Sara Aljamal

Data visualization technical report

|  |  |
| --- | --- |
| Which data visualizations should you choose ? Dashboards, maps... |  |

Table of Contents

[**1** **Fundamentals of Data Visualization** 2](#_Toc168386904)

[1.1 What is the lie factor, and how does it impact Tufte’s graphical integrity rules? 2](#_Toc168386905)

[1.2 What is the lie factor in the following plot considering the values below (actual values) and that the volume of the corresponding cylinder represents values in graphics? 2](#_Toc168386906)

[1.3 What issues do we face when using unjustified 3D plots? 3](#_Toc168386907)

[1.4 Explain the concept of chart Junk and what are its advantages and disadvantages? 3](#_Toc168386908)

[1.5 What is data ink ratio, and should it be maximized or minimized? 4](#_Toc168386909)

[1.6 Describe what semantics can be understood from the following graphical codes. 4](#_Toc168386910)

[1.7 Describe visual association and semantic association and how they would help in data visualization. 5](#_Toc168386911)

[**2** **Techniques of Data Visualization** 5](#_Toc168386912)

[2.1 What are the marks and channels, and number of attributes encoded in the following two plots. 5](#_Toc168386913)

[2.2 Discuss the importance of using the following interaction techniques in data visualization: selection, change over time, navigation, filtering, brush and zoom, brush and link, and aggregation. 6](#_Toc168386914)

[2.3 Mention 4 interactive visualization techniques used in Google Maps. 10](#_Toc168386915)

[**3** **Static and Interactive Visualizations (project)** 11](#_Toc168386916)

[3.1 Provide a detailed description of each of the datasets you selected, and the attributes of interest, and the attributes you would like to visualize (data abstraction). 11](#_Toc168386917)

[3.2 Describe why you are looking at the data (task abstraction), and design rules for each plot you generated in R. 13](#_Toc168386918)

[3.3 Describe the encoding methods for each plot (i.e., choosing the plot based on data and tasks, and marks and channels etc.). 13](#_Toc168386919)

[3.4 Assess the methods (e.g., proper design and encoding, Tufte’s rules, Schneiderman Mantra, etc.) used to build your visualizations and how they helped (static and interactive). 13](#_Toc168386920)

[3.5 Analyze and assess the insights and findings for each plot you generated in R. 13](#_Toc168386921)

[3.6 Critically evaluate how the project visualizations impact the organization and decision-making. 28](#_Toc168386922)

[**4** **Communication of Results and Findings** 28](#_Toc168386923)

[4.1 Prepare PowerPoint slides to demonstrate the use of data visualization and provide insights and knowledge. (slides and presentation) 28](#_Toc168386924)

[4.2 Provide a presentation to communicate your findings and insights (slides and presentation) 28](#_Toc168386925)

[4.3 Evaluate how good your project visualizations can tell the story and provide insights to the organization. (report) 28](#_Toc168386926)

[**5** **References** 30](#_Toc168386927)

# **Fundamentals of Data Visualization**

* 1. ***What is the lie factor, and how does it impact Tufte’s graphical integrity rules?***

The lie factor is concept was introduced by Edward Tufte in his book “The visual display of quantitative information” in which he introduced some principles and rules to ensure graphical integrity when visualizing data and information. Tufte stated that “The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured.”, This highlights the significance of the lie factor, which is a measure that assesses whether a visualization accurately conveys the data it represents. The lie factor can be calculated through the following formula:

Lie factor= Size of effect shown in graphic/ Size of effect in data

A value very close to or equal to 1 indicates high accuracy, which means the visual representation is very close to the actual data, whereas a value less than 1 indicates understating, which means we are attempting to show less difference than the difference that exists in the actual data, and a value greater than 1 indicates overstating, which means we are attempting to manipulate the visualization to show a larger difference between values than the actual difference. Both understating and overstating go against Tufte's rule and are considered distortions of the actual data.

Through the lie factor, we quantify the difference between the visual representation and the actual data in order to reach the perfect value, which is 1. By doing so, we ensure graphical integrity, honesty, and clarity, avoiding manipulation and lying to the audience, which supports and aligns Tufte's graphical integrity rules, so we can say that the lie factor takes our visualizations a step forward in applying Tufte's graphical integrity rules for better visualization.

## *A cylinder and cylinder diagram Description automatically generated****What is the lie factor in the following plot considering the values below (actual values) and that the volume of the corresponding cylinder represents values in graphics?***

Lie factor= Size of effect shown in graphic/ Size of effect in data

1. Size of effect shown in graphic (image):

* First cylinder volume= r^2\*h\* π= (1.5)^2\*5\* π= 35.34
* Second cylinder volume= r^2\*h\* π= (1)^2\*2.5\* π= 7.85
* Size of effect shown in graphic (image) = (35.34-7.85)/7.85= 3.5=3.5%

1. Size of effect shown in data:

* First cylinder real volume= 3
* Second cylinder real volume= 1
* Size of effect shown in data = (3-1)/1=2=2%

1. Lie factor = 3.5%/2%= **1.75** (>1, which indicates overstating).
   1. ***What issues do we face when using unjustified 3D plots?***

As Tufte stated that “The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.”, adding a 3rd dimension to a plot when it’s not needed or unnecessary introduces several issues such as:

1- Occlusion: When adding a third dimension, the positions of data points will change, which may cause overlapping or hiding some points, lines, or regions of the plot affecting its clarity as it results in an incomplete view for the viewer.

2- Lie factor: The addition of a third dimension for a plot means adding one more dimension for the represented data (depth), which affects the appearance of the points or lines or the relationship between them, so it will exaggerate or diminish the appearance of datapoints in a way that violates Tufte's graphical integrity rules and will introduce a lie factor that is significantly greater or lesser than one.

3- Increased complexity: More dimensions mean more details, and when data can only be represented in two dimensions, adding one more dimension adds unnecessary details and information that overwhelms the viewer and makes capturing and extracting insights and relationships more challenging and complicated, which goes against the goal of visualization of trying to make things easier to understand.

4- Accuracy reduction: Dimensionality has a strong effect on accuracy, as the comparison between datapoints gets more complex and harder, especially when the datapoints are close to each other, the difference might not be noticeable. In addition to the lie factor and occlusion issues noted above, the overall plot becomes more difficult to understand, thus diminishing its accuracy and hindering effective data interpretation.

## ***Explain the concept of chart Junk and what are its advantages and disadvantages?***

Chart Junk refers to any irrelevant ink or element that is included in a plot which adds zero new information about the data, they are elements that may confuse or distract the viewer from the real data, for example, adding unnecessary grid lines, colors, 3D effects, shadows, decorations, and icons.

***Chart Junk advantages:***

* Chart Appearance: Adding elements or visual embellishments related to the chart topic results in a better, and engaging appearance for the plot.
* Memorable: Chart Junk allows for adding special items that could be easily remembered and memorized by humans rather than numbers or basic charts.
* Allow for branding and customization: With the addition of Chart Junk, we might customize the charts as we want in a way that benefits the goal of that chart, as well as adding brand- specific design elements related to the brand or organization.

***Chart Junk disadvantages:***

* Reducing data ink ratio: The addition of Chart Junk reduces the data ink ratio (The ratio of the ink used to deliver information and is directly related to the data being represented) which we try to maximize for better visualizations according to Tufte.
* Clutter: Chart Junk causes clutter in the chart causing increased complexity and viewer distraction which makes it challenging for the viewer to extract the wanted information from the crowd.
* Reduce accuracy: Adding unnecessary elements leads to misinterpretation in some cases and reduces the whole chart accuracy as the data ink ratio is reduced and the complexity is increased.

Although Chart Junk might add clutter and complexity to a visualization, it can also improve its appearance so it's not always a bad thing, however, it all depends on the audience, task, and context of that chart.

* 1. ***What is data ink ratio, and should it be maximized or minimized?***

The data ink is the amount of ink required to represent the data in a chart, while the total ink is the total amount of ink used to represent the whole chart including the decorations, grids, and so on. The data ink ratio is the ratio of the data ink over the whole ink used in the chart.

For better charts and visualizations, we should maximize the data ink ratio and minimize or even erase the non-data ink ratio, because the data ink ratio is what we are trying to show in charts, so we need to maximize it as it indicates that large proportion of the chart is used to represent the actual data resulting in better and comprehensive understanding, as well as efficient information delivery. Furthermore, according to Tufte, to achieve graphical excellence we should show data above all, maximize the data ink ratio, erase non-data ink and redundant data ink.

* 1. ***Describe what semantics can be understood from the following graphical codes.***

1. ***Nested regions and partitioned regions:***

Objects or regions that seem nested indicate some kind of hierarchical concepts or ideas.

1. ***Attached shapes:***

When shapes are attached together it conveys that they are parts of conceptual structure.

1. ***Graphical objects in proximity:***

Graphical objects that are seen close to each other after are understood as related information or similar concepts.

1. ***Shapes enclosed by a contour:***

When there are shapes that look like they are enclosed by a contour usually they indicate a sort of connection, association, or relation between them.

* 1. ***Describe visual association and semantic association and how they would help in data visualization.***

Visual association refers to the way the viewer perceives and remembers visual elements that exist in a visualization and how they are connected or related to each other. Through the concept of visual association, we try to make a visualization more memorable to enable viewers to keep as much information as possible in their minds through adding unique elements or elements that could be remembered at-a-glance such as pictograms, human recognizable objects, and visual annotations.

While visual association focuses on visual elements, semantic association focuses on raising the quality of description for the visualization which aids in better understanding for the visualization content or visual elements. Semantic association is about adding textual descriptions such as titles and label providing more details to make the visualization understandable and help in recalling the message.

