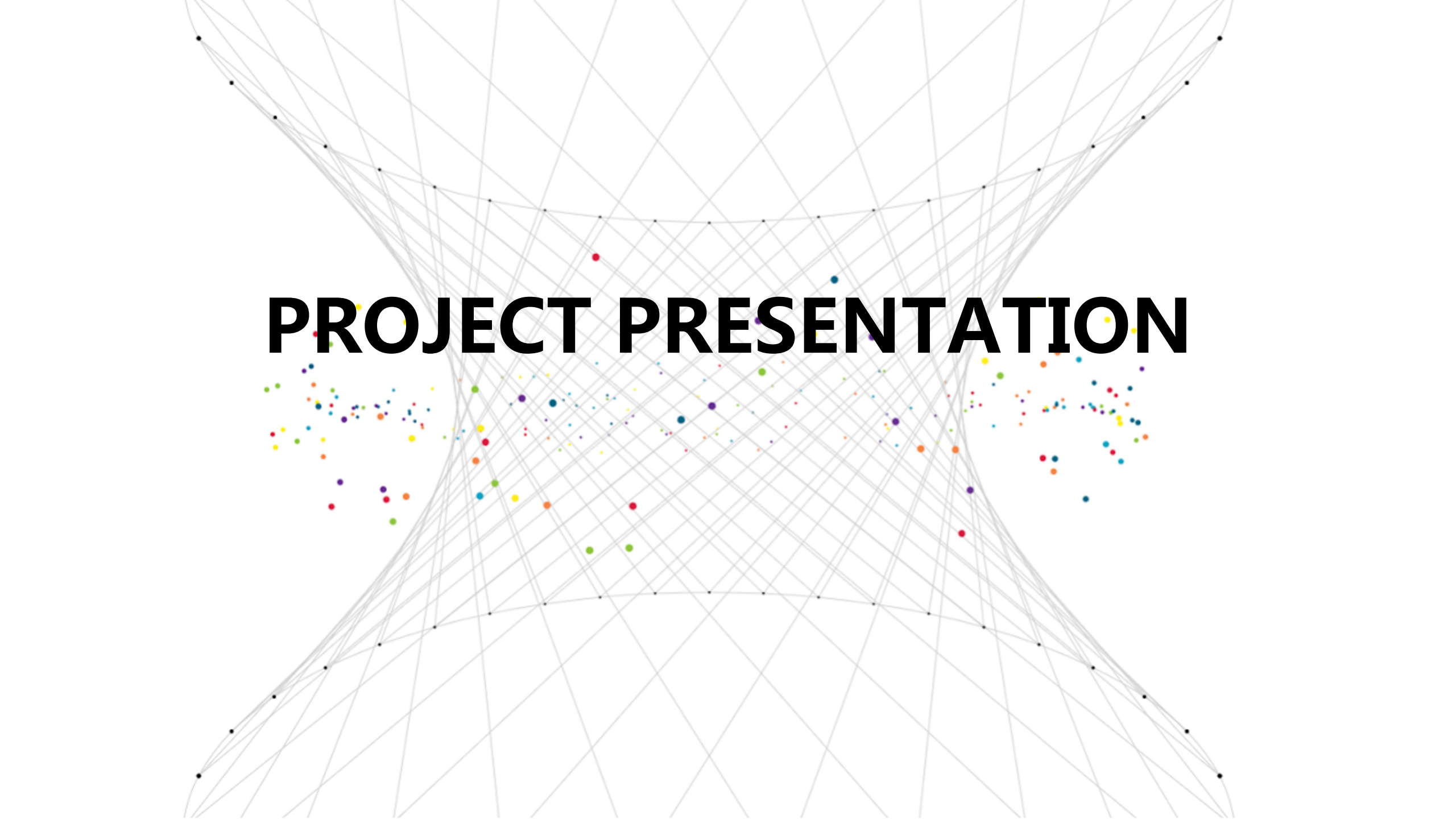
PROJECT PRESENTATION



CONTENTS

Introduction

PART ONE

Descriptive Statistics

PART TWO

Model Selection

PART THREE

Model Diagnostic

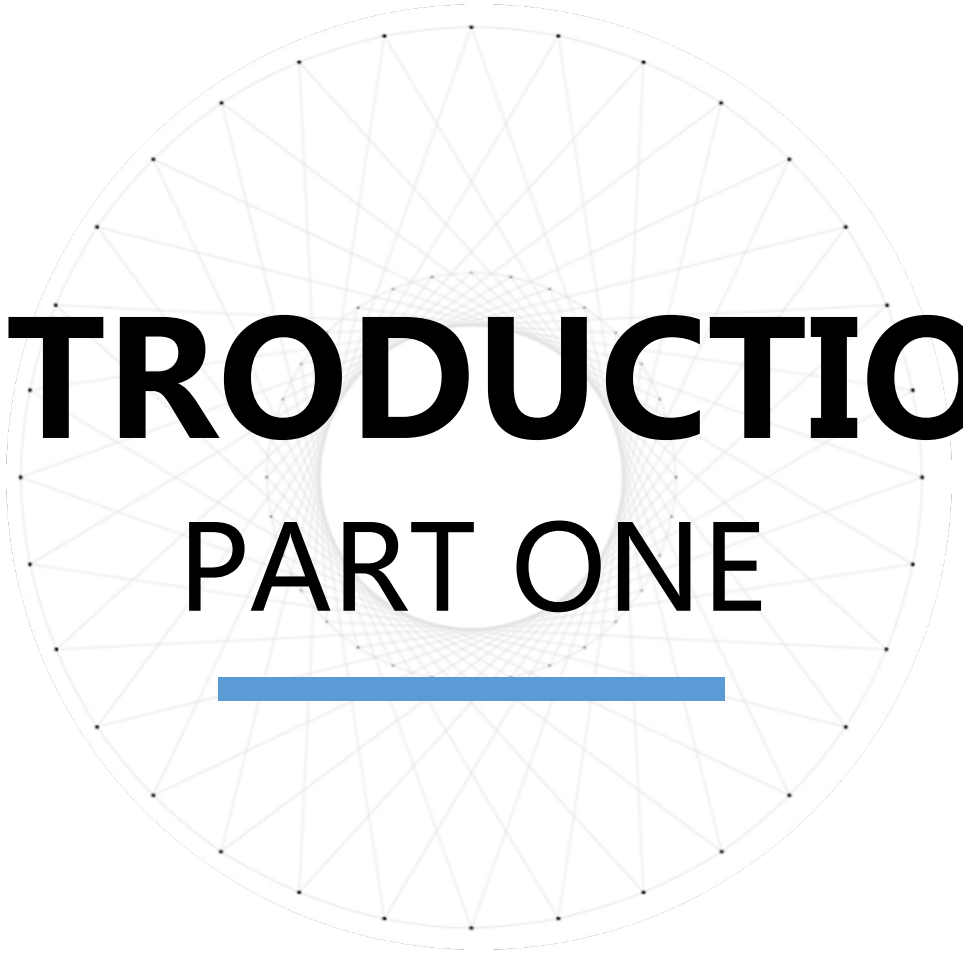
PART FOUR

Result

PART FIVE

Limitations & Future Developments

PART SIX



INTRODUCTION

PART ONE

Obseravational Data: UCI Heart Disease Cleveland

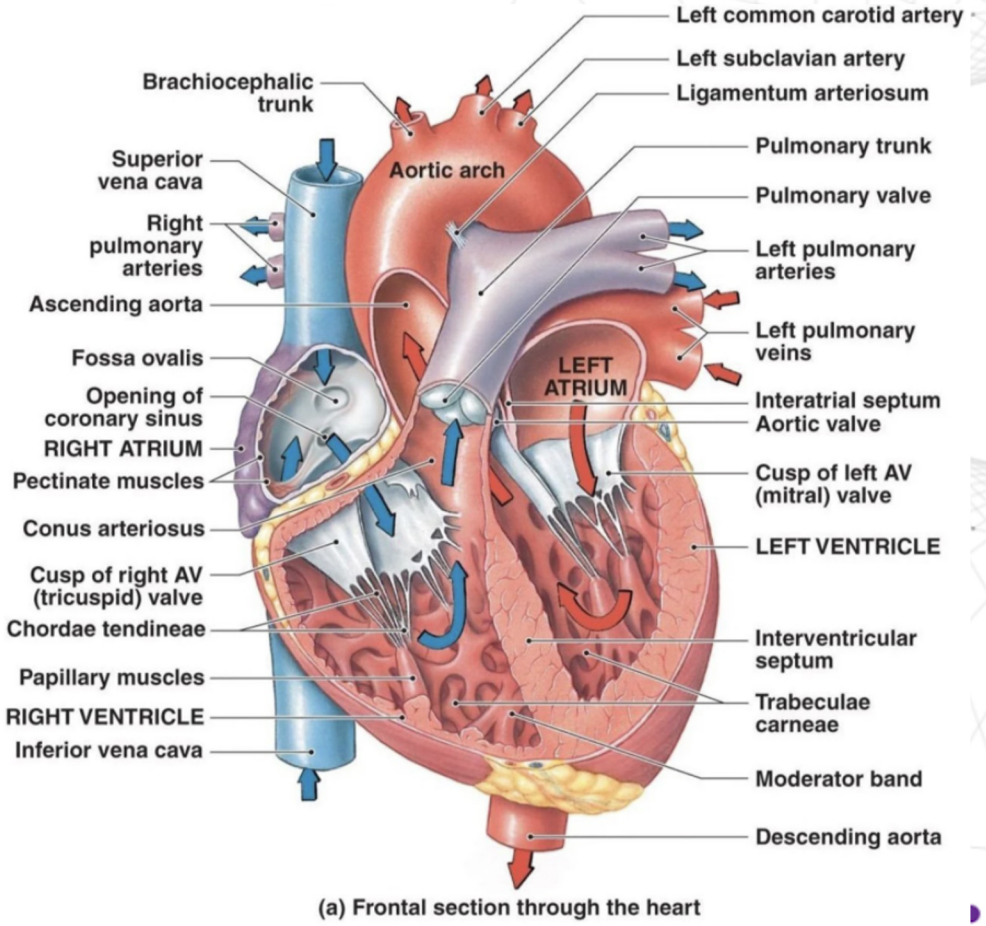
Data:

database contains 76 attributes

all published experiments prefer to study a subset of 14 of them

Goal:

Explore relationship between heart disease and some explanatory variables



(a) Frontal section through the heart

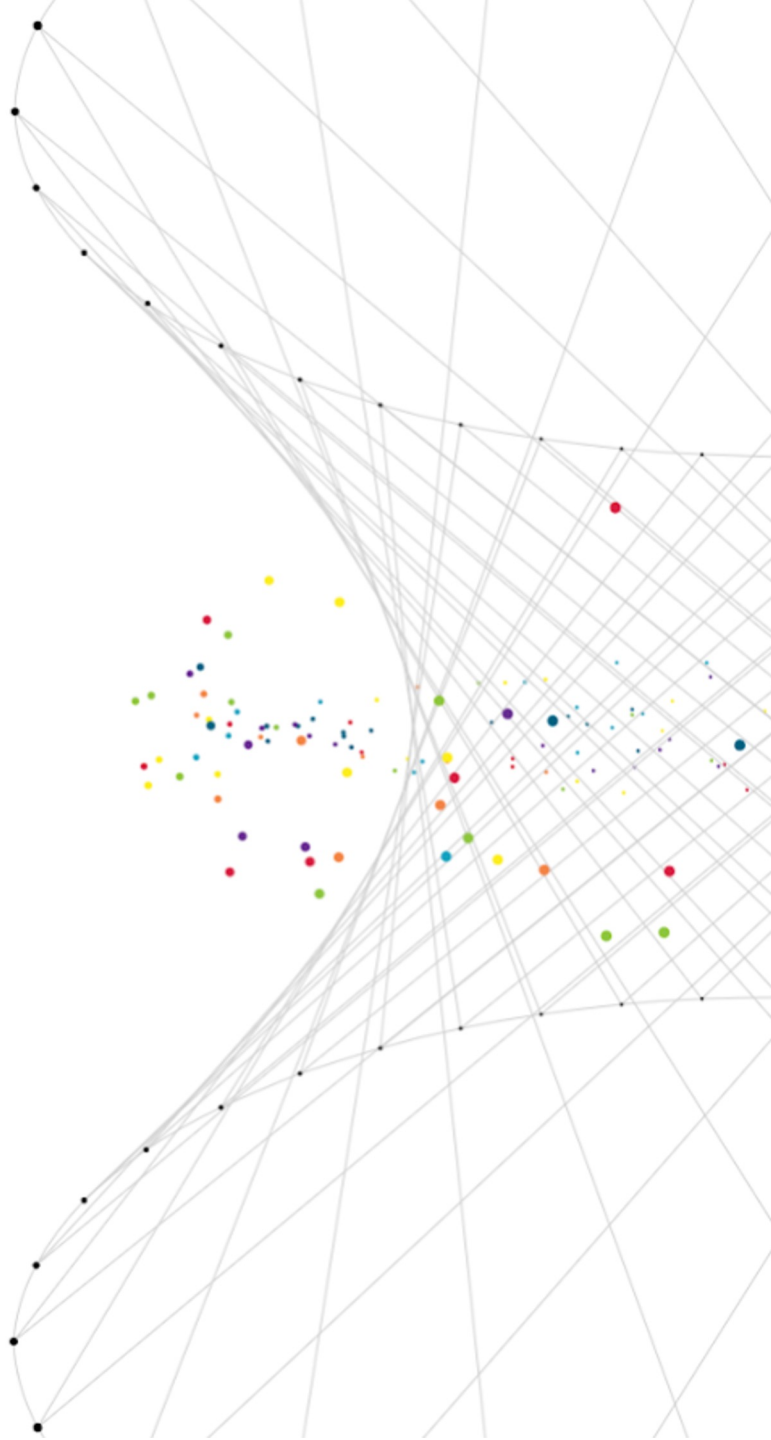
Dependent Variable:

Condition: 0 = no disease, 1 = disease

Predictor Variables:

Numerical (5):

- age: age in years
- trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- chol: serum cholestoral in mg/dl
- thalach: maximum heart rate achieved
- oldpeak: ST depression induced by exercise relative to rest



Predictor Variables:

Categorical (8):

- sex: sex (1 = male; 0 = female)
- cp: chest pain type
  -- Value 0: typical angina
  -- Value 1: atypical angina
  -- Value 2: non-anginal pain
  -- Value 3: asymptomatic
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg: resting electrocardiographic results
  -- Value 0: normal
  -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- exang: exercise induced angina (1 = yes; 0 = no)
- slope: the slope of the peak exercise ST segment
  -- Value 0: upsloping
  -- Value 1: flat
  -- Value 2: downsloping
- ca: number of major vessels (0-3) colored by fluoroscopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversable defect and the label

