UNIVERSITY IN NOVI SAD

FACULTY OF SCIENCES

DEPARTMENT FOR MATHEMATICS AND INFORMATICS

A REPORT FOR BIG DATA IN MEDICINE AND BIOLOGY

Analysis of Visaris statistical data

Professor:

dr Vladimir Petrović

Student:

Nevena Đilas

Novi Sad, April 2023

Introduction

Visaris statistical data is data obtained by an X-ray scanner. It contains information about the time and date when the scan was performed, as well as the anatomy that was scanned, X-ray exposure factors (kVp, mAs, mA, ms), radiation exposure index (EI) and was the obtained image valid or not (rejected).

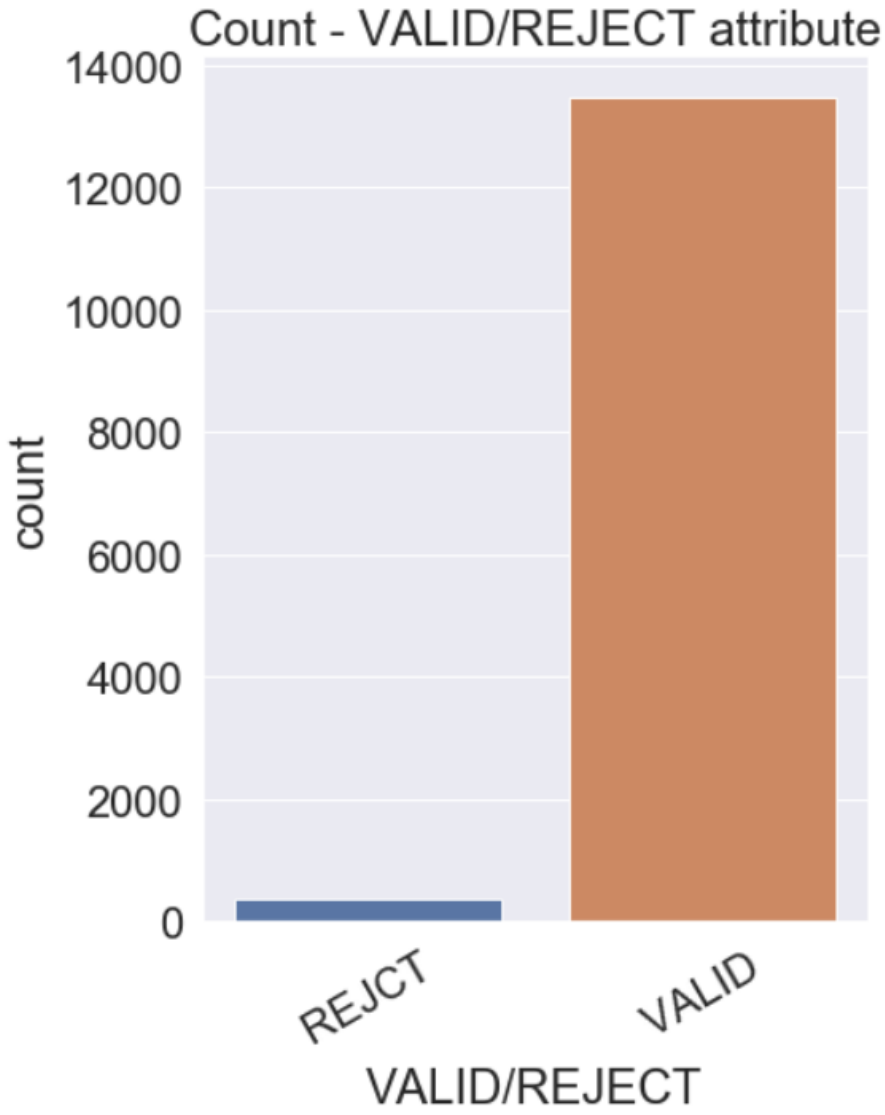
- kVp stands for kilovoltage peak. This is the highest voltage (measured in thousands of volts) that will be produced by the X-ray beam during an exposure.
- mAs stands for milli-ampere-second. It determines how many x-ray photons are produced by the X-ray tube at the setting selected (quantity of X-rays). It has no effect on the strength (penetrating power) of the X-ray photons. mAs is a product of multiplying two factors together: time and milliamperage (mA).
- The exposure index (EI) value is a method of monitoring radiation dose in computed radiography (CR)

Some of the questions that we would like to find an answer to are:

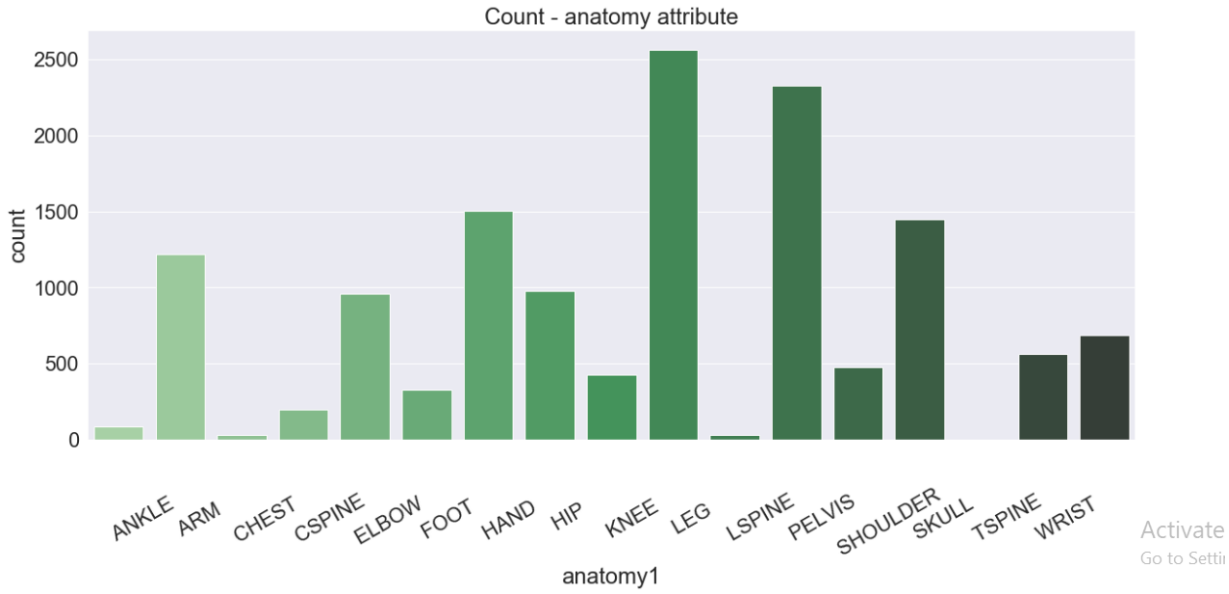
- According to anatomy, how does the exposure index vary on rejected as opposed to valid images?
- What are X-ray exposure factors kVp and mAs for images with the largest EI?
- How is EI changing according to the time of day?

Visualization and description of categorical attributes

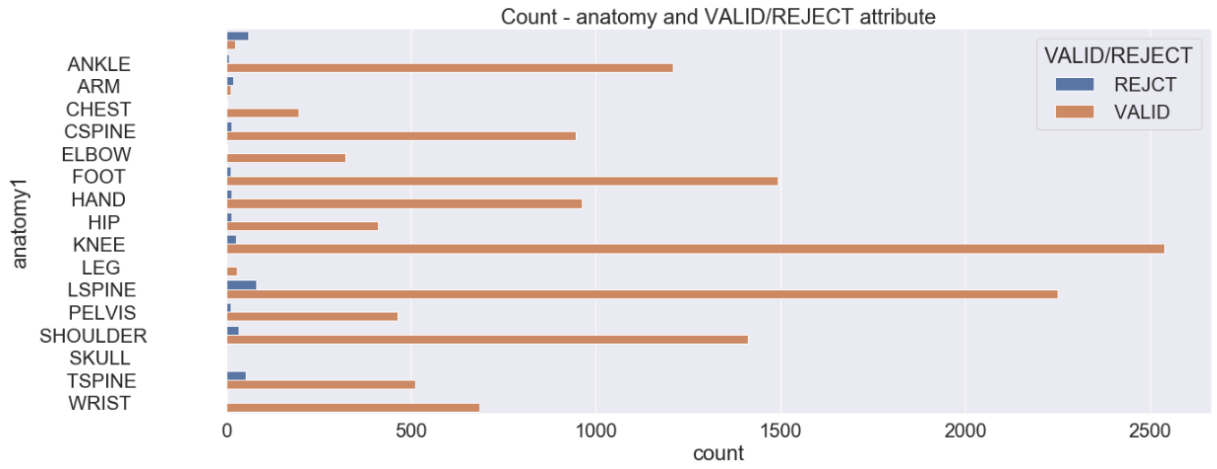
Data consists of 13832 rows, each one of them corresponding to one image obtained with X-ray scanner. Most of the images, 13469, are valid and 363 are rejected.



The highest number of images are scans of knees (2564) and lower spine (2330) followed by these anatomies: foot (1504), shoulder (1446), ankle (1219), hand (978), cervical spine (962), wrist (689), thoracic spine (565), pelvis (477), hip (426), elbow (326), chest (199). Then we have 84 images whose anatomy is unknown. Arm(33), leg (29) and skull (1) are the anatomies with the least samples.



