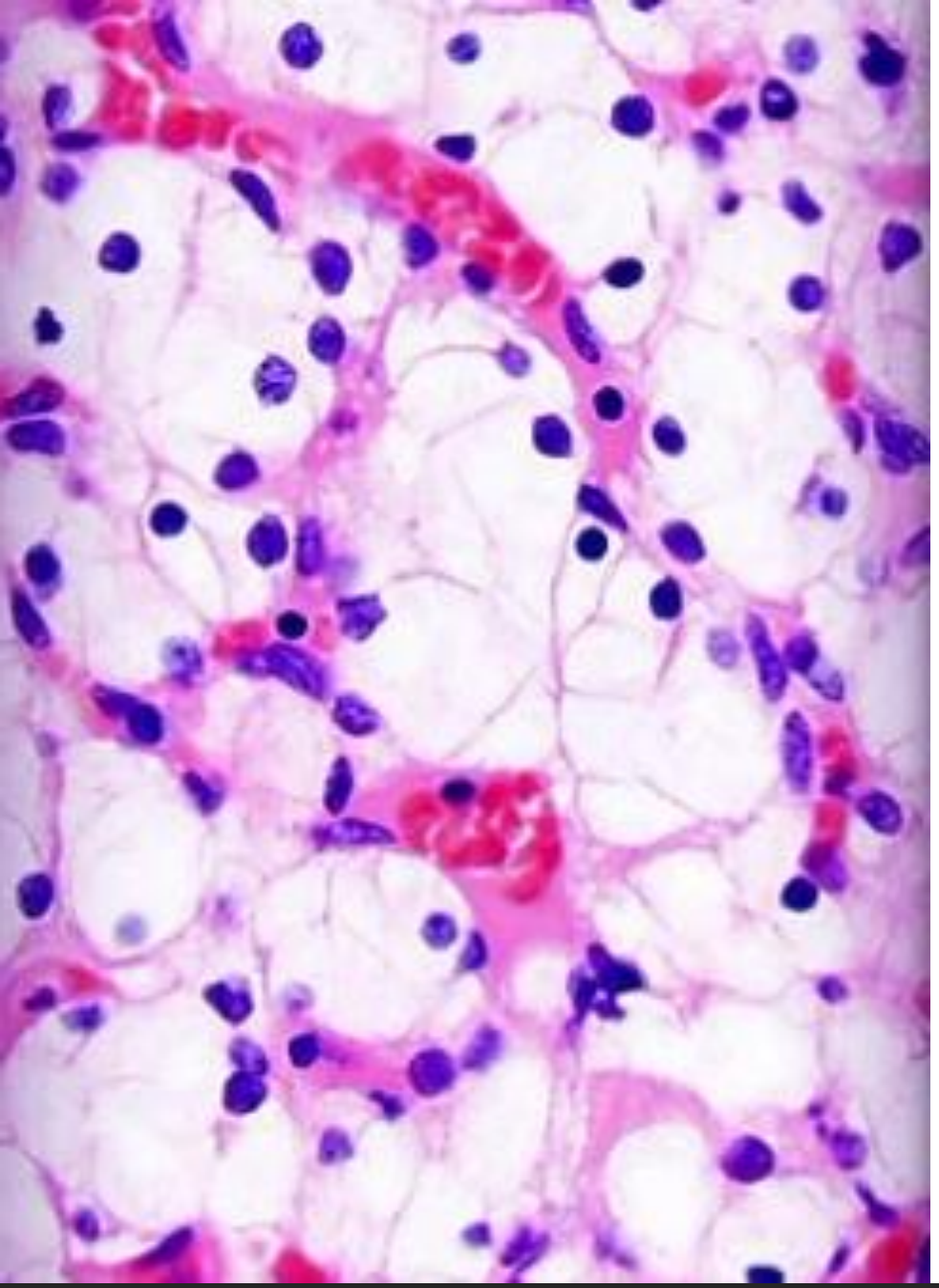


Breast Cancer Biopsy Data

AMY FEMAL

The Dataset

The biopsy dataset includes 569 observations collected by the University of Wisconsin in 1992 from fine needle aspiration biopsies on masses found in breast tissue. During this procedure, a hallow needle attached to a syringe is used to withdraw tissue or fluid from the mass. The variables in the dataset describe the cell nuclei obtained during the biopsies and the diagnosis.

Statistical Questions

What variables are significant to the diagnosis of breast cancer?

Using these significant variables, can I use machine learning to predict diagnosis with high accuracy?

