# Statistical Method for Data analysis

Maryam Mohmmad Alrashdi

443800993

# Table of content

# Table of Figures

# Table of Tables

# INTRODUCTION

Body performance means the condition of being physically and mentally fit with good health. It is the ability to carry out daily tasks with vigor and alertness, without undue fatigue, and with ample energy to enjoy life.

Regular exercise and physical activity promote strong muscles and bones. It improves respiratory, cardiovascular health, and overall health. Staying active can also help you maintain a healthy weight, reduce your risk for type 2 diabetes, heart disease, and reduce your risk for some cancers.

For the body to perform well, it is necessary to know the good exercises for it.

So here we chose a database that confirmed the grade of performance with age and some exercise performance data.

## Dataset

Source: link (Korea Sports Promotion Foundation)

Some post-processing and filtering have done from the raw data

So, it is data collected by the Korea Sports Promotion Foundation. It consists of measurement data for each physical fitness measurement item.

This database was used as csv file from Kaggle

## Dataset content

| | |
|---|---|
| Age | The ages of people whose data was collected |
| Gender | The gender of people whose data was collected. |
| height_cm | The heights of people whose data was collected. |
| weight_kg | The weights of people whose data was collected. |
| body fat_% | The body fats of the people whose data was collected |
| Diastolic | Diastolic blood pressure (min) |
| systolic | Systolic blood pressure (min) |
| gripForce | Train the grip |
| sit and bend forward_cm | Type of exercise |
| sit-ups counts | Type of exercise |
| broad jump_cm | Type of exercise |

| | |
|---|---|
| Class | A,B,C,D ( A: best) / stratified |

Table 1 Dataset content

# Part1:

## Data munging

Data munging is considered as an important process in which data gets cleaned, extracted, and identified. It is also a way of integrating a perfect dataset together

### Explore the data

Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, to better understand the nature of the data.

Data exploration techniques include both manual analysis and automated data exploration software solutions that visually explore and identify relationships between different data variables, the structure of the dataset, the presence of outliers, and the distribution of data values to reveal patterns and points of interest, enabling data analysts to gain greater insight into the raw data.

In this task various functions were used to explore the data.

Overview of the functions used in this part.

- **str():** to Know the structure of the dataset.

```
> str(bf)
'data.frame':   13393 obs. of  12 variables:
 $ age                    : num  27 25 31 32 28 36 42 33 54 28 ...
 $ gender                 : chr  "M" "M" "M" "M" ...
 $ height_cm              : num  172 165 180 174 174 ...
 $ weight_kg              : num  75.2 55.8 78 71.1 67.7 ...
 $ body.fat_.             : num  21.3 15.7 20.1 18.4 17.1 22 32.2 36.9 27.6 14.4 ...
 $ diastolic              : num  80 77 92 76 70 64 72 84 85 81 ...
 $ systolic               : num  130 126 152 147 127 119 135 137 165 156 ...
 $ gripForce              : num  54.9 36.4 44.8 41.4 43.5 23.8 22.7 45.9 40.4 57.9 ...
 $ sit.and.bend.forward_cm: num  18.4 16.3 12 15.2 27.1 21 0.8 12.3 18.6 12.1 ...
 $ sit.ups.counts         : num  60 53 49 53 45 27 18 42 34 55 ...
 $ broad.jump_cm          : num  217 229 181 219 217 153 146 234 148 213 ...
 $ class                  : chr  "C" "A" "C" "B" ...
```

As we can see here in the database structure, only two of the 12 columns are char (gender and class), and the rest are numeric.

- **View ():** View the data set.
- **Head():** See the first 6 rows.
- **Tail():** See the last 6 rows.
- **Names():** Lists name of variables in a dataset.

## Descriptive statistics

Descriptive statistics are brief descriptive coefficients that summarize a given data set.

Overview of the functions used in this part.

