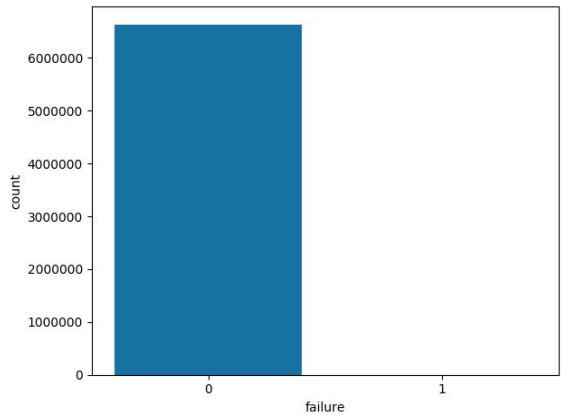
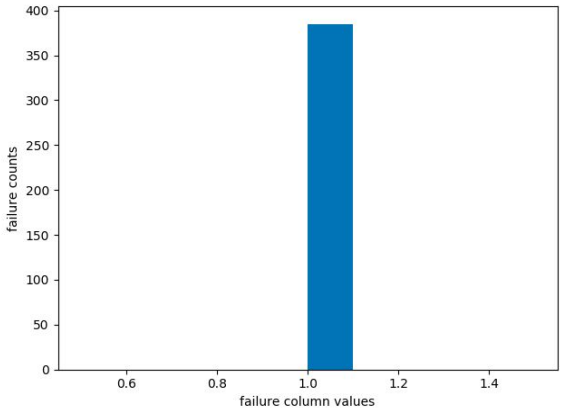
Data Exploration for the year 2017

Overall data

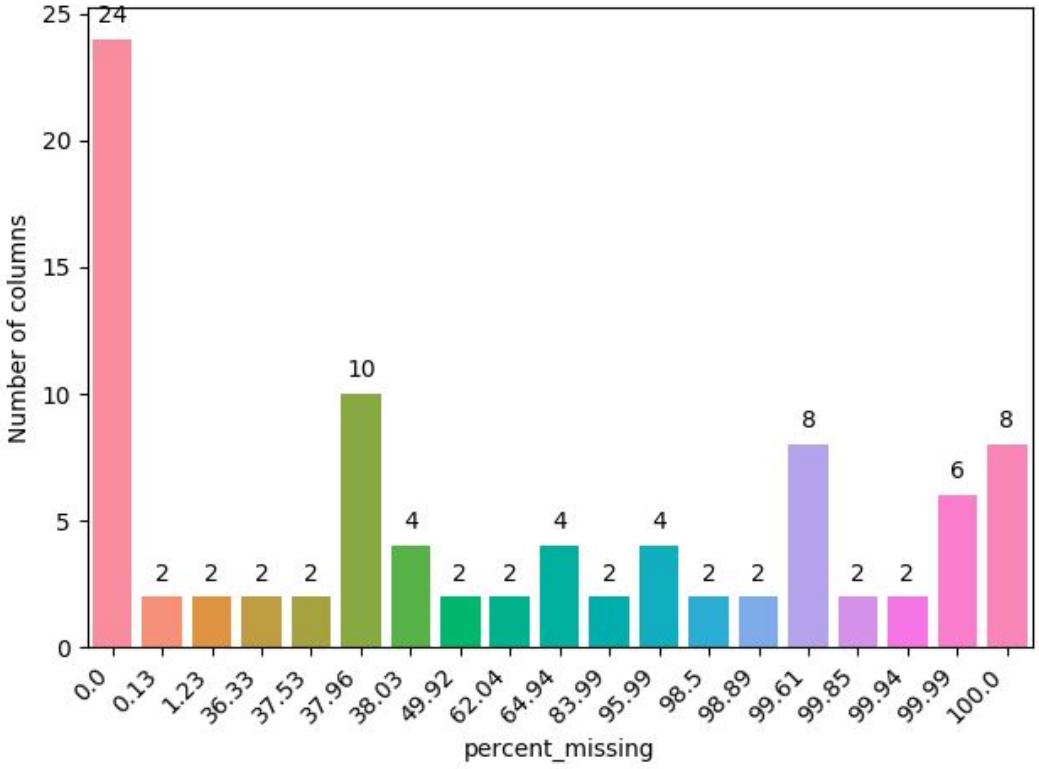
- In Q1 of 2017 - size of data is 6 million, 632 thousand and 104 records, with 95 columns, as shown in below histograms. This tells that the data is very imbalanced as is the case generally with IOT data.
- There are 36 columns found including both raw and normalised among smart stats which have percent missing of more than 75%.
- They contain data for 85301 unique serial number drive.
- There are 385 failures in Q1 part of 2017.
- In one month I have 1989462 - 1 million 989 thousand and 462 records, with 73,885 drives data, failure cases are 101.





Emptiness of Smart Stats columns

Below histogram is produced using one month data(1 million records), 2017-Q1, first month. From the below plot, just 14 columns we can take if we put threshold of 30% emptiness.



column_name	percent_missing
smart_1_normalized	0.00000
smart_194_normalized	0.00000
smart_194_raw	0.00000
smart_12_raw	0.00000
smart_12_normalized	0.00000
smart_197_normalized	0.00000
smart_197_raw	0.00000
smart_10_raw	0.00000
smart_10_normalized	0.00000
smart_9_raw	0.00000
smart_9_normalized	0.00000
smart_198_raw	0.00000
smart_198_normalized	0.00000
smart_7_normalized	0.00000
smart_5_raw	0.00000
smart_5_normalized	0.00000
smart_4_raw	0.00000
smart_4_normalized	0.00000
smart_3_raw	0.00000
smart_3_normalized	0.00000
smart_199_normalized	0.00000
smart_199_raw	0.00000
smart_1_raw	0.00000
smart_7_raw	0.00000
smart_192_raw	0.13000
smart_192_normalized	0.13000
smart_193_raw	1.23000
smart_193_normalized	1.23000
smart_191_normalized	36.33000

The **14 smart stats** which have emptiness percentage to be less than 30% are -

Smart 1
Smart 3
Smart 4
Smart 5
Smart 7
Smart 9
Smart 10
Smart 12
Smart 192
Smart 193
Smart 194
Smart 197
Smart 198
Smart 199

These 14 are the columns which can be used for baseline model and then we can introduce the columns with more emptiness percent using imputation of KNN later and experiment with extra columns and their imputed values for missing ones. But, we can also match the ranges of these 14 columns in below whisker plots of Smart stats. Among the 14 columns, Smart 5, 10, 197, 198 are marked as critical for failure in https://en.wikipedia.org/wiki/S.M.A.R.T. Also for almost all of these smart values the ideal values are described to be low ones.

smart_191_normalized	36.33000
smart_191_raw	36.33000
smart_240_normalized	37.53000
smart_240_raw	37.53000
smart_189_normalized	37.96000
smart_184_normalized	37.96000
smart_184_raw	37.96000
smart_187_normalized	37.96000
smart_187_raw	37.96000
smart_188_normalized	37.96000
smart_188_raw	37.96000
smart_189_raw	37.96000
smart_190_normalized	37.96000
smart_190_raw	37.96000
smart_242_normalized	38.03000
smart_242_raw	38.03000
smart_241_raw	38.03000
smart_241_normalized	38.03000
smart_183_normalized	49.92000
smart_183_raw	49.92000

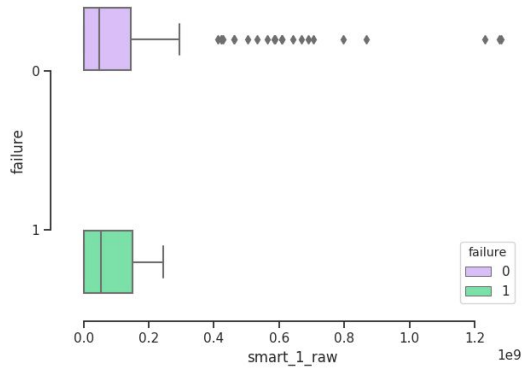
The columns with 36 - 40% missing values are below **9 more columns** which can be considered for careful imputations :

Smart 184
Smart 187
Smart 188
Smart 189
Smart 190
Smart 191
Smart 240
Smart 241
Smart 242

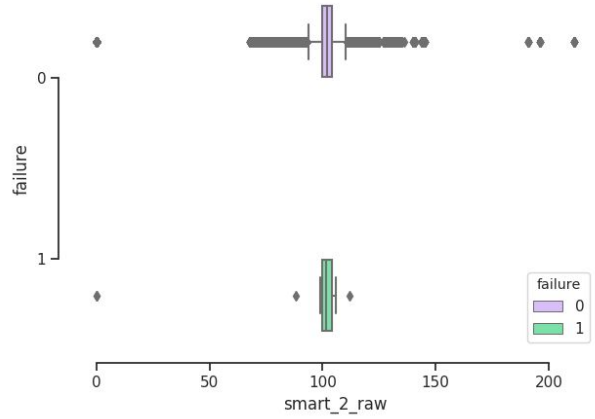
Smart 183 is not that helpful because having 49% missing it has most values as outliers.

