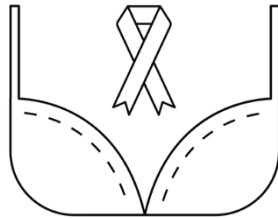


# BREAST CANCER ANALYTICS



Autores:

Matias Trovatto y Val Isetta

Institución:

CoderHouse

Fecha de presentación:

31/10/2022

# Índice

## Contenido

0. Introducción.....	3
1. Tabla de versiones.....	4
2. Herramientas tecnológicas implementadas.....	4
3. Descripción de la temática de los datos.....	4
4. Objetivo de análisis.....	4
5. Data Set.....	5
6. Diagrama entidad-relación.....	8
7. Mención de las tablas.....	8
8. Listado de tablas.....	9
9. Objetivo.....	9
10. Alcance y usuario final.....	10
11. Nivel de aplicación del análisis.....	10
12. Medidas y columnas calculadas y sus fórmulas.....	10
13. Segmentaciones elegidas.....	11
14. Transformaciones realizadas.....	11
15. Visualizaciones PowerBi.....	11
16. Futuras líneas.....	14
17. Drive.....	14
18. Bibliografía.....	14

# Introducción

Siendo octubre el mes de concientización mundial sobre el cáncer de mama, en el presente trabajo se decidió abordar esta temática, con la finalidad de comprender el impacto que tiene esta enfermedad en la población femenina, en que edades predominan los tamaños y grados de diferenciaciones más graves, y la esperanza de vida en relación a este tipo de tumores. dado que consideremos que se trata de un tema de actualidad que afecta a la población mundial y del que se debe contar con datos y tableros de control limpios y exactos para la toma de decisiones.

Todas las personas tienen cierto riesgo de padecer cáncer de mama, el riesgo de una persona de padecer cáncer de mama puede ser mayor o menor, en función de factores específicos de riesgo.

Muchas personas piensan que el cáncer de mama es una enfermedad heredada. Sin embargo, se cree que entre el 5 y el 10 % de los casos de cáncer de mama son hereditarios, lo que significa que son causados por cambios anómalos (o mutaciones) en ciertos genes, que se transmiten de padre o madre a hijos. Gran parte de las personas que tienen cáncer de mama no tienen antecedentes familiares, lo que indica que otros factores pueden entrar en acción, como el entorno y el estilo de vida.

Nosotros nos enfocamos en el comportamiento de la enfermedad, mostrando sus grados y diferenciaciones y sus respectivos impactos en cada una de las pacientes que lo padecen.

También nos enfocamos en analizar la esfera económica, ya que la enfermedad no solo afecta su salud, sino que también impacta su vida en general.

## 1. Tabla de versiones

<u>Versión</u>	<u>Fecha</u>
<u>1</u>	<u>Mié, 24 de agosto de 2022</u>
<u>2</u>	<u>Mié, 14 de septiembre de 2022</u>
<u>3</u>	<u>Mié, 19 de octubre de 2022</u>
<u>4</u>	<u>Lun, 31 de octubre de 2022</u>

## 2. Herramientas tecnológicas implementadas

Para el presente trabajo se utilizaron los siguientes programas:

- Excel para la lectura y limpieza de los datasets.
- ERDplus para la creación del diagrama entidad-relación (<https://erdplus.com/standalone>).
- Power BI Desktop para la creación del tablero de control.

## 3. Descripción de la temática de los datos

Para esta entrega presentamos este conjunto de datos de pacientes con cáncer de mama, que se obtuvo de la actualización de noviembre de 2017 del Programa SEER del NCI, que brinda información sobre estadísticas de cáncer. El conjunto de datos involucró a pacientes mujeres con cáncer de mama de carcinoma lobular y de conducto infiltrante diagnosticado en 2006-2010. Se excluyeron los pacientes con tamaño tumoral desconocido, LN regionales examinados, LN regionales positivos y pacientes cuyos meses de supervivencia fueron inferiores a 1 mes; por lo tanto, finalmente se incluyeron 4005 pacientes.

## 4. Objetivos de análisis

Comprender el impacto que tiene esta enfermedad en la población femenina, en que edades predominan los tamaños y grados de diferenciación más graves, y la esperanza de vida en relación a este tipo de tumores.

## 5. Dataset

A continuación, se adjuntan imágenes de las tablas del set de datos.

