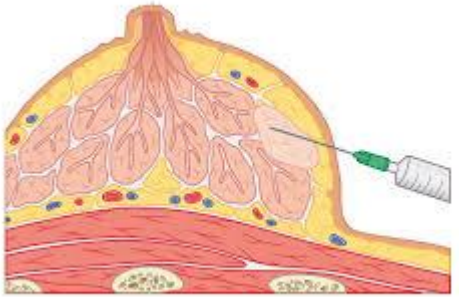
Breast Cancer Wisconsin

--Tumor Cells Classification

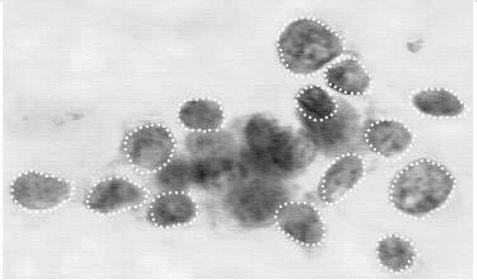
Shiyu Liu
Data 1030 | Brown University
10/22/2019
github.com/shiyuliu1/data1030_project.git

Intro: Problem Narrative

- Background
  - Breast cancer is the second most prevalent cancer among women worldwide
  - Early diagnosis of cancer type promote following treatment and facilitates patient management
- Data Intro.
  - Originally collected by Dr. William H. Wolberg
  - Cell images: sample cells obtained from breast mass tissue using Fine needle aspirations (FNAs)
  - Cytological(cell shape) view: dataset features are closely related to cell morphology
  - Classification problem: Cancer Diagnosis outcome as malignant or benign
  - Dataset from UCI Machine Learning Repository



Data preprocessing--X

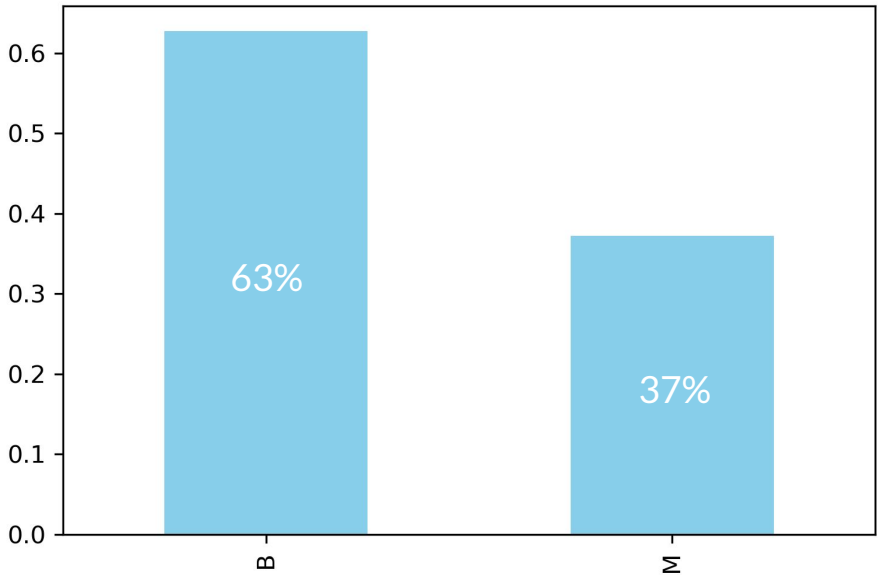


- 30 features: three types of measures for ten cell features
  - Mean
  - Standard Error
  - Worst Case
- No missing Value
- Numerical values are continuous
- Standard scaling for all features

Variables	Explanation
id	Patient ID
diagnosis	“M” as malignant and "B" as benign
radius	mean of distances from center to points on the perimeter
texture	standard deviation of gray-scale values
perimeter	cell perimeter
area	cell area
smoothness	local variation in radius lengths
compactness	perimeter^2 / area - 1.0
concavity	severity of concave portions of the contour
concave points	number of concave portions of the contour
symmetry	length difference between lines perpendicular to major axis
fractal dimension	"coastline approximation" - 1

