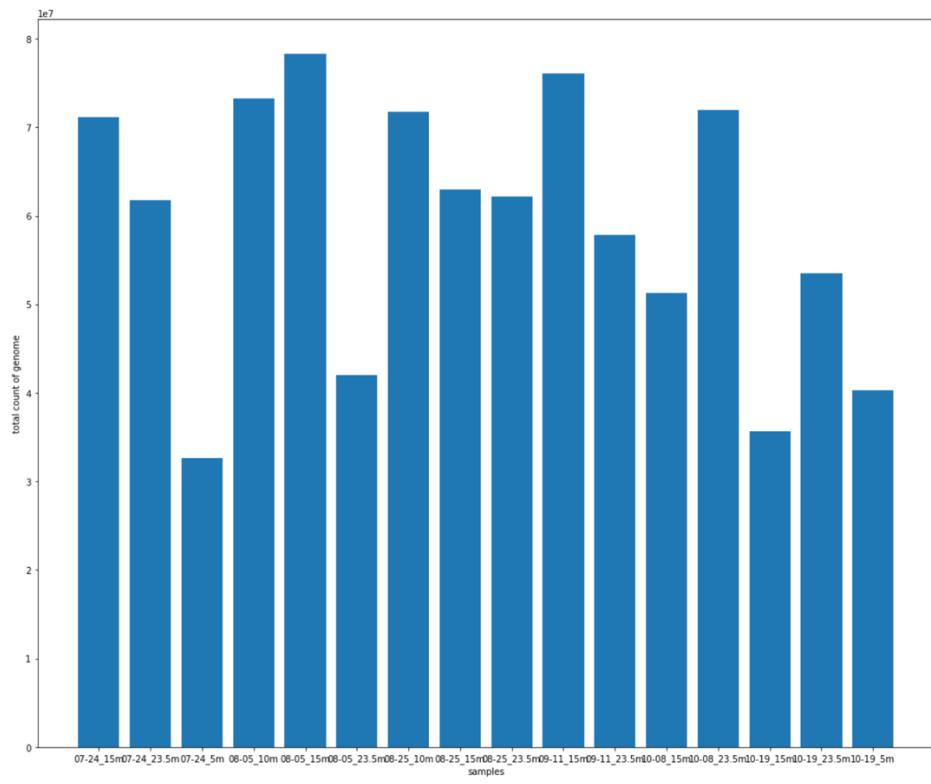


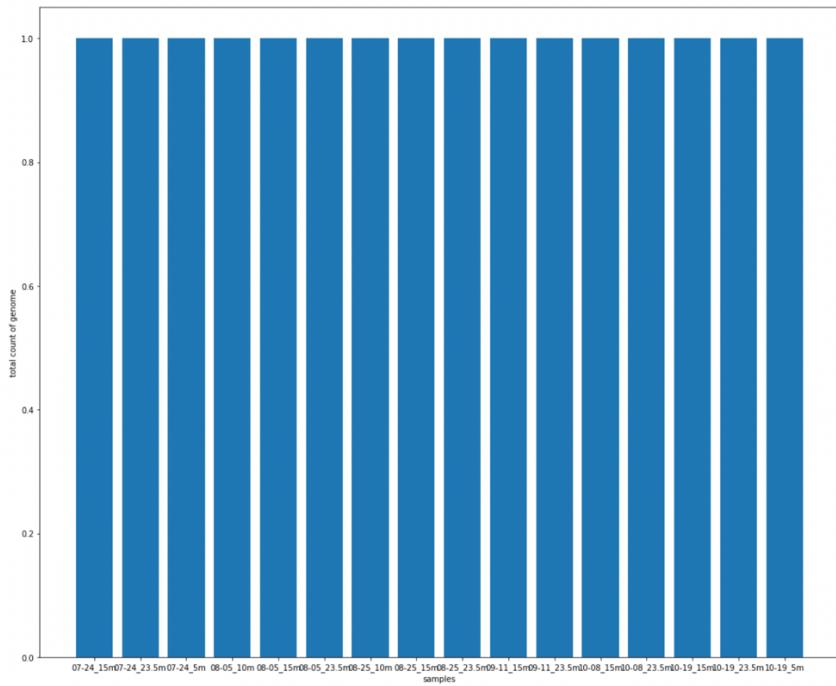
lake-microbiome

by Sophia Yang
April 9th, 2023

All the analysis were done on top the count dataset(being normalized)
("coverm_431_MAGS_metagenomes_reads_count.csv")



This is the count for all the genome per sample before normalization



This is the count for all the genome per sample after normalization (sanity check, should sum to 1)

All the other normalized datasets, since different normalization method is being applied, this might be the reason that the row sum is not equal to 1, and some of the datasets have a vast majority of zeros.

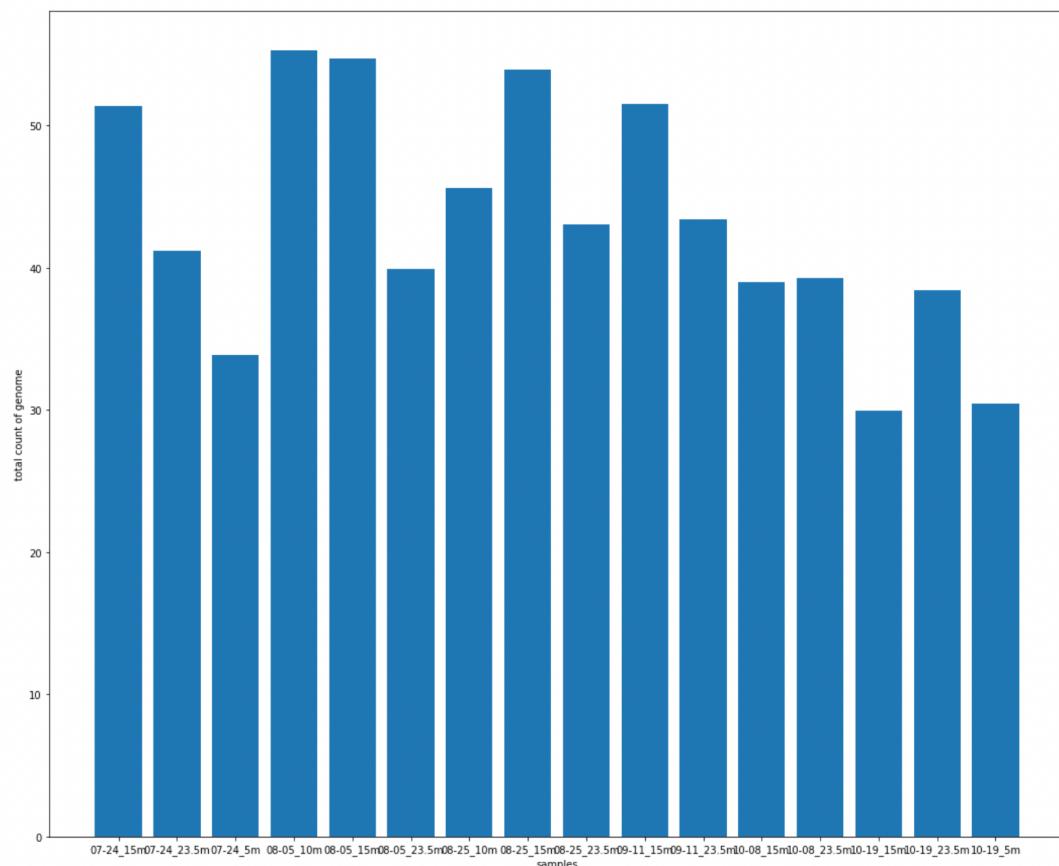
- Fw_lake mendota data(envirrmental)
 - MAG_abundance_matrix_rel_abund.tsv
 - MAG_RPKM_normalized.tsv
 - phages_abundance_matrix_metagenomes.tsv
 - phages_rpkm_normalized_matrix.tsv