- **Summary():** to find the summary of the dataset.

```
      age            gender            height_cm       weight_kg
 Min.   :21.00   Length:13393       Min.   :125.0   Min.   : 26.30
 1st Qu.:25.00   Class :character   1st Qu.:162.4   1st Qu.: 58.20
 Median :32.00   Mode  :character   Median :169.2   Median : 67.40
 Mean   :36.78                      Mean   :168.6   Mean   : 67.45
 3rd Qu.:48.00                      3rd Qu.:174.8   3rd Qu.: 75.30
 Max.   :64.00                      Max.   :193.8   Max.   :138.10
   body.fat_.      diastolic        systolic       gripForce
 Min.   : 3.00   Min.   :  0.0   Min.   :  0.0   Min.   : 0.00
 1st Qu.:18.00   1st Qu.: 71.0   1st Qu.:120.0   1st Qu.:27.50
 Median :22.80   Median : 79.0   Median :130.0   Median :37.90
 Mean   :23.24   Mean   : 78.8   Mean   :130.2   Mean   :36.96
 3rd Qu.:28.00   3rd Qu.: 86.0   3rd Qu.:141.0   3rd Qu.:45.20
 Max.   :78.40   Max.   :156.2   Max.   :201.0   Max.   :70.50
 sit.and.bend.forward_cm sit.ups.counts  broad.jump_cm      class
 Min.   :-25.00          Min.   : 0.00   Min.   :  0.0   Length:13393
 1st Qu.: 10.90          1st Qu.:30.00   1st Qu.:162.0   Class :character
 Median : 16.20          Median :41.00   Median :193.0   Mode  :character
 Mean   : 15.21          Mean   :39.77   Mean   :190.1
 3rd Qu.: 20.70          3rd Qu.:50.00   3rd Qu.:221.0
 Max.   :213.00          Max.   :80.00   Max.   :303.0
```

We can see here that it is a summary of the database that shows us the statistical values for each column, for example, we can say that the average age of the people whose data were collected in this database is 36.78, and we can say the minimum length of people who do these exercises to know the performance of their bodies It is 125.0, and we can know that the highest value of the gripForce exercise is 70.50, so as we said, this summary shows a lot of statistical information that will help us in analyzing the data.

- **Mean():**

for the age column: 36.77511

The average age of the people who performed these exercises and whose data were collected in this database is 36.77511

for the broad.jump_cm column: 190.1296

The average result for this broad.jump cm workout is 190.1296

for the body.fat_. column: 23.24016

The average body fat of the subjects whose exercise data was collected is 23.24016

❖ **Median():**

for the age column: 32

The median of the age of the people who performed these exercises and whose data were collected in this database is 32

for the broad.jump_cm column: 193

The median result of broad.jump cm workout is 193

for the body.fat_. column: 22.8

The median of body fat of the subjects whose exercise data was collected is 22.8

❖ **Max():**

for the age column: 64

The oldest age of those who performed these exercises and whose data were collected in this database is 64.

for the broad.jump_cm column: 303

The highest score of broad.jump cm workout is 303

for the body.fat_. column: 78.4

The highest body fat percentage for people whose exercise data was collected was 78.4

❖ **Min()**

for the age column: 21

The youngest age of those who performed these exercises and whose data were collected in this database is 21.

for the broad.jump_cm column: 0

The lowest score of broad.jump cm workout is 0

for the body.fat_. column: 3

The lowest body fat percentage for people whose exercise data was collected was 3

- **max()-min() = range**

for the age column: 43

The age range of those who performed these exercises and whose data were collected in this database is 43.

for the broad.jump_cm column: 303

The score range of broad.jump cm workout is 303

for the body.fat_. column: 75.4

The range body fat percentage for people whose exercise data was collected was 75.4

## Data Visualization

Overview of the types used in this part.

