which development indicators of a country increase and decrease happiness.

These macro trends may reveal how to create more proactive policies in government for the good of a country's population.

2. Data Sources

2.1 World Development Indicators Dataset (WDI)

The World Development Indicators dataset is a vast store of information on the current development status of most countries, measuring economic, social, and familial metrics that can help identify the country's current status. The information is gathered and kept by the World Bank.

- **Location**: https://databank.worldbank.org/reports.aspx?source=world-development-indicators
- **Format**: CSV file
- **Time Period**: 2015 - 2019
- **Size**: 50MB
- **Total Records**: <1.2M
- **Important Fields**: Country name, indicator Name, indicator value by year (measuring unemployment Rate, GDP per capita, accessibility to running water, etc.)

2.2 World Happiness Report (WHR)

The World Happiness Report measures the happiness of countries using data from the Gallup World Poll. These 5 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.

- **Location**: https://www.kaggle.com/unsdsn/world-happiness
- **Format**: 5 separate CSV files
- **Time Periods**: 2015, 2016, 2017, 2018, and 2019
- **Size**: 20KB each
- **Total Records**: 200 rows for each file
- **Important Fields**: Country or Region, Happiness Rank, GDP per capita, etc.

3. Data Exploration and Aggregation

3.1 Importing Libraries and Data

This project was completed using an interactive online Python Notebook environment called Google Colaboratory. Once opened, we imported Pandas and NumPy, two Python data processing libraries, for the aggregation. We then imported each file into a Pandas dataframe. This included 5 files representing the World Happiness Reports from 2015 to 2019 and the single file representing the World Development Indicators.

3.2 Happiness Report Aggregation

After importing the files, we began our aggregation on the 5 happiness dataframes. We were able to concatenate all 5 reports using the `'Country or region'` columns from each dataframe. This process was completed by:

1) Adding an additional field `'year'` to each dataframe to represent the year the survey was taken

2) Creating uniform column titles for each column type. This includes dropping some fields from 2015, 2016, and 2017 that are not represented in all 5 happiness reports.

3) Aggregating the 5 files vertically using `pandas.concat()` function

4) Consolidating the `'Country or region'` field to only have countries that are represented in all 5 files. This included reformatting country names year-to-year and ultimately dropping countries (i.e. Djibouti, Suriname, Swaziland, etc.) that were not data dense

3.3 Exploratory Data Analysis

Prior to aggregating our two main sources, it was helpful to explore each source on its own to gather information about their descriptive statistics and general state of the data.

Once the happiness reports were combined into one dataframe, we able to start our exploration by examining the set of countries located in the set. This was already briefly touched on when country names were reformatted and sparse countries were dropped. Our combined happiness dataframe

listed 147 unique country names throughout our set and an overall length of 735 rows (147 countries listed once for each year of data). Applying `.describe()` helped to provide a quick look at the mean values of each column which met expectations, and `.isna().sum()` showed that the data was without any NaN values.

Exploration of the World Development Indicator source was a bit more complex due to the size of the dataset and state of the contained data. Descriptive statistics of the dataframe showed 219 unique country names and 217 country codes; as such, we planned to condense this list down to match the countries contained in our happiness data. `.describe()` also showed us that the data contained 1443 unique indicators throughout the 313,131 rows. These numbers led us to a few options as to how to handle finding the best indicator data. First, we were able to split the ['Series Code'] column into smaller components and check those against a dictionary of topics and subjects available from the World Bank. The series codes contained indicator topics, general subjects, specific subjects, and extensions. Given the amount of potential options, we settled on looking to the roughly 60 topics represented for possible further examination.

One major factor in the WDI data was the missing data throughout the set. If we even just looked at a general sum of the NaN values, a given year's data was missing anywhere from 139,000 to 182,000 values. Given the wide number of missing values, we made sure to stay aware of these missing counts as the sets were aggregated and analyzed.

3.4 **Combining World Development Indicators with World Happiness**

- 3.4.1 **Finalizing List of Countries and Their Rank**

Similarly to the aggregation of the happiness reports, there are a large number of mismatching countries between the WDI and WHR data. We took a look at these and reformatted the countries represented in both dataframes to match the format of the WHR data. We then queried both datasets for only the countries represented in both datasets. With our final list of 144 countries, we fixed the `'Overall rank'` field in the WHR dataframe to represent data rank for each year from 1 to 144, overwriting the previous ranking's missing numbers.