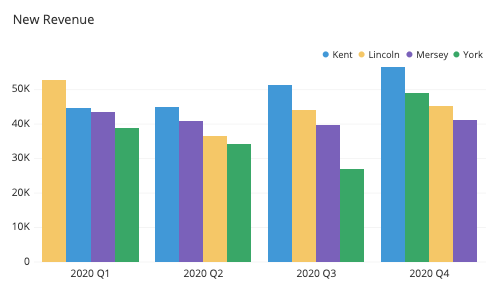
In data visualization, both visual and semantic associations are needed for more efficient and memorable visualizations. Integrating attractive visual elements alongside concise descriptions enhances understanding and visualization quality, ensuring that viewers recall the entire message.

# **Techniques of Data Visualization**

## ***What are the marks and channels, and number of attributes encoded in the following two plots.***

A) B)

A graph of a graph showing the number of sales

Description automatically generated with medium confidence

Marks:

Marks:

* Line
* Area

Channels:

* Position (Both horizantal and vertical)
* Color
* Size (length and area)

Number of attributes encoded: (3)

* Quarter (Date)
* Revenue
* Region
* Points

Channels:

* Position (Both horizantal and vertical)
* Color
* Size (Area)

Number of attributes encoded: (4)

* Market share
* Sales
* Region
* Sales growth

## ***Discuss the importance of using the following interaction techniques in data visualization: selection, change over time, navigation, filtering, brush and zoom, brush and link, and aggregation.***

Interaction techniques are techniques that can be used to enhance the representation of data allowing viewers to explore the story and interact with the data by themselves. Each interactive technique allows for different ways to explore and manipulate data.

***Selection:***

Selection technique enables viewer to highlight specific datapoint/s to explore relevant information to it by directing the whole focus to that point only through highlighting it or kind of turning any non-relevant information for it off. Also, in section we might show more information related to the selected datapoint. This is important because sometimes we want to focus on some points rather than others or to view a specific point without other points in crowded charts. For example, selecting one city out of many to explore its happiness level.

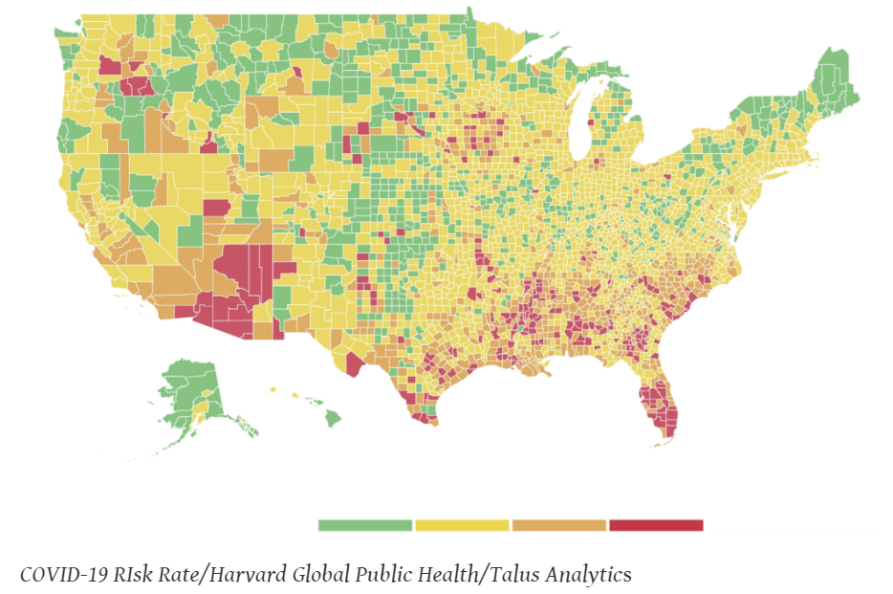
***Change overtime:***

A screenshot of a computer

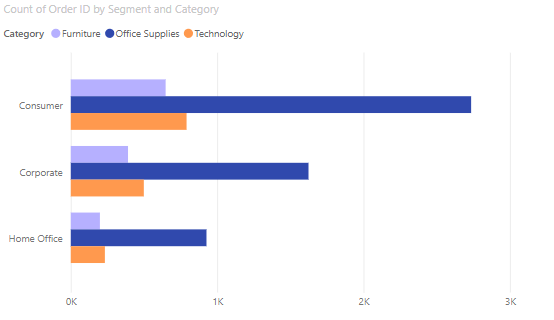
Description automatically generatedA number with colorful circles

Description automatically generated with medium confidenceThis technique animates the change that happens to the data over time, which shows how the data changes and how often. It’s important as it allows for monitoring the change that happens, allowing to explore the times when trends or changes happens. For example, change overtime might be used to show how the life expectancy changed over the years for each country.

***Navigation:***

Navigation is a very important interactive technique that allows viewers to explore data through zooming, scrolling, and moving around the visualization as they want. This technique is widely used in geospatial data or maps as viewers might need to move and zoom. Such a technique is very important as it provide viewers with details they want, allow them to focus on specific regions and explore them better. For example, applying navigation technique on the following map allows viewers to zoom on specific regions or states to explore them more or get more details.

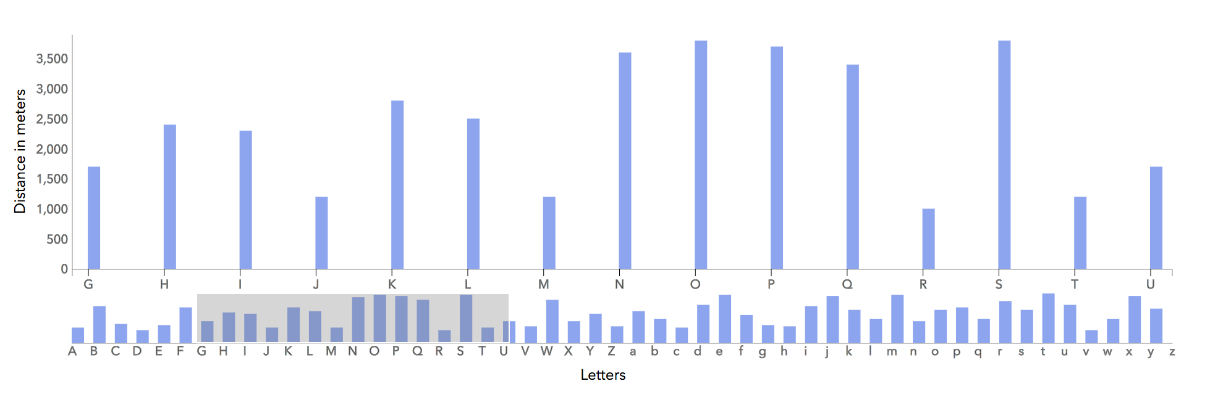
***Filtering:***

Filtering technique enables viewers to filter the data or show only the data points that meets a defined criteria by the viewer, it excludes all other points that doesn’t meet the criteria, this technique dynamically adjusts the displayed visualization according to the defined filter. Such a technique is extremely important in focusing the research on a specific group or segment, exploring different segments by using different filters, and uncovering hidden correlations. For example, filtering the following bar plot to only include home offices.

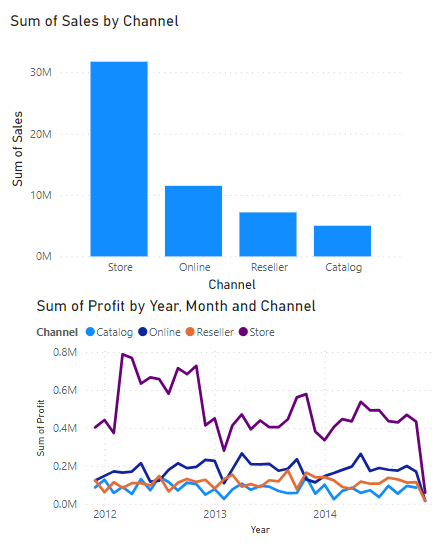
A graph with a bar and a line

Description automatically generated with medium confidence

***Brush and zoom:***

Brush and zoom technique allow for focused and specific views for regions in a visualization, by enabling the viewer to select (brush) a specific area, it creates a visualization for that selected subset of data showing more details. This technique is important especially when dealing with large and complex datasets where the overall visualization might not represent trends in an obvious way or capture fine details. For example, in the figure below, when brushing the letters from G to U, only their bar plots were displayed which provided better view to see and compare the differences between them.

***Brush and link:***

When having multiple visualizations linked together, brush and link technique is the perfect tool, this technique allows for dynamic change in multiple visualizations according to a selection or highlighting (brushing) on one of them, in other words, an action on one visualization affects other visualizations, it aids when we want to explore correlations and dependencies across multiple visualizations, and it’s widely used in dashboards to observe corresponding changes to an action. For example, figure A shows 2 charts related to sales data, the bar plot shows the sum of sales for each channel, while the line plot shows the sum of profit for each channel across different years. Figure B shows how this technique works when selecting only the “online” channel for example.A screenshot of a graph

Description automatically generated

B.

A.

***Aggregation:***

Aggregation technique allows for dynamic summarization for the data to see the bigger picture with the ability to dig deeper and explore the details. Aggregation is important as sometimes viewers may need to view the overall distribution and then the details, it also gives the viewer the ability to control the details wanted to be shown in the graph. For example, a decomposition tree might be used to show how sales are distributed over products, which can be summed up to the overall sales or detailed more to include the sales distribution over the sub products.

A diagram of a product company

Description automatically generated

## ***Mention 4 interactive visualization techniques used in*** [***Google Maps***](https://www.google.com/maps/@31.9628747,35.9032147,12.1z?entry=ttu)***.***

***1- Navigation:*** This technique allows for zooming, street viewing, rotating, and more, helps in finding locations, understanding the map, or exploring a specific region.

