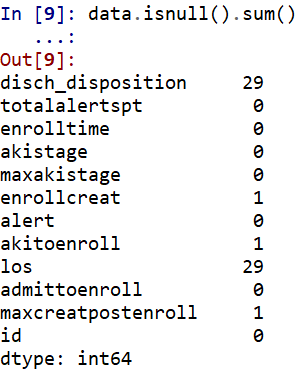
There are two groups: the test group with alert = 1 and control group with alert = 0.

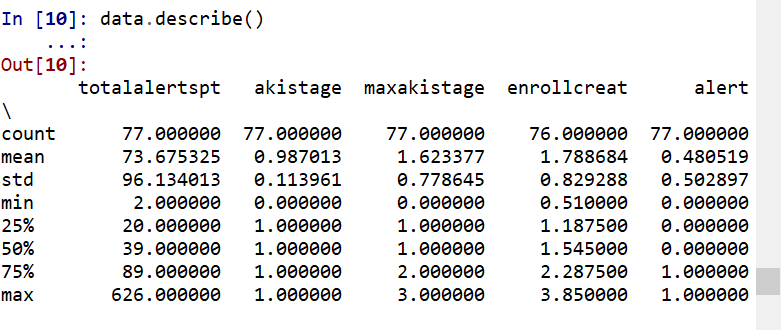
**Data Understanding and preprocessing:**

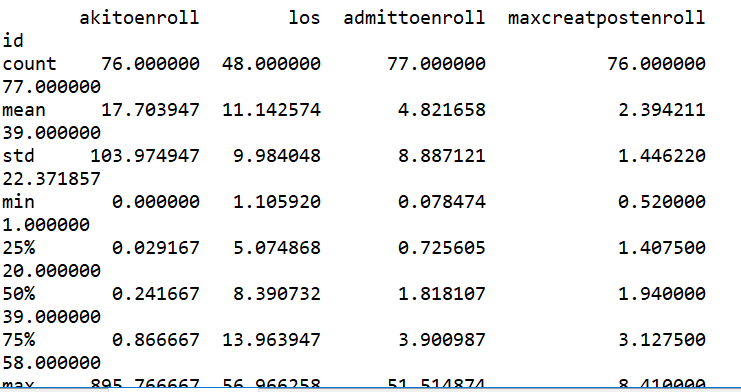
Checking if there are any nulls in each feature of the dataset.



We observe that the variables ‘disch\_disposition’ and ‘los’ has 29 null values each, enrollcreat, akitoenroll, maxcreatpostenroll have 1 null value each.

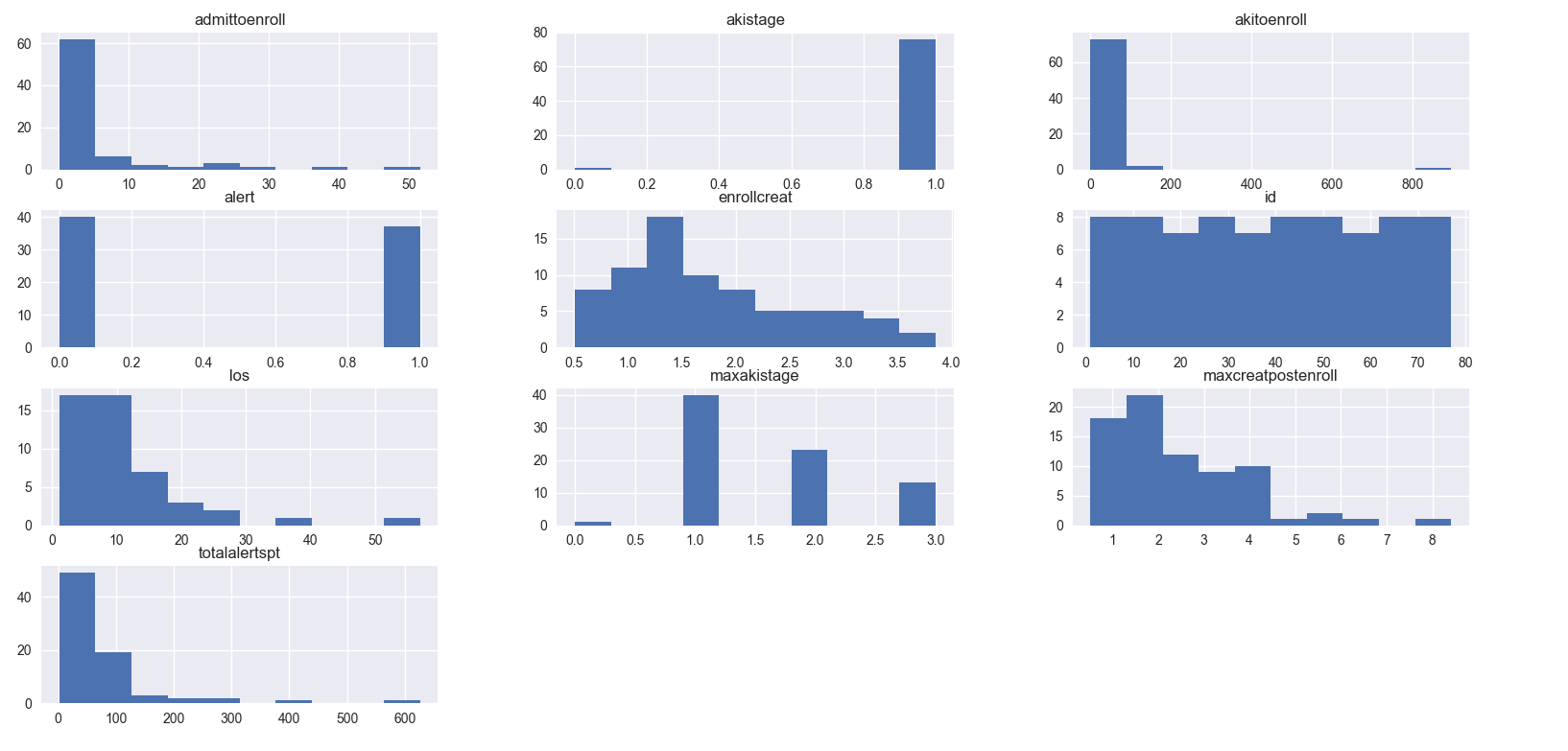
Checking the distribution of numerical variables and imputing missing values for enrollcreat, akitoenroll, maxcreatpostenroll variables.



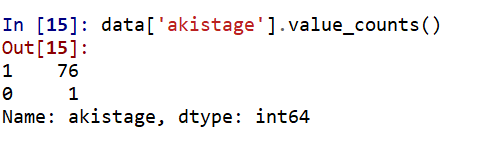


data[data.dtypes[(data.dtypes=="float64")|(data.dtypes=="int64")]

.index.values].hist(figsize=[11,11])

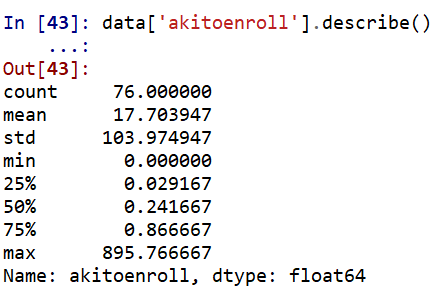


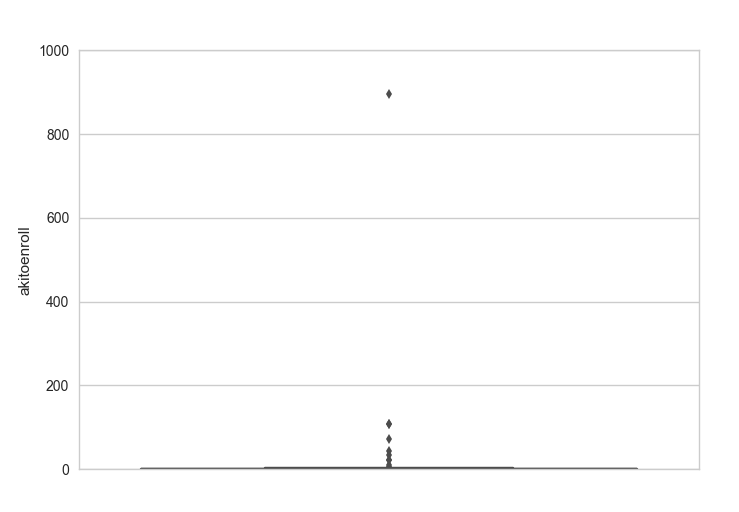
When we look at the above distributions we see that the variable ‘akistage’ has very few variables having value as zero and most of them having 1 as its value. Let’s closely look at each value\_counts for akistage variable.



The above statistics show that ‘akistage’ feature is not useful for our study as it has only 1 value as zero and all others as 1, which is not a random variable. So, we can remove this variable from our study.

The variable akitoenroll also shows similar distribution where majority of the values are 0 and it has very few values after 800. Let’s have a close look at this variable using box plot. But when we see the value counts some many values are in between 0 and 1.



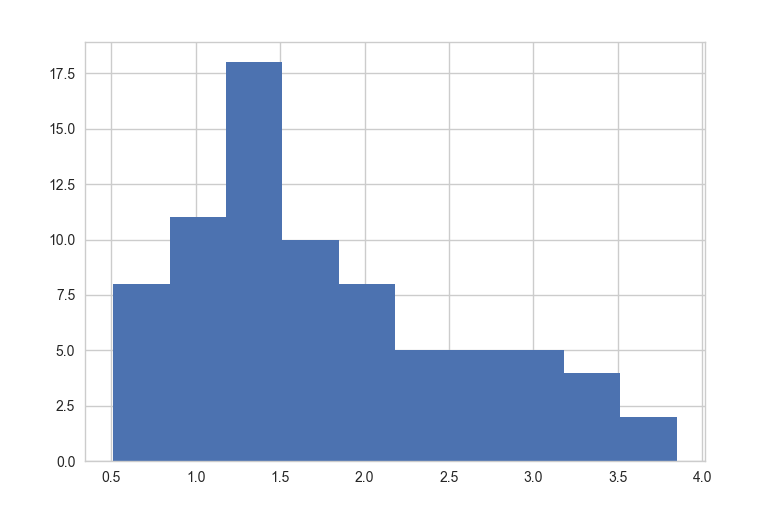
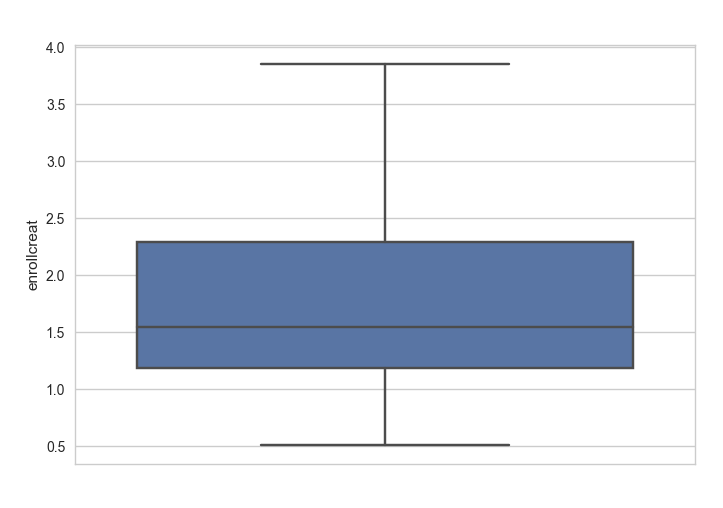


We found that there is only one outlier for this variable whose value is 895.766667.

Let’s only look at the distribution of this variable whose values are below 200. When we look closely at the histogram most of the values in between 0 and 1, also evident in the above describe statement that 75% of the data is also less than 1. As we have one of its value is null, let’s impute the missing value to be the median of that column as the data is skewed.

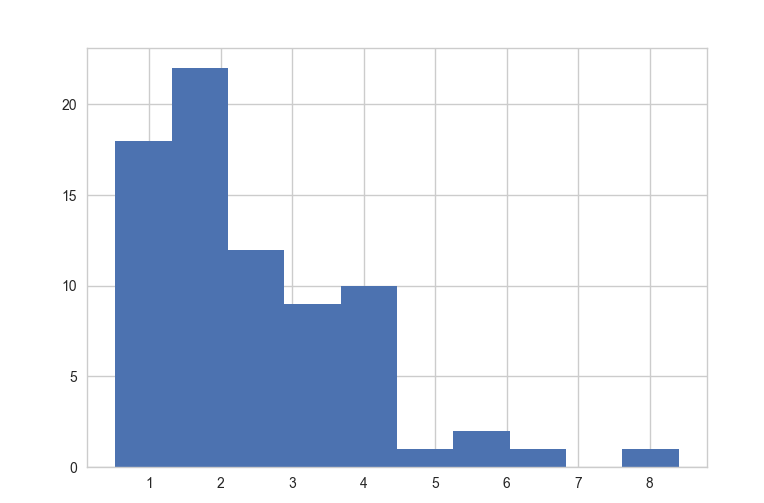
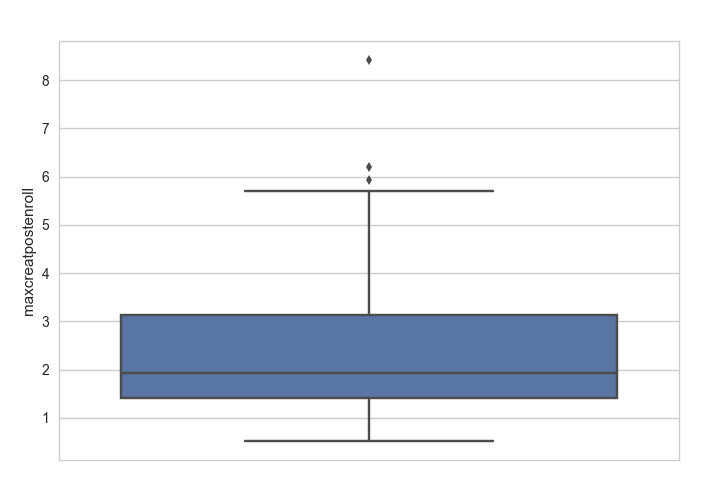
There are two other variables with 1 null value in each of them. They are ‘enrollcreat’ and ‘maxcreatpostenroll’. Let’s impute these variables.

