Product Usage Data Analysis Report

Kristofer Siimar

2020

Introduction

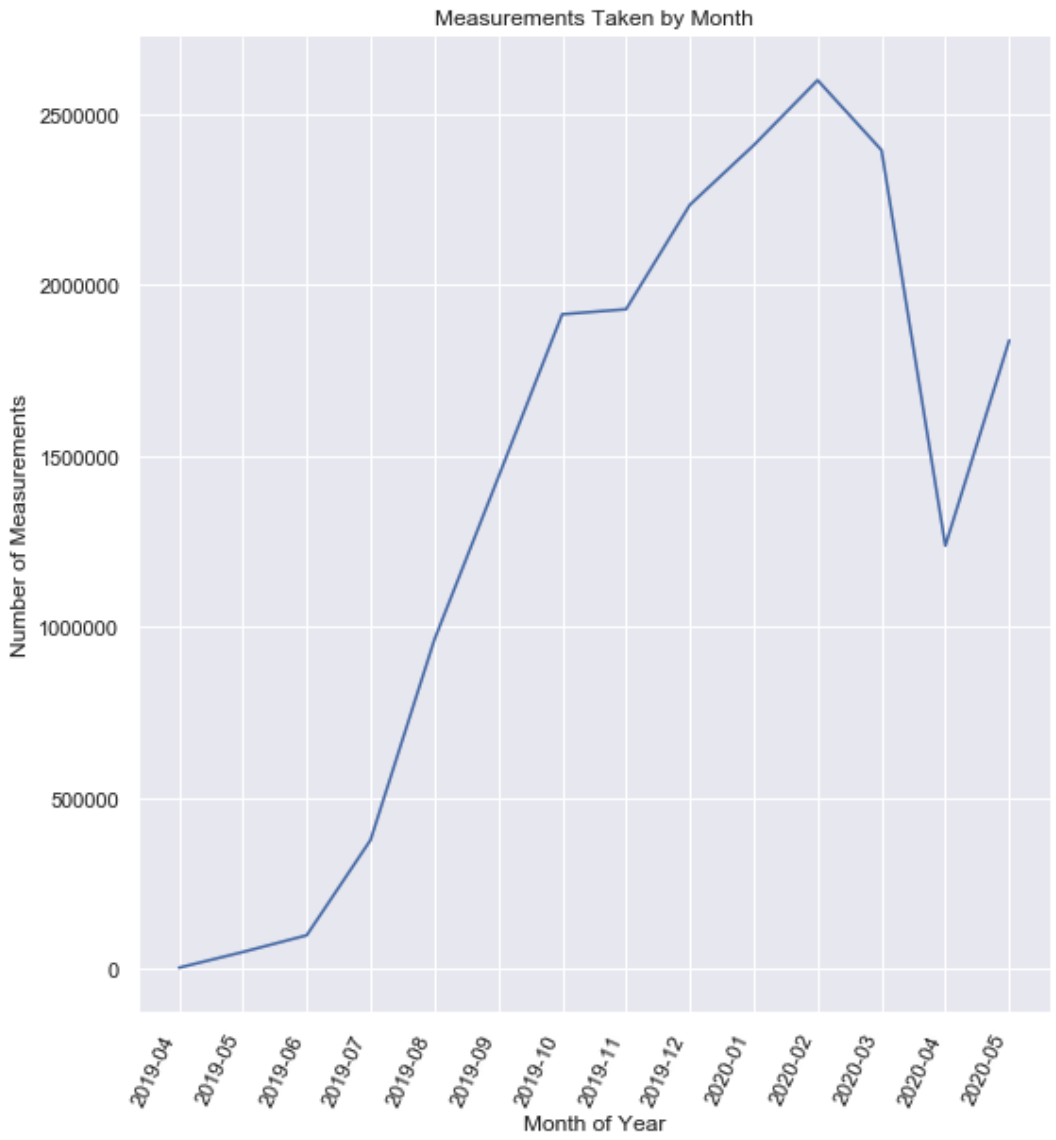
For the analysis project, I worked with CompanyX's current customers' data. The task for the project was to develop an understanding of what kind of data CompanyX currently has, what kind of insight we can get from the data, and what kind of additional data is required to be able to develop an analytics product in the near future.

Analysis

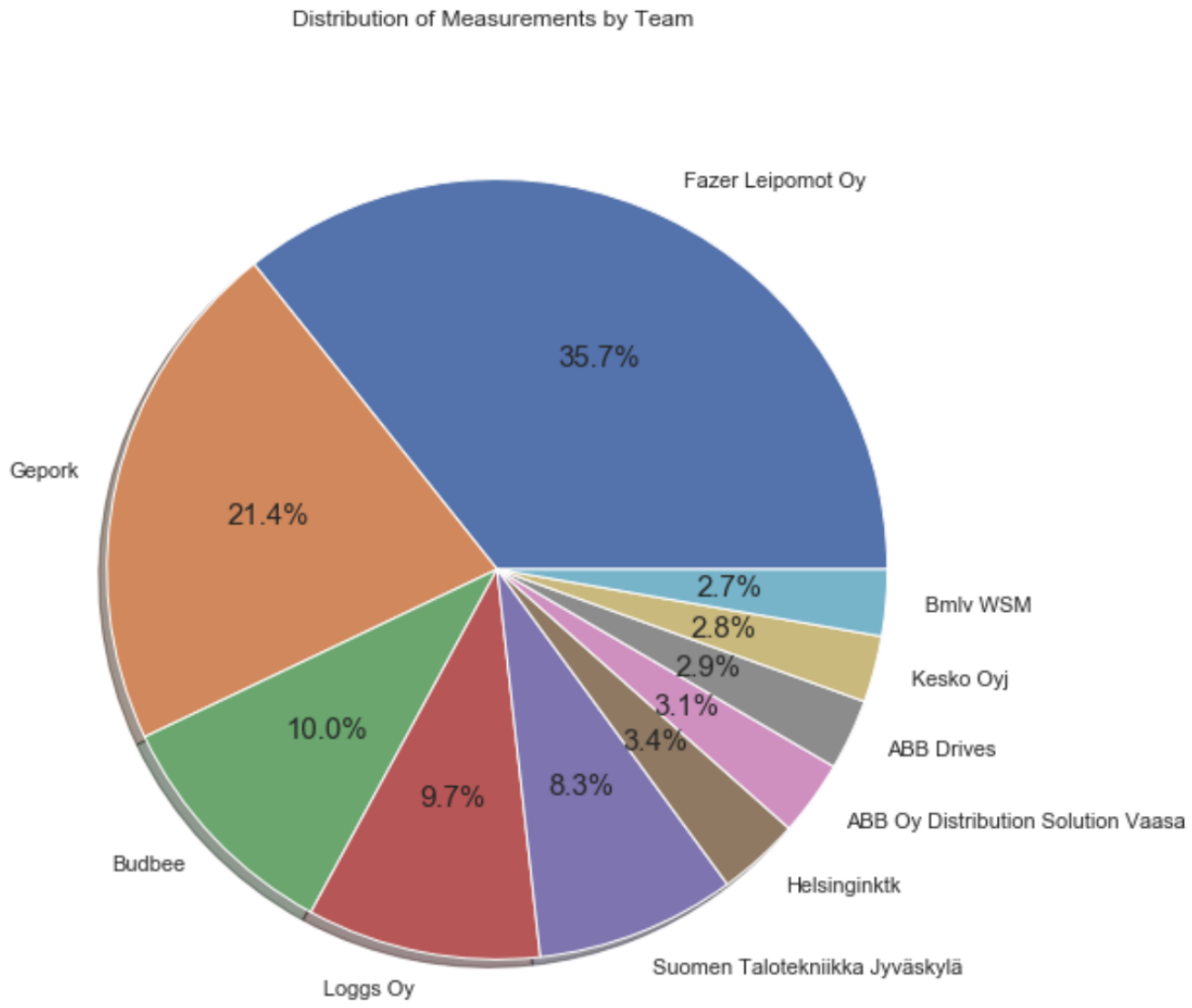
The analysis is structured around CompanyX's data collections, including measurements, loggers, missions, teams, scan events, and alert hits. CompanyX also has other collections that are not as relevant for this project. To eliminate testing and bot data points from the collections, I used data that has been created by loggers with serial numbers between 100 000 and 399 999.