***2- Filtering:*** Allows for more than one view for the map by applying filters, which aids in using the map and viewing specific regions according to specific filters like the traffic.

***3- Selection:*** Enables users to select specific locations to get more information about them, such as reviews, directions, and more.

***4- Query or search******:*** This technique is beneficial when the user wants to search for a specific place or location that they don’t now it’s place on the map, it also shows choices if the user isn’t sure about the place name or location.

# **Static and Interactive Visualizations (project)**

## ***Provide a detailed description of each of the datasets you selected, and the attributes of interest, and the attributes you would like to visualize (data abstraction).***

Liver Cirrhosis stage classification dataset, the dataset was sourced from a study by Mayo Clinic Study on primary biliary cirrhosis (PBC) of the liver carried out from 1974 to 1984, in addition, synthetic data was used to increase samples, it data contains 25000 observations and 19 features.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute name | Data type | Description | Effect |
| N\_days | discrete | Number of days between registration and the earlier of death, transplantation, or study analysis time in 1986 |  |
| Status | Categorical (3 values) | Patient’s status, C (censored), CL (censored due to liver transplantation), or D (death) |  |
| Drug | Categorical (2 values) | type of drug D-penicillamine or placebo  NOTE: A placebo is an inactive substance that has no therapeutic effect used to examine if patients experience real changes in their condition simply because they believe they are receiving treatment. |  |
| Age | Discrete | Age in days, transformed to years |  |
| Sex | Categorical (2 values) | Female or Male |  |
| Ascites | Categorical (binary values) | Accumulation of fluid in the peritoneal cavity | Presence of Ascites may indicate liver damage or cirrhosis. |
| Hepatomegaly | Categorical (binary values) | Enlarged liver | Presence of Hepatomegaly may indicate liver damage or cirrhosis. |
| Spiders | Categorical (binary values) | Spider-like blood vessels visible under the skin | Presence of Spiders may indicate liver cirrhosis. |
| Edema | Categorical (3 values) | Swelling due to fluid accumulation, N (no Edema and no diuretic therapy for edema), S (Edema resolved by diuretics), or Y (Edema despite diuretic therapy) | Presence of Edema may indicate liver damage or cirrhosis. |
| Bilirubin | Continuous | Serum bilirubin in [mg/dl]  Normal range: [0.1-1.2] | Higher levels indicate more severe liver dysfunction. |
| Cholesterol | Continuous | Serum cholesterol in [mg/dl]  Normal range<200 | Can be an indicator of liver function |
| Albumin | Continuous | Albumin in [gm/dl]  Normal range: [3.5-5] | Lower levels may indicate poor liver function as liver cirrhosis stops or lessens its production |
| Copper | Continuous | Urine copper in [ug/day]  Normal range: [70-140] | High levels may indicate liver disease. |
| Alk\_Phos | Continuous | Alkaline phosphatase in [U/liter]  Normal range: [45-150] | High levels may indicate liver disease. |
| SGOT | Continuous | SGOT in [U/ml], also known as AST.  Normal range: [10-40] | High levels may indicate liver disease, it also indicates high alcohol consumption. |
| Triglycerides | Continuous | Triglycerides in [mg/dl], indicates fats level.  Normal range<150 | High levels may cause liver cirrhosis or indicate the existence of liver cirrhosis. |
| Platelets | Continuous | platelets per cubic [ml/1000], an indictor for the clotting factors rate.  Normal range: [150-450] | Lower levels indicate liver disfunction as cirrhosis impairs the production of platelets. |
| Prothrombin | Continuous | Prothrombin time in seconds [s], time to produce clotting factors.  Normal range: [11,13.5] seconds | Longer time indicates a reduced ability to form clotting factors, which is associated with liver disease. |
| Stage (LABEL) | Categorical (2 values) | Histologic stage of disease, early, intermediate, advanced. |  |

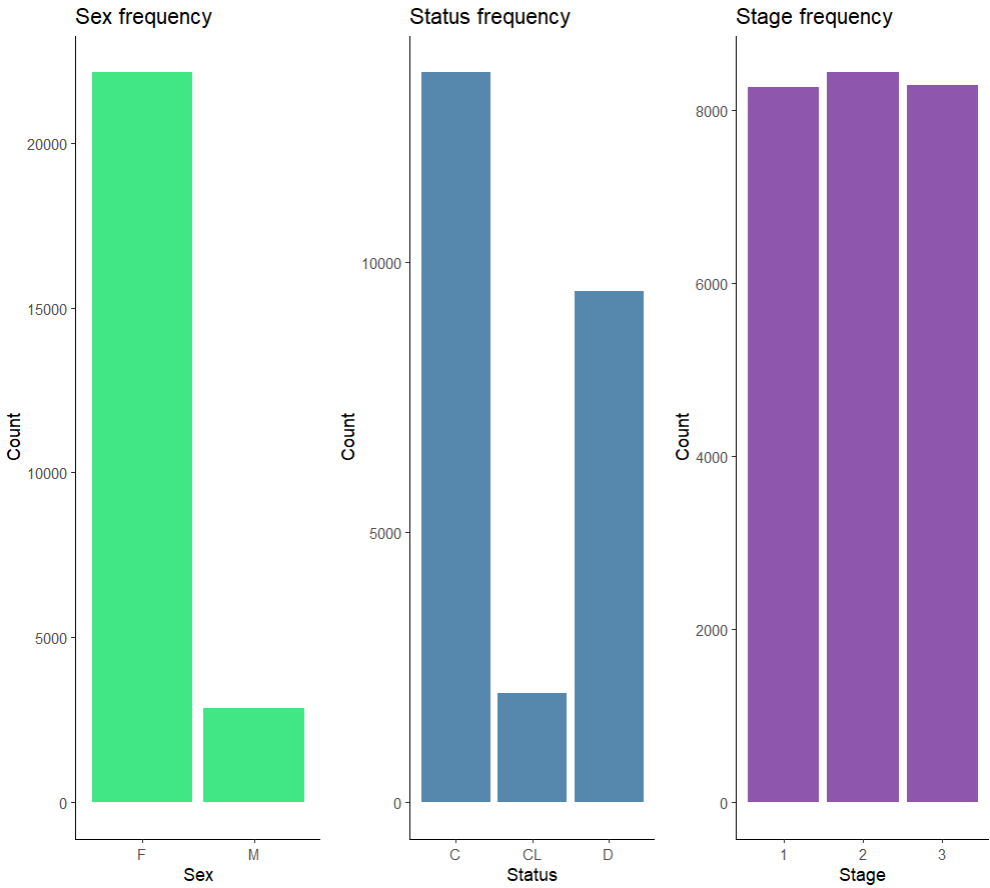
## ***Describe why you are looking at the data (task abstraction), and design rules for each plot you generated in R.***

## ***Describe the encoding methods for each plot (i.e., choosing the plot based on data and tasks, and marks and channels etc.).***

## ***Assess the methods (e.g., proper design and encoding, Tufte’s rules, Schneiderman Mantra, etc.) used to build your visualizations and how they helped (static and interactive).***

## ***Analyze and assess the insights and findings for each plot you generated in R.***

***Static visualizations:***

1- Sex, Status, and Stage bar plots:

***Task:***

Analyze and discover data.

***Num of variables encoded:***

1 variable for each plot

***Marks:***

Line and area for each plot.

***Channels:***

Position (Both horizantal and vertical),

Size (length and area) for each plot

***Tufte’s rules applied:***

Graphical integrity is ensured through the axes labels, titles, avoiding over or under stating in the lie factor, no existence for unjustified 3d plots, the data ink ration is maximized while the non-data ink is erased, no presence for chart junk.

***Insights:***

1st plot: The data is imbalanced when it comes to sex, so we can’t rely on insights related to it.

2nd plot: Most of the patient’s status are C (censored), then D (dead), and the least one is CL (censored due to liver transplantation) which means liver transplantation is the least status that exist among patients as it’s hard.

3rd plot: Showing that the data is balanced in the target, which is the stage, so this is great for extracting features from plot relying on the stage, it also shows that no stage dominates all cases.

A graph of different colored bars

Description automatically generated with medium confidence2- Drug, with stage and status:

***Task:***

Analyze and discover data.

***Num of variables encoded:***

1 variable in the 1st plot, 2 in the 2nd

and 3rd plots

***Marks:***

Line and area for each plot.

***Channels:***

Position (Both horizantal and vertical),

size (length and area) for each plot,

color for the 2nd and 3rd plot.

***Tufte’s rules applied:***

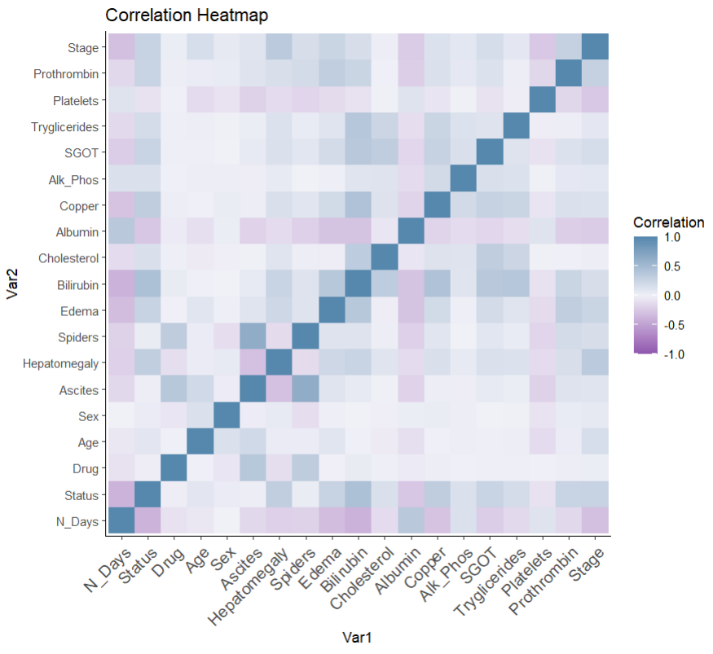
Graphical integrity is ensured through the axes labels, titles, avoiding over or under stating in the lie factor, no existence for unjustified 3d plots, the data ink ration is maximized while the non-data ink is erased, no presence for chart junk, and the scale for both axes is great.

