Breast Cancer Analysis using Visualizations

A Swedish Case study

Visual Data Analysis IT740A

Group 4

Hiwot G. Lassa
Welemhret W. Baraki
M. Usama Shehzad
Aimal Khan

2021-03-25

Report

Contents

Abstract

Breast cancer is the leading cause of cancer among women. Importance for its early detection and understanding about the disease has always been emphasized by researchers. The purpose of this project is to make the users and women in particular to better understand the disease through visualization of the statistical breast cancer data gathered over the last 15 years in Sweden by the Swedish National Breast Cancer Registry.

1. Introduction

Cancer is the name given to a collection of related diseases which are caused when some of the body's cells begin to divide without stopping and spread into surrounding tissues and develop into tumors. Breast cancer is the most common cancer type and is worth studying and visualizing to be able to extract patterns and give insights for different audiences.

Breast cancer is the leading cause of cancer in women and according to the International Agency for Research on Cancer (IARC), it has overtaken lung cancer as the most commonly diagnosed cancer with about 2.3 million women being diagnosed with breast cancer in 2020 [1]. Breast cancer was also the most common cause of cancer death in women with about 685 000 deaths in 2020. Breast cancer was also the largest cancer diagnosed among women in Sweden in 2019 where 8,288 were diagnosed and 1,353 women died with breast cancer as the underlying cause of death in 2019.

Although the diagnosis for breast cancer cases have increased, the mortality rates in developed countries have decreased with more advanced methods of treatment. However, early diagnosis, relevant awareness and information relating to the disease still plays a huge role. Visualization techniques can play an eminent role to give insights and understandings for patients, care staff, researchers and decision-makers.

Analysis and visualization of statistical breast cancer data of Sweden will provide insight to health professionals, the government and individuals on coverage, Severity, and other breast cancer factors and indicators. We will produce the visuals using Spotfire so that they can be presented to the mentioned end-users

1.1 Research questions/hypotheses

According to different research and statistical results breast cancer commonly occurs in women than in men. The main problem is the lack of coherent visualizations in the health system to treat breast cancer patients and taking smart decisions. Healthcare professionals, concerned management and government officials of the health system face challenges in understanding the coverage and severity of breast cancer due to lack of visual representations. As a result human and capital resources can't be efficiently distributed.

Research Questions

1. What are the statistics based on age, gender, mortality across regions in Sweden? (RQ1)
    - Which age group is most vulnerable to breast cancer?
    - How are men and women affected by breast cancer?
    - Has the mortality rate changed over the years?
2. How did the breast cancer patient's pattern change over Sweden for Men and Women? Can we make predictions for future years based on the previous trends? (RQ2)
3. Analysis of different regions across Sweden based on diagnosis (RQ3)

- How did morality change over the years across the regions and in Sweden in general?
- Which Genomic Characteristics i.e. Biological Subtype of Breast Cancer is common?
- How are the Oncological treatments used in the regions in Sweden?
- Distant Metastases

4. Analysis of different regions across Sweden based on oncological treatment? (RQ4)
    - Chemotherapy
    - Radiotherapy
    - Hormone therapy (Anti HER2)
    - Endocrine treatment
5. Analysis of different regions across Sweden based on surgery? (RQ5)
    - Breast reconstructive surgery
    - Mastectomy
6. How did the habits (patterns) of breast cancer patients of Individual Care plan change in Sweden? Is it increasing or decreasing? (RQ6)
7. Has early diagnosis improved over the years (2005 - 2019)? (RQ7)

2. Background

Fundamentally visual data analysis is critically important and plays an eminent role in the health sector. The data in the health sector that is exponentially growing from time to time needs special tools to analyse and manage for smart decision making for doctors and clinical personnel. Sanyour, et al [2] designed and developed a real time tool for cancer disease data analysis and visualization using K-nearest Neighbours, Support Vector machine and Naïve Bayes machine learning algorithms for classification to determine whether the breast lump is benign or malignant.

As a result many researches and technological developments have been adopted and used in recent years. In the era of Big data, where huge amounts of data is accumulated in the health repositories so that utilizing and visualizing the historical data of patients and developing inferences and statistical data is difficult for health professionals and decision makers. To support health professionals and decision maker's smart data analytics and visualization [3] makes health professionals draw smart and accurate decisions and conclusions for different problems. Health related data can be analysed using different methods, approaches and tools like Hidden Markov Models[4], deep learning [2] and other machine learning algorithms. The bio-psycho-social model [2] designed was to consider not only the biological aspects of patients but also their psychological and social aspects that matter(affects) the biological aspect of a patient. Understanding the psychological and social aspects of the patients designed to help and visualize the patients as well as the health professional for effective health treatment through effective interactive visualization. Our visual design can be standalone application, or web-based [5] according to the requirements.

We planned to solve the problem by collecting and cleaning the data set from different sources, designing and developing modern dashboards, bar charts, pie charts and others. The data repositories are the International Agency for Research on cancer and Swedish statistical repositories. The Spotfire data visualization and analysis software will be used. The prototypical representation of the system will be designed with the spotfire. For the analysis part, regression analysis will be conducted and show current and future predicted trends of breast cancer incidents in Sweden. The visualization process starts with pre-processing cancer data, transforming the data and creating visual mappings, finally creating views and presentations for the health professionals and decision makers to be able to extract knowledge. This visualization project on breast cancer is required for health professionals to pay attention to the most vulnerable ages, regions and give insight for their future tasks and actions to minimize and mitigate breast cancer.

3. Data

The statistical data sets of breast cancer are extracted from the national quality register for breast cancer (NKBC)[1] and Socialstyrelsen[2] which are collected from all of Sweden's healthcare providers. Data sets from the International agency for Cancer research (IARC) [3] for different countries gathered for comparison. The swedish breast cancer data set is in the National Quality Register for Breast Cancer (NKBC), coming from all of Sweden's healthcare providers. The purpose is to be a source of knowledge about Swedish care for patients, care staff, researchers and decision-makers. It has many indicators and variables like gender, age ranges, region, health care, tumor size, lymph node status, biological subtype, invasiveness, data after and before diagnosis, survival rates of each year according to each region and other variables. The data sets are produced from 2005 - 2019.

A few important attributes from the Swedish Dataset are listed below

Attribute	Value
Region	Blekinge, Dalarna, Gävleborg, Gotland, Halland, Jämtland, Jönköping, Kalmar, Kronoberg, Norrbotten, Örebro, Östergötland, Skåne, Södermanland, Stockholm, Uppsala, Värmland, Västerbotten, Västernorrland, Västmanland, Västra Götaland
Diagnosis Years	2008 - 2019
Women diagnosed	count
Age at diagnosis	count < = 65 years, count > 65 years
Screening detected breast cancer	count = yes / count = total
Invasive cancer at diagnosis	count = yes, count = No
Spread to lymph nodes at diagnosis	count(cN-) = No, count(cN+) = yes
Distant metastases at diagnosis	count(M0) = No, count(M1) = yes
Tumour size at diagnosis	count <=20mm (T0/T1) , count >20mm (T2-T4)
Biological subtype at diagnosis	Triple Negative = count HER 2 positive = count Luminal = count
Chemotherapy	Pre-Operative = count Pre and post operative = count Post operative = count

[1] National quality register for breast cancer (NKBC): https://statistik.incanet.se/brostcancer/ (accessed on Feb 2021)
[2] Socialstyrelsen: https://www.socialstyrelsen.se/statistik-och-data/statistik/statistikamnen/cancer/ (accessed on Feb 2021)
[3]IARC (International Agency for Cancer Research) for world countries: https://gco.iarc.fr/today/home (accessed on Feb 2021)

Types of surgery	Breast conserving = count
	Breast reconstruction = count
	Mastectomy = count
	Axillary = count

5 year survival	count = yes / count = total

4. Approach

4.1 Data preparation

After collecting data from different sources we cleaned and transformed the data into suitable data features and formats for the Spotfire software package.

The following activities were performed during the data preparation process:

Step-1- Data Collection –The breast cancer data sets were collected from the National Quality Register for breast cancer, socialstyrelsen and International Agency for Cancer Research (IARC).

Step-2- Pre-process and clean The data sets were collected from the above mentioned resources and relevant features were identified and filtered for the breast cancer analysis and visualization.

Step-3-Transformation The pre-processed and cleaned data was transformed by merging the data in some of the required tables and also pivoting of the data was transformed on some of the tables in Spotfire.

After conducting the data preparations process, the data analysis was performed and is described in the section below.

4.2 Data analysis

The Spotfire tool is used for data analysis and visualization of breast cancer from the datasets. The spotfire software package tool has plenty of features which allows users to integrate data into a single framework and to achieve a cohesive view with an integrated visualization. It speeds up visualization across an organization for faster, confident and much accurate decision making. Additionally, the Spotfire software also has built-in implementation features for different machine learning algorithms like regression analysis, classification, clustering and many more features. We have used regression modelling to predict future expected breast cancer patients in each region or the country as a whole.

4.3 Visualization and interaction

The collected data visualizes effectively to avoid biases and minimize risks by easily supporting the different stakeholders. We have visualized patterns, trends and information in the form of line charts, barcharts, maps, scatterplots, pie charts, tree maps, spider charts and others.

The visualizations answer the research questions mentioned above by interpreting the results and also by the results from the evaluation strategy.

4.4 Evaluation

The evaluation strategy will be based on usability testing and heuristic evaluation techniques. Usability and heuristic evaluation will be performed based on the results. Heuristic evaluation is conducted according to the system specifications and the problems associated are identified to check if the system functionalities are included.

Finally Usability testing will be conducted to evaluate the system on the visualization of breast cancer by distributing some questions to express the users how and what they feel based on the System Usability Scale distributed from strongly disagree to strongly agree. The Users responses will be analyzed and present the result.

5. Results

The results from the visualizations are given in the following section.

5.1 Data preparation

During the data preparation, we tried to collect and compile the data collected from different sources.

5.2 Data analysis

The data analysis for the problem can be visually assessed using the tool. The visualizations will be useful to find the patterns from the available breast cancer data.

5.3 Visualization and interaction

The results from the visualizations based on the research questions are as follows.

RQ 1 - 1. What are the statistics based on age, gender, mortality across regions in Sweden?

1.1 Has the mortality rate of women changed over the years?

> Nowadays the number of incidence case and death in breast cancer leading increase in the world. In Sweden the mortality rate of breast cancer for women highly decreasing over the last 10 years (from 2009 to 2019) in figure 1. From figure 2 we can observe that the number of affected cases has increased for women the last 10 years. The total number of mortality rate of Swedish women is 17, 014 for year 2008-2019. The number affected by breast cancer cases in Swedish women is 109,109 for the year 2008-2019. In Sweden still increase women the risk of breast cancer. However, we can observed that the two visualization graphs women to get better medical treatment in Sweden. For example medical treatments like women mammography screening, follow-up and also awareness.

