Statistic Theory Final

Presenting: Inbar Fabian, Emily Bederov & Shira Lavi

Abstract:

Our research question: How does the characteristics of the cell nuclei of a breast mass affects its diagnosis (Malignant\Benign)?

In this paper, we used several statistical tools to explore our dataset and to reach an answer to our question. In addition, we expanded our research to another datasets and created a function which determines for us the diagnosis based on our characteristics.

Our main dataset was taken from Kaggle:

https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

We used an additional datasets:

https://www.kaggle.com/datasets/zgrcemta/world-gdpgdp-gdp-per-capita-and-annual-growths (gdp dataset)

https://www.kaggle.com/datasets/antimoni/cancer-deaths-by-country-and-type-1990-2016 (breast cancer 2nd dataset)

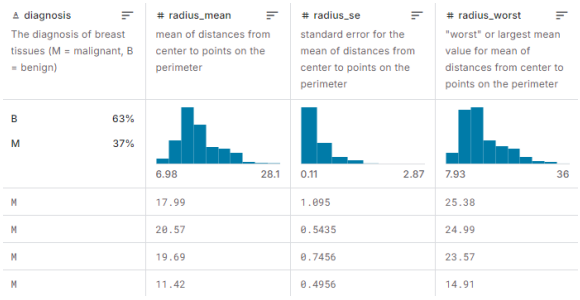
Results: we can determine the diagnosis based mostly on the features:

- Radius_mean
- symmetry_worst
- fractal_dimantion_worst

In conclusion, the most important feature to determine the diagnosis is “radius_mean”.

Introduction:

Let’s start by reviewing the data, we have a table with 31 columns (if we remove the “id”). The columns include the diagnosis(M\B), and 10 more different characteristics of the mass, when each is divided into 3 subcategories: mean, se (standard error), worst.



diagnosis The diagnosis of breast tissues (M = malignant, B = benign)	# radius_mean mean of distances from center to points on the perimeter	# radius_se standard error for the mean of distances from center to points on the perimeter	# radius_worst "worst" or largest mean value for mean of distances from center to points on the perimeter
B 63% M 37%	6.98 28.1	0.11 2.87	7.93 36
M	17.99	1.095	25.38
M	20.57	0.5435	24.99
M	19.69	0.7456	23.57
M	11.42	0.4956	14.91

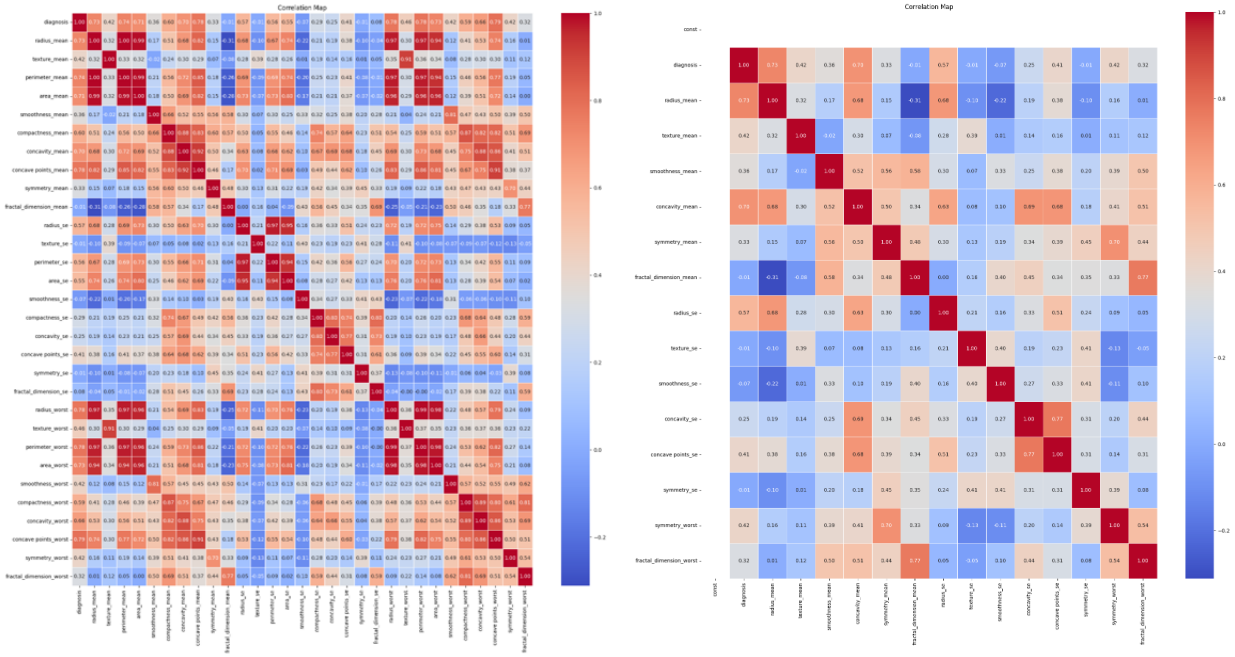
Our relatively large number of features comes with a problem, because the measurements are highly corelated, we noticed there are a lot of dependencies. Those dependencies can cause problems when picking the features which affect the diagnosis the most. This is because two highly correlated features will split their effect on the diagnosis in the measurement therefore reduce its significance.

