Data Mining

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib as mp
        %matplotlib inline
        from sklearn.model_selection import train_test_split
        from sklearn.linear
```

```
Out[2]: ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
         'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome']
```

```
In [3]: data.head()
```

```
Out[3]:   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
        0            6      148             72             35        0  33.6
        1            1       85             66             29        0  26.6
        2            8      183             64              0        0  23.3
        3            1       89             66             23       94  28.1
        4            0      137             40             35      168  43.1

           DiabetesPedigreeFunction  Age  Outcome
        0                     0.627   50        1
        1                     0.351   31        0
        2                     0.672   32        1
        3                     0.167   21        0
        4                     2.288   33        1
```

```
In [4]: data.columns
```

```
Out[4]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness',
               'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
               dtype='object')
```

```
In [5]: data.shape
```

```
Out[5]: (768, 9)
```

```
In [6]: data.info()
```

```
Out[6]: <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 768 entries, 0 to 767
        Data columns (total 9 columns):
        Pregnancies                 768 non-null int64
        Glucose                     768 non-null int64
        BloodPressure               768 non-null int64
        SkinThickness               768 non-null int64
        Insulin                     768 non-null int64
        BMI                         768 non-null float64
        DiabetesPedigreeFunction    768 non-null float64
        Age                         768 non-null int64
        Outcome                     768 non-null int64
        dtypes: float64(2), int64(7)
        memory usage: 54.0 KB
```

```
In [7]: data.describe()
```

```
Out[7]:       Pregnancies     Glucose  BloodPressure  SkinThickness     Insulin  \
        count   768.000000  768.000000     768.000000     768.000000  768.000000
        mean      3.845052  120.894531      69.105469      20.536458   79.799479
        std       3.369578   31.972618      19.355807      15.952218  115.244002
        min       0.000000    0.000000       0.000000       0.000000    0.000000
        25%       1.000000   99.000000      62.000000       0.000000    0.000000
        50%       3.000000  117.000000      72.000000      23.000000   30.500000
        75%       6.000000  140.250000      80.000000      32.000000  127.250000
        max      17.000000  199.000000     122.000000      99.000000  846.000000

                  BMI  DiabetesPedigreeFunction         Age     Outcome
        count  768.000000                768.000000  768.000000  768.000000
        mean    31.992578                  0.471876   33.240885    0.348958
        std     7.884160                   0.331329   11.760232    0.476951
        min     0.000000                   0.078000    21.000000    0.000000
        25%    27.300000                   0.243750    24.000000    0.000000
        50%    32.000000                   0.372500    29.000000    0.000000
        75%    36.600000                   0.626250    41.000000    1.000000
        max    67.100000                   2.420000    81.000000    1.000000
```

```
In [8]: data.tail()
```

```
Out[8]:   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
        0            6      148             72             35        0  33.6
        1            1       85             66             29        0  26.6
        2            8      183             64              0        0  23.3
        3            1       89             66             23       94  28.1
        4            0      137             40             35      168  43.1

           DiabetesPedigreeFunction  Age  Outcome
        0                     0.627   50        1
        1                     0.351   31        0
        2                     0.672   32        1
        3                     0.167   21        0
        4                     2.288   33        1
```

```
In [9]: data.info()
```

```
Out[9]: <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 768 entries, 0 to 767
        Data columns (total 9 columns):
        Pregnancies                 768 non-null int64
        Glucose                     768 non-null int64
        BloodPressure               768 non-null int64
        SkinThickness               768 non-null int64
        Insulin                     768 non-null int64
        BMI                         768 non-null float64
        DiabetesPedigreeFunction    768 non-null float64
        Age                         768 non-null int64
        Outcome                     768 non-null int64
        dtypes: float64(2), int64(7)
        memory usage: 54.0 KB
```

```
In [10]: data.columns
```

```
Out[10]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness',
                'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
                dtype='object')
```

```
In [11]: data['Insulin'].value_counts()
```

```
In [12]: data['Glucose'].value_counts()
```

```
In [13]: data['BloodPressure'].value_counts()
```

```
In [14]: data['SkinThickness'].value_counts()
```

```
In [15]: data['DiabetesPedigreeFunction'].value_counts()
```

```
In [16]: data['Age'].value_counts()
```

```
In [17]: data.dtypes
```

```
Out[17]: Pregnancies                 int64
         Glucose                     int64
         BloodPressure               int64
         SkinThickness               int64
         Insulin                     int64
         BMI                         int64
         DiabetesPedigreeFunction    int64
         Age                         int64
         Outcome                     int64
         dtype: int64
```