***Insights:***

1st plot: Shows the frequency for the 2 drugs used, it’s important to note that Placebo isn’t a real drug it’s just used to test the phycological effect of drugs on the patient status and recovery.

2nd plot: To test if a drug is used more than the other on different status, but it looks like the status doesn’t affect the choice of the drug.

3rd plot: Just as the 2nd plot but to test it on the stage, but it also seems like there is no relationship between the stage and the choice of the drug.

3- Correlation matrix heatmap:

***Task:*** Analyze and discover data.

***Num of variables encoded:***

19.

***Marks:***

Area

***Channels:***

Position (Both horizantal and vertical),

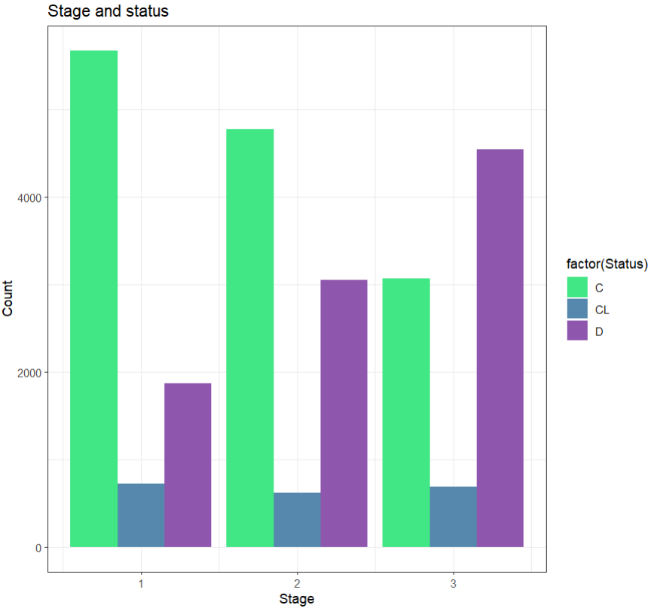
size (length and area), color

***Tufte’s rules applied:***

Graphical integrity is ensured through the axes labels, titles, and legend, no existence for unjustified 3d plots, the data ink ration is maximized while the non-data ink is erased, and no presence for chart junk.

***Insights:***

The correlation or relationship between each variable and all other variables appears in the plot, where the color indicates the strength or weakness of the relationship and if it’s positive or negative relationship. For example, the Bilirubin appears to be correlated with Edema, and so on.

4- Stage and status:

***Task:*** Analyze and discover data.

***Num of variables encoded:*** 2

***Marks:***

Line and area

***Channels:***

Position (Both horizantal and vertical),

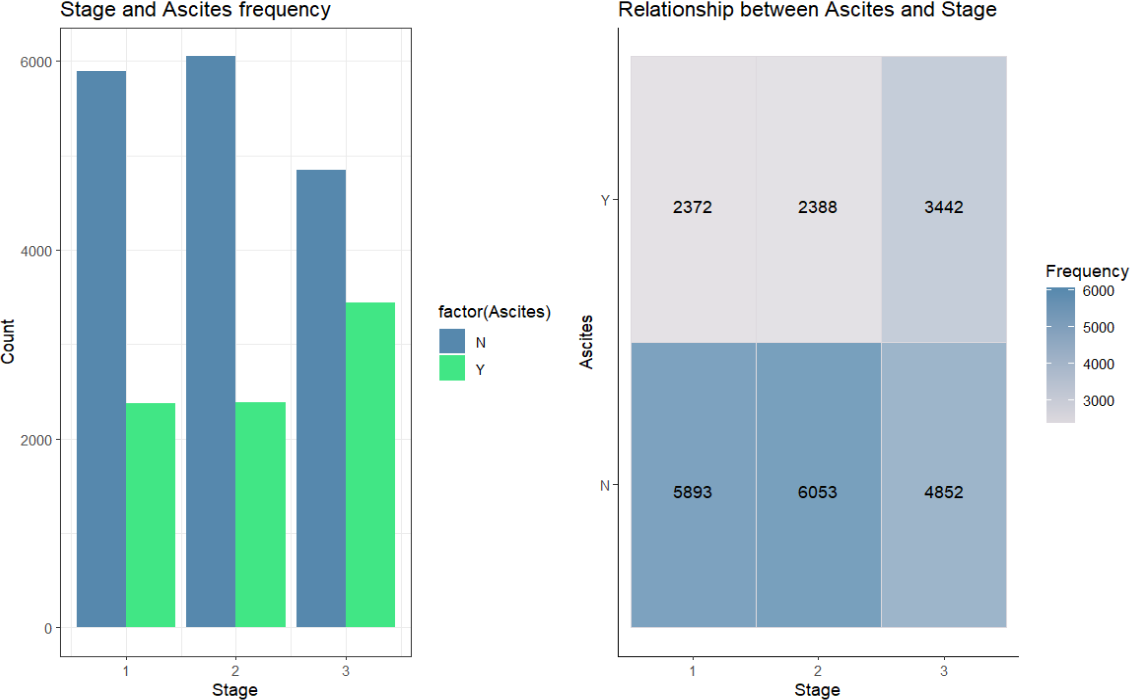
size (length and area), color

***Tufte’s rules applied:***

Graphical integrity is ensured through the axes labels, titles, and legend, no existence for unjustified 3d plots, the data ink ration is maximized while the non-data ink is erased, and no presence for chart junk.

***Insights:***

Through the plot, we can conclude that most death cases are patients with stage 3 of liver cirrhosis, which is normal as it is the most advanced level of the disease. Also, it appears that the liver transplantation has no relation with the stage as it has the almost the same frequency among all stages.

5- Stage and Ascites:

***Task:*** Analyze and discover data.

***Num of variables encoded:*** 2

***Marks:*** Line and area for the 1st plot, area for the 2nd plot

***Channels:***

Position (Both horizantal and vertical) for both plots,

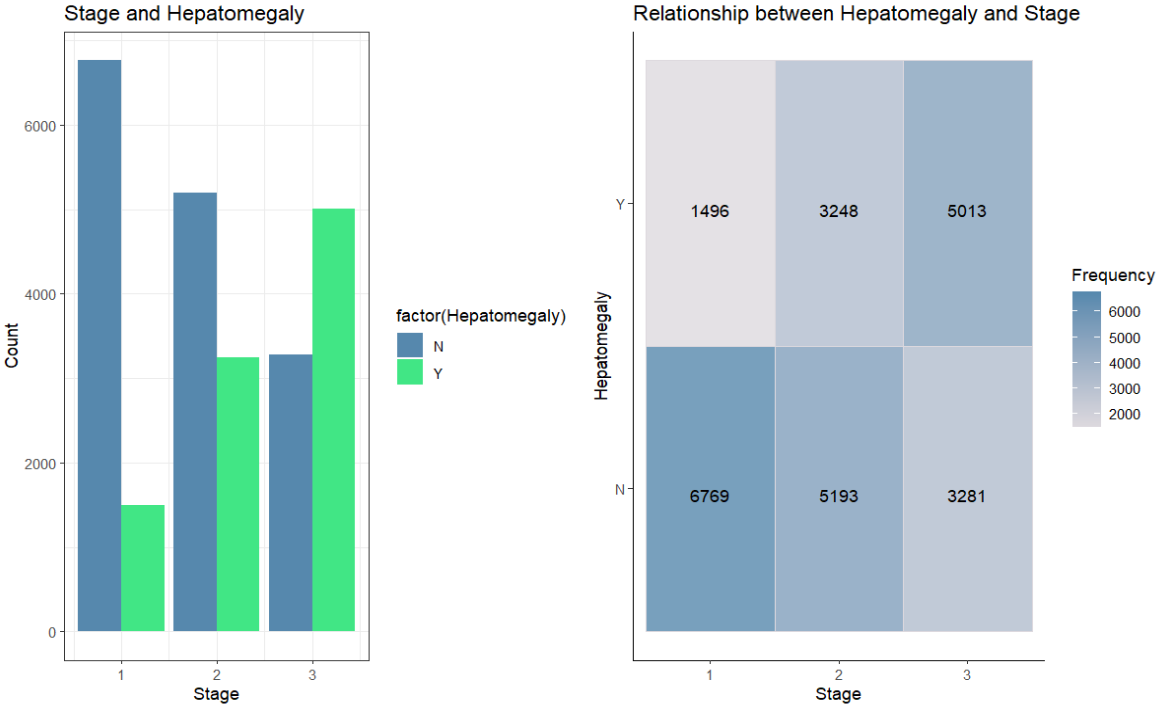
size (length and area) for the 1st plot, color for both plots

***Tufte’s rules applied:***

Graphical integrity is ensured through the axes labels, titles, and legend, no existence for unjustified 3d plots, no occlusion, the data ink ration is maximized while the non-data ink is erased, and no presence for chart junk.

***Insights:***

From both plots we can see that Ascites presence in not strong over all stages, but it has the strongest presence in patients with stage 3 of liver cirrhosis with 3442 values.

