Breast Cancer Data Analysis

Annie Zhang

Tableau Dashboard

GitHub Repository

Project Overview

- Introduction & Project Motivation
- Dataset & Initial Exploration
- SQL Data Profiling
- Statistical Analysis (R)
- Visualization Dashboard (BI Perspective)
- Clinical Interpretation of Findings

Objectives

- Analyze tumor morphology features to distinguish **malignant** vs. **benign** cases
- Apply statistical and BI methods to generate evidence-based findings
- Demonstrate SQL data profiling + R statistical modeling workflow
- Demonstrate data-driven reasoning relevant to healthcare decision-making

Dataset & Tools:
SQL: data validation, aggregations, top-k selection
R: correlation analysis, visualization, hypothesis testing
BI Thinking: categorical grouping, KPI construction, insight synthesis

---

Goal: Identify tumor characteristics that distinguish malignant (M) from benign (B) breast cancer cases.

Key Feature Definitions

Understanding the Key Tumor Features

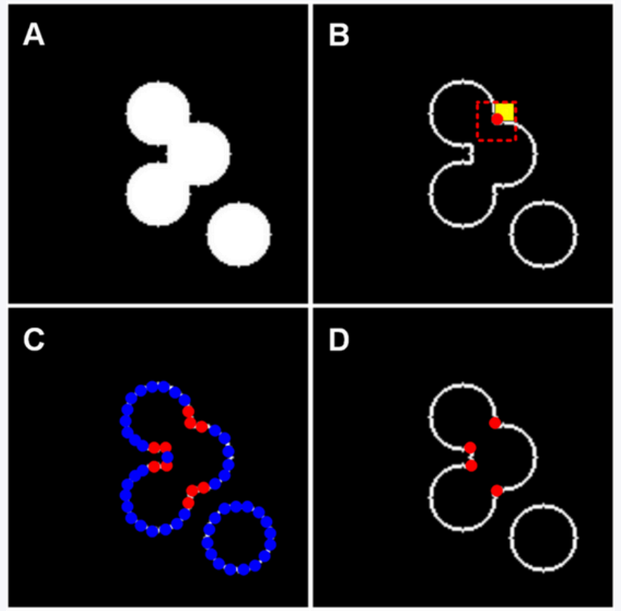
Size & Texture Features

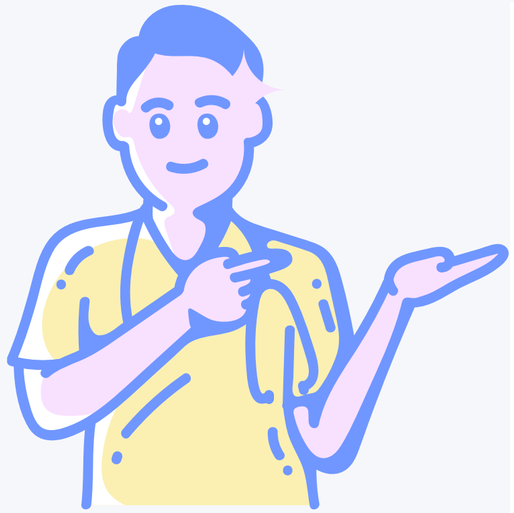
1. ***Radius***: Mean distance from the tumor center to its perimeter points.
2. ***Texture***: Standard deviation of gray-scale intensity values (tissue heterogeneity).
3. ***Perimeter***: Length of the tumor boundary.
4. ***Area***: Total enclosed area inside the boundary.
5. ***Smoothness***: Local variation in radius lengths; lower smoothness = rougher edges.

