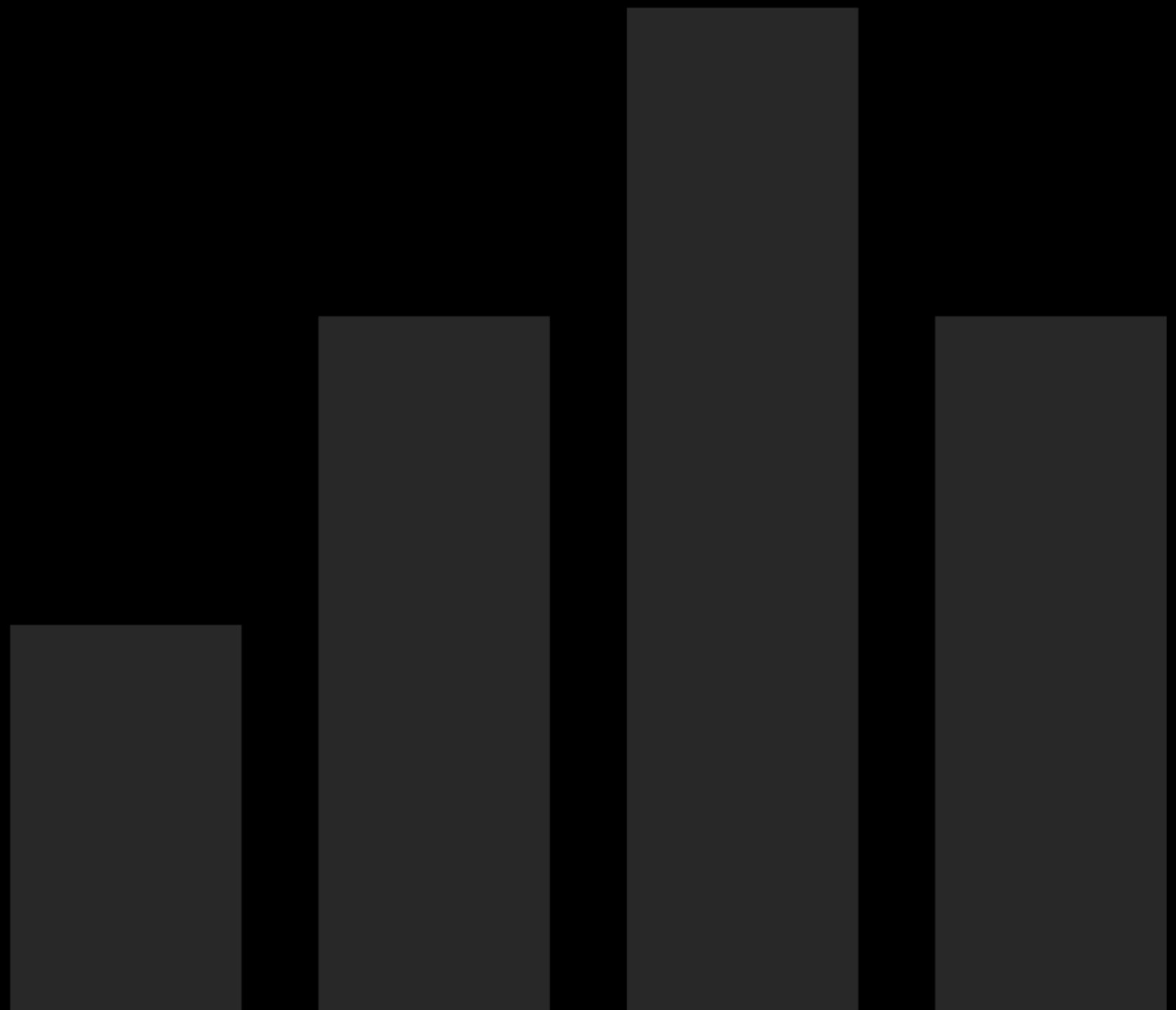
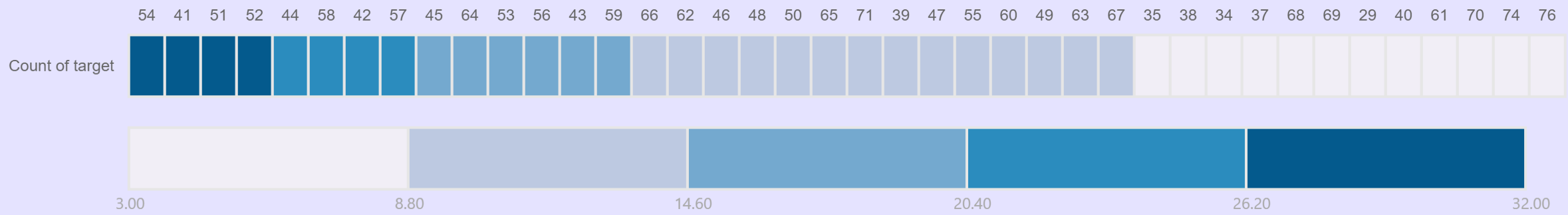
Heart Disease

View in Power BI ↗



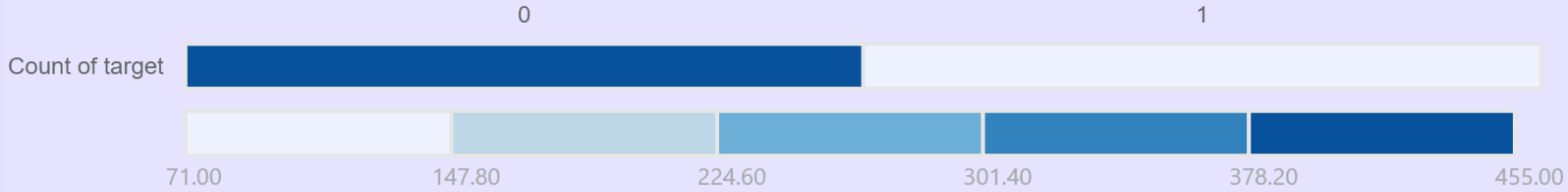
Correlation between heart disease and age



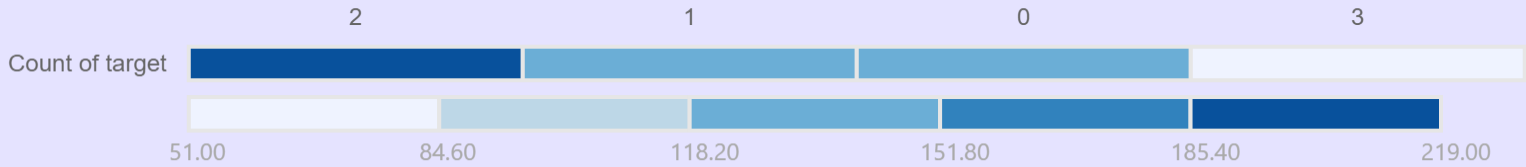
Number of people diagnosed with a heart disease

0 target	499 Count of Index
1 target	526 Count of Index

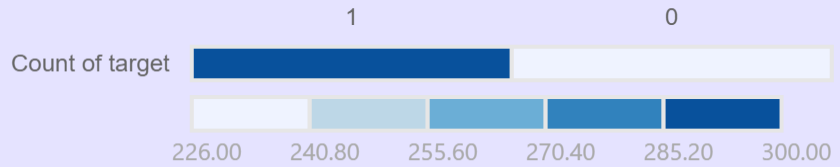
Correlation between heart disease and fasting blood sugar >120



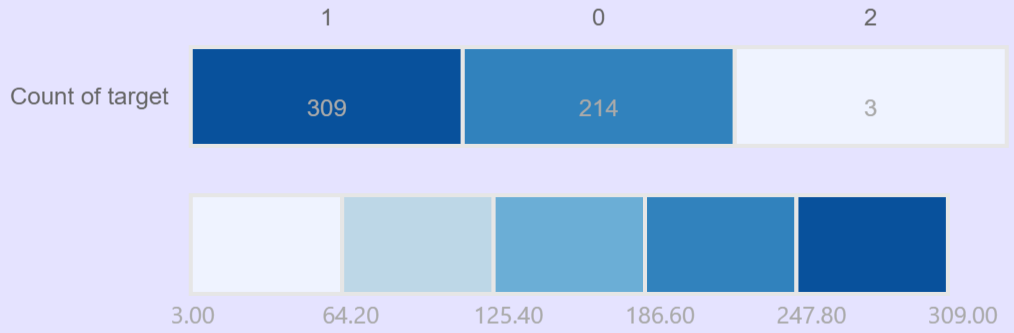
Correlation between heart disease and chest pain type



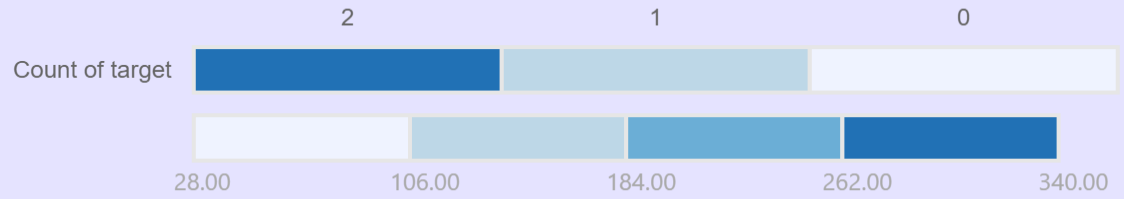
Correlation between heart disease and sex



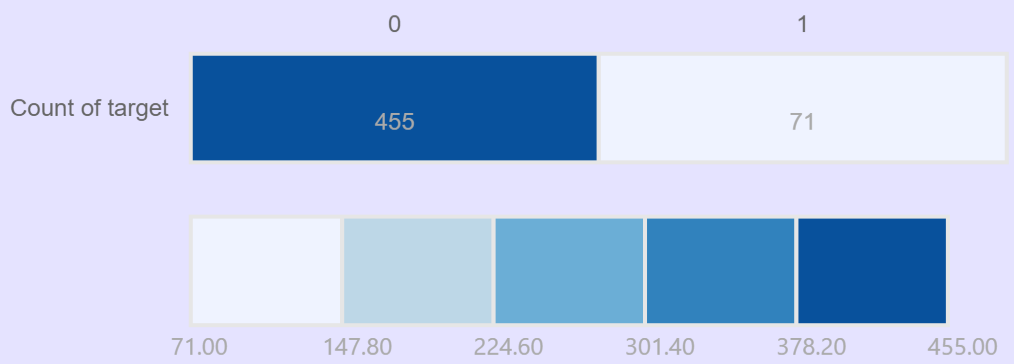
Correlation between heart disease and Resting electrocardiographic results



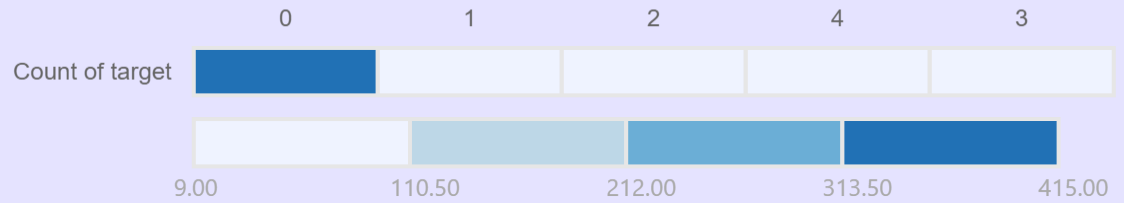
Correlation between heart disease and Slope of the peak exercise ST segment.



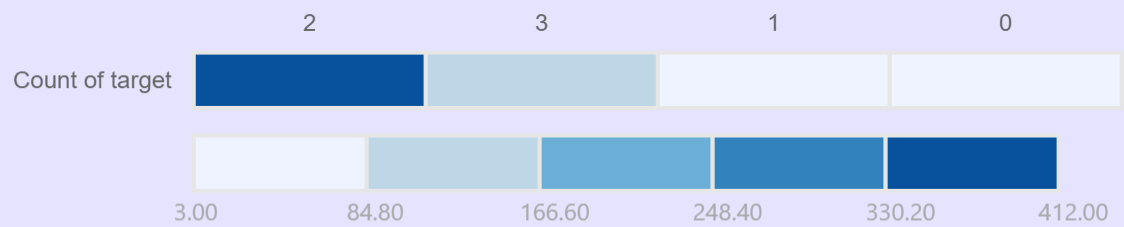
Correlation between heart disease and Exercise-induced angina

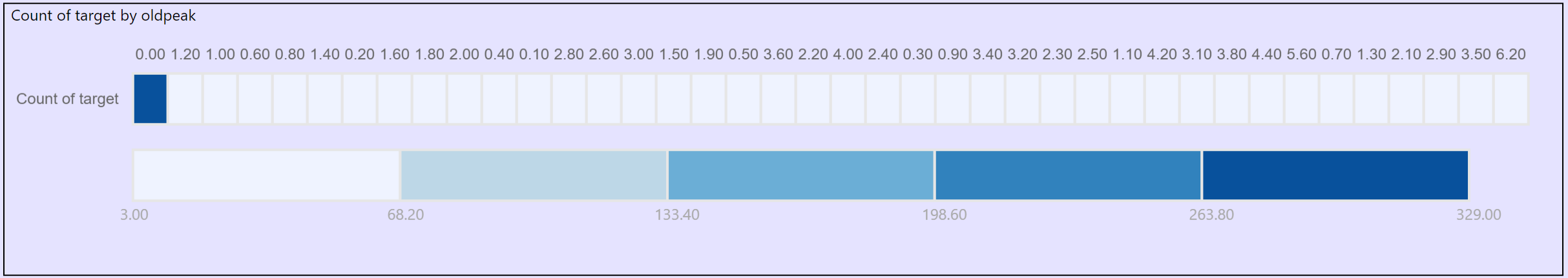


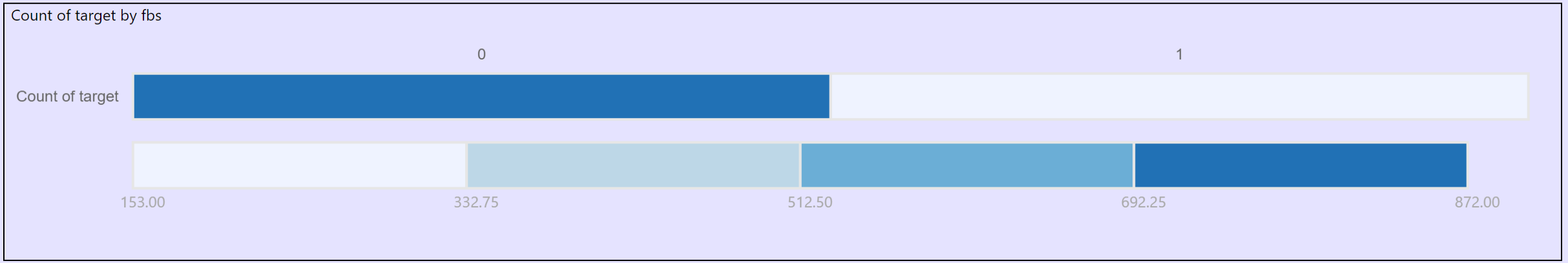
Correlation between heart disease and Number of major vessels colored by fluoroscopy

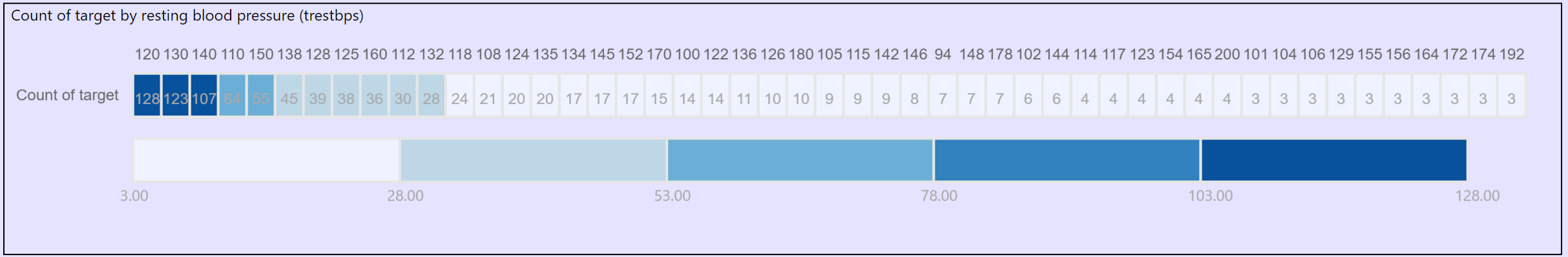


Correlation between heart disease and Thalassemia

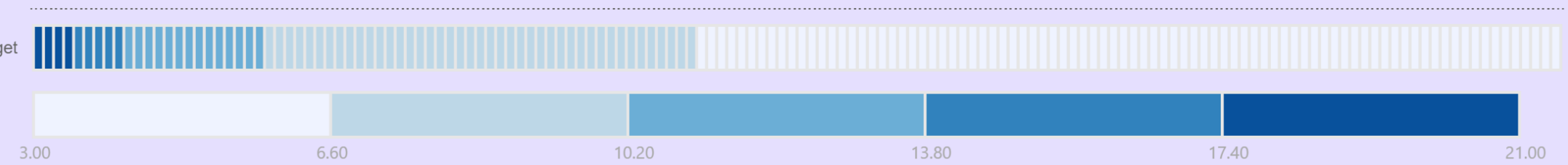



Count of target by oldpeak
0.00 1.20 1.00 0.60 0.80 1.40 0.20 1.60 1.80 2.00 0.40 0.10 2.80 2.60 3.00 1.50 1.90 0.50 3.60 2.20 4.00 2.40 0.30 0.90 3.40 3.20 2.30 2.50 1.10 4.20 3.10 3.80 4.40 5.60 0.70 1.30 2.10 2.90 3.50 6.20
Count of target
3.00
68.20
133.40
198.60
263.80
329.00

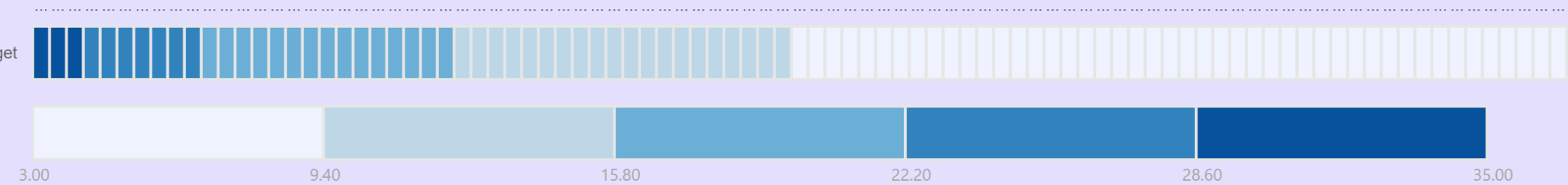

Count of target by fbs
0
1
Count of target
153.00
332.75
512.50
692.25
872.00


Count of target by resting blood pressure (trestbps)
120 130 140 110 150 138 128 125 160 112 132 118 108 124 135 134 145 152 170 100 122 136 126 180 105 115 142 146 94 148 178 102 144 114 117 123 154 165 200 101 104 106 129 155 156 164 172 174 192
Count of target
128 123 107 64 55 45 39 38 36 30 28 24 21 20 20 17 17 17 15 14 14 11 10 10 9 9 9 8 7 7 7 6 6 4 4 4 4 4 4 3 3 3 3 3 3 3 3 3 3
3.00
28.00
53.00
78.00
103.00
128.00

