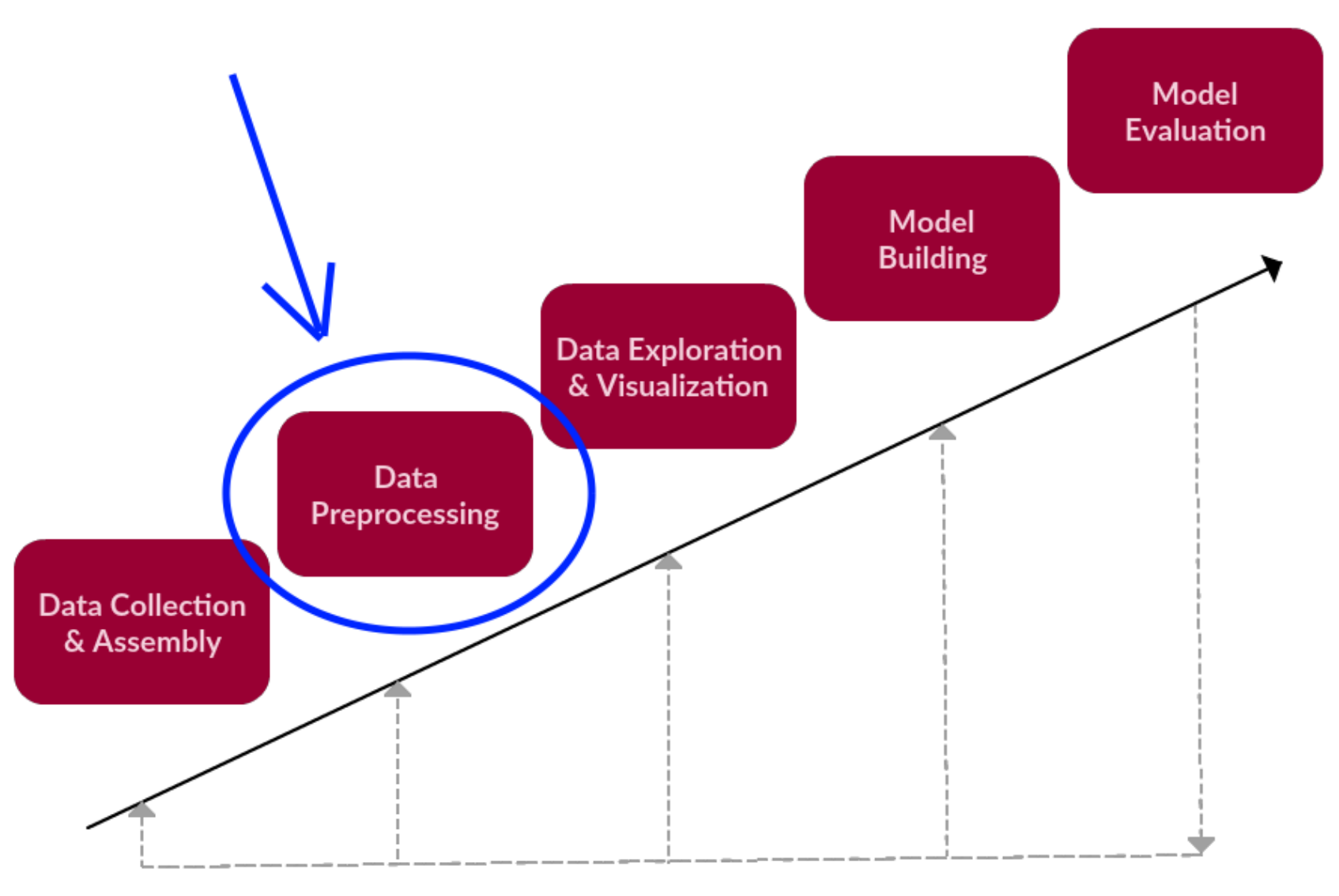
Data Preprocessing - Data Analysis

Data Preprocessing

It is used for maintaining the **Quality** of the data. It includes important factors like -

- Selecting the valid data variables
- Data editing is important in some aspects
- Maintaining uniformity in data values
- Manipulation of the data for achieving the above factors (Data Wrangling)



Credits - Image from Internet

In [ ]:

Classroom Management

Imagine you are the new teacher freshly appointed to manage the classroom.

You want to know how many students are good in -

```
1. sports
2. academics
3. creative work
4. marketing
```

Based on the students data, you have to conclude **who** can do well in **what**.