- **Pie chart**

for the class column

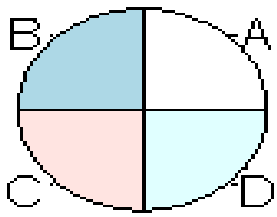


Figure 1 pie chart for class

for the gender column

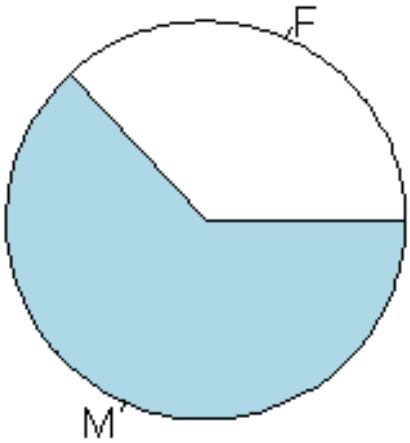


Figure 2 pie chart for gender

- Barplot

for the age column

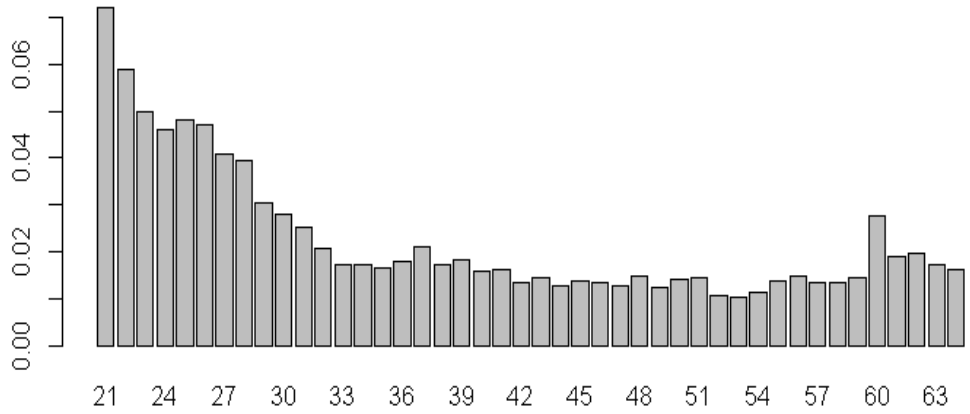


Figure 3 Barplot for age

for the gender column

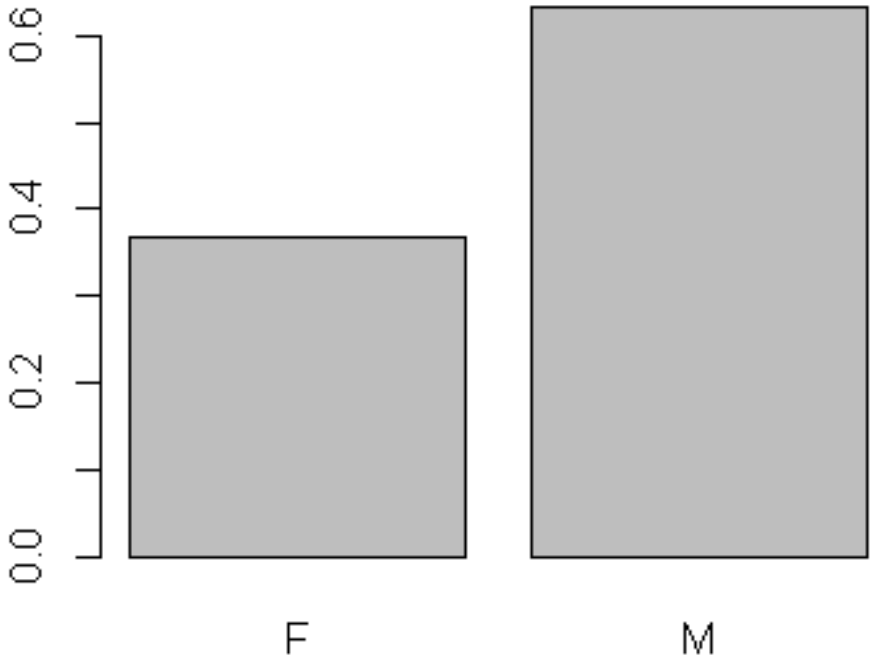


Figure 4 Barplot for gender

❖ **Plot**

for the height and weight variables

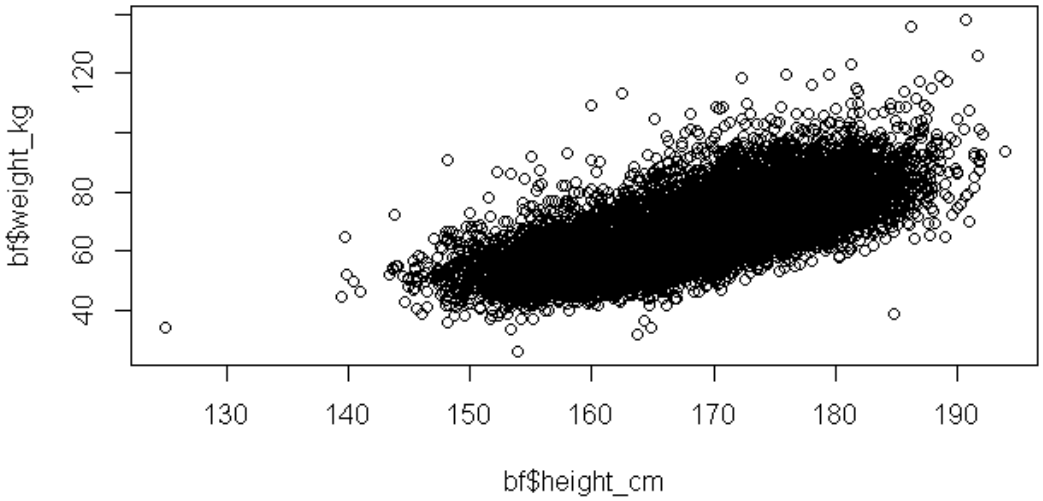


Figure 5 Plot for height and weight

# Linear Regression

At first, the database will be divided into train and test, then variables will be selected to be used to predict a variable.

## Linear Regression with two variables

- I will use sit.ups.counts  to predict broad.jump_cm
- Sum of Squared Errors: 9367715
- R-squared for training set: 0.5599122
- R-squared for testing set: 0.6456779

Predicting by using the Linear Regression with these two columns is good, so there is a relationship between them, i.e., the more one of them increases, the more the other will increase, or vice versa.

## Linear regression with many variables

- I will use weight_kg , height_cm , sit.ups.counts , age and systolic to predict body.fat_.
- Sum of Squared Errors: 276160.3
- R-squared for training set: 0.6084202
- R-squared for testing set: 0.5596265

Prediction by using linear regression with several variables was Not so bad, meaning that these variables influence each other.

# Conclusions

From the previous data analysis, we can draw some conclusions

- The greater the height, the greater the weight
- The age range of the people whose training and characteristics were collected to test their physical performance ranged from 21 to 64 years
