USC Viterbi
School of Engineering

FINAL REPORT
ISE 543, SPRING 2023

Bruce Wilcox, PhD

OVERVIEW

Project objective - build a classification model to predict the likelihood of a patient developing Chronic Heart Disease (CHD) in the coming ten years:

Age
Male
Education
Income
Current Smoker
Cigarettes per Day
BP Meds
Prevalent Stroke
Prevalent Hyp
Diabetes
Total Chol
Sys BP
Dia BP

BMI
Heart Rate
Glucose
A 1 c
Ten Year CHD

EXPLORATORY DATA ANALYSIS/ DATA PREPARATION

ATTRIBUTE SUMMARY

Response Variable

- Ten Year CHD

Categories

- Male
- Education
- Current Smoker
- BP Meds
- Prevalent Stroke
- Prevalent Hyp
- Diabetes

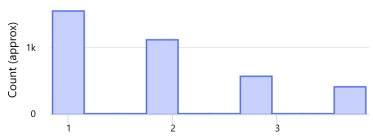
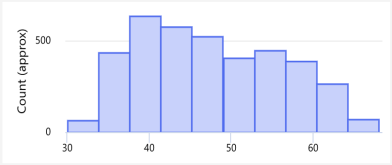
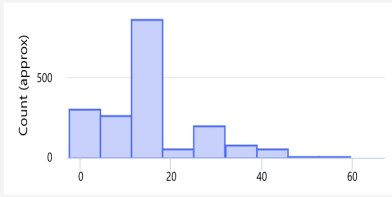
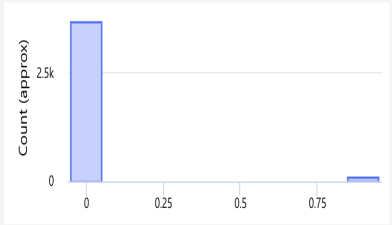
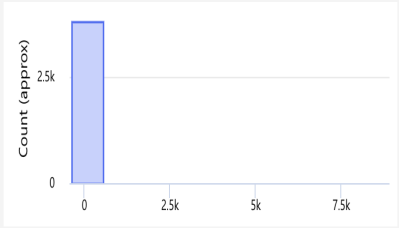
Measures

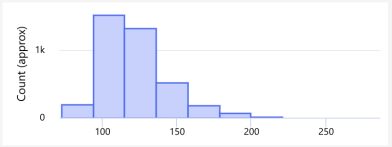
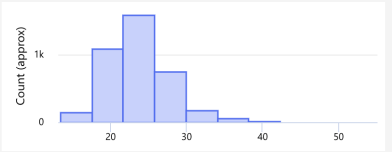
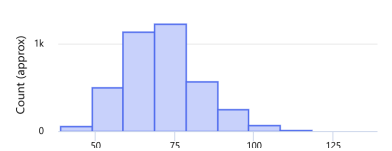
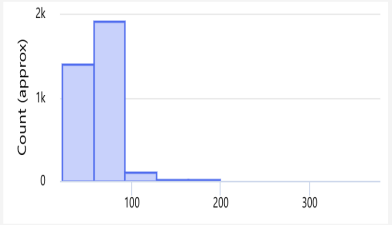
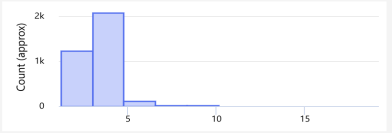
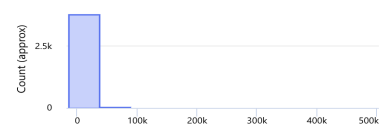
- Income
- Age
- Cigarettes per Day
- Total Chol
- Sys BP
- Dia BP
- BMI
- Heart Rate
- Glucose
- A 1 c

} Need to determine how to variable

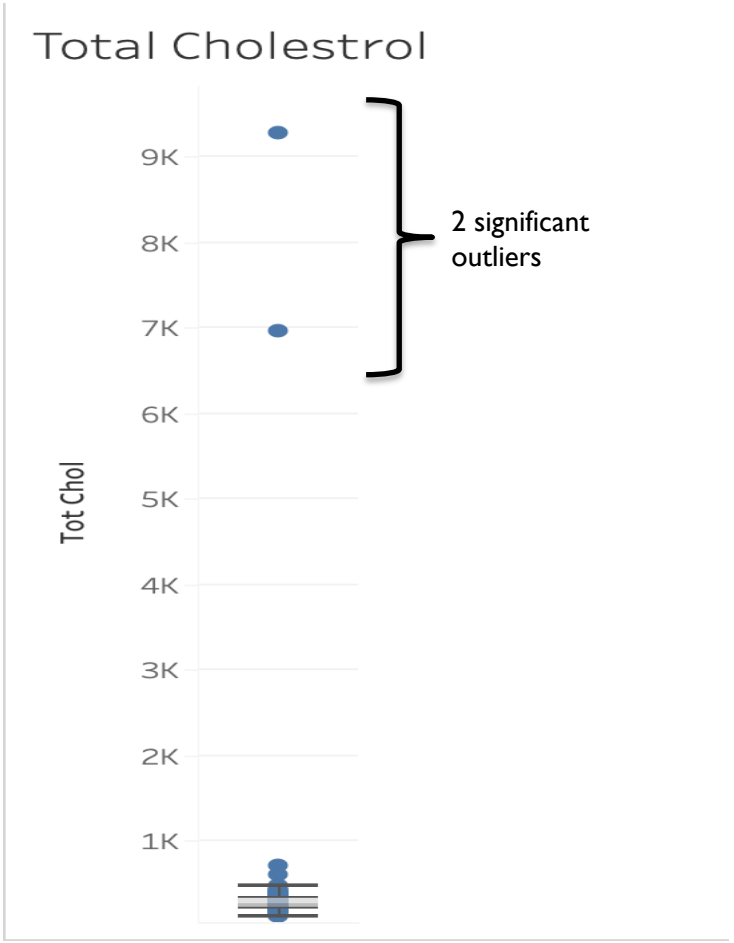
USC Viterbi School of Engineering

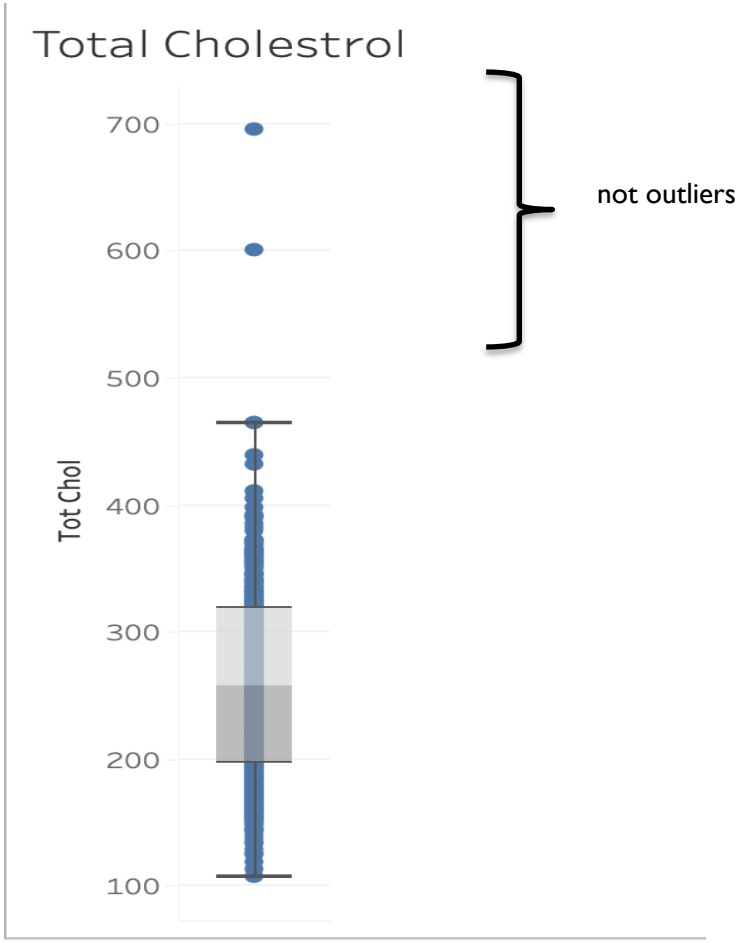
DATA CLEANSING

Column	Profile	Type	Min	Max	Count	Missing count	Empty count	Error count	Mean	Std dev...	Variance	Skewness	Kurtosis
education		Decimal	1	4	3816	93	0	0	1.97	1.02	1.04	0.7	-0.69
age		Integer	32	70	3816	0	0	0	49.57	8.6	73.92	0.24	-0.99
cigsPerDay		Decimal	1	70	3816	1975	0	0	18.5	10.93	119.37	0.76	0.96
BPMeds		Decimal	0	1	3816	45	0	0	0.03	0.17	0.03	5.59	29.29
totChol		Decimal	107	9,280	3816	47	47	0	0	240.85	188.93	35,695.2	41	1,811.61

sysBP		Decimal	83.5	295	3816	0	0	0	132.26	22.12	489.28	1.18	2.29
diaBP		Decimal	50	142.5	3816	0	0	0	82.87	11.91	141.79	0.72	1.23
BMI		Decimal	15.54	56.8	3816	19	0	0	25.81	4.1	16.81	1	2.74
heartRate		Decimal	44	143	3816	1	0	0	75.78	12.04	144.89	0.66	0.97
glucose		Decimal	40	394	3816	361	0	0	81.86	23.57	555.6	6.42	63.34
a1c		Decimal	2.13	19.92	3816	361	361	0	0	4.3	1.19	1.42	6.21	60.59
income		Decimal	12,000	524,494	3816	0	0	0	0	20,355.89	17,865.58	319,178,936.09	13.26	281.55