To deal with this issue we will use two methods:

First, divide the features by correlation. We will make groups of features which are correlated to one another higher than 0.8. these are the groups:

> ['diagnosis'] , ['radius_mean', 'perimeter_mean', 'area_mean', 'concave points_mean', 'radius_worst', 'perimeter_worst', 'area_worst'], ['texture_mean', 'texture_worst'], ['smoothness_mean', 'smoothness_worst'], ['compactness_mean', 'concavity_mean', 'compactness_worst', 'concavity_worst', 'concave points_worst'], ['symmetry_mean'], ['fractal_dimension_mean'], ['radius_se', 'perimeter_se', 'area_se'], ['texture_se'], ['smoothness_se'], ['compactness_se', 'concavity_se', 'fractal_dimension_se'], ['concave points_se'], ['symmetry_se'], ['symmetry_worst'], ['fractal_dimension_worst']

To find the best feature to represent each group we will use ‘VIF,’ for each group we choose the feature with the highest VIF score. meaning, the feature who correlates best to the other features in the group.
Next, we changed our dataset to include only the heads of each group and now we will try to find the affect they have on the diagnosis.
old VS new correlation matrix:

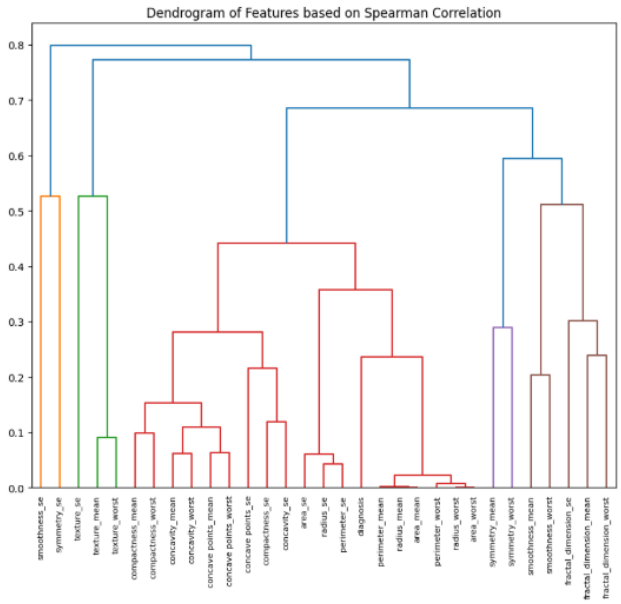


To make sure we are getting the best results, we chose another way to split the data into groups using **Hierarchical Clustering**.

