결측자료분석 중간 프로젝트 발표

2020020336 박수희
2020021202 이혜원
2020021206 최홍석

1. 사용데이터

사용 데이터 개요

wine quality dataset
(from UCI machine learning repository)

- red wine 데이터와 white wine 데이터로 구성
- red wine quality data: 1599 rows, 13 variables
- white wine quality data: 4898 rows, 13 variables

→

wine_total dataset 생성(red wine + white wine)

- 6497 rows, 13 variables

변수 정보

Fixed acidity	volatile acidity	Citric.acid	Residual.sugar	chlorides	Free.sulfur.dioxide	Total.sulfulr.dioxide
float	float	float	float	float	integer	integer

density	pH	sulphates	alcohol	quality	kinds
float	float	float	float	categorical	categorical

2. 결측자료 발생 시나리오

(1) 결측 발생시킬 변수 선택

```
Call:
lm(formula = alcohol ~ fixed.acidity + volatile.acidity + citric.acid +
    residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
    pH + sulphates + density + quality + kinds, data = wine_total)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4228 -0.2944 -0.0319  0.2555 15.1398

Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)
(Intercept)           6.495e+02  5.282e+00  122.968  < 2e-16 ***
fixed.acidity         5.221e-01  8.551e-03   61.053  < 2e-16 ***
volatile.acidity      7.886e-01  5.582e-02   14.128  < 2e-16 ***
citric.acid           5.263e-01  5.375e-02    9.790  < 2e-16 ***
residual.sugar        2.285e-01  2.910e-03   78.534  < 2e-16 ***
chlorides            -9.119e-01  2.270e-01   -4.018 5.94e-05 ***
free.sulfur.dioxide  -3.394e-03  5.204e-04   -6.522 7.48e-11 ***
total.sulfur.dioxide -1.001e-04  2.202e-04   -0.455    0.649
pH                    2.607e+00  5.248e-02   49.681  < 2e-16 ***
sulphates             9.958e-01  5.065e-02   19.659  < 2e-16 ***
density              -6.564e+02  5.397e+00 -121.636  < 2e-16 ***
quality               1.027e-01  8.338e-03   12.320  < 2e-16 ***
kindswhite           -1.146e+00  3.595e-02  -31.865  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4979 on 6484 degrees of freedom
Multiple R-squared:  0.826,     Adjusted R-squared:  0.8257
F-statistic:  2566 on 12 and 6484 DF,  p-value: < 2.2e-16
```

모든 변수들에 대한 적합 결과

```
Call:
lm(formula = alcohol ~ fixed.acidity + residual.sugar + +pH +
    density + quality + kinds, data = wine_total)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5279 -0.3023 -0.0361  0.2582 16.1122

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.450e+02  5.038e+00  128.03   <2e-16 ***
fixed.acidity   5.534e-01  8.126e-03   68.11   <2e-16 ***
residual.sugar  2.235e-01  2.913e-03   76.73   <2e-16 ***
pH              2.666e+00  5.358e-02   49.76   <2e-16 ***
density        -6.513e+02  5.139e+00 -126.75   <2e-16 ***
quality         1.054e-01  8.356e-03   12.61   <2e-16 ***
kindswhite     -1.428e+00  2.506e-02  -56.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5235 on 6490 degrees of freedom
Multiple R-squared:  0.8076,    Adjusted R-squared:  0.8074
F-statistic:  4539 on 6 and 6490 DF,  p-value: < 2.2e-16
```

선택된 변수들에 대한 적합 결과

Wine data의 각 변수를 반응변수로, 나머지 변수를 설명변수로 하여 Linear regression을 적용한 결과를 비교했을 때,
alcohol을 반응변수로 설정한 경우 residual standard error가 가장 작았다.

결측(missing)을 발생시킬 변수를 alcohol로 설정하고,
사용할 설명변수는 alcohol에 가장 유의한 영향을 미치는 변수 4개에,
변수의 다양성을 위해 두 개의 범주형 변수를 추가했다.

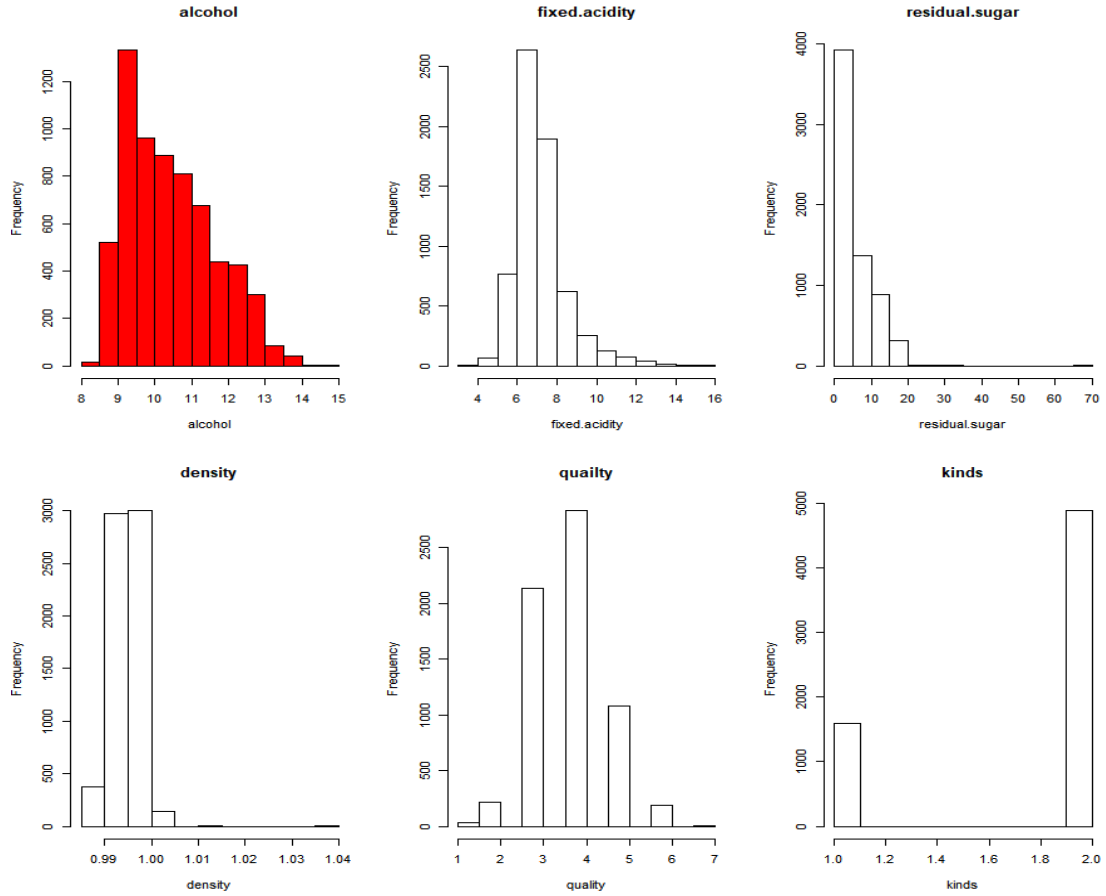
Imputation에 사용하는 변수

- 결측 발생 변수: alcohol
- 선택한 설명 변수: fixed.acidity, residual.sugar, density, pH, quality, kinds

```
 fixed.acidity    residual.sugar      density            pH           alcohol        quality     kinds
Min.   : 3.800   Min.   : 0.600   Min.   :0.9871   Min.   :2.720   Min.   : 8.00   3:  30   red  :1599
1st Qu.: 6.400   1st Qu.: 1.800   1st Qu.:0.9923   1st Qu.:3.110   1st Qu.: 9.50   4: 216   white:4898
Median : 7.000   Median : 3.000   Median :0.9949   Median :3.210   Median :10.30   5:2138
Mean   : 7.215   Mean   : 5.443   Mean   :0.9947   Mean   :3.219   Mean   :10.49   6:2836
3rd Qu.: 7.700   3rd Qu.: 8.100   3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:11.30   7:1079
Max.   :15.900   Max.   :65.800   Max.   :1.0390   Max.   :4.010   Max.   :14.90   8: 193
                                                                                  9:   5
```

2. 결측자료 발생 시나리오

결측 발생 변수로 alcohol 선택 이유



1. 변수별 히스토그램 비교한 결과,
   alcohol 변수의 분포가 비교적 골고루 퍼져 있다.

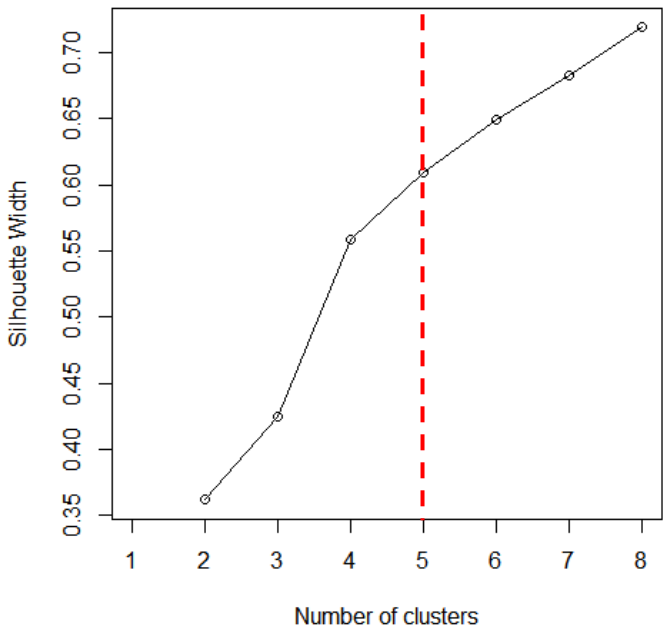
2. Linear regression 결과에서 높은 R-squared 값을 보인다.

2. 결측자료 발생 시나리오

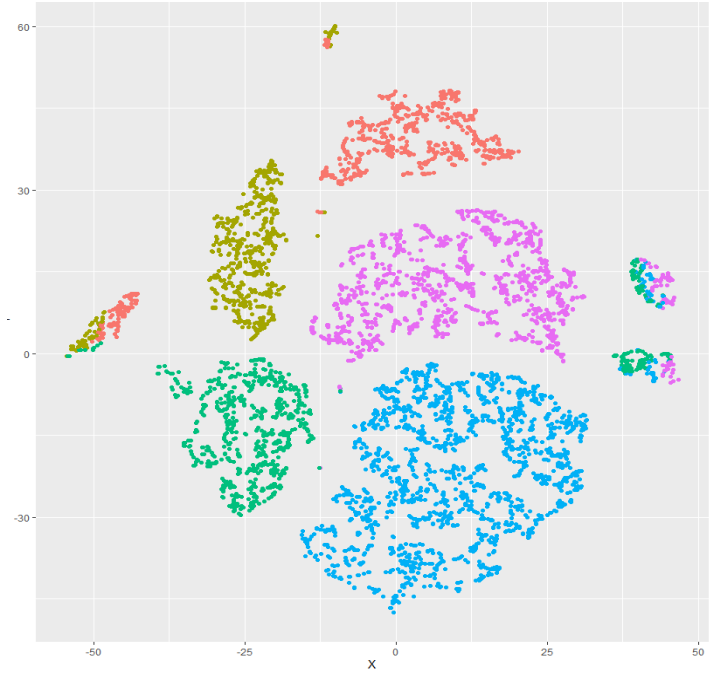
(2) 잠재변수: 클러스터 생성

먼저, 결측을 발생시키는 기준이 되는 잠재변수를 생성하고자 했다.
MAR 가정을 만족하기 위해 결측을 발생시킬 alcohol 변수는 제외하고, 설명변수 6개만을 사용하여 **클러스터링**을 진행했다.
비슷한 특성을 가진 데이터끼리 군집화하여, 특정 군집에 해당하는 데이터에서 결측을 발생시키고자 하였다.
Cluster 1, cluster 2, cluster 3, cluster 4, cluster 5를 값을 가지는 잠재변수 **cluster**을 생성했다.

1) 클러스터의 개수 선택 기준
Silhouette score의 증가세가 감소하기 시작하는 지점을
최적의 클러스터 개수로 결정했다.



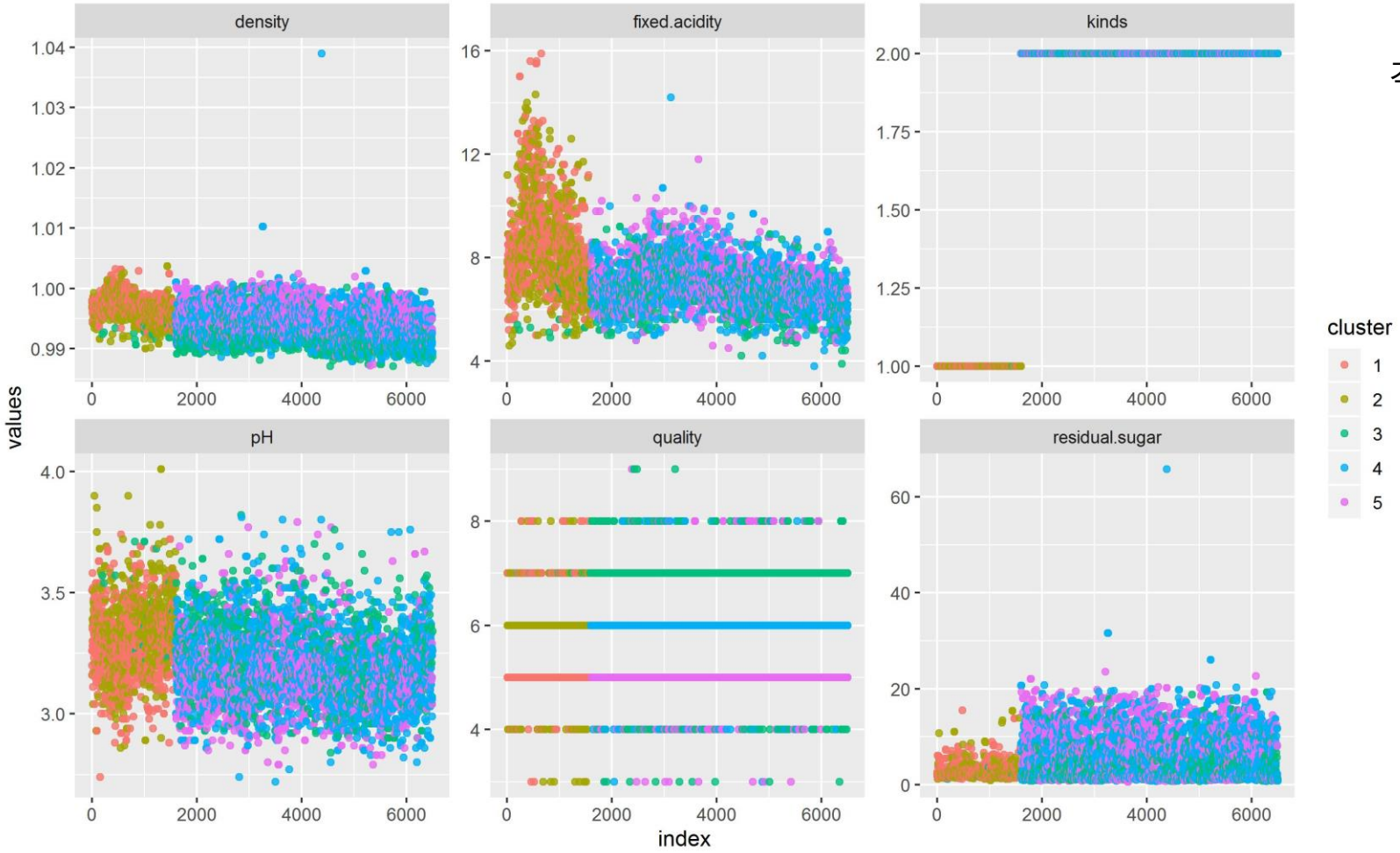
2) 클러스터 생성 방법
6개의 선택된 변수에 대해 K-means 알고리즘을 이용하여,
데이터를 5개의 클러스터로 군집화했다.