Variables in the Dataset

radius - the mean of distances from the center of the cell nuclei to points on the perimeter

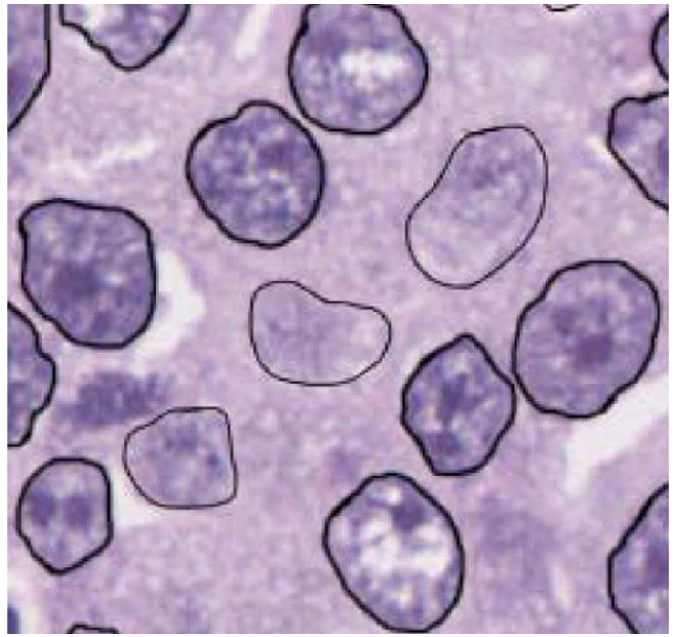
texture - the mean texture of the nuclei, described by the spatial arrangement and variation of grey values observed

perimeter - the mean distance around the nuclei

area - the mean area of the nuclei

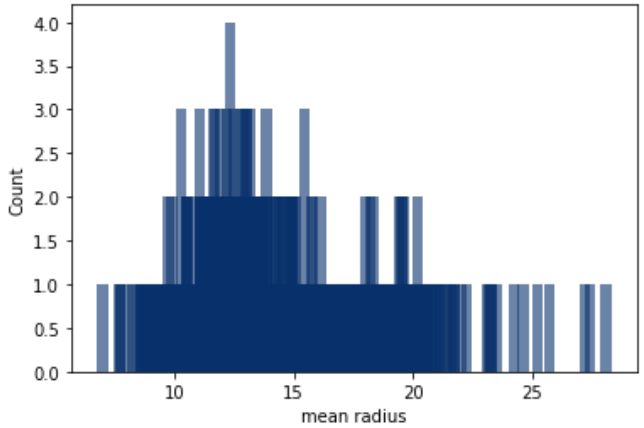
smoothness - the mean of local variation in radius lengths

diagnosis - the diagnosis of the mass, where 0 = malignant and 1 = benign.



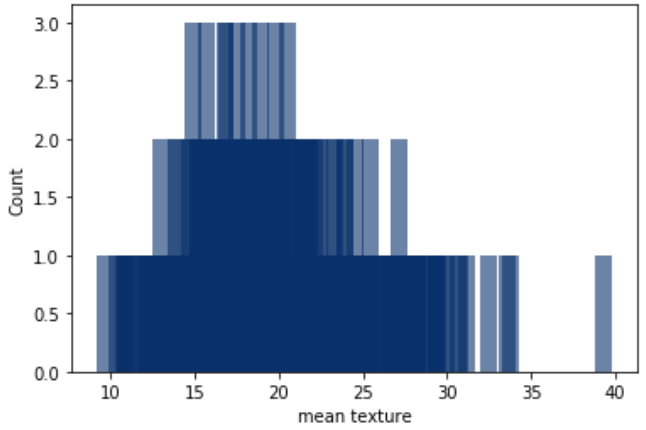
Variable Distribution

radius



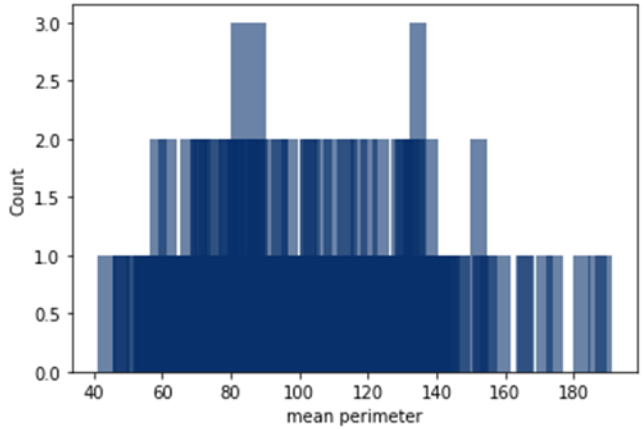
mean	14.13
mode	12.34
std	3.52
min	6.98
25%	11.70
50%	13.37
75%	15.78
max	28.11

texture



mean	19.29
mode	multimodal
std	4.30
min	9.71
25%	16.17
50%	18.84
75%	21.80
max	39.28