Dataset description

You are given the **data of the students** that included the following variables -

- Age
- Gender
- Address
- Father's occupation
- Mother's occupation
- Place of birth
- Height - ft
- Weight - kg
- Prev sports performance
- Prev academics performance
- Voluntary experience
- Extra co-curricular activities
- Arts and Design

Note - For our conveneince all the data values are numericals.

In [ ]:

How do we convert a **" feeling "** into a number?

- We can measure a **" feeling "** into a number through a **" scale range "**
- If the scale is **1** to **4**, then we can term -
  - 1 → Not Satisfied
  - 2 → Slightly Satisfied
  - 3 → Satisfied
  - 4 → Highly Satisfied

In [ ]:

Sports

Scenario 1

Considering the variables that are direclty related -

- Height
- Weight
- Prev sports performance

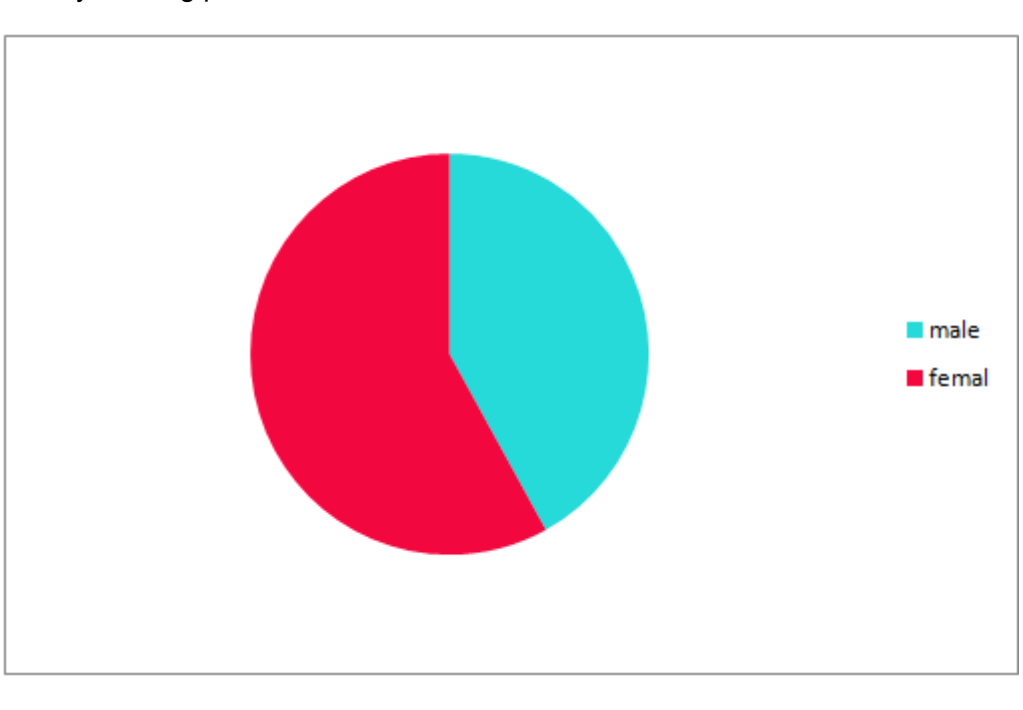
Based on this, you can only get the information of a student irrespective of **gender**.

Scenario 2

Considering the other important factors like **gender** in order to categorize as per **Male related sports** and **Female related sports**.

- Gender
- Height
- Weight
- Prev sports performance

1. Based on this, you can categorize the performance of students in sports by **Male** and **Female**.
2. Visually, you can represent it by drawing pie chart.



Credits - Image from Internet

Scenario 3

If you want to do further research on how good the person is performing in other areas, you can do so by considering -

- Gender
- Height
- Weight
- Prev sports performance
- Voluntary experience
- Extra co-curricular activities
- Prev academics performace (may be or may not be)

1. With this, you conclude the overall students performance on sports
2. Since you are a kind teacher and well wisher of student, you can give the student a proper career guidance.