생성 클러스터(5개)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5

2. 결측자료 발생 시나리오

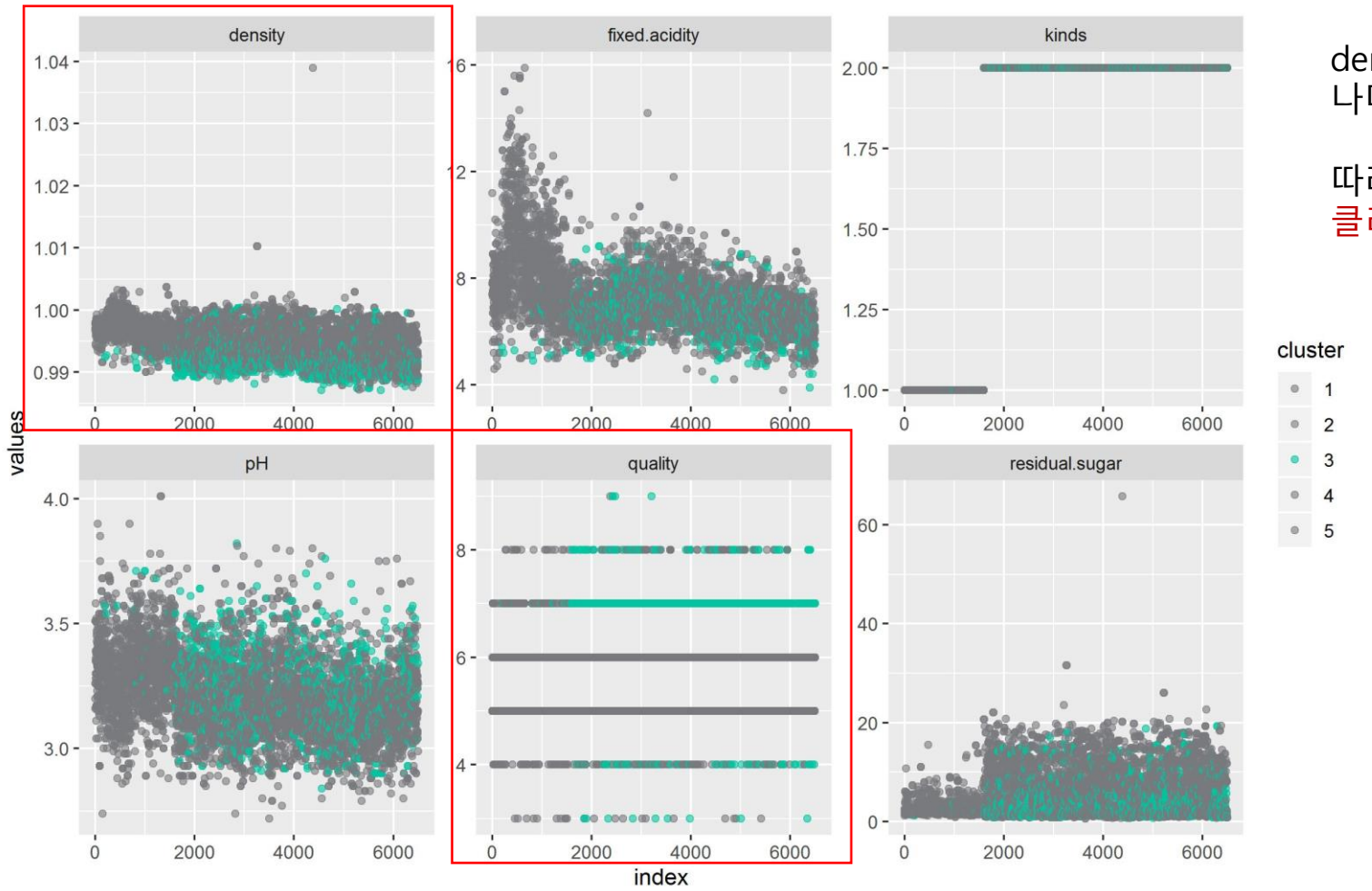
(2) 잠재변수: 클러스터 생성



각 변수의 클러스터별 분포는 다음과 같다.

2. 결측자료 발생 시나리오

(3) 결측을 발생시키는 클러스터 결정

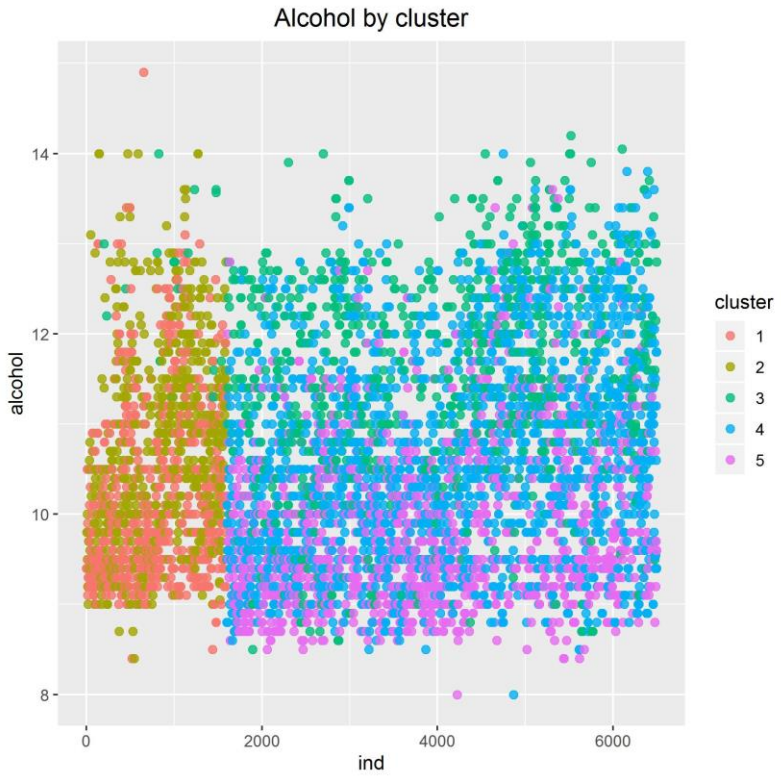


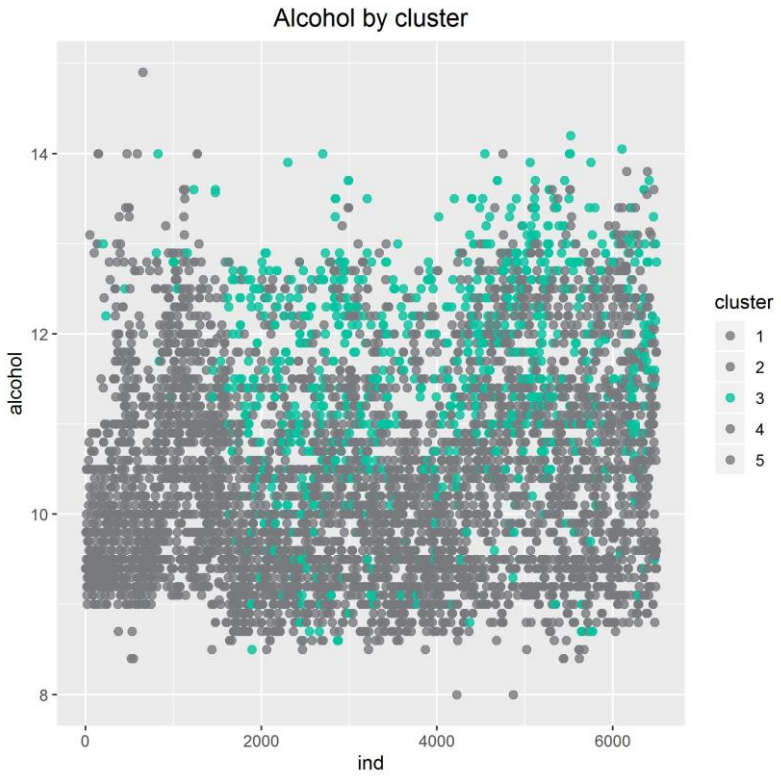
density와 quality 분포 양상에서 Cluster3과 나머지 Cluster들이 비교적 가시적인 차이를 보인다.

따라서 cluster3을 결측을 발생시키는 클러스터로 결정했다.

2. 결측자료 발생 시나리오

(3) 결측을 발생시키는 클러스터 결정





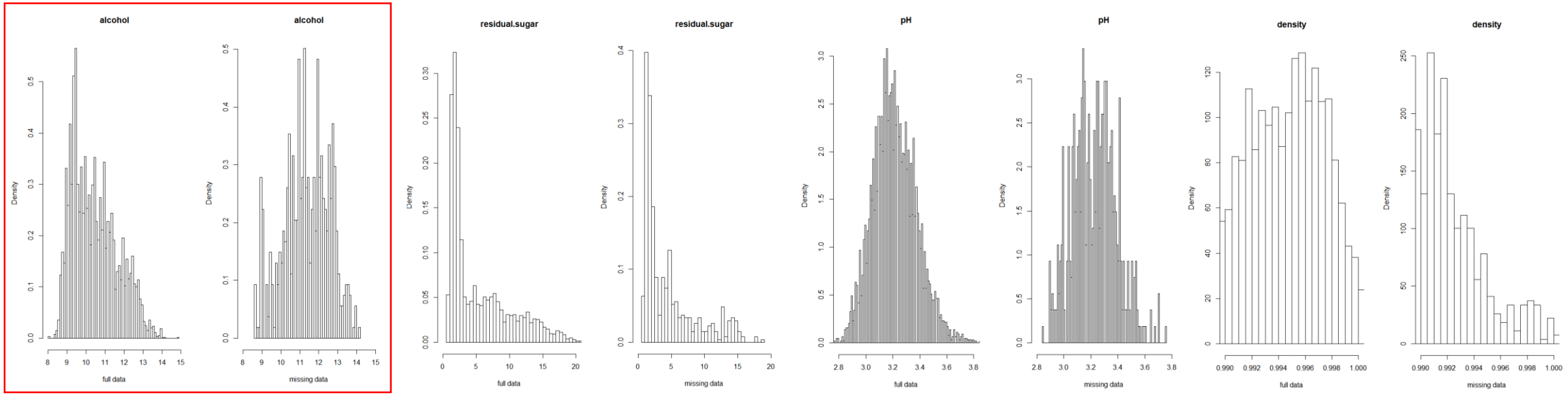
Alcohol 변수를 확인해보면, 클러스터3에 해당하는 값들이 대체로 Alcohol에서 높은 값에 해당하는 것을 확인할 수 있다.

따라서, 클러스터3에서 결측을 발생시키면, Alcohol 변수에서 결측으로 인한 Bias가 생길 것으로 짐작할 수 있었다.

2. 결측자료 발생 시나리오

(4) 결측 발생

alcohol 변수에서 cluster3에 해당하는 unit 중, 결측을 랜덤으로 50% 발생시켰다.
즉, 1076개의 데이터가 존재하는 cluster3의 값에서 538개의 데이터를 결측 처리하였다.



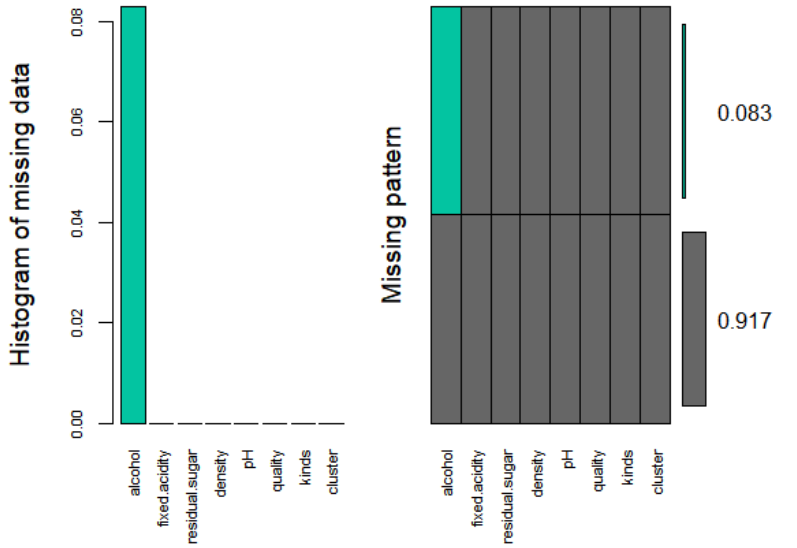
각 변수마다 full data 와 missing data 분포를 비교하는 그래프이다.
alcohol에서 full data와 missing data의 분포가 확연히 다른 것을 확인할 수 있다.
다른 변수들도 마찬가지로 대체로 full data와 missing data의 분포가 상이했다.