Measurements Data

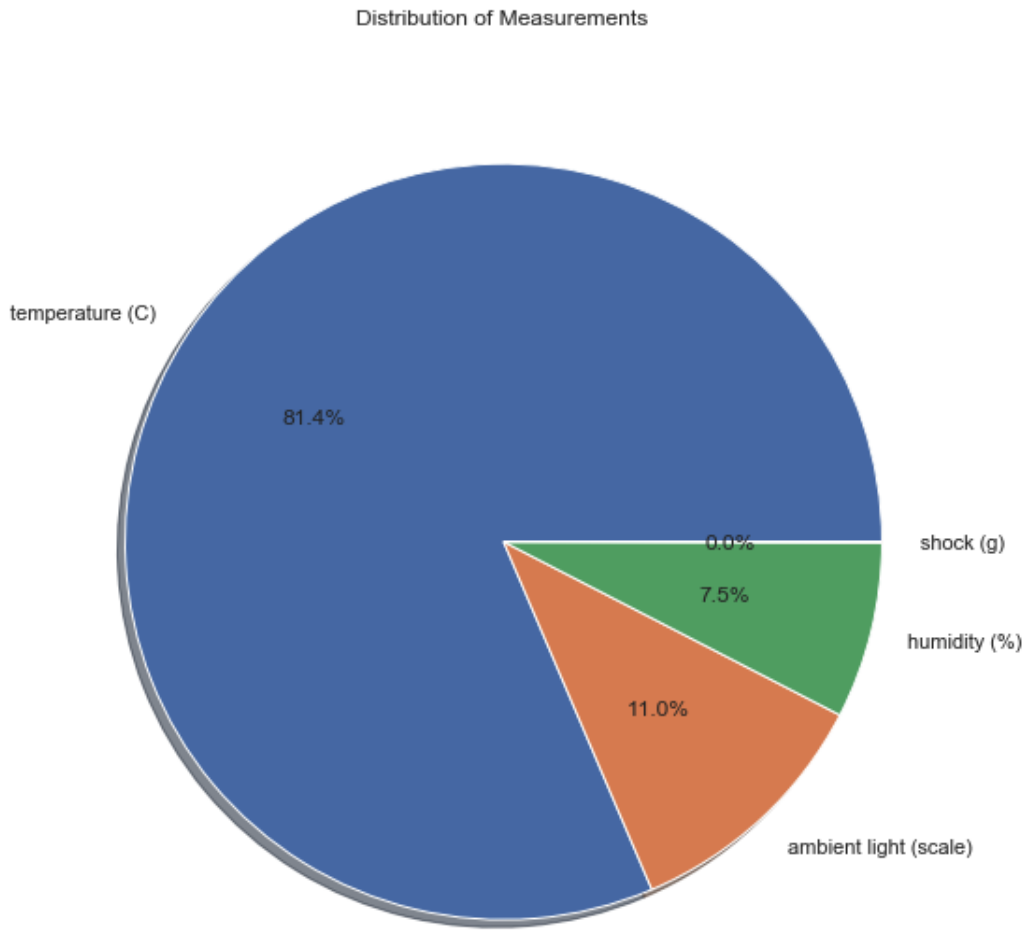
CompanyX currently has 19 485 078 data points of measurements, including temperature, shock, humidity, and ambient light data. On average, each logger has gathered 9279 measurements with a median of 2547 measurements. The number of measurements collected by the fleet of loggers per month has been inconsistently rising since 2018. The most significant amount of measurements - 25 598 294 data points - were gathered in February 2020. Since then, the amount has declined due to the global pandemic. However, the amount of data collected in May is showing rising trends. The measurement data collection by month is demonstrated in the graph below.



From the customer perspective, CompanyX currently has around 220 teams registered in the cloud. Out of all teams, Fazer Leipomot Oy has gathered the most measurements, 3 395 435 data points, which forms 36% of total measurements. Gepork stands second with 21% of the measurements, and the third is Budbee with 10% of the measurements.



