Team Progress Report 1 – Bladder Cancer Detection Team
04/06/19

- Data is obtained from the NIH, imported and ready for wrangling. Due to the manageable size of the data it is currently stored on our computers until we chose a suitable database.

- Doing background research and learning about the data as well as the variables of interest

- Researching possible statistical analysis we will likely use such as K-mean Clustering, Hierarchical Clustering

- Having meetings with team members and discussing the structure of our project
    - Prepared the design of our project as well as a flowchart. Decided on writing a report rather than building an app. Brainstorming about what type of analysis to conduct.
    - All our documents are uploaded to our GitHub depository.
    - The architecture of our project is presented to our TA. Went over the feedback from the TA with the team.

- Wrangling with Pandas. Transposed the data, relabeled columns, indexed uniquely.

- Data exploration, identify potential correlations among features, PCA and hierarchical clustering, any sub-groups among data, any patterns recognizable,
  any pattern associated by age/gender/locations, etc. (on going)

Member Progress Report 1

Guozhen (subject expert):

- Came up with the data.
- Data storage and wrangling. Leading data clean and structuring.
- Briefing the team on the data as well the variables as the team has questions
- Taking online courses and tutorials to learn python and necessary tools for the project.

Leyan:

- Read some research papers about bladder cancer, Bladder Cancer miRNA marker.pdf located in our repository
- Review our project dataset and get an understanding of the data variables.
- Earned intermediate Python on DataCamp and get ready for project coding
    - Intermediate Python
- Built a unit test module template for our project, BladderCancerMar13b.ipynb, located in our repository.
- Researching possible analysis methods
    - Created a comparison table to display how K-mean clustering and hierarchical clustering works and the difference between these two methods so that our team can better understand the two methods.
- Preparing Python code to express the different statistic models (on going)
- Preparing questions/concerns to be discussed with the team:
    - Need to figure out a better way to handle “N/A” values and “uncertain” values for the same variable?
    - Since we will use K mean clustering model, how do we define K?

Fang:

- Working on domain knowledge and bladder cancer scientific background.
    - Reading and understanding bladder cancer related literature as parts of references of our project.

- Working on data cleaning and wrangling

- Grouping samples into different categories to compare the mRNA value and find the significant mRNA candidates.

- Working on the program that processes the above two bullets.

- Figured out how to clone, push, pull, and modify project related files to/from our group's Github repository.

Sertan:

- Completed recommended tutorials on SQL,
    - https://sqlbolt.com
    - http://www.sqlitetutorial.net

- Completed recommended tutorials on git and Github,
    - https://www.datacamp.com/courses/introduction-to-git-for-data-science
    - Another one provided by one the instructor but could not locate the link.

- Taking online courses to learn Python to do the required data ingestion, wrangling, and statistical analysis for our project. Completed 24 hours of online Python courses from datacamp.com,
    - Introduction to Python
    - Intermediate Python
    - Python Tool Box 1
    - Python Tool Box 2
    - Cleaning Data in Python
    - pandas Foundation

- Learning about our data and the meaning of the variables. Since our data is a sample of micoRNAs with close to a thousand indicators for each sample from variety of cancer patients, it requires some reading on the subject to better understand the type of analysis to be conducted.

- Working on how to structure the data. It needs to be transformed and possibly some variables need to be combined.