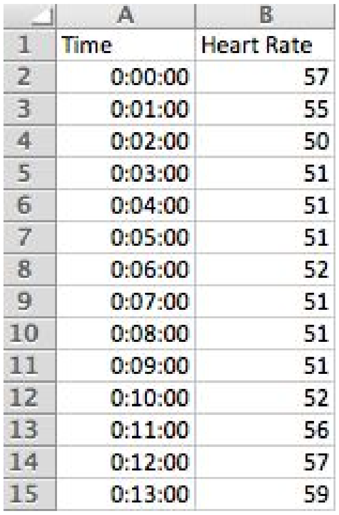
Capstone 1: In-Depth Analysis

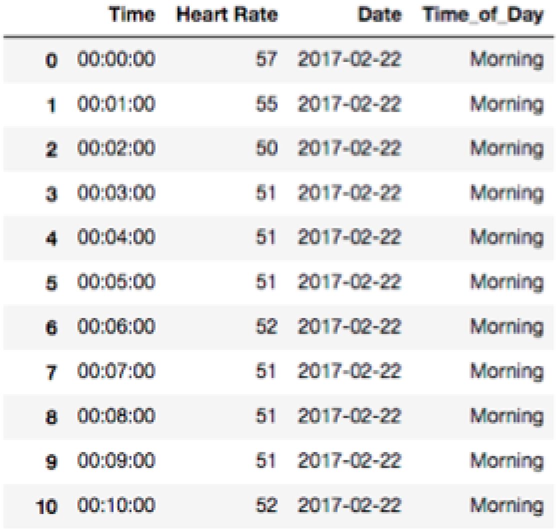
Data Preparation

Most, if not all, Fitbit devices are able to track a user's heart rate. For such a common feature, Fitbit surprisingly does not give its consumers the option to export that data from their account. Through an outside resource, I was able to gather Tracy's, the subject and Fitbit user, intraday heart rate data (by minute). With this newly acquired data, I immediately thought of the different possibilities of using her heart rate to predict features such as how many calories she will burn in a day.

The data comes in the format of a CSV file, and contains simply two columns of data--a column for minutes and another for her heart rate at the time. To be able to make this data more useful, a few features were added (e.g., date, time of day). Since the Fitbit devices are constantly tracking your heart rate, there are no instances of missing data.

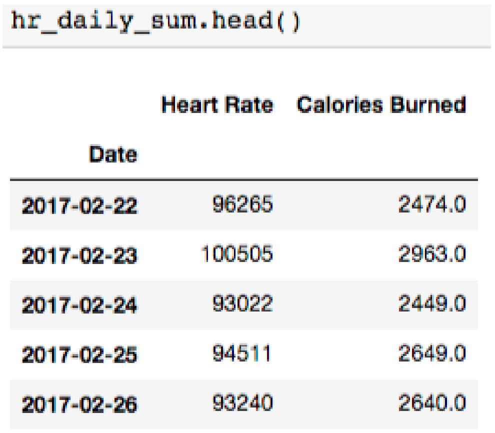


	A	B
1	Time	Heart Rate
2	0:00:00	57
3	0:01:00	55
4	0:02:00	50
5	0:03:00	51
6	0:04:00	51
7	0:05:00	51
8	0:06:00	52
9	0:07:00	51
10	0:08:00	51
11	0:09:00	51
12	0:10:00	52
13	0:11:00	56
14	0:12:00	57
15	0:13:00	59



	Time	Heart Rate	Date	Time_of_Day
0	00:00:00	57	2017-02-22	Morning
1	00:01:00	55	2017-02-22	Morning
2	00:02:00	50	2017-02-22	Morning
3	00:03:00	51	2017-02-22	Morning
4	00:04:00	51	2017-02-22	Morning
5	00:05:00	51	2017-02-22	Morning
6	00:06:00	52	2017-02-22	Morning
7	00:07:00	51	2017-02-22	Morning
8	00:08:00	51	2017-02-22	Morning
9	00:09:00	51	2017-02-22	Morning
10	00:10:00	52	2017-02-22	Morning

As mentioned above, the idea of predicting calories burned for a day using heart rate data was tested. In a typical predictive model, several feature variables are gathered to predict a target variable, all the variables belonging to an instance. The first problem that arose was the structure of the data. The idea is to use one day's worth of heart rate data to predict how many calories was burned, which in terms of the data means using all values of one column to predict a single row value. As a workaround, heart rates were aggregated by summation, and a new dataframe containing the sum of heart rates by day, calories burned, and date was created.



