Practical Data Visualization & Modeling Methods

VSSR 2018 CRASH COURSE
DAY 1

KATHERINA NGUYEN

WHAT WE'LL BE LEARNING

- **Data Visualization End-to-End Process** *(with perspective from UI/UX and product design)*
- **Core analysis skills & techniques to prepare data and design visualizations from scratch**
- **Foundational overview of information design theory**
- **Understanding when/where to apply visualization skills**
- **Practice storytelling and framing compelling presentations** *(leveraging visualizations to share insights and inspire action)*

EXPECTATIONS IN THE CLASS

- Be active class participants and good teammates
- Do the in-class labs and the homework
- Be open-minded and respectful of other's opinions and of the class learning environment

EXPECTATIONS IN THE CLASS

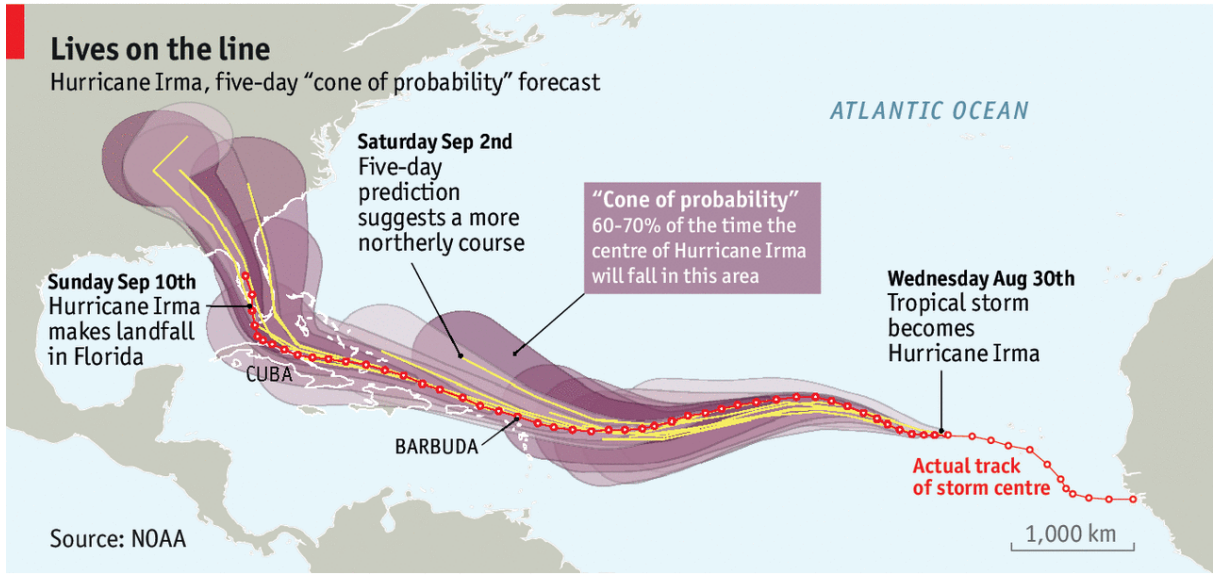
- **1 Individual Data Visualization Presentation**
  (Saturday 8/4)

- **1 Group Project Data Visualization Project**
  (Sunday 8/5)

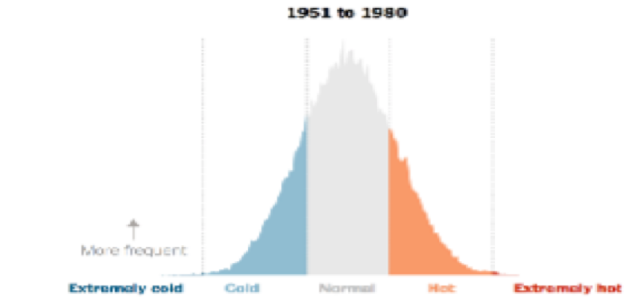
LESSON PLAN

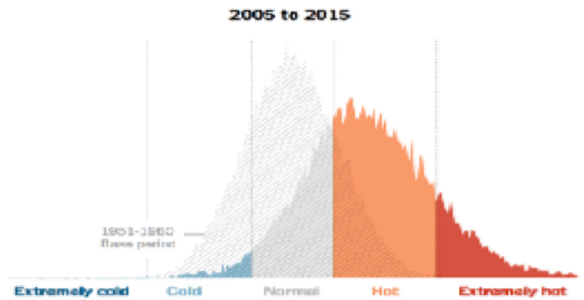
Overview of the Data Visualization Process

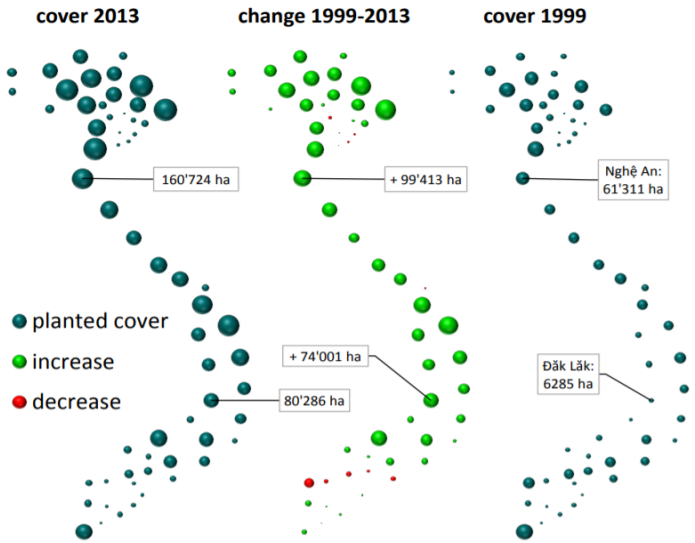
- Data Prep & Understanding (DAY 1)
- Data Exploration (DAY 1, DAY 2)
- Design Strategy (DAY 2)
- Visualization Design & Information Mapping (DAY 2, DAY 3)
- Presenting Insights (DAY 3)

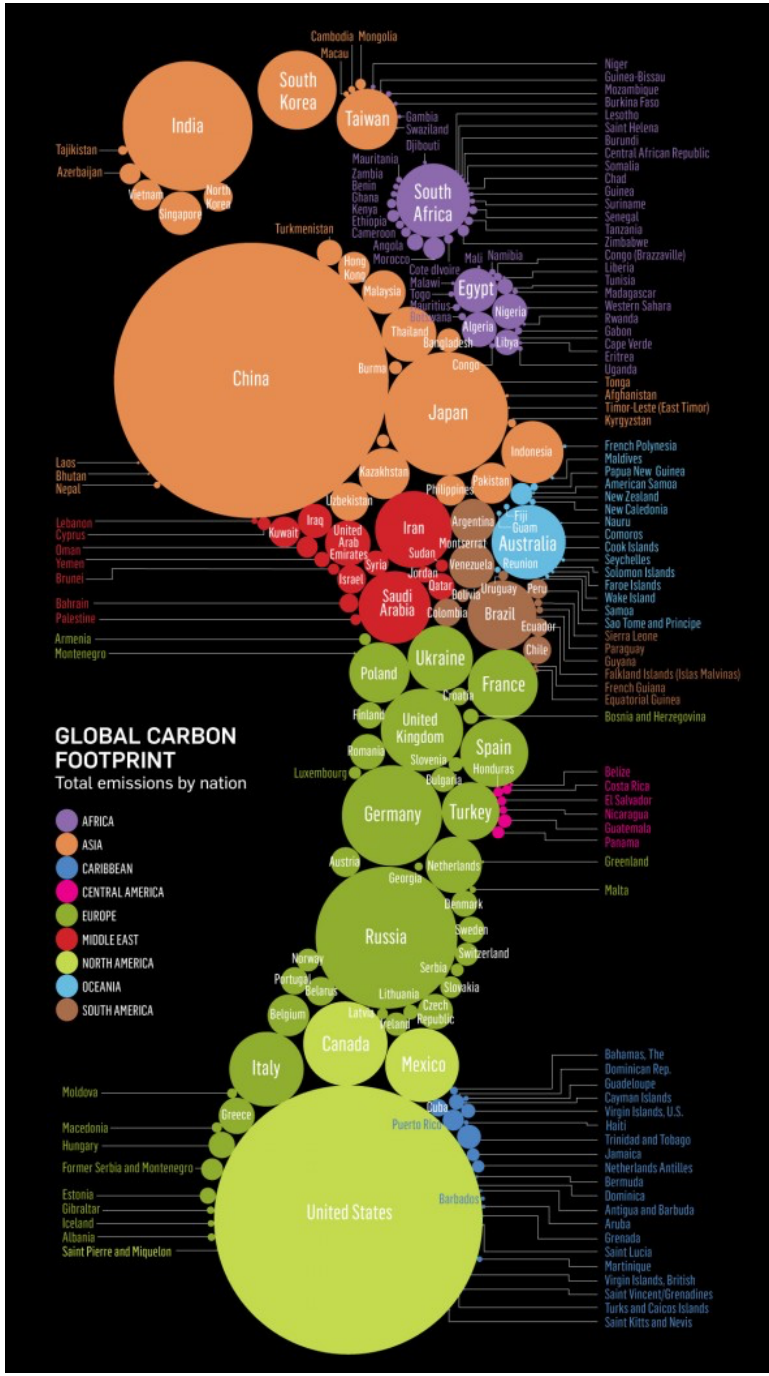


Economist.com

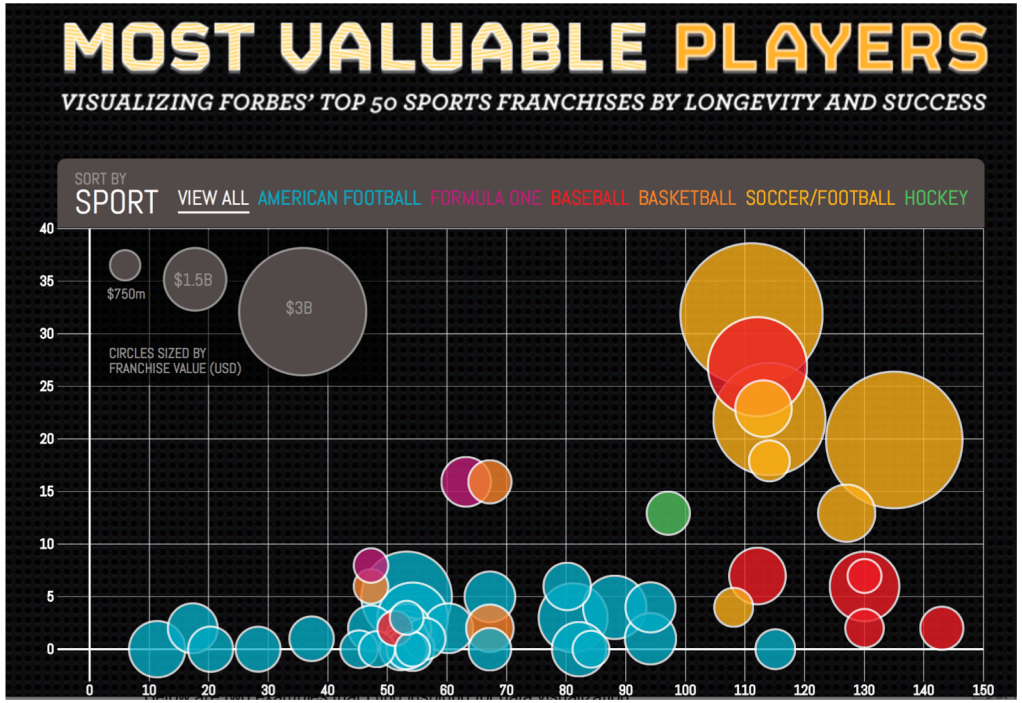


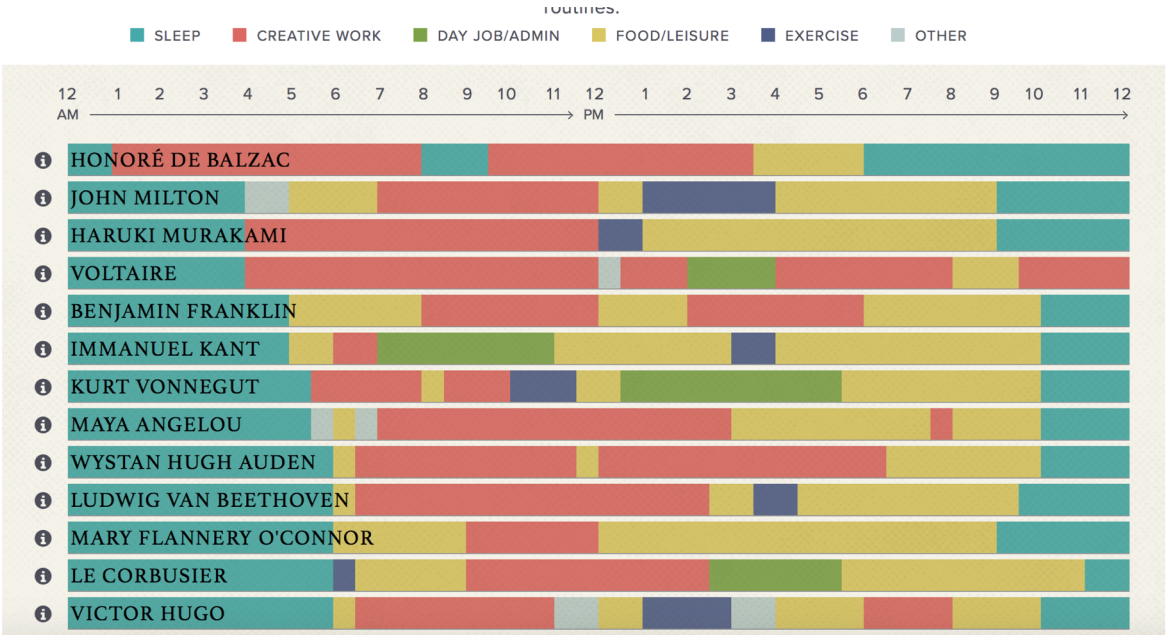







Analysis
Laps
Best distances
Distance
Duration
Pace
Altitude
Heart Rate
03:00
04:00
06:00
12:00
min/km
170
165
160
155
150
145
140
135
130
125
120
bpm
1 km
2 km
3 km
Heart Rate Analysis
Z0 Rest 58 - 123 0m:41s 2%
Z1 Warm Up 124 - 136 9m:40s 34%
Z2 Fat Burning 137 - 149 8m:33s 30%
Z3 Aerobic 150 - 163 9m:21s 33%
Z4 Anaerobic 164 - 176 0m:0s 0%
Z5 Maximum 177 - 190 0m:0s 0%

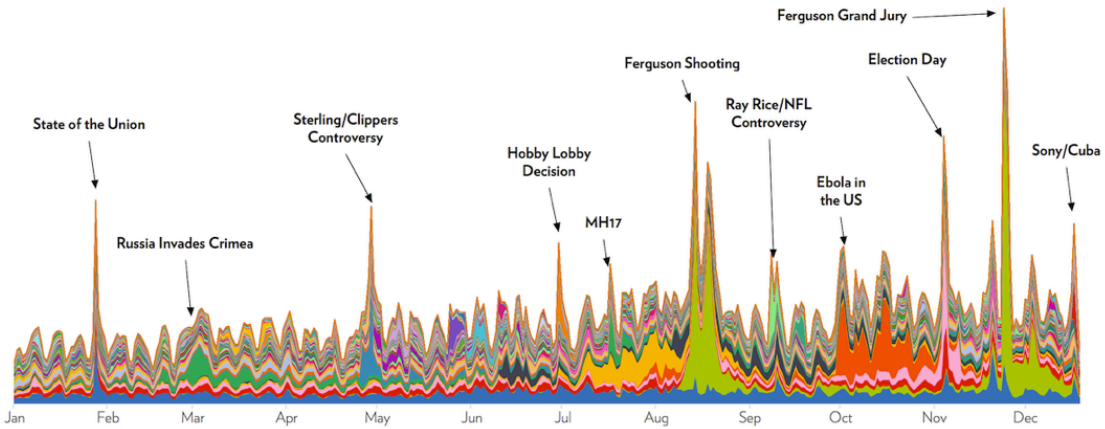

MOST VALUABLE PLAYERS
VISUALIZING FORBES' TOP 50 SPORTS FRANCHISES BY LONGEVITY AND SUCCESS
SORT BY
SPORT
VIEW ALL
AMERICAN FOOTBALL
FORMULA ONE
BASEBALL
BASKETBALL
SOCCER/FOOTBALL
HOCKEY
$750m
$1.5B
$3B
CIRCLES SIZED BY FRANCHISE VALUE (USD)
0
5
10
15
20
25
30
35
40
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150

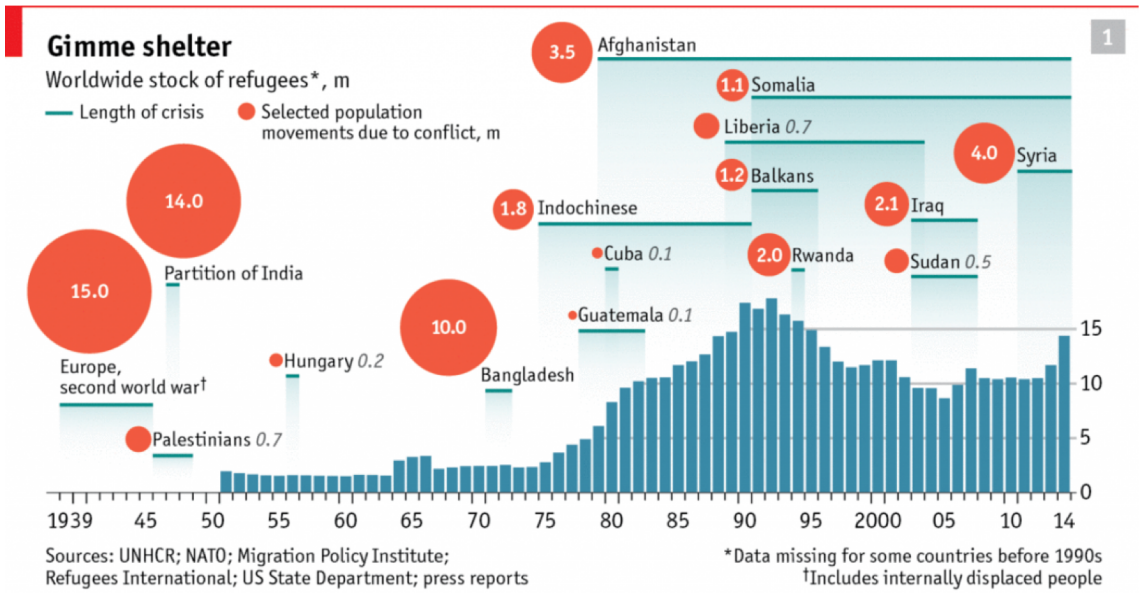

routines.
SLEEP
CREATIVE WORK
DAY JOB/ADMIN
FOOD/LEISURE
EXERCISE
OTHER
12 AM 1 2 3 4 5 6 7 8 9 10 11 12 PM 1 2 3 4 5 6 7 8 9 10 11 12
HONORÉ DE BALZAC
JOHN MILTON
HARUKI MURAKAMI
VOLTAIRE
BENJAMIN FRANKLIN
IMMANUEL KANT
KURT VONNEGUT
MAYA ANGELOU
WYSTAN HUGH AUDEN
LUDWIG VAN BEETHOVEN
MARY FLANNERY O'CONNOR
LE CORBUSIER
VICTOR HUGO


1193
1143
1093
1043
993
943
893
843
793
Nov 2017
Dec 2017
Jan 2018
Feb 2018
Mar 2018
Apr 2018

THE YEAR IN NEWS from ECHELON INSIGHTS

What America talked about in 2014, as viewed through 184.5 million Twitter mentions.