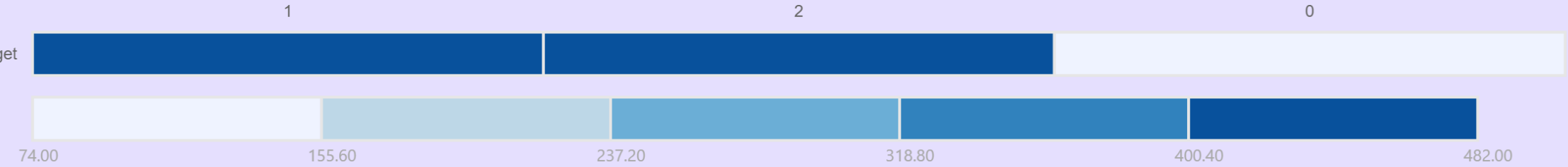
Count of target by chol



Count of target by max heart rate achieved (thalach)

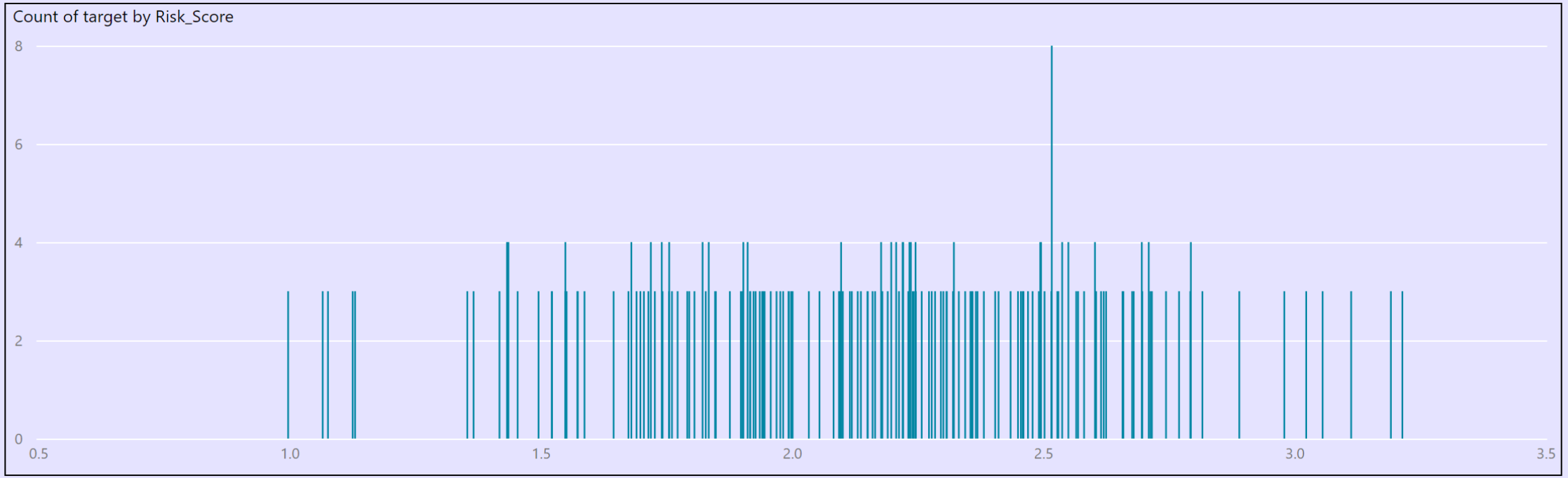


Count of target by slope



HeartDiseaseCount

526



INTRODUCTION

1.1 ABOUT THE DATA

THE DATASET HAS 14 COLUMNS AND 1026 ROWS CONTAINING DATA ABOUT PEOPLE, THEIR RESPECTIVE READINGS FOR BLOOD PRESSURE, CHOLESTEROL, HEART RATE, ETC. AND WHETHER THEY WERE DETECTED WITH HEART DISEASE. IT IS SORT OF A SAMPLE DATA THAT CAN BE USED TO DRAW INSIGHTS ABOUT UNDERLYING PATTERNS AND TO PREDICT FUTURE OUTCOMES.

