REPORT SERIES WITH DLOOKR

Data Quality Diagnosis Report

Author:
dlookr package

Version:
0.3.12

January 9, 2020

Contents

Chapter 1

Diagnose Data

1.1 Overview of Diagnosis

1.1.1 List of all variables quality

Table 1.1: Data quality overview table

variables	type	missing (n)	missing (%)	unique (n)	uniq
subject_id	integer	0	0.000	5,607	
sex	factor	222	3.959	4	
age_years	numeric	0	0.000	91	
bmi	numeric	0	0.000	1,650	
height_cm	numeric	0	0.000	185	
weight_kg	numeric	0	0.000	168	
country_of_birth	factor	22	0.392	90	
census_region	factor	1,053	18.780	5	
economic_region	factor	1,053	18.780	9	
state	factor	791	14.107	93	
country_residence	logical	5,607	100.000	1	
diet_type	factor	59	1.052	6	
multivitamin	factor	88	1.569	3	
probiotic_frequency	factor	2,990	53.326	6	
vitamin_b_supplement_frequency	factor	3,005	53.594	6	
vitamin_d_supplement_frequency	factor	3,010	53.683	6	
other_supplement_frequency	factor	93	1.659	3	
specialized_diet_exclude_dairy	logical	5,607	100.000	1	
specialized_diet_exclude_nightshades	logical	5,607	100.000	1	
specialized_diet_exclude_refined_sugars	logical	5,607	100.000	1	
specialized_diet_fodmap	logical	5,607	100.000	1	
specialized_diet_halaal	logical	5,607	100.000	1	
specialized_diet_i_do_not_eat_a_specialized_diet	logical	5,607	100.000	1	
specialized_diet_kosher	logical	5,607	100.000	1	
specialized_diet_modified_paleo_diet	logical	5,607	100.000	1	
specialized_diet_other_restrictions_not_described_here	logical	5,607	100.000	1	
specialized_diet_paleodiet_or_primal_diet	logical	5,607	100.000	1	
specialized_diet_raw_food_diet	logical	5,607	100.000	1	
specialized_diet_unspecified	logical	5,607	100.000	1	
specialized_diet_westenprice_or_other_lowgrain_low_processed_fo	logical	5,607	100.000	1	

Table 1.1: Data quality overview table *(continued)*

variables	type	missing (n)	missing (%)	unique (n)	uniq
consume_animal_products_abx	factor	2,979	53.130	4	
drinking_water_source	factor	38	0.678	6	
race	factor	26	0.464	6	
last_move	factor	2,968	52.934	6	
last_travel	factor	140	2.497	6	
roommates	factor	4,959	88.443	5	
roommates_in_study	factor	3,237	57.731	4	
livingwith	factor	89	1.587	4	
dog	factor	83	1.480	3	
cat	factor	128	2.283	3	
pets_other	logical	5,607	100.000	1	
dominant_hand	factor	120	2.140	4	
level_of_education	factor	3,060	54.575	8	
exercise_frequency	factor	48	0.856	6	
exercise_location	factor	191	3.406	6	
nail_biter	factor	158	2.818	3	
pool_frequency	factor	50	0.892	6	
smoking_frequency	factor	46	0.820	6	
alcohol_consumption	factor	49	0.874	3	
alcohol_frequency	factor	49	0.874	6	
alcohol_types_beercider	factor	0	0.000	4	
alcohol_types_red_wine	factor	0	0.000	2	
alcohol_types_sour_beers	factor	0	0.000	2	
alcohol_types_spiritshard_alcohol	factor	0	0.000	2	
alcohol_types_unspecified	factor	0	0.000	2	
alcohol_types_white_wine	factor	0	0.000	2	
drinks_per_session	logical	5,607	100.000	1	
teethbrushing_frequency	factor	58	1.034	6	
flossing_frequency	factor	40	0.713	6	
cosmetics_frequency	factor	55	0.981	6	
deodorant_use	factor	56	0.999	5	
sleep_duration	factor	34	0.606	6	
softener	factor	157	2.800	5	
bowel_movement_frequency	factor	3,021	53.879	7	
bowel_movement_quality	factor	3,033	54.093	8	
antibiotic_history	factor	71	1.266	6	
flu_vaccine_date	factor	101	1.801	6	
contraceptive	factor	1,638	29.213	7	
pregnant	factor	2,159	38.505	4	
weight_change	factor	91	1.623	4	
tonsils_removed	factor	126	2.247	4	
appendix_removed	factor	114	2.033	4	
chickenpox	factor	93	1.659	4	
acne_medication	factor	113	2.015	3	
acne_medication_otc	factor	91	1.623	3	
csection	factor	78	1.391	4	
fed_as_infant	factor	2,986	53.255	5	
add_adhd	factor	3,008	53.647	5	

Table 1.1: Data quality overview table *(continued)*

variables	type	missing (n)	missing (%)	unique (n)	uniq
alzheimers	factor	2,969	52.952	4	
lung_disease	factor	387	6.902	5	
asd	factor	2,984	53.219	5	
autoimmune	factor	3,010	53.683	5	
fungal_overgrowth	factor	3,052	54.432	5	
cdiff	factor	3,028	54.004	5	
cardiovascular_disease	factor	2,977	53.094	5	
mental_illness	logical	5,607	100.000	1	
mental_illness_type_anorexia_nervosa	logical	5,607	100.000	1	
mental_illness_type_bipolar_disorder	logical	5,607	100.000	1	
mental_illness_type_bulimia_nervosa	logical	5,607	100.000	1	
mental_illness_type_depression	logical	5,607	100.000	1	
mental_illness_type_ptsd_posttraumatic_stress_disorder	logical	5,607	100.000	1	
mental_illness_type_schizophrenia	logical	5,607	100.000	1	
mental_illness_type_substance_abuse	logical	5,607	100.000	1	
mental_illness_type_unspecified	logical	5,607	100.000	1	
diabetes	factor	165	2.943	5	
diabetes_type	logical	5,607	100.000	1	
epilepsy_or_seizure_disorder	factor	2,984	53.219	5	
ibs	factor	3,026	53.968	5	
ibd	factor	329	5.868	5	
ibd_diagnosis	logical	5,607	100.000	1	
ibd_diagnosis_refined	logical	5,607	100.000	1	
migraine	factor	467	8.329	5	
kidney_disease	factor	2,991	53.344	5	
liver_disease	factor	2,996	53.433	5	
pku	factor	149	2.657	5	
sibo	factor	3,079	54.914	5	
skin_condition	factor	618	11.022	5	
thyroid	factor	2,996	53.433	5	
seasonal_allergies	factor	171	3.050	5	
non_food_allergies_beestings	factor	0	0.000	2	
non_food_allergies_drug_eg_penicillin	factor	0	0.000	2	
non_food_allergies_pet_dander	factor	0	0.000	2	
non_food_allergies_poison_ivyoak	factor	0	0.000	2	
non_food_allergies_sun	factor	0	0.000	2	
non_food_allergies_unspecified	factor	0	0.000	2	
lactose	factor	143	2.550	3	
gluten	factor	798	14.232	5	
allergic_to_i_have_no_food_allergies_that_i_know_of	factor	0	0.000	2	
allergic_to_other	factor	0	0.000	2	
allergic_to_peanuts	factor	0	0.000	2	
allergic_to_shellfish	factor	0	0.000	2	
allergic_to_tree_nuts	factor	0	0.000	2	
allergic_to_unspecified	factor	0	0.000	2	
vivid_dreams	logical	5,607	100.000	1	
breastmilk_formula_ensure	factor	3,107	55.413	4	
meat_eggs_frequency	factor	2,987	53.273	6	

