HEALTHCARE READMISSIONS
ISE-543 PROJECT

EDA REPORT

OUTLINE

- Dataset overview
- Data quality summary
- Statistical summary of dataset
- Univariate analysis
- Bivariate analysis

EDA REPORT

DATASET OVERVIEW

- "healthcare_readmissions_dataset_train.csv" – dataset with 19 variables and 8,038 observations
- Dataset contains 1 response variables:

Variable	Type	Description
Readmission within 30 Days	Categorical (Binary)	1 = patient was readmitted 0 = patient was not readmitted

- Dataset contains 2 identifiers (not for modeling):

Variable	Type	Description
PatientID	Categorical (Nominal, ID)	Unique patient identifier
Hospital ID	Categorical (Nominal)	Hospital identifier with values "Hosp1", "Hosp2", and "Hosp3"

DATASET OVERVIEW

DEMOGRAPHIC VARIABLES

Demographic variables describe the profile of each patient’s age, gender and ethnicity etc.

Name	Type	Description
Age	Numerical (Discrete)	Patient age
Gender	Categorical (Nominal)	String variable with values “Male” or “Female”
Ethnicity	Categorical (Nominal)	String variable with values “Caucasian”, “Hispanic”, “African American” and “Other”

DATASET OVERVIEW

CLINICAL VARIABLES

Clinical variables describe each patient’s health status, such as body status, disease history and clinical records.

Name	Type	Description
Height(m)	Numerical (Continuous)	Patient height in meters
Smoker	Categorical (Boolean)	Boolean indicating if patient is a current smoker 1=patient is a current smoker 0=patient is not a current smoker
BMI	Numerical (Continuous)	Patient Body Mass Index
Weight(kg)	Numerical (Continuous)	Patient weight in kg
Adjusted Weight(kg)	Numerical (Continuous)	Health system-specific adjustments to patient weight (in kg)
Has Diabetes	Categorical (Binary)	Boolean indicating if patient has diabetes 1=patient has diabetes 0=patient does not have diabetes
Has Hypertension	Categorical (Binary)	Boolean indicating if patient has hypertension 1=patient has hypertension 0=patient does not have hypertension

DATASET OVERVIEW

BEHAVIORAL VARIABLES

Behavioral variables describe each patient's lifestyle such as eating and exercising

Name	Type	Description
Exercise Frequency	Categorical (Ordinal)	String variable with values “None”, “Occasional”, or “Regular”
Diet Type	Categorical (Nominal)	String variable with values “Balanced”, “High-fat”, “Vegetarian”, “Other”

DATASET OVERVIEW

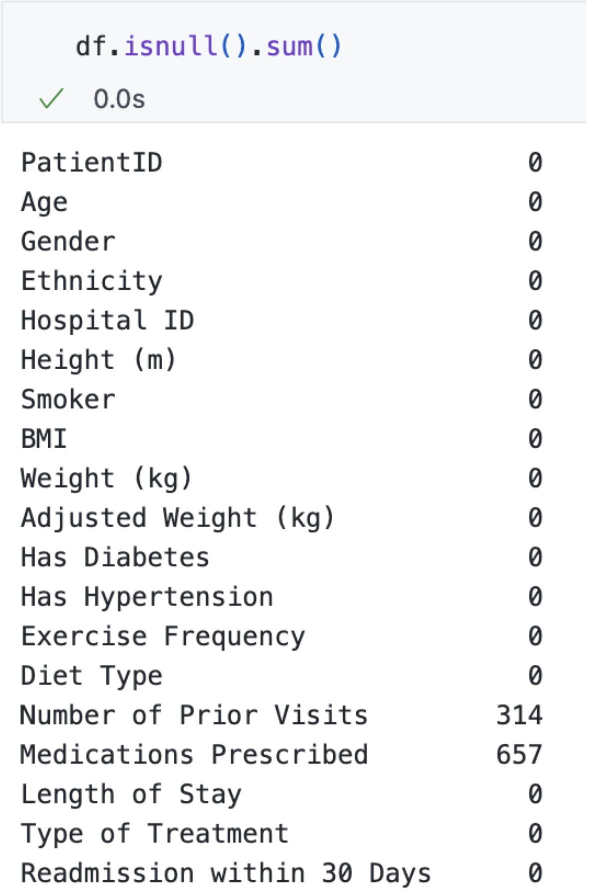
TREATMENT-RELATED VARIABLES

Treatment-related variables describe each patient's medical treatment in the hospital

Name	Type	Description
Number of Prior Visits	Numerical (Discrete)	Number of previous hospitalizations of the patient
Medications Prescribed	Numerical (Discrete)	Number of different prescription medications patient is currently taking
Length of Stay	Numerical (Discrete)	Length of the hospital stay in days
Type of Treatment	Categorical (Nominal)	String variable with values "None", "Minor Surgery", "Major Surgery", "Other Treatment"

DATA QUALITY SUMMARY

Missing values:



```
df.isnull().sum()
```

✓ 0.0s

```
PatientID                   0
Age                         0
Gender                      0
Ethnicity                   0
Hospital ID                 0
Height (m)                  0
Smoker                      0
BMI                         0
Weight (kg)                 0
Adjusted Weight (kg)        0
Has Diabetes                0
Has Hypertension            0
Exercise Frequency          0
Diet Type                   0
Number of Prior Visits    314
Medications Prescribed    657
Length of Stay              0
Type of Treatment           0
Readmission within 30 Days  0
```

Two variables have missing values:

- Number of Prior Visits
- Medications Prescribed

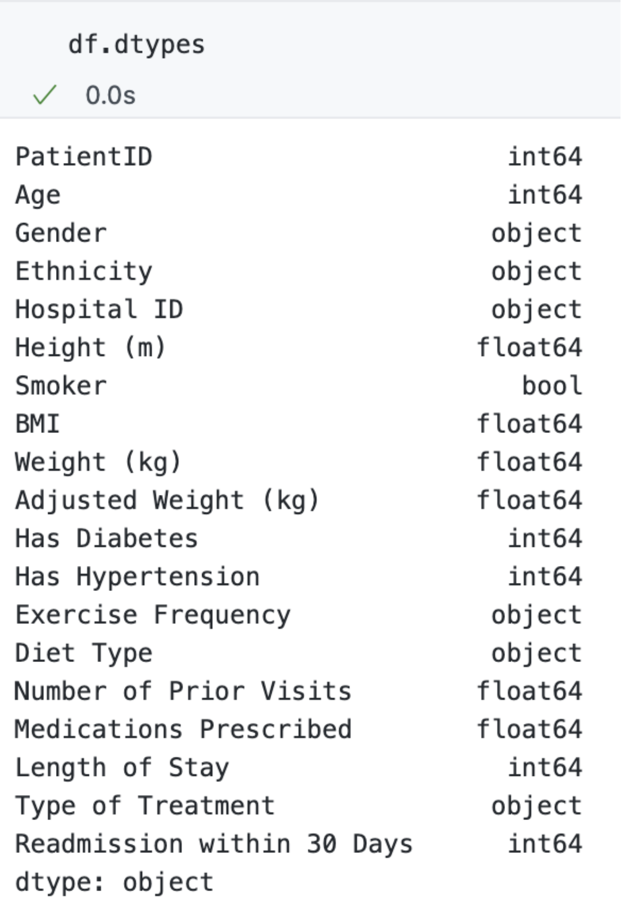
-> As they are discrete and important to training model, it is more appropriate to use mode to impute missing values

