Data Analytics for Cities:

From Analysis to Action

Richard Dunks

Datapolitan
Data Solutions for the Modern Metropolis

Follow along at:
http://www.datapolitan.com/CenterForGovernmentExcellence

Introductions

- Name
- Hometown (or place you call home)
- One thing you hope to get out of class today

Goals

- Discuss the data-driven decision making process as it relates to city government
- Provide hands-on experience using Excel to clean and summarize data, including useful tips and tricks to working with city data
- Introduce advanced functionality within Excel as it relates to analyzing city data
- Discuss best practices when analyzing and visualizing data in Excel
- Review basic statistical concepts and how to perform statistical analysis in Excel

Outcomes

- You will better understand how to leverage Excel in the analytics process
- You will be more proficient using Excel for cleaning, analyzing, and visualizing data
- You will be familiar with Excel functions and other advanced features of Excel for analyzing data
- You will be familiar with fundamental best practices for visualizing data in Excel
- You will understand the key challenges city employees face in using Excel for analysis

Challenge to you

- There's a range of skills in this room
- If you don't know the material, learn the techniques but focus on the intention
- If you know the material, learn the intention but focus on how it's best communicated to others
- Let's all learn from each other as we go through this material

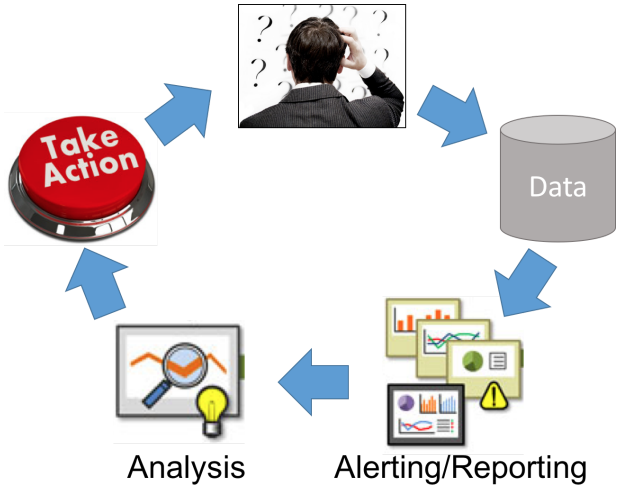
Why Do We Collect Data?

- Accountability
- Transparency
- "If you can't measure it, you can't manage it"
- "..and you can't fix it"

What good are laws if we don't know how they're implemented?

Which is what data tells us

Analytics Value Chain



Benefits of Excel

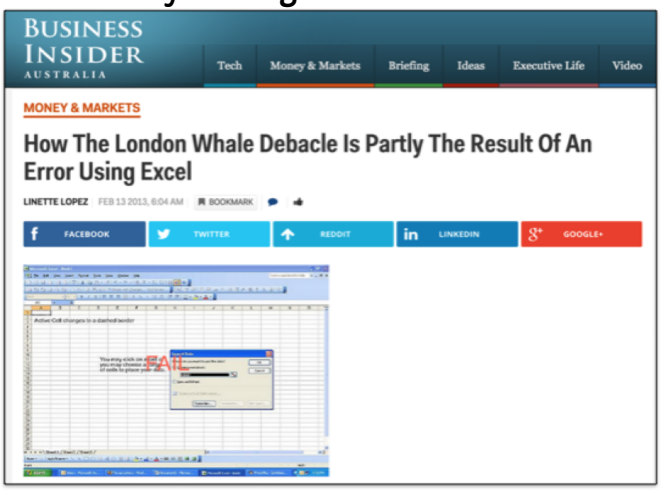
- Easy to use
- Very visual
- Lots of features and functions
- Easy to make charts
- Does a lot of formatting for you

Drawbacks of Excel

- Not very intuitive
- Hard to find what you're looking for (and they keep moving things around)
- Lots of features and functions
- Easy to make (bad) charts
- Does a lot of formatting for you

Cautionary Tale - London Whale

- $6.2 billion lost by JP Morgan Chase & Co



http://www.businessinsider.com.au/excel-partly-to-blame-for-trading-loss-2013-2

Excel Problems

- Manual data errors
- Manual copy and paste
- Simple formula error that hid volatility

Fined over $1 billion for poor internal oversight of trading activities

ata Analytics for Cities by Richard Dunks is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

