Eddieb Sadat
Professor Zadbood
EM622-WS Decision Making via Data Analysis
10/19/22

Homework 3

Introduction:

Data visualization has been utilized in research to identify patterns in data that may have gone unnoticed through quantitative analysis alone. This is no different in the medical field where many different biological processes have correlations to each other. In disease research, identifying these relationships can help develop solutions to cures and mitigations. This is no different for breast cancer research, which will be the focus for this project. Using the Breast Cancer data from UCI's machine-learning archive[1], an advanced visual plot will be developed to better understand the relationship between number and location of tumors on treatment outcomes. This report will follow the CRISP-DM methodology.

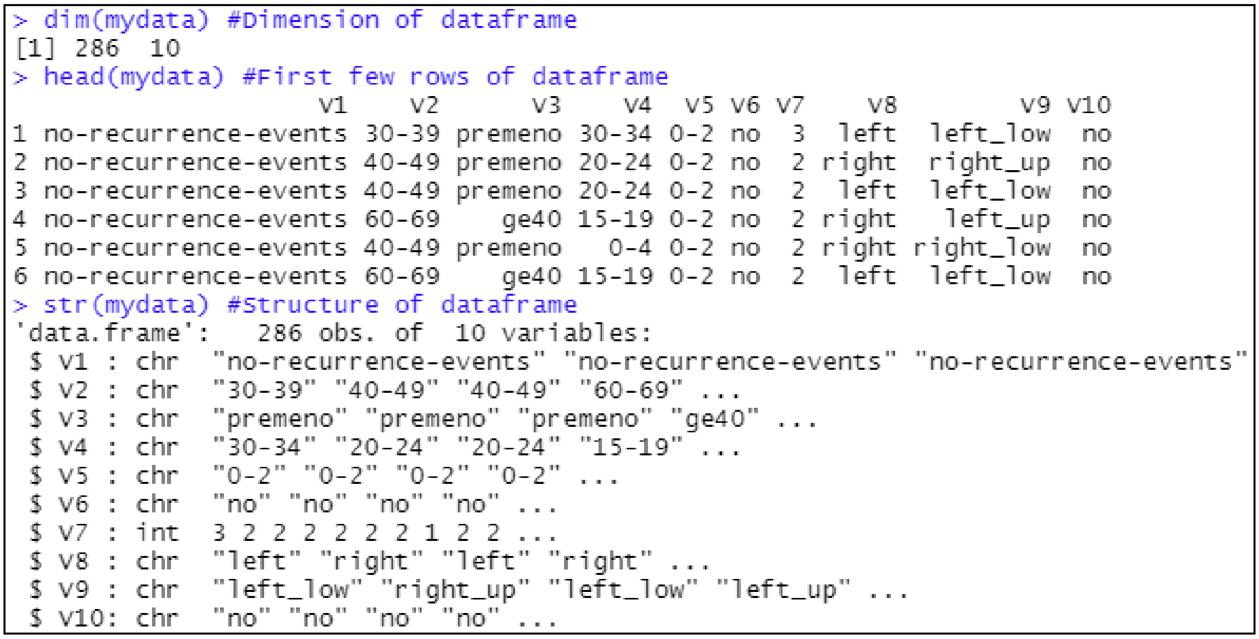
Business Understanding:

This project is being made to complete the Homework 3 assignment for EM-622, with the primary objective of answering the question, "Is there a relationship between the number of tumors and location of tumors on the outcome of treatment?". The other requirements are to use RStudio and the ggplot2 library to develop an advanced plot (specifically replicating the example shown on slide 25 of the Week 6 presentation[2]) and to document the process using the CRISP-DM methodology. To successfully complete this project, the following steps must be taken:

1. Analyze and understand the Breast-Cancer Data
   a. Utilize the UCI archive site to identify column data
   b. Use RStudio to explore the data even more
2. Identify the columns required for the visual
   a. Filter out the other columns
3. Check the integrity of the remaining columns
   a. Look for NA's and fix if needed
   b. Explore column distributions
4. Create the scatter plot
   a. Modify to replicate the given example
5. Develop an answer to the question based on the plot

Data Understanding:

First, the data is uploaded into the RStudio script where the dimension, the first few rows, and the structure are examined. From the figure below, the dataset contains 286 rows and 10 columns. Each column is named V1-V10, where all columns are character data types, except for V7 which is an integer data type.



```
> dim(mydata) #Dimension of dataframe
[1] 286  10
> head(mydata) #First few rows of dataframe
                    V1    V2      V3    V4  V5 V6 V7    V8        V9 V10
1 no-recurrence-events 30-39 premeno 30-34 0-2 no  3  left  left_low  no
2 no-recurrence-events 40-49 premeno 20-24 0-2 no  2 right  right_up  no
3 no-recurrence-events 40-49 premeno 20-24 0-2 no  2  left  left_low  no
4 no-recurrence-events 60-69    ge40 15-19 0-2 no  2 right   left_up  no
5 no-recurrence-events 40-49 premeno   0-4 0-2 no  2 right right_low  no
6 no-recurrence-events 60-69    ge40 15-19 0-2 no  2  left  left_low  no
> str(mydata) #Structure of dataframe
'data.frame':   286 obs. of  10 variables:
 $ V1 : chr  "no-recurrence-events" "no-recurrence-events" "no-recurrence-events"
 $ V2 : chr  "30-39" "40-49" "40-49" "60-69" ...
 $ V3 : chr  "premeno" "premeno" "premeno" "ge40" ...
 $ V4 : chr  "30-34" "20-24" "20-24" "15-19" ...
 $ V5 : chr  "0-2" "0-2" "0-2" "0-2" ...
 $ V6 : chr  "no" "no" "no" "no" ...
 $ V7 : int  3 2 2 2 2 2 2 1 2 2 ...
 $ V8 : chr  "left" "right" "left" "right" ...
 $ V9 : chr  "left_low" "right_up" "left_low" "left_up" ...
 $ V10: chr  "no" "no" "no" "no" ...
```

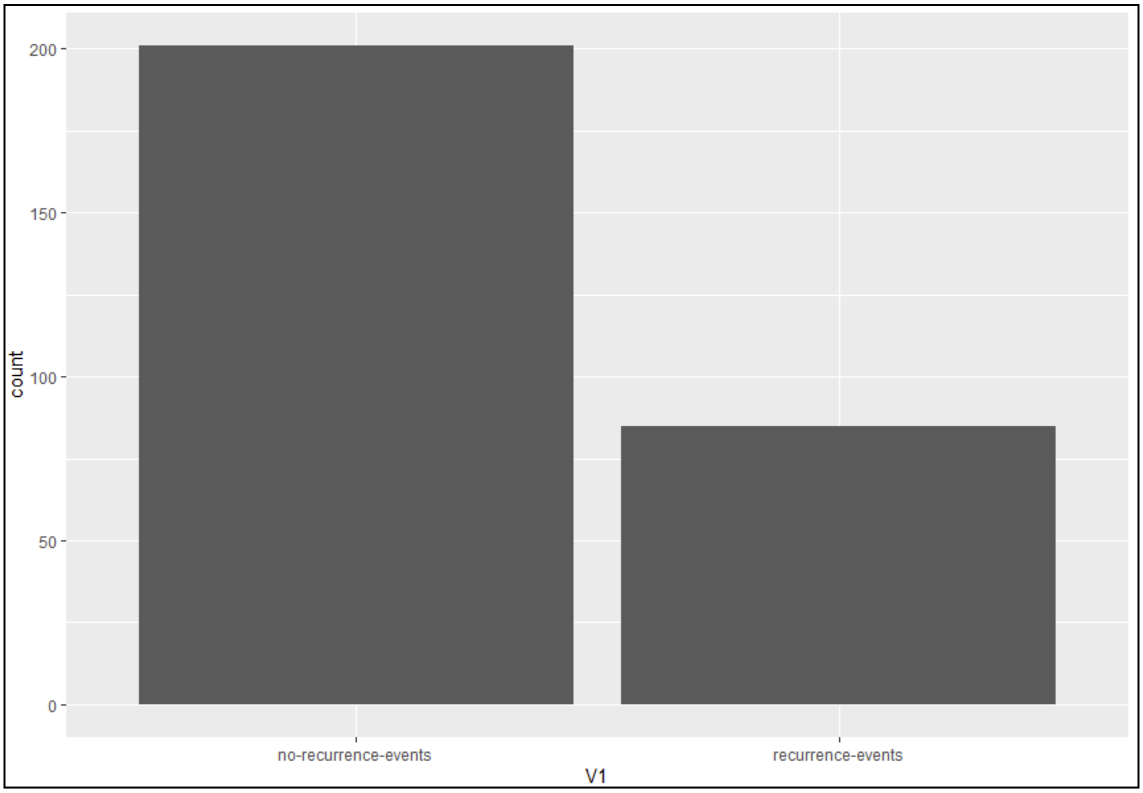
The UCI archive[1] contains information on what each column represents (The UCI page lacks detailed explanations of the data columns. A quick search found a research paper[3] with the same variables that includes better explanations of the variables):

- V1: Does the cancer come back? → Yes = recurrence-events, No = no-recurrence events
- V2: Age → 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
- V3: Menopause → lt40, ge40, premeno
- V4: Tumor-Size (in mm) → 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
- V5: Inv-Nodes (Number of lymph nodes showing Breast Cancer[3]) → 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
- V6: Node-Caps (Does the tumor penetrate the lymph node caps?[3]) → yes, no
- V7: Deg-Malig (Degree of malignancy[3]) → 1, 2, 3
- V8: Breast (which breast) → left, right
- V9: Breast-Quad (where on the breast) → left-up, left-low, right-up, right-low, central
- V10: irradiat (Is radiation treatment administered?) → yes, no.