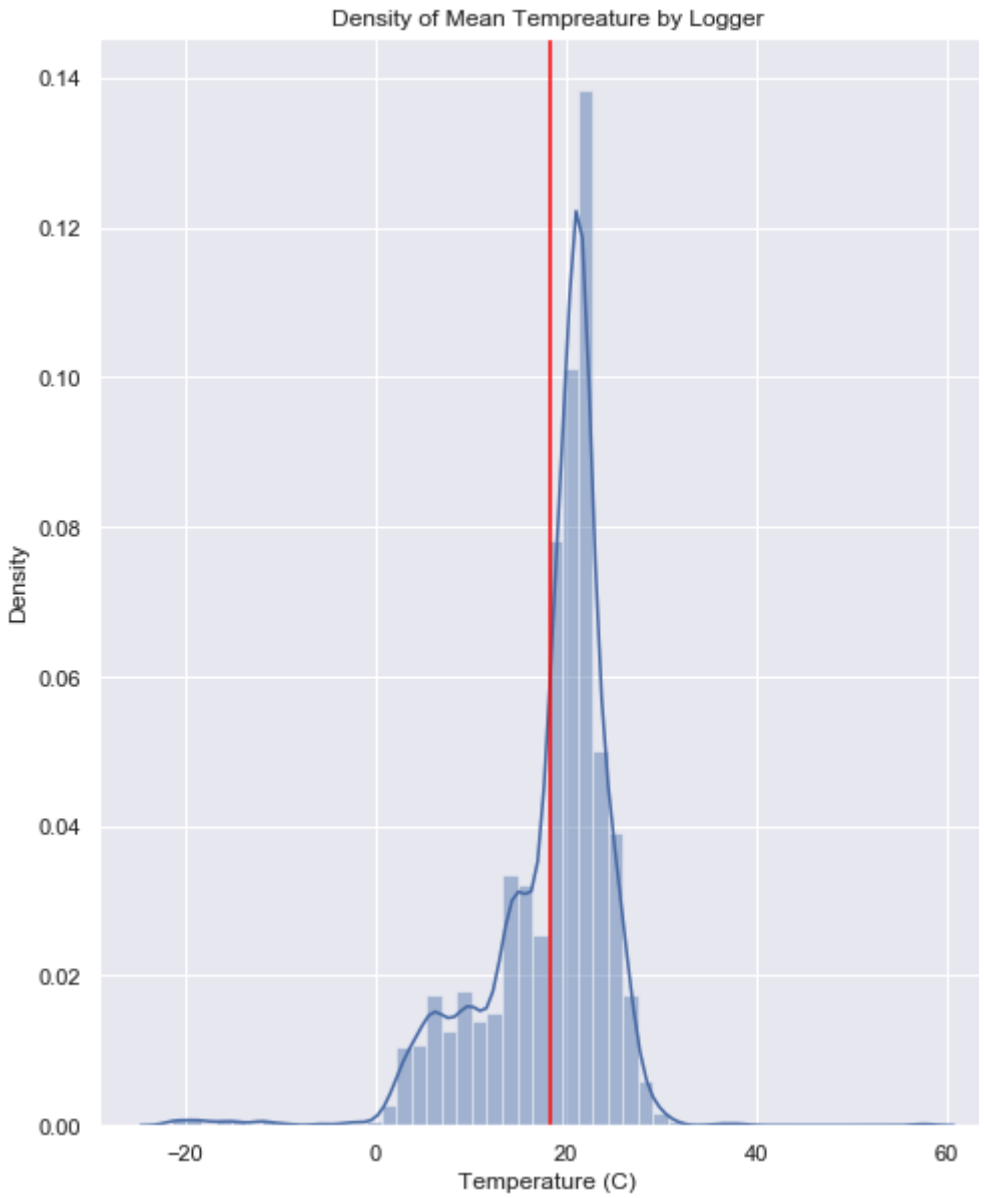
Furthermore, I have analyzed the number of measurements gathered by each sensor type. Most measurements are collected about temperature, making 81% (15 862 535) of all measurements. The second and third are ambient light with 11 % (2 145 299) and humidity with 7.5 % (1 470 670). The fewest measurements we currently have are about shock with 6574 data points.



Besides just counts, I have created general statistics about the measurements collection. All the statistics about each sensor data is displayed in the table below.

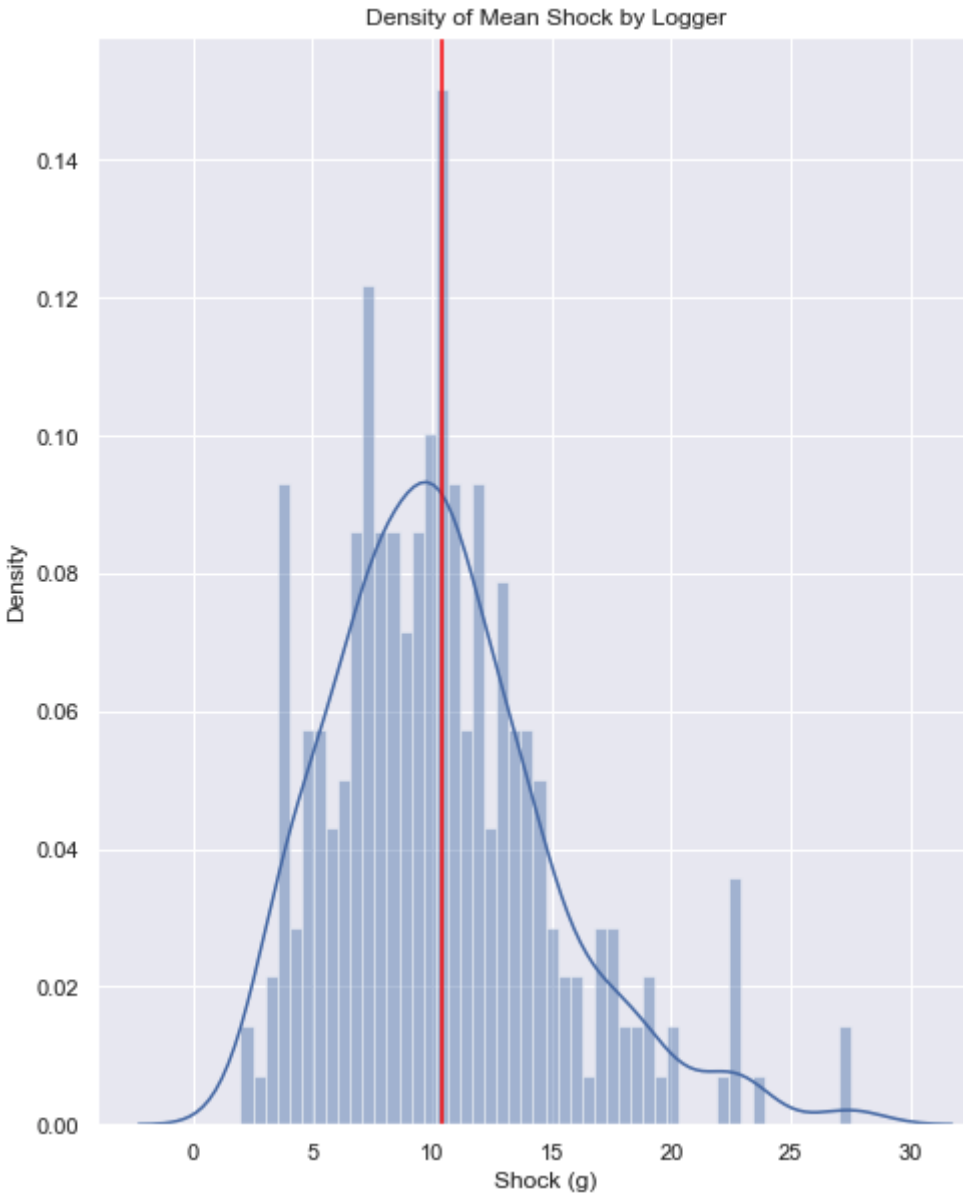
	count	mean	min	max	std
temperature (C)	15862535	14.317138	-30.0	80.000000	8.516959
ambient light (scale)	2145299	67.939571	0.0	15000.000000	653.375233
humidity (%)	1470670	38.561184	1.0	100.000000	18.257025
shock (g)	6574	7.934549	0.9	27.655198	5.320642

From a deeper level, I have analyzed sensor measurements by each logger. Such analysis gives us a better understanding of the environments in which loggers are used. First of all, the density plot below shows the mean temperature gathered by each logger.