Enrollcreat: Let’s look at the box plot distribution of Enrollcreat, it also a little skewed but almost looks like a normal distribution. So let’s impute the missing value by taking the mean of this variable.



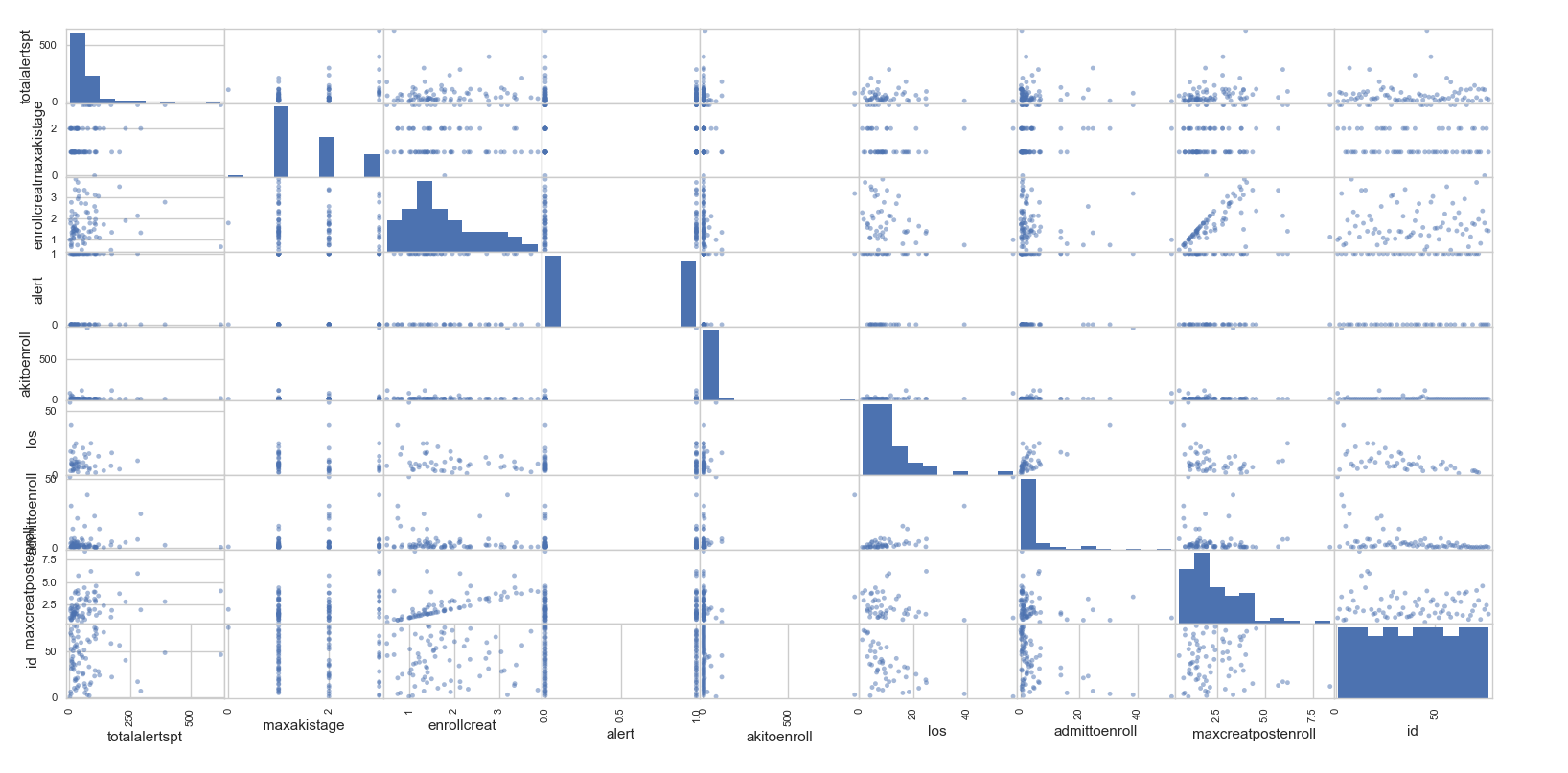
Imputing Maxcreatpostenroll:

The distribution of this variable is much skewed, so median would be a good statistic to impute. So, imputing ‘Maxcreatpostenroll’ variable missing values with its median value of it non-null values.

