FINAL PROJECT
ISE 543, SPRING 2023

XIAOYI WANG
2403234885

INSTRUCTIONS

- For this exam, you are to create a predictive model in Azure ML Studio for the attached dataset and turn in a report as specified in the following pages. You should use whichever data preparation, modeling, and model assessment techniques that were covered in this portion of the class that you believe result in the best model.
- You will be performing an Exploratory Data Analysis, Model Development and Training, and Model Deployment activities and preparing a report in PowerPoint form
- See the sample report that is part of this assignment for a template and example
- When you are complete, save this file as a PDF and upload it to Gradescope
- It is due at 11:00PM on Monday, May 8
- As a reminder, the work that you submit must be done individually. Unlike the homework assignments, working together is not permitted and the graders will be looking for identical solutions.

INSTRUCTIONS

For this exam, you 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 you will be using has been distributed with this exam and consists of the variables on the following page

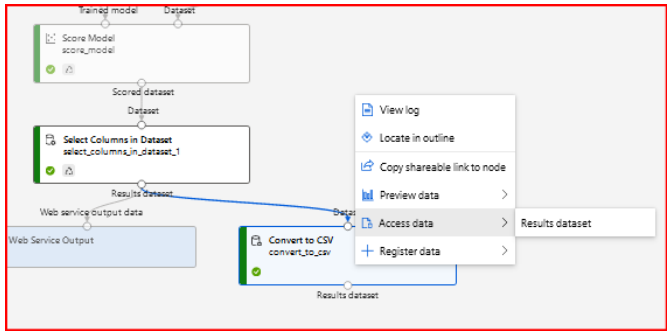
USC Viterbi School of Engineering

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)

NOTE ON MODEL DEPLOYMENT - UPDATE

- When complete, do a final run of your inference pipeline by copying test data from the file “Final Project Evaluation Dataset.csv” into the “Enter Data Manually” component (see Sample Final Report for an example)
  - Do this by opening the file in WordPad or a similar text editor and then copying the data from there
- Include a “Convert to CSV” component at the end of your inference pipeline (see Sample Final Report for an example)
- Download the CSV file containing your labels (predictions) by right-clicking the Convert to CSV component, selecting Access data / Results Dataset



NOTE ON MODEL DEPLOYMENT - UPDATE

- This will take you to an Azure data folder that will contain a file called data.csv. Download this file (right-click, "Download") to your laptop and rename it to be "Final scored dataset – xxxxxx.csv" where xxxxxx is your student ID number
- Upload this file to this Google folder:
  - https://drive.google.com/drive/folders/1rIrbiYMkQHoWlA3Km_moODAnsfHIksQS?usp=sharing
- The TAs will use this file to calculate the AUC for your model (by comparing it to the actual values of the response variable)

FINAL REPORT STRUCTURE

Please follow the provided template/example and structure your final report into the following three sections:

- Exploratory Data Analysis
- Model Development
- Model Deployment

FINAL REPORT OUTLINE/GRADING RUBRIC

Report contents: 80%

- Attribute summary
- Data cleansing – summary of decisions made
- Data cleansing pipeline (portion of your overall pipeline)
- Univariate analysis
- Bivariate analysis (each variable vs the response variable)
- Feature section/engineering decisions
- Model pipeline screenshot
- Model evaluation results screenshot
- Inference pipeline screenshot
- REST Endpoint URL and authentication key (in PPT and in Google drive spreadsheet)
- Screenshot of scored test dataset

Model performance: 20%

- Based on TAs calling your endpoint with test data

ATTRIBUTE SUMMARY

Response Variable

- Ten Year CHD

Categories

- PatientID (UID)
- Male
- Education
- Current Smoker
- BP Meds
- Prevalent Stroke
- Pravalent Hyp
- Diabetes

Measures

- Age
- Income
- Cigarettes per Day
- Total Chol
- Sys BP
- Dia BP
- BMI
- Heart Rate
- Glucose
- A1c

DATA CLEANSING

Rows 19 Columns 23

Feature	Count	Unique Value Count	Missing Value Count	Min	Max	Mean	Mean Deviation	1st quantile	Median	3rd quantile	Mode	Range	Sample Variance	S S D
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,	899824	66073719998.68935	2
male	3816	2	0	0	1	0.427673	0.489538	0	0	1	0	1	0.244833	0
age	3816	39	0	32	70	49.567348	7.350087	42	49	56	40	38	73.918923	8
education	3723	4	93	1.0	4.0	1.974483	0.812462	1	2	3	1.0	3.0	1.037769	1
currentSmoker	3816	2	0	0	1	0.489518	0.49978	0	0	1	0	1	0.249956	0
cigsPerDay	1841	31	1975	1.0	70.0	18.500272	8.12733	10	20	20	20.0	69.0	119.365353	1
BPMeds	3771	2	45	0.0	1.0	0.02917	0.056638	0	0	0	0.0	1.0	0.028327	0
prevalentStroke	3816	2	0	0	1	0.006027	0.011982	0	0	0	0	1	0.005992	0
prevalentHyp	3816	2	0	0	1	0.306604	0.425196	0	0	1	0	1	0.212654	0
diabetes	3816	2	0	0	1	0.024895	0.048551	0	0	0	0	1	0.024282	0
totChol	3769	246	47	107.0	9280.0	240.852746	39.378735	205	234	263	240.0	9173.0	35695.202567	1
sysBP	3816	232	0	83.5	295.0	132.260089	16.890304	117	128	143.5	130.0	211.5	489.279059	2
diaBP	3816	142	0	50.0	142.5	82.874214	9.158874	75	82	89.5	80.0	92.5	141.791251	1
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
heartRate	3815	73	1	44.0	143.0	75.775098	9.288791	68	75	82	75.0	99.0	144.885431	1
glucose	3455	134	361	40.0	394.0	81.856151	12.250119	71	78	87	75.0	354.0	555.598004	2
TenYearCHD	3816	2	0	0	1	0.15173	0.257415	0	0	0	0	1	0.128741	0
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
income	3816	3282	0	12000.0	524494.0	20355.886792	7743.282613	13562.5	16055	21395.5	14623.0	512494.0	319178936.087836	1

many missing values

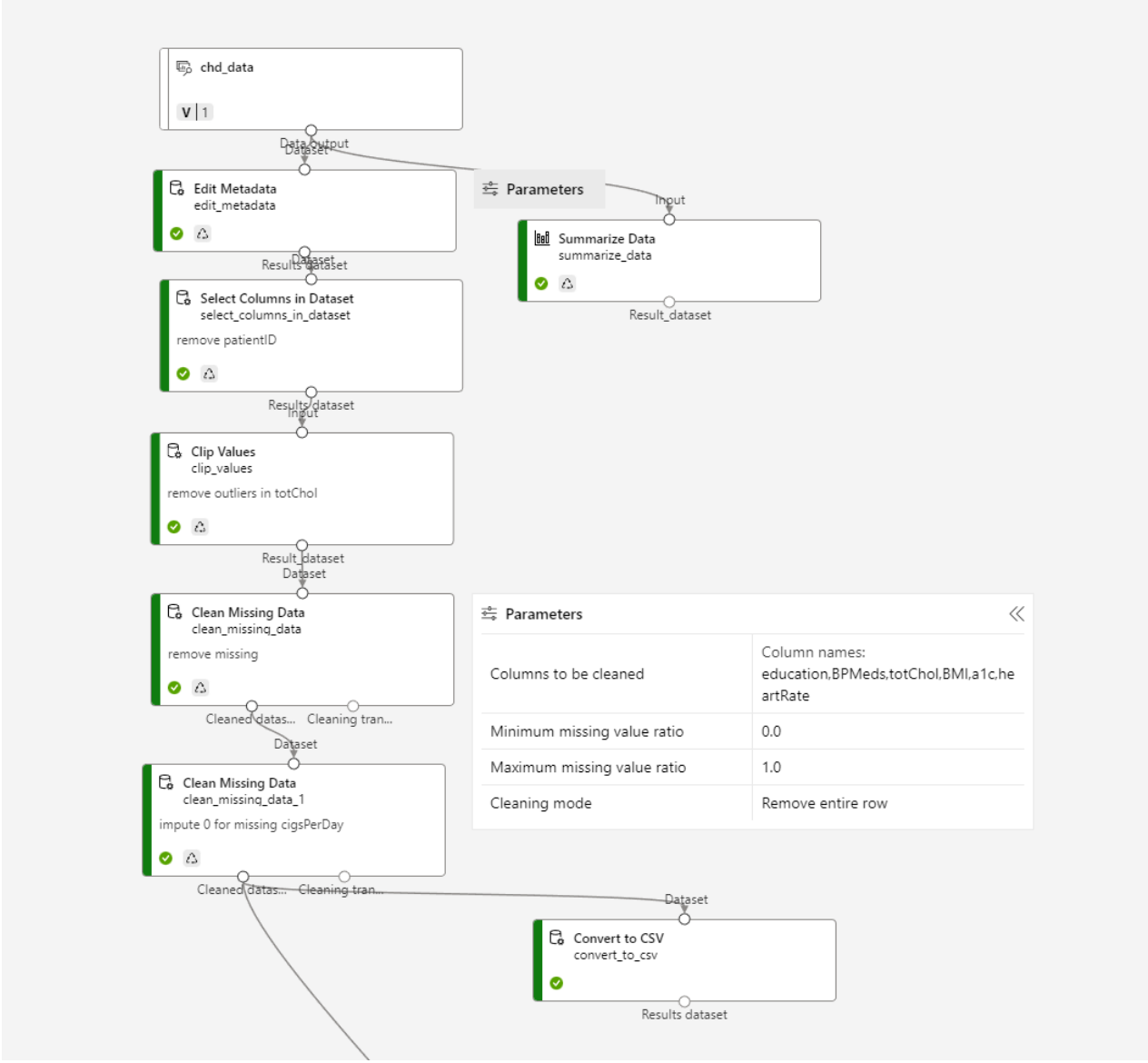
large outliers

DATA CLEANSING

- change response variable TenYearCHD to categorical
- outliers in totChol: seem to be erroneous data. clip outliers in totChol at a threshold of 600
- cigarettes per day: impute missing values with 0 because these non-responses are mostly due to non current smokers
- remove rows with missing values: education, BPMeds, totChol, BMI, heartRate, glucose, a1c
- (Note: normalization and transformation are done in feature engineering)

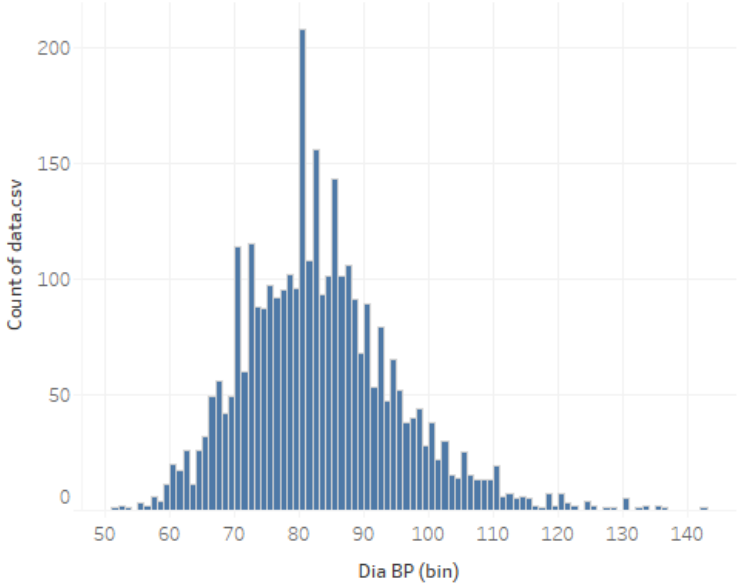
USC Viterbi School of Engineering

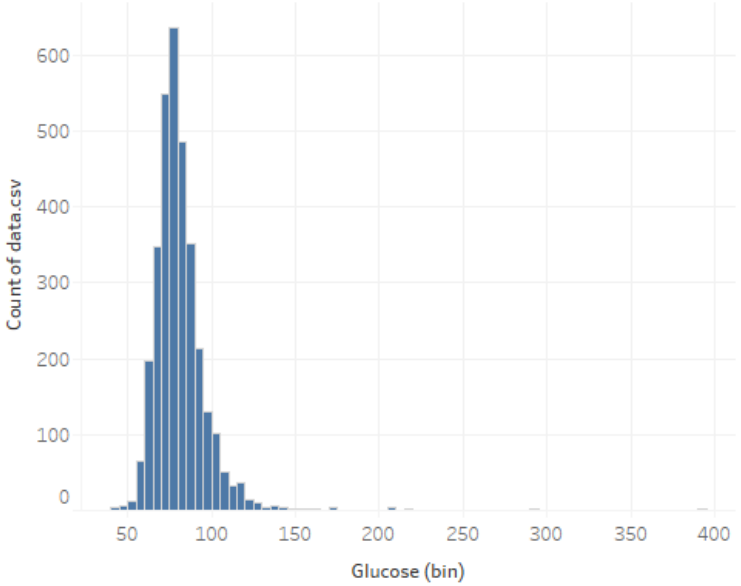
DATA CLEANSING PIPELINE

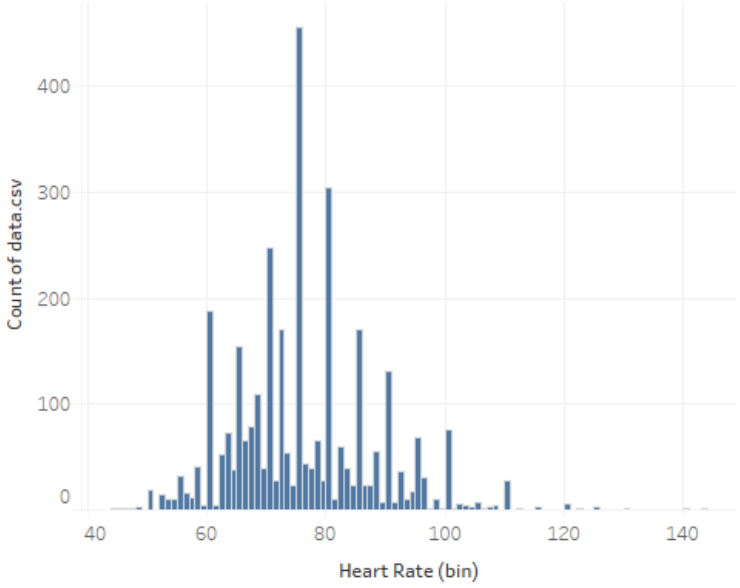


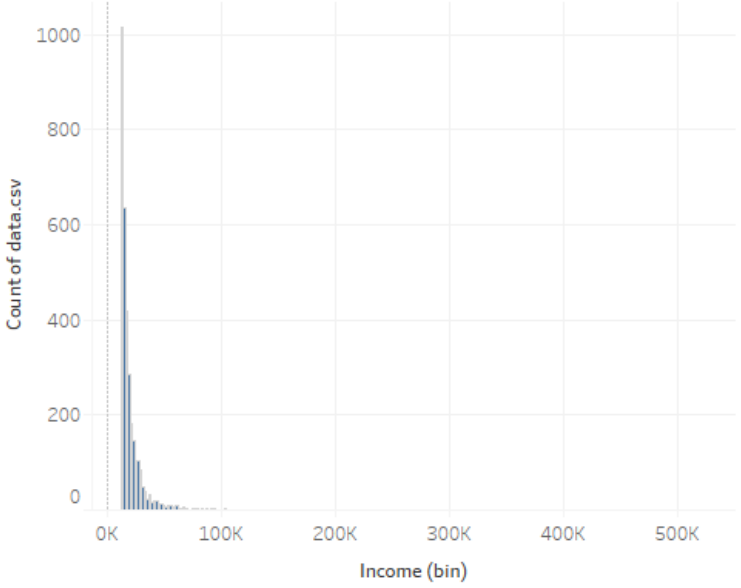
USC Viterbi School of Engineering

UNIVARIATE ANALYSIS



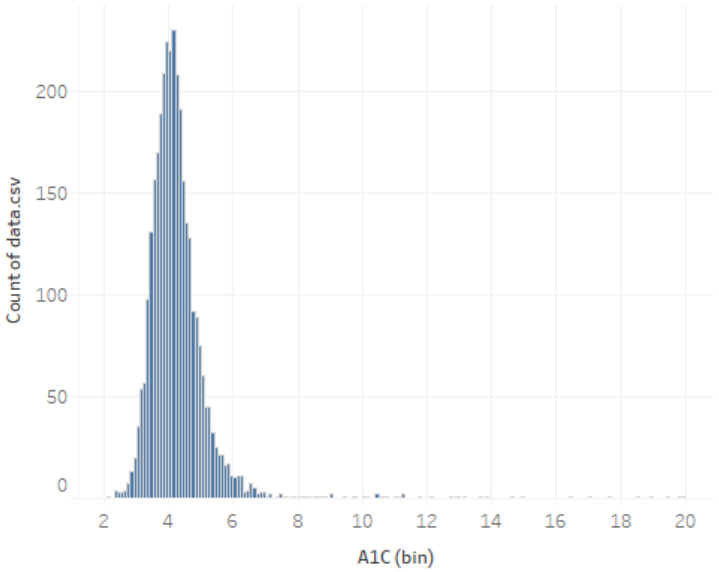


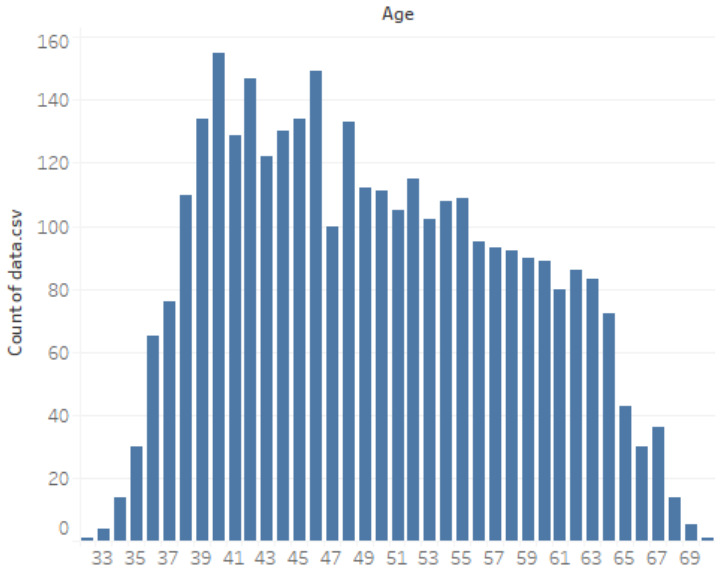


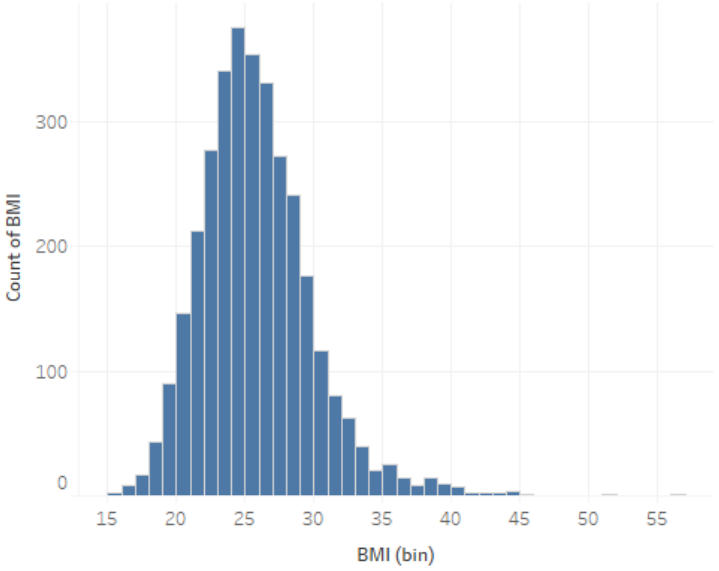


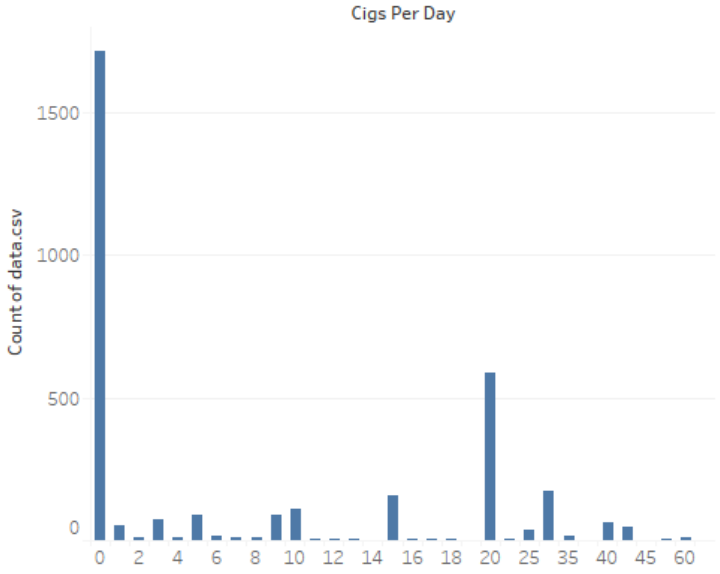
USC Viterbi School of Engineering

UNIVARIATE ANALYSIS

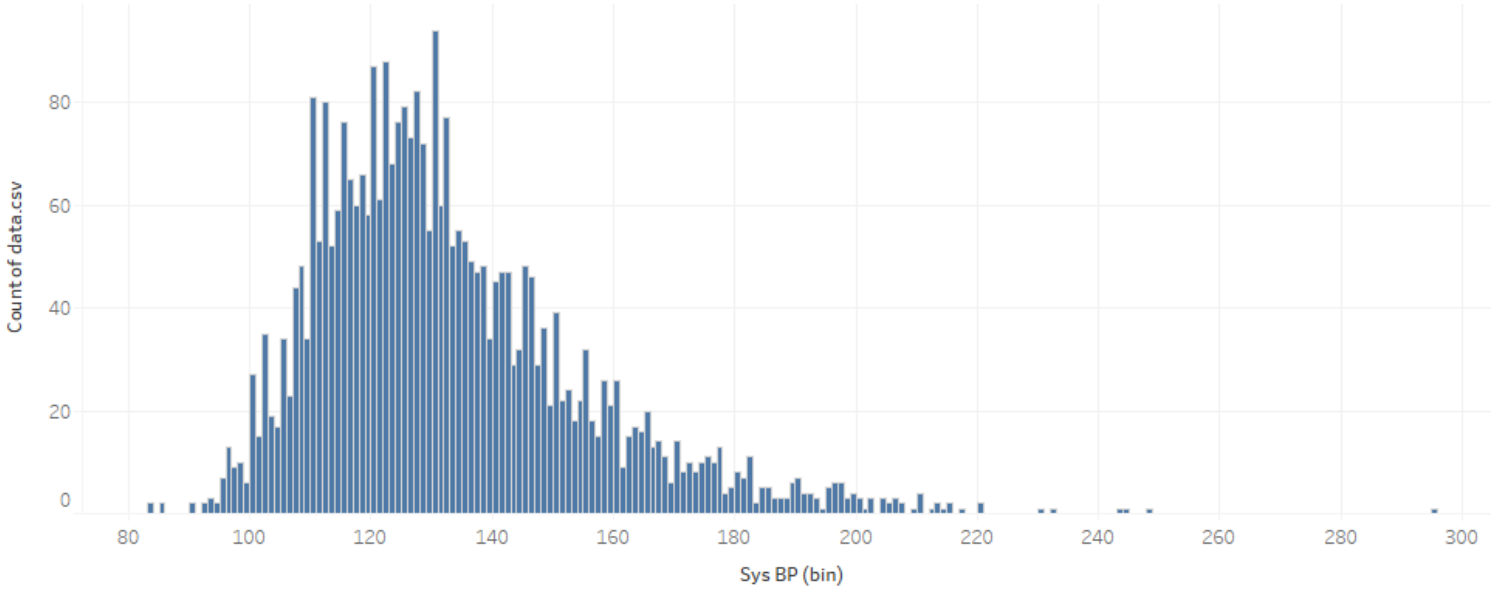


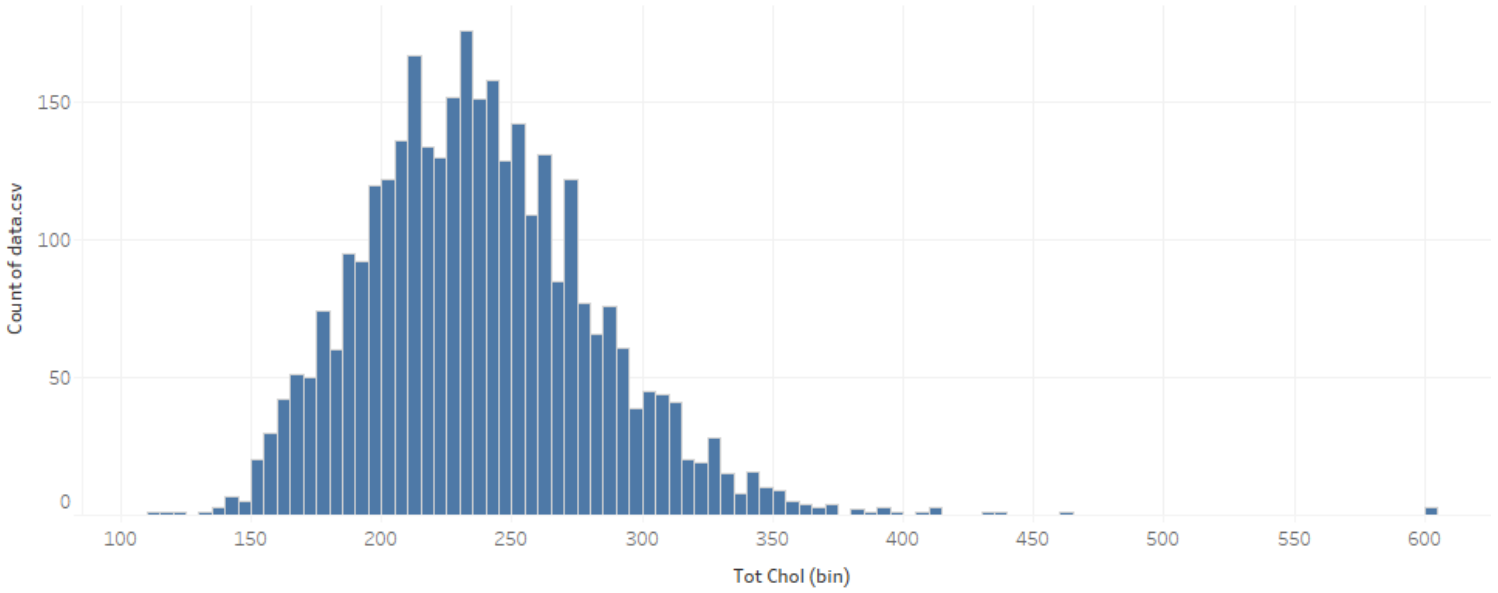






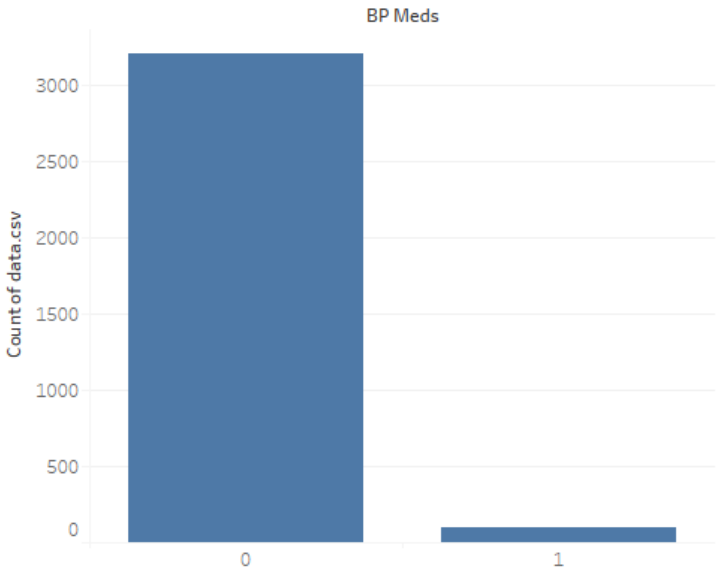
UNIVARIATE ANALYSIS

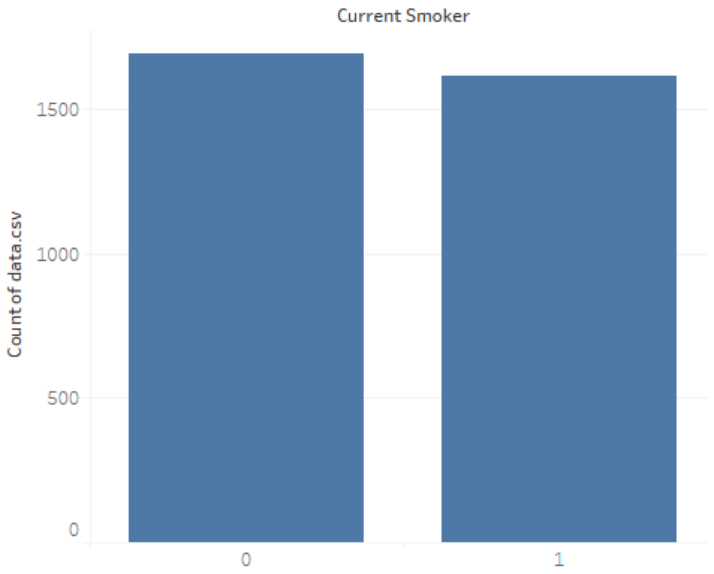


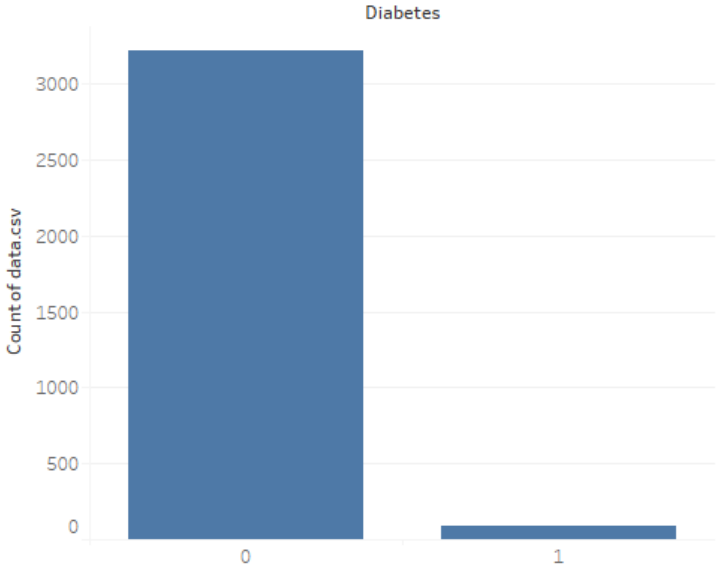


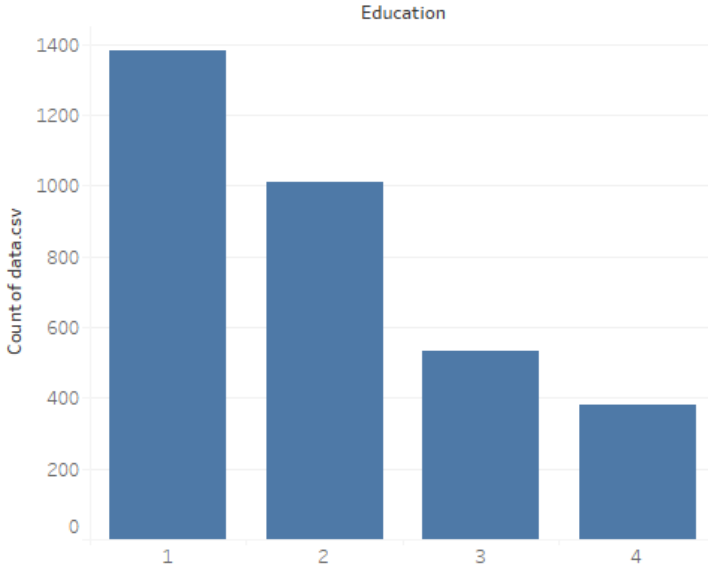
USC Viterbi School of Engineering

UNIVARIATE ANALYSIS

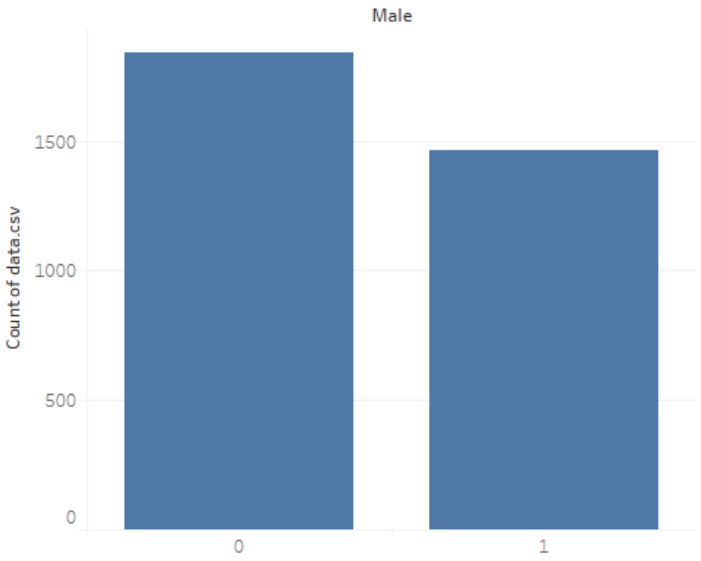


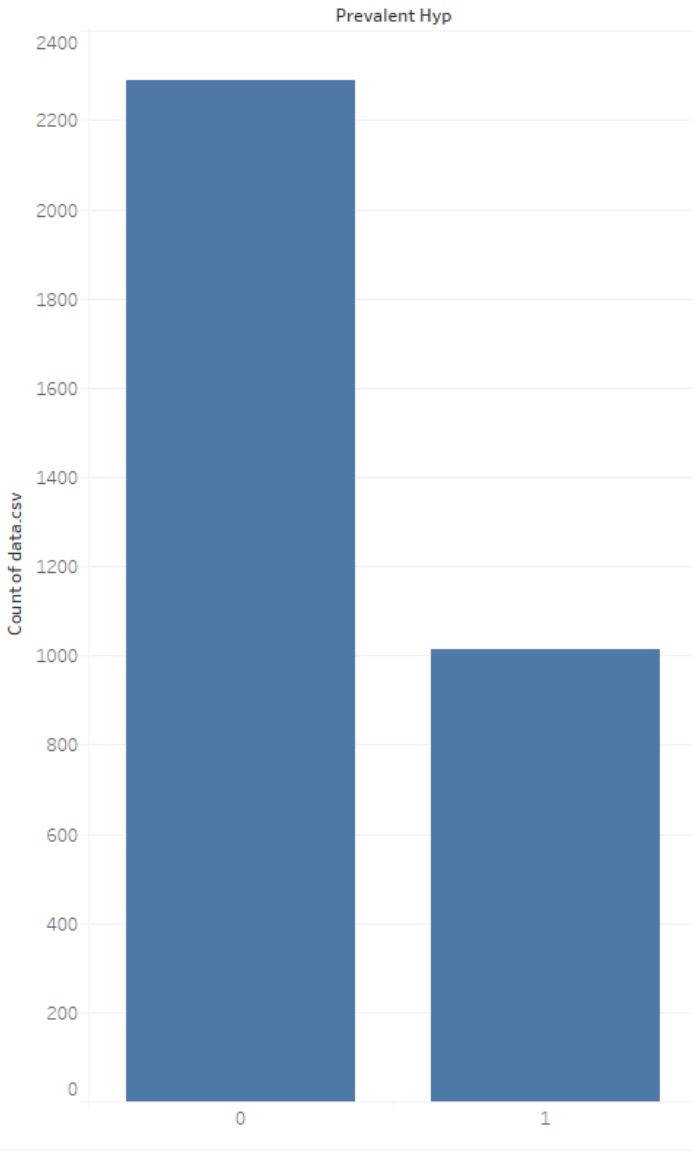


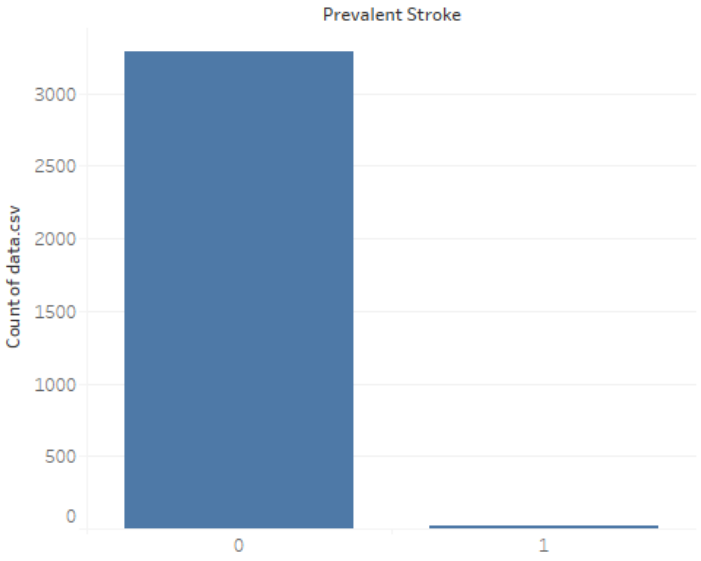




UNIVARIATE ANALYSIS

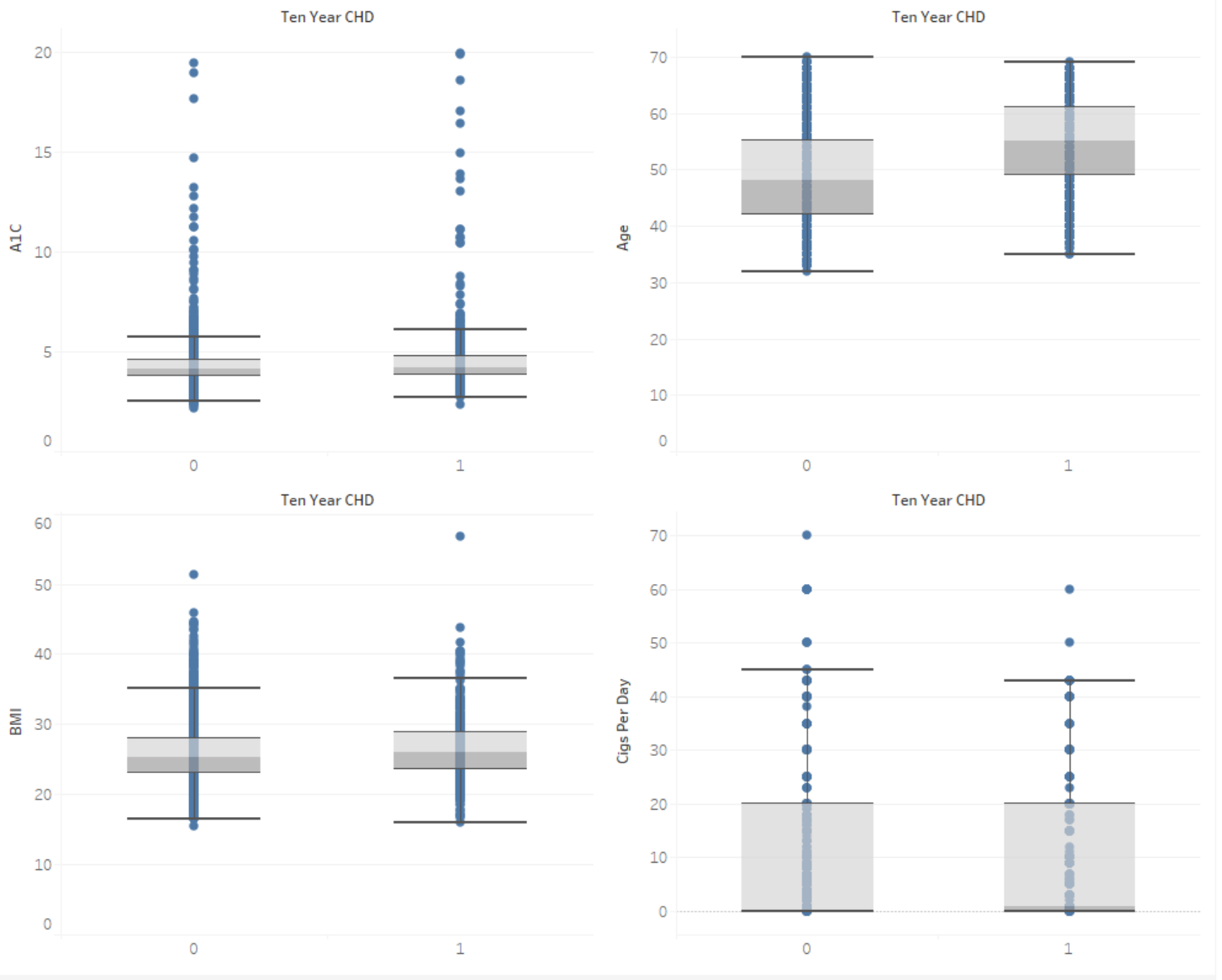




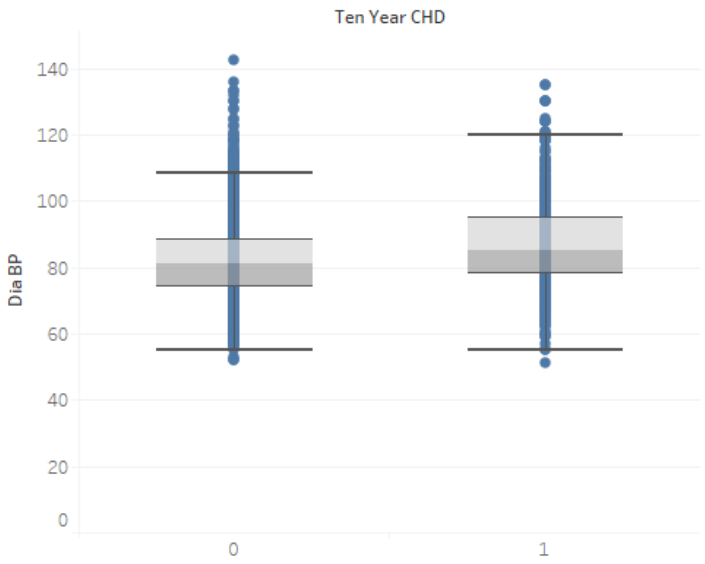


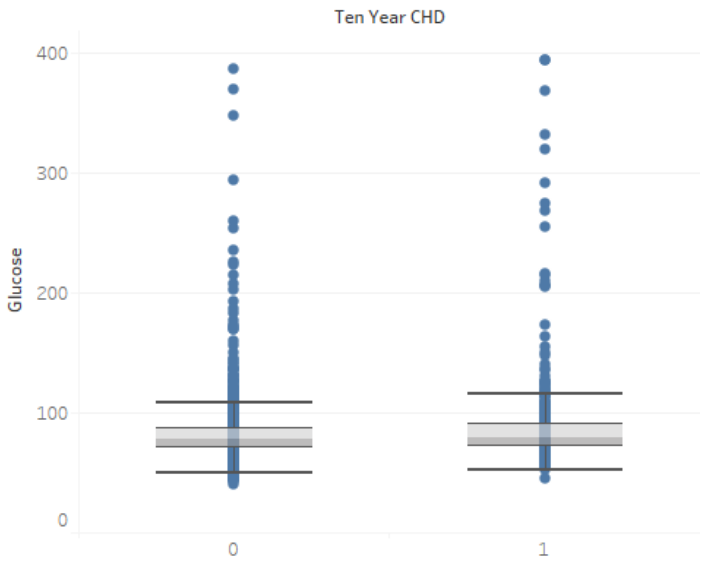
USC Viterbi School of Engineering

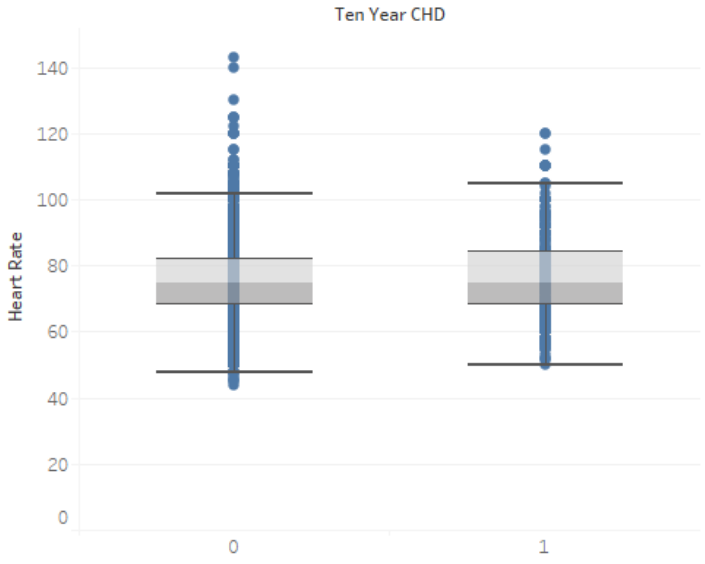
BIVARIATE ANALYSIS

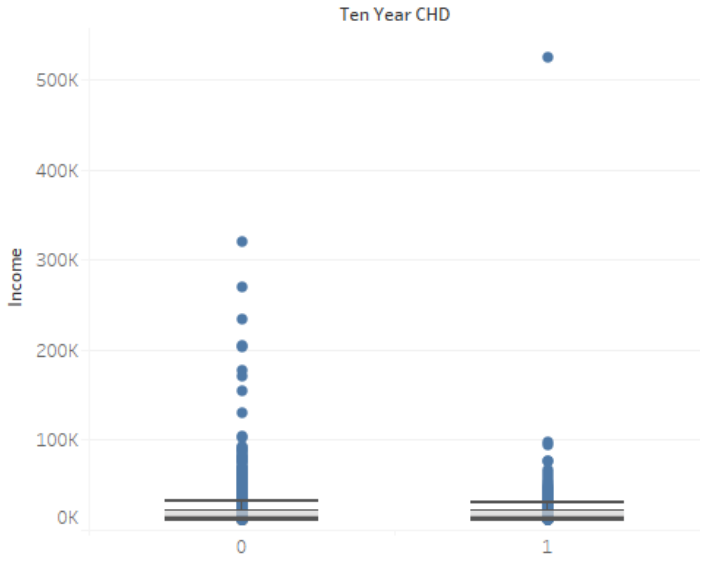


BIVARIATE ANALYSIS

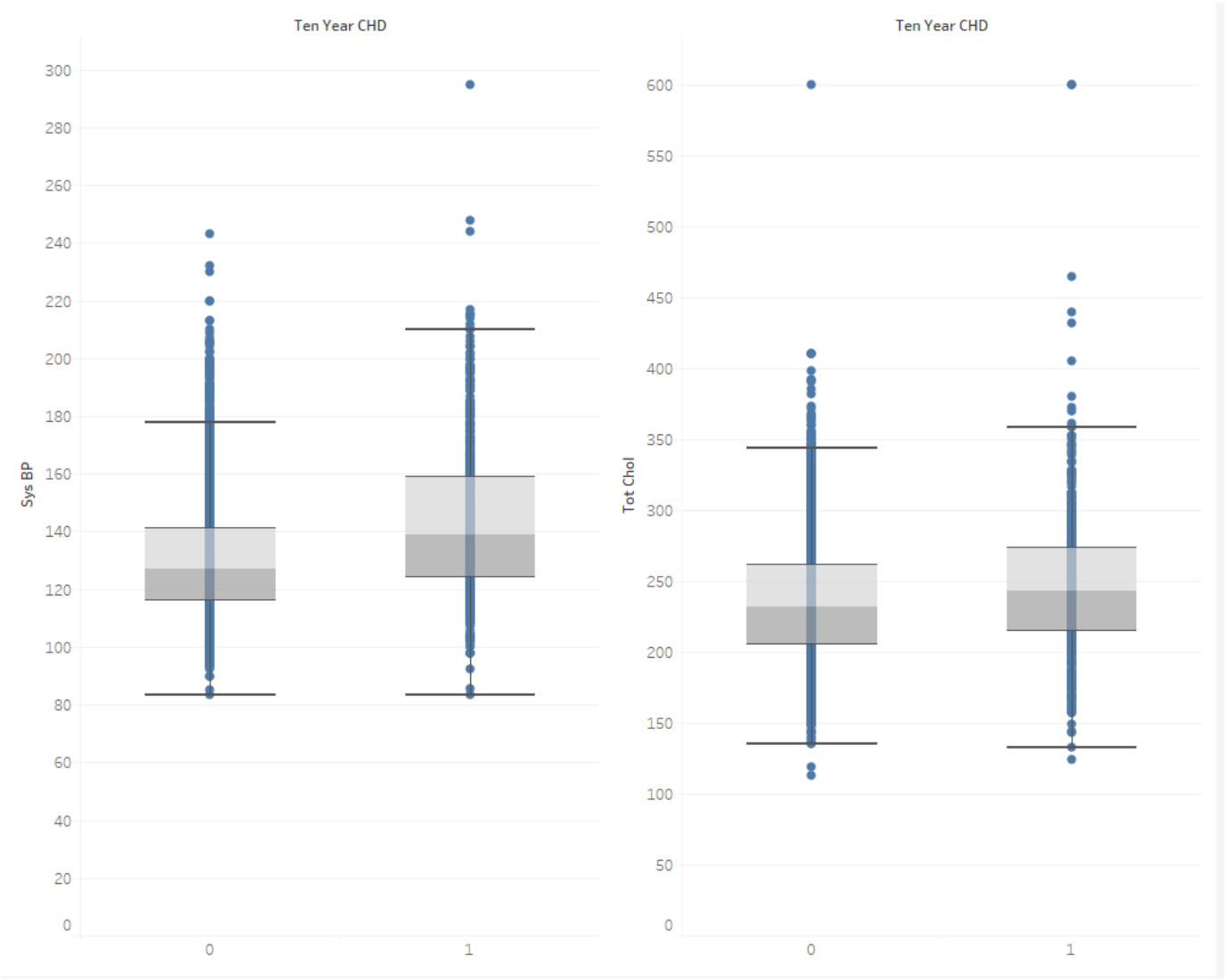






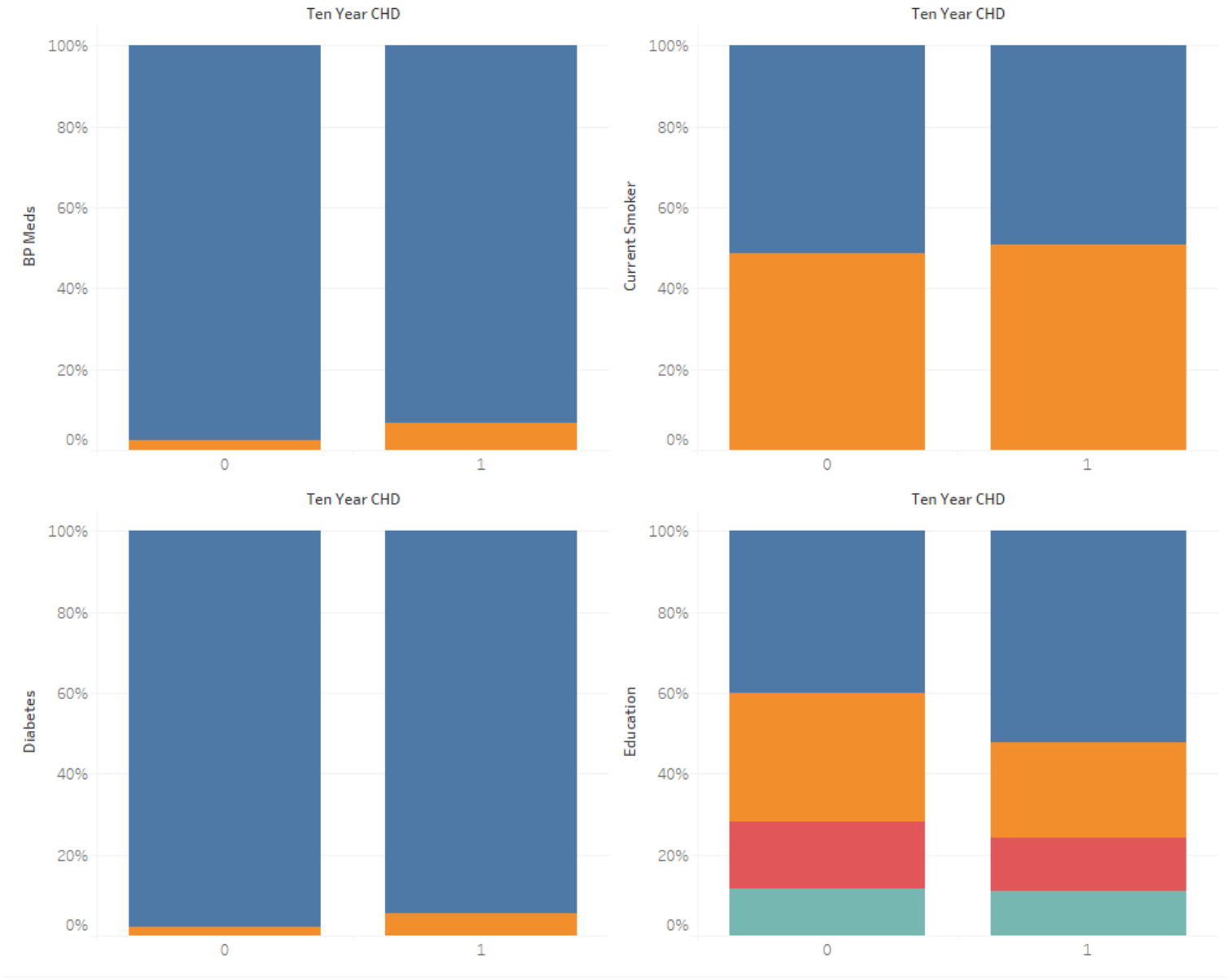


BIVARIATE ANALYSIS

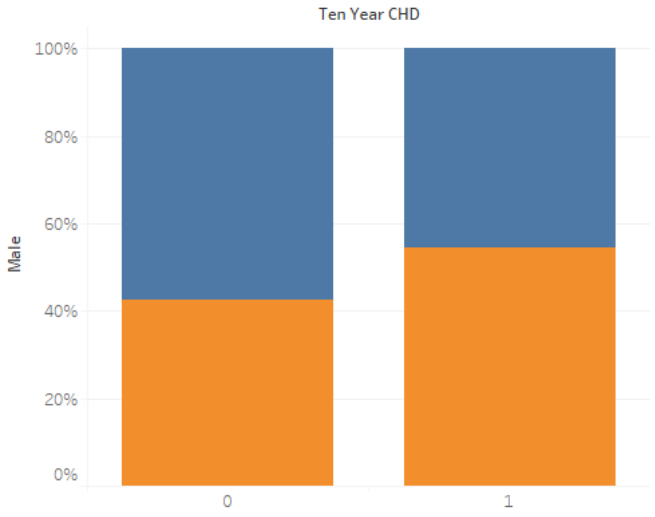


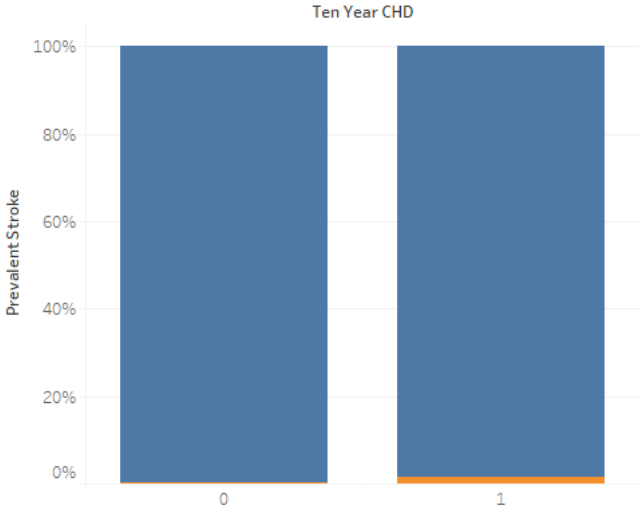
USC Viterbi School of Engineering

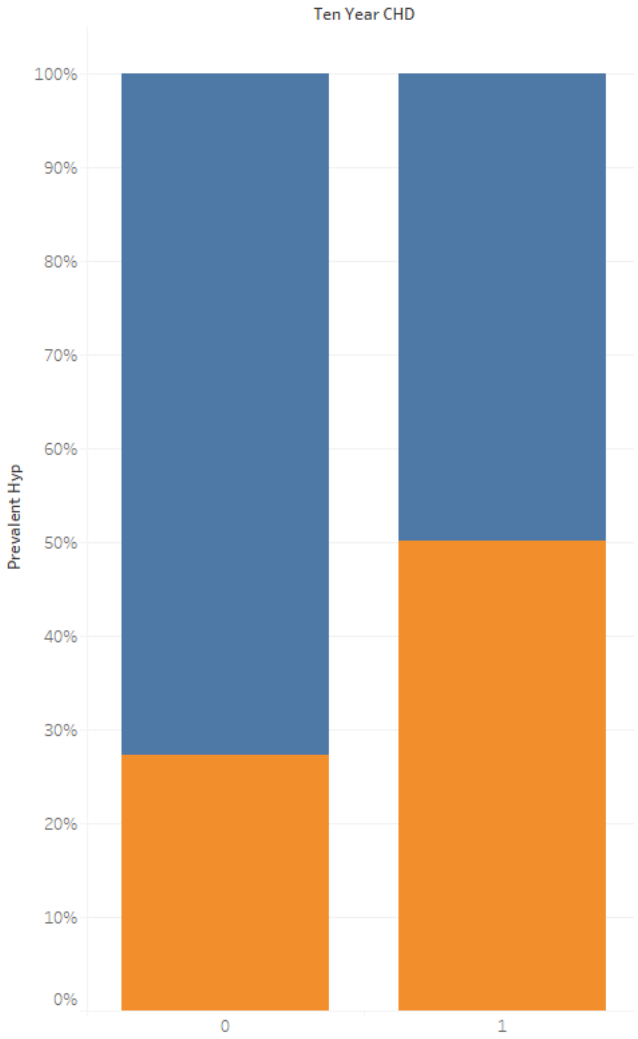
BIVARIATE ANALYSIS



BIVARIATE ANALYSIS

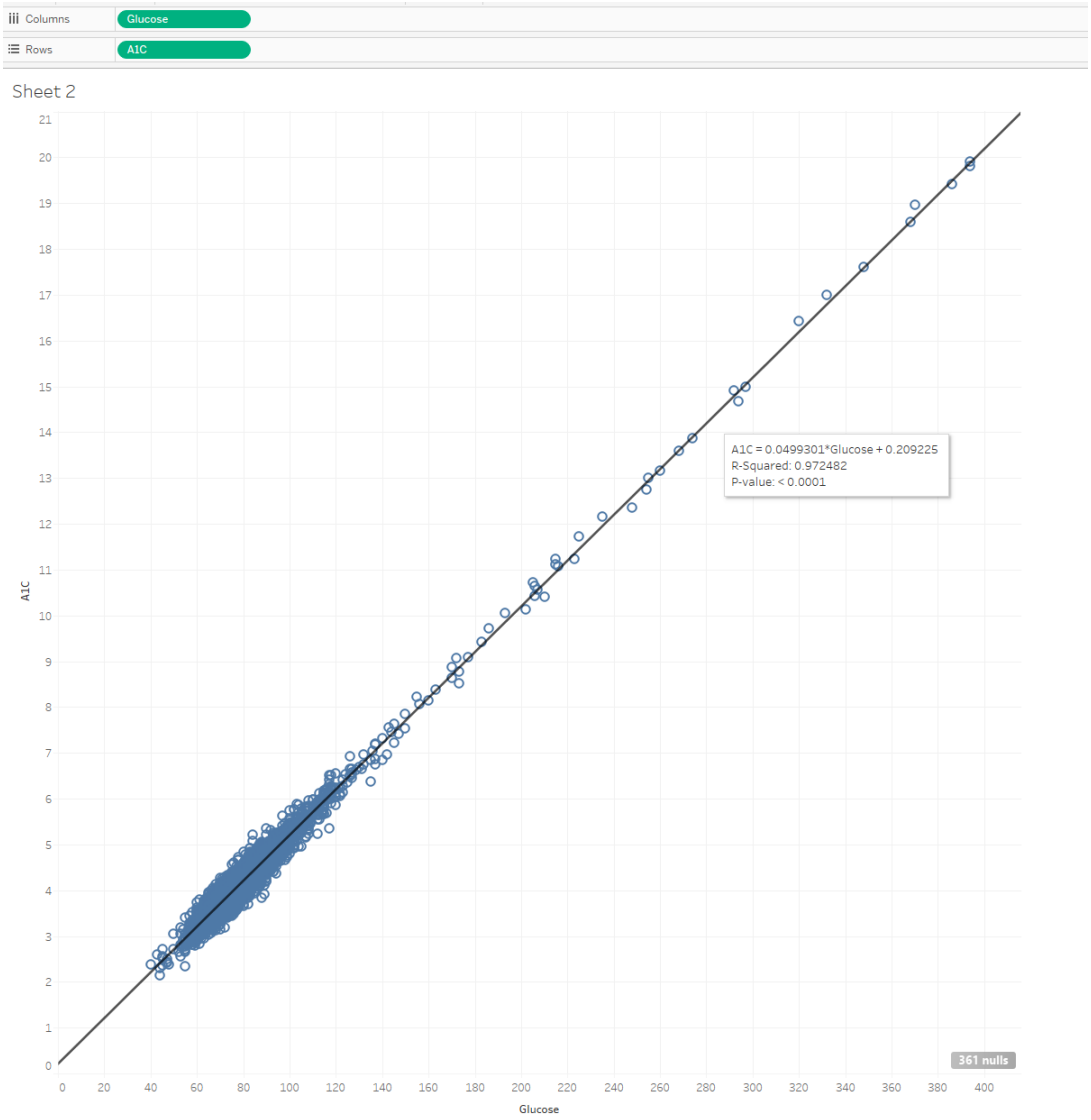






CORRELATION ANALYSIS

- glucose vs. A1C: very strong correlation



CORRELATION ANALYSIS

- Dia BP vs. Sys BP: strong correlation

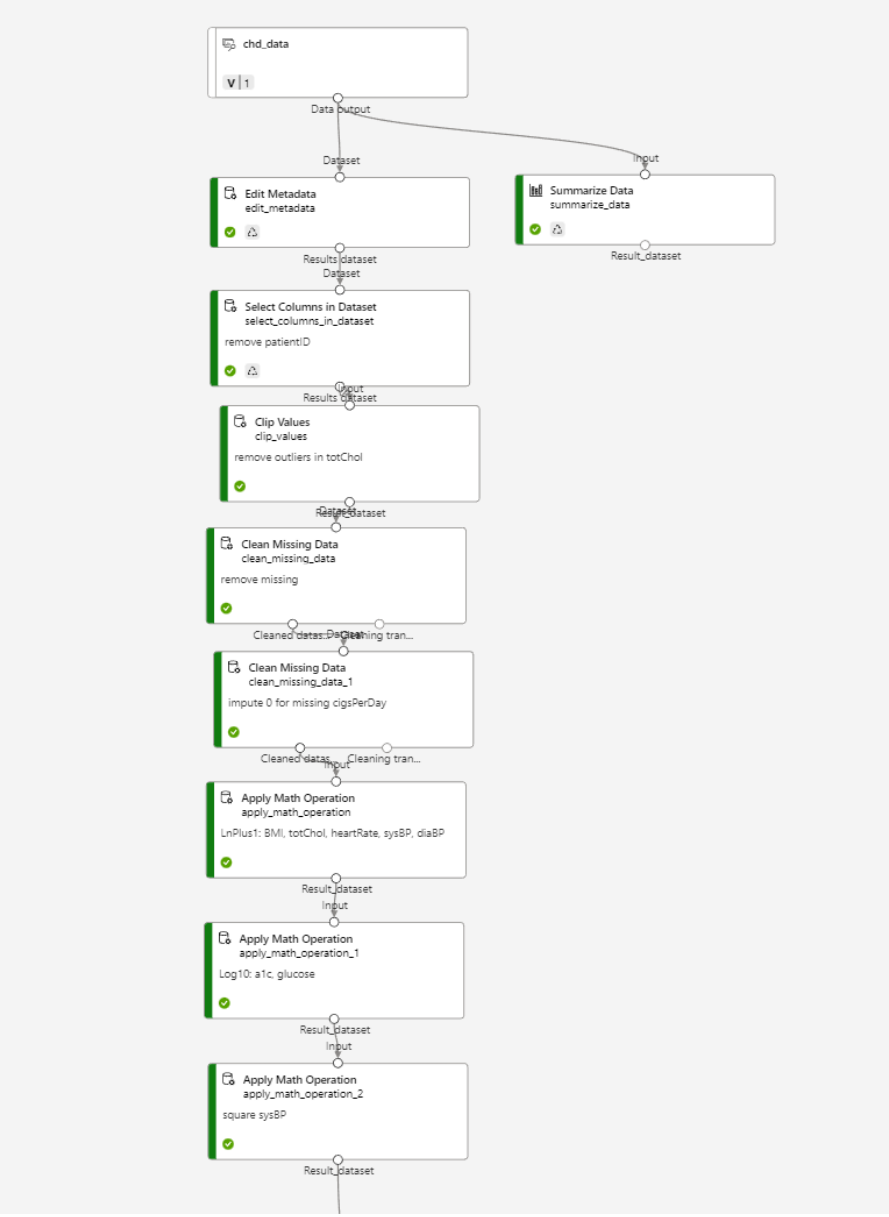
FEATURE SELECTION/ENGINEERING

- Features removed
  - patient ID – irrelevant for prediction
  - highly correlated variables: despite the high correlation between two pairs of predictors (glucose vs. a1c, dia bp vs. sys bp), I decide not to remove those two predictors because doing so will significantly decrease the model performance

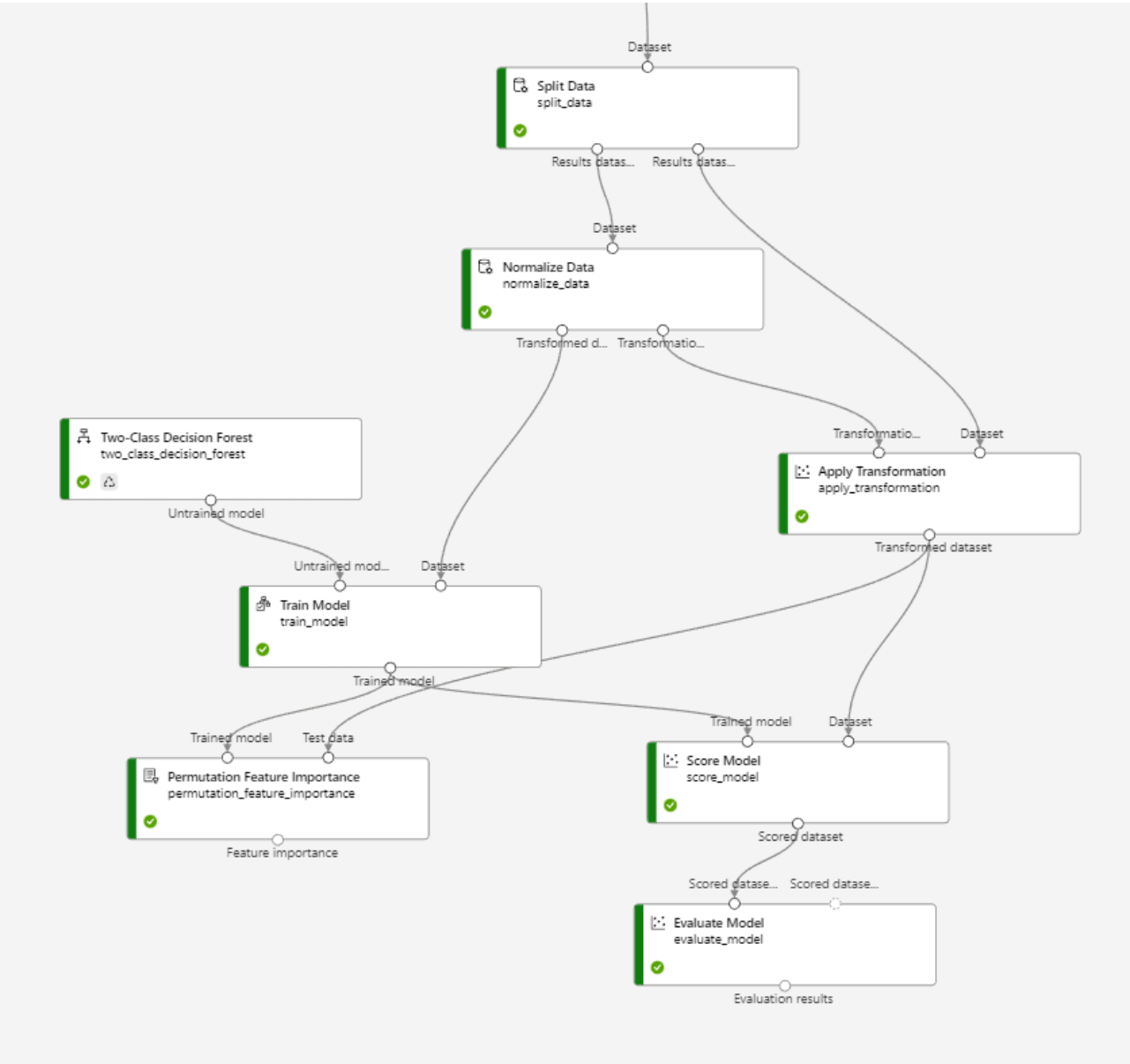
- Feature engineering
  - create square of sysBP as a new predictor
  - normalize income
  - log transformation:
    - very strong skewness (Log10): glucose, a1c
    - moderate skewness (LnPlus1): diaBP, sysBP, BMI, heartRate

USC Viterbi School of Engineering

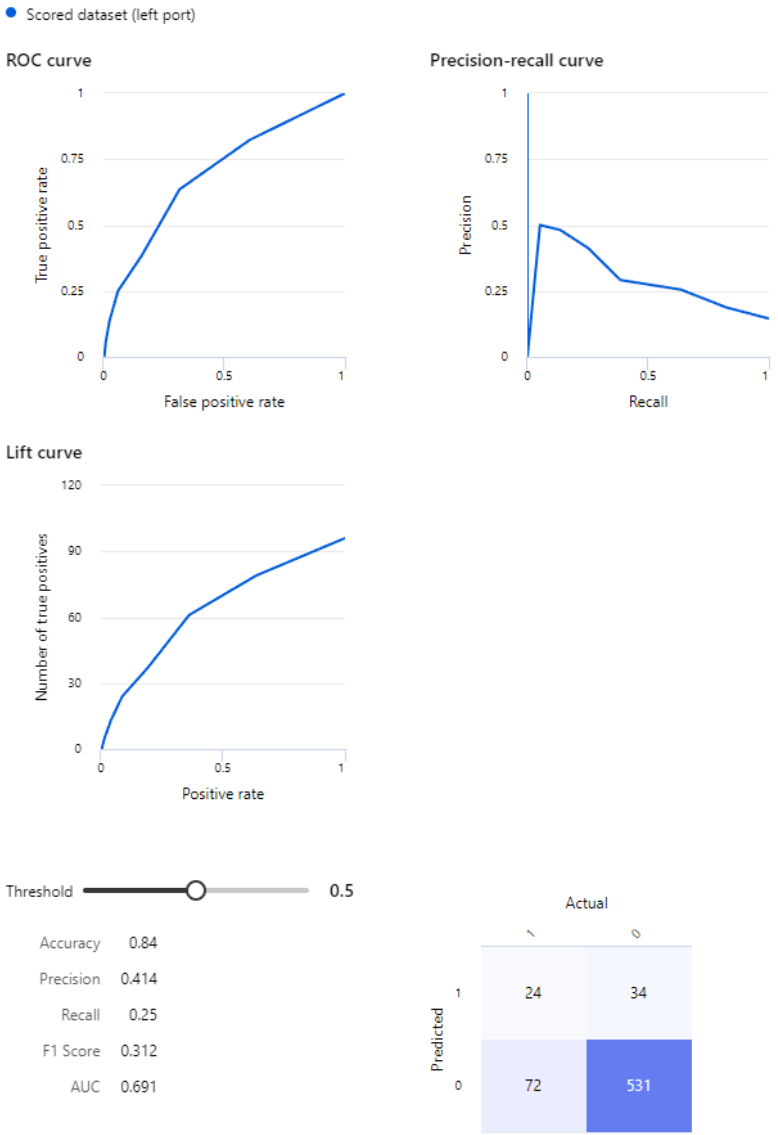
MODEL PIPELINE



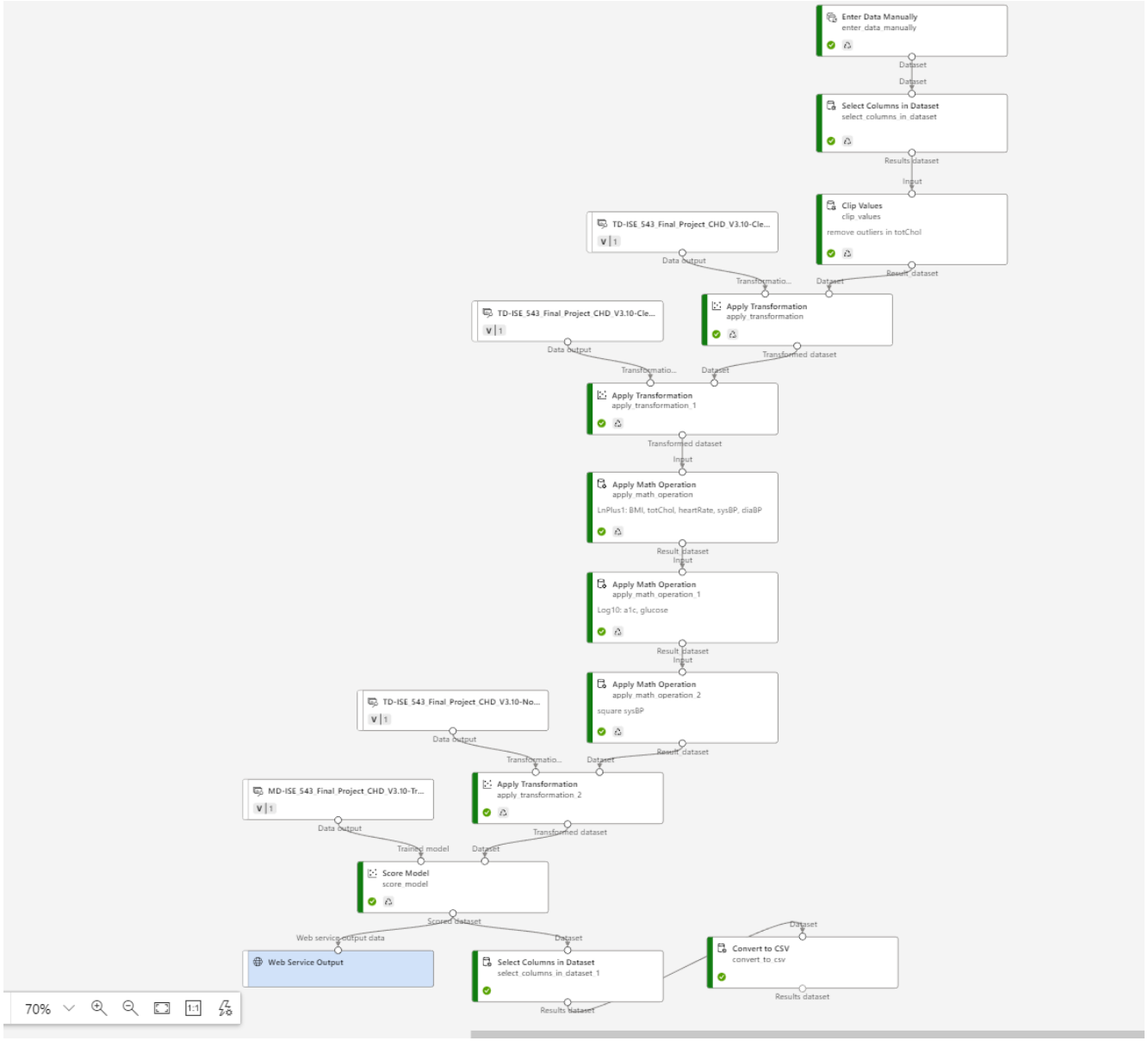
MODEL PIPELINE



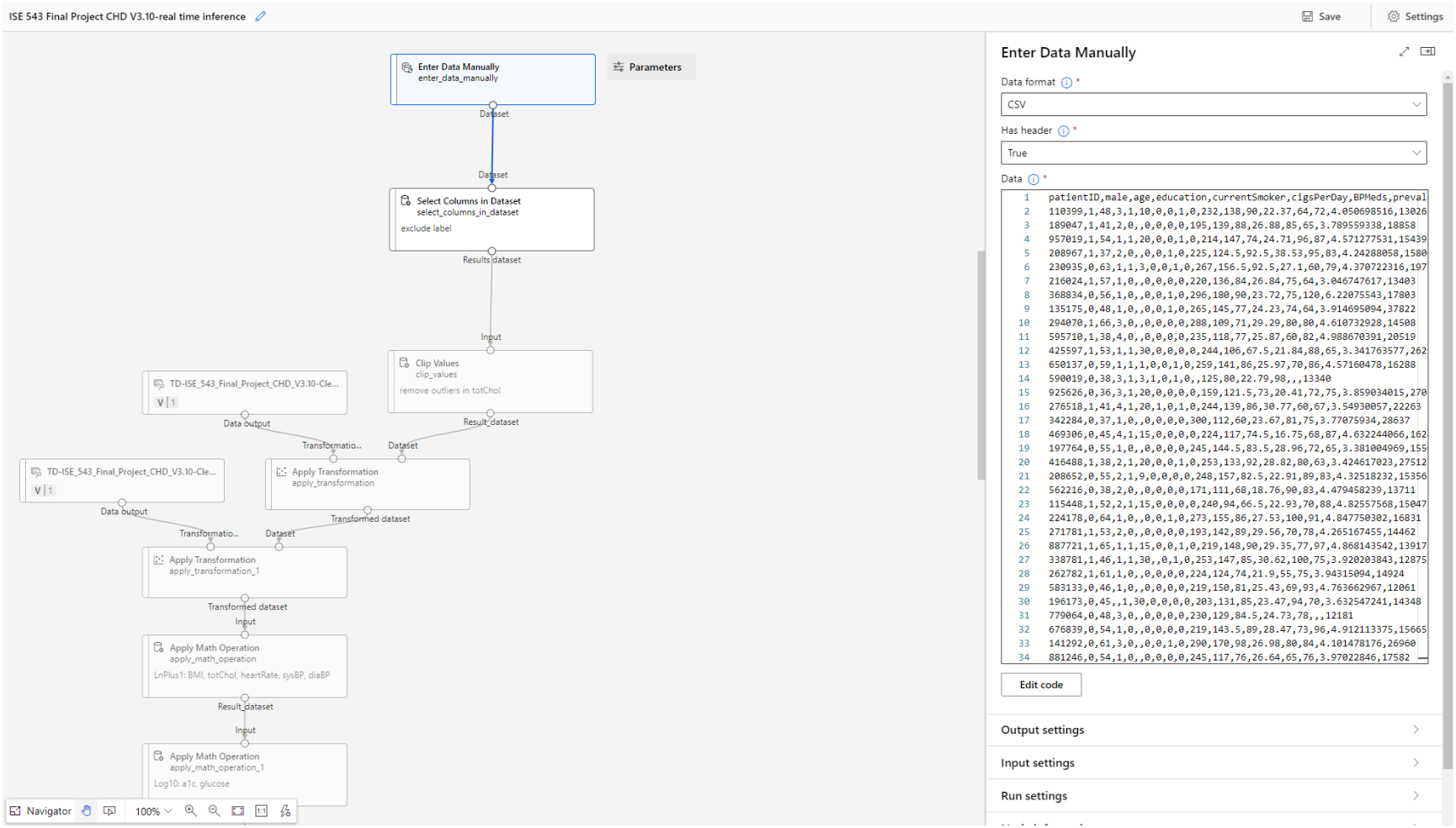
MODEL EVALUATION RESULTS



INFERENCE PIPELINE



MANUAL TEST DATA



USC Viterbi School of Engineering

TEST DATA SCORING