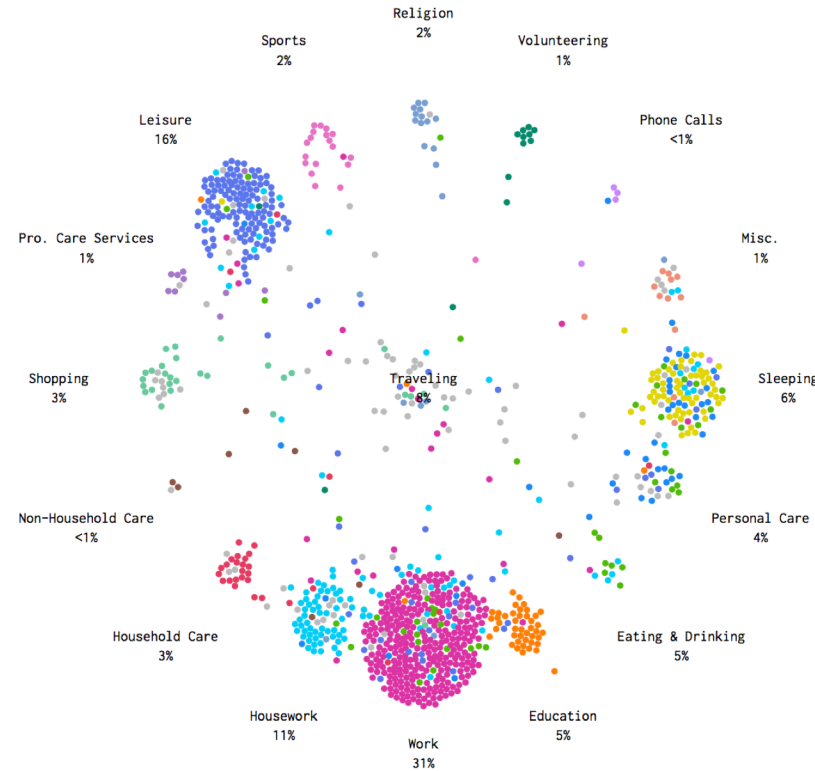


10:02am

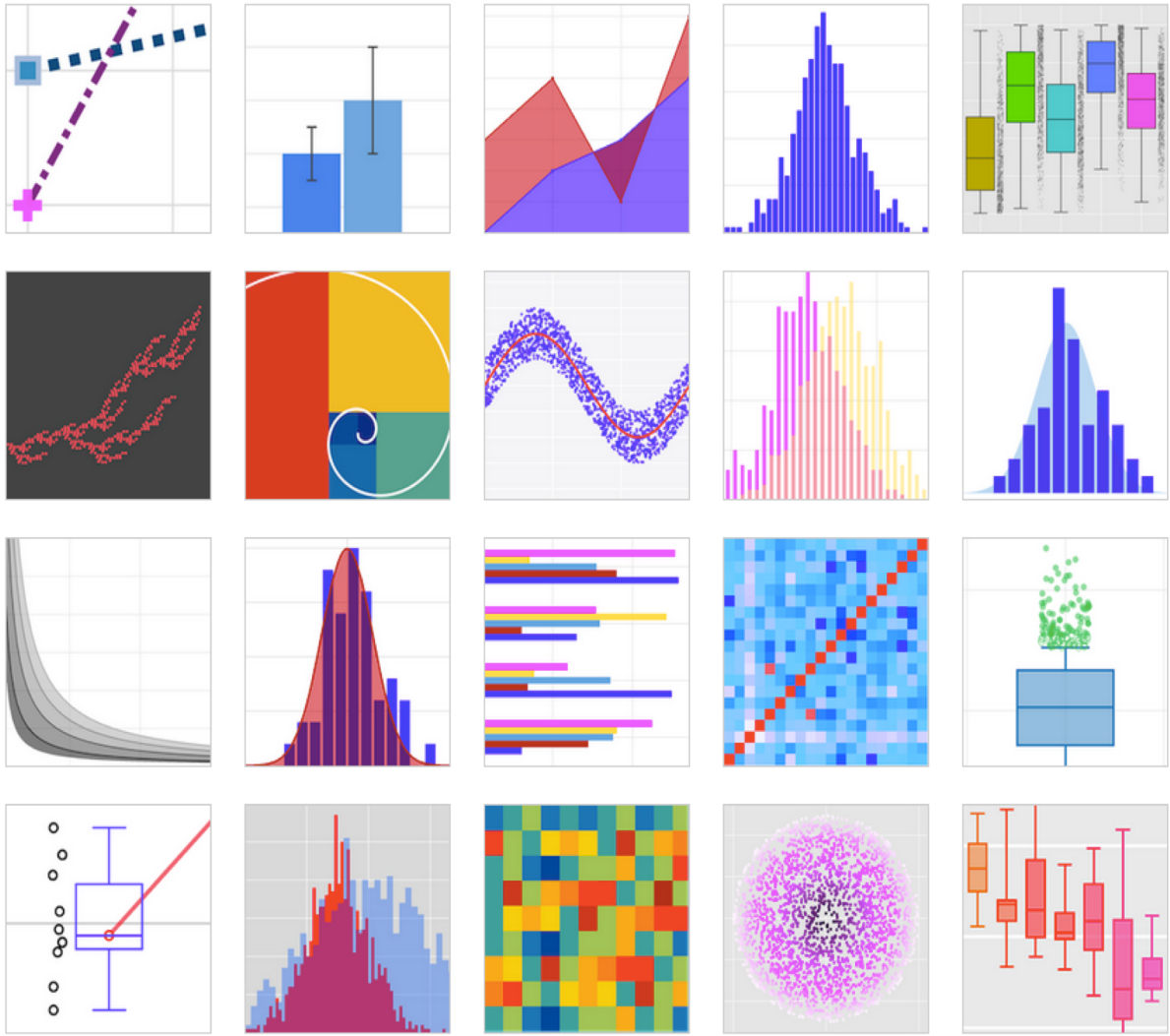
SLOW | MEDIUM | FAST

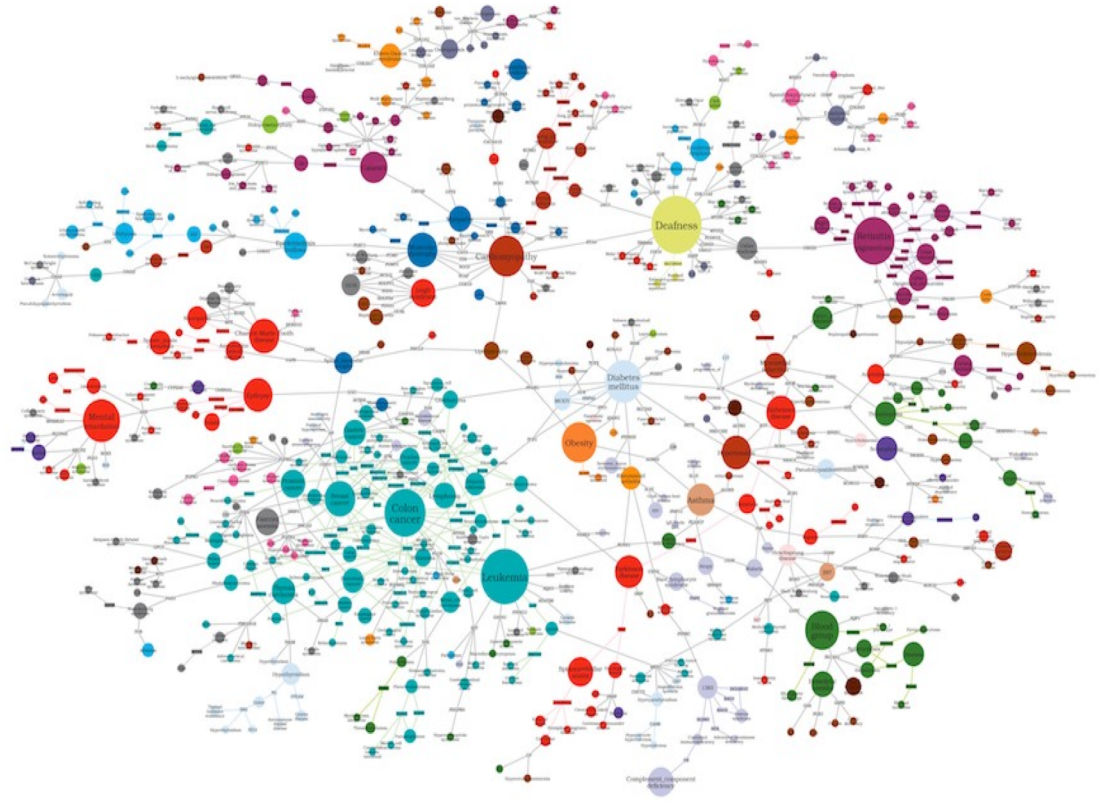
The day is in full swing with work or housework. Stores and services are open so people can run errands, and they take various forms of transportation to get there.

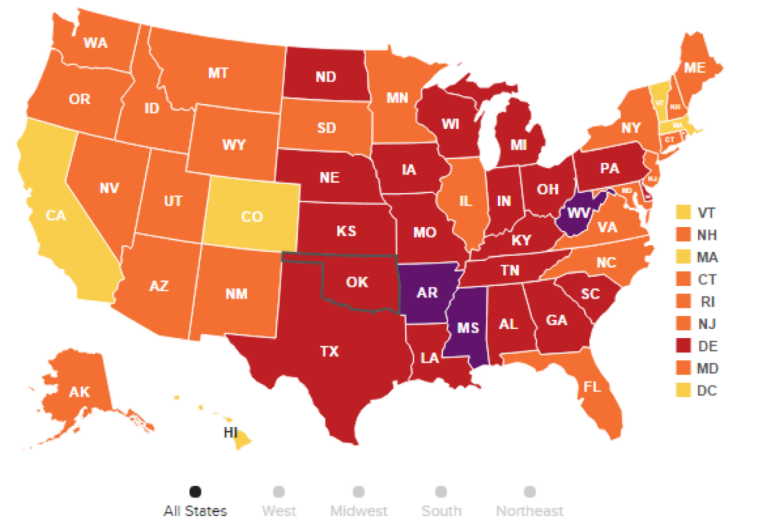
This is a simulation of 1,000 people's average day. It's based on 2014 data from the American Time Use Survey, made way more accessible by the ATUS Extract Builder.



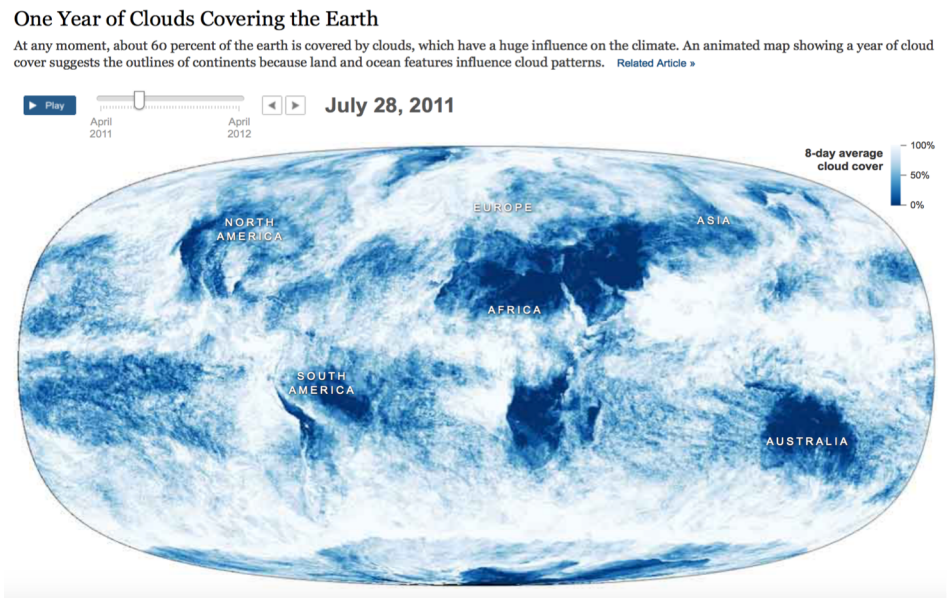
What is a data visualization?