To Get Started

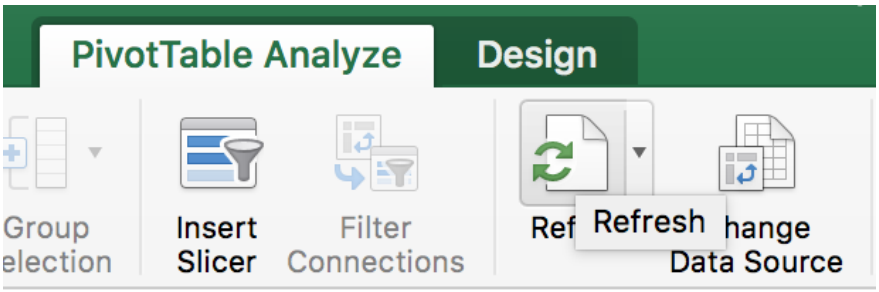
- Autosize all columns
- How many rows do we have?
- What values do we have in the fields?
- Filter by certain values

What Hour of the Day Does 311 Receive the Most Requests?

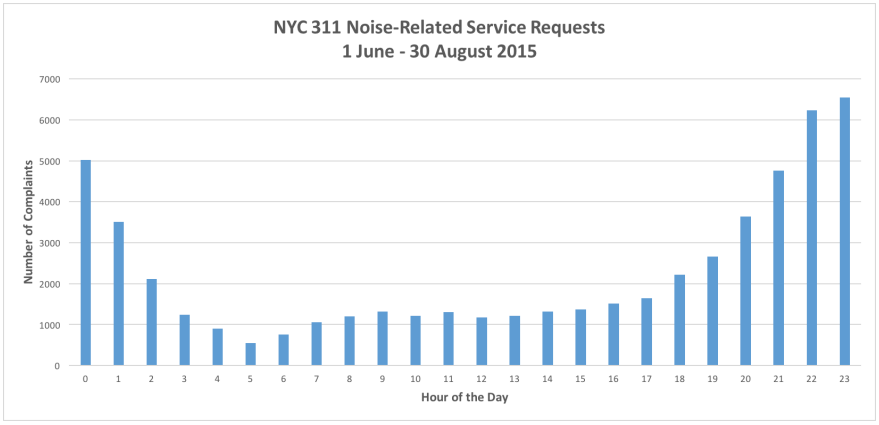
A	B	C
Unique Key	Created Date	hour
31423138	8/30/15 23:59	=HOUR(B2)
31423341	8/30/15 23:56	
31425003	8/30/15 23:56	

HOUR(serial_number)

What Hour of the Day Does 311 Receive the Most Requests?



What Hour of the Day Does 311 Receive the Most Requests?

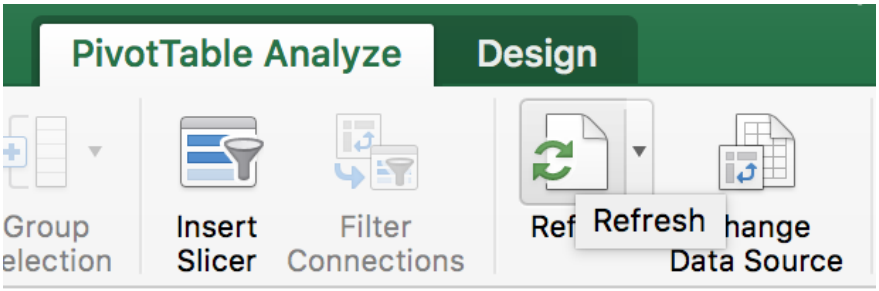


What Day of the Week Does 311 Receive the Most Requests?

A	B	C	D	E
Unique Key	Created Date	hour	dow	Closed D
31423138	8/30/15 23:59	23	=WEEKDAY(B2)	
31423341	8/30/15 23:56	23		
31425003	8/30/15 23:56	23		8/31/

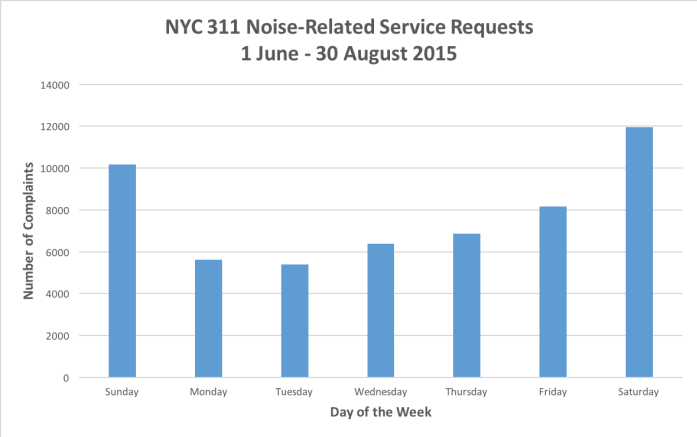
WEEKDAY(serial_number, [return_type])

What Day of the Week Does 311 Receive the Most Requests?



ata Analytics for Cities by Richard Dunks is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

What Day of the Week Does 311 Receive the Most Requests?