```
In [18]: data.hist(figsize=(7,7))
```

```
In [19]: data.head()
```

```
Out[19]:   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
         0            6      148             72             35        0  33.6
         1            1       85             66             29        0  26.6
         2            8      183             64              0        0  23.3
         3            1       89             66             23       94  28.1
         4            0      137             40             35      168  43.1

            DiabetesPedigreeFunction  Age  Outcome
         0                     0.627   50        1
         1                     0.351   31        0
         2                     0.672   32        1
         3                     0.167   21        0
         4                     2.288   33        1
```

```
In [20]: data.dtypes
```

```
Out[20]: Pregnancies                 int64
         Glucose                     int64
         BloodPressure               int64
         SkinThickness               int64
         Insulin                     int64
         BMI                         int64
         DiabetesPedigreeFunction    int64
         Age                         int64
         Outcome                     int64
         dtype: int64
```

```
In [21]: data.hist(figsize=(7,7))
```

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

```
Out[22]:   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
         0            6      148             72             35        0  33.6
         1            1       85             66             29        0  26.6
         2            8      183             64              0        0  23.3
         3            1       89             66             23       94  28.1
         4            0      137             40             35      168  43.1

            DiabetesPedigreeFunction  Age  Outcome
         0                     0.627   50        1
         1                     0.351   31        0
         2                     0.672   32        1
         3                     0.167   21        0
         4                     2.288   33        1
```

```
In [23]: data.shape
```

```
Out[23]: (768, 9)
```

```
In [24]: data.info()
```

```
Out[24]: <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 768 entries, 0 to 767
         Data columns (total 9 columns):
         Pregnancies                 768 non-null int64
         Glucose                     768 non-null int64
         BloodPressure               768 non-null int64
         SkinThickness               768 non-null int64
         Insulin                     768 non-null int64
         BMI                         768 non-null float64
         DiabetesPedigreeFunction    768 non-null float64
         Age                         768 non-null int64
         Outcome                     768 non-null int64
         dtypes: float64(2), int64(7)
         memory usage: 54.0 KB
```

```
In [25]: data.describe()
```

```
Out[25]:       Pregnancies     Glucose  BloodPressure  SkinThickness     Insulin  \
         count   768.000000  768.000000     768.000000     768.000000  768.000000
         mean      3.845052  120.894531      69.105469      20.536458   79.799479
         std       3.369578   31.972618      19.355807      15.952218  115.244002
         min       0.000000    0.000000       0.000000       0.000000    0.000000
         25%       1.000000   99.000000      62.000000       0.000000    0.000000
         50%       3.000000  117.000000      72.000000      23.000000   30.500000
         75%       6.000000  140.250000      80.000000      32.000000  127.250000
         max      17.000000  199.000000     122.000000      99.000000  846.000000

                   BMI  DiabetesPedigreeFunction         Age     Outcome
         count  768.000000                768.000000  768.000000  768.000000
         mean    31.992578                  0.471876   33.240885    0.348958
         std     7.884160                   0.331329   11.760232    0.476951
         min     0.000000                   0.078000    21.000000    0.000000
         25%    27.300000                   0.243750    24.000000    0.000000
         50%    32.000000                   0.372500    29.000000    0.000000
         75%    36.600000                   0.626250    41.000000    1.000000
         max    67.100000                   2.420000    81.000000    1.000000
```

```
In [26]: data.columns
```

```
Out[26]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness',
                'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
                dtype='object')
```

```
In [27]: data['Pregnancies'].value_counts()
```

```
In [28]: data['Glucose'].value_counts()
```

```
In [29]: data['BloodPressure'].value_counts()
```

```
In [30]: data['SkinThickness'].value_counts()
```

```
In [31]: data['Insulin'].value_counts()
```

```
In [32]: data['BMI'].value_counts()
```

```
In [33]: data['DiabetesPedigreeFunction'].value_counts()
```

```
In [34]: data['Age'].value_counts()
```

```
In [35]: data.dtypes
```

```
Out[35]: Pregnancies                 int64
         Glucose                     int64
         BloodPressure               int64
         SkinThickness               int64
         Insulin                     int64
         BMI                         int64
         DiabetesPedigreeFunction    int64
         Age                         int64
         Outcome                     int64
         dtype: int64
```

```
In [36]: data.hist(figsize=(7,7))
```

```
In [37]: data.head()
```

```
Out[37]:   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
         0            6      148             72             35        0  33.6
         1            1       85             66             29        0  26.6
         2            8      183             64              0        0  23.3
         3            1       89             66             23       94  28.1
         4            0      137             40             35      168  43.1

            DiabetesPedigreeFunction  Age  Outcome
         0                     0.627   50        1
         1                     0.351   31        0
         2                     0.672   32        1
         3                     0.167   21        0
         4                     2.288   33        1
```

```
In [38]: data.shape
```

```
Out[38]: (768, 9)
```

```
In [39]: data.info()
```

```
Out[39]: <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 768 entries, 0 to 767
         Data columns (total 9 columns):
         Pregnancies                 768 non-null int64
         Glucose                     768 non-null int64
         BloodPressure               768 non-null int64
         SkinThickness               768 non-null int64
         Insulin                     768 non-null int64
         BMI                         768 non-null float64
         DiabetesPedigreeFunction    768 non-null float64
         Age                         768 non-null int64
         Outcome                     768 non-null int64
         dtypes: float64(2), int64(7)
         memory usage: 54.0 KB
```

```
In [40]: data.describe()
```

```
Out[40]:       Pregnancies     Glucose  BloodPressure  SkinThickness     Insulin  \
         count   768.000000  768.000000     768.000000     768.000000  768.000000
         mean      3.845052  120.894531      69.105469      20.536458   79.799479
         std       3.369578   31.972618      19.355807      15.952218  115.244002
         min       0.000000    0.000000       0.000000       0.000000    0.000000
         25%       1.000000   99.000000      62.000000       0.000000    0.000000
         50%       3.000000  117.000000      72.000000      23.000000   30.500000
         75%       6.000000  140.250000      80.000000      32.000000  127.250000
         max      17.000000  199.000000     122.000000      99.000000  846.000000

                   BMI  DiabetesPedigreeFunction         Age     Outcome
         count  768.000000                768.000000  768.000000  768.000000
         mean    31.992578                  0.471876   33.240885    0.348958
         std     7.884160                   0.331329   11.760232    0.476951
         min     0.000000                   0.078000    21.000000    0.000000
         25%    27.300000                   0.243750    24.000000    0.000000
         50%    32.000000                   0.372500    29.000000    0.000000
         75%    36.600000                   0.626250    41.000000    1.000000
         max    67.100000                   2.420000    81.000000    1.000000
```

```
In [41]: data.columns
```

```
Out[41]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness',
                'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
                dtype='object')
```

```
In [42]: data['Pregnancies'].value_counts()
```

```
In [43]: data['Glucose'].value_counts()
```

```
In [44]: data['BloodPressure'].value_counts()
```

```
In [45]: data['SkinThickness'].value_counts()
```

```
In [46]: data['Insulin'].value_counts()
```

```
In [47]: data['BMI'].value_counts()
```

```
In [48]: data['DiabetesPedigreeFunction'].value_counts()
```

```
In [49]: data['Age'].value_counts()
```

```
In [50]: data.dtypes
```

```
Out[50]: Pregnancies                 int64
         Glucose                     int64
         BloodPressure               int64
         SkinThickness               int64
         Insulin                     int64
         BMI                         int64
         DiabetesPedigreeFunction    int64
         Age                         int64
         Outcome                     int64
         dtype: int64
```

```
In [51]: data.hist(figsize=(7,7))
```

```
In [52]: data.head()
```

```
Out[52]:   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
         0            6      148             72             35        0  33.6
         1            1       85             66             29        0  26.6
         2            8      183             64              0        0  23.3
         3            1       89             66             23       94  28.1
         4            0      137             40             35      168  43.1

            DiabetesPedigreeFunction  Age  Outcome
         0                     0.627   50        1
         1                     0.351   31        0
         2                     0.672   32        1
         3                     0.167   21        0
         4                     2.288   33        1
```

```
In [53]: data.shape
```

```
Out[53]: (768, 9)
```

```
In [54]: data.info()
```

```
Out[54]: <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 768 entries, 0 to 767
         Data columns (total 9 columns):
         Pregnancies                 768 non-null int64
         Glucose                     768 non-null int64
         BloodPressure               768 non-null int64
         SkinThickness               768 non-null int64
         Insulin                     768 non-null int64
         BMI                         768 non-null float64
         DiabetesPedigreeFunction    768 non-null float64
         Age                         768 non-null int64
         Outcome                     768 non-null int64
         dtypes: float64(2), int64(7)
         memory usage: 54.0 KB
```