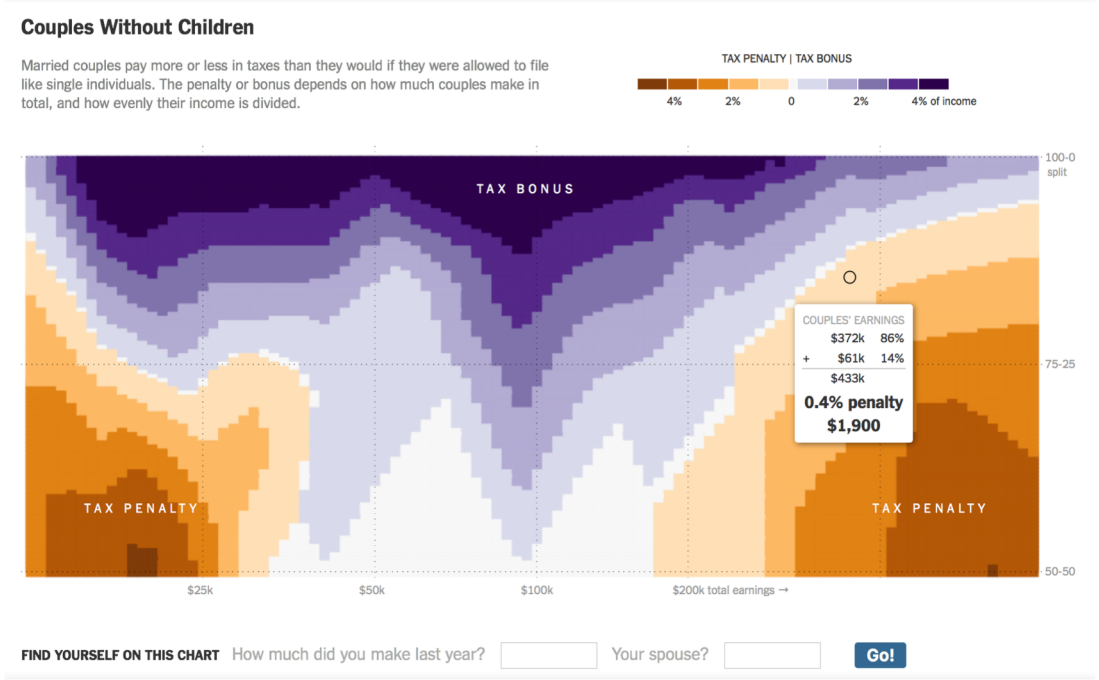


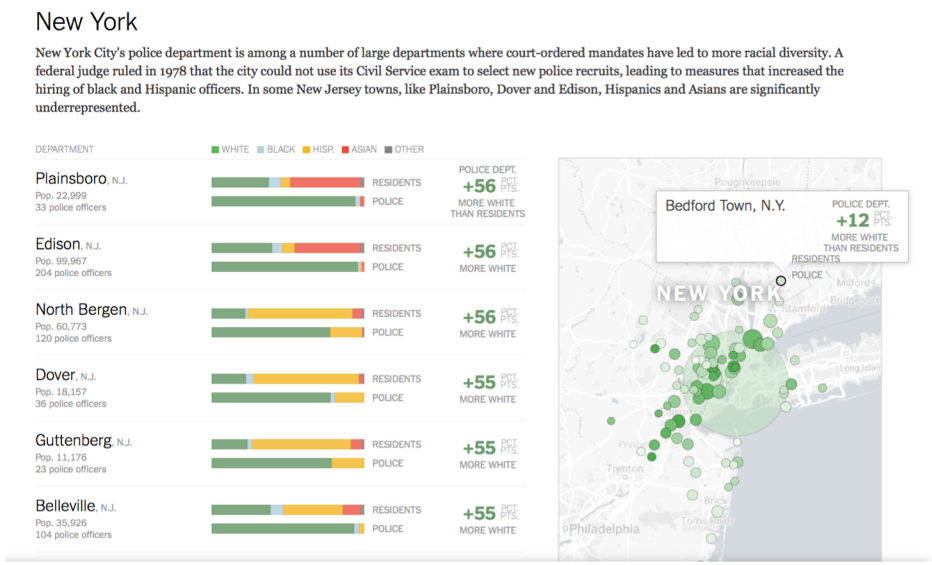

Deafness
Diabetes mellitus
Obesity
Asthma
Colon cancer
Leukemia

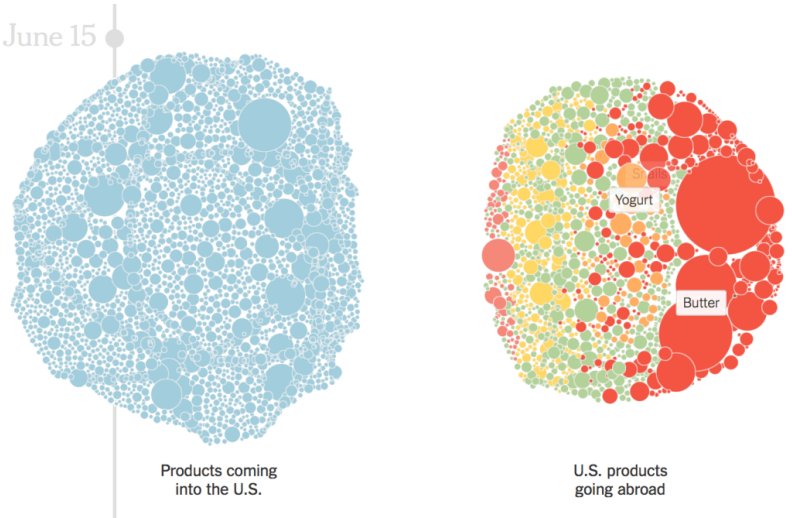

WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
VA
NC
SC
GA
FL
PA
NY
ME
AK
HI
VT
NH
MA
CT
RI
NJ
DE
MD
DC
All States
West
Midwest
South
Northeast

NYTimes Interactive Data Articles





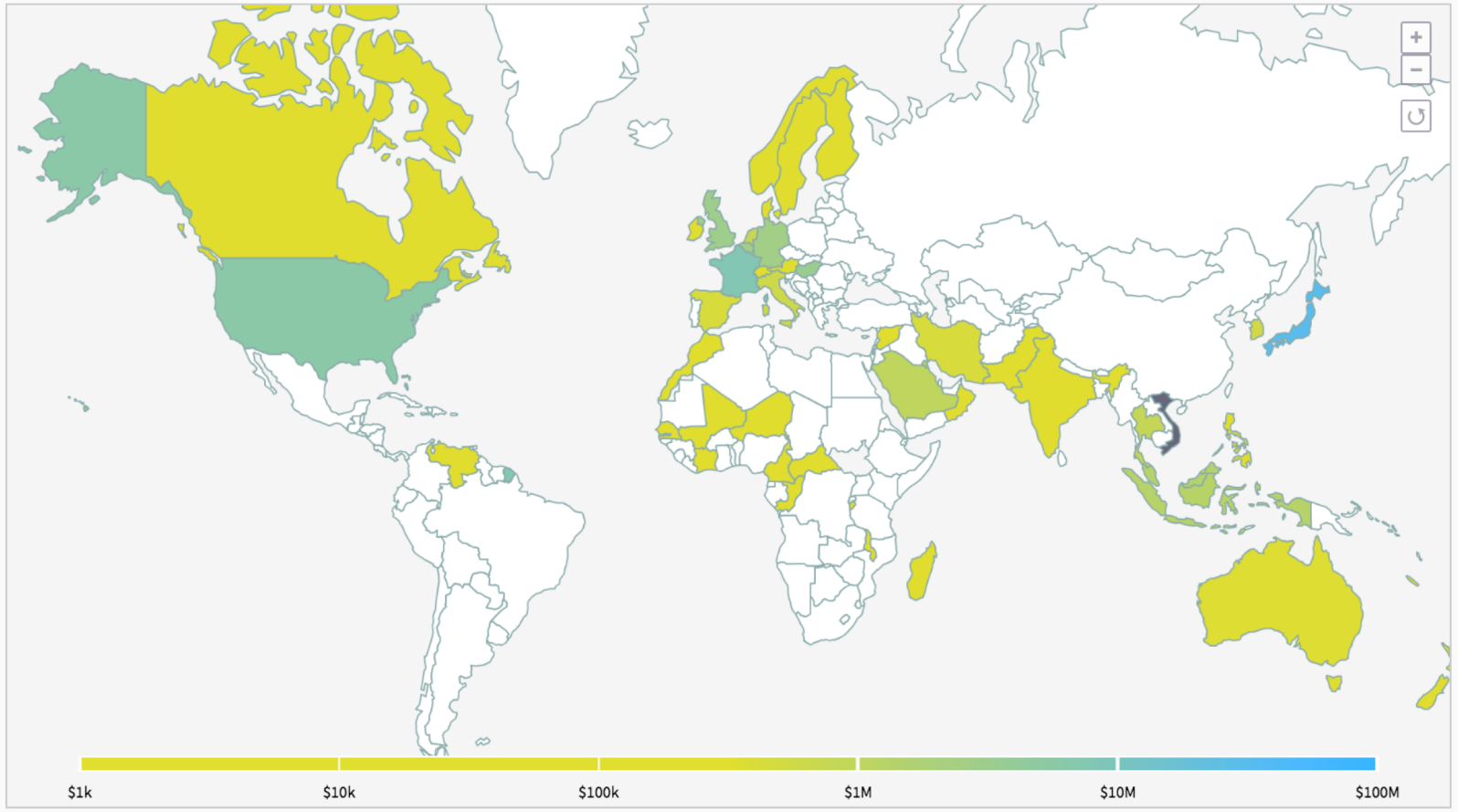




Where did Vietnam export to in 1975?

Share Download Feedback

$116M



Please select a country, a product, or a combination of the two.

Vietnam (VNM)

Exports | Imports

Select a SITC product

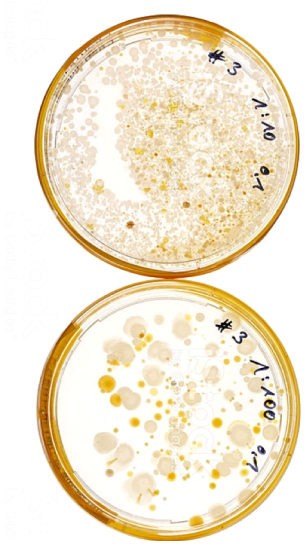
Trade Data Visualization

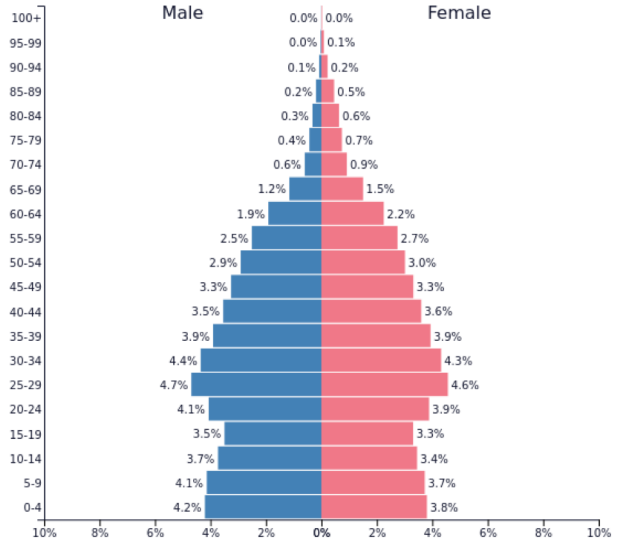
Complexity Visualization

Geo map

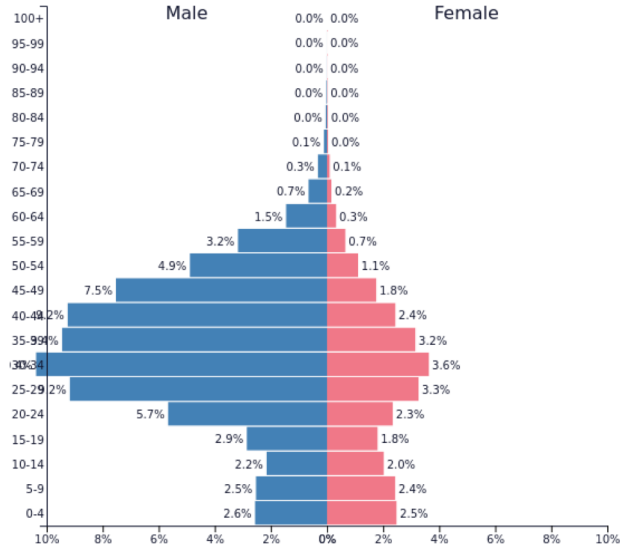
The geo map displays the shares of country trade with other countries or by product, in a given year. **Click on a country for more information or generate variations using the options below.**

Learn more

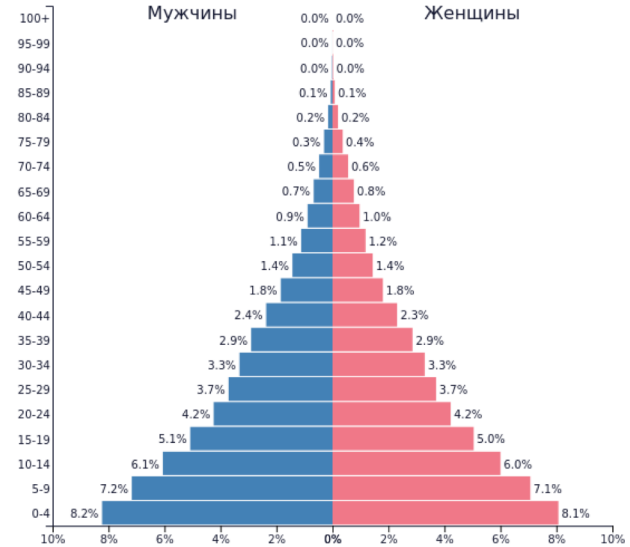




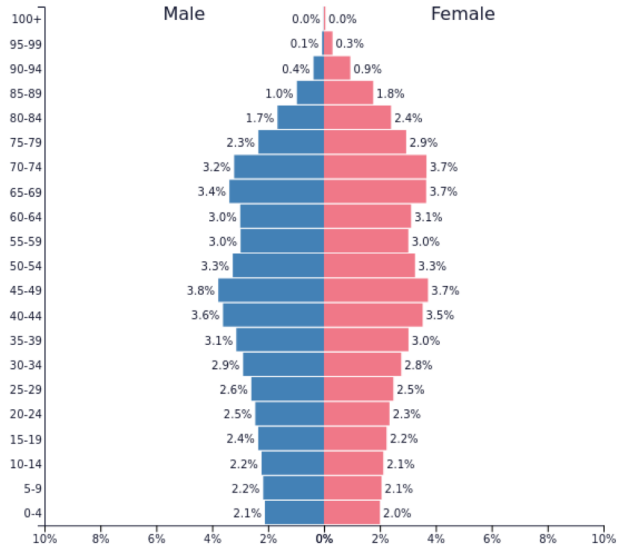
Viet Nam - 2018
Population: **96,356,744**



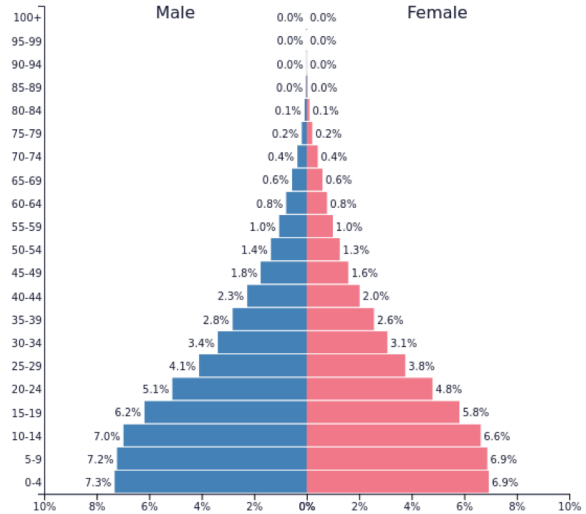
United Arab Emirates - 2018
Population: **9,542,843**



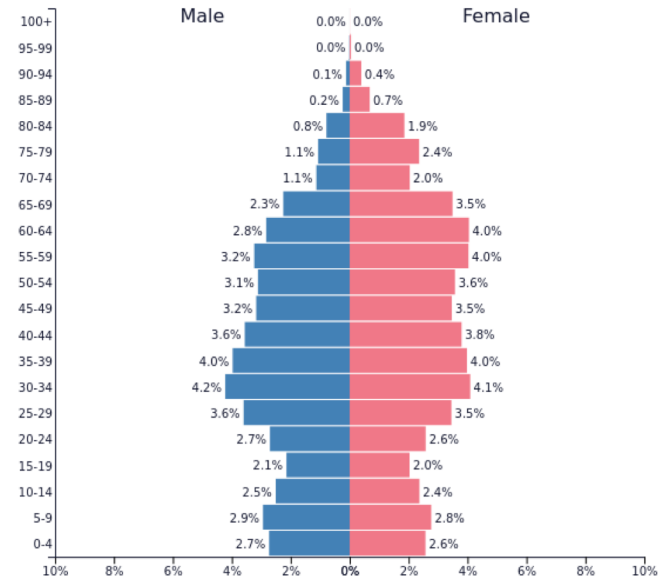
Конго - 2016
Население: **4,740,992**



Japan - 2018
Population: **125,738,357**



Afghanistan - 2018
Population: **34,935,712**



Ukraine - 2018
Population: **44,170,002**

XXII СЪЕЗД КПСС

ПРЕВРАЩЕНИЕ СИСТЕМЫ СОЦИАЛИЗМА В РЕШАЮЩИЙ ФАКТОР МИРОВОГО РАЗВИТИЯ

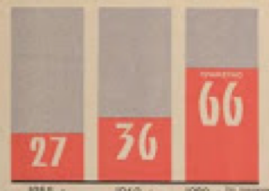
Мировая социалистическая система успешно развивается, крепнет и становится определяющим фактором прогресса человеческого общества.

