Postgraduate Coursework Template

What are the research questions that your data visualization will help you to answer?

Research Questions

Health care costs in the United States are some of the most expensive among developed countries. Many hospitals and insurance providers negotiate prices independently, resulting in large variations in charges for the same procedure at different healthcare centers within the same city. I would like to explore these charges to see if medical charges are truly random, or if there exist any patterns or geographic trends.

- Do charges for the same procedure vary consistently across regions (cities, states, geographic regions) or are there clusters of higher/lower variance in pricing?

Data

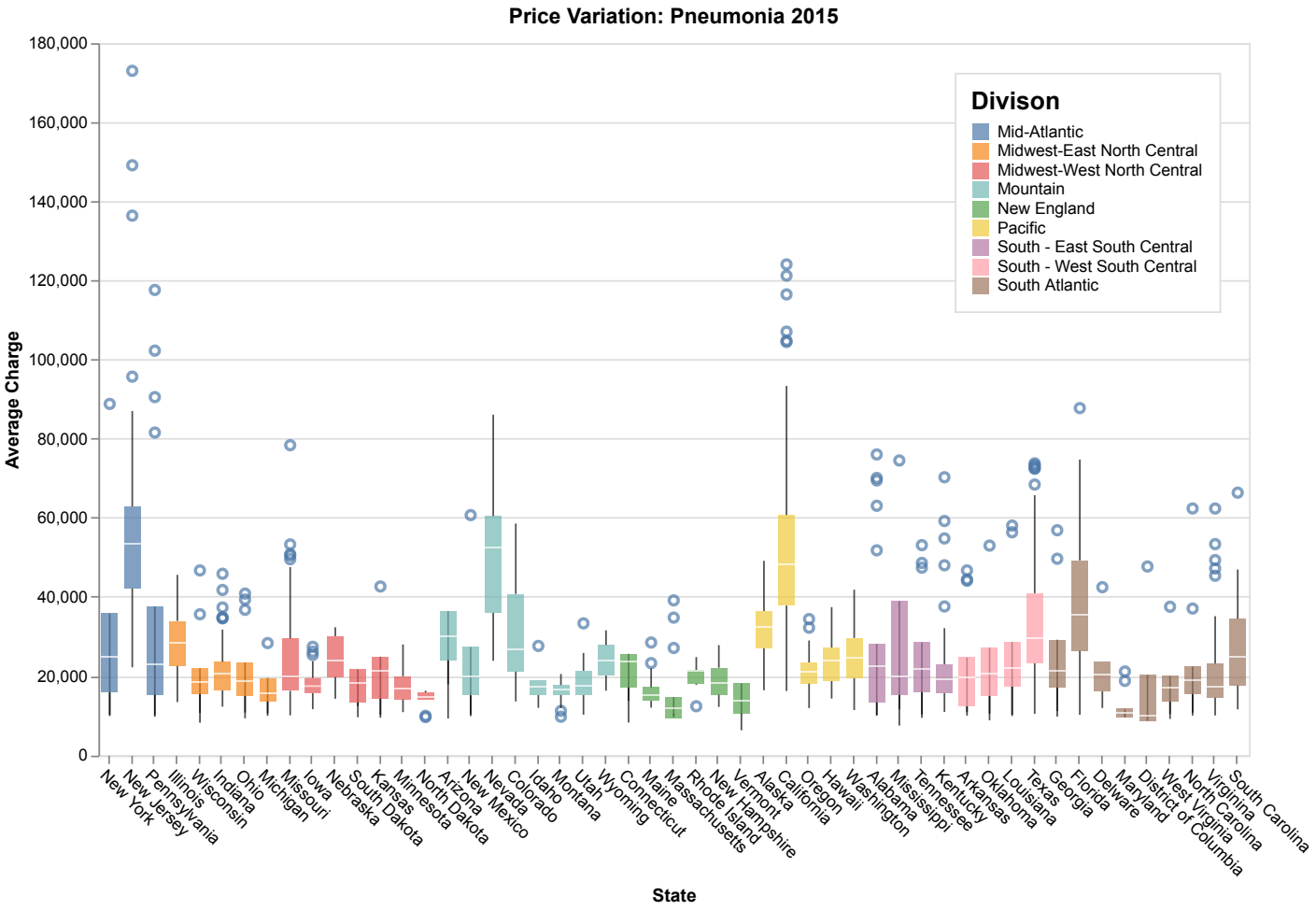
The Centers for Medicare and Medicaid Services (CMS) provides public records of Inpatient and Outpatient Charge Data from health care providers from every state between 2011 - 2016. These files contain information about the DRG (Diagnosis Related Group), health care provider name, city, zip code, hospital referral region (HRR), and the number of discharges (cases), Average Charge, Average Medicare Payment received, and Averaged Total Payment received. To narrow the scope of this project, I used the 2015 data and considered only the Average Charges for a few of the most common diagnoses:

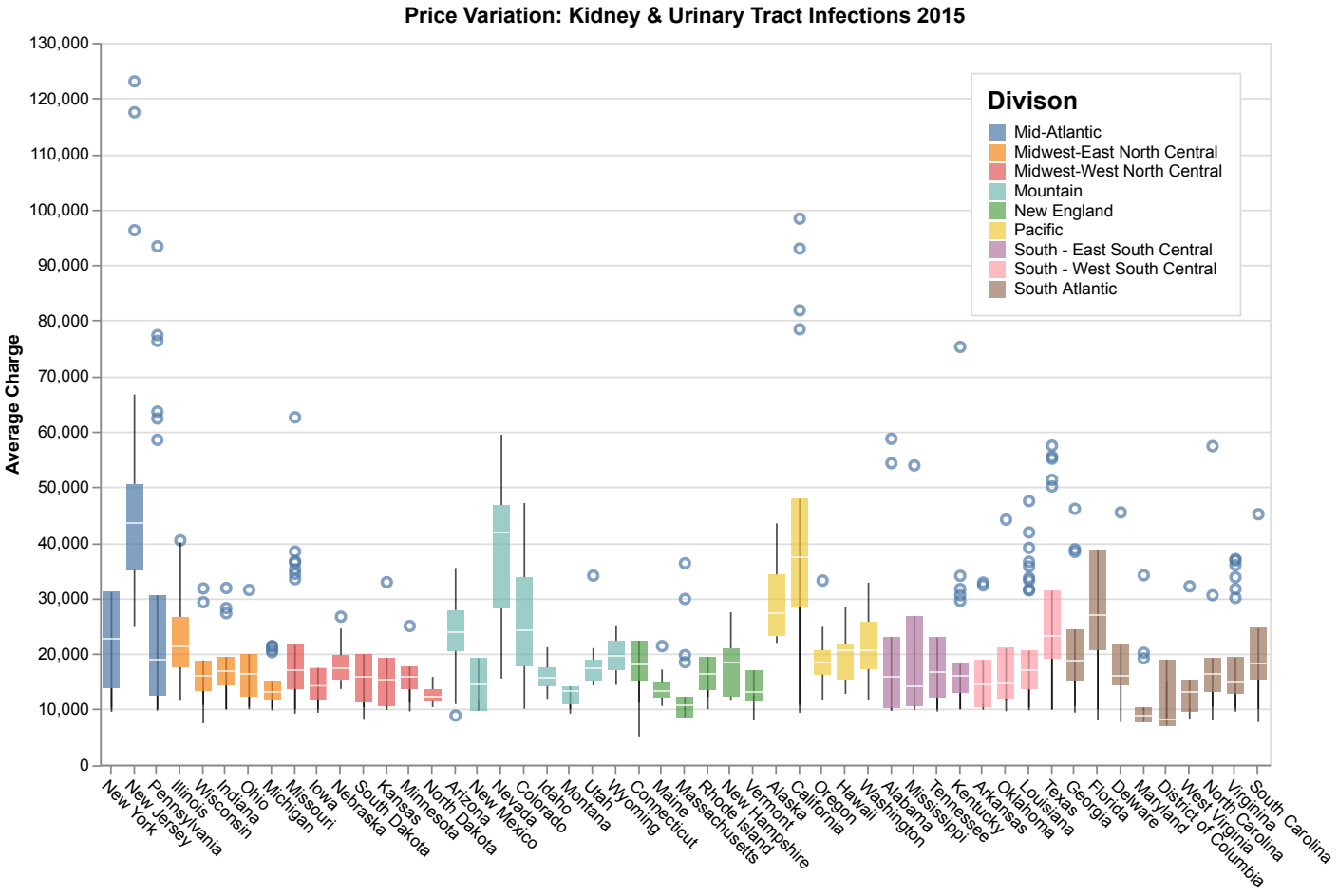
- Simple Pneumonia
- Kidney and Urinary Tract Infections
- Septicemia
- Heart Shock and Failures