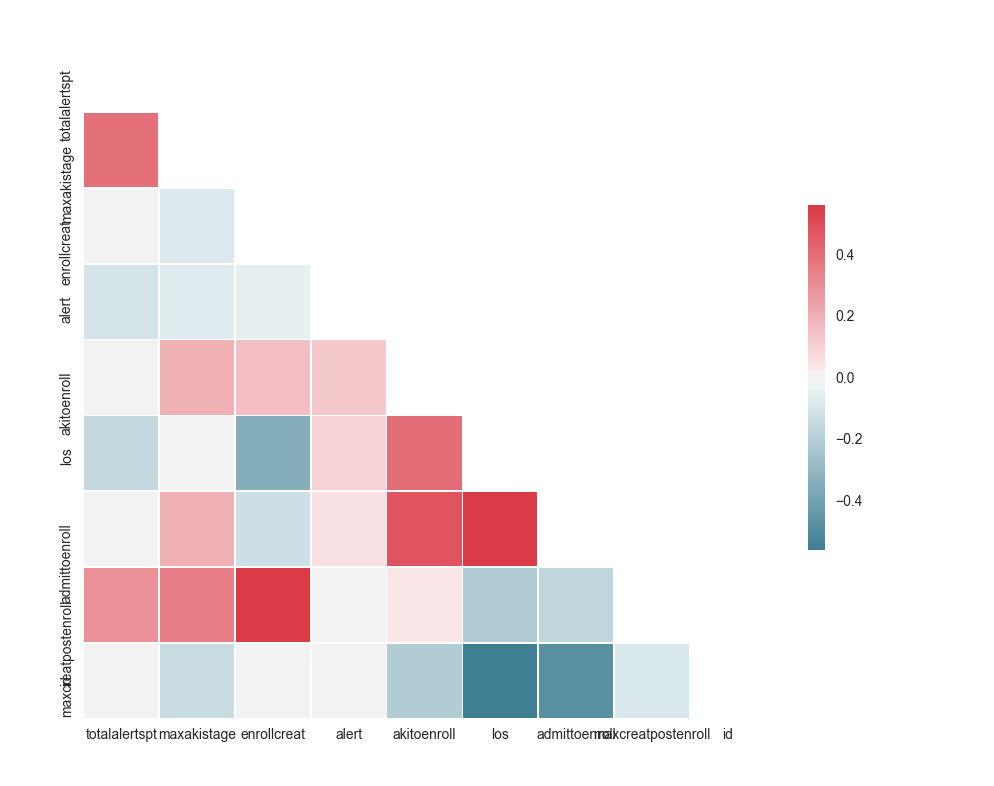


As we can cannot impute the variables which are missing more than 30%. Let’s keep them aside for now.

Correlation matrix for the dataset:



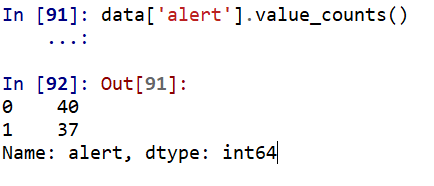
From the above correlation matrix, it is evident that enrollcreat and maxcreatpostenroll are positively correlated. Others correlations cannot be described clearly. Let’s find it graphically.



From fig the variables ‘los’ and ‘admittoenroll’ are highly correlated as these variables are almost representing similar information. When considering for modelling only one among these two should be considered. Also, ‘enrollcreat’ is positively correlated with ‘maxcreatpostenroll’, as both are measuring the same creatinine level. The ‘ID’ is negatively correlated with ‘admittoenroll’ and ‘los’. As we do not use ‘ID’ column for any tests/models we can ignore these correlations.

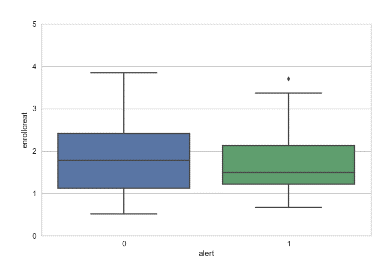
Answers to the questionnaire:

1. **How many individuals were assigned to each randomization group**



40 patients are assigned to control group and 37 test group.

1. **How would you go about comparing the creatinine level at enrollment across the two groups? Are they different?**

#Box plot code

sns.set\_style("whitegrid")

import matplotlib.pyplot as plt

plt.ylim(0, 5)

ax = sns.boxplot(x="alert", y="enrollcreat", data=data)

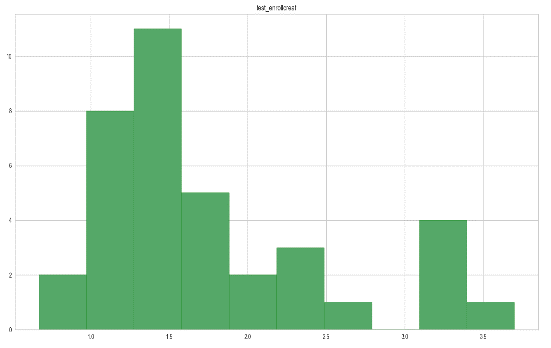
#Histogram code

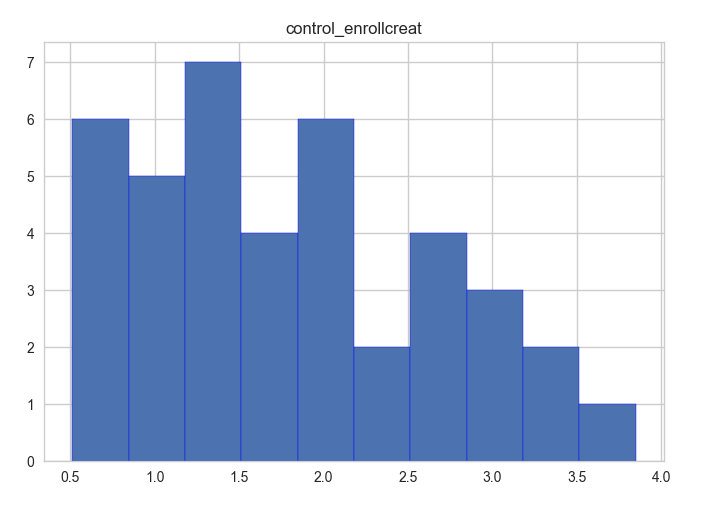
plt.hist(test\_enrollcreat, color= None, edgecolor= 'green')

plt.title(‘test\_enrollcreat’)

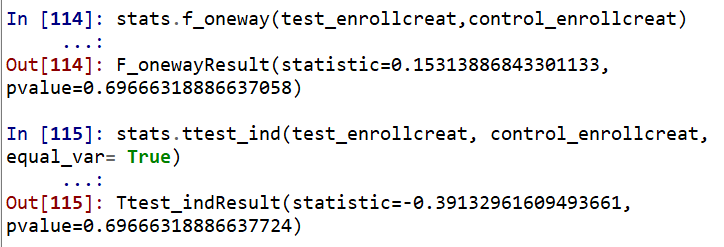
plt.hist(control\_enrollcreat, color= None, edgecolor= 'blue')

plt.title(‘control\_enrollcreat’)



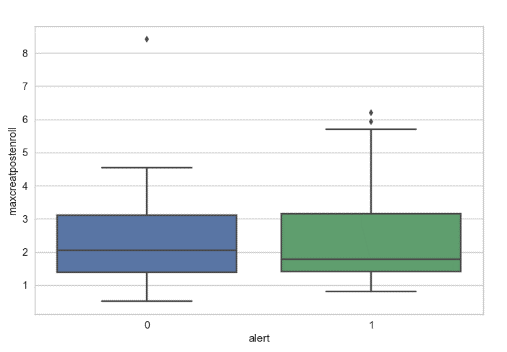


Running ftest and ttest to check if the control group and test group have different creatinine levels at the time of enrollment.



From the results of f test, the two groups variance is equal as the p-value is greater than 0.05, Interpretation of F\_test: we fail to reject the null hypothesis that the variances of two variables are equal. Hence, we accept that the variances of the two groups are equal.

As we have equal variance and approximately normal distributed data, the results for t test are shown in the above. The value of p is greater than 0.05 which fails to reject the null hypothesis that test and random groups are statistically equal. Hence, the two groups have statistically equal ‘creatinine level at enrollment’.