Из Резолюции XXII съезда Коммунистической партии Советского Союза

ДОЛЯ СОЦИАЛИСТИЧЕСКОГО МИРА В ТЕРРИТОРИИ И НАСЕЛЕНИИ ЗЕМНОГО ШАРА

ДОЛЯ СОЦИАЛИСТИЧЕСКИХ СТРАН В МИРОВОМ ПРОМЫШЛЕННОМ ПРОИЗВОДСТВЕ

(В ПРОЦЕНТАХ)

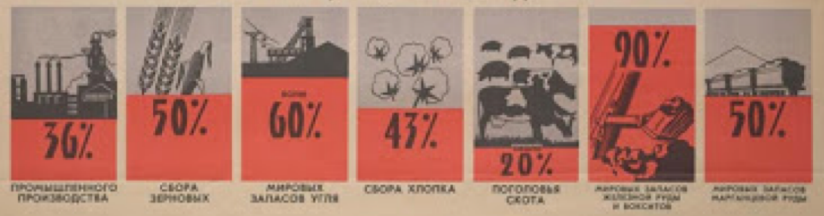


ДОЛЯ ПРОМЫШЛЕННОЙ ПРОДУКЦИИ ВО ВСЕМ НАРОДНОМ ХОЗЯЙСТВЕ В СТРАНАХ СОЦИАЛИСТИЧЕСКОГО ЛАГЕРЯ СОСТАВЛЯЕТ В СРЕДНЕМ ОКОЛО 75%

УДЕЛЬНЫЙ ВЕС СОЦИАЛИСТИЧЕСКОГО СЕКТОРА ВО ВСЕЙ ПЛОЩАДИ СЕЛЬСКОХОЗЯЙСТВЕННЫХ УГОДИЙ СОЦИАЛИСТИЧЕСКИХ СТРАН СОСТАВЛЯЕТ СВЫШЕ 90%

РОСТ ПРОМЫШЛЕННОГО ПРОИЗВОДСТВА В СТРАНАХ СОЦИАЛИЗМА И КАПИТАЛИЗМА

МИРУ СОЦИАЛИЗМА ПРИНАДЛЕЖИТ:



ГОСПОЛИТИЗДАТ · 1963

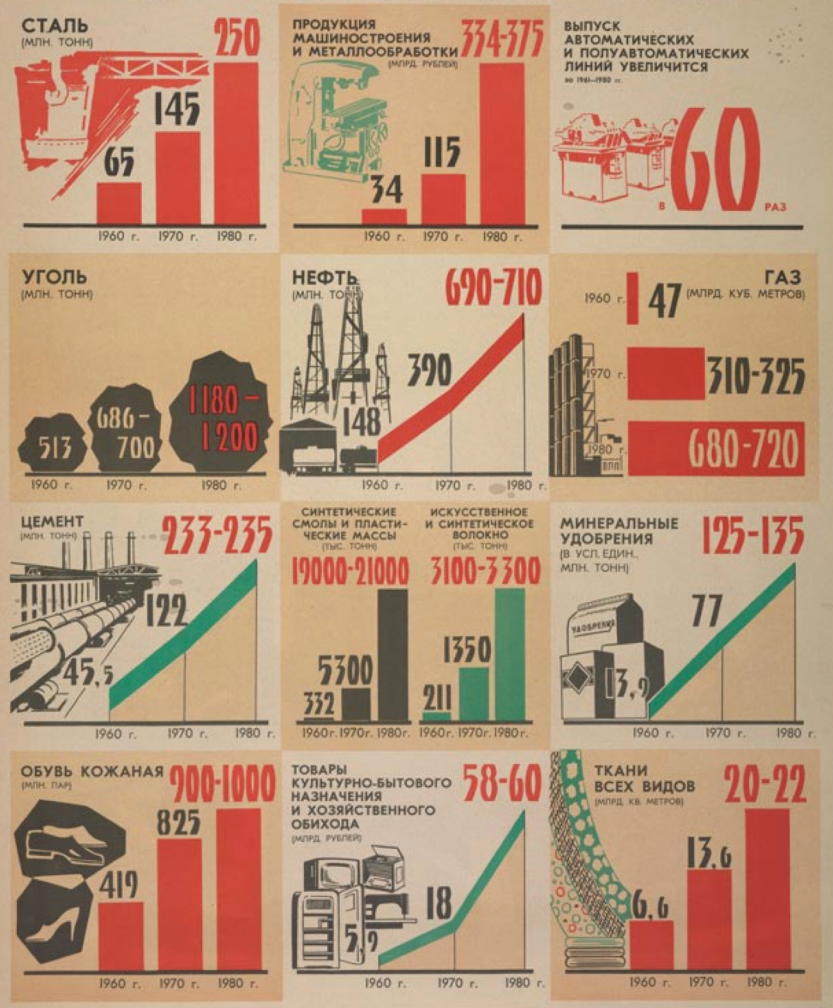
XXII СЪЕЗД КПСС

ГЕНЕРАЛЬНАЯ ПЕРСПЕКТИВА РАЗВИТИЯ ПРОМЫШЛЕННОСТИ СССР. 1961-1980 гг.

Тяжелая промышленность всегда играла и будет играть ведущую роль в расширенном воспроизводстве. Партия и впредь будет неустанно заботиться о ее росте, видя в этом решающее условие создания материально-технической базы, быстрого технического прогресса, основу укрепления обороноспособности социалистического государства. В то же время партия приложит все усилия к тому, чтобы тяжелая индустрия во все возрастающей степени обеспечивала увеличение производства предметов потребления.

Н. С. Хрущев

ПРОИЗВОДСТВО ОТДЕЛЬНЫХ ВИДОВ ПРОМЫШЛЕННОЙ ПРОДУКЦИИ



Через 20 лет СССР будет производить почти в 2 раза больше промышленной продукции, чем в 1961 году произведено во всех странах несоциалистического мира.

ГОСПОЛИТИЗДАТ · 1963

XXII СЪЕЗД КПСС

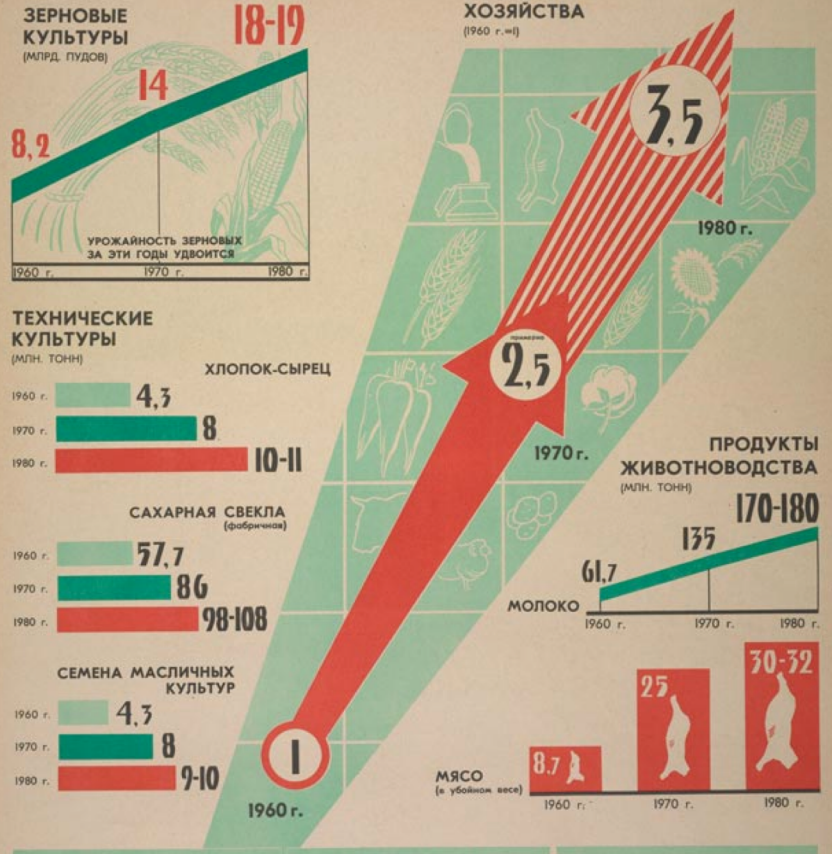
ВЕЛИКИЙ ПЛАН РАЗВИТИЯ СЕЛЬСКОГО ХОЗЯЙСТВА СССР. 1961–1980 гг.

Создание, наряду с могучей промышленностью, процветающего, всесторонне развитого и высокопродуктивного сельского хозяйства—обязательное условие построения коммунизма. Партия организует мощный подъем производительных сил сельского хозяйства, который позволит решить две основные, тесно связанные между собой задачи: а) достигнуть изобилия высококачественных продуктов питания для населения и сырья для промышленности; б) обеспечить постепенный переход советской деревни к коммунистическим общественным отношениям и ликвидировать в основном различия между городом и деревней.

Из Программы Коммунистической партии Советского Союза

ОБЩИЙ ОБЪЕМ ПРОДУКЦИИ СЕЛЬСКОГО ХОЗЯЙСТВА

(1960 г.=1)



ЗЕРНОВЫЕ КУЛЬТУРЫ

(МЛРД. ПУДОВ)

ТЕХНИЧЕСКИЕ КУЛЬТУРЫ

(МЛН. ТОНН)

ХЛОПОК-СЫРЕЦ

1960 г. 4,3
1970 г. 8
1980 г. 10-11

САХАРНАЯ СВЕКЛА (фабричная)

1960 г. 57,7
1970 г. 86
1980 г. 98-108

СЕМЕНА МАСЛИЧНЫХ КУЛЬТУР

1960 г. 4,3
1970 г. 8
1980 г. 9-10

ПРОДУКТЫ ЖИВОТНОВОДСТВА

(МЛН. ТОНН)

УВЕЛИЧЕНИЕ ПЛОЩАДИ ОРОШАЕМЫХ ЗЕМЕЛЬ

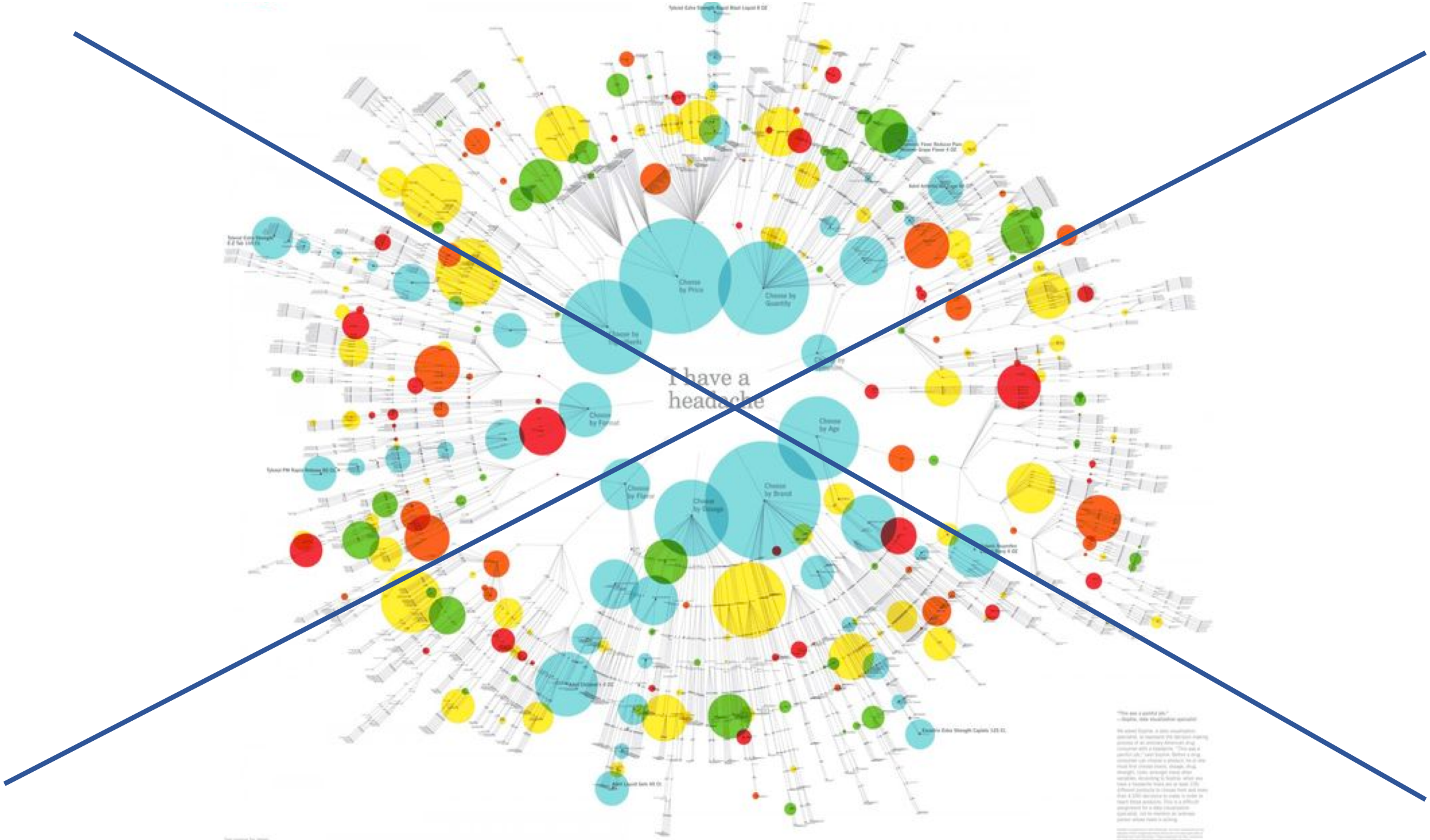
(МЛН. ГЕКТАРОВ)

ПОСТАВКА МИНЕРАЛЬНЫХ УДОБРЕНИЙ СЕЛЬСКОМУ ХОЗЯЙСТВУ

(МЛН. ТОНН)

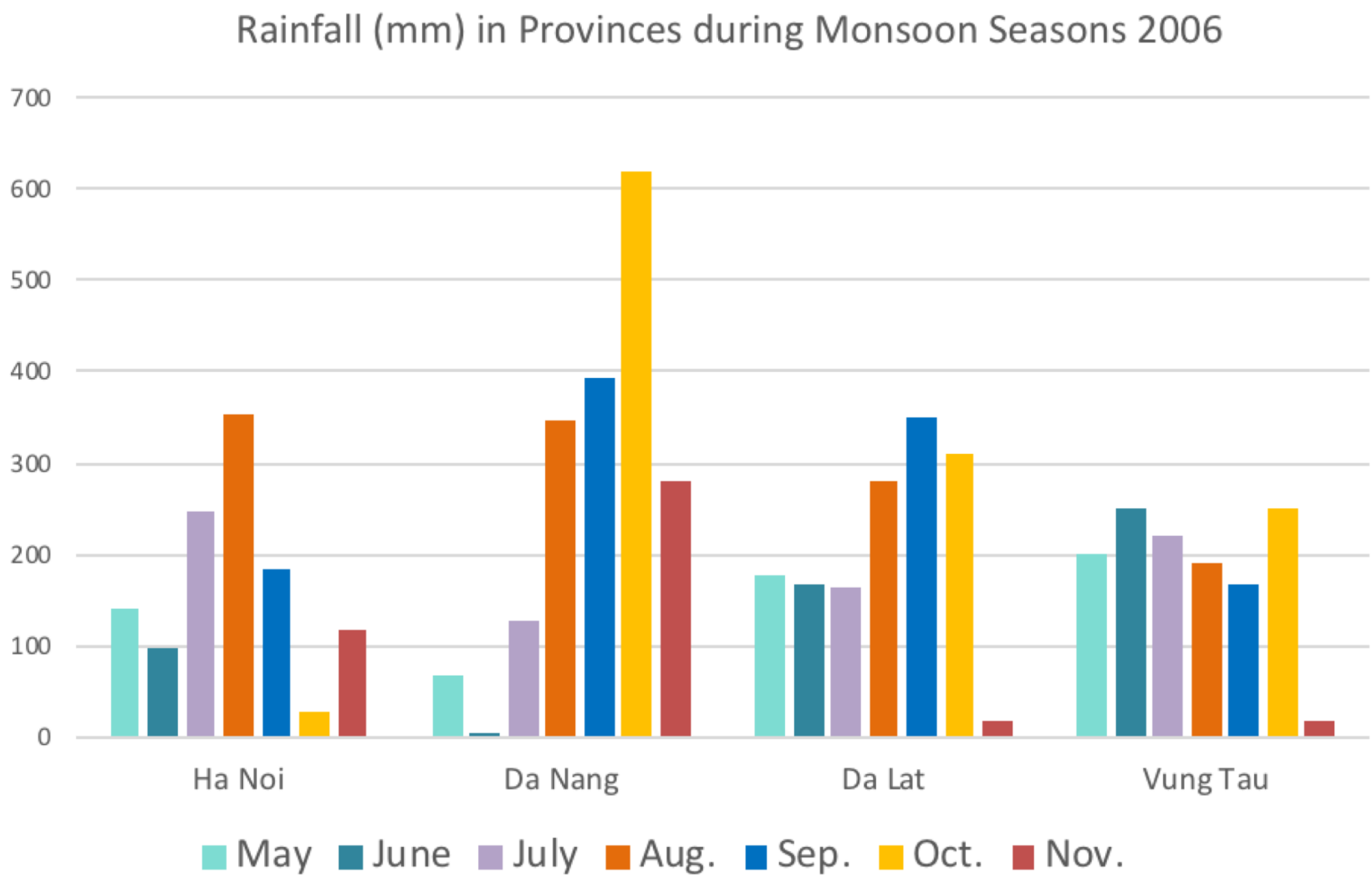
РОСТ ПРОИЗВОДИТЕЛЬНОСТИ ТРУДА В СЕЛЬСКОМ ХОЗЯЙСТВЕ

ГОСПОЛИТИЗДАТ · 1963

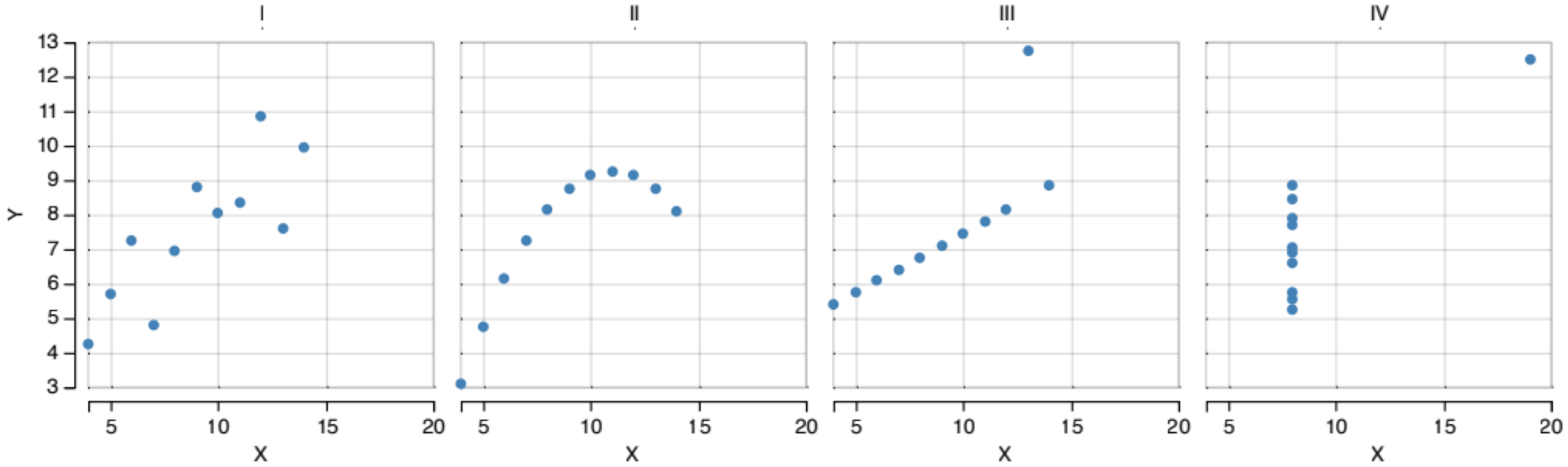

I have a
headache
Choose
by Price
Choose by
Quantity
Choose by
Brand
Choose
by Age
Choose
by Format
Choose
by Flavor
Choose
by Dosage

Data visualization is the art, practice, and strategy of effectively communicating information to humans.

