**Data Analysis on Indian Liver Patient Records**

**Name: Venkata Anil Kumar Thota**

**Sjsu Id: 012525734**

**Overview**:

**Discuss your interests and why you pursued the selected dataset:**

I have been interested in the Healthcare domain from a very long time. Health Care is a revolutionary and promising industry in which data science solutions are implemented. Data analysis moves medical science to a completely new level, from computerization of medical records to drug discovery and exploration of genetic diseases. And it's only the beginning.

In various sectors of the health care industry, data science is progressing. I will be quoting some of them here, the Collection of enormous amounts of data to monitor and prevent health problems, improve diagnostic accuracy and efficiency, reduce hospital readmissions and reduce costs for healthcare.

Despite huge amounts of health data, the diagnostic failure rates are still high. According to a research by the National Academics of Sciences, Medicine, about 5 percent of adult patients are misdiagnosed each year in the US. This totals over 12 million people. Moreover, the postmortem examination results research shows that diagnostic errors cause approximately 10 percent of patient deaths.These research and statistics increased my interest in this Domain and made me pick Dataset related to Diagnosis.

Diagnosis is determination of the nature of a cause of a disease. It is a concise technical description of the cause , nature, of a condition or a situation.

**Why is the subject matter important to you?**

In the process of setting goals Godsey quoted an example for the confusion of correlation and causation. So, looking at subject in detail is essential part in setting goals, as it will lead to some fallacious conclusions. In addition to that in United States, liver cancer ranks as the fifth most common cause of cancer death in Men and the ninth in Women. As per recent stats 71% of liver diseases were in Men. I want to explore the cause of this dangerous disease and find out person’s risk of getting Lung Cancer. To get answers to my questions I should know about the Subject matter in specific.

**What are you hoping to achieve? (Your initial purpose)**

Ultimately, I want to detect relationship between the Variables and make a Data Model so that it will reduce the Burden on the Doctors. As stated by Godsey in his book “Think Like a Data Scientist”, to achieve my goal first I need to stay close to the data, which mean I should Explore the data, so that the Methods which I am applying will not be More complex. Therefore, my initial purpose is to study about the Data.

**Dataset Studied:**

Dataset Which I selected for the analysis is “Indian Liver Patient Dataset”. The Dataset is available in Machine Learning repository in the below address. It contains Patient records collected from North East of Andhra Pradesh, India. These Patient Record are used to which patients have liver disease and which does not have, to reduce burden on the doctors.

Web Link : <https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)>.

**Description of & Structure of the Data:**

This section involves Layout, Metadata analysis of the dataset.

This Dataset contains 583 records and 11 Columns. All the 11 Columns are listed below.

* Age
* Gender
* Total Bilirubin
* Direct Bilirubin
* Alkaline Phosphotase
* Alamine Aminotransferase
* Aspartate Aminotransferase
* Total Protiens
* Albumin
* Albumin and Globulin Ratio
* Dataset.

Any patient whose age exceeded 89 is listed as being of age "90". Last Column i.e.., Dataset is used to split the data into two sets (Patient with liver disease and Patient without liver disease).

Among 11 Columns, 5 columns are of float Datatype, 5 columns are of integer datatype and one column(age) is an object. All the columns in the Dataset are explained below.

1)Age- Age of the patient

2)Gender- Gender of the patient. Contains values either Female/Male.

3) Total\_Bilirubin

4) Direct\_Bilirubin

Column 3 and Column 4 are explained below.

A yellow pigment which is present in everyone’s Blood and Stool is termed as Bilirubin. Bilirubin attached by the liver to glucuronic acid which is a glucose-derived acid is called as Direct Bilirubin or Conjugated Bilirubin. Bilirubin no attached to glucuronic acid is termed as indirect Bilirubin or unconjugated bilirubin. All the Bilirubin in the Blood together is called Total Bilirubin.

5) Alkaline\_Phosphotase – It is an enzyme found in the Blood stream. Abnormal Levels of this Phosphotase results in liver Problems. It varies person to person based on age, blood type etc. The normal range for Alkaline\_Phosphotase is 20 to 140.

6) Alamine\_Aminotransferase – This Variable measures the amount of this enzyme in Blood. Generally low levels of Alamine\_Aminotransferase are found in Blood if there are high levels then it is more likely to cause liver Damage.

7) Aspartate\_Aminotransferase – This is a blood test that checks for liver Damage. The Normal values vary based on Gender, for Males it ranges from 10 to 40 units/L on the other hand for Females it ranges from 9 to 32 units/L.