```
In [55]: data.describe()
```

```
Out[55]:       Pregnancies     Glucose  BloodPressure  SkinThickness     Insulin  \
         count   768.000000  768.000000     768.000000     768.000000  768.000000
         mean      3.845052  120.894531      69.105469      20.536458   79.799479
         std       3.369578   31.972618      19.355807      15.952218  115.244002
         min       0.000000    0.000000       0.000000       0.000000    0.000000
         25%       1.000000   99.000000      62.000000       0.000000    0.000000
         50%       3.000000  117.000000      72.000000      23.000000   30.500000
         75%       6.000000  140.250000      80.000000      32.000000  127.250000
         max      17.000000  199.000000     122.000000      99.000000  846.000000

                   BMI  DiabetesPedigreeFunction         Age     Outcome
         count  768.000000                768.000000  768.000000  768.000000
         mean    31.992578                  0.471876   33.240885    0.348958
         std     7.884160                   0.331329   11.760232    0.476951
         min     0.000000                   0.078000    21.000000    0.000000
         25%    27.300000                   0.243750    24.000000    0.000000
         50%    32.000000                   0.372500    29.000000    0.000000
         75%    36.600000                   0.626250    41.000000    1.000000
         max    67.100000                   2.420000    81.000000    1.000000
```

```
In [56]: data.columns
```

```
Out[56]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness',
                'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
                dtype='object')
```

```
In [57]: data['Pregnancies'].value_counts()
```

```
In [58]: data['Glucose'].value_counts()
```

```
In [59]: data['BloodPressure'].value_counts()
```

```
In [60]: data['SkinThickness'].value_counts()
```

```
In [61]: data['Insulin'].value_counts()
```

```
In [62]: data['BMI'].value_counts()
```

```
In [63]: data['DiabetesPedigreeFunction'].value_counts()
```

```
In [64]: data['Age'].value_counts()
```

```
In [65]: data.dtypes
```

```
Out[65]: Pregnancies                 int64
         Glucose                     int64
         BloodPressure               int64
         SkinThickness               int64
         Insulin                     int64
         BMI                         int64
         DiabetesPedigreeFunction    int64
         Age                         int64
         Outcome                     int64
         dtype: int64
```

```
In [66]: data.hist(figsize=(7,7))
```

```
In [67]: data.head()
```

```
Out[67]:   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
         0            6      148             72             35        0  33.6
         1            1       85             66             29        0  26.6
         2            8      183             64              0        0  23.3
         3            1       89             66             23       94  28.1
         4            0      137             40             35      168  43.1

            DiabetesPedigreeFunction  Age  Outcome
         0                     0.627   50        1
         1                     0.351   31        0
         2                     0.672   32        1
         3                     0.167   21        0
         4                     2.288   33        1
```

outerlier detection and treatment and plot boxplot for detecting outerlier

```
In [68]: data.columns
```

```
Out[68]: Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness',
                'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
                dtype='object')
```

```
In [69]: data.Pregnancies
```

```
Out[69]: 0      6
         1      1
         2      8
         3      1
         4      0
         ..     ..
         763    10
         764     2
         765     5
         766     1
         767     1
         Name: Pregnancies, Dtype: int64
```

```
In [70]: data.Pregnancies[(data.Pregnancies > 13)]
```

```
Out[70]: 759    14
         761    13
         Name: Pregnancies, Dtype: int64
```

```
In [71]: data.Pregnancies[(data.Pregnancies < 1)]
```

```
Out[71]: 4    0
         Name: Pregnancies, Dtype: int64
```

```
In [72]: data.Pregnancies[(data.Pregnancies > 13)] = 13
```

```
In [73]: data.Pregnancies[(data.Pregnancies < 1)] = 1
```

```
In [74]: data.Pregnancies
```

```
Out[74]: 0      6
         1      1
         2      8
         3      1
         4      1
         ..     ..
         763    10
         764     2
         765     5
         766     1
         767     1
         Name: Pregnancies, Dtype: int64
```

```
In [75]: data.Glucose
```

```
Out[75]: 0      148
         1       85
         2      183
         3       89
         4      137
         ..     ...
         763    101
         764    122
         765    121
         766    126
         767     93
         Name: Glucose, Dtype: int64
```

```
In [76]: data.Glucose[(data.Glucose > 190)]
```

```
Out[76]: 681    197
         684    195
         753    198
         759    199
         Name: Glucose, Dtype: int64
```

