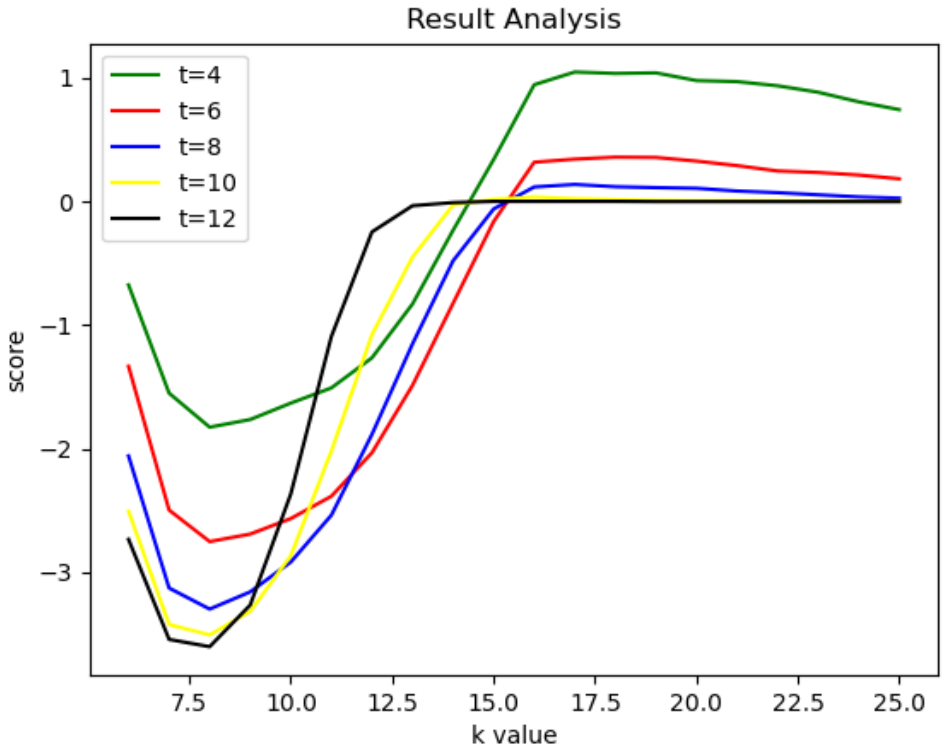
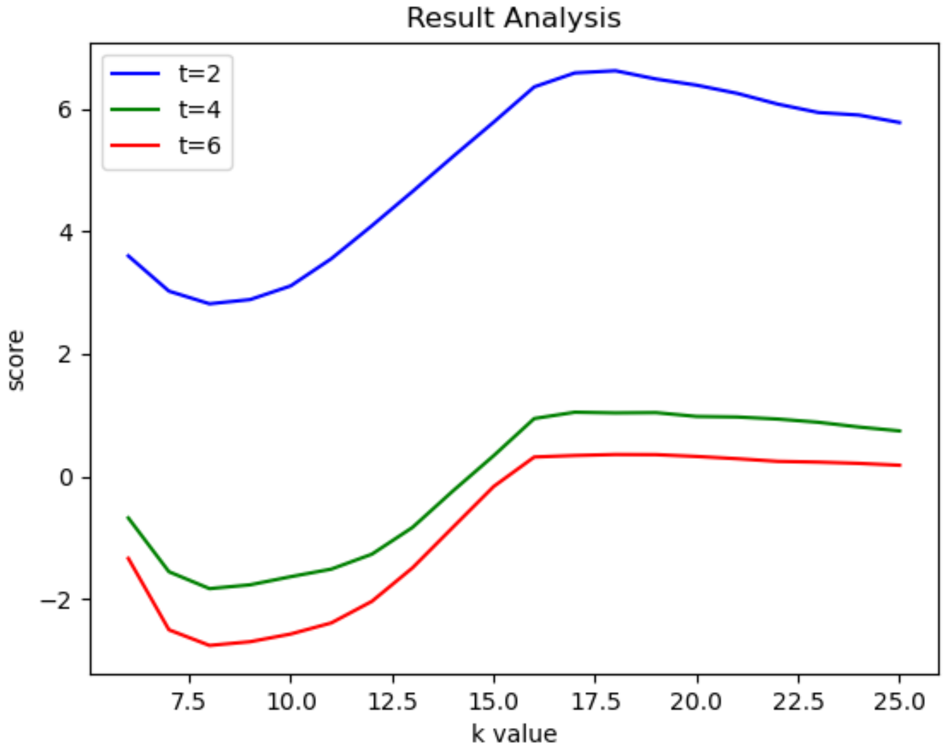
# Q0. the values of Nmin, Sm, max(Sk), and max(St)

