Part_I_exploration_maternal_health

January 9, 2023

1 Part I - Exploration of data on Maternal Health Risk

1.1 by Tova Sonntag

1.2 Introduction

This project focuses on the data set on maternal health risk, which can be found on the Kaggle website: https://www.kaggle.com/datasets/csafrit2/maternal-health-risk-data The data has been collected from hospitals, clinics and maternal health care centers. The main goal of the survey was to find which health conditions can be considered indicators for high risks during pregnancy.

1.3 Preliminary Wrangling

```
In [2]: # import all packages and set plots to be embedded inline
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sb

        %matplotlib inline

In [3]: data_original = pd.read_csv("maternal_health_risk.csv")

In [4]: data_original.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1014 entries, 0 to 1013
Data columns (total 7 columns):
Age            1014 non-null int64
SystolicBP     1014 non-null int64
DiastolicBP    1014 non-null int64
BS             1014 non-null float64
BodyTemp       1014 non-null float64
HeartRate      1014 non-null int64
RiskLevel      1014 non-null object
dtypes: float64(2), int64(4), object(1)
memory usage: 55.5+ KB
```

In [5]: data_original.describe()

Out[5]:

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	\
count	1014.000000	1014.000000	1014.000000	1014.000000	1014.000000	
mean	29.871795	113.198225	76.460552	8.725986	98.665089	
std	13.474386	18.403913	13.885796	3.293532	1.371384	
min	10.000000	70.000000	49.000000	6.000000	98.000000	
25%	19.000000	100.000000	65.000000	6.900000	98.000000	
50%	26.000000	120.000000	80.000000	7.500000	98.000000	
75%	39.000000	120.000000	90.000000	8.000000	98.000000	
max	70.000000	160.000000	100.000000	19.000000	103.000000	

	HeartRate
count	1014.000000
mean	74.301775
std	8.088702
min	7.000000
25%	70.000000
50%	76.000000
75%	80.000000
max	90.000000

In [6]: data_original.head()

Out[6]:

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
0	25	130	80	15.0	98.0	86	high risk
1	35	140	90	13.0	98.0	70	high risk
2	29	90	70	8.0	100.0	80	high risk
3	30	140	85	7.0	98.0	70	high risk
4	35	120	60	6.1	98.0	76	low risk

1.3.1 What is the structure of your dataset?

The data is available from 1014 pregnant women. The information given is the age, systolic blood pressure, distolic blood pressure, blood sugar, body temperture, heart rate and the risk level. All the data is numeric with the exception of the last column, which is the risk level, which can be either high, mid or low. I will add an 8.variable, the pulse pressure, which is the difference between the systolic and diastolic blood pressure and is also used as an indicator of medical conditions.

1.3.2 What is/are the main feature(s) of interest in your dataset?

I want to see which health conditions have the highest impact on the risk level during pregnancy.

1.3.3 What features in the dataset do you think will help support your investigation into your feature(s) of interest?

I assume that advanced maternal age, high blood pressure and elevated blood sugar levels will have a strong correlation with high risk pregnancies.

Before the data exploration, I will copy the original data frame and add the pulse pressure column:

```
In [7]: data = data_original.copy()
```

```
In [8]: data["PulseP"] = data["SystolicBP"] - data["DiastolicBP"]
```

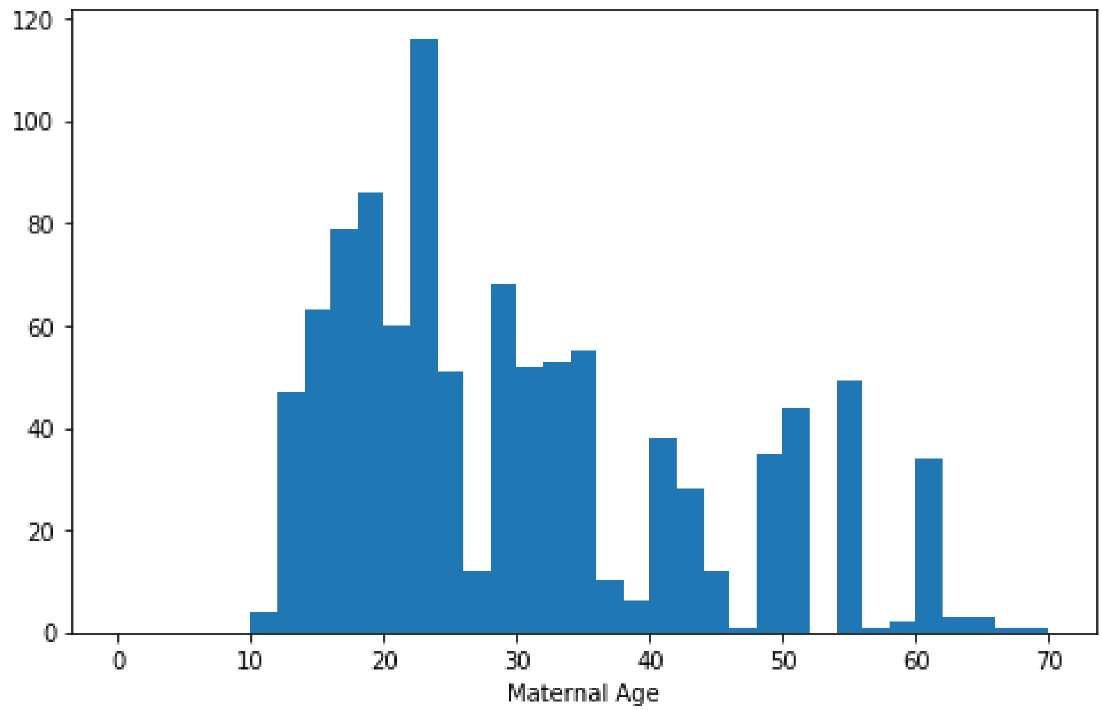
```
In [9]: data.head()
```

```
Out[9]:    Age  SystolicBP  DiastolicBP    BS  BodyTemp  HeartRate  RiskLevel  PulseP
        0   25         130           80  15.0      98.0         86  high risk      50
        1   35         140           90  13.0      98.0         70  high risk      50
        2   29          90           70   8.0     100.0         80  high risk      20
        3   30         140           85   7.0      98.0         70  high risk      55
        4   35         120           60   6.1      98.0         76   low risk      60
```

1.4 Univariate Exploration

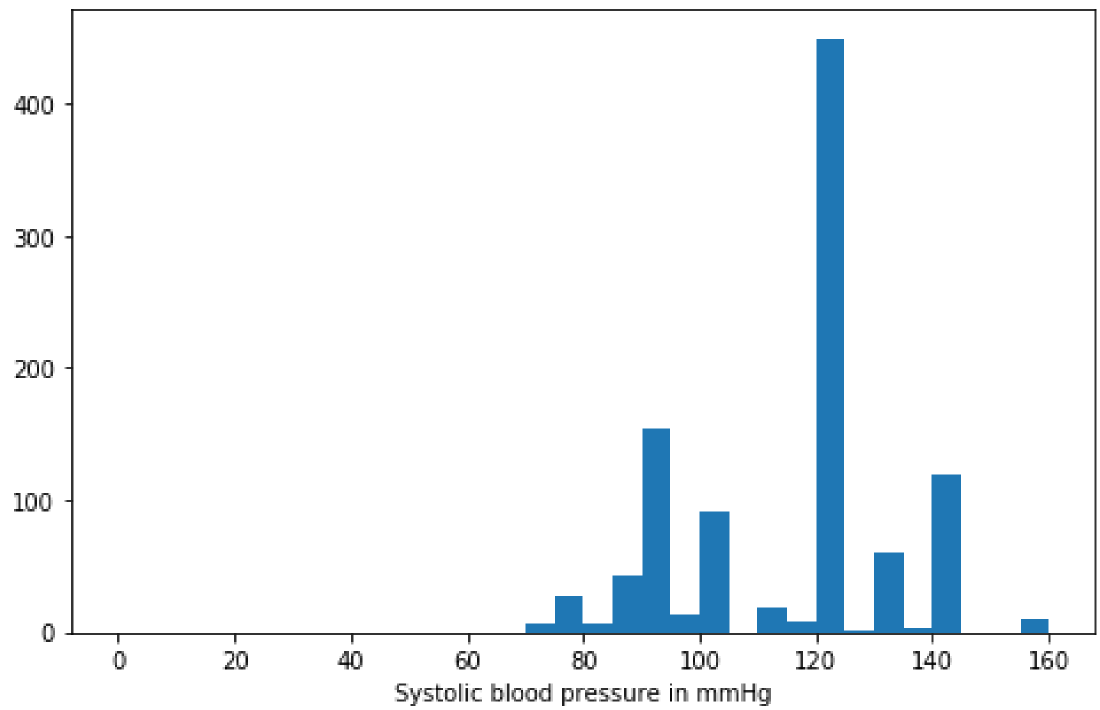
I will start by plotting the distribution of the variables to have an understanding of the ranges and see if there are outliers.

```
In [10]: #The first variable I will focus on is the maternal age

         def plot_hist(binsize, data, x, plt_title):
             binsize = binsize
             bins = np.arange(0, data[x].max()+binsize, binsize)

             plt.figure(figsize=[8, 5])
             plt.hist(data = data, x = x, bins = bins)
             plt.xlabel(plt_title)
             plt.show()

         plot_hist(2, data, 'Age', 'Maternal Age')
```



We see that in this data set, most of the participants were in their early twenties as well as teenage mothers. We also find a significant number of women in their early thirties who took part in the survey. The youngest pregnant woman was just 10 years old, the oldest mother that participated was 70, which is an extreme outlier, but still possible with IVF technology that exists nowadays.

```
In [11]: #Next, I would like to have a look at the distribution of the systolic blood pressure (

         plot_hist(5, data, 'SystolicBP', 'Systolic blood pressure in mmHg')
```




```
In [12]: data.SystolicBP.value_counts()
```

