Crowdedness at the Campus Gym

Group # 5

Cameron Milligan

Mariuxi Leon

Jorge Gonzalez

Abstract

A study was made in the gymnasium of the University Campus, if the gym is crowded then it is not possible to exercise but if the gym is not so full it is the right time.

The measurement focuses on the number of people who attend the gym one time every 10 minutes over the last year.

A prediction model will be developed and the variables that compose it will be selected. Accuracy is the factor that will be taken into account when selecting the corresponding model. The prediction will be to determine how crowded the gym will be in the future.

Crowdedness at the Campus Gym

1. Objectives

1.1. Given a time of day (and maybe some other features, including weather), predict how crowded the gym will be.

1.2. Figure out which features are actually important, which are redundant, and what features could be added to make the predictions more accurate.

2. Participants

Participants are all those who attend the University Campus gym.

3. Assessments and Measures

It will be made:

- Graphs that allow visualizing the behavior of the data, analysis of the measures of central tendency, etc.
- Proposal of a prediction model.

4. Data

The data that we selected for the analysis is a "**Crowdedness at the Campus Gym**" .

https://www.kaggle.com/nsrose7224/crowdedness-at-the-campus-gym

The dataset consists of 26,000 people counts (about every 10 minutes) over the last year.

The 11 columns of the data set are the following: Number of people, date(datetime of data), timestamp(number of seconds since beginning of day), day_of_week (Monday to

Sunday), is_weekend (Saturday or Sunday), is_holiday (i.e federal holiday) , temperature (degrees fahrenheit), Is_start_of_semester, month (1 january - 12 december), hour (0-23). This data was collected with the consent of the university and the gym in question.

4.1. Data Preparation:

The data preparation includes:

- Load the Crowdedness at the Campus Gym data into the Pandas data frames.
- Display only the first row of the data frame to take a look at the data.

```
[ ] df.head(1)
```

	number_people	date	timestamp	day_of_week	is_weekend	is_holiday	temperature	is_start_of_semester	is_during_semester	month	hour
0	37	2015-08-14 17:00:11-07:00	61211	4	0	0	71.76	0	0	8	17

- Show some basic statistical details like percentile, mean, standard deviation etc. of the data frame.

```
[ ] df.describe()
```

	number_people	timestamp	day_of_week	is_weekend	is_holiday	temperature	is_start_of_semester	is_during_semester	month	hour
count	62184.000000	62184.000000	62184.000000	62184.000000	62184.000000	62184.000000	62184.000000	62184.000000	62184.000000	62184.000000
mean	29.072543	45799.437958	2.982504	0.282870	0.002573	58.557108	0.078831	0.660218	7.439824	12.236460
std	22.689026	24211.275891	1.996825	0.450398	0.050660	6.316396	0.269476	0.473639	3.445069	6.717631
min	0.000000	0.000000	0.000000	0.000000	0.000000	38.140000	0.000000	0.000000	1.000000	0.000000
25%	9.000000	26624.000000	1.000000	0.000000	0.000000	55.000000	0.000000	0.000000	5.000000	7.000000
50%	28.000000	46522.500000	3.000000	0.000000	0.000000	58.340000	0.000000	1.000000	8.000000	12.000000
75%	43.000000	66612.000000	5.000000	1.000000	0.000000	62.280000	0.000000	1.000000	10.000000	18.000000
max	145.000000	86399.000000	6.000000	1.000000	1.000000	87.170000	1.000000	1.000000	12.000000	23.000000

4.2. Clean Data:

The dataset is clean because:

