

Predicting Chronic Heart Disease using Azure ML Studio

Shan

CONTENT

- Project Goal
- Exploratory Data Analysis/Data Preparation
- Model Development
- Model Deployment

PROJECT GOAL

For this project, I will use Azure ML Studio Designer to build a classification model to predict the likelihood of a patient developing Chronic Heart Disease (CHD) in the coming ten years. The dataset will be using has been distributed with this project and consists of the variables on the following page


DATA DICTIONARY

Variable	Description
Age	age of the participant at the time of examination
Male	gender of the participant (male =1, female = 0)
Education	Educational level of the patient (1 = less than high school, 2 = completed high school or equivalent, 3 = some college, 4= completed college or higher)
Income	Income of the patient
Current Smoker	whether the participant is currently a smoker (yes or no)
Cigarettes per Day	the average number of cigarettes smoked per day by current smokers
BP Meds	whether the participant is taking blood pressure medication (yes or no)
Prevalent Stroke	whether the participant has a history of stroke (yes or no)
Prevalent Hyp	whether the participant has a history of hypertension (yes or no)
Diabetes	whether the participant has diabetes (yes or no)
Total Chol	total cholesterol level in milligrams per deciliter
Sys BP	systolic blood pressure in millimeters of mercury
Dia BP	diastolic blood pressure in millimeters of mercury
BMI	body mass index in kilograms per square meter
Heart Rate	resting heart rate in beats per minute
Glucose	Blood glucose level in milligrams per deciliter
A1c	Hemoglobin A1c (%)
Ten Year CHD	whether the participant developed coronary heart disease (CHD) within 10 years of the examination (yes or no)



EXPLORATORY DATA ANALYSIS/ DATA PREPARATION

ATTRIBUTE SUMMARY

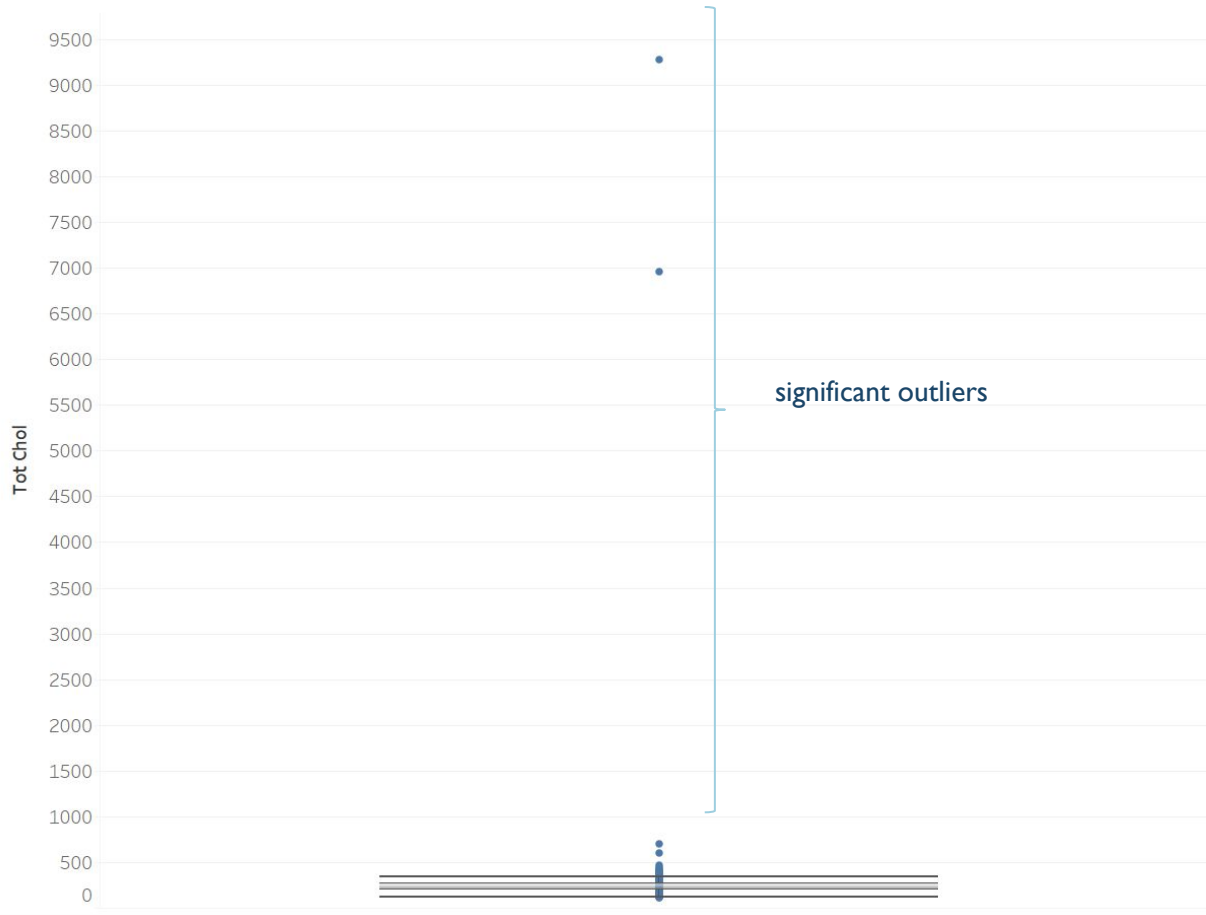
- Response Variable
 - Ten Year CHD
 - Category
 - Male
 - Education
 - Current Smoker
 - BP Meds
 - Prevalent Stroke
 - Prevalent Hyp
 - Diabetes
 - › Measure
 - » Age
 - » Income
 - » Cigarettes per Day
 - » Total Chol
 - » Sys BP
 - » Dia BP
 - » BMI
 - » Heart Rate
 - » Glucose
 - » A1c
- 

DATASET

Feature	Count	Unique Value Count	Missing Value Count	Min	Max	Mean	Mean Deviation	1st quantile	Median	3rd quantile	Mode	Range	Sample Variance	Sample Standard Deviation	Sample Skewness	Sample Kurtosis
patientID	3816	3816	0	100002	999826	554019.06499	222606.258181	336251.75	555421.5	772812.25	(100002, 100012, 100130, 100214, 100754, 101056, 101294, 101524, 102052, 102439, 102712, 102873, 102935, 103840, 103955, 103992, 104276, 104985, 104993, 105054, 105480, 105652, 105670, 106313, 106323, 106806, 106859, 107031, 107211, 107565,	899824	66073719998.68935	257048.088884	-0.005479	-1.179744
	Missing Value															
male	3816	2	0	0	1	0.427673	0.489538	0	0	1	0	1	0.244833	0.494806	0.292498	-1.915449
age	3816	39	0	32	70	49.567348	7.350087	42	49	56	40	38	73.918923	8.597611	0.237195	-0.989866
education	3723	4	93	1.0	4.0	1.974483	0.812462	1	2	3	1.0	3.0	1.037769	1.018709	0.704027	-0.686453
currentSmoker	3816	2	0	0	1	0.489518	0.49978	0	0	1	0	1	0.249956	0.499956	0.041954	-1.999288
cigsPerDay	1841	31	1975	1.0	70.0	18.500272	8.12733	10	20	20	20.0	69.0	119.365353	10.925445	0.761183	0.970104
BPMeds	3771	2	45	0.0	1.0	0.02917	0.056638	0	0	0	0.0	1.0	0.028327	0.168305	5.597927	29.352359
prevalentStroke	3816	2	0	0	1	0.006027	0.011982	0	0	0	0	1	0.005992	0.077411	12.768997	161.131728
prevalentHyp	3816	2	0	0	1	0.306604	0.425196	0	0	1	0	1	0.212654	0.461144	0.839207	-1.296411
diabetes	3816	2	0	0	1	0.024895	0.048551	0	0	0	0	1	0.024282	0.155826	6.101083	35.241683
totChol	3769	246	47	107.0	9280.0	240.852746	39.378735	205	234	263	240.0	9173.0	35695.202567	188.93174	41.03155	1814.986243
sysBP	3816	232	0	83.5	295.0	132.260089	16.890304	117	128	143.5	130.0	211.5	489.279059	22.119653	1.178872	2.298285
diaBP	3816	142	0	50.0	142.5	82.874214	9.158874	75	82	89.5	80.0	92.5	141.791251	11.907613	0.717316	1.232449
BMI	3797	1319	19	15.54	56.8	25.814791	3.113207	23.07	25.4	28.04	22.91	41.26	16.807305	4.099671	0.998709	2.743934
heartRate	3815	73	1	44.0	143.0	75.775098	9.288791	68	75	82	75.0	99.0	144.885431	12.036836	0.656227	0.973439
glucose	3455	134	361	40.0	394.0	81.856151	12.250119	71	78	87	75.0	354.0	555.598004	23.571126	6.424396	63.473602
TenYearCHD	3816	2	0	0	1	0.15173	0.257415	0	0	0	0	1	0.128741	0.358806	1.942295	1.773438
	Outliers															
a1c	3455	3455	361	2.134768768113766	19.917371285750395	4.296312	0.631685	3.738947	4.126325	4.564732	(2.134768768113766, 2.3023530613653507, 2.349277968047848, 2.355203600535006, 2.3861003033933663, 2.38903897942708, 2.4155486171620075, 2.4374983635432623, 2.4946768795983005, 2.51440307446826,	17.78260251763663	1.424309	1.193444	6.218681	60.721181
	High Skew															
income	3816	3282	0	12000.0	524494.0	20355.886792	7743.282613	13562.5	16055	21395.5	14623.0	512494.0	319178936.087836	17865.579646	13.274848	282.066261

DATA CLEANSING - OUTLIER ANALYSIS

Tot Chol Bos Plot



Summary

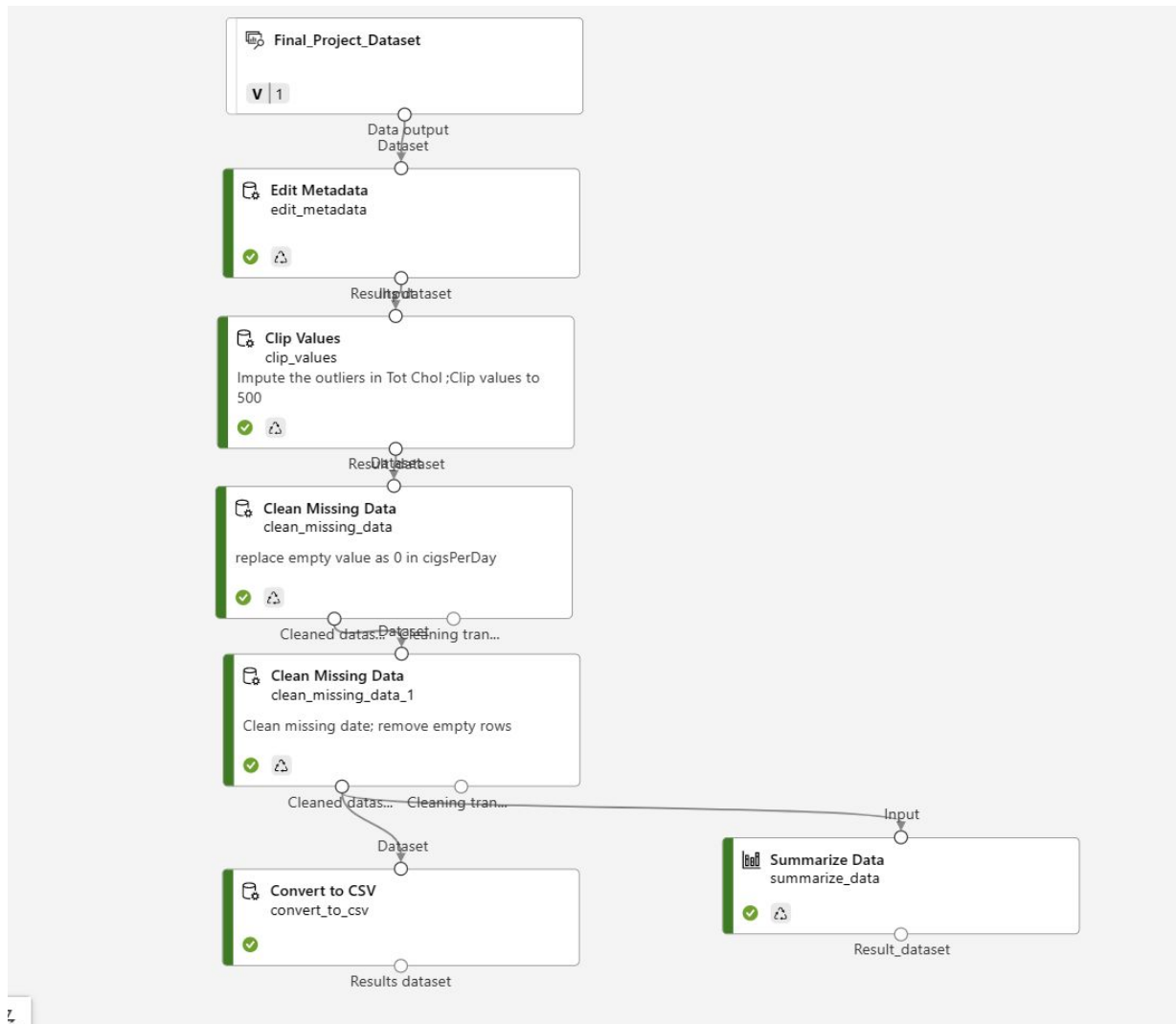
Count:	3769
SUM(Tot Chol)	
Sum:	907,774
Average:	241
Minimum:	107
Maximum:	9,280
Median:	234
Standard deviation:	189
First quartile:	205
Third quartile:	263
Skewness:	41.02
Excess Kurtosis:	1,812.58

DATA CLEAN

- Sets the metadata for the dataset to indicate that TenYearCHD, male, education ,currentSmoker, BPMeds, prevalentStroke, prevalentHyp, diabetes column are categorical variables.
- Impute the outliers in Tot Chol as they appear to clearly be bad data. Clip their values to 500.
- Replace the rows with missing values for 'Cigs Per Day' with 0, which 'Cigs Per Day' is MAR (Missing at Random).
- Delete the rows with missing values for education, BPMeds, Totchol, BMI, HeartRate, glucose, and a1c

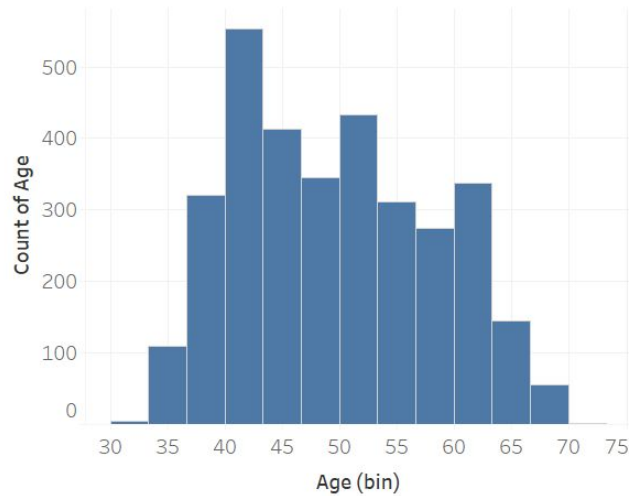


DATA CLEAN

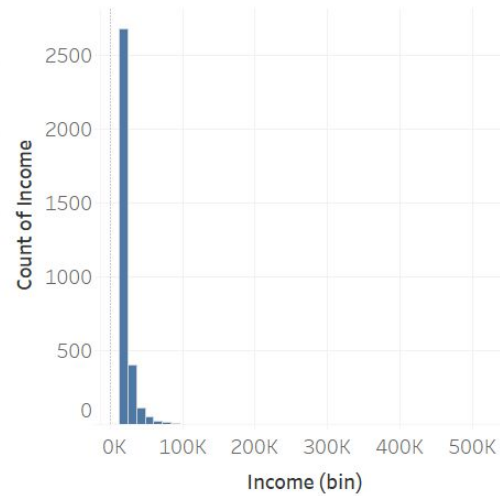


UNIVARIATE ANALYSIS MEASURES

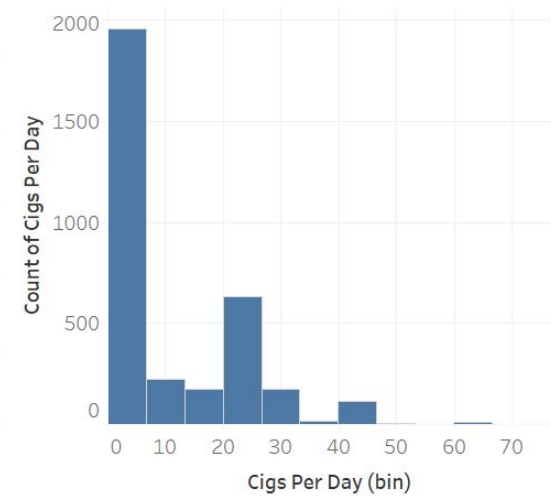
Age



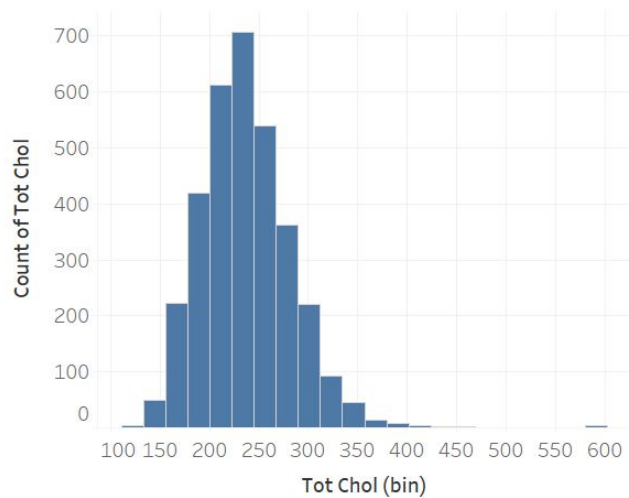
Income



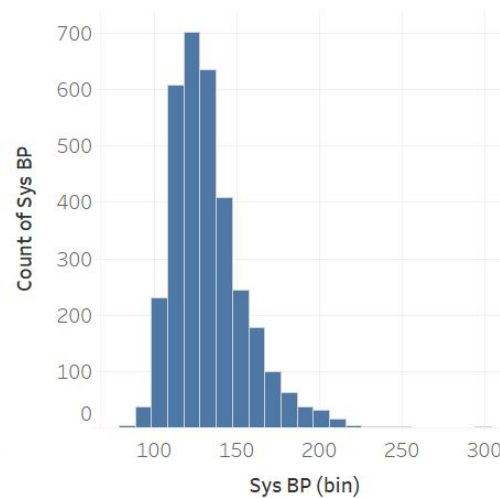
CigsPerDay



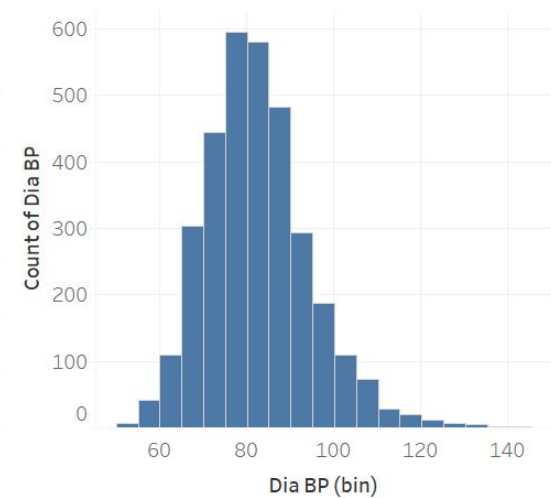
TotChol



SysBP

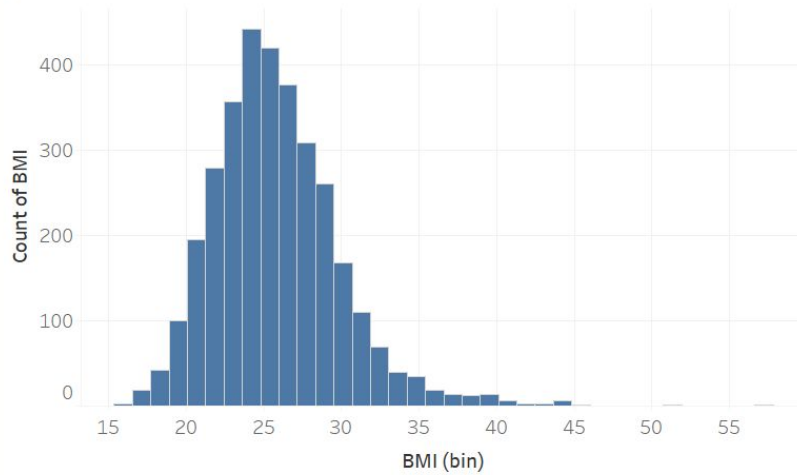


DiaBP

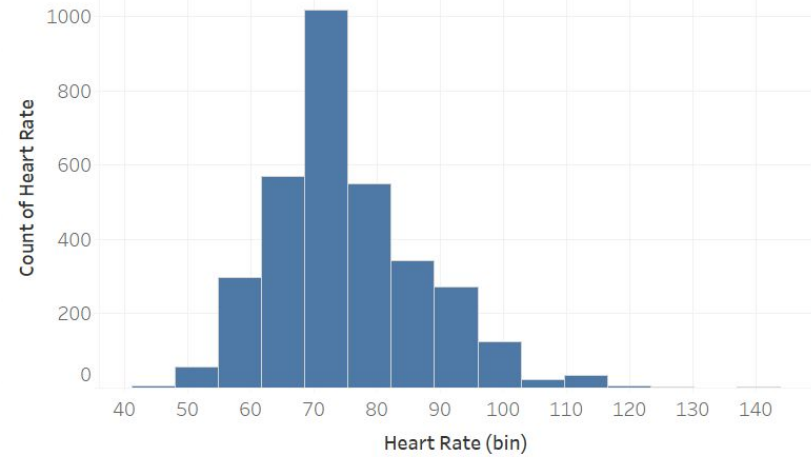


UNIVARIATE ANALYSIS MEASURES

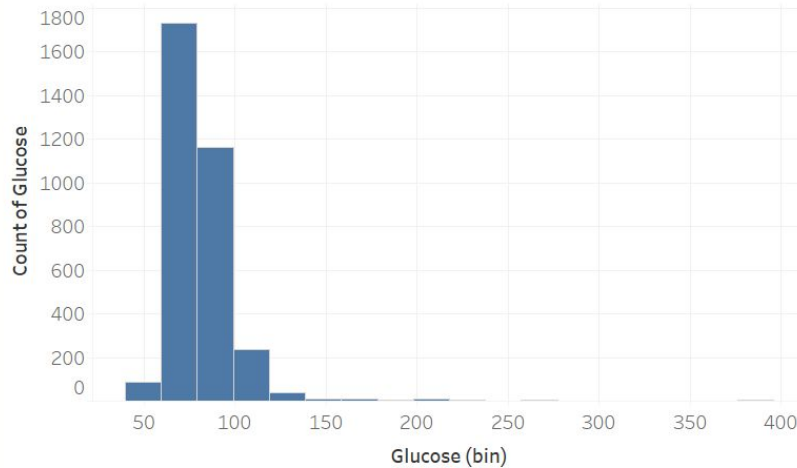
BMI



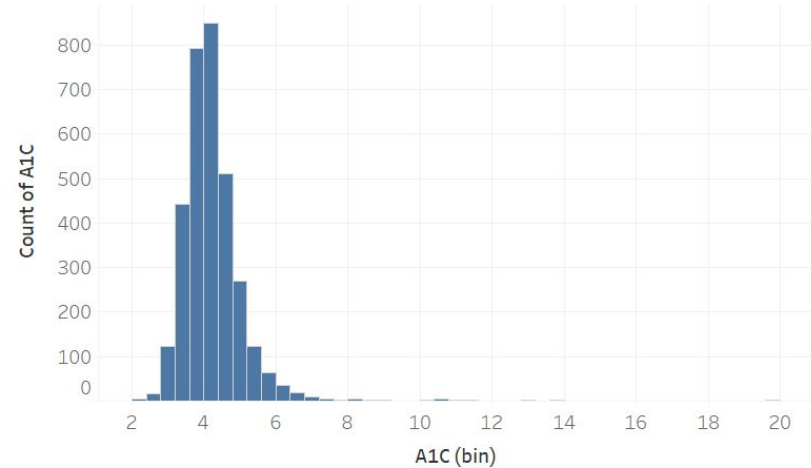
HeartRate



Glucose

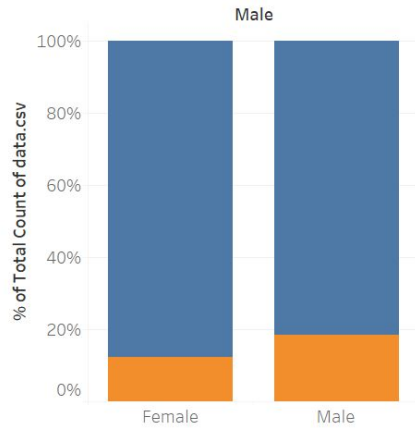


A1C

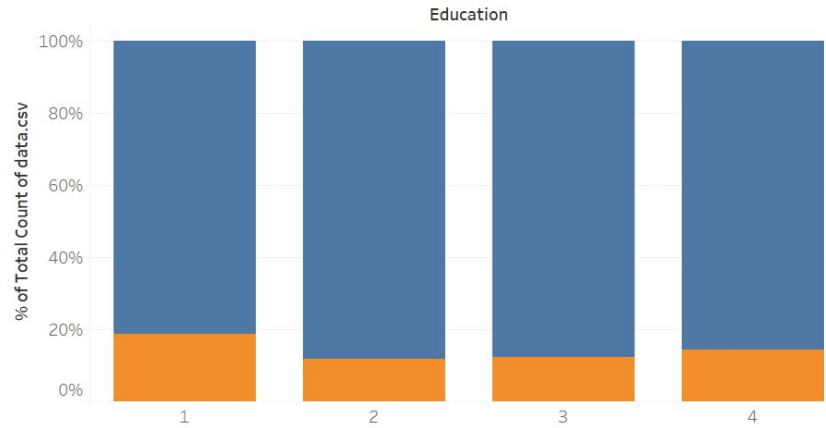


BIVARIATE ANALYSIS VS RESPONSE VARIABLE

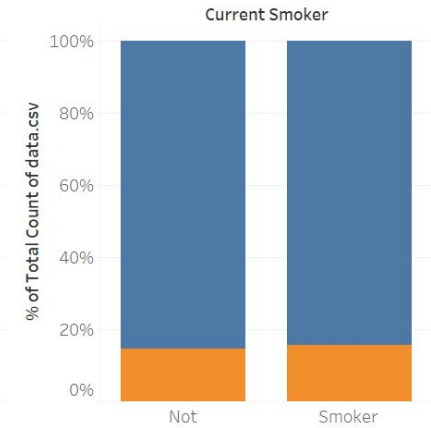
Gender vs TenYearCHD



Education vs TenYearCHD

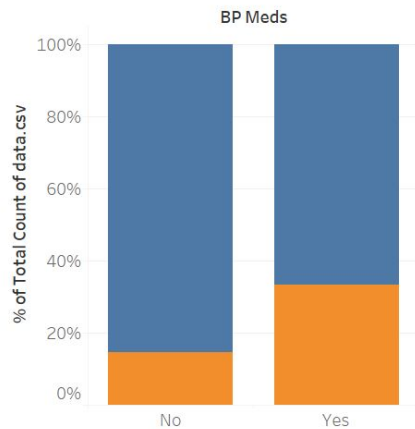


Smoker vs TenYearCHD

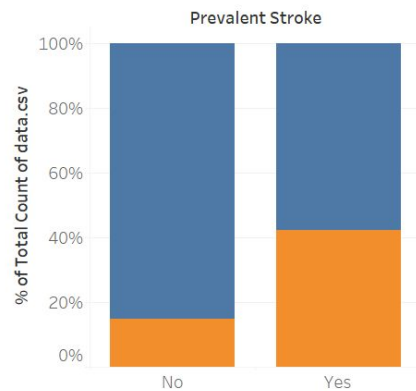


Ten Year CHD
■ No
■ Yes

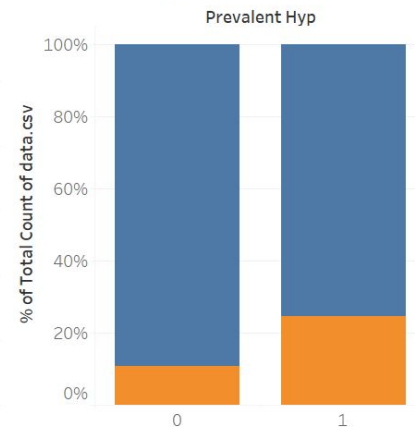
BP Meds vs TenYearCHD



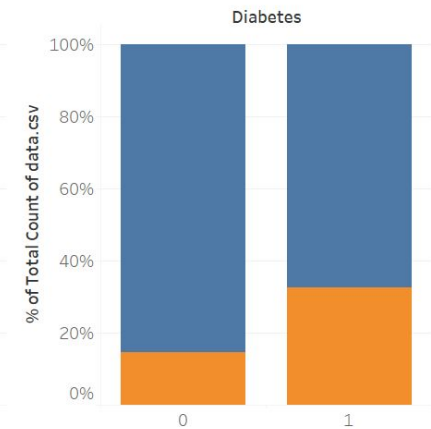
Prevalent Stroke vs TenYearCHD



Prevalent hyp TenYearCHD

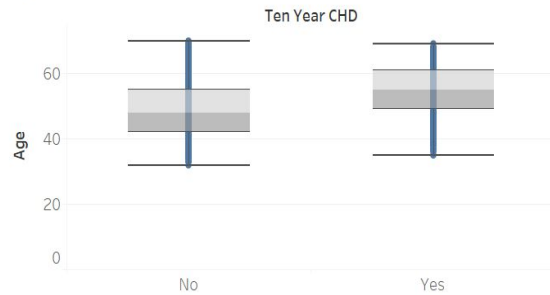


Diabetes vs TenYearCHD

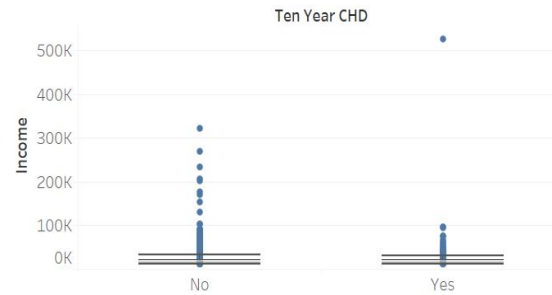


BIVARIATE ANALYSIS VS RESPONSE VARIABLE

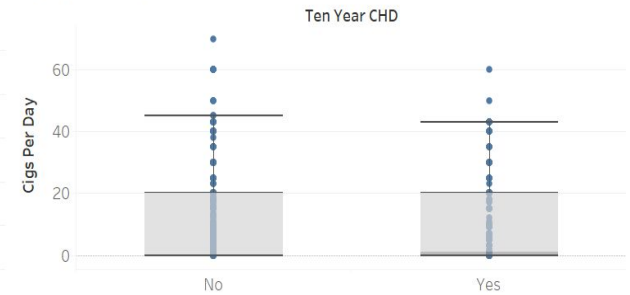
Age vs TenYearCHD



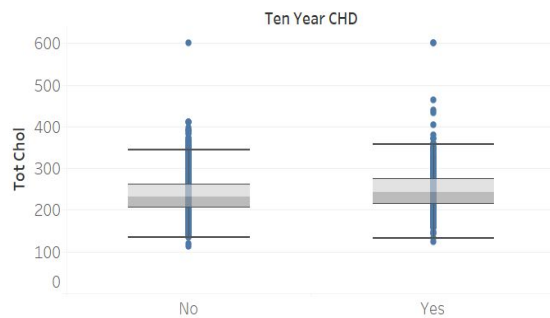
Income vs TenYearCHD



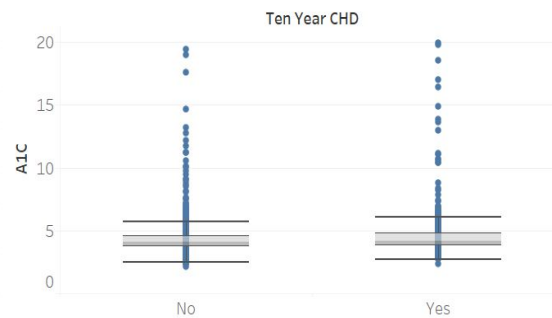
cigs per day vs TenYearCHD



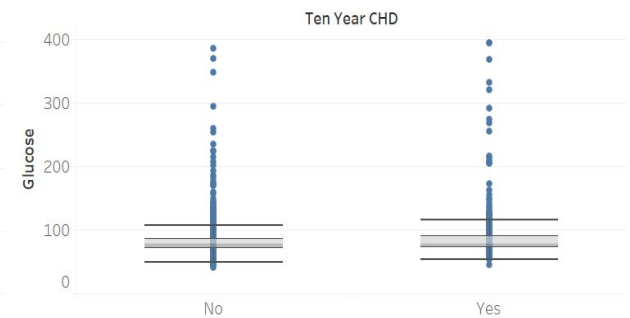
TotChol vs TenYearCHD



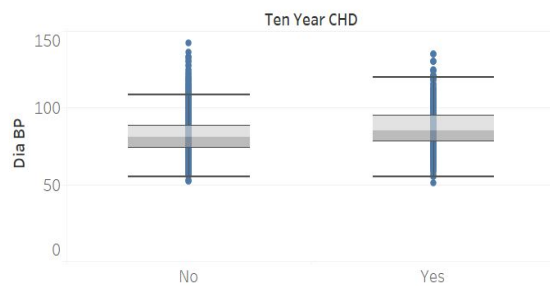
A1C vs TenYearCHD



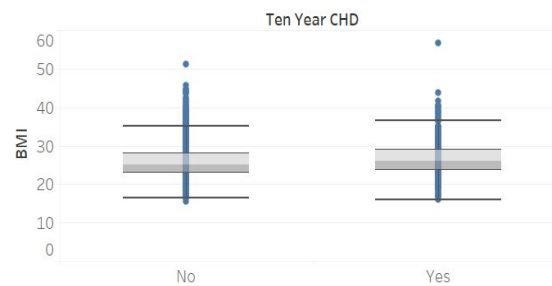
Glucose vs TenYearCHD



Dia BP vs TenYearCHD



BMI vs TenYearCHD

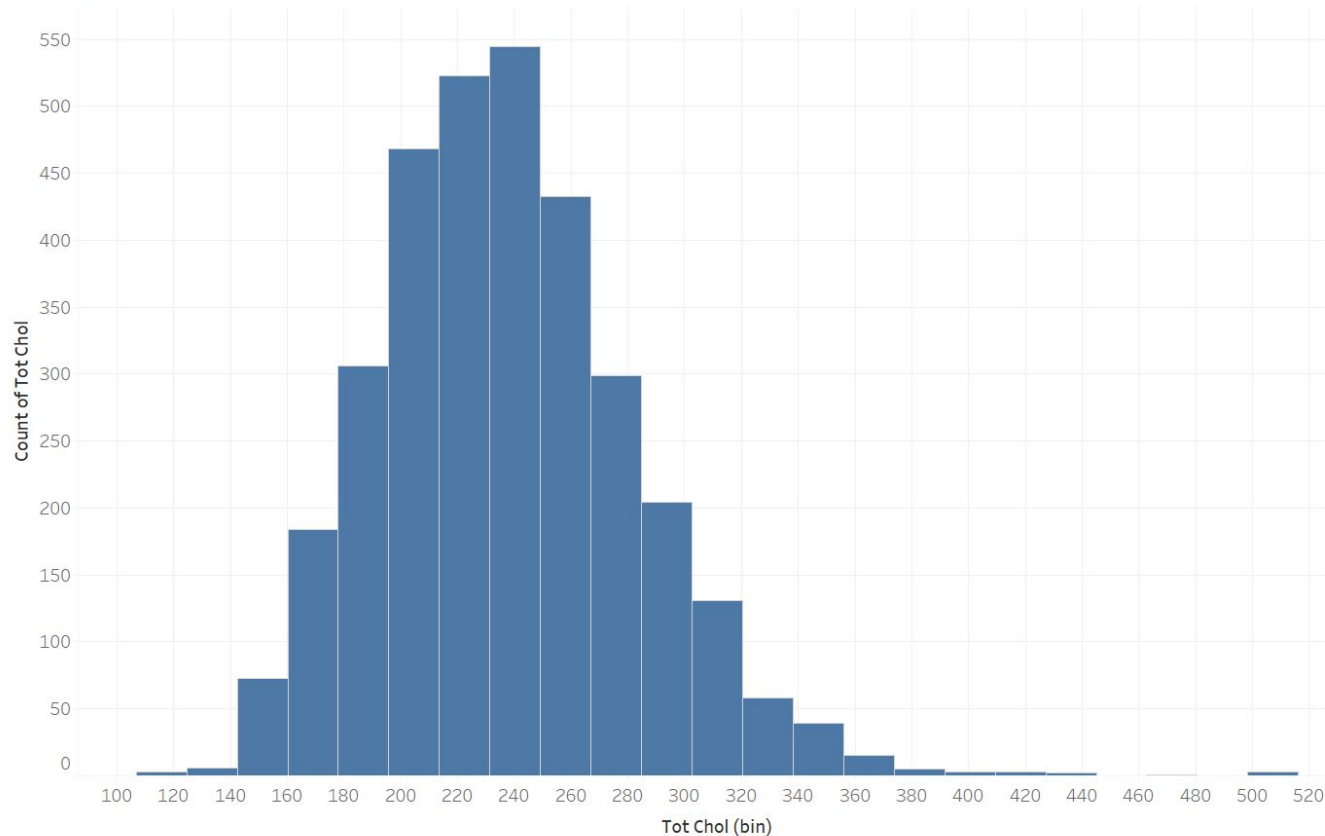


Sys BP vs TenYearCHD



SKEWNESS ANALYSIS

Sheet 4



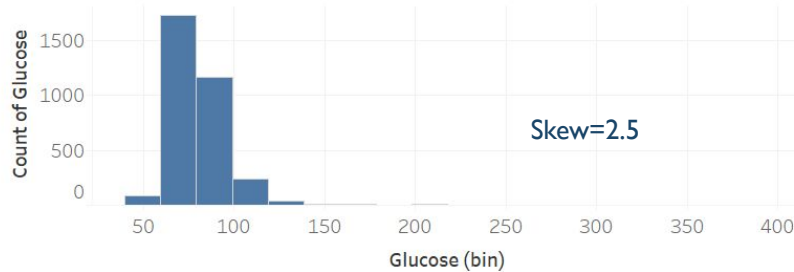
Summary

Count:	21
CNT(Tot Chol)	
Sum:	3,304
Average:	157.33
Minimum:	1
Maximum:	545
Median:	58.00
Skewness:	0.92

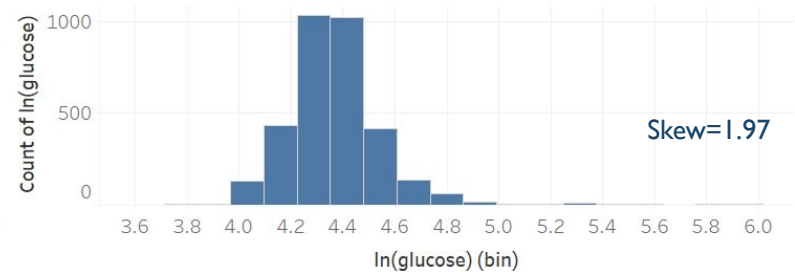
After impute the outliers of Tot Chol, Skewness is 0.92. Therefore, I do not need to transform Tot Chol.

SKEWNESS ANALYSIS

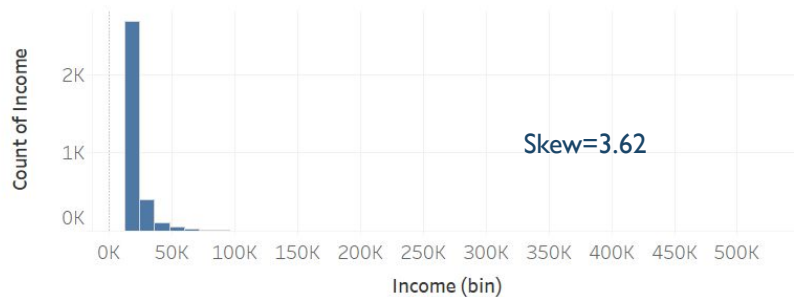
glucose Distribution



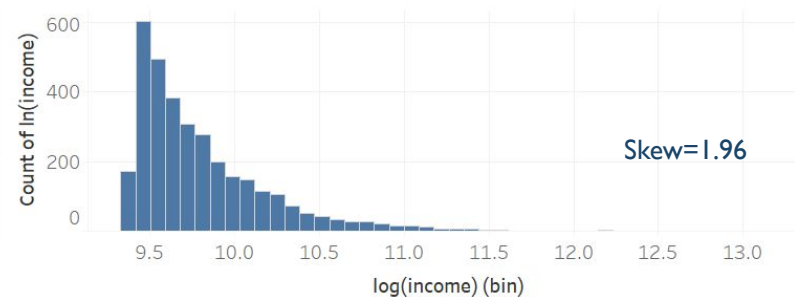
ln(glucose) Distribution



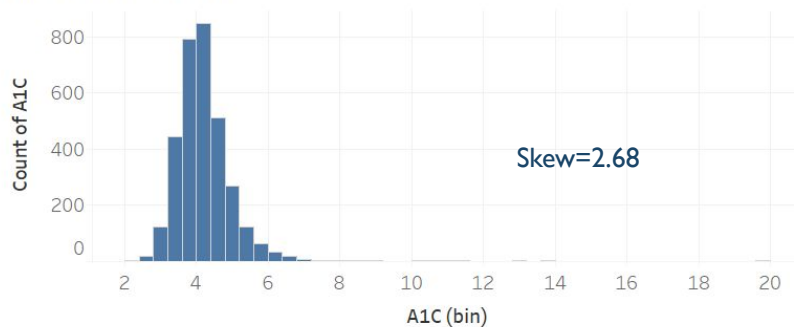
Income Distribution



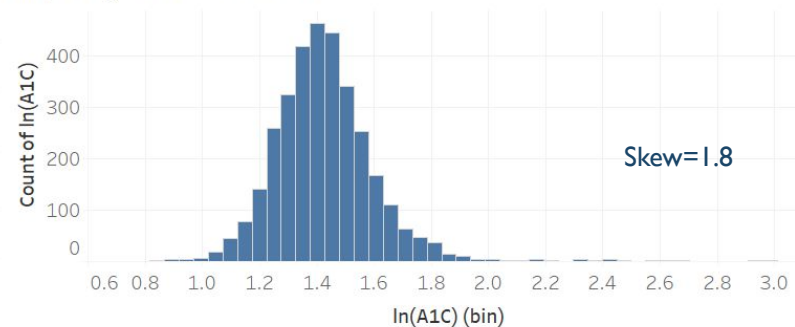
ln(income) Distribution



A1C Distribution

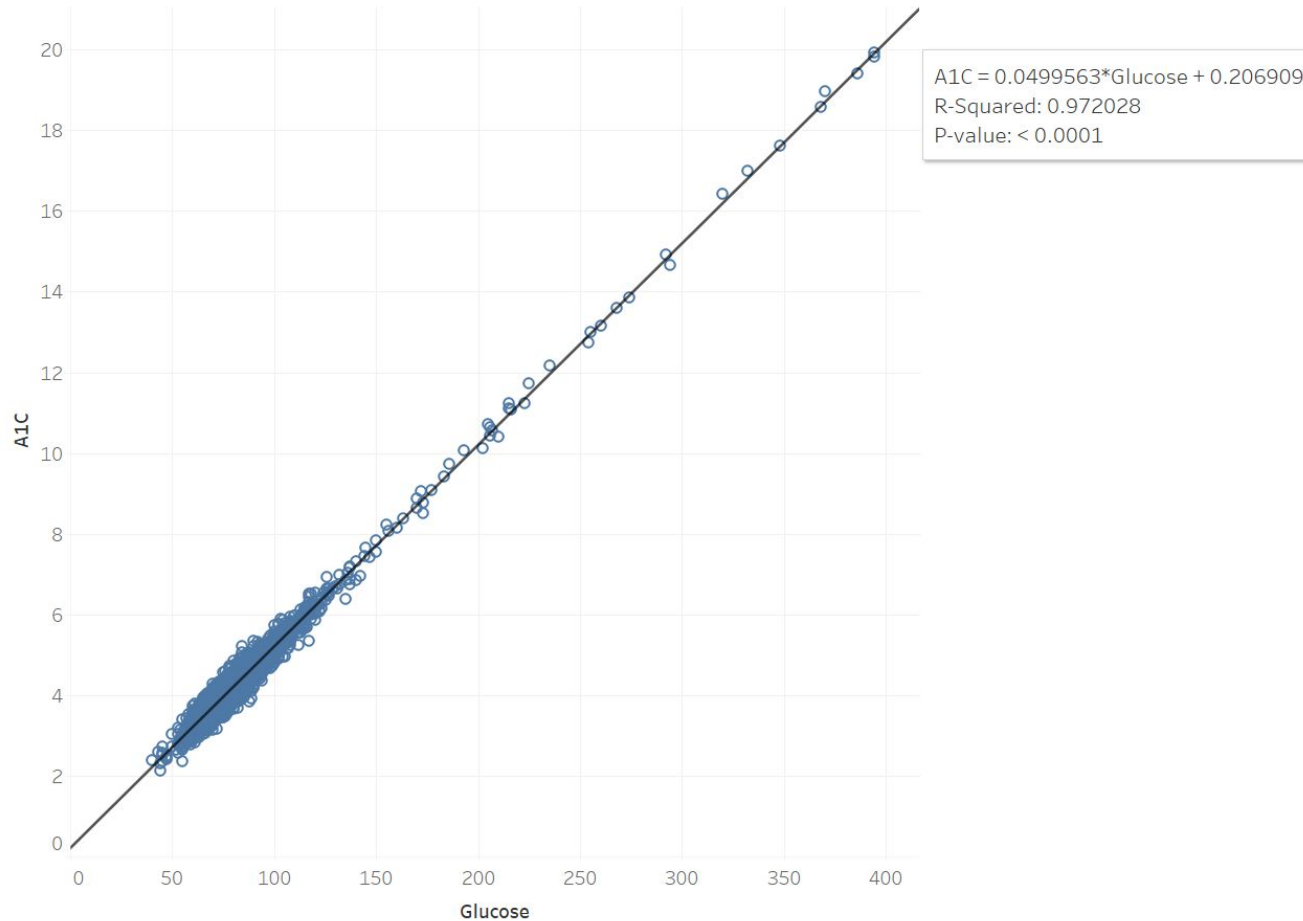


ln(A1C) Distribution



It appears that ln() function helps measure to reduce skewness.

CORRELATION ANALYSIS



After dropping the missing value, Glucose and A1C are perfectly correlated.
Decision: drop A1C from the analysis

FEATURE SELECTION/ENGINEERING DECISIONS


Features removed

- patientID – irrelevant for prediction purposes
 - I did not patientID, because I want to keep patientID in my final score dataset.
- AIC – highly correlated with glucose

Features engineering

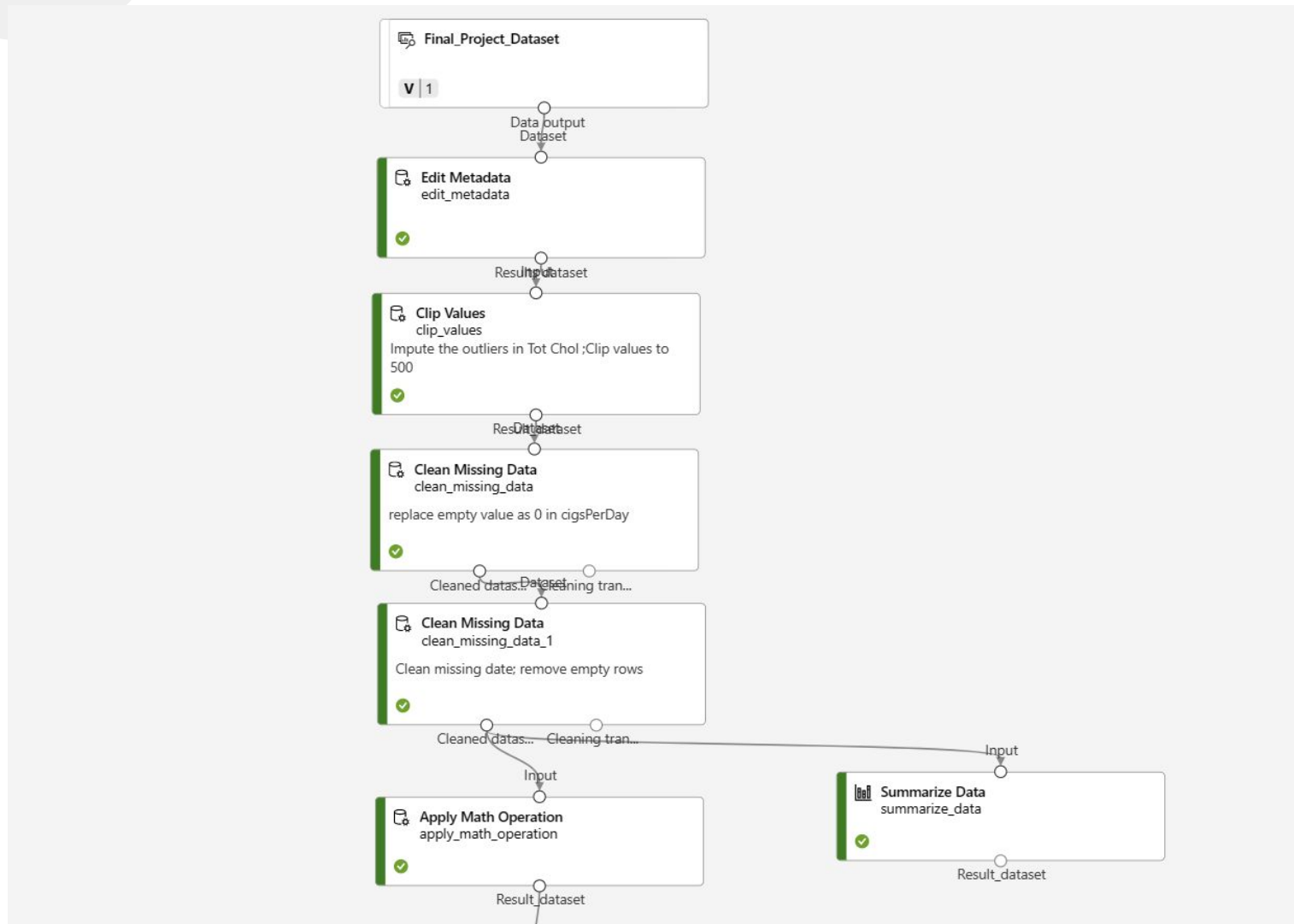
- Created \ln of glucose as predictor to reduce the high skewness
- Created \ln of income as predictor to reduce the high skewness



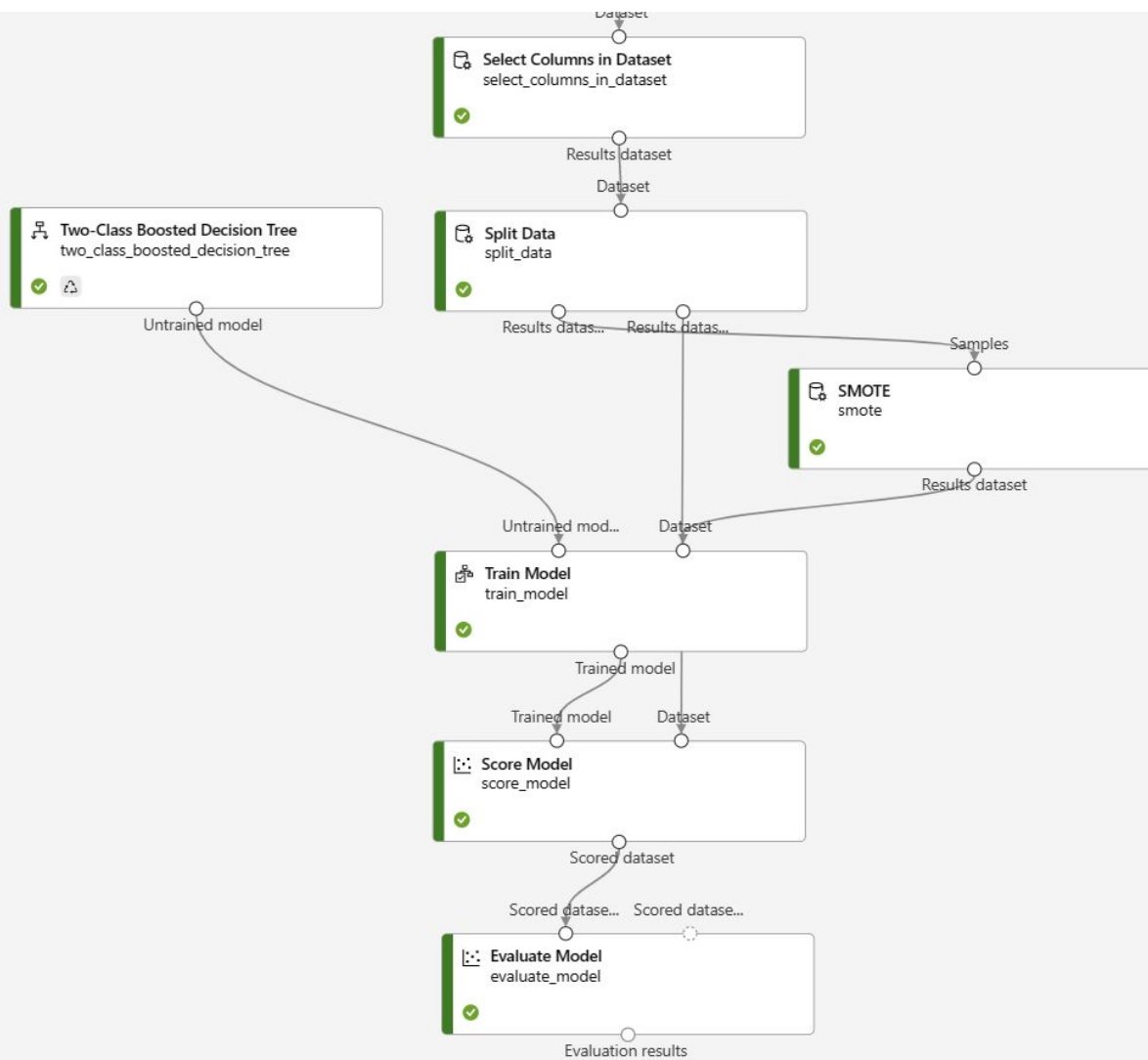


MODEL DEVELOPMENT

MODEL



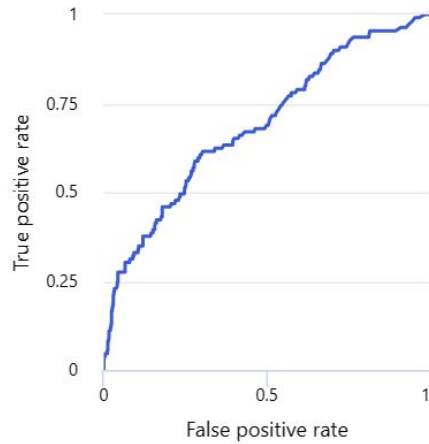
MODEL



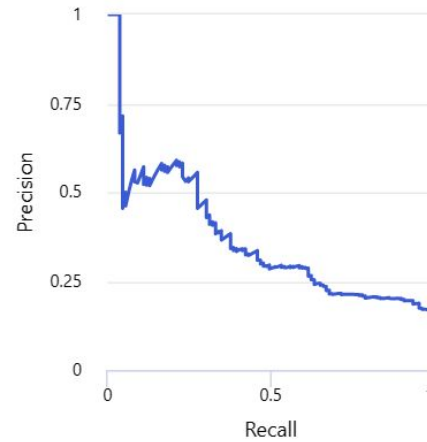
Result

● Scored dataset (left port)

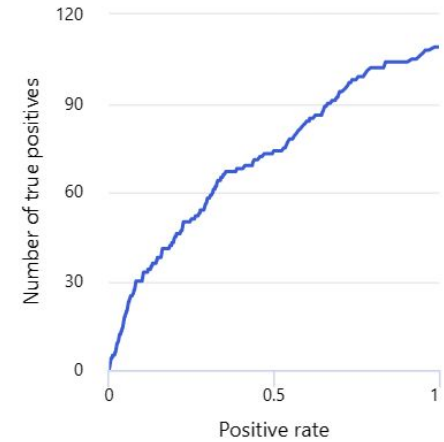
ROC curve



Precision-recall curve



Lift curve



Threshold 0.5

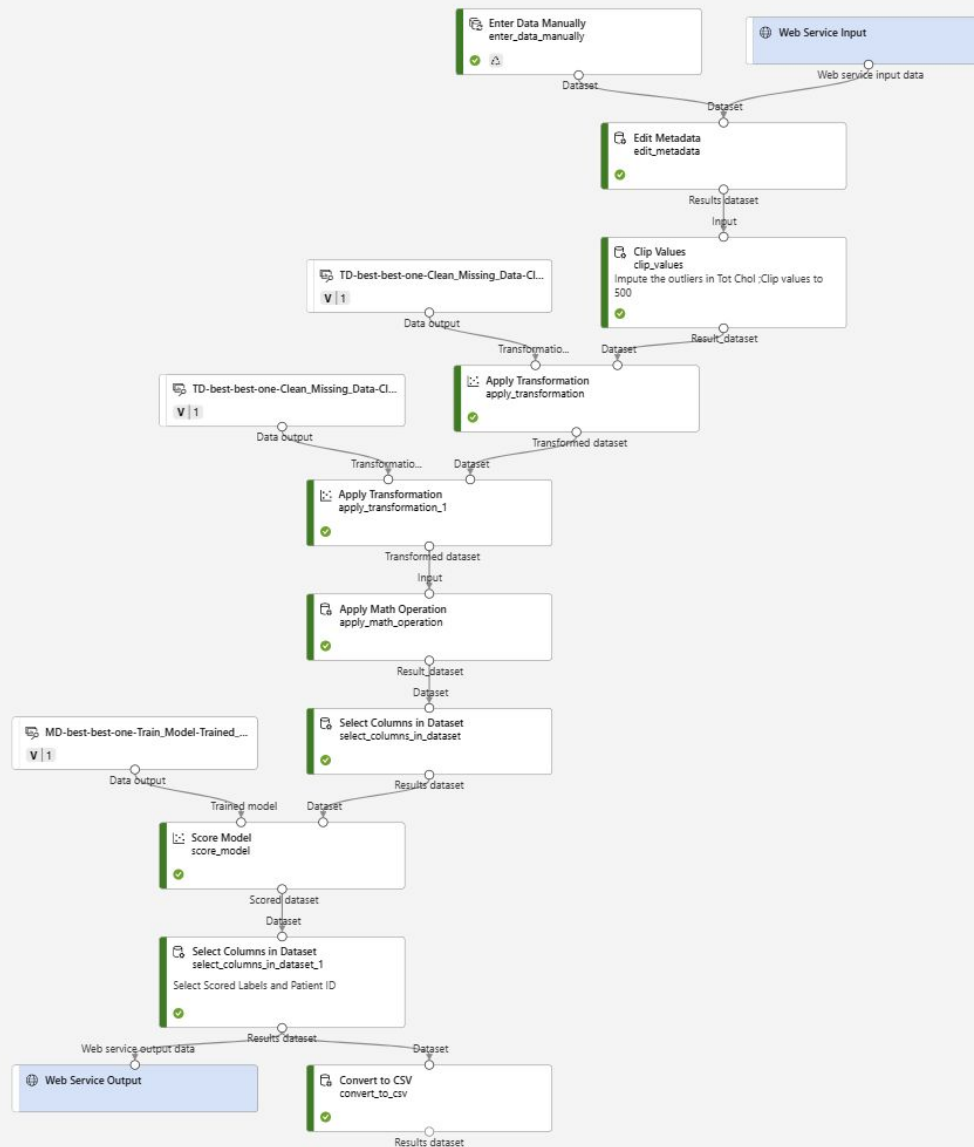
Accuracy 0.846
Precision 0.585
Recall 0.22
F1 Score 0.32
AUC 0.688

	Actual	
	1	0
Predicted 1	24	17
Predicted 0	85	535

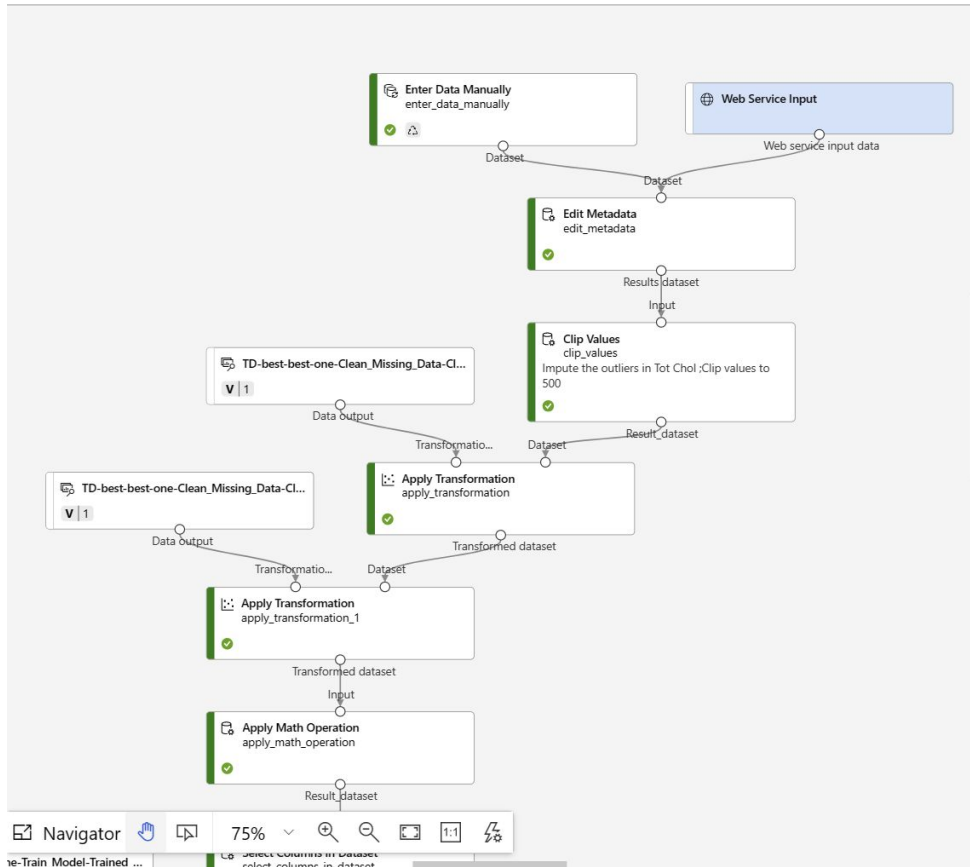


MODEL DEPLOYMENT

MODEL DEPLOYMENT



MANUAL TEST DATA



Enter Data Manually

Overview **Parameters** Outputs + logs Metrics Child jobs Images ...

Refresh + Register model Debug and monitor

Data format ⓘ *

CSV

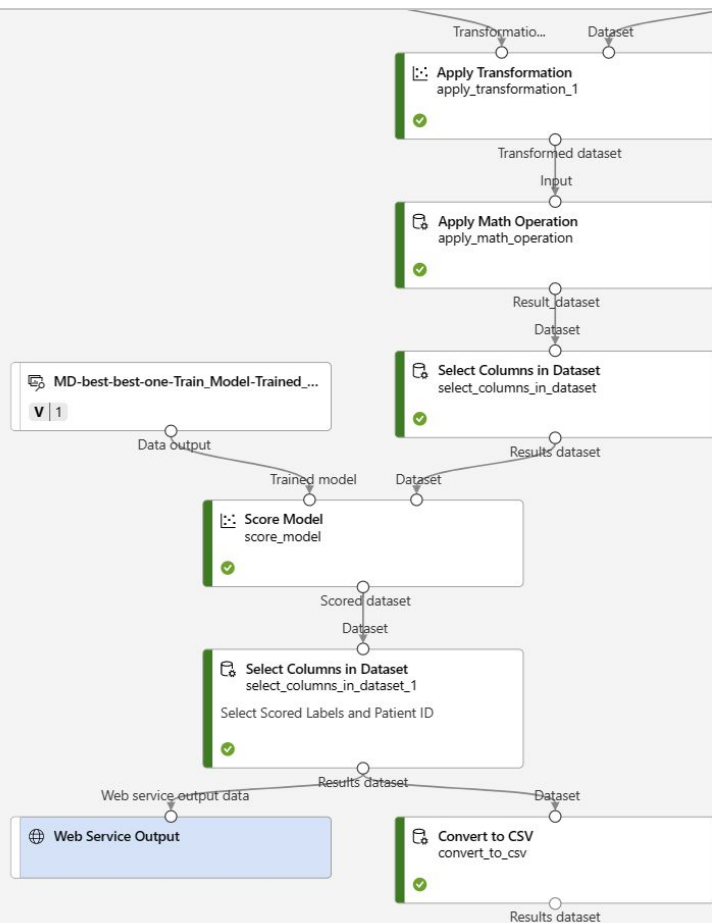
Has header ⓘ *

True

Data ⓘ *

```
1 patientID,male,age,education,currentSmoker,cigsPerDay,BPMeds,prevalence
2 110399,1,48,3,1,10,0,0,1,0,232,138,90,22.37,64,72,4.050698516,13026
3 189047,1,41,2,0,,0,0,0,0,195,139,88,26.88,85,65,3.789559338,18858
4 957019,1,54,1,1,20,0,0,1,0,214,147,74,24.71,96,87,4.571277531,15439
5 208967,1,37,2,0,,0,0,1,0,225,124.5,92.5,38.53,95,83,4.24288058,15806
6 230935,0,63,1,1,3,0,0,1,0,267,156.5,92.5,27.1,60,79,4.370722316,1977
7 216024,1,57,1,0,,0,0,0,0,220,136,84,26.84,75,64,3.046747617,13403
8 368834,0,56,1,0,,0,0,1,0,296,180,90,23.72,75,120,6.22075543,17803
9 135175,0,48,1,0,,0,0,1,0,265,145,77,24.23,74,64,3.914695094,37822
10 294070,1,66,3,0,,0,0,0,0,288,109,71,29.29,80,80,4.610732928,14508
11 595710,1,38,4,0,,0,0,0,0,235,118,77,25.87,60,82,4.988670391,20519
12 425597,1,53,1,1,30,0,0,0,0,244,106,67.5,21.84,88,65,3.341763577,2625
13 650137,0,59,1,1,1,0,0,1,0,259,141,86,25.97,70,86,4.57160478,16288
14 590019,0,38,3,1,3,1,0,1,0,,125,80,22.79,98,,13340
15 925626,0,36,3,1,20,0,0,0,0,159,121.5,73,20.41,72,75,3.859034015,2702
16 276518,1,41,4,1,20,1,0,1,0,244,139,86,30.77,60,67,3.54930057,22263
17 342284,0,37,1,0,,0,0,0,0,300,112,60,23.67,81,75,3.77075934,28637
18 469306,0,45,4,1,15,0,0,0,0,224,117,74.5,16.75,68,87,4.632244066,1624
19 197764,0,55,1,0,,0,0,0,0,245,144.5,83.5,28.96,72,65,3.381004969,1556
```

TEST DATA SCORING



Results_dataset

Rows ? Columns ?
162 2

patientID Scored Labels

110399	0
189047	0
957019	0
208967	0
230935	0
216024	0
368834	0
135175	0
294070	1
595710	0
425597	0
650137	0
925626	0
276518	0
342284	0
469306	0
197764	0
416488	0
208652	0
500010	0