Medians & Quartiles of Smart Stats

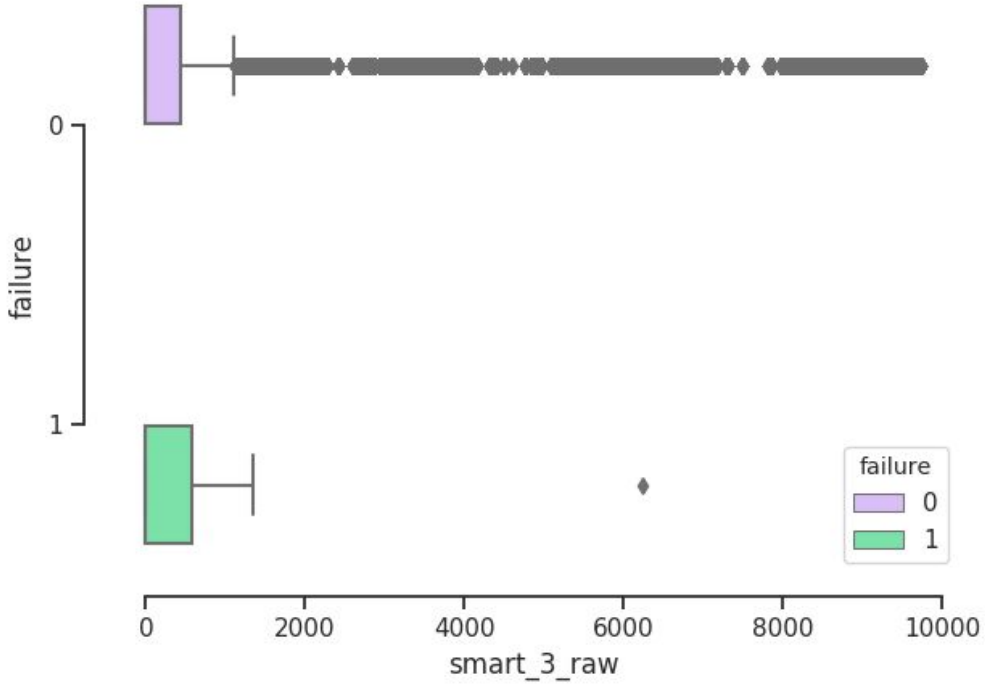
Below whisker plots show the medians, quartiles and range of values for failure and operational drives for all smart stats non-empty for more than 75% for 1105153 (around 1 million records) data.



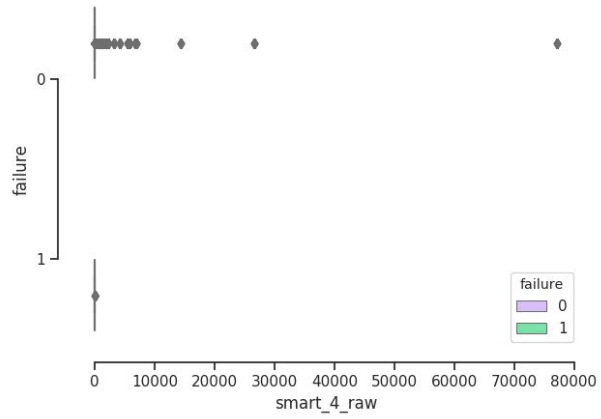
Smart 1 can be useful with showing variation and having 0% missing values

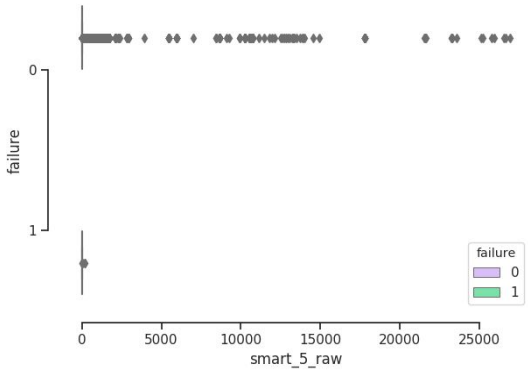


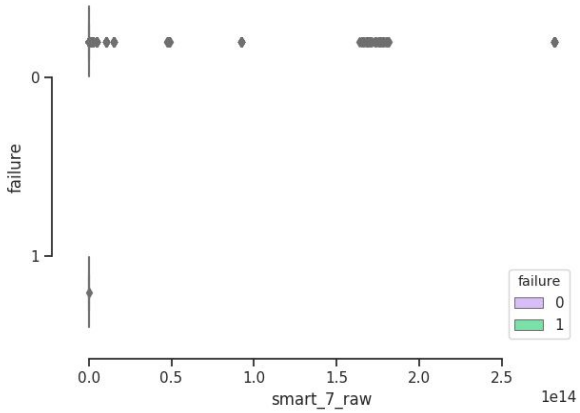
Smart 2 has 64.9 % missing values and appears from above plot very less variation, making difficult to differentiate between its values for failure and operational drives. Thus, it does not seem to be very important.

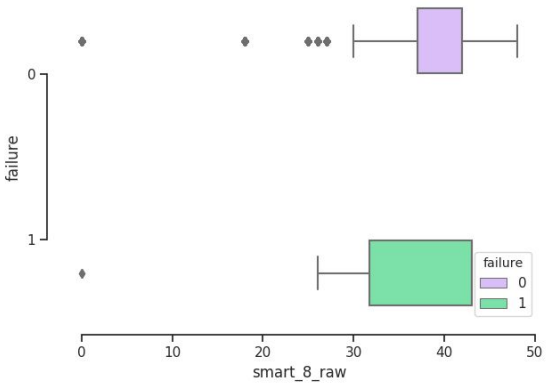


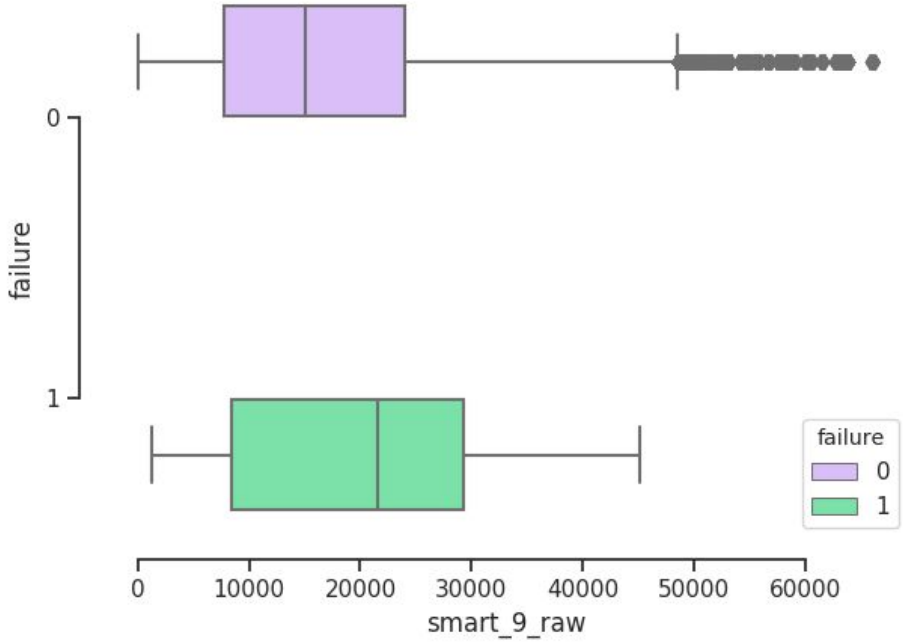
Smart 3 also appears to be important for the same reason as Smart 1 stat.



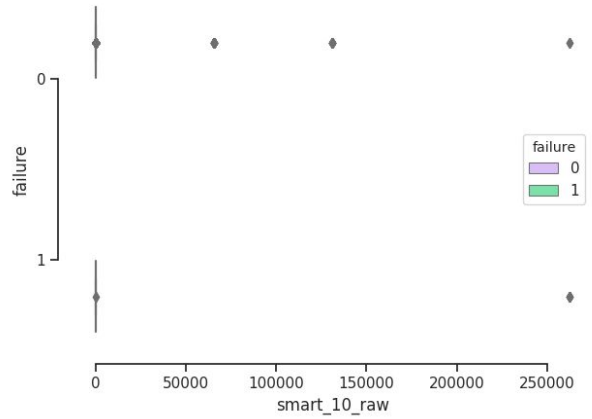


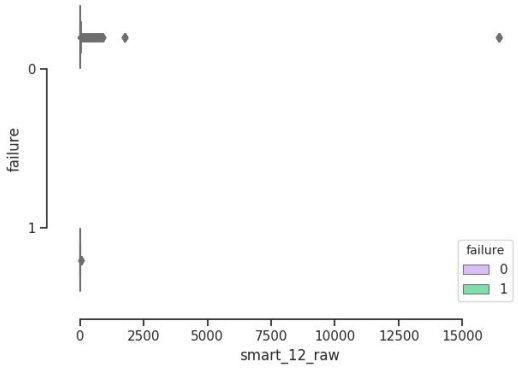




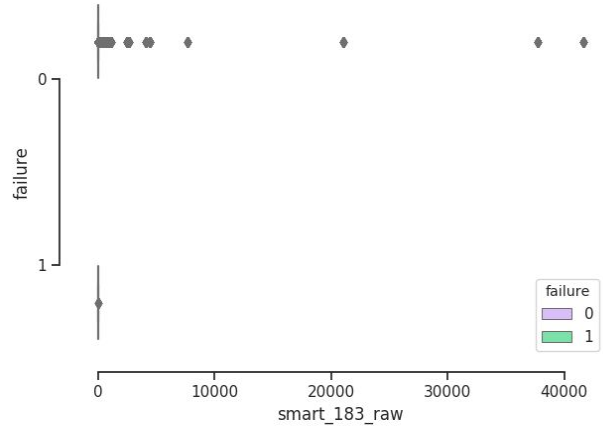


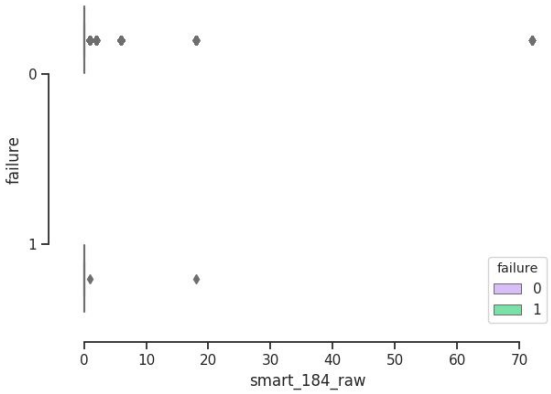
Smart 9 shows to have higher median as well as more variation for failure drives.