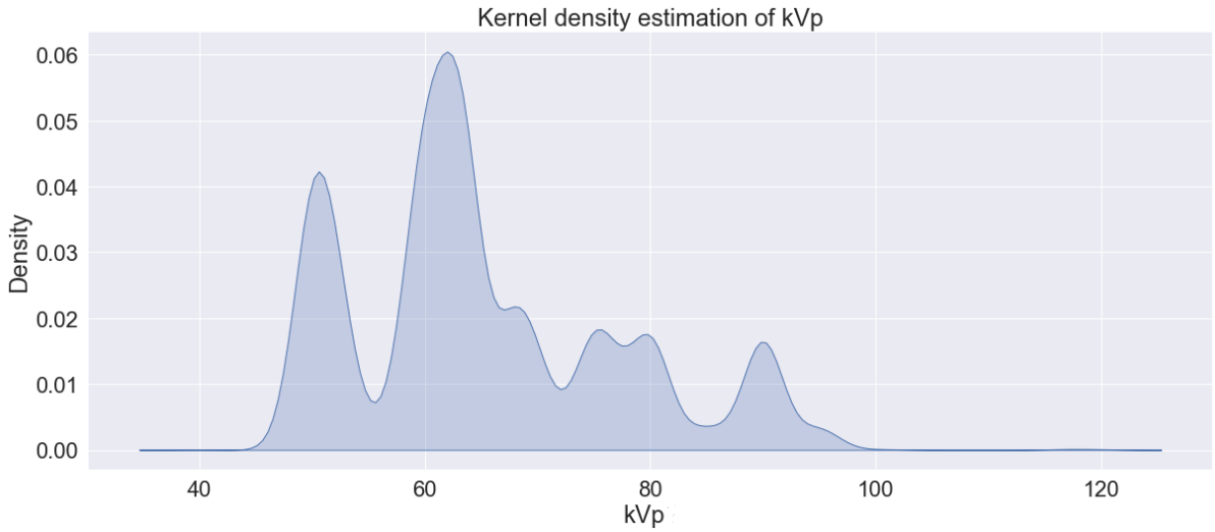
We can also see the valid/rejected scans by anatomy:



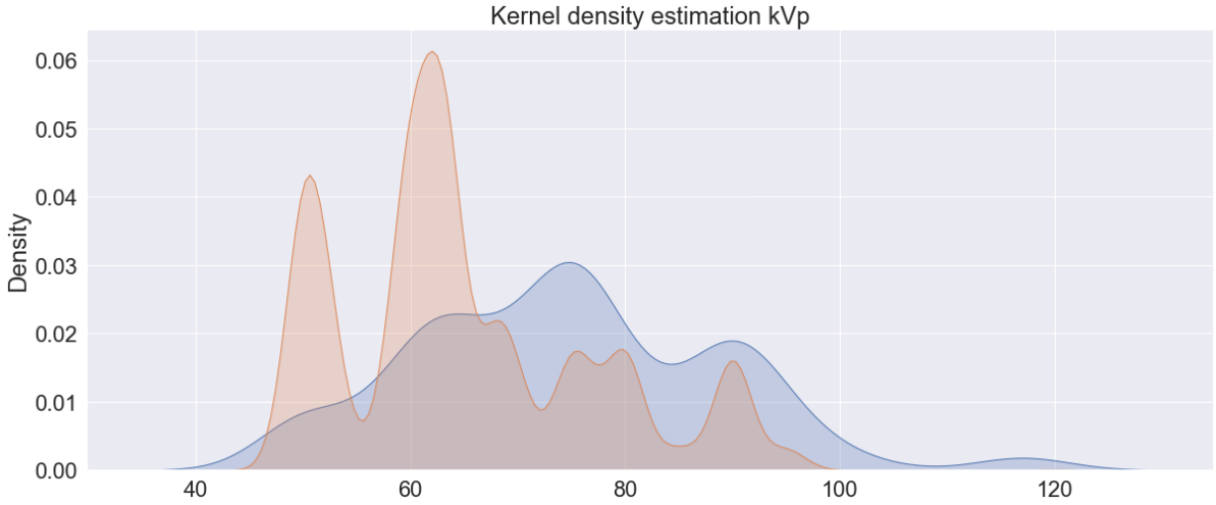
It is important to note that for leg and skull we don't have rejected samples, so it would be impossible to run any test statistics for those anatomies. For most anatomies there are far more valid then rejected samples like, for example, ankle has 1210 valid and 9 rejected samples and knee has 2538 valid and 26 rejected samples. More rejected than valid samples have only arm (12 valid and 21 rejected) and the unknown anatomies (24 valid and 60 rejected)

Visualization and description of X-ray exposure factors and dose

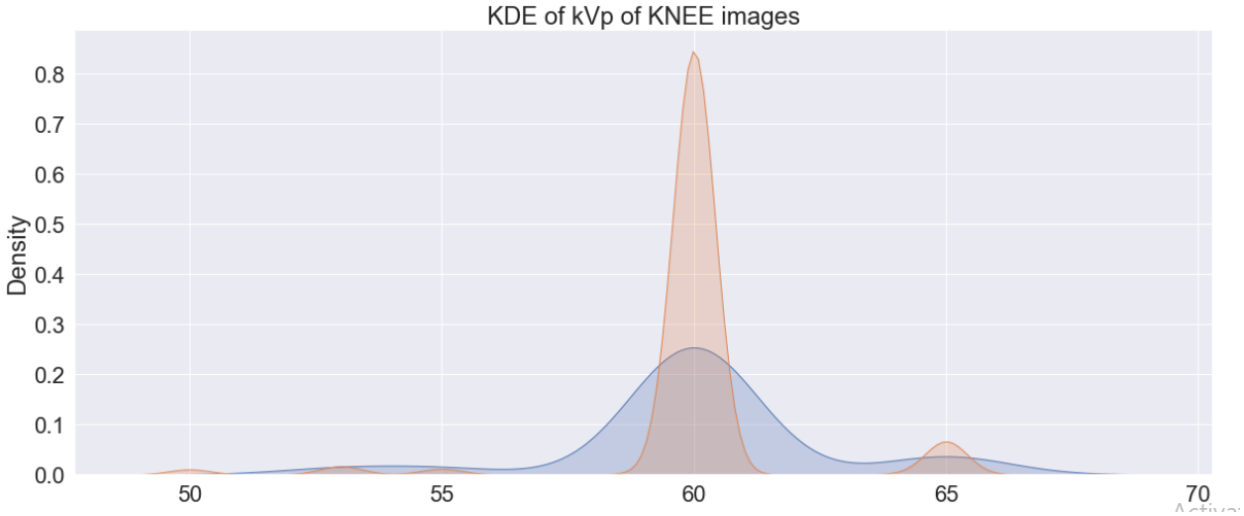
KVp (kilovoltage peak) is in interval [40, 120]. It's mean value is 65.37883169462117. Most of the data (the middle half of the data between 1st and 3rd quartile) is in interval [60.0, 75.0]. Interquartile range is 15. Distribution of kVp is estimated using KDE (kernel density estimation with Gaussian kernel) and it is shown on the graph below:



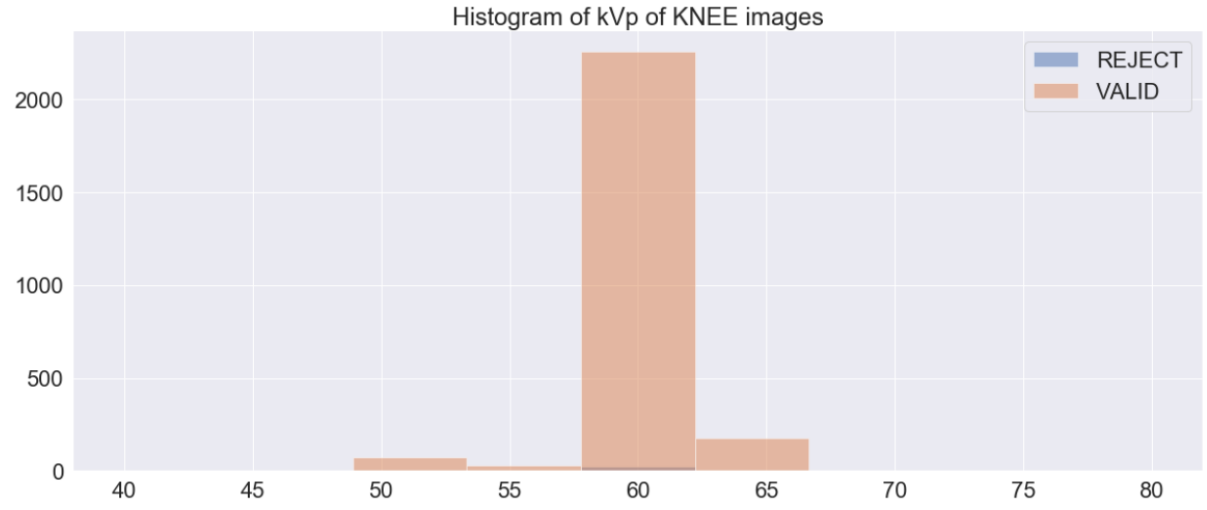
If we divide the data for kVp on rejected and valid samples and then look on the estimated distributions of those two groups, we can see that KDE for valid samples is very similar to the KDE estimated for all kVp compared to the group that is rejected. (valid is red and reject is blue):