In [ ]:

Academics

Scenario 1

Considering the variables that are directly related -

- Prev academics peroformance

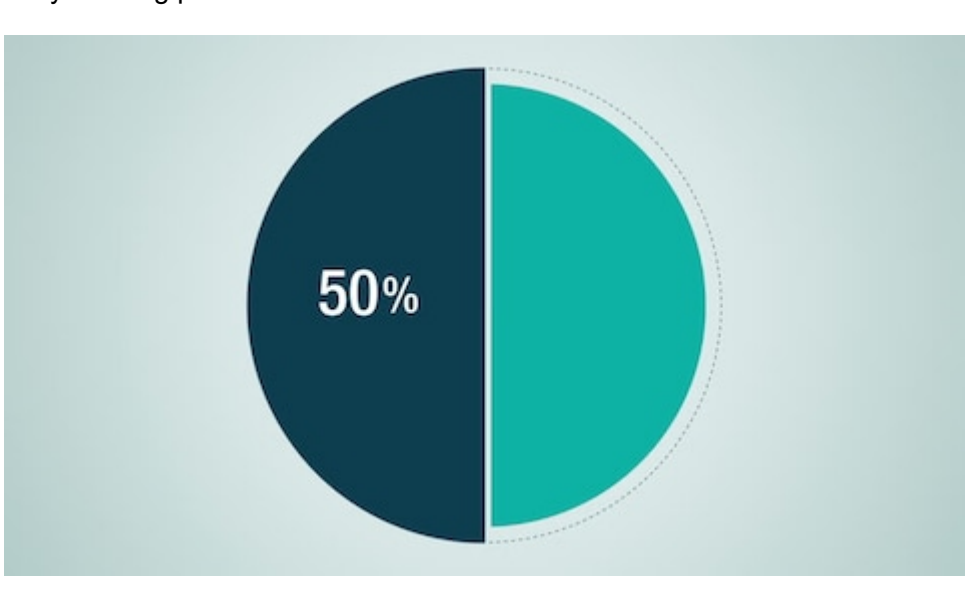
Based on this, you can only get the information of student irrespective of **gender**.

Scenario 2

Considering the other important factors like **gender** to categorize **Male** and **Female** separately.

- Gender
- Prev academics performance

1. Based on this, you can categorize the performance of students in academics by **Male** and **Female**.
2. Visually, you can represent it by drawing pie chart.



Credits - Image from Internet

Scenario 3

If you want to further research on why a particular student is **lagging behind** or **excelling ahead**, you can do so by considering -

- Address
- Father's occupation
- Mother's occupation
- Prev academics performance
- Gender (for categorizing in terms of gender)

and later on, you can decide whether to change your teaching methodology or not.

In [ ]:

Note -

- Data Analyst should be wise enough to select the important variables.
- This helps to get proper insights pertaining to the problem statement that he/she is assigned to do.

In [ ]:

Let's make our hands dirty

data source → http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights

In [1]:
```
import pandas as pd
```

In [2]:
```
df = pd.read_csv('students_hw.csv')
df.head()
```

Out[2]:

	Height(Inches)	Weight(Pounds)
0	65.78	112.99
1	71.52	136.49
2	69.40	153.03
3	68.22	142.34
4	67.79	144.30

Check the length of the `df`

In [3]:
```
df.shape
```

Out[3]: (200, 2)

Data Preprocessing

Check for `NaN`

In [4]:
```
def check_for_nan(dframe):
    """
    dframe → pandas data frame object
    returns `nan_places` a dictionary of column names and the `nan_indices`
    """
    nan_frame = dframe.isna()
    d_cols = dframe.columns

    nan_places = {}
    for col in d_cols:
        col_lvals = nan_frame[col].to_list()
        nan_indices = [index for index, val in enumerate(col_lvals) if val == True]
        if nan_indices:
            nan_places[col] = nan_indices
        else:
            nan_places[col] = None

    return nan_places
```

In [5]:
```
check_for_nan(dframe=df)
```

Out[5]: {'Height(Inches)': [10, 32], 'Weight(Pounds)': [19]}

- In the column `Height(Inches)`, there are two `NaN` values at indices `10` and `32`.
- In the column `Weight(Pounds)`, there is one `NaN` value at index `19`.

What can we do for those?

- Remove the entire row which ever column has a `NaN`.

For this, we will remove the rows which ever column has `NaN`. In total, there are 3 rows that need to be removed.

Remove 3 rows

In [6]:
```
pdf = df.dropna(axis=0)
```

check the length of `pdf`

In [7]:
```
pdf.shape
```

Out[7]: (197, 2)

In [8]:
```
# pdf.head(31)
```

Since the index of the data frame is not in order, we need to reindex the index values to get the perfect order.

Reindex the index

In [9]:
```
new_index = list(range(pdf.shape[0]))
rdf = pdf.reindex(new_index)
```

In [10]:
```
# rdf.head(32)
```

In [11]:
```
rdf.head()
```

Out[11]:

	Height(Inches)	Weight(Pounds)
0	65.78	112.99
1	71.52	136.49
2	69.40	153.03
3	68.22	142.34
4	67.79	144.30

Check if `Height(Inches)` < 40

In [12]:
```
inch_thresh = 40
```

In [13]:
```
rdf[rdf['Height(Inches)'] < inch_thresh]
```

Out[13]:

	Height(Inches)	Weight(Pounds)
71	30.84	134.02
96	36.29	120.03

Remove the rows where `Height(Inches)` < 40

- In the above case, we can see two values where height is less than 40.
- We remove by specifying the index values in `drop()` method.

