Deloitte.

Data Visualization for Data QC

March, 2012

Steve Berman
Deloitte Consulting

Agenda

- Overview
- Reasons to Use Data Visualization in Data Cleansing
- Examples
- Tools
- Wrap-Up

How QC has worked up to now

- Checking data for reasonability is an important part of any modeling process
  - Can't be avoided – no such thing as "perfect data"
  - Can be time intensive – the "preprocess" portion of any project takes 60-80% of the total project time
- QC often relies on individual queries that test for specific conditions, or filters in Excel

```
SELECT policy_no, pol_eff_date, tot_incurred_loss
FROM claim_data
WHERE tot_incurred_loss) > 0
ORDER BY DESCENDING tot_incurred_loss;
```

```
proc summary data=pol_info;
   by company state;
   var pol_ct earned_prem written_prem;
   output out=pol_info_sum (drop=_type_ _freq_) sum=;
run;
```

Introducing data visualization

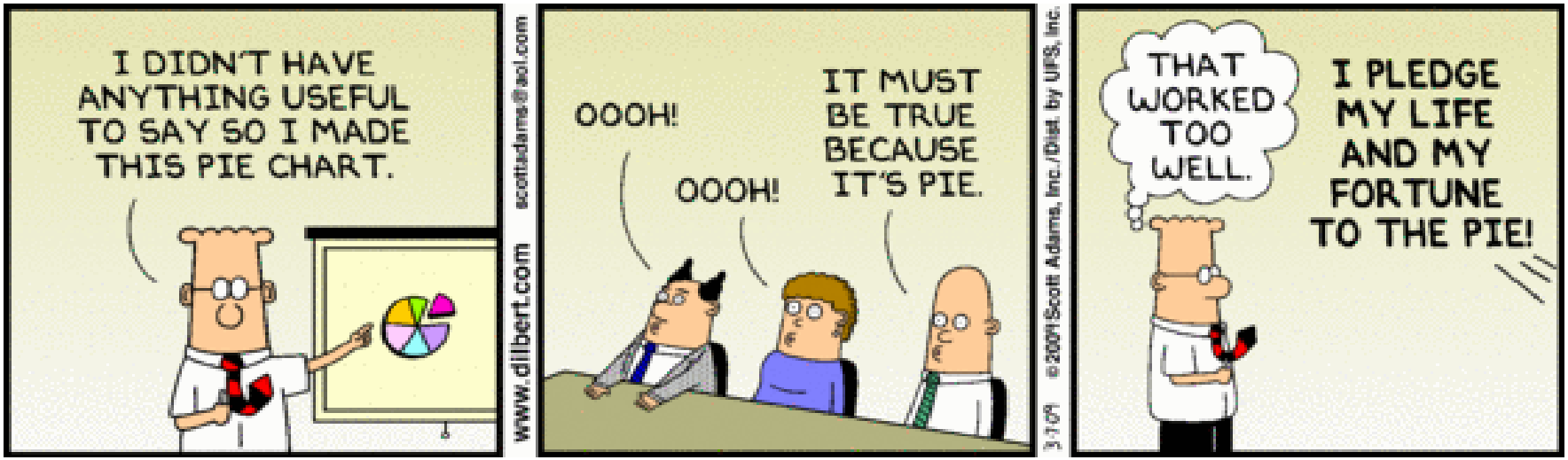
Data Visualization is the communication of information using graphical representations

"A picture is worth a thousand words."

"The greatest value of a picture is when it forces us to notice what we never expected to see." – John Tukey

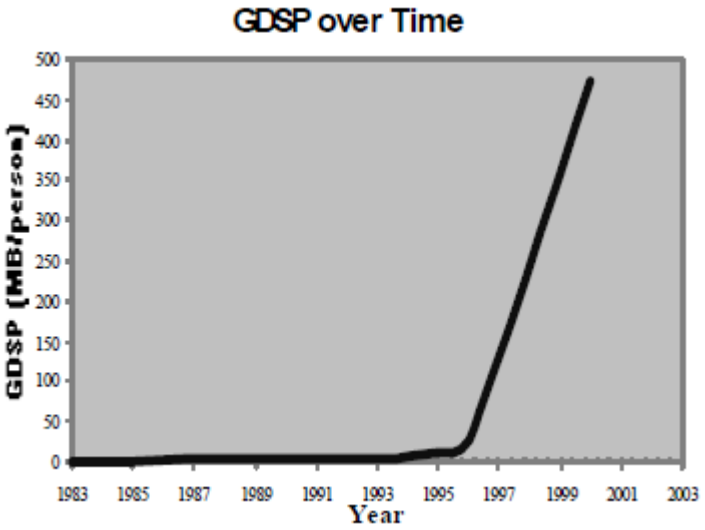
- Data visualization is an alternative and a supplement to the more traditional querying methods
- Use is becoming more prevalent in presentation
- Not yet widely used by actuaries in QC process – more used to seeing numbers

Foundations of data visualization



Evolution in data visualization

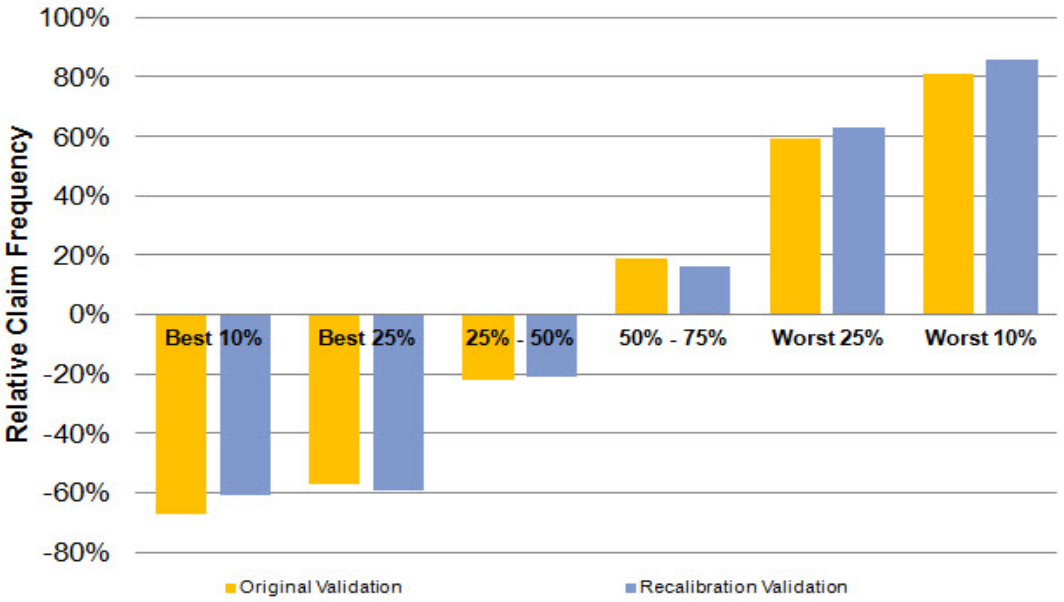
- Amount of data is growing
- Information harder to be expressed numerically in simple spreadsheets
- Improved computing power
- Improved tools and visualization methods, which are easier to use
- Increased demand from clients / management