Table 1.1: Data quality overview table *(continued)*

variables	type	missing (n)	missing (%)	unique (n)	uniq
homecooked_meals_frequency	factor	2,977	53.094	6	
ready_to_eat_meals_frequency	factor	2,977	53.094	6	
prepared_meals_frequency	factor	2,974	53.041	6	
whole_grain_frequency	factor	2,989	53.308	6	
fruit_frequency	factor	2,995	53.415	6	
vegetable_frequency	factor	2,993	53.380	6	
types_of_plants	factor	2,584	46.085	6	
fermented_plant_frequency	factor	3,058	54.539	6	
milk_cheese_frequency	factor	2,984	53.219	6	
milk_substitute_frequency	factor	2,984	53.219	6	
frozen_dessert_frequency	factor	2,976	53.077	6	
red_meat_frequency	factor	2,992	53.362	6	
high_fat_red_meat_frequency	factor	2,993	53.380	6	
poultry_frequency	factor	2,969	52.952	6	
seafood_frequency	factor	2,979	53.130	6	
salted_snacks_frequency	factor	2,974	53.041	6	
sugary_sweets_frequency	factor	2,968	52.934	6	
olive_oil	factor	2,990	53.326	6	
whole_eggs	factor	2,968	52.934	6	
sugar_sweetened_drink_frequency	factor	2,995	53.415	6	
artificial_sweeteners	logical	5,607	100.000	1	
one_liter_of_water_a_day_frequency	factor	3,004	53.576	6	

1.1.2 Diagnosis of missing data

Table 1.2: Variables that include missing values

variables	type	missing (n)	missing (%)	unique (n)	uniqu
country_residence	logical	5,607	100.000	1	
specialized_diet_exclude_dairy	logical	5,607	100.000	1	
specialized_diet_exclude_nightshades	logical	5,607	100.000	1	
specialized_diet_exclude_refined_sugars	logical	5,607	100.000	1	
specialized_diet_fodmap	logical	5,607	100.000	1	
specialized_diet_halaal	logical	5,607	100.000	1	
specialized_diet_i_do_not_eat_a_specialized_diet	logical	5,607	100.000	1	
specialized_diet_kosher	logical	5,607	100.000	1	
specialized_diet_modified_paleo_diet	logical	5,607	100.000	1	
specialized_diet_other_restrictions_not_described_here	logical	5,607	100.000	1	
specialized_diet_paleodiet_or_primal_diet	logical	5,607	100.000	1	
specialized_diet_raw_food_diet	logical	5,607	100.000	1	
specialized_diet_unspecified	logical	5,607	100.000	1	
specialized_diet_westenprice_or_other_lowgrain_low_processed_fo	logical	5,607	100.000	1	
pets_other	logical	5,607	100.000	1	
drinks_per_session	logical	5,607	100.000	1	
mental_illness	logical	5,607	100.000	1	
mental_illness_type_anorexia_nervosa	logical	5,607	100.000	1	
mental_illness_type_bipolar_disorder	logical	5,607	100.000	1	
mental_illness_type_bulimia_nervosa	logical	5,607	100.000	1	

Table 1.2: Variables that include missing values *(continued)*

variables	type	missing (n)	missing (%)	unique (n)	uniqu
mental_illness_type_depression	logical	5,607	100.000	1	
mental_illness_type_ptsd_posttraumatic_stress_disorder	logical	5,607	100.000	1	
mental_illness_type_schizophrenia	logical	5,607	100.000	1	
mental_illness_type_substance_abuse	logical	5,607	100.000	1	
mental_illness_type_unspecified	logical	5,607	100.000	1	
diabetes_type	logical	5,607	100.000	1	
ibd_diagnosis	logical	5,607	100.000	1	
ibd_diagnosis_refined	logical	5,607	100.000	1	
vivid_dreams	logical	5,607	100.000	1	
artificial_sweeteners	logical	5,607	100.000	1	
roommates	factor	4,959	88.443	5	
roommates_in_study	factor	3,237	57.731	4	
breastmilk_formula_ensure	factor	3,107	55.413	4	
sibo	factor	3,079	54.914	5	
level_of_education	factor	3,060	54.575	8	
fermented_plant_frequency	factor	3,058	54.539	6	
fungal_overgrowth	factor	3,052	54.432	5	
bowel_movement_quality	factor	3,033	54.093	8	
cdiff	factor	3,028	54.004	5	
ibs	factor	3,026	53.968	5	
bowel_movement_frequency	factor	3,021	53.879	7	
vitamin_d_supplement_frequency	factor	3,010	53.683	6	
autoimmune	factor	3,010	53.683	5	
add_adhd	factor	3,008	53.647	5	
vitamin_b_supplement_frequency	factor	3,005	53.594	6	
one_liter_of_water_a_day_frequency	factor	3,004	53.576	6	
liver_disease	factor	2,996	53.433	5	
thyroid	factor	2,996	53.433	5	
fruit_frequency	factor	2,995	53.415	6	
sugar_sweetened_drink_frequency	factor	2,995	53.415	6	
vegetable_frequency	factor	2,993	53.380	6	
high_fat_red_meat_frequency	factor	2,993	53.380	6	
red_meat_frequency	factor	2,992	53.362	6	
kidney_disease	factor	2,991	53.344	5	
probiotic_frequency	factor	2,990	53.326	6	
olive_oil	factor	2,990	53.326	6	
whole_grain_frequency	factor	2,989	53.308	6	
meat_eggs_frequency	factor	2,987	53.273	6	
fed_as_infant	factor	2,986	53.255	5	
asd	factor	2,984	53.219	5	
epilepsy_or_seizure_disorder	factor	2,984	53.219	5	
milk_cheese_frequency	factor	2,984	53.219	6	
milk_substitute_frequency	factor	2,984	53.219	6	
consume_animal_products_abx	factor	2,979	53.130	4	
seafood_frequency	factor	2,979	53.130	6	
cardiovascular_disease	factor	2,977	53.094	5	
homecooked_meals_frequency	factor	2,977	53.094	6	
ready_to_eat_meals_frequency	factor	2,977	53.094	6	