I did analysis based on these four datasets to check the row sum and number of zeros...

MAG_abundance_matrix_rel_abund.tsv

```
1 row_sums_MAG_abundance = MAG_abundance.sum(axis=1)
2 row_sums_MAG_abundance
```

2020-07-24_15m	51.388456
2020-07-24_23.5m	41.211519
2020-07-24_5m	33.847230
2020-08-05_10m	55.278206
2020-08-05_15m	54.719929
2020-08-05_23.5m	39.912086
2020-08-25_10m	45.643490
2020-08-25_15m	53.954127
2020-08-25_23.5m	43.067711
2020-09-11_15m	51.559964
2020-09-11_23.5m	43.448078
2020-10-08_15m	38.991339
2020-10-08_23.5m	39.313672
2020-10-19_15m	29.948284
2020-10-19_23.5m	38.422739
2020-10-19_5m	30.432611

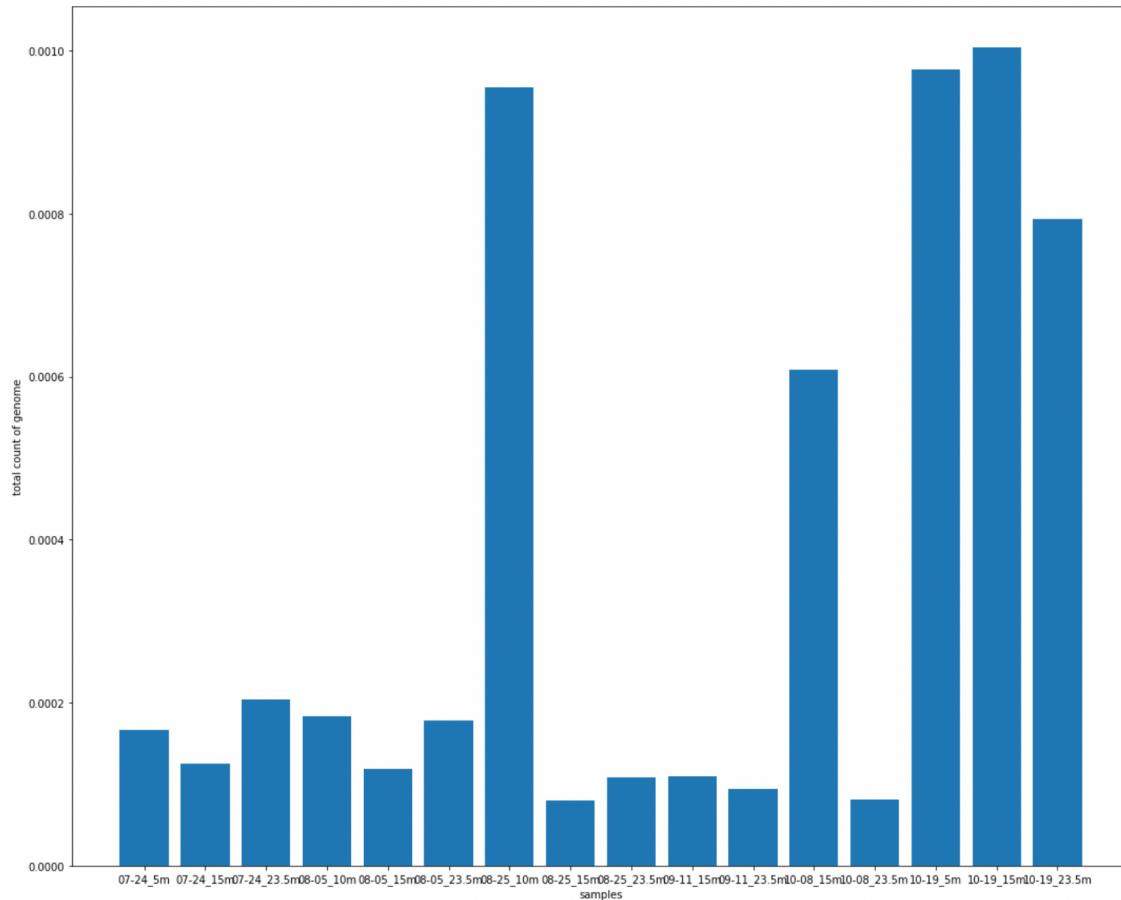


the number of zeros: 1384
Percentage of zeros: 20.07%

MAG_RPKM_normalized.tsv

```
1 row_sums_MAG_RPKM_normalized = MAG_RPKM_normalized.sum(axis=1)  
2 row_sums_MAG_RPKM_normalized
```

2020-07-24_5m	0.000166
2020-07-24_15m	0.000125
2020-07-24_23.5m	0.000204
2020-08-05_10m	0.000183
2020-08-05_15m	0.000118
2020-08-05_23.5m	0.000178
2020-08-25_10m	0.000955
2020-08-25_15m	0.000080
2020-08-25_23.5m	0.000108
2020-09-11_15m	0.000110
2020-09-11_23.5m	0.000094
2020-10-08_15m	0.000608
2020-10-08_23.5m	0.000082
2020-10-19_5m	0.000977
2020-10-19_15m	0.001004
2020-10-19_23.5m	0.000793

