



# 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



**1**

Why Visualize?

**2**

How We Think Visually

**3**

Things To Avoid

**4**

Problems & Tools to Solve Them

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

# 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?

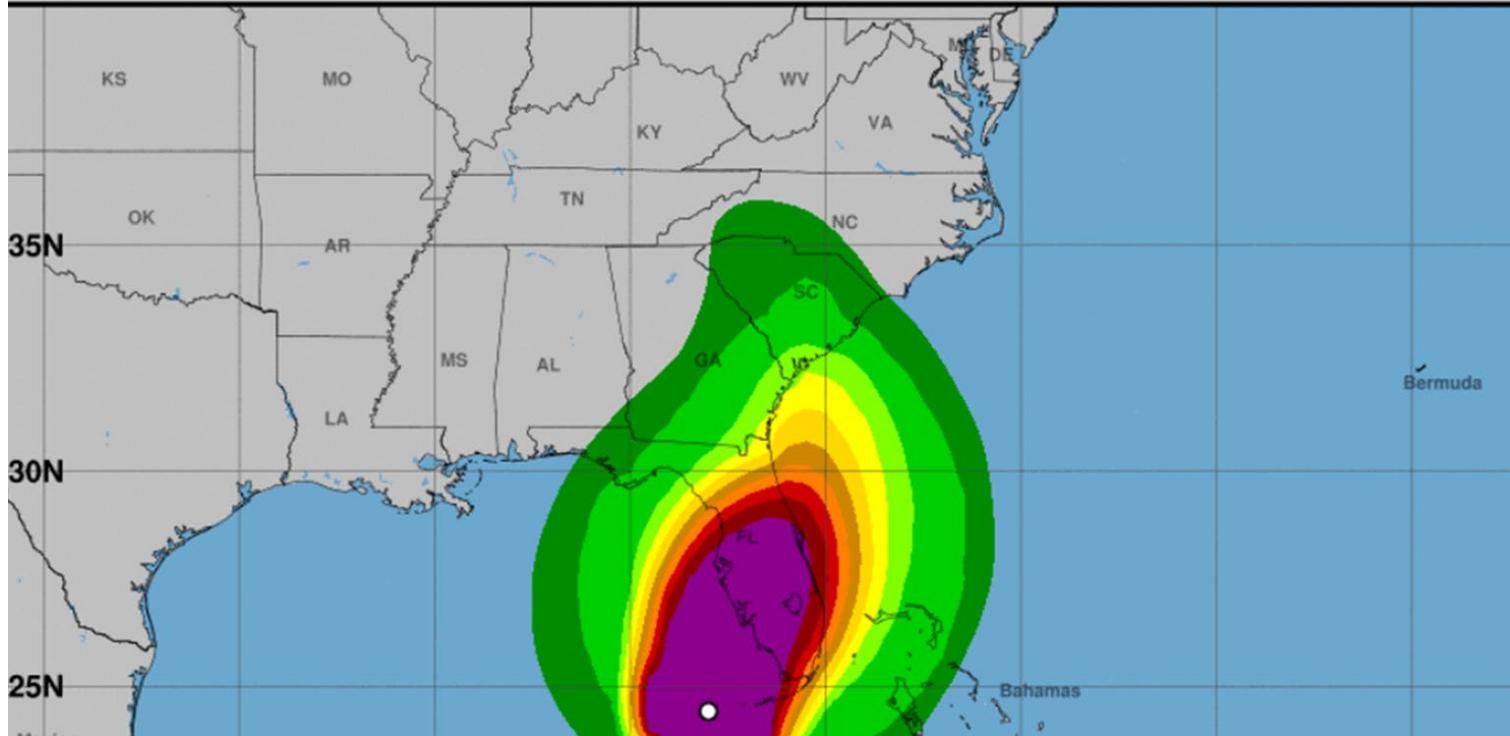


Beware of the road ahead. It might have rained or snowed and it could be slippery, which might cause an incident. Be careful.



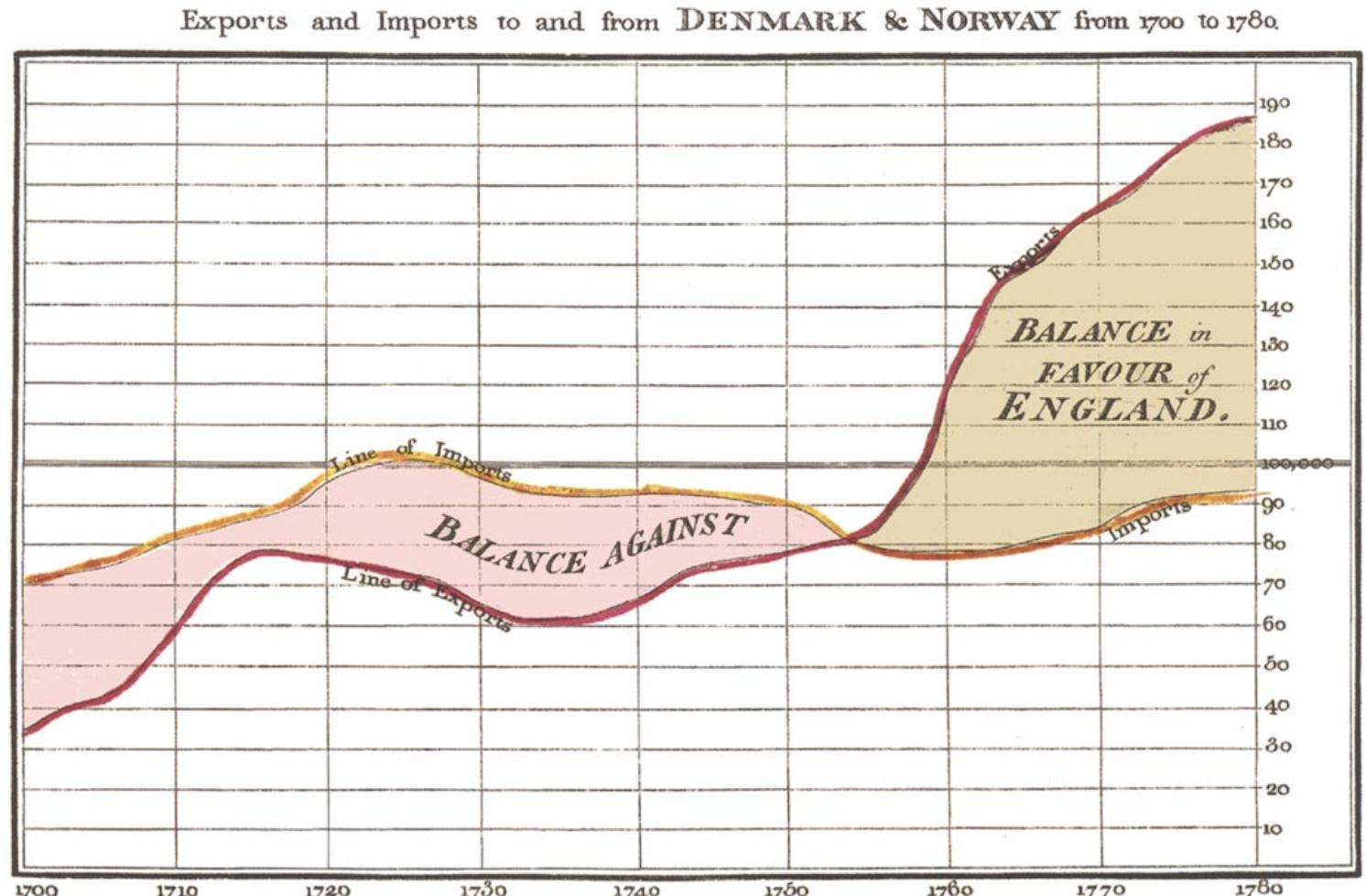
## Tropical-Storm-Force Wind Speed Probabilities (Preliminary)

For the 120 hours (5.0 days) from 8 AM EDT WED SEP 28 to 8 AM EDT MON OCT 03



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

“The Commercial and  
Political Atlas”  
- William Playfair (1786)



The Bottom line is divided into Years, the Right hand line into £10,000 each.  
Published as the Act directs, 1<sup>st</sup> May 1786. by W<sup>m</sup> Playfair  
Neale sculpt 352, Strand, London.

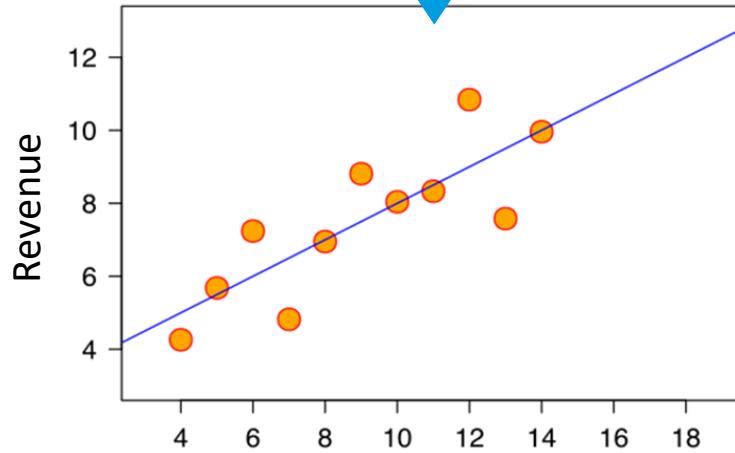
“

Well, just show me the numbers.

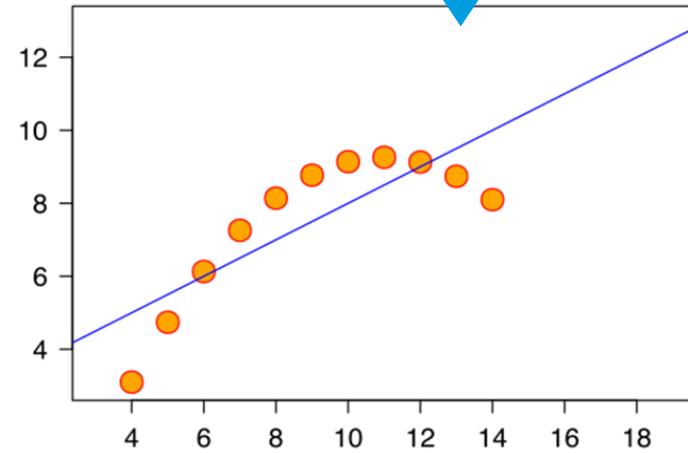
Let's take a look at an example.

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

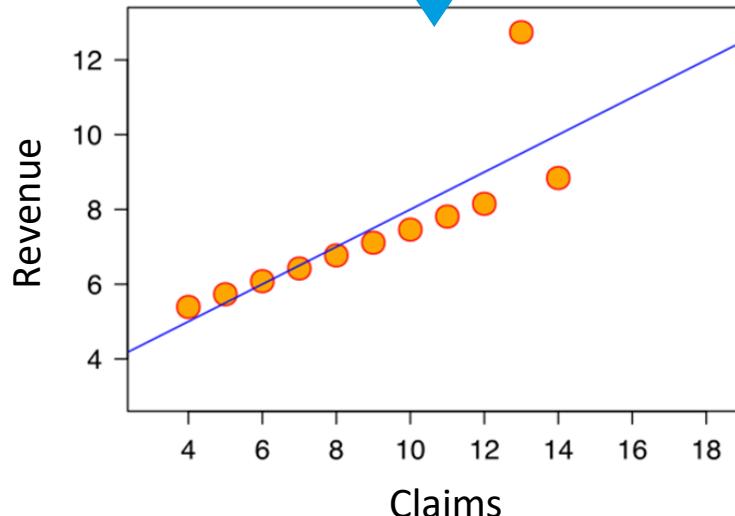
Revenue increases with claims.



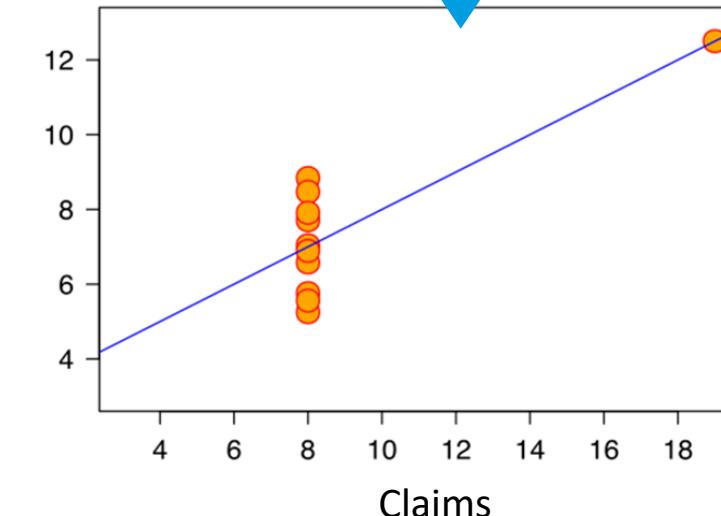
Revenue capped.



Revenue outlier.