8) Total\_Protiens - Albumin and globulin are two types of protein in the body. The total protein test measures the total amount albumin and globulin in the body. The total protein is between 6 and 8.3 grams per deciliter. This value varies based on age, gender etc.

9) Albumin – This test measures amount of Albumin protein in the Blood. Albumin test can help to evaluate liver condition. A normal albumin ranges from 3.4 to 5.4 g/dl.

10) Albumin\_and\_Globulin\_Ratio – This ratio is used as an index of disease state. The normal A/G ratio 0.8-2.0. The A/G ratio can be decreased in response to a low albumin or to elevated globulins.

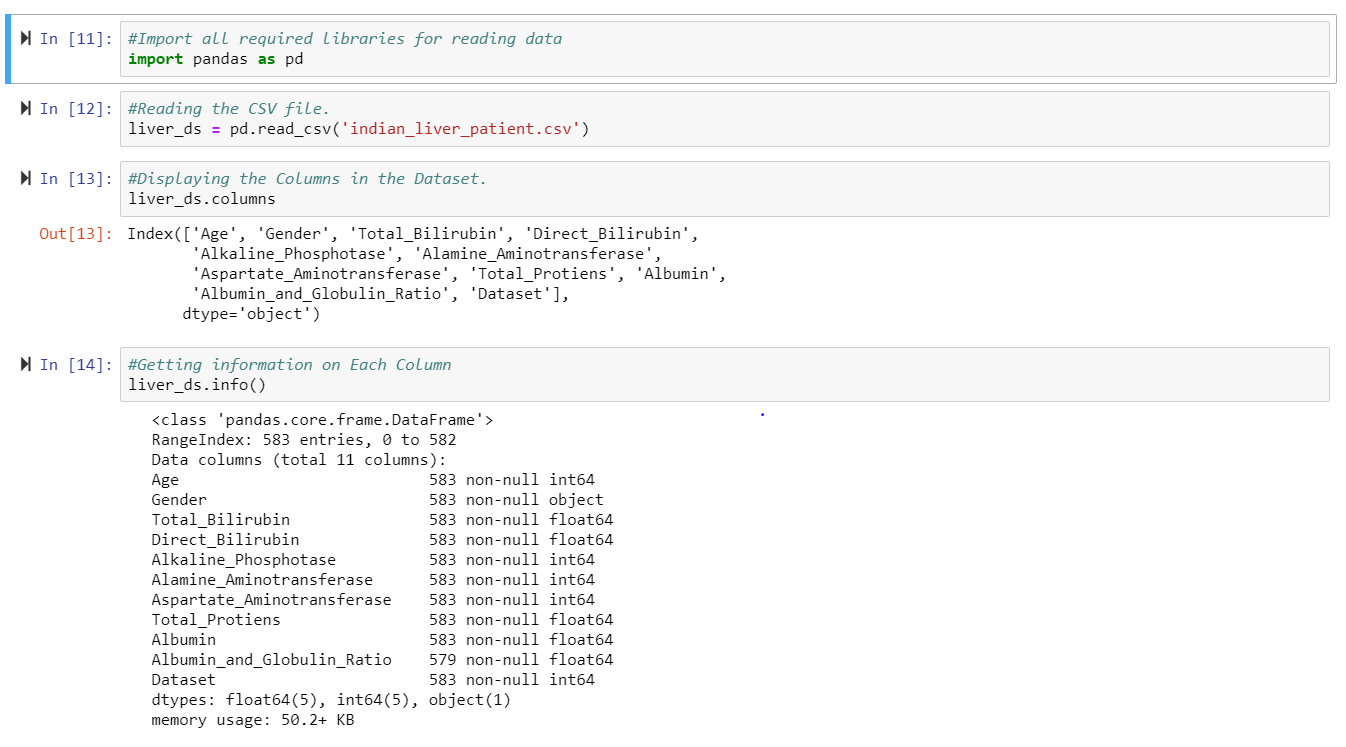
11) Dataset – The final Column(Dataset) is output variable. It contains two values either 1 or 2. Value 1 indicates that the patient has liver disease and 2 indicates the patient does not have liver disease.