Table 1.2: Variables that include missing values *(continued)*

variables	type	missing (n)	missing (%)	unique (n)	uniqu
frozen_dessert_frequency	factor	2,976	53.077	6	
prepared_meals_frequency	factor	2,974	53.041	6	
salted_snacks_frequency	factor	2,974	53.041	6	
alzheimers	factor	2,969	52.952	4	
poultry_frequency	factor	2,969	52.952	6	
last_move	factor	2,968	52.934	6	
sugary_sweets_frequency	factor	2,968	52.934	6	
whole_eggs	factor	2,968	52.934	6	
types_of_plants	factor	2,584	46.085	6	
pregnant	factor	2,159	38.505	4	
contraceptive	factor	1,638	29.213	7	
census_region	factor	1,053	18.780	5	
economic_region	factor	1,053	18.780	9	
gluten	factor	798	14.232	5	
state	factor	791	14.107	93	
skin_condition	factor	618	11.022	5	
migraine	factor	467	8.329	5	
lung_disease	factor	387	6.902	5	
ibd	factor	329	5.868	5	
sex	factor	222	3.959	4	
exercise_location	factor	191	3.406	6	
seasonal_allergies	factor	171	3.050	5	
diabetes	factor	165	2.943	5	
nail_biter	factor	158	2.818	3	
softener	factor	157	2.800	5	
pku	factor	149	2.657	5	
lactose	factor	143	2.550	3	
last_travel	factor	140	2.497	6	
cat	factor	128	2.283	3	
tonsils_removed	factor	126	2.247	4	
dominant_hand	factor	120	2.140	4	
appendix_removed	factor	114	2.033	4	
acne_medication	factor	113	2.015	3	
flu_vaccine_date	factor	101	1.801	6	
other_supplement_frequency	factor	93	1.659	3	
chickenpox	factor	93	1.659	4	
weight_change	factor	91	1.623	4	
acne_medication_otc	factor	91	1.623	3	
livingwith	factor	89	1.587	4	
multivitamin	factor	88	1.569	3	
dog	factor	83	1.480	3	
csection	factor	78	1.391	4	
antibiotic_history	factor	71	1.266	6	
diet_type	factor	59	1.052	6	
teethbrushing_frequency	factor	58	1.034	6	
deodorant_use	factor	56	0.999	5	
cosmetics_frequency	factor	55	0.981	6	
pool_frequency	factor	50	0.892	6	

Table 1.2: Variables that include missing values *(continued)*

variables	type	missing (n)	missing (%)	unique (n)	uniqu
alcohol_consumption	factor	49	0.874	3	
alcohol_frequency	factor	49	0.874	6	
exercise_frequency	factor	48	0.856	6	
smoking_frequency	factor	46	0.820	6	
flossing_frequency	factor	40	0.713	6	
drinking_water_source	factor	38	0.678	6	
sleep_duration	factor	34	0.606	6	
race	factor	26	0.464	6	
country_of_birth	factor	22	0.392	90	

1.1.3 Diagnosis of unique data(Text and Category)

No variable with a high proportion greater than 0.5

1.1.4 Diagnosis of unique data(Numerical)

Table 1.3: Variables where the proportion of unique data is less than 0.1

variables	type	missing (n)	missing (%)	unique (n)	unique (n/N)
height_cm	numeric	0	0	185	0.033
weight_kg	numeric	0	0	168	0.030
age_years	numeric	0	0	91	0.016

1.2 Detailed data diagnosis

1.2.1 Diagnosis of categorical variables

Table 1.4: Categorical variable level top 10

variables	levels
sex	female
sex	male
sex	NA
sex	other
country_of_birth	United States
country_of_birth	United Kingdom
country_of_birth	Australia
country_of_birth	Canada
country_of_birth	Germany
country_of_birth	France
country_of_birth	India
country_of_birth	Ireland
country_of_birth	Russian Federation
country_of_birth	NA
census_region	West
census_region	South

Table 1.4: Categorical variable level top 10 *(continued)*

variables	levels
census_region	NA
census_region	Northeast
census_region	Midwest
economic_region	Far West
economic_region	NA
economic_region	Mideast
economic_region	Southeast
economic_region	New England
economic_region	Rocky Mountain
economic_region	Great Lakes
economic_region	Southwest
economic_region	Plains
state	CA
state	NA
state	CO
state	NY
state	MA
state	WA
state	TX
state	MD
state	IL
state	VA
diet_type	Omnivore
diet_type	Omnivore but do not eat red meat
diet_type	Vegetarian but eat seafood
diet_type	Vegetarian
diet_type	Vegan
diet_type	NA
multivitamin	false
multivitamin	true
multivitamin	NA
probiotic_frequency	NA
probiotic_frequency	Never
probiotic_frequency	Daily
probiotic_frequency	Rarely (a few times/month)
probiotic_frequency	Regularly (3-5 times/week)
probiotic_frequency	Occasionally (1-2 times/week)
vitamin_b_supplement_frequency	NA
vitamin_b_supplement_frequency	Never
vitamin_b_supplement_frequency	Daily
vitamin_b_supplement_frequency	Rarely (a few times/month)
vitamin_b_supplement_frequency	Regularly (3-5 times/week)
vitamin_b_supplement_frequency	Occasionally (1-2 times/week)
vitamin_d_supplement_frequency	NA
vitamin_d_supplement_frequency	Never
vitamin_d_supplement_frequency	Daily
vitamin_d_supplement_frequency	Regularly (3-5 times/week)
vitamin_d_supplement_frequency	Rarely (a few times/month)

Table 1.4: Categorical variable level top 10 *(continued)*

variables	levels
vitamin_d_supplement_frequency	Occasionally (1-2 times/week)
other_supplement_frequency	true
other_supplement_frequency	false
other_supplement_frequency	NA
consume_animal_products_abx	NA
consume_animal_products_abx	Yes
consume_animal_products_abx	Not sure
consume_animal_products_abx	No
drinking_water_source	City
drinking_water_source	Filtered
drinking_water_source	Well
drinking_water_source	Bottled
drinking_water_source	Not sure
drinking_water_source	NA
race	Caucasian
race	Asian or Pacific Islander
race	Other
race	Hispanic
race	African American
race	NA
last_move	NA
last_move	I have lived in my current state of residence for more than a year.
last_move	Within the past year
last_move	Within the past 6 months
last_move	Within the past 3 months
last_move	Within the past month
last_travel	I have not been outside of my country of residence in the past year.
last_travel	1 year
last_travel	3 months
last_travel	6 months
last_travel	Month
last_travel	NA
roommates	NA
roommates	One
roommates	Two
roommates	Three
roommates	More than three
roommates_in_study	NA
roommates_in_study	No
roommates_in_study	Yes
roommates_in_study	Not sure
livingwith	No
livingwith	Yes
livingwith	Not sure
livingwith	NA
dog	false
dog	true
dog	NA