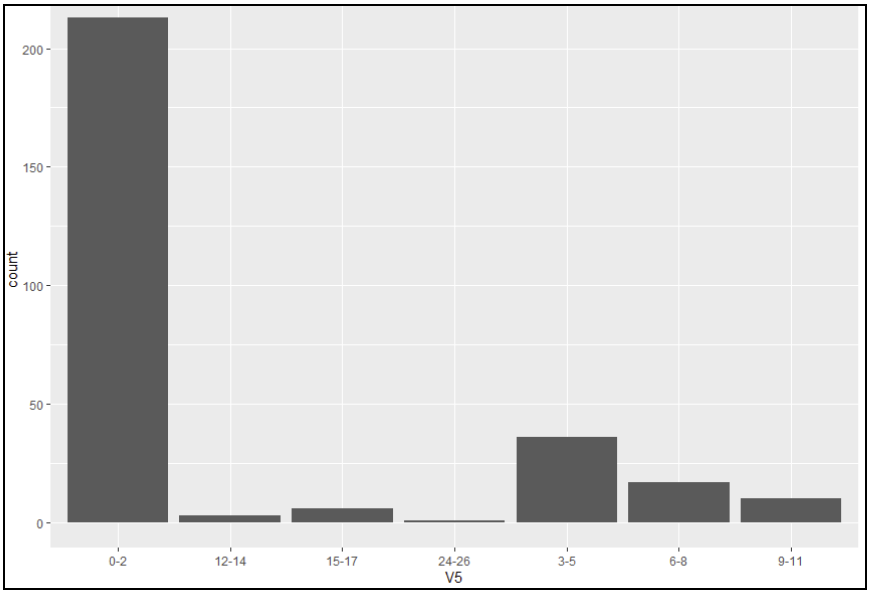
Since the objective is to identify the relationship between number and location of tumors, the columns that will be used are V5 (Number of lymph nodes showing cancer), V8 (Which breast, left or right, are the tumors located), V9 (The location on the breast where the tumors are located), and V10 (Whether radiation treatment is administered). As an extra piece of information, V1 (Recurrence, yes or no, of the cancer) will be included in the visual as well.

One thing to note is that the example on slide 25 of the week 6 presentation creates a scatter plot using V3 (Menopause) as the x-axis values. However, I believe that this variable is unrelated to answering our question, so I will replace the x-axis with V8 (Which breast, left or right, the tumors are located). Similarly, the example uses V4 (Tumor Size) as the y-axis; It will be replaced with V5 (Number of lymph nodes displaying breast cancer).

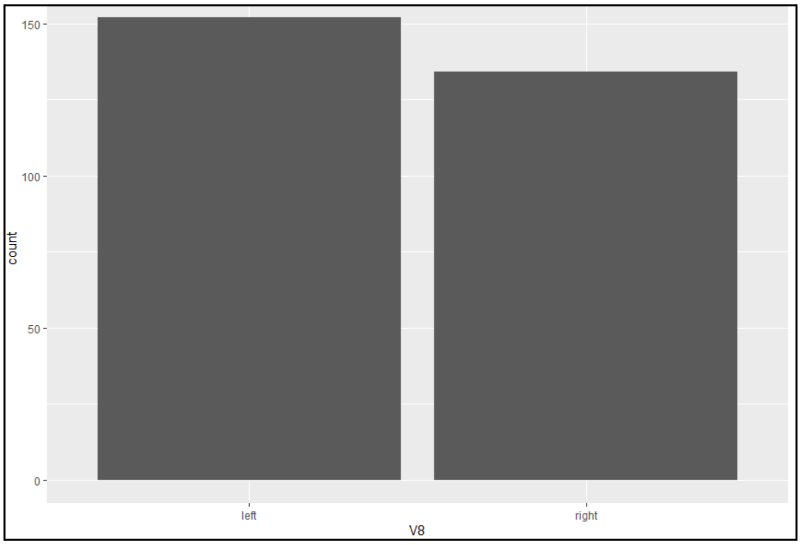
A quick look at the distributions of each of these columns (V1, V5, V8, V9, V10) is useful to quickly see how the data set is represented by each category. Since all of these columns are categorical, a barplot will be used to display their distributions.



