Breast Cancer Wisconsin Data Analysis

CellSight Diagnostics

Introduction

Briefly introduce your "company" and the diagnostic challenge

Your firm has been awarded a contract to analyze breast cancer diagnostic imaging data derived from fine needle aspirate (FNA) samples.
Your client is a radiology network exploring computer-aided diagnosis (CAD) systems for breast cancer screening. They've provided you with cell nuclei measurements from digitized FNA images and need your team to identify which cellular characteristics best distinguish malignant from benign tumors.

Dataset Description

Wisconsin Diagnostic Breast Cancer Dataset (WDBC)

- **Source**: UCI ML Repository / Kaggle
- **File**: data.csv
- **Rows / Columns**: 569 rows, 33 columns

Target Variable:

- **Diagnosis (Categorical)**
  - M = Malignant

    B = Benign
- **Purpose**:
  - Binary classification to predict if a tumor is malignant or benign.

Features:

- 30 numeric tumor measurements: **Radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry , fractal dimension.**
- Each of the 10 attributes is recorded in all three groups, giving 30 total numeric features
- Measurement Groups: **Mean, Standard Error, Worst values**
- One empty column: **Unnamed: 32**

Data Dictionary

Column	Description	Feature Type	Valid Values/Range	Notes/Issues
id	Unique patient/sample identifier	Identifier	8670-911320502 integers	Variable length, inconsistent formatting
diagnosis	Tumor classification	Binary Categorical	M (malignant), B (benign)	Target variable
radius_mean	Mean distance from center to nucleus perimeter	Continuous	6.98 - 28.11	Larger = more likely malignant
texture_mean	Mean gray-scale variation in nucleus	Continuous	9.71-39.28	Higher = more irregular surface
perimeter_mean	Mean nucleus perimeter	Continuous	43.79 - 188.50	Correlated with radius
area_mean	Mean area of cell nucleus	Continuous	143.50 - 2501.00	Malignant cells typically larger
smoothness_mean	Mean variation in radius lengths	Continuous	0.05 - 0.16	Lower values indicate smoother borders, higher values indicate irregular borders
compactness_mean	Mean nucleus compactness	Continuous	0.02 - 0.35	0 = perfect circle; higher = more irregular
concavity_mean	Mean severity of concave contour portions	Continuous	0.00 - 0.43	Higher = more indentations
concave points_mean	Mean number of concave contour points	Continuous	0.00 - 0.20	Malignant tumors have more concave points
symmetry_mean	Mean nucleus symmetry	Continuous	0.11 - 0.30	Lower = more symmetric = likely benign
fractal_dimension_mean	Mean boundary complexity	Continuous	0.05 - 0.10	Higher = more irregular border

Analysis of 10-Point Inspection

What did the 10-Point Inspection reveal? Any data quality surprises?

Tumor Size Categories

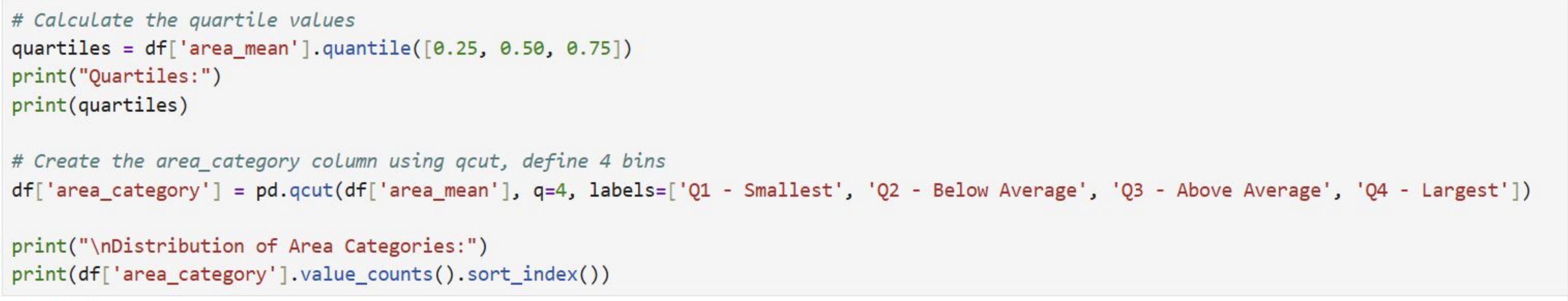
```
##Step 1, create the tumor size category column

df['tumor_size_category'] = pd.cut(
    df['radius_mean'],
    bins=[0, 10, 12, 15, 20, float('inf')],
    labels=['Very Small', 'Small', 'Medium',
'Large', 'Very Large'],
    include_lowest=True
)

print("Tumor Size Category Counts:")
print(df['tumor_size_category'].value_counts().sort
_index())
```

```
Tumor Size Category Counts:
tumor_size_category
Very Small     47
Small         124
Medium        225
Large         128
Very Large     45
Name: count, dtype: int64
```

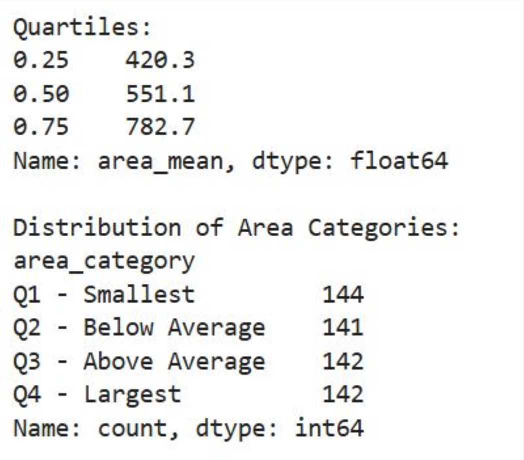
```
 size_analysis =
df.groupby('tumor_size_category')['diagnosis'].value_counts(
).unstack(fill_value=0)
df.groupby('tumor_size_category')['diagnosis']


size_analysis['Total'] = size_analysis.sum(axis=1)
size_analysis['Malignant_Rate_%'] = (size_analysis['M'] /
size_analysis['Total'] * 100).round(2)

print(size_analysis)
```

```
diagnosis              B    M  Total  Malignant_Rate_%
tumor_size_category
Very Small            47    0     47              0.00
Small                118    6    124              4.84
Medium               180   45    225             20.00
Large                 12  116    128             90.62
 Very Large             0   45     45            100.00
```

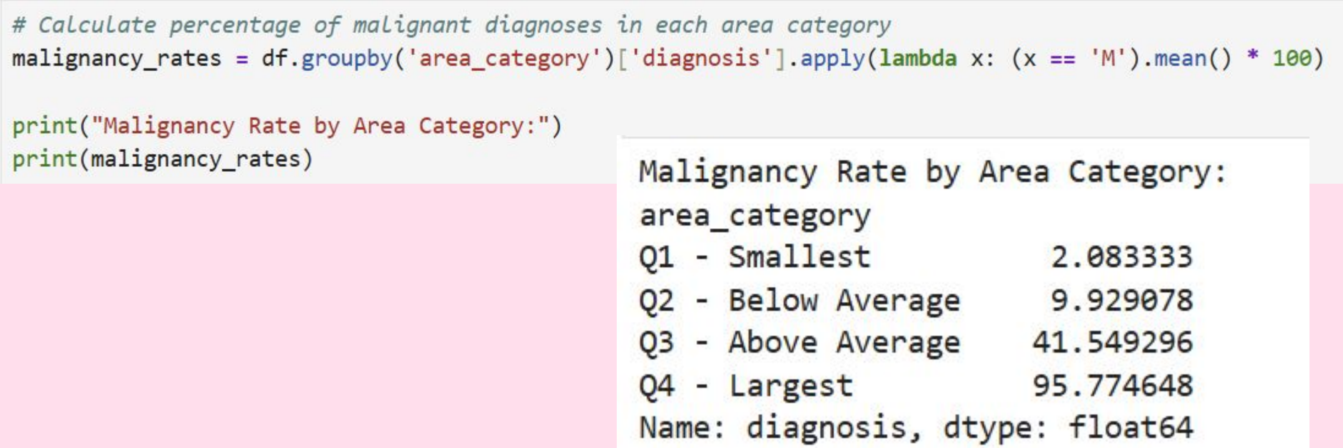
Tumor Area Categories



```
# Calculate the quartile values
quartiles = df['area_mean'].quantile([0.25, 0.50, 0.75])
print("Quartiles:")
print(quartiles)

# Create the area_category column using qcut, define 4 bins
df['area_category'] = pd.qcut(df['area_mean'], q=4, labels=['Q1 - Smallest', 'Q2 - Below Average', 'Q3 - Above Average', 'Q4 - Largest'])

print("\nDistribution of Area Categories:")
print(df['area_category'].value_counts().sort_index())
```