```
In [77]: data.Glucose[(data.Glucose < 40)]
```

```
Out[77]: 75    0
         Name: Glucose, Dtype: int64
```

```
In [78]: data.Glucose[(data.Glucose > 190)] = 190
```

```
In [79]: data.Glucose[(data.Glucose < 40)] = 40
```

```
In [80]: data.Glucose
```

```
Out[80]: 0      148
         1       85
         2      183
         3       89
         4      137
         ..     ...
         763    101
         764    122
         765    121
         766    126
         767     93
         Name: Glucose, Dtype: int64
```

```
In [81]: data.BloodPressure
```

```
Out[81]: 0      72
         1      66
         2      64
         3      66
         4      40
         ..     ..
         763    76
         764    70
         765    72
         766    60
         767    70
         Name: BloodPressure, Dtype: int64
```

```
In [82]: data.BloodPressure[(data.BloodPressure > 110)]
```

```
Out[82]: 43    114
         84    122
         Name: BloodPressure, Dtype: int64
```

```
In [83]: data.BloodPressure[(data.BloodPressure < 40)]
```

```
Out[83]: 7    0
         Name: BloodPressure, Dtype: int64
```

```
In [84]: data.BloodPressure[(data.BloodPressure > 110)] = 110
```

```
In [85]: data.BloodPressure[(data.BloodPressure < 40)] = 40
```

```
In [86]: data.BloodPressure
```

```
Out[86]: 0      72
         1      66
         2      64
         3      66
         4      40
         ..     ..
         763    76
         764    70
         765    72
         766    60
         767    70
         Name: BloodPressure, Dtype: int64
```

```
In [87]: data.SkinThickness
```

```
Out[87]: 0      35
         1      29
         2       0
         3      23
         4      35
         ..     ..
         763    48
         764    27
         765    23
         766    31
         767    41
         Name: SkinThickness, Dtype: int64
```

```
In [88]: data.SkinThickness[(data.SkinThickness > 60)]
```

```
Out[88]: 579    63
         579    99
         Name: SkinThickness, Dtype: int64
```

```
In [89]: data.SkinThickness[(data.SkinThickness < 10)]
```

```
Out[89]: 2    0
         5    0
         7    0
         9    0
         Name: SkinThickness, Dtype: int64
```

```
In [90]: data.SkinThickness[(data.SkinThickness > 60)] = 60
```

```
In [91]: data.SkinThickness[(data.SkinThickness < 10)] = 10
```

```
In [92]: data.SkinThickness
```

```
Out[92]: 0      35
         1      29
         2      10
         3      23
         4      35
         ..     ..
         763    48
         764    27
         765    23
         766    31
         767    41
         Name: SkinThickness, Dtype: int64
```

```
In [93]: data.Insulin
```

```
Out[93]: 0      0
         1      0
         2      0
         3     94
         4    168
         ..     ...
         763    0
         764    0
         765    0
         766    0
         767    0
         Name: Insulin, Dtype: int64
```

```
In [94]: data.Insulin[(data.Insulin > 300)]
```

```
Out[94]: 8    846
         Name: Insulin, Dtype: int64
```

```
In [95]: data.Insulin[(data.Insulin < 40)]
```

```
Out[95]: 0    0
         1    0
         2    0
         5    0
         7    0
         9    0
         Name: Insulin, Dtype: int64
```

```
In [96]: data.Insulin[(data.Insulin > 300)] = 300
```

```
In [97]: data.Insulin[(data.Insulin < 40)] = 40
```

```
In [98]: data.Insulin
```

```
Out[98]: 0      40
         1      40
         2      40
         3      94
         4     168
         ..     ...
         763    40
         764    40
         765    40
         766    40
         767    40
         Name: Insulin, Dtype: int64
```

```
In [99]: data.BMI
```

```
Out[99]: 0      33.6
         1      26.6
         2      23.3
         3      28.1
         4      43.1
         ...     ...
         763    32.9
         764    36.8
         765    26.2
         766    30.1
         767    30.4
         Name: BMI, Dtype: float64
```

```
In [100]: data.BMI[(data.BMI > 60)]
```

```
Out[100]: 9    67.1
          Name: BMI, Dtype: float64
```

```
In [101]: data.BMI[(data.BMI < 18)]
```

```
Out[101]: 32    0.0
          Name: BMI, Dtype: float64
```

```
In [102]: data.BMI[(data.BMI > 60)] = 60
```

```
In [103]: data.BMI[(data.BMI < 18)] = 18
```

```
In [104]: data.BMI
```

```
Out[104]: 0      33.6
          1      26.6
          2      23.3
          3      28.1
          4      43.1
          ...     ...
          763    32.9
          764    36.8
          765    26.2
          766    30.1
          767    30.4
          Name: BMI, Dtype: float64
```