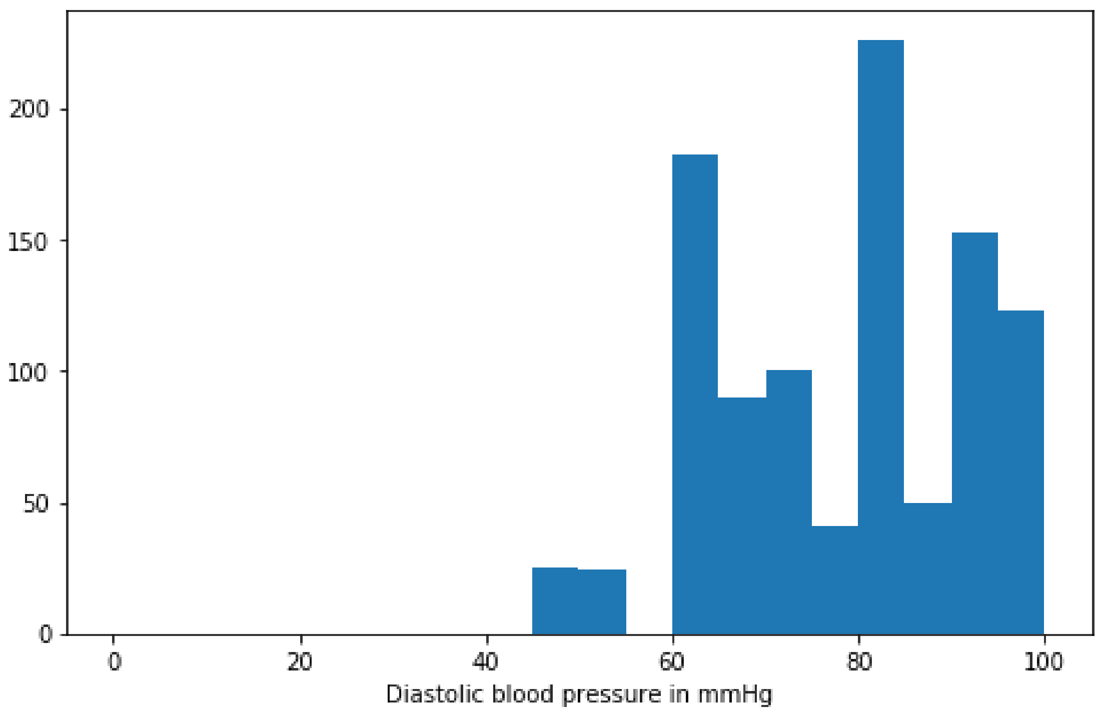
```
Out[12]: 120    449
         90     154
         140    120
         100     92
         130     60
         85      43
         110     19
         76      16
         95      12
         160     10
         75       8
         115      8
         70       7
         80       5
         78       3
         135      3
         83       2
         99       2
         129      1
         Name: SystolicBP, dtype: int64
```

We can see that the lowest systolic blood pressure recorded in the data is 70, the maximum value is 160. The most common value is 120.

According to the American College of Obstetricians and Gynecologists (ACOG), values of over 140 are considered hypertension and indicate potentially serious health conditions such as preeclampsia.

See: https://www.acog.org/womens-health/experts-and-stories/the-latest/whats-the-concern-about-high-blood-pressure-during-pregnancy-an-ob-gyn-explains#:~:text=Your%20blood%20pressure%20is%20measured,120%2F80%20is%20considered%20normal.

In [13]:
```
#Plotting the distolic blood pressure (mmHg)

plot_hist(5, data, 'DiastolicBP', 'Diastolic blood pressure in mmHg')
```



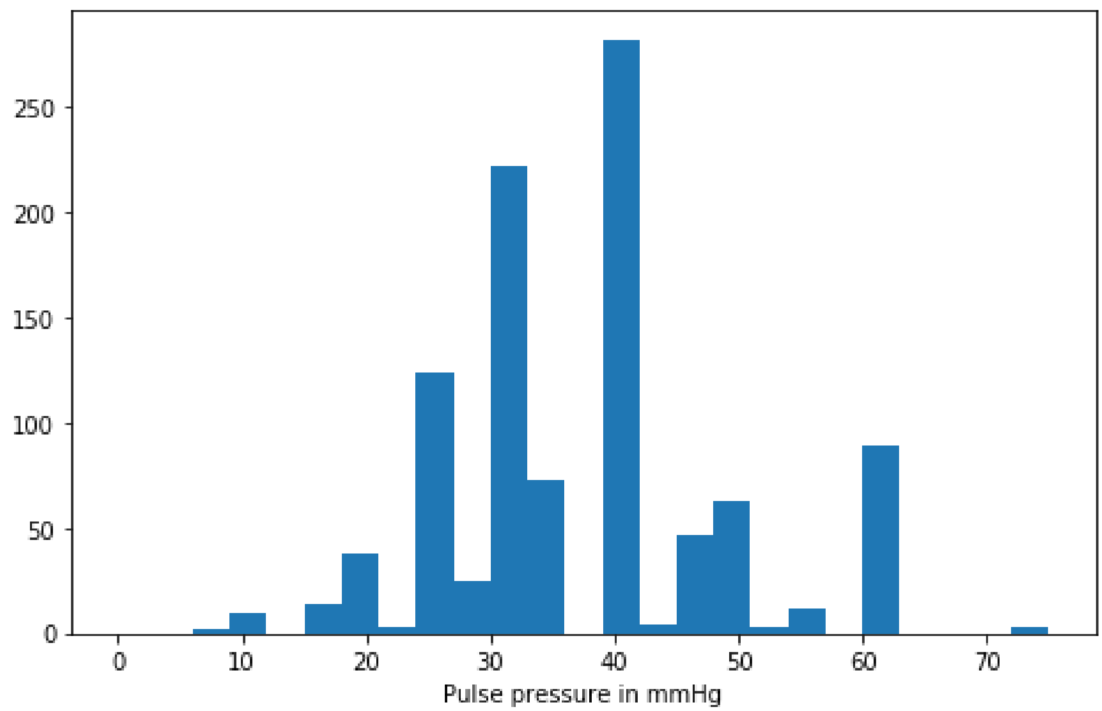
In [14]:
```
data.DiastolicBP.value_counts()
```

Out[14]:
```
80     226
60     174
90     153
70     100
100     87
65      87
85      49
75      38
95      36
49      25
50      24
```

```
63        8
76        3
68        2
89        1
69        1
Name: DiastolicBP, dtype: int64
```

In the diastolic blood pressure levels, the most common value is 80, which is a healthy level, the minimum is 50 and the maximum is 100. Values over 90 are considered indicators of high risk pregnancies according to the ACOG (see link above).

In [15]:
```
#Plotting the pulse pressure

plot_hist(3, data, 'PulseP', 'Pulse pressure in mmHg')
```



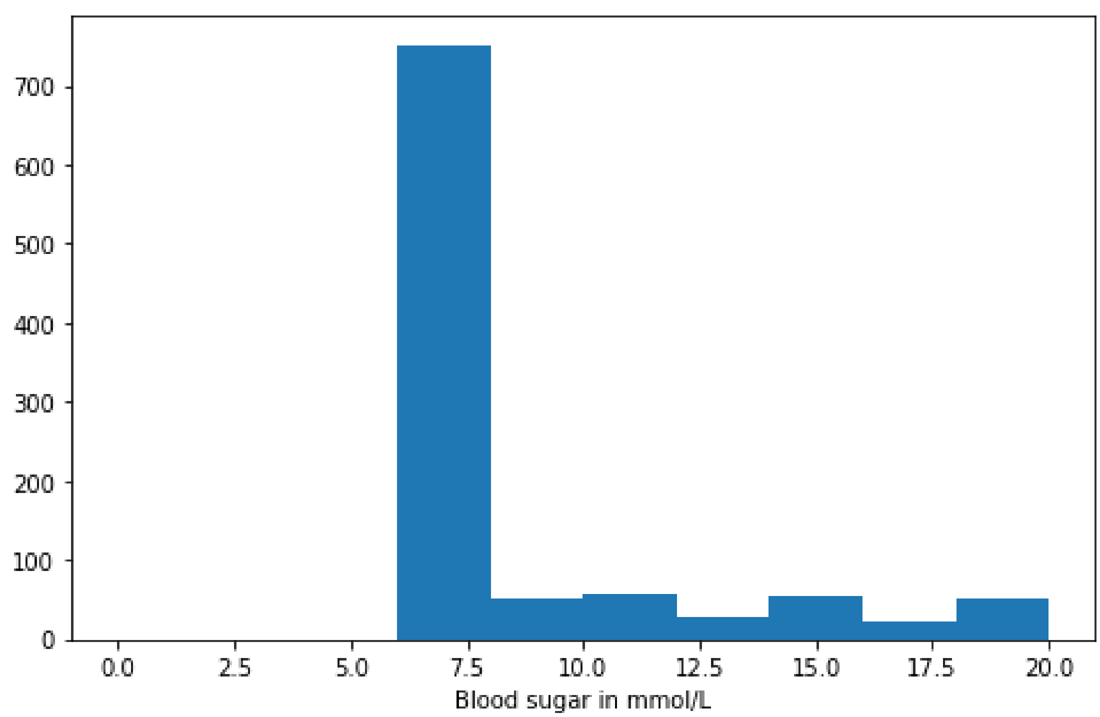
In [16]:
```
data.PulseP.value_counts()
```

Out[16]:
```
40    276
30    222
25    123
60     89
35     73
50     63
45     47
```

```
20    38
27    22
15    13
55    12
10    10
44     4
41     4
75     3
29     3
51     3
23     2
39     2
8      2
26     1
21     1
16     1
Name: PulseP, dtype: int64
```

