Assignment 1 - Getting familiar with Data types and Visualization

Step 1: **Download** the dataset files belong to the following data formats from internet. The files may belong to any dataset available online.

Step 2: **Read** these files inside the python code. Some of the file formats cannot be read using default python packages. In this case, explore the python packages suitable for reading the files.

Step 3: **Print** the properties of the data files such as size, shape, dimensions, etc.

Step 4: **Visualize** each of these data files using graphs, diagrams, etc.

- Table data visualization: line graph, bar graph, histogram chart, pie chart, scatter plot
- Image visualization: image plot, 3d plot
- Video visualization: video player
- Audio visualization: audio player, spectrogram
- Text visualization: Word cloud, bubble cloud (some more in http://vallandingham.me/textvis-talk/)

1. Tabular, Spreadsheet and Interchange Data Formats

- "Table" — generic tabular data (.dat), "CSV" — comma-separated values (.csv), "TSV" — tab-separated values (.tsv), “ARFF” - Attribute-Relation File Format (.arff) – Read and visualize the data
- "XLS" — Excel spreadsheet (.xls), "XLSX" — Excel 2007 format (.xlsx), "ODS" — OpenDocument spreadsheet (.ods), "SXC" — OpenOffice 1.0 spreadsheet file (.sxc), "DIF" — VisiCalc data interchange format (.dif) – Read and visualize the data
- "JSON" — JavaScript Object Notation (.json), "UBJSON" — Universal Binary JSON (.ubj), “HTML” – Hypertext Markup Language (.html), “XML” - eXtensible Markup Language (.xml) - Read and Parse the data

2. Data File Formats

- PKL – Pickle format, HDF5, Zip, SQL, MAT, NPY, NPZ – Read and display the data

3. Image Data Formats

- JPG, PNG, BMP, TIFF – Read and display the image
- 3D medical Images: DICOM, MHA – Read and display the image

4. Video Data Formats

- MP4, AVI, MPEG – Read and play the video

5. Audio Data Formats

- MP3, MIDI, WAV – Read and play the audio

6. Text Data Formats

- TXT, PDF, DOC – Read and parse the data

Suggested Platform: Python: Azure Notebook/Google Colab Notebook, packages such as Numpy, Pandas, Sklearn

Marking: Marking is based on both **performance during the lab hours** as well as **complete submission**.