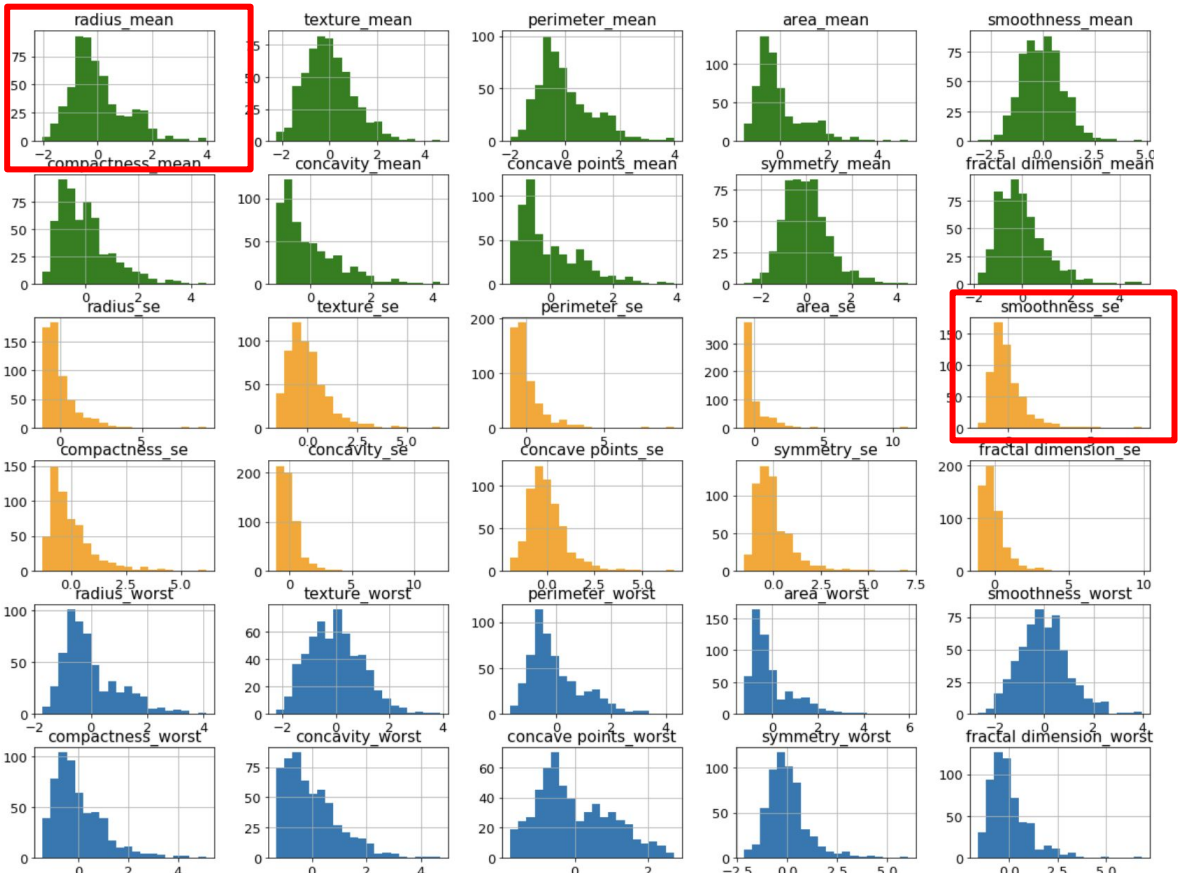
Data preprocessing--Target(diagnosis)

- Among 569 Observations:
  - 212 ‘M’ as malignant
  - 357 ‘B’ as benign
- Label encoder: labeled as ‘1’ and ‘0’