1. **How would you go about comparing the maximum achieved creatinine level after enrollment across the two groups? Are they different?**

#Box plot code

sns.set\_style("whitegrid")

import matplotlib.pyplot as plt

ax = sns.boxplot(x="alert", y="maxcreatpostenroll", data=data)

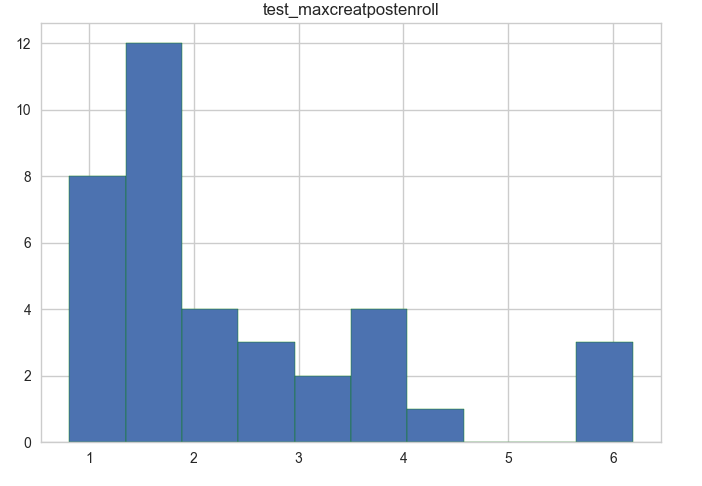
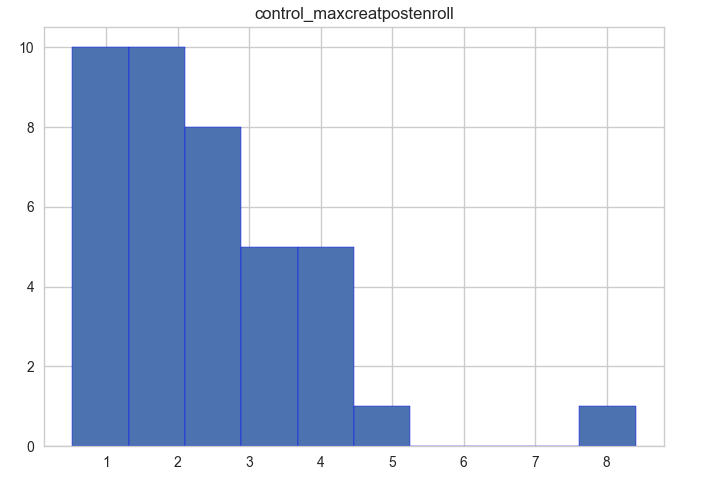
#Plotting histogram to check if the distribution is approximate to normal

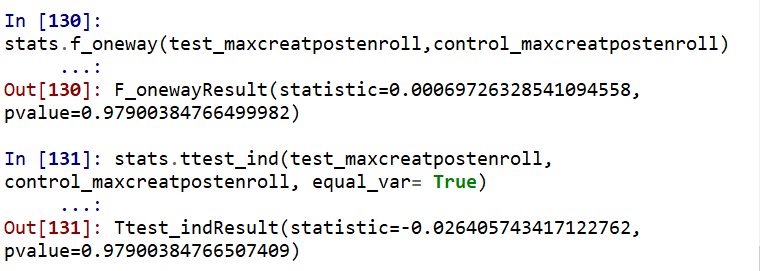
plt.hist(test\_maxcreatpostenroll, color= None, edgecolor= 'green')

plt.title('test\_maxcreatpostenroll')

plt.hist(control\_maxcreatpostenroll, color= None, edgecolor= 'blue')

plt.title('control\_maxcreatpostenroll')

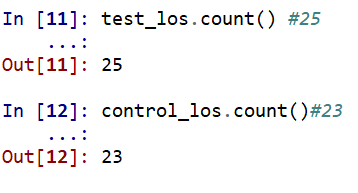


From the results of f test and t test the two groups have statistically equal ‘maximum achieved creatinine level after enrollment’. As we fail to reject the null hypothesis for both f-test, followed by T-test.

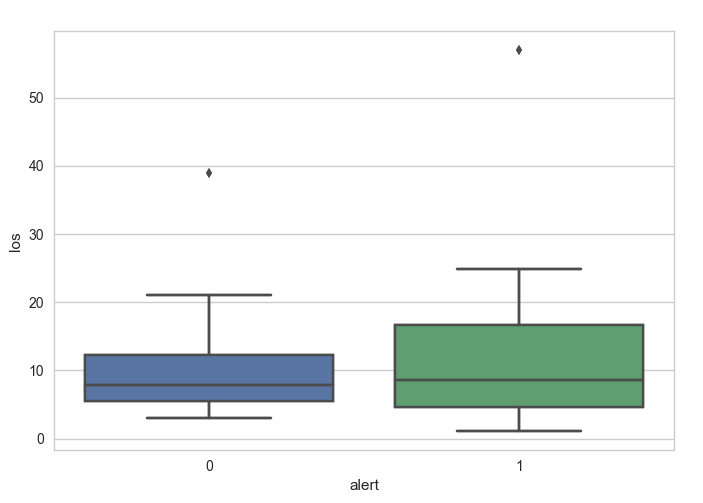
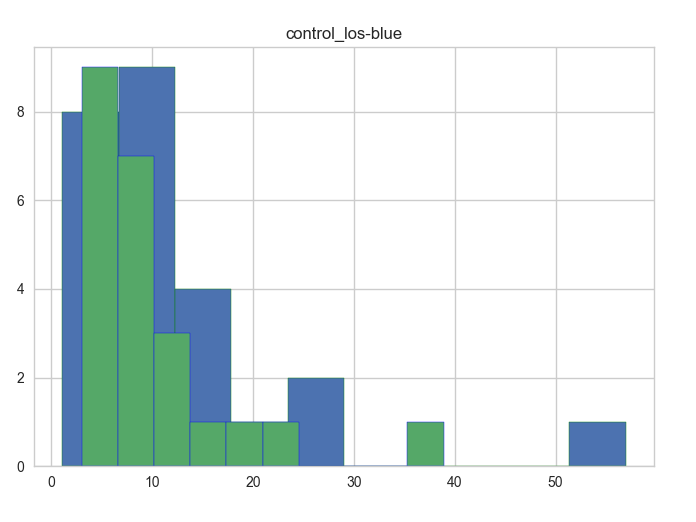
1. **How does the length of stay compare between the two intervention groups? How would you go about comparing these outcomes?**

As there are many nulls in the data, we are only considering the records with not-null in ‘los’ values.