Tabla Pacientes

	A	B	C	D	E	F
1	id_paciente	Race	Marital Status	Survival Months	Status	Date of admission
2	1	White	Married	60	Alive	6/5/2004
3	2	White	Married	62	Alive	4/9/2004
4	3	White	Divorced	75	Alive	26/8/2004
5	4	White	Married	84	Alive	11/11/2006
6	5	White	Married	50	Alive	14/10/2001
7	6	White	Single	89	Alive	22/10/2000
8	7	White	Married	54	Alive	17/9/2003
9	8	White	Married	14	Dead	4/4/2003
10	9	White	Divorced	70	Alive	31/5/2010
11	10	White	Married	92	Alive	5/6/2008
12	11	White	Widowed	64	Dead	5/12/2000
13	12	White	Married	92	Alive	2/11/2009
14	13	White	Married	56	Alive	13/4/2001
15	14	White	Married	38	Alive	16/7/2006
16	15	White	Divorced	64	Alive	24/4/2000
17	16	White	Married	49	Alive	27/7/2004
18	17	White	Single	105	Alive	21/9/2010
19	18	White	Married	62	Alive	2/7/2000
20	19	Black	Divorced	107	Alive	3/2/2007
21	20	White	Divorced	77	Alive	23/2/2002
22	21	Other	Married	81	Alive	11/7/2010
23	22	White	Married	50	Alive	1/3/2008
24	23	White	Single	78	Alive	10/9/2000
25	24	White	Married	102	Alive	17/3/2002

◀ ▶
Pacientes
Tratamiento
Edad
Hormonas
Tumor
region
+

Tabla Tratamiento

	A	B	C	D
1	id_paciente	Cost per month	Monthly Incom	months of treatment
2	1	\$ 1.568,00	\$ 8.033,00	12
3	2	\$ 2.451,00	\$ 8.740,00	24
4	3	\$ 480,00	\$ 8.075,00	2
5	4	\$ 2.233,00	\$ 3.928,00	14
6	5	\$ 832,00	\$ 10.338,00	19
7	6	\$ 2.927,00	\$ 11.249,00	15
8	7	\$ 1.392,00	\$ 8.159,00	20
9	8	\$ 1.855,00	\$ 11.696,00	4
10	9	\$ 1.909,00	\$ 10.848,00	17
11	10	\$ 2.254,00	\$ 11.035,00	2
12	11	\$ 1.534,00	\$ 4.578,00	15
13	12	\$ 1.224,00	\$ 11.734,00	17
14	13	\$ 2.557,00	\$ 5.309,00	10
15	14	\$ 1.199,00	\$ 9.473,00	20
16	15	\$ 1.878,00	\$ 11.628,00	4
17	16	\$ 578,00	\$ 6.565,00	2
18	17	\$ 1.703,00	\$ 8.746,00	4
19	18	\$ 1.663,00	\$ 6.737,00	23
20	19	\$ 1.151,00	\$ 2.295,00	14
21	20	\$ 2.580,00	\$ 1.302,00	22
22	21	\$ 2.230,00	\$ 8.803,00	15

◀ ▶
Pacientes
Tratamiento
Edad
Hormonas
Tumor
region
+

Tabla edad

	A	B	C
1	id_paciente ▾	Age ▾	
2	1	68	
3	2	50	
4	3	58	
5	4	58	
6	5	47	
7	6	51	
8	7	51	
9	8	40	
10	9	40	
11	10	69	
12	11	68	
13	12	46	
14	13	65	
15	14	48	
16	15	62	
17	16	61	
18	17	56	
19	18	43	
20	19	48	
21	20	60	

◀ ▶ Pacientes Tratamiento **Edad**

Tabla hormonas

	A	B	C
1	id_paciente ▾	Estrogen Status ▾	Progesterone Status ▾
2	1	Positive	Positive
3	2	Positive	Positive
4	3	Positive	Positive
5	4	Positive	Positive
6	5	Positive	Positive
7	6	Positive	Positive
8	7	Positive	Positive
9	8	Positive	Positive
10	9	Positive	Positive
11	10	Positive	Positive
12	11	Positive	Positive
13	12	Negative	Negative
14	13	Positive	Positive
15	14	Positive	Positive
16	15	Positive	Positive
17	16	Positive	Positive
18	17	Positive	Positive
19	18	Positive	Positive
20	19	Positive	Positive
21	20	Negative	Negative