The Visualization

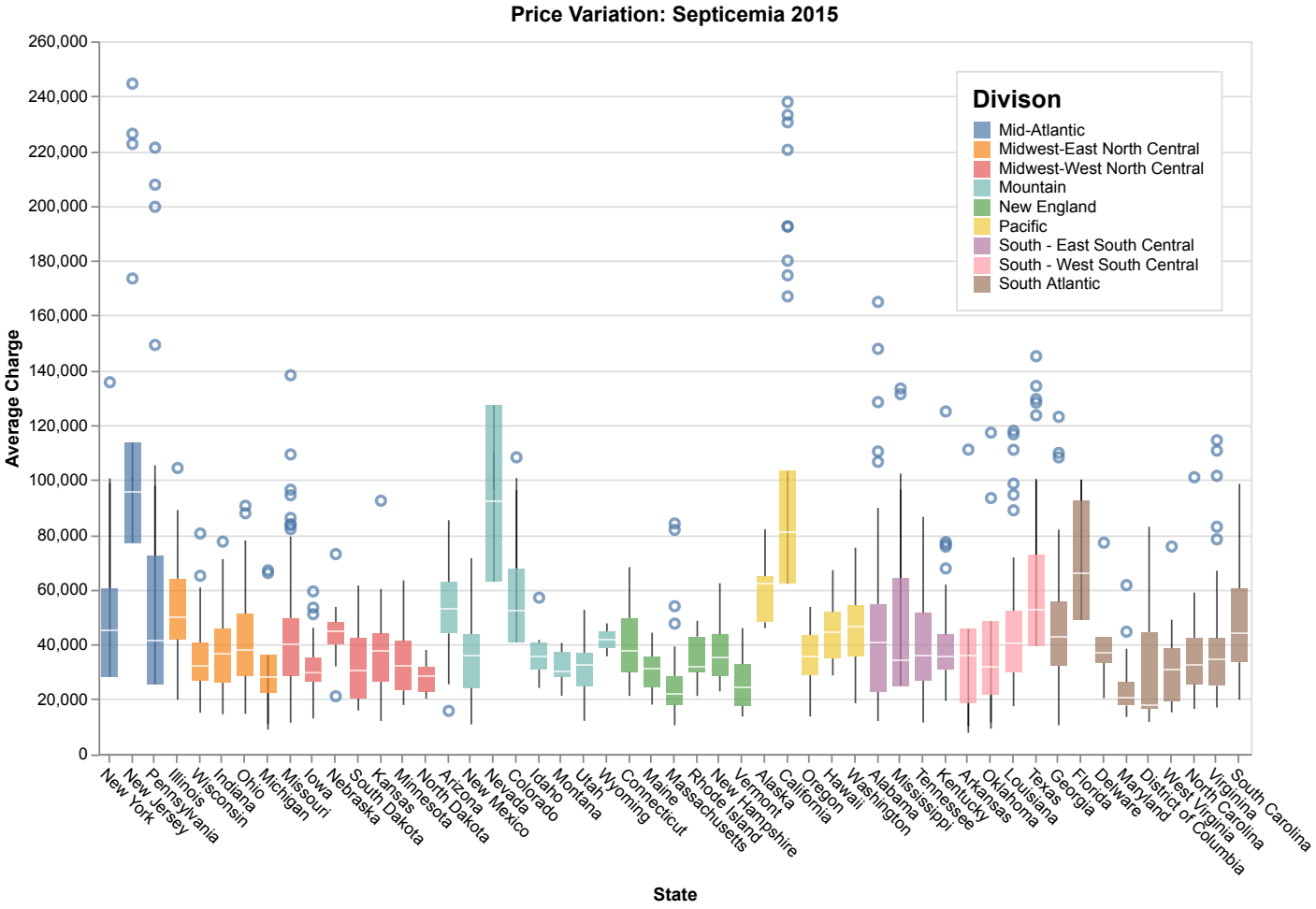
Insert your visualization(s) here.

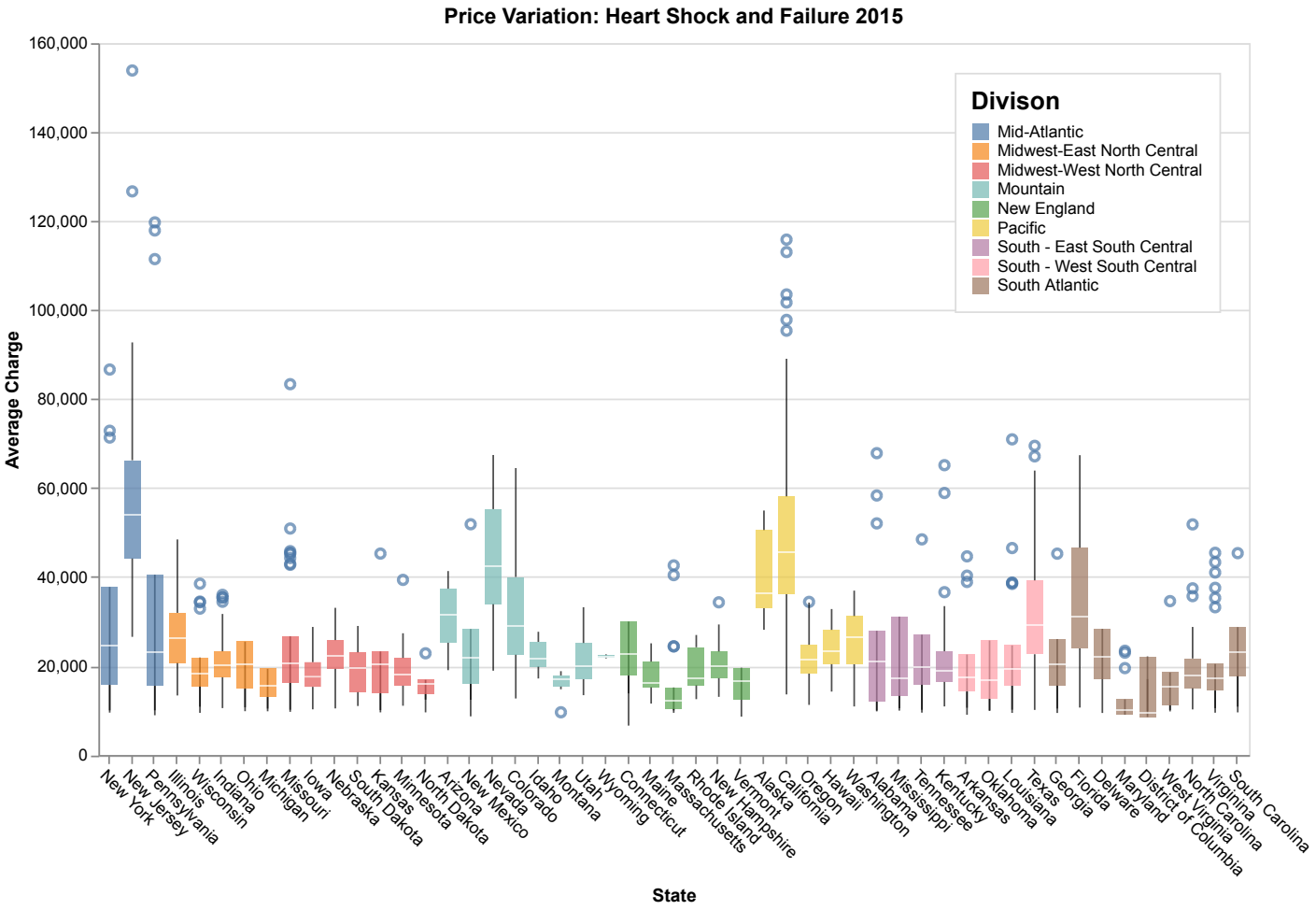
I created four charts, one for each of the most common diagnoses.