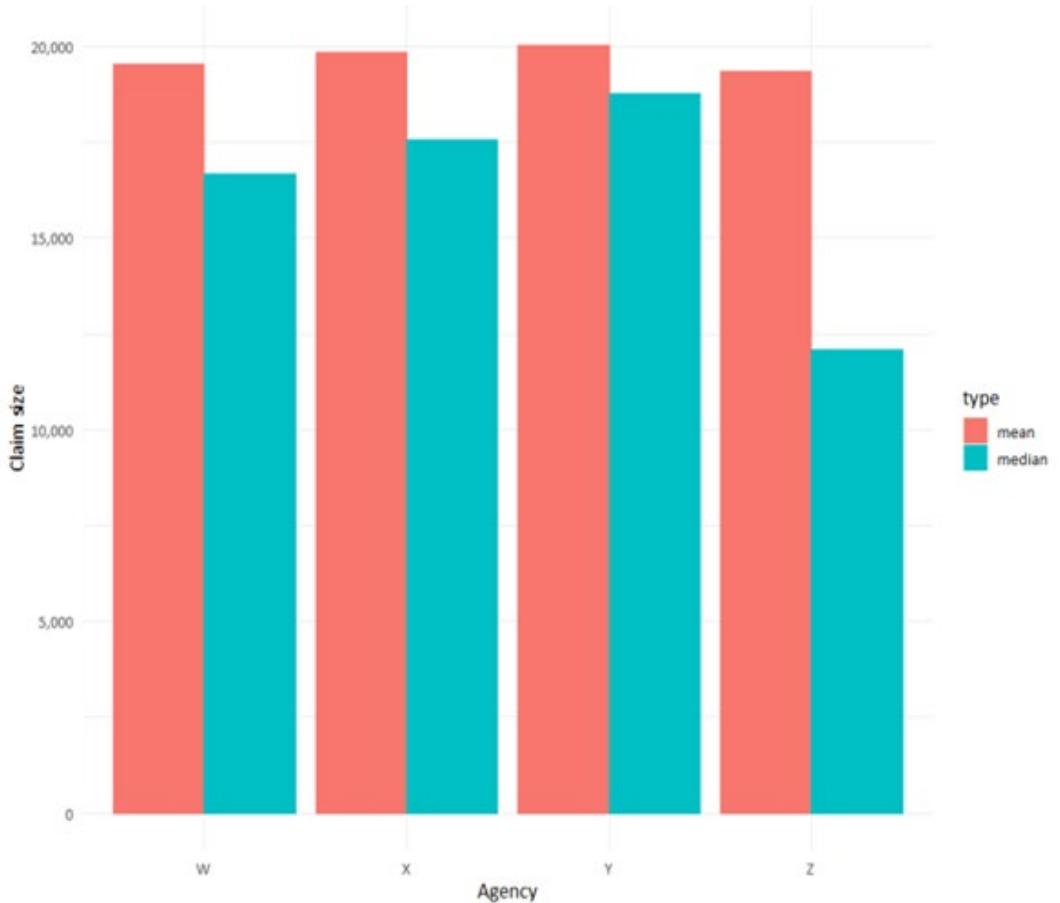
Revenue and amount outlier.



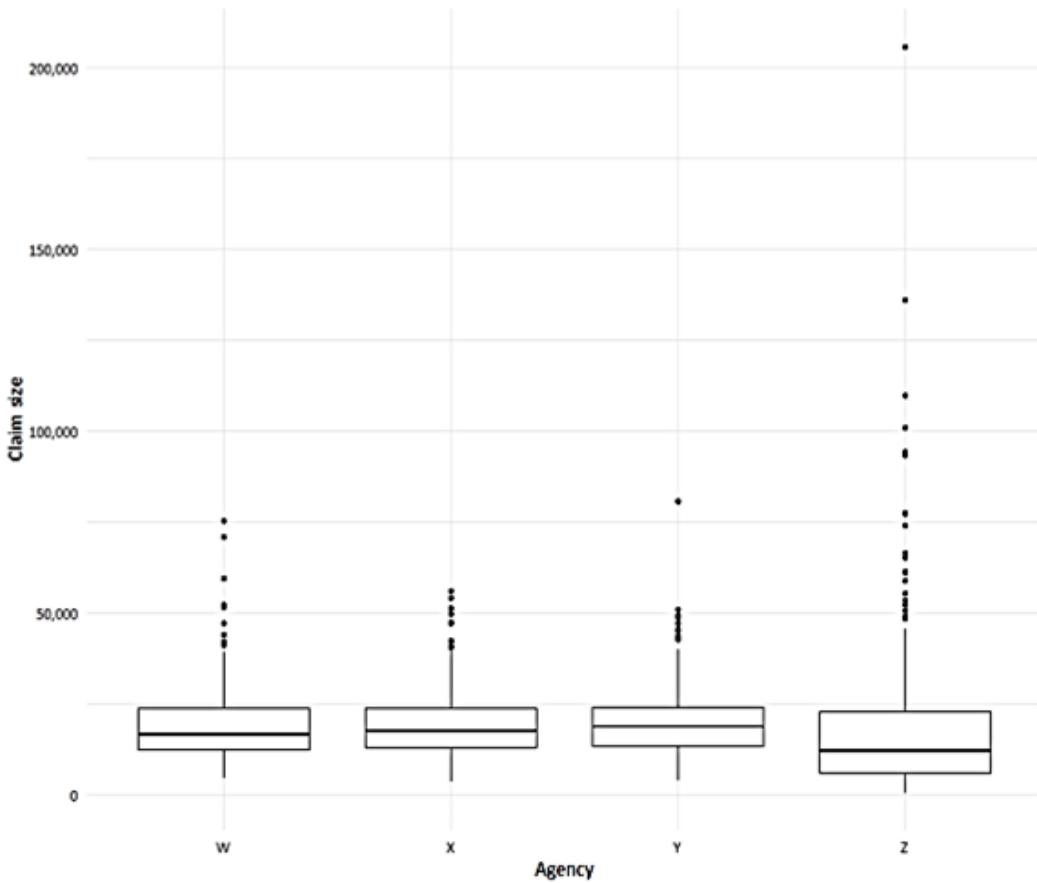
“

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?**

**99999999**

**9999999999**

**999,999,999**

**99,999,999,999**

**999,999,999**

**99,999,999,999**

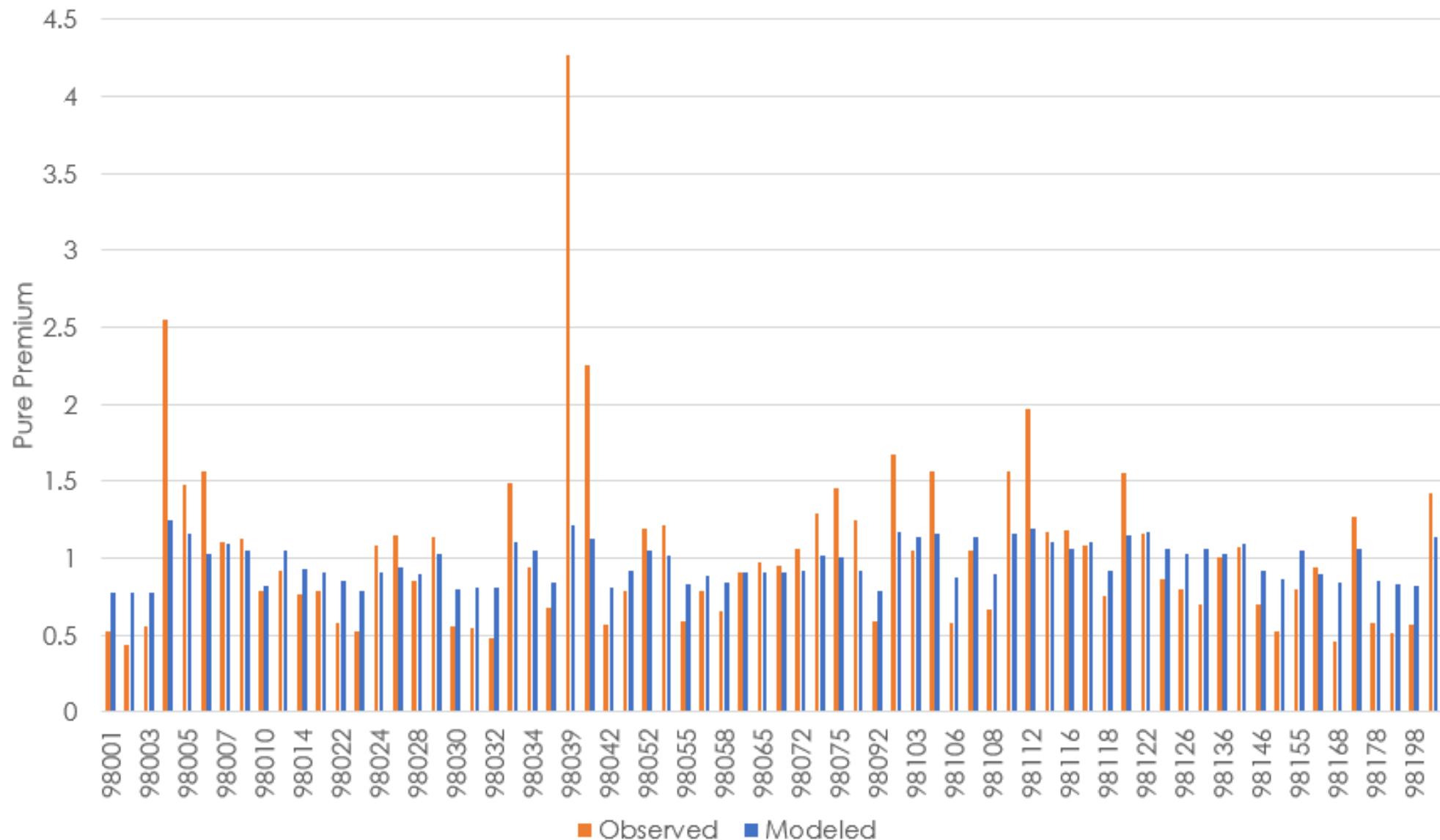


Mapping the values in a visual space helps us understand the enormity of the difference!