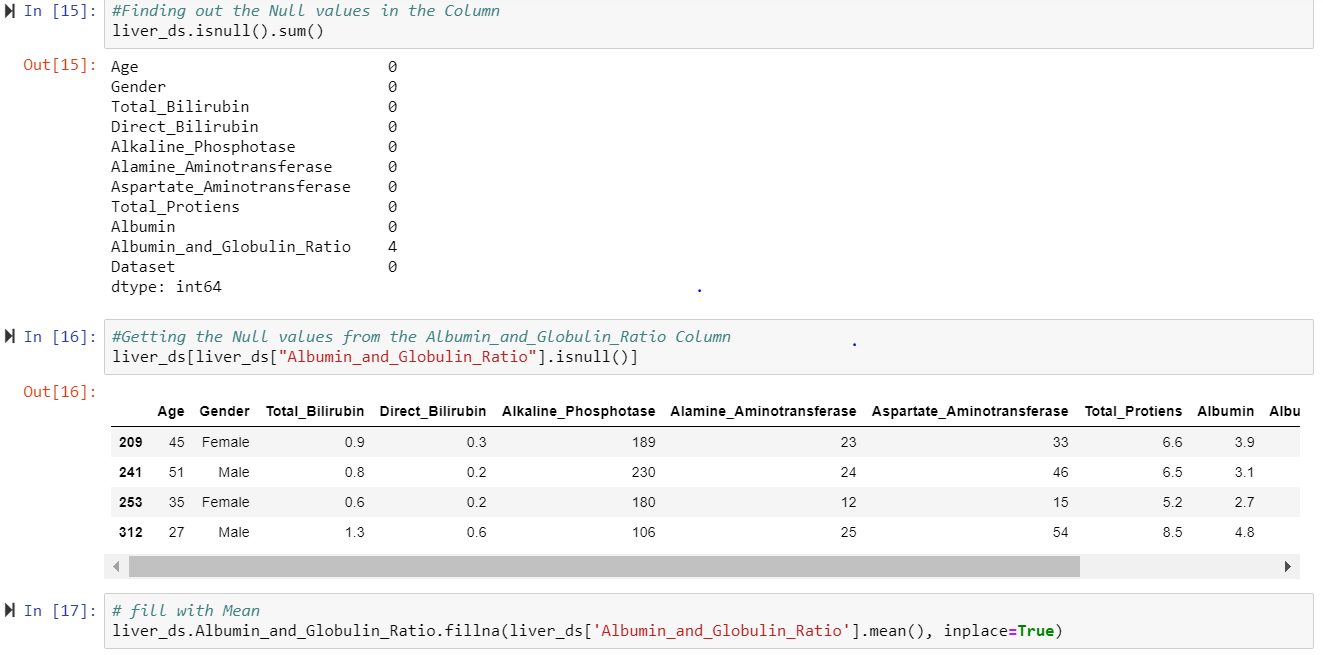
**Notes on Exploration of the Dataset:**

