

Imputation Class 형성 방법에 따른 Hot-deck 성능 비교

–Survey Data에의 적용–

2024.10.29

2024020409 권휘준
2024021609 정윤주
2021150470 백소윤

Contents

1. Background

문제제기

DATA 소개

DATA 전처리

2. Objective

Pipeline

Generating missing

3. Imputation

X imputation

Y imputation

4. Result

Performacne metric

Model Comparsion

& Selection

Apply to raw data

5. Conclusion

Background

“We have found that no consensus exists as to the best way to apply the hot deck and obtain inferences from the completed data set.”

Andridge, R. R., & Little, R. J. A. (2010)

핫덱을 적용하고 Complete Data에서 추론을 도출하는 최적의 방법에 대한 합의는 존재하지 않음

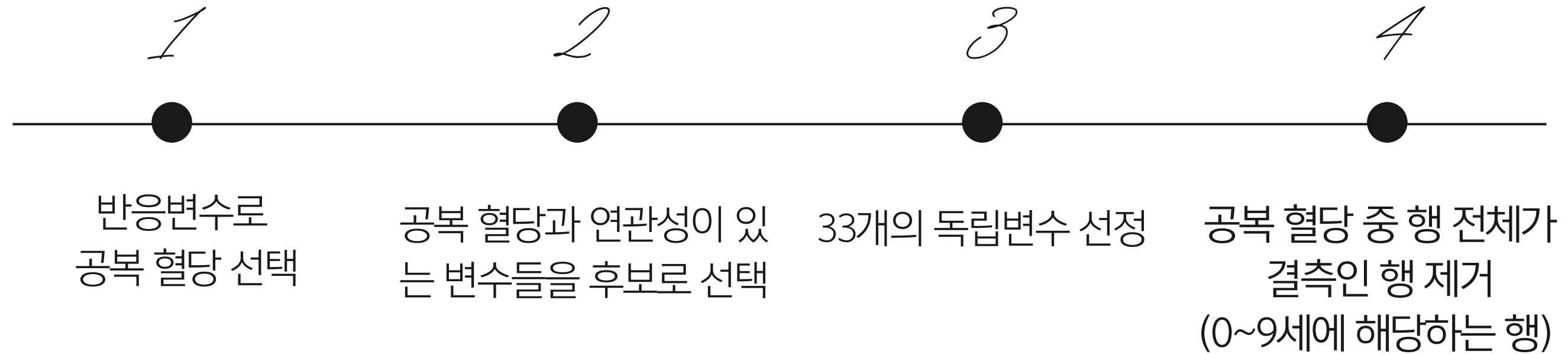


Imputation class (Donor pool) 구성 방법에 따른
Hot-deck의 성능 비교

Raw data | 2022 국민건강영양조사 원시자료 (질병관리청, 2024-05-21)

우리나라 국민 1만명에 대한 건강수준, 건강
관련 의식 및 행태, 식품 및 영양 섭취실태
조사를 통해 국가단위 통계를 산출하는 전국
규모의 조사

조사 년도 : 2022
조사 대상 : 전국 만1세 이상 약 1만명
표본 추출 : 192개 지역
조사 방식 : 순환표본조사
층화 변수 : 시도, 동읍면, 주택유형



DATA 전처리

● 1

반응변수로 공복 혈당 선택

● 2

공복 혈당과 연관성이 있는 변수들을 후보로 선택

1. 국민건강영양조사 결과발표회 기준, 당뇨 및 공복 혈당 분석에 활용된 변수로 1차 선택
2. raw data 기준 HE_glu complete과 missing인 data로 나누었을 때 histogram 분포차이가 큰 것들을 우선적으로 선택
(histogram은 appendix)
3. HE_glu 자체의 연관성이 높은 값들을 우선 선택
4. 변수간 같은 의미를 내포할 경우 하나만 선택
ex) BMI, 각 신체부위의 지방량 변수들 -> BMI 하나로 선택

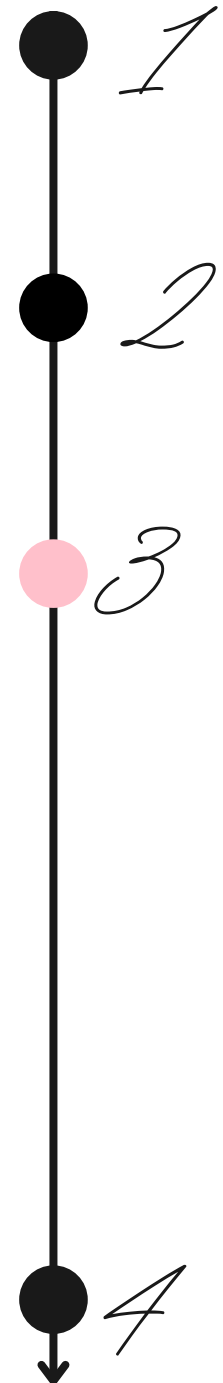
● 3

33개의 독립변수 선정

● 4
↓

공복 혈당 중 행 전체가 결측인 행 제거(0~9세에 해당하는 행)

DATA 전처리



반응변수로 공복 혈당 선택

공복 혈당과 연관성이 있는 변수들을 후보로 선택

33개의 독립변수 선정

survey data 특성을 고려하기 위해 index 및 psu, region 변수를 select
아래가 최종 선택된 결과

"HE_Ucrea" "HE_Uph" "HE_Ubld" "HE_Unitr" "HE_Uro" "HE_Upro" "DE1_dg" "DE1_ag" "HE_Usg"
"HE_Uglu" "HE_HbA1c" "HE_anem" "age" "ID" "region" "psu" "sex", "incm5" "edu"
"occp" "marri_1" "genertn" "HE_fst" "HE_mens" "HE_prg" "HE_DMfh1" "HE_DMfh2" "HE_DMfh3"
"HE_rPLS" "HE_sbp" "HE_dbp" "HE_BMI"

공복 혈당 중 행 전체가 결측인 행 제거(0~9세에 해당하는 행)

- 1 반응변수로 공복 혈당 선택
- 2 공복 혈당과 연관성이 있는 변수들을 후보로 선택
- 3 33개의 독립변수 선정

4 **공복 혈당 중 행 전체가 결측인 행 제거(0~9세에 해당하는 행)**
신체 및 혈액검사 중 나이 제한이 있는 경우가 많아, 전체적으로 0~9세 행에는 검사 관련 행들이 통으로 결측 donor를 활용할 hot-deck으로 대체할 수 없는 경우가 많으며, 최종적 비교도 어려울 것이라 판단 해당 연령대 데이터는 drop

HE_anem	빈혈 유병여부	0. 없음
		1. 있음
	【분자】 헤모글로빈(g/dL)이 10~11세 11.5미만, 12~14세 12미만, 15세이상 비임신 여성 12미만, 임신여성 11미만, 남성 13미만에 해당하는 사람 수	
	【분모】 10세이상 대상자 수	

DATA 전처리

데이터 예시

Background

반응변수 (Y)	독립변수 (X1, ..., X33)				
공복 혈당 (HE_glu)	성별 (sex)	나이 (age)	요크레아티닌 (HE_Ucrea)	...	BMI (HE_BMI)
94	2	56	84.6	...	26.507
84	1	30	54.3	...	27.152
87	2	25	192.4	...	21.308
...	

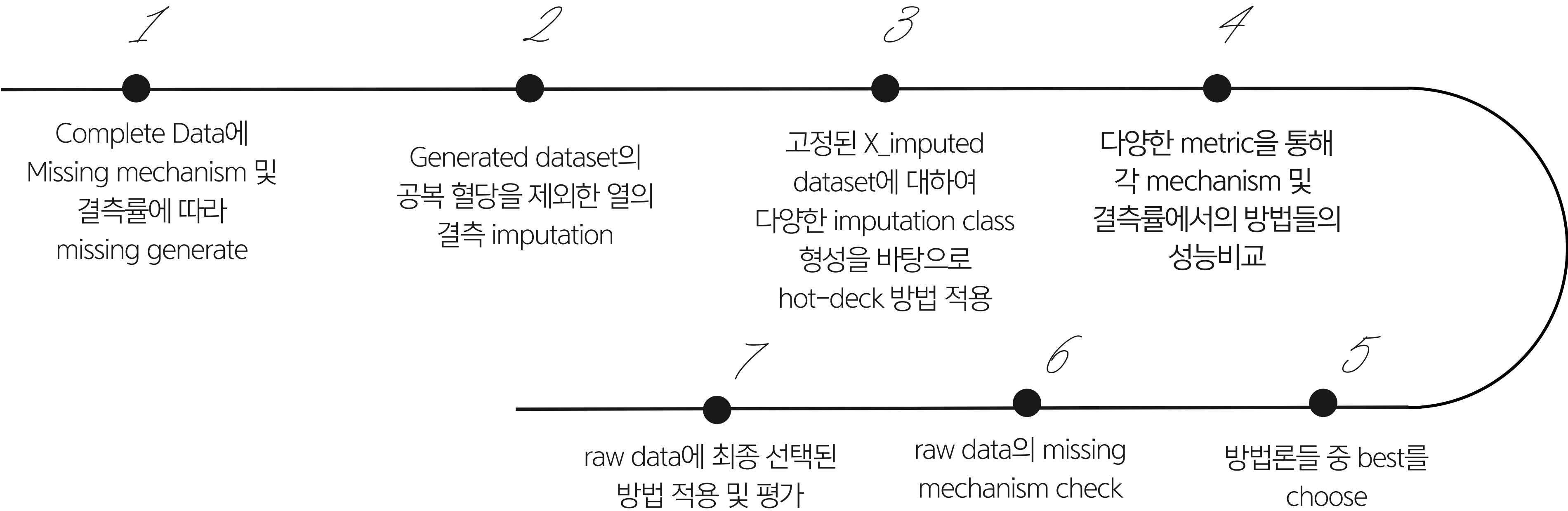
반응변수 : 공복 혈당 (HE_glu)

- 독립변수 : 33개의 변수
- 인적 사항(성별, 나이)
 - 혈액 검사 결과 중 일부
 - 신체 검사 결과 중 일부

Objective

Pipeline

Objective



Generating Missing

결측 매커니즘과 결측 비율에 따라 결측 데이터 생성

- 결측 매커니즘 : MCAR, MAR, MNAR
- 결측 비율 : 10%, 40%

MCAR 가정

- sample() function을 활용한 random sampling

MAR 가정


- 모든 변수를 고려할 수 없음, 사전지식과 연관성 높은 변수 : “age”, “DE1_ag”, “incm5”, “HE_anem” 선택
- raw data에서 logistic regression 결과를 반영해서 propensity model형성
- propensity의 quantile을 기준으로 class를 나눠 가중치 다르게 missing index 선택

MNAR 가정

- HE_glu의 값이 클수록 결측되는 비율이 높아지도록, HE_glu를 quantile로 나눠 각각 1,2,4,8 가중치를 가지고 sampling

Imputation

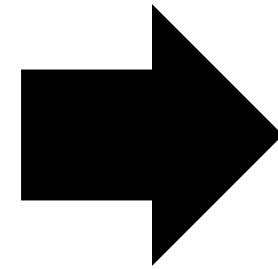
X imputation for Complete Data

- 
- 1 NA 비율이 높고, 연관성이 높은 변수들을 파악하기 위해 column별 결측비율 및 HE_glu와의 연관성을 파악
관련 자료 : Appendix참고
 - 2 각 변수를 3가지 group으로 나눠 imputation 방법 적용
 - NA 비율이 낮고 HE_glu와도 연관성이 낮은 변수군
 - HE_Ucrea, HE_Uph, HE_Ubld, HE_Unitr, HE_Uro, HE_Upro, HE_Usg, HE_rPLS : Hot deck imputation
 - NA 비율은 높으나 HE_glu와 연관성이 낮은 변수군
 - HE_DMfh3, HE_DMfh2, DE_DMfh1 : 가족력을 나타냄, age, sex, region을 class로 Hot deck imputation
 - occp, edu : age, incm5를 class로 hot-deck
 - NA 비율이 높고 HE_glu와의 연관성도 높은 변수군 (혹은 HE_glu와의 연관성이 매우 높은 변수)
 - DE1_ag, HE_BMI, HE_HbA1c, HE_Uglu, HE_sbp, HE_dbp : pmm을 metric으로 설정해 hot deck imputation

Y imputation

Imputation class 형성

1. Adjustment Cell
2. Propensity-Based Stratification
3. Predictive Mean Matching (PMM)
4. Tree – based (Random forest)
5. Clustering – based (K-means, PAM)



**Multiple Hot-deck
Imputation (D=5) 진행**

Y imputation

– Adjustment Cell

- 가장 기본적인 Hot-deck Imputation
- 보조 변수(auxiliary variables) 를 사용하여 응답자와 비응답자를 여러 개의 대체군(imputation classes, adjustment cells)으로 분류

STEP 1 : 범주형 변수를 기반으로, 연속형 변수는 구간을 나눠 범주형 변수로 변환하여 대체군 형성

STEP 2 : 반응변수와의 상관성과 반응변수의 결측여부의 상관성(*)을 모두 고려하여 대체군 형성 변수 선택

STEP 3 : 동일한 대체 셀 내에서 Hot-deck Imputation 수행

* 송주원 (2011) 핫덱대체의 대체군 형성 변수 선택, 한국자료분석학회

Y imputation

– Propensity-Based Stratification

목적: Hot Deck Imputation을 넘어, 성향점수(Propensity)를 활용하여 응답여부에 따른 대체군 간의 불균형을 처리

***성향 점수(Propensity Score):** 개체가 특정 처리(-> 응답)를 받을 확률을 관찰된 변수들에 기반해 추정

STEP 1 : 응답 상황 점수 추정

주어진 보조 변수로 로지스틱 회귀 모델을 사용하여 각 개체의 응답 성향 점수 추정

STEP 2 : 층화

추정된 응답 성향 점수를 기반으로 개체를 여러 개의 층으로 나눔 (3개 층)

STEP 3 : 층 내 분석

각 층 내에서 Hot-deck Imputation 수행

Y imputation

– Predictive Mean Matching (PMM) / RandomForest

PMM

```
impute_pmm <- function(data, target, m = 5) {
  # mice를 이용하여 imputation 수행
  imputed_data <- mice(data, m = m, method = 'pmm', maxit = 5, printFlag = FALSE)

  # 대체 값을 저장할 데이터프레임 생성
  imputed_values_df <- data.frame(matrix(ncol = m, nrow = nrow(data)))
  colnames(imputed_values_df) <- paste0(target, "_imputed_", 1:m)

  for (i in 1:m) {
    complete_data <- complete(imputed_data, i)
    # HE_glu 대체 값 저장
    imputed_values_df[, i] <- ifelse(is.na(data[[target]]), complete_data[[target]], data[[target]])
  }

  # HE_glu 대체 값만 포함된 데이터프레임 반환
  return(imputed_values_df)
}
```

1. 변수가 많다고 판단

2. class 개수가 많아지면 donor가 적어지는 문제가 발생 가능성 높음

=> 해당 문제를 보완하는 방법이 PMM

RandomForest

```
rf_impute <- function(data, target, n, m) {
  # 결측치의 인덱스 추출
  na_indices <- which(is.na(data[[target]]))

  # 대체 값을 저장할 매트릭스 생성
  imputed_values_matrix <- matrix(NA, nrow = nrow(data), ncol = m)

  for (i in 1:m) {
    # mice.impute.rf로 결측치 대체
    y <- data[[target]]
    ry <- !is.na(y)
    x <- data[, !(names(data) %in% target)]

    imputed_values <- mice.impute.rf(y = y, ry = ry, x = x, ntree = n)

    # 대체 값을 매트릭스에 저장
    imputed_values_matrix[, i] <- ifelse(is.na(data[[target]]), imputed_values, data[[target]])
  }

  # 데이터프레임으로 반환
}
```

1. 과적합이 잘 일어나지 않음

2. 결측치나 이상치에 강하며, scaling이나 정규화 과정이 필요없음

3. 변수가 많고 다양한 scale이 있는 데이터에서 적합하다고 판단

=> 우리 데이터의 similar donor 선택에 적합하다고 판단.
(tree개수가 중요하지만, 컴퓨터 성능 이슈로 n=10(default)만 활용)

Y imputation

– Clustering based (K-means / PAM)

STEP 1 : clustering 진행을 위한 데이터의 차원 축소

- y값과 상관계수 절대값 0.1인 변수들만 select 후 진행

STEP 2 : 하이퍼 파라미터 K 설정

- 이상치에 robust한 ‘Dunn index’ 기준 최적의 파라미터 값 $\Rightarrow K=5$
 - 해당 파라미터는 얼마나 clustering 군집화가 잘되는가 기준
- Donor pool의 절대적인 ‘수’에 따른 Hot-deck 결과를 비교하기 위해 (K=9, 13) 함께 확인

STEP 3 : K-means (Euclidean)과 PAM (Gower) 모델로 K = 5, 9, 13에 대해 clustering

- 유클리디안 거리에는 범주형 변수를 고려할 수 없음
- 설명변수에 범주형 변수 비율이 높으므로, 범주형 변수까지 고려할 수 있는 Gower distance를 사용
- 이상치에 robust한 PAM 모델을 사용하여 함께 확인

Results

Performance Metrics

Results

Bias

참 평균값과 결측 대체 후 추정된
평균값의 차이

RMSE

대체값의 평균제곱근오차

**Total
Variacne**

결측된 데이터의 총 분산

FMI

결측 대체 과정에서 사용된 데이터의
총 분산 중 결측으로 인한 분산 비율
→ 작을수록 good

Performance Metrics

Results

Methods	Imputation Cell Variables	Number of Cells
Adjustment cells	HE_HbA1c, age, HE_sbp	12
Propensity cells	HE_HbA1c, age, HE_sbp, Propensity Score	14
Predictive mean cells	All variables are used	–
Random Forest	All variables are used	–
K-means PAM	All variables correlation > 0.2	K = 5, 9, 13