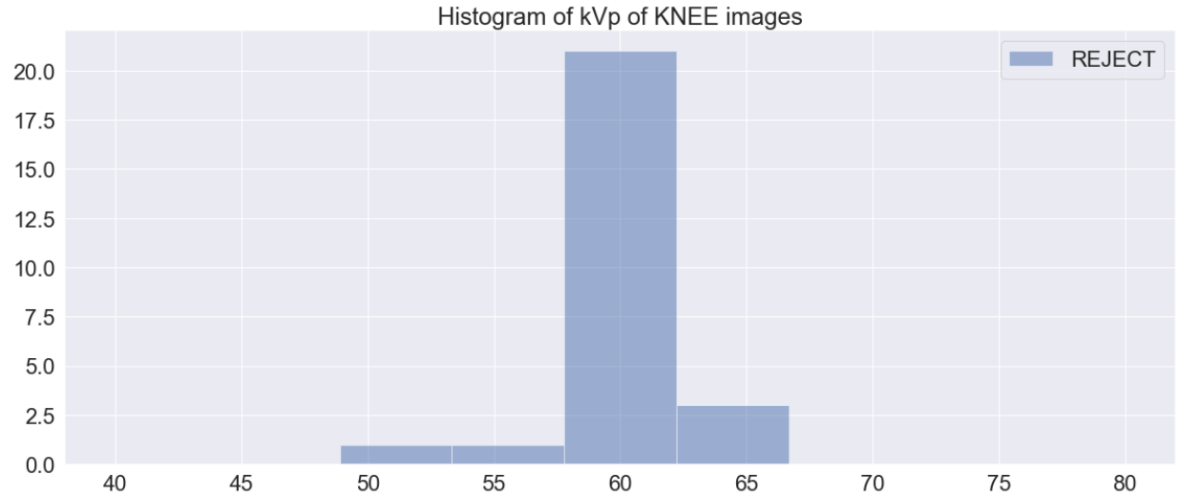


However, since the values for kVp are different for every anatomy, the best way to see if there are big differences between kVp for valid and rejected samples is to sort them by anatomy. Graph shown below is for knees (it's the most accurate because of large number of samples):

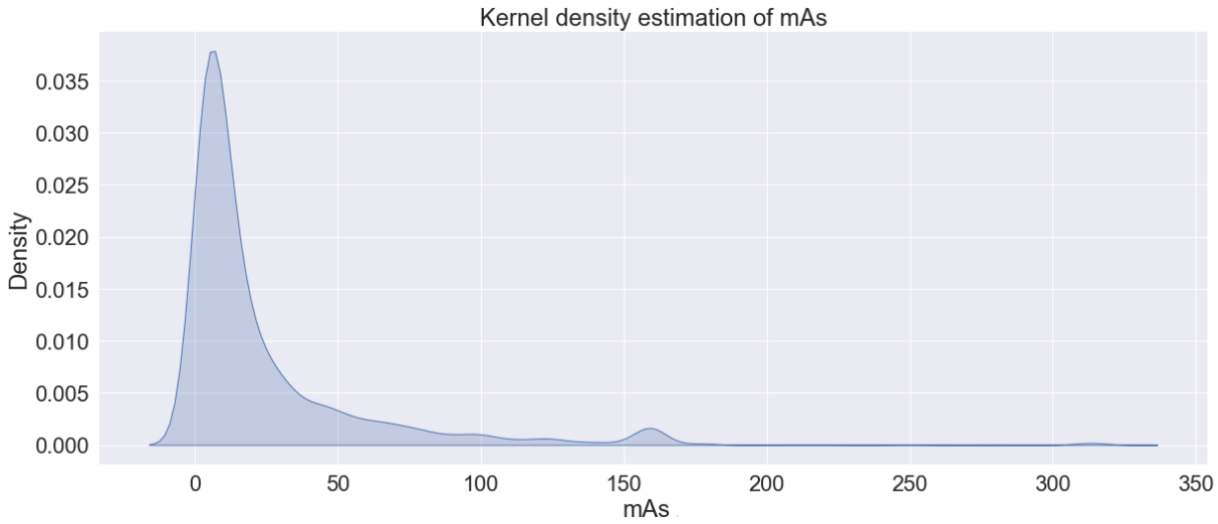


It looks like rejected images have the bigger dispersion, but that is mostly a consequence of the smaller number of samples. That can be better spotted on histogram:

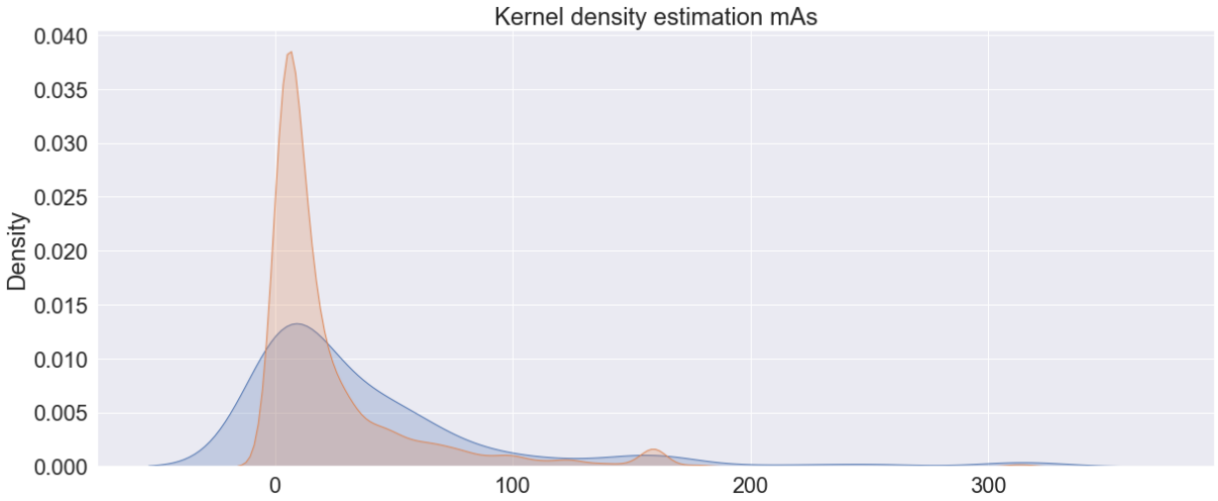




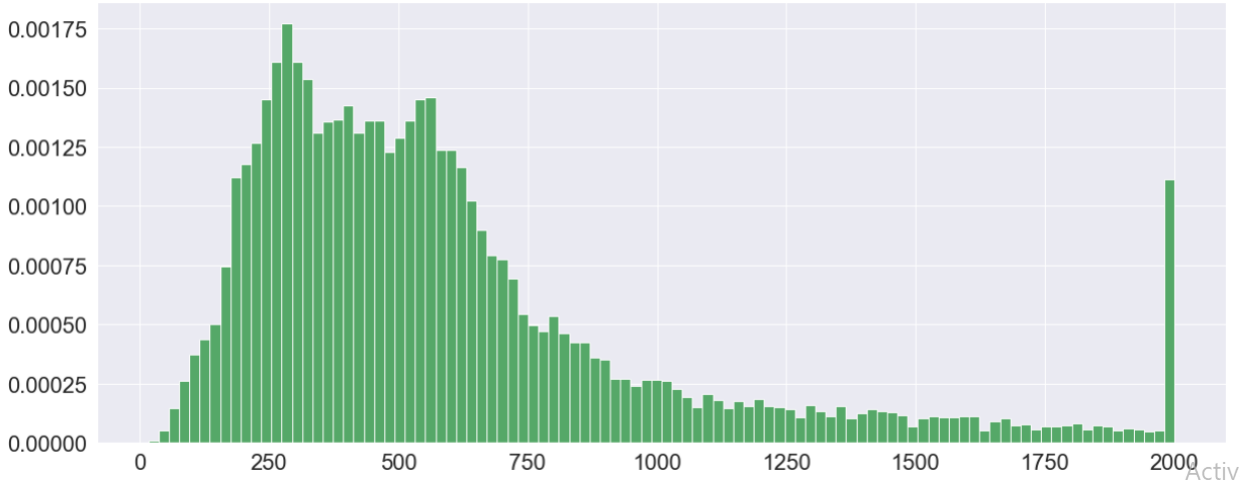
MAs (milli-ampere-second) is in interval [0.56, 320.0]. Mean value of mAs is 24.900925390399074. Data between 1st and 3rd quartile: [4.21, 27.5]. Interquartile range: 23.29. It's distribution is estimated using KDE:



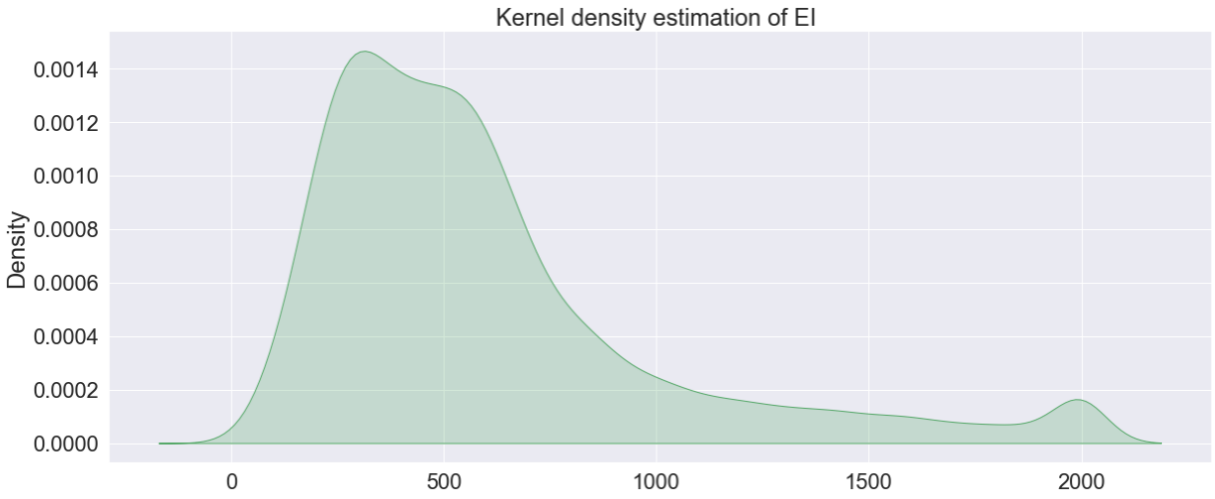
KDE for mAs divided by two groups – valid (red) and rejected (blue):