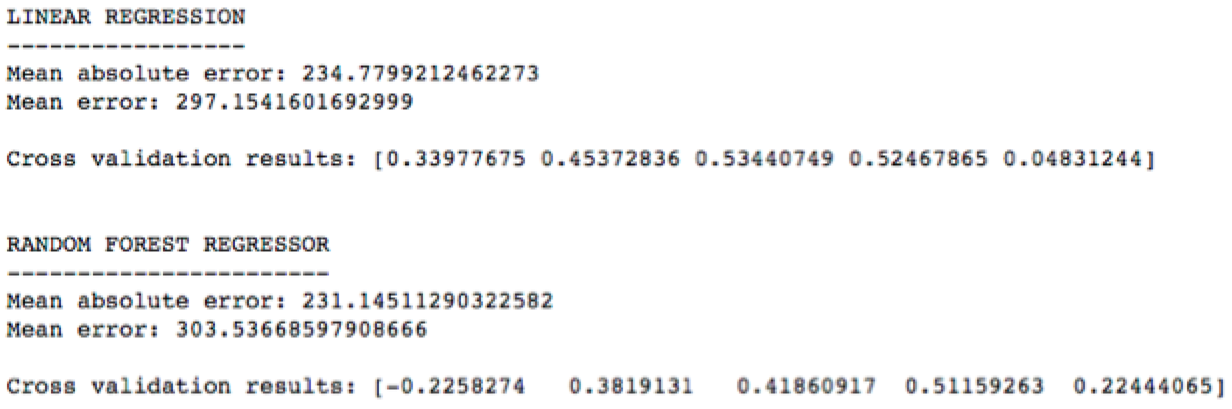
```
hr_daily_sum.head()
```

	Heart Rate	Calories Burned
Date		
2017-02-22	96265	2474.0
2017-02-23	100505	2963.0
2017-02-24	93022	2449.0
2017-02-25	94511	2649.0
2017-02-26	93240	2640.0

Supervised Learning

With the new dataframe, a model was tested using only the ‘Heart Rate’ feature to predict ‘Calories Burned’. Both a linear regression and random forest regressor were used. Increased heart rate burns more calories, so intuitively, the greater the sum of heart rates, the more calories are burned. This suggests that there is probably a somewhat linear relationship between sum of heart rates and calories burned, which is why a Linear Regression model was chosen. However, the relationship between the two variables might not be completely linear, so a Random Forest Regressor is a good alternative to test out.

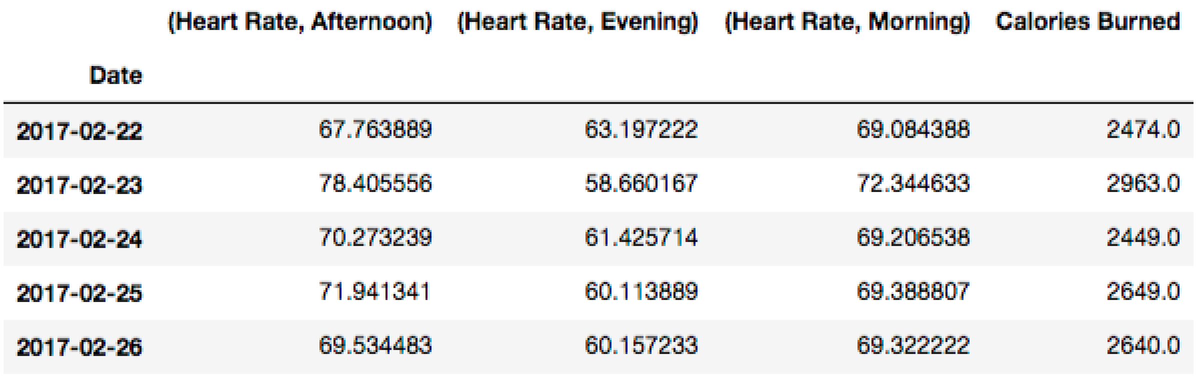
The datafame containing heart rate sums and calories burned is split into training and test sets (70-30 ratio), where the training set is used to train the model and test set to evaluate the performance of the model. To evaluate the performance, both the mean absolute error and mean error (square root of mean squared error) were calculated. The following are the results of the predictive models:



```
LINEAR REGRESSION
-----------------
Mean absolute error: 234.7799212462273
Mean error: 297.1541601692999

Cross validation results: [0.33977675 0.45372836 0.53440749 0.52467865 0.04831244]


RANDOM FOREST REGRESSOR
-----------------------
Mean absolute error: 231.14511290322582
Mean error: 303.53668597908666

Cross validation results: [-0.2258274   0.3819131   0.41860917  0.51159263  0.22444065]
```