1.2 OBJECTIVES

THE MAIN OBJECTIVES OF THIS PROJECT ARE TO:

VISUALISE THE DATA TO SEE MEANINGFUL CORRELATIONS USING POWER BI.

TO FIND THE CORRELATING FACTORS FOR HEART DISEASE BY USING HEATMAPS.

TO NORMALISE THE FACTORS AS THEY BELONG TO DIFFERENT SCALES.

CREATING A PREDICTIVE MODEL THAT PREDICTS A PERSON'S RISK SCORE BASED ON THE FACTORS.

1.3 DATA CLEANING

THERE ARE NO MISSING VALUES OR NULL VALUES IN THE DATA.

METHODOLOGY

FIRST, WE CREATE HEATMAPS BETWEEN VARIOUS FACTORS AND THE DETECTION OF HEART DISEASE TO SEE WHICH FACTORS ARE RESPONSIBLE FOR HEART DISEASE.

AGES BETWEEN 40 AND 60 ARE HIGHLY CORRELATED WITH HEART DISEASE. (CHART 1)

MALES ARE MORE PRONE TO HEART DISEASE. (CHART 5)

CHEST PAIN OF ALL TYPES HAS ABOUT THE SAME CORRELATION WITH HEART DISEASE. (CHART 4)

STEP 1: CREATING A RISK SCORE BASED ON MULTIPLE FACTORS IN POWER BI INVOLVES USING DAX (DATA ANALYSIS EXPRESSIONS) TO COMBINE RELEVANT VARIABLES INTO A SINGLE CALCULATED MEASURE OR COLUMN REPRESENTING THE OVERALL RISK.

- AGE
- CHOLESTEROL LEVELS
- RESTING BLOOD PRESSURE
- MAXIMUM HEART RATE ACHIEVED
- CHEST PAIN TYPE (CP)
- EXERCISE-INDUCED ANGINA (EXANG)
- NUMBER OF MAJOR VESSELS COLOURED BY FLUOROSCOPY (CA)
- THALASSEMIA TYPE (THAL)

STEP 2: NORMALIZE OR STANDARDIZE THE FACTORS

SINCE DIFFERENT FACTORS ARE ON DIFFERENT SCALES WE NORMALISE OR STANDARDISE THESE VALUES BEFORE COMBINING THEM. WE DO THIS TO CONVERT CONTINUOUS NUMERICAL VARIABLES TO ENSURE THEY ARE ON A COMPARABLE SCALE.

FOR INSTANCE, NORMALIZED_AGE = ([AGE] - MIN([AGE])) / (MAX([AGE]) - MIN([AGE])).

SIMILARLY, WE CAN NORMALISE OTHER NUMERIC FACTORS IN THE SAME WAY. OTHER NUMERIC FACTORS INCLUDE RESTING BLOOD PRESSURE (TRESTBPS), CHOLESTEROL, MAXIMUM HEART RATE ACHIEVED (THALACH), ST DEPRESSION INDUCED BY EXERCISE RELATIVE TO REST (OLDPEAK).

HOWEVER, FOR CATEGORICAL DATA, WE DO NOT NORMALISE AS THEY ARE ALREADY ON A STANDARD SCALE. IN THE DATA, WE HAVE TWO TYPES OF CATEGORICAL DATA:

BINARY CATEGORICAL VARIABLES: HAVING ONLY TWO CATEGORICAL VALUES, FOR EXAMPLE, SEX, FASTING BLOOD SUGAR (FBS), EXERCISE-INDUCED ANGINA (EXANG), AND DIAGNOSIS OF HEART DISEASE (TARGET). WE CAN USE THESE DIRECTLY FOR CALCULATING RISK SCORES.

MULTI-CATEGORY VARIABLES: HAVING MORE THAN TWO CATEGORICAL VALUES, FOR EXAMPLE, CHEST PAIN TYPE (CP), RESTING ELECTROCARDIOGRAPHIC RESULTS (RESTECG), SLOPE OF THE PEAK EXERCISE ST SEGMENT (SLOPE), NUMBER OF MAJOR VESSELS (0-3) COLOURED BY FLUOROSCOPY (CA), THALASSEMIA (THAL). WE CAN ASSIGN NUMERIC SCORES TO EACH CATEGORY BASED ON HOW MUCH IT CONTRIBUTES TO HEART DISEASE RISK.

