```
# Load the data file
# Reading the data file to see its content
# Reading the names file to see its content
# Displaying the first few lines of each file to understand the structure
```

Result

```
('842302,M,17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.07871,1.095,0.9053,8.589,153.4,0.006399,0.04904,0.05
373,0.01587,0.03003,0.006193,25.38,17.33,184.6,2019,0.1622,0.6656,0.7119,0.2654,0.4601,0.1189\n842517,M,20.57,17.77,132.9,
1326,0.08474,0.07864,0.0869,0.07017,0.1812,0.05667,0.5435,0.7339,3.398,74.08,0.005225,0.01308,0.0186,0.0134,0.01389,0.0035
32,24.99,23.41,158.8,1956,0.1238,0.1866,0.2416,0.186,0.275,0.08902\n84300903,M,19.69,21.25,130,1203,0.1096,0.1599,0.1974,0
.1279,0.2069,0.05999,0.7456,0.7869,4.585,94.03,0.00615,0.04006,0.03832,0.02058,0.0225,0.004571,23.57,25.53,152.5,1709,0.144
4,0.4245,0.4504,0.243,0.3613,0.08758\n84348301,M,11.42,20.38,77.58,386.1,0.1425,0.2839,0.2414,0.1052,0.2597,0.09744,0.4956
,1.156,3.445,27.23,0.00911,0.07458,0.05661,0.01867,0.05963,0.009208,14.91,26.5,98.87,567.7,0.2098,0.8663,0.6869,0.2575,0.663
8,0.173\n84358402,M,20.29,14.34,135.1,1297,0.1003,0.1328,0.198,0.1043,0.1809,0.05883,0.7572,0.7813,5.438,94.44,0.01149,0.0
2461,0.05688,0.01885,0.017', '1. Title: Wisconsin Diagnostic Breast Cancer (WDBC)\n\n2. Source Information\n\na) Creators:
\n\n\tDr. William H. Wolberg, General Surgery Dept., University of\n\tWisconsin, Clinical Sciences Center, Madison, WI
53792\n\twolberg@eagle.surgery.wisc.edu\n\n\tW. Nick Street, Computer Sciences Dept., University of\n\tWisconsin, 1210 West
Dayton St., Madison, WI 53706\n\tstreet@cs.wisc.edu 608-262-6619\n\n\tOlvi L. Mangasarian, Computer Sciences Dept.,
University of\n\tWisconsin, 1210 West Dayton St., Madison, WI 53706\n\tolvi@cs.wisc.edu \n\nb) Donor: Nick Street\n\nc) Date:
November 1995\n\n3. Past Usage:\n\nfirst usage:\n\n\tW.N. Street, W.H. Wolberg and O.L. Mangasarian \n\tNuclear feature
extraction for breast tumor diagnosis.\n\tIS&T/SPIE 1993 International Symposium on Electronic Imaging: Science\n\tand
Technology, volume 1905, pages 861-870, San Jose, CA, 1993.\n\nOR literature:\n\n\tO.L. Mangasarian, W.N. Street and W.H.
Wolberg. \n\tBreast cancer diagnosis and prognosis via linear programming. \n\tOperations Research, 43(4), pages 570-577, July-
August 1995.\n\nMedical literature:\n\n\tW.H. Wolberg, W.N. Street, and O.L. Mangasarian. \n\tMachine learning techniques to
diagnose breast cancer from\n\tfine-needle aspirates. \n\tCancer Letters 77 (1994) 163-171.\n\n\tW.H. Wolberg, W.N. Street, and
O.L. Mangasarian. \n\tImage analysis and machine learning applied to breast cancer\n\tdiagnosis and prognosis. \n\tAnalytical and
Quantitative Cytology and Histology, Vol. 17\n\tNo. 2, pages 77-87, April 1995. \n\n\tW.H. Wolberg, W.N. Street, D.M. Heisey, and
O.L. Mangasarian. \n\computerised breast cancer diagnosis and prognosis from fine\n\tneedle aspirates. \n\tArchives of Surgery
1995;130:511-516.\n\n\tW.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. \n\tComputer-derived nuclear features