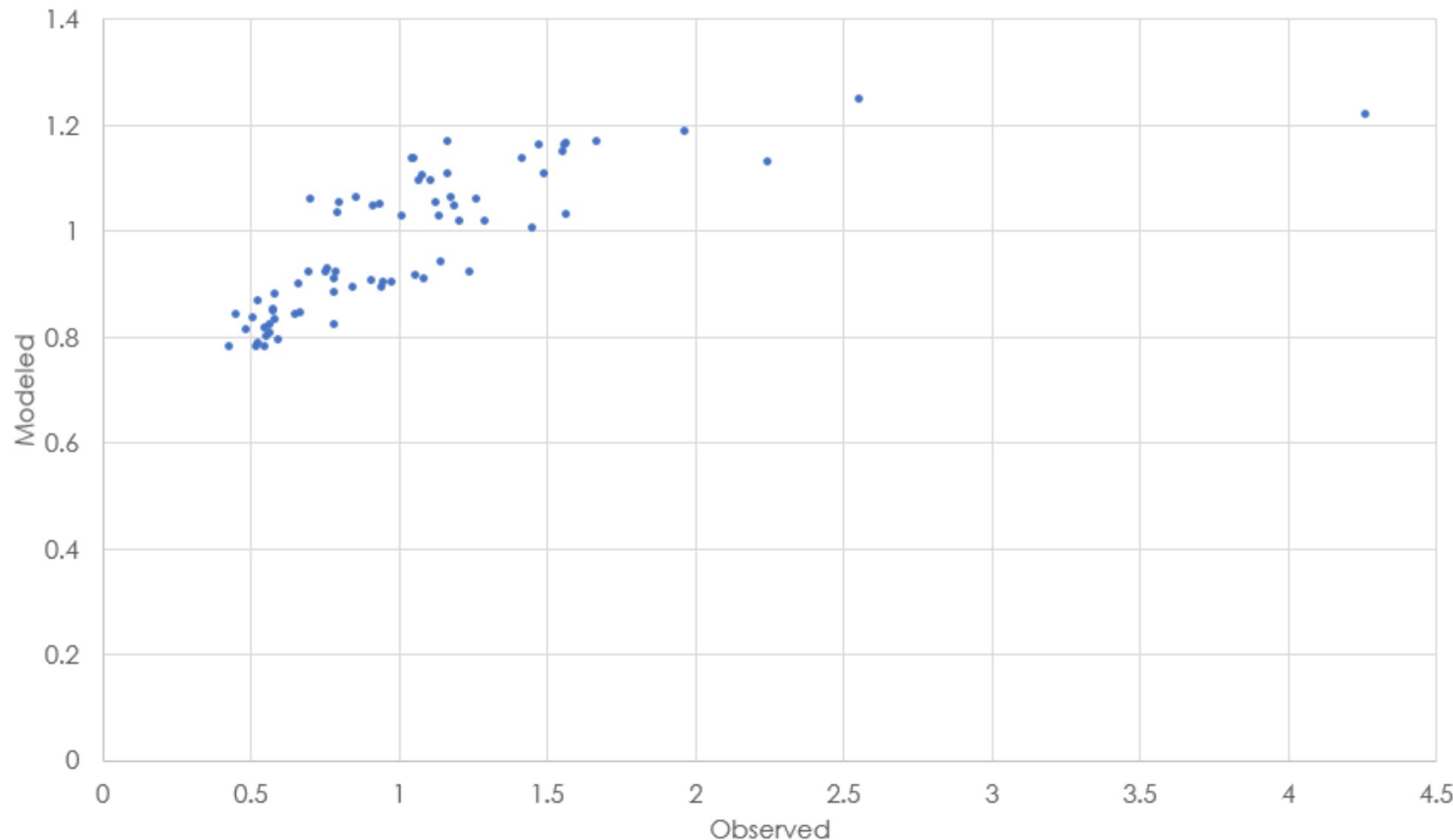
“

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

## Observed vs. Modeled Pure Premium by ZIP Code

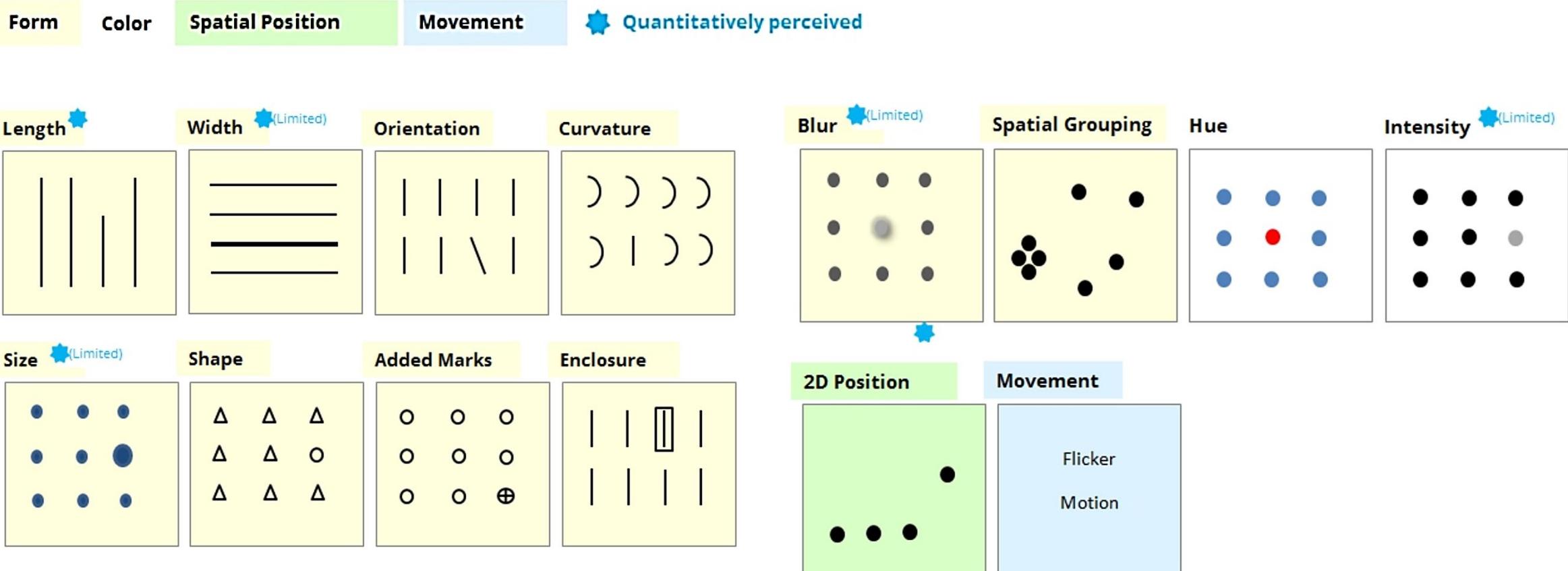


Observed vs. Modeled Pure Premium by ZIP Code

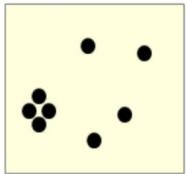


# PRE-ATTENTIVE ATTRIBUTES

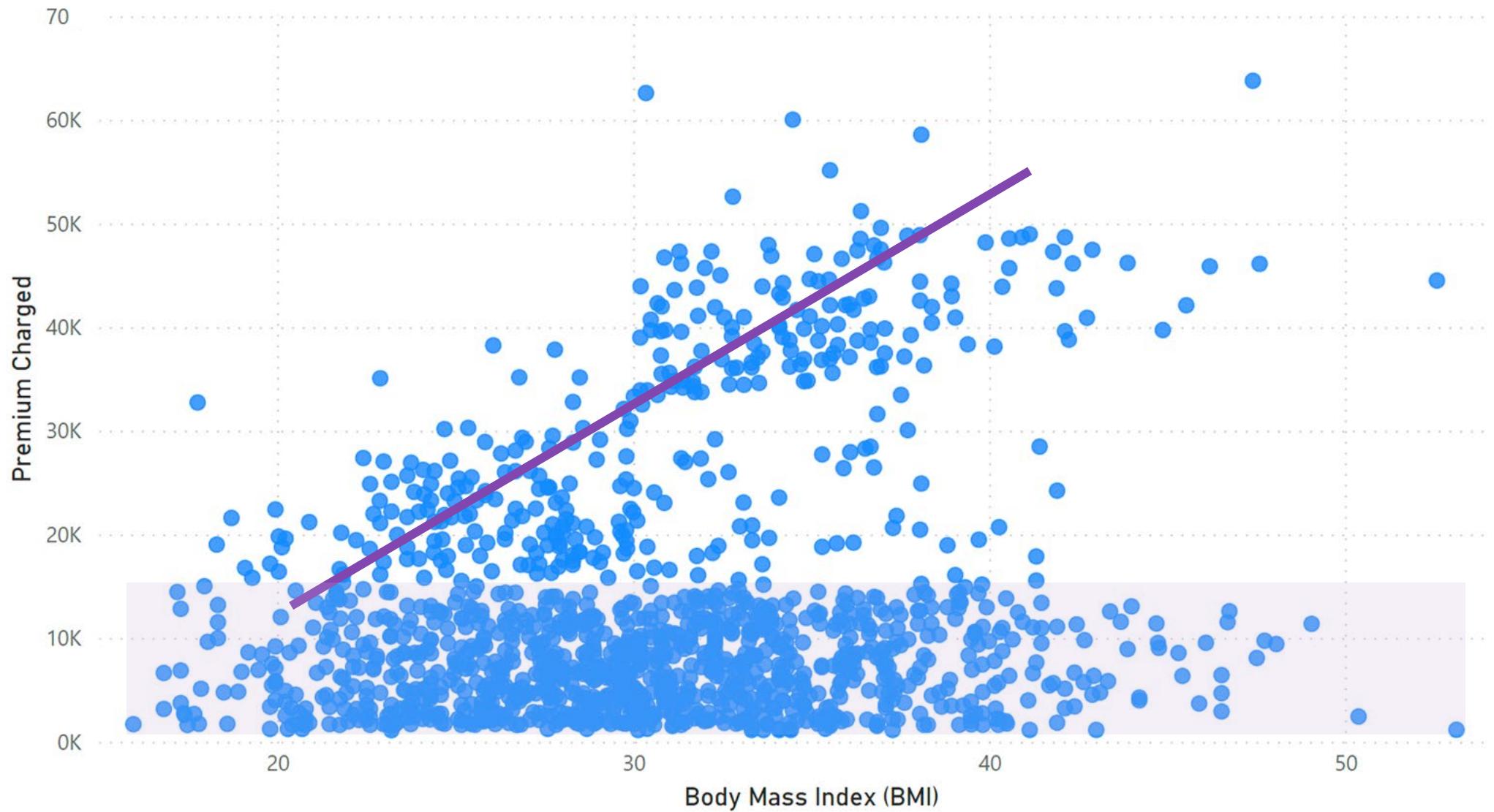
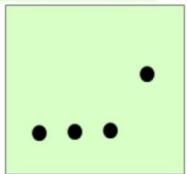
# TOOLS FOR COMMUNICATING VISUALLY

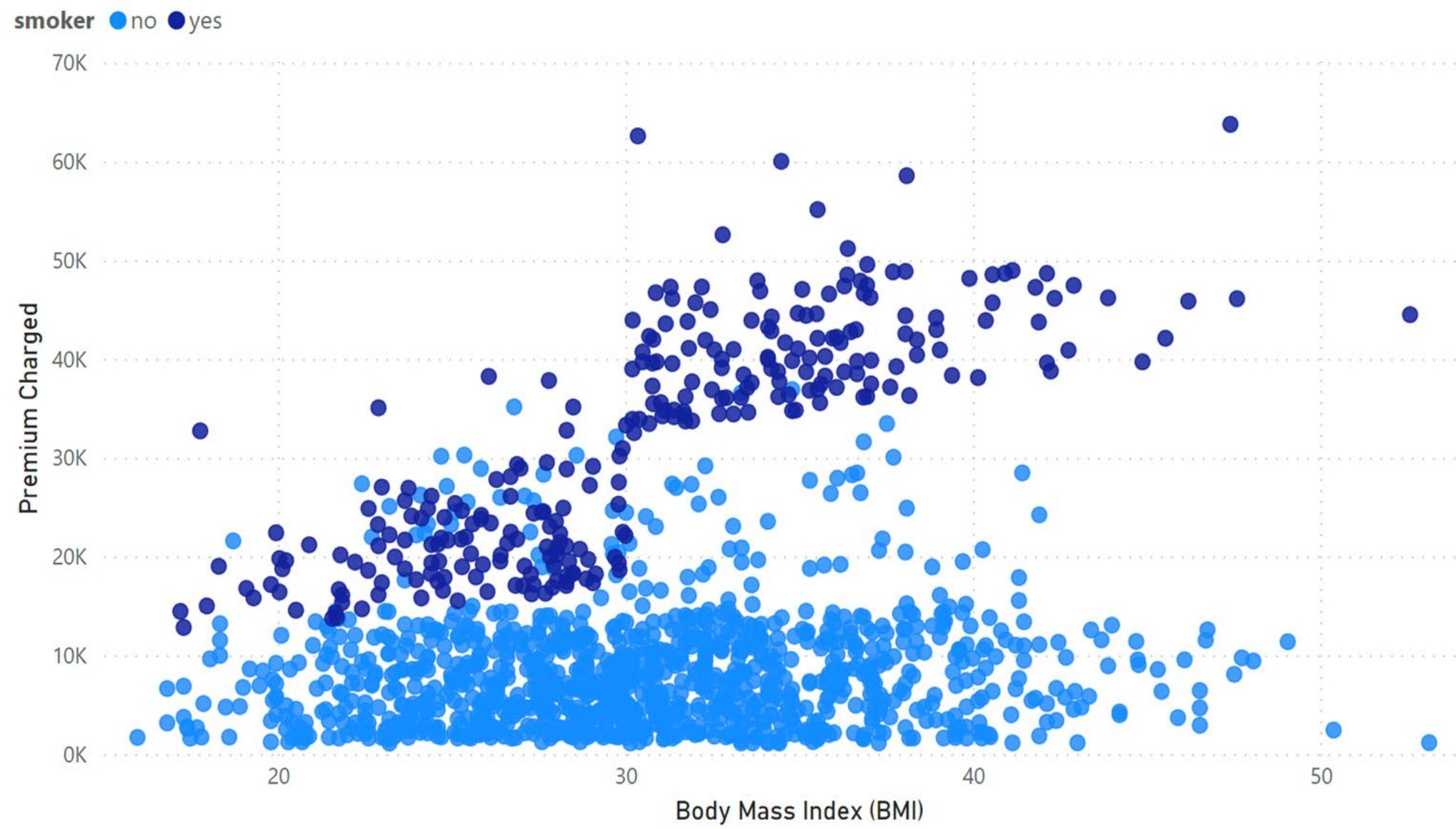
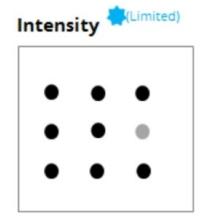
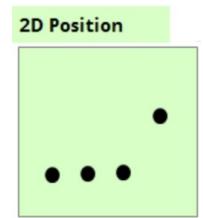
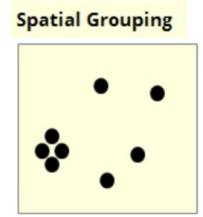


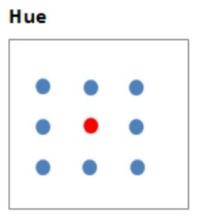
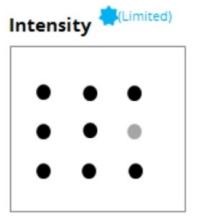
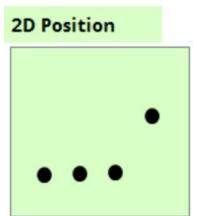
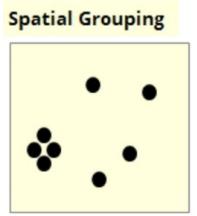
Spatial Grouping