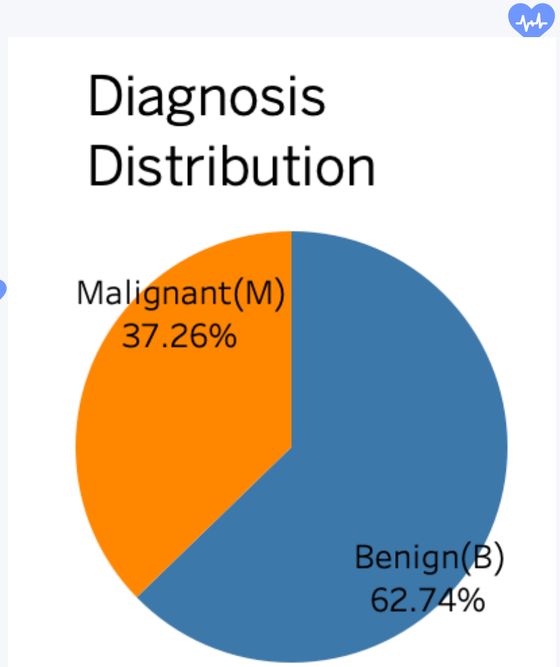
Shape Irregularity Features

(These often distinguish malignant tumors more strongly than size alone.)

1. ***Compactness*** : (perimeter² / area – 1.0)
   - Higher values indicate a less regular shape.
2. ***Concavity*** : Severity of inward-curving sections of the contour.
   - Higher = more structural distortion.
3. ***Concave Points*** : Number of concave indentations on the boundary.
   - More concave points = more irregular/perhaps malignant growth.







Understanding the dataset

To understand the diagnostic patterns, I first profiled the dataset using SQL and R:

- 569 total tumor records
- Each record contains 30+ morphological features (radius, area, concavity, smoothness, etc.)
- Data grouped by Diagnosis: Benign (B) vs Malignant (M)

Description: df [6 × 32]

	id <int>	diagnosis <chr>	radius_mean <dbl>	texture_mean <dbl>	perimeter_mean <dbl>	area_mean <dbl>
1	842302	M	17.99	10.38	122.80	1001.0
2	842517	M	20.57	17.77	132.90	1326.0
3	84300903	M	19.69	21.25	130.00	1203.0
4	84348301	M	11.42	20.38	77.58	386.1
5	84358402	M	20.29	14.34	135.10	1297.0
6	843786	M	12.45	15.70	82.57	477.1

6 rows | 1–7 of 32 columns

DIAGNOSIS DISTRIBUTION SHOWING MODERATE CLASS IMBALANCE.

Key Finding: Diagnosis Imbalance

During this step, I discovered a class imbalance:

- ***Benign: ~63%***
- ***Malignant: ~37%***

This led me to explore how feature distributions differ between these two diagnostic categories.

Intuition → First Hypothesis

Hypothesis:
"Tumor radius is a key factor in determining malignancy."

This intuition comes from clinical reasoning: malignant tumors often grow larger and faster.
To check this hypothesis, I analyzed ***size-related metrics*** such as radius, area, and perimeter using SQL exploratory queries and early BI-style plots.

Why This Hypothesis Matters Clinically

If size truly differentiates malignant tumors, it can support:

- Faster risk triage
- Prioritizing screening for high-risk cases
- Streamlining diagnostic workflows
- Simplifying clinical decision dashboards

Immediate Evidence from SQL

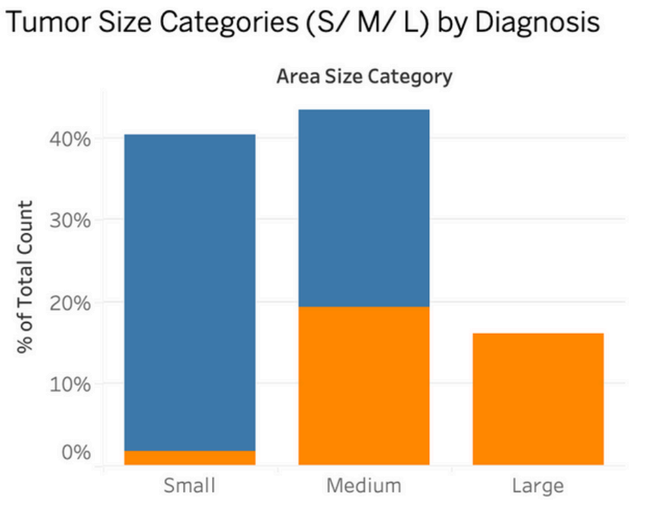
- The **Top 10 largest** tumors (by ***area_mean***) were all malignant.
- **Benign** tumors appeared mostly in the Small category, after grouping tumors into Small/Medium/Large.
- **Malignant** cases dominated the Medium and Large categories.

