

1. Wykorzystane biblioteki
2. Kod zapewniający powtarzalność
3. Wczytanie danych z pliku
4. Wstępne czyszczenie danych
5. Kod przetwarzający brakujące dane
6. Podsumowanie wartości w kolumnach
7. Zliczenie 50 najpopularniejszych
8. Korelacje między zmiennymi
9. Liczba przypadków dla każdej z klas
10. Wykresy rozkładu liczby atomów i elektronów
11. Klasy z największą niezgodnością liczby atomów i elektronów
12. Rozkład wartości kolumn part_01
13. Interaktywny wykres
14. Przewidywanie liczby elektronów i atomów na podstawie innych kolumn
15. Klasyfikator

W danych znajdowało się bardzo dużo wartości NA. W obliczeniach na zależności co było liczone zostały one zamienione lub pominięte - szczególnie podczas obliczania korelacji (sposób obsługi wartości NA podczas obliczania korelacji został opisany w punkcie 5).

Zauważono także, że kolumny fo_col fo_col zawierają tylko 1 wartość. Powinny one zostać usunięte ze zbioru przy próbie utworzenia klasyfikatora.

Zauważono także, że kolumny fo_col fc_col zawierają tylko 1 wartość. Powinny one zostać usunięte ze zbioru przy próbie utworzenia klasyfikatora.

Podczas rysowania wykresów dla kolumn zaczynających się od part_01 zauważono, że dużo wartości jest skupionych w okolicy zera.

Została podjęta próba wykonania klasyfikatora, ale niestety nie zakończyła się powodzeniem. W ostatnim punkcie zostały opisane podjęte kroki, które zostały podjęte oraz napotkane błędy.

```
library(knitr)
library(ggplot2)
library(dplyr)
library(ggExtra)
library(caret)
library(corrplot)
library(plotly)
library(randomForest)
```

Wczytanie danych z pliku.

```
rawData <- data.table::fread("all_summary.csv", header="auto", sep="auto")

## Warning in data.table::fread("all_summary.csv", header="auto", sep
## = "auto"): Bumped column 6 to type character on data row 13920, field
## contains '260G'. Coercing previously read values in this column from
## logical, integer or numeric back to character which may not be lossless;
## e.g., if '00' and '000' occurred before they will now be just '00', and
## there may be inconsistencies with treatment of ',', and 'NA', too (if they
## occurred in this column before the bump). If this matters please rerun and
## set 'colClasses' to 'character' for this column. Please note that column
## type detection uses a sample of 1,000 rows (100 rows at 10 points) so
## hopefully this message should be very rare. If reporting to datatable-help,
## please rerun and include the output from verbose=TRUE.
```

Załadowano 591042 wierszy, które mają 412 zmiennych.

```
popular <- rawData[, .N, by = res_name]
popular <- popular[order(-N)]
popular <- popular[1:50]
pop_names <- popular$res_name
rawData <- select(filter(rawData, res name %in% pop names), matches(".*"))
```

Pozostało 382720 wierszy z 50 najpopularniejszych grup.

Usuwanie z danych wiersze posiadające wartość zmiennej res_name równą: "UNK", "UNX", "UNL", "DUM", "N", "BLOB", "ALA", "ARG", "ASN", "ASP", "CYS", "GLN", "GLU", "GLY", "HIS", "ILE", "LEU", "LYS", "MET", "MSE", "PHE", "PRO", "SEC", "SER", "THR", "TRP", "TYR", "VAL", "DA", "DG", "DT", "DC", "DU", "A", "G", "T", "C", "U", "HOH", "H2O". WAT lub NAN ## Wstępne czyszczenie danych

```
selectedData <- selectedData <- rawData %>% filter(!res_name %in% c("UNK", "UNX", "UNL", "DUM", "N", "BLOB",
"ALA", "ARG", "ASN", "ASP", "CYS", "GLN", "GLU", "GLY", "HIS", "ILE", "LEU", "LYS", "MET", "MSE", "PHE", "PRO",
"SEC", "SER", "THR", "TRP", "TYR", "VAL", "DA", "DG", "DT", "DC", "DU", "A", "G", "T", "C", "U", "HOH", "H2O", "WA
TC", "NAN", "", "NA", "NA"))
```

[illegible]

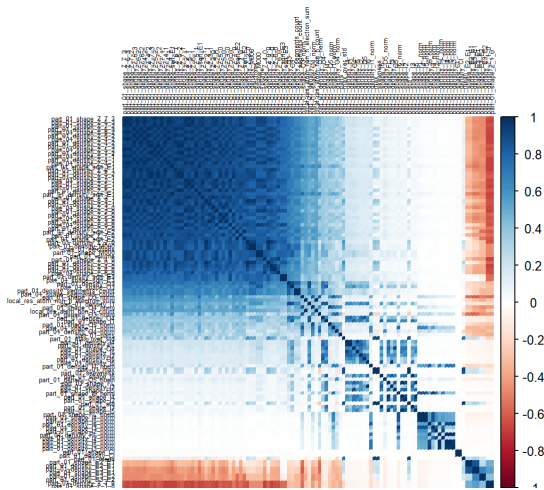
Korelacje między zmiennymi

Do obliczenia korelacji użyto funkcji `>cor<` z parametrem `>use = "pairwise.complete.obs"<`, który ignoruje w obliczeniach korelacji dla danej pary wartości NA.

Podczas obliczania korelacji zauważono, że dla wszystkich kolumn `part_XX` jest ona bardzo podobna. W przedstawionej graficznej reprezentacji korelacji zabrano zmienne z `part_01`, by ograniczyć liczbę danych. Poza kolumnami `part_01` w macierzy widzimy kolumny `local_res_atom_non_h_electron_sum`, `local_res_atom_non_h_count`, `solvent_mask_count`, `void_mask_count`, `modeled_mask_count`, `solvent_ratio`. Wybrano takie kolumny, ponieważ te kolumny będą brały w wyznaczaniu klasyfikatora.

Zmienne zostały posortowane wg algorytmu "FPC" (First Principal Component).

```
correlation_data <- cor(data, use = "pairwise.complete.obs")
corrplot(correlation_data, method = "color", tl.cex = 0.4, order = "FPC", tl.col="black")
```



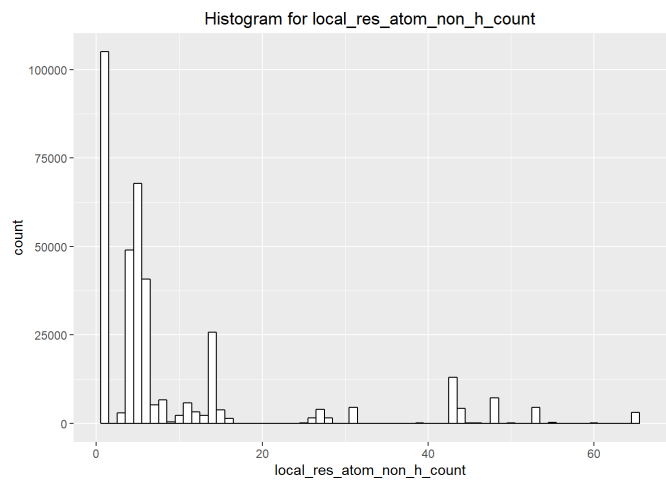
Liczba przypadków dla każdej z klas

count	class
1587	SAH
1589	GDP
1594	PLP
1596	NO3
1602	FE
1609	ACY
1637	NI
1647	SF4
1656	TRS
1905	PGE
1917	HEC
1933	EPE
2084	FMN
2106	NDP
2127	BR
2136	1PE
2183	COA
2296	ATP
2353	CU
2697	MES
2768	PG4
2841	MAN
2918	FMT
3221	MPD
3242	CD
3505	NAP
3509	MLY
3819	ADP
4215	MN
4501	NAD
4555	FAD
4706	K
4784	CLA
4987	PEG
6317	IOD
6633	DMS
8096	ACT
11090	PO4
11192	HEM

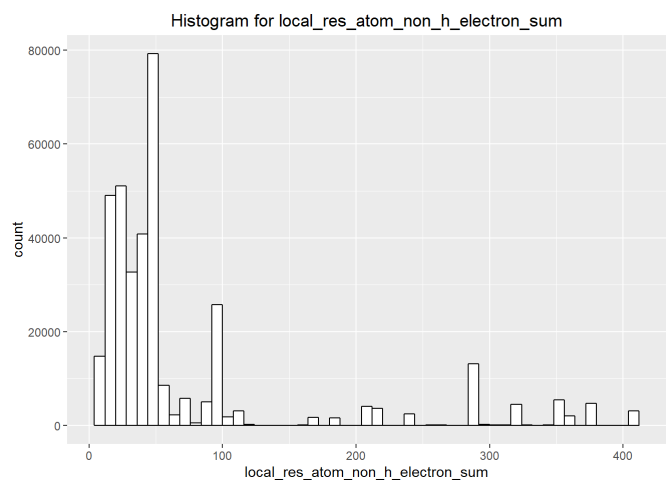
count	class
14779	MG
19826	ZN
21038	CA
23223	CL
26360	NAG
30825	EDO
40606	GOL
56572	SO4

Wykresy rozkładu liczby atomów i elektronów

Rozkład atomów



Rozkład elektronów



Klasy z największą niezgodnością liczby atomów i elektronów

Klasy z największą niezgodnością liczby atomów

res_name	local_res_atom_non_h_count	dict_atom_non_h_count	odds
NAG	369564	395400	25836
CLA	285867	310960	25093
1PE	28487	34176	5689
MLY	37453	42108	4655
NAP	163742	168240	4498
COA	100779	104784	4005
NAD	194945	198044	3099
PG4	33094	35984	2890
MAN	31551	34092	2541
NDP	99007	101088	2081

Klasy z największą niezgodnością liczby elektronów

res_name	local_res_atom_non_h_electron_sum	dict_atom_non_h_electron_sum	odds
NAG	2508452	2715080	206628
CLA	1809084	1961440	152356