Table 1.4: Categorical variable level top 10 *(continued)*

variables	levels
cat	false
cat	true
cat	NA
dominant_hand	I am right handed
dominant_hand	I am left handed
dominant_hand	I am ambidextrous
dominant_hand	NA
level_of_education	NA
level_of_education	Graduate or Professional degree
level_of_education	Bachelor's degree
level_of_education	Some college or technical school
level_of_education	Some graduate school or professional
level_of_education	High School or GED equilivant
level_of_education	Did not complete high school
level_of_education	Associate's degree
exercise_frequency	Regularly (3-5 times/week)
exercise_frequency	Daily
exercise_frequency	Occasionally (1-2 times/week)
exercise_frequency	Rarely (a few times/month)
exercise_frequency	Never
exercise_frequency	NA
exercise_location	Both
exercise_location	Outdoors
exercise_location	Indoors
exercise_location	Depends on the season
exercise_location	NA
exercise_location	None of the above
nail_biter	false
nail_biter	true
nail_biter	NA
pool_frequency	Never
pool_frequency	Rarely (a few times/month)
pool_frequency	Occasionally (1-2 times/week)
pool_frequency	Regularly (3-5 times/week)
pool_frequency	NA
pool_frequency	Daily
smoking_frequency	Never
smoking_frequency	Rarely (a few times/month)
smoking_frequency	Daily
smoking_frequency	NA
smoking_frequency	Occasionally (1-2 times/week)
smoking_frequency	Regularly (3-5 times/week)
alcohol_consumption	true
alcohol_consumption	false
alcohol_consumption	NA
alcohol_frequency	Rarely (a few times/month)
alcohol_frequency	Never
alcohol_frequency	Occasionally (1-2 times/week)

Table 1.4: Categorical variable level top 10 *(continued)*

variables	levels
alcohol_frequency	Regularly (3-5 times/week)
alcohol_frequency	Daily
alcohol_frequency	NA
alcohol_types_beercider	false
alcohol_types_beercider	true
alcohol_types_beercider	No
alcohol_types_beercider	Yes
alcohol_types_red_wine	false
alcohol_types_red_wine	true
alcohol_types_sour_beers	false
alcohol_types_sour_beers	true
alcohol_types_spiritshard_alcohol	false
alcohol_types_spiritshard_alcohol	true
alcohol_types_unspecified	true
alcohol_types_unspecified	false
alcohol_types_white_wine	false
alcohol_types_white_wine	true
teethbrushing_frequency	Daily
teethbrushing_frequency	Regularly (3-5 times/week)
teethbrushing_frequency	NA
teethbrushing_frequency	Occasionally (1-2 times/week)
teethbrushing_frequency	Never
teethbrushing_frequency	Rarely (a few times/month)
flossing_frequency	Daily
flossing_frequency	Regularly (3-5 times/week)
flossing_frequency	Rarely (a few times/month)
flossing_frequency	Occasionally (1-2 times/week)
flossing_frequency	Never
flossing_frequency	NA
cosmetics_frequency	Never
cosmetics_frequency	Daily
cosmetics_frequency	Regularly (3-5 times/week)
cosmetics_frequency	Rarely (a few times/month)
cosmetics_frequency	Occasionally (1-2 times/week)
cosmetics_frequency	NA
deodorant_use	I use deodorant
deodorant_use	I do not use deodorant or an antiperspirant
deodorant_use	I use an antiperspirant
deodorant_use	Not sure, but I use some form of deodorant/antiperspirant
deodorant_use	NA
sleep_duration	7-8 hours
sleep_duration	6-7 hours
sleep_duration	8 or more hours
sleep_duration	5-6 hours
sleep_duration	Less than 5 hours
sleep_duration	NA
softener	false
softener	true

Table 1.4: Categorical variable level top 10 *(continued)*

variables	levels
softener	NA
softener	No
softener	Yes
bowel_movement_frequency	NA
bowel_movement_frequency	One
bowel_movement_frequency	Two
bowel_movement_frequency	Less than one
bowel_movement_frequency	Three
bowel_movement_frequency	Four
bowel_movement_frequency	Five or more
bowel_movement_quality	NA
bowel_movement_quality	I tend to have normal formed stool
bowel_movement_quality	I tend to be constipated (have difficulty passing stool)
bowel_movement_quality	I tend to have diarrhea (watery stool)
bowel_movement_quality	I don't know, I do not have a point of reference
bowel_movement_quality	I tend to have normal formed stool - Type 3 and 4
bowel_movement_quality	I tend to be constipated (have difficulty passing stool) - Type 1 and 2
bowel_movement_quality	I tend to have diarrhea (watery stool) - Type 5, 6 and 7
antibiotic_history	I have not taken antibiotics in the past year.
antibiotic_history	Year
antibiotic_history	6 months
antibiotic_history	Month
antibiotic_history	Week
antibiotic_history	NA
flu_vaccine_date	I have not gotten the flu vaccine in the past year.
flu_vaccine_date	Year
flu_vaccine_date	6 months
flu_vaccine_date	Month
flu_vaccine_date	NA
flu_vaccine_date	Week
contraceptive	No
contraceptive	NA
contraceptive	Yes, I am taking the pill
contraceptive	Yes, I use a hormonal IUD (Mirena)
contraceptive	Yes, I use the NuvaRing
contraceptive	Yes, I use an injected contraceptive (DMPA)
contraceptive	Yes, I use a contraceptive patch (Ortho-Evra)
pregnant	false
pregnant	NA
pregnant	true
pregnant	Not sure
weight_change	Remained stable
weight_change	Decreased more than 10 pounds
weight_change	Increased more than 10 pounds
weight_change	NA
tonsils_removed	false
tonsils_removed	true
tonsils_removed	NA

Table 1.4: Categorical variable level top 10 *(continued)*

variables	levels
tonsils_removed	Not sure
appendix_removed	false
appendix_removed	true
appendix_removed	NA
appendix_removed	Not sure
chickenpox	Yes
chickenpox	No
chickenpox	Not sure
chickenpox	NA
acne_medication	false
acne_medication	true
acne_medication	NA
acne_medication_otc	false
acne_medication_otc	true
acne_medication_otc	NA
csection	false
csection	true
csection	Not sure
csection	NA
fed_as_infant	NA
fed_as_infant	Primarily breast milk
fed_as_infant	Primarily infant formula
fed_as_infant	A mixture of breast milk and formula
fed_as_infant	Not sure
add_adhd	NA
add_adhd	I do not have this condition
add_adhd	Diagnosed by a medical professional (doctor, physician assistant)
add_adhd	Self-diagnosed
add_adhd	Diagnosed by an alternative medicine practitioner
alzheimers	NA
alzheimers	I do not have this condition
alzheimers	Diagnosed by a medical professional (doctor, physician assistant)
alzheimers	Self-diagnosed
lung_disease	I do not have this condition
lung_disease	NA
lung_disease	Diagnosed by a medical professional (doctor, physician assistant)
lung_disease	Self-diagnosed
lung_disease	Diagnosed by an alternative medicine practitioner
asd	NA
asd	I do not have this condition
asd	Diagnosed by a medical professional (doctor, physician assistant)
asd	Self-diagnosed
asd	Diagnosed by an alternative medicine practitioner
autoimmune	NA
autoimmune	I do not have this condition
autoimmune	Diagnosed by a medical professional (doctor, physician assistant)
autoimmune	Self-diagnosed
autoimmune	Diagnosed by an alternative medicine practitioner