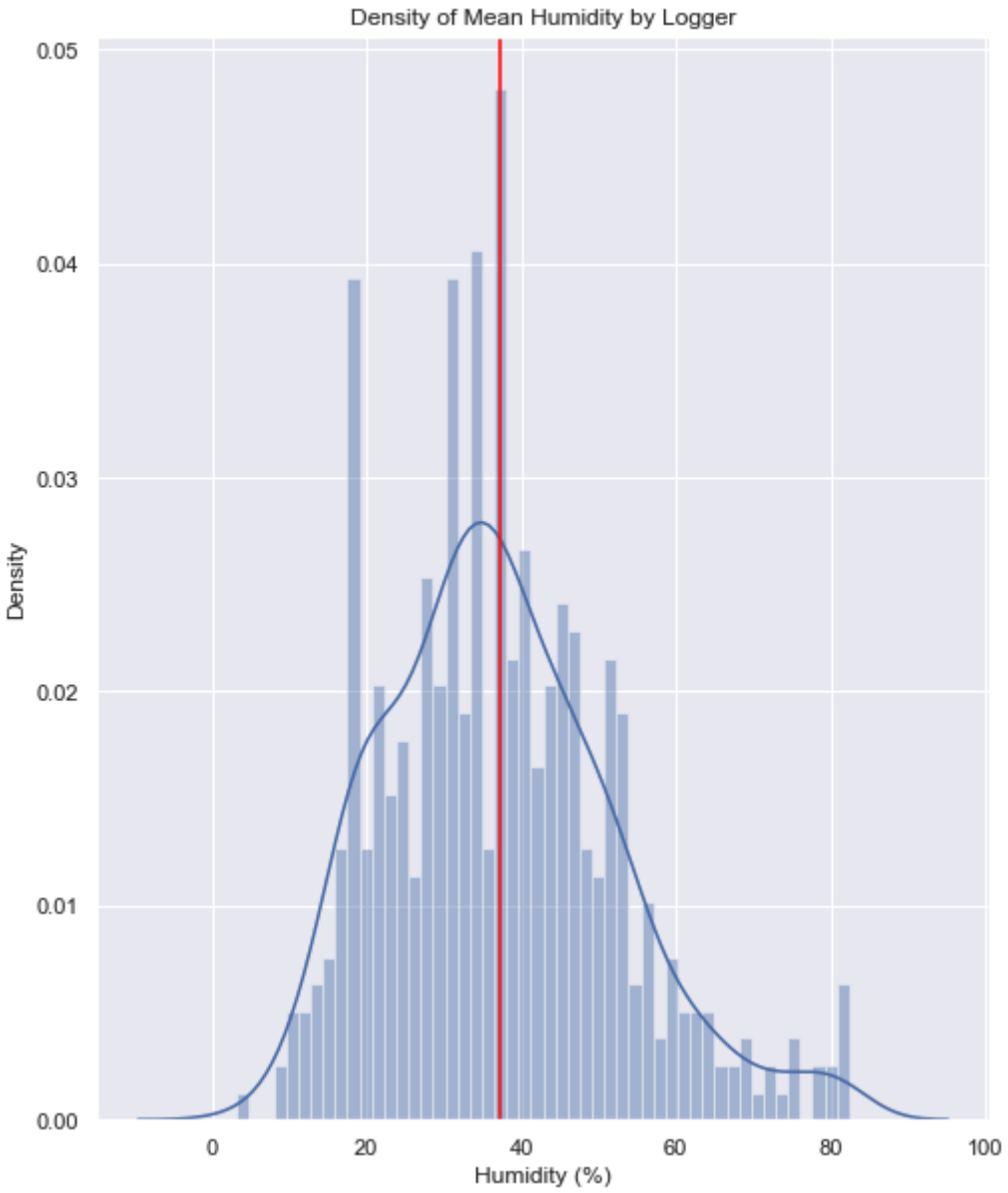


On average, each logger's mean temperature is around 18 degrees Celsius, and over half of the loggers' mean temperatures stay between 16 to 22 degrees Celsius. The median temperature is 20 C. Considering that, most of the loggers are used in environments with usual outdoor temperatures. However, a noticeable number of loggers are also used in colder environments, where the temperature stays between 6 and 13 degrees Celcius.

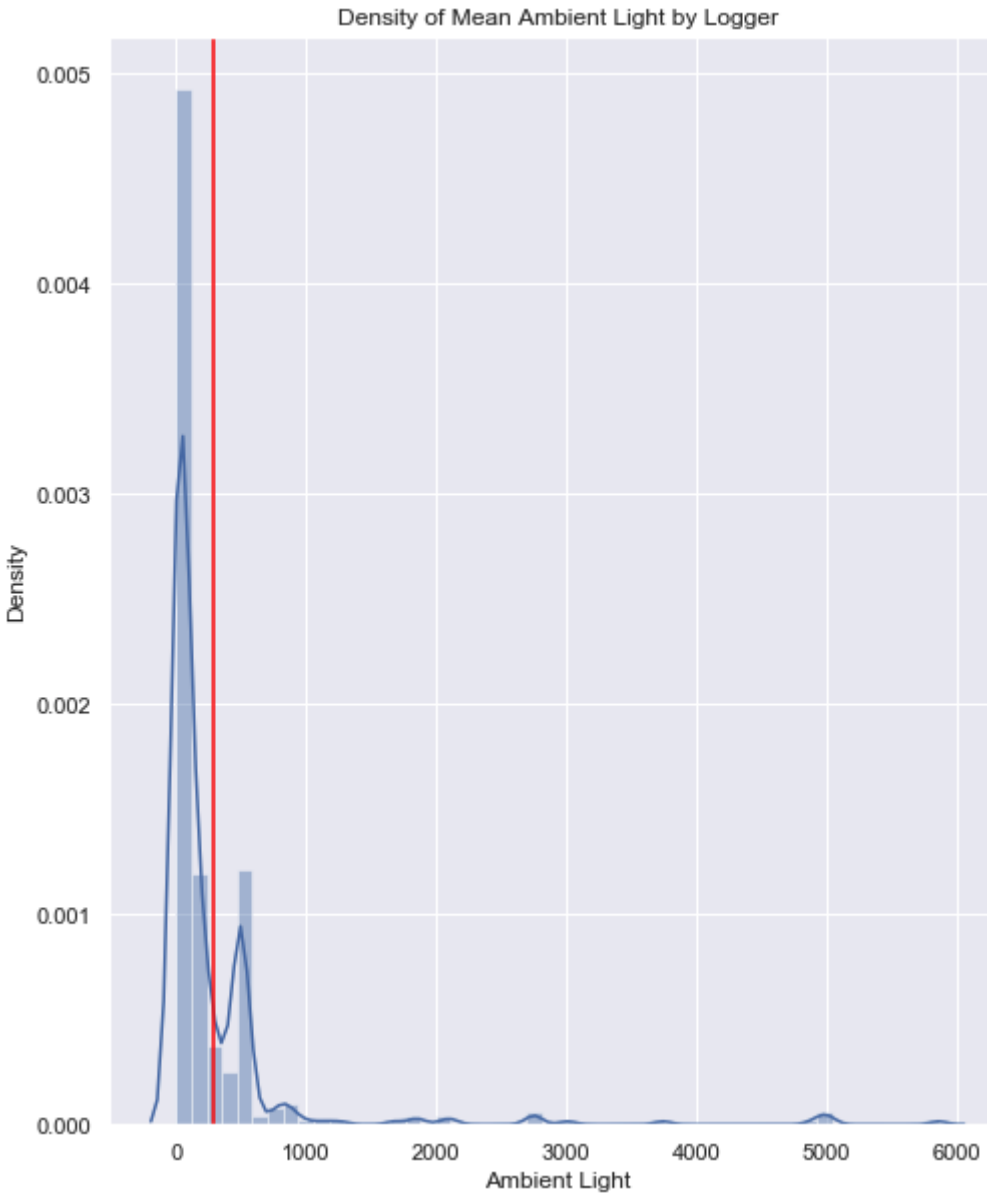
The graph below visualizes the density of mean shock detection of loggers. On average, the mean shock in g-force for loggers is 10 g, and the values range between 2 g and 27 g; the median, in this case, is 10 g. Considering the relatively high deviation, we can indicate that some packages experience wilder transport methods than others.