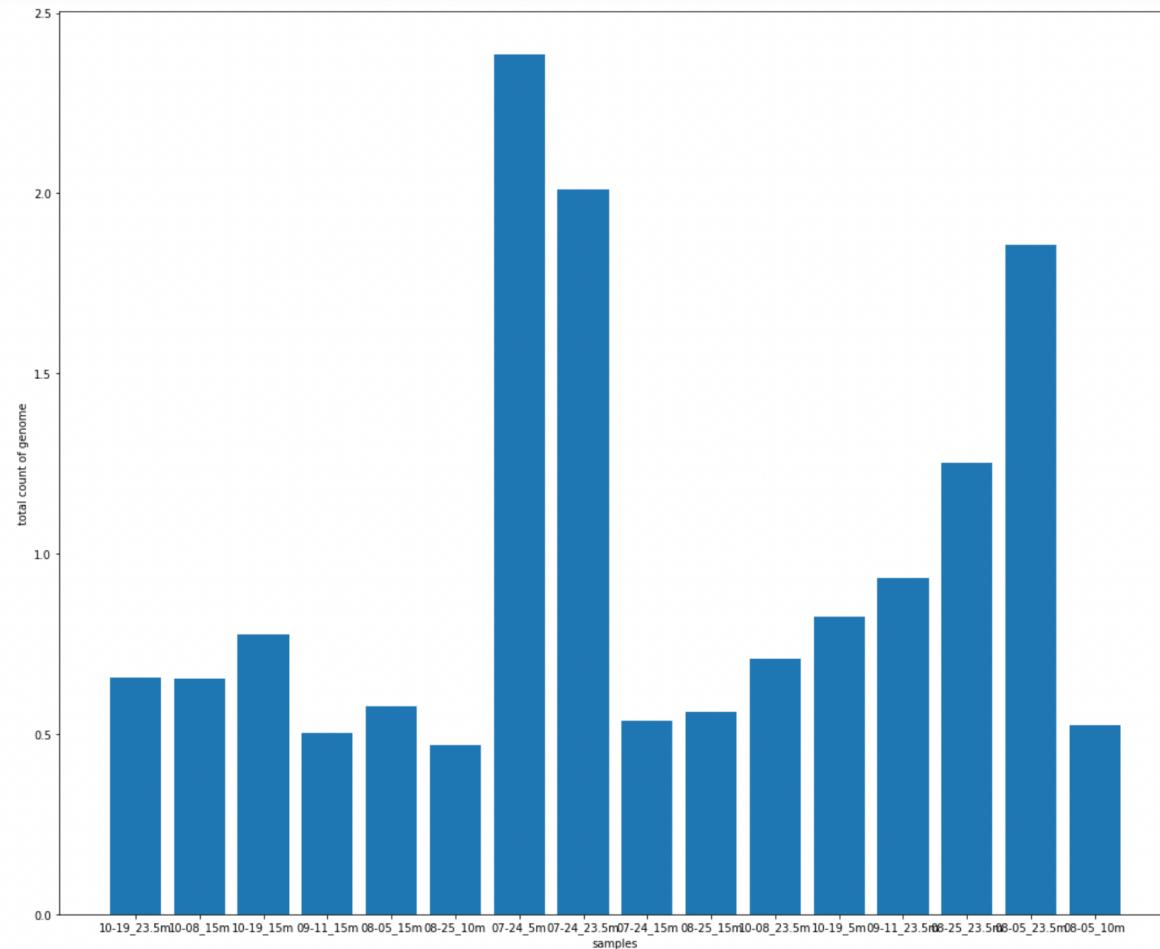


the number of zeros: 5893
Percentage of zeros: 85.46%

phages_abundance_matrix_metagenomes.tsv

```
1 | row_sums_phages_abundance_matrix_metagenomes= phages_abundance_matrix_metagenomes.sum(axis=1)  
2 | row_sums_phages_abundance_matrix_metagenomes
```

```
2020-10-19_23.5m      0.655490  
2020-10-08_15m       0.653947  
2020-10-19_15m       0.777846  
2020-09-11_15m       0.504348  
2020-08-05_15m       0.577252  
2020-08-25_10m       0.468715  
2020-07-24_5m        2.384052  
2020-07-24_23.5m     2.011944  
2020-07-24_15m       0.536729  
2020-08-25_15m       0.560599  
2020-10-08_23.5m     0.709240  
2020-10-19_5m        0.826225  
2020-09-11_23.5m     0.932316  
2020-08-25_23.5m     1.253326  
2020-08-05_23.5m     1.858483  
2020-08-05_10m       0.525313
```