```
In [105]: data.DiabetesPedigreeFunction
```

```
Out[105]: 0      0.627
          1      0.351
          2      0.672
          3      0.167
          4      2.288
          ...      ...
          763    0.171
          764    0.340
          765    0.245
          766    0.349
          767    0.315
          Name: DiabetesPedigreeFunction, Dtype: float64
```

```
In [106]: data.DiabetesPedigreeFunction[(data.DiabetesPedigreeFunction > 1.5)]
```

```
Out[106]: 12    2.329
          13    1.232
          228   1.731
          228   2.288
          Name: DiabetesPedigreeFunction, Dtype: float64
```

```
In [107]: data.DiabetesPedigreeFunction[(data.DiabetesPedigreeFunction < 0.1)]
```

```
Out[107]: 291    0.085
          291    0.088
          Name: DiabetesPedigreeFunction, Dtype: float64
```

```
In [108]: data.DiabetesPedigreeFunction[(data.DiabetesPedigreeFunction > 1.5)] = 1.5
```

```
In [109]: data.DiabetesPedigreeFunction[(data.DiabetesPedigreeFunction < 0.1)] = 0.1
```

```
In [110]: data.DiabetesPedigreeFunction
```

```
Out[110]: 0      0.627
          1      0.351
          2      0.672
          3      0.167
          4      2.288
          ...      ...
          763    0.171
          764    0.340
          765    0.245
          766    0.349
          767    0.315
          Name: DiabetesPedigreeFunction, Dtype: float64
```

```
In [111]: data.Age
```

```
Out[111]: 0    50
          1    31
          2    32
          3    21
          4    33
          ..   ..
          763  41
          764  26
          765  30
          766  47
          767  23
          Name: Age, Dtype: int64
```

```
In [112]: data.Age[(data.Age > 70)]
```

```
Out[112]: 459    81
          Name: Age, Dtype: int64
```

```
In [113]: data.Age[(data.Age < 18)]
```

```
Out[113]: Series([], Name: Age, Dtype: int64)
```

```
In [114]: data.Age[(data.Age > 70)] = 70
```

```
In [115]: data.Age[(data.Age < 18)] = 18
```

```
In [116]: data.Age
```

```
Out[116]: 0    50
          1    31
          2    32
          3    21
          4    33
          ..   ..
          763  41
          764  26
          765  30
          766  47
          767  23
          Name: Age, Dtype: int64
```

```
In [117]: data.dtypes
```

```
Out[117]: Pregnancies                 int64
          Glucose                     int64
          BloodPressure               int64
          SkinThickness               int64
          Insulin                     int64
          BMI                         int64
          DiabetesPedigreeFunction    int64
          Age                         int64
          Outcome                     int64
          dtype: int64
```

```
In [118]: data.hist(figsize=(7,7))
```

```
In [119]: data.head()
```

```
Out[119]:   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
          0            6      148             72             35        0  33.6
          1            1       85             66             29        0  26.6
          2            8      183             64              0        0  23.3
          3            1       89             66             23       94  28.1
          4            0      137             40             35      168  43.1

             DiabetesPedigreeFunction  Age  Outcome
          0                     0.627   50        1
          1                     0.351   31        0
          2                     0.672   32        1
          3                     0.167   21        0
          4                     2.288   33        1
```

```
In [120]: data.shape
```

```
Out[120]: (768, 9)
```

```
In [121]: data.info()
```

```
Out[121]: <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 768 entries, 0 to 767
          Data columns (total 9 columns):
          Pregnancies                 768 non-null int64
          Glucose                     768 non-null int64
          BloodPressure               768 non-null int64
          SkinThickness               768 non-null int64
          Insulin                     768 non-null int64
          BMI                         768 non-null float64
          DiabetesPedigreeFunction    768 non-null float64
          Age                         768 non-null int64
          Outcome                     768 non-null int64
          dtypes: float64(2), int64(7)
          memory usage: 54.0 KB
```

```
In [122]: data.describe()
```

```
Out[122]:       Pregnancies     Glucose  BloodPressure  SkinThickness     Insulin  \
          count   768.000000  768.000000     768.000000     768.000000  768.000000
          mean      3.845052  120.894531      69.105469      20.536458   79.799479
          std       3.369578   31.972618      19.355807      15.952218  115.244002
          min       0.000000    0.000000       0.000000       0.000000    0.000000
          25%       1.000000   99.000000      62.000000       0.000000    0.000000
          50%       3.000000  117.000000      72.000000      23.000000   30.500000
          75%       6.000000  140.250000      80.000000      32.000000  127.250000
          max      17.000000  199.000000     122.000000      99.000000  846.000000

                    BMI  DiabetesPedigreeFunction         Age     Outcome
          count  768.000000                768.000000  768.000000  768.000000
          mean    31.992578                  0.471876   33.240885    0.348958
          std     7.884160                   0.331329   11.760232    0.476951
          min     0.000000                   0.078000    21.000000    0.000000
          25%    27.300000                   0.243750    24.000000    0.000000
          50%    32.000000                   0.372500    29.000000    0.000000
          75%    36.600000                   0.626250    41.000000    1.000000
          max    67.100000                   2.420000    81.000000    1.000000
```

again apply boxplot

```
In [123]: data.Pregnancies
```

```
Out[123]: 0      6
          1      1
          2      8
          3      1
          4      1
          ..     ..
          763    10
          764     2
          765     5
          766     1
          767     1
          Name: Pregnancies, Dtype: int64
```

```
In [124]: data.Pregnancies[(data.Pregnancies > 13)]
```

```
Out[124]: Series([], Name: Pregnancies, Dtype: int64)
```

```
In [125]: data.Pregnancies[(data.Pregnancies < 1)]
```

```
Out[125]: Series([], Name: Pregnancies, Dtype: int64)
```

```
In [126]: data.Pregnancies[(data.Pregnancies > 13)] = 13
```

```
In [127]: data.Pregnancies[(data.Pregnancies < 1)] = 1
```

```
In [128]: data.Pregnancies
```

```
Out[128]: 0      6
          1      1
          2      8
          3      1
          4      1
          ..     ..
          763    10
          764     2
          765     5
          766     1
          767     1
          Name: Pregnancies, Dtype: int64
```

```
In [129]: data.Glucose
```

```
Out[129]: 0      148
          1       85
          2      183
          3       89
          4      137
          ..     ...
          763    101
          764    122
          765    121
          766    126
          767     93
          Name: Glucose, Dtype: int64
```

```
In [130]: data.Glucose[(data.Glucose > 190)]
```

```
Out[130]: 681    190
          684    190
          753    190
          759    190
          Name: Glucose, Dtype: int64
```

```
In [131]: data.Glucose[(data.Glucose < 40)]
```

```
Out[131]: 75    40
          Name: Glucose, Dtype: int64
```

```
In [132]: data.Glucose[(data.Glucose > 190)] = 190
```

```
In [133]: data.Glucose[(data.Glucose < 40)] = 40
```

```
In [134]: data.Glucose
```

```
Out[134]: 0      148
          1       85
          2      183
          3       89
          4      137
          ..     ...
          763    101
          764    122
          765    121
          766    126
          767     93
          Name: Glucose, Dtype: int64
```

```
In [135]: data.BloodPressure
```

```
Out[135]: 0      72
          1      66
          2      64
          3      66
          4      40
          ..     ..
          763    76
          764    70
          765    72
          766    60
          767    70
          Name: BloodPressure, Dtype: int64
```

```
In [136]: data.BloodPressure[(data.BloodPressure > 110)]
```

```
Out[136]: 43    110
          84    110
          Name: BloodPressure, Dtype: int64
```

```
In [137]: data.BloodPressure[(data.BloodPressure < 40)]
```

```
Out[137]: 7    40
          Name: BloodPressure, Dtype: int64
```

```
In [138]: data.BloodPressure[(data.BloodPressure > 110)] = 110
```

```
In [139]: data.BloodPressure[(data.BloodPressure < 40)] = 40
```

```
In [140]: data.BloodPressure
```

```
Out[140]: 0      72
          1      66
          2      64
          3      66
          4      40
          ..     ..
          763    76
          764    70
          765    72
          766    60
          767    70
          Name: BloodPressure, Dtype: int64
```

```
In [141]: data.SkinThickness
```

```
Out[141]: 0      35
          1      29
          2      10
          3      23
          4      35
          ..     ..
          763    48
          764    27
          765    23
          766    31
          767    41
          Name: SkinThickness, Dtype: int64
```

```
In [142]: data.SkinThickness[(data.SkinThickness > 60)]
```

```
Out[142]: 579    60
          579    60
          Name: SkinThickness, Dtype: int64
```

