Getting Combined Files

So all the parquet files were split into batches of 603

- We need to make a combined fle to get the embeddings.
- The code file can be found in `readings_clean.ipynb`
- The code used was using pyarrow to avoid memory overload.
- The results were obtained as follows: (the quality column is not exactly needed since we already filtered phred for being > 30, check document 5 for the same.)

Forward reads:

```
(28354061, 2)
              id                                                sequence
0  SRR5177930.19  GCCATAGCCATTGCCATTGCCACTTGGGGCAAAGCCATTTCCCCCA...
1  SRR5177930.28  ATGTGGGATTTTGATATTTATGGTACTGTGTCTATGTGCTGATTGT...
2  SRR5177930.38  ACCTTTATAGGTGGGGATTAGGAGTCCCTTCTGGGCTGGGTGTGGT...
3  SRR5177930.39  GCACAGGTAGCCAGACTCTGATCATGGCTCTGAGGAGGAGCCCTGG...
4  SRR5177930.58  ATCCTGGGTTTTAATGCTAGGGTGGAAAGGTATTTCTGAAGCCTTG...
```

Backward reads:

```
(9081535, 2)
              id                                                sequence
0   SRR5177930.6  ATAGAGCCCTAAACTCCCTTGCTGGTTTCTCTGAGCTCTCTGATTT...
1   SRR5177930.9  ATTTTAATCACATACTATCTCCACCTTCTGTAGAAGCTTAGGCCCC...
2  SRR5177930.11  ATCTACCTACCTTATAGCAAGACTTCTGGGGTTTCCAGCATCACCA...
3  SRR5177930.16  ATTTCATTGTGTCCTCGTTTCTAACCCCCAGTACTTCCATTTTCCC...
4  SRR5177930.20  ATAGATTTTTATGTGGATAGTATAATAATTCAATATCAAATAAAAA...
```

The total number of reads have dropped *significantly*.