OliverWyman

IABA
INTERNATIONAL ASSOCIATION OF BLACK ACTUARIES

DATA VISUALIZATION IN BUSINESS COMMUNICATION

Theory, Methods, and Tools

International Association of Black Actuaries
2023 IABA Annual Meeting
Chicago, IL

SPEAKERS

Dalesa Bady, ACAS, MAAA
Actuary, GuideOne Insurance

Associate of the Casualty Actuarial Society (CAS)

12 years of experience in the property and casualty insurance industry.

Fav areas of work: ratemaking, product development, and predictive analytics.

Passionate about coaching and development.

Co-Chair of IABA Student Programs, CAS Leadership Development Committee.

Likes to travel and watch, active in sports.

Bryce Chamberlain, ASA, MSCA
Principal, Oliver Wyman Actuarial Consulting

Associate of the Society of Actuaries (SOA)

Master of Science in Analytics, University of Chicago

Leads a team at building business intelligence apps for the web using R Shiny.

Passionate about data visualization, user-friendly design, and efficiency.

Lead developer for R packages: easyr, hcslim.

Likes to play video games and meet friends at cocktail bars.

AGENDA

The last slide will show a link to this deck + all the resources mentioned.

1

Why Visualize?

2

How We Think Visually

3

Things To Avoid

4

Problems & Tools to Solve Them

WHY VISUALIZE?

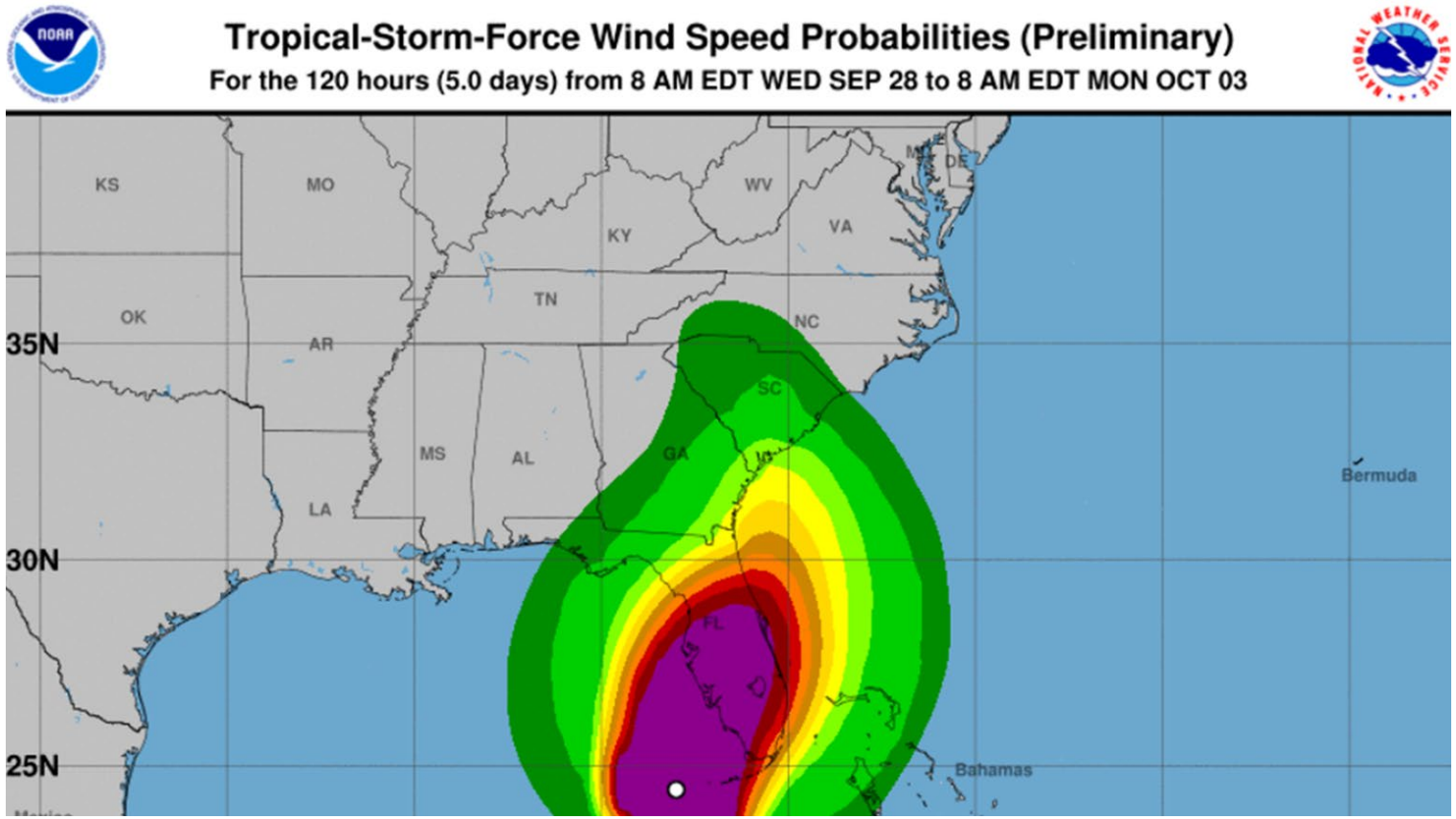
“

How many of you pay attention to road signs while driving?

How much harder would it be if road signs were a full body of text?

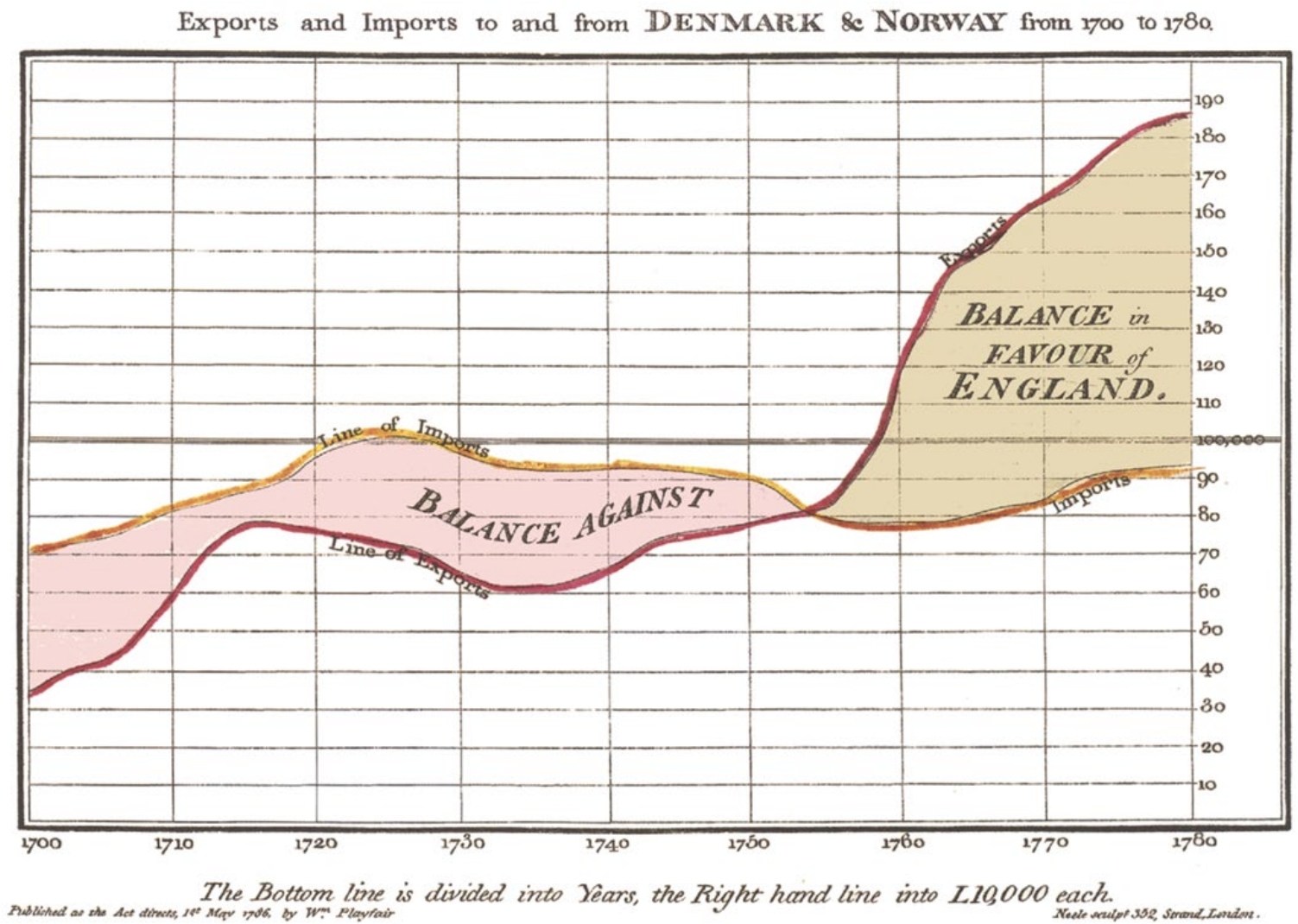
Beware of the road ahead. It might have rained or snowed and it could be slippery, which might cause an incident.
Be careful.

https://piktochart.com/blog/why-people-love-great-visuals-science



Would this message be as powerful with a table of numbers?

"The Commercial and Political Atlas"
- William Playfair (1786)



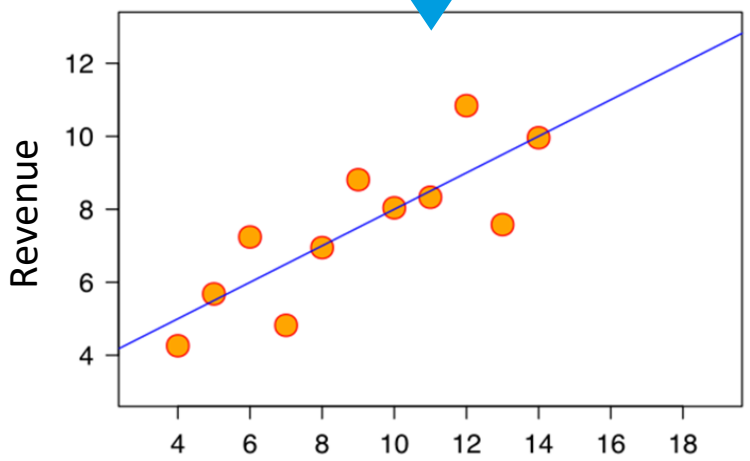
“

Well, just show me the numbers.

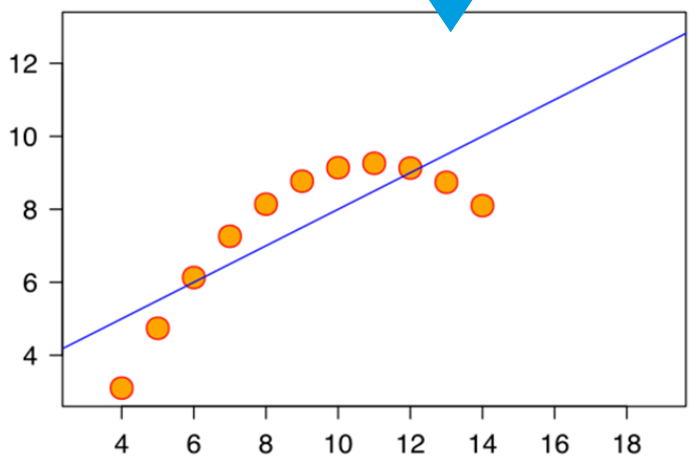
Let’s take a look at an example.

Agent	Mean Claims ($millions)	Mean Revenue ($millions)	Correlation
A	9	7.5	0.816
B	9	7.5	0.816
C	9	7.5	0.816
D	9	7.5	0.816

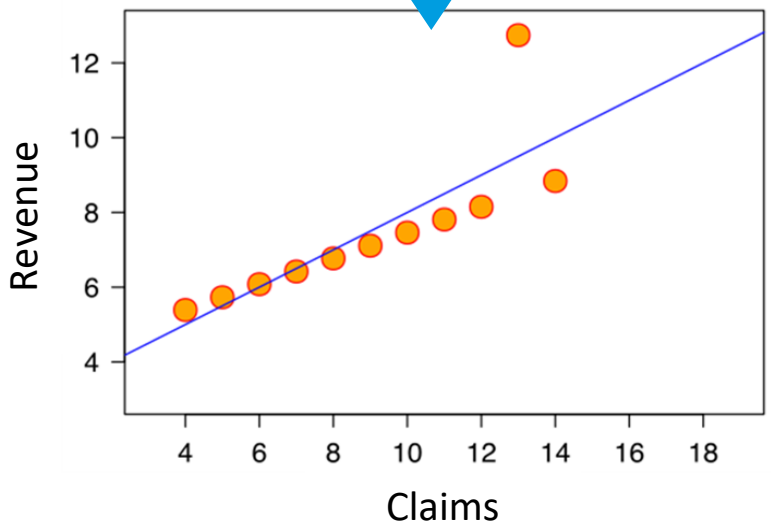
Revenue increases with claims.



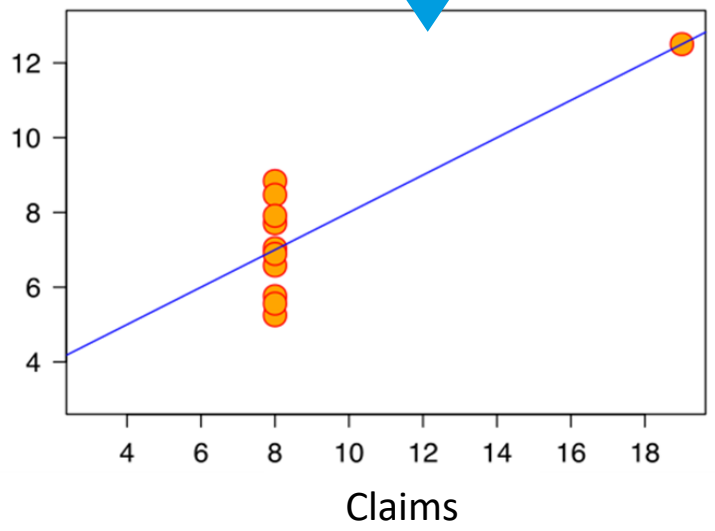
Revenue capped.



Revenue outlier.