Table 1.4: Categorical variable level top 10 *(continued)*

variables	levels
fungal_overgrowth	NA
fungal_overgrowth	I do not have this condition
fungal_overgrowth	Diagnosed by an alternative medicine practitioner
fungal_overgrowth	Self-diagnosed
fungal_overgrowth	Diagnosed by a medical professional (doctor, physician assistant)
cdiff	NA
cdiff	I do not have this condition
cdiff	Diagnosed by a medical professional (doctor, physician assistant)
cdiff	Self-diagnosed
cdiff	Diagnosed by an alternative medicine practitioner
cardiovascular_disease	NA
cardiovascular_disease	I do not have this condition
cardiovascular_disease	Diagnosed by a medical professional (doctor, physician assistant)
cardiovascular_disease	Self-diagnosed
cardiovascular_disease	Diagnosed by an alternative medicine practitioner
diabetes	I do not have this condition
diabetes	NA
diabetes	Diagnosed by a medical professional (doctor, physician assistant)
diabetes	Self-diagnosed
diabetes	Diagnosed by an alternative medicine practitioner
epilepsy_or_seizure_disorder	NA
epilepsy_or_seizure_disorder	I do not have this condition
epilepsy_or_seizure_disorder	Diagnosed by a medical professional (doctor, physician assistant)
epilepsy_or_seizure_disorder	Self-diagnosed
epilepsy_or_seizure_disorder	Diagnosed by an alternative medicine practitioner
ibs	NA
ibs	I do not have this condition
ibs	Diagnosed by a medical professional (doctor, physician assistant)
ibs	Self-diagnosed
ibs	Diagnosed by an alternative medicine practitioner
ibd	I do not have this condition
ibd	NA
ibd	Diagnosed by a medical professional (doctor, physician assistant)
ibd	Self-diagnosed
ibd	Diagnosed by an alternative medicine practitioner
migraine	I do not have this condition
migraine	NA
migraine	Diagnosed by a medical professional (doctor, physician assistant)
migraine	Self-diagnosed
migraine	Diagnosed by an alternative medicine practitioner
kidney_disease	NA
kidney_disease	I do not have this condition
kidney_disease	Diagnosed by a medical professional (doctor, physician assistant)
kidney_disease	Self-diagnosed
kidney_disease	Diagnosed by an alternative medicine practitioner
liver_disease	NA
liver_disease	I do not have this condition
liver_disease	Diagnosed by a medical professional (doctor, physician assistant)

Table 1.4: Categorical variable level top 10 *(continued)*

variables	levels
liver_disease	Diagnosed by an alternative medicine practitioner
liver_disease	Self-diagnosed
pku	I do not have this condition
pku	NA
pku	Diagnosed by a medical professional (doctor, physician assistant)
pku	Diagnosed by an alternative medicine practitioner
pku	Self-diagnosed
sibo	NA
sibo	I do not have this condition
sibo	Self-diagnosed
sibo	Diagnosed by a medical professional (doctor, physician assistant)
sibo	Diagnosed by an alternative medicine practitioner
skin_condition	I do not have this condition
skin_condition	Diagnosed by a medical professional (doctor, physician assistant)
skin_condition	NA
skin_condition	Self-diagnosed
skin_condition	Diagnosed by an alternative medicine practitioner
thyroid	NA
thyroid	I do not have this condition
thyroid	Diagnosed by a medical professional (doctor, physician assistant)
thyroid	Diagnosed by an alternative medicine practitioner
thyroid	Self-diagnosed
seasonal_allergies	false
seasonal_allergies	true
seasonal_allergies	NA
seasonal_allergies	No
seasonal_allergies	Yes
non_food_allergies_beestings	false
non_food_allergies_beestings	true
non_food_allergies_drug_eg_penicillin	false
non_food_allergies_drug_eg_penicillin	true
non_food_allergies_pet_dander	false
non_food_allergies_pet_dander	true
non_food_allergies_poison_ivyoak	false
non_food_allergies_poison_ivyoak	true
non_food_allergies_sun	false
non_food_allergies_sun	true
non_food_allergies_unspecified	true
non_food_allergies_unspecified	false
lactose	false
lactose	true
lactose	NA
gluten	No
gluten	NA
gluten	I do not eat gluten because it makes me feel bad
gluten	I was diagnosed with gluten allergy (anti-gluten IgG), but not celiac d
gluten	I was diagnosed with celiac disease
allergic_to_i_have_no_food_allergies_that_i_know_of	false

Table 1.4: Categorical variable level top 10 *(continued)*

variables	levels
allergic_to_i_have_no_food_allergies_that_i_know_of	true
allergic_to_other	false
allergic_to_other	true
allergic_to_peanuts	false
allergic_to_peanuts	true
allergic_to_shellfish	false
allergic_to_shellfish	true
allergic_to_tree_nuts	false
allergic_to_tree_nuts	true
allergic_to_unspecified	true
allergic_to_unspecified	false
breastmilk_formula_ensure	NA
breastmilk_formula_ensure	false
breastmilk_formula_ensure	true
breastmilk_formula_ensure	I eat both solid food and formula/breast milk
meat_eggs_frequency	NA
meat_eggs_frequency	Regularly (3-5 times/week)
meat_eggs_frequency	Daily
meat_eggs_frequency	Occasionally (1-2 times/week)
meat_eggs_frequency	Rarely (less than once/week)
meat_eggs_frequency	Never
homecooked_meals_frequency	NA
homecooked_meals_frequency	Daily
homecooked_meals_frequency	Regularly (3-5 times/week)
homecooked_meals_frequency	Occasionally (1-2 times/week)
homecooked_meals_frequency	Rarely (less than once/week)
homecooked_meals_frequency	Never
ready_to_eat_meals_frequency	NA
ready_to_eat_meals_frequency	Never
ready_to_eat_meals_frequency	Rarely (less than once/week)
ready_to_eat_meals_frequency	Occasionally (1-2 times/week)
ready_to_eat_meals_frequency	Regularly (3-5 times/week)
ready_to_eat_meals_frequency	Daily
prepared_meals_frequency	NA
prepared_meals_frequency	Rarely (less than once/week)
prepared_meals_frequency	Occasionally (1-2 times/week)
prepared_meals_frequency	Regularly (3-5 times/week)
prepared_meals_frequency	Never
prepared_meals_frequency	Daily
whole_grain_frequency	NA
whole_grain_frequency	Regularly (3-5 times/week)
whole_grain_frequency	Occasionally (1-2 times/week)
whole_grain_frequency	Rarely (less than once/week)
whole_grain_frequency	Daily
whole_grain_frequency	Never
fruit_frequency	NA
fruit_frequency	Regularly (3-5 times/week)
fruit_frequency	Daily