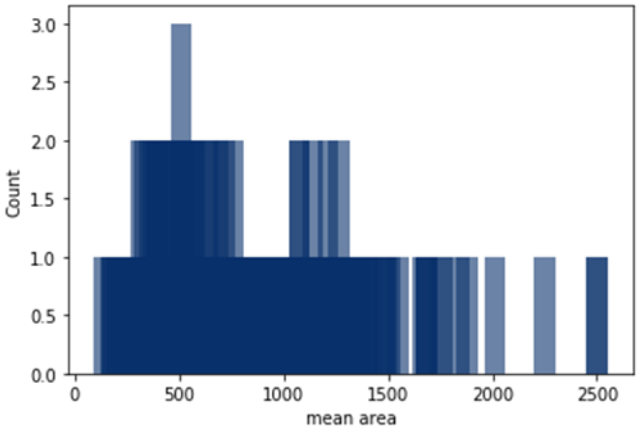
perimeter



mean	91.97
mode	multimodal
std	24.30
min	43.79
25%	75.17
50%	86.24
75%	104.10
max	188.50

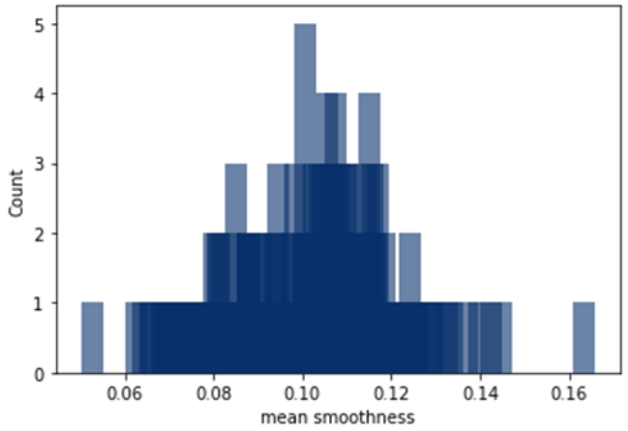
Variable Distribution

area



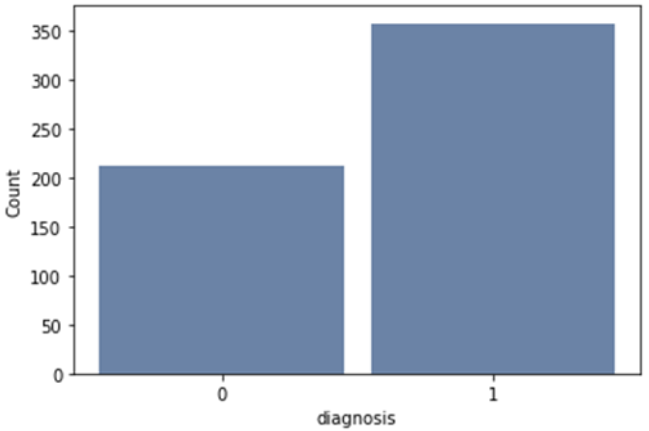
mean	654.89
mode	512.20
std	351.91
min	142.50
25%	420.30
50%	551.10
75%	782.70
max	2501.00

smoothness



mean	0.096
mode	0.101
std	0.014
min	0.053
25%	0.086
50%	0.096
75%	0.105
max	0.163