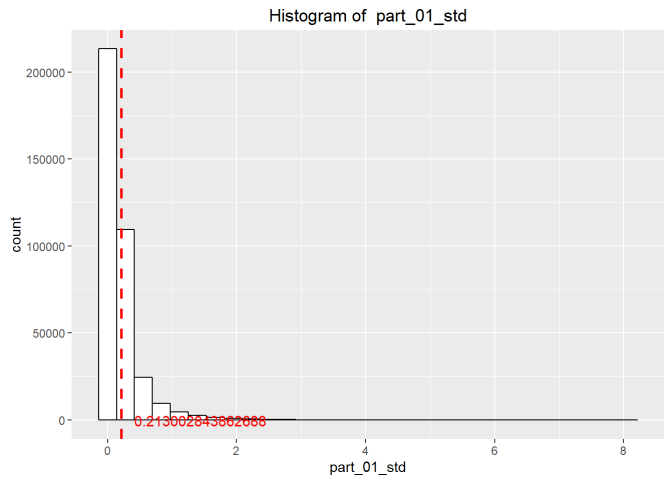
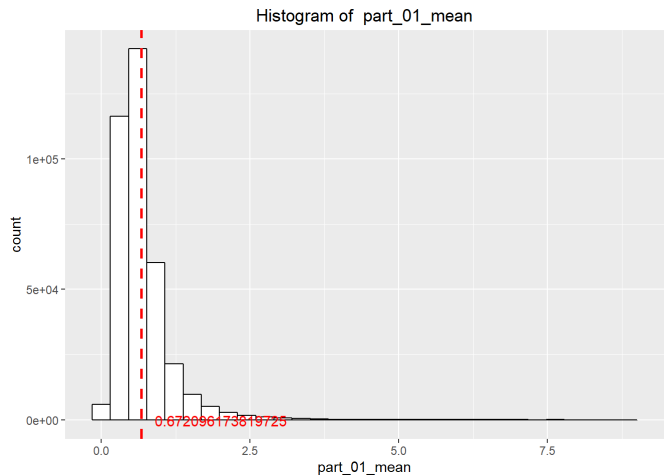
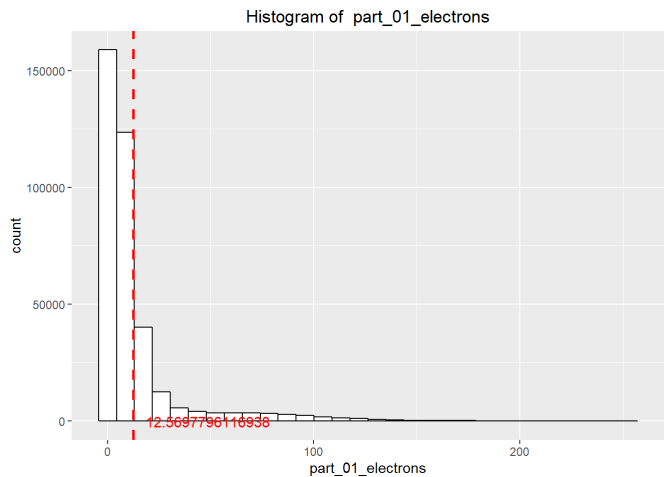
res_name	local_res_atom_non_h_electron_sum	dict_atom_non_h_electron_sum	odds
1PE	191894	230688	38794
MLY	238696	273702	35006
NAP	1217531	1247780	30249
COA	764523	794612	30089
NAD	1406030	1426817	20787
MAN	218316	238644	20328
PG4	224042	243584	19542
NDP	735538	749736	14198

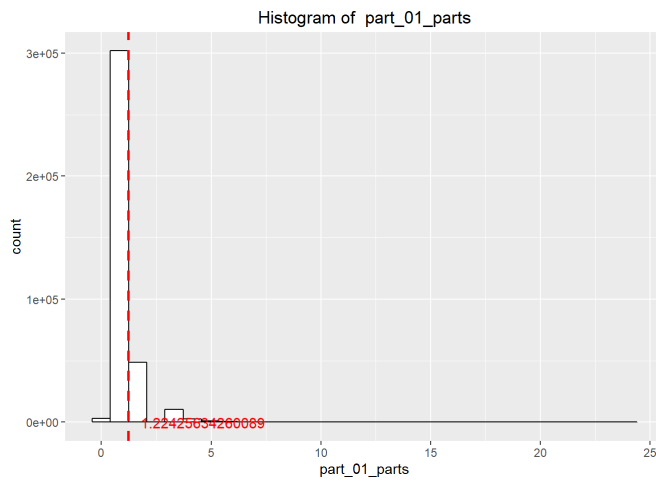
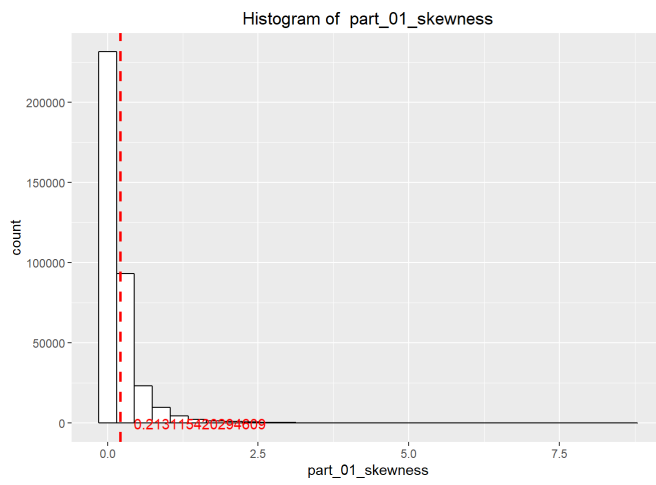
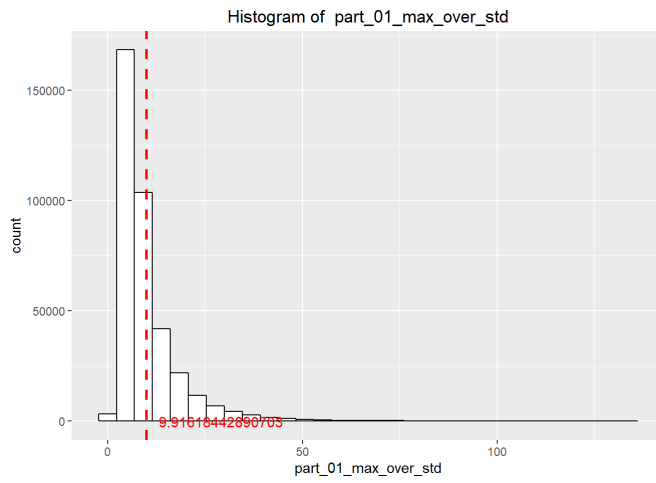
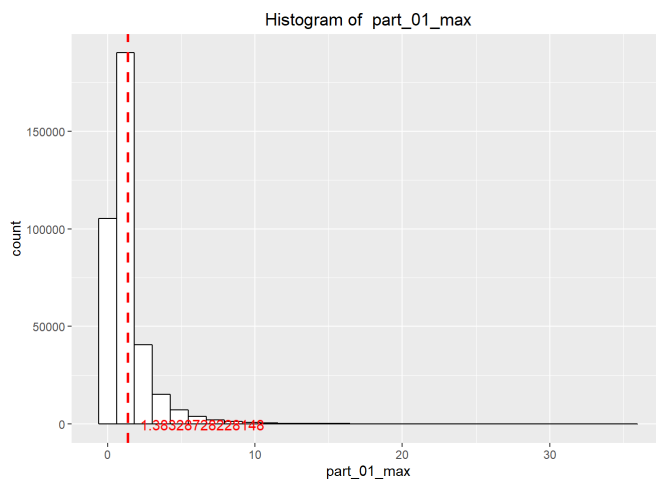
Rozkład wartości kolumn part_01