DATA CLEANSING – OUTLIER ANALYSIS





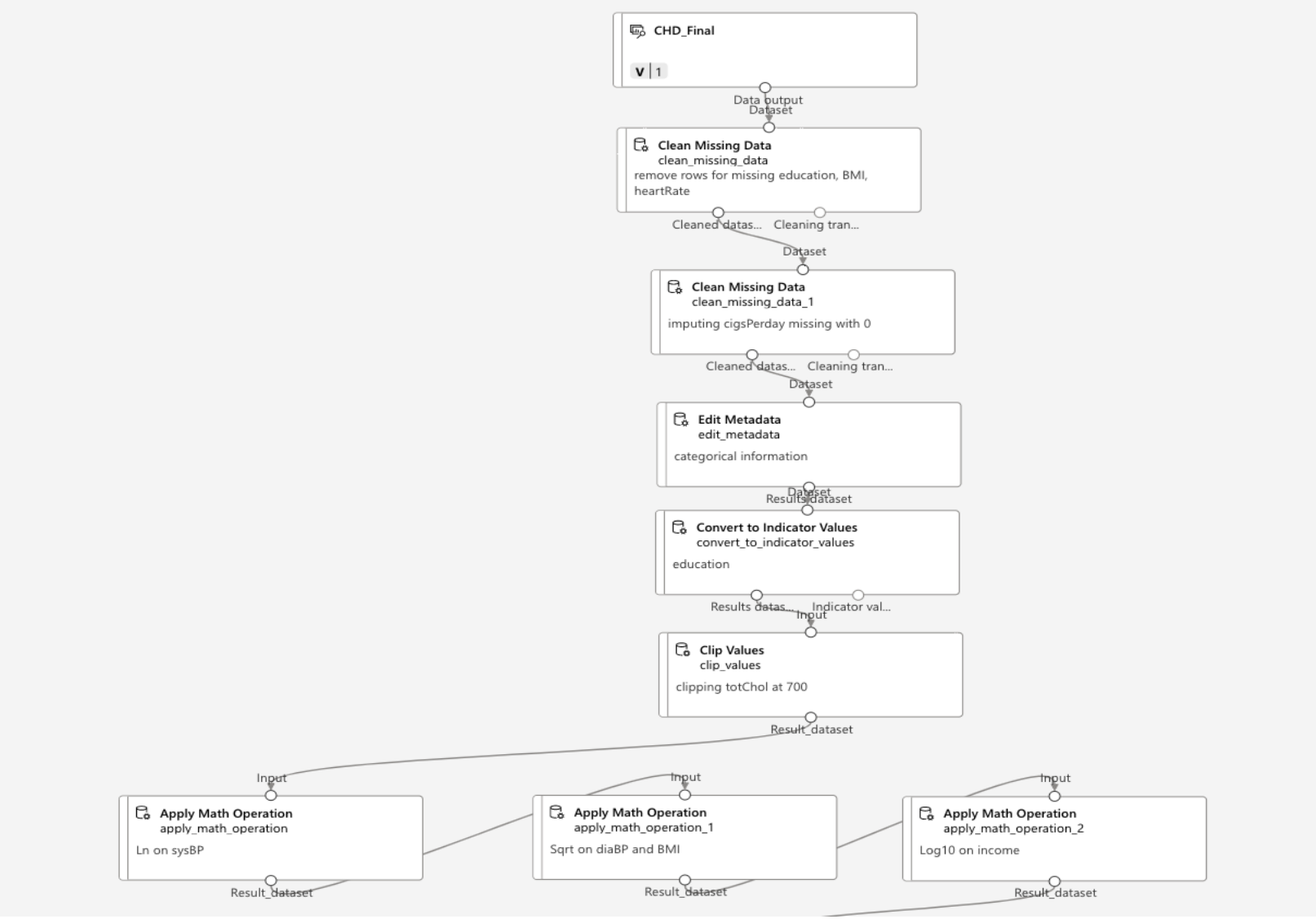
DATA CLEANSING - DECISIONS

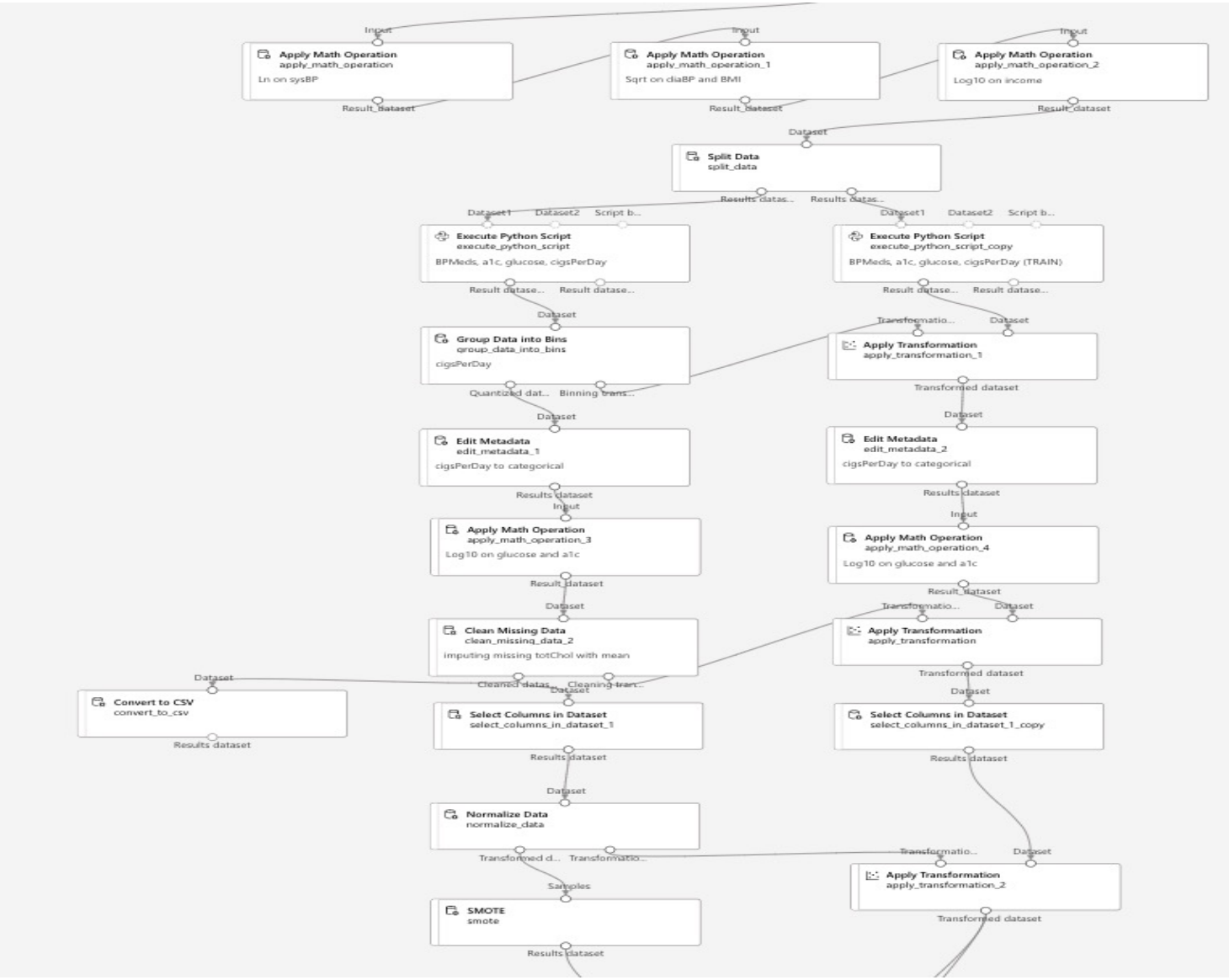
- Remove rows with missing values for education, BMI and heartrate because which do not have any information which indicates the education or BMI of the patient and imputing summary statistic in these cases can be misleading.
- Impute 0 for missing values in cigsPerDay variable as most of the missing variables correspond to patients that do are not current smokers.
- Remove outliers from totChol variable by clipping values greater than 646
- Impute missing values in BPMeds as 0 if the corresponding sysBP is greater than 140 and 0 otherwise. This is called stage 2 high BP and it is highly likely that they are on medication.
- Impute missing values in glucose and a1c by taking the average glucose or a1c of the group (diabetic or non-diabetic) that the patient belongs to as both glucose and a1c are related to diabetes

- For patients that are current smokers and have missing values in cigsPerDay variables, impute with average cigsPerDay
- Apply binning on cigsPerDay as 0,1,10,31,40,50 and convert it into a categorical variable
- Impute missing values in totChol with average

USC Viterbi School of Engineering

DATA CLEANSING PIPELINE

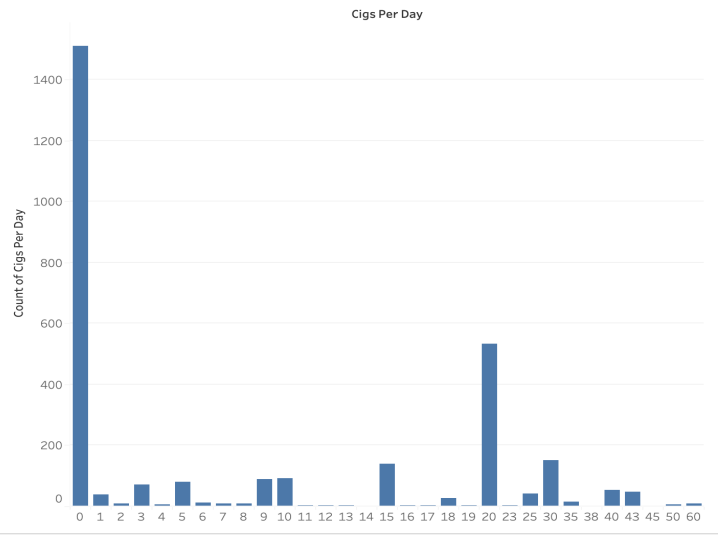



Apply Math Operation
apply_math_operation
Ln on sysBP
Result dataset
Apply Math Operation
apply_math_operation_1
Sqrt on diaBP and BMI
Result dataset
Apply Math Operation
apply_math_operation_2
Log10 on income
Result dataset
Split Data
split_data
Execute Python Script
execute_python_script
BPMeds, a1c, glucose, cigsPerDay
Execute Python Script
execute_python_script_copy
BPMeds, a1c, glucose, cigsPerDay (TRAIN)
Group Data into Bins
group_data_into_bins
cigsPerDay
Apply Transformation
apply_transformation_1
Edit Metadata
edit_metadata_1
cigsPerDay to categorical
Edit Metadata
edit_metadata_2
cigsPerDay to categorical
Apply Math Operation
apply_math_operation_3
Log10 on glucose and a1c
Apply Math Operation
apply_math_operation_4
Log10 on glucose and a1c
Clean Missing Data
clean_missing_data_2
imputing missing totChol with mean
Apply Transformation
apply_transformation
Convert to CSV
convert_to_csv
Select Columns in Dataset
select_columns_in_dataset_1
Select Columns in Dataset
select_columns_in_dataset_1_copy
Normalize Data
normalize_data
Apply Transformation
apply_transformation_2
SMOTE
smote

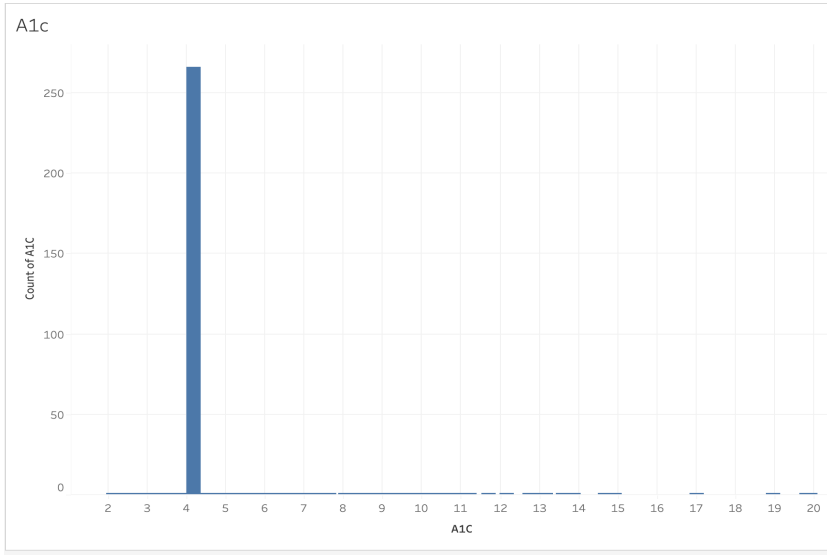
USC Viterbi School of Engineering

UNIVARIATE ANALYSIS MEASURES

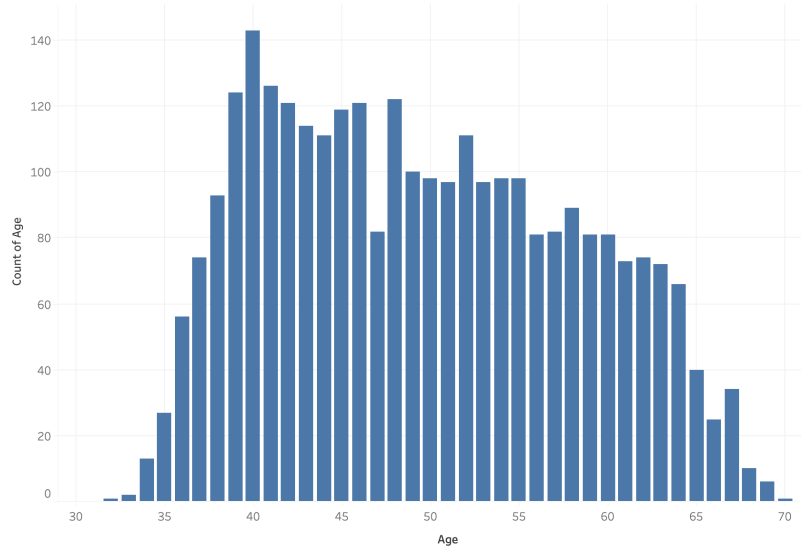
CigsPerDay



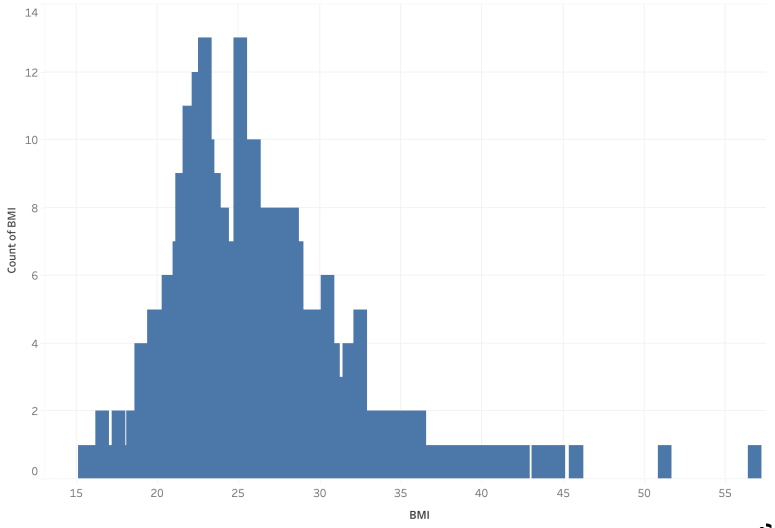
A1c



Age

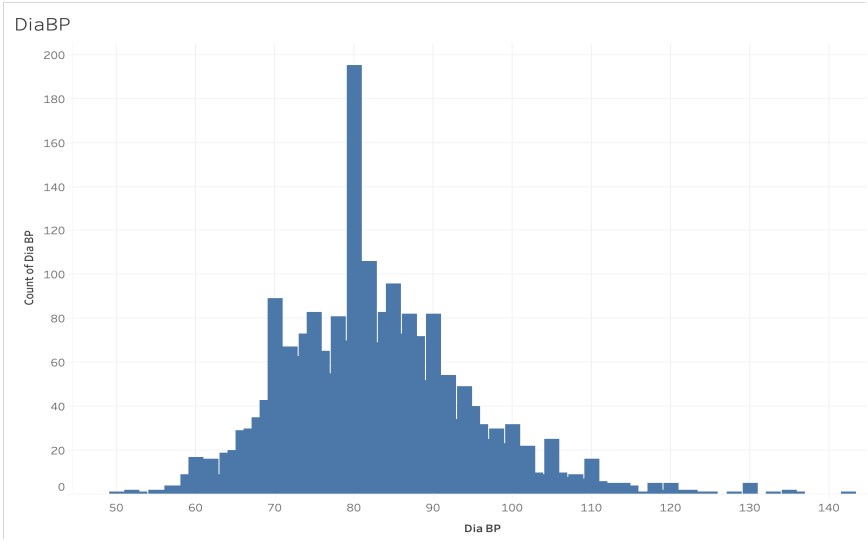


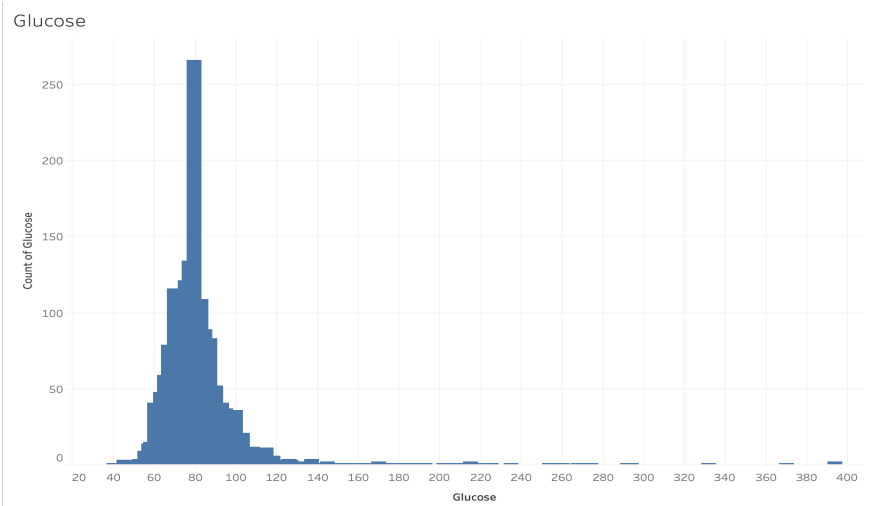
BMI

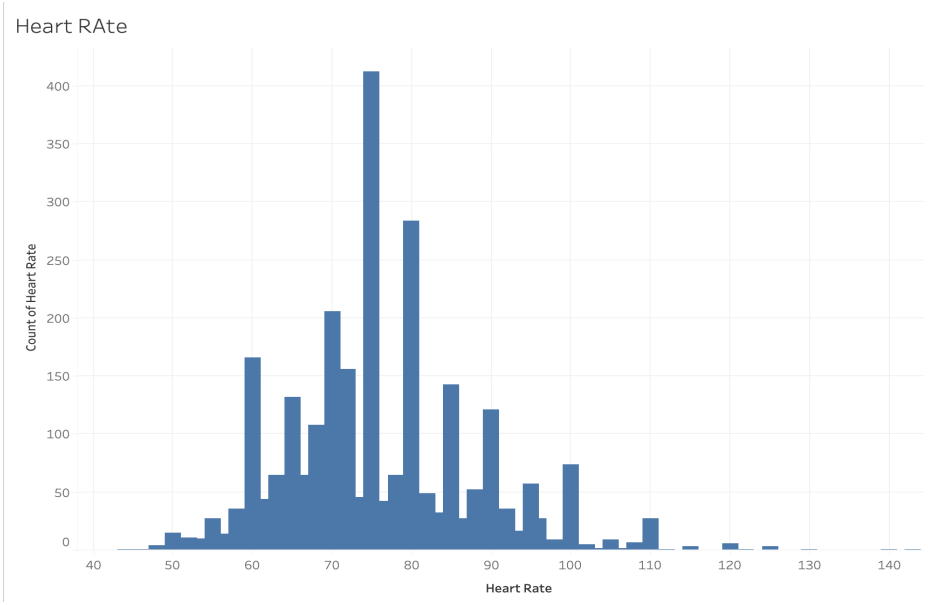


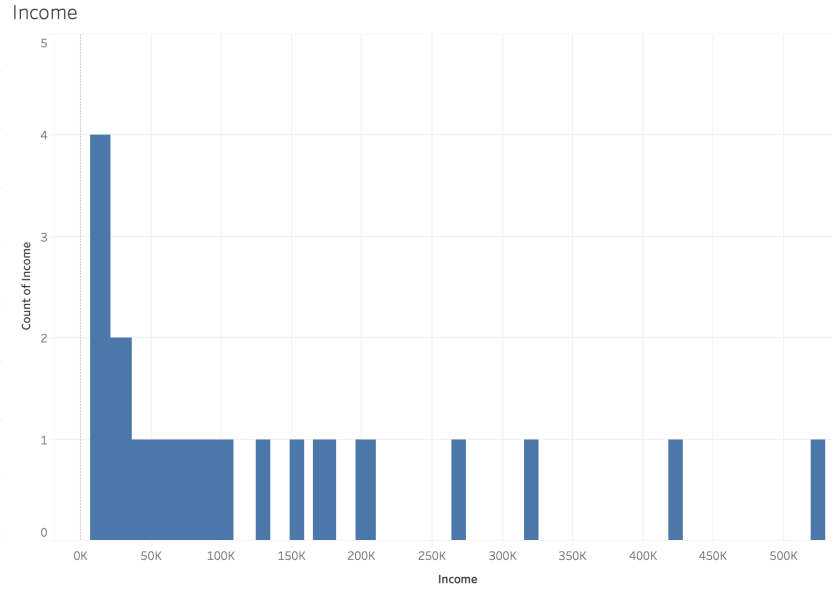
USC Viterbi School of Engineering

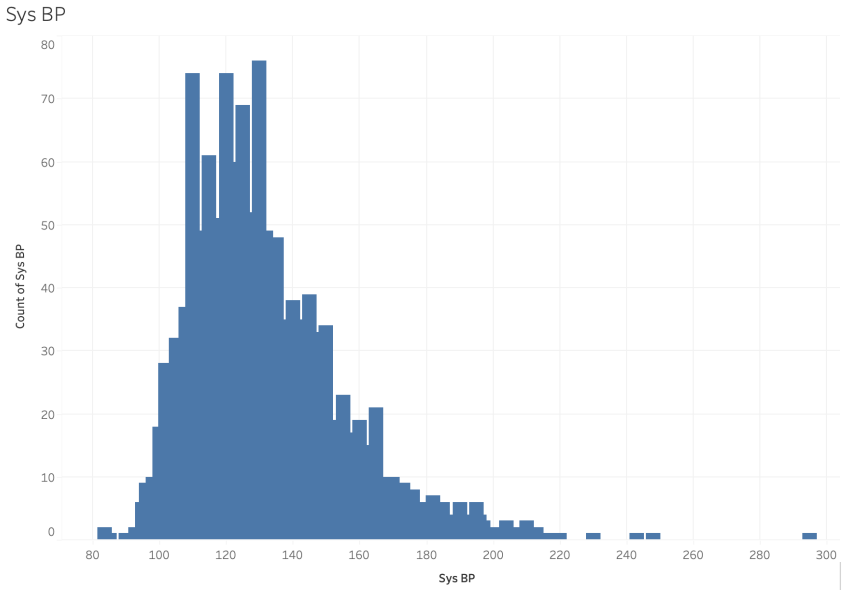
UNIVARIATE ANALYSIS MEASURES

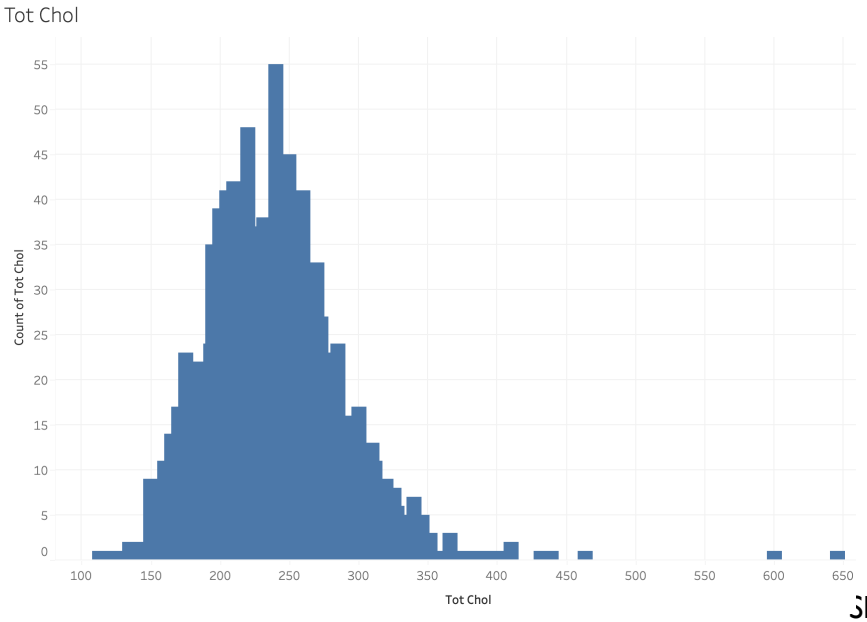






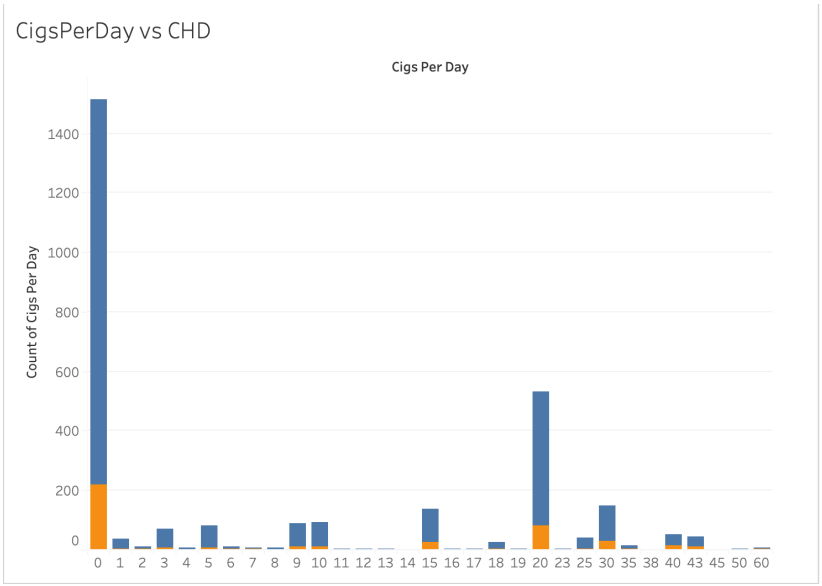


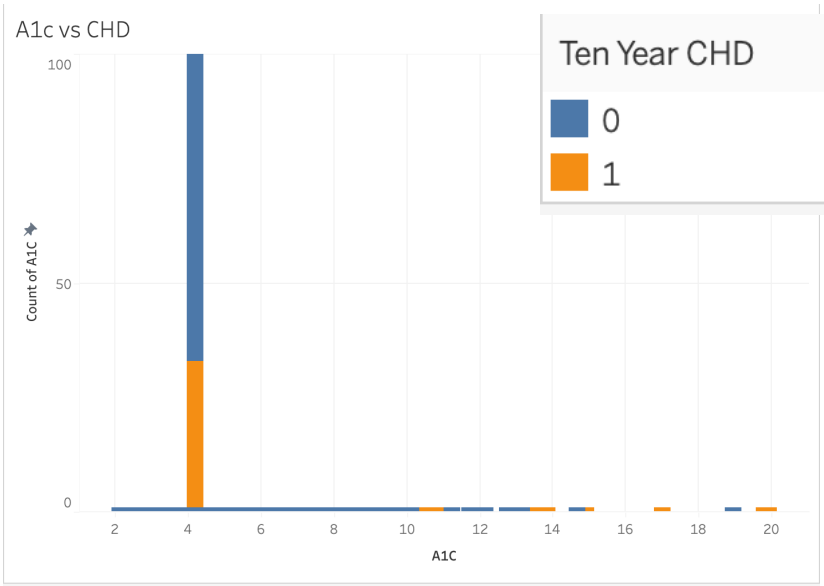

Sys BP
Count of Sys BP
Sys BP
0
10
20
30
40
50
60
70
80
80
100
120
140
160
180
200
220
240
260
280
300

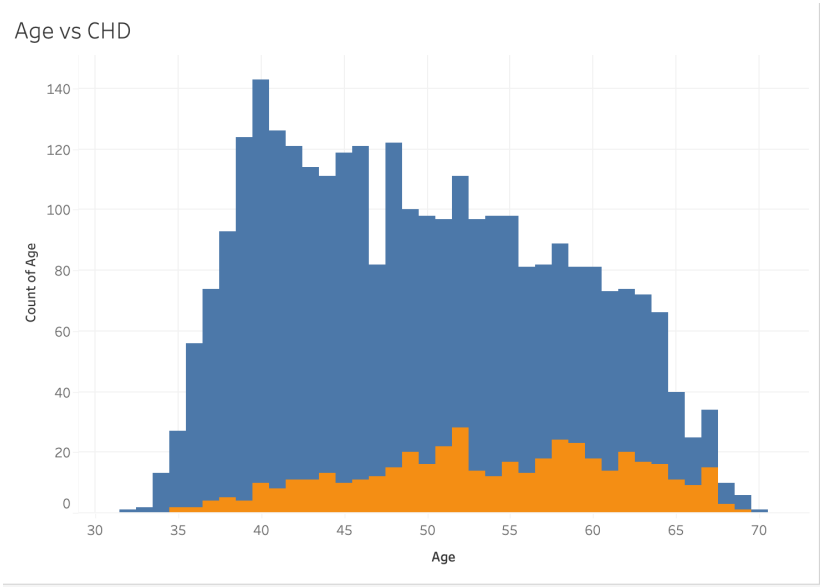

Tot Chol
Count of Tot Chol
Tot Chol
0
5
10
15
20
25
30
35
40
45
50
55
100
150
200
250
300
350
400
450
500
550
600
650

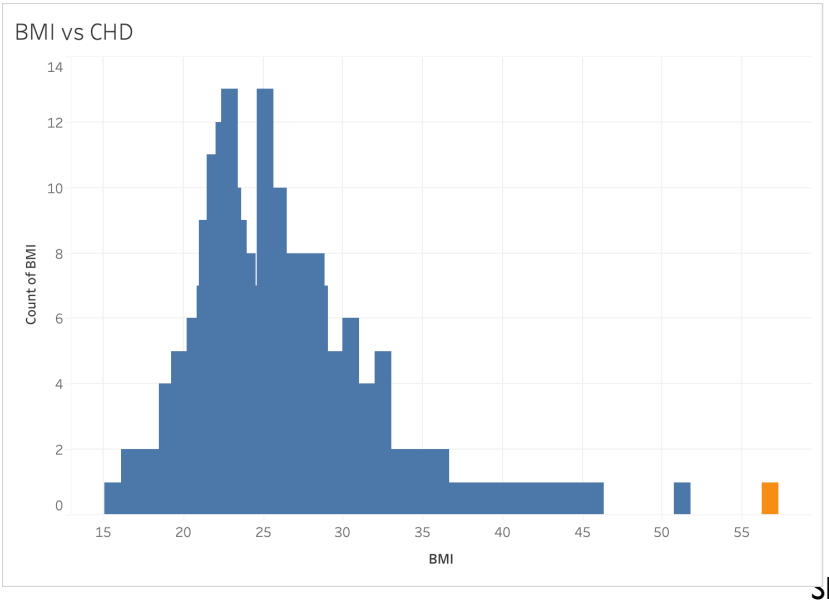
USC Viterbi School of Engineering

BIVARIATE ANALYSIS VS RESPONSE VARIABLE



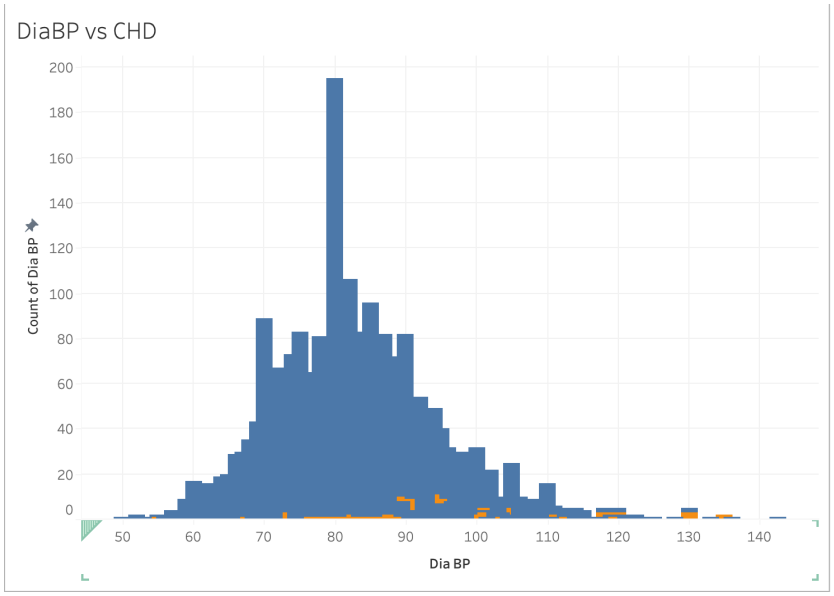


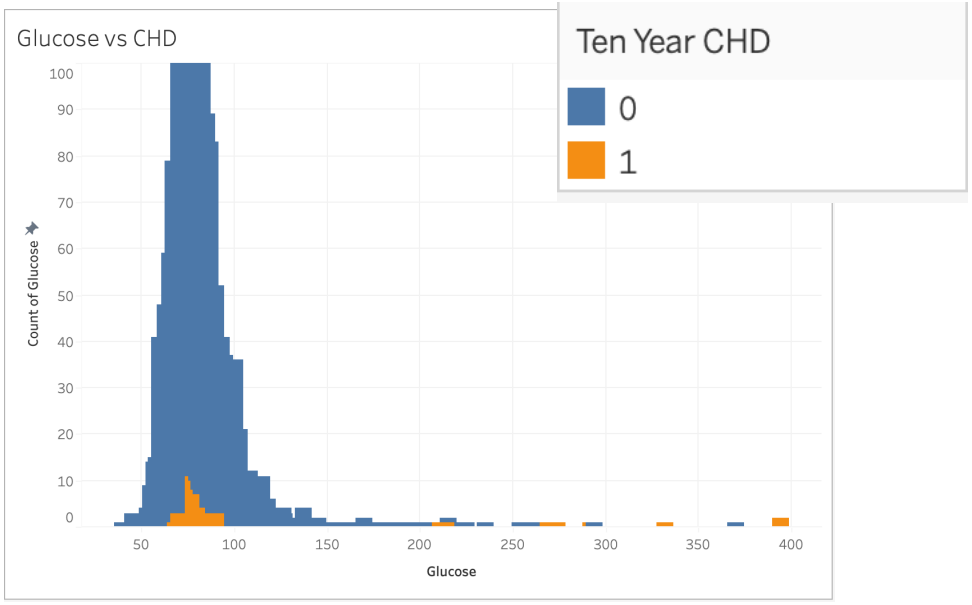


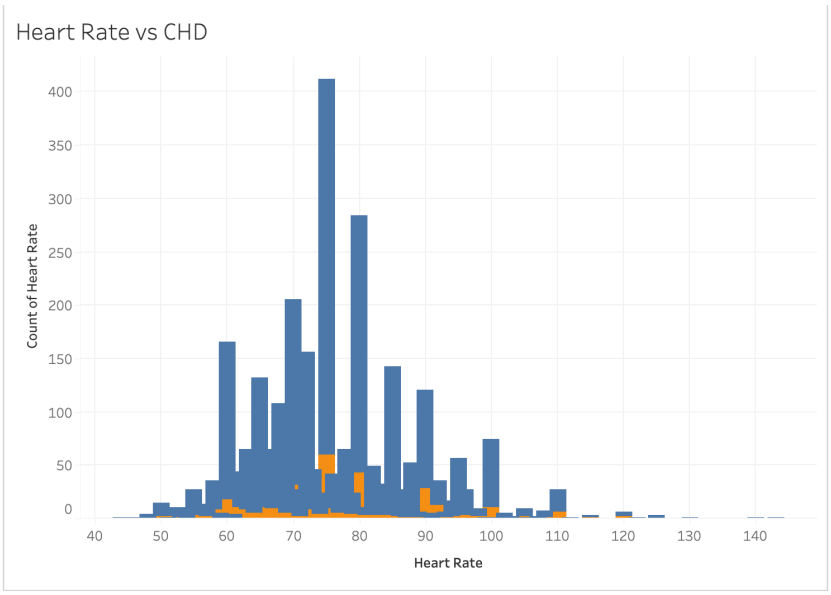


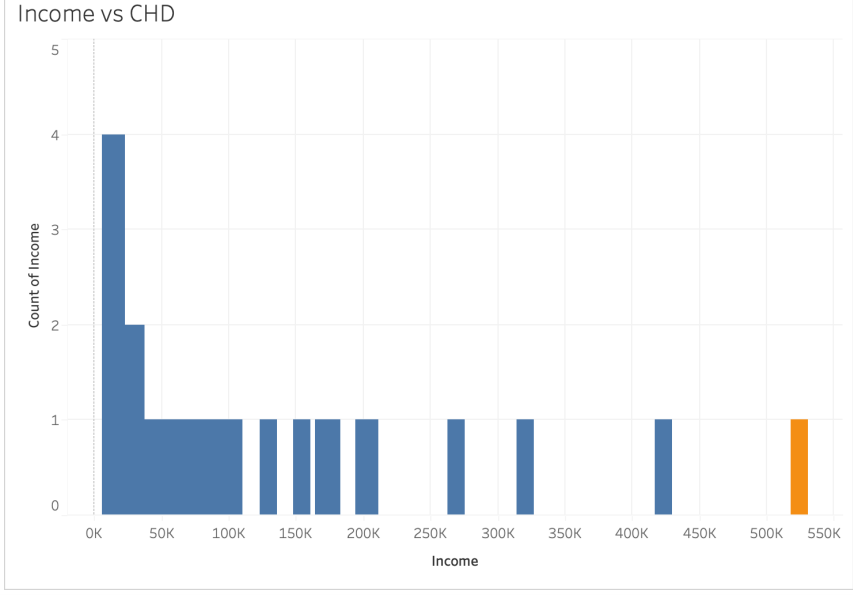
USC Viterbi School of Engineering

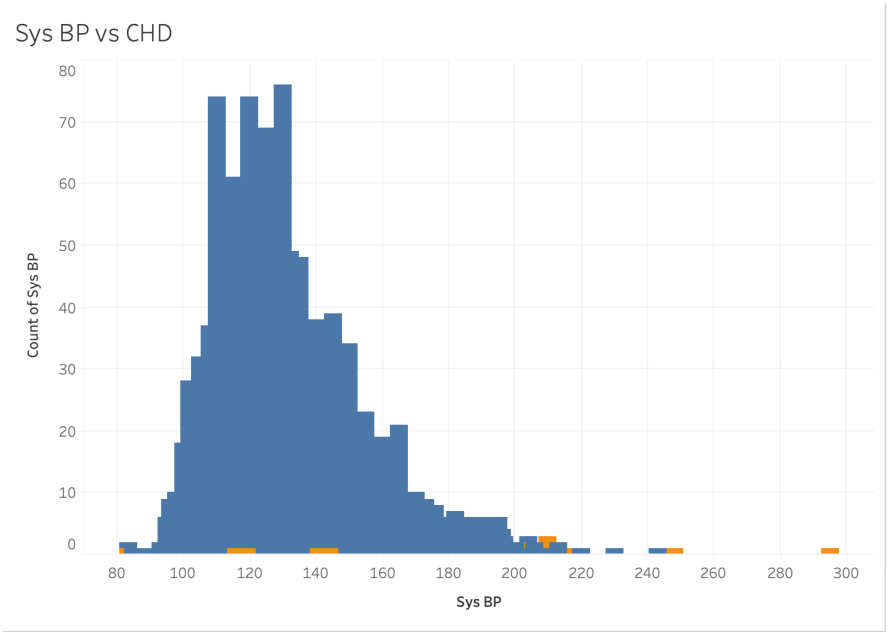
BIVARIATE ANALYSIS VS RESPONSE VARIABLE

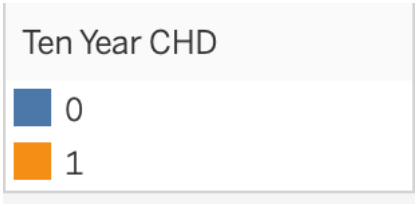




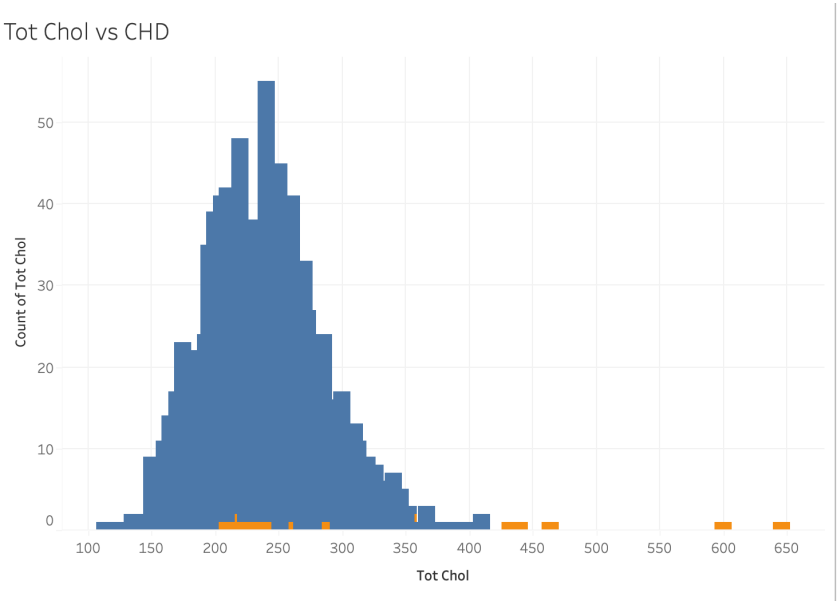








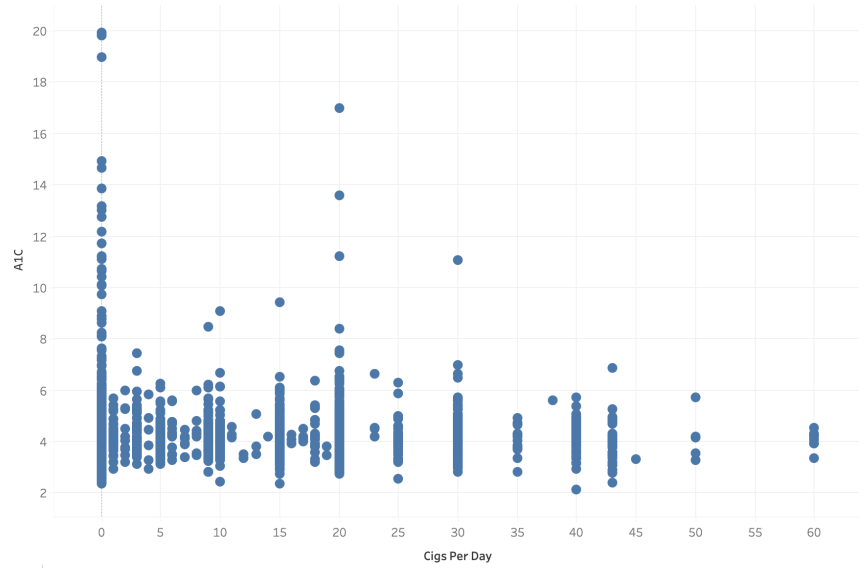
Determined that outliers are not to be deleted considering the relationship to response variable and business knowledge (nature of the variable) i.e. it is not uncommon to see such large/extreme values



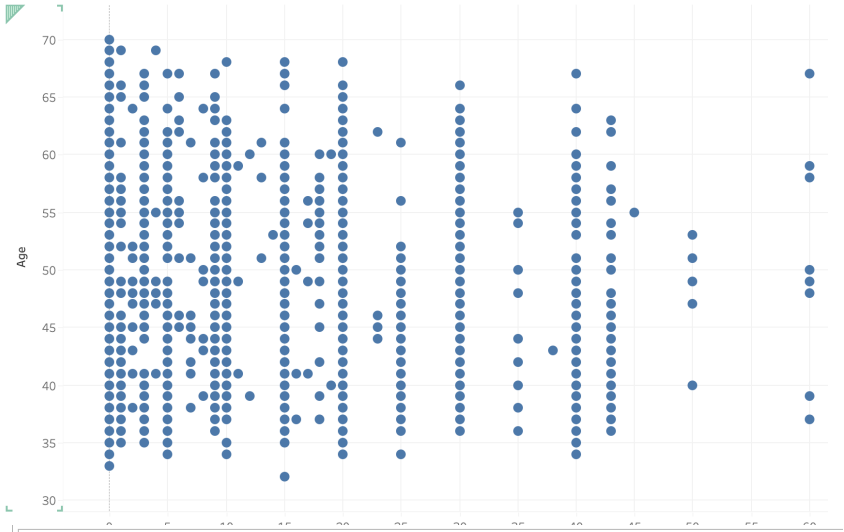
USC Viterbi School of Engineering

CORRELATION ANALYSIS

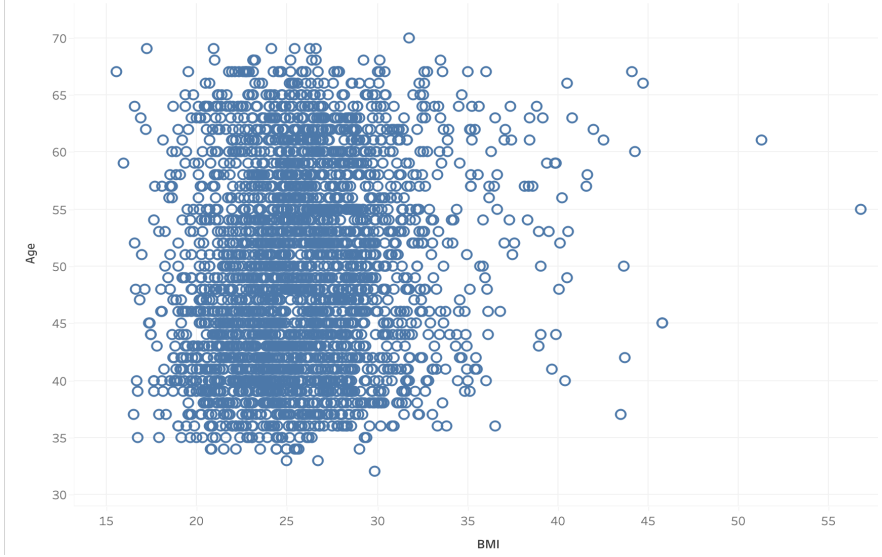
CigsPerDay vs a1c



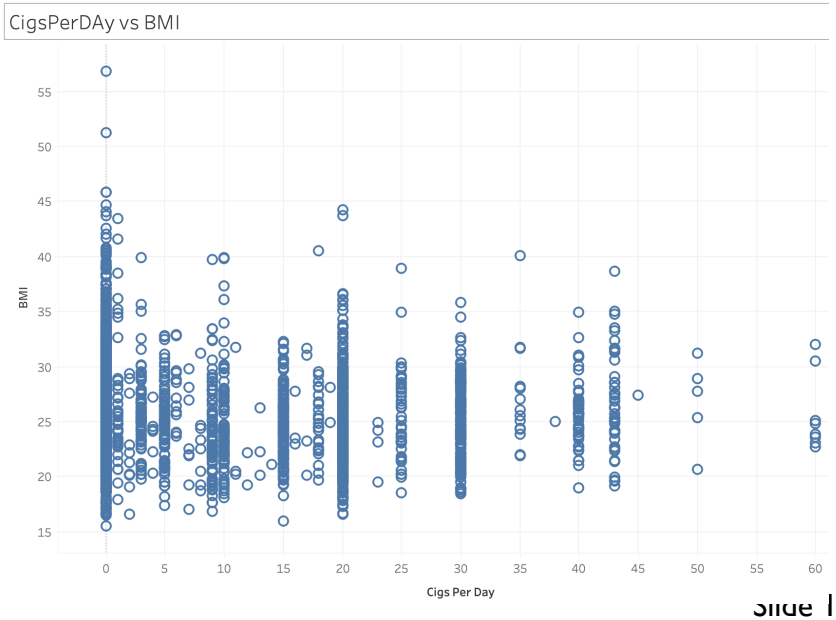
cigsPerDay vs Age



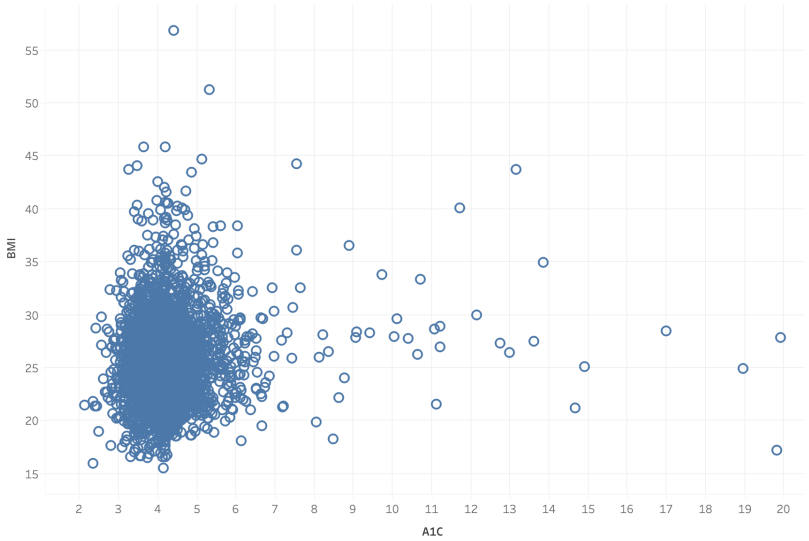
BMI vs Age



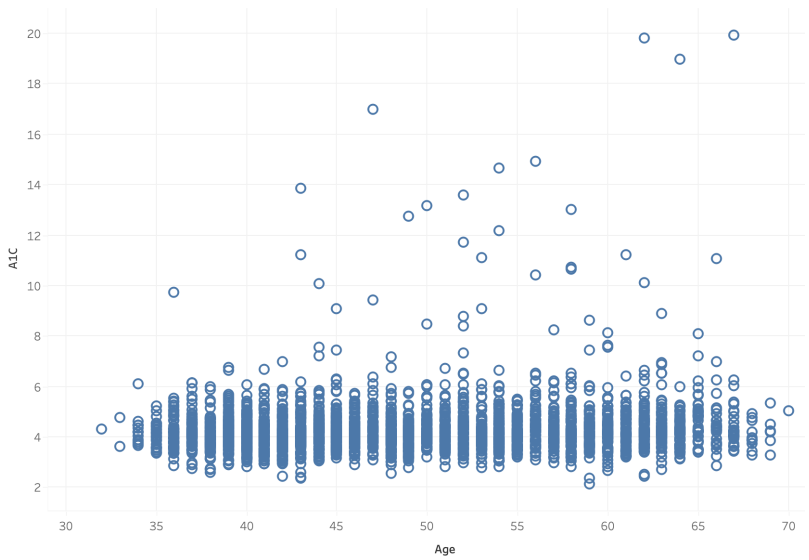
CigsPerDAy vs BMI



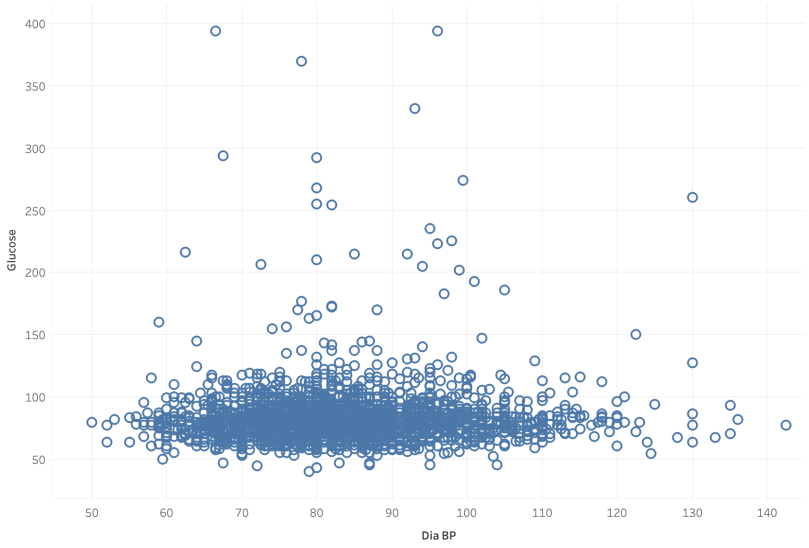
a1c vs BMI



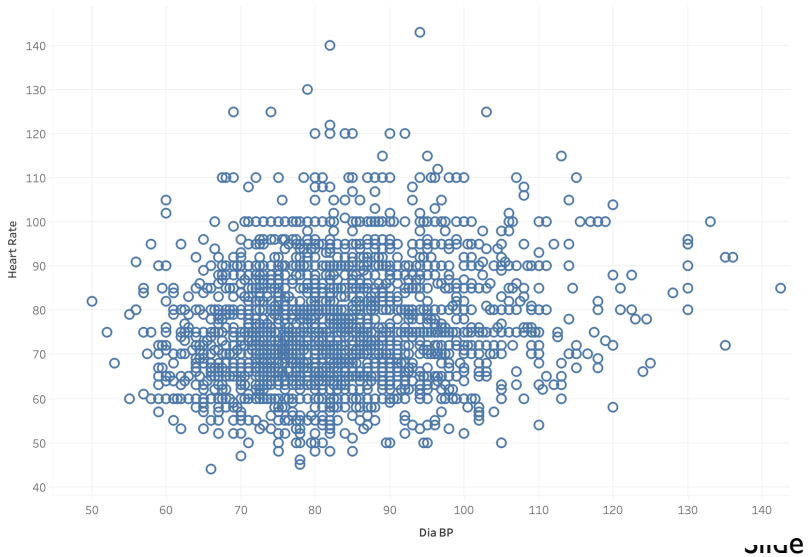
a1c vs age



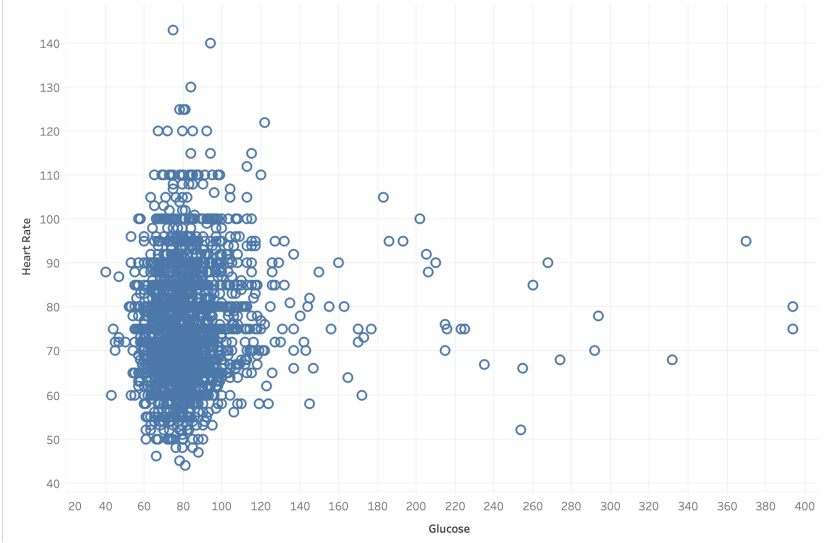
DiaBP vs Glucose



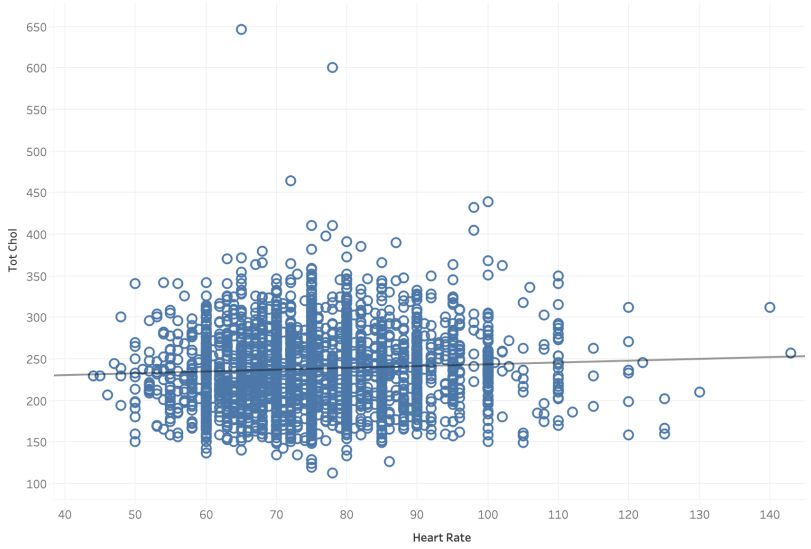
DiaBP vs Heart Rate



glucose vs heart rate



heart rate vs tot Chol



No Significant correlations found

FEATURE SELECTION/ENGINEERING DECISIONS

Features engineering

- Convert education into a categorical variable
- Apply Ln function on sysBP
- Apply Sqrt function on diaBP and BMI
- Apply Log10 function on income, glucose a1c
- Bin cigsPerDay variavble into 6 bins after looking at bivariate analysis and convert them into categories

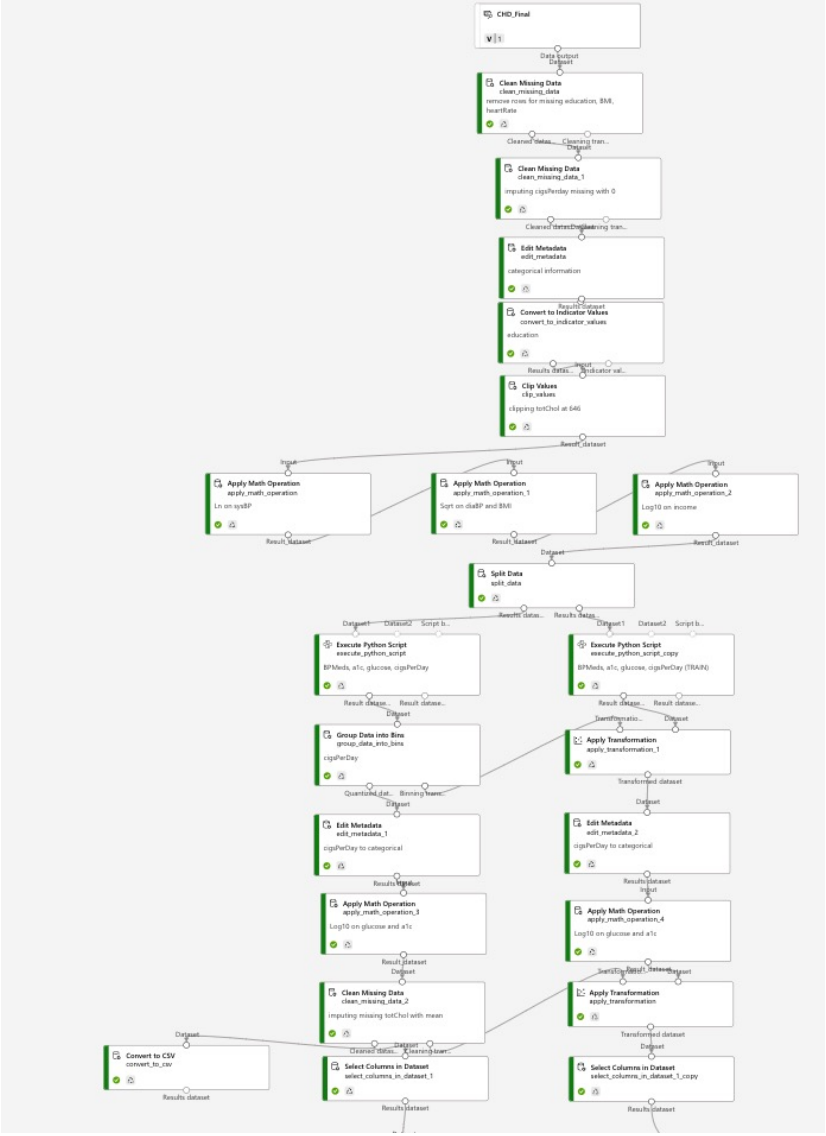
Other Decisions

- Split data into 80-20 and prevent information leakage
- Normalize all measures (except patientID) using Zscore transformation
- Use SMOTE, with 200% and 3 nearest neighbors, as positive obs in response variable are only 15%

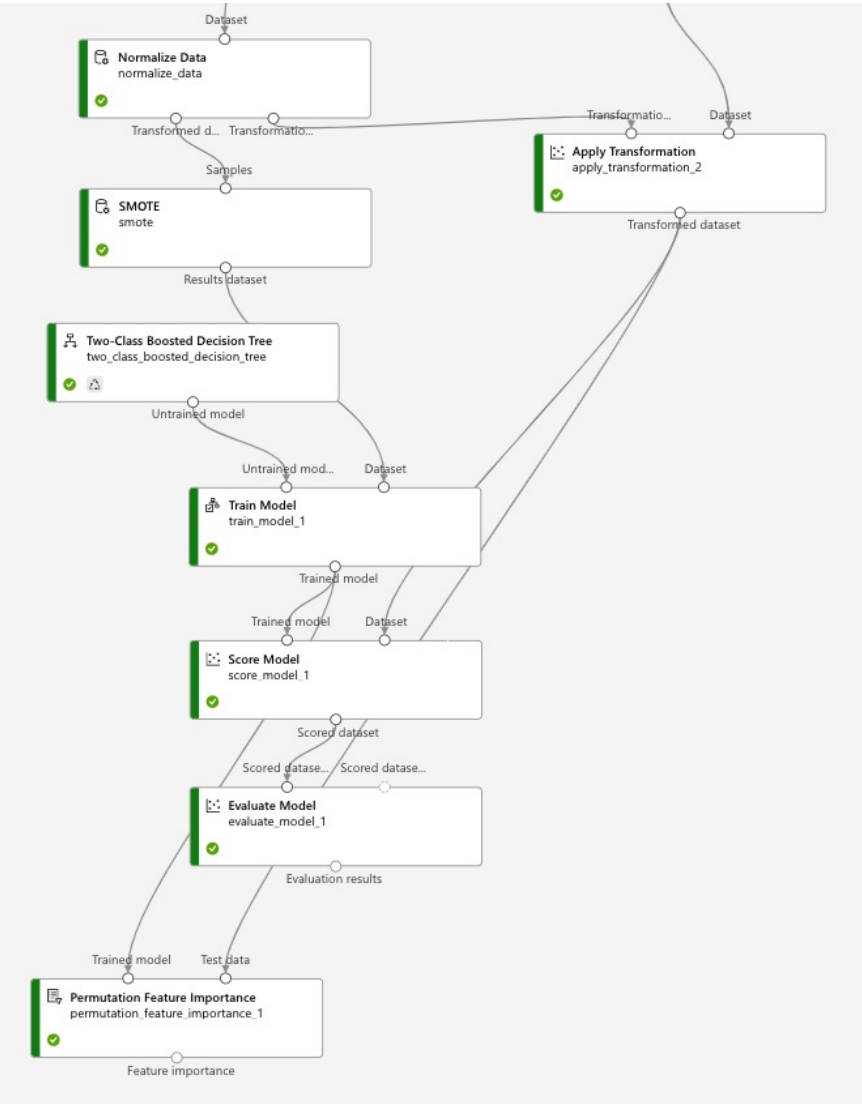
MODEL DEVELOPMENT

USC Viterbi School of Engineering

BUILD MODEL PIPELINE



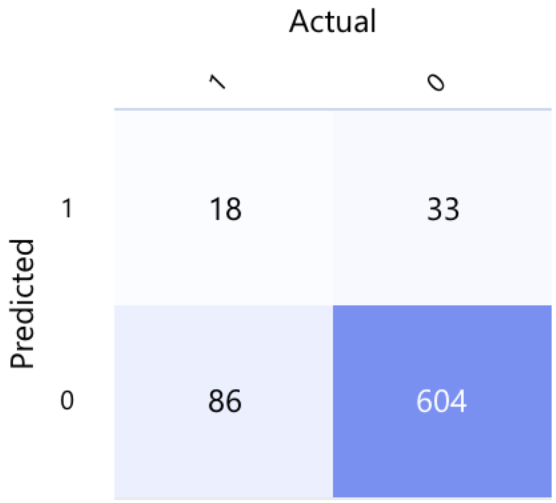
PIPELINE


Dataset
Normalize Data
normalize_data
Transformed d...
Transformatio...
Transformatio...
Dataset
Apply Transformation
apply_transformation_2
Transformed dataset
Samples
SMOTE
smote
Results dataset
Two-Class Boosted Decision Tree
two_class_boosted_decision_tree
Untrained model
Untrained mod...
Dataset
Train Model
train_model_1
Trained model
Trained model
Dataset
Score Model
score_model_1
Scored dataset
Scored datase...
Scored datase...
Evaluate Model
evaluate_model_1
Evaluation results
Trained model
Test data
Permutation Feature Importance
permutation_feature_importance_1
Feature importance

EVALUATION RESULTS

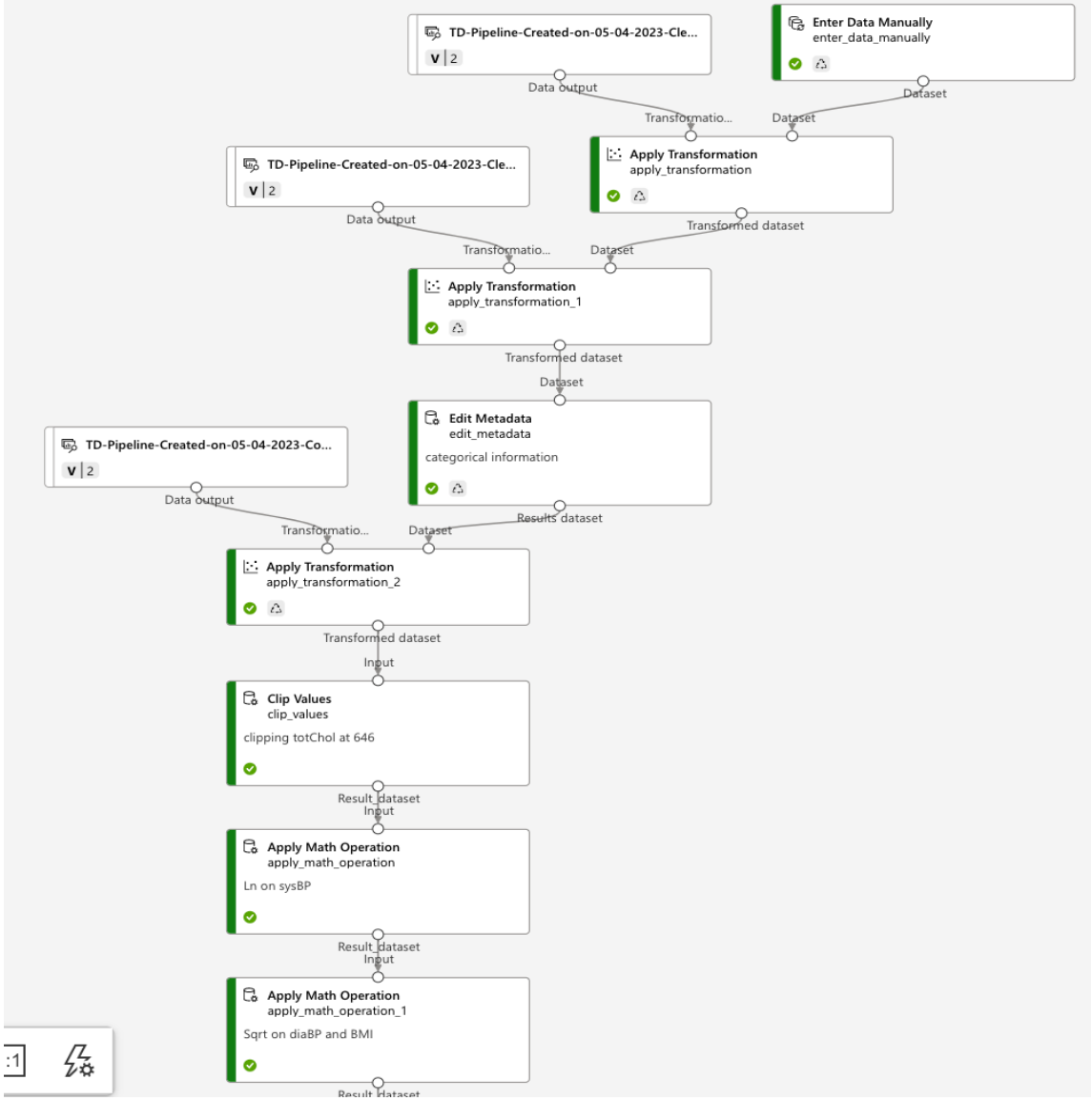
Threshold 0.5

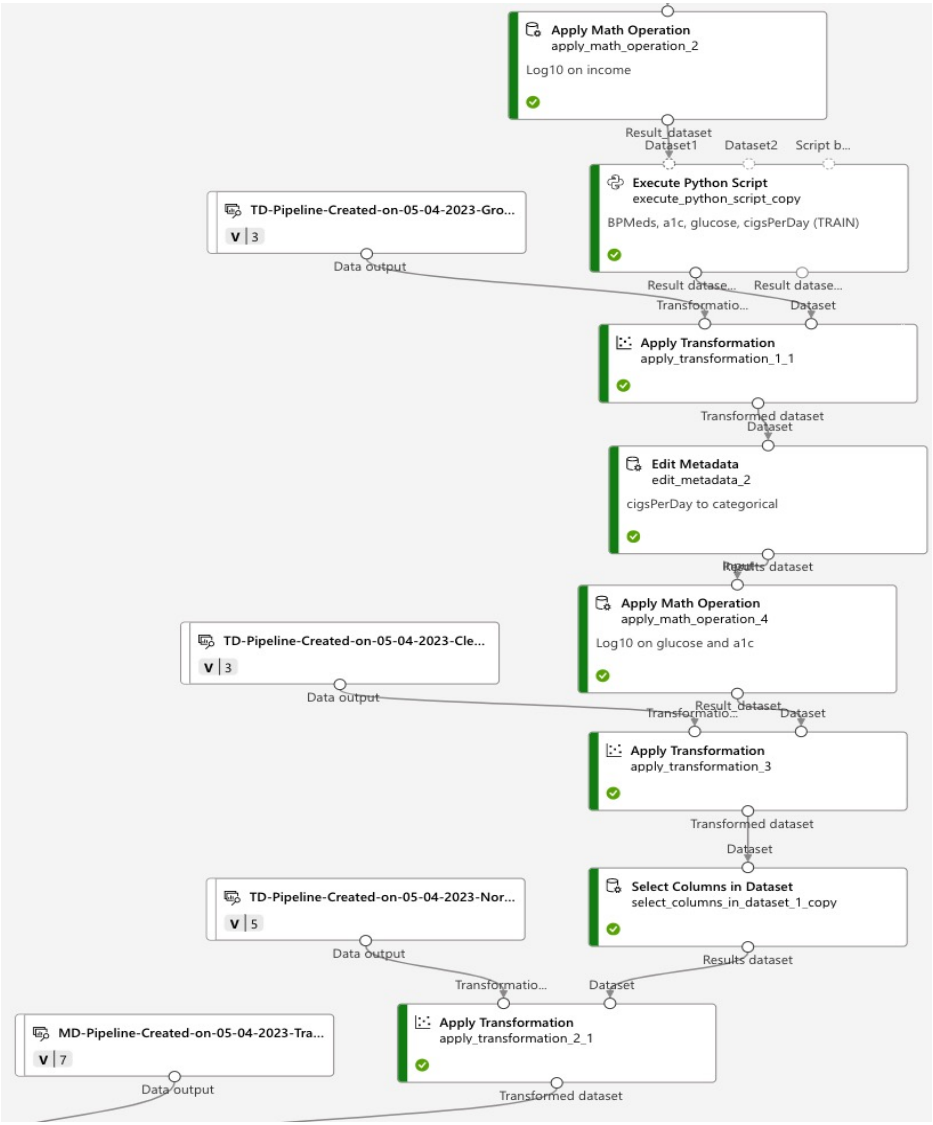
Accuracy	0.839
Precision	0.353
Recall	0.173
F1 Score	0.232
AUC	0.662

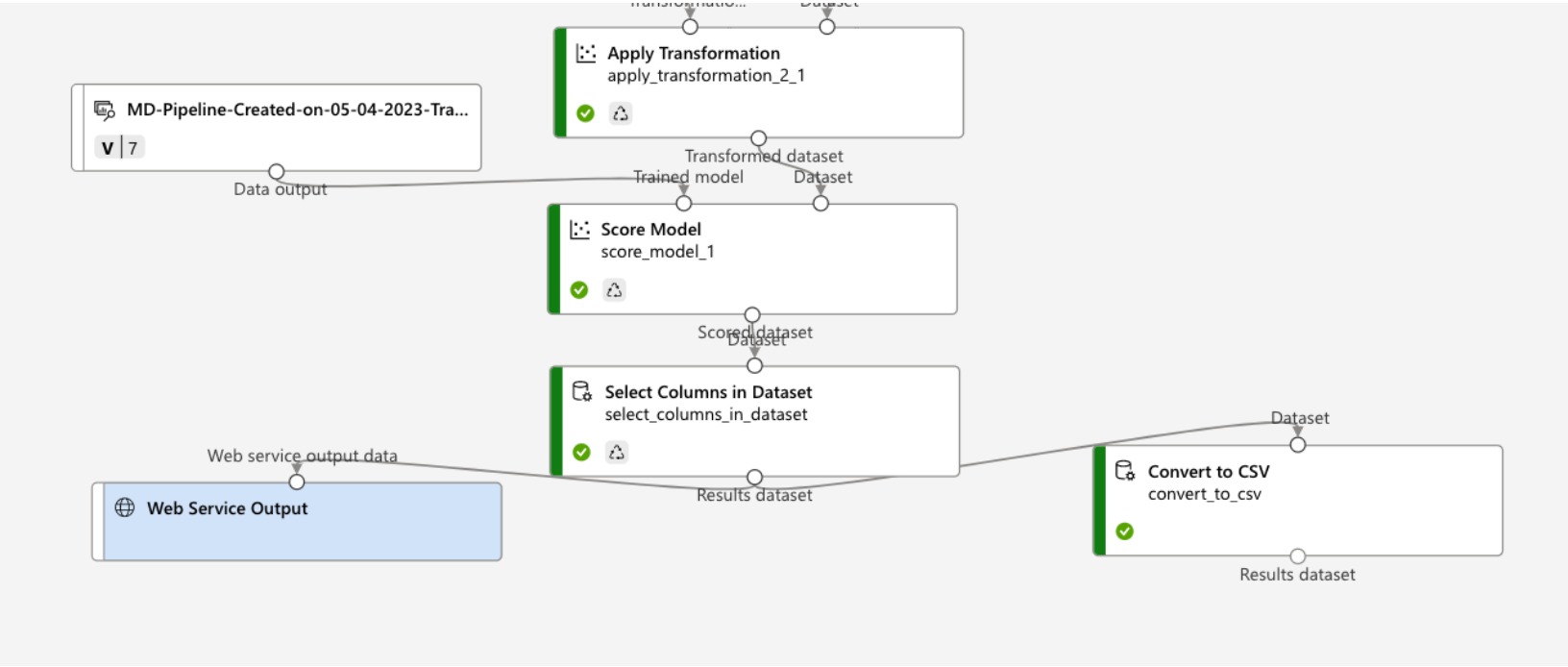


MODEL DEPLOYMENT

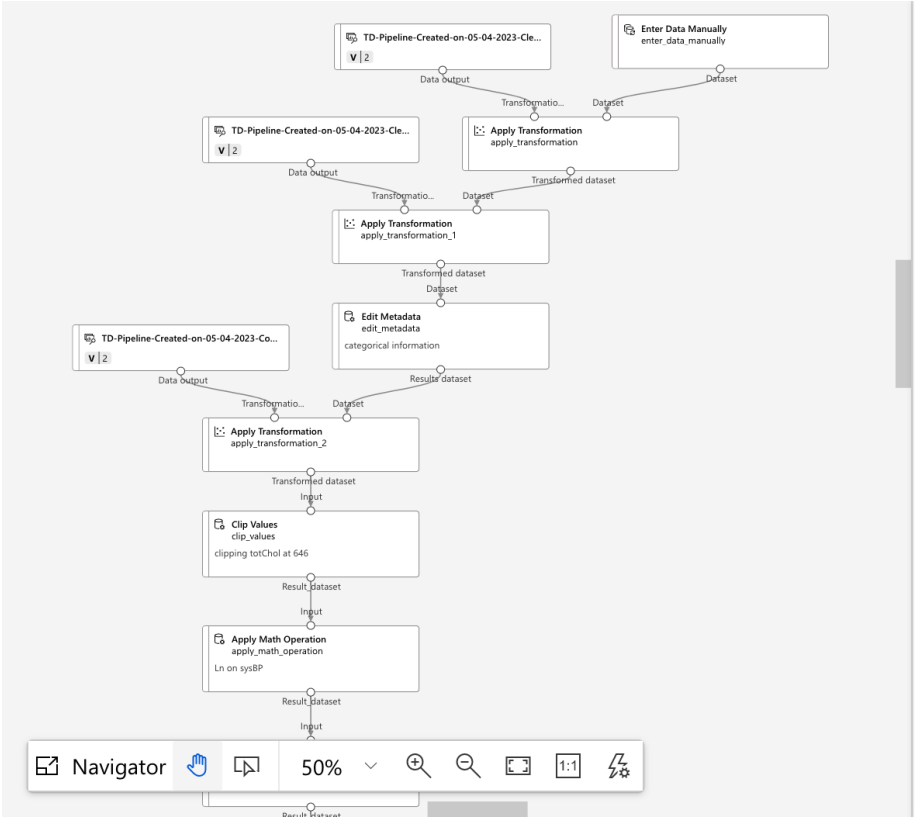
MODEL DEPLOYMENT




Apply Math Operation
apply_math_operation_2
Log10 on income
Result dataset
Dataset1
Dataset2
Script b...
Execute Python Script
execute_python_script_copy
BPMeds, a1c, glucose, cigsPerDay (TRAIN)
TD-Pipeline-Created-on-05-04-2023-Gro...
v | 3
Data output
Result datase...
Result datase...
Transformatio...
Dataset
Apply Transformation
apply_transformation_1_1
Transformed dataset
Dataset
Edit Metadata
edit_metadata_2
cigsPerDay to categorical
Results dataset
Apply Math Operation
apply_math_operation_4
Log10 on glucose and a1c
TD-Pipeline-Created-on-05-04-2023-Cle...
v | 3
Data output
Result dataset
Transformatio...
Dataset
Apply Transformation
apply_transformation_3
Transformed dataset
Dataset
Select Columns in Dataset
select_columns_in_dataset_1_copy
TD-Pipeline-Created-on-05-04-2023-Nor...
v | 5
Data output
Results dataset
Transformatio...
Dataset
Apply Transformation
apply_transformation_2_1
MD-Pipeline-Created-on-05-04-2023-Tra...
v | 7
Data output
Transformed dataset


Apply Transformation
apply_transformation_2_1
MD-Pipeline-Created-on-05-04-2023-Tra...
v | 7
Data output
Transformed dataset
Trained model
Dataset
Score Model
score_model_1
Scored dataset
Dataset
Select Columns in Dataset
select_columns_in_dataset
Web service output data
Web Service Output
Results dataset
Dataset
Convert to CSV
convert_to_csv
Results dataset

MANUAL TEST DATA



Enter Data Manually

Data format *

CSV

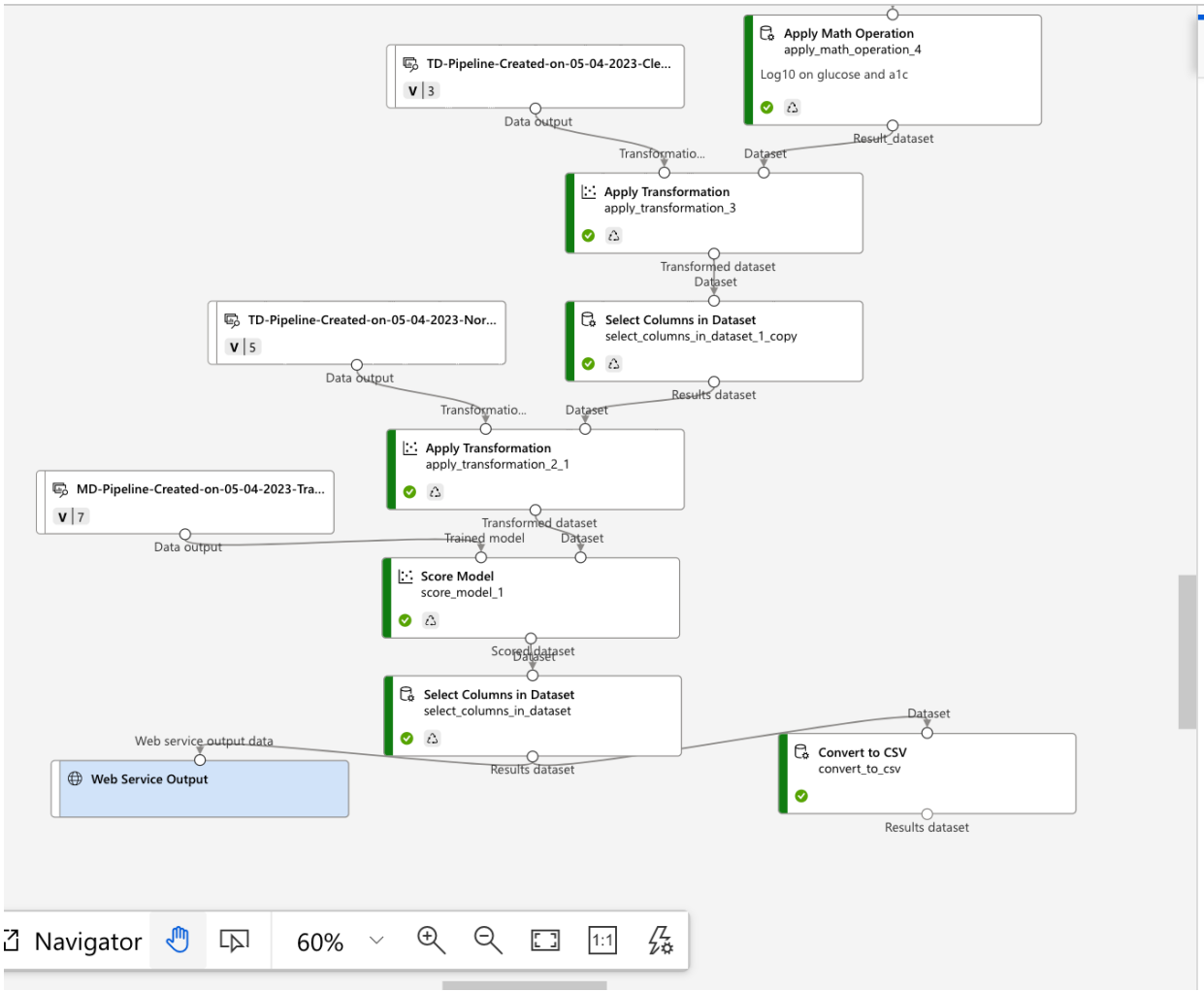
Has header *

True

Data *

```
1   patientID,male,age,education,currentSmoker,cigsPer
2   110399,1,48,3,1,10,0,0,1,0,232,138,90,22.37,64,72,
3   189047,1,41,2,0,,0,0,0,0,195,139,88,26.88,85,65,3.
4   957019,1,54,1,1,20,0,0,1,0,214,147,74,24.71,96,87,
5   208967,1,37,2,0,,0,0,1,0,225,124.5,92.5,38.53,95,8
6   230935,0,63,1,1,3,0,0,1,0,267,156.5,92.5,27.1,60,7
7   216024,1,57,1,0,,0,0,0,0,220,136,84,26.84,75,64,3.
8   368834,0,56,1,0,,0,0,1,0,296,180,90,23.72,75,120,6
9   135175,0,48,1,0,,0,0,1,0,265,145,77,24.23,74,64,3.
10  294070,1,66,3,0,,0,0,0,0,288,109,71,29.29,80,80,4.
11  595710,1,38,4,0,,0,0,0,0,235,118,77,25.87,60,82,4.
12  425597,1,53,1,1,30,0,0,0,0,244,106,67.5,21.84,88,6
13  650137,0,59,1,1,1,0,0,1,0,259,141,86,25.97,70,86,4
14  590019,0,38,3,1,3,1,0,1,0,,125,80,22.79,98,,,13340
15  925626,0,36,3,1,20,0,0,0,0,159,121.5,73,20.41,72,7
16  276518,1,41,4,1,20,1,0,1,0,244,139,86,30.77,60,67,
17  342284,0,37,1,0,,0,0,0,0,300,112,60,23.67,81,75,3.
```

TEST DATA SCORING



Results_dataset

Rows 174

Columns 2

patientID	Scored Labels
110399	1
189047	0
957019	1
208967	0
230935	0
216024	0
368834	0
135175	1
294070	0
595710	0
425597	0
650137	1
590019	0