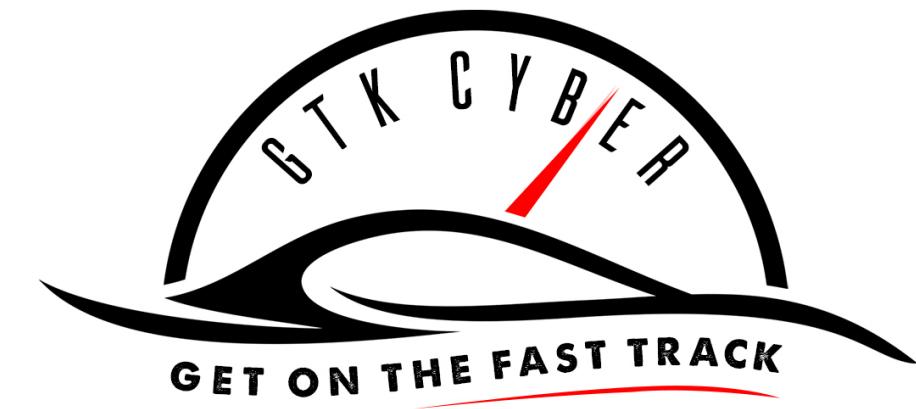




Module 4: Data Visualization

GET ON THE FAST TRACK

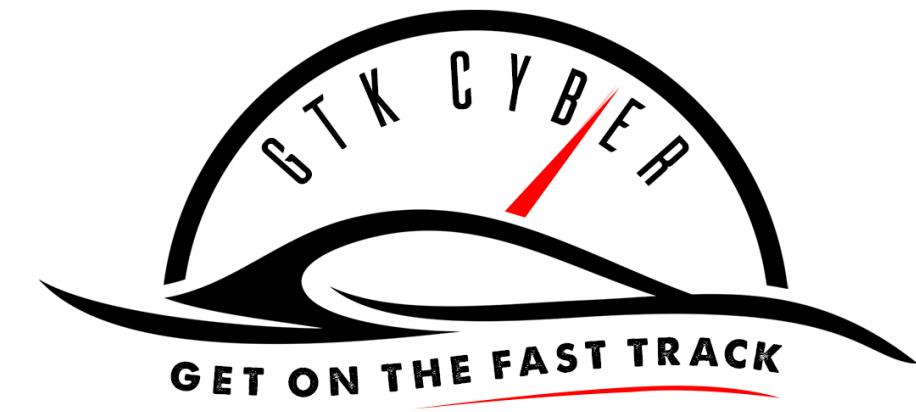


Our Lawyers Make Us Say This



All materials presented in this training and those provided as an adjunct to the program are copyrighted 2019 by GTK Cyber LLC.

They are intended solely for the use of registered program participants and may not be reproduced or redistributed in any manner for any other reason.



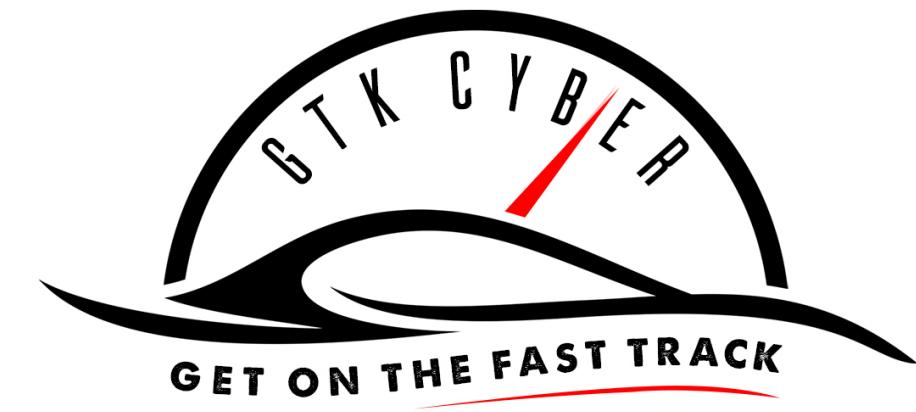
Exploring Data Through Visualization

Exploratory Visualizations

Explanatory Visualizations

Why Visualize Data?

Visualizing data can inspire you to ask new and more refined questions of your data and ultimately lead to better analysis.



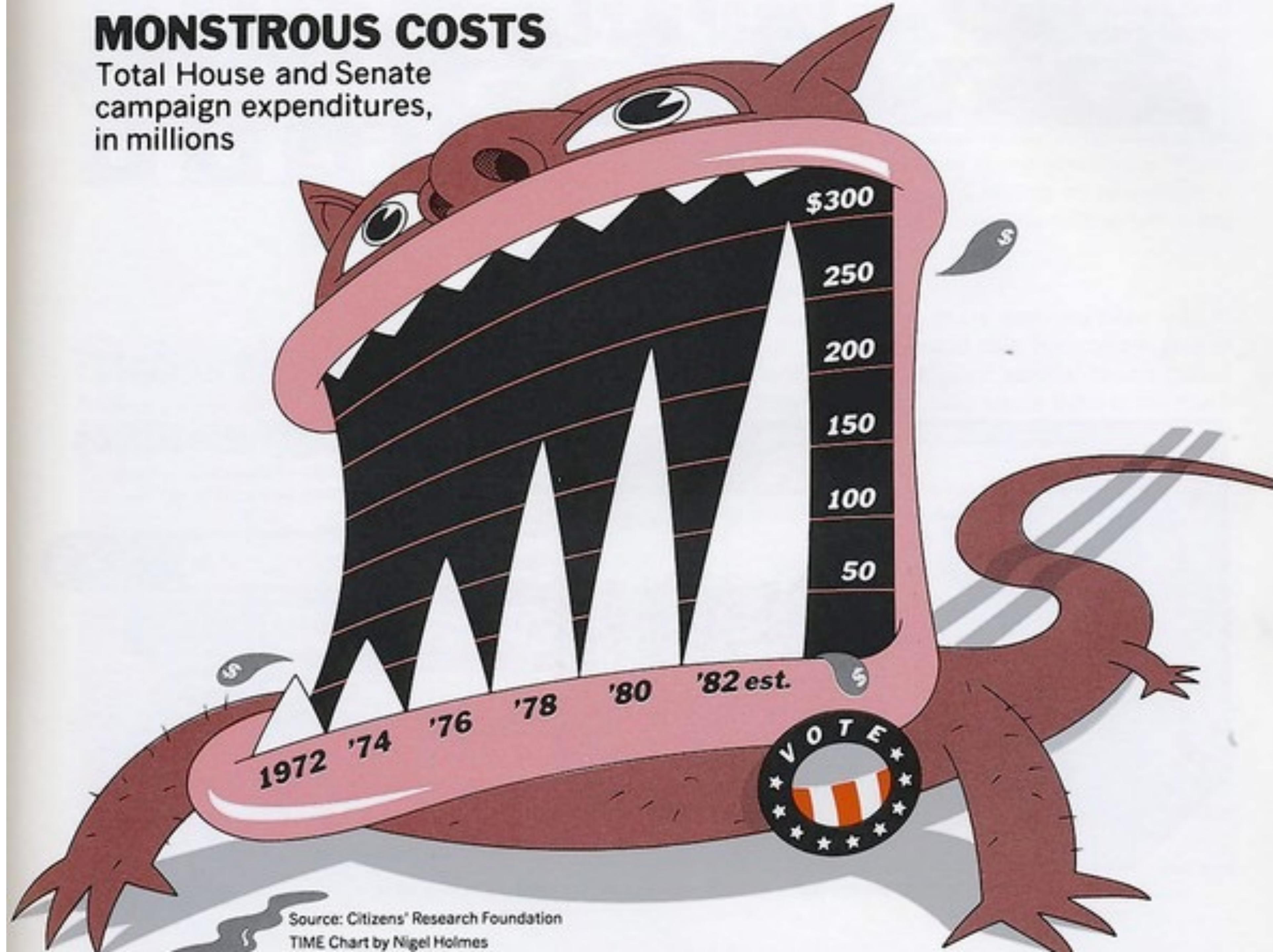
“The greatest value of a picture is when it forces us to notice what we never expected to see.”

–John Tukey, 1977

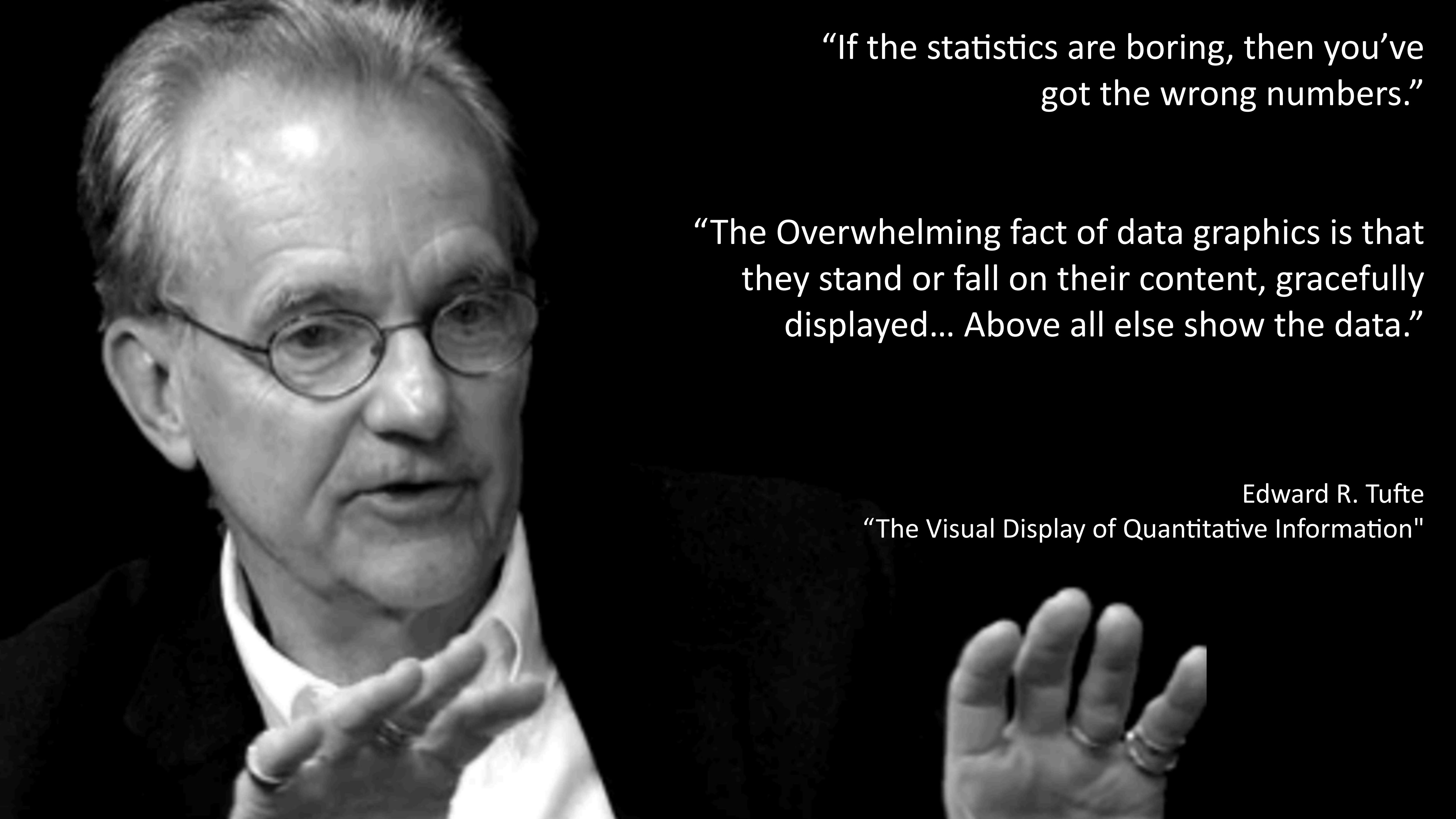
The power of visualization comes from
illustrating relationships, contrasts and
comparisons between many different dimensions
of data.

MONSTROUS COSTS

Total House and Senate campaign expenditures,
in millions



Source: Citizens' Research Foundation
TIME Chart by Nigel Holmes



“If the statistics are boring, then you’ve got the wrong numbers.”

“The Overwhelming fact of data graphics is that they stand or fall on their content, gracefully displayed... Above all else show the data.”

Edward R. Tufte

“The Visual Display of Quantitative Information”

Remove
to improve
(the **data-ink** ratio)

Show Comparisons,
Contrasts and Differences

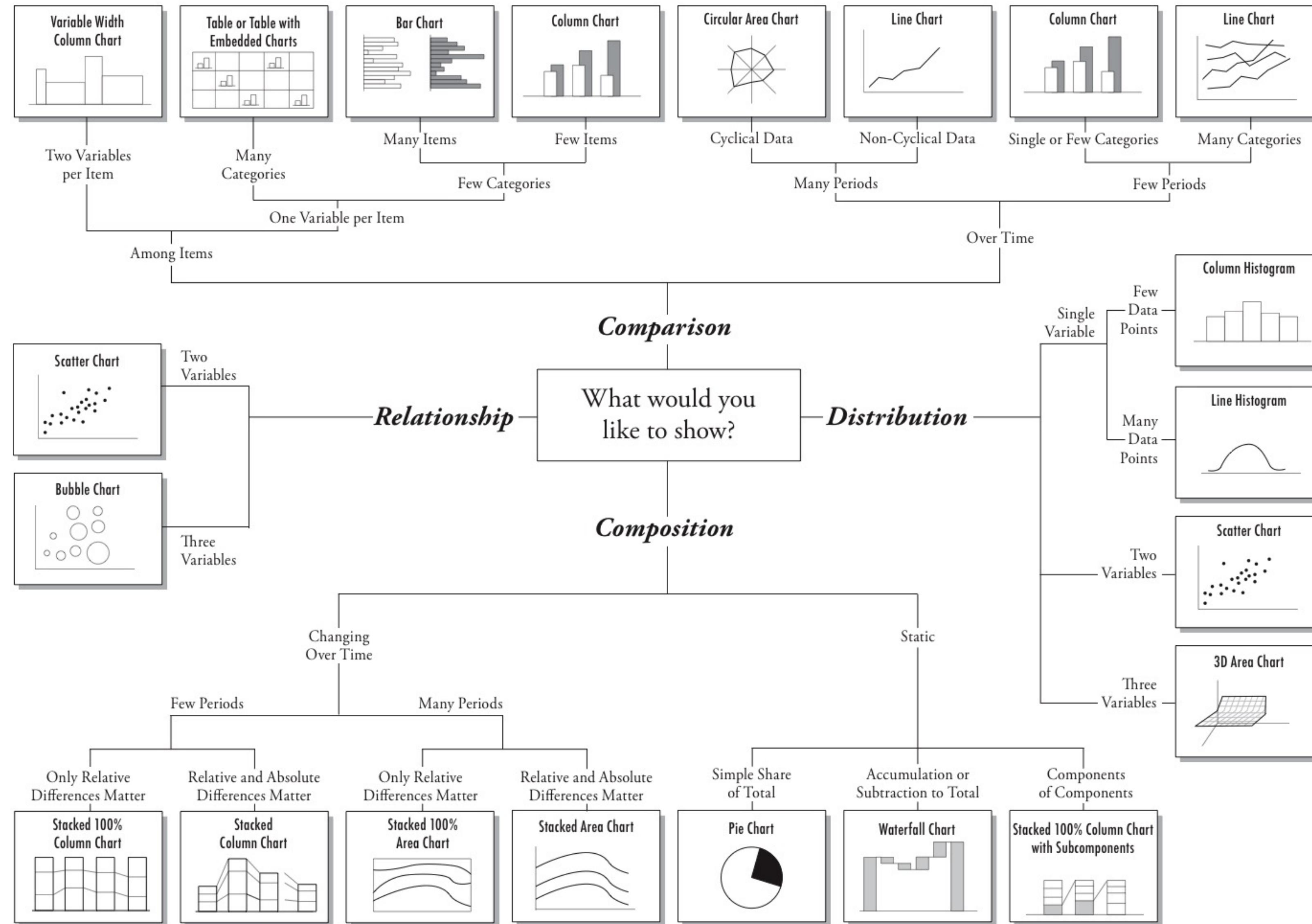
Show Multivariate Data

Integrate words, numbers,
images and diagrams

Document your
Evidence

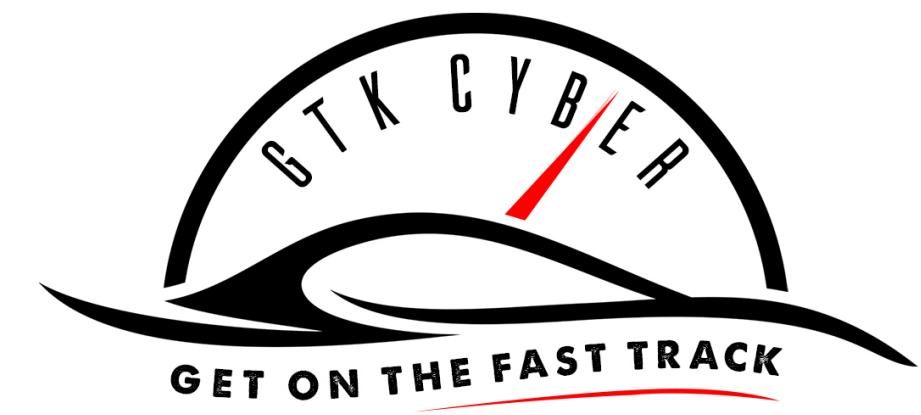
Ultimately, the quality,
relevance and integrity of the
content is most important.

Chart Suggestions—A Thought-Starter



Visualization Goals

- Analyze
- Explore
- Assess
- Determine
- Decide
- Communicate
- Explain
- Present
- Prove
- Persuade



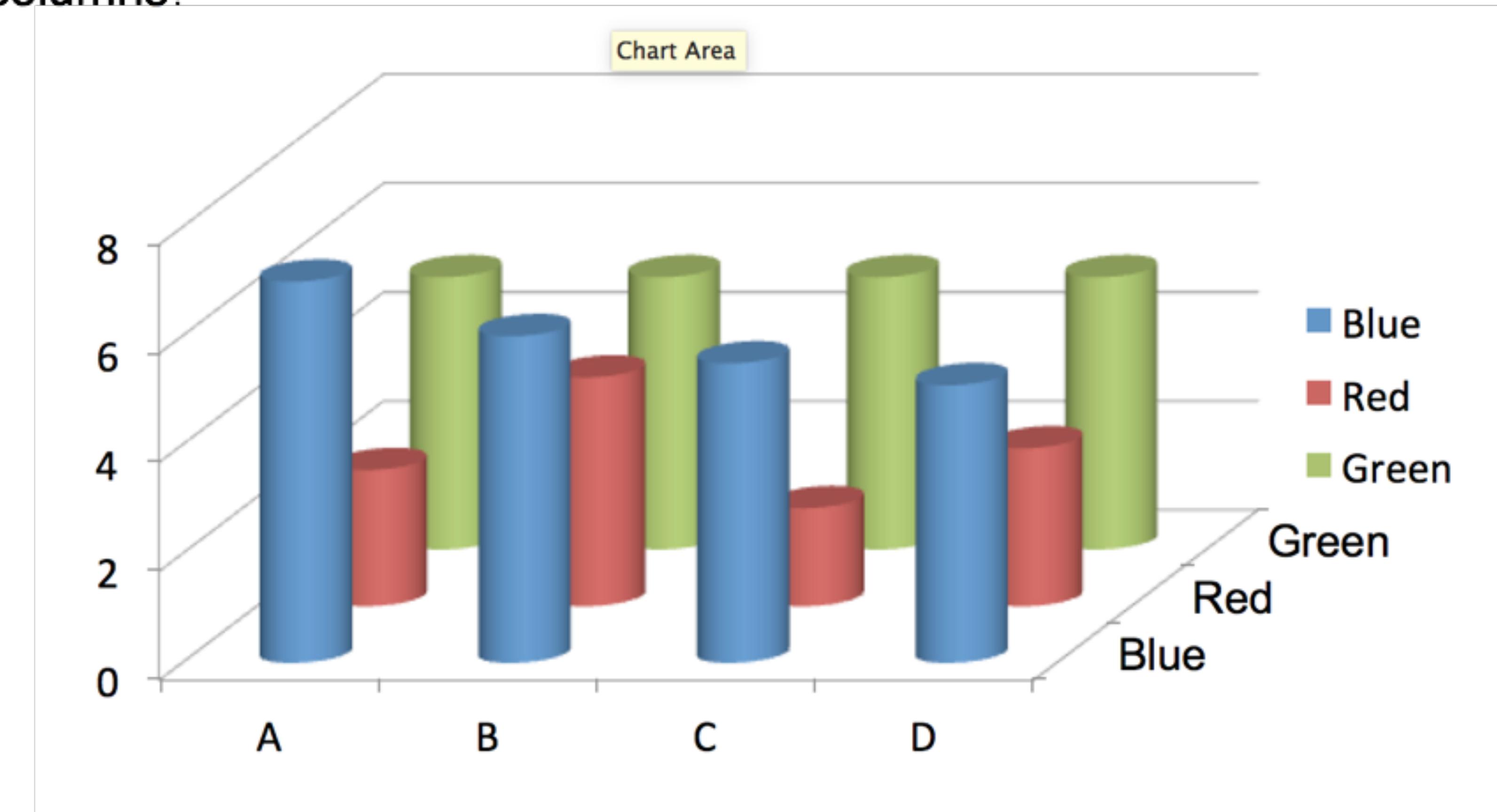
Elements of Good Visualizations

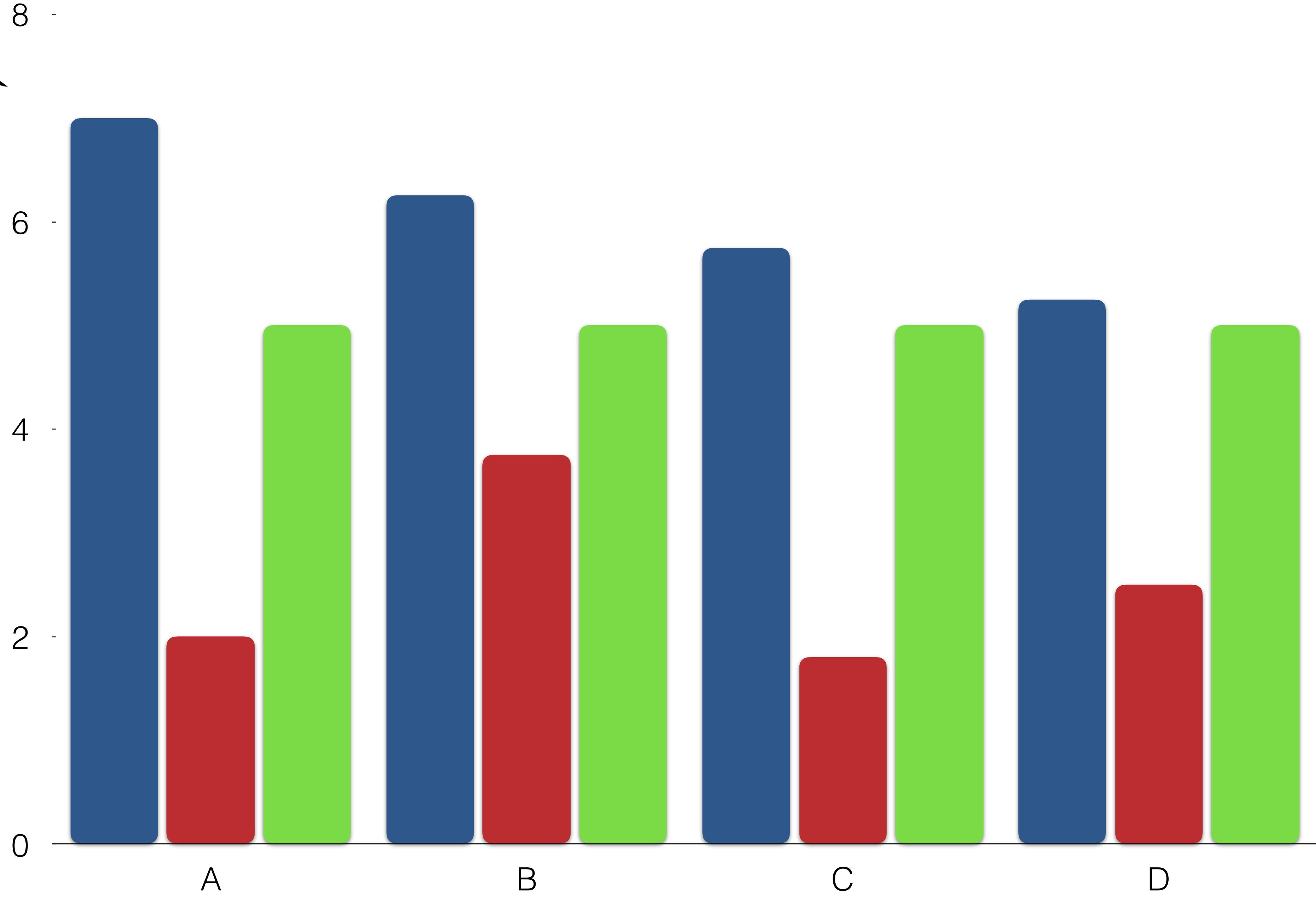
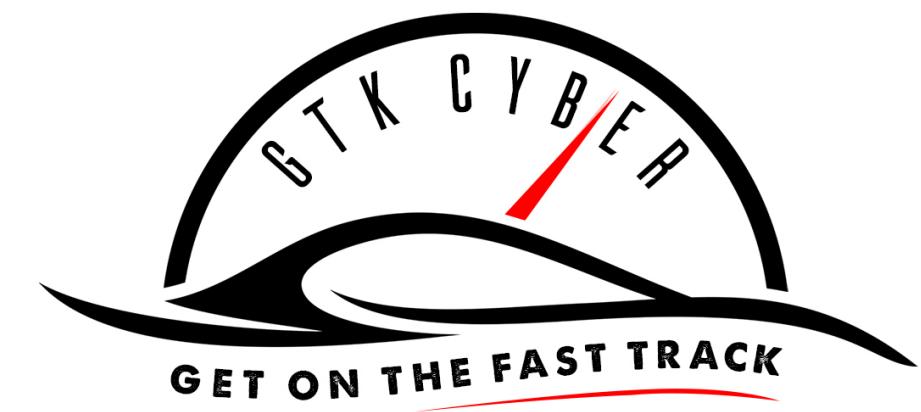
1. Graphical Integrity
2. Simple
3. Proper Display
4. Proper Color
5. Tells a story

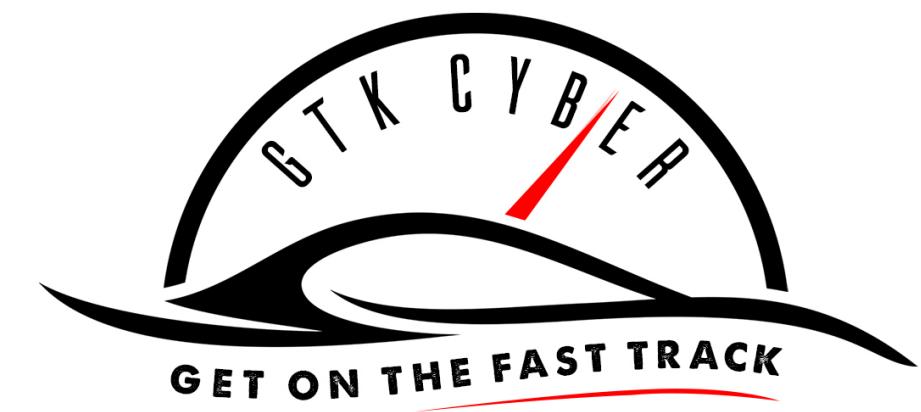


Proper Display

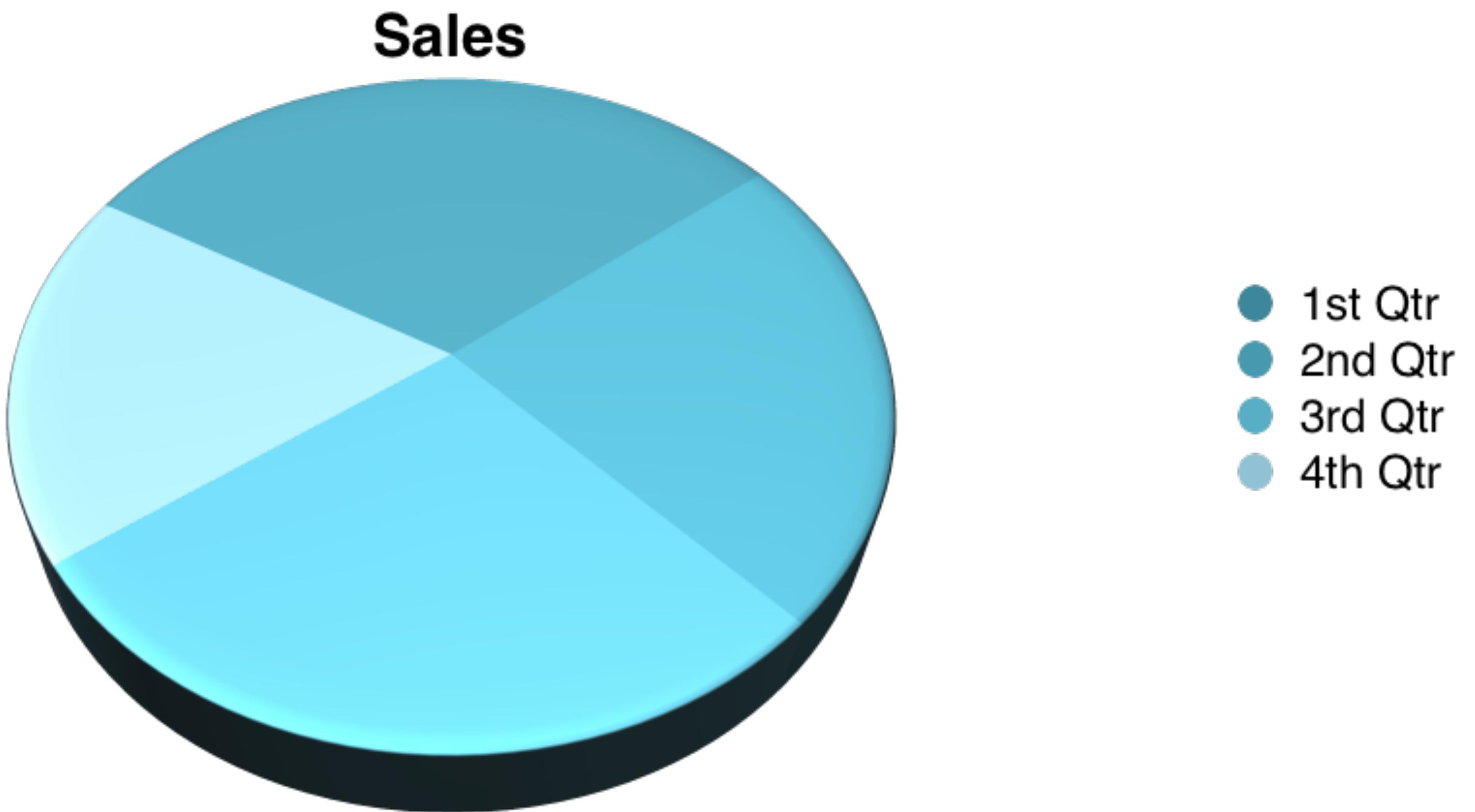
Questions: What is the height of the green columns? For which categories (A,B,C,D) are the blue columns taller than the green columns?

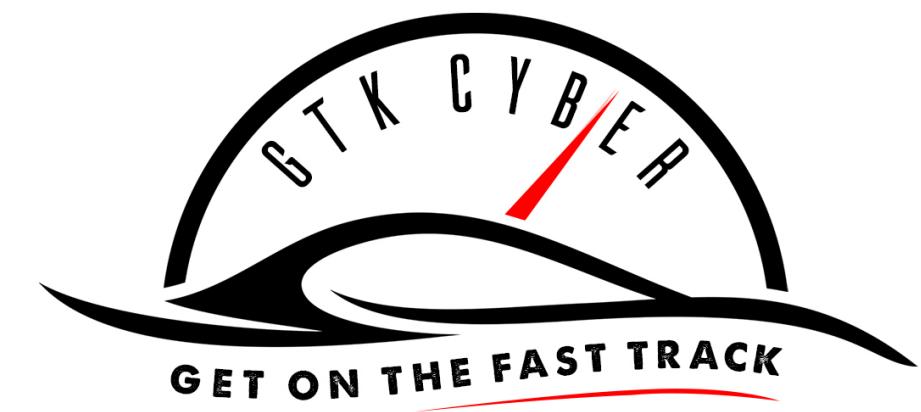




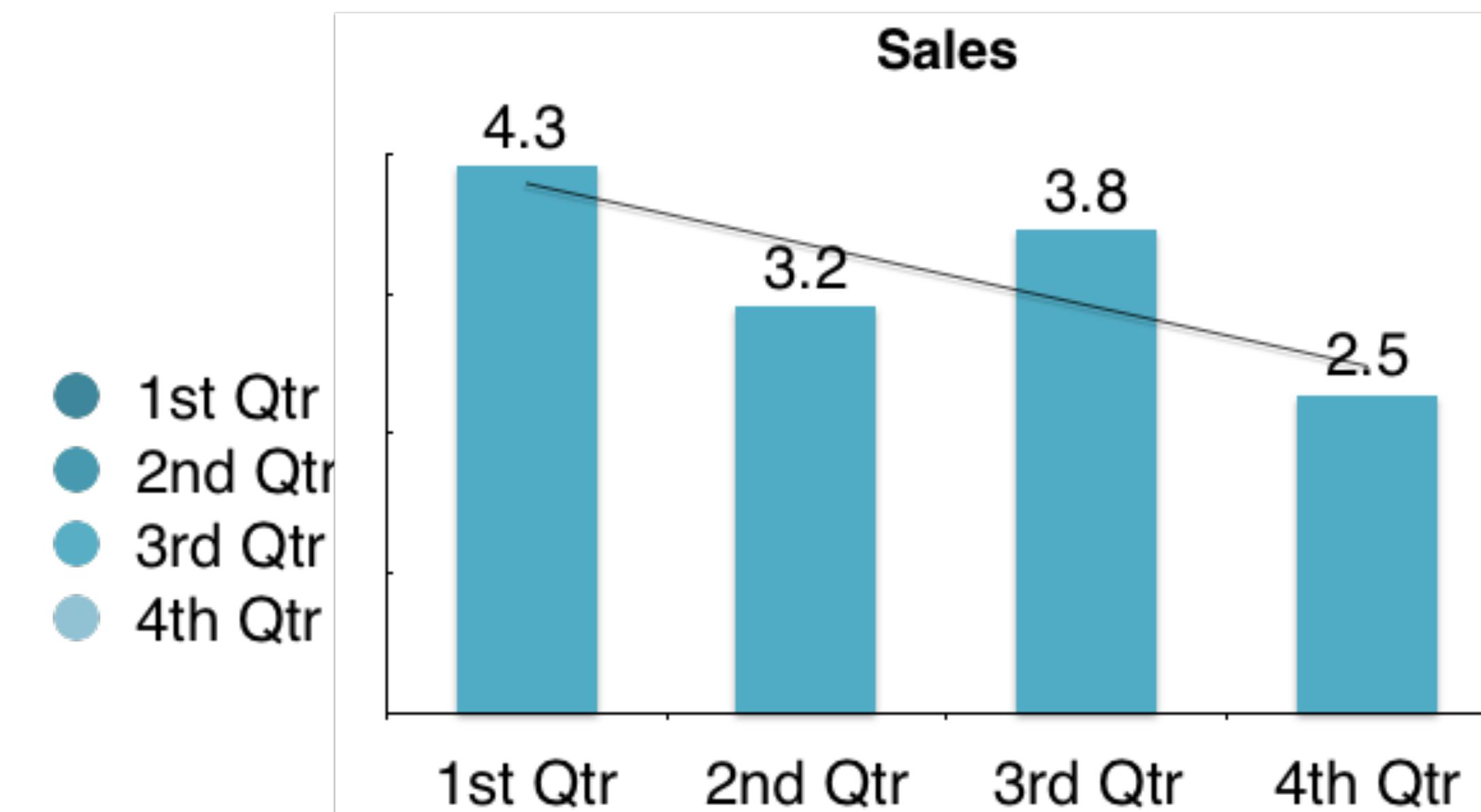


Proper Display



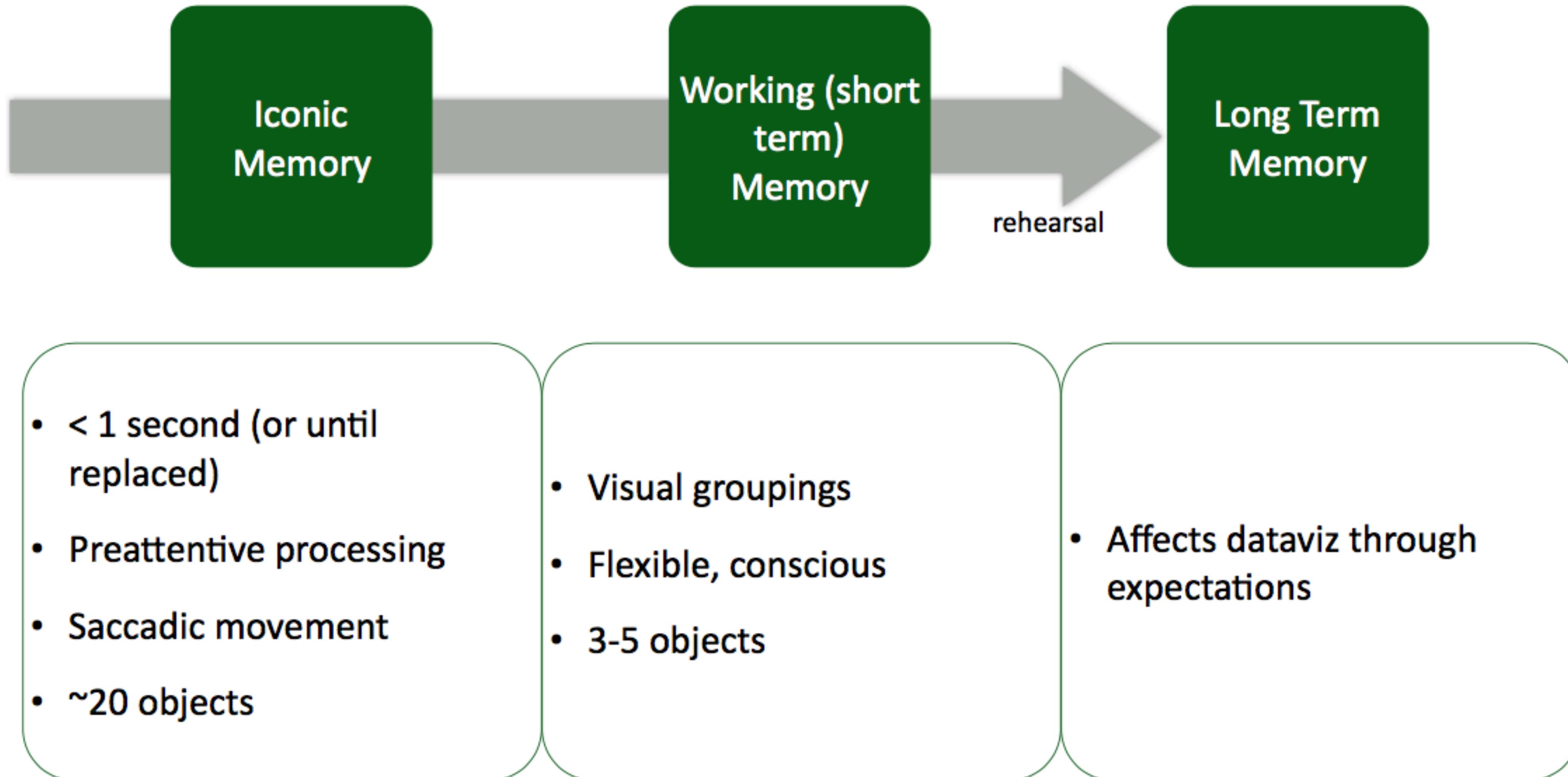


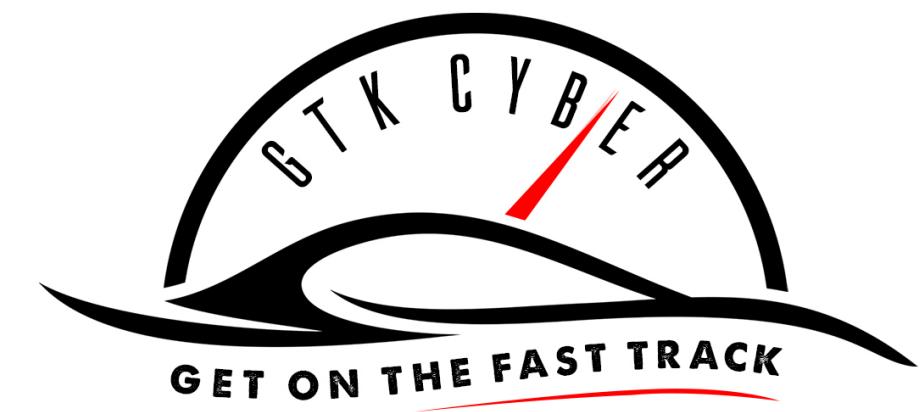
Proper Display





Visual Processing System

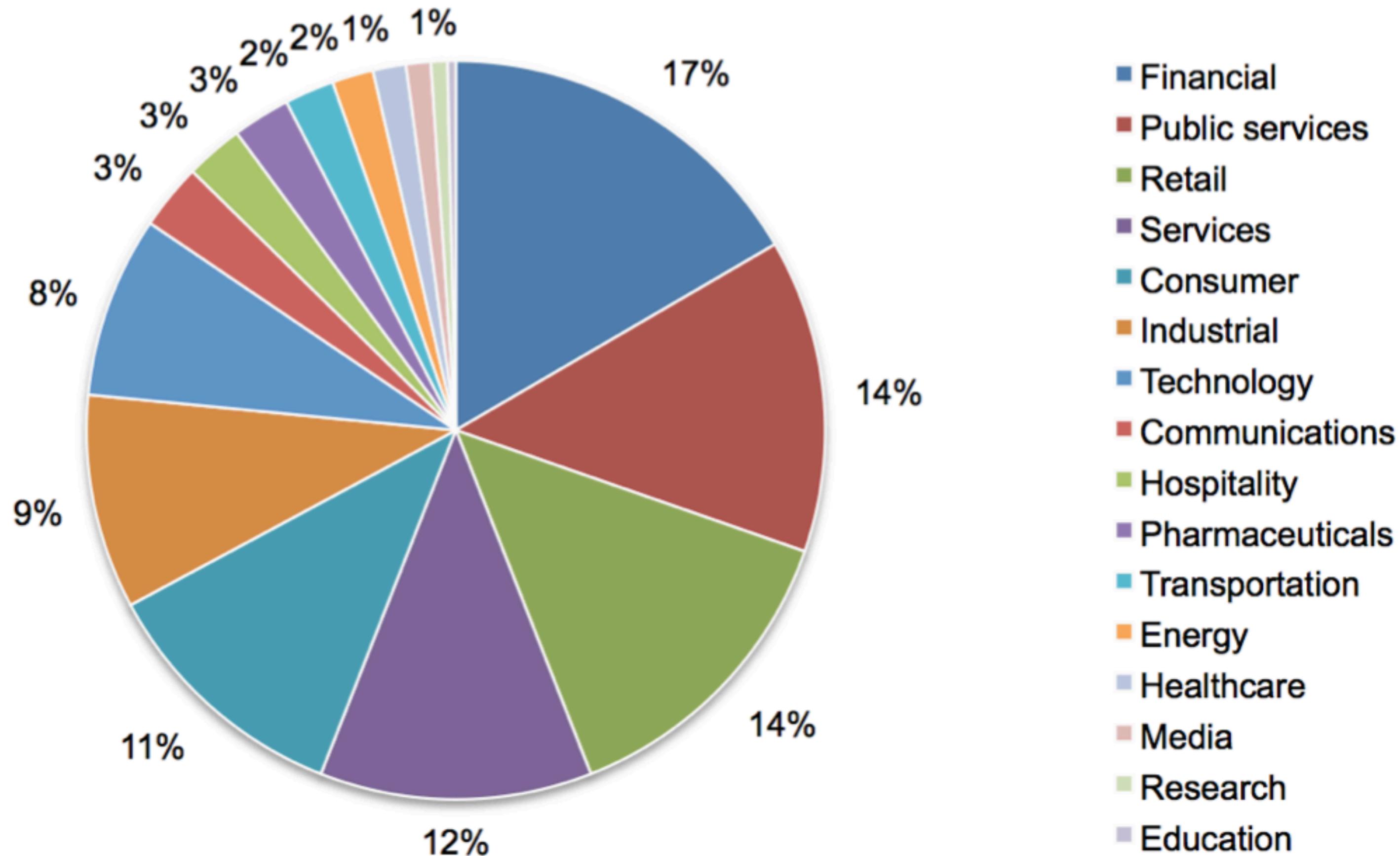


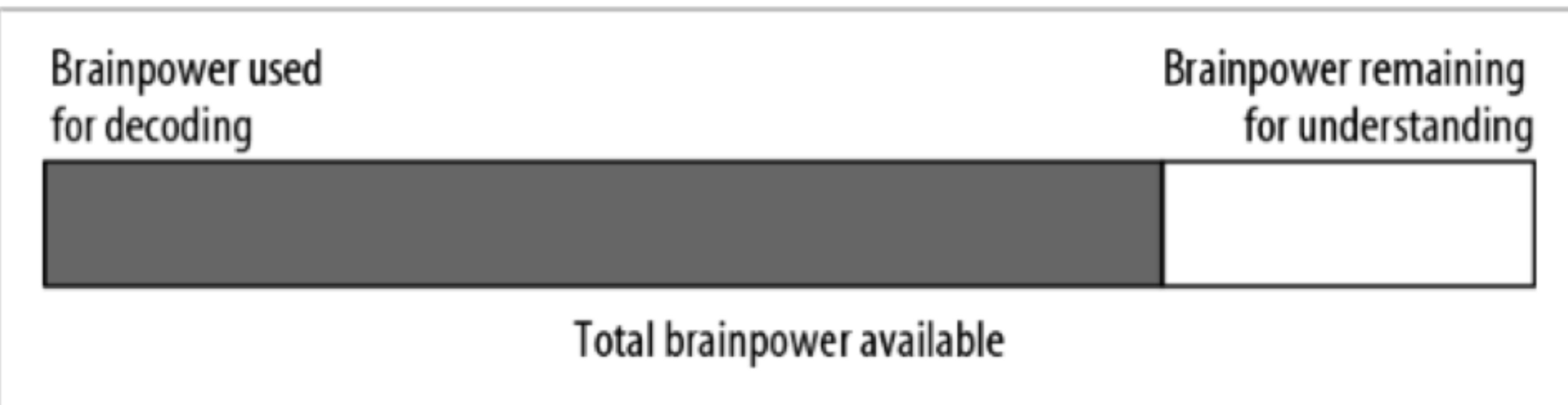


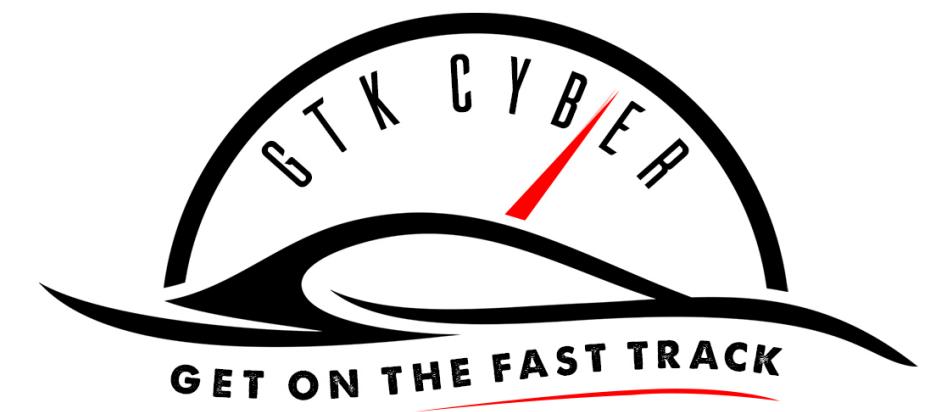
Overworking Visual Memory

Figure 20. Distribution of the benchmark sample by industry segment

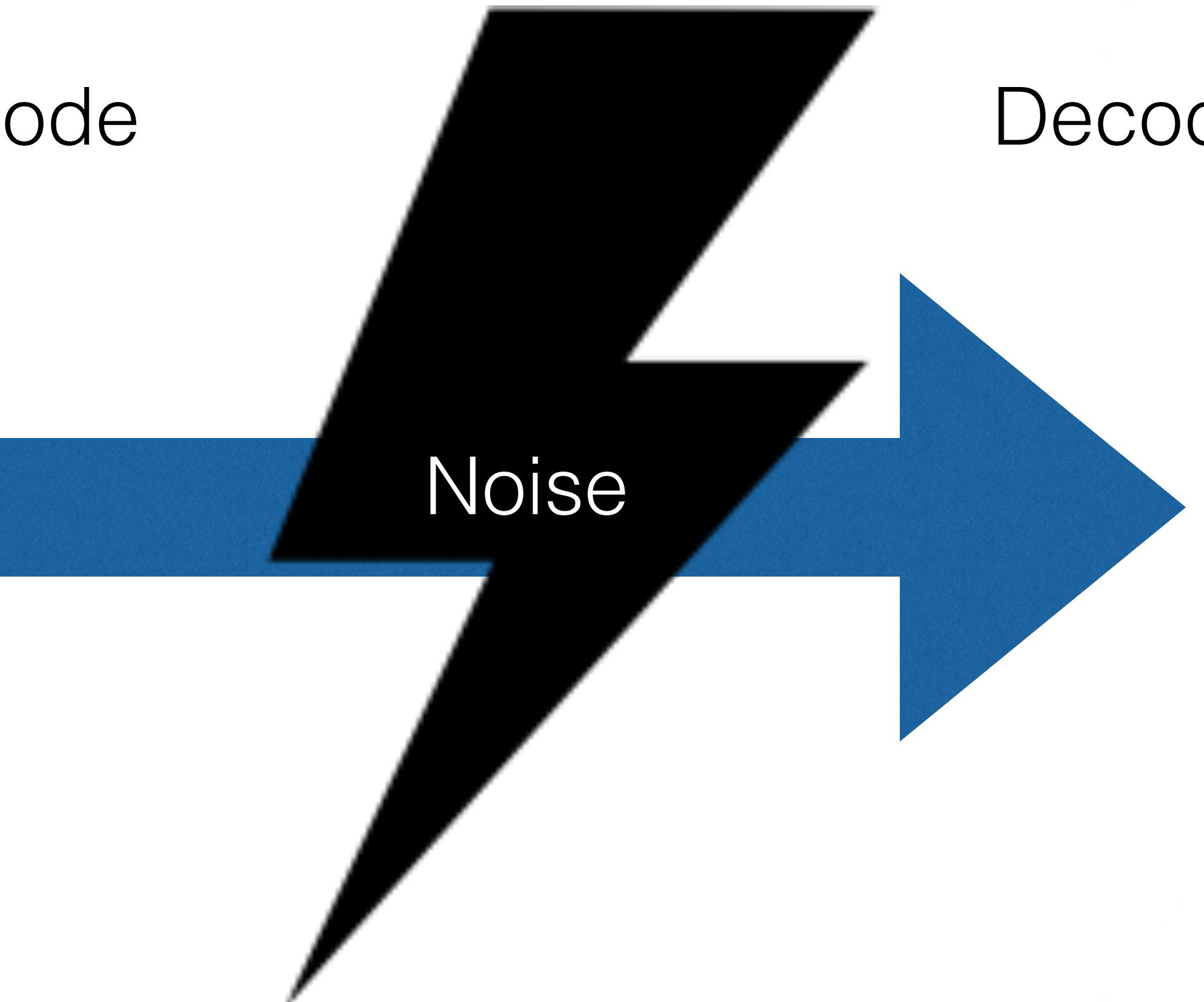
Consolidated (n = 277 organizations)

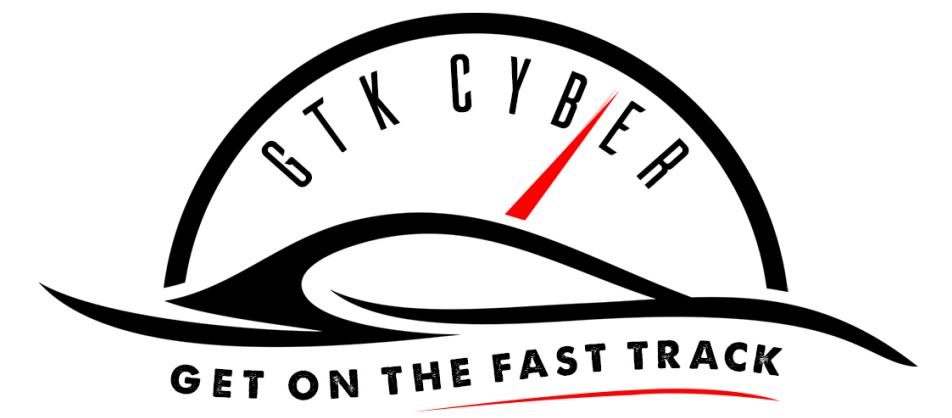




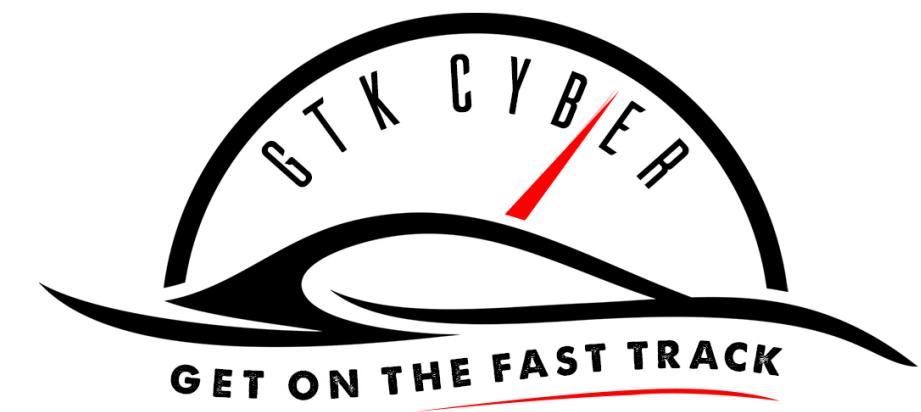


Encode



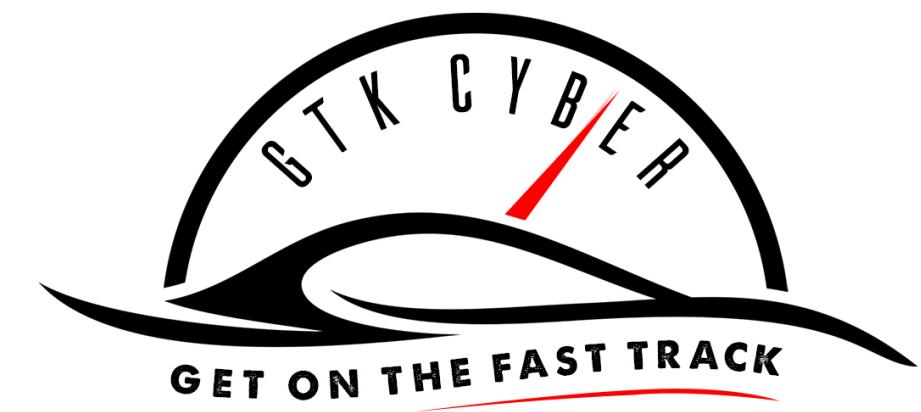


The Bad...

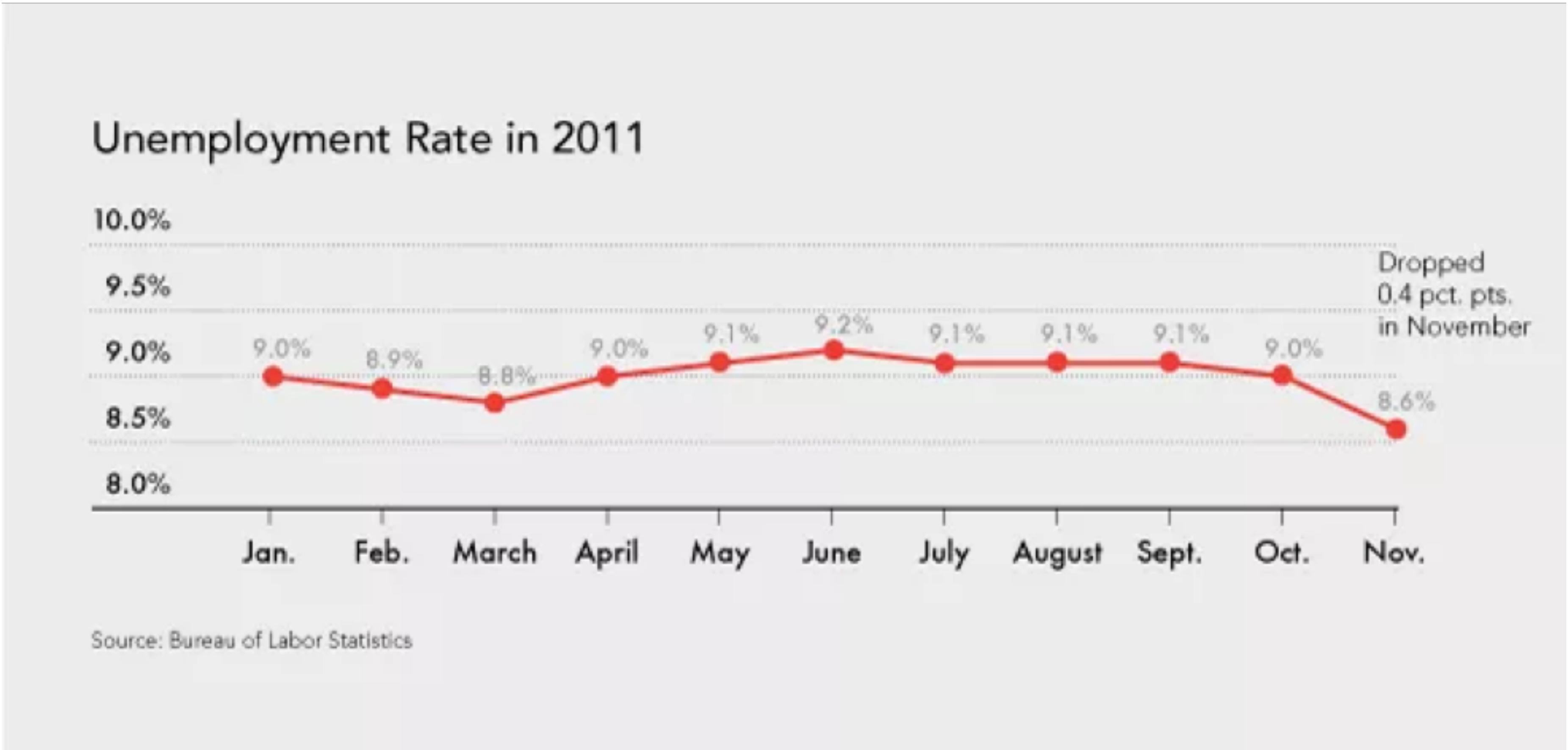


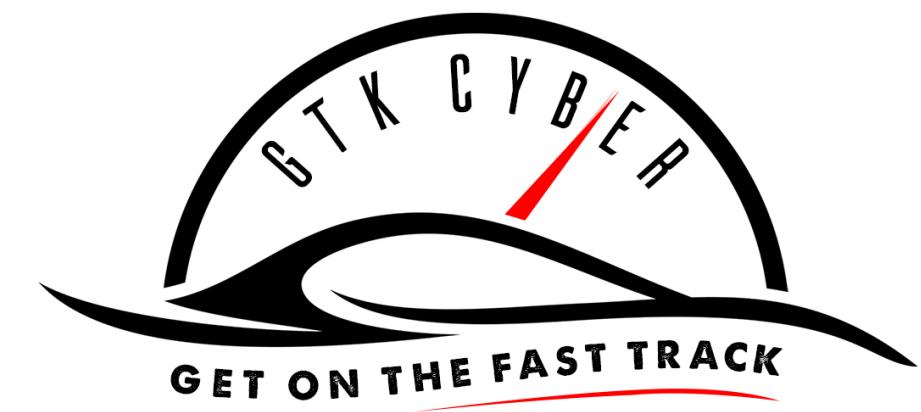
Graphical Integrity



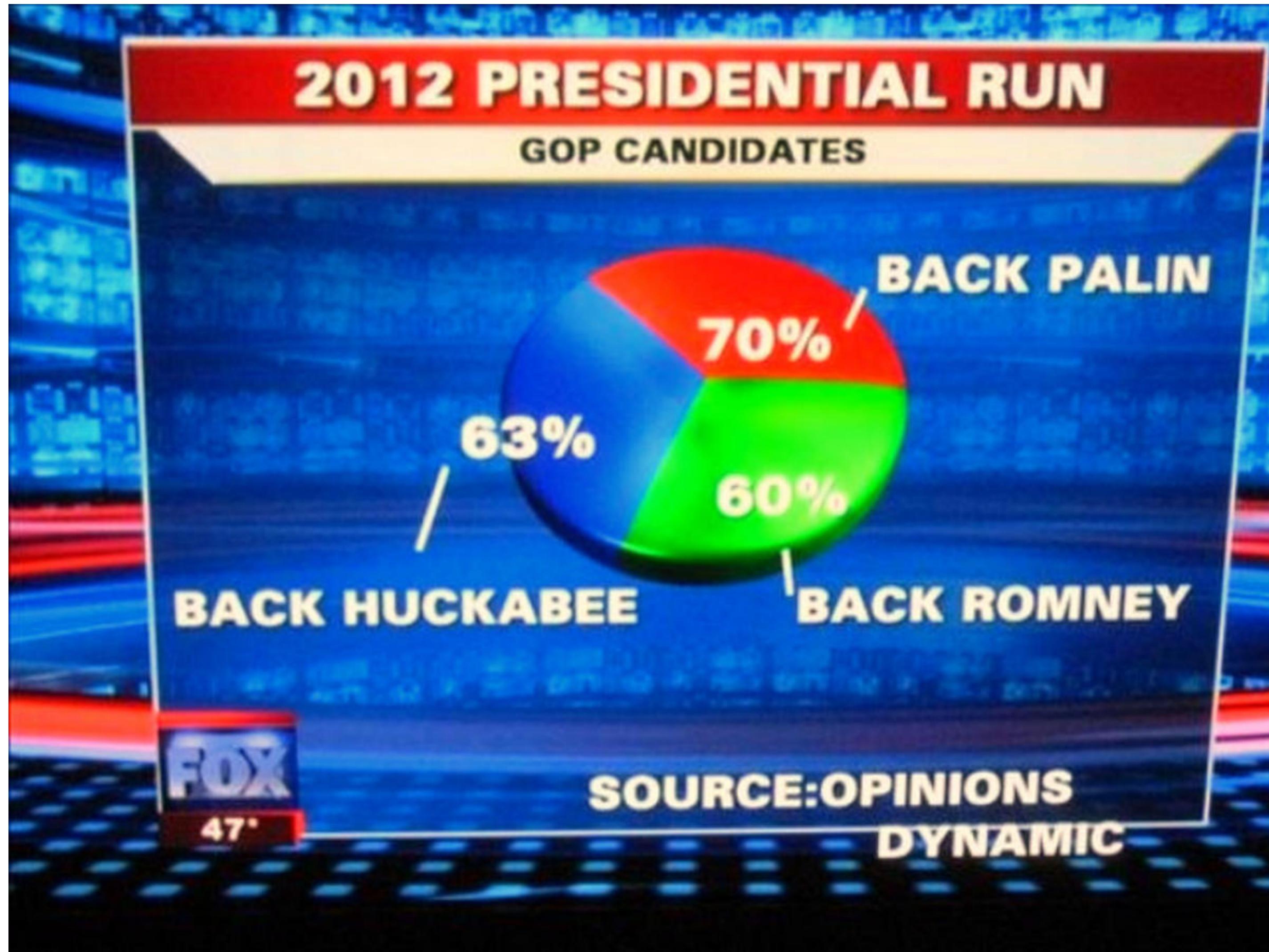


Graphical Integrity

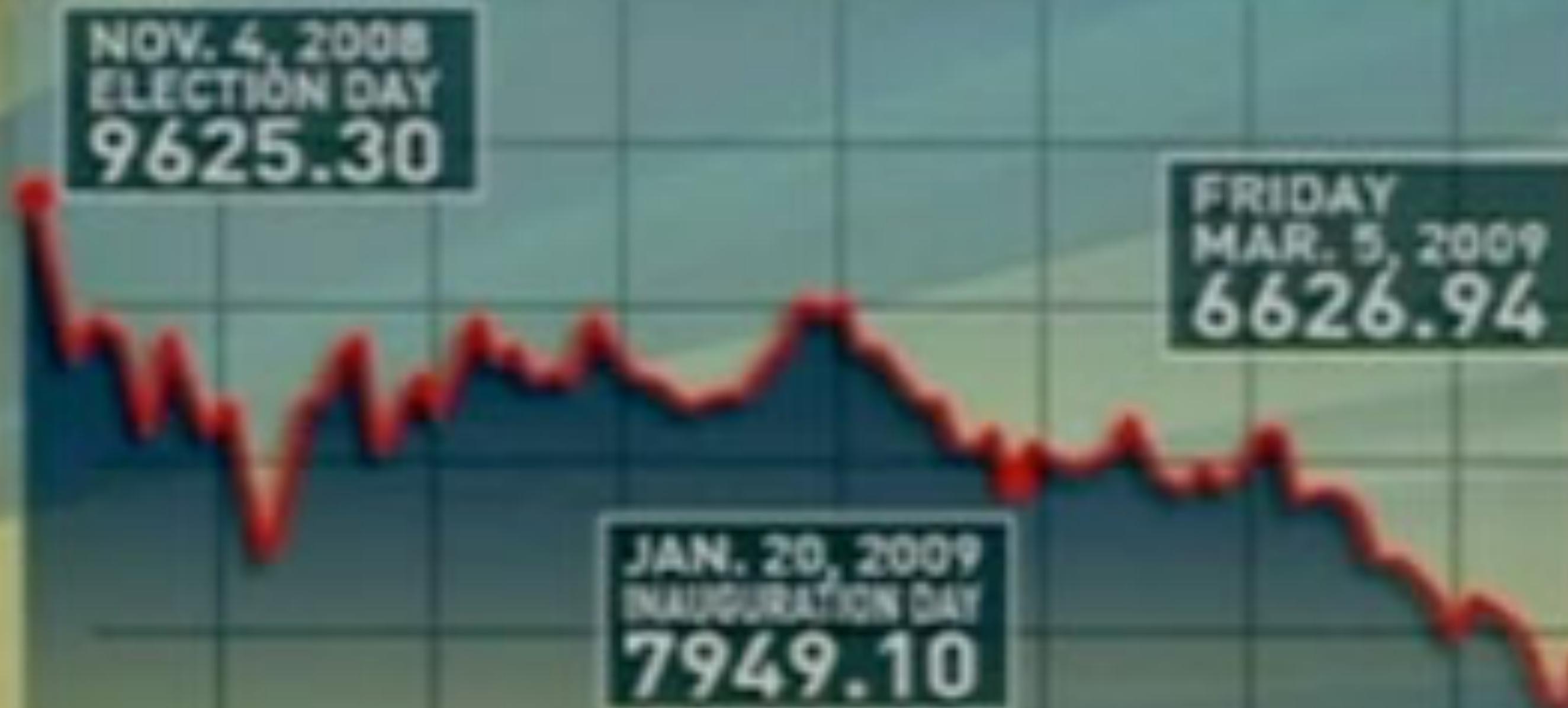




Graphical Integrity?