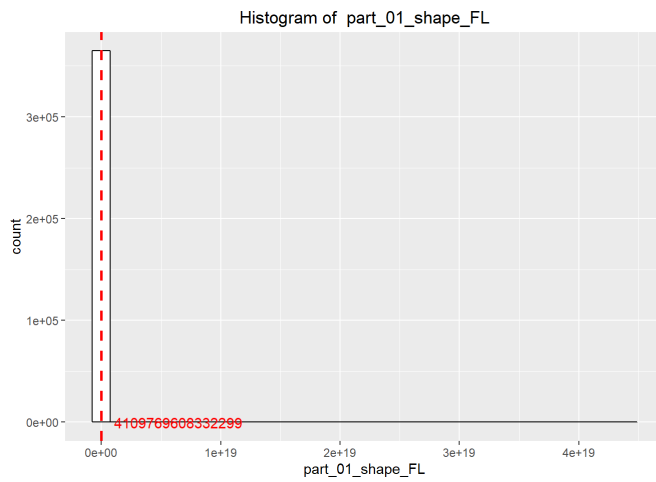
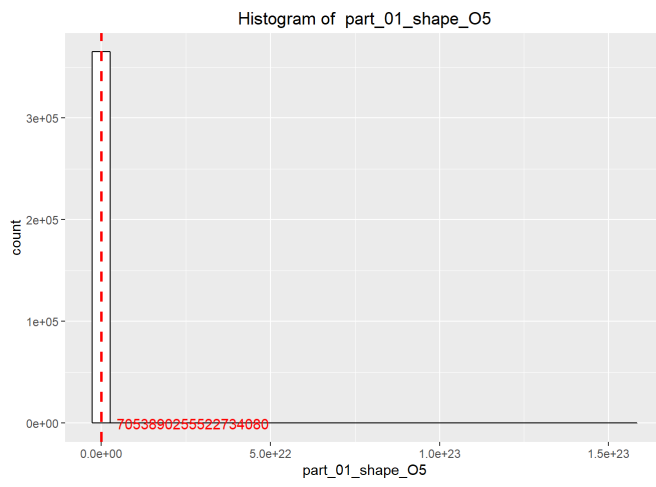
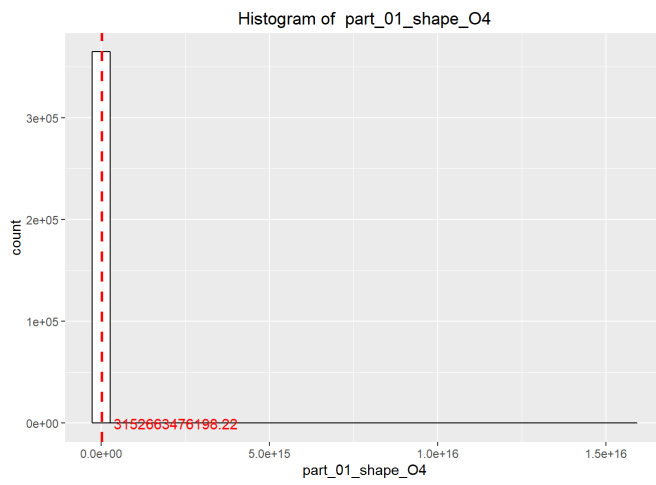
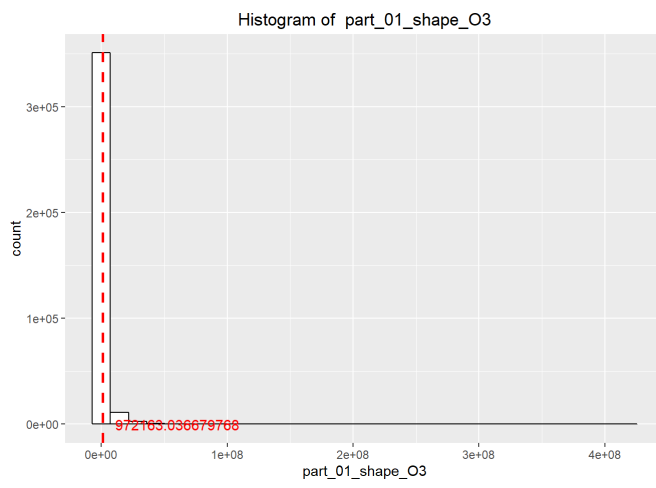
Sekcja przedstawia rozkład wartości wszystkich kolumn zaczynających się od part_01

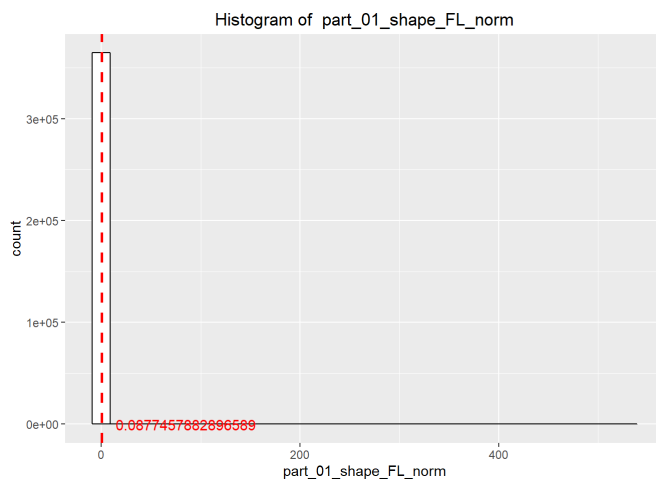
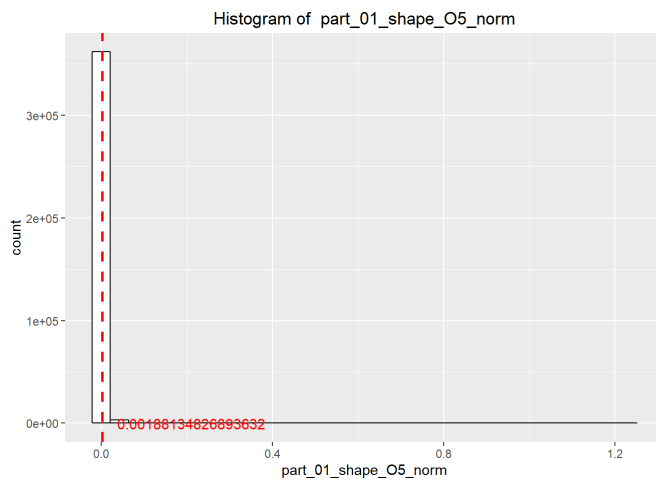
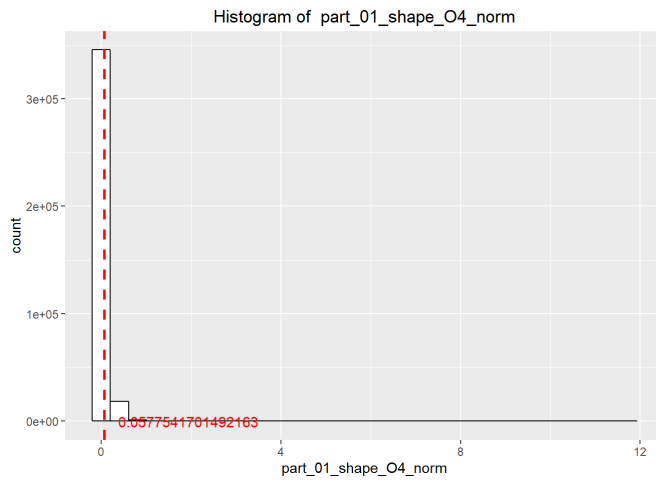
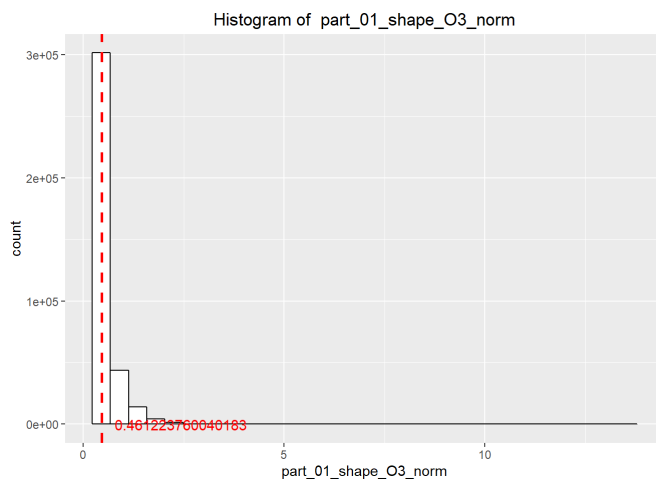
Usunięto wartości NA dla każdej kolumny z osobna. Nie zostały zamienione na wartość 0 by nie zaburzać rozkładu zmiennych.

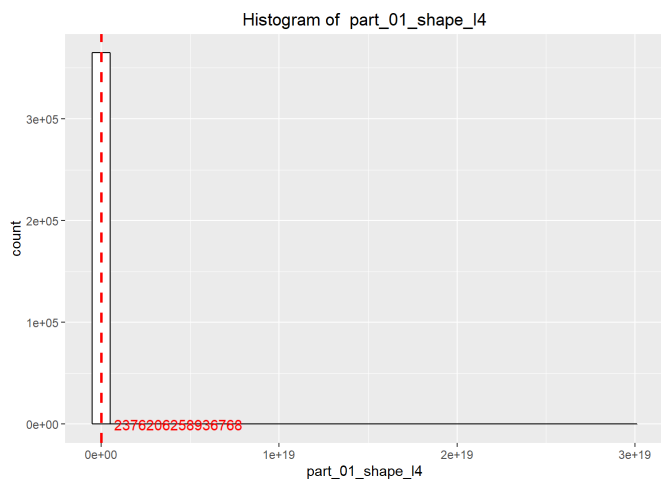
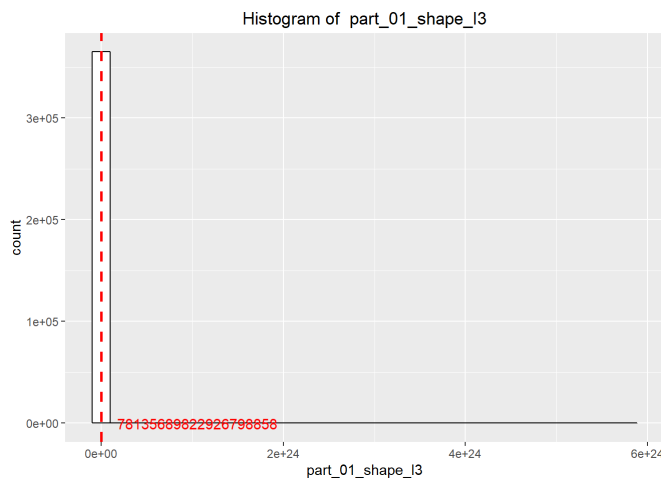
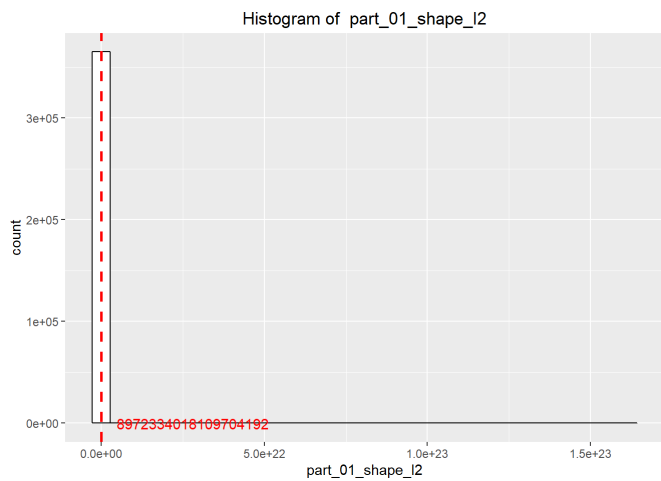
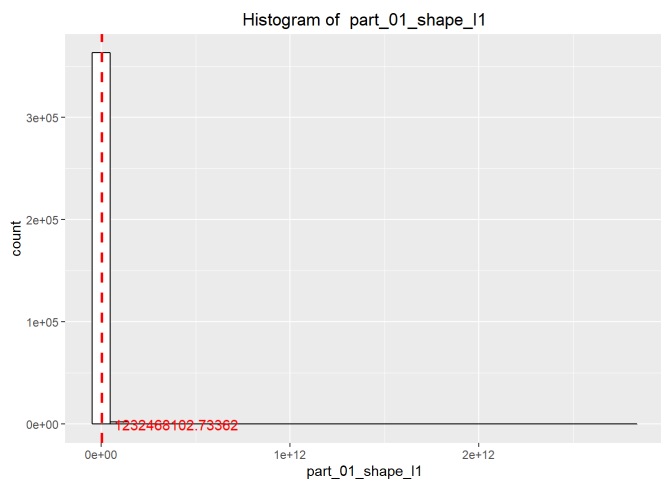
Na wykresach zaznaczono średnią wartość zmiennej (w formie graficznej oraz liczbowej).