Interpretation:

Early on, the data supports size as a meaningful clinical indicator.

Top 10 Largest Tumors

Id	Area M..	Radius Me..	Diagnosis	
911296202	2501	27.42	Malignant(M)	
8810703	2499	28.11	Malignant(M)	
873592	2250	27.22	Malignant(M)	
899987	2010	25.73	Malignant(M)	
8611555	1878	25.22	Malignant(M)	
91762702	1841	24.63	Malignant(M)	
865423	1761	24.25	Malignant(M)	
89812	1747	23.51	Malignant(M)	
8712289	1686	23.27	Malignant(M)	
878796	1685	23.29	Malignant(M)	



Statistical Validation (Hypothesis Testing)

```
t.test(radius_mean ~ diagnosis, data = data)
```

Conclusion from Test

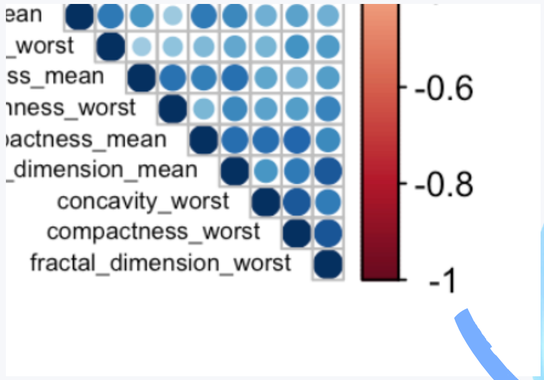
- $p < 2.2 \times 10^{-16}$
- Confirms *malignant tumors* have significantly larger radius
- 95% CI entirely negative → strong statistical certainty
- Supports using tumor size as a screening KPI

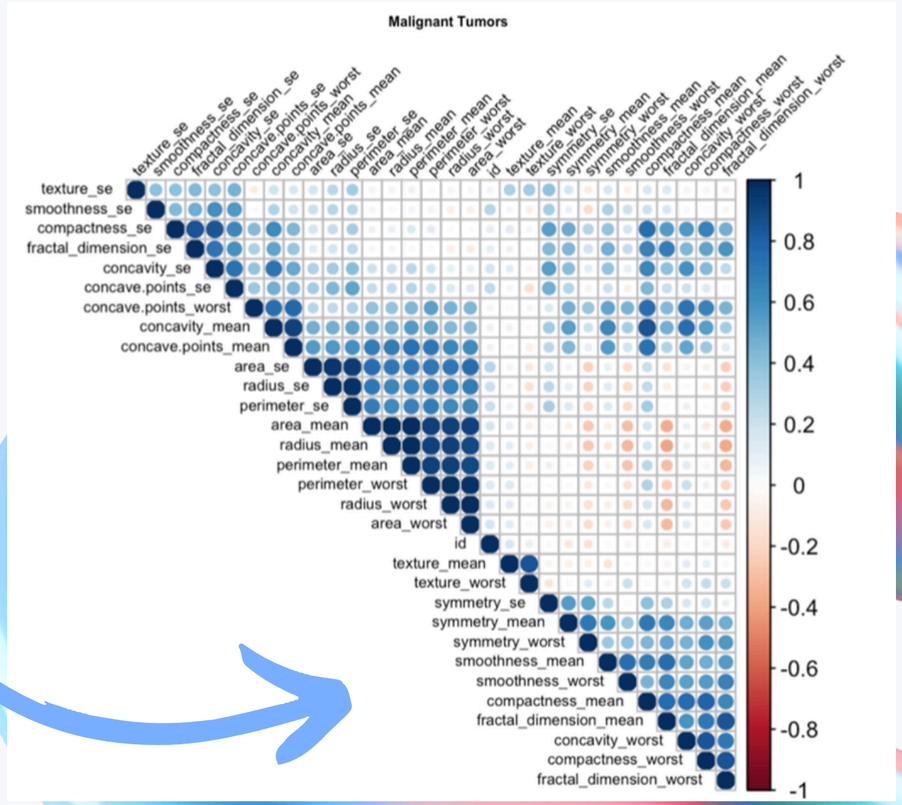
```
        Welch Two Sample t-test

data:  radius_mean by diagnosis
t = -22.209, df = 289.71, p-value < 2.2e-16
alternative hypothesis: true difference in means between group B and group M is not equal
to 0
95 percent confidence interval:
 -5.787448 -4.845165
sample estimates:
mean in group B mean in group M
       12.14652        17.46283
```