- 3.4.2 **Reformat Happiness Data to Match WDI Fields**

The column formats of the WDI and WHR datasets mismatch. On the WHR dataset, we used functions `pd.DataFrame.stack()` and `pd.DataFrame.unstack()` to get most of the columns to match their counterparts in WDI. We also created two new columns in the WHR dataframe, `'Country Code'` and `'Series Code'`, for their counterparts in the WDI dataframe.

- 3.4.3 **Aggregating and Cleaning the Final Dataset**

With matching fields in the WDI and WHR datasets, we once again utilized `pd.concat()` to vertically concatenate the two dataframes.

3.5 Outputting the Data

We outputted three CSV files using the `pd.DataFrame.to_csv()` function:

- `wdi_happiness` (**the most important**, representing both WDI data and happiness data)
- `wdi_formatted` (WDI data only)
- `happiness_formatted` (happiness data only)

3.6 Comments on Aggregation

The final output file `wdi_happiness` is data dense. This density came at a cost, as many fields were dropped in the aggregation.

In the aggregation of the happiness reports in 3.2, we had to drop a number of columns due to the fact that they were not represented in all 5 reports. The bulk of these dropped fields were measurements of uncertainty in a country's happiness score calculation. The omission of these uncertainty metrics makes each country's happiness score look like a flat value, when it may be likely that some countries have a lot of uncertainty around the calculated happiness score. This is an important consideration because results about countries with a smaller standard error tend to be more true to the real world status of a country than those with a larger standard error.

We dropped many rows in both datasets due to countries which were not represented in all files in 3.2 and 3.4.1. This included countries such as Lesotho, Belize, Comoros, and Suriname. Because our analysis covers the correlation between happiness scores and development indicators, we needed to drop any countries in the WDI dataset that were not represented in the WHR dataset. We also dropped countries not represented in all 5 happiness reports, as it is difficult to build a reliable correlation between the many indicators and happiness when there is significant missing data among the happiness scores. Dropping these countries is important to consider, as many of the countries that are not represented in the final dataset tend to be from either in the third world or small island nations. This may imply that our results are not applicable to the entire world, only larger, wealthier nations.

4. Analysis of Correlated Fields and Visualizations

4.1 Creating the `correlate_happiness_and_indicators()` Function

Our next goal was to create a meaningful correlation between happiness and our desired indicators. Our created function, `correlate_happiness_and_indicators()`, helps create meaningful correlation matricies that we can use to measure our happiness correlations.

This function takes 3 arguments - `data`, `bins`, and `bin_by_year`. First, we input the data queried for our desired indicators from our `wdi_happiness` dataframe. This should typically be somewhere between 10-30 indicators that we want to study. Since we only have 5 values for each country's happiness score, we decided to bin the data by the countrys happiness of a given year. For instance, if we decide `bins=5` and `bin_by_year='2017'`, the function will create 5 sorted bins of countries, the first bin representing the first happiest fifth of the countries in 2017, the second bin representing the second happiest fifth of countries, and so on. This helps generalize about the trends in the most and least happy countries. For each bin, we then calculate the correlation between happiness score and our indicators. We then output a dataframe with the indicator correlations sorted into equal bins based on happiness.

4.2 Creating a Correlation Heatmap

Now that the correlation dataframe has been created, we were able to use Seaborn's heatmap function to visualize our calculated correlations. The happiness bins are encoded to the x-axis and indicators to the y-axis, with the corresponding correlation encoded to a red/yellow/blue color range. We selected the indicators with the strongest negative and positive correlations with happiness for a given year - 2017.

From this data, we can see that generally, negative and positive correlations go together year-to-year, but strength of a correlation tends to be an outlier. Due to the variance in the level of correlation year-to-year, it is difficult to conclude that happiness has a consistent correlation with these indicators.