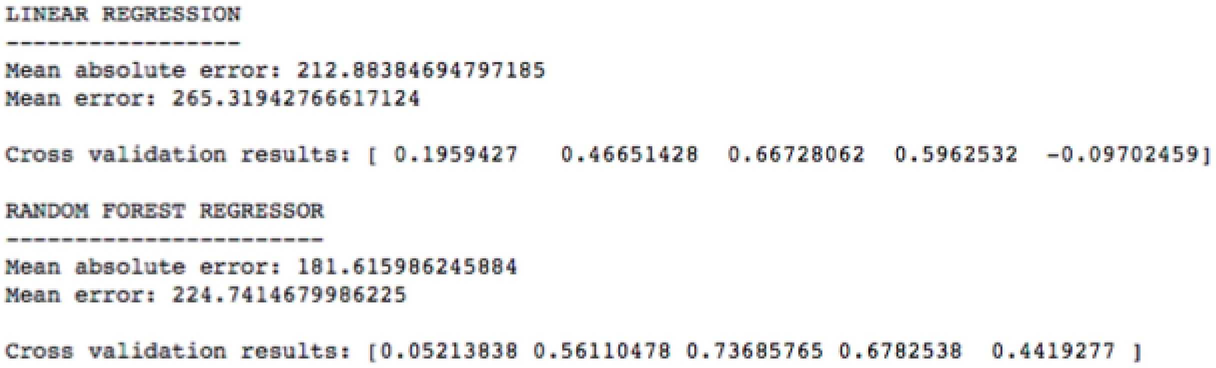
The mean absolute errors suggests that on average, the predictive models will be roughly 230 calories off. The error is quite substantial--an individual would have to spend 15-20 minutes jump roping or 30 minutes jogging to burn off roughly 200 calories. The models are, to a certain degree, underfitting the training data.

While still utilizing just the heart rate data, new features were created in an attempt to create better models. Heart rate data were split by time of day (morning, afternoon, and evening) and averaged. Now three predictor variables are used to predict amount of calories burned.



	(Heart Rate, Afternoon)	(Heart Rate, Evening)	(Heart Rate, Morning)	Calories Burned
Date				
2017-02-22	67.763889	63.197222	69.084388	2474.0
2017-02-23	78.405556	58.660167	72.344633	2963.0
2017-02-24	70.273239	61.425714	69.206538	2449.0
2017-02-25	71.941341	60.113889	69.388807	2649.0
2017-02-26	69.534483	60.157233	69.322222	2640.0

Both Linear Regression and Random Forest Regression models were used again to compare performance against the first two models.

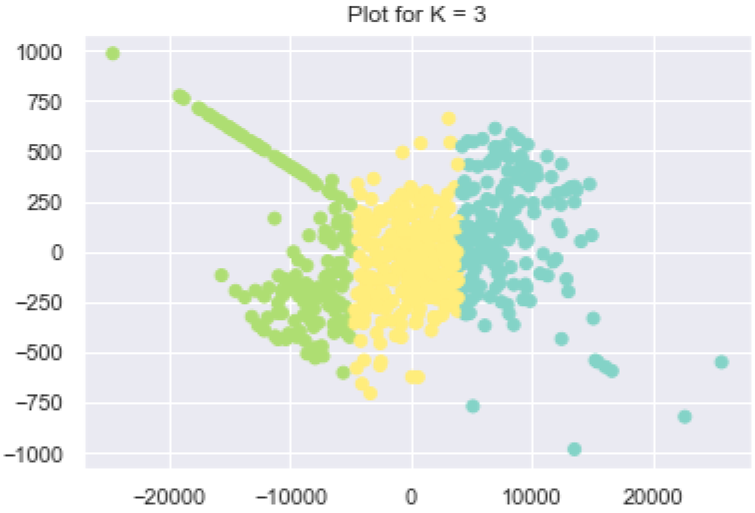


```
LINEAR REGRESSION
-----------------
Mean absolute error: 212.88384694797185
Mean error: 265.31942766617124

Cross validation results: [ 0.1959427   0.46651428  0.66728062  0.5962532  -0.09702459]

RANDOM FOREST REGRESSOR
-----------------------
Mean absolute error: 181.615986245884
Mean error: 224.7414679986225

Cross validation results: [0.05213838 0.56110478 0.73685765 0.6782538  0.4419277 ]
```

