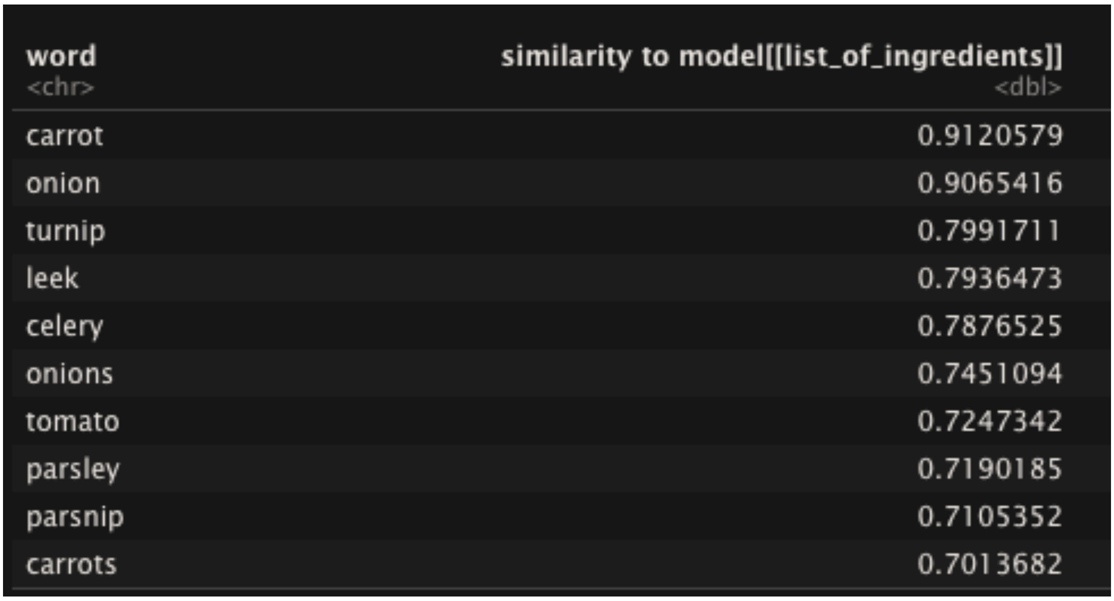
H3P2

Guanzhi Wang
4/12/2020"
Collaborate with Shaoyu Feng

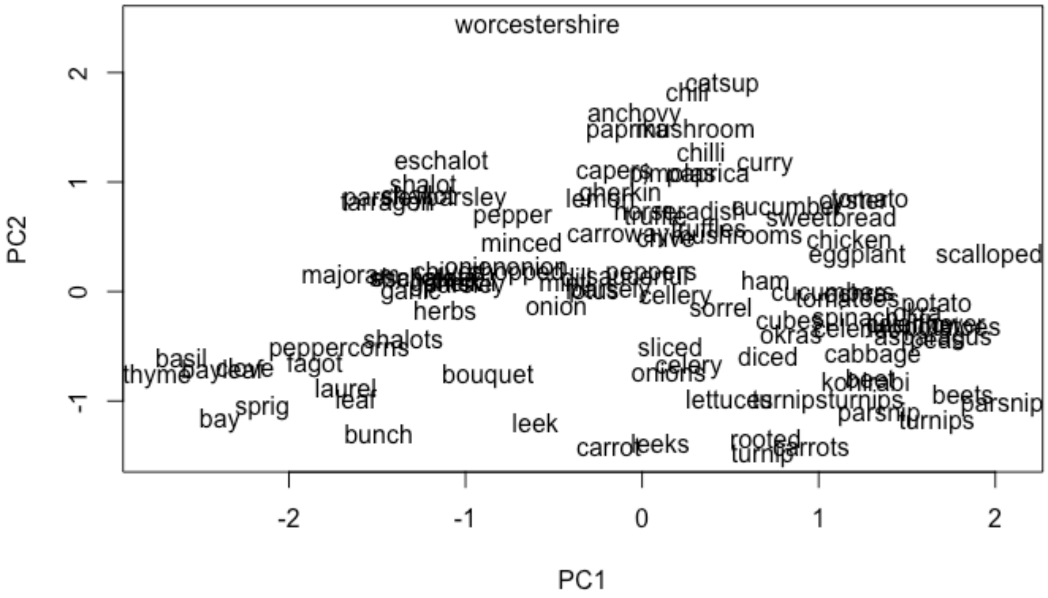
Question 3

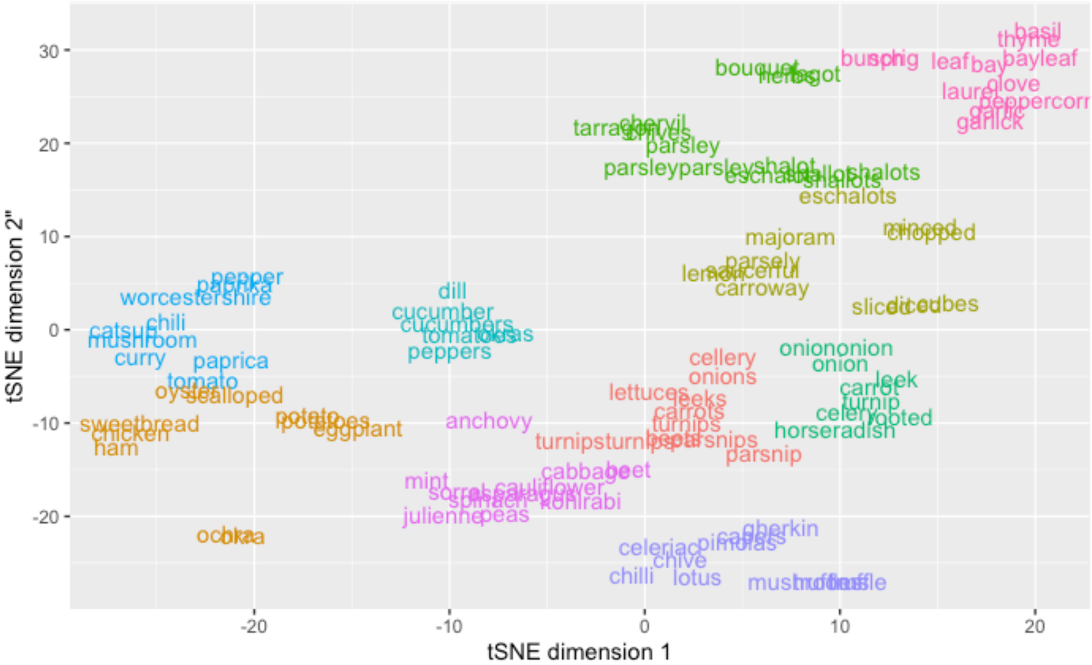
1. Remove the stop word, and to the stemming and lemmatization.
2. I picked tomato, onion and carrot.



word	similarity to model[[list_of_ingredients]]
<chr>	<dbl>
carrot	0.9120579
onion	0.9065416
turnip	0.7991711
leek	0.7936473
celery	0.7876525
onions	0.7451094
tomato	0.7247342
parsley	0.7190185
parsnip	0.7105352
carrots	0.7013682