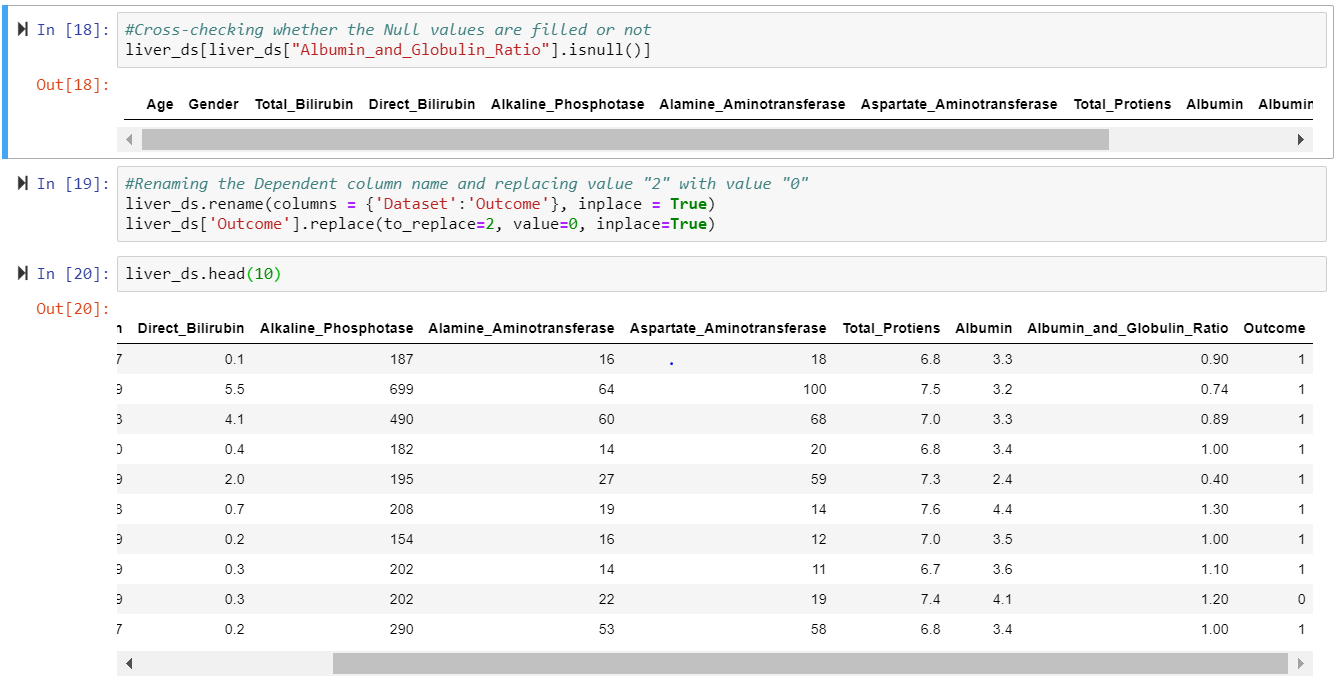
After going through the Godsey’s Textbook, I got a clear picture of the Steps to follow in the preparation Phase. After setting the Goals, we should Explore the data. Here comes the concept of Data Wrangling. Data Wrangling is an important part of any Data Analysis. We should make sure that the data is in perfect Shape and ready for consumption before we apply any algorithms on it. As Cautioned by Author in Chapter 3 I decided to spend extra time on wrangling so that it will save a lot of pain later.

Dataset I selected is in CSV (Comma- Separated Values) format. Columns Names are given in the description, but the actual CSV file doesn’t have the Column names. So, I manually added columns to File. Then I used Jypter Notebook to perform data wrangling on the Dataset.

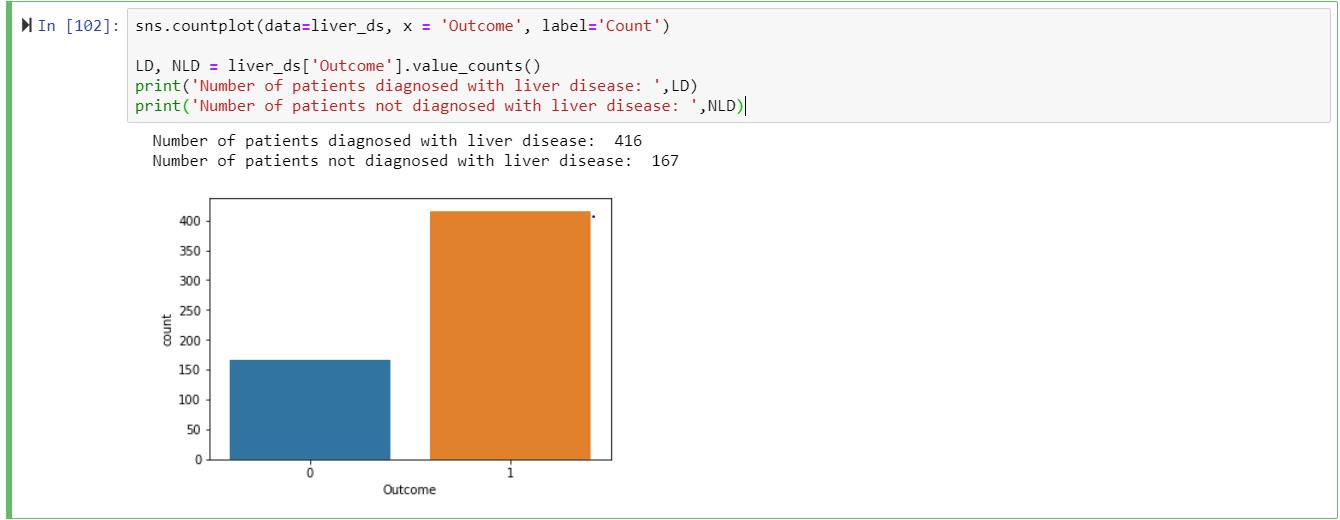
First, I imported the pandas Library and used it to read the CSV file. I used available commands in python to get the Column names and its information like Datatypes. The prominent step here is Cleaning Null Values. By using the isnull() function in python I found that “Albumin\_and\_Globulin\_Ratio” column has 4 Null values, which are marked as “NaN”. Out of 583 records available only 4 rows have null values. So, I replaced those Null values with mean of the column. This replacement will not have much effect on the results. The dependent variable in the dataset is named as “Dataset”, I changed it to “Outcome” so that it makes some sense. The values of the dependent variable are given in terms of “1” and “2”, value “1” indicates patient has liver disease and “2” indicates the opposite side. To be consistent changed “2” with “0” in the Outcome Column. Here are some of the snaps taken from the notebook. I will also attach the Notebook with the report.