BASED ON THE HEATMAPS, WE ANALYSE WHICH FACTORS HAVE A HIGH CORRELATION WITH HEART DISEASE.

AGE HAS A HIGH CORRELATION WITH HEART DISEASE, ESPECIALLY IN AGE GROUPS BETWEEN 40 AND 60. (CHART 1)

CHEST PAIN ALSO HAS A HIGH CORRELATION WITH HEART DISEASE. (CHART 4)

RESTING ELECTROCARDIOGRAPHIC RESULTS OF CATEGORIES 0 AND 1. (CHART 6)

PEAK EXERCISE ST SEGMENTS 1 AND 2. (CHART 8)

MEN ARE MORE PRONE TO HEART DISEASE. (CHART 5)

SLOPE OF THE PEAK EXERCISE ST SEGMENT (SLOPE) (CHART 16)

HOWEVER, THALASSEMIA (CHART 10) , RESTING BOOD PRESSURE (TRESTBPS) (CHART 13) , CHOLESTEROL (CHART 14) , MAXIMUM HEART RATE ACHIEVED (THALACH) (CHART 15) , ST DEPRESSION INDUCED BY EXERCISE RELATIVE TO REST (OLDPEAK) (CHART 11) AND NUMBER OF MAJOR BLOOD VESSELS COLOURED BY FLUOROSCOPY (CA) (CHART 9) ARE COMPARATIVELY NOT HIGHLY CORRELATED WITH HEART DISEASE. EXERCISE-INDUCED ANGINA (EXANG) OF ONLY CATEGORY 0 IS CORRELATED WITH HEART DISEASE.

ON THE BASIS OF THIS, WE CAN CURATE AN EXPRESSION TO CALCULATE RISK SCORE AS FOLLOWS:

RISK SCORE = 0.3*NORMALISED AGE + 0.3*SEX + 0.3*CHEST PAIN + 0.3* RESTECG+ 0.3*SLOPE + 0.2*NORMALISED_CHOLESTEROL + 0..2*FBS + 0.2*NORMALISED THALACH + 0.2*NORMALISED RESTING BLOOD PRESSURE (TRESTBPS) + 0.2*THALASSEMIA + 0.1*NORMALISED OLDPEAK

```
Risk_Score = 0.3*'Heart Disease data'[Normalised_Age]+0.3*'Heart Disease data'[sex]+0.3*'Heart Disease data'[cp]+0.3*'Heart Disease data'[restecg]+0.3*'Heart Disease data'[slope]+0.2*'Heart Disease data'[Normalised_Chol]+0.2*'Heart Disease data'[fbs]+0.2*'Heart Disease data'[Normalised_Thalach]+0.2*'Heart Disease data'[Normalised_trestbps]+0.2*'Heart Disease data'[thal]+0.1*'Heart Disease data'[oldpeak]
```

A NEW COLUMN IS CREATED FOR CALCULATING THE RISK OF EACH PERSON BASED ON THEIR CORRESPONDING VALUES FOR EACH FACTOR. THE SAME RISK SCORE FORMULA CAN ALSO BE USED TO ESTIMATE RISK SCORED IN A PREDICTIVE MODEL.

THE TOTAL NUMBER OF TARGET=1, I.E., DETECTION OF HEART DISEASE IS 526. (CHART 17) WHEN WE PLOT A BAR GRAPH (CHART 18) OF ALL RISK SCORES, WHICH HAVE TARGET=1, WE SEE THAT MOST OF THE BARS ARE CONCENTRATED BETWEEN 2 AND 3 WHICH COULD BE AN INDICATOR THAT PEOPLE WITH RISK SCORES ABOVE 2 HAVE A HIGHER CHANCE OF HAVING A HEART DISEASE.