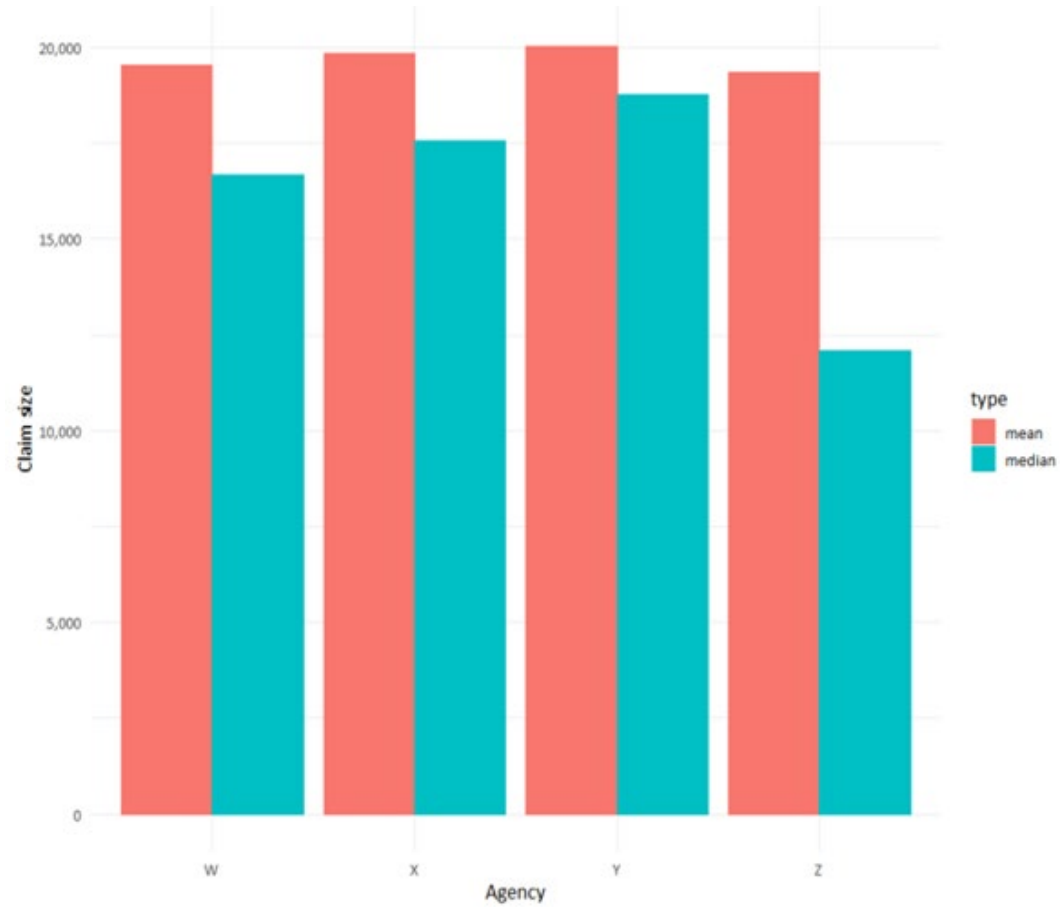
Revenue and amount outlier.



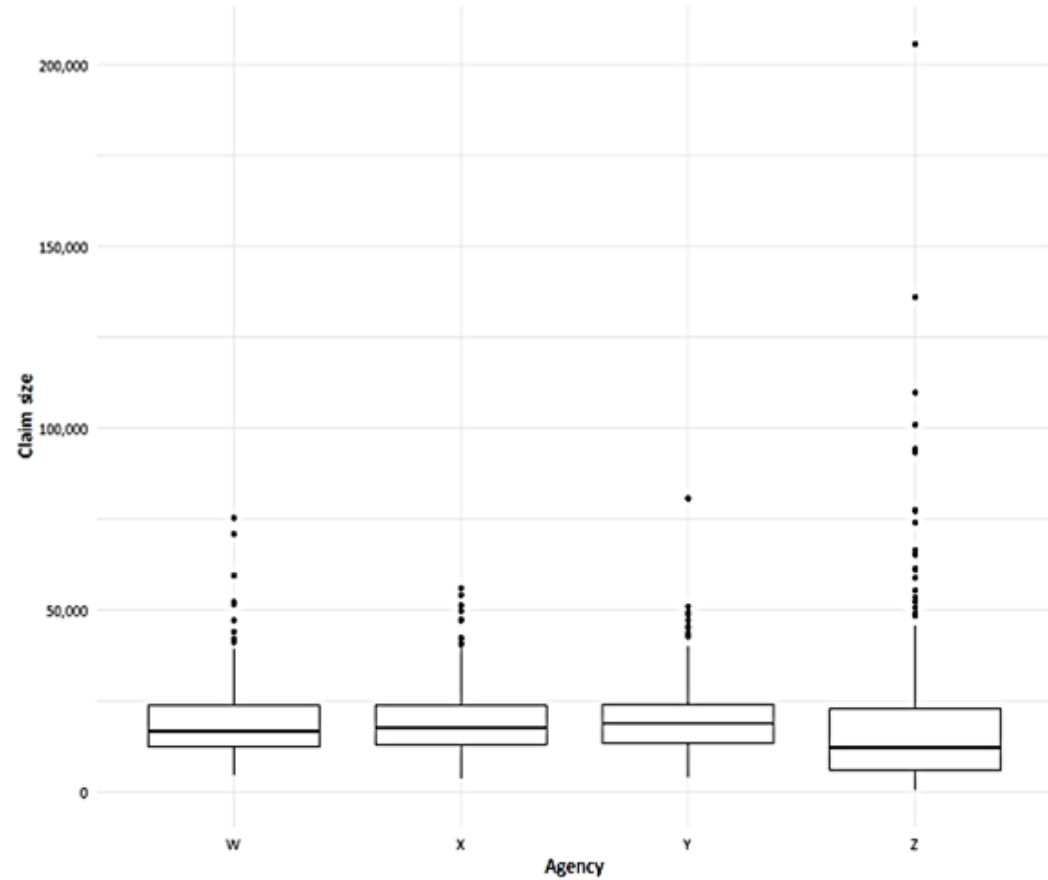
“

Visualization is useful when communicating to a non-technical person or audience.

Agency	Mean Claim Size	Median Claim Size
W	19,533	16,675
X	19,829	17,577
Y	20,039	18,758
Z	19,363	12,097



Agency	Mean Claim Size	Median Claim Size
W	19,533	16,675
X	19,829	17,577
Y	20,039	18,758
Z	19,363	12,097



HOW DO WE THINK VISUALLY?

999999999

99999999999

999,999,999

99,999,999,999

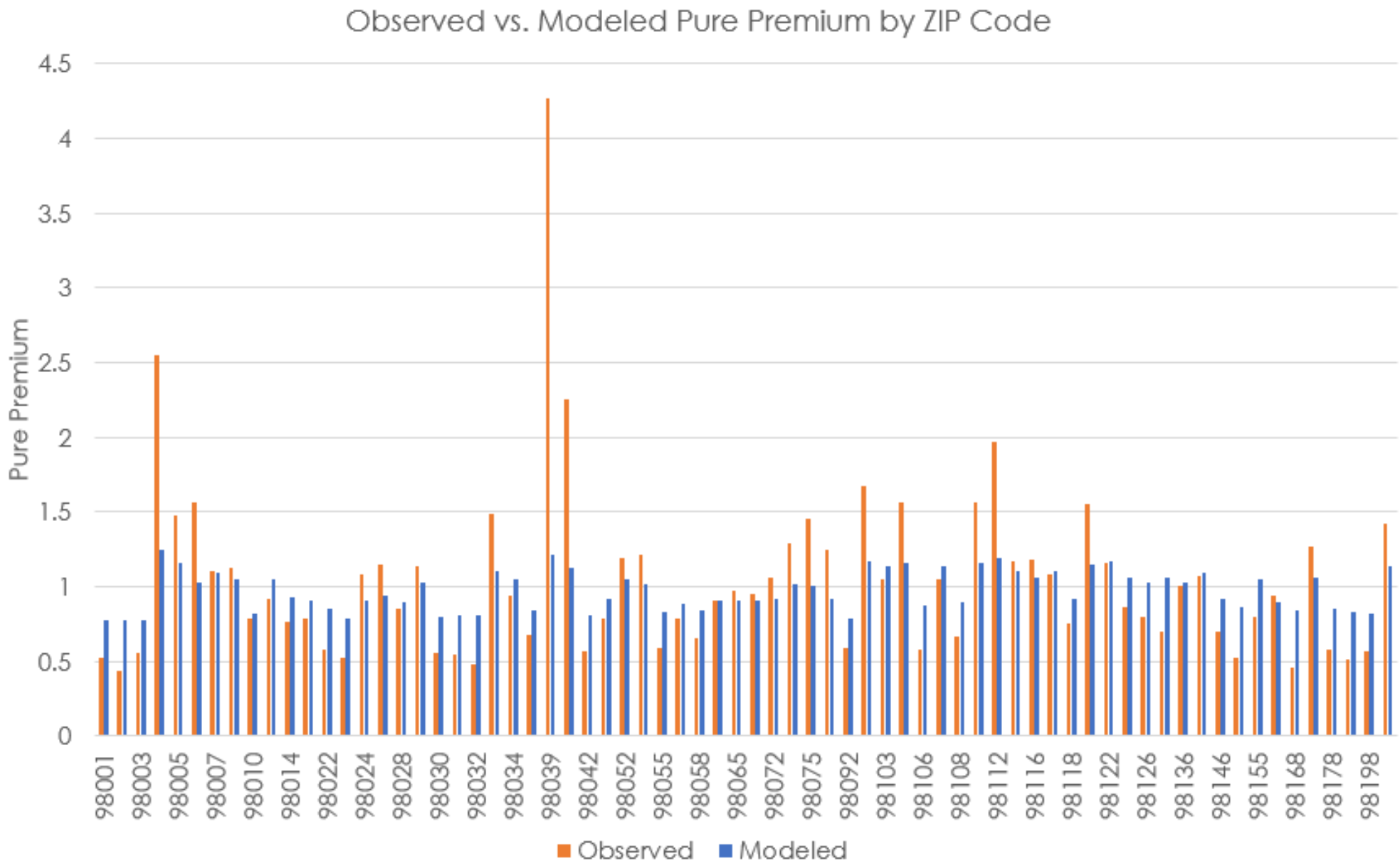
999,999,999

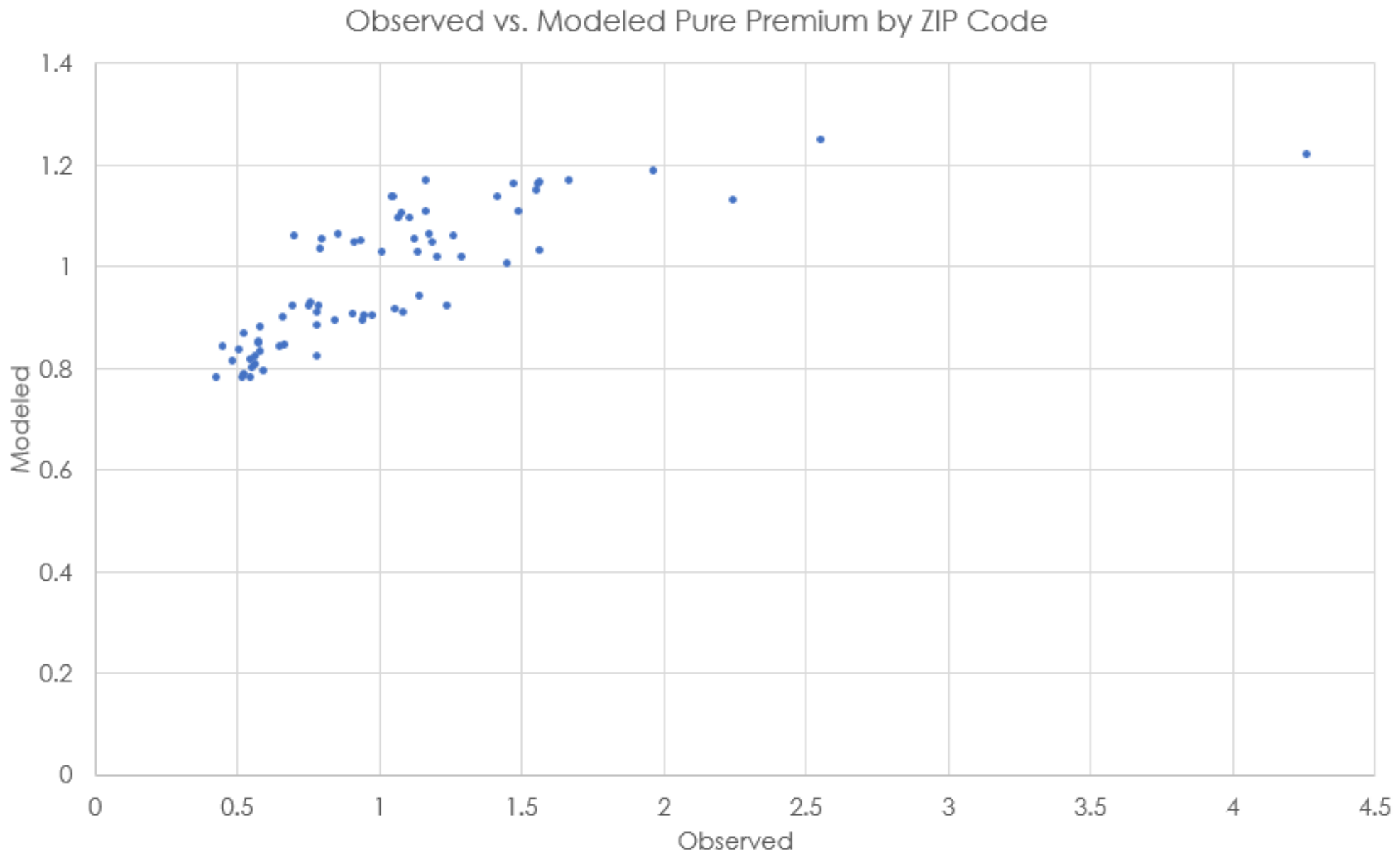
99,999,999,999

Mapping the values in a visual space helps us understand the enormity of the difference!

“

Let’s take a look at an example of an insurance dataset visualized in two ways.

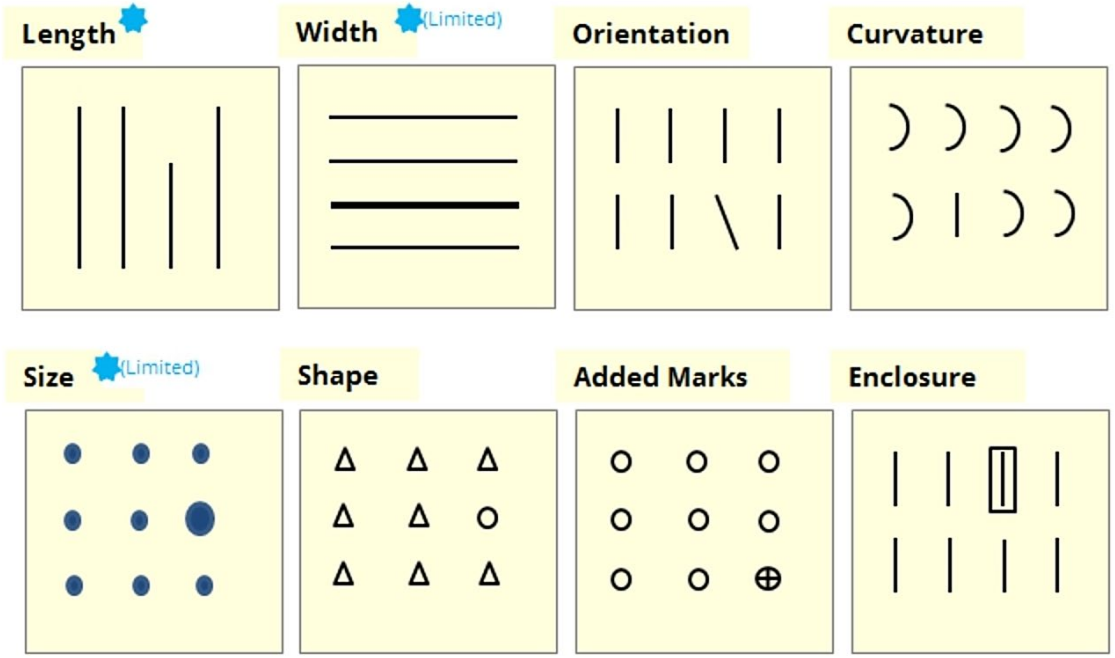

Observed vs. Modeled Pure Premium by ZIP Code
Pure Premium
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
98001
98003
98005
98007
98010
98014
98022
98024
98028
98030
98032
98034
98039
98042
98052
98055
98058
98065
98072
98075
98092
98103
98106
98108
98112
98116
98118
98122
98126
98136
98146
98155
98168
98178
98198
Observed
Modeled