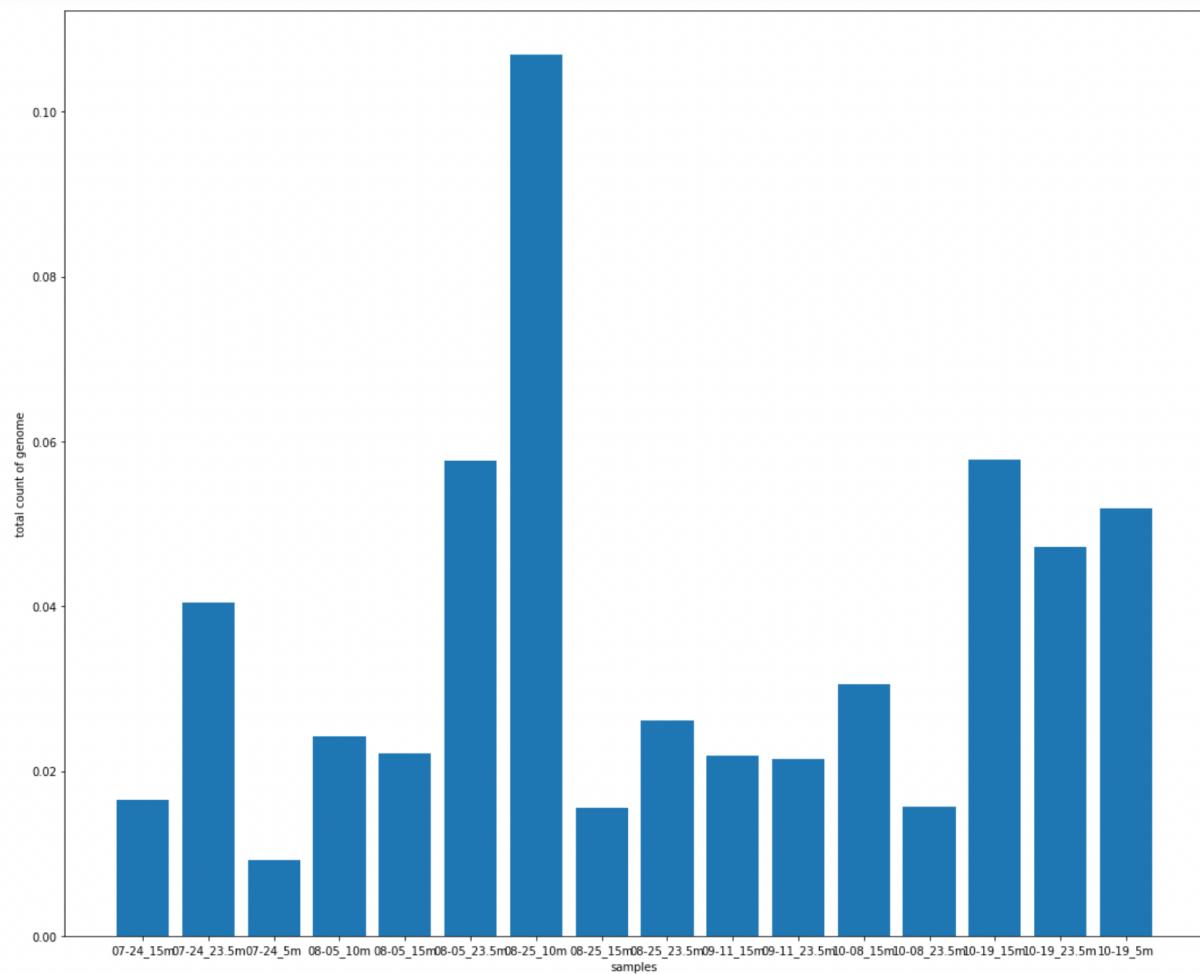


the number of zeros: 13125
Percentage of zeros: 36.52%

phages_rpkm_normalized_matrix.tsv

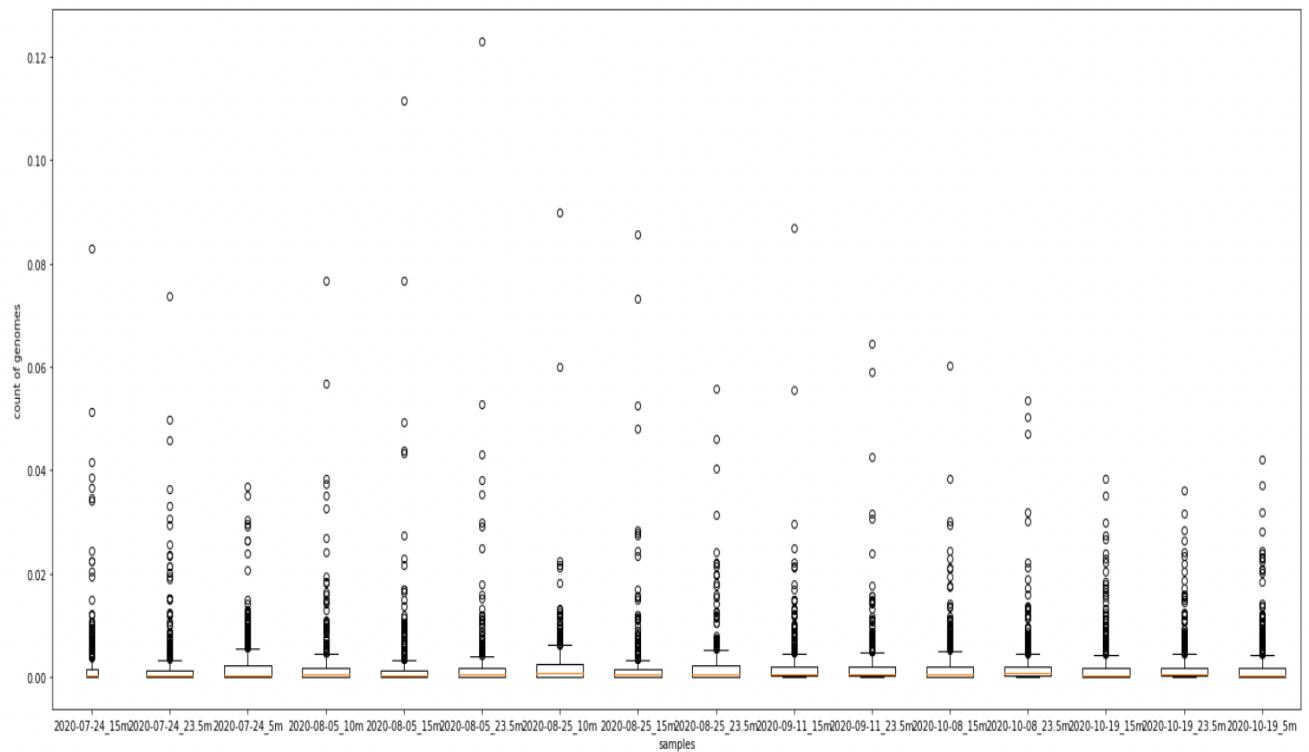
```
1 row_sums_phages_rpkm_normalized_matrix= phages_rpkm_normalized_matrix.sum(axis=1)
2 row_sums_phages_rpkm_normalized_matrix
```

2020-07-24_15m	0.016504
2020-07-24_23.5m	0.040392
2020-07-24_5m	0.009160
2020-08-05_10m	0.024225
2020-08-05_15m	0.022192
2020-08-05_23.5m	0.057709
2020-08-25_10m	0.106863
2020-08-25_15m	0.015531
2020-08-25_23.5m	0.026158
2020-09-11_15m	0.021904
2020-09-11_23.5m	0.021430
2020-10-08_15m	0.030538
2020-10-08_23.5m	0.015622
2020-10-19_15m	0.057807
2020-10-19_23.5m	0.047207
2020-10-19_5m	0.051827



the number of zeros: 34983
Percentage of zeros: 97.35%

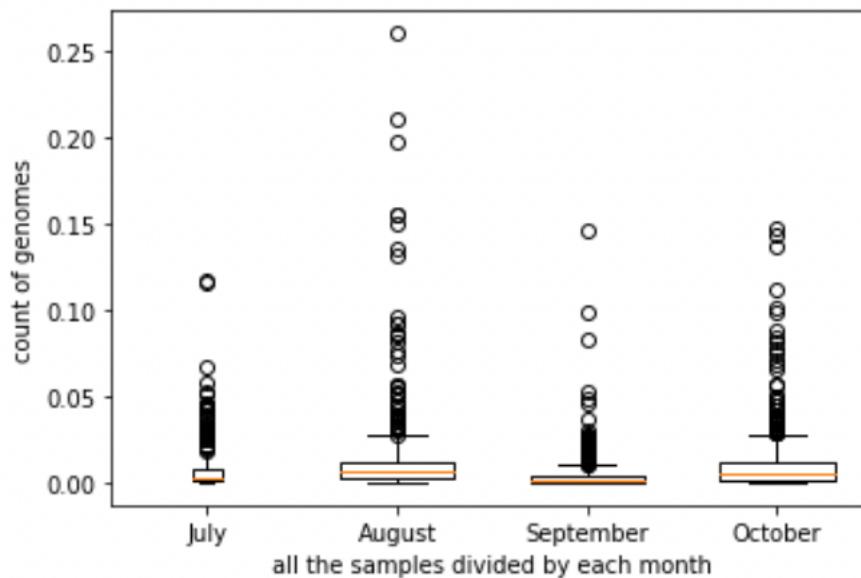
**Boxplot for distribution of count of genome per sample
("coverm_431_MAGS_metagenomes_reads_count.csv")**



