Codebook - Doing Data Science Case Study 2

Introduction

Joe's: github

The following datasets are referenced in case study rmarkdown. Some are simulated for instructional purposes. Others are obtained from publicly accessible avenues, or scraped for educational purposes. Please be sure to give credit if warranted.

Disclaimer: These data sets are not meant to be used as a substitute for real data or construed as advice for shaping legal, business, or political decisions. They are intended to be used for educational purposes only.

This data folder contains the following:

- procrastination.csv - a data frame concerning survey responses around the the world regarding various subject matter including demographic and procrastination information about each respondent.
- Human Development Index tables online (https://en.wikipedia.org/wiki/List_of_countries_by_Human_Development_Index#Comple te_list_of_countries). This is a list of all the countries by the Human Development Index as included in a United Nations Development Programme's Human Development Report. The Human Development Index (HDI) is a composite statistic of life expectancy, education, and income per capita indicators. A country scores higher HDI when the life expectancy at birth is longer, the education period is longer, and the income per capita is higher. It is used to distinguish whether the country is a developed, a developing or an underdeveloped country.
- Finalized HDI table download as hdicomb.csv
- Tidied version of the original input to be output in the repository, including the merged HDI data downloaded as procrastcomb.csv
- A dataset that shows the Top 15 nations as well based on 2 of their HDI scores. Downloaded as Top15aip.csv and top15dp.csv

procrastination.csv details

CSV concerning survey responses around the the world regarding various subject matter including demographic and procrastination information about each respondent.

.

- Type: Comma-separated value file
- Dimensions: 4264 observations X 61 variables
- Unit Used: Case Study 2
- Public: unknown

Variable Information:

Renamed Variables	Variables from original file
Age	Age
Gender	Gender
Kids	Kids
Edu	Edu
WorkStatus	Work Status
AnnualIncome	Annual Income
CurrentOccup	Current Occupation
Timeinposyr	How long have you held this position?: Years
Timeinposmn	How long have you held this position?: Months
Communitysize	Community size
Countryres	Country of residence
MaritalStatus	Marital Status
Numsons	Number of sons
Numdaughters	Number of daughters
DP1int	(DP 1) I waste a lot of time on trivial matters before getting to the final decisions
DP2int	(DP 2) Even after I make a decision I delay acting upon it

DP3int	(DP 3) I don’t make decisions unless I really have to
DP4int	(DP 4) I delay making decisions until it’s too late
DP5int	(DP 5) I put off making decisions until it’s too late
AIP1	(AIP 1) I pay my bills on time
AIP2	(AIP 2)I am prompt and on time for most appointments.
AIP3	(AIP 3)I lay out my clothes the night before I have an important appointment, so I won’t be late
AIP4	(AIP 4) I find myself running later than I would like to be
AIP5	(AIP 5) I don’t get things done on time
AIP6	(AIP 6) If someone were teaching a course on how to get things done on time, I would attend
AIP7	(AIP 7) My friends and family think I wait until the last minute.
AIP8	(AIP 8) I get important things done with time to spare
AIP9	(AIP 9) I am not very good at meeting deadlines
AIP10	(AIP 10) I find myself running out of time.
AIP11	(AIP 11) I schedule doctor’s appointments when I am supposed to without delay
AIP12	(AIP 12) I am more punctual than most people I know
AIP13	(AIP 13) I do routine maintenance (e.g., changing the car oil) on things I own as often as I should
AIP14	(AIP 14)When I have to be somewhere at a certain time my friends expect me to run a bit late
AIP15	(AIP 15)Putting things off till the last minute has cost me money in the past
GP1int	(GP 1)I often find myself performing tasks that I had intended to do days before
GP2int	(GP2) I often miss concerts, sporting events, or the like because I don’t get around to buying tickets on time
GP3int	(GP 3) When planning a party, I make the necessary arrangements well in advance
GP4int	(GP 4) When it is time to get up in the morning, I most often get right out of bed
GP5int	(GP 5) A letter may sit for days after I write it before mailing it possible
GP6int	(GP 6) I generally return phone calls promptly
GP7int	(GP 7) Even jobs that require little else except sitting down and doing them, I find that they seldom get done for days
GP8int	(GP 8) I usually make decisions as soon as possible

GP9int	(GP 9) I generally delay before starting on work I have to do
GP10	(GP 10) When traveling, I usually have to rush in preparing to arrive at the airport or station at the appropriate time
GP11	(GP 11) When preparing to go out, I am seldom caught having to do something at the last minute
GP12	(GP 12) In preparation for some deadlines, I often waste time by doing other things
GP13	(GP 13) If a bill for a small amount comes, I pay it right away
GP14	(GP 14) I usually return a “RSVP” request very shortly after receiving it
GP15	(GP 15) I often have a task finished sooner than necessary
GP16	(GP 16) I always seem to end up shopping for birthday gifts at the last minute
GP17	(GP 17) I usually buy even an essential item at the last minute
GP18	(GP 18) I usually accomplish all the things I plan to do in a day
GP19	(GP 19) I am continually saying “I’ll do it tomorrow”
GP20	(GP 20) I usually take care of all the tasks
SWLS1	(SWLS 1) In most ways my life is close to my ideal
SWLS2	(SWLS 2)The conditions of my life are excellent
SWLS3	(SWLS 3) I am satisfied with my life.
SWLS4	(SWLS 4) So far I have gotten the important things I want in life
SWLS5	(SWLS 5) If I could live my life over, I would change almost nothing
UCONSPRO	Do you consider yourself a procrastinator?
OTHCONSPRO	Do others consider you a procrastinator?