diagnosis



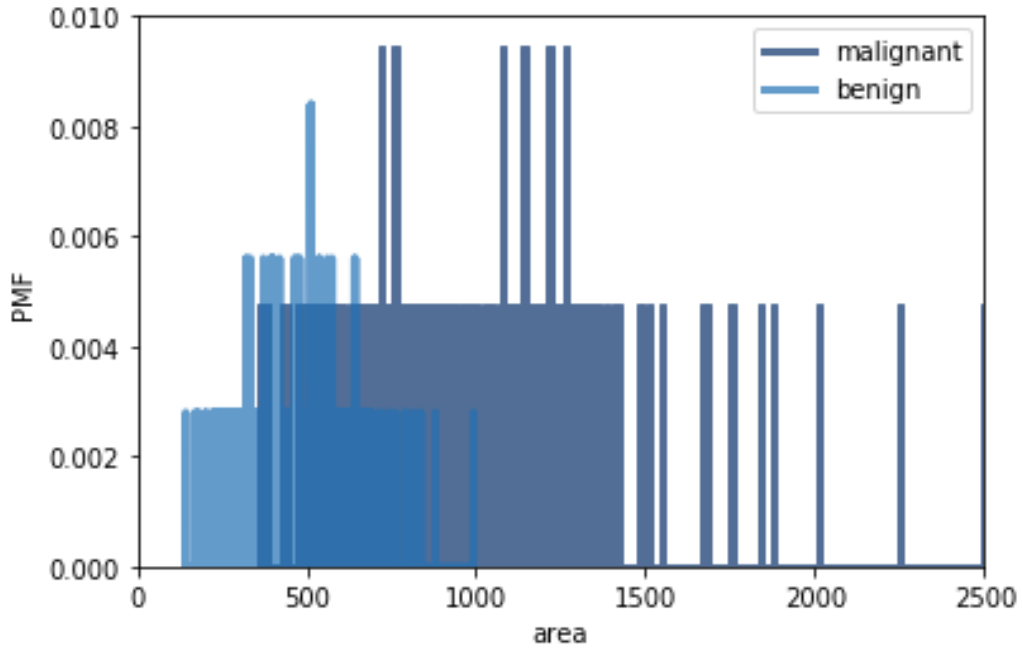
mean	0.627
mode	1
std	0.484
min	0
25%	0
50%	1
75%	1
max	1

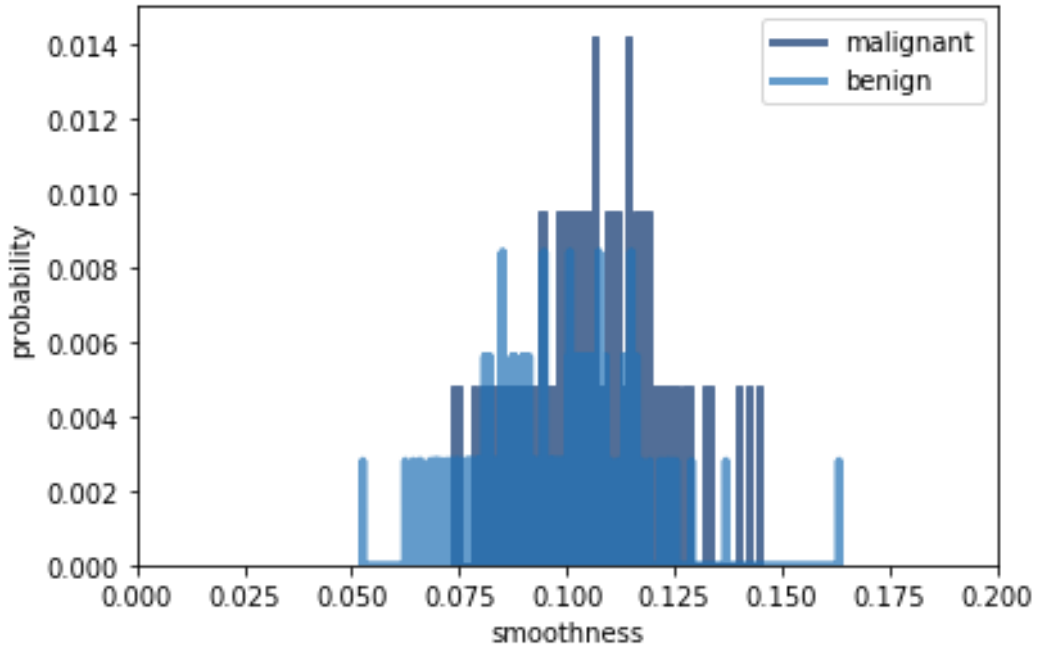
Variable Distribution

All variables, except diagnosis, have right-skewed distributions. Though many outliers are evident, they do not appear to be errors within the data. Therefore, I will use all observations in my analysis as they have been recorded.