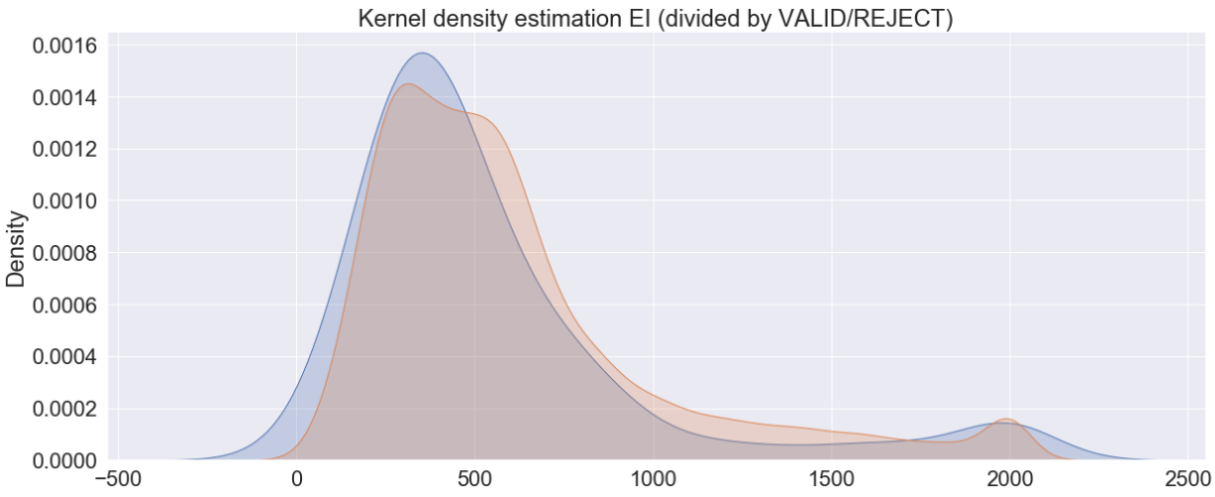
EI (Exposure index, measure of the dose of radiation that patient recives) is in interval [16.4859011769295, 2000.0]. Mean value for EI is 604.339268138448. Interval containing half of the data is interval between 1st and 3rd quartile, [314.8173928260805, 718.7985777854923]. Interquartile range is 403.9811849594118. Histogram of EI:



KDE of EI:



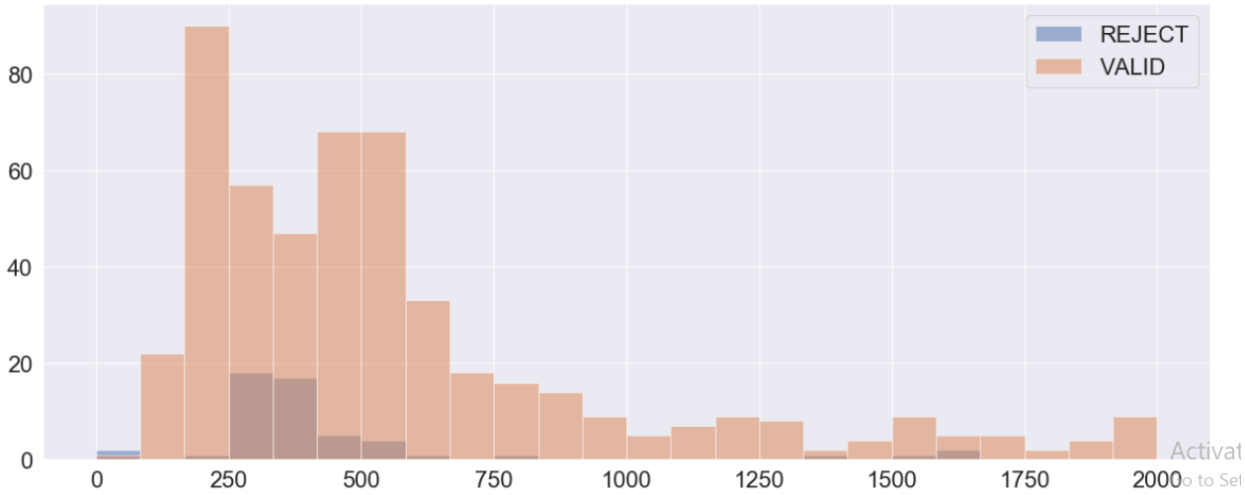
KDE of EI for valid (red) and rejected (blue) samples:



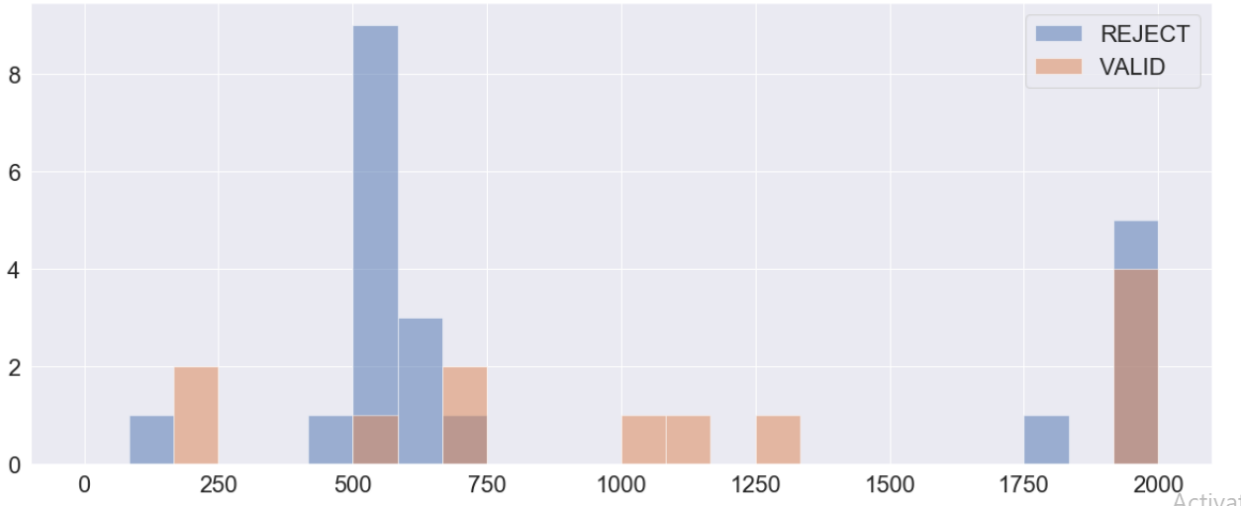
EI on different anatomies divided according to the validation

Let's first see the histograms of EI on some anatomies divided on valid/rejected samples:

```
(' TSPINE              ', ' REJCT ')
(' TSPINE              ', ' VALID ')
```




```
(' ARM                 ', ' REJCT ')
(' ARM                 ', ' VALID ')
```



We would like to know are there significant differences between valid and rejected samples. Since we have two groups of samples that don't overlap we can use a two sample z-test or a two sample t-test. The difference between them is that for z-test we need to know the standard deviation or variance of the population (or we can aproximate it with the sample variance for a sample with large sample size) and for t-test it isn't needed, since t-test uses sample variance. For samples that are large, with sample size bigger then 30, both tests yield similar results. The difference between the z score and t score is negligible. So, since the number

of rejected samples is less then 30 for many anatomies I used two sample t-test on all of them. I also performed z-test on those with bigger number of samples in both groups (lower spine, shoulder, thoracic spine and unknown anatomy).

The null hypothesis is that 2 independent samples (valid and rejected) have identical average (expected) values. T-test assumes that the populations have identical variances by default. To test if the expected values are not identical, we set α value (significance level, the probability of rejecting the null hypothesis when it is true) to be 0.05. This means that there is a 5% risk of making the wrong conclusion. If p-value is less than or equal to the significance level 0.05, then we have sufficient evidence to reject the null hypothesis.