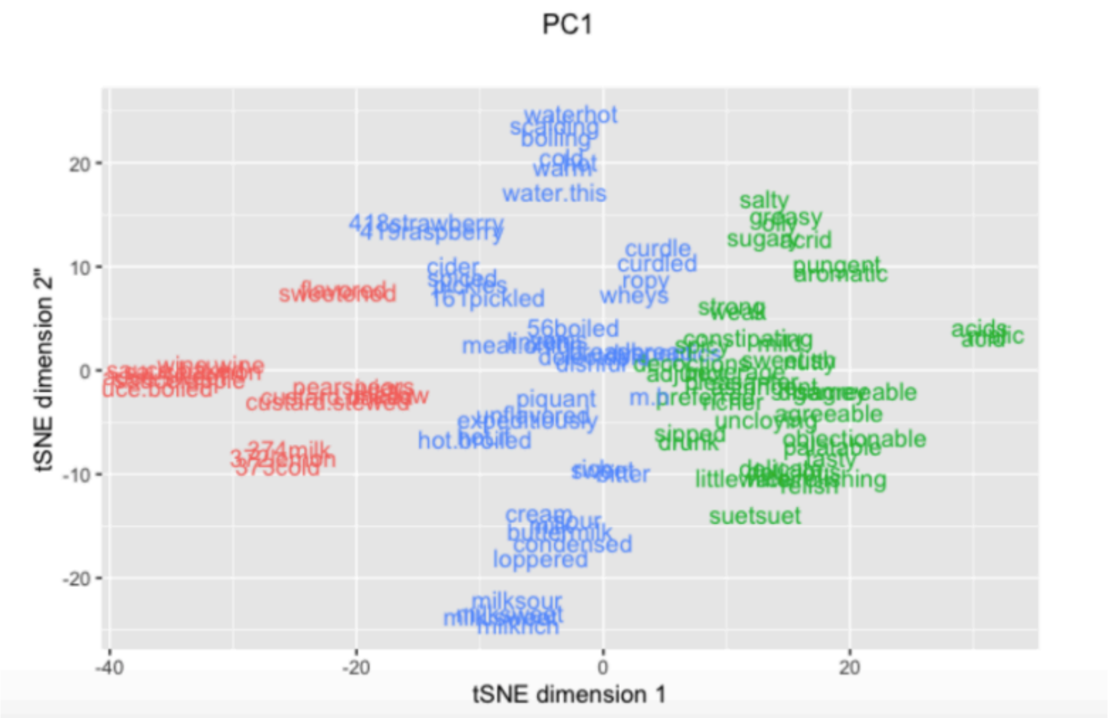
3.





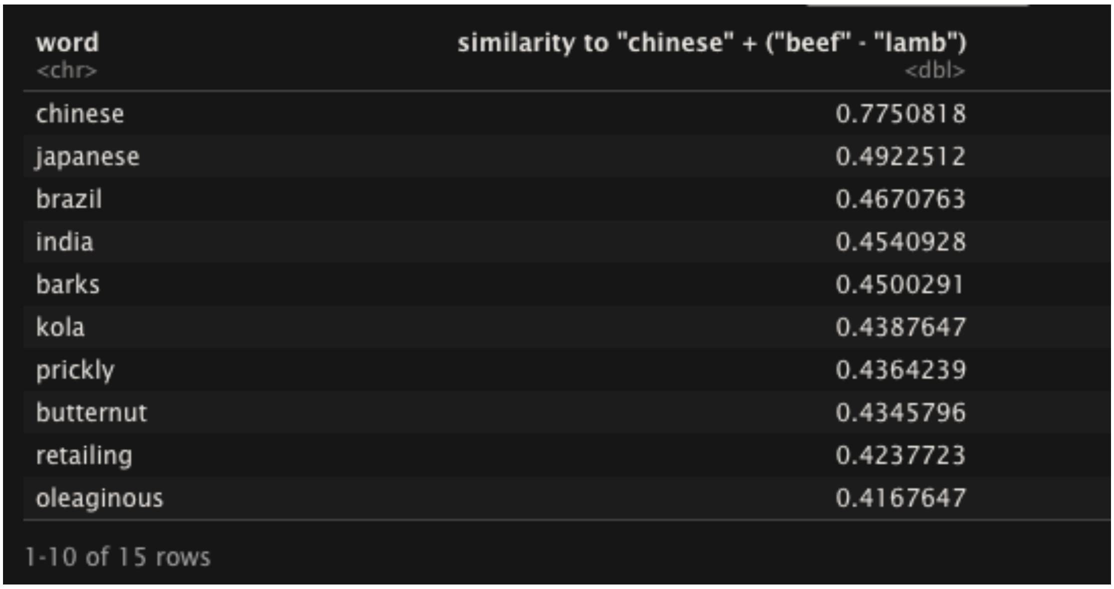
4.