```
In [143]: data.SkinThickness[(data.SkinThickness < 10)]
```

```
Out[143]: 2    10
          5    10
          7    10
          9    10
          Name: SkinThickness, Dtype: int64
```

```
In [144]: data.SkinThickness[(data.SkinThickness > 60)] = 60
```

```
In [145]: data.SkinThickness[(data.SkinThickness < 10)] = 10
```

```
In [146]: data.SkinThickness
```

```
Out[146]: 0      35
          1      29
          2      10
          3      23
          4      35
          ..     ..
          763    48
          764    27
          765    23
          766    31
          767    41
          Name: SkinThickness, Dtype: int64
```

```
In [147]: data.Insulin
```

```
Out[147]: 0      40
          1      40
          2      40
          3      94
          4     168
          ..     ...
          763    40
          764    40
          765    40
          766    40
          767    40
          Name: Insulin, Dtype: int64
```

```
In [148]: data.Insulin[(data.Insulin > 300)]
```

```
Out[148]: 8    300
          Name: Insulin, Dtype: int64
```

```
In [149]: data.Insulin[(data.Insulin < 40)]
```

```
Out[149]: 0    40
          1    40
          2    40
          5    40
          7    40
          9    40
          Name: Insulin, Dtype: int64
```

```
In [150]: data.Insulin[(data.Insulin > 300)] = 300
```

```
In [151]: data.Insulin[(data.Insulin < 40)] = 40
```

```
In [152]: data.Insulin
```

```
Out[152]: 0      40
          1      40
          2      40
          3      94
          4     168
          ..     ...
          763    40
          764    40
          765    40
          766    40
          767    40
          Name: Insulin, Dtype: int64
```

```
In [153]: data.BMI
```

```
Out[153]: 0      33.6
          1      26.6
          2      23.3
          3      28.1
          4      43.1
          ...     ...
          763    32.9
          764    36.8
          765    26.2
          766    30.1
          767    30.4
          Name: BMI, Dtype: float64
```

```
In [154]: data.BMI[(data.BMI > 60)]
```

```
Out[154]: 9    60.0
          Name: BMI, Dtype: float64
```

```
In [155]: data.BMI[(data.BMI < 18)]
```

```
Out[155]: 32    18.0
          Name: BMI, Dtype: float64
```

```
In [156]: data.BMI[(data.BMI > 60)] = 60
```

```
In [157]: data.BMI[(data.BMI < 18)] = 18
```

```
In [158]: data.BMI
```

```
Out[158]: 0      33.6
          1      26.6
          2      23.3
          3      28.1
          4      43.1
          ...     ...
          763    32.9
          764    36.8
          765    26.2
          766    30.1
          767    30.4
          Name: BMI, Dtype: float64
```

```
In [159]: data.DiabetesPedigreeFunction
```

```
Out[159]: 0      0.627
          1      0.351
          2      0.672
          3      0.167
          4      2.288
          ...      ...
          763    0.171
          764    0.340
          765    0.245
          766    0.349
          767    0.315
          Name: DiabetesPedigreeFunction, Dtype: float64
```

```
In [160]: data.DiabetesPedigreeFunction[(data.DiabetesPedigreeFunction > 1.5)]
```

```
Out[160]: 12    2.329
          13    1.232
          228   1.731
          228   2.288
          Name: DiabetesPedigreeFunction, Dtype: float64
```

```
In [161]: data.DiabetesPedigreeFunction[(data.DiabetesPedigreeFunction < 0.1)]
```

```
Out[161]: 291    0.085
          291    0.088
          Name: DiabetesPedigreeFunction, Dtype: float64
```

```
In [162]: data.DiabetesPedigreeFunction[(data.DiabetesPedigreeFunction > 1.5)] = 1.5
```

```
In [163]: data.DiabetesPedigreeFunction[(data.DiabetesPedigreeFunction < 0.1)] = 0.1
```

```
In [164]: data.DiabetesPedigreeFunction
```

```
Out[164]: 0      0.627
          1      0.351
          2      0.672
          3      0.167
          4      2.288
          ...      ...
          763    0.171
          764    0.340
          765    0.245
          766    0.349
          767    0.315
          Name: DiabetesPedigreeFunction, Dtype: float64
```

```
In [165]: data.Age
```

```
Out[165]: 0    50
          1    31
          2    32
          3    21
          4    33
          ..   ..
          763  41
          764  26
          765  30
          766  47
          767  23
          Name: Age, Dtype: int64
```