Province	May	June	July	Aug.	Sep.	Oct.	Nov.
Ha Noi	140	97	247	354	183	28	116
Da Nang	69	2	127	346	394	619	279
Da Lat	176	166	165	281	349	309	19
Vung Tau	202	249	219	190	169	252	19

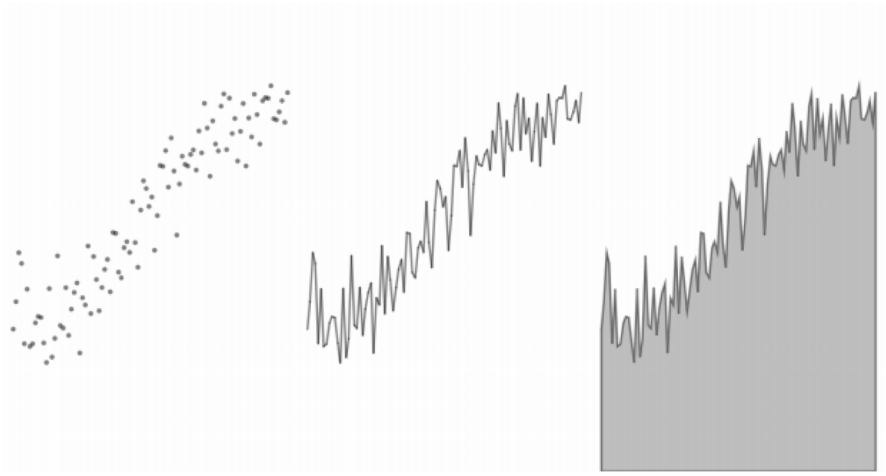

Rainfall (mm) in Provinces during Monsoon Seasons 2006
700
600
500
400
300
200
100
0
Ha Noi
Da Nang
Da Lat
Vung Tau
May
June
July
Aug.
Sep.
Oct.
Nov.

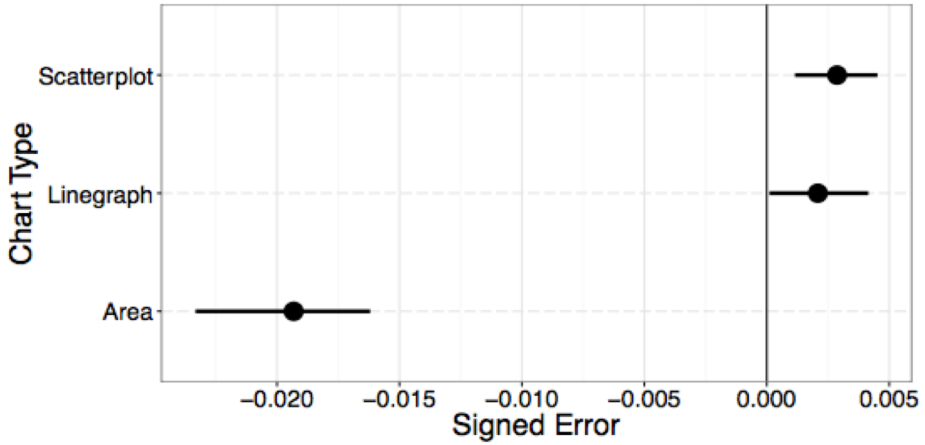
Regression by Eyes : Estimating Trends

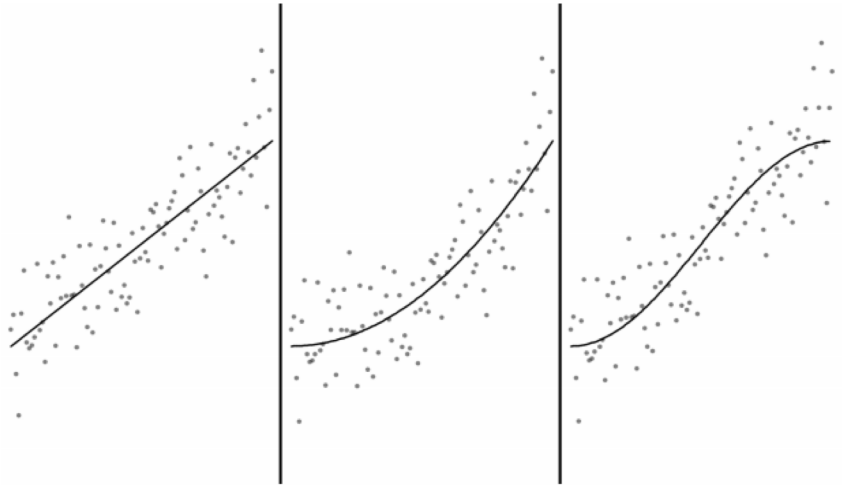


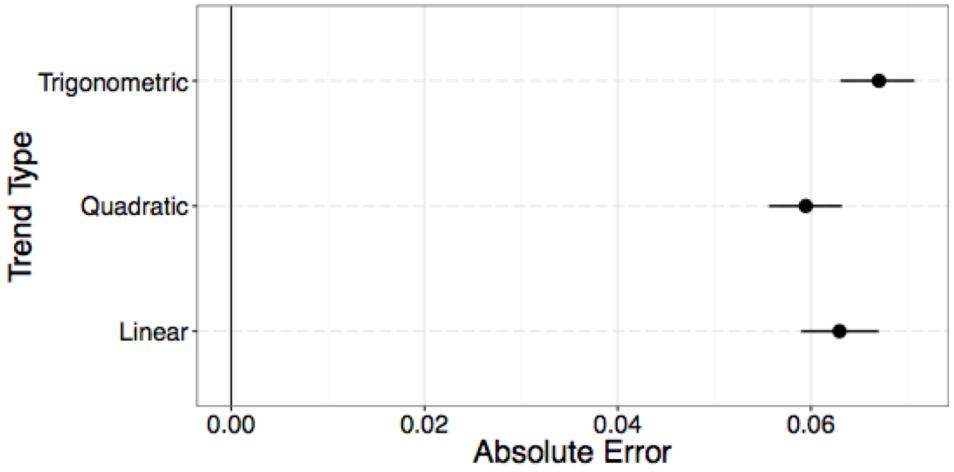
ANSCOMBE'S QUARTET

- *Same mean, standard deviation, and linear fit*









(CORELL, HEER 2017)

- Our eyes cannot estimate area graph regressions well.
- Our eyes can estimate more complex trends than just linear fits.

Data may be empirical but data interpretation is not.

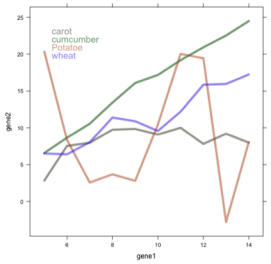
The person and the public reading your research has their own set of biases and former knowledge/experiences that may cloud their ability to understand your research or interpret it accurately.

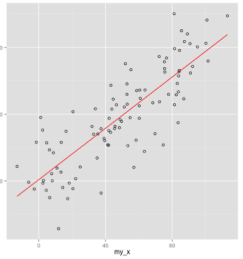
Why care about data visualization?

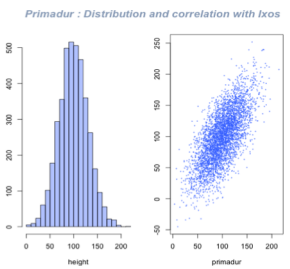
- See trends and relationships in data
- Spot errors and outliers otherwise unnoticeable in number format
- **Assure that results are interpreted correctly by key decision-makers**
- **Share your work and have it be properly understood and appreciated**
- **Make collaboration faster & easier**

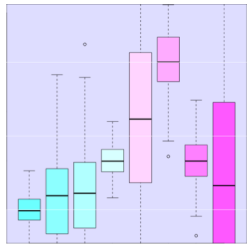
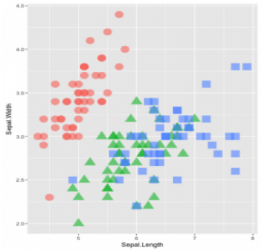
Common combined visualization + analysis platforms

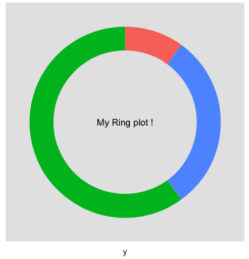
R – ggplot2, shiny etc.

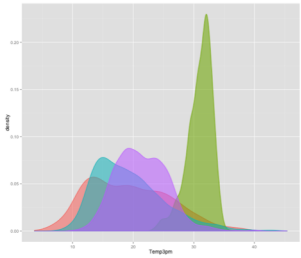


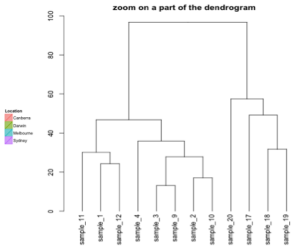




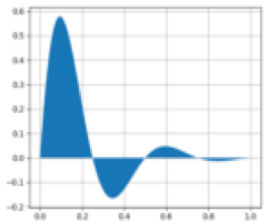
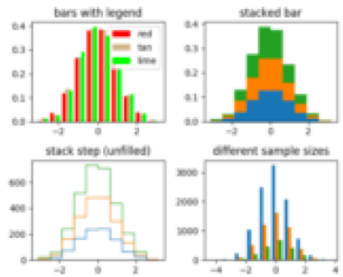



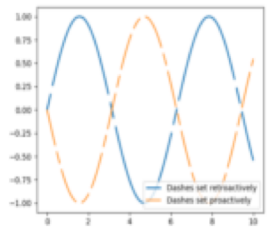


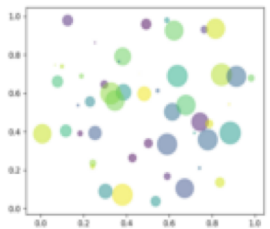
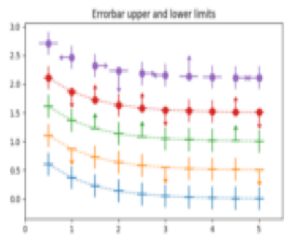


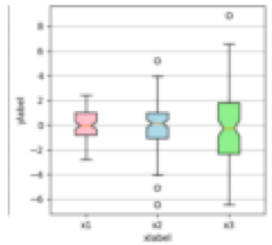


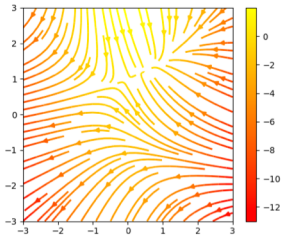
Python - matplotlib, seaborn, bokeh etc.







Other data visualization technologies (varying degrees of researcher-friendliness)

- Python, R, Stata, Matlib
- Networks : Cytoscape(js), Gephi, Netdraw
- D3.js
- CartoDB, Mapbox, OpenMaps (custom layer-based geomaps)
- 3JS for 3D visualizations
- Processing / P5.js for simulations
- Visual graph editor platforms : Excel, Lucidchart, Tableau, Plot.ly

PRACTICE ACTIVITY

For each prompt, sketch what you think is a possible way to represent the data described.

1. daily crop yield over weeks in May
2. daily crop yield of 4 types of fruit & vegetable crops over weeks in May
3. 2 years worth of daily crop yield of 3 types of fruit & vegetable crops over weeks in May
4. 2 years worth of daily crop yield of 3 types of fruit & vegetable crops, categorized by whether they are profitable, over weeks in May
5. * 2 years worth of daily crop yield of 3 types of fruit & vegetable crops, categorized by whether they are profitable, over weeks in May and August, between 5 farms in Hanoi, Hue, and Vung Tau

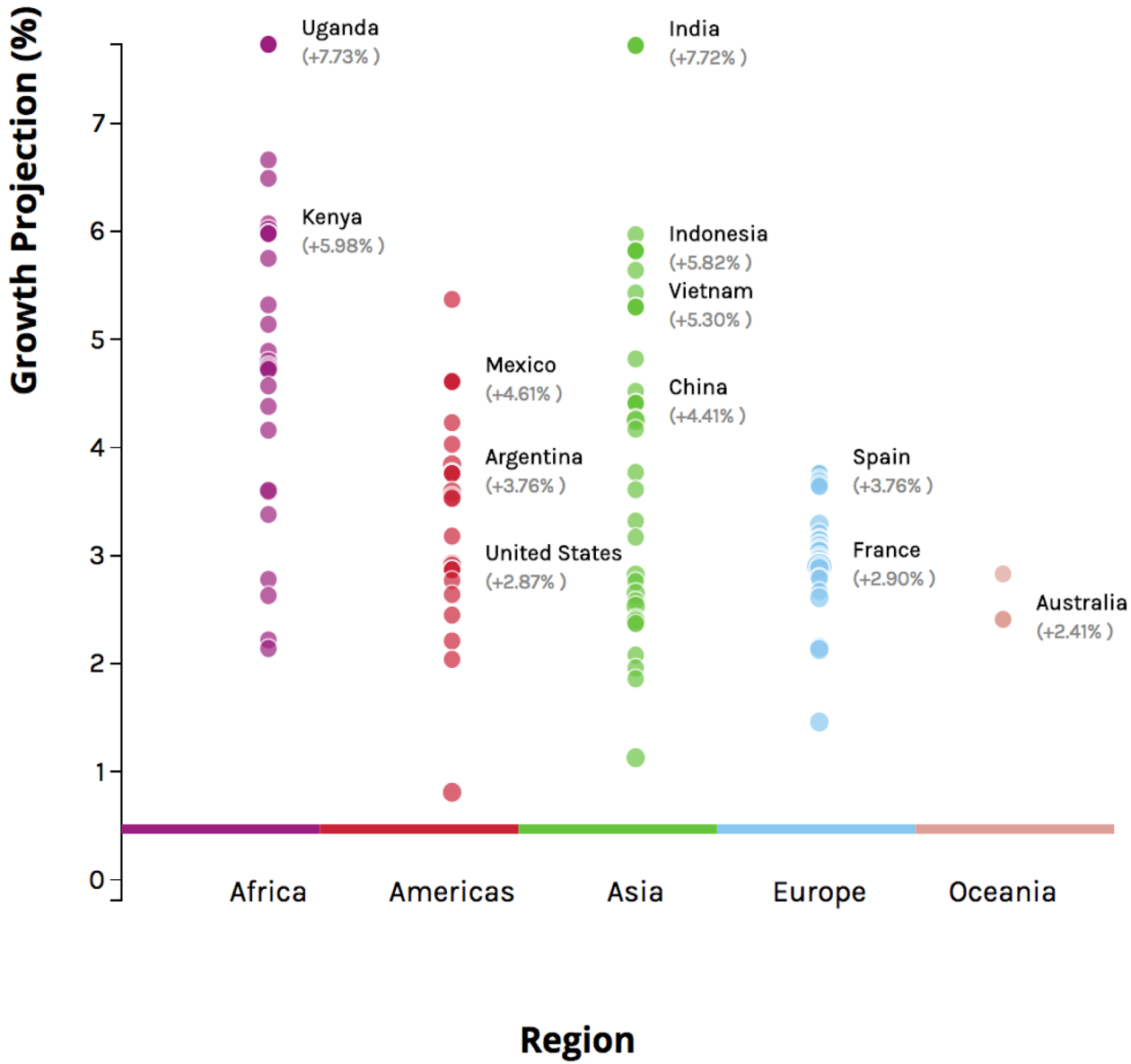
- 1x7 = **2D** data matrix *(daily crop yield over a week)*
- 3x2x4x7 = **4D** data matrix *(daily crop yield of 3 types of fruit & vegetable crops over weeks in May)*
- 2x3x2x4x7 = **5D** data matrix *(2 years worth of daily crop yield of 3 types of fruit & vegetable crops over weeks in May)*
- 2x2x3x2x4x7 = **6D** data matrix *(2 years worth of daily crop yield of 3 types of fruit & vegetable crops, categorized by whether they are profitable, over weeks in May)*
- 3x5x2x2x2x3x2x4x7 = **9D** data matrix *(2 years worth of daily crop yield of 3 types of fruit & vegetable crops, categorized by whether they are profitable, over weeks in May and August, between 5 farms in Hanoi, Hue, and Vung Tau)*