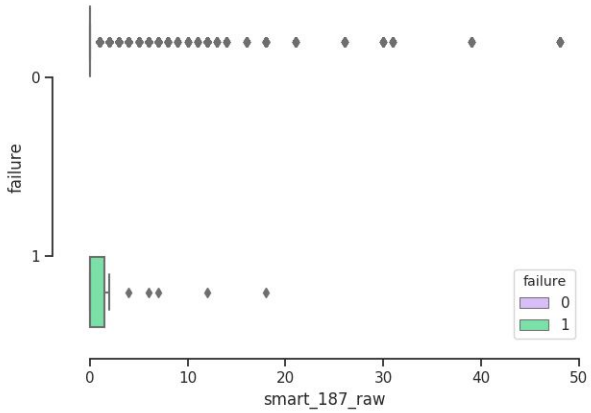


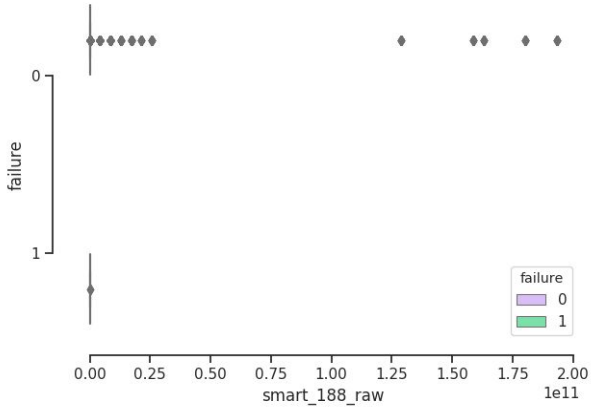


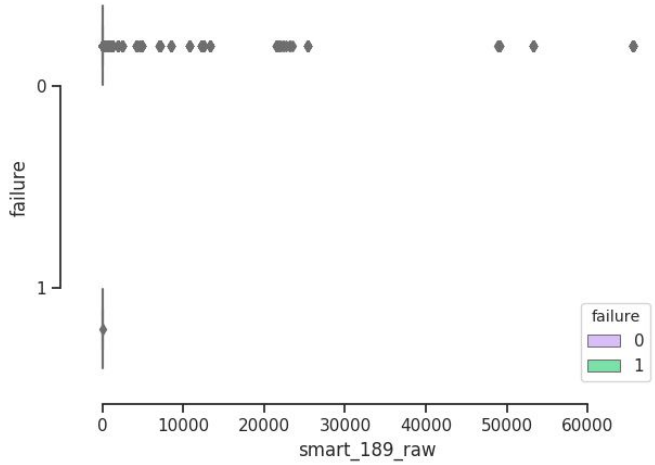
Smart stats 4, 5, 7, 10, 12 appear to have either outliers or no variation. Hence they do not prove to be important for failure prediction even though they are among the least percent missing value columns. Smart 8 even though showing variation has high amount of missing percent values.



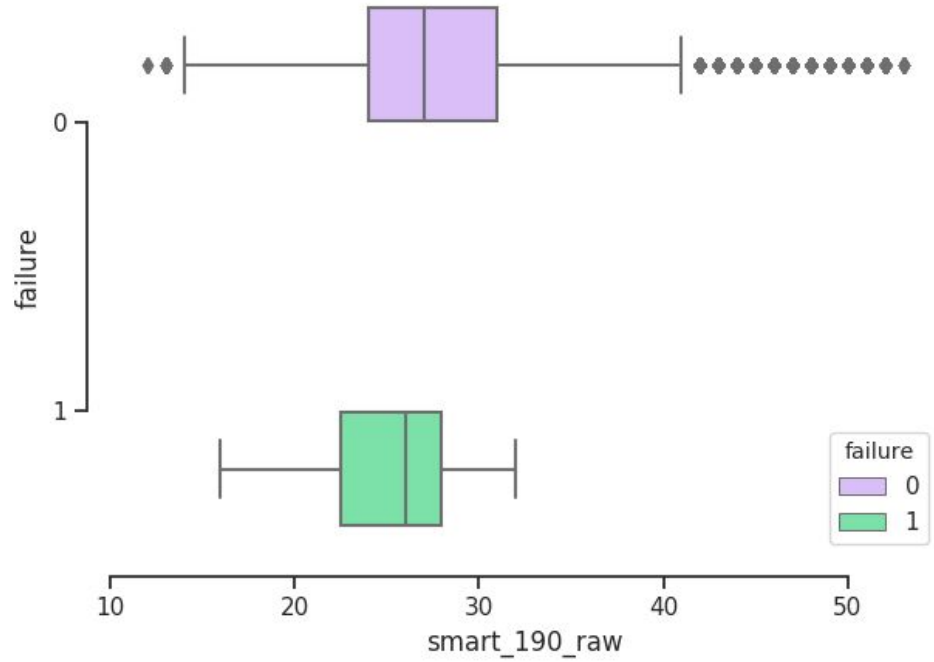




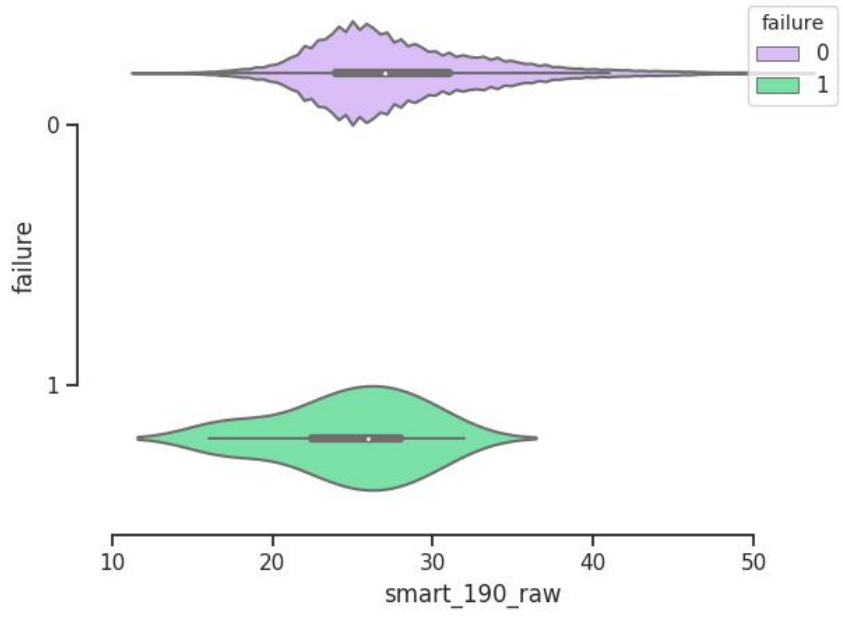




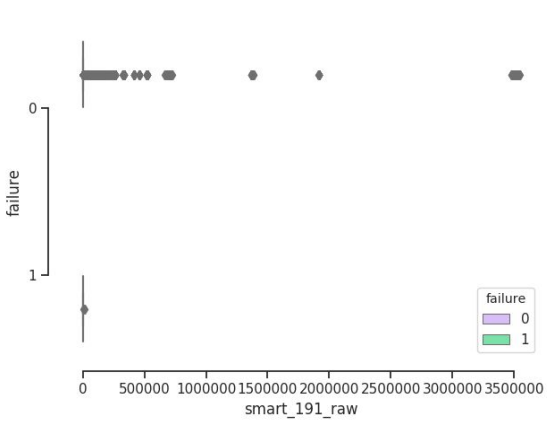
We can notice that all of the plots of 184, 187, 188, 189 smart stats are around 36% empty but as from their whisker plots we can observe they show no variations in values, absolutely zero variation, they do not seem to be helpful. Smart 183 not only has higher percent missing(49%) but also shows no variation so not appears to be useful for consideration.

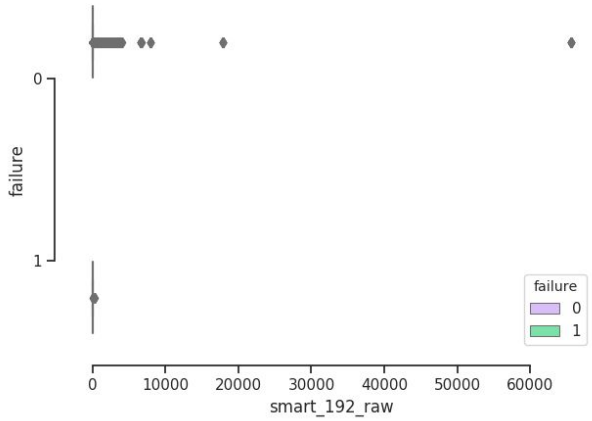


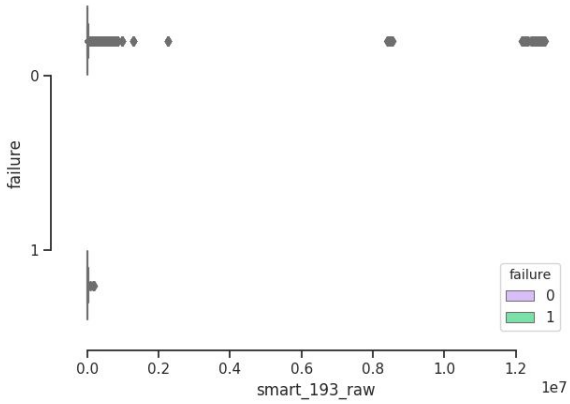
Smart 190 looks to be a promising stat as it not only has 37.9 % missing value but it also shows that it has clearly lower range and median value for failure drives than that for operational ones. Hence we will also draw its violin plot in order to notice its probability density. We can see here that failure has more data distribution for lower values of smart 190 values while for operational drives its opposite.Violin plot = Box plot + density plot with white dot in middle as median. Thick black bar as interquartile range with thin extended line as min and max. The width = frequency in violin plot.