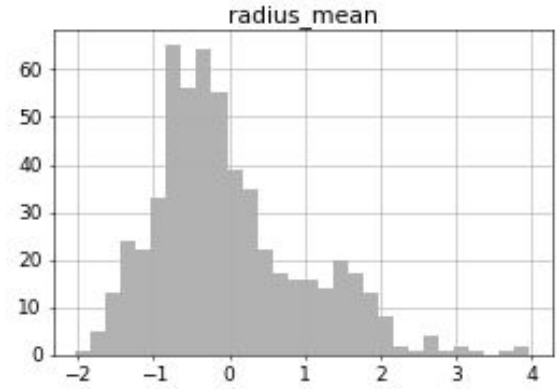


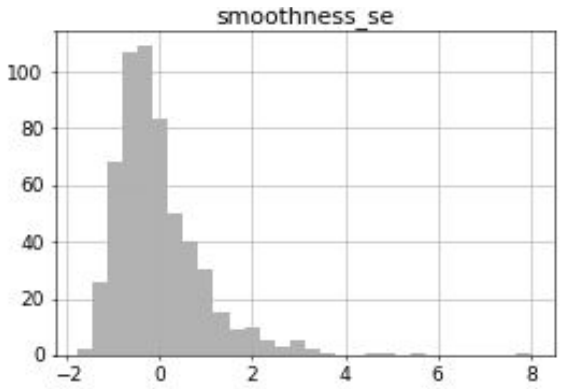
Exploratory Data Analysis

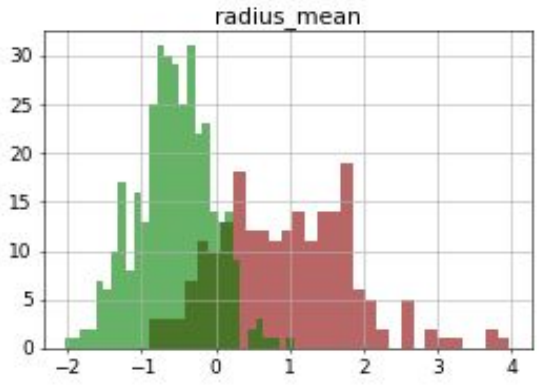
- Histogram for each feature
- Gives us how features distributed but not enough information relating the target variable