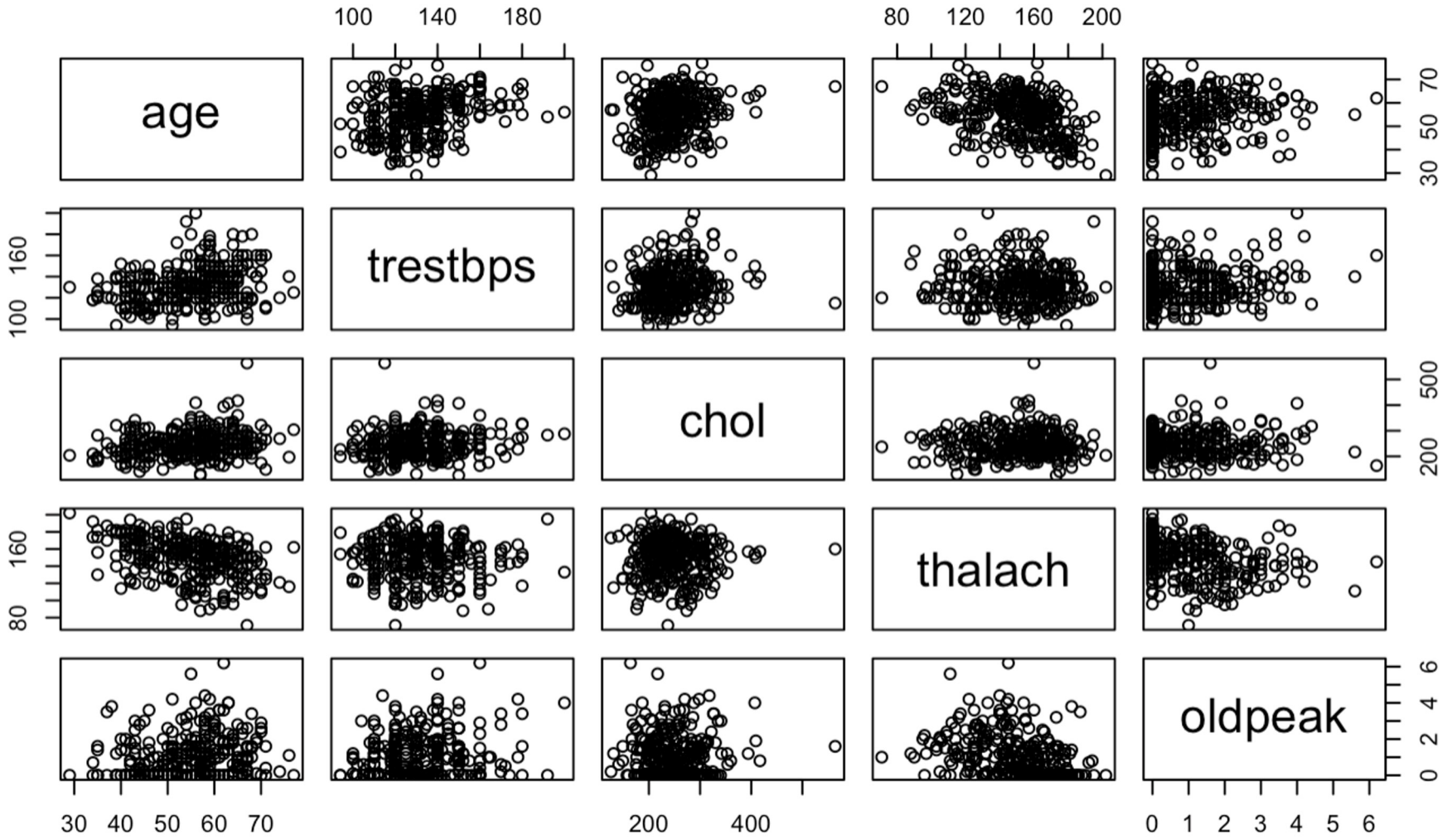
Descriptive Statistics

PART TWO

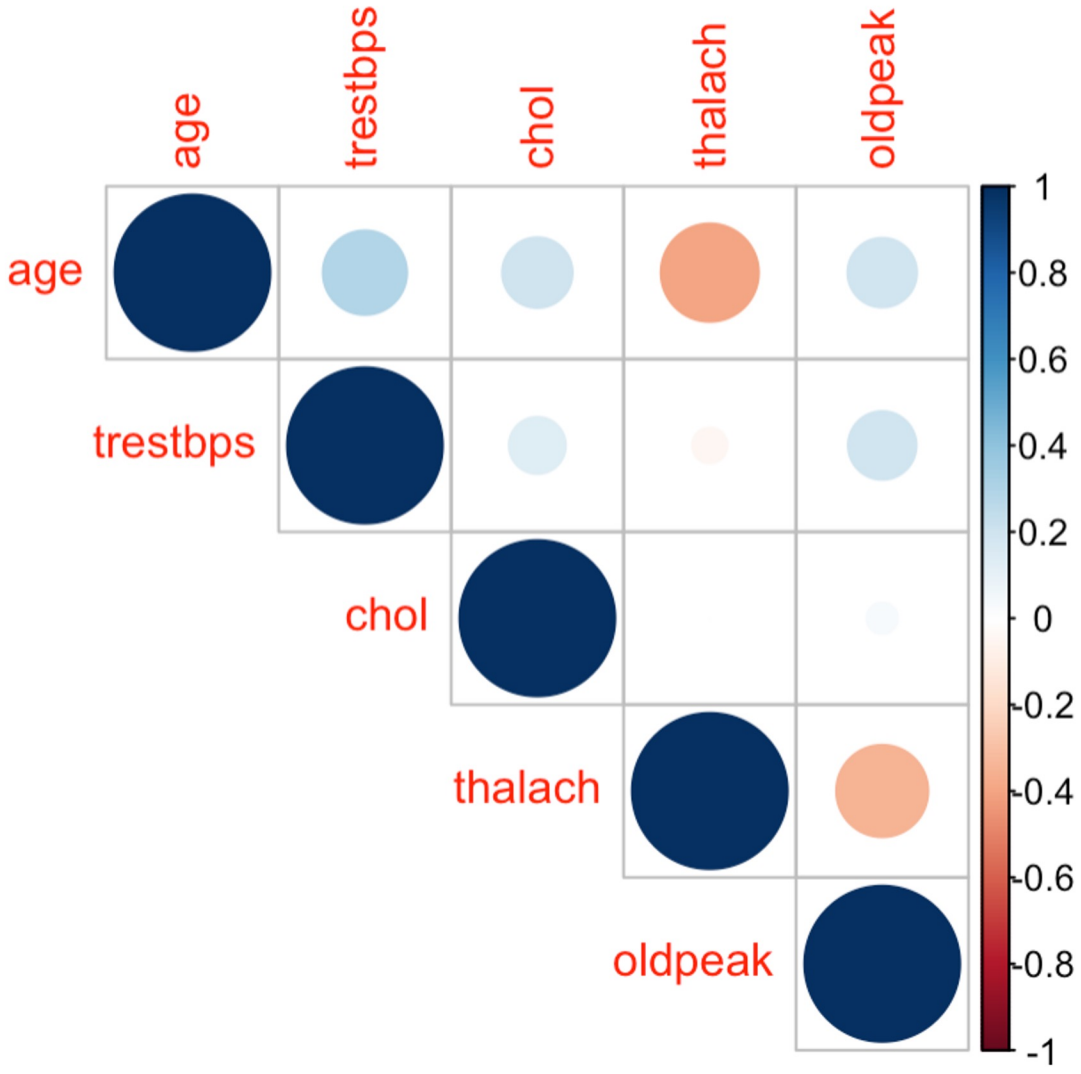
Summary Statistics

```
      age             sex      cp          trestbps          chol        fbs      restecg     thalach       exang
 Min.   :29.00   0: 96   0: 23   Min.   : 94.0   Min.   :126.0   0:254   0:147   Min.   : 71.0   0:200
 1st Qu.:48.00   1:201   1: 49   1st Qu.:120.0   1st Qu.:211.0   1: 43   1:  4   1st Qu.:133.0   1: 97
 Median :56.00           2: 83   Median :130.0   Median :243.0           2:146   Median :153.0
 Mean   :54.54           3:142   Mean   :131.7   Mean   :247.4                   Mean   :149.6
 3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:276.0                   3rd Qu.:166.0
 Max.   :77.00                   Max.   :200.0   Max.   :564.0                   Max.   :202.0
    oldpeak        slope    ca        thal     condition
 Min.   :0.000   0:139   0:174   0:164   0:160
 1st Qu.:0.000   1:137   1: 65   1: 18   1:137
 Median :0.800   2: 21   2: 38   2:115
 Mean   :1.056           3: 20
 3rd Qu.:1.600
 Max.   :6.200
```

Pair Plot (Numerical Variables Only):



Correlation Plot (Numerical Variables Only):