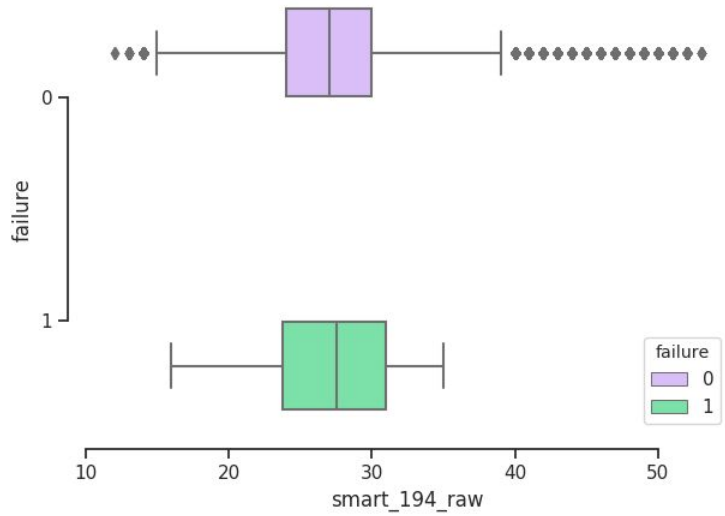


On each side is kernel density estimation to show distribution shape of data. Wider section of violin plot represent higher probability that member of population will take on that value while skinnier section represent lower probability.

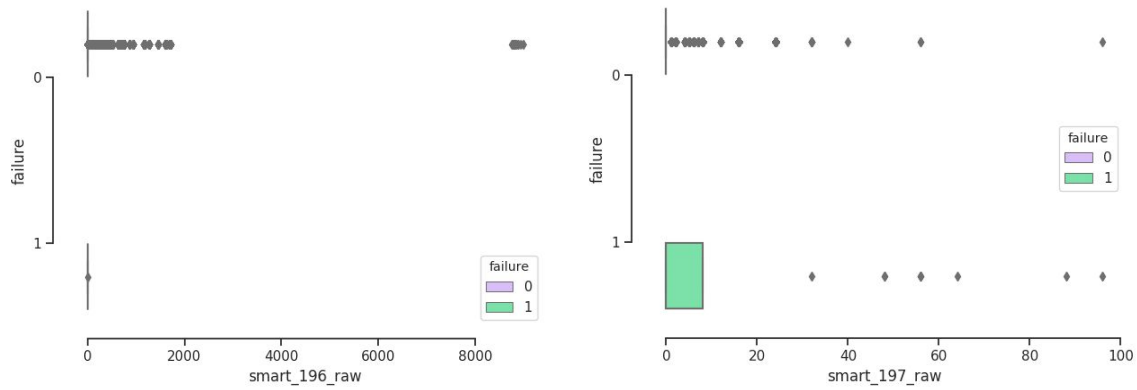




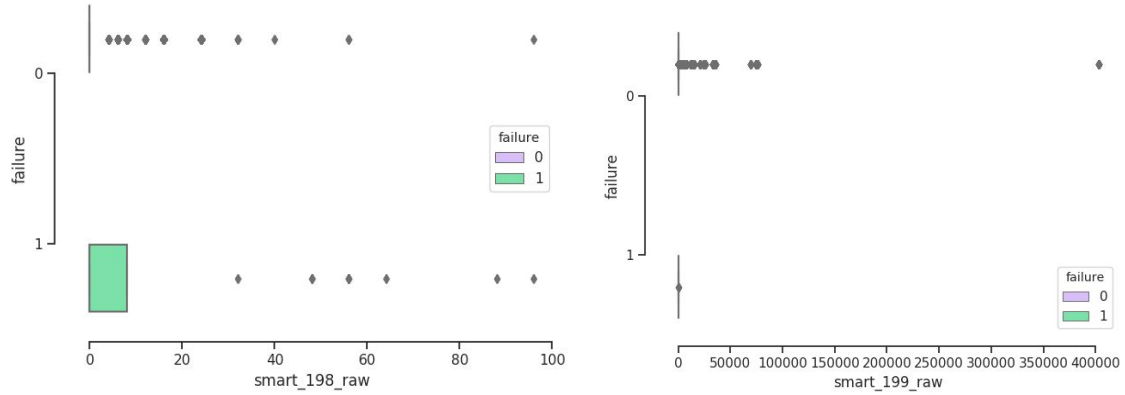




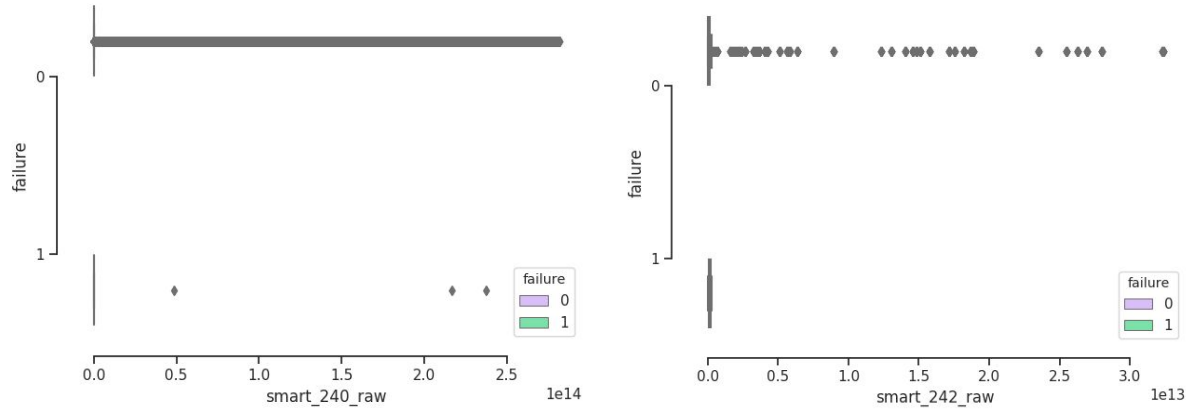
The plot of **Smart 194** does shows that lower value of it is better for not having failure.

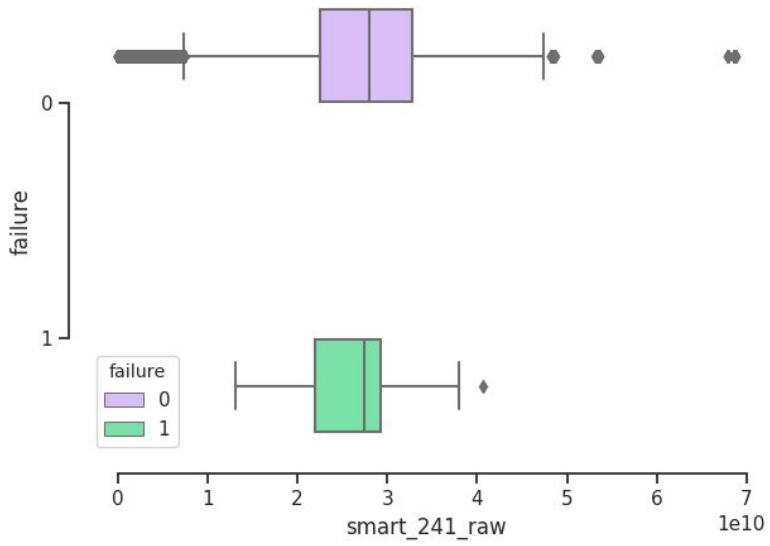


Again for even **Smart 197**, from above whisker plot we can observe that its higher value falls in failure whisker but also more variation is seen in that. And even same is seen in plot of **Smart 198** below. Both Smart 197 and 198 have more outlying values in case of operational drives than in failure.



Even though Smart 199 has 0% missing value it appears to have a lot of outliers and no variation.

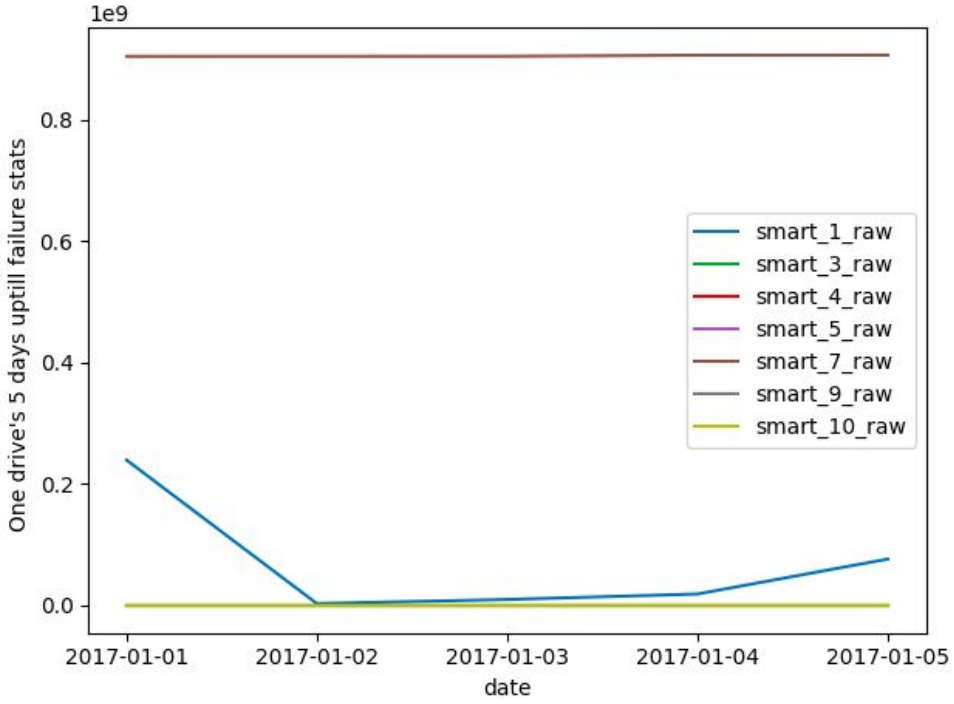




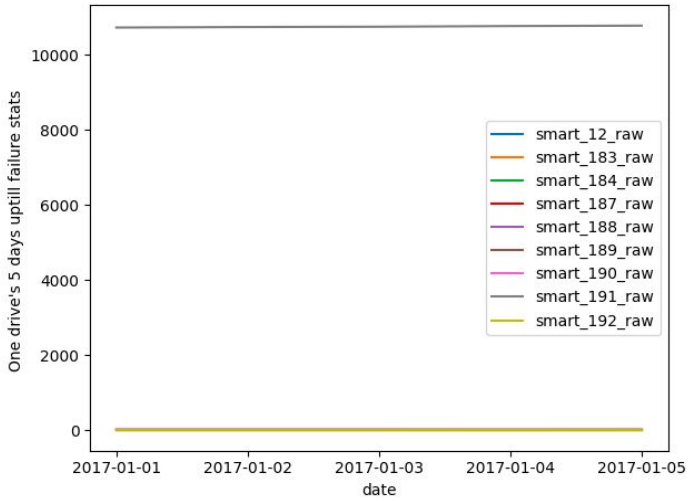
Columns Smart 240, 242 and 242 have emptiness greater than 30% so their whisker plot cannot be of much help. Moreover, from dataframe with missing percent, we found that Smart 240, 241 and 242 have missing percent around 38% and from above plots, Smart 240 and 242 appear to have a lot of outliers.

Analysing per drive

The data was taken up till Jan,5th, 2017 for a drive that failed on that day to observe the behaviour of various smart stats for this particular drive.



On finding what were the values below are the results for above stats
Smart_7 -> 1
Smart_1 -> varies
Smart_3 -> 0
Smart_4 -> 0
Smart_5 -> 0
Smart_8 -> NaN
Smart_2 -> NaN
Smart_9 -> 0
Smart_10 -> 0



Smart_12 ->13

Among Smart stats 1 -12 , none of them show any useful trend of increase or decrease.

Smart_183-> 0

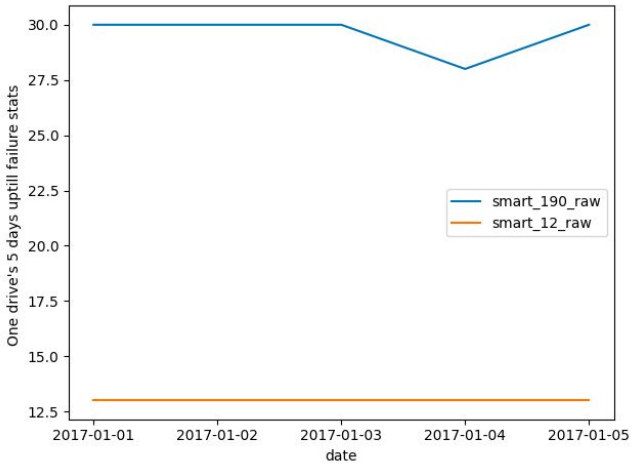
Smart_184 -> 0

Smart_187 -> 0

Smart_188 -> 0

Smart_189 -> 10

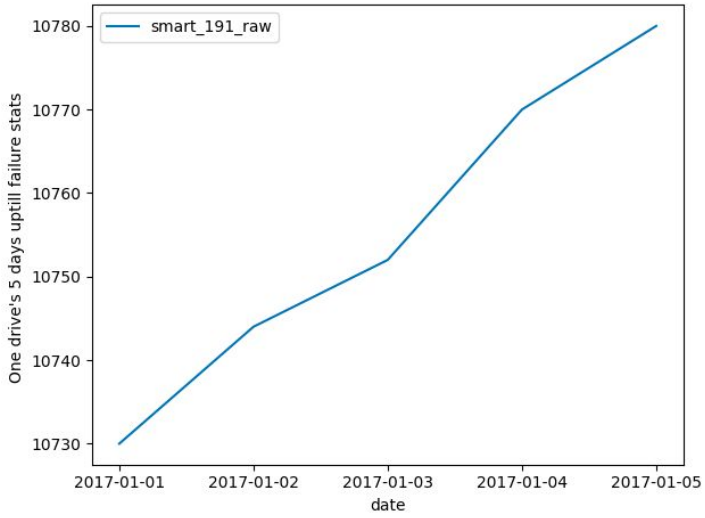
Smart_190 -> 30, 30,30,28,30

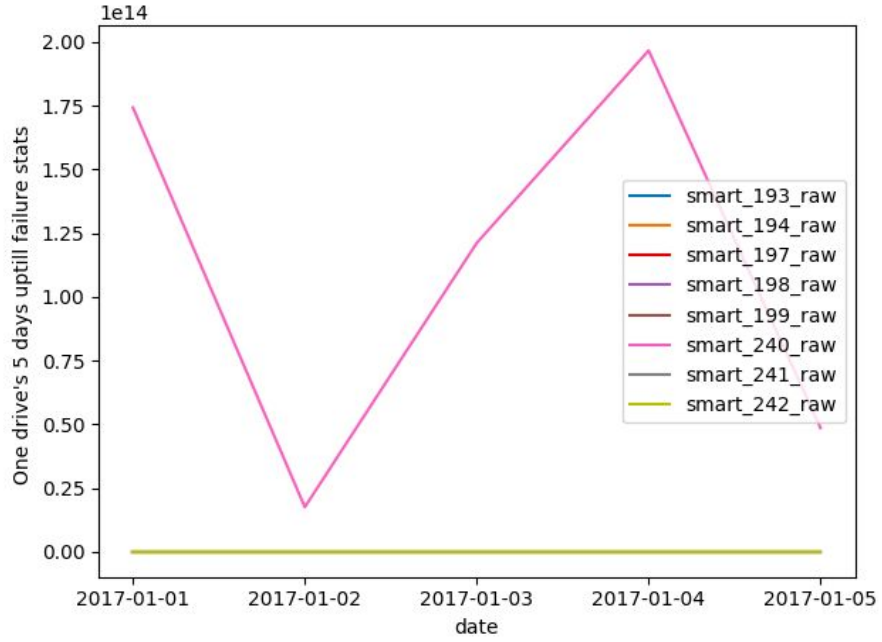


Smart_191 -> 10730 , 10780

Again **190 and 191** can be used as useful features because they have around 37% missing values they do show variation in values before failure. Also their first derivative suggest us important information, with first derivative of Smart 191 is greater than 190 among two plots.

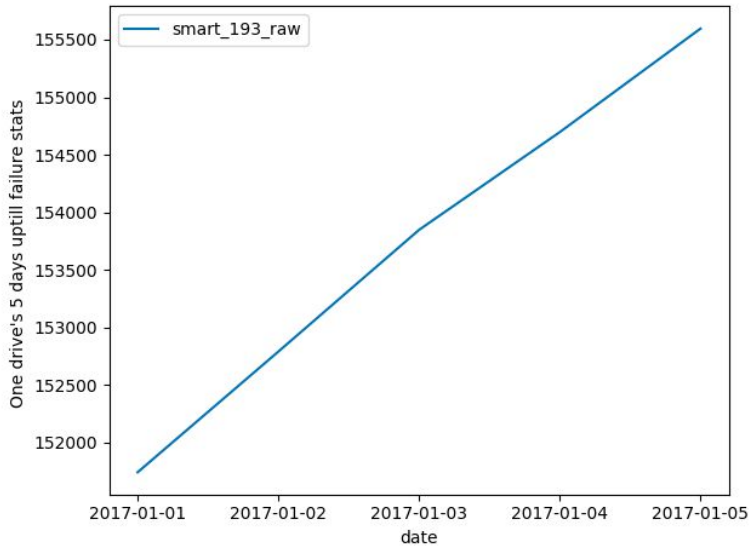
Smart_192 -> 0





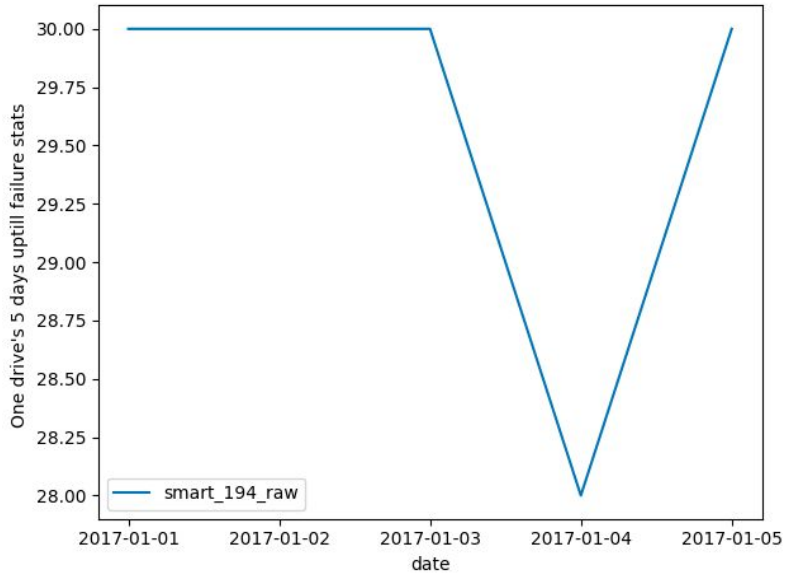
Smart_193 -> 151743, 155596

Smart 193 can be useful as in our data it shows percent emptiness to be less than 30%. As well in this drive's behaviour increase in value when drive approaches to fail on the fifth day.



