Pull requests Issues Marketplace Explore

kmendonsa / CaseStudy2

Watch 0 | 0 | 1

Code | Issues 0 | Pull requests 1 | Projects 0 | Wiki | Insights | Settings

CaseStudy2 / README.md

Find file | Copy path

Rmarkas Update README.md d582a9a 25 seconds ago20 seconds ago

162 lines (92 sloc) | 11.8 KB

Raw | Blame | History

CODEBOOK

Introduction to the basis of the data and its analysis

Procrastination - A global view .

This analysis takes a closer look into the behaviors of procrastination globally, and its correlation if any to the Human Development Index (HDI) and the variables that may contribute to these behaviors. Procrastination is an intentional delay of a course of action despite expecting to be worse. The main purpose of this study is to analyze surveys that measure procrastination. The various surveys used for this analysis are based on the below scales and further information is provided later in this codebook.

Decisional Procrastination Scale (Mann, 1982)

Adult Inventory of Procrastination (McCown & Johnson, 1989)

General Procrastination scale (Lay, 1986)

Satisfaction with Life Scale (Diener et al., 1985)

Repo Organization

The Git Hub is logically organized to make review, commits, and accessibility easy for all contributors and clients. The following folders can be found in the root directory:

All these repositories are contained within the parent repo, CaseStudy2

Data - This repo contains all the data related to the project. This data will be described in detail in later sections but the two key files are the procrastination.csv file and the data scraped from wikipedia on the Human Development Index. The files contained in this repo are:

- Copy of Procrastination.xlsx: Copied file of the raw procrastination survey
- DataDictionary.csv: naming convention overrides of all the survey data, character length, data format, and number of missing variables
- HDI_Data.csv: File containing the HDI category and HDI score per country
- HDI_Merged_Data.csv: File containing sampled procrastination and HDI data
- OccupationsMapping.CSV: Naming convention transformation of all the occupation types
- ProcrastinationCLEAN.csv: Using the transformed variables names
- ProcrastinationRAW.csv: Does not use the transformed variable names
- WebScrape.CSV: HDI data scraped from Wikipedia

Paper - The objectives of the case study are stored in this repo.

Source - This repo contains the final Rmarkdown and HTML outputs for the case study.

Procrastination Data Set

File is labeled Procastination.csv and can be found in the Data repo.

Methodology used to clean and structure the data.

- Missing values in fields are converted explicitly to "***Missing***" to indicate the presence of an absence.
- Blanks are also considered as missing values and converted explicitly to "***Missing***".
- Factors were converted to character fields and/or numeric fields as appropriate based on the field.
- Numeric fields where appropriate were rounded to 2 digits unless required for the analysis.
- Some fields were "trimmed" of leading and trailing "whitespace" to enable matches and joins to other fields or data sets.
- Where appropriate, invalid values in fields were considered as '***Missing***' i.e. 999,0 in character fields.
- "NA" is also considered as ***Missing*** value and in character fields have been explicitly changed.
- A file named DataDictionary.csv is provided in the Data folder that provides an in-depth data dictionary of the fields, original names, new names, label length etc.
- A file named CleanData.csv of the "CLEANSED" data is also added to this folder for delivery to the client.

An explanation of each variables is provided below:

- Age: The participant’s age in years. Participants 18 years of age and under and 80 years of age and above have been exlcuded from the analysis and study of the data.

- Gender: The gender the participant identifies as (Male or Female). There are some missing values in this field.

- Kids: Binary, whether they the respondent/participant has kids or not. The field was cleand to remove the word "kids" and retain a simple Yes/No to improve readability and enable better analysis.

- Edu: Education levels are Phd, Masters, Degree, Diploma,Left University, High School, Left High School, Elementary school. There are missing values in this field that have been converted to an explicit ***Missing***. The original values are retained and a new field called EducationAlt (alternate) is crested so the original field can be used for reference ir required for clarification.
- Work Status: The current work status of the participant and work type. This is either Full-time, Part-time, student, Retired, Unemployed. There are missing values in this field that have been converted to an explicit ***Missing***
- Annual Income: This has been converted to dollars in the field IncomeCurr. The regular field was retained as Income to conduct further analysis as a numeric field.
- Current Occupation:What kind of job are they working? The data in this field required considerable "scrubbing" as many respondents appear to have free-typed in their occupations making it quite difficult to do deeper analysis without further cleansing and categorizing the data.
    - The approach that was adopted was to try and "Normalize" and "Standardize" the occupations into meaningful commonly known professions so that they could be grouped/arranged as necessary to provide better insight.
    - Unique professions were left untouched but similar occupations were standardized i.e. technology roles converted to IT so in a datasort they would be grouped together.
    - Self employed individuals were classified as Busines Owner / Self employed.
    - School teachers and other teachers were classified as Teachers while University professors were tagged as Professor.
    - Many values were more of Job Levels than occupational roles i.e.Manager, CEO, Director etc. than the true occupations.
    - A complete listing of the "Renaming" of occupations is visible in the markdown document.
    - Additionally, we retained the Original Occupation field but added a new field called OccupationAlt(alternate) so the original value is available for reference if required.
    - The field was converted to UPPER case to improve readability and consistency to all professions i.e. CRNA, CEO, MD etc.
    - While considerable more work could be done to further standardize and normalize the data, it may be conducted should deeper analysis be required.
- How long have you held this position?: Years: Number of years in this job.
- How long have you held this position?: Months: Number of months in this job.
- Community: Size of community
- Country of Residence: The country where the person holds citizenship. There were missing values in this field.
- Marital Status: Single, Married, Divorced, Separated, etc.
- Number of Sons/Number of daughters: integer number of children. In this field we have explicitly converted the number 2 to "Female" and the number 1 to "Male" as it appears those are the true values that should have been present.
- All variables starting with DP – the Decisional Procrastination Scale (Mann, 1982)
- All variables starting with AIP – Adult Inventory of Procrastination (McCown & Johnson, 1989)
- All variables starting with GP – the General Procrastination scale (Lay, 1986)
- All variables starting with SWLS – the Satisfaction with Life Scale (Diener et al., 1985)
- Do you consider yourself a procrastinator?: a binary response
- Do others consider you a procrastinator?: a binary response

