Technical Annexe PWD CW2:

Phase 1 Function output:

2a Function output:

```
abnormaldict = abnormalSignAnalytics(mydict,50)
print(abnormaldict)

There are 50 records selected
Time taken to run function: 0.00032711029052734375 seconds
{'abnormal_pulse_count': 9, 'abnormal_values': [[87, 100], [596, 57], [332, 57], [479, 59], [531, 55], [815, 55], [94, 59], [211, 57], [258, 59]]}
```

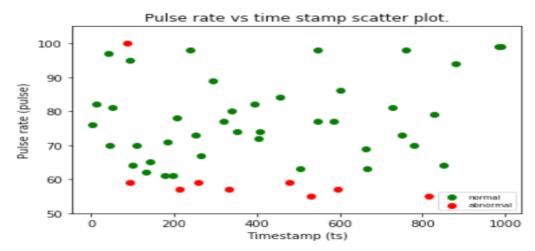
2c) 'abnormalSignAnalytics' algorithm utilises a random function to generate random numbers within range of the number of samples generated in phase 1. This chooses n number of time stamps to use as records. This is a non-deterministic approach which creates an assumption-free collection.

Also, a linear search is used which has a time complexity worst case scenario of O(n) which is fairly efficient. There are a few 'for loops' used but nothing more complicated (e.g. nested for loops) that would increase the time complexity. This also applies to 'frequencyAnalytics' and 'frequencyAnalytics2'. The results are stored in a dictionary format within 'mydict2'. More details of my code can be seen in the comments within the '.ipynb' file submitted.

Space complexity is very minimal as most of the data printed is stored in a dictionary. Many small lists are also used within the functions.

Although binary search would have produced a lower time complexity of O (log n), Logarithmic time, it would require a sort algorithm to achieve this as the keys in the dictionary from phase 1 are not ordered. If I were to implement the lowest time complexity sort algorithm explored in lectures (Merge sort or Timsort) they, at best, would provide a time complexity of O (n log n) and O(n) respectively. However, they would also produce a worst-case scenario of O (n log n) which is worse than my current time complexity. This also applies to 'frequencyAnalytics' and 'frequencyAnalytics'.

From the scatter plot generated below of function (2a), it is evident that there is no correlation between time stamps and pulse rate chosen randomly. Very few values (9 red values) are abnormal out of the 50 selected. Again, emphasizing a non-deterministic approach for selection of records.



As seen in the histogram for function 'frequencyAnalytics' and frequencyAnalytics2' (2b), there appears to be a random but slight normal distribution where majority of the pulse rates is between 60 and 99. The most common pulse rates in that sample seem to be between 75 and 80. There was only 1 record with pulse rate of 100. A total of 9 abnormal records and 41 normal records.

A random function is used to generate random numbers, within range of the number of samples generated in phase 1, to choose n number of time stamps to use as records, much like the previous function in 2a.

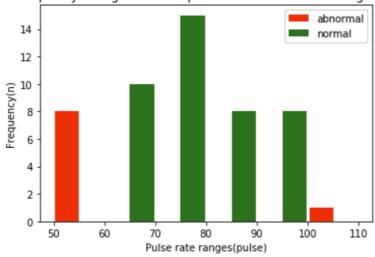
Space complexity is minimal as most of the data printed is stored in a dictionary. Many small lists are also used within the functions but there is the plotting and preparation ('for' loop) of the data for the histogram plot.

2b function output:

```
In [89]: frequencydict = frequencyAnalytics(mydict,50)
    print(frequencydict)
```

There are 50 samples selected.

Pulse frequency histogram of sample chosen from abnormalSignAnalytics.



```
Time taken to run function: 0.31235694885253906 seconds {100: 1, 72: 1, 76: 1, 95: 1, 57: 3, 82: 2, 64: 2, 94: 1, 63: 2, 59: 3, 55: 2, 61: 2, 77: 3, 73: 2, 71: 1, 65: 1, 62: 1, 74: 2, 67: 1, 70: 3, 80: 1, 78: 1, 81: 2, 99: 2, 98: 3, 86: 1, 79: 1, 89: 1, 69: 1, 97: 1, 84: 1}
```