Price Variation: Pneumonia 2015
Average Charge
180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
Divison
Mid-Atlantic
Midwest-East North Central
Midwest-West North Central
Mountain
New England
Pacific
South - East South Central
South - West South Central
South Atlantic
New York
New Jersey
Pennsylvania
Illinois
Wisconsin
Indiana
Ohio
Michigan
Missouri
Iowa
Nebraska
South Dakota
Kansas
Minnesota
North Dakota
Arizona
New Mexico
Nevada
Colorado
Idaho
Montana
Utah
Wyoming
Connecticut
Maine
Massachusetts
Rhode Island
New Hampshire
Vermont
Alaska
California
Oregon
Hawaii
Washington
Alabama
Mississippi
Tennessee
Kentucky
Arkansas
Oklahoma
Louisiana
Texas
Georgia
Florida
Delware
Maryland
District of Columbia
West Virginia
North Carolina
Virginnia
South Carolina
State


Price Variation: Kidney & Urinary Tract Infections 2015
Average Charge
130,000
120,000
110,000
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Divison
Mid-Atlantic
Midwest-East North Central
Midwest-West North Central
Mountain
New England
Pacific
South - East South Central
South - West South Central
South Atlantic
New York
New Jersey
Pennsylvania
Illinois
Wisconsin
Indiana
Ohio
Michigan
Missouri
Iowa
Nebraska
South Dakota
Kansas
Minnesota
North Dakota
Arizona
New Mexico
Nevada
Colorado
Idaho
Montana
Utah
Wyoming
Connecticut
Maine
Massachusetts
Rhode Island
New Hampshire
Vermont
Alaska
California
Oregon
Hawaii
Washington
Alabama
Mississippi
Tennessee
Kentucky
Arkansas
Oklahoma
Louisiana
Texas
Georgia
Florida
Delware
Maryland
District of Columbia
West Virginia
North Carolina
Virginnia
South Carolina

State


Price Variation: Septicemia 2015
Average Charge
260,000
240,000
220,000
200,000
180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
Divison
Mid-Atlantic
Midwest-East North Central
Midwest-West North Central
Mountain
New England
Pacific
South - East South Central
South - West South Central
South Atlantic
New York
New Jersey
Pennsylvania
Illinois
Wisconsin
Indiana
Ohio
Michigan
Missouri
Iowa
Nebraska
South Dakota
Kansas
Minnesota
North Dakota
Arizona
New Mexico
Nevada
Colorado
Idaho
Montana
Utah
Wyoming
Connecticut
Maine
Massachusetts
Rhode Island
New Hampshire
Vermont
Alaska
California
Oregon
Hawaii
Washington
Alabama
Mississippi
Tennessee
Kentucky
Arkansas
Oklahoma
Louisiana
Texas
Georgia
Florida
Delware
Maryland
District of Columbia
West Virginia
North Carolina
Virginina
South Carolina
State



Insights

What has your visualization allowed you to discover about your data that help you answer your research question(s)?

Across the four diagnoses, the states of New Jersey, Nevada, California, and Florida consistently have the highest charges, followed by Arizona, Texas, Alaska, and Colorado. Pennsylvania and New York consistently have a large IQR range. Excluding the most expensive states, the average charge for Pneumonia and Heart Shock and Failure treatments ranges between $15,000 and $35,000, between $20,000 and $60,000 for Septicemia, and between $10,000 and $30,000 for Kidney and Urinary Tract Infections. This proves that significant variation in pricing exists in every state, however, few discernable patterns related to geography are evident. In the case of New Jersey, California, and Florida, the higher charges could be potentially be explained by their large populations. This suggests that geographic region may not be strongly correlated with pricing and thus a poor explanatory factor for the variation in charges, but more statistical analysis is needed to support this hypothesis.

Design Justification

Why have you designed your visualizations this way? Consider design approaches (e.g. session 2) and your use of visual variables (channels), layout and interaction.

Visual Variable

I wanted to avoid using an average to summarize the charges in a state since they can hide variation. With averages, one state with several low-cost procedures and one high-cost procedure could appear similar to another state with just a few medium cost procedures. A box-and-whisker plot is a quantitative visual variable shows the range in which the majority of values lie, the range between the min and max value, and any outliers.

I felt that this would be a more accurate representation of aggregated data for each state. Additionally, box plots facilitate additional comparisons of the spread of values, and number and extent of outliers when juxtaposed against each other.

Color

The United States can be divided into four major statistical regions: the Northeast, the Midwest, the South, and the West, and nine underlying geographic regions: New England, Mid-Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain and Pacific (U.S. Census Bureau, 2013). The box plots, however, represent charges that are aggregated by state. Munzner (2015) suggests that spatial location and hue are the most effective channels for representing categorical attributes. Following this logic, I used the color of a box plot to indicate what region the state belongs to. I also placed box plots of states in the same region next to each other on the graph. I used the Tableau10 distinct color scheme since the regions are categorical values and added a legend as a guide. This way, I am able to create visual groupings to support analysis. I also set the opacity to 0.7 for a softer look.

Orientation

I changed the orientation of the state labels to be angled at 45 degrees for ease of reading. Generally, horizontal text is preferable to vertical text which can difficult to read (Tufte, 2001). In the default vertical orientation, state names must be read from bottom to top, which may involve turning one's head and scanning along a jagged edge of text since the only the ends of the words were aligned with the axis. While it was not possible to orient the text horizontally due to space constraints, the 45-degree orientation aligned the start of the state name with the x-axis, hopefully making it easier to scan across and read.

I also had to decide which on which axis to have the box plots. Designer Ann Emery (2017) suggests that sorted, horizontal bar charts are better for the comparison of categorical values, and vertical column bar charts are better for ordinal values since they can be ordered by their natural progression. Even though the states are categorical values, I instead chose to have the states on the horizontal axis resulting in vertical box plots. This made it easier to compare the box plots against the qualitative axis (Average Charge) when it appears on the left instead of at the very bottom of the chart.

Layout

A majority of the box plot data is located at the bottom of the chart. The presence of outliers, however, increased the upper bound of the qualitative scale. As as a result, I had a great deal more whitespace that I was comfortable with at the top of the chart. To utilize this space and create a more compact design, I moved to the legend inside the boundaries of the chart. To make it distinct from the rest of the chart, I kept the white background and added a border outline to the box.

Validation

Evaluate the degree to which your visualization has helped to answer your research questions. What are the strengths and limitations of your design in answering your questions? What might you do differently if you were to do this exercise again?