6- Stage and Hepatomegaly:

***Task:*** Analyze and discover data.

***Num of variables encoded:*** 2

***Marks:*** Line and area for the 1st plot, area for the 2nd plot

***Channels:***

Position (Both horizantal and vertical) for both plots,

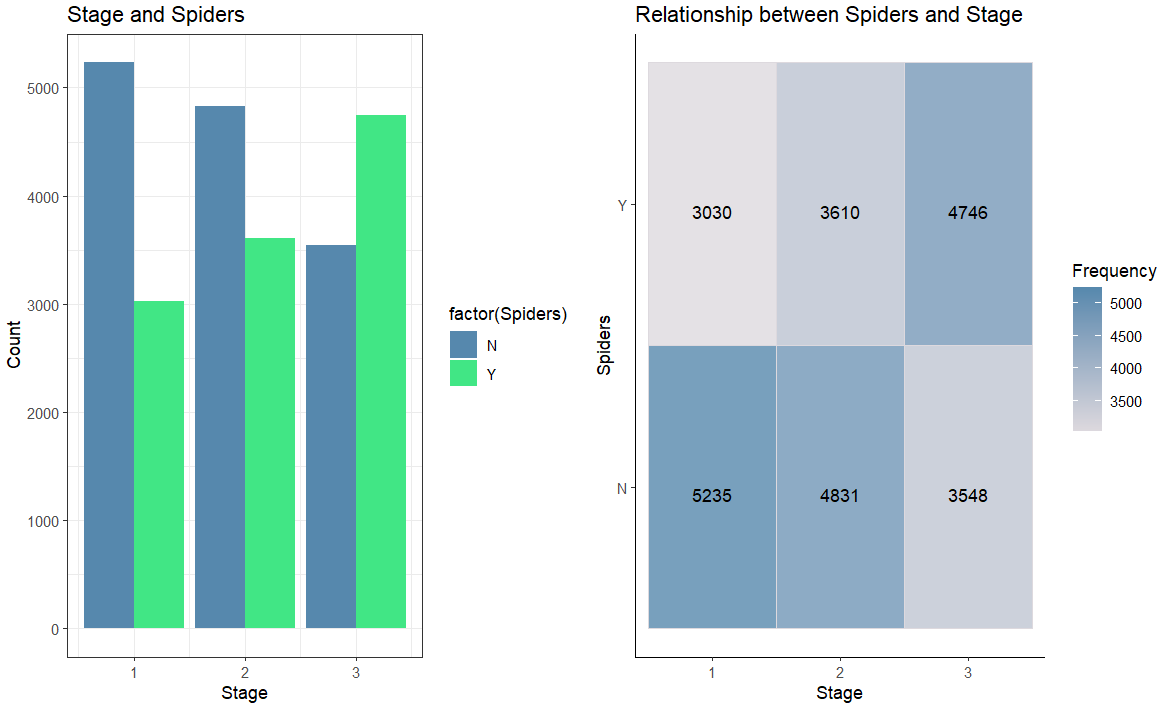
size (length and area) for the 1st plot, color for both plots

***Tufte’s rules applied:***

Graphical integrity is ensured through the axes labels, titles, and legend, no existence for unjustified 3d plots, no occlusion, the data ink ration is maximized while the non-data ink is erased, and no presence for chart junk.

***Insights:***

From both plots we can Conclude that the presence of Hepatomegaly is highly correlated with the stage, which means that most cases where Hepatomegaly is present are classified as stage 3 (5013), while the cases with no presence of Hepatomegaly most of the times are cases with stage 1 of liver cirrhosis (6769).

7- Stage and Spiders

***Task:***

Analyze and discover data.

***Num of variables encoded:*** 2

***Marks:***

Line and area for the 1st plot, area for the 2nd plot

***Channels:***

Position (Both horizantal and vertical) for both plots,

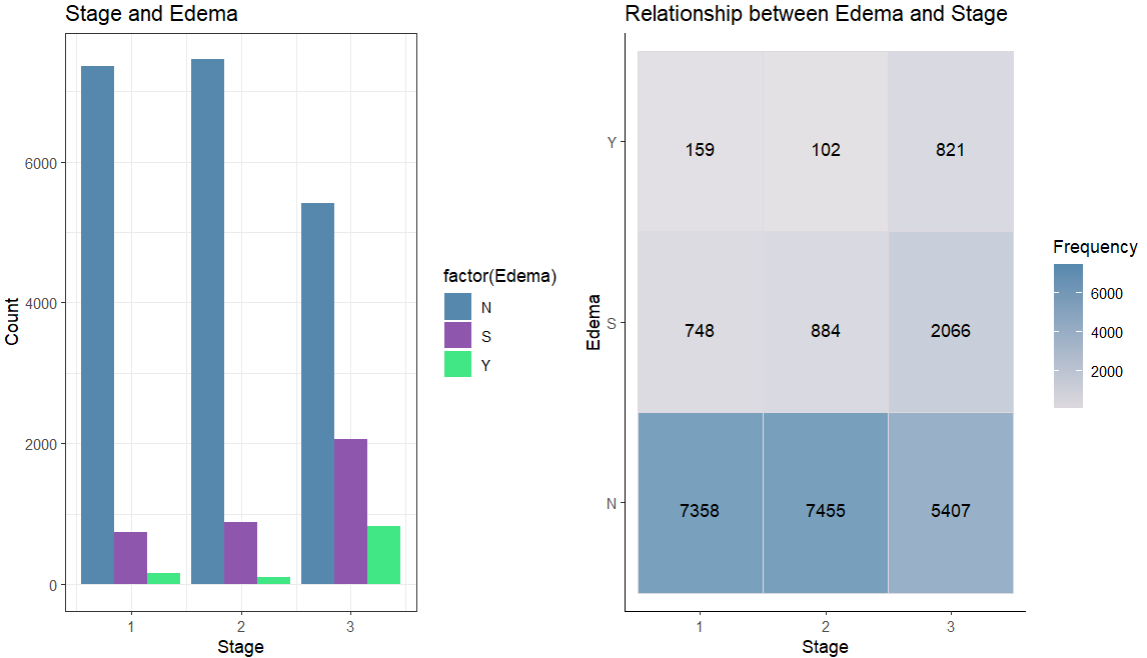
size (length and area) for the 1st plot, color for both plots

***Tufte’s rules applied:***

Graphical integrity is ensured through the axes labels, titles, and legend, no existence for unjustified 3d plots, no occlusion, the data ink ration is maximized while the non-data ink is erased, and no presence for chart junk.

***Insights:***

From the plots we can notice that the presence of spiders can be an indicator for liver cirrhosis especially for stage 3, where 4746 cases with stage 3 have spiders, while most stage 1 cases have no presence of Spiders.

7- Stage and Edema:

***Task:***

Analyze and discover data.

