

# Untitled

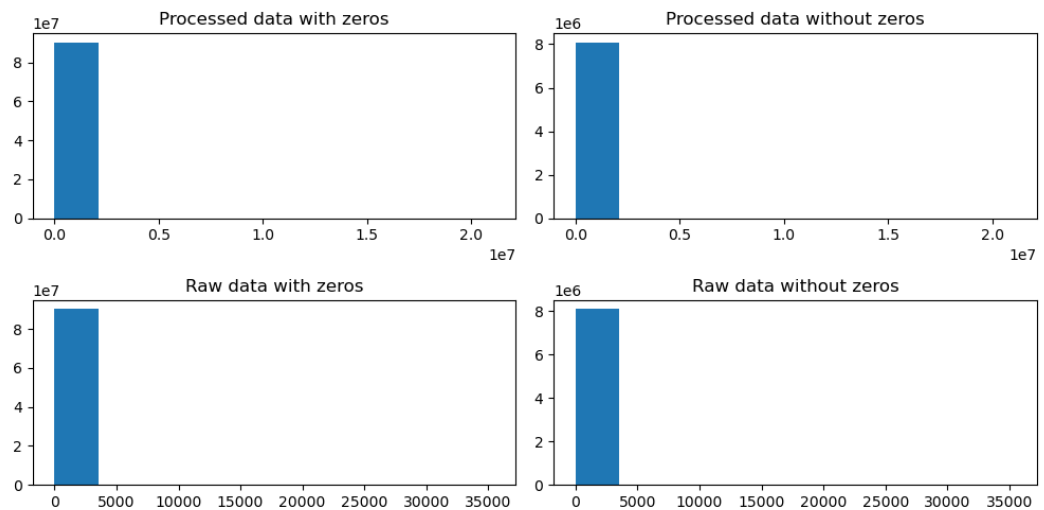
December 15, 2022

## 0.1 Projekt 1

### 0.1.1 Paulina Kucharewicz

#### 0.1.2 1. Exploration

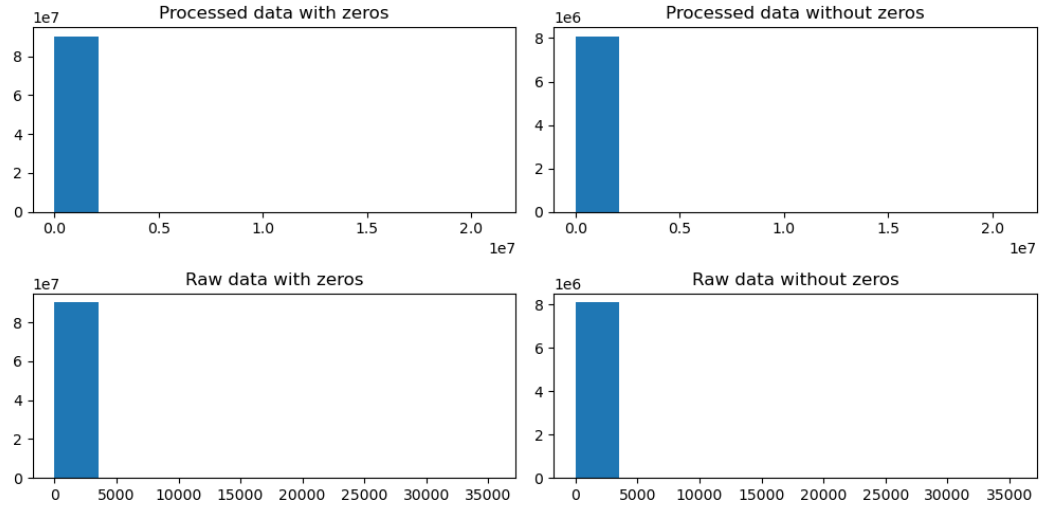
1. There are 72208 observations and 5000 variables in the train data set, and 18052 observations and 5000 variables in the test data set.



#### 2. Histograms |

| |:-:| | Figura 1: |

3. Data was not normalised to 10k reads and was not log1p transformed. Data also was not scaled to unit variance.



4. Histograms |  
| |:-:| | Figura 1: |

5. This data set has count data. Such data can be described with Poisson or negative binomial distributions. However, in Poisson distribution mean and variance are equal and for this data set these values are very different. Our data have negative binomial distribution, possibly zero-inflated. Abundance of zero can be explained by no transcription of genes or absence of mRNA (caused by faster mRNA degradation than transcription). Gene expression is different in different type of cells, some genes are very rarely expressed in general and some are not expressed in certain cell types. Many genes are also not expressed constantly, but only under certain conditions (gene regulation). All of the above can lead to elevated levels of read counts equal to zero.
6. Object `adata.obs` is an annotation of observations. Each row corresponds to one of the cells from experiment and contains information such as type of a cell, its donor, batch etc.. From this data frame it can be concluded that there were: