Download Tableau & H-1B petition data

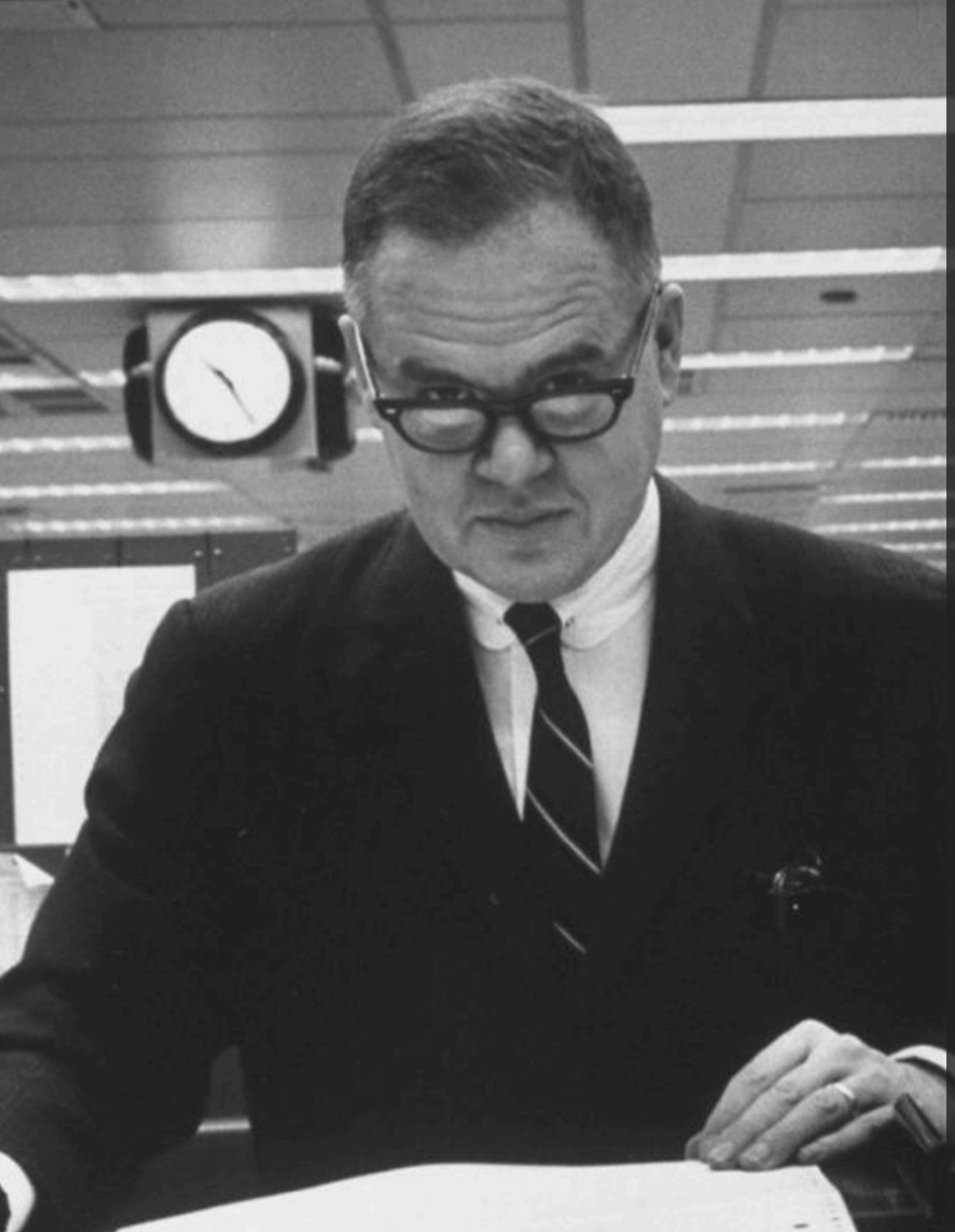
Exploratory Data Analysis

Nam Wook Kim

Mini-Courses — January @ GSAS
2018

Goal

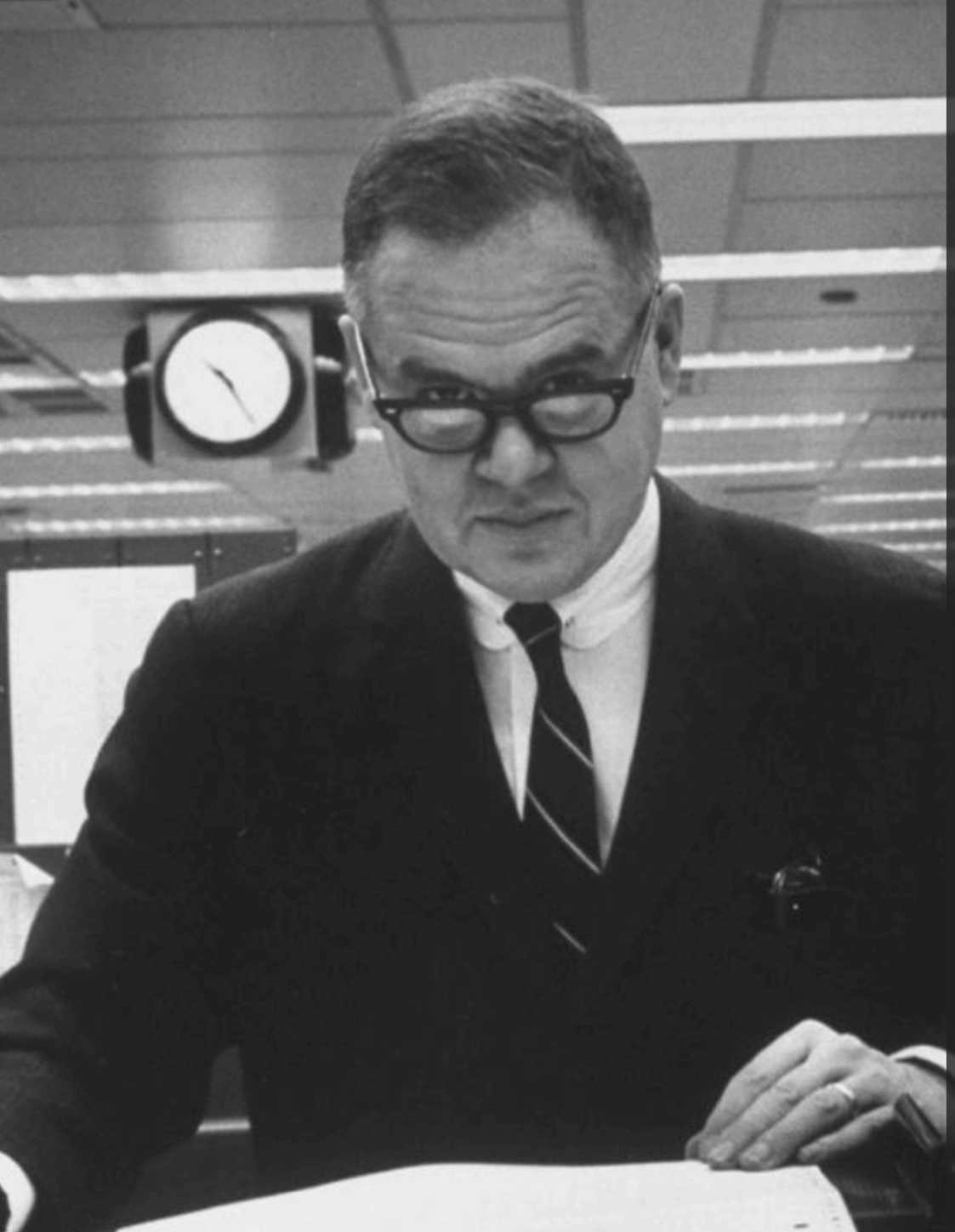
Learn the Philosophy of Exploratory Data Analysis



Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis...

It is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.

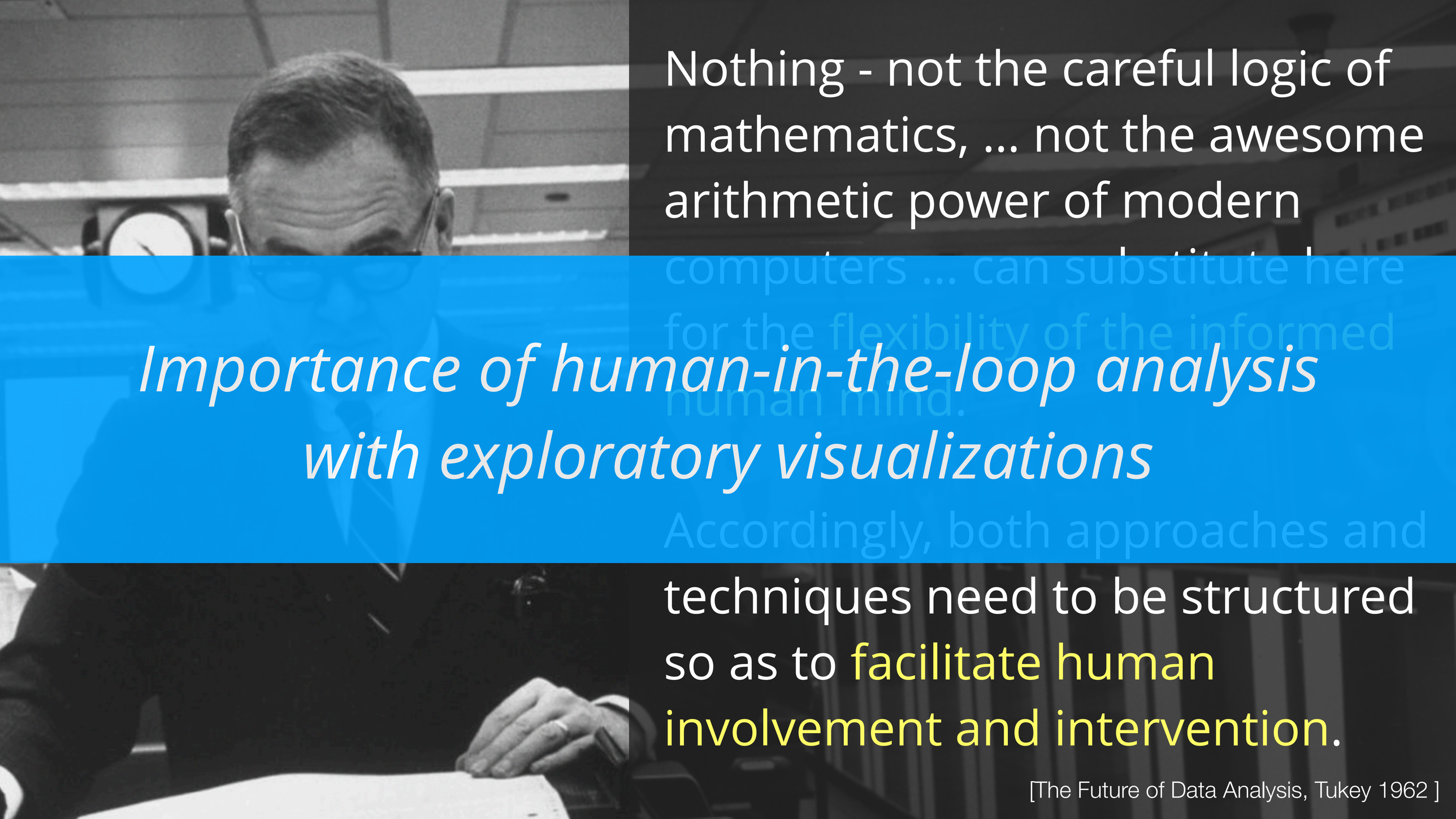
[The Future of Data Analysis, Tukey 1962]



Nothing - not the careful logic of mathematics, ... not the awesome arithmetic power of modern computers ... can substitute here for the flexibility of the informed human mind.

Accordingly, both approaches and techniques need to be structured so as to facilitate human involvement and intervention.

[The Future of Data Analysis, Tukey 1962]



Nothing - not the careful logic of mathematics, ... not the awesome arithmetic power of modern computers ... can substitute here for the flexibility of the informed human mind.

Accordingly, both approaches and techniques need to be structured so as to facilitate human involvement and intervention.

Importance of human-in-the-loop analysis with exploratory visualizations

[The Future of Data Analysis, Tukey 1962]

Anscombe's Quartet

A		B		C		D	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.8

Summary Statistics

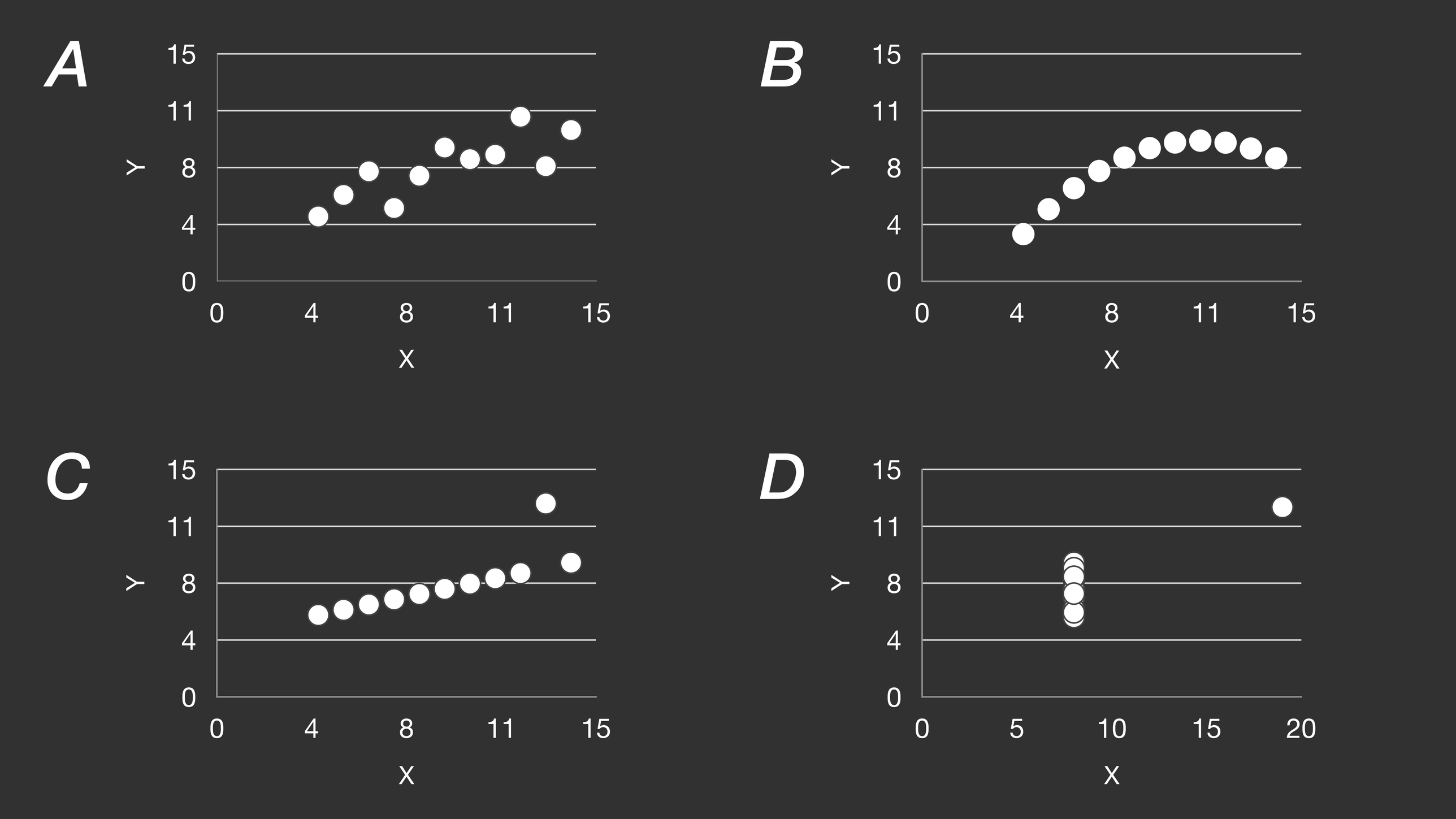
$u_X = 9.0$ $\sigma_X = 3.317$

$u_Y = 7.5$ $\sigma_Y = 2.03$

Linear Regression

$Y = 3 + 0.5\,X$

$R^2 = 0.67$


A
15
11
8
4
0
Y
0
4
8
11
15
X
B
15
11
8
4
0
Y
0
4
8
11
15
X
C
15
11
8
4
0
Y
0
4
8
11
15
X
D
15
11
8
4
0
Y
0
5
10
15
20
X

Topics

- What is exploratory analysis
- Stages of data analysis
- Exploratory analysis with Tableau

What is Exploratory Data Analysis?

An philosophy for data analysis that employs a variety of techniques (mostly graphical):

1. maximize insight into a data set
2. uncover underlying structure
3. extract important variables
4. detect outliers and anomalies
5. test underlying assumptions

http://www.itl.nist.gov/div898/handbook/eda/eda.htm

It’s Iterative Process

Ask questions

Construct graphics to address questions

Inspect “answer” and derive new questions

Repeat…

“Show data variation, not design variation” —Tufte

Acquisition

↓

Cleaning

↓

Integration

↓

Visualization

↓

Modeling

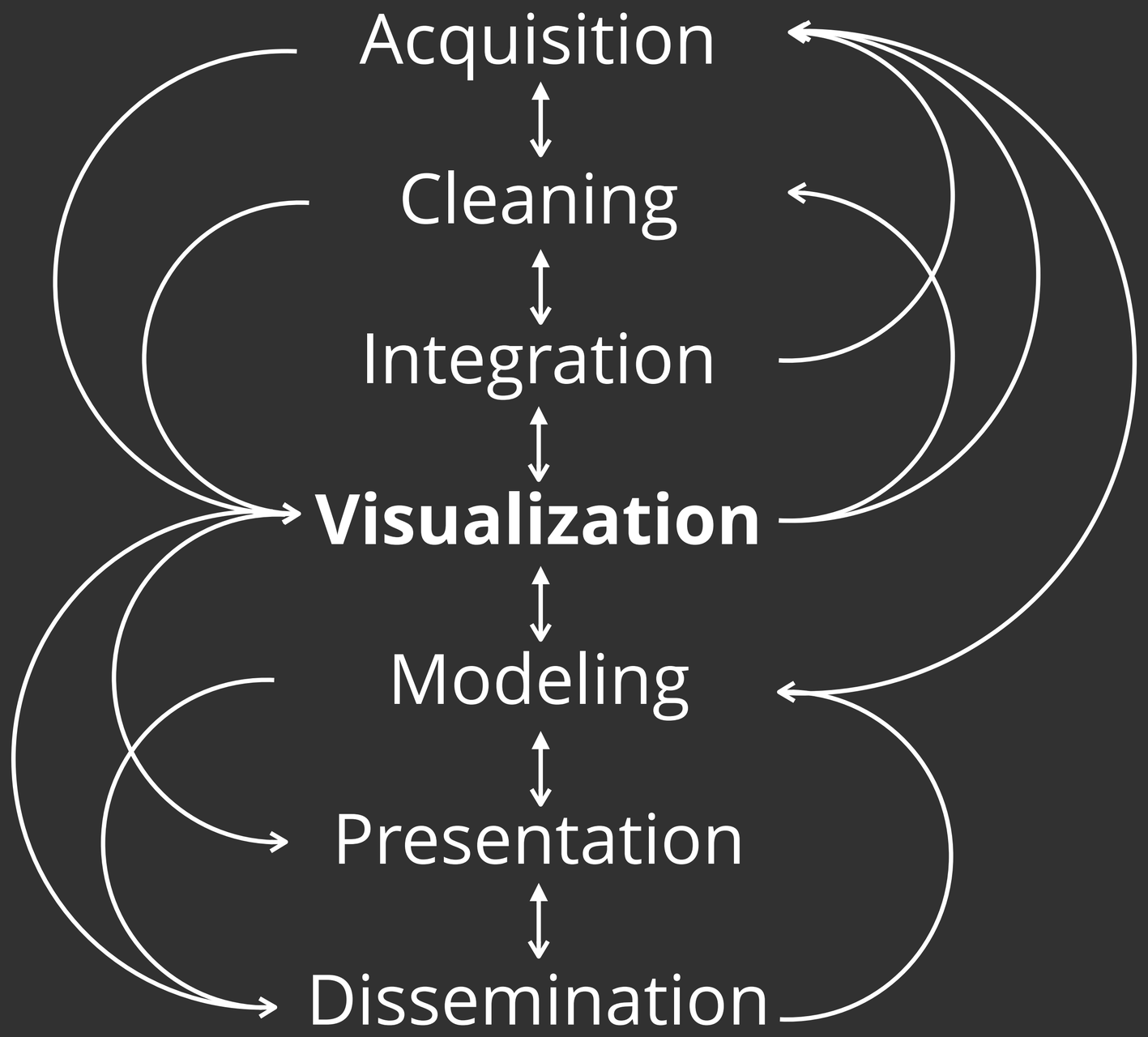
↓

Presentation

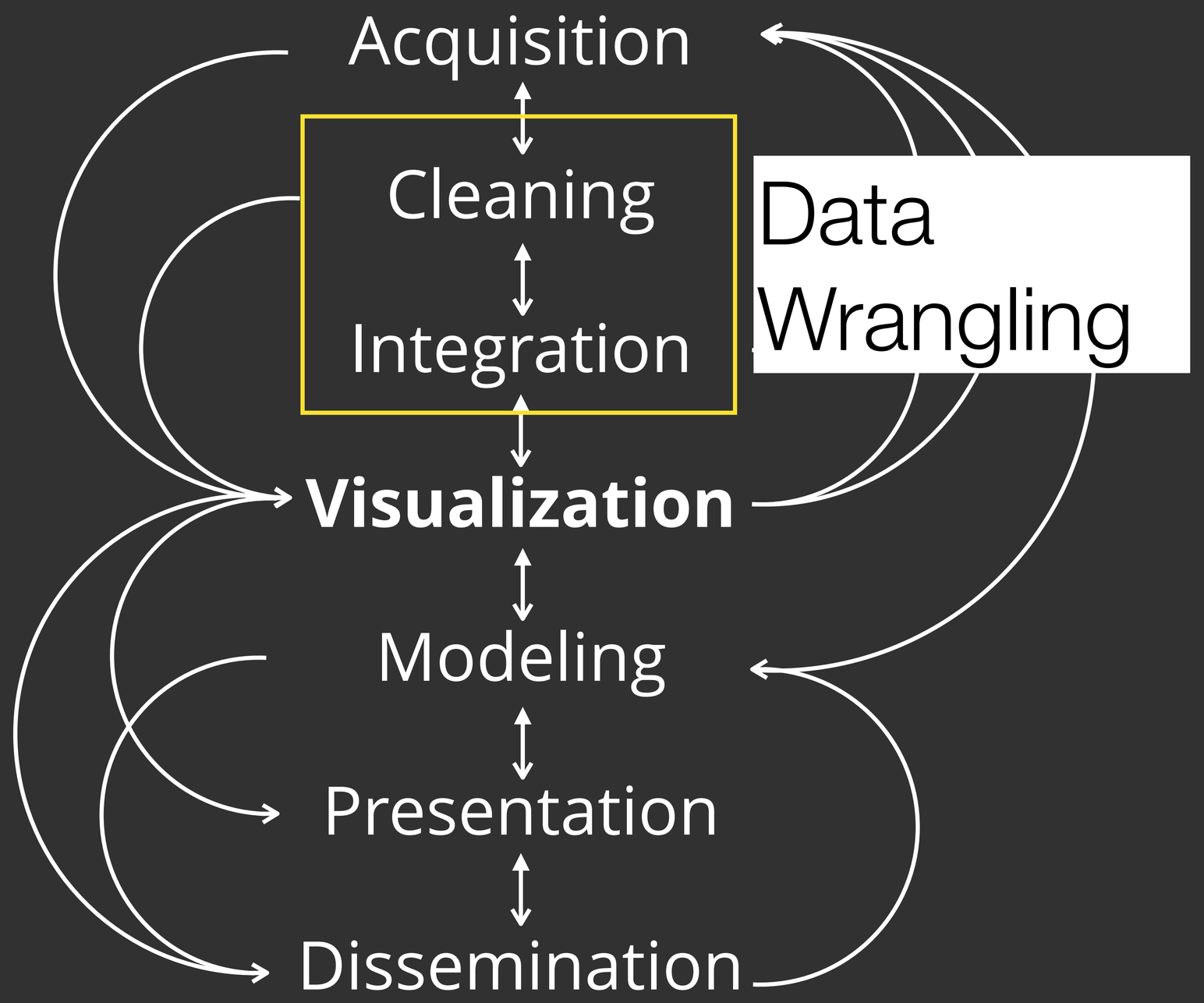
↓

Dissemination

[J. Heer]


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

[J. Heer]


Acquisition
Cleaning
Integration
Data Wrangling
Visualization
Modeling
Presentation
Dissemination

[J. Heer]

Big Data Borat
@BigDataBorat

Follow

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

6:47 PM - 26 Feb 2013

Bureau of Justice Statistics - Data Online
http://bjs.ojp.usdoj.gov/

Reported crime in Alabama

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4525375	4029.3	987	2732.4	309.9
2005	4548327	3900	955.8	2656	289
2006	4599030	3937	968.9	2645.1	322.9
2007	4627851	3974.9	980.2	2687	307.7
2008	4661900	4081.9	1080.7	2712.6	288.6

Reported crime in Alaska

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	657755	3370.9	573.6	2456.7	340.6
2005	663253	3615	622.8	2601	391
2006	670053	3582	615.2	2588.5	378.3
2007	683478	3373.9	538.9	2480	355.1
2008	686293	2928.3	470.9	2219.9	237.5

Reported crime in Arizona

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	5739879	5073.3	991	3118.7	963.5
2005	5953007	4827	946.2	2958	922
2006	6166318	4741.6	953	2874.1	914.4
2007	6338755	4502.6	935.4	2780.5	786.7
2008	6500180	4087.3	894.2	2605.3	587.8

Reported crime in Arkansas

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	2750000	4033.1	1096.4	2699.7	237
2005	2775708	4068	1085.1	2720	262
2006	2810872	4021.6	1154.4	2596.7	270.4
2007	2834797	3945.5	1124.4	2574.6	246.5
2008	2855390	3843.7	1182.7	2433.4	227.6

Reported crime in California

Bureau of Justice Statistics - Data Online
http://bjs.ojp.usdoj.gov/

Reported crime in Alabama

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4525375	4029.3	987	2732.4	309.9
2005	4548327	3900			
2006	4599030	3937			
2007	4627851	3974.			
2008	4661900	4081.			

Reported crime in Ala

Year	Population				vehicle theft rate
2004	657755	3370.			
2005	663253	3615			
2006	670053	3582			
2007	683478	3373.			
2008	686293	2928.			

Reported crime in Ari

Year	Population				vehicle theft rate
2004	5739879	5073.			
2005	5953007	4827			
2006	6166318	4741.			
2007	6338755	4502.			
2008	6500180	4087.			

Reported crime in Ark

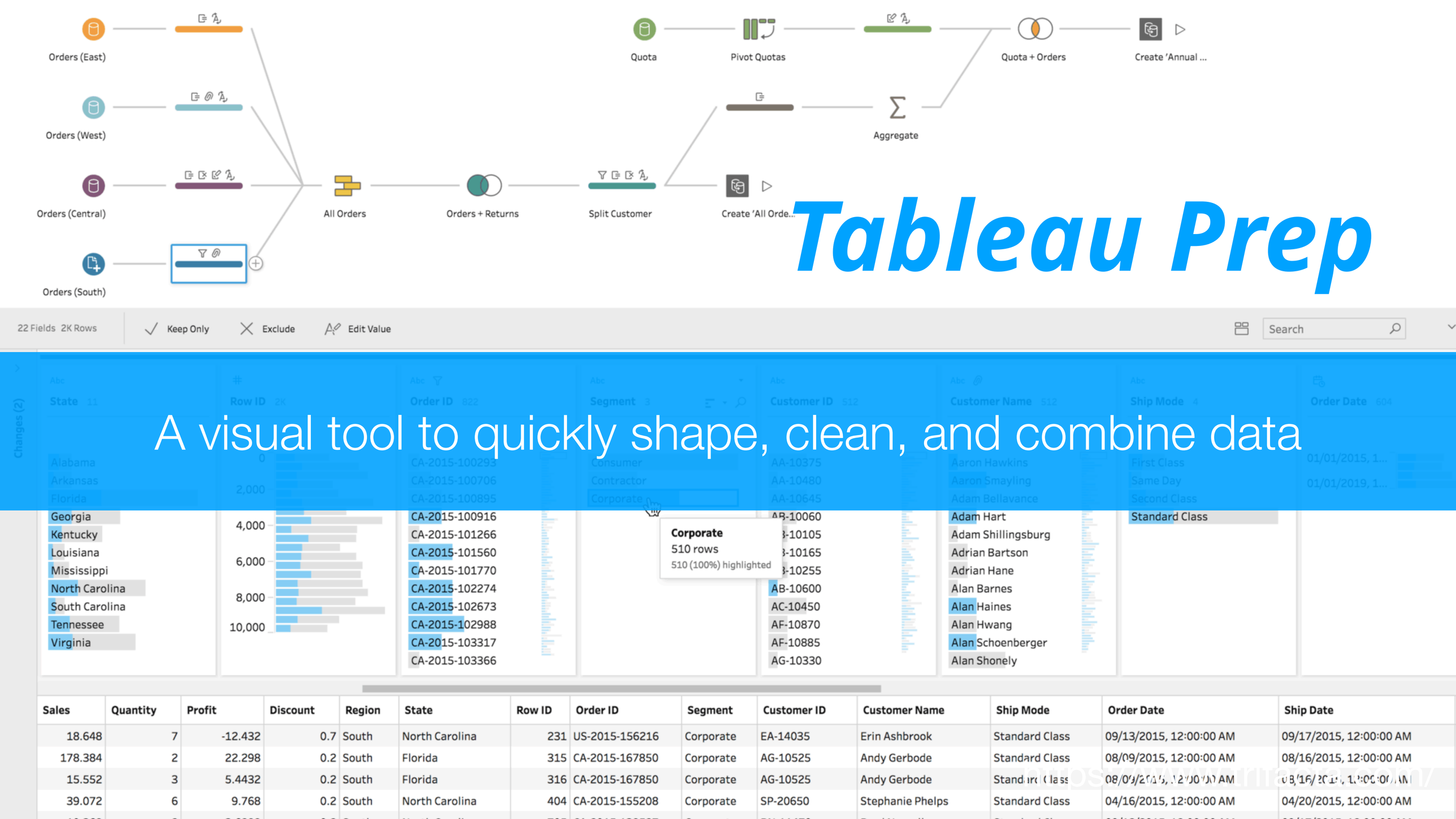
Year	Population				vehicle theft rate
2004	2750000	4033.1	1096.4	2699.7	237
2005	2775708	4068	1085.1	2720	262
2006	2810872	4021.6	1154.4	2596.7	270.4
2007	2834797	3945.5	1124.4	2574.6	246.5
2008	2855390	3843.7	1182.7	2433.4	227.6

Reported crime in California

	Year	State	Property_crime_rate
0	2004	Alabama	4029.3
1	2005	Alabama	3900
2	2006	Alabama	3937
3	2007	Alabama	3974.9
4	2008	Alabama	4081.9
5	2004	Alaska	3370.9
6	2005	Alaska	3615
7	2006	Alaska	3582
8	2007	Alaska	3373.9
9	2008	Alaska	2928.3
10	2004	Arizona	5073.3
11	2005	Arizona	4827
12	2006	Arizona	4741.6
13	2007	Arizona	4502.6
14	2008	Arizona	4087.3

Data Quality Hurdles

Missing Data	no measurements, redacted, ...?
Erroneous Values	misspelling, outliers, ...?
Type Conversion	e.g., zip code to lat-lon
Entity Resolution	diff. values for the same thing?
Data Integration	effort/errors when combining data



A visual tool to quickly shape, clean, and combine data

https://www.tableau.com/

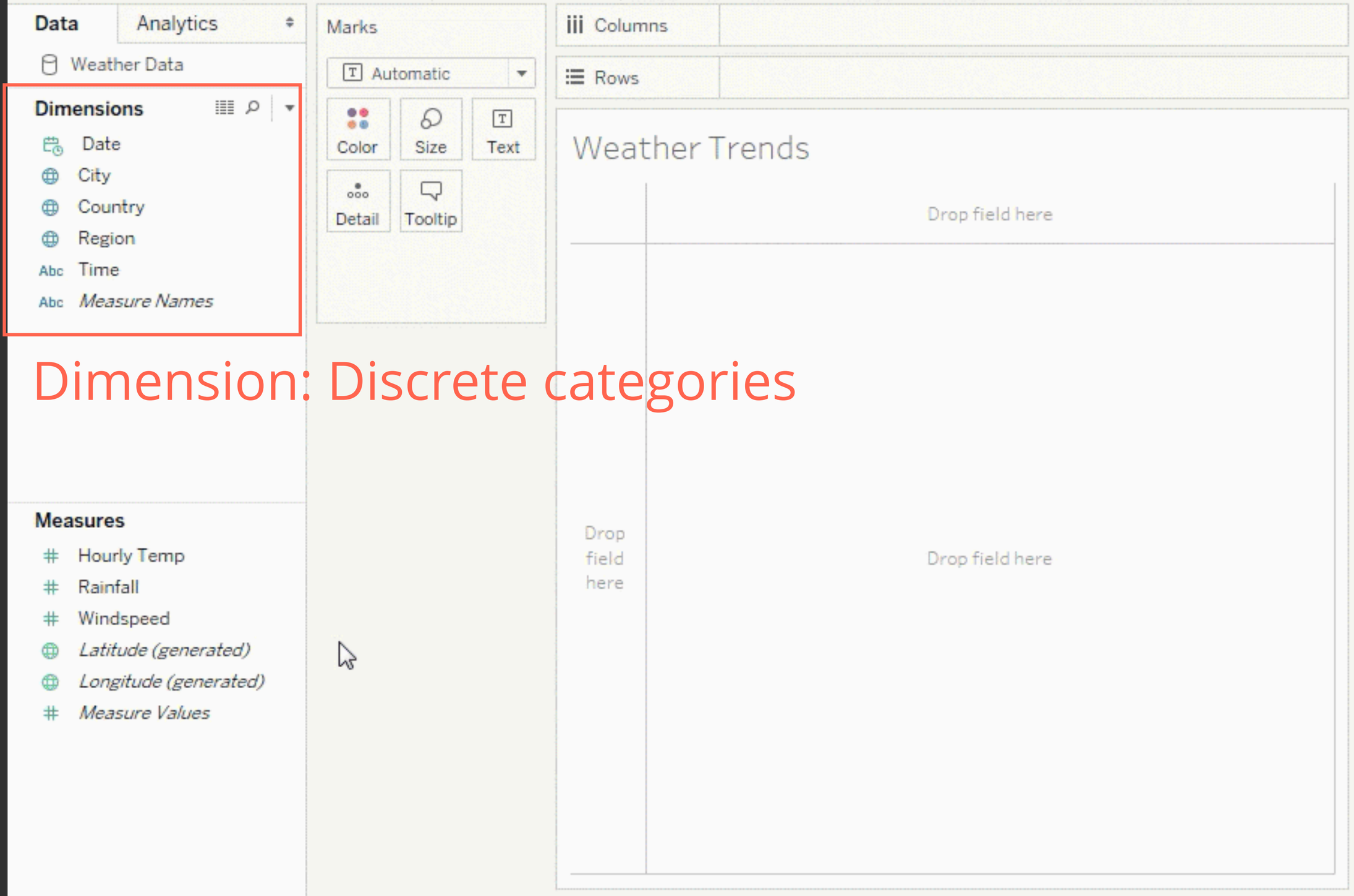
Exploratory Analysis with Tableau

What is Tableau?

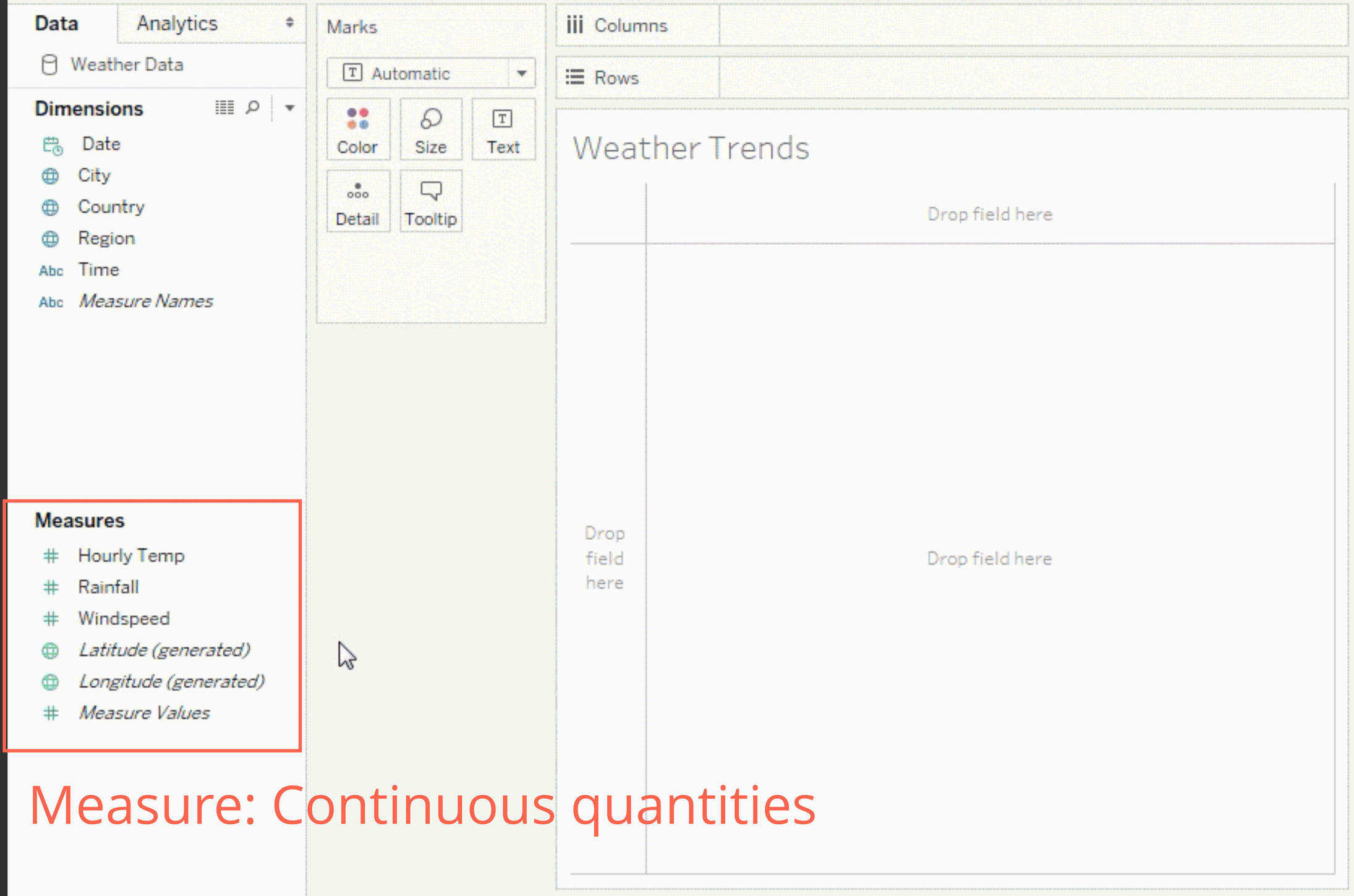
Software to rapidly construct visualizations of data and perform exploratory analysis of data

Download: https://public.tableau.com

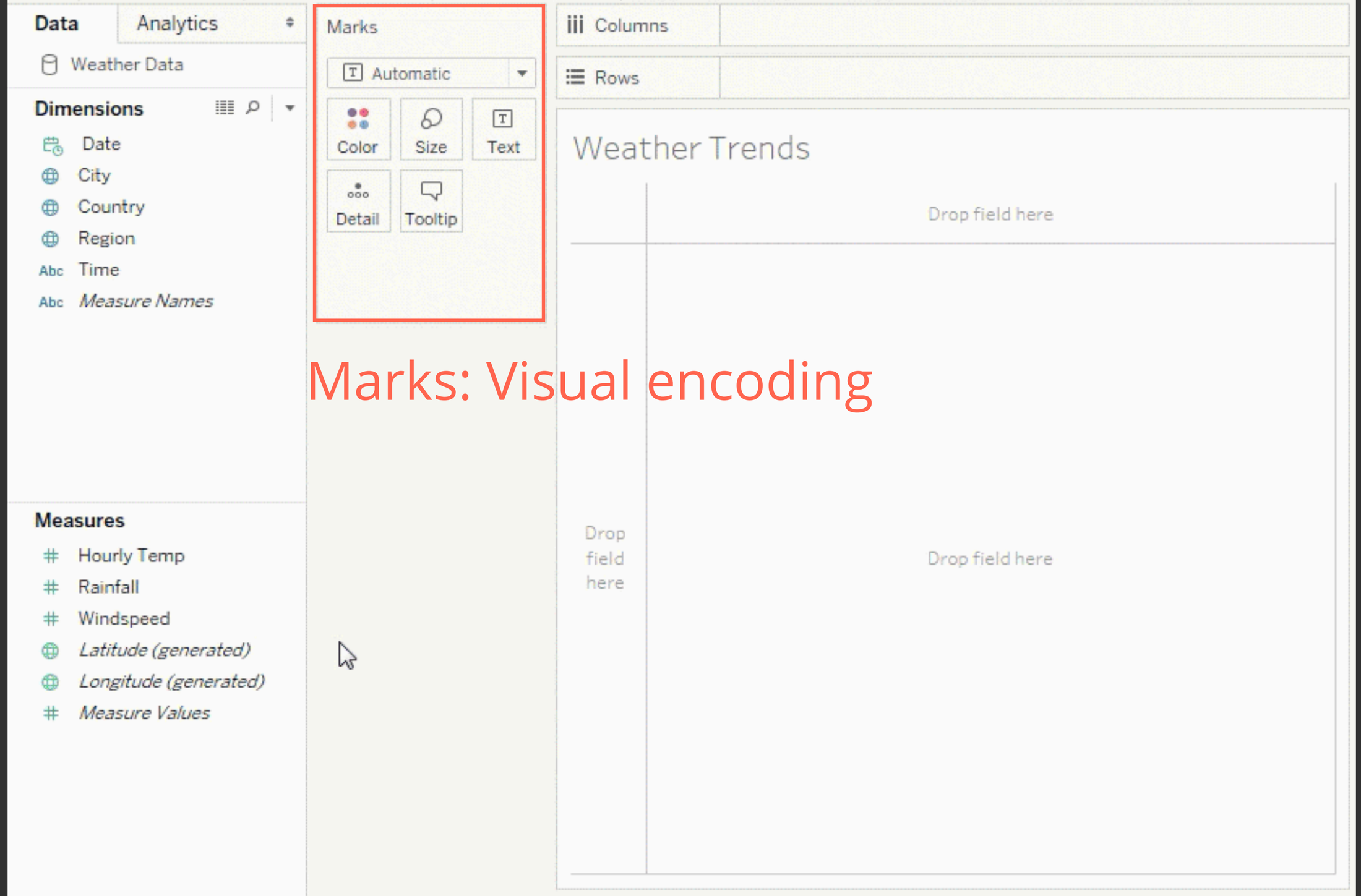
Dataset: http://www.namwkim.org/datavis/h1b_kaggle_sample.csv



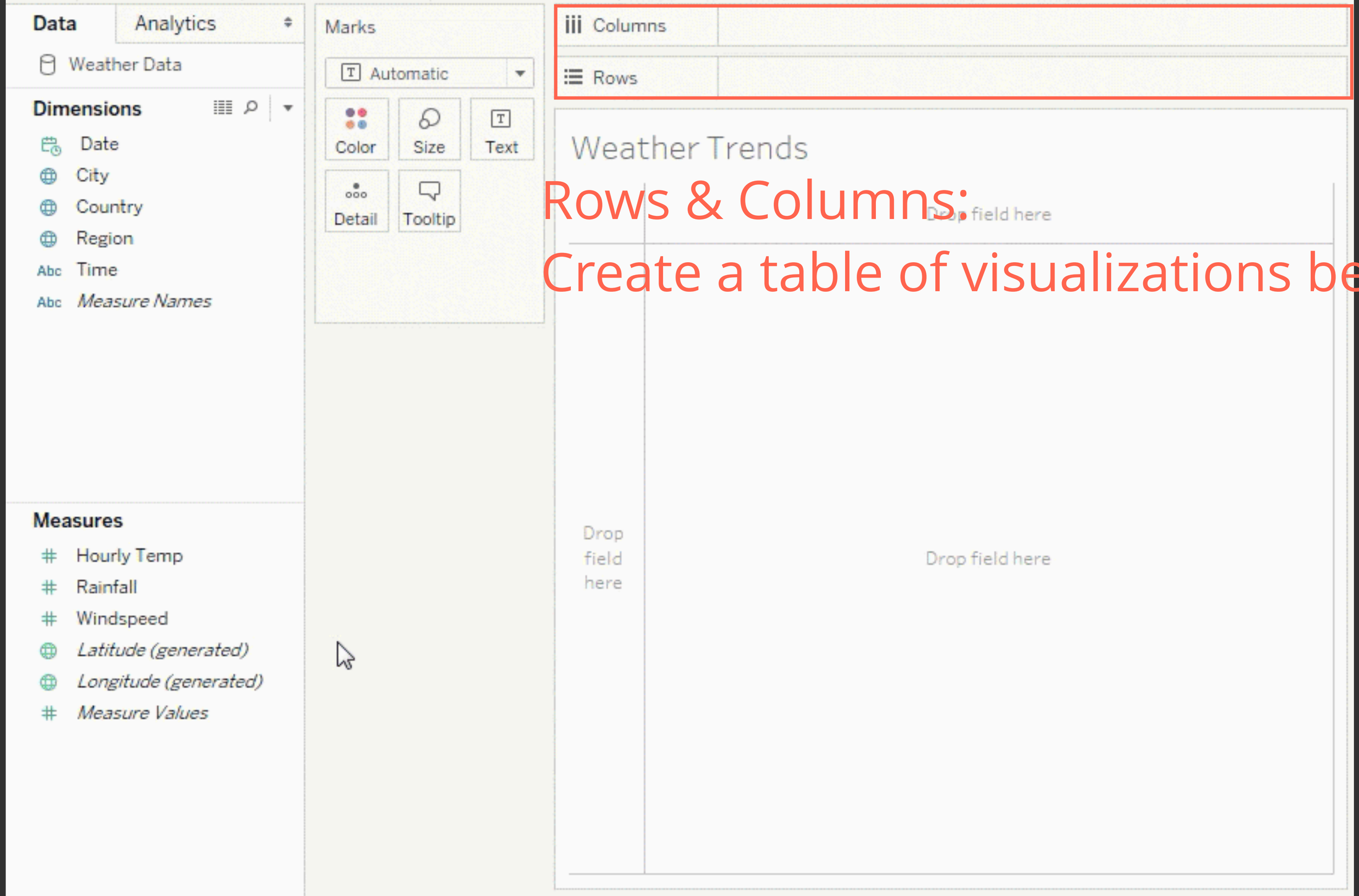
Dimension: Discrete categories



Measure: Continuous quantities

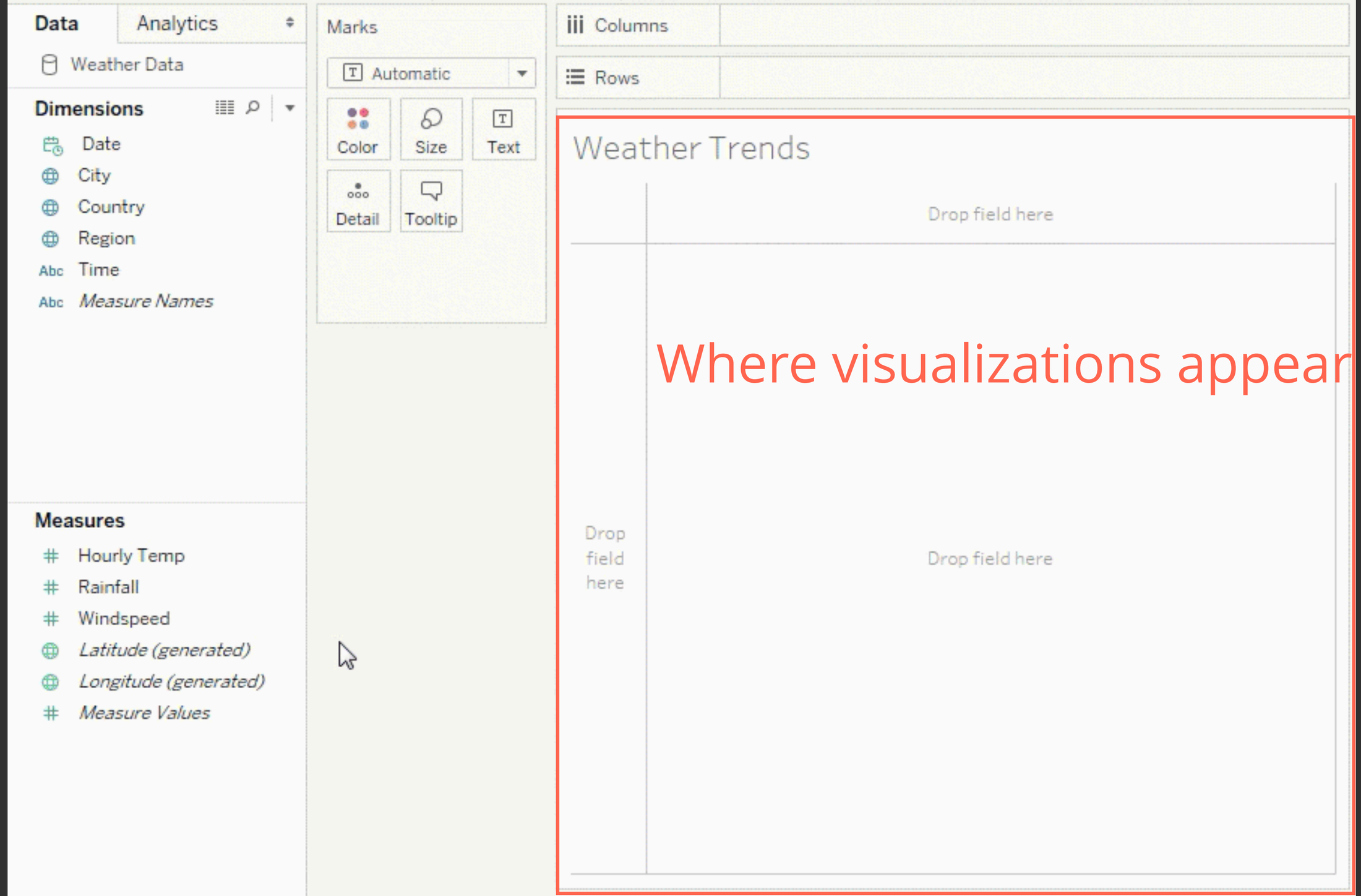


Marks: Visual encoding



Rows & Columns:

Create a table of visualizations below


Data
Analytics
Weather Data
Dimensions
Date
City
Country
Region
Time
Measure Names
Measures
Hourly Temp
Rainfall
Windspeed
Latitude (generated)
Longitude (generated)
Measure Values
Marks
Automatic
Color
Size
Text
Detail
Tooltip
Columns
Rows
Weather Trends
Drop field here
Drop field here
Drop field here
Where visualizations appear

Data | Analytics

Weather Data

Dimensions

- Date
- City
- Country
- Region
- Time
- *Measure Names*

Measures

- Hourly Temp
- Rainfall
- Windspeed
- *Latitude (generated)*
- *Longitude (generated)*
- *Measure Values*

Marks

Automatic

Color | Size | Text

Detail | Tooltip

Columns

Rows

Weather Trends

Drop field here

Drop field here

Drop field here

Analysis Example:
H-1B Visa Petitions 2011-2016

Dataset: H1B Visa Petitions (2011-16)

H1B is a Employment-based, non-immigrant visa category for temporary foreign workers

The raw data was published by The Office of Foreign Labor Certification (OFLC)

The data was cleaned by Sharan Naribole, featured on Kaggle: https://www.kaggle.com/nsharan/h-1b-visa

Dataset: H1B Visa Petitions (2011-16)

CASE_STATUS (N): “Certified” (means eligible not approved) “Denied”....

EMPLOYER_NAME (N) — Company submitting this petition

SOC_NAME (N) — Standard occupational name

JOB_TITLE (N) — Title of the job

FULL_TIME_POSITION (N) — Y = Full Time Position; N = Part Time Position

PREVAILING_WAGE (Q) — the average wage paid to similar workers in the company

YEAR (O): Year in which the H-1B visa petition was filed

WORKSITE (N): City and State information of the foreign worker's intended area of employment

lon (Q): longitude of the Worksite

lat (Q): latitude of the Worksite

Dataset: H1B Visa Petitions (2011-16)

CASE_STATUS (N): “Certified” (means eligible not approved) “Denied” ....

EMPLOYER_NAME (N) — Company submitting this petition

SOC_NAME (N) — Standard Occupational Name

JOB_TITLE (N) — Job Title of the position

FULL_TIME_POSITION (N) — Y = Full Time Position; N = Part Time Position

PREVAILING_WAGE (Q) — the average wage paid to similar workers in the company

YEAR (Q): Year in which the H-1B visa petition was filed

WORKSITE (N): City and State information of the foreign worker's intended area of employment

lon (Q): longitude of the Worksite

lat (Q): latitude of the Worksite

3 million records of H-1B Visa Petitions

492MB!!

Dataset: H1B Visa Petitions (2011-16)

~~**CASE_STATUS (N)**: "**Certified**" (means eligible not approved) "Denied"....~~

EMPLOYER_NAME (N) — Company submitting this petition

SOC_NAME (N) — Standard occupational name

JOB_TITLE (N) — Title of the job

~~**FULL_TIME_POSITION (N)** —~~ **Y = Full Time Position**~~; N = Part Time Position~~

PREVAILING_WAGE (Q) — the average wage paid to similar workers in the company

YEAR (O): Year in which the H-1B visa petition was filed

~~**WORKSITE (N)**:~~ **City and State** ~~information of the foreign worker's intended area of employment~~

City (N)

State (N)

~~**lon (Q)**: longitude of the Worksite~~ **Tableau can infer this from worksite**

~~**lat (Q)**: latitude of the Worksite~~

Dataset: H1B Visa Petitions (2011-16)

~~**CASE_STATUS (N)**: "**Certified**" (means eligible not approved) "Denied"....~~

EMPLOYER_NAME (N) — Company submitting this petition

SOC_NAME (N) — Standard occupational name

JOB_TITLE (N) — Title of the job

~~**FULL_TIME_POSITION (N)** —~~ **Y = Full Time Position**~~; N = Part Time Position~~

PREVAILING_WAGE (Q) — the average wage paid to similar workers in the company

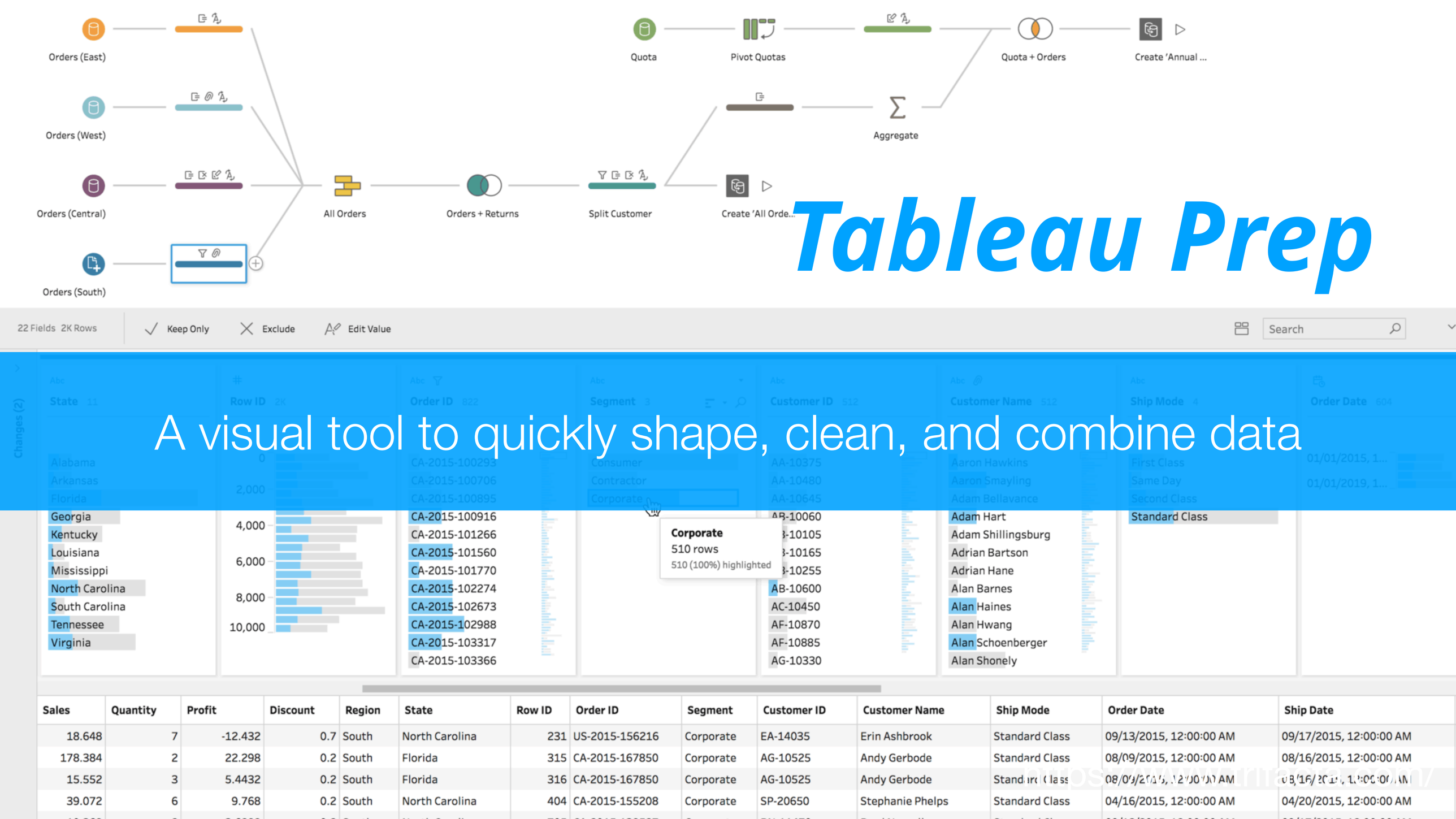
YEAR (O): Year in which the H-1B visa petition was filed

~~**WORKSITE (N)**: **City and State** information of the foreign worker's intended area of employment~~

And removed rows of missing data
and randomly sampled 40% of the whole data

~~**lon (Q)**: longitude of the Worksite~~

~~**lat (Q)**: latitude of the Worksite~~



A visual tool to quickly shape, clean, and combine data

https://www.tableau.com/

Dataset: H1B Visa Petitions (2011-16)

EMPLOYER_NAME (N) — Company submitting this petition

SOC_NAME (N) — Standard occupational name

JOB_TITLE (N) — Title of the job

PREVAILING_WAGE (Q) — the average wage paid to workers

YEAR (O): Year in which the H-1B visa petition was filed

City (N): City of the worksite

State (N): State of the worksite

~20MB

Questions

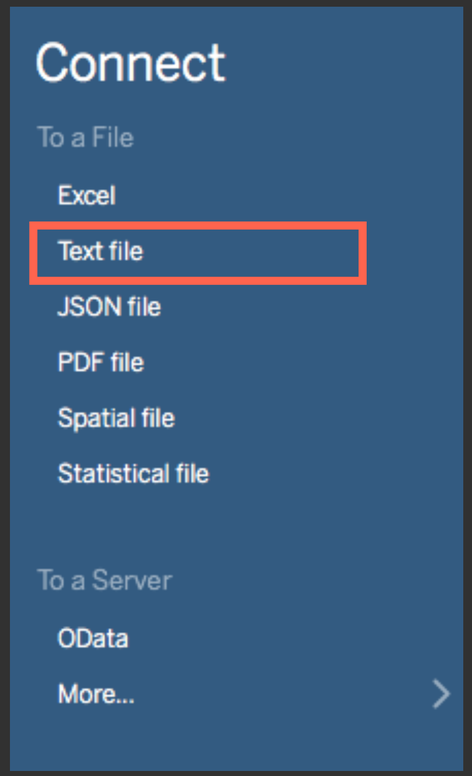
What might we learn from this data?

- Do petitions increase over time?
- Which company files petitions the most?
- What kind of job is the most applied?
- Which company offers the highest salary?
- What kind of job is offered the highest salary?
- Which states/cities file petitions the most?
- What are differences in salaries across states & cities?
- What is the relationship between salaries and petitions?

Tableau Demo

Load data

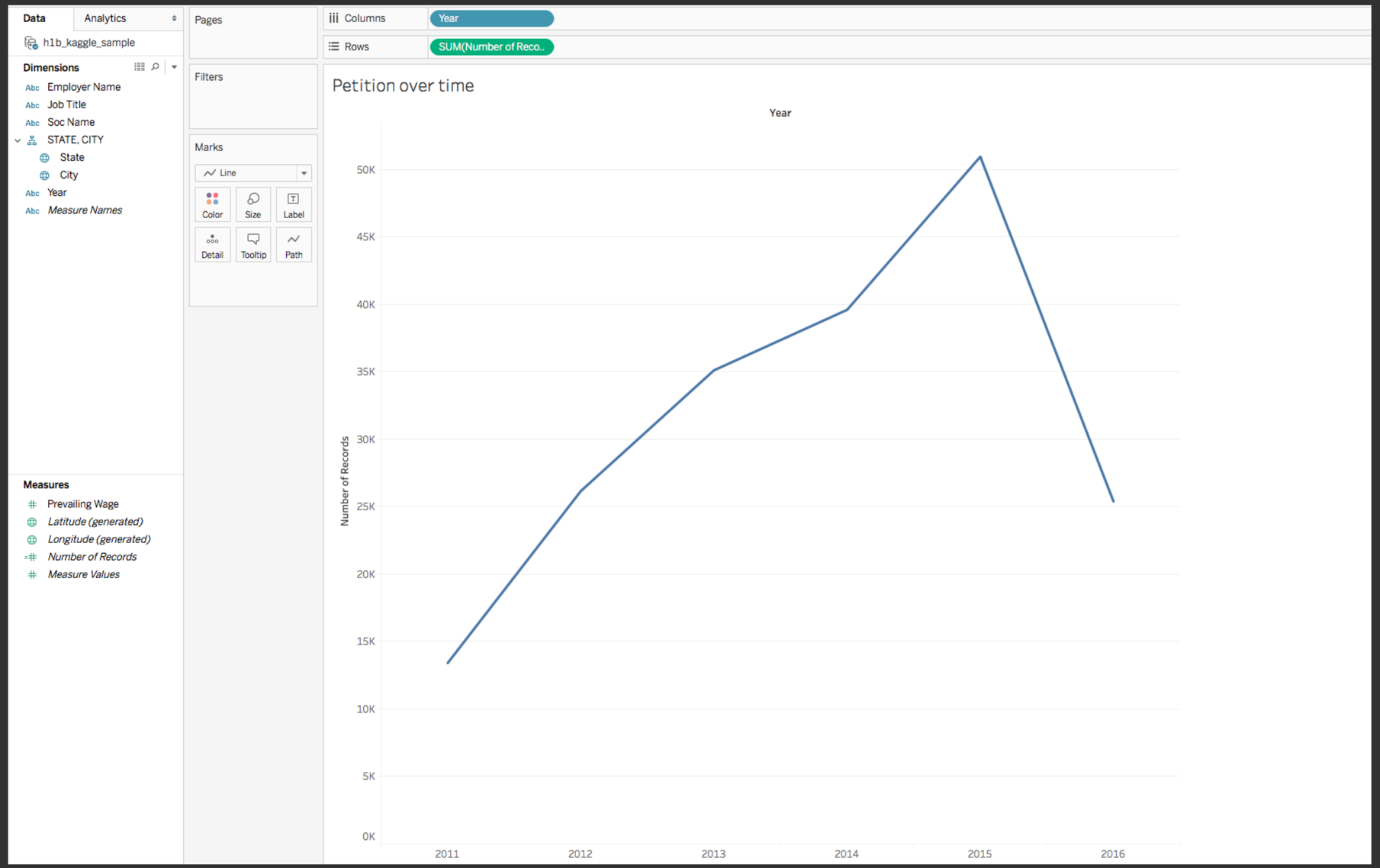
Change Year to String Type



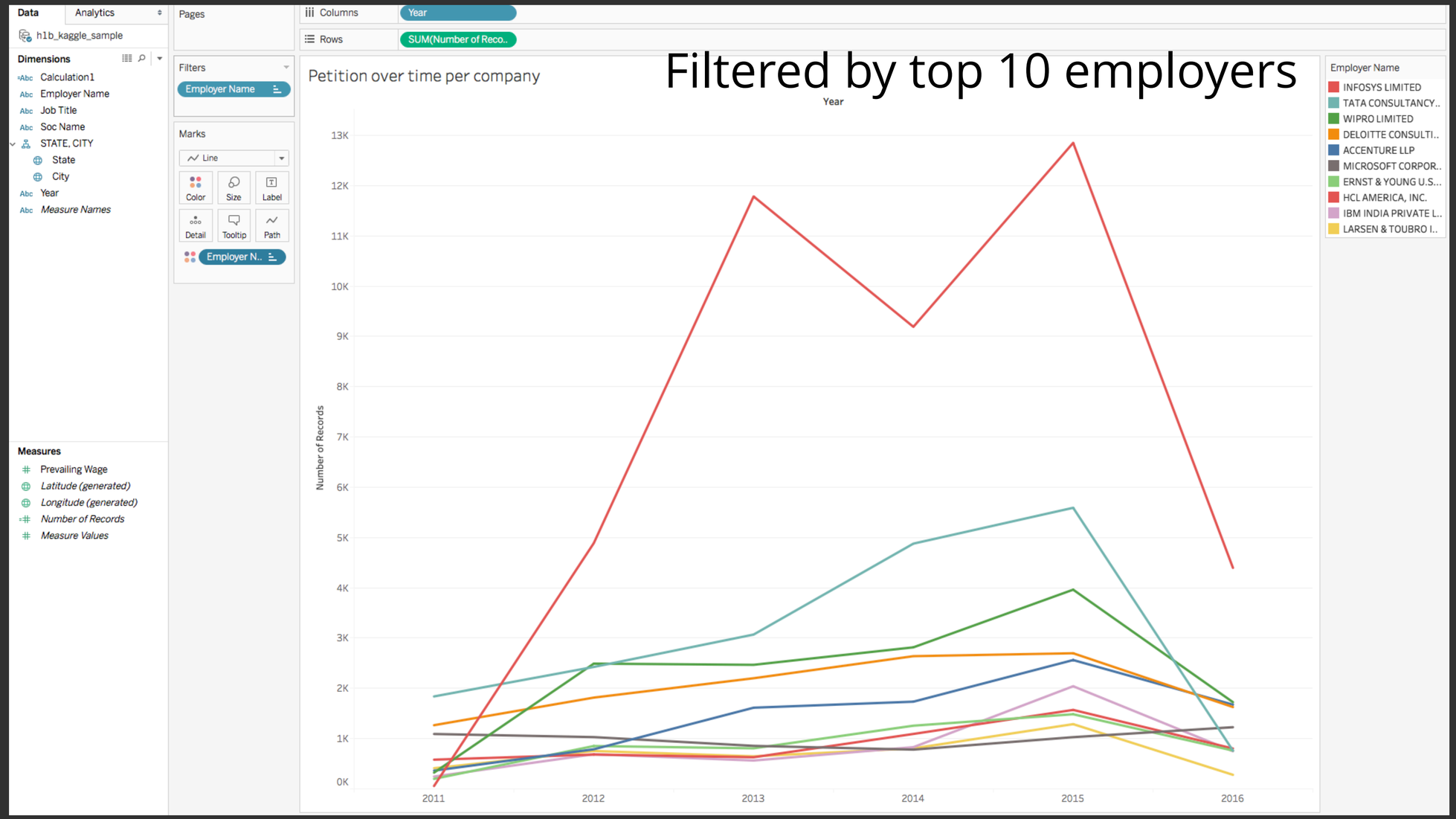
Sort fields Data source order

Abc h1b_kaggle_sample.csv Employer Name	Abc h1b_kaggle_sample.csv Soc Name	Abc h1b_kaggle_sample.csv Job Title	# h1b_kaggle_sample.csv Prevailing Wage	Abc h1b_kaggle_sa... Year	h1b_kaggle_sample.csv City	h1b_kaggle_sampl... State
WAL-MART ASSOCIA...	Computer Systems Analysts	PROGRAMMER ANALYST	40,061.00	2011	BENTONVILLE	ARKANSAS
KPMG LLP	Accountants and Auditors	MANAGER	81,640.00	2011	SAN FRANCISCO	CALIFORNIA
LARSEN & TOUBRO LI...	Commercial and Industrial De...	INDUSTRIAL DESIGNER	39,437.00	2011	PLAYA VISTA	CALIFORNIA
LARSEN & TOUBRO I...	Computer Programmers	COMPUTER PROGRAMMER	54,870.00	2011	SAN DIEGO	CALIFORNIA
GOOGLE INC.	Computer Software Engineers...	SOFTWARE ENGINEER	90,480.00	2011	SAN BRUNO	CALIFORNIA
MICROSOFT CORPOR...	Computer Software Engineers...	SOFTWARE DEVELOPMENT ENGI...	98,530.00	2011	MOUNTAIN VIEW	CALIFORNIA
CAPGEMINI U.S. LLC	Computer Software Engineers...	CONSULTANT	66,602.00	2011	BURBANK	CALIFORNIA
DELOITTE CONSULTI...	Computer Software Engineers...	SENIOR CONSULTANT	83,512.00	2011	IRWINDALE	CALIFORNIA
DELOITTE CONSULTI...	Computer Software Engineers...	SPECIALIST SENIOR	71,490.00	2011	RANCHO CORDOVA	CALIFORNIA
INTEL CORPORATION	Computer Software Engineers...	SOFTWARE ENGINEER	124,363.00	2011	SANTA CLARA	CALIFORNIA
MICROSOFT CORPOR...	Computer Software Engineers...	SOFTWARE DEVELOPMENT ENGI...	85,904.00	2011	MOUNTAIN VIEW	CALIFORNIA
HCL AMERICA, INC.	Computer Systems Analysts	SYSTEMS ANALYST	58,427.00	2011	SAN JOSE	CALIFORNIA
PERSISTENT SYSTEM...	Computer Systems Analysts	PROGRAMMER ANALYST	63,107.00	2011	REDWOOD CITY	CALIFORNIA
UST GLOBAL INC.	Computer Systems Analysts	SYSTEMS ANALYST	68,682.00	2011	WOODLAND HILLS	CALIFORNIA
INTEL CORPORATION	Electronics Engineers, Except ...	HARDWARE ENGINEER	86,732.00	2011	SANTA CLARA	CALIFORNIA
LARSEN & TOUBRO I...	Management Analysts	BUSINESS SYSTEMS ANALYST	44,387.00	2011	SANTA ANA	CALIFORNIA
LARSEN & TOUBRO LI...	Commercial and Industrial De...	INDUSTRIAL DESIGNER	34,278.00	2011	NORTH HAVEN	CONNECTICUT
ACCENTURE LLP	Computer Programmers	COMPUTER PROGRAMMER/CON...	71,885.00	2011	HARTFORD	CONNECTICUT
V-SOFT CONSULTING	Computer Systems Analysts	SYSTEMS ANALYST	63,648.00	2011	WINDSOR	CONNECTICUT

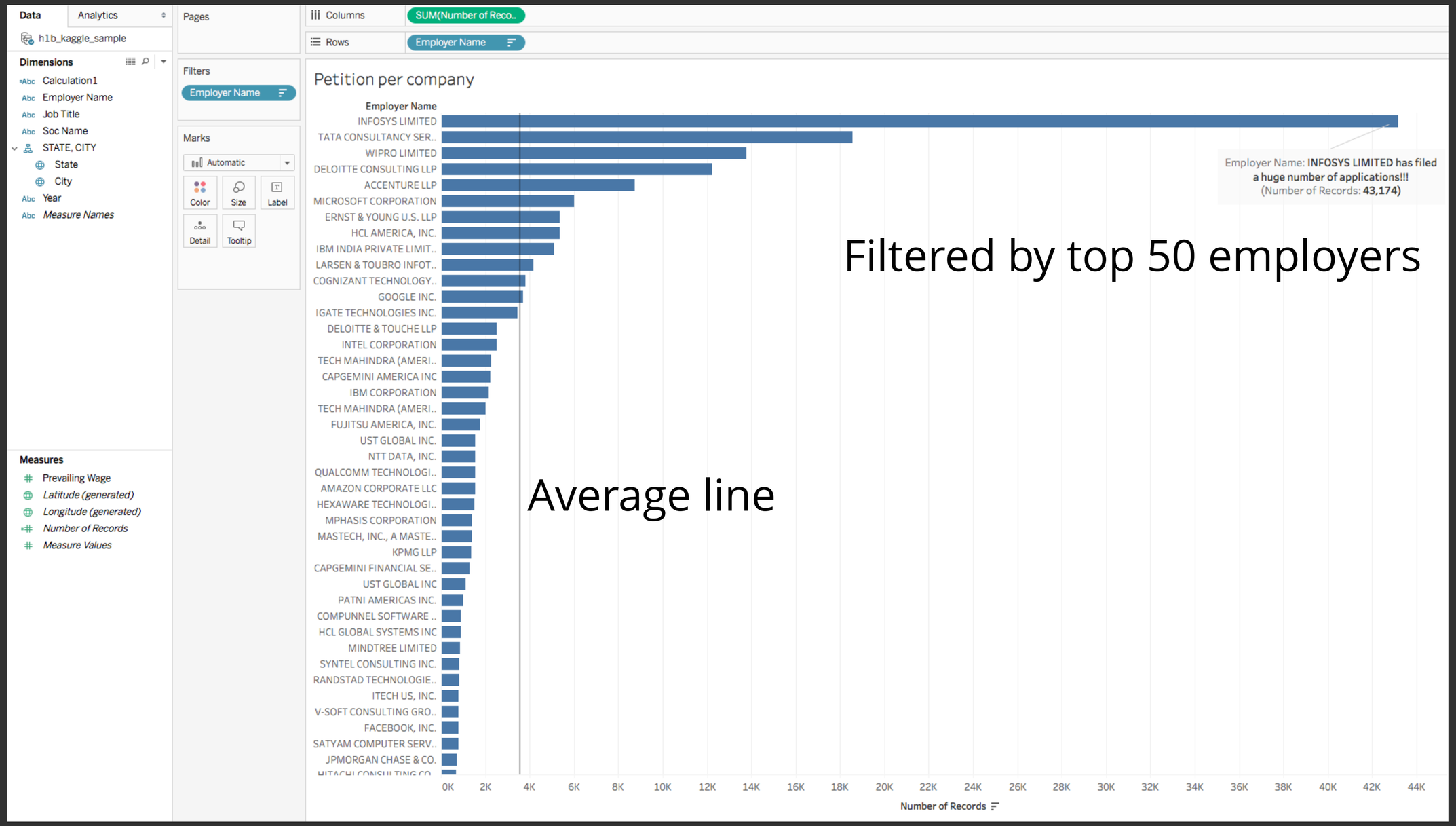
Do petitions increase over time?



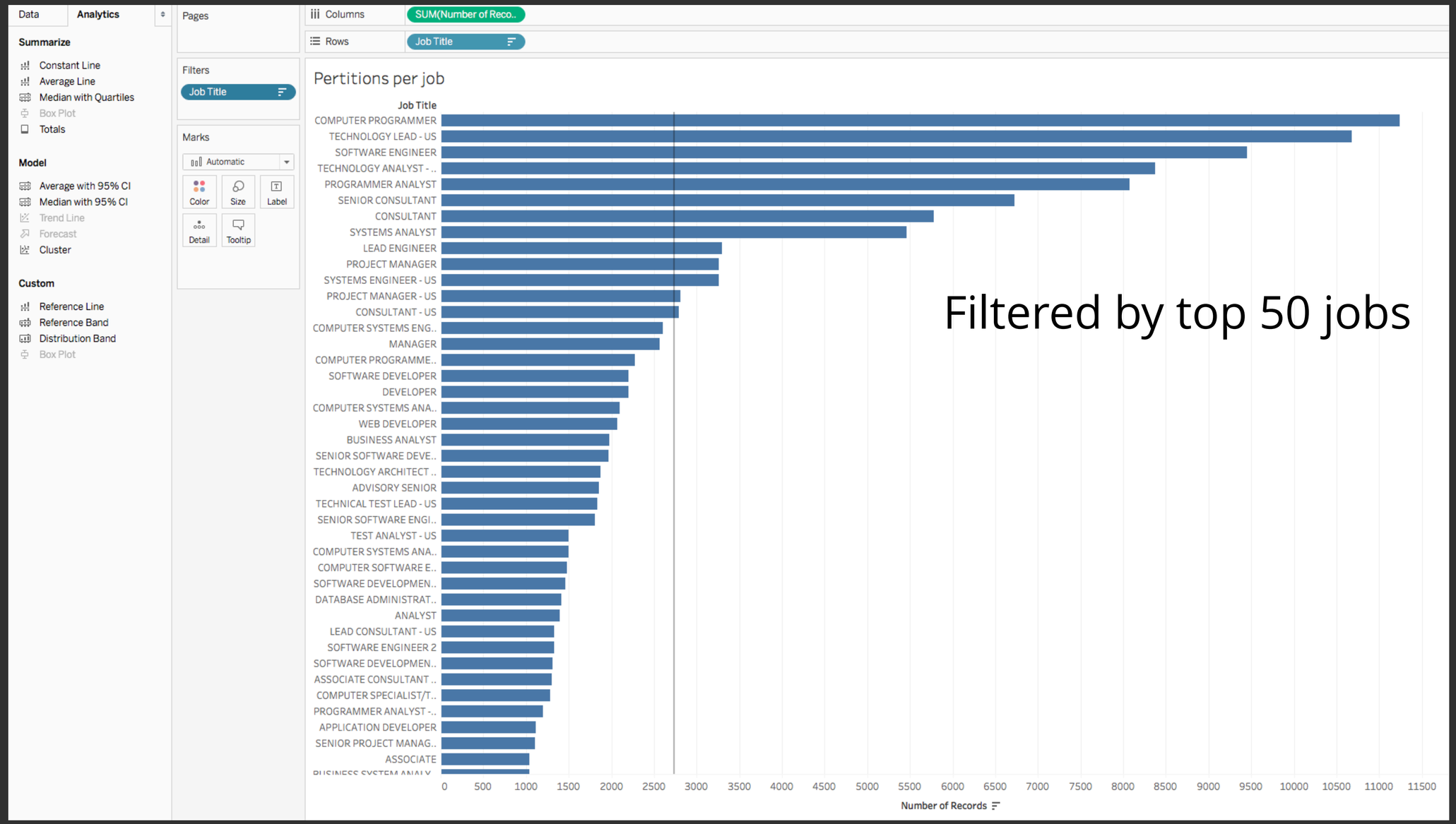
Do petitions increase over time?



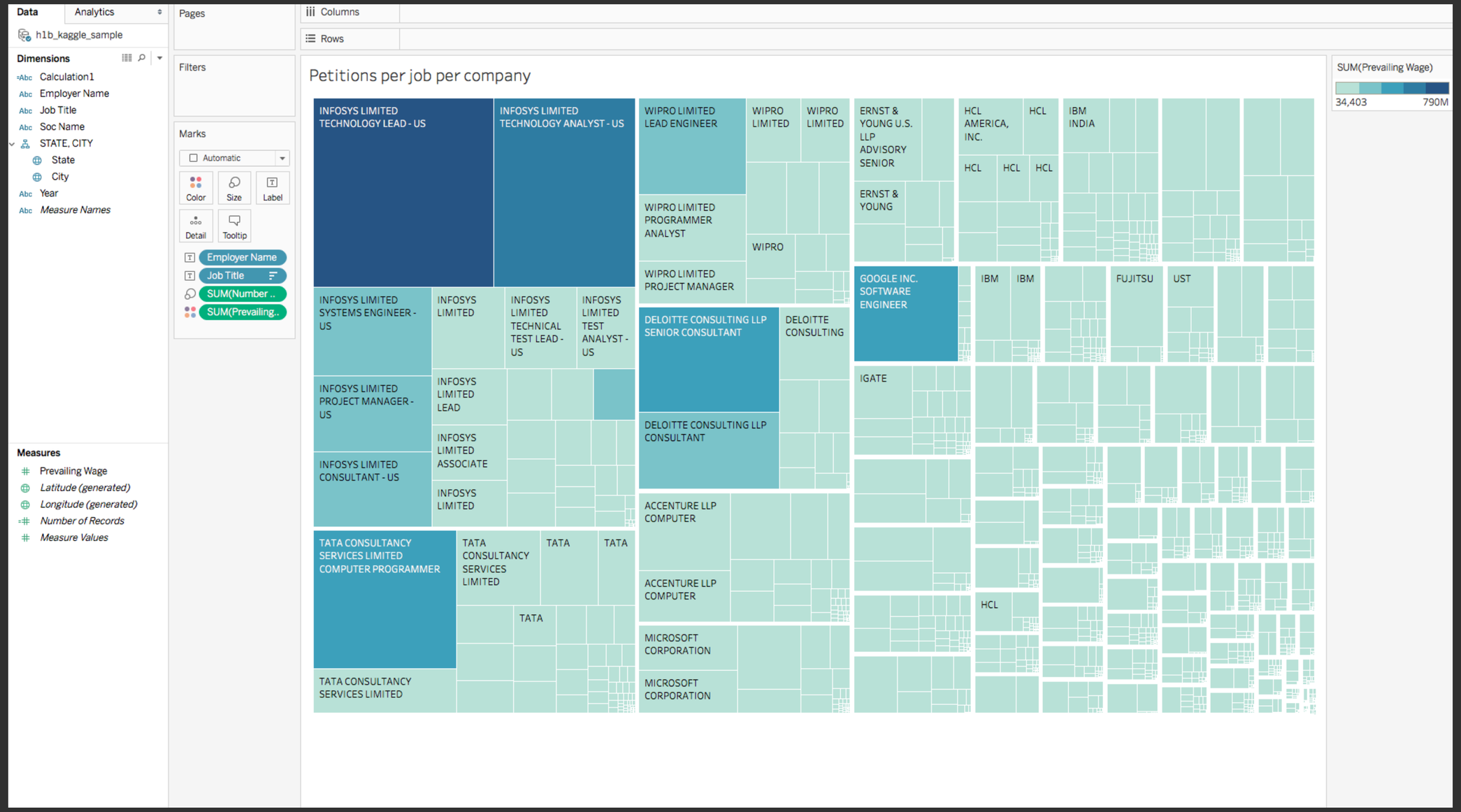
Which company files petitions the most?



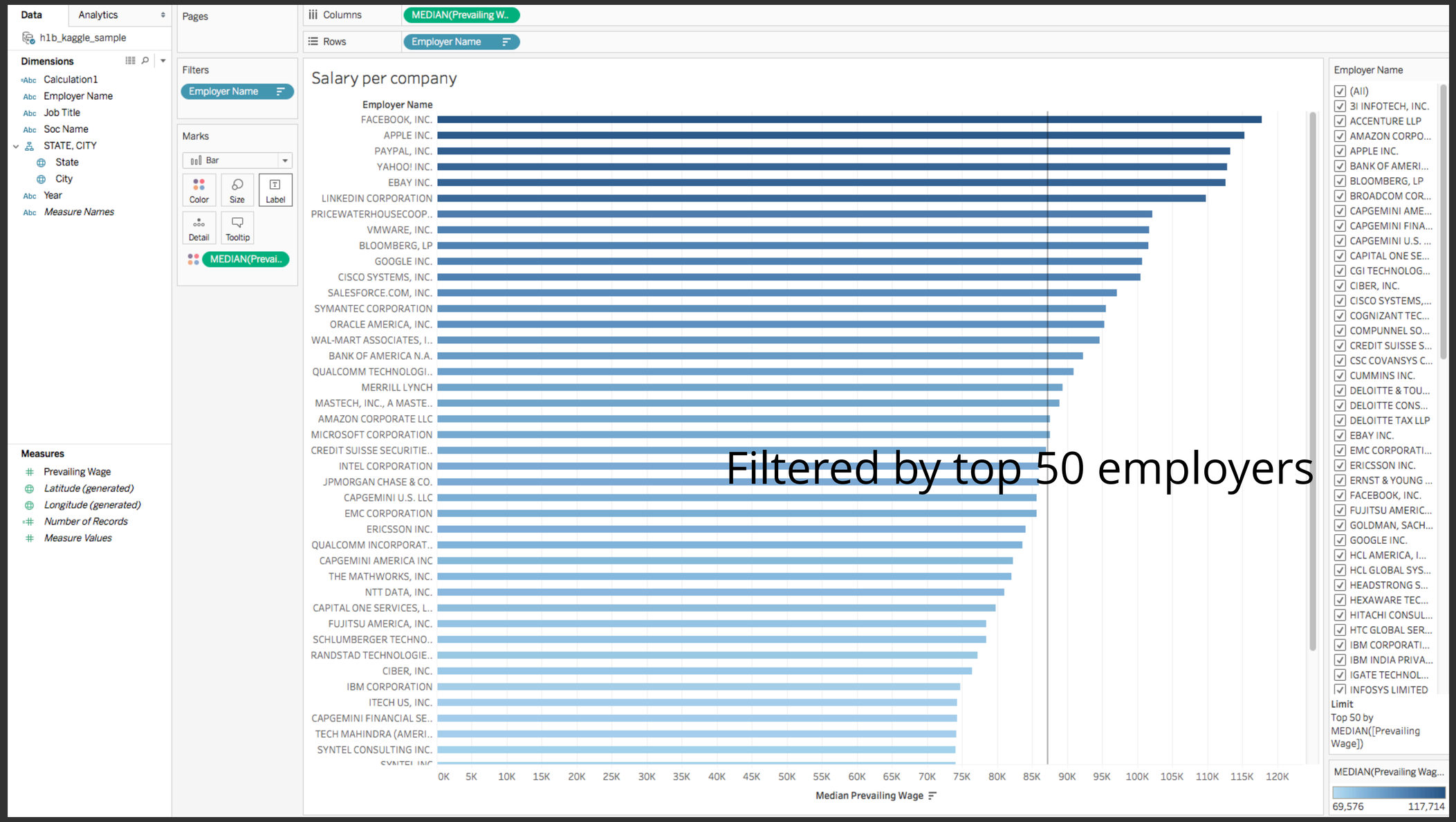
What kind of job is the most applied?



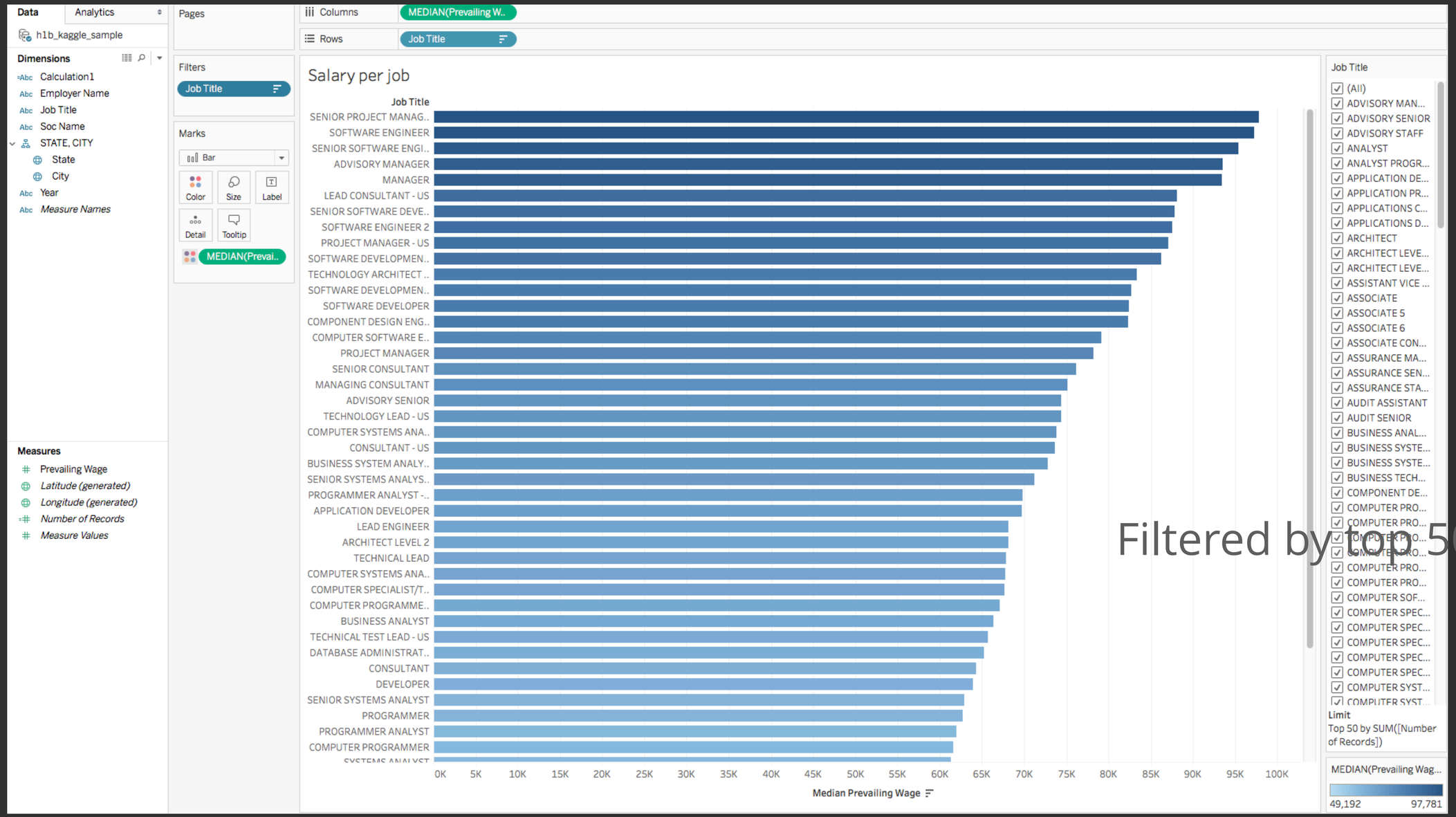
What kind of job is the most applied?



Which company offers the highest salary?

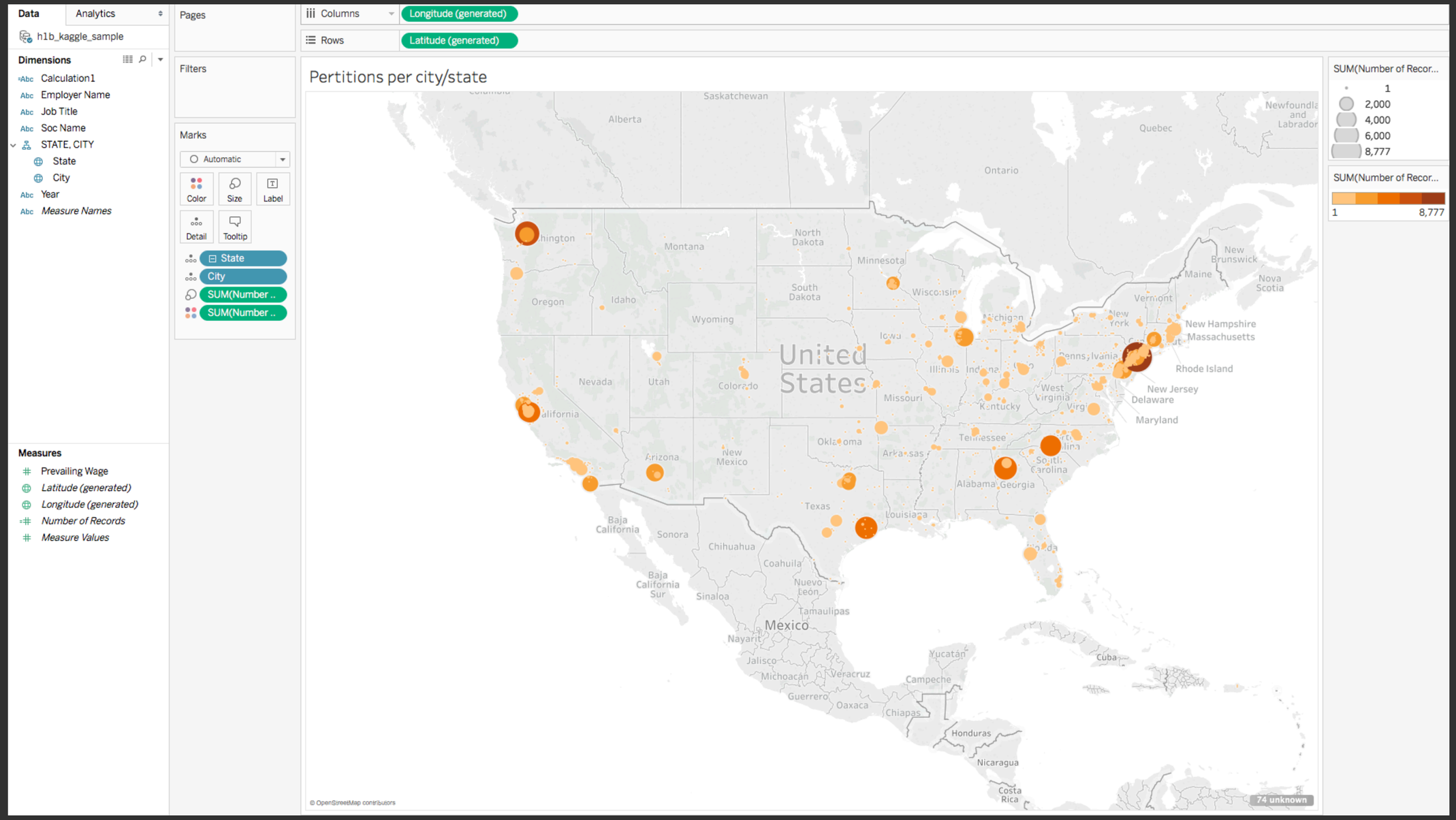


What kind of job is offered the highest salary?

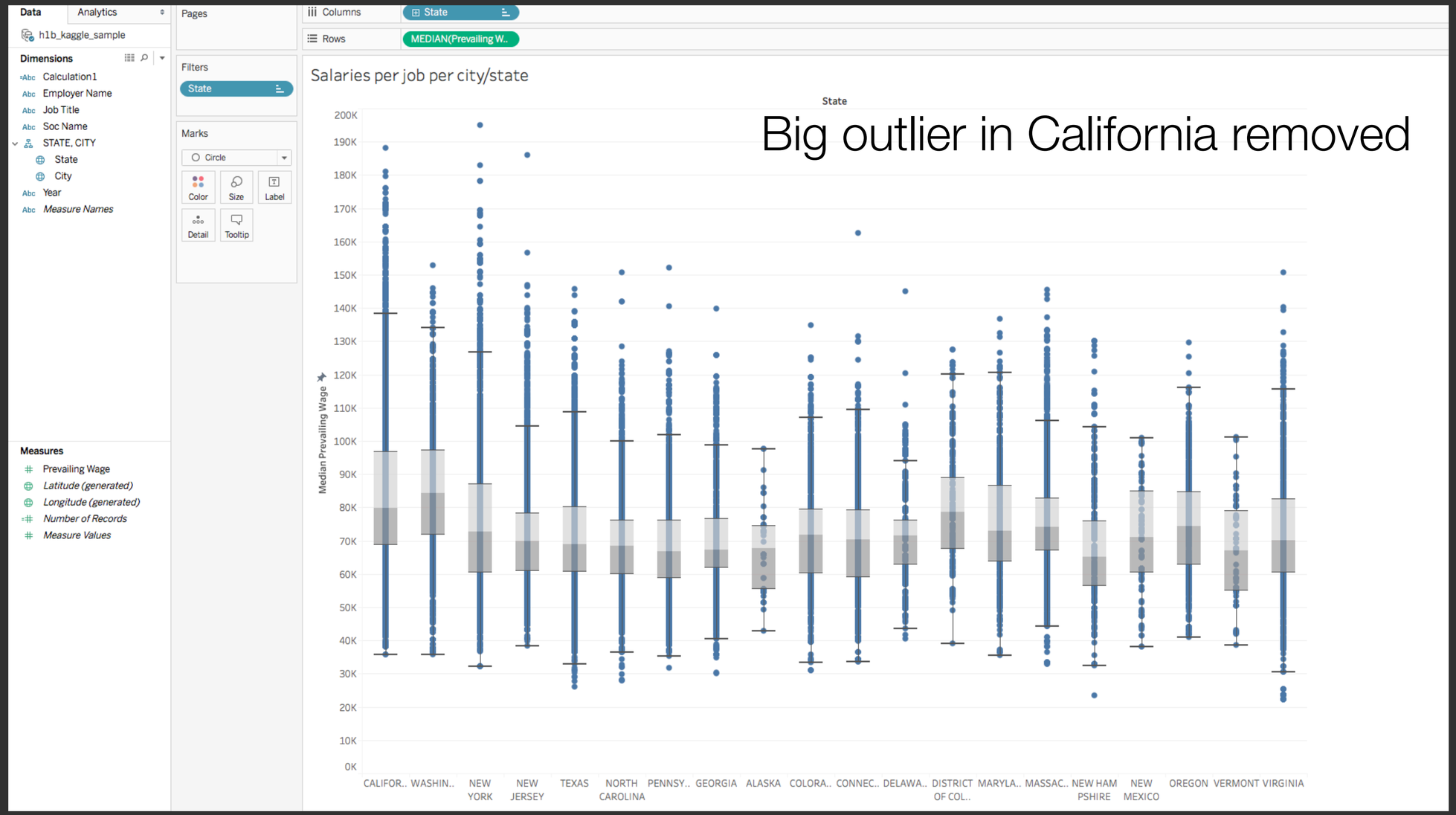


Filtered by top 50 jobs

Which states/cities files petitions the most?



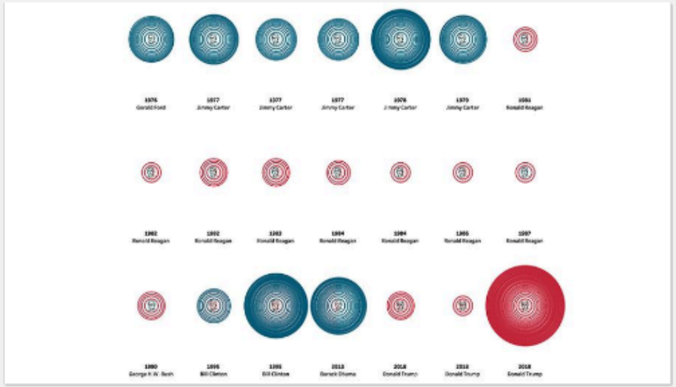
What are differences in salaries across states & cities?



What is the relationship between salaries and petitions?

Tableau Gallery

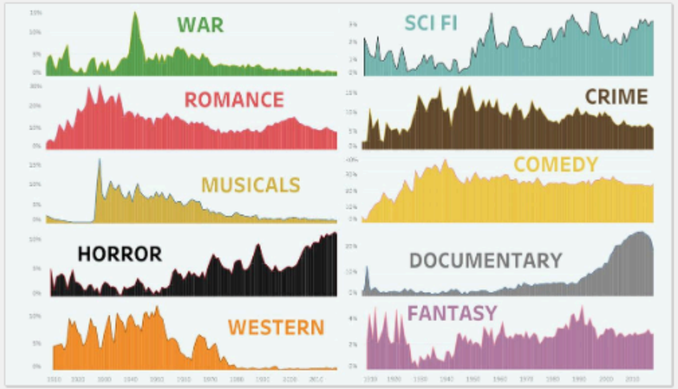
https://public.tableau.com/en-us/s/gallery



A History of US Government Shutdowns

In this visualization by Lilian Hoang, explore the issues, congressional makeup, and presidents behind all twenty-one US government shutdowns.

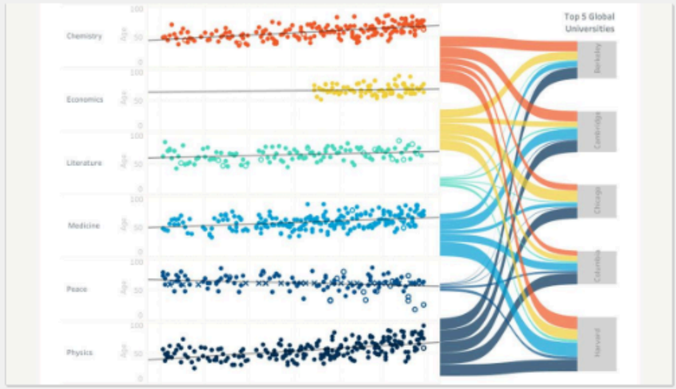
January 23, 2019



Film Genre Popularity since 1910

Explore the rising and falling popularity of film genres since 1910 in this visualization by Bo McCready. See what you can create using the same IMDB dataset!

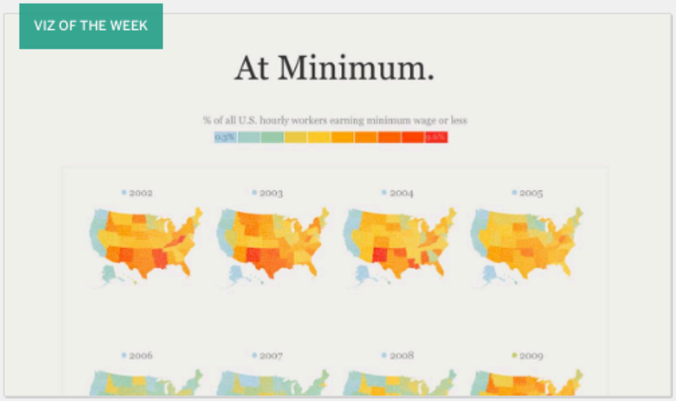
January 18, 2019



Nobel Prizes and Laureates

Ratnesh Pandey uses data from the Nobel Foundation to break down who has been awarded the Nobel Prize.

January 22, 2019



US Workers At or Below Minimum Wage

Using data from the US Bureau of Labor Statistics, Justin Davis visualizes the percentage of all US hourly workers earning minimum wage or less.

January 17, 2019

Next

Storytelling with Data

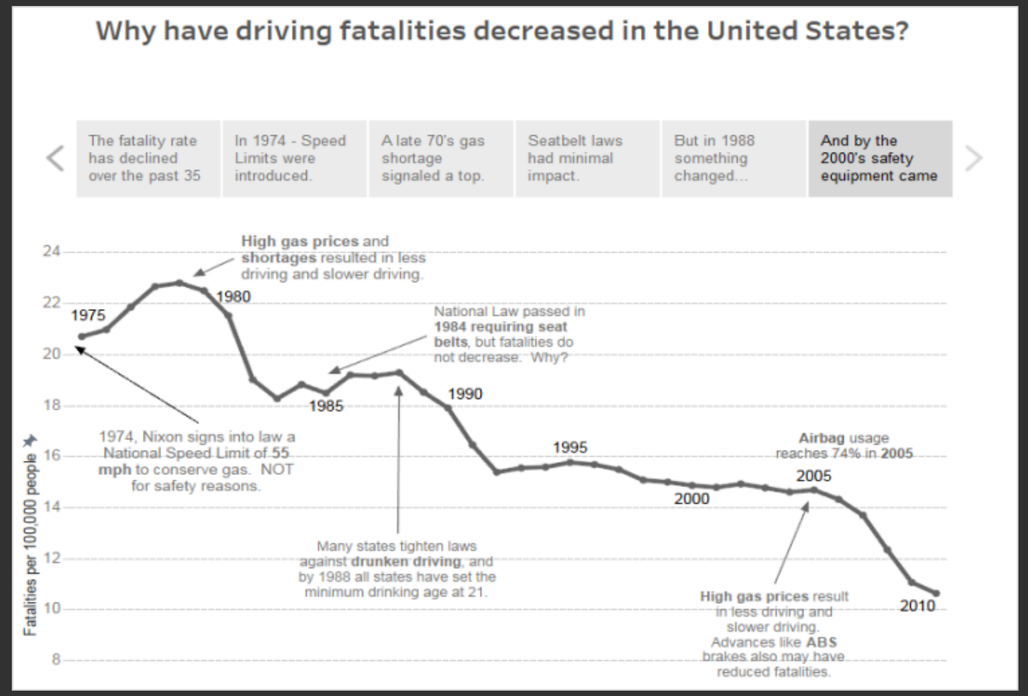


Tableau Story Points

10 min break