This method gave us those results:

Meaning, the groups:

['smoothness_se'], ['symmetry_se'], ['texture_mean', 'texture_worst'], ['texture_se'], ['compactness_worst', 'compactness_mean', 'concavity_mean', 'concave points_mean', 'concave points_se', 'concave points_worst', 'compactness_se', 'concavity_se', 'concavity_worst'],
['radius_se', 'perimeter_se', 'area_se'], ['diagnosis', 'area_worst', 'radius_worst', 'perimeter_worst', 'area_mean', 'perimeter_mean', 'radius_mean'],
['symmetry_worst', 'symmetry_mean'], ['smoothness_mean', 'smoothness_worst'], ['fractal_dimension_mean', 'fractal_dimension_worst'], ['fractal_dimension_se']



In this method diagnosis is part of a cluster, meaning the features in this cluster are highly correlated to it, thus they are valuable for its prediction, we will not choose only one representative, we will use them all.

To test which of those methods is better, we Performed cross-validation for VIF-selected features and for Hierarchical Clustering-selected features, then we checked the scores of the cross validation and used Sign Test (which is a statistical test for consistent differences between pairs of observations), and with p value = 0.3 which is very high, then we rejected the null hypothesis.

Results:

We can choose a method to our liking. Therefore, from now on we will use the first feature selection method. Then we will use statistic tests to understand what features can contribute mostly, then we will perform regression.

Statistical Analysis

We performed statistic tests such as Kolmogorov-Smirnov & Shapiro-Wilk to determine whether a feature distribute normally.

- texture_mean
- smoothness_mean
- symmetry_mean

those are the normally distributed features, thus we checked using two-sample T-test, if there is a significant difference between the cancer group (diagnosis == 'M') and the non-cancer group (diagnosis == 'B') for individuals that for them, those features are above the mean.

We conclude that:

1. texture_mean - There is no significant difference in texture values between the cancer and non-cancer groups.
2. smoothness_mean - here is a significant difference in smoothness values between the cancer and non-cancer groups.
3. symmetry_mean - There is a significant difference in symmetry values between the cancer and non-cancer groups.

Which helps learn that smoothness_mean & symmetry_mean are valuable features to help determine type of cancer.
As for the non-normally distributed we firstly used the Goodness-of-fit method, we tried to fit features to distributions – normal, exponential, gamma, lognormal & beta.
Then, to decide if the fit is actually good, for normal fits, we performed Kolmogorov-Smirnov test, to check the credibility of the fit, and as for non-normal distributions we performed Anderson-Darling test (The Anderson–Darling test is a statistical test of whether a given sample of data is drawn from a given probability distribution).
Results:

- concave points_se (Beta distribution): The relatively high p-value indicates that the beta distribution is a good fit for this data.
- symmetry_se (Lognormal distribution): The high p-value indicates that the lognormal distribution is a good fit for this data.

we couldn't fit any feature to normal distribution; hence we will now use non-parametric tests.
Using Mann-Whitney U Test Statistic, for the parameters:

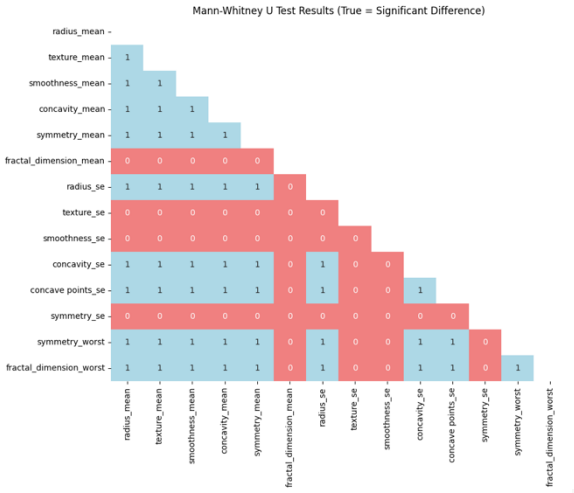
- Radius_mean
- symmetry_se
- symmetry_worst
- fractal_dimantion_worst