Human Development Index

This is a list of all the countries by the Human Development Index as included in a United Nations Development Programme's Human Development Report. The Human Development Index (HDI) is a composite statistic of life expectancy, education, and income per capita indicators.

- Type: HTML Web page tables
- Dimensions: 188 observations X 6 variables
- Unit Used: Case Study 2
- Public: yes (https://en.wikipedia.org/wiki/List_of_countries_by_Human_Development_Index#Complete_list_of_countries)

Variable Information:

Rank	
RankChng	
Country	
HDI	
HDICHNG	

Finalized HDI table

- Name: hdicomb.csv
- Type: CSV
- Dimensions: 188 observations X 6 variables
- Public: yes

Tidied version of the original input to be output in the repository, including the merged HDI data

- Name: procrastcomb.csv
- Type: CSV
- Dimensions: 188 observations X 6 variables
- Public: yes

A dataset that shows the Top 15 nations as well based on 2 of their HDI scores.

- Name: Top15aip.csv and top15dp.csv
- Type: CSV
- Dimensions: 3620 observations X 67 variables
- Public: yes

General R Code Flow

Clean your Raw Data (10%)

a Read the csv into R. Outputted how many rows and columns the data.frame has.

b The column names were either too much or not enough. Changed the column names so that they do not have spaces, underscores, slashes, and the like. All column names should be under 12 characters.

c Some columns are, due to Qualtrics, malfunctioning. Prime examples are the following columns: “How long have you held this position?: Years”, Country of residence, Number of sons, and Current Occupation.

i Some have impossible data values. For example fixed improbably years on job.

ii Somehow, “Number of sons” was labeled with Male (1) and Female (2). Changed these incorrect labels back to integers.

iii There are no “0” country of residences. Treated them as missing.

iv Current Occupation has no “please specify” or “0.” Treated them as missing. Some jobs are quite similar. Used judgment calls to overwrite them into the same category. It's not 100% accurate, but for example “ESL Teacher” would not be counted as “teacher” if there were unique counts.

d Made sure your columns are the proper data types (i.e., numeric, character, etc.). If they were incorrect, converted them.

e Each variable that starts with either DP, AIP, GP, or SWLS is an individual item on a scale. Calculated a DPMean column, an AIPMean column, a GPMean column, and a SWLSMean column. This represents the individual’s average decisional procrastination (DP), procrastination behavior (AIP), generalized procrastination (GP), and life satisfaction (SWLS).

3. Scraped the Human Development Index tables online (20%) (https://en.wikipedia.org/wiki/List_of_countries_by_Human_Development_Index#Comple te_list_of_countries).

a There are 8 tables, cleaned them as to be usable, and then bound them into one singular table. Only Country and 2016 Estimates for 2015 columns were needed for the final table.

b Created a new column for this final scraped table which categories the Countries like the original page (Very high human development, High human development, Medium human development, Low human development). After these categories, outputted a csv file of this table to the repository.

c Merged HDI data frame to the Country of Residence column of Procrastination.csv so that data now has an HDI column and HDI categories (Very high human development, etc.).

4. Preliminary Analysis

a Removed all observations where the participant is under age 18. No further analysis of underage individuals is permitted by the client.

b Provided (in pretty-fied table format or similar), descriptive statistics on Age, Income, HDI, and the four mean columns (DP, etc.). Created a simple histogram for two of these seven variables. Commented on the shape of the distribution in your markdown.

c Gave the frequencies (in table format or similar) for Gender, Work Status, and Occupation.

Gave the counts (again, pretty table) of how many participants per country in descending order.

e Analysis on perceptions completed. There are two variables in the set: whether the person considers themselves a procrastinator (yes/no) and whether others consider them a procrastinator (yes/no). Calculated how many people matched their perceptions to others' (so, yes/yes and no/no)? To clarify: how many people said they felt they were procrastinators and also said others thought they were procrastinators? Likewise, how many said they were not procrastinators and others also did not think they were procrastinators?

5. Deeper Analysis and Visualization

A. Created a barchart in ggplotr which displays the top 15 nations in average procrastination scores, using one measure of the following: DP. The bars are in descending order, with the number 1 most procrastinating nation at the top and 15th most procrastinating at the bottom. Omitted all other nations. Colored the bars by HDI category (see 3B).

B. Created another barchart identical in features to 5a, but with AIP. Examined how many nations showed up both in 5A0's plot and 5B's?

C. Examined relationship between Age and Income. Created a scatterplot and made an assessment of whether there is a relationship. Colored each point based on the Gender of the participant. Used statistical functions to validate.

D. Examined Life Satisfaction and HDI? Create another scatterplot. Is there a discernible relationship there?   What about if you used the HDI category instead and made a barplot?

6. Outputting to CSV format – Make sure there are no index numbers (10%) a The client would like the finalized HDI table (3A and 3B) b The client would like the Tidied version of the original input to be output in the repository, including the merged HDI data (3C). c The client would like a dataset (or two) that shows the Top 15 nations (in 5B and 5C), as well as their HDI scores. d All output should be in plain English or translated in the Codebook.