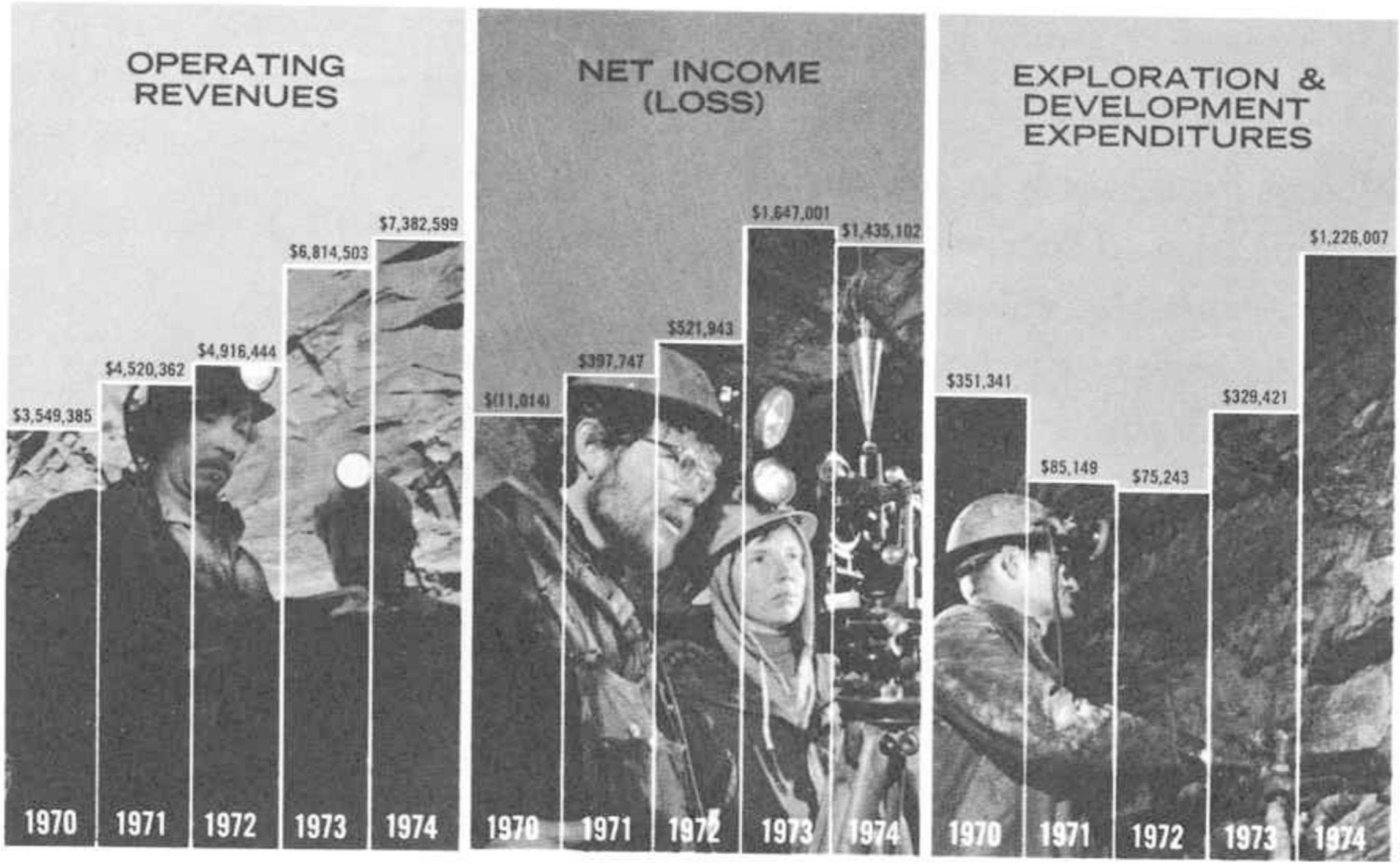
WHAT GOES INTO A DATA VISUALIZATION

1 | Source Data	*Metadata : header columns, data type, amount of data* *Data source context, subset selection*
2 | Intention	*What is the purpose of this data visualization?* *What question is it answering and to whom?*
3 | Visual Logic	*How should this visualization be read and in what order?* *Basics: Title, Axis, Legend + data elements* *Visual Hierarchy : What should be standing out first? What can be de-emphasized?*
4 | Interaction	*Where do people encounter this visualization and how should they interact with it?* *What kind of behavior does the visualization medium allow and encourage?*
5 | Takeaway	*What key learnings should the reader be persuaded of or want to learn more about?* *Does the data visualization inspire decisions or actions?*

Setting up data

	rank	value	name	iso
1	1	7.73	Uganda	UGA
2	2	7.72	India	IND
3	3	6.98	Egypt, Arab Rep.	EGY
4	4	6.66	Tanzania	TZA
5	5	6.49	Senegal	SEN
6	6	6.07	Madagascar	MDG
7	7	6.01	Malawi	MWI
8	8	5.98	Kenya	KEN
9	9	5.97	Pakistan	PAK
10	10	5.82	Indonesia	IDN
11	11	5.77	Kyrgyz Republic	KGZ
12	12	5.75	Mali	MLI
13	13	5.64	Turkey	TUR
14	14	5.43	Philippines	PHL
15	15	5.37	Guatemala	GTM
16				




OPERATING REVENUES
$3,549,385
$4,520,362
$4,916,444
$6,814,503
$7,382,599
1970
1971
1972
1973
1974
NET INCOME (LOSS)
$(11,014)
$397,747
$521,943
$1,647,001
$1,435,102
1970
1971
1972
1973
1974
EXPLORATION & DEVELOPMENT EXPENDITURES
$351,341
$85,149
$75,243
$329,421
$1,226,007
1970
1971
1972
1973
1974

OVERVIEW OF THE DATA VISUALIZATION PROCESS

- **Data Prep & Understanding**
- **Data Exploration**
- Design Strategy
- Visualization Design & Information Mapping
- Presentation

Setting up data

- Readable data file, convert to correct format to open
- Correct format for data type (is “6” an ‘Int/Number’ or a ‘String/Text’?)
- Data within reasonable range of graph axes, check for outlier skew or scaling issues
- Check for data duplicates or missing entries
- Are the units / proportional reference of the data comparable?
- What are the relationships set up between columns and rows?
- Combine datasets on common dimensions

		May	June	July	Aug.	Sep.	Oct.	Nov.
2005	Lai Chau	73	513	322	334	30	38	21
	Son La	65	150	267.32	403	147	58	21
	Tuyen Quang	110	280	167	340	172	11	44
	Ha Noi	221	278	278	377	366	18	92
	Bai Chay	247	340	628	364	167	92	87
	Nam Dinh	73	67	241	324	496	63	210
	Vinh	119	50	228	424	647	258	106
	Hue	42	113	129	189	350		485
	Da Nang	20	22	136	210	236	510	432
	Qui Nhon	49	27	13	20	362	914	488
	Playku	46	182	479	610	314	187	45
	DaLat	172	182	200	259	354	263	92
	Nha Trang	0	32	42	11	258	487	355
	Vung Tau	119	147	170	155	189	71	7
	Ca Mau	213	227	400	166	380	497	207
2006	Lai Chau	243	402	378	291	163	42	25
	Son La	152	223	262	305	58	39	12
	Tuyen Quang	263	115	459	455	94	58	50
	Ha Noi	140	97	247	354	183	28	116
	Bai Chay	49	198	464	666	80	50	86
	Vinh	100	57	171	547	254	518	58
	Nam Dinh	220	124	186	3270	102	60	low
	Vinh	100	57	171	547	254	518	58
	Huee	61	13	54	476	510	406	239
	Da Nang	69	2	127	346	394	619	279
	Qui Nhon	106	30	70	46	219	191	138
	Playku	152	202	649	526	330	202	2
	Da Lat	"176"		165	281	349	309	19
	Nha Trang	24	5	7	68	158	179	-61
	Vung Tau	202	249	219	190	169	252	19
	Ca Mau	231	324	475	450	374	241	80

Rainfall in Vietnam Provinces during rain season 2005 – 2006

Source : Vietnam Statistics Office Website

		May	June	July	Aug.	Sep.	Oct.	Nov.
2005	Lai Chau	73	513	322	334	30	38	21
	Son La	65	150	267.32	403	147	58	21
	Tuyen Quang	110	280	167	340	172	11	44
	Ha Noi	221	278	278	377	366	18	92
	Bai Chay	247	340	628	364	167	92	87
	Nam Dinh	73	67	241	324	496	63	210
	Vinh	119	50	228	424	647	258	106
	Hue	42	113	129	189	350		485
	Da Nang	20	22	136	210	236	510	432
	Qui Nhon	49	27	13	20	362	914	488
	Playku	46	182	479	610	314	187	45
	DaLat	172	182	200	259	354	263	92
	Nha Trang	0	32	42	11	258	487	355
	Vung Tau	119	147	170	155	189	71	7
	Ca Mau	213	227	400	166	380	497	207
2006	Lai Chau	243	402	378	291	163	42	25
	Son La	152	223	262	305	58	39	12
	Tuyen Quang	263	115	459	455	94	58	50
	Ha Noi	140	97	247	354	183	28	116
	Bai Chay	49	198	464	666	80	50	86
	Vinh	100	57	171	547	254	518	58
	Nam Dinh	220	124	186	3270	102	60	low
	Vinh	100	57	171	547	254	518	58
	Huee	61	13	54	476	510	406	239
	Da Nang	69	2	127	346	394	619	279
	Qui Nhon	106	30	70	46	219	191	138
	Playku	152	202	649	526	330	202	2
	Da Lat	"176"		165	281	349	309	19
	Nha Trang	24	5	7	68	158	179	-61
	Vung Tau	202	249	219	190	169	252	19
	Ca Mau	231	324	475	450	374	241	80

Rainfall in Vietnam Provinces during rain season 2005 – 2006

Source : Vietnam Statistics Office Website

Data formats

```
Year;"Cities";"Jan.";"Feb.";"March”
2016;"Lai Chau";176.1;130.0;181.8
2016;"Son La";148.2;120.2;193.3
```

```
Year    "Cities”        "Jan.”   "Feb.”   "March”
2016    "Lai Chau”      176.1    130.0    181.8
2016    "Son La”        148.2    120.2    193.3
```

		Jan.	Feb.	March
2016	**Lai Chau**	176.1	130	181.8
	Son La	148.2	120.2	193.3

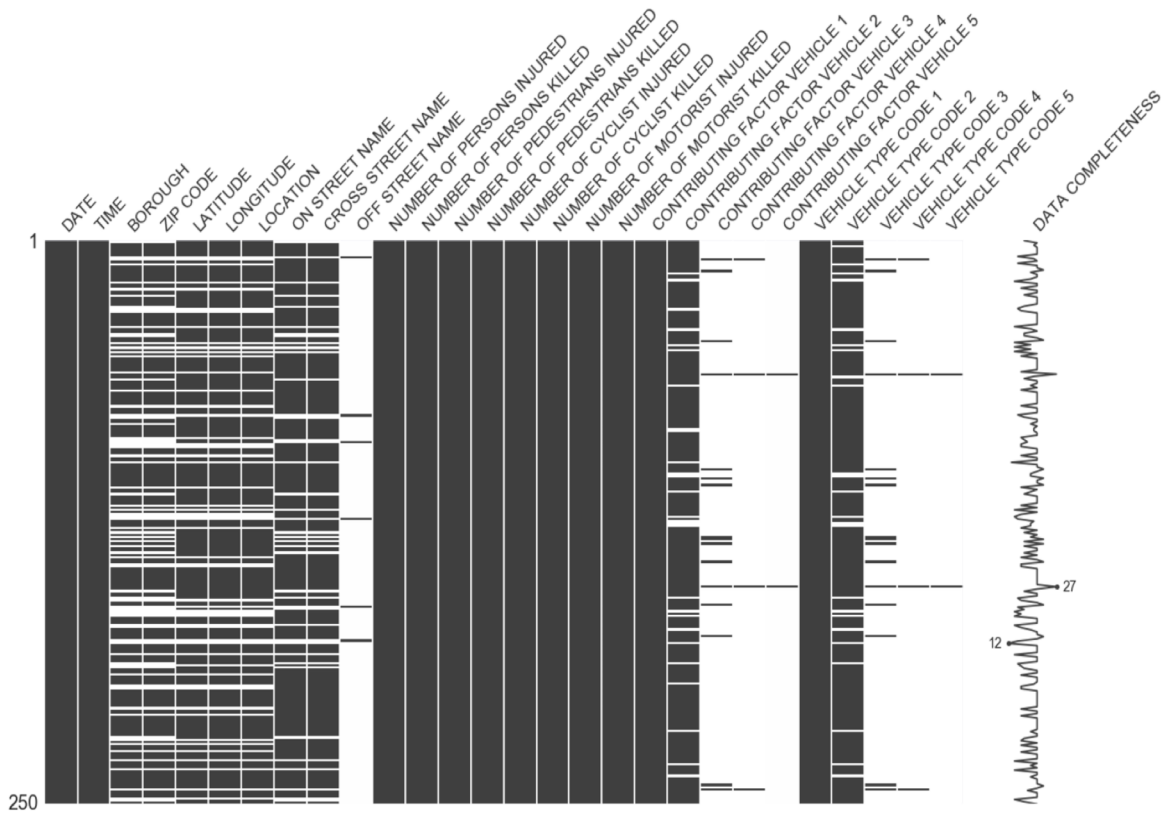
CSV (Comma Separated Values), TSV (Tab Separated Values)

Excel Table

```
[
  {
    "Year": 2016,
    "Cities": "Lai Chau",
    "Jan.": 176.1,
    "Feb.": 130,
    "March": 181.8
  },
  {
    "Year": 2016,
    "Cities": "Son La",
    "Jan.": 148.2,
    "Feb.": 120.2,
    "March": 193.3
  }
]
```

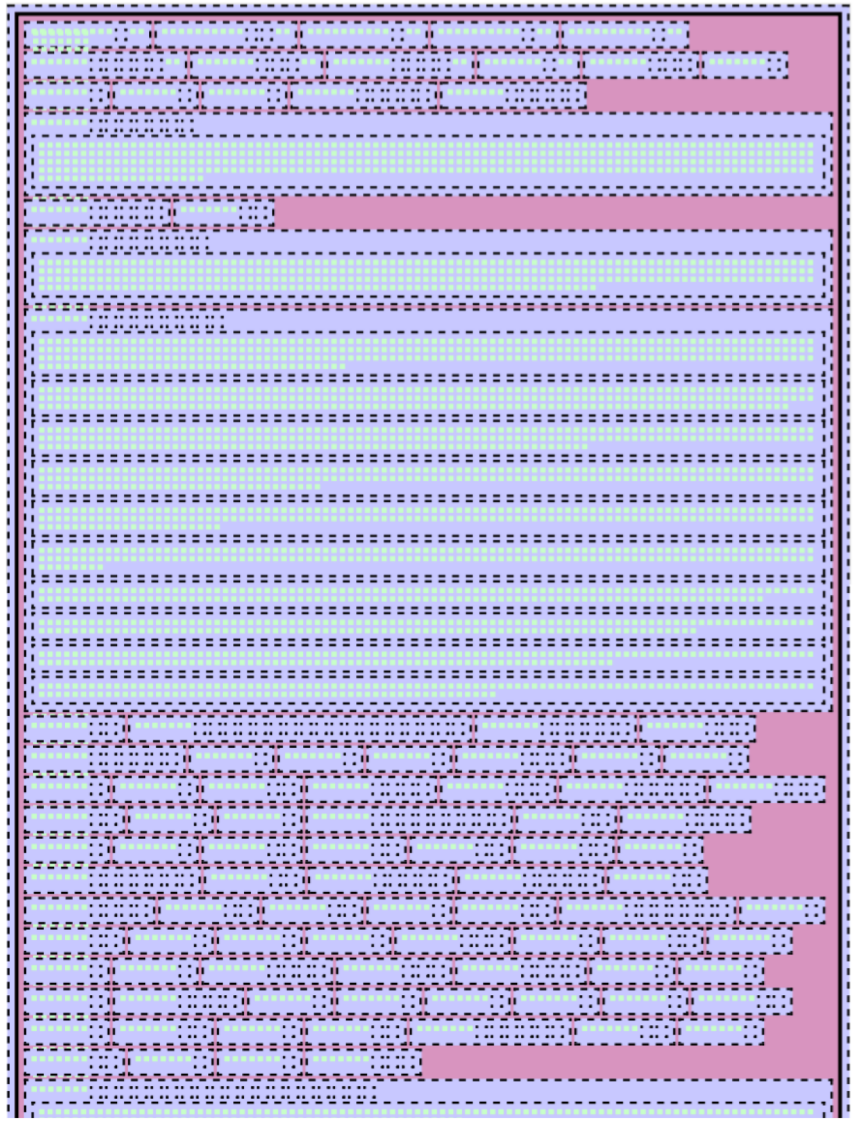
JSON (Javascript Object Notation)

Data structures



See gaps in data file

Missingno project :
https://github.com/ResidentMario/missingno#matrix



Visualization of a single JSON object with encompassing architectures

(embedded arrays, objects, text/number value)

Exploratory Data Visualization

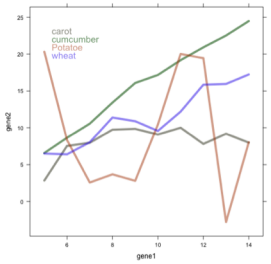
versus (and)

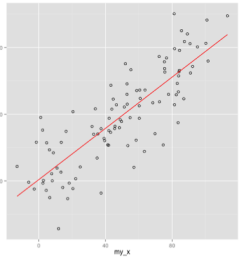
Persuasive Data Visualization

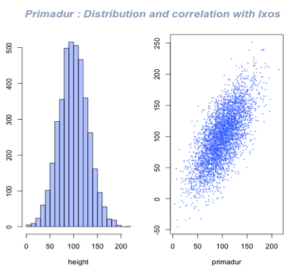
When to use each?