The next density plot demonstrates the density of mean humidity levels collected by loggers. On average, loggers have a mean humidity level of 37 %. Compared to other sensors, the mean humidity data by loggers is much more volatile. Specifically, the mean humidity levels range from 3 % to 82 %. Considering that, loggers are used in diverse environments, where humidity levels vary.

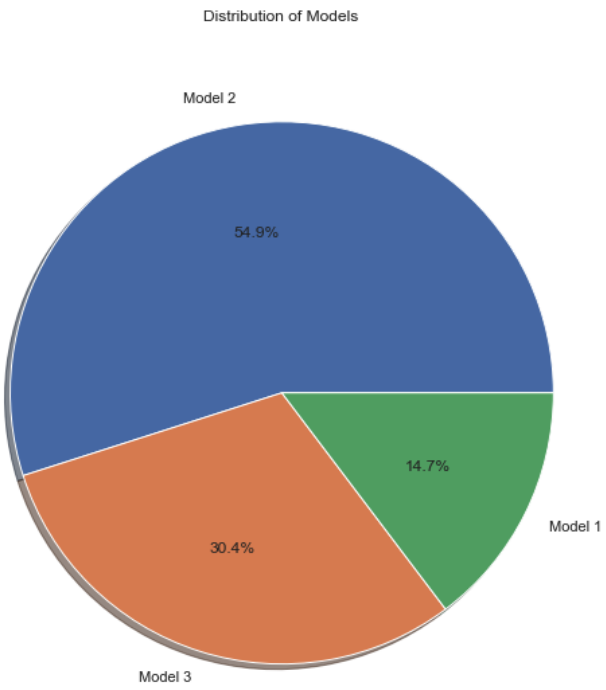


The below density graph demonstrates the density of mean ambient light by loggers. According to the data, loggers have a mean ambient light around 283 scale points, which is described as dim light where the human eye can see colors and shapes. However, the mean, in this case, might not give the best overview as the standard deviation is relatively high, 670 scale points. Therefore, the median of 83 scale points describes better the average light metric, which indicates that most of the loggers are used in dark environments.


Density of Mean Ambient Light by Logger
0.005
0.004
0.003
0.002
0.001
0.000
Density
0
1000
2000
3000
4000
5000
6000
Ambient Light

Loggers Data

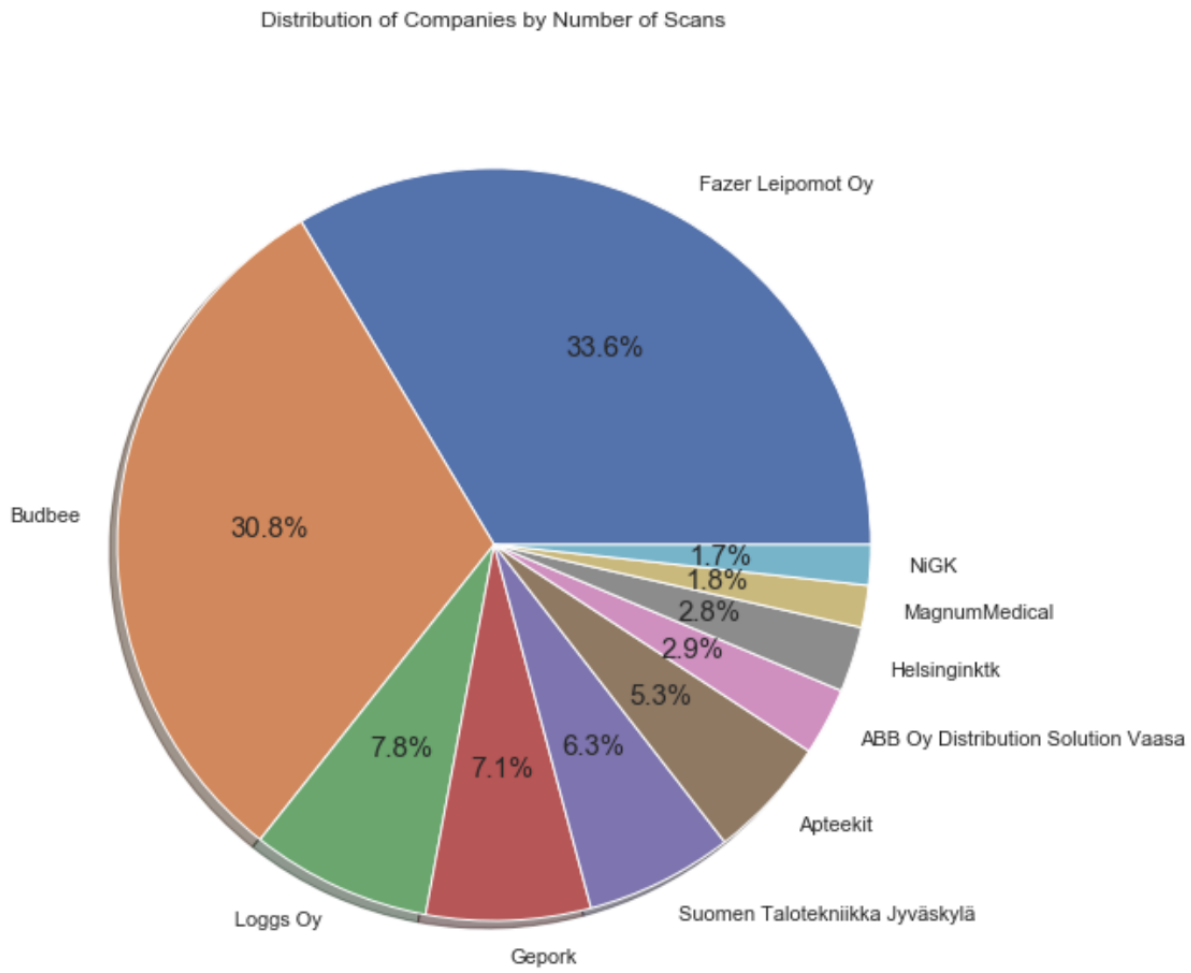
Currently, CompanyX's customers have 4618 loggers in total, from which 55% are Model 2's, 30% are Model 3's, and 15% are Model 1's. Furthermore, Suomen Talotekniikka Jyväskylä has the most loggers - 286, Fazer Leipomot Oy has 166 loggers, and Schenker AG has 71 loggers. Further information about teams and loggers are shown in the graph on the right. This data might not represent the actual situations because many loggers don't have a team attached to them in the cloud.



	team_name
Suomen Talotekniikka Jyväskylä	286
Fazer Leipomot Oy	166
Loggs Oy	113
Schenker AG – Corporate Head Office	71
Protogen	65
ABB Drives	55
ABB Oy Distribution Solution Vaasa	55
Helsinginktk	46
Budbee	42
Fresh	22

Scan Events Data

Next up comes an overview of CompanyX's customers' scanning behaviors. At the moment, users have scanned a total of 48 673 times. With each new scan, users insert about 500 new data points to the cloud, on average. However, the number of data points added varies from 1 to 18 148. As the deviation is substantial, the median of 284 data points might give us a better understanding of the amount of new data inserted with each scan. On to of that, I have analyzed time periods over which users have scanned their loggers. First, a new data point is inserted into the cloud by any user over 12 minutes, on average. The shortest period has been 1 second and longest 6 days and 2 hours when a new measurement has been sent to the cloud. Narrowing down, I have processed how different companies have scanned their loggers.



