2.2. Case Study 1: The Happiness Report

Before proceeding, you need to have read the required readings about the data science processing pipeline, exploratory **data analysis**, and the other resources on spreadsheets and summary statistics linked in the previous section.

2.2.1. Introducing the Happiness Report

The World Happiness Report is a landmark survey of the state of global happiness. The World Happiness Report 2018 ranks 156 countries by their happiness levels, and 117 countries by the happiness levels of their immigrants. Many factors may contribute to the happiness of a country, and we will use spreadsheets to explore and analyze what factors may be most important in determining a country's happiness.

We will start by loading the happiness_2017.csv (../_static/happiness_2017.csv) file into Google Sheets. The list below gives a bit of detail about each of the columns on the spreadsheet.

The following definitions are reproduced from World Happiness Report 2018 (http://worldhappiness.report/ed/2018/).

1. GDP per capita is in terms of Purchasing Power Parity (PPP) adjusted to constant 2011 international dollars, taken from the World Development Indicators (WDI) released by the World Bank in September 2017. See Appendix 1 for more details. GDP data for 2017 are not yet available, so we extend the GDP time series from 2016 to 2017 using country-specific forecasts of real GDP growth from the OECD Economic Outlook No. 102 (Edition November 2017) and the World Bank's Global Economic Prospects (Last Updated: 06/04/2017), after adjustment for population growth. The equation uses the natural log of GDP per capita, as this form fits the data significantly better than GDP per capita.
2. The time series of healthy life expectancy at birth are constructed based on data from the World Health Organization (WHO) and WDI. WHO publishes the data on healthy life expectancy for the year 2012. The time series of life expectancies, with no adjustment for health, are available in WDI. We adopt the following strategy to construct the time series of healthy life expectancy at birth. First, we generate the ratios of healthy life expectancy to life expectancy in 2012 for countries with both data. We then apply the country-specific ratios to other years to generate the healthy life expectancy data.
3. Social support is the national average of the binary responses (either 0 or 1) to the Gallup World Poll (GWP) 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?"
4. Freedom to make life choices is the national average of binary responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
5. Generosity is a function of the national average of GWP responses to the question "Have you donated money to a charity in the past month?" on GDP per capita.
6. Perceptions of corruption are the average of binary answers to two GWP questions: "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?". Where data for government corruption are missing, the perception of business corruption is used as the overall corruption-perception measure.
7. Positive affect is defined as the average of previous-day affect measures for happiness, laughter, and enjoyment for GWP waves 3-7 (years 2008 to 2012, and some in 2013). It is defined as the average of laughter and enjoyment for other waves where the happiness question was not asked.
8. Negative affect is defined as the average of previous-day affect measures for worry, sadness, and anger for all waves.

In this first part, we will review and practice some spreadsheet calculations by doing some exploratory data analysis. If you have never used a spreadsheet before, don't worry, you will catch on quickly. Remember that we are just exploring at this point, so there isn't necessarily a right answer. Most of the time, we don't know what the right answers are while we are in exploring mode. You might even be wondering what it means to be in exploring mode. The main thing we do is look at the data and seek out things that look like 'interesting' bits of statistics that stand out. We also think about how things might correlate or what variables might be interdependent on others. Two of the primary tools we use in this exploring mode are summary statistics and visualization.

Q-1: Which of the columns in the happiness spreadsheet represent categorical (nominal) data? ✅

8

- [x] A. Country
- [x] B. Region
- [ ] C. Rank
- [ ] D. Generosity

Check Me Compare me

✓
A. Correct
B. Correct

Activity: 2.2.1.1 Multiple Choice (dat_cat)

Q-2: Which of the columns in the happiness spreadsheet represent ordinal data? ✅

- ○ A. Generosity
- ○ B. Happiness Score
- ◉ C. Rank
- ○ D. Country

Check Me Compare me

✓
Correct

Activity: 2.2.1.2 Multiple Choice (dat_ord)

Q-3: Which of the columns in the happiness spreadsheet represent ratio data? ✅

- [x] A. Happiness Score
- [x] B. Generosity
- [ ] C. Rank
- [x] D. Confidence in national Government

Check Me Compare me

✓
A. Correct
B. Correct
D. Correct

Activity: 2.2.1.3 Multiple Choice (dat_ratio)

2.2.2. Happiness Index Research Questions

1. What are the different factors that lead to the happiness of a country?
2. What role does the economy play in determining the happiness of a country?
3. Which factor, on average, contributes most/least to happiness?
4. What similarities and differences do the countries experiencing the highest/lowest WHS have? Are there any countries where their scores for some factor are very different than those of the countries around it in the rankings?
5. Does being in a certain region (continent) have any correlation to the average score of countries?
6. How have the happiness numbers changed over time? Which countries have increased the most? Which countries have decreased the most?

7. For the countries with the largest increase which factors changed the most? Are those factors the same as you identified in the first 3 questions?