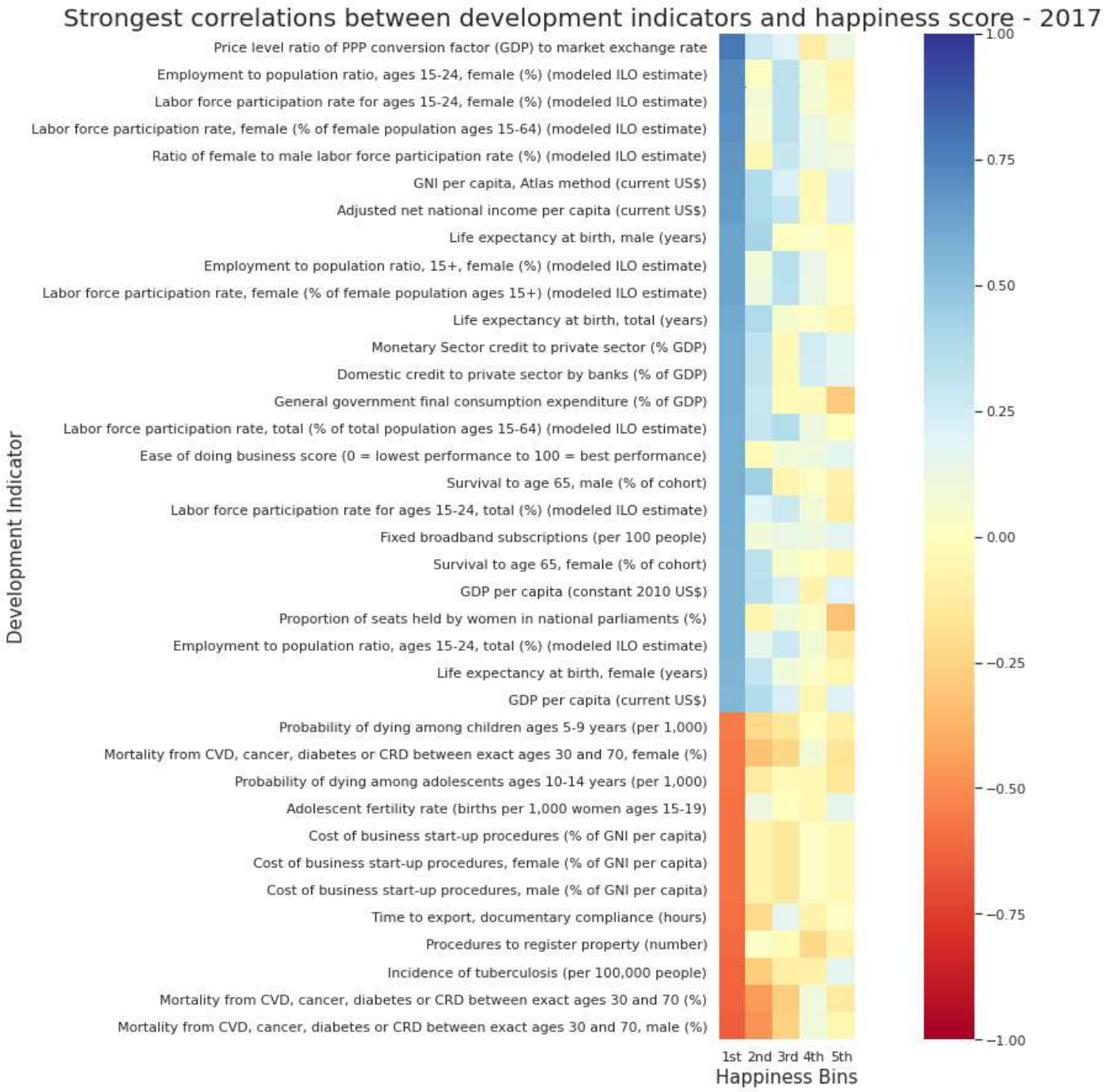


Figure 1. Heatmap showing the strongest correlations between development indicators and happiness scores in 2017. Values were taken from the correlation matrix described in Section 4.1, and selected by a threshhold of above 0.55 or below -0.55.

