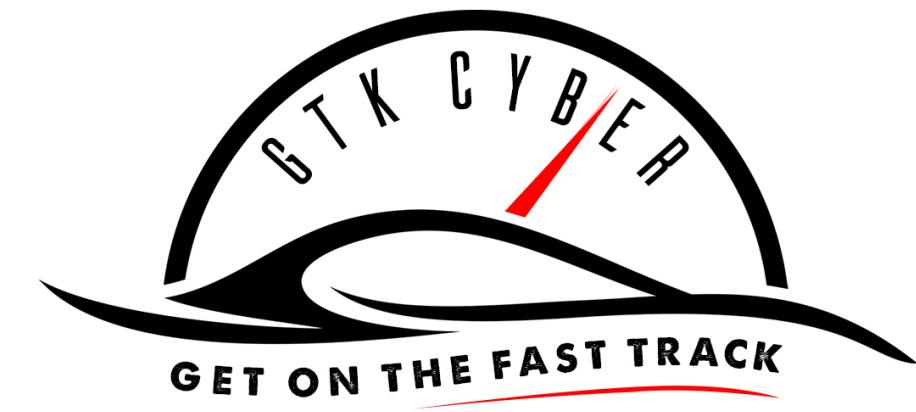


GTK CYBER

GET ON THE FAST TRACK

Module 4: Data Visualization

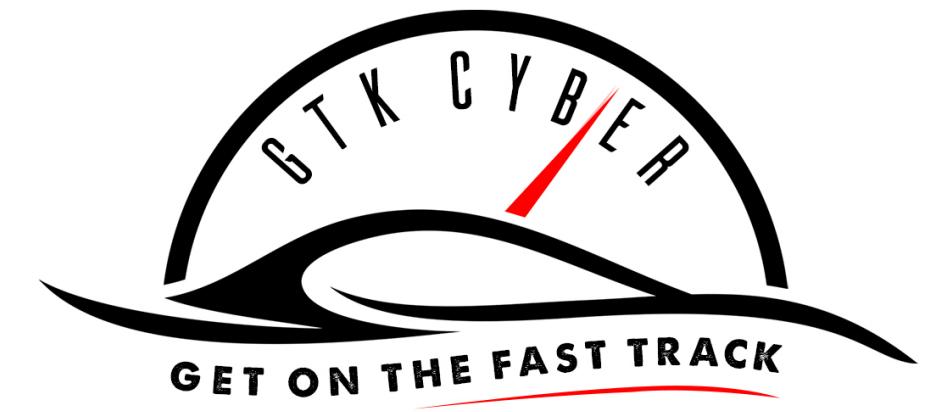
GET ON THE FAST TRACK



Agenda for the Course

Day 2

- Data Visualization
- Machine Learning: Supervised
 - Lab: URL Classification Challenge
- Machine Learning: Unsupervised Learning
- Introduction to Big Data Tools
 - Hadoop/Spark/Drill
- Hacking Machine Learning Models



Exploring Data Through Visualization

Exploratory Visualizations

Explanatory Visualizations

Why Visualize Data?

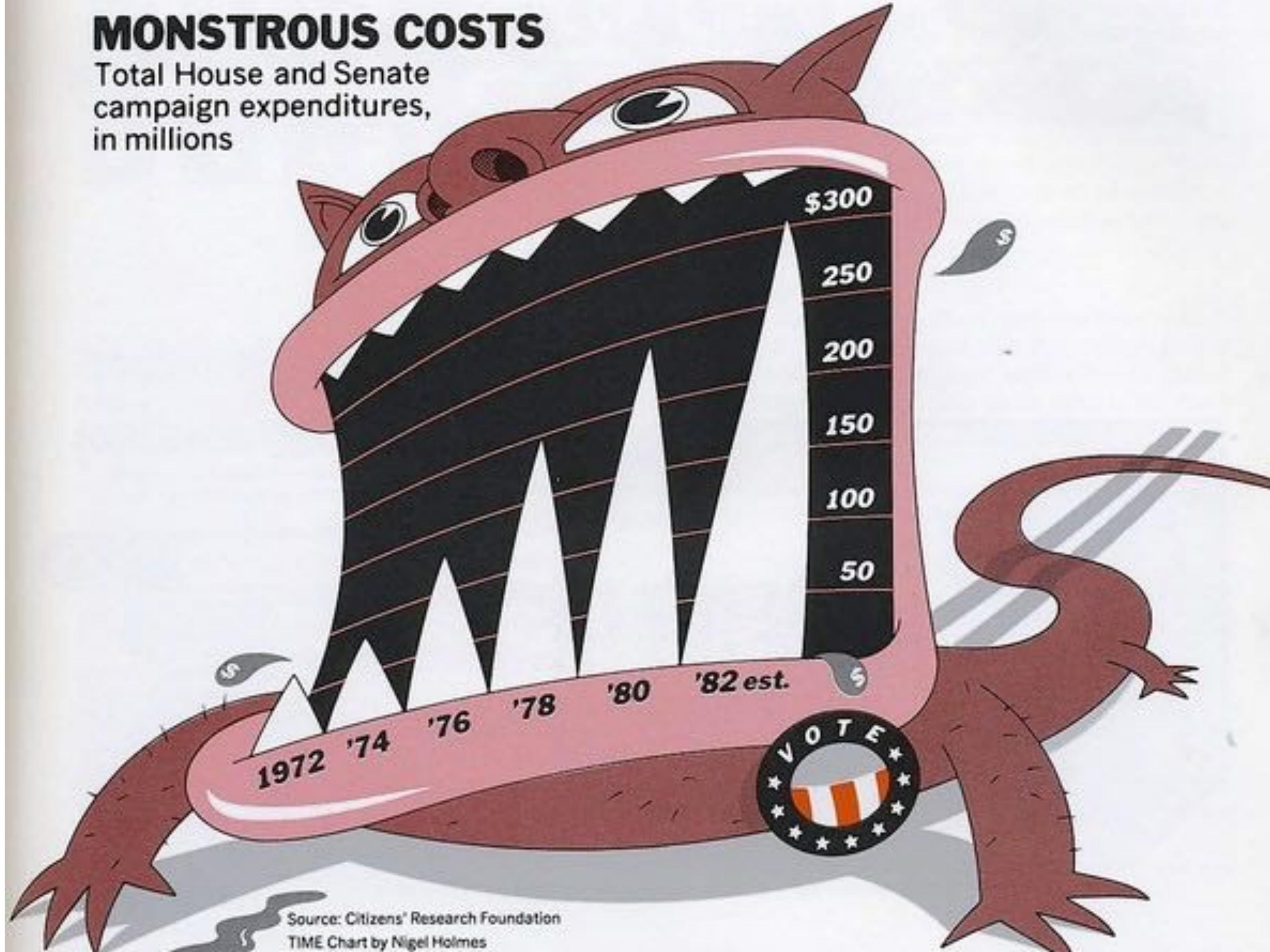
Visualizing data can inspire you to ask new and more refined questions of your data and ultimately lead to better analysis.

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

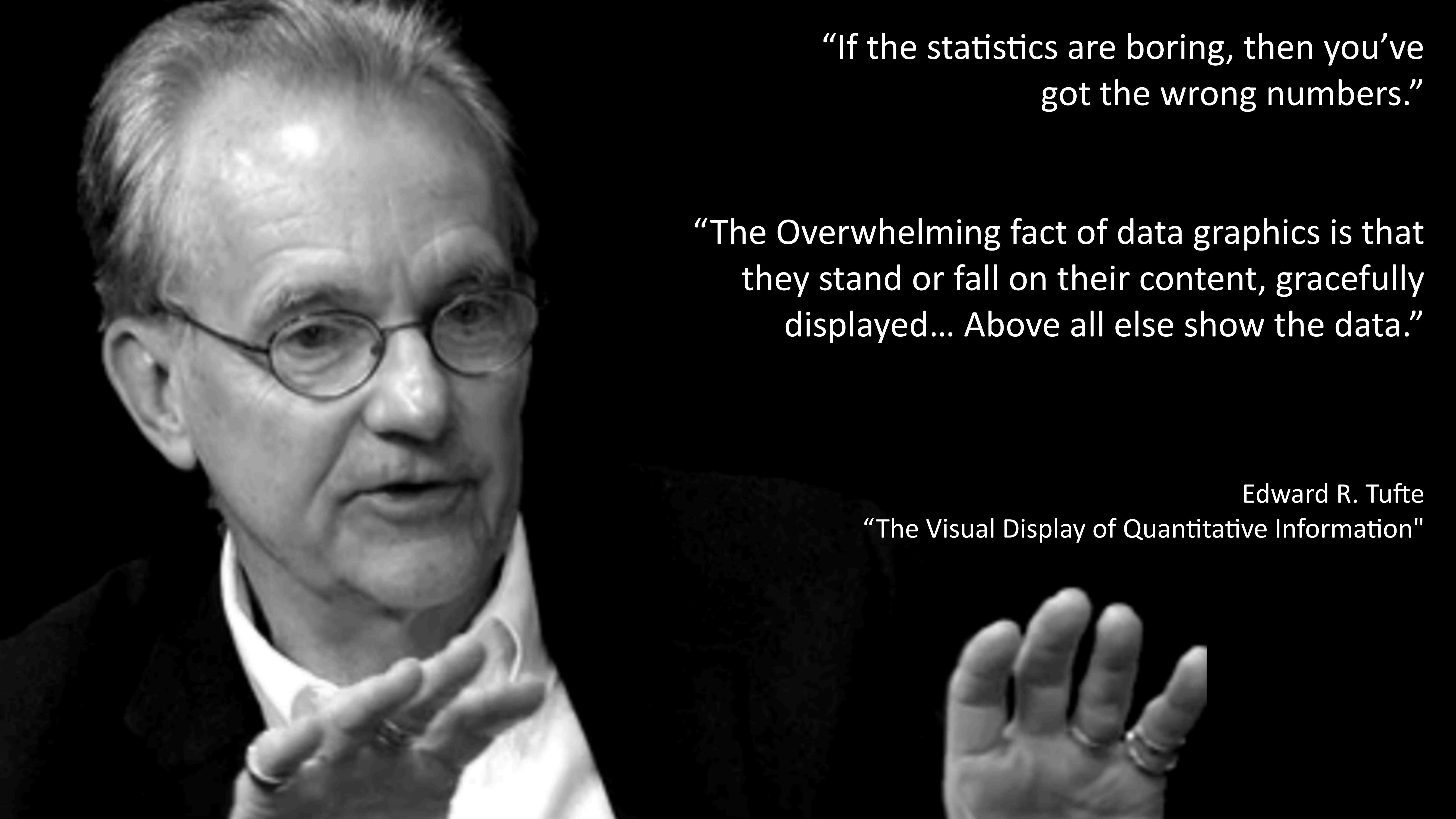
Remove
to improve
(the **data-ink** ratio)

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”

Show Comparisons,
Contrasts and Differences

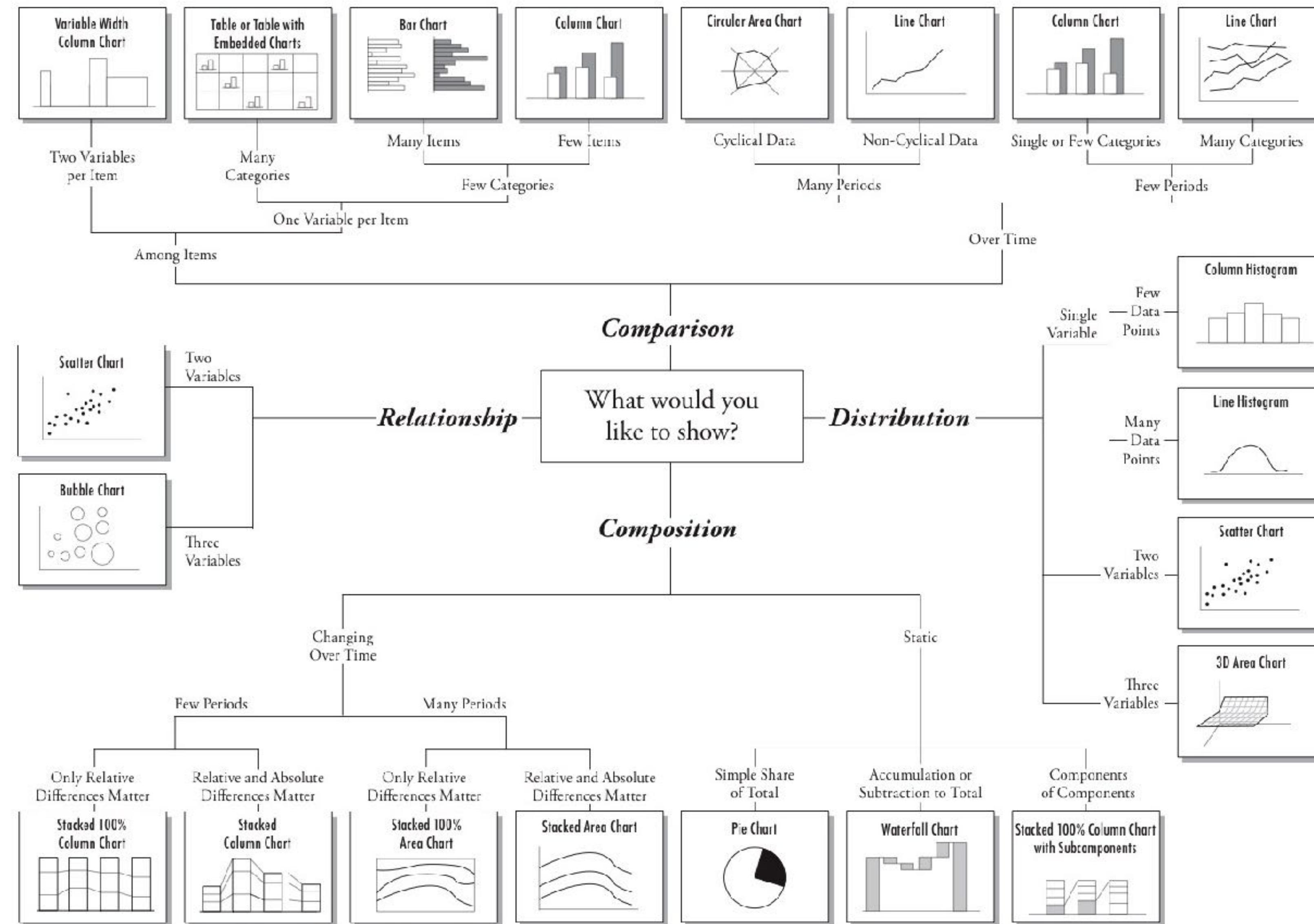
Show Multivariate Data

Integrate words, numbers,
images and diagrams

Document your
Evidence

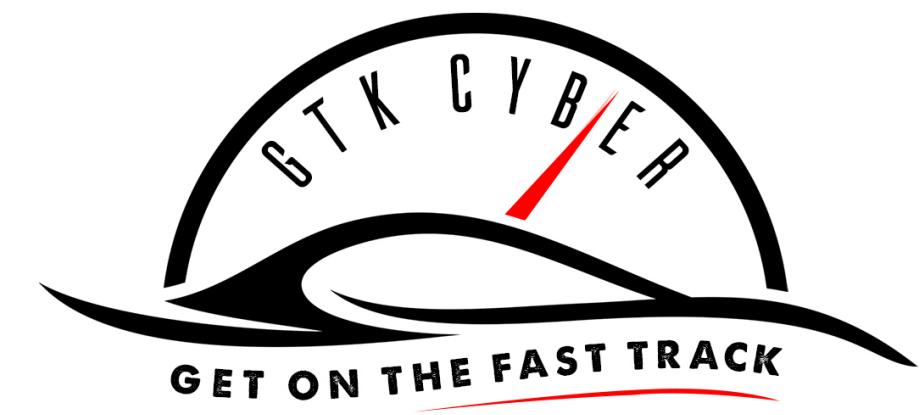
Ultimately, the quality,
relevance and integrity of the
content is most important.

Chart Suggestions—A Thought-Starter



Visualization Goals

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



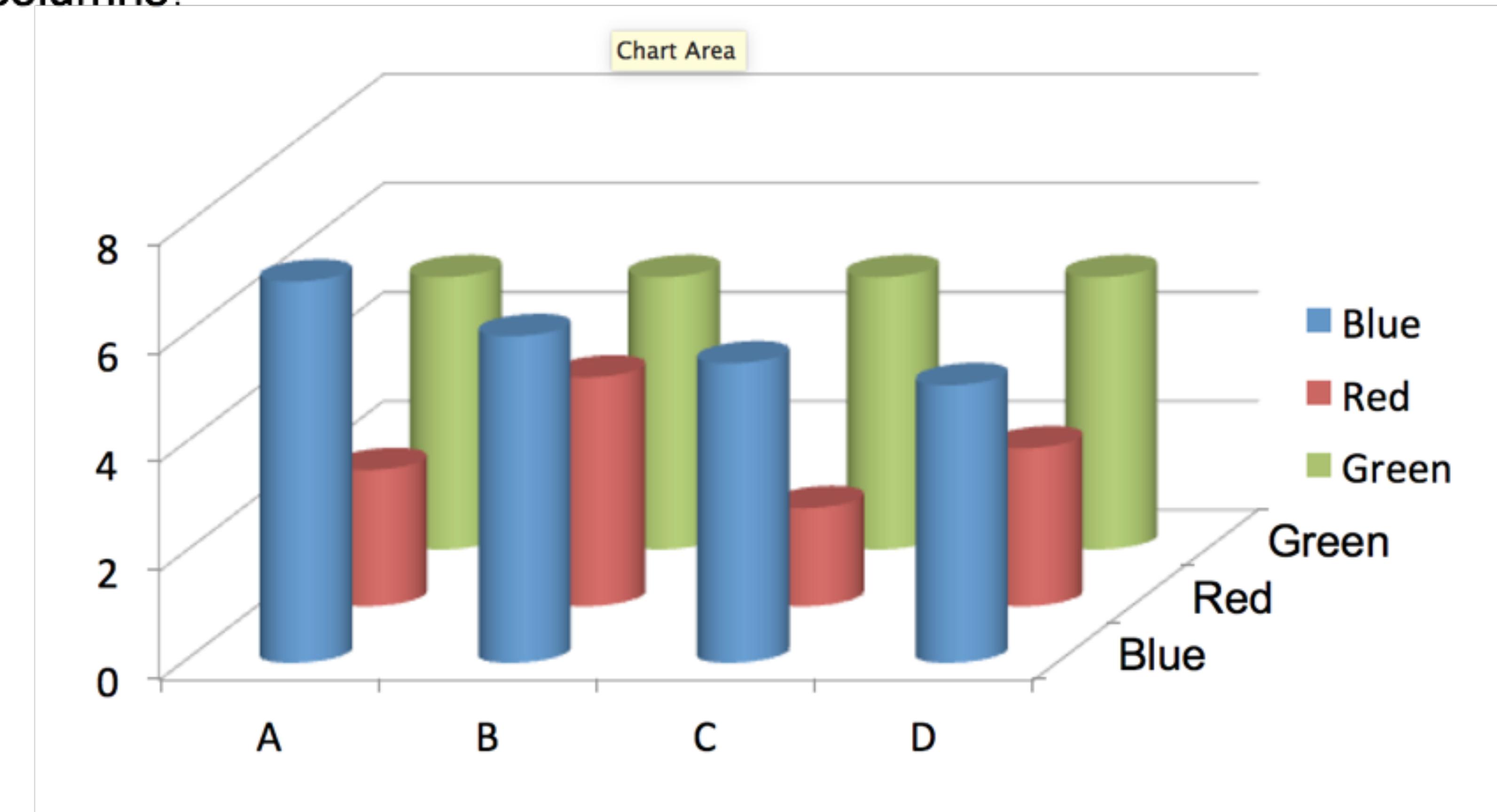
Elements of Good Visualizations

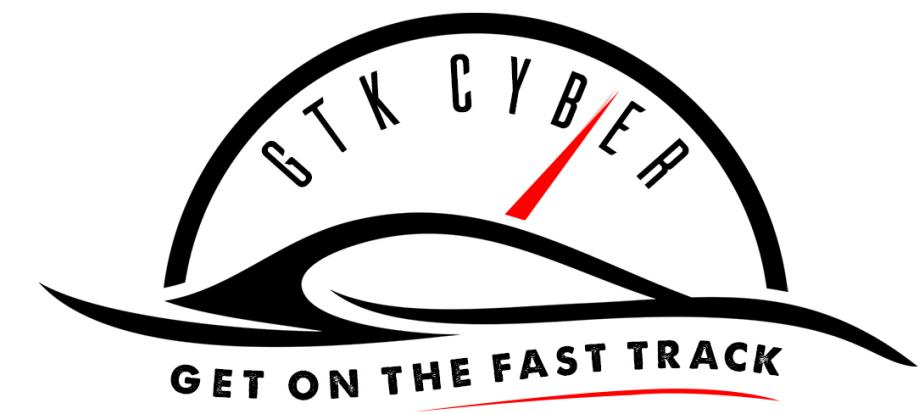
1. Graphical Integrity
2. Simple
3. Proper Display
4. Proper Color
5. Tells a story



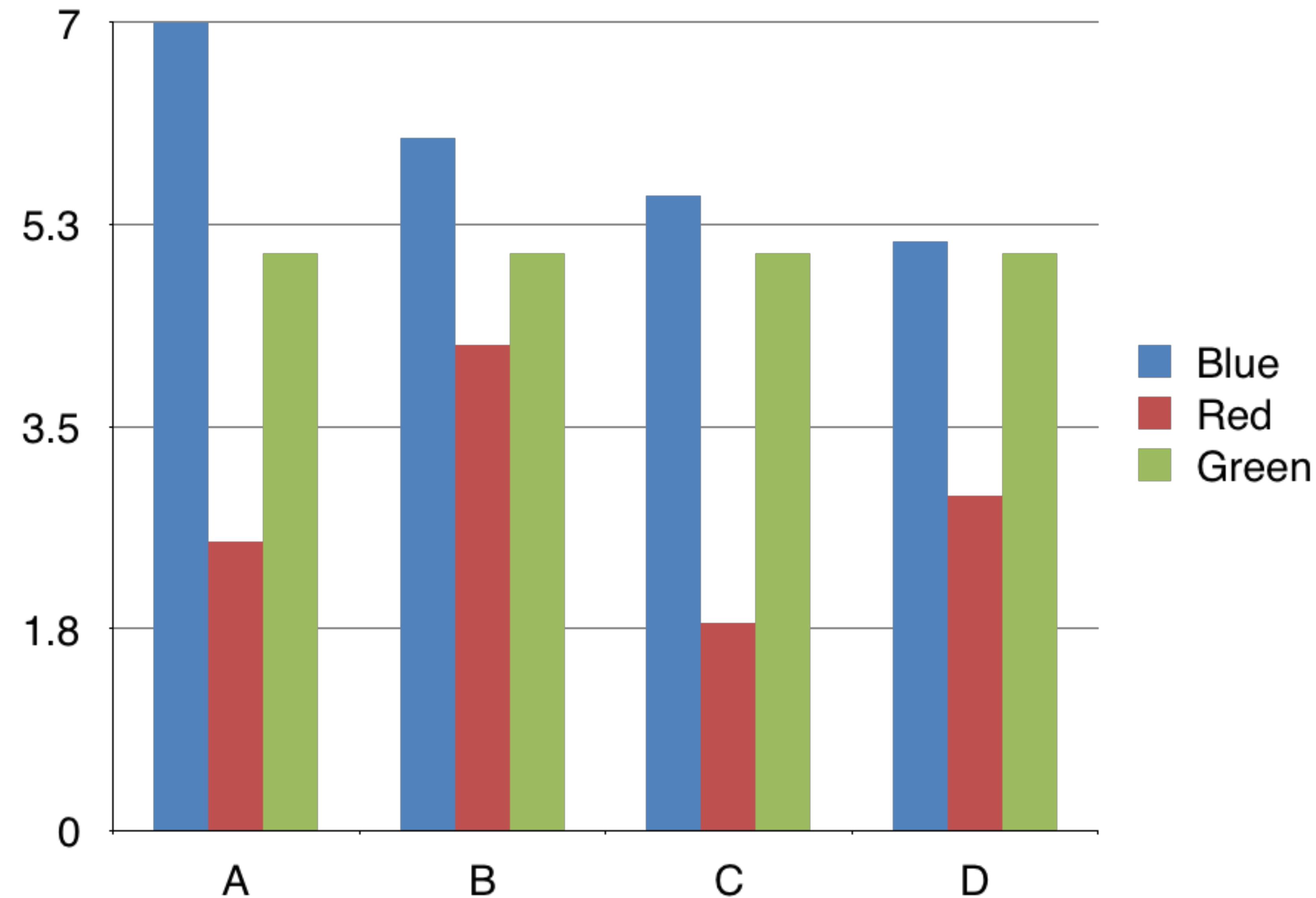
Proper Display

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?



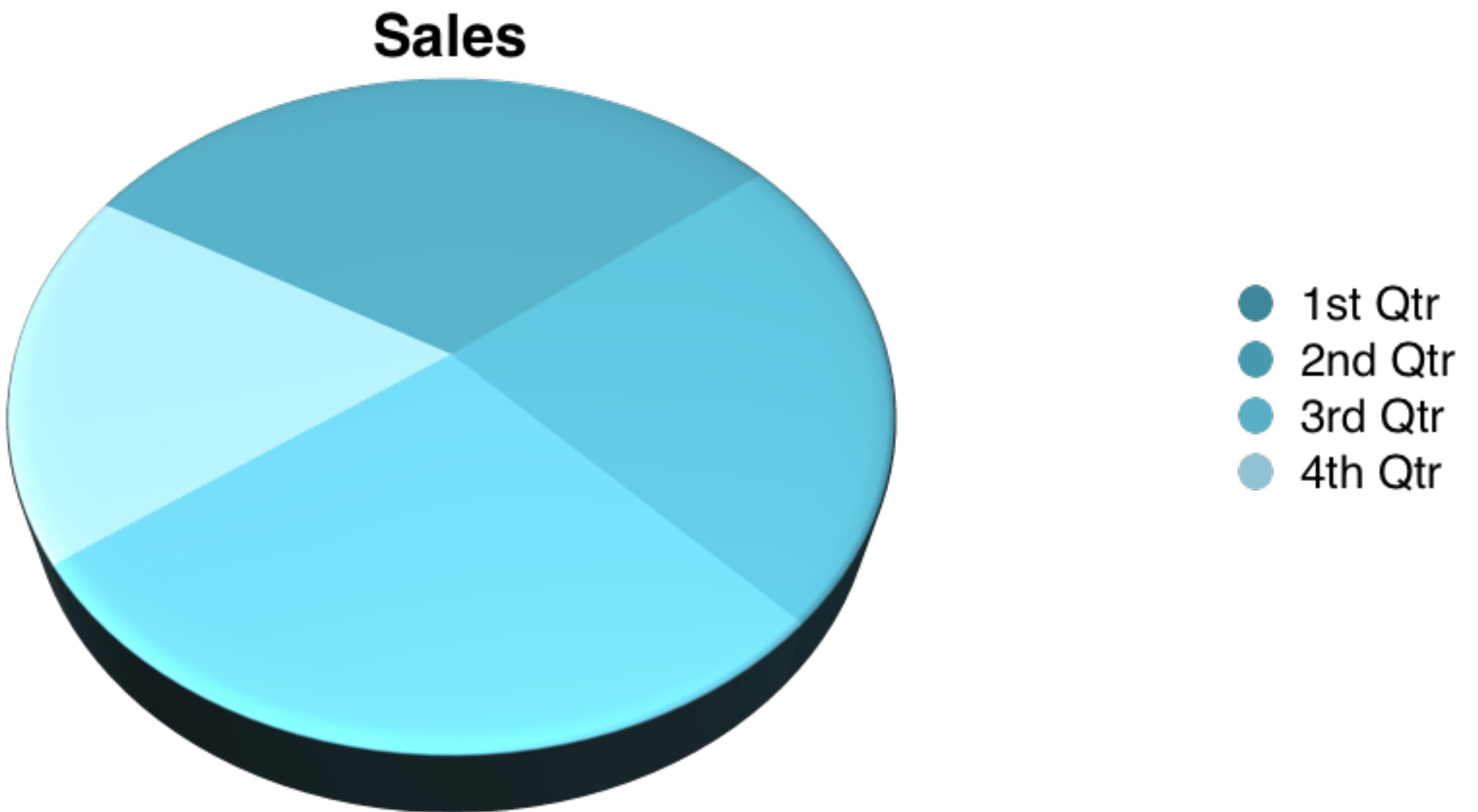


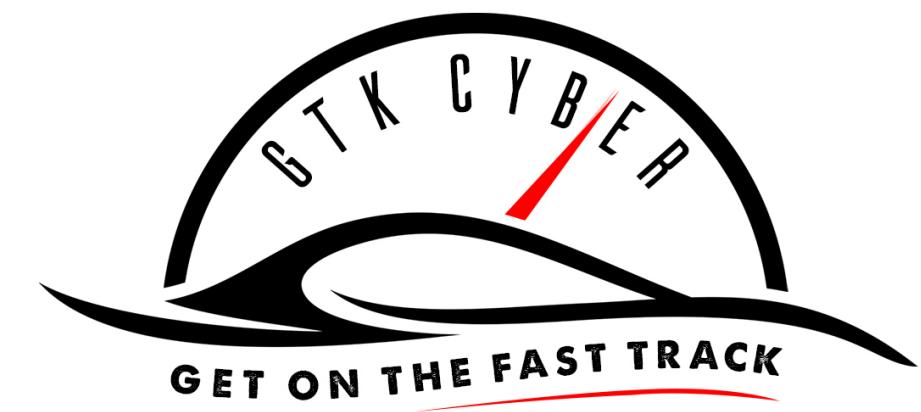
Proper Display



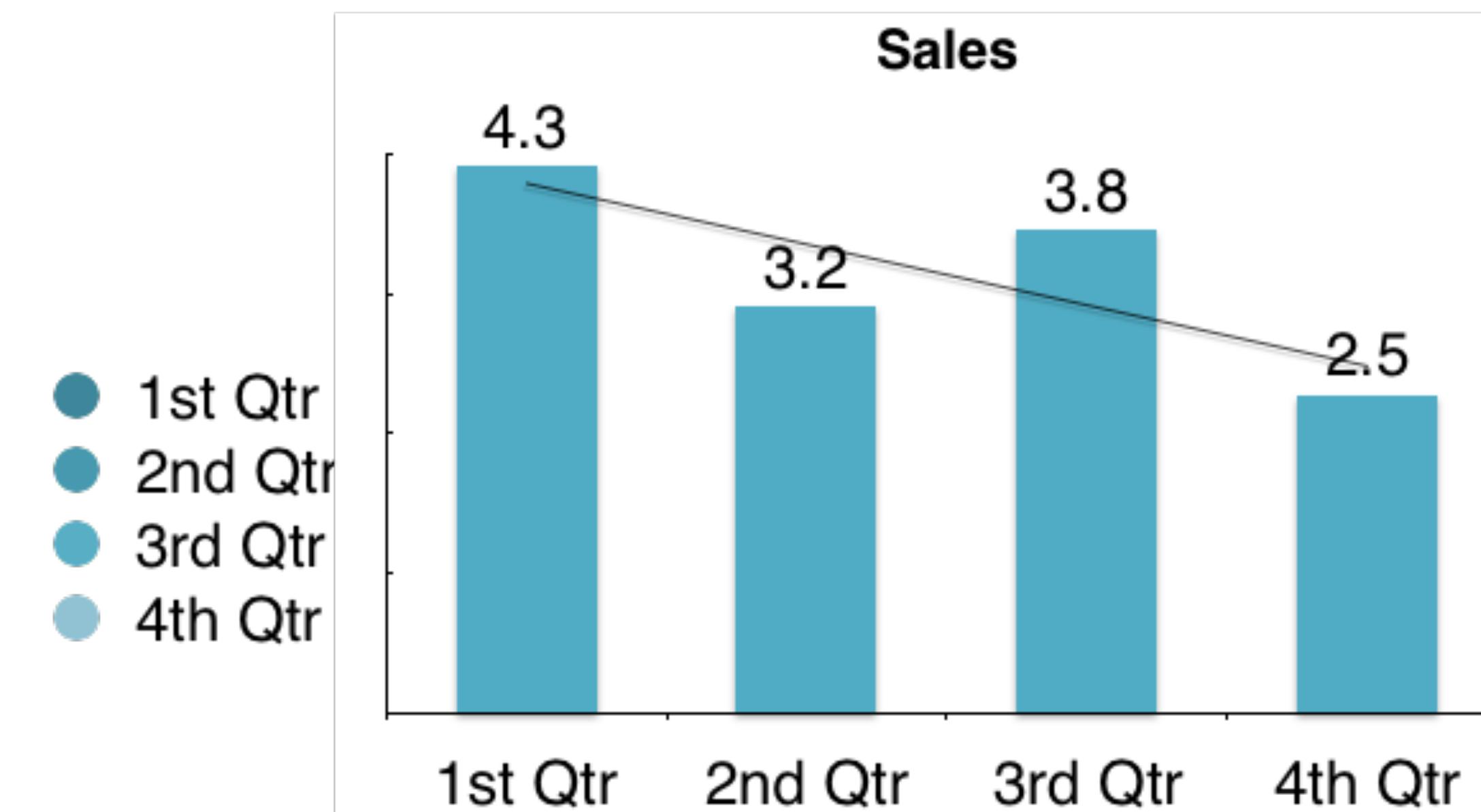


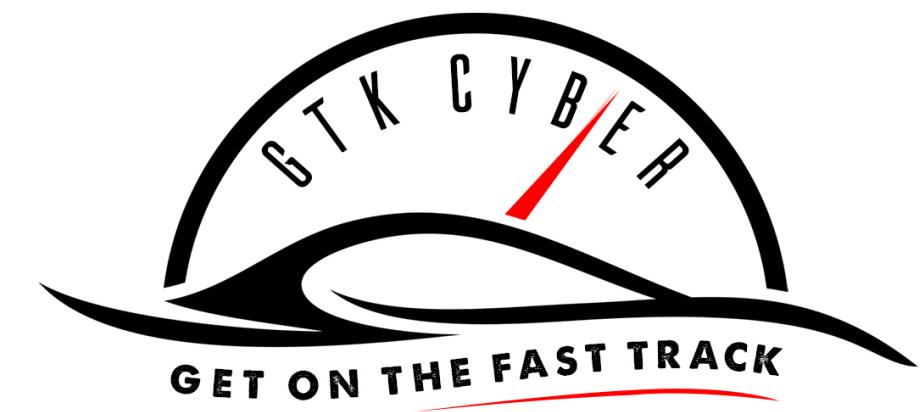
Proper Display



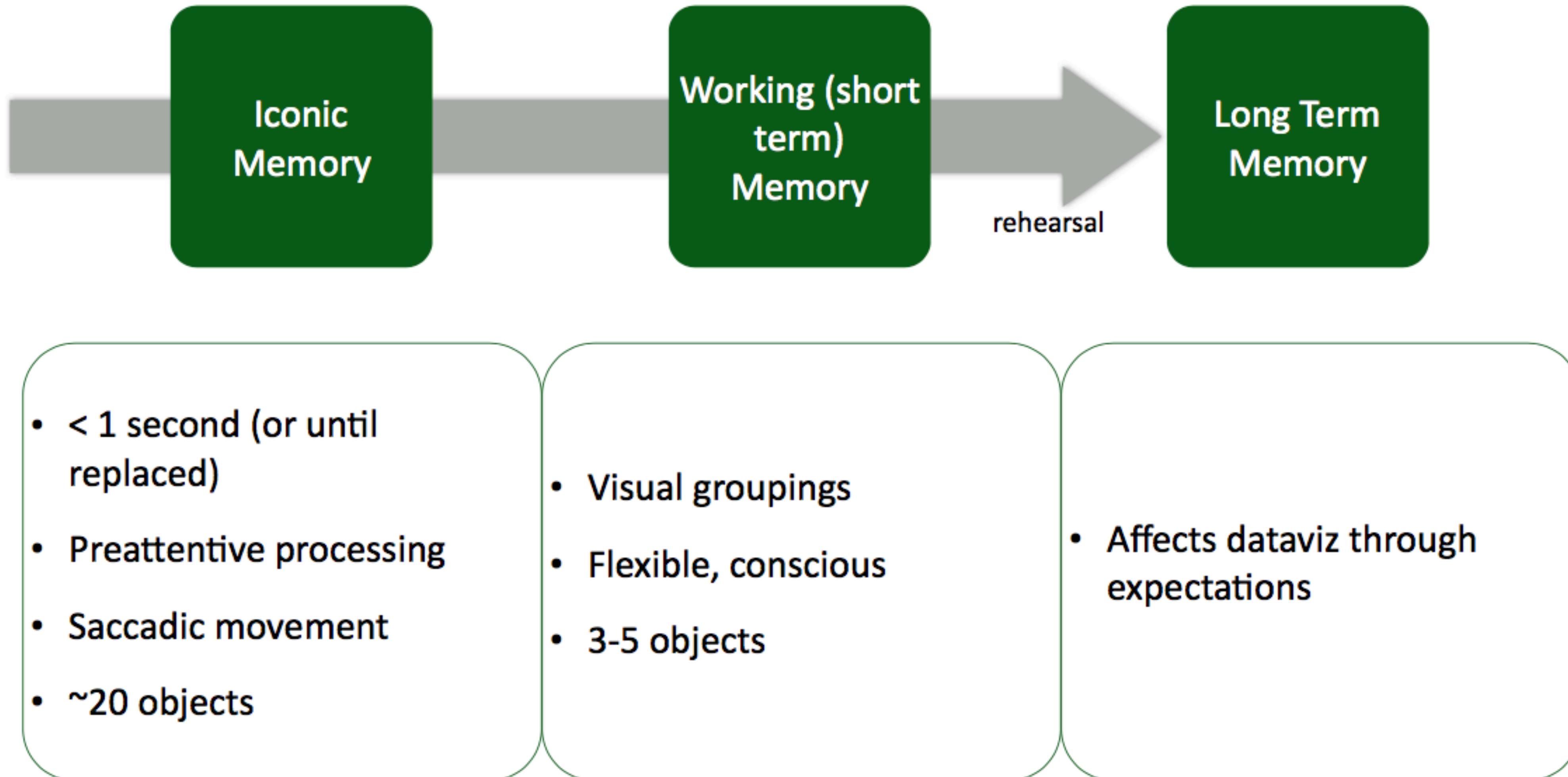


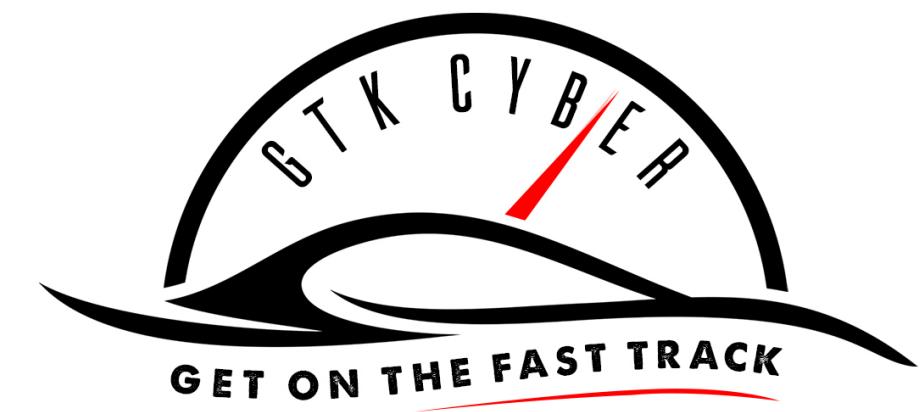
Proper Display





Visual Processing System

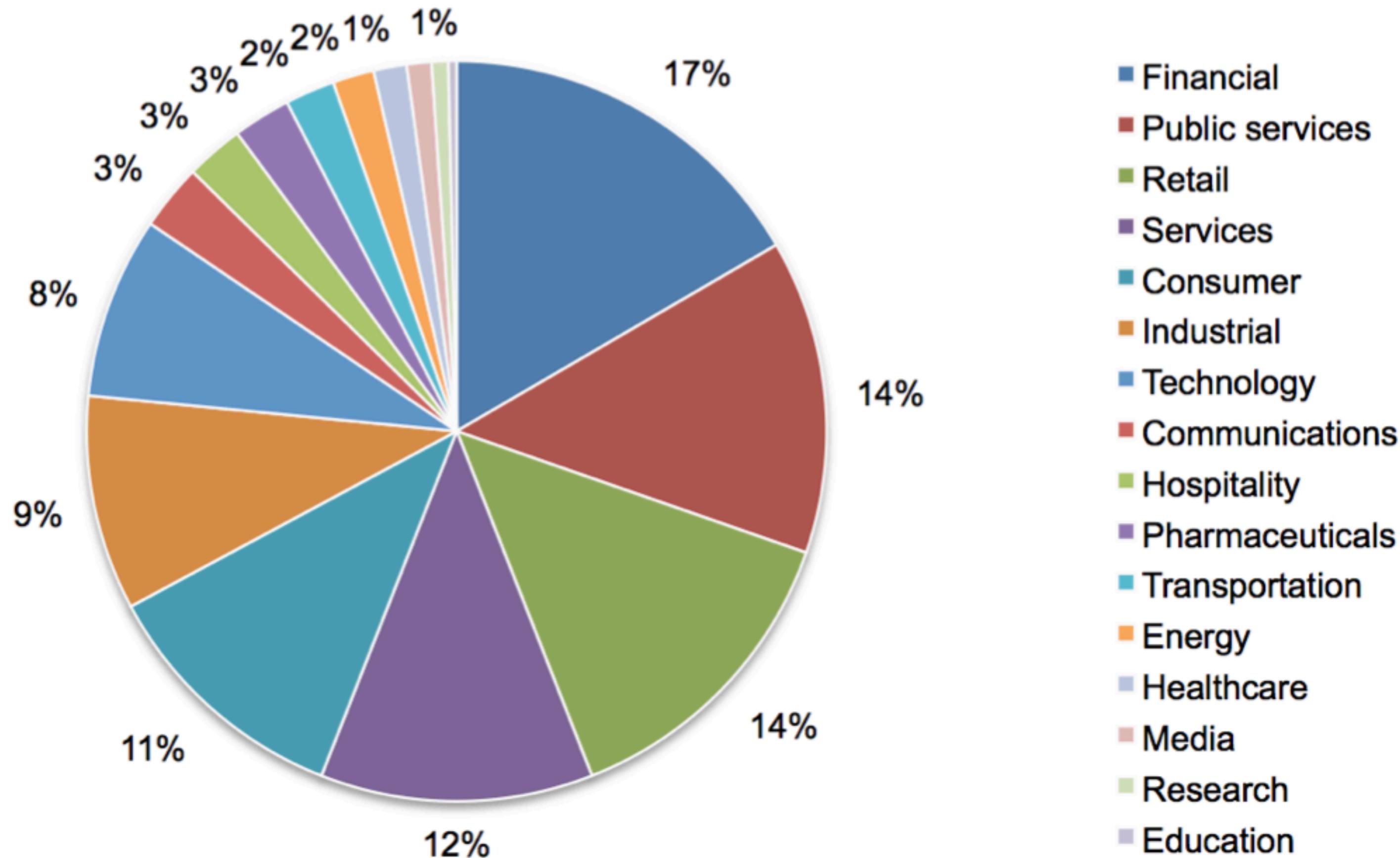


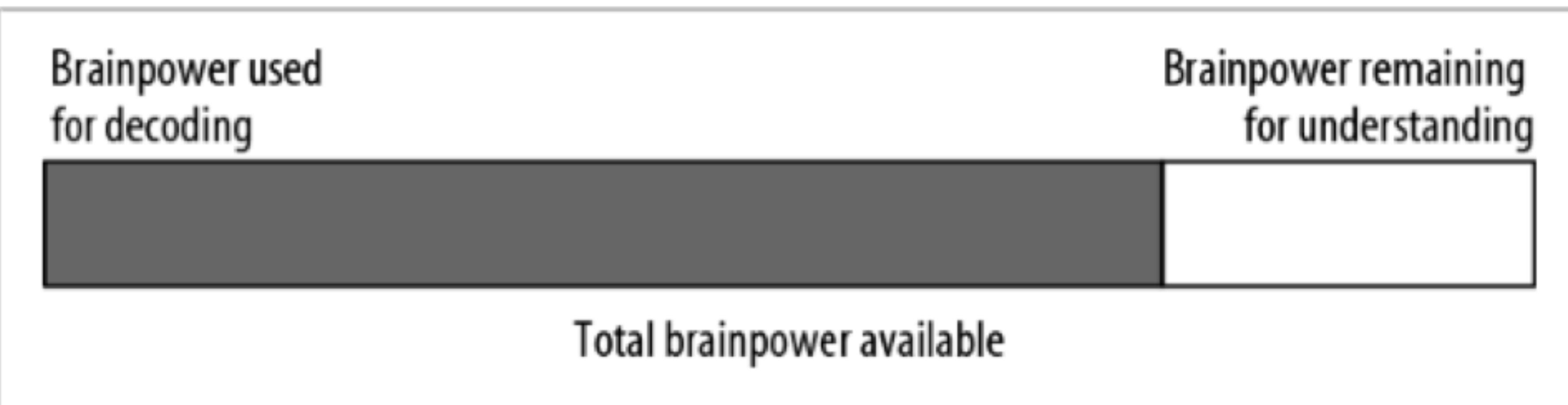
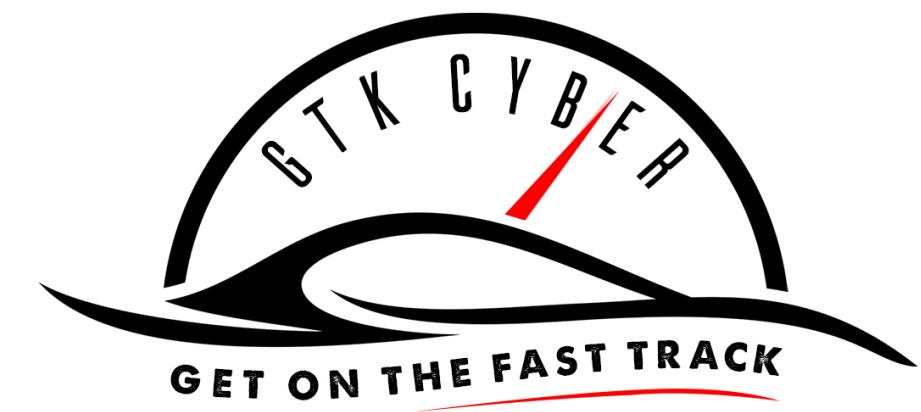


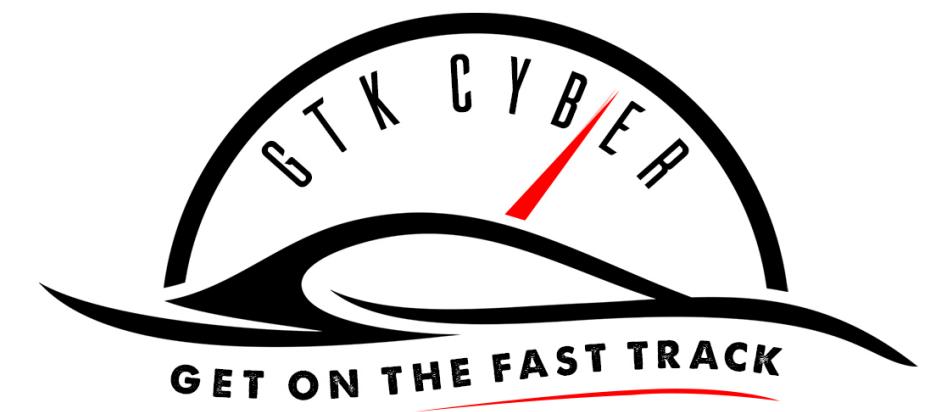
Overworking Visual Memory

Figure 20. Distribution of the benchmark sample by industry segment

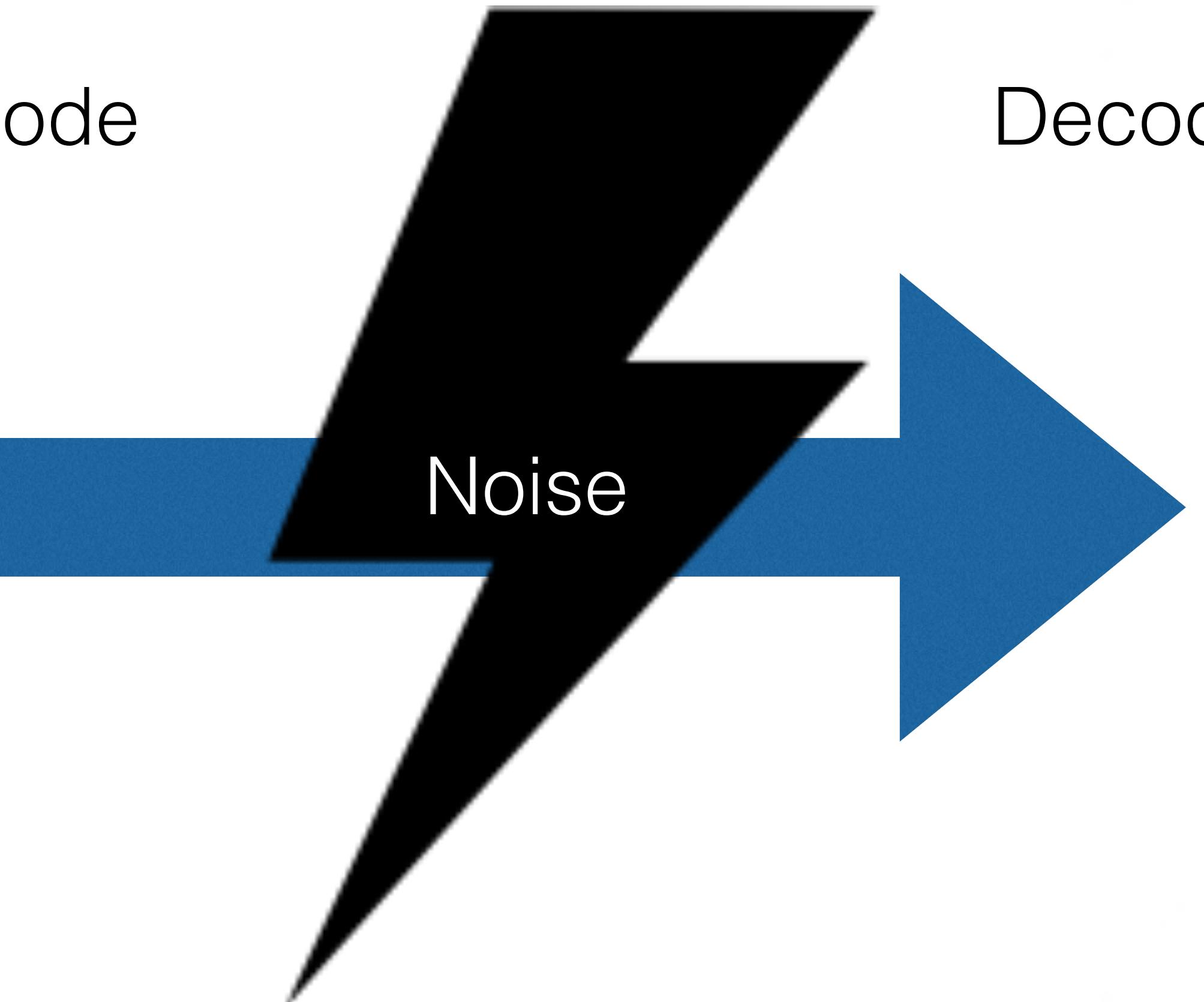
Consolidated (n = 277 organizations)

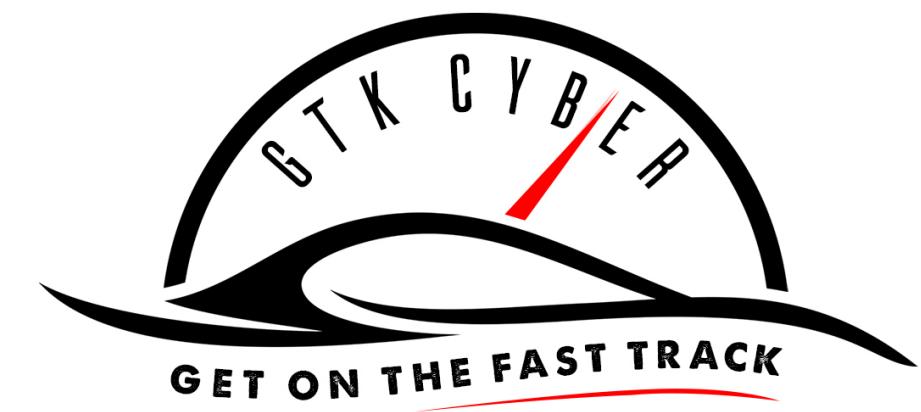




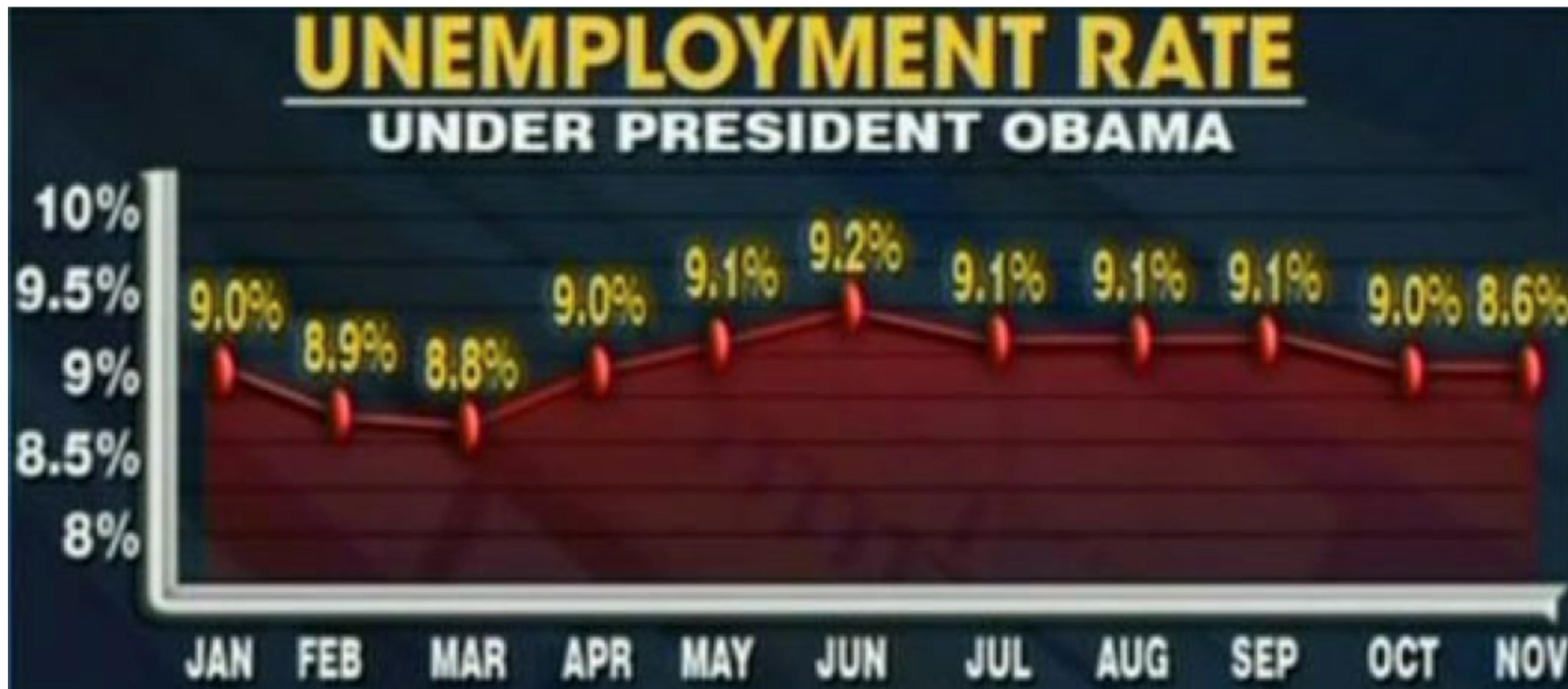


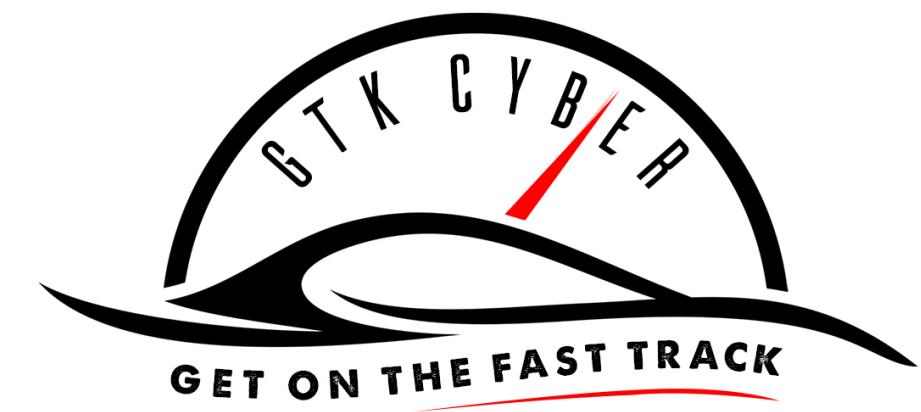
Encode



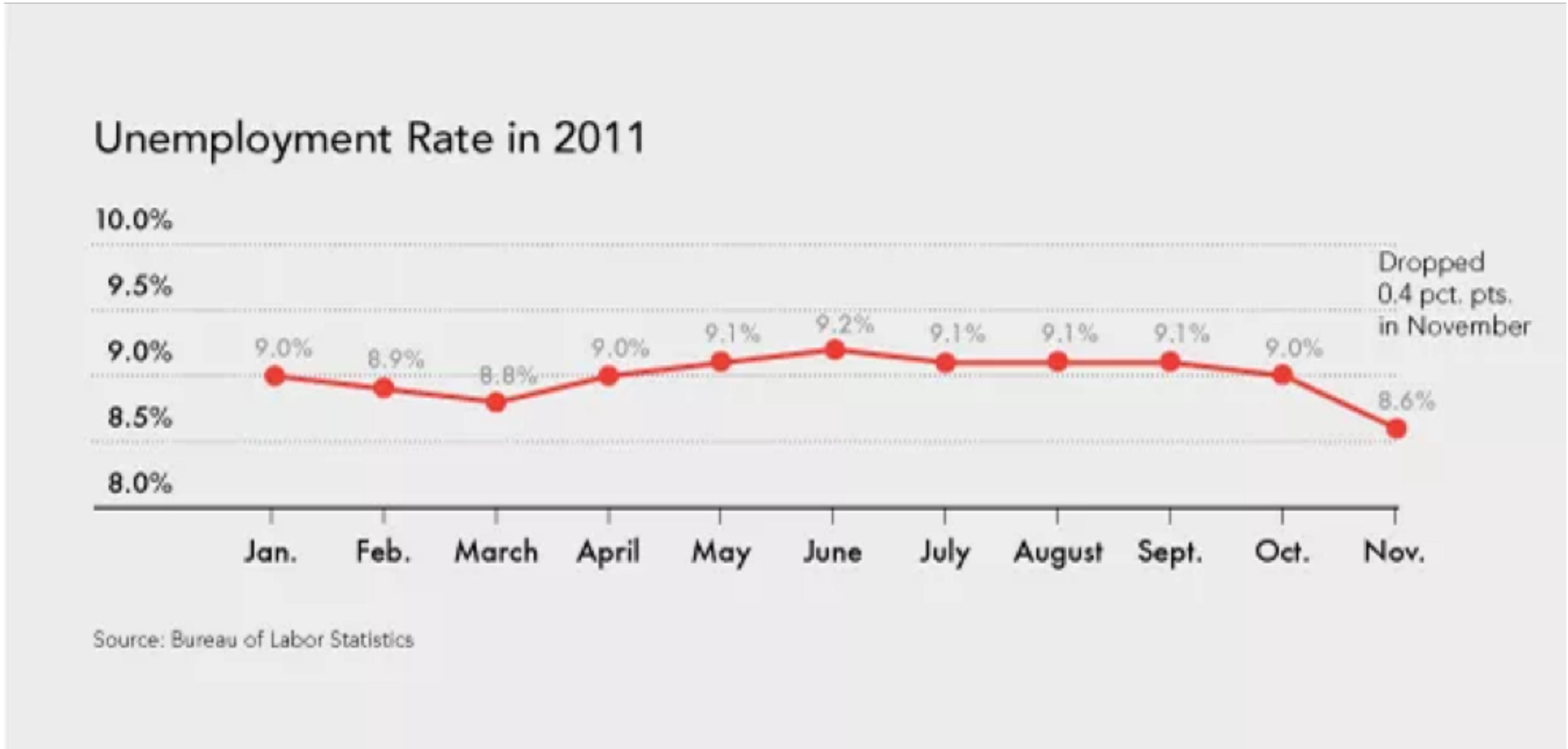


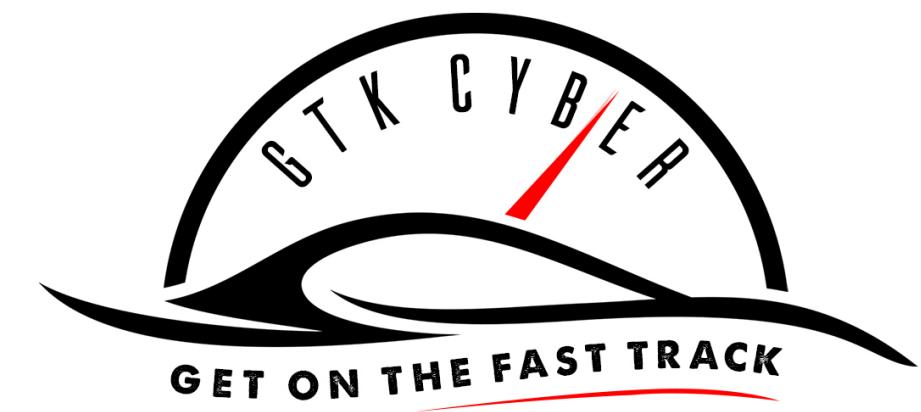
Graphical Integrity



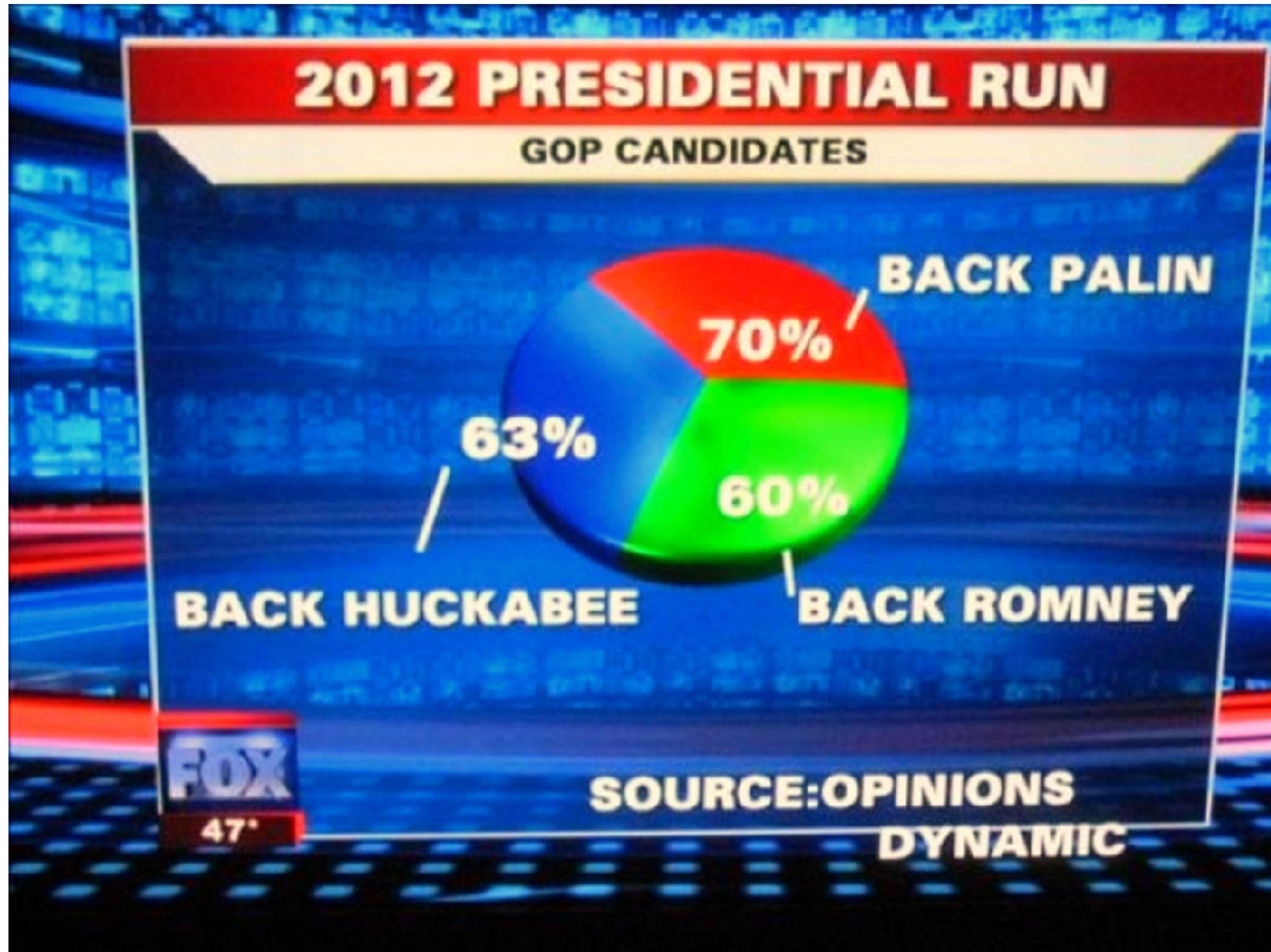


Graphical Integrity





Graphical Integrity?



STOCK MARKET SLIDE

NOV. 4, 2008
ELECTION DAY
9625.30

FRIDAY
MAR. 5, 2009
6626.94

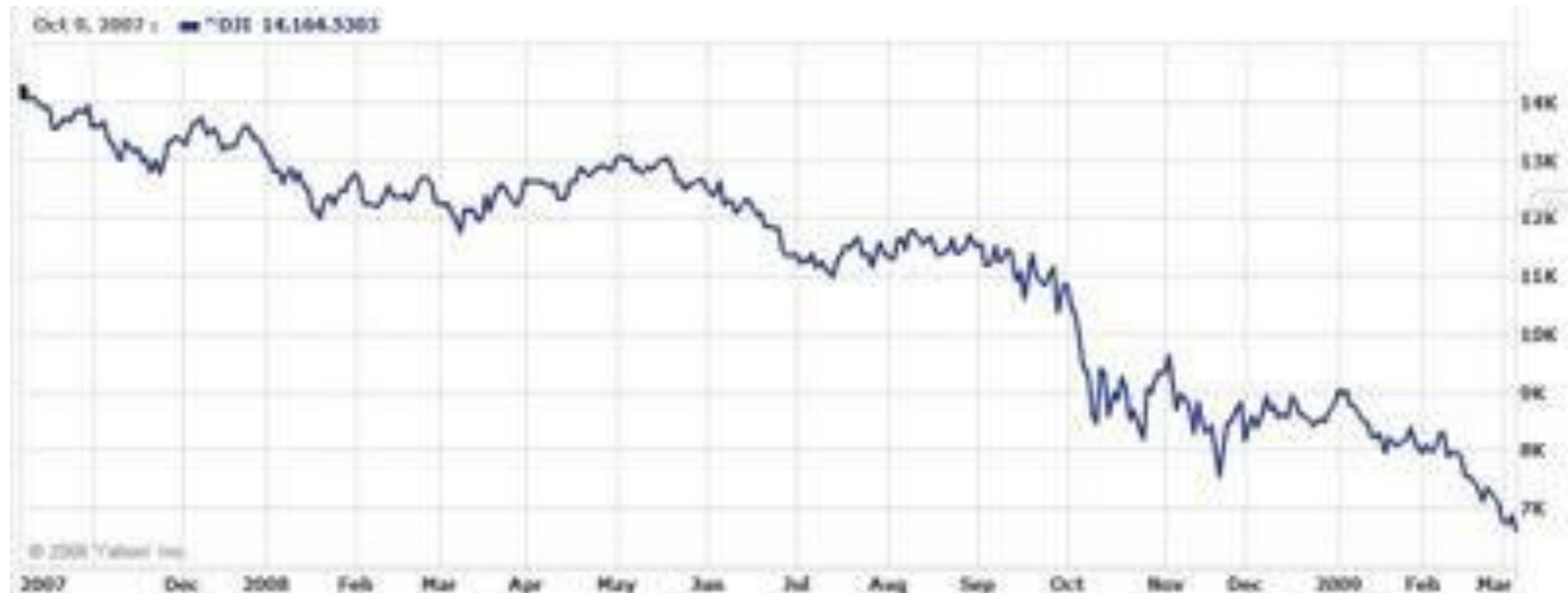
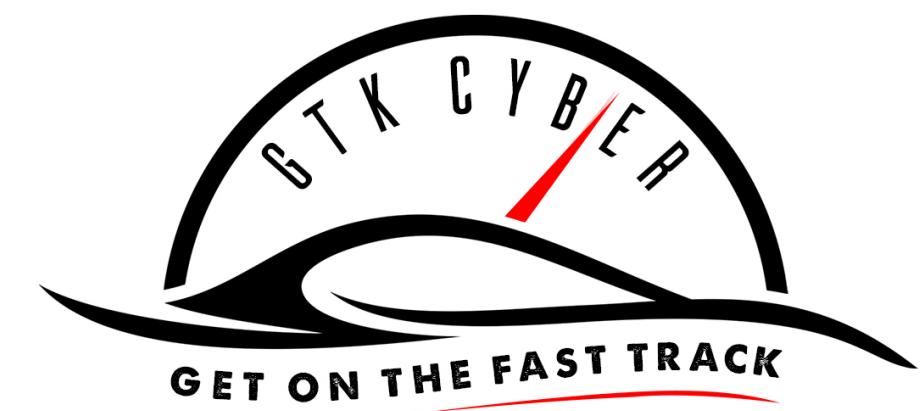
JAN. 20, 2009
INAUGURATION DAY
7949.10

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

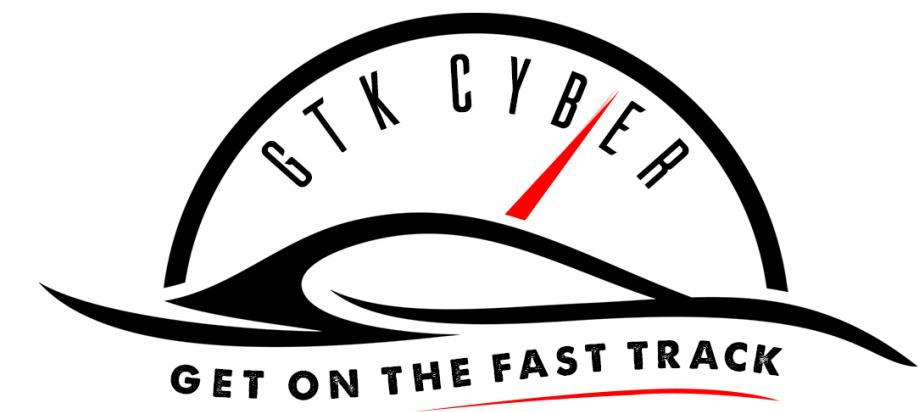
IN PRESIDENT

R TRUMP ✓
D CLINTON

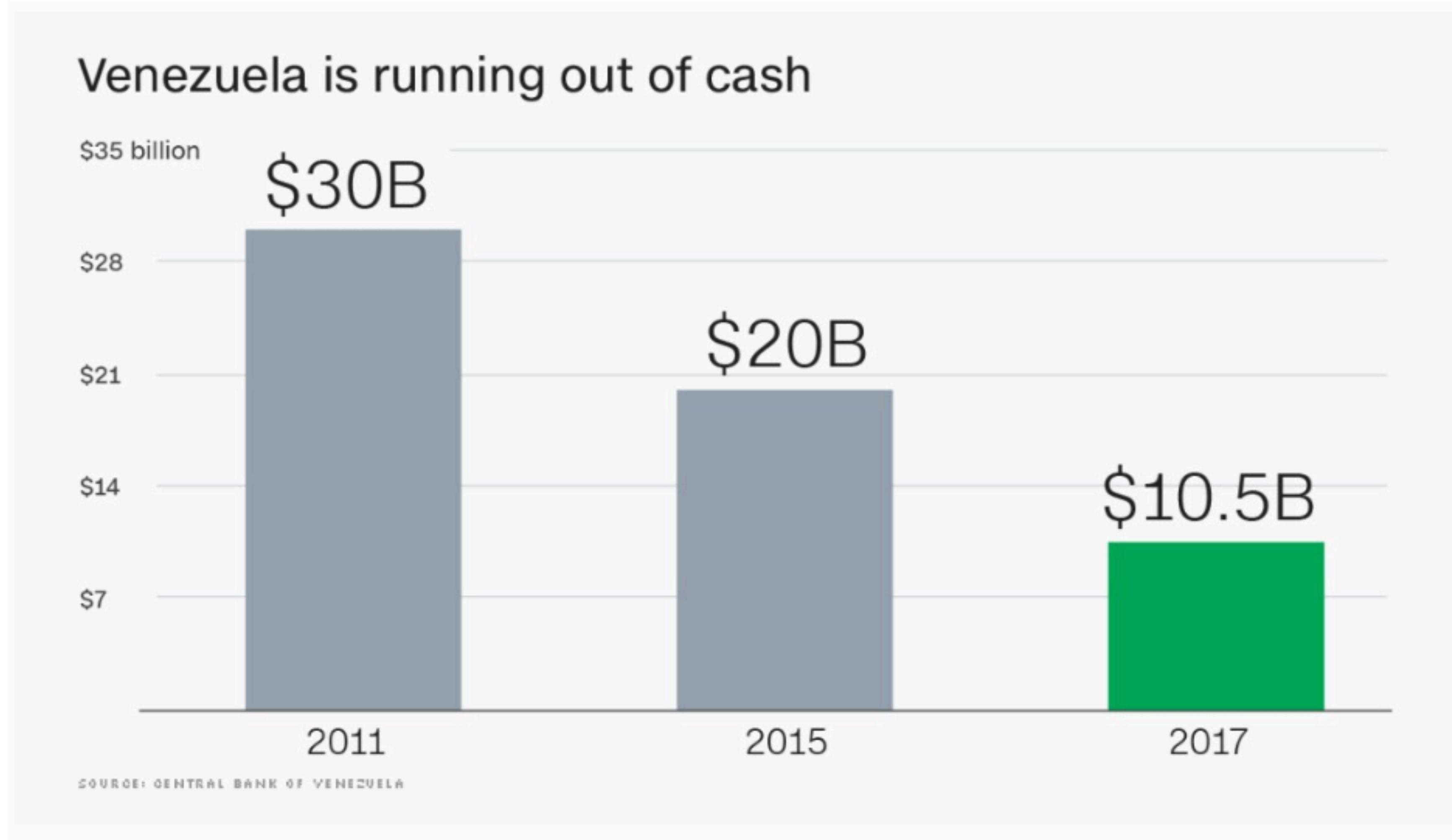
126,791 65.4%
59,284 30.6%

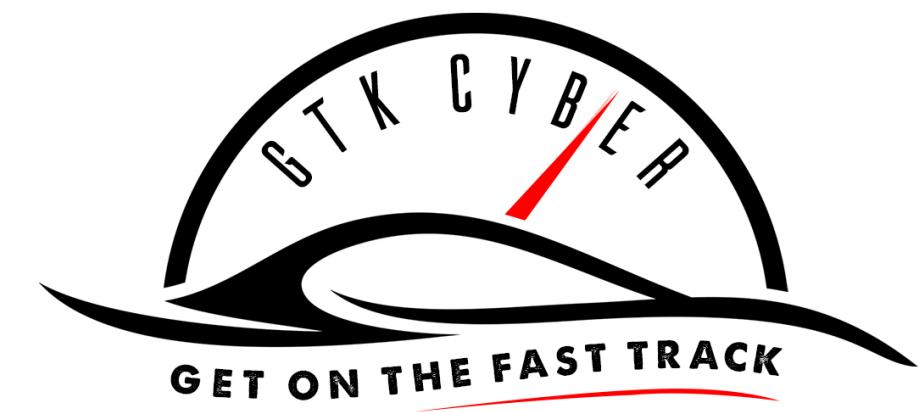
26:23
NEXT POLLS CLOSE
LIVE
CNN

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% satisfied (D Clinton), 14% dissatisfied (R Trump), and 4% neutral (J 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: R TRUMP ✓ (126,791, 65.4%) and D CLINTON (59,284, 30.6%). The time 26:23 and the text "NEXT POLLS CLOSE" are also visible.

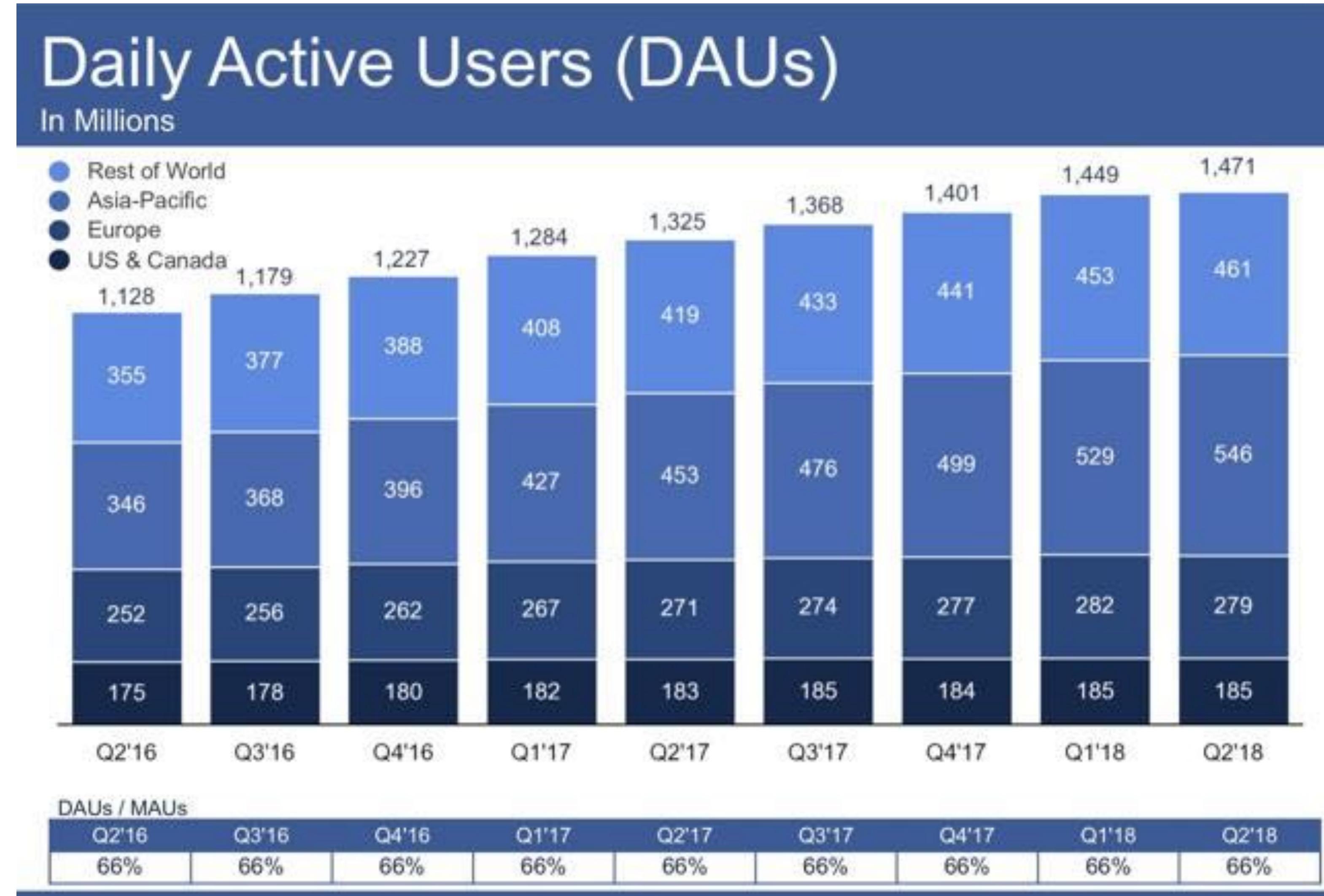


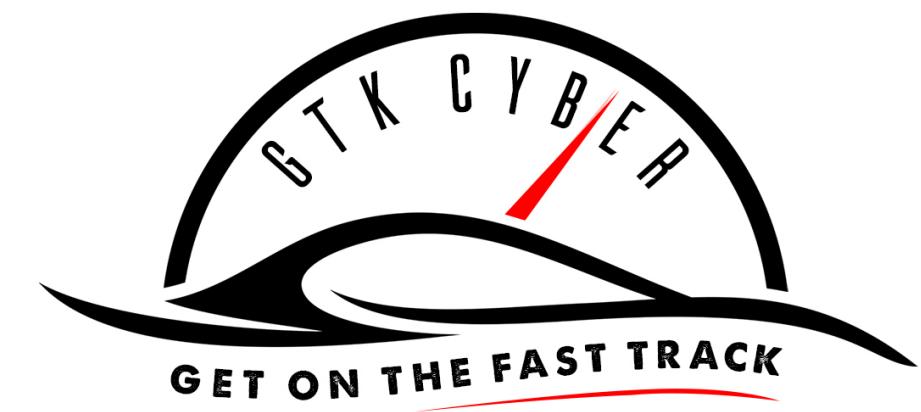
Graphical Integrity



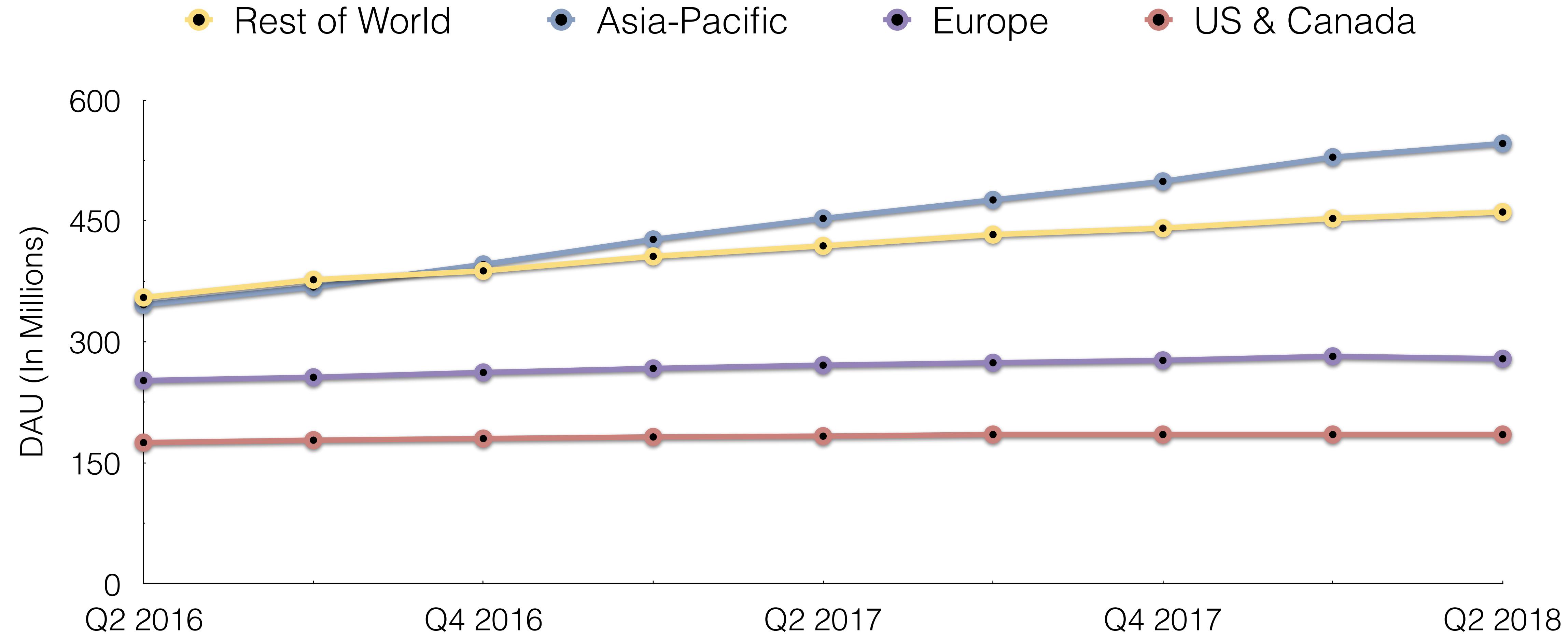


Proper Display



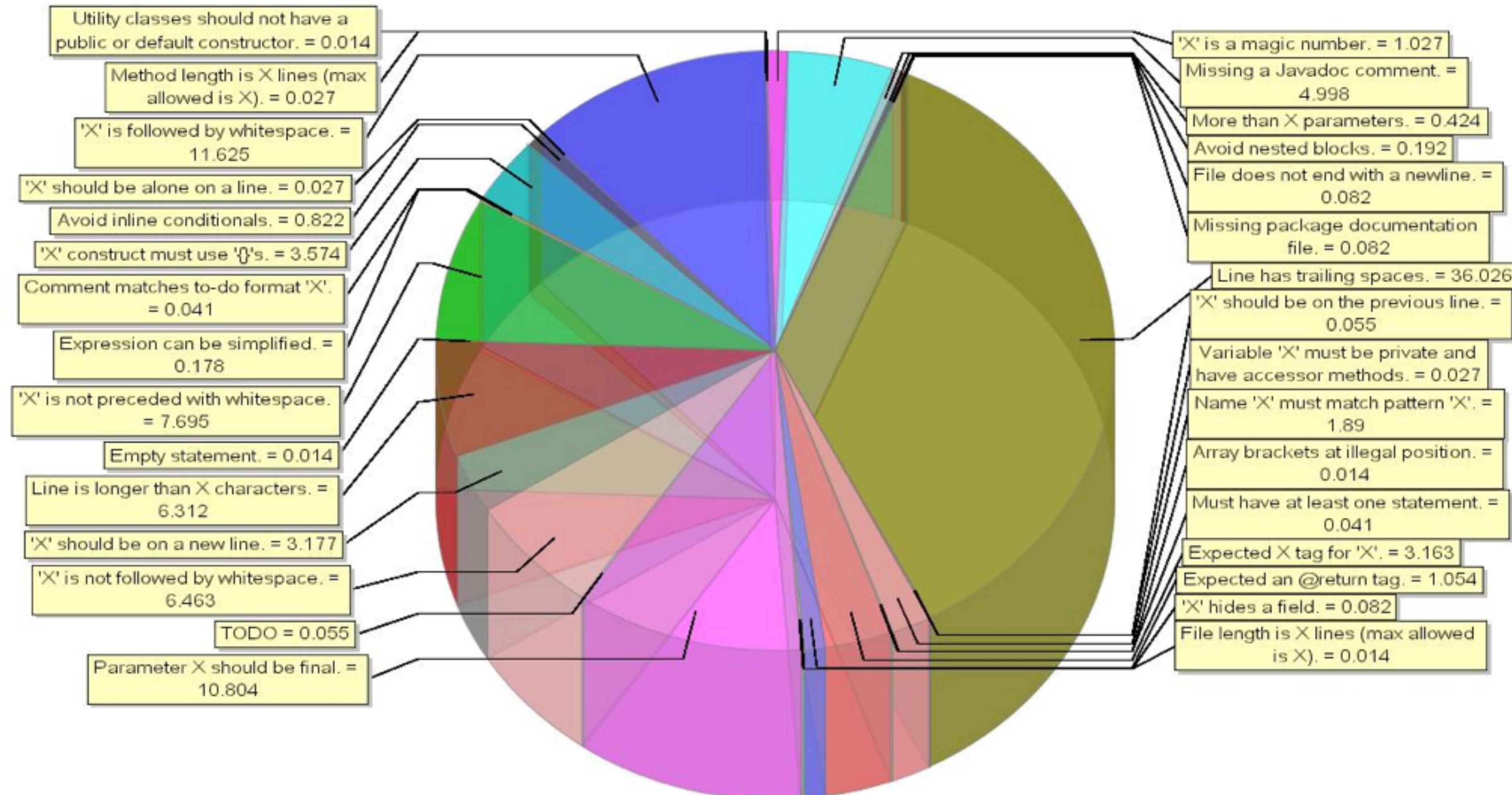


Proper Display



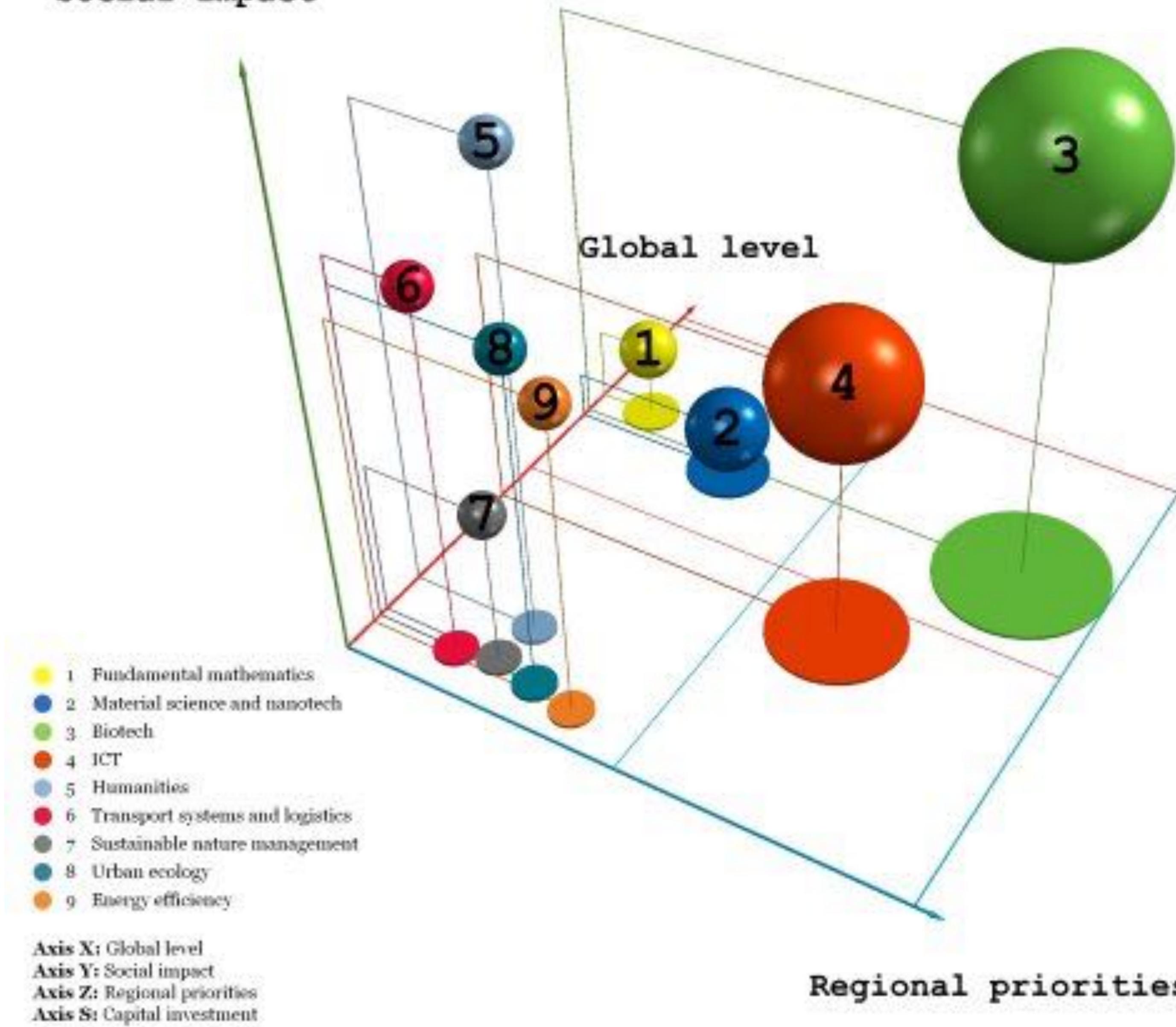


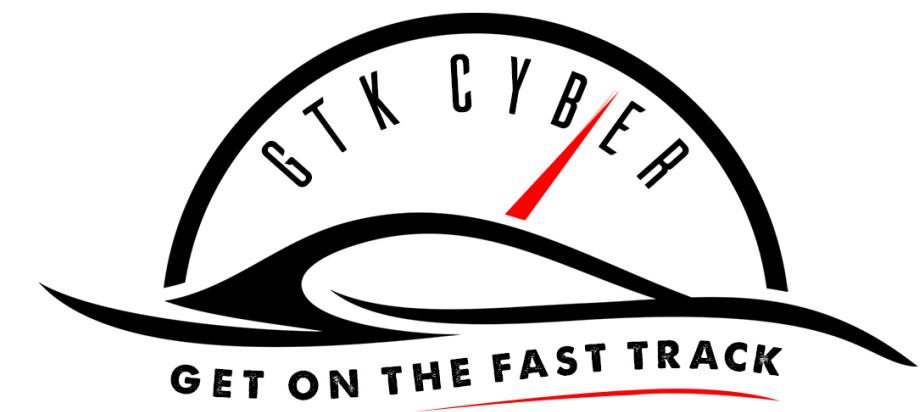
Simple





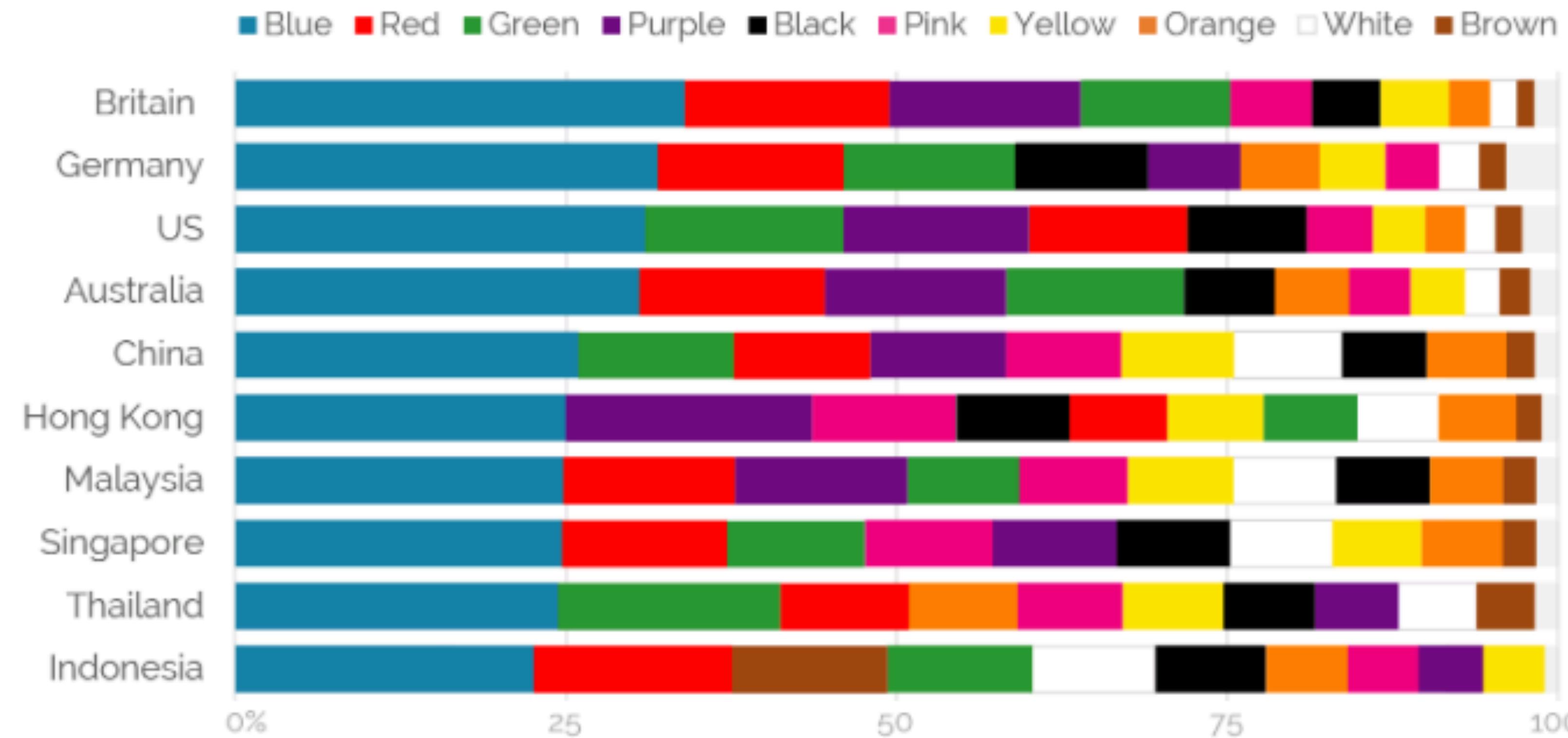
Social impact





Blue planet

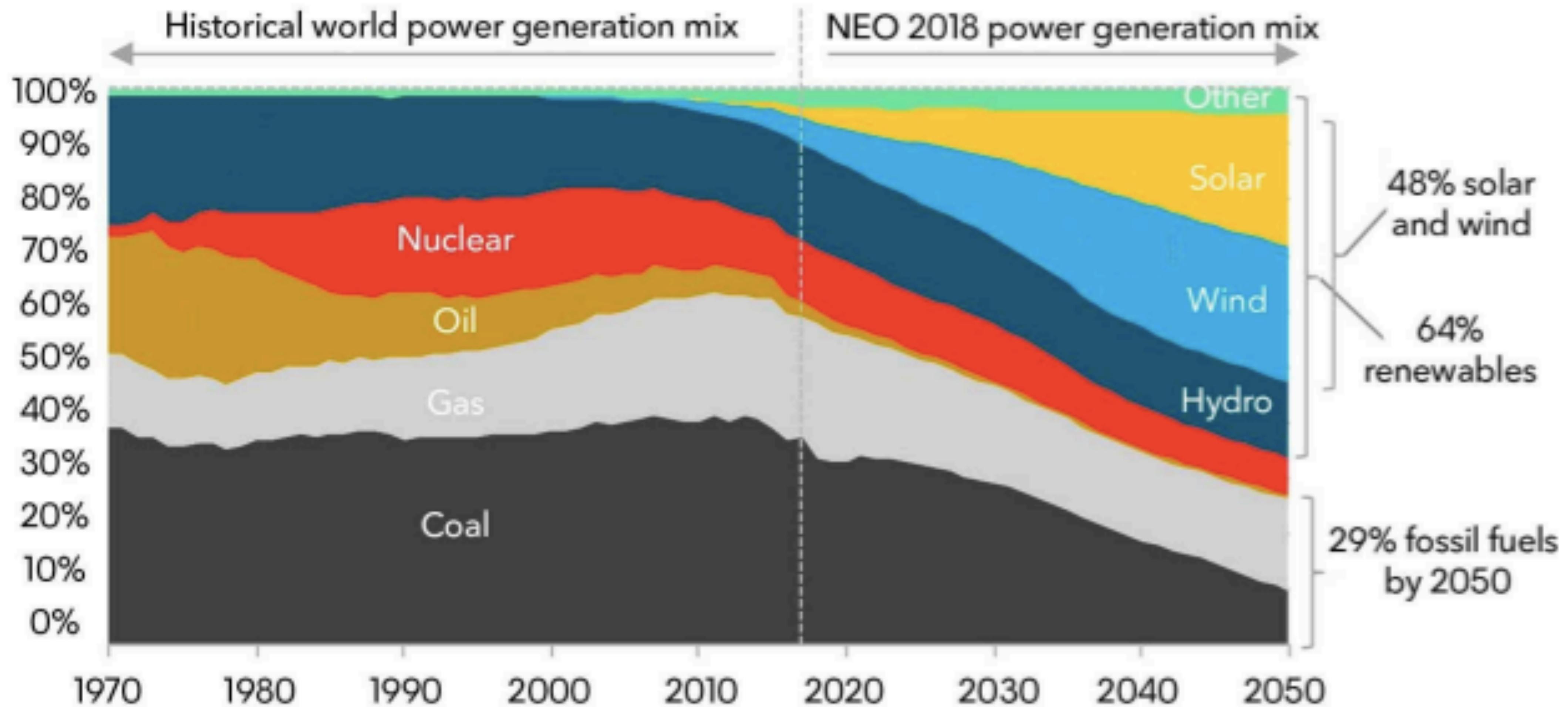
Which one of the colors listed below do you like the most?



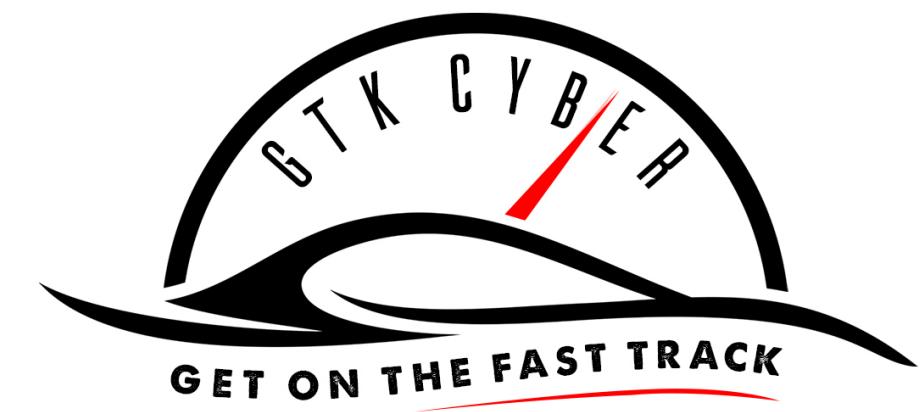
YouGov | yougov.com



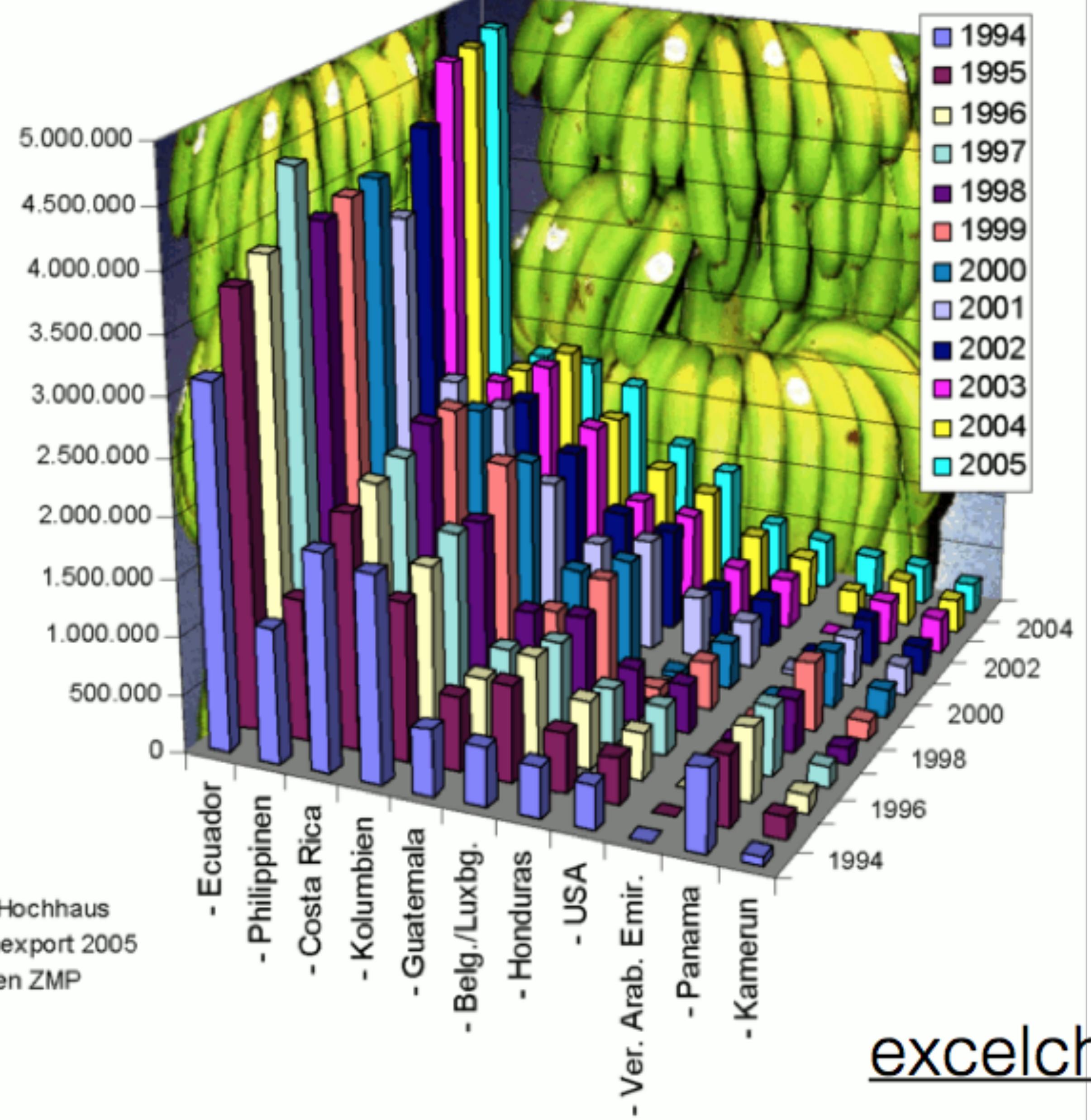
Power generation mix



Source: Bloomberg NEF

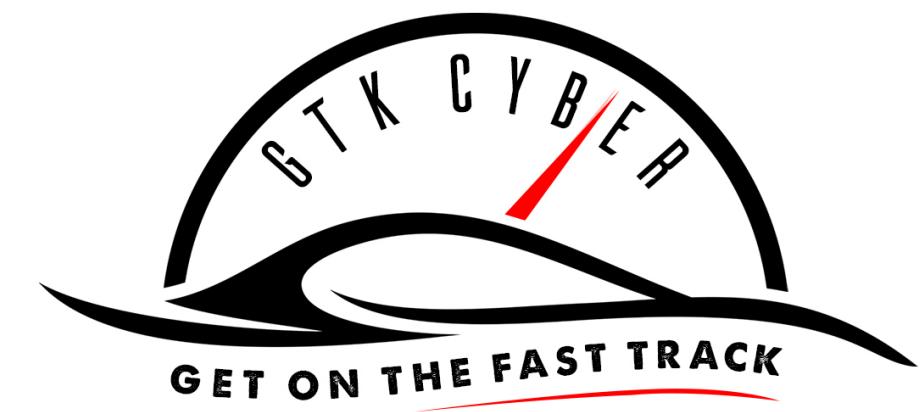


Export von Bananen in Tonnen von 1994-2005



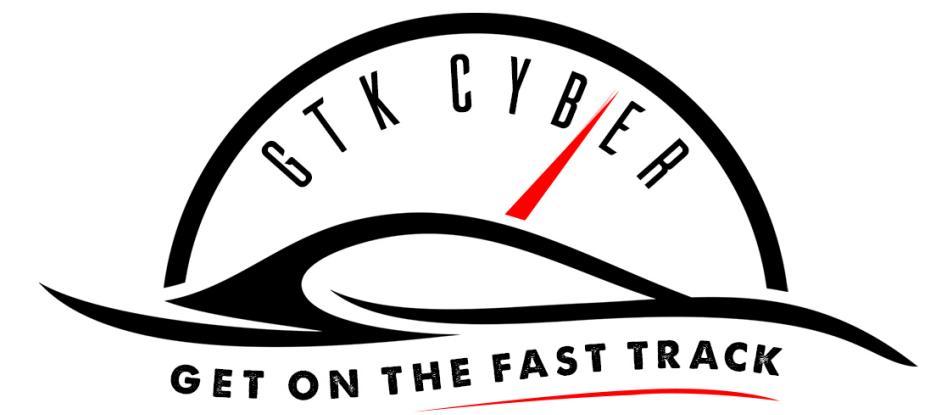
Dr. Hochhaus
Banlexport 2005
Daten ZMP

excelcharts.com

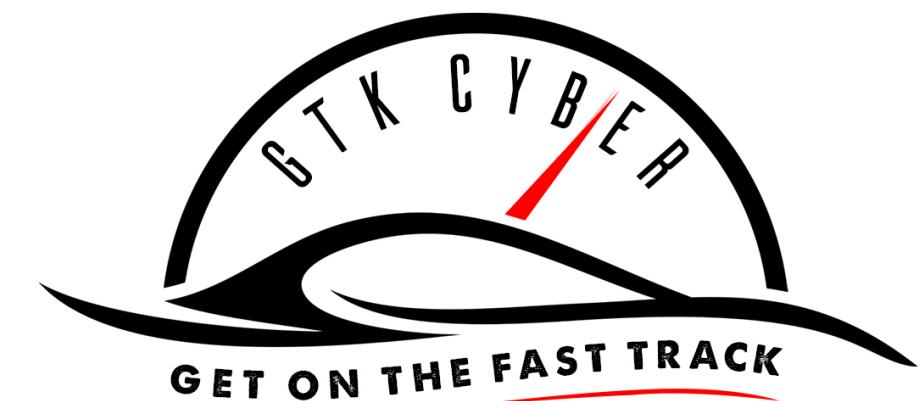


Proper Display





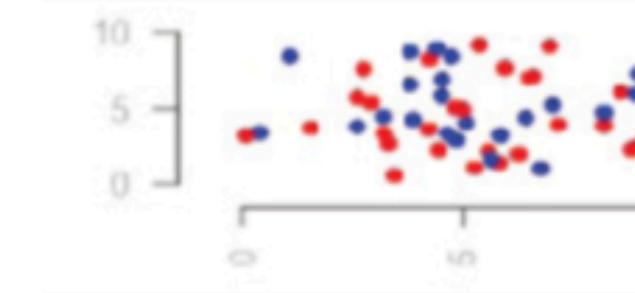
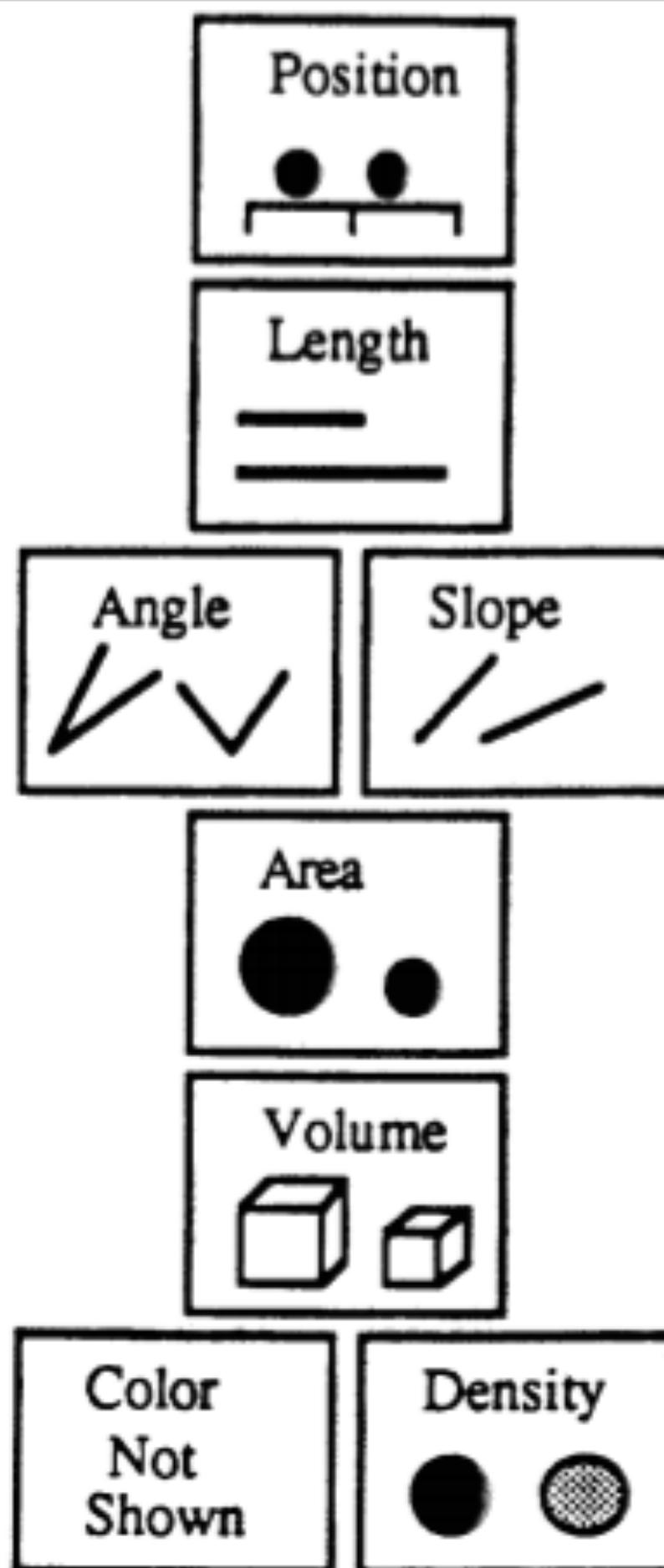
Visual Encoding



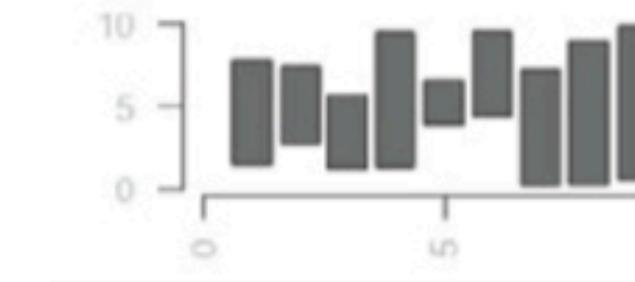
More accurate



Less accurate



Position



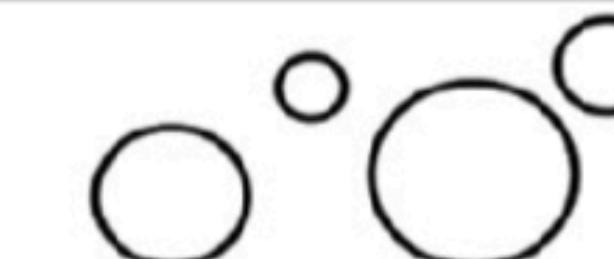
Length



Direction/Slope



Angle



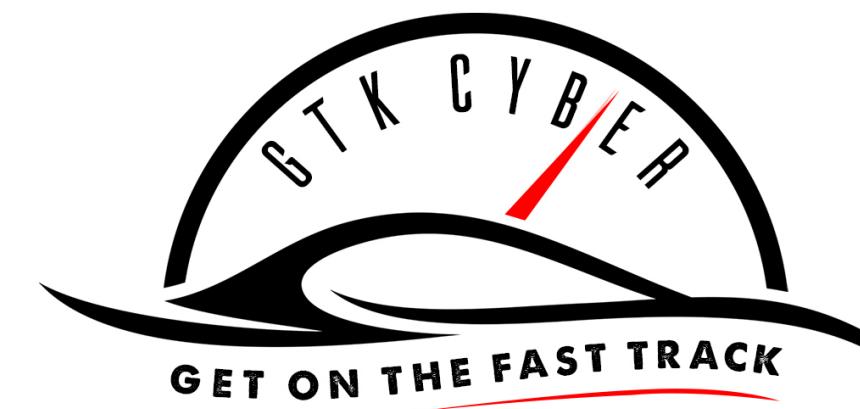
Area



Density/Saturation

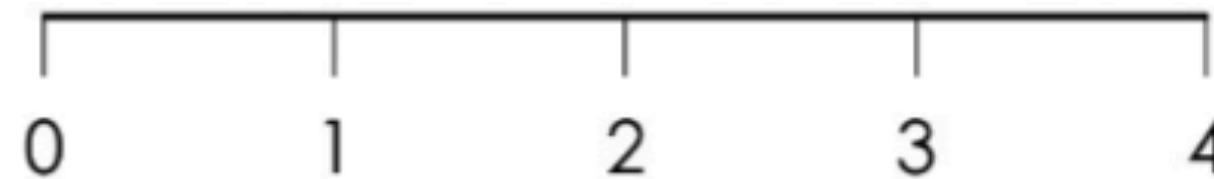


Color Hue



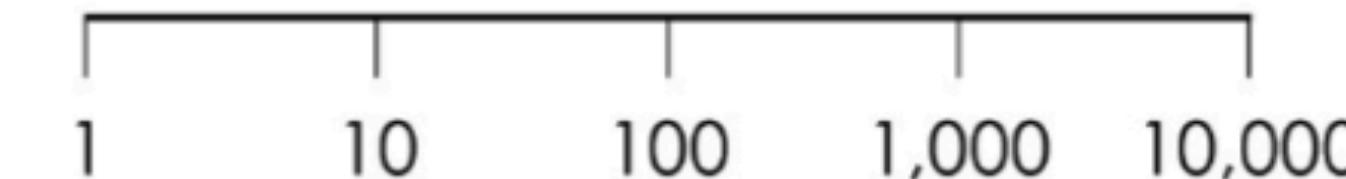
Linear

Values are evenly spaced



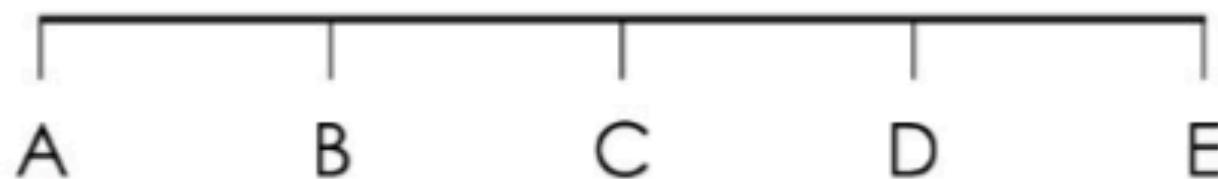
Logarithmic

Focus on percent change



Categorical

Discrete placement in bins



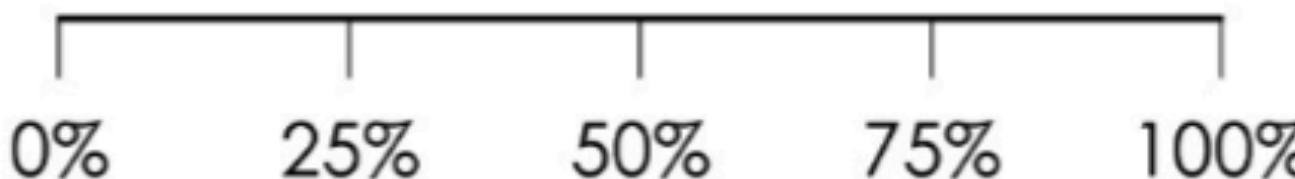
Ordinal

Categories where order matters



Percent

Representing parts of a whole

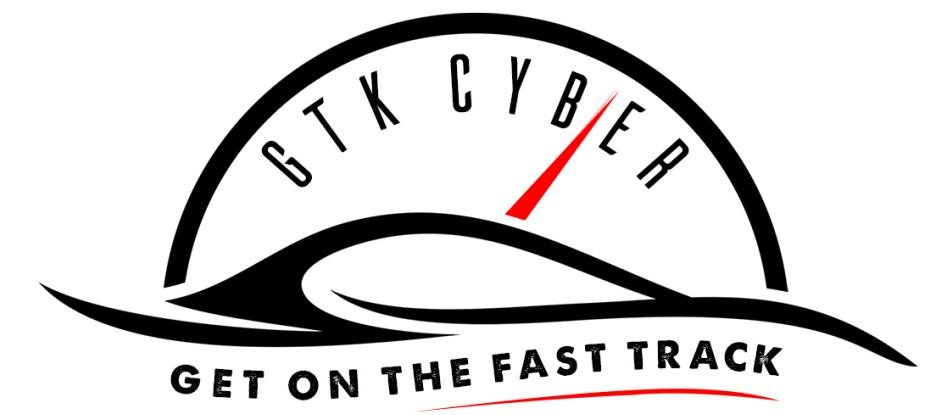


Time

Units of months, days, or hours

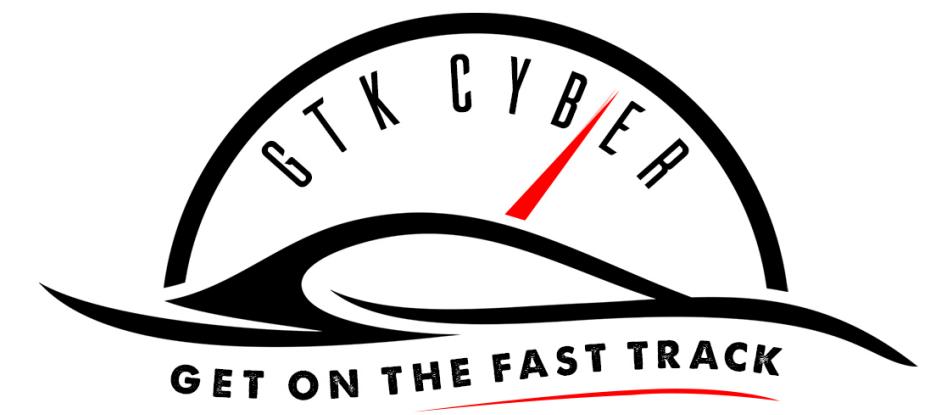


Source: Nathan Yau, Data Points

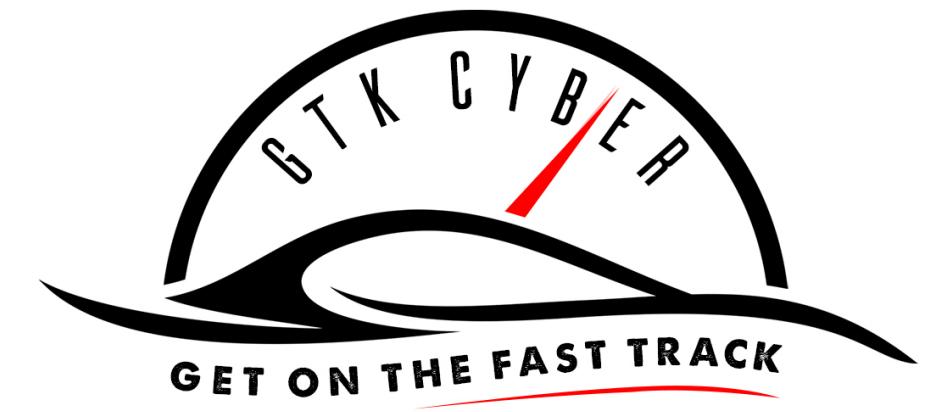


Color

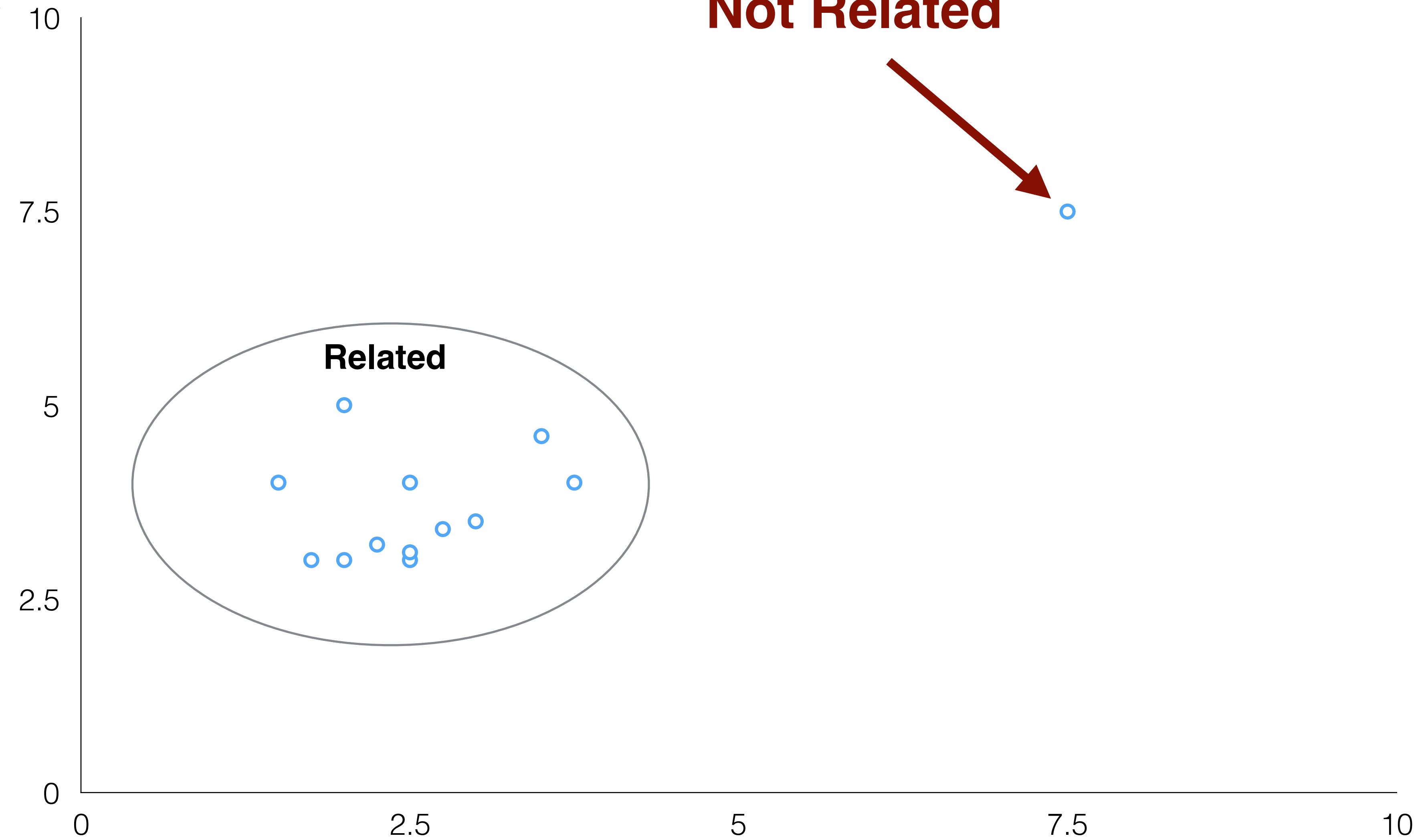
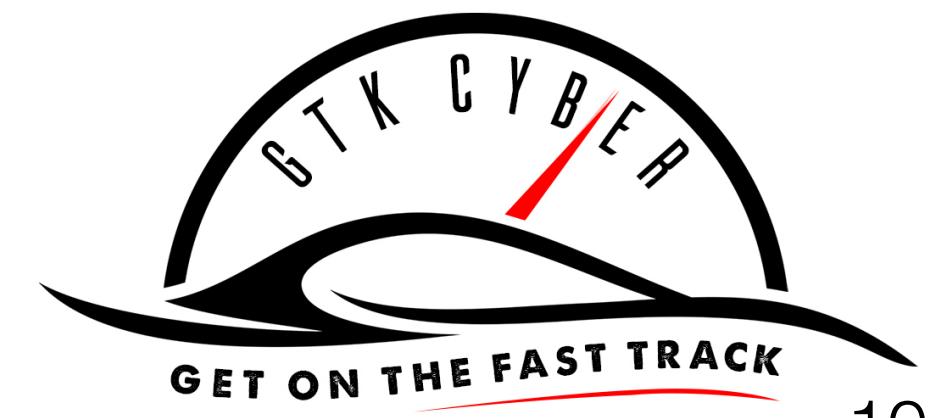


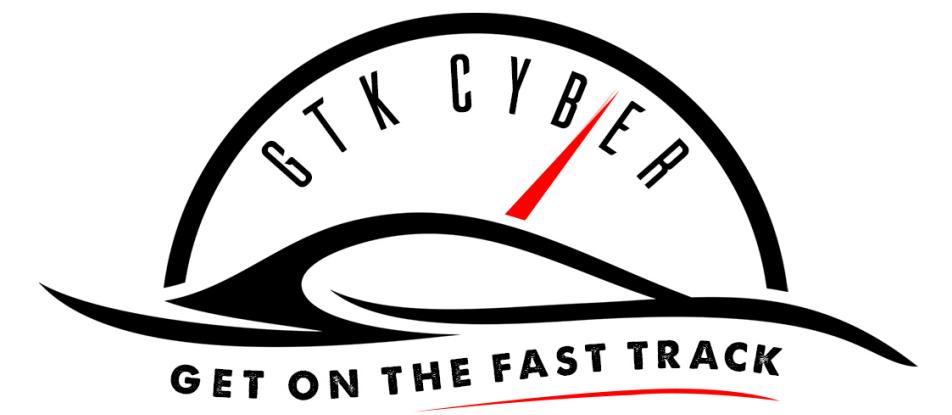


Visual Encodings

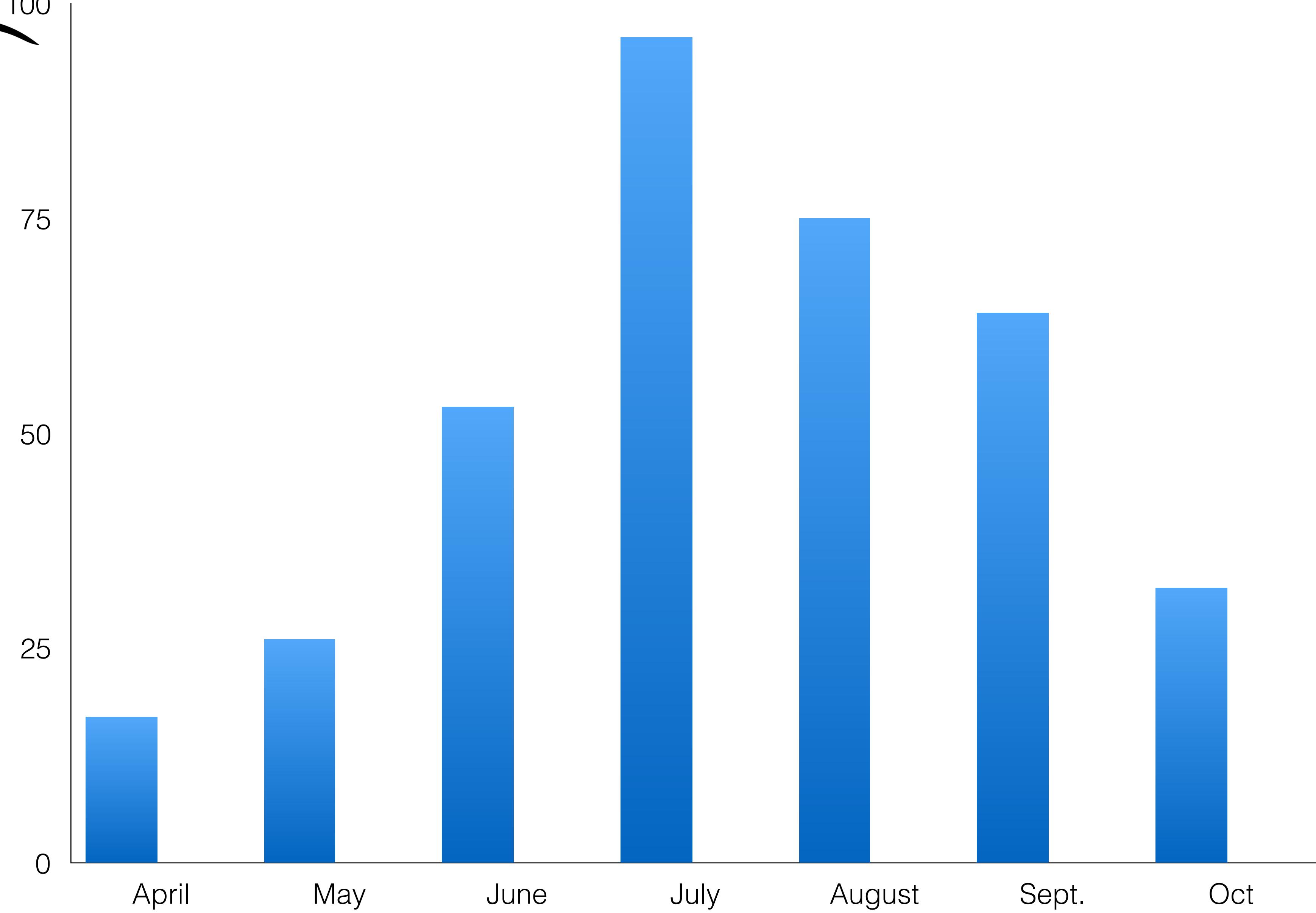
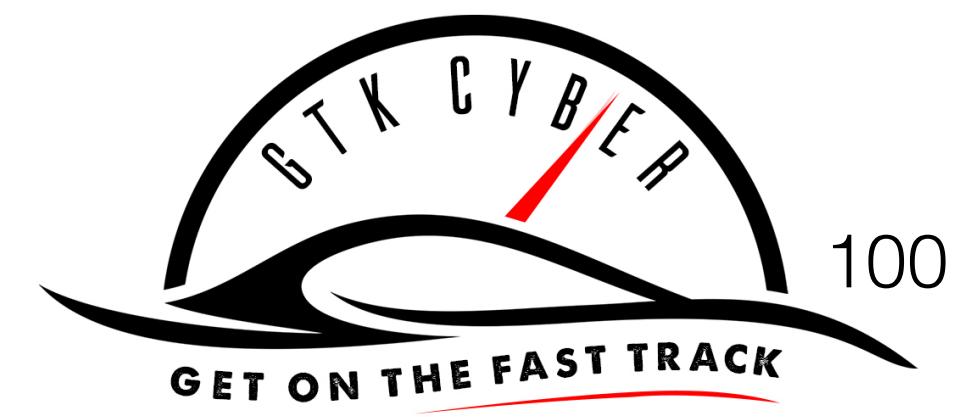


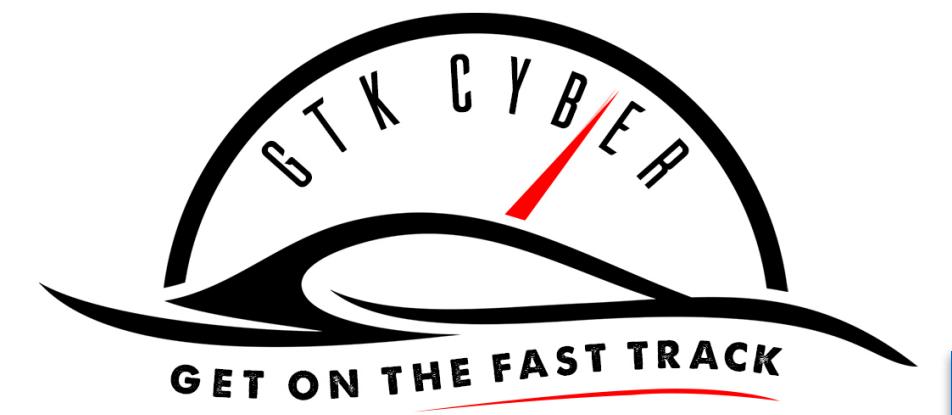
Visual Encodings: Position

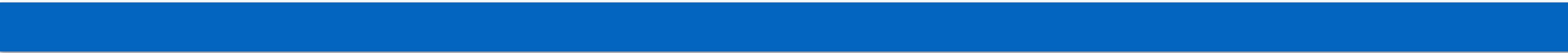
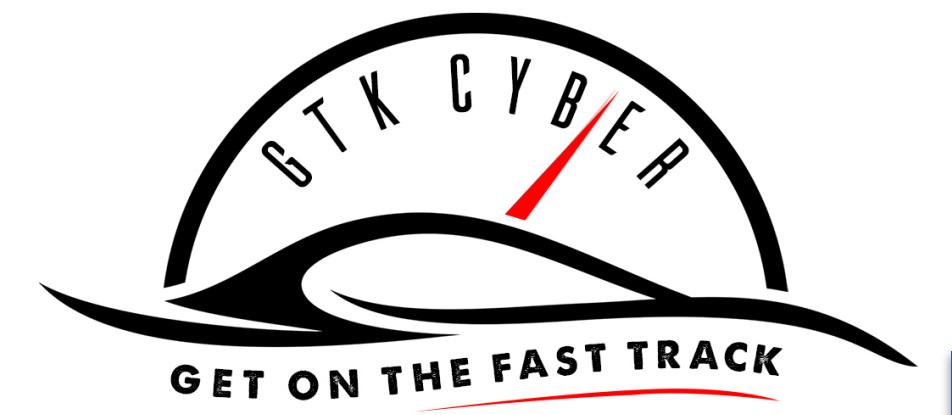


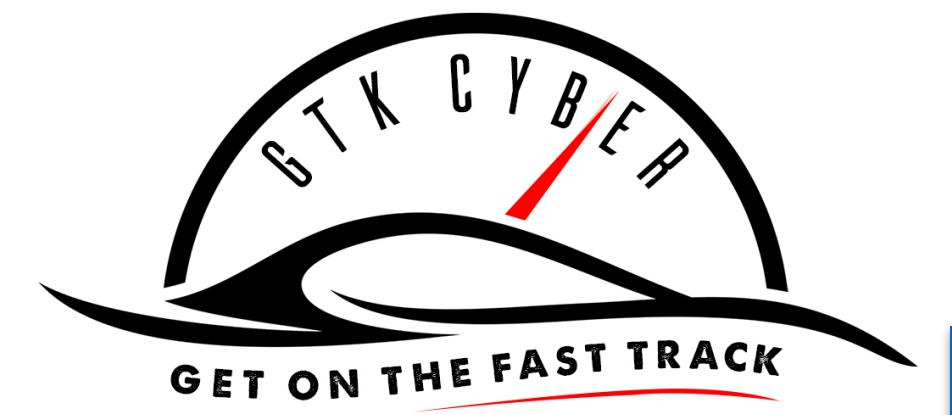


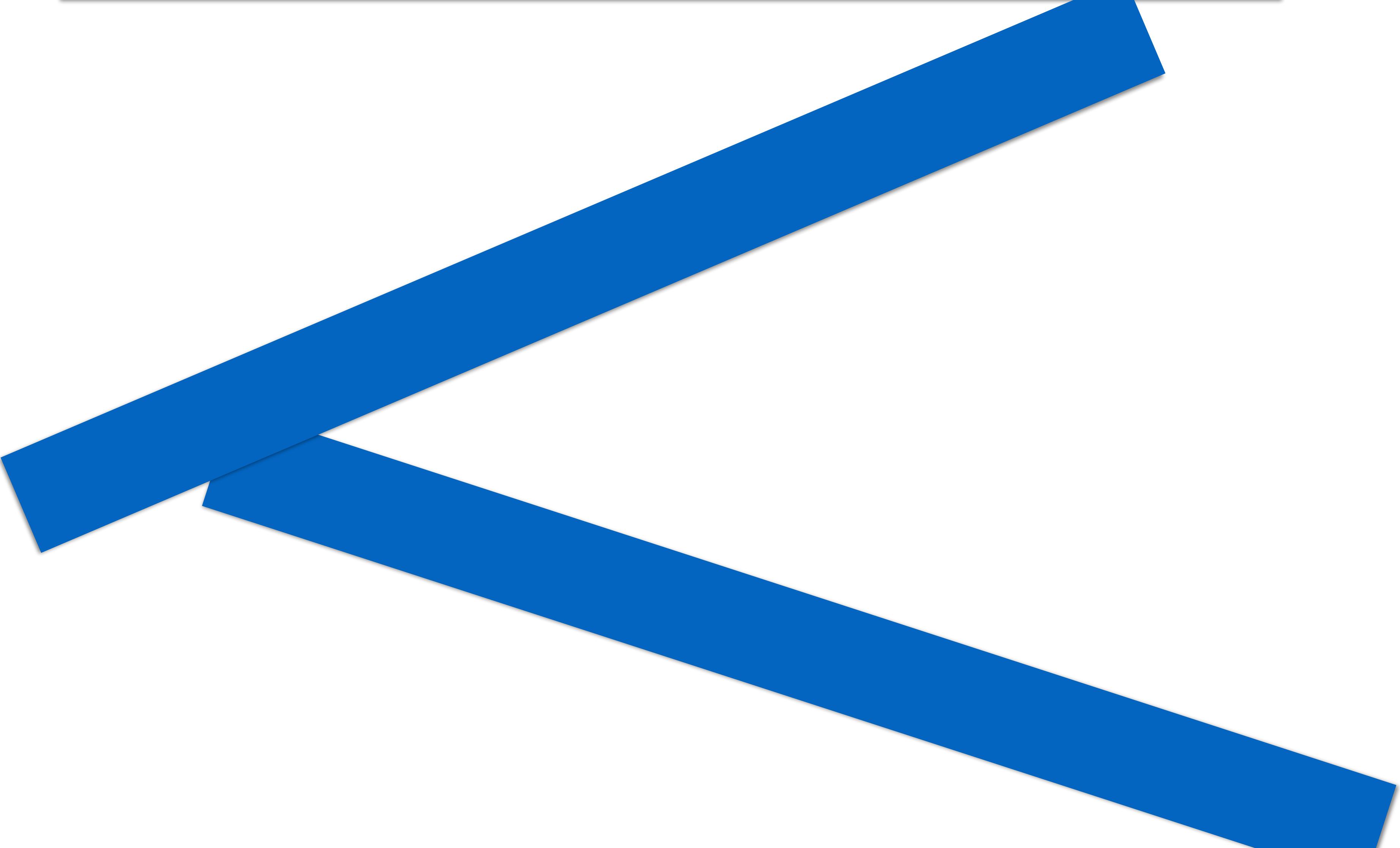
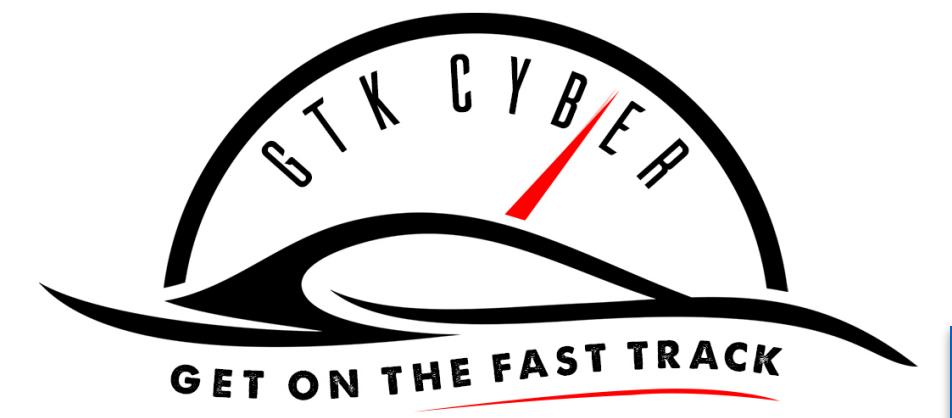
Visual Encodings: Length

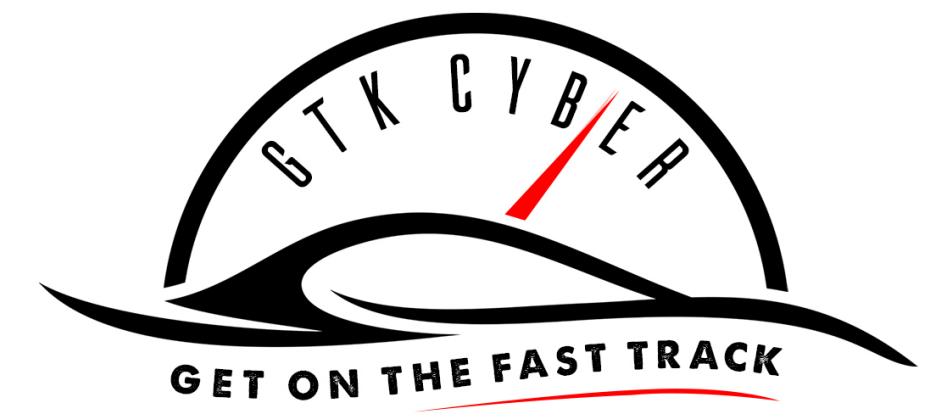


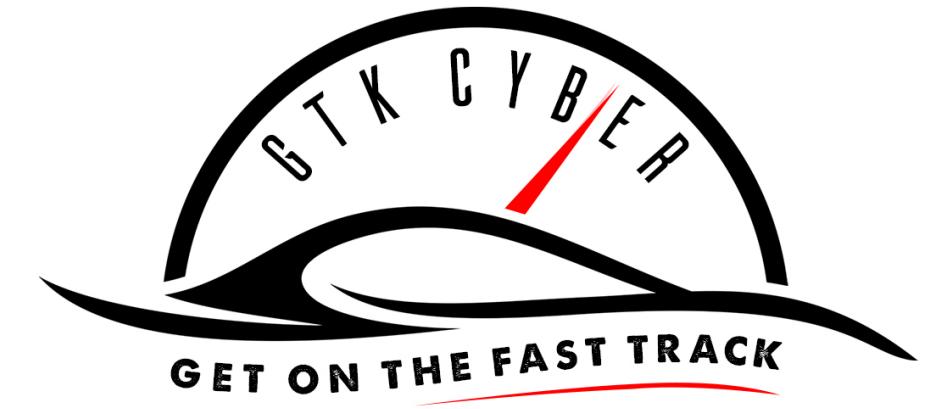




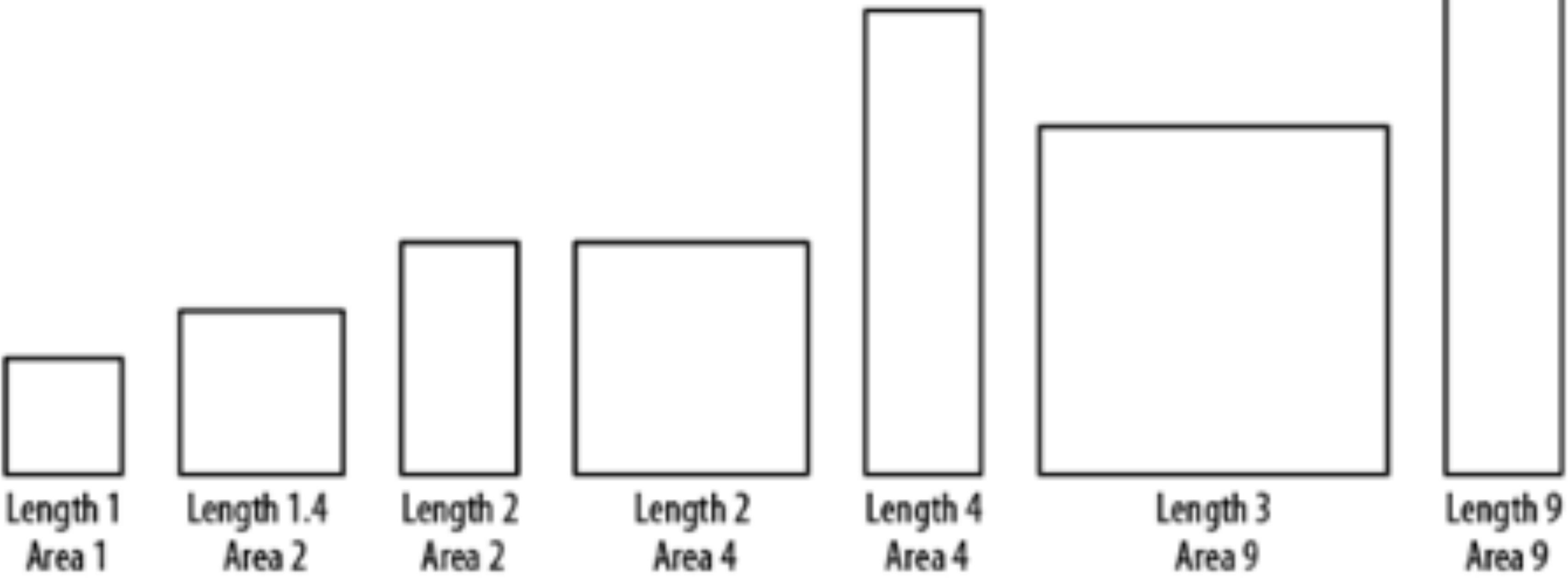


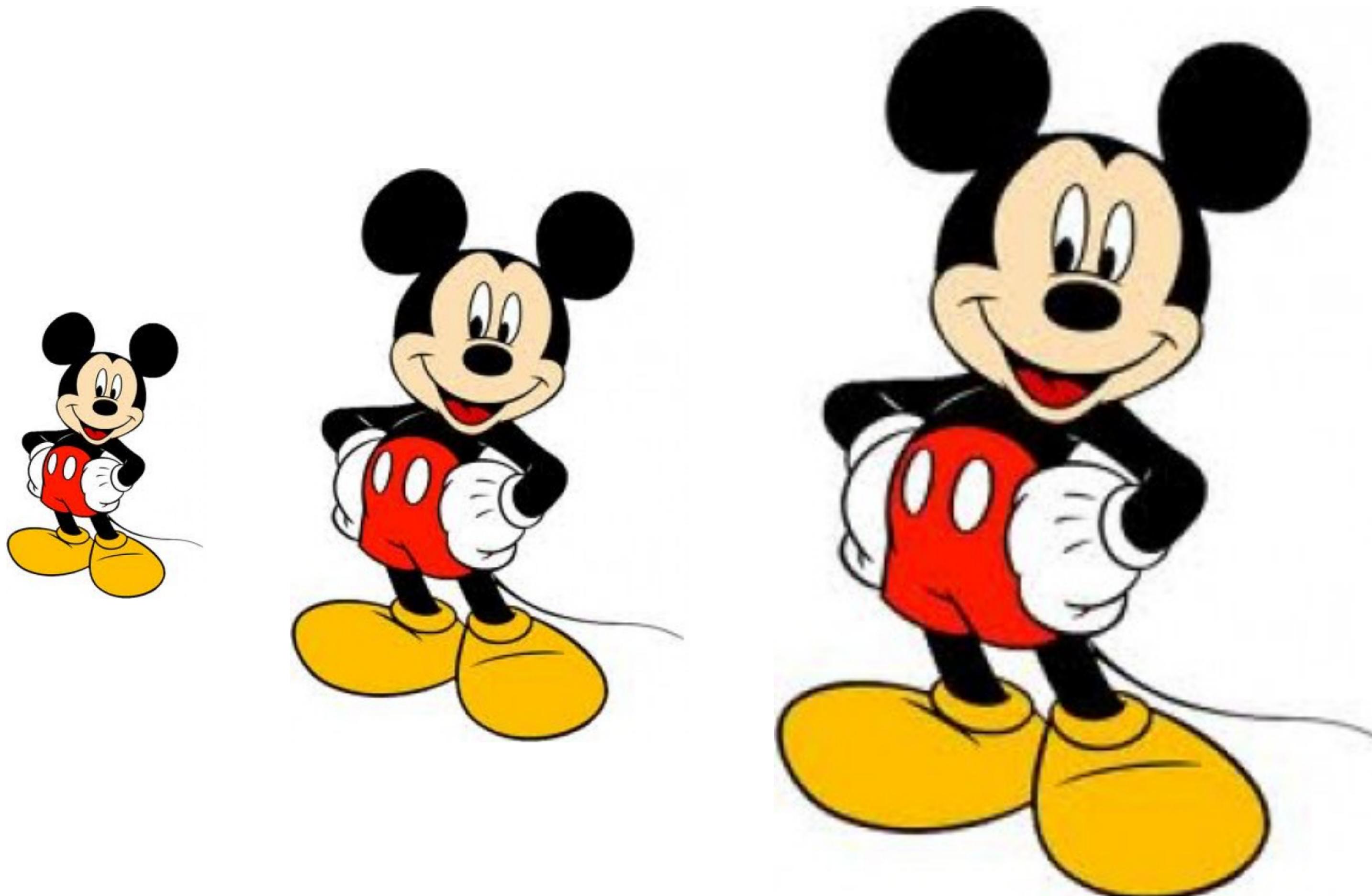
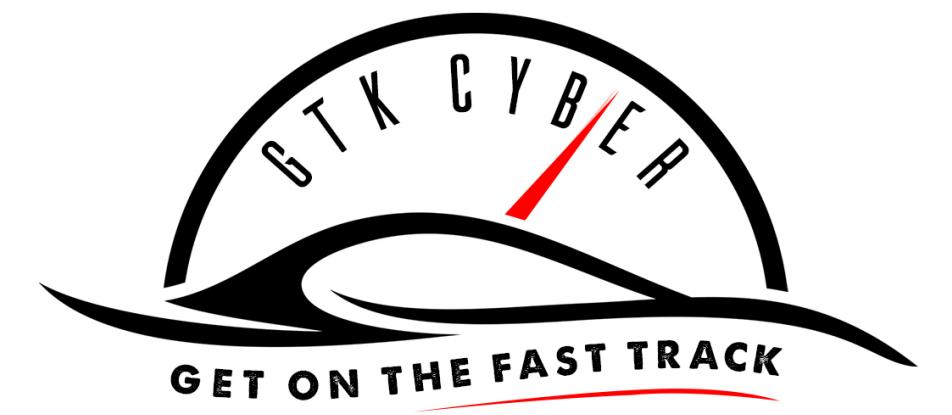


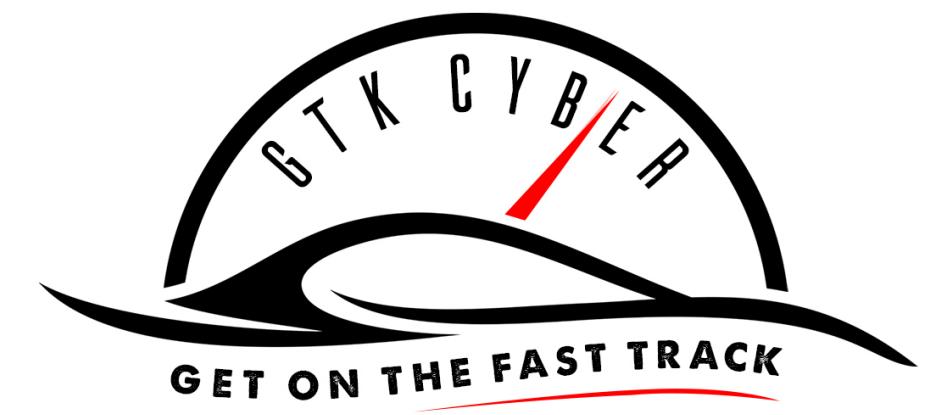




Visual Encodings: **Size**



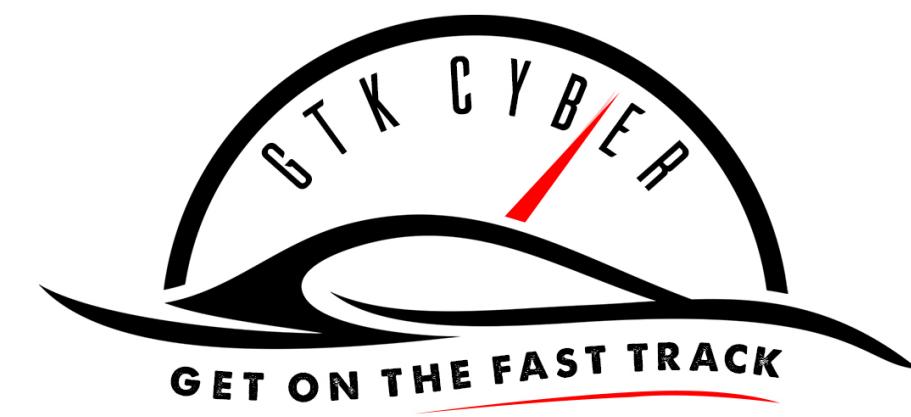




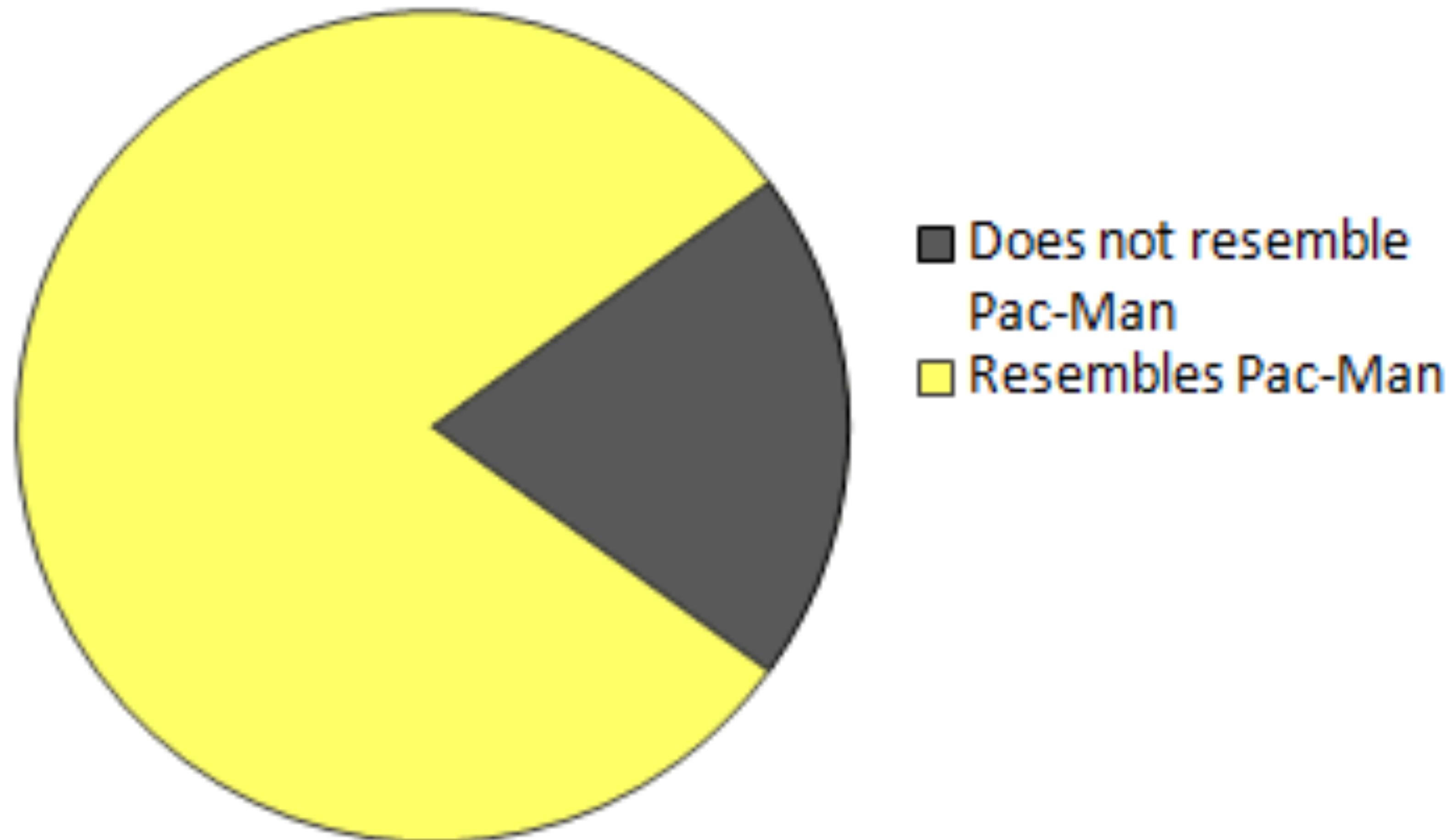


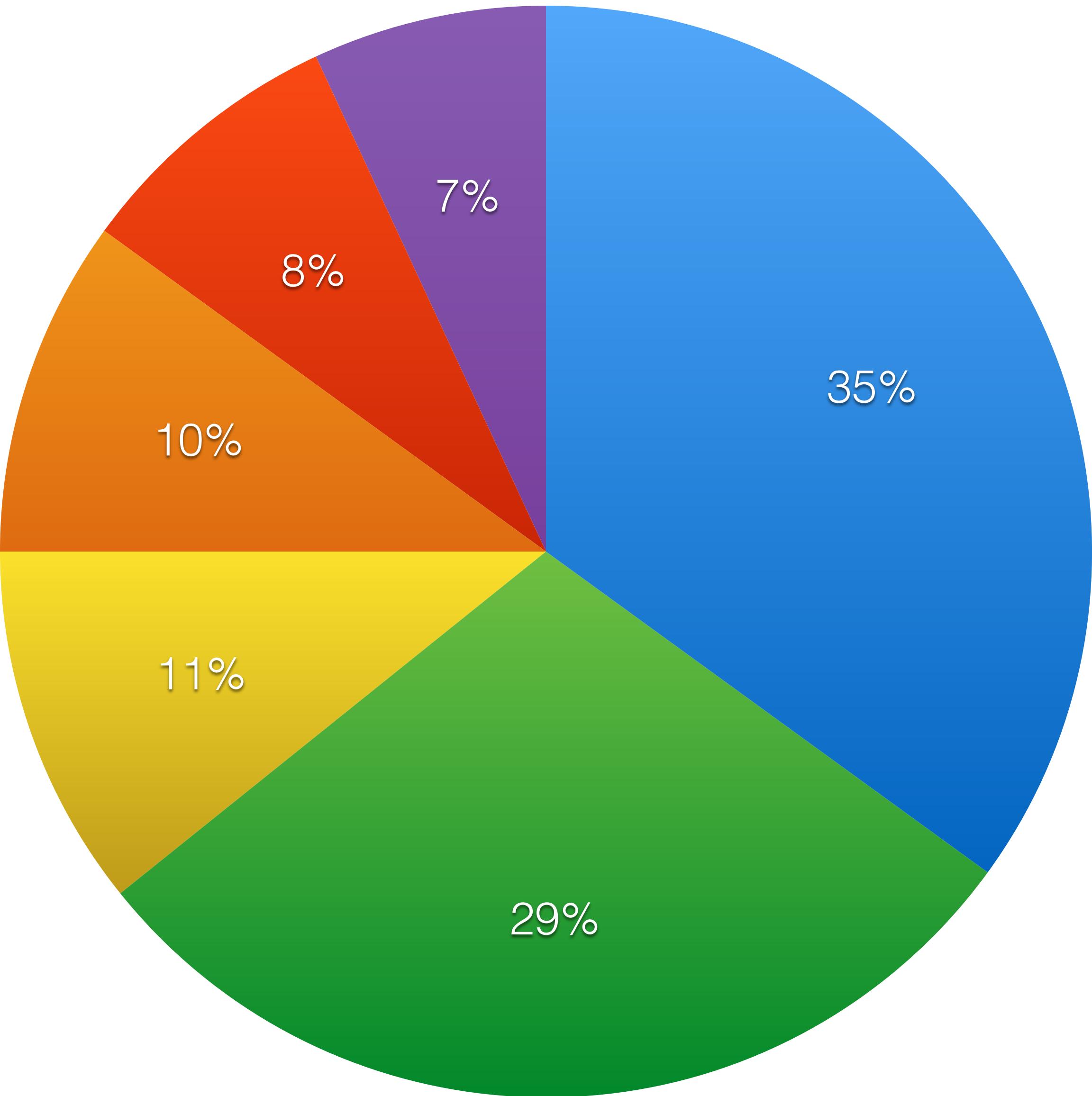
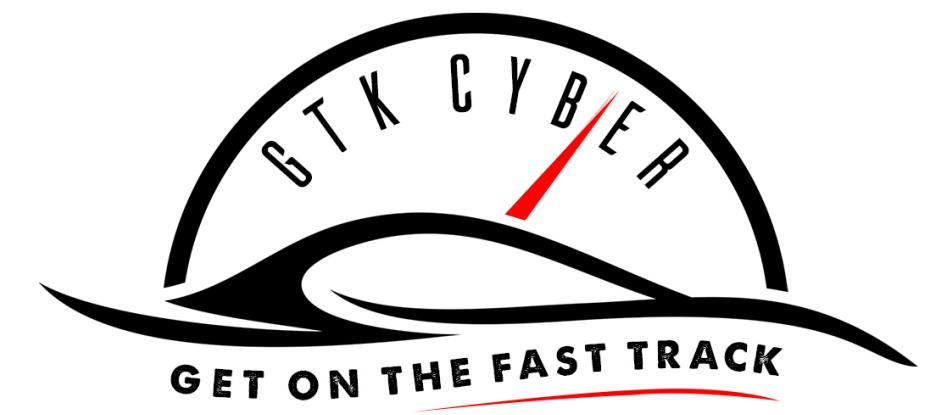
A brief interlude: Why
never to use a Pie Chart

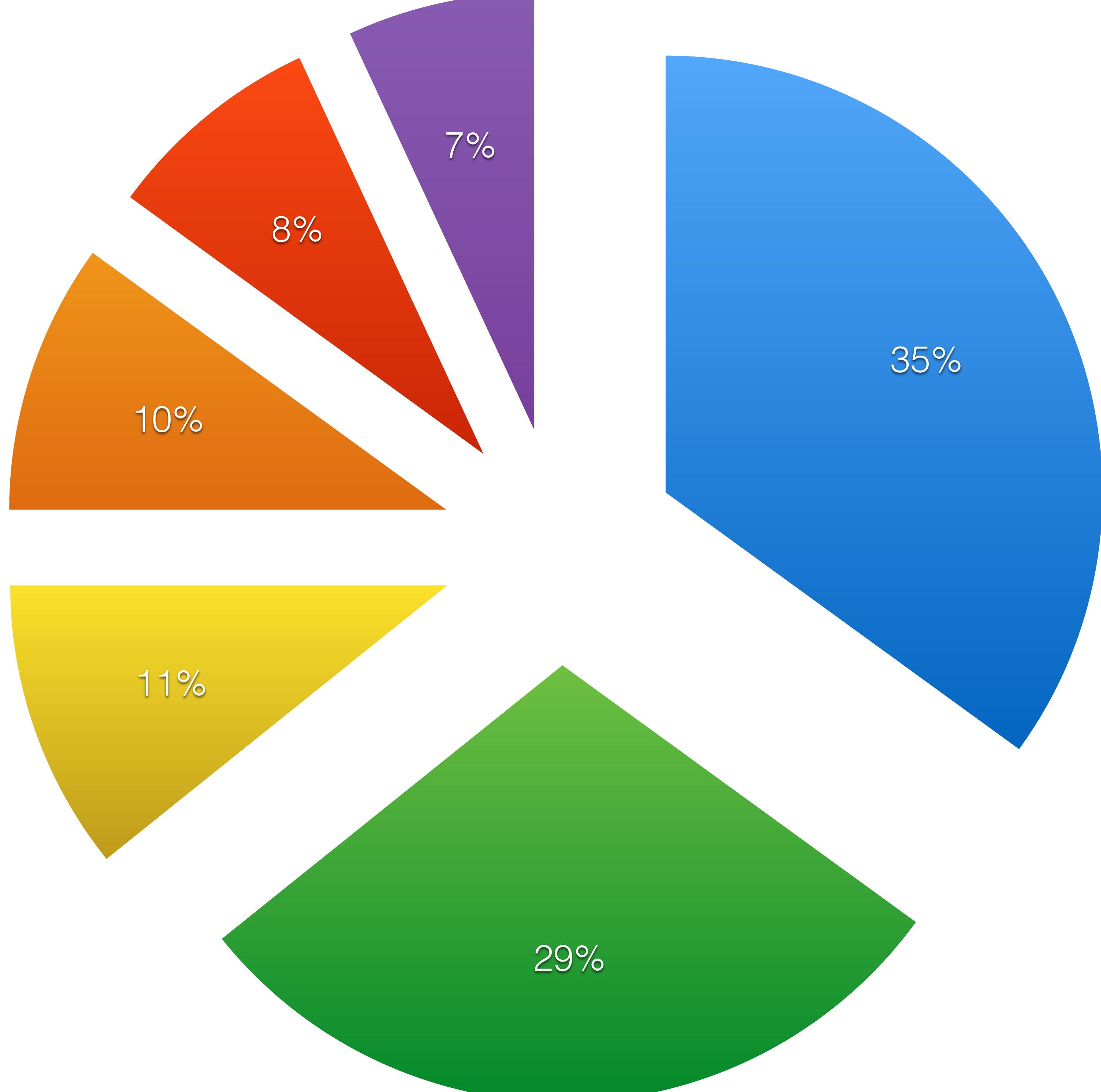
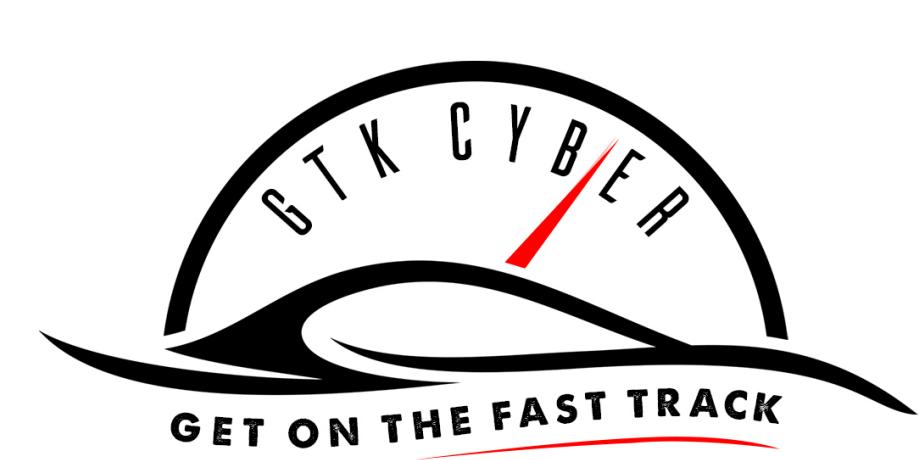


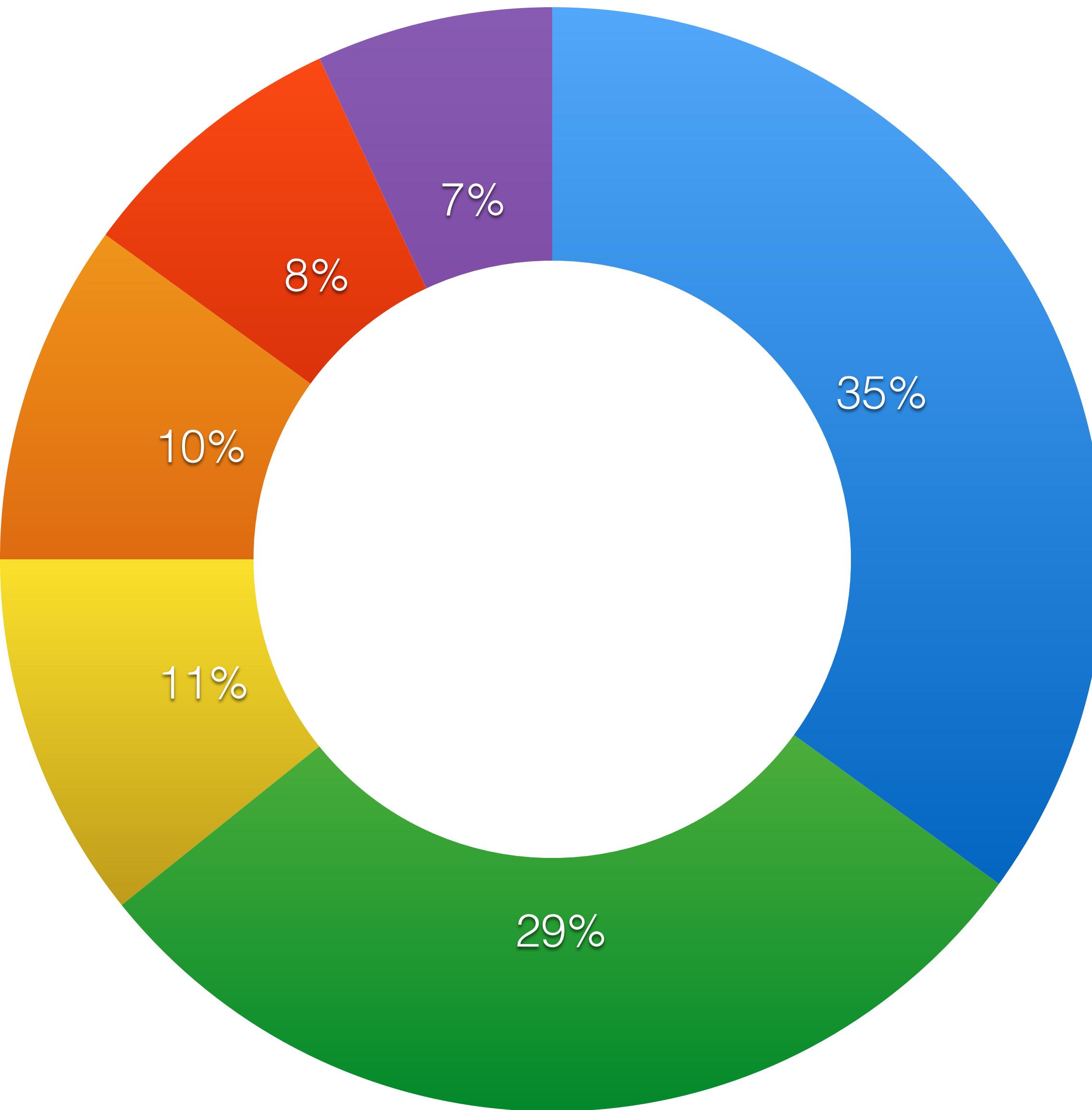
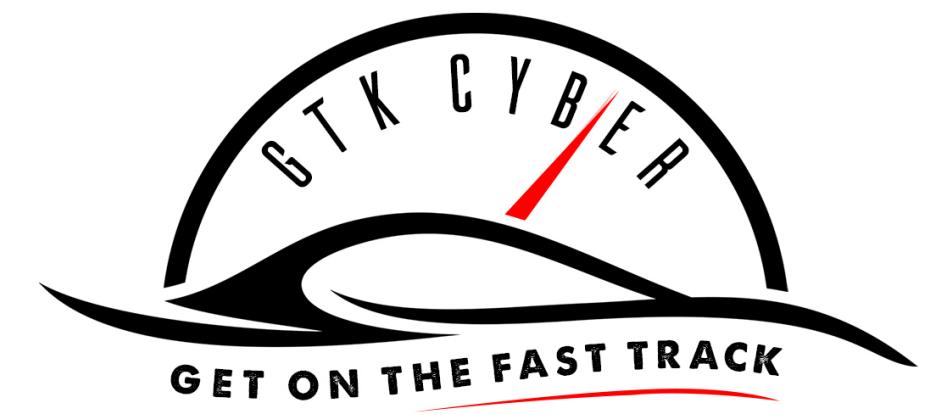


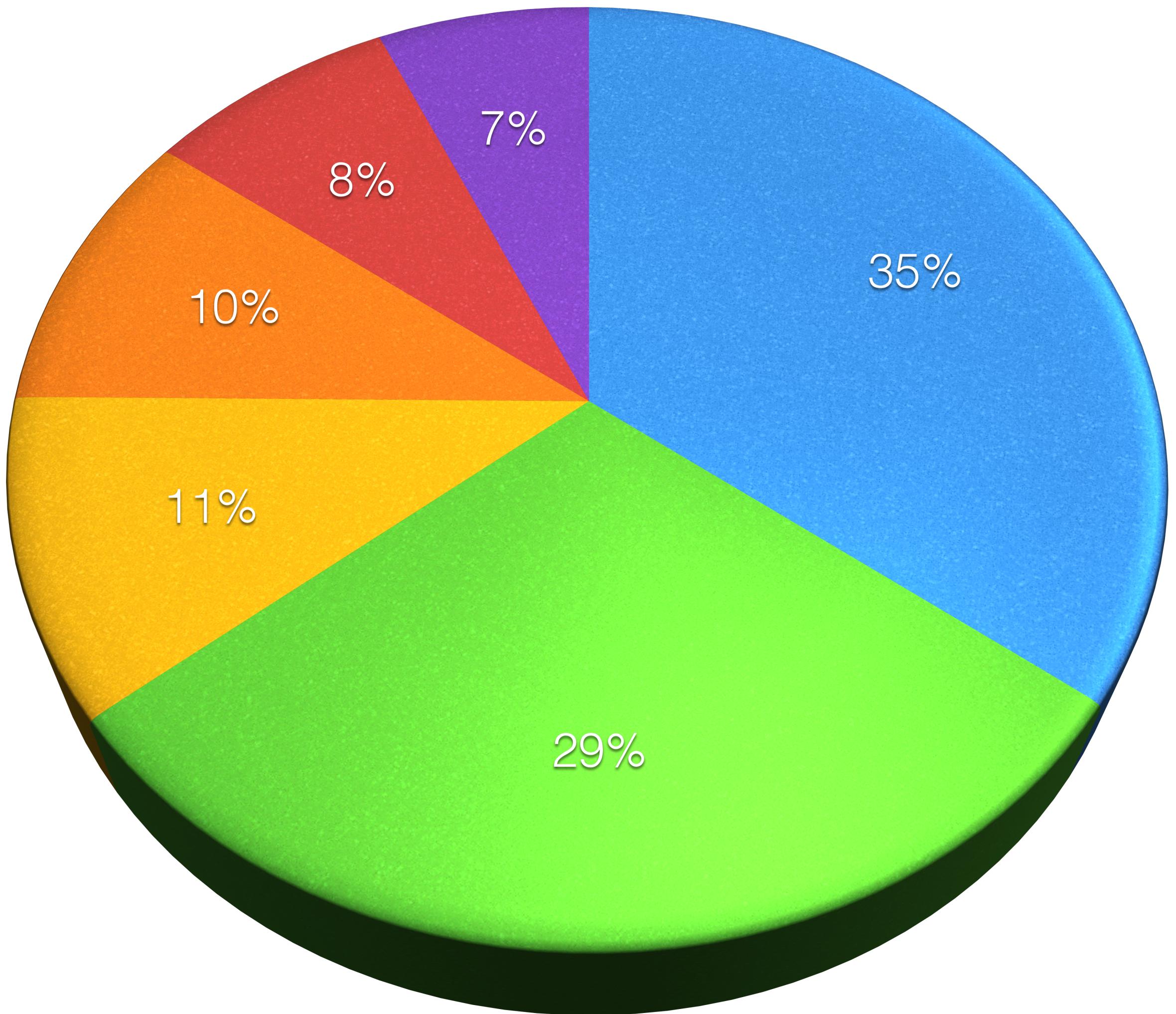
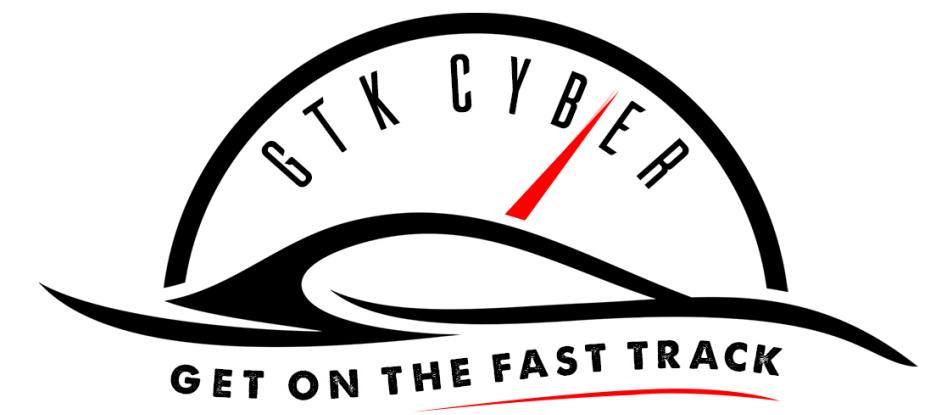
Percentage of Chart Which Resembles Pac-Man

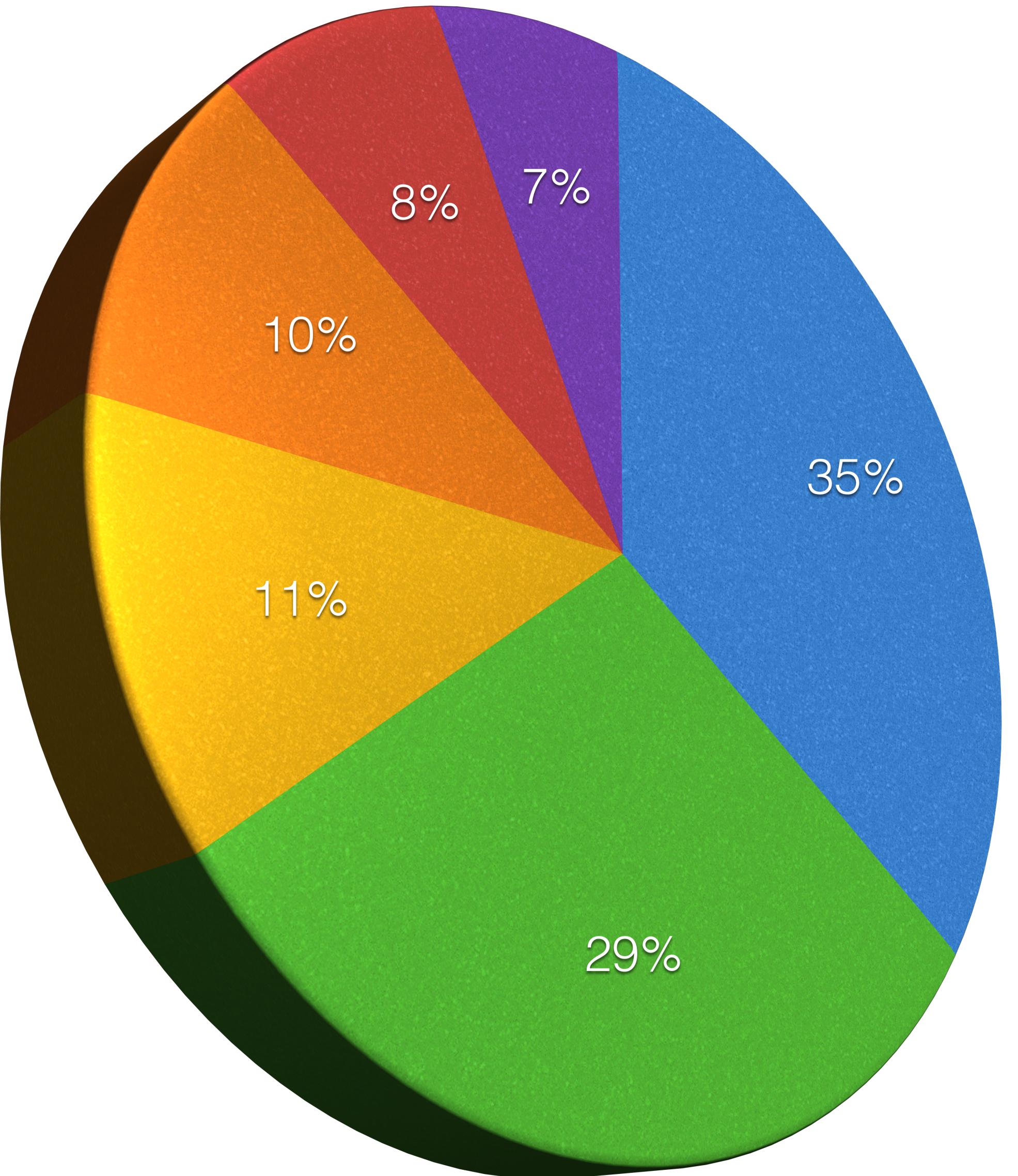
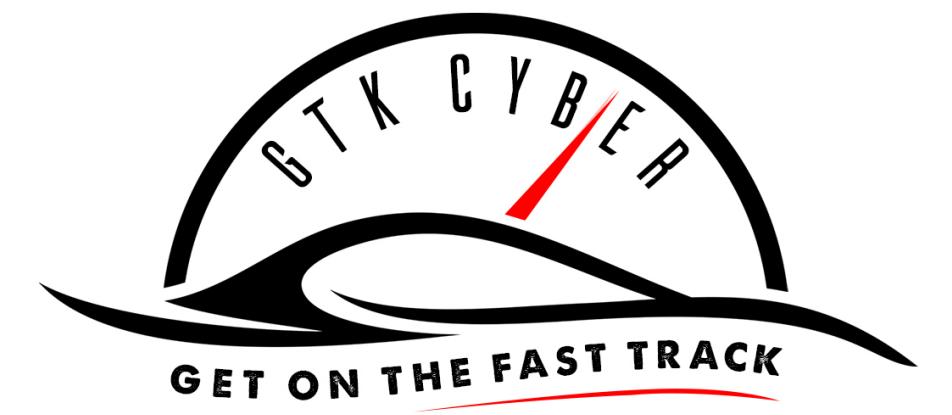


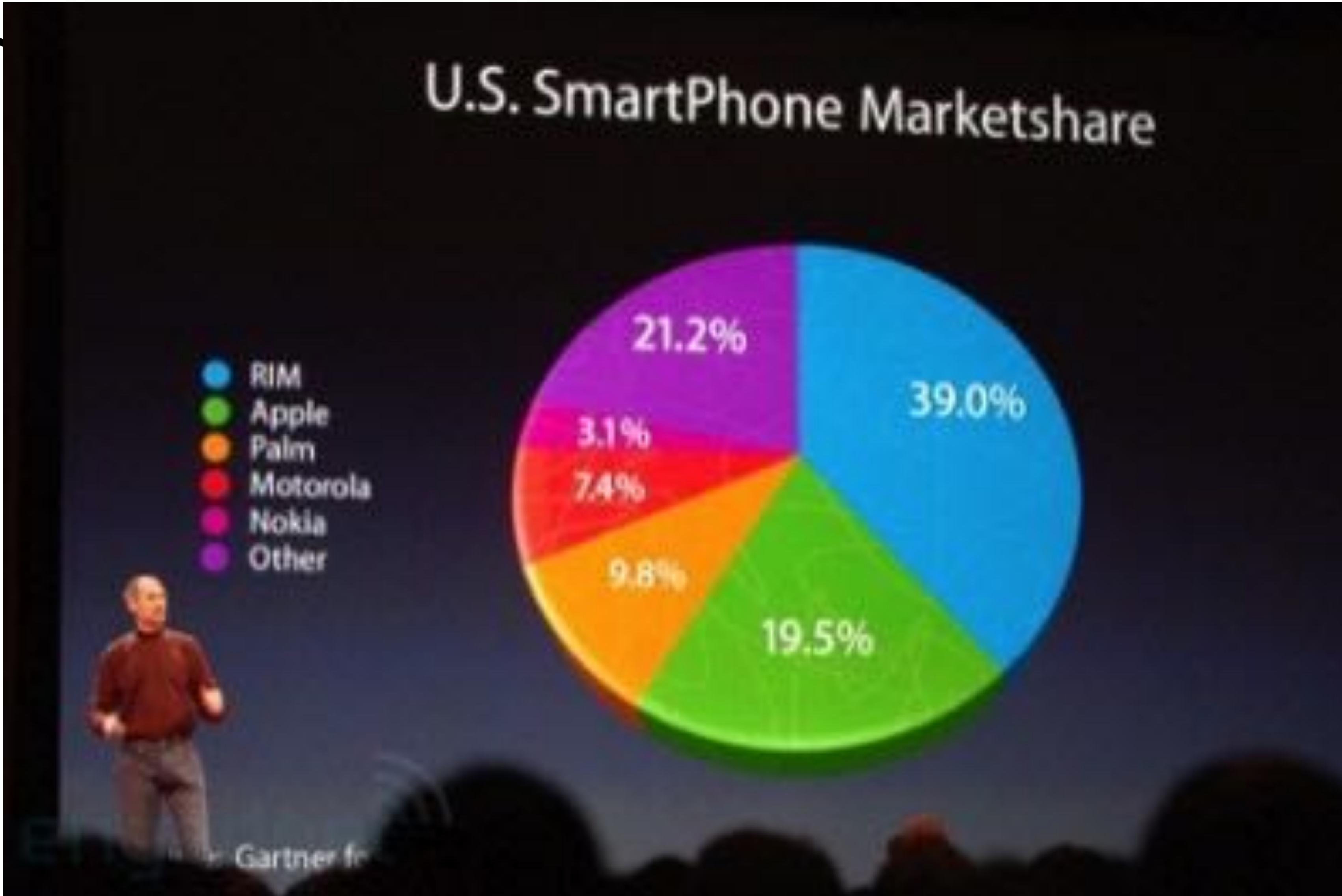




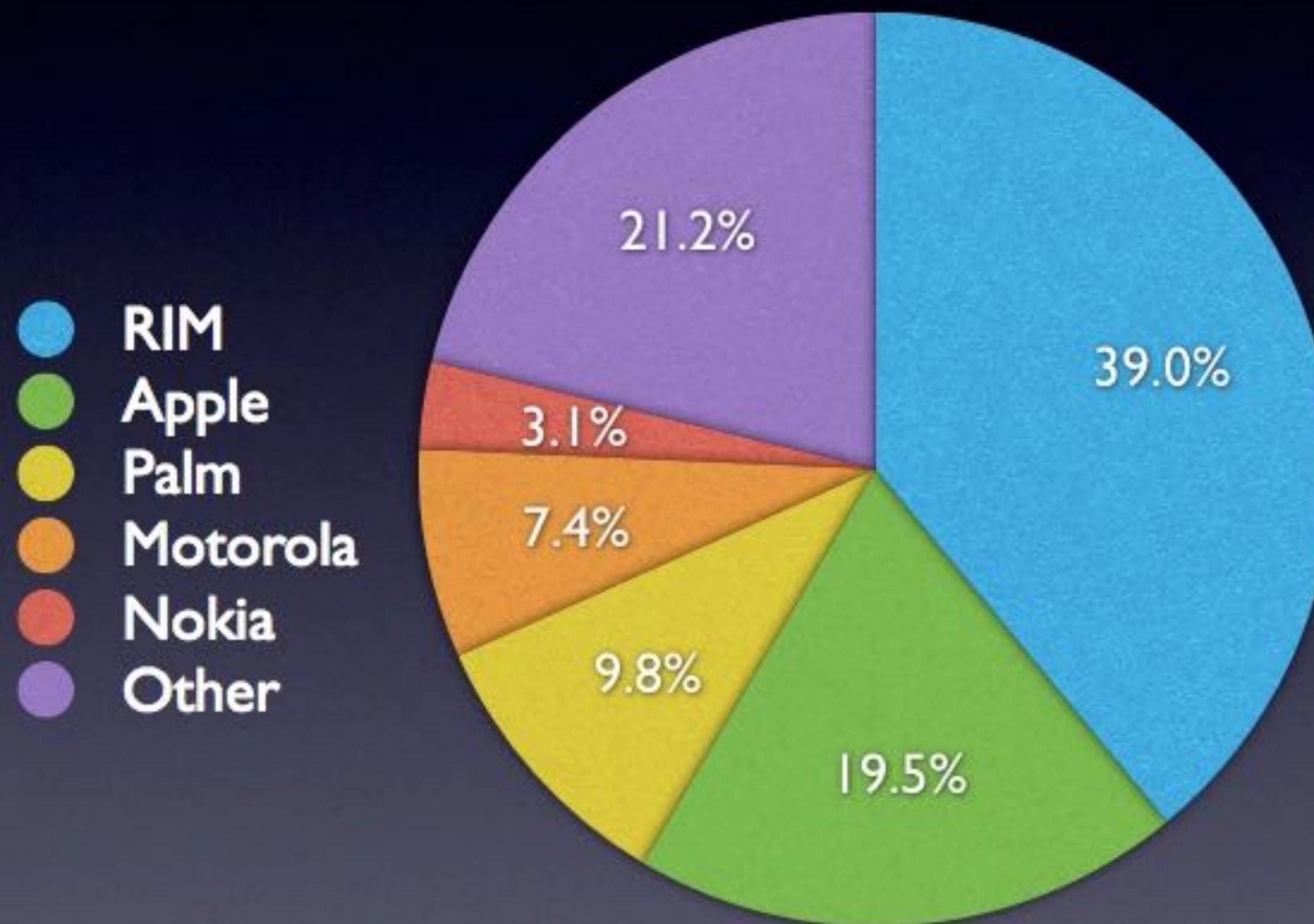


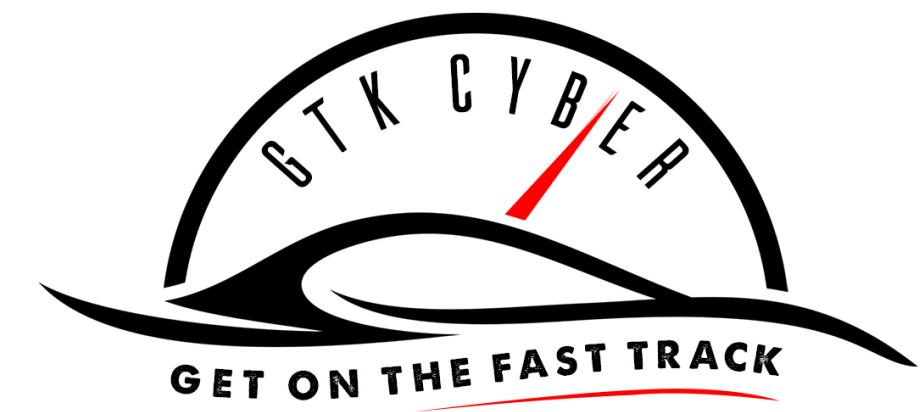






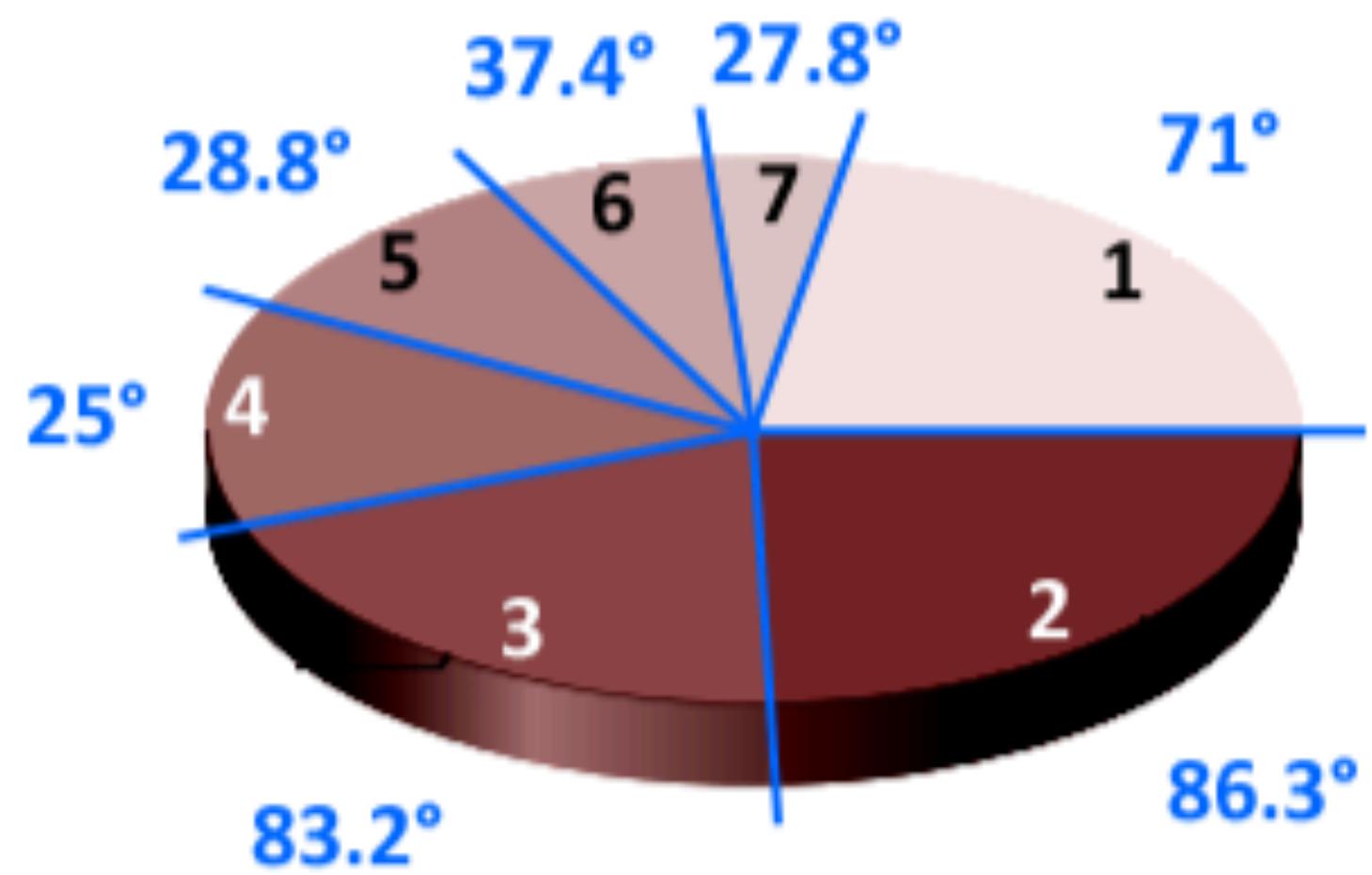
U.S. SmartPhone Marketshare



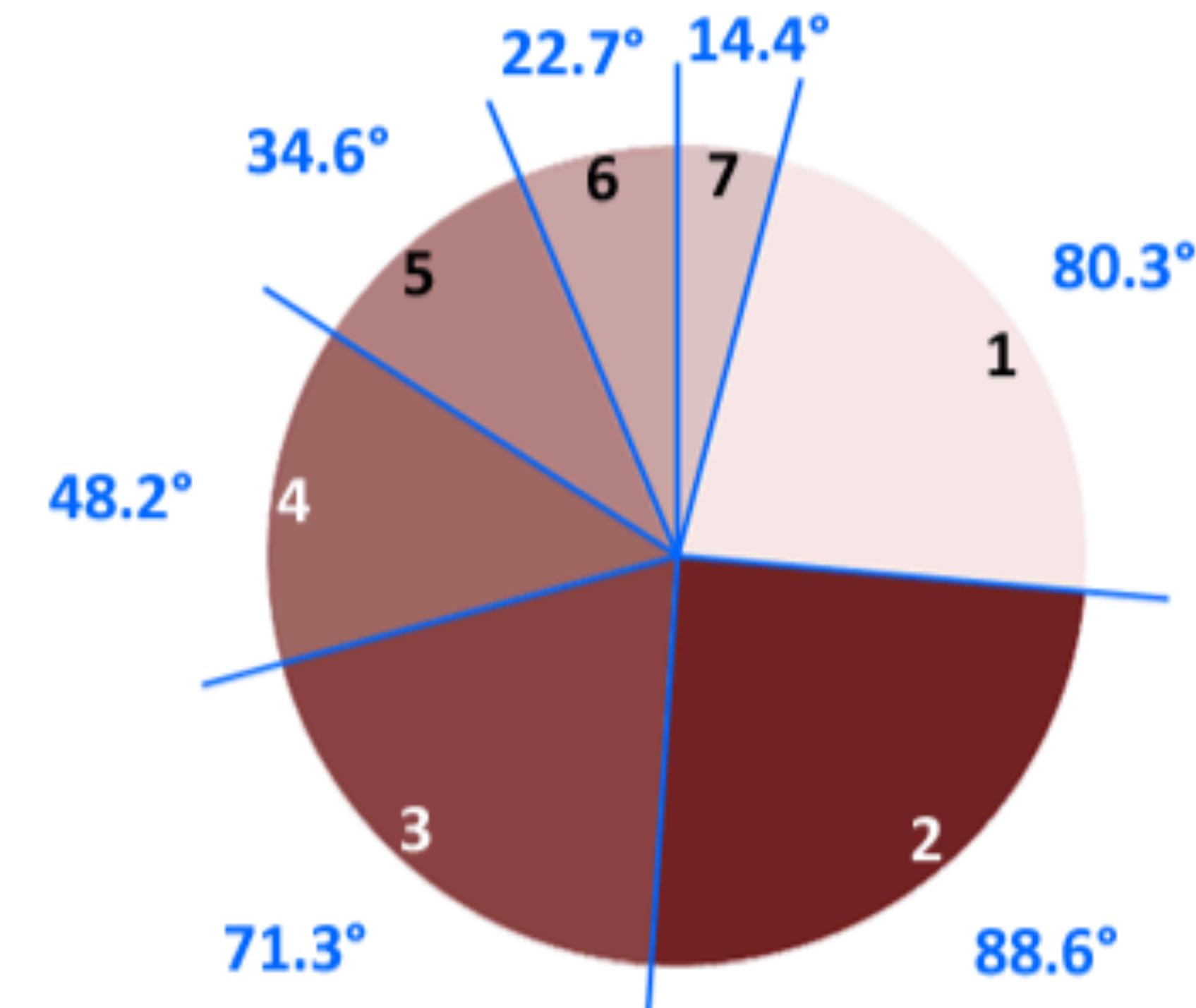


3D Pie Charts are BAD!

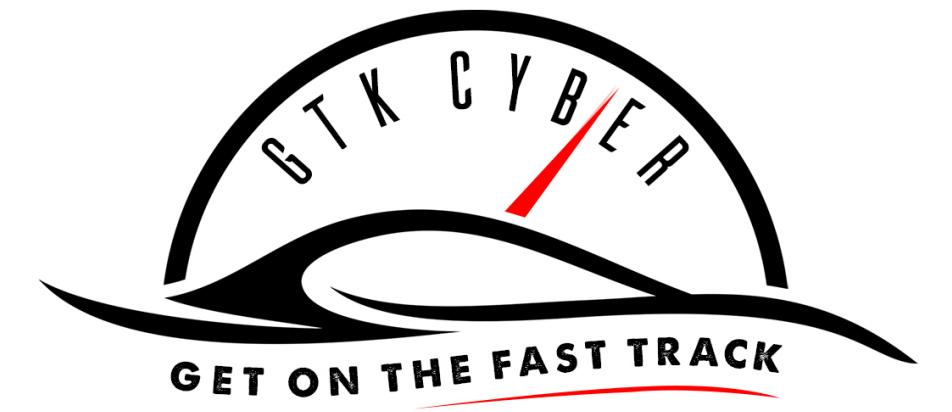
3D Pie Charts Distort Angles



Angles on the original pie chart



Angles on a non-3D pie chart

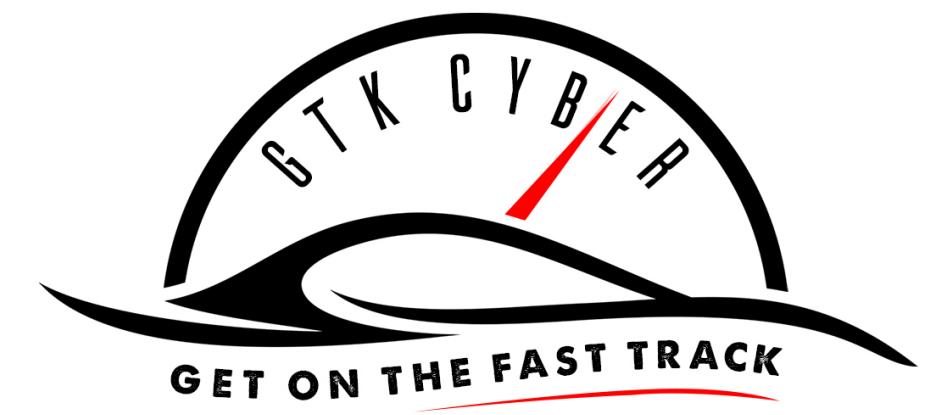


Visual Encodings: Color

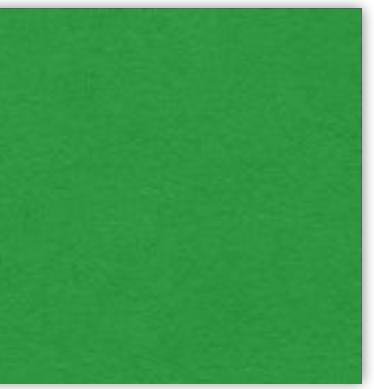


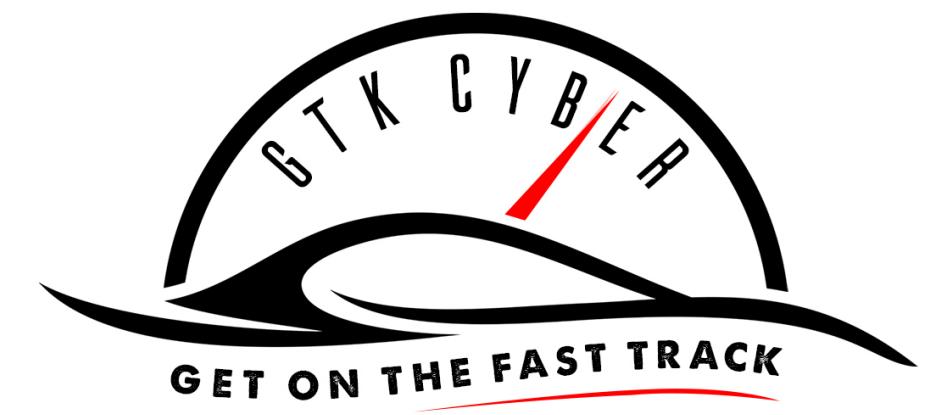
Visual Encodings: Color

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

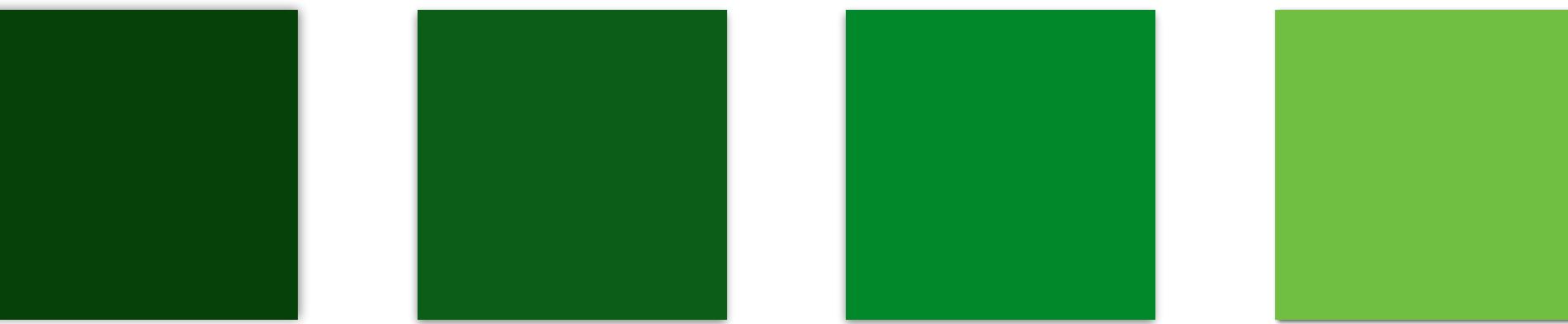


Hue

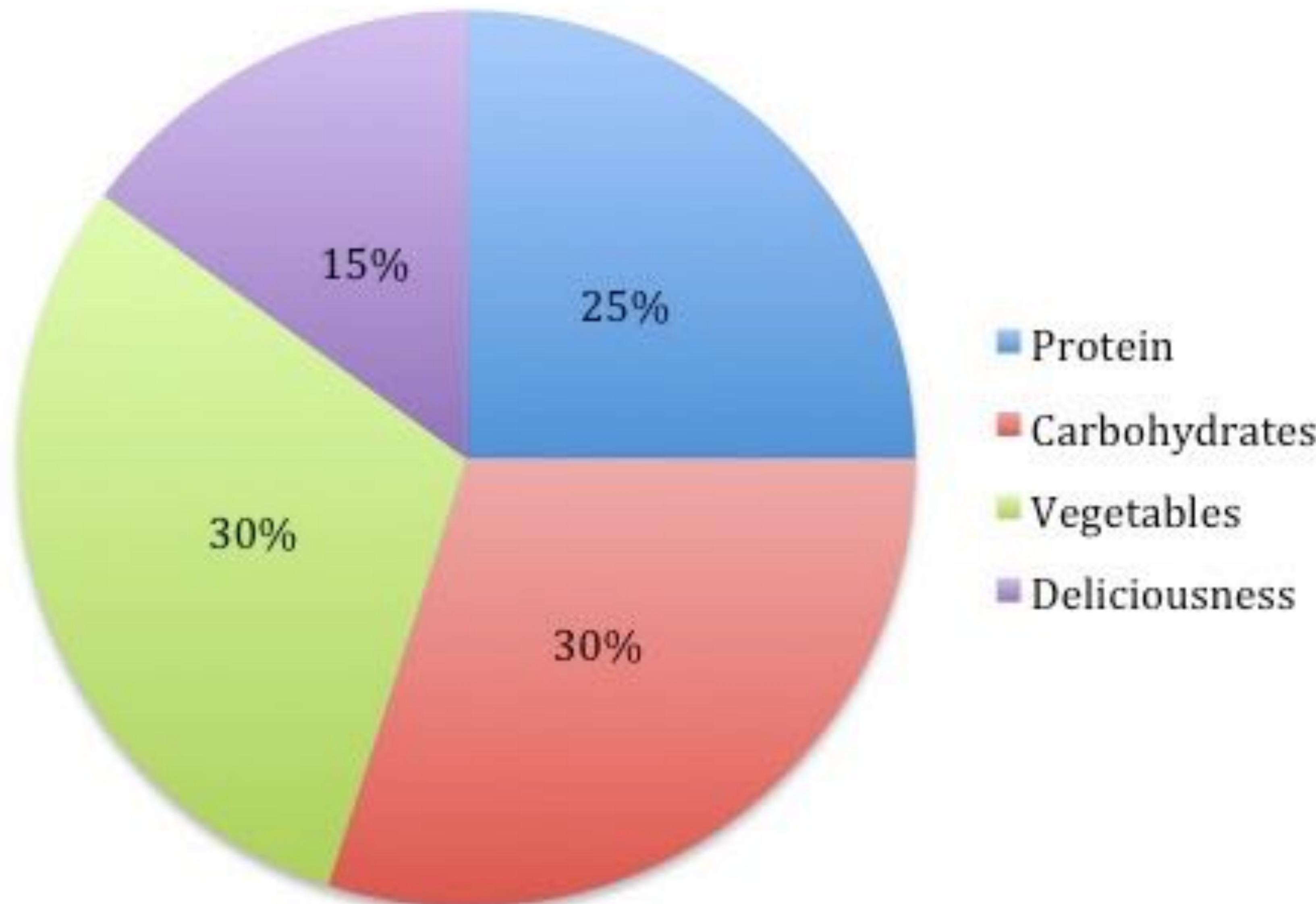




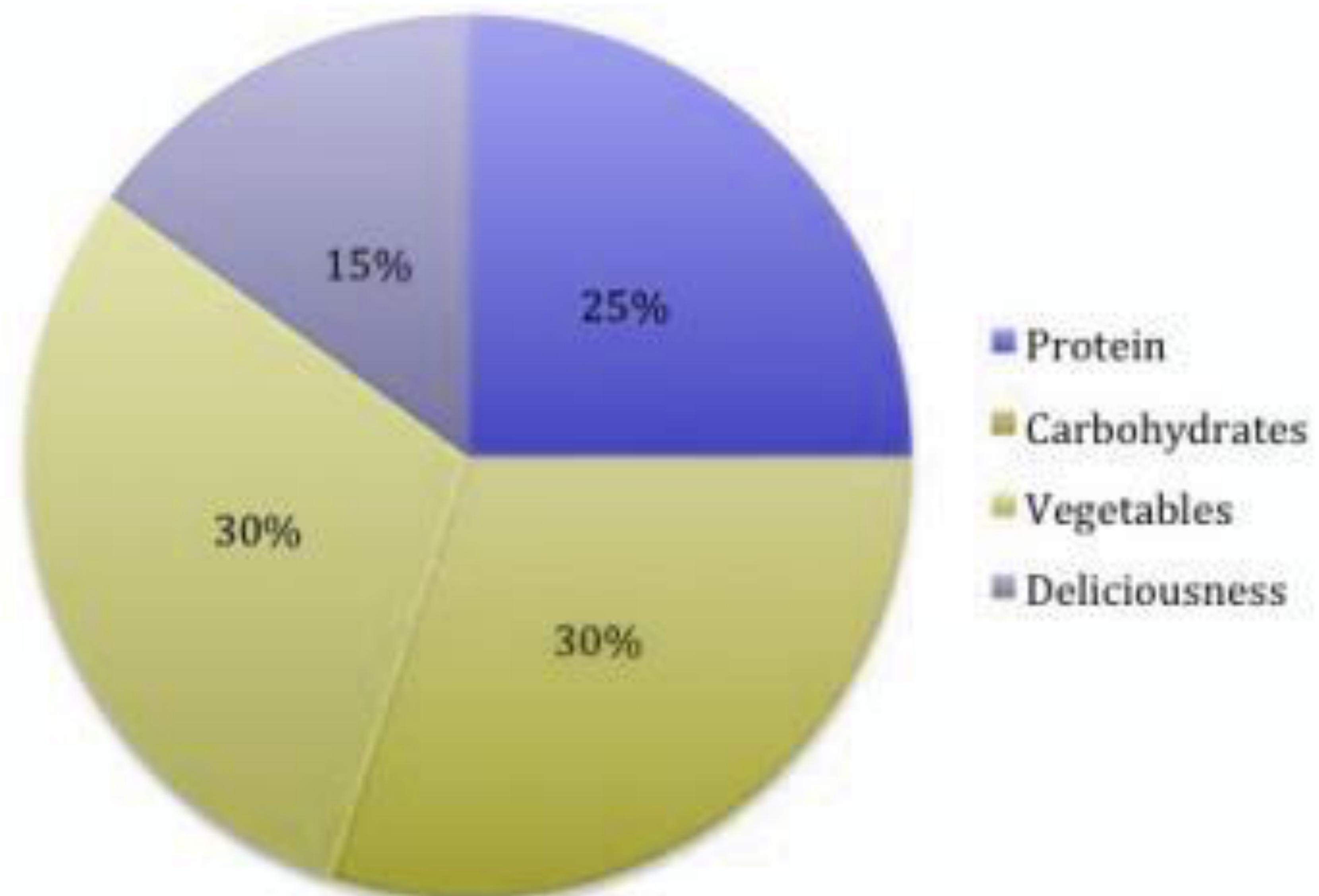
Saturation



A Healthy Meal



A Healthy Meal





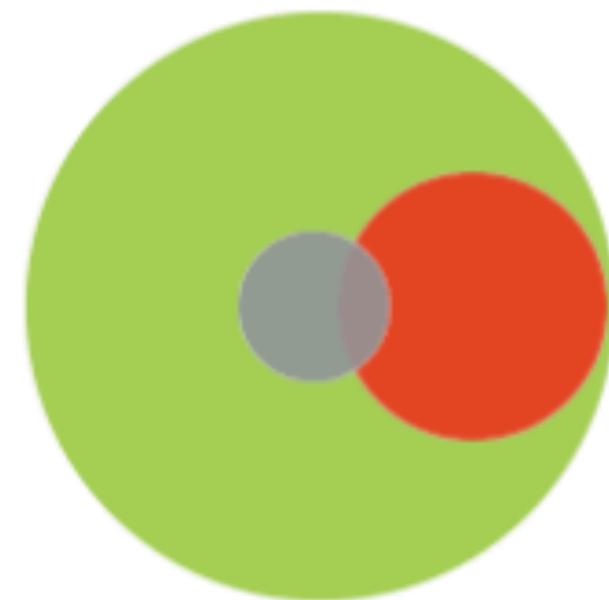
elastica



Dashboard

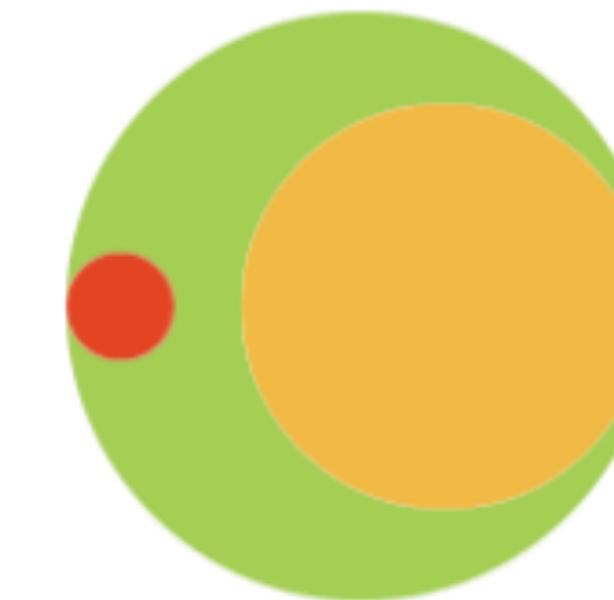
Users

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

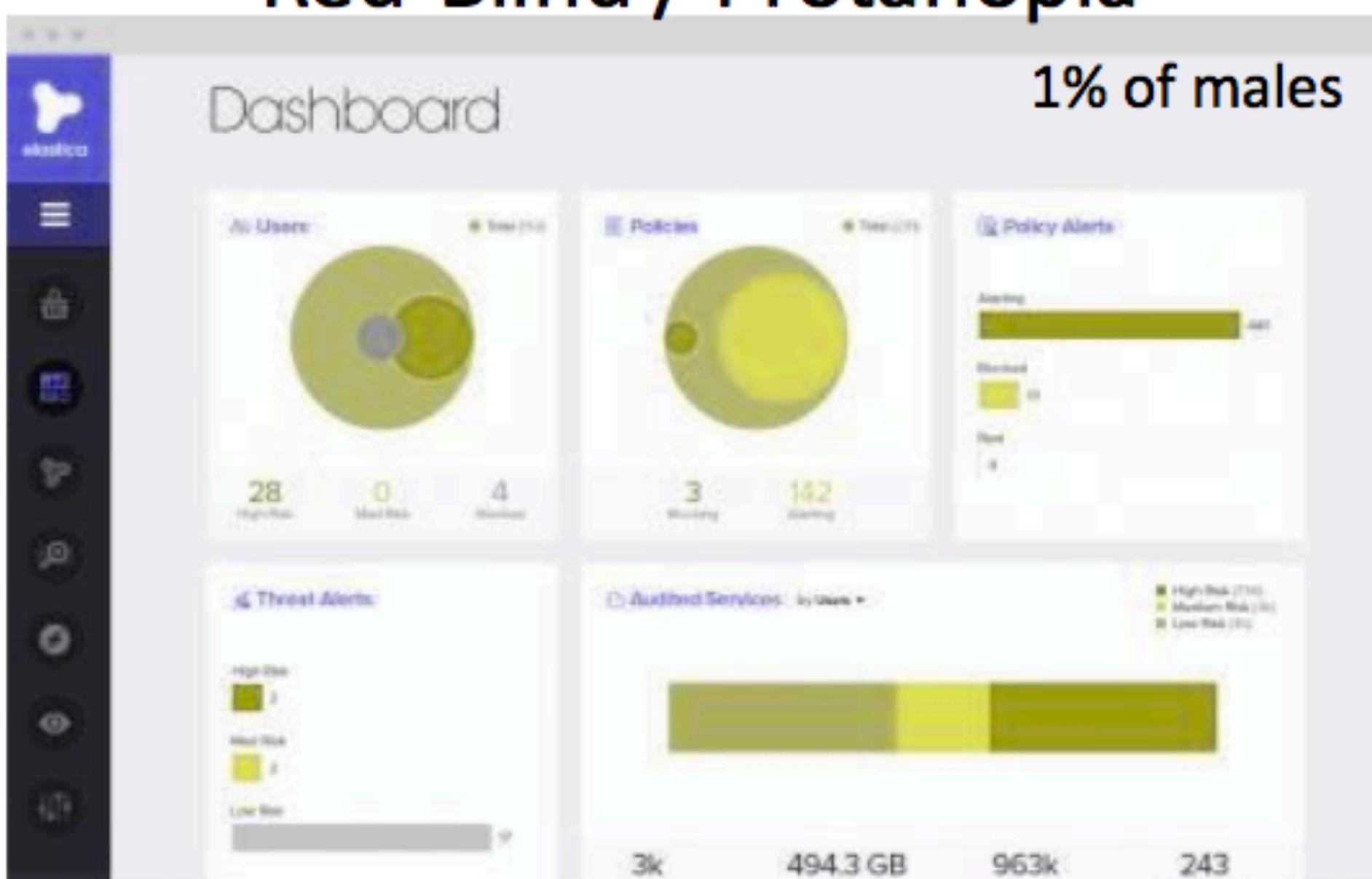
Original



Green-Blind / Deutanopia

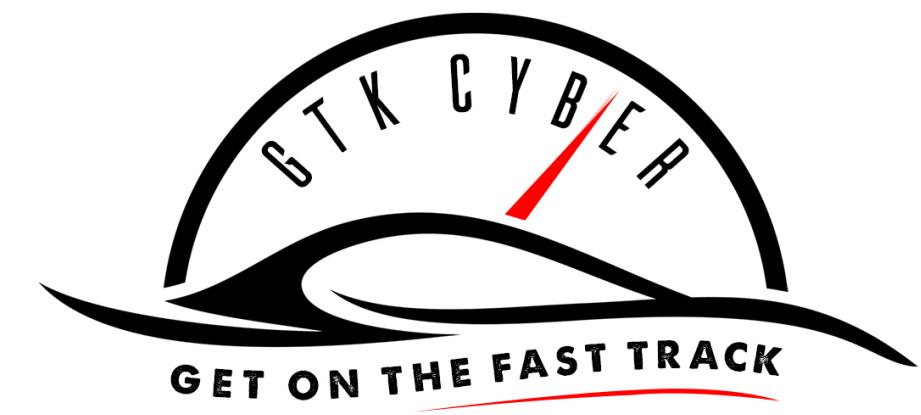


Red-Blind / Protanopia

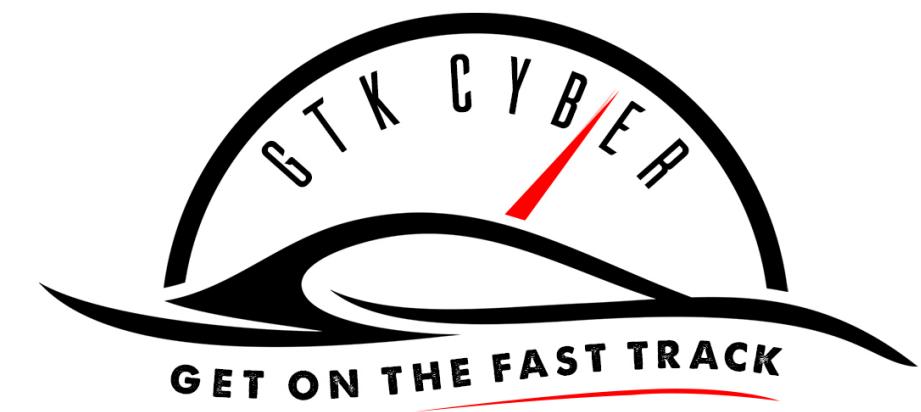


Blue-Blind / Tritanopia





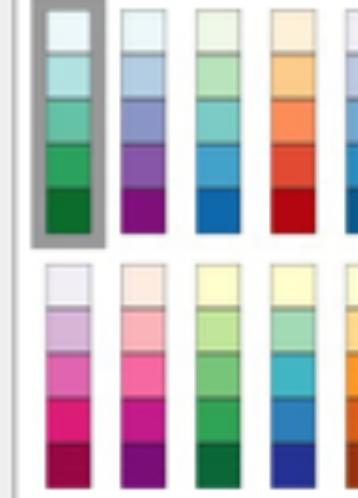
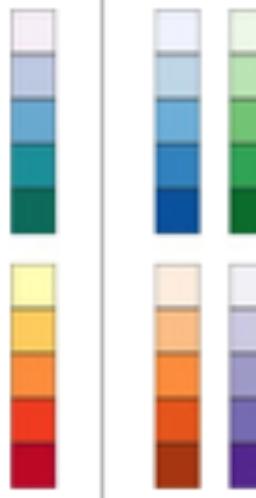
<http://www.color-blindness.com/coblis-color-blindness-simulator/>



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

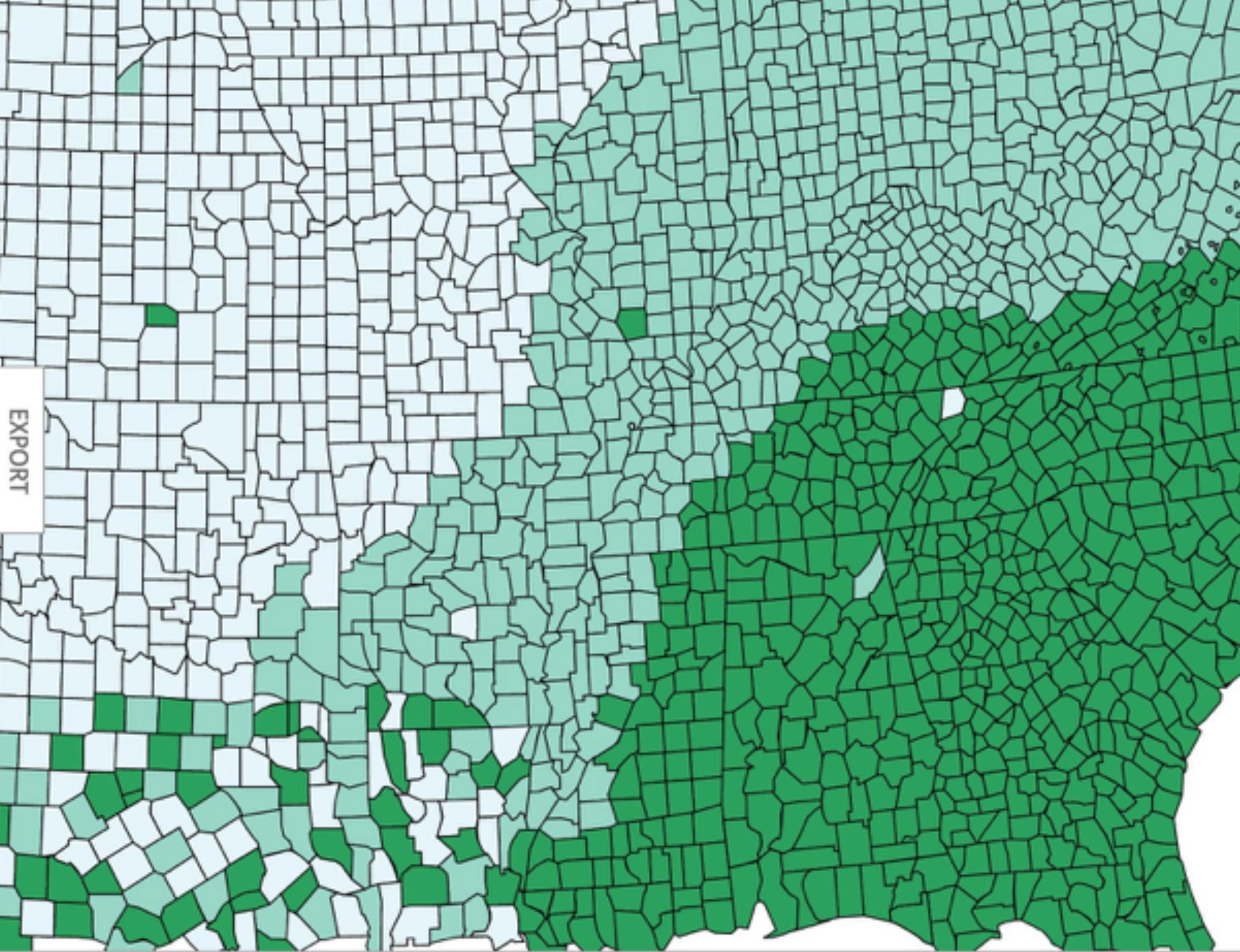
how to use | updates | downloads | credits

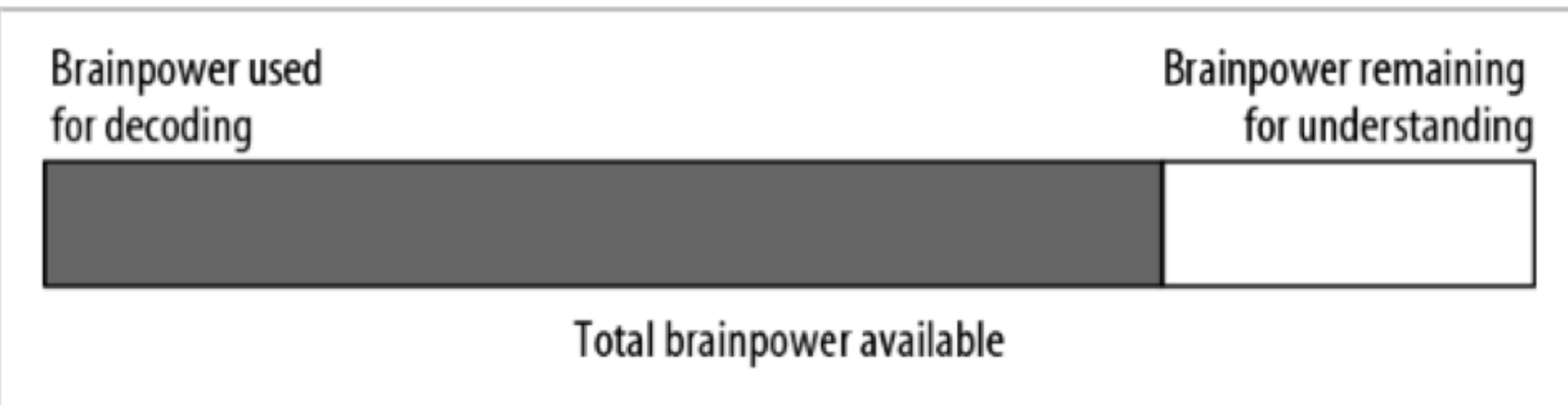
COLORBREWER 2.0
color advice for cartography

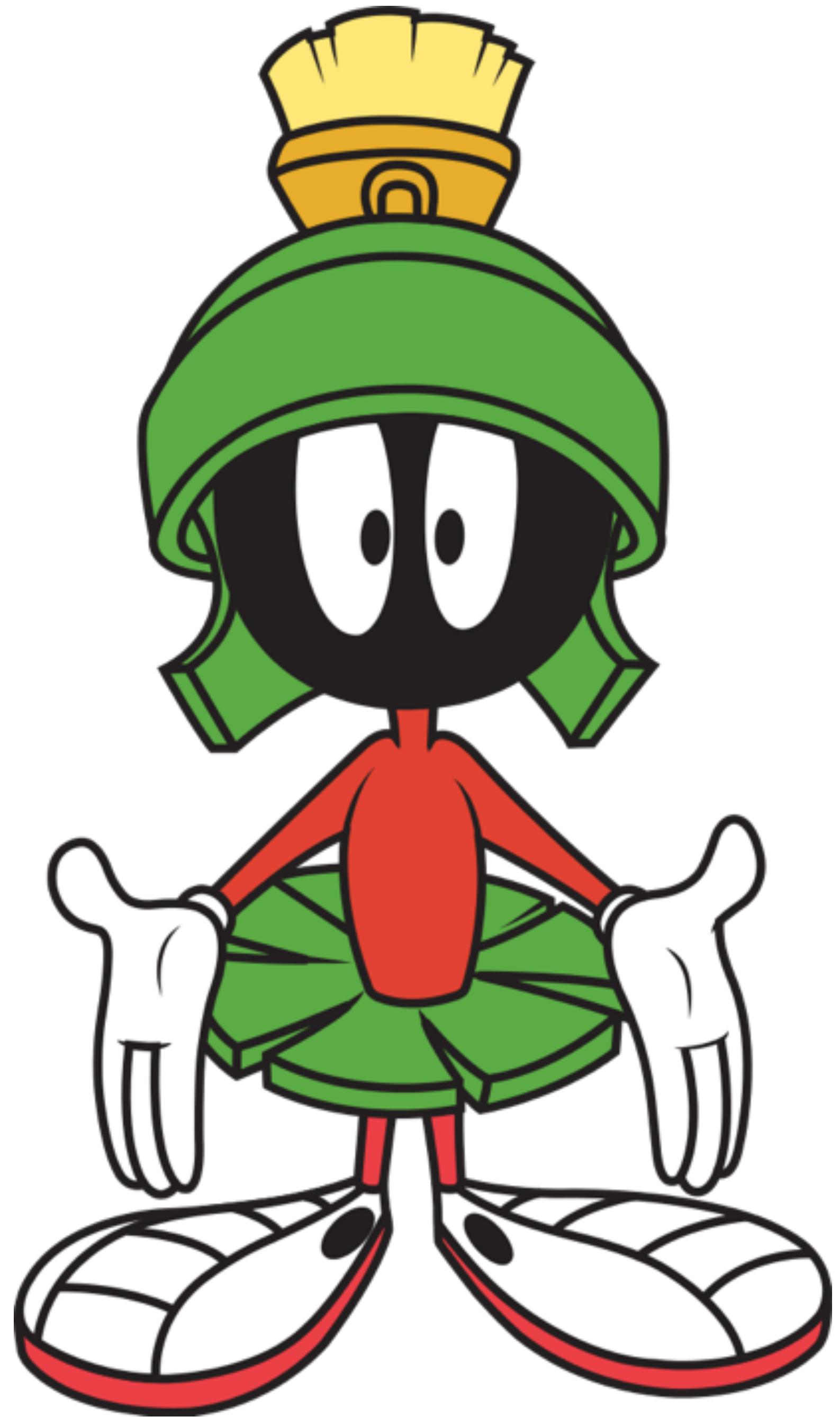
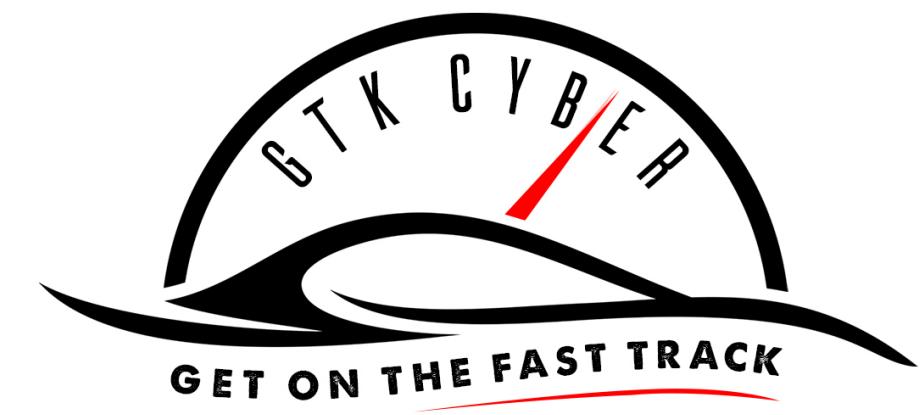
EXPORT

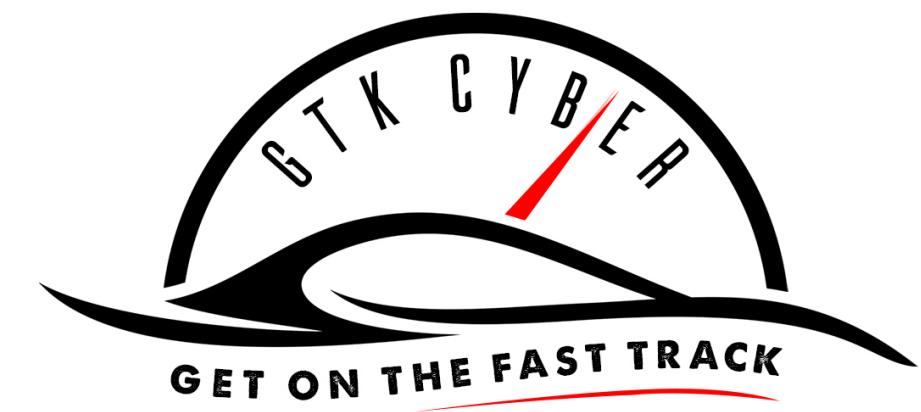
3-class BuGn

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









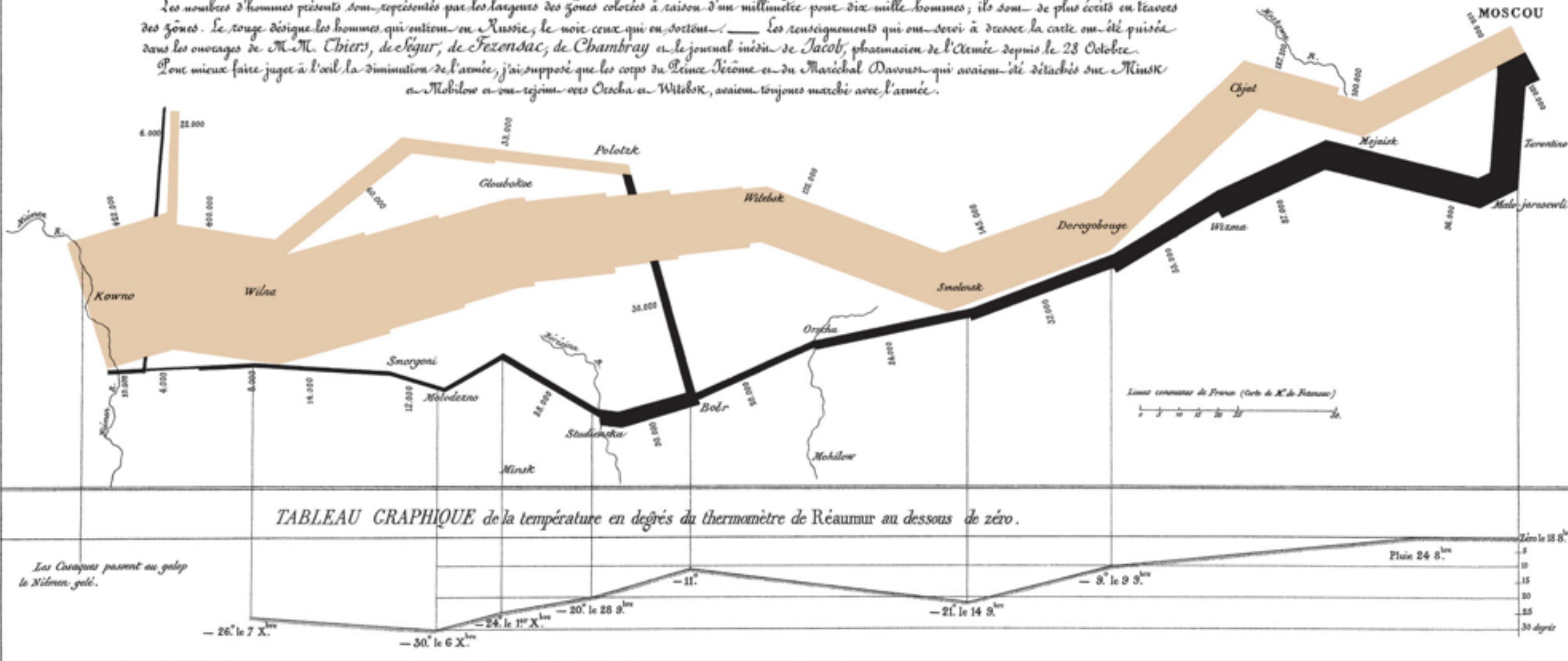
Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.

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

Paris, le 20 Novembre 1869.

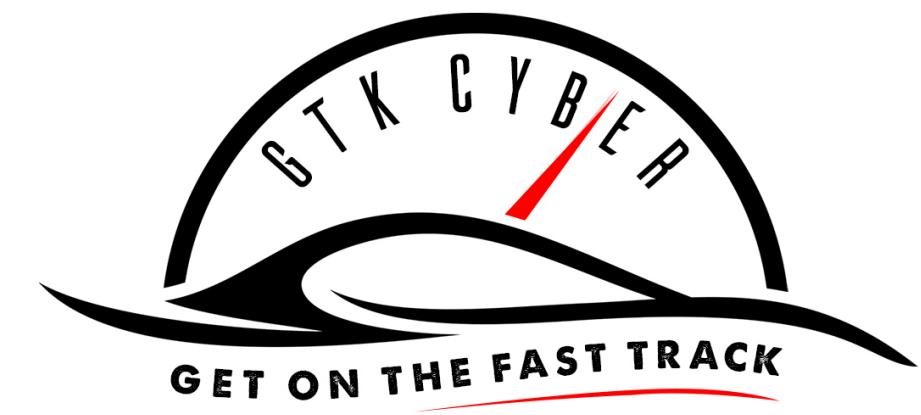
Les nombres d'hommes présents sont représentés par les larges des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en lettres des zones. Le rouge désigne les hommes qui ont été en Russie; le noir ceux qui en sortirent. — Les renseignements qui ont servi à dresser 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.



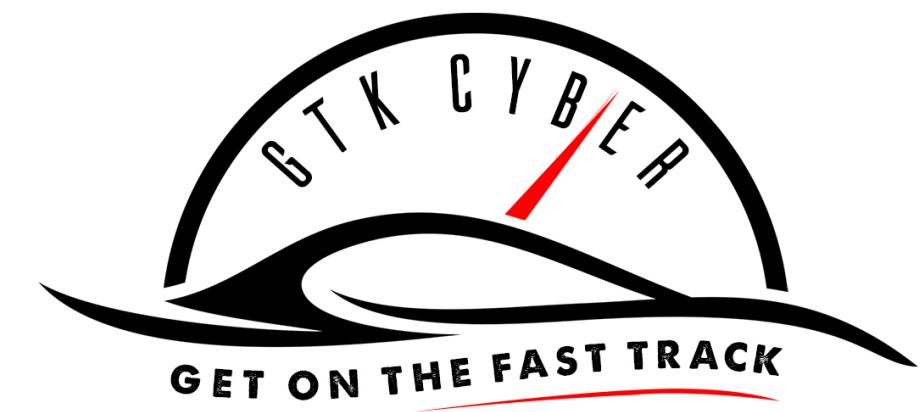
Avec par Regnier, 8, Rue J^e Marie S^e G^e à Paris.

Imp. Lib. Regnier et Bourdet.

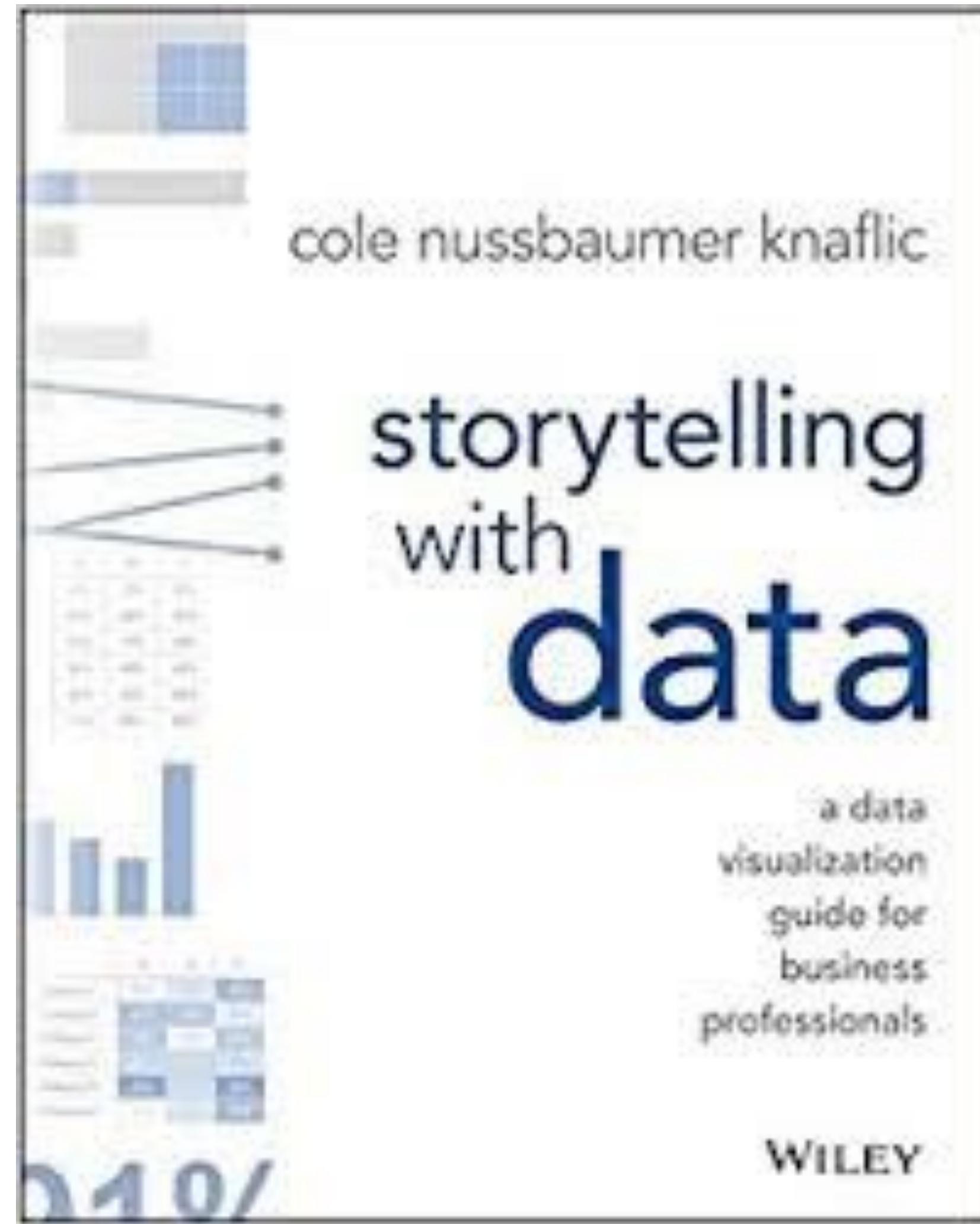


Recommended Reading



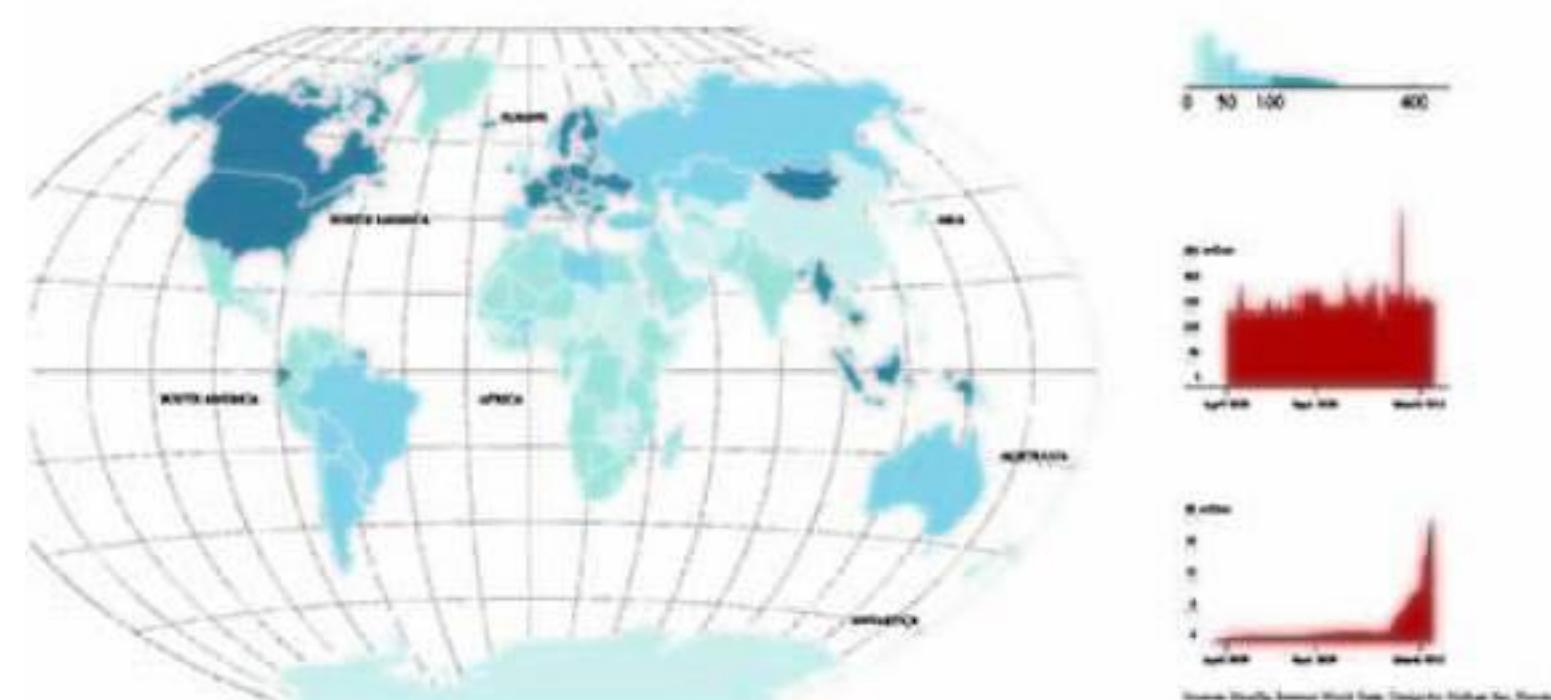
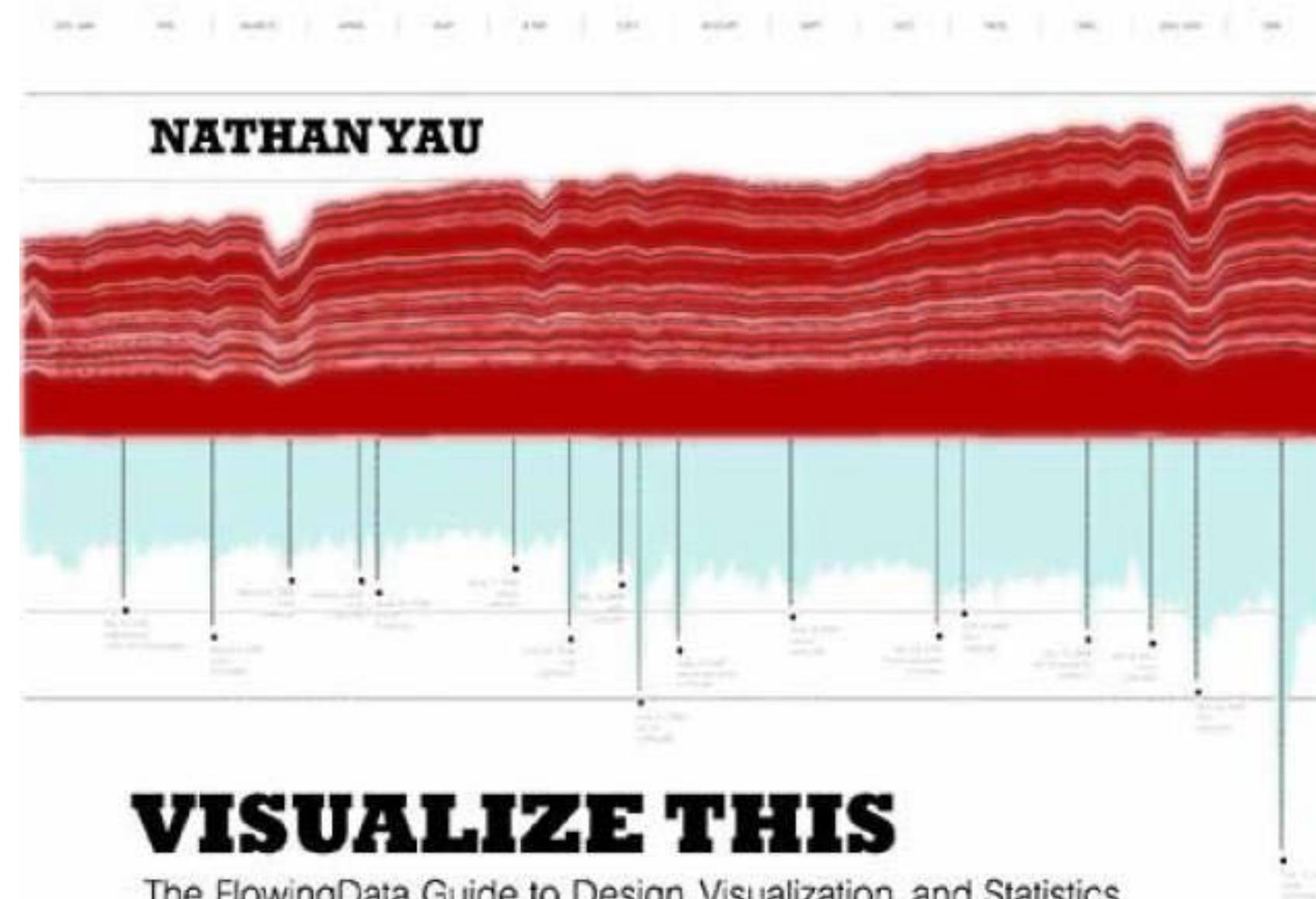


Recommended Reading





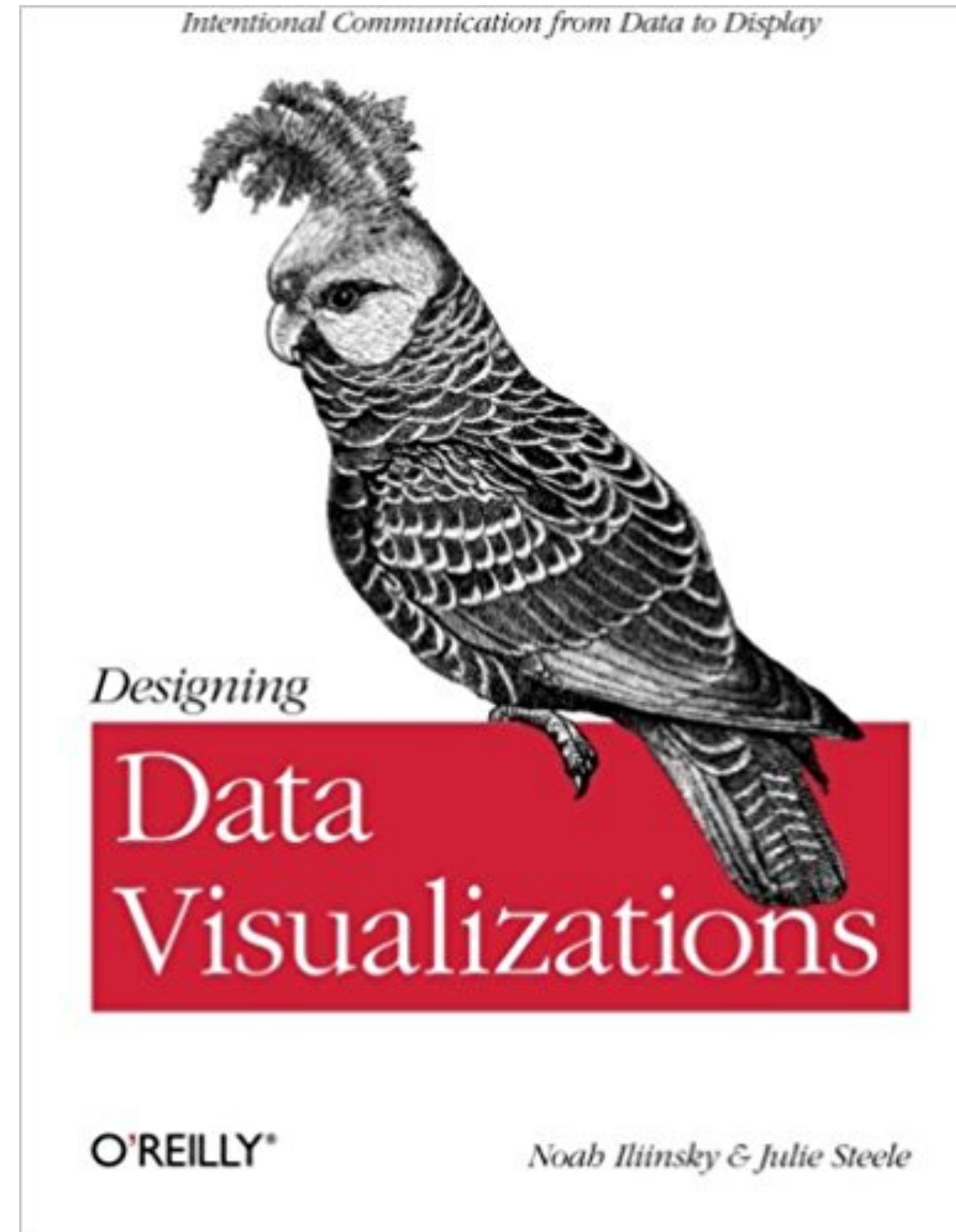
Recommended Reading



WILEY

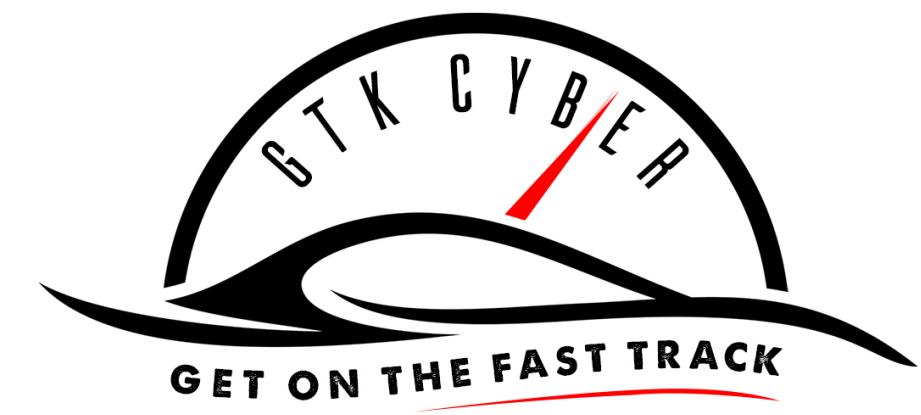


Recommended Reading



Data Visualization in Python

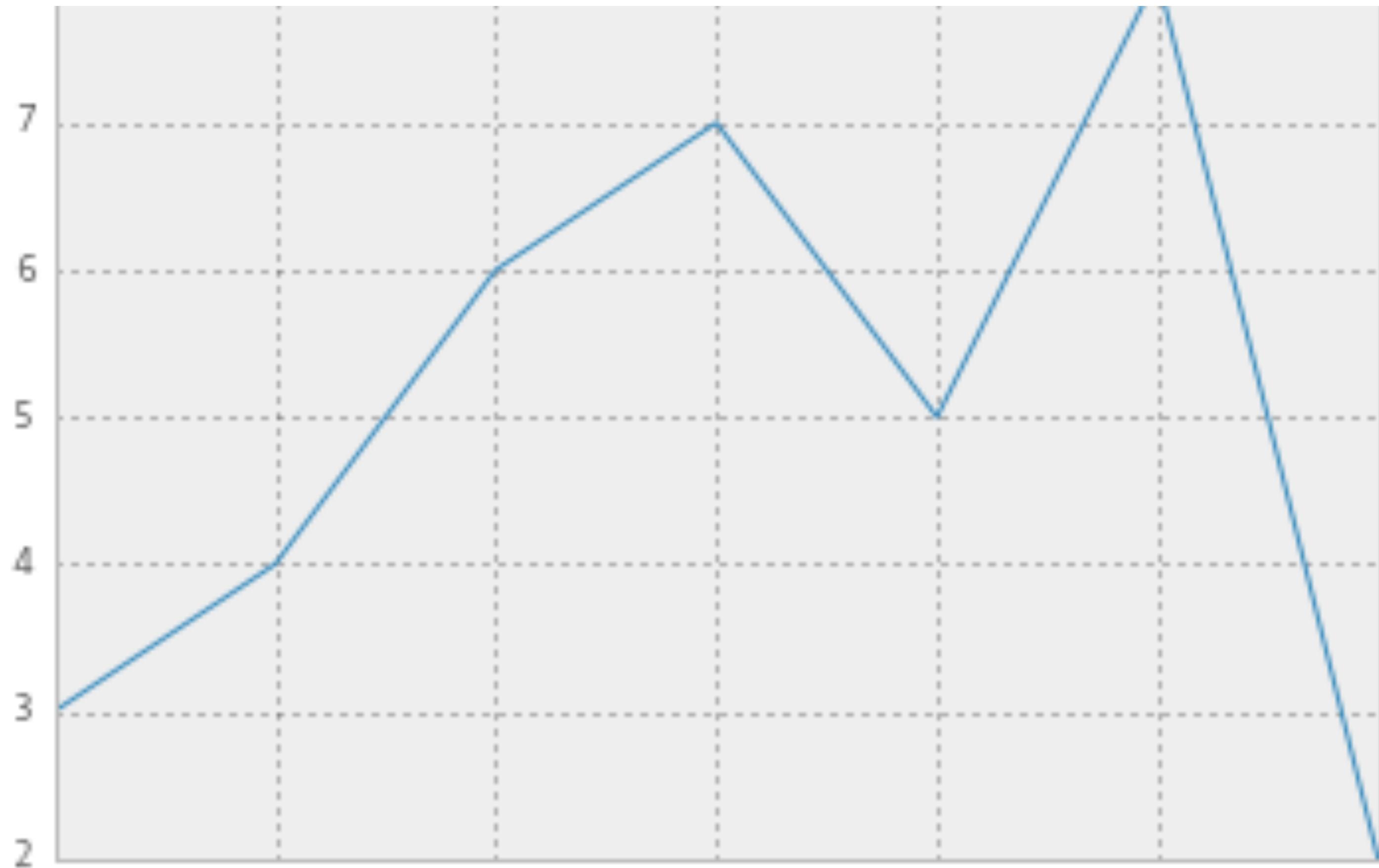
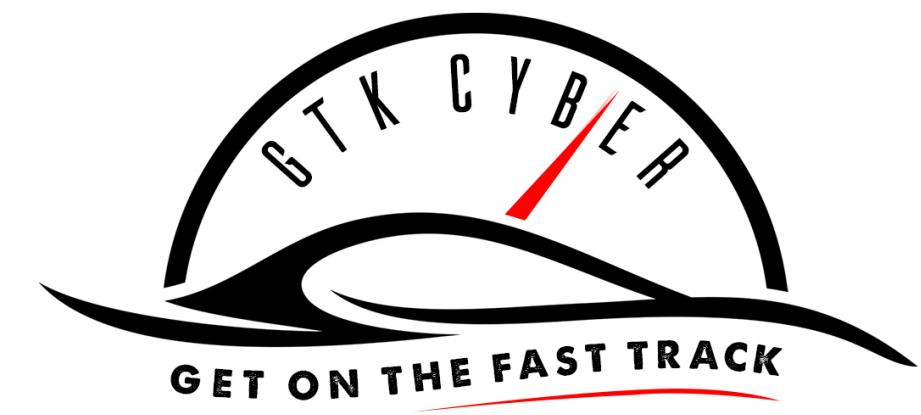
GET ON THE FAST TRACK



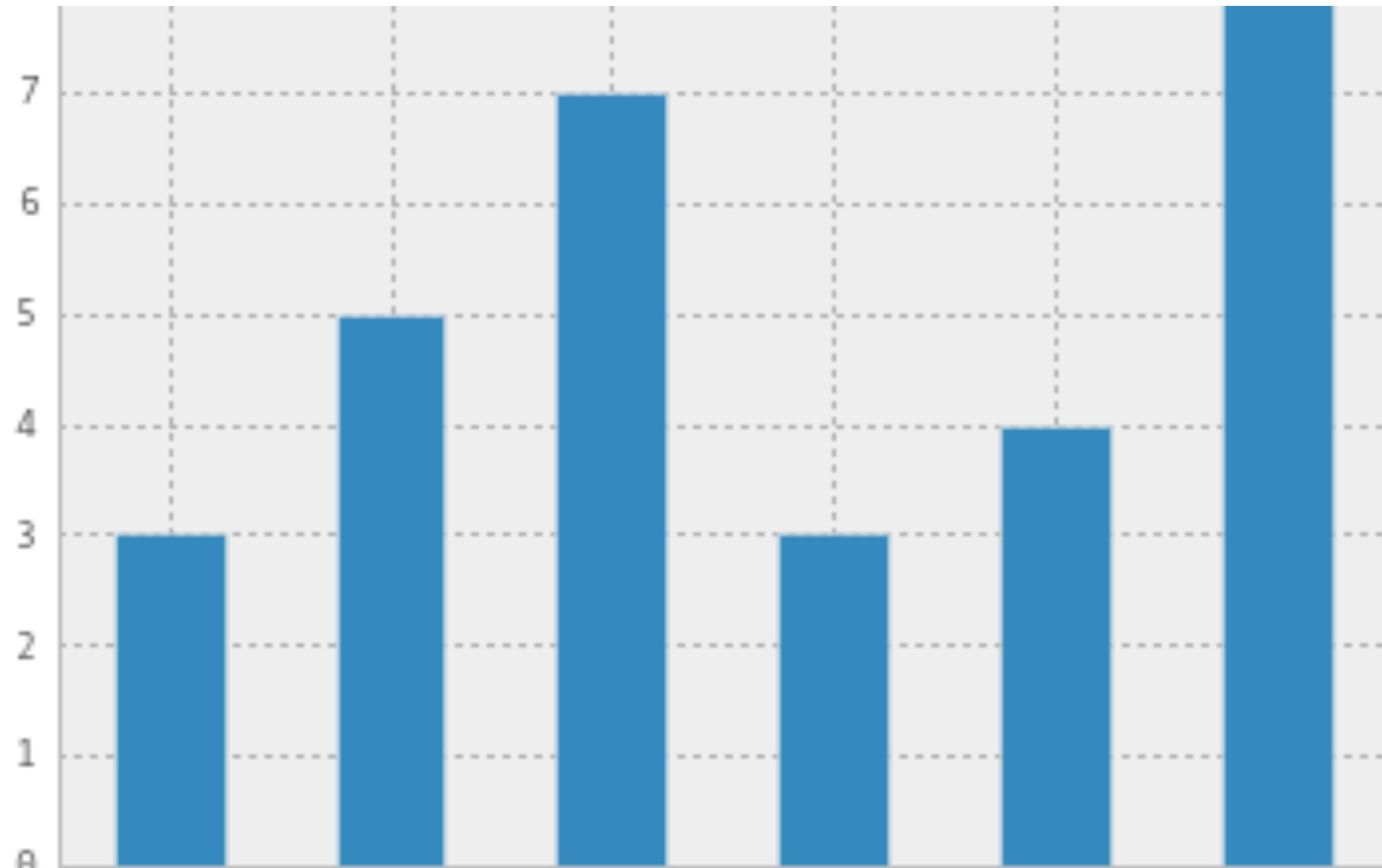
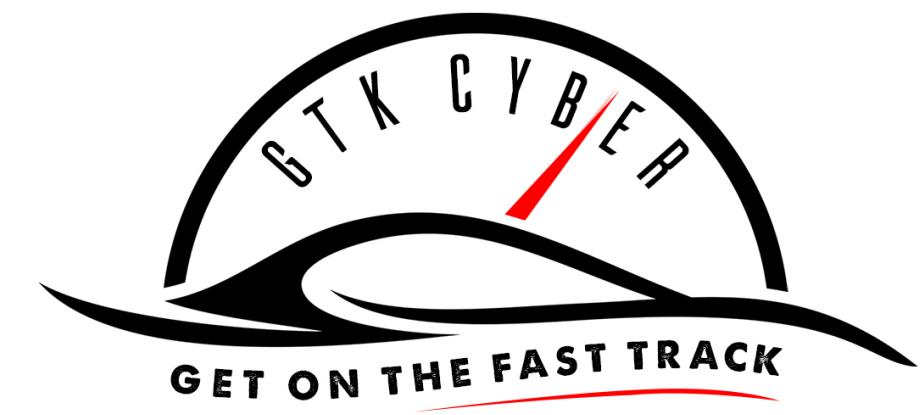
Python Visualization Libraries

- **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!

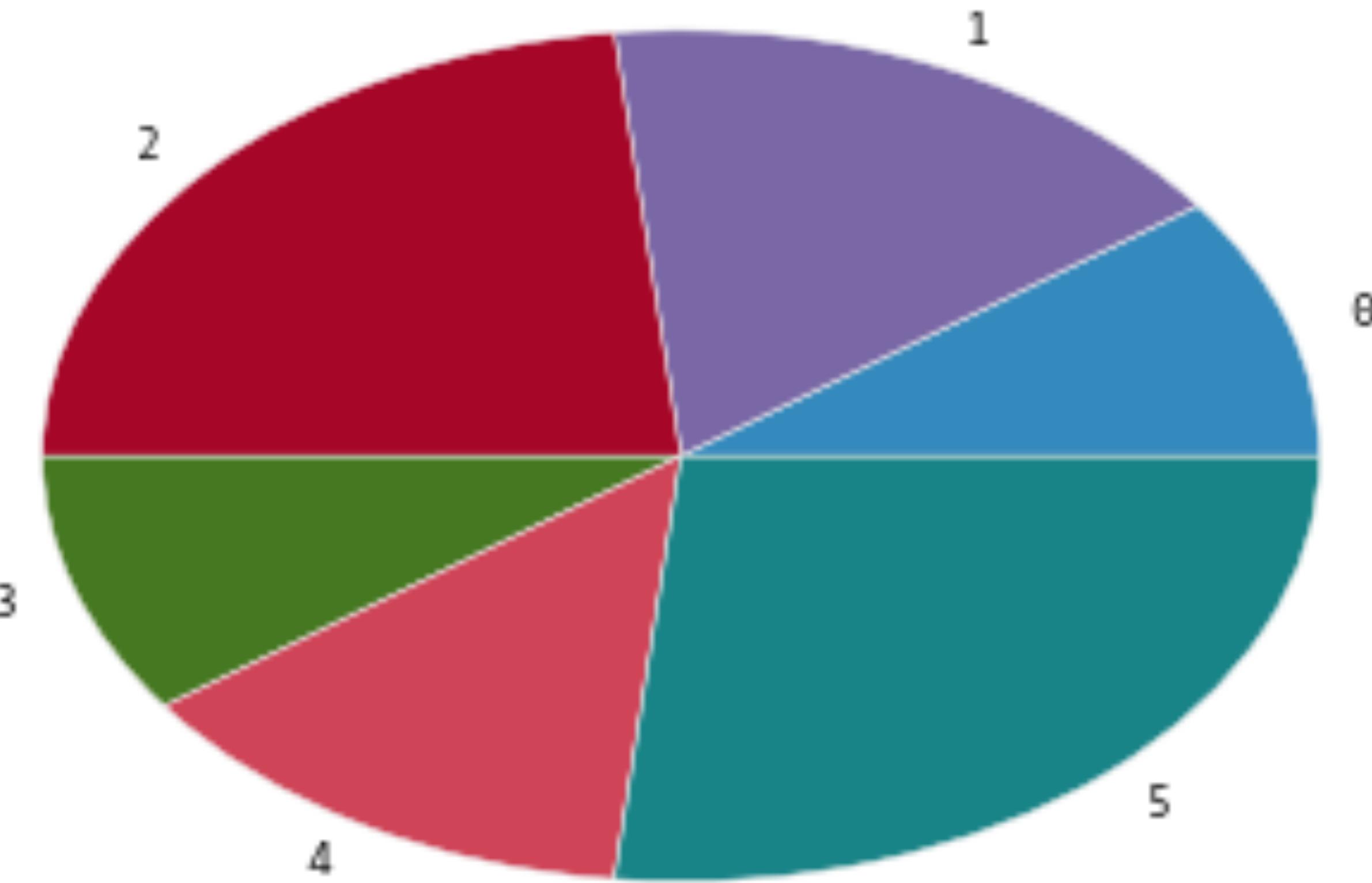
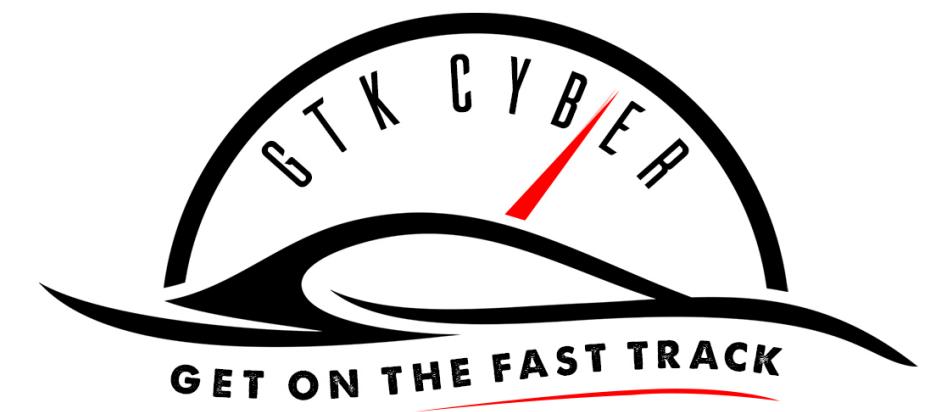
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
%matplotlib inline
pd.options.display.mpl_style = 'default'
```



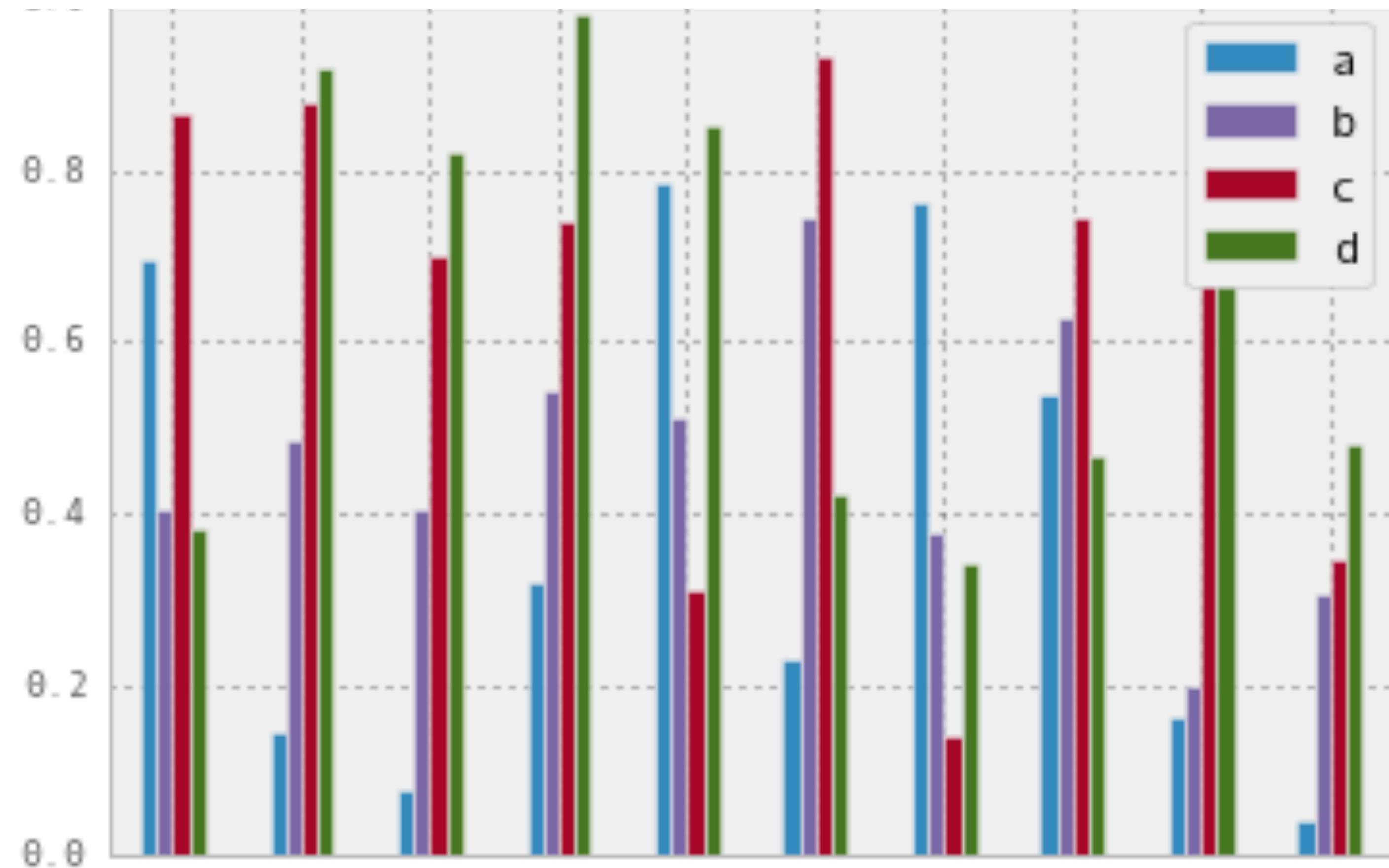
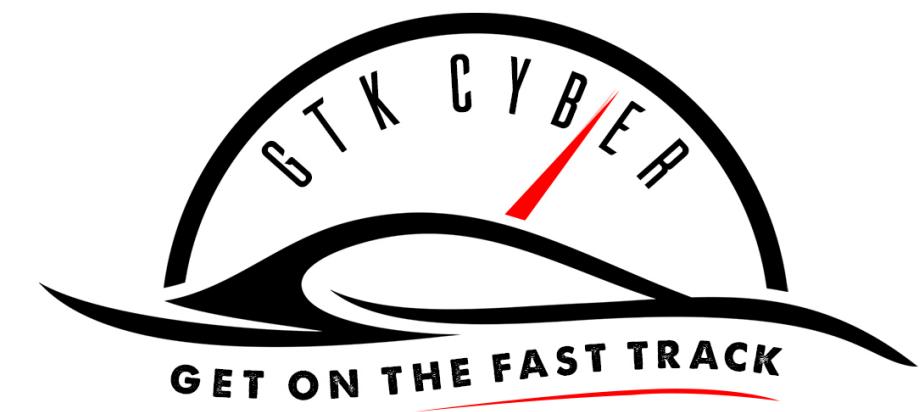
```
data = pd.Series( [ 3, 4, 6, 7, 5, 8, 2 ] )
graph = data.plot()
```



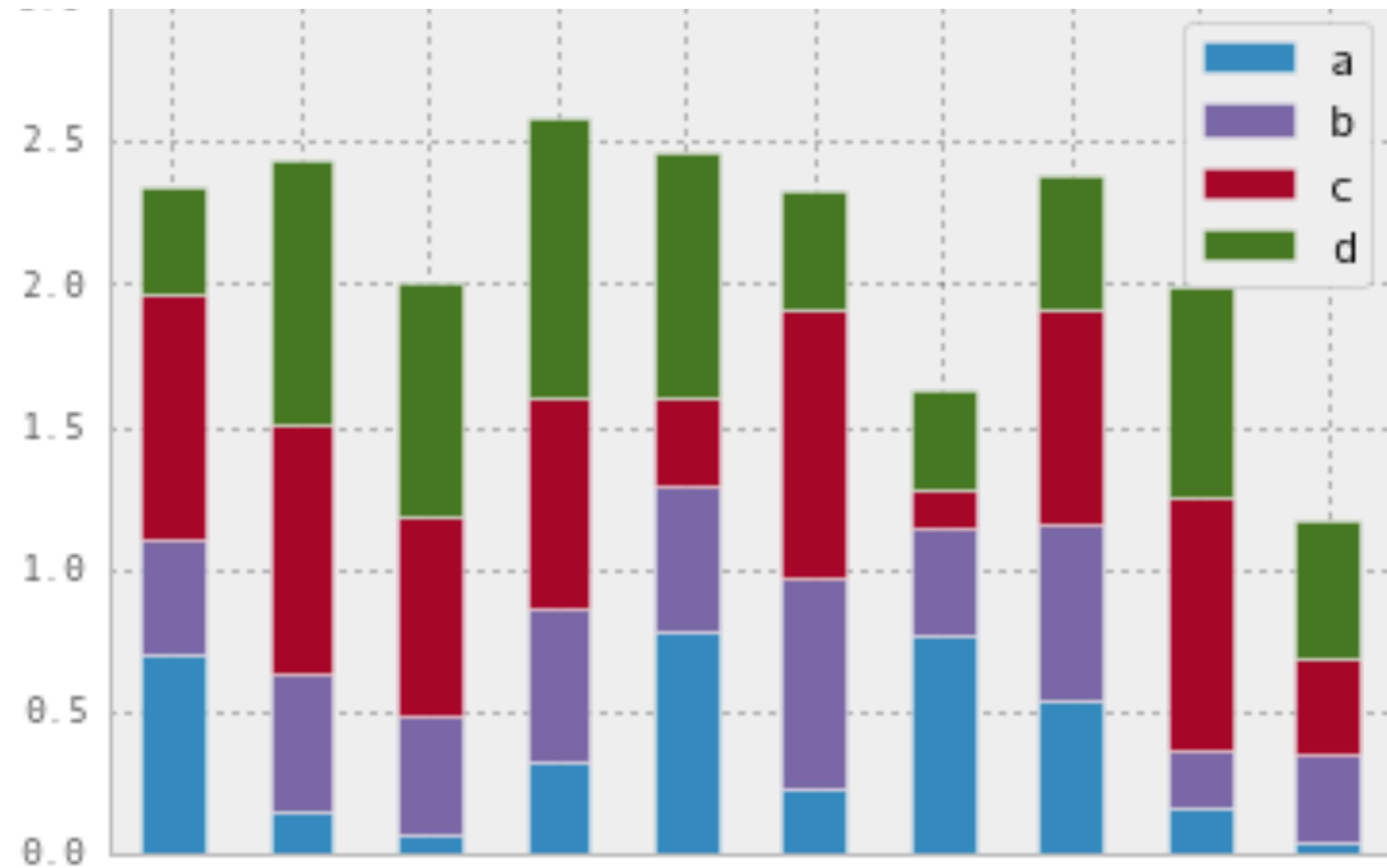
```
barchart = data.plot( kind="bar" )
```



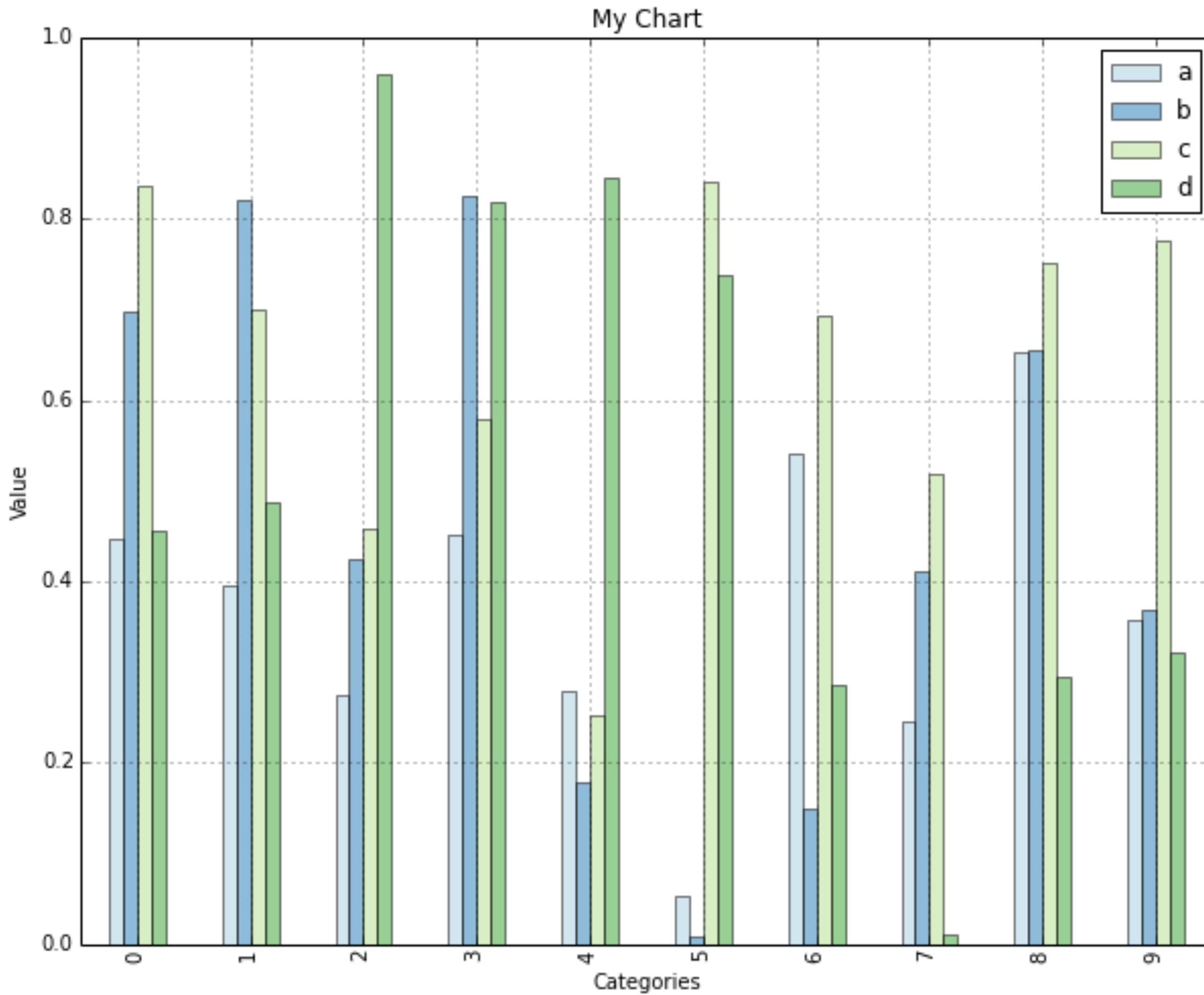
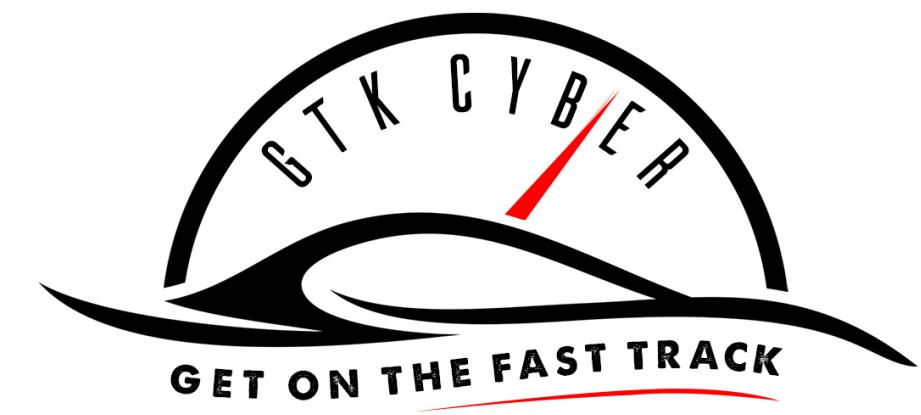
```
piechart = data.plot( kind="pie" )
```



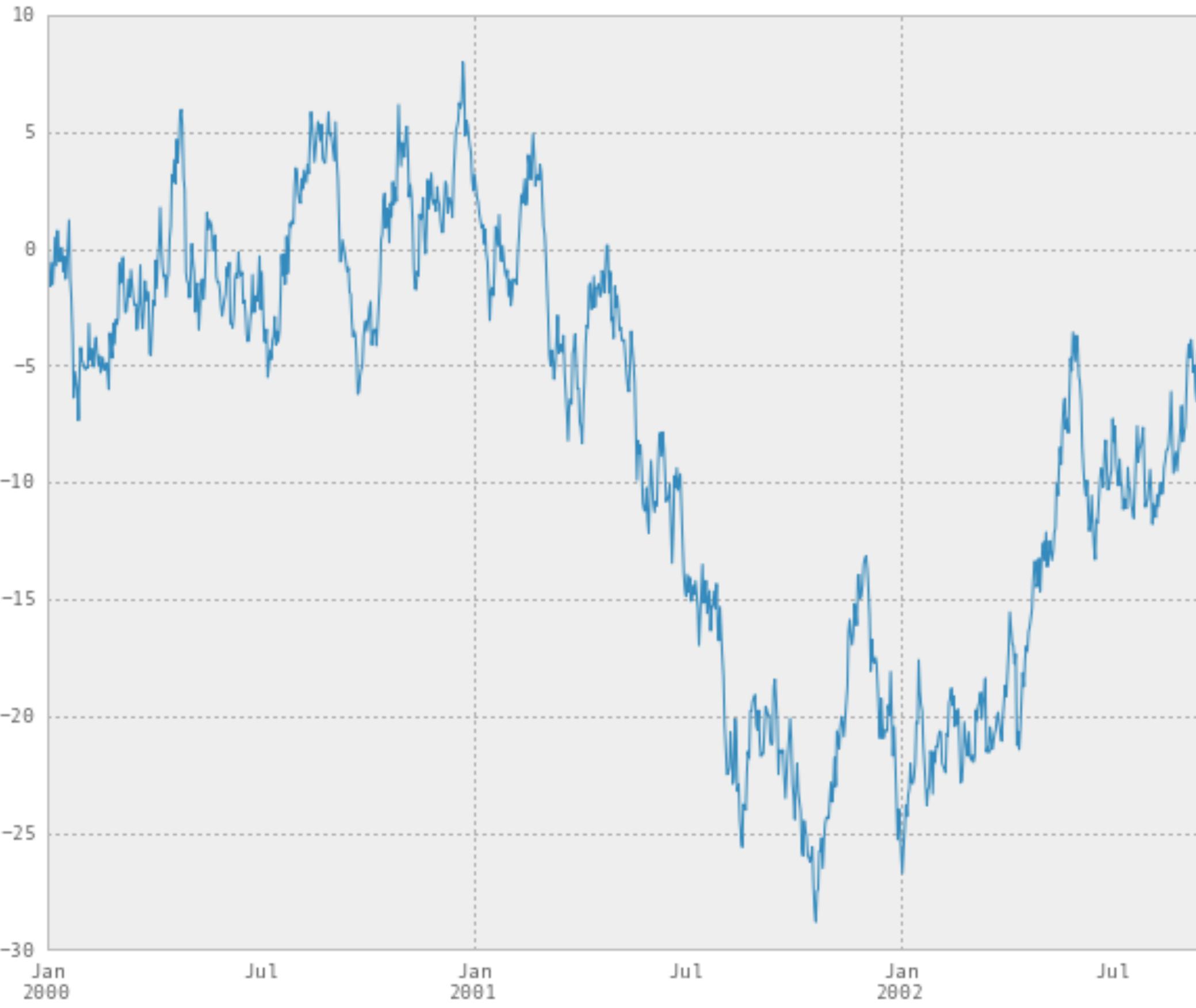
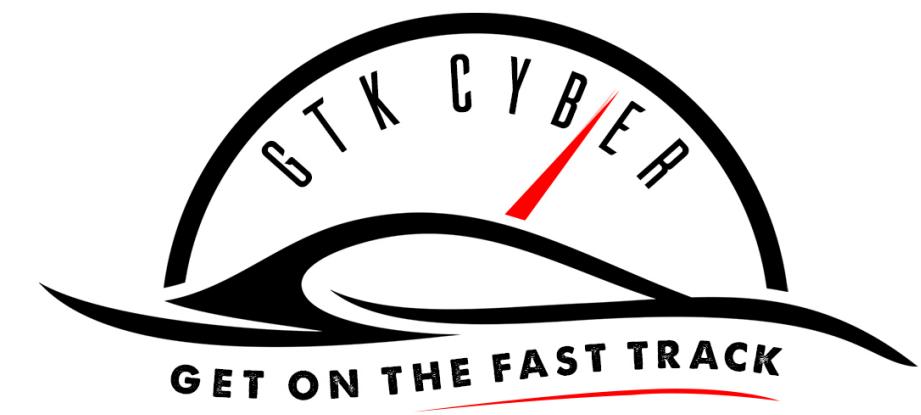
```
df2 = pd.DataFrame(np.random.rand(10, 4),  
columns=[ 'a' , 'b' , 'c' , 'd' ] )  
df2.plot( kind='bar' )
```



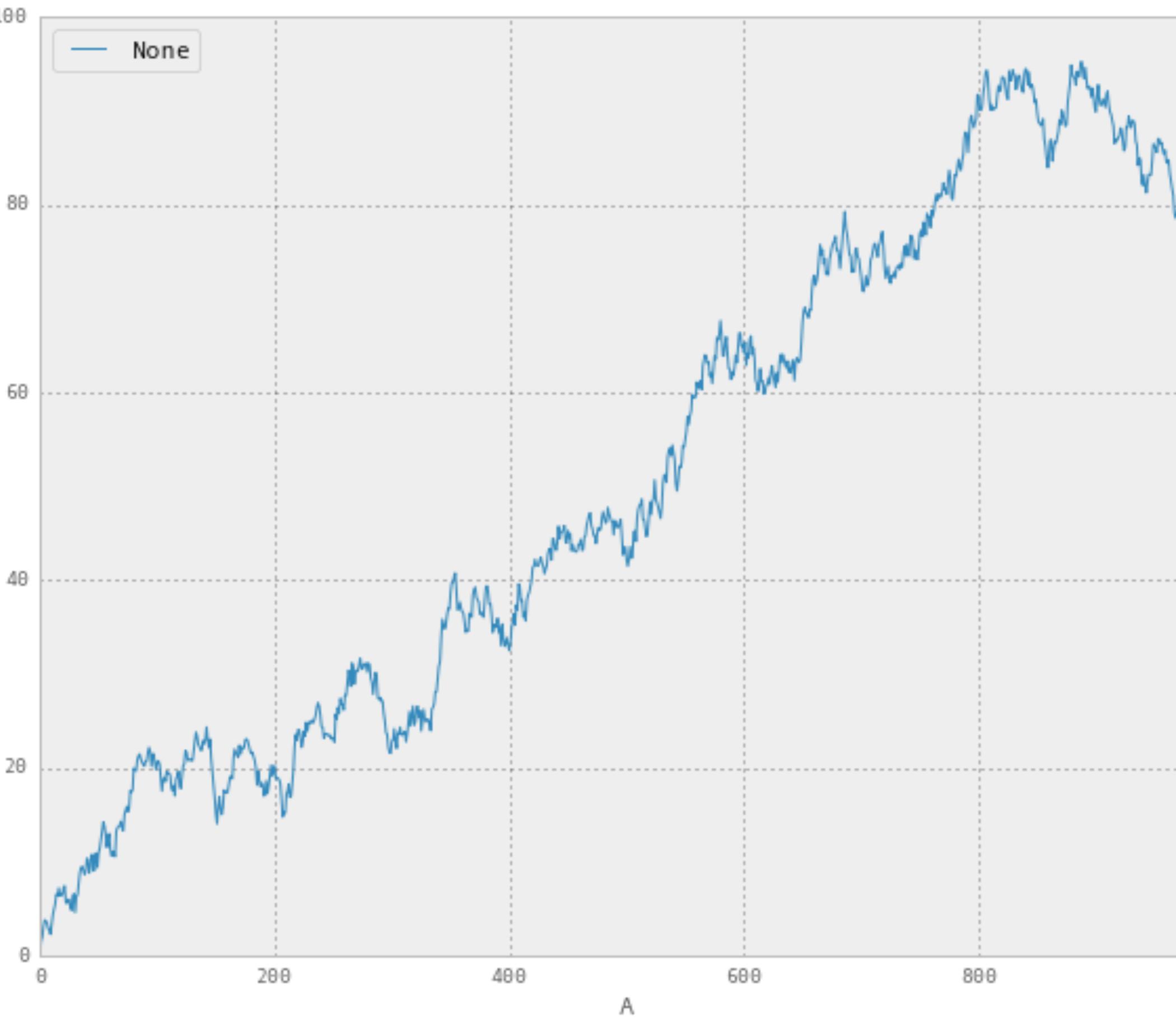
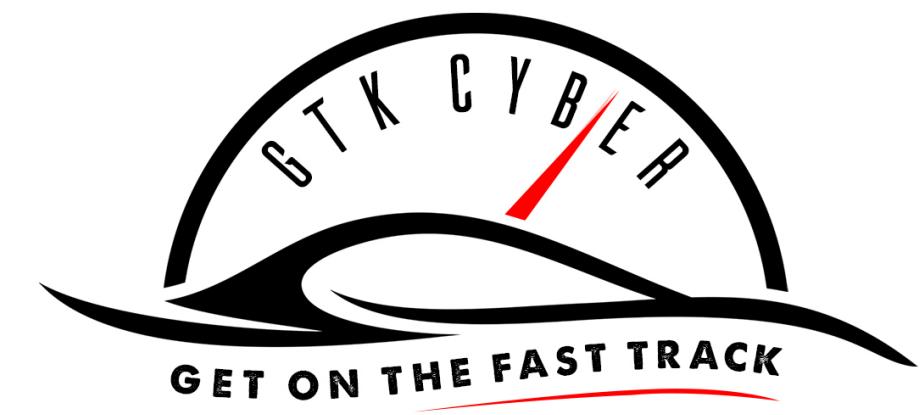
```
df2.plot( kind='bar', stacked=True )
```



```
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" )
```

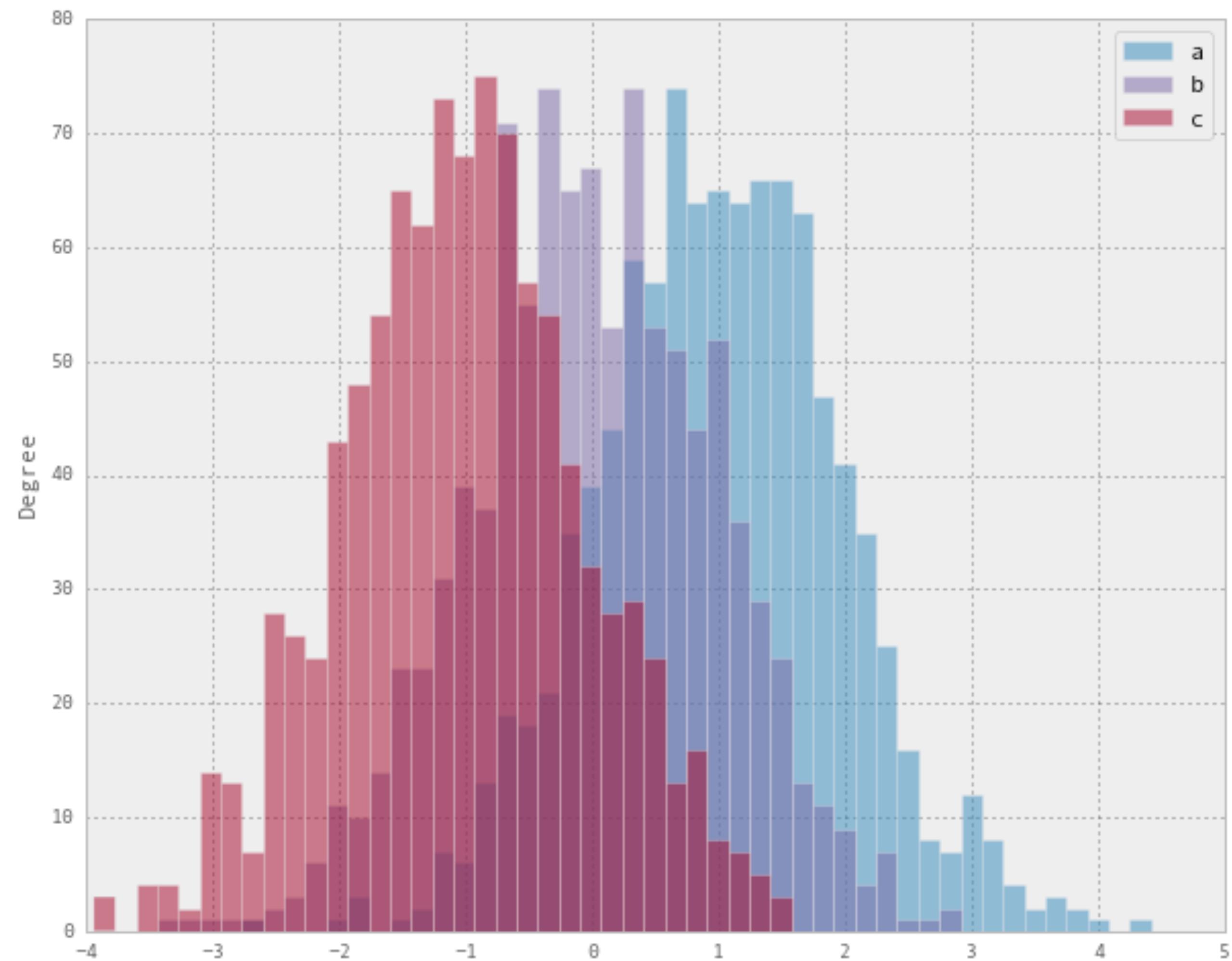
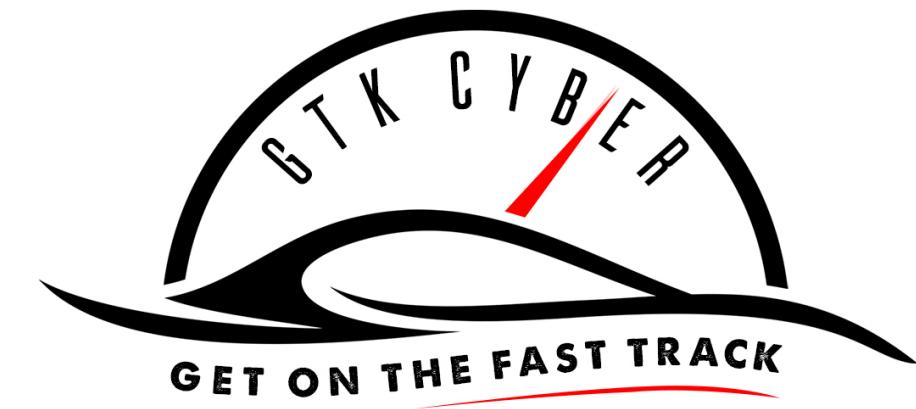


```
ts = pd.Series(np.random.randn( 1000 ),  
index=pd.date_range('1/1/2000', periods=1000))  
ts = ts.cumsum()  
timeseriesChart = ts.plot( figsize=(10, 8) )
```

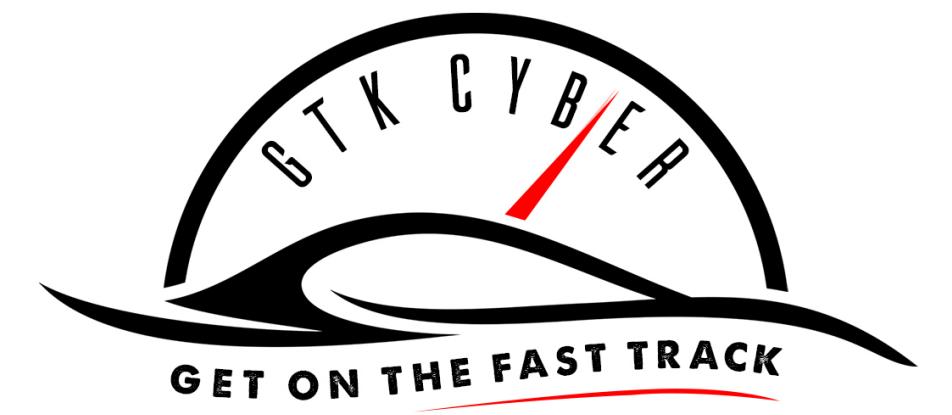


```
df3 = pd.DataFrame(np.random.randn(1000, 2),  
columns=[ 'B' , 'C' ]).cumsum()  
df3[ 'A' ] = pd.Series(list(range(len(df3))))
```

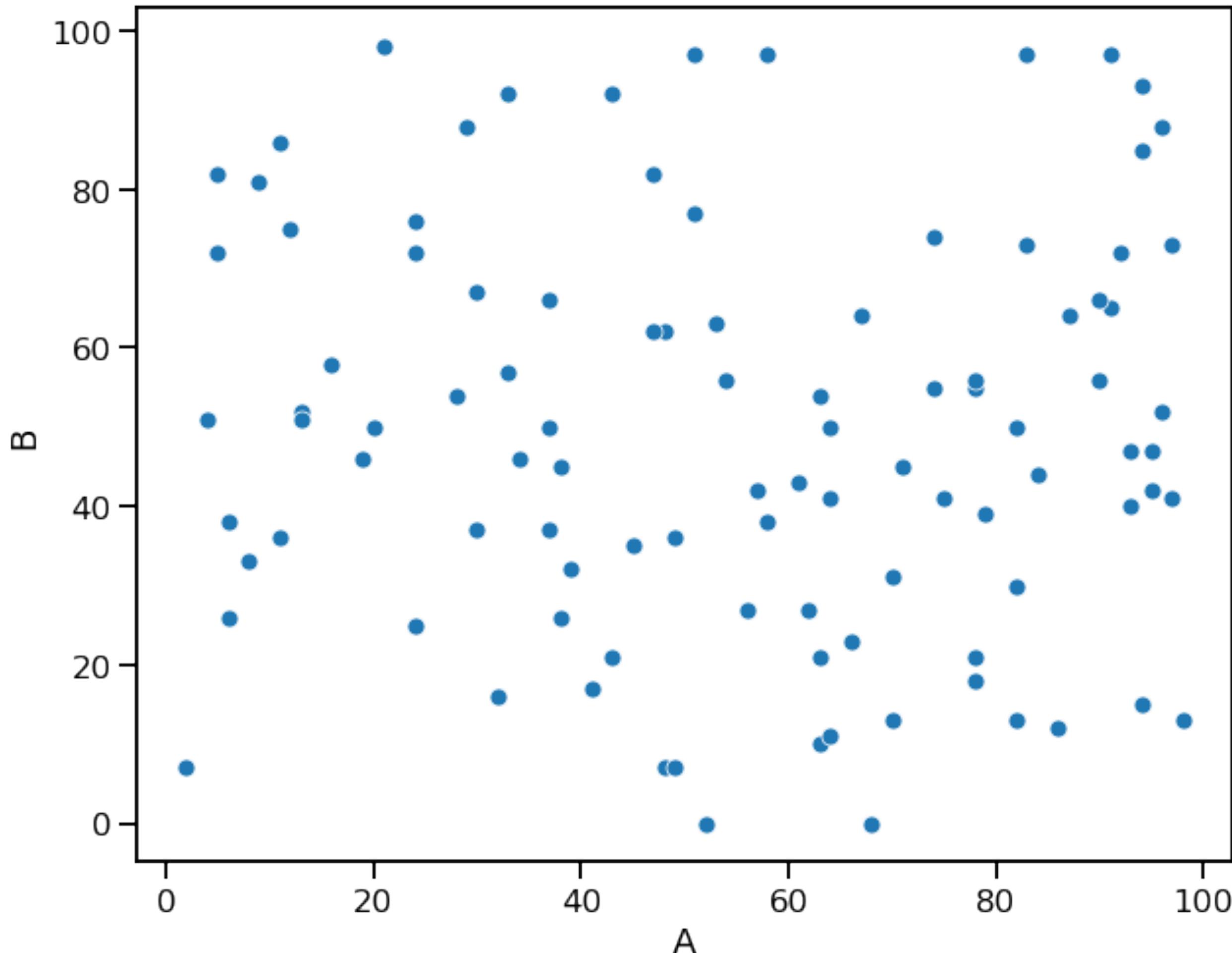
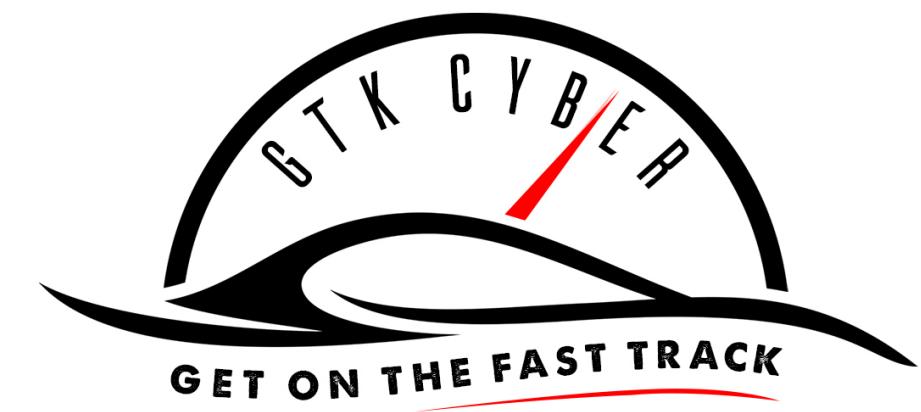
```
df3.plot( x='A' , y='B' )
```



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

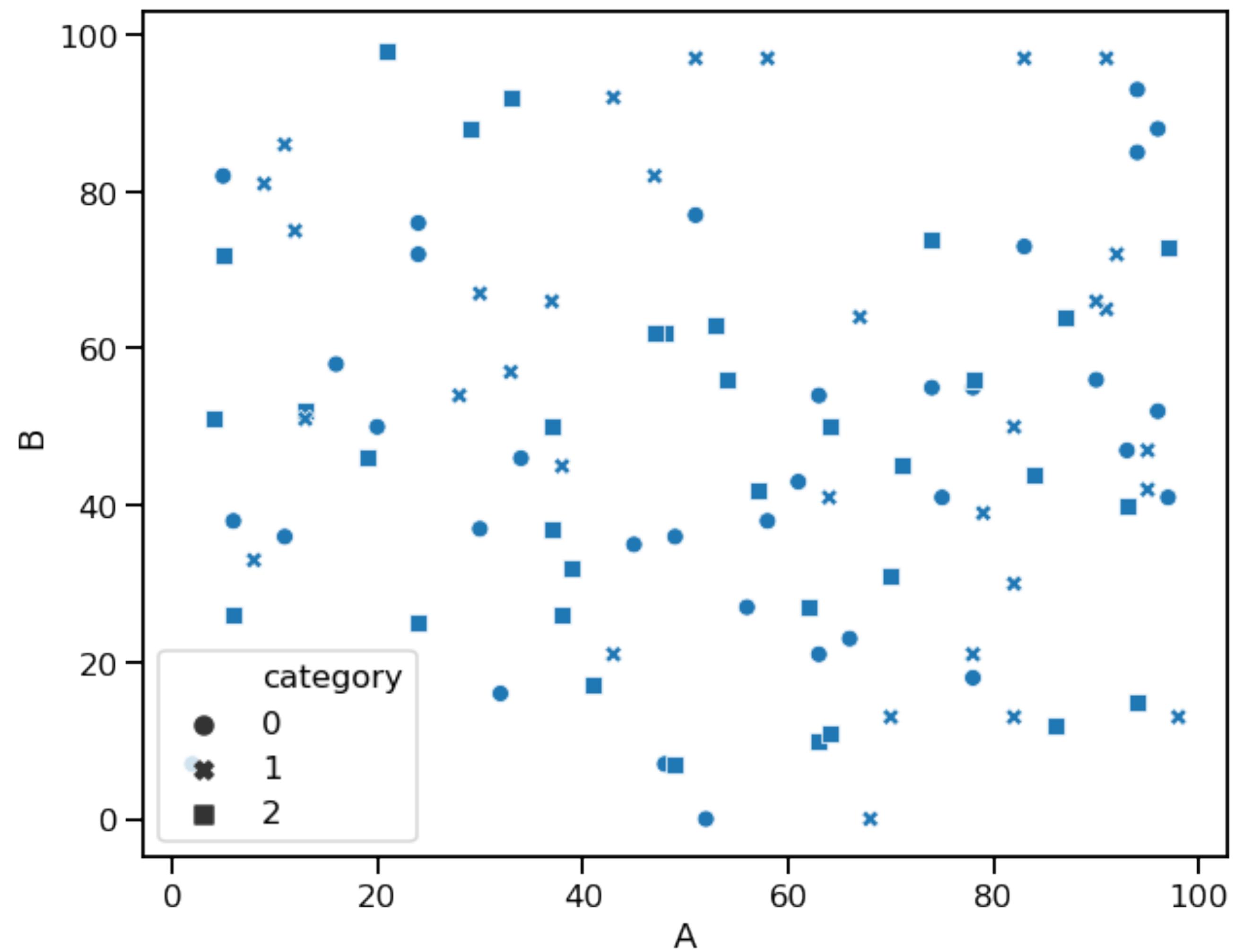
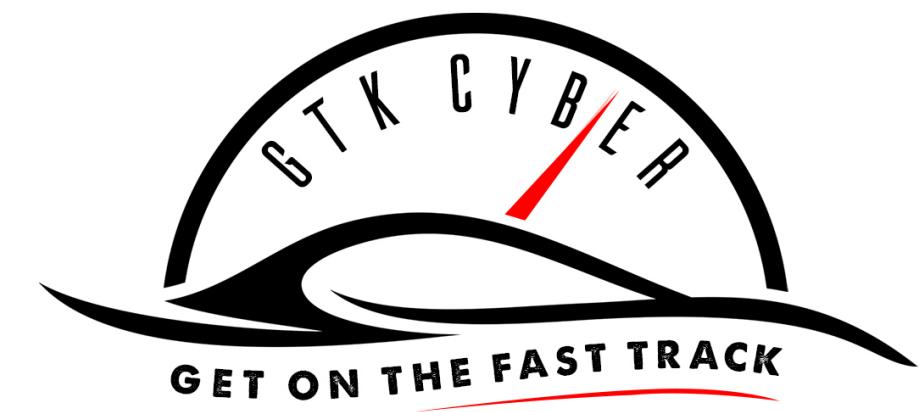


import seaborn as sns

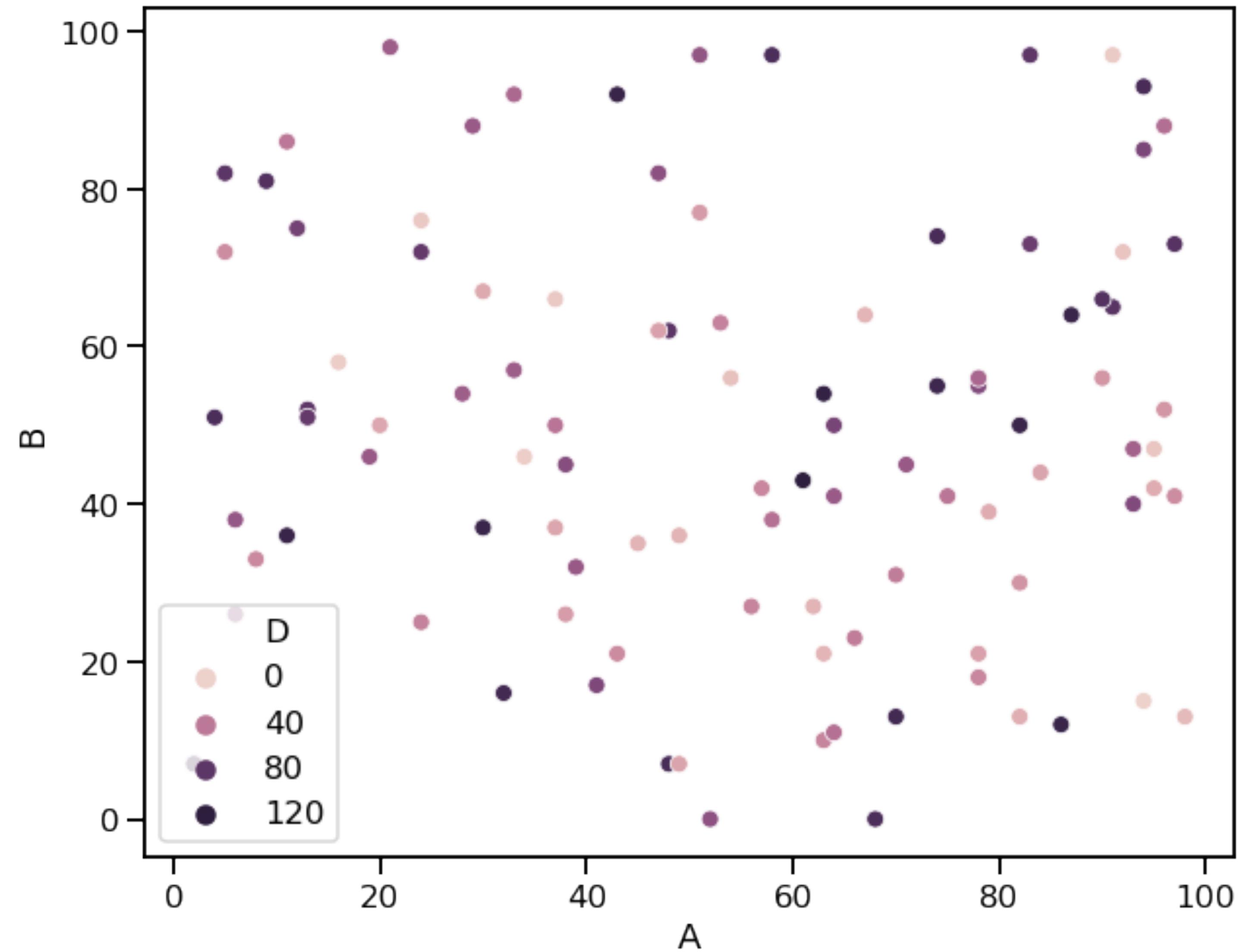


```
df = pd.DataFrame(np.random.randint(0,100,size=(100,  
4)), columns=list('ABCD'))  
df['category'] = df['C'] % 3
```

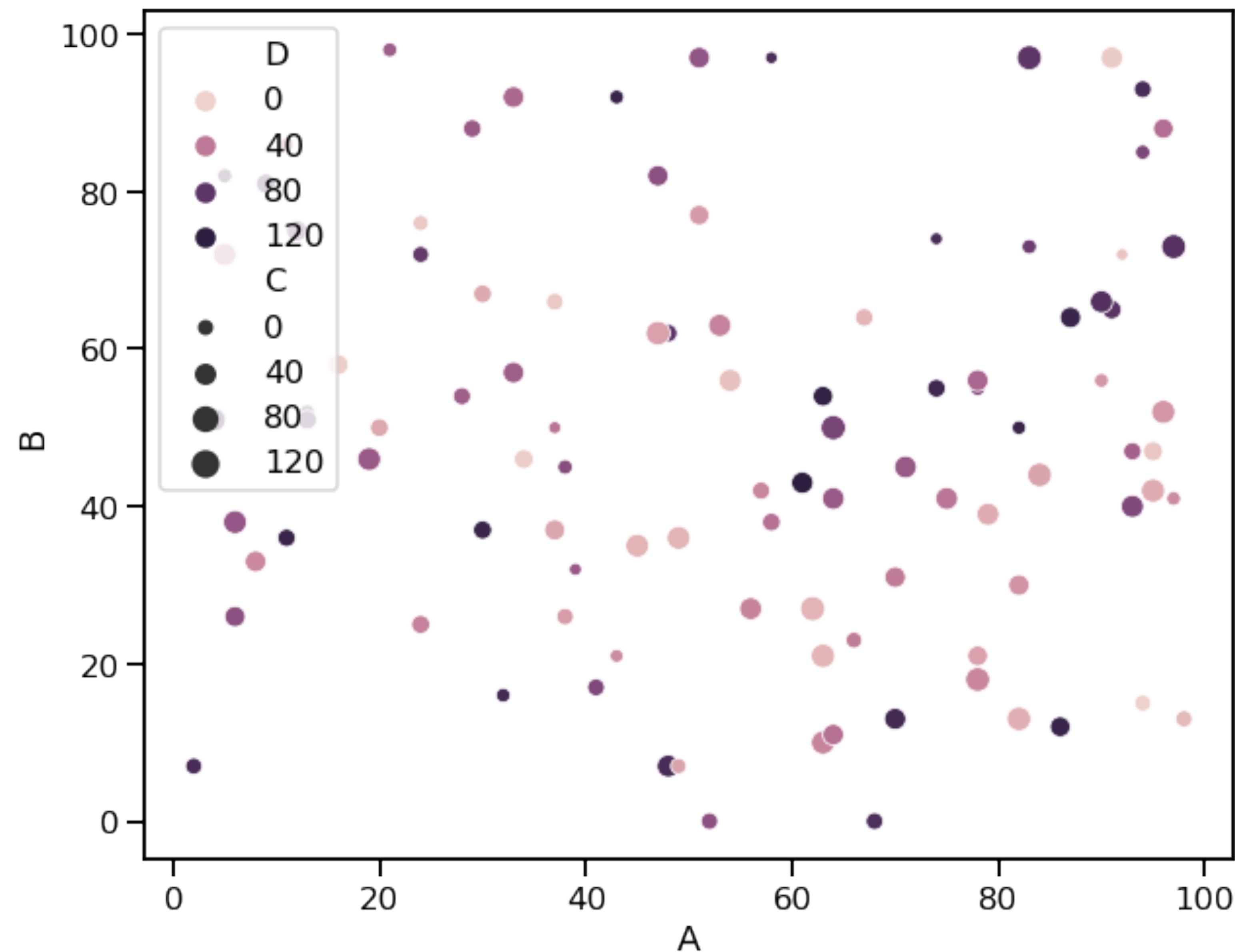
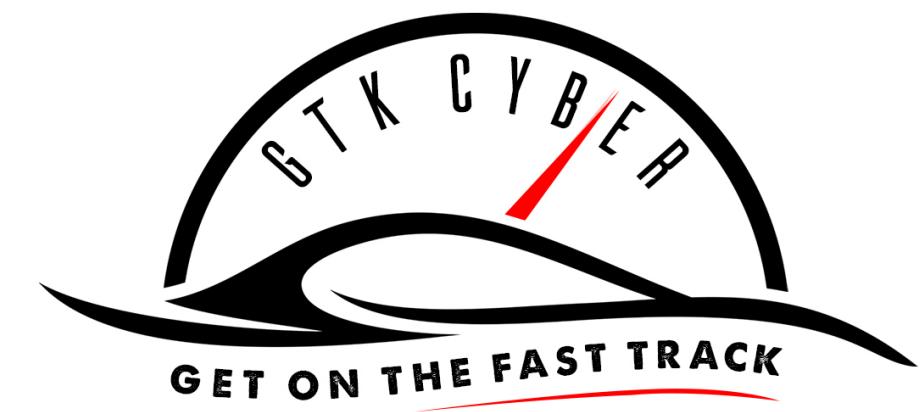
```
sns.scatterplot(x=df['A'], y=df['B'], data=df)
```



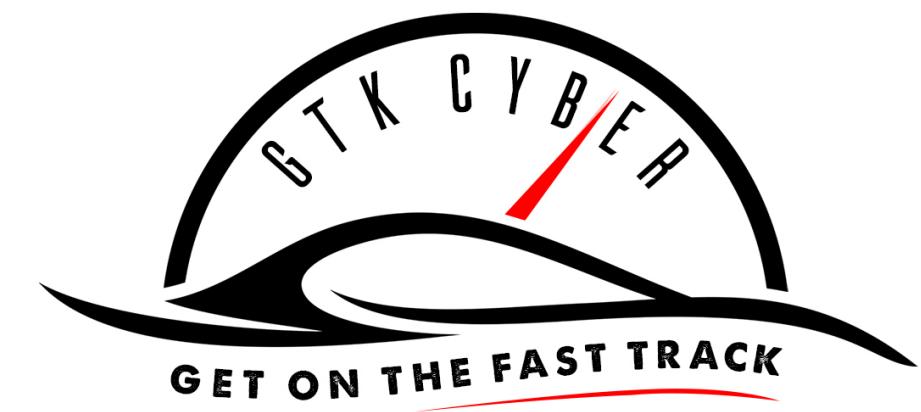
```
sns.scatterplot(x=df[ 'A' ], y=df[ 'B' ], style=df[ 'category' ])
```



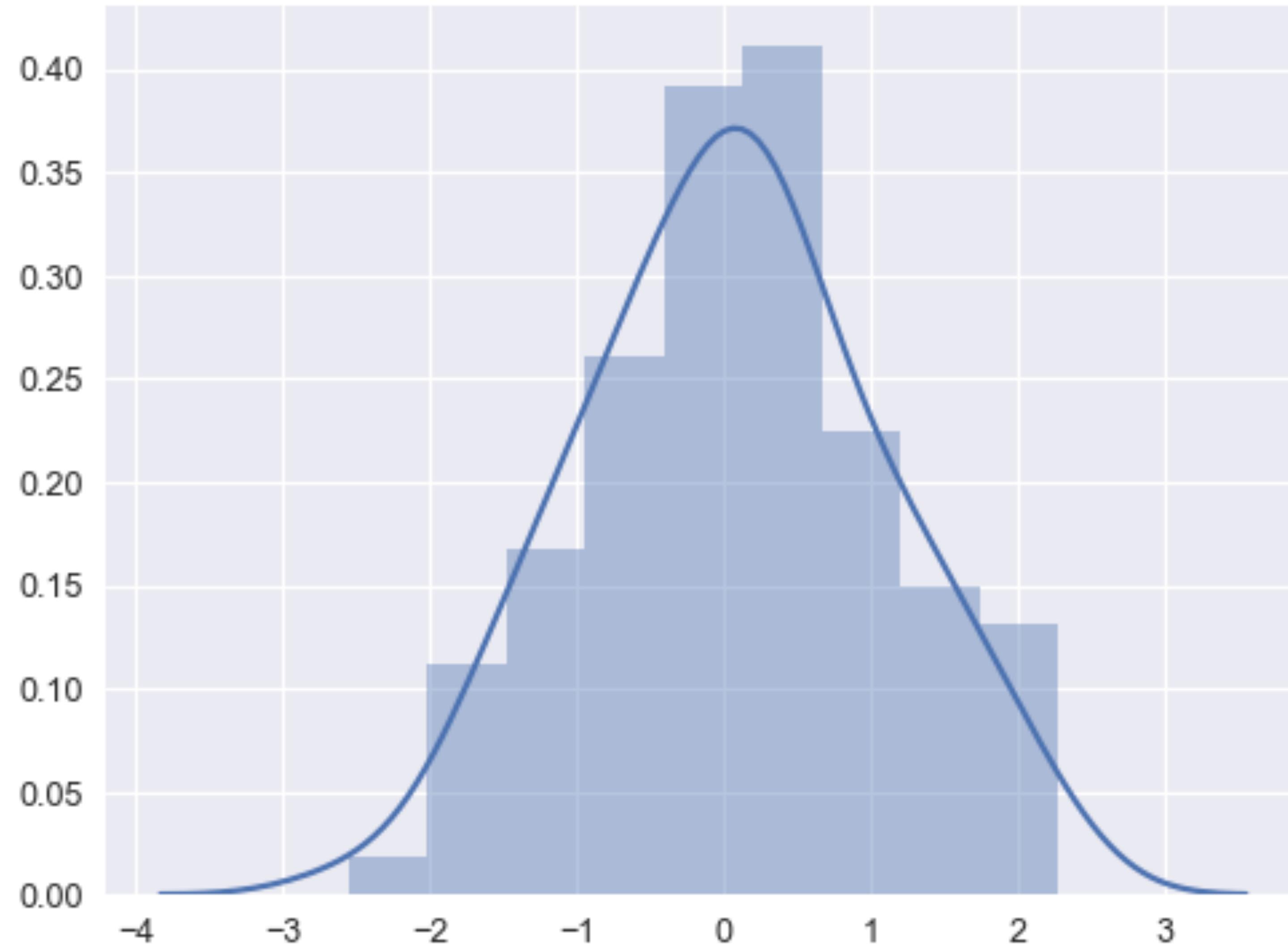
```
sns.scatterplot(x=df['A'], y=df['B'], hue=df['D'])
```



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

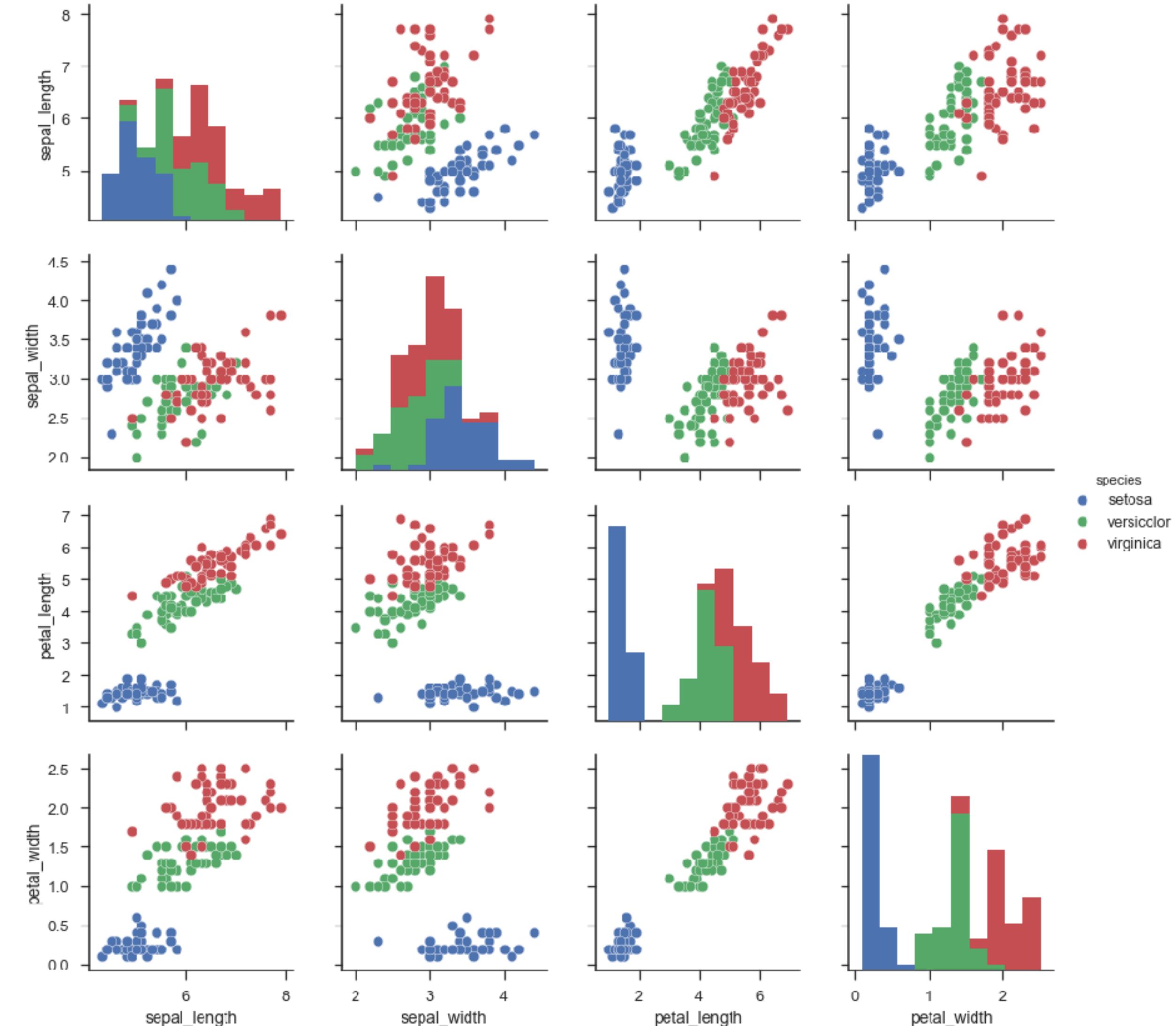


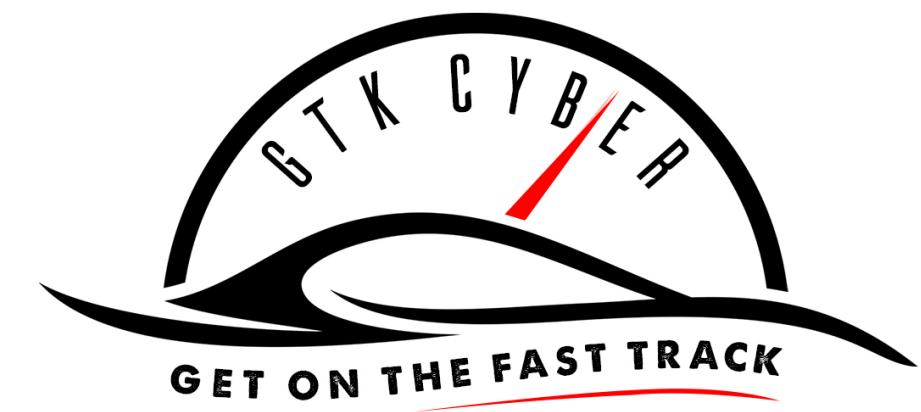
ax = sns.distplot(<data>)



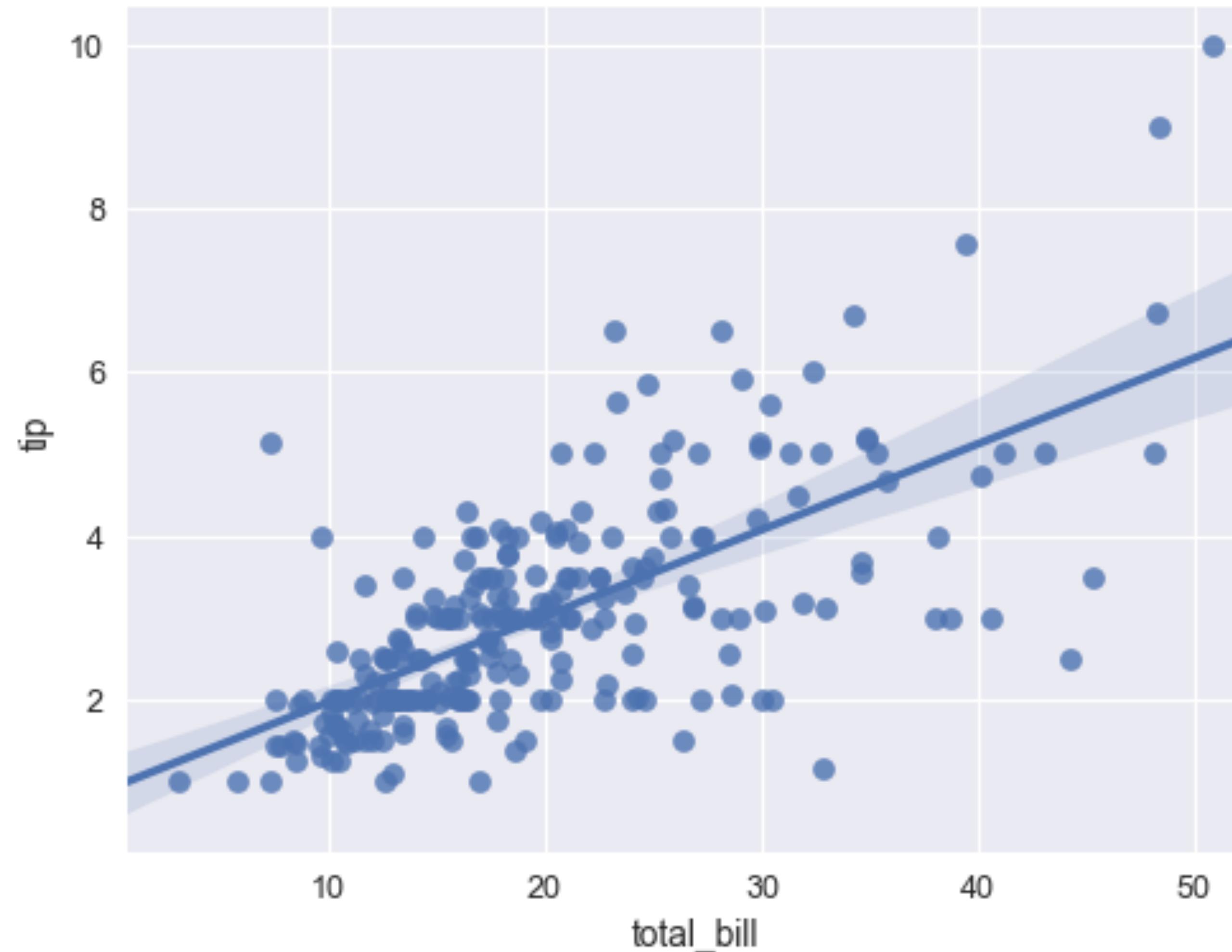


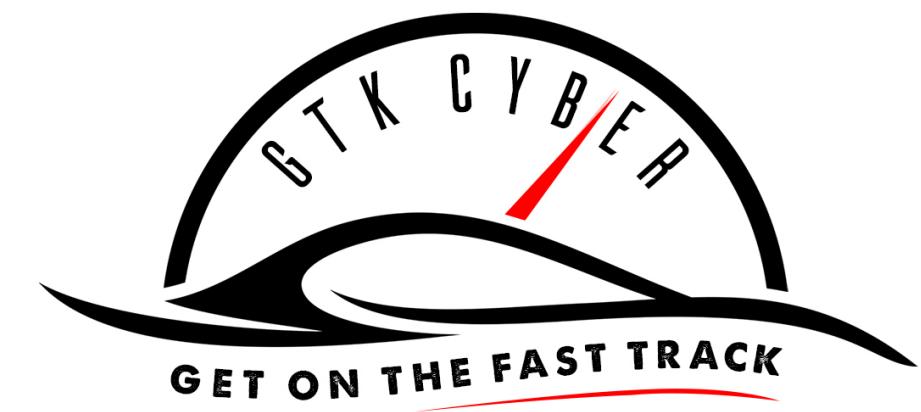
`sns.pairplot(<data>, hue="<target>")`



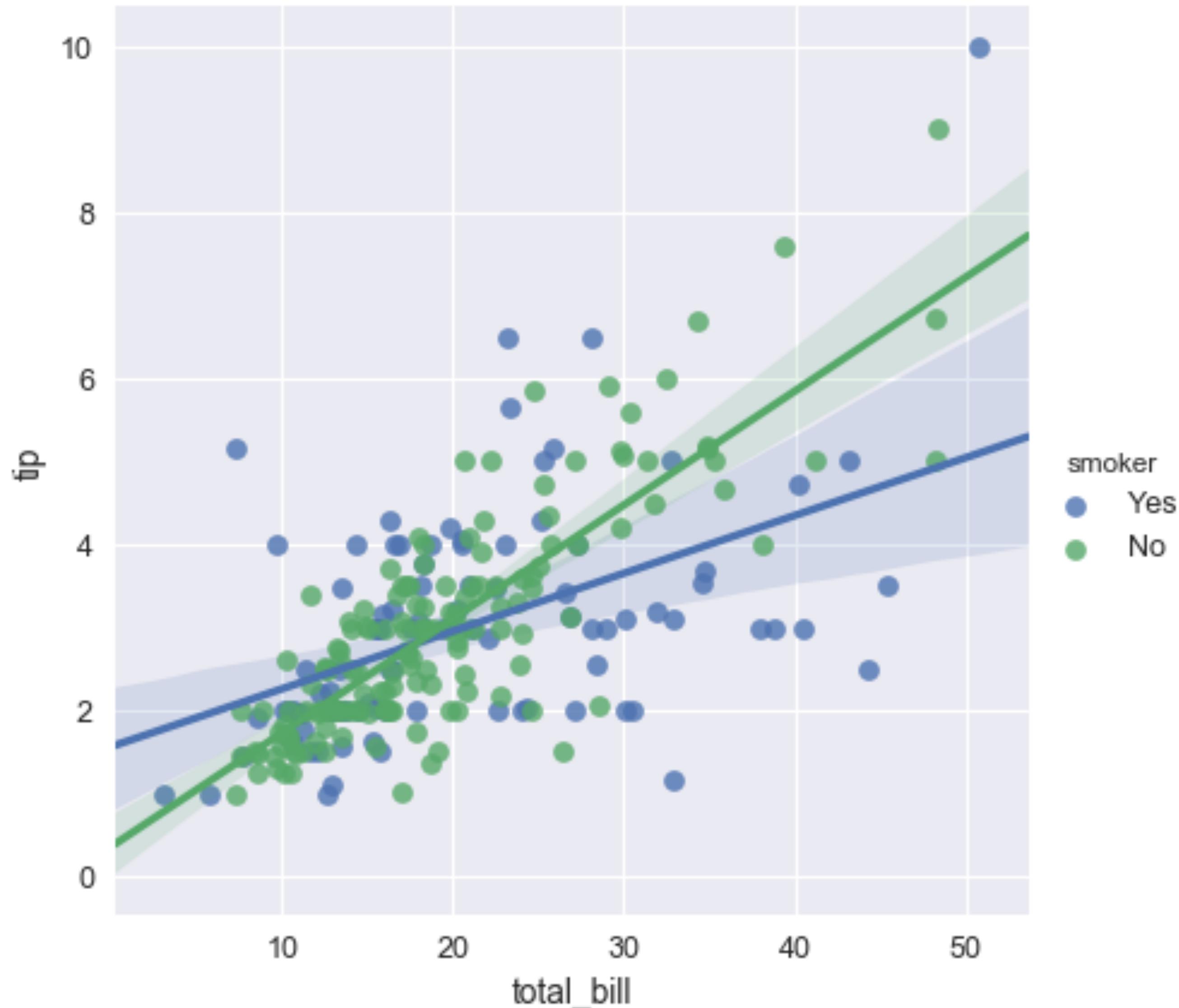


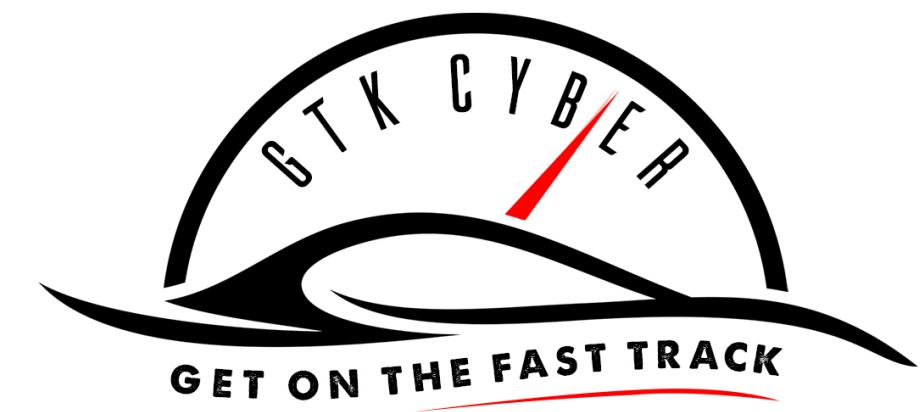
```
ax = sns.regplot(x="total_bill",  
                  y="tip", data=tips)
```



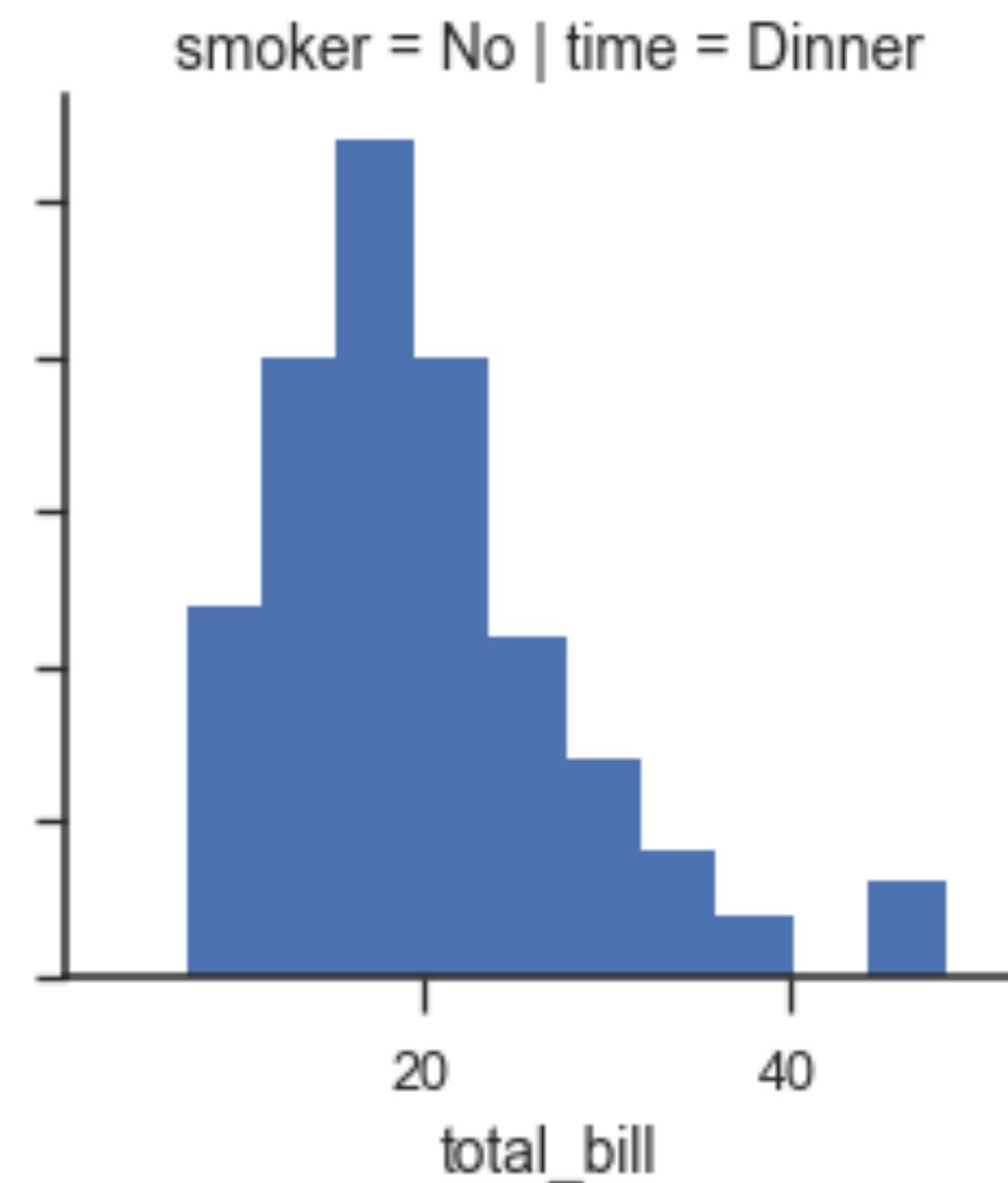
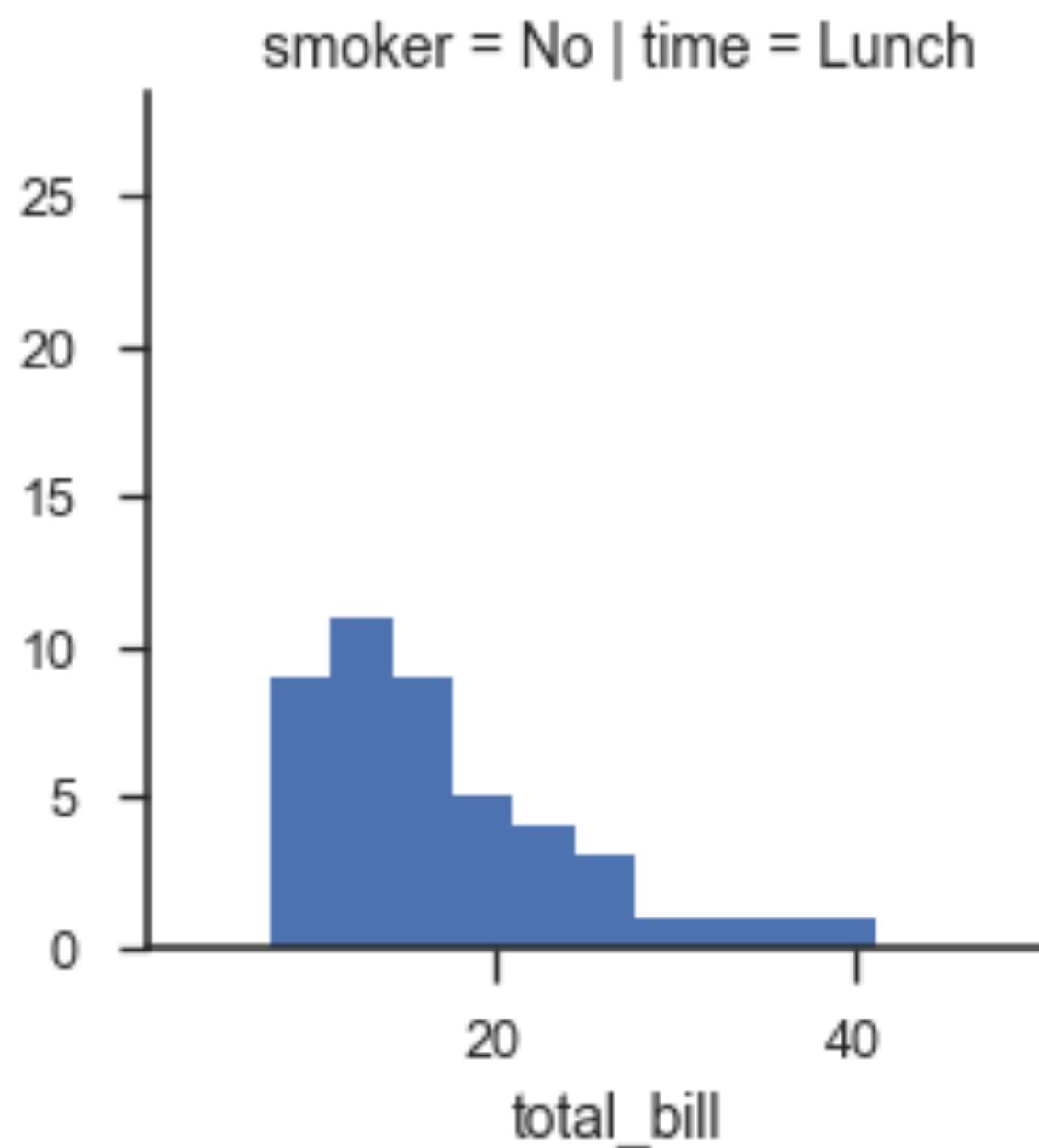
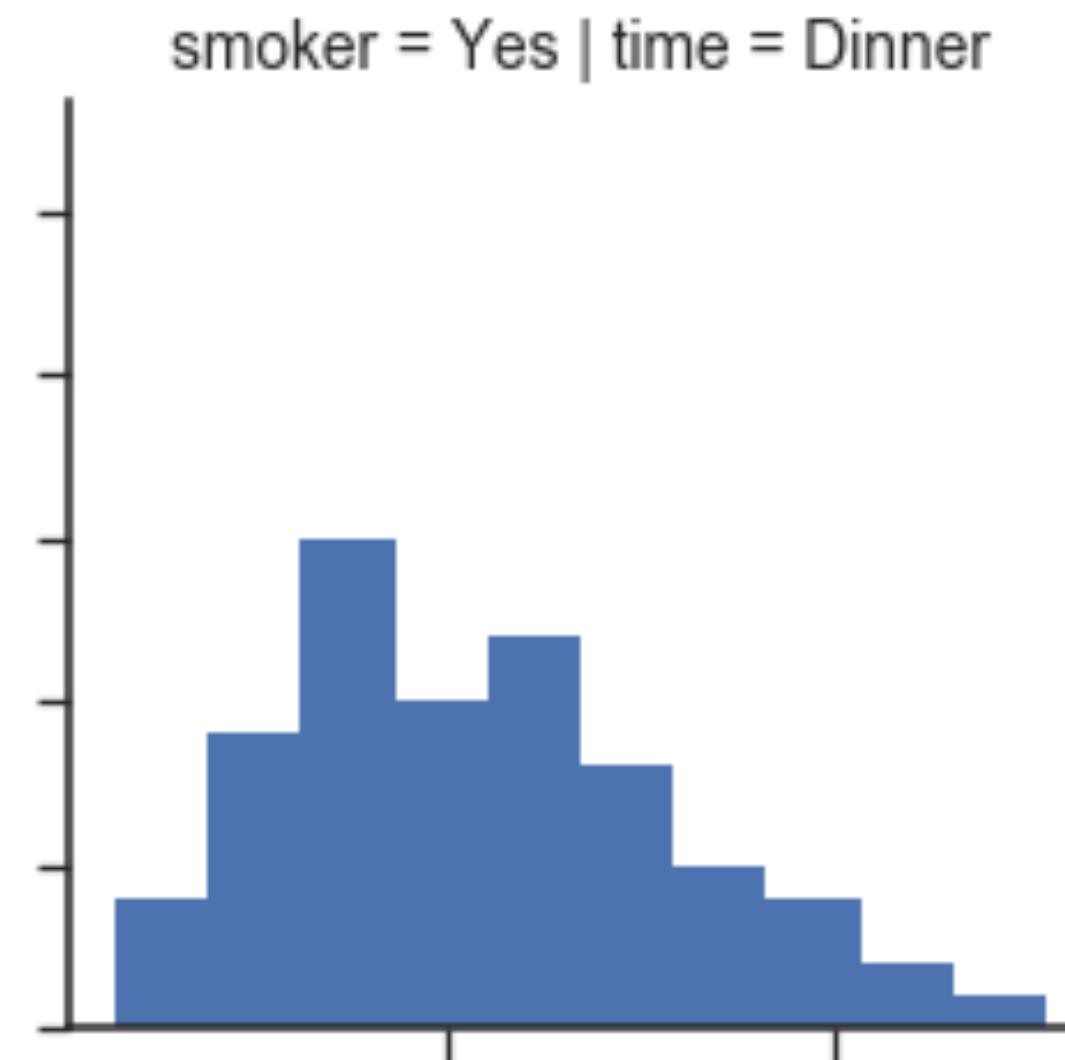
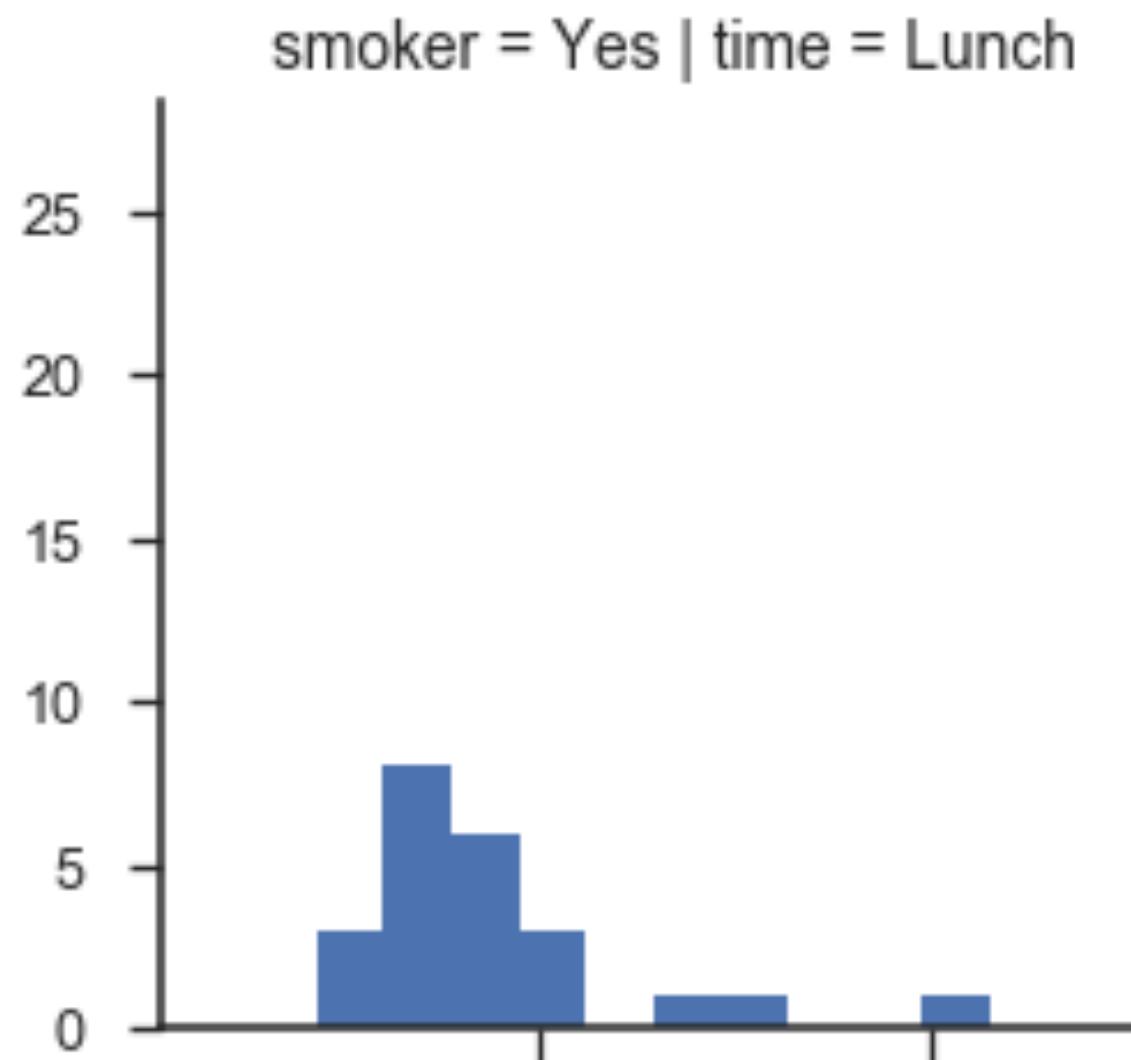


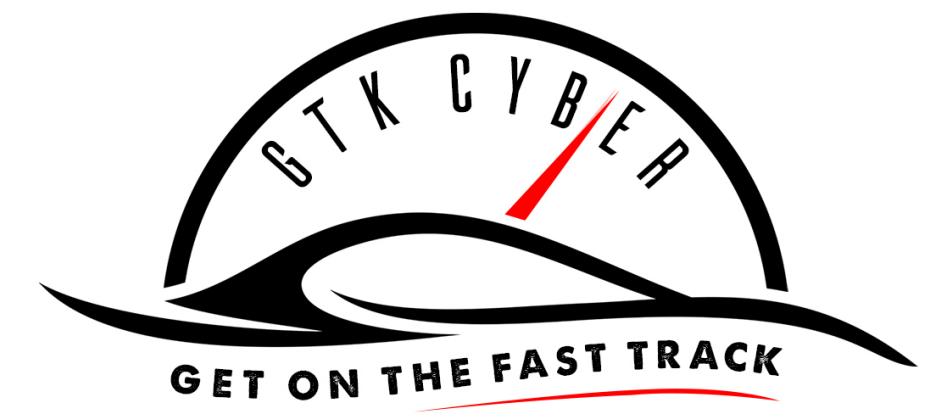
```
g = sns.lmplot(x="total_bill",  
y="tip", hue="smoker", data=tips)
```



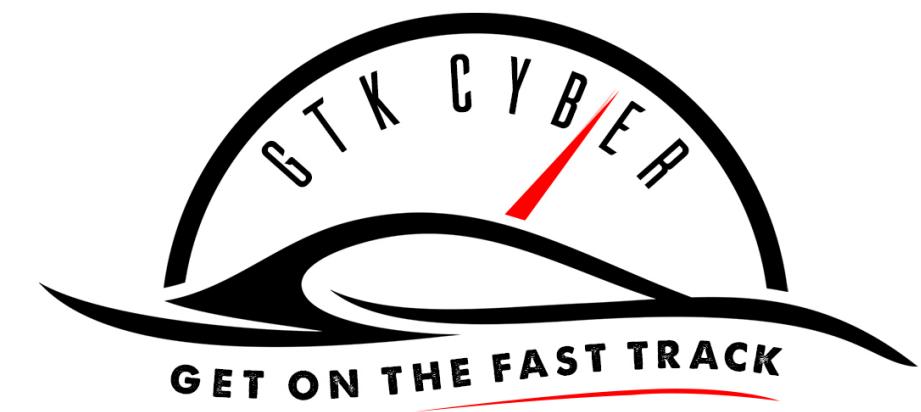


```
g = sns.FacetGrid(tips, col="time", row="smoker")
>>> g = g.map(plt.hist, "total_bill")
```



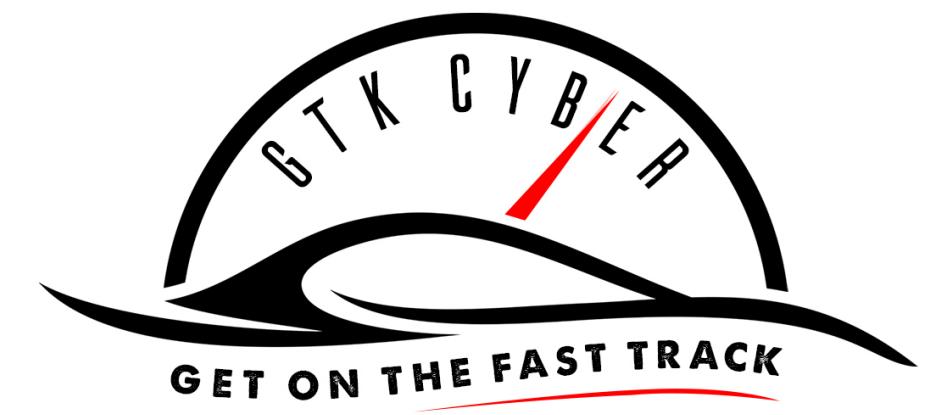


```
scatterPlot.get_figure().savefig( "scatterPlot.png" )
```

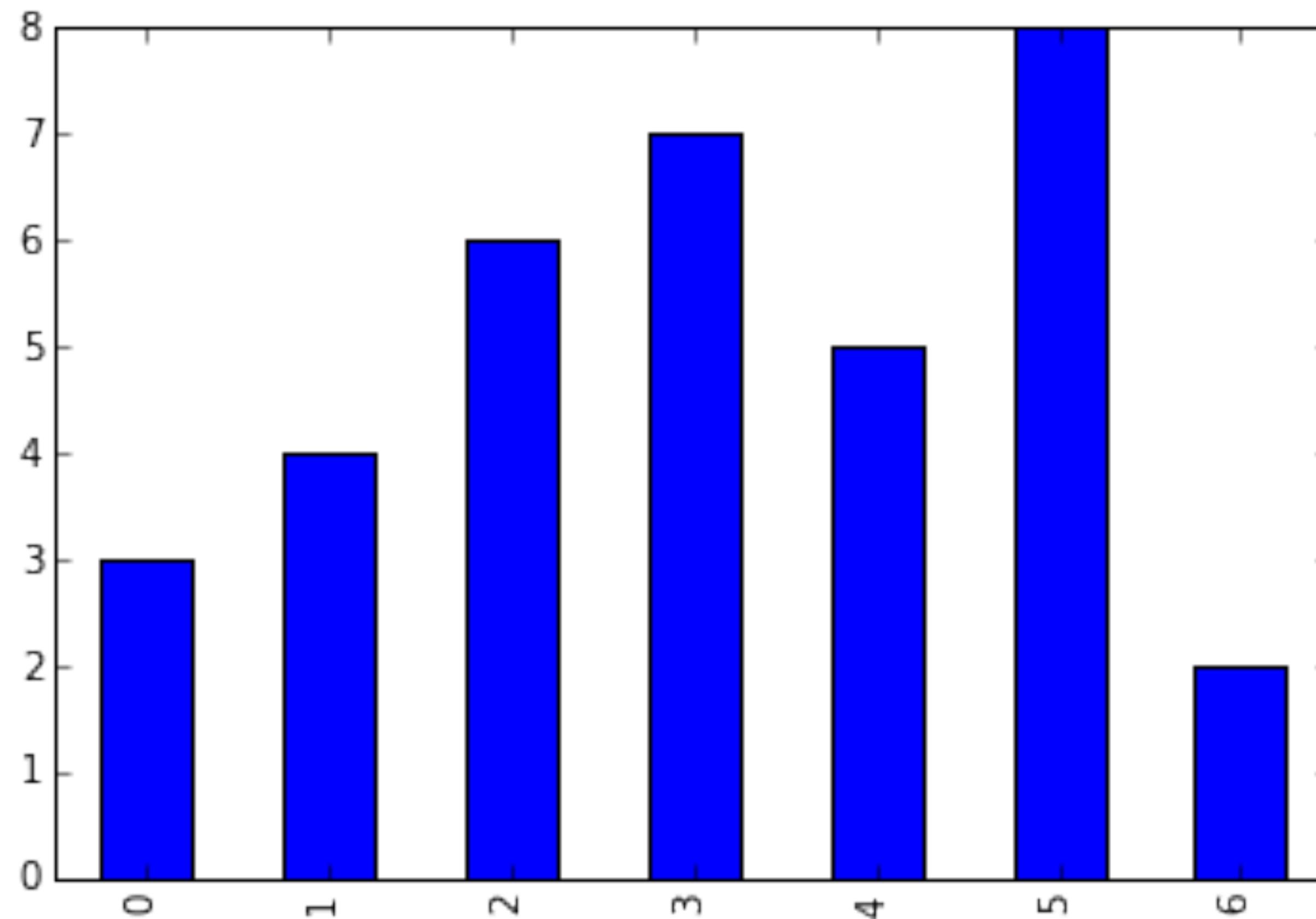
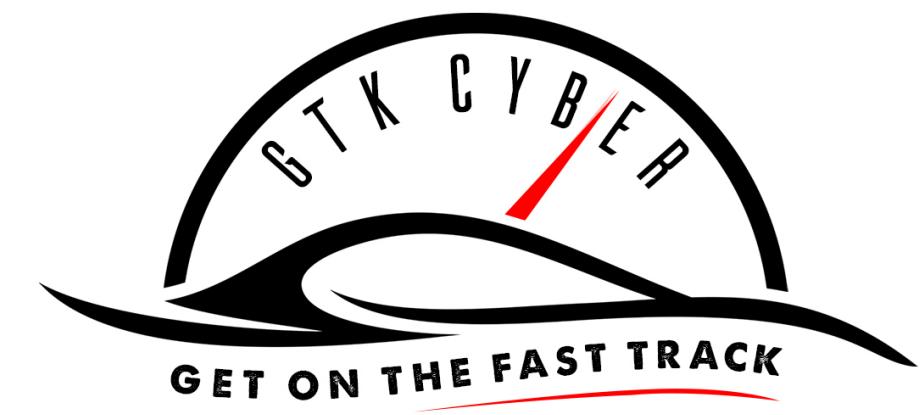


```
print(plt.style.available)

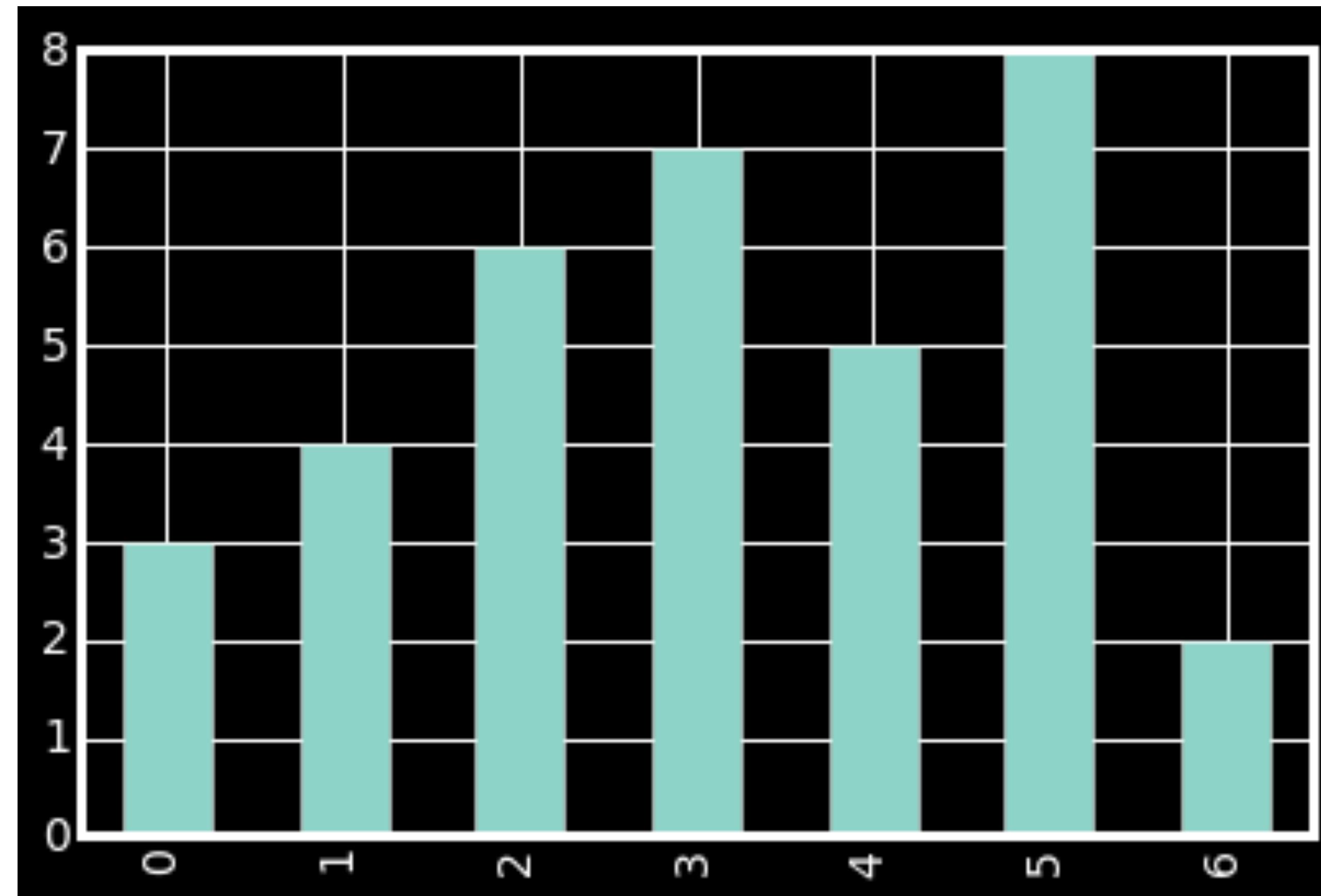
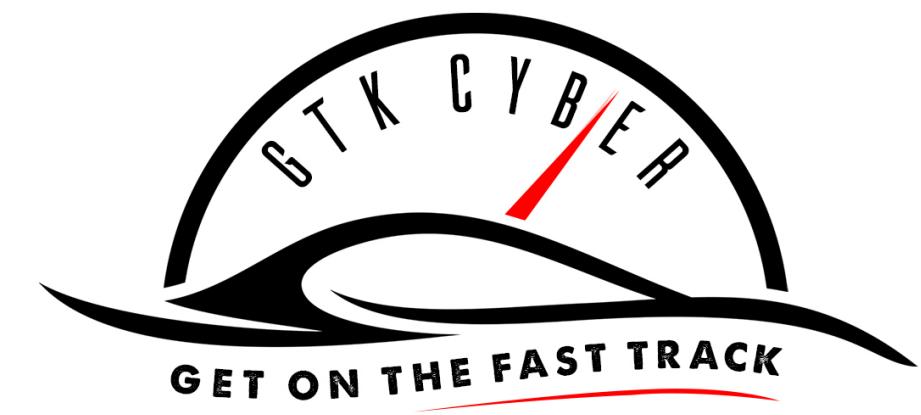
['dark_background', 'grayscale', 'ggplot',
 'bmh', 'fivethirtyeight']
```



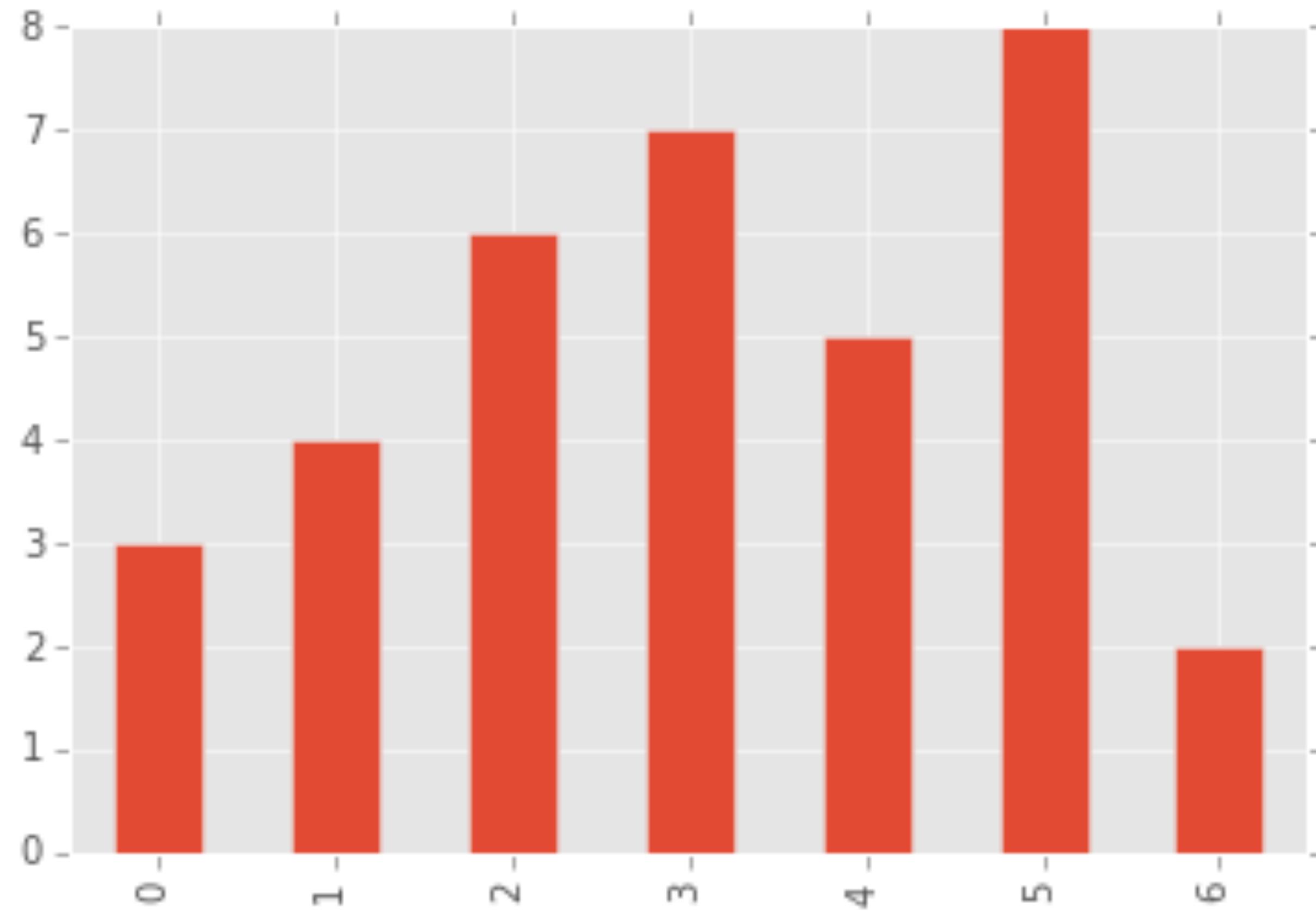
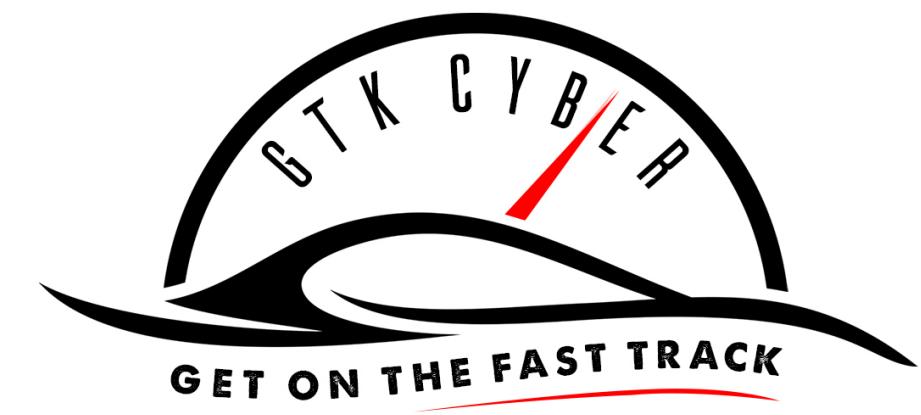
UGLY



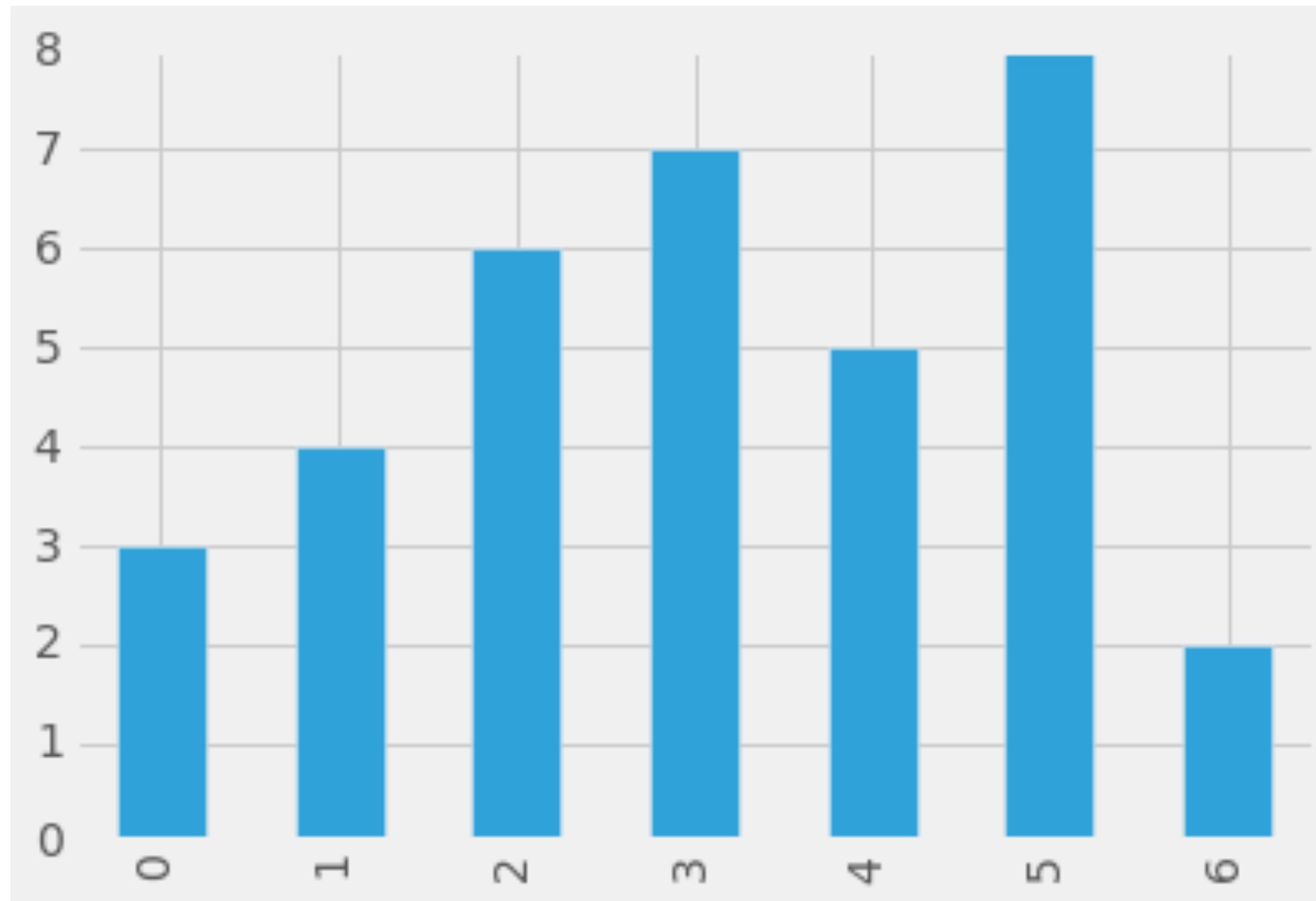
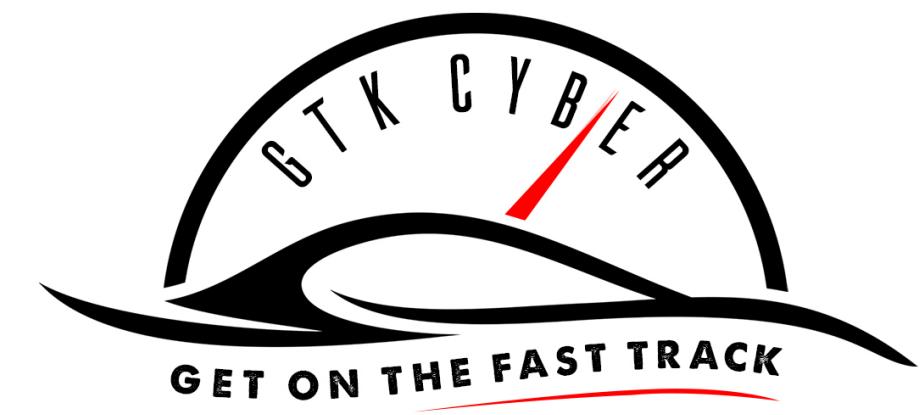
```
data = pd.Series( [3,4,6,7,5,8,2] )
barchart = data.plot( kind="bar" )
```



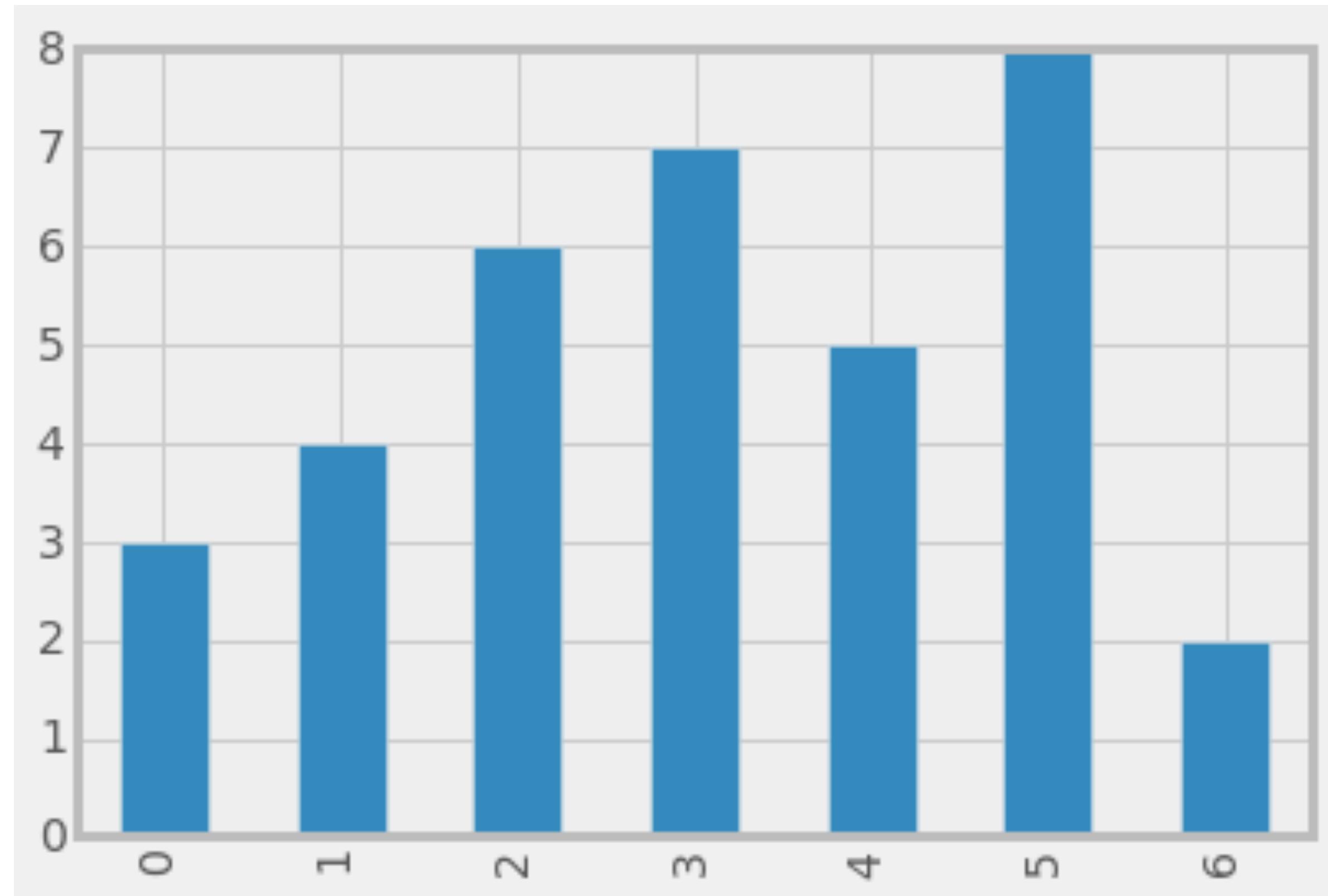
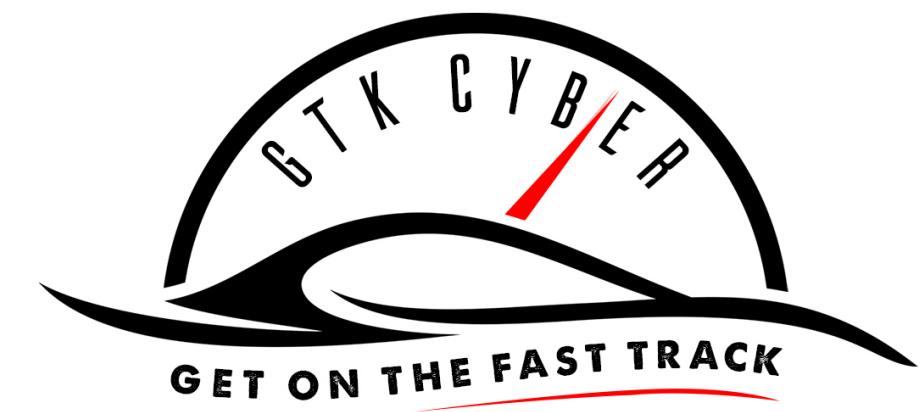
```
plt.style.use('dark_background')
barchart = data.plot( kind="bar" )
```



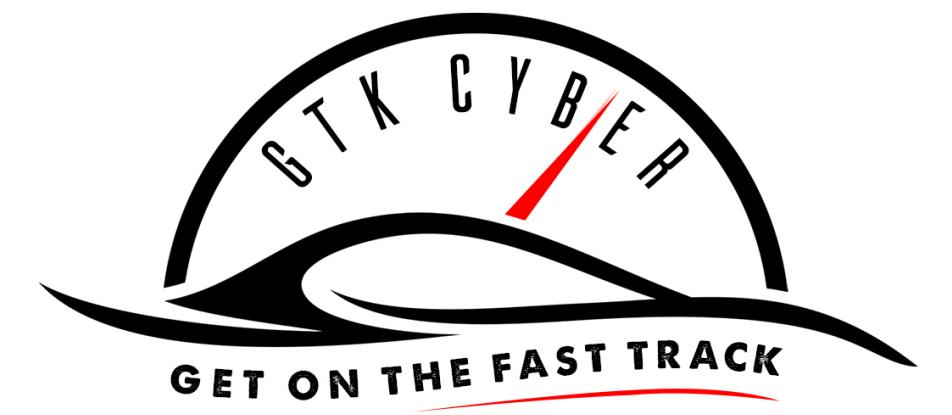
```
plt.style.use('ggplot')
barchart = data.plot( kind="bar" )
```



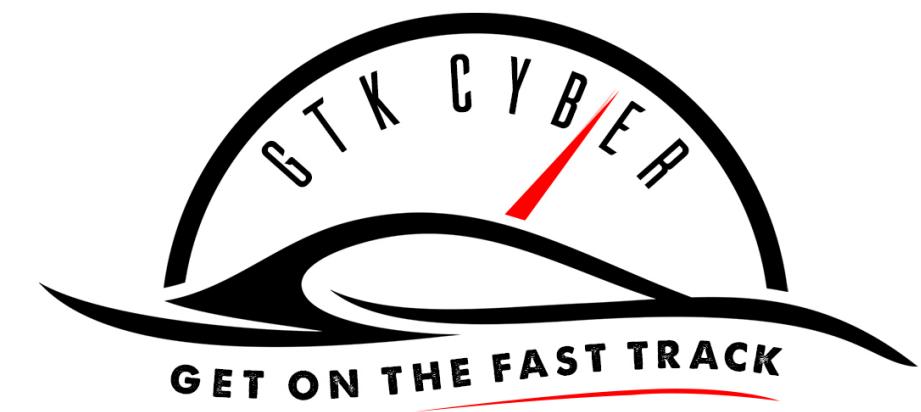
```
plt.style.use('fivethirtyeight')
barchart = data.plot( kind="bar" )
```



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



Interactive Visualizations

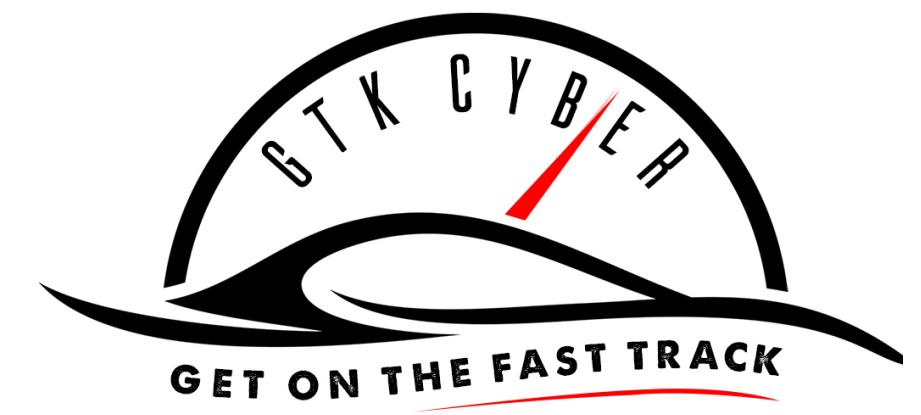


QlikView

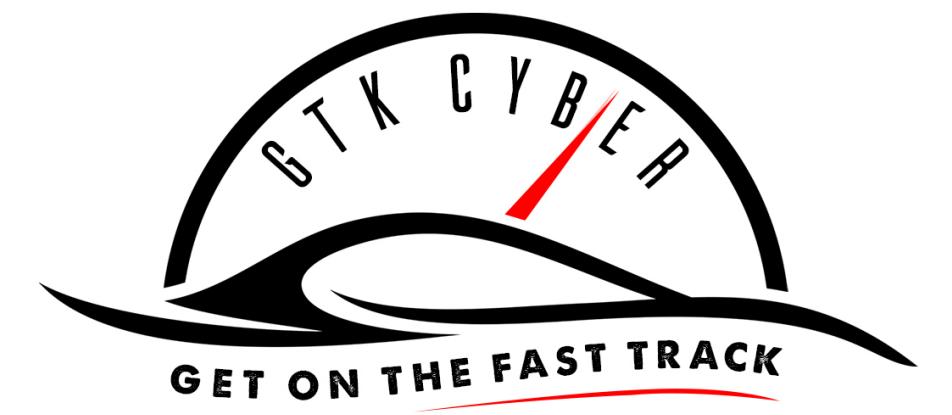




JavaScript



Easy to use... if you know R

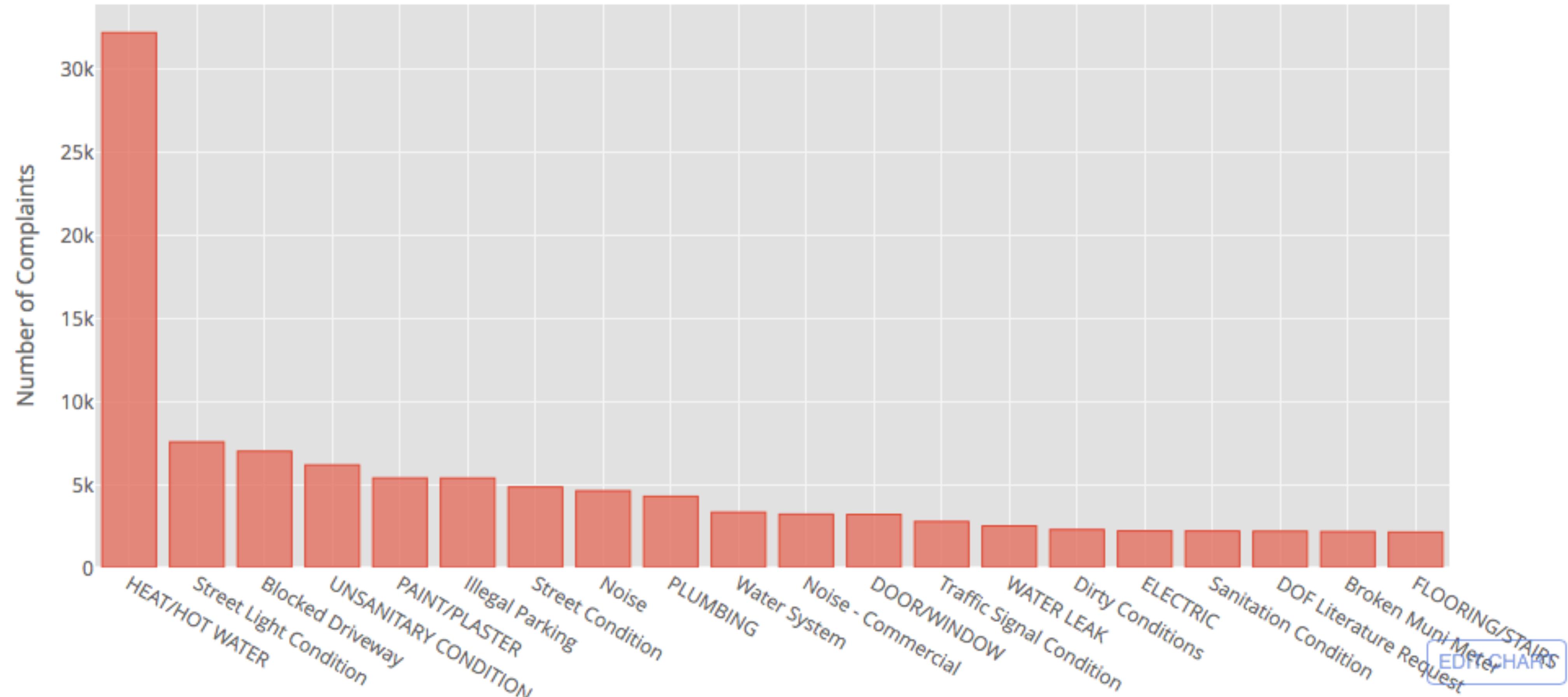


Introducing Cufflinks

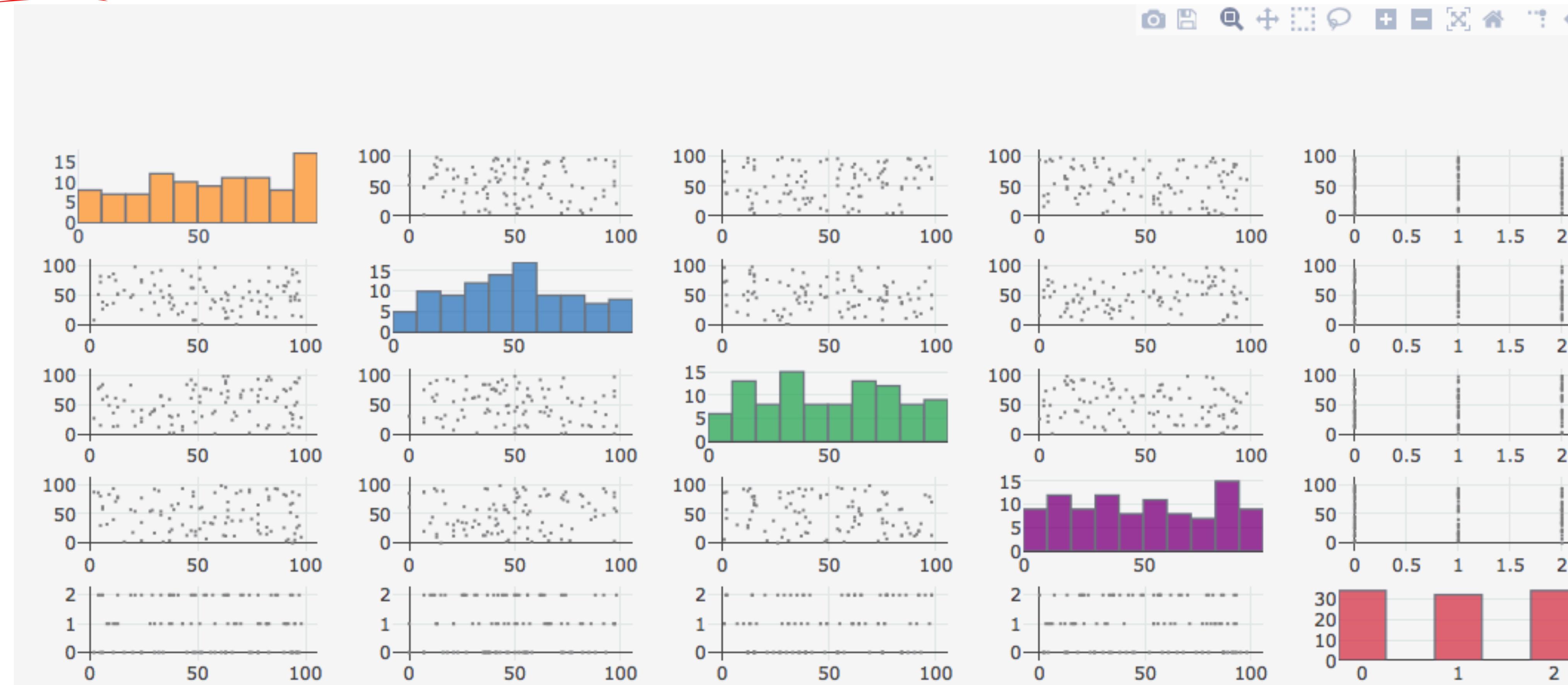




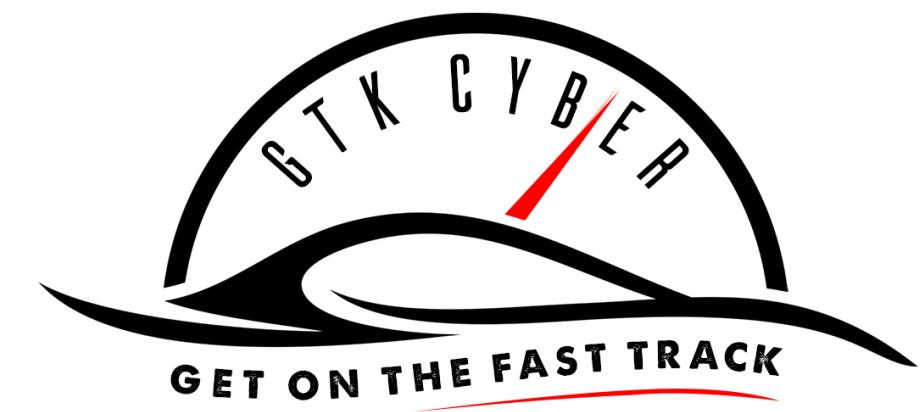
NYC 311 Complaints



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



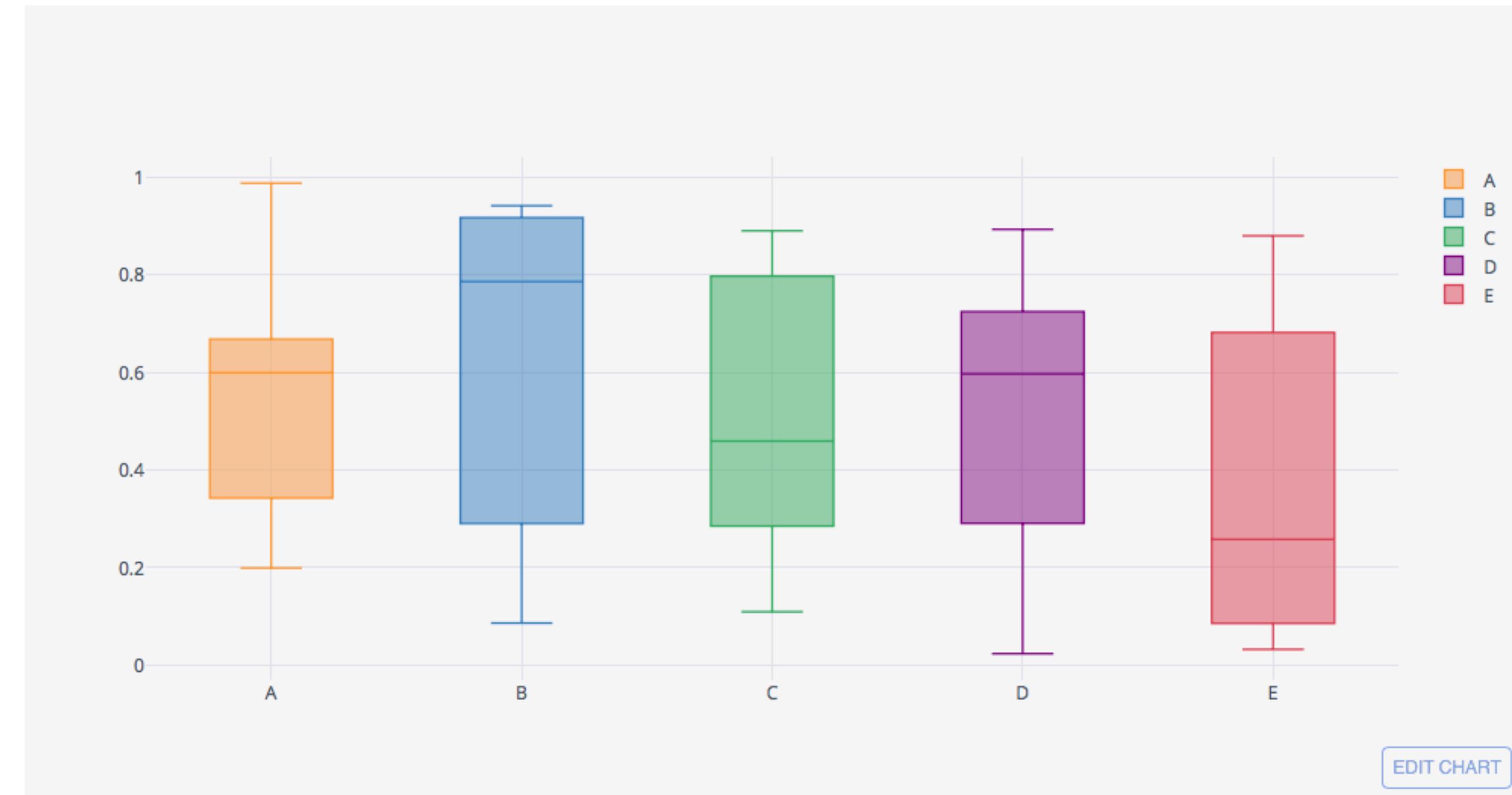
`df.scatter_matrix()`

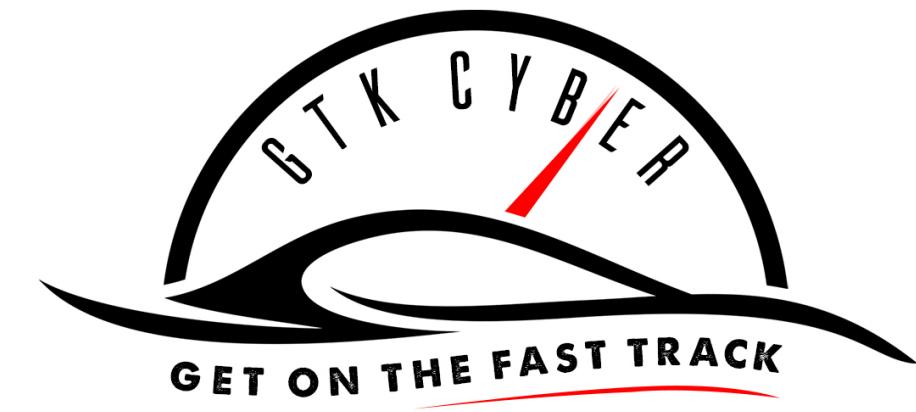


Supported Chart Types

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

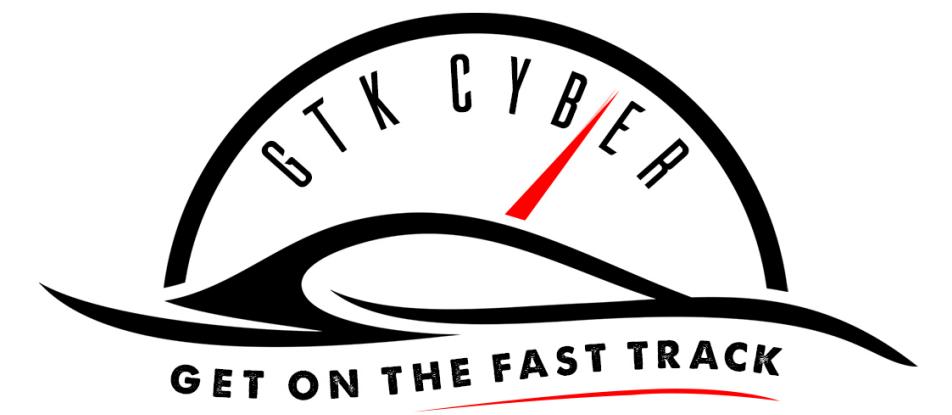
- line
- bar
- histogram
- box (boxplot)
- area
- scatter
- bubble
- heat map





In Class Exercise

Please complete Worksheet 4: Data Visualization



Questions?