The pulse pressure recorded in our data ranges between 8 and 75 with 40 as the mode. There are studies that show that an elevated pulse pressure prior or during pregnancy is an indicator of higher cardiovascular risks (see: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3118522/). It would be interesting to explore if the pregnancies in the present data that were categorized as "high risk" have higher values of pulse pressure as well.

```
In [17]: #Blood sugar levels in mmol/L

         plot_hist(2, data, 'BS', 'Blood sugar in mmol/L')
```



In [18]: data.BS.describe()

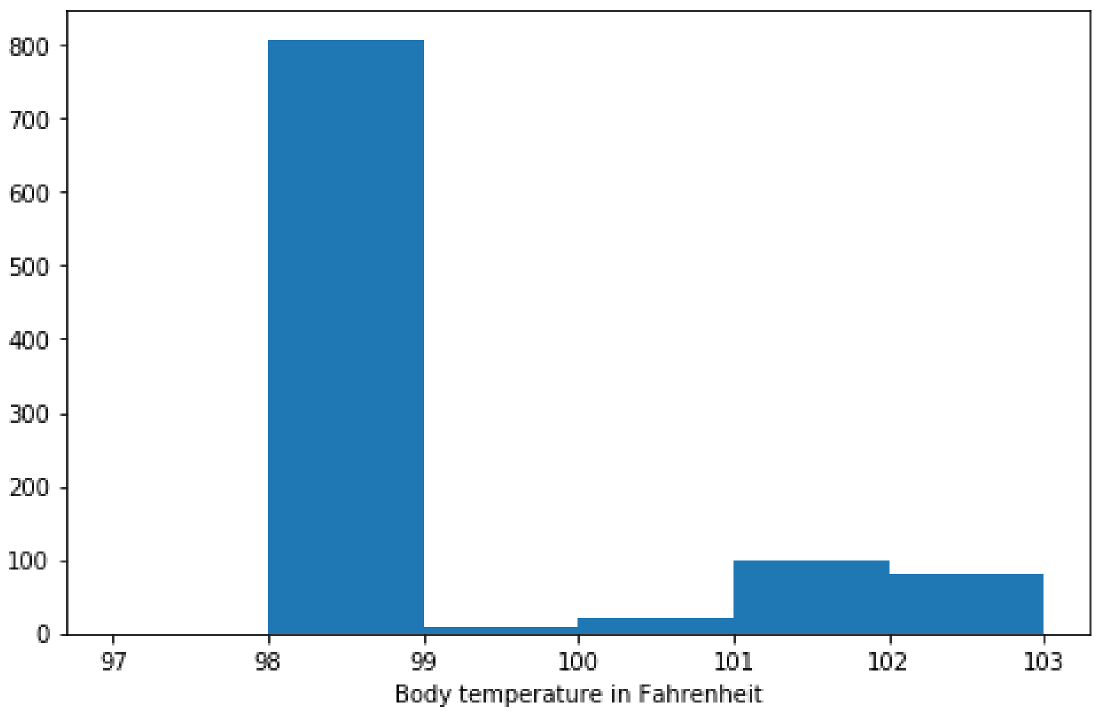
Out[18]:
```
count    1014.000000
mean        8.725986
std         3.293532
min         6.000000
25%         6.900000
50%         7.500000
75%         8.000000
max        19.000000
Name: BS, dtype: float64
```

In [19]: data.BS.value_counts()

Out[19]:
```
7.50     176
6.90     113
6.80      88
7.00      79
7.90      60
15.00     54
6.10      53
11.00     52
7.80      45
6.70      33
9.00      31
18.00     29
7.70      24
19.00     22
8.00      22
6.00      21
7.20      20
12.00     18
16.00     17
7.01      15
6.40      10
13.00      9
7.10       8
17.00      5
10.00      4
6.30       2
6.60       2
7.60       1
6.50       1
Name: BS, dtype: int64
```

The mode is 7.5, which is considered a healthy level. The maximum is 19, which is extremely high since all levels above 11.0 mmol/L are indicators of gestational diabetes, a dangerous condition both for the mother and the baby. Among the many risks are a premature birth and birth complications. Also, the mother is more likely to develop diabetes later in life. See: https://www.diabete.qc.ca/en/understand-diabetes/all-about-diabetes/types-of-diabetes/diabetes-in-pregnancy/

In [20]:
```
#Plotting the body temperature

binsize = 1
bins = np.arange(97, data['BodyTemp'].max()+binsize, binsize)

plt.figure(figsize=[8, 5])
plt.hist(data = data, x = 'BodyTemp', bins = bins)
plt.xlabel('Body temperature in Fahrenheit')
plt.show()
```



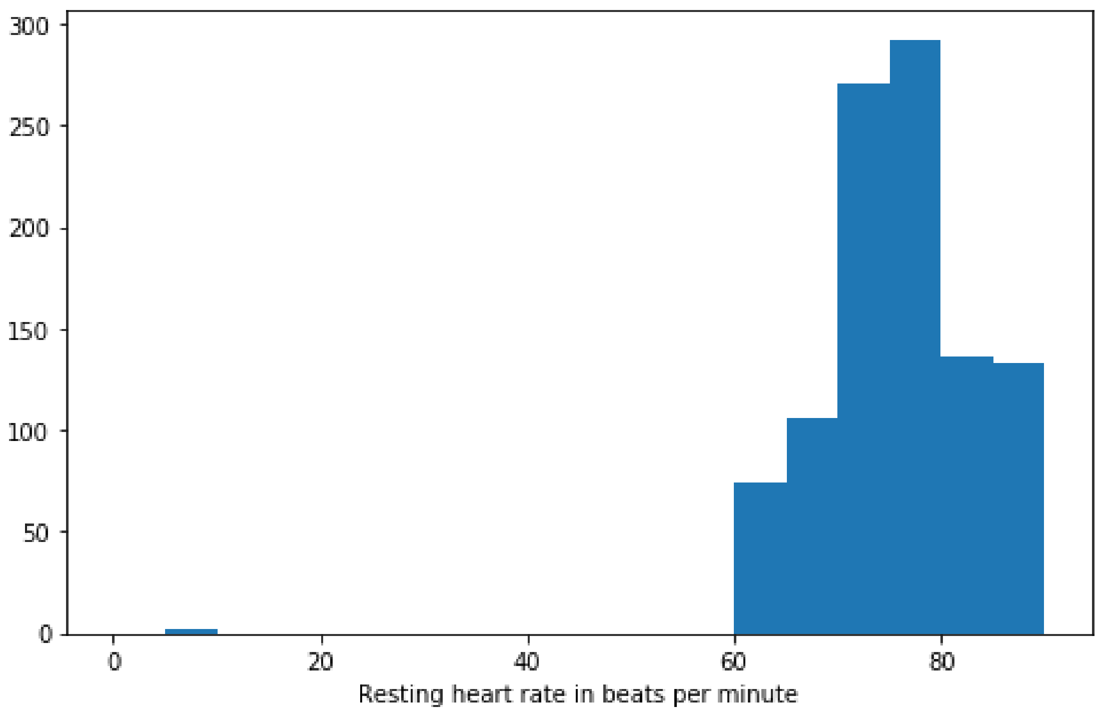
In [21]:
```
data.BodyTemp.describe()
```

Out[21]:
```
count    1014.000000
mean       98.665089
std         1.371384
min        98.000000
25%        98.000000
```

```
50%         98.000000
75%         98.000000
max        103.000000
Name: BodyTemp, dtype: float64
```

The lowest body temperature recorded is 98řF. So is the mode, which is a healthy value. The maximum is 103řF, a fever probably caused by some illness. Elevated body temperature during pregnancy can be dangerous, especially in the early stages and if it lasts for a long period of time. See: https://www.ncbi.nlm.nih.gov/books/NBK582757/

In [22]:
```
#Plotting the heart rate

plot_hist(5, data, 'HeartRate', 'Resting heart rate in beats per minute')
```



In [23]:
```
data.HeartRate.describe()
```

Out[23]:
```
count    1014.000000
mean       74.301775
std         8.088702
min         7.000000
25%        70.000000
50%        76.000000
75%        80.000000
max        90.000000
Name: HeartRate, dtype: float64
```

```
In [24]: data.HeartRate.value_counts()
```

```
Out[24]: 70    271
         76    131
         80    117
         77     96
         66     87
         60     74
         88     59
         86     55
         78     46
         90     19
         82     19
         75     19
         67     12
         65      5
         68      2
         7       2
         Name: HeartRate, dtype: int64
```

```
In [25]: # 2 of the records were 7, which are outliers and seem improbable for a heart rate of t

         data.loc[data['HeartRate']==7]
```

```
Out[25]:      Age  SystolicBP  DiastolicBP   BS  BodyTemp  HeartRate RiskLevel  PulseP
         499   16         120           75  7.9      98.0          7  low risk      45
         908   16         120           75  7.9      98.0          7  low risk      45
```

Considering that the 2 patients for whom the heart rates were recorded as 7 have a normal body temperature, healthy blood pressure and low risk levels, we can assume that there was some mistake when the data was collected or recorded. Since these values do not pass the sanity check, I would replace them with 70 (probably the real original values), which is also the mode in this data set.