STOCK MARKET SLIDE

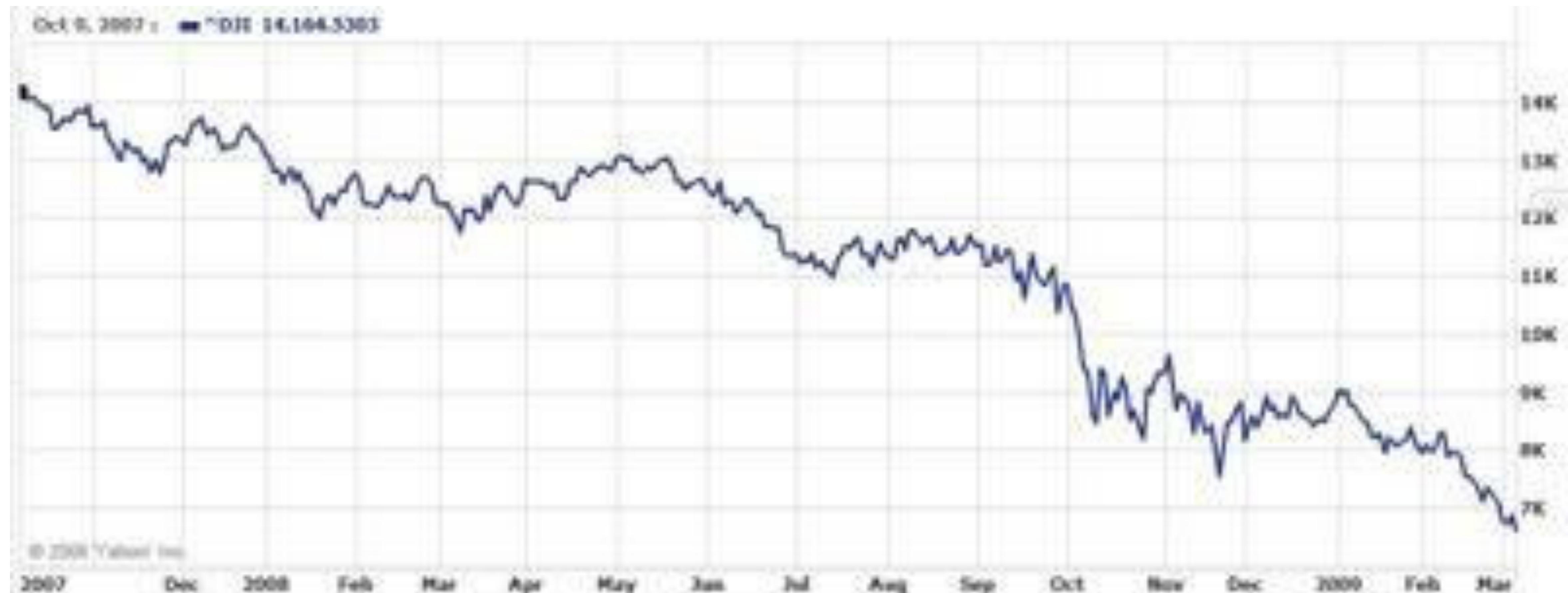
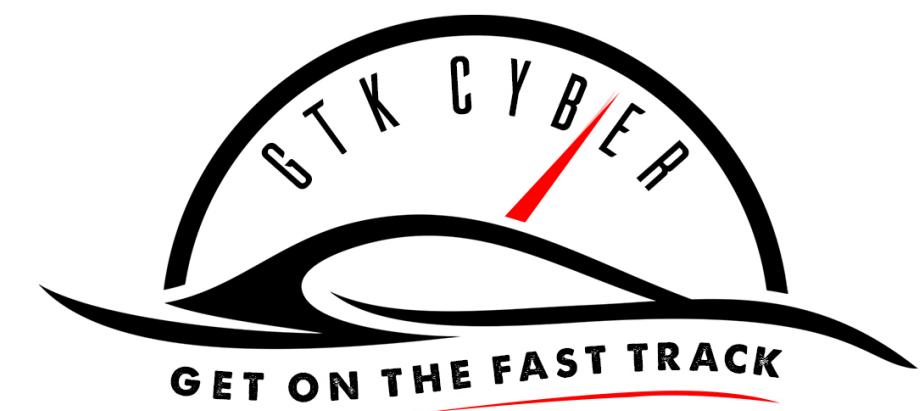


SOURCE: BUSINESSWEEK

6:31 CT

MSNBC

"A PRESIDENTT. HAS MINDD ABBAS-S IT WILL TAKE FFFFCCT .0





GEORGIA
FEELINGS ABOUT THE FEDERAL GOVERNMENT

PRESIDENT

SATISFIED

D CLINTON 81%

R TRUMP 14%

J JOHNSON 4%

CNN EXIT POLL

ELECTORAL MAP

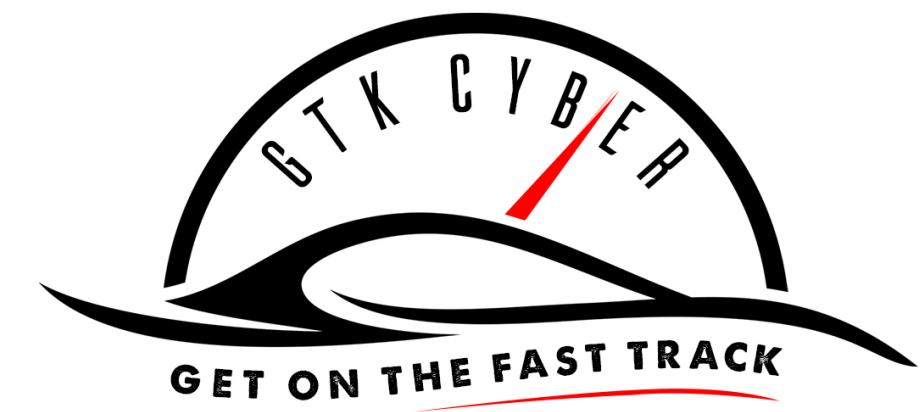
IN PRESIDENT

R TRUMP ✓
D CLINTON

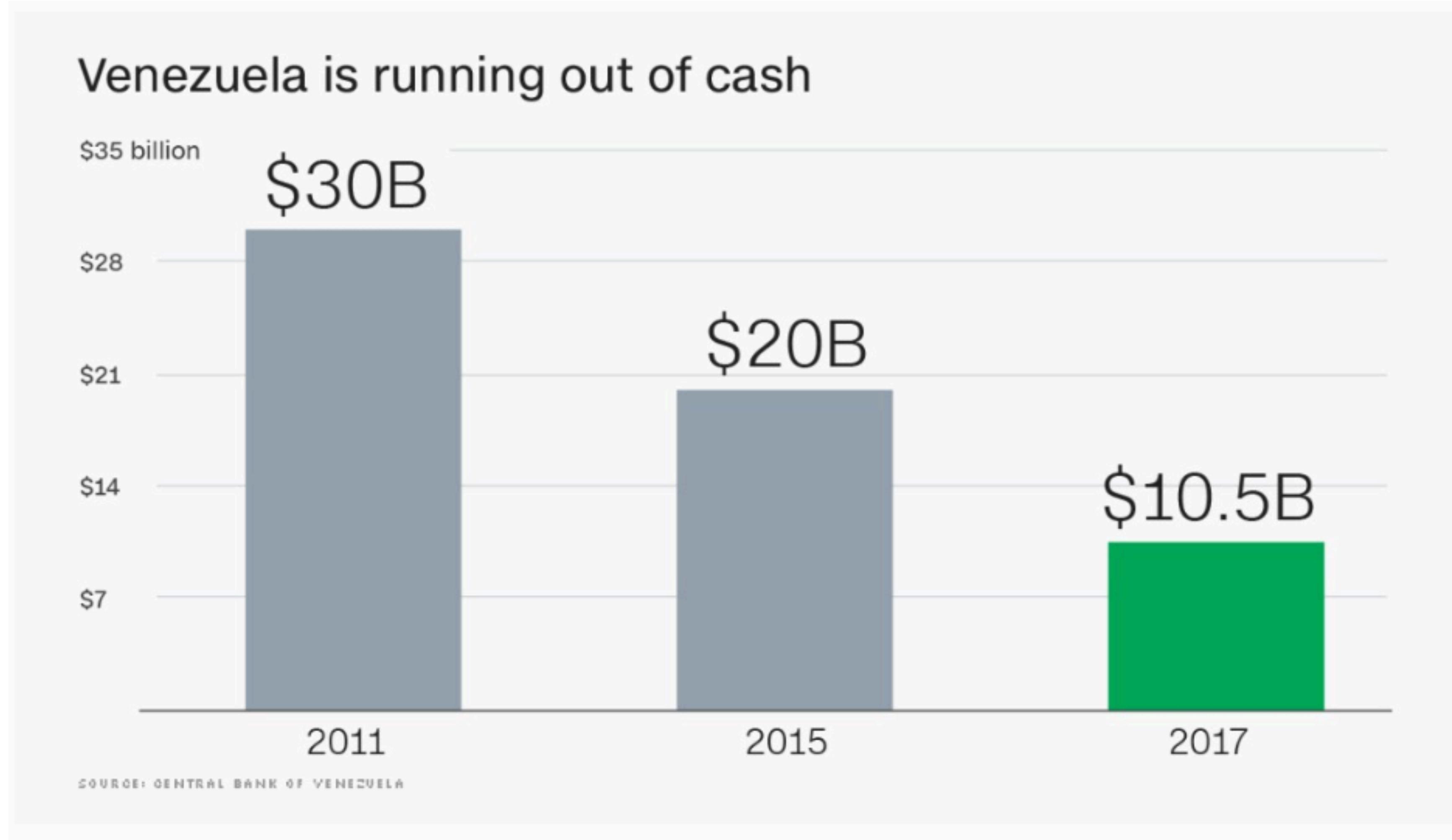
126,791 65.4%
59,284 30.6%

26:23
NEXT POLLS CLOSE
LIVE
CNN

The image shows a CNN news broadcast from a studio. On the left, a large graphic displays exit poll results for Georgia's feelings about the federal government, with a pie chart showing 81% for Clinton, 14% for Trump, and 4% for Johnson. Below this is an electoral map showing 19 electoral votes for Trump and 3 for Clinton. On the right, a male news anchor in a suit and glasses is speaking. A lower third graphic shows the results for President, with Trump leading Clinton by 126,791 votes (65.4%) to 59,284 (30.6%). The time 26:23 is displayed, along with a note that "NEXT POLLS CLOSE" and the word "LIVE". The CNN logo is in the bottom right corner.

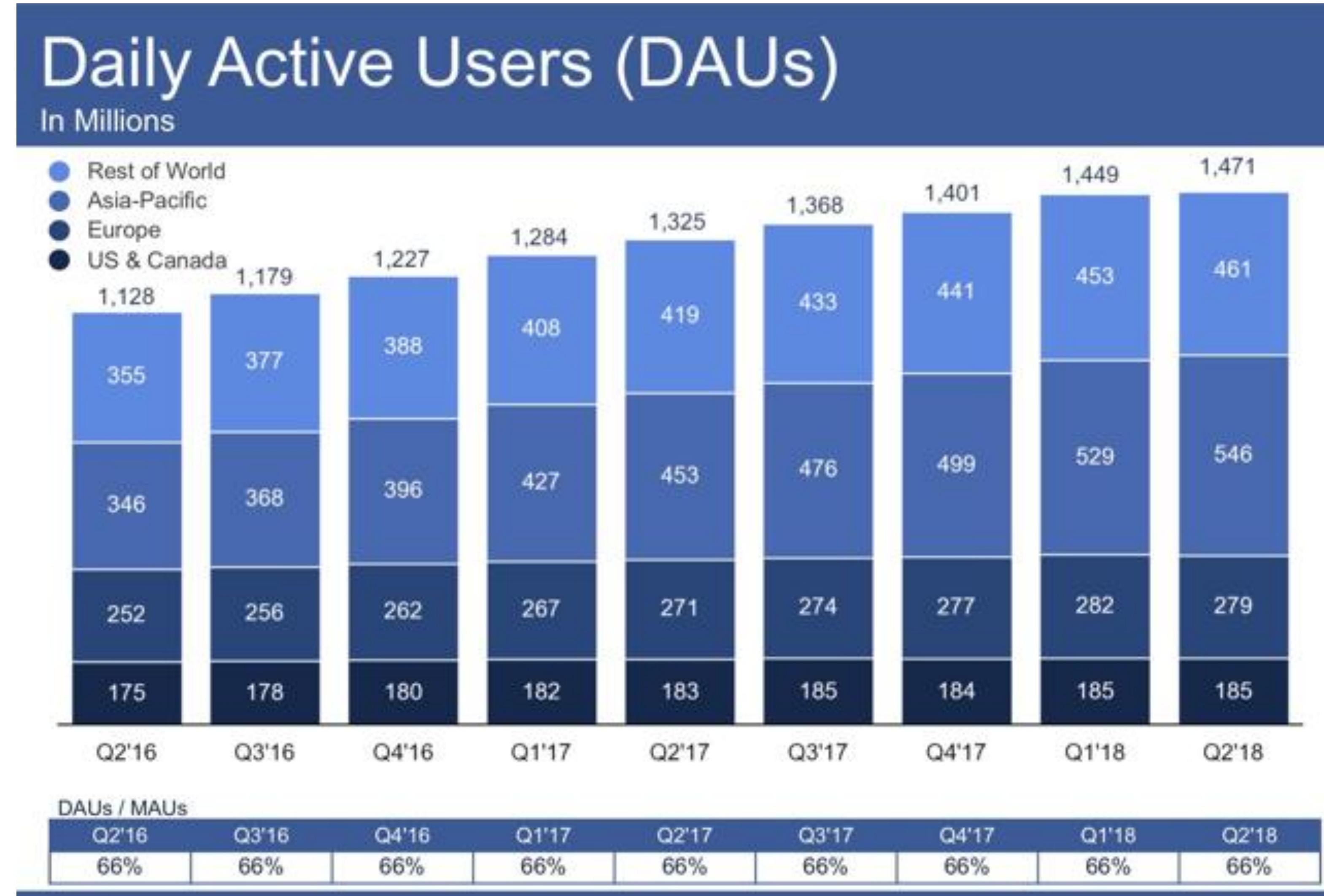


Graphical Integrity



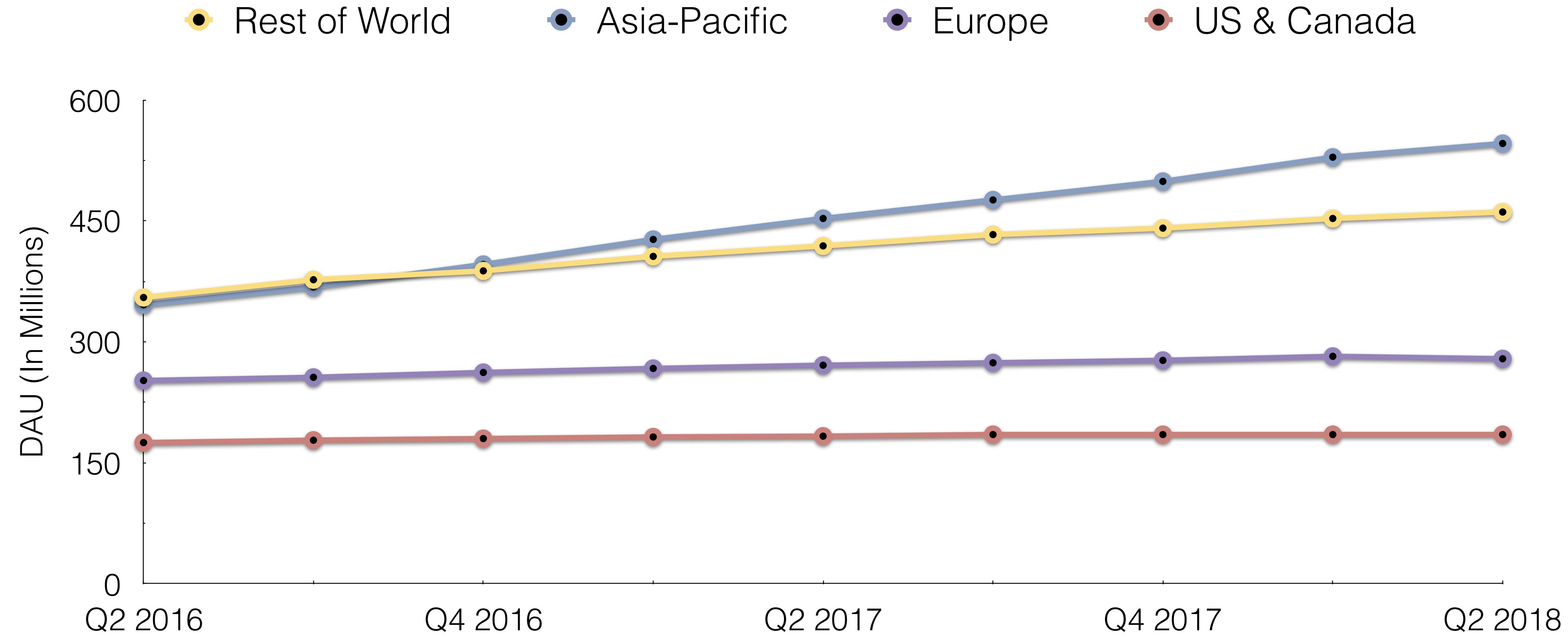


Proper Display



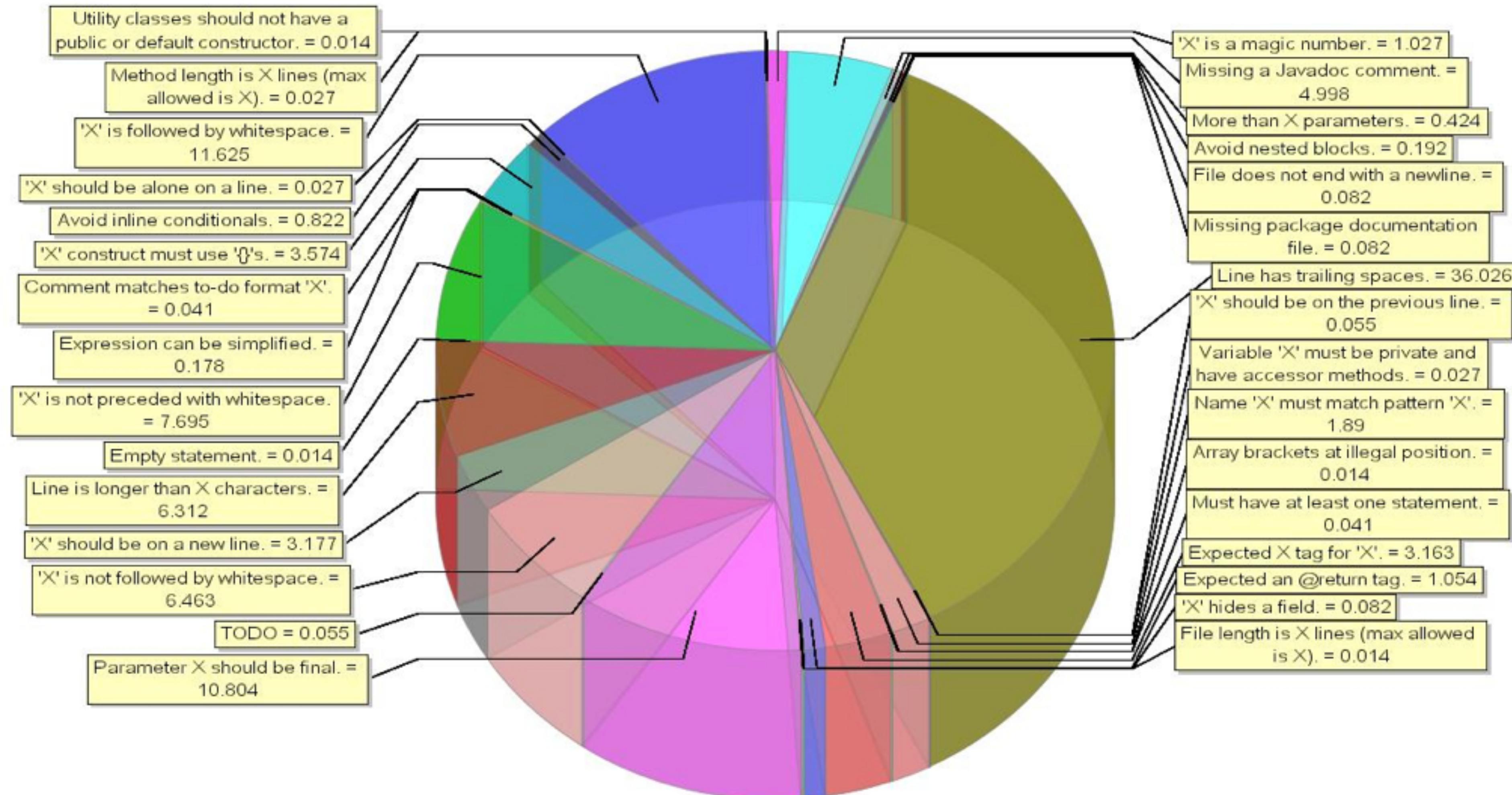


Proper Display



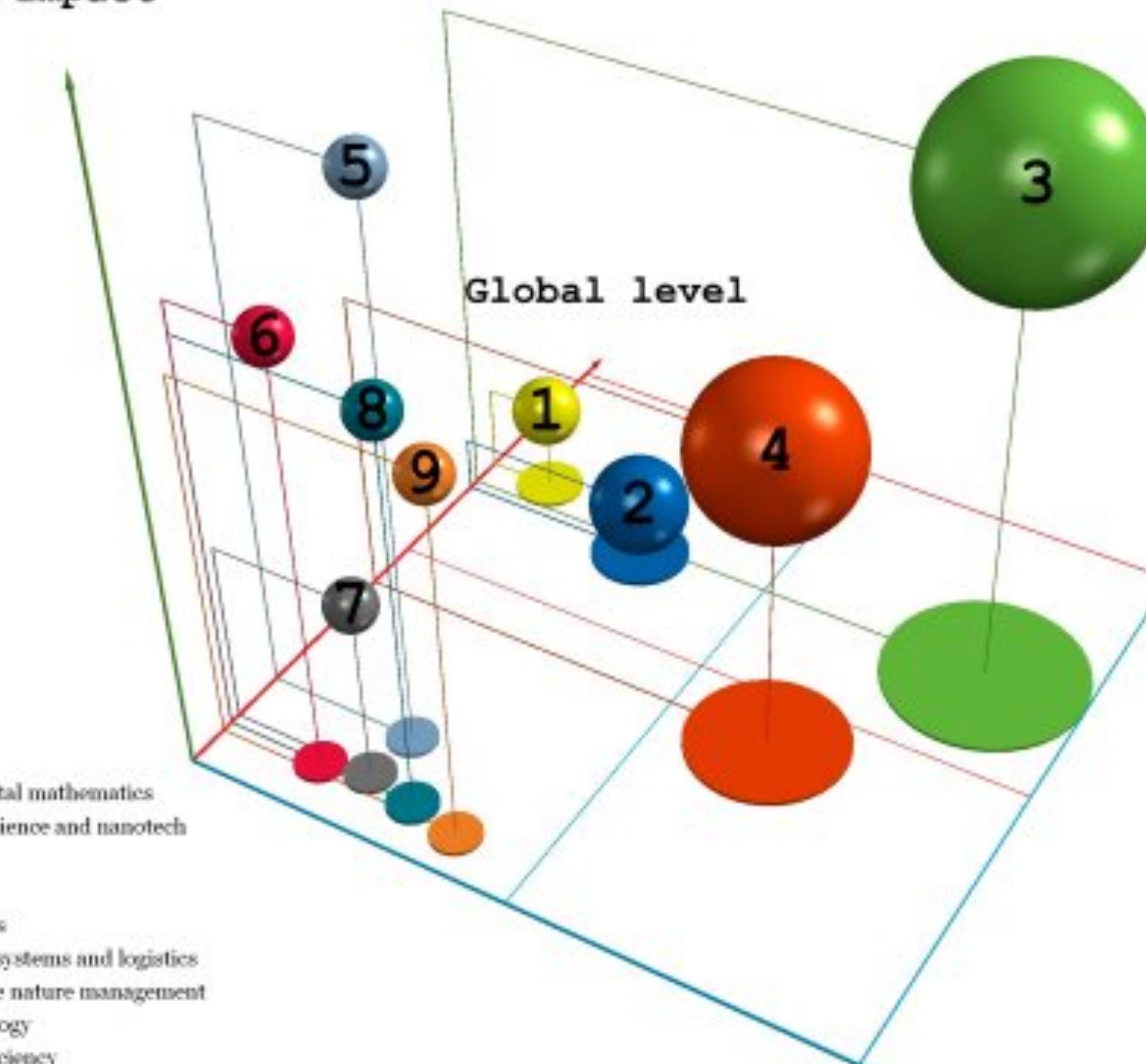


Simple





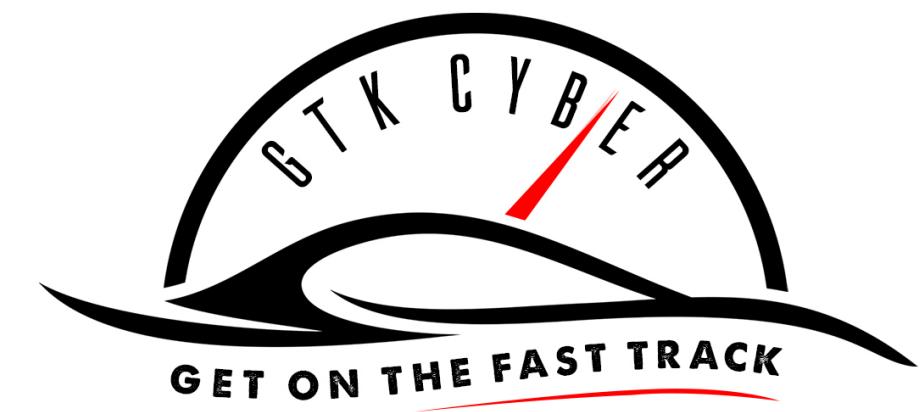
Social impact



- 1 Fundamental mathematics
- 2 Material science and nanotech
- 3 Biotech
- 4 ICT
- 5 Humanities
- 6 Transport systems and logistics
- 7 Sustainable nature management
- 8 Urban ecology
- 9 Energy efficiency

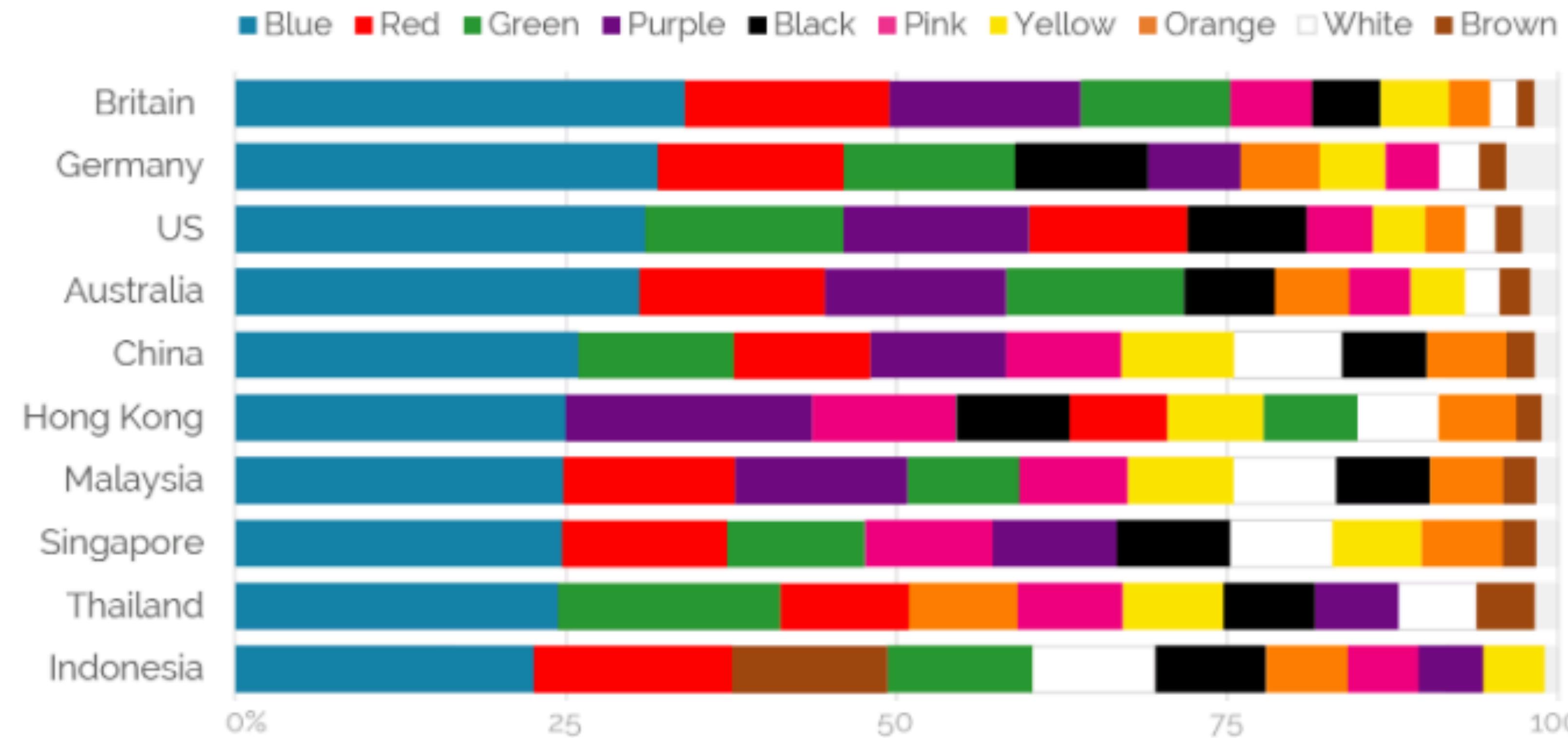
Axis X: Global level
Axis Y: Social impact
Axis Z: Regional priorities
Axis S: Capital investment

Regional priorities



Blue planet

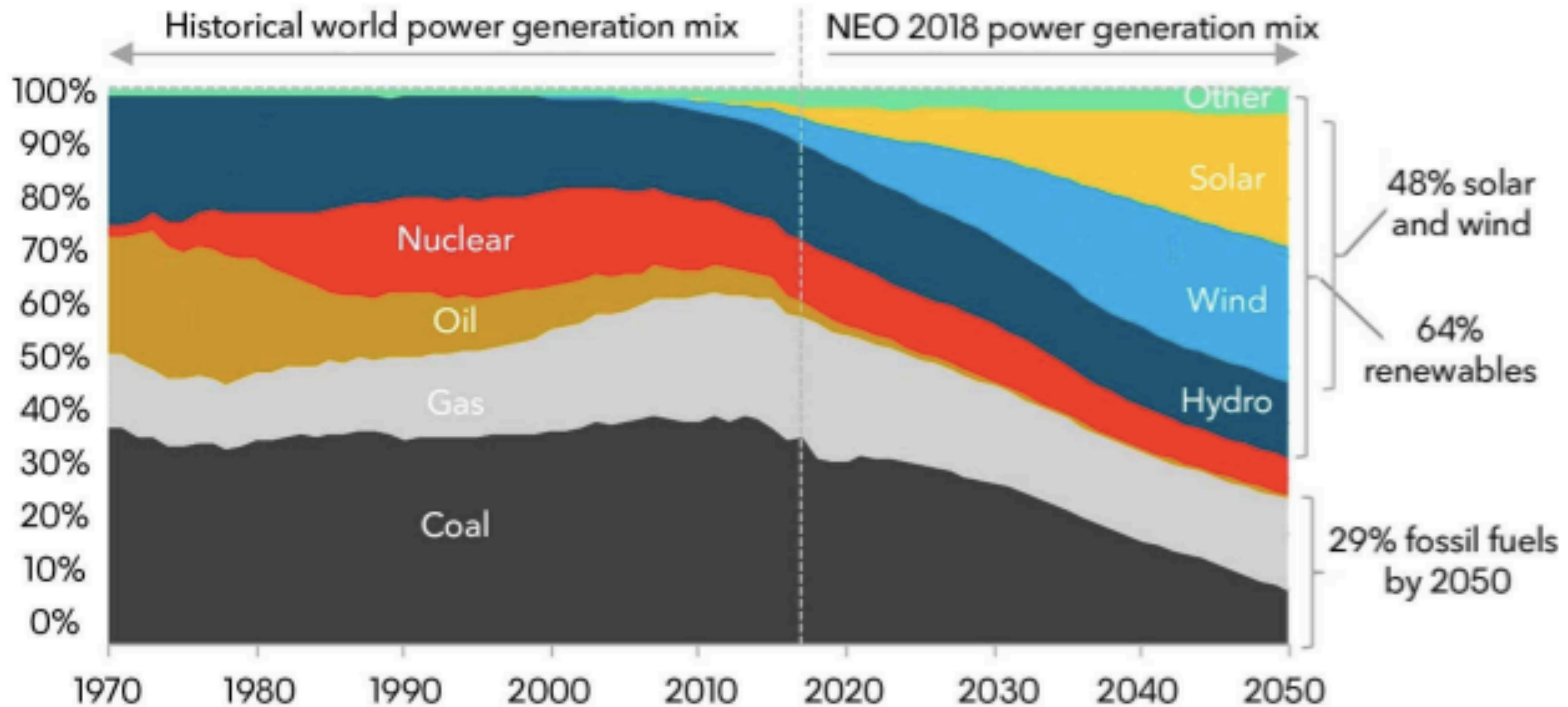
Which one of the colors listed below do you like the most?



YouGov | yougov.com



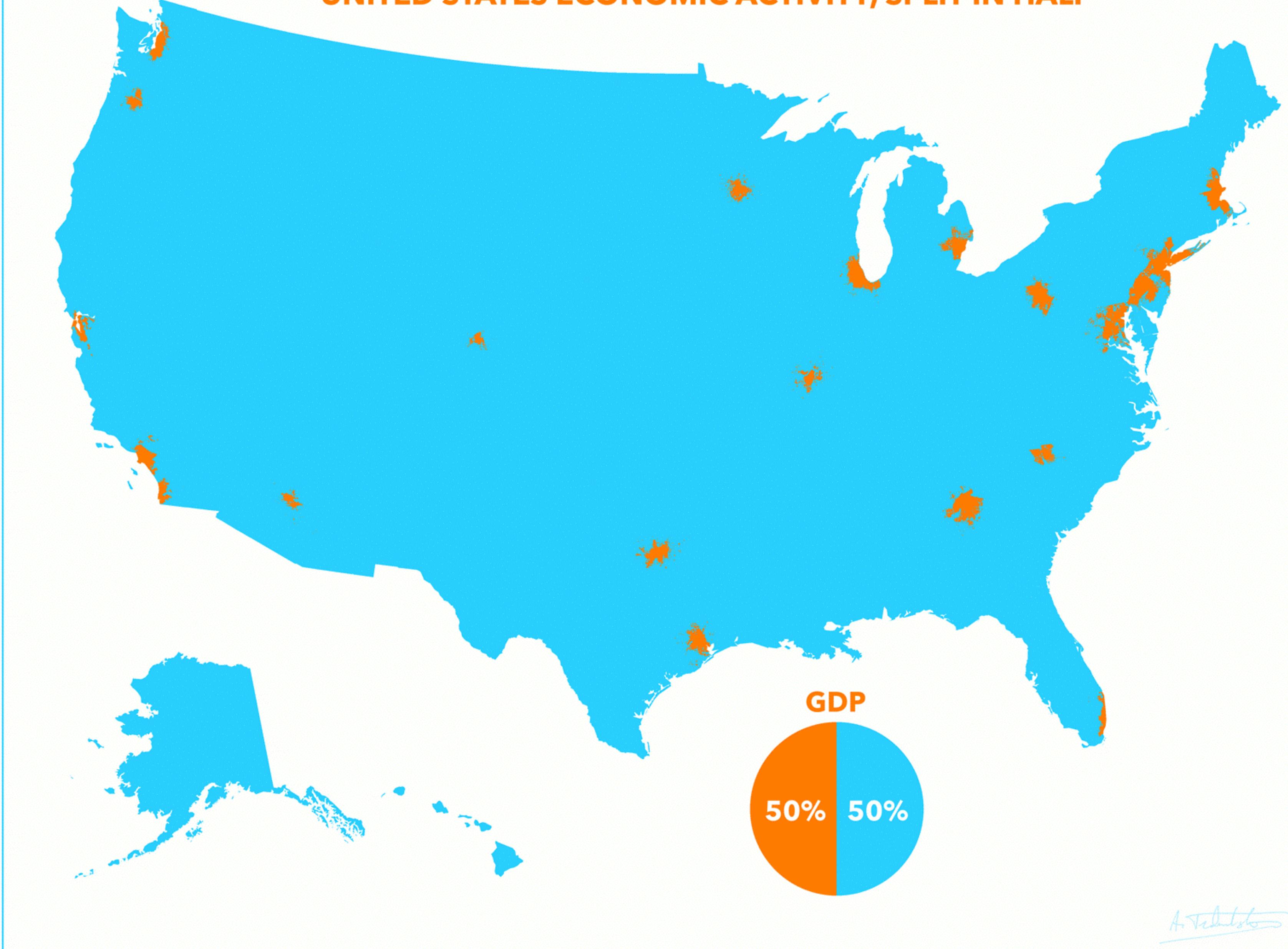
Power generation mix



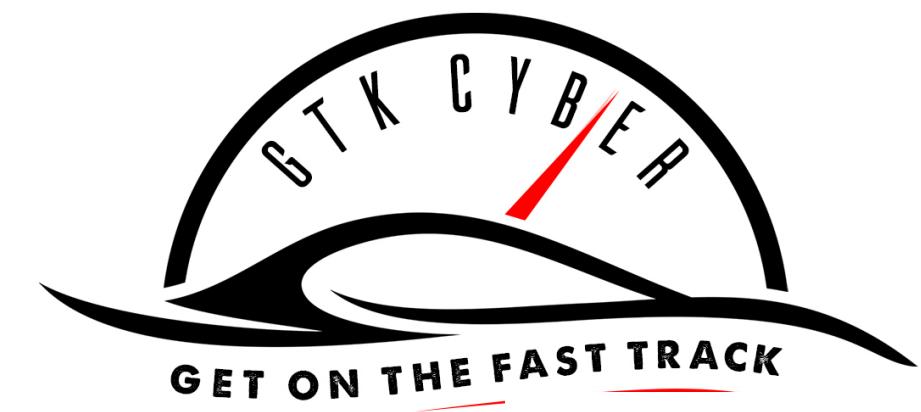
Source: Bloomberg NEF



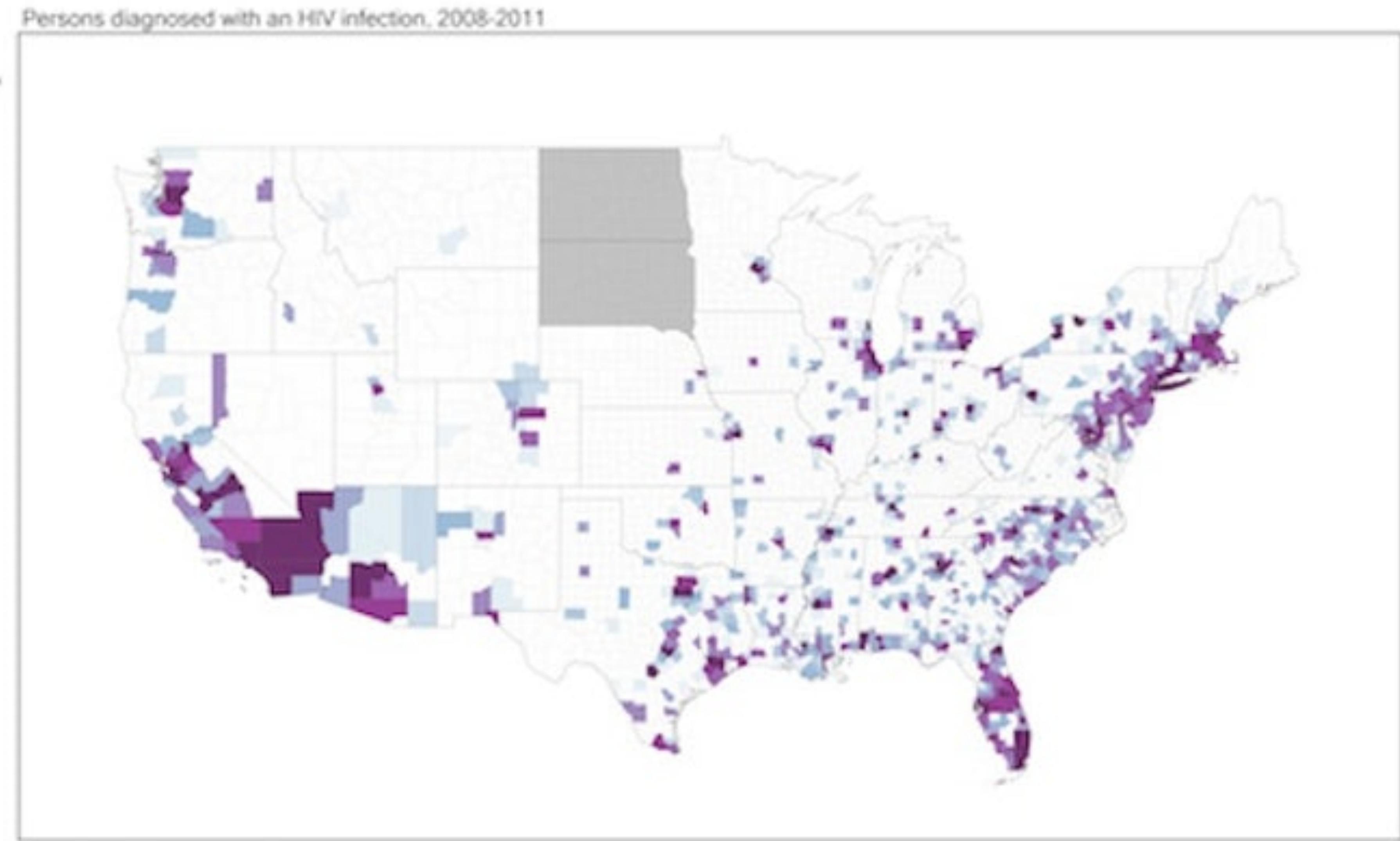
UNITED STATES ECONOMIC ACTIVITY, SPLIT IN HALF



<http://www.thefunctionalart.com/2014/02/the-incredible-gdp-map-that-shows-that.html>

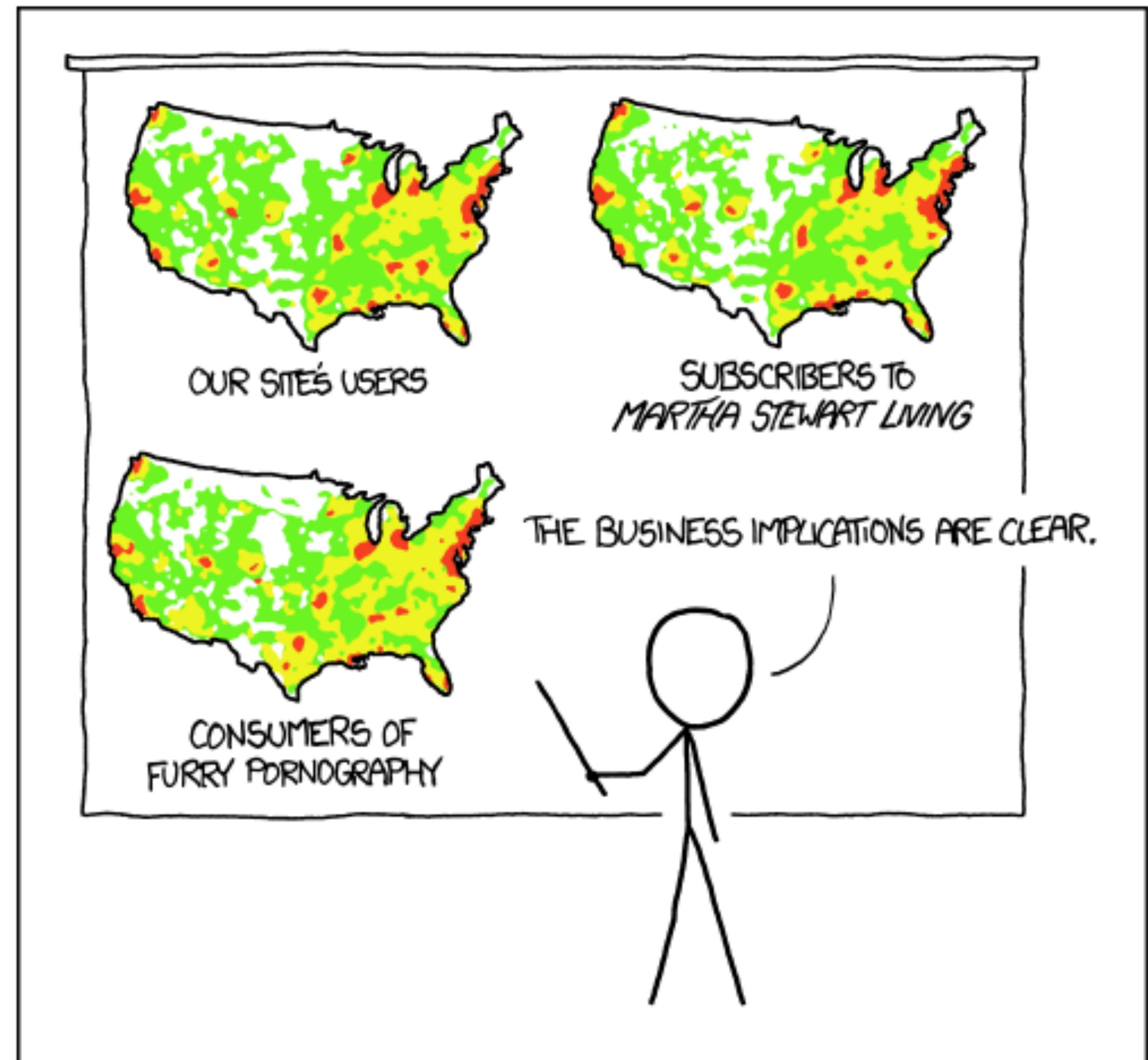
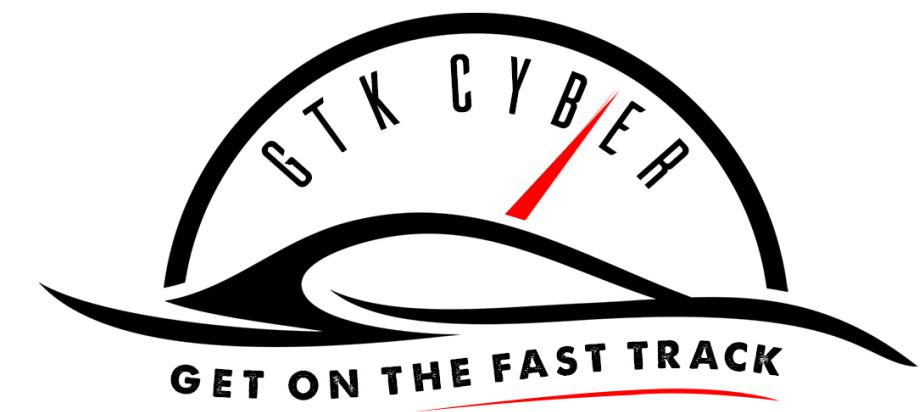


Startling New Map: 92 Percent of New HIV Cases are in 25 Percent of Counties

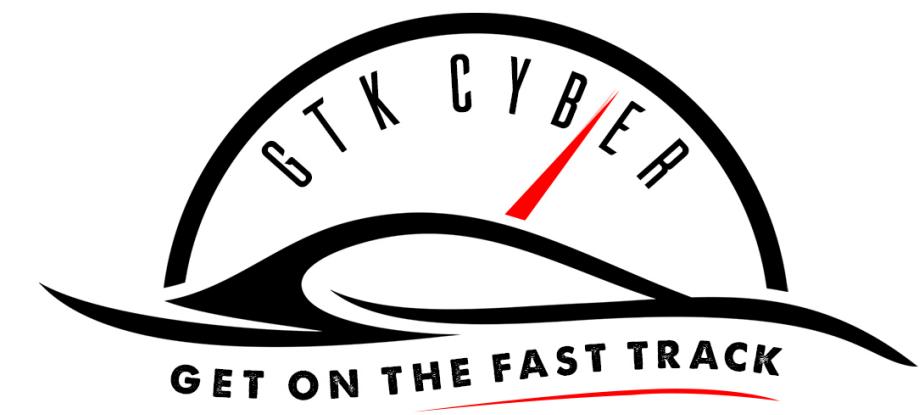


** Data are not shown to protect privacy.

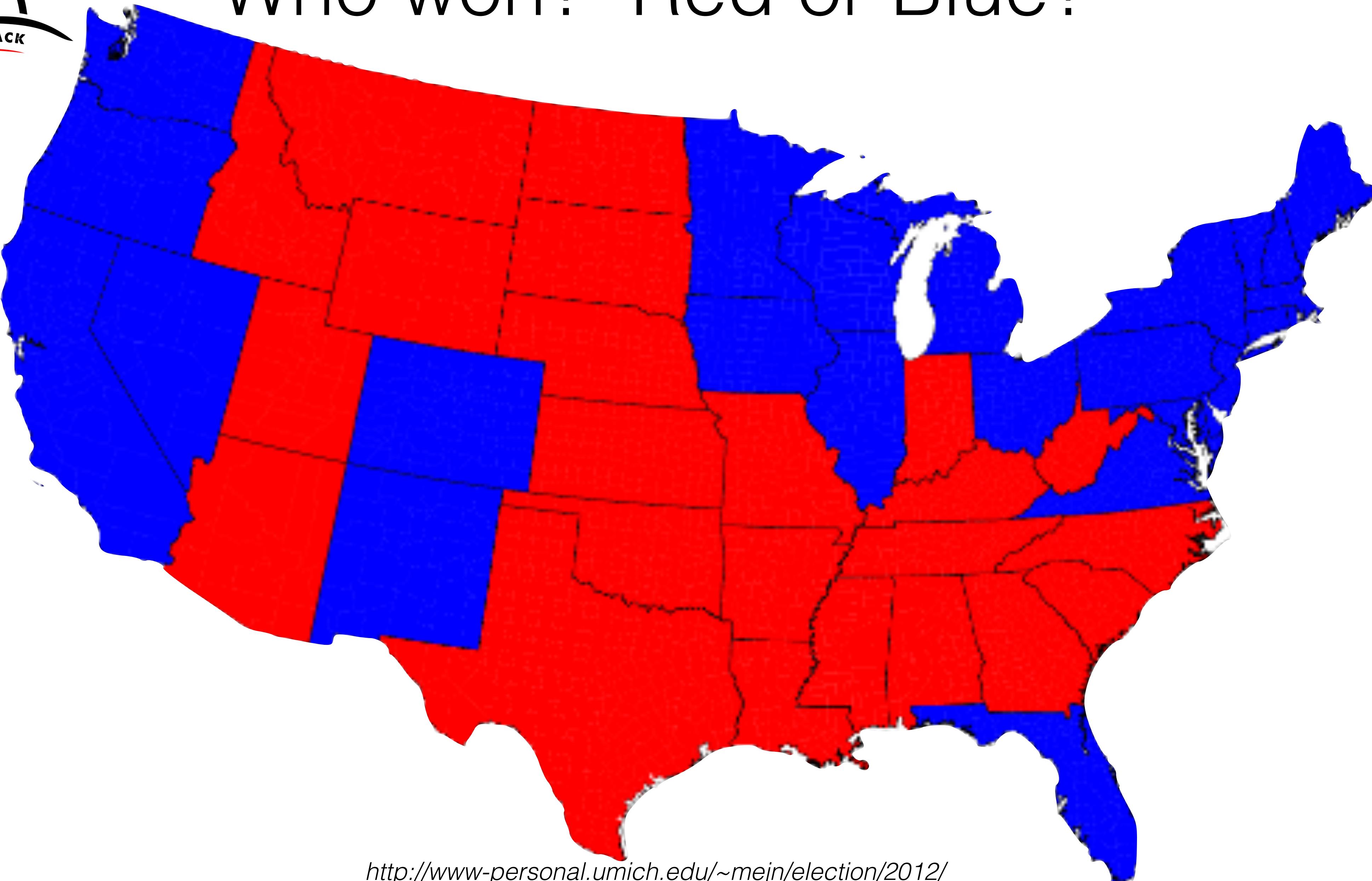
*** State health department, per its HIV data re-release agreement with CDC, requested not to release data to AIDSVu.



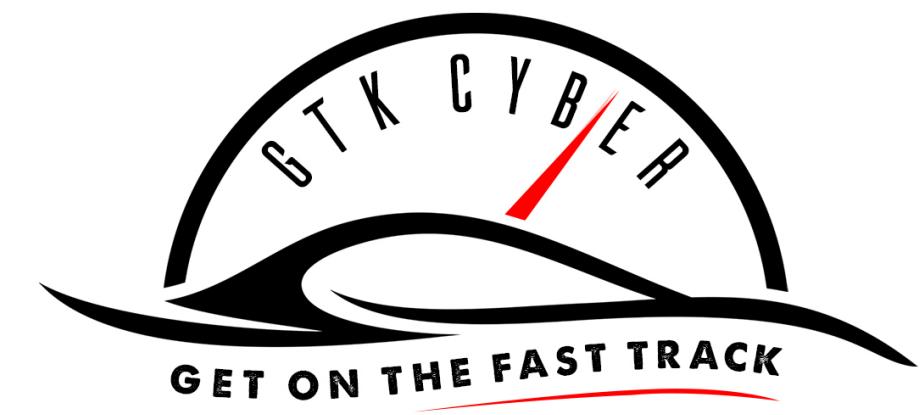
PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE
BASICALLY JUST POPULATION MAPS



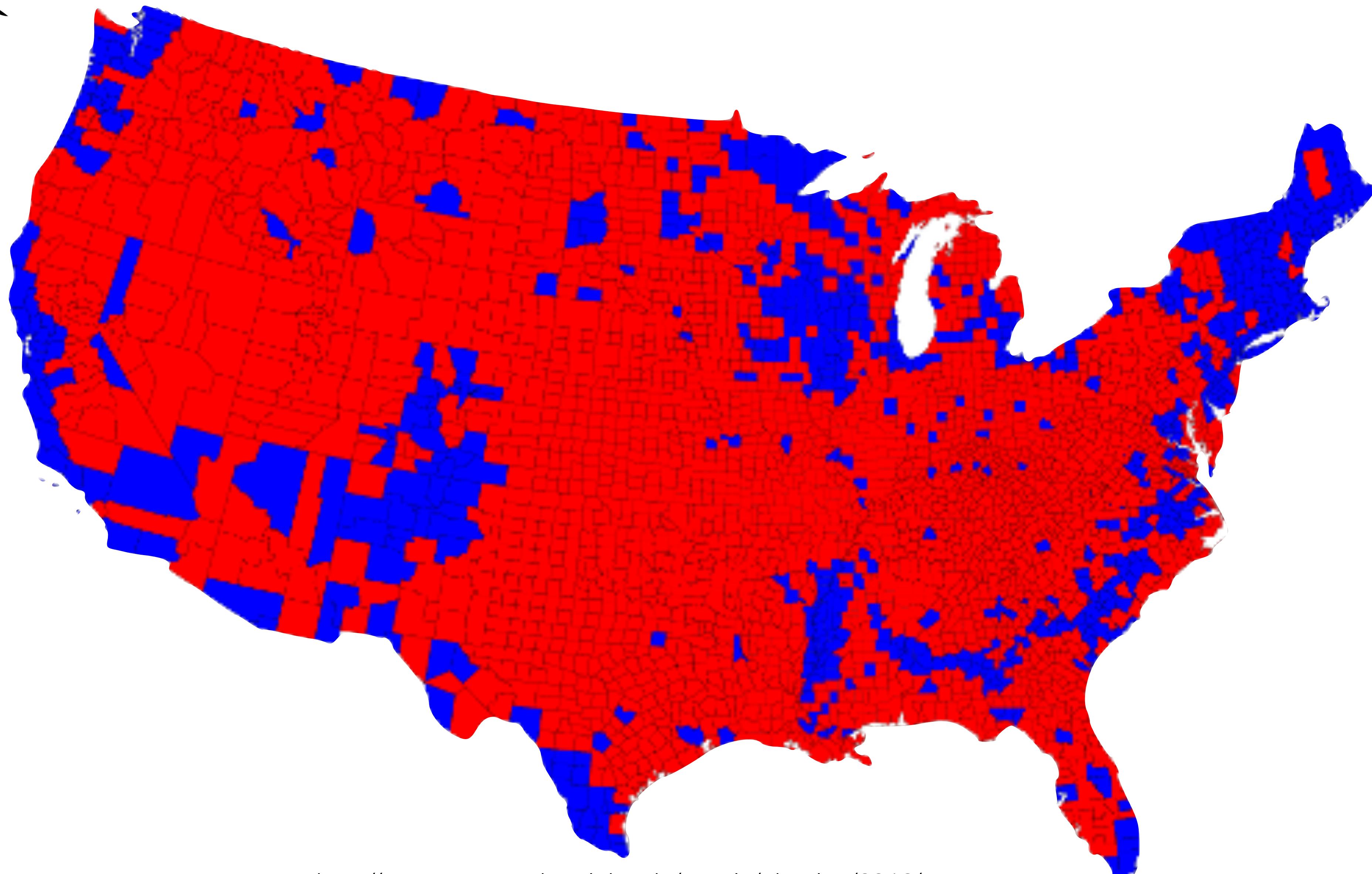
Who won? Red or Blue?



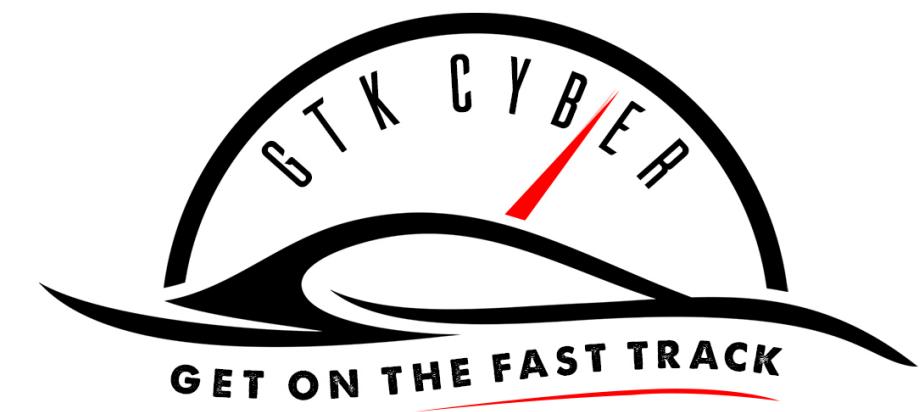
<http://www-personal.umich.edu/~mejn/election/2012/>



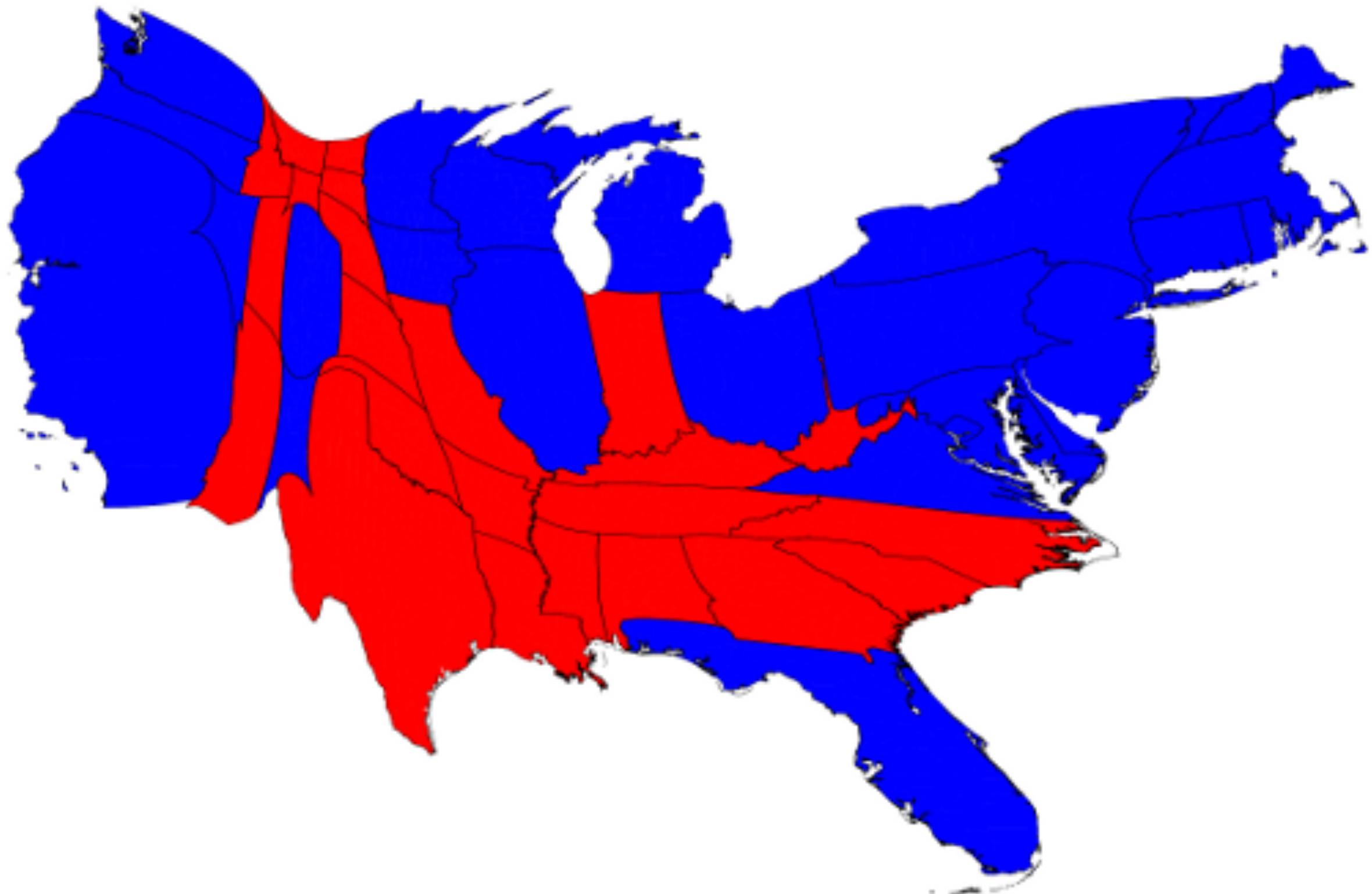
Who won? Red or Blue?