As described by Godsey in his text book to know better about the data I started to calculate descriptive statistics on the dataset. Got the summary of the dataset and the maximum, minimum and mean values of every column. The mean Age is about 44, the oldest person in the data is 90, while the youngest person is of 4 Years. There are more males than females in the whole dataset. For better understanding I used other available libraries in python and drew few graphs. I got to know few interesting things when I started Assessing the Data. Out of 583 patient records 416 are found to have liver disease, the rest 167 are not diagnosed with liver disease.

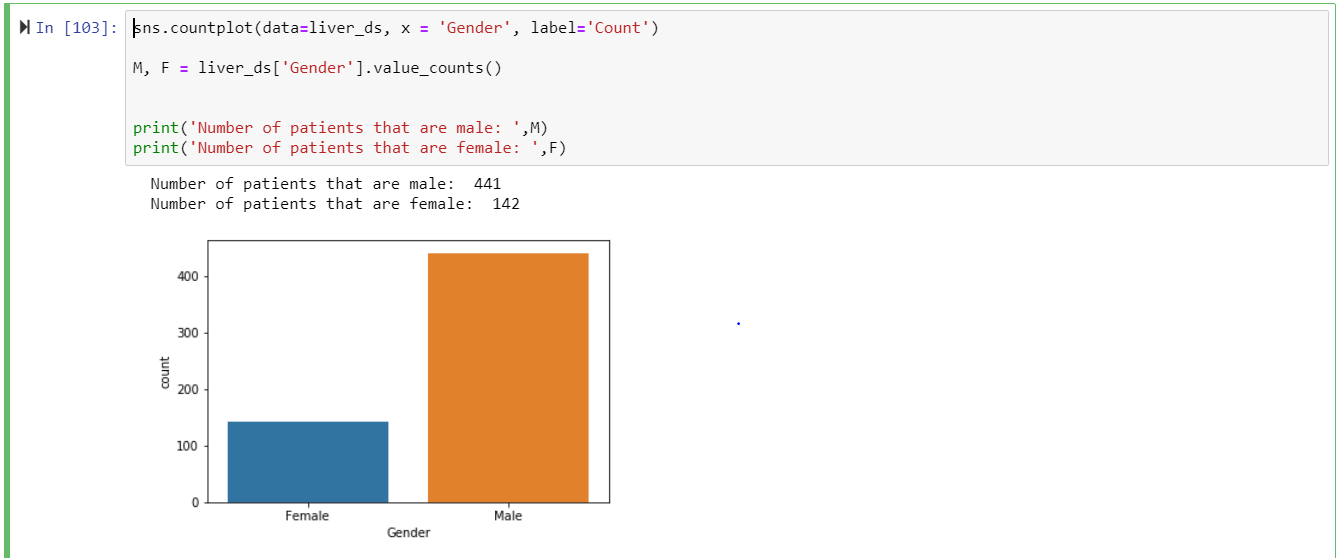


**Questions to Ask of the Dataset:**

Based on the Exploration that I did on the dataset. I started to ask few questions on the dataset.

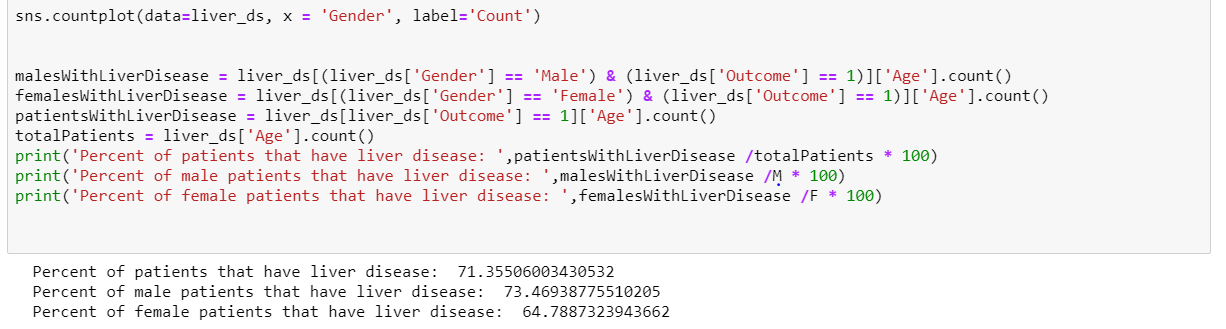
1)Out of 583 Patient records considered for the analysis how many are Male and how many are Female?