Benefits of data visualization

- Which one is easiest to convey the message?

	Original Validation	Recalibration Validation
Best 10%	-67%	-61%
Best 25%	-57%	-59%
25% - 50%	-22%	-21%
50% - 75%	19%	16%
Worst 25%	59%	63%
Worst 10%	81%	86%



Benefits of data visualization in QC

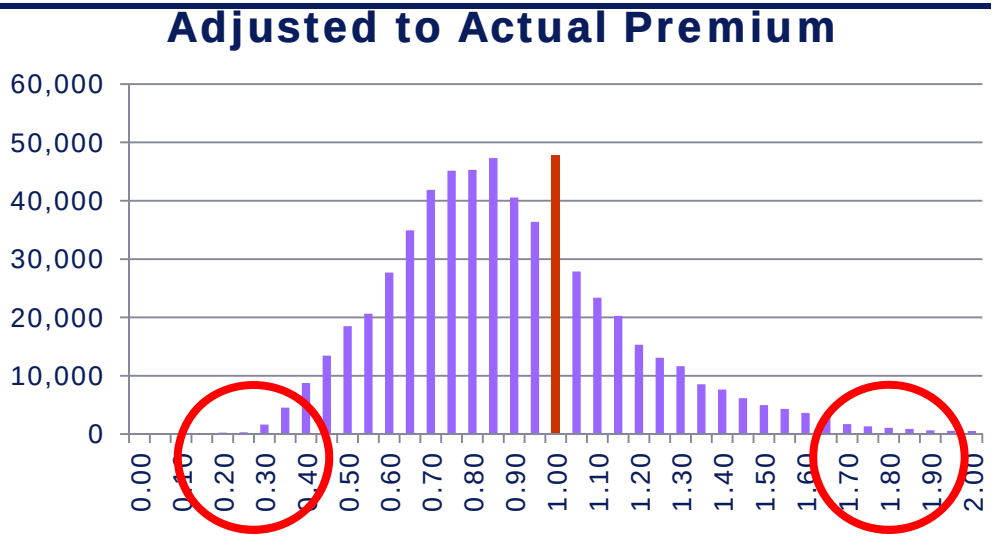
- Data visualization allows users see several different perspectives of the data.
- Data visualization makes it possible to interpret vast amounts of data
- Data visualization offers the ability to note exceptions in the data
- Data visualization allows the user to analyze visual patterns in the data
- Data visualization equips users with the ability to see influences that would otherwise be difficult to find
- Packages allow drag and drop functionality, limited need for coding
- Also allows for easy drill down, filtering, grouping
- Data visualization allows a natural progression for QC to insight in modeling

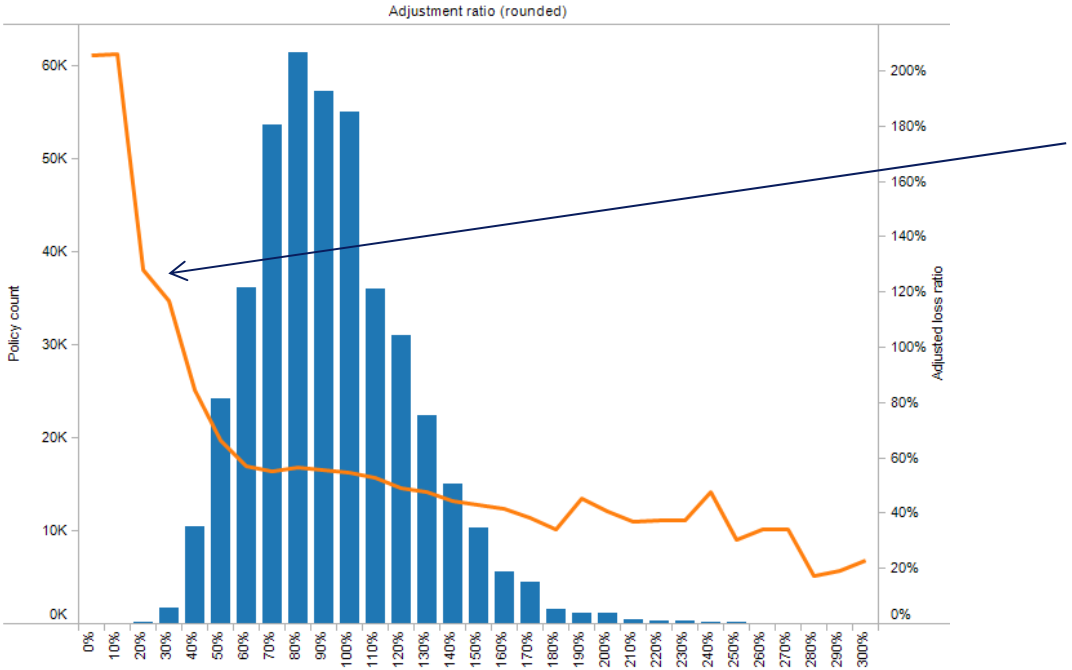
Uses of data visualization

- **QC, Data Mining, Exploratory Data Analysis (EDA)**
- Model interpretation
- Results / Presentation