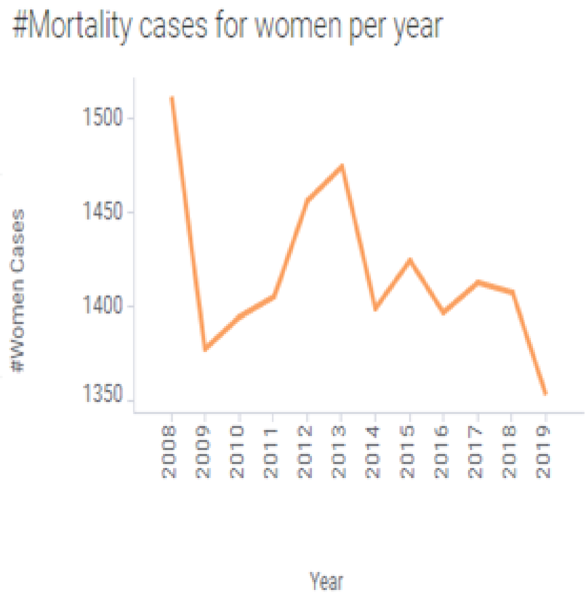


Figure - 1. The number of mortality in 2008-2019

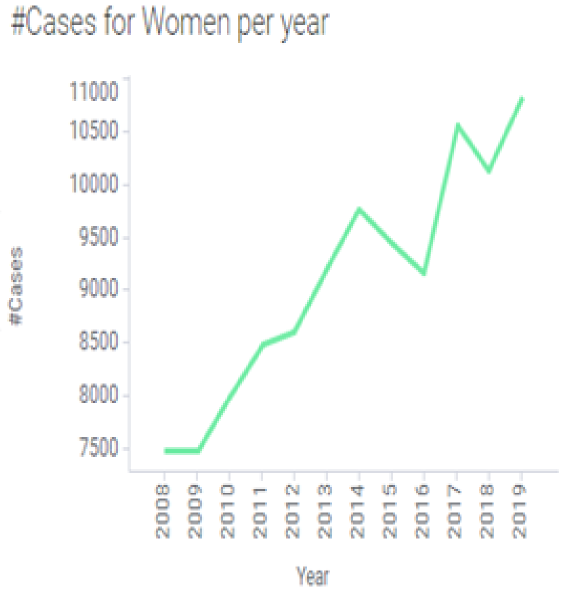


Figure - 2 The number of breast cancer incidence cases 2008-2019

1.2 Is Breast cancer a higher risk for men than women in Sweden?

The mortality percentage of men in some regions are lower than women such as **Gotlands län and Västerbottens län. The lower percentage of men versus women is 0% vs 15.32%, 0% vs 13.86.** Some regions the mortality rates of men **unexpectedly** higher than women these are **Gävleborgs län, Jönköpings län, and Östergötlands län.** The reason for men mortality rates higher than women are, traditionally breast cancer is most commonly thought of women's disease and lack of awareness of its occurrence in men may lead to diagnoses at later age and more advanced stages than in women. In the graph the color of men represented by blue and women represented by red for in this case the blue color is some regions larger than the red.For example in **Gävleborgs län** registered men affected 15 and death 9 the percentage of mortality for men is 60% and also women affected the same region is 3,002 and death 470 the percentage mortality of women is 15.656%. The first higher percentage of mortality registered in the region is **Gävleborgs län** the year of 2008 to 2019. Therefore, in this visualization observed how many men and women are deaths individually out of the the total number of affected cases in each region.

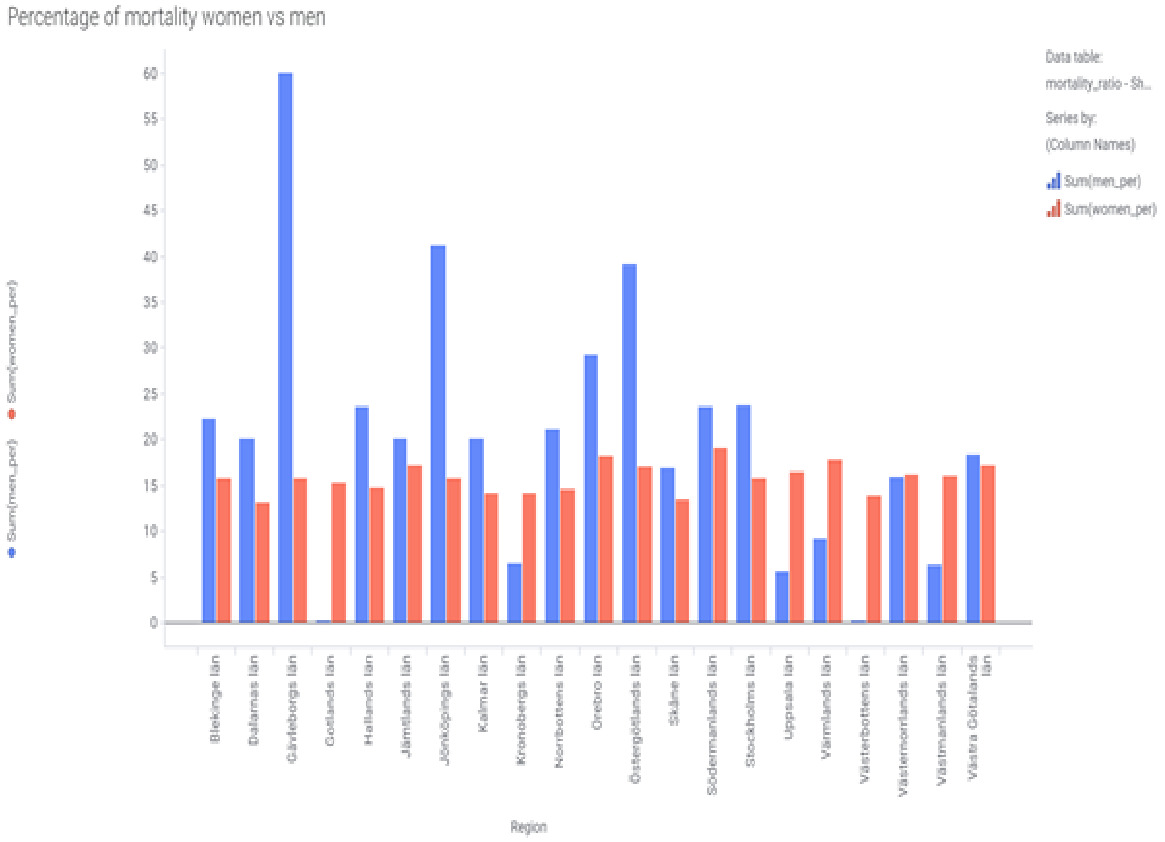
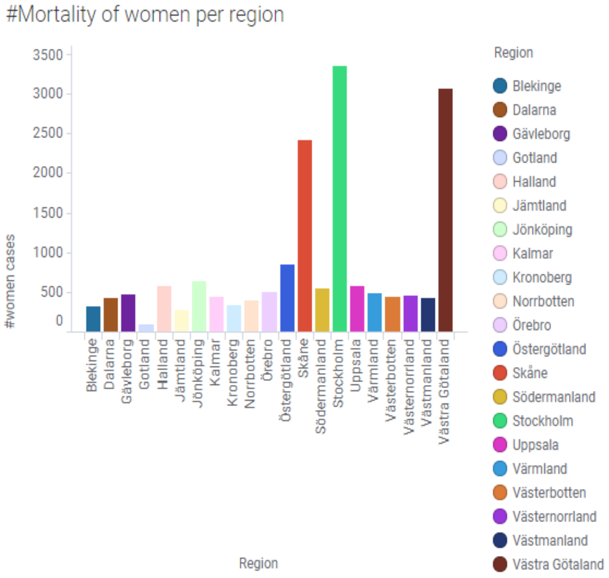


Figure - 3. The percentage of mortality women and men per region

1.3 How about the mortality rate of breast cancer for both genders in Sweden different regions?

In Sweden the number of mortality in breast cancer on the three regions (**Stockholms län**, **Skåne län,** and **Västra Götalands**) are higher than compare to the remaining. In **Stockholms län**, **Skåne län,** and **Västra Götalands**) is both genders are similarly more highly affected by breast cancer cases are registered compared to the other region in Sweden. In figure 4a and 4b observed that the first, second and the third higher mortality cases of men and women found in **Stockholms län, Skåne län,** and **Västra Götalands** these values are 27, 19, and 15 for men and 3347, 3066 and 2405 for women respectively. In general, the two visualization graphs understand the number of mortality of women higher than men.



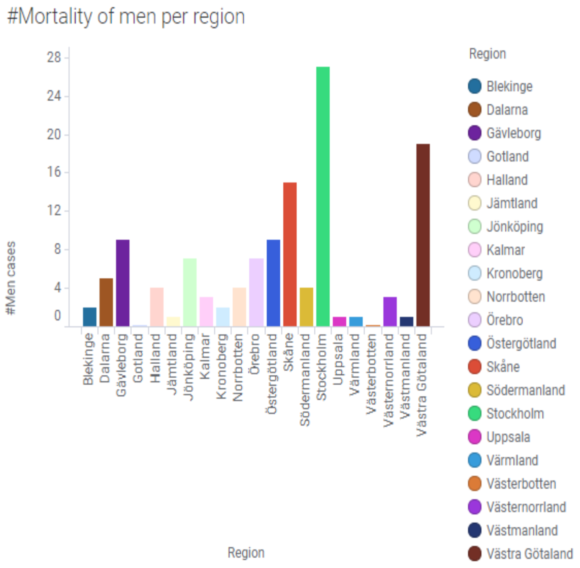
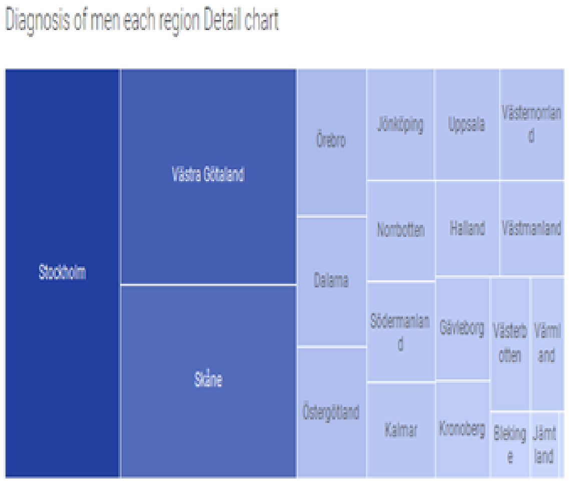


Figure - 4a. No.Mortality of women per region **Figure - 4b**. No. Mortality rate of men per region

The Diagnosis of men in different regions but three regions with the highest number of men breast cancer diagnosis during the period 2008 to 2019. We can observe those two visualization high registrations in those three regions compare to the other **Stockholms län**, **Skåne län,** and **Västra Götalands län.**



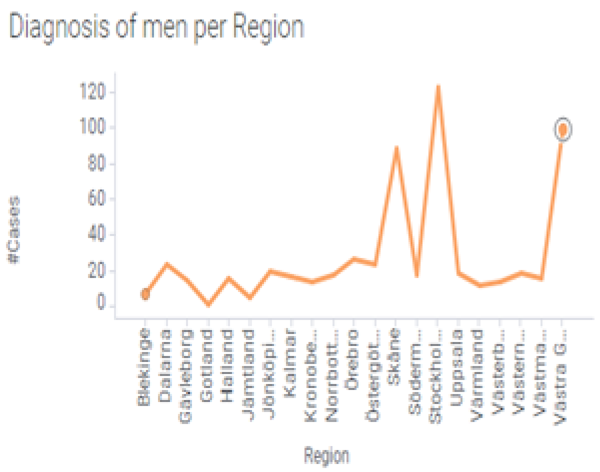


Figure 5 Detail visualization Figure 5 No. of Diagnosis of men per region

1.4 Which age group in different regions is more affected by breast cancer?

Breast cancer cases expectedly highly affected a person above age 40 years and older. In this study, different region men and women affected by breast cancer the year of 2008 to 2019.In addition, compared to the all age groups to more affected are the ranging (50-59, 60-69 and 70-79) highly affected than the other age group.From the figure_6 observed that highly affected find in the three regions are Stockholms län, Skåne län, and Västra Götalands län. The green, magenta and yellow colors represent the age group of 60-69, 70-79 and 50-59 respectively. Those highly affected age groups are found in the all-region except **Västmanlands län** region in Sweden. In **Västmanlands län** the two age groups are the same as the other but the third one is a different 80-85+ age group.