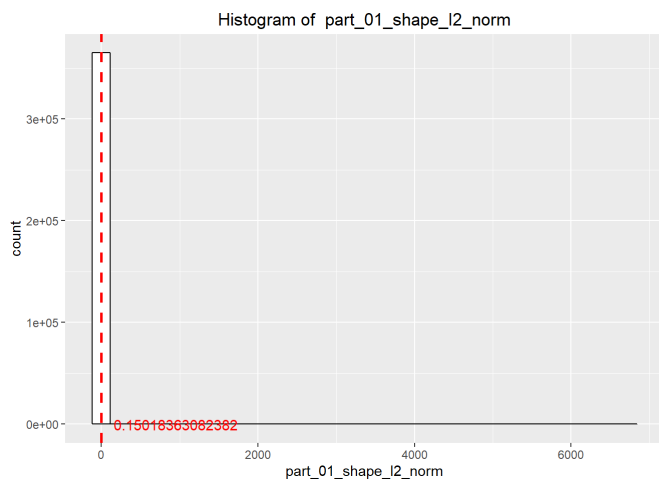
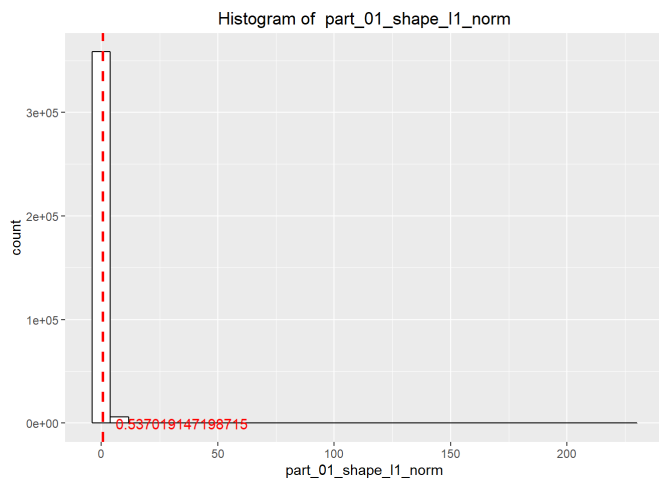
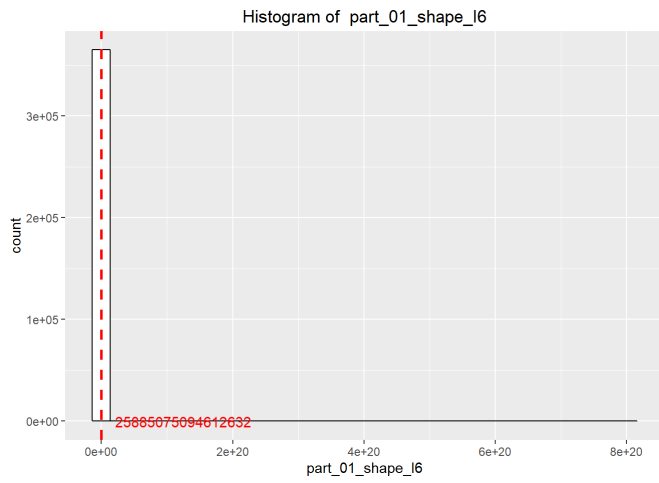
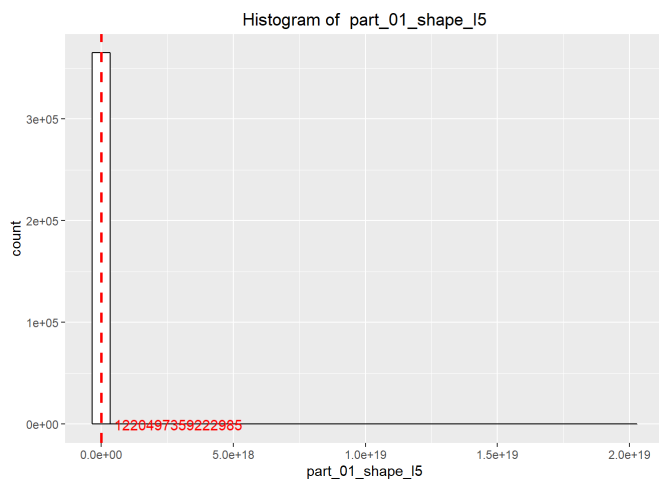


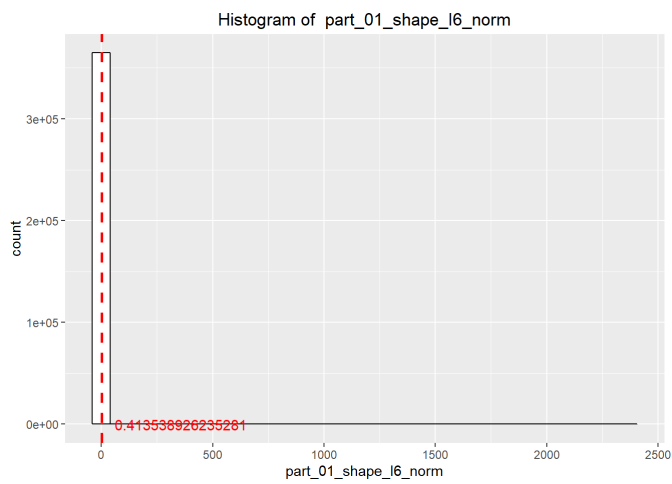
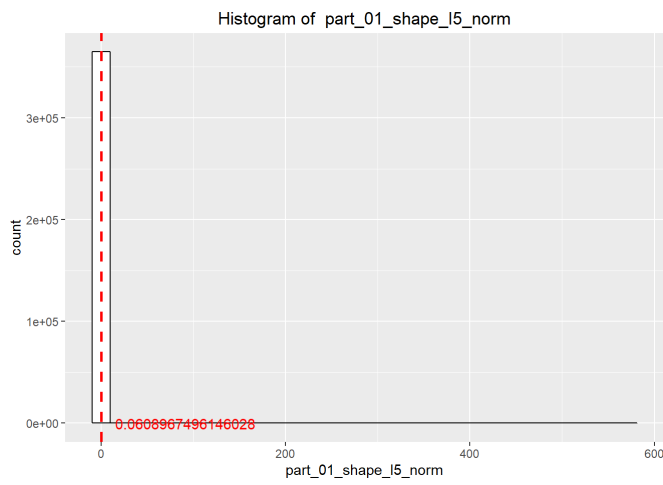
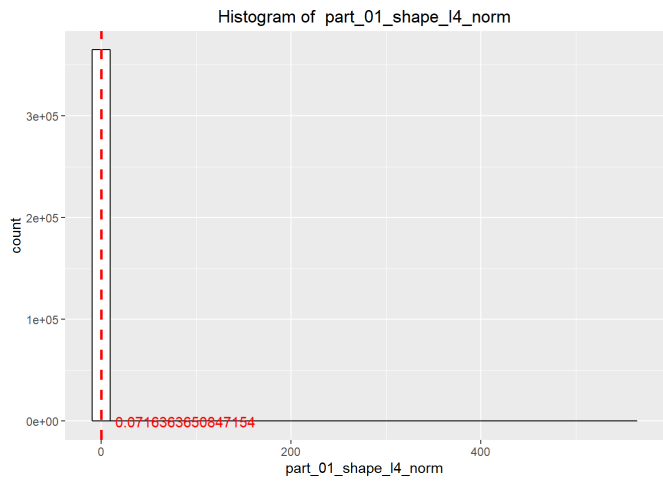
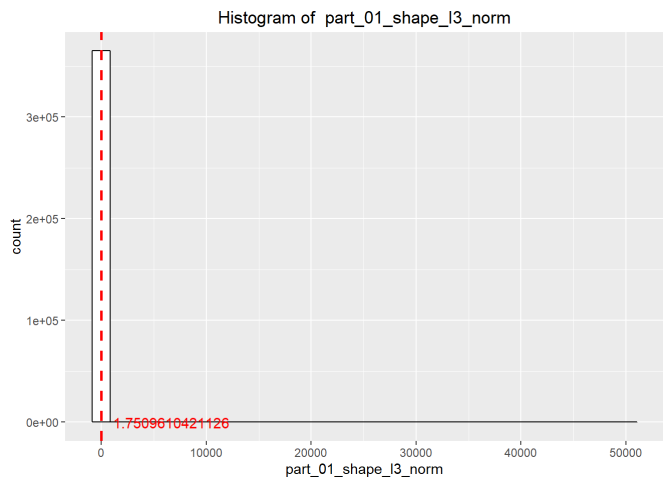