distinguish malignant from\n\tbenign breast cytology. \n\tHuman Pathology, 26:792--796, 1995.\n\nSee
also:\n\thttp://www.cs.wisc.edu/~olvi/uwmp/mpml.html\n\thttp://www.cs.wisc.edu/~olvi/uwmp/cancer.html\n\nResults:\n\
n\t- predicting field 2, diagnosis: B = benign, M = malignant\n\t- sets are linearly separable using all 30 input features\n\t- best
predictive accuracy obtained using one separating plane\n\t\tin the 3-D space of Worst Area, Worst Smoothness and\n\t\tMean
Texture. Estimated accuracy 97.5% using repeated\n\t\t10-fold crossvalidations. Classifier has correctly\n\t\tdiagnosed 176
consecutive new patients as of November\n\t\t1995. \n\n4. Relevant information\n\n\tFeatures are computed from a digitised
image of a fine needle\n\taspirate (FNA) of a breast mass. They describe\n\tcharacteristics of the cell nuclei present in the
image.\n\tA few of the images can be found at\n\thttp://www.cs.wisc.edu/~street/images/\n\n\tSeparating plane described
above was obtained using\n\tMultisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree\n\tConstruction Via Linear
Programming." Proceedings of the 4th\n\tMidwest Artificial Intelligence and Cognitive Science Society,\n\tpp. 97-101, 1992], a
classification method which uses linear\n\tprogramming to construct a decision tree. Relevant features\n\twere selected using an
exhaustive search in the space of 1-4\n\tfeatures and 1-3 separating planes.\n\n\tThe actual linear program used to obtain the
separating plane\n\tin the 3-dimensional space is that described in:\n\t[K. P. Bennett and O. L. Mangasarian: "Robust
Linear\n\tProgramming Discrimination of Two Linearly Inseparable Sets",\n\optimisation Methods and Software 1, 1992, 23-
34].\n\n\n\tThis database is also available through the UW CS ftp server:\n\n\tftp ftp.cs.wisc.edu\n\tcd math-prog/cpo-
dataset/machine-learn/WDBC/\n\n5. Number of instances: 569 \n\n6. Number of attributes: 32 (ID, diagnosis, 30 real-valued
input features)\n\n7. Attribute information\n\n1) ID number\n2) Diagnosis (M = malignant, B = benign)\n3-32)\n\nTen real-
valued features are computed for each cell nucleus:\n\n\ta) radius (mean of distances from center to points on the
perimeter)\n\tb) texture (standard deviation of gray-scale values)\n\tc) perimeter\n\td) area\n\te) smoothness (local variation in
radius lengths)\n\tf) compactness (perimeter^2 / area - 1.0)\n\tg) concavity (severity of concave portions of the contour)\n\th)
concave points (number of concave portions of the contour)\n\ti) symmetry \n\tj) fractal dimension ("coastline approximation" -
1)\n\nSeveral of the papers listed above contain detailed descriptions of\nhow these features are computed. \n\nThe mean,
standard error, and "worst" or largest (mean of the three\nlargest values) of these features were computed for each
image,\nresulting in 30 features. For instance, field 3 is Mean Radius, field\n13 is Radius SE, field 23 is Worst Radius.\n\nAll feature
values are recoded with four significant digits.\n\n8. Missing attribute values: none\n\n9. Class distribution: 357 benign, 212
malignant')
```