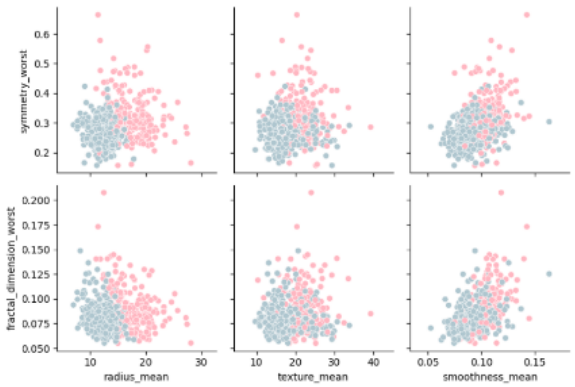
We conclude that:

1. Radius_mean - significant difference
2. symmetry_se – not significant difference
3. symmetry_worst - significant difference
4. fractal_dimention_worst - significant difference

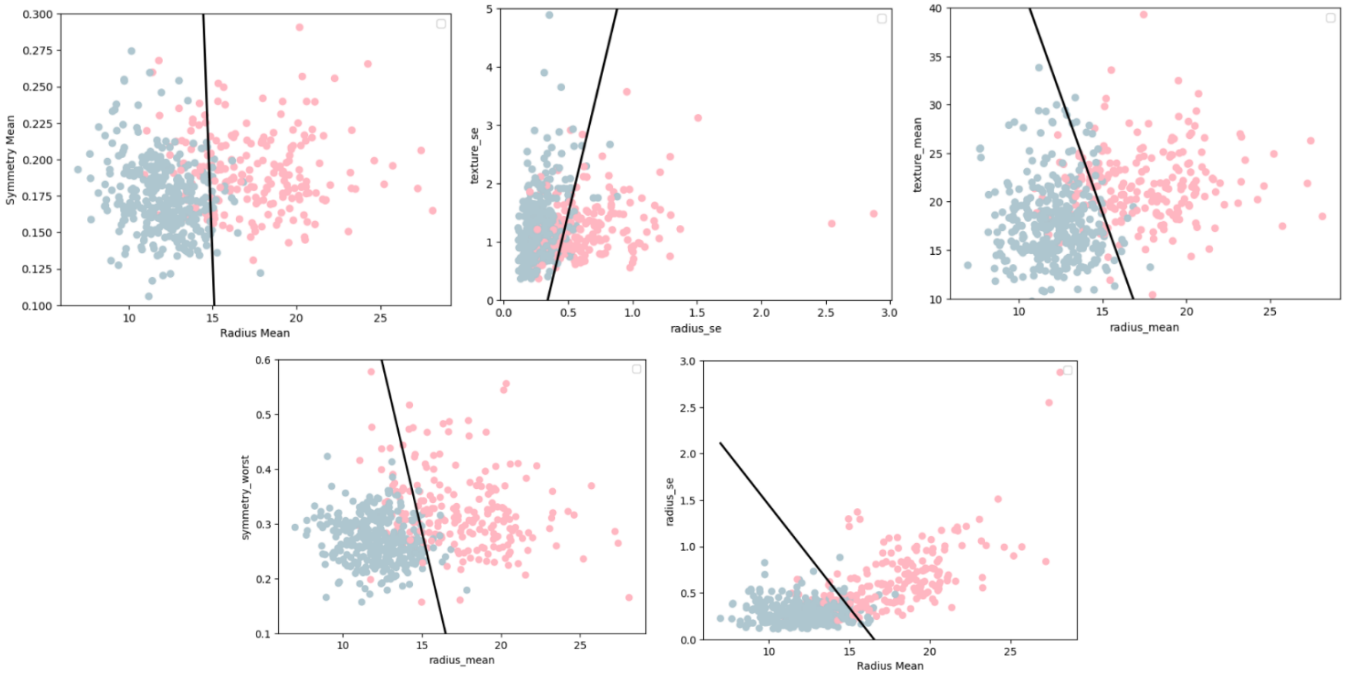
<u>Regression analysis:</u>
To find out how the different representatives interact with each other, we will plot the histograms of the new data-frame when in the X-axis there is one feature, the Y-axis another, and each sample is represented by a dot colored: blue-benign, pink-malignant.
For example:
We can see from these few samples that some pairs mix the dots, but some pairs divide the data to pink and blue very good, up until the point we can almost divide them completely

using a line. Based on the plotting, we used Mann-Whitney U Test on each pair to determine if the mean values of the blue and pink points are far away from each other enough so it will be optimal to separate them by a line.
The results to this test are:





Now we will make for every one of them a SVM linear model, the visual results are:



To make sure we only take the best models we will check for accuracy and keep the ones with over 0.85 accuracy. These are the ones that we are left with:

- SVM model for texture_mean and radius_mean - Accuracy: 0.9064
- SVM model for smoothness_mean and radius_mean - Accuracy: 0.9181
- SVM model for concavity_mean and radius_mean - Accuracy: 0.9240
- SVM model for concavity_mean and texture_mean - Accuracy: 0.8713

- SVM model for concavity_mean and smoothness_mean - Accuracy: 0.8538
- SVM model for symmetry_mean and radius_mean - Accuracy: 0.9181
- SVM model for symmetry_mean and concavity_mean - Accuracy: 0.8538
- SVM model for radius_se and radius_mean - Accuracy: 0.9064
- SVM model for radius_se and concavity_mean - Accuracy: 0.8596
- SVM model for concavity_se and radius_mean - Accuracy: 0.9181
- SVM model for concave points_se and radius_mean - Accuracy: 0.9181
- SVM model for concave points_se and concavity_mean - Accuracy: 0.8538
- SVM model for symmetry_worst and radius_mean - Accuracy: 0.9357
- SVM model for fractal_dimension_worst and radius_mean - Accuracy: 0.9181

Next we will make a model which takes the mean value of all the models above and returns 1 if the prediction is greater than 0.5 and 0 if the prediction is less than 0.5

The final model accuracy came out as 0.918.

Combining Datasets:

Our goal was to find out what is the relationship of the economic status of a country to the number of deaths from breast cancer in relation to the population, in order to check whether the fact that Wisconsin is a small state and not in a dire economic situation increases or decreases the mortality rate from breast cancer in 1992.
The dataset we started with, is a dataset of 27 types of cancer around the world.

	Country	Code	Year	Liver cancer	Kidney cancer	Larynx cancer	Breast cancer	Thyroid cancer	Stomach cancer	Bladder cancer	...	Non-melanoma skin cancer	Lip and oral cavity cancer	Brain and nervous system cancer
0	Afghanistan	AFG	1990	243.663716	39.470495	109.334207	766.535431	79.820167	923.495208	148.139204	...	26.446156	53.599636	163.869062
1	Afghanistan	AFG	1991	261.241824	41.376024	117.311719	823.233932	85.111020	989.709648	156.977412	...	28.275271	57.148890	174.183219
2	Afghanistan	AFG	1992	284.443630	44.106315	128.071634	901.022100	92.240603	1078.459037	168.990462	...	30.718152	61.876100	188.382296
3	Afghanistan	AFG	1993	313.136816	47.424854	141.429604	996.432762	101.206726	1192.064525	184.347737	...	33.835442	67.504857	205.250430
4	Afghanistan	AFG	1994	343.229715	50.710951	155.754606	1097.895223	110.679923	1316.505674	200.246949	...	37.103370	73.175879	222.383572
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5989	Zimbabwe	ZWE	2012	1218.763107	56.966136	162.131298	783.959361	115.203608	936.013607	420.658042	...	212.201798	95.507275	222.048414
5990	Zimbabwe	ZWE	2013	1252.747896	58.735014	161.039807	790.077464	115.846418	928.096553	423.397684	...	213.262823	97.058488	229.271375
5991	Zimbabwe	ZWE	2014	1308.483454	61.575167	161.512234	839.938132	121.251540	949.321368	434.314706	...	215.100202	99.866860	240.975514
5992	Zimbabwe	ZWE	2015	1357.611713	63.757395	162.909299	862.707637	123.638675	956.138239	442.122976	...	218.011570	102.734862	250.586202
5993	Zimbabwe	ZWE	2016	1411.242274	66.764985	164.929997	891.630167	126.800977	969.794783	453.711001	...	222.509943	106.303546	263.215193