Performance Comparsion

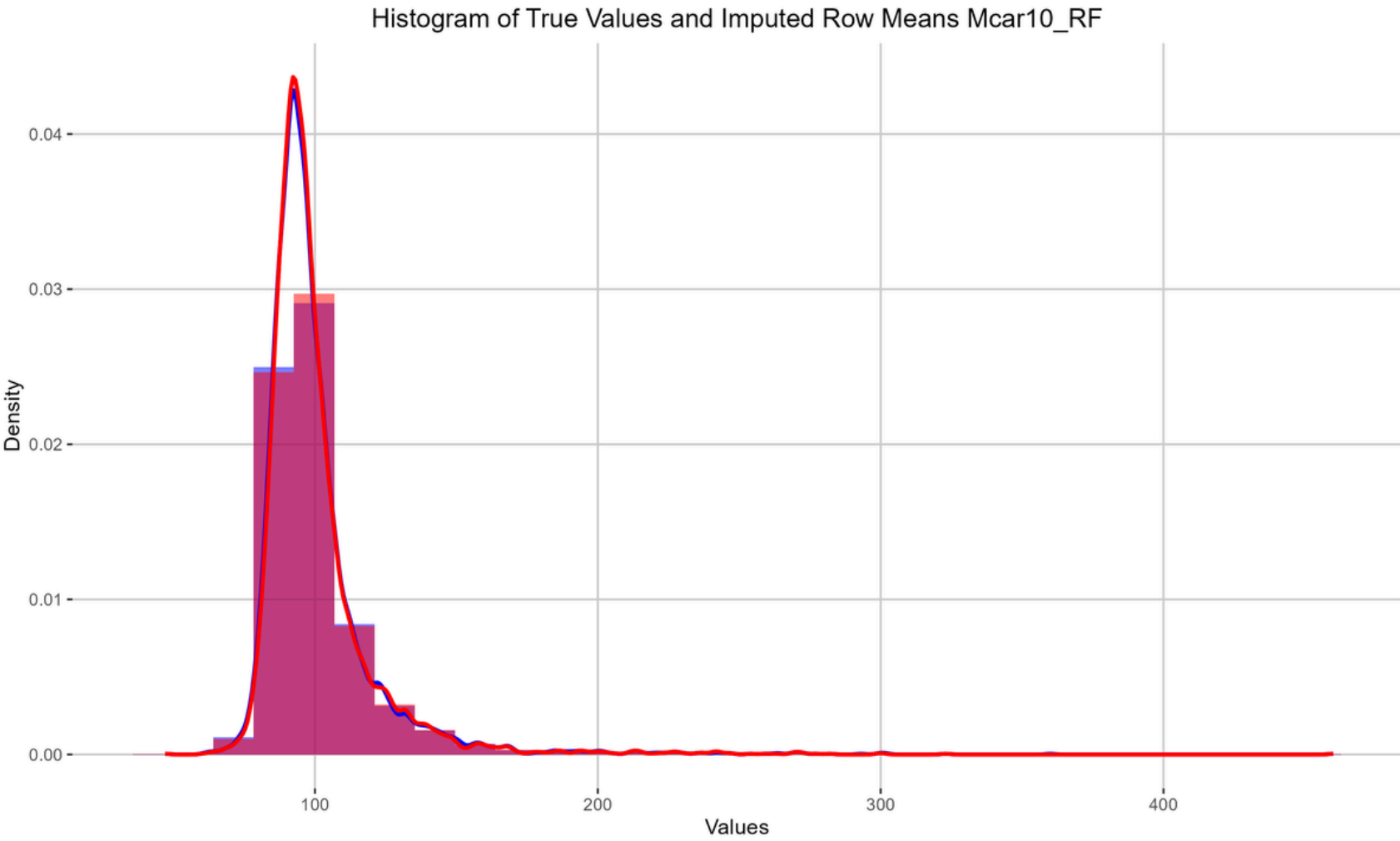
– MCAR10

*true(complete) data’s variance : 472.022

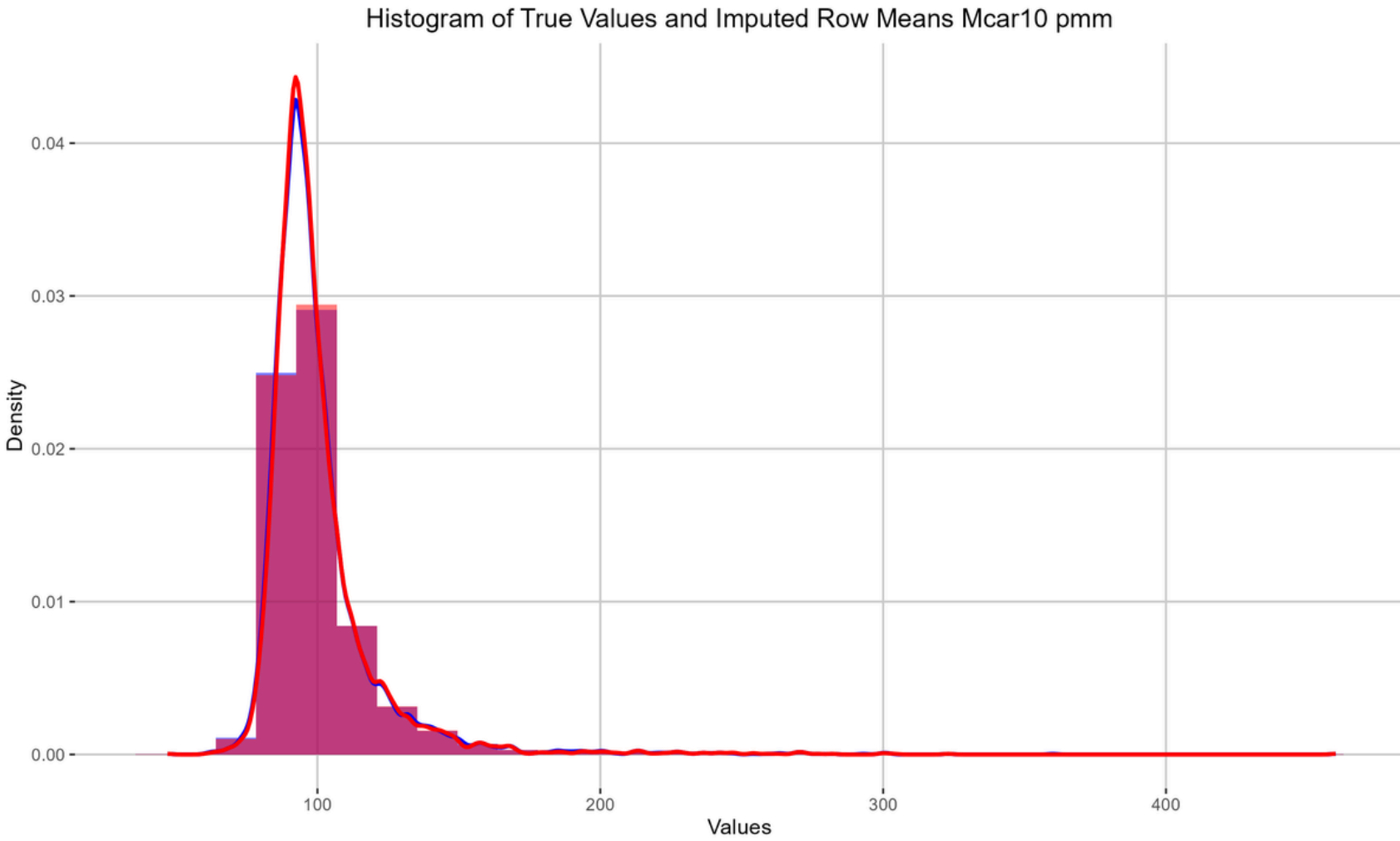
Method	parameter	Bias	RMSE	Variance	FMI
Adjustment cell		0.05	31.37605	457.8337	1.316929e-05
Propensity		0.09834532	28.68899	449.1145	2.860802e-05
PMM		0.084856115	20.34569	455.6508	1.146671e-05
Random forest		0.097769784	33.49781	453.8444	1.825363e-05
K-means	k = 5	0.1313669065	29.05445	458.5510	2.353602e-05
	k = 9	0.0648561151	26.52297	458.9373	1.545308e-05
	k = 13	0.1113669065	24.79078	454.5134	5.707954e-06
PAM	k = 5	0.085071942	32.57473	456.2318	2.580617e-05
	k = 9	0.104136691	29.46416	460.8521	1.037027e-05
	k = 13	0.149244604	29.04573	449.5111	3.535441e-05

Performance Comparision

– MCAR10



Random Forest



PMM

Performance Comparsion

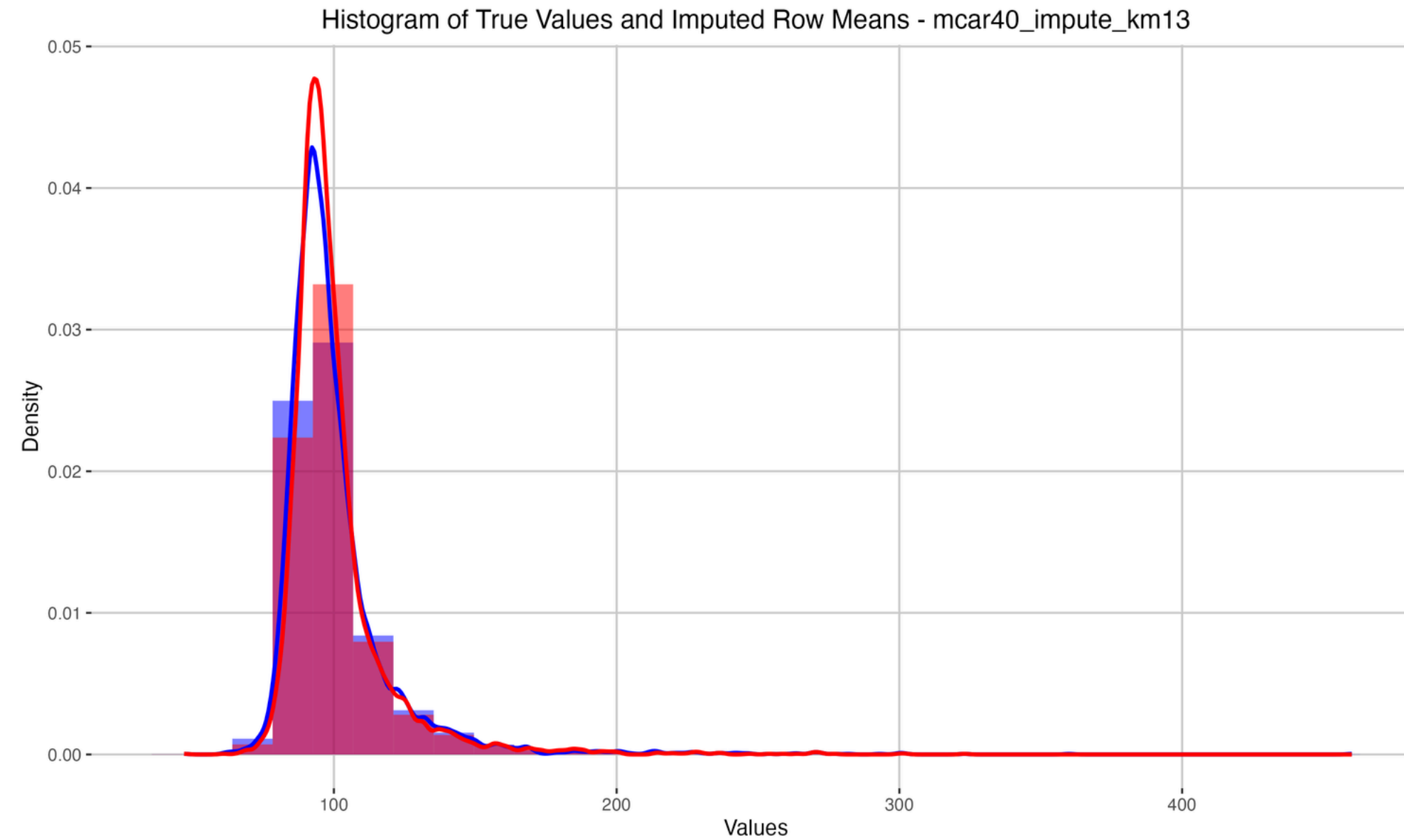
– MCAR40

*true(complete) data’s variance : 472.022

Method	parameter	Bias	RMSE	Variance	FMI
Adjustment cell		0.05935252	28.19844	460.7285	1.158227e-04
Propensity		0.17456835	26.87162	431.4044	1.136150e-04
PMM		0.03165468	19.01169	446.3843	8.215084e-06
Random forest		0.22384892	29.60871	435.8378	1.359673e-04
K-means	k = 5	-0.063669065	25.88204	462.4396	2.525040e-05
	k = 9	0.040863309	23.69414	445.5957	2.508552e-05
	k = 13	-0.008309353	24.35698	457.3773	7.796988e-05
PAM	k = 5	-0.07510791	30.24407	463.1853	1.704827e-04
	k = 9	0.03442446	26.21323	443.9528	1.731442e-04
	k = 13	-0.08751799	26.71547	455.8272	3.467459e-05

Performance Comparison

– MCAR40



K-means (K = 13)

Performance Comparsion

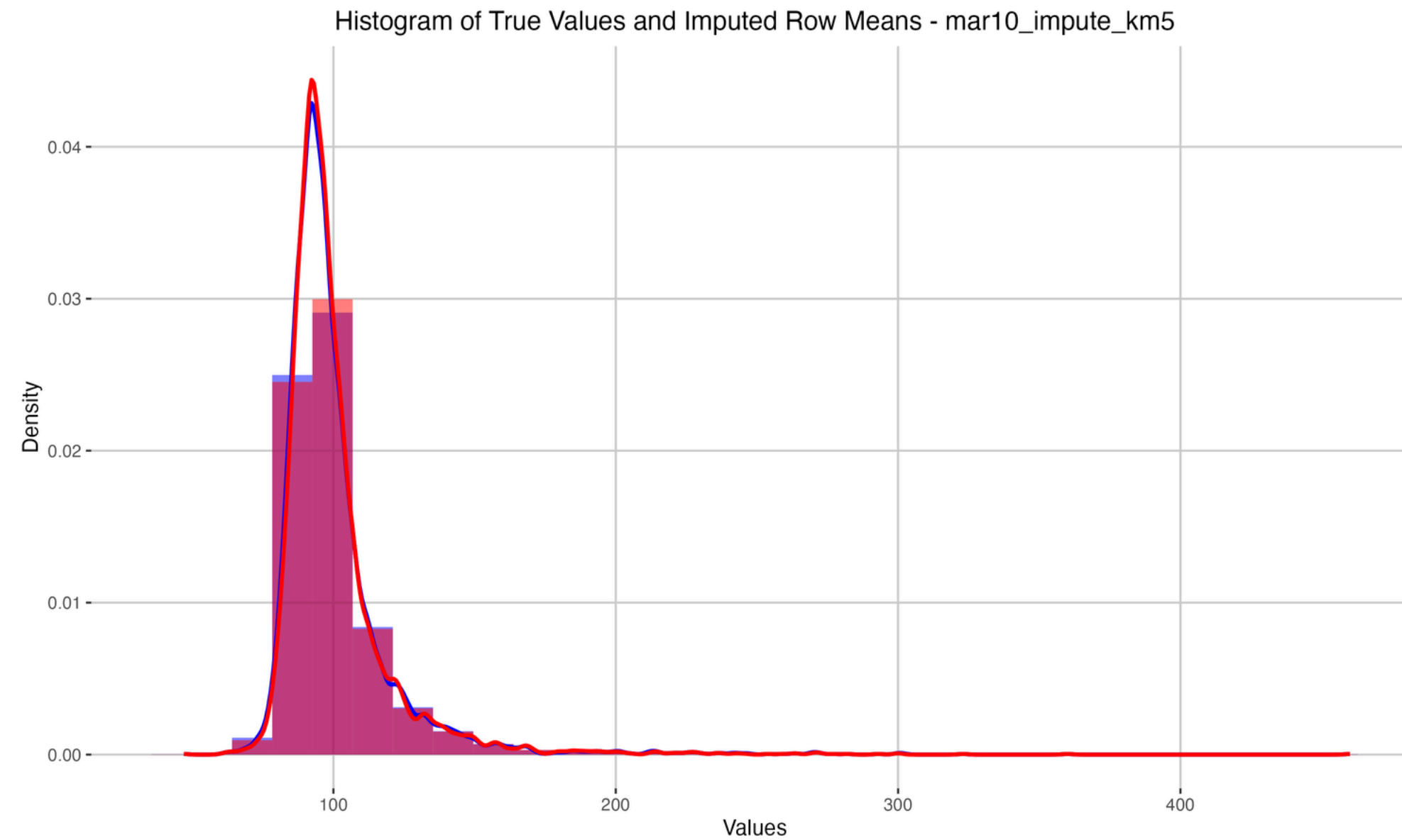
– MAR10

*true(complete) data’s variance : 472.022

Method	parameter	Bias	RMSE	Variance	FMI
Adjustment cell		0.18032374	28.29717	457.4687	2.835152e-05
Propensity		0.19953237	27.16735	450.6193	8.350334e-06
PMM		-0.020035971	20.14545	462.1551	2.423102e-06
Random forest		0.005539568	31.48486	464.6991	2.198620e-05
K-means	k = 5	0.0008273381	27.40626	476.9160	1.296829e-05
	k = 9	-0.0331294964	26.78425	478.0871	7.287190e-06
	k = 13	-0.0015827338	25.55444	473.2848	1.451043e-05
PAM	k = 5	0.176043165	30.69765	459.8989	1.453071e-05
	k = 9	0.025719424	27.78536	468.0834	2.625420e-06
	k = 13	0.004892086	27.20459	469.0524	3.274535e-05

Performance Comparison

– MAR10



K-means (K = 5)

Performance Comparsion

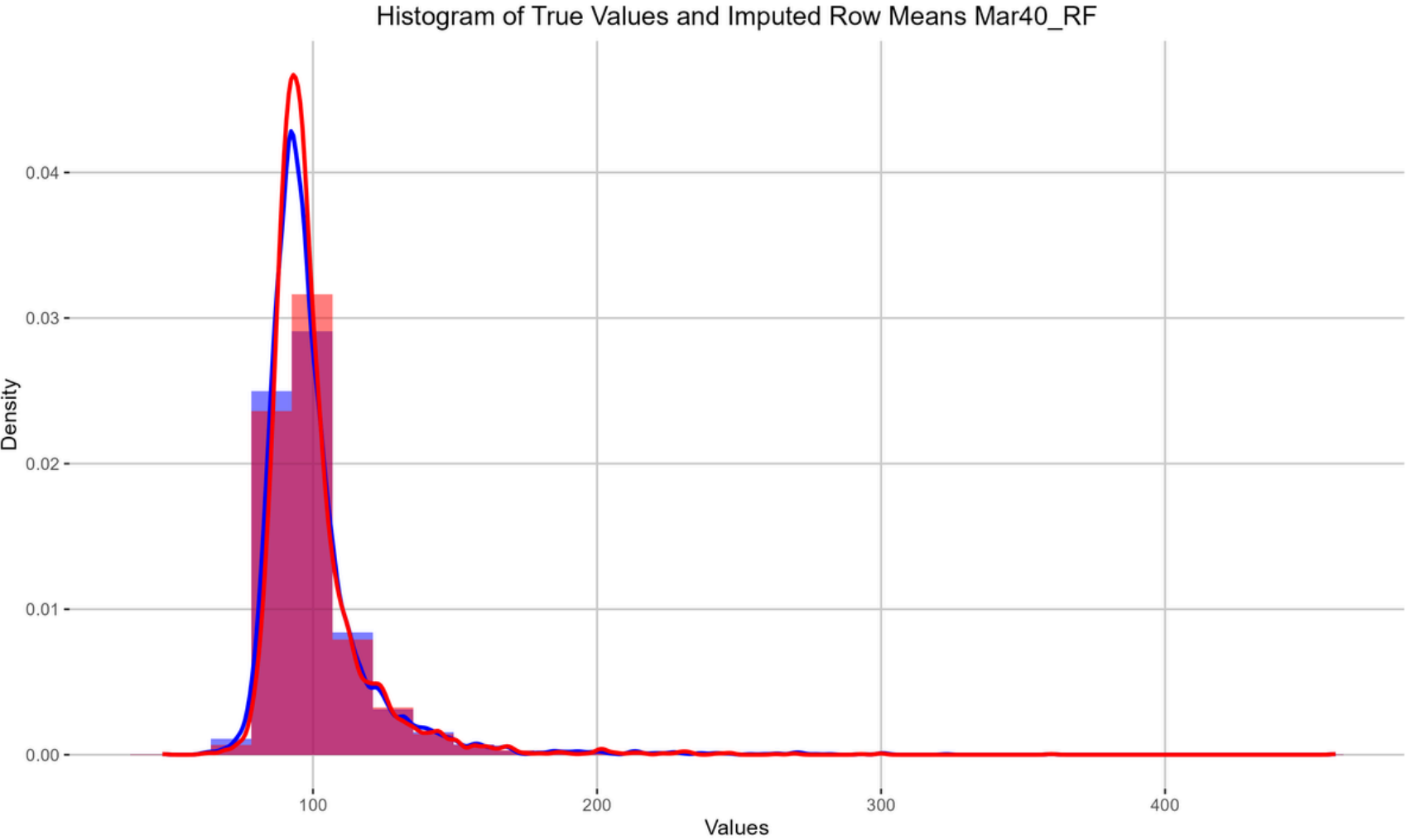
– MAR40

*true(complete) data’s variance : 472.022

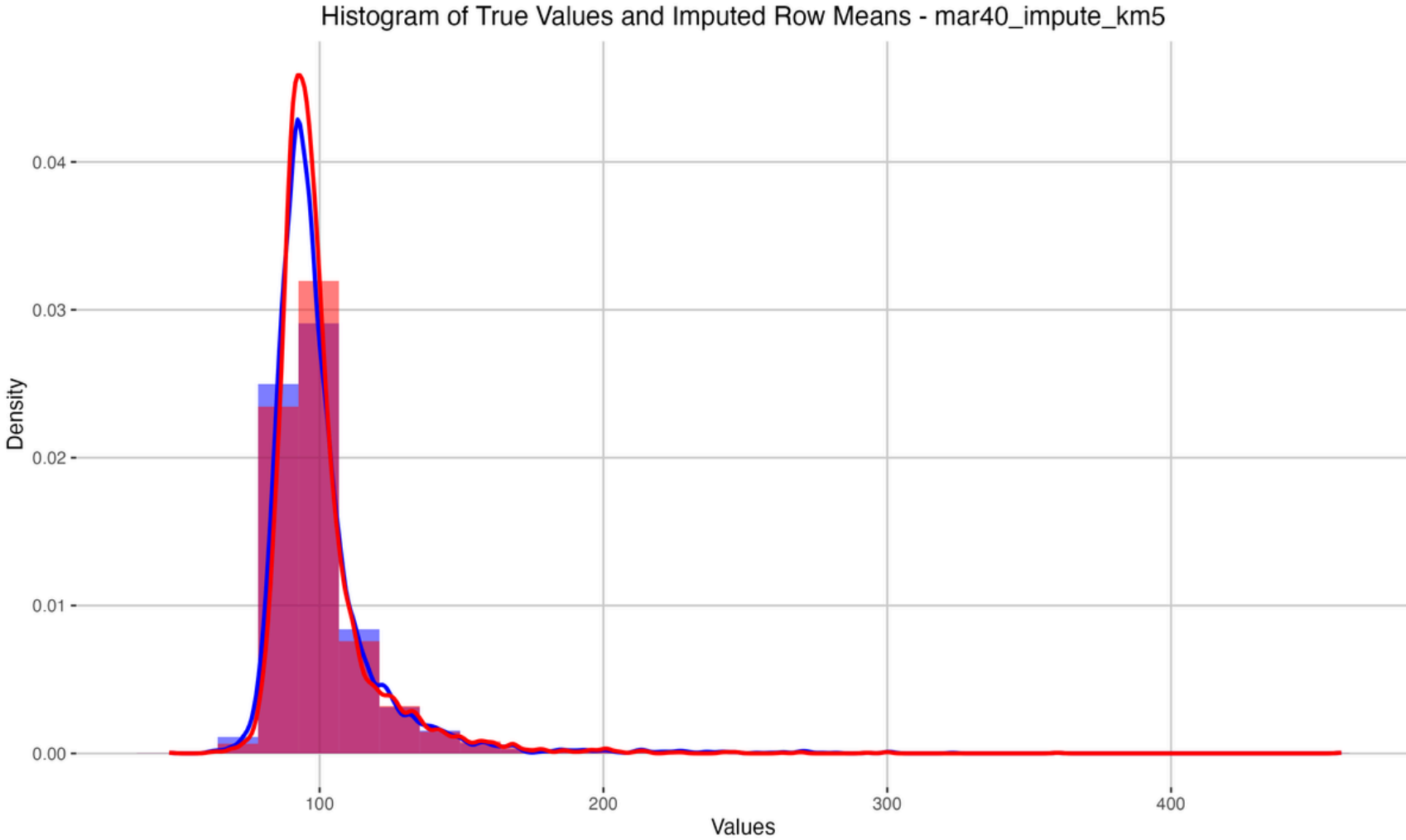
Method	parameter	Bias	RMSE	Variance	FMI
Adjustment cell		0.835575	28.03473	421.4917	1.095848e-05
Propensity		0.46208633	25.29535	470.7456	1.750341e-04
PMM		-0.17582734	20.29983	513.3139	2.540314e-05
Random forest		0.08133094	33.68006	487.0741	1.099072e-04
K-means	k = 5	-0.003201439	28.62518	528.3782	7.647621e-05
	k = 9	0.227194245	27.22896	479.8190	1.094357e-04
	k = 13	0.146618705	26.32626	495.5778	1.096851e-04
PAM	k = 5	0.78755396	30.05226	425.0273	1.330300e-04
	k = 9	0.08571942	27.64677	473.3437	2.422717e-04
	k = 13	0.21143885	27.19775	452.9332	3.456973e-05

Performance Comparsion

– MAR40



Random Forest



K-means (K = 5)

Performance Comparsion

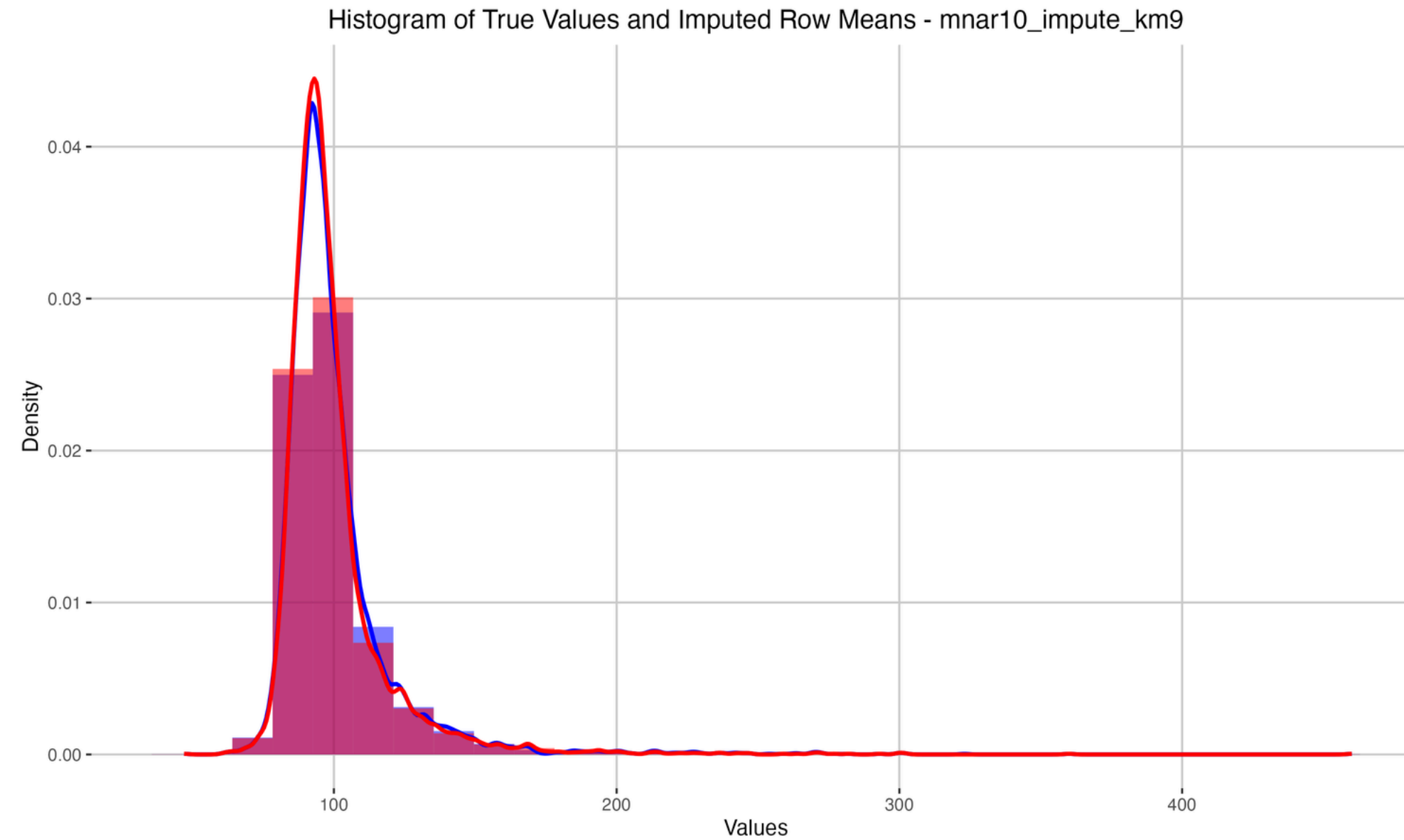
– MNAR10

*true(complete) data’s variance : 472.022

Method	parameter	Bias	RMSE	Variance	FMI
Adjustment cell		0.84809353	33.82461	443.5945	9.571428e-06
Propensity		0.92758993	30.84684	434.9595	5.062846e-05
PMM		2.706510791	46.08117	397.7060	5.941570e-05
Random forest		2.813633094	73.59988	409.4718	1.566147e-05
K-means	k = 5	0.6666187050	33.02707	468.4944	1.422460e-05
	k = 9	0.6136330935	33.89301	477.2872	4.674968e-05
	k = 13	0.6465467626	32.49579	471.9853	1.045123e-05
PAM	k = 5	1.116330935	34.67836	432.2006	6.120758e-05
	k = 9	0.764892086	32.95078	459.4443	6.308382e-05
	k = 13	0.696330935	34.80261	469.7883	3.891455e-05

Performance Comparison

– MNAR10



K-means (K = 9)

Performance Comparsion

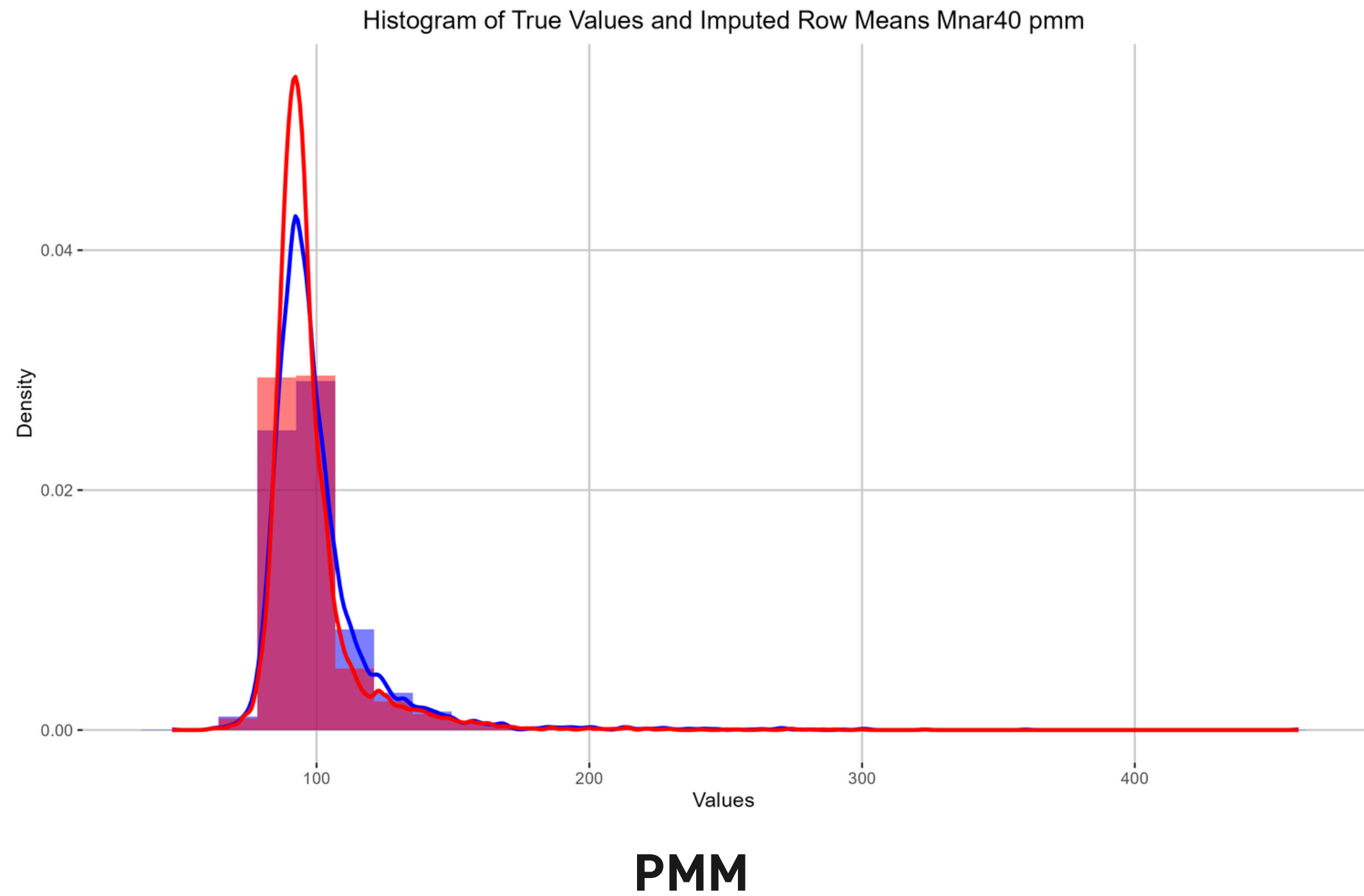
– MNAR40

*true(complete) data’s variance : 472.022

Method	parameter	Bias	RMSE	Variance	FMI
Adjustment cell		4.03805755	32.25771	296.8728	1.789921e-05
Propensity		3.21676259	27.46390	378.8853	3.585285e-04
PMM		2.64805755	23.39085	394.0622	7.611573e-05
Random forest		3.15438849	35.22856	352.3334	1.310594e-04
K-means	k = 5	3.540719424	30.74845	373.8021	1.085215e-04
	k = 9	3.545287770	29.31931	366.6103	2.512483e-04
	k = 13	3.549928058	29.92430	376.4065	1.420678e-04
PAM	k = 5	5.02384892	32.90586	249.5920	1.204730e-04
	k = 9	3.56438849	30.92888	356.0307	4.242527e-04
	k = 13	3.82701439	31.06026	322.3779	2.097425e-04

Performance Comparision

– MNAR40



Performance Comparision

-Result

결측비율 : 10%

- MCAR, MAR 가정에 비해 MNAR일 때, 어떠한 class 형성 방법을 활용하더라도, 성능이 그리 좋지 않았음
 - MNAR이 변수자체에 결측 비가 의존하는 형태이기 때문일것
- MCAR 10%인 경우 전체 방법론에 대해 전반적으로 좋은 성능
- sampling에서의 error로 인해 우리 데이터의 outlier가 결측이 된 경우가 MCAR에서가 MAR에서보다 조금 더 많았기에, MCAR임에도 bias가 조금 더 높았음
 - Appendix참고
- 특히 Adjustment Cell 적용에 있어 한계가 있었음
 - 반응변수(HE_glu)와 연관성이 높은 변수와 반응변수의 결측여부와 연관성이 높은 변수가 유사한 경우가 발생
 - 연관성이 높은 변수들을 사용하면 Class 불균형 문제가 생김

결측비율 : 40%

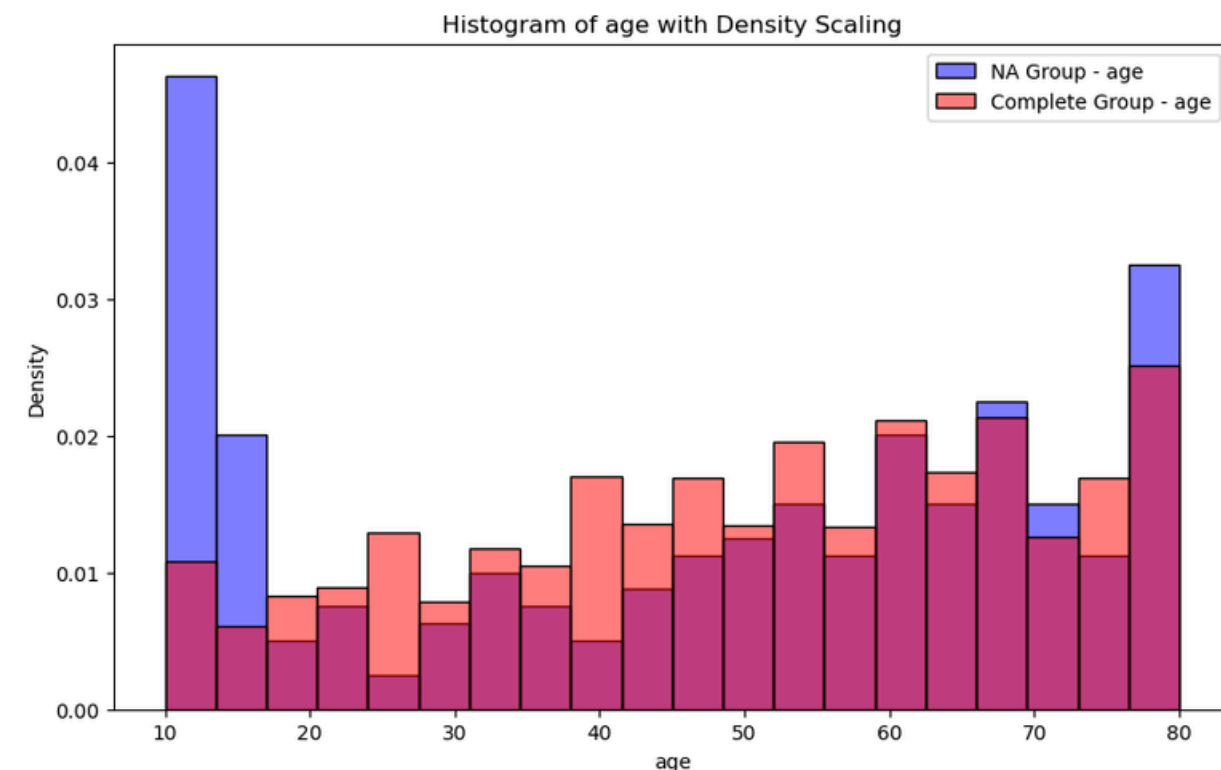
- MCAR, MAR 가정에 비해 MNAR일 때, 어떠한 class 형성 방법을 활용하더라도, 성능이 그리 좋지 않았음
 - MNAR이 변수자체에 결측 비가 의존하는 형태이기 때문일것
- 결측비율이 10%일 때보다 40%일 때, donor의 문제로 전반적인 성능에서 10%보다 낮게 나왔음
 - hot-deck method를 사용했기 때문에 결측율이 높으면 hot-deck이 적절하지 않을 수도 있음을 시사
- MNAR 40%에서는 전체 방법론에 대해 variance가 매우 낮았음
 - 과소 추정된 분산이 아닐지 의심할 필요 있음

Apply to Raw data

- 매커니즘 가정

등을 제외한 모든 가구원을 조사대상자로 선정하였다.
대상가구 및 대상자에 대한 응답률 곱의 역수를 무응답
조정가중치로 반영하여 무응답 편향을 보정하였다.

$$\text{무응답조정가중치 (응답률 역수)} = \frac{\text{조사대상 가구수}}{\text{참여 가구수}} \times \frac{\text{조사대상 가구원수}}{\text{참여 가구원수}}$$



국민건강영양조사 설명에서 무응답조정가중치를 통해 무응답을 보정하고 있음

이는 MAR 가정을 바탕으로 계산하는 과정이며,

final data의 CC와 IC 간의 차이를 histogram으로 체크했을 때, CC와 IC간 분포차이가 있었기 때문에, MAR로 판단했음.

```
> sum(is.na(final$HE_glu))
[1] 228
> length(final$HE_glu)
[1] 5788
> 228/5788
[1] 0.03939185
```

Apply to Raw data

- 적용 결과

앞선 Simulation MAR10에서 선택한 model : K-means, Random Forest, PMM

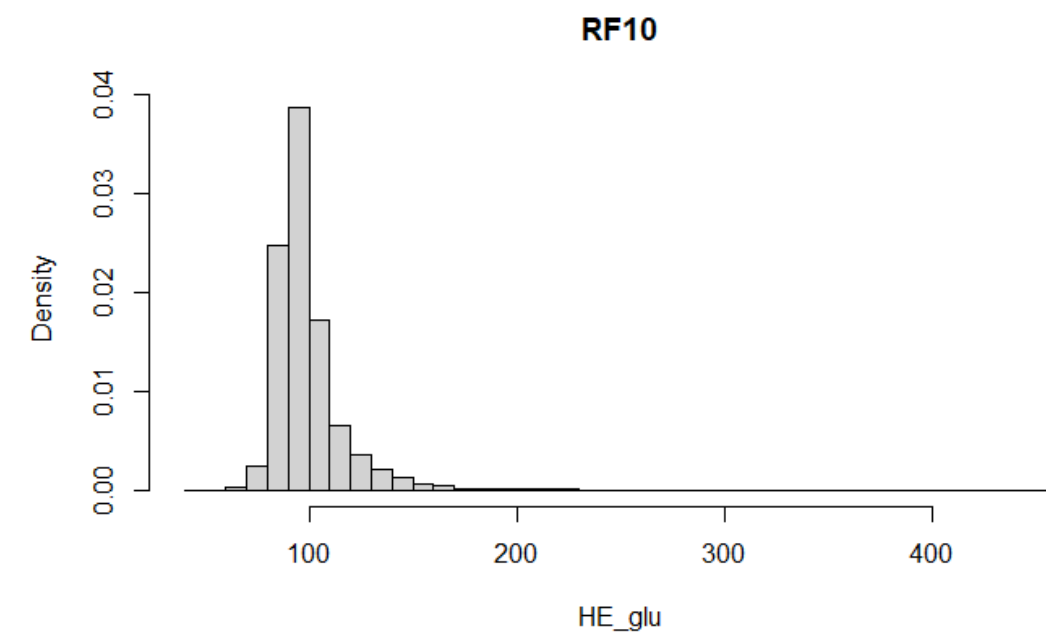
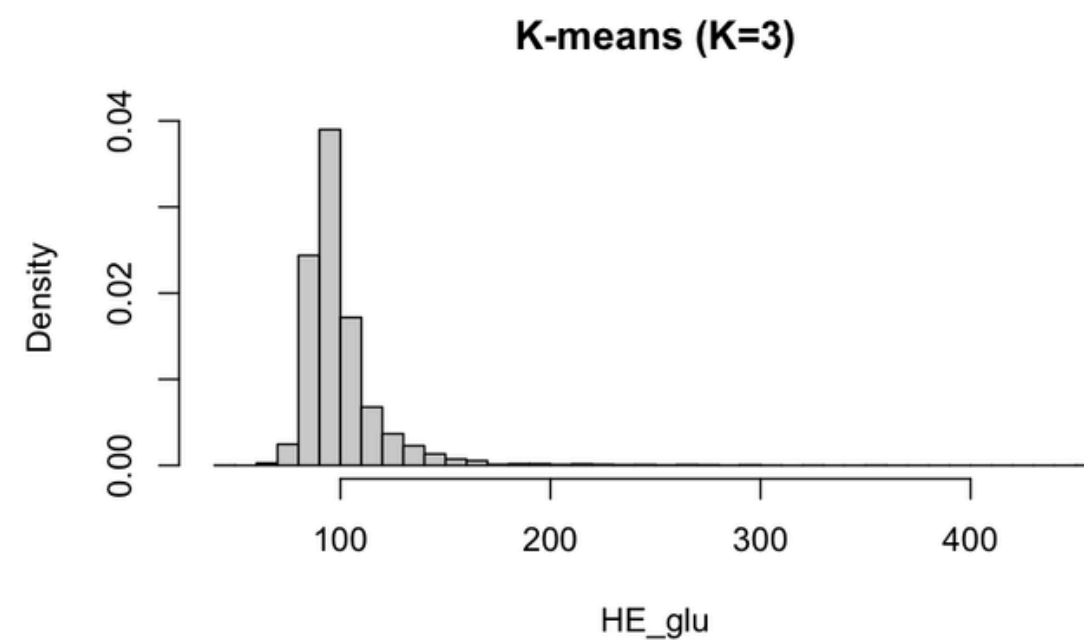
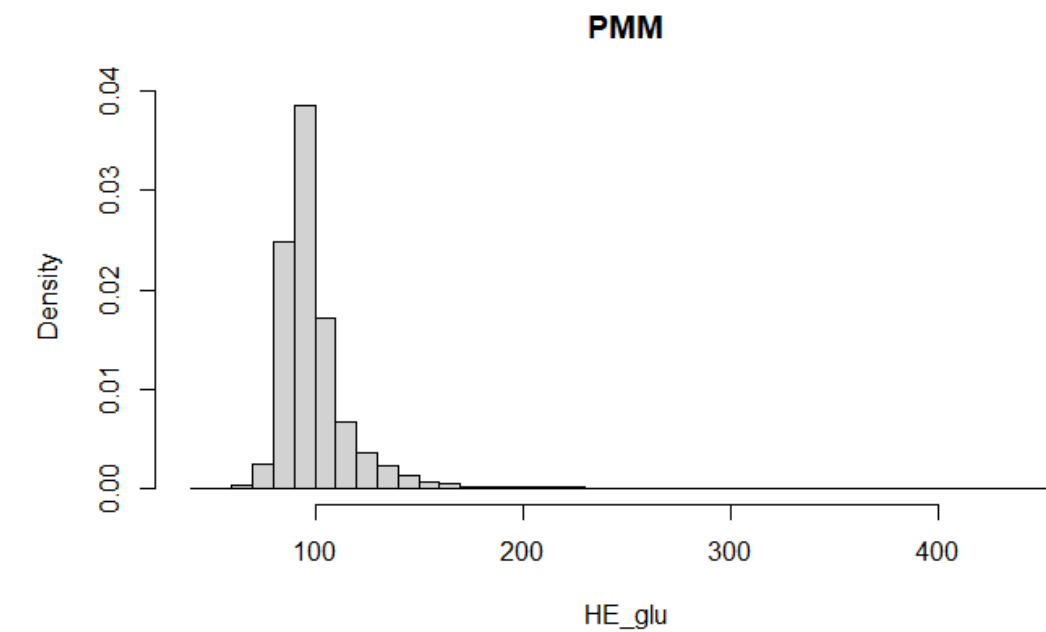
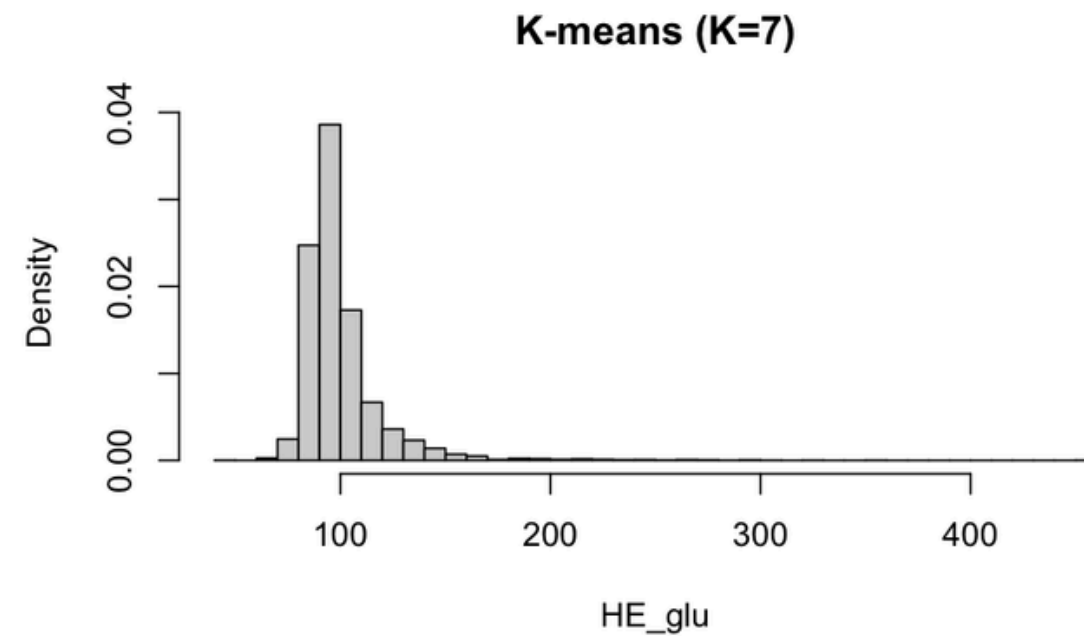
Method	parameter	Variance	FMI
PMM		479.0132	6.091135e-06
Random forest		476.9133	2.411057e-06
K-means	k = 3	476.7084	4.334942e-06
	k = 7	478.3161	2.956655e-06

Apply to Raw data

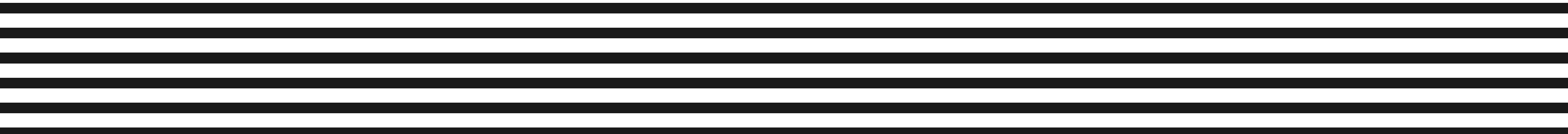
- 적용 결과

앞선 Simulation MAR10에서 선택한 model : K-means, Random Forest, PMM

*Dunn index 기준 K=3, 7 good



Conclusion



연구 의의 & 한계점

연구 의의

- 결측률이 40%인 데이터셋에서 볼 수 있듯, 결측률이 높다면 Hot-deck 방법은 적절하지 않음
- MCAR, MAR 가정에 비해 MNAR일 때, 어떠한 class 형성 방법을 활용하더라도, 성능이 그리 좋지 않았음
- data의 missing이 outlier에 많이 생성된다면, 이를 hot-deck을 통해 처리하는 데에는 어려움이 있음
- 우리 데이터에서는 MCAR10에서는 randomforest, pmm, MCAR40, MAR10에서는 k-means, MAR40은 randomforest와 kmeans, MNAR10 k-means, MNAR40은 pmm방법이 우수했다고 판단함. (bias 측면)

연구 한계점

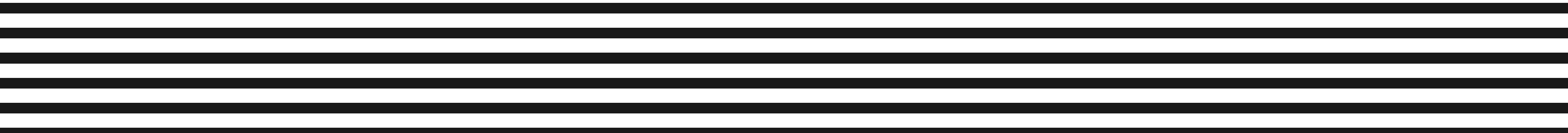
- 모델마다 변수 선택의 영향에 대한 제어
 - RF, PMM에서 variable selection을 진행하지 않았음
- RandomForest에서 tree 수를 computation cost로 인해 여러 가지 실험 불가
- skewed target distribution임에도 mean을 estimate으로 활용한 점
 - median을 시도해보았으나, 성능 지표 구성 등에서 문제 발생

Future works

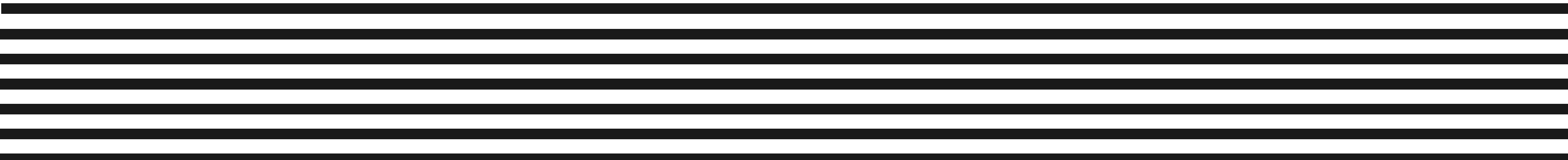
- 결측 매커니즘 데이터셋을 하나로 고정했으나, 추후 연구에서는 반복 실험
- x imputation 결과에 따른 변동이 있으므로 추후 연구에서는
 - x imputation에 대해 반복 실험
 - x imputation ordering을 여러 조합으로 실험
- Hotdeck Imputation을 Multiple Imputation만 적용해본 점
 - Bayesian Bootstrap (BB)과 Approximate Bayesian Bootstrap (ABB) 등을 통해 더 복잡한 분포에 대한 모델링

Q&A

질의 응답

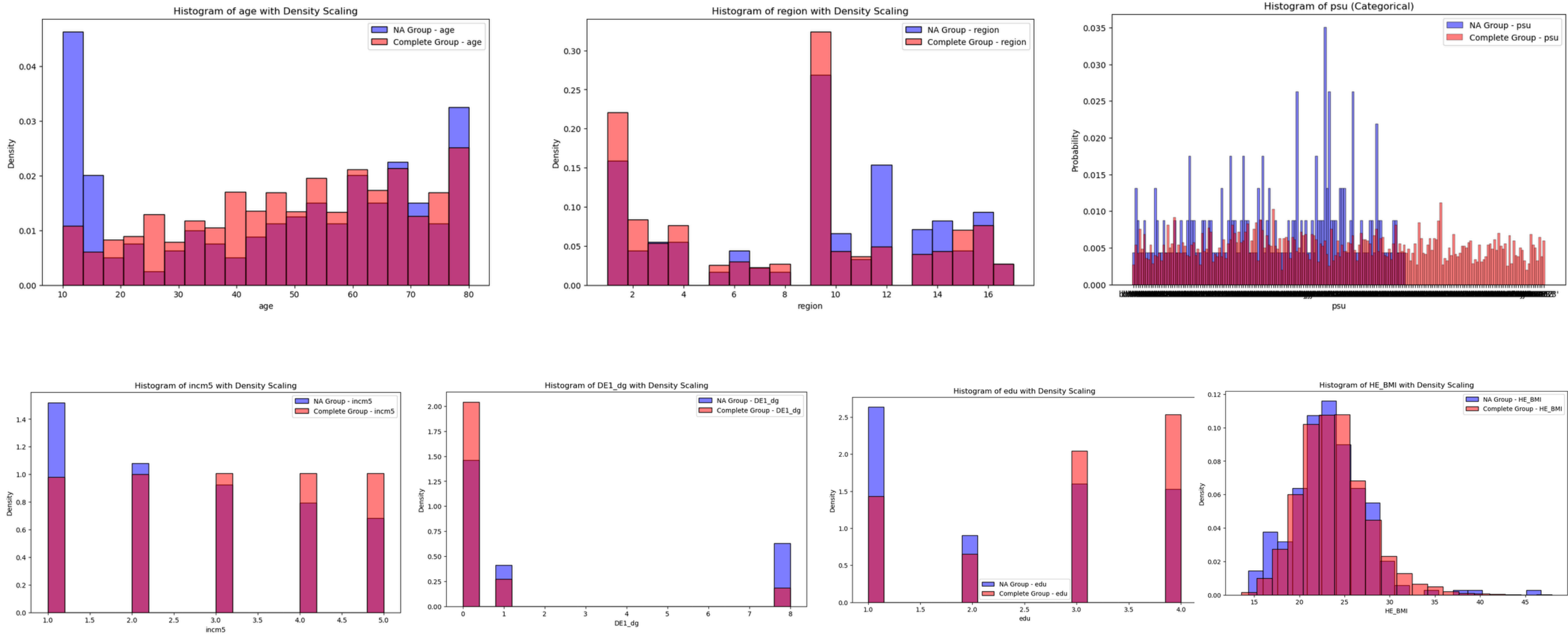


Appendix



변수 선택

Appendix



X imputation

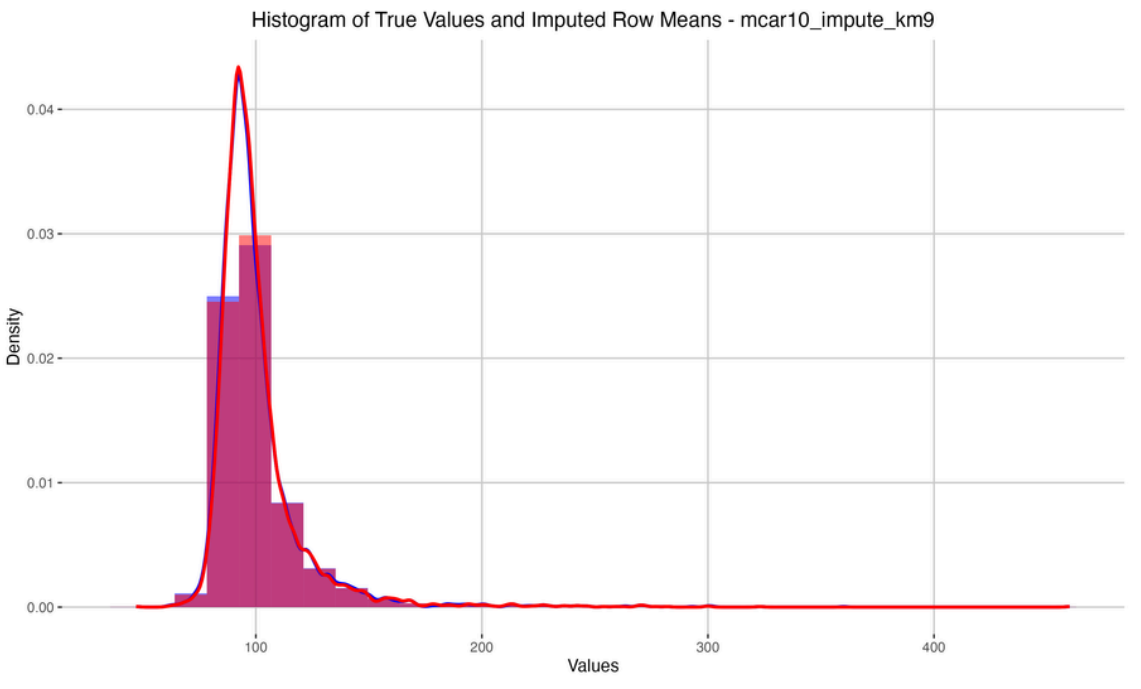
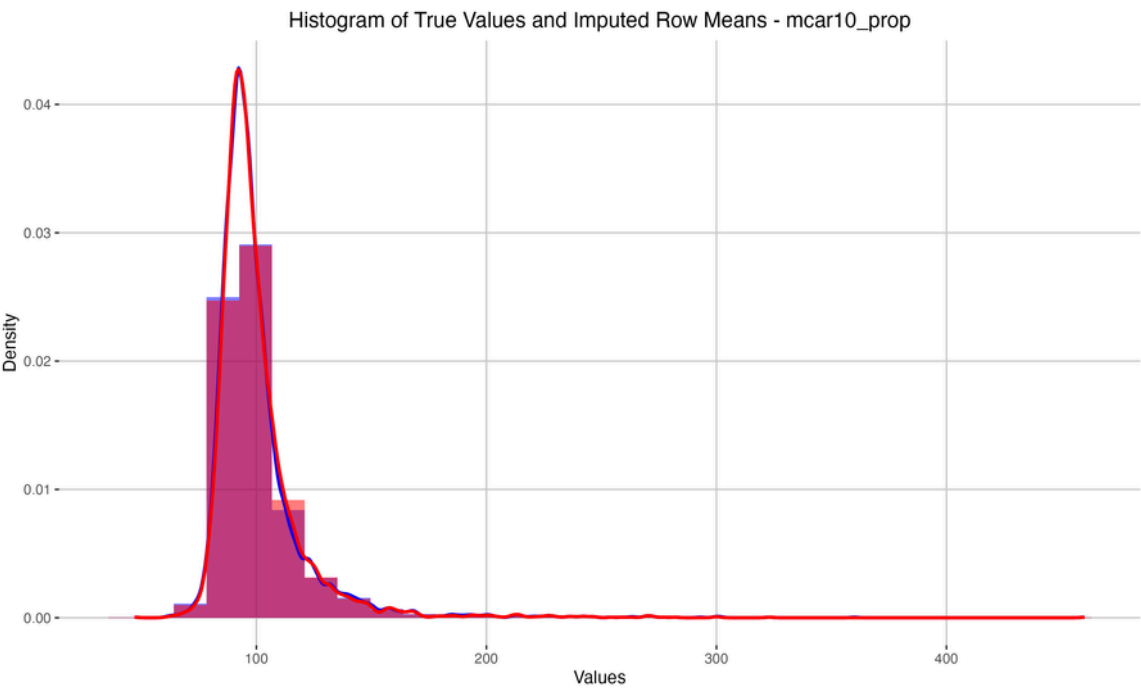
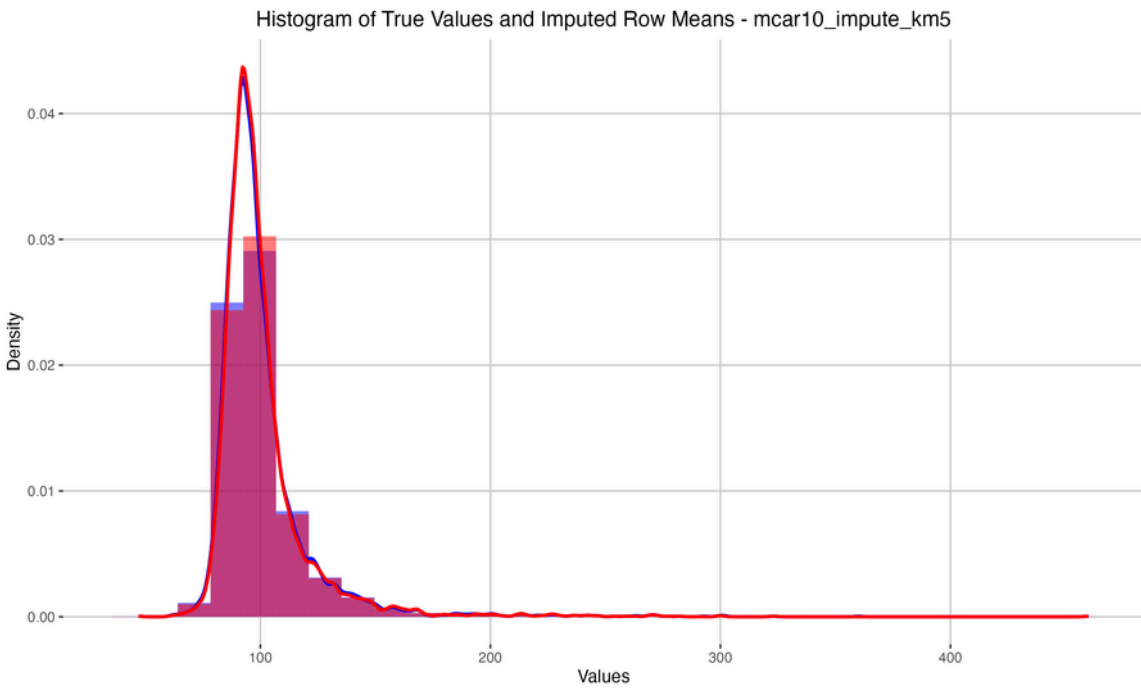
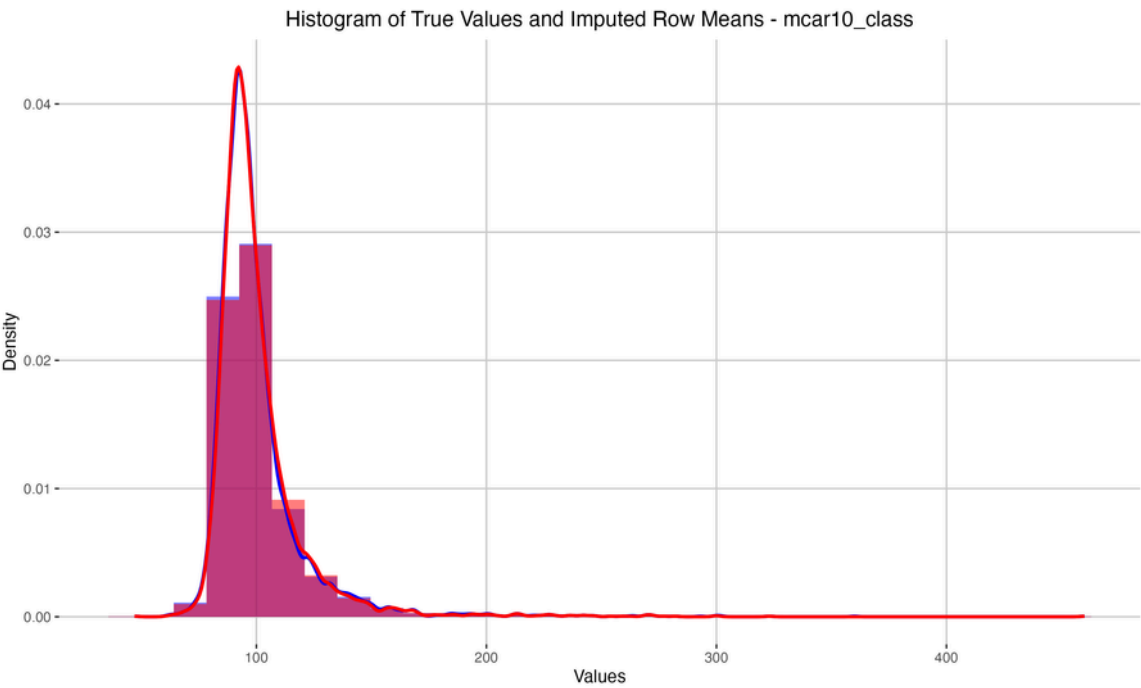
NA 개수

HE_Ucrea	HE_Uph	HE_Ubld	HE_Unitr	HE_Uro	HE_Upro	DE1_dg	DE1_ag	HE_Usg	HE_Uglu	HE_HbA1c	HE_anem	age	ID	region	psu	sex	incm5	edu	occp
131	131	131	131	131	131	0	5	131	131	15	0	0	0	0	0	0	0	187	406
marri_l	genertn	HEfst	HEmens	HEprg	HE_DMfh1	HE_DMfh2	HE_DMfh3	HE_rPLS	HE_sbp	HE_dbp	HE_BMI	HE_glu	glu	weight	m				
0	0	0	0	0	480	374	302	119	120	120	94	2224	0	0	0				

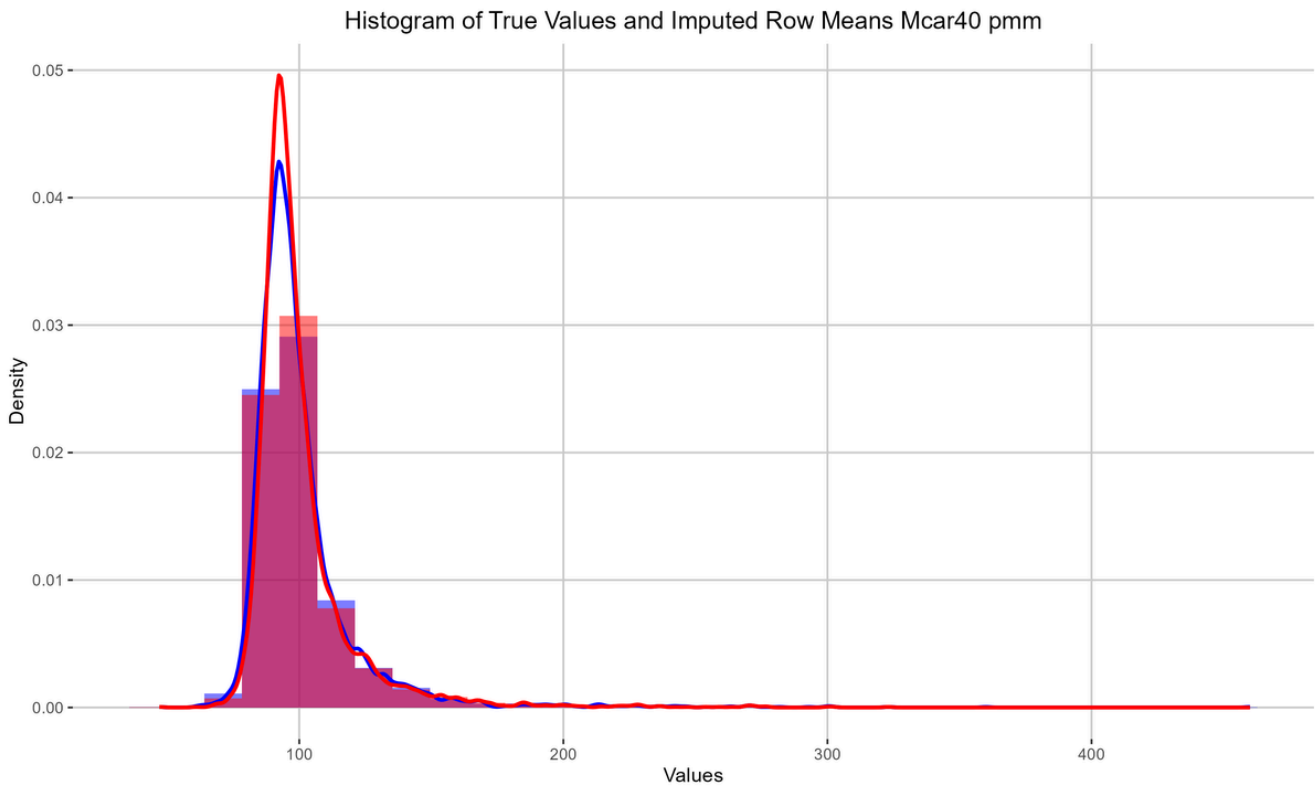
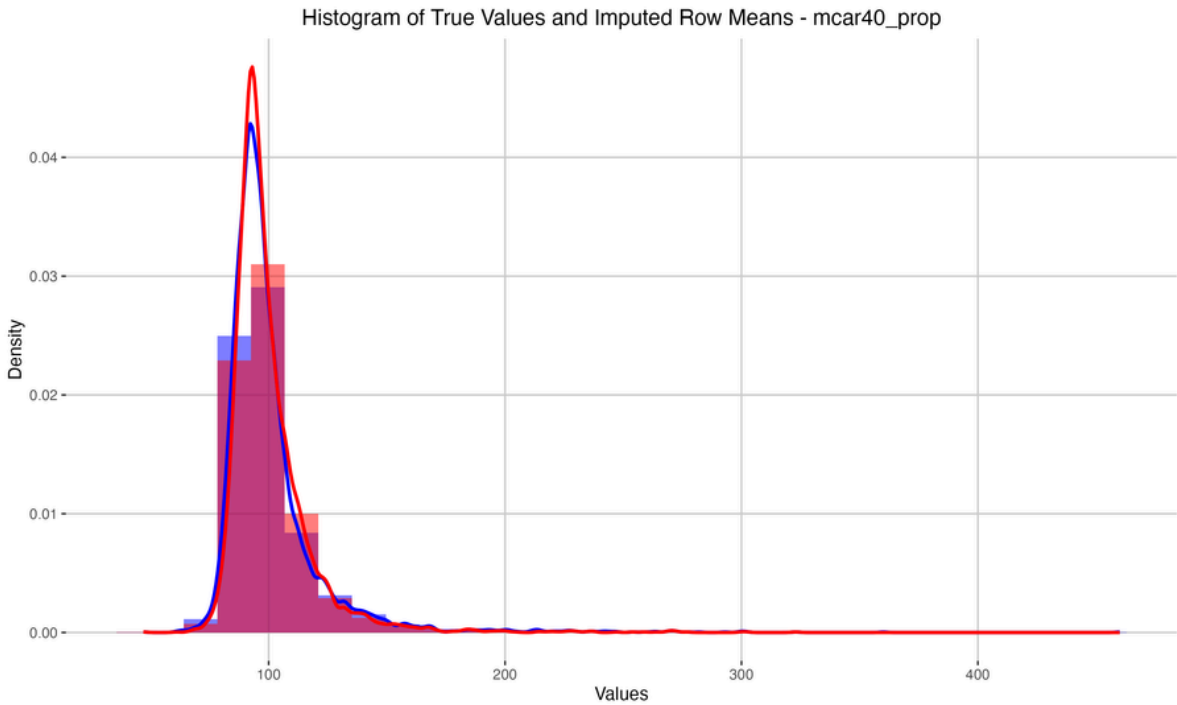
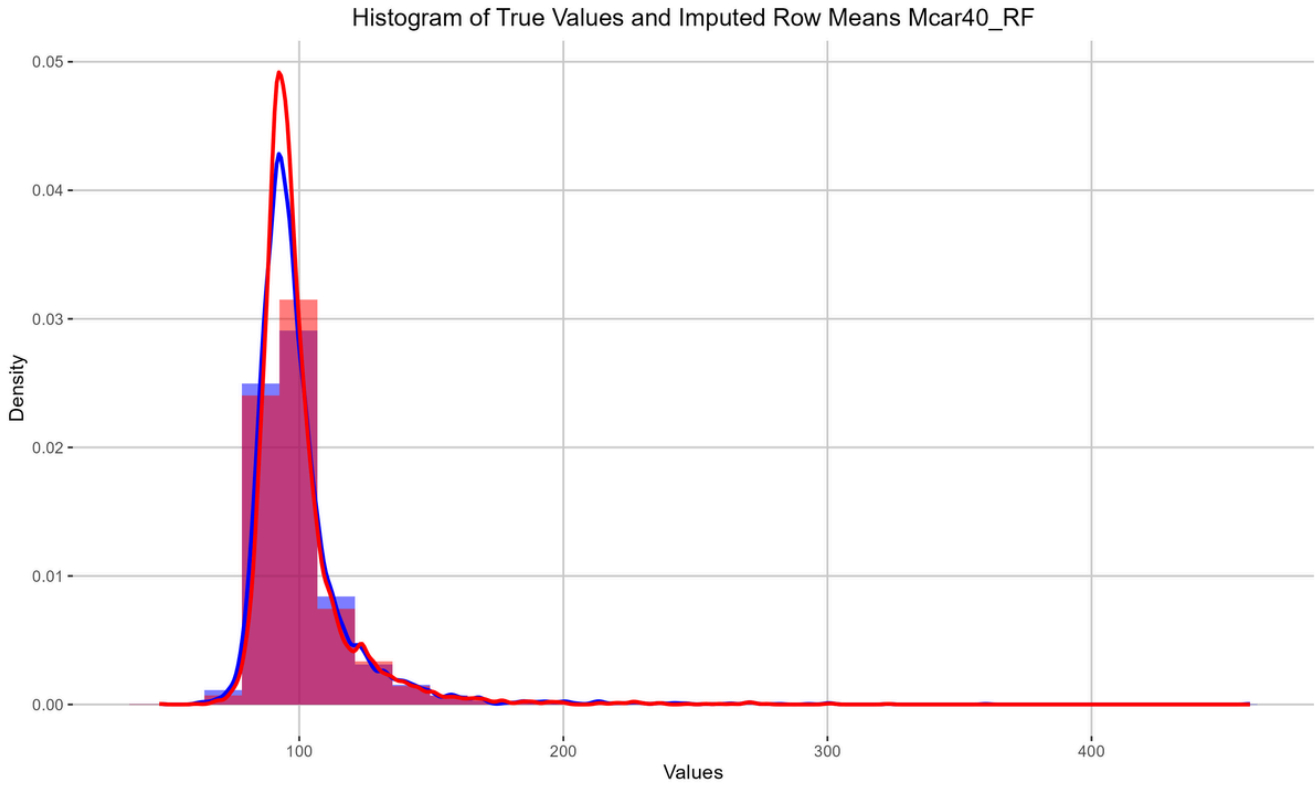
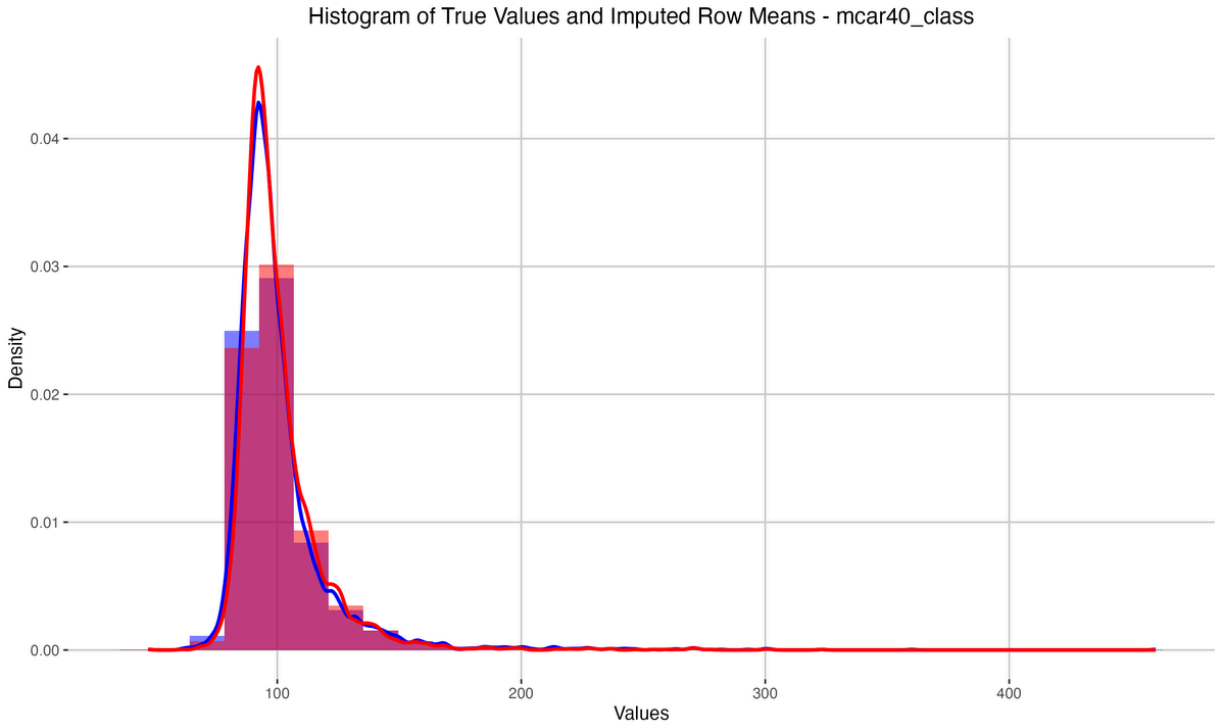
HE_glu간 관계

HE_Ucrea	HE_Uph	HE_Ubld	HE_Unitr	HE_Uro	HE_Upro	DE1_ag	HE_Usg	HE_Uglu	HE_HbA1c	age	region	incm5	edu
-0.140854136	-0.033622510	-0.008024170	0.030804935	-0.005788269	0.082963813	-0.486714922	0.064253505	0.517230665	0.783173470	0.265807089	0.035823840	-0.034983008	-0.083114148
occp	genertn	HEfst	HEmens	HEprg	HE_DMfh1	HE_DMfh2	HE_DMfh3	HE_rPLS	HE_sbp	HE_dbp	HE_BMI	HE_glu	
0.044472018	-0.087912205	-0.070621872	0.115498850	0.116842394	0.052397903	0.086323111	-0.013872251	0.028755966	0.220041437	0.149468115	0.218381607	1.000000000	

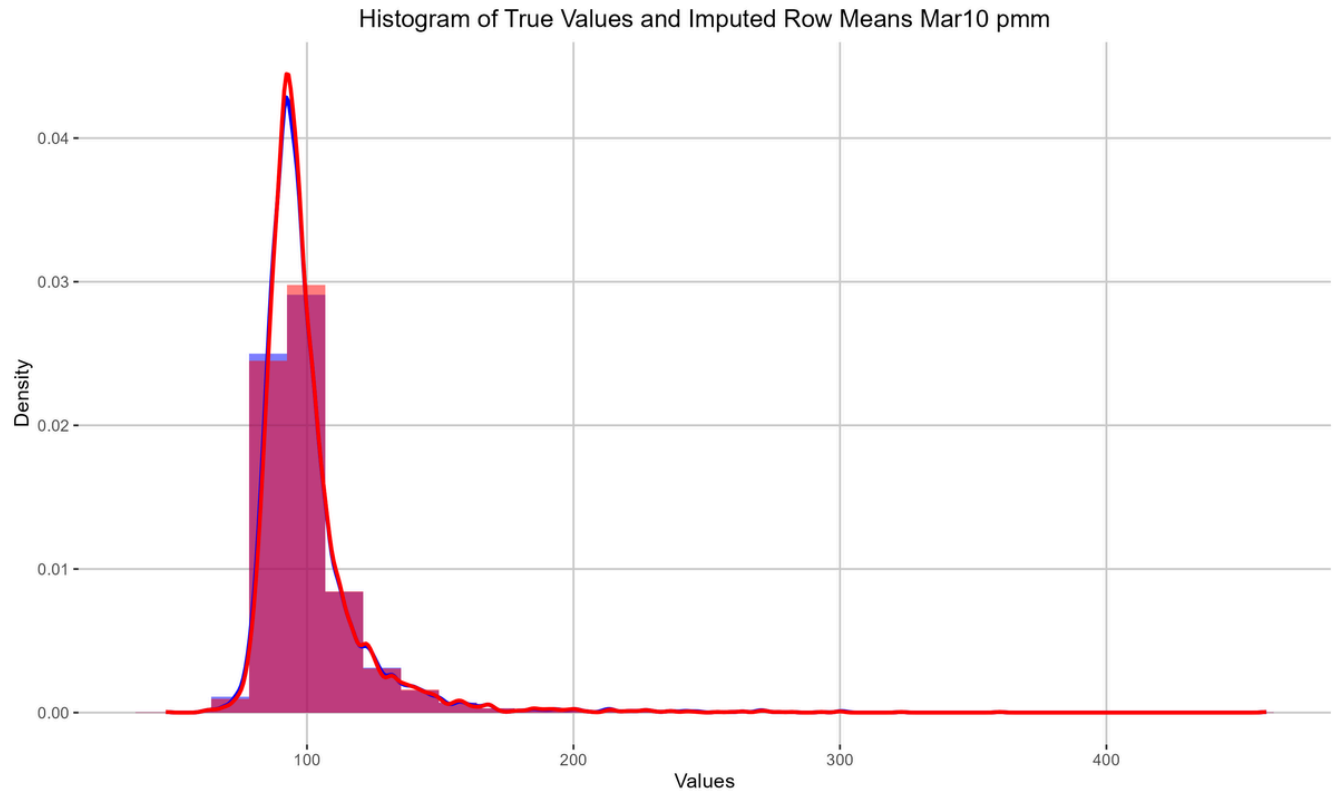
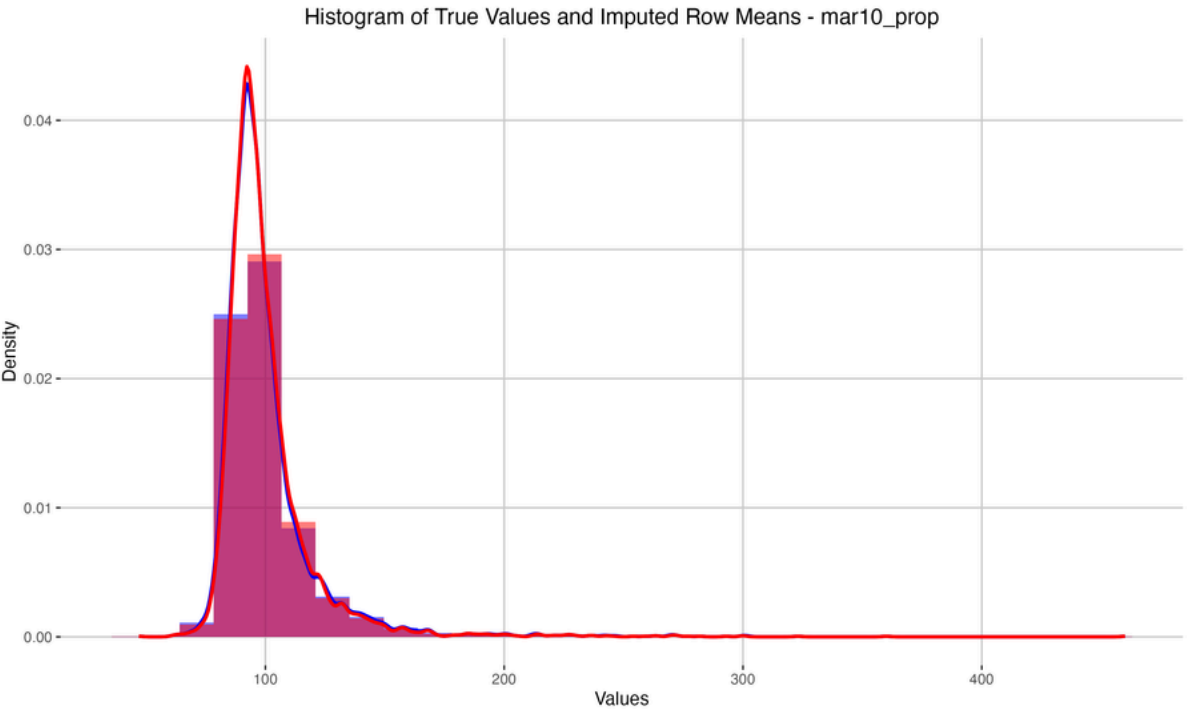
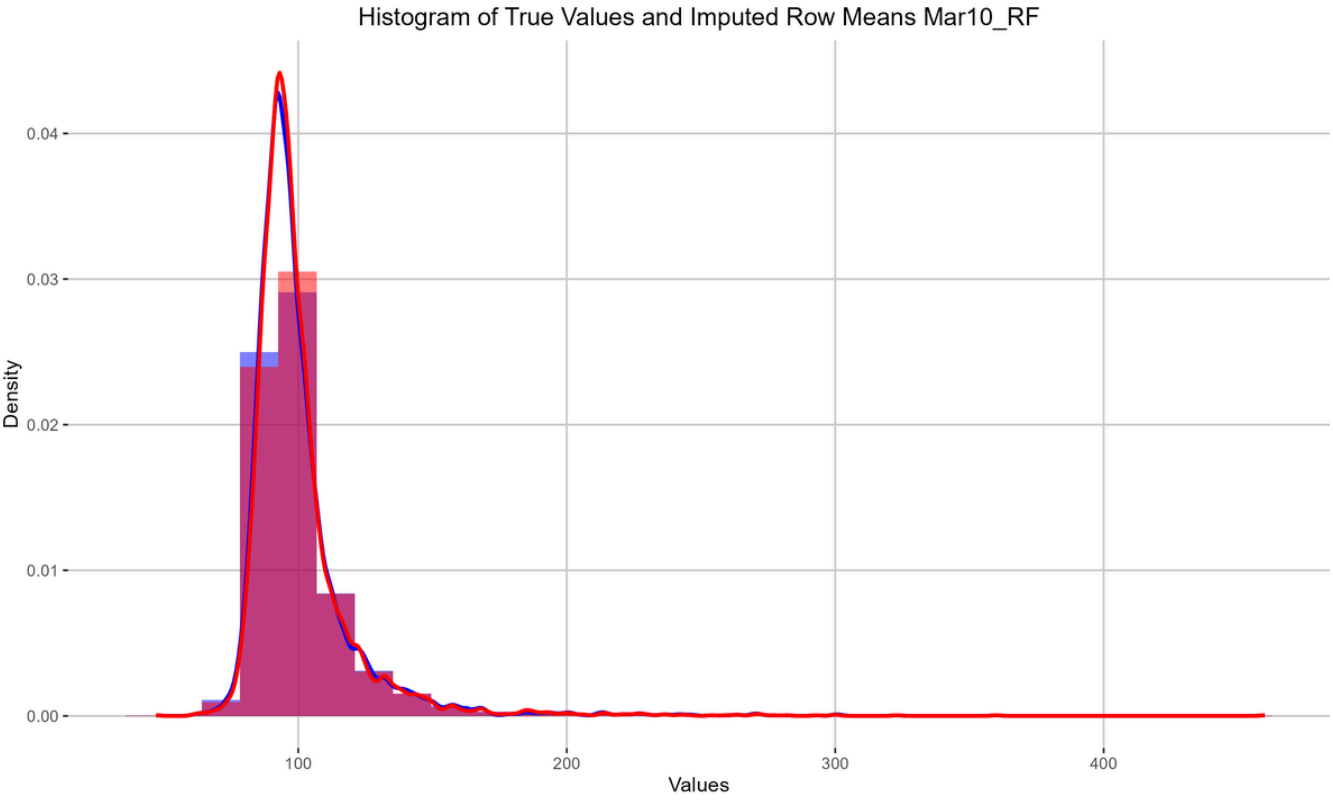
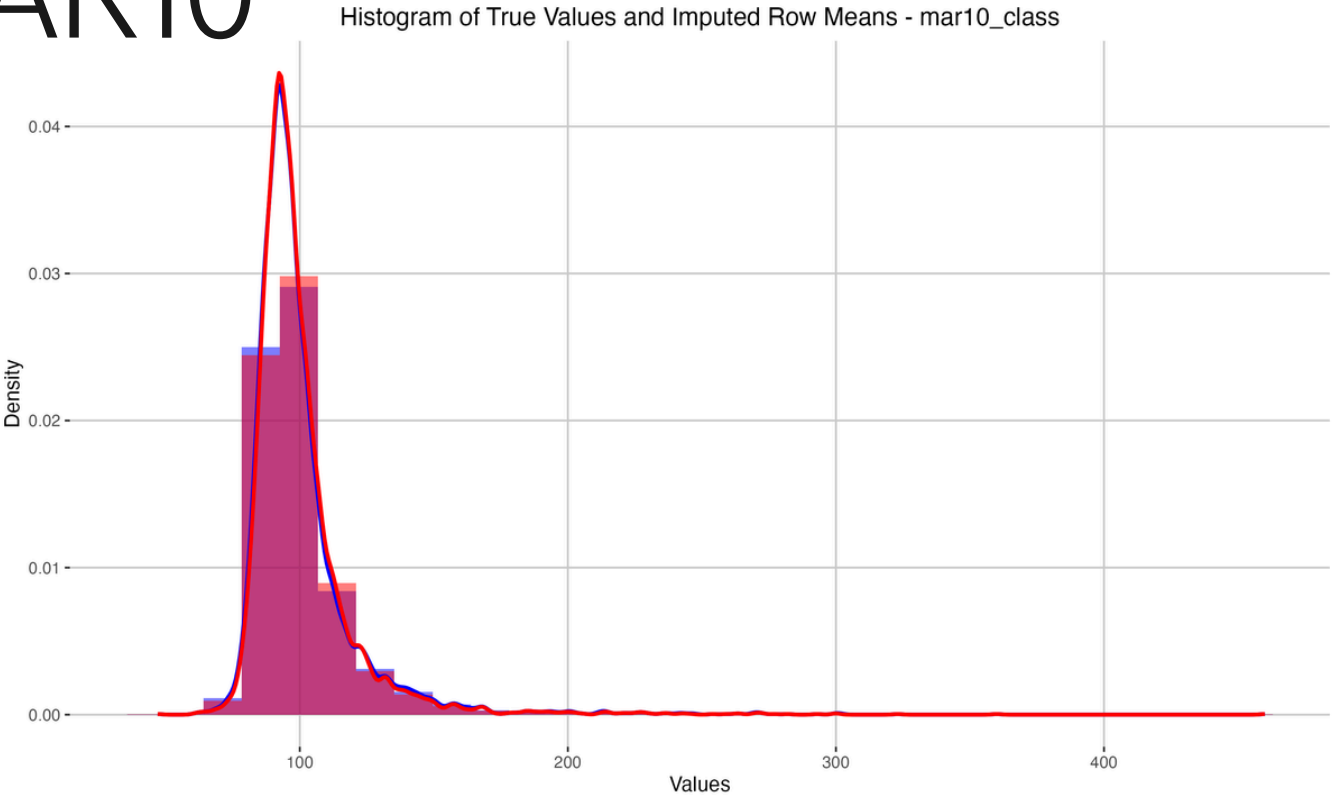
MCAR10



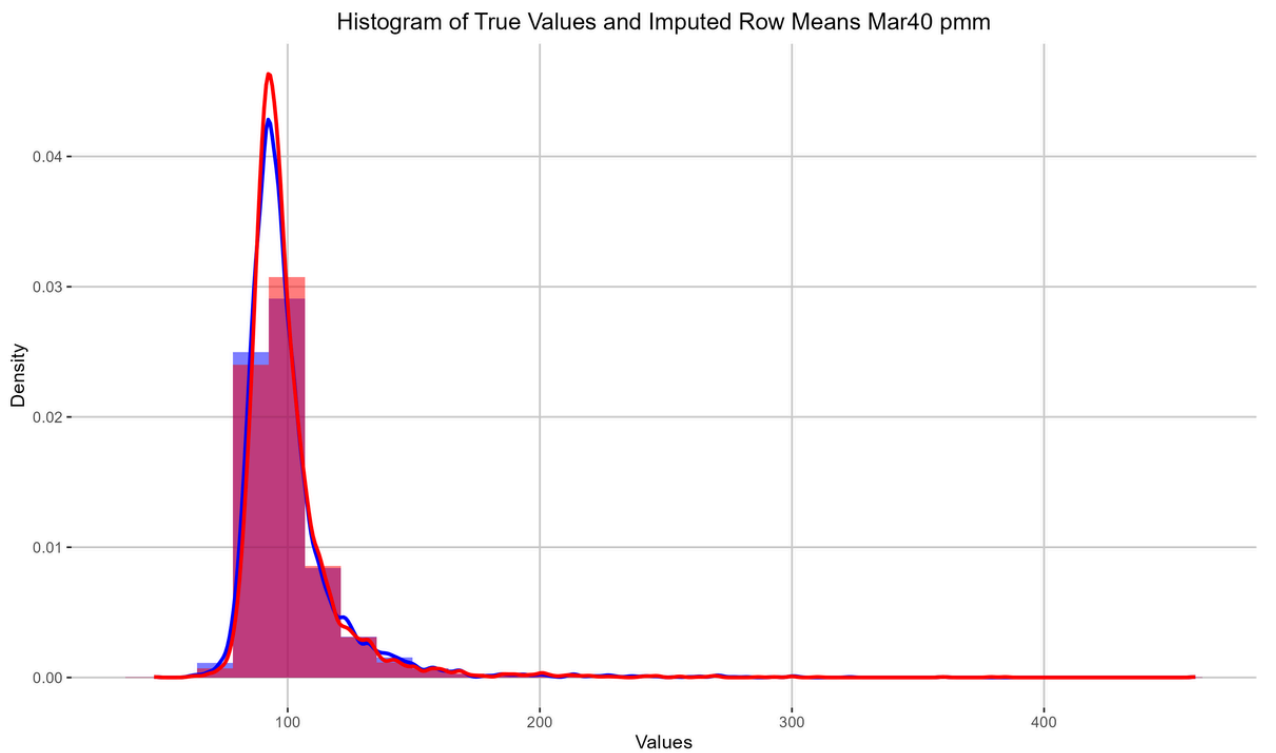
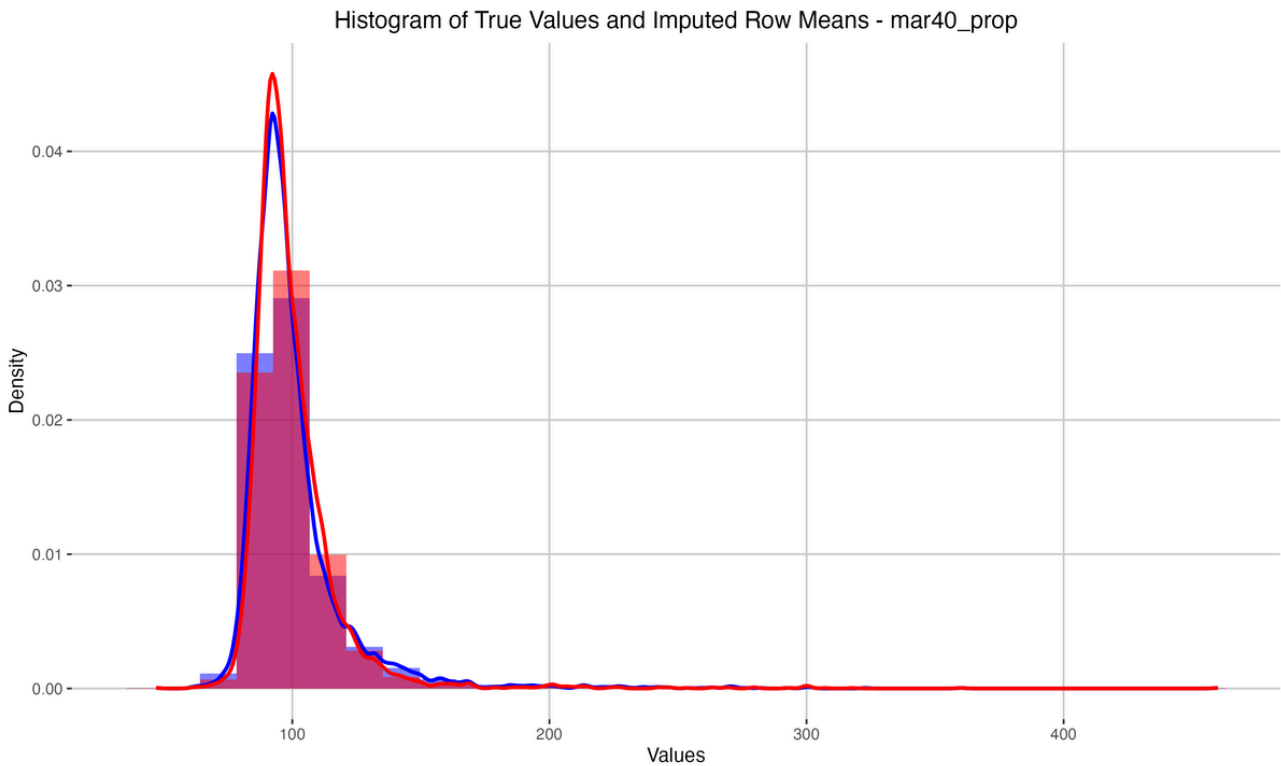
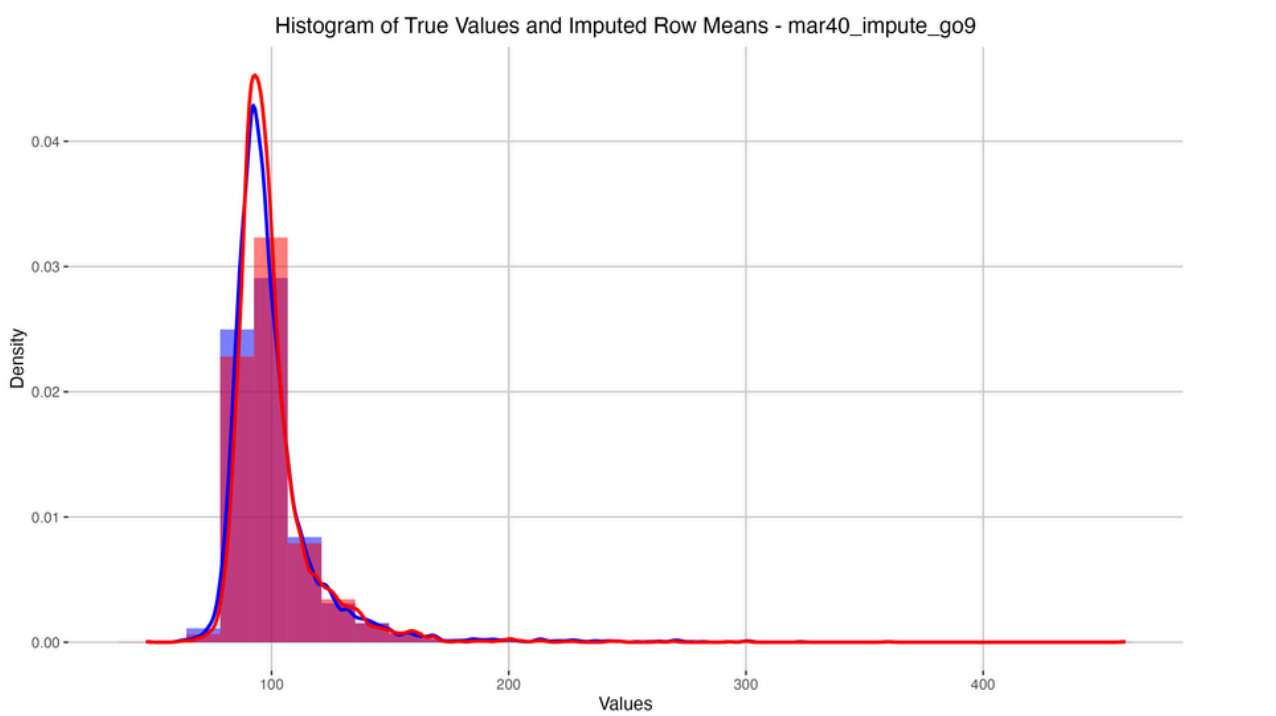
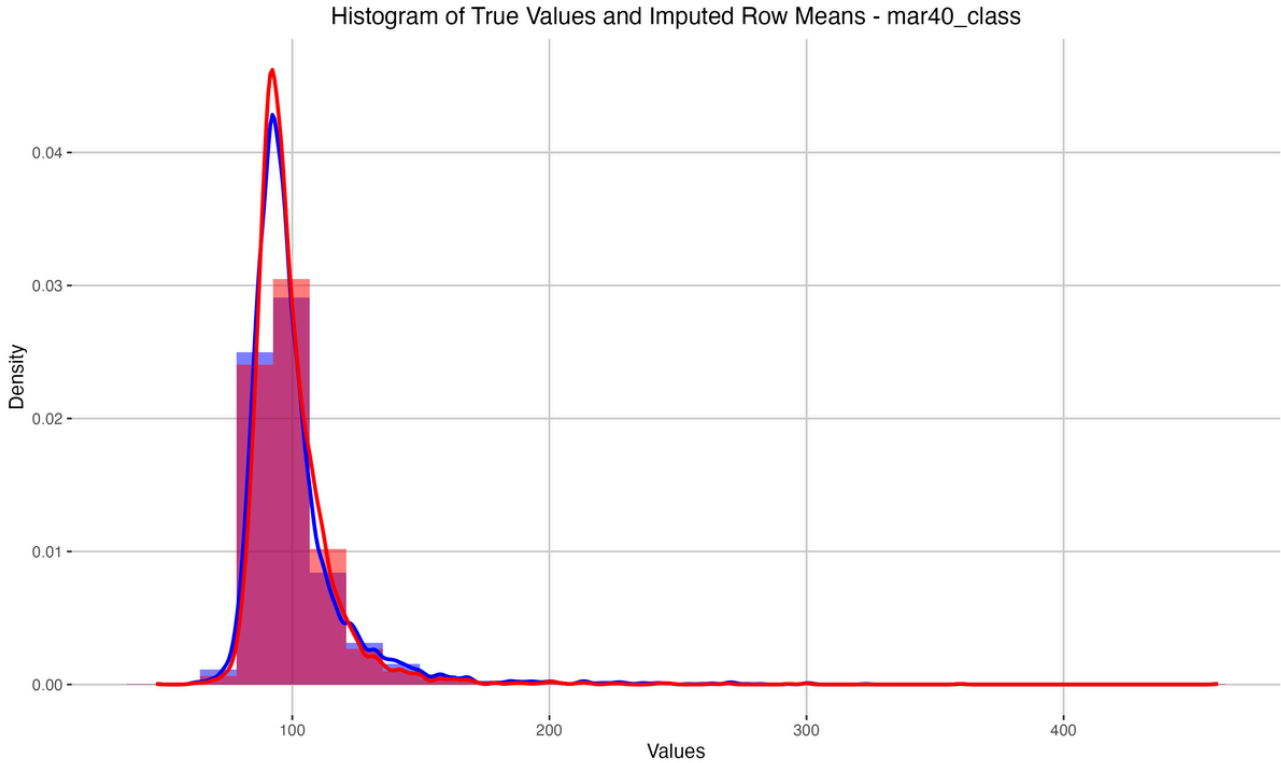
MCAR40



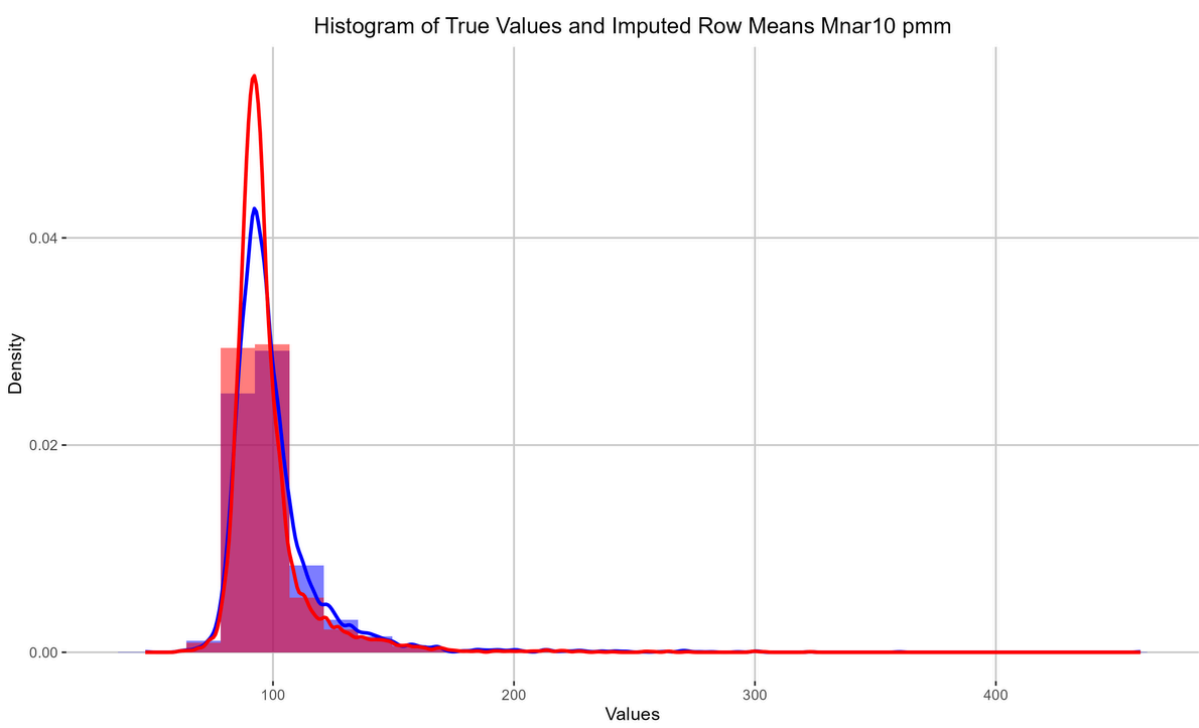
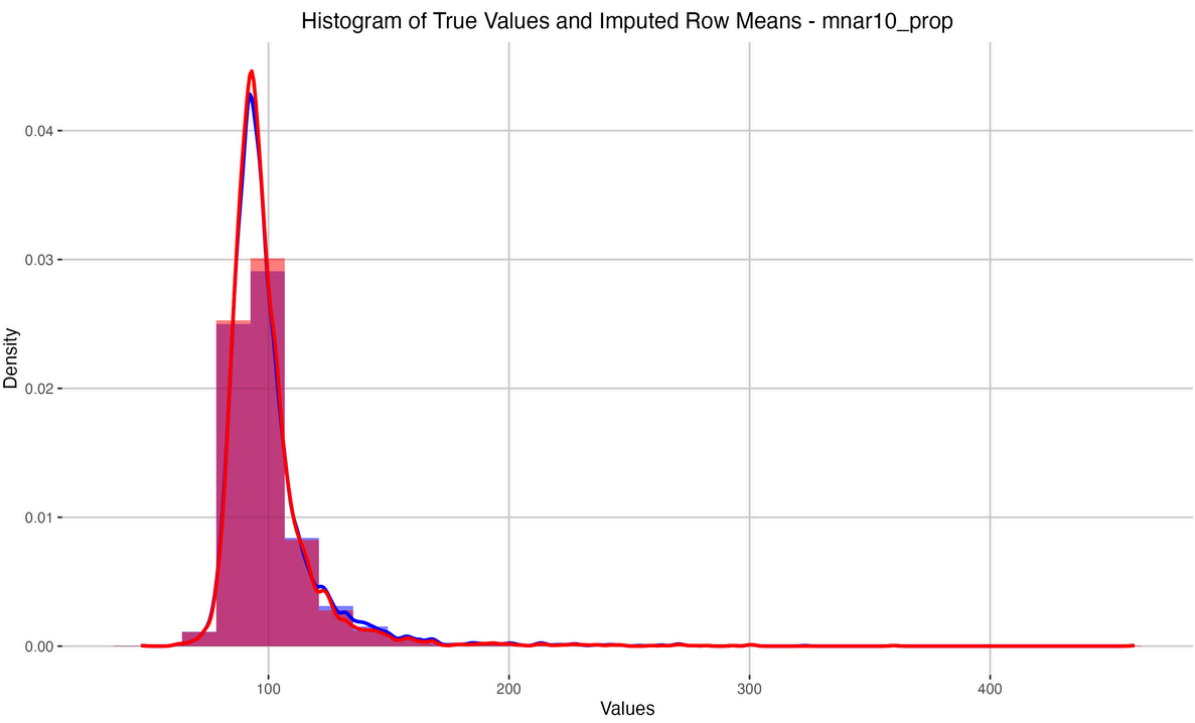
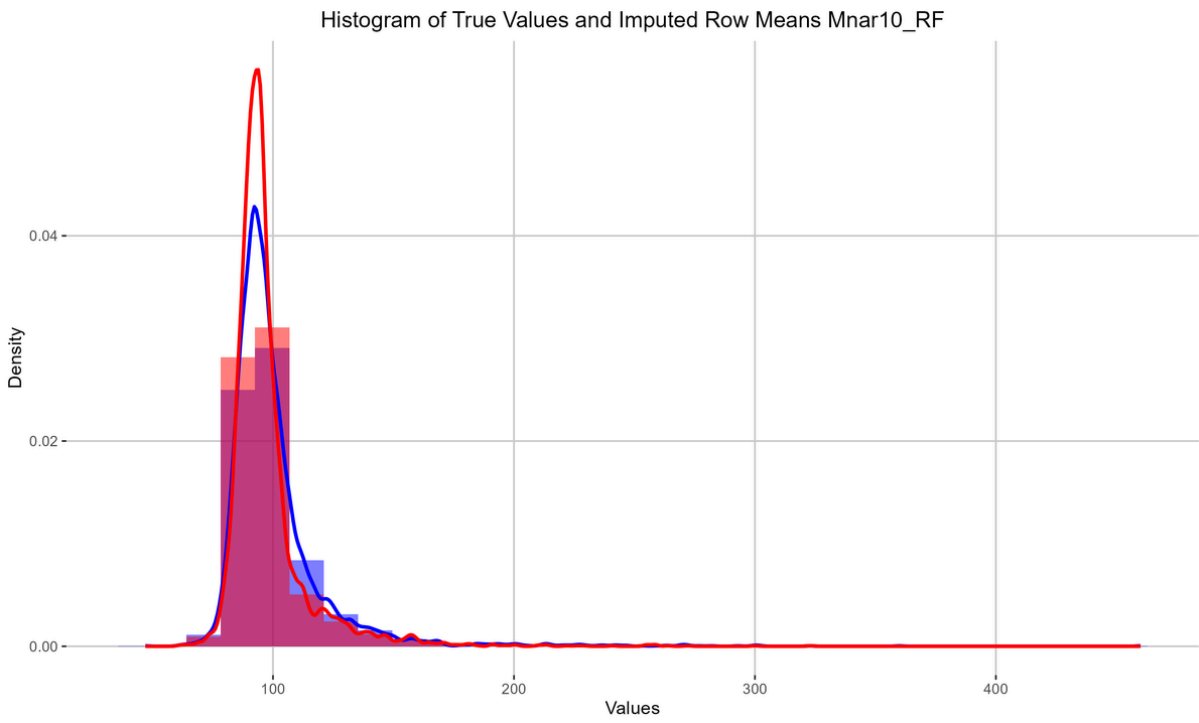
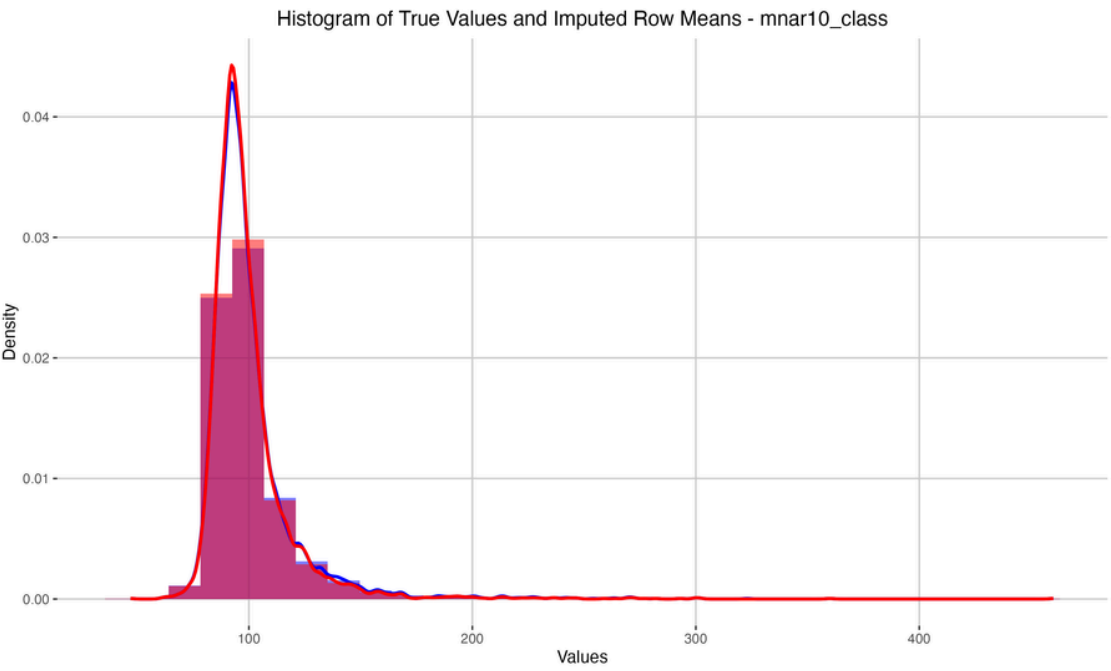
MAR10



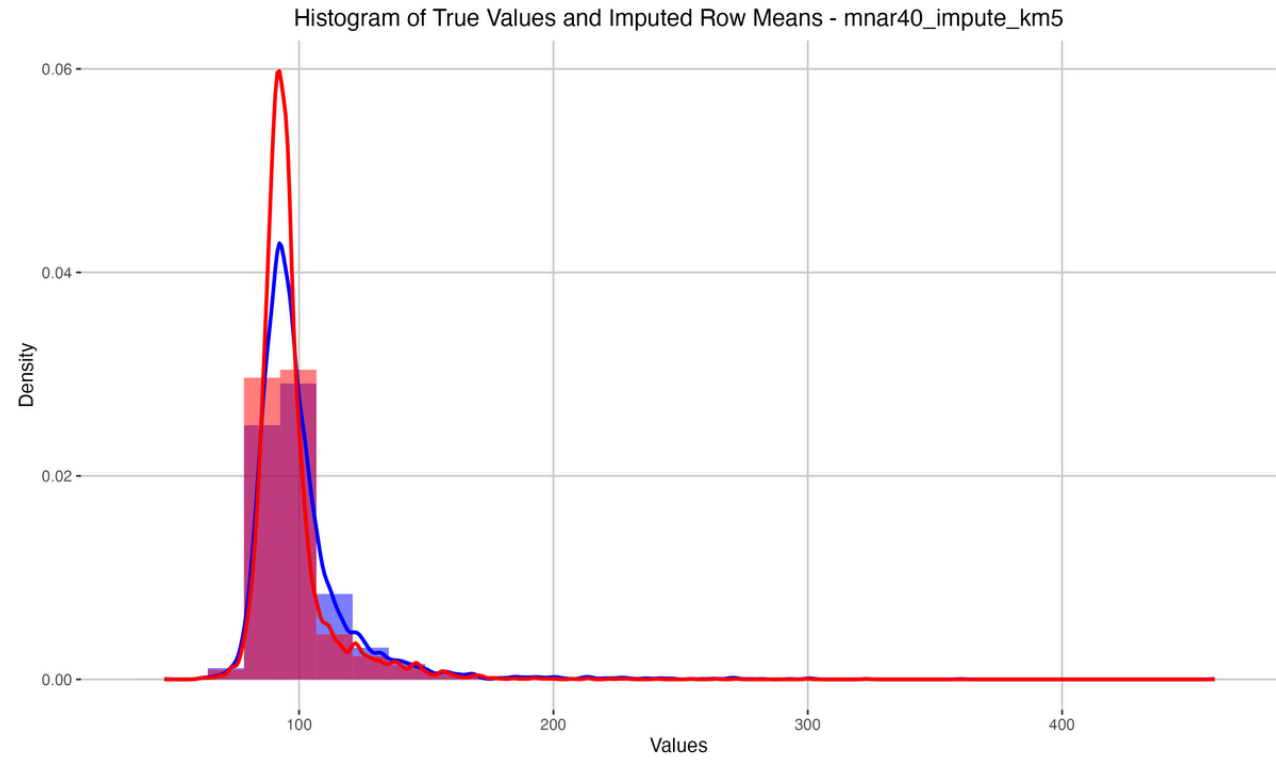
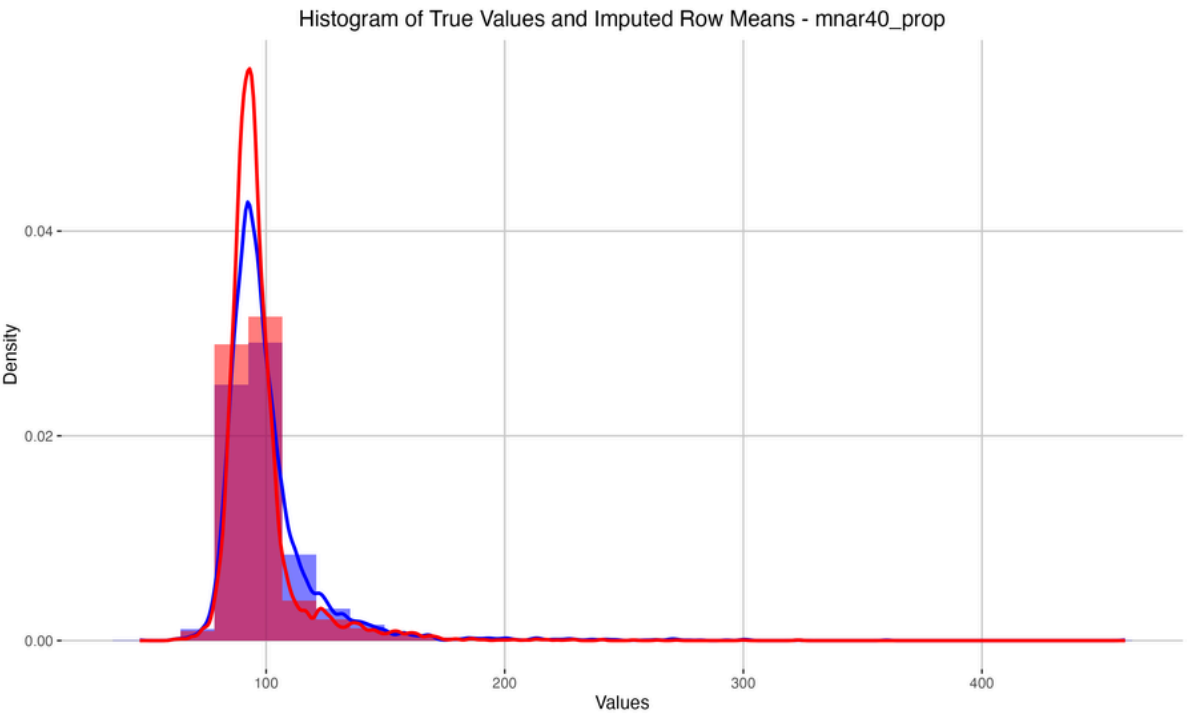
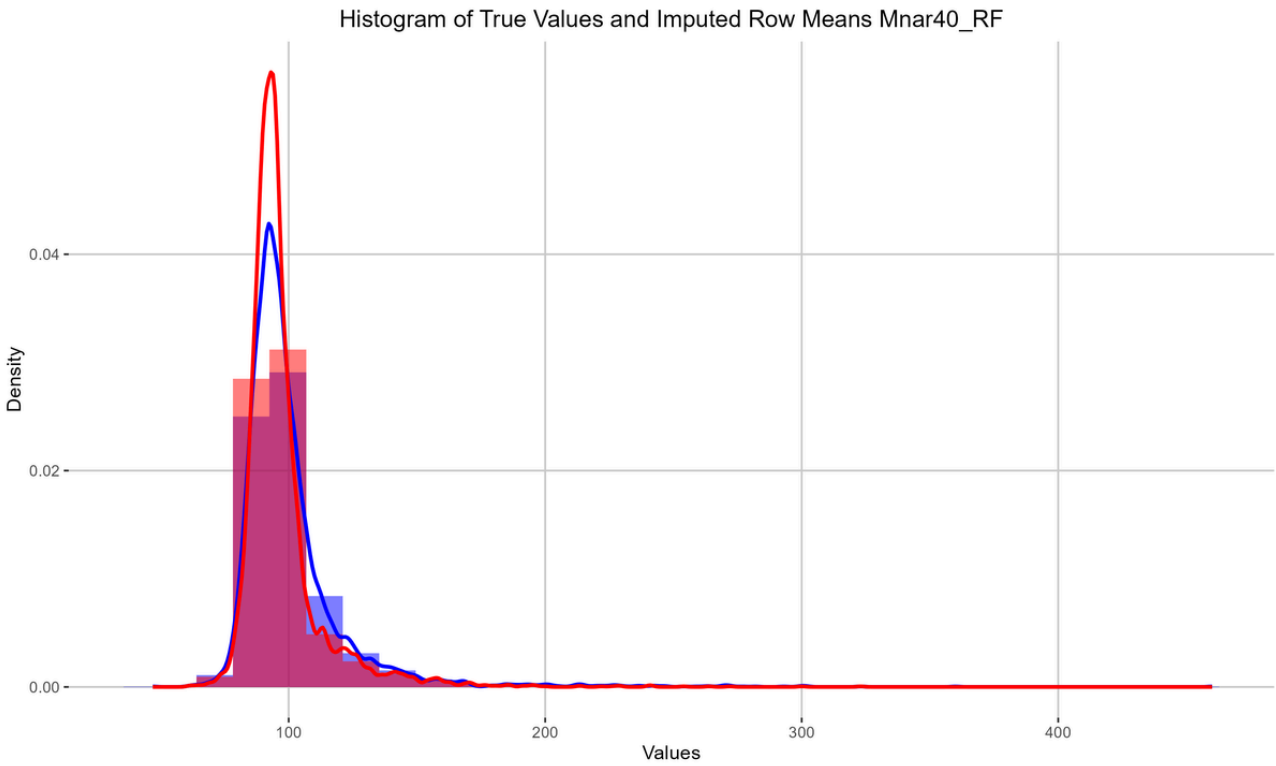
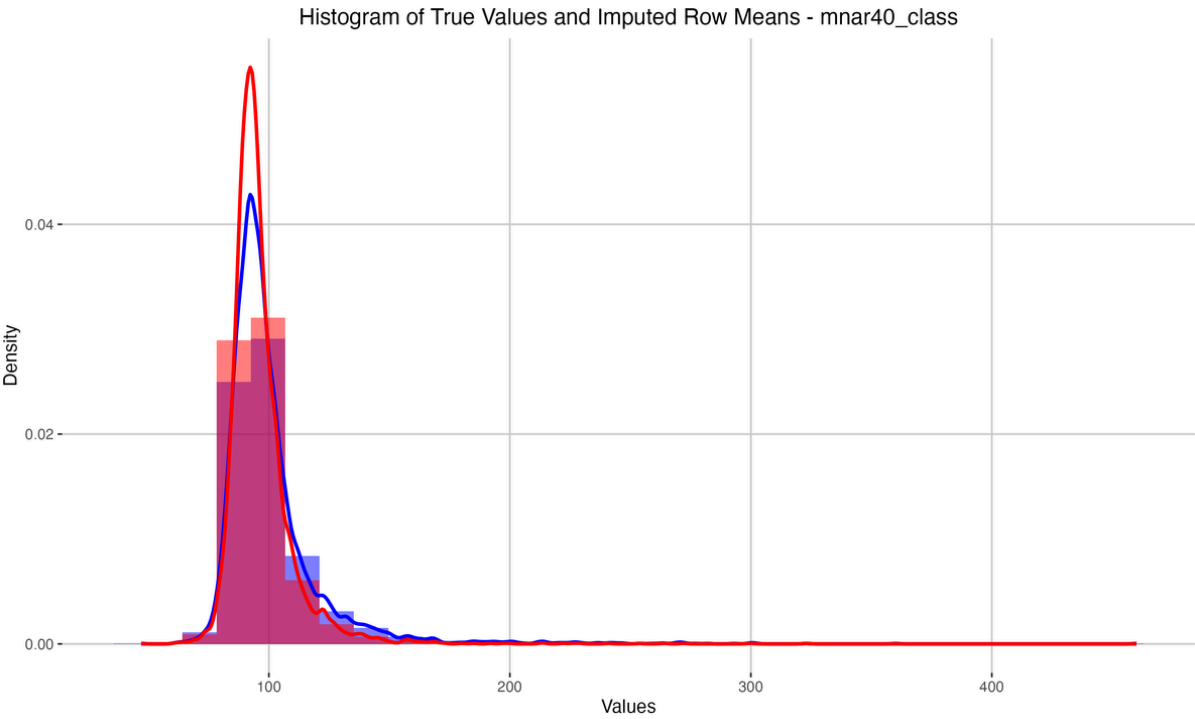
MAR40



MNAR10



MNAR40



Performance Comparision

–Result

```
> complete[complete$HE_glu>250,"HE_glu"]
```

```
# A tibble: 18 × 1
```

```
  HE_glu
```

```
  <dbl>
```

```
1    270
2    300
3    323
4    272
5    270
6    360
7    300
8    282
9    259
10   264
11   460
12   263
13   293
14   269
15   254
16   277
17   302
18   272
```

```
> sum(mar10$HE_glu>250,na.rm=TRUE)
```

```
[1] 16
```

```
> sum(mar10$HE_glu>400,na.rm=TRUE)
```

```
[1] 1
```

```
> sum(mar40$HE_glu>400,na.rm=TRUE)
```

```
[1] 1
```

```
> sum(mar40$HE_glu>250,na.rm=TRUE)
```

```
[1] 8
```

```
> sum(mcar10$HE_glu>250,na.rm=TRUE)
```

```
[1] 14
```

```
> sum(mcar10$HE_glu>400,na.rm=TRUE)
```

```
[1] 1
```

```
> sum(mnar10$HE_glu>250,na.rm=TRUE)
```

```
[1] 14
```

```
> sum(mnar10$HE_glu>400,na.rm=TRUE)
```

```
[1] 1
```

Reference

- 김혜인, 송주원 (2019) A comparison of imputation methods using nonlinear models, The Korean Journal of Applied Statistics, 32(4),543–559
- 고길곤, 탁현우(2016.12) 설문자료의 결측치 처리방법에 관한 연구: 다중대체법과 재조사법을 중심으로, 행정논총 291–319
- 이경재, 임현우(2024.02) 건물 에너지 데이터 분석에서 결측치 처리방식에 따른 차원 축소 및 모델 예측 성능 비교, 한국태양에너지학회, 59–75
- 송주원 (2011) 핫덱대체의 대체군 형성 변수 선택, 한국자료분석학회
- 송주원 (2009) 비정규성 변수에 대하여 잠재변수를 이용한 다중대체법의 적용, Journal of The Korean Data Analysis Society, 11(3), 1377–1387
- 통계청(2011.10) [국민건강영양조사] 품질개선지원 최종결과보고서 -무응답 현황분석 및 대체
- 한국보건의료연구원 (2013) 측정된 교란요인을 고려한 성과분석 방법
- Guangyu Zhang & Roderick Little (2011) A comparative study of doubly robust estimators of the mean with missing data, Journal of Statistical Computation and Simulation, 81, 2039–2058
- Joseph D.Y. Kang and Joseph L.Schafer (2007) Demystifying Doubly Robustness: A Comparison of Alternative Strategies for Estimating a Population MEan from Incomplete Data, Statistical Science, 22, 523–539