Diagnosis – Malignant vs Benign

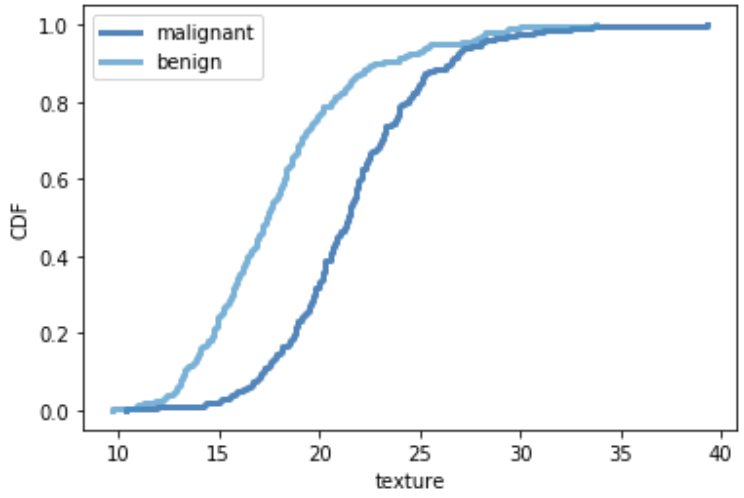
I have divided the data between the two diagnosis groups, malignant and benign. The PMFs of area and smoothness for malignant and benign diagnosis indicate cell nuclei with larger areas and smoothness are more likely to be malignant.

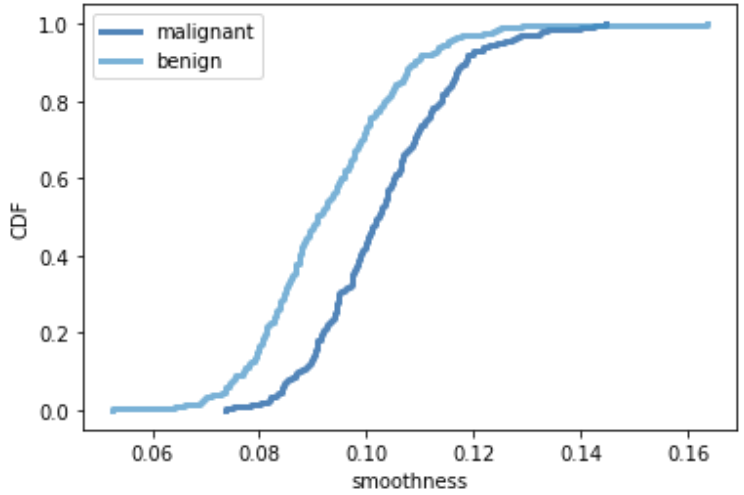


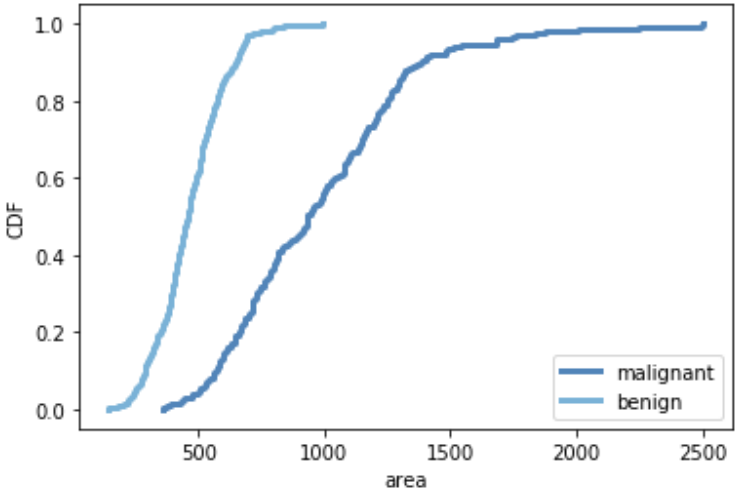


Comparing CDFs

The CDFs show texture, area, and smoothness are significantly larger throughout the distribution for malignant diagnosis.