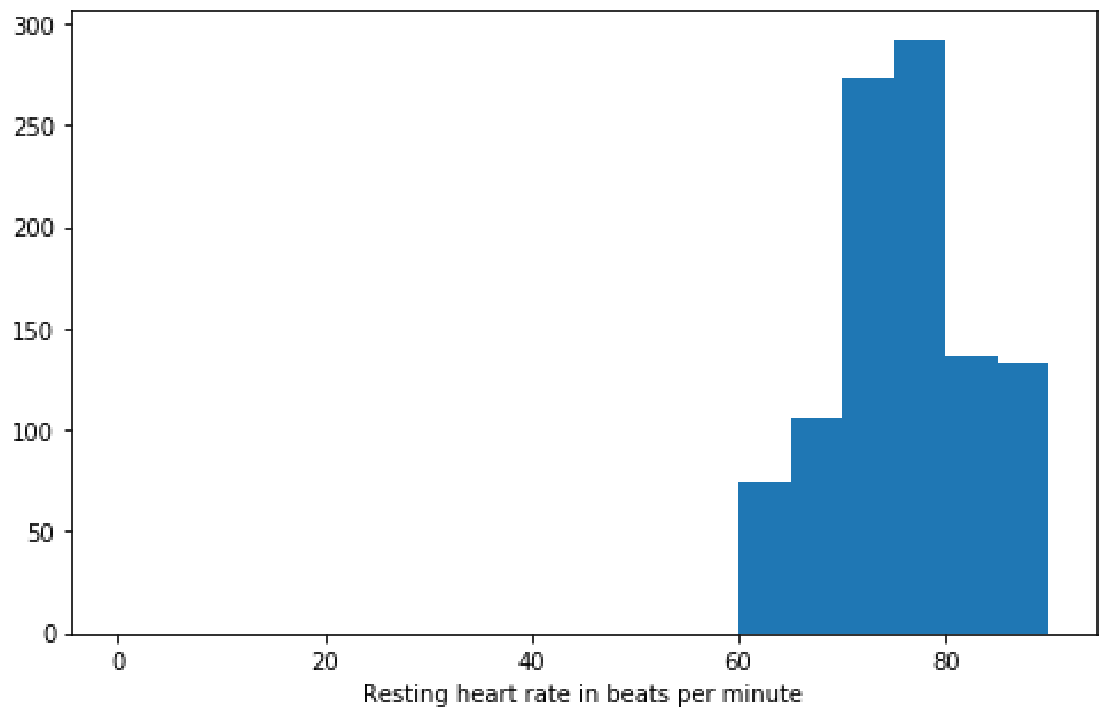
```
In [26]: #replacing the outliers 7 with 70

         data.loc[data.HeartRate == 7, 'HeartRate'] = 70
```

```
In [27]: #checking if the outliers were removed

         data.HeartRate.value_counts()
```

```
Out[27]: 70    273
         76    131
         80    117
         77     96
         66     87
         60     74
         88     59
```

```
86      55
78      46
90      19
82      19
75      19
67      12
65       5
68       2
Name: HeartRate, dtype: int64
```

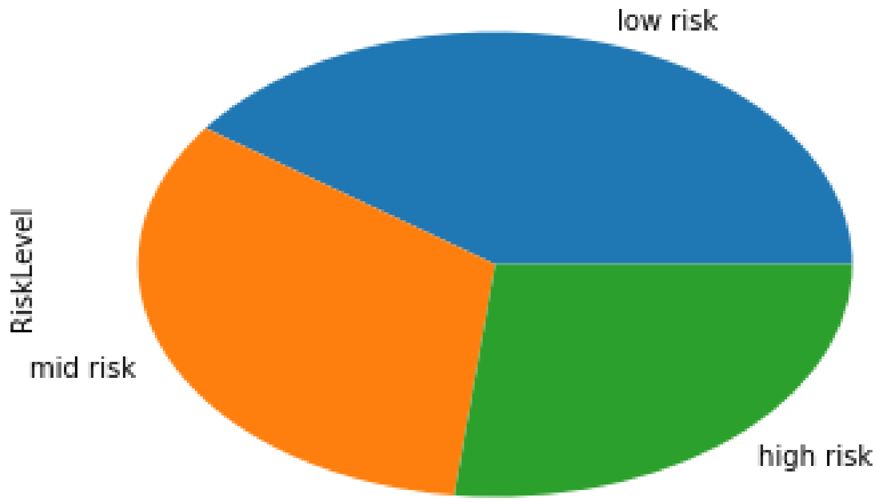
In [28]:
```
#Plotting the heart rate again

plot_hist(5, data, 'HeartRate', 'Resting heart rate in beats per minute')
```



Now we can see the distribution of the heart rates, ranging from 60 to 90. During pregnancy, the maternal heart rate tends to increase so that neither of the values from this data set seem to be indicators of a higher risk. See: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8439506/

In [29]:
```
#Plotting the risk levels.

data['RiskLevel'].value_counts().plot.pie()
```

Out[29]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7efc530a5d30>
```



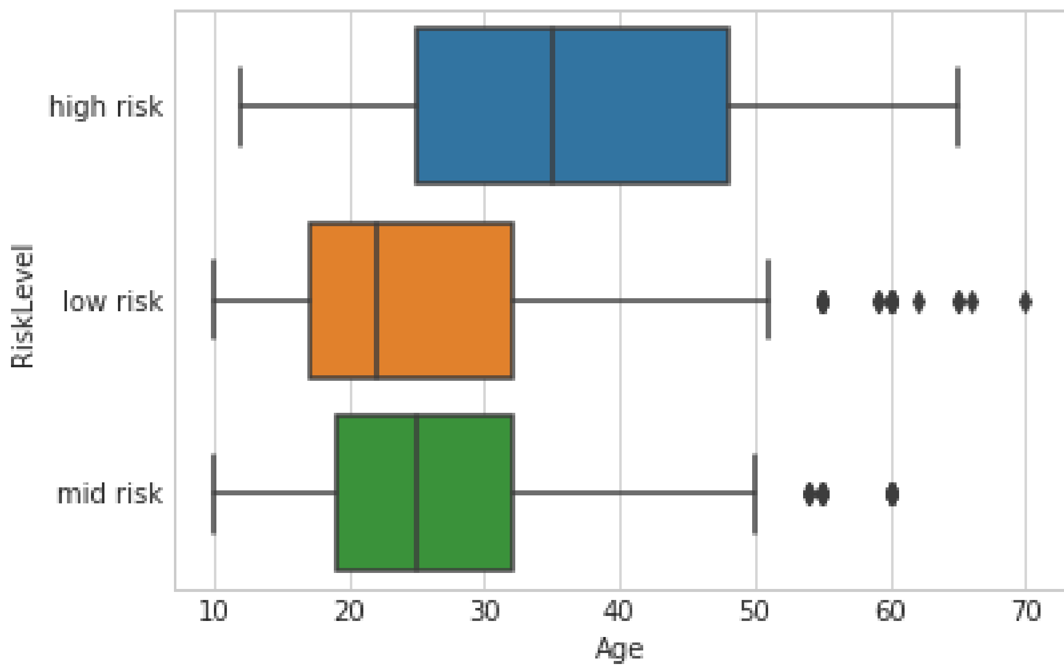
We can see that the majority of the pregnancies from our data set are categorized as low risk, however, there is also a significant number or mid and high risk cases. In the next part, the bivariate exploration, I will try to find correlations between the different measurements and the pregnancy risk levels.

1.5 Bivariate Exploration

The first relationship I would like to explore is between the maternal age and the pregnancy risk level.

```
In [31]: #Creating a boxplot to see if there is a correlation between the age and the risk level