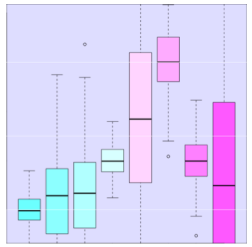
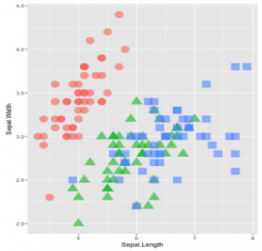
Common combined visualization + analysis platforms

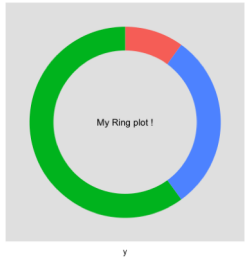
R – ggplot2, shiny etc.

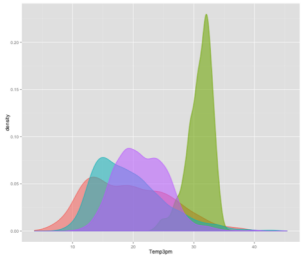


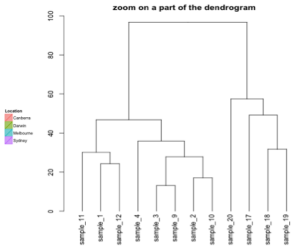




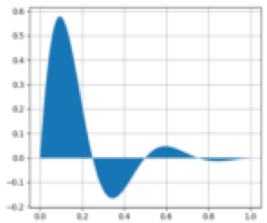
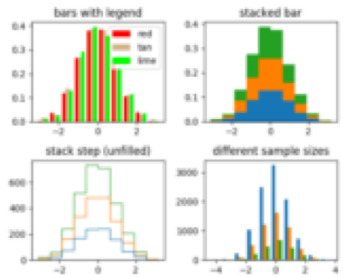



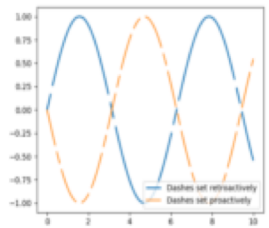


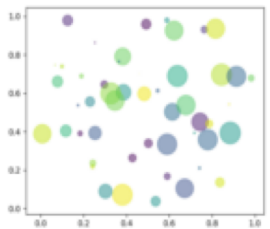
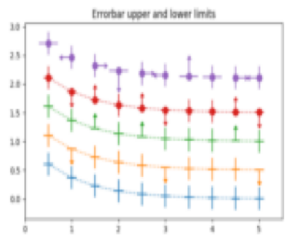


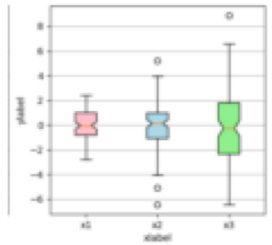


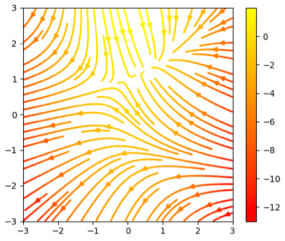
Python - matplotlib, seaborn, bokeh etc.







Common data exploration visualizations

Histogram

Bar plot

Boxplot

Scatterplot

Line plot

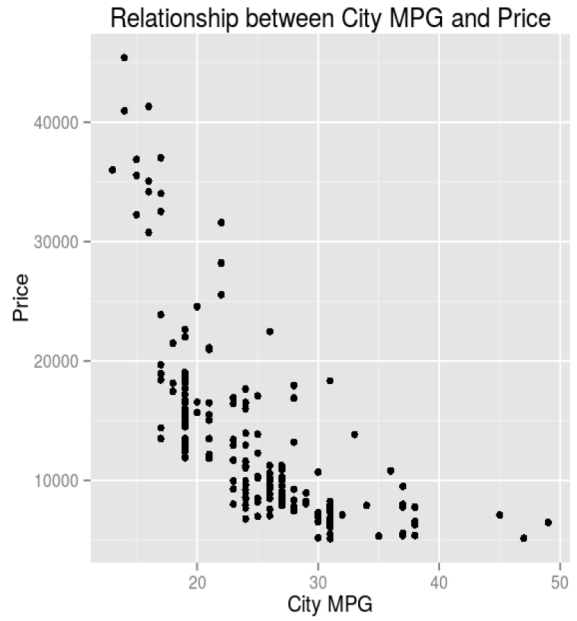
Conditioned plots

Facet plots

Exploring a dataset with basic visualizations

R

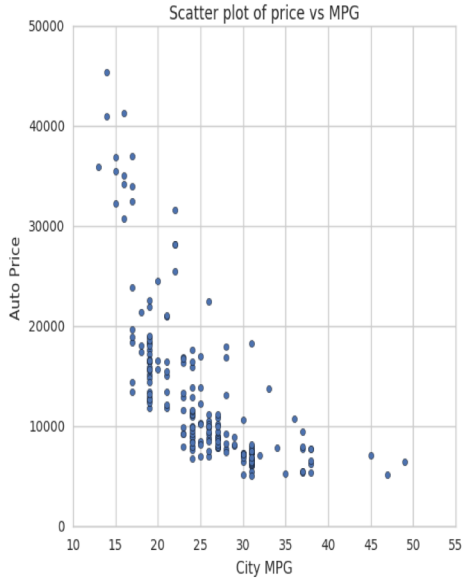
```
In [67]: options(repr.plot.width=5, repr.plot.height=5)
ggplot(auto.price, aes(city.mpg, price)) + geom_point() +
    xlab('City MPG') + ylab('Price') +
    ggtitle('Relationship between City MPG and Price')
```



Python

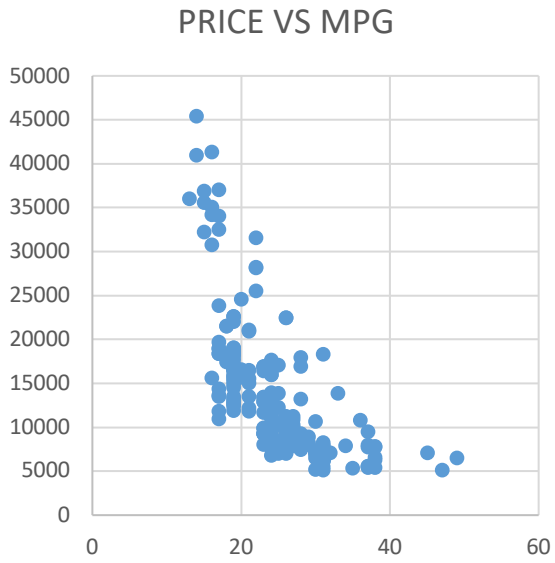
```
In [11]: import matplotlib.pyplot as plt
fig = plt.figure(figsize=(6,6)) # define plot area
ax = fig.gca() # define axis
auto_prices.plot(kind = 'scatter', x = 'city-mpg', y = 'price', ax = ax)
ax.set_title('Scatter plot of price vs MPG') # Give the plot a main title
ax.set_xlabel('City MPG') # Set text for the x axis
ax.set_ylabel('Auto Price')# Set text for y axis
```

```
Out[11]: <matplotlib.text.Text at 0x7f10a32fbf10>
```



Excel

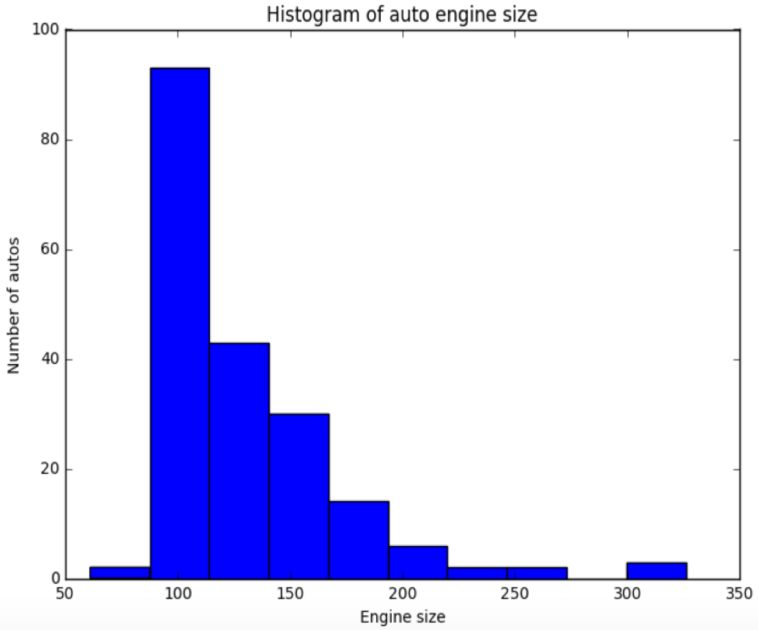
1. Select data ranges
2. *INSERT : Graph*
3. *Select : Scatterplot*



Exploring a dataset with basic visualizations

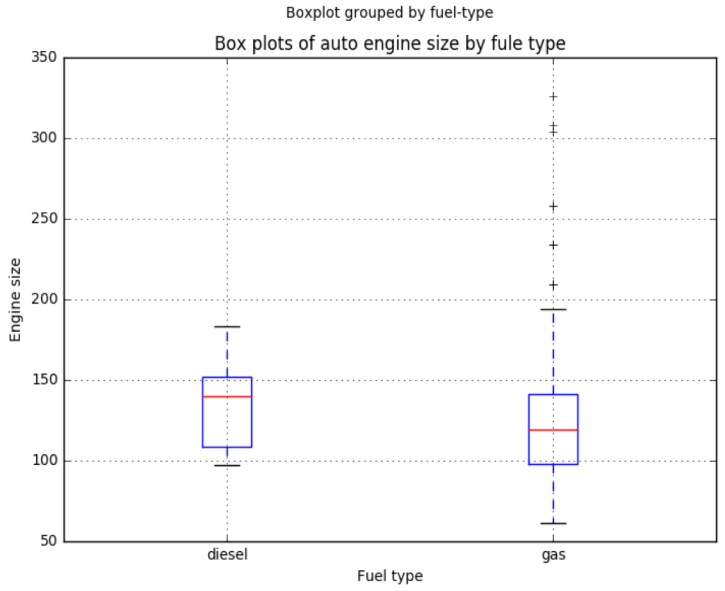
```
In [5]: fig = plt.figure(figsize=(8,6)) # define plot area
        ax = fig.gca() # define axis
        auto_prices['engine-size'].plot.hist(ax = ax) # Use the plot.hist method on subset of the data frame
        ax.set_title('Histogram of auto engine size') # Give the plot a main title
        ax.set_xlabel('Engine size') # Set text for the x axis
        ax.set_ylabel('Number of autos')# Set text for y axis
```

```
Out[5]: <matplotlib.text.Text at 0x7f10af5d0610>
```



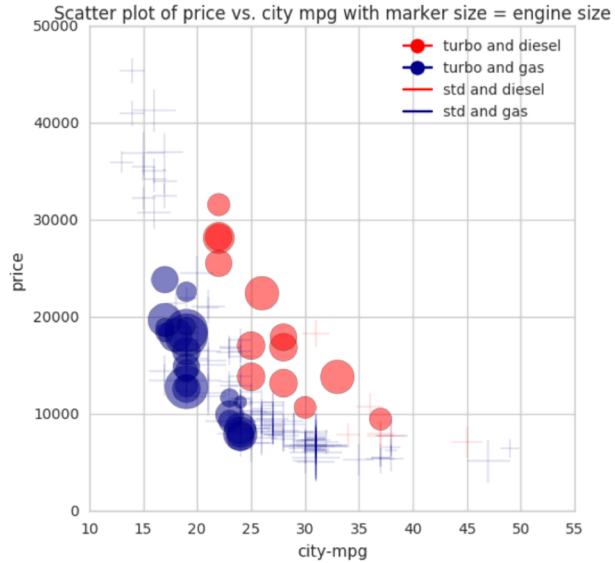

```
In [6]: fig = plt.figure(figsize=(8,6)) # define plot area
        ax = fig.gca() # define axis
        auto_prices[['engine-size','fuel-type']].boxplot(by = 'fuel-type', ax = ax) # Use the plot
        ax.set_title('Box plots of auto engine size by fule type') # Give the plot a main title
        ax.set_xlabel('Fuel type') # Set text for the x axis
        ax.set_ylabel('Engine size')# Set text for y axis
```

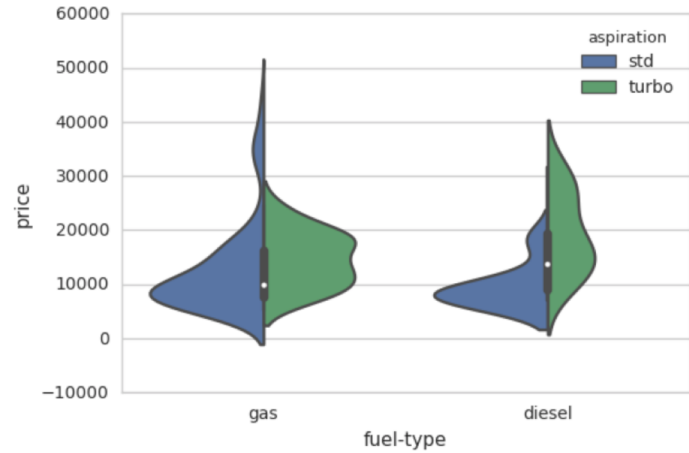
```
Out[6]: <matplotlib.text.Text at 0x7f10af6742d0>
```

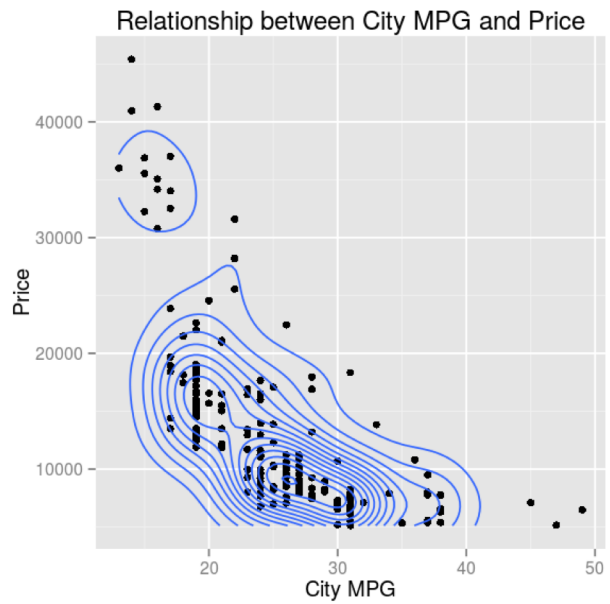


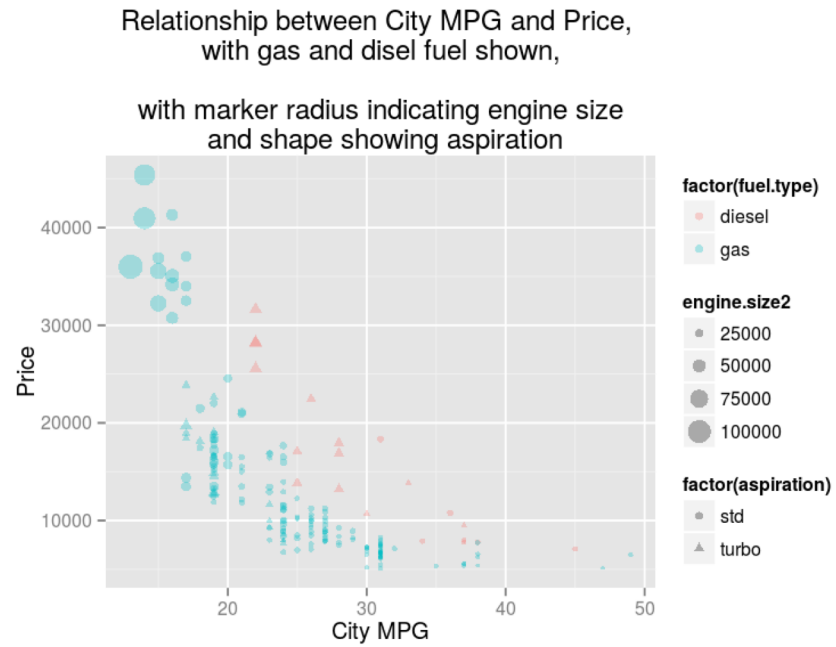
Basic Data Analysis Statistical Tests

- T-test (checks the null hypothesis that two ranges of data do not have a significant difference)
  - T-statistic vs. Critical value (1.96)
  - P-value vs. Confidence level (0.05)
  - Low/High 95% Confidence Interval
- Linear Correlation