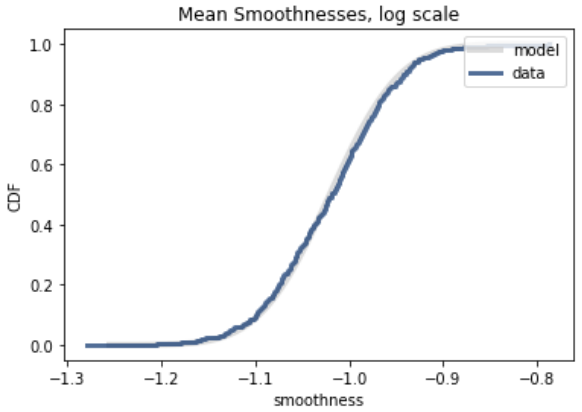


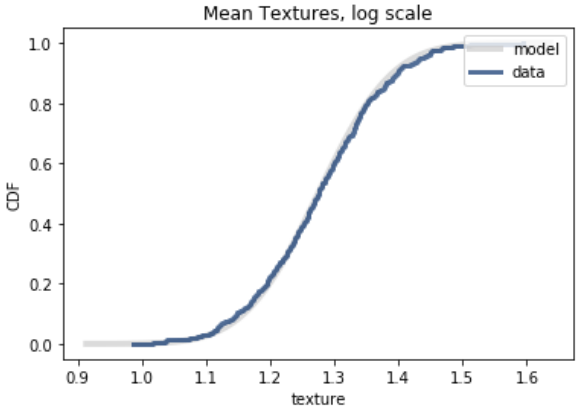


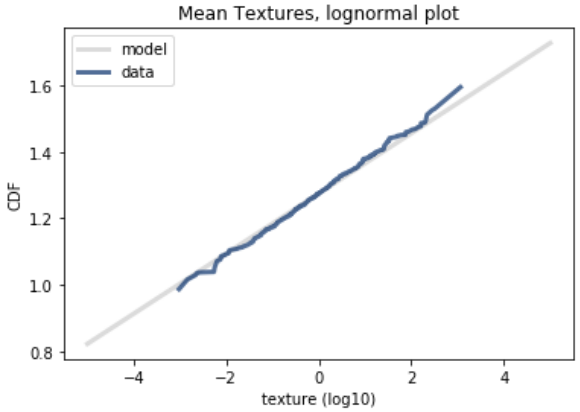


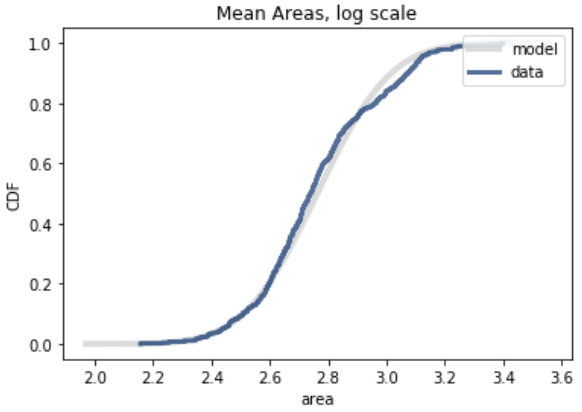
Lognormal Distribution

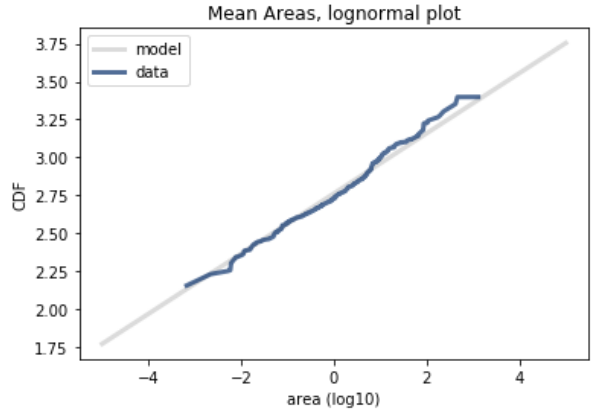
The CDFs of texture, area, and smoothness have the shape of a lognormal distribution. The lognormal model fits the data very well.