The dataset provided contains 569 instances with 32 attributes, including:

1. **ID number**
2. **Diagnosis** (M = malignant, B = benign)
3. **30 real-valued features** describing various characteristics of the cell nuclei in digitised images of fine needle aspirate (FNA) of breast masses. These features are:
   - Radius (mean, standard error, and worst)
   - Texture (mean, standard error, and worst)
   - Perimeter (mean, standard error, and worst)

  - Area (mean, standard error, and worst)
  - Smoothness (mean, standard error, and worst)
  - Compactness (mean, standard error, and worst)
  - Concavity (mean, standard error, and worst)
  - Concave points (mean, standard error, and worst)
  - Symmetry (mean, standard error, and worst)
  - Fractal dimension (mean, standard error, and worst)

Following steps for statistical analysis:

1. Load the data into a Pandas DataFrame.
2. Clean and preprocess the data if necessary.
3. Perform descriptive statistics and visualisations.
4. Conduct inferential statistics to understand relationships and differences between benign and malignant tumors.

```
# Define column names based on the information provided
# Load the dataset
# Display the first few rows of the data frame
```

Result

```
         ID Diagnosis  mean_radius  mean_texture  mean_perimeter  mean_area
\
0    842302         M        17.99         10.38          122.80     1001.0
1    842517         M        20.57         17.77          132.90     1326.0
2  84300903         M        19.69         21.25          130.00     1203.0
3  84348301         M        11.42         20.38           77.58      386.1
4  84358402         M        20.29         14.34          135.10     1297.0

   mean_smoothness  mean_compactness  mean_concavity  mean_concave_points
\
0          0.11840           0.27760          0.3001              0.14710
1          0.08474           0.07864          0.0869              0.07017
2          0.10960           0.15990          0.1974              0.12790
3          0.14250           0.28390          0.2414              0.10520
4          0.10030           0.13280          0.1980              0.10430

   ...  worst_radius  worst_texture  worst_perimeter  worst_area  \
0  ...         25.38          17.33           184.60      2019.0
1  ...         24.99          23.41           158.80      1956.0
2  ...         23.57          25.53           152.50      1709.0
3  ...         14.91          26.50            98.87       567.7
4  ...         22.54          16.67           152.20      1575.0

   worst_smoothness  worst_compactness  worst_concavity
worst_concave_points  \
0            0.1622             0.6656           0.7119
0.2654
1            0.1238             0.1866           0.2416
0.1860
2            0.1444             0.4245           0.4504
0.2430
3            0.2098             0.8663           0.6869
0.2575
```

```
4            0.1374             0.2050           0.4000
0.1625

   worst_symmetry  worst_fractal_dimension
0          0.4601                  0.11890
1          0.2750                  0.08902
2          0.3613                  0.08758
3          0.6638                  0.17300
4          0.2364                  0.07678

[5 rows x 32 columns]
```

Start by looking at the summary statistics, checking for missing values, and visualising some of the features.

Steps:

1. **Summary statistics**: Get an overview of the mean, median, standard deviation, etc., of the features.
2. **Class distribution**: Check the distribution of benign and malignant cases.
3. **Correlation matrix**: Identify relationships between the features.
4. **Visualisations**: Create some visual plots like histograms, box plots, and scatter plots to visualise the data.

```
# Summary statistics of the dataset
# Distribution of the 'Diagnosis' variable
```

Result

```
(                 ID  mean_radius  mean_texture  mean_perimeter
mean_area  \
 count  5.690000e+02   569.000000    569.000000      569.000000
569.000000
 mean   3.037183e+07    14.127292     19.289649       91.969033
654.889104
 std    1.250206e+08     3.524049      4.301036       24.298981
351.914129
 min    8.670000e+03     6.981000      9.710000       43.790000
143.500000
 25%    8.692180e+05    11.700000     16.170000       75.170000
420.300000
 50%    9.060240e+05    13.370000     18.840000       86.240000
551.100000
 75%    8.813129e+06    15.780000     21.800000      104.100000
782.700000
 max    9.113205e+08    28.110000     39.280000      188.500000
2501.000000

        mean_smoothness  mean_compactness  mean_concavity
mean_concave_points  \
 count       569.000000        569.000000      569.000000
569.000000
 mean          0.096360          0.104341        0.088799
0.048919
```

```
 std          0.014064          0.052813        0.079720
0.038803
 min          0.052630          0.019380        0.000000
0.000000
 25%          0.086370          0.064920        0.029560
0.020310
 50%          0.095870          0.092630        0.061540
0.033500
 75%          0.105300          0.130400        0.130700
0.074000
 max          0.163400          0.345400        0.426800
0.201200

        mean_symmetry  ...  worst_radius  worst_texture  worst_perimeter  \
 count     569.000000  ...    569.000000     569.000000       569.000000
 mean        0.181162  ...     16.269190      25.677223       107.261213
 std         0.027414  ...      4.833242       6.146258        33.602542
 min         0.106000  ...      7.930000      12.020000        50.410000
 25%         0.161900  ...     13.010000      21.080000        84.110000
 50%         0.179200  ...     14.970000      25.410000        97.660000
 75%         0.195700  ...     18.790000      29.720000       125.400000
 max         0.304000  ...     36.040000      49.540000       251.200000

         worst_area  worst_smoothness  worst_compactness  worst_concavity
\
 count   569.000000        569.000000         569.000000       569.000000
 mean    880.583128          0.132369           0.254265         0.272188
 std     569.356993          0.022832           0.157336         0.208624
 min     185.200000          0.071170           0.027290         0.000000
 25%     515.300000          0.116600           0.147200         0.114500
 50%     686.500000          0.131300           0.211900         0.226700
 75%    1084.000000          0.146000           0.339100         0.382900
 max    4254.000000          0.222600           1.058000         1.252000

        worst_concave_points  worst_symmetry  worst_fractal_dimension
 count            569.000000      569.000000               569.000000
 mean               0.114606        0.290076                 0.083946
 std                0.065732        0.061867                 0.018061
 min                0.000000        0.156500                 0.055040
 25%                0.064930        0.250400                 0.071460
 50%                0.099930        0.282200                 0.080040
 75%                0.161400        0.317900                 0.092080
 max                0.291000        0.663800                 0.207500

 [8 rows x 31 columns],
 B    357
 M    212
 Name: Diagnosis, dtype: int64)
```