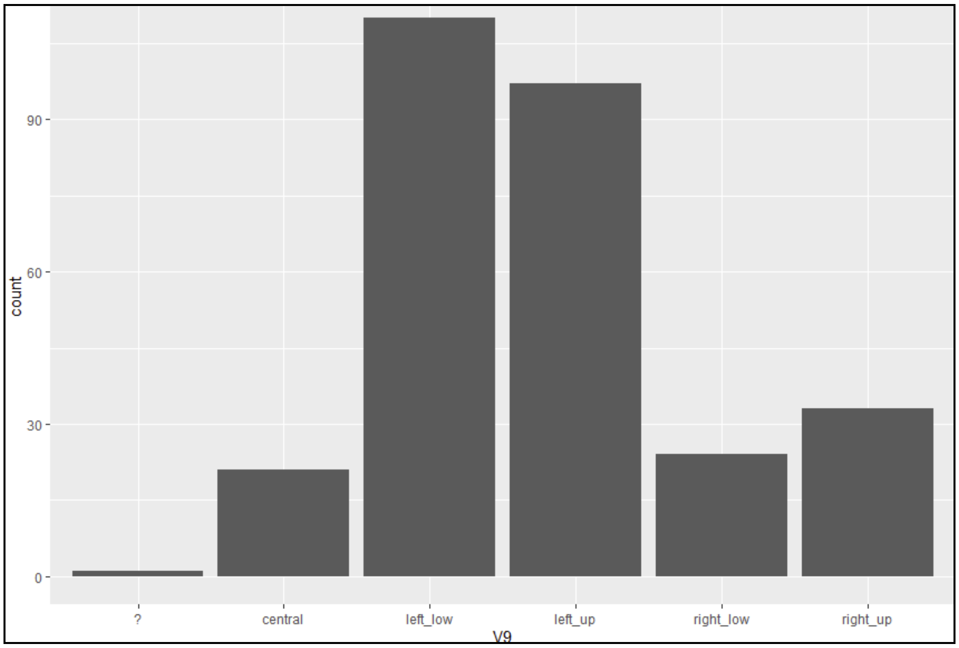
The figure above is a bar graph of V1, and it is clear that no-recurrence-events (~65% of the data) is more represented than recurrence-events (~35% of the data). For the purposes of this assignment, this is a perfectly acceptable distribution, and there are likely enough data points in both categories to display a pattern, if it exists. There appears to be no bar labeled ‘?’, which indicates this column is likely void of NA’s.



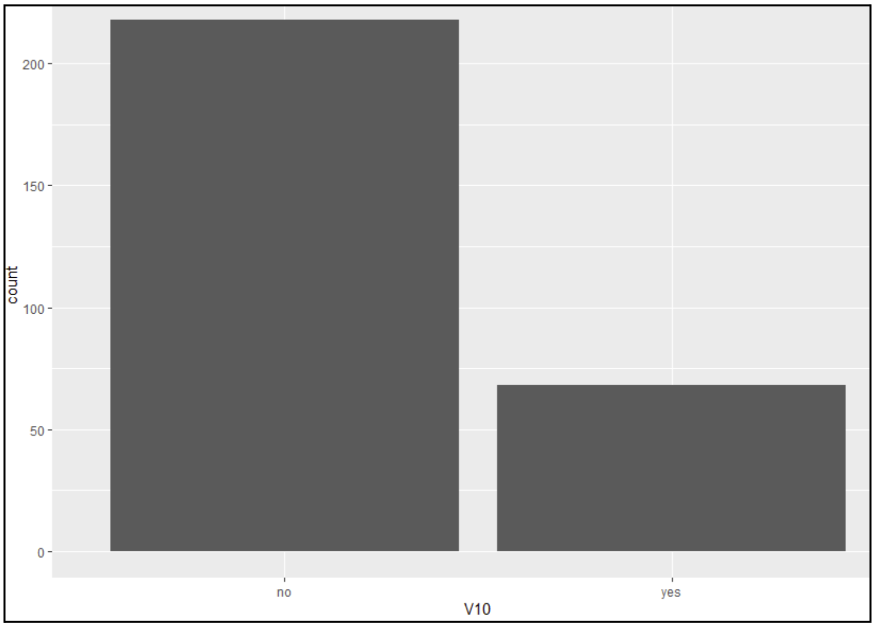
The figure above is a bar graph of V5, and it is clear that an overwhelming portion of the data is of 0-2 lymph nodes that show cancer. In other words, there is a clear over representation of 0-2, and a severe underrepresentation of all other categories, with some of them noticeably negated from the graph, indicating that there are no data points with those values (e.g. there are zero points with 36-39). It is important to keep in mind that this severe representation discrepancy may make it very difficult to identify relationships by number of tumors for categories other than 0-2. In other words, the results of this report may not yield a realistic conclusion. There is no bar labeled '?', indicating that there are no NA's in this column.



The figure above is a bar graph of V8, and it is clear that there is a relatively equal number of data points for left and right breasts. This is perfectly acceptable for this project, though, it may indicate that recurrence/treatment is independent of which breast (which makes sense, logically). There is no bar labeled '?', indicating no NA's in this column.



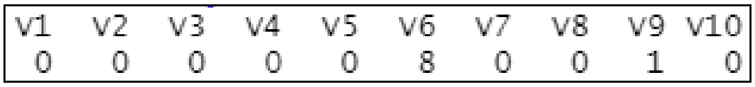
The figure above is a bar graph of V9, and it is clear that the left_low and left_up categories have significantly more points than the rest. However, each category has enough points that it is perfectly acceptable for this assignment, and it will likely be enough to display any patterns, if any exist. There is a bar labeled '?', which means there are NA values that need to be addressed.