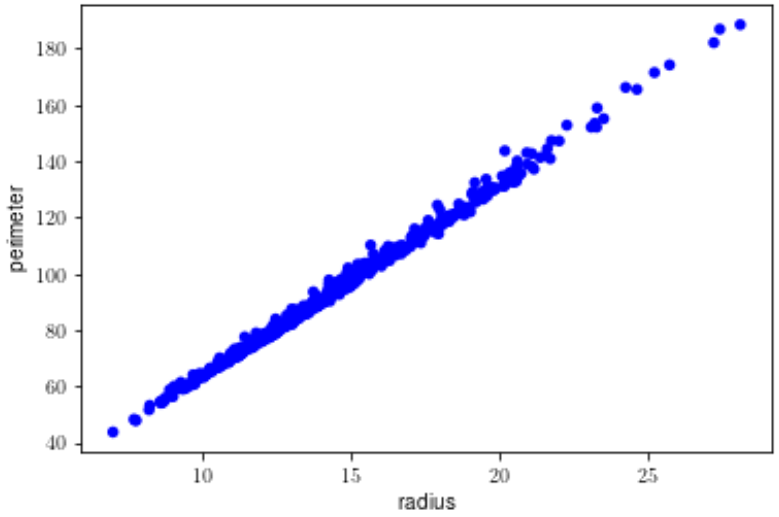




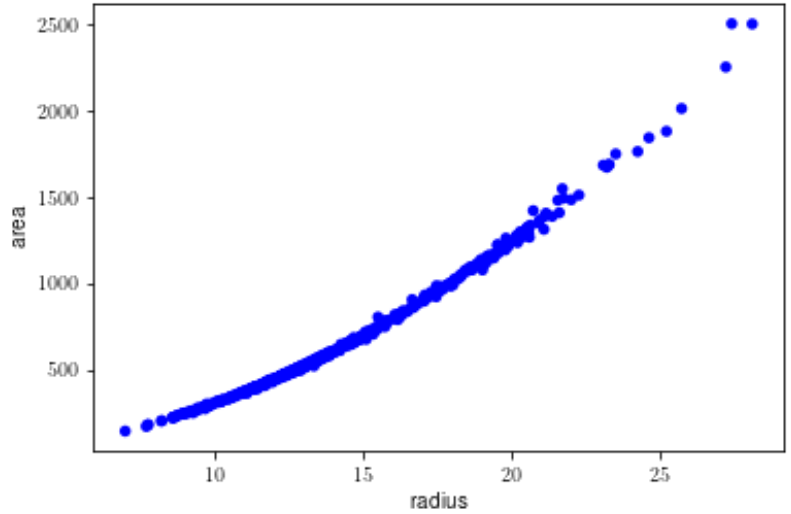


Scatter Plots

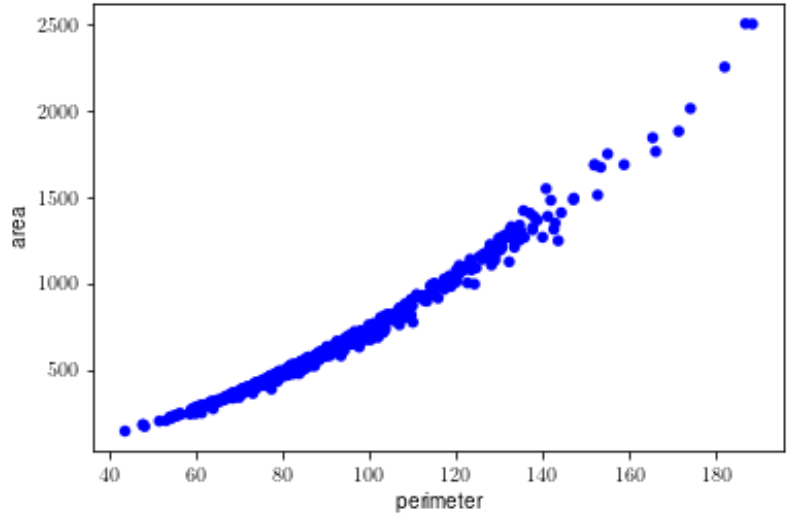
Correlation vs Causation



Correlation 0.99786



Correlation 0.98738

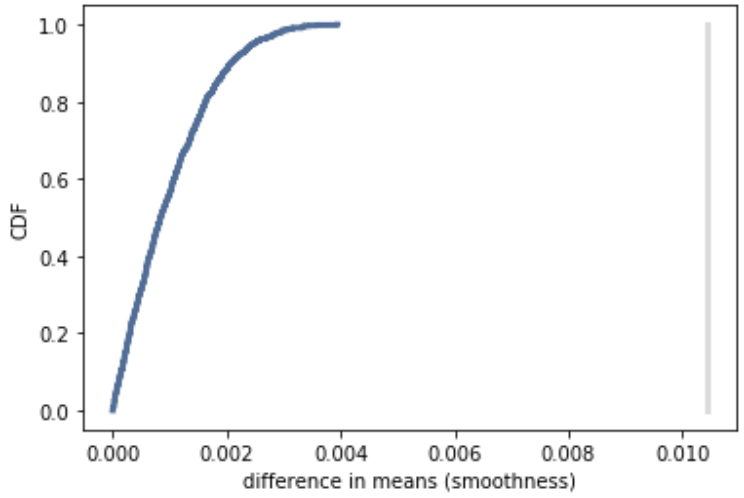


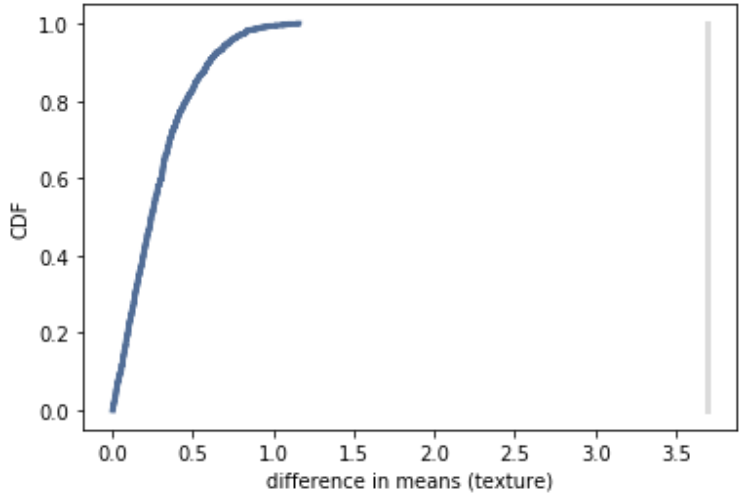
Correlation 0.98651

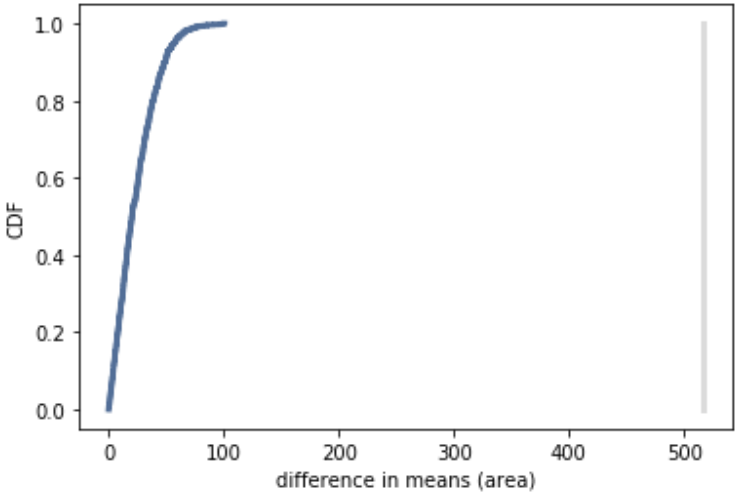
Like expected, perimeter, radius, and area of cell nuclei are highly correlated. Considering the geometry of a circle, the perimeter (circumference) is $2\pi r$ and the area is πr^2. Therefore, I conclude that radius is causally correlated with perimeter and area.

Hypothesis Test

Permutation hypothesis tests on texture, area, and smoothness yield p-values of 0. These results confirm the difference in means of texture, area, and smoothness between malignant and benign diagnosis are statistically significant. The plots below show that the CDFs never intersect the observed differences.







Logistic Regression