4.3 Conclusions from our heatmaps

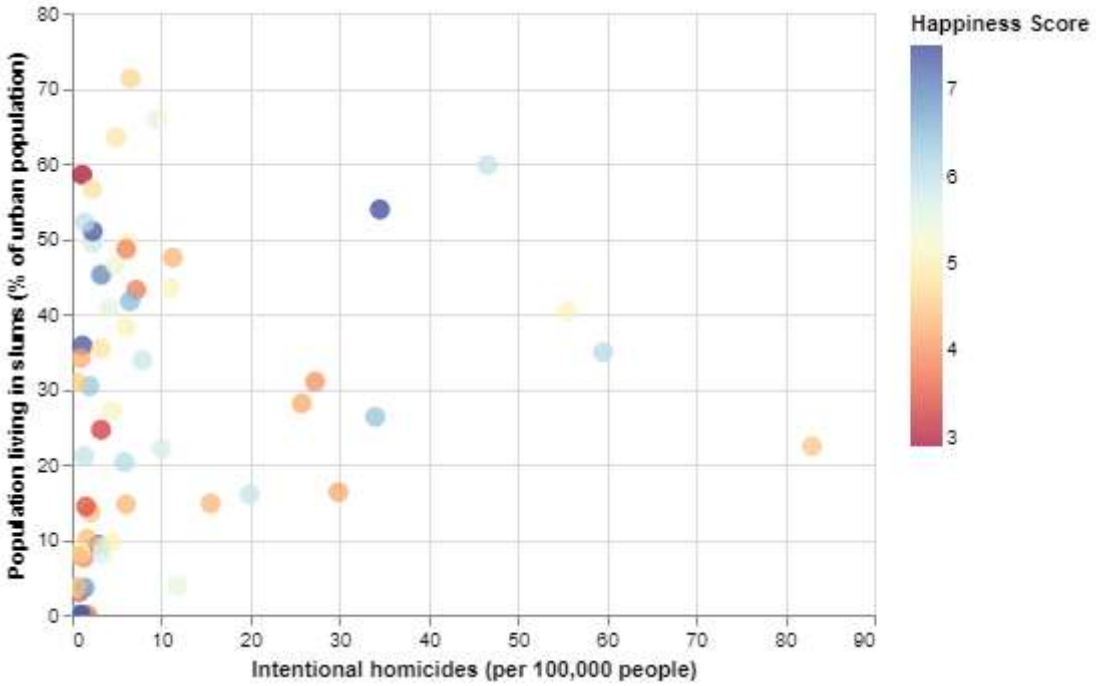
Once we visualize the strongest correlations among development indicators and happiness, a few trends become apparent. Most of the indicators with a strong influence on a country's happiness tend to be either employment or health related.

On the positively correlated side of the data, we see indicators that include employment rates, life expectancy, or gross income. Negatively correlated, we see indicators that include things like mortality rates, serious diseases, or teen pregnancy. These relationships are maintained somewhat across our different happiness bins, but come closer to zero as we move down the bins.

4.4 Creating Scatter Visualizations

From our main combined dataset, we were also able to create a series of scatter plots that examine the relationship between development indicators while also looking into patterns in a country's happiness score. To achieve this, we created a pair of functions that work together to pull relevant values together and create an Altair visualization. `vis_frame` and `vis_frame_vis` take arguments for our aggregated dataframe, the two indicator series codes of interest, and the intended year to output a scatter plot which encodes the measures on the x- and y-axes, with the country's happiness score encoded on a color scale.

An example of this scatterplot visualization is below, vizualizing the difference between Population living in slums and intentional homicides. In the interactive Jupyter environment, one can scroll over a specific point for each country, realizing the relationship between these two indicators for a given year. Here we see that all of the happier countries have less than 55% of the population living in slums (shockingly!) and that homicides don't have a huge correlation with happiness in the World Happiness Report. With this function, we could make more scatter plots of the same type, comparing two indicators of interest.



5. Conclusion

After examining our data through the use of our heatmap and scatter visualizations, we feel that heatmaps provided us with the most useful information taken from the combination of these two datasets. Selection based on series code topics might make the process more manageable, but overall the roughly one million combinations make it much more likely that significant results might come from cherry-picking the strongest relationships. That would not be dissimilar from our heatmap that pulls the strongest correlations, but searching by results should be kept to a minimum wherever possible. All in all, heatmaps provide a great visualization for the analysis of correlations between fields when exploring a dataset.

6. Statement of Work

- Motivation: Nick and Ben
- Data Sources: Nick and Ben
- Data Manipulation Methods: Ben
- Analysis and Vizualization: Nick