Table 1.4: Categorical variable level top 10 *(continued)*

variables	levels
fruit_frequency	Occasionally (1-2 times/week)
fruit_frequency	Rarely (less than once/week)
fruit_frequency	Never
vegetable_frequency	NA
vegetable_frequency	Daily
vegetable_frequency	Regularly (3-5 times/week)
vegetable_frequency	Occasionally (1-2 times/week)
vegetable_frequency	Rarely (less than once/week)
vegetable_frequency	Never
types_of_plants	NA
types_of_plants	11 to 20
types_of_plants	6 to 10
types_of_plants	21 to 30
types_of_plants	More than 30
types_of_plants	Less than 5
fermented_plant_frequency	NA
fermented_plant_frequency	Rarely (less than once/week)
fermented_plant_frequency	Never
fermented_plant_frequency	Occasionally (1-2 times/week)
fermented_plant_frequency	Regularly (3-5 times/week)
fermented_plant_frequency	Daily
milk_cheese_frequency	NA
milk_cheese_frequency	Regularly (3-5 times/week)
milk_cheese_frequency	Occasionally (1-2 times/week)
milk_cheese_frequency	Rarely (less than once/week)
milk_cheese_frequency	Never
milk_cheese_frequency	Daily
milk_substitute_frequency	NA
milk_substitute_frequency	Never
milk_substitute_frequency	Rarely (less than once/week)
milk_substitute_frequency	Daily
milk_substitute_frequency	Regularly (3-5 times/week)
milk_substitute_frequency	Occasionally (1-2 times/week)
frozen_dessert_frequency	NA
frozen_dessert_frequency	Rarely (less than once/week)
frozen_dessert_frequency	Never
frozen_dessert_frequency	Occasionally (1-2 times/week)
frozen_dessert_frequency	Regularly (3-5 times/week)
frozen_dessert_frequency	Daily
red_meat_frequency	NA
red_meat_frequency	Occasionally (1-2 times/week)
red_meat_frequency	Regularly (3-5 times/week)
red_meat_frequency	Rarely (less than once/week)
red_meat_frequency	Never
red_meat_frequency	Daily
high_fat_red_meat_frequency	NA
high_fat_red_meat_frequency	Rarely (less than once/week)
high_fat_red_meat_frequency	Occasionally (1-2 times/week)

Table 1.4: Categorical variable level top 10 *(continued)*

variables	levels
high_fat_red_meat_frequency	Never
high_fat_red_meat_frequency	Regularly (3-5 times/week)
high_fat_red_meat_frequency	Daily
poultry_frequency	NA
poultry_frequency	Occasionally (1-2 times/week)
poultry_frequency	Regularly (3-5 times/week)
poultry_frequency	Rarely (less than once/week)
poultry_frequency	Never
poultry_frequency	Daily
seafood_frequency	NA
seafood_frequency	Occasionally (1-2 times/week)
seafood_frequency	Rarely (less than once/week)
seafood_frequency	Regularly (3-5 times/week)
seafood_frequency	Never
seafood_frequency	Daily
salted_snacks_frequency	NA
salted_snacks_frequency	Rarely (less than once/week)
salted_snacks_frequency	Occasionally (1-2 times/week)
salted_snacks_frequency	Regularly (3-5 times/week)
salted_snacks_frequency	Never
salted_snacks_frequency	Daily
sugary_sweets_frequency	NA
sugary_sweets_frequency	Occasionally (1-2 times/week)
sugary_sweets_frequency	Rarely (less than once/week)
sugary_sweets_frequency	Regularly (3-5 times/week)
sugary_sweets_frequency	Daily
sugary_sweets_frequency	Never
olive_oil	NA
olive_oil	Regularly (3-5 times/week)
olive_oil	Daily
olive_oil	Occasionally (1-2 times/week)
olive_oil	Rarely (less than once/week)
olive_oil	Never
whole_eggs	NA
whole_eggs	Occasionally (1-2 times/week)
whole_eggs	Regularly (3-5 times/week)
whole_eggs	Rarely (less than once/week)
whole_eggs	Daily
whole_eggs	Never
sugar_sweetened_drink_frequency	NA
sugar_sweetened_drink_frequency	Never
sugar_sweetened_drink_frequency	Rarely (less than once/week)
sugar_sweetened_drink_frequency	Occasionally (1-2 times/week)
sugar_sweetened_drink_frequency	Regularly (3-5 times/week)
sugar_sweetened_drink_frequency	Daily
one_liter_of_water_a_day_frequency	NA
one_liter_of_water_a_day_frequency	Daily
one_liter_of_water_a_day_frequency	Regularly (3-5 times/week)

Table 1.4: Categorical variable level top 10 *(continued)*

variables	levels
one_liter_of_water_a_day_frequency	Occasionally (1-2 times/week)
one_liter_of_water_a_day_frequency	Rarely (less than once/week)
one_liter_of_water_a_day_frequency	Never

1.2.2 Diagnosis of numerical variables

Table 1.5: General list of numerical diagnosis

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
subject_id	1,000	2,593.50	4,285.984	4,359.00	5,951.500	7,573	0	0	0
age_years	1	35.00	46.523	48.00	60.000	101	0	0	1
bmi	0	20.82	2,560.858	23.31	26.575	1,130,000	1	0	496
height_cm	1	162.00	168.859	170.00	178.000	16,540	0	0	570
weight_kg	2	58.00	69.769	68.00	79.000	932	0	0	354

1.2.3 List of numerical diagnosis (zero)

Table 1.6: List of numerical diagnosis (zero)

variables	min	median	max	zero	zero ratio(%)
bmi	0	23.31	1,130,000	1	0.018

1.2.4 List of numerical diagnosis (minus)

No numeric variable with negative value

Chapter 2

Diagnose Outliers

2.1 Overview of Diagnosis

2.1.1 Diagnosis of numerical variable outliers

Table 2.1: Diagnosis of numerical variable outliers

variables	min	median	max	outlier	outlier ratio(%)
height_cm	1	170.00	16,540	570	10.166
bmi	0	23.31	1,130,000	496	8.846
weight_kg	2	68.00	932	354	6.314
age_years	1	48.00	101	1	0.018

2.2 Detailed outliers diagnosis

variable : height_cm

Table 2.2: Outliers information of height_cm

Measures	Values
Outliers count	570.00
Outliers ratio (%)	10.17
Mean of outliers	150.15
Mean with outliers	168.86
Mean without outliers	170.98

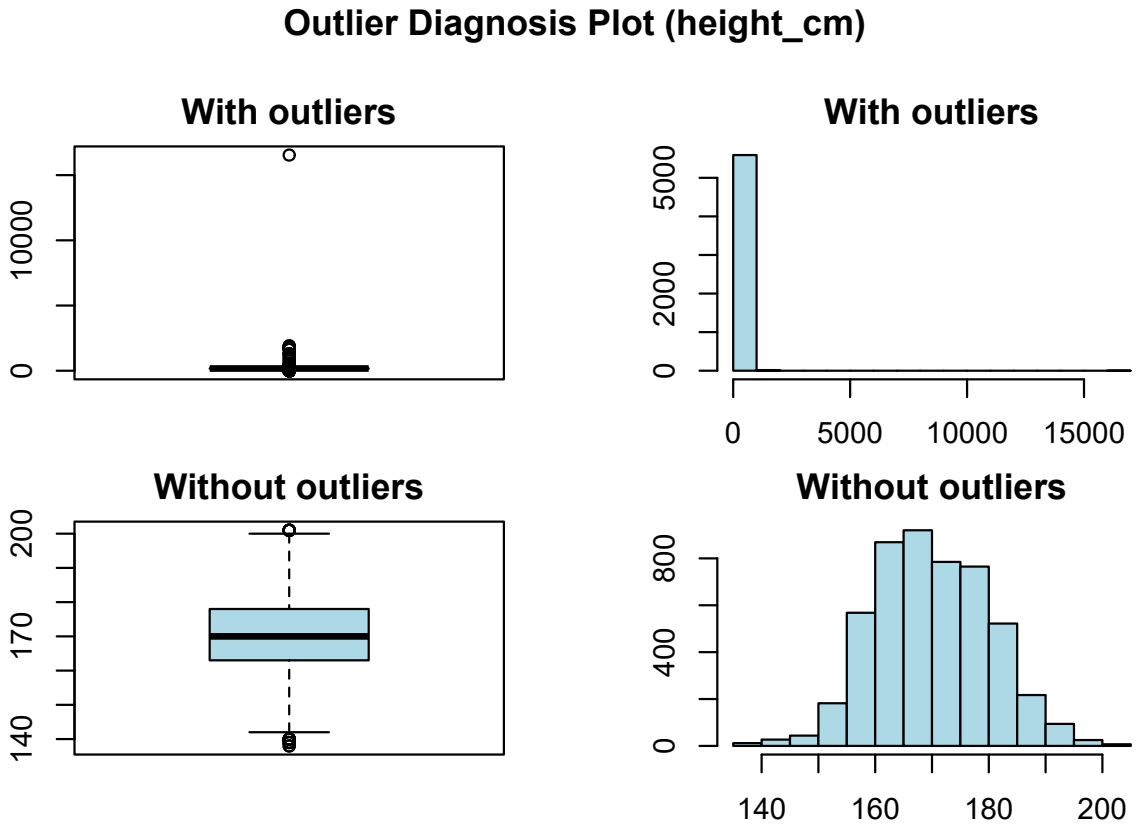

Figure 2.1: Distribution of height_cm

variable : bmi

Table 2.3: Outliers information of bmi

Measures	Values
Outliers count	496.00
Outliers ratio (%)	8.85
Mean of outliers	28,708.85
Mean with outliers	2,560.86
Mean without outliers	23.31

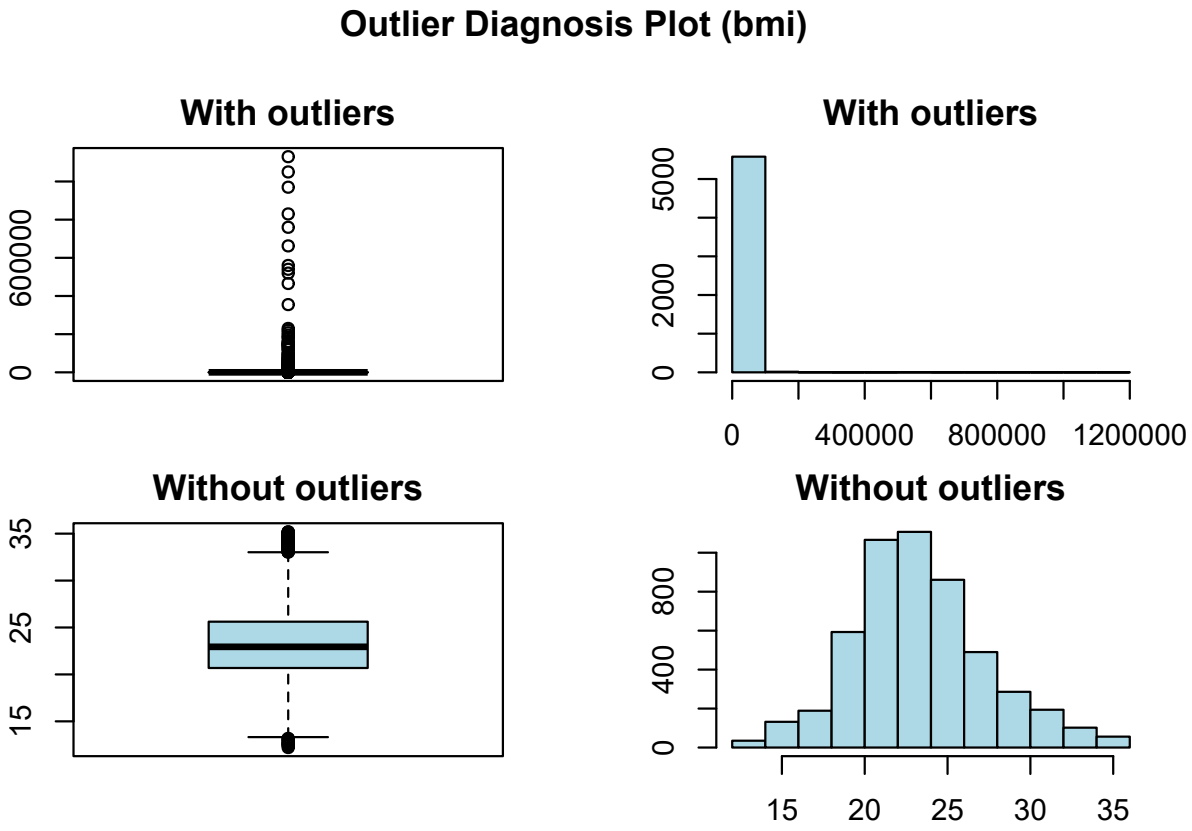

Figure 2.2: Distribution of bmi

variable : weight_kg

Table 2.4: Outliers information of weight_kg

Measures	Values
Outliers count	354.00
Outliers ratio (%)	6.31
Mean of outliers	74.03
Mean with outliers	69.77
Mean without outliers	69.48

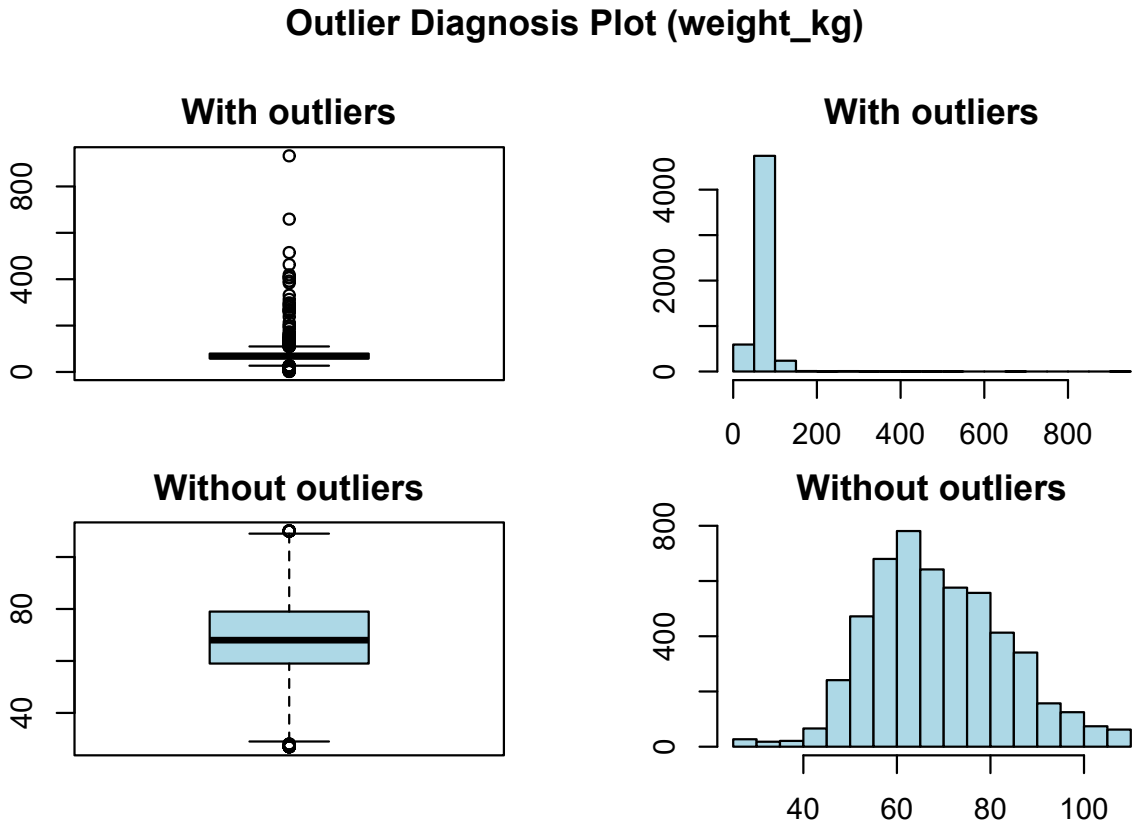

Figure 2.3: Distribution of weight_kg

variable : age_years

Table 2.5: Outliers information of age_years

Measures	Values
Outliers count	1.00
Outliers ratio (%)	0.02
Mean of outliers	101.00
Mean with outliers	46.52
Mean without outliers	46.51

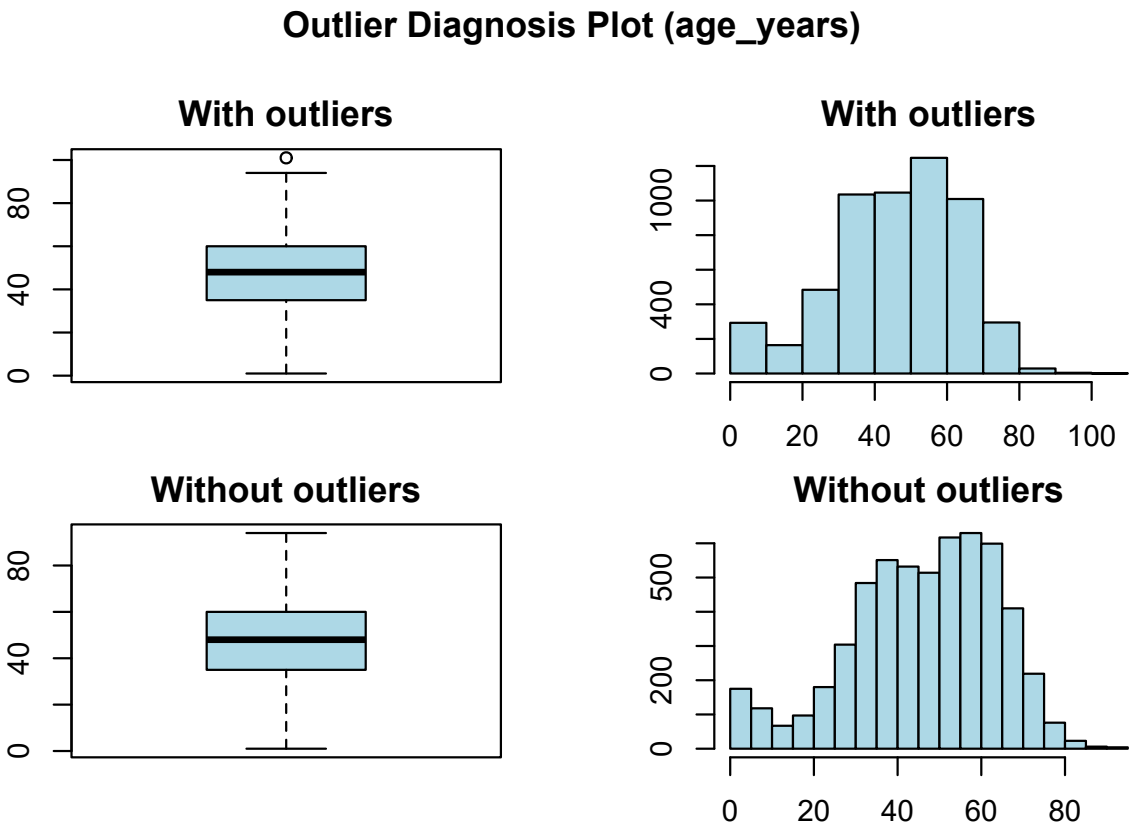

Figure 2.4: Distribution of age_years