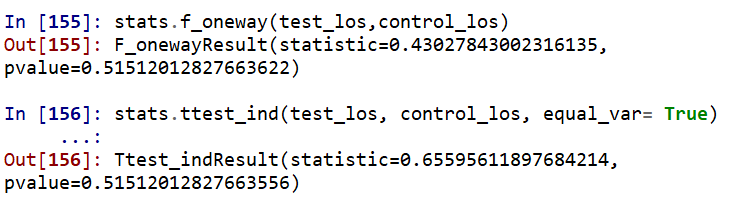
Verifying the data after removing the nulls if there are sufficient and random data in each of the randomized groups.



See the distribution over box plot and combined histogram where the control group is colored in blue.

The distribution in the grapgh shows that mean of the test groups ‘los’ is close to mean of the control group ‘los’. So, running the f test and t-test to check if the means of ‘los’ are different.

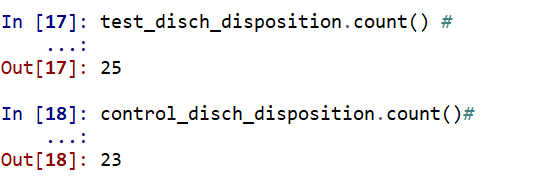


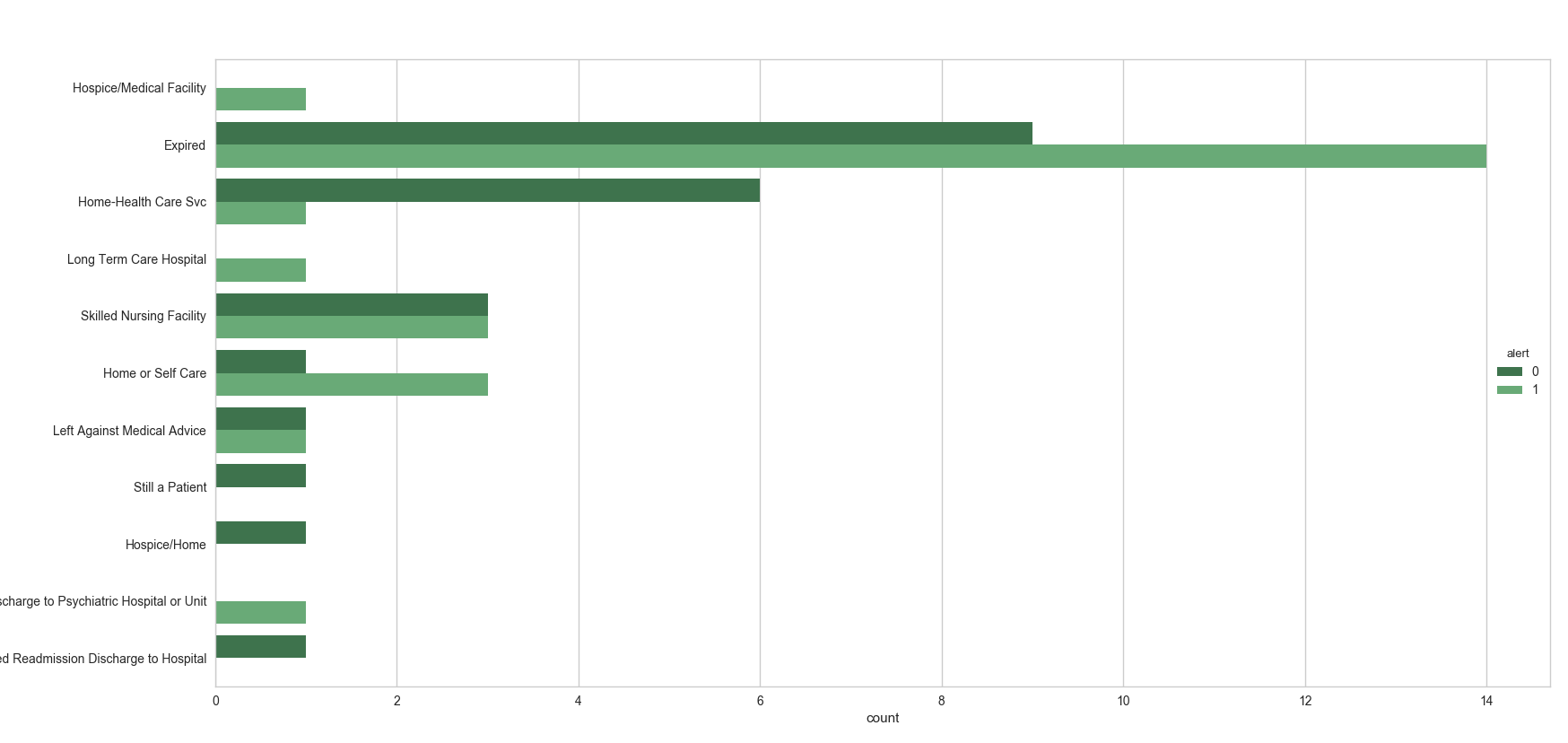
The f-test results in equal variance of test ‘los’ and control ‘los’, as the p value is greater than 0.05. The two sided t-test with equality of two means has p value is greater than alpha 0.05, the mean of the los for the test group and control group are equal.

1. **What about inpatient death? What are the findings there?**

There are 38% nulls in the disch\_disposition variable. Assuming that the null records are missing data.

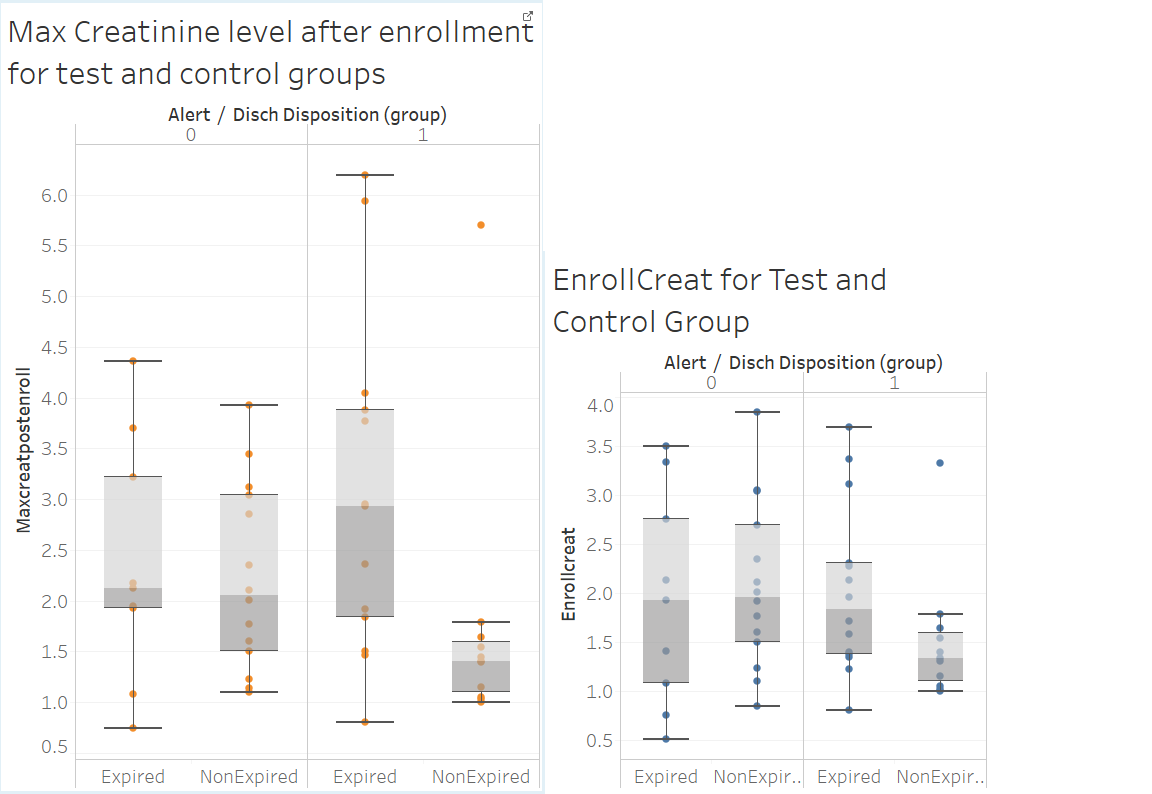
Both the control and test group have nearly same number of records.