2.2.3. Summary Statistics

Q-4: Match the Term on the left with the description on the right.

	The sum of all values, divided by the number of values
	Mean
	The middle value in all your data
	Median
	The value that occurs most often
	Mode
	A measure of how spread out your data is
	Standard Deviation

Check me Reset

Activity: 2.2.3.1 Drag-N-Drop (dd_summarystats)

1. Although the countries are ranked from most happy to least happy, we might want to start by looking at some summary statistics for the happiness score.
    a. Use the `AVERAGE` function of sheets to calculate the mean in column D. Scroll down and click in cell D143. That should be an empty cell below the column of numbers for the happiness score. Now type `=AVERAGE(D2:D141)`. You can also type `=AVERAGE(` and then click and drag the numbers you want. D2:D141 specifies a range, from Column D Row 2 down to Column D Row 141.

Q-5: Calculating the average happiness score. You should include three digits to the right of the decimal point.|blank|

5.399

Check me Compare me

✓ Is the correct answer

Activity: 2.2.3.2 Fill in the Blank (fb_avghappiness)

Since you are going to be entering numbers to 3 digits, you can use a custom number format under the Format menu, to have Sheets automatically display your values correctly rounded to just three digits to the right of the decimal point.

b. Many formulas in Sheets use ranges. Ranges can span cells in a single column like we did in a. Or, they can span cells in a single row such as A1:L1. They can even span rows and columns to form a rectangle such as A1:L141.

c. Now calculate the `STDEV` and `MEDIAN` for the Happiness Score column as well. If you are fuzzy about **standard deviation**, this article (https://towardsdatascience.com/intro-to-descriptive-statistics-252e9c464ac9) is a nice intuitive explanation.

Q-6: What is the standard deviation of the happiness score? ✅

1.11 Again, you only need to include three digits to the right of the decimal point.

Check me Compare me

✓
Is the correct answer

Activity: 2.2.3.3 Fill in the Blank (fb_stdhappiness)

d. We can calculate the same statistics for the other columns by copying and pasting the formula to the cells under the other columns. As a shortcut, you can also click on the square in the lower right corner of the currently selected cell and drag it.

e. After you have copied and pasted the formula for `AVERAGE` to cells E143 to N143 click in N143. The formula there looks like `=AVERAGE(N2:N141)`. Notice that Sheets is smart about changing the cell references when you copy/paste a formula.

Q-7: The mean value for healthy life expectancy is 63.441 and ✅

the standard deviation is 7.596 .

Check me Compare me

- ✓
  Is the correct answer
- ✓
  Is the correct answer

Activity: 2.2.3.4 Fill in the Blank (fb_life_exp)

f. If you do NOT want Sheets to change the cell references when you are copy/pasting you can use a $ in front of the row or the column, which tells sheets to “leave this reference alone”. We see some examples of this later.

2.2.4. Visualizing Happiness

1. The STDDEV value tells us that the majority of the values are between 4.0 and 6.6 so let us check that graphically. It is easy to make a **histogram** of the values in Sheets. Note: To do this same thing in Excel you would need to install an extension.
   a. Click on the insert graph icon.
   b. Choose chart type of histogram.

c. Enter or drag the rows in column D. It should look like most of the bars are between 4 and 6.6 on your histogram?
d. Try editing the details of the histogram to look at the distribution in other columns.

2. Because we are exploring you might also wonder "which country has the largest GDP, or which country scores the highest on Family, or Generosity? Learning about minimum and maximum values can definitely lead you in interesting directions. It is also a great chance to learn a couple of other really powerful functions. Let's explore which country has the highest score in the Generosity column.

a. Start by finding the maximum value in the generosity column, putting the result in cell J146.

Q-8: What is the maximum value in the generosity column? ✅

0.629

Check me Compare me

✓
Is the correct answer

Activity: 2.2.4.1 Fill in the Blank (gen_max)

b. Knowing the maximum is one thing, but that does not tell us which country it corresponds to. For that, we will use the `MATCH` and `INDEX` functions. `MATCH` allows us to search for a value in a range of cells, just like the search function in a word processor. In cell J147, type `=MATCH(J146, J2:J141, 0)`. The `MATCH` function looks for the value in cell J146 in the range J2:J141 and the 0 tells it that the data is not sorted. If you leave out the 0, Sheets will assume that the data is sorted, stop searching, and return the first cell it finds that is greater than the value in J146.

Q-9: The index of the row containing the maximum value is ✅

105 .

Check me Compare me

✓
Is the correct answer

Activity: 2.2.4.2 Fill in the Blank (gen_max_row)

c. In cell J148, type `=INDEX(A2:A141, J147)`. This tells Sheets to return the value from the range A2:A141 in the row specified by the value in J163. As we will see later, `INDEX` is a really powerful tool for doing all kinds of things, but for now we will primarily think of the combination of `MATCH` and `INDEX` as being our search and retrieve power tools.

Q-10: The name of the country that is most generous is ✅

Myanmar .

Check me Compare me

✓
Is the correct answer

Activity: 2.2.4.3 Fill in the Blank (gen_max_country)

d. We broke this process into three steps to make it clear what we were doing. But they can be combined into a single cell by nesting the functions. Let's figure out which country gets the lowest score for Generosity, but in one cell. In J165 enter `=INDEX($A2:$A141, MATCH(MIN(J2:J141), J2:J141, 0))`. Here we are using the fact that `MATCH` and `MIN` each return values, and rather than have them visible in a cell for us to look at, we can just use them directly as parameters to another function. That probably seems pretty logical to you since you have done this in Python many times.

Q-11: The country with the lowest generosity score is Greece .

Check me Compare me

✓
Correct

Activity: 2.2.4.4 Fill in the Blank (gen_min)

e. Now you should practice by finding the names of the countries that have the minimum and maximum values for some other columns.

Q-12: What is the name of the country that has the highest confidence in their national government? Uzbekistan

Check me Compare me

✓
Is the correct answer

Activity: 2.2.4.5 Fill in the Blank (gov_conf)

Q-13: What is the name Ukraine and happiness score 4.096 of the country with the lowest confidence in their national government?

Check me Compare me

- ✓
  Is the correct answer
- ✓
  Is the correct answer

Activity: 2.2.4.6 Fill in the Blank (gov_conf_min)

f. If you tried to copy/paste the functions from above you likely ran into some errors. Check the ranges carefully and remember what Sheets do when you copy and paste. If you insist on copy/pasting, then you are going to have to use $ to get it right. We'll leave it to you to figure that out.

3. One great way to get an overview of the data visually is to make a **choropleth**. A choropleth combines the geographic data with some other data such as the happiness score. Sheets makes it very easy to graph data by country.
    a. Click on the insert graph icon.
    b. Choose Geo Chart.
    c. Use the country column and the happiness score column.
    d. Experiment with using other columns such as freedom or generosity.
4. The exploration of the happiness scores and the different factors related may have you wondering which factors lead to some people being happier than others. Is it their level of freedom, or their level of wealth? One way we can answer this question is to calculate a correlation between the happiness index and the various factors. This will create a small table that computes a correlation score between of our columns of data. Happiness score to Economy, Happiness score to Family, etc.
    a. First, let's calculate a correlation between happiness score and each other factor.
    b. To do this, we can use the `CORREL` function, which calculates a **Pearson correlation** between two ranges of data. Because we want to always keep the happiness index as one of the columns, we will anchor that column using $ and but not the other columns. This will allow us to copy the formula across.
5. We might now try to focus in on the characteristics of the most happy countries and the least happy countries.

    a. Recompute the correlation scores, but don't do it for all of the countries. Do it only for the top 25 and bottom 25. What stands out for you?

    b. Calculate the mean value for each of the factors for the 25 and bottom 25 countries and then calculate a difference between these values. Which have the largest and smallest difference?
6. Another interesting exercise we could do is to identify some countries where their scores in some category like generosity are significantly different from the countries around them. For example the country of Myanmar stands out on a Choropleth as being the most generous country. Yet its happiness rank is 114. Its Generosity score is 0.8 but the country right above it has a score of 0.3 and the country below 0.1.
    a. For each country, compute the total difference between its score and the country above it and below it.
    b. Then, you can apply some conditional formatting to help visually pick out the outliers.
    c. You can also sort the region containing the rankings based on this column to gather together the countries with significant differences from their neighbors. WARNING: Sorting by a calculated column like this will lead to unexpected results. Copy this column and do a paste special where you paste only the values before sorting.

Lesson Feedback

During this lesson I was primarily in my...
- ○ 1. Comfort Zone
- ◉ 2. Learning Zone
- ○ 3. Panic Zone

Activity: 2.2.4.7 Poll (LearningZone_2_1)

Completing this lesson took...
- ◉ 1. Very little time
- ○ 2. A reasonable amount of time
- ○ 3. More time than is reasonable

Activity: 2.2.4.8 Poll (Time_2_1)

Based on my own interests and needs, the things taught in this lesson...
- ○ 1. Don't seem worth learning
- ○ 2. May be worth learning
- ◉ 3. Are definitely worth learning

Activity: 2.2.4.9 Poll (TaskValue_2_1)

For me to master the things taught in this lesson feels...

- (•) 1. Definitely within reach
- ( ) 2. Within reach if I try my hardest
- ( ) 3. Out of reach no matter how hard I try

Activity: 2.2.4.10 Poll (Expectancy_2_1)

You have attempted 18 of 18 activities on this page

✔ Completed. Well Done!

(introduction.html) (cs1_more_happir

This page is not part of the last reading assignment you visited.