NEW DISCOVERY: **IRREGULARITY** MATTERS MORE THAN EXPECTED

My interpretation evolved from:
“Malignant tumors are big.”
to
“Malignant tumors are big and structurally irregular.”

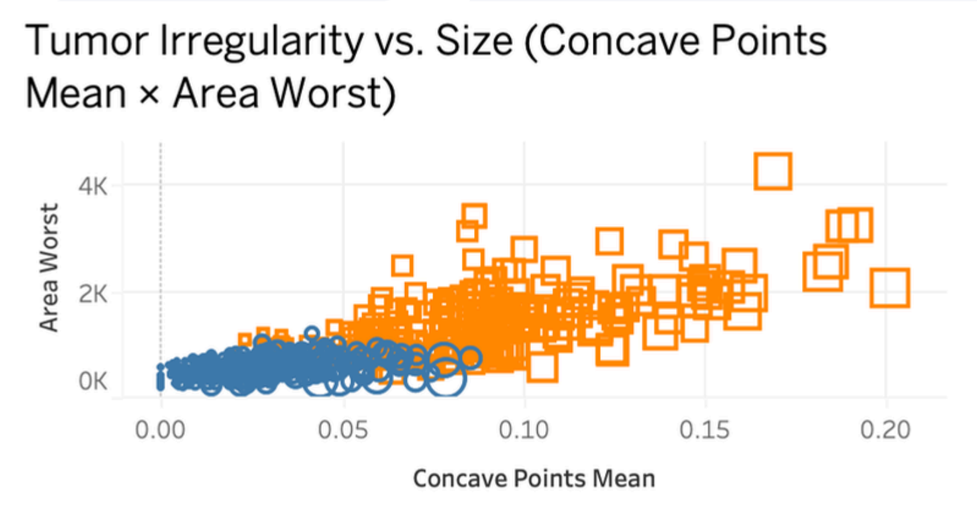
BEYOND SIZE: WHAT CORRELATION REVEALED



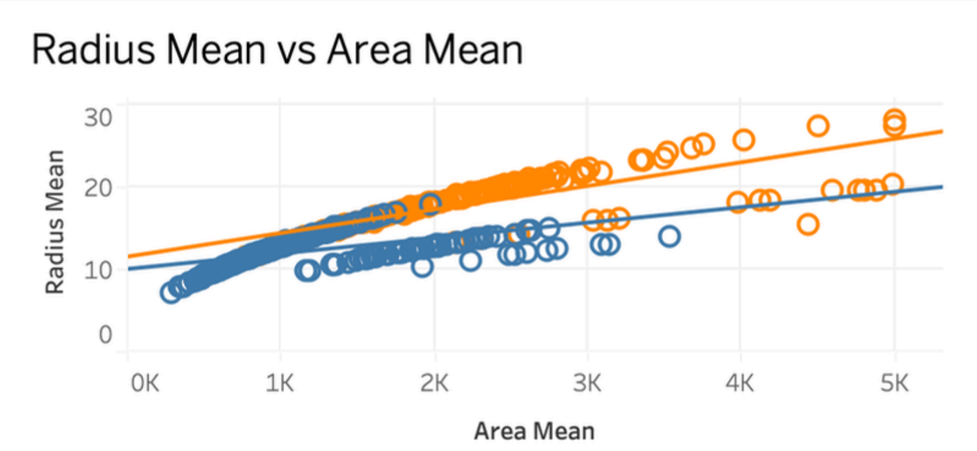


- **Size features** cluster tightly, confirming that malignant tumors are generally larger.
- But **concavity-related features** stand out as even more strongly correlated, revealing that irregularity, not just size, is a defining characteristic of malignant tumors.