◀ ▶ Pacientes Tratamiento Edad **Hormonas** Tumor regio

Tabla tumor

	A	B	C	D	E
1	id_paciente ▾	Grade ▾	Discovery Date ▾	Tumor Size ▾	differentiate ▾
2	1	3	6/12/2009	4	Poorly differentiated
3	2	2	24/4/2002	35	Moderately differentiated
4	3	2	25/7/2003	63	Moderately differentiated
5	4	3	10/4/2009	18	Poorly differentiated
6	5	3	7/1/2010	41	Poorly differentiated
7	6	2	7/4/2005	20	Moderately differentiated
8	7	1	7/1/2003	8	Well differentiated
9	8	2	26/1/2009	30	Moderately differentiated
10	9	3	19/7/2003	103	Poorly differentiated
11	10	1	6/4/2002	32	Well differentiated
12	11	2	23/12/2001	13	Moderately differentiated
13	12	3	10/12/2003	59	Poorly differentiated
14	13	3	27/4/2005	35	Poorly differentiated
15	14	3	25/7/2009	15	Poorly differentiated
16	15	2	18/9/2009	35	Moderately differentiated
17	16	2	23/6/2010	19	Moderately differentiated
18	17	2	29/11/2007	46	Moderately differentiated
19	18	2	1/1/2007	24	Moderately differentiated
20	19	2	11/2/2009	25	Moderately differentiated
21	20	2	18/12/2002	29	Moderately differentiated

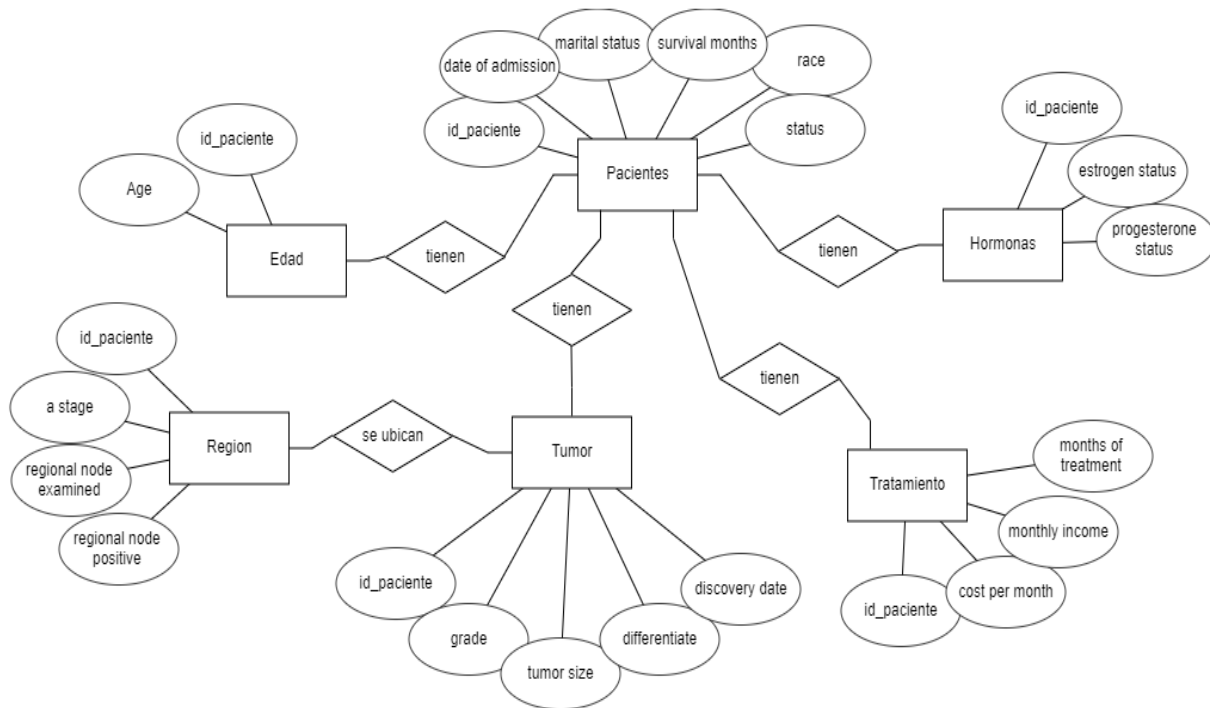
◀ ▶
Pacientes
Tratamiento
Edad
Hormonas
Tumor
region
⊕