In [14]:
```
rdf.drop(index=[71, 96], inplace=True)
```

In [15]:
```
rdf.shape
```

Out[15]: (195, 2)

Reindex the index

In [16]:
```
new_index = list(range(rdf.shape[0]))
hdf = rdf.reindex(new_index)
```

Since the index of the data frame is not in order, we need to reindex the index values to get the perfect order.

Categorize the data

In [17]:
```
avg_height = hdf['Height(Inches)'].mean()
```

In [18]:
```
avg_height
```

Out[18]: 67.957

In [19]:
```
height_cat = []
for i in hdf['Height(Inches)'].to_list():
    if i < avg_height:
        height_cat.append('short')
    elif i > avg_height:
        height_cat.append('tall')
    else:
        height_cat.append('average')
```

In [20]:
```
len(height_cat)
```

Out[20]: 195

In [21]:
```
hdf['height_category'] = height_cat
```

In [22]:
```
hdf.head()
```

Out[22]:

	Height(Inches)	Weight(Pounds)	height_category
0	65.78	112.99	short
1	71.52	136.49	tall
2	69.40	153.03	tall
3	68.22	142.34	tall
4	67.79	144.30	short

Plotting the pie chart to show

- how many are short
- how many are tall

In [23]:
```
pie_df = hdf.groupby(by='height_category').sum()
```

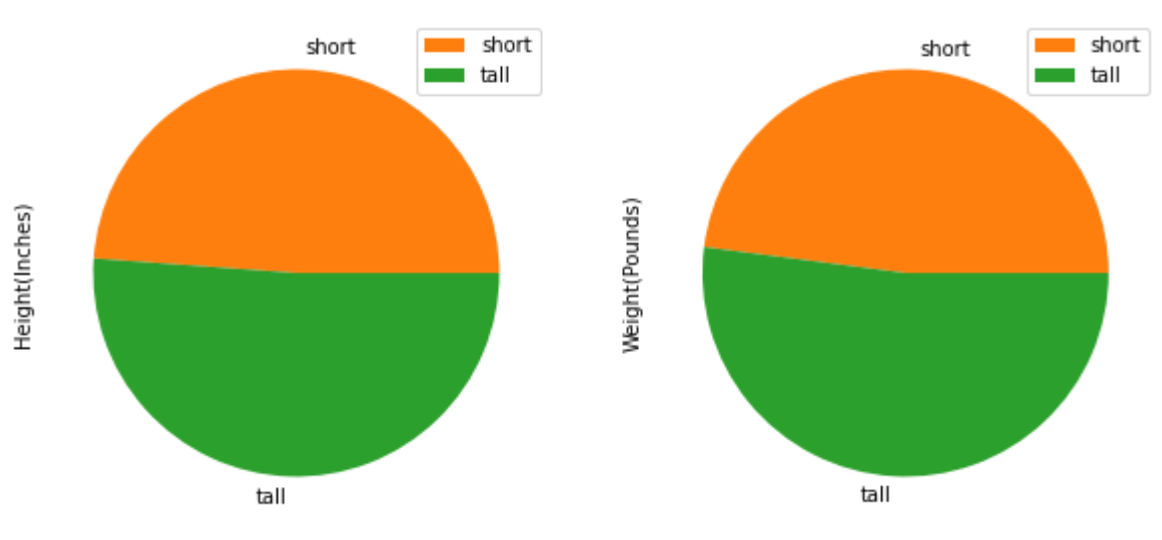
In [24]:
```
pie_df
```

Out[24]:

	Height(Inches)	Weight(Pounds)
height_category		
average	0.00	0.00
short	6308.72	11604.05
tall	6603.11	12588.74

In [25]:
```
pie_df.plot(kind='pie', subplots=True, figsize=(10, 15))
```

Out[25]:
```
array([<AxesSubplot:ylabel='Height(Inches)'>,
       <AxesSubplot:ylabel='Weight(Pounds)'>], dtype=object)
```



In [ ]:

What did we learn?

- Data Preprocessing
- Real life scenario Example
- Classroom Management problem
  - sports
  - academics
  - creative work
  - marketing
- Getting hands dirty by writing code
- Plotting the pie chart for showing categorization