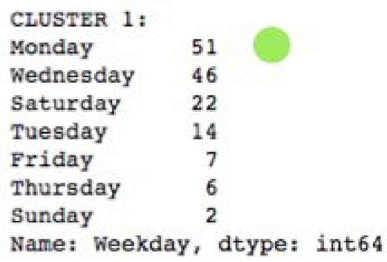
With more features, we can see an improvement in predictive power. The new Random Forest Regressor model will have an average (absolute) predictive error of less than 200 calories.

Unsupervised Learning

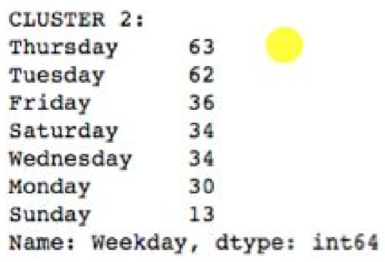
In addition to supervised learning, unsupervised learning, specifically K-Means Clustering, was also utilized on the Fitbit data. Using the elbow method, it was determined the optimal number of clusters is 3, which yields the plot below:



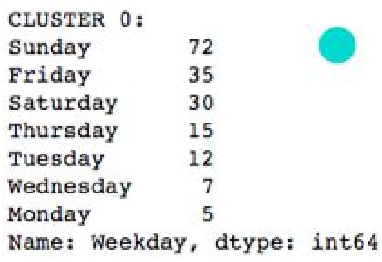
Analyzing deeper into each cluster, many variables were investigated, but one of the more interesting ones to look at is ‘Weekday’.



```
CLUSTER 1:
Monday       51
Wednesday    46
Saturday     22
Tuesday      14
Friday        7
Thursday      6
Sunday        2
Name: Weekday, dtype: int64
```

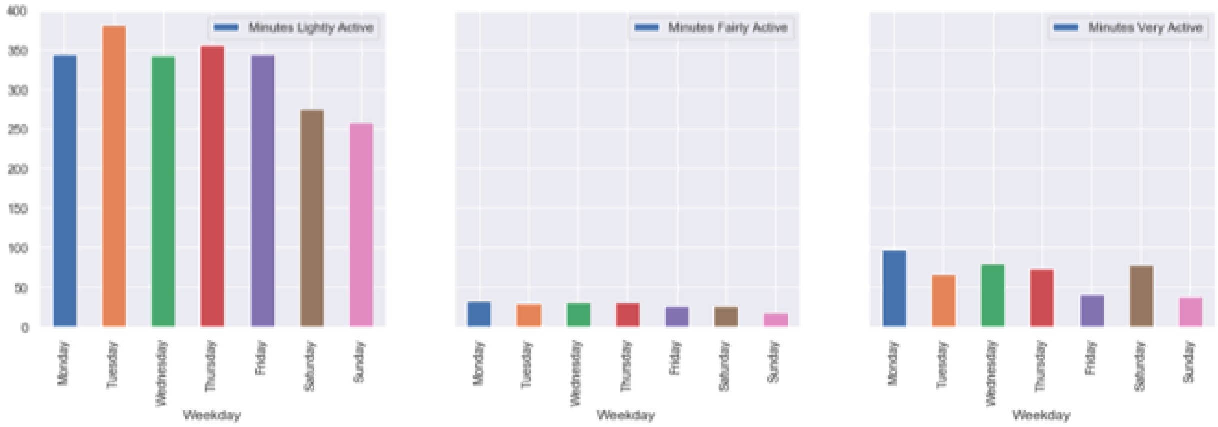


```
CLUSTER 2:
Thursday     63
Tuesday      62
Friday       36
Saturday     34
Wednesday    34
Monday       30
Sunday       13
Name: Weekday, dtype: int64
```