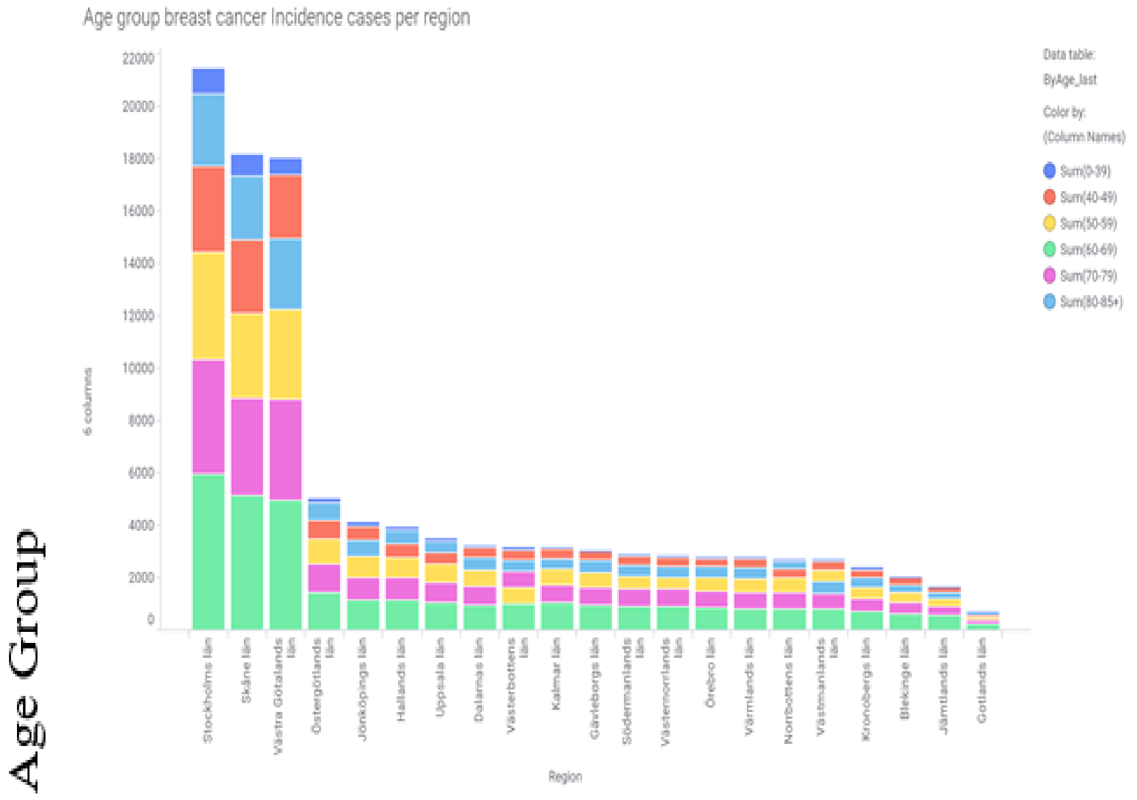


Figure 6 Age group breast cancer Incidence cases per region

RQ 1 What are the statistics based on age, gender, mortality across regions in Sweden?

1.1 Which age group is most vulnerable to breast cancer?

As you get older, the chances of having cancer increase. Fortunately, Men and Women affected over the age of 40. Here in the diagram different colors show the age groups that are vulnerable to breast cancer. Between the age of 60 and 74, the majority of deaths from cancer as the primary cause of death occur. Young women have a higher cancer prevalence and mortality rate than young men. Before the age of 50, there is a significant gap between the sexes, owing to the fact that breast cancer and gynecological diseases can strike at any age. In this visualization we can see how different age groups showed which age group more affected.

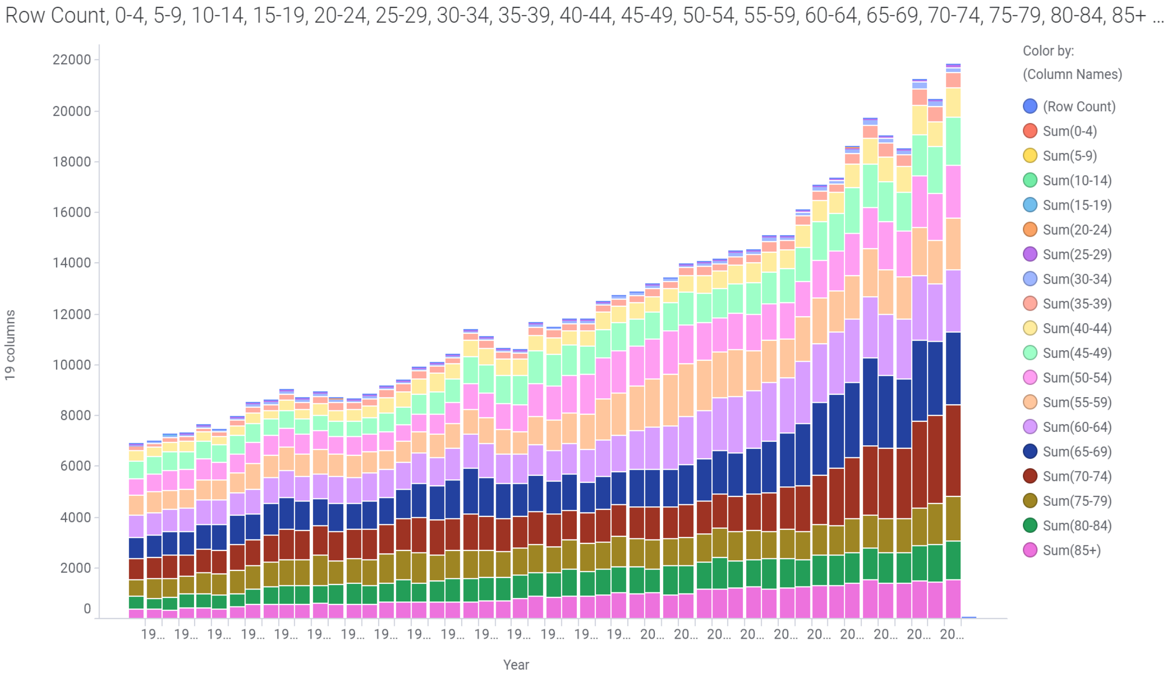


Figure 7: Age group breast cancer Incidence**.**

1.2 How about the mortality rate of breast cancer for men and women in different regions?

Here is the diagram that shows the age diagnosis proportion within 65 years in different regions of Sweden. In this visualization one region with the highest number of men and women breast cancer diagnosis during the period 2008 to 2019 is stockholm.

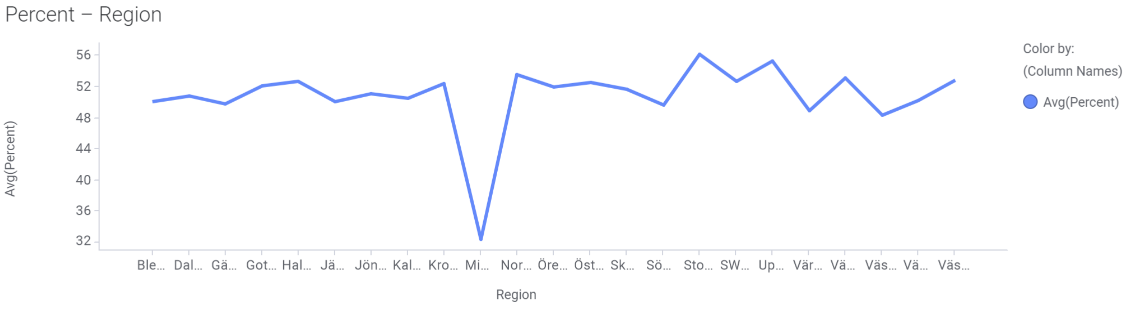


Figure 8: Number of diagnosis within proportion 65 years

RQ 7 Has early diagnosis improved over the years (2008-2019)?

Breast cancer survival has steadily increased in recent decades, but population-based studies of patients diagnosed at various stages of the diseases and in various age groups have yielded little information on the progress. Small-sized cancers have increased dramatically, whereas large-sized cancers have decreased just slightly. Large cancers may be avoided with an intervention that increases the likelihood of self-detection. Every year, over 7000 women out of Sweden's nine million people are diagnosed with breast cancer, with 1500 dying as a result.

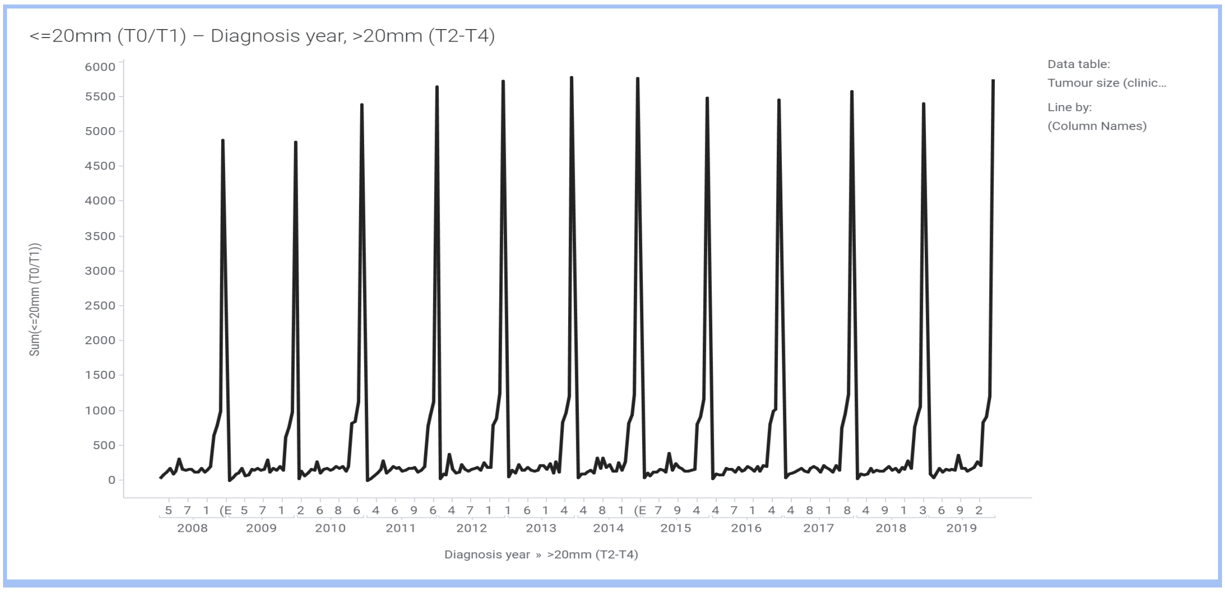


Figure 9: Diagnosis with tumor size

Question 1: **How did the breast cancer patient's pattern change over Sweden for Men and Women? Can we make predictions for future years based on the previous trends? (RQ2)**

Since this research question is mainly concerned to the management body, the forecasts for each region is not included and is forecasting the number of future breast cancer patients for women and men in Sweden(country wise). Most of the time the higher government officials may find it useful, than the lower health professionals and managers. To show the distribution and trends of data that change over time, line charts are commonly used. As a result, a line chart is used to visualize the historical distribution and predicted trends of women and men breast cancer patients between the years of 2008 to 2023. The predicted curve (dotted) figure below shows the predicted values of each woman and men 4 years ahead. In our data the previous year's breast cancer data is not included, that's why included in the predicted curve.