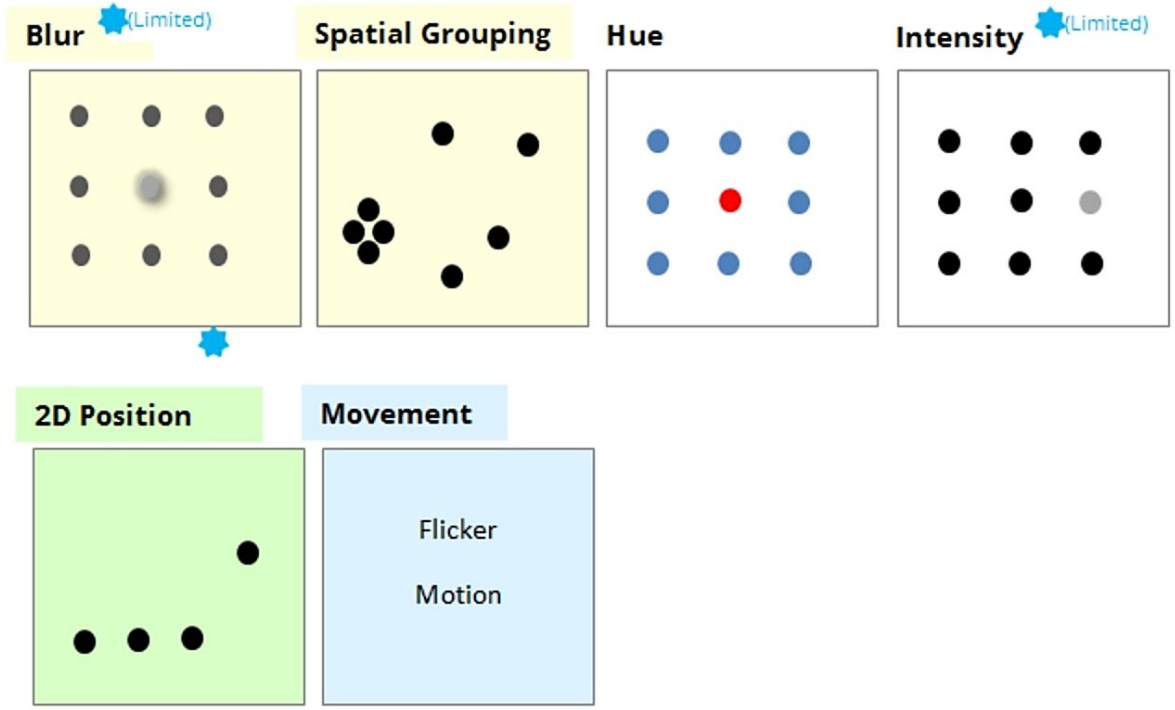

Observed vs. Modeled Pure Premium by ZIP Code
Modeled
Observed
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

PRE-ATTENTIVE ATTRIBUTES

TOOLS FOR COMMUNICATING VISUALLY

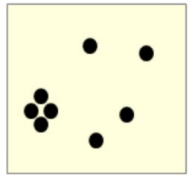
Form | Color | Spatial Position | Movement | Quantitatively perceived



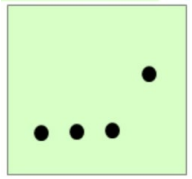


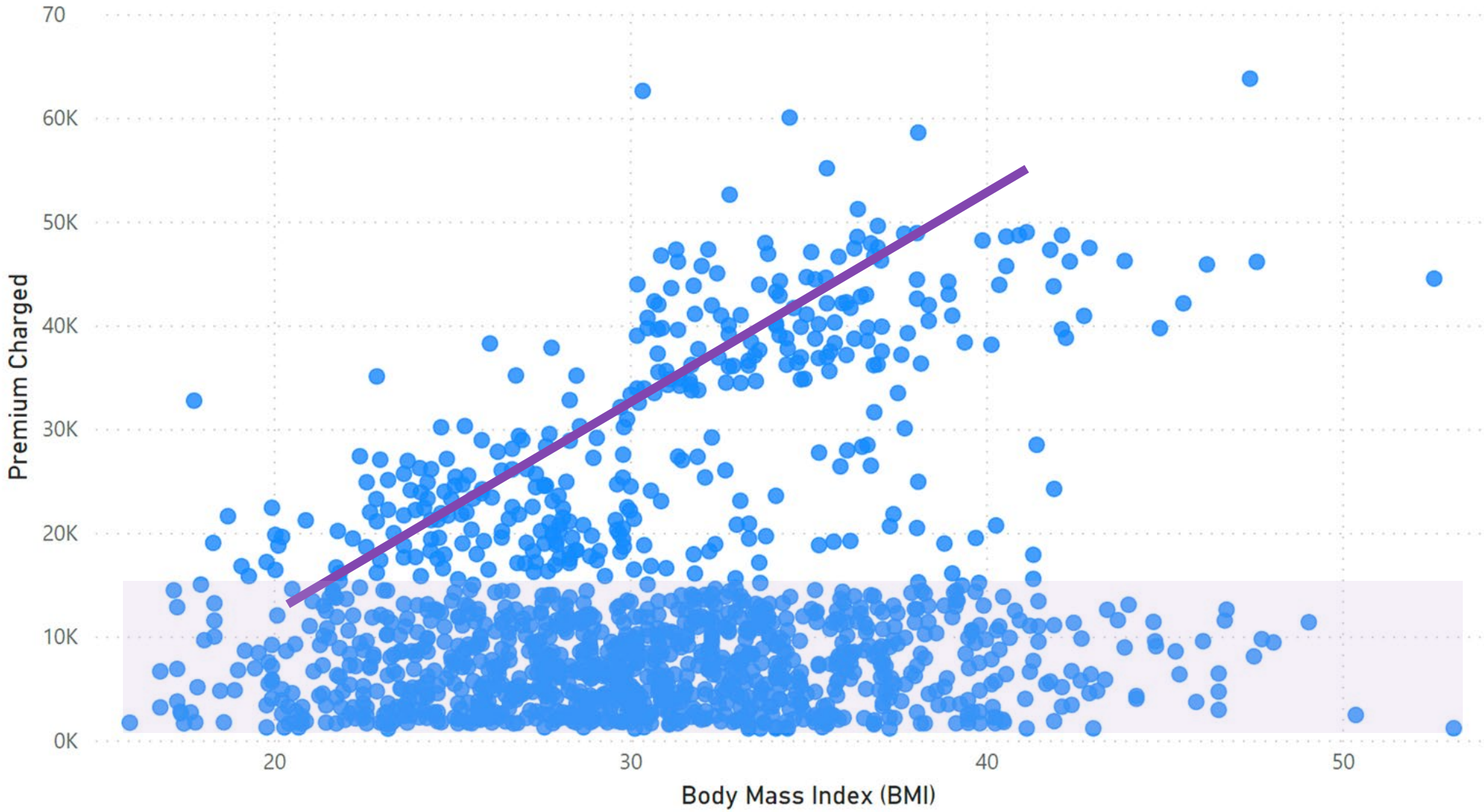
Colin Ware, Information Visualization, Perception for Design, 2012
Images from http://daydreamingnumbers.com/blog/preattentive-attributes-example