More outliers in the middle seasons

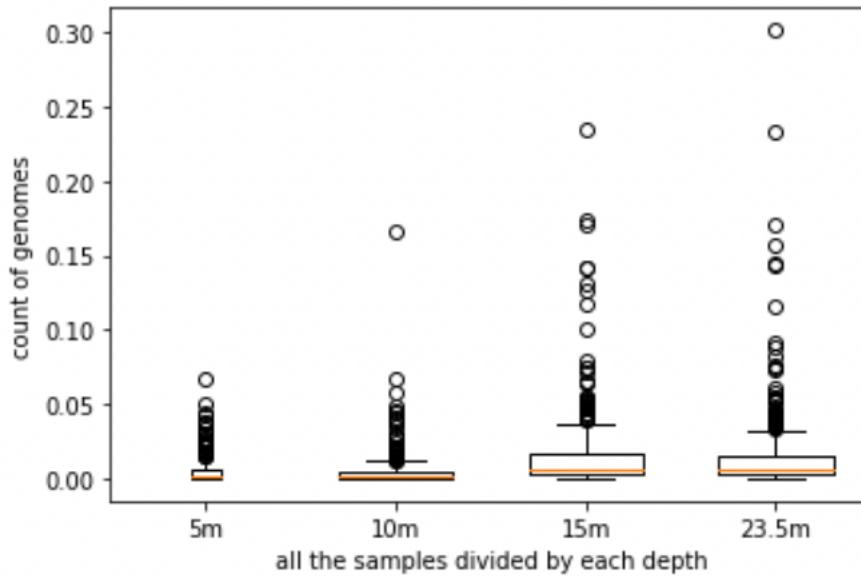
Sum of count of each genome categorized by each month

These figures are generated based on normalized data set (row sum method)



August has more outliers than all the other months. However more samples in August.... Outlier might be expected.

Sum of count of each genome categorized by each depth



As the depth became deeper, there are more outliers. However, since we have less sample data on depth 5 and 10m(2 for each), less count is expected.

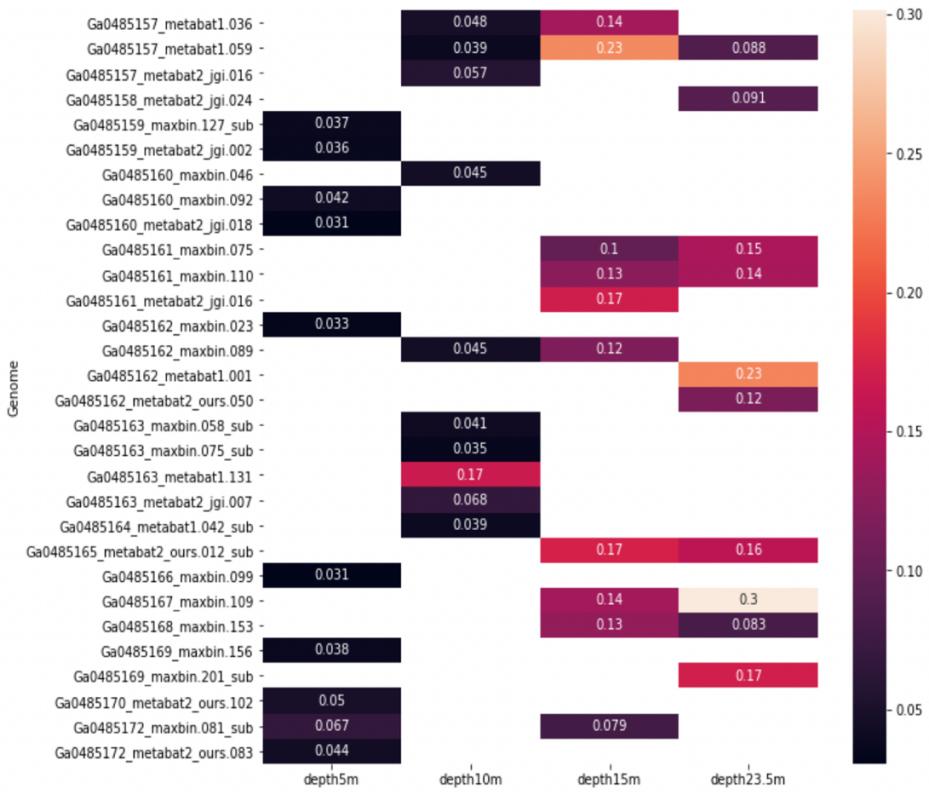
Below here is a dictionary that is the count for how many times that the genomes appear in the top10(percentage)

This time I am only focusing on the most popular(count) top ten genomes that have appeared in all the **depths**.