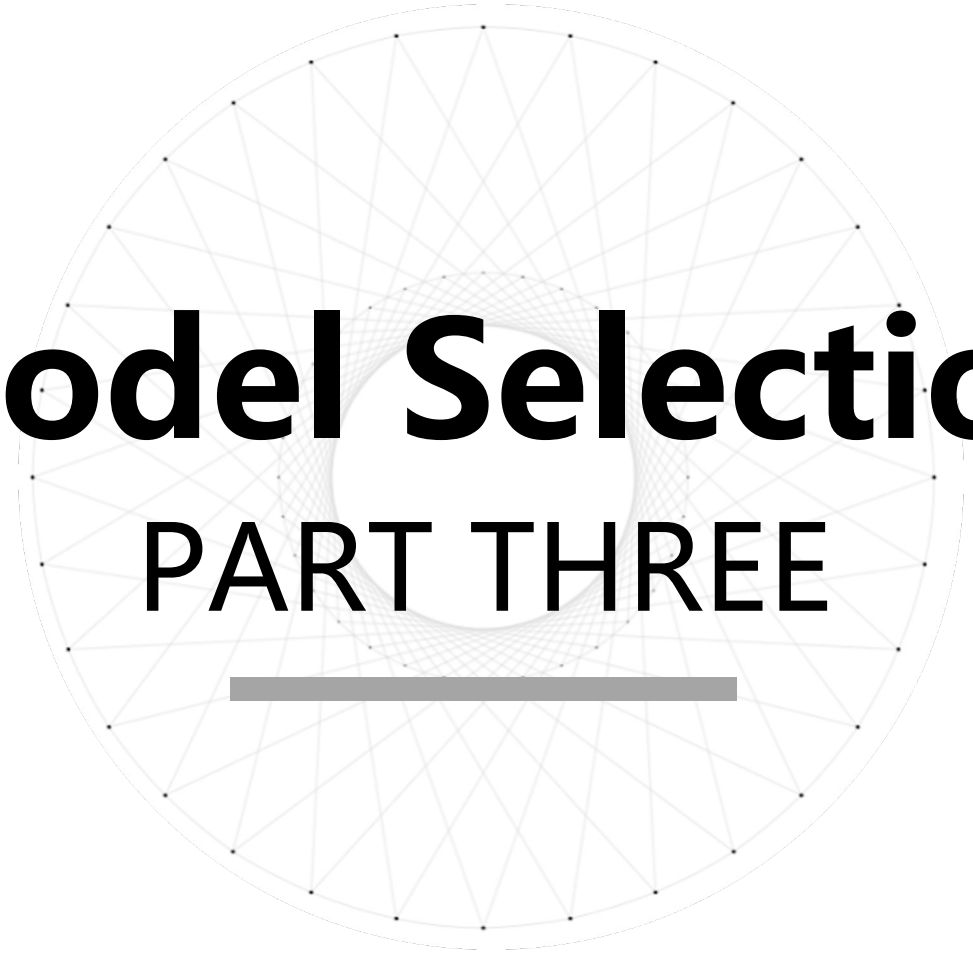
Variance Inflation Factor (VIF):

Quantifies the severity of multicollinearity in an ordinary least squares regression analysis

Provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity

VIF:

age	trestbps	chol	thalach	oldpeak
1.347358	1.137604	1.056714	1.330874	1.178751



Model Selection

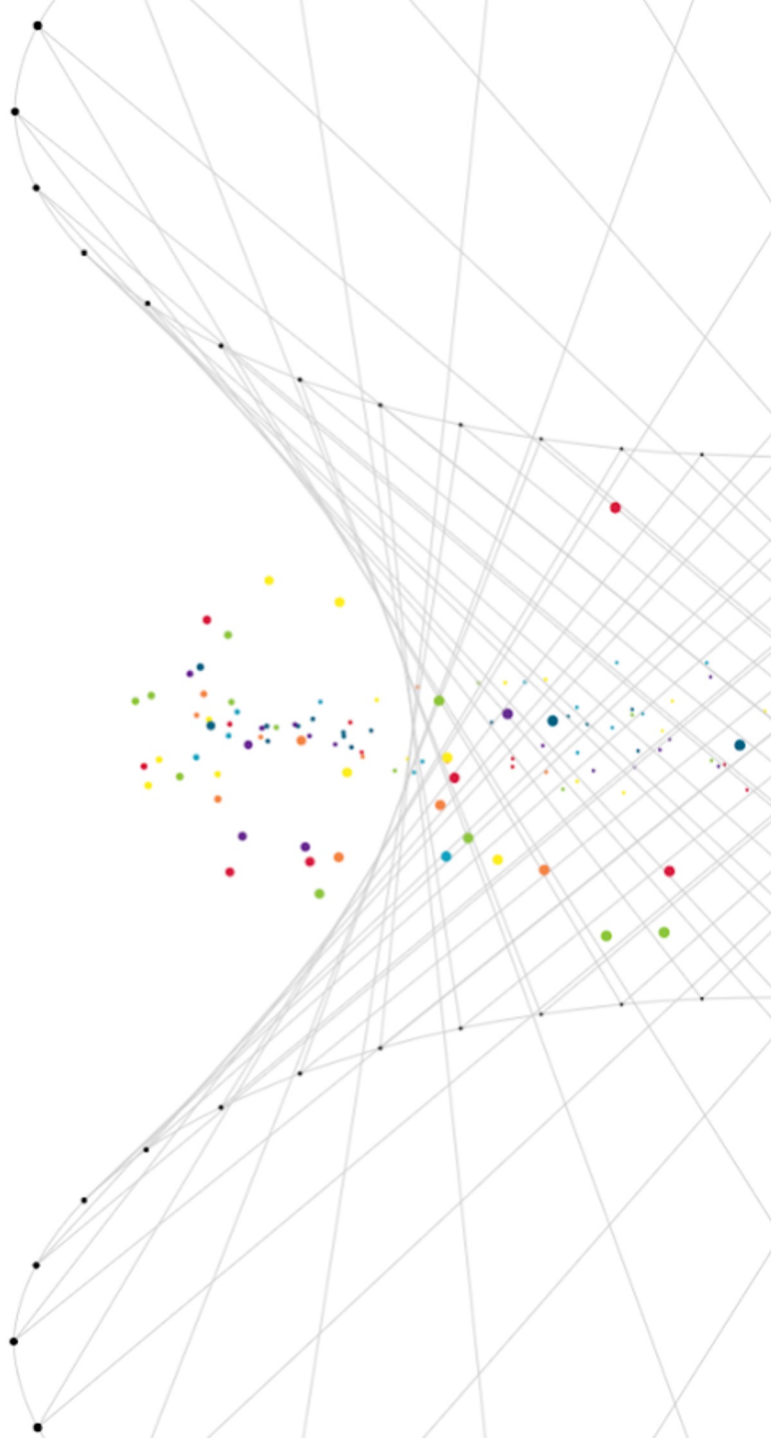
PART THREE

Logistic Regression:

Binary logistic model

Dependent variable: condition: 0/1

Predictor variables: numerical and categorical

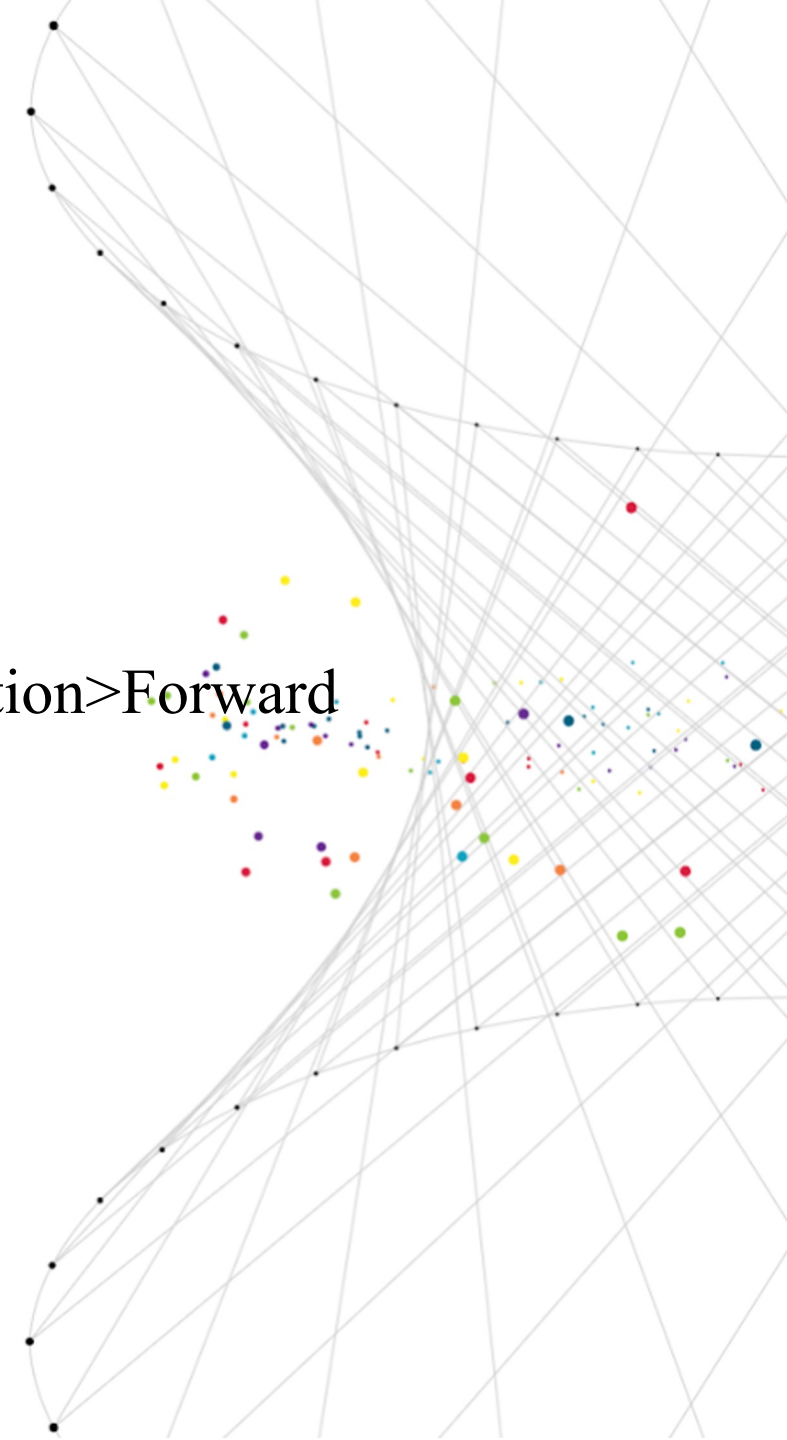


Automated Model Selection:

Forward selection: start from model with intercept only
Backward elemination: start from full model
Stepwise selection: start from main effect model

Runtime comparsion: Backward elemination>>Stepwise selection>Forward selection

AIC comparison:
Forward selection: AIC: 206.41, df: 36
Stepwise selection: AIC: 150.40, df: 61



Manual Model Selection:

Start from model selected by Stepwise:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.290e+03	8.090e+02	1.595	0.111
age	1.036e+00	7.504e-01	1.380	0.168
sex1	-8.980e+01	7.136e+01	-1.258	0.208
cp1	-1.612e+02	1.633e+02	-0.987	0.324
cp2	-7.771e+01	1.572e+02	-0.494	0.621
cp3	-1.652e+02	1.374e+02	-1.202	0.229
trestbps	-7.613e+00	5.197e+00	-1.465	0.143
chol	-3.882e-01	2.902e-01	-1.338	0.181
fbs1	-2.464e+02	4.748e+03	-0.052	0.959
restecg1	-2.915e+01	1.042e+02	-0.280	0.780
restecg2	4.474e+00	4.620e+00	0.968	0.333
thalach	-7.661e+00	4.904e+00	-1.562	0.118
exang1	-4.022e+01	3.222e+01	-1.248	0.212
oldpeak	-1.383e+02	1.055e+02	-1.310	0.190
slope1	-5.454e+02	4.036e+02	-1.351	0.177
slope2	1.382e+03	3.120e+05	0.004	0.996
ca1	-1.181e+02	1.024e+02	-1.154	0.249
ca2	8.367e+00	1.082e+02	0.077	0.938
ca3	1.512e+03	3.576e+05	0.004	0.997
thal1	-1.111e+02	1.462e+05	-0.001	0.999
thal2	1.634e+02	1.110e+02	1.473	0.141
age:thal1	1.868e+00	2.775e+03	0.001	0.999
age:thal2	-3.079e+00	2.084e+00	-1.477	0.140
exang1:slope1	6.878e+01	5.347e+01	1.286	0.198
exang1:slope2	-2.726e+02	7.876e+04	-0.003	0.997
trestbps:slope1	2.720e+00	2.045e+00	1.330	0.183
trestbps:slope2	-8.920e+00	2.863e+03	-0.003	0.998
oldpeak:slope1	1.057e+02	8.137e+01	1.298	0.194
oldpeak:slope2	2.766e+01	6.743e+04	0.000	1.000
thalach:ca1	1.730e+00	1.277e+00	1.354	0.176
thalach:ca2	-9.410e-01	1.103e+00	-0.853	0.394
thalach:ca3	-1.073e+01	3.744e+03	-0.003	0.998
oldpeak:thal1	-2.436e+01	1.544e+04	-0.002	0.999
oldpeak:thal2	9.500e+01	7.141e+01	1.330	0.183
exang1:ca1	1.024e+01	1.508e+03	0.007	0.995
exang1:ca2	1.710e+02	4.404e+04	0.004	0.997
exang1:ca3	3.659e+02	8.051e+04	0.005	0.996
cp1:thalach	5.499e-01	1.017e+00	0.541	0.589
cp2:thalach	1.073e-01	1.040e+00	0.103	0.918
cp3:thalach	7.704e-01	8.021e-01	0.961	0.337
fbs1:ca1	1.745e+02	1.739e+04	0.010	0.992
fbs1:ca2	2.305e+02	4.747e+03	0.049	0.961
fbs1:ca3	-7.612e+01	4.059e+05	0.000	1.000
fbs1:oldpeak	1.166e+02	1.928e+03	0.060	0.952
cp1:slope1	2.191e+02	1.612e+02	1.359	0.174
cp2:slope1	1.556e+02	1.077e+02	1.445	0.149
cp3:slope1	1.414e+02	9.974e+01	1.418	0.156
cp1:slope2	-3.197e+02	4.745e+07	0.000	1.000
cp2:slope2	-2.750e+02	1.476e+05	-0.002	0.999
cp3:slope2	2.815e+02	1.024e+05	0.003	0.998
sex1:oldpeak	4.689e+01	3.357e+01	1.397	0.163
chol:ca1	-6.574e-01	4.592e-01	-1.432	0.152
chol:ca2	5.103e-01	3.957e-01	1.290	0.197
chol:ca3	-2.465e-01	6.959e+02	0.000	1.000
sex1:chol	3.828e-01	3.034e-01	1.262	0.207
trestbps:thalach	4.240e-02	3.032e-02	1.398	0.162
sex1:ca1	3.604e+01	1.508e+03	0.024	0.981
sex1:ca2	9.526e+01	7.625e+01	1.249	0.212
sex1:ca3	-5.164e+01	9.378e+04	-0.001	1.000
trestbps:chol	3.724e-03	2.372e-03	1.570	0.116
exang1:oldpeak	1.322e+01	1.261e+01	1.048	0.294

Very large p-values after interaction terms added even after automated model selection

Suggest to stay with main effect model

Manual Model Selection:

Main effect model:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.253978   2.960399  -2.113 0.034640 *
age         -0.023508   0.025122  -0.936 0.349402
sex1         1.670152   0.552486   3.023 0.002503 **
cp1          1.448396   0.809136   1.790 0.073446 .
cp2          0.393353   0.700338   0.562 0.574347
cp3          2.373287   0.709094   3.347 0.000817 ***
trestbps     0.027720   0.011748   2.359 0.018300 *
chol         0.004445   0.004091   1.087 0.277253
fbs1        -0.574079   0.592539  -0.969 0.332622
restecg1     1.000887   2.638393   0.379 0.704424
restecg2     0.486408   0.396327   1.227 0.219713
thalach     -0.019695   0.011717  -1.681 0.092781 .
exang1       0.653306   0.447445   1.460 0.144267
oldpeak      0.390679   0.239173   1.633 0.102373
slope1       1.302289   0.486197   2.679 0.007395 **
slope2       0.606760   0.939324   0.646 0.518309
ca1          2.237444   0.514770   4.346 1.38e-05 ***
ca2          3.271852   0.785123   4.167 3.08e-05 ***
ca3          2.188715   0.928644   2.357 0.018428 *
thal1       -0.168439   0.810310  -0.208 0.835331
thal2        1.433319   0.440567   3.253 0.001141 **
```

Find largest p-value:
restecg: resting electrocardiographic results

Remove the corresponding explanatory variable from the model

A reduced model has been built

Manual Model Selection:

Conduct a Deviance test:

Null Hypothesis: The reduced model provides a better fit
Alternative Hypothesis: The main effect model provides a better fit

p-value: 0.45>0.05

Fail to reject null hypothesis, reduced model provides a better fit

Repeat the above steps until the final model has been achieved

Manual Model Selection:

Final Model:

```
glm(formula = condition ~ sex + cp + trestbps + oldpeak + slope +
    ca + thal, family = binomial(link = "logit"), data = heart_disease)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8068  -0.5022  -0.1270   0.3829   2.9998

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.86579    1.89438  -4.680 2.87e-06 ***
sex1         1.39946    0.49810   2.810 0.004960 **
cp1          1.45775    0.79649   1.830 0.067219 .
cp2          0.31192    0.69677   0.448 0.654392
cp3          2.75456    0.68831   4.002 6.28e-05 ***
trestbps     0.02469    0.01071   2.305 0.021182 *
oldpeak      0.49800    0.22399   2.223 0.026194 *
slope1       1.53979    0.45813   3.361 0.000776 ***
slope2       0.65270    0.86245   0.757 0.449172
ca1          2.30806    0.48468   4.762 1.92e-06 ***
ca2          2.83132    0.71991   3.933 8.39e-05 ***
ca3          2.19538    0.90437   2.428 0.015203 *
thal1       -0.08340    0.74388  -0.112 0.910736
thal2        1.52373    0.42081   3.621 0.000294 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Explanatory variables included in the final model:

Numerical: (2)

trestbps: resting blood pressure

oldpeak: ST depression induced by exercise relative to rest

Categorical: (5)

sex: sex

cp: chest pain type

slope: the slope of the peak exercise ST segment

ca: number of major vessels (0-3) colored by fluorosopy

thal: level of defect

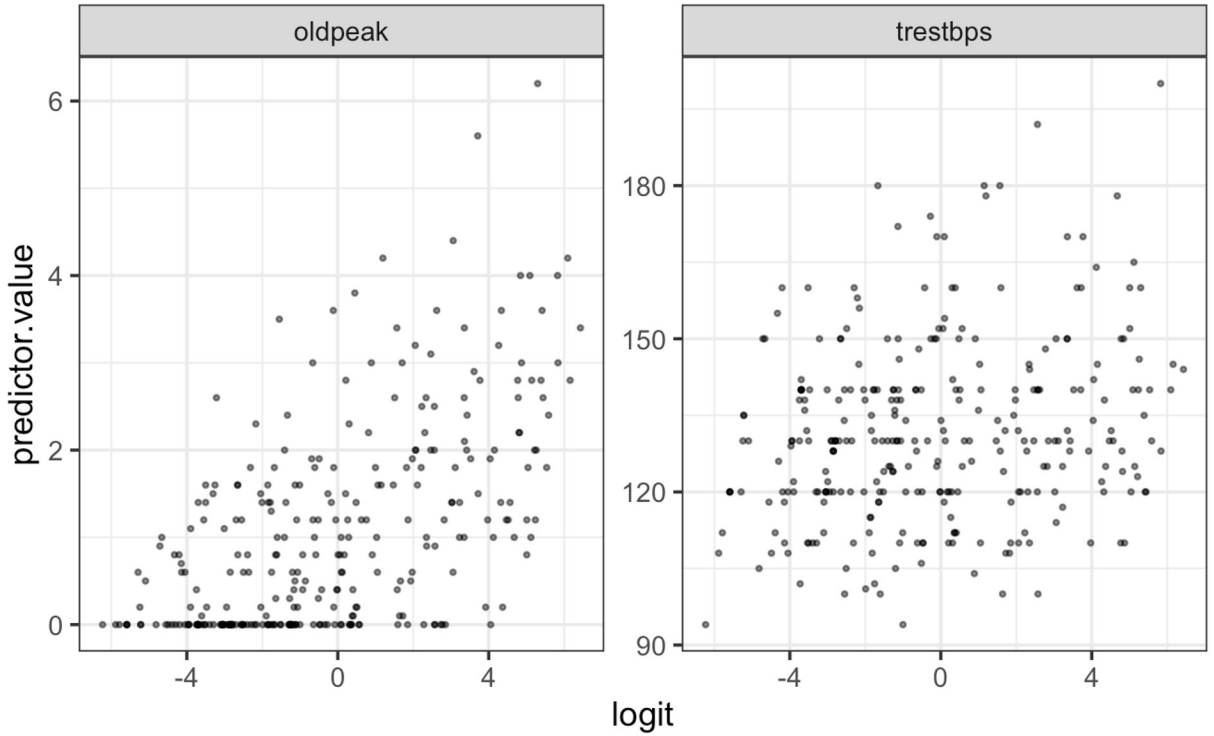
Model Diagnostic

PART FOUR

Linearity Assumption:

A linear relationship between continuous independent variables and the logit of the outcome variable.

The logits must be linearly related to the continuous numeric independent variables.



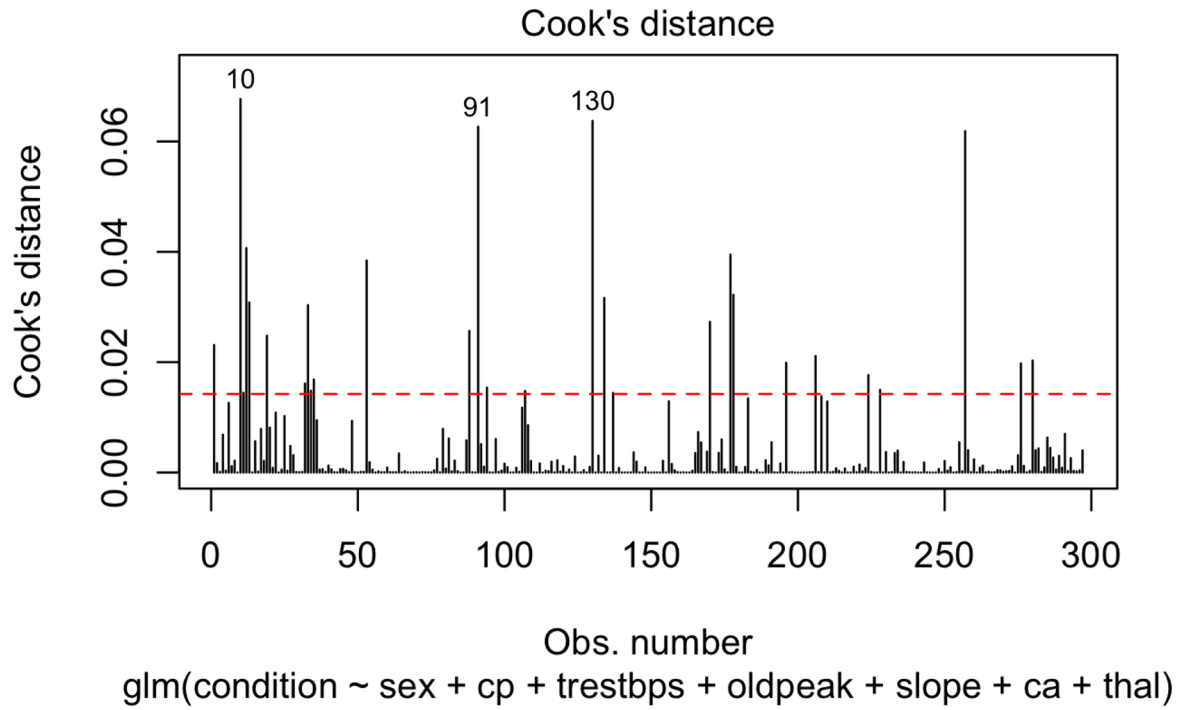
Continuous numeric independent variable : **oldpeak** and **trestbps**

y-axis: predictor value
x-axis: logit of the outcome.

Both variables are **quite linearly** associated with the heart disease outcome in logit scale since the points are randomly distributed.

Influential values:

We do not want our sample to include observations that might be outliers which influence our model's results.




```{r, label = "outliers and influence points", warning=FALSE}
which(influence.measures(Mfinal)$is.inf[,'cook.d'] )
```

```
named integer(0)
```

There is no influential observations in our data.

Multicollinearity:

Smaller than 2
No Multicollinearity problem!!

	GVIF	Df	GVIF^(1/(2*Df))
sex	1.439324	1	1.199719
cp	1.605205	3	1.082069
trestbps	1.128841	1	1.062469
oldpeak	1.519753	1	1.232783
slope	1.704172	2	1.142558
ca	1.436789	3	1.062263
thal	1.354106	2	1.078731

Goodness of fit:

The null hypothesis, H0, is that the model fits.
The alternative hypothesis, H1, is that the model does not fit.

```
df = 283
deviance = 192.57
p_val = pchisq(deviance, df=df, lower.tail=FALSE)
p_val
```

```
[1] 0.9999904
```

The p-value is larger than significant level 0.05

There is no evidence to reject the null hypothesis that the model fits

Result

PART FIVE

MODEL:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = X_i\beta$$

$$Y_i \sim Binomial(N_i, \mu_i)$$

For the *ith* observation, where μi is the proportion of having heart disease

The model we use here is logistic regression, where our response (proportion of having heart disease) is linked to a linear combination of covariates with a logit link.