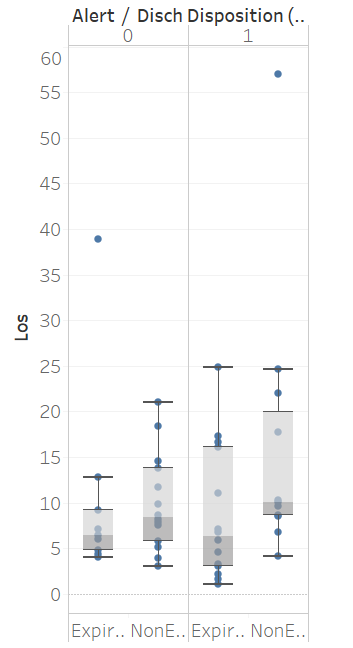
As disch\_disposition is a categorical variable, we can use count plots to view the distribution. From the count plot for test and control group, it is evident that no of deaths is more for test group compared to control group. 

The percentage of deaths in each alert group is 56% where as in control group it is 39.13%.

Comparison of Max creatinine level post enrollment, enroll creatinine level for expired and non-expired for each of the test and control groups is given below. When comparing the box plots of max creatinine level, enroll creatinine for expired people in alert group, it is observed that the max creatinine level for the expired people in alert group has increased on average.



Comparison of Los for control and alert groups for expired vs non-expired. From the below comparison the drug has not shown any significant differences in length of stay for control and test groups.



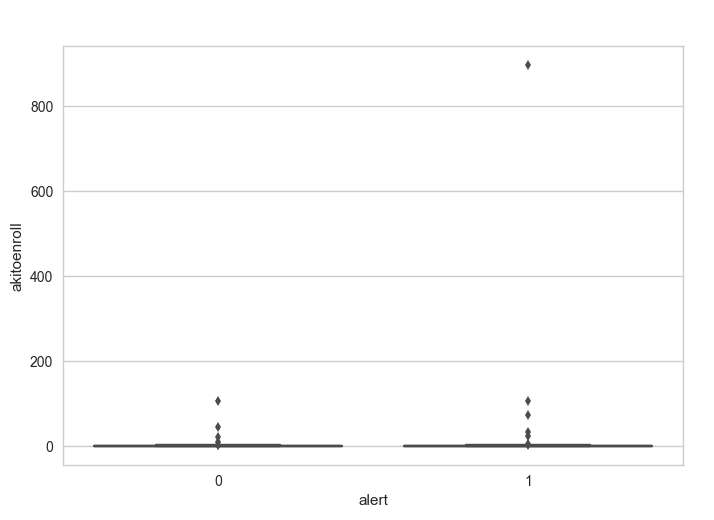
1. Are there any particular outliers with regards to the time between acute kidney injury and trial enrollment? How would you identify them?

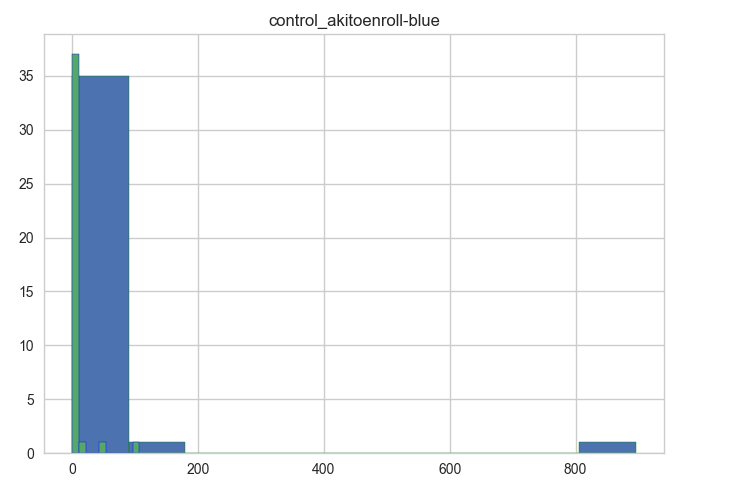
sns.set\_style("whitegrid")

import matplotlib.pyplot as plt

ax = sns.boxplot(x="alert", y="akitoenroll", data=data)

#Histogram code showing the distribution of akitenroll for test and control group. The second figure shows the distribution of akitenroll values less than 120, as there is an outlier of akitenroll for test group.

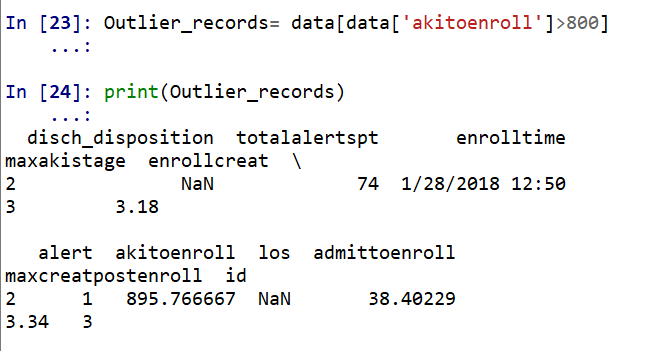




From the above distribution of data the record with values of akitoenroll greater than 800 are outliers. Outliers can be found using the below code:

Outlier\_records= data[data['akitoenroll']>800]