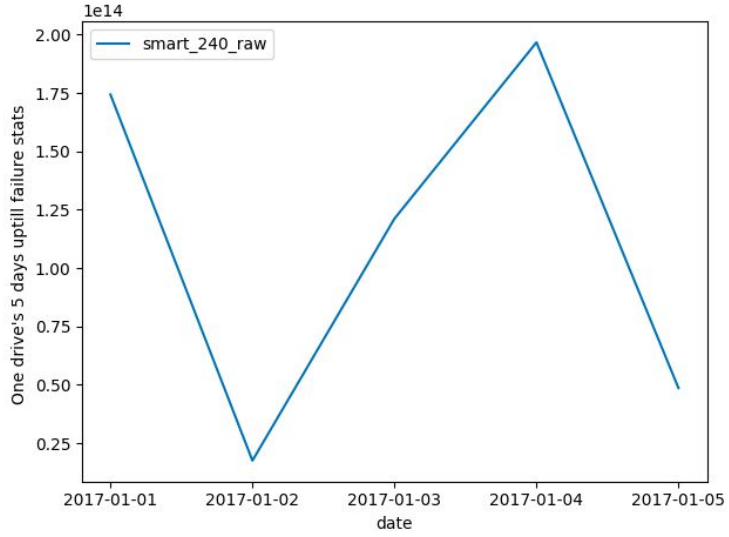
Smart_194 -> 30, 30,30,28,30

Again, due to the same reason of percent emptiness, Smart 194 can be very useful to track as it shows decrease in value just before the day of failure.



Smart_197 -> 0
Smart_198 -> 0
Smart_199 -> 0
Smart_240 ->

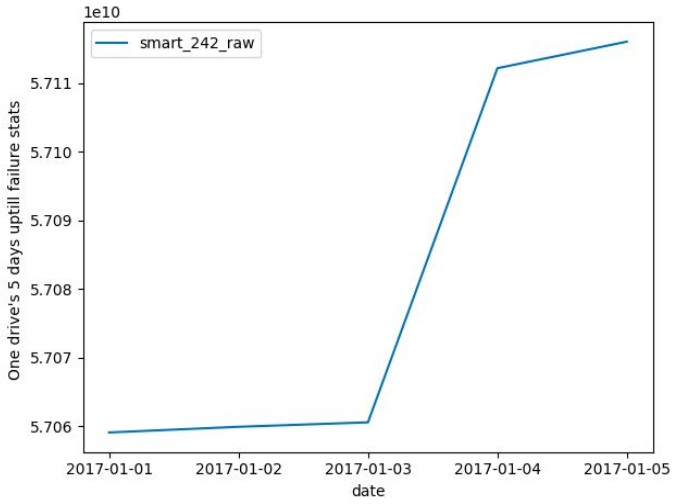
```
20314    1.743155e+14
20314    1.757501e+13
20314    1.211739e+14
20329    1.966666e+14
20329    4.861044e+13
```

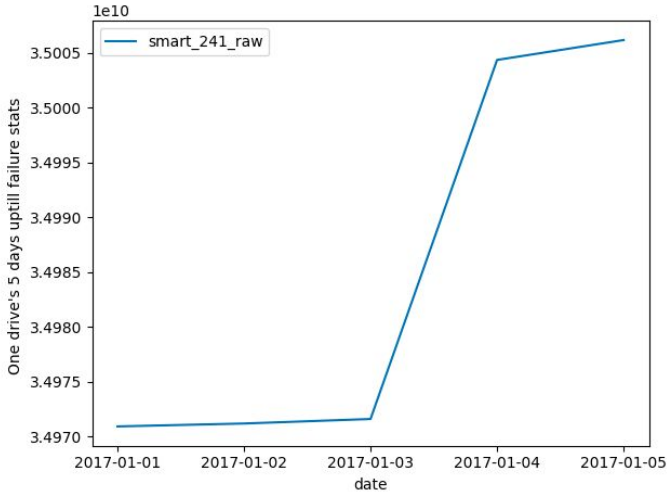


Smart_241 and Smart_242 ->

```
In[140]: f_drive["smart_241_raw"]
Out[140]:
20314    3.497092e+10
20314    3.497118e+10
20314    3.497160e+10
20329    3.500434e+10
20329    3.500615e+10
Name: smart_241_raw, dtype: float64
In[141]: f_drive["smart_242_raw"]
Out[141]:
20314    5.705911e+10
20314    5.705994e+10
20314    5.706059e+10
20329    5.711214e+10
20329    5.711601e+10
```

We can notice that Smart 241, Smart 242 even though show variation it is significant increase in value to be considered important. The first derivative/ increase from day-3 to day-5 in smart 241 is 34550000.0. Similarly for smart 242 it is 55420000.0 from day-3 to day-5. Overlapping with whisker plots we can notice that Smart 241 can be considered useful with 38% missing value and its difference in values is too large between first and day before failure of drive data.

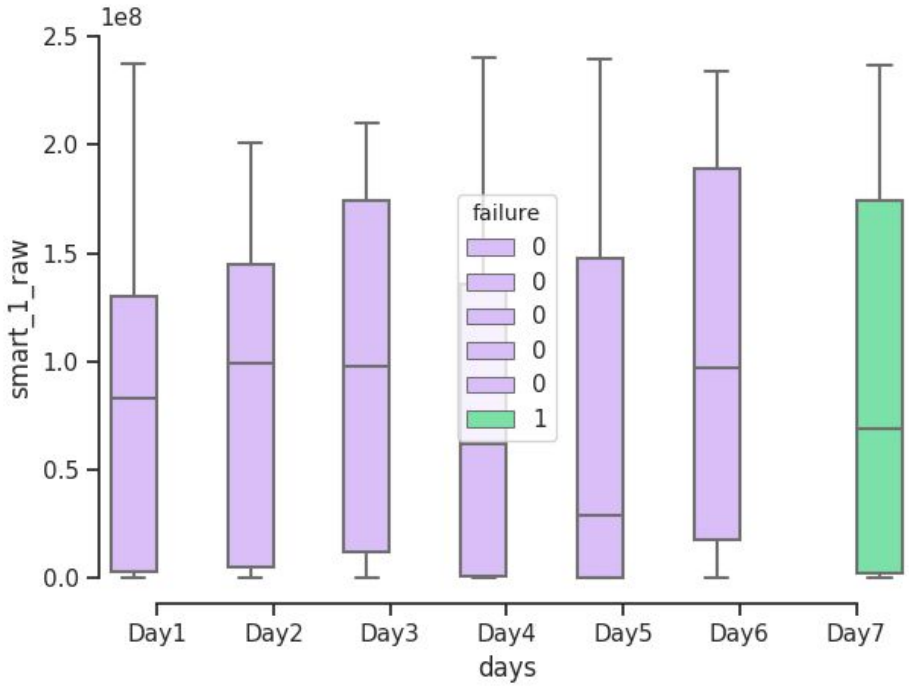


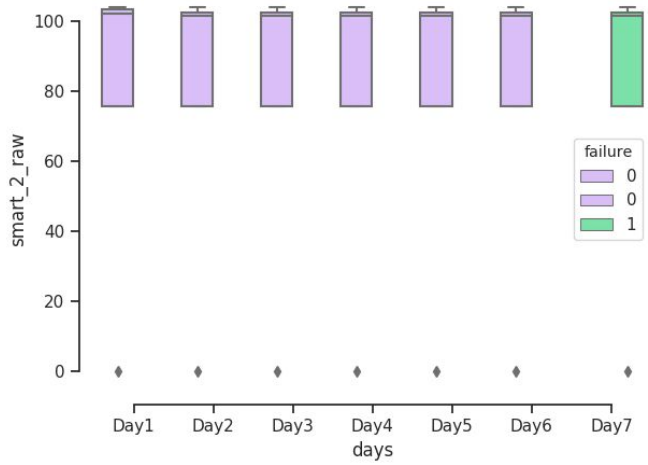


Overlapping results of emptiness of columns, whisker plots and analysis per drive values before failure, we can say that among 0-2 % missing values, column 1, 3, 9 , 190, 194, 197, 198 are important and among columns with 36-40 % missing values, smart stats 190, 191, 240, 241, 242 can still be considered important.

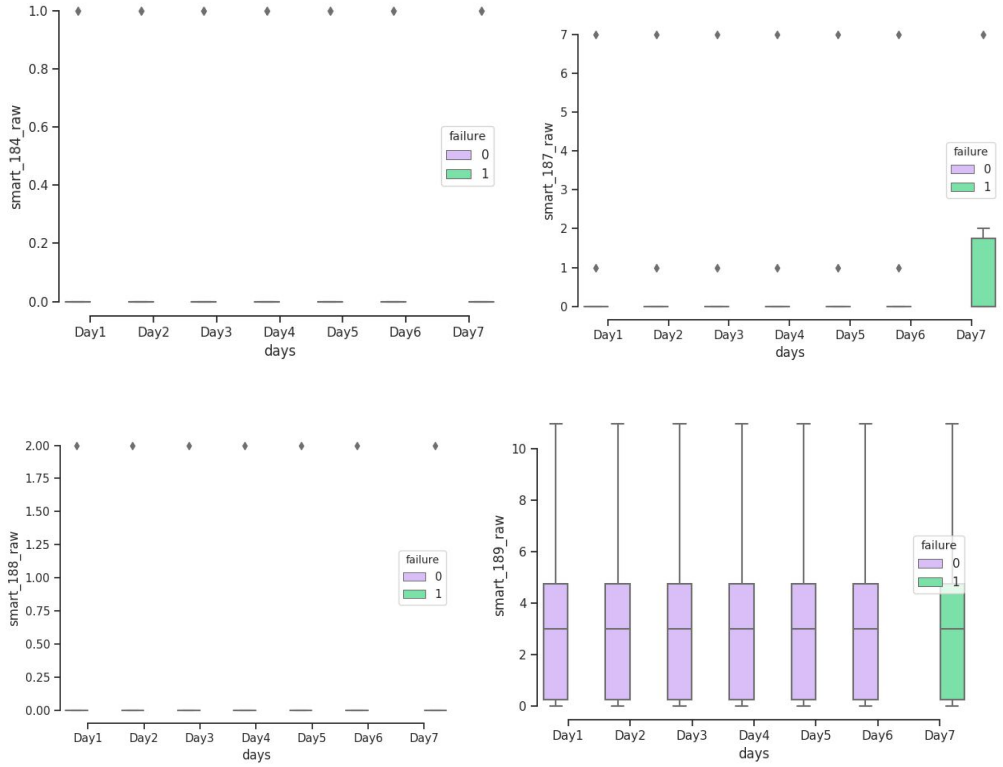
Tracking behaviour for group of drives

Then, drives data were plotted for 14 drives, such that Day -7 is the when these drives fail and stats can be observed for 6 days prior to their failure.