Strenths

I believe this visualisation is effective in depicting regional variation in charge data among states and regions. The use of a box plot as a visual variable directly aids in visualising and comparing those variations. Secondly, the use of color and colocation for grouping is an effective means of creating distinct categories.

Limitations

First of all, I am still dissatisfied with the level of clutter on the visualisation. The outliers, while interesting artefacts, perhaps serve to create more "chart junk" -- as termed by Tufte (2001) -- than do serve as explanatory factors. In order to improve my data-ink ratio, I would consider removing the outliers, as well as the horizontal grid lines.

Second, the state names are still rather small and difficult to read. This is not easy to resolve, however, as increasing the font size causes the labels to overlap and increasing the space between box plots to allow for a bigger label font results in a bigger graph. I was thinking of allowing

the user to interactively toggle of state names by region to reduce clutter.

Third, I feel having nine regional categories, and thus nine different colors, adds unneeded complexity to the chart. In their study of the limits of human attention, Haroz and Whitney (2012) recommend fewer categories in visualisations to avoid overwhelming the user. Following this advice, I would go back to a design that has only 4 or 5 regions to simplify the visualisation.

Finally, this design doesn't support effective comparison between different DRG types (i.e. Pneumonia versus Septicemia). Currently, the four charts are arranged vertically, and the y-axes are scaled independently. This kind of juxtaposition requires the user to scroll to see each chart and remember important differences. As a result, the user can fail to notice certain details like the fact that Septicemia has much more expensive charges than do the other three DRGs. I would question if it was really necessary to compare different DRGs against one another since the research question is focused on variations between geographic regions. If such a comparison is still desired, a butterfly chart may a more effective tool for comparison.

