# 115_impute

December 26, 2018

## 1 Replace (impute) missing numerical data with median of column values

When we import data into NumPy or Pandas, any empty cells of numerical data will be labelled np.NaN on import. In techniques such as machine learning we may wish to either 1) remove rows with any missing data, or 2) fill in the missing data with a set value, often the median of all other values in that data column. The latter has an advantage that the technique can be used both in training the machine learning model, and in predicting output when we are given examples with some missing data.

Here we define a function that goes through data columns in a Pandas DataFrame, looks to see if there is any missing data and, of there is, replaces np.NaN with the median of all other values in that data column.

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

        def impute_with_median (df):
            """Iterate through columns of Pandas DataFrame.
            Where NaNs exist replace with median"""

            # Get list of DataFrame column names
            cols = list(df)
            # Loop through columns
            for column in cols:
                # Transfer column to independent series
                col_data = df[column]
                # Look to see if there is any missing numerical data
                missing_data = sum(col_data.isna())
                if missing_data > 0:
                    # Get median and replace missing numerical data with median
                    col_median = col_data.median()
                    col_data.fillna(col_median, inplace=True)
                    df[column] = col_data
            return df
```

We will mimic importing data with missing numerical data.

```
In [2]: name = ['Bob', 'Jim', 'Anne', 'Rosie', 'Ben', 'Tom']
        colour = ['red', 'red', 'red', 'blue', 'red', 'blue']
        age = [23, 45, np.NaN, 21, 18, 20]
        height = [1.80, np.NaN, 1.65, 1.71, 1.61, 1.76]

        data =pd.DataFrame()
        data['name'] = name
        data['colour'] = colour
        data['age'] = age
        data['height'] = height
```

View the data with missing values.

```
In [3]: data
```

```
Out[3]:     name colour   age  height
        0    Bob    red  23.0    1.80
        1    Jim    red  45.0     NaN
        2   Anne    red   NaN    1.65
        3  Rosie   blue  21.0    1.71
        4    Ben    red  18.0    1.61
        5    Tom   blue  20.0    1.76
```

Call the function to replace missing data with the median, and re-examine data.

```
In [4]: data = impute_with_median(data)

In [5]: data
```

```
Out[5]:     name colour   age  height
        0    Bob    red  23.0    1.80
        1    Jim    red  45.0    1.71
        2   Anne    red  21.0    1.65
        3  Rosie   blue  21.0    1.71
        4    Ben    red  18.0    1.61
        5    Tom   blue  20.0    1.76
```