         sb.set_style("whitegrid")
         sb.boxplot(x='Age', y='RiskLevel', data=data)
```

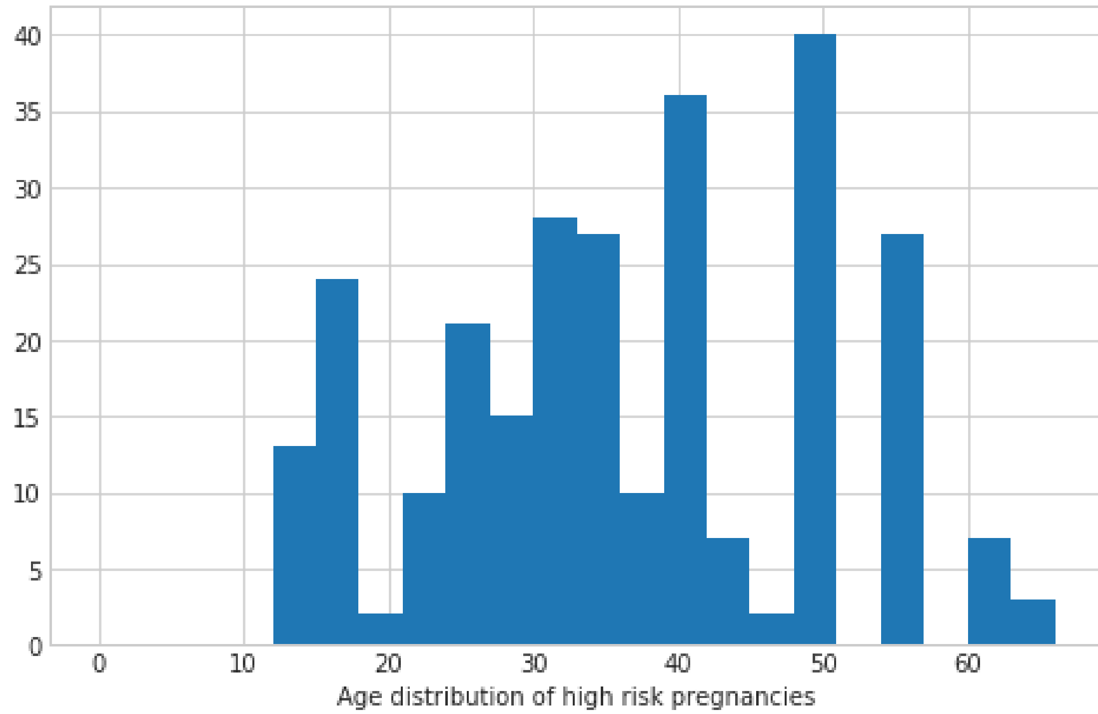
```
Out[31]: <matplotlib.axes._subplots.AxesSubplot at 0x7efc530df940>
```



A quick look at boxplot reveals that in this data set there seems to be a strong correlation between the maternal age and the risk level. The median age of the high risk pregnancies is in the mid 30s and the 3rd quartile is in the late 40s. Most women from the low and mid risk categories are between 20 and 30 years old, while the median age of low risk pregnancies is a bit smaller than the mid risk value. Surprisingly, there are a few outliers in the low risk group who are in their 60s and someone even 70 years old. There is a big gap between the ages of high risk and mid/low risk mothers, while the difference between mid and low risk is much smaller. It would be interesting to have a closer look at the age distribution in the different risk level groups.

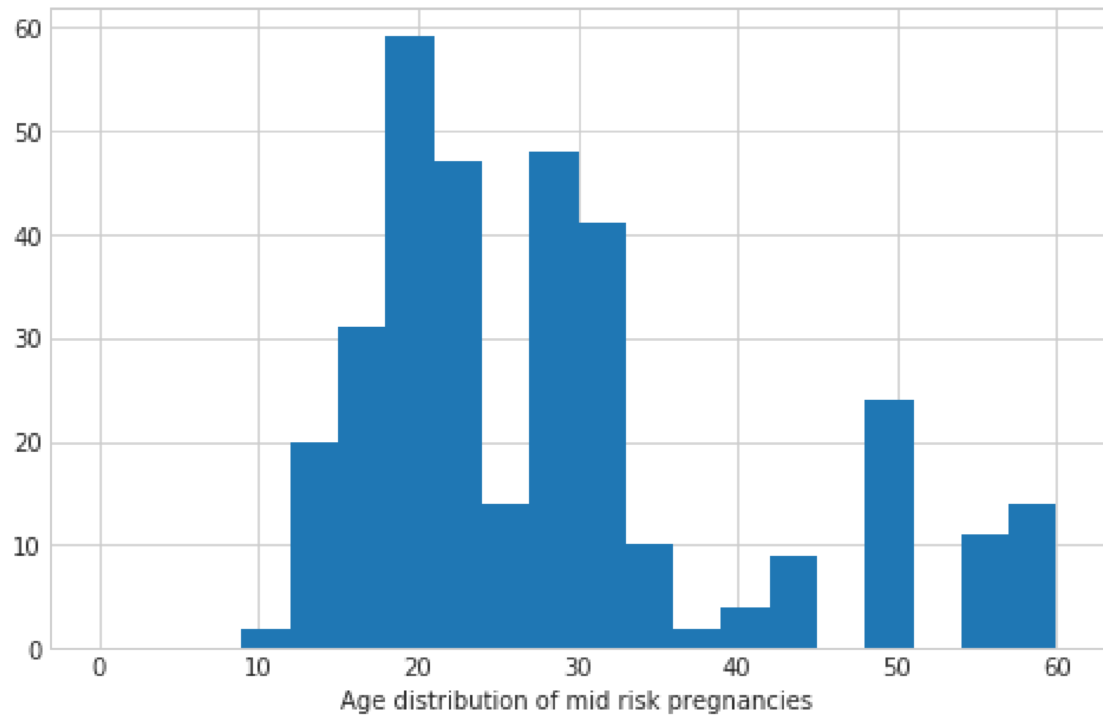
```
In [32]: #Extract the 2 relevant variables from the data set:

         data_age_risk = data[['Age','RiskLevel']]

In [33]: #Filter data by the risk levels for further analysis:

         data_age_high_risk = data_age_risk.loc[data_age_risk['RiskLevel']=='high risk']

In [34]: data_age_mid_risk = data_age_risk.loc[data_age_risk['RiskLevel']=='mid risk']

In [35]: data_age_low_risk = data_age_risk.loc[data_age_risk['RiskLevel']=='low risk']

In [36]: #Plotting the distribution of the ages of high risk pregnancies

         binsize = 3
         bins = np.arange(0, data_age_high_risk['Age'].max()+binsize, binsize)

         plt.figure(figsize=[8, 5])
```

```
plt.hist(data = data_age_high_risk, x = 'Age', bins = bins)
plt.xlabel('Age distribution of high risk pregnancies')
plt.show()
```



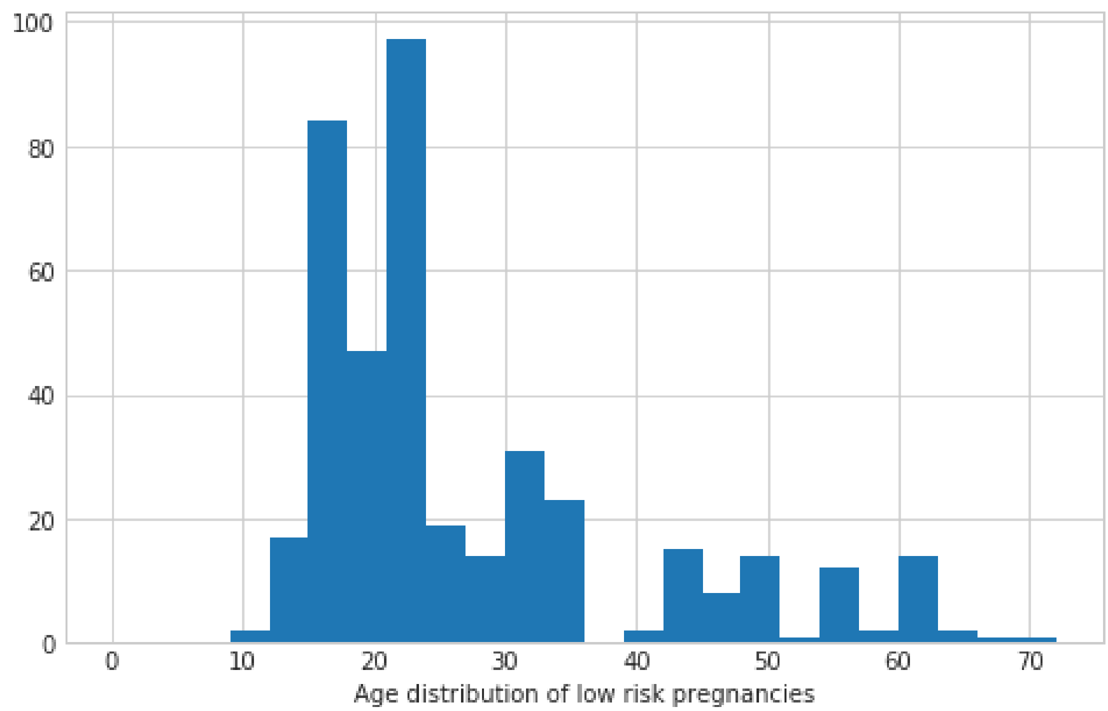
We see that the highest number of high risk pregnancies is in the age group of about 50 years, followed by women who are about 40. There is an unexpected peak in teenage high risk pregnancies, which could be related to specific health conditions (to be explored later) or an unhealthy lifestyle (e.g. smoking, alcohol). However, we do not have data about the lifestyle of the mothers, so we cannot draw any conclusions on this topic.

```
In [37]: binsize = 3
         bins = np.arange(0, data_age_mid_risk['Age'].max()+binsize, binsize)

         plt.figure(figsize=[8, 5])
         plt.hist(data = data_age_mid_risk, x = 'Age', bins = bins)
         plt.xlabel('Age distribution of mid risk pregnancies')
         plt.show()
```



In the mid risk category, most values are between 12 and 35 with 2 peaks - 1 is in the late teenage age group and 1 is in the late 20s. There are outliers in the groups of approximately 50 years and 55-60.

```
In [38]: #Plotting the distribution of the ages of low risk pregnancies

         binsize = 3
         bins = np.arange(0, data_age_low_risk['Age'].max()+binsize, binsize)