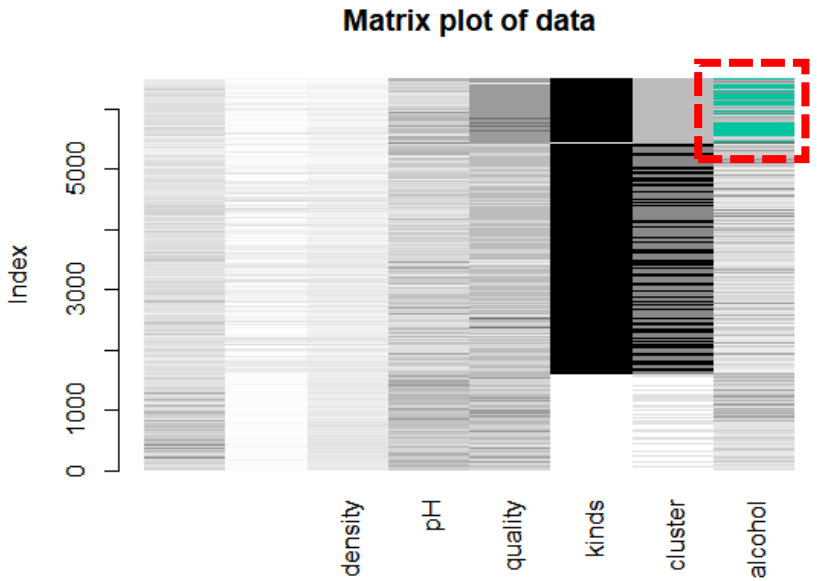
2. 결측자료 발생 시나리오

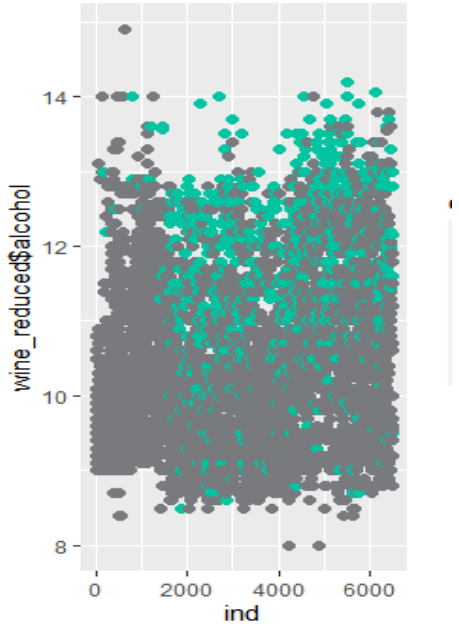
(5) 결측 패턴 시각화

전체적인 missing data 발생 패턴을 시각화했다.

Missing data의 alcohol이 전반적으로 높은 값에 형성되어 있음을 볼 수 있다.







3. Imputation Method

Imputation Method

- Unconditional mean imputation
- Regression imputation
- Stochastic regression imputation
- Mean imputation
  - Imputation cell
  - Nearest neighbor
- Hotdeck imputation
  - Imputation cell
  - Nearest neighbor

Evaluation

Missing data
- Original mean vs Imputed mean (complete data)
- Original variance vs Imputed variance (complete data)

Total data
- Original mean vs Imputed mean (complete data)
- Original variance vs Imputed variance (complete data)

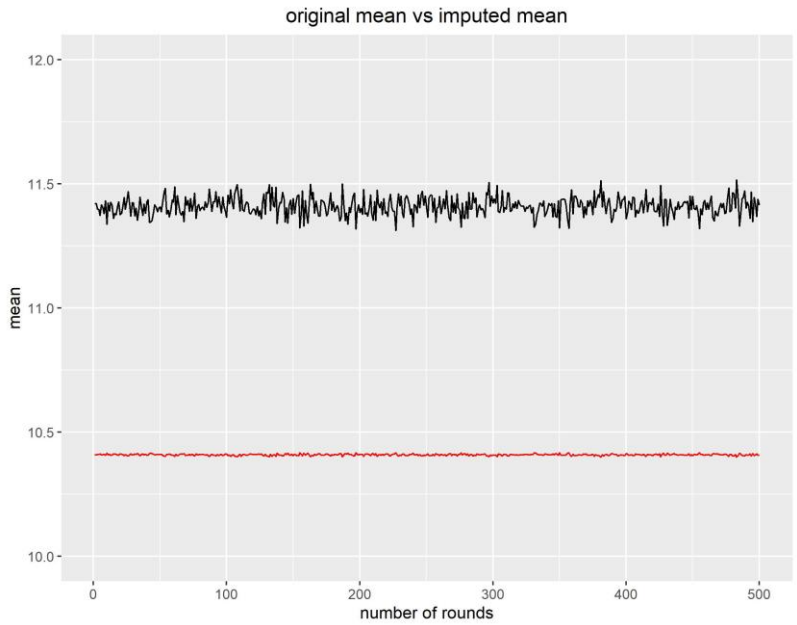
결측을 발생시키고 7개의 Imputation 방법을 각각 적용해보는 것을 **500번** 반복하여,
분포 및 평균, 분산을 비교하고자 하였다.

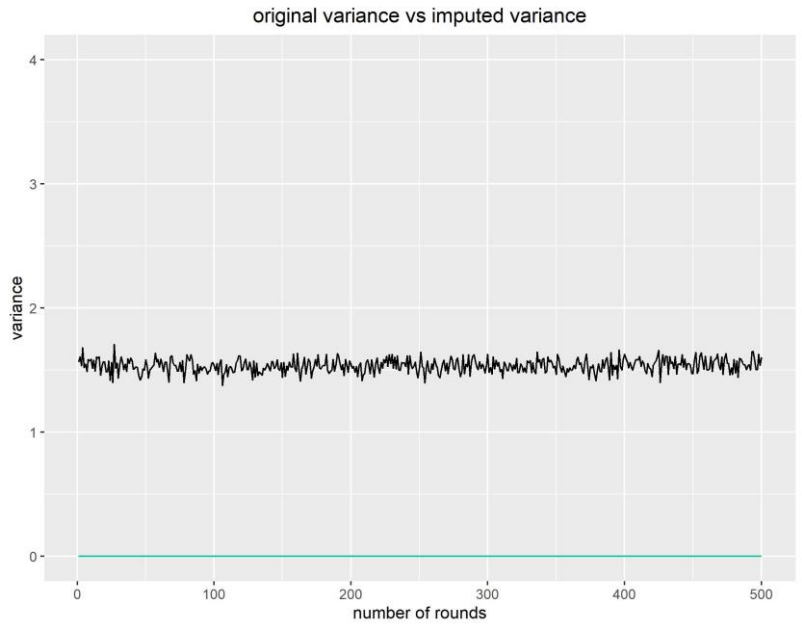
3. Imputation Method

(1) Unconditional mean imputation

결측치를 alcohol의 관측된 값들의 평균으로 대체했다.

(1) missing data의 original mean/variance 와 imputed missing data의 mean/variance 비교





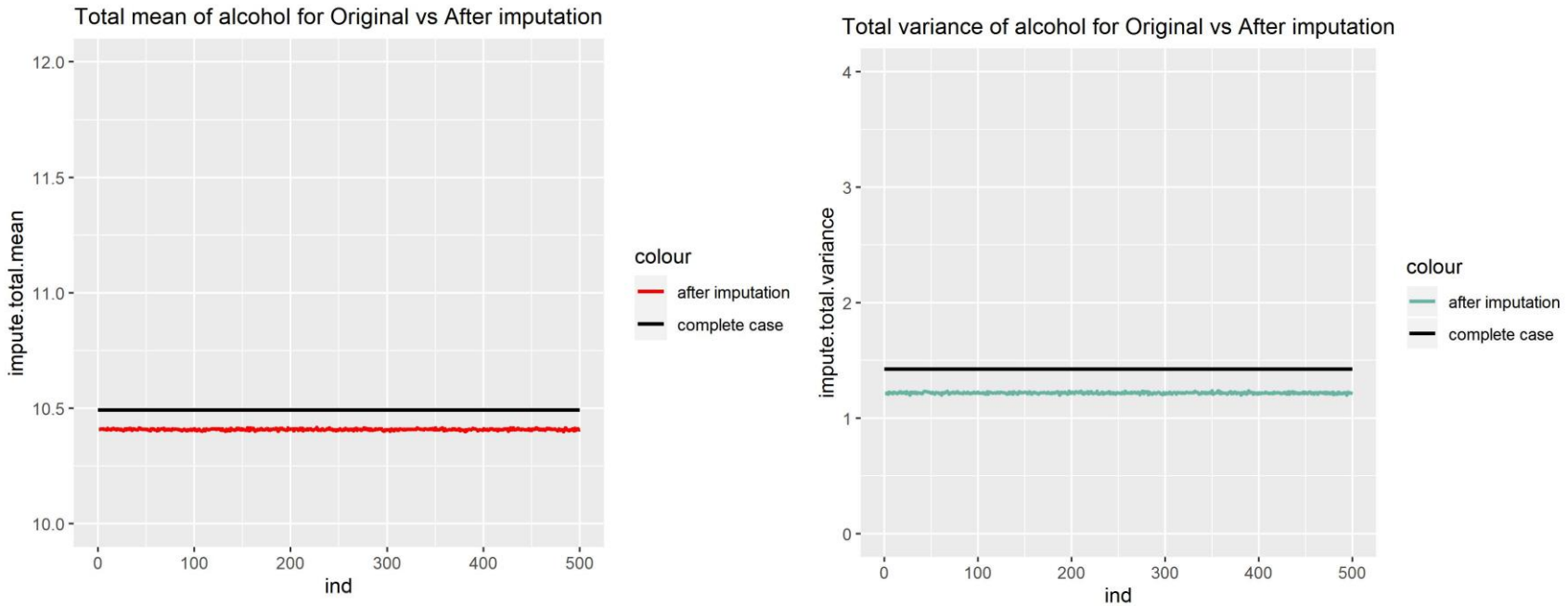
: Original data : Imputed data

- Unconditional mean imputation 방법으로 missing data를 impute 하면, missing data의 평균은 전체 데이터의 평균과 같고, 대체한 값들의 분산은 0이다.
- 앞서 예측한대로, alcohol 변수에서 결측이 발생한 값들이 비교적 큰 값을 가지므로, 관측된 값들의 평균이 그보다 작게 된다. 따라서, imputed data의 mean이 original mean보다 확연히 작은 것을 볼 수 있었다.

3. Imputation Method

(1) Unconditional mean imputation

(2) Total data의 original mean/variance 와 imputed data의 mean/variance 비교



: Original data : Imputed data

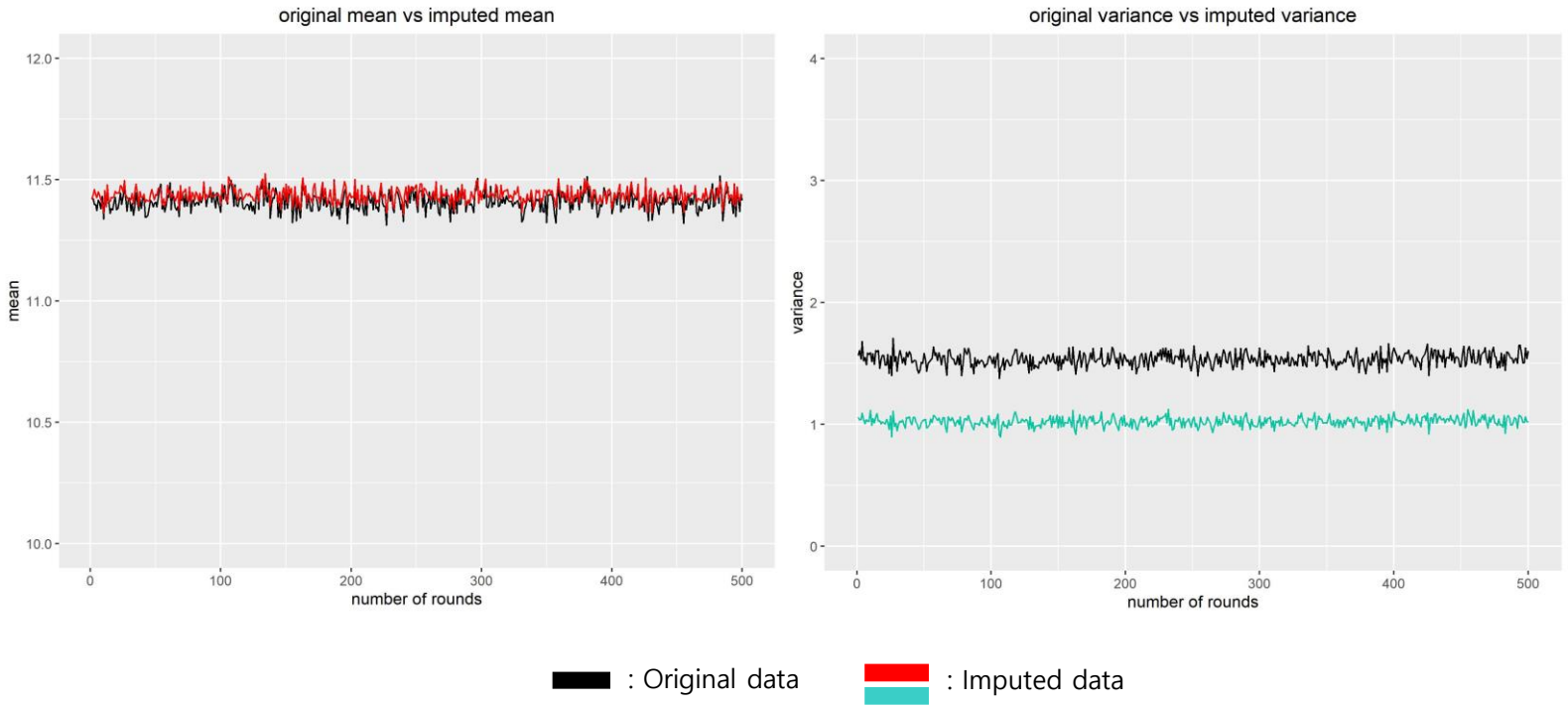
- Umean imputation을 적용하면 missing data 가 모두 complete mean으로 대체되기 때문에 분산이 과소추정 된다.