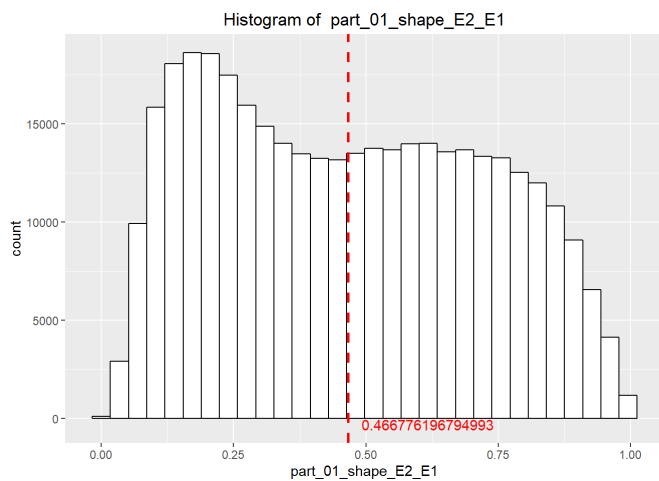
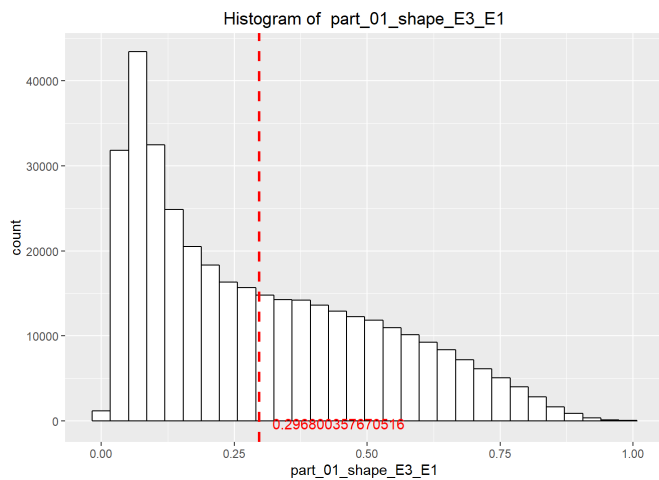
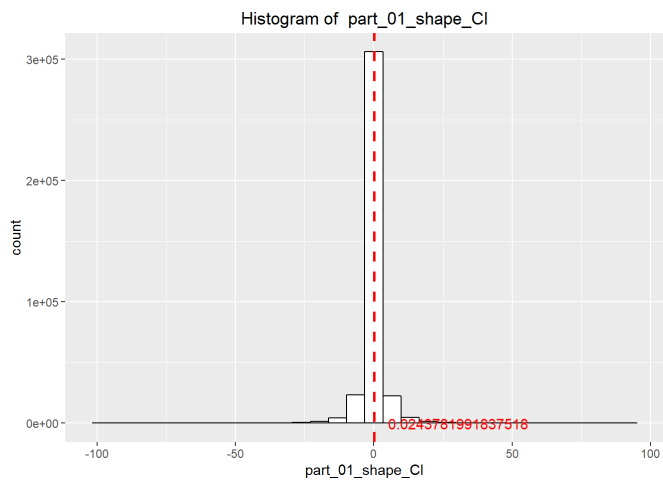
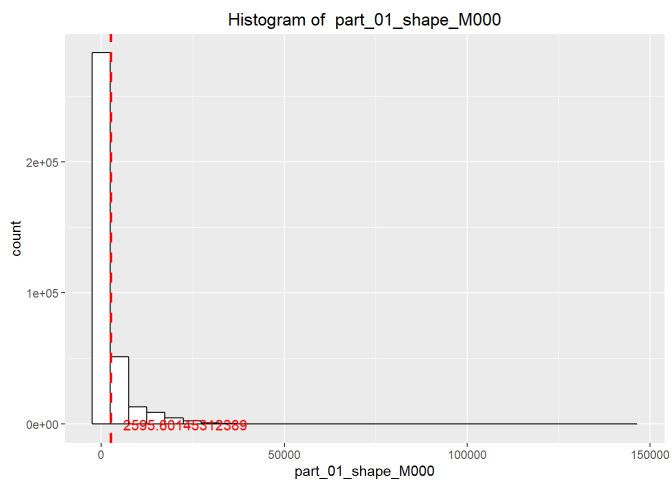


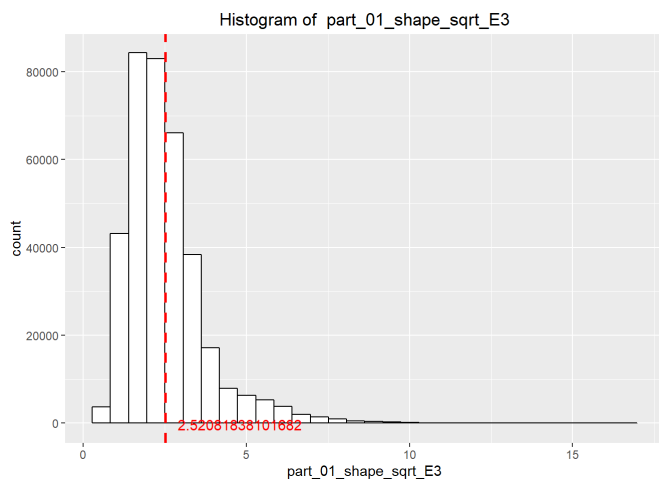
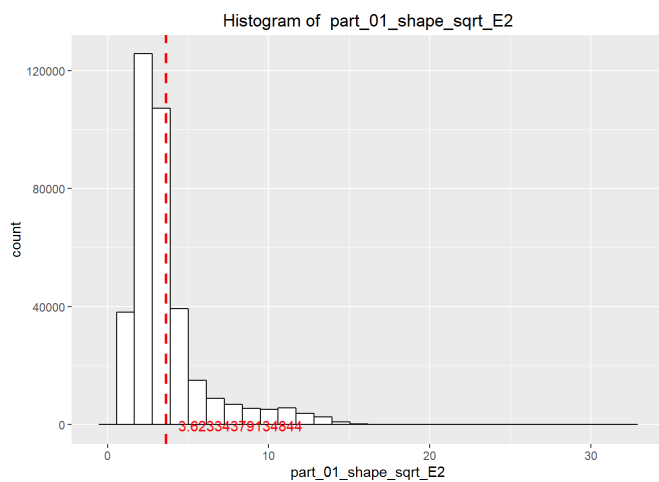
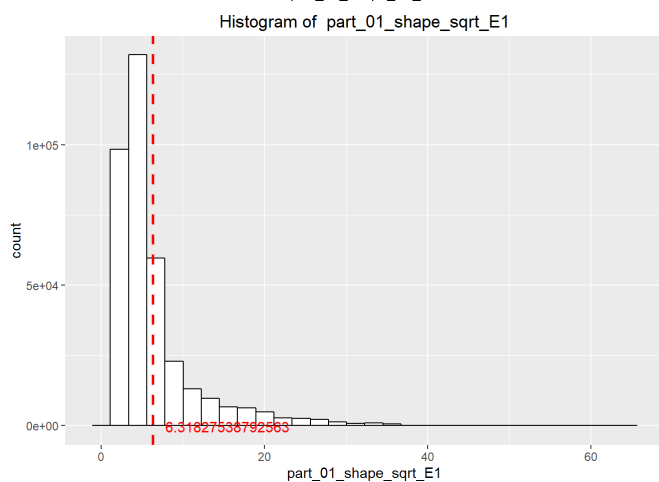
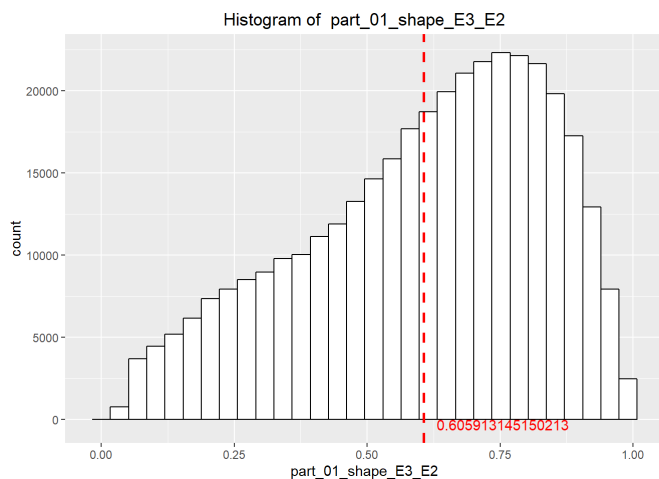