         plt.figure(figsize=[8, 5])
         plt.hist(data = data_age_low_risk, x = 'Age', bins = bins)
         plt.xlabel('Age distribution of low risk pregnancies')
         plt.show()
```

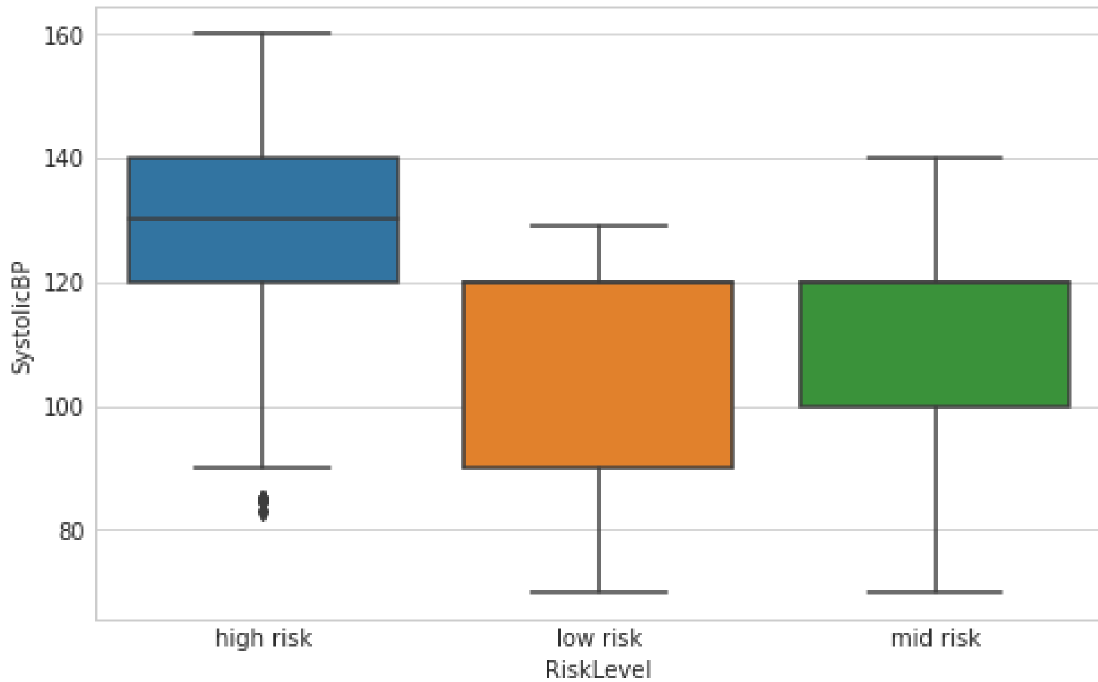


The low risk range covers all ages from 10 to 70, a rather surprising maximum value in the low risk category. Most women with low risk pregnancies, however, are in the age group between approximately 15 and 25.

The next variable of interest was the blood pressure, which is known for being an important indicator of risk levels during pregnancy.

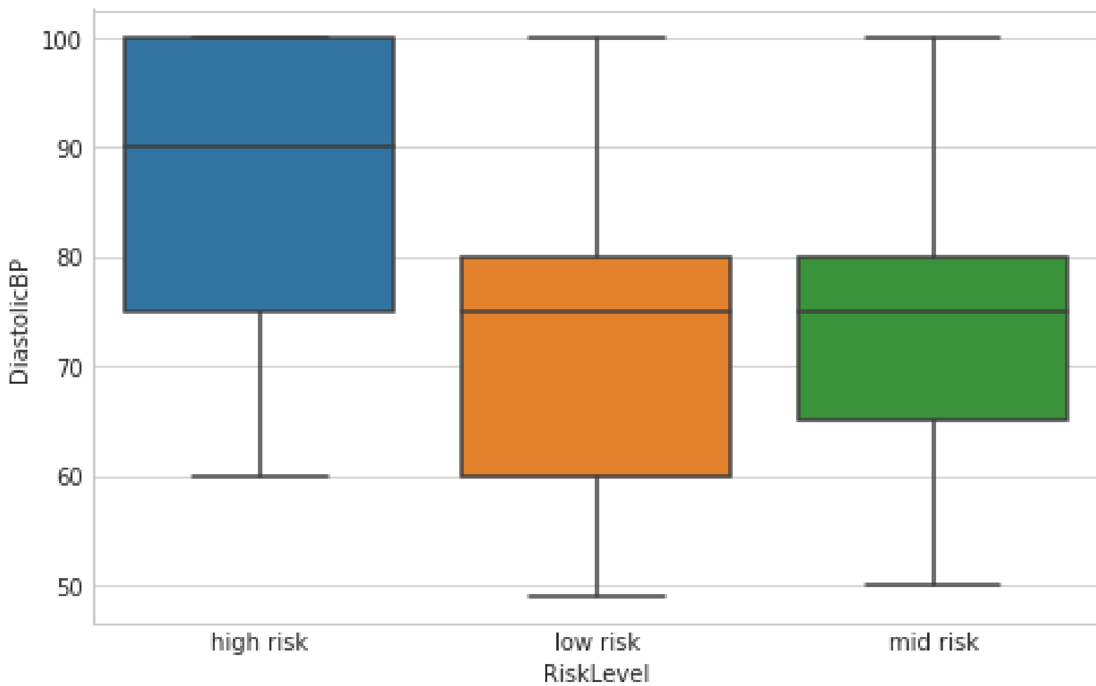
In [39]:
```
#Creating a box plot to see if there is a correlation between the systolic blood pressu

plt.figure(figsize=[8, 5])
sb.set_style("whitegrid")
sb.boxplot(x='RiskLevel', y='SystolicBP', data=data)
```

Out[39]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7efc533af6d8>
```



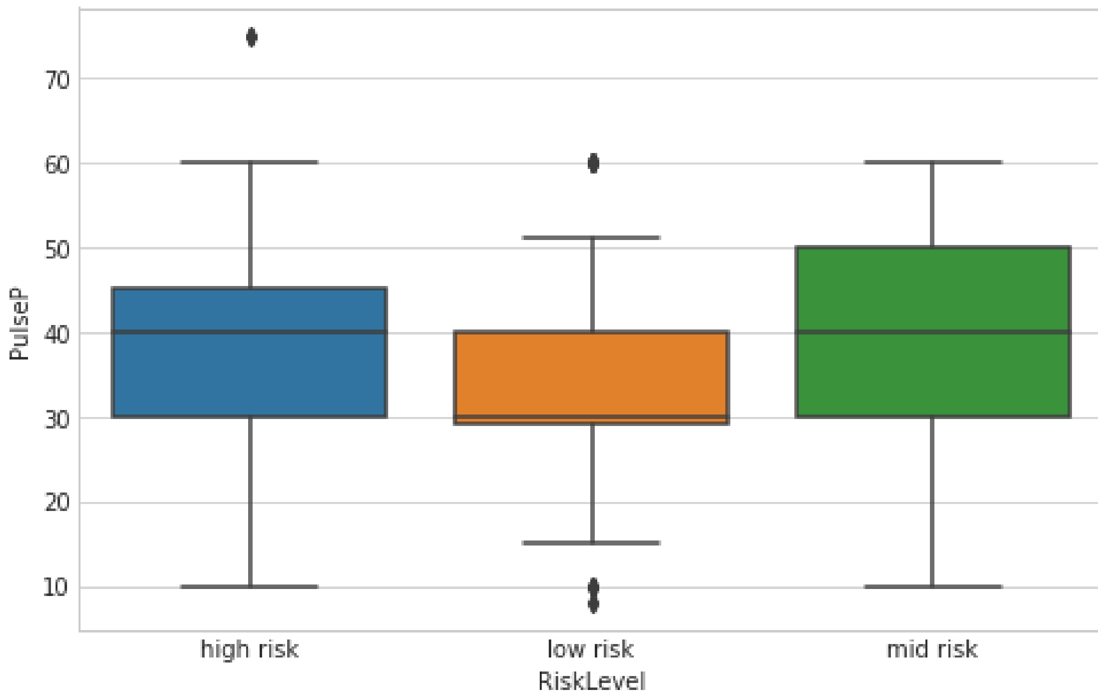
We can clearly observe that indeed, elevated systolic blood pressure is an indicator of high risk levels in pregnancy. Most values in this category are between 120 and 140, while the 3.quartiles in the mid and low risk categories are 120, which is the same value as the 1.quartile in the high risk group.

```
In [40]: #Creating a box plot to see if there is a correlation between the diastolic blood press

         plt.figure(figsize=[8, 5])
         sb.set_style("whitegrid")
         sb.boxplot(x='RiskLevel', y='DiastolicBP', data=data)

Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0x7efc806c6240>
```



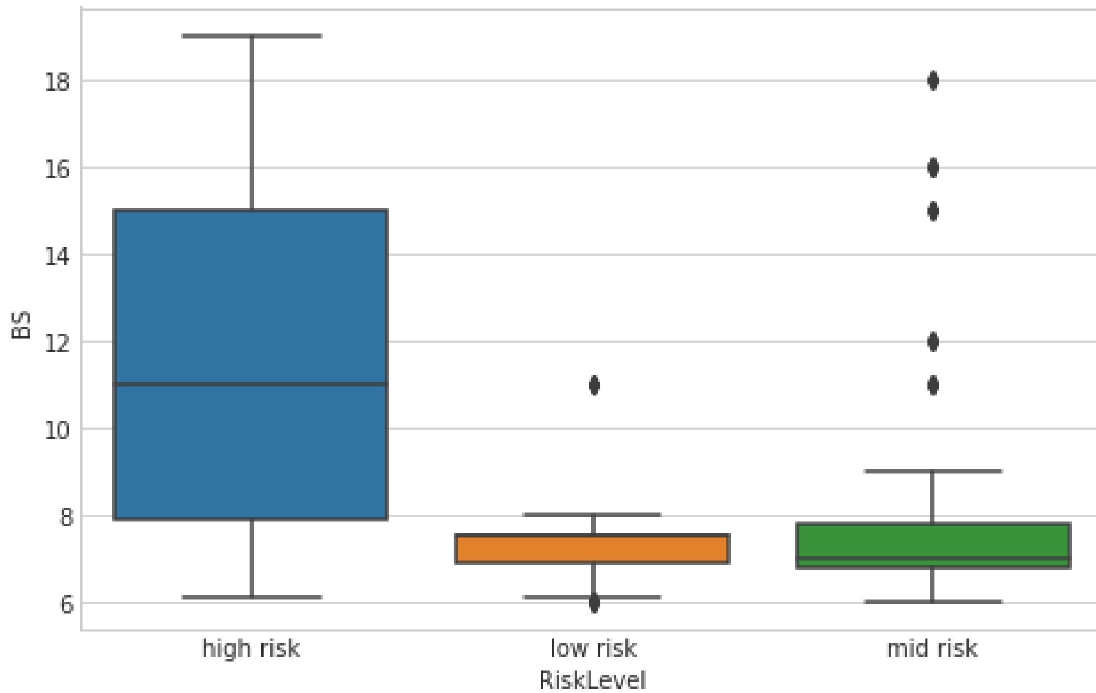
When looking at the diastolic blood pressure, we see a similar picture: most of the values in the high risk group are higher than in the low and mid risk groups. For instance, the median in the high risk is 90, while the median in the low and mid risk is about 75, which is the 1.quartile in the high risk category. There is a clear correlation between elevated diastolic blood pressure and a higher risk level.

```
In [41]: #Pulse pressure and risk level:

         plt.figure(figsize=[8, 5])
         sb.set_style("whitegrid")
         sb.boxplot(x='RiskLevel', y='PulseP', data=data)

Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x7efc53437780>
```