3. Imputation Method

(2) Regression imputation(Cmean)

Wine data의 complete set에서 반응변수를 alcohol, 설명변수를 나머지 변수들로 하는 linear regression을 적합하였다.
이 모델을 이용하여 결측치를 예측하고, 예측값으로 결측치를 대체하였다.

(1) missing data의 original mean/variance 와 imputed missing data의 mean/variance 비교

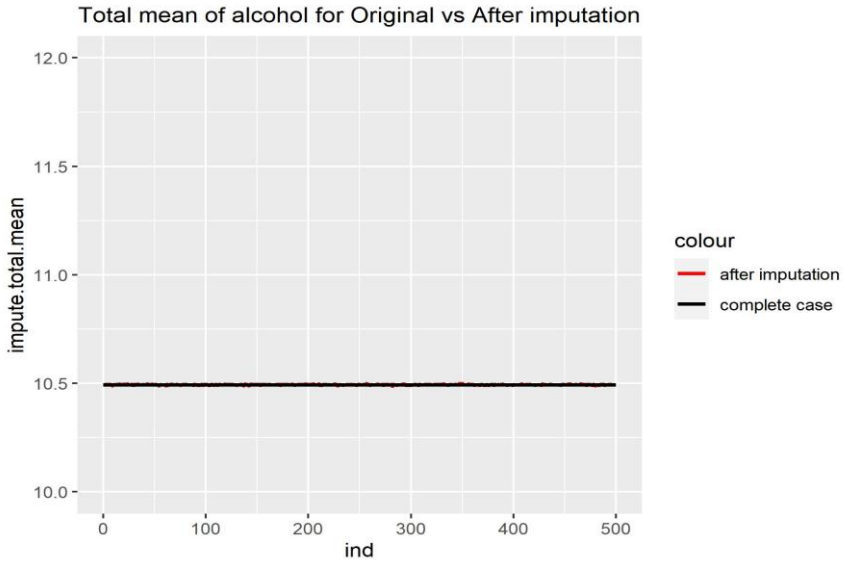


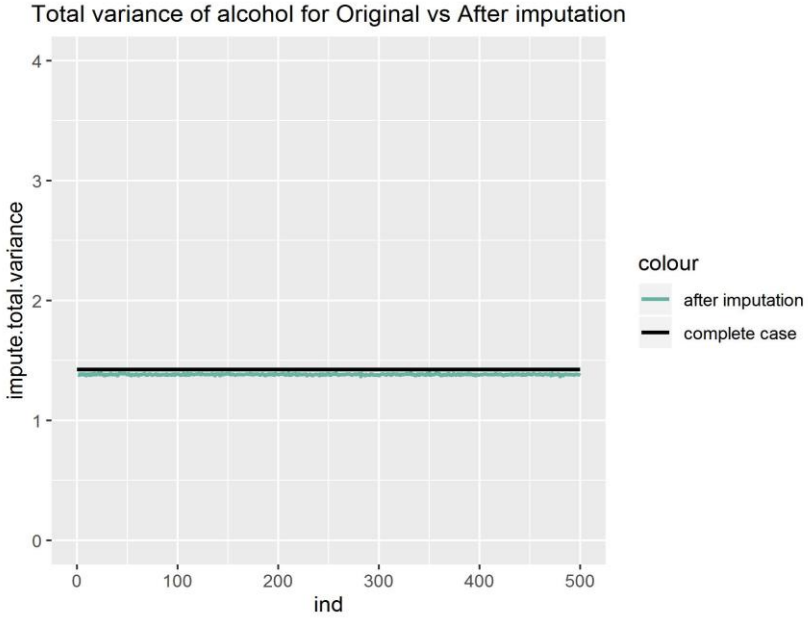
- Regression imputation으로 결측치를 대체하면, original missing data의 mean과 imputed mean의 차이가 거의 없는 것을 확인할 수 있었다.
- Imputed variance는 original variance에 비해 감소한 것을 확인할 수 있다. Regression line을 통해 예측된 alcohol의 값들은 나머지 설명변수들을 조건으로 하는 Conditional Mean에 해당하므로, Variability는 실제보다 감소하게 될 여지가 많다.

3. Imputation Method

(2) Regression imputation(Cmean)

(2) Total data의 original mean/variance 와 imputed data의 mean/variance 비교





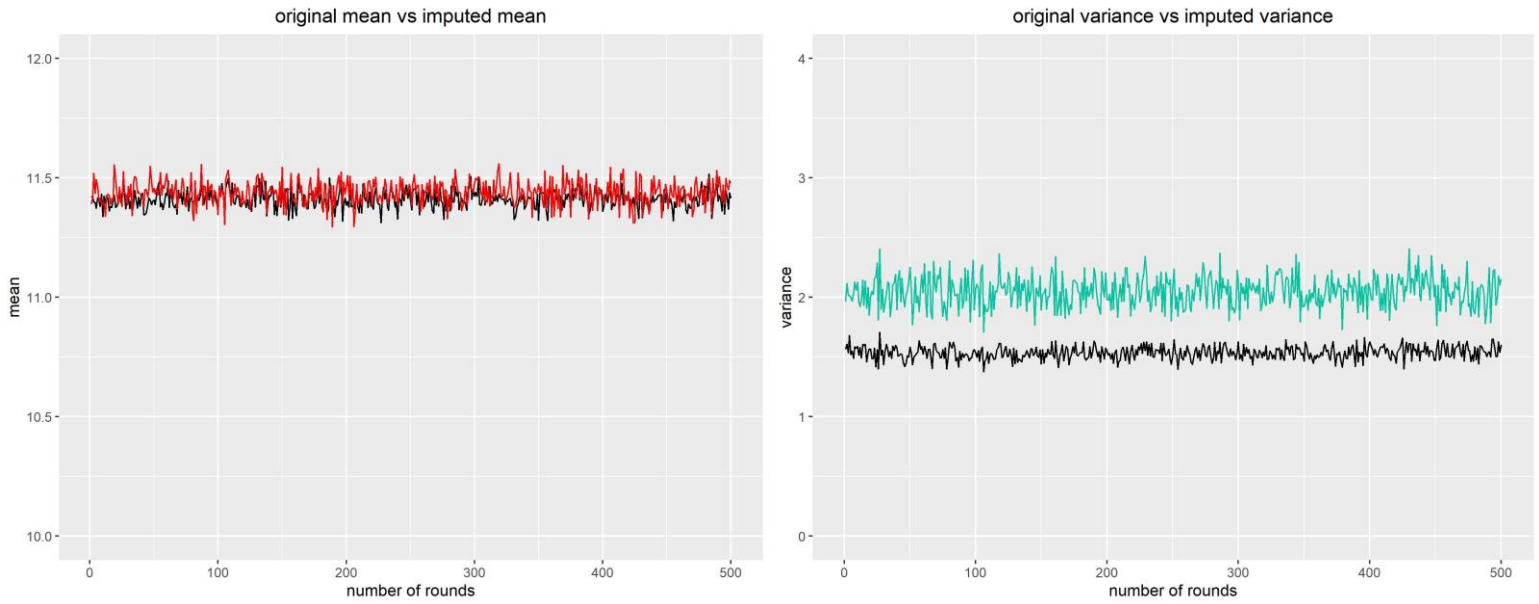
: Original data : Imputed data

3. Imputation Method

(3) Stochastic regression imputation(Cdraw)

Method (2)와 비슷하지만, linear regression model에 residual variance를 추가하여 missing data를 예측했다.

(1) missing data의 original mean/variance 와 imputed missing data의 mean/variance 비교

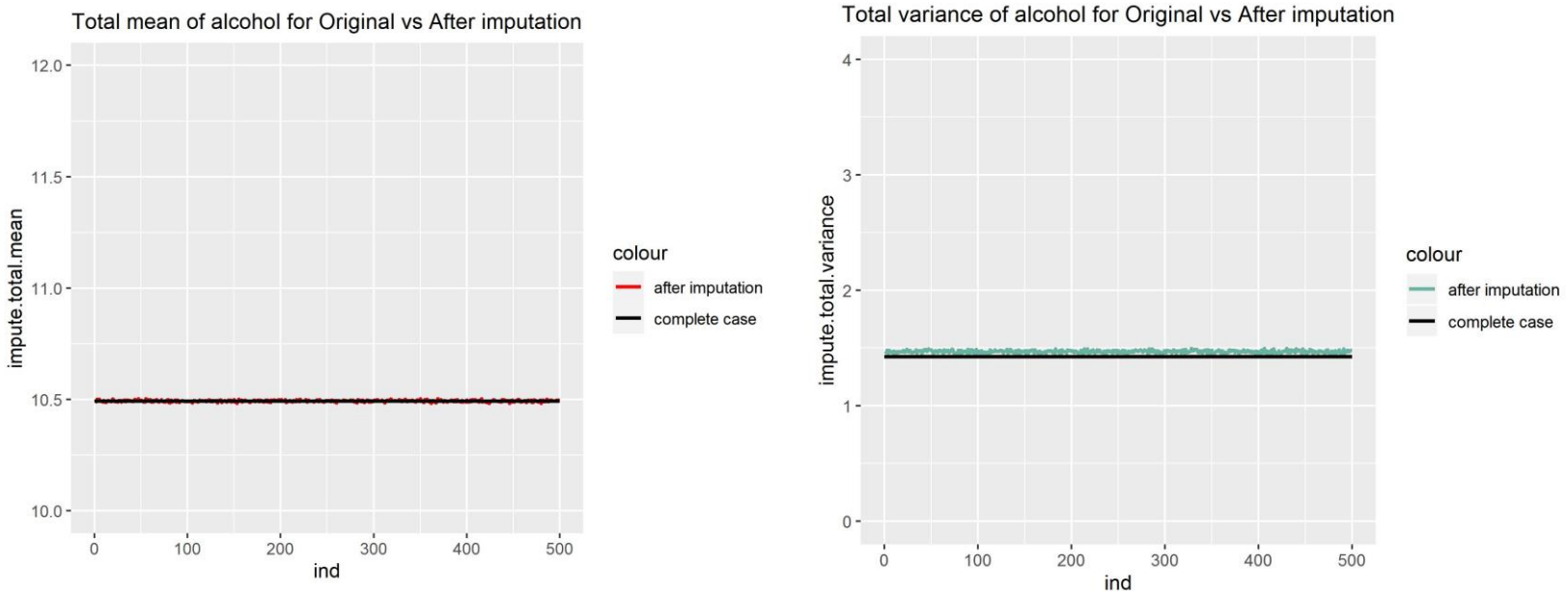


- Stochastic regression imputation으로 missing data를 처리하면 original missing data의 mean과 imputed mean의 차이는 거의 없는 것을 확인할 수 있다.
- Imputed variance는 original variance 보다 크다. Stochastic regression은 Cmean에 에러를 부여함으로써 분산의 과소추정을 방지한다.

3. Imputation Method

(3) Stochastic regression imputation(Cdraw)

(2) Total data의 original mean/variance 와 imputed data의 mean/variance 비교



: Original data : Imputed data

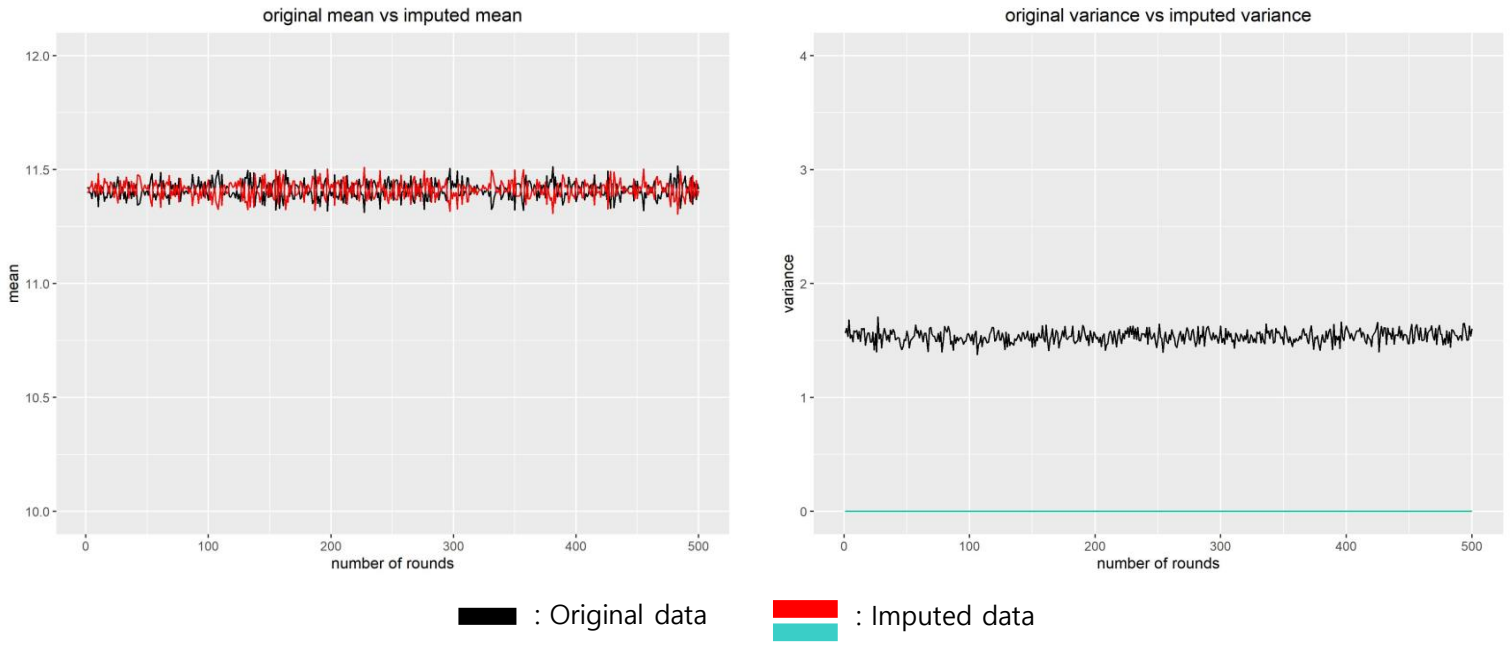
3. Imputation Method

(4) Mean imputation within imputation cell

Imputation cell의 선정은 alcohol을 제외한 나머지 변수들로 클러스터링을 진행한 결과로 결정했다.
이 클러스터 결과는 결측 발생했던 결과와 동일한데, 결측이 발생한 데이터는 모두 cluster3에 해당했으므로,
결측값을 cluster3의 complete units의 alcohol의 평균으로 대체했다.