USC Viterbi
School of Engineering

EDA REPORT

DATA QUALITY SUMMARY

Data types:



```
df.dtypes
✓ 0.0s
```

```
PatientID                      int64
Age                            int64
Gender                        object
Ethnicity                     object
Hospital ID                   object
Height (m)                   float64
Smoker                          bool
BMI                          float64
Weight (kg)                  float64
Adjusted Weight (kg)         float64
Has Diabetes                   int64
Has Hypertension               int64
Exercise Frequency            object
Diet Type                     object
Number of Prior Visits       float64
Medications Prescribed       float64
Length of Stay                 int64
Type of Treatment             object
Readmission within 30 Days     int64
dtype: object
```

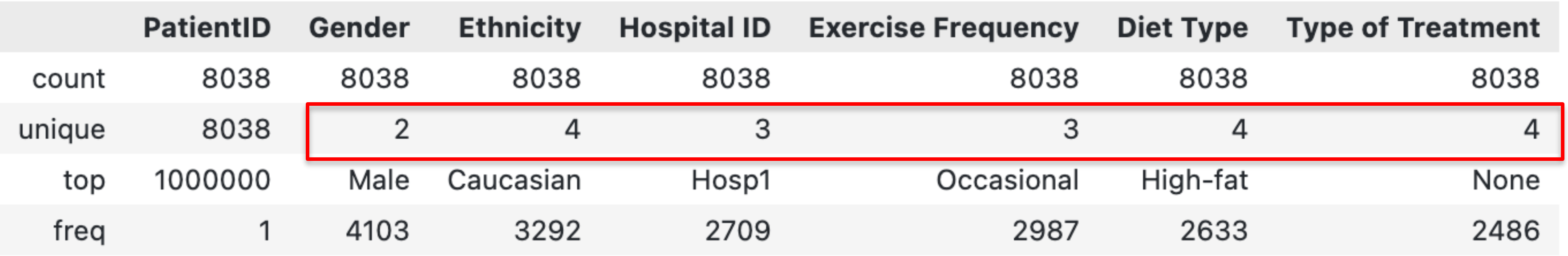
Several string variables:

object variables that must be encoded before modeling

- Gender
- Ethnicity
- Hospital ID
- Exercise Frequency
- Diet Type
- Type of Treatment

DATASET STATISTICAL SUMMARY

CATEGORICAL VARIABLES

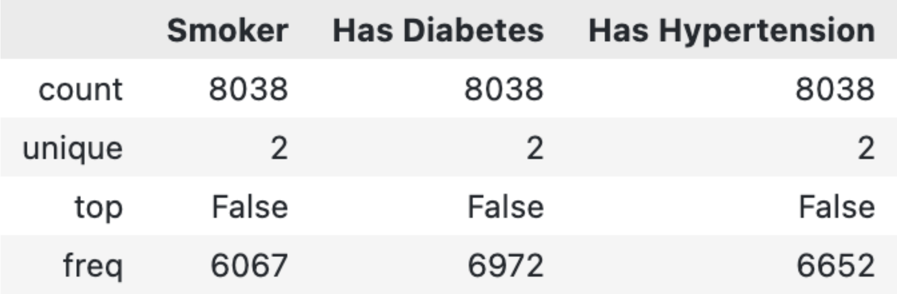


	PatientID	Gender	Ethnicity	Hospital ID	Exercise Frequency	Diet Type	Type of Treatment
count	8038	8038	8038	8038	8038	8038	8038
unique	8038	2	4	3	3	4	4
top	1000000	Male	Caucasian	Hosp1	Occasional	High-fat	None
freq	1	4103	3292	2709	2987	2633	2486

- No high dimensionality data

DATASET STATISTICAL SUMMARY

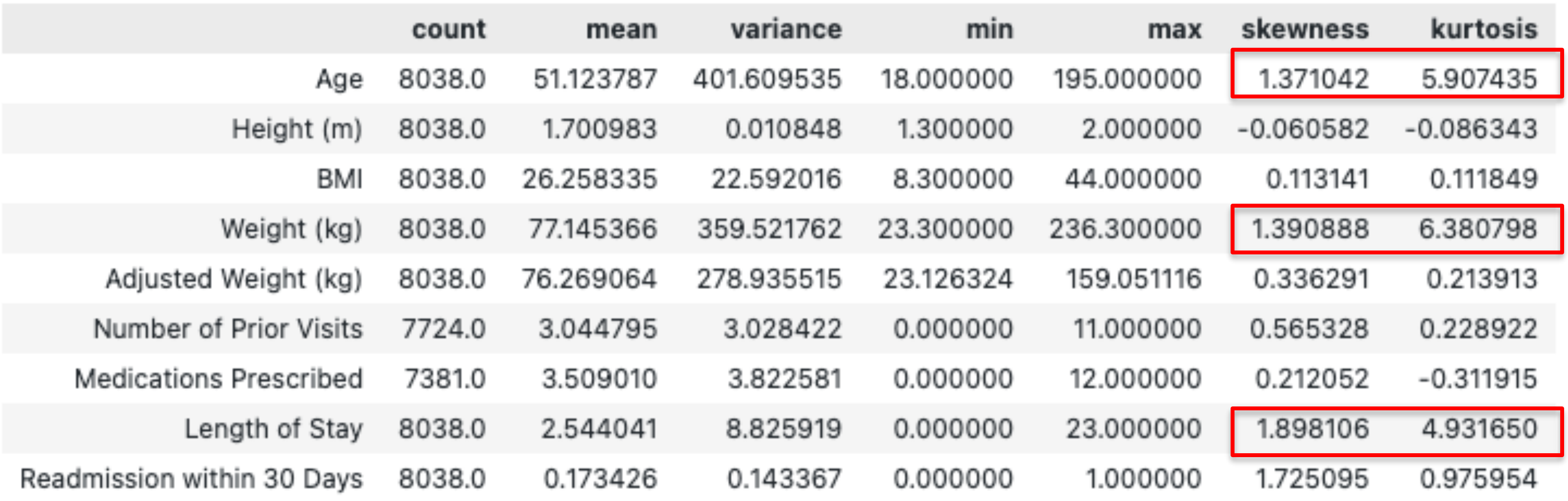
BOOLEAN VARIABLES



	Smoker	Has Diabetes	Has Hypertension
count	8038	8038	8038
unique	2	2	2
top	False	False	False
freq	6067	6972	6652