- This database shows that the proportion of males is higher than the proportion of females
- It also shows that people in their twenties and thirties train more than people of other ages
- The weight range of the subjects whose training and characteristics were collected to test their body performance is between 26-138
- The height range of the subjects whose training and characteristics were collected to test their body performance ranged from 125-193
- Also, the lowest percentage for the sit and bend forward_cm exercise can be negative, as it appeared to be -25

# Part 2

We will use the Diabetes file to find the following:

## The average of BMI

The average BMI for diabetic patients is 31.99258

## The maximum and minimum value for BloodPressure

The maximum value for blood pressure for people with diabetes is 122

The minimum value for blood pressure for people with diabetes is 0

## The age in the bottom 25% of Glucose and the top 25% of the Glucose

Ages of people with glucose bottom 25%

```
  [1] 50 32 53 34 59 51 41 51 43 57 28 26 54 40 25 29 58 42 41 44 24 27 21 40
 [25] 33 22 37 24 46 39 61 25 33 30 23 65 42 23 43 36 47 36 22 36 47 40 41 60
 [49] 36 29 29 57 52 41 24 60 38 21 66 61 24 31 22 26 51 23 32 29 49 23 24 51
 [73] 34 27 63 28 28 47 34 65 28 29 30 25 37 47 27 29 59 43 24 37 41 26 33 41
 [97] 49 29 29 63 67 30 25 32 58 27 31 25 41 39 28 22 21 22 37 36 31 38 29 28
[121] 41 25 38 51 27 28 35 31 58 67 66 55 39 35 28 53 37 53 50 37 28 37 23 62
[145] 21 41 52 22 38 54 36 22 27 36 40 45 50 30 33 22 25 54 22 43 40 70 45 49
[169] 31 53 26 46 44 43 31 52 45 24 34 31 42 22 22 24 39 27 36 50 26 45 66 43
```