The figure above is a bar graph of V10, and it is clear that 'no' dominates most of the data points (~70% of the data). Although 'yes' has significantly less representation, this column is perfectly healthy because there are likely still enough data points to show a pattern, if it exists. There is no bar labeled '?', indicating no NA's in this column.

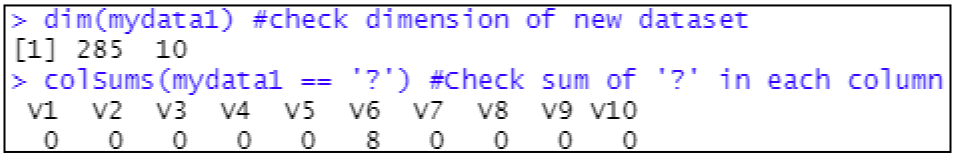
Data Preparation:

From the bar graphs, there was an indication that at least one of the variables contains NA values. To address this, the number of NA's per column is first identified.



```
 V1   V2   V3   V4   V5   V6   V7   V8   V9  V10
  0    0    0    0    0    8    0    0    1    0
```

From the figure above, the only column (within the scope of this project) that contains NA's is V9, and there is only one NA. Because these rows are categorical, it would require a predictive model to determine what category can best replace the NA. However, since this method is out-of-scope for this assignment (and frankly, the effort is not worth the reward), the row containing that NA will be removed.



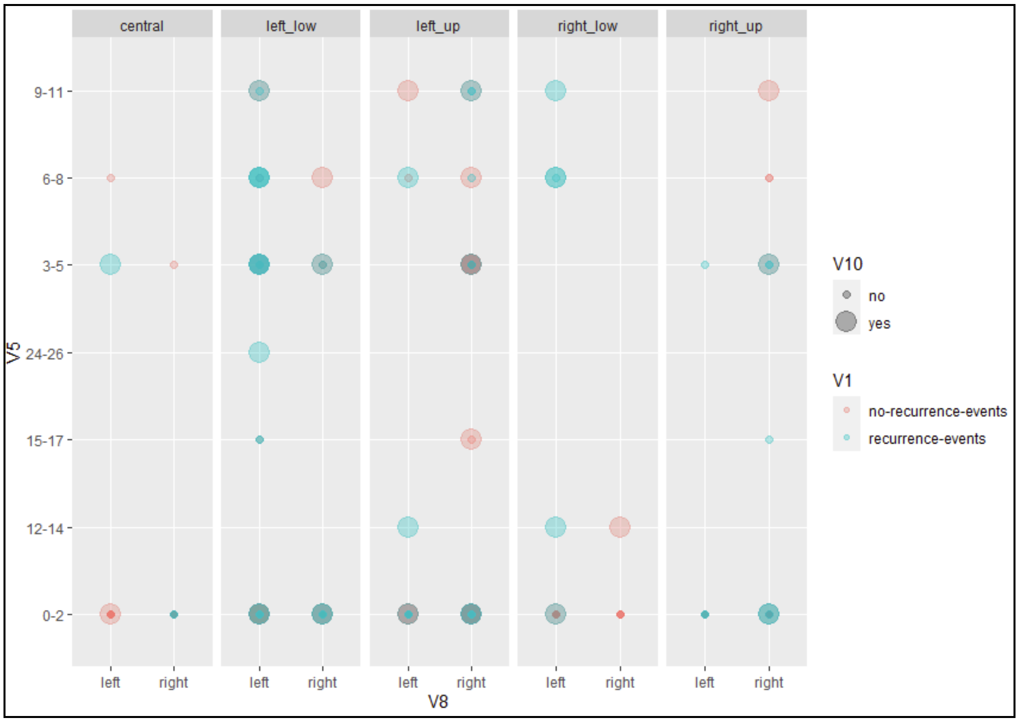
```
> dim(mydata1) #check dimension of new dataset
[1] 285  10
> colSums(mydata1 == '?') #Check sum of '?' in each column
 V1  V2  V3  V4  V5  V6  V7  V8  V9 V10
  0   0   0   0   0   8   0   0   0   0
```

After removing the row, the dataset now contains 285 rows with no NA's in any relevant column. As such, the data is ready to be used for visualization.