- $N_{min}$ = 1269
  - $1 - e^{-a} = cover$, where $a = \frac{NL}{G}, L = 50, G = 9181$
  - $0.999 = 1 - e^{-\frac{NL}{G}}$
  - $6.907755279 = \frac{N \times 50}{9181}$
  - take the integer, $N_{min} = 1269$
- $S_m = 49.02758077226162$
- As I mentioned in README, I inplemented multiple algorithms for comparison
  - stack_correction:
    - $max(S_k) = 49.65692$, with t = 4, k = 17
    - $max(S_t) = 49.997$, with t = 2, k = 18
  - simple_correction
    - $max(S_k) = 49.57324$, with t = 4, k = 17
    - $max(S_t) = 49.985$, with t = 2, k = 17
  - merge_correction:
    - test in process

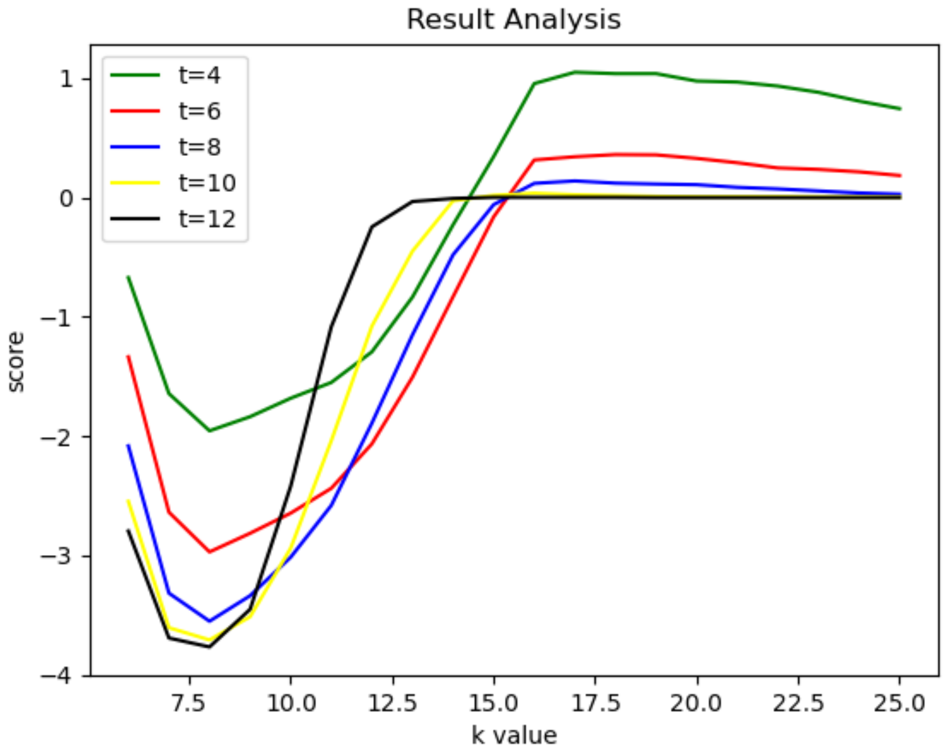
# Q1. the two plots described above

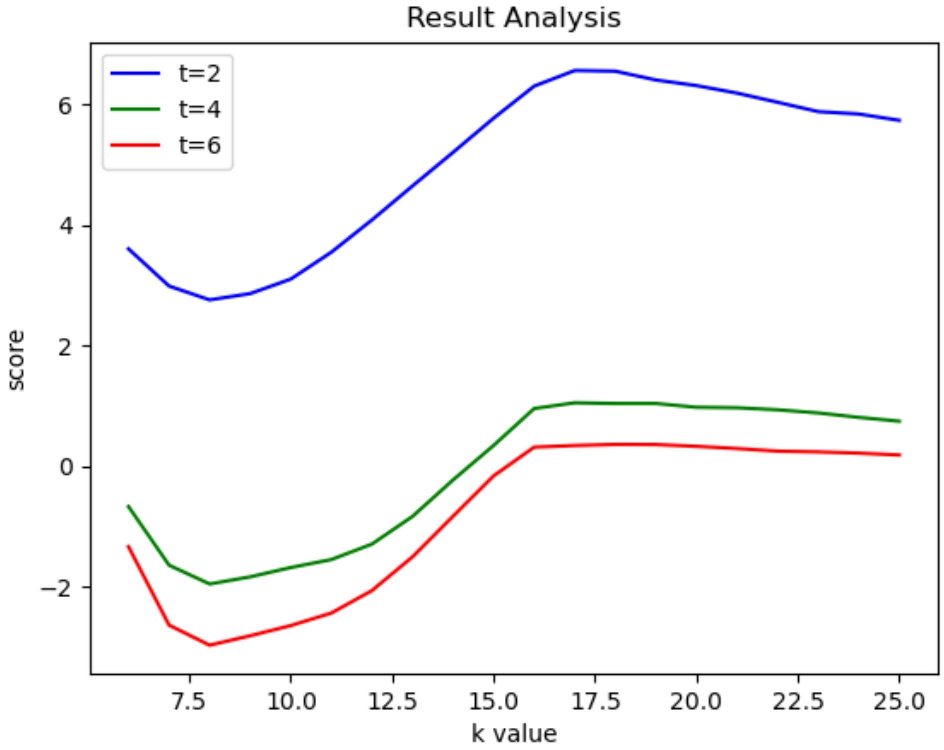
## For stack correction



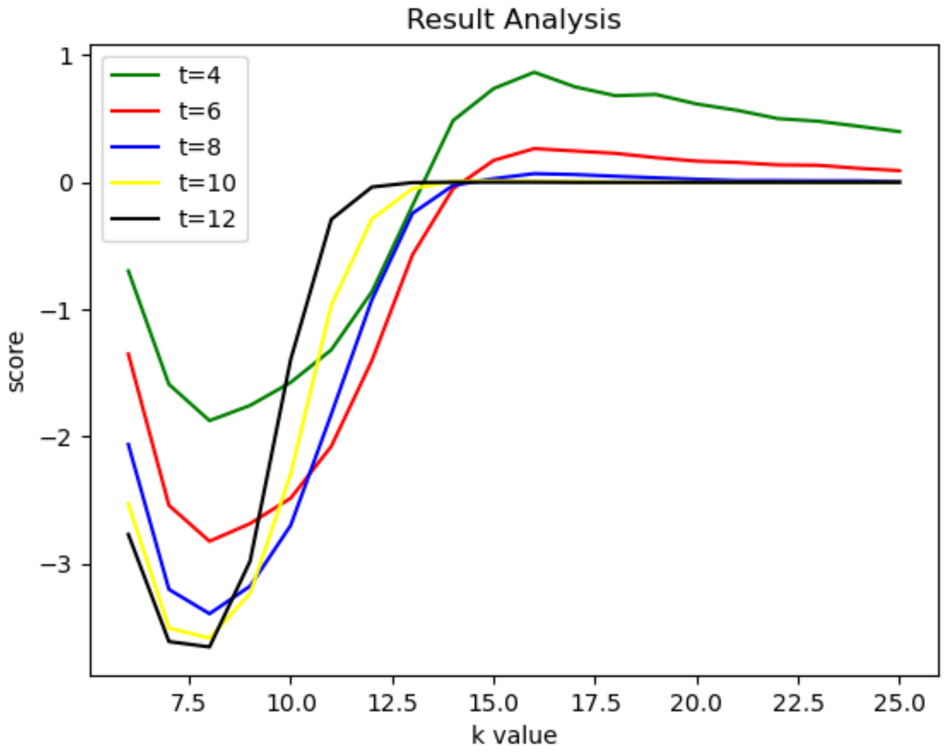


## For Simple Correction



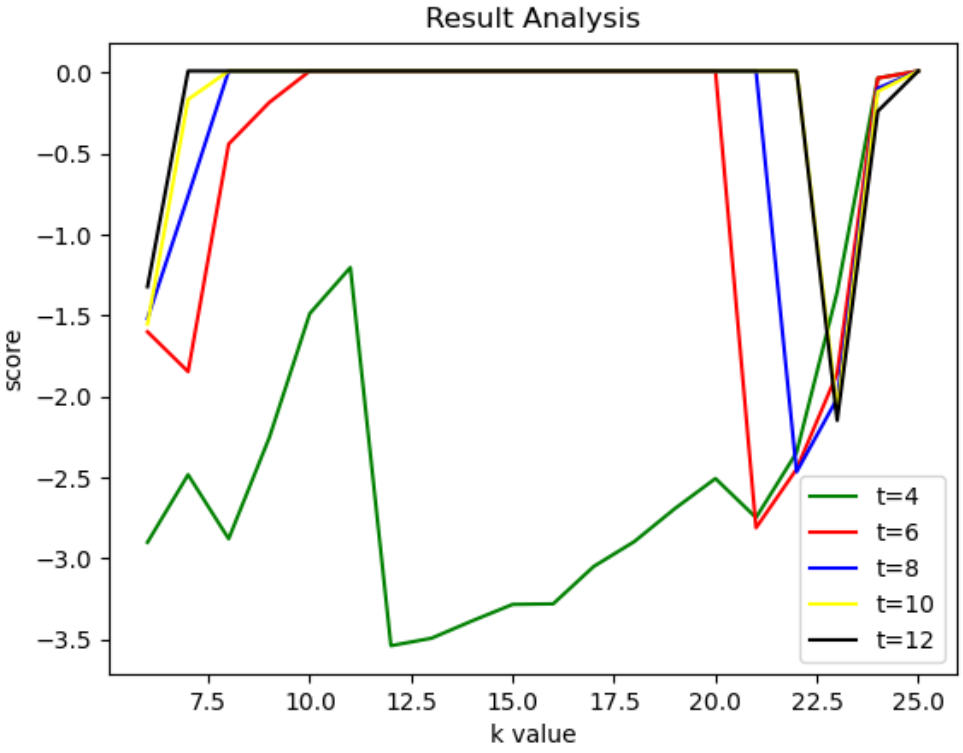


## For Naive Correction (only experiment on part1)



## For Merge Correction

time limitation only experiment on $N_{min} = 100$, you can notice lots of 0 value because can not find frequent k-mer with low sample rate.



## Q3.1 What general trends do you observe in varying k and t? Justify your claims by referring to both of the plots generated above. Give an explanation (either mathematical or intuitive) for why modulating each variable produced these results.