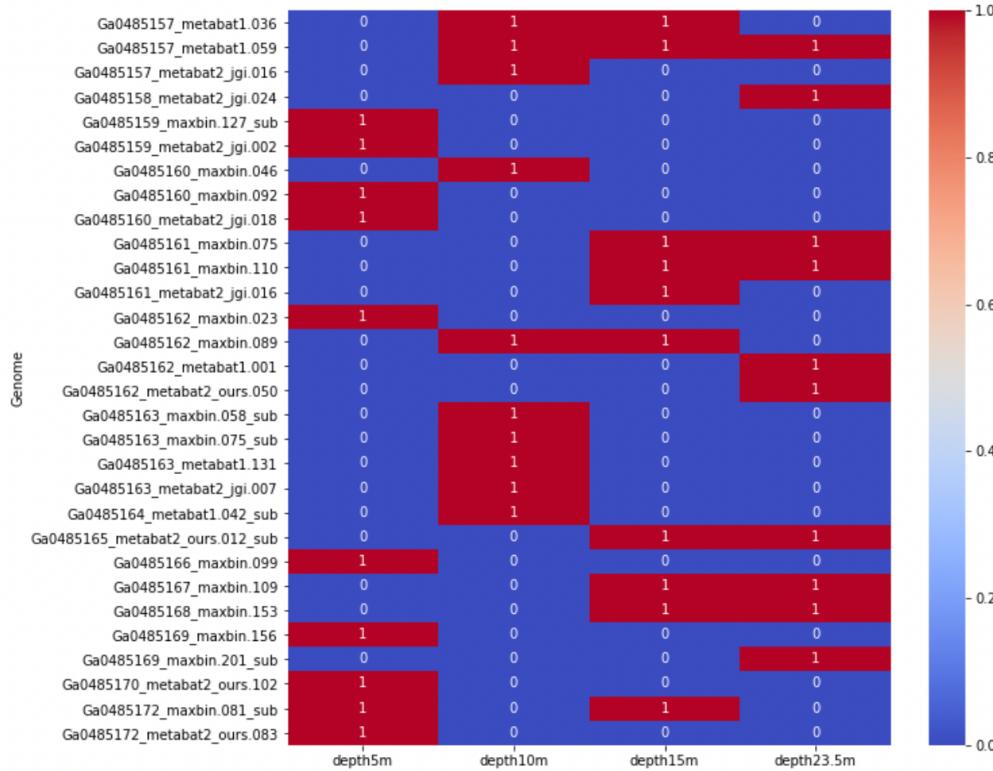
```
{'Ga0485157_metabat1.059': 3,
 'Ga0485157_metabat1.036': 2,
 'Ga0485162_maxbin.089': 2,
 'Ga0485165_metabat2_ours.012_sub': 2,
 'Ga0485167_maxbin.109': 2,
 'Ga0485161_maxbin.110': 2,
 'Ga0485168_maxbin.153': 2,
 'Ga0485161_maxbin.075': 2,
 'Ga0485172_maxbin.081_sub': 1,
 'Ga0485172_metabat2_ours.083': 1,
 'Ga0485170_metabat2_ours.102': 1,
 'Ga0485160_maxbin.092': 1,
 'Ga0485169_maxbin.156': 1,
 'Ga0485162_maxbin.023': 1,
 'Ga0485159_metabat2_jgi.002': 1,
 'Ga0485159_maxbin.127_sub': 1,
 'Ga0485166_maxbin.099': 1,
 'Ga0485172_maxbin.051_sub': 1,
 'Ga0485163_metabat1.131': 1,
 'Ga0485163_metabat2_jgi.007': 1,
 'Ga0485157_metabat2_jgi.016': 1,
 'Ga0485160_maxbin.046': 1,
 'Ga0485163_maxbin.058_sub': 1,
 'Ga0485164_metabat1.042_sub': 1,
 'Ga0485163_maxbin.075_sub': 1,
 'Ga0485161_metabat2_jgi.016': 1,
 'Ga0485167_maxbin.132': 1,
 'Ga0485162_metabat1.001': 1,
 'Ga0485169_maxbin.201_sub': 1,
 'Ga0485162_metabat2_ours.050': 1,
 'Ga0485158_metabat2_jgi.024': 1}
```

Ga0485157_metabat1.059 exists three times across all the dep

ths(3/4).



plotting heatmap for all of the genome vs. depth. However, we are only focusing on the top 10 this time.... this would actually be the all of the top10 genomes that appear in all of the depths. Generally speaking, shallower depths have less percentage, which means less count of the genome.

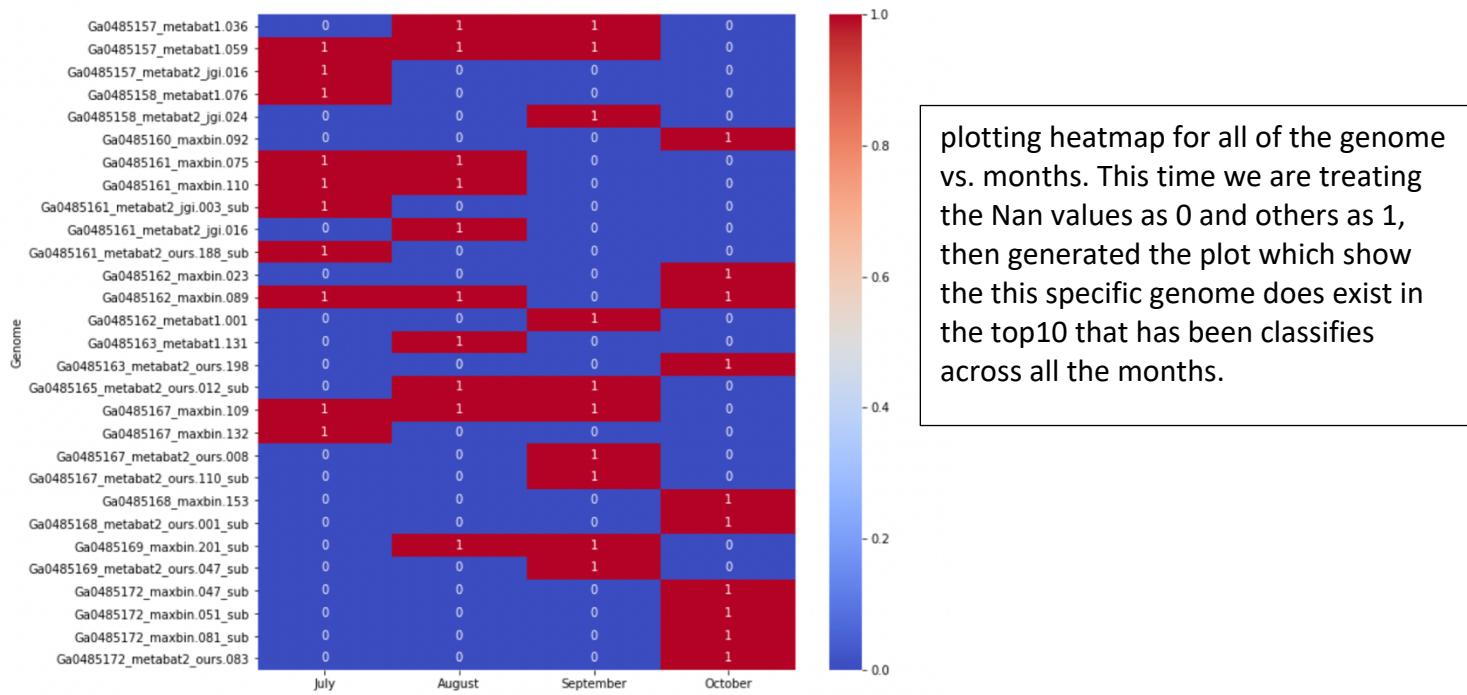
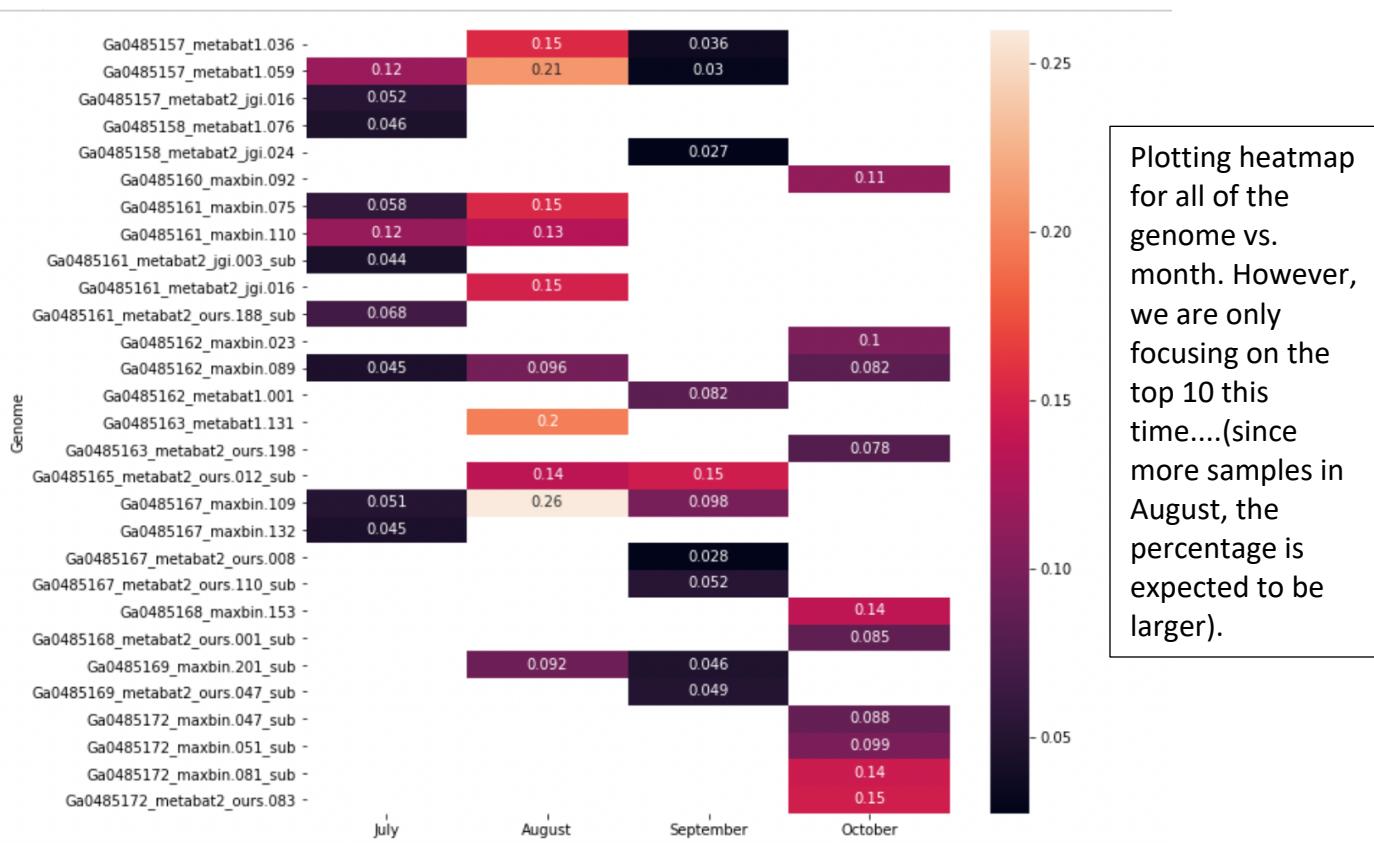