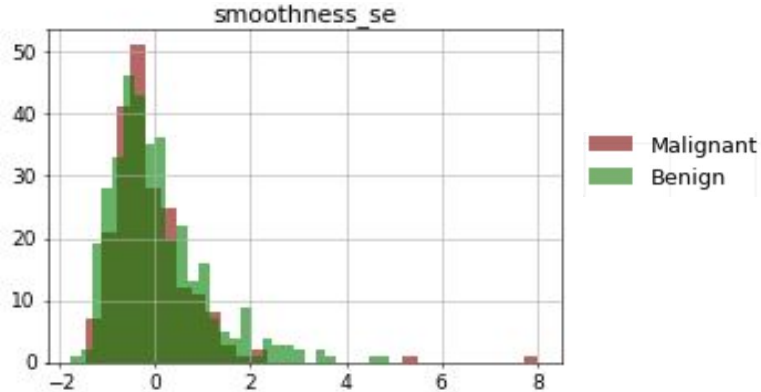


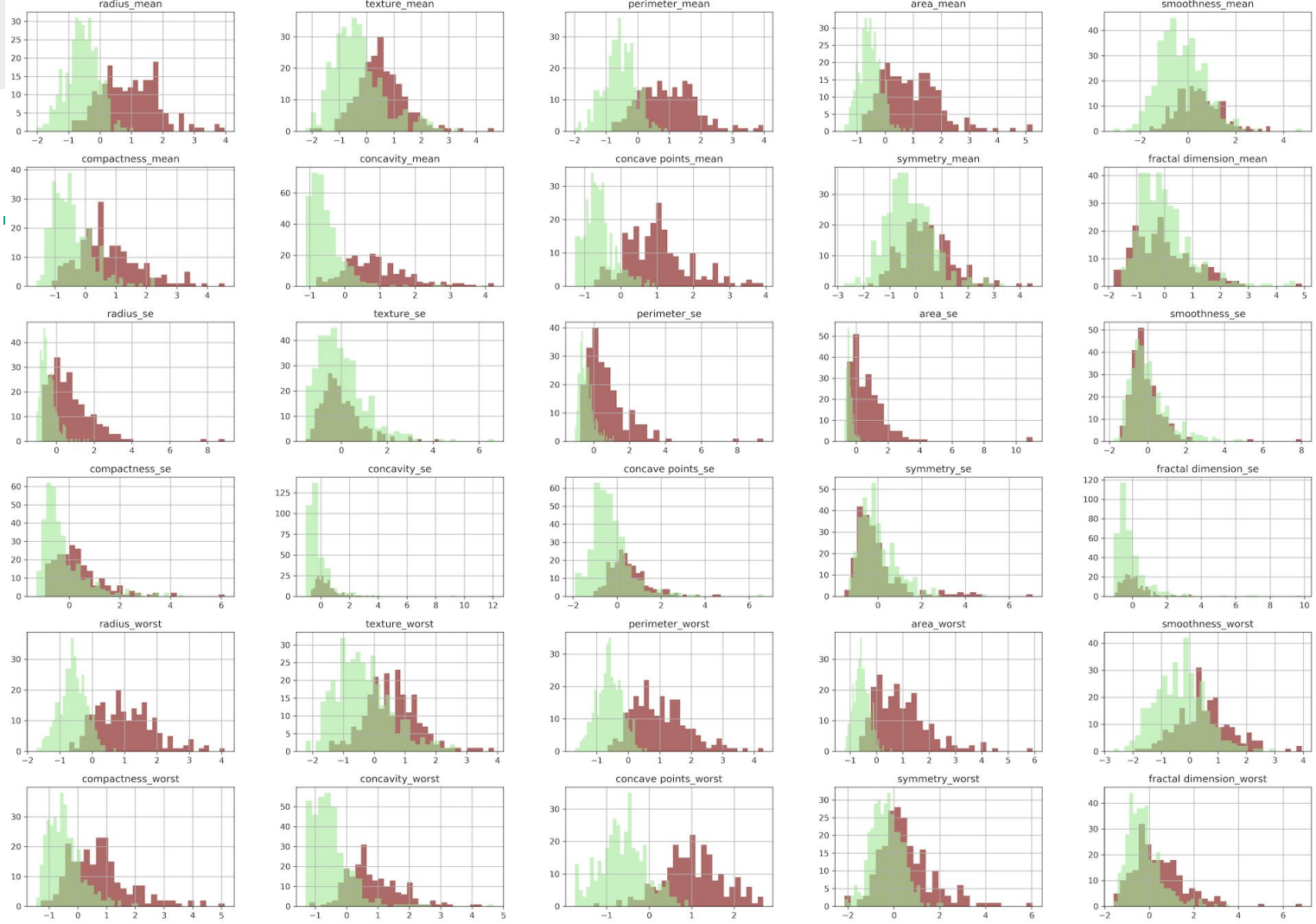
Exploratory Data Analysis--Feature examples