Here are the test results:

```
T-test: Ttest_indResult(statistic=-1.3429891500888154, pvalue=0.1829809969
8981246)

Z-test:(-1.3429891500888154, 0.1792754871196537) (test statistic, p-value)

 ANKLE
T-test: Ttest_indResult(statistic=1.347658812985604, pvalue=0.178018963843
3711)

 ARM
T-test: Ttest_indResult(statistic=0.8428412034469935, pvalue=0.40577353755
21486)

 CHEST
T-test: Ttest_indResult(statistic=1.079908047969275, pvalue=0.281503466463
29855)

 CSPINE
T-test: Ttest_indResult(statistic=0.10765422429538393, pvalue=0.9142924763
180287)

 ELBOW
T-test: Ttest_indResult(statistic=0.9195014887919398, pvalue=0.35851737688
910745)

 FOOT
T-test: Ttest_indResult(statistic=-0.22172148831828384, pvalue=0.824560841
4508426)

 HAND
T-test: Ttest_indResult(statistic=-0.8611469522725997, pvalue=0.3893686167
519005)

 HIP
T-test: Ttest_indResult(statistic=0.6537987031095991, pvalue=0.51359591927
28178)

 KNEE
```

```
T-test: Ttest_indResult(statistic=0.6527946873292986, pvalue=0.5139472189201811)

 LEG
T-test: Ttest_indResult(statistic=nan, pvalue=nan)

 LSPINE
T-test: Ttest_indResult(statistic=-0.5136056986279987, pvalue=0.6075764352351509)

Z-test: (-0.5136056986279987, 0.6075276945710196) (test statistic, p-value)

 PELVIS
T-test: Ttest_indResult(statistic=0.010780641175109486, pvalue=0.9914029857488563)

 SHOULDER
T-test: Ttest_indResult(statistic=0.17495442156661184, pvalue=0.8611399745462298)

Z-test:(0.17495442156661187, 0.8611154474381499) (test statistic, p-value)

 SKULL
T-test: Ttest_indResult(statistic=nan, pvalue=nan)

 TSPINE
T-test: Ttest_indResult(statistic=2.093698416907239, pvalue=0.03673356749158495)

Z-test: (2.093698416907239, 0.03628685405412929) (test statistic, p-value)

 WRIST
T-test: Ttest_indResult(statistic=0.830260274605334, pvalue=0.40668006551593516)
```

Here we see that we have sufficient evidence to reject the null hypothesis only for thoracic spine. We conclude that there is not enough evidence to prove that the expected value of EI for rejected samples differs from expected value of EI for valid samples for all other anatomies.

If we take values of EI for all samples (not divided by anatomy) and perform t-test on valid and rejected samples, this is the result:

```
Ttest_indResult(statistic=2.484028826876062, pvalue=0.013002276686998165)
```

Here, the conclusion is that there is enough evidence to reject the hypothesis that states that rejected and valid samples have the same mean value because $p < 0.05$. However, since values of EI differ for each anatomy, the results conducted by anatomy are more valuable and in most cases there wasn't enough evidence to prove that the expected value of valid and rejected samples are significantly different.

X-ray exposure factors kVp and mAs on images with large EI

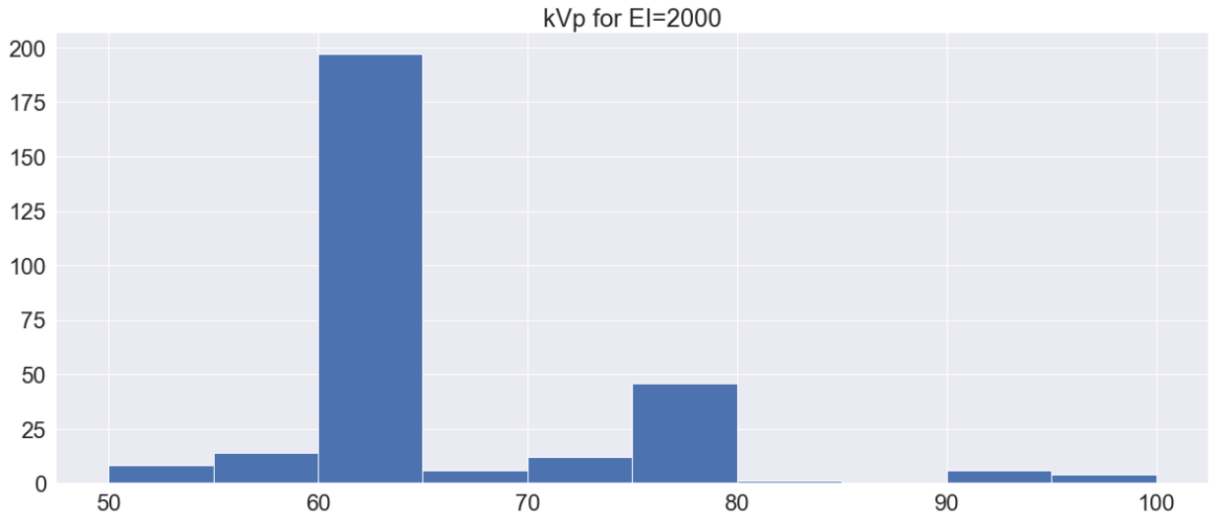
Largest value of EI is 2000 and in our dataset there are 294 images with that value. Let's look at kVp and mAs on images with EI $= 2000$:

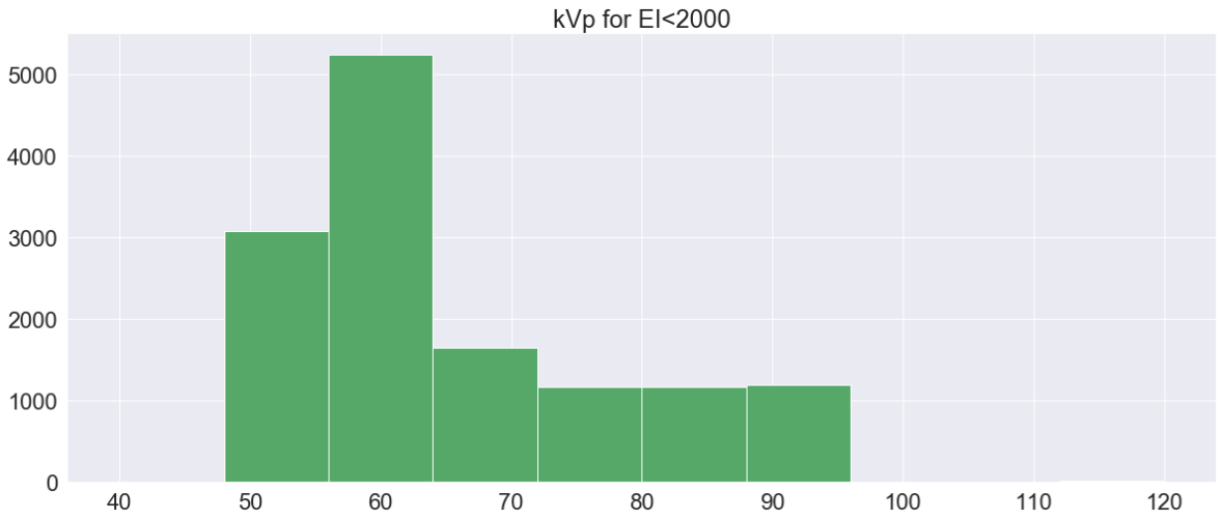
```
KVp and mAs where EI=2000 have these properties:
Range kVp: [ 50 , 100 ]
Range mAs: [ 1.56 , 221.1 ]
Mean value for kVp: 65.37414965986395
Mean value for mAs: 26.603537414965988
```

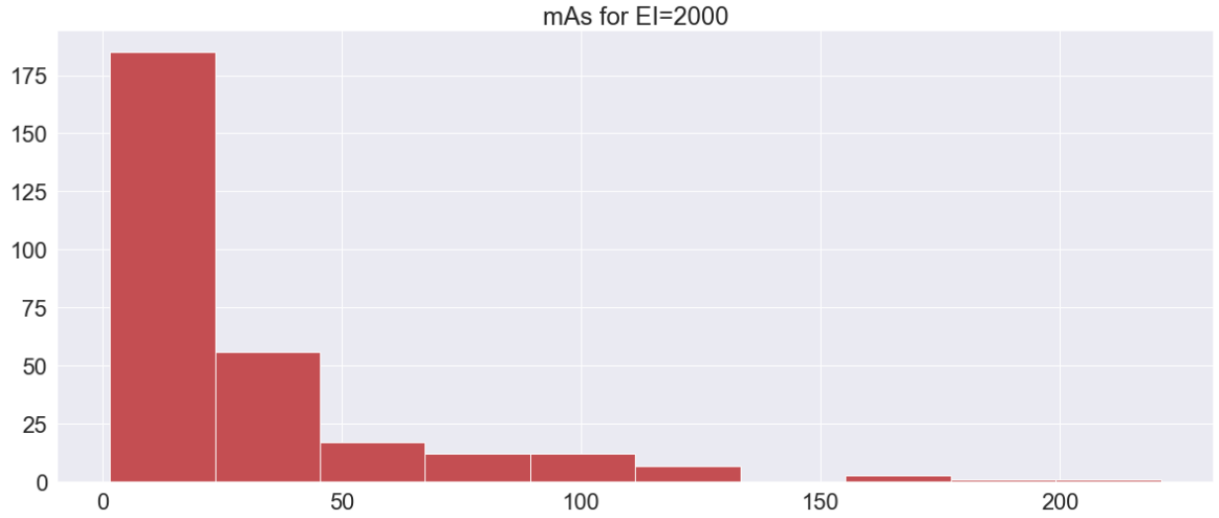
Compared to kVp and mAs for all samples, we see that mean value for kVp is almost the same and that mean value for mAs for EI $= 2000$ is slightly higher then mAs for all samples: $26.603537414965988 > 24.900925390399074$.