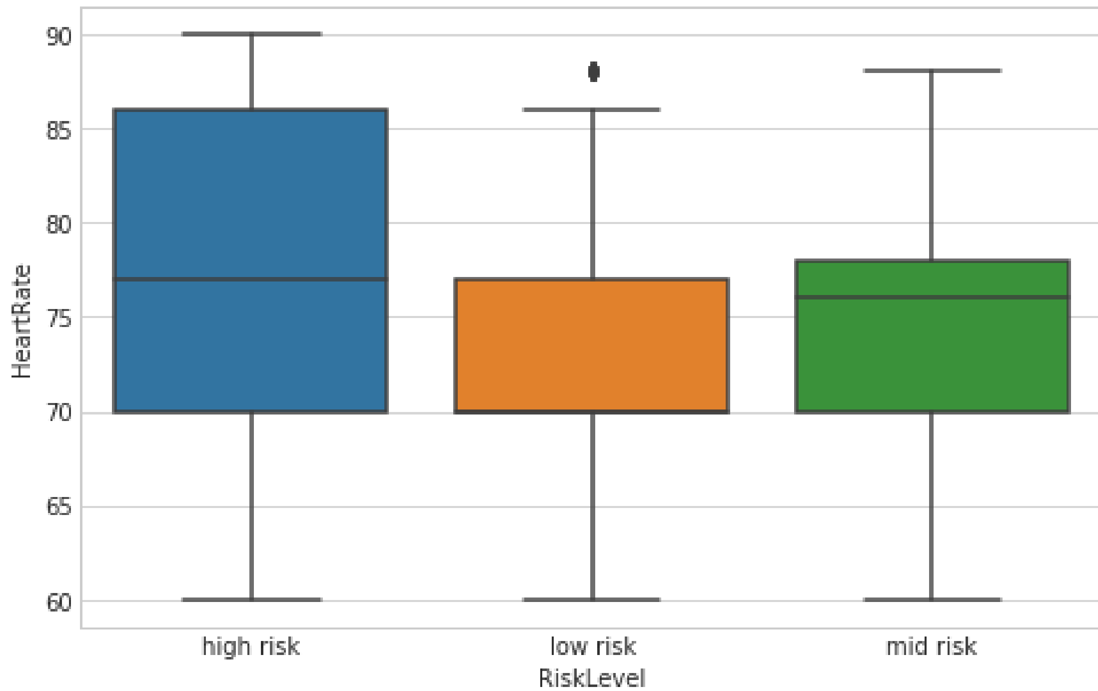
In the case of pulse pressure it is not as obvious as in the other 2 box plots: both the median of the high and the mid risk groups is 40, while the 3.quartile in the mid risk is 50, higher than the high risk group (about 45). Most low risk pulse pressure values are between 30 and 40. We can conclude that the pulse pressure may be an indicator of higher chances of a mid or high risk pregnancy, but the correlation does not seem as strong as in the cases of systolic and diastolic blood pressure levels.

```
In [42]: #Blood sugar and risk level:

         plt.figure(figsize=[8, 5])
         sb.set_style("whitegrid")
         sb.boxplot(x='RiskLevel', y='BS', data=data)

Out[42]: <matplotlib.axes._subplots.AxesSubplot at 0x7efc53449eb8>
```



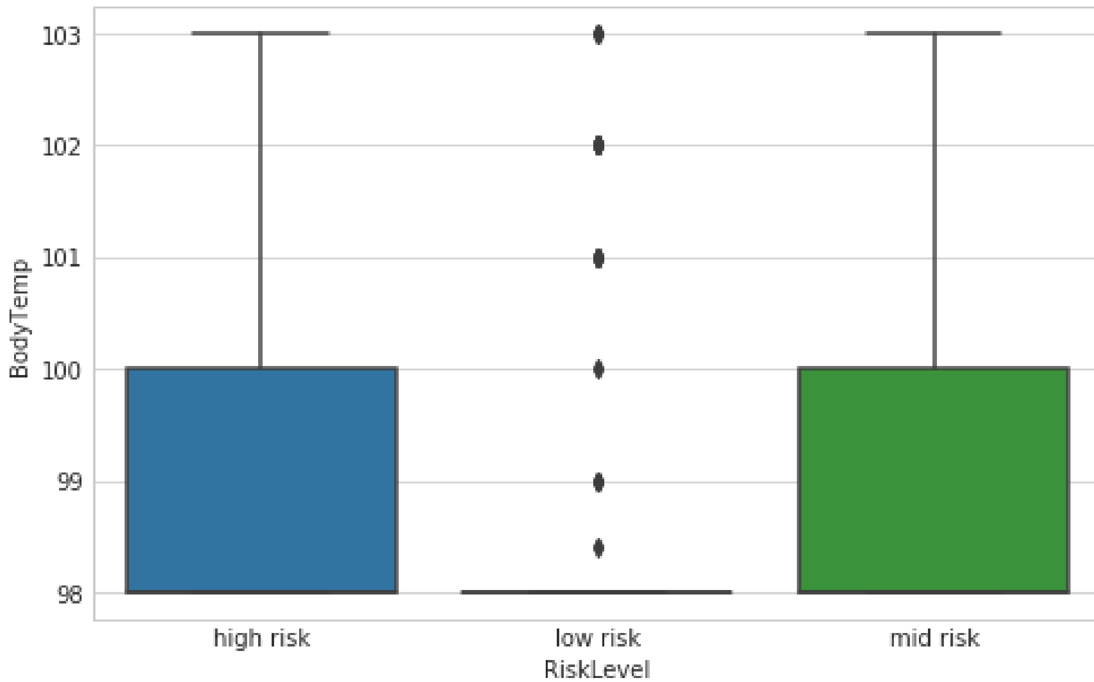
The blood sugar levels of most high risk mothers are much higher than in the mid and low risk groups: the median is 11 and the 3.quartile reaches 15. The 3.quartiles of the low and mid risk groups do not reach 8, which is the 1.quartile of the high risk group. Clearly, the blood sugar levels are a major indicator of pregnancy risk levels.

```
In [43]: #Heart rate:

         plt.figure(figsize=[8, 5])
         sb.set_style("whitegrid")
         sb.boxplot(x='RiskLevel', y='HeartRate', data=data)

Out[43]: <matplotlib.axes._subplots.AxesSubplot at 0x7efc53023ba8>
```



Even though we saw in the first part of the analysis that the heart rates reaching up to 90 beats per minute are not considered abnormal during pregnancy, we can observe a correlation between the risk level and the heart rate. The 3.quartile in the high risk group is about 86, while the 3.quartiles in the other groups are about 77/78, approximately as high as the high risk median.

```
In [44]: #Body temperature and risk level:

         plt.figure(figsize=[8, 5])
         sb.set_style("whitegrid")
         sb.boxplot(x='RiskLevel', y='BodyTemp', data=data)

Out[44]: <matplotlib.axes._subplots.AxesSubplot at 0x7efc53016048>
```



In the last part of the bivariate exploration we will focus on the correlation between the body temperature and the risk level. The box plot reveals an interesting picture: the high and the mid risk levels seem to have the same maximum temperatures (103) and the same quartiles. In the low risk group, the vast majority has a normal temperature level of 98 and only a few outliers showed higher values. We do not have any data on the causes of the fever, which can be ranging from a harmless cold to a dangerous bacterial or viral infection. We can only observe that elevated temperature can be an indicator of mid and high risk levels in pregnancy.

We have seen strong correlations between the risk levels, a categorical variable, and the different numeric variables such as the blood sugar, the age and the blood pressure levels. Now I would like to see if there is a correlation between the numeric variables in the high risk pregnancy group. In order to do that, we need to create a heatmap that represents the Pears correlation.

In [45]:
```
#Creating a data frame just for the high risk pregnancies:

data_high_risk = data.loc[data['RiskLevel']=='high risk']
```

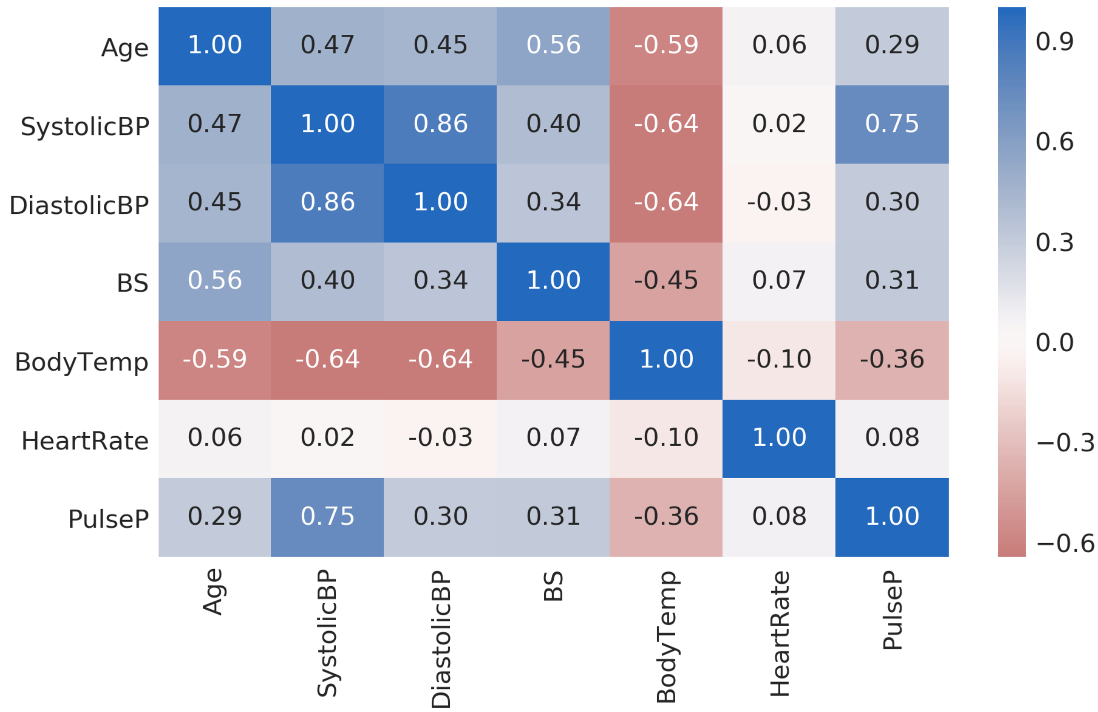
In [54]:
```
data_high_risk.head()
```

Out[54]:

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel	PulseP
0	25	130	80	15.00	98.0	86	high risk	50
1	35	140	90	13.00	98.0	70	high risk	50
2	29	90	70	8.00	100.0	80	high risk	20
3	30	140	85	7.00	98.0	70	high risk	55
5	23	140	80	7.01	98.0	70	high risk	60

In [85]:
```
# heatmap with Pears correlation
```

```
num_vars = ["Age", "SystolicBP", "DiastolicBP", "BS", "BodyTemp", "HeartRate", "PulseP"

plt.figure(figsize=[7, 4], dpi=300)
sb.heatmap(
    data_high_risk[num_vars].corr(), annot=True, fmt=".2f", cmap="vlag_r", center=0
);
```



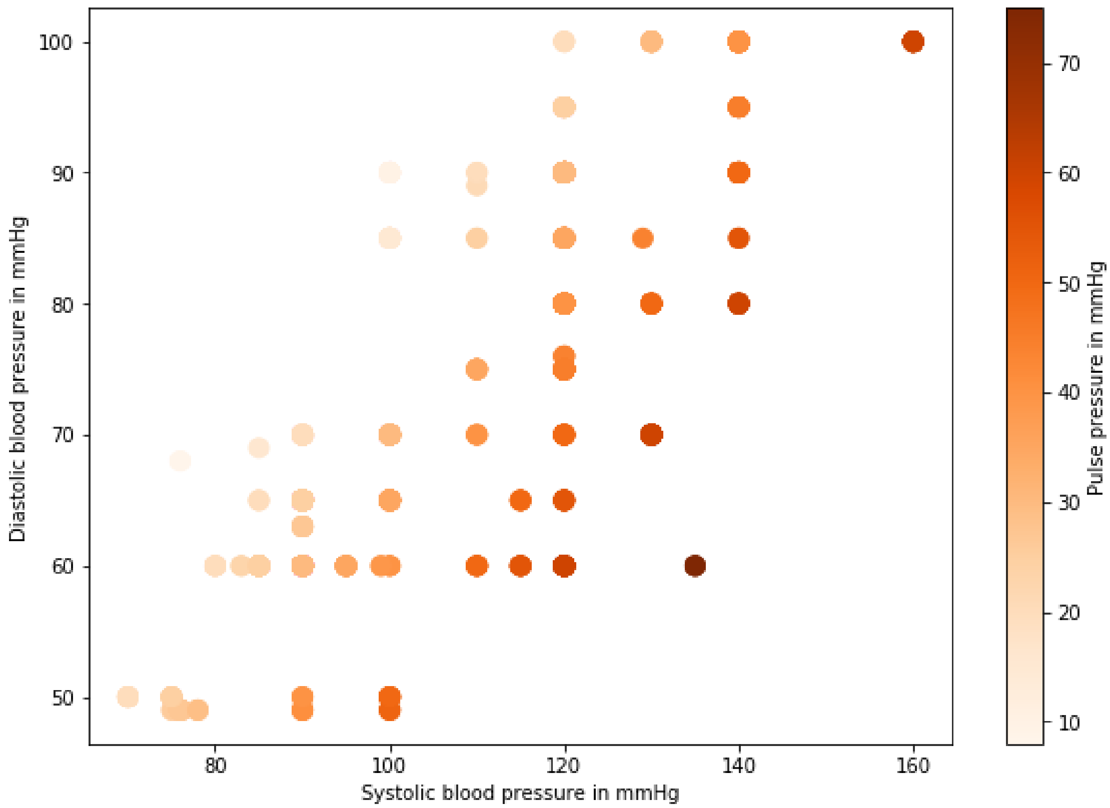
The heatmap reveals some interesting correlations: It seems that when the systolic blood pressure is higher, so is the diastolic blood pressure (with a correlation coefficient of 0.86) and the pulse pressure as well (with 0.75). There seems to be a relatively strong negative correlation between the body temperature, the systolic and the diastolic blood pressure levels (-0.64). It would be interesting to ask a medical professional if there is a real connection between these variables or it just happened to be so in this data set. The blood sugar levels seem to be correlated with the age (0.56), so are the blood pressure levels with a relatively high coefficient of 0.45/0.47. Interestingly, the heart rate does not seem to be dependent on any other variable.

1.6 Multivariate Exploration

Now I would like to plot the variables with the highest correlation coefficients: the systolic, diastolic and the pulse pressure levels.

```
In [30]: #Scatterplot of the systolic, diastolic and the pulse pressure levels:

         plt.figure(figsize=[10, 7])
```

```
x = data.SystolicBP
y = data.DiastolicBP
z = data.PulseP

plt.scatter(x, y, s=100, c=z, cmap="Oranges")
plt.xlabel('Systolic blood pressure in mmHg')
plt.ylabel('Diastolic blood pressure in mmHg')
plt.colorbar(label='Pulse pressure in mmHg')

plt.show()
```

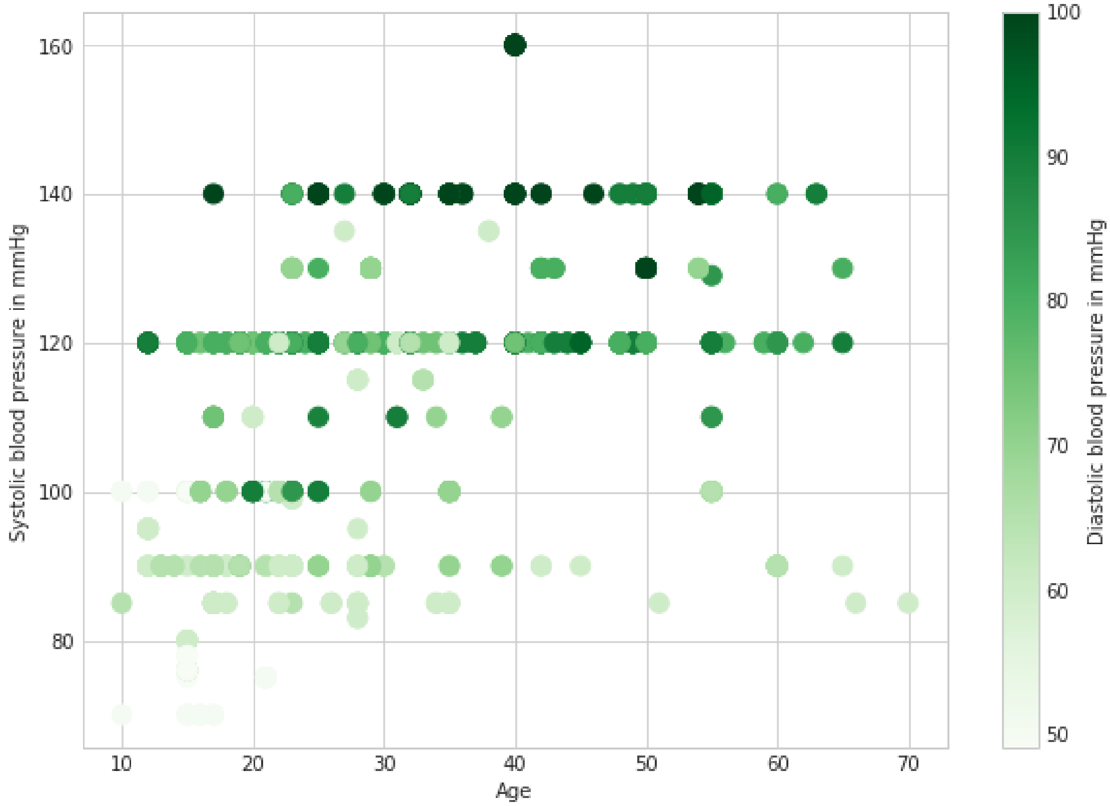


Here we have a visual representation of the correlation of the 3 variables. We can observe that as the systolic blood pressure (BP) levels grow, so do the diastolic levels. The orange colors represent the pulse pressure levels that become darker when the x values rise, so that most of the darker dots can be found at systolic BP levels of 120 and higher.

Let's see if there are more correlations between the different variables.

```
In [47]: #Scatterplot of the age, systolic and diastolic pressure levels:

         plt.figure(figsize=[10, 7])

         x = data.Age
```

```
y = data.SystolicBP
z = data.DiastolicBP

plt.scatter(x, y, s=100, c=z, cmap="Greens")
plt.xlabel('Age')
plt.ylabel('Systolic blood pressure in mmHg')
plt.colorbar(label='Diastolic blood pressure in mmHg')

plt.show()
```



There seems to be a moderate correlation between the age and the systolic and diastolic blood pressure levels. The darker points appear mainly after the age of approximately 25. Most of the darker green points are along the lines of 120 and 140, which shows that there is a strong correlation between the systolic and diastolic blood pressure levels and less with the age since they are distributed relatively evenly along the x axis.

```
In [48]: #Scatterplot of the age, systolic and diastolic pressure levels:

         plt.figure(figsize=[10, 7])

         x = data.Age
         y = data.BS
```

```python
z = data.HeartRate

plt.scatter(x, y, s=100, c=z, cmap="Purples")
plt.xlabel('Age')
plt.ylabel('Blood sugar')
plt.colorbar(label='HeartRate')

plt.show()
```

There seems to be a moderate correlation between the age and the blood sugar levels since most of the points located at the level of 16 and above apply to women in their 30s and older. However, the heart rates do not seem to correlate with the age.