ROSEBELLA CAPIO

STAT 517

FINAL PROJECT DRAFT

25TH OCTOBER 2018

**Data Description**

The data is a subset of the collection of the RNA-seq (HiSeq) PANCAN data set. It is a random extraction of gene expressions of patients having different types of tumors, specifically:

* BRCA
* KIRC
* COAD
* LUAD
* PRAD.

There is a total of 801 observations and 16384 features with no missing values. The variables of each tumor type are RNA-Seq gene expression levels measured by illumine HiSeq platform

**Source**

Samuele Fiorini, samuele.fiorini **'@'** dibris.unige.it, University of Genoa, redistributed under Creative Commons license (<http://creativecommons.org/licenses/by/3.0/legalcode>) from <https://www.synapse.org/>#!Synapse:syn4301332.

**Research Goals**

* Perform predictive analysis on the data by
* Perform clustering analysis on the data
* Perform data classification
* Check for associations