I plotted a histogram, and used value\_counts() to get answer for this question. Out of 583 records, 441 are Male patients and 142 are Female patients.



2)What is the percentage of Male and Female that have liver disease.

Out of all the patient records collected 71% patients have liver disease. Coming to the Male category percentage of patients who are with liver disease are around 73%, on the other hand this percentage is 64 for Female category.



3)How Gender and Age affect the Outcome (Whether the person is diagnosed with Liver disease or Not)

4)Is there any relationship between Total\_Bilirubin, Direct\_Bilirubin and Gender.

5)Is there any relationship between Aspartate\_Aminotransferase, Alamine\_Aminotransferase and Gender.

6)Is there any relationship between Albumin\_and\_Globulin\_Ratio and Albumin and Gender.

7) What are the correlation coefficients for all dataset. Can we Exclude any of the features?

8) Are there any other additional columns required to diagnose the liver disease?

I will attach Jypter notebook where I tried to get answers for remaining questions. After completing this step, I got to know how the attributes are closely related. Based on these relationships I got a clear picture about the data and Subject Matter. In case of considering other factors I, am interested to ask customer about the weight of the patients, will the weight of the patient have any impact on the liver?

**Summary**:

After completing the preparation phase of the project, concluded that, to build a better model for the dataset first we should spend more time on the preparation phase. Also, I came to know how data analysis is used in Healthcare domain specifically in Diagnosis and what are the skills required to crack a Job as Data Scientist in the Health Care Field. Going through Godsey “Think Like a Data Scientist”, it gave me a path how the data science process arrives to a practical solution to the real-life data centric problems. Starting a data-centric project as soon as the wrangling phase completes is a bad design, even though it takes some time we should spend time on knowing about data. It also increased my Python programming skills combined with Stats.

Being a Data Scientist, one should know statistics and software. Data scientists need to gather, organize and analyze data helping people from every part of the industry. Data scientists work on system where one component which is involved in the analysis is not defined properly. Communication is one of the most important skill needed to become a data scientist. We should ask questions on data and learn things from another domain expertise. Being a data scientist involves some contradictory skill sets: intelligence to handle data processing and create useful models; and an intuitive understanding of the business problem, the structure and nuances of the data, and how the models work.

Data science is an interdisciplinary field that contains methods, process and involves extraction of knowledge from the data. Critical thinking is the fundamental skill I need to develop. So that I can apply objective analysis of facts on the given topic or problem. To be successful data scientist one should have good programming skills, I should increase my programming skills ad become master in one of the programming language like Python or Java. Last but not the least Mathematics and Statistics are the principal skills required to be successful in this field. I am very much excited to build the model for the dataset. If I get time I will start developing a data model for this dataset. Logistic regression, Random Forest method will be the appropriate statistical methods for this data. Overall doing this Mini-project gave me a better understanding how actually data analysis is performed on a real time data.

**References**:

1. Godsey, B. (2017). Think like a data scientist: Tackle the data science process step-by-step. Shelter Island: Manning.
2. A Comprehensive Introduction to Data Wrangling. (2018, September 12). Retrieved November 8, 2018, from <https://www.springboard.com/blog/data-wrangling/>
3. The Stages of Liver Disease. (n.d.). Retrieved from <https://liverfoundation.org/for-patients/about-the-liver/the-progression-of-liver-disease/#1503432878616-a25d5b59-3a75>
4. Liver Failure. (n.d.). Retrieved November 06, 2018, from <https://www.webmd.com/digestive-disorders/digestive-diseases-liver-failure#1>
5. Python, R. (n.d.). Python Data Science – Real Python. Retrieved from <https://realpython.com/tutorials/data-science/>
6. DataSets. (n.d.). Retrieved from <https://archive.ics.uci.edu/ml/datasets.html>