- Computed column: DPMean - mean of the DP survey variable
- Computed column: AIPMean - mean of the AIP survey variable
- Computed column: GPMean - mean of the GP survey variable
- Computed column: SWLSMean - mean of the SWLS survey variable
- There were several survey questionnaire types that are not defined in the data set: DP, AIP, GP, and SWLS. We will explain the concepts of these questions so that it may be properly interpreted. The exact questions can be found within the Procrastination.csv file or the data dictionary that is published in the "data" repo.

GP:General Procrastination Scale

The General Procrastination scale is used for people to describe a wide variety of activites associated with procrastinatoin. It is a 5-point scale: (1 - Extremely uncharacteristic 3 - Netural 5 - Extremeley Characterisitc). An Example of this question type includes, "When it is time to get up in the morning, I most often get right out of bed. "

Source: http://www.sciencedirect.com/science/article/pii/0092656686901273?via%3Dihub

DP: Decisoinal Procrastination Scale

This scale is often regarded as the only reliable measure of indecision. The objective of this method is to measure the the tendency of participants to put off decisions by some other task or activity. An example used in this survey includes, "I don't make decisions unless I really have to."

Source:
https://www.researchgate.net/profile/Joseph_Ferrari3/publication/232562755_Decisional_procrastination_Examining_personality_correlates/links/5556ae4d08ae6943a8734cc2/Decisional-procrastination-Examining-personality-correlates.pdf

AIP: Adult Inventory of Procrastination

The AIP measures the tendency for individuals to postpone tasks under differing circumstances. It is most often used as a measure of procrastination due to fear, procrastination due to a lack of skills, or procrastination to protect ones self-esteem from failure. Answers are submitted on a 5 point scale: (1 = strongly disagree; 5 = strongly agree). Examples of qeustions include, "I am not very good at meeting deadlines."

Source:
https://www.researchgate.net/publication/279532373_Adult_Inventory_of_Procrastination_Scale_AIP_A_comparison_of_models_with_an_Italian_sample

SWLS: Satisfied With Life Scale

SWLS is typically a concise 5-item insturment used to underantd and measure global cognitice judgements pertaining to the satification of a participants time. These survey questions do not weight the imporatnce of any one particular area of an individuals life, but rather allow individuals to factor in all aspects of thier life. An example used in the procrastination survey, "I am satisfied with my life" doesn't ask about a specific aspect, but rather a broad representation.

Source: https://internal.psychology.illinois.edu/~ediener/SWLS.html

WebScrape.CSV

This data was scraped from Wikipedia and contains Human Development Attibutes for 189 countries.
https://en.wikipedia.org/wiki/List_of_countries_by_Human_Development_Index

The variables are described below:

- HDICategory: Categorization of the coutnries into 4 variables of HDI (Very High, High, Medium, Low)

- Rank2016_15: Individual countries rank for the calendar year '15-'16
- RankChange: How many positions the country changed in overall ranking
- Country: Countries involved in the study
- HDI2016_15: Human Development Index (HDI) score for the calendar year '15-'16
- HDIChngYoY: Meaure (%) as to the change in HDI over the last year

Summary

This study focuses on the procrastination behaviors of a wide spectrum of regions and demographic behavior. As the United States is the leading polled country, it can be inferred (assumption we are making pretending that we actually conducted the survey on behalf of the client) that this is a representative view of the US procrastination behavior. Some descriptive statistics of the participants give us a better understanding of the polled demographics: ~57% are female and the remaining 43% are male, mean age is ~37 and the mean income is $58,916. Of all the countries surveyed, Taiwan had the highest GP and AIP mean, and Brunei had the highest DP mean. Top 15 country rankings by AIP and DP can be found in the main analysis. The survey gave us a deep look into procrastination behavior but it was interesting to assess whether broader factors, such as HDI, contribute to these behaviors. After scraping wikipedia and associating HDI scores and categories to country origin of surveyed participants we were able to dig in futher. Our finding is that HDI and socio-economic attributes are a driving factor of procrastination behavior.