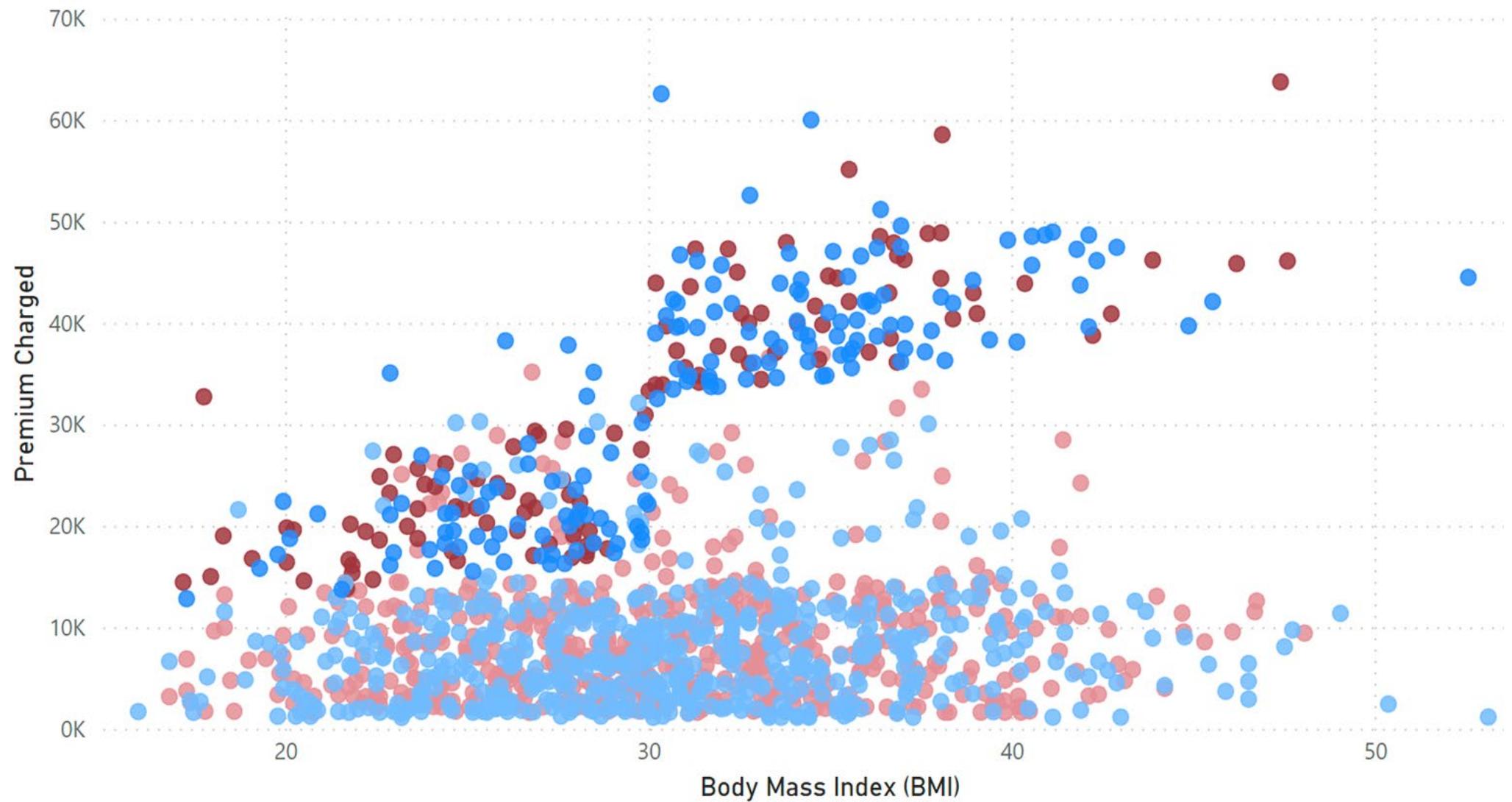
2D Position

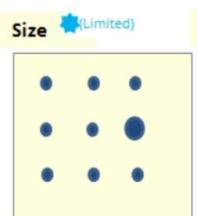
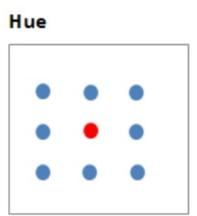
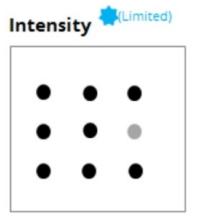
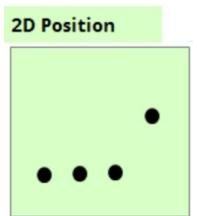
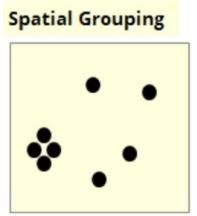




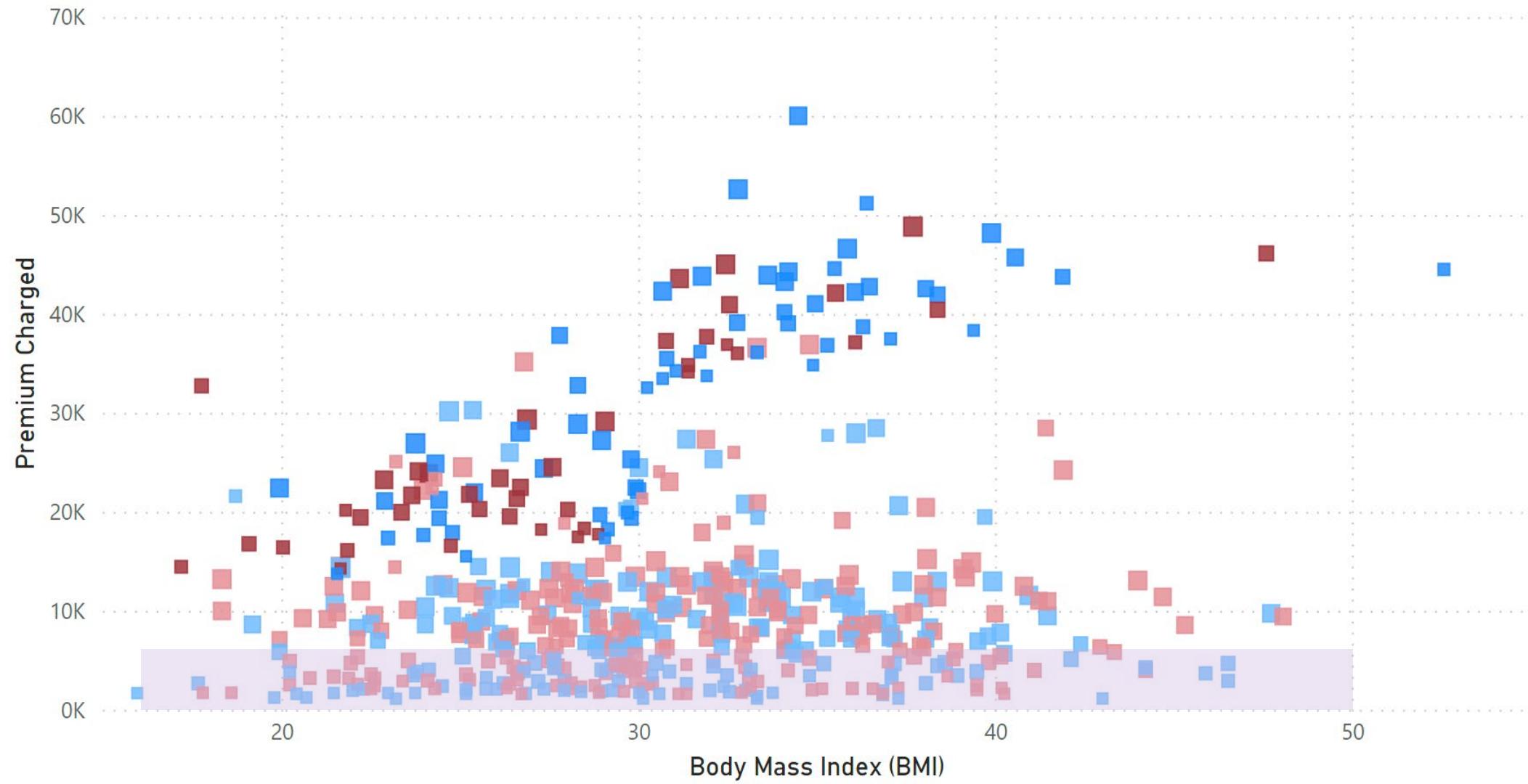


**Sex & Smoker** ● female-no ● female-yes ● male-no ● male-yes





**Sex & Smoker** ■ female-no ■ female-yes ■ male-no ■ male-yes

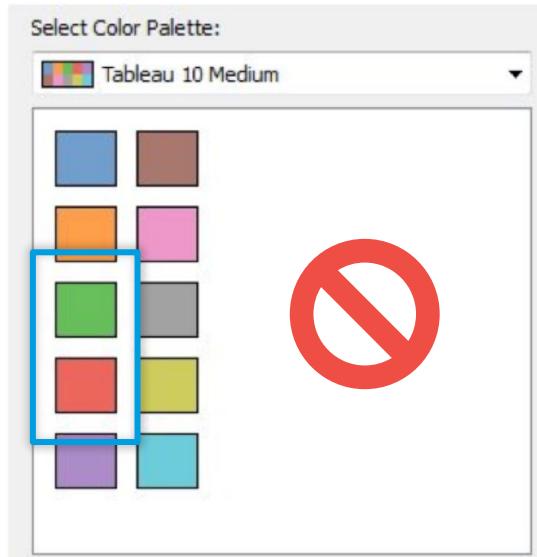


Size indicates Age (18 – 64)