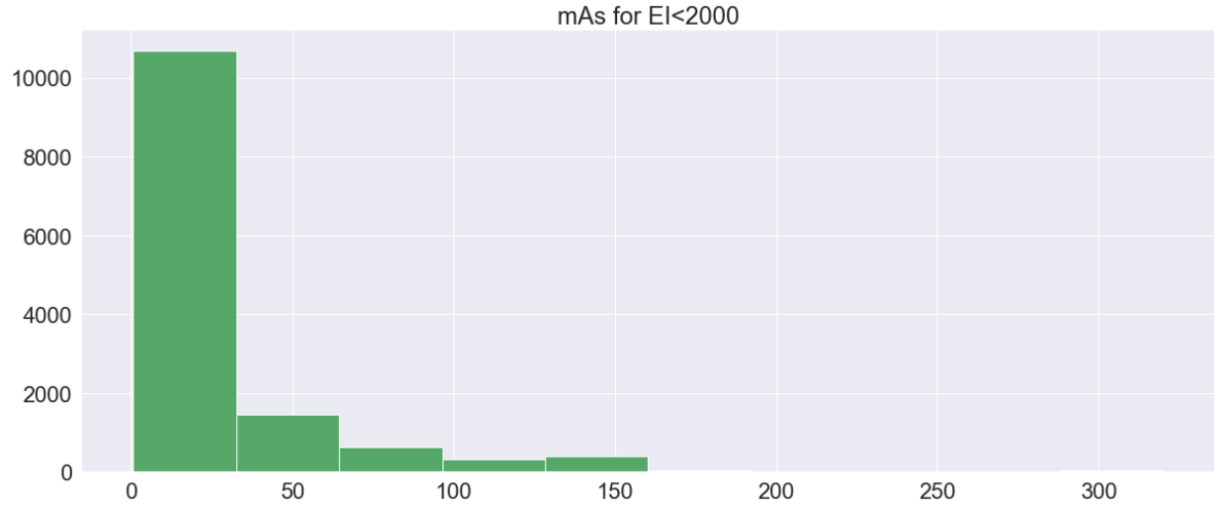
Interval of kVp for EI $= 2000$ is wide, but not as wide as interval for kVp for all samples: $[\,50\,,100\,] \subset [\,40,120\,]$. Interval of mAs for EI $= 2000$ is definitely a lot narrower then mAs for all samples: $[\,1.56\,,221.1\,] \subset [\,0.56,320.0\,]$.

Histograms can help us visualize the distribution of kVp and mAs for maximal EI:




kVp for EI<2000
5000
4000
3000
2000
1000
0
40
50
60
70
80
90
100
110
120


mAs for EI=2000
175
150
125
100
75
50
25
0
0
50
100
150
200

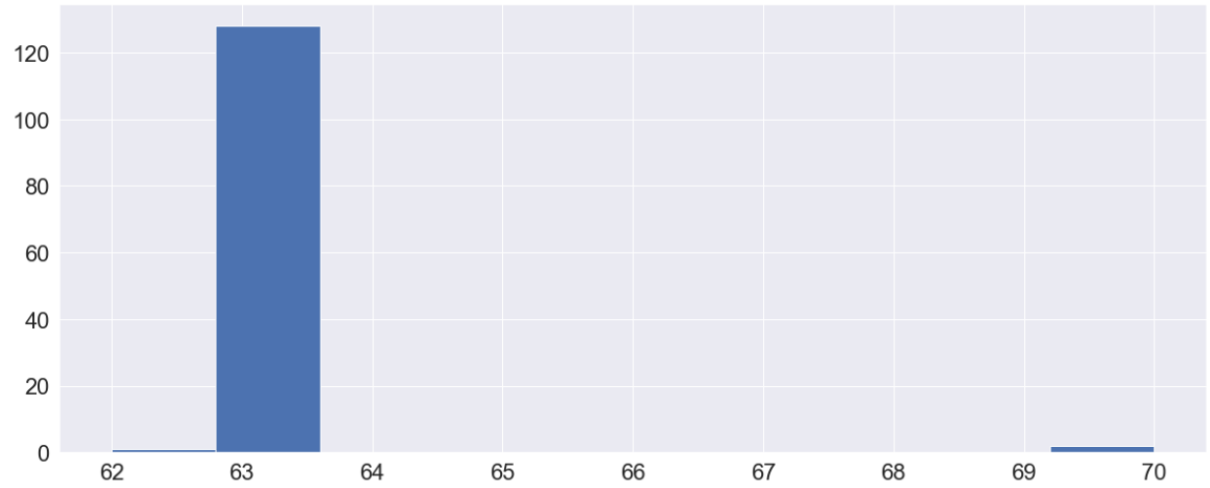

mAs for EI<2000
10000
8000
6000
4000
2000
0
0
50
100
150
200
250
300

Now we will divide these samples with the largest EI by anatomy and then look at the kVp and mAs. There are 294 images in the dataset with EI = 2000, 131 of them show a foot, 39 knee, 29 shoulder, 27 pelvis, 18 hip, 14 elbow, 9 arm, 9 lspine, 7 tspine, 4 cspine, 2 wrist, 2 hand, one chest and for two remaining the anatomy is unknown.

Since the most of images with EI = 2000 are of anatomy of foot, let's look at their kVp, mAs and compare it with kVp and mAs for all images of a foot.

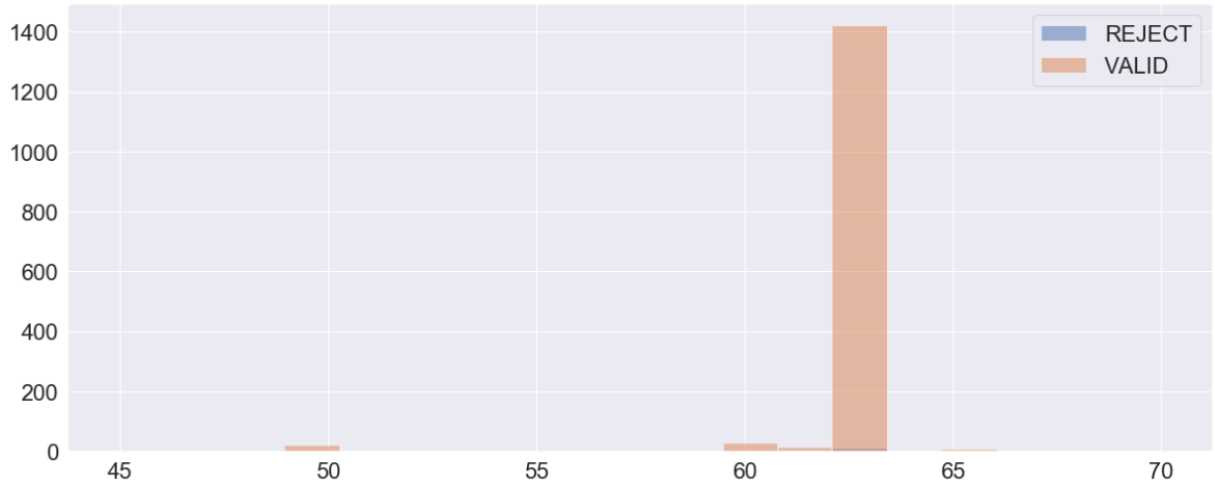
KVp of foot where EI = 2000:

```
 FOOT
63    128
70      2
62      1
Name: kVp , dtype: int64
```



KVp of all foot images:

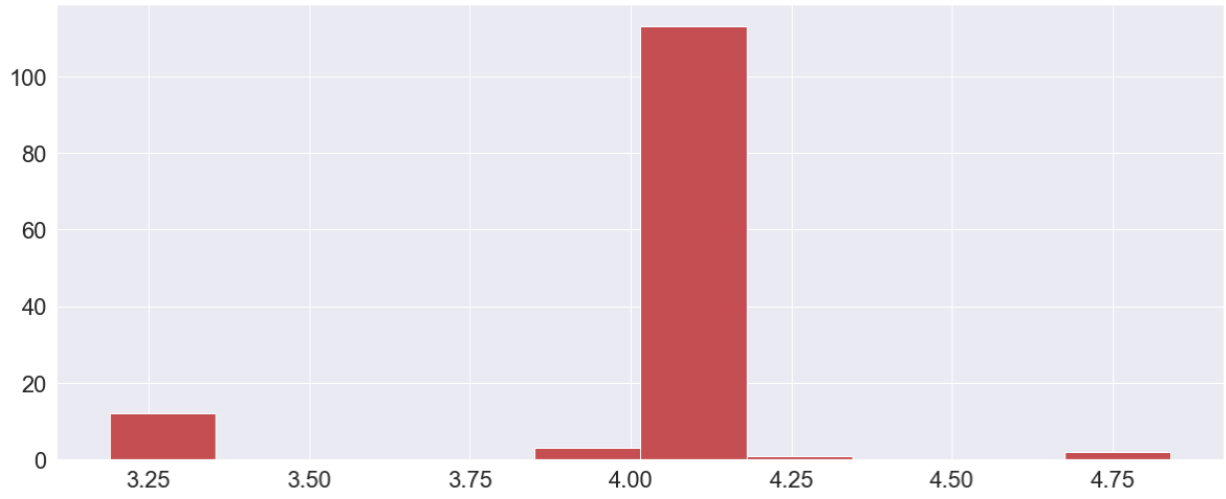
```
(' FOOT                ', ' REJCT ')
(' FOOT                ', ' VALID ')
```



Here we see that 128 images of foot with the largest EI have the most common value of kVp for foot – 63. One has value 62 and two have kVp = 70 which is not that common and is the upper bond for kVp of a foot.