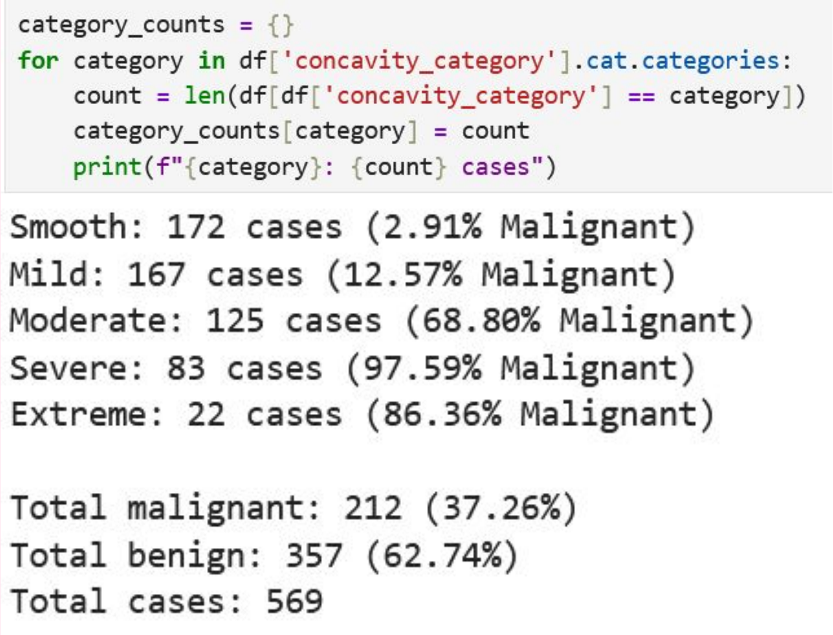
```
Quartiles:
0.25    420.3
0.50    551.1
0.75    782.7
Name: area_mean, dtype: float64

Distribution of Area Categories:
area_category
Q1 - Smallest         144
Q2 - Below Average    141
Q3 - Above Average    142
Q4 - Largest          142
Name: count, dtype: int64
```



```
# Calculate percentage of malignant diagnoses in each area category
malignancy_rates = df.groupby('area_category')['diagnosis'].apply(lambda x: (x == 'M').mean() * 100)

print("Malignancy Rate by Area Category:")
print(malignancy_rates)
```

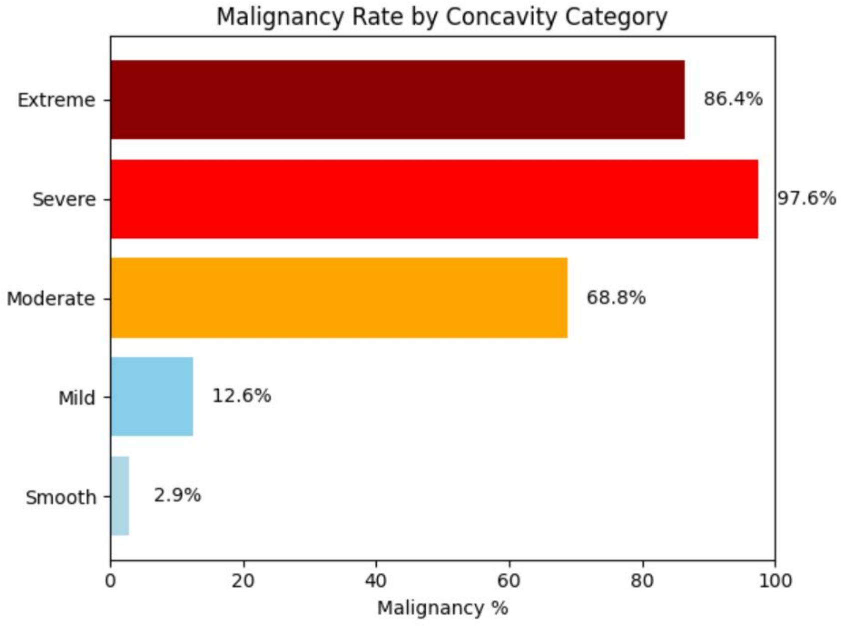
```
Malignancy Rate by Area Category:
area_category
Q1 - Smallest          2.083333
Q2 - Below Average     9.929078
Q3 - Above Average    41.549296
Q4 - Largest          95.774648
Name: diagnosis, dtype: float64
```

Cell Irregularity Categories



```
category_counts = {}
for category in df['concavity_category'].cat.categories:
    count = len(df[df['concavity_category'] == category])
    category_counts[category] = count
    print(f"{category}: {count} cases")
```

```
Smooth: 172 cases (2.91% Malignant)
Mild: 167 cases (12.57% Malignant)
Moderate: 125 cases (68.80% Malignant)
Severe: 83 cases (97.59% Malignant)
Extreme: 22 cases (86.36% Malignant)

Total malignant: 212 (37.26%)
Total benign: 357 (62.74%)
Total cases: 569
```