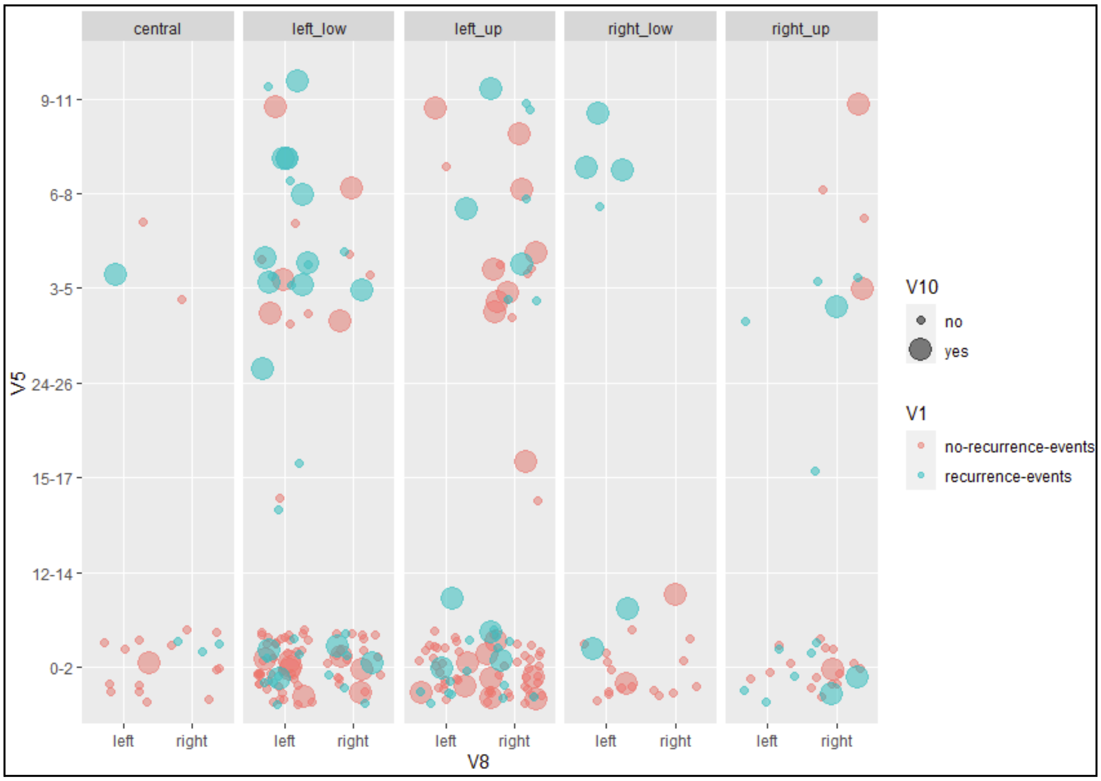
Modeling:

Choosing the correct visual that best displays the relationships between each variable is important. Because this report is replicating the example displayed on slide 25 of the week 6 presentation, it will simply use the same visualization - an advanced scatterplot using ggplot. This method will nicely display the relationship between tumor location (breast and breast quadrant) and tumor quantity to treatment (radiation) and recurrence.

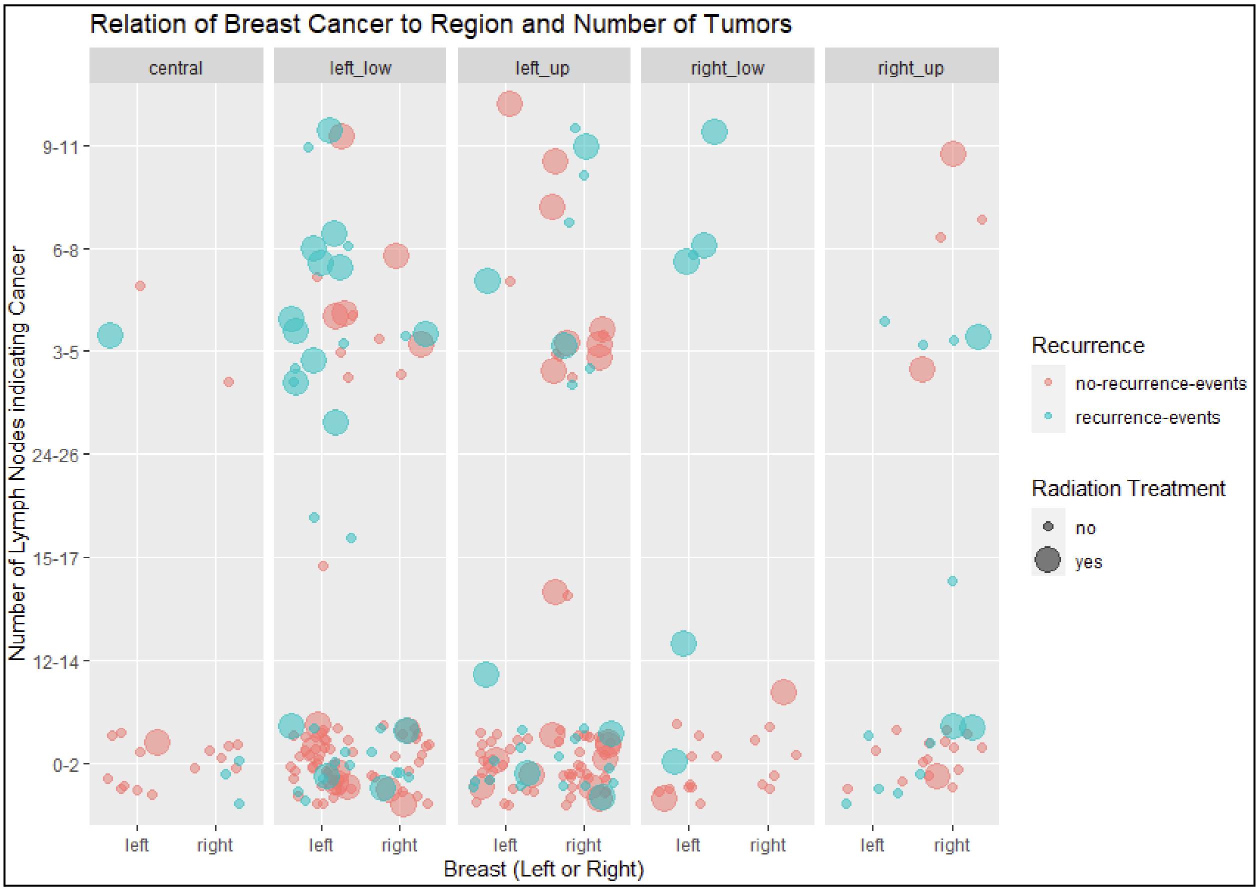
V8 (left/right breast) was used as the x-axis, V5 (number of lymph nodes showing cancer) was used as the y-axis, V9 (breast location) was used as the faceted columns, V1 (recurrence) was used with a color differentiator, and V10 (radiation) was used with a size differentiator. However, this graph was very difficult to read because many of the data points overlap, even when an alpha is placed to make the points transparent.