When pick hot, spicy and sour, we have the following pic



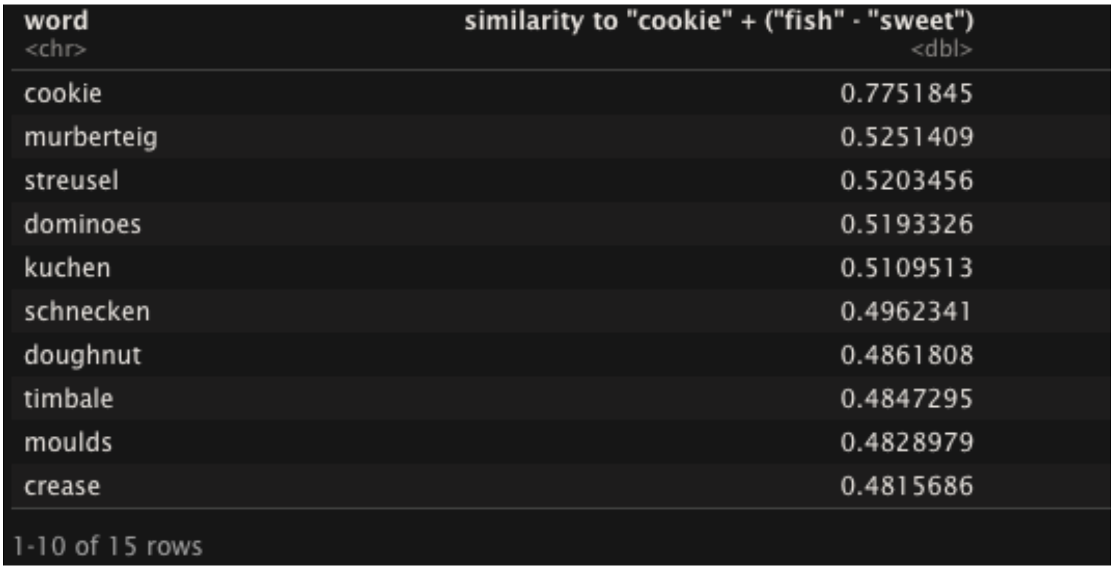
This looks reasonable.

5.