VISUAL EVIDENCE: SHAPE IRREGULARITY & VARIABILITY



- Malignant tumors cluster at higher concavity + larger area
- Benign tumors cluster lower on both axes



- Malignant tumors follow a steeper growth pattern
- Benign tumors are smaller with narrower spread

Clinical Interpretation

Malignant tumors tend to be:

- Larger
- More irregular (concave indentations)
- More heterogeneous (variation metrics like area_worst)

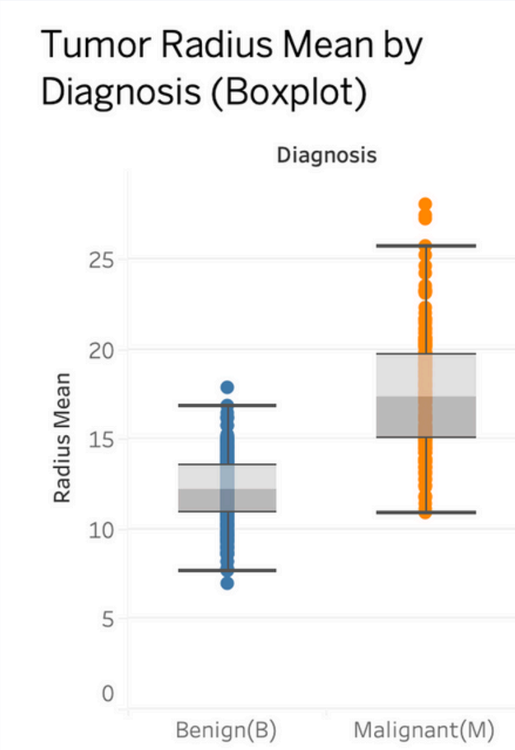
BOXPLOT + FEATURE BREAKDOWN

Feature Breakdown by Diagnosis

	Diagnosis	
	Benign(B)	Malignant(M)
Area Mean	463	978
Area Worst	559	1,422
Concave Points Me..	0.0257	0.0880
Concavity Mean	0.0461	0.1608
Radius Mean	12.1	17.5

Malignant tumors have:

- 3× larger Area Worst
- 3× higher Concave Points Mean
- Higher concavity overall
- Substantially larger Radius Mean



The *median radius* for malignant tumors is dramatically higher, with wider IQR and many extreme values.

Interpretation

Both *size* and *shape irregularity* consistently differentiate malignant tumors.

Key Clinical Insights

My findings mirror clinically validated diagnostic patterns used in radiology and oncology practice:

Tumor size as an early risk flag

Malignant tumors tend to be larger due to rapid, uncontrolled growth (Elmore et al., 2005; Guray & Sahin, 2006).

Clinical suggestion: Integrate simple size-based thresholds into early-screening triage tools to flag patients who may need accelerated imaging or biopsy..

Irregular Shape as a Key Diagnostic Marker

High concavity and irregular borders strongly indicate malignancy (ACR, 2013; Chen et al., 2014).

Clinical suggestion: Use ***concavity-derived*** features to refine BI-RADS scoring, especially for borderline cases where size alone is inconclusive.

a standardized scoring system created by the American College of Radiology (ACR)

Structural Heterogeneity as a Sign of Aggressive Disease

Malignant tumors show greater variability in “worst” and SE metrics, reflecting biological heterogeneity (Burrell et al., 2013; Polyak, 2011).

Clinical suggestion: Incorporate heterogeneity metrics into decision-support dashboards to help clinicians identify tumors that may require more urgent intervention.

THANKS :)

Tableau Dashboard