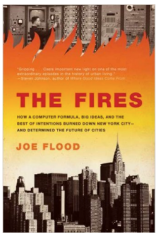
Exploratory Data Analysis

- Goal -> Discover patterns in the data
- Understand the context
- Summarize fields
- Use graphical representations of the data
- Explore outliers

Tukey, J.W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

ata Analytics for Cities by Richard Dunks is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Telling a Story with Data



Preparing the data

COMMUNITY DISTRICT NUMBER	COMMUNITY DISTRICT NAME	Total Population 1970	Total Population 1980
BRONX COMMUNITY DISTRICTS			
1	Melrose, Mott Haven, Port Morris	138,557	78,441
2	Hunts Point, Longwood	99,493	34,399
3	Morrisania, Crotona Park East	150,636	53,635
4	Highbridge, Concourse Village	144,207	114,312
5	University Hts., Fordham, Mt. Hope	121,807	107,995
6	East Tremont, Belmont	114,137	65,016
7	Bedford Park, Norwood, Fordham	113,764	116,827
8	Riverdale, Kingsbridge, Marble Hill	103,543	98,275
9	Soundview, Parkchester	166,442	167,627
10	Throgs Nk., Co-op City, Pelham Bay	84,948	106,516
11	Pelham Pkwy, Morris Park, Laconia	105,980	99,080
12	Wakefield, Williamsbridge	135,010	128,226
	Total Borough Population	=SUM(C3:C14)	
	Population Change		
	Percentage Change		

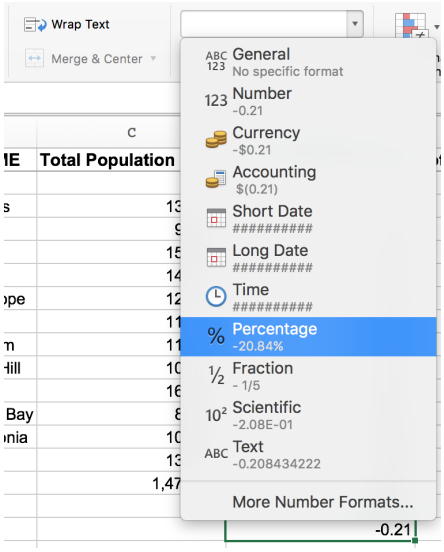
Preparing the data

COMMUNITY DISTRICT NUMBER	COMMUNITY DISTRICT NAME	Total Population 1970	Total Population 1980
BRONX COMMUNITY DISTRICTS			
1	Melrose, Mott Haven, Port Morris	138,557	78,441
2	Hunts Point, Longwood	99,493	34,399
3	Morrisania, Crotona Park East	150,636	53,635
4	Highbridge, Concourse Village	144,207	114,312
5	University Hts., Fordham, Mt. Hope	121,807	107,995
6	East Tremont, Belmont	114,137	65,016
7	Bedford Park, Norwood, Fordham	113,764	116,827
8	Riverdale, Kingsbridge, Marble Hill	103,543	98,275
9	Soundview, Parkchester	166,442	167,627
10	Throgs Nk., Co-op City, Pelham Bay	84,948	106,516
11	Pelham Pkwy, Morris Park, Laconia	105,980	99,080
12	Wakefield, Williamsbridge	135,010	128,226
	Total Borough Population	1,478,524	1,170,349
	Population Change		=D15-C15
	Percentage Change		

Preparing the data

COMMUNITY DISTRICT NUMBER	COMMUNITY DISTRICT NAME	Total Population 1970	Total Population 1980
BRONX COMMUNITY DISTRICTS			
1	Melrose, Mott Haven, Port Morris	138,557	78,441
2	Hunts Point, Longwood	99,493	34,399
3	Morrisania, Crotona Park East	150,636	53,635
4	Highbridge, Concourse Village	144,207	114,312
5	University Hts., Fordham, Mt. Hope	121,807	107,995
6	East Tremont, Belmont	114,137	65,016
7	Bedford Park, Norwood, Fordham	113,764	116,827
8	Riverdale, Kingsbridge, Marble Hill	103,543	98,275
9	Soundview, Parkchester	166,442	167,627
10	Throgs Nk., Co-op City, Pelham Bay	84,948	106,516
11	Pelham Pkwy, Morris Park, Laconia	105,980	99,080
12	Wakefield, Williamsbridge	135,010	128,226
	Total Borough Population	1,478,524	1,170,349
	Population Change		-308,175
	Percentage Change		=D16/C15

Preparing the data



Why do we visualize data?

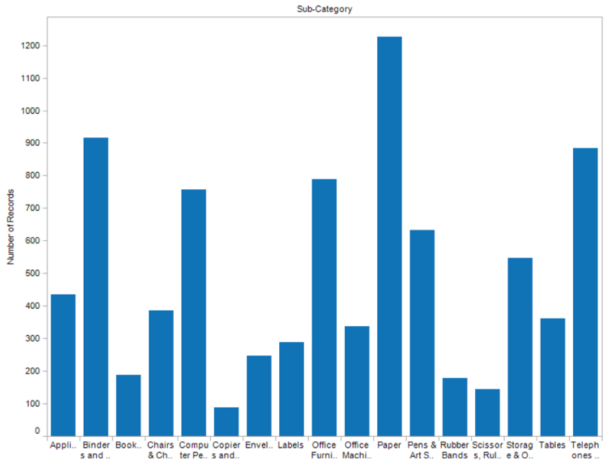
- Explore
- Explain
- Persuade

ata Analytics for Cities by Richard Dunks is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

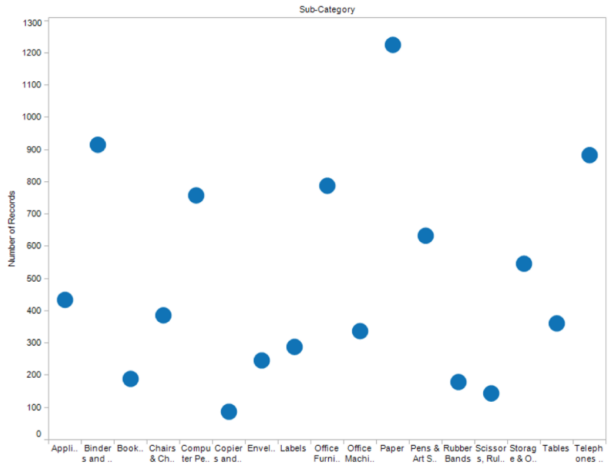
Encoding

- Visual encoding is the process of encoding information visually
- Every visualization is a mapping between data objects and features to visual objects and features

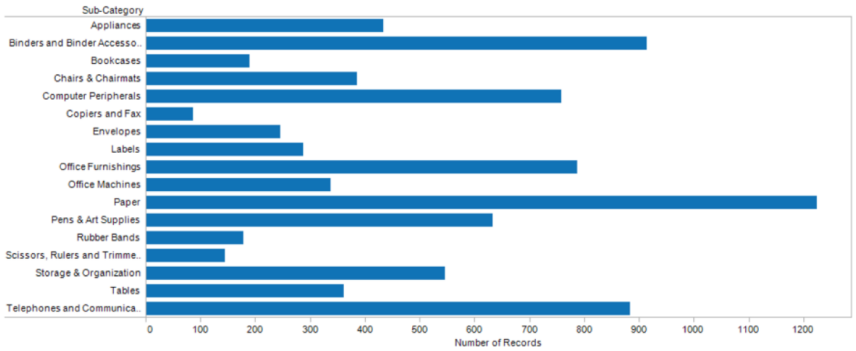
Bar Chart



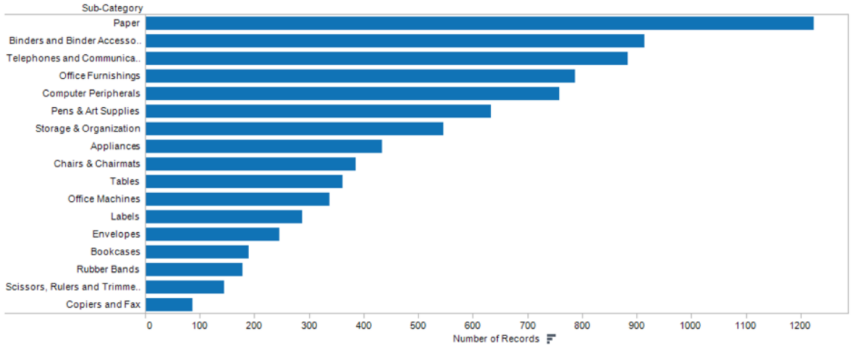
"Barless" Bar Chart



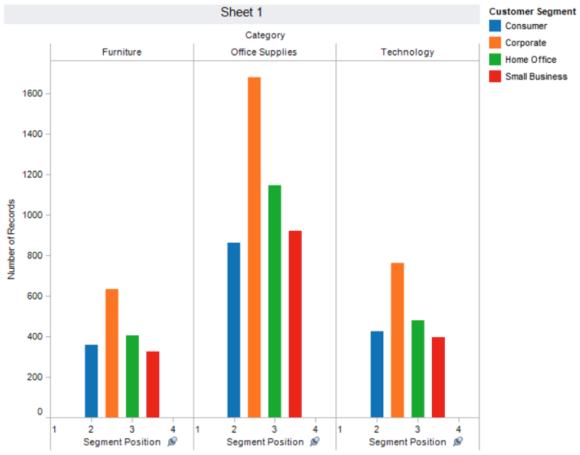
Horizontal Bar Chart



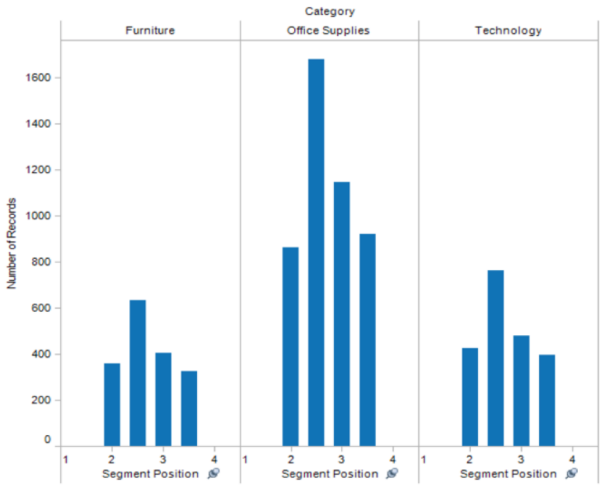
Ranked Horizontal Bar Chart



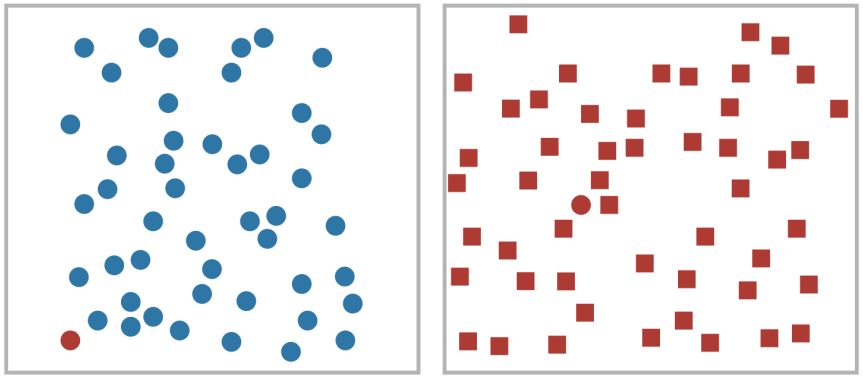
Grouped Bar Chart



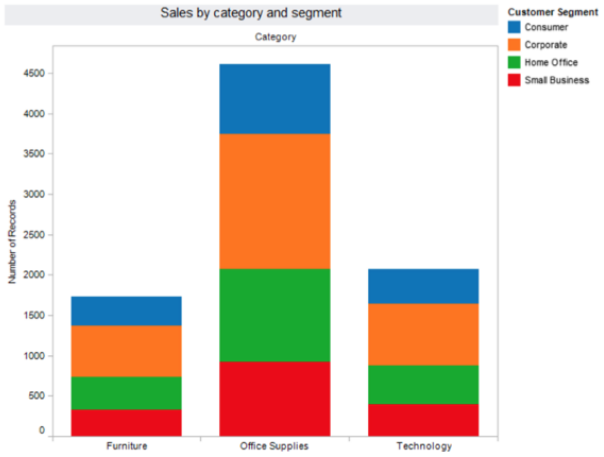
Grouped Bar Chart



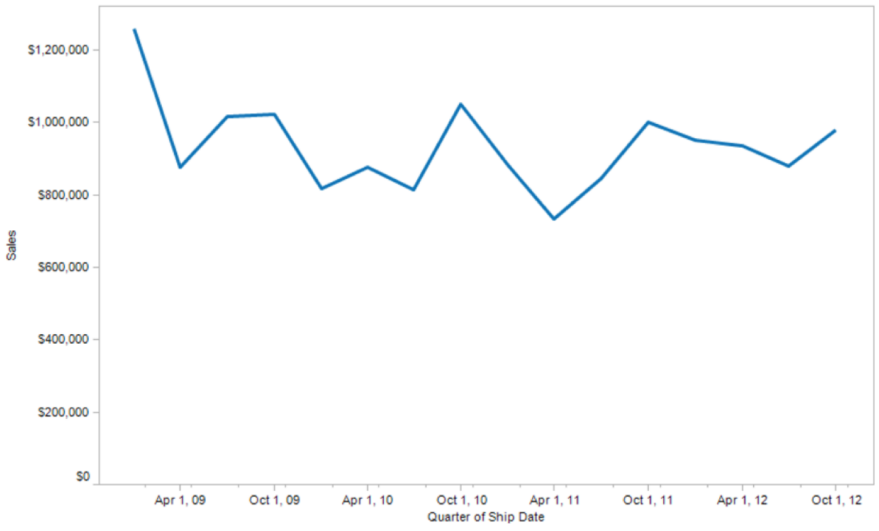
Precognitive Processing



Stacked Bar Chart

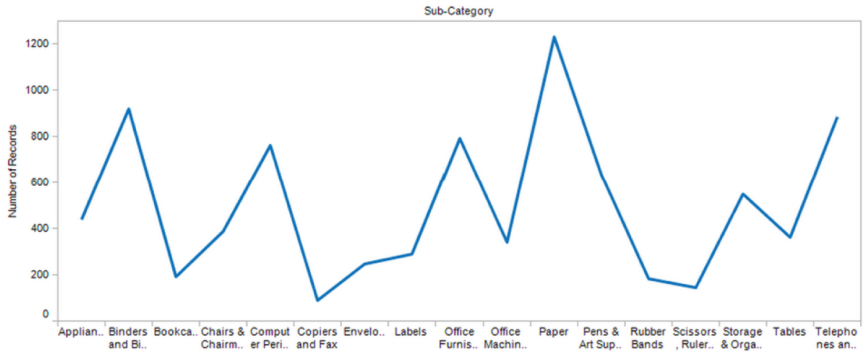


Line Chart



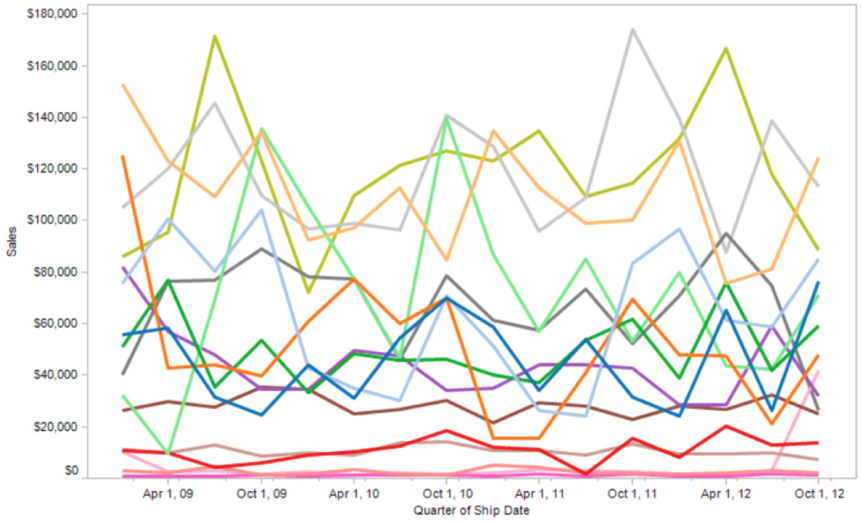
Why use a line chart?

Line Chart with Categorical Data

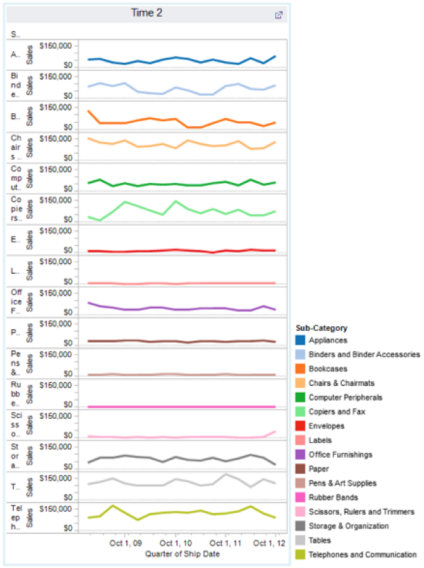


Line Chart

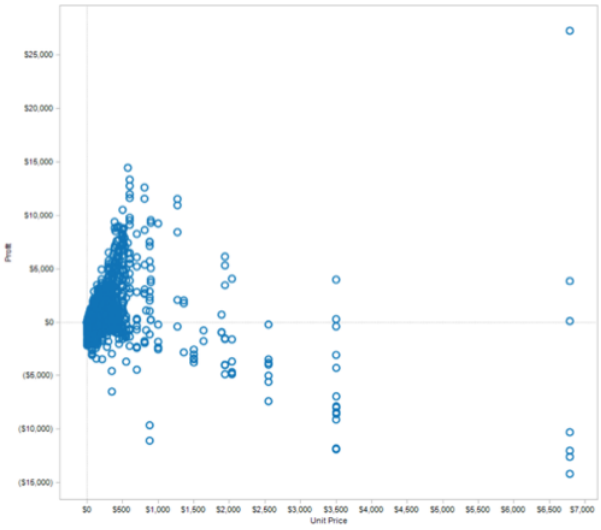
Line Chart



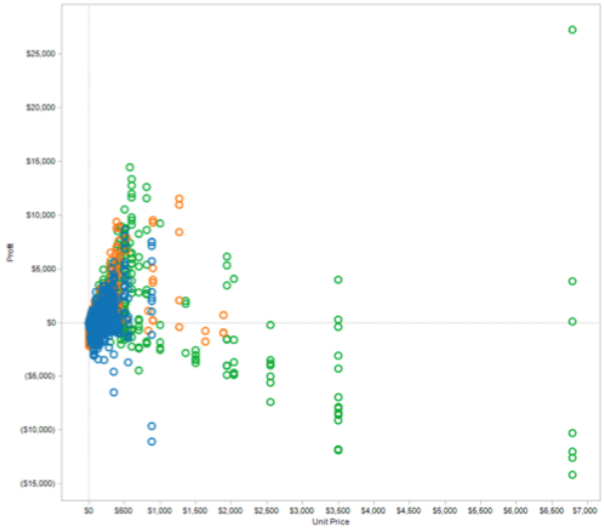
Small Multiples



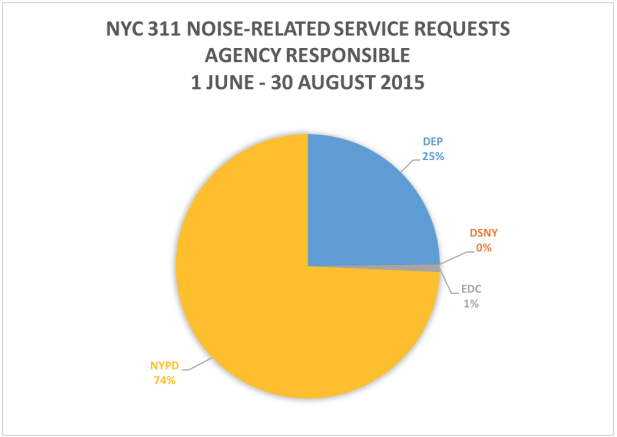
Scatter Plot



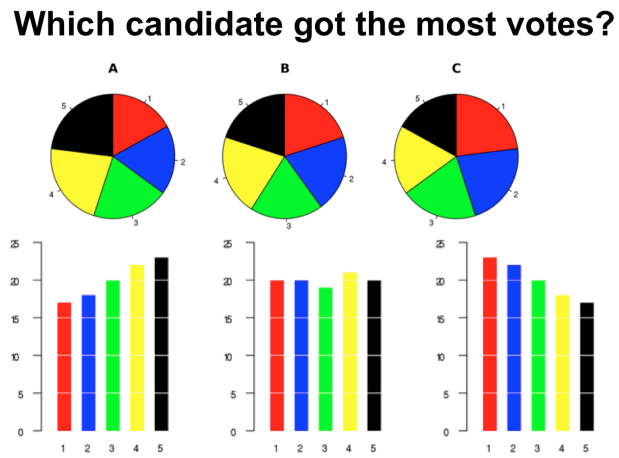
Scatter Plot



Pie Charts

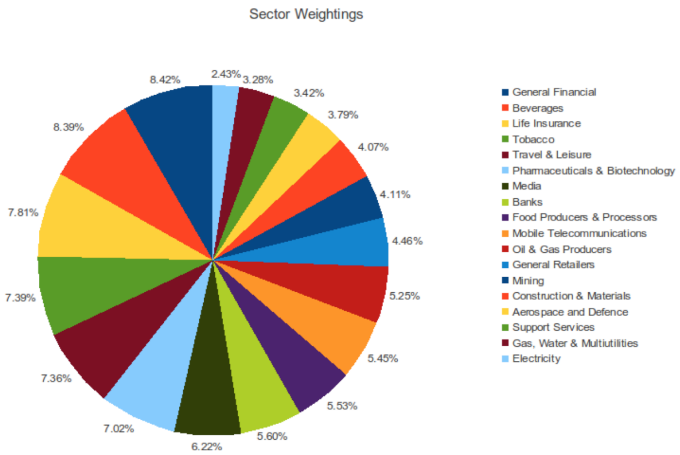


Pie Charts



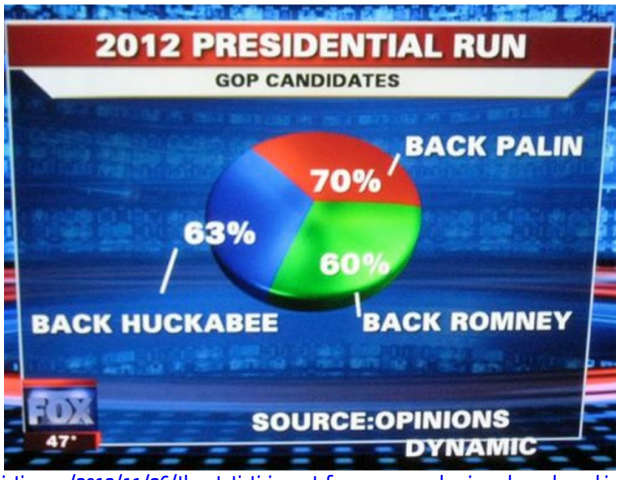
"Piecharts". Licensed under CC BY 1.0 via Wikimedia Commons