Based on the high correlation of radius, perimeter, and area, I have decided to remove radius and perimeter from the model. As the data follows a lognormal distribution, I have used the log transformation of texture, area, and smoothness in a logistic regression model to predict diagnosis with 94% accuracy.

Logit Regression Results

Dep. Variable:	malignant	**No. Observations:**	569
Model:	Logit	**Df Residuals:**	565
Method:	MLE	**Df Model:**	3
Date:	Thu, 06 Aug 2020	**Pseudo R-squ.:**	0.7670
Time:	19:26:20	**Log-Likelihood:**	-87.533
converged:	True	**LL-Null:**	-375.72
Covariance Type:	nonrobust	**LLR p-value:**	1.333e-124

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-51.6957	6.994	-7.391	0.000	-65.404	-37.987
log_textures	18.4502	2.771	6.658	0.000	13.019	23.882
log_smoothnesses	36.8540	4.929	7.477	0.000	27.194	46.514
log_areas	23.2399	2.604	8.925	0.000	18.136	28.344

References

Fine Needle Aspiration (FNA) Biopsy of the Breast: Breast Aspiration. (n.d.). Retrieved August 06, 2020, from https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast.html

Singh Suwal, M. (2018, September 26) Breast Cancer Prediction Dataset. Retrieved July 07, 2020 from, www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset