Homework 3

Fall 2016, Math 107, Prof. Adam Loy

Due Friday, September 23 by 4:00 p.m.

Problem 1

For each of the following, describe whether you expect the distribution to be symmetric, right-skewed, or left-skewed. Also specify whether the mean or median would best represent a typical observation, and whether the variability of observations would be best represented using the standard deviation or IQR.

a. Housing prices in a country where 25% of the houses cost below \$350,000, 50% of the houses cost below \$450,000, 75% of the houses cost below \$1,000,000 and there are a meaningful number of houses that cost more than \$6,000,000.

b. Housing prices in a country where 25% of the houses cost below \$300,000, 50% of the houses cost below \$600,000, 75% of the houses cost below \$900,000 and very few houses that cost more than \$1,200,000.

c. Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

d. Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than all the other employees.

Problem 2

The website TED.com offers free short presentations, called TED Talks, on a variety of interesting subjects. One of the talks is called "The Happy Planet Index," by Nic Marks. [1] Marks comments that we regularly measure and report economic data on countries, such as Gross National Product, when we really ought to be measuring the well-being of the people in the countries. He calls this measure Happiness, with larger numbers indicating greater happiness, health, and well-being.

You can find a tidy version of the 2012 Happy Planet Index (`hpi-tidy.csv`) on the course webpage (An messy version can be downloaded from http://www.happyplanetindex.org/data/). In this homework problem you will explore the Happy Planet Index using R. **Please use R markdown to complete this part of the assignment.** You should **knit to Word or PDF** and submit this document. You can find an R markdown template on the course webpage.

A basic description of all of the variables included in the data set is given below:

Variable	Description
HPIRank	HPI rank for the country
Country	Name of country
LifeExpectancy	Average life expectancy (in years)
Wellbeing	"Ladder of Life" index from the Gallup World Poll (0 = worst possible life, 10 = best possible life)
HappyLifeYears	Index variable combining life expectancy and well-being
Footprint	Ecological footprint–a measure of the per capita ecological impact
HappyPlanetIndex	Happy Planet Index (0-100 scale)
Population	Population (in millions)
GDPcapita	Gross Domestic Product (per capita)
GovernanceRank	Governance ranking (1 = highest)

[1] Marks, N. "The Happy Planet Index," www.TED.com/talks, August 29, 2010.

Variable	Description
Region	Region of the world

a. What are the cases?

b. List each variable in the data set and classify it as either quantitative or categorical.

c. Create a histogram of the Happy Planet Index scores and decribe the distribution, mentioning the number of modes, the shape, and the absence/presence of outliers. Remember that you should experiment with the bin-width until you find one that works well.

d. Create a density plot of the ecological footprint and decribe the distribution, mentioning the number of modes and the shape.

e. Calculate the mean and standard deviation of the Happy Planet Index scores. Why do we prefer to report the mean and standard deviation in this situation?

f. Calculate the five-number summary for ecological footprint. Why do we prefer to report the five-number summary in this situation?

g. Create side-by-side boxplot of life expectancy by region and briefly describe what you learn from the plot.

h. Create overlaid density plots of life expectancy by region. What aspects of the distributions are easier to see using this these density plots than the boxplots you created in the previous part? What aspects are harder/impossible to see?

i. Create a scatterplot of happy life years (on the y-axis) against ecological footprint (on the x-axis). Describe the relationship between happy life years and ecological footprint, commenting on the direction, form, and strength, along with any outlying observations.

j. Create another scatterplot of happy life years against ecological footprint, but this time use different colors and shapes to represent the region for each country. What do you learn by adding this extra information to the plot?