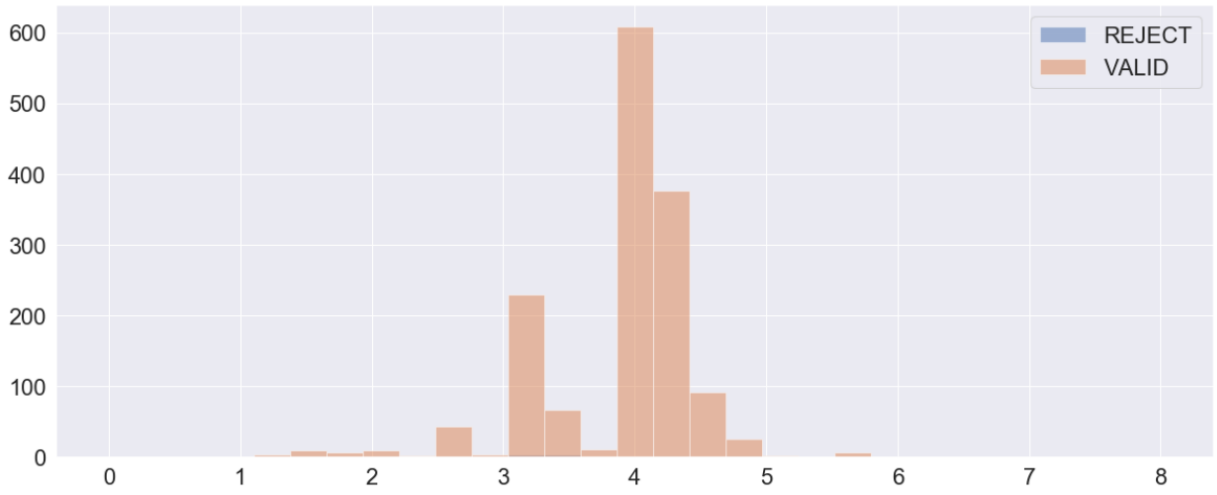
mAs of foot where EI = 2000:

```
 FOOT
4.05    26
4.04    25
4.06    24
4.07    13
4.03     9
4.08     9
3.21     4
3.20     4
4.01     3
4.02     3
4.09     3
4.84     2
3.23     2
3.24     1
4.10     1
3.19     1
4.25     1
Name: mAs , dtype: int64
```



mAs for all images of a foot:

```
(' FOOT                ', ' REJCT ')
(' FOOT                ', ' VALID ')
```



Here, mAs of images with the maximal EI is in interval [3.20 , 4.84]. Interval in which is half of all the samples are is [3.56 , 4.2] with interquartile range 0.6400000000000001. Mas for EI = 2000 is in wider interval, but it doesn't take extrem values.

I performed two samples t-test for mAs and kVp on sample sets with EI = 2000 and with EI < 2000. Zero hypothesis is that they have the same expected value. This test will show if there exist significant differences between X-ray exposure factors kVp and mAs on images with large EI as opposed to images with smaller EI.

Results of two sample t-test for kVp:

```
Ttest_indResult(statistic=-3.5810122323694347, pvalue=0.000578779244384304
4)

 ANKLE
Ttest_indResult(statistic=nan, pvalue=nan)

 ARM
Ttest_indResult(statistic=-1.2092038814048958, pvalue=0.23572632991525205)

 CHEST
Ttest_indResult(statistic=nan, pvalue=nan)

 CSPINE
Ttest_indResult(statistic=0.21315906829230225, pvalue=0.8312481502813376)

 ELBOW
Ttest_indResult(statistic=1.3915510365927184, pvalue=0.16501303083092783)

 FOOT
Ttest_indResult(statistic=2.406906514660287, pvalue=0.01620832770228678)
```

```
 HAND
Ttest_indResult(statistic=-0.08898940368373667, pvalue=0.9291085878900869)

 HIP
Ttest_indResult(statistic=0.6114534151763664, pvalue=0.5412271386532377)

 KNEE
Ttest_indResult(statistic=0.9020916859740038, pvalue=0.36709296322207885)

 LEG
Ttest_indResult(statistic=nan, pvalue=nan)

 LSPINE
Ttest_indResult(statistic=2.249823622057545, pvalue=0.024553198337461786)

 PELVIS
Ttest_indResult(statistic=-0.9954866502436879, pvalue=0.3200064754437999)

 SHOULDER
Ttest_indResult(statistic=-1.3273547763534734, pvalue=0.18460116871342225)

 SKULL
Ttest_indResult(statistic=nan, pvalue=nan)

 TSPINE
Ttest_indResult(statistic=-1.6686271479225228, pvalue=0.09574727179182123)

 WRIST
Ttest_indResult(statistic=0.05790468156946368, pvalue=0.9538413917494271)
```

We can conclude that with confidence level 0.05 we have enough evidence to reject the null hypothesis for foot, lspine and unknown anatomy. However, we don't have sufficiant evidence to reject the null hypothesis for anatomies: arm, cspine, elbow, hand, hip, knee, pelvis, shoulder, tspine and wrist.

Results of two sample t-test for mAs:

```
Ttest_indResult(statistic=0.3513298909814573, pvalue=0.7262419980305554)

 ANKLE
Ttest_indResult(statistic=nan, pvalue=nan)

 ARM
Ttest_indResult(statistic=-1.5046485595762682, pvalue=0.14253593887874952)

 CHEST
Ttest_indResult(statistic=nan, pvalue=nan)

 CSPINE
Ttest_indResult(statistic=5.0583800126137, pvalue=5.064664463199981e-07)
```

```
 ELBOW
Ttest_indResult(statistic=-0.10960433500415323, pvalue=0.9127910257973383)

 FOOT
Ttest_indResult(statistic=1.64783528028154, pvalue=0.0995956307716126)

 HAND
Ttest_indResult(statistic=2.9804873703787558, pvalue=0.0029492592679197854
)

 HIP
Ttest_indResult(statistic=2.1181257061153245, pvalue=0.03474564650898367)

 KNEE
Ttest_indResult(statistic=9.321147016063945, pvalue=2.4048652819096094e-20
)

 LEG
Ttest_indResult(statistic=nan, pvalue=nan)

 LSPINE
Ttest_indResult(statistic=2.4151793643680572, pvalue=0.015804056481903247)

 PELVIS
Ttest_indResult(statistic=2.0699840457480647, pvalue=0.03899377070192052)

 SHOULDER
Ttest_indResult(statistic=8.269228498295139, pvalue=3.0242433225963137e-16
)

 SKULL
Ttest_indResult(statistic=nan, pvalue=nan)

 TSPINE
Ttest_indResult(statistic=4.302047334685248, pvalue=1.995279217764903e-05)

 WRIST
Ttest_indResult(statistic=0.2066550593837283, pvalue=0.8363404887894182)
```

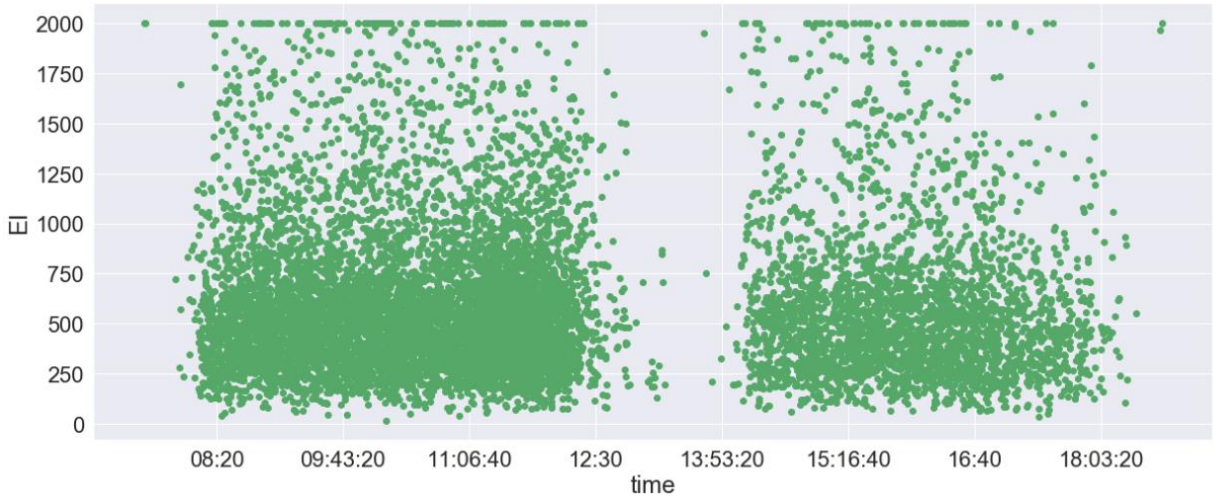
Taking confidence level 0.05, we have sufficient evidence to reject the null hypothesis for hand, hip, lspine and pelvis. Also, we don't have sufficiant evidence to reject the null hypothesis for anatomies: ankle, arm, chest, cspine, elbow, foot, knee, shoulder, tspine, wrist and unknown anatomy.

Only for lower spine with $EI = 2000$ both kVp and mAs don't have the same mean value as the samples with $EI < 2000$ with confidence of 0.05.