**For constant k**

- **for small k (range 6 to 10)**,
    - **for the score**: when k increases, the score decreases, but the decreasing speed is becoming smaller and smaller.
    - **for the t:** In this range, smaller t will give relatively  better performance (though they are making the read worse)
- **for middle k (range 11 to 15),**
    - **for the score**: when k increases, the score increases
        - **Explaination**: this is the only range when k increases, the score increases. One possible explaination is the length of this range is meaningful for the sequence. Like the meaningful motif areas in hiv are usually from length 11 to 15. Therefore, in this range, we have the possibility to get the true frequent motif, which makes the frequent k-mers we find are the correct answer during the correction.
    - **for the t:** in this range, larger t will have more rapid increasing, and better performance
- **for large k (range 15 to 25),**
    - **for the score**: when k increases, the score converges, with slightly decreasing
    - **for the t:** in this range, smaller t will give better score.

**For constant t**

- all the t-curves have the tendency: decreasing in small range of k, increasing in middle range of k, converging in large range of k
- generally, in the range of [4,12], smaller t will give better performance.

**Explaination**:

when k is small, since the total combo of the k-mer will be small, and we can image we ca get more repeated k-mers; on the contrary, when k is large, the repeated k-mers will be less.

- for the small k, though we have a lot of frequent k-mer, these repeated k-mers might

contain noise, since the repetition might not be meaningful and only because the combo size is small. That might be the reason why when k is in small range, increasing frequency threshold is not helpful to select the potential correct k-mer but including noise to the data, which makes correction be less efficiency.
- for middle k,  we can see the performance is strictly increasing. This might because when k is increasing, the noise frequent k-mer is gradually excluded and increase the performance.
- for large k, after k reaches a meaningful value, like close to the motif length in the original sequence, at this time,  increasing k might not be useful. Even the performance might decrease becasue of it is harder to generate frequent k-mers for correction. For example, we can notice for t in [6,12] and large k, the score is almost zero, that is a significant flag for there might not have any correction operation happens since the infrequent k-mer cannot find any close frequent k-mer and directly ends the program.

And for t, when t is small, it means the there will be more k-mer be marked as frequent; on the contrary, large t will result in less frequent k-mer.

- small t usually performs better in small k and large
  - for large k, it is quite hard to find large quantity of k-mers, therefore, if t is large, it will eliminate some meaningful but relatively low frequecy k-mers
  - for small k, a lot of k-mers are frequent but without biological meaning. Under this situation, if we want to raise k threshold to get the k-mer, we might end up with noise (noise amount might be larger than meaningful k-mer). Therefore,  with small k-mer, we keep them all, but let the distance threshold to pick up the right k-mer.
- large k usually performs better in middle range
  - as we mentioned, we believe the middle range k might be close to the motif length in the hiv sequence, therefore, we have the k-mers with little noise, this is the time we want t to eliminate the infrequent one which has lower possibility to be the motif.

# Q3.2 How is the choice of t dependent on k? Are there any values of k for which varying t does not produce a significant difference in correction efficiency? Why might this be the case?

**How is the choice of t dependent on k?**

- for small k, like [6,10], we should choose small k to obtain better performance
- for middle k, like [11, 15], we should choose large k to obtain better performance
- for large k, like [16, 25], we should choose small k to obtain better performance
- however, with my implementation,  the correction starts to work (larger than 0) in large k, so we can choose large k in practice to obtain better correction efficiency.

**Are there any values of k for which varying t does not produce a significant difference in correction efficiency? Why?**

- Yes, there are. When k is in range 14 to 16, varying t will not have a significant difference in correction efficiency. Also notice, apart from t=4, for t in range [6, 12], the, for k in range [17, 25], t also does not have a major difference.
- My explaination is
  - for the k in range [14, 16], it might because current k value is close to the meaningful motif length, therefore, with certain elimination brings by t, we can have the relatively correct correction.
  - for the k in range [17, 25], with the large k-mer length, it is hard to find the frequent, or the frequency will not be high. Therefore, above certain threshold like 4, for t in range [6, 12], it exceeds the highest frequency, and makes infrequent cannot find close frequent as we mentioned in Q3.1.

# Q3.3 What feature(s) of the data are the parameters k and t dependent on? Give an explanation (either mathematical or intuitive). How could you modify your code to take advantage of this dependence and improve your correction procedure?