Ages of people with glucose top 25%

```
  [1] 31 21 26 22 22 28 27 22 30 21 21 36 25 22 22 41 27 22 36 21 42 24 22 23
 [25] 24 27 27 25 24 23 25 25 22 26 39 26 22 41 22 22 25 23 33 42 32 21 27 42
 [49] 21 25 32 22 22 23 21 22 24 25 44 24 43 21 26 21 35 41 25 22 36 22 57 37
 [73] 29 46 24 21 48 22 21 22 22 35 25 39 25 31 38 22 27 21 25 43 38 22 36 21
 [97] 22 26 23 25 24 26 39 37 22 25 28 23 24 42 22 25 28 41 41 40 38 21 46 58
[121] 22 41 25 21 21 29 31 24 67 56 25 29 25 23 28 30 35 24 27 22 46 39 22 22
[145] 25 25 46 21 28 25 23 28 21 29 21 31 23 32 31 23 27 43 47 68 25 22 22 25
[169] 28 23 56 52 34 22 28 42 32 22 33 23
```

The distribution of the variables

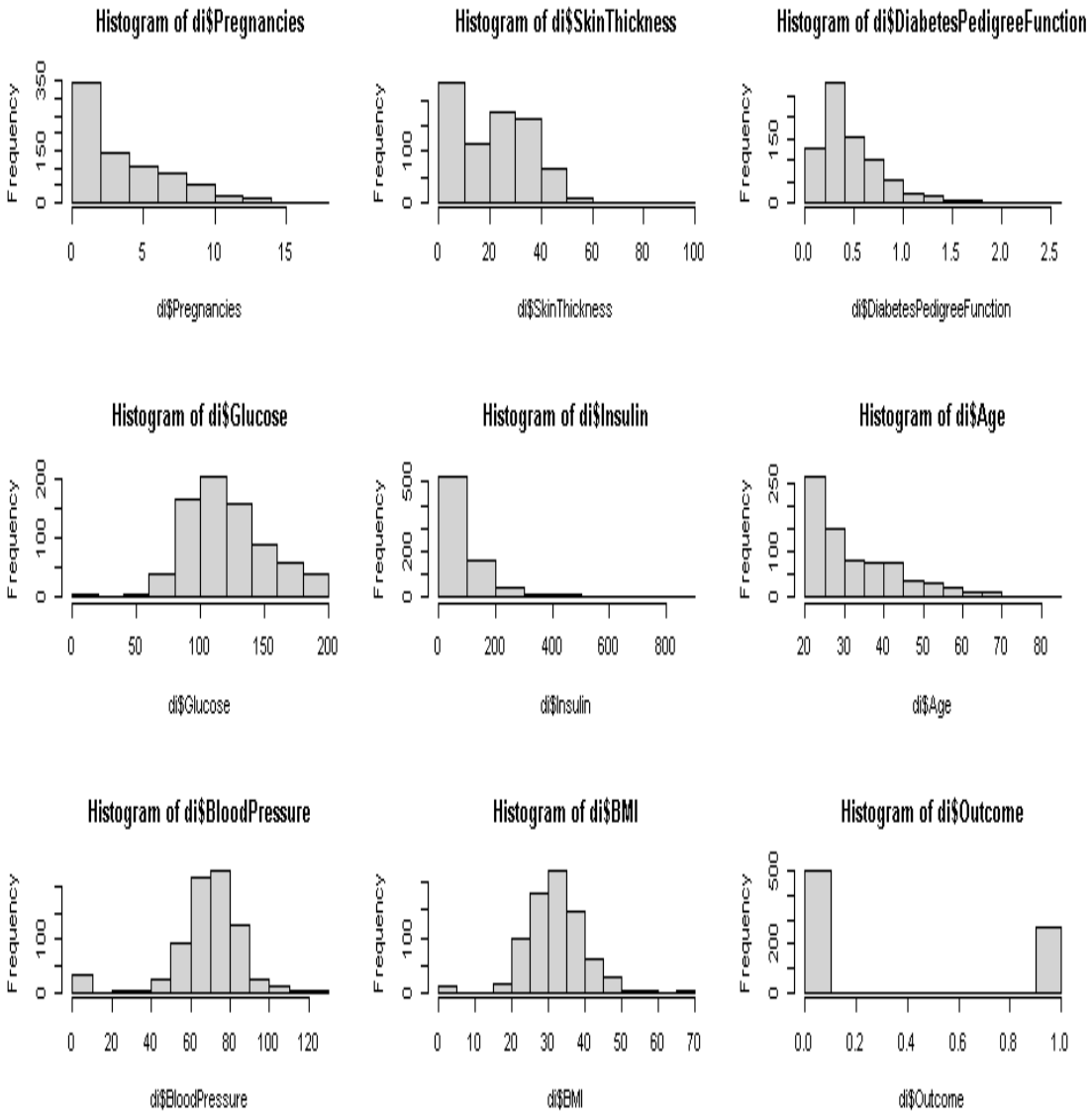


Figure 6 Distribution of the variables

- Distribution of the Pregnancies

A right-skewed distribution (positively skewed distribution.)