Cluster 결과를 imputation cell로 설정한 이유:
클러스터 결과는 나머지 설명변수들을 모두 반영하여 데이터를 비슷한 특성끼리 그룹화한 것이므로, 적절한 imputation cell 선정 방법이라고 판단했다.

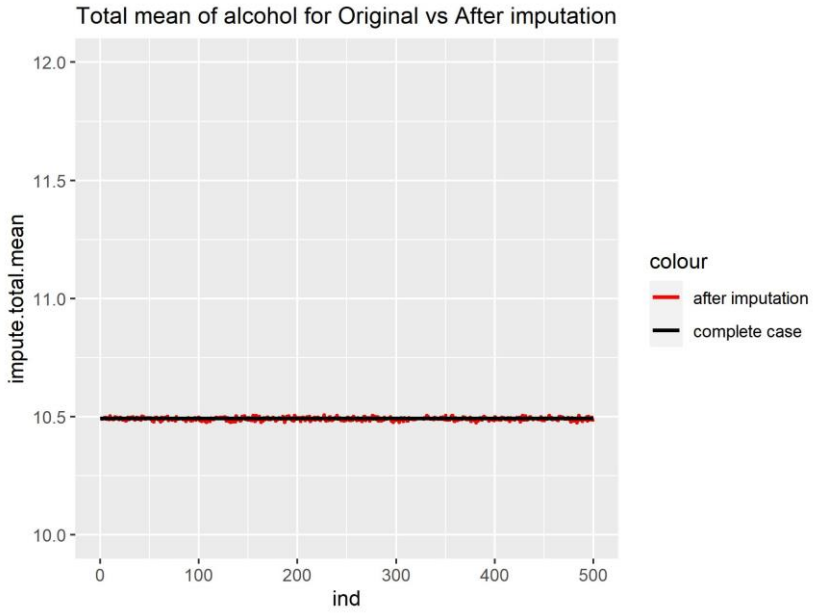
(1) missing data의 original mean/variance 와 imputed missing data의 mean/variance 비교

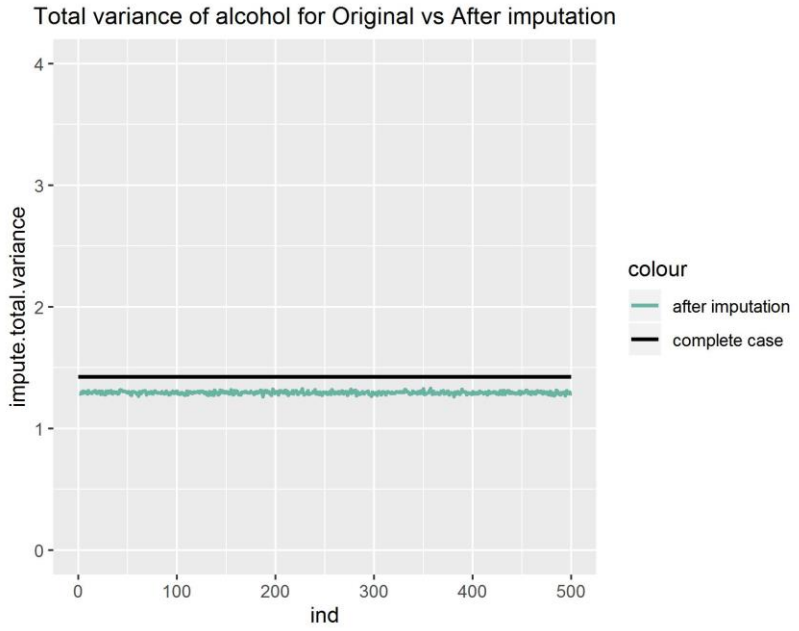


3. Imputation Method

(4) Mean imputation within imputation cell

(2) Total data의 original mean/variance 와 imputed data의 mean/variance 비교



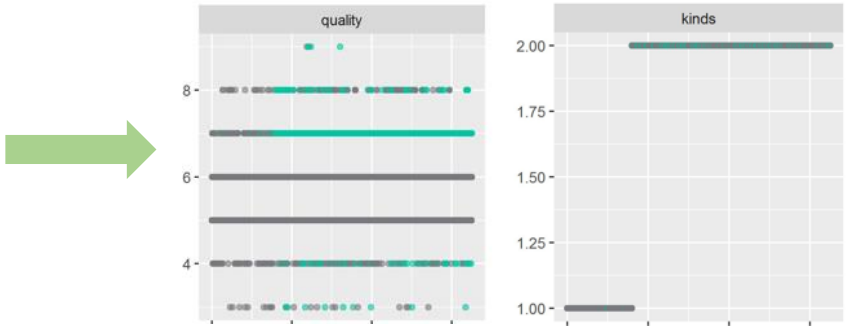


: Original data : Imputed data

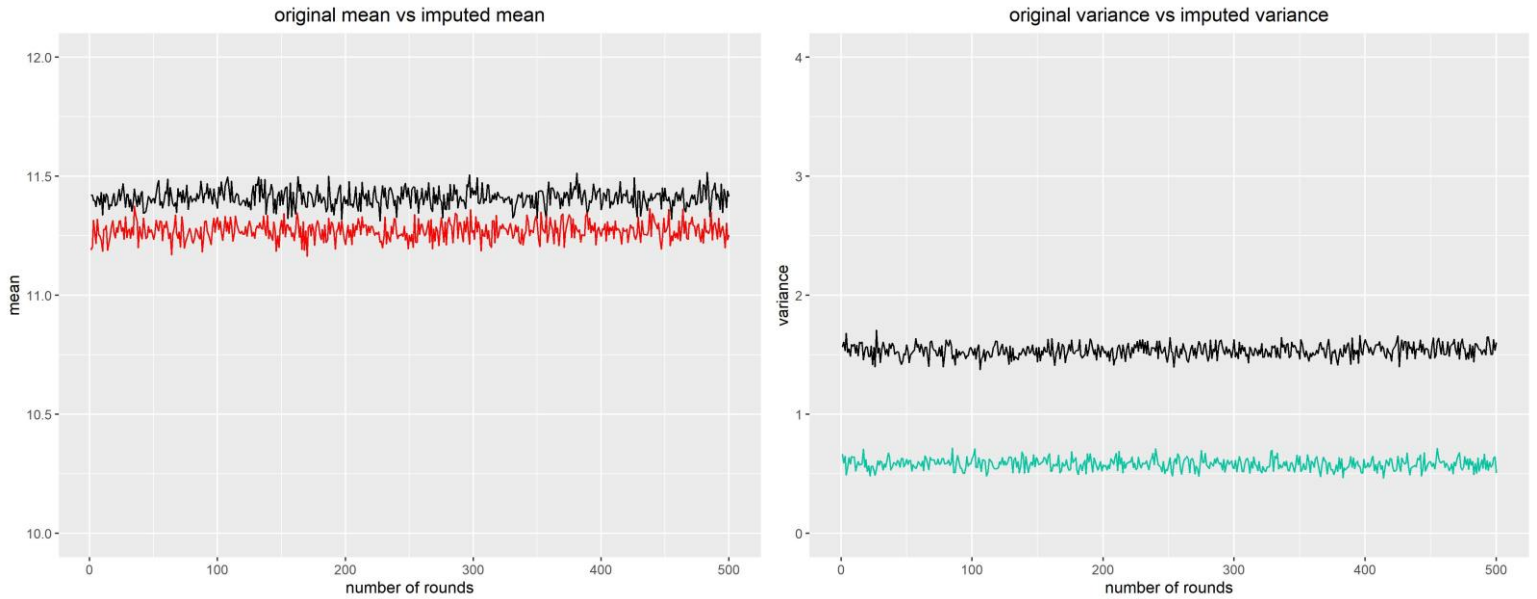
3. Imputation Method

(5) Mean imputation by nearest neighbor

Step 1: 거리 계산을 위해 데이터 표준화
Step 2: cluster 3이 속한 범주형 데이터(quality=3,4,7,8,9, kinds=white)만 선택
Step 3: Euclidean distance로 변수 간 거리 계산
Step 4: Euclidean distance가 작은(가까운) 데이터 20개를 선택
Step 5: 이웃 데이터 20개의 alcohol mean을 구해 결측값 대체



(1) missing data의 original mean/variance 와 imputed missing data의 mean/variance 비교

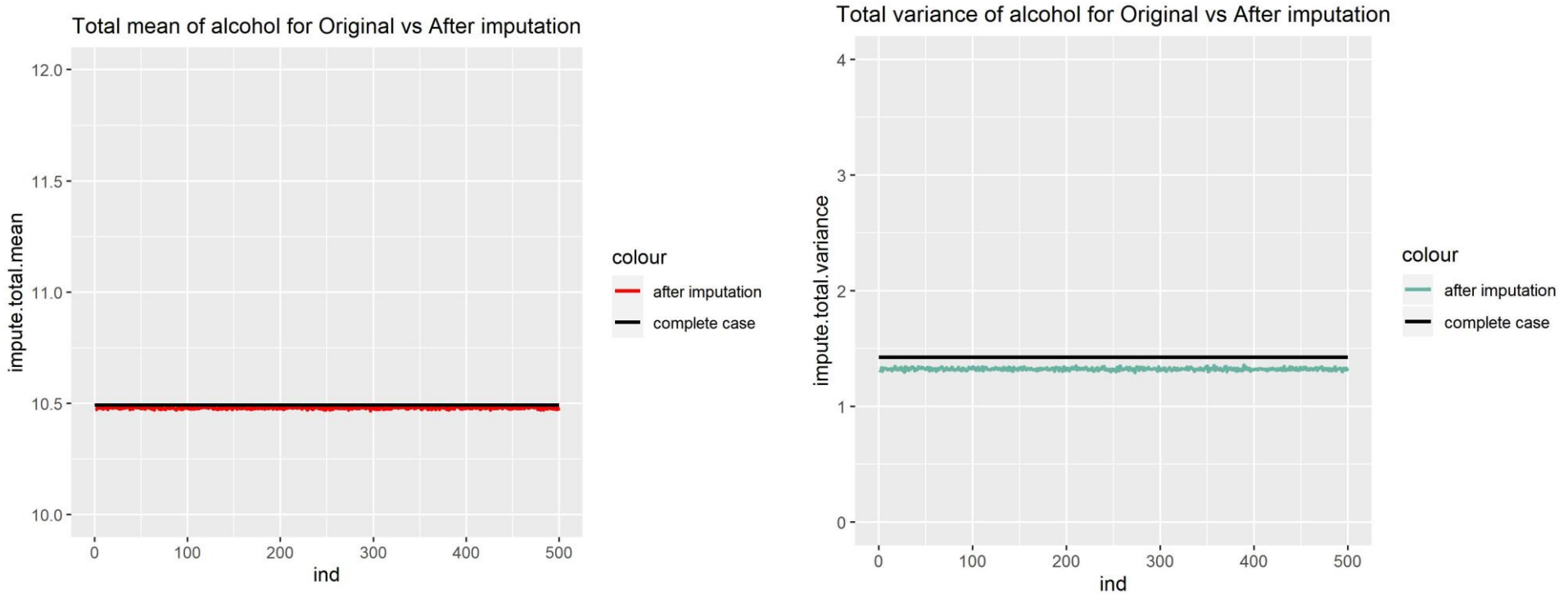


: Original data : Imputed data

3. Imputation Method

(5) Mean imputation by nearest neighbor

(2) Total data의 original mean/variance 와 imputed data의 mean/variance 비교



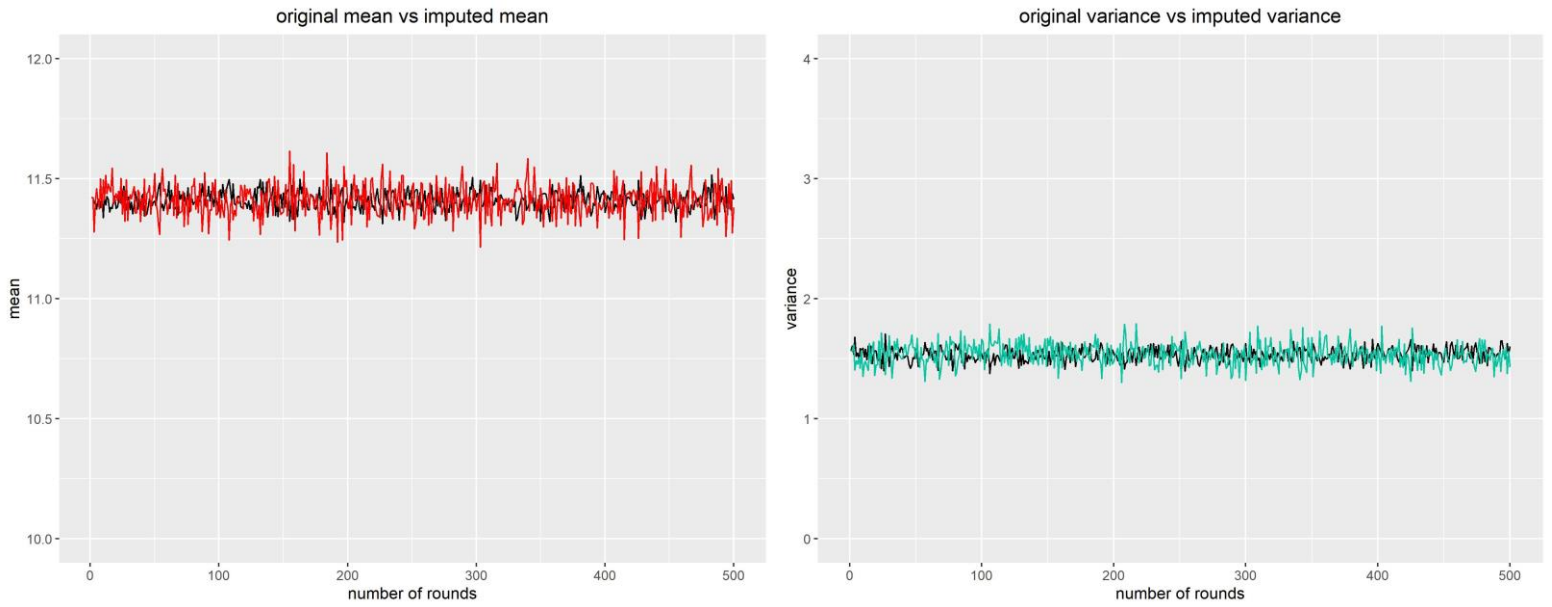
: Original data : Imputed data

3. Imputation Method

(6) Hotdeck imputation within imputation cell

Mean imputation within imputation cell방법과 마찬가지로, 클러스터 결과를 imputation cell로 선정했다.
Cluster3 alcohol의 complete units에서 랜덤으로 doner을 선택하고 결측값을 doner의 값으로 대체했다.