There are seven covariates:

- Gender (sex)
- Chest pain type (cp)
- Resting blood pressure (trestbps)
- ST depression induced by exercise relative to rest (oldpeak)
- The slope of the peak exercise ST segment (slope)
- Number of major vessels (0-3) colored by flourosopy (ca)
- Whether have detective (thal)

Preliminary insights:

How odds ratio increase or decrease between groups

Table 1: Output of Pmisc:: coefTable(heart_disease_table) - the odds and CI for baseline for all variables

	est	2.5	97.5
(Intercept)	0.000	0.000	0.006
sex1	4.053	1.527	10.759
cp1	4.296	0.902	20.468
cp2	1.366	0.349	5.352
cp3	15.714	4.078	60.559
trestbps	1.025	1.004	1.047
oldpeak	1.645	1.061	2.552
slope1	4.664	1.900	11.447
slope2	1.921	0.354	10.413
ca1	10.055	3.889	25.998
ca2	16.968	4.138	69.569
ca3	8.983	1.526	52.874
thal1	0.920	0.214	3.953
thal2	4.589	2.012	10.470

95% confidence interval of each variable

Preliminary insights:

Table 1: Output of Pmisc:: coefTable(heart_disease_table) - the odds and CI for baseline for all variables

	est	2.5	97.5
(Intercept)	0.000	0.000	0.006
sex1	4.053	1.527	10.759
cp1	4.296	0.902	20.468
cp2	1.366	0.349	5.352
cp3	15.714	4.078	60.559
trestbps	1.025	1.004	1.047
oldpeak	1.645	1.061	2.552
slope1	4.664	1.900	11.447
slope2	1.921	0.354	10.413
ca1	10.055	3.889	25.998
ca2	16.968	4.138	69.569
ca3	8.983	1.526	52.874
thal1	0.920	0.214	3.953
thal2	4.589	2.012	10.470

The odds for male patients to have heart disease is 4.05 times of females.

The odds of the patient with asymptomatic chest pain suffering from heart disease is 15.71 times that of the patient who has typical angina

The odds of a patient with a flat slope of the peak exercise ST-segment having heart disease is 4.664 times that of the patient who is upsloping.

The patient with one major vessel coloured by fluoroscopy has the lowest odds to suffer from heart disease, while the patient with two major vessels coloured has the highest odds.

The odds of patient with reversable defect to suffer in heart disease is 4.59 times of the patient who have normal defect.

Preliminary insights:

Table 1: Output of Pmisc:: coefTable(heart_disease_table) - the odds and CI for baseline for all variables

	est	2.5	97.5
(Intercept)	0.000	0.000	0.006
sex1	4.053	1.527	10.759
cp1	4.296	0.902	20.468
cp2	1.366	0.349	5.352
cp3	15.714	4.078	60.559
trestbps	1.025	1.004	1.047
oldpeak	1.645	1.061	2.552
slope1	4.664	1.900	11.447
slope2	1.921	0.354	10.413
ca1	10.055	3.889	25.998
ca2	16.968	4.138	69.569
ca3	8.983	1.526	52.874
thal1	0.920	0.214	3.953
thal2	4.589	2.012	10.470

Both resting blood pressure and ST depression induced by exercise relative to rest are significant variables in this model because their confidence intervals exclude 1.

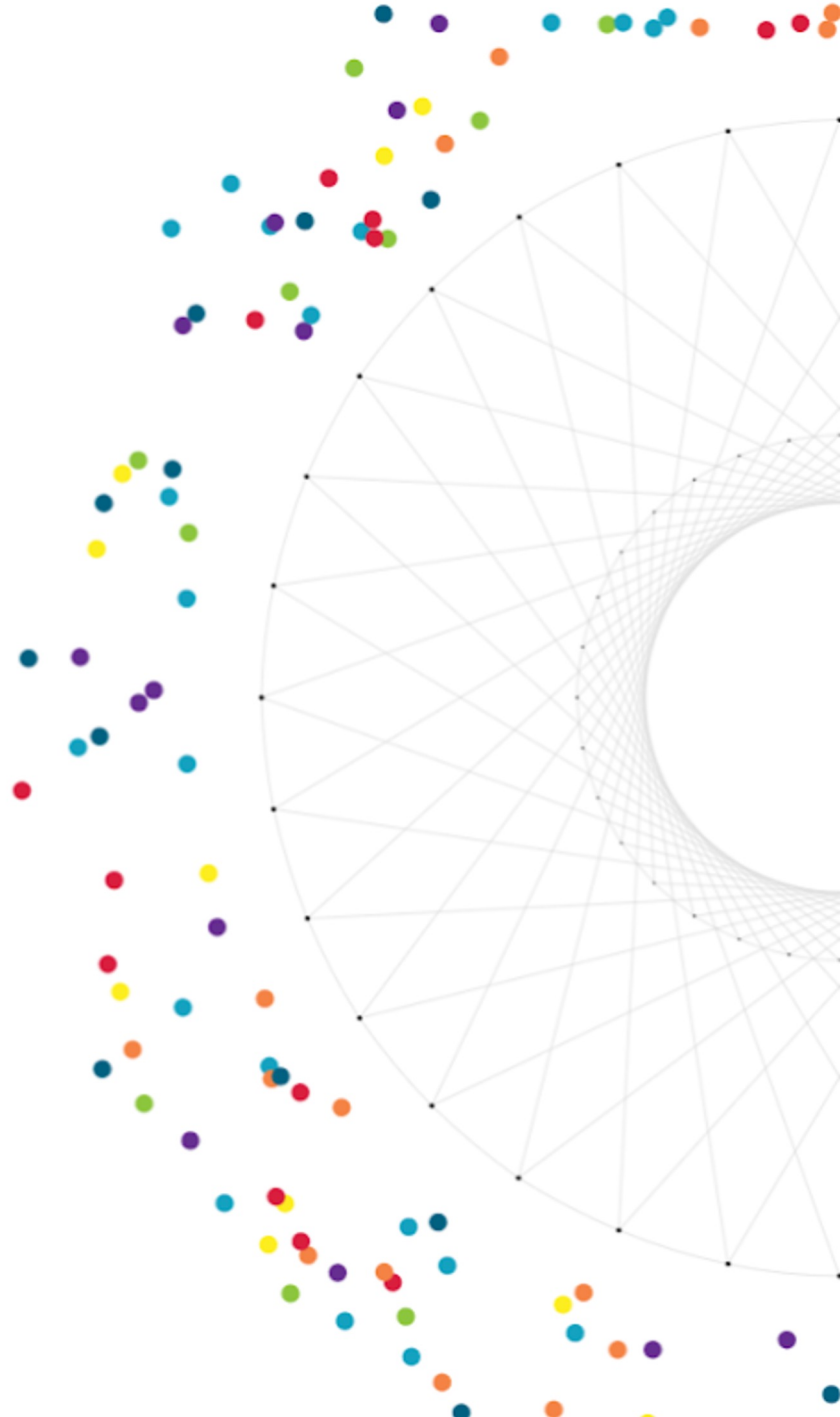
The odds of the patient having heart disease increase by 1.02% for each additional unit of resting blood pressure, while the odds increase by 1.65% for each additional unit of ST depression induced by exercise relative to rest.

Conclusion:

In conclusion, for the numerical variables, resting blood pressure and ST depression induced by exercise relative to rest all significantly influence the odds of having heart diseases.

Among all the categorical variables, patients with different gender and a different number of significant vessels coloured by fluoroscopy will significantly affect the odds of having heart disease.

The most significant difference occurs when we compare the patient who has asymptomatic chest pain and typical angina. This tells us that asymptomatic chest pain might be the most severe factor of heart disease.