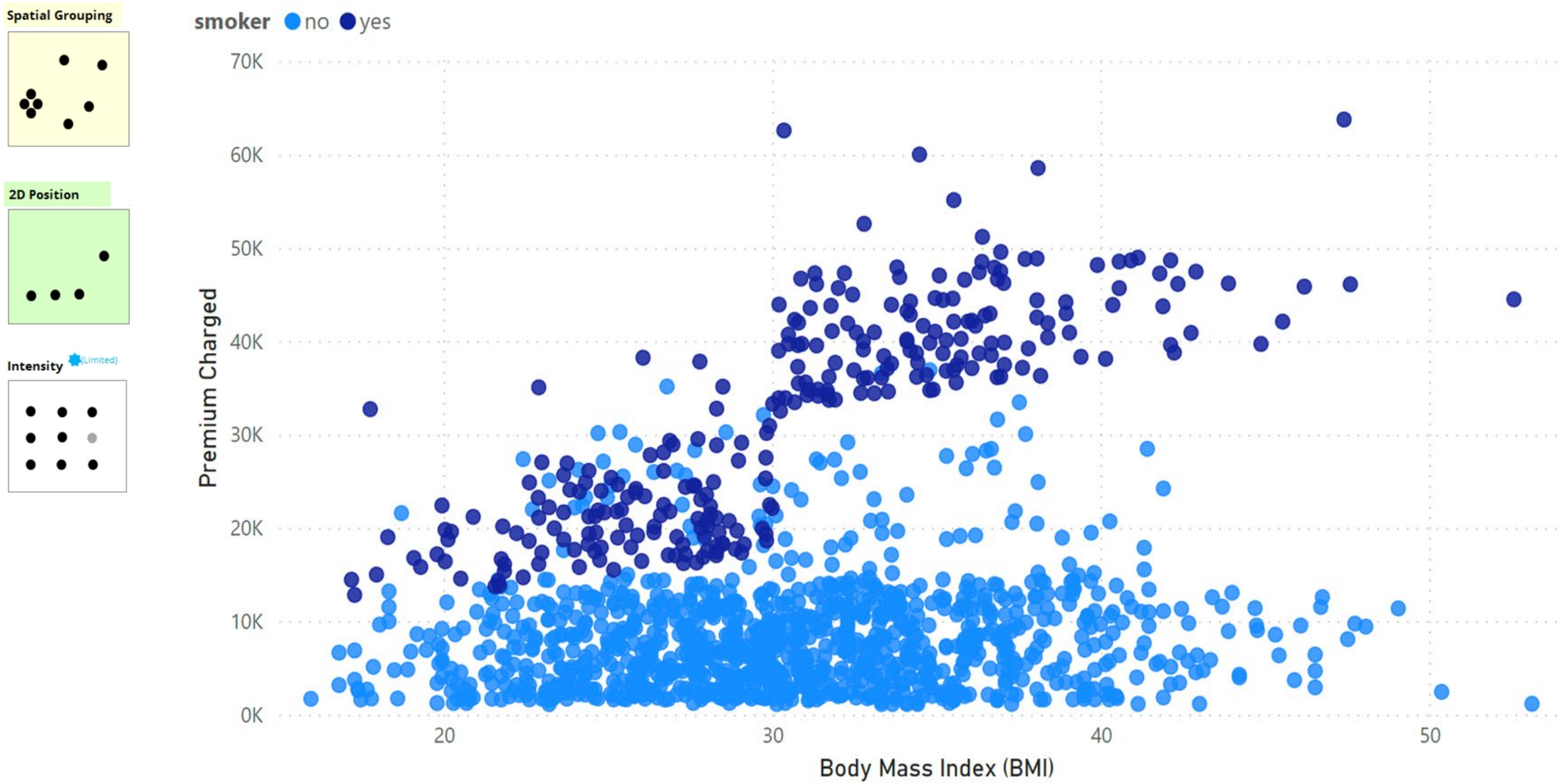
Spatial Grouping

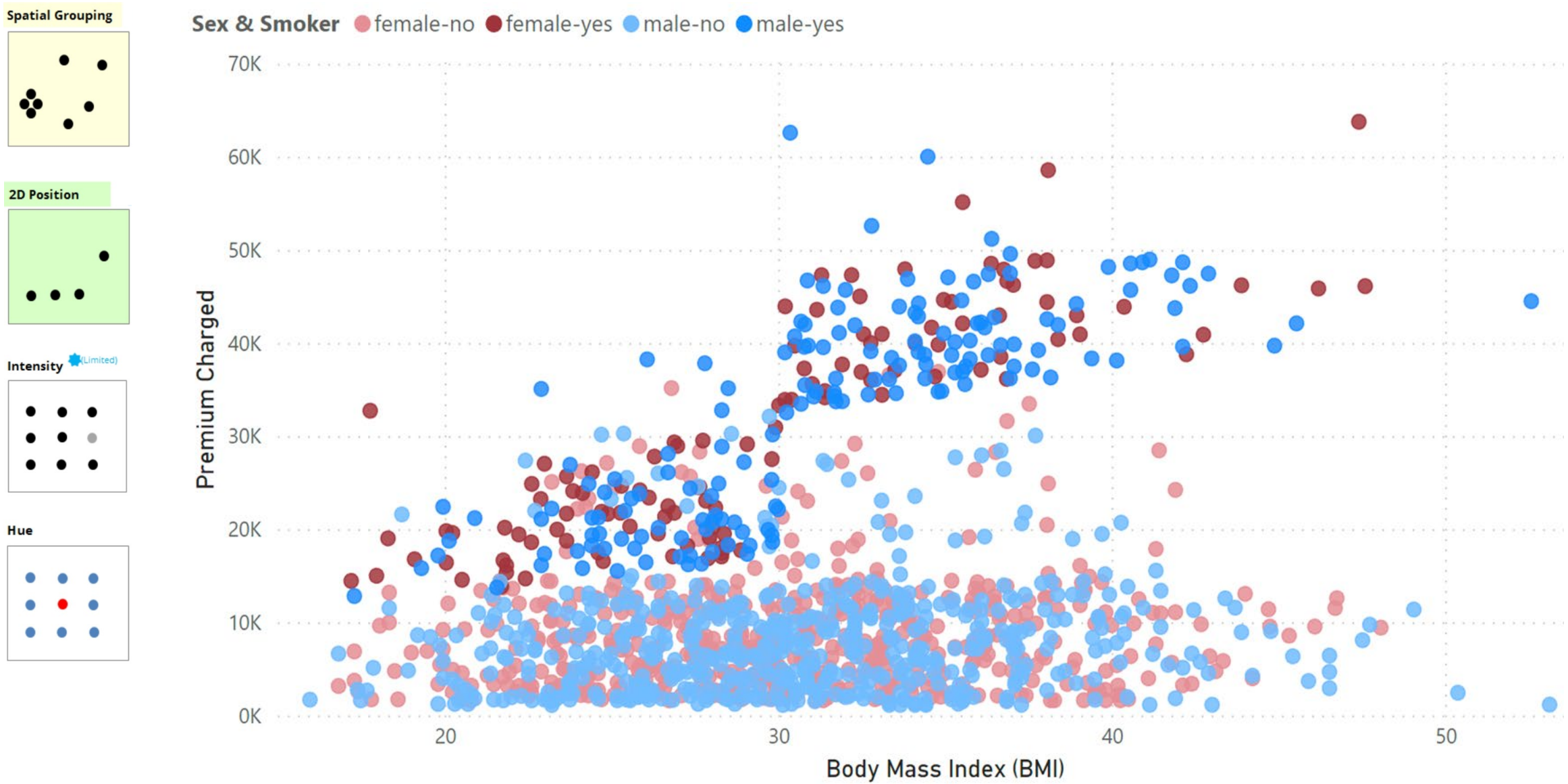


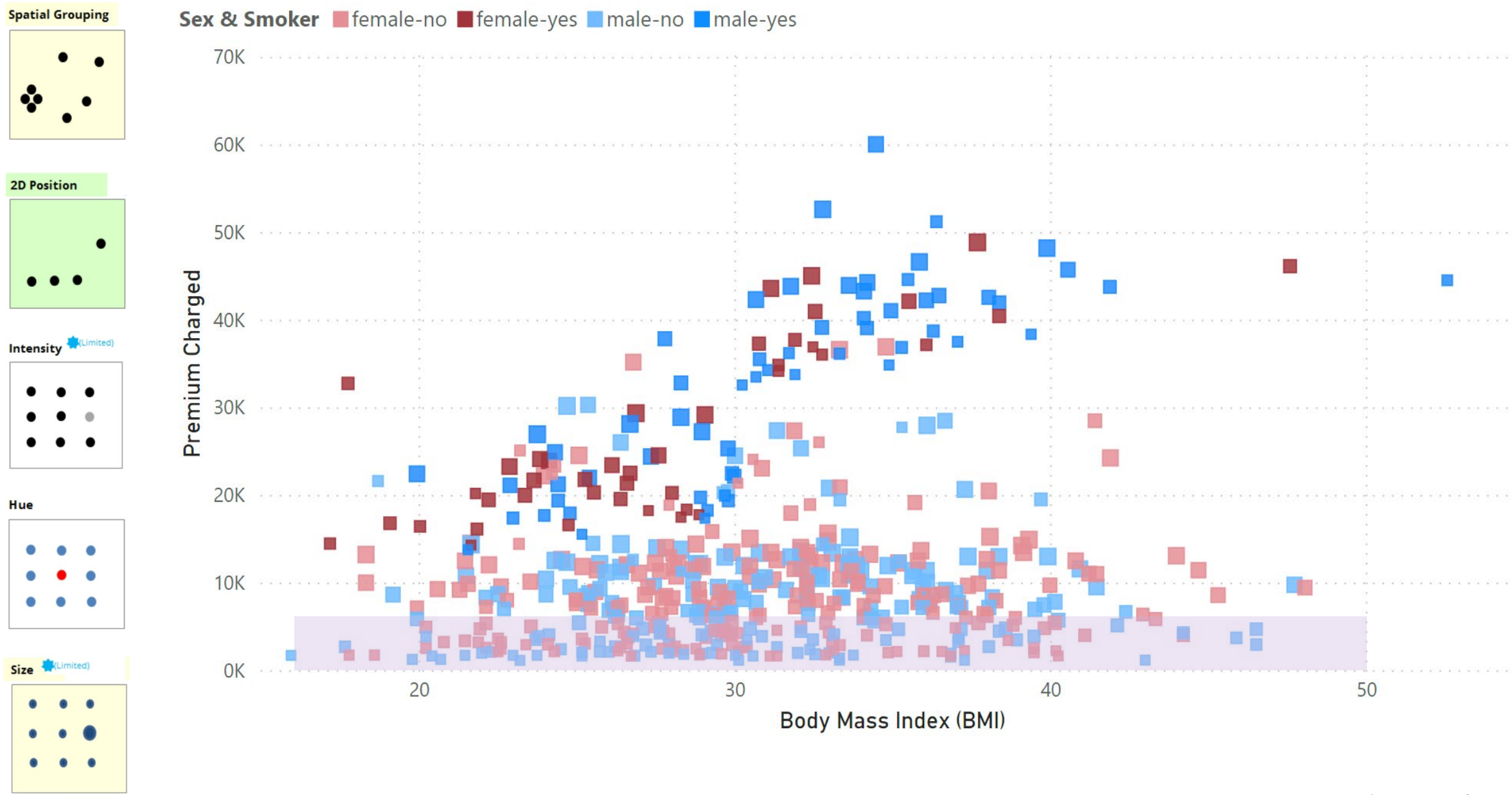
2D Position






Spatial Grouping
2D Position
Intensity
(Limited)
smoker
no
yes
70K
60K
50K
40K
30K
20K
10K
0K
Premium Charged
20
30
40
50
Body Mass Index (BMI)

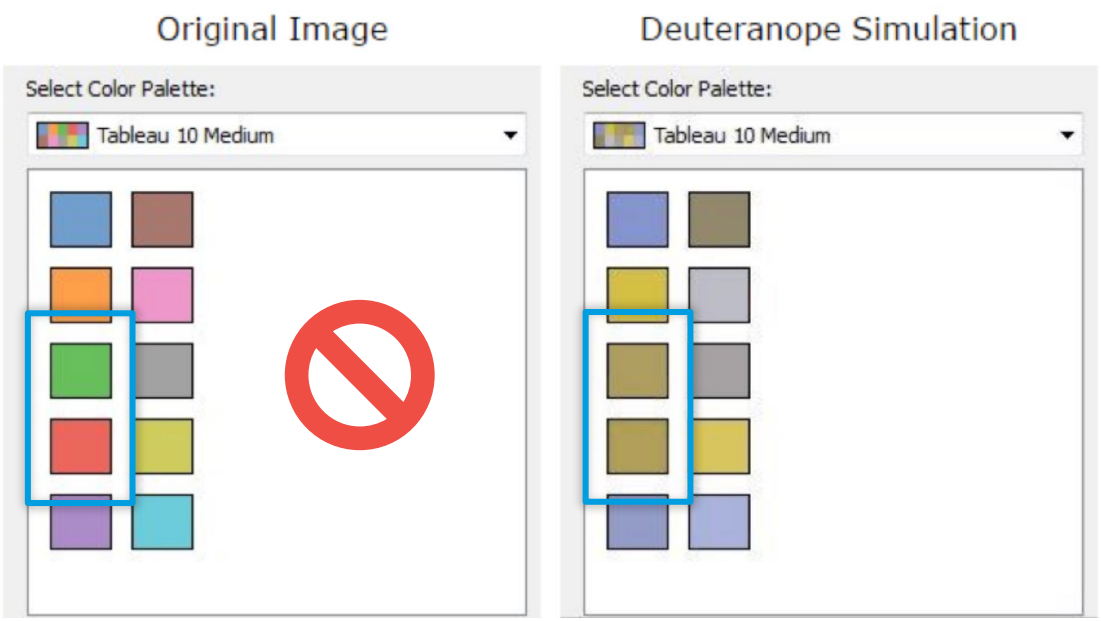

Spatial Grouping
2D Position
Intensity
(Limited)
Hue
Sex & Smoker
female-no
female-yes
male-no
male-yes
70K
60K
50K
40K
30K
20K
10K
0K
Premium Charged
20
30
40
50
Body Mass Index (BMI)



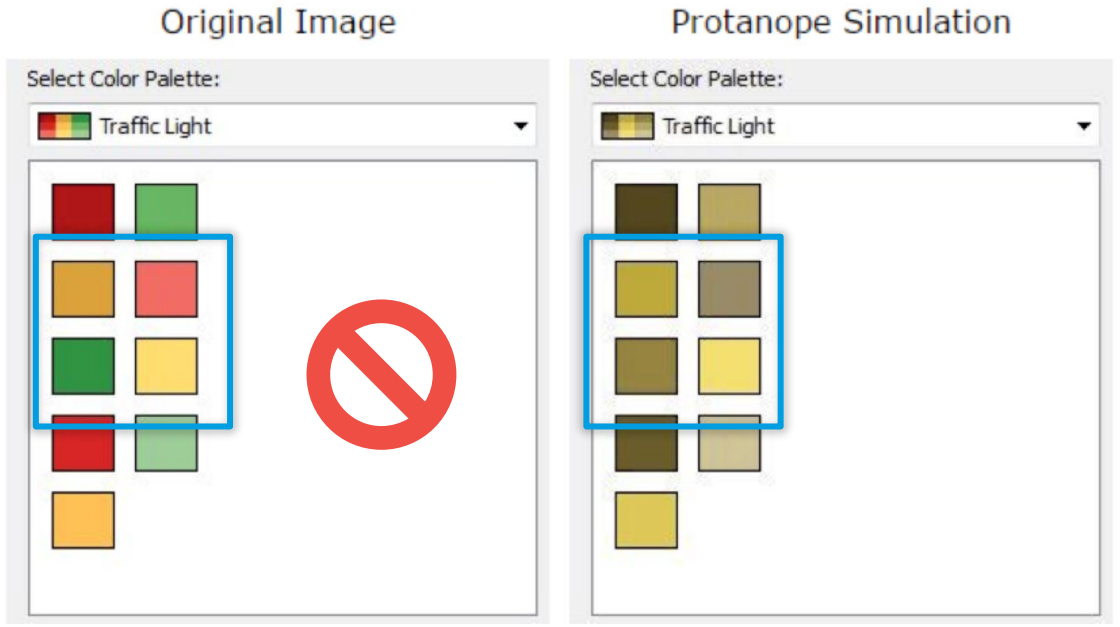
Size indicates Age (18 – 64)

THINGS TO AVOID

Red/Green **Deuteranopia** (6% of males)

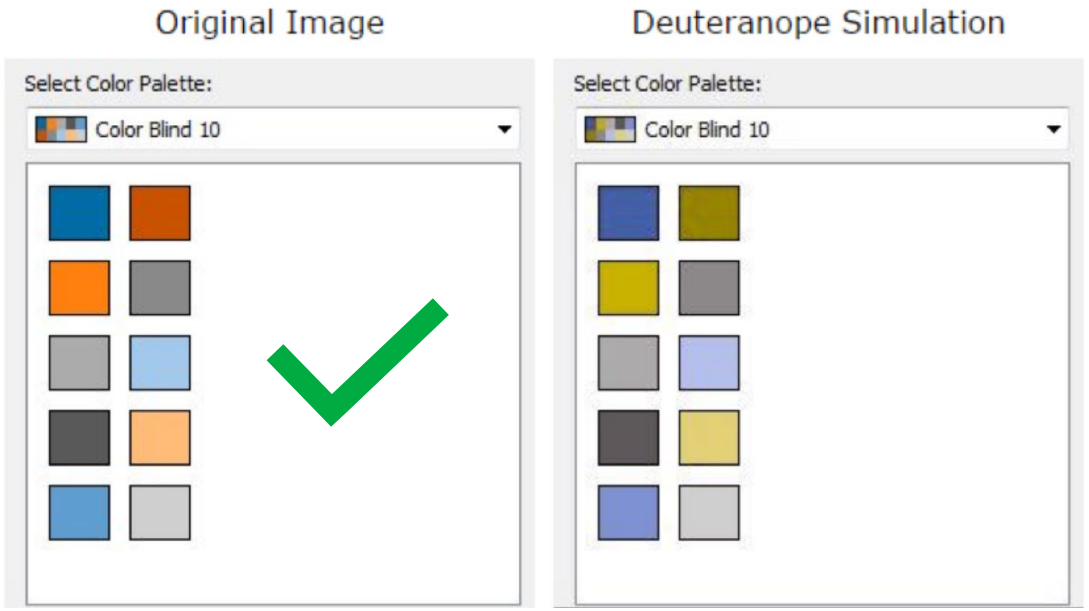


Red/Green **Protanopia** (2% of males)

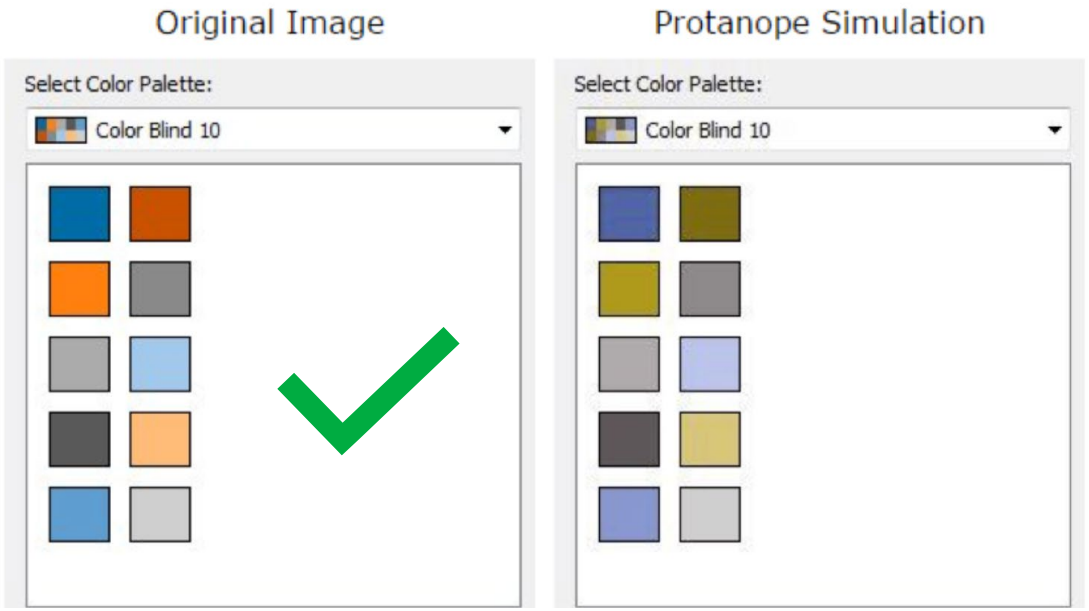


https://www.tableau.com/blog/examining-data-viz-rules-dont-use-red-green-together
https://en.wikipedia.org/wiki/Color_blindness

Red/Green **Deuteranopia** (6% of males)

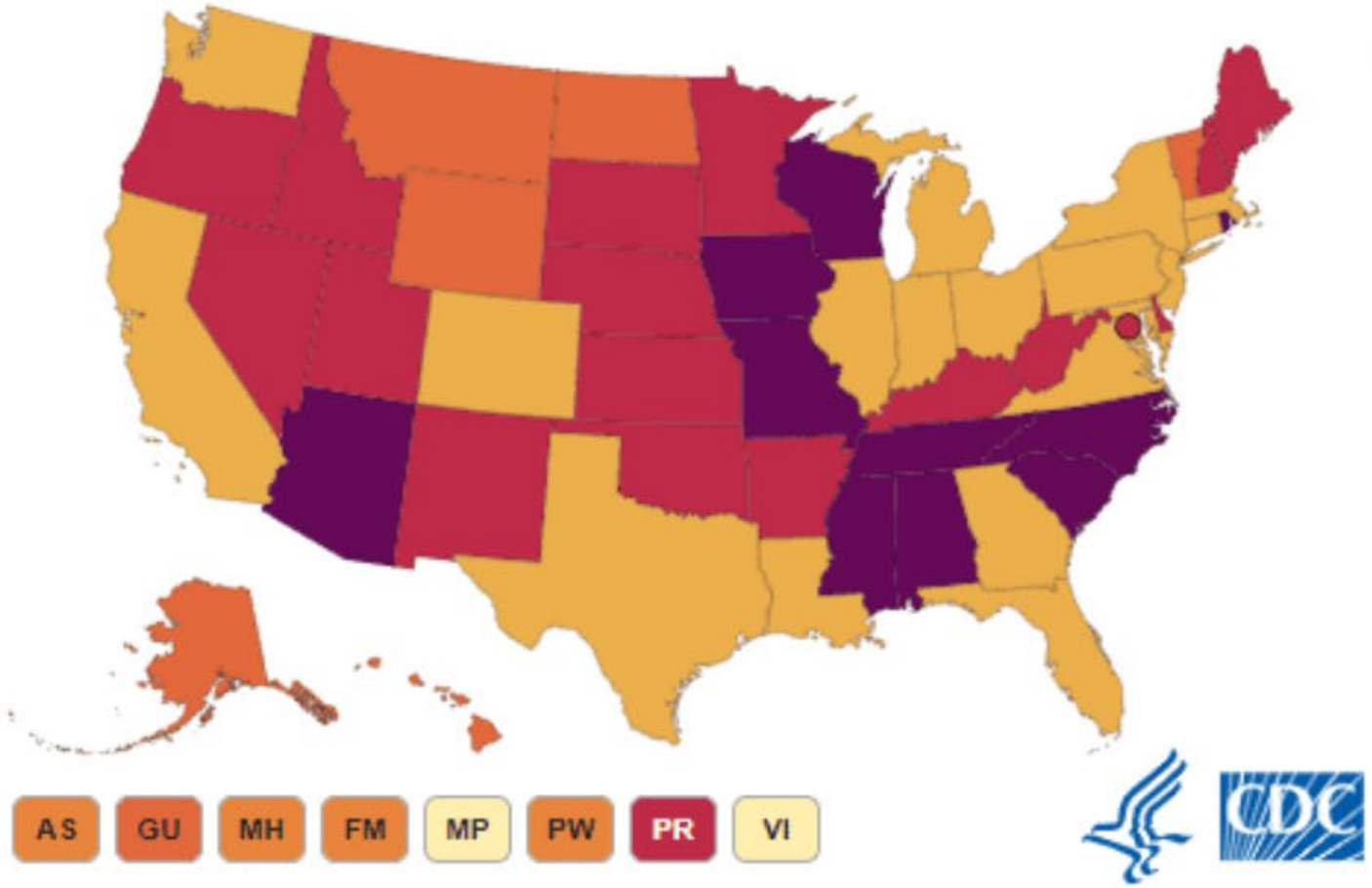


Red/Green **Protanopia** (2% of males)



https://www.tableau.com/blog/examining-data-viz-rules-dont-use-red-green-together
https://en.wikipedia.org/wiki/Color_blindness