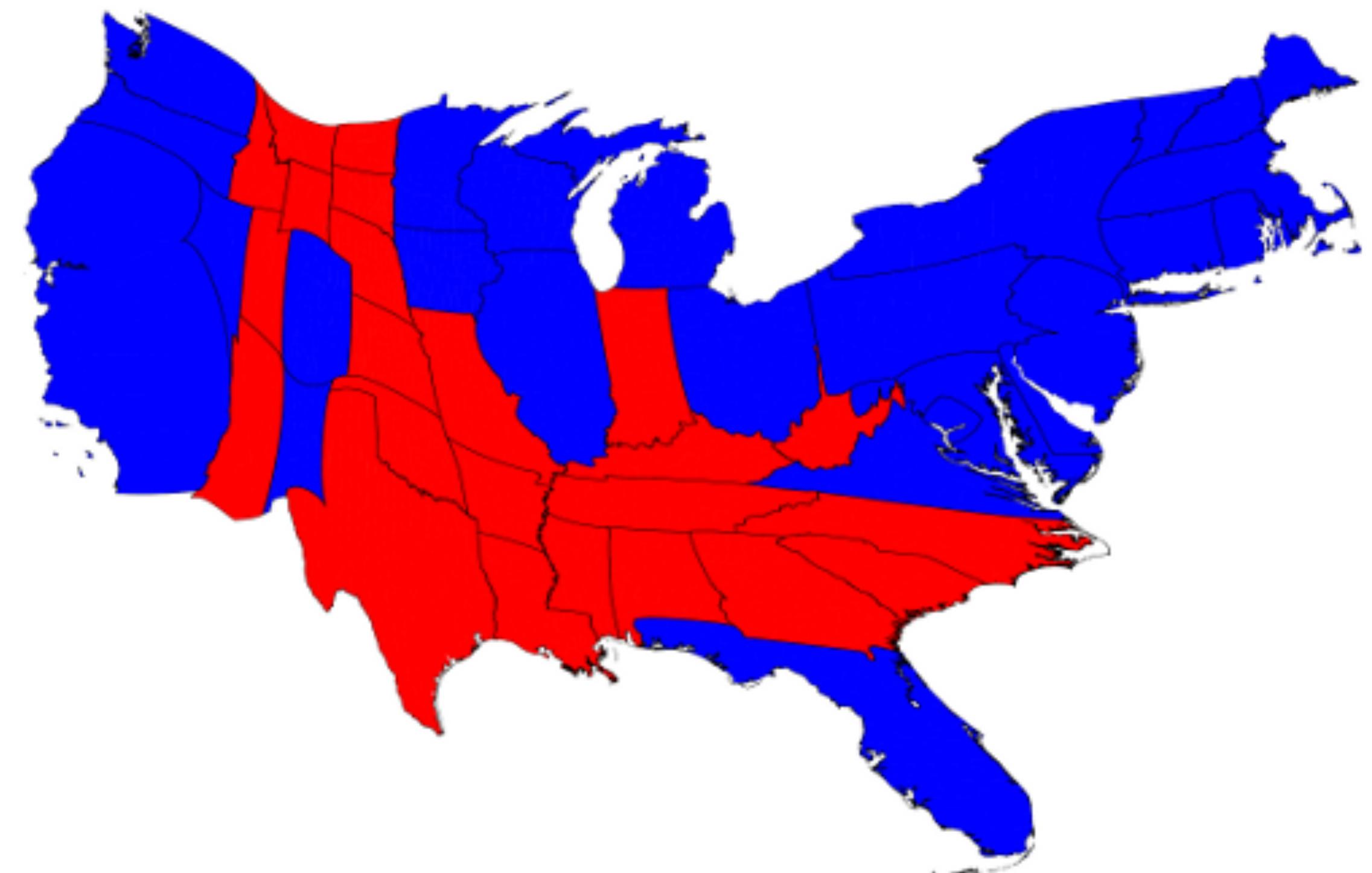
<http://www-personal.umich.edu/~mejn/election/2012/>



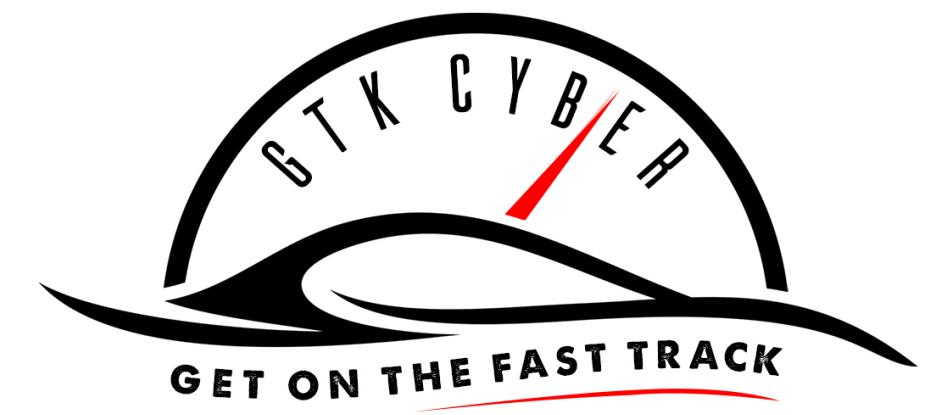
Who won? Red or Blue?



Scaled by Population



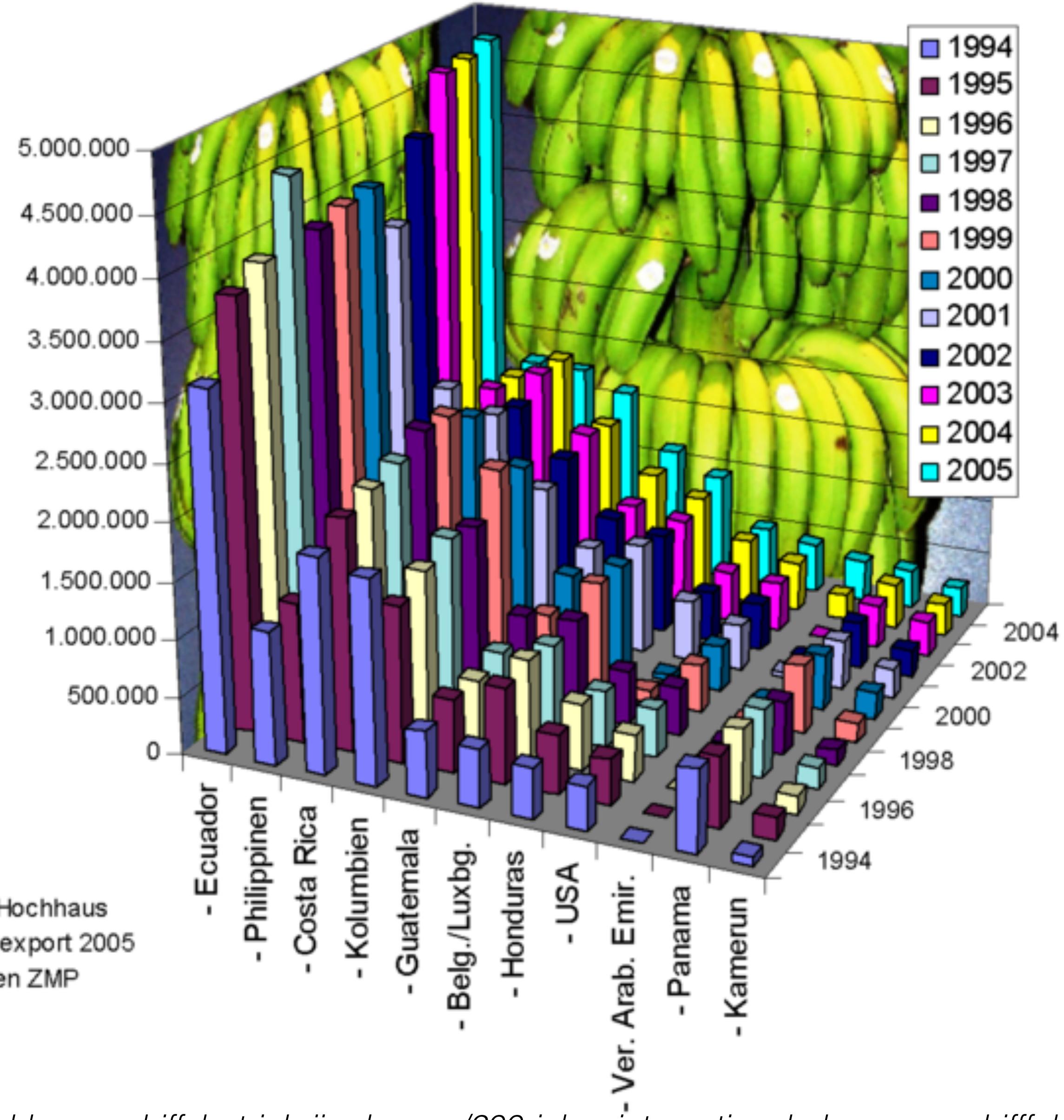
Scaled by Electoral Votes



The ugly



Export von Bananen in Tonnen von 1994-2005

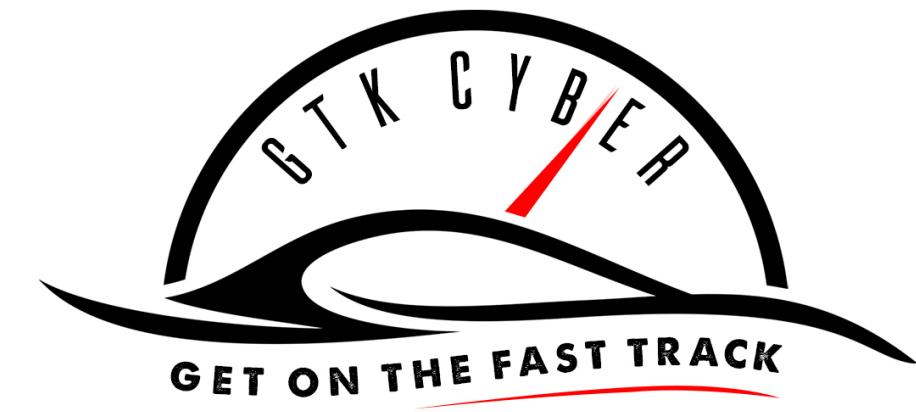


Dr. Hochhaus
Banlexport 2005
Daten ZMP

<https://hochhaus-schiffsbetrieb.jimdo.com/200-jahre-internationale-bananenschifffahrt-273/>

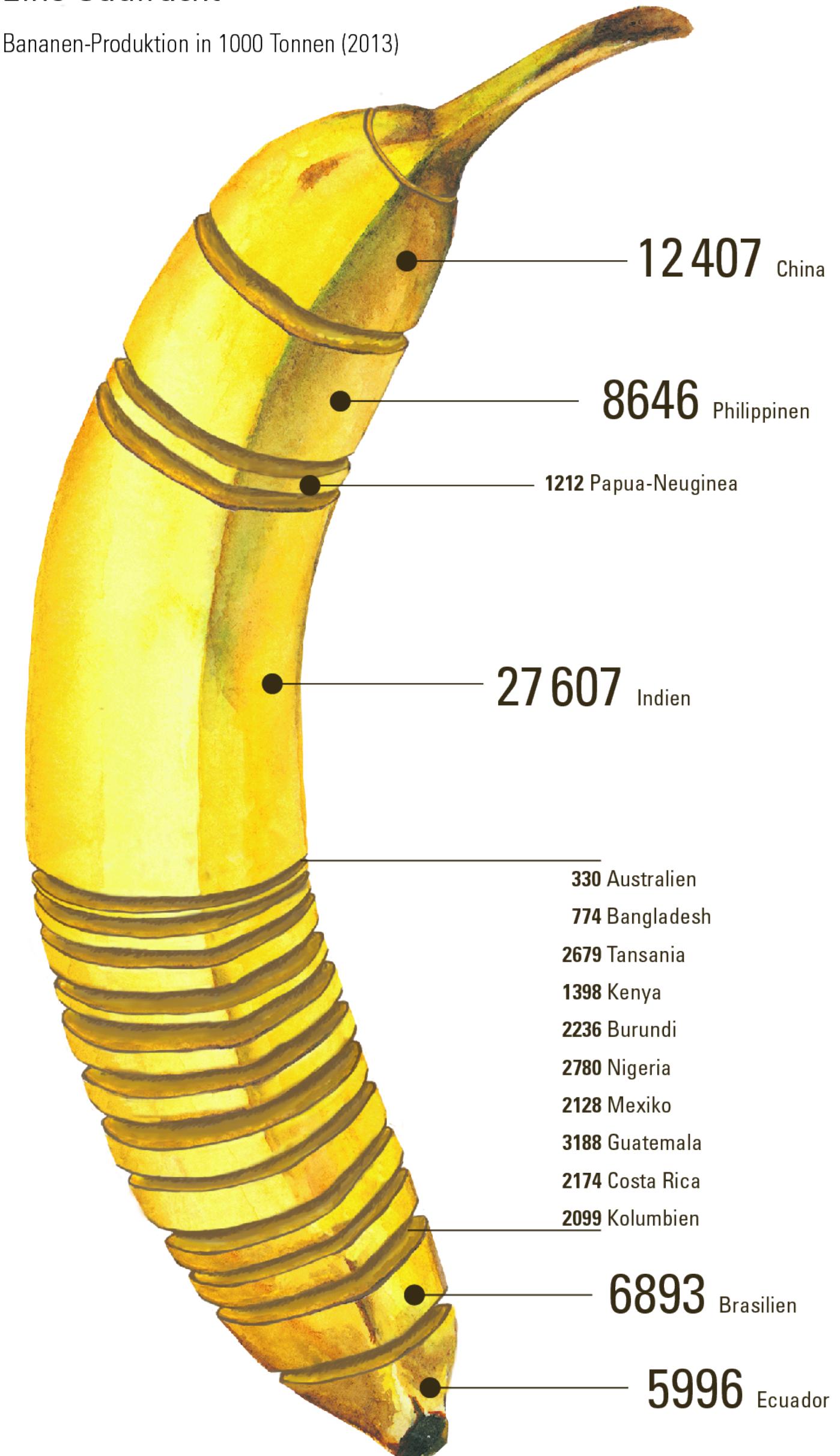
<https://excelcharts.com/change-bad-charts-in-the-wikipedia/>

<http://www.fao.org/faostat/en/#data/QC>



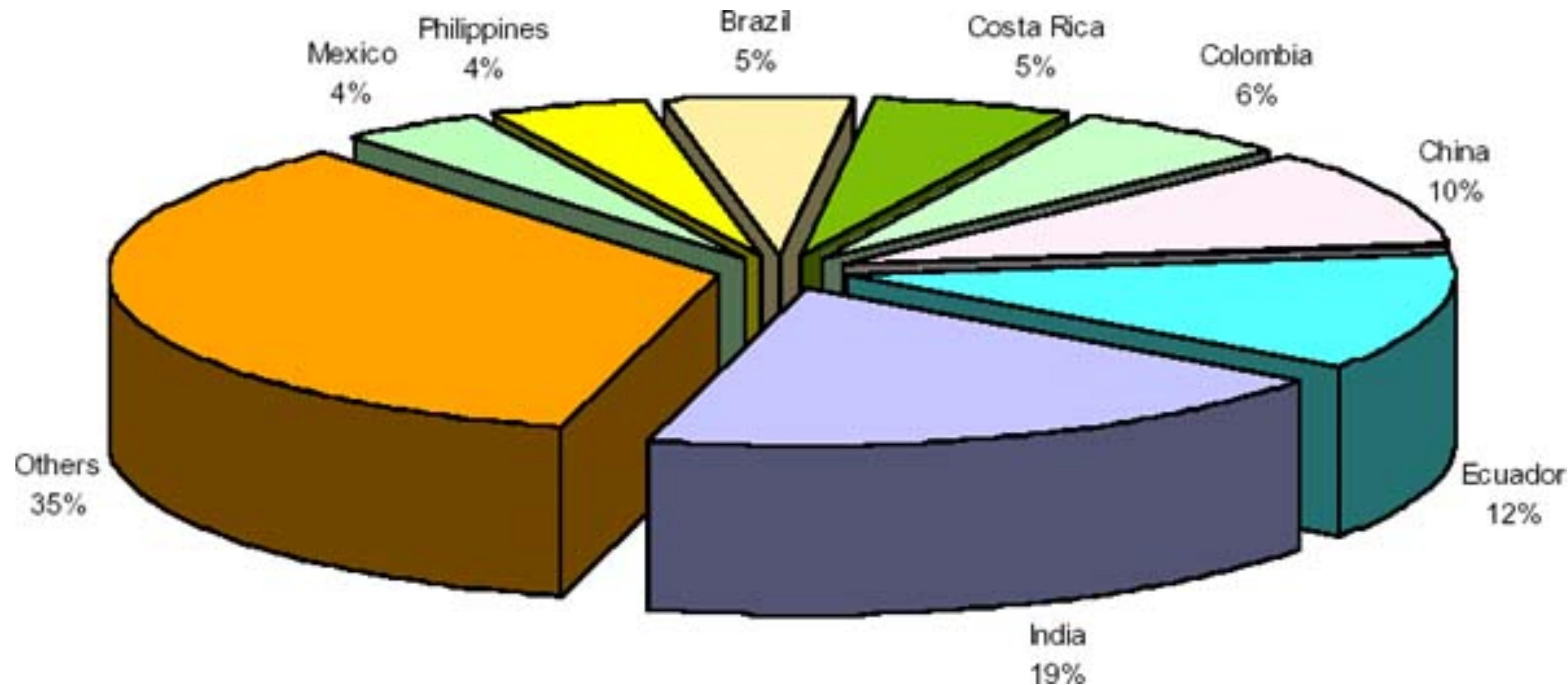
Eine Südfrucht

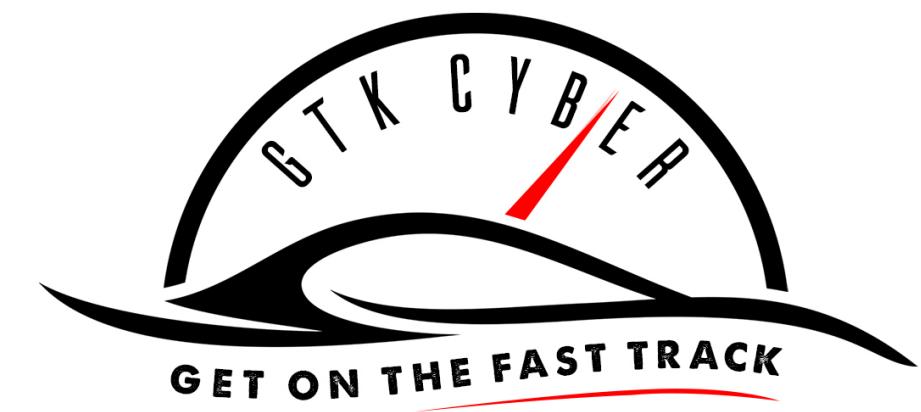
Bananen-Produktion in 1000 Tonnen (2013)



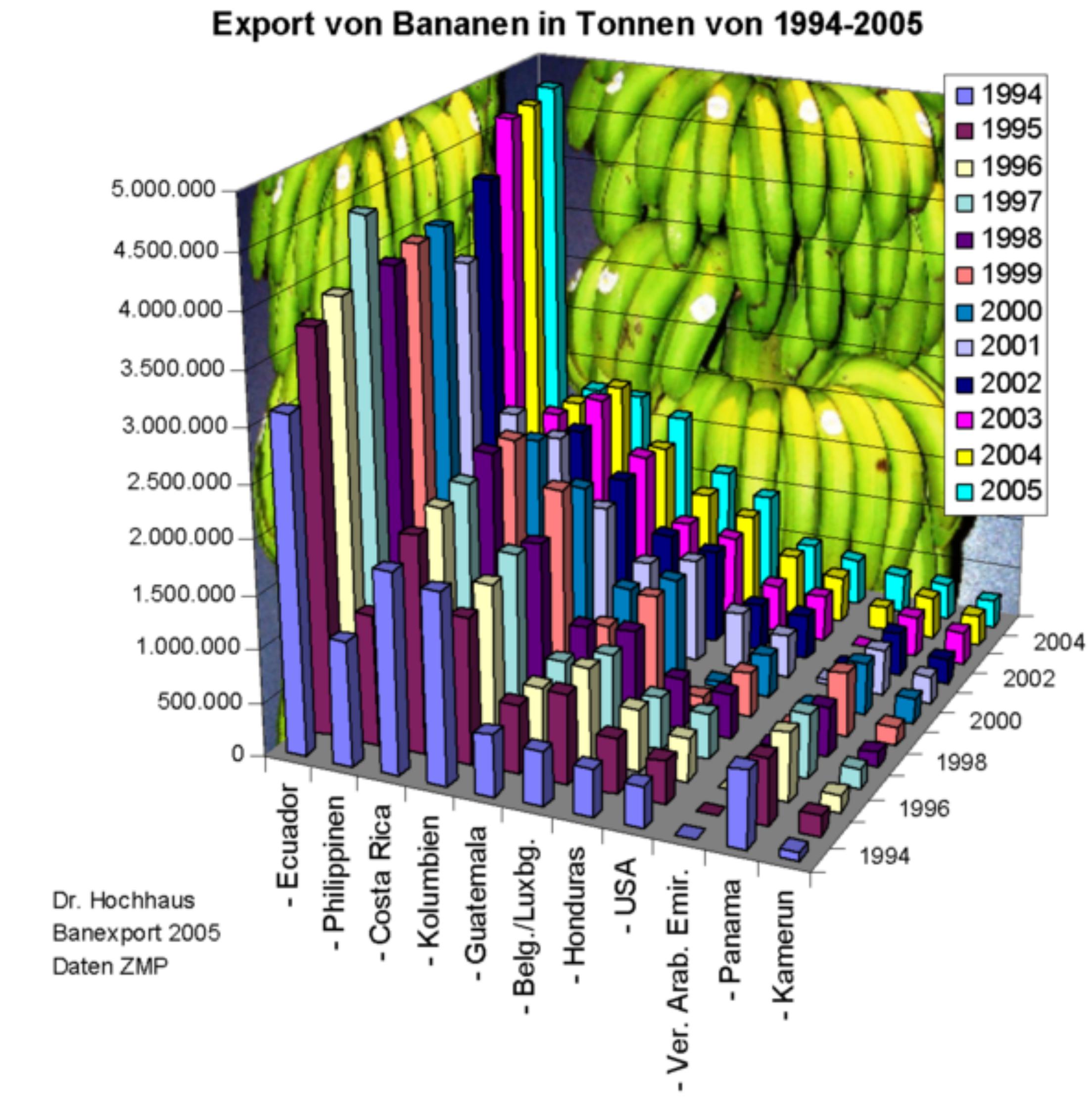
QUELLE: FAOSTAT

NZZ-Infografik/lea.



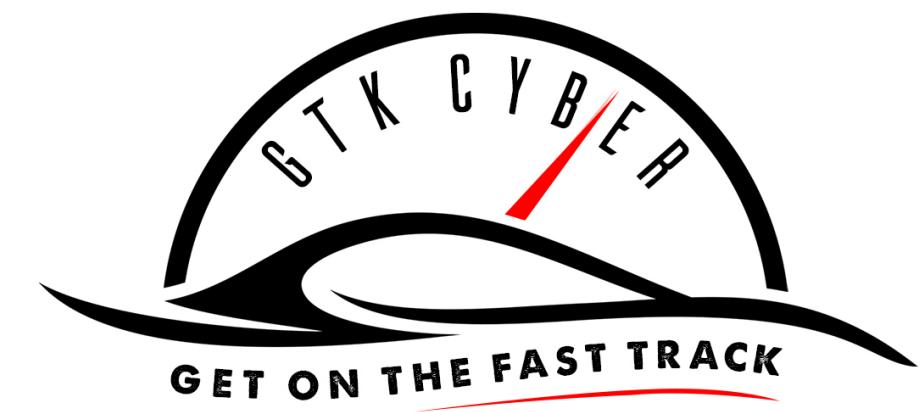


How could we fix this?

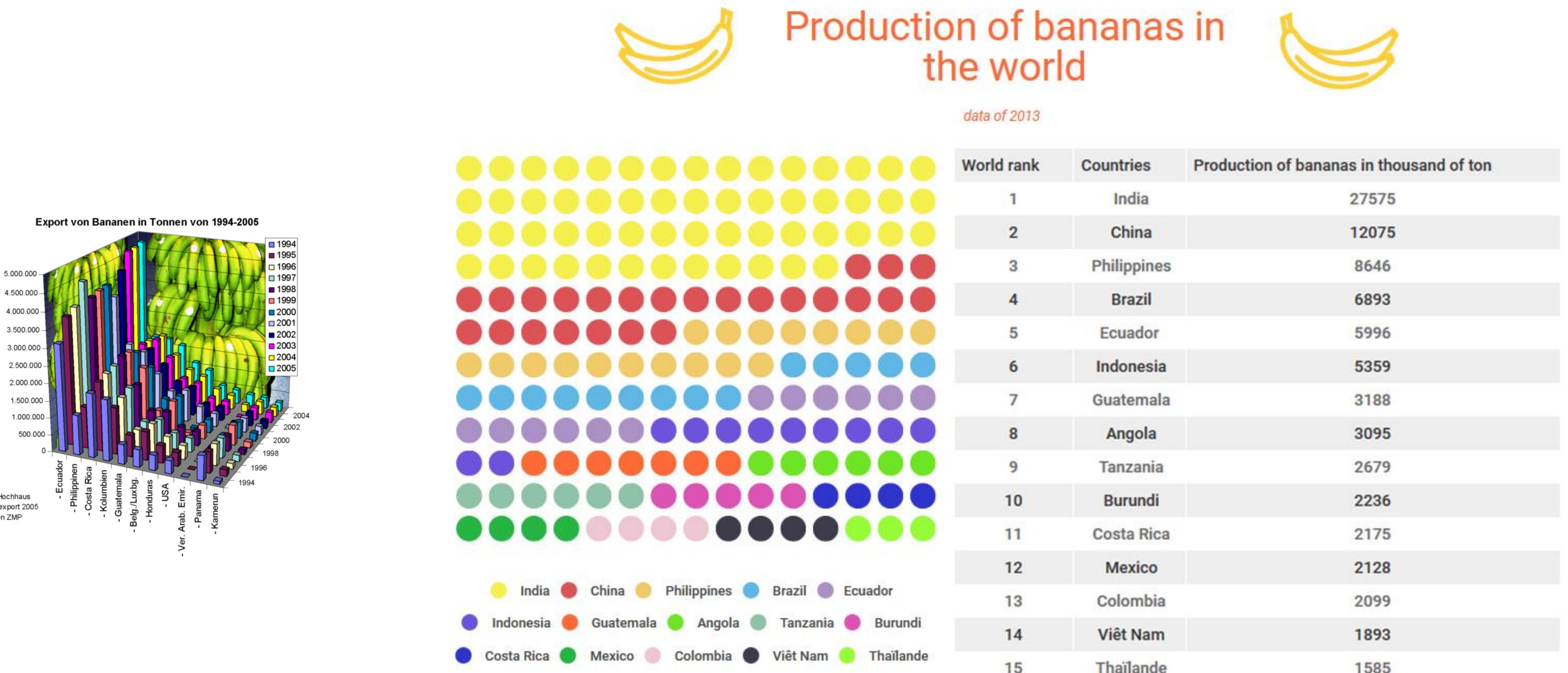


<https://hochhaus-schiffsbetrieb.jimdo.com/200-jahre-internationale-bananenschifffahrt-273/>

<https://excelcharts.com/change-bad-charts-in-the-wikipedia/>

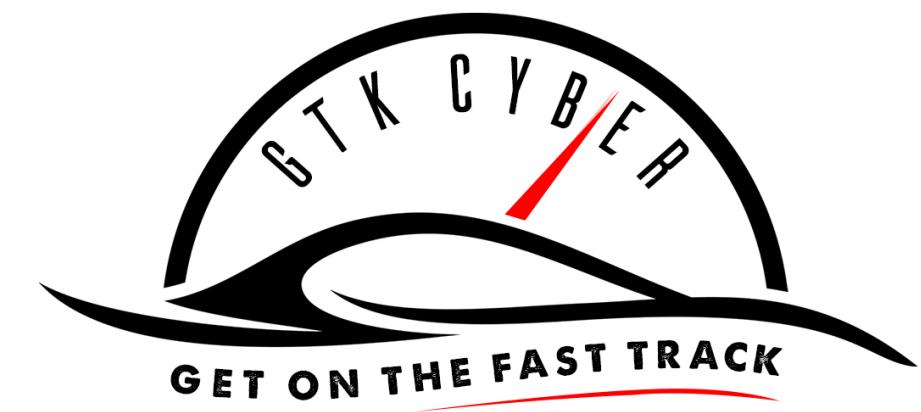


How could we fix this?

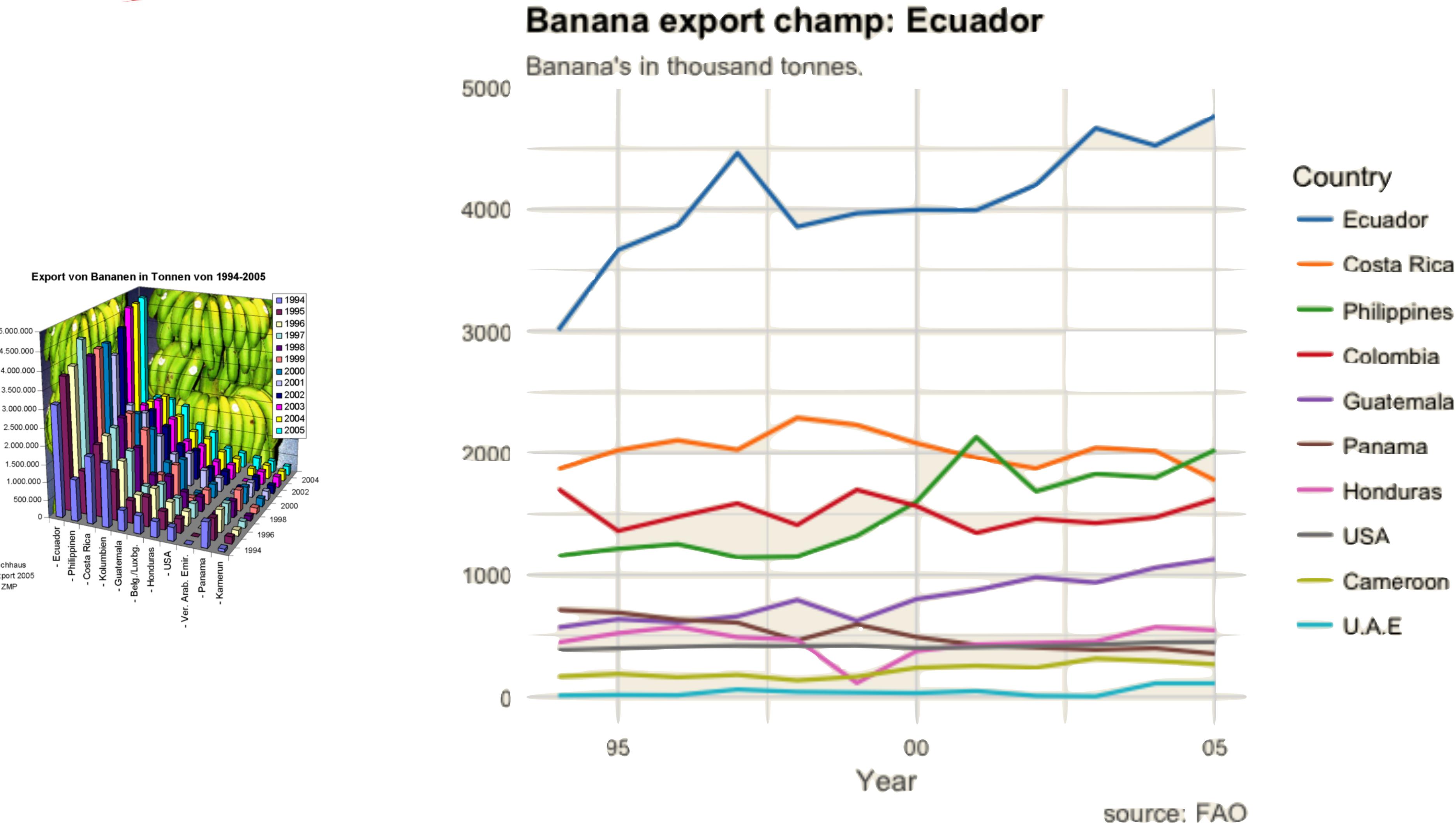


Post by PacoJoke

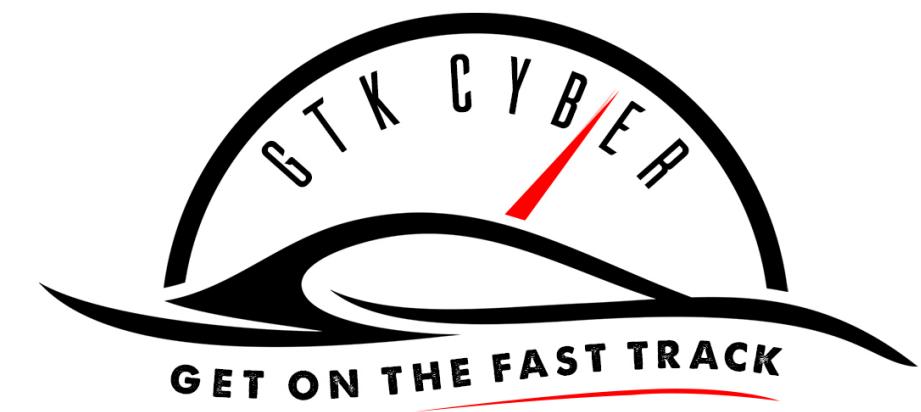
https://www.reddit.com/r/dataisbeautiful/comments/6dv56j/production_of_bananas_in_the_world_oc/



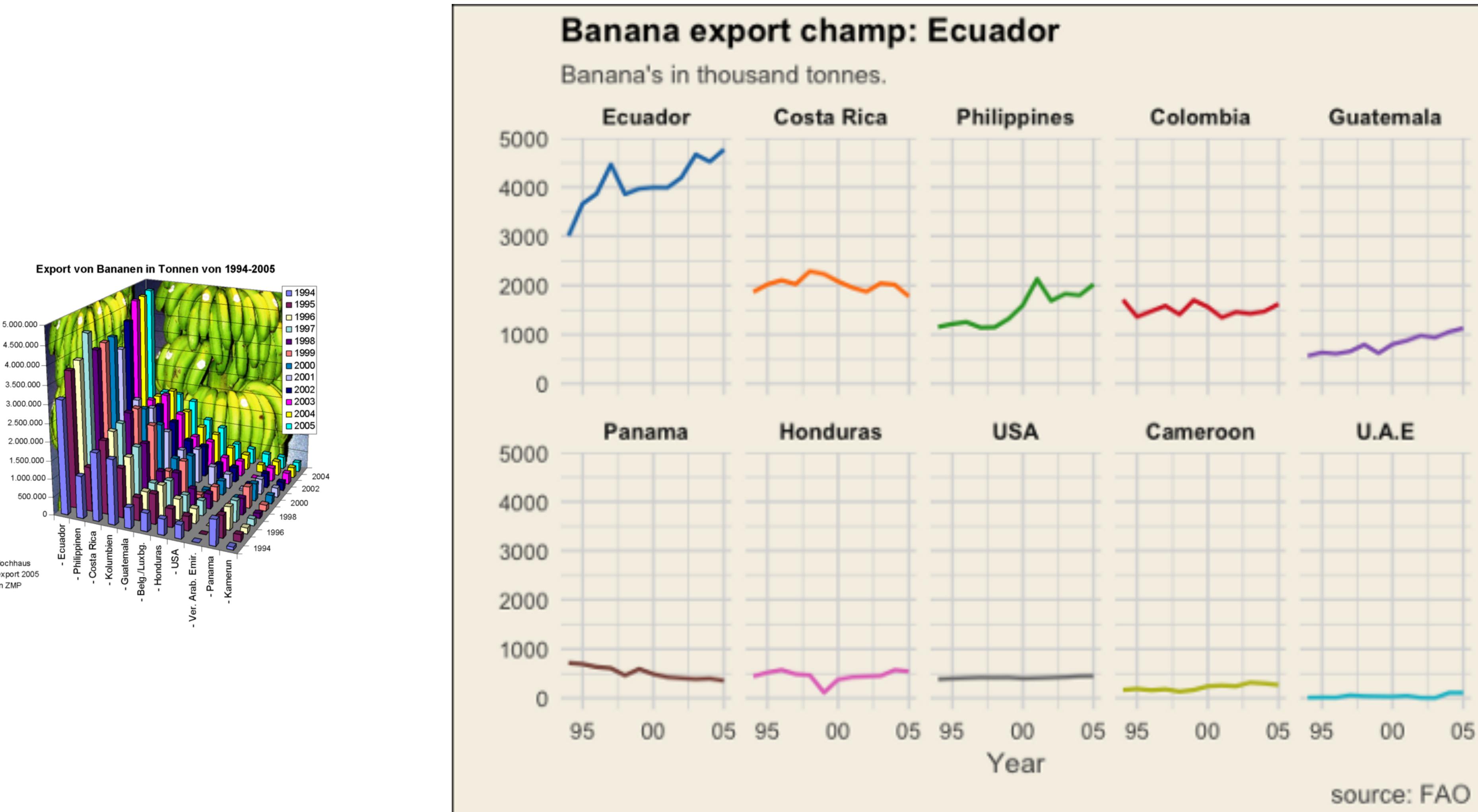
How could we fix this?



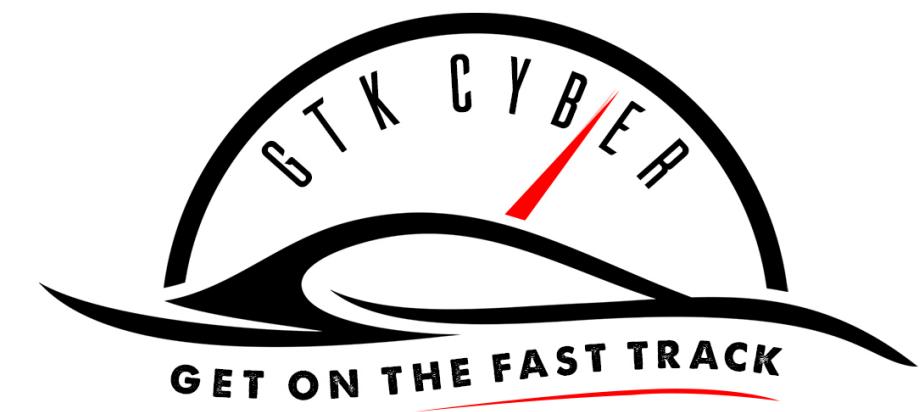
<https://medium.com/tdebeus/redesign-of-a-truly-bananas-chart-1617f930808d>



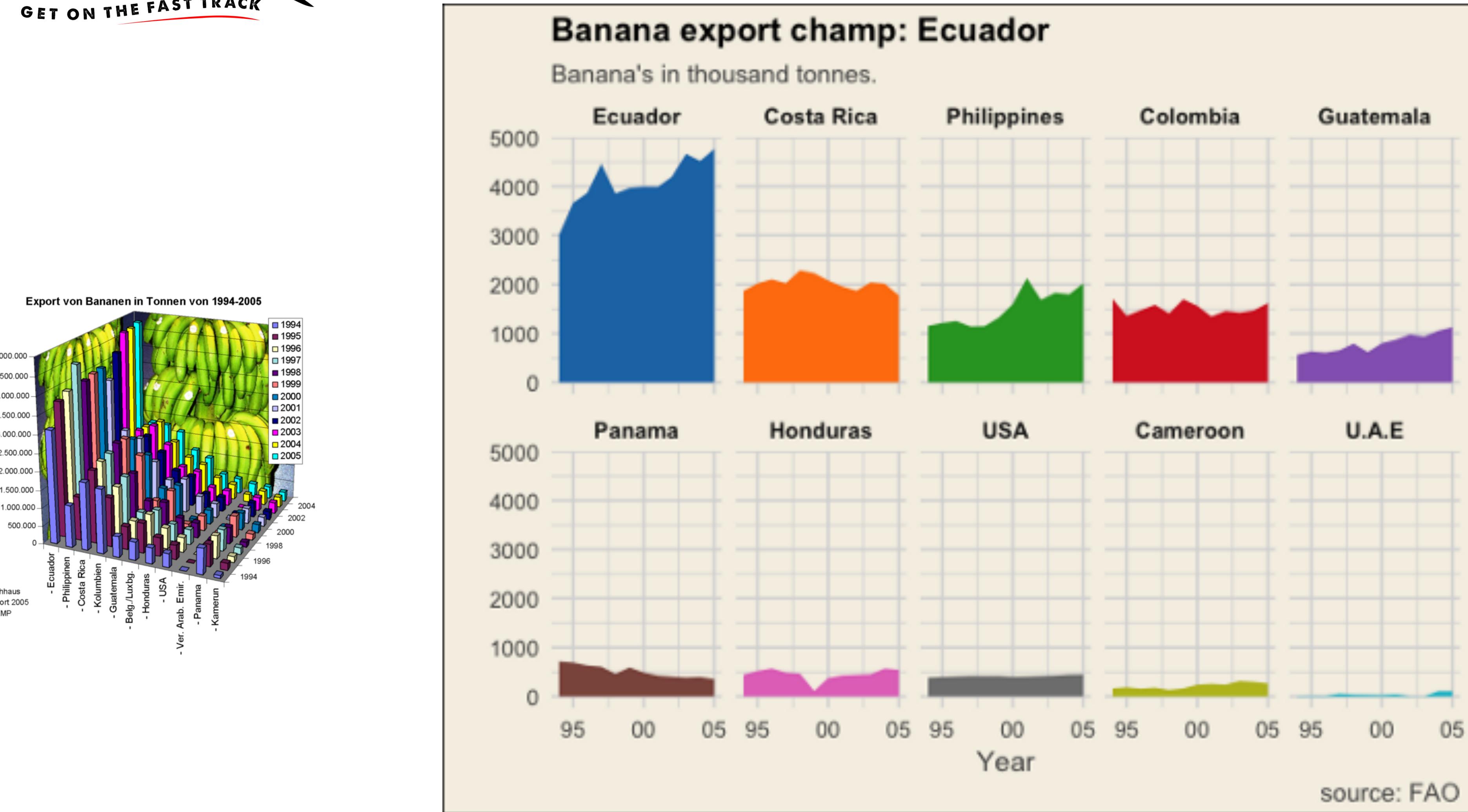
How could we fix this? Small Multiples



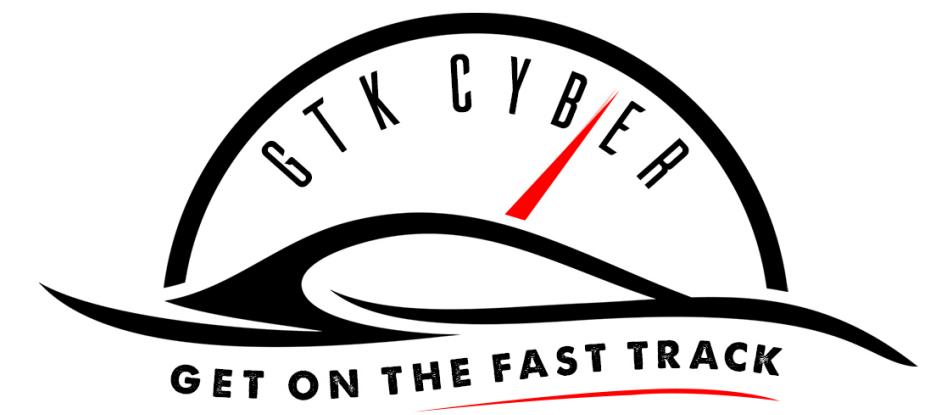
<https://medium.com/tdebeus/redesign-of-a-truly-bananas-chart-1617f930808d>



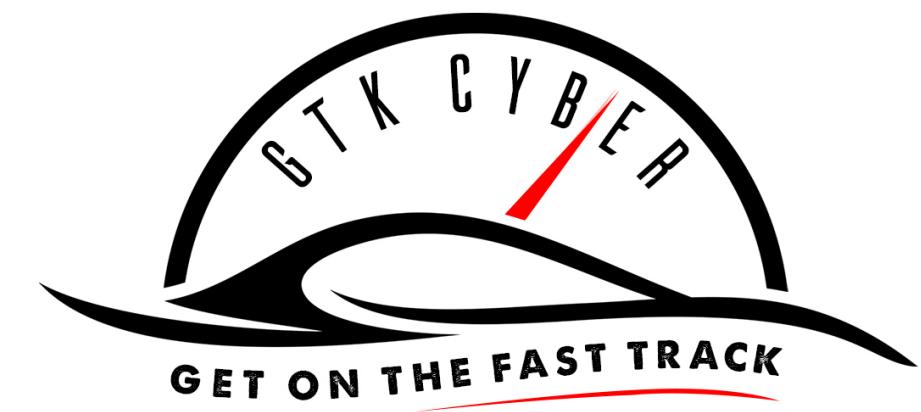
How could we fix this? Small Multiples



<https://medium.com/tdebeus/redesign-of-a-truly-bananas-chart-1617f930808d>

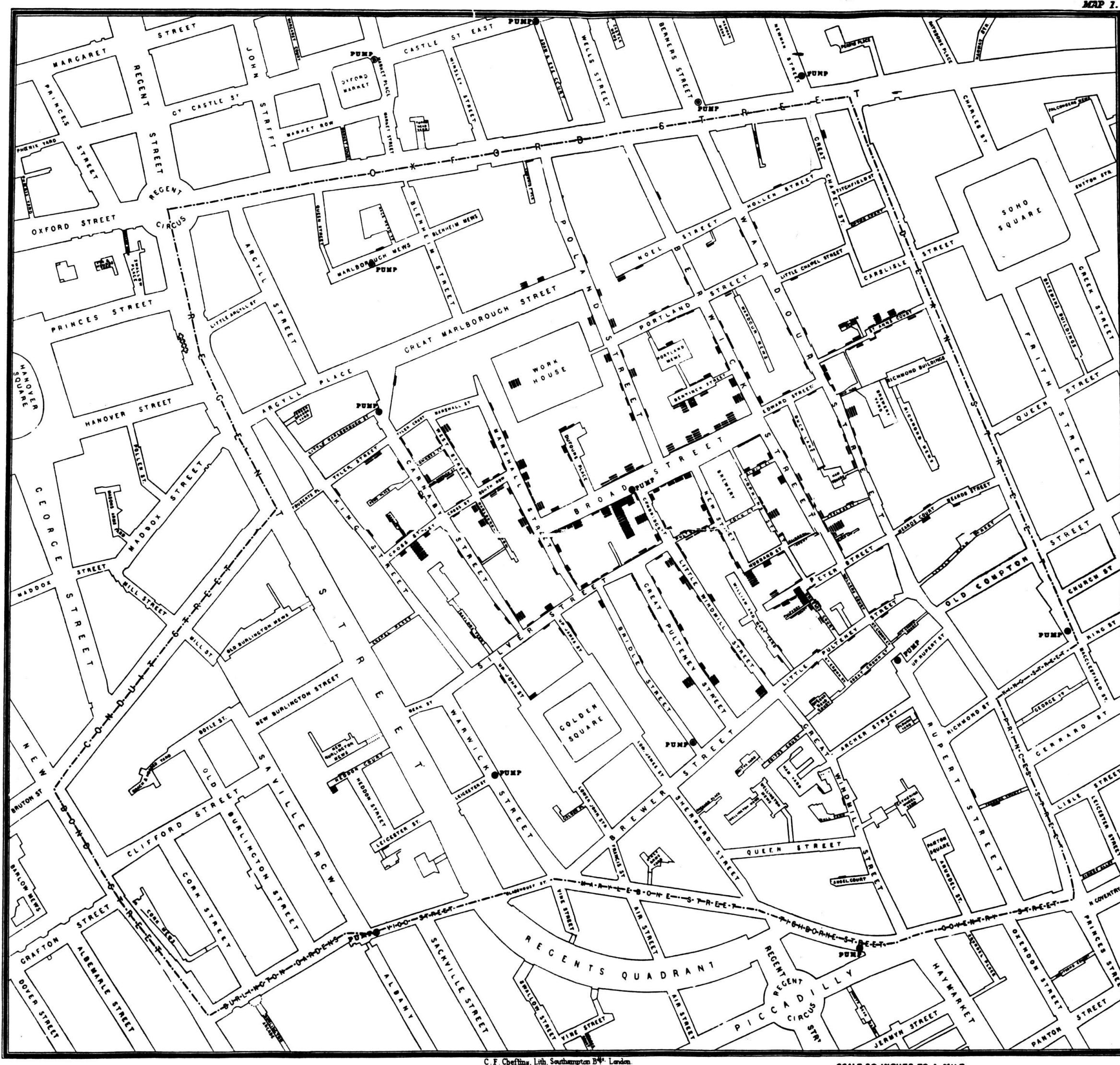


The good



Proper Display





- Created by John Snow in 1854
- This visualization proved that cholera was water-borne
- Snow identified the source of the outbreak as a water pump on Broad Street
- Spurred city to create a true sewage system

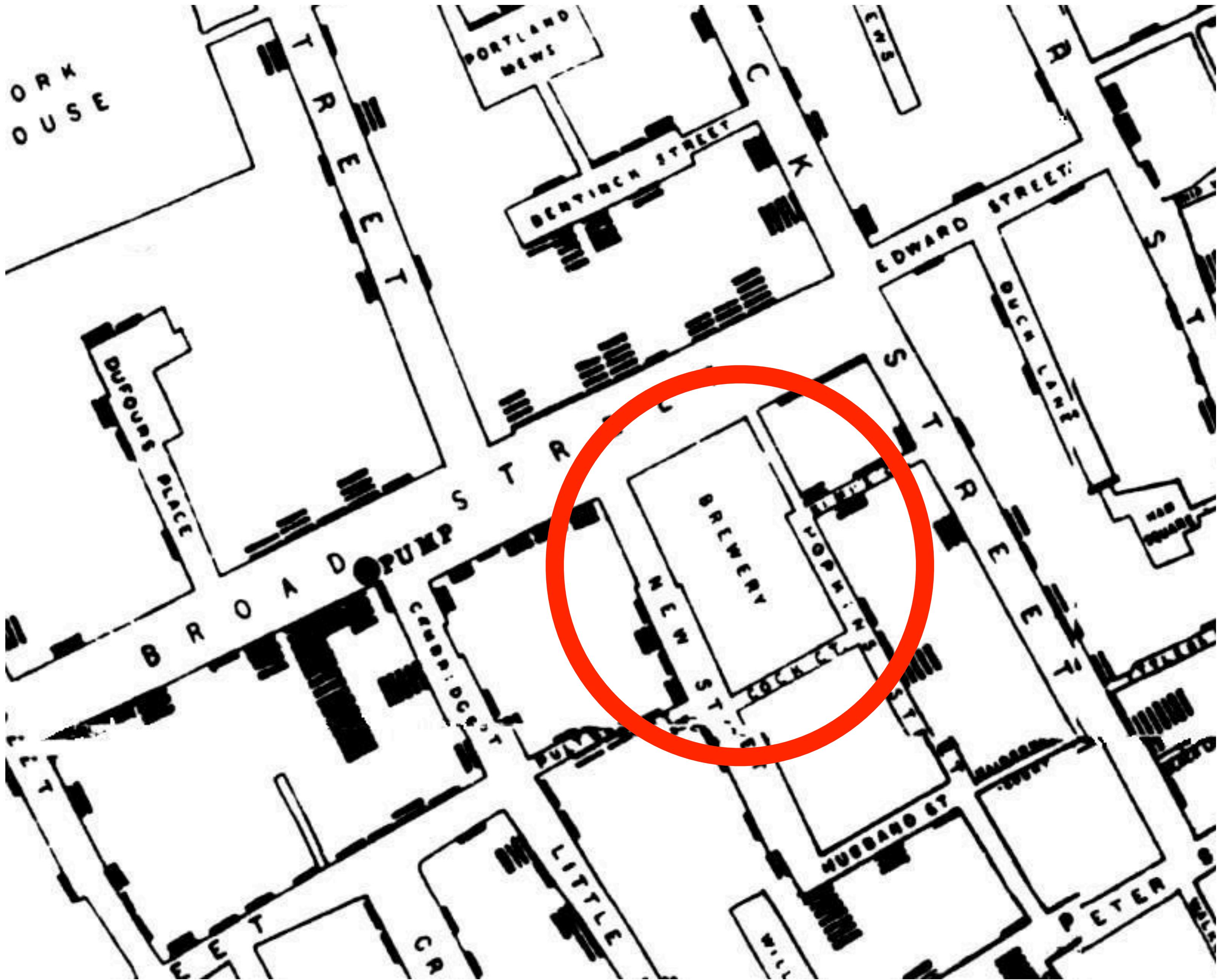
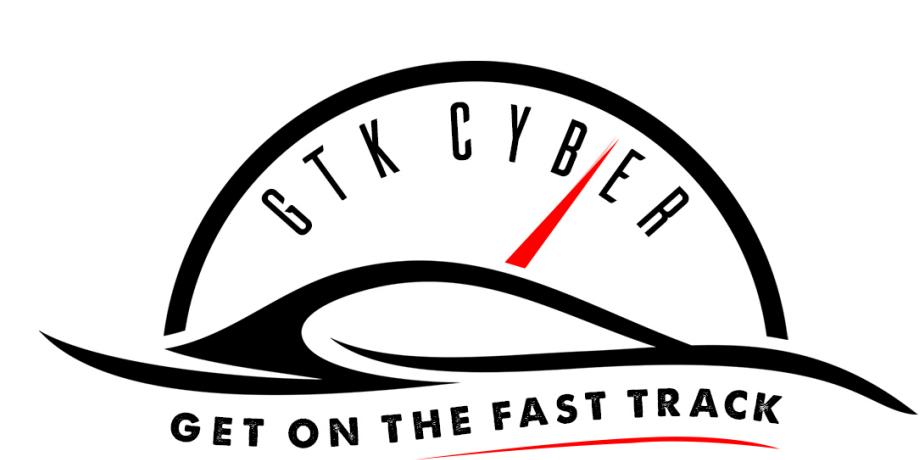
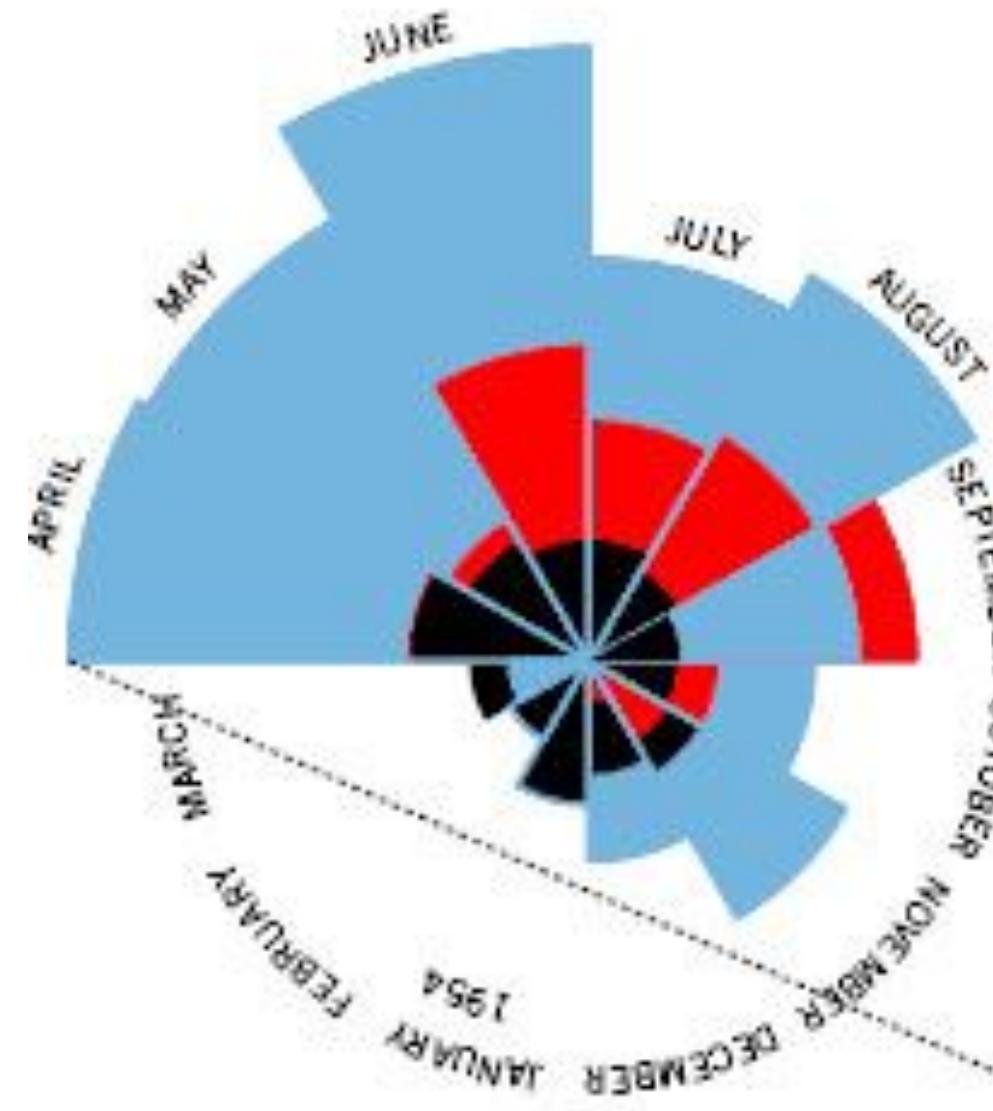




DIAGRAM OF THE CAUSES OF MORTALITY IN THE ARMY IN THE EAST

2.
APRIL 1855 TO MARCH 1856



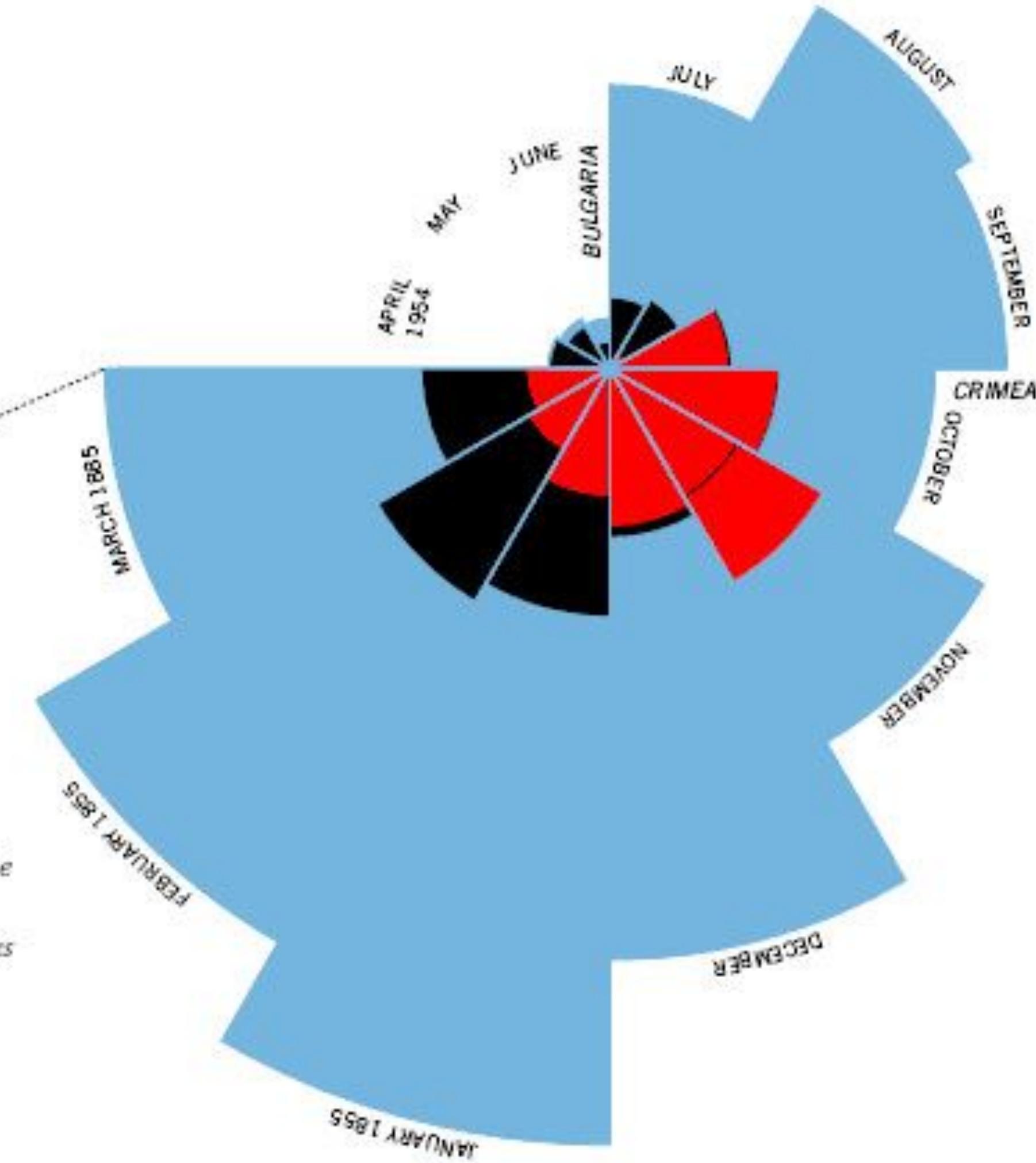
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex

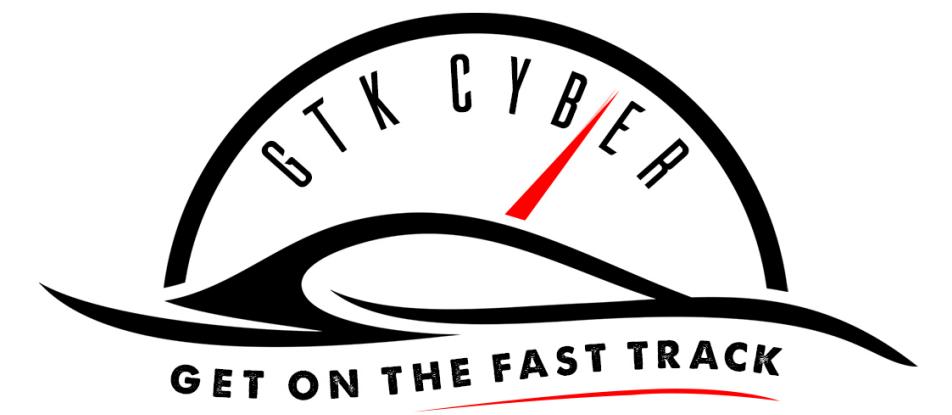
*The blue wedges measured from the centre of the circle represent area
for area the deaths from Preventible or Mitigable Zymotic Diseases, the
red wedges measured from the centre the deaths from wounds, & the
black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov' 1854 marks the boundary
of the deaths from all other causes during the month.*

*In October 1854, & April 1855, the black area coincides with the red;
in January & February 1856, the blue coincides with the black*

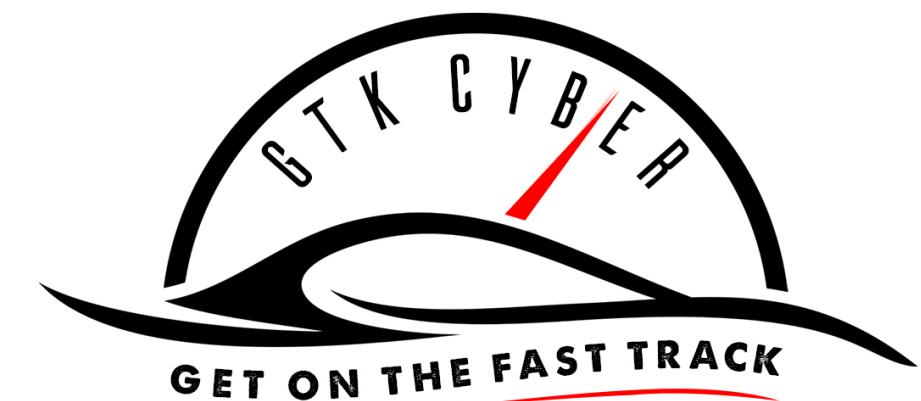
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

1.
APRIL 1854 TO MARCH 1855





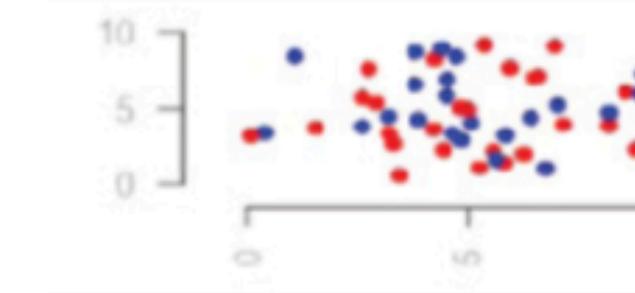
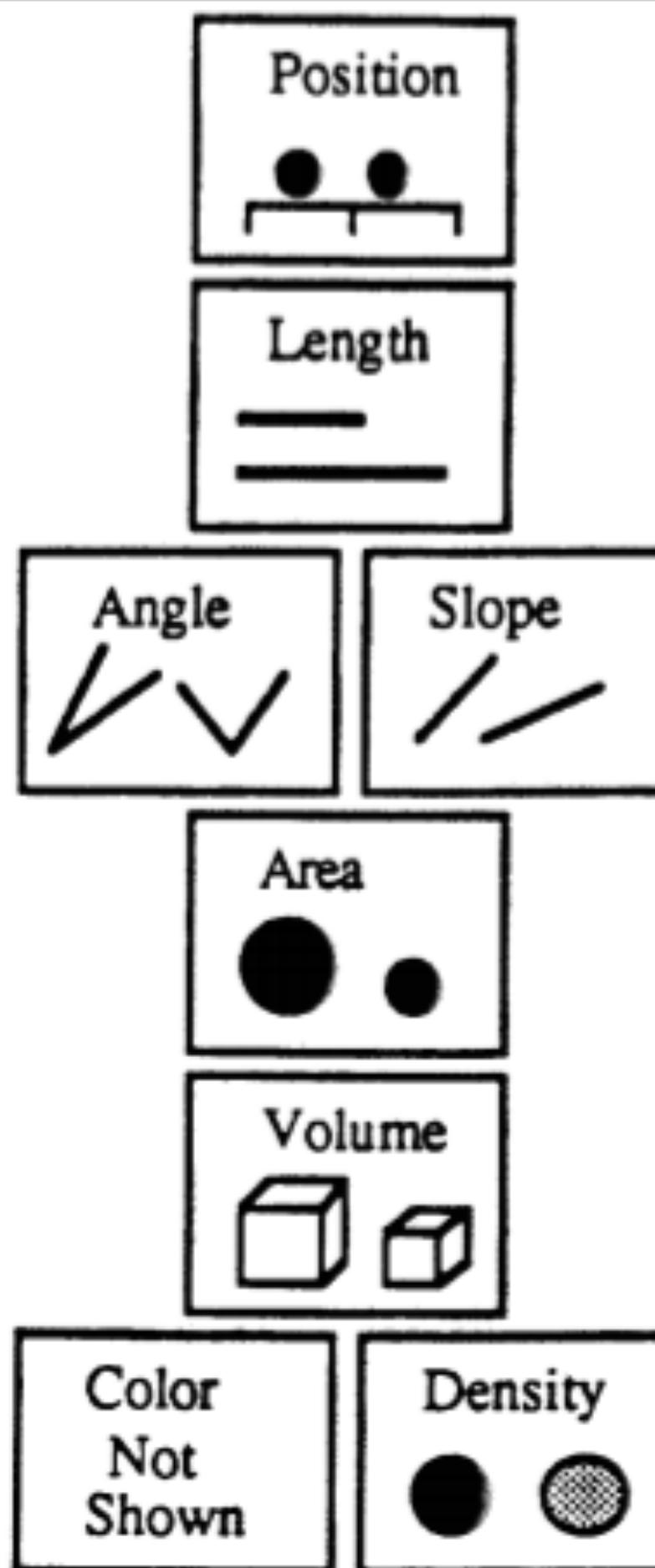
Visual Encoding



More accurate



Less accurate



Position



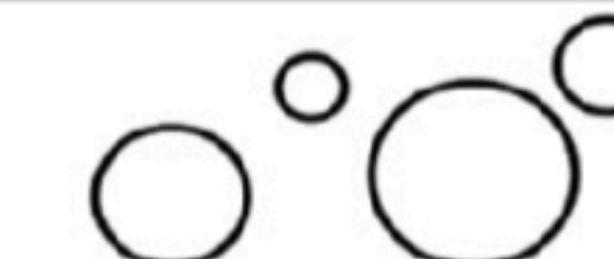
Length



Direction/Slope



Angle



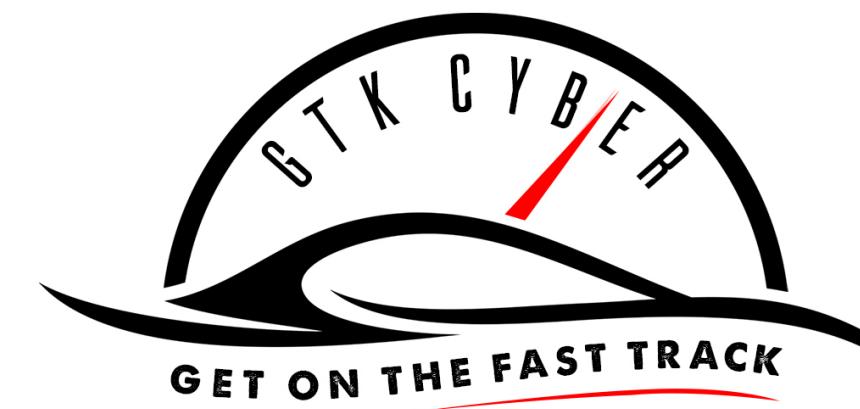
Area



Density/Saturation

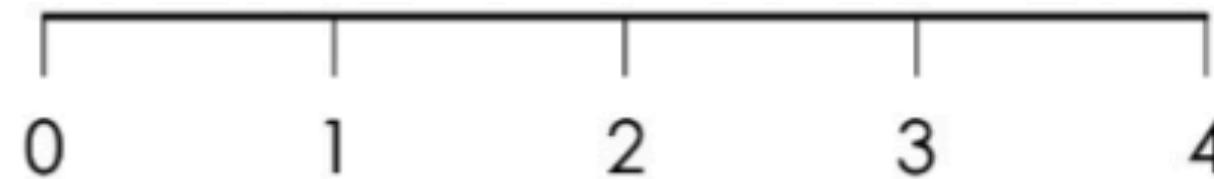


Color Hue



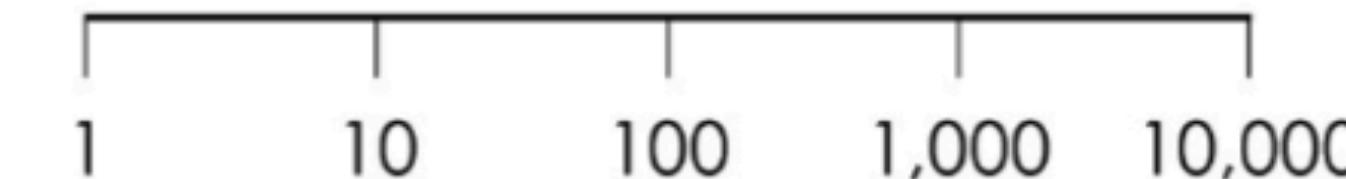
Linear

Values are evenly spaced



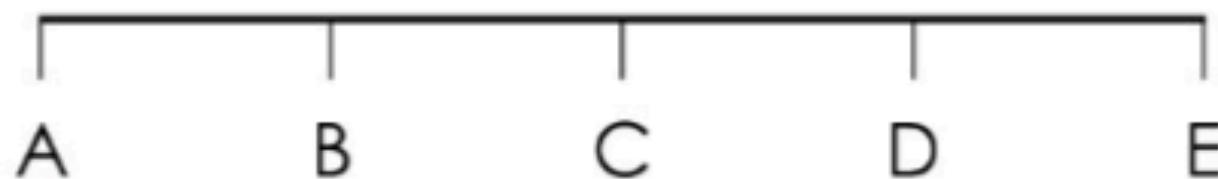
Logarithmic

Focus on percent change



Categorical

Discrete placement in bins



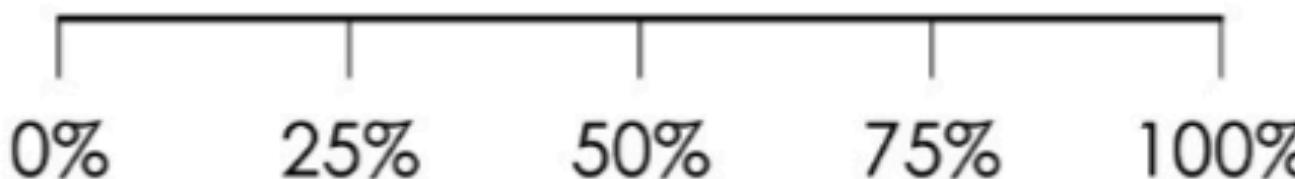
Ordinal

Categories where order matters



Percent

Representing parts of a whole

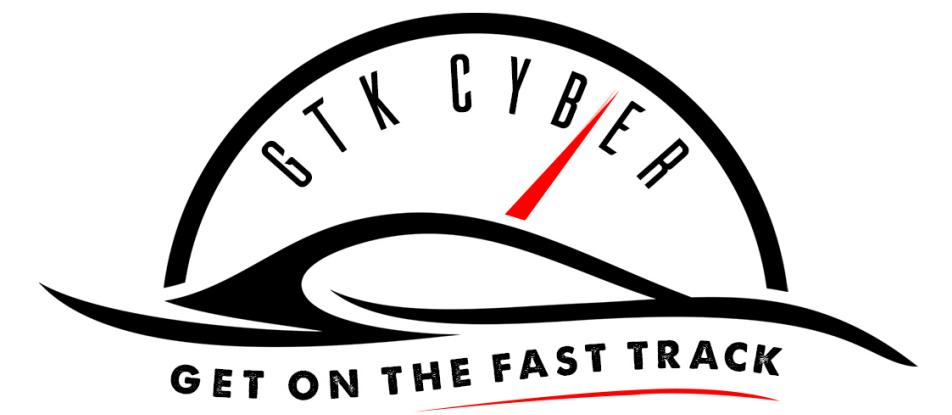


Time

Units of months, days, or hours



Source: Nathan Yau, Data Points

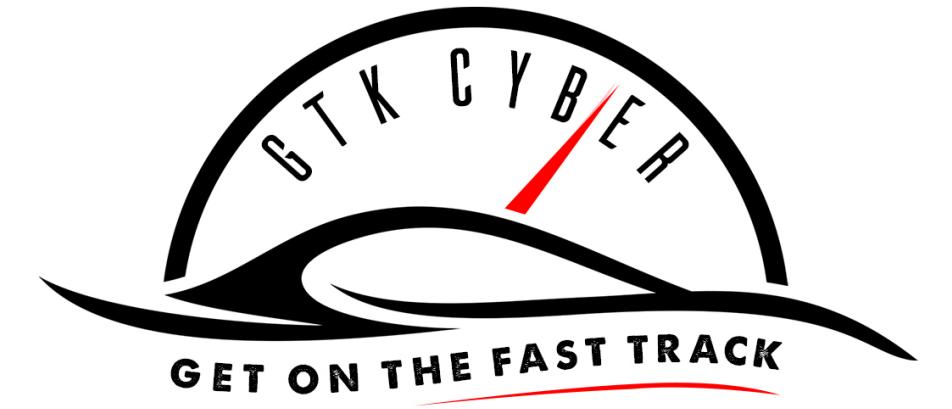


Color

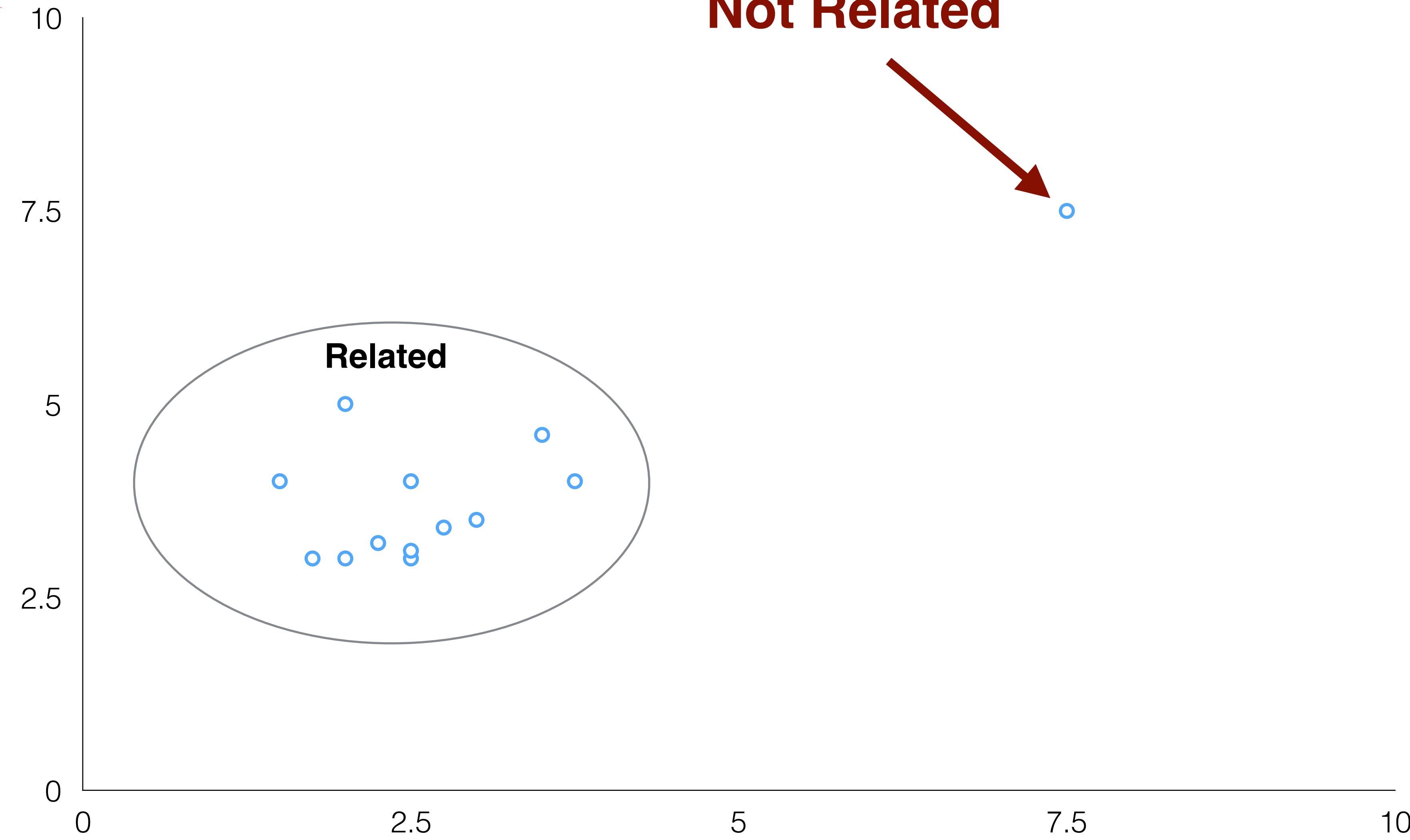
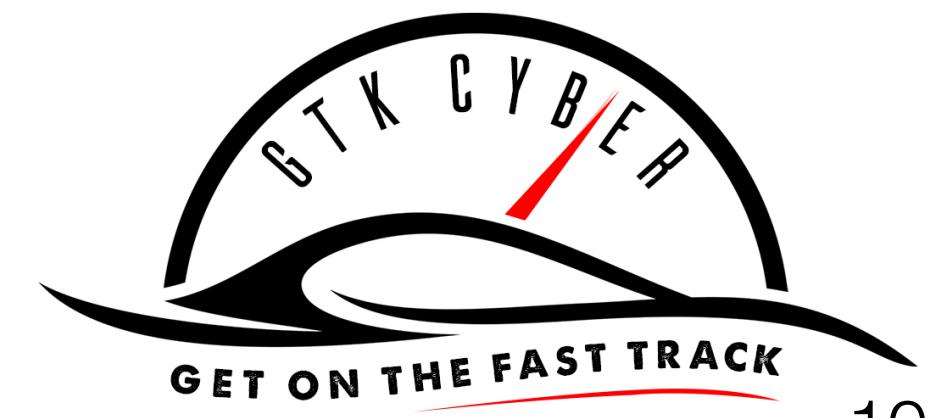


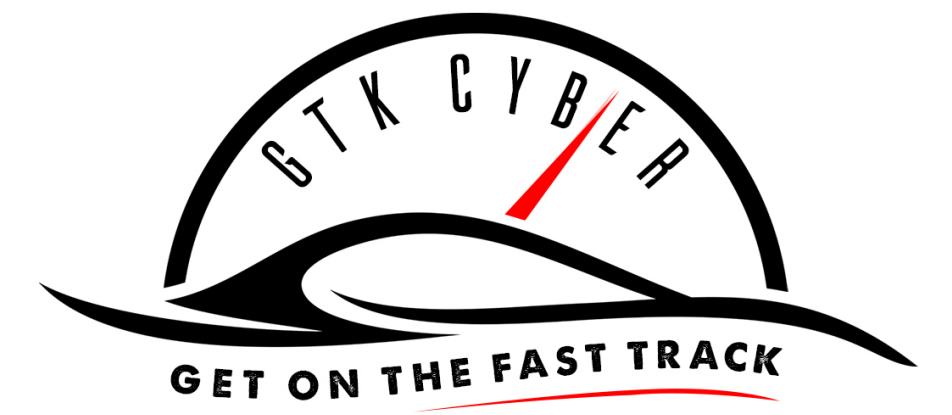


Visual Encodings: Translating Data into Visual Form

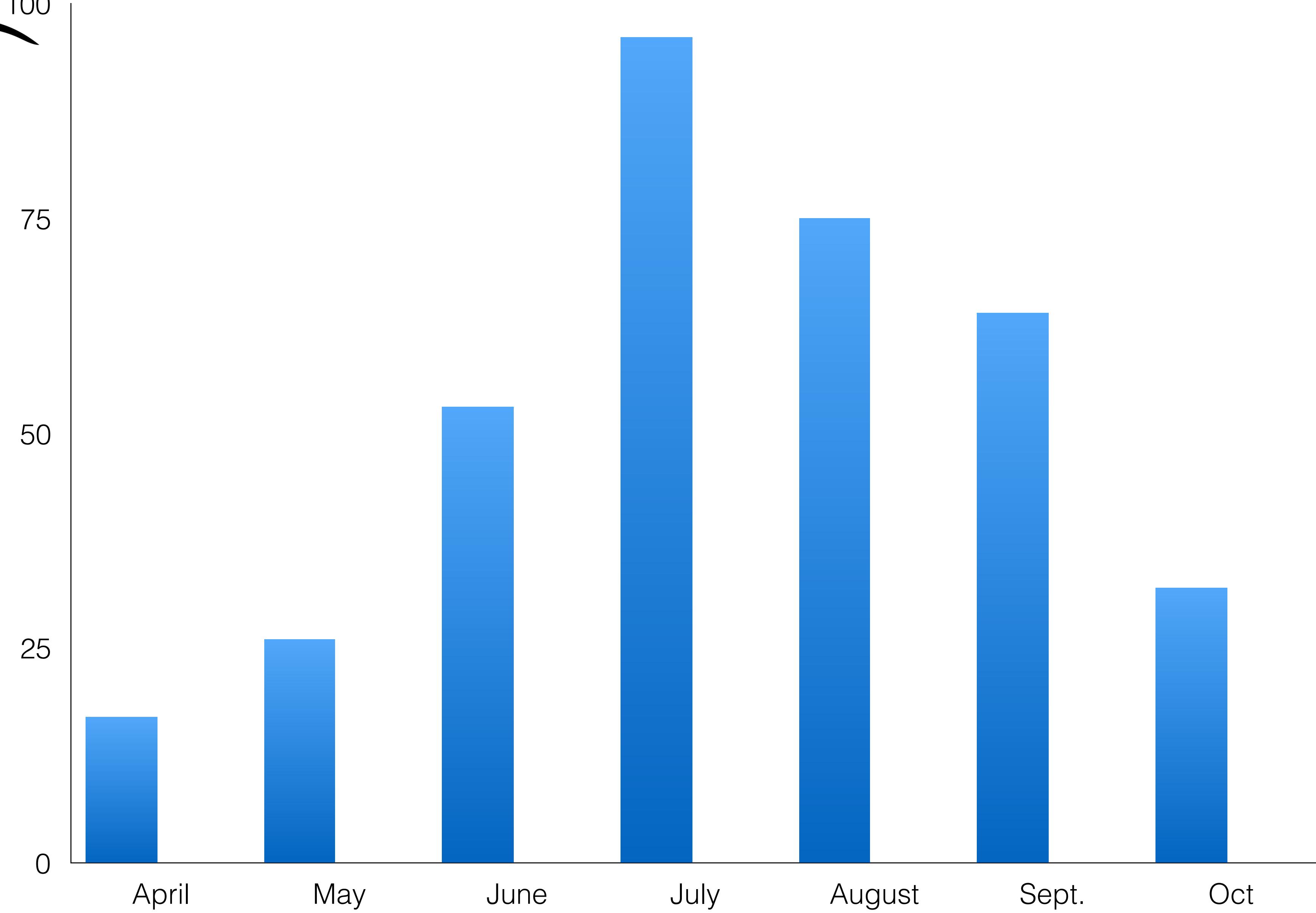
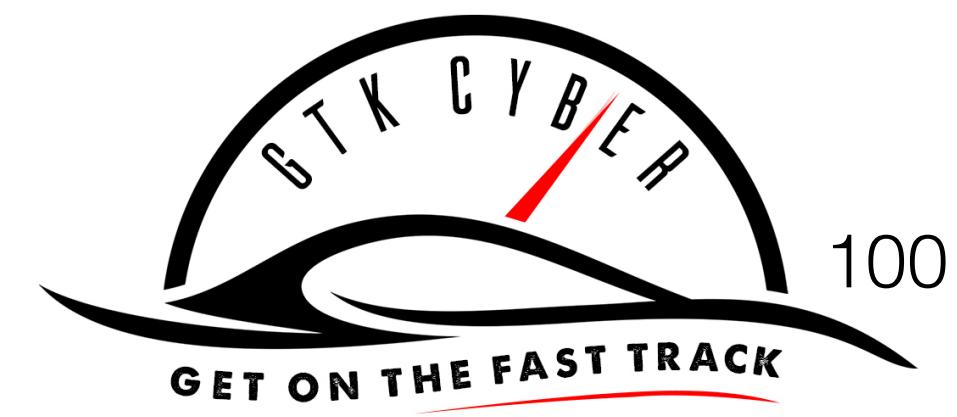


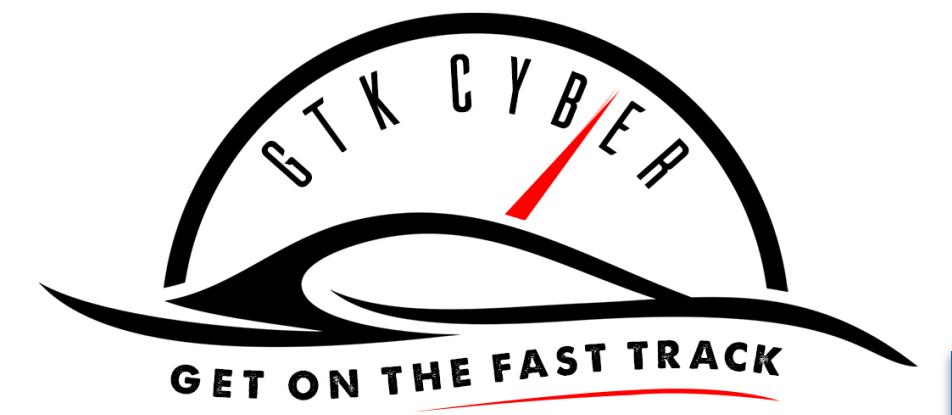
Visual Encodings: Position

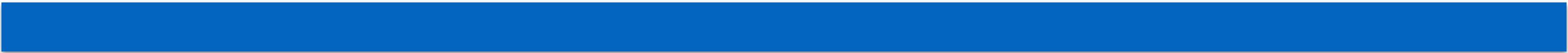
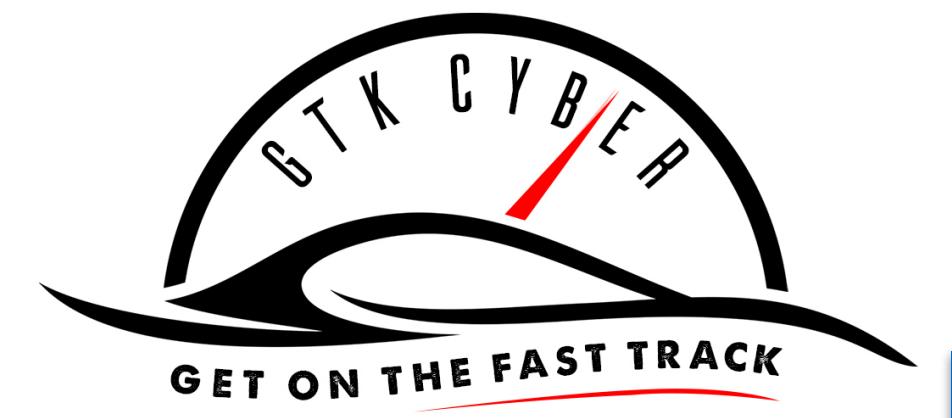


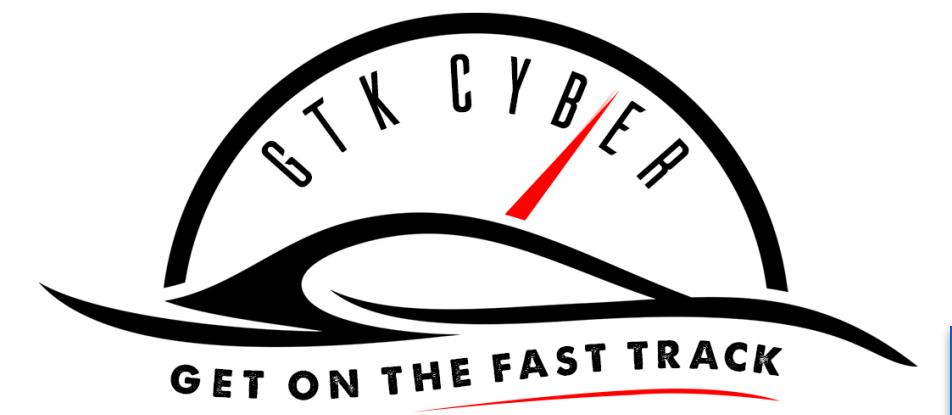


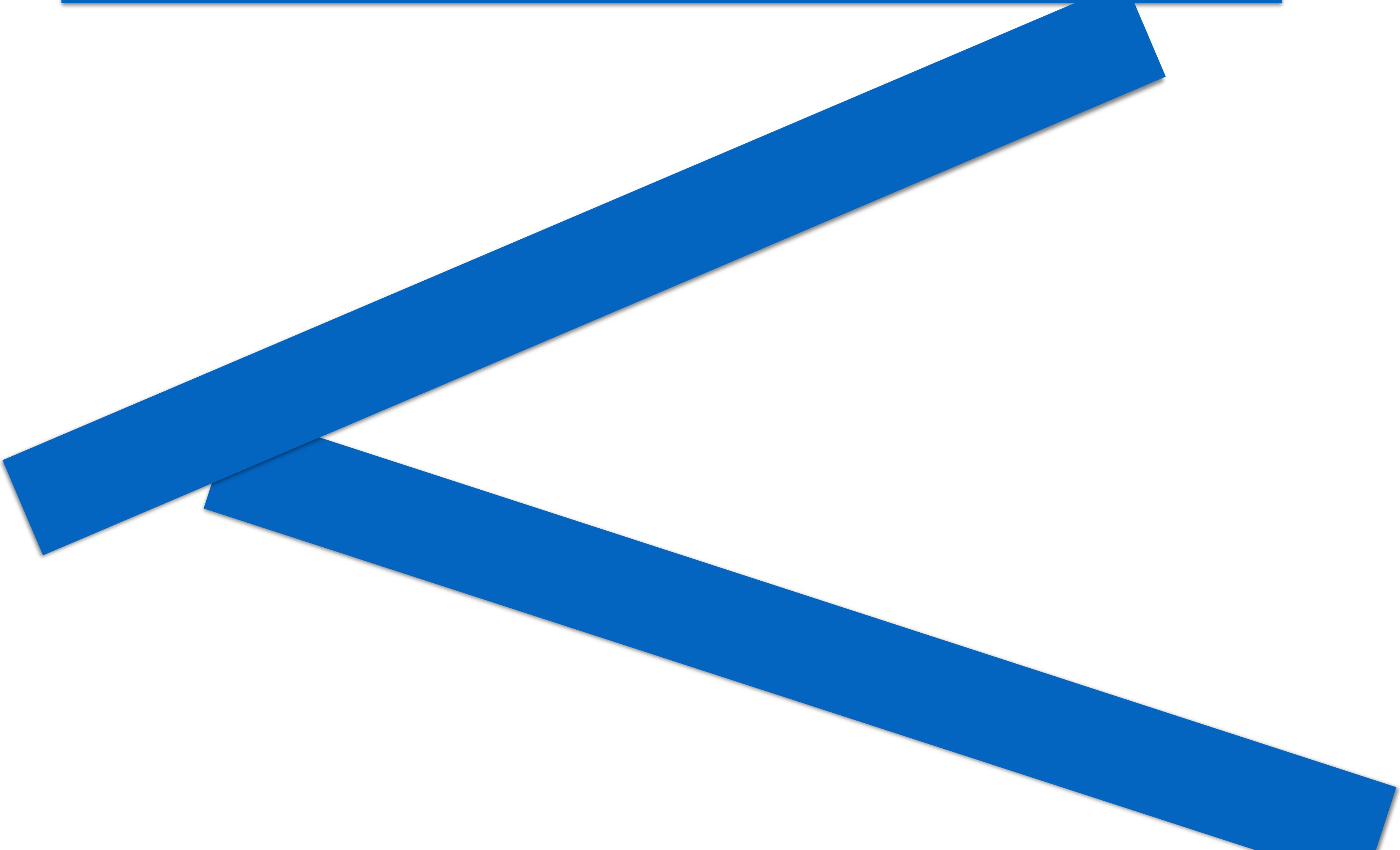
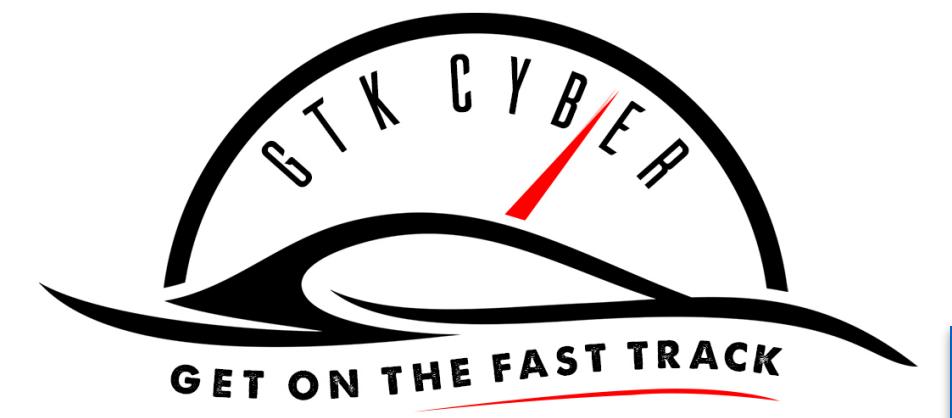
Visual Encodings: Length

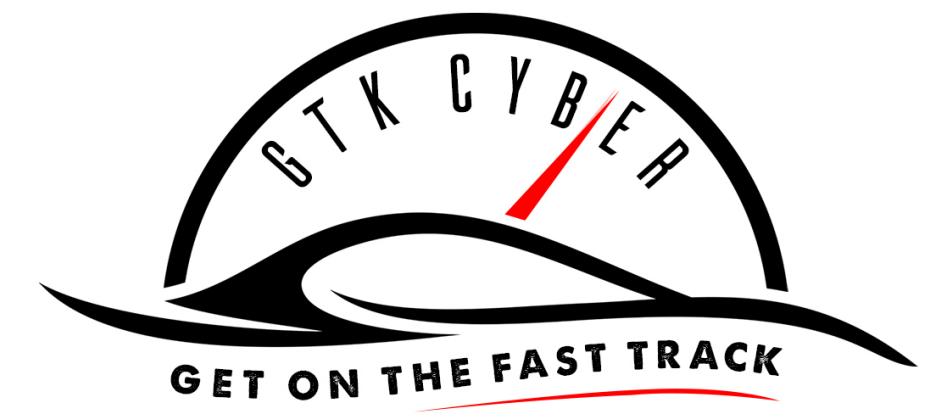


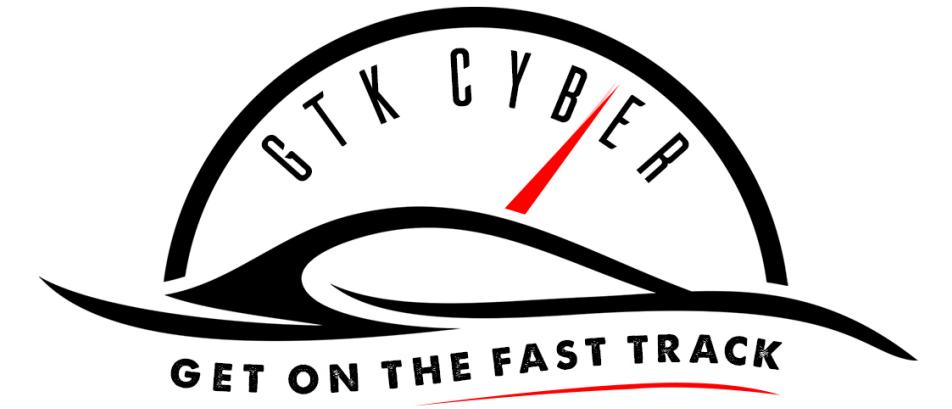




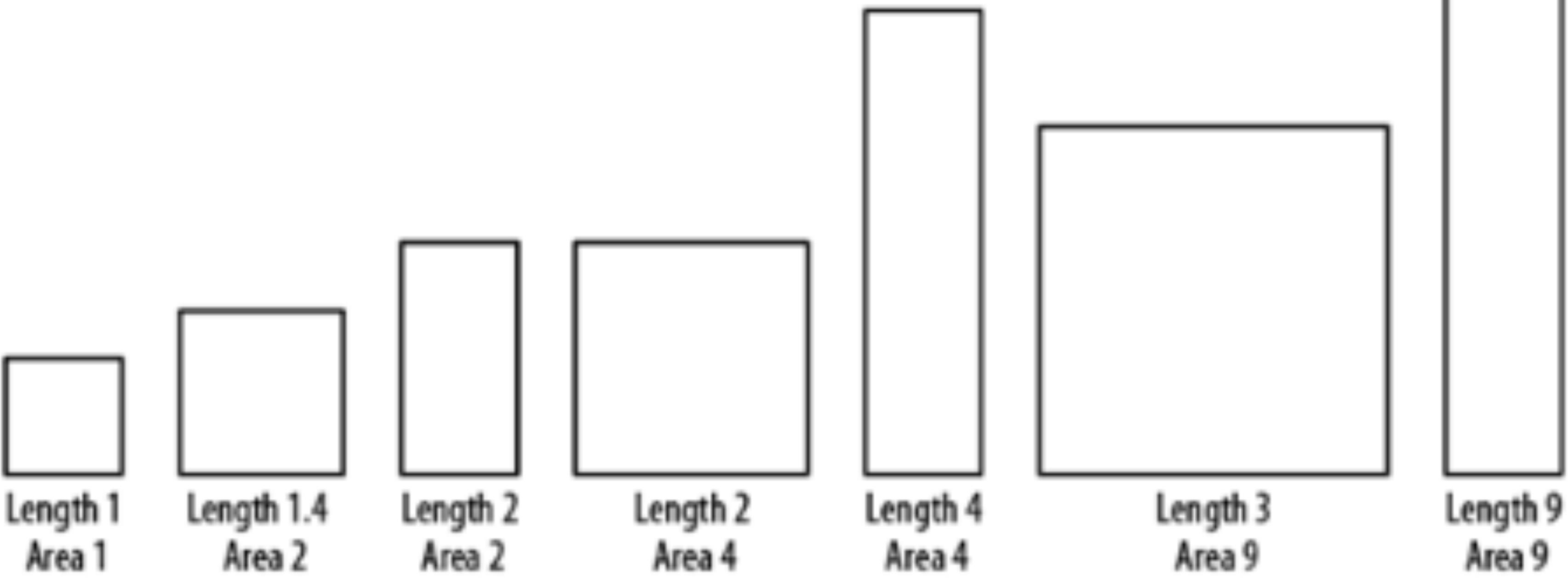


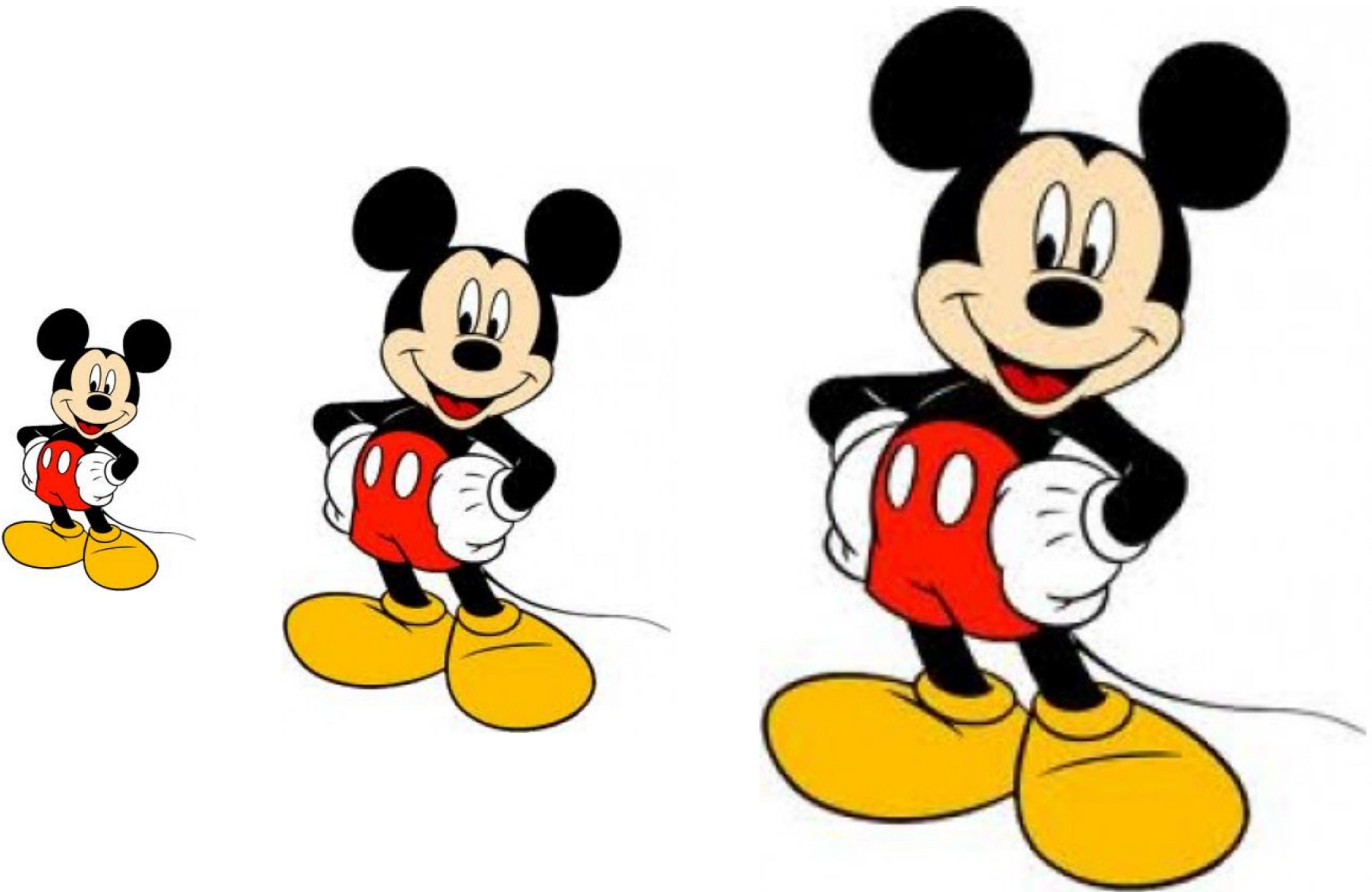
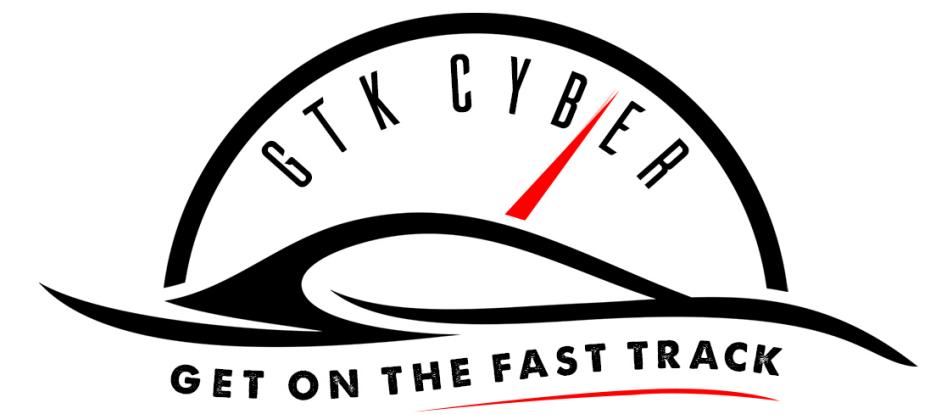


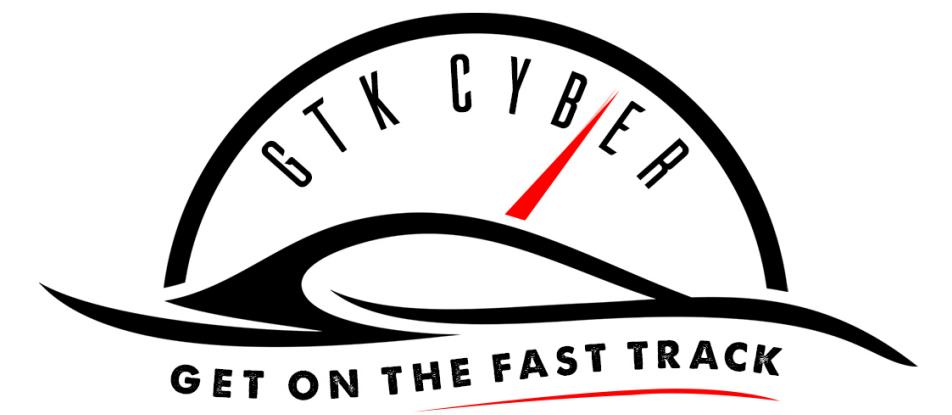


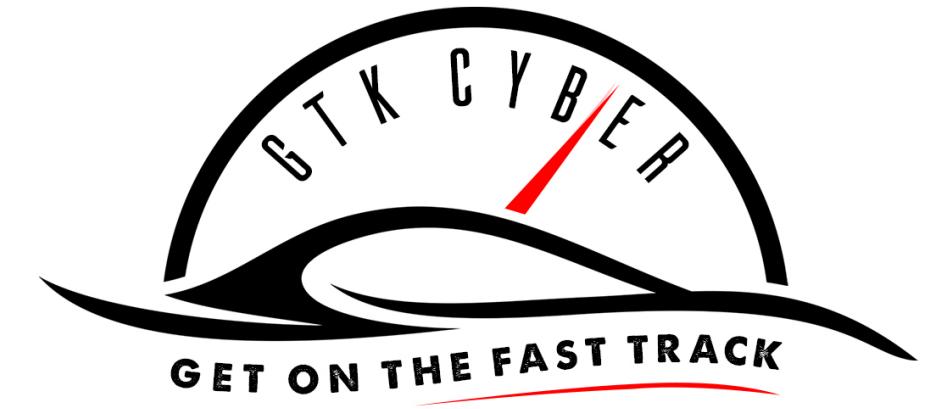


Visual Encodings: **Size**

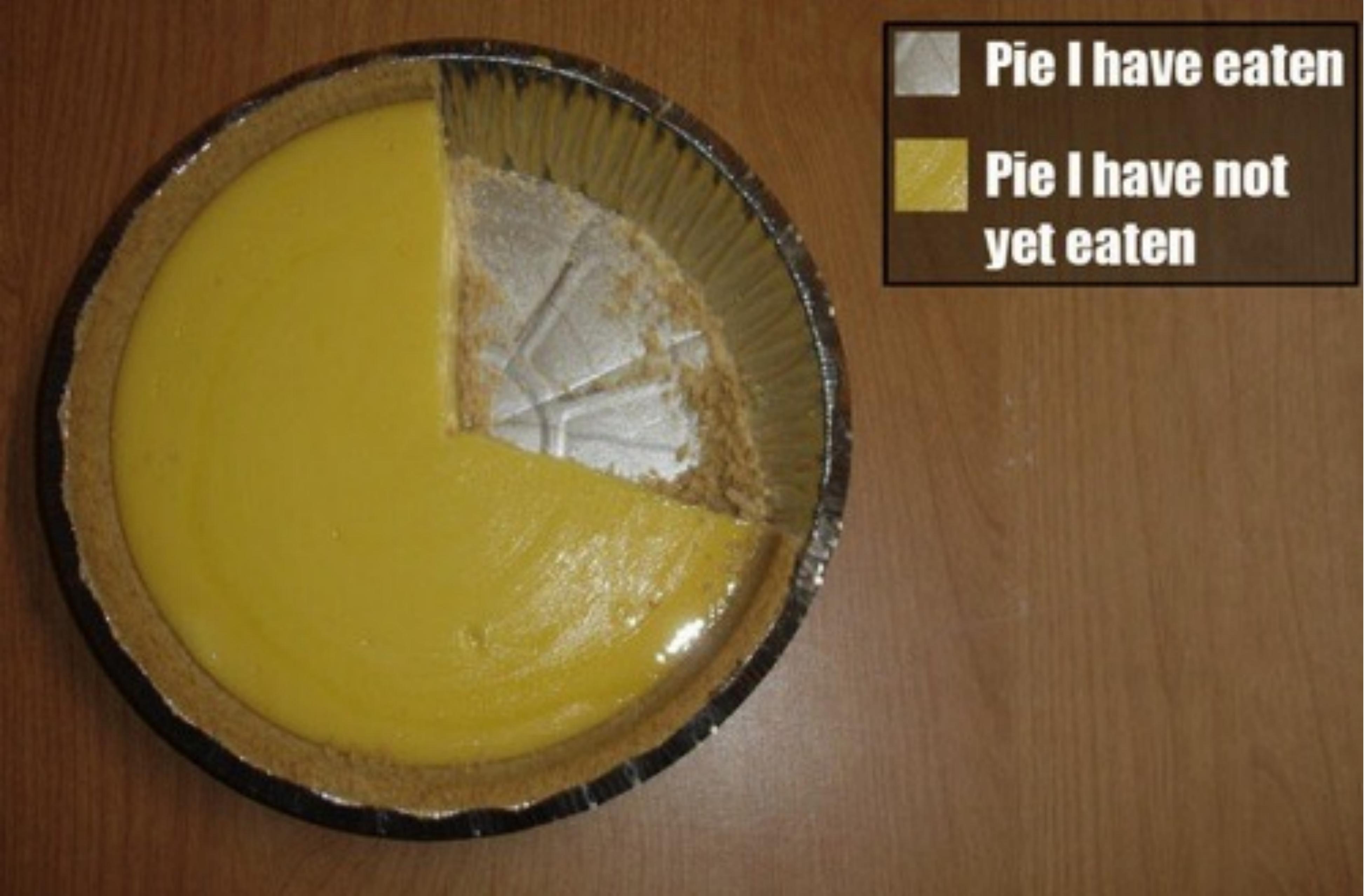


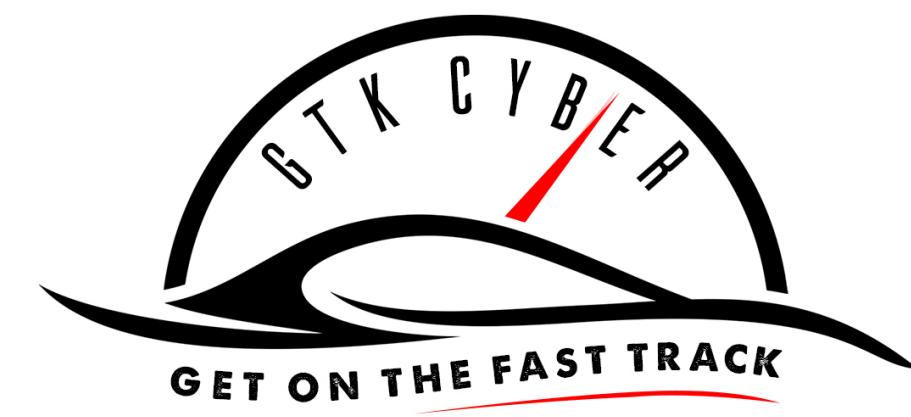




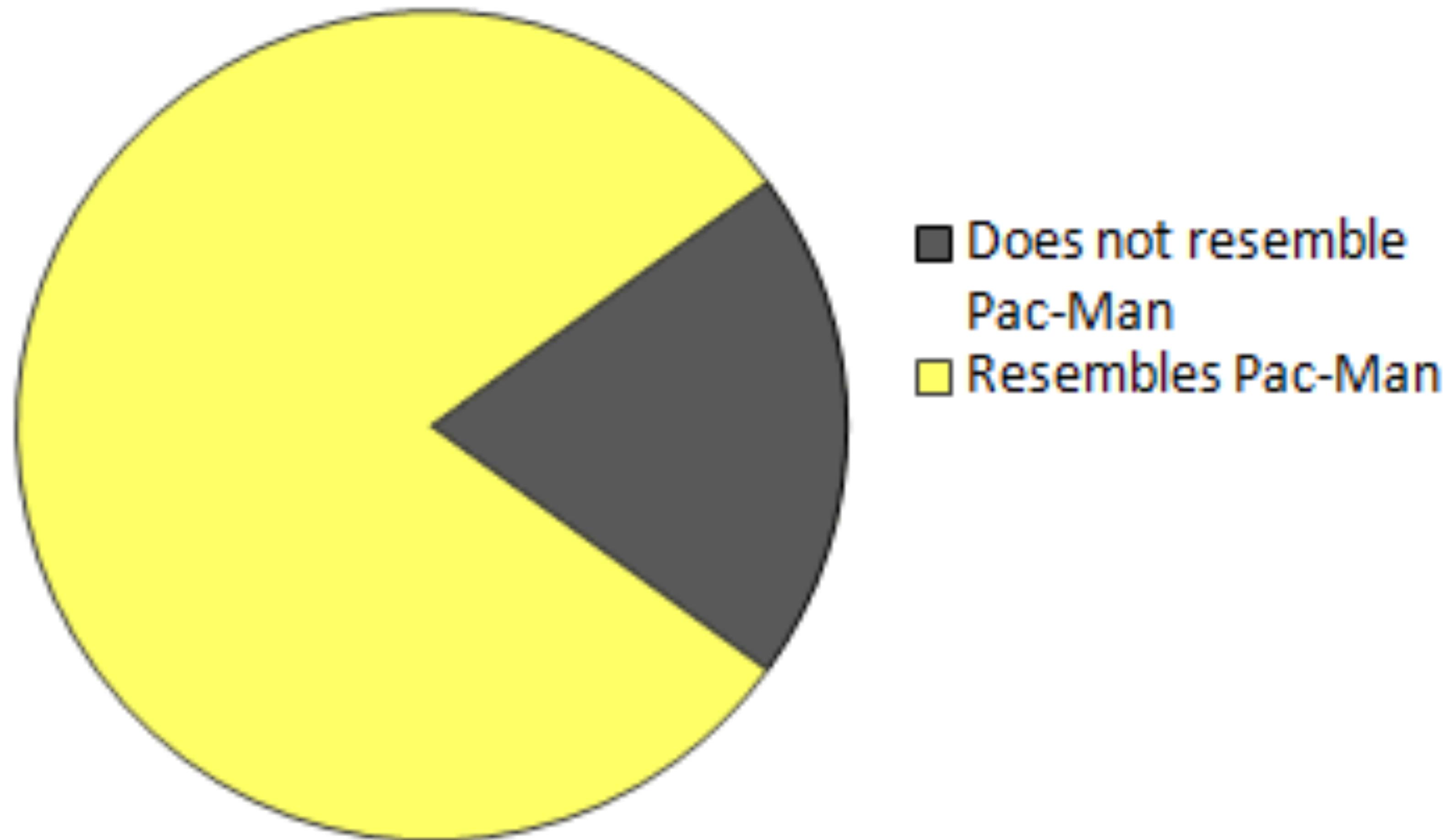


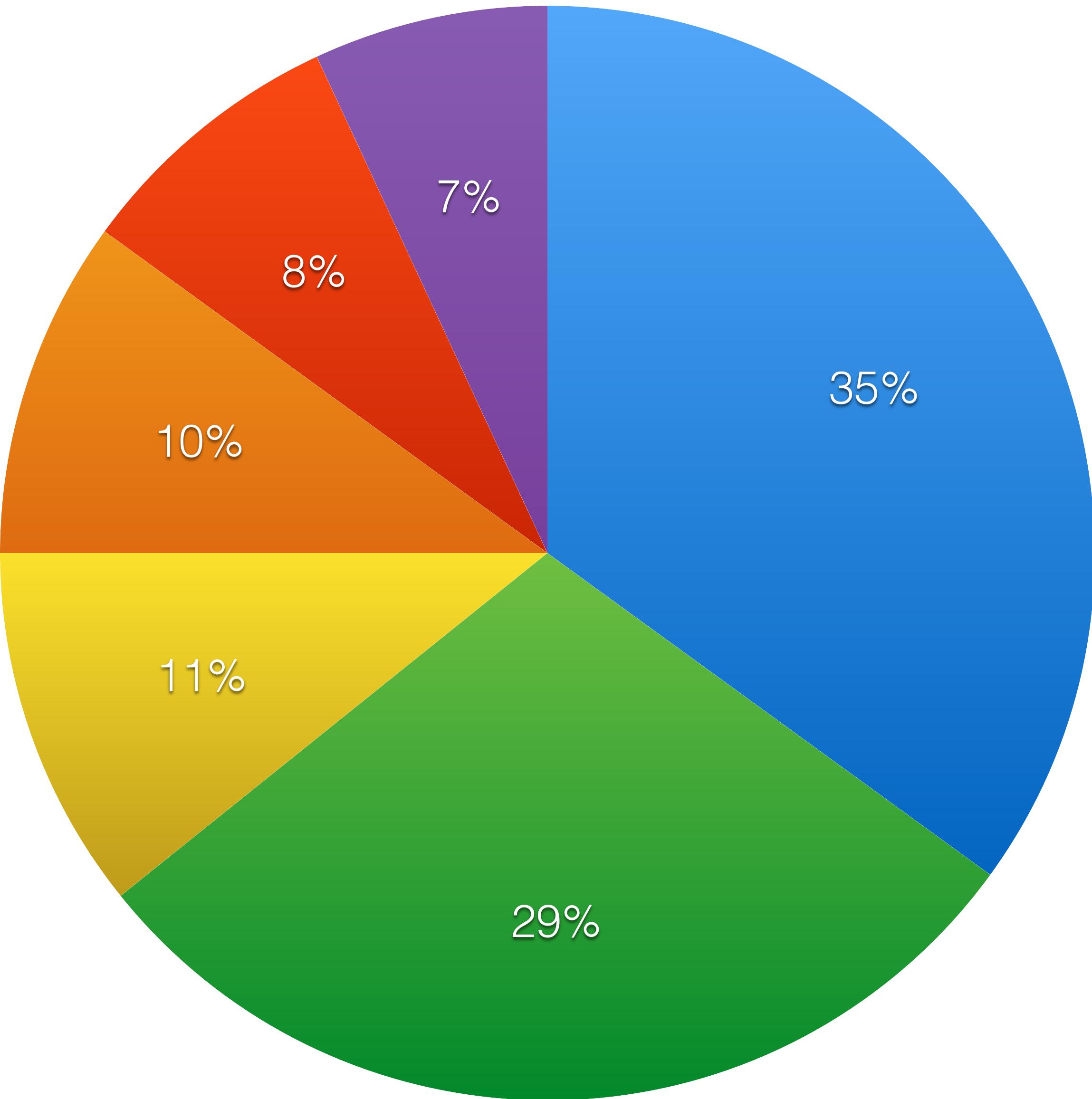
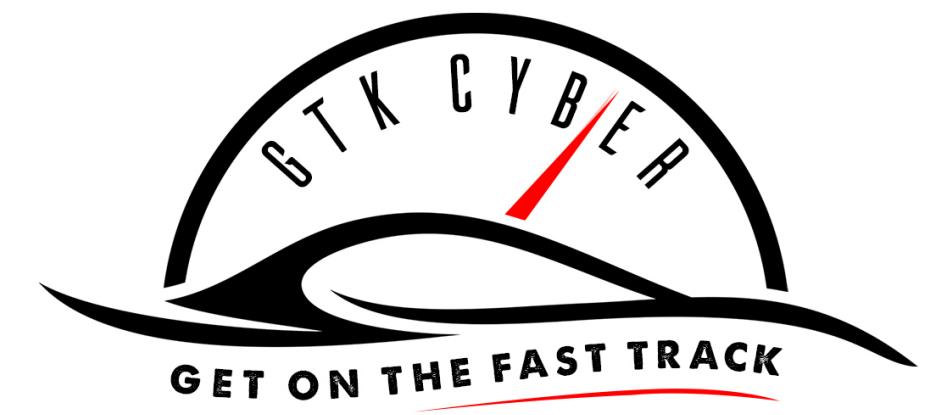
A brief interlude: Why
never to use a Pie Chart

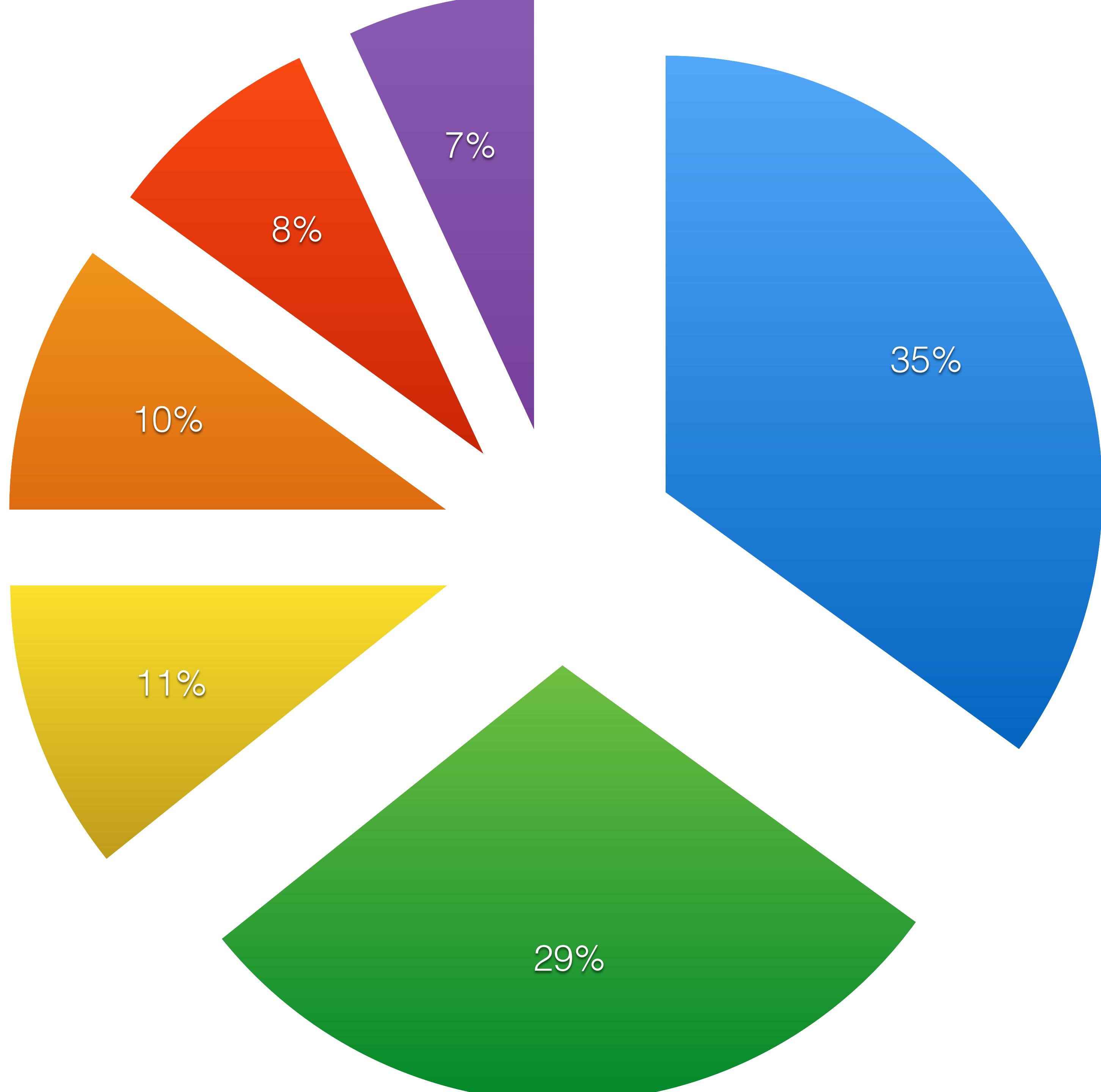
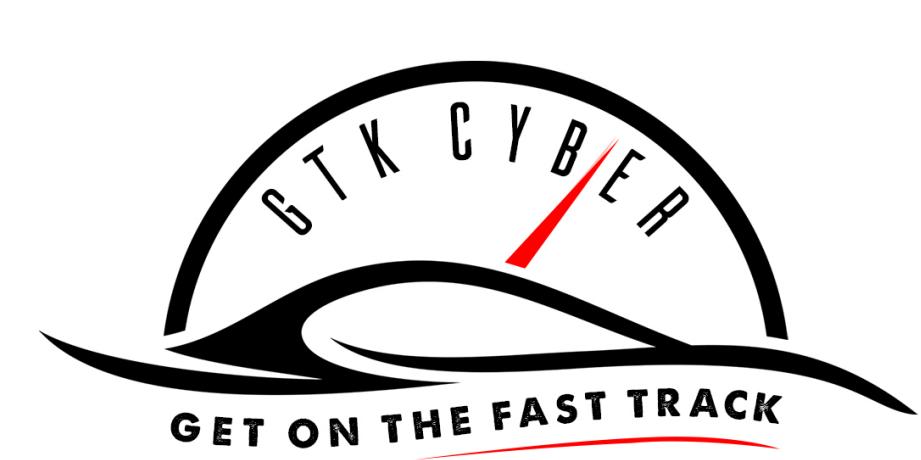


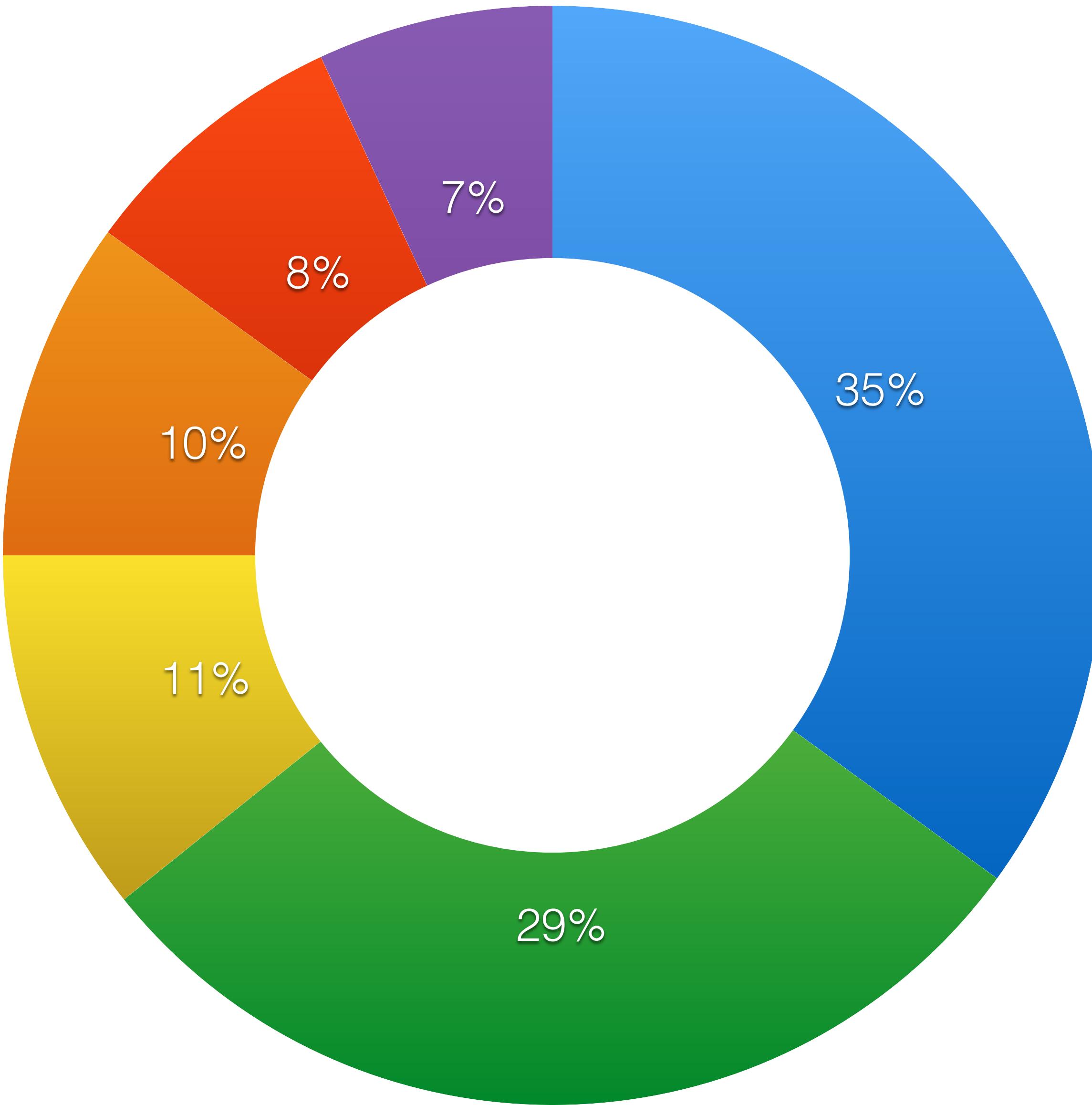
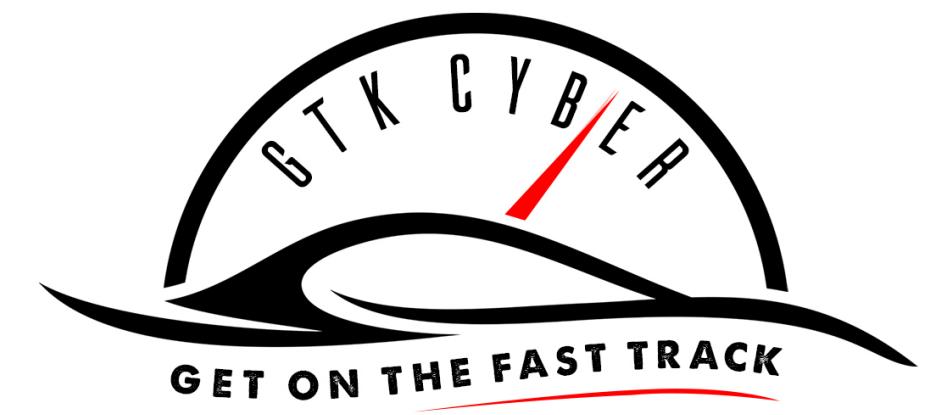


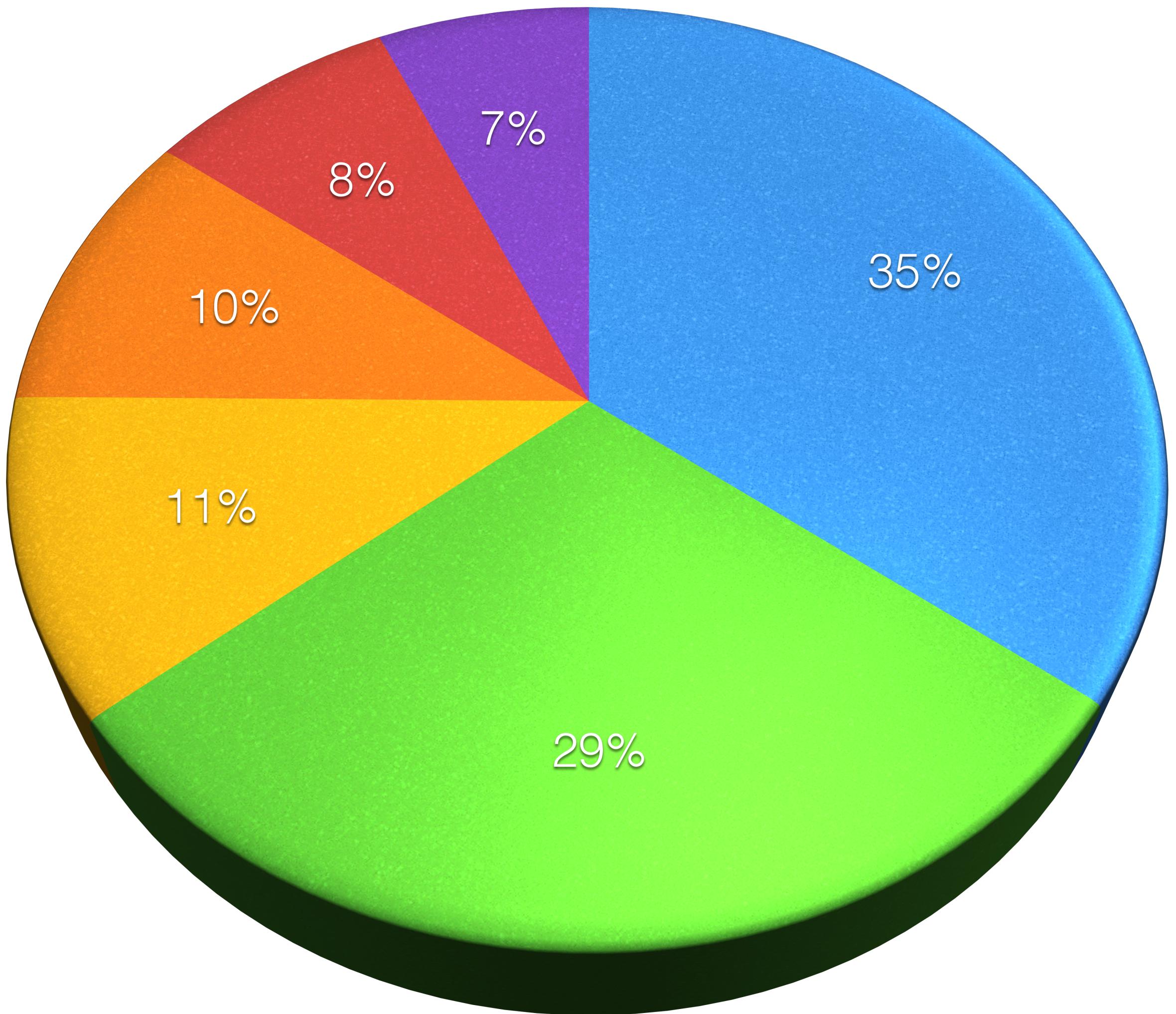
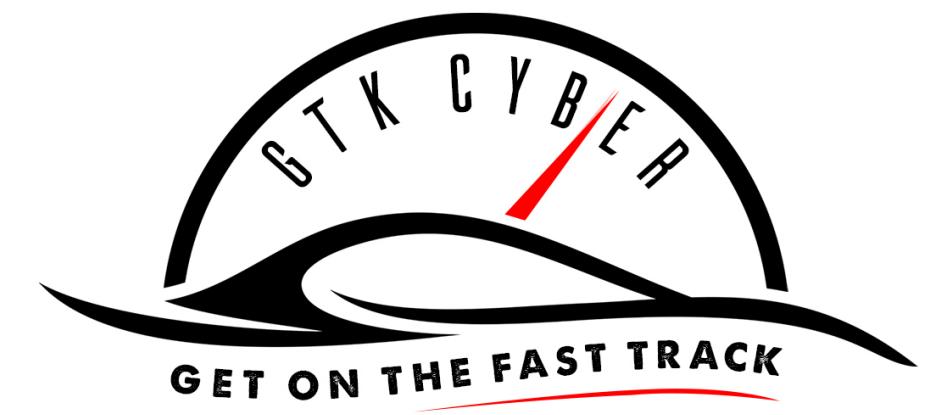
Percentage of Chart Which Resembles Pac-Man

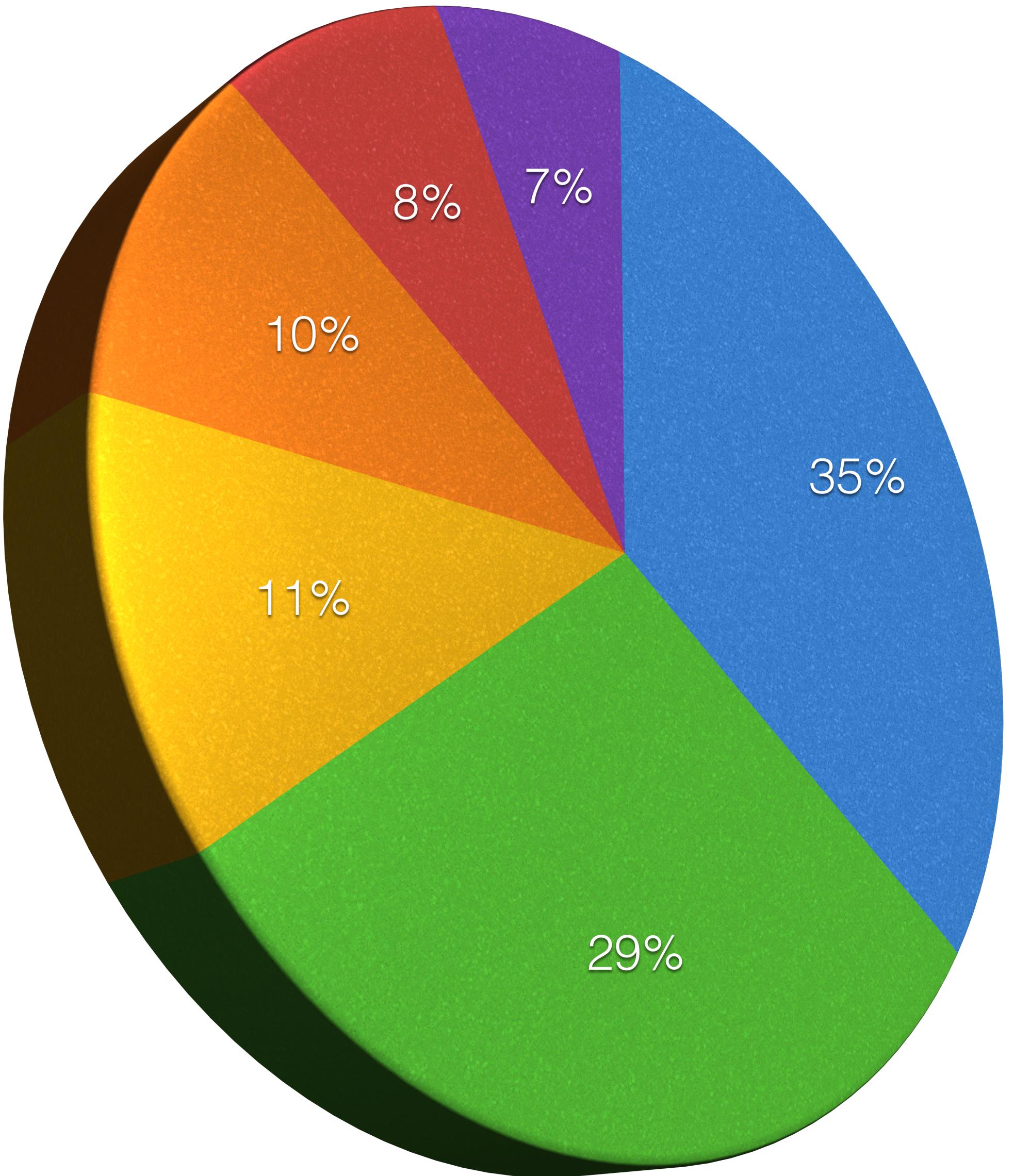
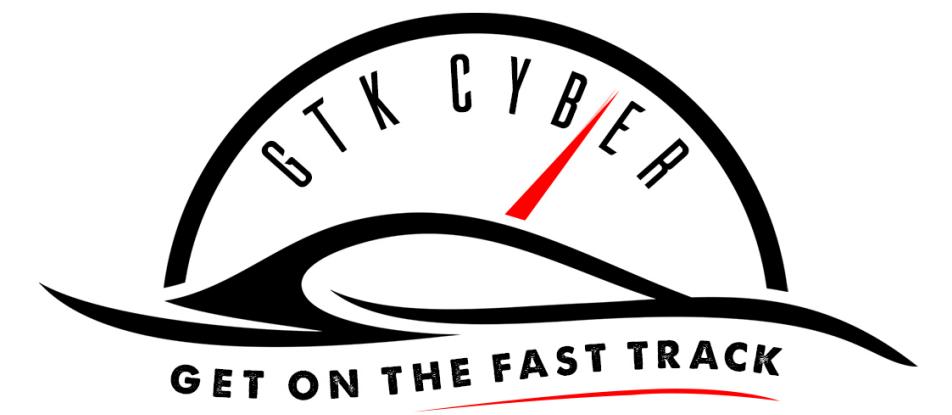


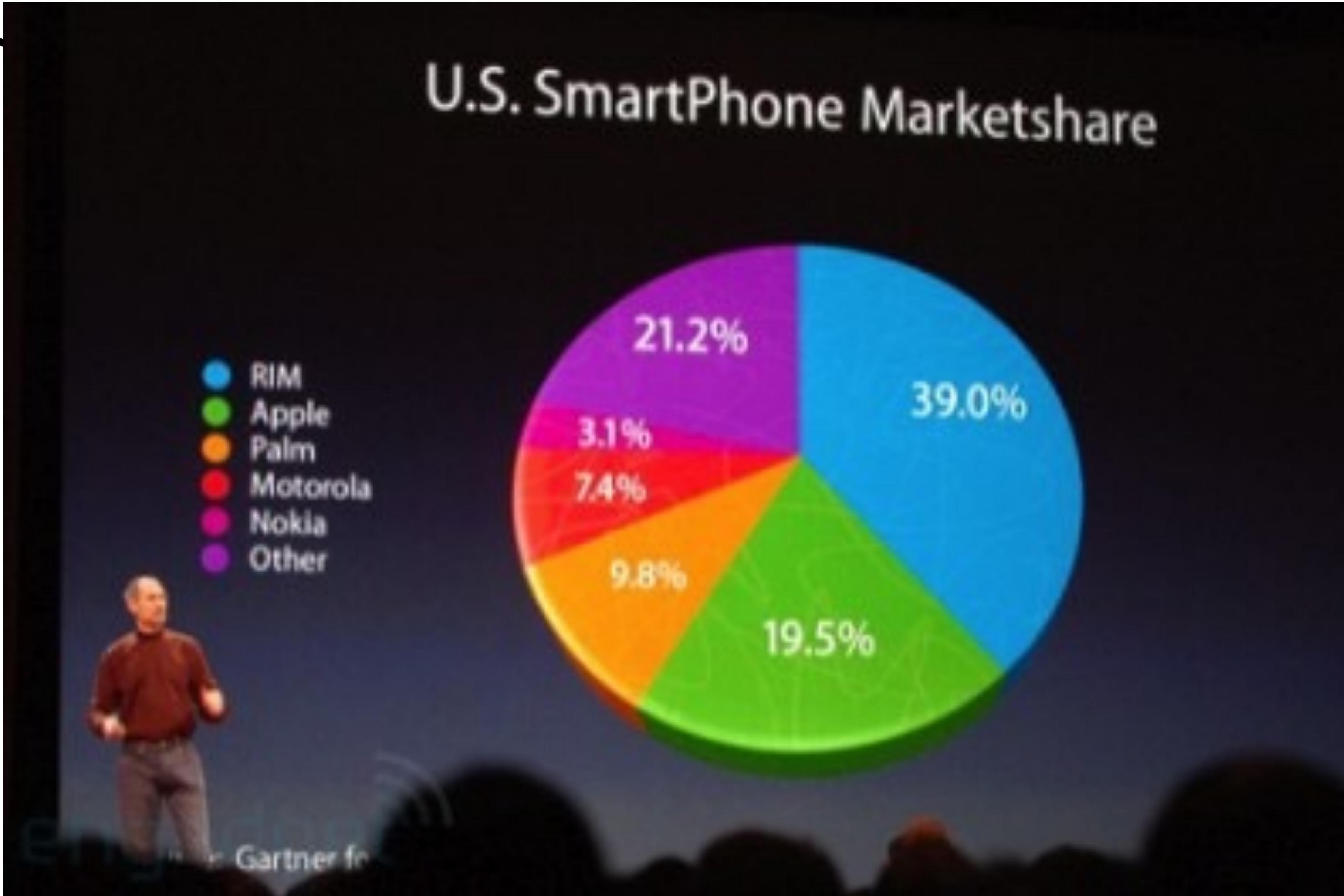




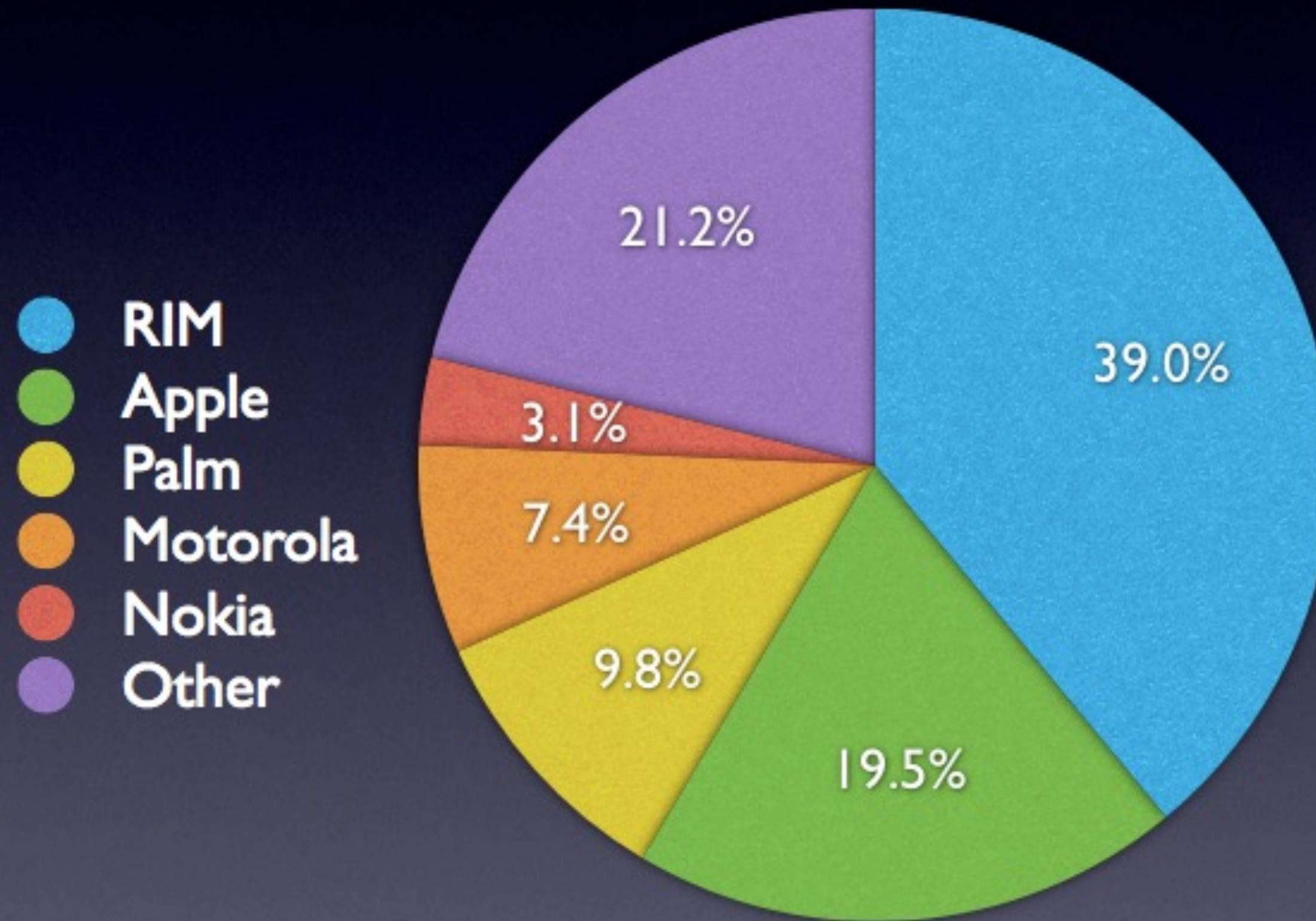








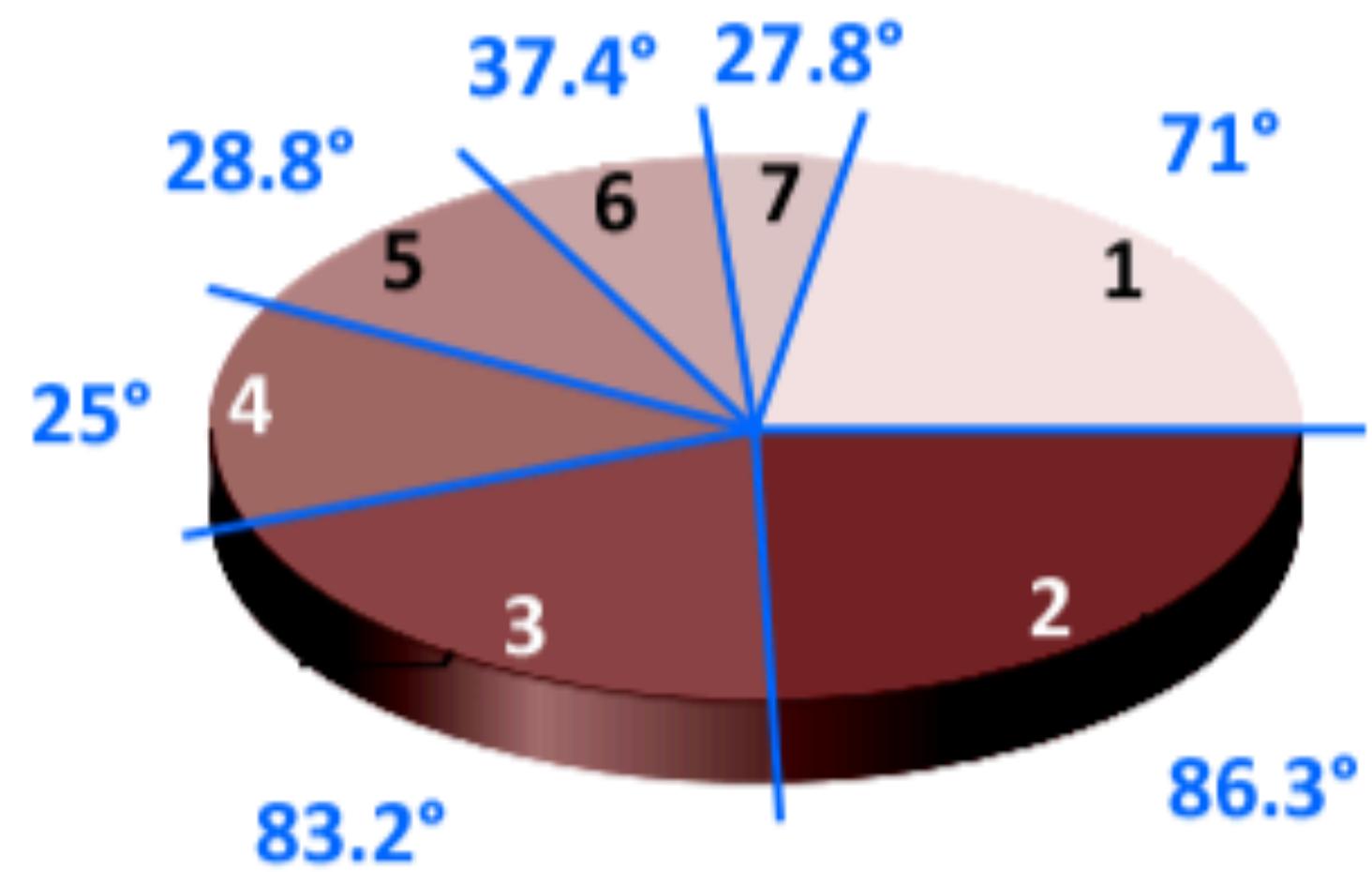
U.S. SmartPhone Marketshare



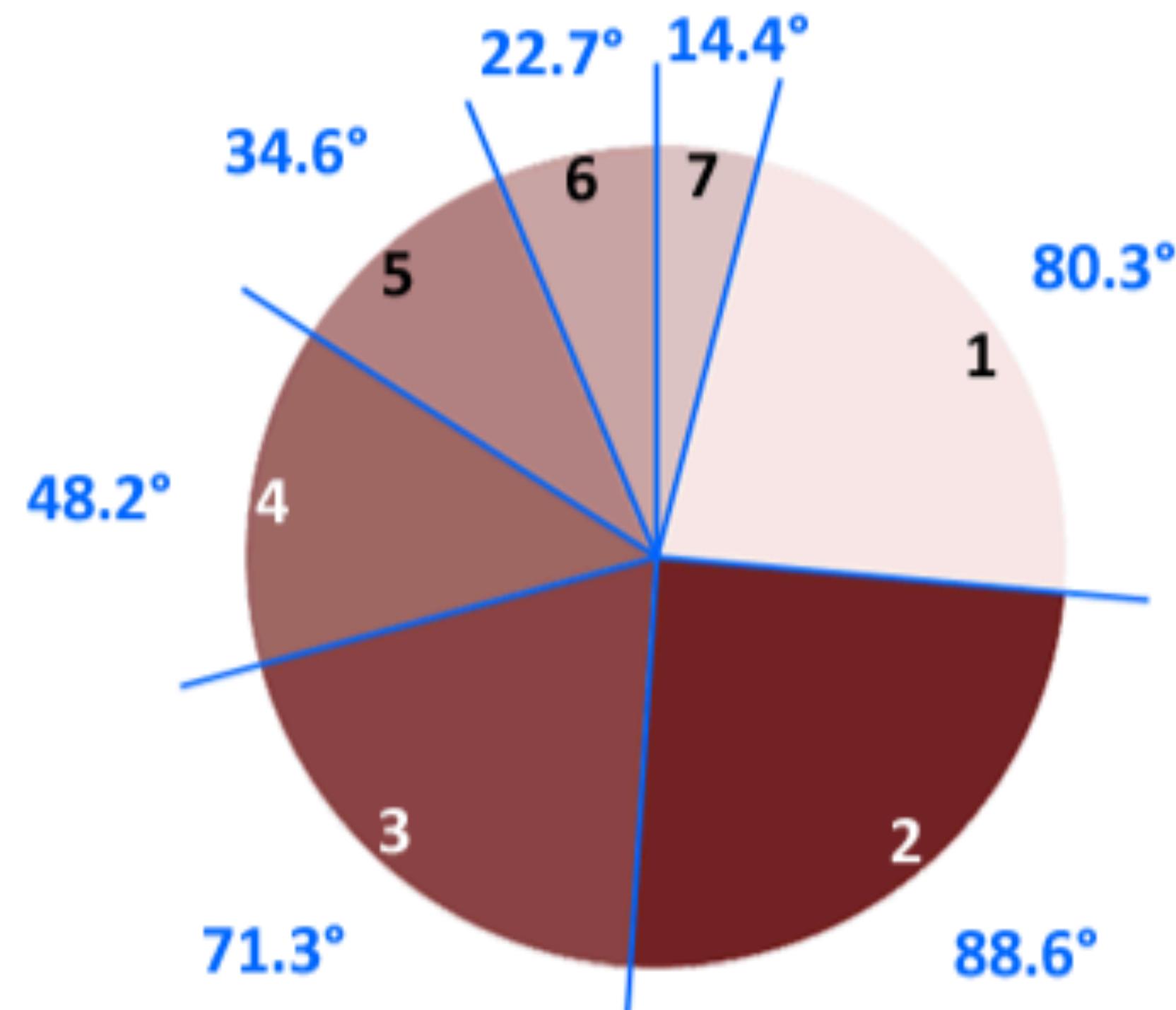


3D Pie Charts are BAD!

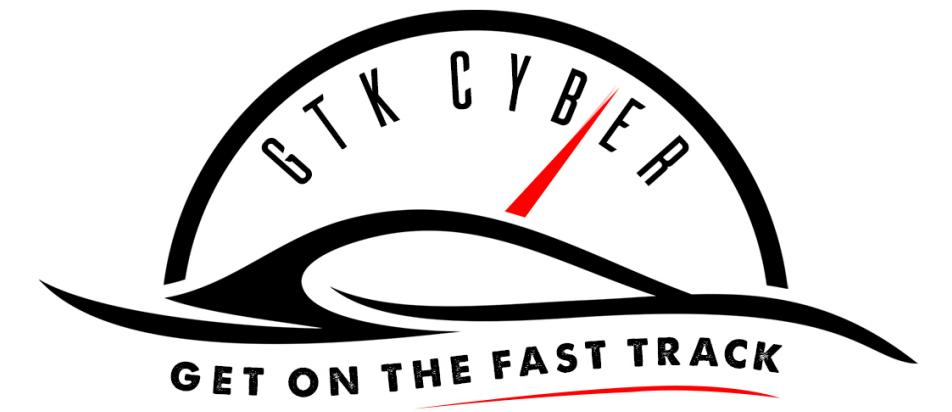
3D Pie Charts Distort Angles



Angles on the original pie chart



Angles on a non-3D pie chart

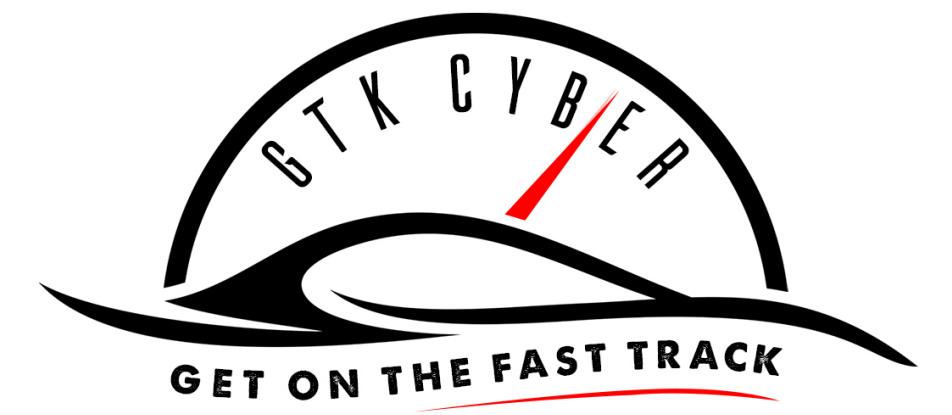


Visual Encodings: Color

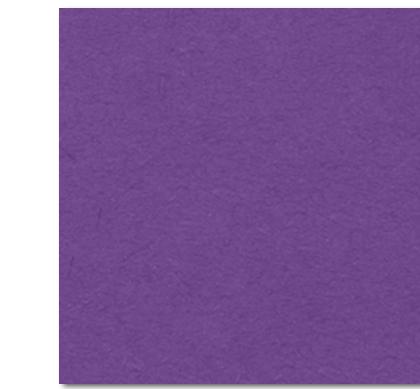
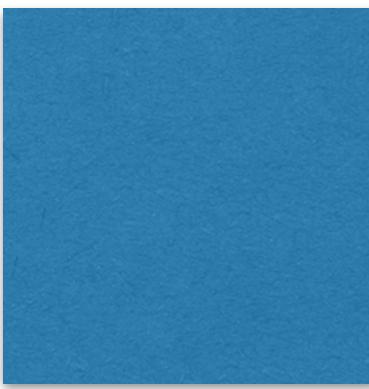
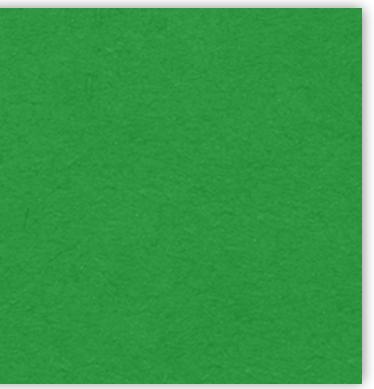


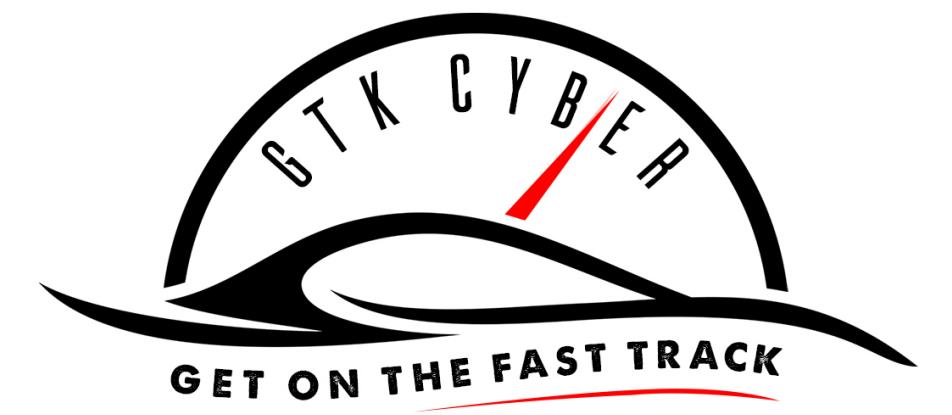
Visual Encodings: Color

- **Hue** should be used to encode categorical data
- **Saturation** should be used to encode intensity or a continuous value

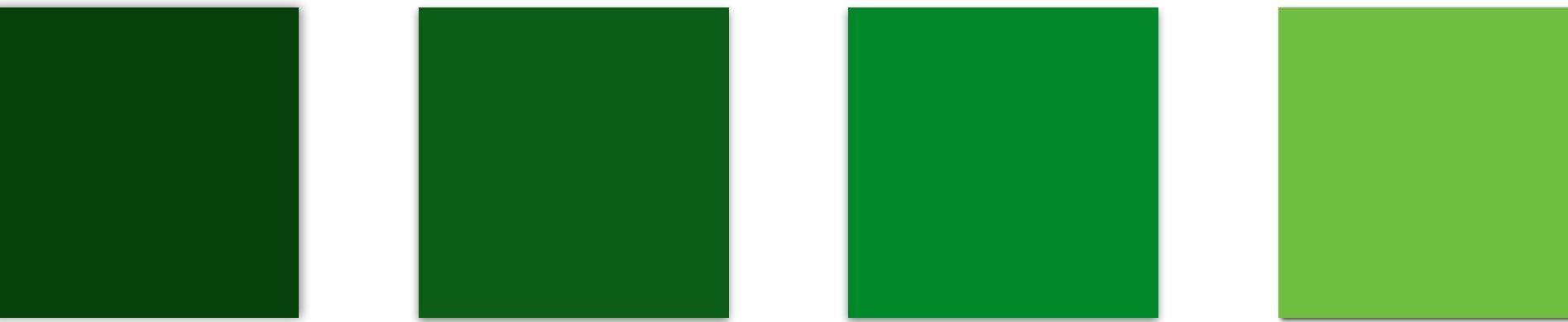


Hue

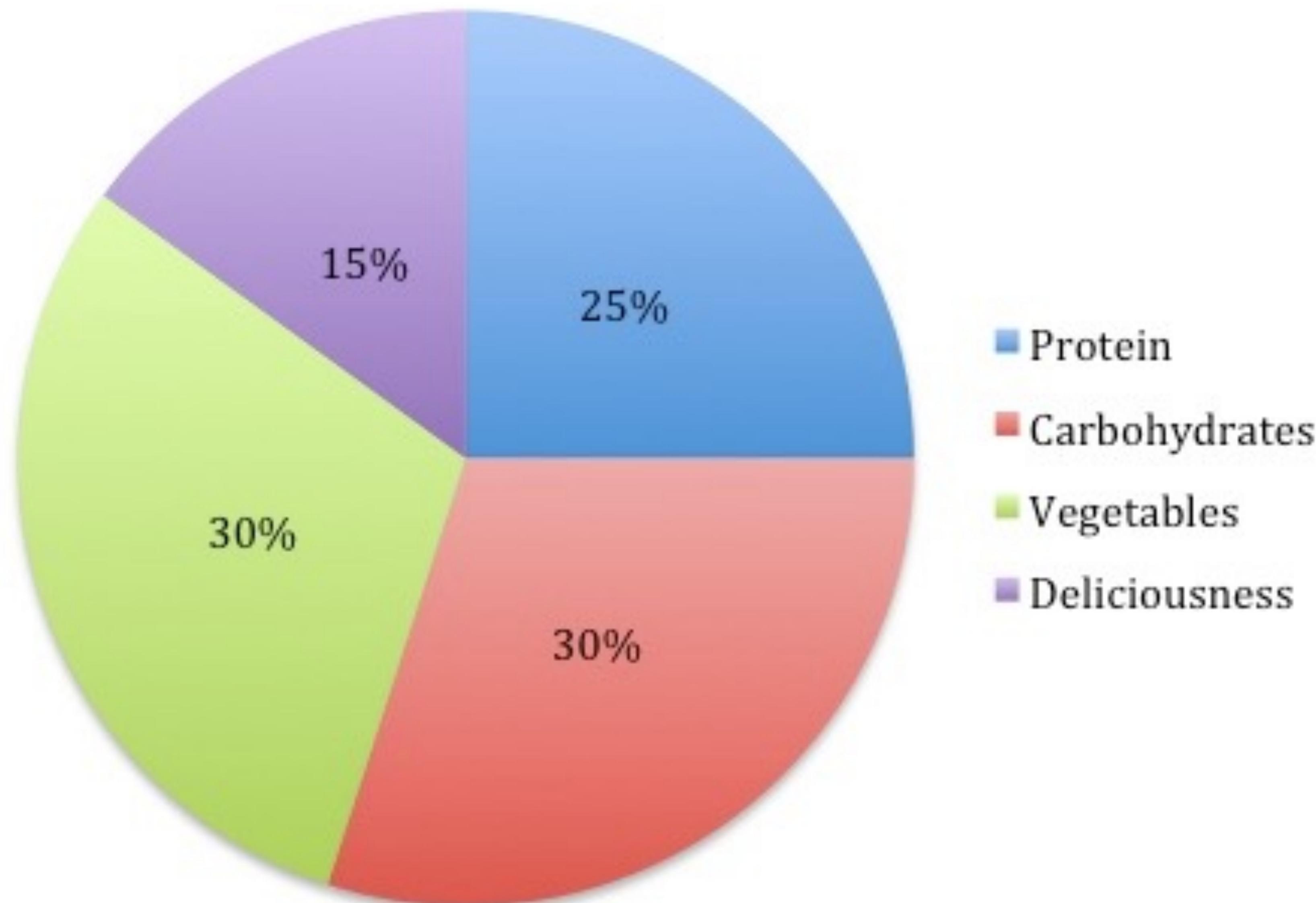




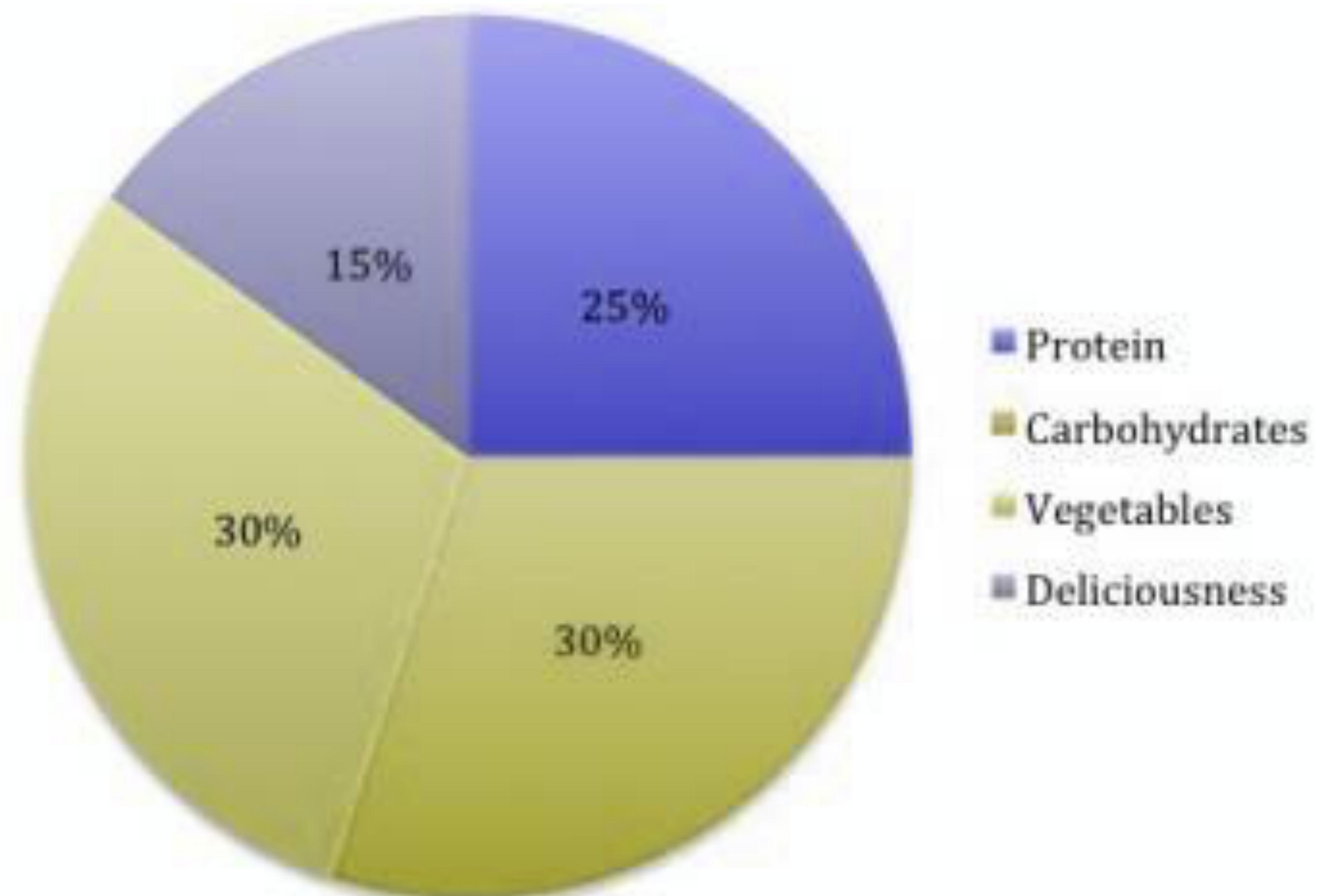
Saturation



A Healthy Meal

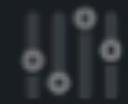


A Healthy Meal





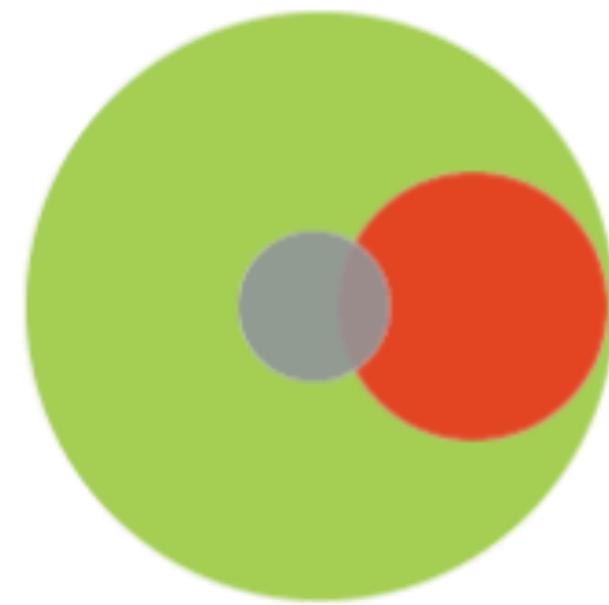
elastica



Dashboard

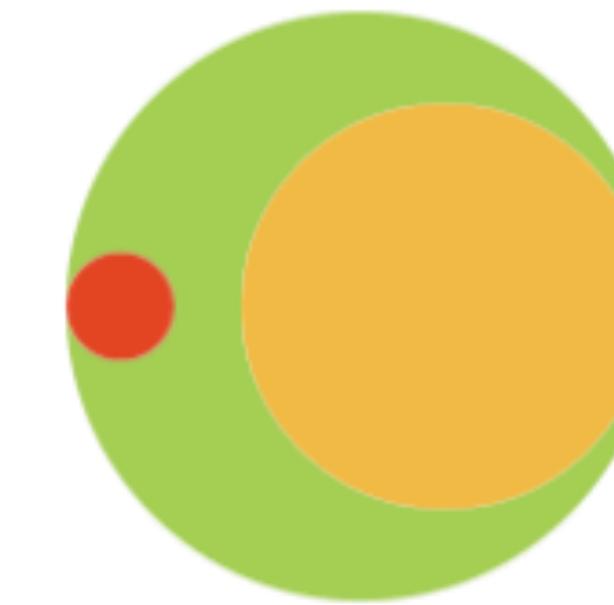
Users

Total (192)

28
High Risk0
Med Risk4
Blocked

Policies

Total (231)

3
Blocking142
Alerting

Policy Alerts

Alerting



Blocked



Rest



Threat Alerts

High Risk



Med Risk



Low Risk



Audited Services

by Users ▾

High Risk (736)

Medium Risk (3k)

Low Risk (3k)

3k
Users494.3 GB
Traffic963k
Sessions243
Destinations

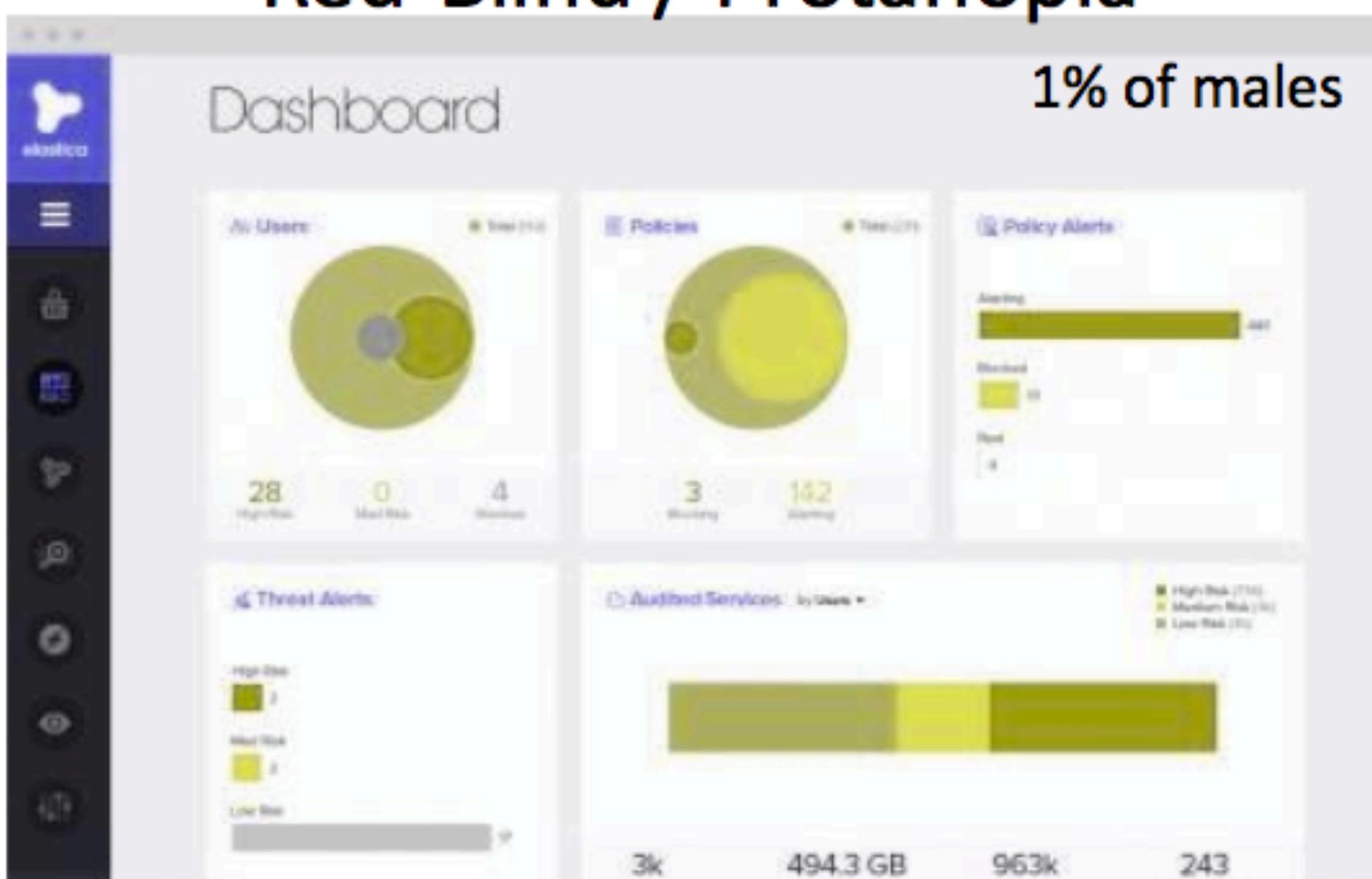
Original



Green-Blind / Deutanopia

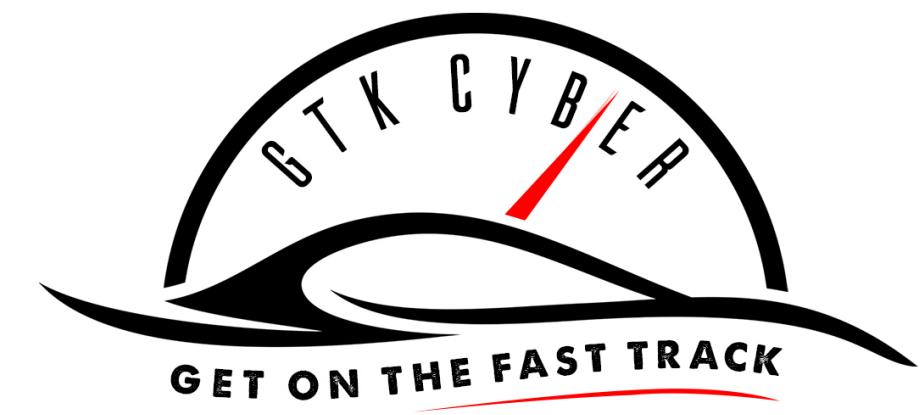


Red-Blind / Protanopia

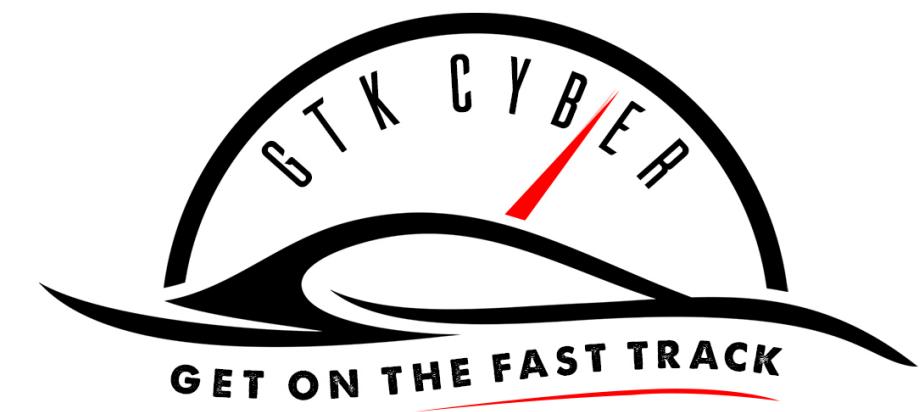


Blue-Blind / Tritanopia





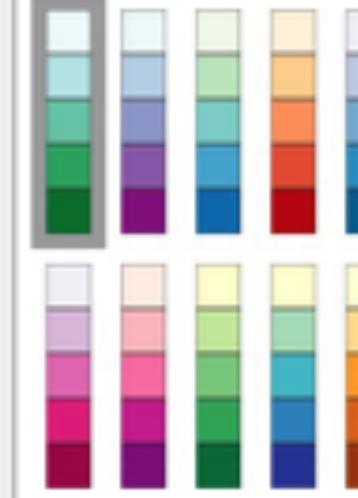
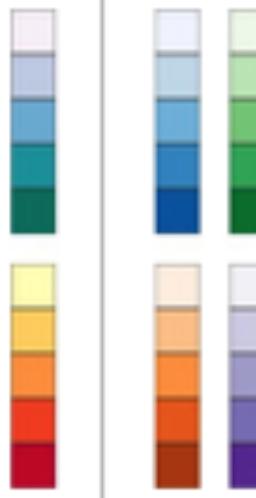
<http://www.color-blindness.com/coblis-color-blindness-simulator/>



Color

Number of data classes: 3

Nature of your data:
 sequential diverging qualitative

Pick a color scheme:
Multi-hue:  Single hue: 

Only show:
 colorblind safe
 print friendly
 photocopy safe

Context:
 roads
 cities
 borders

Background:
 solid color terrain
 color transparency

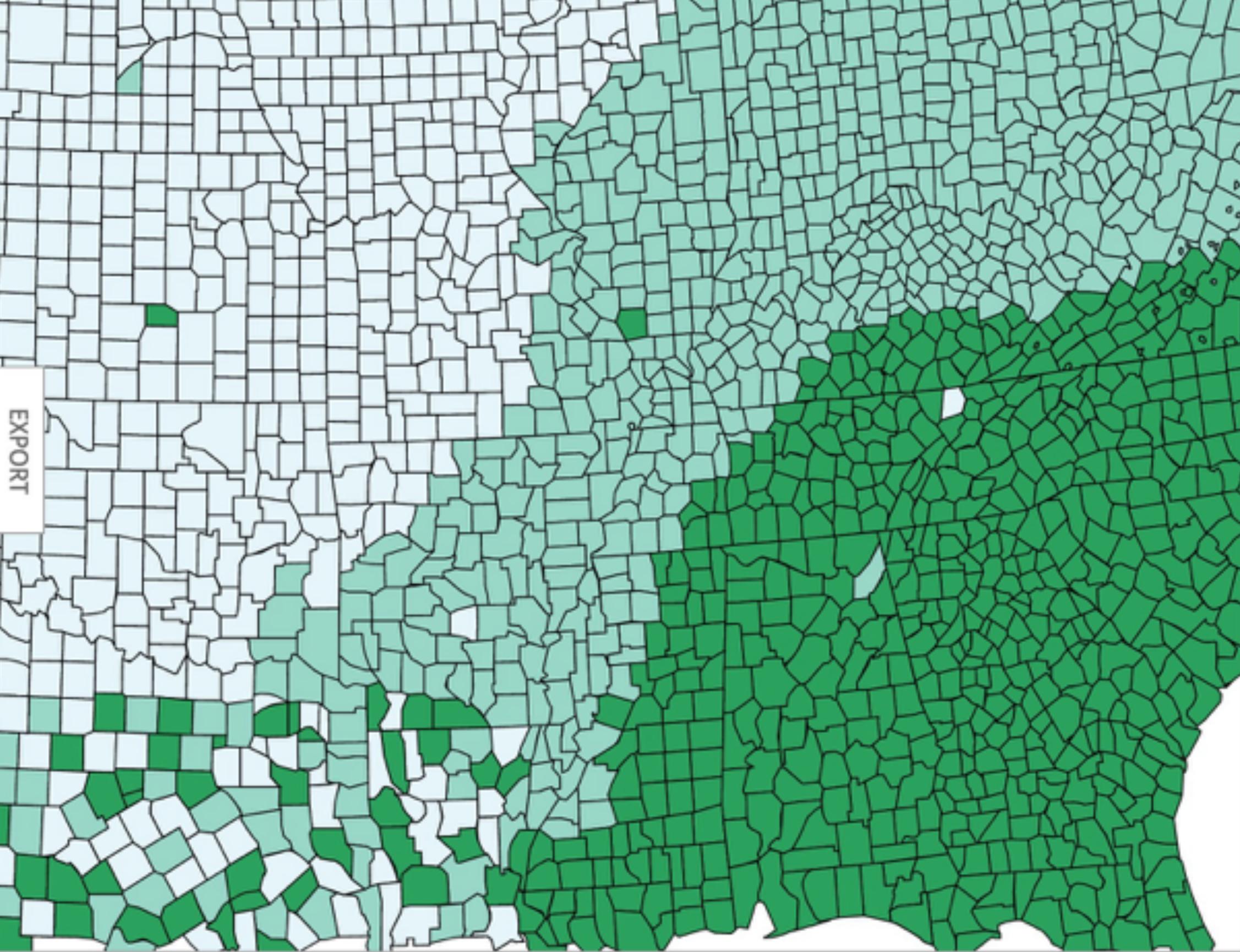
how to use | updates | downloads | credits

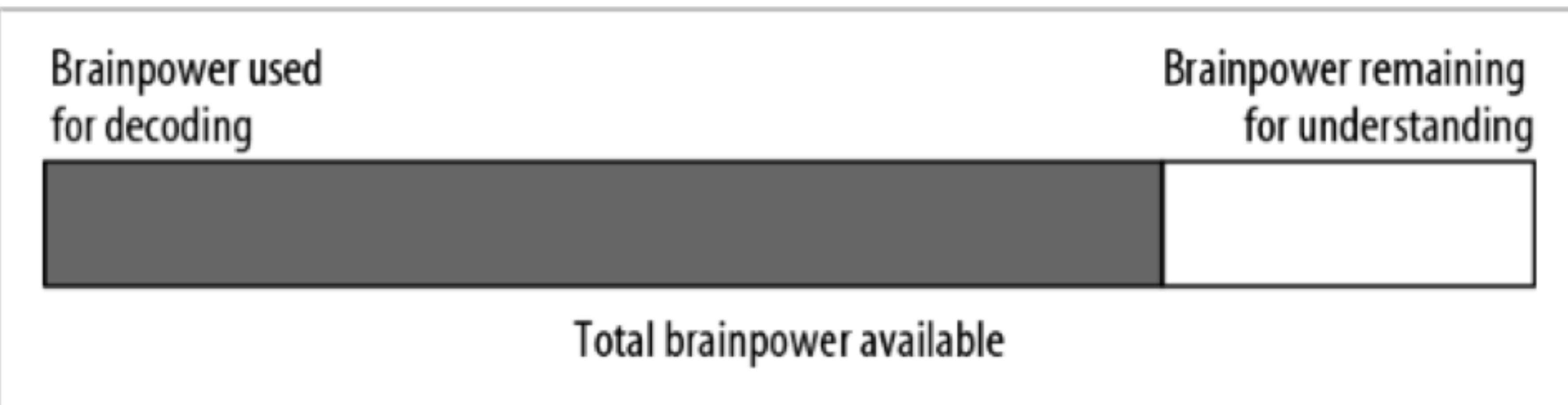
COLORBREWER 2.0
color advice for cartography

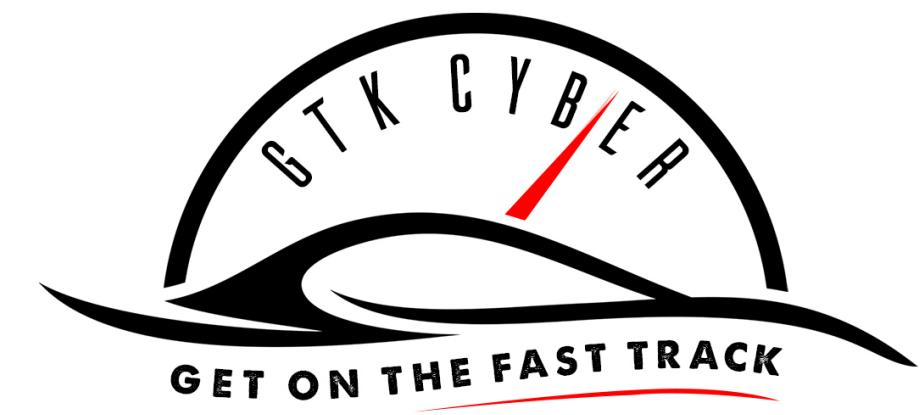
EXPORT

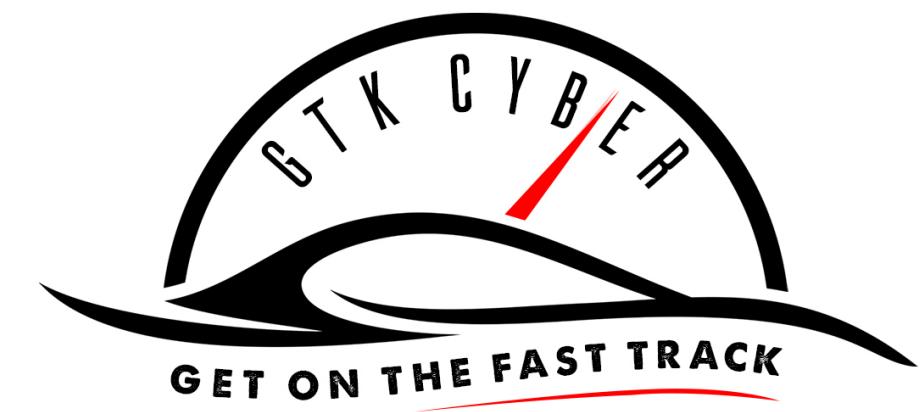
3-class BuGn

HEX
`#e5f5f9`
`#99d8c9`
`#2ca25f`









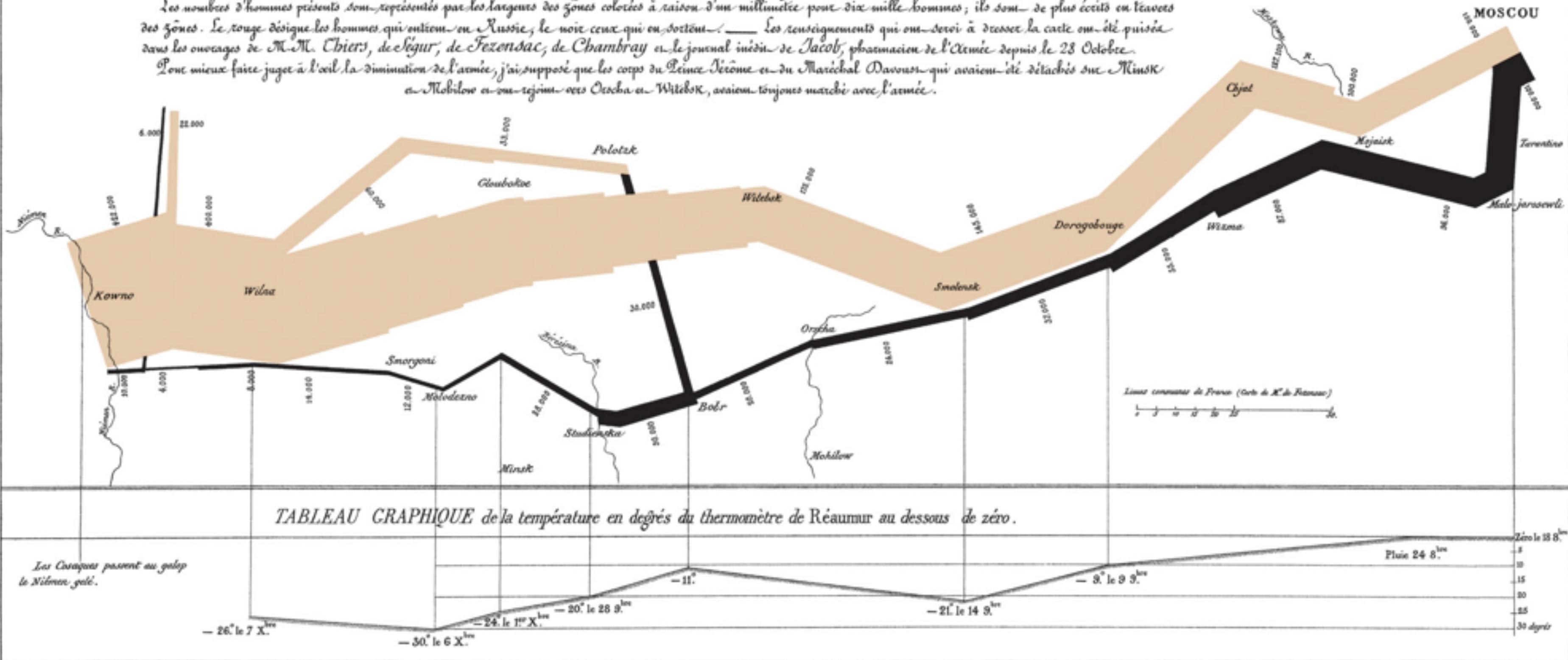
Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.

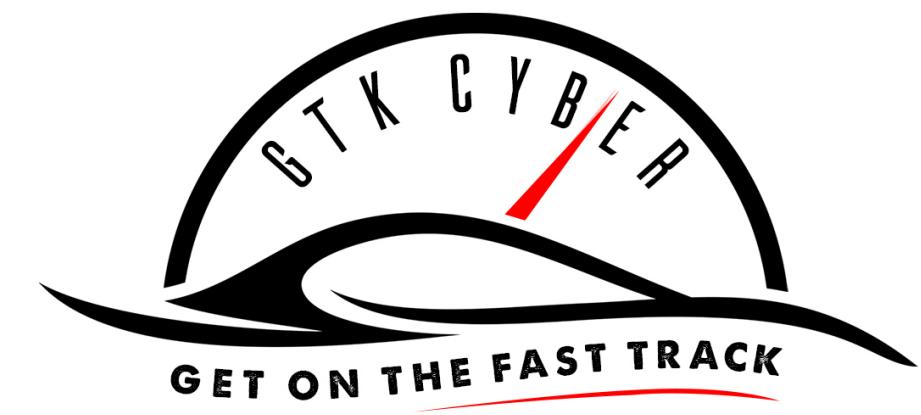
Dessiné par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite

Paris, le 20 Novembre 1869.

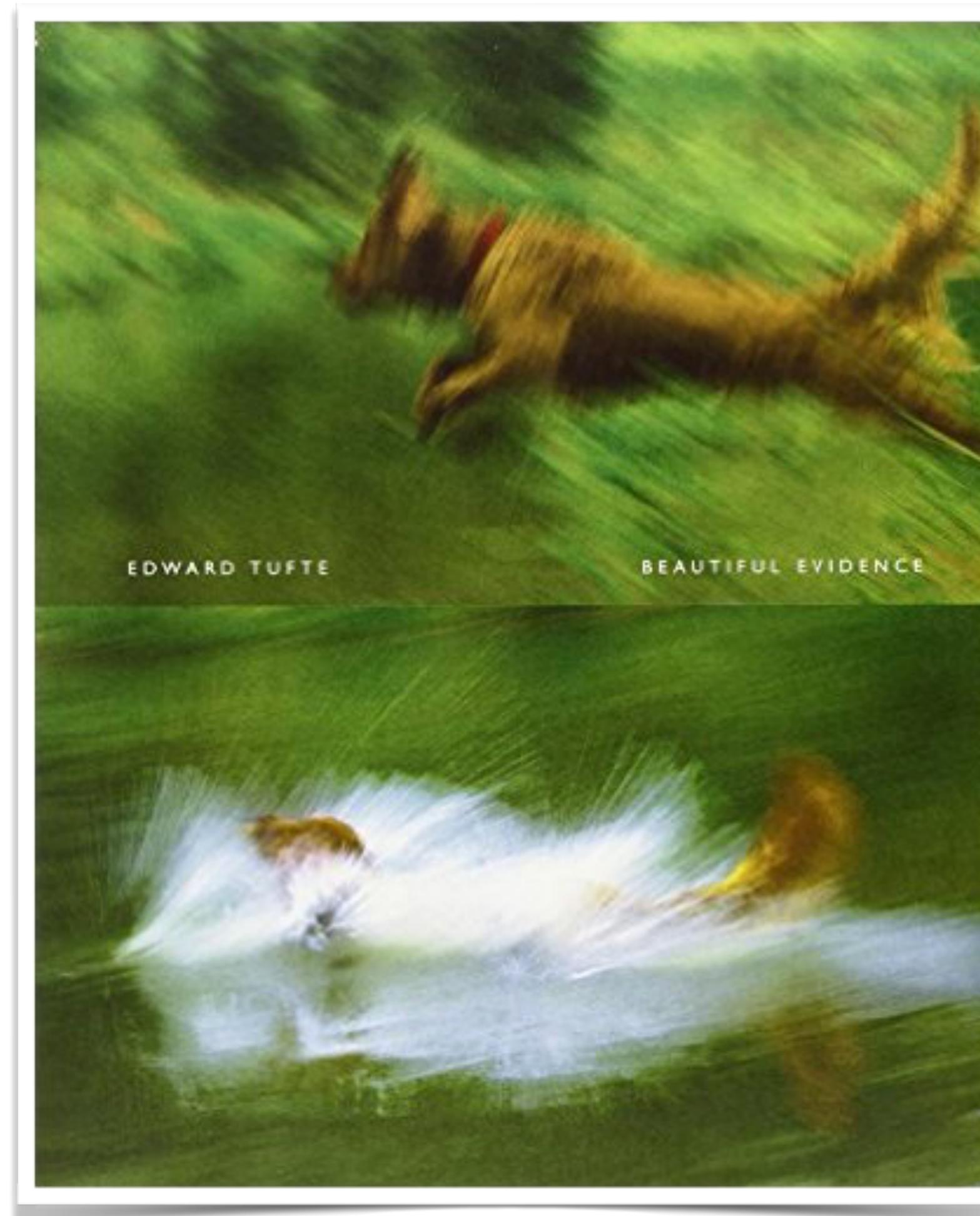
Les nombres d'hommes présents sont représentés par les larges des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en lettres des zones. Le rouge désigne les hommes qui ont été en Russie; le noir ceux qui en sortirent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chier, de Léger, de Fezensac, de Chambray et le journal médical de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

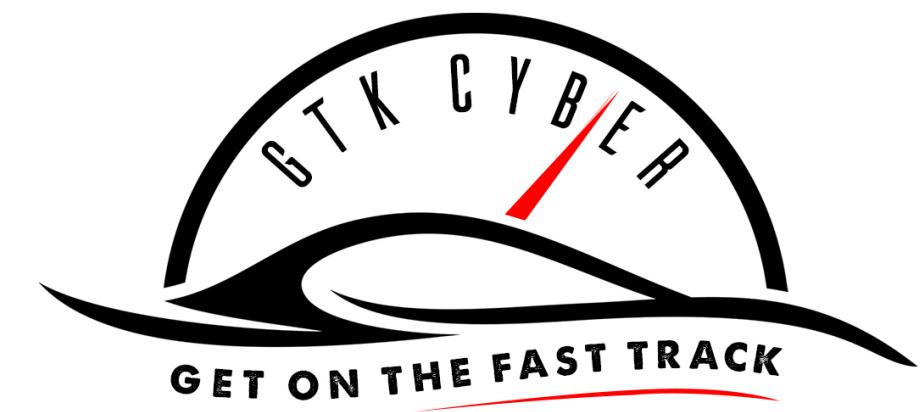
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Napoléon et du Maréchal Davout qui avaient été détachés sur Minsk et Malibor se sont rejoints vers Orsha et Vitebsk, avaient toujours marché avec l'armée.



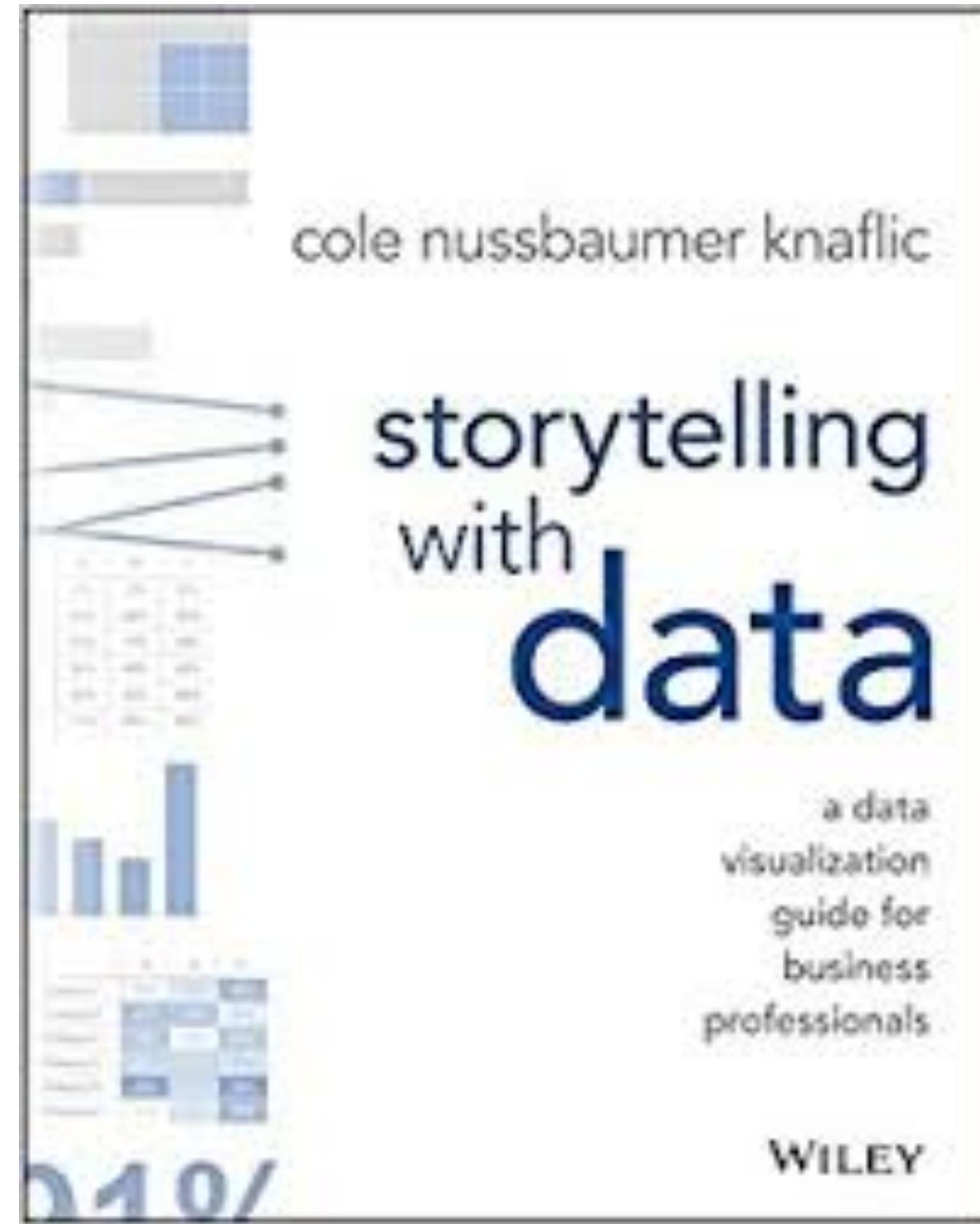


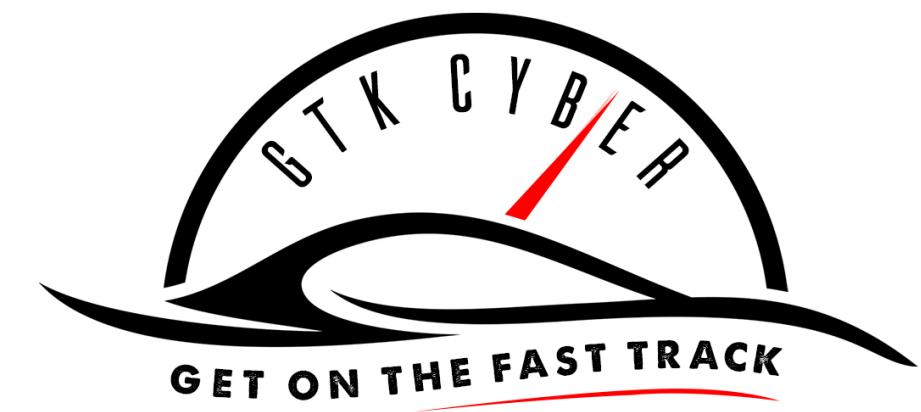
Recommended Reading



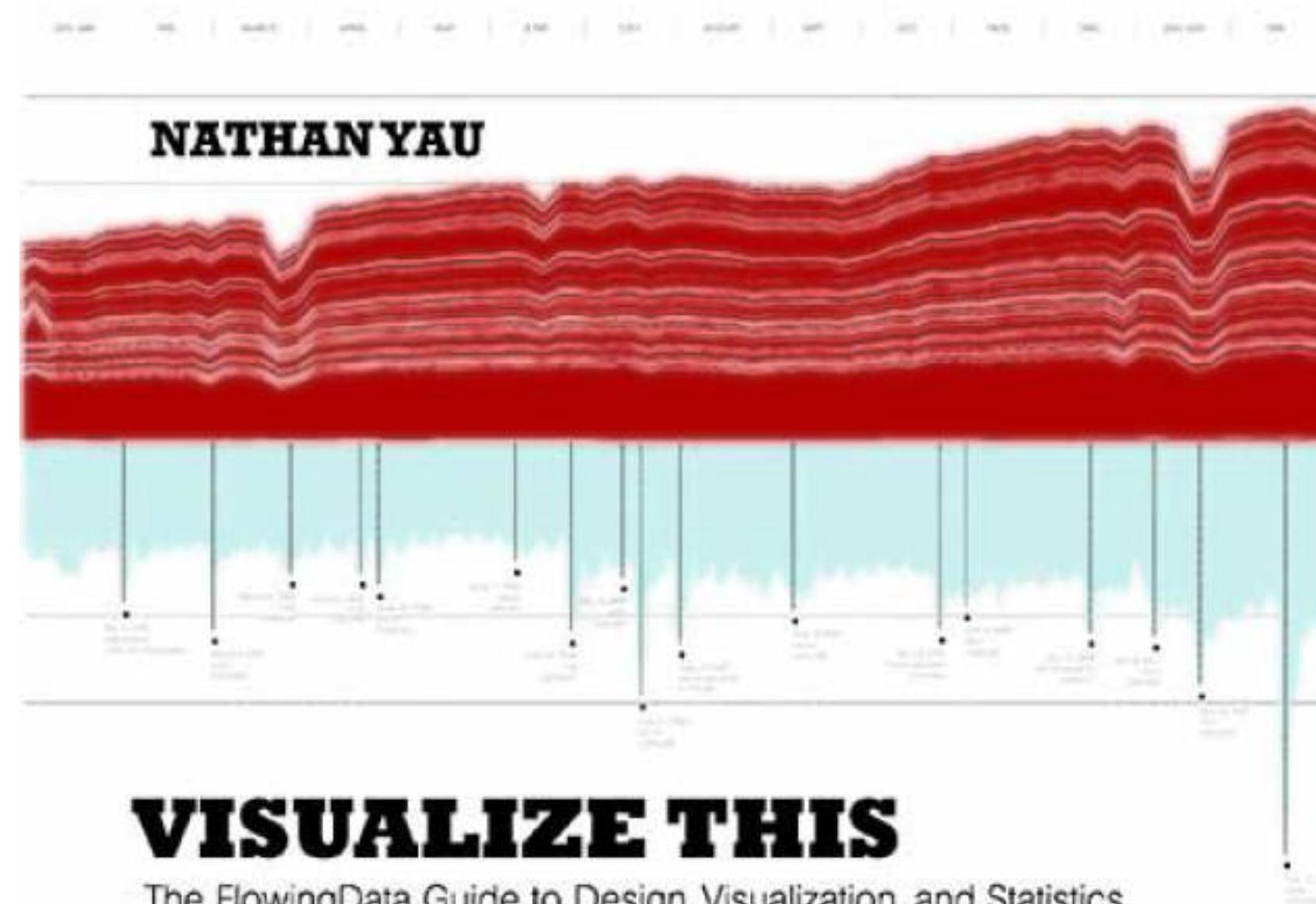


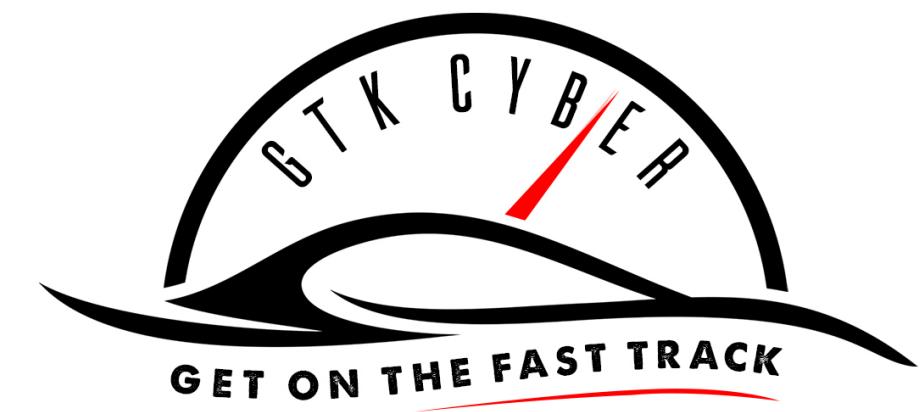
Recommended Reading



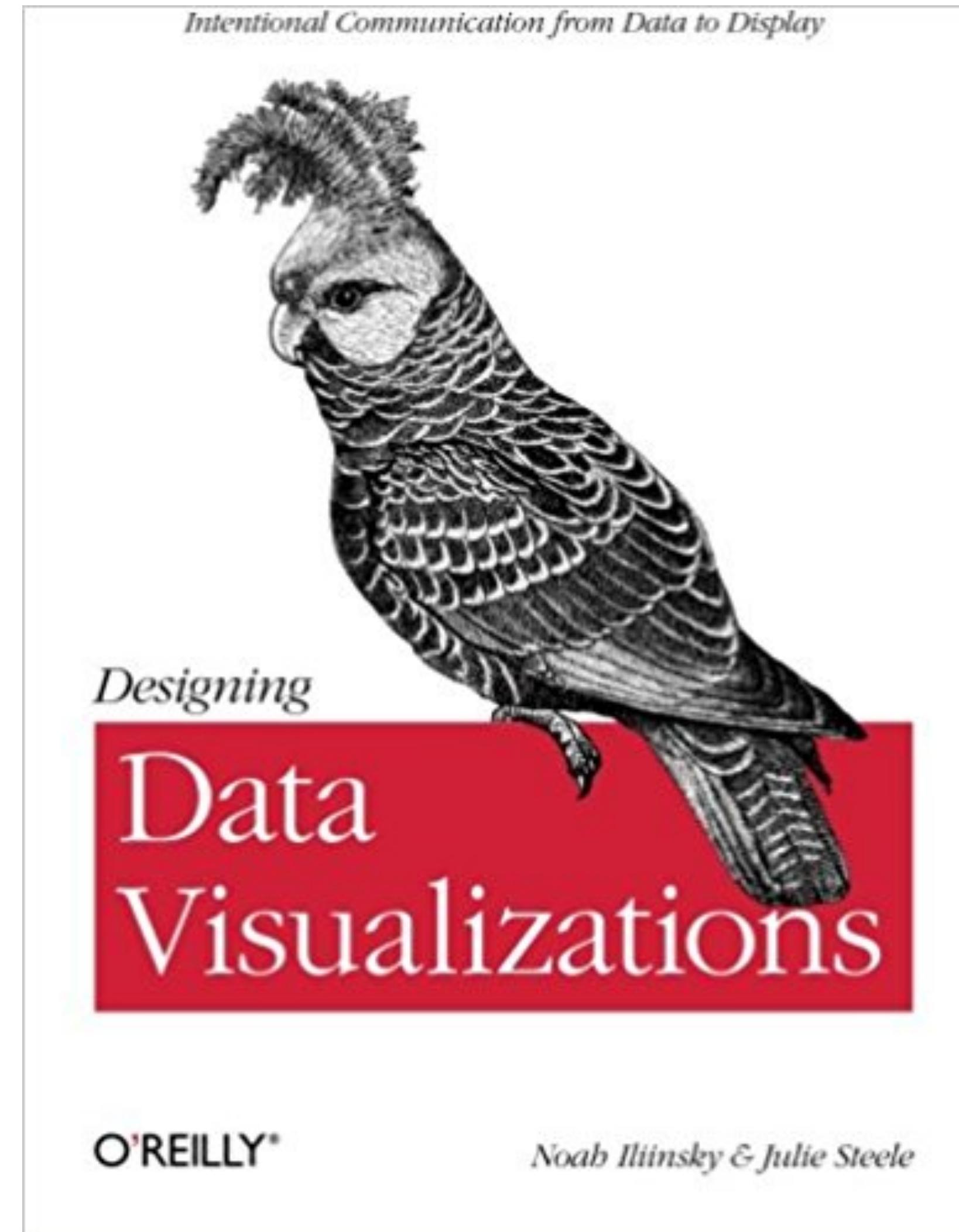


Recommended Reading



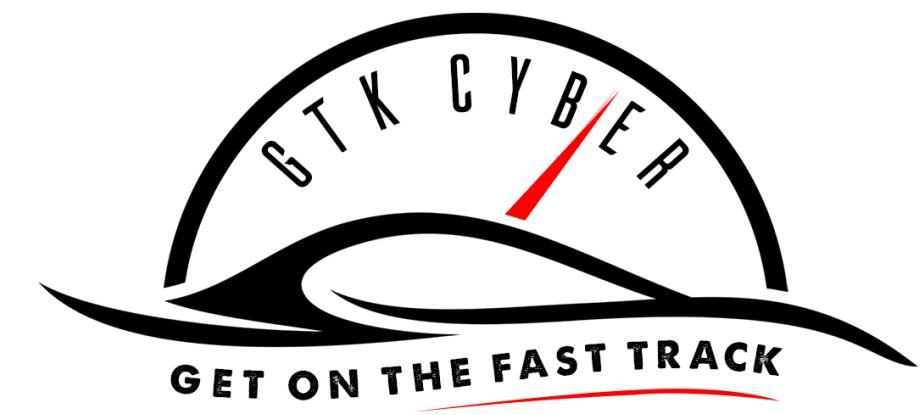


Recommended Reading



Data Visualization in Python

GET ON THE FAST TRACK

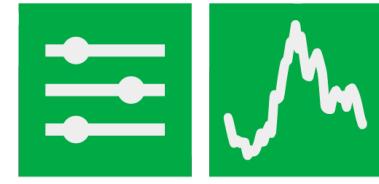


Python Visualization Libraries

- **Matplotlib**: “Lowest” level visualization library. Very powerful, but little abstraction
- **Pandas/Pyplot**: Easier, but requires transformation for complex visualizations
- **Seaborn**: Good for advanced statistical visualizations
- **Altair**: New library: declarative statistical visualization library
- **Plotly / Cufflinks**: Make it interactive!



Other Visualization Libraries

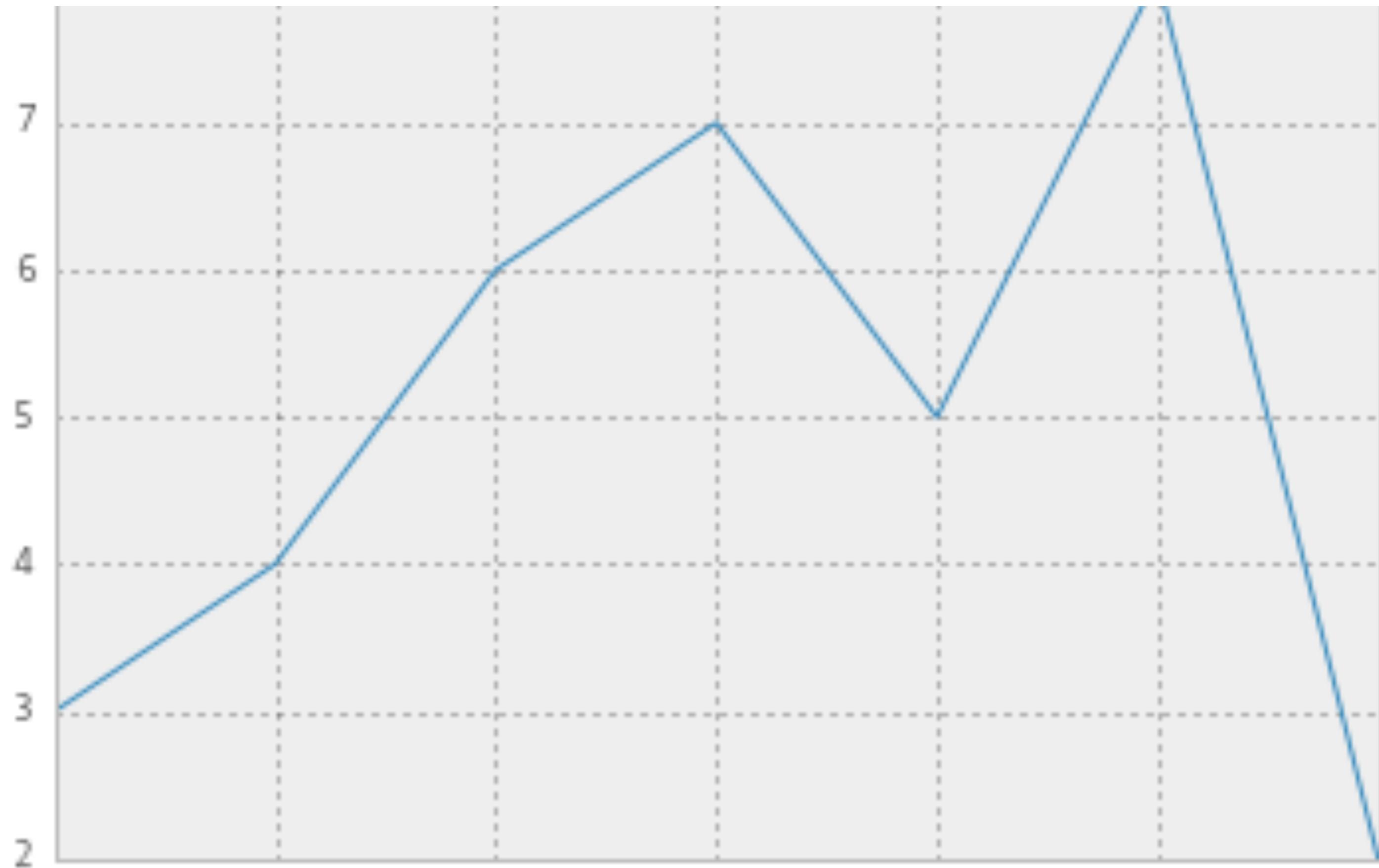
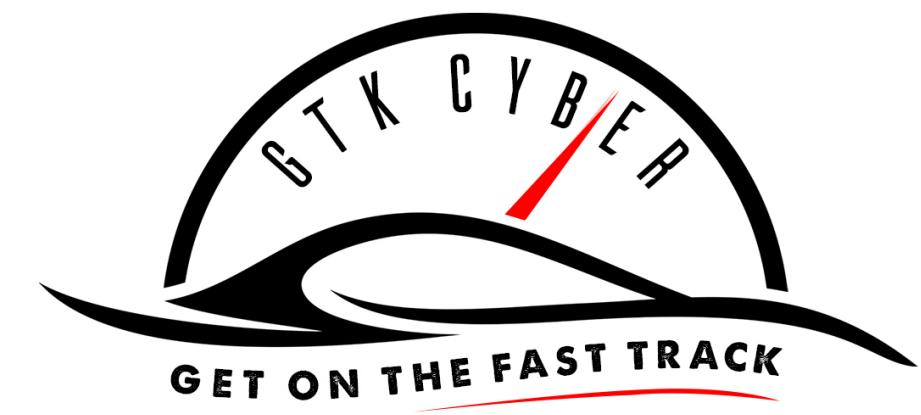


Panel

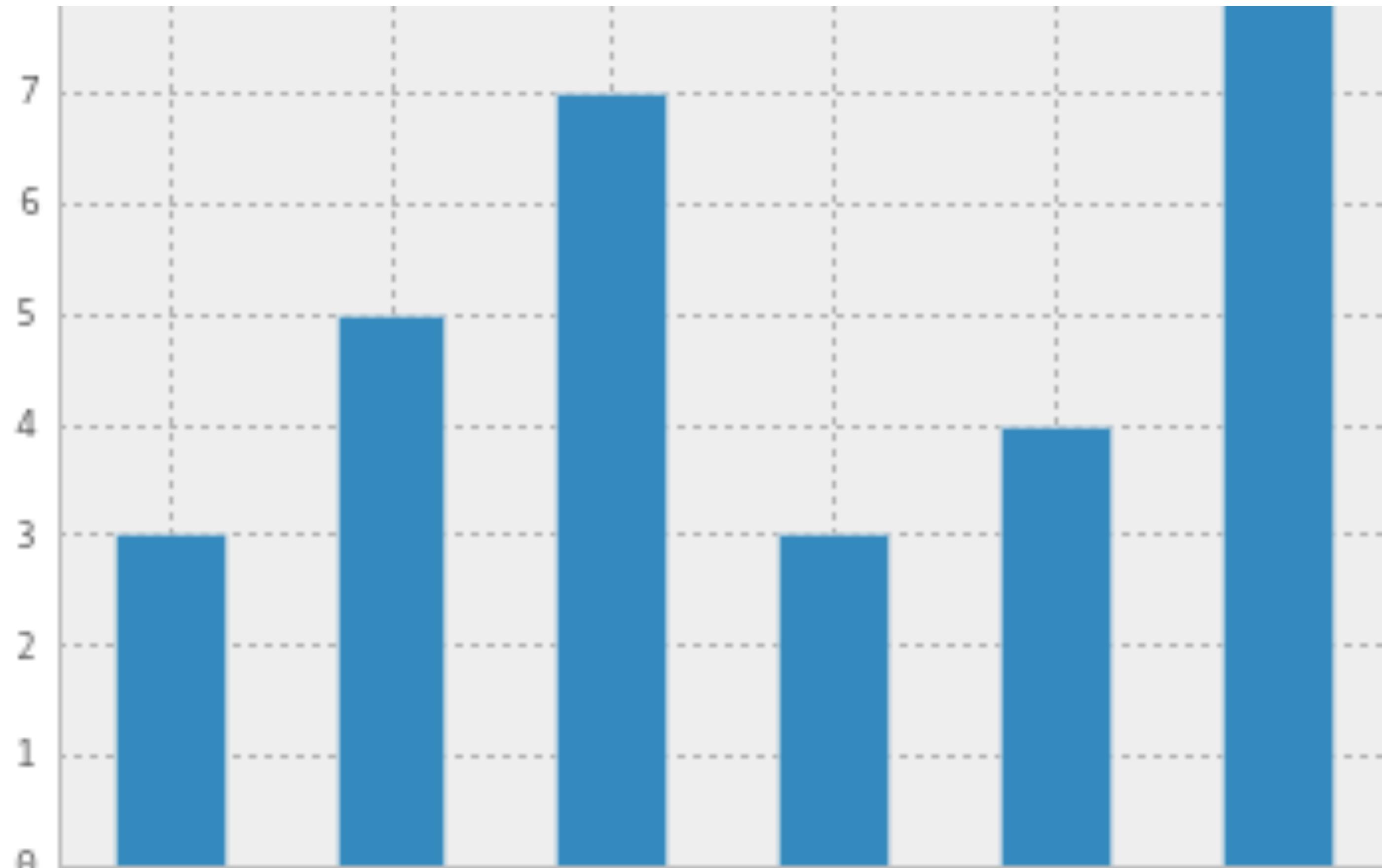
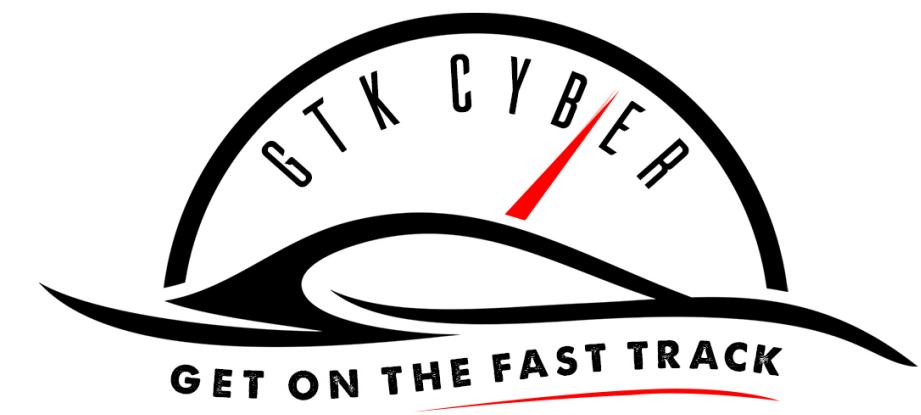
- **Panel**: Powerful library for creating apps and dashboards
- **hvPlot**: Quickly generate interactive plots from data
- **HoloViews**: Helps you make all data instantly visualizable
- **GeoViews**: HoloViews for geo-spatial data
- **ggplot2/PlotNine**: This is Python's implementation of R's ggplot2
- For more useful info about Python visualization, check out
pyviz.org

<https://pyviz.org/tools.html>

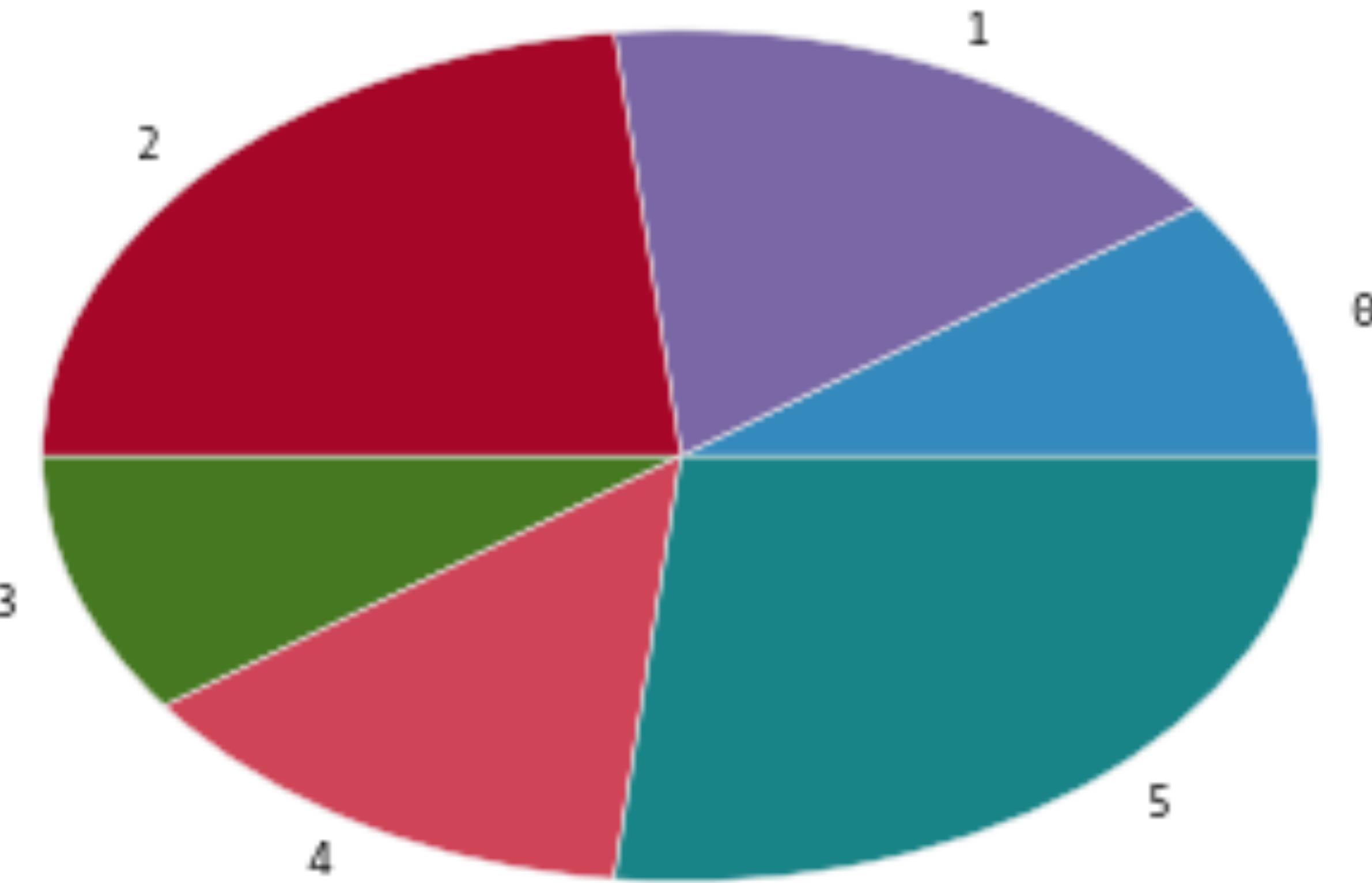
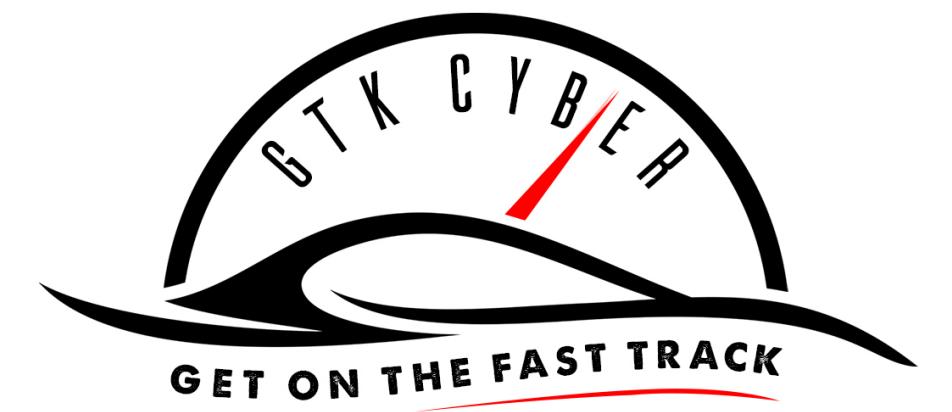
```
import pandas as pd  
import matplotlib.pyplot as plt  
import numpy as np  
import seaborn as sns  
%matplotlib inline
```



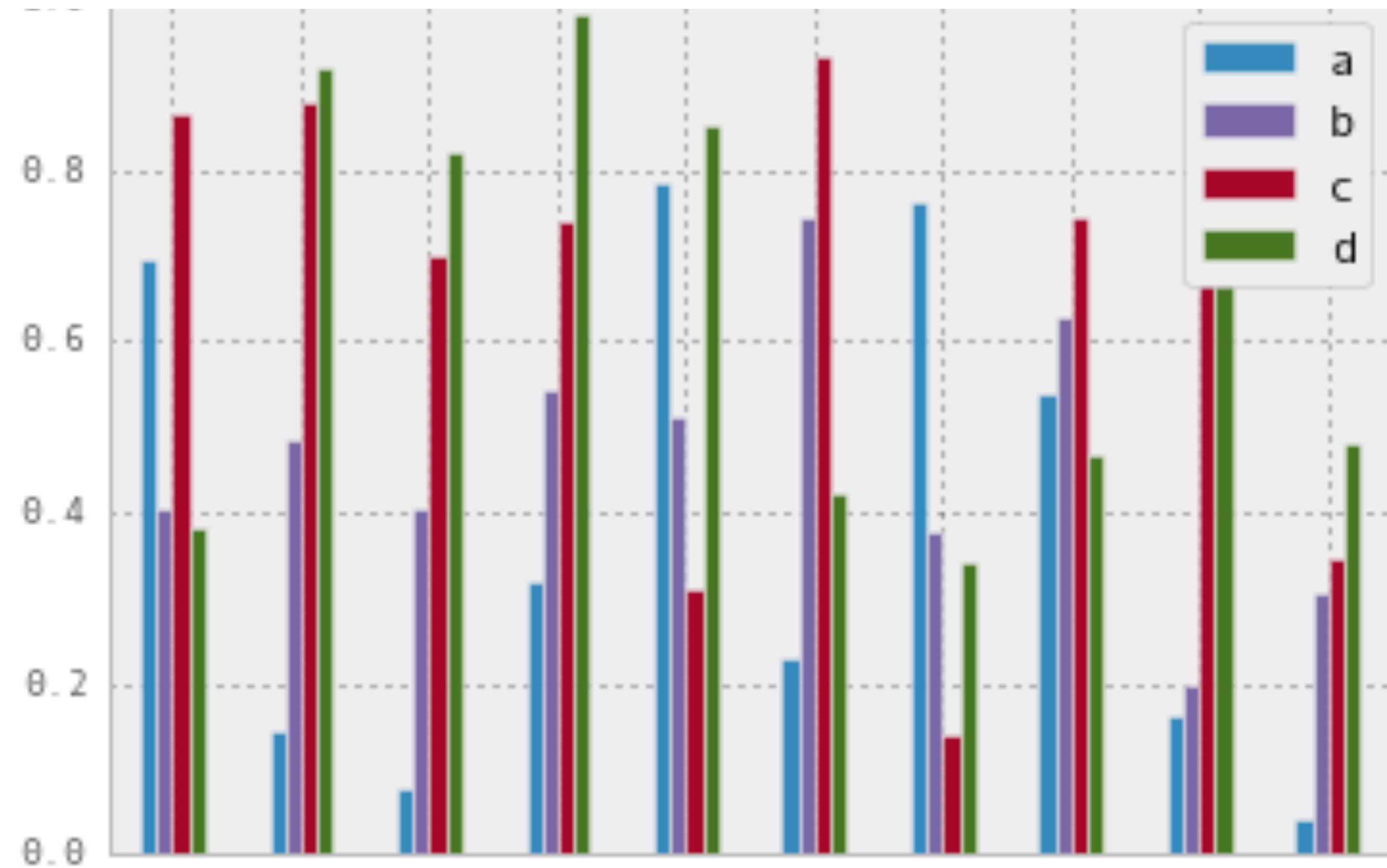
```
data = pd.Series( [ 3 , 4 , 6 , 7 , 5 , 8 , 2 ] )  
graph = data.plot()
```



```
barchart = data.plot( kind="bar" )
```

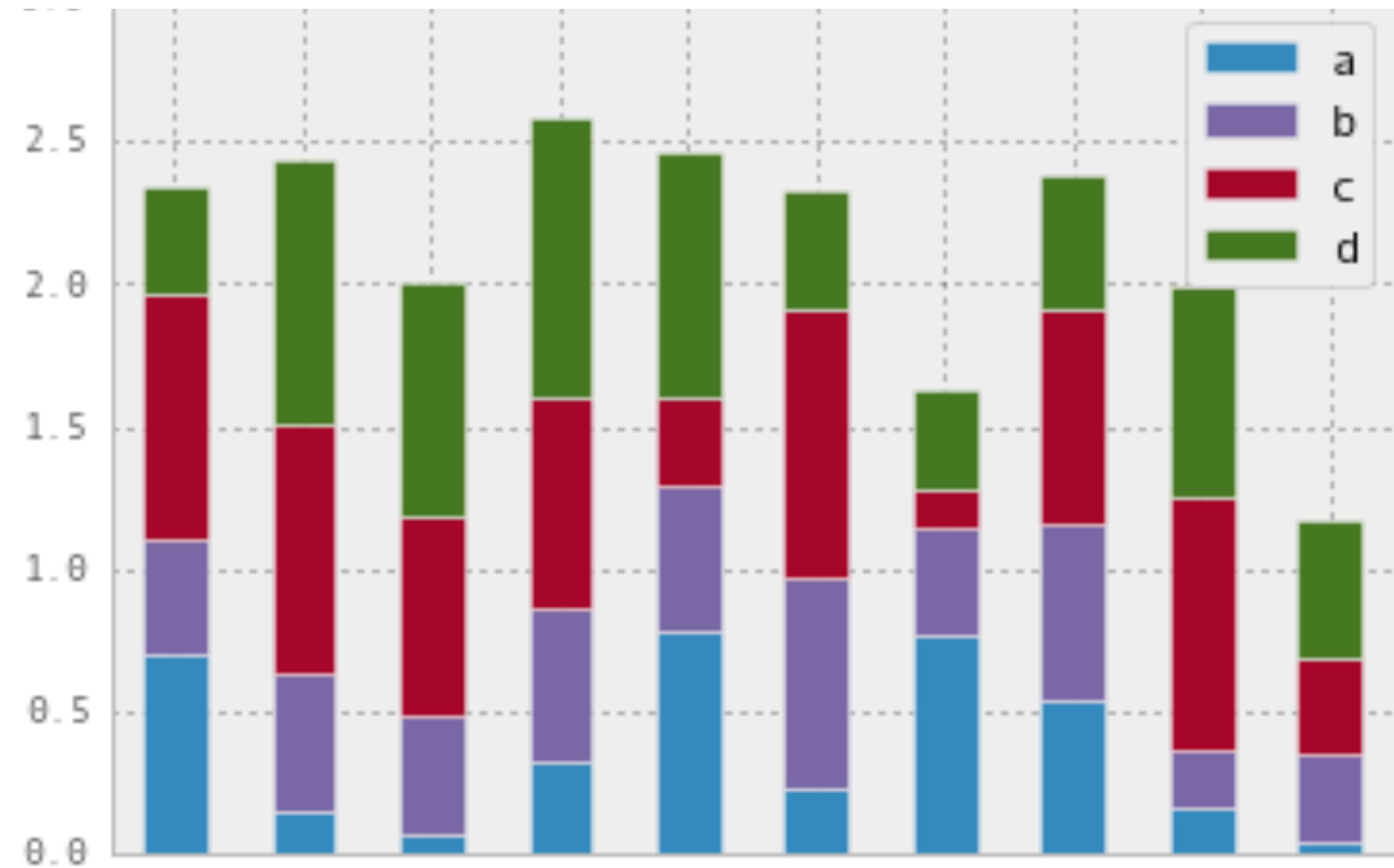
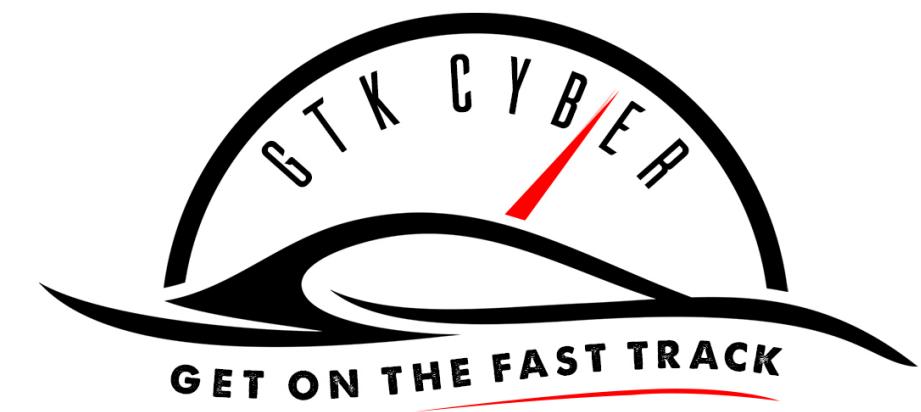


```
piechart = data.plot( kind="pie" )
```

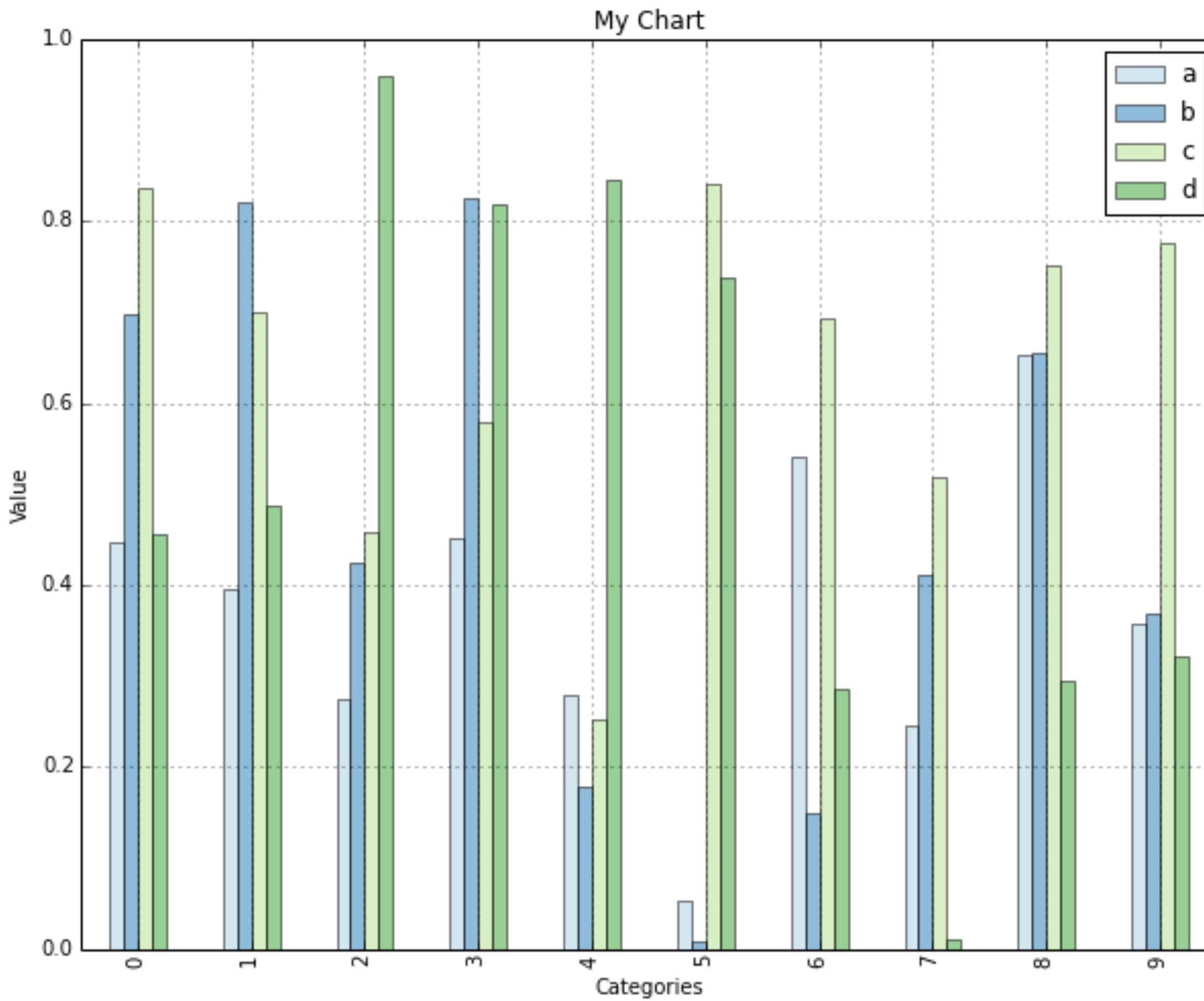
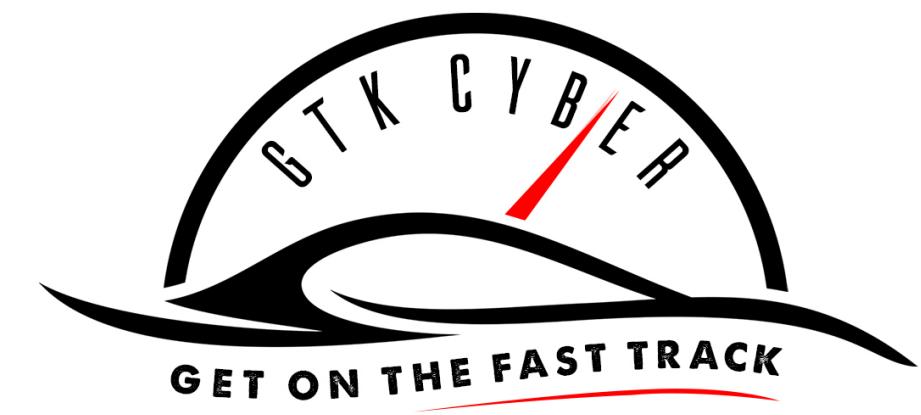


```
df2 = pd.DataFrame(np.random.rand(10, 4), \
columns=[ 'a' , 'b' , 'c' , 'd' ] )
```

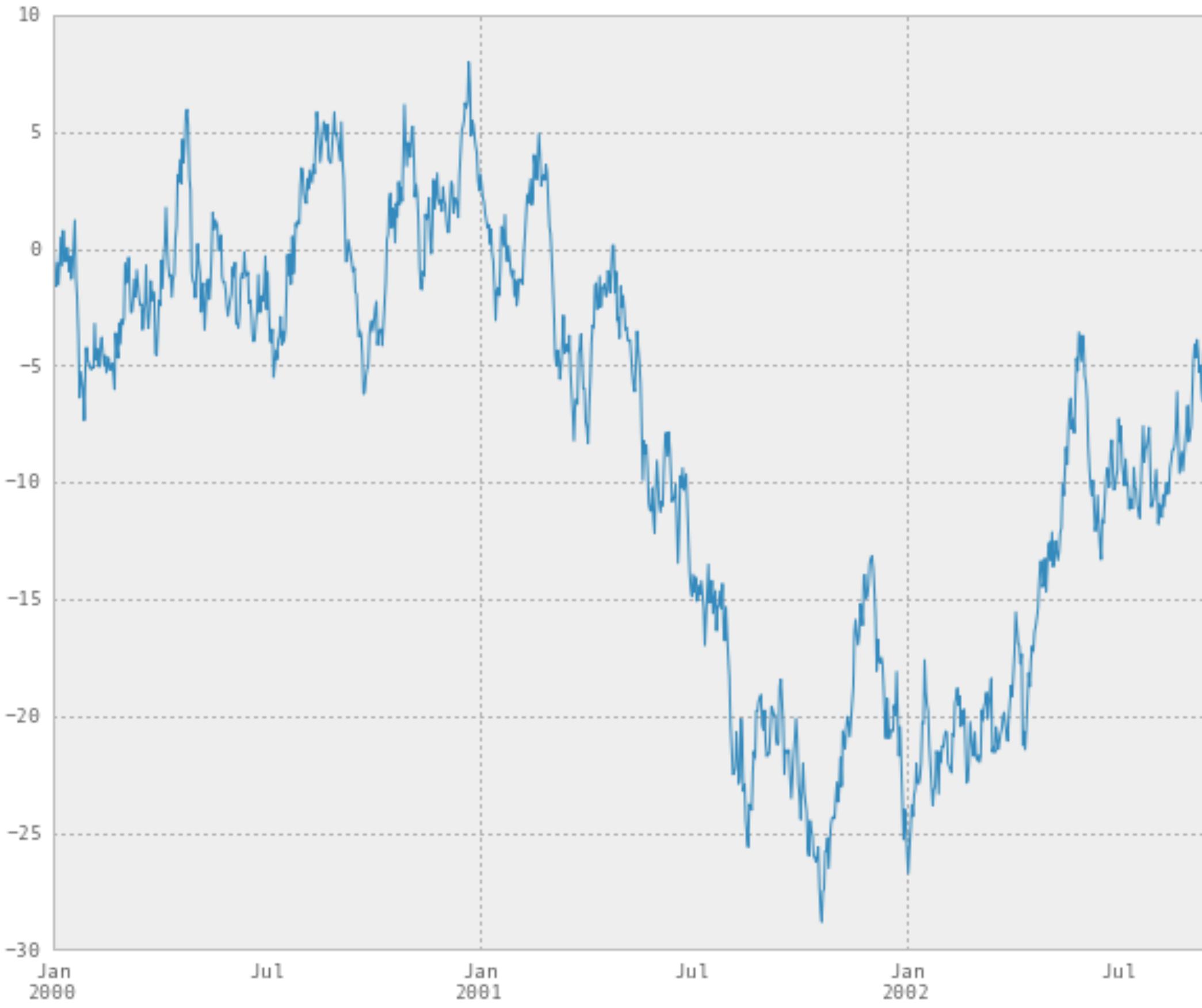
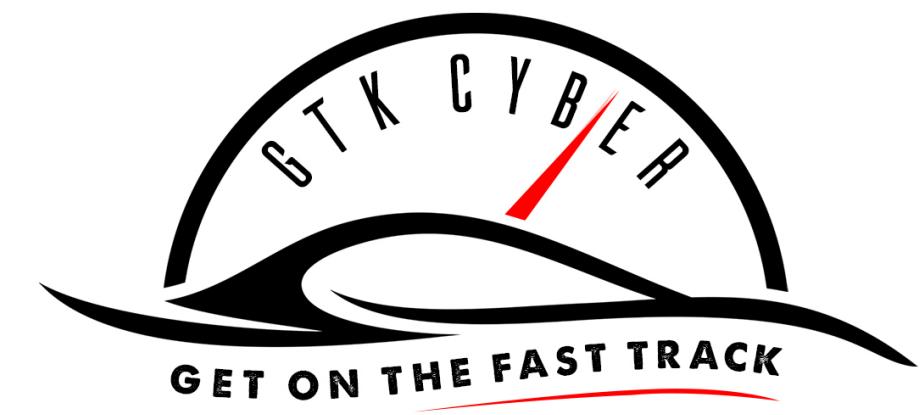
```
df2.plot( kind='bar' )
```



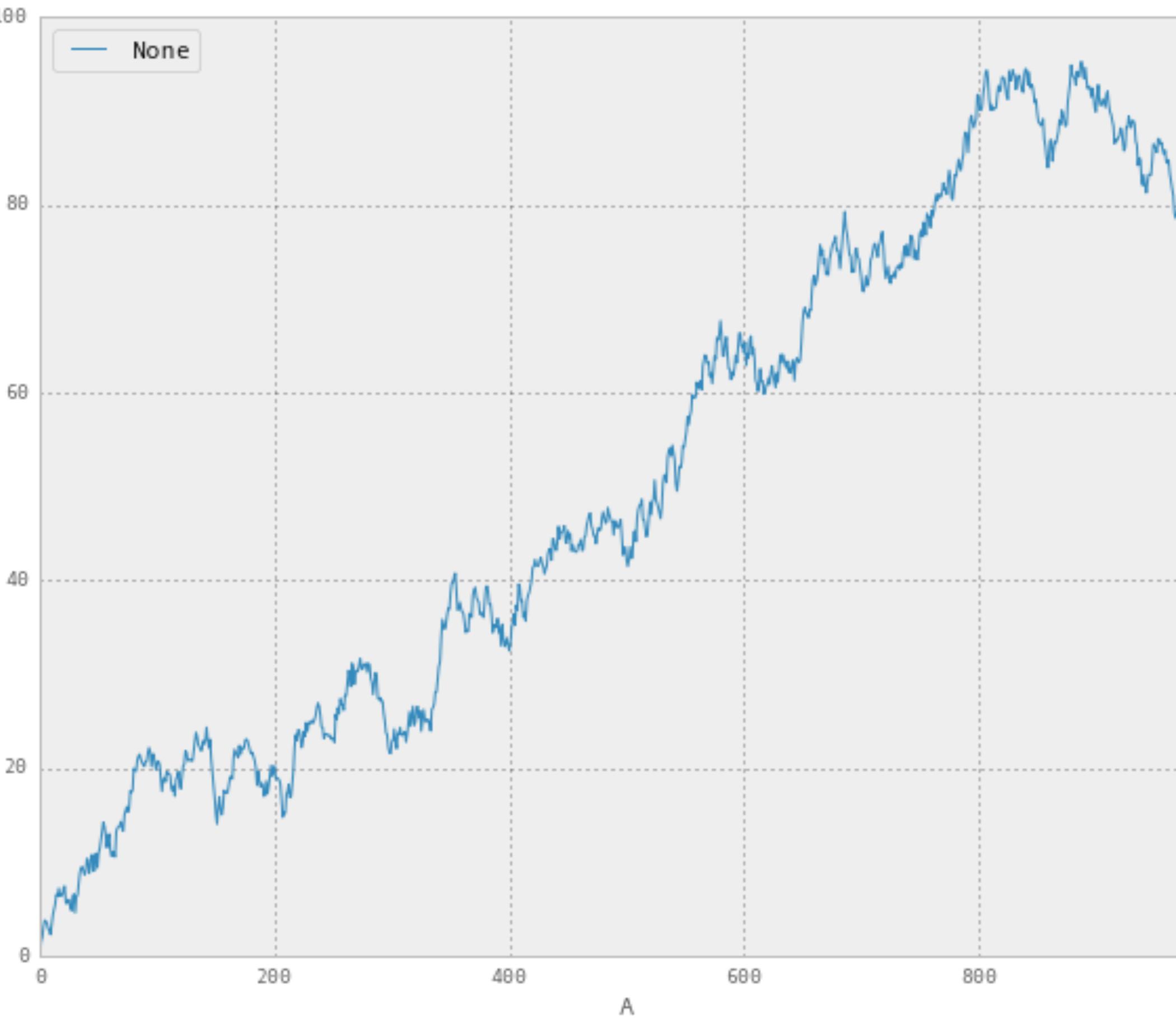
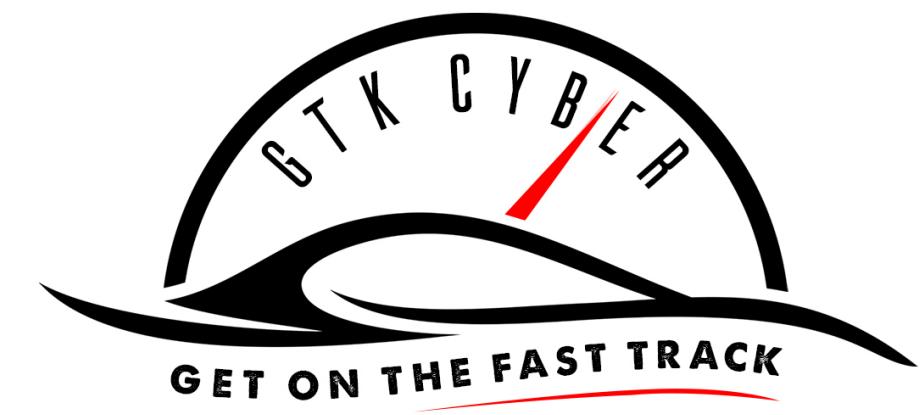
```
df2.plot( kind='bar', stacked=True )
```



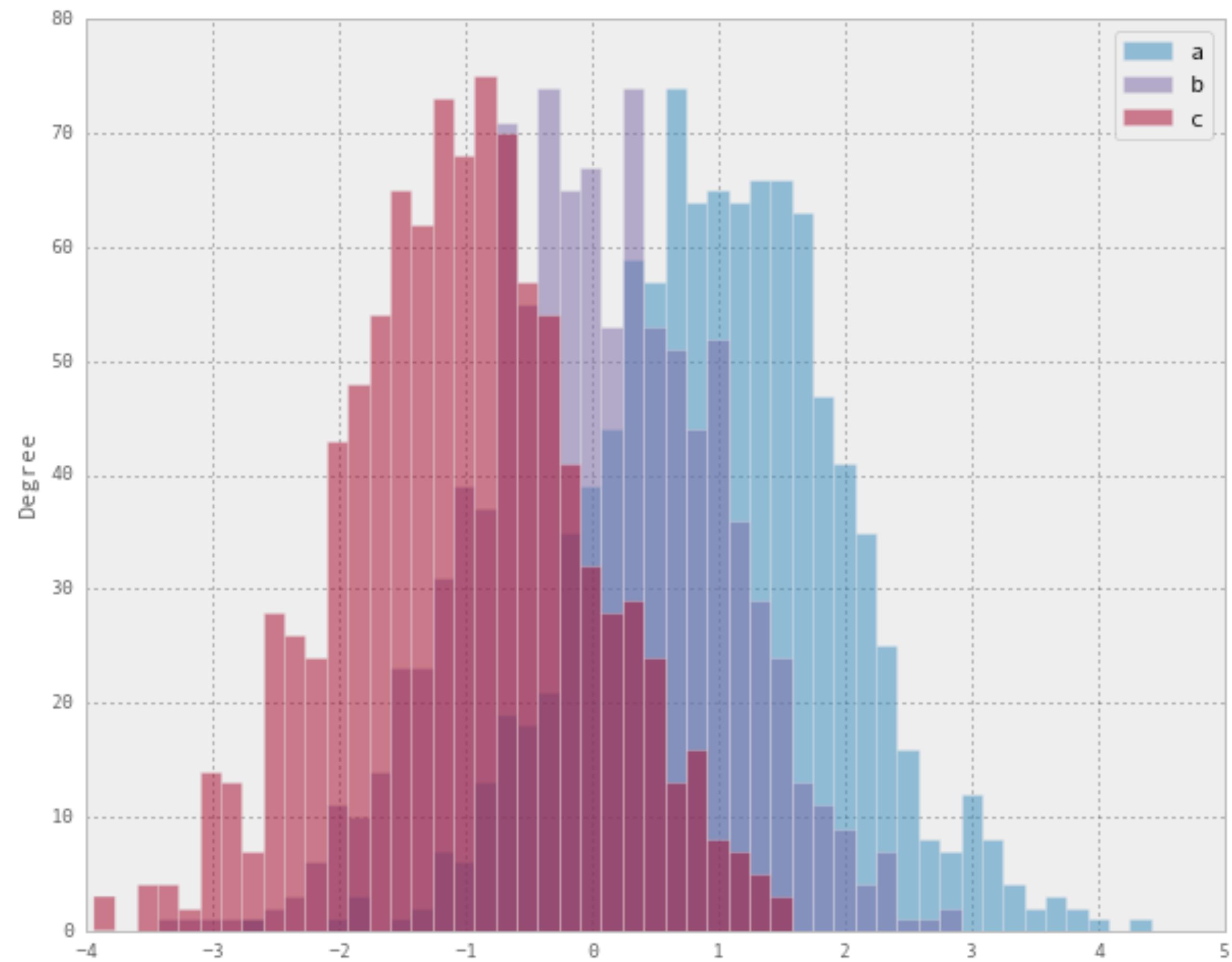
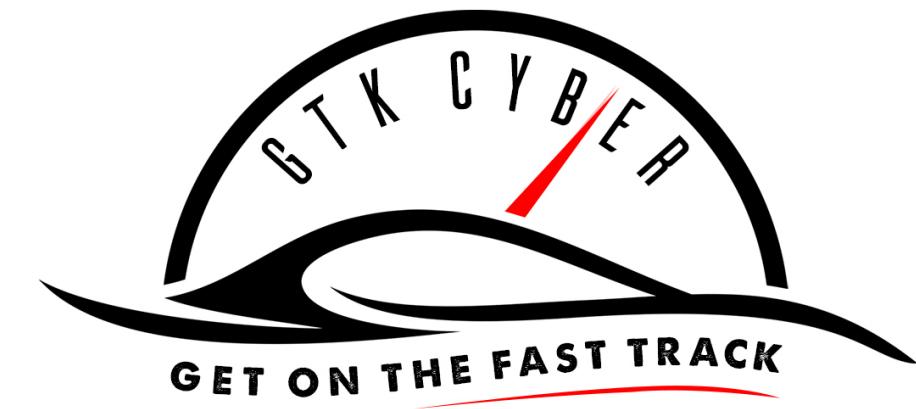
```
df2.plot( kind='bar' ,  
          color=('#a6cee3', '#1f78b4', '#b2df8a', '#33a02c') ,  
          alpha=0.5 ,  
          width=0.5 ,  
          figsize=(10,8))  
plt.title( "My Chart" )  
plt.xlabel( "Categories" )  
plt.ylabel( "Value" )
```



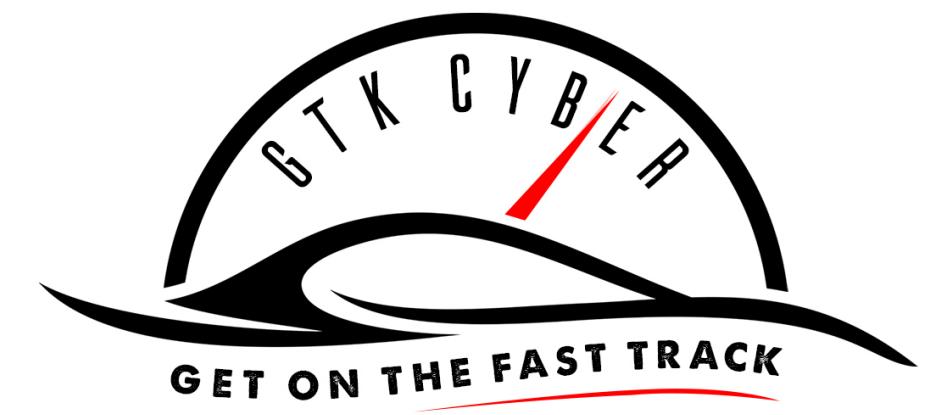
```
ts = pd.Series(np.random.randn( 1000 ),  
index=pd.date_range('1/1/2000', periods=1000))  
ts = ts.cumsum()  
timeseriesChart = ts.plot( figsize=(10, 8) )
```



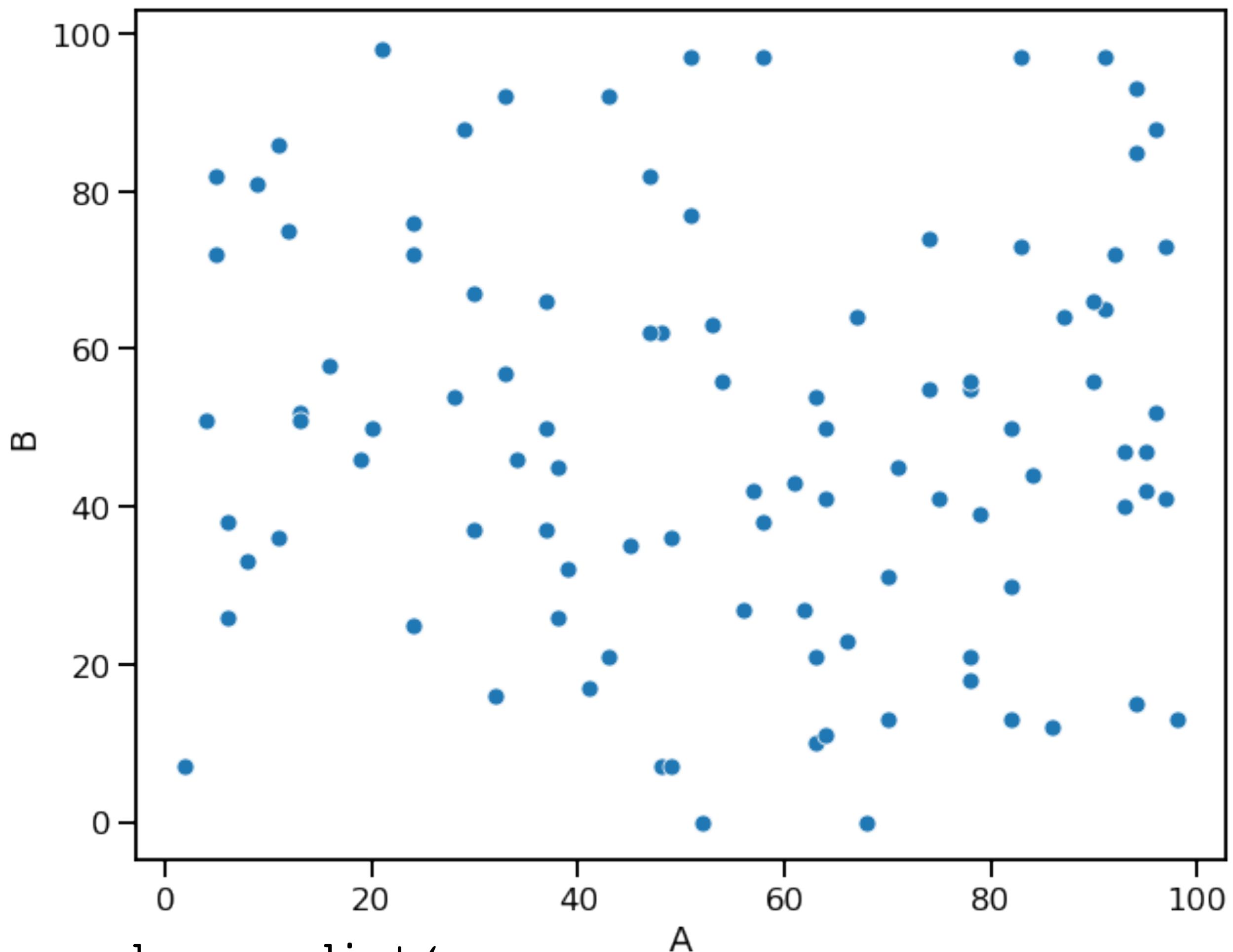
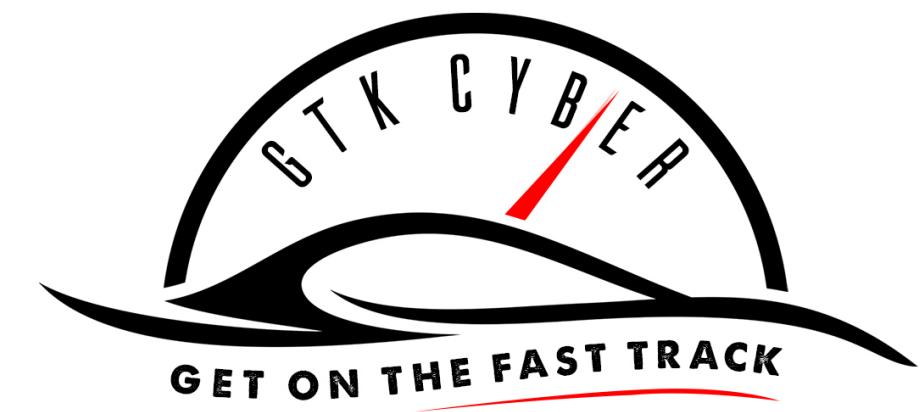
```
df3 = pd.DataFrame(np.random.randn(1000, 2),
                   columns=[ 'B' , 'C' ]).cumsum()
df3[ 'A' ] = pd.Series(list(range(len(df3))))  
  
df3.plot( x='A' , y='B' )
```



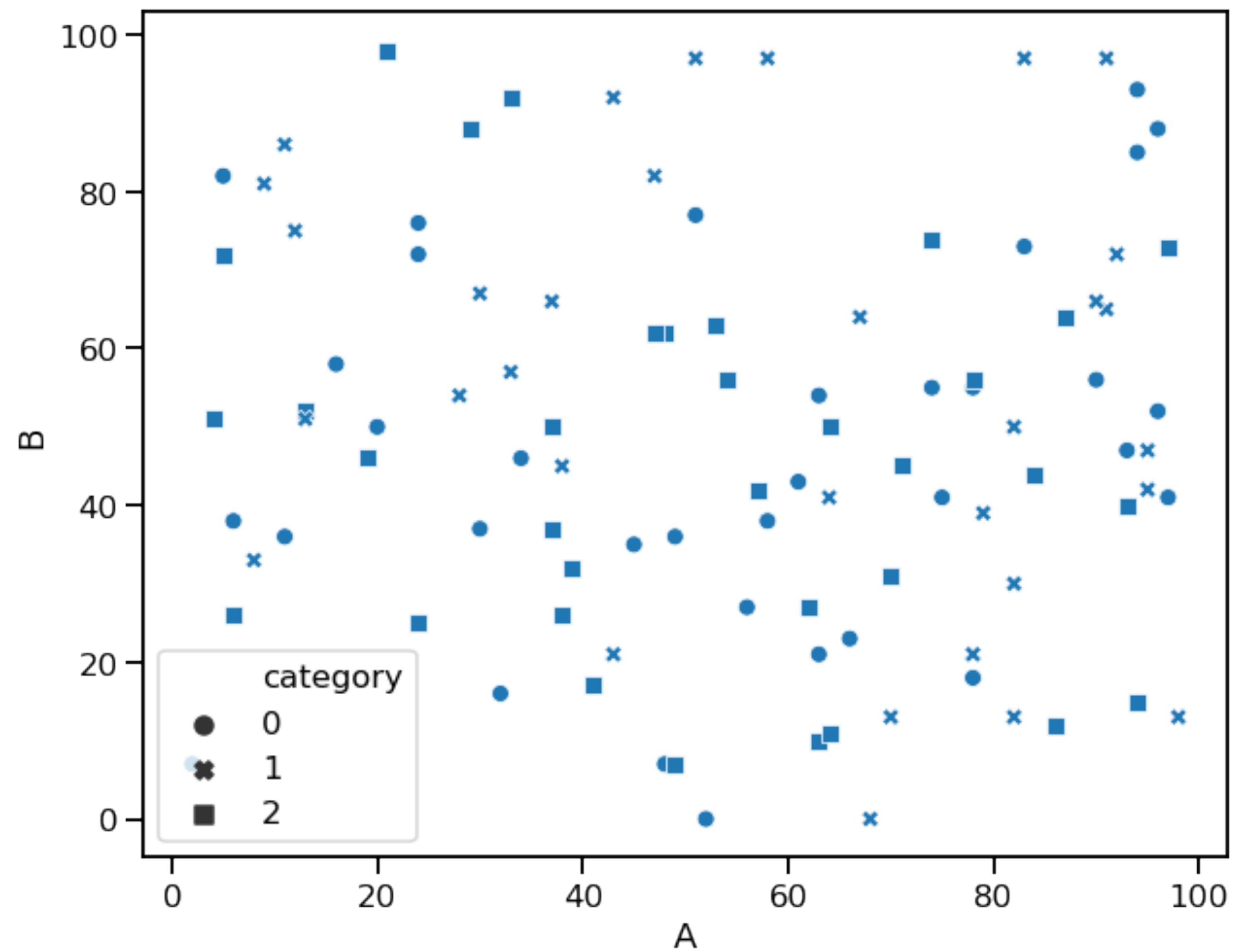
```
df4.plot(kind='hist',  
         alpha=0.5,  
         bins=50 )
```



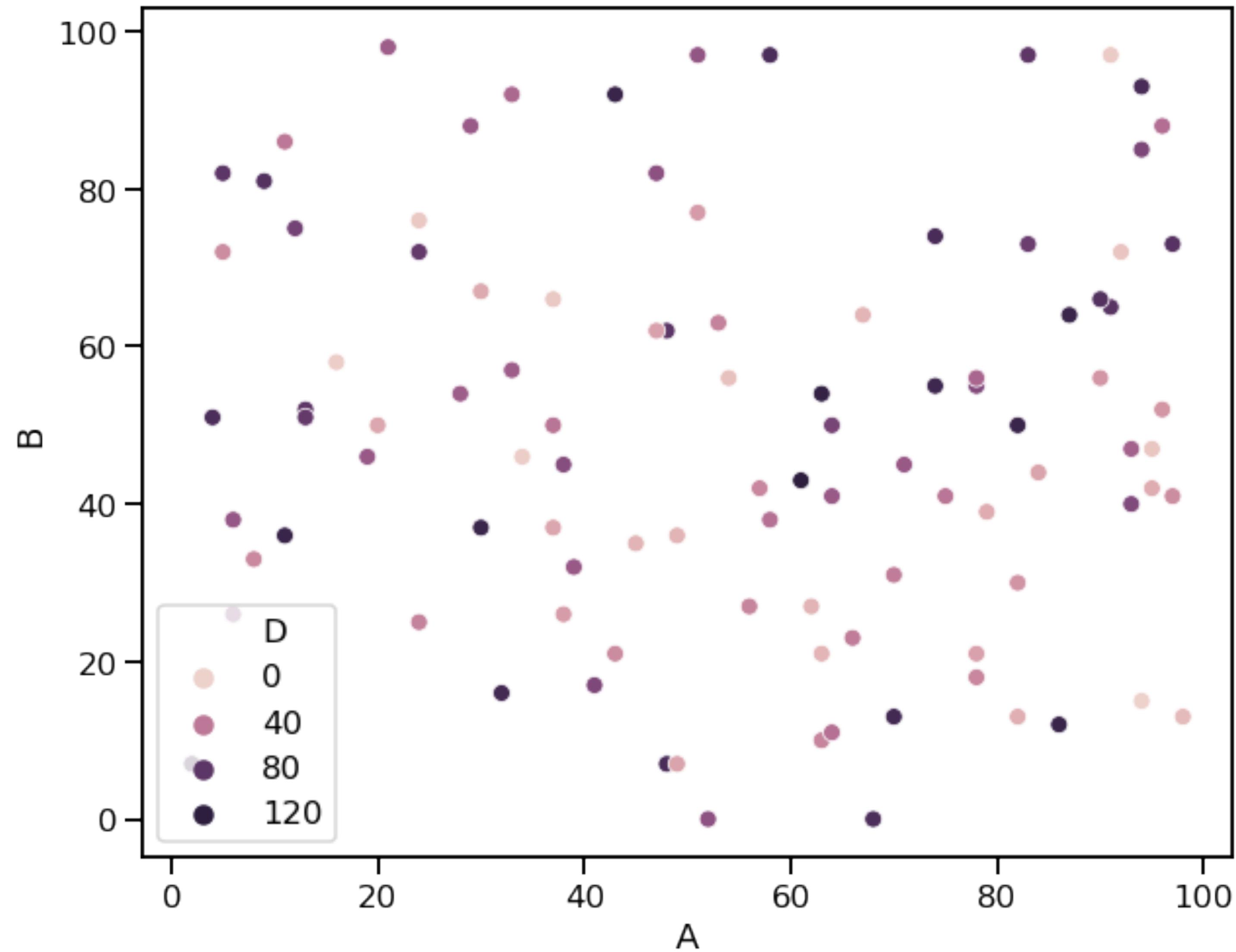
import seaborn as sns



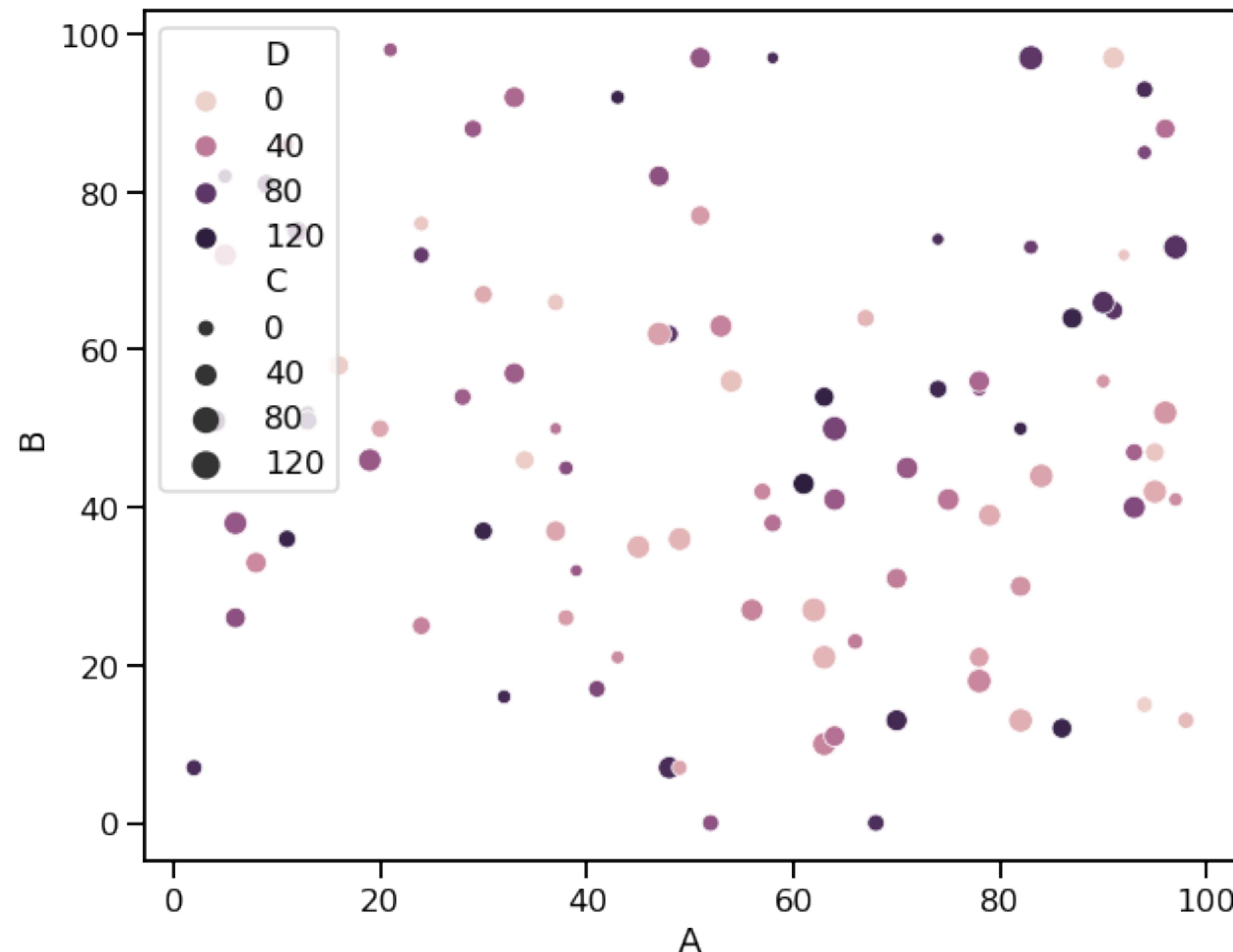
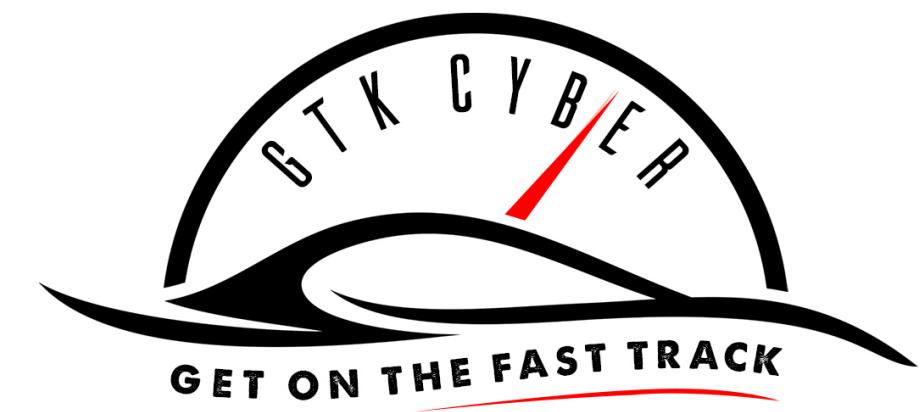
```
df = pd.DataFrame(np.random.randint(0,100,size=(100, 4)),  
                  columns=list('ABCD'))  
df['category'] = df['C'] % 3  
  
sns.scatterplot(x=df['A'], y=df['B'], data=df)
```



```
sns.scatterplot(x=df[ 'A' ], y=df[ 'B' ], style=df[ 'category' ])
```



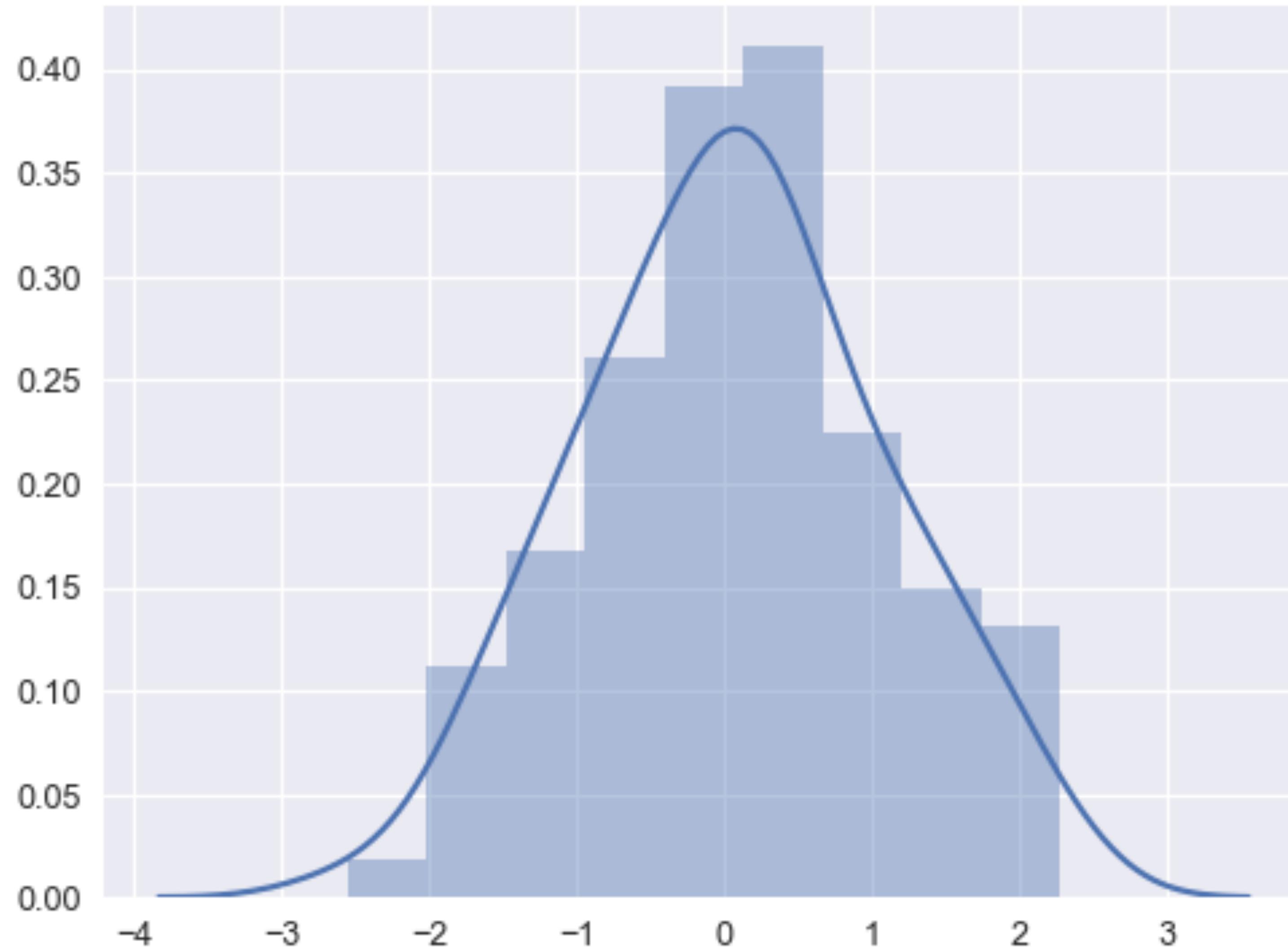
```
sns.scatterplot(x=df['A'], y=df['B'], hue=df['D'])
```



```
sns.scatterplot(x=df['A'], y=df['B'], size=df['C'],
hue=df['D'])
```

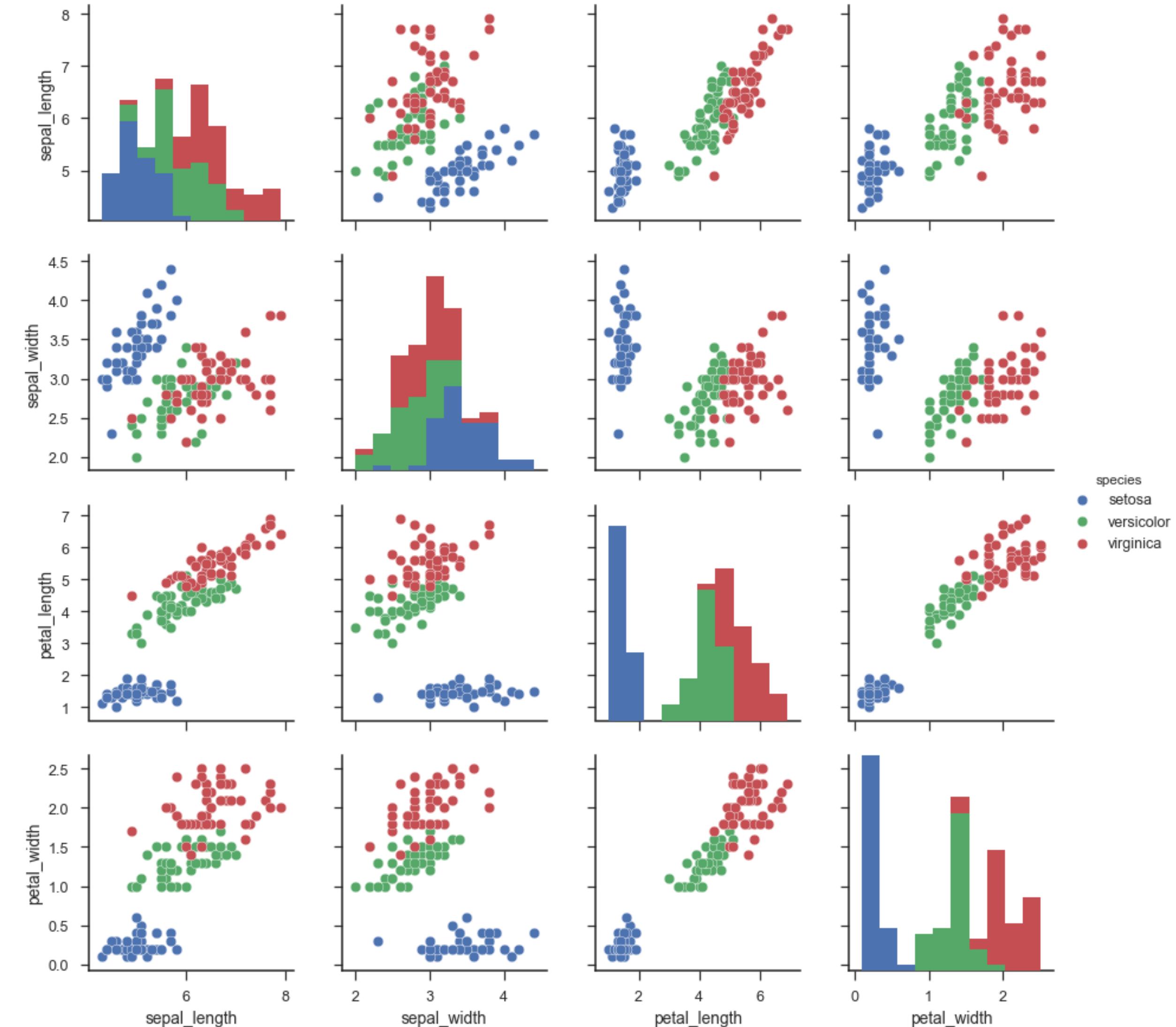


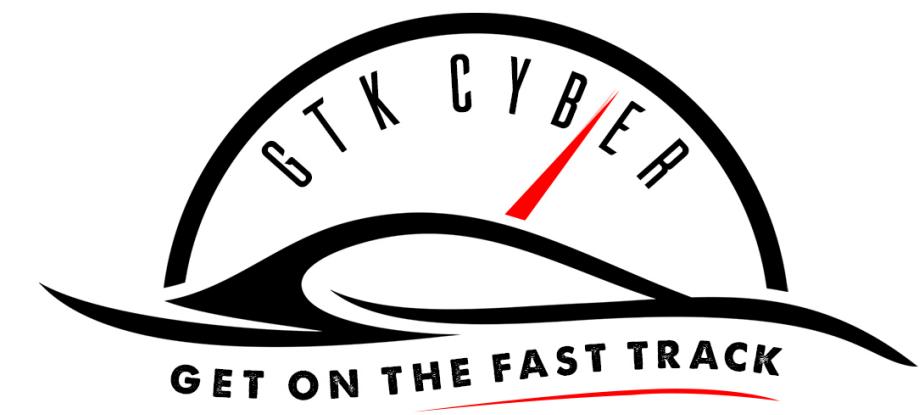
`ax = sns.distplot(<data>)`



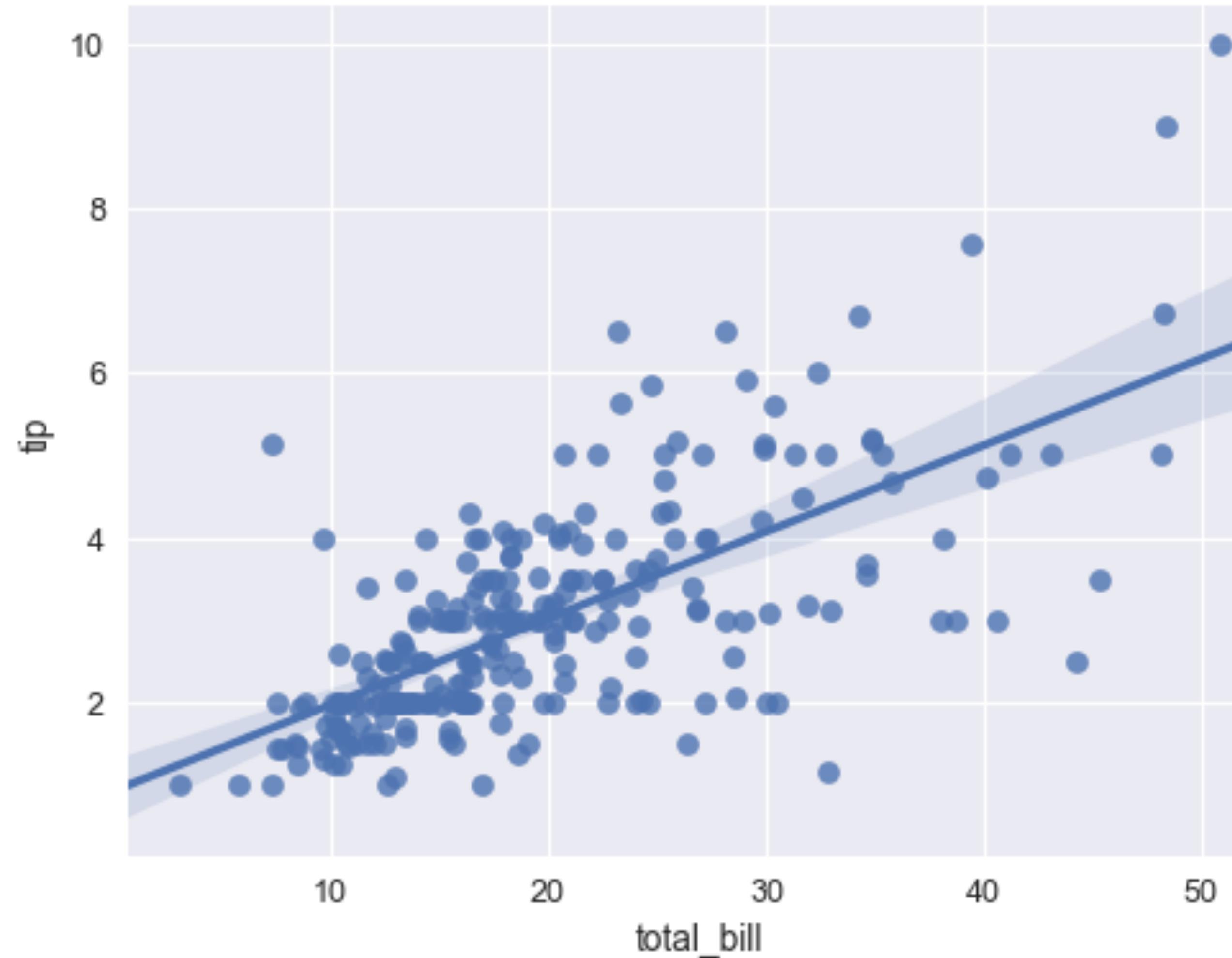


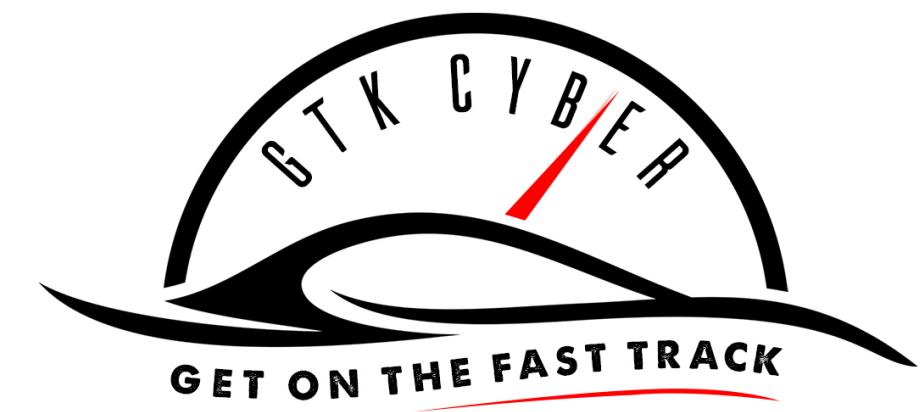
`sns.pairplot(<data>, hue="<target>")`



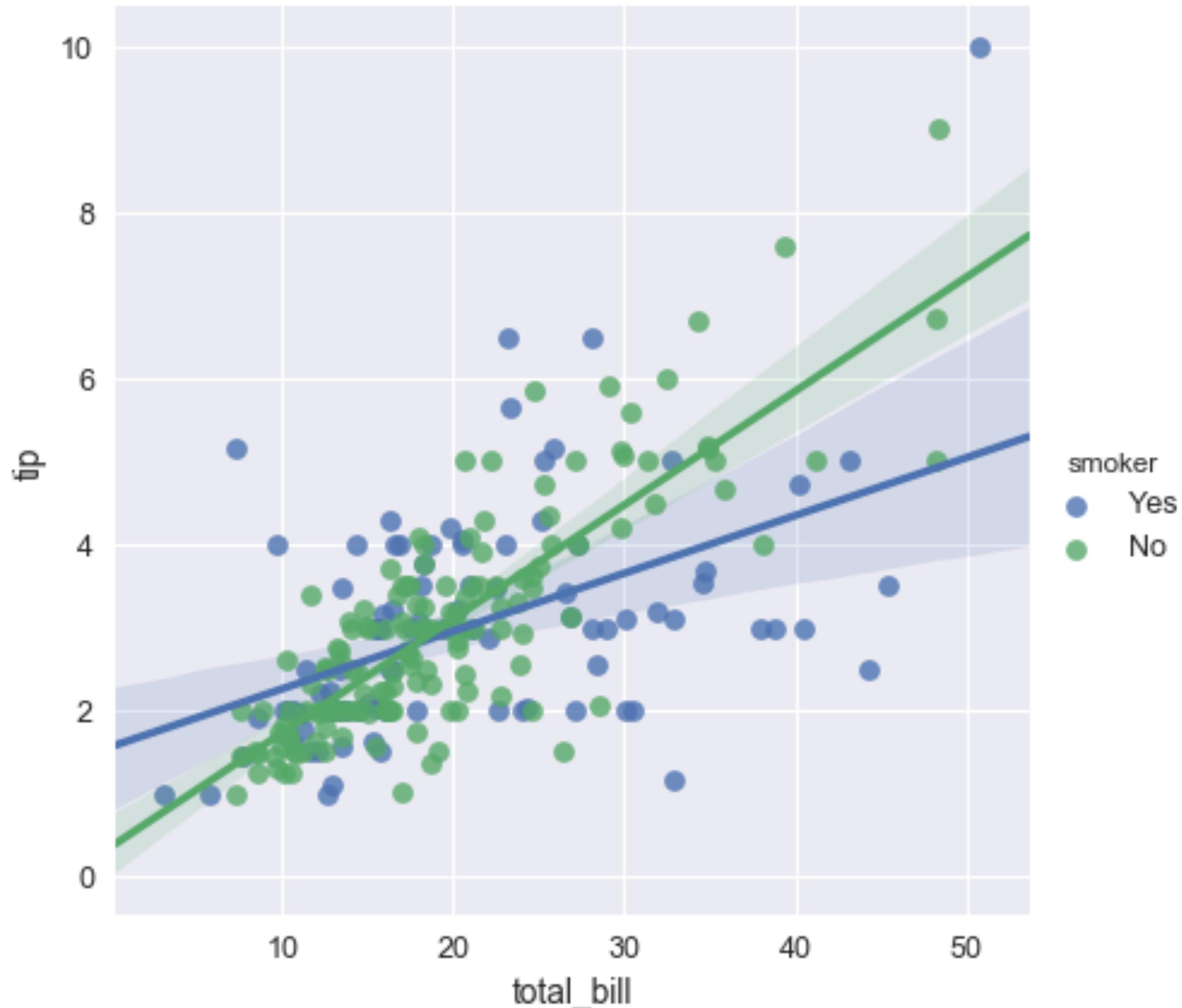


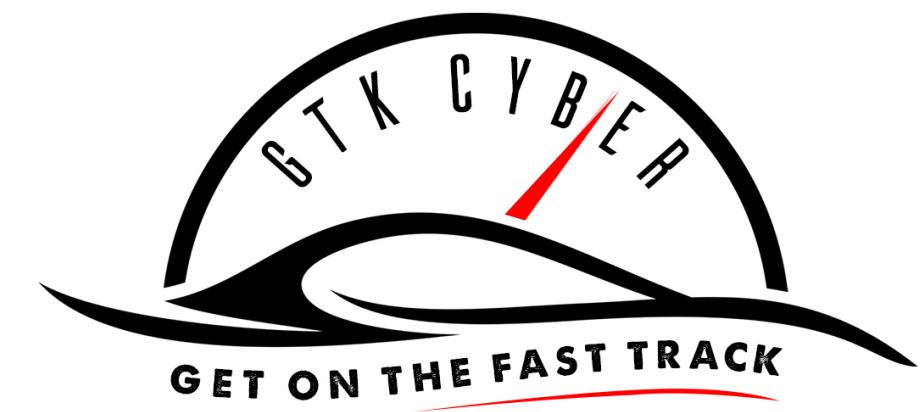
```
ax = sns.regplot(x="total_bill",  
                  y="tip", data=tips)
```



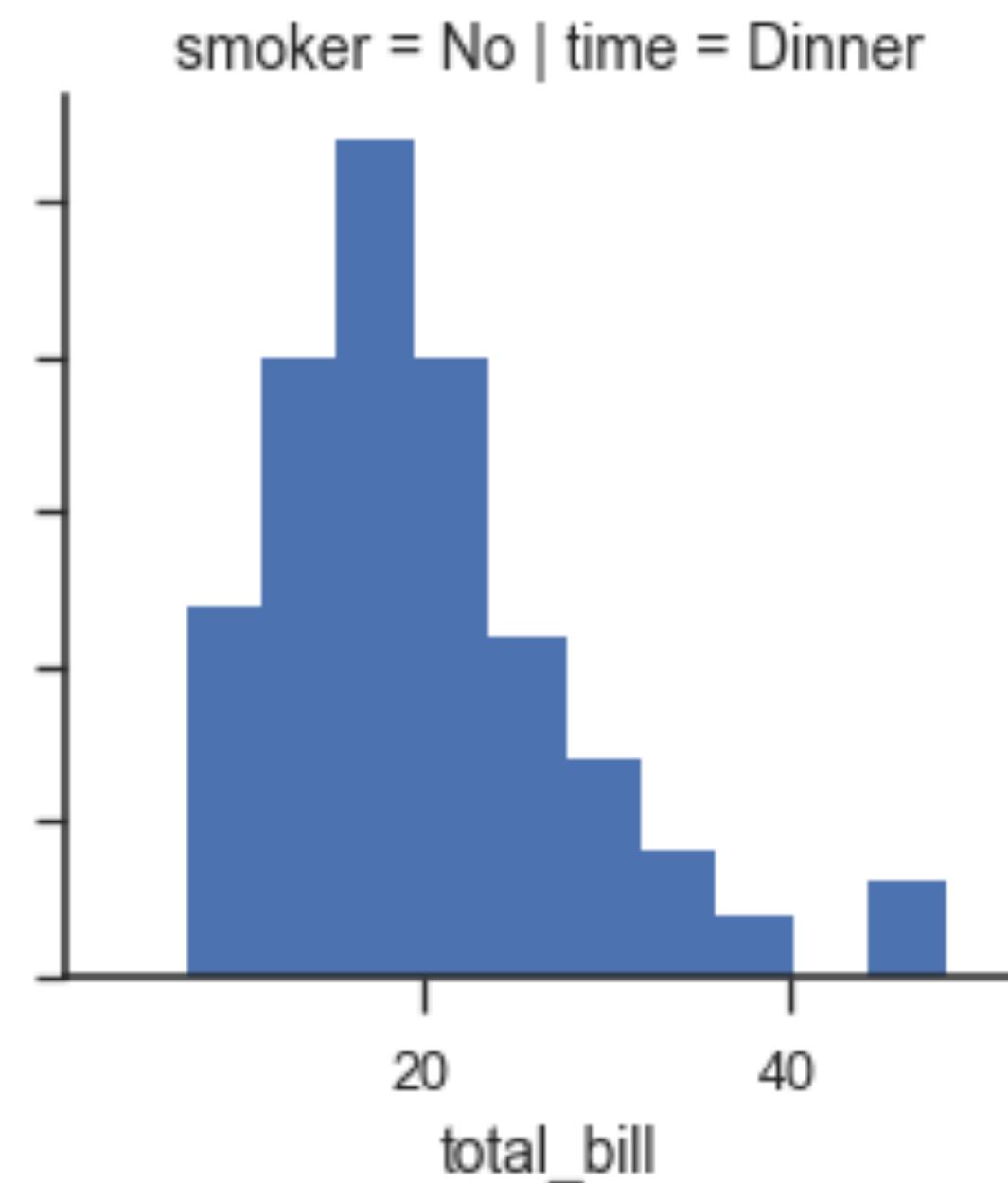
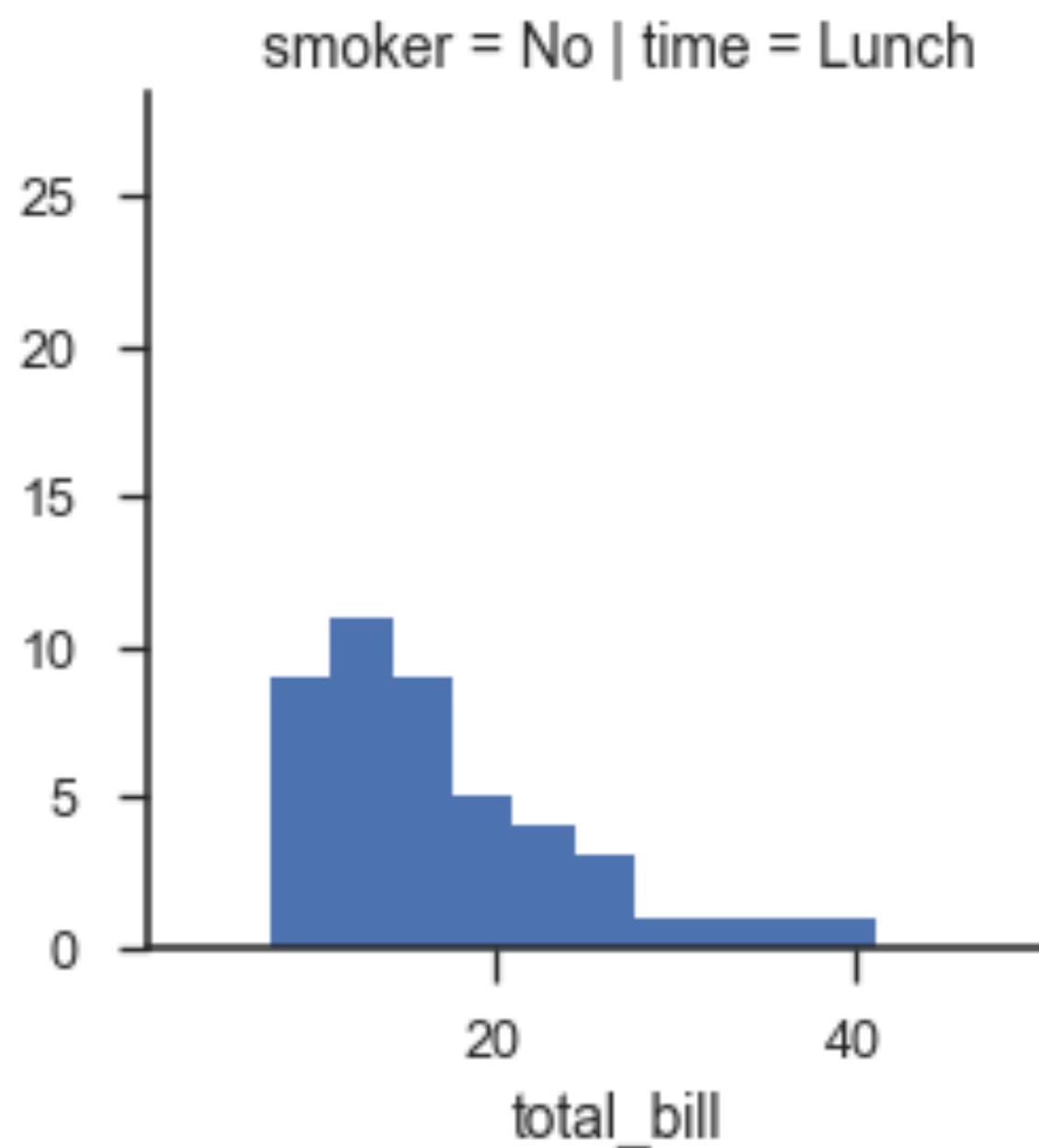
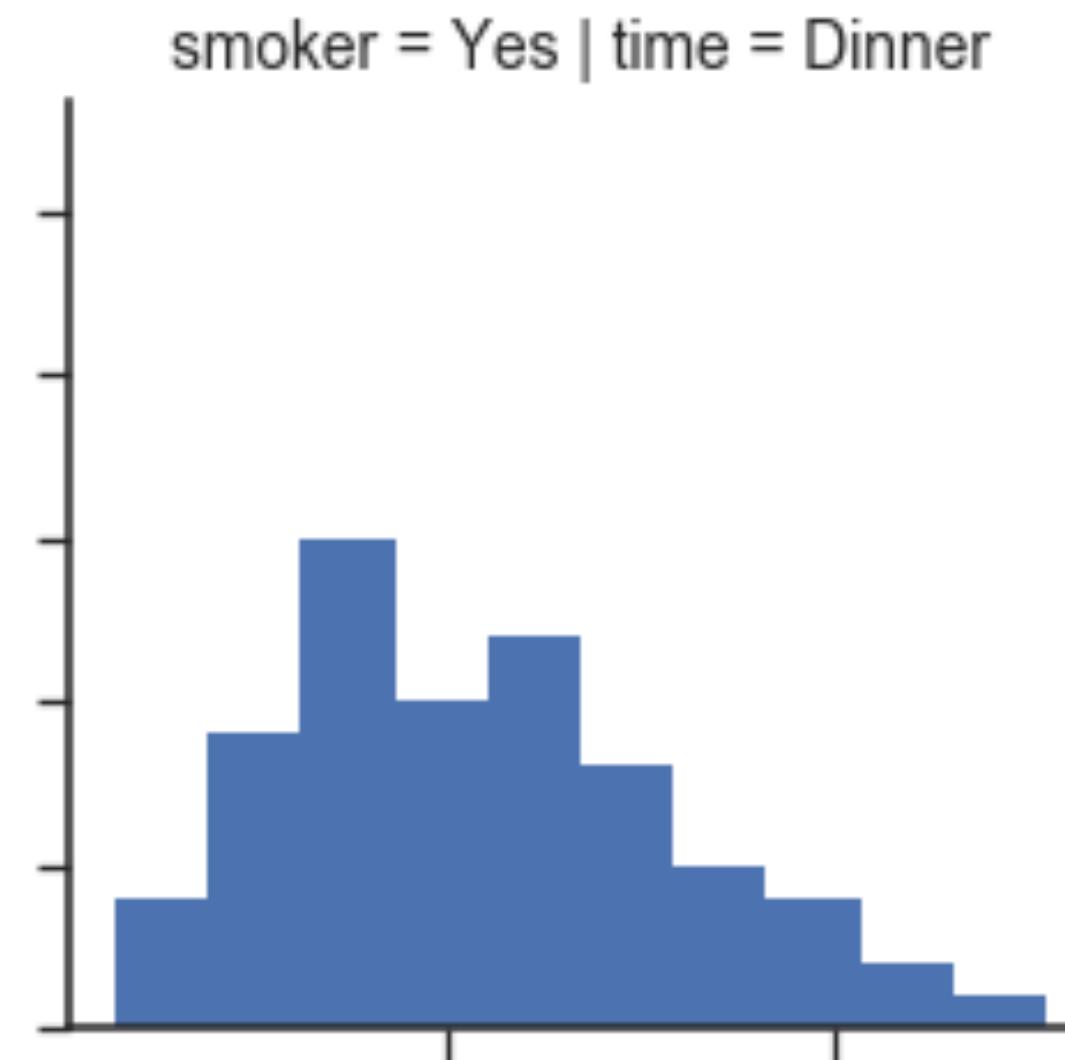
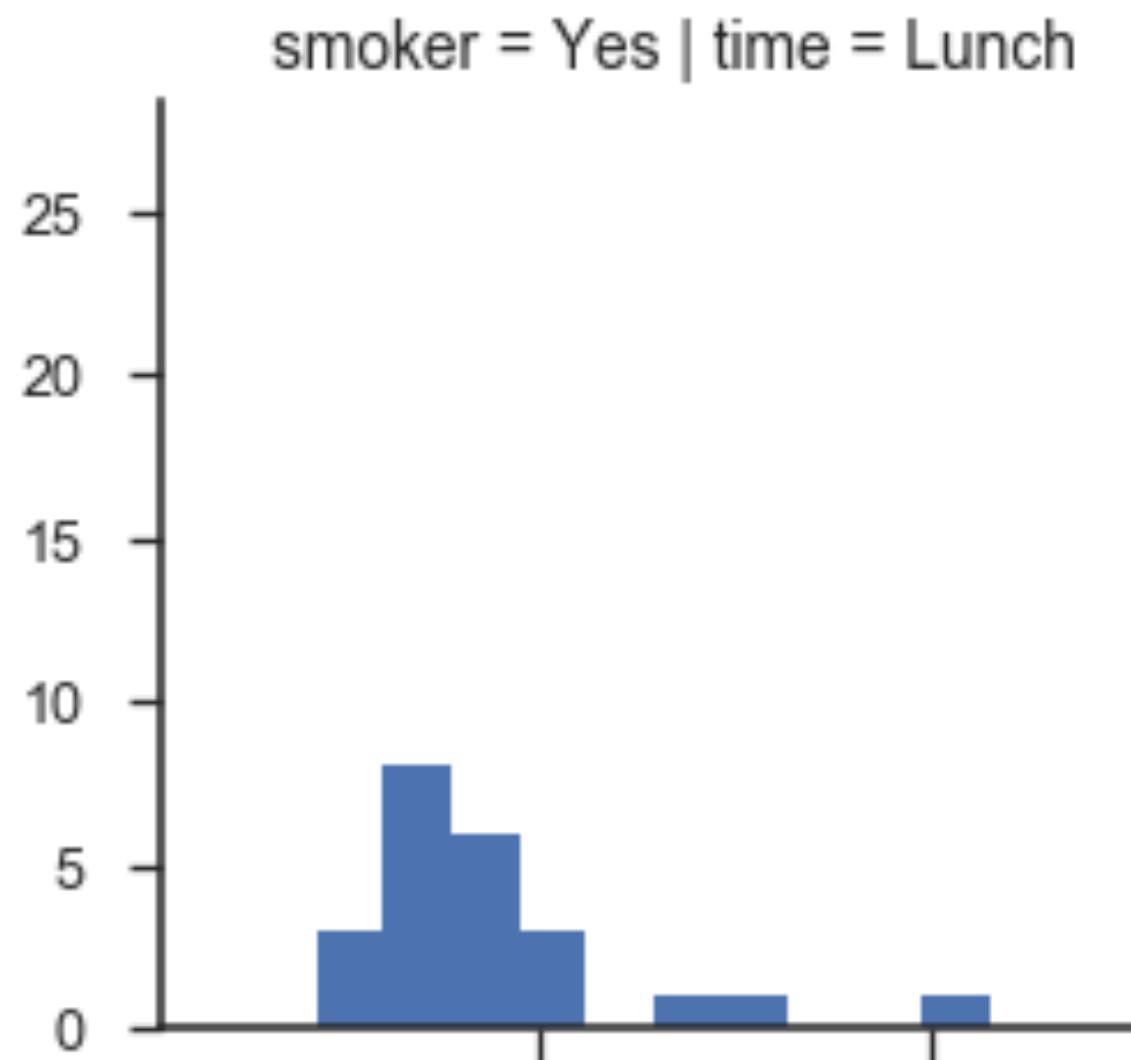


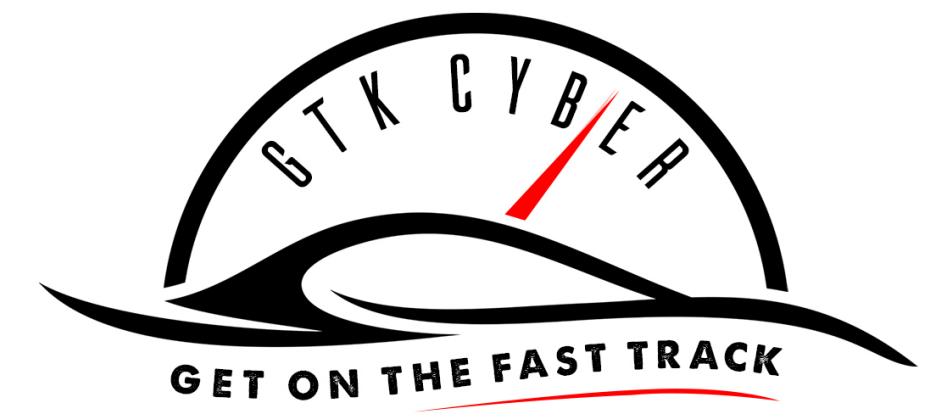
```
g = sns.lmplot(x="total_bill",  
y="tip", hue="smoker", data=tips)
```



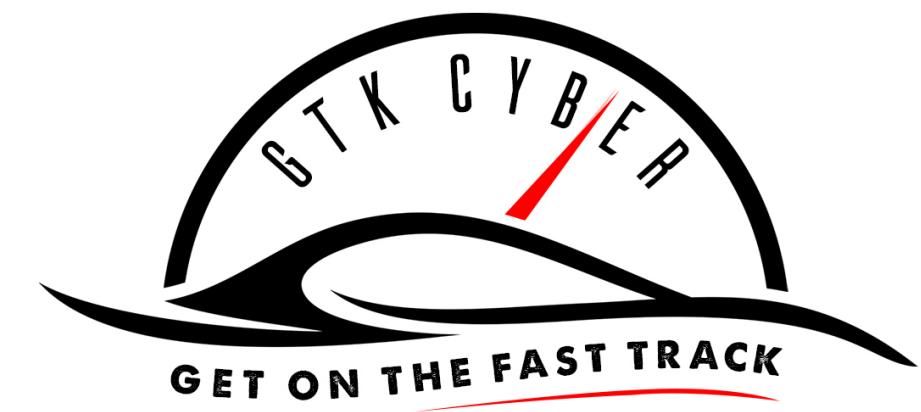


```
g = sns.FacetGrid(tips, col="time", row="smoker")
>>> g = g.map(plt.hist, "total_bill")
```



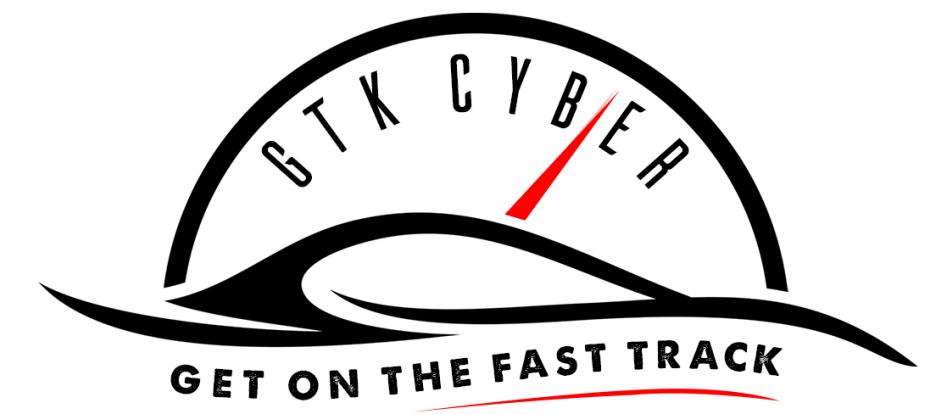


```
scatterPlot.get_figure().savefig( "scatterPlot.png" )
```

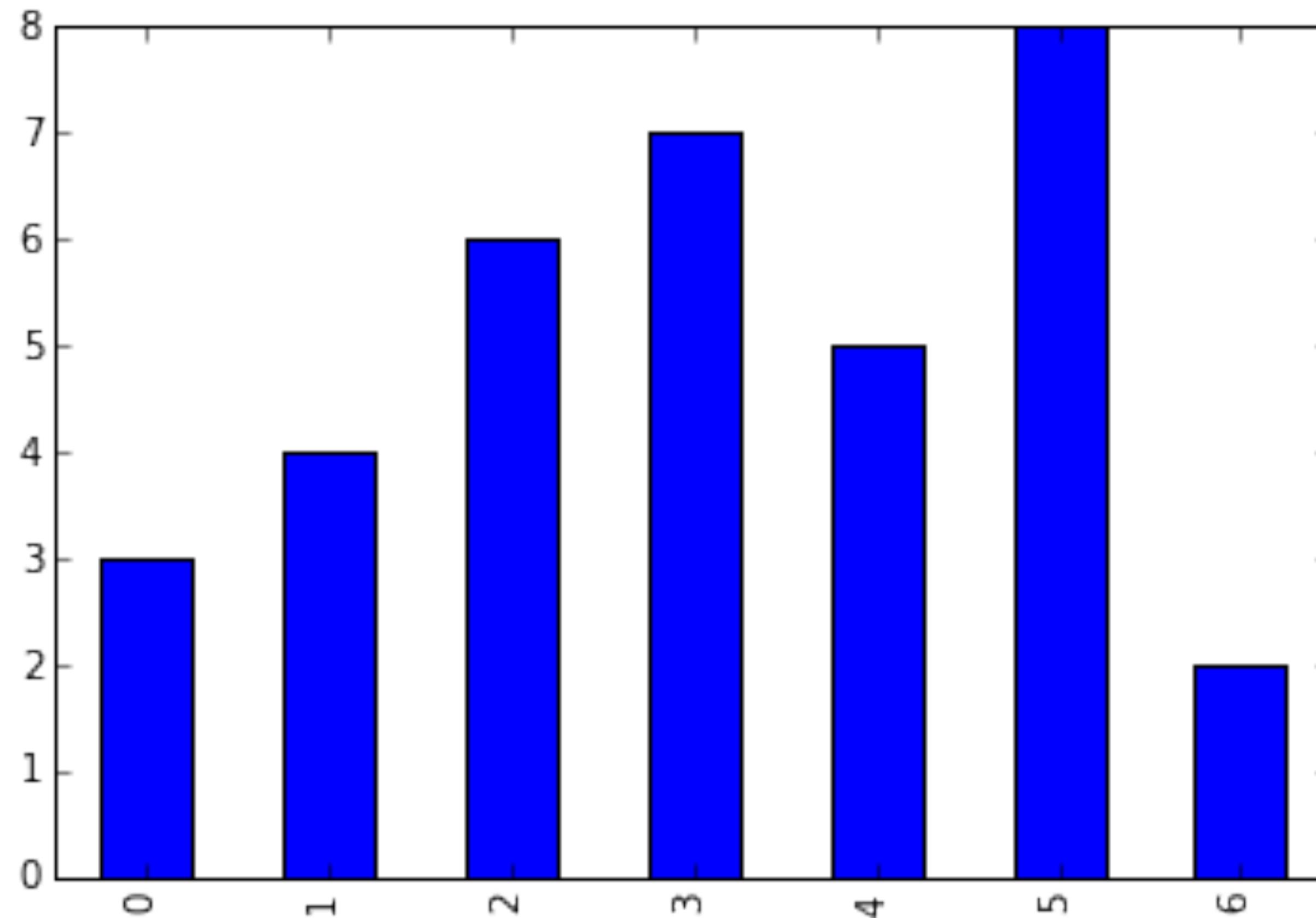
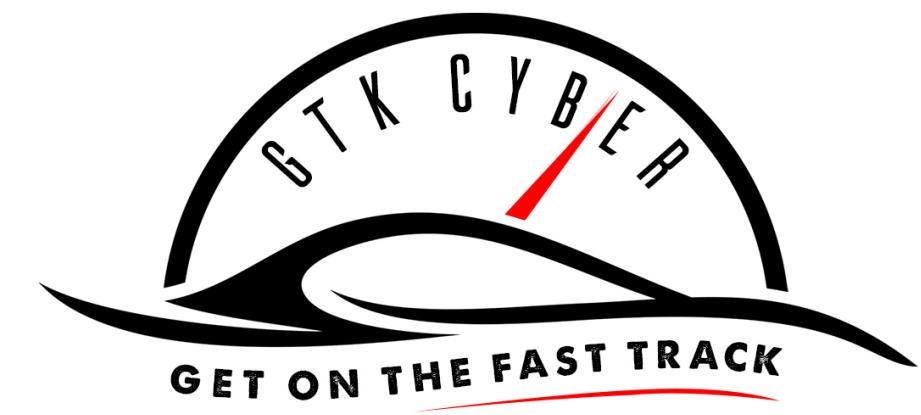


```
print(plt.style.available)

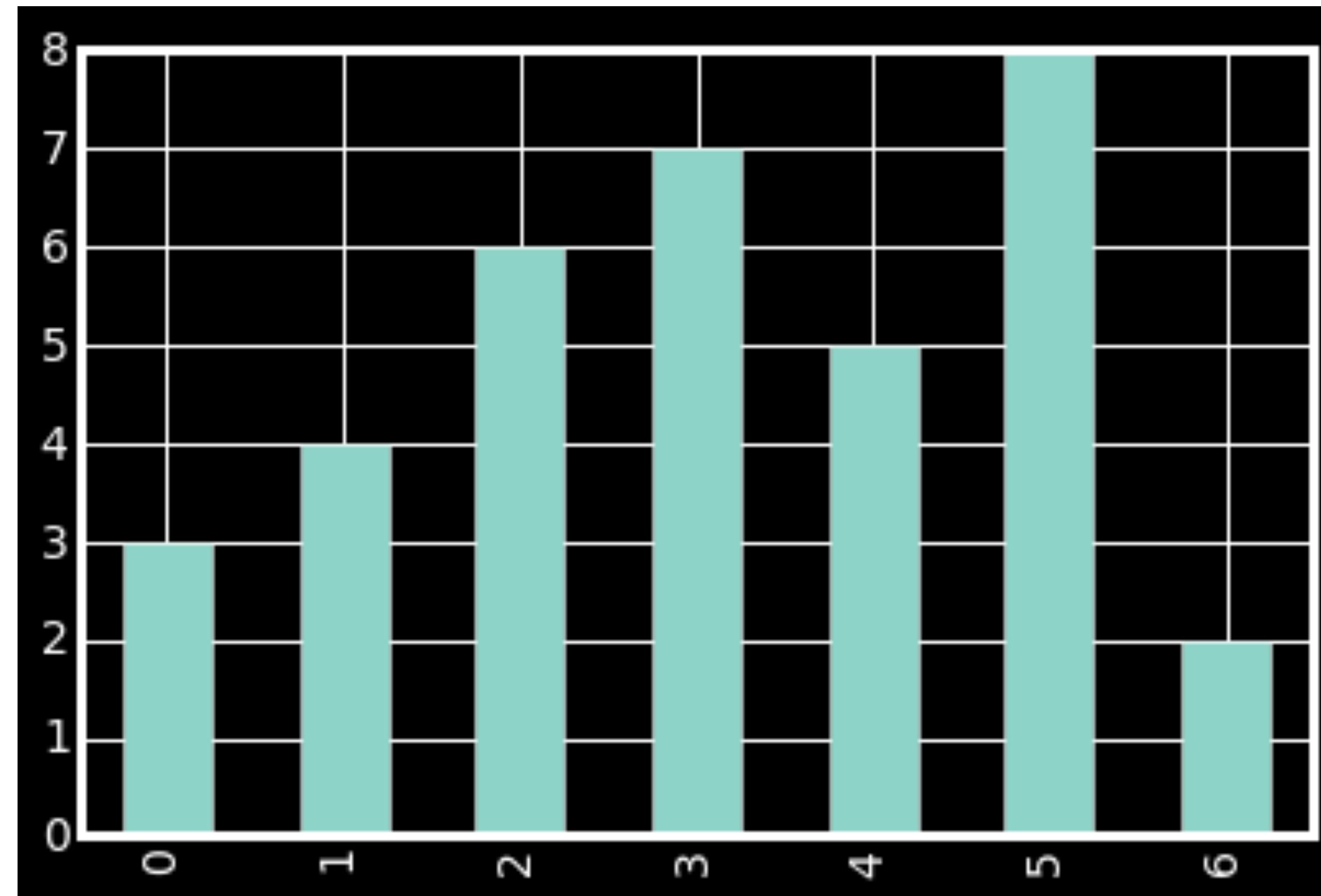
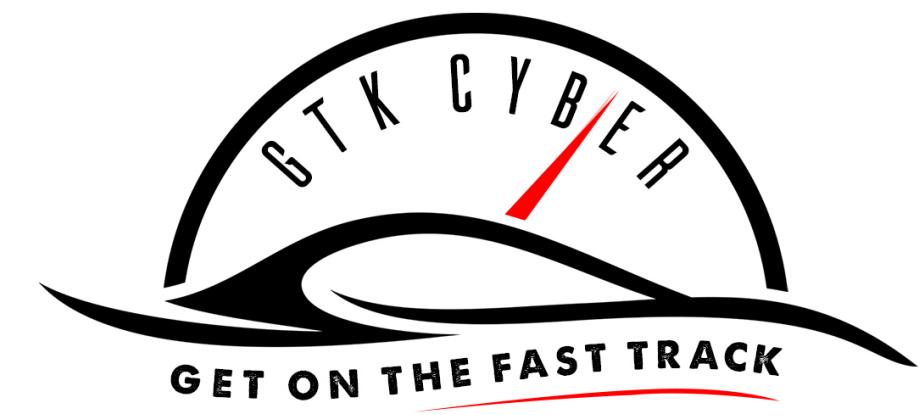
['dark_background', 'grayscale', 'ggplot',
 'bmh', 'fivethirtyeight']
```



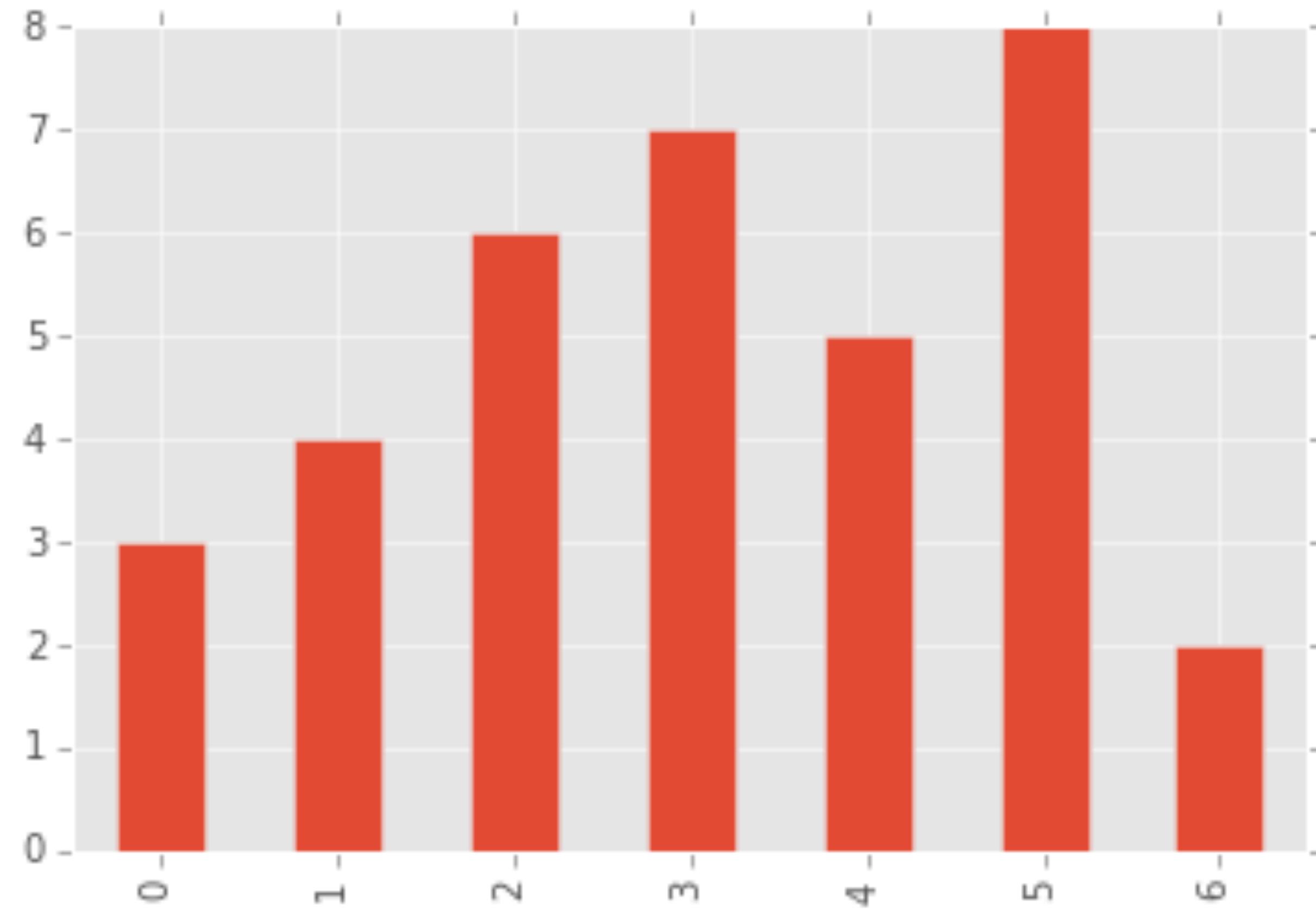
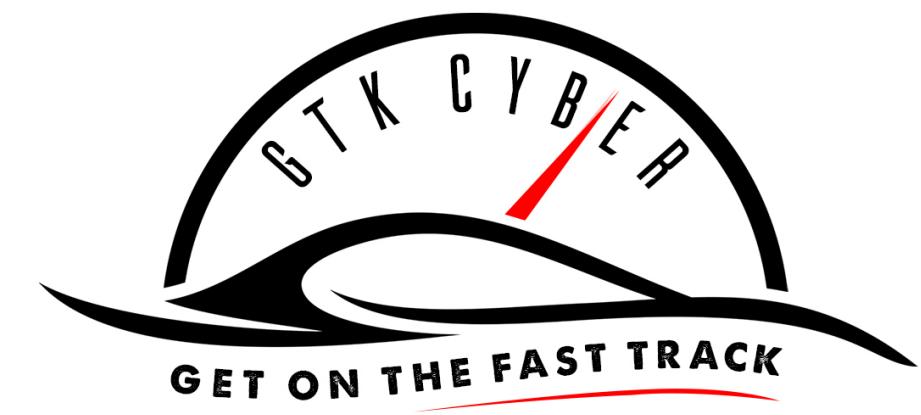
UGLY



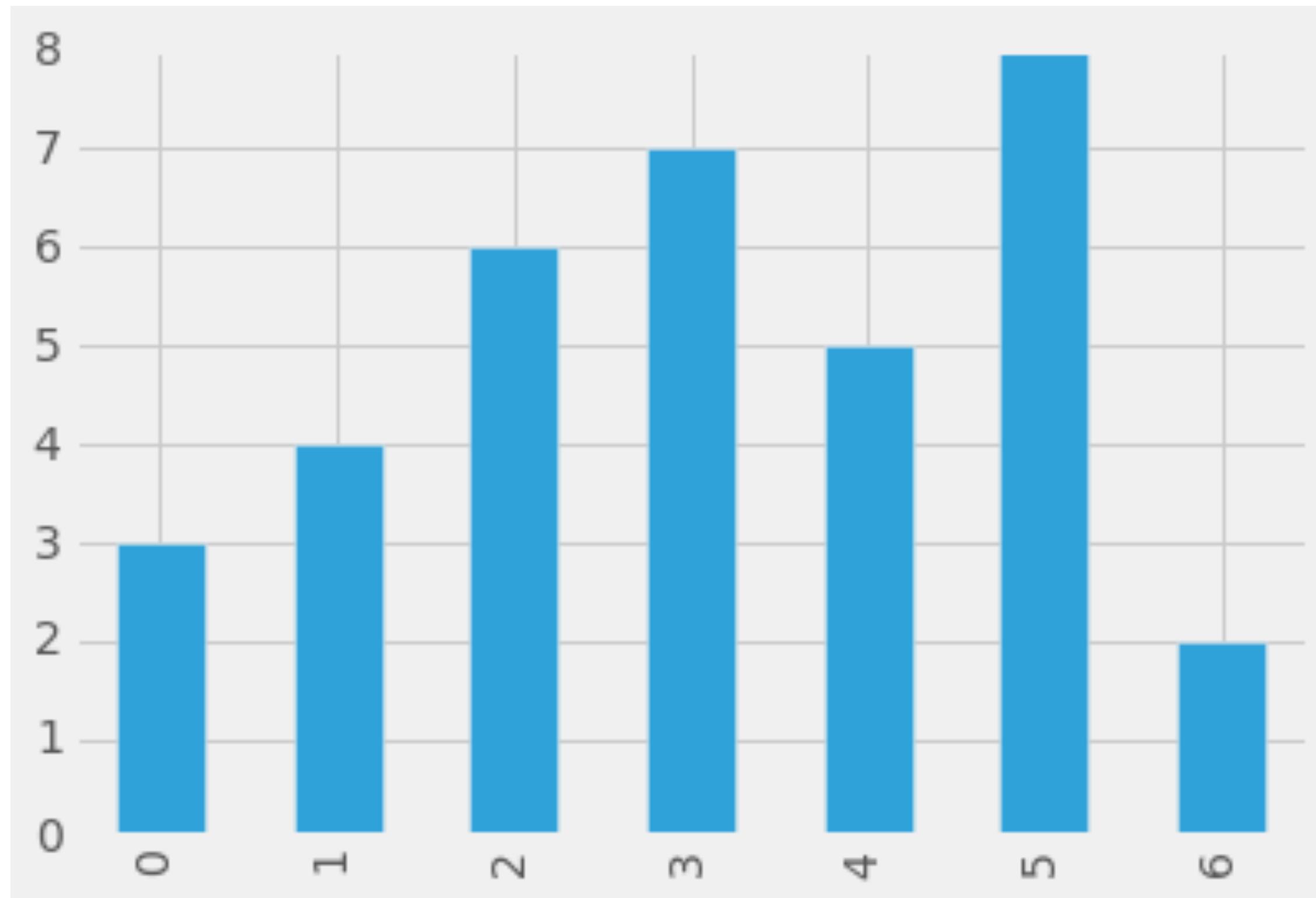
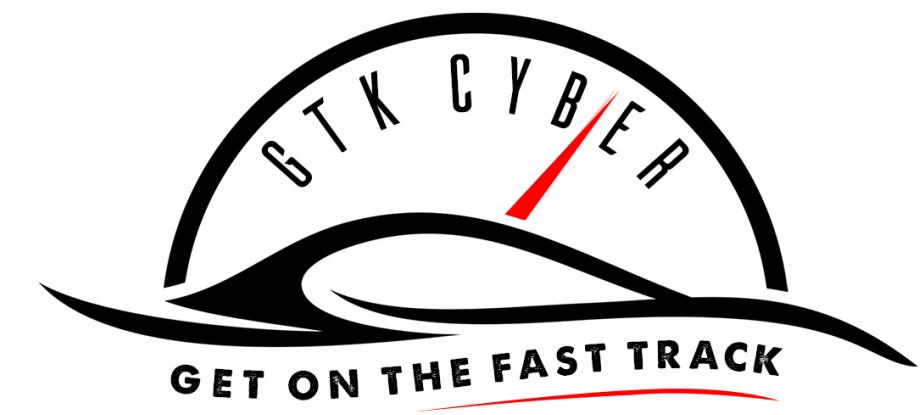
```
data = pd.Series( [3,4,6,7,5,8,2] )
barchart = data.plot( kind="bar" )
```



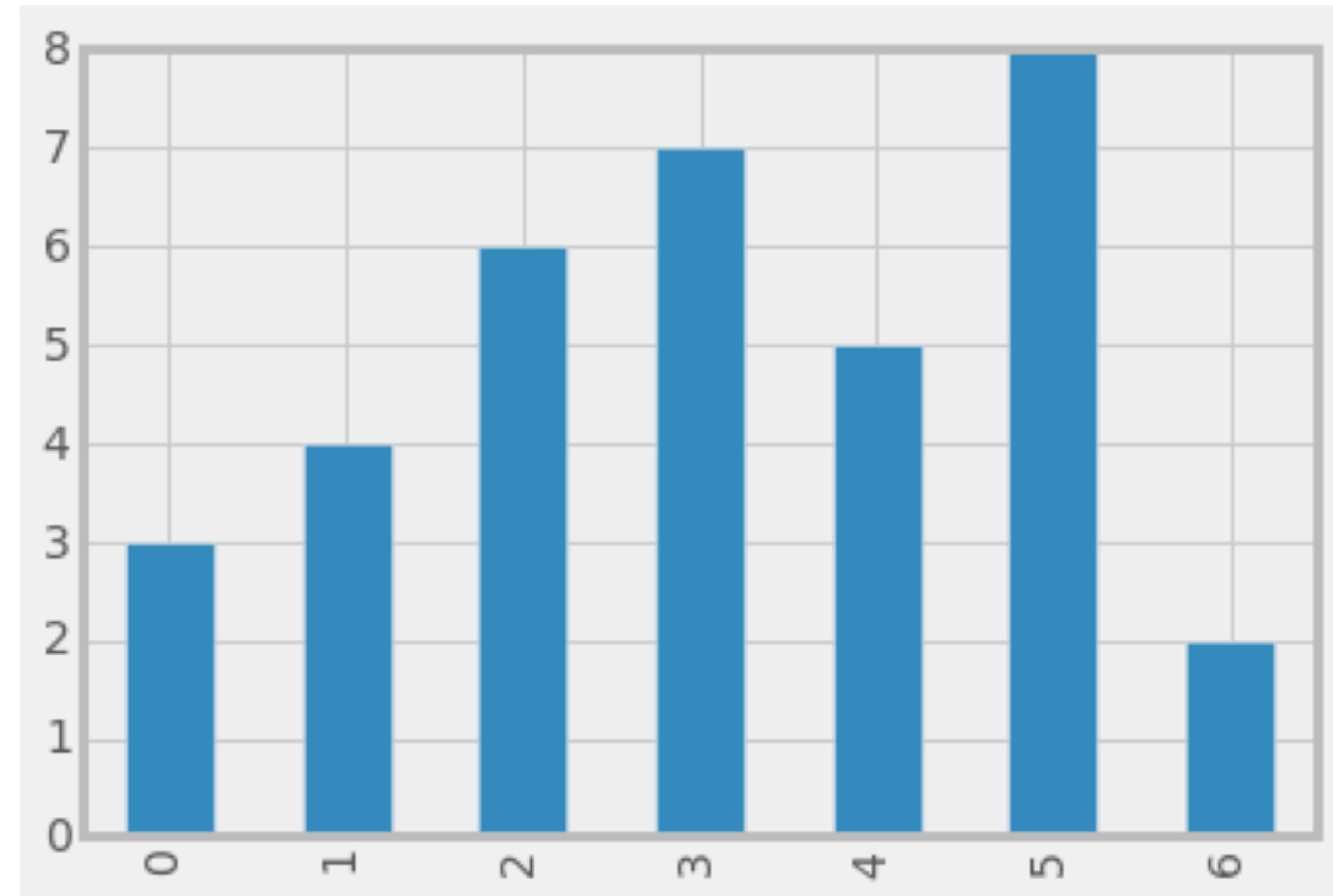
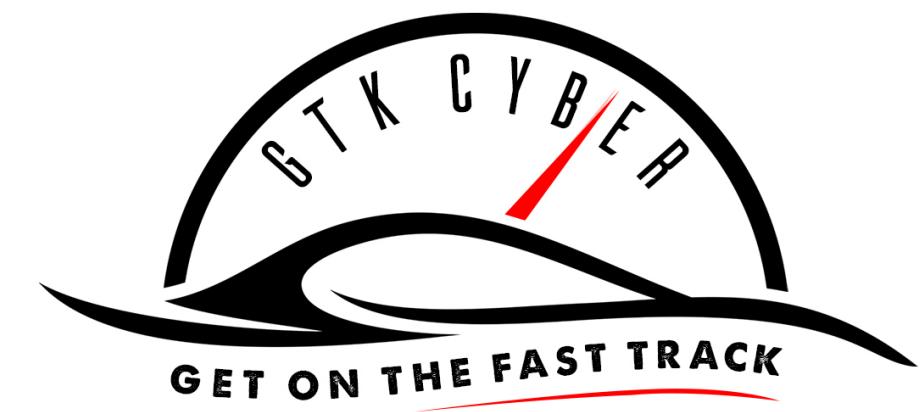
```
plt.style.use('dark_background')
barchart = data.plot( kind="bar" )
```



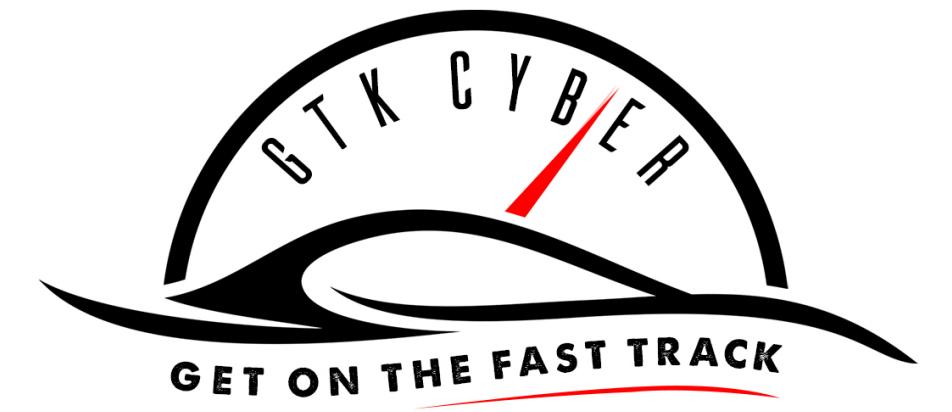
```
plt.style.use('ggplot')
barchart = data.plot( kind="bar" )
```



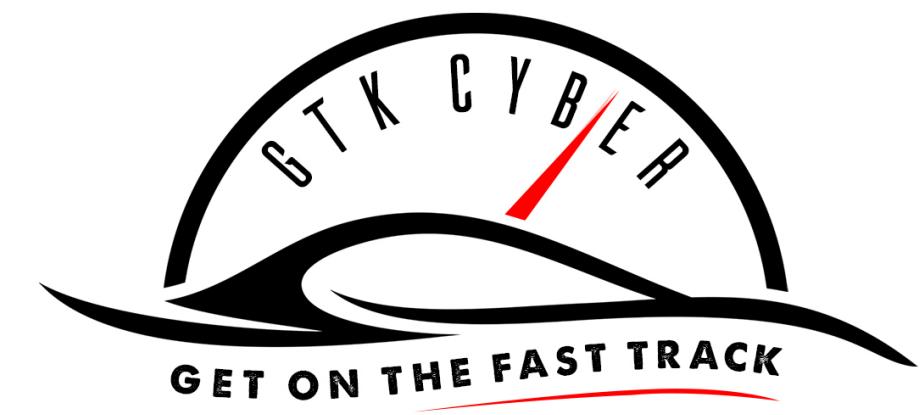
```
plt.style.use('fivethirtyeight')
barchart = data.plot( kind="bar" )
```



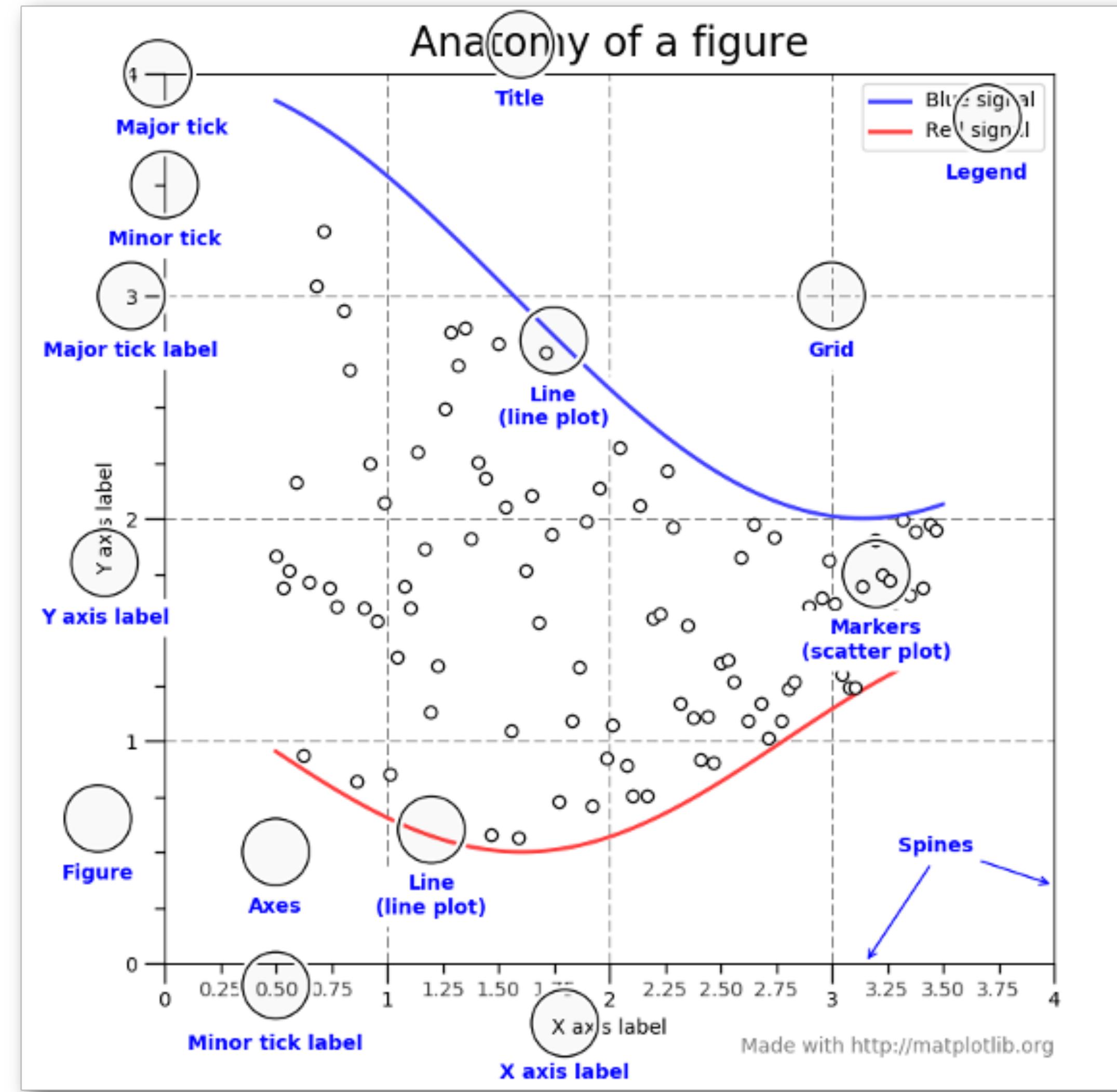
```
plt.style.use('bmh')
barchart = data.plot( kind="bar" )
```

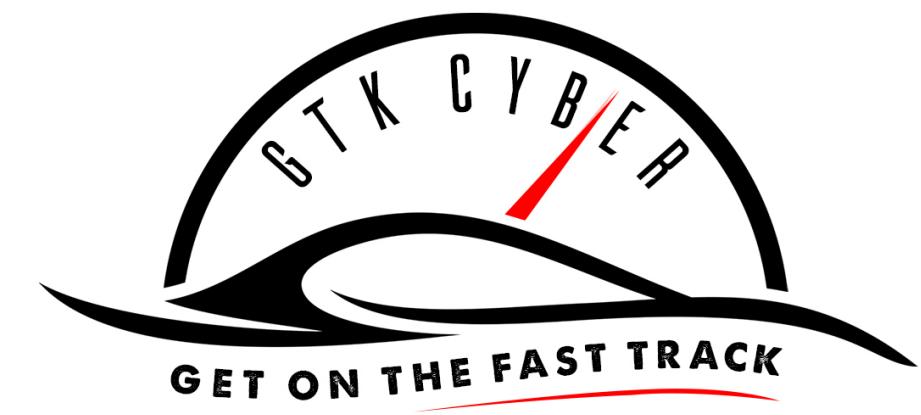


Customizing your Charts



Customizing your Charts





Customizing your Charts

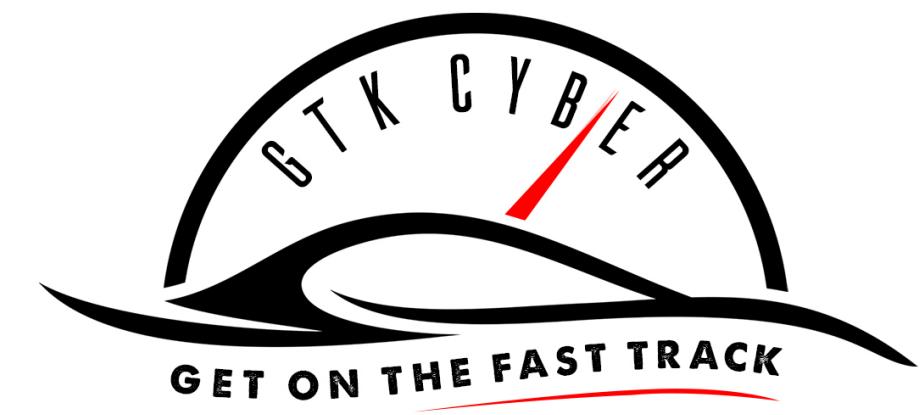
- **Axes:** Axes represent an individual plot or chart.
- **Figure:** This is the final complete image and may contain 1 or more axes.



Customizing your Charts



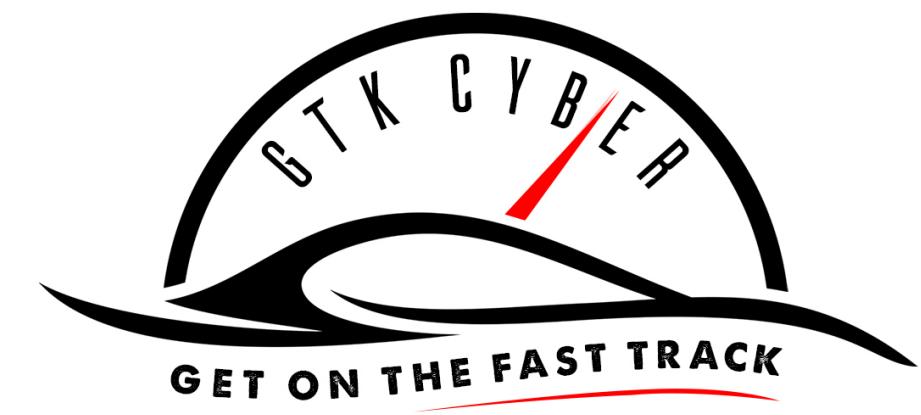
- Matplotlib has both an object-oriented and state-based interface. This can be confusing when you look up answers on Stack Overflow.
- Use basic pandas plotting for simple plots
- Seaborn, YellowBrick and others for more complex visualizations
- Use the OO interface in MatPlotLib to customize simple visualizations.



Customizing your Charts

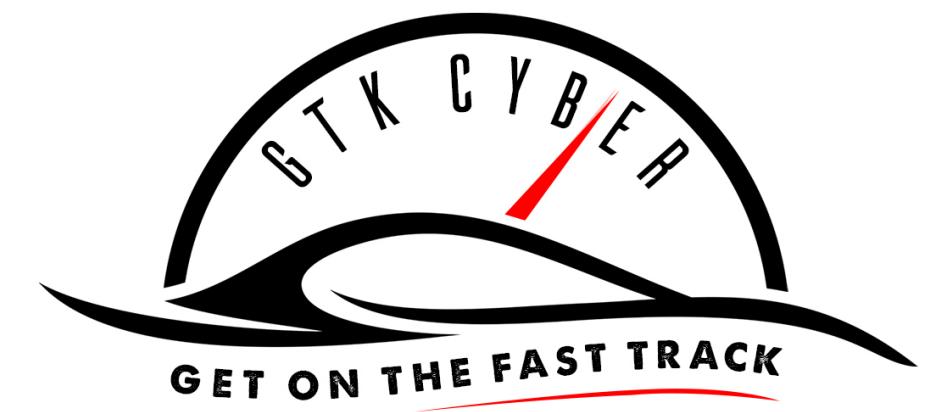
```
fig, ax = plt.subplots()  
df.plot(..., ax=ax)
```

This method gets you access to the **figure** and **axes**.



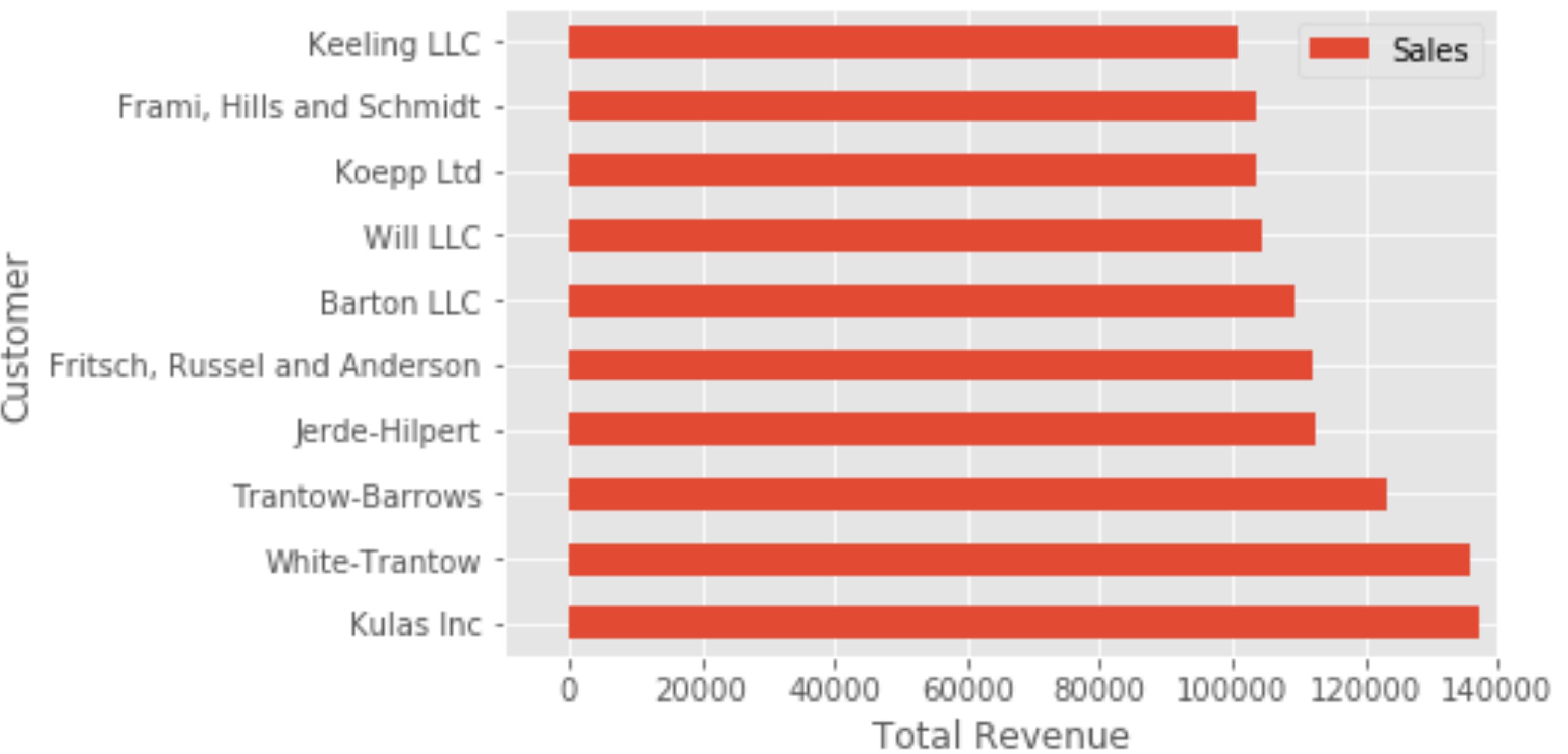
Customizing your Charts

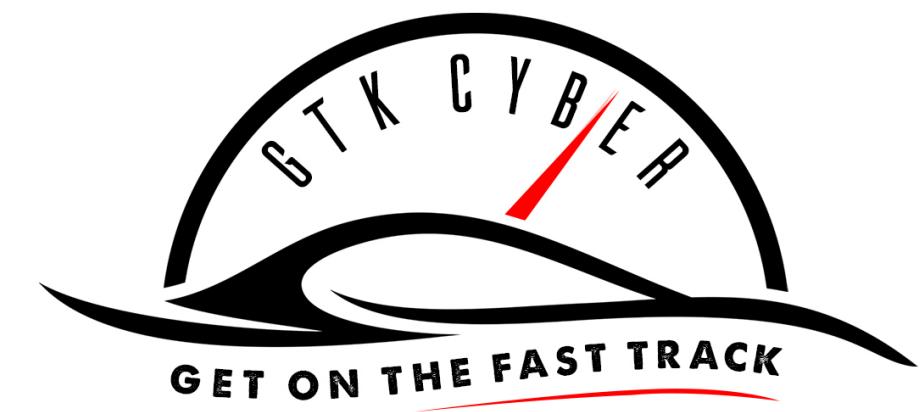
- **Axes:** Long list of possible options to configure your chart: [**https://matplotlib.org/3.1.0/api/axes_api.html#the-axes-class**](https://matplotlib.org/3.1.0/api/axes_api.html#the-axes-class)



Add Axis Labels

```
fig, ax = plt.subplots()  
df.plot(kind='barh', ax=ax)  
ax.set_xlim([-10000, 140000])  
ax.set_xlabel('Total Revenue')  
ax.set_ylabel('Customer');
```



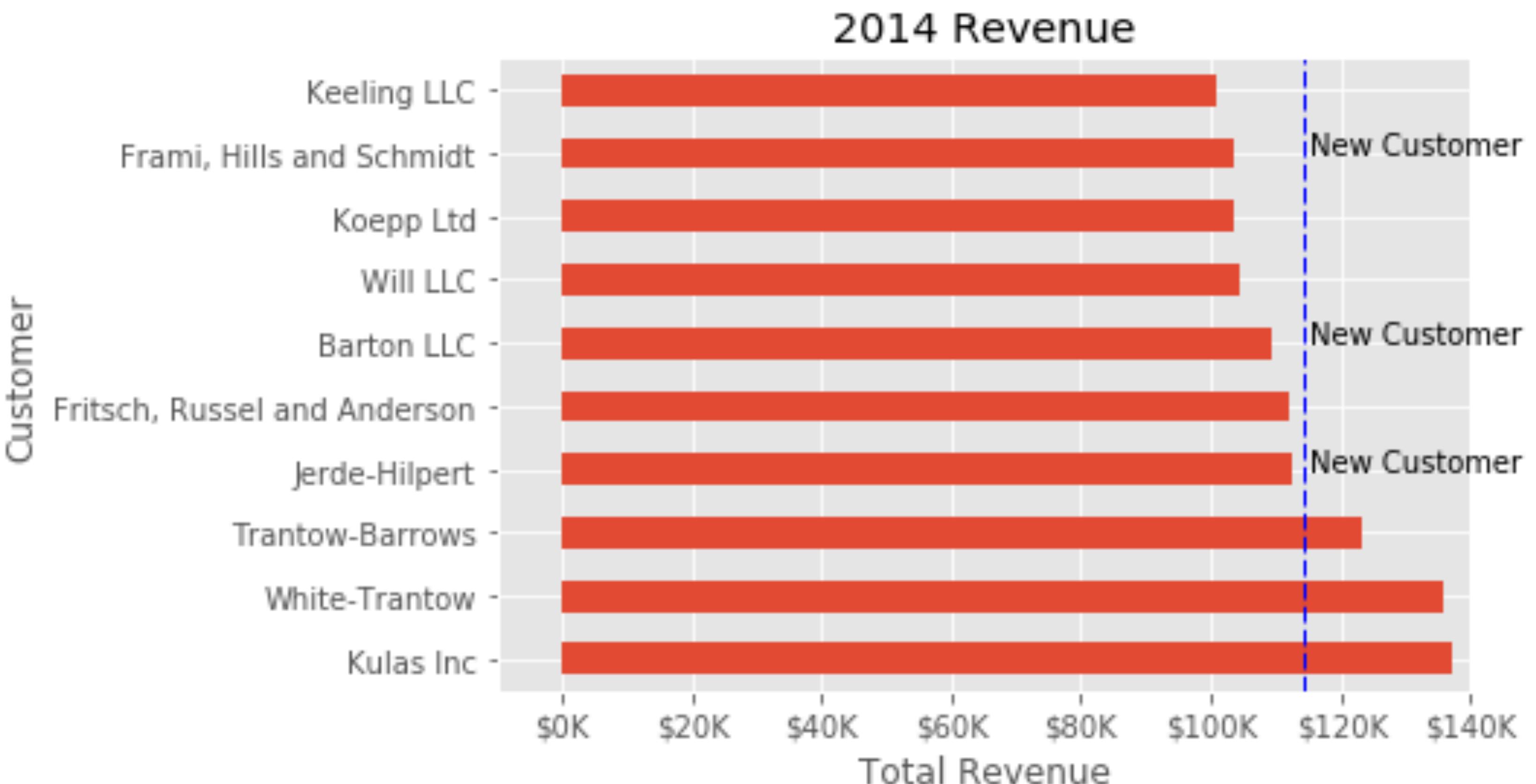


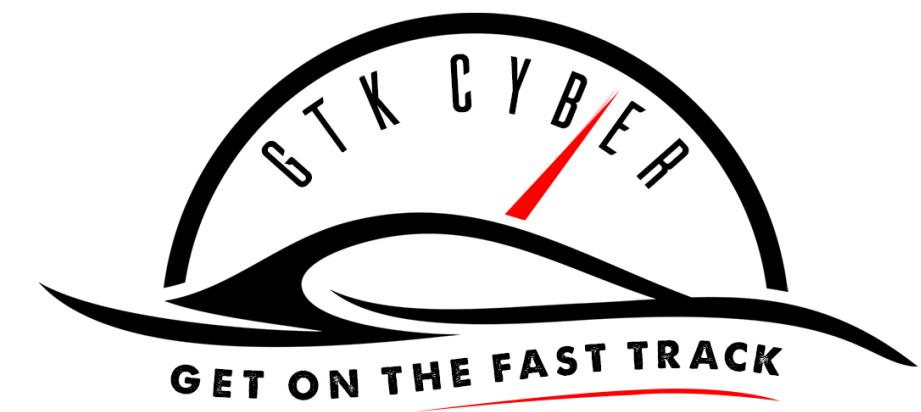
Add Annotations

```
fig, ax = plt.subplots()
df.plot(kind='barh', ax=ax)
avg = df['Sales'].mean()

ax.axvline(x=avg,
            color='b',
            label='Average',
            linestyle='--',
            linewidth=1)

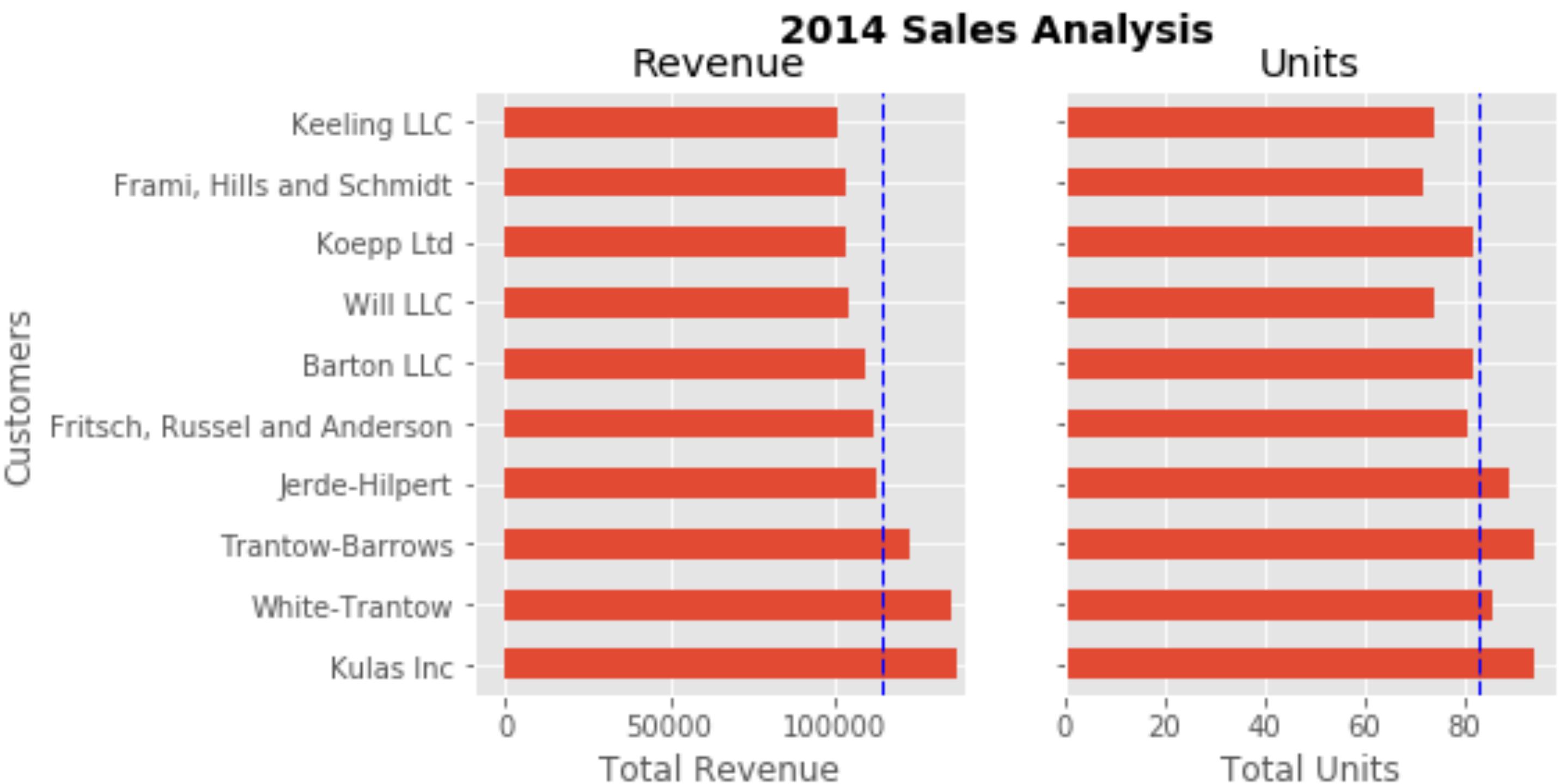
for cust in [3, 5, 8]:
    ax.text(115000, cust,
            "New Customer")
```

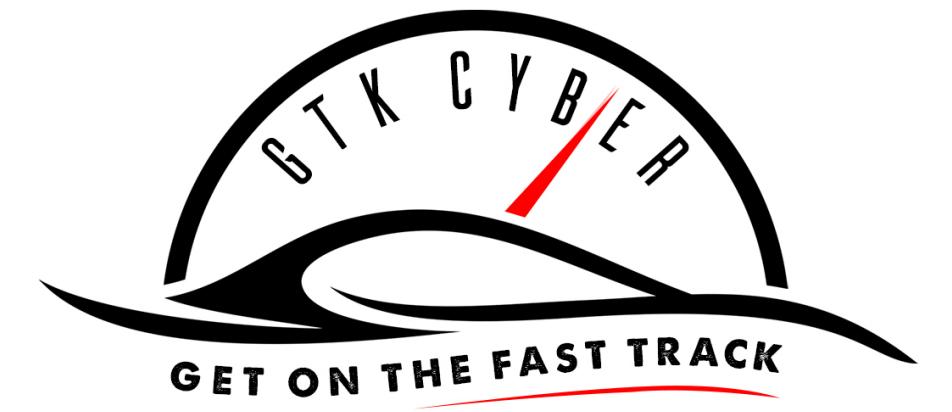




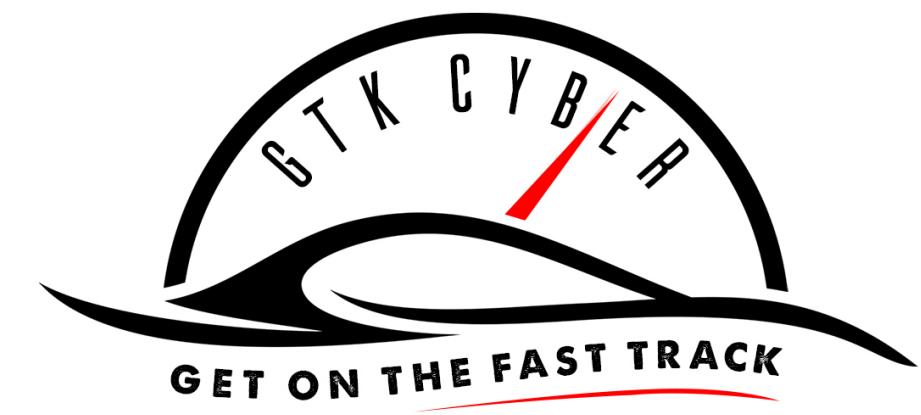
Subplots

```
fig, (ax0, ax1) = plt.subplots(nrows=1,
                               ncols=2,
                               sharey=True,
                               figsize=(7, 4))
df.plot(kind='barh', ax=ax0)
ax0.set_xlim([-100000, 140000])
...
df.plot(kind='barh', ax=ax1)
avg = df['Purchases'].mean()
ax1.set(title='Units',
        xlabel='Total Units',
        ylabel='')
ax1.axvline(x=avg, color='b',
             label='Average',
             linestyle='--',
             linewidth=1)
```



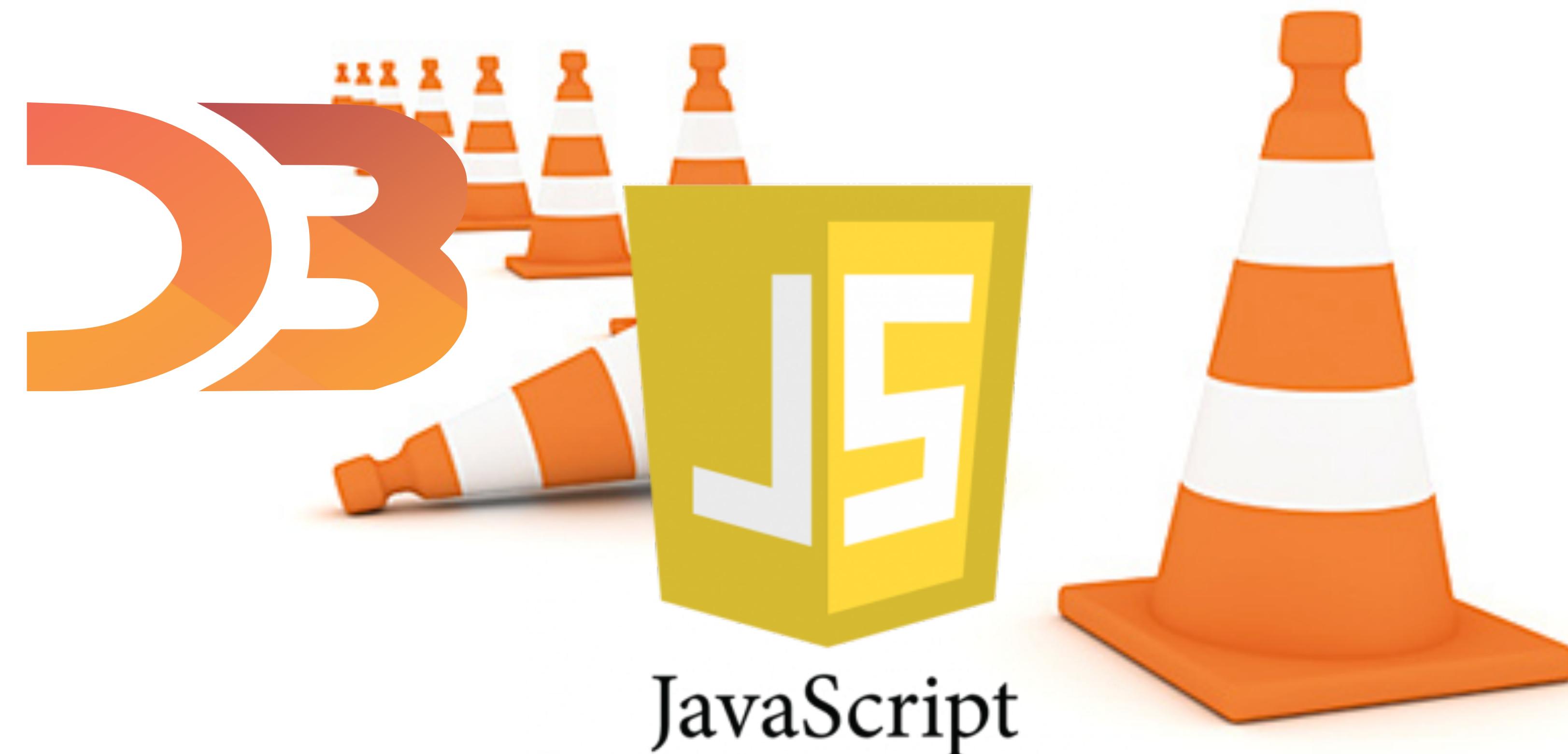


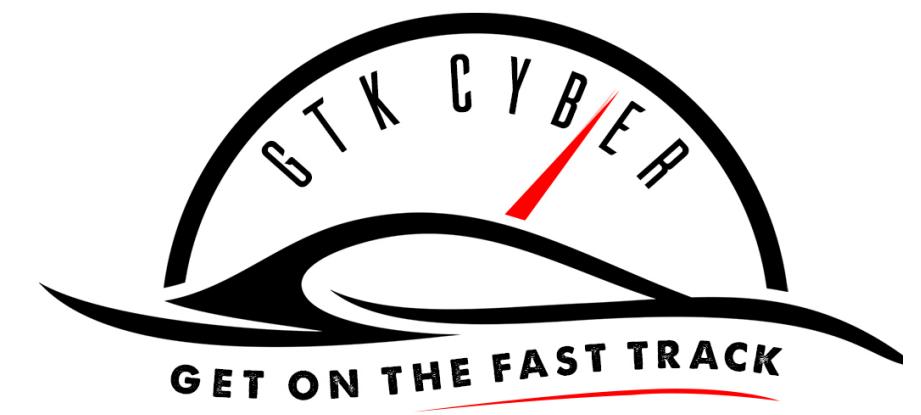
Interactive Visualizations



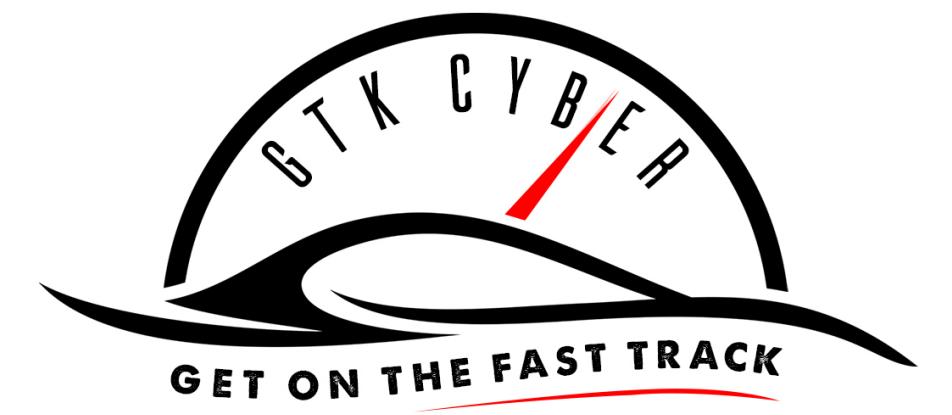
QlikView







Easy to use... if you know R

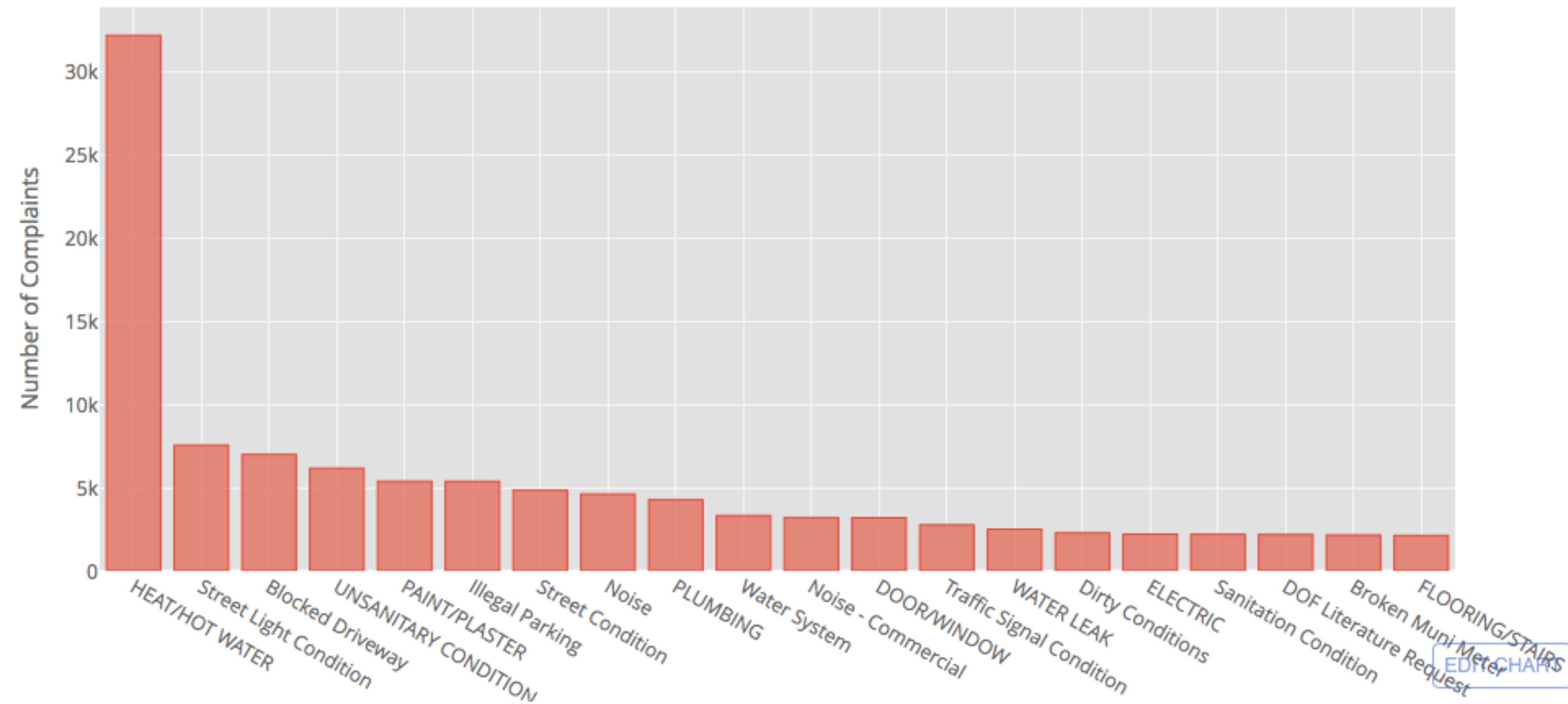


Introducing Cufflinks

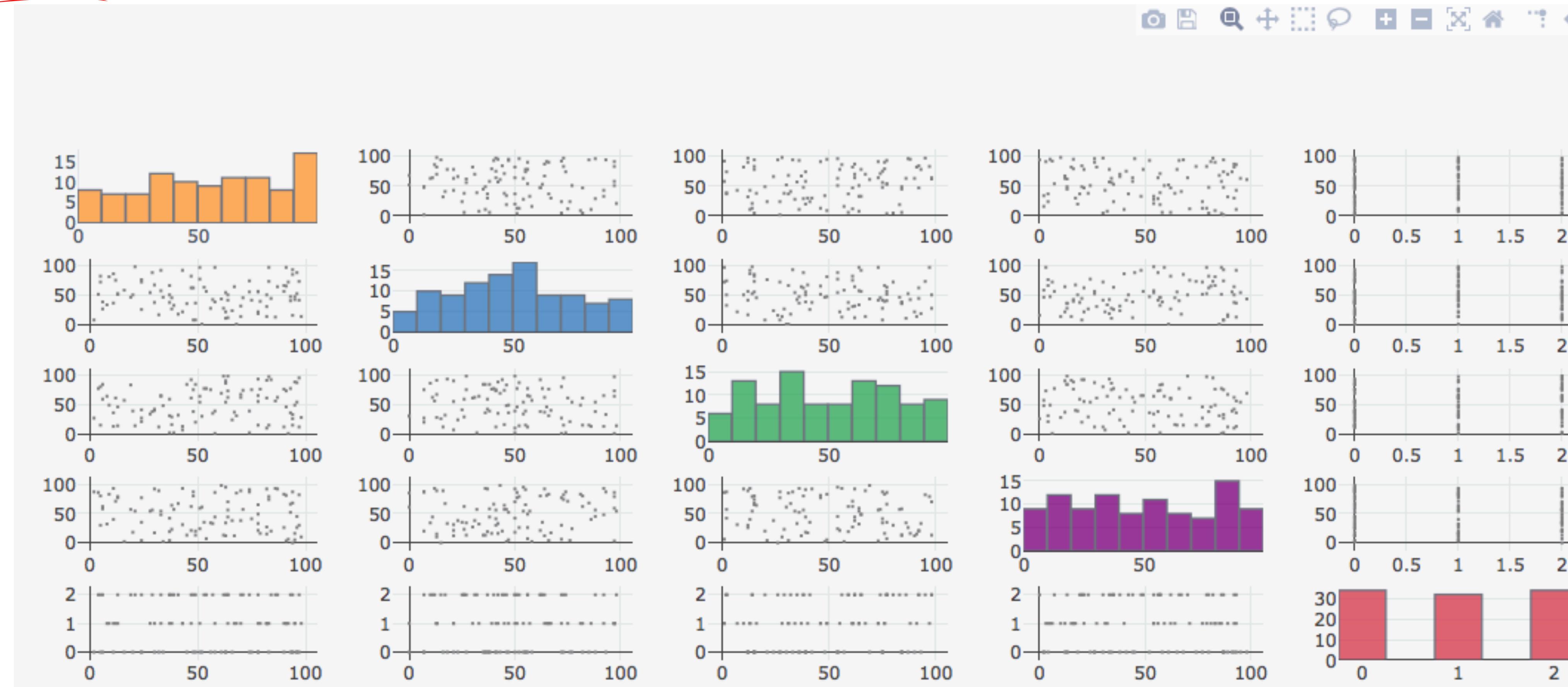




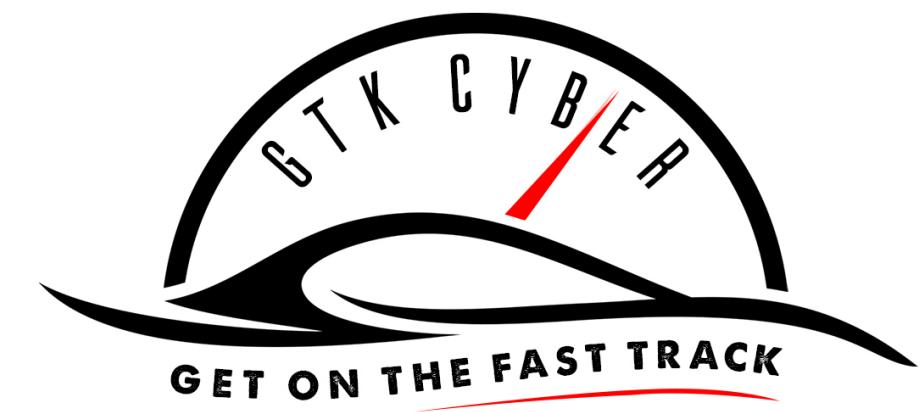
NYC 311 Complaints



```
series.iplot(kind='bar', yTitle='Number of Complaints', title='NYC 311 Complaints')
```



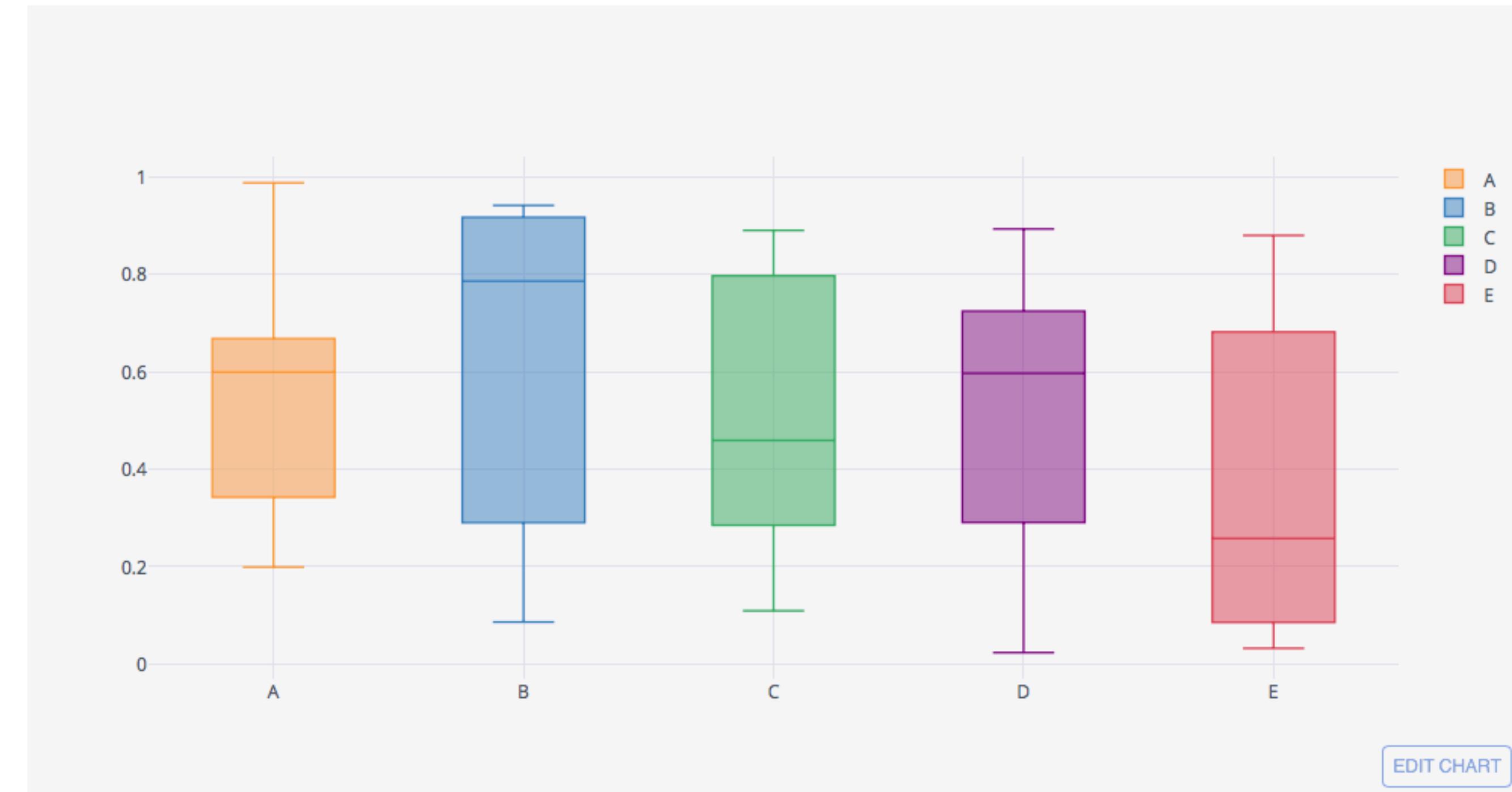
`df.scatter_matrix()`

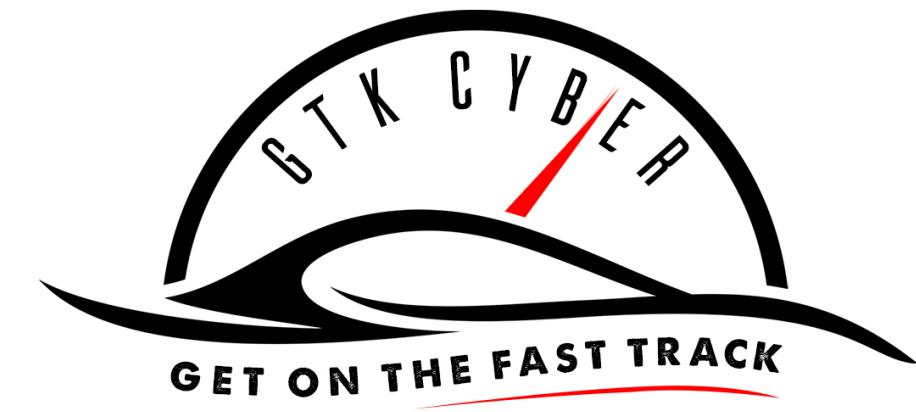


Supported Chart Types

`df.iplot(kind='<type>')`

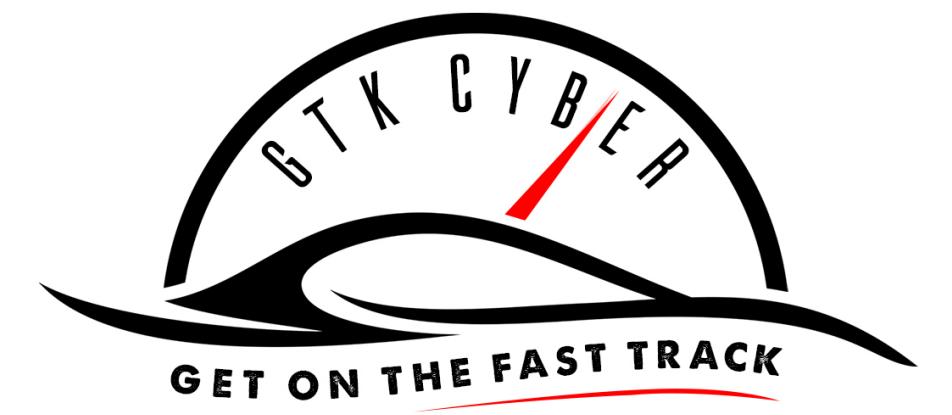
- line
- bar
- histogram
- box (boxplot)
- area
- scatter
- bubble
- heat map





In Class Exercise

Please complete Worksheet 4: Data Visualization



Questions?