Texture Variability Categories

- Tumors in each texture category

Low	91
Below Average	215
Above Average	175
High	65
Very High	23

- Malignancy Rate

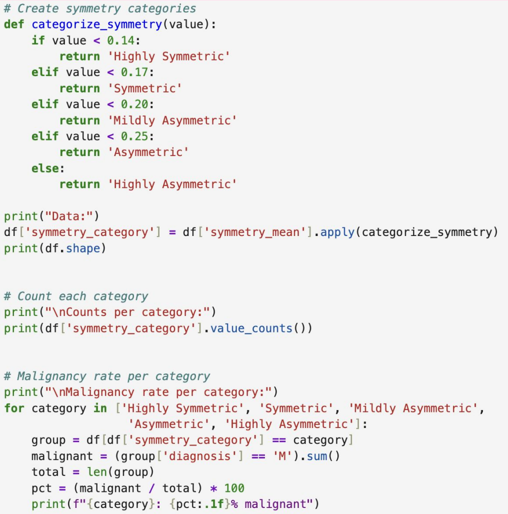
Low	4.4%
Below Average	22.3%
Above Average	60.6%
High	66.2%
Very High	47.8%

- The mean value for texture_mean was 19.29 and the standard deviation was 4.30
- Texture categories were made based on mean and standard deviation
- Direct relationship between higher texture variability and malignancy

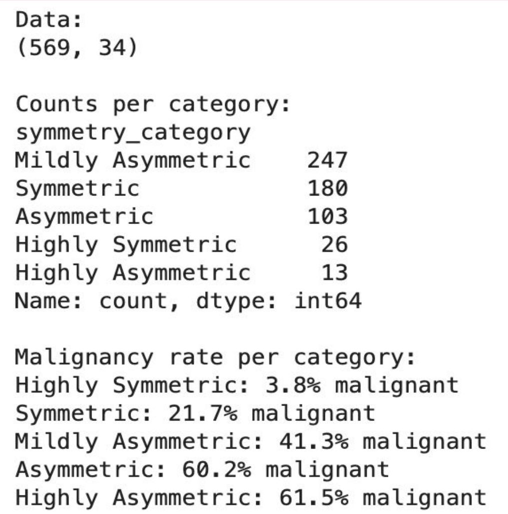
Symmetry-Based Categories

- Asymmetric Cells = Higher Malignancy Risk:
  - Cell division, chromosomal instability, unequal cell division, gene mutations all affect nuclear shape

- Most tumors (43.4%) fall in Mildly Asymmetric
- Malignancy rates rise as asymmetry increases
  - 3.8% -> 61.5%
- Symmetry is a useful diagnostic feature but cannot be used alone



```python
# Create symmetry categories
def categorize_symmetry(value):
    if value < 0.14:
        return 'Highly Symmetric'
    elif value < 0.17:
        return 'Symmetric'
    elif value < 0.20:
        return 'Mildly Asymmetric'
    elif value < 0.25:
        return 'Asymmetric'
    else:
        return 'Highly Asymmetric'

print("Data:")
df['symmetry_category'] = df['symmetry_mean'].apply(categorize_symmetry)
print(df.shape)


# Count each category
print("\nCounts per category:")
print(df['symmetry_category'].value_counts())


# Malignancy rate per category
print("\nMalignancy rate per category:")
for category in ['Highly Symmetric', 'Symmetric', 'Mildly Asymmetric',
                 'Asymmetric', 'Highly Asymmetric']:
    group = df[df['symmetry_category'] == category]
    malignant = (group['diagnosis'] == 'M').sum()
    total = len(group)
    pct = (malignant / total) * 100
    print(f"{category}: {pct:.1f}% malignant")
```



```
Data:
(569, 34)

Counts per category:
symmetry_category
Mildly Asymmetric    247
Symmetric            180
Asymmetric           103
Highly Symmetric      26
Highly Asymmetric     13
Name: count, dtype: int64

Malignancy rate per category:
Highly Symmetric: 3.8% malignant
Symmetric: 21.7% malignant
Mildly Asymmetric: 41.3% malignant
Asymmetric: 60.2% malignant
Highly Asymmetric: 61.5% malignant
```

Observations

- Which features best predict malignancy? How do they relate to each other?
  - Very large tumor size = 100% malignancy rate
  - High texture variability = 66.62% malignancy rate
    - Malignant tumors had a higher average texture mean
  - Highly asymmetric = 61.5% malignancy rate
  - Large area category = 95.77% malignancy rate

- Size and area are the best indicators of tumor malignancy
- Size and area are directly proportional

Recommendations & Conclusion

Key Finding:

- Tumor **size** and **area** are the most accurate and strongest predictors of malignancy.
  - Tumors in the **very large** category were **always** malignant
  - Tumors in the **large** category were malignant in more than 95% of cases

What features should the CAD system prioritize? What are the limitations?

- CAD system should prioritize tumor size and area
  - The sample data indicated that those features were the best indicators of tumor malignancy
- This dataset is limited to non-temporal data