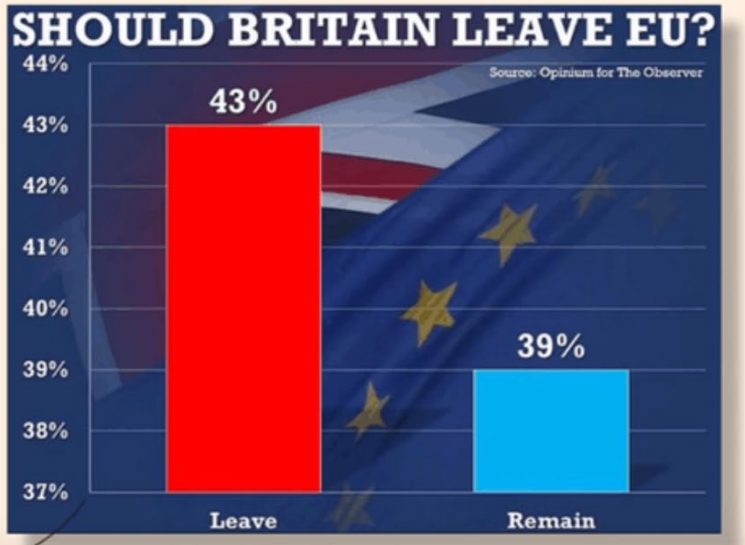
18 states report more than 10,000 cases of COVID-19.



https://www.oldstreetsolutions.com/good-and-bad-data-visualization

Graphics that are accurate but misleading

Baseline should start at zero,



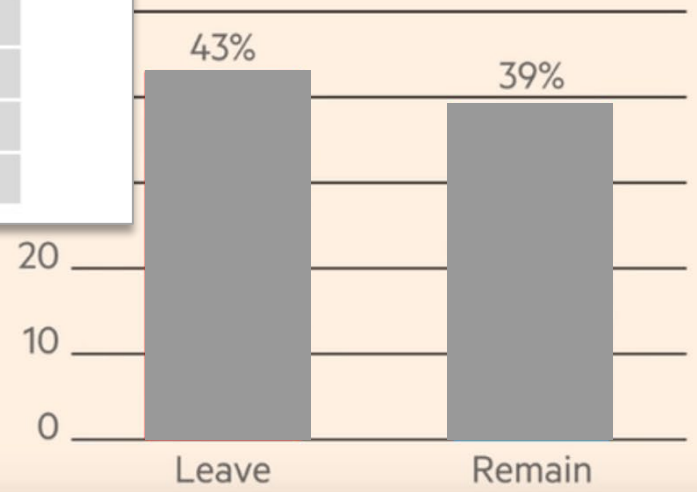
How Britain voted

Older people with fewer formal qualifications most likely to have voted Leave

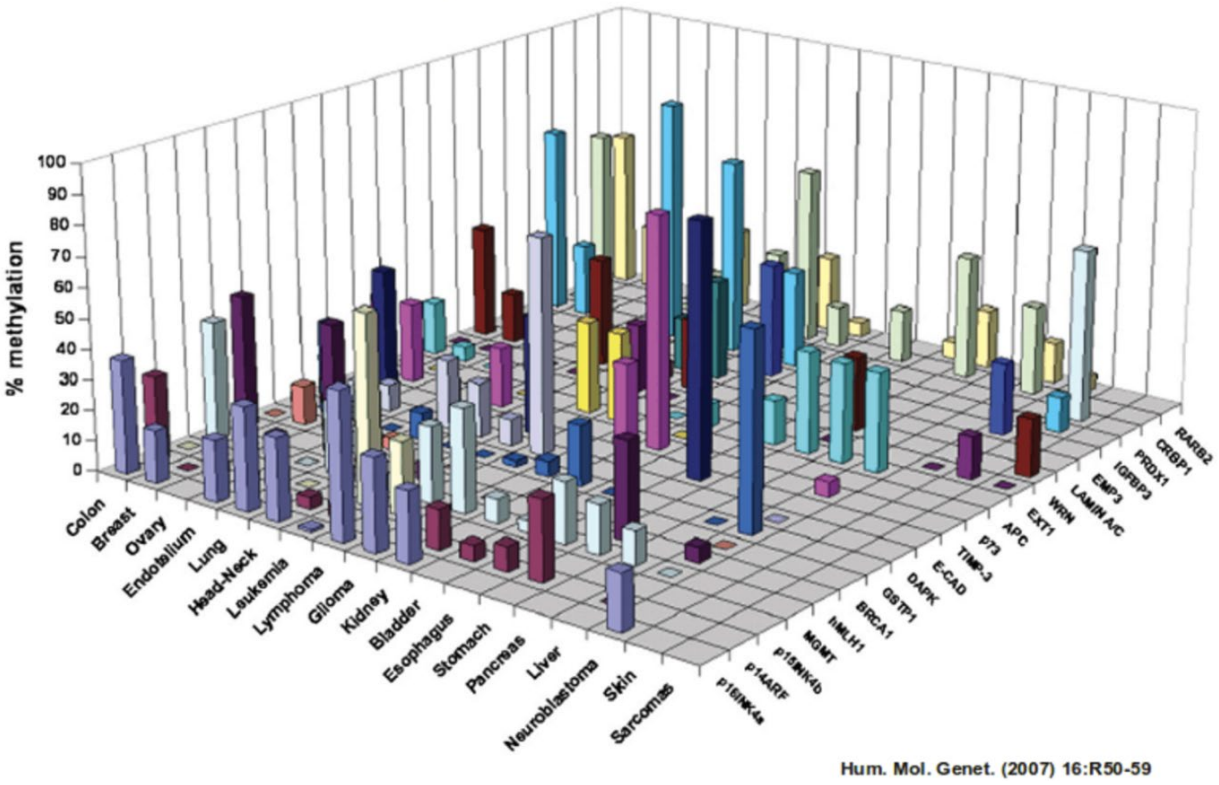
2015 vote	Remain	Leave
Conservatives	39	61
Labour	65	35
Liberal Democrat	68	32
UKIP	5	95
Green	80	20

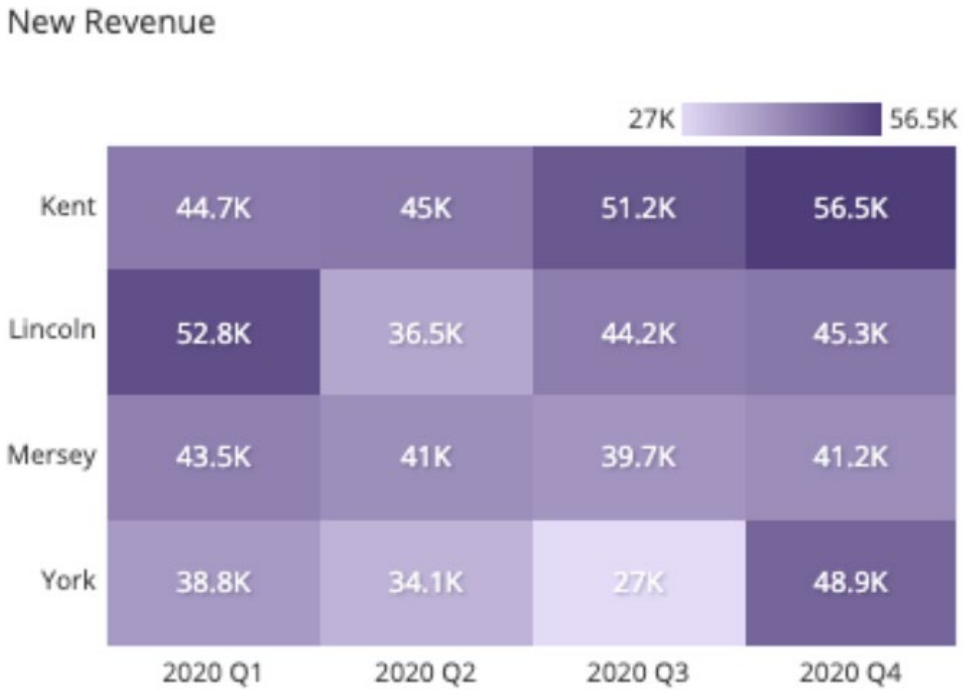
but misleading

A better chart of the same data



https://www.oldstreetsolutions.com/good-and-bad-data-visualization
https://yougov.co.uk/topics/politics/articles-reports/2016/06/27/how-britain-voted





https://www.oldstreetsolutions.com/good-and-bad-data-visualization

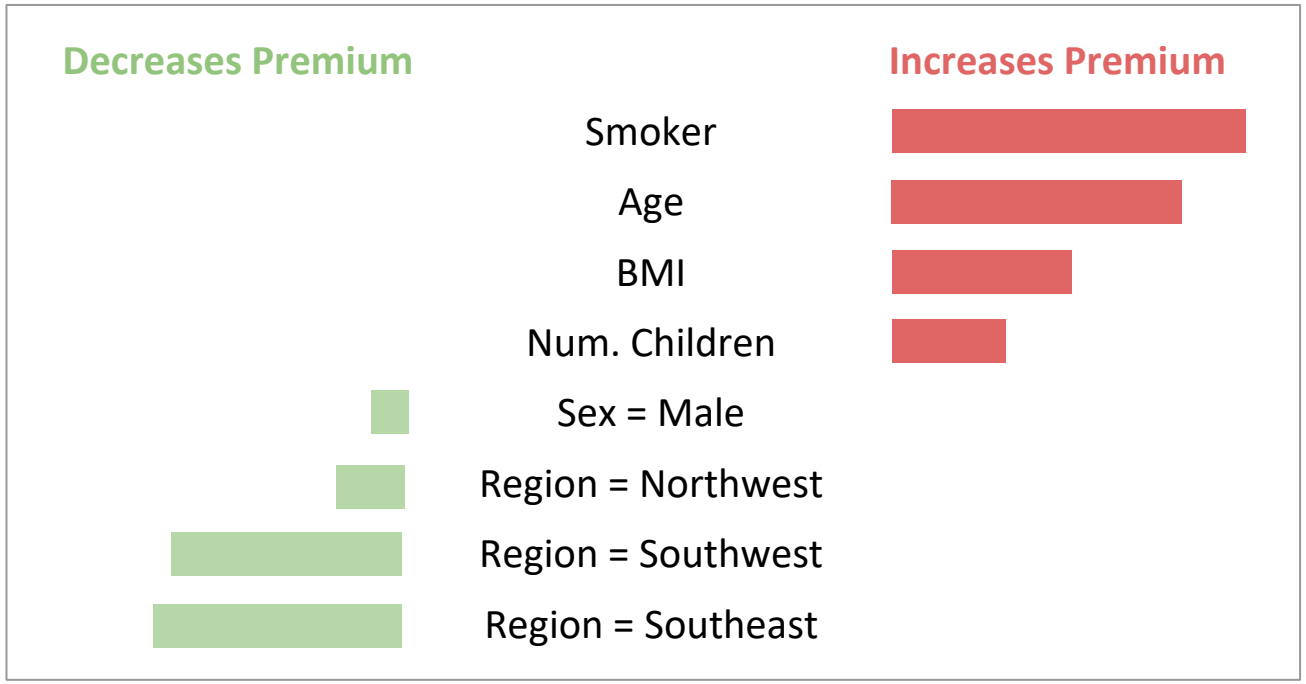
PROBLEMS & TOOLS TO FIX THEM

I HAVE DATA, BUT I DON'T KNOW WHAT STORY TO TELL.

- Build a predictive model and look at important features.
- Do this quickly with an automated machine learning tool: RapidMiner (point/click), storyteller (code/R), etc.

```r
# run model a model to find a story about charges (premiums)
dt %>%
  correlatedfeatures_address(
    target = 'charges'
  ) %>%
  fitmodel() %>%
  summary()
```

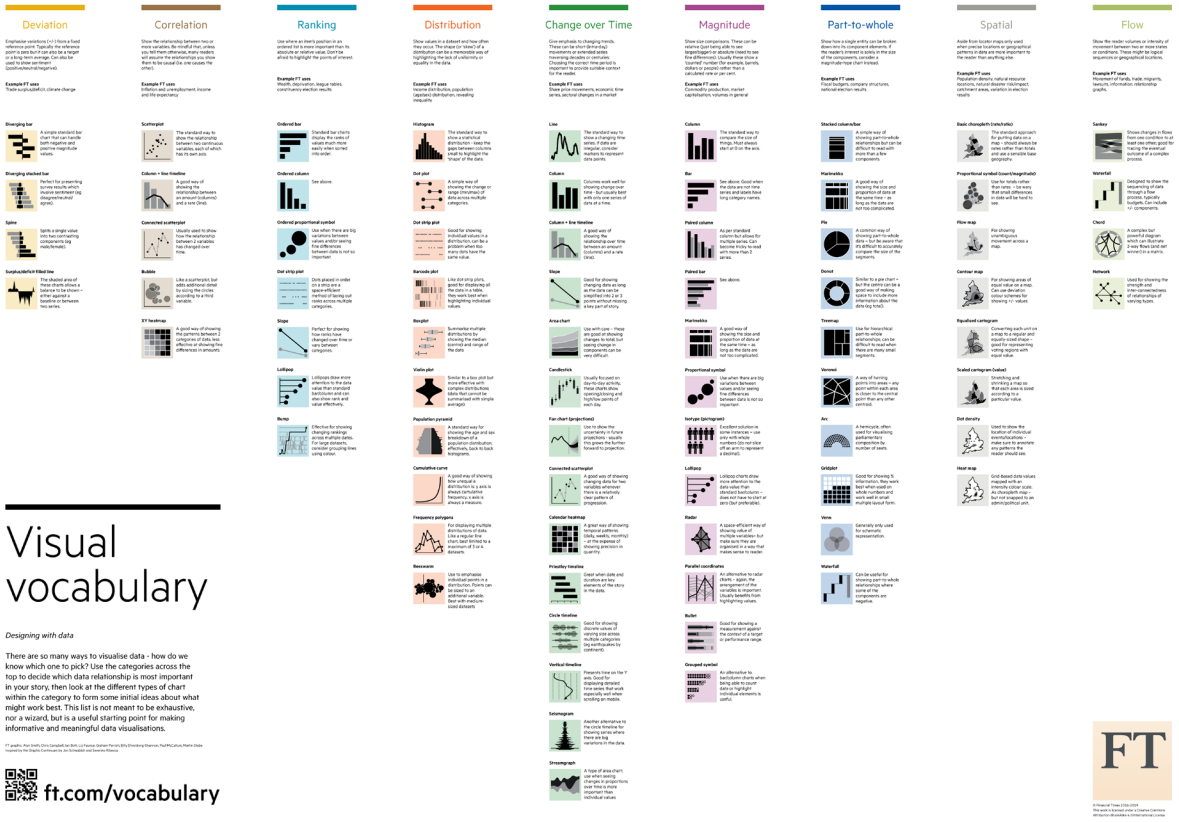
```
[1] "charges"

Call:
lm(formula = y ~ ., data = yX)

Residuals:
   Min     1Q Median     3Q    Max
-10584  -2748  -1068   1092  24373

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -11003.70     965.52 -11.397  < 2e-16 ***
age                 250.67      11.48  21.841  < 2e-16 ***
bmi                 317.37      28.15  11.273  < 2e-16 ***
children            519.40     132.24   3.928 9.03e-05 ***
smokerTRUE        22885.98     403.69  56.692  < 2e-16 ***
sex.male           -106.99     320.02  -0.334   0.7382
region.southeast  -1072.18     460.65  -2.328   0.0201 *
region.northwest   -444.34     456.46  -0.973   0.3305
region.southwest  -1021.20     458.71  -2.226   0.0262 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5787 on 1309 degrees of freedom
Multiple R-squared:  0.7448,    Adjusted R-squared:  0.7432
F-statistic: 477.4 on 8 and 1309 DF,  p-value: < 2.2e-16
```

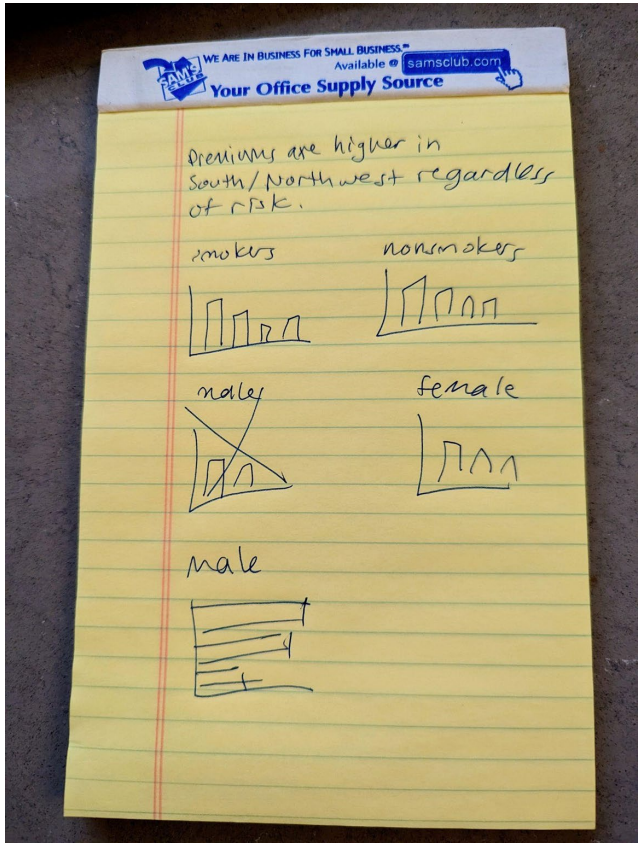


https://www.kaggle.com/code/brycechamberlain/data-explore-automl/

I KNOW THE STORY I WANT TO TELL, BUT WHAT VIZ SHOULD I USE?

- Get inspiration from the Financial Times Visual Vocabulary
- Use hand-drawing to quickly prototype an idea.

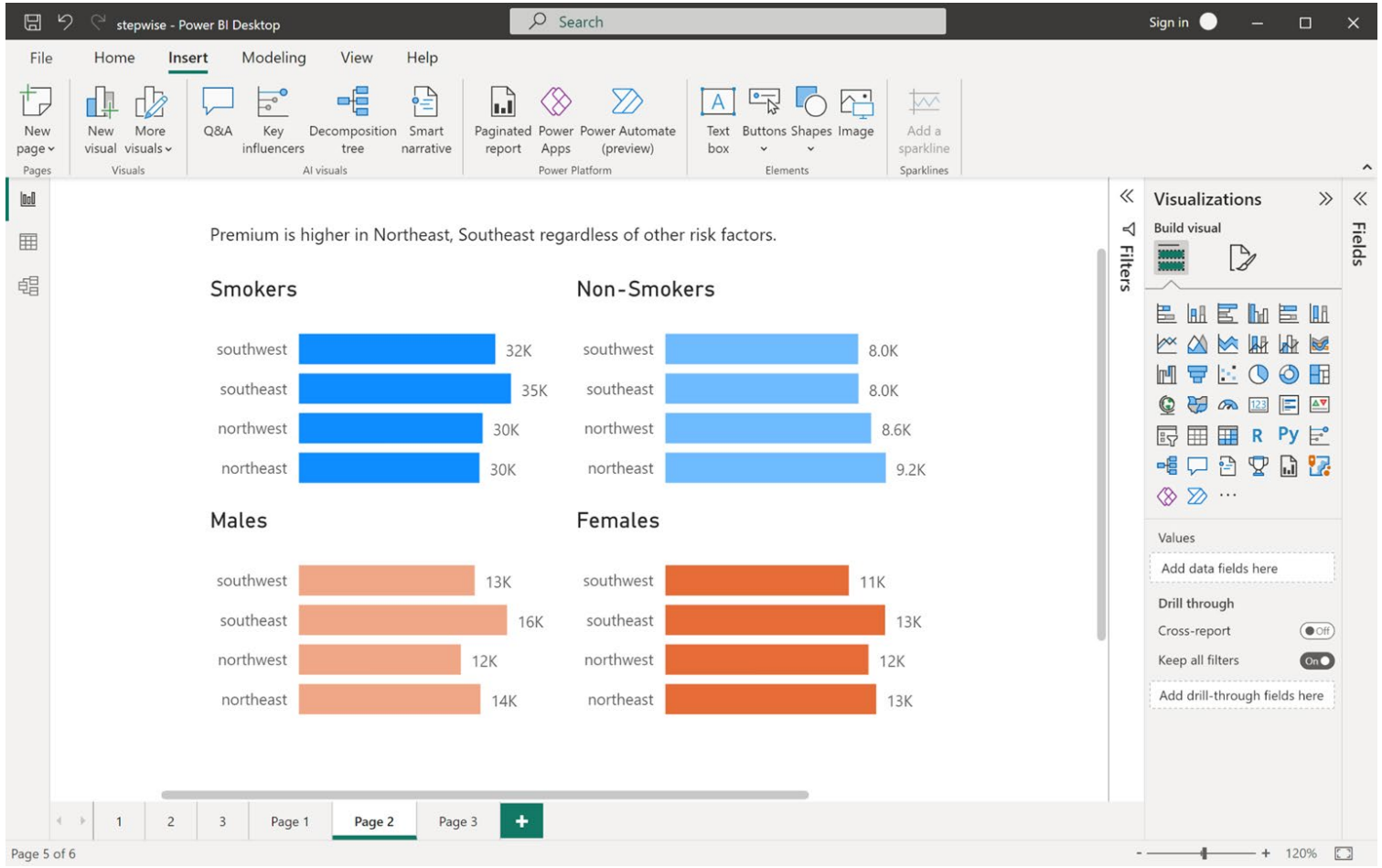




https://www.ft.com/vocabulary

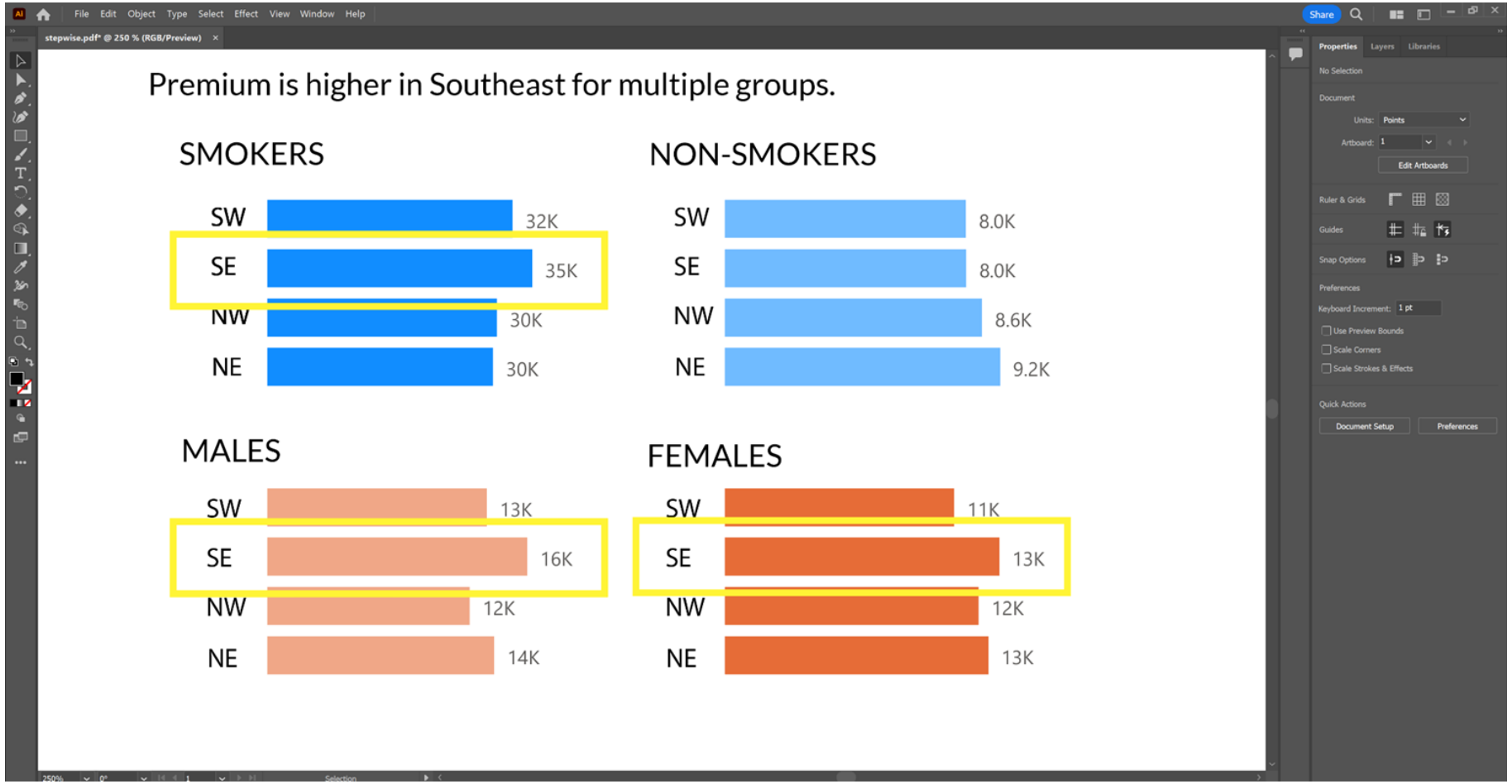
I AM READY TO MAKE A VISUALIZATION, BUT I AM SHORT ON TIME.

- Use a click/drag tool: PowerBI Free Desktop, Excel PivotChart, etc.
- Avoid code-based solutions that can take a long time.



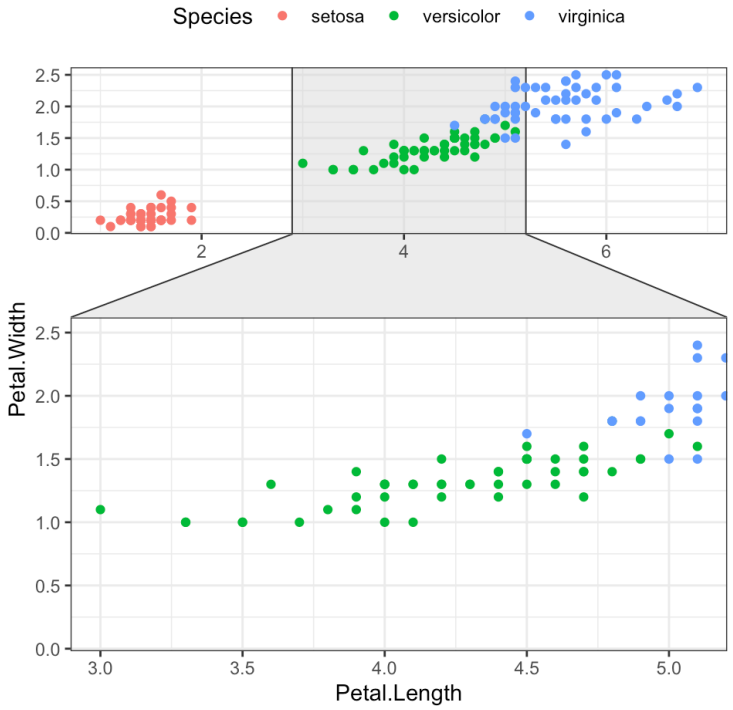
I'M NOT HAPPY WITH THE DESIGN OPTIONS IN POWER BI

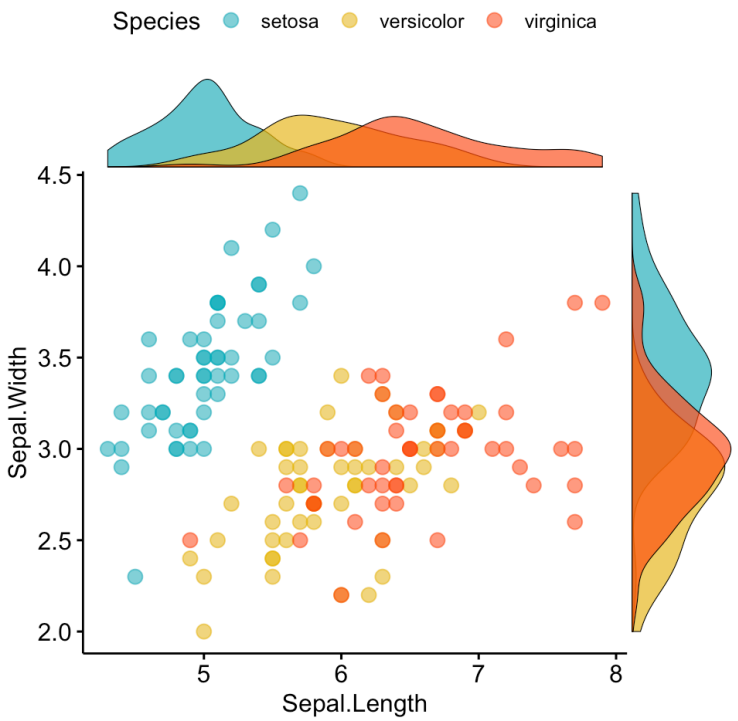
- Export from GUI to PDF and edit with a design tool like Adobe Illustrator, Inkscape, or Fiji.
- This is not intuitive, so watch a video to see how (link on last slide).

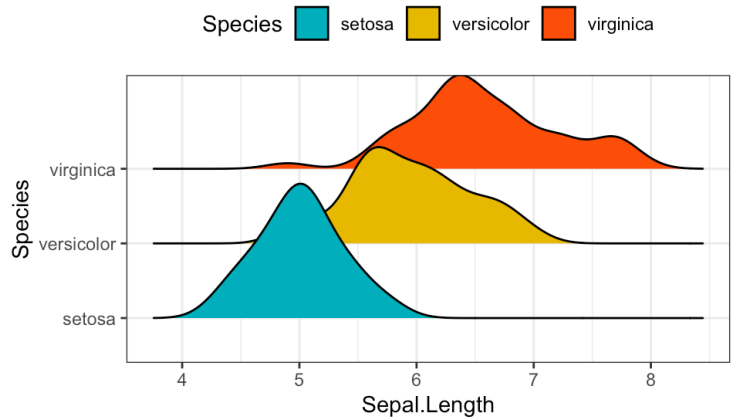


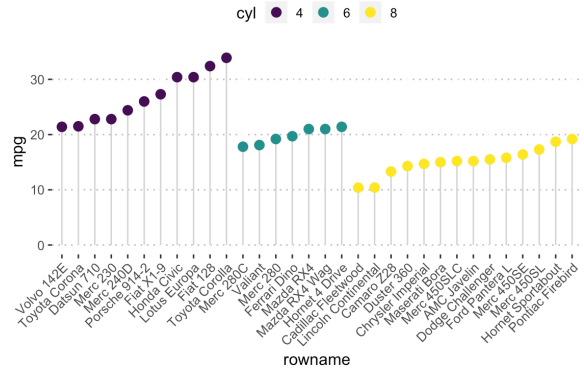
I AM NOT HAPPY WITH CHART TYPE / NEED TO MAKE MANY SIMILAR CHARTS

- Use a code-based solution so you can automate the process or use new chart types.
- R: ggplot, RMarkdown; Python: seaborn, Jupyter.





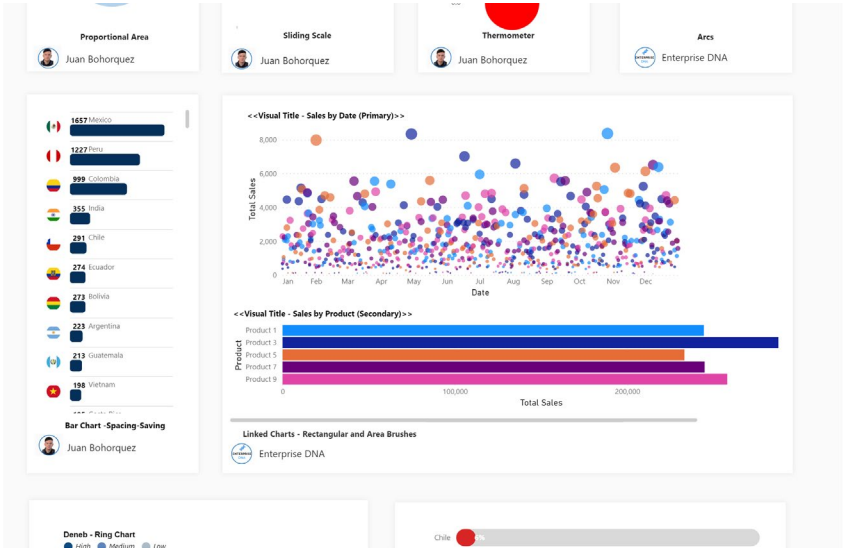


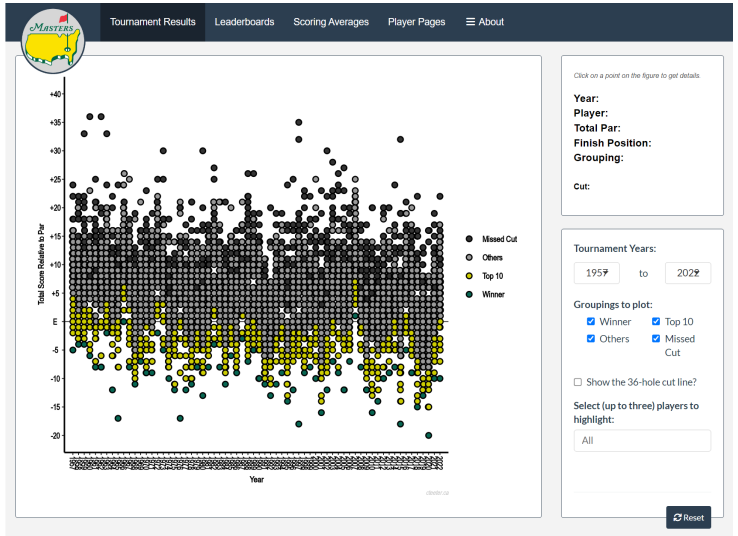


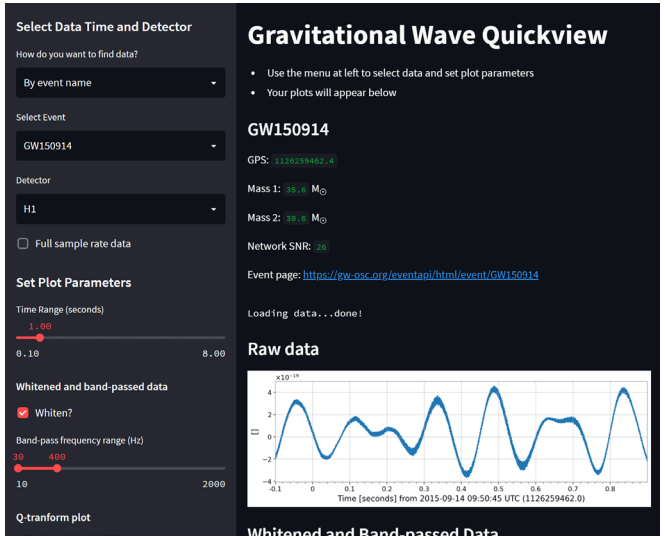
https://www.datanovia.com/en/blog/ggplot-examples-best-reference/

I NEED TO BUILD A DYNAMIC APPLICATION.

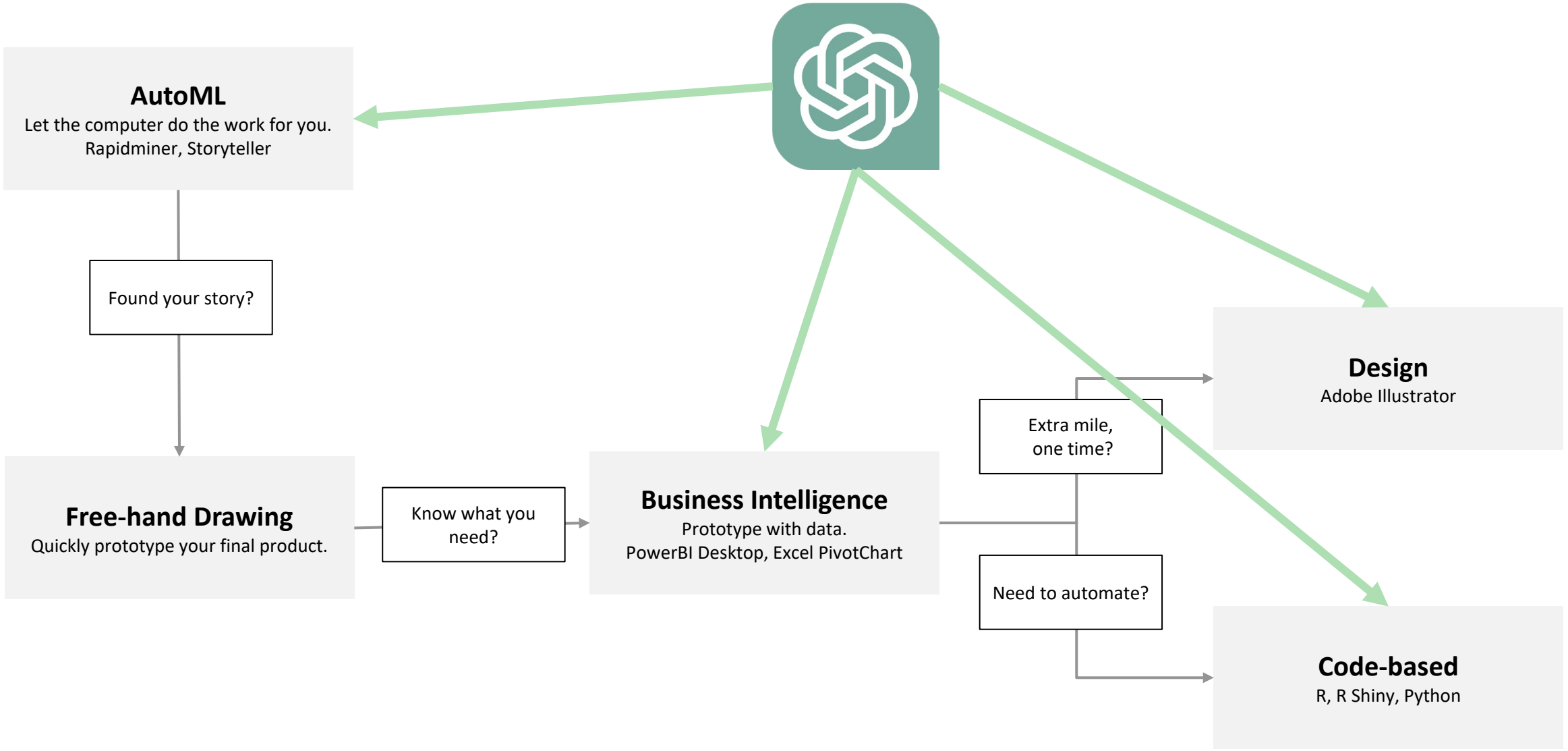
- Use a tool you can publish to the web and give users power to filter, etc.
- Power BI, Tableau (click/drag), R Shiny (code, R) or Streamlit (code, Python)







https://community.powerbi.com/t5/Data-Stories-Gallery/My-own-Gallery/td-p/3054132
https://shiny.rstudio.com/gallery/masters.html
https://gw-quickview.streamlit.app/


AutoML
Let the computer do the work for you.
Rapidminer, Storyteller
Found your story?
Free-hand Drawing
Quickly prototype your final product.
Know what you need?
Business Intelligence
Prototype with data.
PowerBI Desktop, Excel PivotChart
Extra mile, one time?
Need to automate?
Design
Adobe Illustrator
Code-based
R, R Shiny, Python

CHAT GPT: FINDING A STORY

GPT recommends regression analysis: let's ask it to do that.

*Prompt 1: In the attached dataset, **what are the important features in***

Response: "[term-by-term correlation indicating important features: smoker, age, bmi, children, sex, region. Includes python script] … For more complex relationships, **other methods, such as regression analysis** or machine learning, may be necessary."

We get correct explanations of important features with just a few prompts!

*Prompt 2: Please update your response to include cleaning the data and **using a regression analysis**.*

Response:

"[Python script] After cleaning the data and performing a linear regression analysis, we obtained the following results: …

R-squared (R^2), is approximately 0.753 … means that about **75.3% of the variability in charges can be explained** by the model …

being a smoker has the largest impact on charges. On average, **being a smoker increases charges by about $23,261** compared to non-smokers.

`children` has a positive effect on charges. **Each additional child increases charges by about $515**, all else being equal.

`bmi` also has a positive effect on charges. **Each additional unit of BMI increases charges by about $314**, all else being equal.

`age` has a positive effect on charges. **Each additional year of age increases charges by about $260**, all else being equal.

`sex_male` has a negative effect on charges. **Being male decreases charges by about $205** compared to females, all else being equal.

The different regions also have a negative effect on charges compared to the northeast region. **Being in the northwest, southwest, or southeast regions decreases charges by about $677, $929, and $1022 respectively**, all else being equal.

…"

CHAT GPT: ANOTHER TEST

Create dummy data with 1,000 rows:

- v1 uniform mean 15
- v2 normal with mean 100
- v3, v4, v5, lognormal x 1, 10, 30
- target = 80 + rnorm(mean = 30, n = 1000) + v1 * **10** + v2 * **300**

Prompt 1: Please clean this data and use a linear regression model to explain the impact of each variable on the column y.

Response: [Python code and results that didn't make sense until I realized it had scaled all values to between 0 and 1.

Prompt 2: please do the same analysis but without scaling or normalizing the data.

It gets the correct values! *We know they are correct because we created them.*

Response: [Python code] …

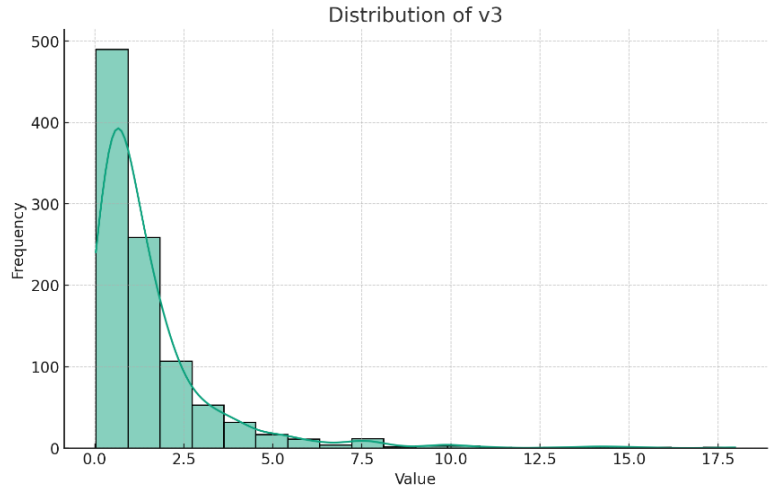
- *V1: A one-unit increase in v1 is associated with an increase of approximately **9.99** in the target variable.*
- *V2: A one-unit increase in v2 is associated with an increase of approximately **300.00** in the target variable. This variable seems to have the most significant impact, consistent with the previous analysis.*
- *v3: The coefficient is close to zero, indicating that v3 has almost no effect on the target variable.*
- *v4: A one-unit increase in v4 is associated with a decrease of approximately 0.0016 in the target variable.*
- *v5: A one-unit increase in v5 is associated with an increase of approximately 0.00026 in the target variable.*

TAKING IT FURTHER: IDENTIFY A DISTRIBUTION

Prompt 3: Please identify the distribution of column v3.

We know it is lognormal with shape = 1, location = 0, scale = 1

Response: [Python code to gather stats and plot histogram] …




```
RESULT
count    1000.000000
mean        1.556172
std         1.912439
min         0.033280
25%         0.486068
50%         0.948349
75%         1.837909
max        17.977436
Name: v3, dtype: float64
```

… The right-skewed shape might suggest that the distribution of "v3" could follow a log-normal, exponential, or gamma distribution…

Gets very close! A bit of error is expected with only 1,000 simulations.

Prompt 4: please identify if it follows a log-normal, exponential, or gamma distri meters of the selected distribution.