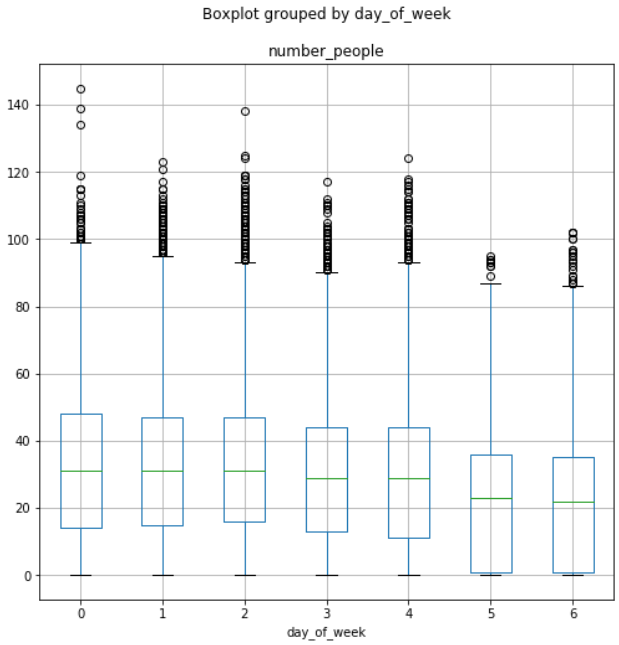
- It can visualize that all the columns (attributes) have the same number of rows or registers, therefore there are no missing values.
- We can observe in the results, that the maximum number of people is 145 and minimum is 0 .
- Maximum timestampvalue is 86399 which is close to 24 hours which is reasonable
- Maximum day of the week is 6 and minimum is 0.

5. Analysis or Models

Using this dataset interesting insights into attendance of the gym shall be derived. Also a predictive model shall be developed which will predict the number of people attending the gym given the values of other factors.

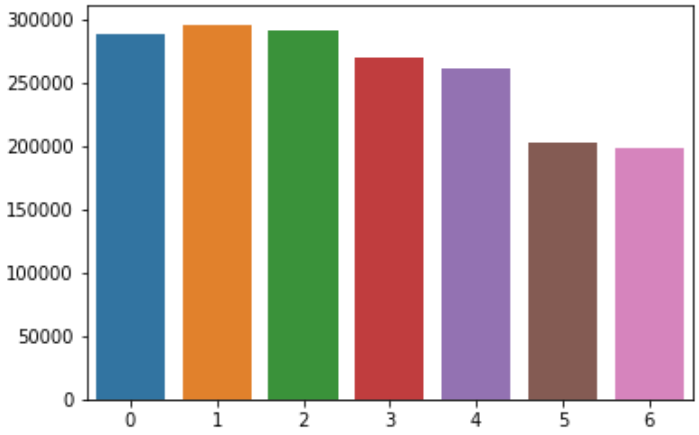
Next, different graphs will be made that will allow us to analyze the data from different points of view:

Boxplot Grouped by day_of_week



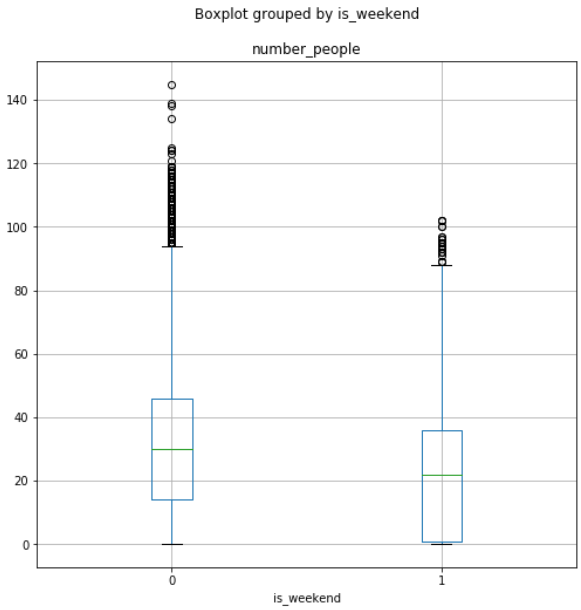
In this graph, it can be visualized the number of people that attend the gym in the different days of the week:

- The largest attendance at the gym are: Monday, Tuesday and Wednesday. On Monday, the 25% (Quartile 1) have values less than or equal to 18, the 50% (Quartile 2 or Median) have values less than or equal to 30 and 75% (Quartile 3) have values less than or equal to 50.
- The weekends (Saturday and Sunday) are less popular. If we analyze on Saturday, the 25% (Quartile 1) have values less than or equal to 1, the 50% (Quartile 2 or Median) have values less than or equal to 25 and the 75% (Quartile 3) have values less than or equal to 35.



