Geographical Analysis of Patient Data in North America for the testing of protein PD-L1

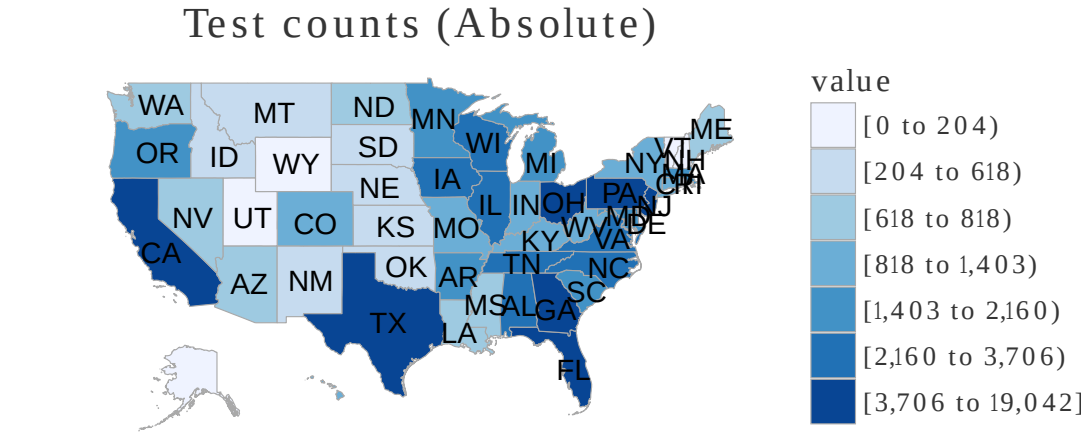
Niall Tyndall

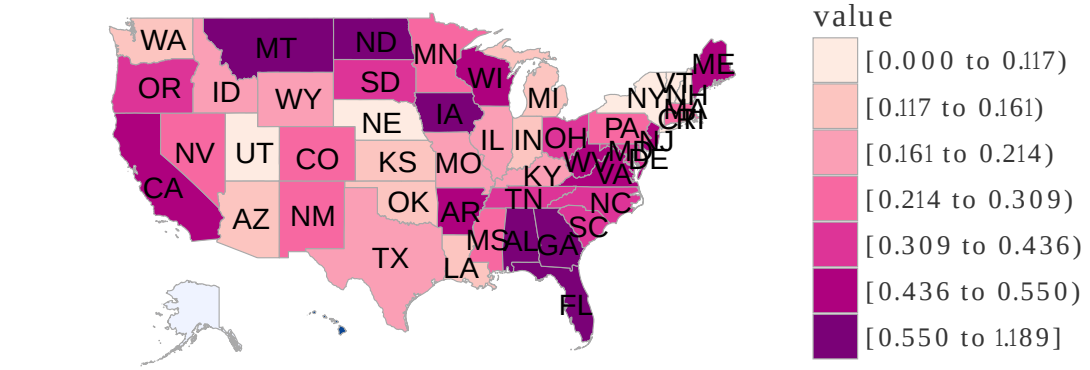
Introduction

This report outlines a brief investigation of 100581 rows of patient data, spanning over a date range of 384 days. The majority of this data set has been collected from Northern America with a few exceptions (after some analysis), such as the country of Peurto Rico, and other unknown zip codes from the **Client.Zip** header. With this in mind, the format of this report is designed to group all the data geographically, so we can investigate further and drill down specific metrics by this *grouping* method. A minor assumption is considered to cast the data into such a form, therefore data that has **no** well defined region from the ordering physician is neglected. We could go through a process to estimate the error by considering the ratio of tests to states and during the analysis we could incorporate this error, although it should be negligible enough to avoid this task. How many records are *thrown* away from the analysis? In total, there are 0 ZIP codes that could not be determined, and as such, have been given a tag of *unknown* and are also considered in the analysis. Finally, 17 could not be determined at all, resulting in a total of 17 missing entries (or 0.02%).

The data set that is used in the analysis is therefore almost identical to the original set, with every record having an additional piece of information, a mapping of the zip code to a US State.

```
# Get ZIP code hash and create map
mapInfo <- map.data$Client.Zip %>%
  pdata::create_map()
```






```
# Append actual states to the main data set
map.data$state <- sapply(
  X = map.data$Client.Zip,
  FUN = function(x) x %>% mapInfo$find()
  )
```

As shown above, we pass all the zip codes into function *create_map*, which takes ZIP codes and maps them to a US state through the *zipcode* package. We obtain a hash lookup of zip codes to states and append this information on through the vectorized *sapply* functionality. This function also generates the American map of individual test counts, where I have generated both one for absolute counts and also counts, normalized by each state population per 1000 people. Generally the two images are quite similar, however the most stark difference is the flip between the central American states after normalization, such as Texas, Montana, and North Dakota. While the coastal counties of California, Florida, and Georgia (generally the whole east coast) remaining quite dominant after normalization.

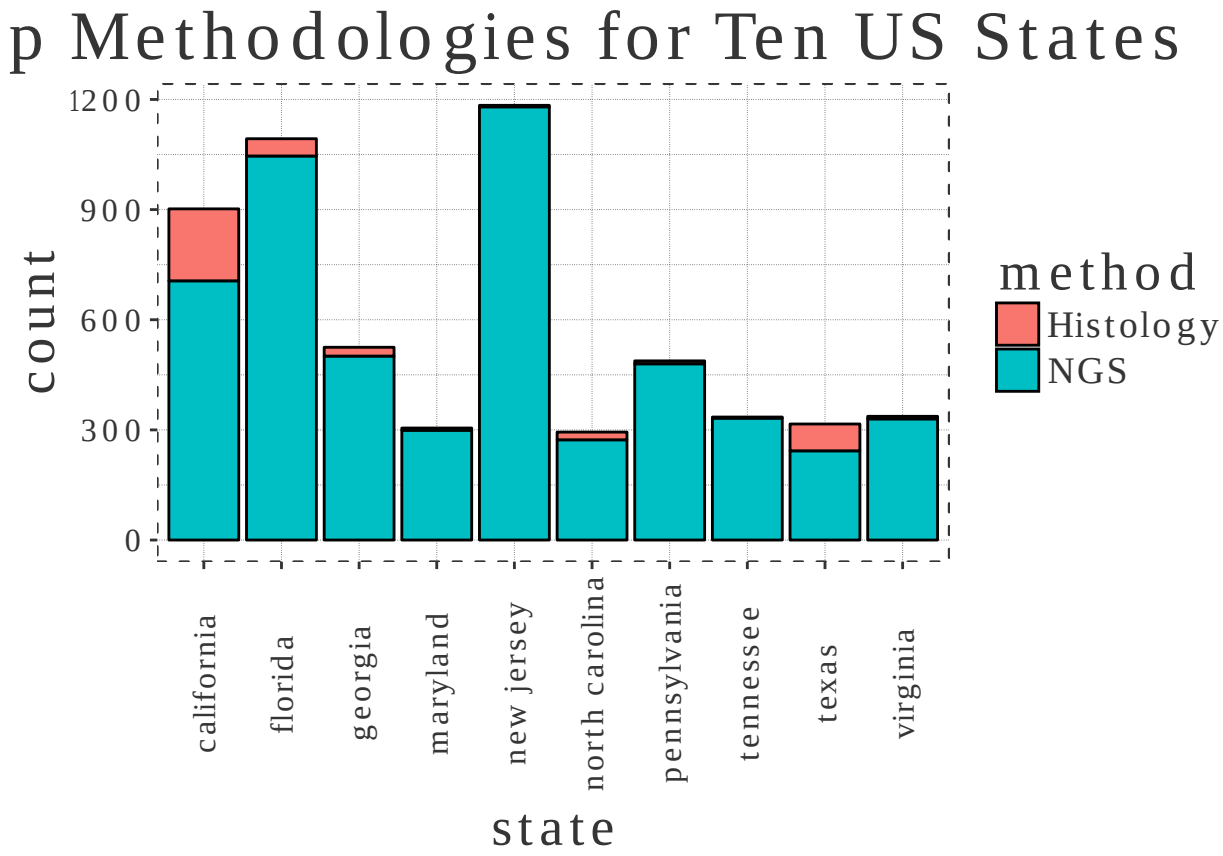
Results

In this section I will outline a few different ideas and functionality built, utilizing the new, augmented data set called *map.data*. All functionality will take this data frame and break it down into individual states for analysis, depending on the analysis being carried out.

1) Methodologies

After some basic manipulation, we can see clearly that the four antibodies used as a diagnostic all use the same methodology; *IHC*, representing 92.2% of the total data set. PD-L1 is tested using 2 other methodologies, defined here by Histology and Next-generation sequencing (NGS).

```
# Return results from top methodologies per state
mResults <- map.data %>%
  pdata::method_plot()

mResults
```



Above, we plot the results as a bar chart for the remaining two methodologies for the 10 highest state counts. In taking this approach we can see the difference between these results, without being swamped by the remaining 92.2% IHC method. We could further this analysis by considering test specifics for these methods as opposed to IHC by looking at the anatomical site the sample was removed from etc. It is clear from the chart that New Jersey and Florida (East coast) have the highest usage of NGS methods, and California and Texas have a higher percentage use of Histology (West coast).

2) Turnaround times

The next investigation revolved around looking at the turn aroundtimes recorded for each test on a US state basis. Nothing overly complex was carried out with the data here, although more effort could be spent to give some sort of errors on the results of the box plots. No data cleansing of such was carried out, as no preliminary evidence to support this would have enhanced the results. One such solution could have analysed each state in turn, and then tried to fill in the missing blanks between the number of days between result reported and number of days between test order date from the specimen collection date. In other words, either calculating averages on the data set available and filling in the blanks with a simple median value, or

looking at a time series approach and looking at surrounding x days of a missing value and fitting some sort of spline to estimate that value. However, due to time constraints, none of this was implemented and more than likely, would have made little difference to the results.

```
rResults <- map.data %>%
  pdata::reporting_time()
```

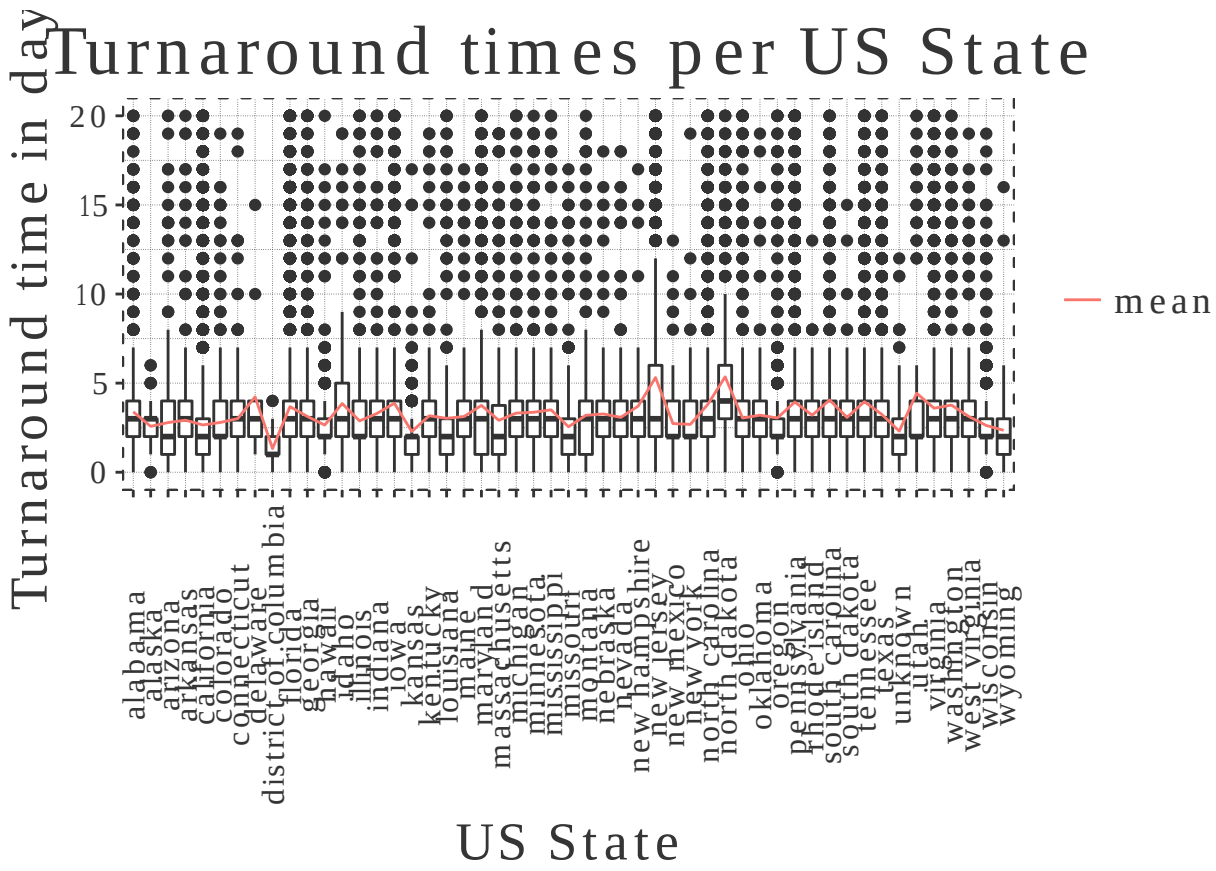
```
## Warning: Removed 985 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 985 rows containing non-finite values (stat_summary).
```

```
rResults
```

```
## Warning: Removed 985 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 985 rows containing non-finite values (stat_summary).
```



This analysis was relatively simple, from the snippet above, a call to *reporting_time* with the map data provides a boxplot for every US state, recording the aggregated statistics of turnaround time in days. The extent of the whiskers here is not the typical 100%, but instead is set to 95% to see some of the outliers from the quartiles. The plot has also been cut to inspect the boxplots easier. Most of the data is infact relatively similar across the board here, with nothing really catching the eye. I have included a summary function of the mean overlaid with the boxplots to see the averages per state. While most median values lie around the 3 / 4 day turnaround period, the states of North Carolina and New Jersey are slightly higher, around 6 days turnaround.

```
tat.anom <- map.data %>% subset(
  map.data$Tat %>% `>`(50),
  select = c(
    "Specimen.Type", "Methodology", "Tat"
```

```
  )
)

knitr::kable(
  x = tat.anom[tat.anom$Tat %>% order %>% rev, ],
  row.names = FALSE,
  align = "c",
  caption = "Possible anomalous values of turn around times that are above 50 days."
)
```

Table 1: Po
above 50 da

Specimen.Type	Methodology
Paraffin Tissue	IHC
Paraffin Tissue	IHC
Paraffin Tissue	IHC
Tissue	IHC
Paraffin Tissue	IHC
Paraffin Tissue	IHC
Paraffin Tissue	IHC
Paraffin Tissue	IHC
Paraffin Tissue	IHC
Paraffin Tissue	IHC

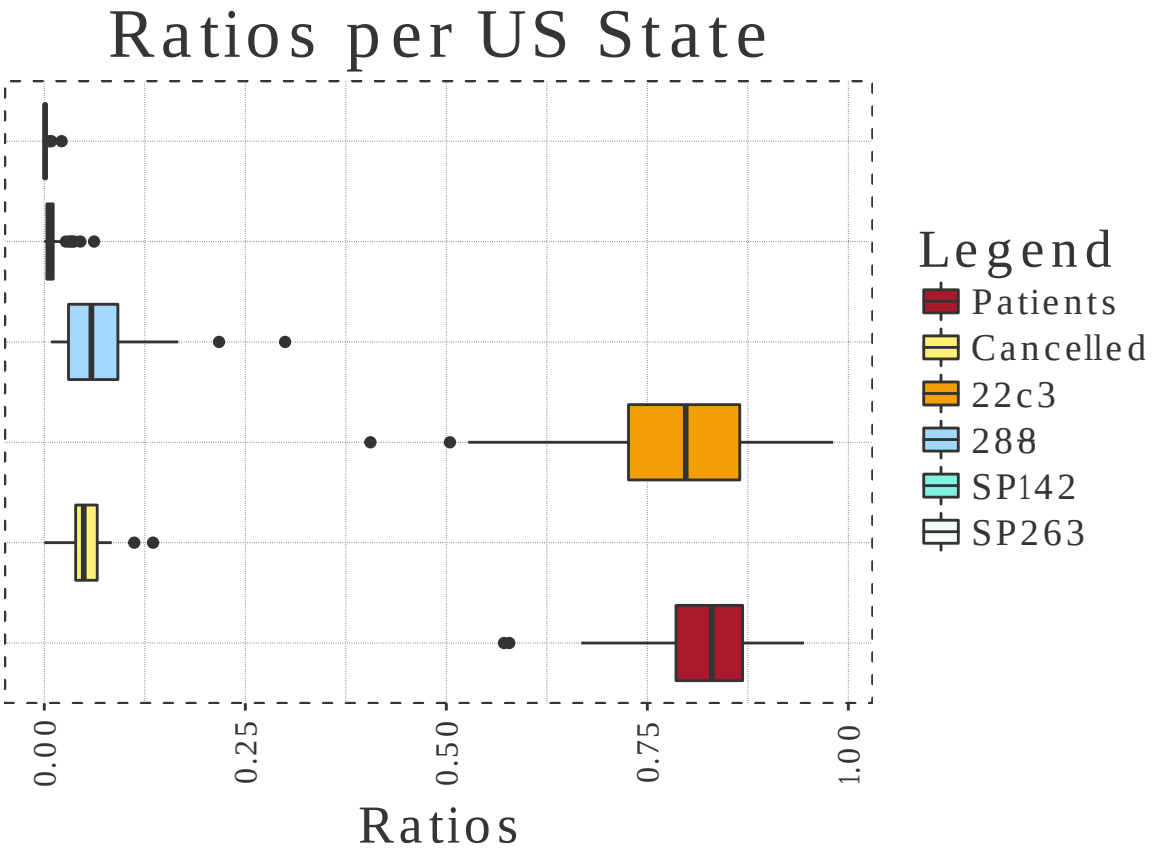
As this plot has be en chopped at r oughly 20 days, there are still values that exceed this. The above snippet provides s

3) State Data Spreads

This section is useful to provide a high level overview of some of the data, again on a state by state basis, by looking at well defined results and reporting on states that deviate largely from the norm. I have chosen to use boxplots again to display the deviation of particular results, and will take simple ratios of insightful data to help us build a better picture of particlar features.

```
ratioResults <- map.data %>%
  pdata::ratio_features()

suppressWarnings(ratioResults$plot)
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x2d

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x2d

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x
## $y, : font width unknown for character 0x2d
```



The above snippet calls some function *ratio_features* with map data to return a boxplot of defined ratios to take per US state, while returning the maximum results in the form of a data frame. This could easily be extended to included, say the 3 highest plus 3 lowest ratios, or even those that closely match the norm for particular states depending on the analysis being carried out. However, for the purpose of this report I will consider the maximum. Looking bottom to top; the first boxplot (red) represents the number of patients actually carrying out tests in each state, and the second boxplot (yellow) is the fraction of tests that have been cancelled. The remaining four boxplots represent the antibody being used as a diagnostic in each test.

```
ratioData <- ratioResults$data
tabNames <- ratioData %>% names

blnk <- data.frame(
  uniqueState = "...",
  patientToTest = "...",
  stringsAsFactors = FALSE
)

for (i in 1:(ratioData %>% length)) {
  ratioData[[i]]$type <- NULL
  ratioData[[i]] <- rbind(
    ratioData[[i]][1:2, ],
    blnk,
    ratioData[[i]][3, ],
    blnk,
    ratioData[[i]][4:5, ]
  )

  names(ratioData[[i]]) <- c(tabNames[i], "Ratio")
```

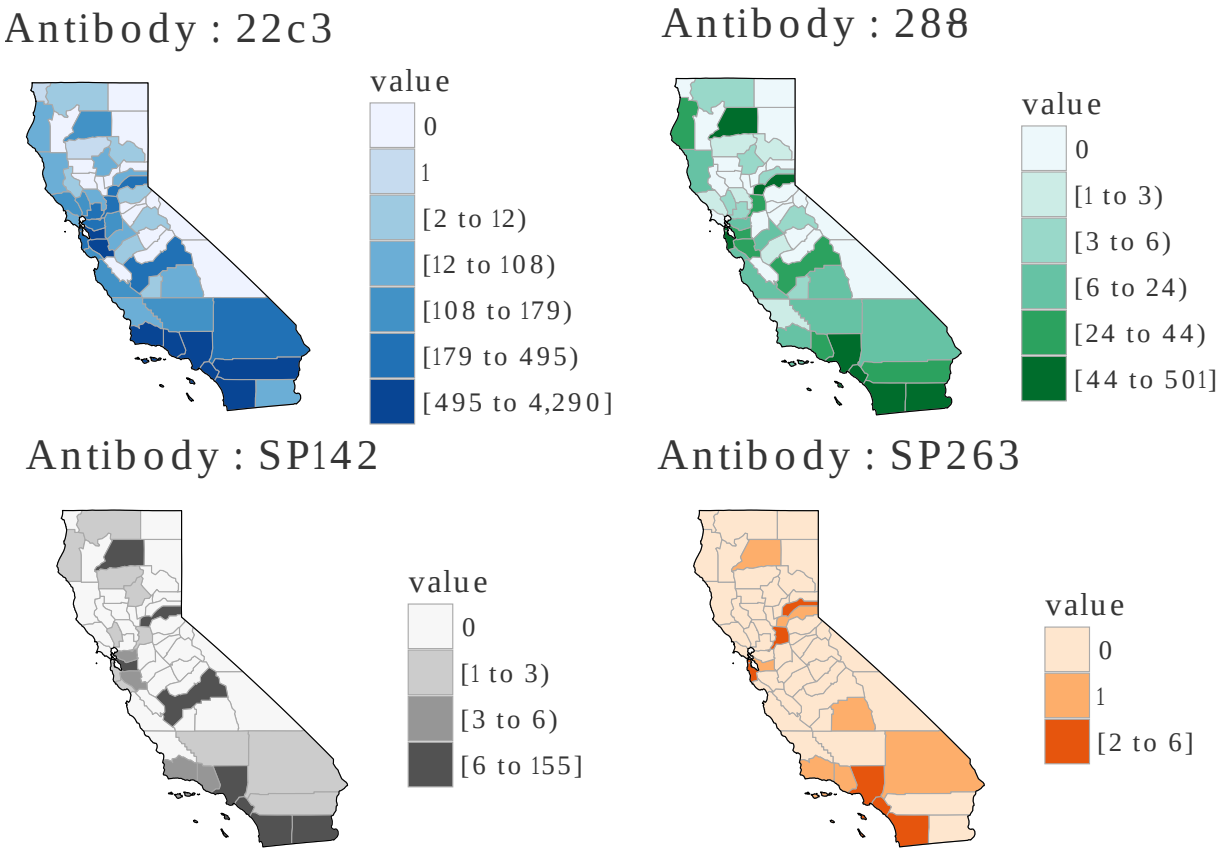
```
}

knitr::kable(
  x = ratioData %>% purrr::reduce(cbind),
  row.names = FALSE,
  align = "c",
  caption = "The table of extremeties of boxplot information that includes the ratio data."
)
```

Patients	Ratio	Cancelled	Ratio	22c3	Ratio
south dakota	0.944805194805195	maryland	0.135215453194651	district of columbia	0.9811320754
wyoming	0.936170212765957	unknown	0.111888111888112	south dakota	0.9707792207
...	...	...	...	...	...
mississippi	0.841843088418431	illinois	0.0496644295302013	california	0.7978153555
...	...	...	...	...	...
new jersey	0.57823596792669	wyoming	0.0106382978723404	new jersey	0.5044673539
arizona	0.571782178217822	delaware	0	unknown	0.4055944055
The above table	represents some of th	e results of	the boxplot, and can g	ive some meaning to thes	e results. For

4) Specific State Information (California)

This final section is to help visualise more of the data on a more granular level. The data set that we have been working with *map.data*, is now not enough for this section, and we must now involve individual county information in order to plot these results. This is probably a little more complex than the previous investigations, as a number of edge cases need to be handled for the general case. A default state is currently employed, and we look at California, as from the world map has the highest absolute tests per US state. However, it is possible to supply the input argument **currentState** to this function with any of the states in map.data (excluding unknown of course). To do so, I have mapped the initial ZIP code, as considered earlier, to a US standard FIPS county code. Further checks to look for missing FIPS information must also be considered to create the data set required for the packaged **choroplethr**. I have written this function to look at the antibodies used in each test geographically across the state, however this could easily be extended.

```
# Look at the most populous counts - california
california <- map.data %>%
  pdata::state_plot()
```



The above snippet will create the 2x2 plot of california for each of the 4 antibodies considered in the data set. It is clear to see from the boxplot of ratios across all states also conforms to the results found here, where Antibody 22c3 is vastly more used in testing compared with the likes of SP263, where only single values are obtained across the counties.

Table 3: The table of highest frequency counts for each of the 4 antibodies considered during testing.

22c3	Value	28-8	Value	SP142	Value	SP263	Value
Los Angeles	4290	Orange	501	Shasta	155	Los Angeles	6
Orange	2404	Los Angeles	311	Orange	61	San Francisco	3
San Diego	1336	Shasta	160	Los Angeles	48	San Diego	3
Santa Clara	818	San Francisco	159	Imperial	20	Orange	3

The patterns across the antibodies used in testing across California are quite consistent, and even from the table you can see the most frequent occur in Los Angeles and Orange County. This helps provide a more granular vision of the data, right down to the county level.

Conclusions

This final section is simply a brief summary of the analysis carried out. I have taken the route of analysing this data based on geographic location, showing some of the possibilities that can be achieved through a rough, high level perspective. The investigation can be carried through further for most of these sections, incorporating other features from the data set and testing hypotheses and trying to identify further correlations.