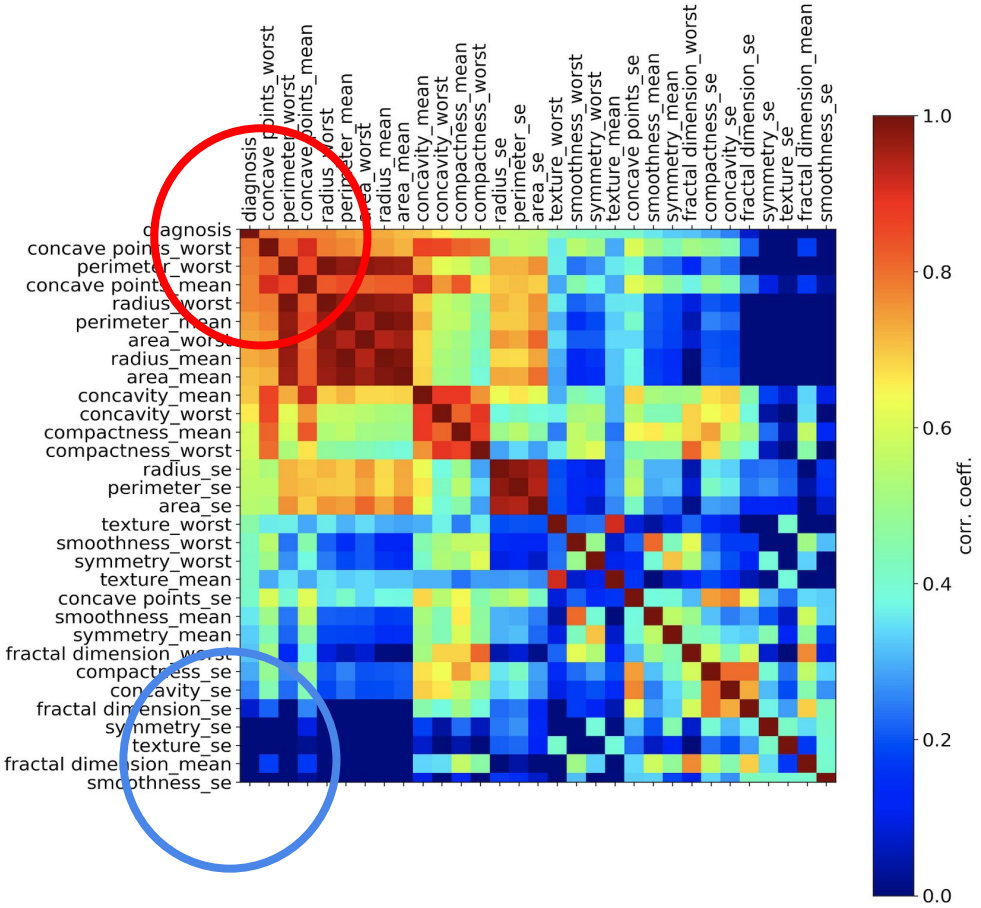





radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave points_mean
symmetry_mean
fractal dimension_mean
radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave points_se
symmetry_se
fractal dimension_se
radius_worst
texture_worst
perimeter_worst
area_worst
smoothness_worst
compactness_worst
concavity_worst
concave points_worst
symmetry_worst
fractal dimension_worst

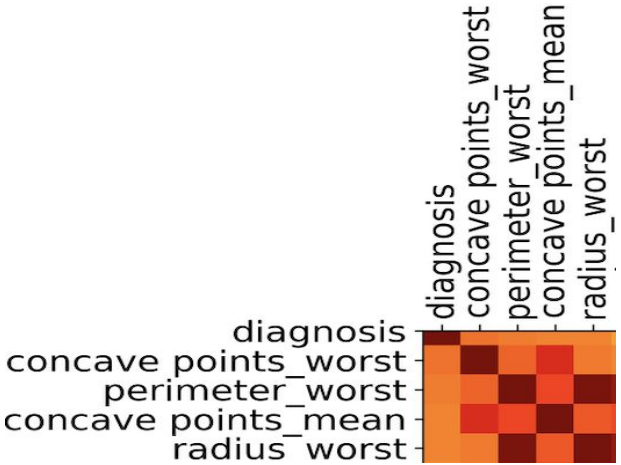
Exploratory Data Analysis--Variable Correlation

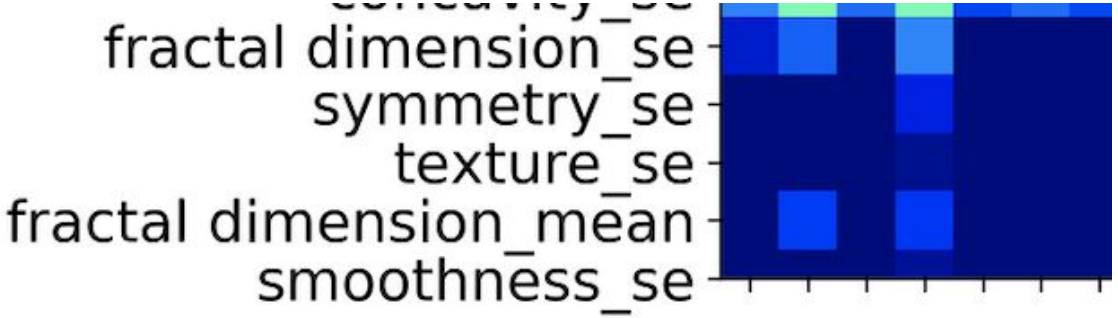
- Order the variables based on its correlation with target.
- Top three correlated features
  - Concave points worst
  - Perimeter worst
  - Concave points mean
- Last three correlated features
  - Smoothness standard error
  - Fractal dimension mean
  - Texture standard error



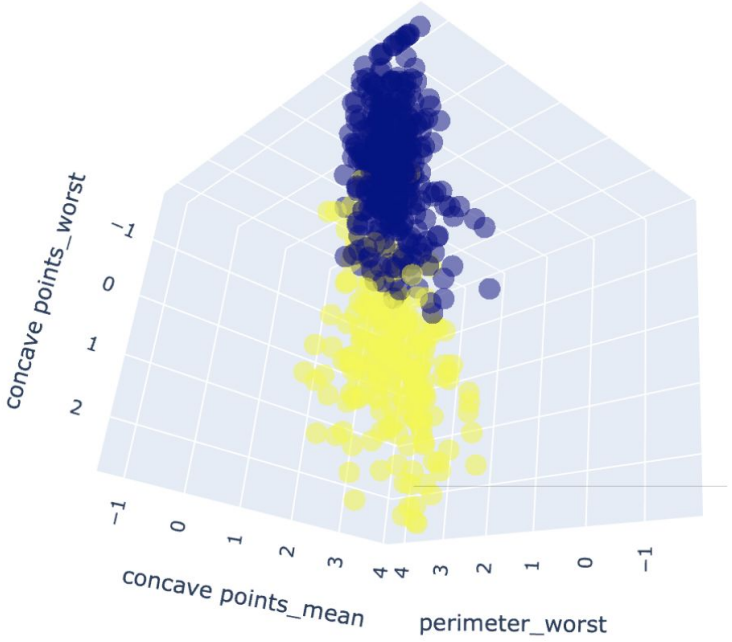
Exploratory Data Analysis--Variable Correlation

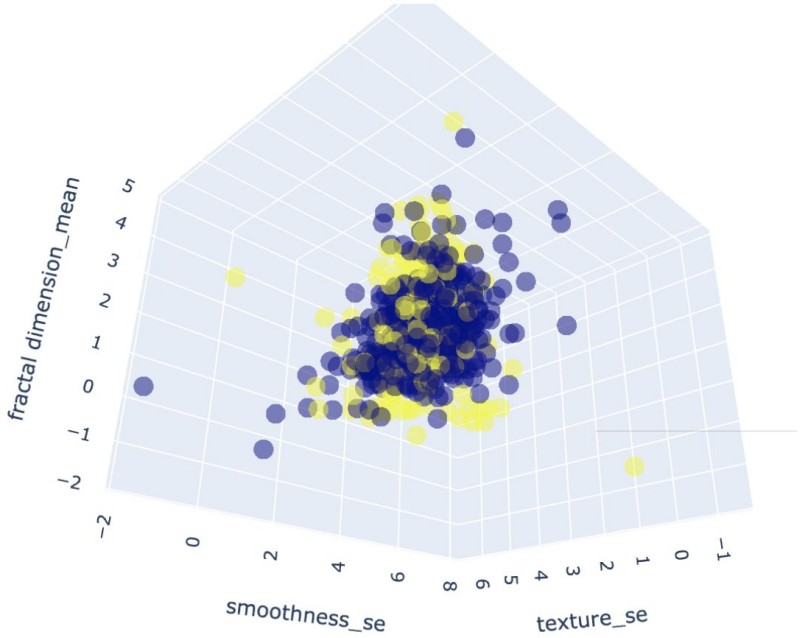
- Order the variables based on its correlation with target.
- Top three correlated features
  - Concave points worst
  - Perimeter worst
  - Concave points mean
- Last three correlated features
  - Smoothness standard error
  - Fractal dimension mean
  - Texture standard error