Tabla región

1	id_paciente ▾	A Stage ▾	Regional Node Examined ▾	Reginol Node Positive ▾
2	1	Regional	24	1
3	2	Regional	14	5
4	3	Regional	14	7
5	4	Regional	2	1
6	5	Regional	3	1
7	6	Regional	18	2
8	7	Regional	11	1
9	8	Regional	9	1
10	9	Regional	20	18
11	10	Distant	21	12
12	11	Regional	9	1
13	12	Regional	11	3
14	13	Regional	13	3
15	14	Regional	23	7
16	15	Regional	16	14
17	16	Regional	20	1
18	17	Regional	1	1
19	18	Regional	22	1
20	19	Regional	16	1
21	20	Regional	20	1

◀ ▶
Pacientes
Tratamiento
Edad
Hormonas
Tumor
region
⊕

## 6. Diagrama entidad-relación



## 7. Mención de las tablas.

A continuación, se encuentra la mención de las tablas del set de datos trabajado, junto con una pequeña descripción y sus respectivas llaves primarias y foráneas.

- **Pacientes:** Contiene los datos de los pacientes, estado marital, meses sobrevividos y status actual.

PK: id\_paciente

- **Edad:** Contiene las edades de los pacientes.

FK: id\_paciente

- **Hormonas:** Contiene la confirmación de las hormonas de progesterona y estrógenos.

FK: id\_paciente

- **Tumor:** Contiene el grado, tamaño y diferenciación del tumor hallado.

FK: id\_paciente



- **Región:** Contiene la región del nodo examinado

FK: id\_paciente

- **Tratamiento:** Contiene los costos del tratamiento por mes, ingresos mensuales de las pacientes y meses de tratamiento.
- FK: id\_paciente

## 8. Listado de tablas.

A continuación, se encuentra el listado de tablas con sus campos, claves y tipo de datos.

Edad		
Campo	Tipo de campo	Tipo de clave
id_paciente	int	FK
age	int	-

Hormonas		
Campo	Tipo de campo	Tipo de clave
id_paciente	int	FK
Estrogen Status	varchar (250)	-
Progesterone Status	varchar (250)	-

Pacientes		
Campo	Tipo de campo	Tipo de clave
id_paciente	int	PK-index
Race	varchar (250)	-
Marital Status	varchar (250)	-
Survival Months	int	-
Status	varchar (250)	-

Tumor		
Campo	Tipo de campo	Tipo de clave
id_paciente	int	FK
Grade	int	-
Tumor Size	int	-
differentiate	varchar (250)	-

Region		
Campo	Tipo de campo	Tipo de clave
id_paciente	int	FK
A Stage	varchar (250)	-
Regional Node Examined	int	-
Regional Node Positive	int	-

Tratamiento		
Campo	Tipo de campo	Tipo de clave
id_paciente	int	FK
Cost per month	int	-
Monthly Income	int	-
Months of treatment	int	-

## 9. Objetivo:

El objetivo de este análisis es mostrar el comportamiento de la enfermedad, orientado a informar a quienes se encargan de tomar medidas de salud pública, y también a quienes padecen la enfermedad en sí. El análisis toca diferentes matices; una parte, dirigida a la comprensión de la enfermedad, sus grados, y el impacto en los pacientes según, por ejemplo, su edad; por otro lado,

también fue analizada la esfera económica, mostrando los aranceles promedio para costear los tratamientos.

#### **10. Alcance y usuario final:**

El dashboard está orientado tanto al sector público encargado de tomar medidas sobre salud pública y las personas que padezcan la enfermedad y sus familias.

#### **11. Nivel de aplicación del análisis:**

El Proyecto está orientado a un nivel táctico, para que permita tomar medidas económicas para poder ayudar a financiar los tratamientos a las personas que padezcan la enfermedad y no puedan acceder a estos mismo.

#### **12. Medidas y columnas calculadas y sus fórmulas:**

**Columna calculada:**

- Diferenciamos rangos de edad con función IF en la tabla de Edad.

(Range = IF(Edad[Age]<=50,"Menor de 50",IF(Edad[Age]<=60,"50-60",IF(Edad[Age]<=80,"60-80", "+80"))))

- Calculamos rango de meses sobrevividos con función IF en la tabla Pacientes.