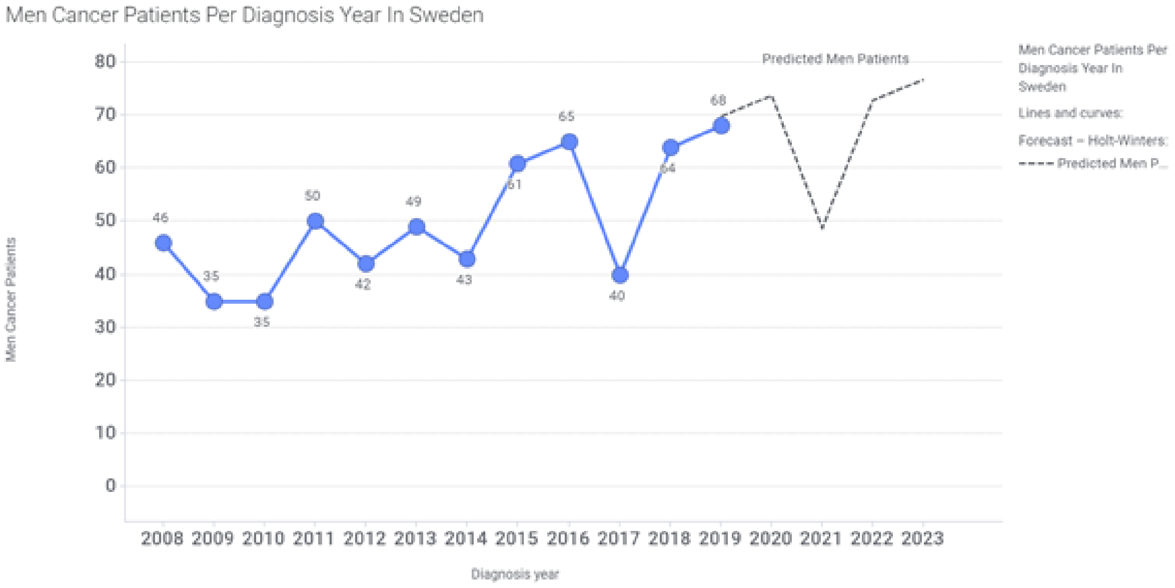


Figure 10: Men Cancer Patients per Diagnosis Year in Sweden

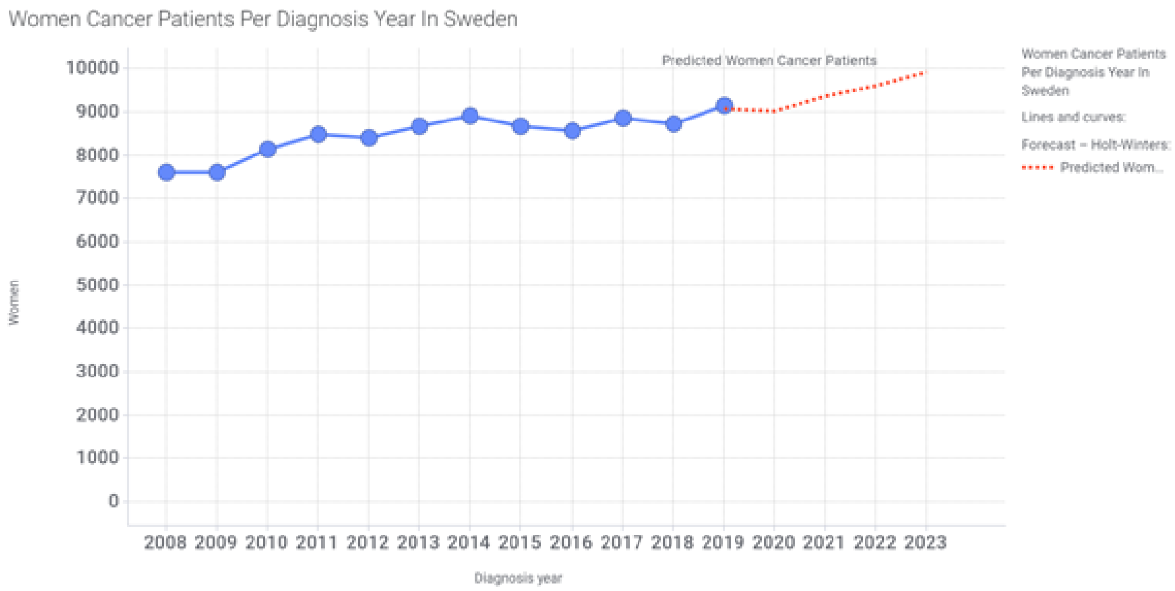


Figure 11: Women Cancer Patients per Diagnosis Year in Sweden

From the above two visualizations the users can easily recognize and track the changes in the number of breast cancer patients over the years in Sweden. The number of breast cancers in both men and women are on the rise. The predictions can be informed and used as input for their future plans and decisions to reduce the risks of breast cancer patients and health recommendations will be given to patients. The main reason that we choose line charts is that, it is easy to interpret and handle trends and changes over time.

Question 2: Analysis of different regions across Sweden based on Mortality, Biological Subtypes and different Oncological Treatments? (RQ4)

The analysis and visualization is mainly focused on the mortality, biological subtypes and oncological treatments of different regions across Sweden. We divided this research question into different manageable sections so as to create the best visualization.

How did Morality change over the years across the regions and in Sweden in general?

The figure below mainly focuses on the visualization of the mortality (death) of breast cancer patients each year in each region in Sweden using the bar charts with a filtering scheme with regions. List of options in-terms of each region is available to filter out and add flexibility to the users to navigate and visualize according to the interest of the users based on the regions. Accordingly the users of our visualization can simply track the changes in breast cancer deaths in each year, per each region or country wide in Sweden. Additionally, the bar chart also shows the gradient of colors to overlook the maximum (dark blue) and minimum (light blue) mortality of a particular selection (region/s) over the years of 2005 to 2019.

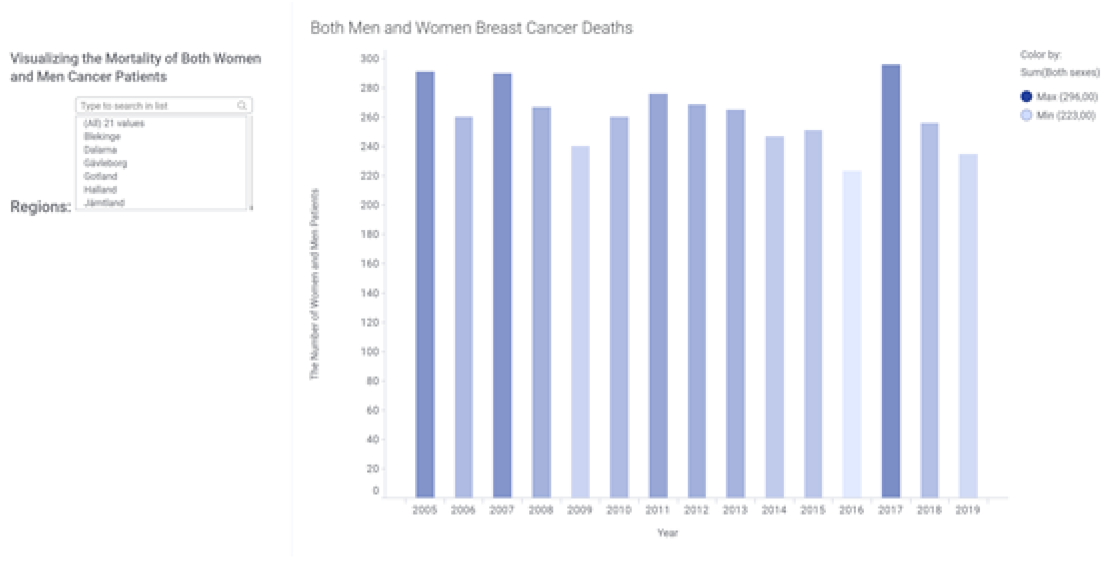


Figure 12: Mortality Västra Götaland

From the above bar chart the user can track the death changes in every year across the regions of Sweden, for example, in 2017, the recorded deaths was 296(highest), in 2016 it was 223(lowest) deaths.

- **Which Genomic Characteristics (Biological Subtype) of Breast Cancer?**

The figure below of the pie chart shows the proportion of the biological subtypes of breast cancer patients in Stockholm. Breast cancer subtypes are defined according to the biologic properties of the invasive cancer and different subtypes respond to different therapies. To choose the appropriate therapy identification of biological subtypes of breast cancer patients helps identification of the biological subtype is important. Different treatment therapies require different professionals, materials and others, knowledge of the proportion of the biological subtypes of breast cancer patients is essential. The pie chart was used to show the percentages and proportions on the three biological categories particularly in the case of the Stockholm breast cancer patients 77.7% have Luminal, 13.7% HER2 and 8.7% Triple Negative biological subtypes. The user can simply recognize the Biological subtypes and infer that Luminal is common in Sweden (Stockholm).

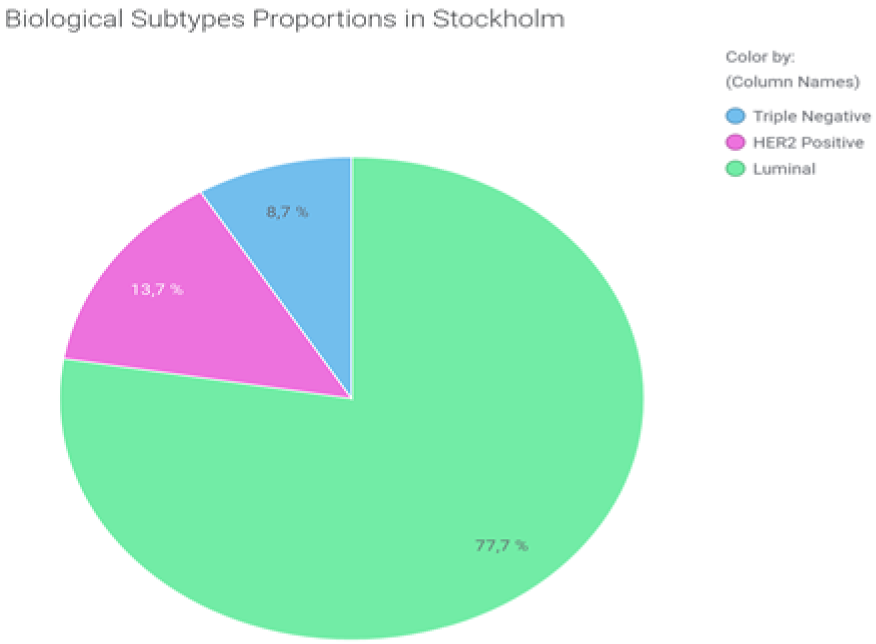


Figure 13: Proportion of the biological subtypes in Stockholm region

- **How is the Oncological treatment used over the regions in Sweden?**

The oncological treatment visualization is depicted using the parallel coordinate plots to understand the patterns of different breast cancer treatment types in each region with respect to the years. The different regions are represented using different colors, treatment types used in the x-axis with their respective percentage of breast cancer. The same region can represent using the same color representation for different years. Based on the different treatment methods each region shows different numbers and when the user hovers on the lines the number of patients treated and the year was displayed for a particular treatment. This can ease the comparison of treatments across different regions. Additionally, it also provides the ability to compare the single region's historical data easily based on the years. The stockholm region have generally the highest number of patients that take different treatment methods for breast cancer.

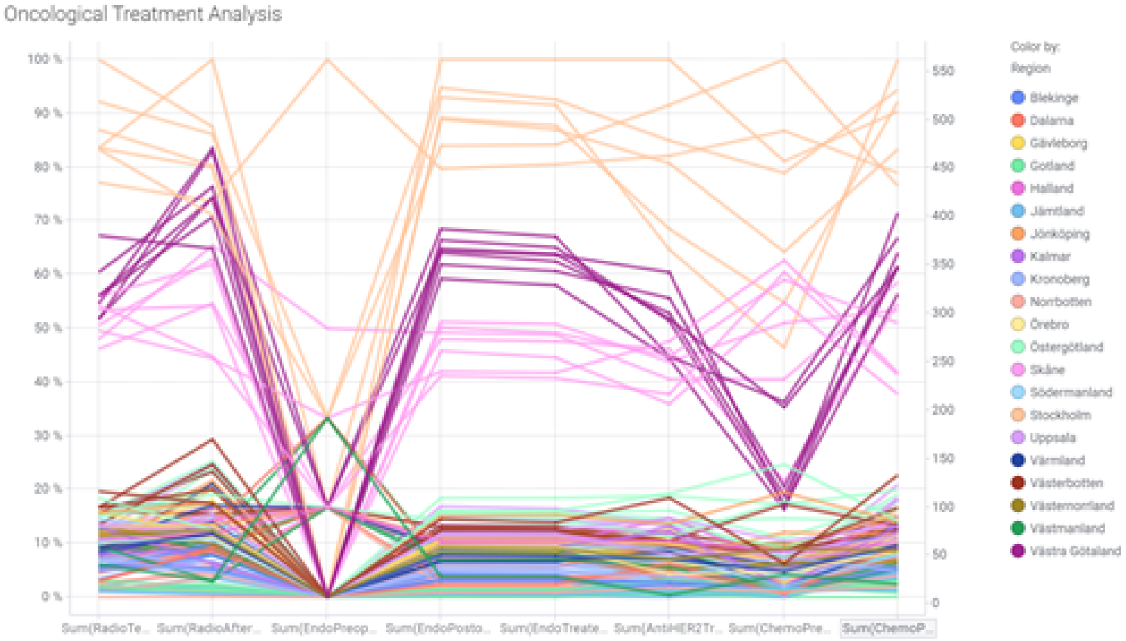


Figure 14: Different Oncological Treatment Analysis

Question 3: How did the patterns change in the Breast Conserving Surgery over the Years in Each Region? (RQ5)

To show the Breast Conserving surgery per diagnosis year per region we use the tree map as in the figure below, in which the area and the orange color shows those regions that are above average and have a high number of patients that conduct breast conserving surgery. The blue color tree map shows those regions that are below average. The treemap is mainly used in hierarchical data representation, in which our data can be represented in a treemap that the year can be considered as root of the tree and regions as branches of the tree.

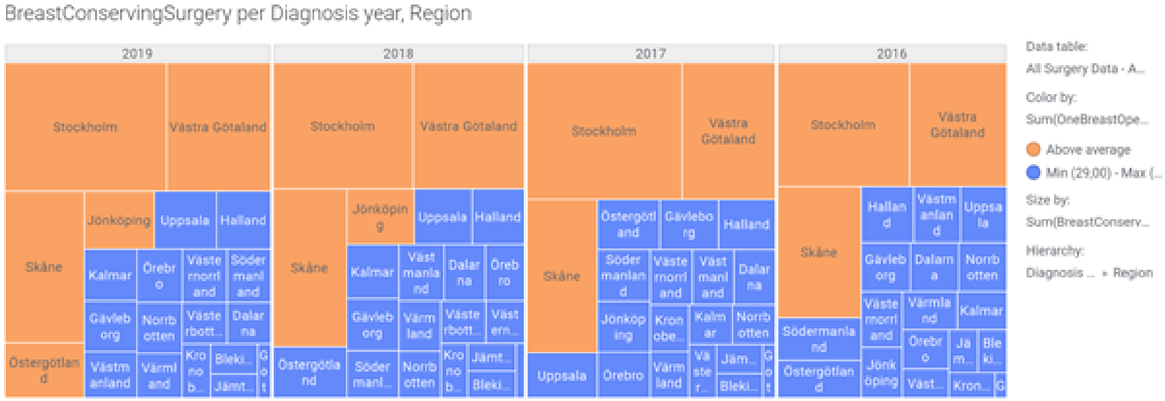


Figure 15: Breast Conserving Surgery per diagnosis year per region

Question 5: How did the habits (patterns) of breast cancer patients in Individual Care plan change in Sweden? Is it increasing or decreasing? (RQ6)

An individual written care plan should be prepared for each patient with cancer according to the National Cancer Strategy and all cancer patients should be offered a contact nurse. In 2019, the regional Cancer Centres published a national description of the contact nurse's assignment. The bar chart below depicts that there is an increase in awareness and use of individual care plans for breast cancer patients. This shows that not all breast cancer patients have individual care plans.

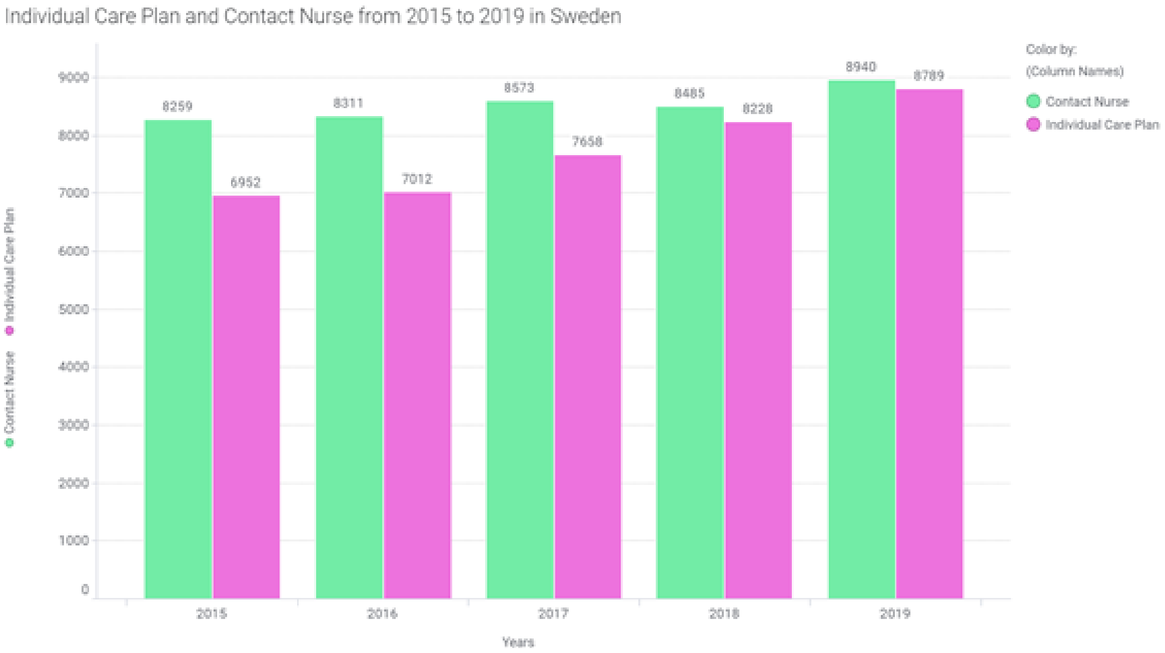


Figure 16 : Individual Care plan and Contact Nurse in Sweden

Overall, the charts are selected based on the nature of data and persuasiveness of the visual representations in different contexts without creating clutter and confusion to the end-users.

RQ3 - Analysis of different regions across Sweden based on diagnosis

Question: Is breast cancer detected through screening and are the regions in compliance with the target levels?

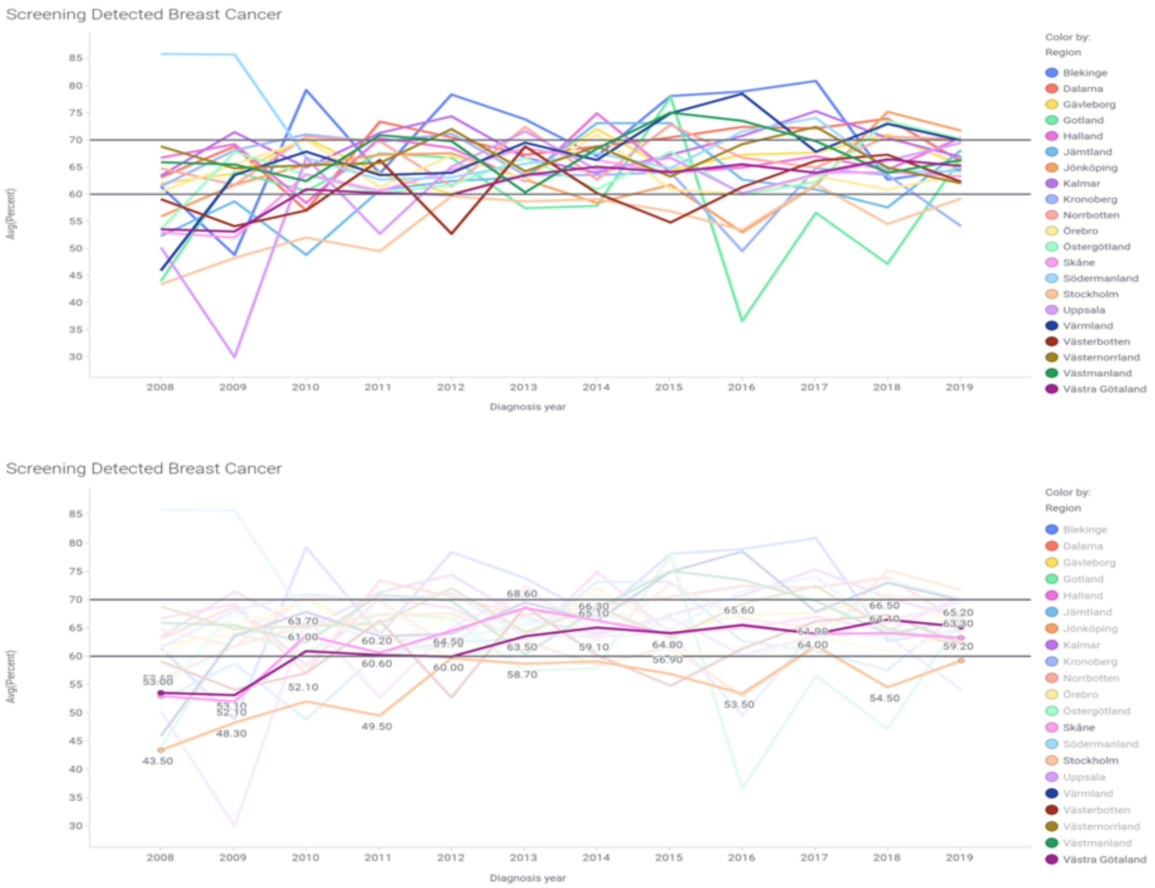


Figure 17: Screening detected breast cancer

As can be seen from the above figure 17, the percentage of screening detected breast cancer has increased over the years. The level of compliance should be 60%, with 70% and above falling in the higher target range. The trend shows that most regions are in compliance with the target levels by improving over the years. Stockholm is the noticeable region which although improved over the years, still does not meet the required target levels.

A question pointed out by the teachers during the final review check was that if Skåne and Västra Götaland are in the compliance target levels then why the mortality rate is high for these regions. Figure 17 above shows the percentage mortality rate across the regions over the years. As it can be seen, the trend for the mortality rate of these two regions is generally on the decline and the gap difference from Stockholm is increasing. Also it is not the only factor that can be considered for the mortality rate. The biological subtype, the tumour size and the treatment pathway are altogether important to reduce the mortality rate. The notable point here is that more work needs to be done in the form of creating awareness among people in Stockholm to go for breast cancer screening whether they have the cancer or not, and get it detected through screening if they have it, which eventually will be a factor in detecting the cancer early and reducing the mortality rate.

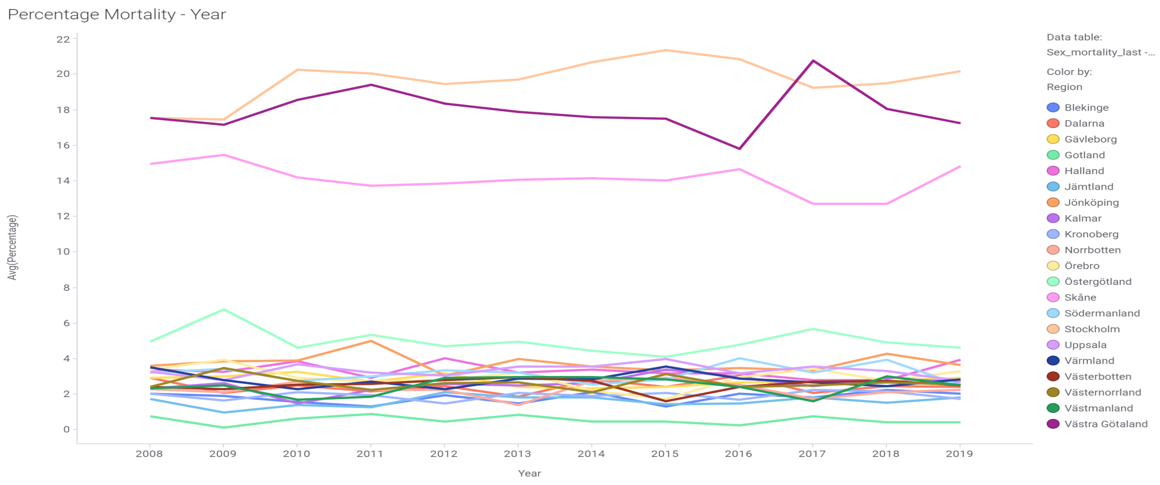


Figure 18: Percentage mortality per year

RQ3
Question: Which biological sub type of breast cancer is more prevalent?

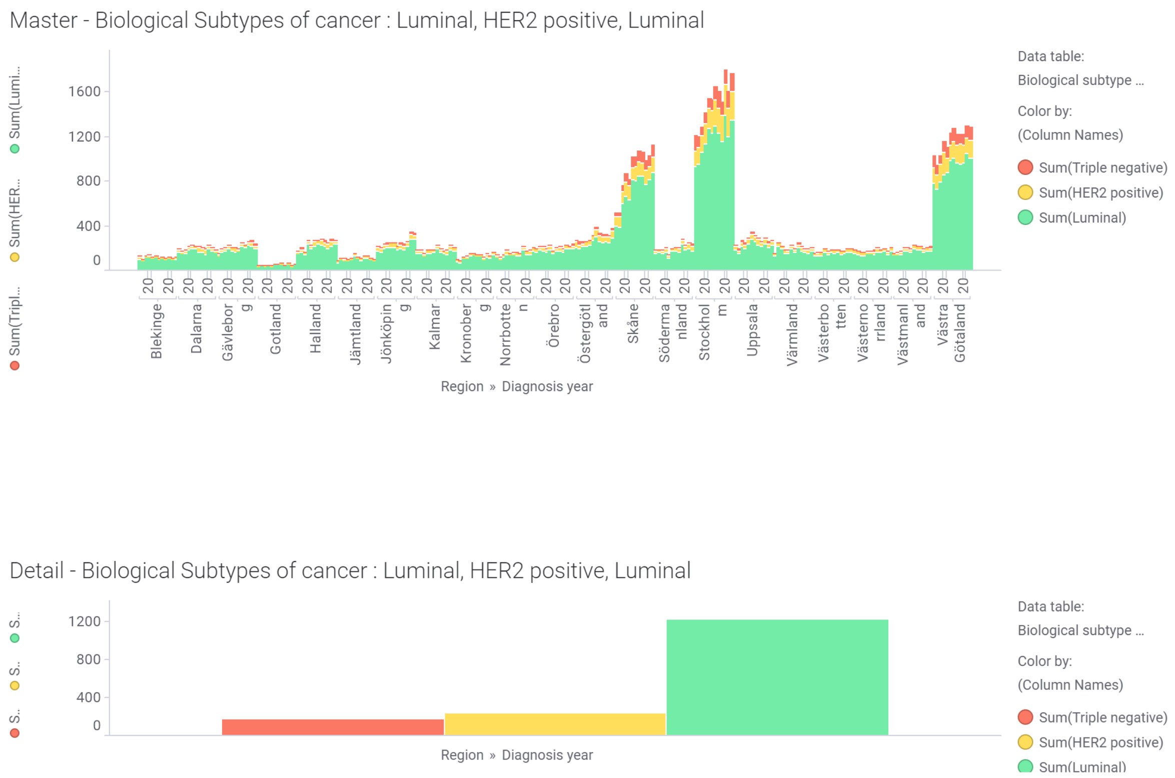


Figure 19: Biological sub type of cancer

There are three common subtypes of breast cancer namely Luminal, Her2 Positive and Triple Negative. It can be seen that Luminal subtype of breast cancer is the most common across Sweden. The main regions with the most number of cases are Stockholm, Västra Götaland and Skåne. The treatment plan for these areas needs to be designed accordingly.

RQ3

Question: How much does the pathology report comply by including all biological markers for the diagnosis?

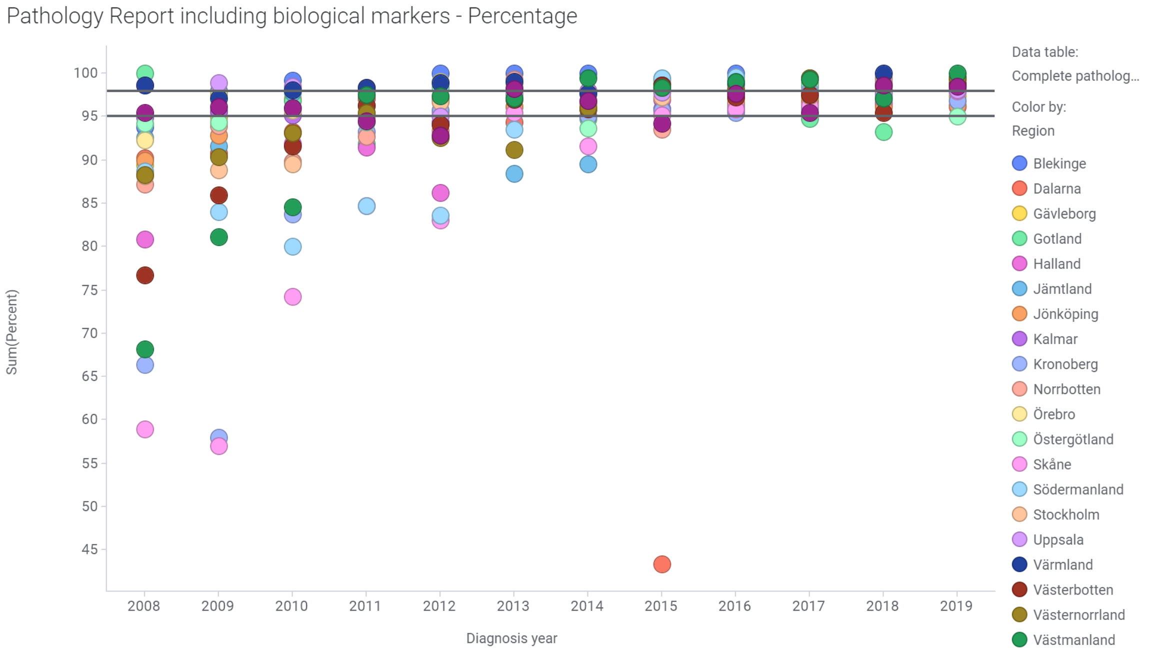


Figure 20: Pathology report

The compliance of the pathology report has improved over the years for all the regions. In the last five years almost all regions have maintained the target level of 95% to include all biological markers for diagnosis in the pathology report, with a few exceptions; Gotaland with 93% in 2018, Gotaland with 94% in 2017, Västra Götaland, Norrbotten and Jonkoping with 94%, 93% and 43% in 2015 respectively.

RQ3

Question: What is the percentage of Tumour size > 20 mm at diagnosis?

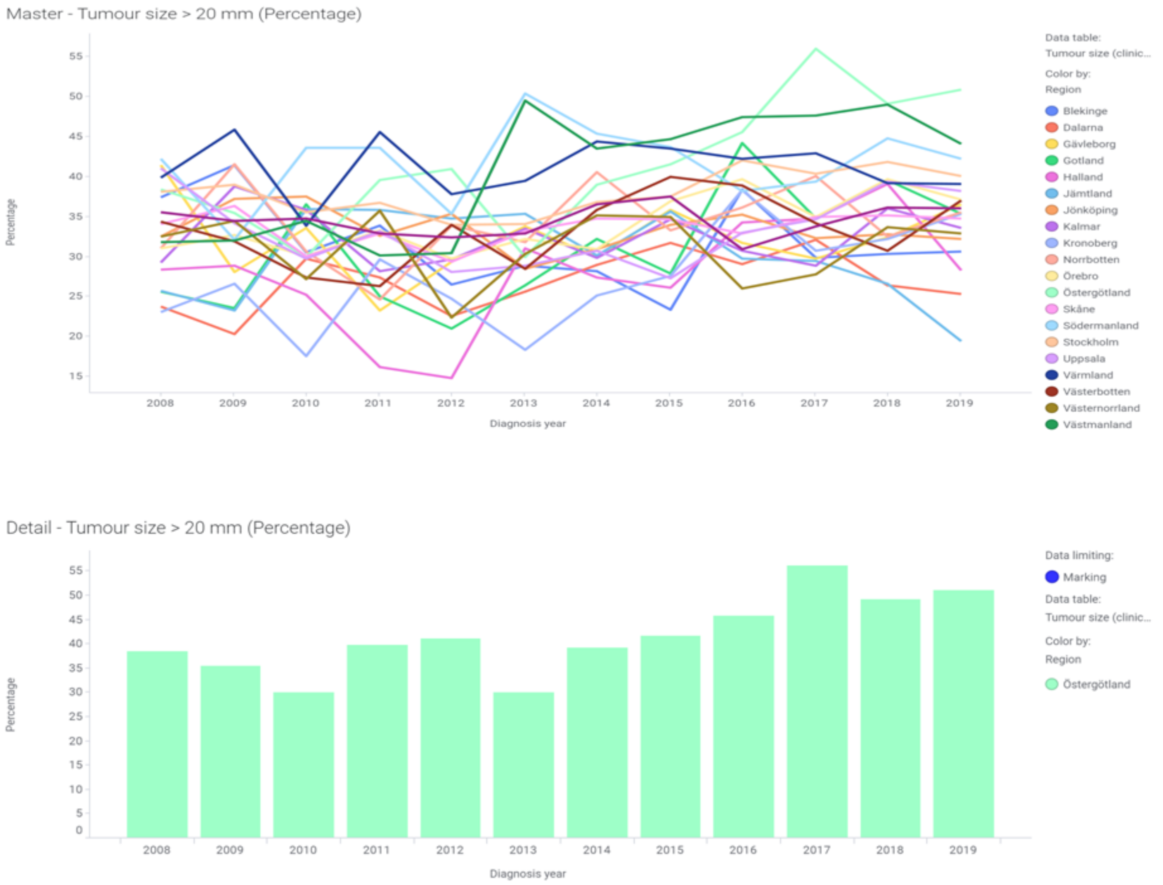


Figure 21: Tumour size

This is an important question, because early diagnosis can be related to the size of tumour. If the tumour size is large, it is inferred that the diagnosis was not early. The percentages of tumour size > 20mm over the years show that the results vary across the regions. A key focus should be that the tumour size should be minimal at diagnosis, as it will then lead to better treatment outcomes.

RQ3

Question 8: How are the significant features of breast cancer related?

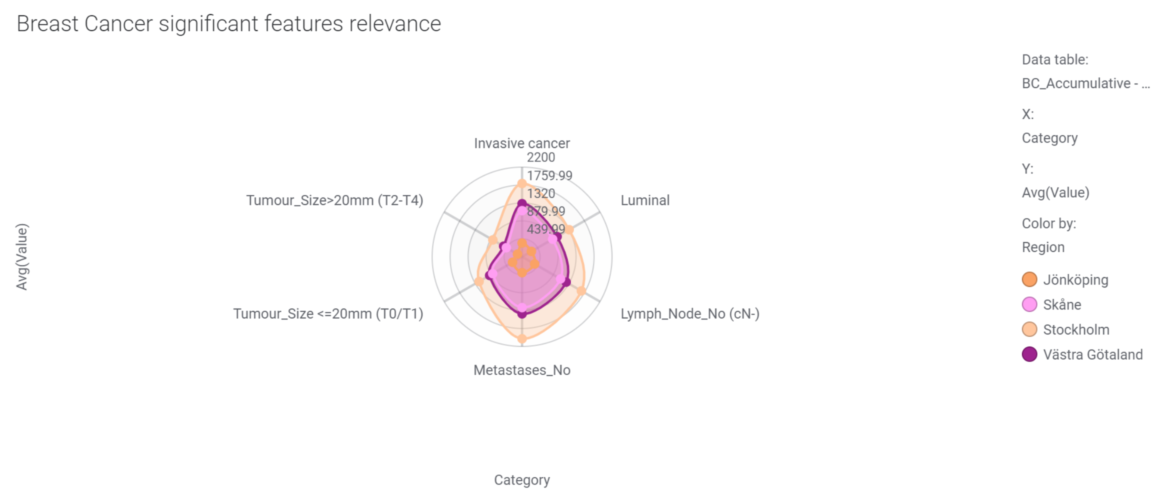


Figure 22: Breast Cancer Significant features relevance

The chart above shows the features that are related and therefore have higher significance than others. The top 3 regions show a similar relation to the features.

RQ4

Question: Has Chemotherapy Treatment been effective?

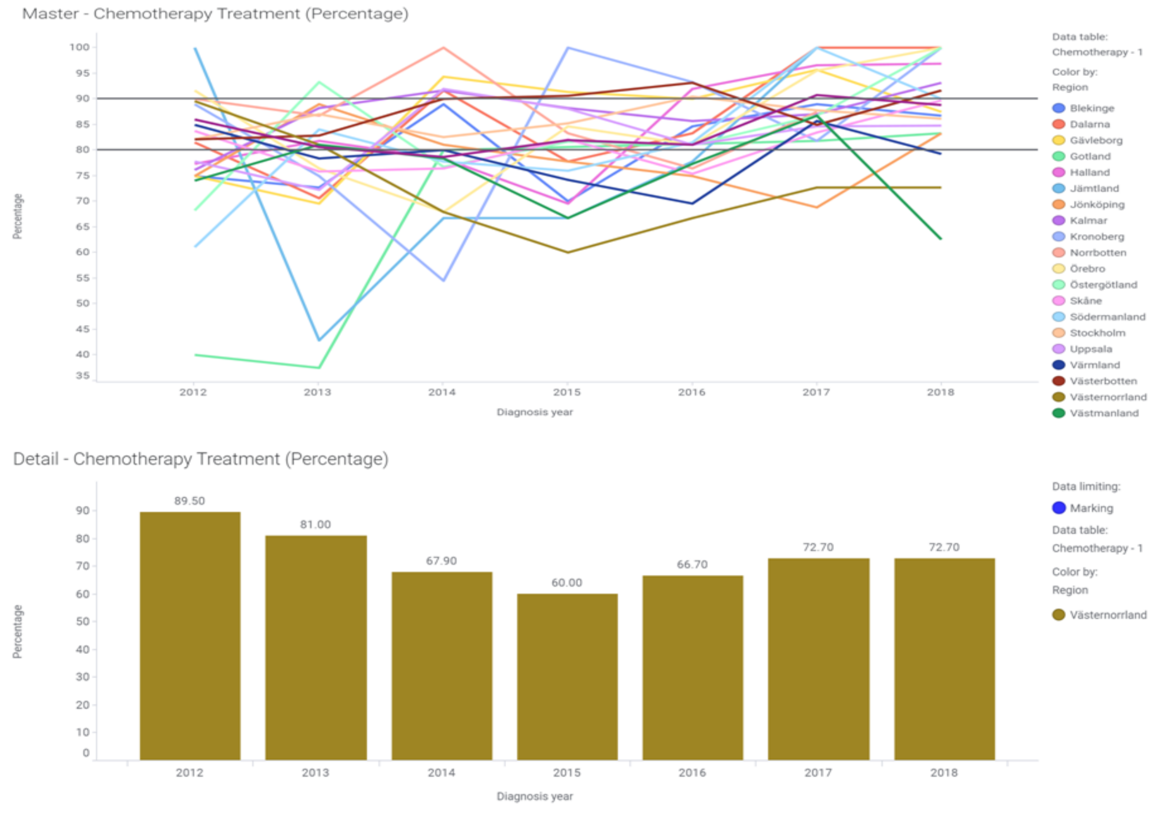


Figure 23: Chemotherapy Treatment

The compliance level for chemotherapy treatment over the years has mostly been met by all the regions except for Västernorrland.

RQ4

Question: How do the different regions relate to the breast cancer features?

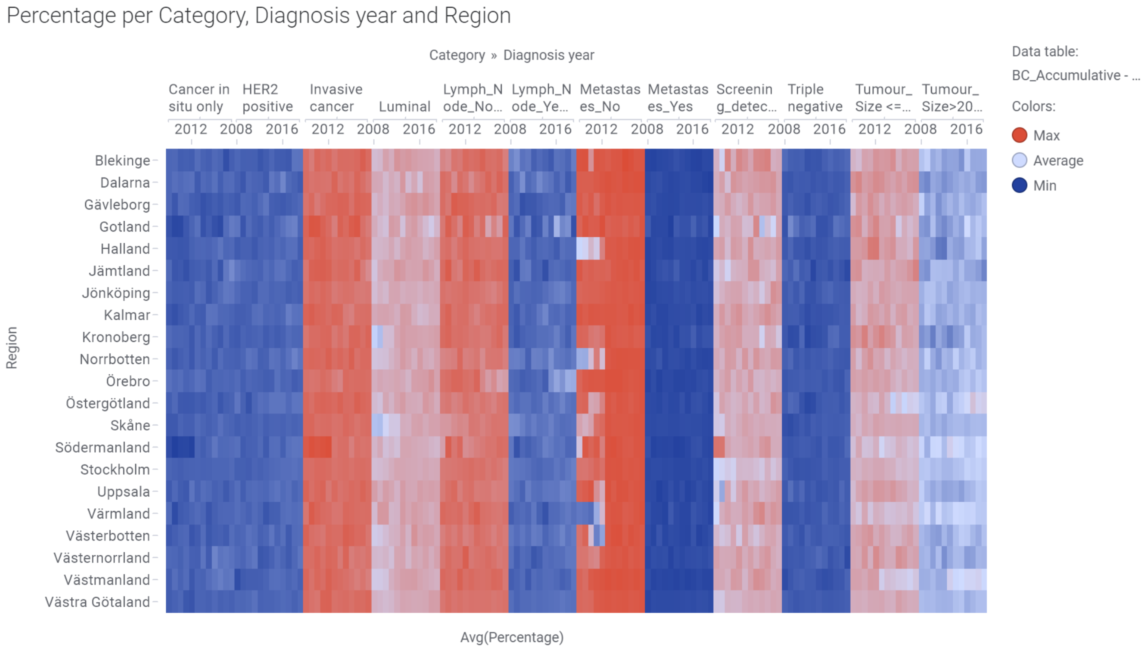


Figure 24: Breast cancer feature relevance

As it can be seen from the above visualization, the combined effects of the breast cancer features on the regions. These results are from having the percentage values instead of the actual values.

5.4 Evaluation

For the evaluations, we planned to perform heuristic evaluation for the suitability of our visualization and usability testing by the user. First, we validated our visualizations using the Nieslson's Usability Heuristics in order to determine its suitability to be used by the user. We based our evaluations on the following heuristics that were applicable to our visualizations from the 10 recommended usability heuristics by Jakob Nielsen.

- **Match between system and the real world:** We have based our visualizations on formats that are easy to understand by the users. These include line charts, bar charts, pie charts, spider charts, heat maps,parallel coordinates and tree maps. As our targeted audience is general users, health professionals and decision makers so we have used these formats because they are simple as well as containing details for the respective users.
- **User control and freedom:** The users have the freedom to navigate the tool particularly by filtering data and viewing visualization as they would desire. They can navigate for the years

and regions and also view results based on the combination of different breast cancer attributes.

- **Consistency and standards:** We have used the same color scheme to depict the regions across the visualizations which makes it easier to follow for the user. Also we have used percentage, average and summation as the standard metrics which again can be easily followed by the users in the visualizations.
- **Recognition over recall:** The users do not have to keep track of the previous details as they navigate through the visualization. The users will be familiarized with the colors used for the regions and also the time frame used when they navigate and therefore will not need to go back to a specific page to look for information.
- **Flexibility and efficiency of use:** The visualization is developed for general users which include women in particular as well as for experts which include health professionals and officials. The cancer terminologies are easy to follow across the visualizations.
- **Aesthetic and minimalist design:** We have tried to use simple design elements and to provide relevant information. Some charts might be a bit complex to comprehend at first but with the ease of use and with an increased eye for detail can be easily understood.

In the initial design, we used visualizations that were not well suited to depict the data. For example we used a bar chart to depict time series data which did not work well. So we had to change it when we evaluated our tool and also from recommendations by the supervisors.

After the heuristic evaluation, we designed a user usability study through a survey in order to test the usability of our visualization. The full set of questions for the usability testing in the survey are found in appendix B.

We received 15 responses from our usability testing survey. The results of the survey were mostly similar to what we had expected with some results not according to our expectations.

From statistical survey investigation of breast cancer on age, gender, mortality and new cases to get the following result.

Nearly 53.3% of participants are aware that men are affected by breast cancer, but some (33.3%) of people are not aware that men are affected by breast cancer and the remaining (13.3%) participants don't think mens affected by breast cancer.

We used a combination chart for identifying which age groups are more affected by breast cancer, so all survey participants understood the visualization graph to answer 100%. The participants simply identified the three regions that are more affected by breast cancer for using the same visualization method of combination chart. The two mostly affected regions are the participant almost 80% answered and the one region 100% answered by the participant.

The line chart is good for time series data visualization because all of the survey participants fully understand the visualization of the mortality rate across Sweden decreasing over the years graph chart.

Out of the 14 respondents, 12(85.7%) of them responded correctly to the question related to the prediction of patients in 2022, whereas 2 (14.3%) them incorrectly missed the prediction.

For the visualization and identification of the most common biological subtype in Stockholm region 12(85.7%) of the respondents correctly answered luminal, 1(7.1) didn't answer and another 1(7.1%) person answered as HER2 Positive, which is wrong.

From the individual care plan and contact nurse all of the respondents observe easily that the individual care plan is increasing.

In the oncological analysis of the parallel coordinates, 12(85.7%) responded correctly and 2(14.3) of the respondents are not responding to our visual.

From the treemap, the number of breast conserving increments is identified by 6(40%) respondents, 4(26.7%) respondents didn't identify, 5(33.3%) of the respondents didn't respond to this visual output.

About patients' predictions, many people gave the right answers to the questions and only some people were incorrect about that. Many people correctly respond to the question of an increase in diagnosis over the years.

For the question relating to compliance of screening detected breast cancer across Sweden, 60% of the respondents answered correctly, with 6.7% wrong answers and 33.3% of the respondents cannot answer from the visualization.

For the question relating to compliance of screening detected breast cancer for Stockholm, 53.3% of the respondents answered correctly, with 26% wrong answers and 20% of the respondents cannot answer from the visualization.

For the question relating to the attribute with high precedence across Sweden, out of the 13 responses, since it was an answer to enter manually, 9 respondents wrote the correct answer but with different wordings, 3 had wrong answers and 1 respondent cannot answer from the visualization.

For the question relating to chemotherapy compliance only one respondent could not answer the question. All the others had one or both answers correct.

For the question relating to chemotherapy compliance, 80% of the respondents answered correctly, with 13.3% answering wrongly and 6.7% did not answer.

The results from the evaluation of the usability study show that most of our visualizations were easy to follow by the users. For some questions where the users cannot answer, these visualizations need to be assessed. We could have given the option to users to write about what difficulty they were having when they were not able to tell.

The usability testing responses when concluded, show positive outlook for the visualization.

6. Discussion

Data visualization and visualization technique is a key tool to data analysis, feature selection, the progression of the disease like breast cancer disease. In this study data visualization has a lot of benefits and applies to many applicable areas like identifying which age groups are more attentively to medical treatment diagnosis(e.g. 50-59, 60-69, 70-79) and which age group needs more awareness diagnosis(e.g. below 40 age). Then which regions are mostly affected by breast cancer and focus on more affected regions to detect early stages. It is better to use for women or men whose age is a very important medical treatment , flow up and awareness.

Data visualization tools used in medical diagnosis like mammography screening are one part of visualization techniques.In general data visualization is very important for data mining, machine learning and other applicable areas.

Some of the results were as expected but data visualization helped them to understand better i.e. Stockholm, Skåne and Västra Götaland are the most populated regions across Sweden so breast cancer incidences were expected to be higher in these regions. But some of the statistics like Stockholm being non compliant with the target levels for screening detected breast cancer were unexpected given that Stockholm is the capital and would probably have more resources than other regions in order to make the concerned people to get screened for breast cancer.

Also the tumour size detected at diagnosis is high for some regions, which shows that the cancer was not diagnosed early. The use of visualization has signified this aspect which is difficult to assess from numerical data in tabular form.

In health center data visualization is a very important tool to assess the large databases to use in identifying important patterns from the large dataset for machine learning methods. Clinical researcher to use visualization tools to see the progression of the disease patterns.

The results mainly shows that most of the users easily understand the visual outputs and answer correctly and eases in understanding the complex relationship of different attributes.

6.1.Limitation and Challenges

As the dataset we used had many different attributes, given more time we could have analyzed the combination of these attributes to each other. Also since most of the data is of numerical value, although we did manage to convert some of it to percentage, it would have been better to analye if all the data was converted to percentage values.

Since we were not familiar with Spotfire to start with, learning it was a challenge. As we have managed to use it better now, given more time we would have utilized it in an effective way, e.g. using the unpivot feature was beneficial for our analysis.

Also we could have investigated with different visualizations for most of the research questions.

The challenge that we face is mainly the difficulty of integrating statistical data.

7. Conclusion

In this project, the data was collected manually and it was not real time data. The data is Swedish breast cancer data and this data collected from published Swedish national registration . The data managed by extracting important features or variables for this project from a massive amount of data. Those data are divided by different parts based on the problem the study like the form of by region, Age group, year and others. The data analysis tools used spotfire data visualization tools gave this good performance to visualize Multidimensional data.

In Sweden the three regions **Stockholms län**, **Skåne län,** and **Västra Götalands län** are most of the time the cases (new case, Diagnosis and mortality) are higher than the remaining regions. For Swedish women get better medical treatments such as mammography screening, follow-up and awareness. The death rate of women and men vary from one region to the other. Some regions have the death rate of men higher than women, and also some of them have no death rate of men but, the death rate of women is higher. The incidence of breast cancer in women is much higher than men.

The breast cancer patients individual care plan is increasing. The visualization can be used as a main decision making input to different managers and government officials if the research project extended and realized.

The issue we are trying to solve is to use statistical analysis and visualization to emphasize the importance of recognizing breast cancer.

We are presenting visualization that illustrate its significance, beginning with an overview and then focusing on diagnostics, treatment mechanisms, and survival rates using statistics.

From the project we gained knowledge about the use of visualizations for problems such as breast cancer understanding. Most people might just know some basic information regarding a problem. Visualizing data helps to uncover hidden features that cannot be analyzed from textual data. As we have seen from this project that some results were different from ones perceived from before e.g. we were not sure that men can be diagnosed to have breast cancer.

Also we think that we could have made the visualization more interactive for the users, and let them analyze the data based on their preferences.

References

[1] iarc.who.int (2021). *World Cancer Day: Breast cancer overtakes lung cancer as leading cause of cancer worldwide. IARC showcases key research projects to address breast cancer – IARC.* [online] Available at: https://www.iarc.who.int/news-events/world-cancer-day-2021/ [Accessed 7 February 2021].

[2] W. So, E. P. Bogucka, S. Scepanovic, S. Joglekar, K. Zhou, and D. Quercia, "Humane visual AI: Telling the stories behind a medical condition," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 678–688, 2021, doi: 10.1109/TVCG.2020.3030391.

[2] C. Leung, Y. Zhang, C. S. H. Hoi, J. Souza, and B. H. Wodi, "Big data analysis and services: Visualization on smart data to support healthcare analytics," *Proc. - 2019 IEEE Int. Congr. Cybermatics 12th IEEE Int. Conf. Internet Things, 15th IEEE Int. Conf. Green Comput. Commun. 12th IEEE Int. Conf. Cyber, Phys. So*, pp. 1261–1268, 2019, doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00212.

[3] B. C. Kwon *et al.*, "DPVis: Visual Analytics with Hidden Markov Models for Disease Progression Pathways," *IEEE Trans. Vis. Comput. Graph.*, pp. 1–15, 2020, doi:

10.1109/TVCG.2020.2985689.

[4] E. Polychronidou, I. Kalamaras, K. Votis, and D. Tzovaras, “Health Vision: An interactive web based platform for healthcare data analysis and visualisation,” *2019 IEEE Conf. Comput. Intell. Bioinforma. Comput. Biol. CIBCB 2019*, 2019, doi: 10.1109/CIBCB.2019.8791462.

[5] B. S. Santos, S. Silva, and P. Dias, “Heuristic evaluation in visualization: An empirical study : ition paper,” *Proc. - 7th Bienn. Work. Eval. Beyond Methodol. Approaches Vis. BELIV 2018*, pp. 78–85, 2019, doi: 10.1109/BELIV.2018.8634108.

Appendix A: Declaration of individual student efforts and time plan

Time plan

#	Activity/Milestone	Responsible	4	5	6	7	8	9	10	11	12	13			
A1	Proposal Writing	All members	🟨	🟨	🟨										
A2	Data Collection	All members		🟦	🟦	🟦									
A3	Data Preparation	All members				🟨	🟨								
A4	Data analysis	All members					🟦	🟦							
M1	Data Visualization	All members						🟨	🟨						
A5	Evaluation	All members							🟦						
A6	Results and conclusions	All members							🟨	🟨					

Appendix B: The Usability Testing and User Survey

Breast Cancer Data Visualization Survey

This user survey is only used for research purposes and your genuine response will be a lot to our work. The research project mainly focuses on the visualization of breast cancer using the Spotfire.

Do you have knowledge on breast cancer

【 】Yes 【 】No

Is your work related to the health sector(Medical doctor, nurse and related fields)

【 】Yes 【 】No

Research and Visualization Related Questions

What are the predicted women and men breast cancer patients in the year 2022? *

Which age group is most vulnerable to breast cancer? *

【 】0-39

【 】40-49

【 】50-59

【 】60-69

【 】70-79

【 】80 and above

How did the habits (patterns) of breast cancer patient of Individual Care plan change in Sweden? *

Which biological Subtype of breast cancer is common in Sweden? *

【 】Triple Negative

【 】Luminal

【 】HER2 Positive

Which regions are highly affected by breast cancer in Sweden?(List three highly affected) ______________________________

Usability Testing

To evaluate the Usability testing, you will be able to rate the questions below on a scale from 1-5 where 1 means Strongly Disagree and 5 means Strongly Agree.

Learnability, satisfaction, and related questions are stated to rate below:

The visualization is easy to use.

It took me short time to learn the visualization

The visualization is easy to interpret.

I am satisfied with the overall visualization of the system

I found the system unnecessarily complex.

I think that I would need the support of a technical person to be able to use this system.

I found the various functions in this system were well integrated.

I thought there was too much inconsistency in this system.

I found the system very cumbersome to use.

I felt very confident using the system.

I needed to learn a lot of things before I could get going with this system.