The dataset contains a variety of features with different ranges and distributions. Here's a brief overview of the summary statistics:

- The mean values for features vary significantly, for example, the mean radius is around 14.13, whereas the mean area is approximately 654.89.
- Standard deviations indicate a considerable spread in values for some features, such as the area (351.91) and worst area (569.36).
- Minimum and maximum values show a wide range for all features, indicating diverse characteristics of the cell nuclei.

Class Distribution:

The class distribution is:

- Benign (B): 357 instances
- Malignant (M): 212 instances

This indicates that the dataset is somewhat imbalanced, with more benign cases than malignant ones.

Next Steps:

1. **Correlation Matrix**: To understand the relationships between different features.
2. **Visualisations**: Generate histograms, box plots, and scatter plots to visualise the data distribution and potential relationships.

```
# Define column names based on the information provided
# Load the dataset
# Splitting the dataset into benign and malignant
# Question 1: Average radius of cell nuclei for malignant vs benign tumors
# Question 2: Ranges (min-max) of the mean area for both classes
# Question 3: Percentage of benign and malignant cases
# Question 4: Features with the highest correlation with the diagnosis
    -   # Encode Diagnosis for correlation
# Question 5: Highly correlated features indicating redundancy
# Question 6: Histograms of the mean radius for benign and malignant tumors
# Question 7: Box plots of mean texture for benign and malignant tumors
# Question 8: Statistical significance of the mean perimeter between benign and malignant tumors
# Compiling the results
```

Result

```
{'Average Radius': {'Benign': 12.146523809523808,
  'Malignant': 17.462830188679245},
 'Range of Mean Area': {'Benign': (143.5, 992.1),
  'Malignant': (361.6, 2501.0)},
 'Class Percentage': {'Benign': 62.741652021089635,
  'Malignant': 37.258347978910365},
 'Top Correlated Features with Diagnosis': Diagnosis               1.000000
 worst_concave_points    0.793566
 worst_perimeter         0.782914
 mean_concave_points     0.776614
 worst_radius            0.776454
 mean_perimeter          0.742636
 worst_area              0.733825
 mean_radius             0.730029
 mean_area               0.708984
 mean_concavity          0.696360
 Name: Diagnosis, dtype: float64,
 'Highly Correlated Feature Pairs': [('mean_perimeter',
   'mean_radius',
   0.9978552814938109),
  ('mean_area', 'mean_radius', 0.9873571700566127),
  ('mean_concavity', 'mean_compactness', 0.8831206701772507),
  ('mean_concave_points', 'mean_concavity', 0.9213910263788588),
  ('se_perimeter', 'se_radius', 0.9727936770160764),
  ('se_area', 'se_radius', 0.951830112110991),
  ('se_concavity', 'se_compactness', 0.8012683427383291),
  ('se_concave_points', 'se_concavity', 0.7718039949238469),
  ('se_fractal_dimension', 'se_compactness', 0.8032688176617679),
  ('worst_radius', 'mean_radius', 0.9695389726112055),
  ('worst_texture', 'mean_texture', 0.912044588840421),
  ('worst_perimeter', 'worst_radius', 0.993707916102949),
  ('worst_area', 'worst_radius', 0.9840145644590742),
  ('worst_smoothness', 'mean_smoothness', 0.8053241954943627),
  ('worst_compactness', 'mean_compactness', 0.8658090398022629),
  ('worst_concavity', 'worst_compactness', 0.8922608987764685),
  ('worst_concave_points', 'mean_concave_points', 0.9101553142985918),
  ('worst_fractal_dimension', 'worst_compactness', 0.8104548563116117)],
 'T-test of Mean Perimeter': {'t-statistic': -22.93531376570585,
  'p-value': 1.0231409970104293e-66}}
```