Examples - Histogram

- Mass points
- Extreme values
- Unreasonable values



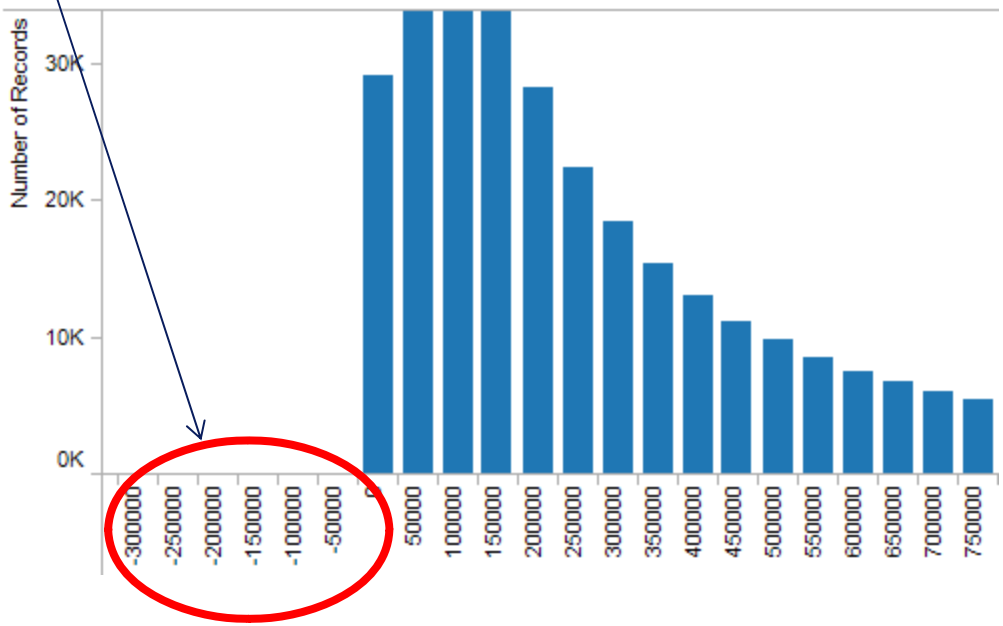


Histogram after correction – but is it fixed?

- Still some outliers
- For extremes, higher loss ratio indicates potential problems with adjusted premium