(1) missing data의 original mean/variance 와 imputed missing data의 mean/variance 비교

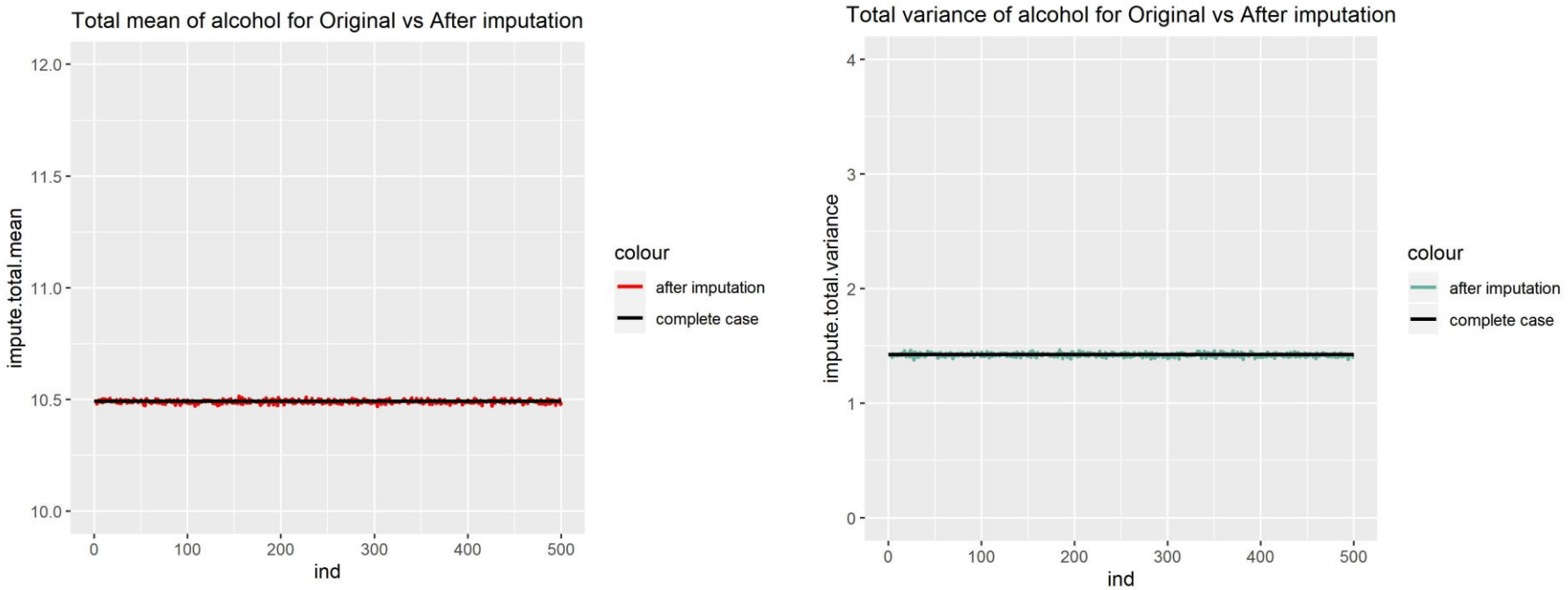


: Original data : Imputed data

3. Imputation Method

(6) Hotdeck imputation within imputation cell

(2) Total data의 original mean/variance 와 imputed data의 mean/variance 비교



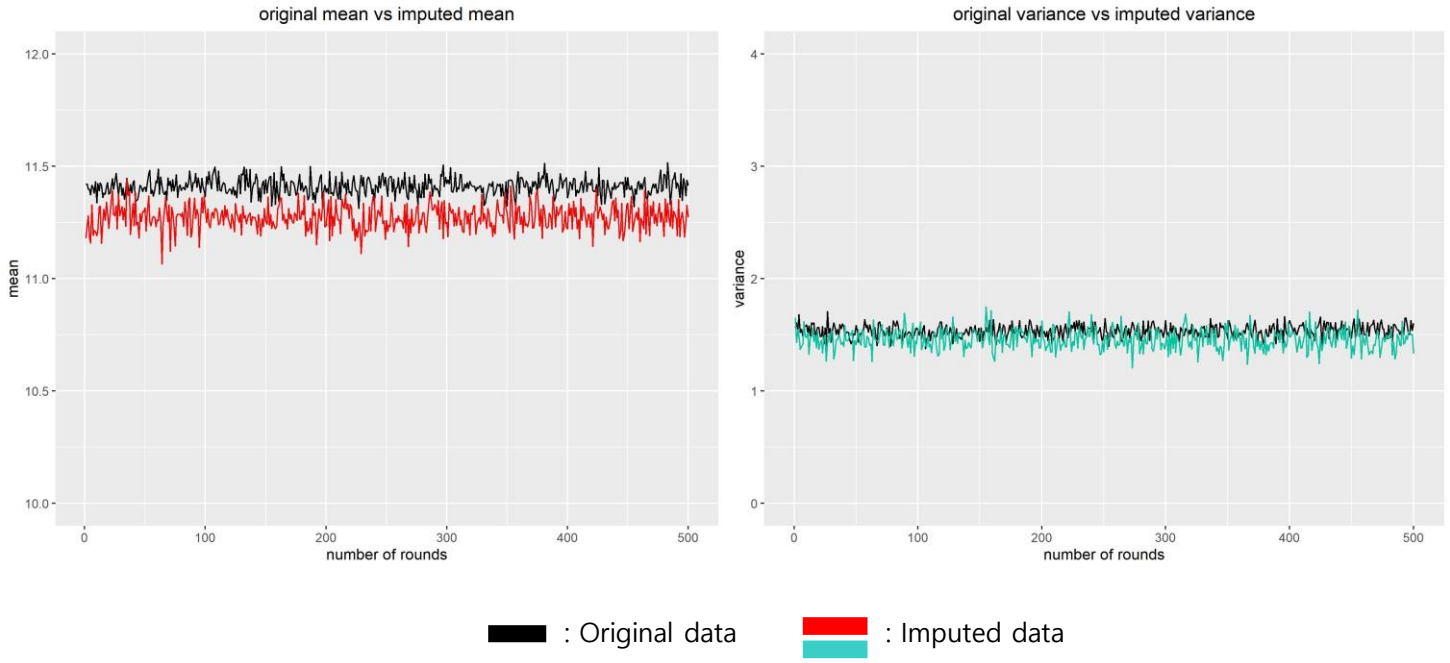
: Original data : Imputed data

3. Imputation Method

(7) Hotdeck imputation by nearest neighbor

Step 1: 거리 계산을 위해 데이터 표준화
Step 2: 결측값이 있는 범주들에 해당하는 범주형 데이터만 선택(원활한 거리계산)
Step 3: Euclidean distance로 변수 간 거리 계산
Step 4: Euclidean distance가 작은(가까운) 데이터 20개를 선택
Step 5: 이웃 데이터 20개의 alcohol 값을 doner로 설정하여 결측값을 doner의 값으로 대체

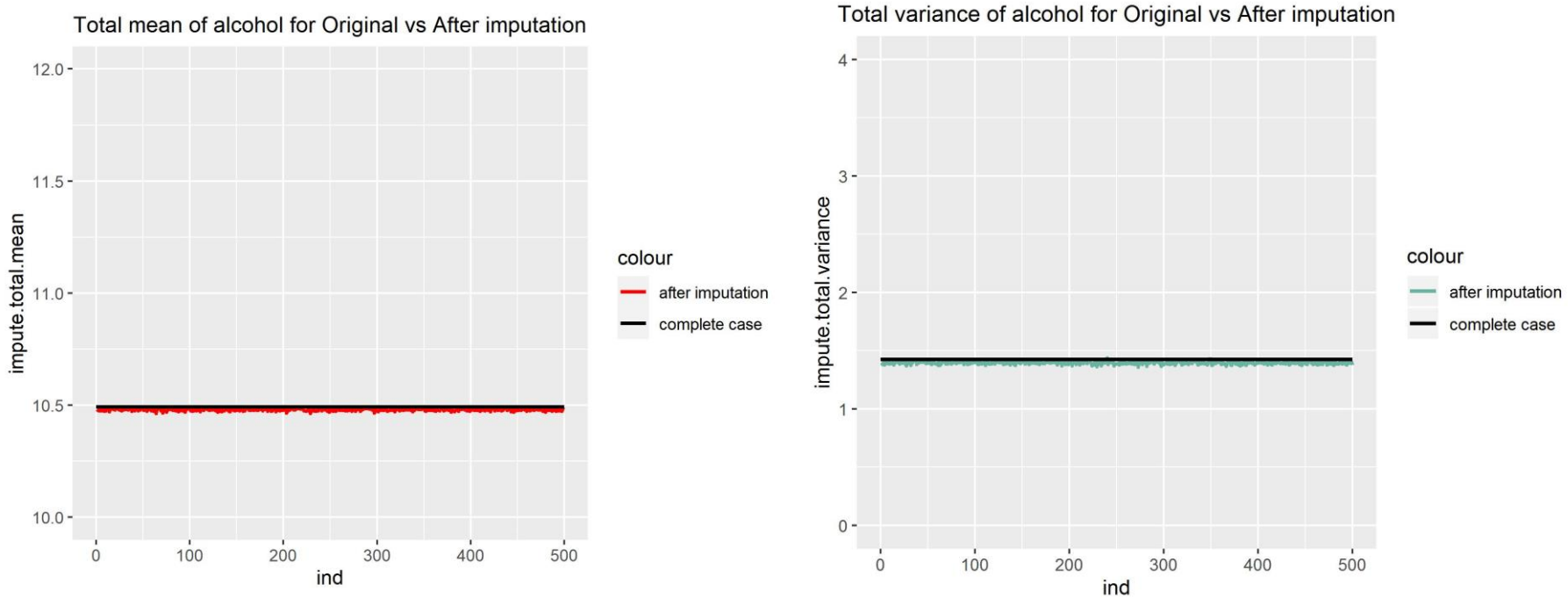
(1) missing data의 original mean/variance 와 imputed missing data의 mean/variance 비교



3. Imputation Method

(7) Hotdeck imputation by nearest neighbor

(2) Total data의 original mean/variance 와 imputed data의 mean/variance 비교



: Original data : Imputed data

4. 결론

(1) Imputation sample data

각 Imputation method로 imputation을 한 데이터 중 5개를 sampling 한 결과이다. (original value와 비교)

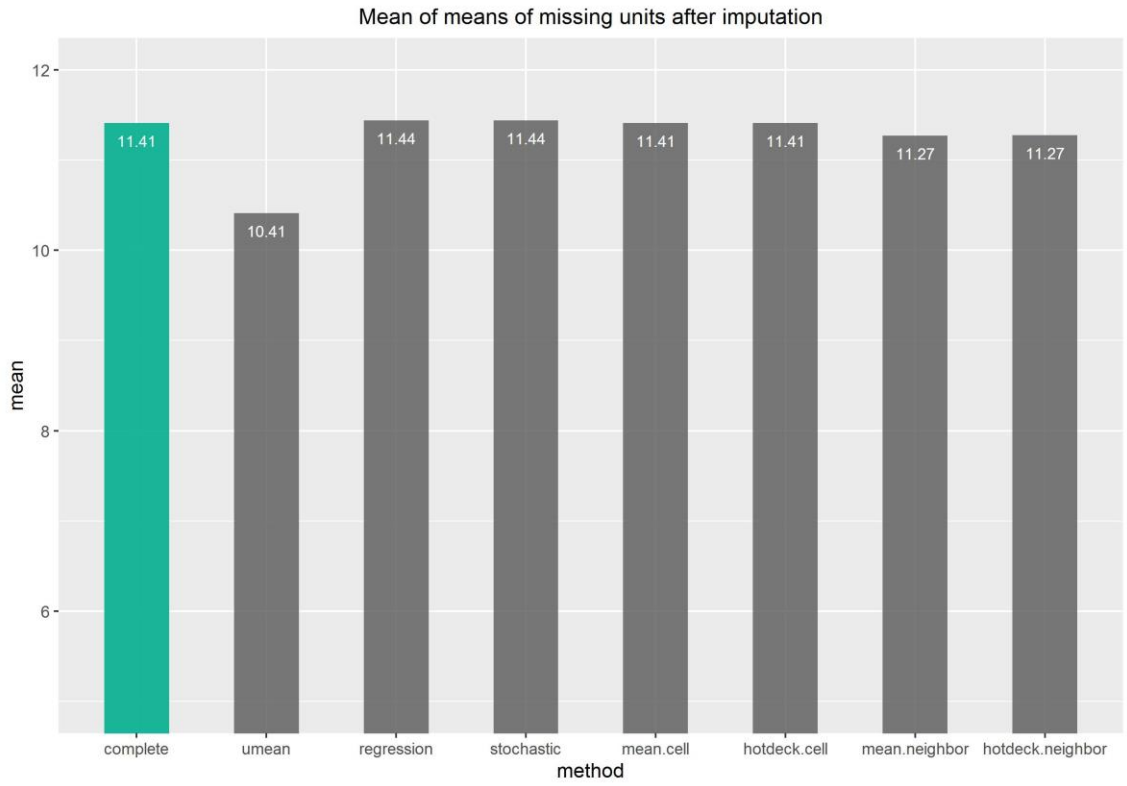
index	original.value	unconditional.mean.value	regression.value	stochastic.value	Mean.cluster.value	Hotdeck.cluster.value	Mean.nearest.neighbor.value	Hotdeck.nearest.neighbor.value
2487	11	10.41176148	11.40019538	12.0445825	11.44304213	10.4	11.635	10.6
5111	11.7	10.41176148	11.70229355	11.30939078	11.44304213	12.3	11.57	10
4493	12	10.41176148	11.66953968	12.71554587	11.44304213	12.3	11.64	11.8
4818	9.9	10.41176148	9.784582497	11.91447036	11.44304213	10	11.92	12.4
3437	11	10.41176148	11.46569313	13.04015325	11.44304213	12.3	11.92	12.6