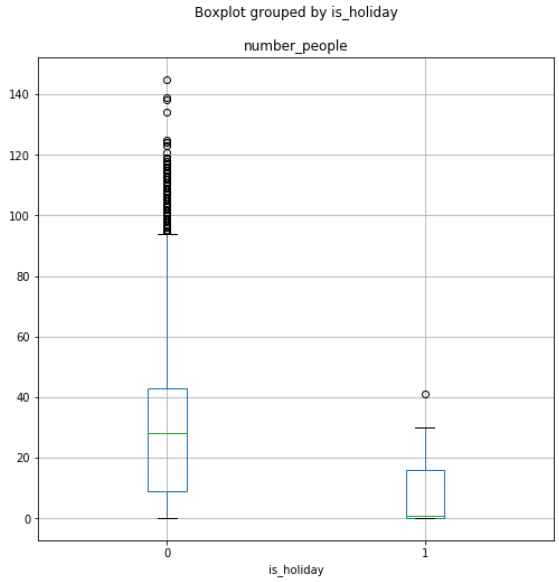
Another way to visualize it is through a bar chart. It shows that early during week more number of people are going to the gym and it goes steadily down until sunday.

Boxplot Grouped by is_weekend



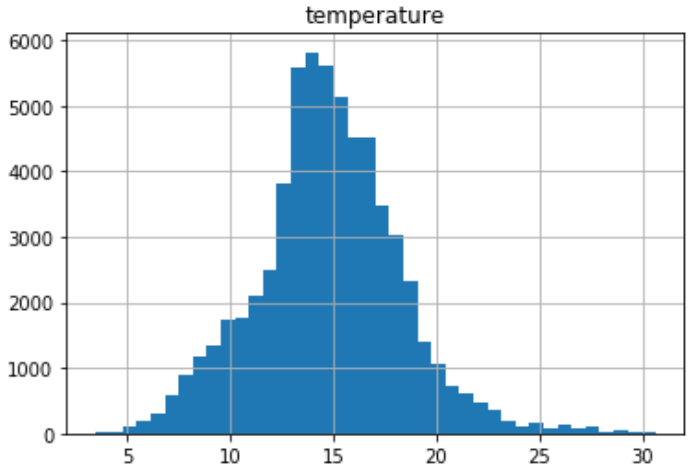
This graphic is another way of visualizing that people prefer to attend the gym on days that are not weekends. As you can see the attendance from Monday to Friday (value of 0 in the graph) is still greater than the assist to the gym during the weekend (value of 1).

Boxplot Grouped by is_holiday



During the holidays, gym attendance also decreases. The 25% (Quartile 1) have values less than or equal to 1, the 50% (Quartile 2 or Median) have values less than or equal to 1 and the 75% (Quartile 3) have values less than or equal to 18.

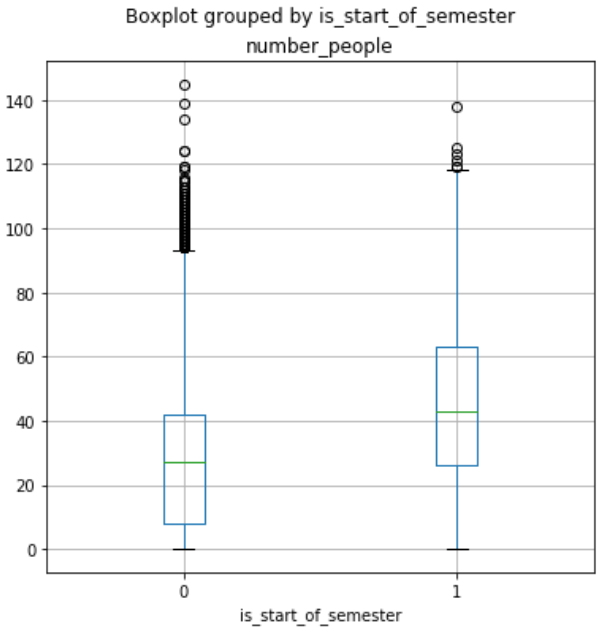
Temperature Histogram

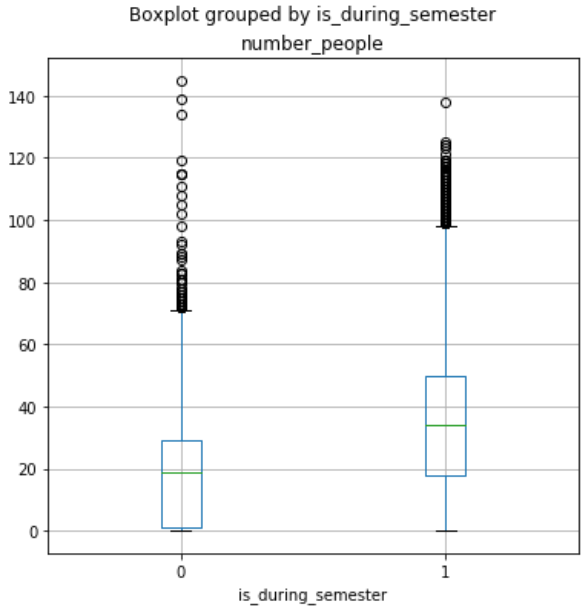


For the analysis, the temperature variable was converted to degrees celsius.

Last year's temperature ranges from 3.41 and 30.65 with a mean of 14.75 degrees celsius.This average temperature could favor people attending the gym.

Boxplot Grouped by is_start_of_semester and is_during_semester

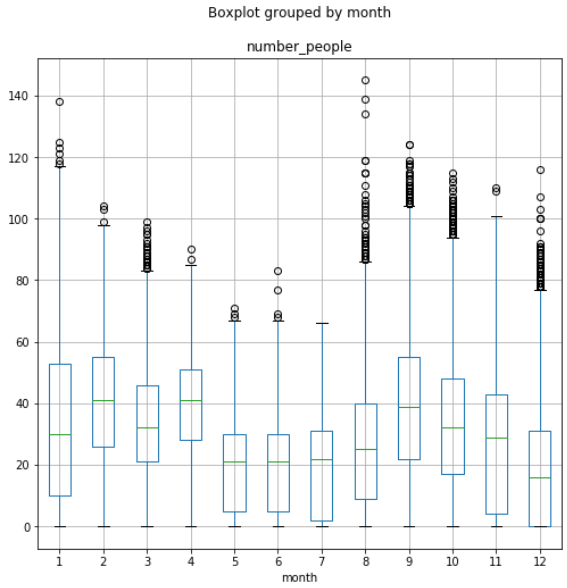




The graph on the left shows that the activity in the gym increases when the semester begins at the University.

If we make a comparison between attending the gym at the beginning of the semester and during the semester, it can be noted that the largest number of people attend at the beginning of the semester.

Boxplot Grouped by month

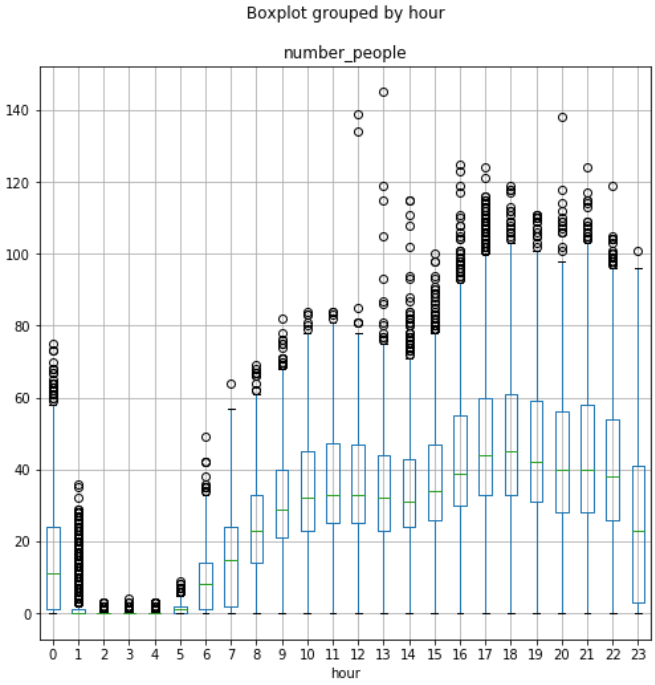


The demand of the gym can be visualized through this graph. It can be noted that the months of the year in which there is a greater influx of people are: January, February, March, April.

The months of the year where a decrease in activities within the gym is denoted: May, June, July.

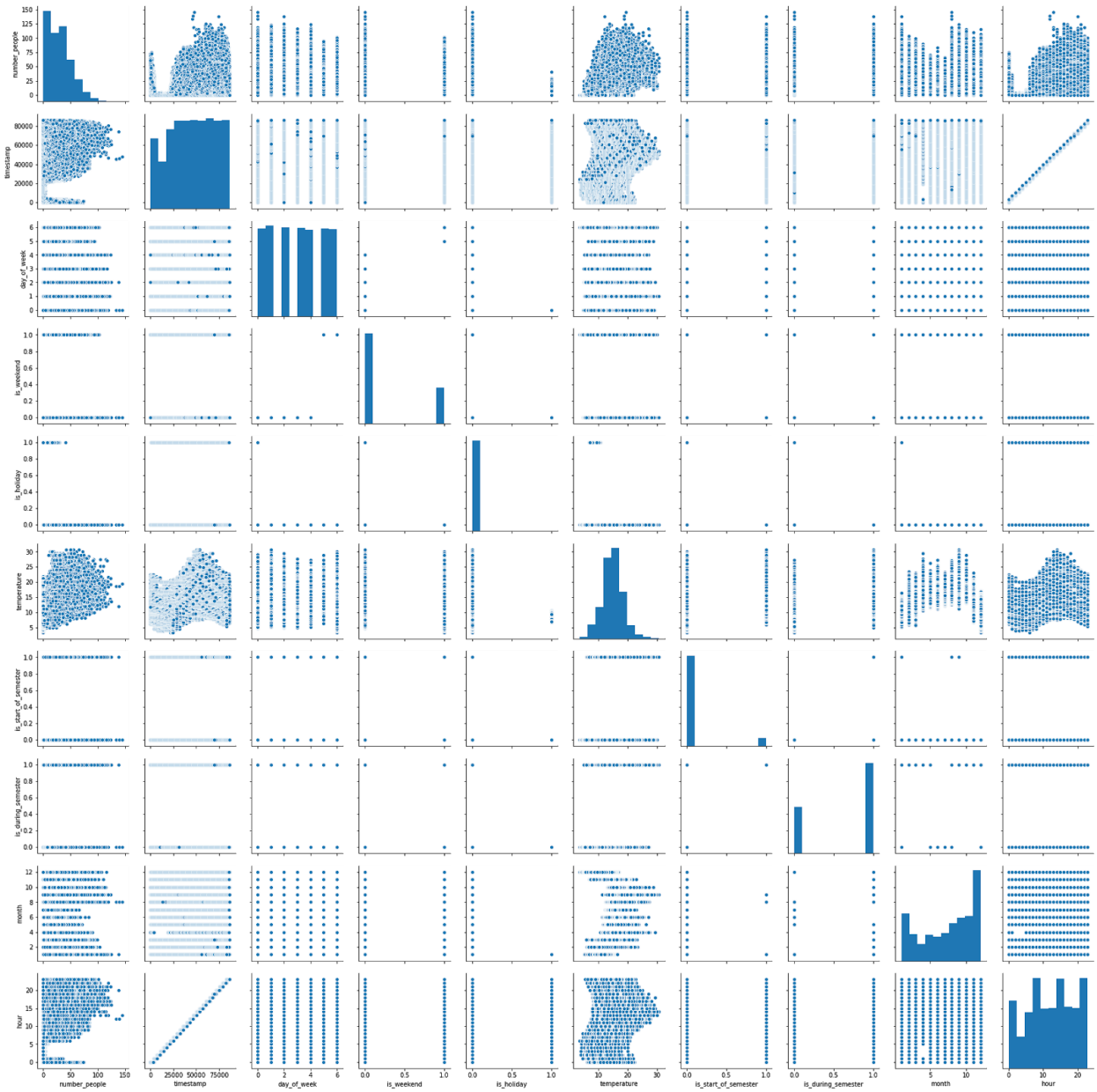
In August, September and October you can see that the activities in the gym are increased again. And in November and December a decrease is noticed again.

Boxplot Grouped by hour



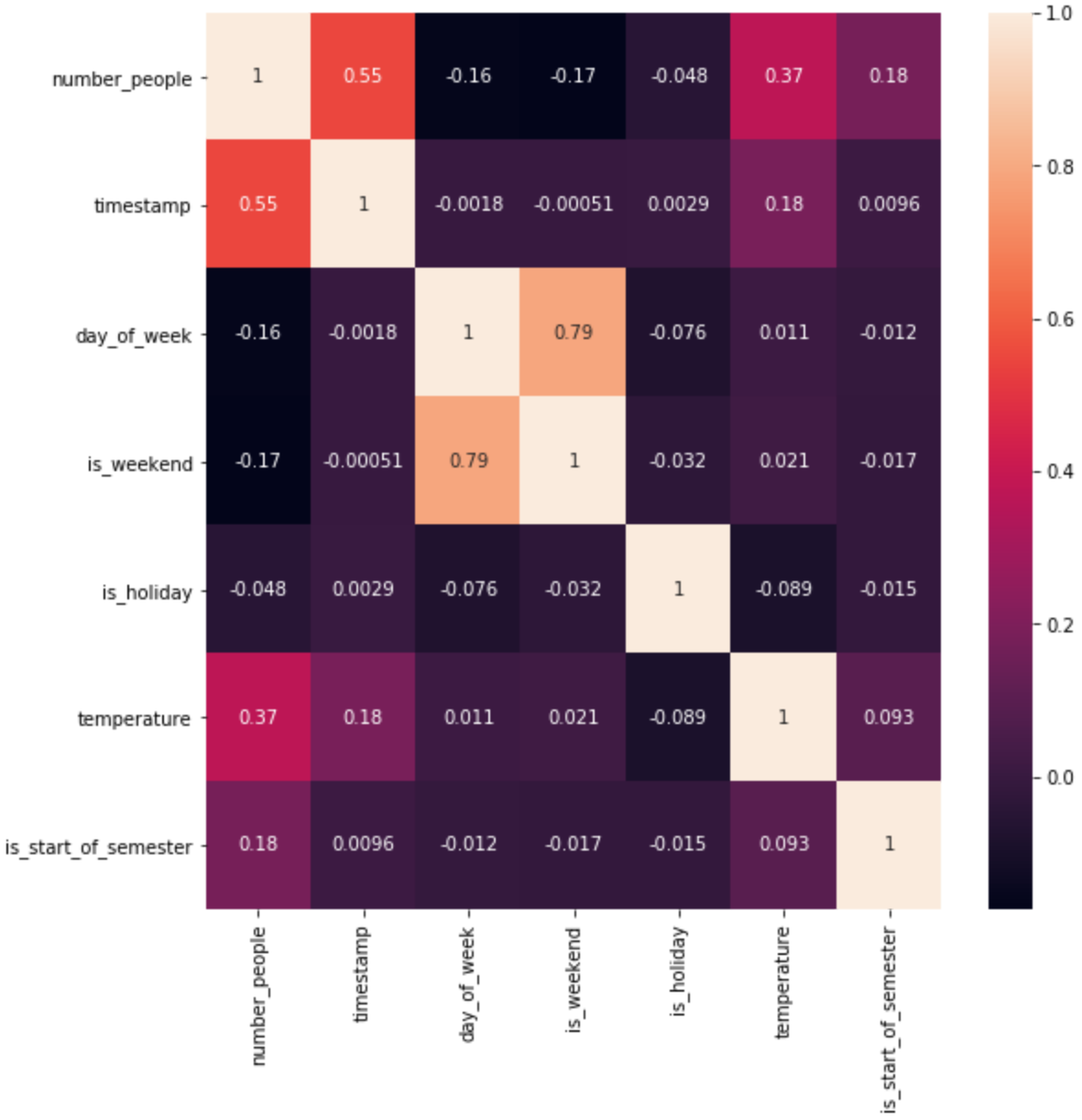
With regard to the hours of attendance at the gym, the box diagram shows that the preference is between 17h00 p.m. to 19h00 p.m. The 25% (Quartile 1) have values less than or equal to 35 personas, the 50% (Quartile 2 or Median) have values less than or equal to 50 and 75% (Quartile 3) have values less than or equal to 60.

6. Pair Plot:



To obtain an overview of the gym assistance data, a pairplot chart is made. This graph shows histograms of the columns in the diagonal of the matrix and pairwise scatter plot of the data.

7. Correlation Matrix:



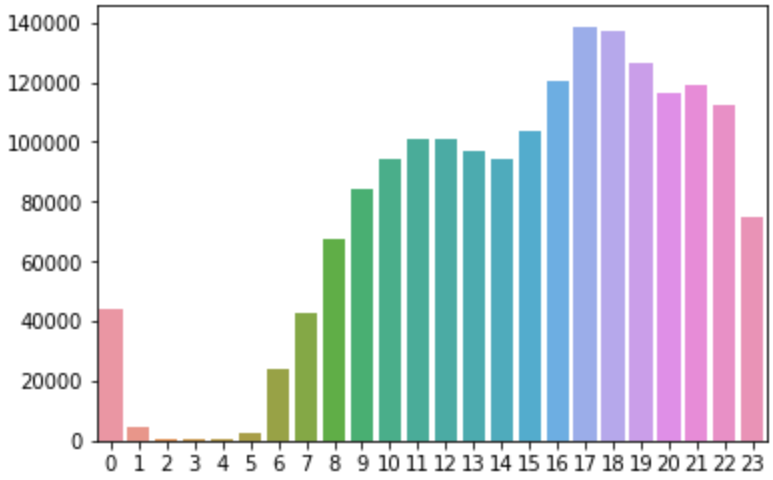
In the correlation matrix you can visualize the relationship that exists between the different variables that make up the dataset.

There is positive correlation between:

1. Number of people and timestamp (0.55)
2. Number of people and temperature (0.37)
3. Number of people and is_start_of_semester (0.18)
4. Timestamp and temperature (0.18)

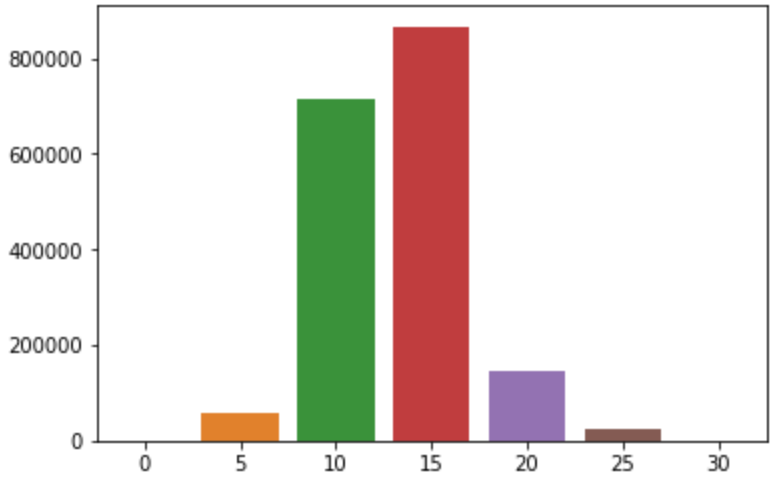
Number of people and timestamp (Correlation Positive 0.55)

If we focus on the highest positive correlation then: Does it mean more number of people are coming at higher timestamp?

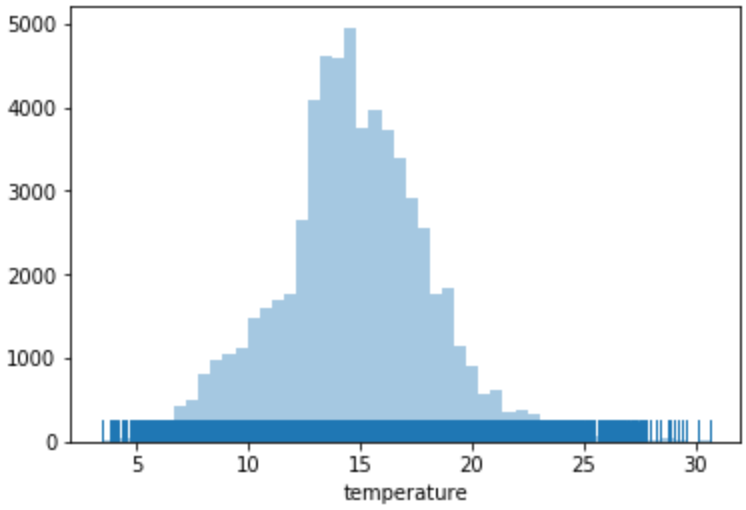


When visualizing the graph we can answer affirmatively to the previous question.

Number of people and temperature (Correlation Positive: 0.37)



From correlation plot, there exists a positive correlation between number of people going to gym and temperature.



So even though from barplot it seems like more people attend gym in warm weather it's not true as most number of days weather remains in the range 10 to 20. So the analysis that more people attend gym during 10 to 20 temperature range is not entirely correct.

8. Prediction Model:

A model will be created to predict the number of people attending the gym. The variables that were selected for the prediction model are the following:

	number_people	date	timestamp	day_of_week	is_weekend	is_holiday	temperature	is_start_of_semester	is_during_semester	month	hour
0	37	2015-08-14 17:00:11-07:00	61211	4	0	0	71.76	0	0	8	17
1	45	2015-08-14 17:20:14-07:00	62414	4	0	0	71.76	0	0	8	17

Two methodologies will be analyzed to determine which one has a higher accuracy. They are:

- SGDRegressor(): Obtain an accuracy of 0.50.
- RandomForestRegressor(): Obtain an accuracy of 0.91.

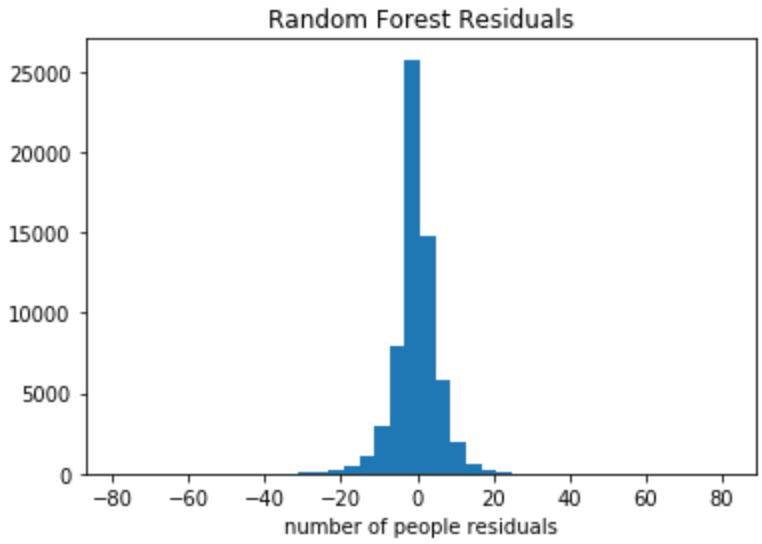
We choose the Random Forest Regressor because it has a greater value of accuracy. Therefore the predictive model can predict the number of people attending the gym with 91% accuracy.

8.1. Feature ranking:

Below is a ranking of the selected characteristics that are part of the model.

1. Feature 6 hour (0.517531)
2. Feature 2 temperature (0.187774)
3. Feature 4 is_during_semester (0.112300)
4. Feature 0 day_of_week (0.088704)

5. Feature 5 month (0.079952)
6. Feature 3 is_start_of_semester (0.013602)
7. Feature 1 is_holiday (0.000137)



The graph shows that the residuals mean is -0.0317.

9. Conclusions:

After analyzing the data about the number of people who attend the gymnasium of the University Campus, the following conclusions are obtained:

- People prefer to attend the gym on days that are not weekends.
- It can be evidenced that people prefer to do activities in the gym when it is not a holiday.

- People show their preference for attending the gym at the beginning of the semester. Since this is related to students, we assume that people go less at the end of the semester because finals/exams are getting close (they have to study).
- The monthly demand of the gymnasium denotes that the months with the assistance of a greater number of people are: January, February, March, April, August, September and October while in the months of May, June, July, November and December there is a decrease in activity in the gym.
- Regarding the exercise schedule preference, it is 17h00 p.m to 19h00 p.m.
- The correlation matrix was an important factor when determining the variables that would make up the model.
- There are factors that affect the number of people attending gym, ranked from top to bottom in decreasing importance manner (1. Hour, 2. Temperature, 3. Is during semester, 4. Day of week, 5. Month, 6. Is the start of semester, 7. Is holiday).
- The model selected, RandomForestRegressor, to predict the number of people that assist the gym in the future has an accuracy of 91%.

10. References

Rose, N. (2017, March 19). Crowdedness at the Campus Gym. Retrieved from https://www.kaggle.com/nsrose7224/crowdedness-at-the-campus-gym