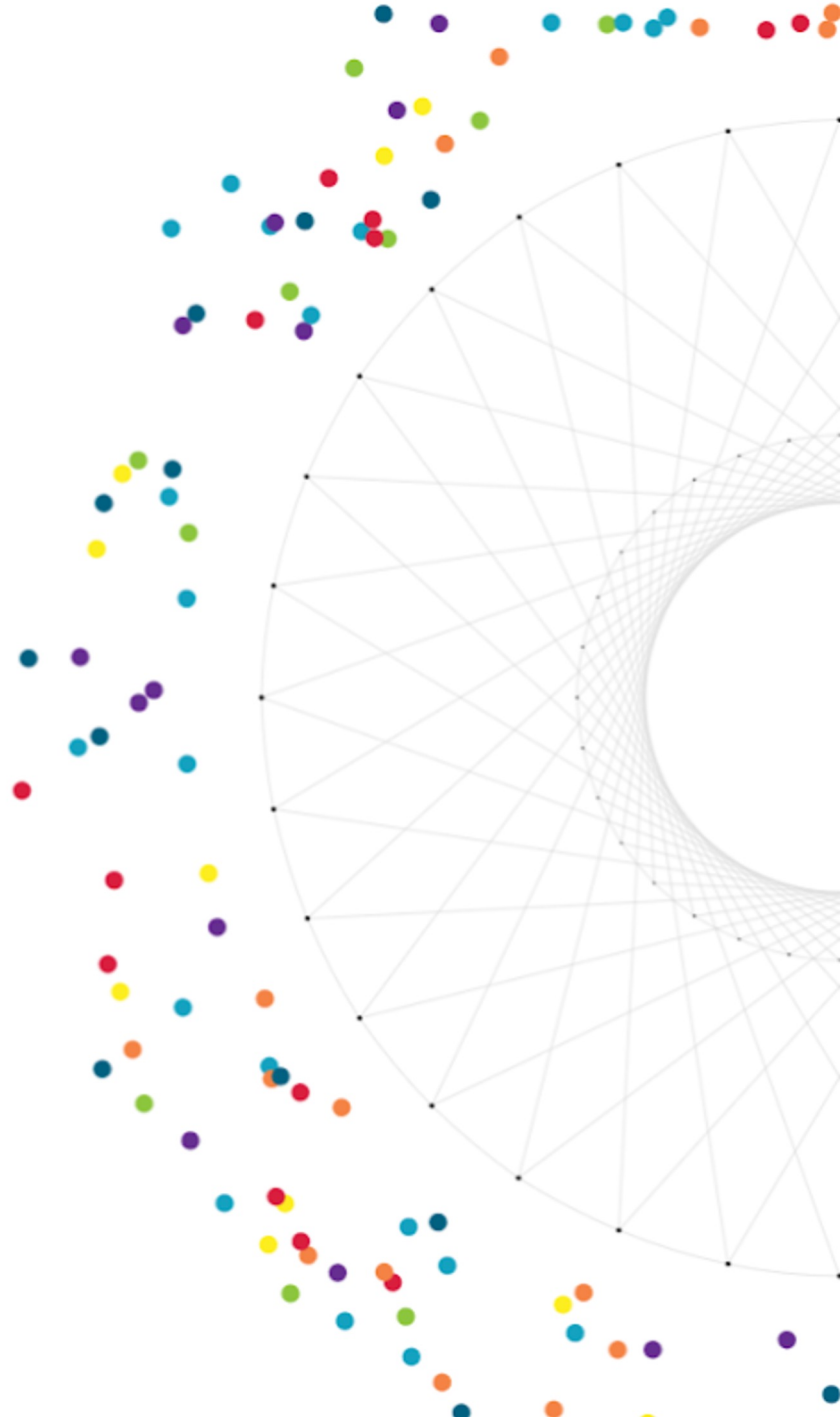
Limitations & Future Developments

PART SIX

Limitation:

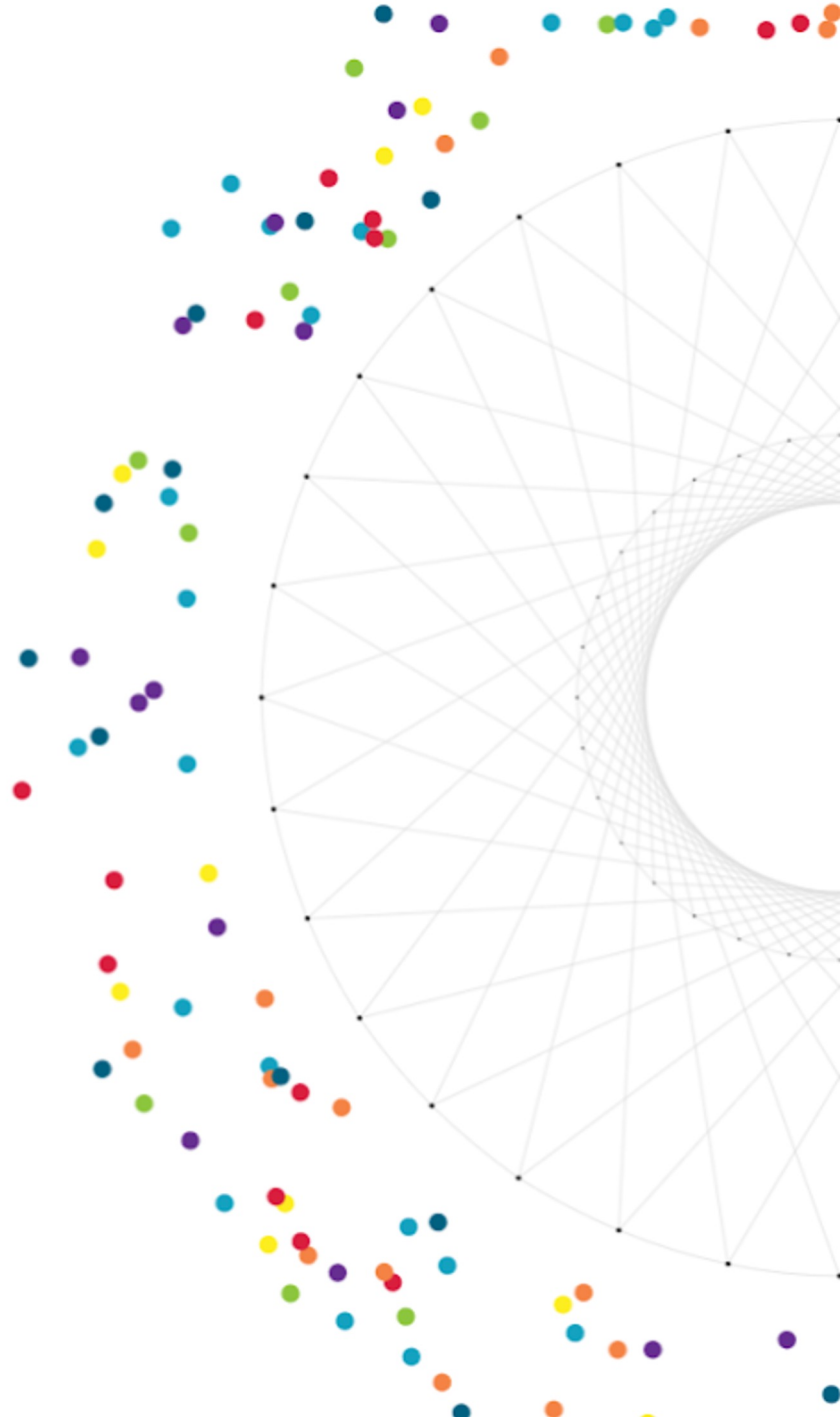
- Lack of data

- The samples can not represents all human beings



Future Development:

- Collecting more data from other countries
- Add other independent variables such as do samples take long-term medications or whether they have regularity of work and rest
- Use other Machine Learning techniques - Synthetic data



References:

García-Portugués, E. (n.d.). 5.7 Model diagnostics | Notes for Predictive Modeling. Retrieved April 9, 2022, from https://bookdown.org/egarpor/PM-UC3M/glm-diagnostics.html

Logistic Regression Assumptions and Diagnostics in R - Articles—STHDA. (n.d.). Retrieved April 9, 2022, from http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/

The Investopedia Team. (2022, February 8). *Variance inflation factor (VIF)*. Investopedia. Retrieved April 10, 2022, from https://www.investopedia.com/terms/v/variance-inflation-factor.asp#:~:text=Variance%20inflation%20factor%20measures%20how,standard%20error%20in%20the%20regression.

The Ultimate Guide to Synthetic Data: Uses, Benefits & Tools. (2018, July 19). https://research.aimultiple.com/synthetic-data/

Triveri, J. D. (2017, June 6). Goodness of Fit and Significance Testing for Logistic. Regression Models. The Pleasure of Finding Things Out. https://jtrive84.github.io/goodness-of-fit-and-significance-testing-for-logistic-regression-models.html

World Health Organization. (2022). *Cardiovascular diseases*. World Health Organization. Retrieved April 10, 2022, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

4.5 Multicollinearity | Practical Econometrics and Data Science. (n.d.). Retrieved. April 9, 2022, from http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/4-5-Multiple-collinearity.html

5 Ways to Deal with the Lack of Data in Machine Learning. (n.d.). KDnuggets. Retrieved April 9, 2022, from https://www.kdnuggets.com/5-ways-to-deal-with-the-lack-of-data-in-machine-learning.html/


THANK YOU FOR WATCHING