Fazer Leipomot Oy has scanned the most times, forming 34 % of all scans. The next two performers are Budbee and Gepork, 31% and 7% of the scans respectively. On the company level, Fazer Leipomot Oy has scanned one of its loggers after 53 minutes, on average. Again, as the deviation varies from 1 second to 13 days and 3 hours, the median of 6 hours can offer us a better understanding of Fazer's scanning behavior. Additionally, Fazer scans its most used logger after 15 hours and 30 minutes. This indicates that Fazer uses its loggers relatively consistently.

In the case of Budbee, the average period after which one of its loggers is scanned is 20 hours. The periods range between 1 minute and 4 days and 17 hours and the median is 1 hour and 20 minutes. Budbee scans its most scanned logger after 9 hours. This means that Budbee's scanning behavior is even more consistent that Fazer's.

Compared to Budbee, ABB scans one of its loggers more frequently, after 6 hours and 15 minutes, on average. On the other hand, scanning times for ABB are much more volatile,

ranging between 1 second and 16 days and 1 hour. ABB's median scanning period is 1 hour and it scans its most used logger after 7 days and 5 hours. Considering that, ABB's scanning behavior is inconsistent - it scans its loggers frequently for some shorter period of time and then does not scan them for a while.

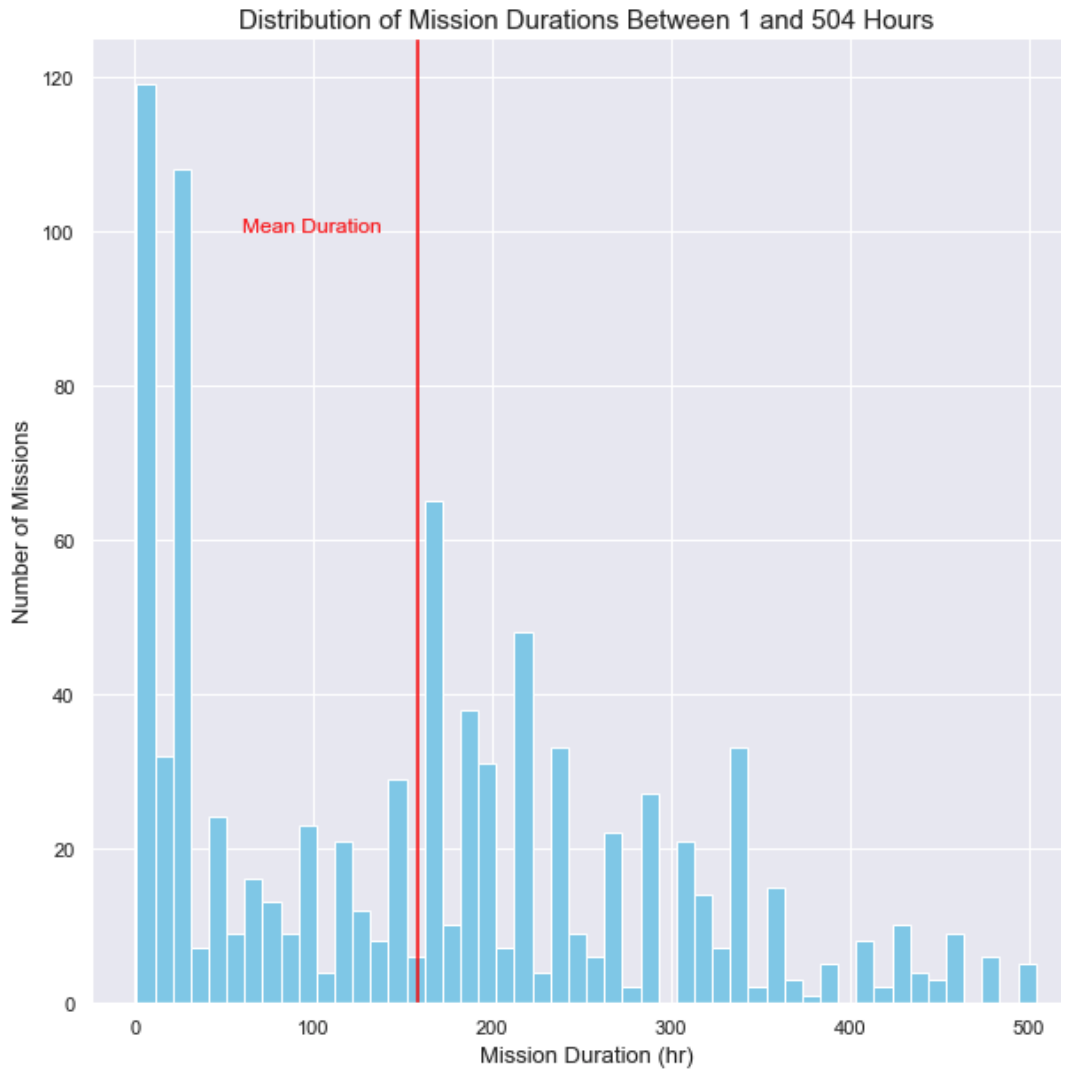
Lastly, Magnum Medical scans one of its loggers after 9 hours and 10 minutes, on average. The period ranges between 4 seconds and 6 days and 1 hour. Furthermore, its median scanning period is 50 minutes. This means that there are many occurrences when Magnum Medical has scanned one of its loggers after short periods or they have greeted a lot of testing data that biases the results. The company scans its most used logger after 4 days and 1 hour, on average.

Missions Data

Currently, CompanyX's customers have created 4174 missions. The top five companies with the most missions are shown in the table on the right. Considering the scans and loggers data from previous pages, we can suggest that more prominent companies with many loggers and scans such as ABB, Budbee, Fazer, Gepork, and Apteekit do not create missions or the data they gather is not linked with their team name on the database.