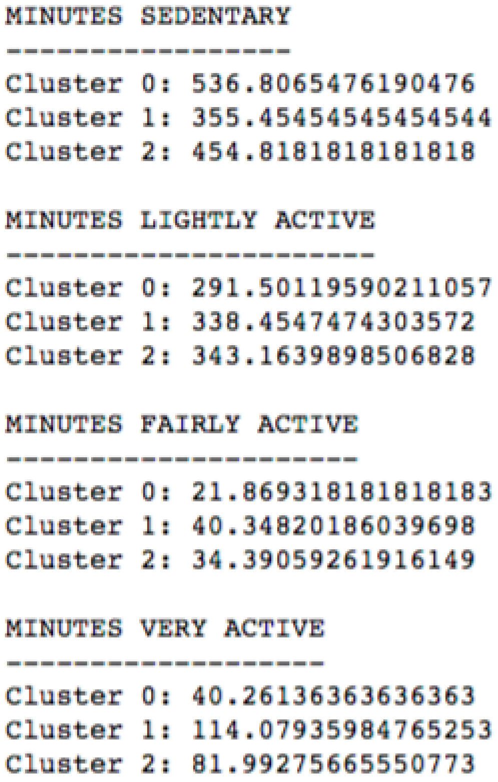


```
CLUSTER 0:
Sunday       72
Friday       35
Saturday     30
Thursday     15
Tuesday      12
Wednesday     7
Monday        5
Name: Weekday, dtype: int64
```

Cluster 0 has many occurrences of Sunday’s, Cluster 1 has Monday’s and Wednesday’s, and Cluster 2 has Tuesday’s and Thursday’s. Having made graphs pertaining to the day of week, something peculiar is noticed.



Looking at the graph depicting ‘Minutes Very Active’, Monday’s and Wednesday’s have the highest mean values, which in turn means those two days are days where Tracy is most active. The assumption is confirmed by Tracy herself--she has, for an extended period of time, taught at least two Insanity classes on both days. Sunday’s, overall, are Tracy’s most laid back days, ranking last in all three activity level categories except ‘Minutes Sedentary’. Tuesday’s and Thursday’s have highest mean values in the ‘Minutes Lightly Active’ category, and relatively high mean values for ‘Minutes Very Active’. This could be an indication that she spends more time at her home health aide profession and spends less time at the gym. To analyze the activity data by clusters, they first needed to be merged into the dataframe.



```
MINUTES SEDENTARY
-----------------
Cluster 0: 536.8065476190476
Cluster 1: 355.45454545454544
Cluster 2: 454.8181818181818

MINUTES LIGHTLY ACTIVE
----------------------
Cluster 0: 291.50119590211057
Cluster 1: 338.4547474303572
Cluster 2: 343.1639898506828

MINUTES FAIRLY ACTIVE
---------------------
Cluster 0: 21.869318181818183
Cluster 1: 40.34820186039698
Cluster 2: 34.39059261916149

MINUTES VERY ACTIVE
-------------------
Cluster 0: 40.26136363636363
Cluster 1: 114.07935984765253
Cluster 2: 81.99275665550773
```

Looking at the ‘Minutes Sedentary’ category, it seems that data points in Cluster 0 are representative of Tracy’s laid back/rest days--cluster 0 has the highest mean value under the ‘Minutes Sedentary’ category. Cluster 1 contains many occurrences of Monday’s and Wednesday’s, and looking at the ‘Minutes Very Active’ chart, Cluster 1 is in fact representative of Tracy’s workout-intensive days.

Overall, there were some limitations with how well the predictive models can perform due to lack of independence between some variables (data pertaining to activity are, to a degree, all dependent of one another). Breaking up the heart rate data by time of day and aggregating them by average increased the performance of both the Linear Regression and Random Forest

Regression models. By performing unsupervised learning, we are able to determine which days Tracy teaches/trains and which days she uses as rest days.