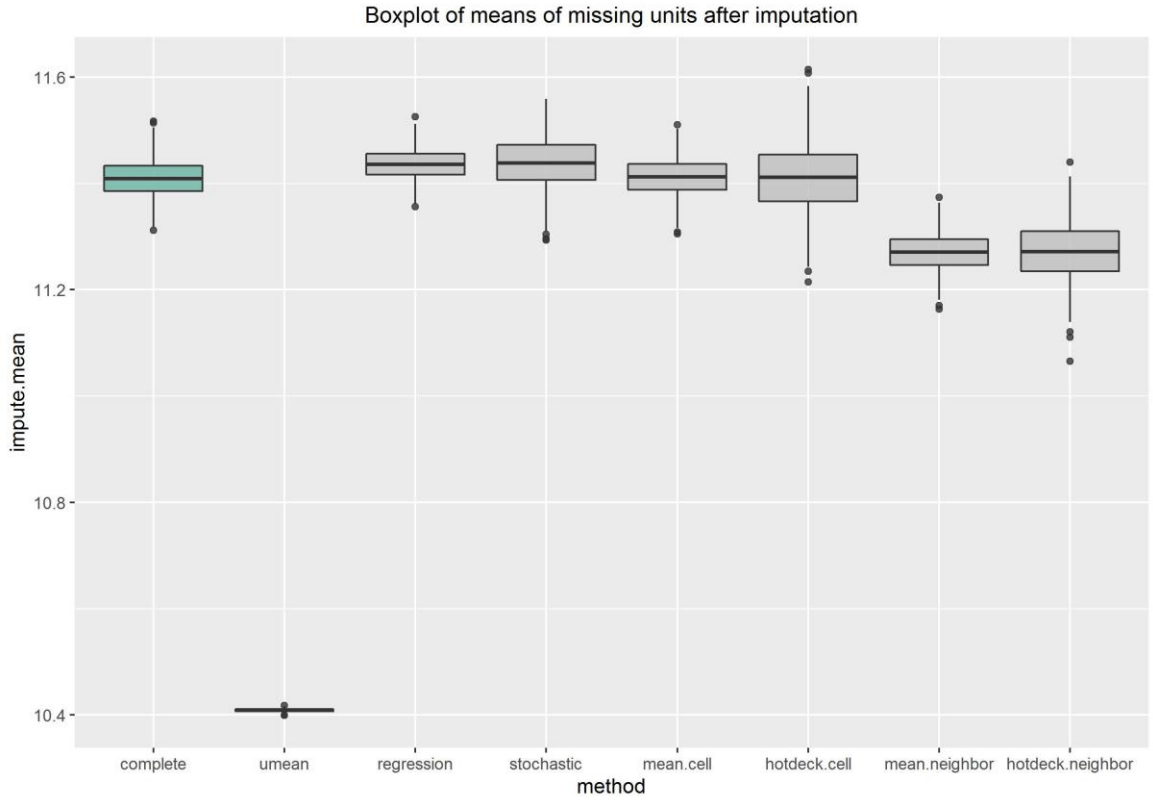
4. 결론

(2) Imputation method 비교

Original mean/variance와 각 imputation method를 적용한 결과를 비교하면 다음과 같다.

- Missing Unit들의 Imputation 후 Mean들의 평균 및 boxplot





4. 결론

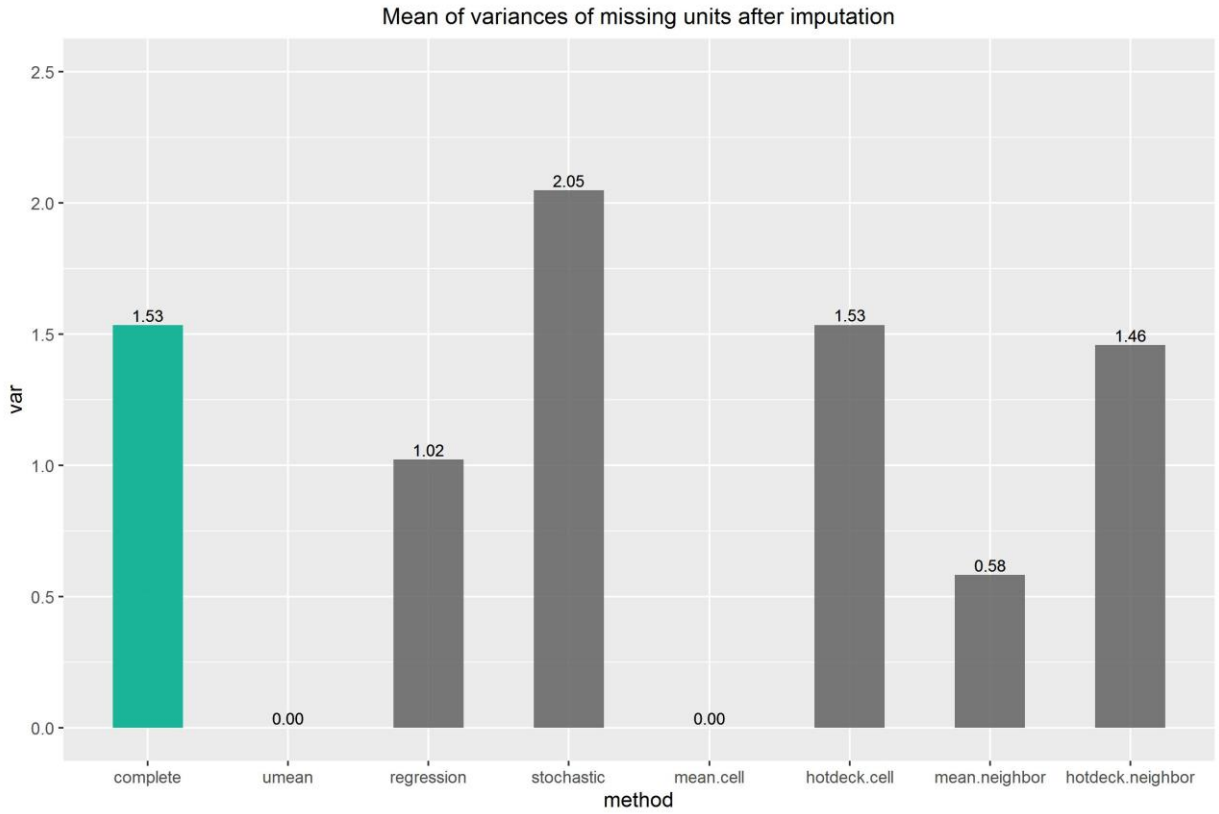
(2) Imputation method 비교

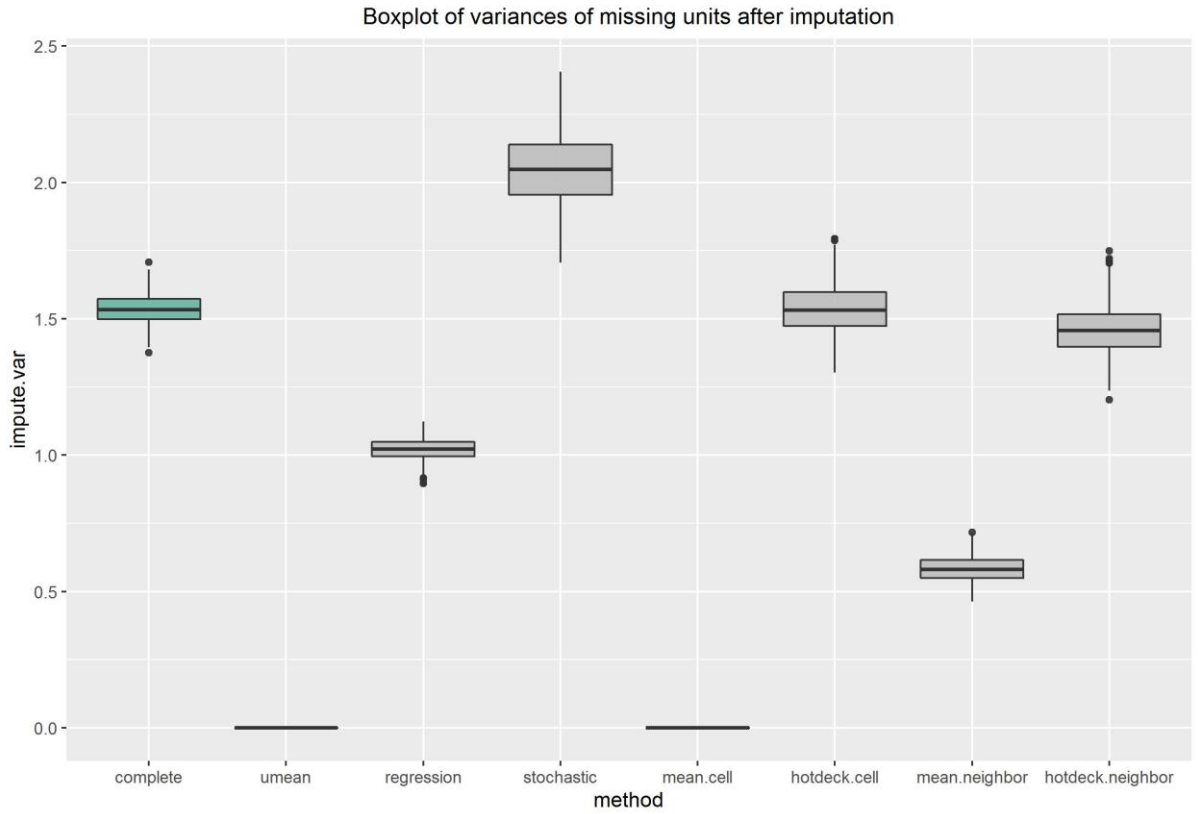
- Missing Unit들의 Imputation 후 Mean들의 Scatterplot