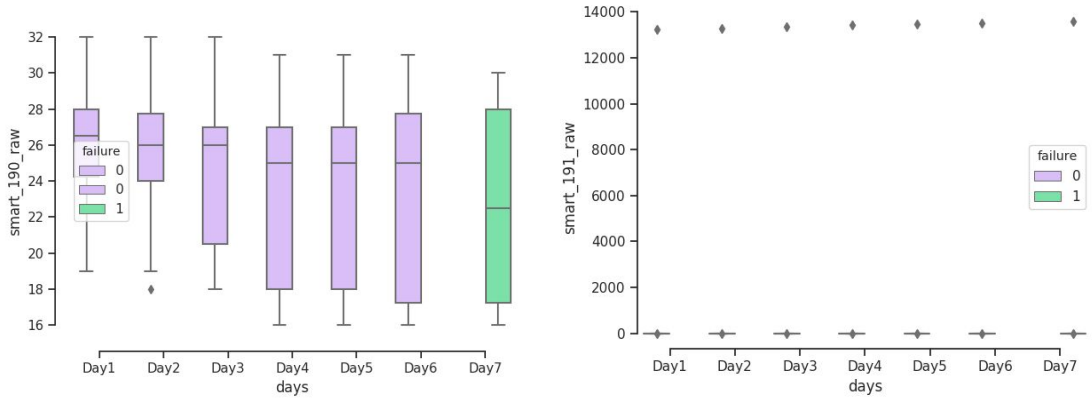


We can notice that values become more variable (or have more variation) from Day-4 to Day-7 and median value goes low.

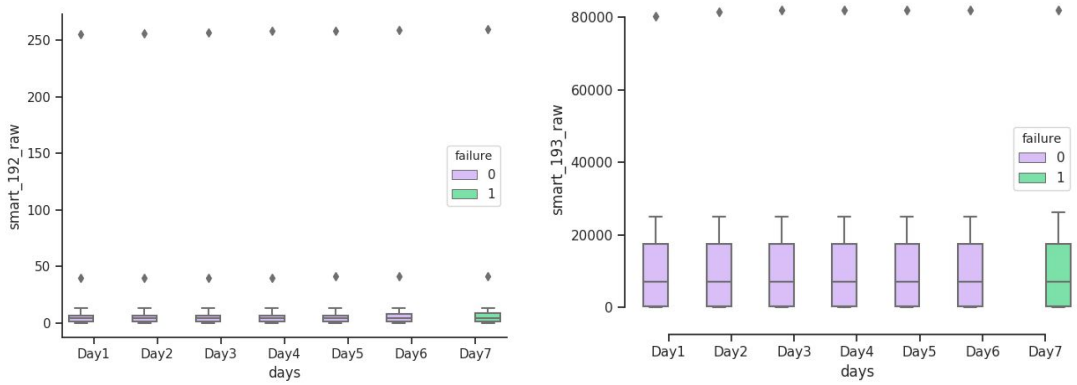


From the plots of Smart 184, 187, 188, 189 we can notice that 184, 188 are following the trend as shown by whisker plots of containing no variation. Moreover, Smart 189 even having variation in value has it constant across all days whether failed or not, hence that also proves to be less relevant. But, Smart 187 having no variation in previous days

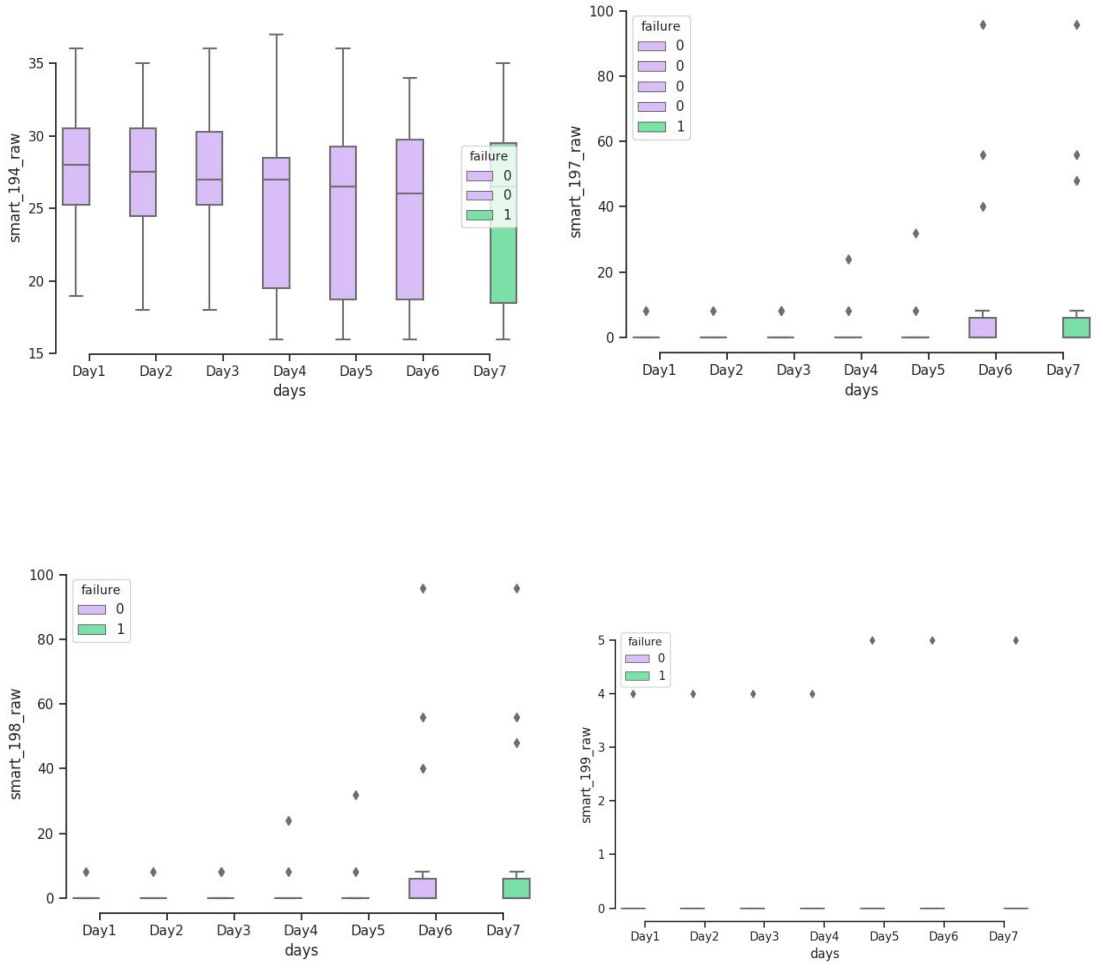
does show variation on day-7 but the whisker plot showed no such indication. From these 4 plots, we may consider Smart 187 to be still thought to be considered among features for models.



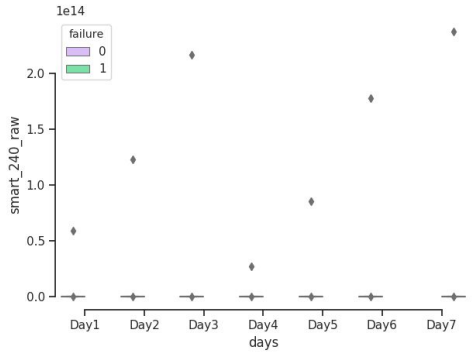
We cannot get much information from plots of Smart 190 and Smart 191 here as in the case of Smart 191, the variation does increases from Day-1 to Day-7 but the median min, max value decreases. For Smart 191, we cannot observe any variation itself in the values.

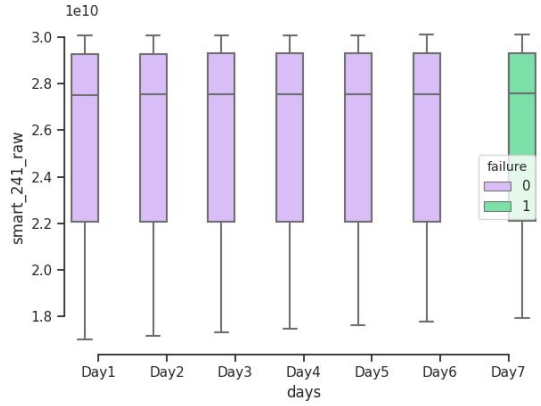


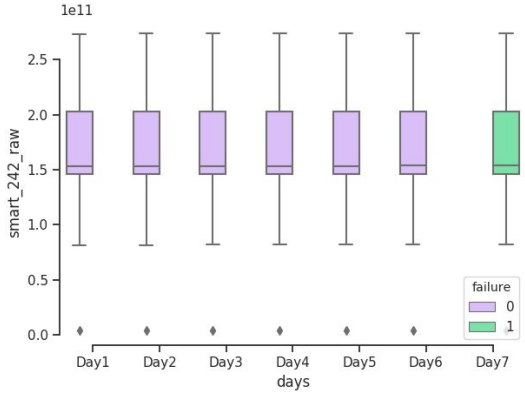
For Smart 192 there is very little variation in values between day-1 to day-7.
For Smart 193, values have equal variation for all days as well almost the same distribution through all the days. Among the two, we can still consider Smart 192.