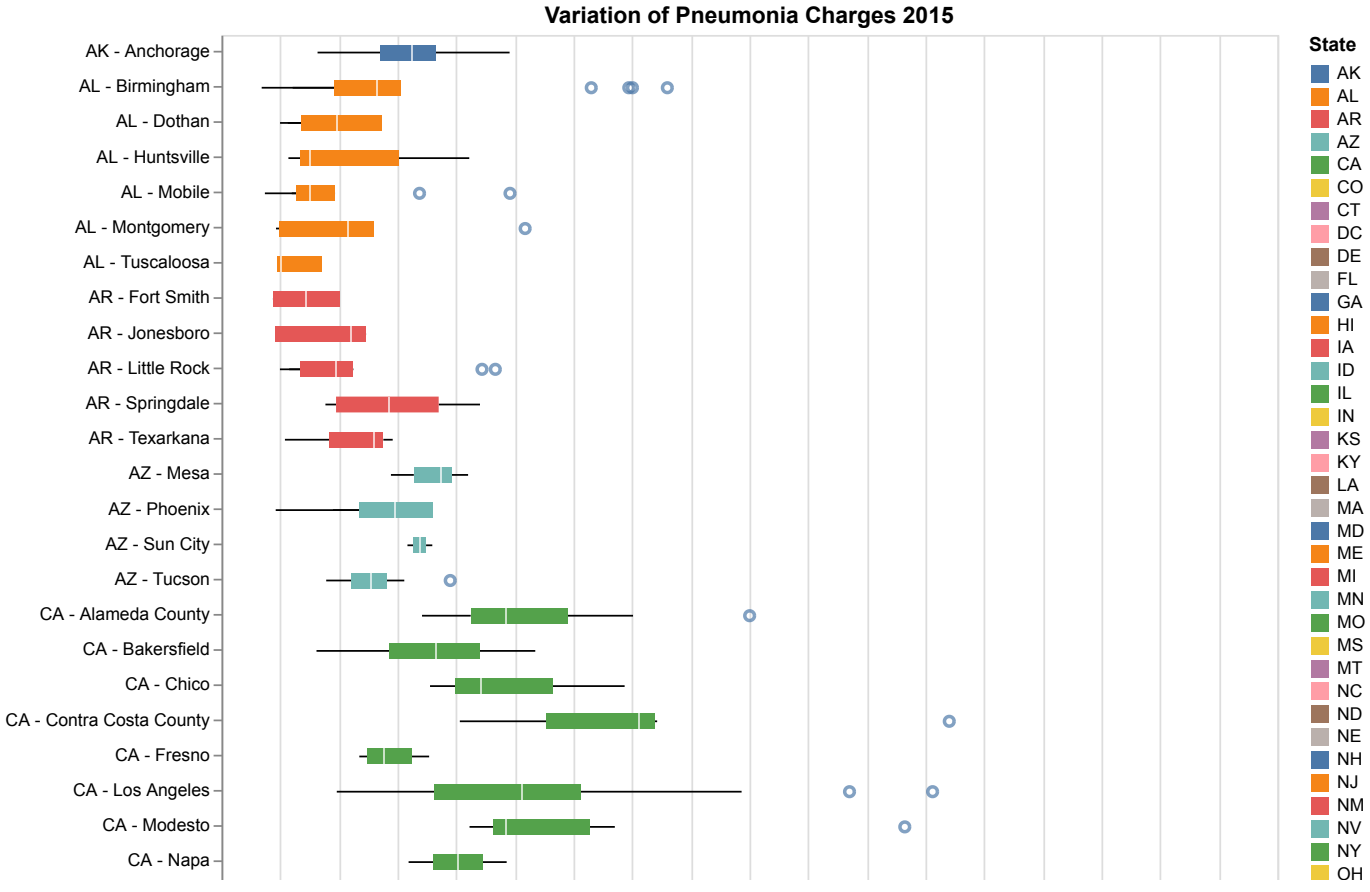
Futher Explorations

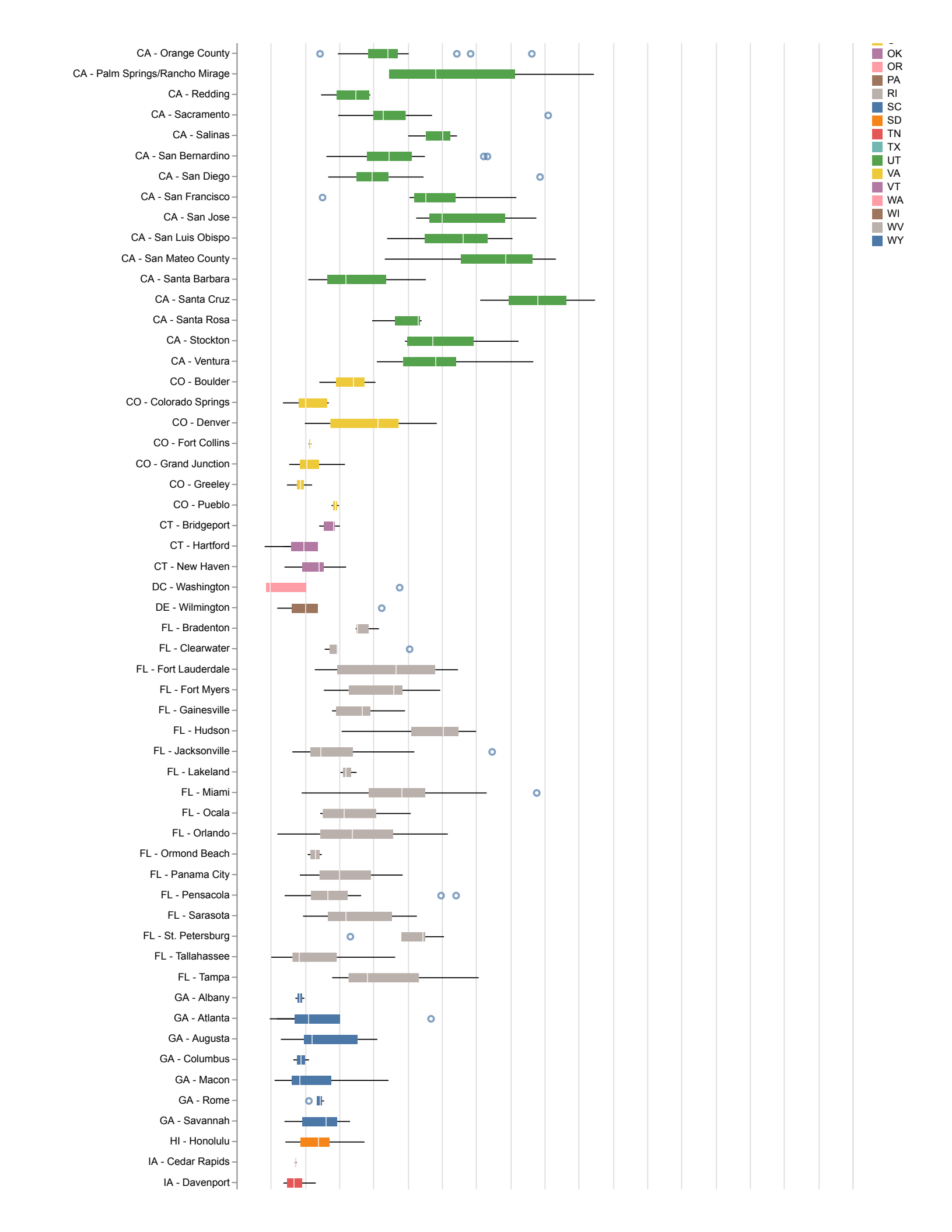
With this visualisation, I have only looked at a small aspect of the available data. The variation in Medicare reimbursements and customer payments could be visualized in a similar manner. Additionally, I would be interested in exploring temporal trends, perhaps using a stream graph to visualize the variation over time. Finally, to add interaction to the existing plots, I believe it would be nice to have a tooltip that displays the precise min, max, median, average, and first and third quartile values when the mouse moves over a box plot.

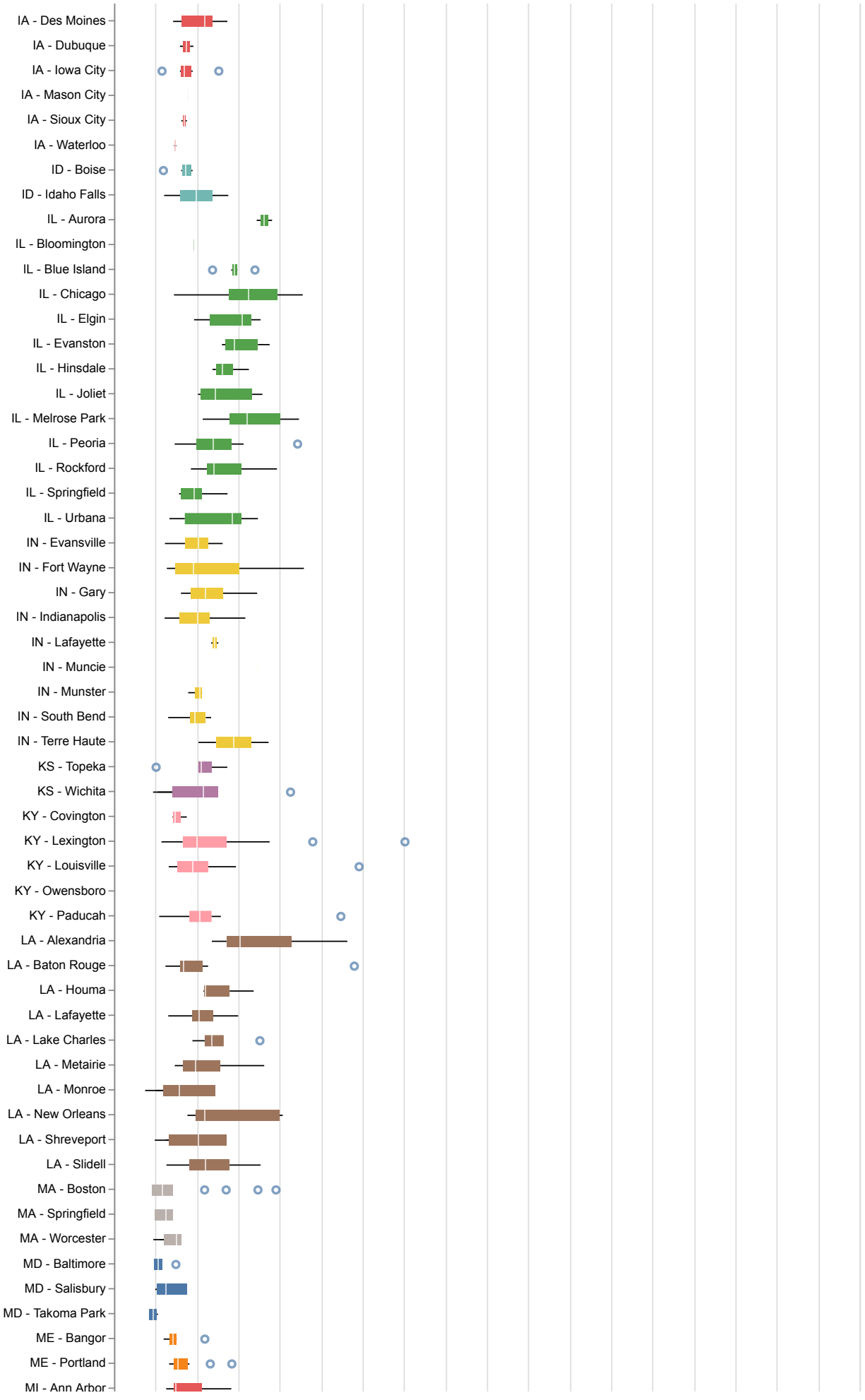
Previous Designs

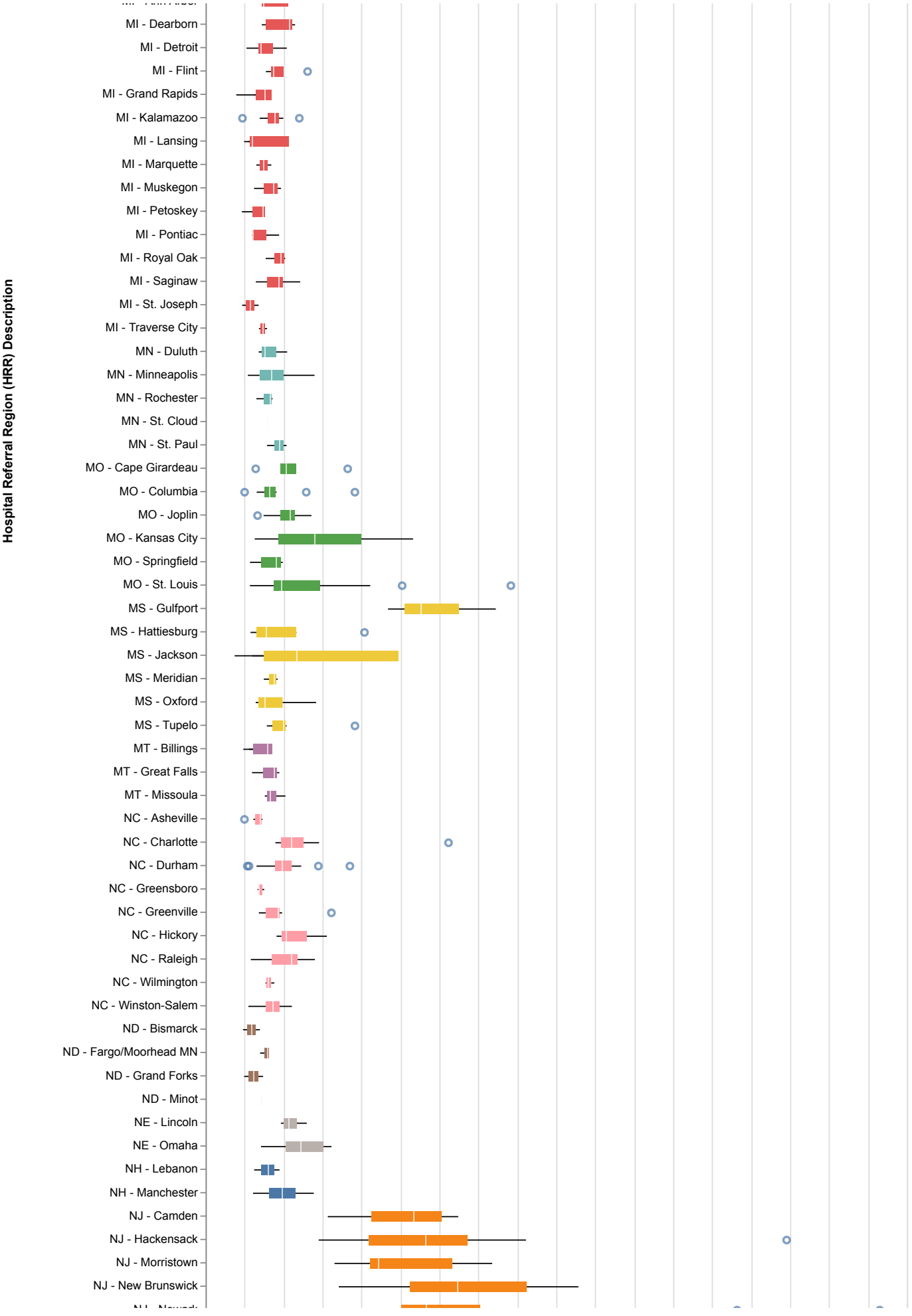
Version 1 : Horizontal Boxplot Chart

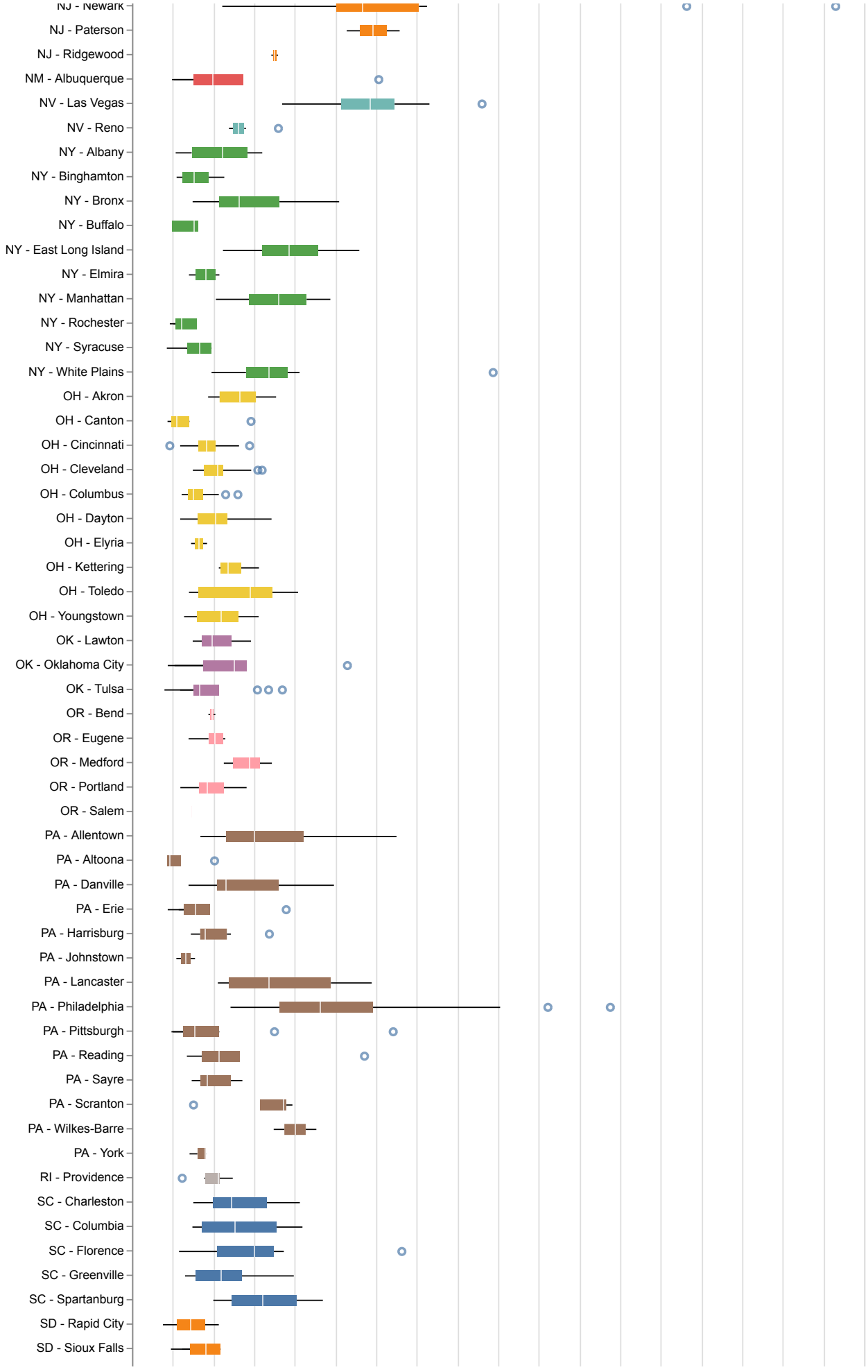
In this first attempt, I compared the variation in charges across the 306 hospital referral regions (HRR) using box plots and used color to group HRR regions by state. While visually interesting, the chart was just too long to be practical. This gave me the idea of trying to arrange the box plots geospatially in a relaxed-spatial grid of the United States as seen in Version 2. (Scroll past this chart)