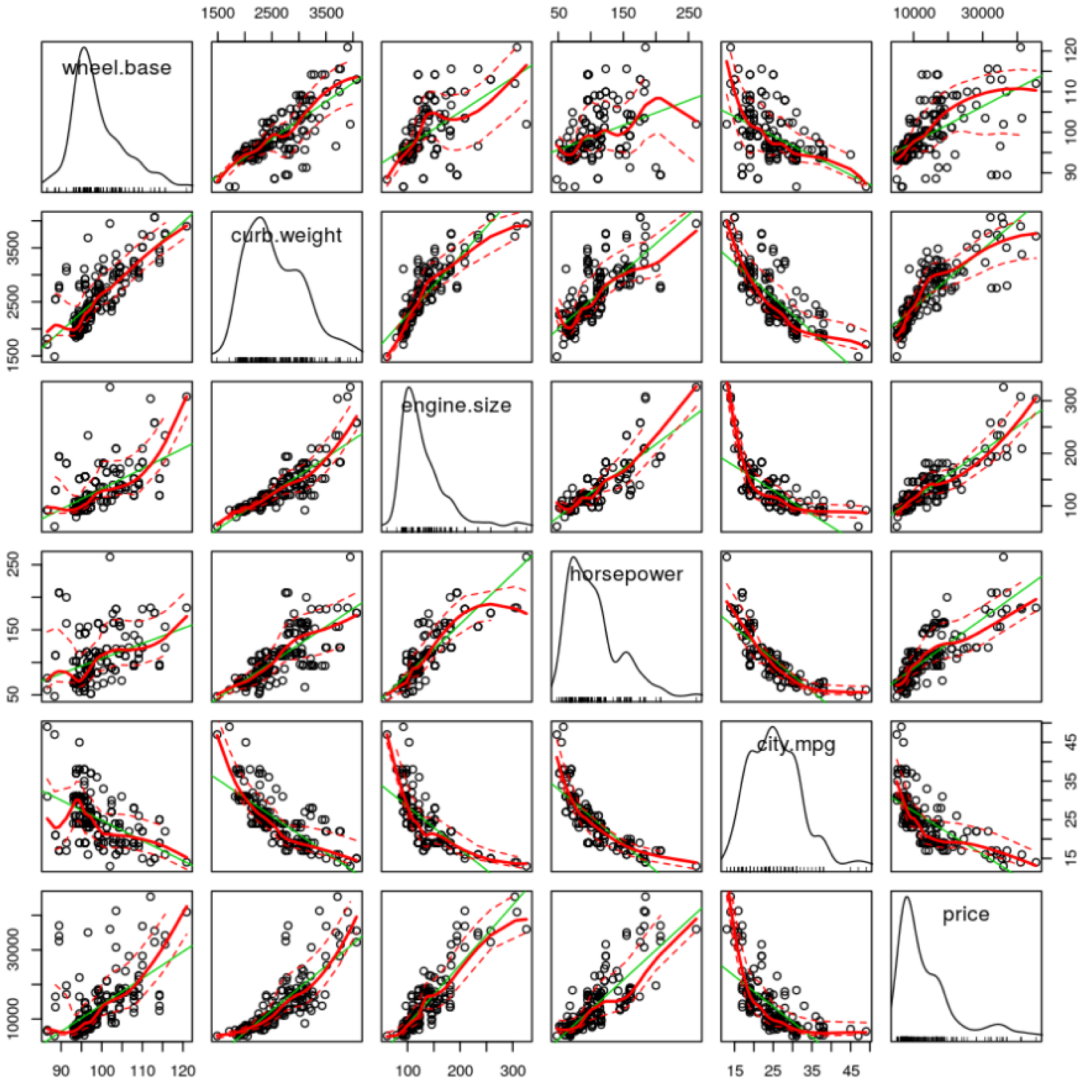

Scatter plot of price vs. city mpg with marker size = engine size
turbo and diesel
turbo and gas
std and diesel
std and gas
price
city-mpg


aspiration
std
turbo
price
gas
diesel
fuel-type

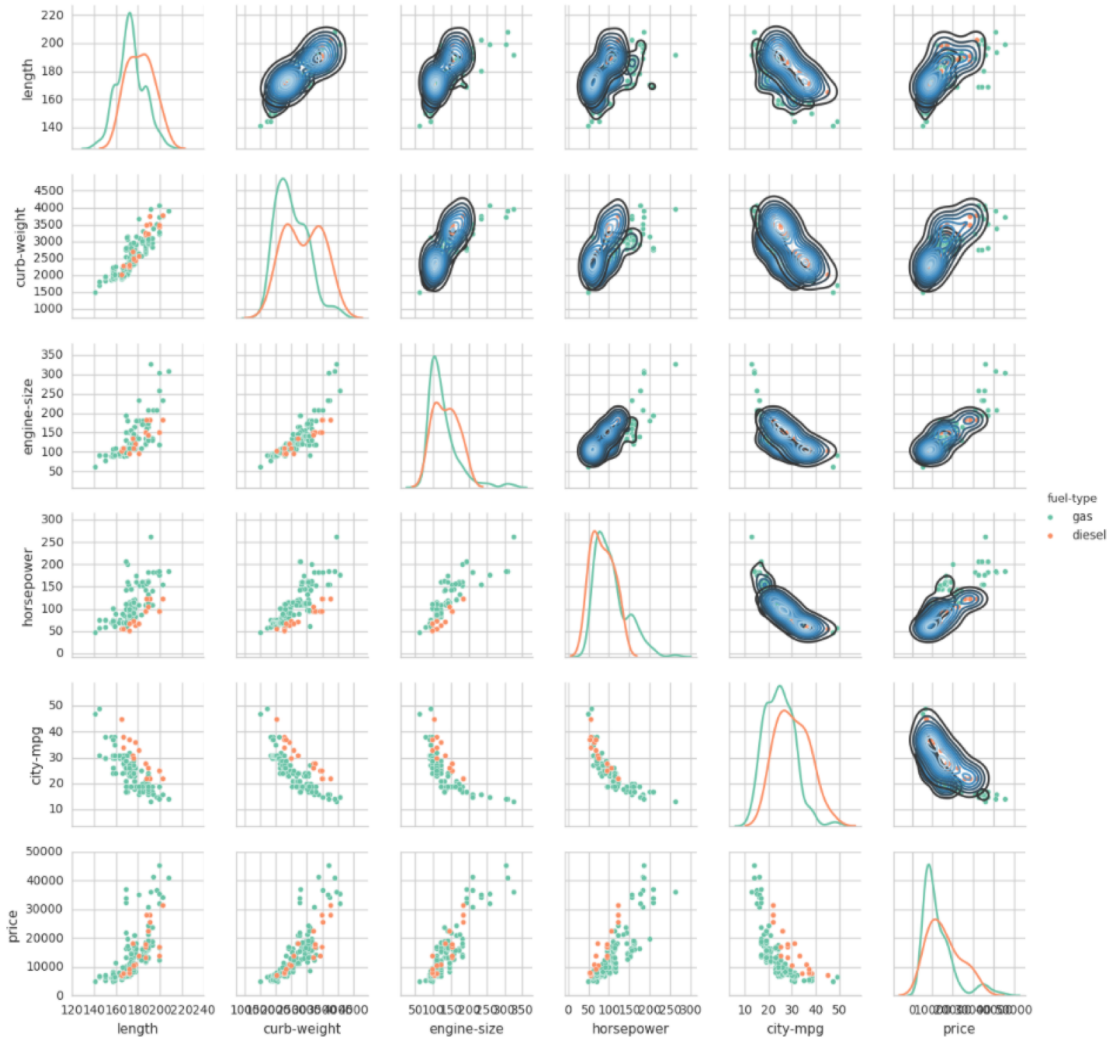

Relationship between City MPG and Price
Price
City MPG

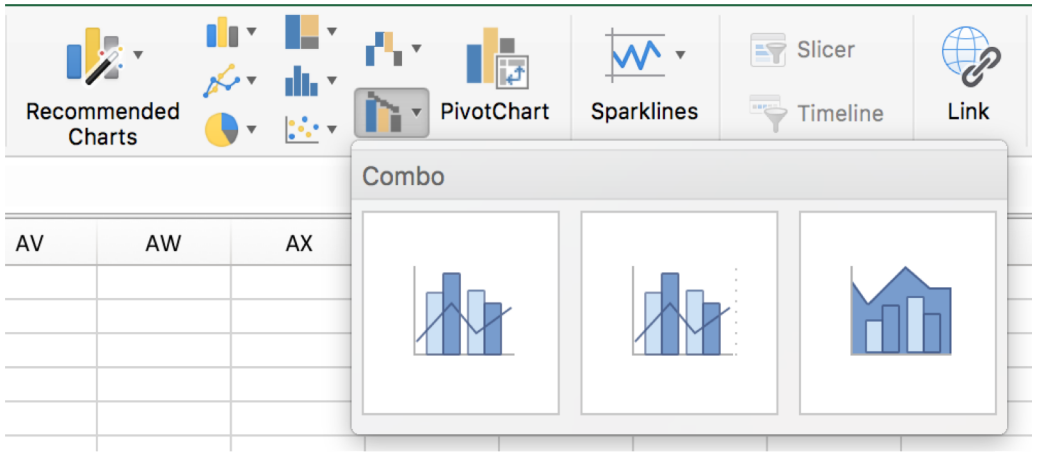

Relationship between City MPG and Price,
with gas and disel fuel shown,
with marker radius indicating engine size
and shape showing aspiration
Price
City MPG
factor(fuel.type)
diesel
gas
engine.size2
25000
50000
75000
100000
factor(aspiration)
std
turbo

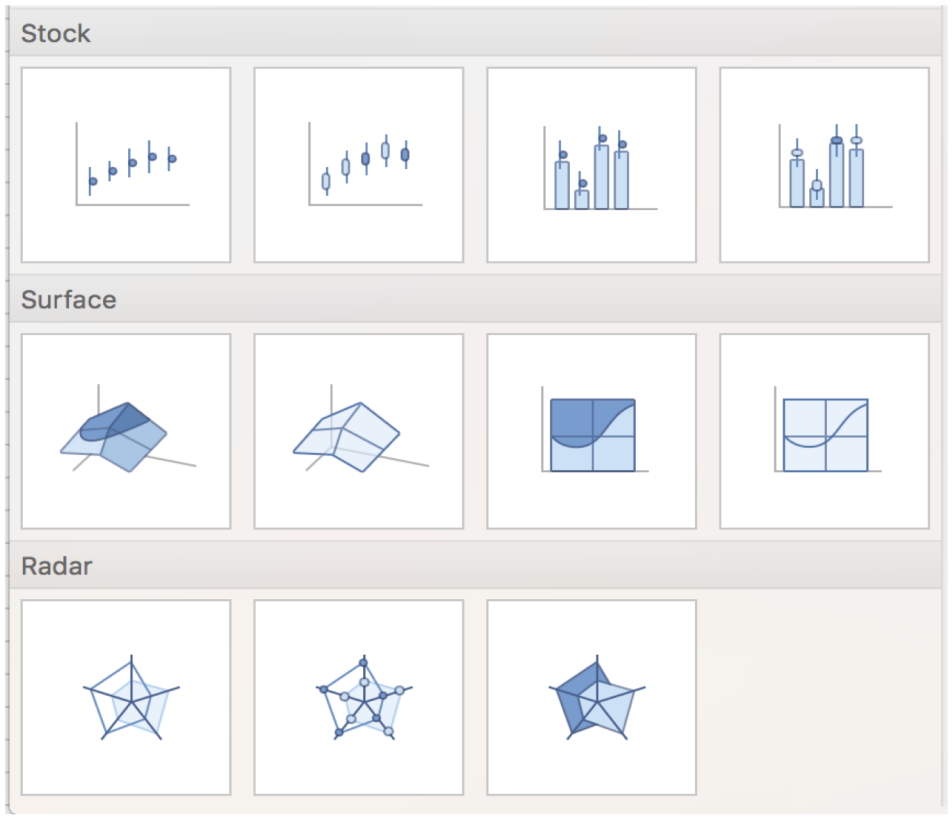
R

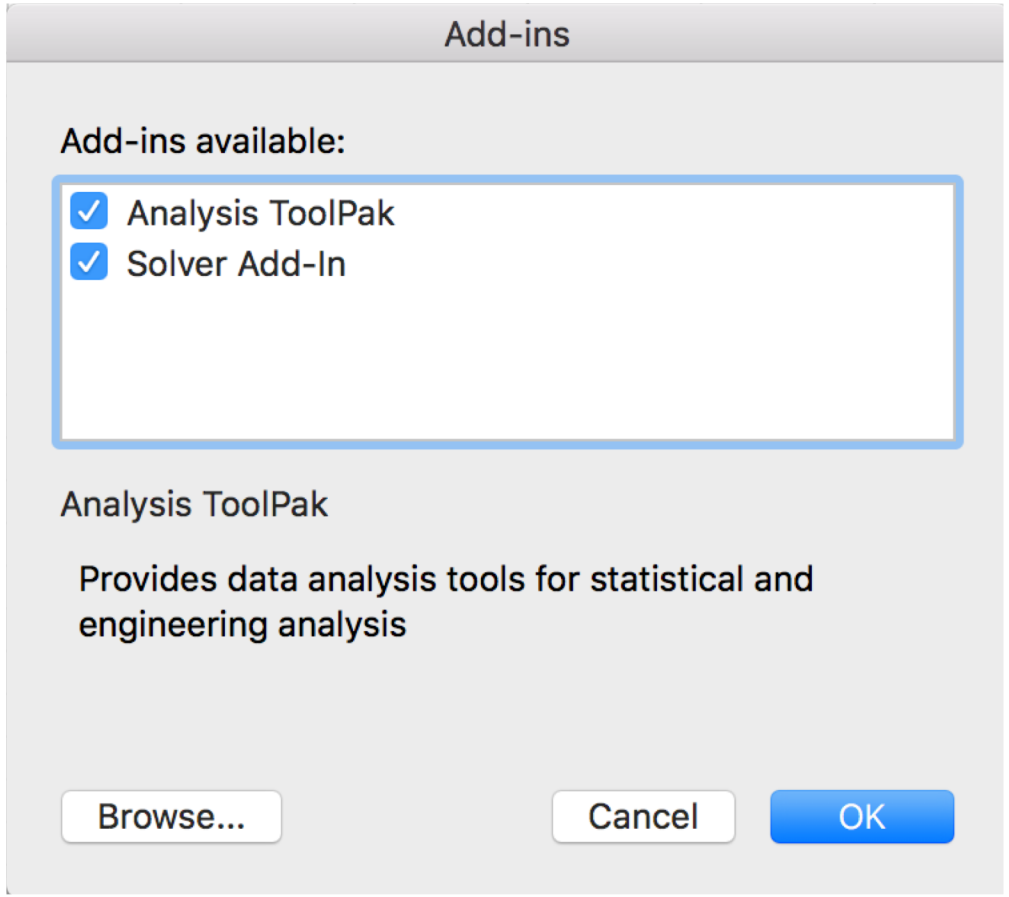

wheel.base
curb.weight
engine.size
horsepower
city.mpg
price

Python


length
curb-weight
engine-size
horsepower
city-mpg
price
fuel-type
gas
diesel


Recommended Charts
PivotChart
Sparklines
Slicer
Timeline
Link
Combo
AV
AW
AX


Stock
Surface
Radar


Add-ins
Add-ins available:
Analysis ToolPak
Solver Add-In
Analysis ToolPak
Provides data analysis tools for statistical and engineering analysis
Browse...
Cancel
OK

Excel + table header filter + conditional formatting = 👍

Year	Province	May	June	July	Aug.	Sep.	Oct.	Nov.
2016	Lai Chau	308.3	446.3	422.7	236.3	281.6	45.6	81.6
2016	Son La	347.3	166.0	154.5	286.1	129.4	32.0	42.3
2016	Tuyen Quang	285.7	74.1	312.0	306.9	86.7	104.4	40.5
2016	Ha Noi	249.0	95.1	280.4	534.5	178.5	45.0	9.3
2016	Bai Chay	205.9	211.5	567.5	497.4	213.8	43.6	18.1
2016	Nam Dinh	116.7	92.0	296.7	446.2	220.9	78.3	7.8
2016	Vinh	85.5	9.7	114.5	177.3	741.4	563.9	287.5
2016	Hue	108.0	102.4	84.4	165.9	661.9	618.6	577.3
2016	Da Nang	59.0	47.0	54.3	145.0	783.3	411.2	336.8
2016	Qui Nhon	41.1	47.7	4.7	183.4	192.4	385.9	762.8
2016	Playku	161.8	195.1	141.6	448.7	524.0	229.1	54.0
2016	Da Lat	133.5	226.8	209.4	83.0	498.7	377.7	116.3
2016	Nha Trang	52.7	87.9	29.7	82.2	123.6	255.3	399.4
2016	Vung Tau	83.0	211.2	136.7	226.7	165.6	373.6	135.4
2016	Ca Mau	161.4	230.8	432.2	271.9	345.0	501.1	183.6

LAB ACTIVITY & HOMEWORK

- Finish cleaning 2005-2006 rainfall dataset.
- Explore complete rainfall dataset using histogram, line chart, stacked bar, and scatterplot to see the change in rainfall between 2005 – 2006 for Hanoi and Nha Trang
- What are learnings from this dataset? Use 2 graphs of your choice to visualize rainfall for all cities during three years of your choice.
- Review the other datasets in the drive and/or research your own dataset for the individual & final group projects
- *Extra credit 1: Disprove a null hypothesis and visualize the relationship.*
- *Extra credit 2 : Do Practice Activity Challenge #5*