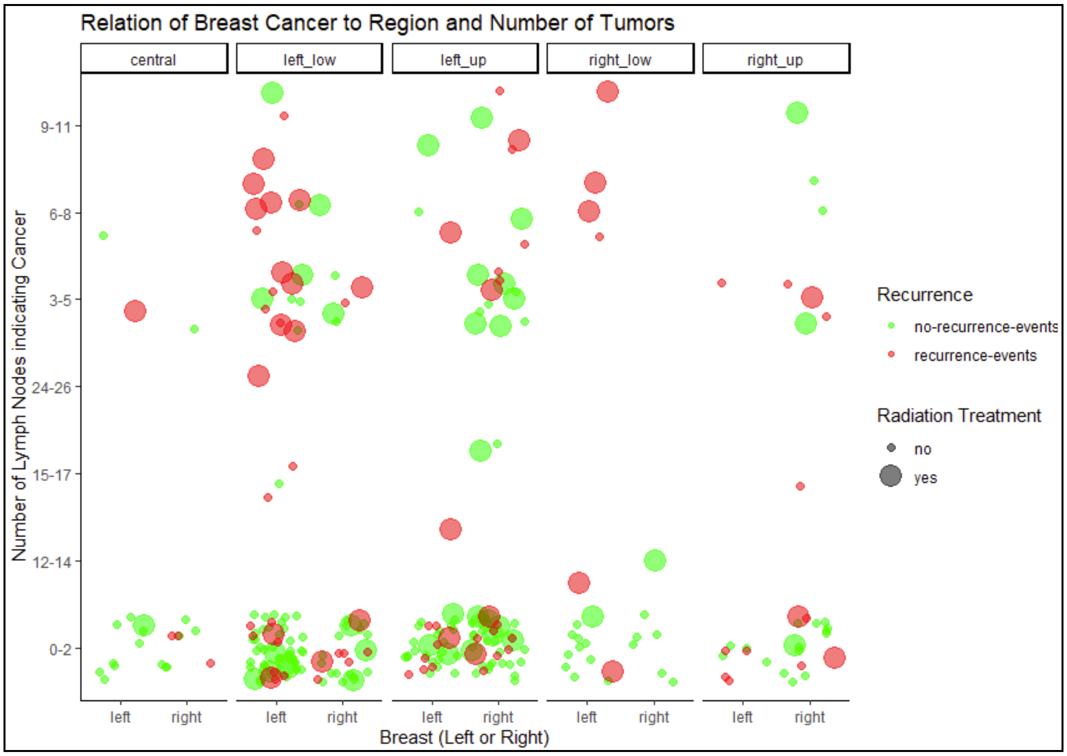
After some research, a technique called jittering can be applied to prevent overlapping of points.[4]



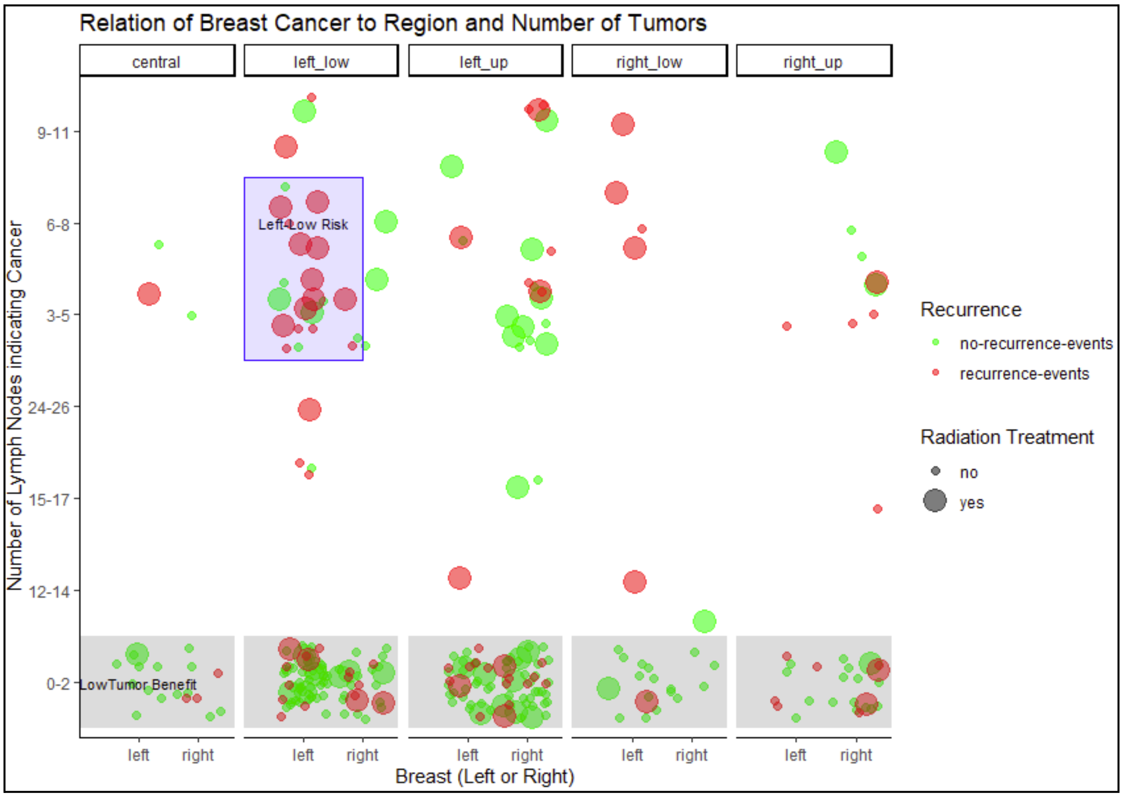
Next, the labels need to be changed to make it more clear what each axis represents. This was difficult, because no matter what I tried, I could not change the legend labels since they were discrete (I guess R did not like that). I employed the guide() method to change the legend titles[5]. I also changed the axis labels, and added a title. Below is a figure of the final graph generated.



The color of the scatterplot points is changed using scale_color_mannual[6]. Additionally, to imitate the example graph, a theme_classic() is applied to apply a white background with no grid lines[7].



Next, I added text to individual facets[8] to identify areas of interest, and created rectangles to also highlight those areas[9].



Evaluation:

The data representation is a little misleading because of the overrepresentation of the number of lymph nodes 0-2, and a severe underrepresentation of all other data points.

According to this visual, it generally seems that cancer recurrence is less likely to occur when there are less tumors; This is concluded by looking at the ratio of red points to green. There are higher ratios of green points for people with 3-5, 6-8, 12-14, and 15-17 lymph nodes indicating cancer, whereas the 0-2 category has a higher red ratio. However, recurrence seems to be fairly equal across all breast locations and quadrants, which may indicate that recurrence has low correlation to tumor location.

It seems that radiation treatment occurs for people with tumors located in the left_low and left_up quadrants of the breast, indicating that these two locations may have a strong relationship to receiving radiation treatment. This is indicated by the ratio of big circles to small circles in the columns. These ratios seem equal among different tumor numbers and in the left/right breasts, indicating low correlation between number of tumors and radiation treatment, and between which-breast and radiation treatment.

Deployment:

A decent conclusion was able to be developed from the data and the advanced scatterplot, however, they should be taken with a grain of salt considering the discrepancies in data representation. For the purpose of this assignment, it was successful because it demonstrated the power of advanced plots in ggplot, and how useful they can be when trying to identify relationships among multiple variables. It would not be recommended to use this data for professional or industry level work.

Sources:

[1] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer
[2] https://sit.instructure.com/courses/62354/files/10372973?module_item_id=1634191
[3] http://dx.doi.org/10.18203/2394-6040.ijcmph20202994
[4] https://ggplot2.tidyverse.org/reference/geom_jitter.html
[5] https://www.geeksforgeeks.org/how-to-change-legend-title-in-ggplot2-in-r/
[6] http://www.sthda.com/english/wiki/ggplot2-colors-how-to-change-colors-automatically-and-manually
[7] https://www.datanovia.com/en/blog/ggplot-theme-background-color-and-grids/
[8] https://stackoverflow.com/questions/11889625/annotating-text-on-individual-facet-in-ggplot2
[9] https://r-graphics.org/recipe-annotate-rect