***Num of variables encoded:*** 2

***Marks:***

Line and area for the 1st plot, area for the 2nd plot

***Channels:***

Position (Both horizantal and vertical) for both plots,

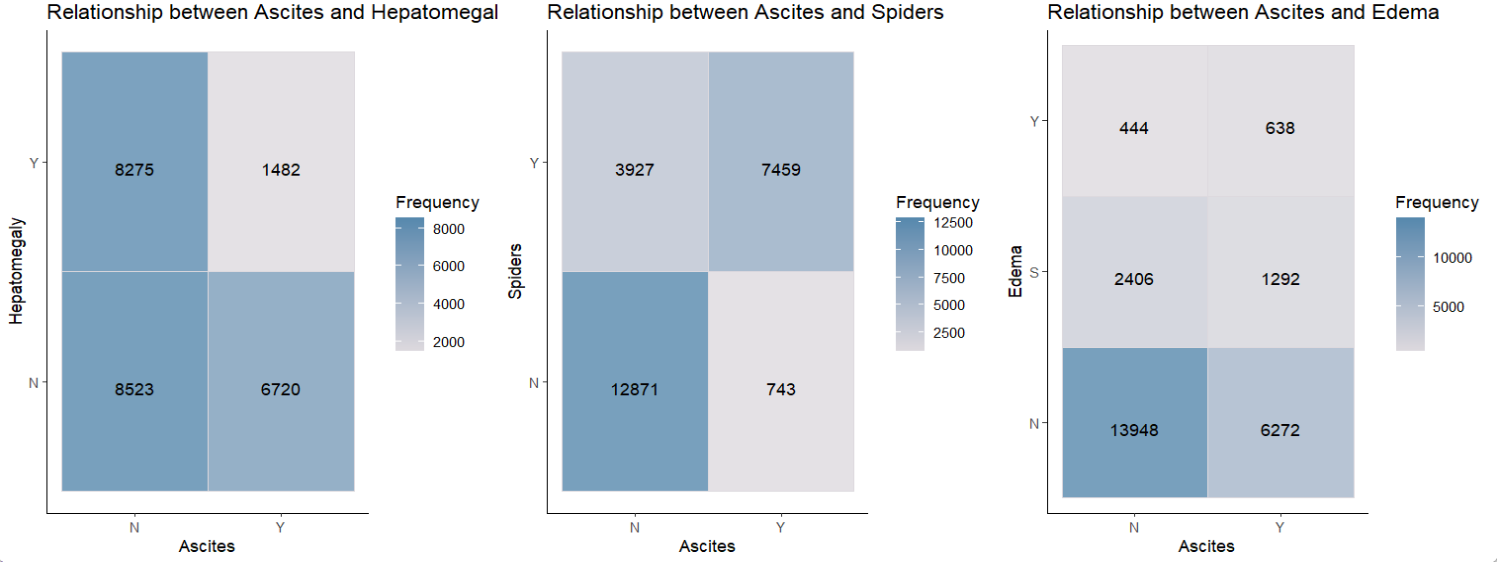
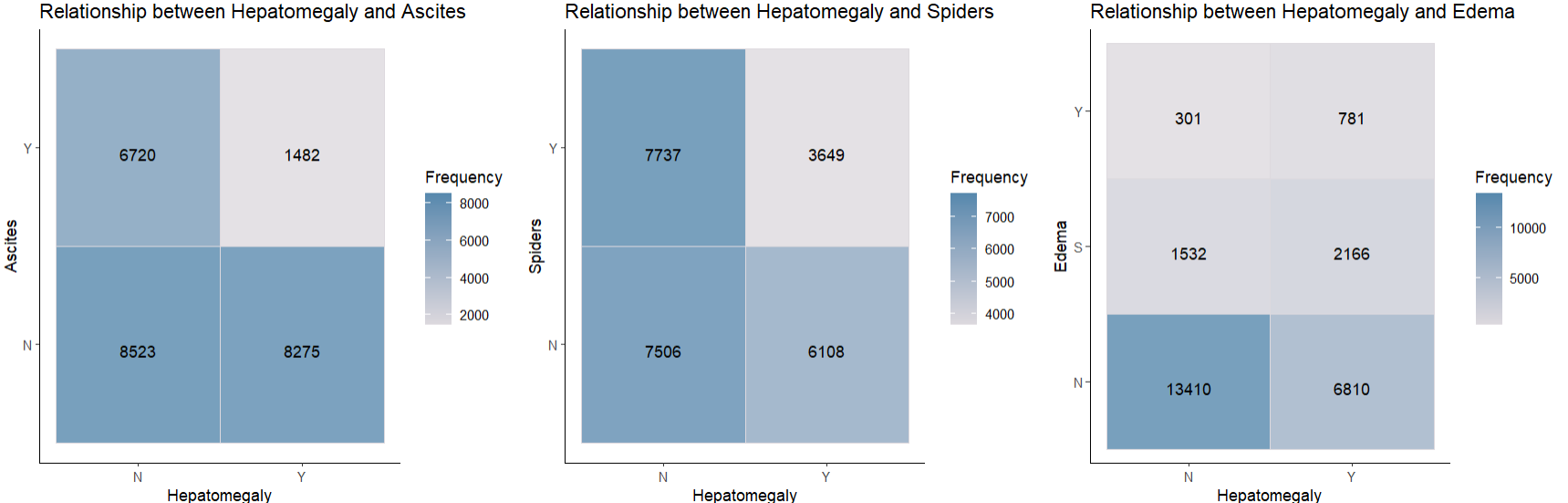
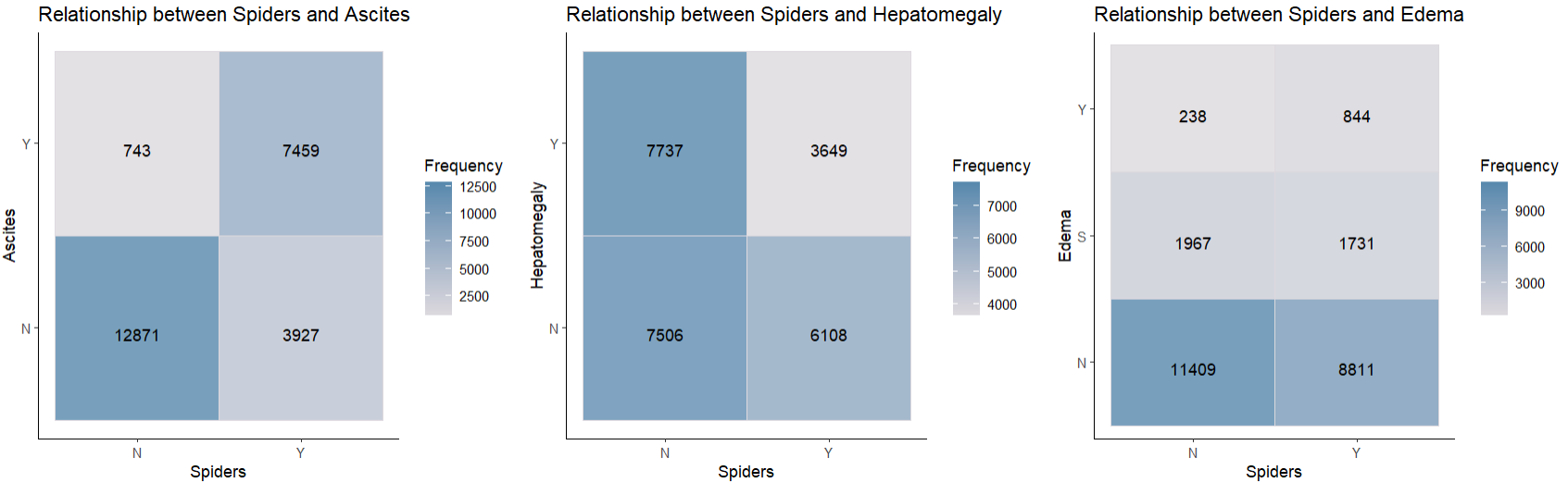
size (length and area) for the 1st plot, color for both plots

***Tufte’s rules applied:***

Graphical integrity is ensured through the axes labels, titles, and legend, no existence for unjustified 3d plots, no occlusion, the data ink ration is maximized while the non-data ink is erased, and no presence for chart junk.

***Insights:***

It’s noticeable from the plots that in most cases there is no presence for Edema or that it has been resolved by using diuretics, it’s also clear that the presence of Edema or resolved Edema has the highest frequency in Stage 3 cases, which indicates a relationship between the 2 variables.

8- Categorical variables relationship through crosstabs:

***Task:*** Analyze and discover data.

***Num of variables encoded:*** 2 for each plot

***Marks:*** Area

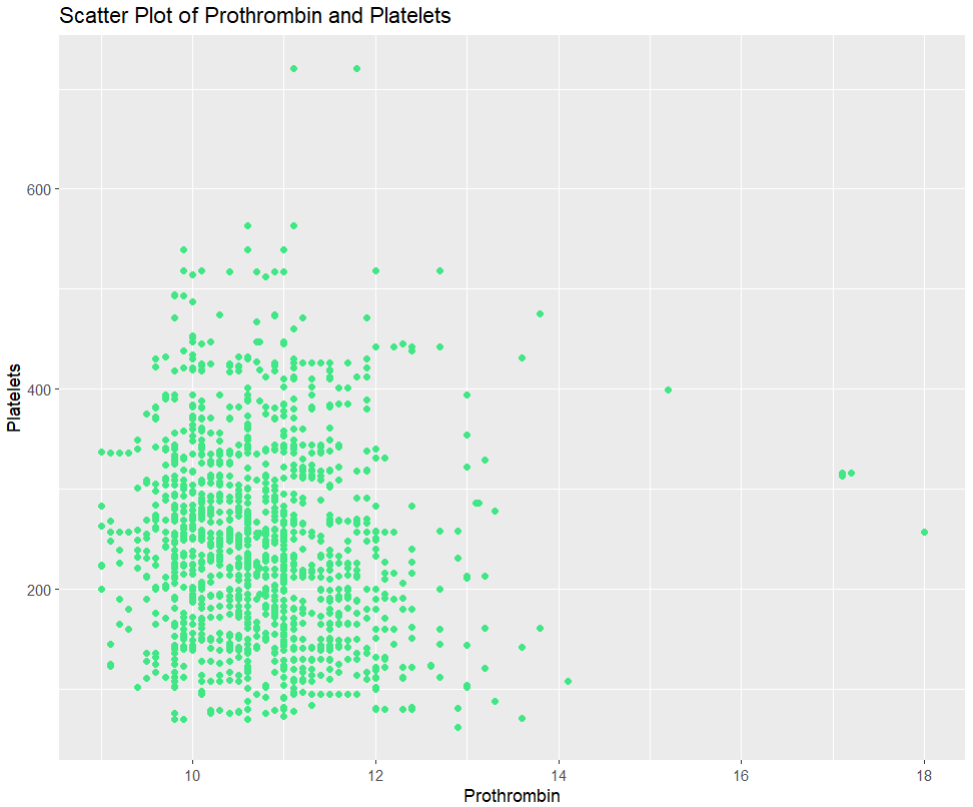
***Channels:*** Position (Both horizantal and vertical), color for both plots

***Tufte’s rules applied:***

Graphical integrity is ensured through the axes labels, titles, and legend, no existence for unjustified 3d plots, no occlusion, the data ink ration is maximized while the non-data ink is erased, and no presence for chart junk.

***Insights:***

The crosstab plots above show the relationship between each categorical variable from (Ascites, Hepatomegaly, Spiders, and Edema) and the other 3 variables. It shows the frequency between each combination of values, which can lead us to concluding the type of relationship between each 2 variables, for example, from the plot we can wee that the presence of Ascites is highly correlated with the presence of Spiders.