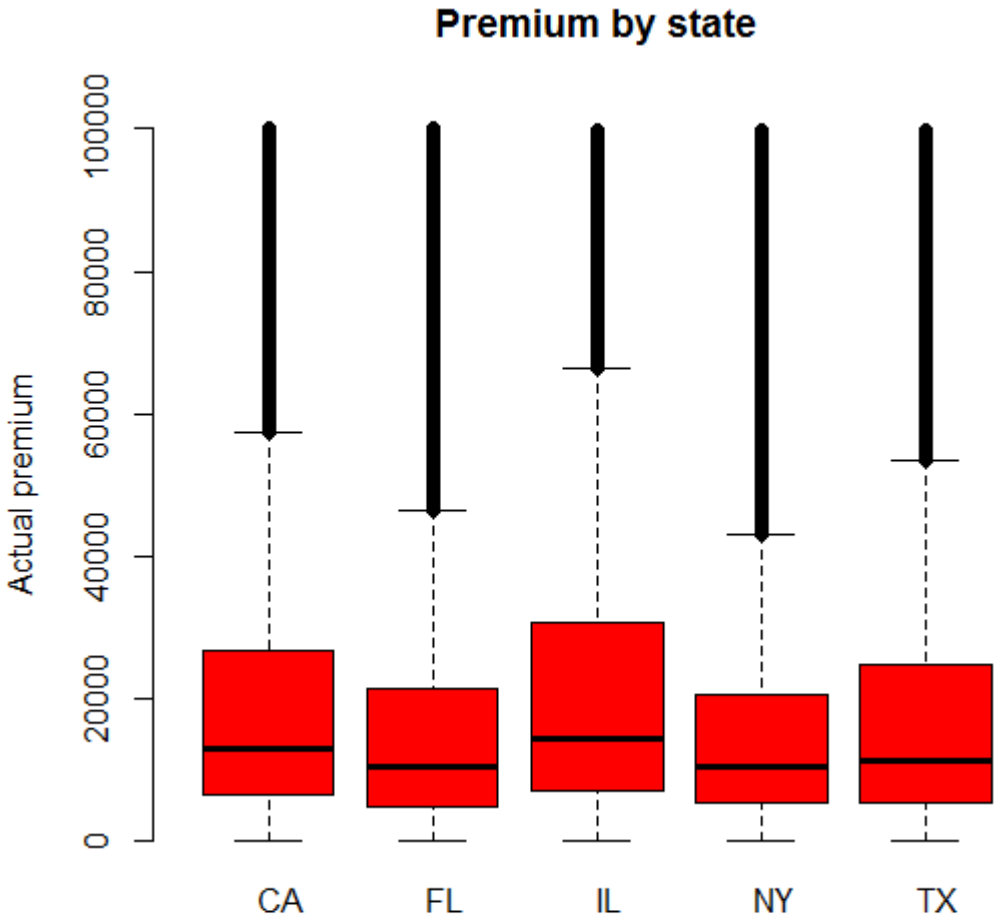
Examples - Histogram

- Incorrectly calculated exposure is one of the causes



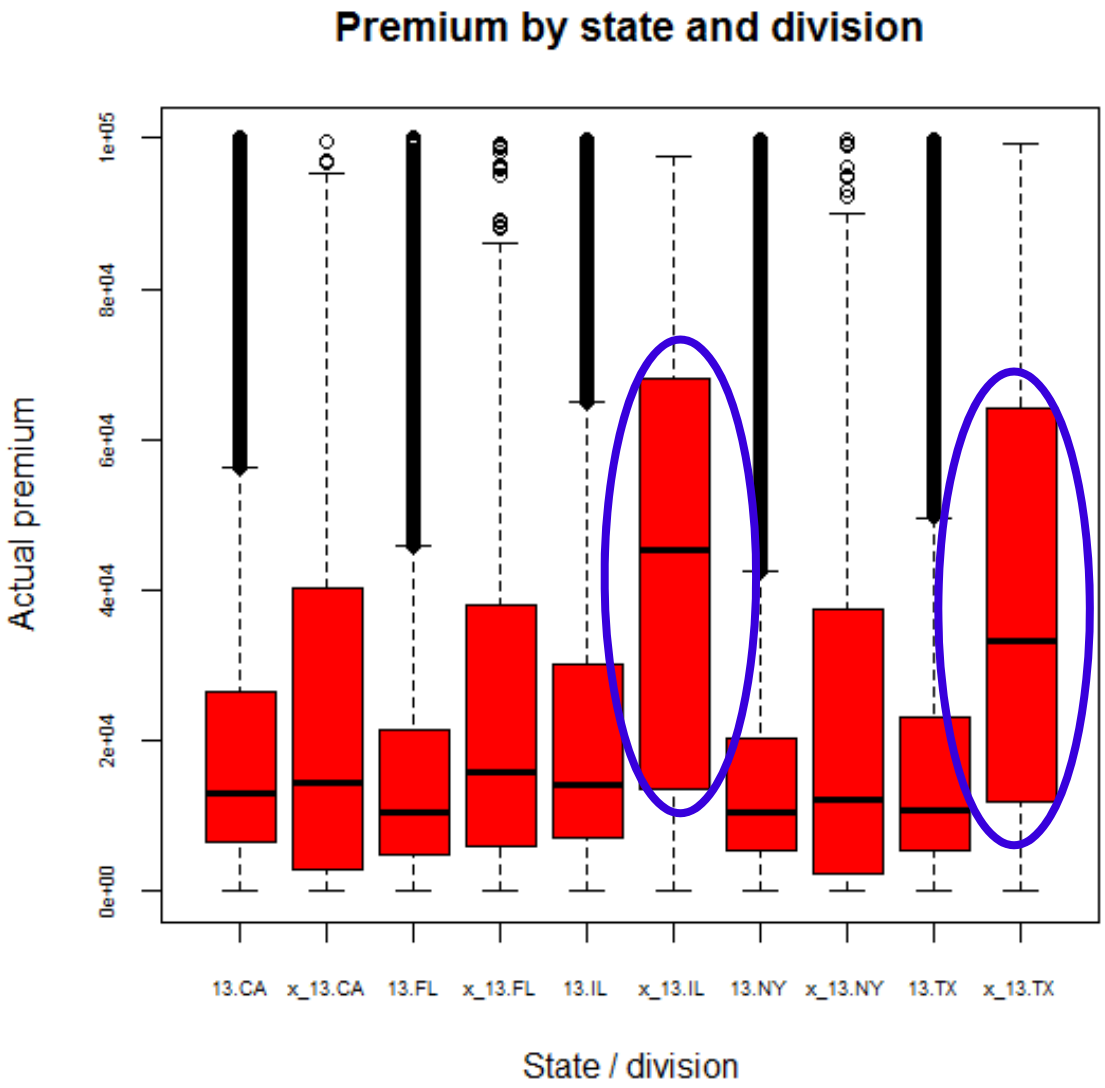
Examples - Boxplots

- Descriptive of median, standard deviation, outliers
- Particularly useful in comparing groups within a variable



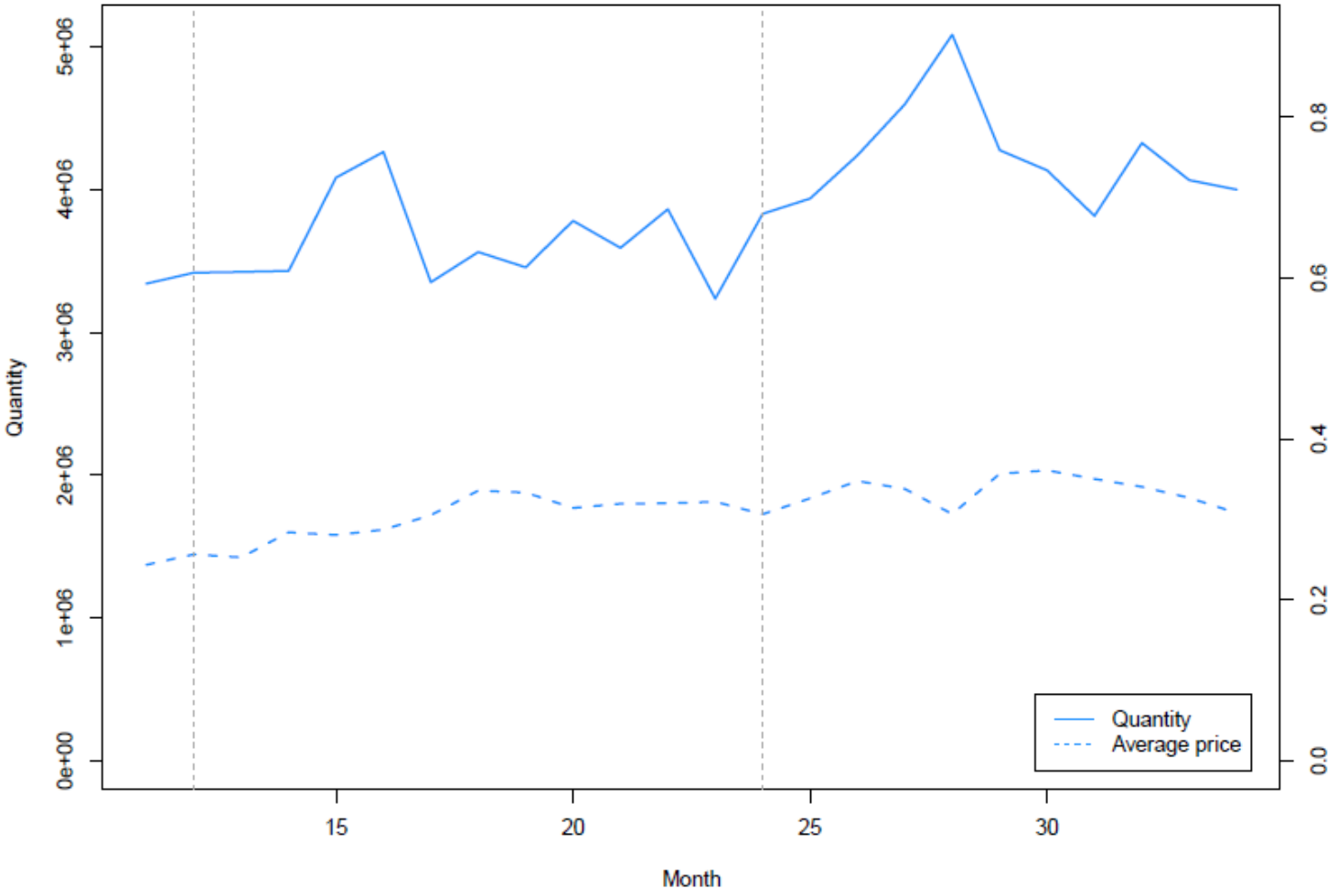
Examples - Boxplots

- Split into subgroups as well



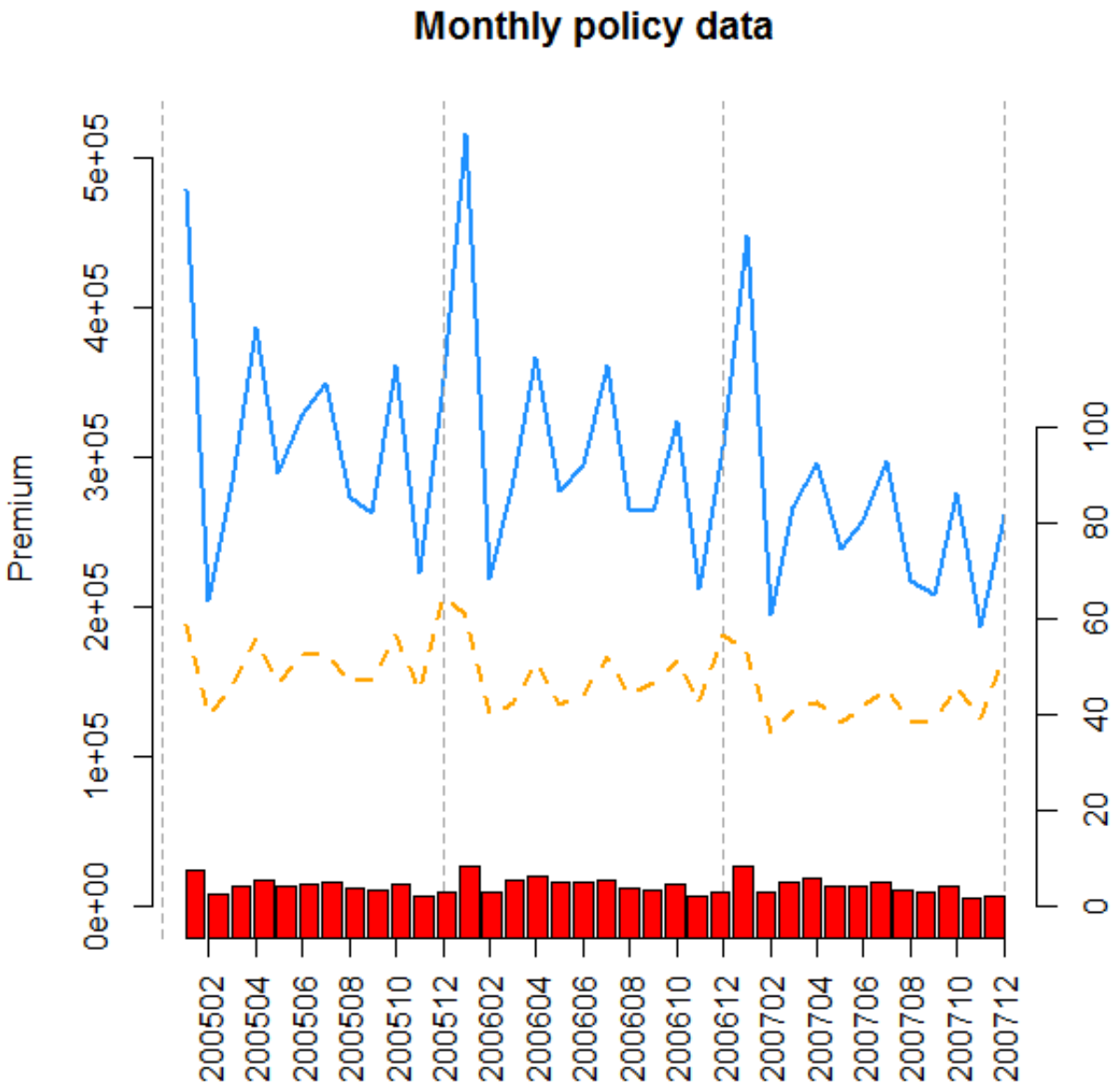
Examples – Line Charts

- Unusual spikes by month?
- Seasonal activity consistent with business knowledge?
- Consistent behavior over time?

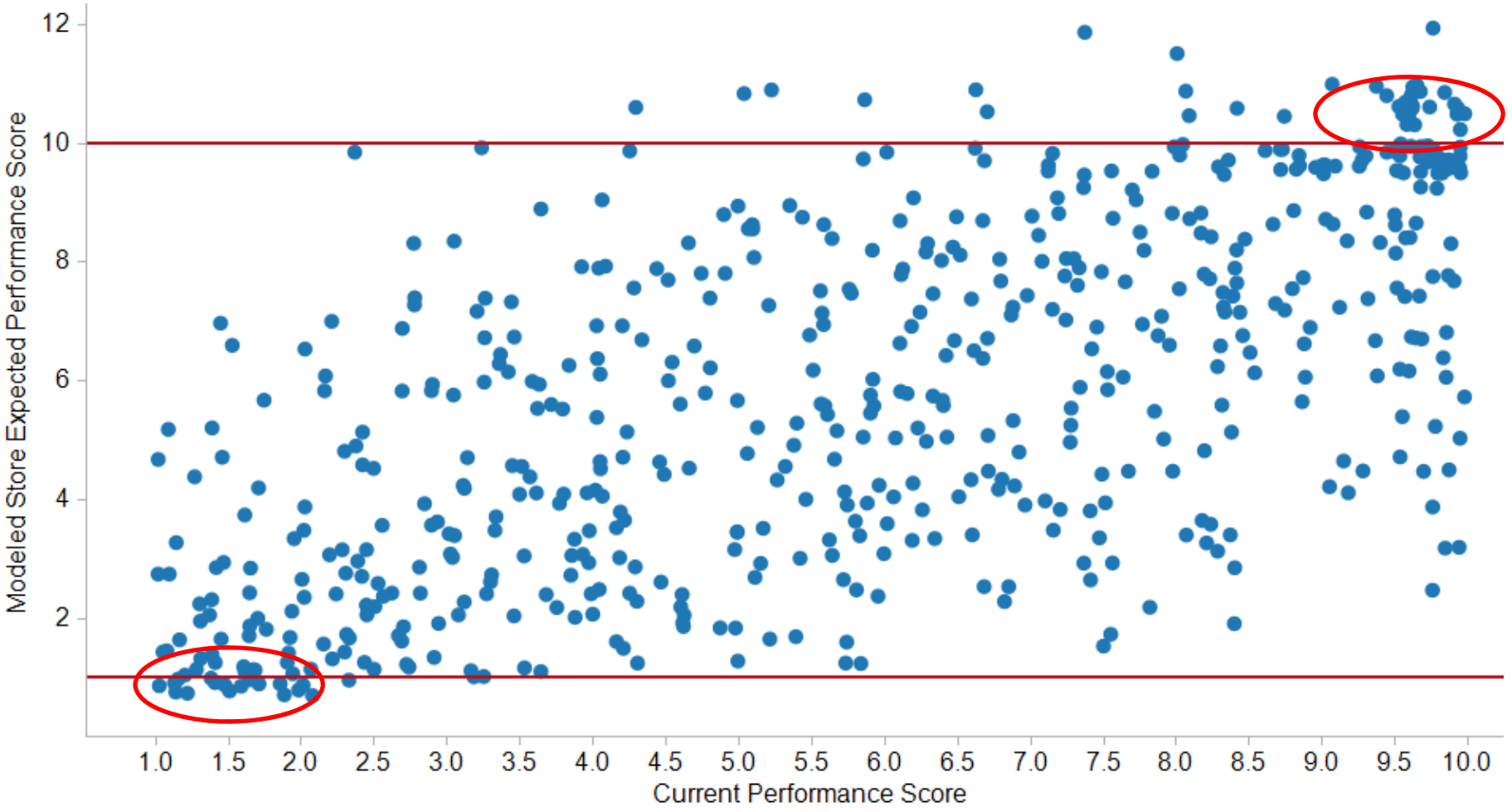


Examples – Line Charts

- Back to insurance: Total premium average premium, policy counts by month
- Is peak at January renewal (both counts and average premium) consistent with business?

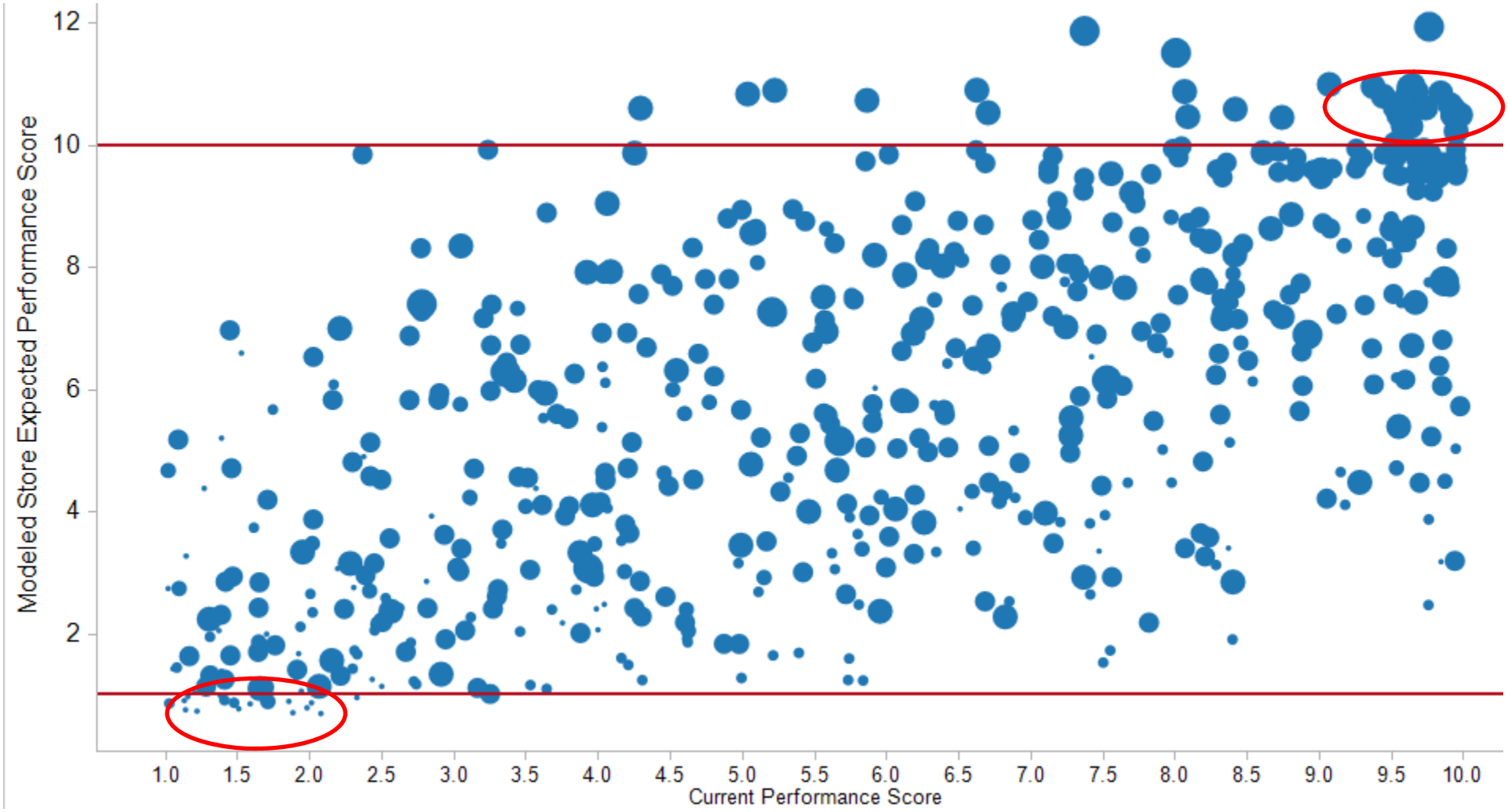


Examples – Scatterplots



Modeled points outside reasonable bounds (1-10)

Examples – Scatterplots



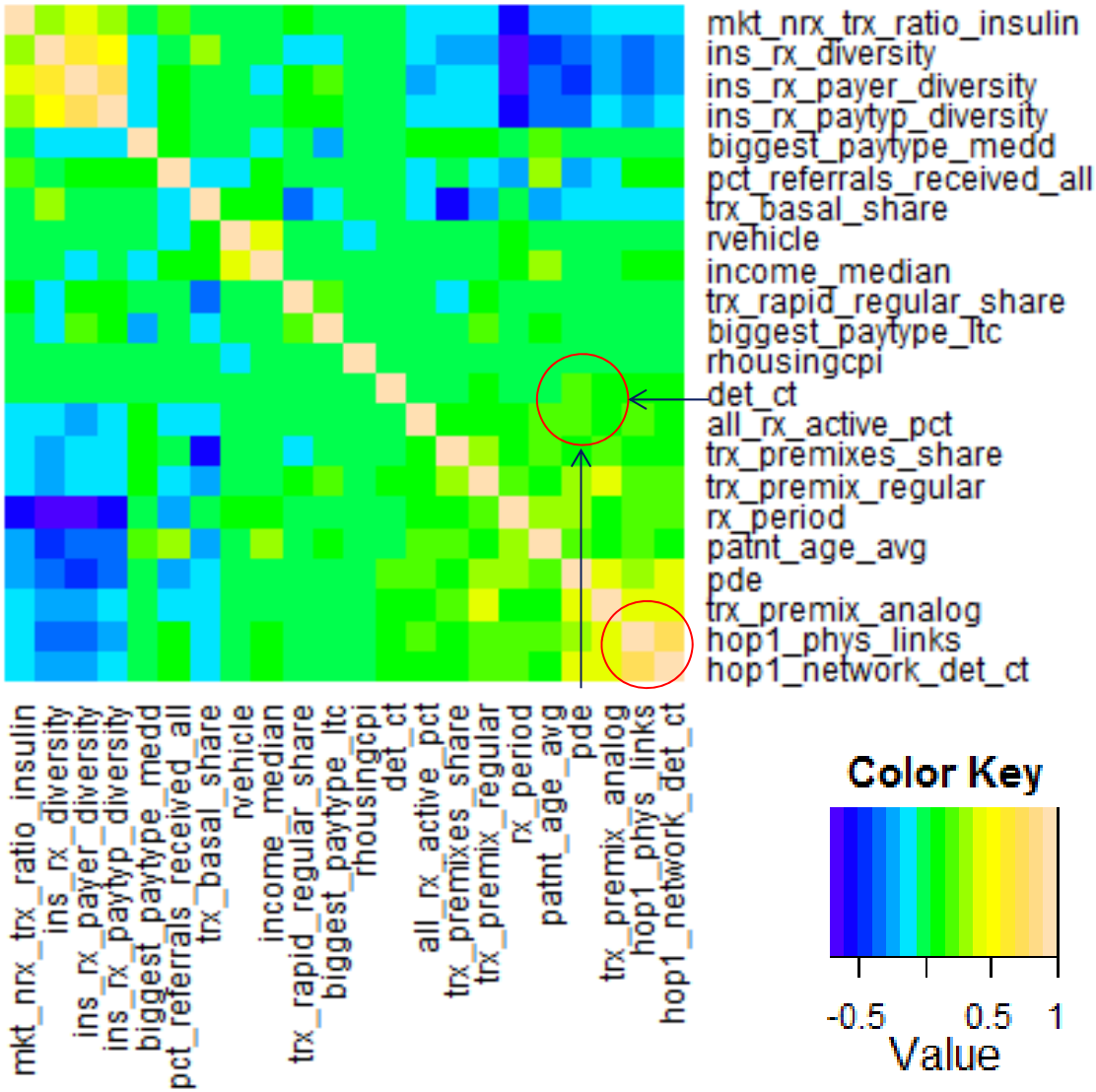
Adding dimension of business age (size of circles) shows that age incorrectly included in model score

Examples - Heatmaps

Correlations between variables

Look for:

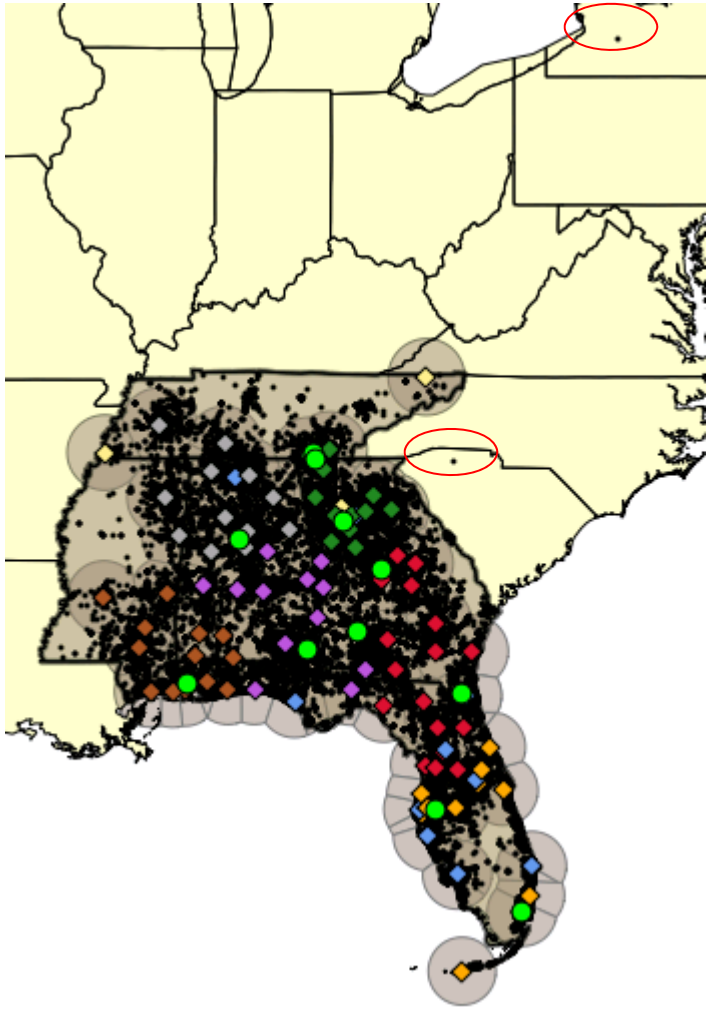
- Very high correlations (near +1 or -1
- Low correlations for variables that should be related



Examples - Geospatial

Customers in a five state area compared to sales locations

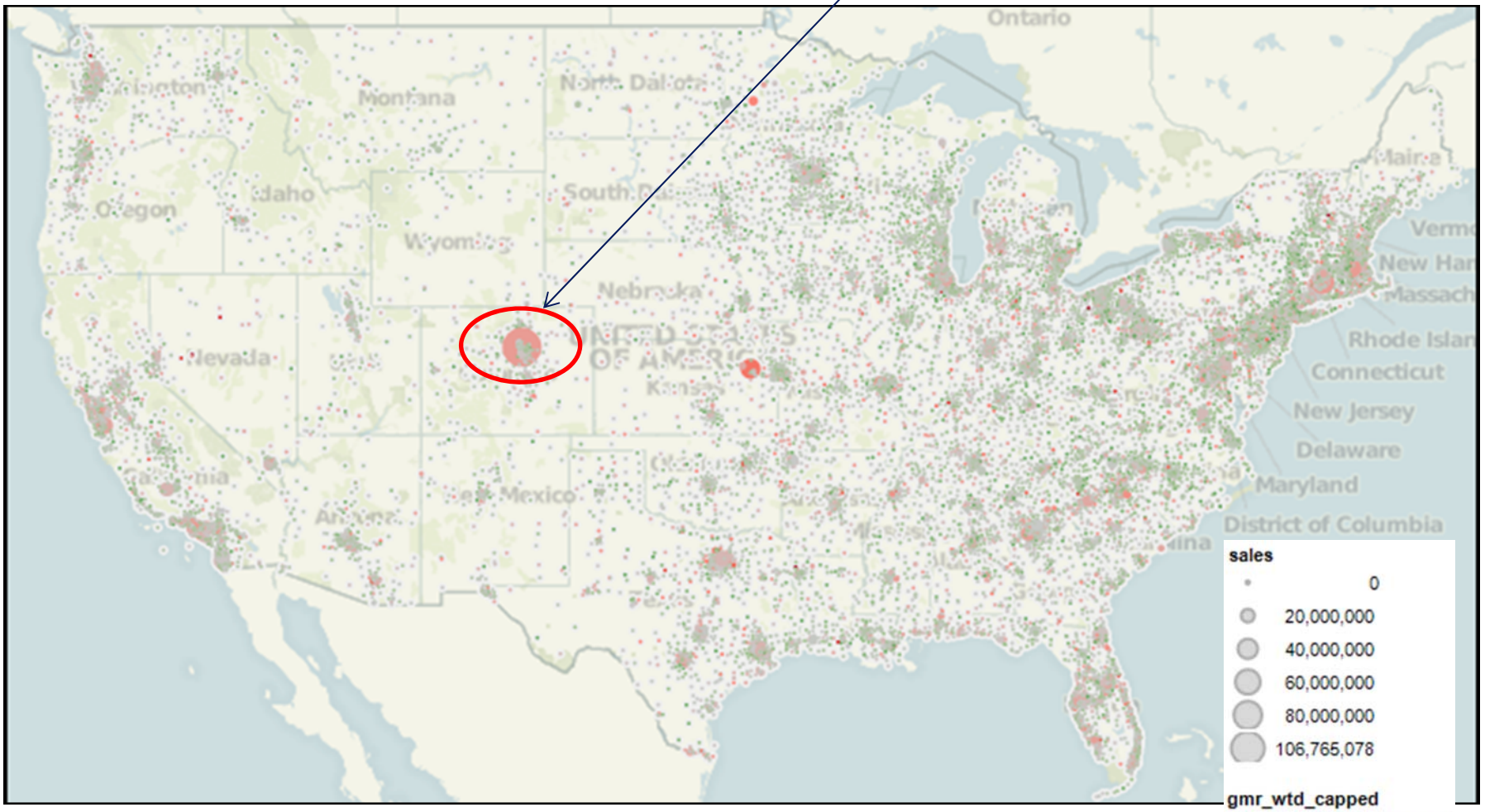
- Customers in area that are not within driving distance
- Customers mapped well outside area



Examples - Geospatial

Sales and gross margin rate by zip code

Large mass point in sales, also large negative GMR



Software (not an exhaustive list!)

Excel	Third party add-ins available
PowerPivot	Microsoft add-in to Excel
R	Open source, multitudes of packages
Tableau	
SAS	SAS/GRAPH module
Spotfire	
Microstrategy	
Geospatial mapping	ESRI, Alteryx, Google API, etc.

Wrapping Up

- Data Visualization allows analysts to communicate with clarity, precision, and efficiency
- However, there are benefits in utilizing visualizations early in the process, as they may uncover data issues more easily than individual queries
- Standard tools, like histograms, scatterplots, maps, are becoming more available and easier to use