5994 rows × 30 columns

from this dataset we only took the cases that are relevant to our main dataset, we filtered the dataset so that it only has a few states in the United States (Canada, Chile, Colombia, Costa Rica, Guatemala, Bahamas, Ecuador, El Salvador, Mexico, Honduras) breast cancer data, and finally we filtered our data to only be from 1992 like the original data from Wisconsin, we checked the number of deaths from breast cancer in Wisconsin in 1992 and added a row of Wisconsin to the data.

	Country	Code	Year	Breast cancer
0	Afghanistan	AFG	1990	766.535431
1	Afghanistan	AFG	1991	823.233932
2	Afghanistan	AFG	1992	901.022100
3	Afghanistan	AFG	1993	996.432762
4	Afghanistan	AFG	1994	1097.895223
...	...	...	...	...
5989	Zimbabwe	ZWE	2012	783.959361
5990	Zimbabwe	ZWE	2013	790.077464
5991	Zimbabwe	ZWE	2014	839.938132
5992	Zimbabwe	ZWE	2015	862.707637
5993	Zimbabwe	ZWE	2016	891.630167

5994 rows × 4 columns

	Country	Code	Year	Breast cancer
378	Bahamas	BHS	1990	26.962018
379	Bahamas	BHS	1991	27.808558
380	Bahamas	BHS	1992	28.811027
381	Bahamas	BHS	1993	30.021995
382	Bahamas	BHS	1994	30.787088
...	...	...	...	...
3370	Mexico	MEX	2012	5998.797400
3371	Mexico	MEX	2013	6335.127044
3372	Mexico	MEX	2014	6610.387303
3373	Mexico	MEX	2015	6695.789585
3374	Mexico	MEX	2016	6848.454072

270 rows × 4 columns

	Country	Code	Year	Breast cancer
0	Bahamas	BHS	1992	28.811027
1	Canada	CAN	1992	4952.458943
2	Chile	CHL	1992	841.069976
3	Colombia	COL	1992	1615.615711
4	Costa Rica	CRI	1992	161.836578
5	Ecuador	ECU	1992	295.522830
6	El Salvador	SLV	1992	164.064214
7	Guatemala	GTM	1992	197.725307
8	Honduras	HND	1992	125.483079
9	Mexico	MEX	1992	2962.225256
10	Wisconsin	WIS	1992	4455.000000

In addition, we crossed another dataset of GDP of countries from the years 1960-2020 to get the GDP of each of the countries in 1992, and we also added the GDP of Wisconsin in 1992

	Country Name	Country Code	1960	1961	1962	1963	1964	1965	1966	1967	...	2011	
0	Africa Eastern and Southern	AFE	1.931311e+10	1.972349e+10	2.149392e+10	2.573321e+10	2.352744e+10	2.681057e+10	2.915216e+10	3.017317e+10	...	9.430000e+11	9.51000
1	Africa Western and Central	AFW	1.040428e+10	1.112805e+10	1.194335e+10	1.267652e+10	1.383858e+10	1.486247e+10	1.583285e+10	1.442643e+10	...	6.710000e+11	7.28000
2	Australia	AUS	1.860679e+10	1.968306e+10	1.992272e+10	2.153993e+10	2.380110e+10	2.597715e+10	2.730989e+10	3.044462e+10	...	1.400000e+12	1.55000
3	Austria	AUT	6.592694e+09	7.311750e+09	7.756110e+09	8.374175e+09	9.169984e+09	9.994071e+09	1.088768e+10	1.157943e+10	...	4.310000e+11	4.09000
4	Burundi	BDI	1.960000e+08	2.030000e+08	2.135000e+08	2.327500e+08	2.607500e+08	1.589950e+08	1.654446e+08	1.782971e+08	...	2.235821e+09	2.33330
...	...	...	...	...	...	...	...	...	...	...	...	...	
115	St. Vincent and the Grenadines	VCT	1.306656e+07	1.399988e+07	1.452488e+07	1.370822e+07	1.475821e+07	1.510821e+07	1.609987e+07	1.583518e+07	...	6.761296e+08	6.92933
116	World	WLD	1.390000e+12	1.440000e+12	1.550000e+12	1.670000e+12	1.820000e+12	1.990000e+12	2.160000e+12	2.290000e+12	...	7.370000e+13	7.53000