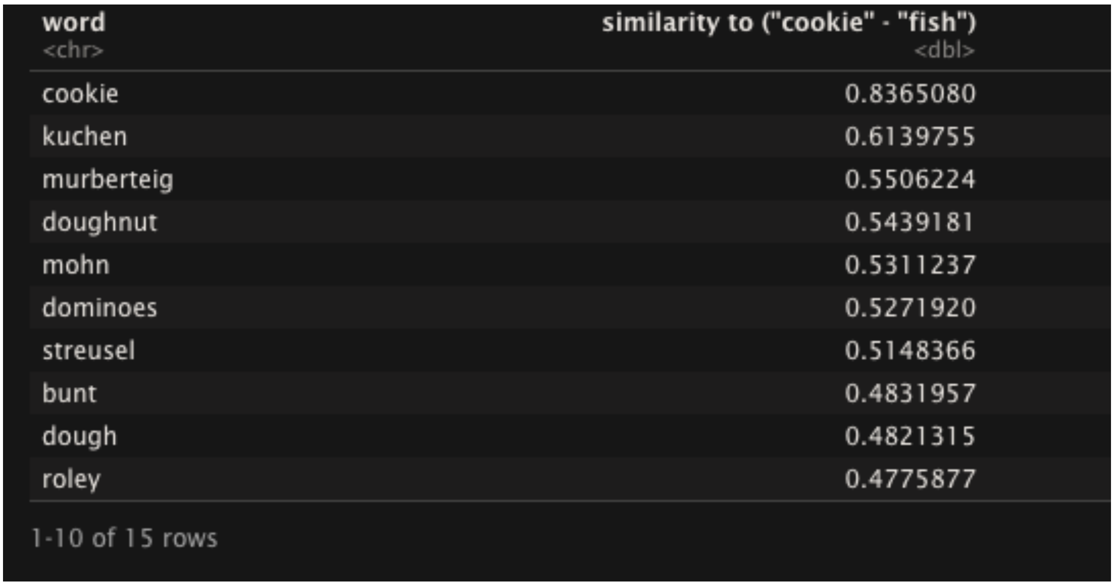
word <chr>	similarity to "chinese" + ("beef" - "lamb") <dbl>
chinese	0.7750818
japanese	0.4922512
brazil	0.4670763
india	0.4540928
barks	0.4500291
kola	0.4387647
prickly	0.4364239
butternut	0.4345796
retailing	0.4237723
oleaginous	0.4167647

1-10 of 15 rows



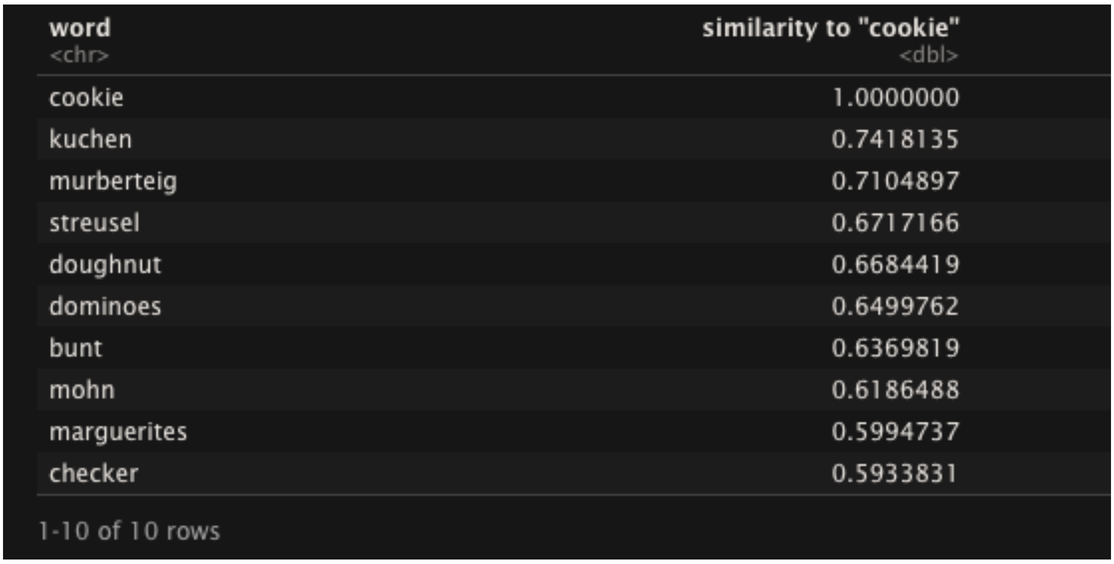
word <chr>	similarity to "cookie" + ("fish" - "sweet") <dbl>
cookie	0.7751845
murberteig	0.5251409
streusel	0.5203456
dominoes	0.5193326
kuchen	0.5109513
schnecken	0.4962341
doughnut	0.4861808
timbale	0.4847295
moulds	0.4828979
crease	0.4815686