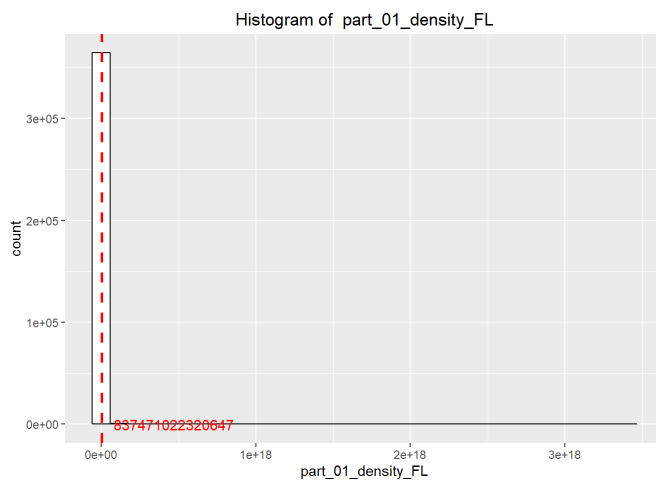
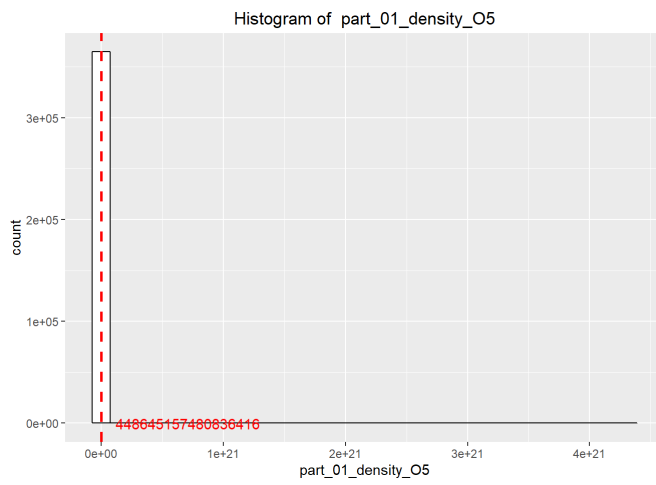
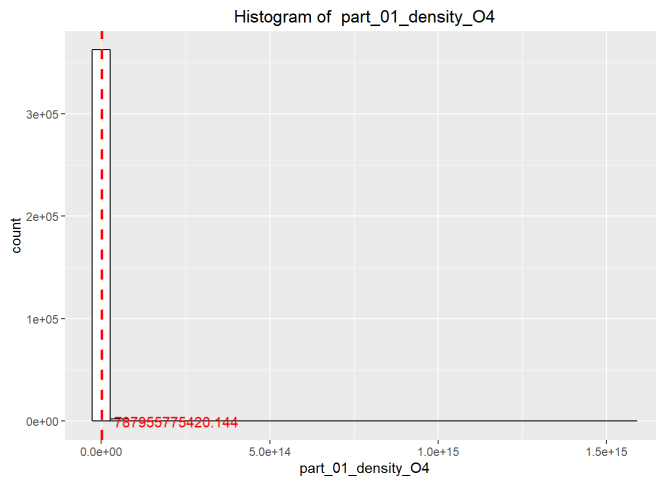
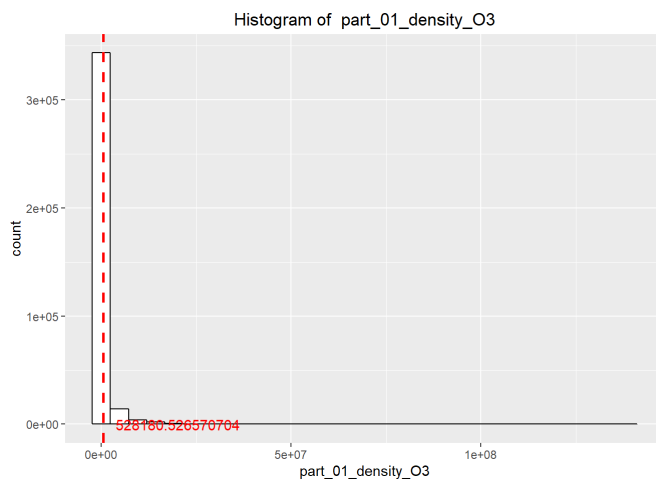


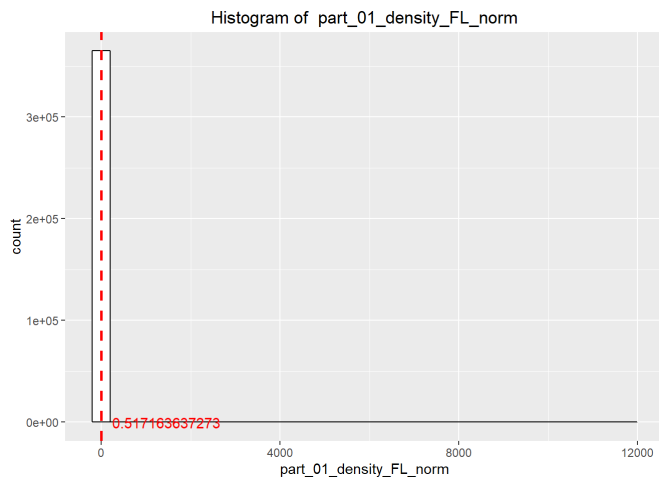
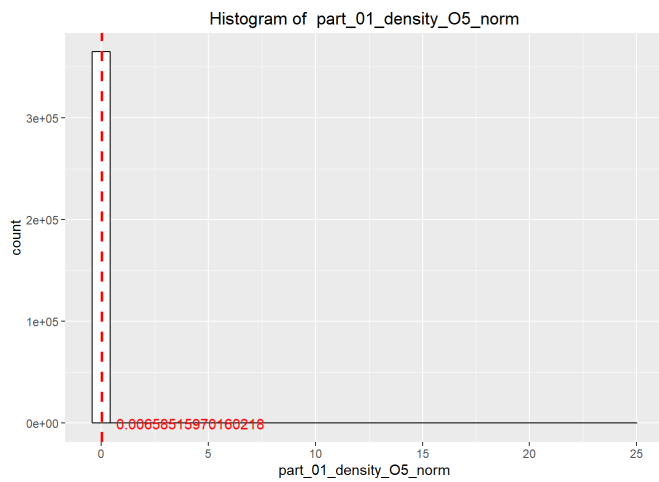
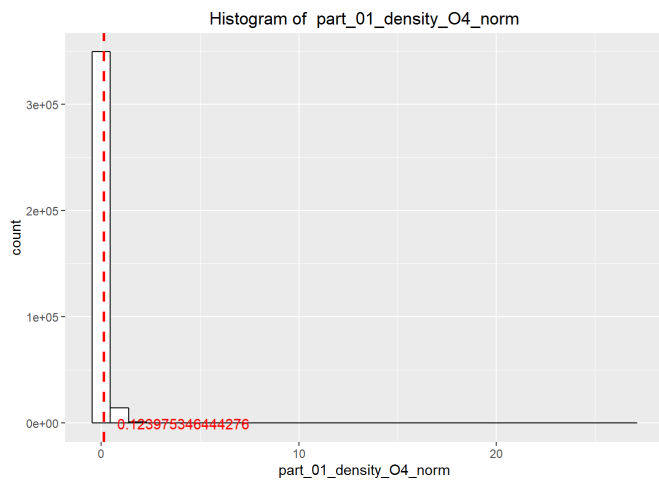
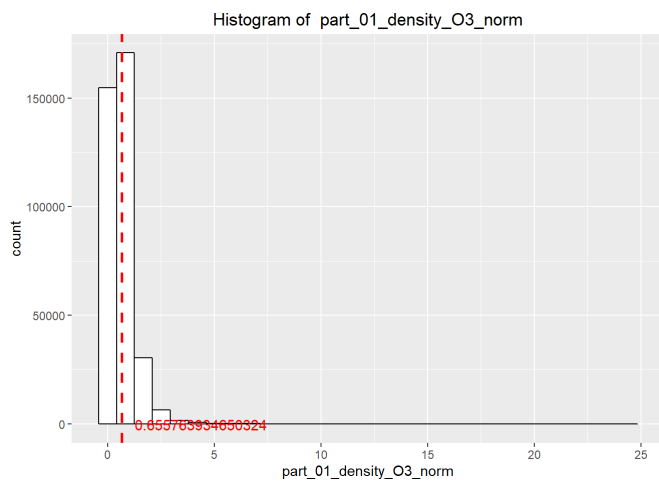


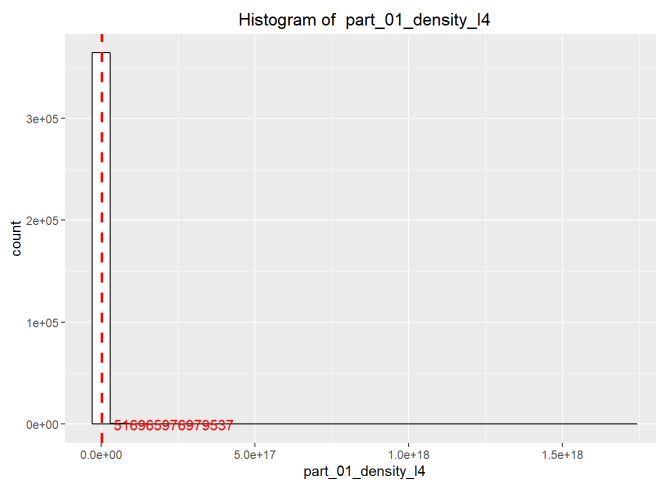
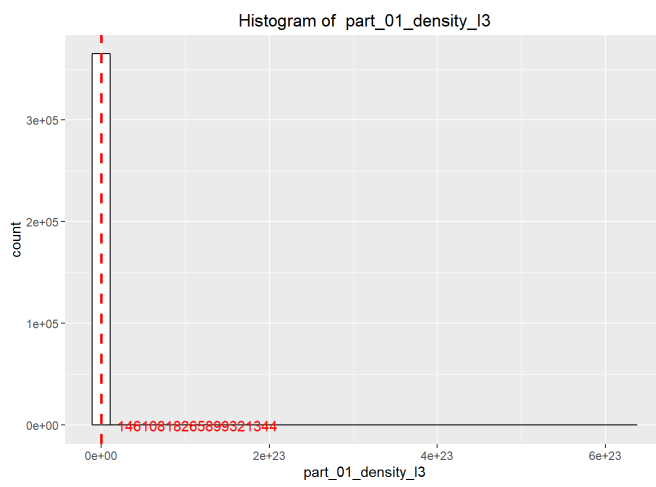
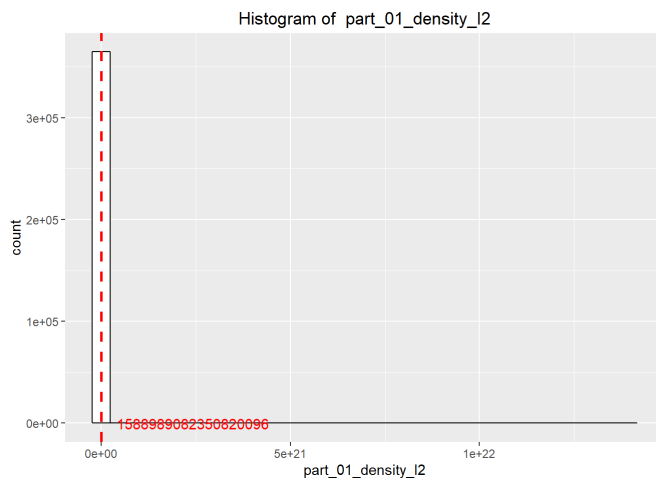
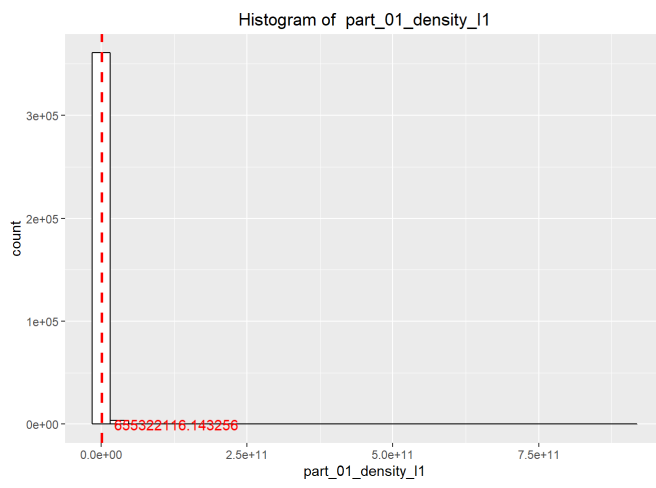


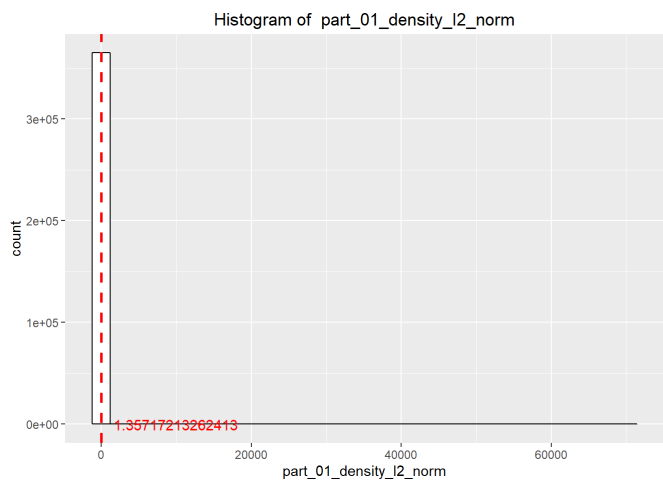
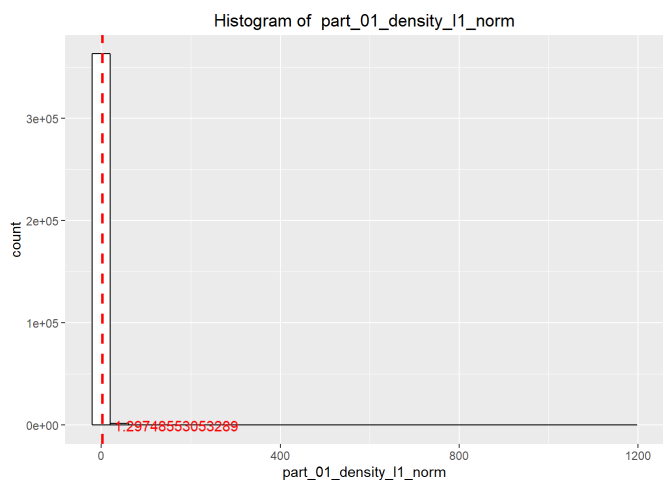
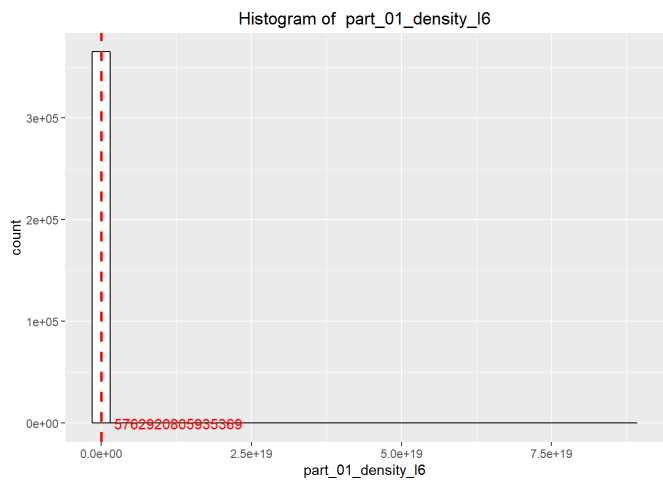
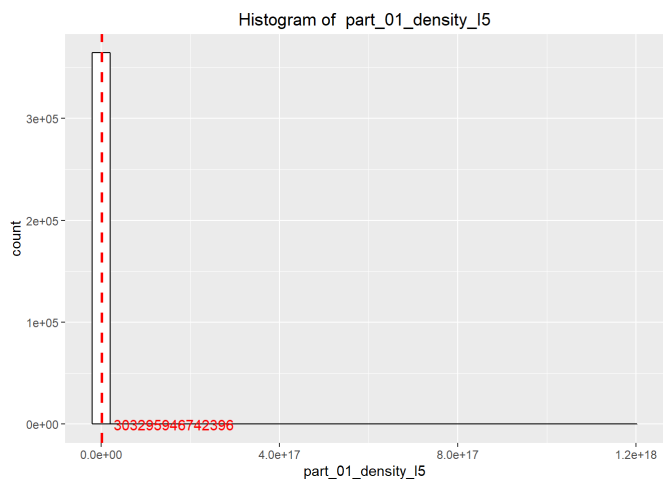


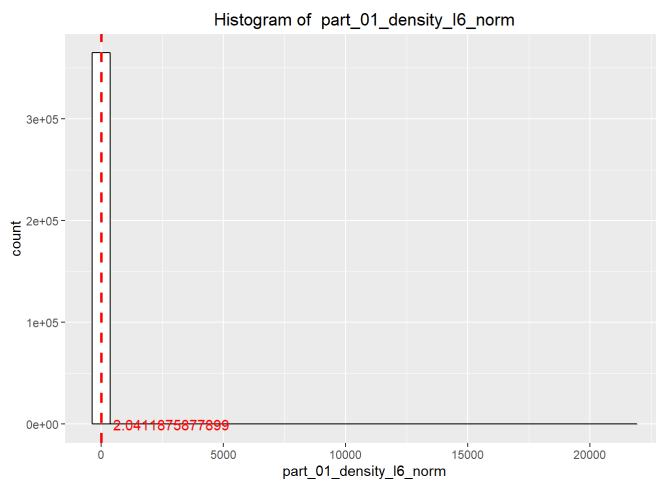
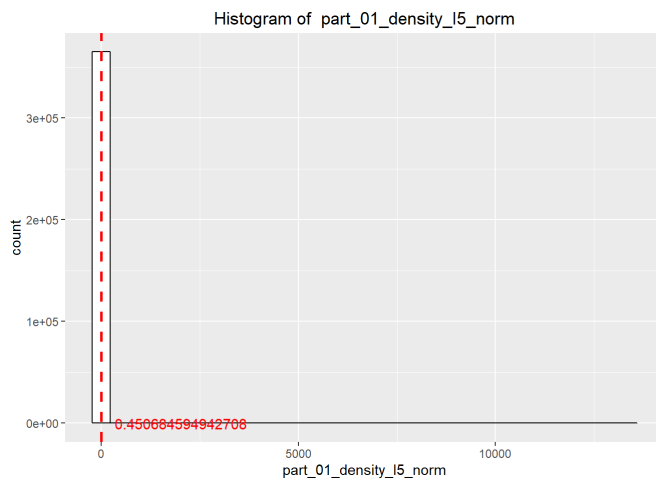
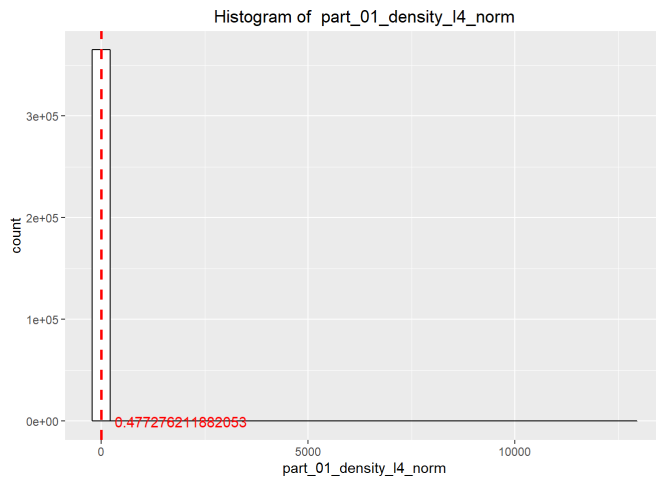
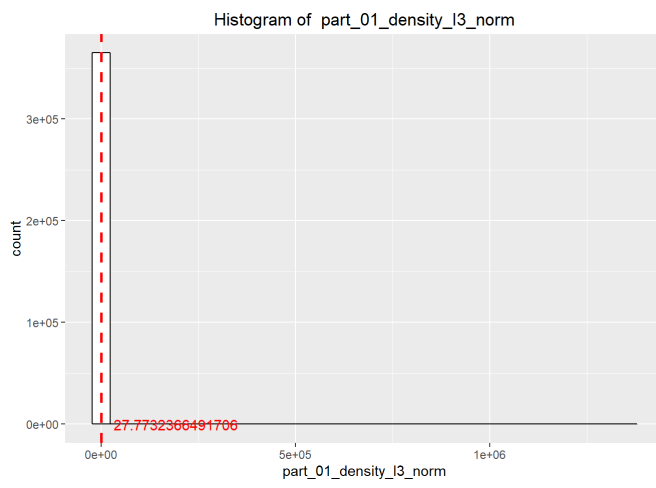