DATASET STATISTICAL SUMMARY

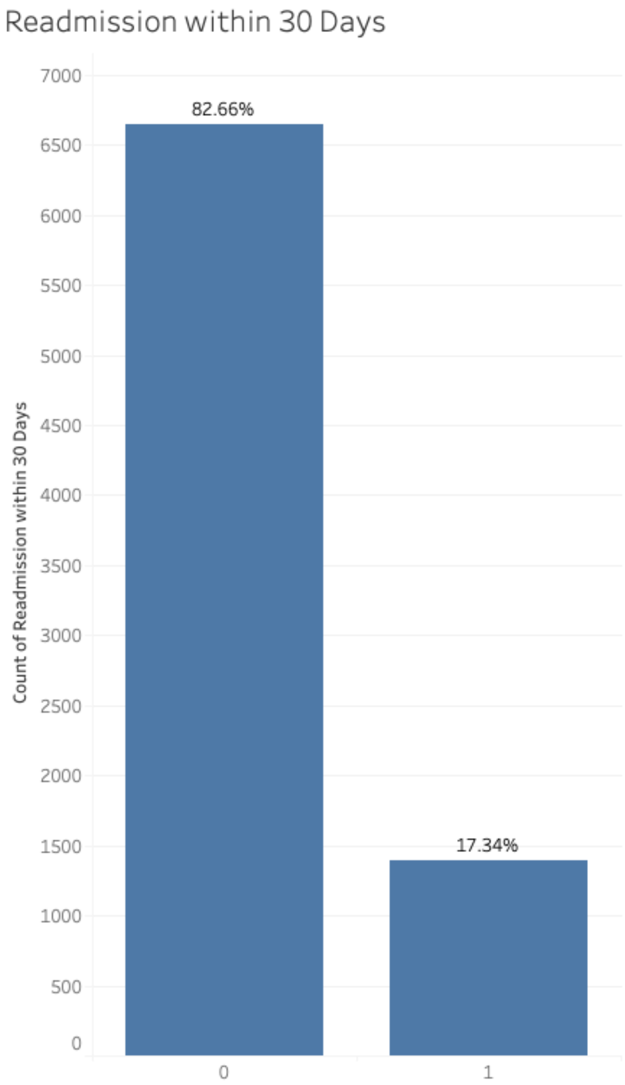
NUMERICAL VARIABLES



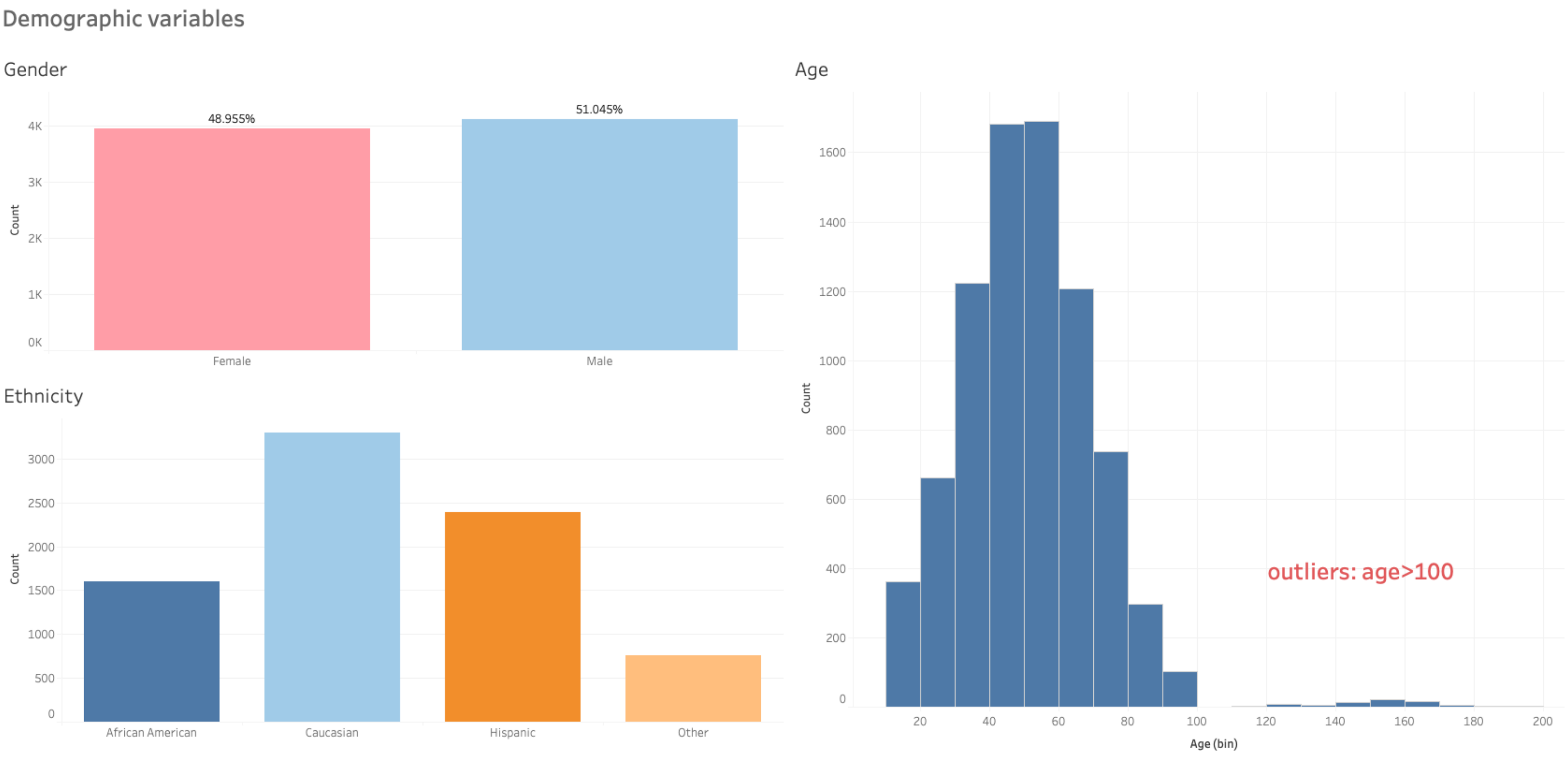
	count	mean	variance	min	max	skewness	kurtosis
Age	8038.0	51.123787	401.609535	18.000000	195.000000	1.371042	5.907435
Height (m)	8038.0	1.700983	0.010848	1.300000	2.000000	-0.060582	-0.086343
BMI	8038.0	26.258335	22.592016	8.300000	44.000000	0.113141	0.111849
Weight (kg)	8038.0	77.145366	359.521762	23.300000	236.300000	1.390888	6.380798
Adjusted Weight (kg)	8038.0	76.269064	278.935515	23.126324	159.051116	0.336291	0.213913
Number of Prior Visits	7724.0	3.044795	3.028422	0.000000	11.000000	0.565328	0.228922
Medications Prescribed	7381.0	3.509010	3.822581	0.000000	12.000000	0.212052	-0.311915
Length of Stay	8038.0	2.544041	8.825919	0.000000	23.000000	1.898106	4.931650
Readmission within 30 Days	8038.0	0.173426	0.143367	0.000000	1.000000	1.725095	0.975954

- Age: Right-skewed with heavily tails
- Weight(kg): Strong right-skew with outliers
- Length of Stay: Very skewed with outliers

EDA REPORT – UNIVARIATE ANALYSIS RESPONSE VARIABLES



EDA REPORT – UNIVARIATE ANALYSIS
DEMOGRAPHIC VARIABLES


Demographic variables

Gender

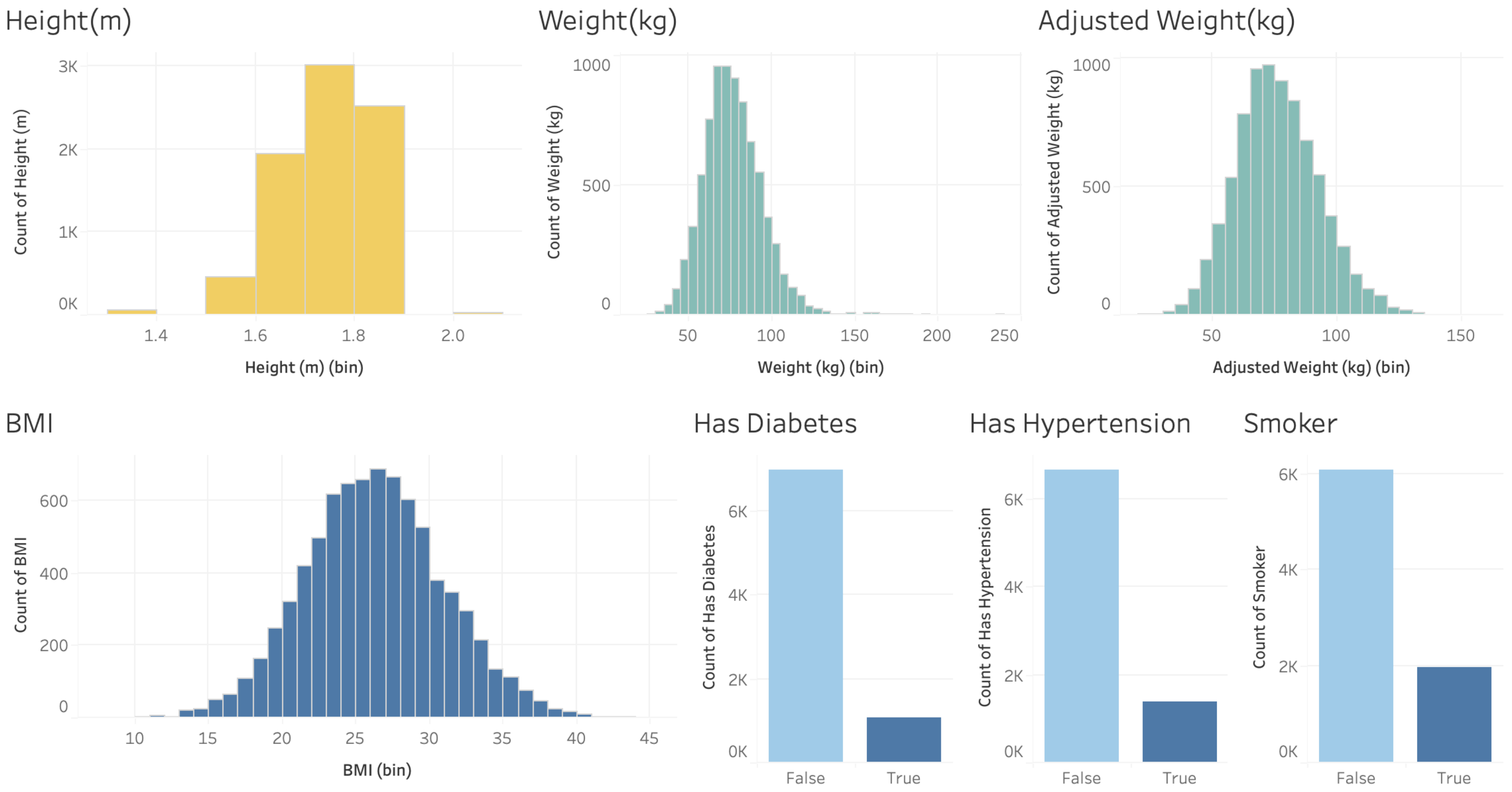
Ethnicity

Age

outliers: age>100

EDA REPORT – UNIVARIATE ANALYSIS
CLINICAL VARIABLES

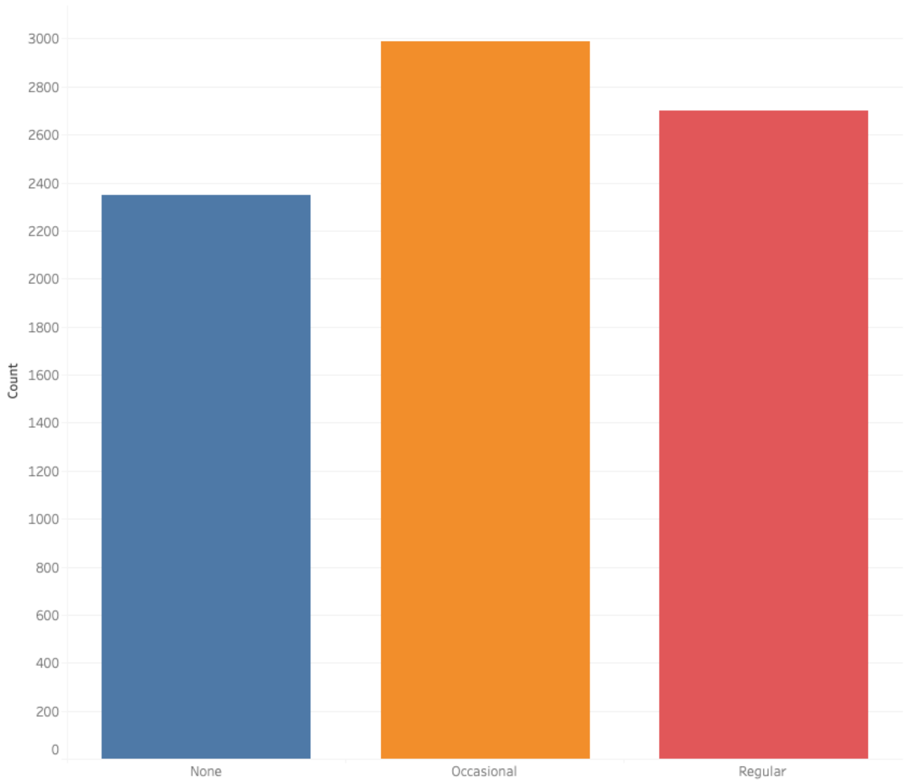
Clinical variables



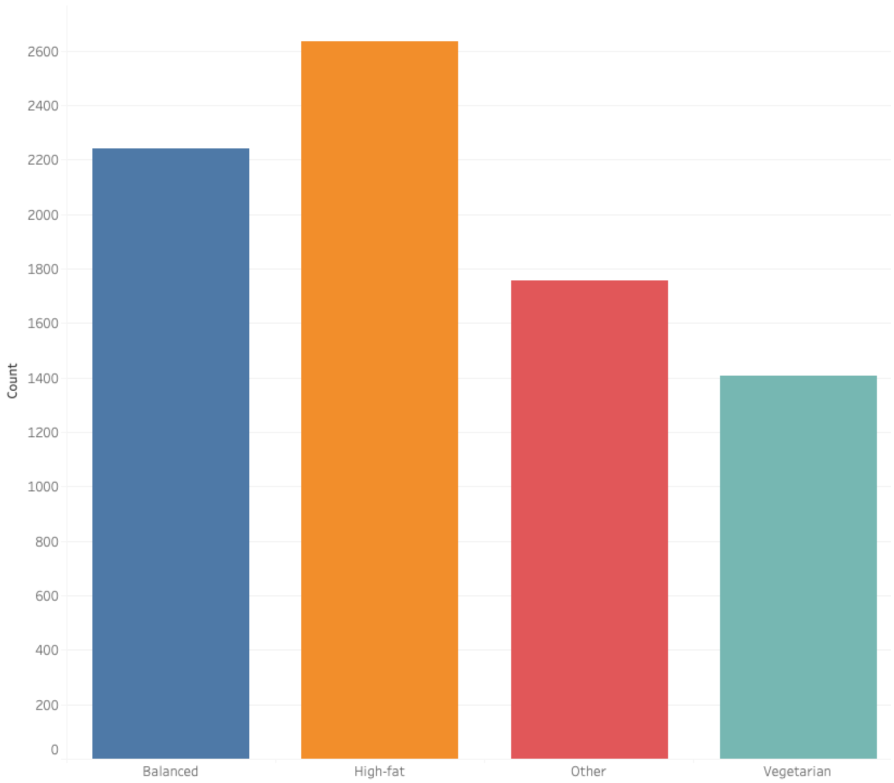
EDA REPORT – UNIVARIATE ANALYSIS
BEHAVIORAL VARIABLES

Behavioral variables

Exercise Frequency



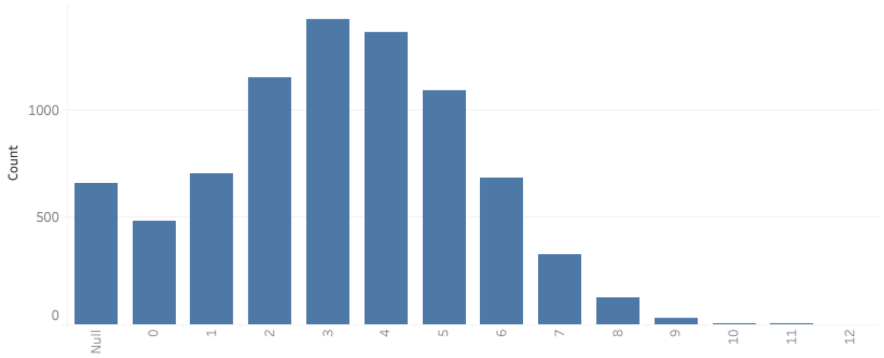
Diet Type



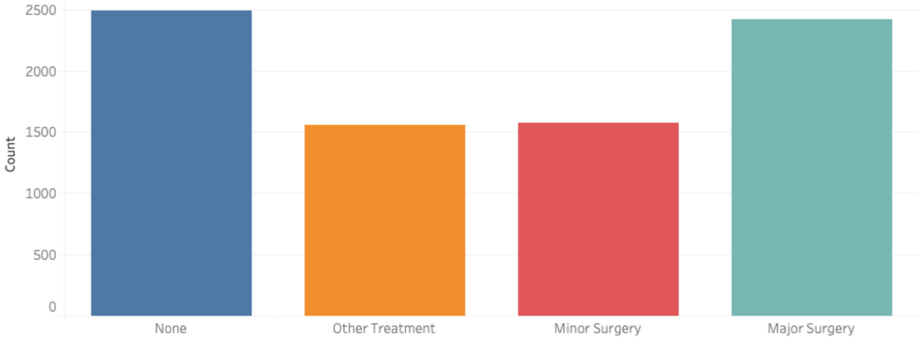
EDA REPORT – UNIVARIATE ANALYSIS
TREATMENT-RELATED VARIABLES

Treatment-related variables

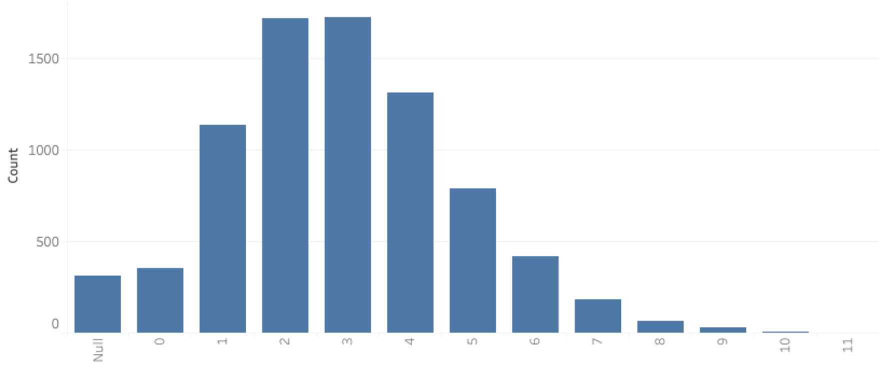
Medications Prescribed



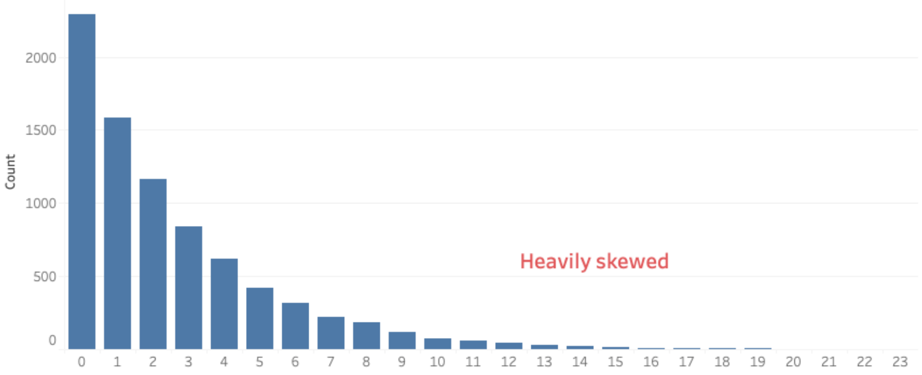
Type of Treatment



Number of Prior Visits



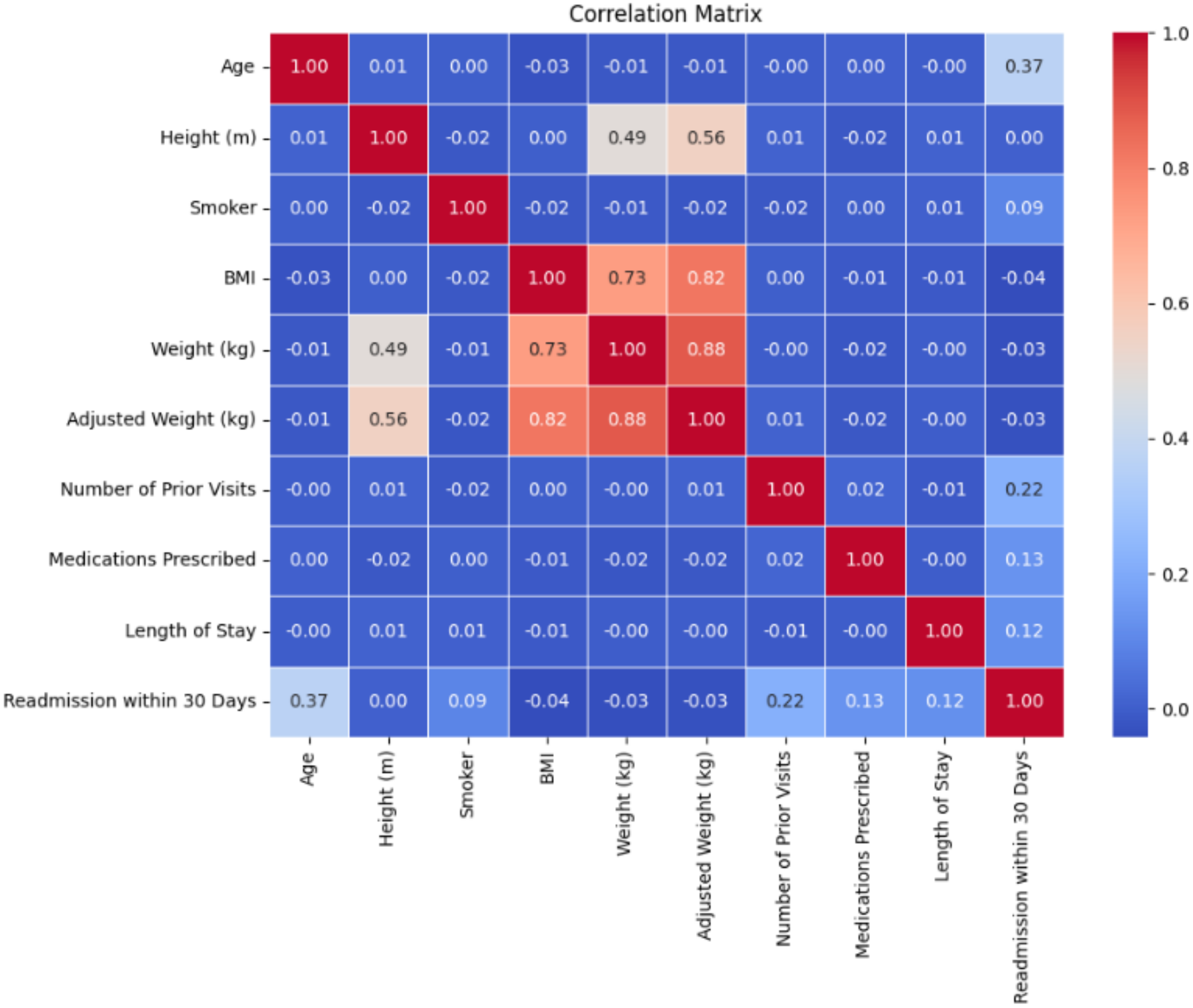
Length of Stay



Heavily skewed

BIVARIATE ANALYSIS
CORRELATION MATRIX

- BMI is highly correlated with Weight and Adjusted Weight



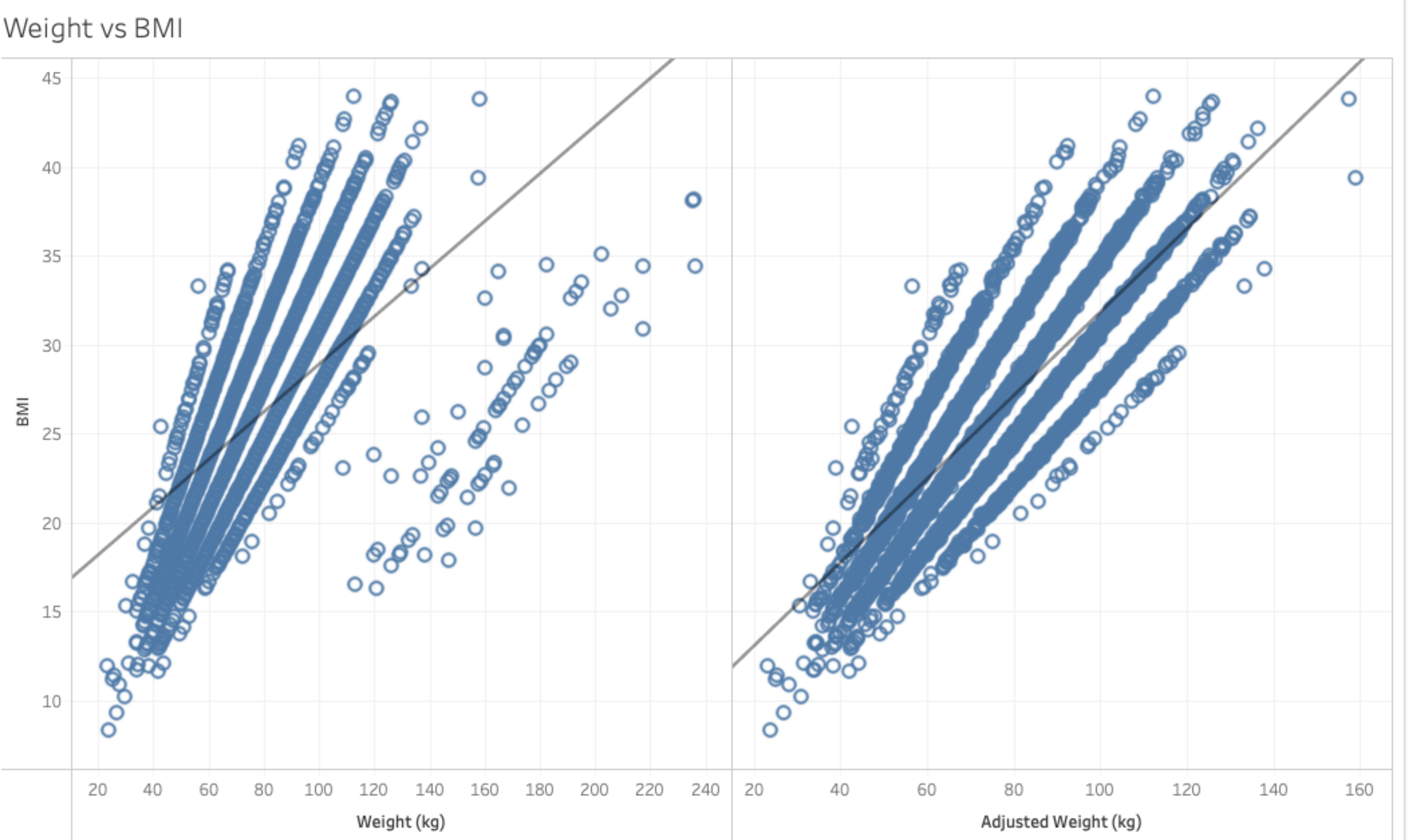
Correlation Matrix

	Age	Height (m)	Smoker	BMI	Weight (kg)	Adjusted Weight (kg)	Number of Prior Visits	Medications Prescribed	Length of Stay	Readmission within 30 Days
Age	1.00	0.01	0.00	-0.03	-0.01	-0.01	-0.00	0.00	-0.00	0.37
Height (m)	0.01	1.00	-0.02	0.00	0.49	0.56	0.01	-0.02	0.01	0.00
Smoker	0.00	-0.02	1.00	-0.02	-0.01	-0.02	-0.02	0.00	0.01	0.09
BMI	-0.03	0.00	-0.02	1.00	0.73	0.82	0.00	-0.01	-0.01	-0.04
Weight (kg)	-0.01	0.49	-0.01	0.73	1.00	0.88	-0.00	-0.02	-0.00	-0.03
Adjusted Weight (kg)	-0.01	0.56	-0.02	0.82	0.88	1.00	0.01	-0.02	-0.00	-0.03
Number of Prior Visits	-0.00	0.01	-0.02	0.00	-0.00	0.01	1.00	0.02	-0.01	0.22
Medications Prescribed	0.00	-0.02	0.00	-0.01	-0.02	-0.02	0.02	1.00	-0.00	0.13
Length of Stay	-0.00	0.01	0.01	-0.01	-0.00	-0.00	-0.01	-0.00	1.00	0.12
Readmission within 30 Days	0.37	0.00	0.09	-0.04	-0.03	-0.03	0.22	0.13	0.12	1.00

BIVARIATE ANALYSIS

WEIGHT - BMI

- BMI is highly correlated with Weight and Adjusted Weight

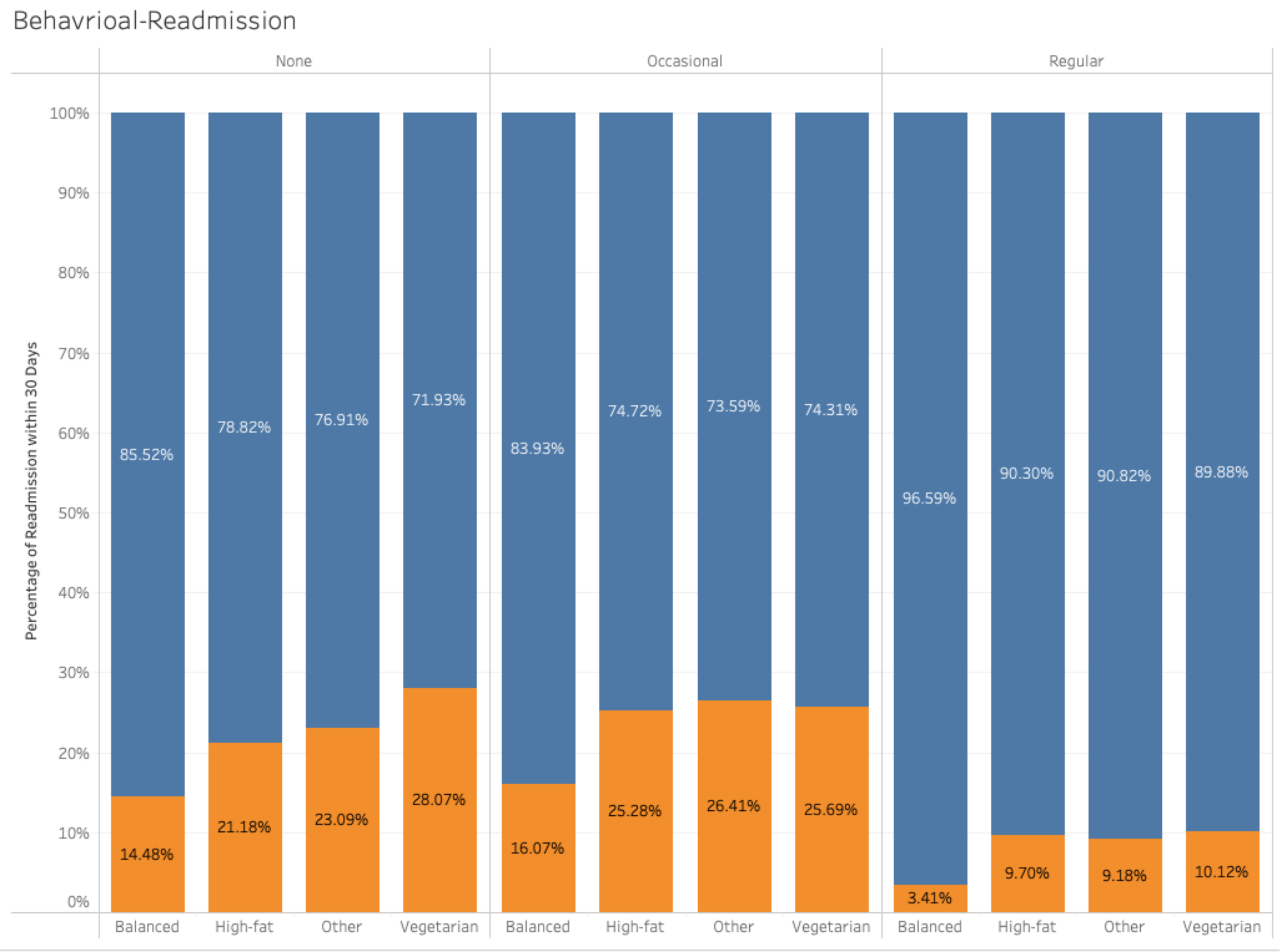


USC Viterbi
School of Engineering

BIVARIATE ANALYSIS

BEHAVIORAL - READMISSIONS

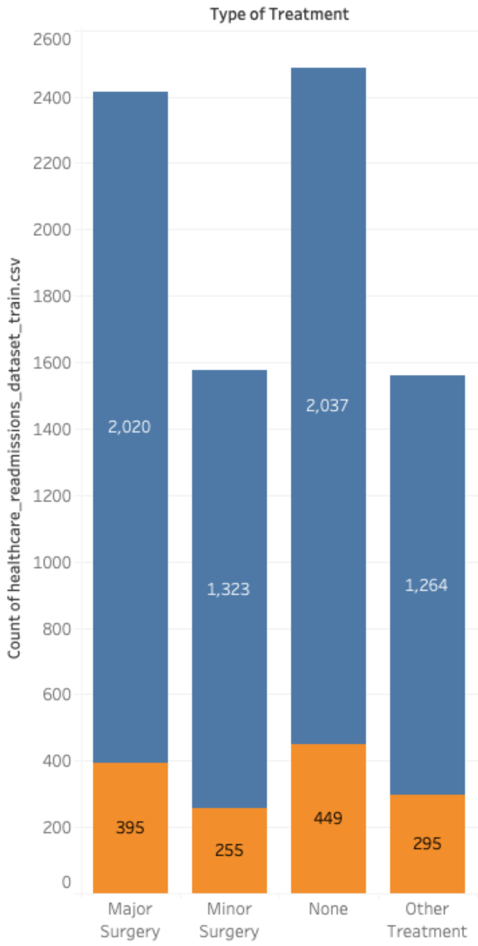
- Regular exercise and balanced diet type have a strong relationship to readmissions

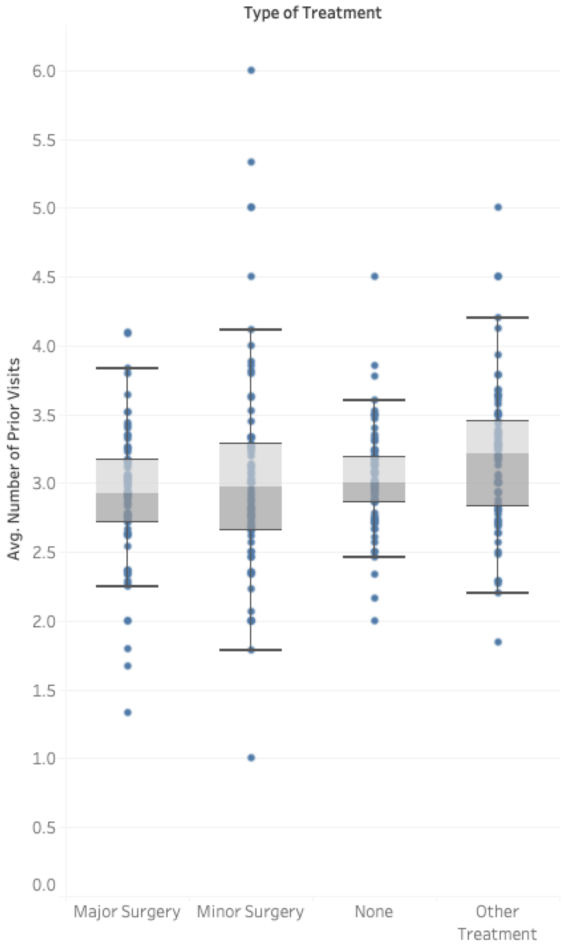


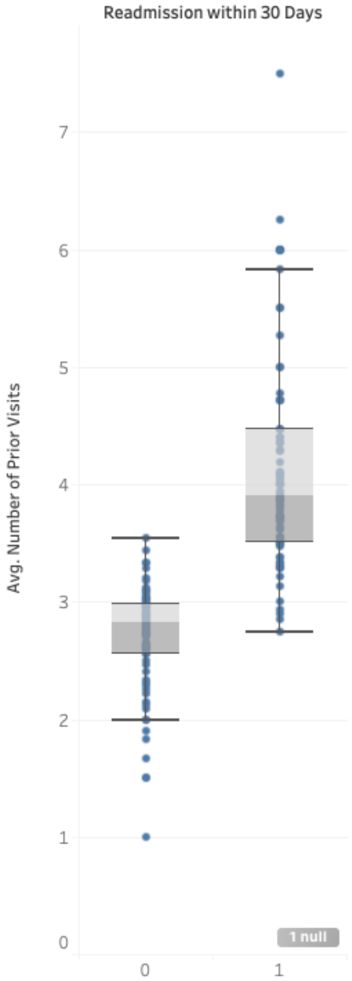
USC Viterbi
School of Engineering

BIVARIATE ANALYSIS
PRIOR VISITS - READMISSIONS

- Patients readmission within 30 Days have a higher number of prior visits
- No obvious different between type of treatment



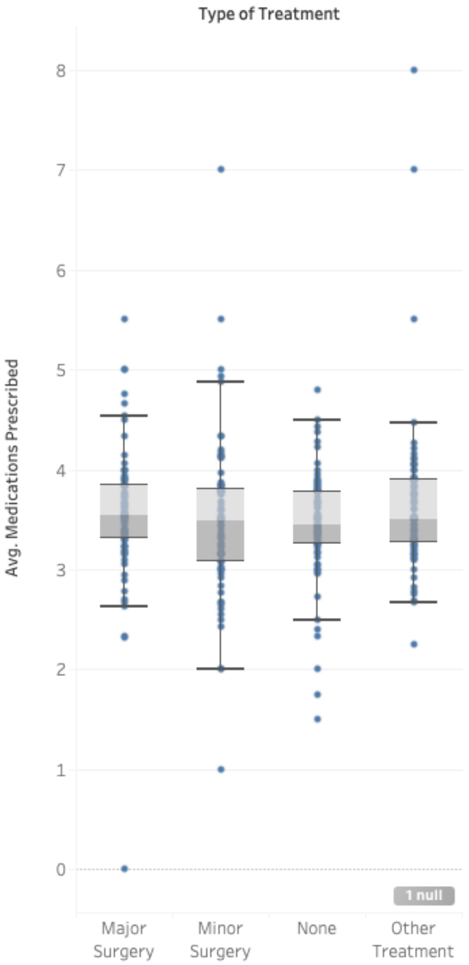


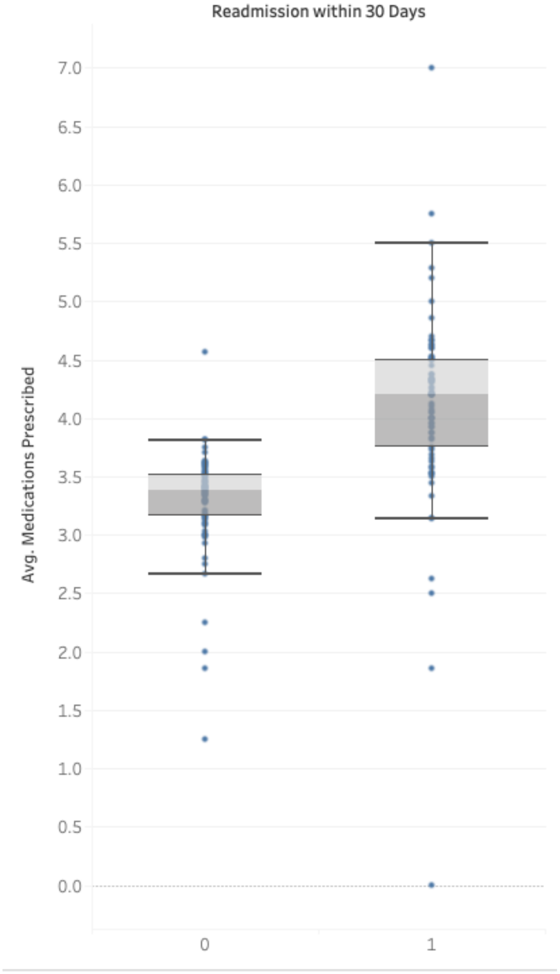


BIVARIATE ANALYSIS

MEDICATIONS PRESCRIBED - READMISSIONS

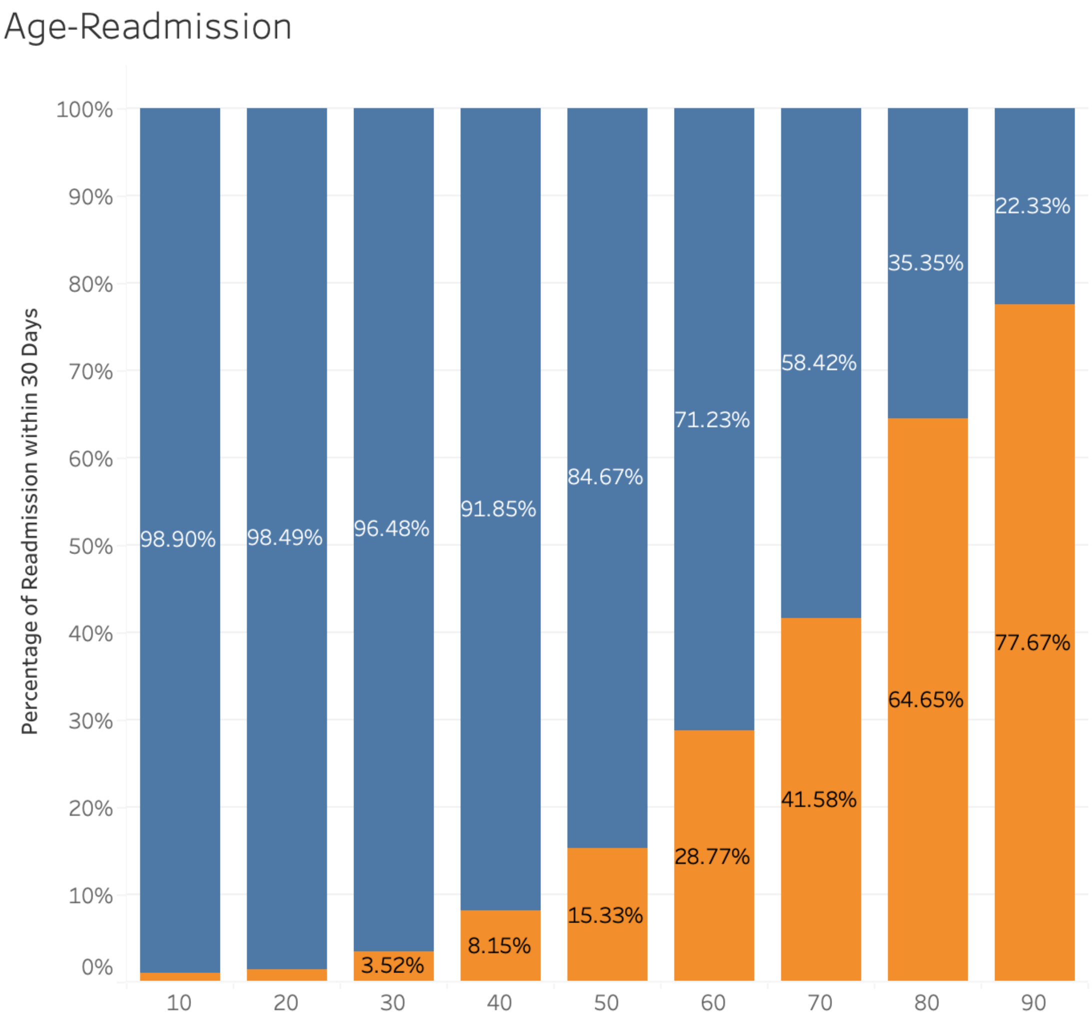
- Patients readmission within 30 Days have a higher number of medications prescribed
- No obvious different between type of treatment





BIVARIATE ANALYSIS
AGE - READMISSION

- Age is an important feature influencing readmissions


Age-Readmission
Percentage of Readmission within 30 Days