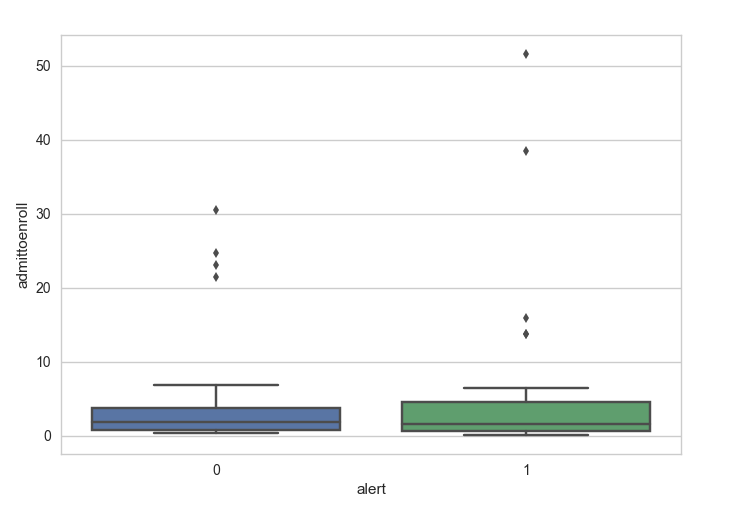
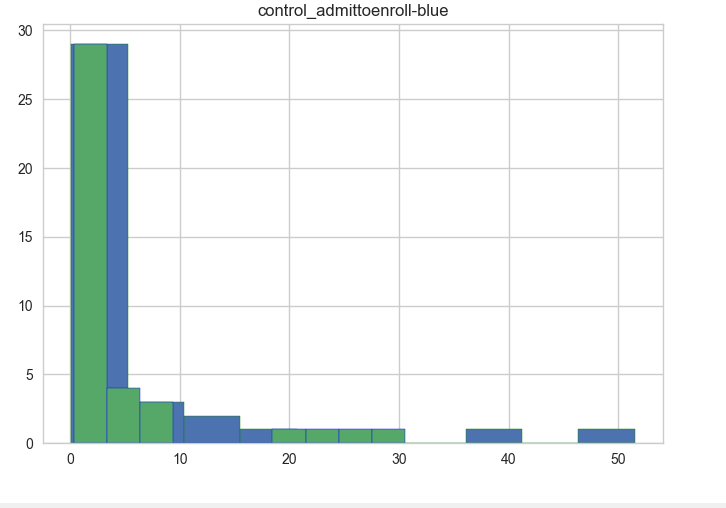
print(Outlier\_records)

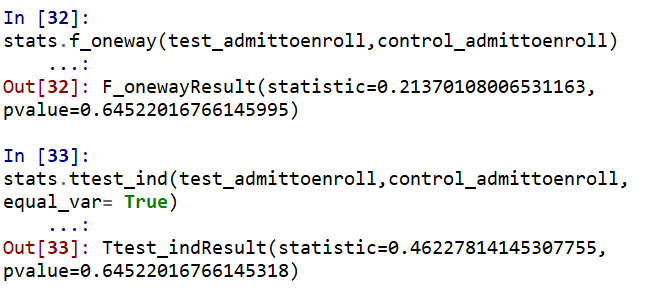


From the above observations it is clear that the initial variables of the test and control group are almost random and represent correct data to test the effectiveness of drug. The variables to be checked for randomness are admittoenroll, akitoenroll, enrollcreat, totalalertspt(As other variables like akitstage and enrollmentdate are known to be not random ). Earlier we have checked for the randomness of akitoenroll, enrollcreat.

Validating the randomness of admittoenroll.

Box plot:

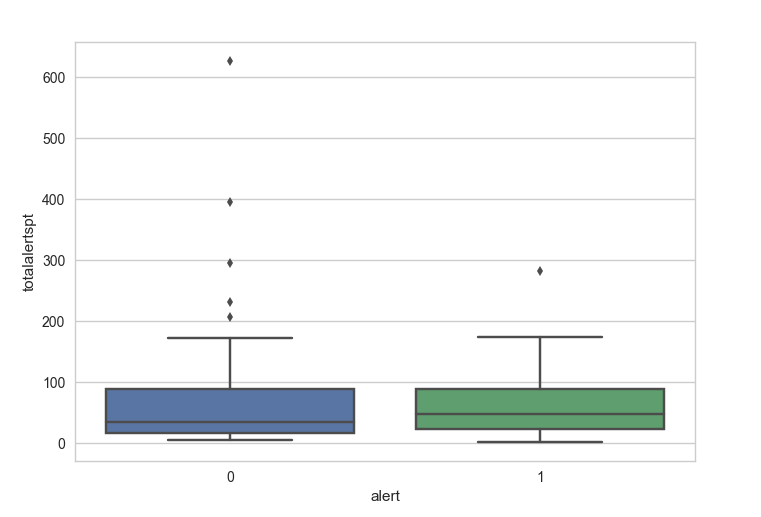
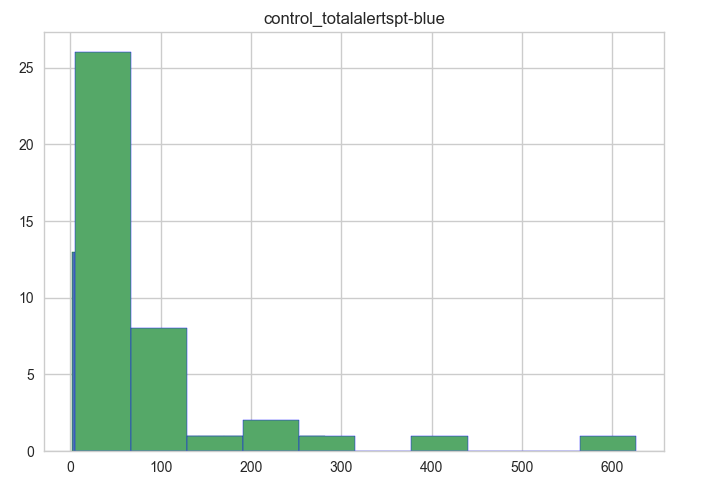
 

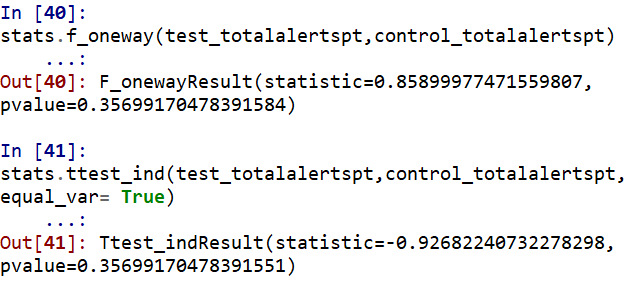


From the above F-test and T-test results, the admittoenroll variable is random.

Validating the randomness of ‘totalalertspt’ variable.

Box plot:





The f-test and t-test results show that the average totalalertspt is equal in both test and control group as there is no enough evidence to prove both are different.