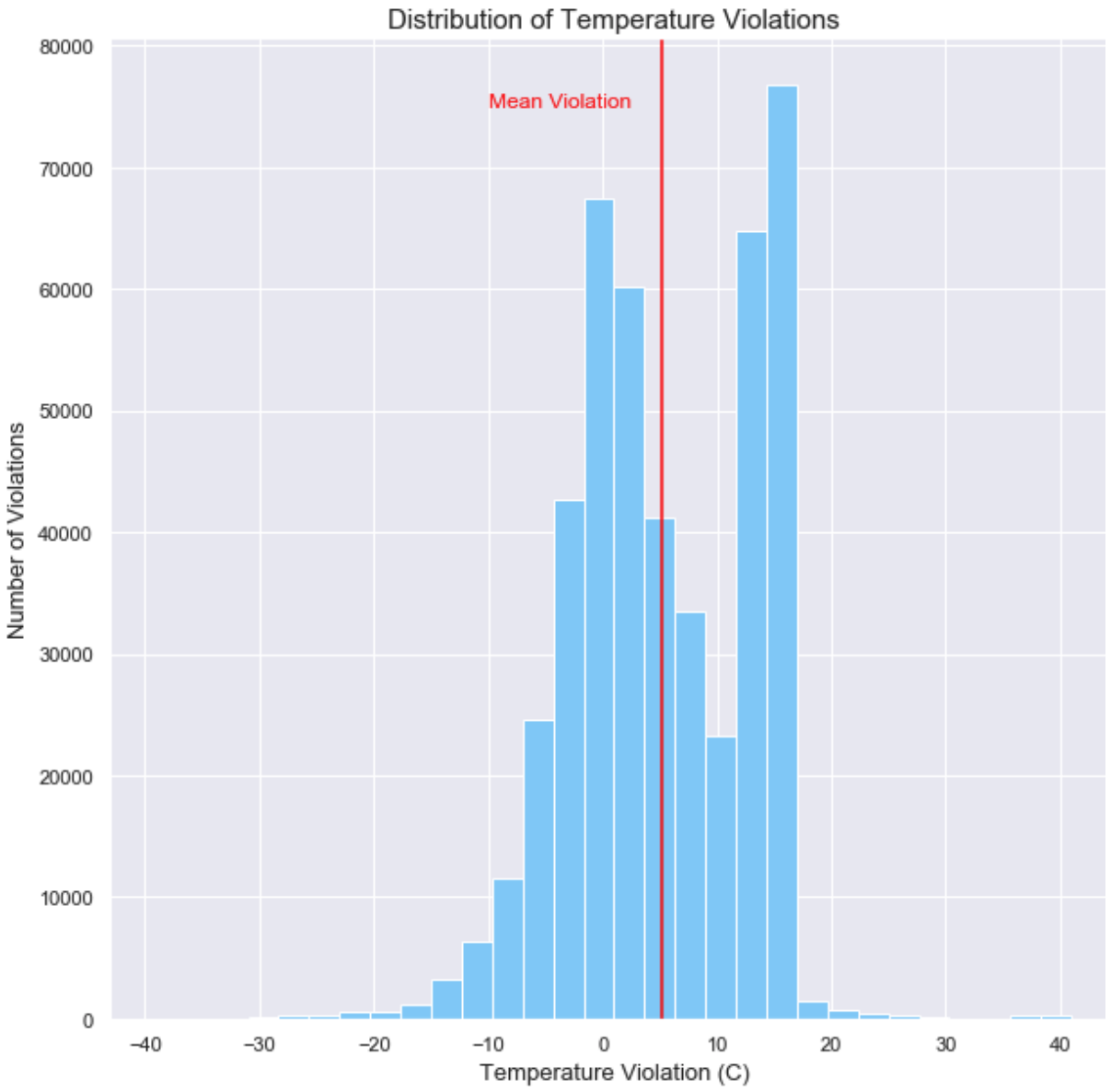
	team_name
MagnumMedical	322
Hoivatiimi	185
Fredman Perfect Kitchen Oy	106
Fresh	100
Pielaveden Apteekki	74

When analyzing missions' durations, we observe the mean mission duration to be 222 hours, which is equivalent to 9 days and 6 hours. The durations vary between 10 minutes (0.16 hr) and 217 days (5207 hr); the median mission duration is 5 days and 23 hours (143 hr). When eliminating the outliers, missions with duration less than an hour, and missions with duration over 3 weeks, we get the average missions time of 6 days and 14 hours (158 hr).



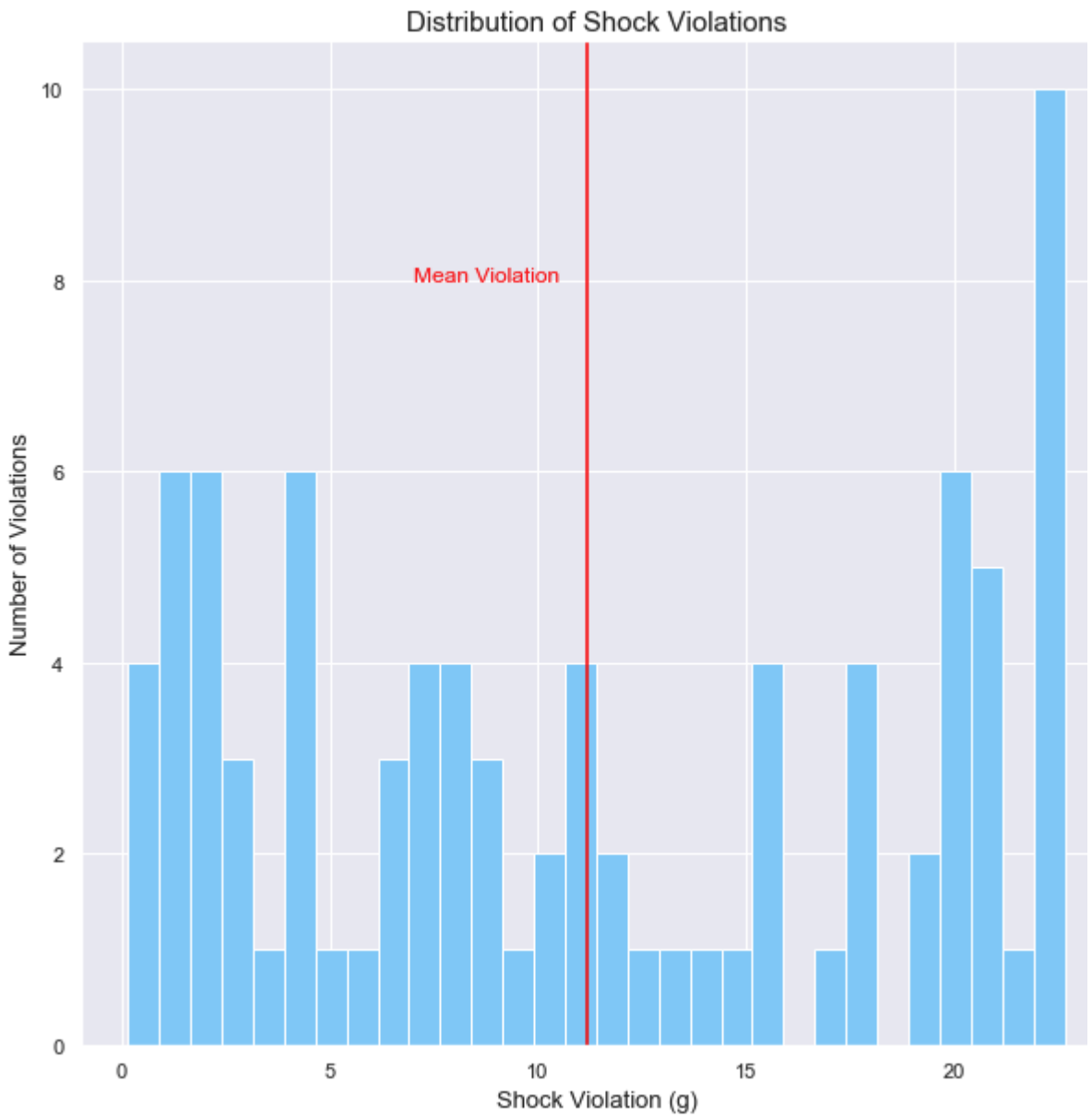
Alerts Data

At this time, CompanyX's customers have received 479 247 alert hits in total; 96 % of them are from temperature sensors, 3 % from humidity sensors, and less than 1 % from light and shock sensors. For temperature, users set the violation limits between 12 and 13 degrees Celsius, on average. Furthermore, temperatures tend to violate their set boundaries by about 5 degrees, on average. The graph below demonstrates the distribution of temperatures' violations.

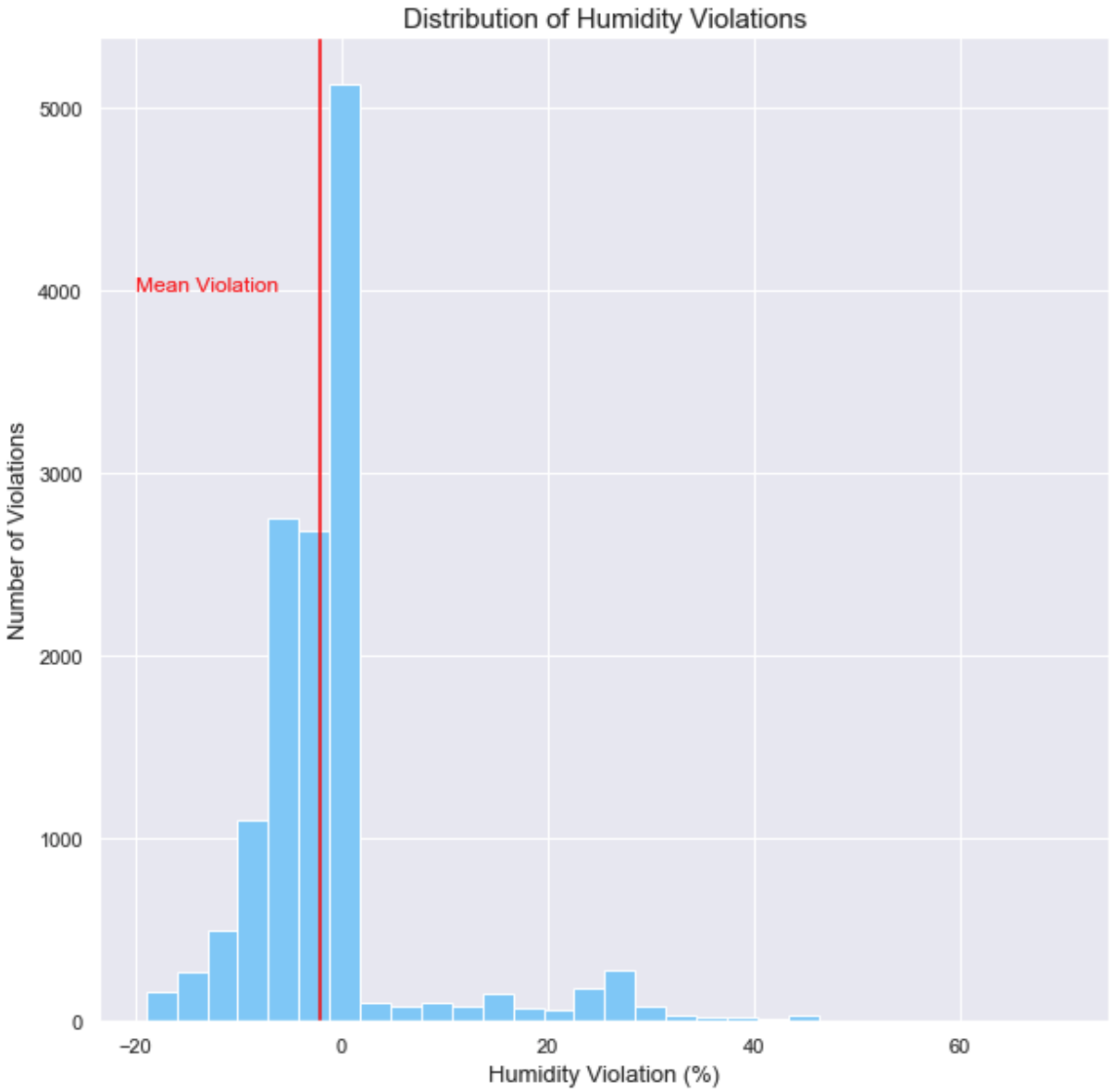


Besides the average, temperature violations tend to vary quite a bit as the standard deviation is 7 degrees celsius, maximum violation of 40 degrees, and minimum violation -15 degrees, meaning that the temperature fell 15 degrees below the lower boundary.

When observing the shock limits, users set the upper boundary to be around 6 g. Besides that, shock sensors detect the violation to be approximately 11 g, on average, which is relatively high. The distribution of shock violations detected by loggers is shown below. The graph shows that similarly to temperature, the violations vary a lot - from 0.1g to 23 g.



Humidity limits are set between 20 % and 30 %. Unlike other sensor violations, humidity violations tend to be less than the set range - the mean humidity violation is around - 2 %.



From the graph, we see that a significantly greater amount of humidity alerts are caused by loggers being in less humid environments than their owners would like them to be.

Lastly, most of the users set the light limit to be 0 as they wish the logger and packages to be unopened. Based on that, the mean violation is also around 0. Interestingly, some of the violations are as large as 500 scale points, but these violations are most probably from the times that packages are opened and the end of their journeys.

Conclusion

In conclusion, I would like to bring up some weak spots on the data perspective that I faced during the analysis. First of all, it can be tricky to filter out all the testing and bot data since some testing loggers have serial numbers in the range of 100 000 and 399 999 (ex. 'Test team' loggers and 'Test team 2' loggers). On top of that, users produce a lot of 'waste' or 'testing' data that bias the analysis. On another topic, some of the data that we have in one collection does not have corresponding data in other collections. For example, 68 % of loggers don't have team data attached to them, and over half of the scanned events don't have location data. To solve these problems, I would first suggest clearly indicating all the data in every collection that is produced by real users. Additionally, I would come up with filters that detect inappropriate data generated by users. Another procedure I would perform to get the data into better shape is to go over each data collection and fill as many documents with appropriate data to unify the whole database. Overall, the analysis is giving a broader overview of Company's users and device usage. A more in-depth analysis of customer behavior, measurements, and missions can be achieved with current data.