plotting heatmap for all of the genome vs. depth. This time we are treating the Nan values as 0 and others as 1, then generated the plot which show the this specific genome does exist in the top10 that has been classified across all the depths.

This would be a dictionary that represent the most popular(count) top ten genomes that have appeared in all the **months**.

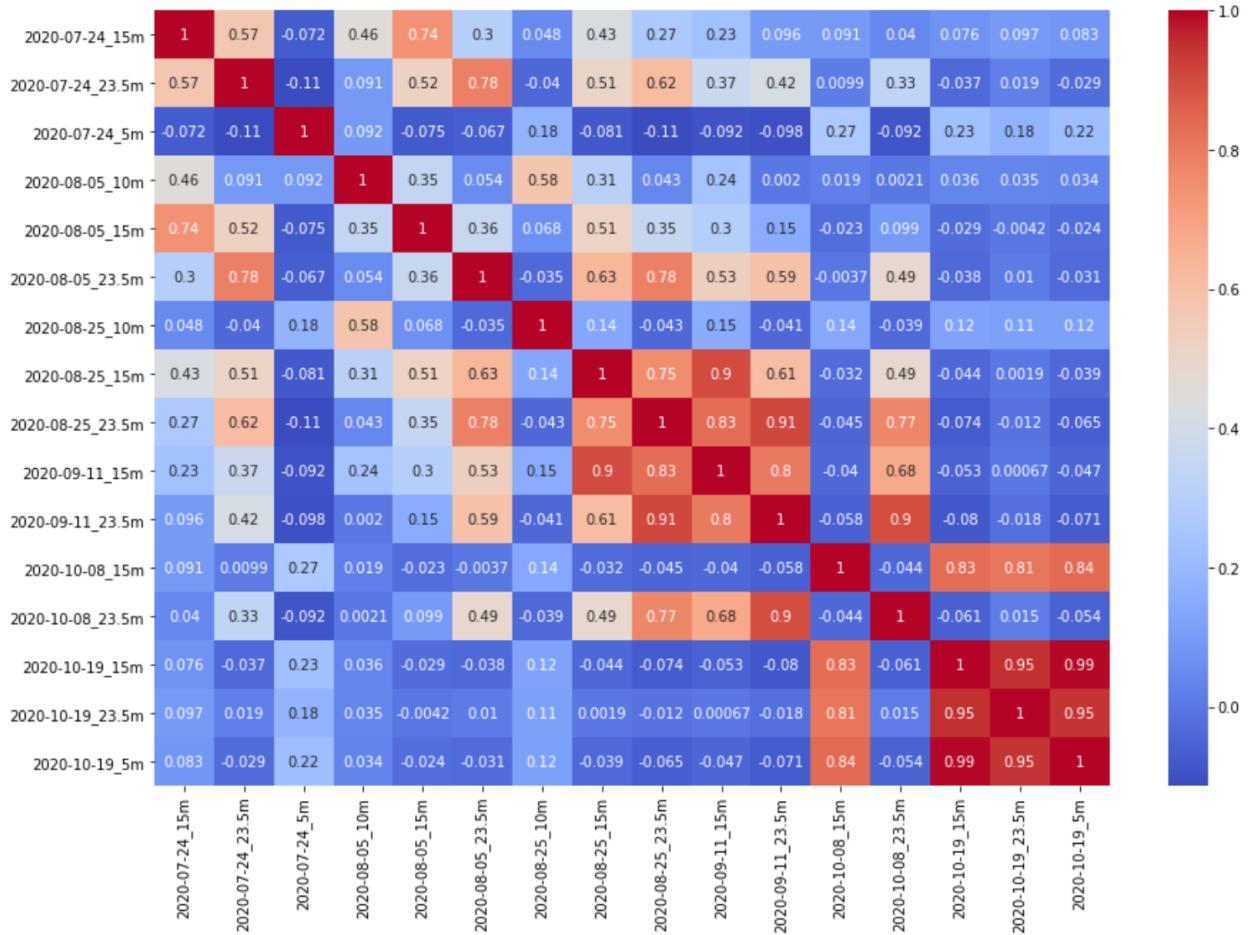
```
{'Ga0485157_metabat1.059': 4,  
 'Ga0485167_maxbin.109': 4,  
 'Ga0485163_metabat1.131': 3,  
 'Ga0485157_metabat1.036': 3,  
 'Ga0485165_metabat2_ours.012_sub': 3,  
 'Ga0485161_maxbin.110': 2,  
 'Ga0485161_maxbin.075': 2,  
 'Ga0485162_maxbin.089': 2,  
 'Ga0485162_metabat1.001': 2,  
 'Ga0485167_metabat2_ours.110_sub': 2,  
 'Ga0485169_metabat2_ours.047_sub': 2,  
 'Ga0485169_maxbin.201_sub': 2,  
 'Ga0485167_metabat2_ours.008': 2,  
 'Ga0485161_metabat2_ours.188_sub': 1,  
 'Ga0485157_metabat2_jgi.016': 1,  
 'Ga0485161_metabat2_jgi.003_sub': 1,  
 'Ga0485158_metabat1.076': 1,  
 'Ga0485167_maxbin.132': 1,  
 'Ga0485161_metabat2_jgi.016': 1,  
 'Ga0485164_metabat2_ours.159': 1}
```

'Ga0485157_metabat1.059', 'Ga0485167_maxbin.109' exist 4 out of 4 times across all the months. 'Ga0485157_metabat1.036', 'Ga0485165_metabat2_ours.012_sub', 'Ga0485169_maxbin.201_sub' exist 3 out of 4 times across all the months.



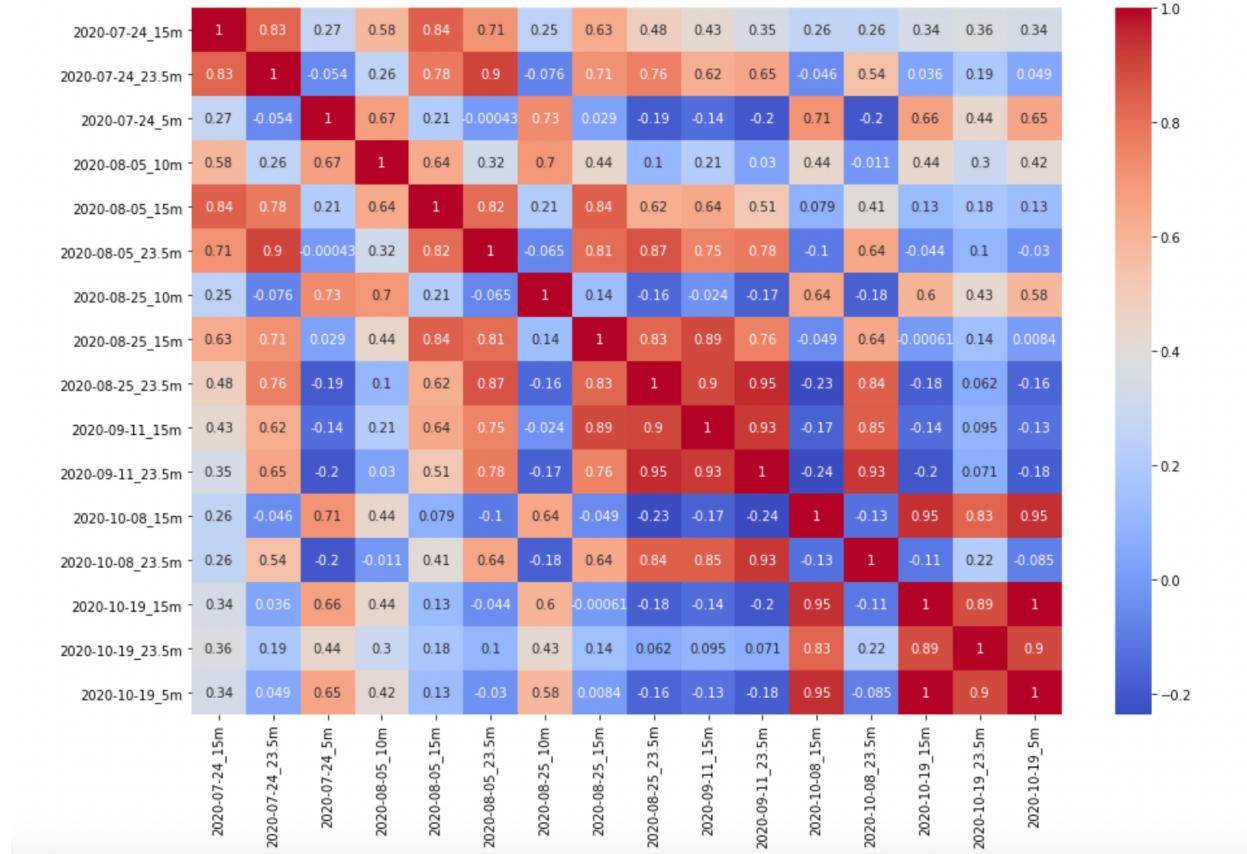
plotting heatmap for all of the genome vs. months. This time we are treating the Nan values as 0 and others as 1, then generated the plot which show the this specific genome does exist in the top10 that has been classified across all the months.

Correlation matrix of the count dataset
 ("coverm_431_MAGS_metagenomes_reads_count.csv")
 #measured by Pearson's correlation



We concluded that sample that are measured at 23.5m on 2020-10-19 is highly correlated with sample that are measured at 15m on 2020-10-19. Sample at 5m on 2020-10-19 is highly correlated with sample that are measured at 15m on 2020-10-19. Sample at 5m on 2020-10-19 is highly correlated with Sample at 23.5m on 2020-10-19.

#measured by spearman's correlation



Since this heatmap is generated by using Spearman correlation which measures the monotonic relationship between two variable. we can conclude that the coefficient tend to increase(or perhaps decrease) shows show that the two variables are more related. The last samples that have been mention in the conclusion of Pearson's correlation are still very correlated. Besides that, Sample at 15m on 2020-10-08 is highly correlated with all of the samples that have been measured on 2020-10-19.