4. 결론

(2) Imputation method 비교

- Missing Unit들의 Imputation 후 Variance들의 평균 및 boxplot

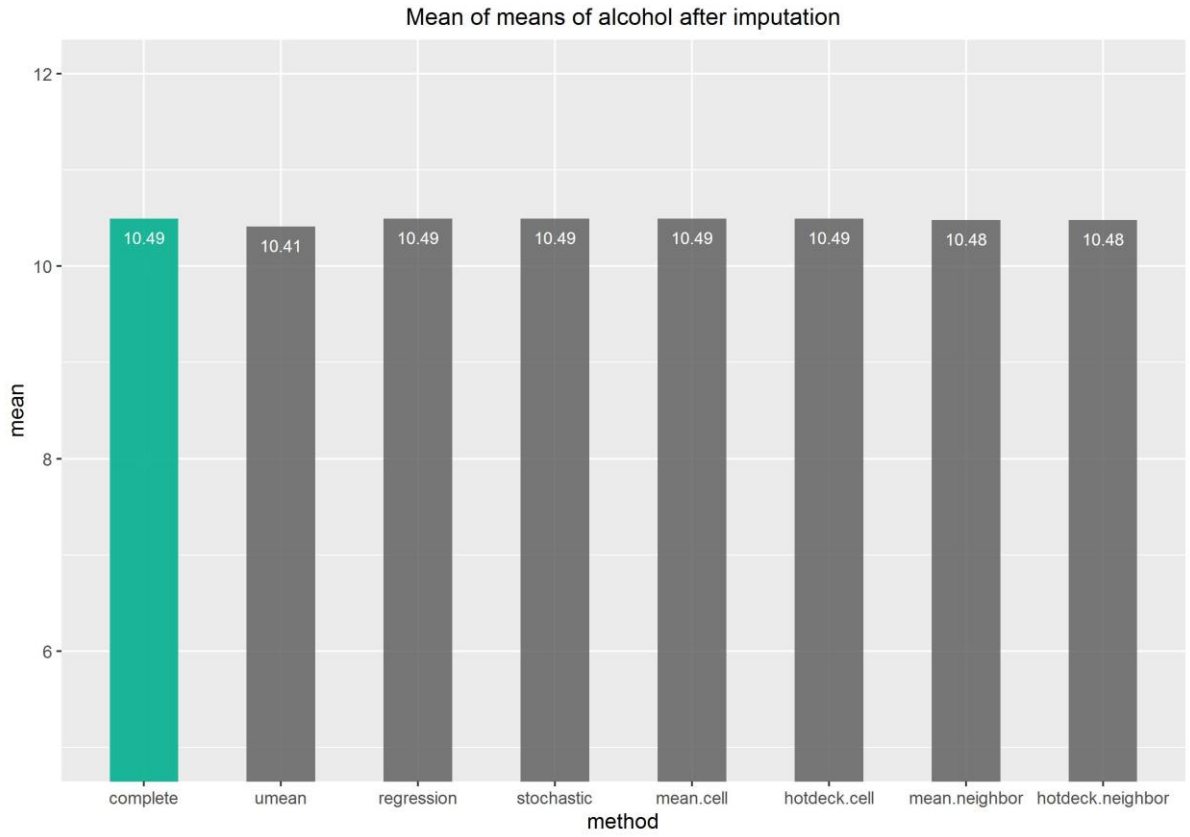


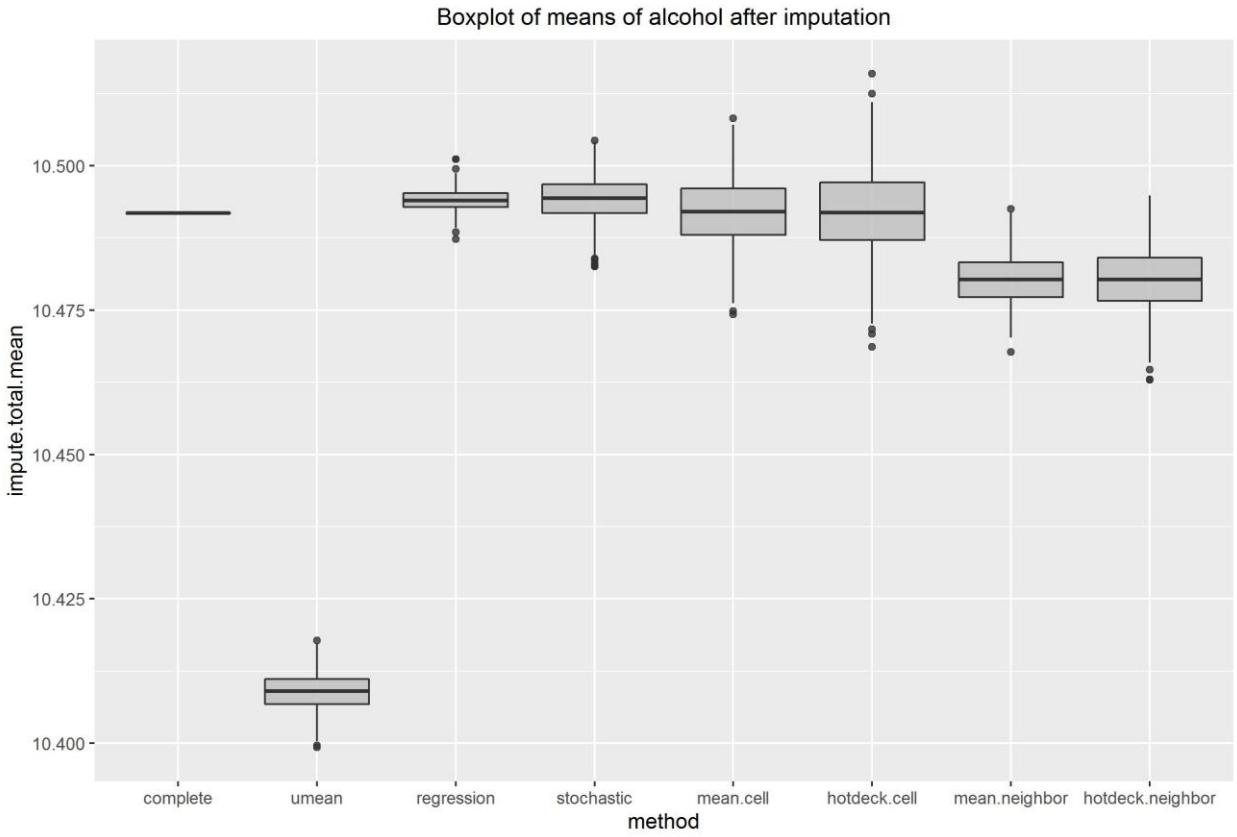


4. 결론

(2) Imputation method 비교

- 전체 Alcohol 변수의 Imputation 후 Mean들의 평균 및 boxplot

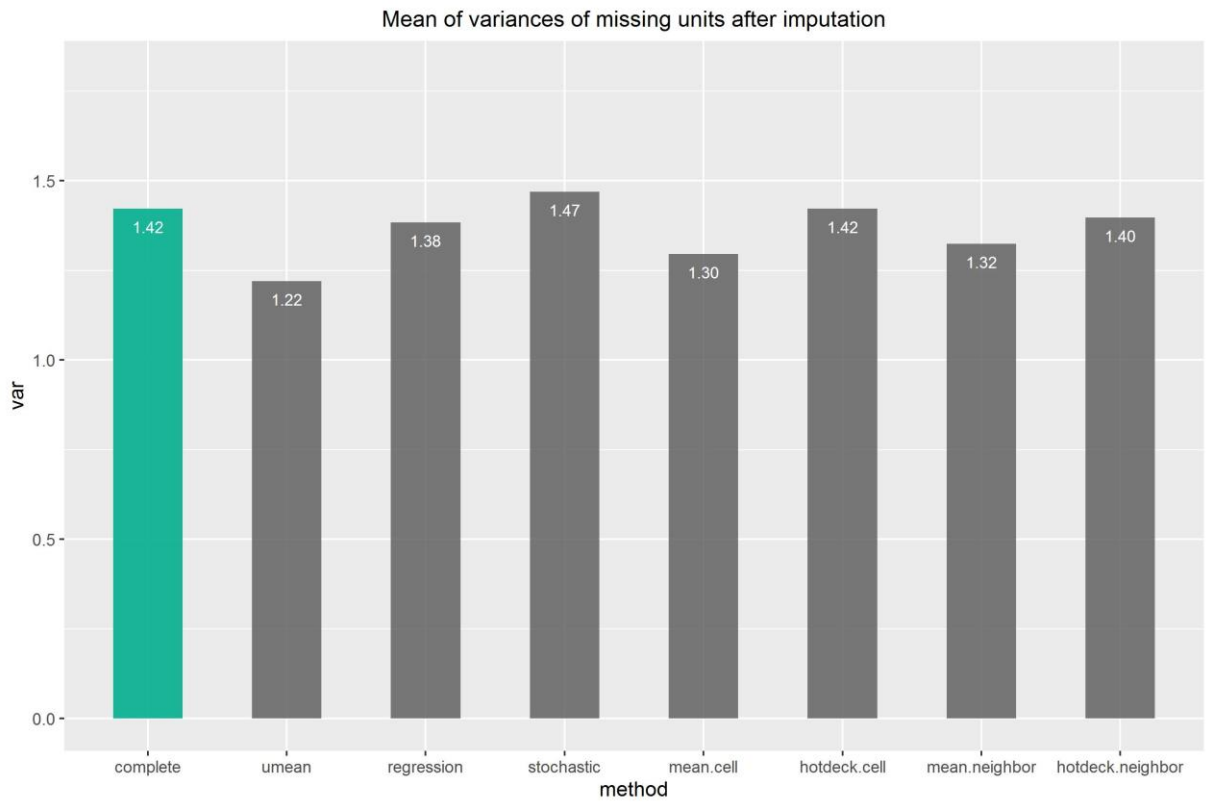


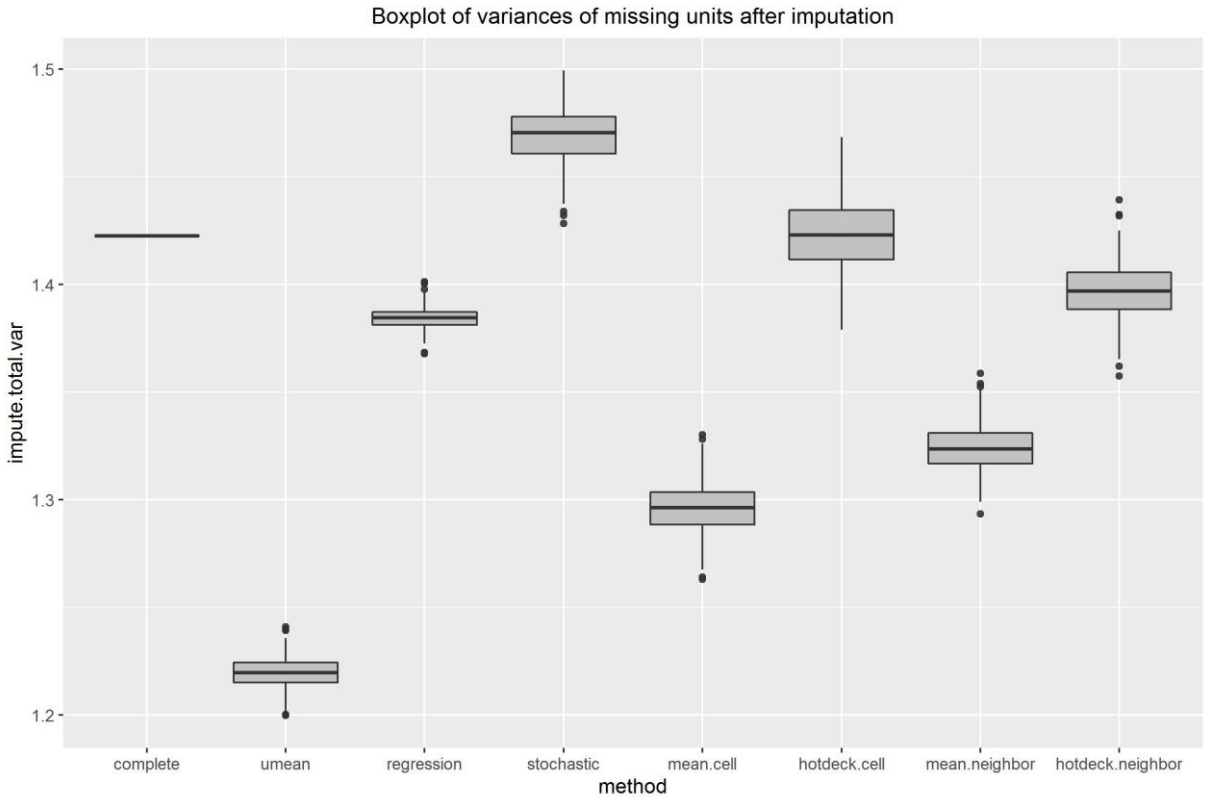


4. 결론

(2) Imputation method 비교

- 전체 Alcohol 변수의 Imputation 후 Variance들의 평균 및 boxplot

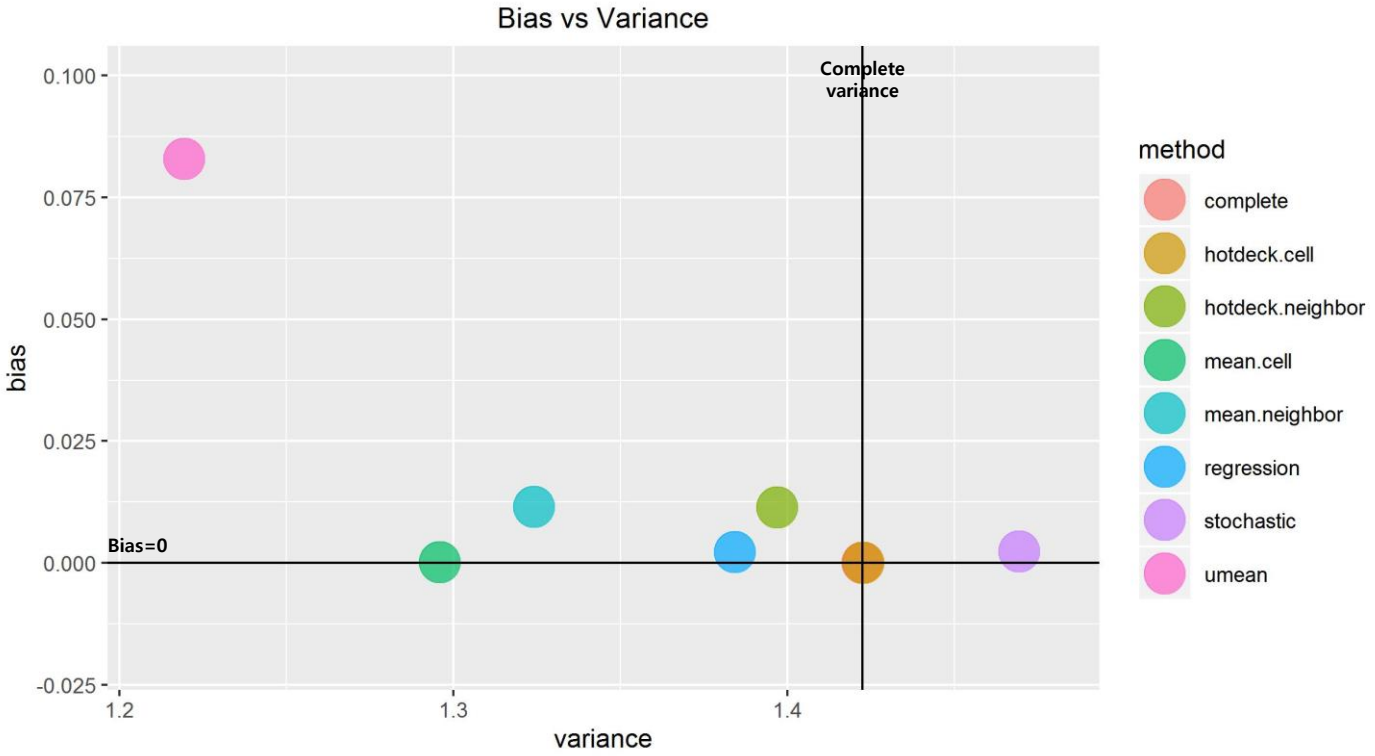




4. 결론

(2) Imputation method 비교

- 각 방법의 Bias(Mean들의 평균- 원래 평균) vs 분산의 평균



Complete(Original) 데이터와 가장 유사한 값을 대체한 방법은 **Hotdeck within imputation cell** 방법이었다. 하지만,
해당 프로젝트에서는 Imputation cell 선정 방법이 결측발생기준과 동일했으므로 가장 유사한 결과를 보였으나,
Imputation cell 선정 방법에 따라 다른 결과 가능성이 존재할 것으로 판단된다.

이를 감안할 때, **Hotdeck within nearest neighbor** 방법은
비교적 variance가 실제와 비슷하고 낮은 bias를 보이며,
결과의 변동성이 낮을 것으로 판단되기에
해당 결측 상황에서 가장 적절한 imputation 방법이라 생각한다.

4. 결론

(2) Imputation method 비교