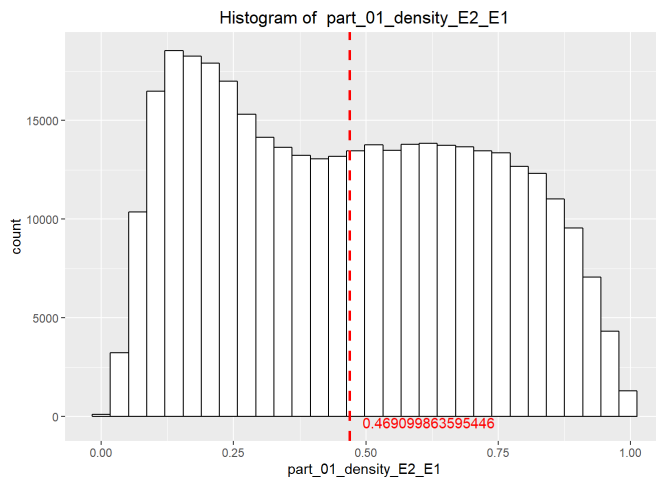
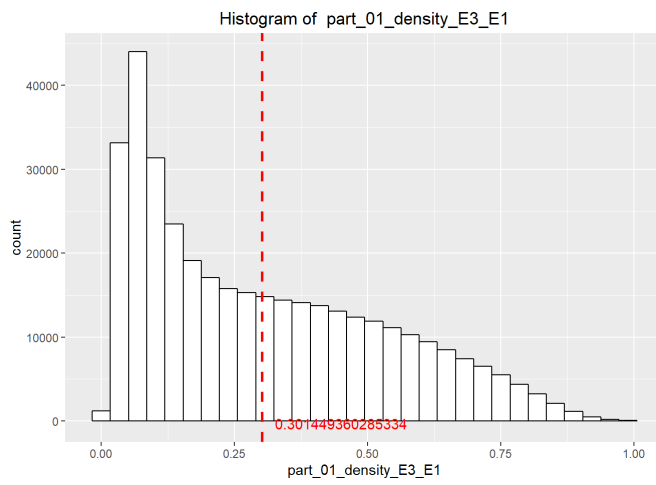
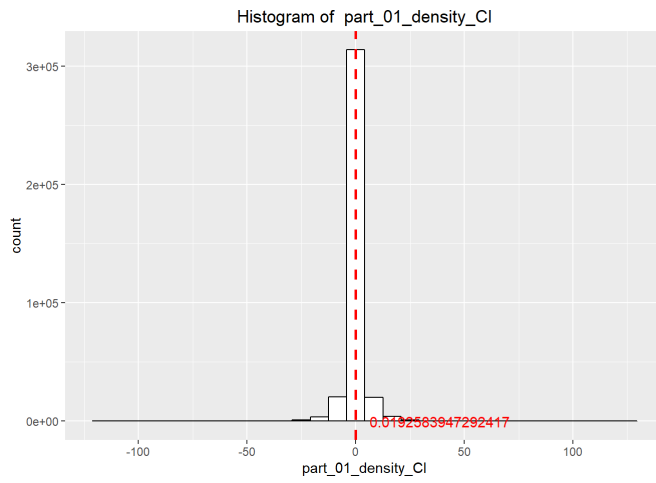
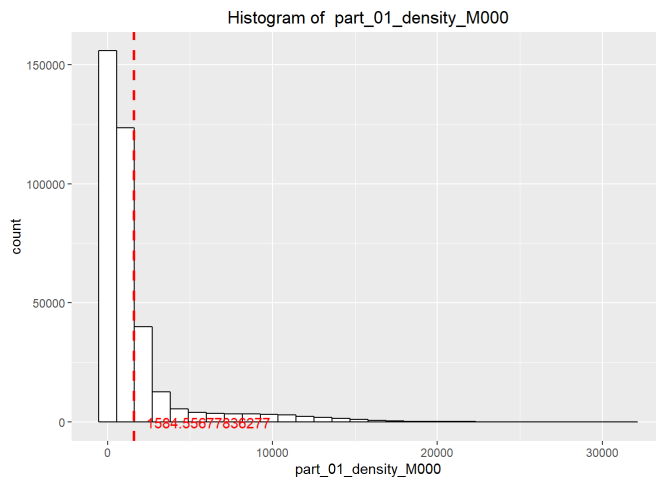


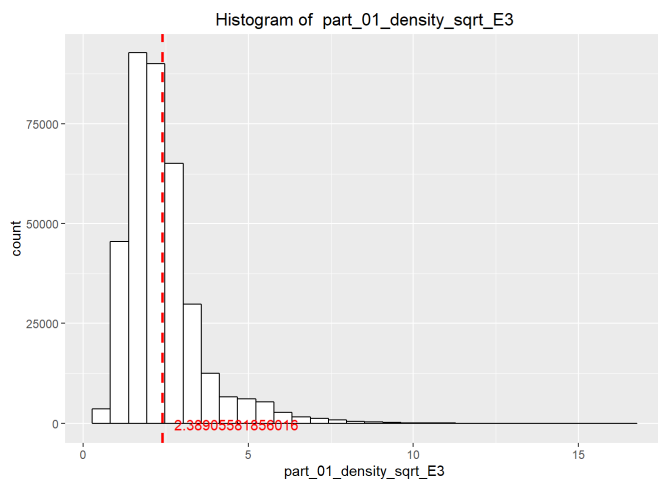
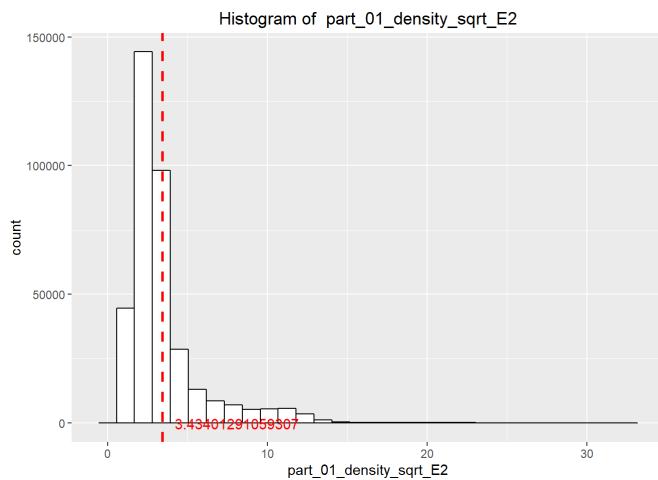
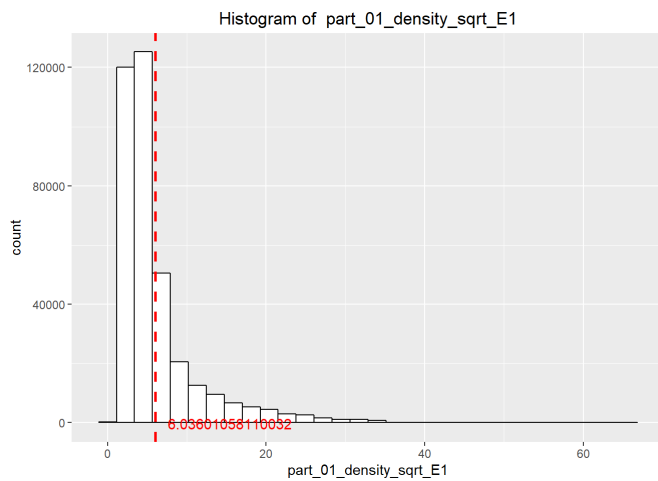
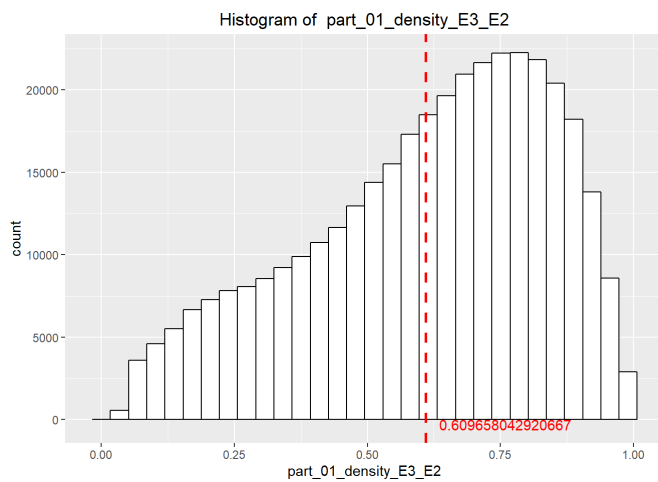












Wykres interaktywny

```
ggplotly(qplot(local_res_atom_non_h_electron_sum, data=pdb_code_res_name))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Przewidywanie liczby elektronów i atomów na podstawie innych kolumn

```
lm_data <- rawData
lm_data[is.na(lm_data)] <- 0
lm_data <- dplyr::select_if(lm_data, is.numeric)

lm_atom_model <- lm(local_res_atom_non_h_count ~ ., lm_data)
lm_atom_summary <- summary(lm_atom_model)

lm_electron_model <- lm(local_res_atom_non_h_electron_sum ~ ., lm_data)
lm_electron_summary <- summary(lm_electron_model)

pdb_code_res_name <- pdb_code_res_name[ , -which(names(pdb_code_res_name) %in% c("blob_coverage", "res_coverage", "
pdb_code", "res_id", "chain_id", "skeleton_data", "fc_col", "fo_col", "weight_col", "title"))]

pdb_code_res_name$res_name <- as.character(pdb_code_res_name$res_name)
pdb_code_res_name$res_name <- as.factor(pdb_code_res_name$res_name)
pdb_code_res_name[is.na(pdb_code_res_name)] <- -1000000
```

Miary dla liczby atomów:

R²: 0.9999915

RMSEL 0.0390963

Miary dla liczby elektronów:

R²: 0.9999877

RMSEL 0.3168246

Klasyfikator