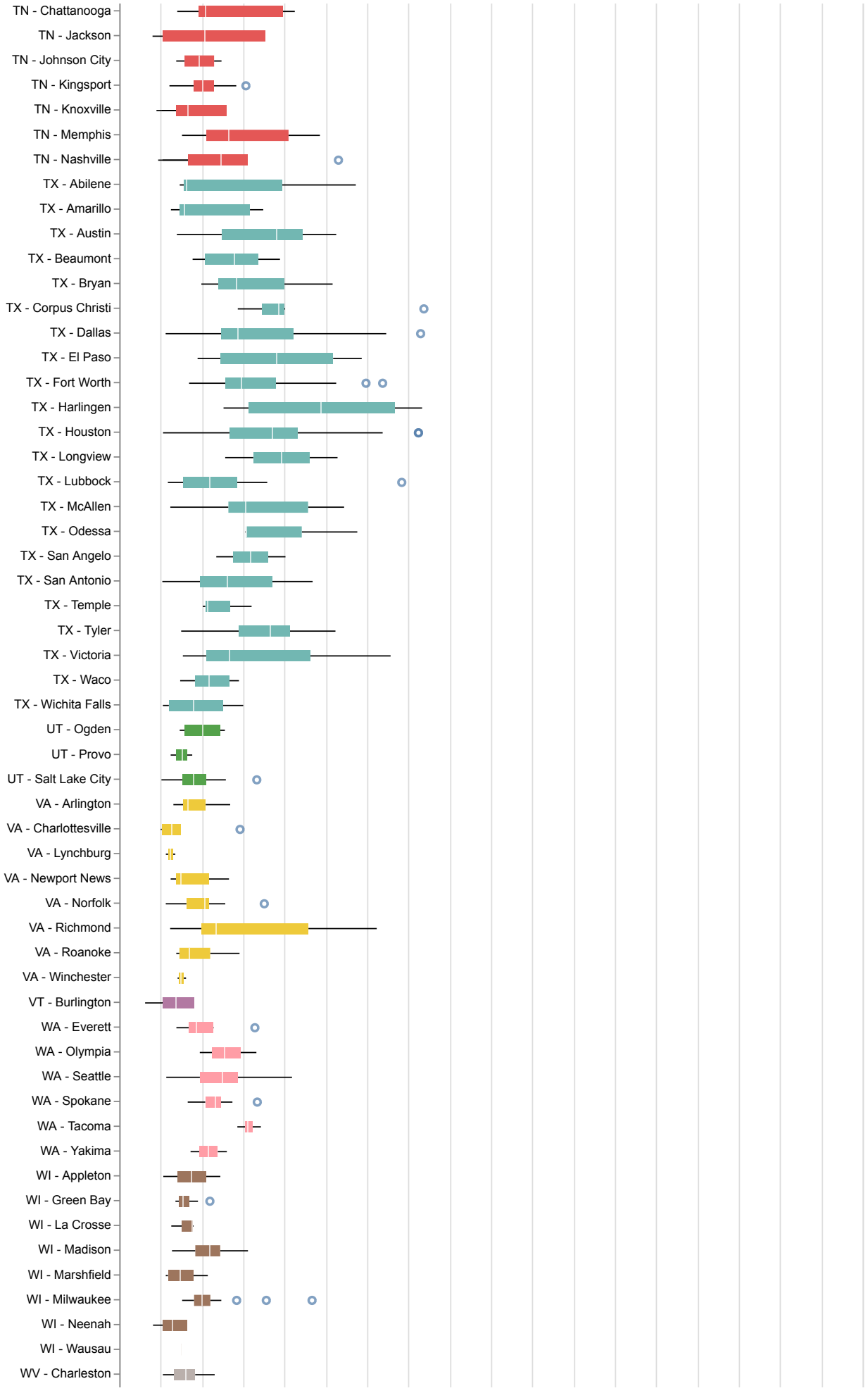


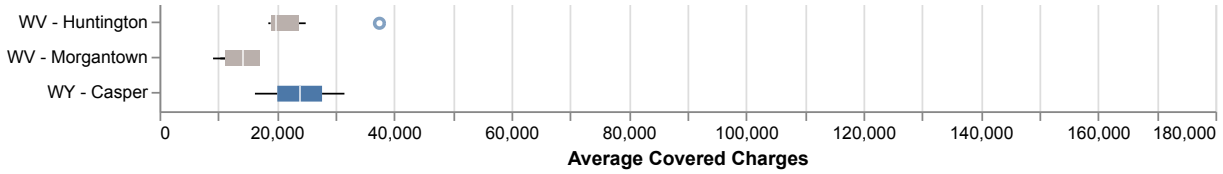

CA - Orange County
CA - Palm Springs/Rancho Mirage
CA - Redding
CA - Sacramento
CA - Salinas
CA - San Bernardino
CA - San Diego
CA - San Francisco
CA - San Jose
CA - San Luis Obispo
CA - San Mateo County
CA - Santa Barbara
CA - Santa Cruz
CA - Santa Rosa
CA - Stockton
CA - Ventura
CO - Boulder
CO - Colorado Springs
CO - Denver
CO - Fort Collins
CO - Grand Junction
CO - Greeley
CO - Pueblo
CT - Bridgeport
CT - Hartford
CT - New Haven
DC - Washington
DE - Wilmington
FL - Bradenton
FL - Clearwater
FL - Fort Lauderdale
FL - Fort Myers
FL - Gainesville
FL - Hudson
FL - Jacksonville
FL - Lakeland
FL - Miami
FL - Ocala
FL - Orlando
FL - Ormond Beach
FL - Panama City
FL - Pensacola
FL - Sarasota
FL - St. Petersburg
FL - Tallahassee
FL - Tampa
GA - Albany
GA - Atlanta
GA - Augusta
GA - Columbus
GA - Macon
GA - Rome
GA - Savannah
HI - Honolulu
IA - Cedar Rapids
IA - Davenport
OK
OR
PA
RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY


IA - Des Moines
IA - Dubuque
IA - Iowa City
IA - Mason City
IA - Sioux City
IA - Waterloo
ID - Boise
ID - Idaho Falls
IL - Aurora
IL - Bloomington
IL - Blue Island
IL - Chicago
IL - Elgin
IL - Evanston
IL - Hinsdale
IL - Joliet
IL - Melrose Park
IL - Peoria
IL - Rockford
IL - Springfield
IL - Urbana
IN - Evansville
IN - Fort Wayne
IN - Gary
IN - Indianapolis
IN - Lafayette
IN - Muncie
IN - Munster
IN - South Bend
IN - Terre Haute
KS - Topeka
KS - Wichita
KY - Covington
KY - Lexington
KY - Louisville
KY - Owensboro
KY - Paducah
LA - Alexandria
LA - Baton Rouge
LA - Houma
LA - Lafayette
LA - Lake Charles
LA - Metairie
LA - Monroe
LA - New Orleans
LA - Shreveport
LA - Slidell
MA - Boston
MA - Springfield
MA - Worcester
MD - Baltimore
MD - Salisbury
MD - Takoma Park
ME - Bangor
ME - Portland
MI - Ann Arbor


Hospital Referral Region (HRR) Description
MI - Dearborn
MI - Detroit
MI - Flint
MI - Grand Rapids
MI - Kalamazoo
MI - Lansing
MI - Marquette
MI - Muskegon
MI - Petoskey
MI - Pontiac
MI - Royal Oak
MI - Saginaw
MI - St. Joseph
MI - Traverse City
MN - Duluth
MN - Minneapolis
MN - Rochester
MN - St. Cloud
MN - St. Paul
MO - Cape Girardeau
MO - Columbia
MO - Joplin
MO - Kansas City
MO - Springfield
MO - St. Louis
MS - Gulfport
MS - Hattiesburg
MS - Jackson
MS - Meridian
MS - Oxford
MS - Tupelo
MT - Billings
MT - Great Falls
MT - Missoula
NC - Asheville
NC - Charlotte
NC - Durham
NC - Greensboro
NC - Greenville
NC - Hickory
NC - Raleigh
NC - Wilmington
NC - Winston-Salem
ND - Bismarck
ND - Fargo/Moorhead MN
ND - Grand Forks
ND - Minot
NE - Lincoln
NE - Omaha
NH - Lebanon
NH - Manchester
NJ - Camden
NJ - Hackensack
NJ - Morristown
NJ - New Brunswick


NJ - Newark
NJ - Paterson
NJ - Ridgewood
NM - Albuquerque
NV - Las Vegas
NV - Reno
NY - Albany
NY - Binghamton
NY - Bronx
NY - Buffalo
NY - East Long Island
NY - Elmira
NY - Manhattan
NY - Rochester
NY - Syracuse
NY - White Plains
OH - Akron
OH - Canton
OH - Cincinnati
OH - Cleveland
OH - Columbus
OH - Dayton
OH - Elyria
OH - Kettering
OH - Toledo
OH - Youngstown
OK - Lawton
OK - Oklahoma City
OK - Tulsa
OR - Bend
OR - Eugene
OR - Medford
OR - Portland
OR - Salem
PA - Allentown
PA - Altoona
PA - Danville
PA - Erie
PA - Harrisburg
PA - Johnstown
PA - Lancaster
PA - Philadelphia
PA - Pittsburgh
PA - Reading
PA - Sayre
PA - Scranton
PA - Wilkes-Barre
PA - York
RI - Providence
SC - Charleston
SC - Columbia
SC - Florence
SC - Greenville
SC - Spartanburg
SD - Rapid City
SD - Sioux Falls


TN - Chattanooga
TN - Jackson
TN - Johnson City
TN - Kingsport
TN - Knoxville
TN - Memphis
TN - Nashville
TX - Abilene
TX - Amarillo
TX - Austin
TX - Beaumont
TX - Bryan
TX - Corpus Christi
TX - Dallas
TX - El Paso
TX - Fort Worth
TX - Harlingen
TX - Houston
TX - Longview
TX - Lubbock
TX - McAllen
TX - Odessa
TX - San Angelo
TX - San Antonio
TX - Temple
TX - Tyler
TX - Victoria
TX - Waco
TX - Wichita Falls
UT - Ogden
UT - Provo
UT - Salt Lake City
VA - Arlington
VA - Charlottesville
VA - Lynchburg
VA - Newport News
VA - Norfolk
VA - Richmond
VA - Roanoke
VA - Winchester
VT - Burlington
WA - Everett
WA - Olympia
WA - Seattle
WA - Spokane
WA - Tacoma
WA - Yakima
WI - Appleton
WI - Green Bay
WI - La Crosse
WI - Madison
WI - Marshfield
WI - Milwaukee
WI - Neenah
WI - Wausau
WV - Charleston



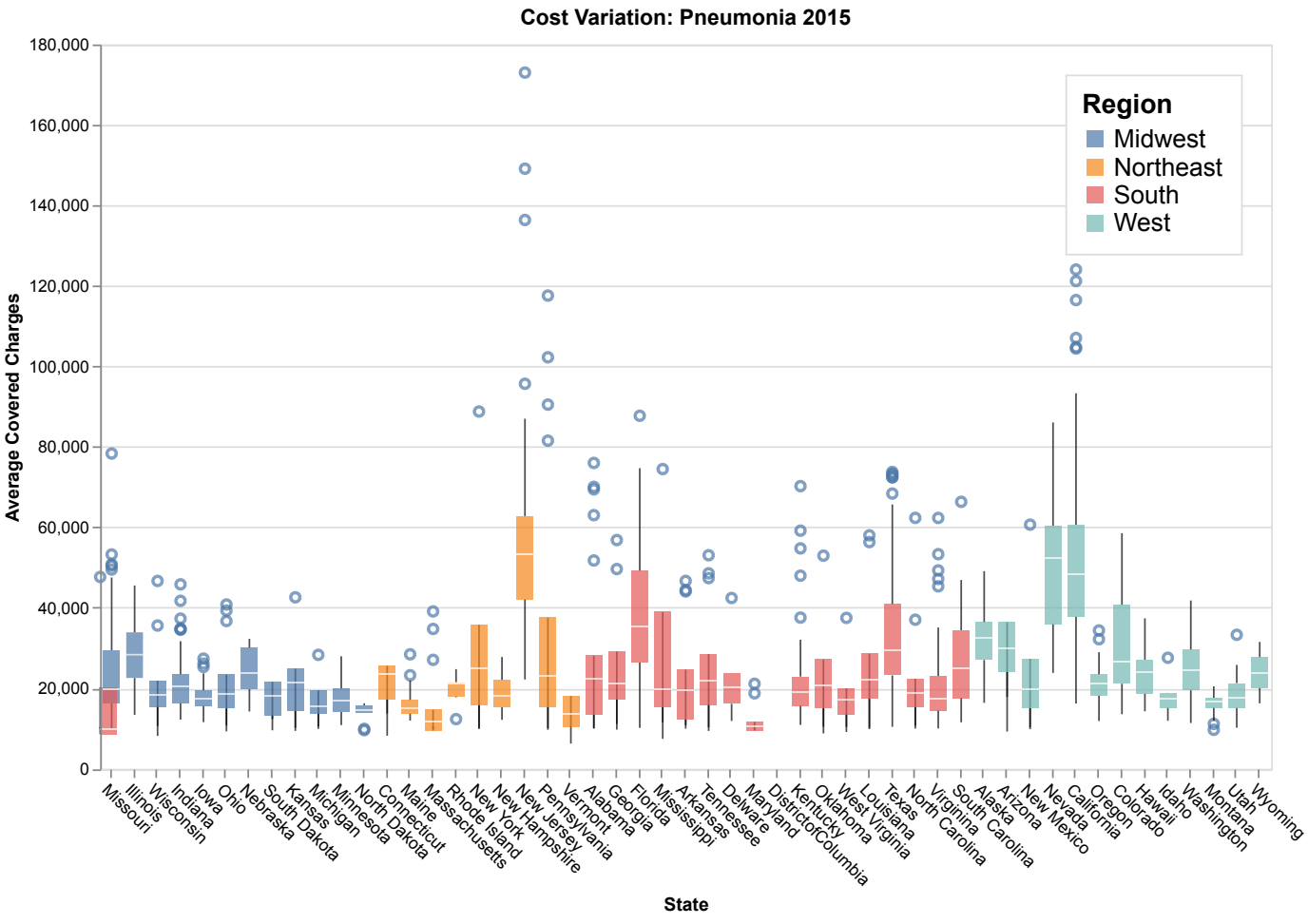
Version 2 : Relaxed Grid Layout

Based on the idea of using small multiples of a plot in a grid layout that only loosely approximates geographical regions (Meulemans et al., 2017), I created an interactive grid version of the United States. Unfortunately, I was unable to get this relaxed-geographically grid layout to display properly. I believe it is in part due to the assumption in Vega that each subplot has the same axis values. So even though each state has only a subset of the data, the y-axis was still scaled for 306 values, resulting in boxplots being squished together. Not only was the size of this graphic a problem, but it would have been difficult to embed the necessary labels without creating even more visual clutter.

```
ReferenceError: vegaEmbed is not defined
```

Version 3: Vertical Column plots

Instead of 306 HRR regions, I finally chose to aggregate data by the 50 U.S. states (and Washington D.C.) for 51 categorical values. States were grouped by color into super-regions (Midwest, West, South, etc). I also made the move to a vertical column plot so that all the data is visible at once. While I like the visual simplicity of four major regions, I was concerned that four was too few and could lead to inappropriate comparisons between dissimilar states. Therefore, in the next iteration, I chose to use nine geographic subregions (Pacific, Mountain, Northeast, etc).



References

Add references from literature that you have used to support your design justification.

Ann K.Emery (2017) When to Use Horizontal Bar Charts vs. Vertical Column Charts [Online]. *Depict Data Studio*. URL http://depictdatastudio.com/when-to-use-horizontal-bar-charts-vs-vertical-column-charts/ (accessed 5.5.19).

Haroz, S., and Whitney, D. (2012) How capacity limits of attention influence information visualization effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 18(12) pp.2402-2410.

Meulemans, W., Dykes, J., Slingsby, A., Turkay, C. and Wood, J. (2017) Small multiples with gaps. *IEEE Transactions on Visualization and Computer Graphics*, 23(1), pp.381-390.

Munzner, T. (2014) Visualization Analysis and Design. CRC Press LLC, Florida, UNITED STATES.

Tufte, E. (2001) The Visual Display of Quantitative Information, Graphics Press.

United States Census Bureau, Geography Division. (2013) "Census Regions and Divisions of the United States" (PDF). Retrieved May 04, 2019.