Response: [Python code] … Log-Normal Distribution: Shape parameter (s): **1.0274** Location parameter (loc): **0.0219** Scale parameter (scale): 0.9853 … The visual comparison shows that **the log-normal distribution** appears to be the closest fit to the data, as the red line follows the histogram more closely… KS Statistic = 0.0214, P-Value = 0.7386 … The exponential and gamma distributions have very low p-values, suggesting that they do not fit the data well…

CHAT GPT: CREATE A VISUALIZATION 1

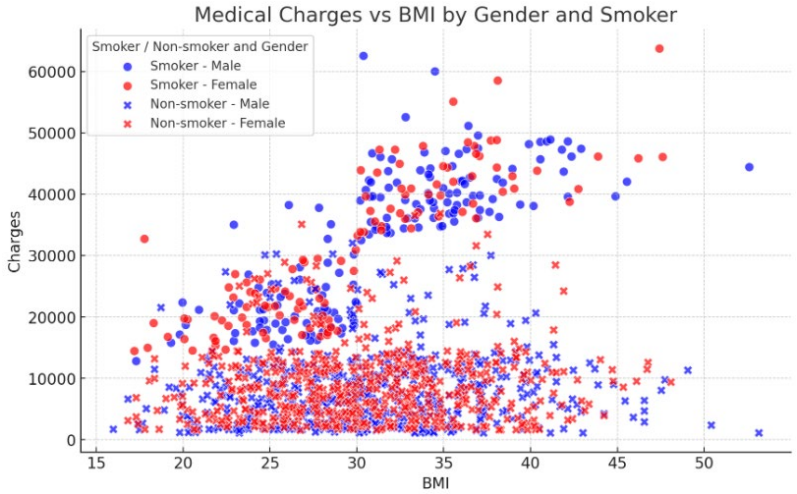
Prompt 3 : Please create a chart that shows the relationship between charges, bmi, gender, and smoker.

Response: [Python script and pair plot that I didn't like]

Prompt 4: Please create one chart that shows the relationship between all these variables.

Response:

[Python script]



Very close to our chart in a fraction of the time to load into PowerBI!

We get some interpretation as a bonus.

"… From the plot, you can observe the following:

Among smokers (red points), there is a clear positive correlation between BMI and charges. That is, smokers with higher BMI tend to have higher charges.

Among non-smokers (blue points), there is no clear correlation between BMI and charges.

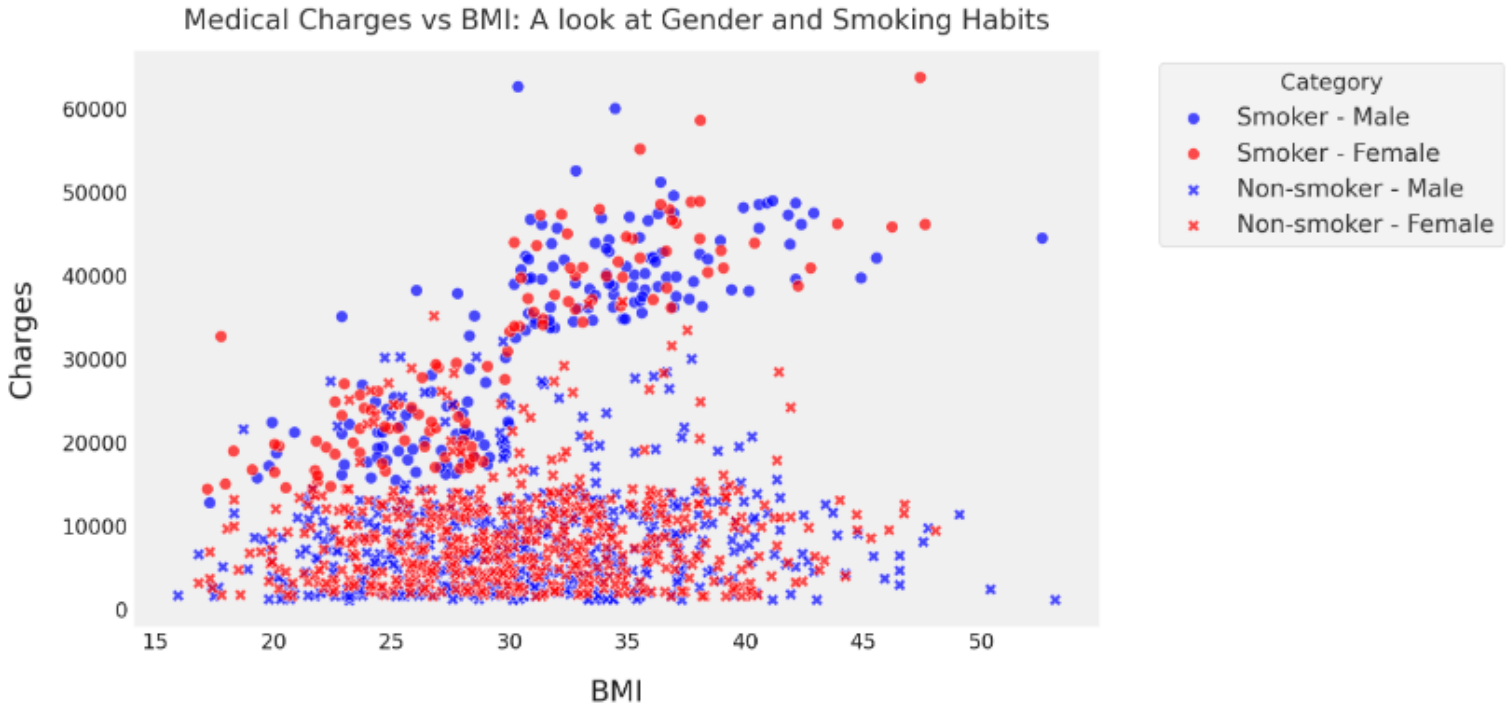
Both men and women follow the same general patterns in terms of the relationship between BMI, smoker status, and charges. However, it's worth noting that there are slightly more male smokers with high charges and high BMI compared to female smokers.

…"

CHAT GPT: CREATE A VISUALIZATION 2

Prompt 4: Please clean this chart up to style it for use in a newspaper.

Response: [Python script]

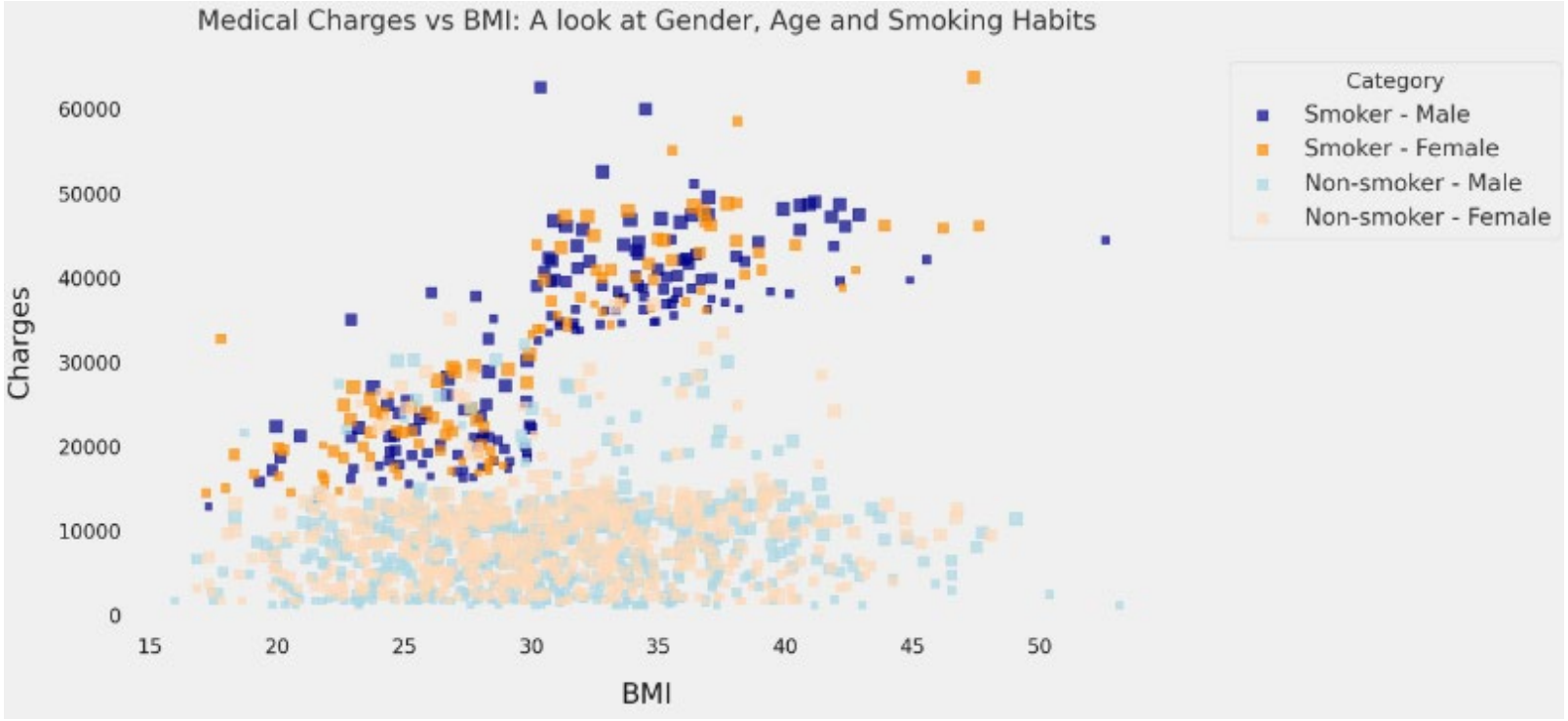


“Here's the updated, stylized chart suitable for a newspaper publication. It has a clear title, labeled axes, and a legend explaining the color and marker coding for the different categories of individuals (smokers/non-smokers and males/females). I have removed the grid lines and spines on the top and right for a cleaner look…”

CHAT GPT: CREATE A VISUALIZATION 3

Prompt 4: please modify this chart to use boxes instead of circles, with dark colors indicating smokers and light indicating non-smokers, and change the colors so that a color-blind individual can easily understand the plot, and change the boxes so that they their size indicates the age.

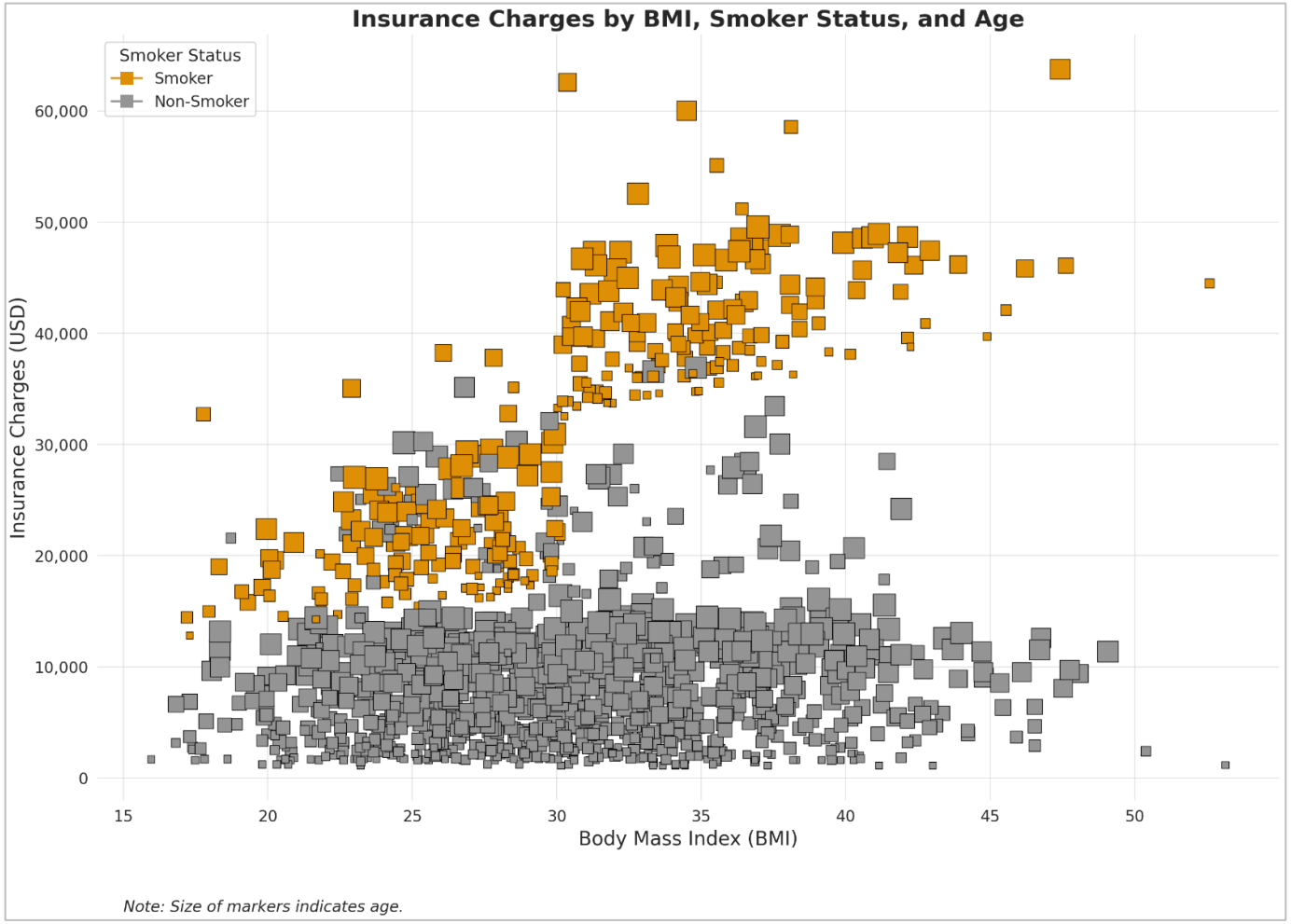
Response: [Python script]



"… The dark colors represent smokers and the light colors represent non-smokers, providing contrast between these two groups. The colors have also been chosen to be distinguishable for individuals with color-blindness..."

CHAT GPT: CREATE A VISUALIZATION 4

After a few more revisions.. we get the below chart which isn't perfect but it's pretty good.



PROBLEM: I CAN'T PUT MY PRIVATE DATA INTO CODE INTERPRETER

Confirm with your company's IT that this is OK!!

- Use the fakeR package (or another method) to create dummy data from your actual data, with the same variable names and types.
- Run Code Interpreter prompts on this dummy data.
- Copy the code into a Jupyter Notebook and replace dummy data with your actual data.
- Replicate the process used by Code Interpreter to find insights.

QUESTIONS

Dalesa Bady, ACAS, MAAA
Actuary, GuideOne Insurance
linkedin.com/in/dalesa-bady-acas-05822336

Bryce Chamberlain, ASA, MSCA
Principal, Oliver Wyman Actuarial Consulting
linkedin.com/in/brycechamberlain

Get this deck + linked resources + Chat GPT prompts and responses @ https://bit.ly/gis23-dataviz

Worksheets
Calculator
Excel, Google Sheets

Most people can use it. Easily manipulate single records.	Easy to make mistakes. Difficult to automate. Slow on large data.

Business Intelligence
Get started now
Power BI, Tableau, PivotCharts

Lots of options quickly. Click & drag	Limited functionality. Difficult to automate.

AutoML
Search for insights
Rapidminer, Storyteller

Find stories across all data.	Limited visualization. Results are complex. Expensive if not open source.

Design
Make it pretty
Adobe Illustrator, Inkscape

Lots of features and options for perfecting the visual.	Very time consuming. Software is complex, difficult to learn.

Free-hand Drawing
Begin with the end in mind
Pen & Paper, Tablet

Fastest method, no interface to slow you down.	Not generated by data. Not fit for delivery.

Code-based
Automate repetitive tasks
RMarkdown, R Shiny

Unlimited functionality. Open source. Git version control.	Time consuming. Need to code.