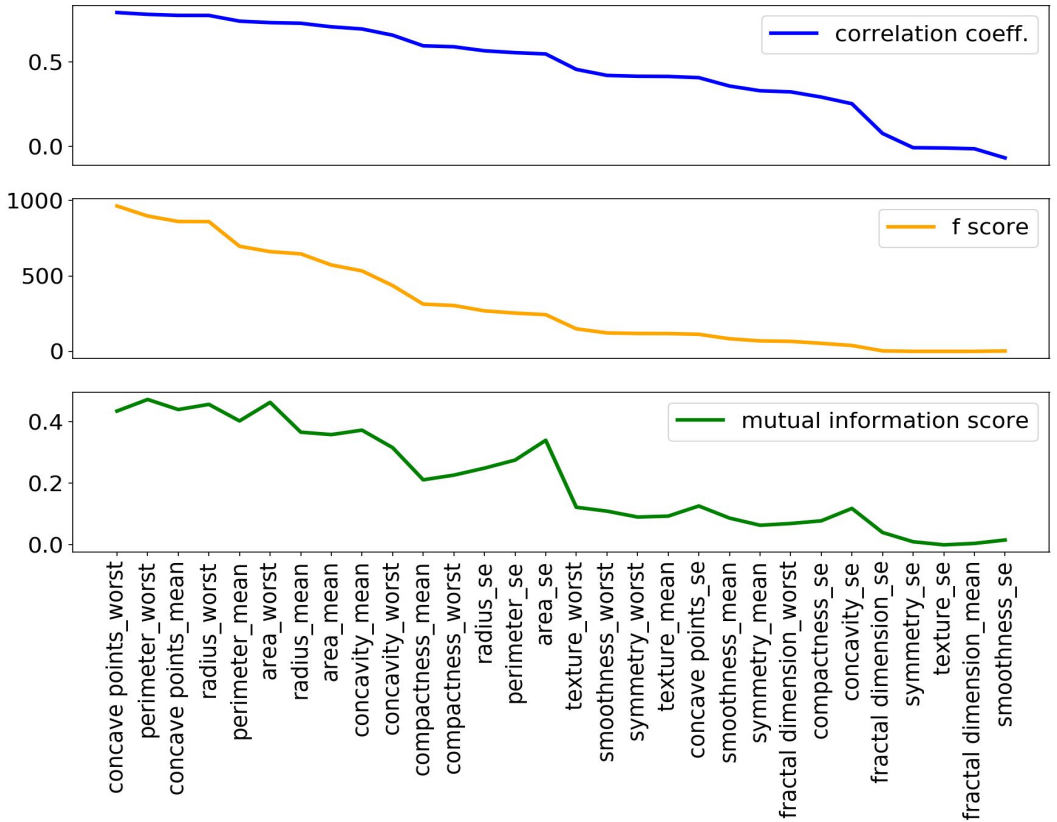
Exploratory Data Analysis-3D Scatterplots





Exploratory Data Analysis--Feature engineering

- Already know correlation matrix
- Conducted F test
- Conducted mutual information calculatio
- Ordered each scores based on the order c correlation coefficients with target variable



Future Work

- Select features and try different combinations based on EDA
- Apply different machine learning algorithms
- Perform sensitivity and specificity analysis and choose appropriate models
    - High sensitivity (recall) scores
    - Reasonable cost of specificity score

Questions?

Acknowledgement: **Andras Zsom, Sida Li**

References:

W.N. Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic
Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.

O.L. Mangasarian, W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577,
July-August 1995.

Yuh-Jeng Lee. Smooth Support Vector Machines. Preliminary Thesis Proposal Computer Sciences Department University of Wisconsin. 2000.