9- Prothrombin and Platelets:

***Task:*** Analyze and discover data.

***Num of variables encoded:*** 2

***Marks:*** Points

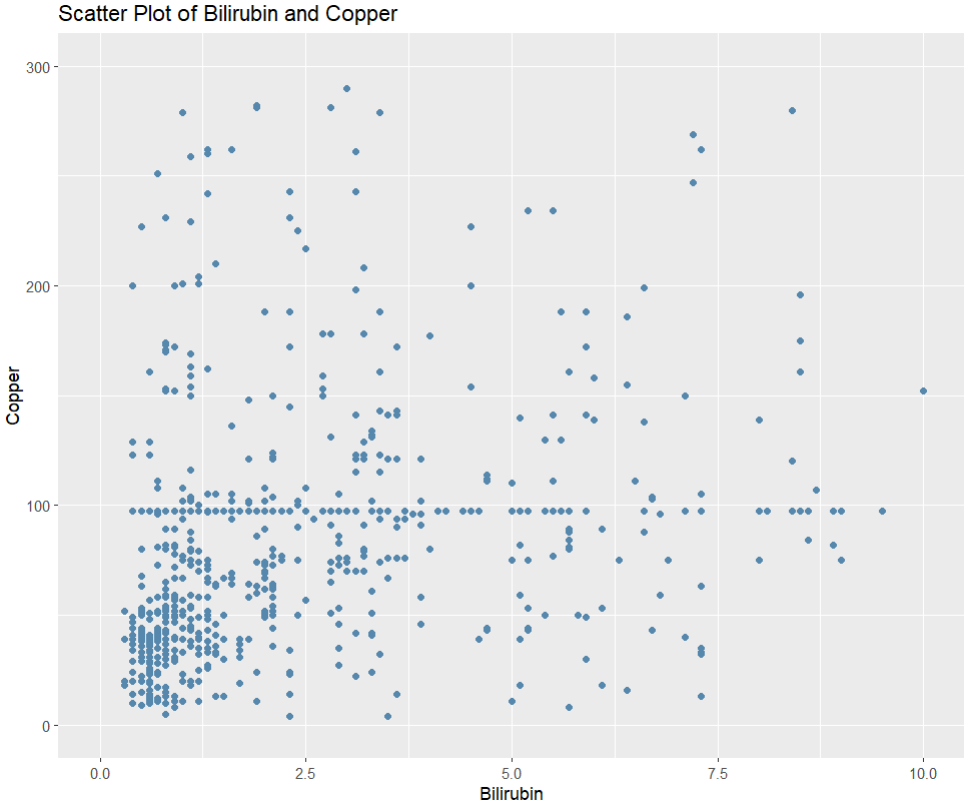
***Channels:*** Position (Both horizantal and vertical)

***Tufte’s rules applied:***

Graphical integrity is ensured through the axes labels, titles, and legend, no existence for unjustified 3d plots, no occlusion, the data ink ration is maximized while the non-data ink is erased, and no presence for chart junk.

***Insights:***

The aim of this plot is to discover weather longer time of Prothrombin affects the number of Platelets through the distribution of data points, it seems like there is no strong relationship between them, as the change in one of them doesn’t reflect on the other one.

******10- Bilirubin and Cooper:

***Task:*** Analyze and discover data.

***Num of variables encoded:*** 2

***Marks:*** Points

***Channels:*** Position (Both horizantal and vertical)

***Tufte’s rules applied:***

Graphical integrity is ensured through the axes labels, titles, and legend, no existence for unjustified 3d plots, no occlusion, the data ink ration is maximized while the non-data ink is erased, and no presence for chart junk.

***Insights:***

This plot shows the distribution of data on both Cooper and Bilirubin, from the plot we can see that most Cooper values are below 100, while most Bilirubin value are below 2, it also seems like sometimes the high values of one variable result in high values for the other variable but not always.

***Interactive visualizations:***

1- Age distribution across stages:

***Task:*** Analyze and discover data.

***Num of variables encoded:*** 2

***Marks:*** line and area

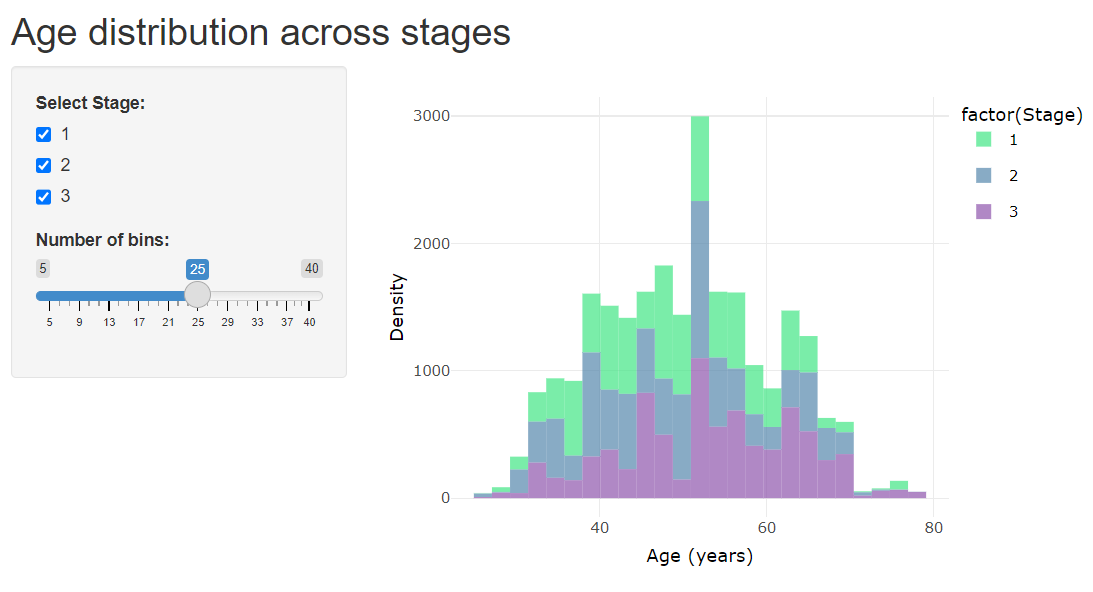
***Channels:*** Position (Both horizantal and vertical), size (length and area), color

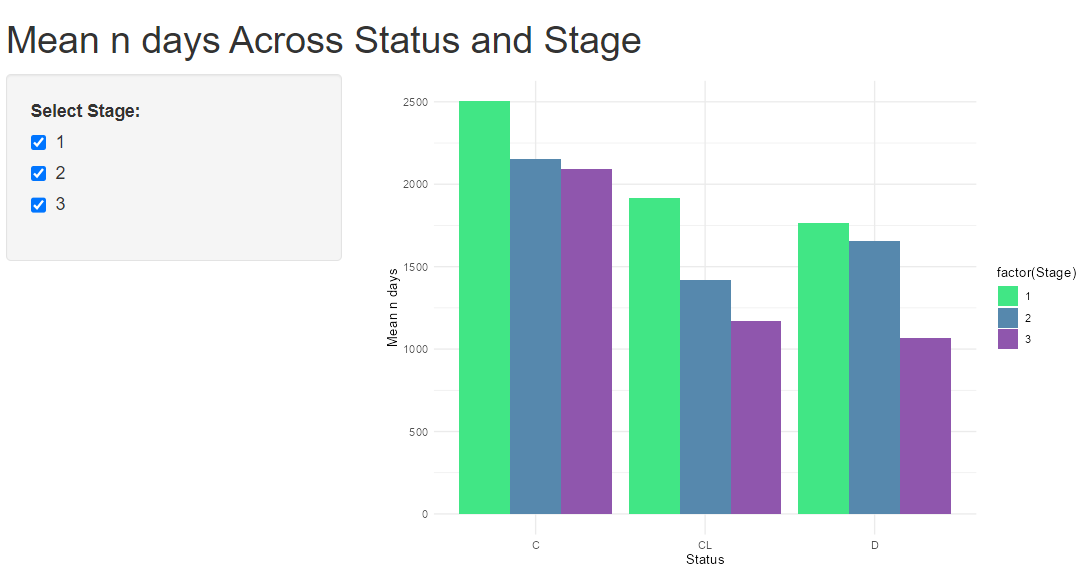
***Schneiderman Mantra rules applied:***

The plot provides summary and high-level view for the age distribution, it also enables users to discover and explore data through zooming and filtering, and the details are on demand, the viewer can select what information they want to review and see instead of overwhelming them with unnecessary details.

***Insights:***

It’s an interactive plot that shows the age distribution among stages, where the user can slide to choose the number of bins ranged from 5-40 bins, and the checkbox can be used for filtering or showing specific stage distribution instead of all stages, in addition, legend filtering is also applicable. From the plot we can see that most patients are aged from 40-60, we can also conclude that stage 1 dominates across all ages.



2- Mean number of days across Status and Stage:

***Task:*** Analyze and discover data.

***Num of variables encoded:*** 3

***Marks:*** line and area

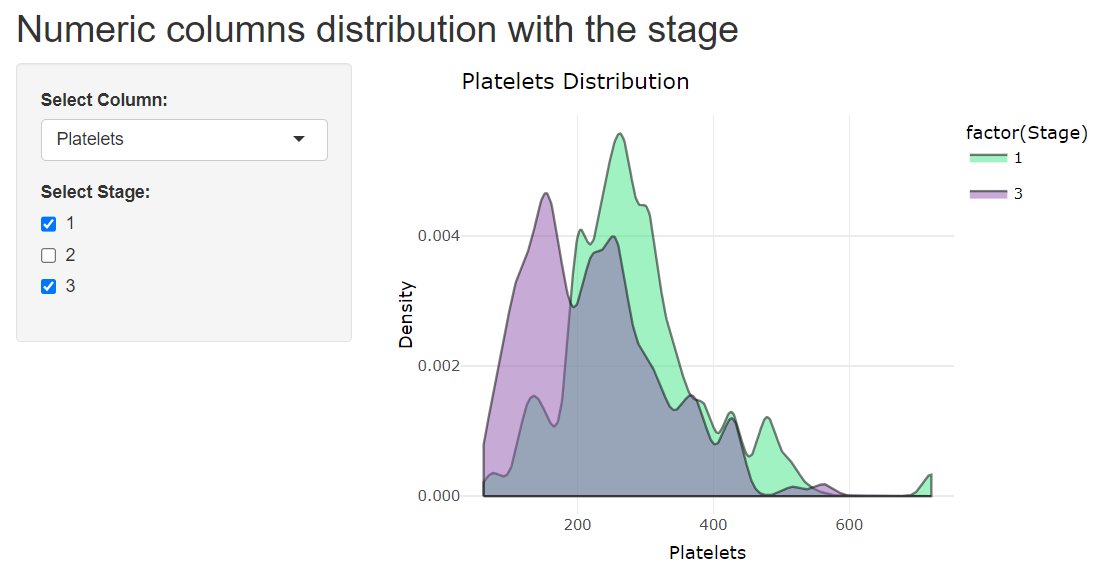
***Channels:*** Position (Both horizantal and vertical), size (length and area), color

***Schneiderman Mantra rules applied:***

The plot provides summary and high-level view for the mean number of days for each status, it also enables users to discover and explore data through zooming and filtering on the stage, and the details are on demand, the viewer can select what information they want to review and see instead of overwhelming them with unnecessary details.