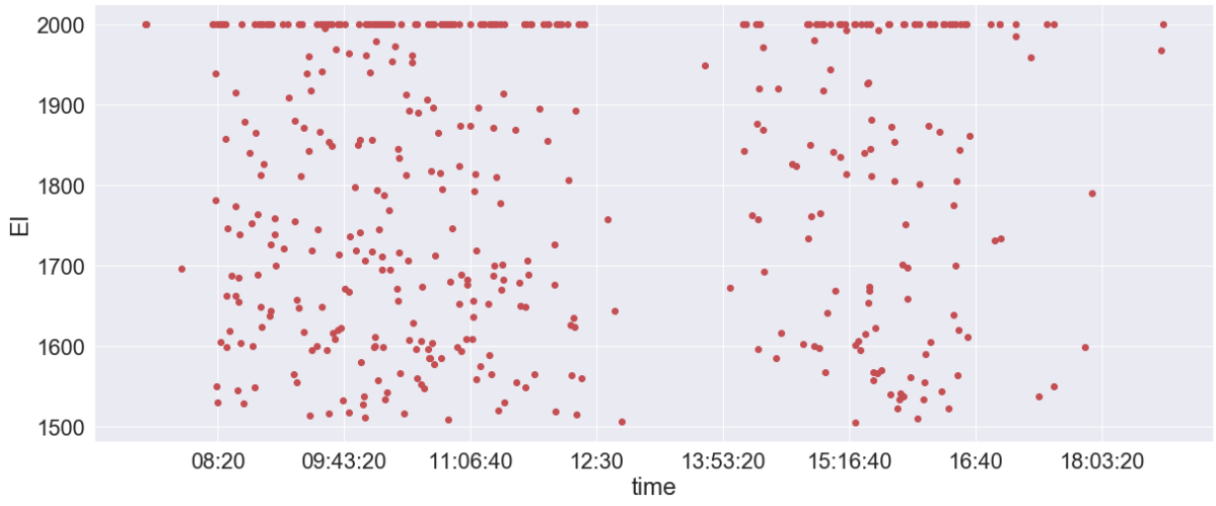
How is EI changing according to the time of day

For each X-ray image we have the date and time when it was taken as well as the dose of radiation that was used to make the image. What we would like to know is do technicians use higher doses at the end of working hour when they are tired? Is there a specific time of the day when the dose is high? In other words, we would like to find a link between time of the day and EI.

The EI in relationship to time plotted as scatterplot: (If we have two or more images taken at the same second, their average is displayed)

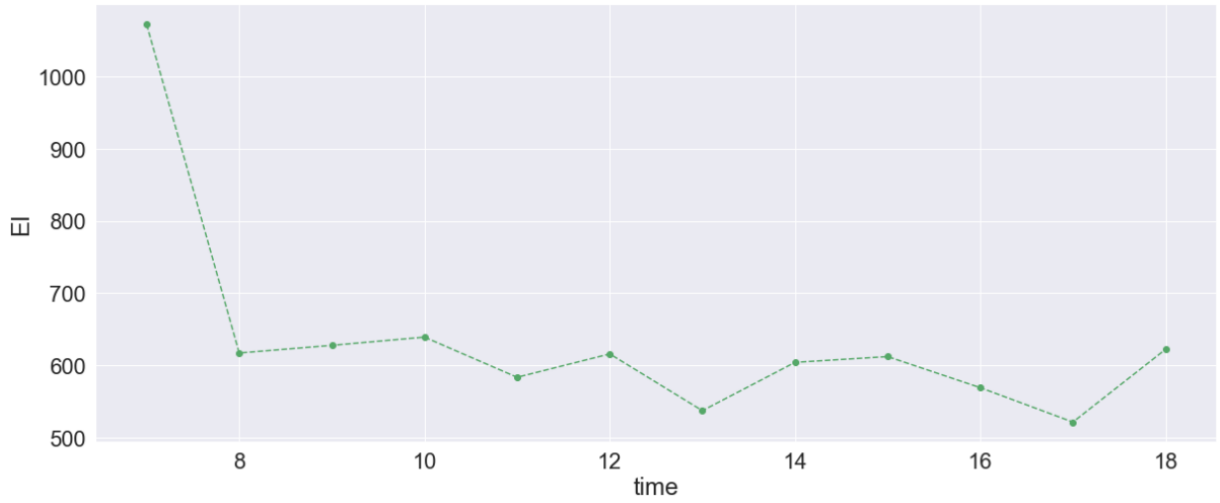


Scatterplot with EI > 1500:



Out of 10362 moments in time, in 459 were captured doses higher than 1500.

Here is a simplified graphic where we took for every hour the mean value of EI taken in that hour and represented it on the graph:

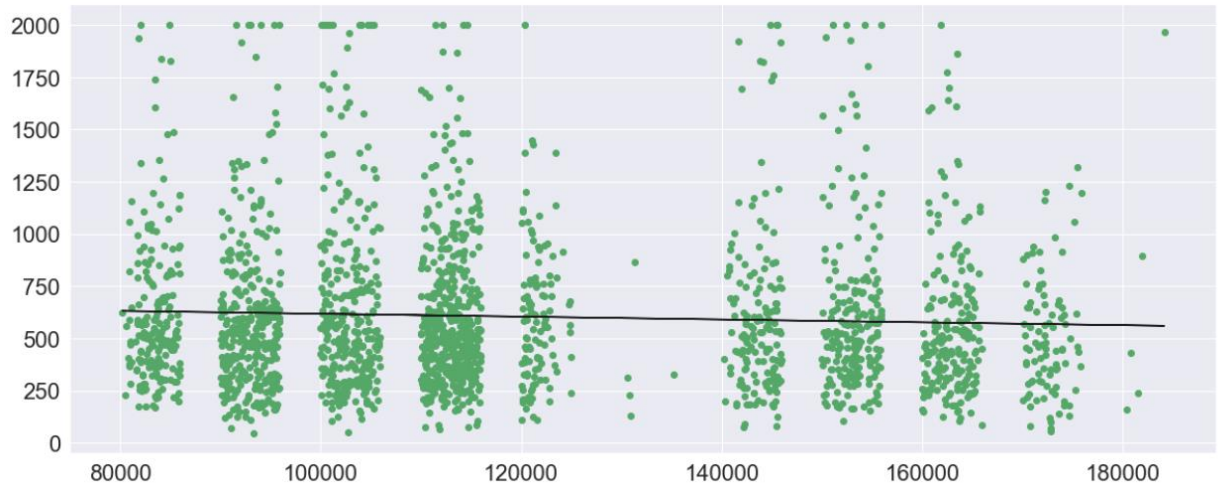


Here we see that the highest average value of EI is early in the morning between 7 and 8 o'clock and the lowest is between 17 and 18 o'clock. However, it is important to note that there is significantly less images taken between 7 and 8 o'clock (7 images), between 13 and 14 o'clock (20 images) and after 18 o'clock (29 images) then in other times of the day.

I have mapped time to integers such that, for example, 08:09:31 is mapped to 80931. So, I used the formula $\text{hour} * 10000 + \text{minute} * 100 + \text{second} * 1$ to calculate these integers.

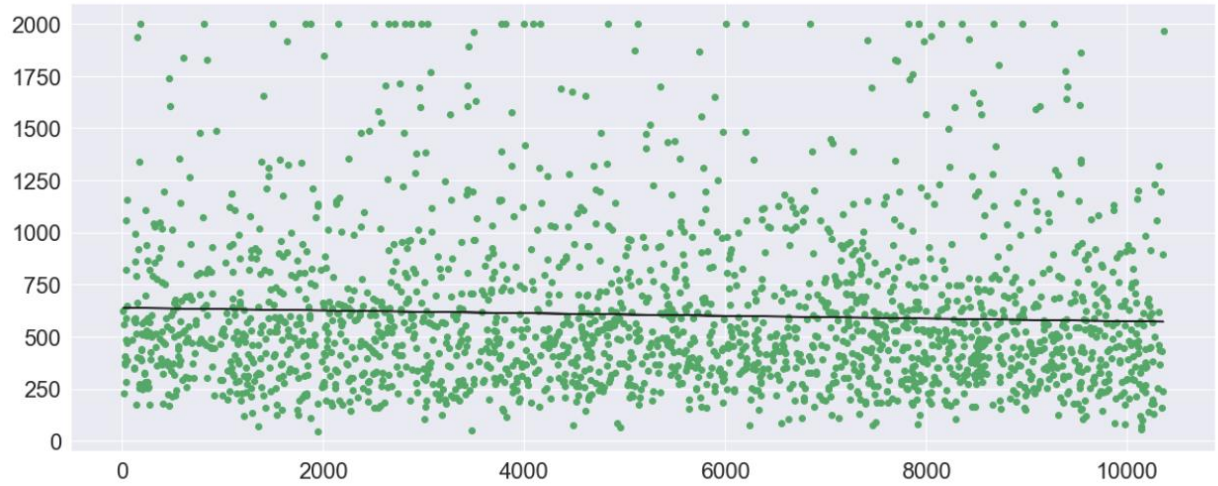
Correlation between EI and mapped time is -0.04183596. It is negative, which means that as one variable increases, the other decreases proportionally. However, since $-0.04183596 > -0.5$, these variables are not strongly correlated. This implies that there is no signifficant link between EI and time. Therefore, it is not strongly advised to use one to predict the value of the other. However, I tried making prediction model using linear regression.

I've split the data on training and test sets and made a linear regression model to see will it be able to predict the EI with given time (mapped to integers). This is the result:



The predicted values are on the black line, while the real values from the test set are the green dots. The line goes through the area on the graph where the dots are dense. However, it would never predict the high values since they are not frequent. Coefficient of determination (the proportion of the variation in the dependent variable, EI, that is predictable from the independent variable, time) is 0.0022974360271222904. Since it is closer to 0, the to 1, this is the indication that it isn't that good of a fit.

I have tried one more way to map the time. The first time instance is mapped to 1, and every next time instance is mapped to the next integer. The results were similar.




```
coefficient of determination: 0.0024511322594605867
```

In conclusion, since time and EI are not strongly correlated it isn't possible to accurately predict when the EI will be high according to the time of day. EI is a bit higher on average in the morning. It is decreasing as time passes, but just slightly.