However, when using 'frequencyAnalytics2', where the code to plot the histogram was taken out and inputted into a new function 'frequencyAnalyticsHisto', the time to run the algorithm greatly improves from 0.321 seconds to 0.00087 seconds. I believe this is due to the time saved not printing and not having to use the additional 'for loop' to separate pulse rates into normal and abnormal lists before plotting.

```
In [38]: frequencydict2 = frequencyAnalytics2(mydict,100)
    print(frequencydict2)

There are 100 samples selected.
    Time taken to run function: 0.0008740425109863281 seconds
    {100: 2, 72: 3, 76: 2, 95: 4, 57: 5, 82: 3, 64: 2, 94: 1, 63: 3, 59: 3, 5
    5: 3, 61: 3, 77: 4, 73: 2, 71: 2, 65: 3, 62: 1, 74: 2, 67: 1, 70: 6, 80:
    1, 78: 1, 81: 2, 99: 3, 98: 4, 86: 5, 79: 1, 89: 1, 69: 1, 97: 1, 84: 1,
    75: 3, 92: 3, 66: 1, 85: 2, 83: 3, 56: 2, 93: 4, 60: 1, 68: 2, 96: 1, 87:
    1, 90: 1}
```

3a function output:

```
In [111]: healthanalyser = healthAnalyzer(mydict, 56)
print(healthanalyser)

Time taken to run function: 0.0007848739624023438 seconds
[(1, 39, 61, 56, 121, 11, 93, 7.2), (75, 36, 83, 56, 120, 12, 99, 7.2), (126, 37, 90, 56, 121, 12, 98, 7.2), (202, 3 9, 80, 56, 121, 14, 99, 7.3), (240, 39, 92, 56, 121, 14, 100, 7.4), (310, 37, 84, 56, 121, 13, 96, 7.5), (337, 36, 7 0, 56, 121, 12, 93, 7.1), (457, 37, 100, 56, 121, 17, 99, 7.3), (514, 39, 76, 56, 121, 11, 93, 7.5), (548, 36, 79, 5 6, 121, 13, 95, 7.2), (551, 37, 98, 56, 120, 16, 96, 7.5), (575, 36, 78, 56, 121, 15, 93, 7.6), (597, 39, 77, 56, 12 1, 15, 94, 7.4), (626, 38, 86, 56, 121, 15, 98, 7.1), (673, 38, 76, 56, 121, 12, 95, 7.2), (700, 38, 89, 56, 120, 12, 95, 7.2), (740, 38, 83, 56, 120, 15, 97, 7.3), (756, 37, 84, 56, 120, 16, 98, 7.4), (786, 39, 70, 56, 120, 15, 93, 7.
4), (808, 37, 82, 56, 121, 17, 98, 7.4), (929, 38, 76, 56, 120, 16, 95, 7.2), (964, 38, 82, 56, 121, 14, 96, 7.4)]
```

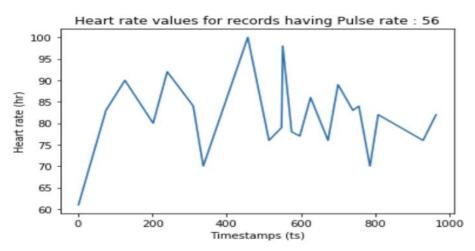
3b) The 'sorted()' function used in python, I believe is defaulted to the Timsort function so time complexity is worst-case O (n log n) and best-case is O (n), space complexity is O (n).

I also use a linear search algorithm which has a worst-case O (n) time complexity.

Overall, worst-case scenario is O (n log n) due to the Timsort function used.

However, if the outputted multidimensional list is not required to be in a sorted order then 'healthAnalyser' would only uses a linear search would have a lower complexity of O(n).

Below is the plotted heart rate values for records having pulse rate 56 which shows little correlation with timestamps.



Another solution, 'healthAnalyzer2', utilises a binary search in addition to the Timsort used in 'healthAnalyzer1'. This has a worst-case scenario of O (log n) which is better than Timsort. It also has a better O (1) space complexity. However, this algorithm's complexity is decided on the sorting of the list which utilises a 'for loop' so the complexity is O(n) overall.

4a function output:

```
print(benchmarking())
#nvalues = [1000,2500,5000, 7500, 10000]

[0.01545095443725586, 0.039411306381225586, 0.09513401985168457, 0.09292006492614746, 0.13114500045776367]
```

4b) Above are the benchmarks taken for 'myHealthcare' and the times are shown in seconds. There shows a positive correlation between time and 'n' number of operations. This function has a O (n) time complexity as there are only 'for loops' used to generate the samples. Data generated is stored in a dictionary.

Below, is the plot of the above results. It generally shows a steep linear increase as operations increase. There is however a small decline between 5000 and 7500 operations.