**What feature(s) of the data are the parameters k and t dependent on?**

- **Repeat**: there are frequently repeated regions in the sequence that have biological meanings, like the concept motif (not sure whether this is correct, let me know if I understand the bio term incorrectly)
  - **Repeat Amount**: for certain repeated pattern, if the amount of this pattern is large, (1) then we can have larger t to have a more confident exclusion of the infrequent k-mers; since we reduce the false positive frequent k-mer amount, then close frequent k-mer space will be narrowed down for infrequent k-mer, which reduces the complexity.
  - **Repeat Length**: repeat length will influence both k and t
    - for k: if repeat length is large, then the k can be relatively large to include these repeats; and vice versa
    - for t: when repeat length is large, then we can image the frequency for the repeats will be relatively small, therefore, we can set a relatively small t to identify the frequent k-mer; on the contrary, t can be larger.
- **Error Rate and Distribution**: here, since our error rate is low (1%) and can be seen as uniformly distributed, our correction strategy is closely related to this: notice the writer developed different strategy to avoid the overlapped situation and long continuous infrequent k-mers - specifically, one strategy is related to k, since we do not want k continuous infrequent k-mers.
  - if the error rate and distribution changed, say the error rate is high,

        - then we would like to choose larger k, since for small k, we might get a k-mer full of the mutated base, and when we try to find the close frequent k-mer to correct it, it might reject the distance threshold d and end up being not fixed
        - and we might want to lower the standard of frequent by using relatively small t, since when mutation is high, we might observe less repeats. If t is too high, we will end up with small frequent k-mer set, and it will be hard for infrequent one find the potential frequent, which might decrease the error correction efficiency.
    - if the error rate is low, the case will be opposite.

- **Coverage**
    - when the data coverage is low, we cannot cover all the sequences, and we can image the frequent k-mer number will decrease.
        - for k: At this time, we might want to use relatively small k to let the data contains all the possible combos of a small k (remember when we deal with gene assembly, if we cannot get all fraction of certain k-mers, we will decrease the k value to cover nearlly all fraction of small k-mer)
        - for t: we might also want to lower the t value
    - However, coverage is really important, what we might really want to do is sample more data and increase the coverage.

**Furture Improvement**

- **know your data:** we can first have an overall idea of our data property, like the data coverage, what would be the general motif length in this sequence, and what is our sampling method, what error rate is this kind of sampling method. After knowing this information, we can try to choose the suitable k and t range for our data.
- **how to identify frequent**:
    - with all of the analysis and assumption the writer proposed, we can notice one key step is finding the frequent and infrequent k-mers. Remember we analyze for the middle range of k, no matter what t value is, we can observe the performance is increasing. And the writer's assumption is "this range might be close to the motif length in the hiv sequence", and if this is true truth, the key step is identifying the length range of meaningful repeat in the given data.
    - this might lead us to a question, that is, if the meaningful repeats have a length range, i.e. not a constant, how can we use only one constant k and tend to identify all the repeats? That's to say, our frequent identification process can be improved by editing k to be dynamic.
    - One naive proposal can be similar as what we do in the application, we try certain range of k, see how the result goes and choose a range of k that makes sense. This is like a grid search algorithm to choose the best model, the problem will be the large time complexity.
    - I recently read a classic paper in BIM :
        - Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. Nature biotechnology, 33(8):831–838, 2015.

- this famous deepbind model proposed in 2015 might have something we can learn
- The paper utilizes the first convolutional layer of its prediction network as motif-finders, i.e., using CNN to find the biological meaningful repeat pattern in the sequence.
- we can also try this, instead of defining certain k to count frequency, we can try to train the filter to identify the frequent repeats.
  - we can simply use the k range/value developed by the network as our best-guess k
  - or we can design more advanced algorithm, like find the distance between the infrequent and the identified motif and try to correct the infrequent.