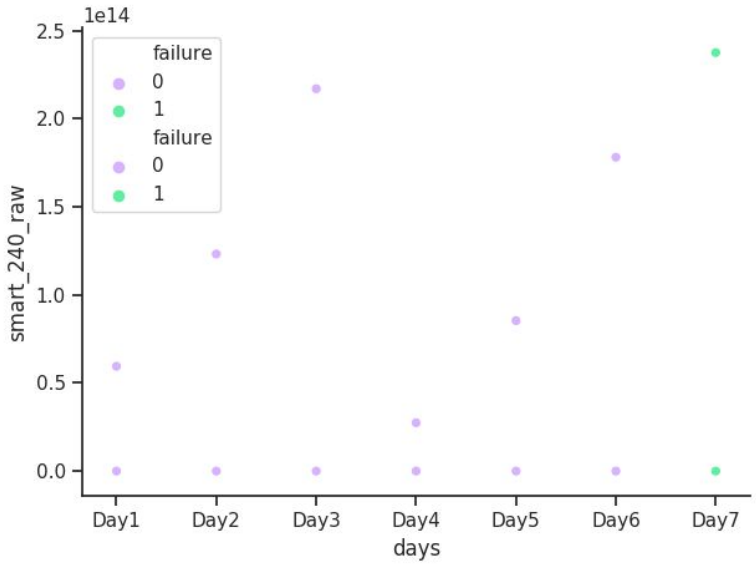
For Smart 194, it does show variation increase from day-1 to day-7 but not very helpful shift in median value can be interpreted. Smart 197 and Smart 198 have very similar plots and have greater variation on day-6 and day-7 which collapses with our results of previous whisker plots and analysis per drive for them to be relevant. Smart 199 continues to prove not at all helpful by showing no variation in whisker plot, analysis per drive as well here for group of drives across all 7 days. Hence among 192, 197, 198 and 199, we can leave 199.

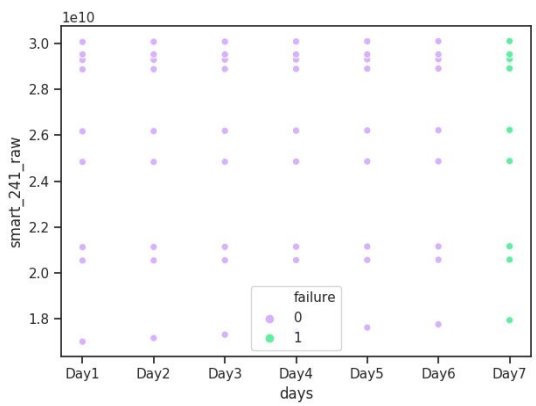


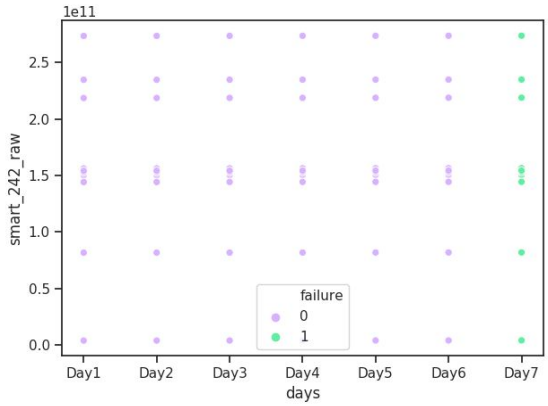




Among Smart 240, 241 and 242 from above plots none looks important as in each either there is constant variation across all days or there is no variation at all with outliers. Thus we can also notice scatterplot of values for these smart stats in case we miss anything from above plots.







We can notice from scatter plot show exact similar result as whisker plots of three stats not being helpful.

Relevant Features from EDA

Therefore, from above data exploration till now relevant ones, we have **8 columns** for baseline model -
Smart 1,
Smart 3,
Smart 9,
Smart 187,
Smart 192,
Smart 194,
Smart 197,
Smart 198
Among these Smart 187 can be considered later as it contains more missing percent than others. With maybe important or 5 columns can be considered later with smart ways of imputation -
Smart 190,
Smart 191,
Smart 240,
Smart 241,
Smart 242

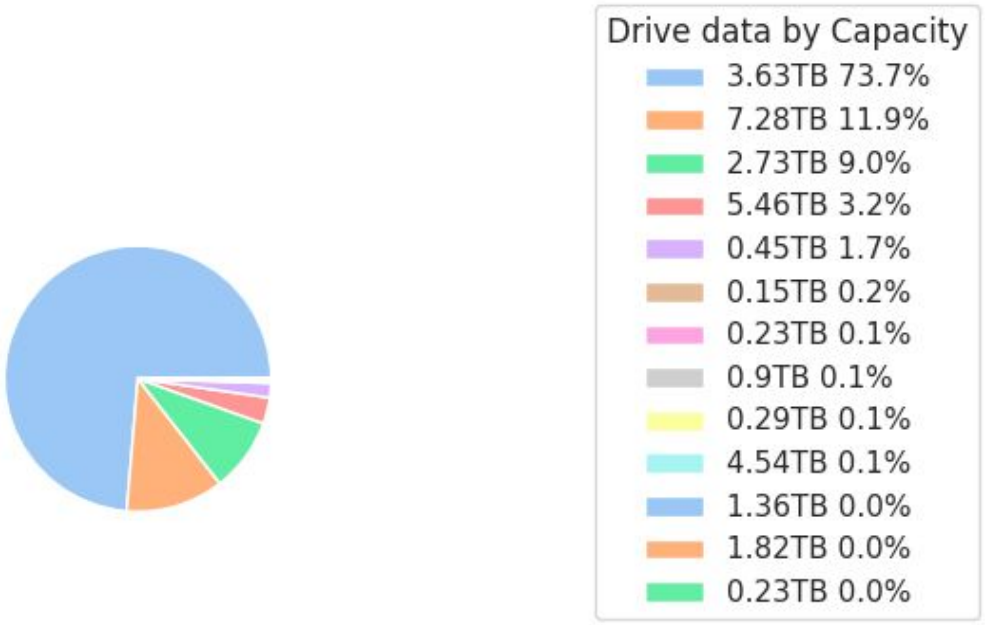
Therefore baseline supervised models can initially use only **7 Smart raw columns as below :**

Smart 1 Read Error Rate
Smart 3 Spin Up Time
Smart 9 Power On Hours
Smart 192 Unsafe Shutdown Count
Smart 194 Temperature or Temperature Celsius
Smart 197 Current Pending Sector Count
Smart 198 Uncorrectable Sector Count

Are they different type of drives ?

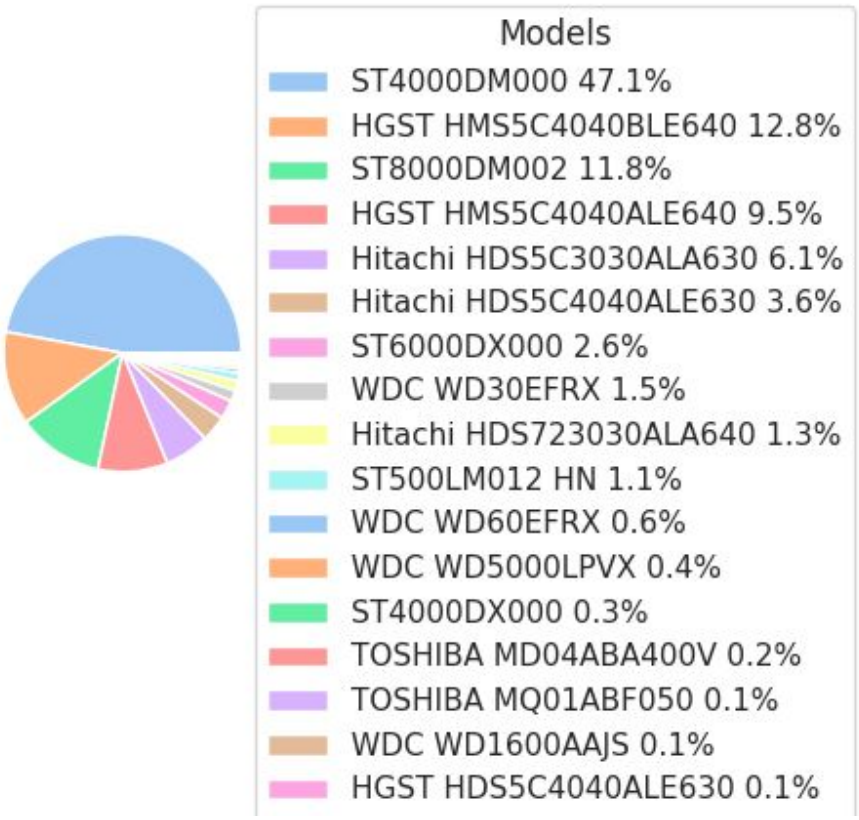
Percentage of various types of capacities in drive data

Capacity bytes of drives - 2.73TB, 3.63 TB, 7.28TB, 0.45TB, 5.46TB, 4.54TB, 0.15 TB, 0.29TB, 0.23TB,1.36TB,0.9TB, 0.23TB,1.82TB



```
In[107]: yr_2017_Q1_data["capacity_
Out[107]:
4000787030016    814070
8001563222016    131475
3000592982016     99240
6001175126016     35100
500107862016      19278
160041885696       2045
250059350016       1410
320072933376        675
5000981078016       675
1500301910016       615
1000204886016       510
2000398934016        45
250000000000         15
Name: capacity_bytes, dtype: int64
```

Percentage of various types of models in drive data



The above two pie charts are created using drive data of first 15 days of year 2017 which contains around 1 million records.