***Insights:***

It’s an interactive plot that shows the mean number of days among status, where the viewer can filter on the stage using the check box or legend filtering. From the plot we can conclude that the mean number of days for death people among all stages is the less that other status, through filtering the viewer can also find other insights related to their interest.

3- Numeric columns distribution with the stage:

***Task:*** Analyze and discover data.

***Num of variables encoded:*** 2 for each plot generated.

***Marks:*** Area

***Channels:*** Position (Both horizantal and vertical), size (length and area), color

***Schneiderman Mantra rules applied:***

The plot provides high-level view for the distribution of each numeric column out of the 9 columns, it also enables users to discover and explore data through zooming and filtering on the stage, and the details are on demand, the viewer can select what stage they want to review or if they want to compare all stages.

***Insights:***

As it’s an interactive plot for many attributes, each one has a different insight, the viewer can select the attribute they want, and then explore the distribution for it among all stages and compare them to get insights according to their attribute of interest. For example, in the figure above, the Platelets attribute was chosen, and only the distribution for it in stage 1 and 3 are selected, we can see that the overall distribution for it in stage 3 differs from stage 1 as stage 3 patient appear to have lower Platelets levels.

A screenshot of a graph

Description automatically generated4- Numeric columns distribution with the normal ranges:

***Task:*** Analyze and discover data.

***Num of variables encoded:*** 1 for each plot generated.

***Marks:*** line and area

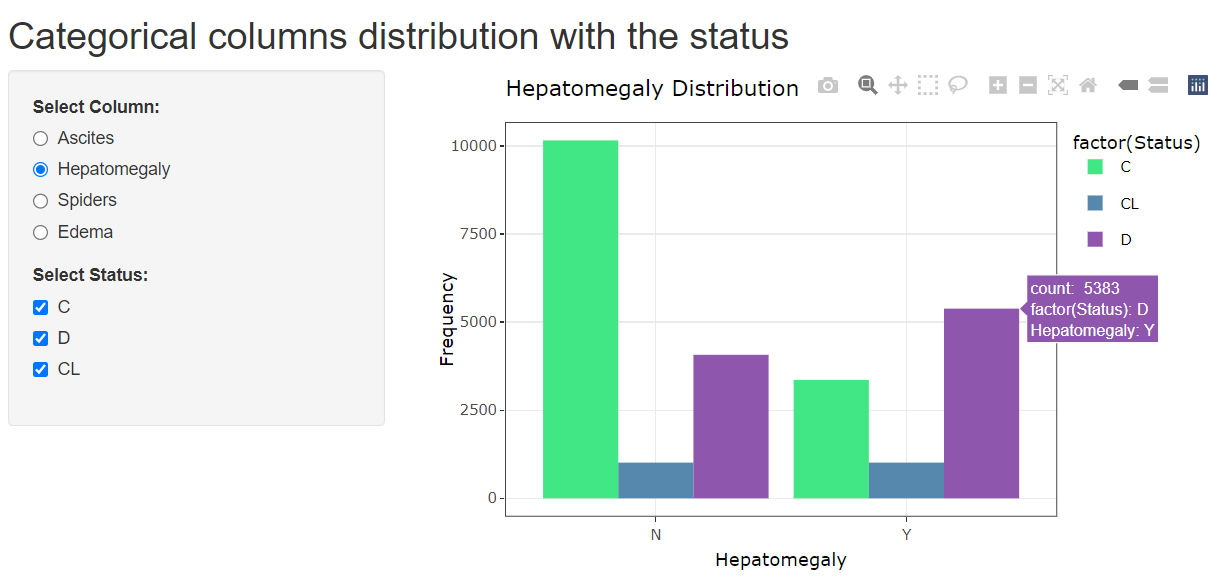
***Channels:*** Position (Both horizantal and vertical), size (length and area), color

***Schneiderman Mantra rules applied:***

The plot provides high-level view for the distribution of each numeric column out of the 9 columns and their normal ranges, and the details are on demand, the viewer can select what attribute they want to view from the dropdown list.

***Insights:***

The plot allows for discovering 9 attributes each one alone with its normal range, each plot provides insights about how liver cirrhosis affects these reading, for example, the plot above shows the cholesterol levels distribution and its normal range between the green line, we can conclude from the plot that most liver cirrhosis patients have higher cholesterol levels than the normal range.

5- Categorical columns distribution with the status:

***Task:*** Analyze and discover data.

***Num of variables encoded:*** 2 for each plot generated.

***Marks:*** line and area

***Channels:*** Position (Both horizantal and vertical), size (length and area), color

***Schneiderman Mantra rules applied:***

The plot provides a view for the distribution of each categorical column selected, zooming and filtering on the status are allowed using the checkbox or the legend. and the details are on demand, the viewer can select what attribute they want to view from the radio button and what status to avoid showing unwanted details., selection is also enabled, and it presents more details such as the count.

***Insights:***

The above interactive plot allows the viewer to get insights related to some categorical columns and the status, from each plot, we can discover and compare the relationship between them, and the count for each combination, the plot above shows how Hepatomegaly is distributed among different status, we can derive the insight that the presence of Hepatomegaly affects death status more than the absence of it. It also shows that most cases where the Hepatomegaly present are death cases.

## ***Critically evaluate how the project visualizations impact the organization and decision-making.***

Visualizing liver cirrhosis data have a significant impact on the organization and decision making positively, the created visualizations transformed complex and detailed data to actionable, accessible, and understandable insights that benefit health organizations, healthcare providers, healthcare researchers, and medicine developers. Though the visualizations, the viewer can easily understand the relationship between each of the attributes in the dataset and the Stage (label), for example, each numeric column distribution can be viewed with each Stage, and each categorical column relationship with the Stage is also available, they provide a high-quality look for the data that enables users to understand the data at a glimpse, quickly identify trends, patterns, and extract useful information and insights that can be hardly understood from the data in its normal form. The visualizations don’t only represent and show insights about the attributes relationship with the Stage but also for attributes between each other’s, such as the crosstab visualization between the categorical columns (Ascites, Hepatomegaly, Spiders, and Edema), and between the Status and the numeric columns like Platelets, these visualizations allow for uncovering hidden correlations that provide deeper understanding for the cases which benefit in choosing the correct treatment, or developing medicines to address liver cirrhosis. Organizations and stakeholders can rely on visualizations to make decisions as the visualizations convey insights that leads to better and data driven decisions, which results in better health outcomes.

# **Communication of Results and Findings**

## ***Prepare PowerPoint slides to demonstrate the use of data visualization and provide insights and knowledge. (slides and presentation)***

## ***Provide a presentation to communicate your findings and insights (slides and presentation)***

## ***Evaluate how good your project visualizations can tell the story and provide insights to the organization. (report)***

My project which is visualizing liver cirrhosis data excel in several aspects, as the healthcare data is a very detailed and complex data that sometimes contain hidden correlations and trends that can only be uncovered in visualizations, my project’s visualizations provided a simpler and clearer view for the data, which results in better representation for the data that could be easily understood and explored by stakeholders and healthcare professionals. Furthermore, the visualizations were created according to Tufte’s and Schneiderman Mantra’svisualization rules which ensured the graphical integrity and improved the visualizations allowing for even better representation and easier way to convey the message to the viewer. In addition, the interactive visualizations present large amount of information in an organized manner and also enables viewers to view the data according to the size of details they want, it allows stakeholder to customize their data exploration to extract specific insights they require and need, this enhances the utility of the generated insights and visualizations. Lastly, the visualizations of my project have covered all attributes in the dataset meaning that all information in the dataset has been presented and utilized in order to create visualizations that empower stakeholders to make data driven decisions. To conclude, I can describe my project as the best narrator to tell the story of liver cirrhosis, as it connects all attributes, provides comprehensive view for the data, follows visualization rules, and allows interactivity, putting all of this together yields a project that assists and empowers stakeholders in making informed decisions, providing better health services to patients, and encouraging healthcare innovation.

# **References**

*Gallery of data visualization* (no date) *Gallery of Data Visualization - The Lie Factor*. Available at: https://www.datavis.ca/gallery/lie-factor.php (Accessed: 16 May 2024).

Raj, A. (2022) *Chart junk: How to avoid it?*, *Code Conquest*. Available at: https://www.codeconquest.com/blog/chart-junk-how-to-avoid-it/ (Accessed: 16 May 2024).

Blaxell, R. (no date) *The manifesto of the data-ink ratio*, *Journal*. Available at: https://blog.liquidinteractive.com.au/the-manifesto-of-the-data-ink-ratio#:~:text=The%20data%2Dink%20ratio%20is,be%20deleted%20everywhere%20where%20possible. (Accessed: 16 May 2024).