Rango de meses = IF(Pacientes[Survival Months]<=20,"Menos de 20",IF(Pacientes[Survival Months]<=40,"20-40",IF(Pacientes[Survival Months]<=60,"40-60",IF(Pacientes[Survival Months]<=80,"60-80",IF(Pacientes[Survival Months]<=100,"80-100", "100+")))))

- Agrupamos los meses de tratamiento con función IF en la tabla Tratamiento.

Agrupamiento = IF(Tratamiento[months of treatment]<=3,"Tratamiento corto",if(Tratamiento[months of treatment]<=6,"Tratamiento medio",IF(Tratamiento[months of treatment]<=12,"Tratamiento extenso", "Tratamiento muy extenso"))))

**Medidas calculadas (tabla Medidas):**

- Cantidad total de pacientes (N° de pacientes = COUNT(Pacientes[id\_paciente]))
- Promedio de edad de pacientes AVG (Promedio de edad = AVERAGE(Edad[Age]))
- Promedio de los costos del tratamiento Promedio:  
(Costos = AVERAGE(Tratamiento[Cost per month]))
- Promedio de ingresos mensuales de los pacientes:  
(Promedio de ingresos de pacientes por mes = AVERAGE(Tratamiento[Monthly Income]))
- Ingresos mensuales:  
(Ingresos mensuales = SUM(Tratamiento[Cost per month]))
- Ingresos anuales:  
(Ingresos Anuales =

**VAR** ingresos\_mensuales = **SUM**(Tratamiento[Cost per month])

**VAR** ingresos\_anuales =  
ingresos\_mensuales \* **12**

**RETURN** ingresos\_anuales)

- Tratamiento más costoso:

Tratamiento más costoso = **MAX**(Tratamiento[Cost per month])

- Tratamiento menos costoso:

Tratamiento menos costoso = **MIN**(Tratamiento[Cost per month])

### 13. Segmentaciones elegidas:

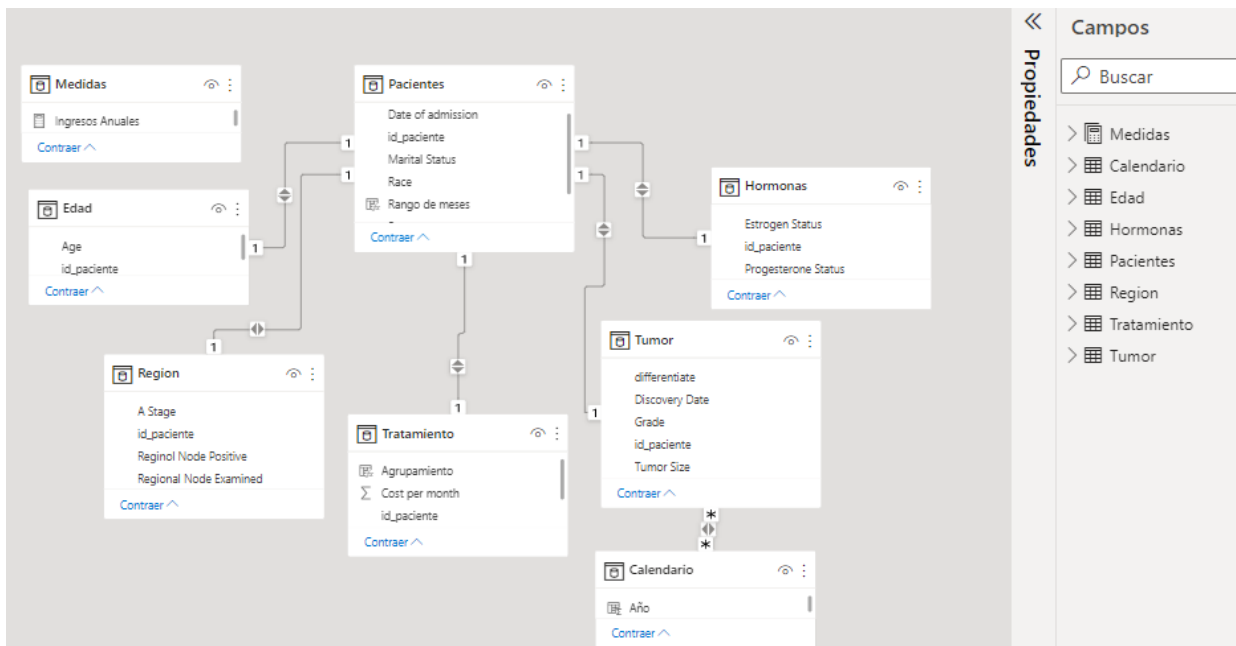
Utilizamos segmentaciones por Año, Tamaño del tumor, Presencia de hormonas, Meses sobrevividos y Edad.

### 14. Transformaciones realizadas:

Las transformaciones realizadas fueron las que hace por defecto Power BI.

### 15. Visualizaciones power BI:

#### Diagrama entidad-relación



## Solapa Presentación

# Breast Cancer Analytics

[Para más información](#) 

Proyecto por: Matias Trovatto y Val Isetta

El objetivo de este análisis es mostrar el comportamiento de la enfermedad, orientado a informar a quienes se encargan de tomar medidas de salud pública, y también a quienes padecen la enfermedad en sí.



El dashboard está orientado tanto al sector público encargado de tomar medidas sobre salud pública y las personas que padezcan la enfermedad y sus familias.




El análisis toca diferentes matices; una parte, dirigida a la comprensión de la enfermedad, sus grados, y el impacto en los pacientes según, por ejemplo, su edad; por otro lado, también fue analizada la esfera económica, mostrando los aranceles promedio para costear los tratamientos.




[→ Índice](#)

## Solapa índice


### Índice




Pacientes



Tratamiento

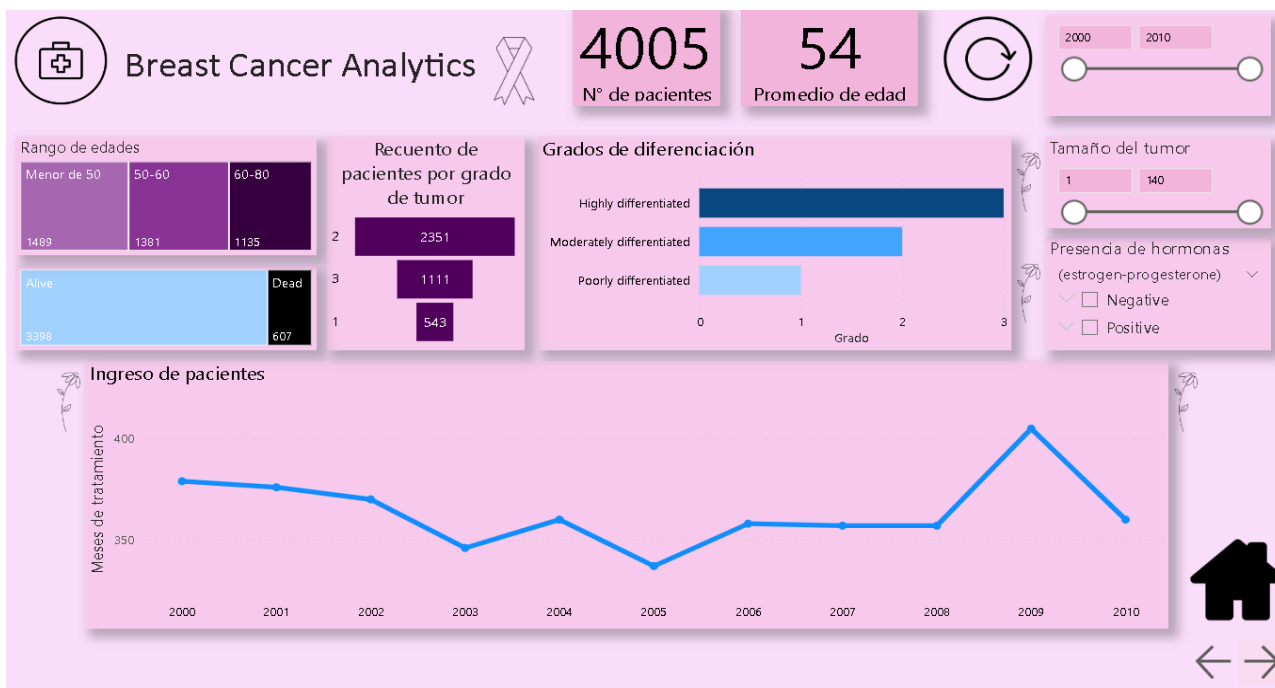


Duración del tratamiento



## Solapa Pacientes:

Contiene datos generales sobre los pacientes, tales como cantidad, edad, grado de tumor y su diferenciación, estado de vida y ingreso por año. Y como filtros, el año, el tamaño del tumor y la presencia de hormonas.



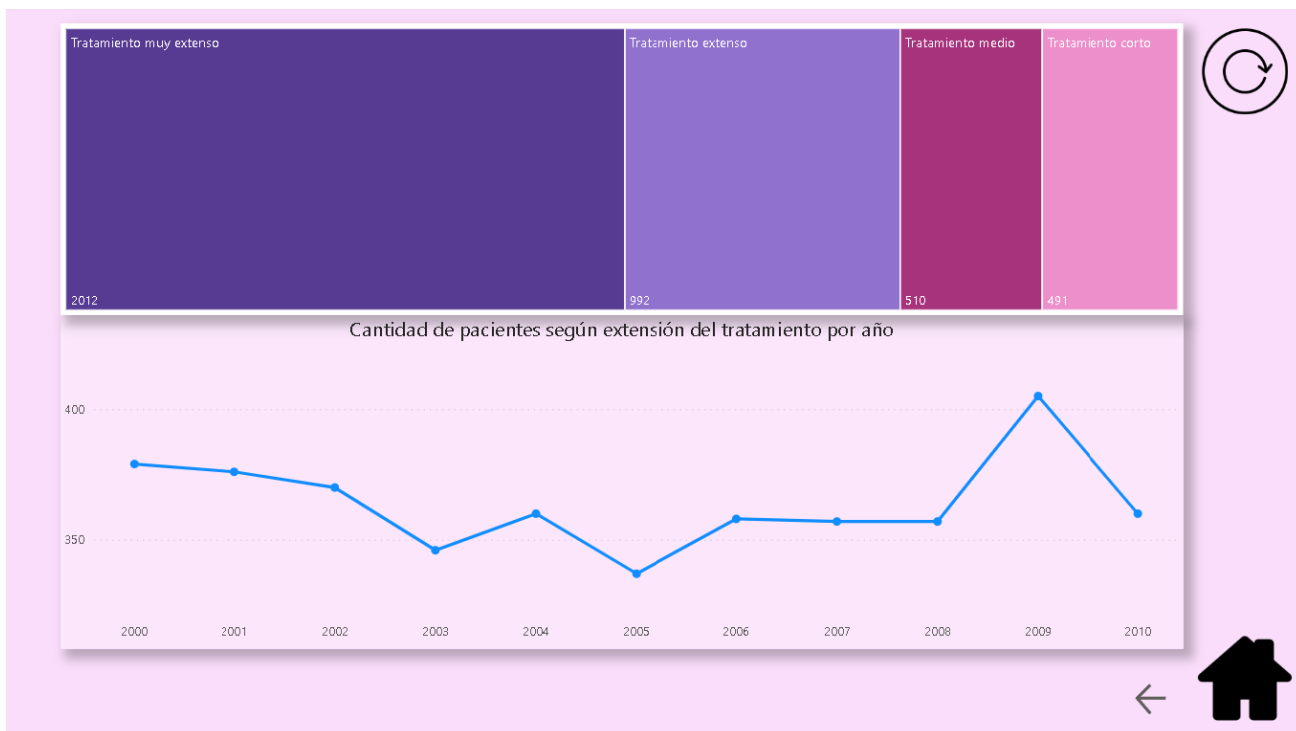
## Solapa Tratamiento:

Contiene las medidas sobre el promedio de los costos del tratamiento por mes, con el precio del tratamiento más costoso y el menos costoso, promedio de los ingresos mensuales de los pacientes y los ingresos mensuales y anuales del hospital.



### Solapa Duración:

Contiene la visualización de la duración de los tratamientos por año y cantidad de pacientes.



### 16. Futuras líneas:

Ahondar en como el tratamiento y la enfermedad afectan la vida personal de las pacientes, analizando el impacto en relación a su recuperación.

También es de interés analizar si los casos aumentan o decrecen a través del tiempo, y cómo afectan las condiciones socio-ambientales en relación a los aumentos de diagnósticos.

### 17. Drive

[https://drive.google.com/drive/folders/1myh0\\_kIU574br4Jr2eQ\\_0LveRj8mdWOF?usp=sharing](https://drive.google.com/drive/folders/1myh0_kIU574br4Jr2eQ_0LveRj8mdWOF?usp=sharing)

(Carpeta del drive donde se encuentran todos los archivos)

### 18. Bibliografía

- <https://www.breastcancer.org/es>