many data values occur on the left side with a fewer number of data values on the right side.

- Distribution of the Glucose

A normal distribution: points on one side of the average are as likely to occur as on the other side of the average.

- Distribution of the BloodPressure

A normal distribution with some outlier.

- Distribution of the SkinThickness

A random distribution:  lacks an apparent pattern.

- Distribution of the Insulin

A right-skewed distribution (positively skewed distribution.)

many data values occur on the left side with a fewer number of data values on the right side.

- Distribution of the BMI

A normal distribution: points on one side of the average are as likely to occur as on the other side of the average.

- Distribution of the DiabetesPedigreeFunction

A right-skewed distribution (positively skewed distribution.)

many data values occur on the left side with a fewer number of data values on the right side.

- Distribution of the age

A right-skewed distribution (positively skewed distribution.)

many data values occur on the left side with a fewer number of data values on the right side.

- Distribution of the Outcome

A bimodal distribution: there are two peaks.

The correlation between variables

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI |
|---|---|---|---|---|---|---|
| Pregnancies | 1.00000000 | 0.12945867 | 0.14128198 | -0.08167177 | -0.07353461 | 0.01768309 |
| Glucose | 0.12945867 | 1.00000000 | 0.15258959 | 0.05732789 | 0.33135711 | 0.22107107 |
| BloodPressure | 0.14128198 | 0.15258959 | 1.00000000 | 0.20737054 | 0.08893338 | 0.28180529 |
| SkinThickness | -0.08167177 | 0.05732789 | 0.20737054 | 1.00000000 | 0.43678257 | 0.39257320 |
| Insulin | -0.07353461 | 0.33135711 | 0.08893338 | 0.43678257 | 1.00000000 | 0.19785906 |
| BMI | 0.01768309 | 0.22107107 | 0.28180529 | 0.39257320 | 0.19785906 | 1.00000000 |
| DiabetesPedigreeFunction | -0.03352267 | 0.13733730 | 0.04126495 | 0.18392757 | 0.18507093 | 0.14064695 |
| Age | 0.54434123 | 0.26351432 | 0.23952795 | -0.11397026 | -0.04216295 | 0.03624187 |
| Outcome | 0.22189815 | 0.46658140 | 0.06506836 | 0.07475223 | 0.13054795 | 0.29269466 |

| | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|
| Pregnancies | -0.03352267 | 0.54434123 | 0.22189815 |
| Glucose | 0.13733730 | 0.26351432 | 0.46658140 |
| BloodPressure | 0.04126495 | 0.23952795 | 0.06506836 |
| SkinThickness | 0.18392757 | -0.11397026 | 0.07475223 |
| Insulin | 0.18507093 | -0.04216295 | 0.13054795 |
| BMI | 0.14064695 | 0.03624187 | 0.29269466 |
| DiabetesPedigreeFunction | 1.00000000 | 0.03356131 | 0.17384407 |
| Age | 0.03356131 | 1.00000000 | 0.23835598 |
| Outcome | 0.17384407 | 0.23835598 | 1.00000000 |

In correlation between variables there is a general rule (if the correlation is 1, there is a strong positive correlation, but if it is -1, there is a strong negative correlation)

Without touching on the relationship of the variables themselves, which of course would be 1

In this database, the correlation between the variables is very weak, as the highest correlation is 0.5, which is a ratio between (age and Pregnancies), which is not a good ratio.

# REFERENCES


- The National Academies of Science Engineering Medicine: //www.nap.edu/read/11203/chapter/4
- Jigsaw academy: www.jigsawacademy.com/blogs/data-science/data-munging/
- Massively Accelerated Analytics: www.omnisci.com/learn/data-exploration