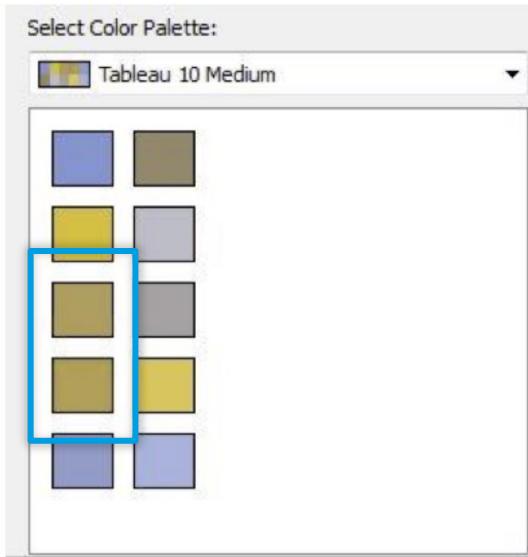
# THINGS TO AVOID

## Red/Green Deutanopia (6% of males)

Original Image

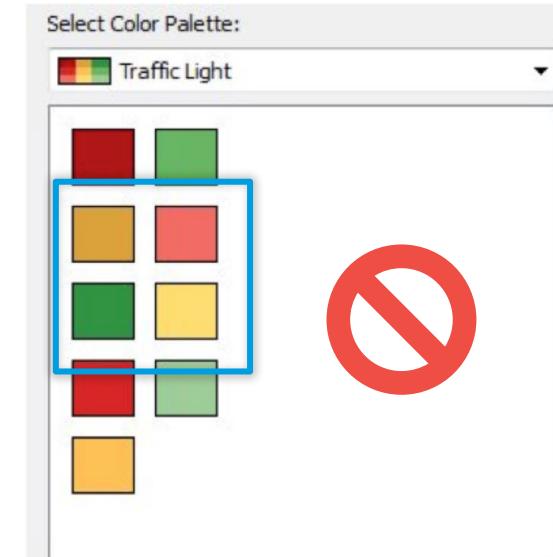


Deutanope Simulation

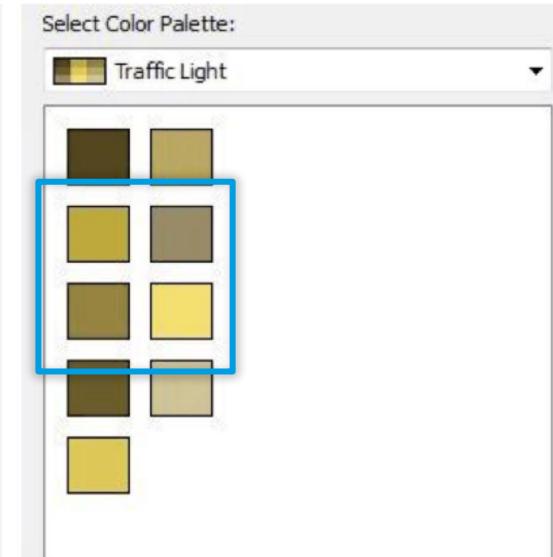


## Red/Green Protanopia (2% of males)

Original Image

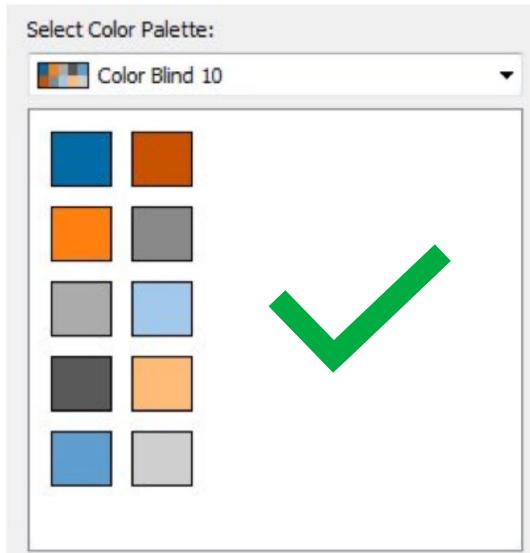


Protanope Simulation

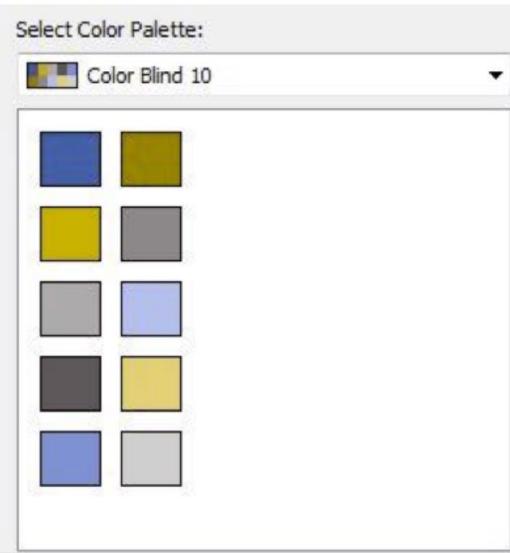


## Red/Green Deutanopia (6% of males)

Original Image

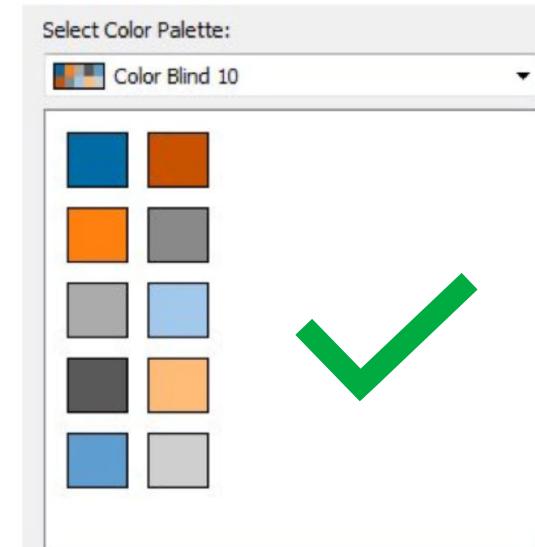


Deutanope Simulation

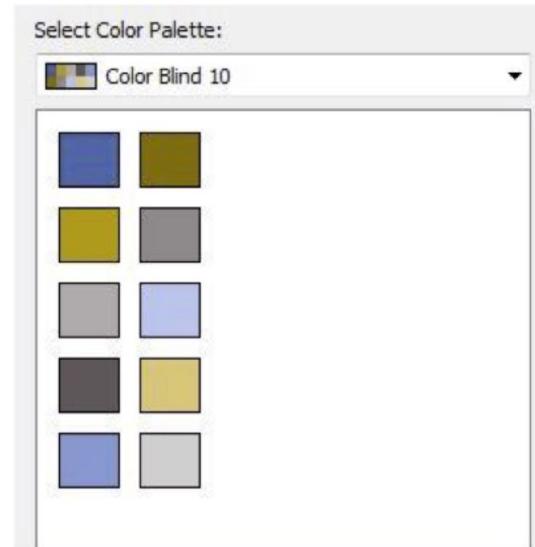


## Red/Green Protanopia (2% of males)

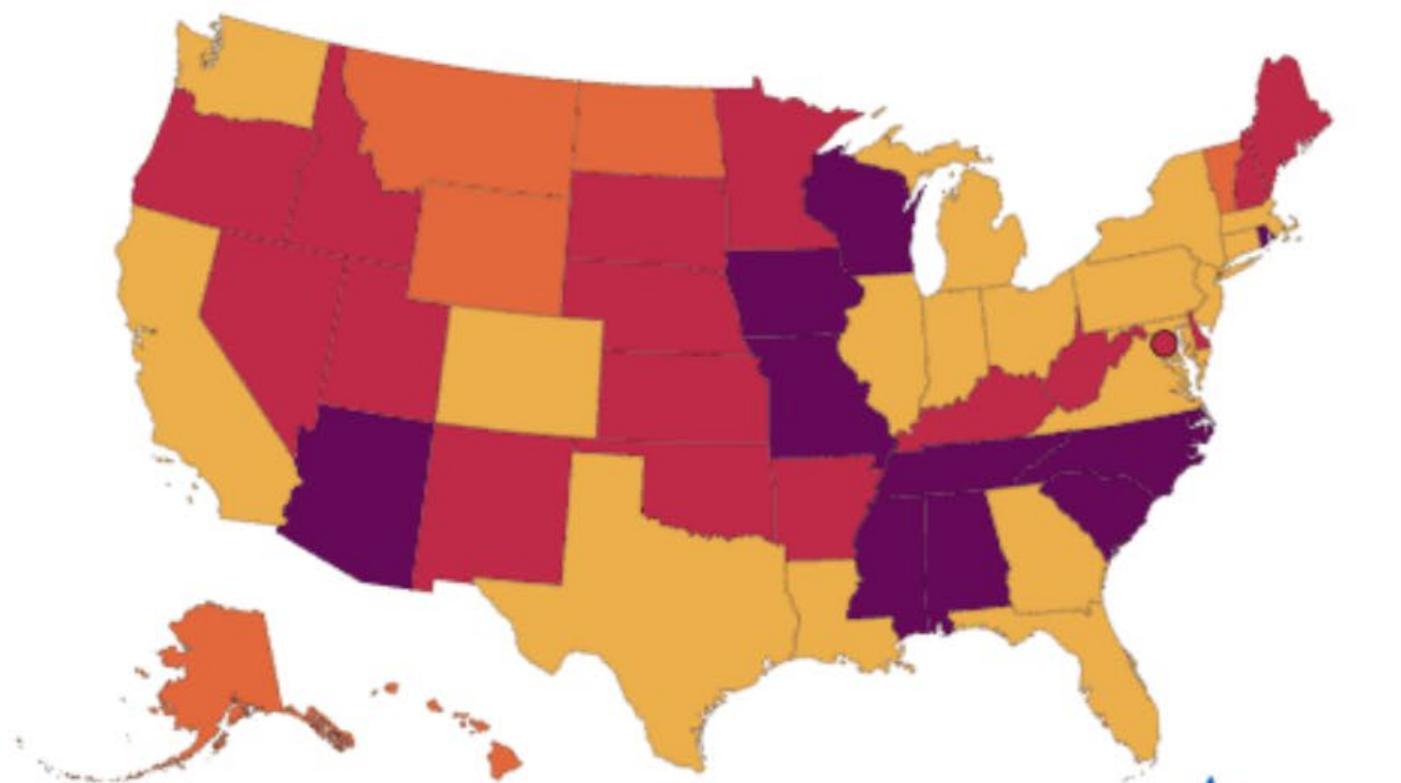
Original Image



Protanope Simulation



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



### Reported Cases

- |                  |                   |
|------------------|-------------------|
| ● 1 to 100       | ● 10,001 or more  |
| ● None           | ● 101 to 1,000    |
| ● 1,001 to 5,000 | ● 5,001 to 10,000 |



Graphics that  
are accurate  
but misleading

Baseline should  
start at zero,

page 77

## SHOULD BRITAIN LEAVE EU?



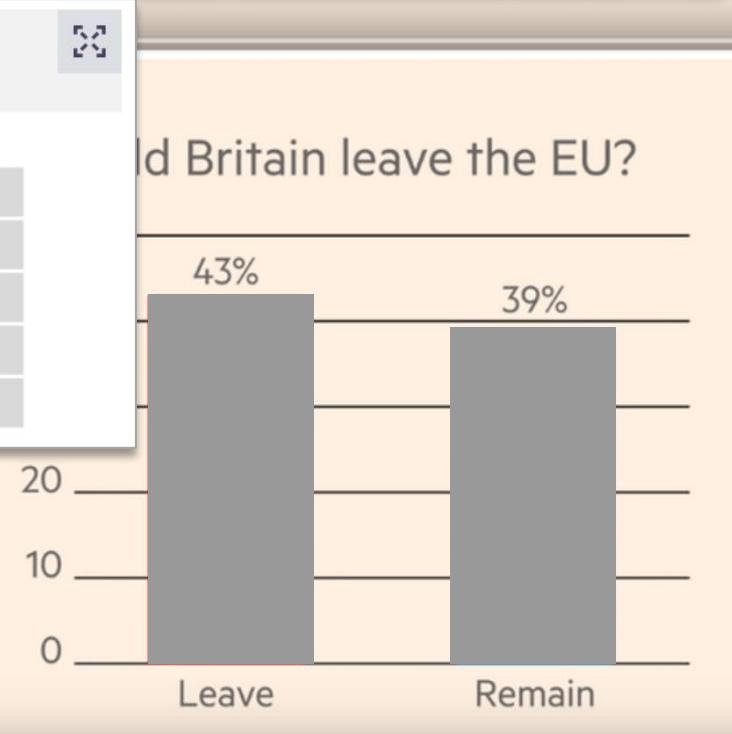
## How Britain voted

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

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

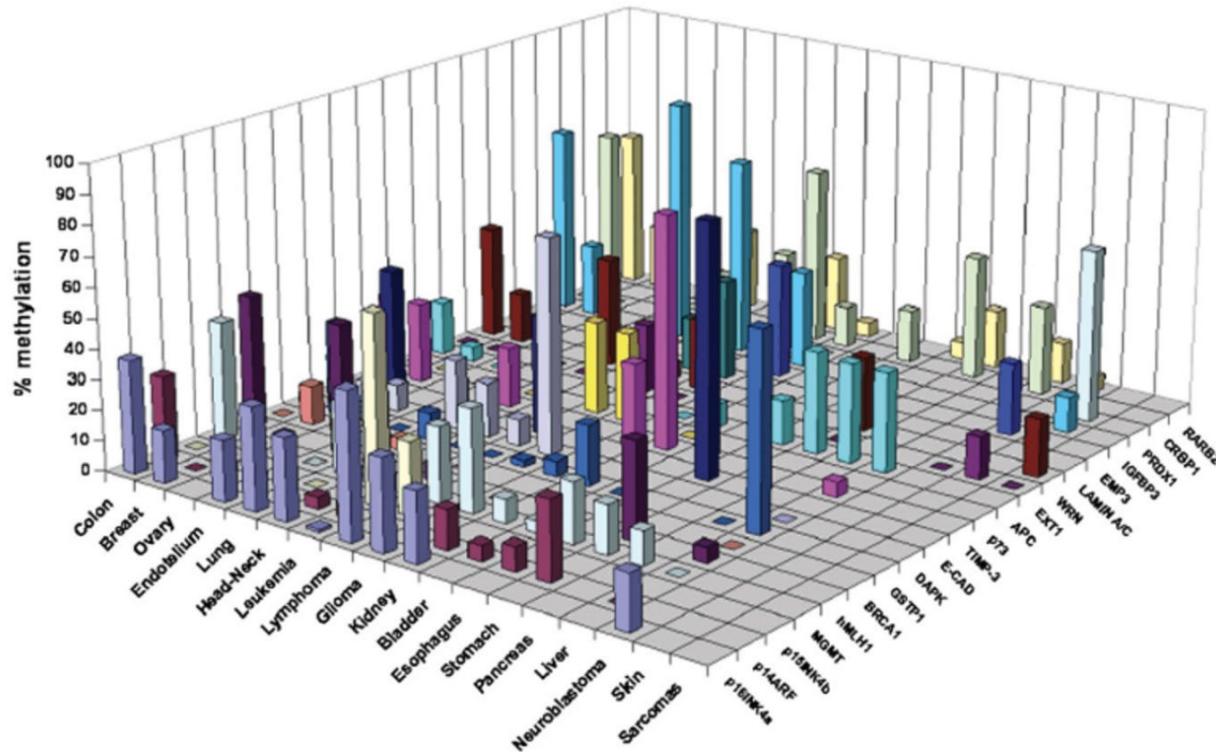
are accurate  
but misleading

A better chart of  
the same data





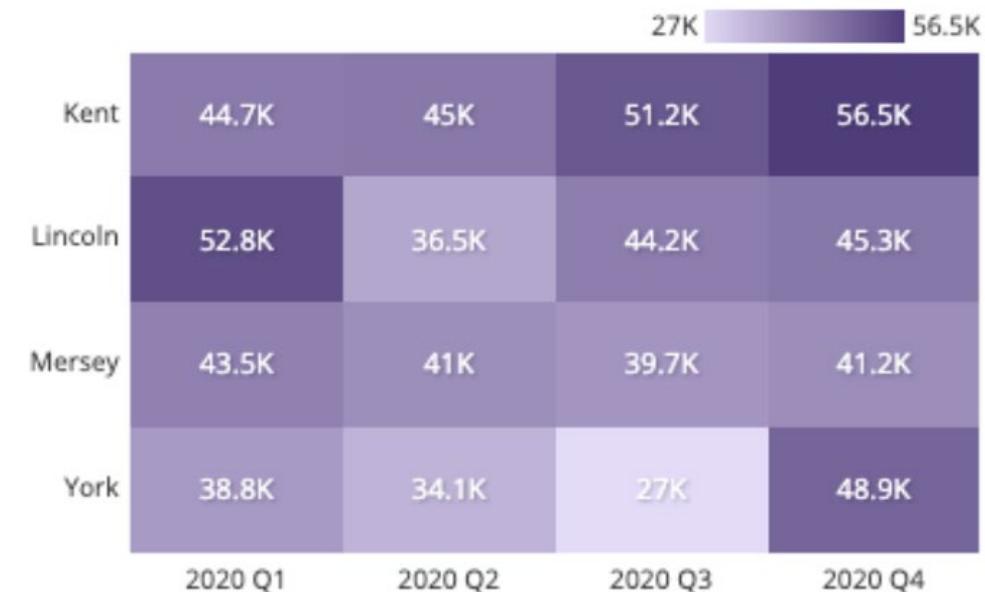
## A CpG Island Hypermethylation Profile of Human Cancer



Hum. Mol. Genet. (2007) 16:R50-59



## New Revenue



# PROBLEMS & TOOLS TO FIX THEM

# I DON'T HAVE ANY DATA

Good, public data can be difficult to find – even in this modern age!

Here are some places you can start looking:

- <https://www.kaggle.com/datasets> - data sets linked to code and analysis
- <https://paperswithcode.com/datasets>
- <https://datasetsearch.research.google.com/>
- <https://archive.ics.uci.edu/datasets>
- <https://www.reddit.com/r/datasets/>
- <https://aws.amazon.com/marketplace/>
- <https://github.com/fivethirtyeight/data>
- College/university library may have data subscriptions available.
- Government (national):
  - <https://data.gov/>
  - <https://www.census.gov/data.html>
  - <https://www.census.gov/programs-surveys/susb.html> counts of businesses and employees by size, location, industry
- Government (local): cities will often have data portals, too! For example: <https://data.cityofchicago.org/>
- Publicly traded companies:
  - <https://www.sec.gov/edgar/search-and-access> for lists of companies, SEC financial disclosures
  - <https://www.simfin.com/en/fundamental-data-download/> for clean company datasets

# 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.

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

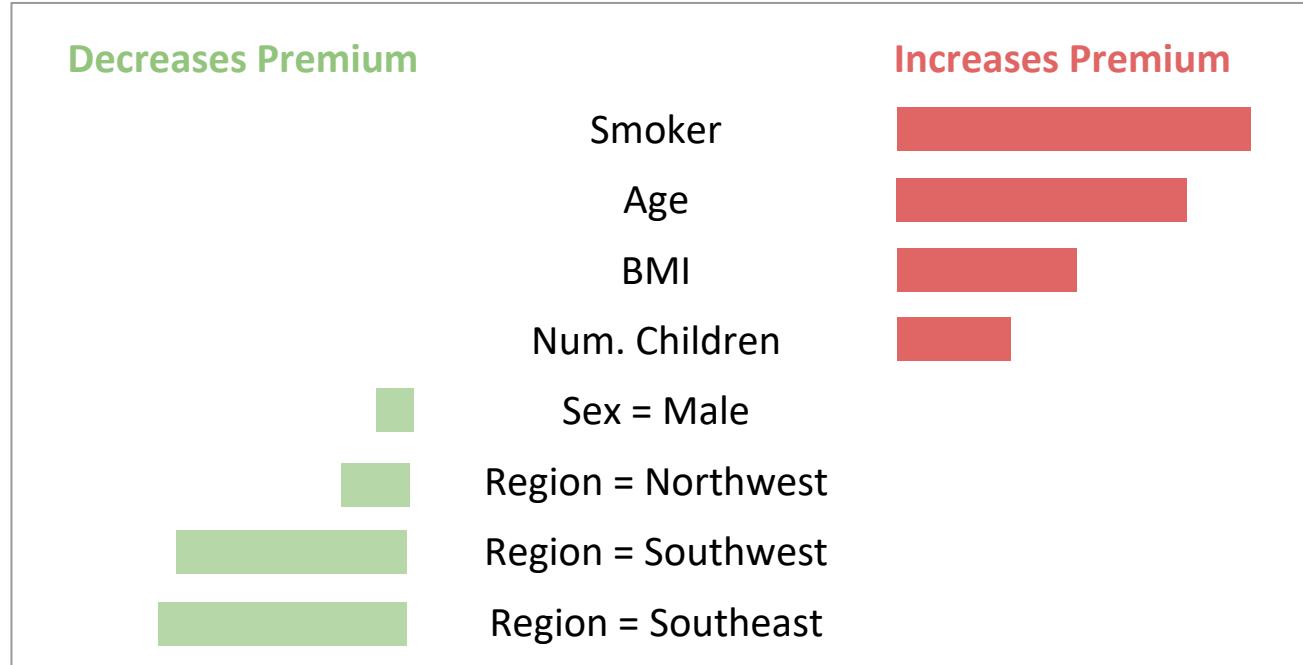
[1] "charges"

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
```



# 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.

**Deviation**  
Emphasize volatility or how far a figure has moved from its mean. Good for showing the range of a single variable or a long-term range. Can also highlight the points of interest.

**Correlation**  
Show the relationship between two variables. Good for showing the relationship between two variables, or how one variable will move in relation to the other.

**Ranking**  
Use when there are three or more variables. Good for ranking items in order of importance or size.

**Distribution**  
Show a dataset and often an underlying distribution. Good for highlighting the points of interest.

**Change over Time**  
Give emphasis to changing trends. Good for showing the evolution of a variable over time.

**Magnitude**  
Show size comparisons. These can be absolute, relative or proportional. Choosing the right scale is important to provide suitable context for the data.

**Part-to-whole**  
Show how a single entity can be broken down into smaller parts. Good for showing the differences, usually there is a whole and a part, and the whole is usually greater than a part.

**Spatial**  
Aside from location, it's also good for showing the relationship between objects in space or time. Good for showing the location of consumers for anything.

**Flow**  
Show the reader volume or intensity of something moving from one place to another. Good for showing the sequence or geographical routes.

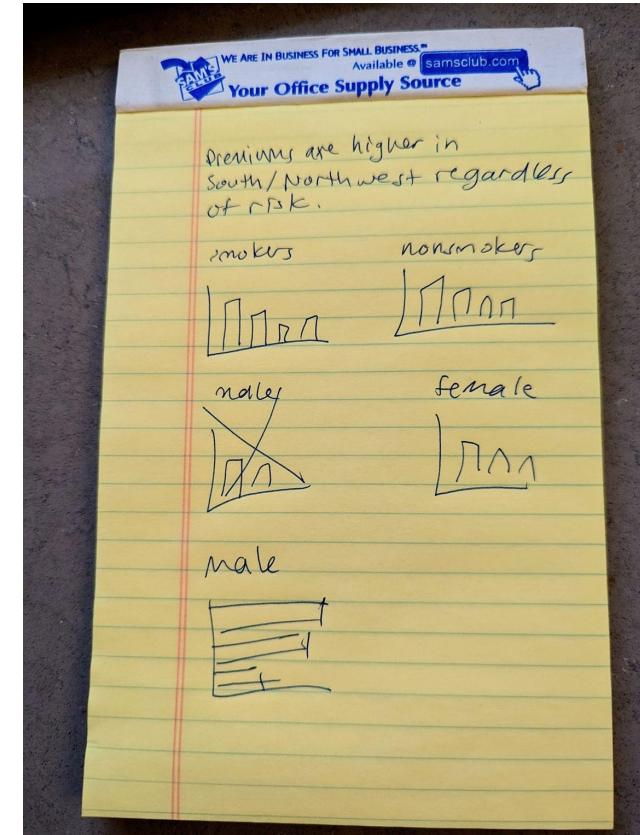
**Visual vocabulary**

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to decide which data dimension is most important to show. You may find that different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

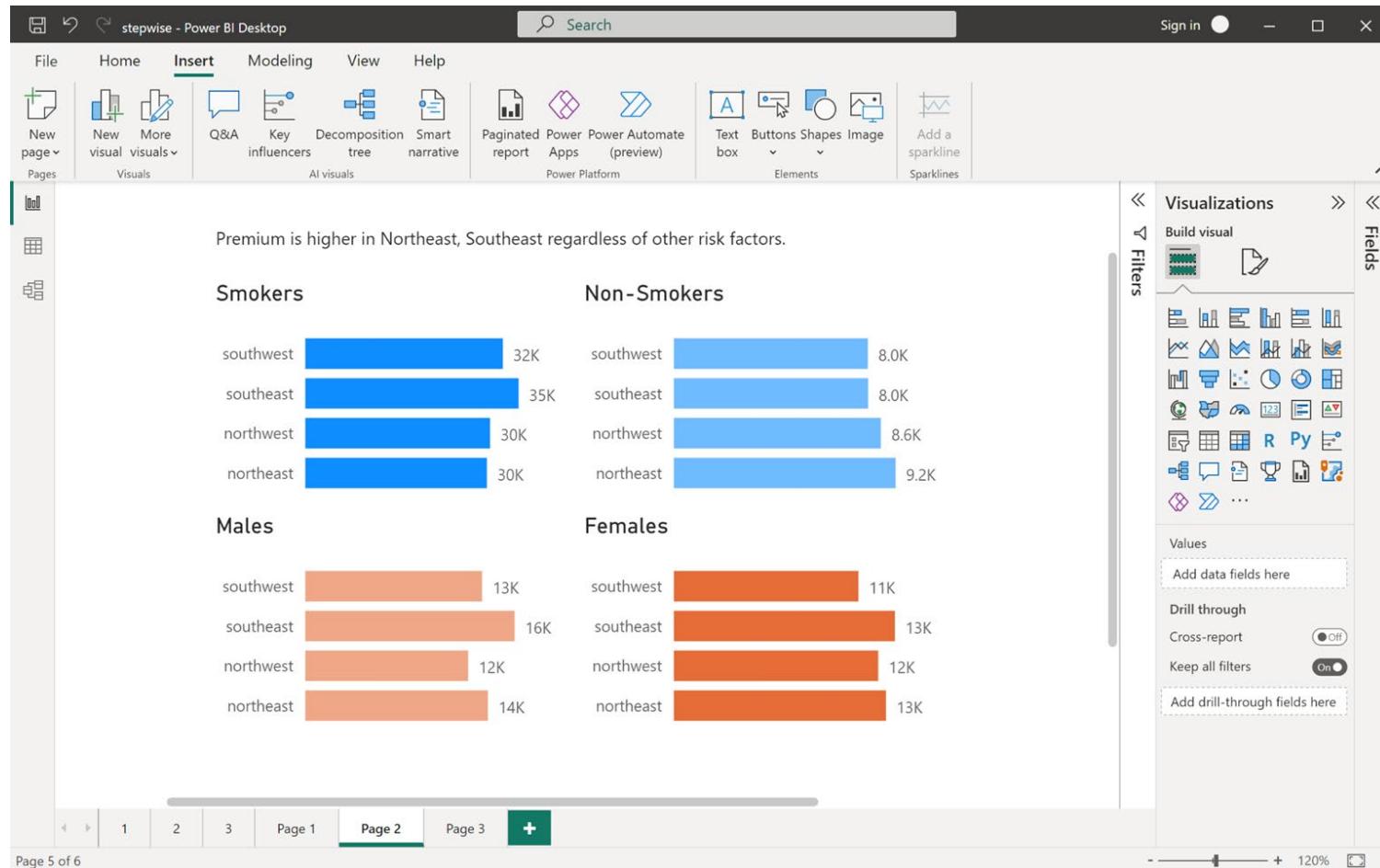
FT graphic: Alex Smith. Chris Collard & Ruth Johnson. Design: Alex Sherriff. Graphic: Paul Pichler. Photo: Steve

[ft.com/vocabulary](https://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.



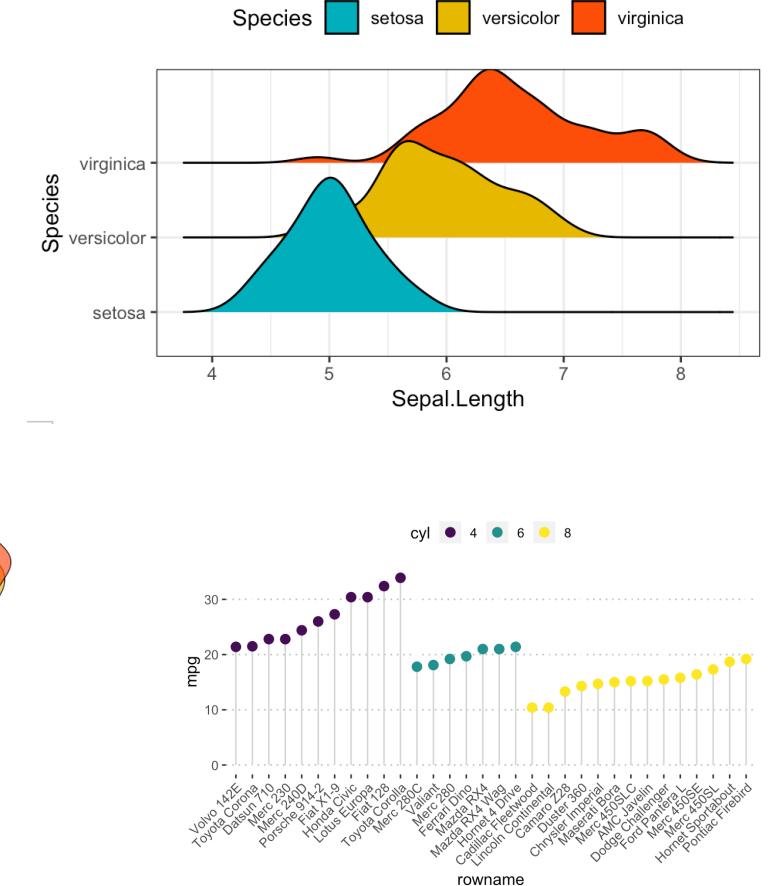
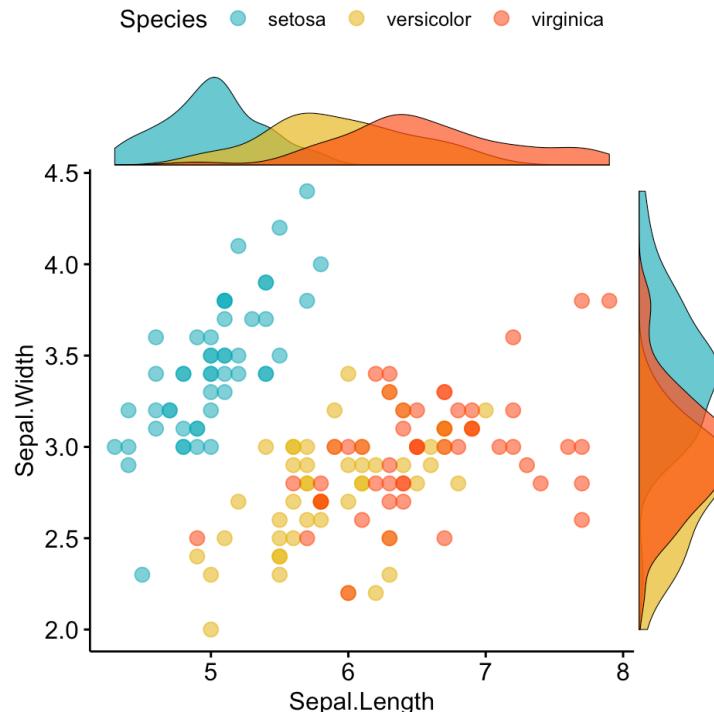
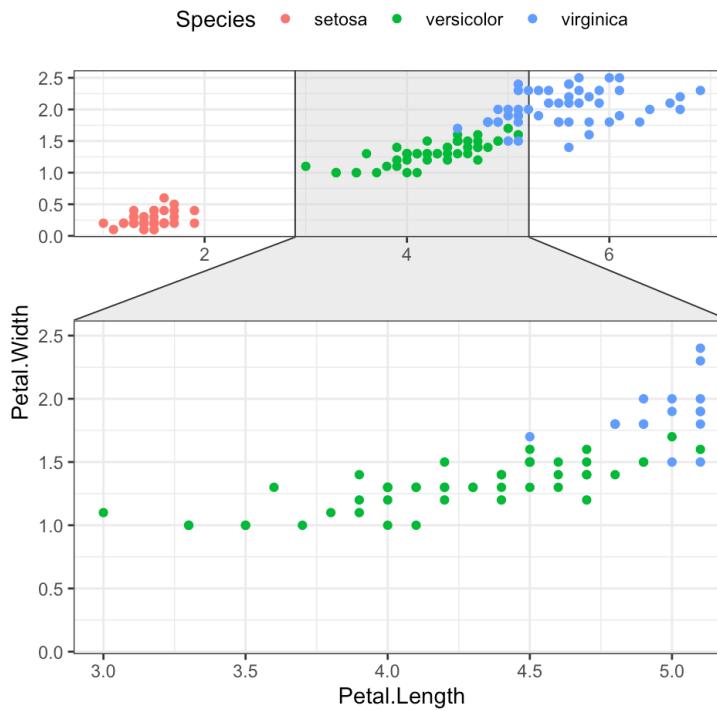
# I'M NOT HAPPY WITH THE DESIGN OPTIONS IN POWER BI

- 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).



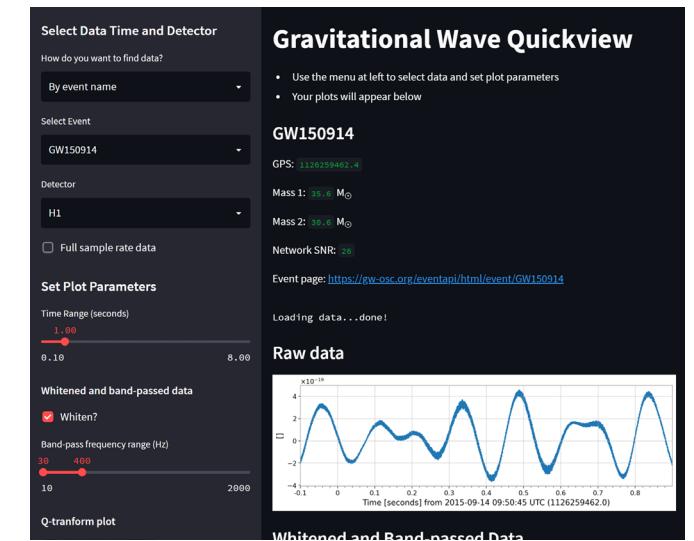
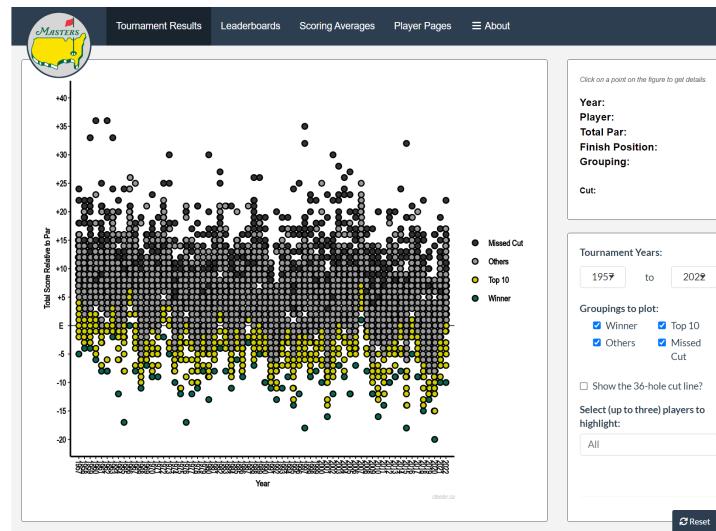
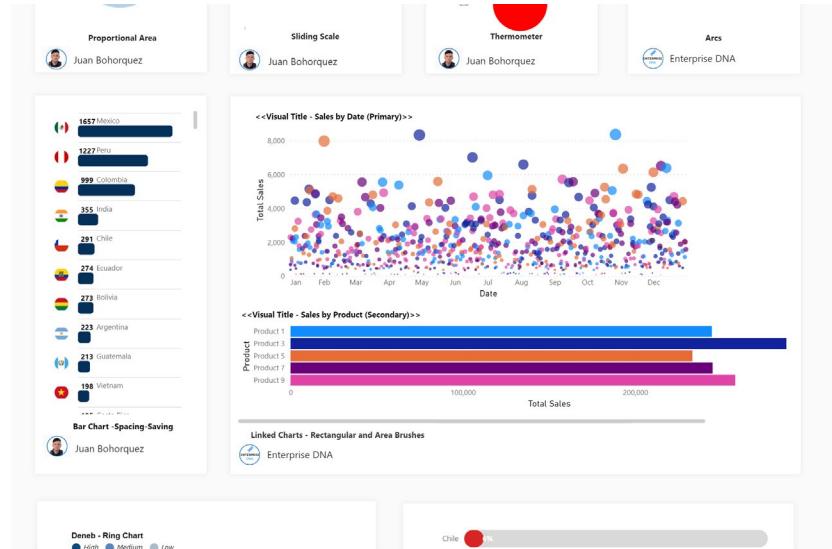
# I AM NOT HAPPY WITH CHART TYPE / NEED TO MAKE MANY SIMILAR CHARTS

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



# 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>

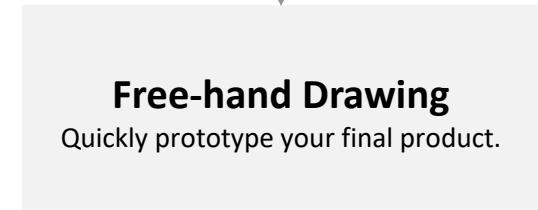
<https://gw-quickview.streamlit.app/>



## AutoML

Let the computer do the work for you.  
Rapidminer, Storyteller

Found your story?



## Business Intelligence

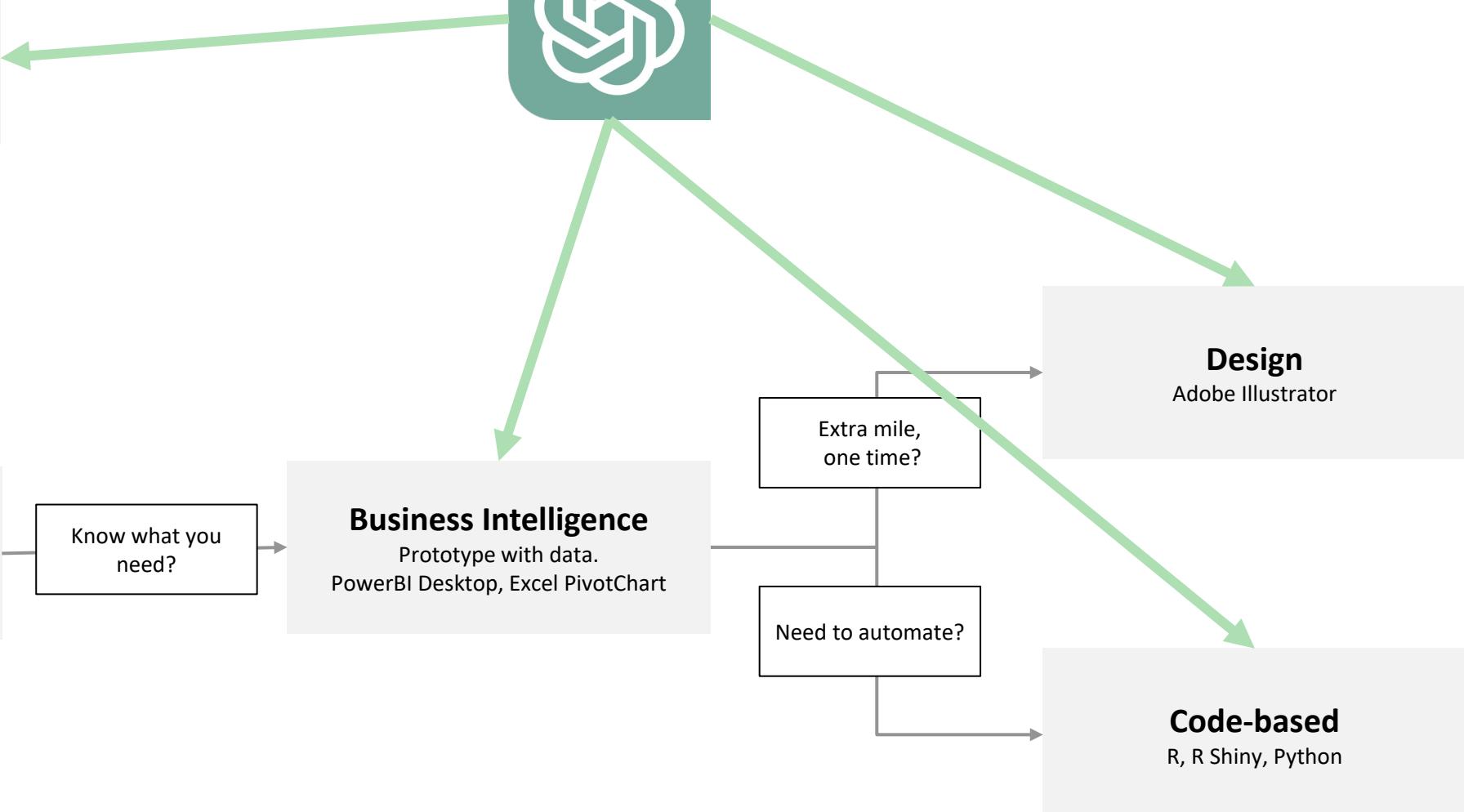
Prototype with data.  
PowerBI Desktop, Excel PivotChart

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.”

*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.

...”

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

# 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.*

Response: [Python code] ...

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

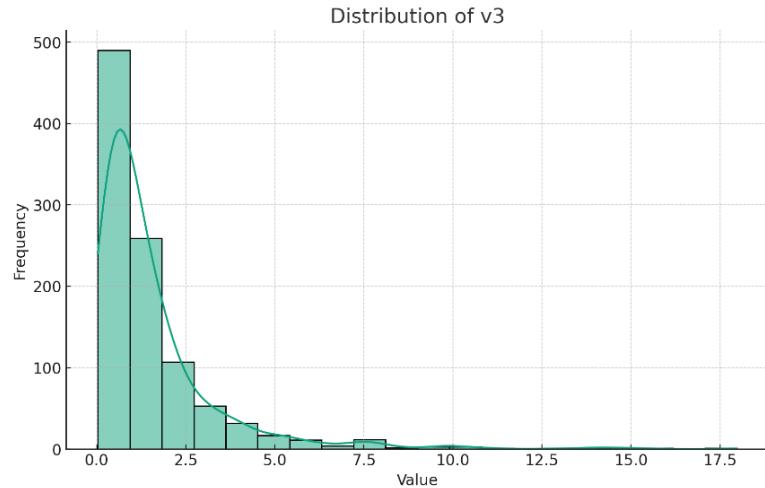
- V1: A one-unit increase in v1 is associated with an increase of approximately **9.99** in the target variable.
- V2: A one-unit increase in v2 is associated with an increase of approximately **300.00** in the target variable. This variable seems to have the most significant impact, consistent with 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.

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

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



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...

Prompt 4: please identify if it follows a log-normal, exponential, or gamma distribution.

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

meters of the selected

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]

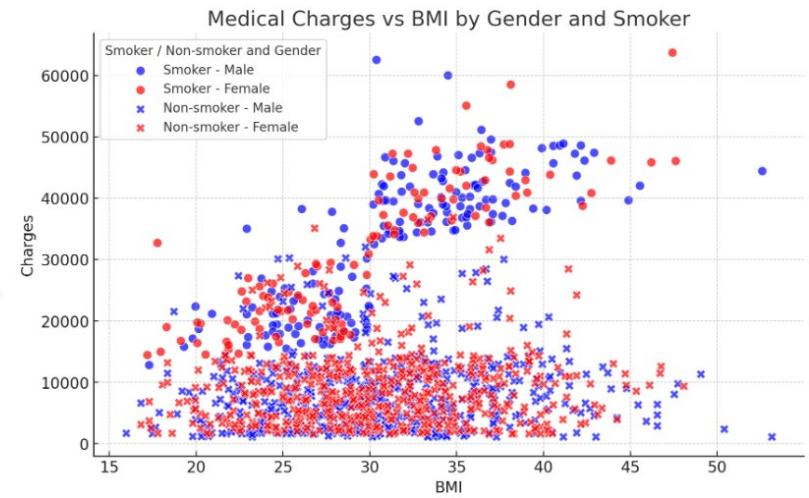
*Prompt 4: Please create one chart that shows the relationship between all these variables.*

Response:

[Python script]

We get some interpretation  
as a bonus.

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



"... 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.

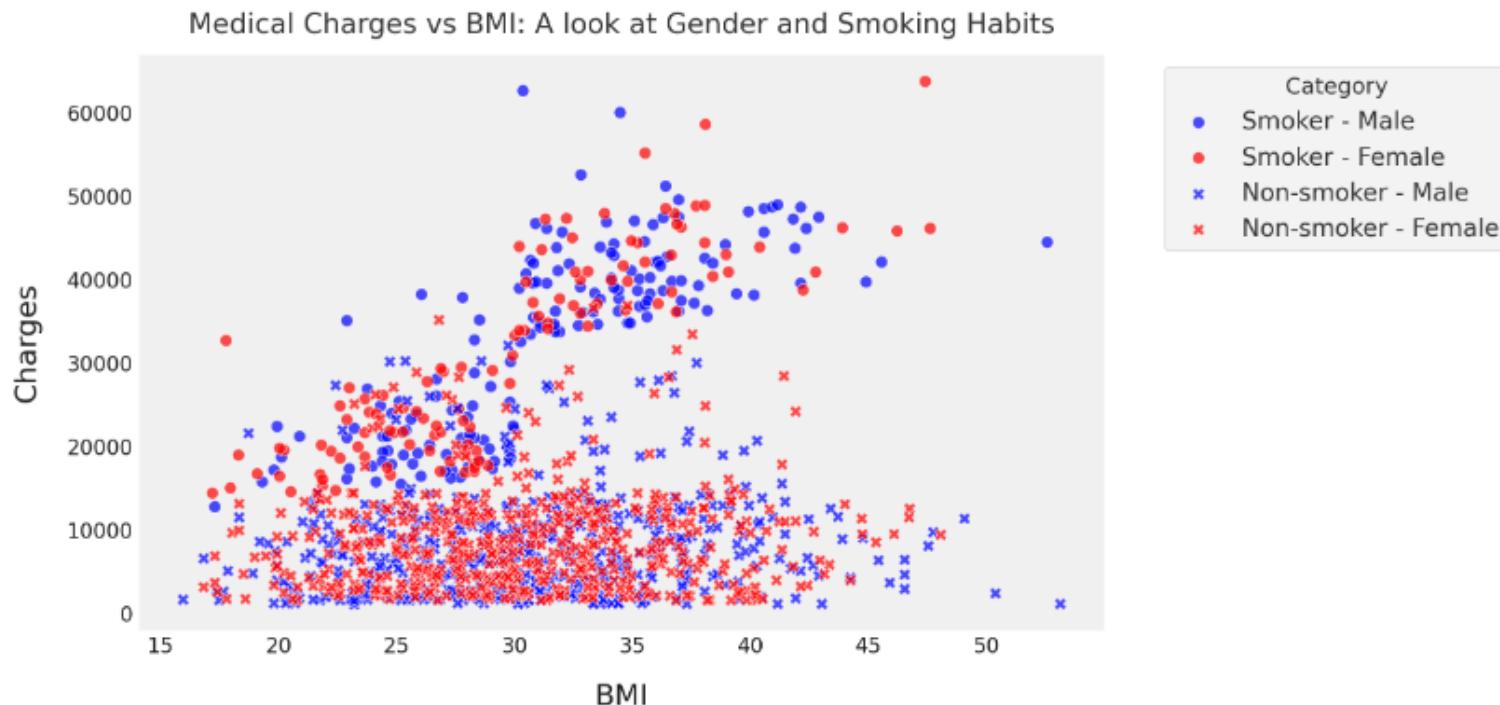
Both men and women follow the same general patterns in terms of the relationship between BMI, smoker status, and charges. However, it's worth noting that there are slightly more male smokers with high charges and high BMI compared to female smokers.

..."

# CHAT GPT: CREATE A VISUALIZATION 2

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

Response: [Python script]

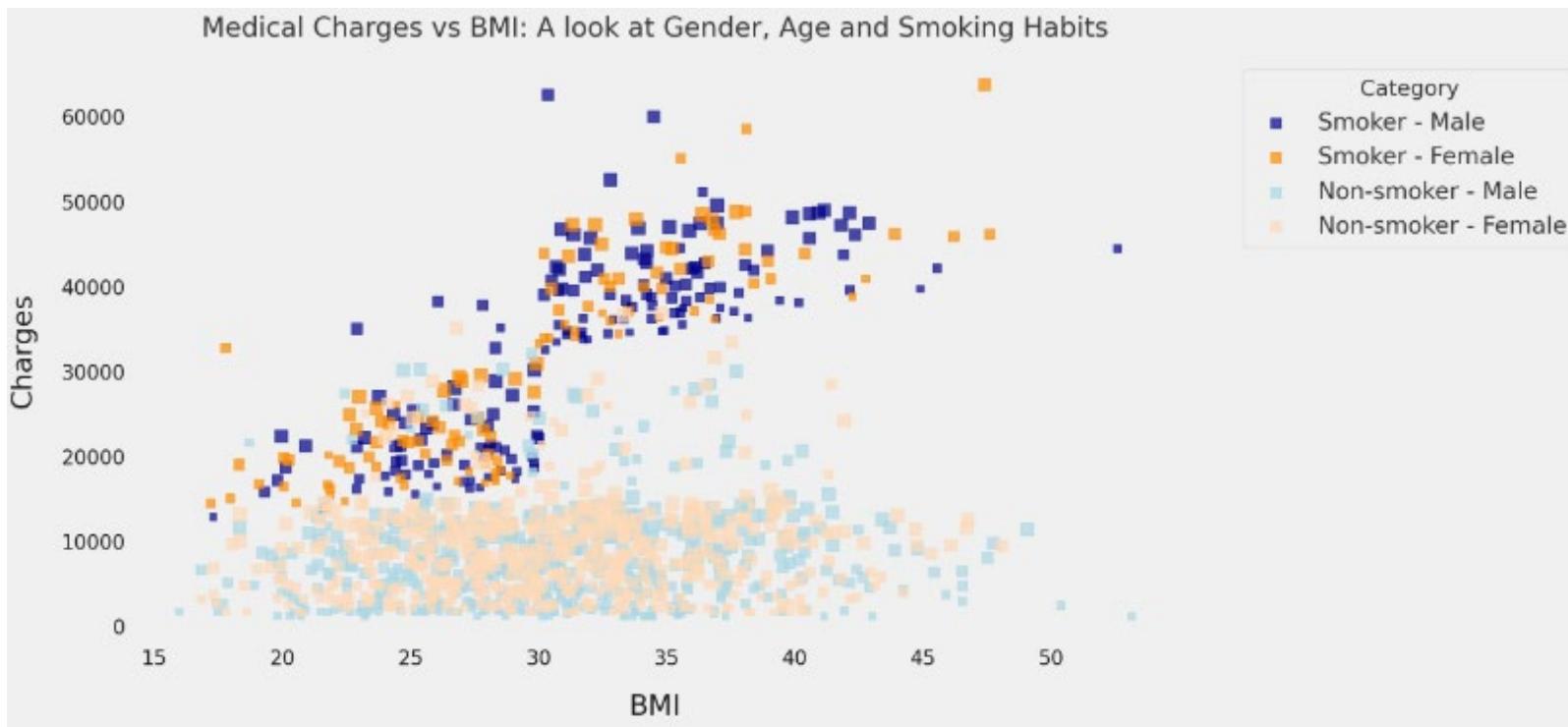


"Here's the updated, stylized chart suitable for a newspaper publication. It has a clear title, labeled axes, and a legend explaining the color and marker coding for the different categories of individuals (smokers/non-smokers and males/females). I have removed the grid lines and spines on the top and 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 their size indicates the age.*

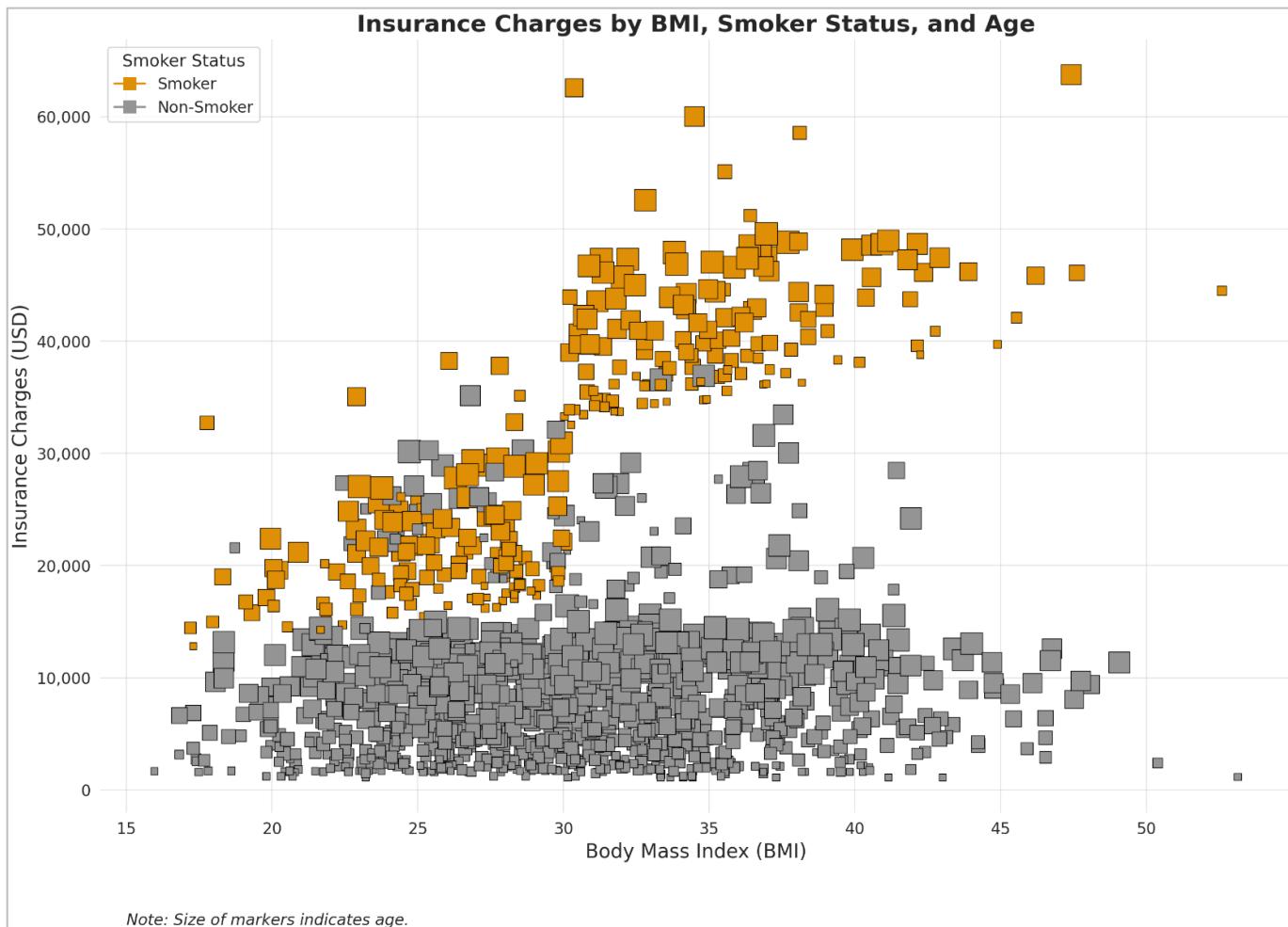
Response: [Python script]



“... 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/dalesabady](https://linkedin.com/in/dalesabady)



**Bryce Chamberlain, ASA, MSCA**  
Principal, Oliver Wyman Actuarial Consulting  
[linkedin.com/in/brycechamberlain](https://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*

R Markdown, R Shiny

Unlimited functionality.

Open source.

Git version control.

Time consuming.

Need to code.