In addition, we created a column with the amount of population that was in each of the countries in 1992, To check the percentage of deaths from breast cancer in the population. We also added a column of the relationship between them.

	Country Name	Code	1992
0	Bahamas, The	BHS	3.109000e+09
1	Canada	CAN	5.923877e+11
2	Chile	CHL	4.596433e+10
3	Colombia	COL	5.841899e+10
4	Costa Rica	CRI	8.564044e+09
5	Ecuador	ECU	1.809424e+10
6	Guatemala	GTM	1.044084e+10
7	Honduras	HND	4.943700e+09
8	Mexico	MEX	3.631576e+11
9	El Salvador	SLV	5.813399e+09
10	Wisconsin	WIS	1.260000e+11

	Country Name	Code	1992	Population	proportion
0	Bahamas, The	BHS	3.109000e+09	281973	0.000102
1	Canada	CAN	5.923877e+11	28370000	0.000175
2	Chile	CHL	4.596433e+10	13782297	0.000061
3	Colombia	COL	5.841899e+10	33940000	0.000048
4	Costa Rica	CRI	8.564044e+09	3321939	0.000049
5	Ecuador	ECU	1.809424e+10	10910000	0.000027
6	Guatemala	GTM	1.044084e+10	9544000	0.000017
7	Honduras	HND	4.943700e+09	5345000	0.000037
8	Mexico	MEX	3.631576e+11	85990000	0.000001
9	El Salvador	SLV	5.813399e+09	5552000	0.000534
10	Wisconsin	WIS	1.260000e+11	5005000	0.000890

We performed normality tests and discovered that our data is not normally distributed, then, we sorted the GDP into categories (GDP< e^10 - **small**, e^10<GDP< e^11 - **medium**, e^11<GDP - **large**) to create data with a nominal variable and use the chi test.

The results of the test showed that there is no relationship between the GDP index of a state and the percentage of deaths from breast cancer, which shows us that even though Wisconsin was a small and economically undeveloped state in 1992, it did not affect the percentage of deaths from breast cancer.

```
(Proportion Category  Low  Medium
 GDP Category
 Small                  2       2
 Medium                 4       0
 Large                  1       2,
 3.797619047619048,
 0.14974678313156617,
 2,
 array([[2.54545455, 1.45454545],
        [2.54545455, 1.45454545],
        [1.90909091, 1.09090909]]))
```

Methods:

The methods we used throughout our project in order are:

Feature selection using VIF, Group selection using hierarchical clustering, Sign Test for cross validation the feature selection methods, Kolmogorov-Smirnov & Shapiro-Wilk to determine whether a feature distribute normally, T-test to see if there is a significant difference in the features between the cancer group (diagnosis == ‘M’) and the non-cancer group (diagnosis == ‘B’), Goodness-of-fit method to try and relate each feature to some distribution, Anderson-Darling test on the non-normally distributed features, Mann-Whitney U Test for every pair of features to determine if we are able to use SVM, Mann-Whitney U Test for the features most related to diagnosis, SVM models on the pairs of features, using the mean value of the best models to determine for a new sample the diagnosis, Chi test to see that there is no relationship between the GDP index of a state and the percentage of deaths from breast cancer.

Discussion:

To sum up the paper, we concluded that the most important features to determine the diagnosis are ‘radius_mean’, ‘symmetry_worst’, ‘fractal_dimension_worst’. We also made a model that determines the diagnosis based of the features. We want to emphasize that those features are just the representation of their group, so it’s the three group which have the most effect over the diagnosis.