1-10 of 15 rows



word <chr>	similarity to ("cookie" - "fish") <dbl>
cookie	0.8365080
kuchen	0.6139755
murberteig	0.5506224
doughnut	0.5439181
mohn	0.5311237
dominoes	0.5271920
streusel	0.5148366
bunt	0.4831957
dough	0.4821315
roley	0.4775877

1-10 of 15 rows



word <chr>	similarity to "cookie" <dbl>
cookie	1.0000000
kuchen	0.7418135
murberteig	0.7104897
streusel	0.6717166
doughnut	0.6684419
dominoes	0.6499762
bunt	0.6369819
mohn	0.6186488
marguerites	0.5994737
checker	0.5933831

1-10 of 10 rows

6.

It is interesting wo see when we have Chinese+(bee-lamb), we have the top 4 words as country names, and the number 5 is kola, which is note even related with neither beef, lamb nor Chinese.

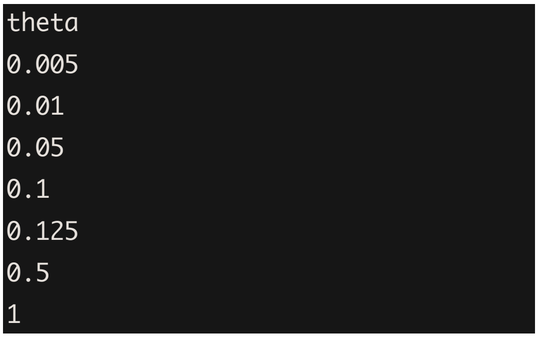
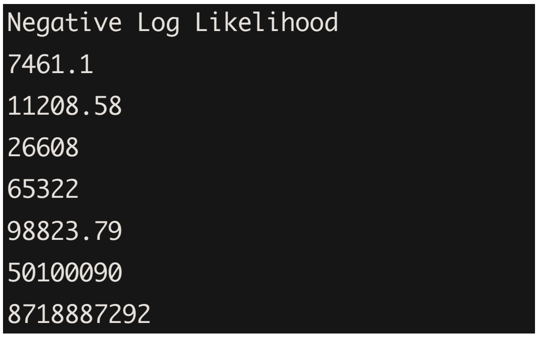
For cookie" + ("fish"- "sweet"), we see doughnut, which is reasonable. However, it’ s hard to understand why bunt and checker will appear in this list.

7.
The most word we got is like “of the” ,” in a” , etc. They are generally the stop word.

Question 4

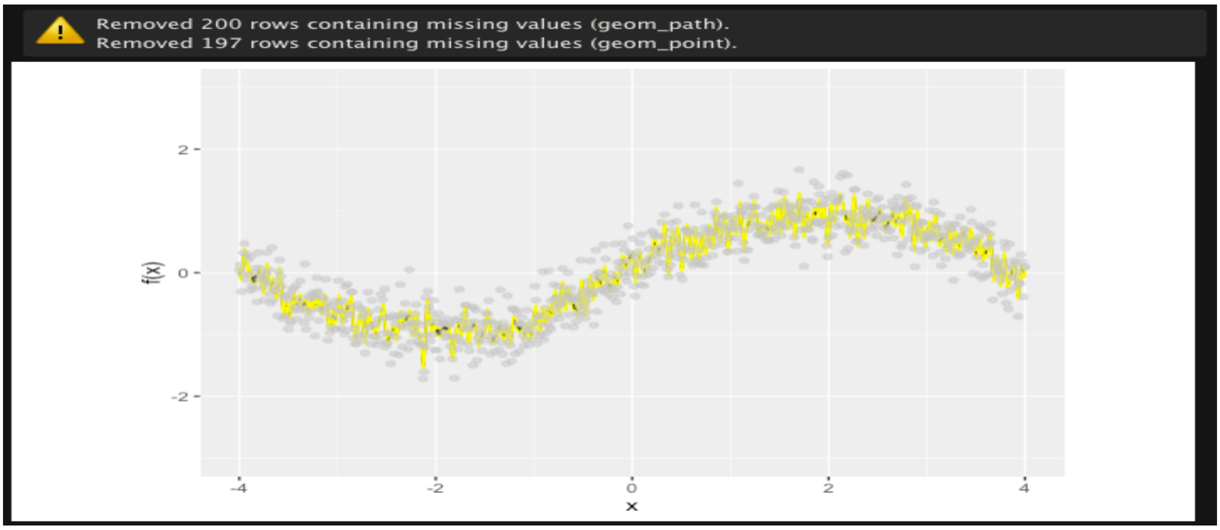
Part 1

2.

	
theta	Negative Log Likelihood
0.005	7461.1
0.01	11208.58
0.05	26608
0.1	65322
0.125	98823.79
0.5	50100090
1	8718887292

From the chart, we have the best theta as 0.005.

3.

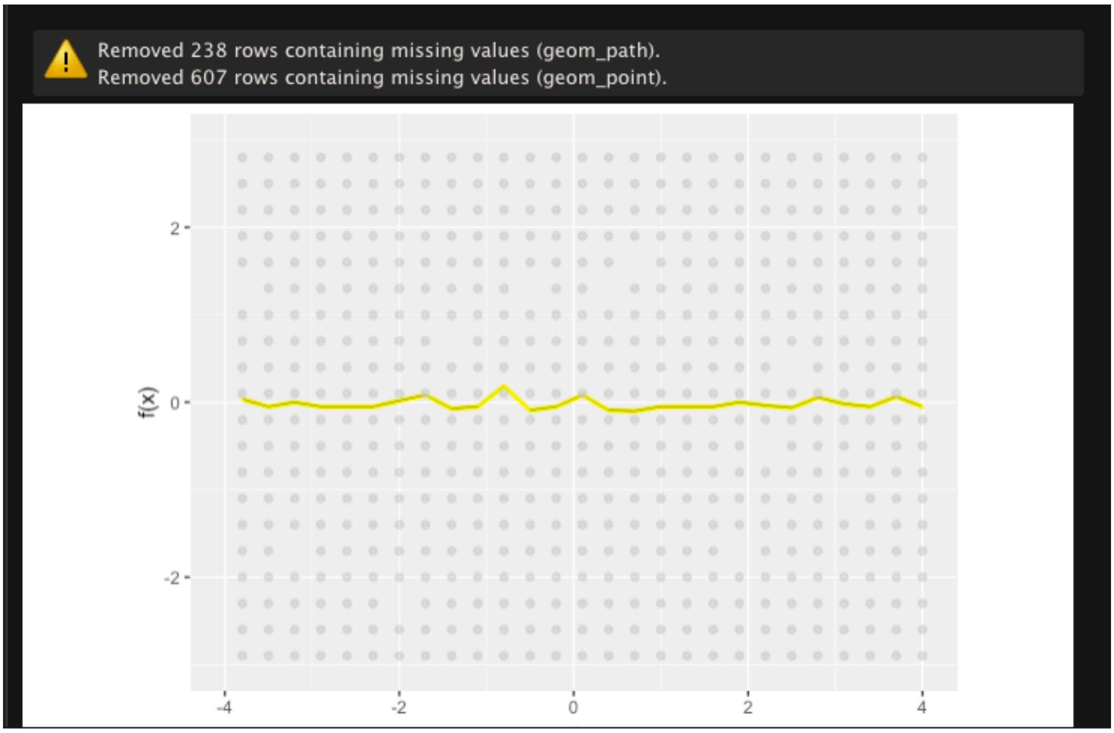


```
Removed 200 rows containing missing values (geom_path).
Removed 197 rows containing missing values (geom_point).
```

Part 2

I choose the question 1.

Refer to the R code, we have the best theta 0.01. And the plot is



```
Removed 238 rows containing missing values (geom_path).
Removed 607 rows containing missing values (geom_point).
```