CS 579 – DATA SCIENCE – ASSIGNMENT 1 – VIGNESH J MURALIDHARAN

DATA COLLECTION PROPOSAL

1. DATA COLLECTION – PROPOSAL OF DATA COLLECTION DETAILS – SPEED DATING DATASET

I choose to research about the modern trends of speed dating around the United States. The data is gathered from participants in experimental speed dating events of the years 2002-2004. All the attendees would have to talk with at least twenty people of opposite sex and were asked with both the partners each time when they meet so that to get an idea of the population. According to https://academic.oup.com/qje/article/121/2/673/1884033 every participant were asked to rate their date based on six attributes like

- Attractiveness
- Sincerity
- Intelligence
- Fun
- Ambition &
- Shared Interest

The dataset also includes fields like demographics, dating habits, self-perception across key attributes, beliefs on what others find valuable in the partner and lifestyle with preference of yours on the attributes, how do you rate the partner? How does partner give importance on the attributes? and How do we rate yourself on the attributes? At the end of the event the decision on both sides were asked and based on the decision the match or not match has been decided.

1.2 DATA DESCRIPTION/METADATA (DUBLIN CORE STANDARD)

- **Digital Object Identifier (DOI) Number**: 10.1162/qjec.2006.121.2.673
- **Title / DOI Related Article**: Gender Difference in Mate Selection: Evidence from a Speed Dating Experiment
- **Creator**: Raymond Fisman, Sheena S.Iyengar, Emir Kamenica, Itamar Simonson
- **Publisher:** The Quarterly Journal of Economics
- **Date Published:** 1st May 2006 (Article Type)
- **Source:** Dataset <https://openml2.win.tue.nl/d/40536>
- **Language:** English
- **Dimension of the actual dataset:** 8378 (Observations) & 123 (Variables)
- **List of Textual Variables:** Gender, Race, and Field of Study – Total 3
- **List of Continuous variables:** Total 122
- **Target (Y):** Variable Name – Match => Type Binary (0 – Not Match, 1 – Match)

2. DATA MANAGEMENT

2.1 **Creation of logical collections**: I will do this by collecting/arranging the data into an excel file, logical analyze using graphs and visualize different aspects of the attributes when I enter the program using different libraries.

2.2 **Physical data handling**: I will back up the data onto the cloud to ensure that the data is available. This would help with ensuring easy access to the data. I will be using GitHub, and or OneDrive/Google Drive.

2.3 **Interoperability Support**: I will be having this data accessible through the cloud which allows for the data to be downloaded on different platforms. I will be using mostly Google Drive or GitHub.

2.4 **Security Support**: The metadata will be updated every time data is changed with the category being 'Last Updated'. Only I will have permission to authorize others to view or edit the data. The original data set will always be secured and for another person editing the document, a copy of the new data will be stored.

2.5 **Data Ownership**: The creator of the data, I, will be responsible for reviewing any edited data to ensure quality and meaning. The ownership will not be shared along the course of study.

2.6 **Metadata Collection, management and access**: The metadata will be included in the excel file or separate PDF document. It will include the creator name, original dataset collection data, dataset title, description, contact information, sources, last updated date, etc. The metadata will be created with standard terminology for the understanding of viewers.

2.7 **Persistence**: Since the data is fixed for a specific time, it will remain the same and will be accessible through the drive or any other means of backup. If the data is updated in future before the project ends and if the numbers will vary, then the old records/file along with the updated once will be available through the means of backup on the cloud like Google drive/ OneDrive.

2.8 **Discovery**: The link to my data will be accessible for others to view. This can be done by adding it to my GitHub website or also by using SEO on google to allow other to more easily find my data.

2.9 **Data Dissemination and Publications**: All edited versions will be stored for the purpose of comparison and accessibility to all recoded data. This will be done in the excel sheet or by using data frame editing using R/Python and the interested parties will be aware of the changes as recorded in the data and the metadata.

3. SURVEY OF DATA STORAGE / FORMATS

I will be using the information provided by article research and the data to understand which data formats would work well for my collection. But as per my understanding since I am planning to use R or Python for future visualizations and analysis, I will collect this data in excel format (Xlsx or CSV). Some of the recommended data formats could be used for quantitative data to record, collect and publish are .mdb, .txt, .xls/.xlsx and .csv.

The data presentation and publications can be done in .ppt, .txt, .html, .doc/.docx, .pdf and .tex. The data formats that could be used for images and visualization are .jpg/jpeg and .pdf. The data formats that could be used for documentation purposes are .txt, .doc/.docx, .xls/.xlsx, .xml, .rtf, .html and .pdf.

4. SURVEY OF METADATA CONVENTIONS, STANDARDS

As discussed in the class lectures, the metadata will comprise of the title, creator, data, description, keywords/subjects, language, publisher, terms of access to the data and the data format. The metadata will be stored along with the data in the same format as the quantitative data format or in a .doc file with table in sighted. The metadata will provide the viewer with the overview of the data, such as the purpose and timeline. For my data collection the creator is not me and the research proposal has been explained clearly in the data collection and the data description headings. The detailed metadata will be included with the research question also and will be included during the data presentation.

Note: The glimpse of the metadata has been provided in the section 1.2 of this document for the research dataset.