Pie Charts



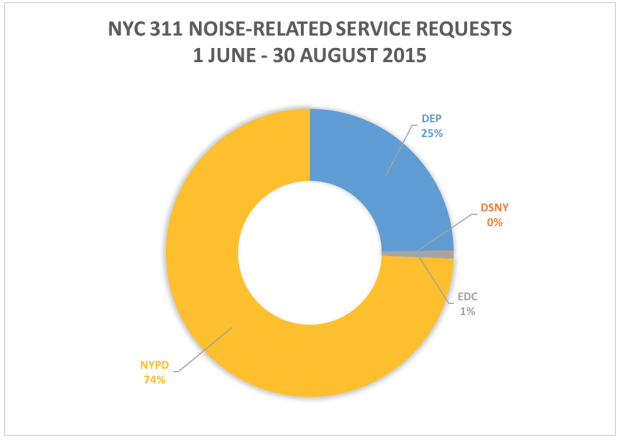
"Badpie" by Gilgongo - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons

Pie Charts

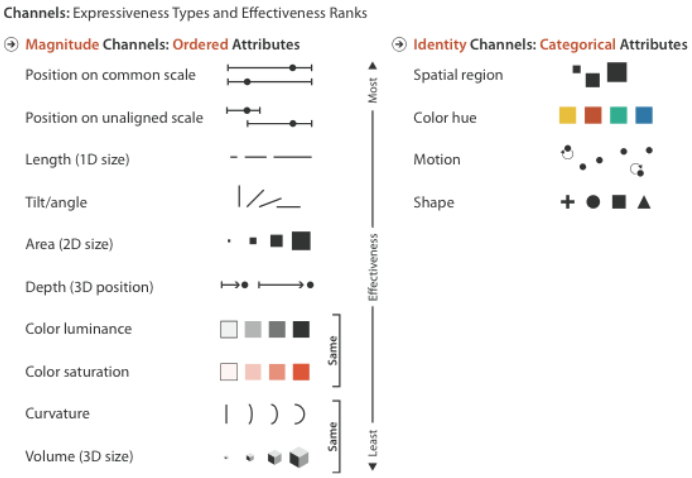


http://simplystatistics.org/2012/11/26/the-statisticians-at-fox-news-use-classic-and-novel-graphical-techniques-to-lead-with-data/

Donut Charts



Channels and Marks

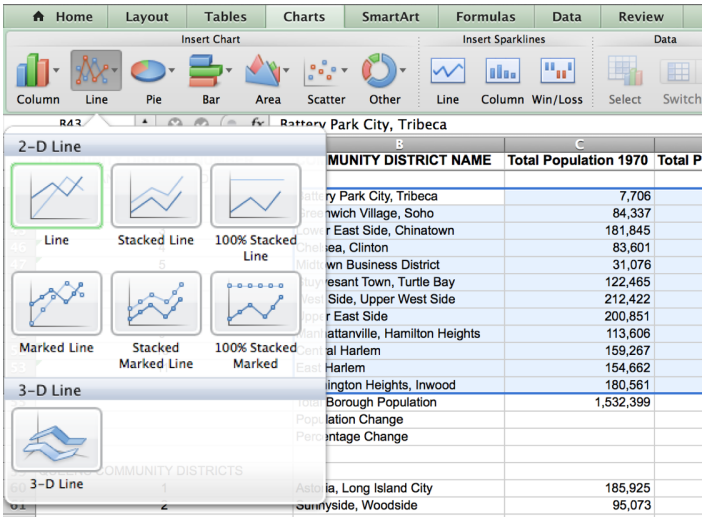


From Tamara Munzner, Visualization Analysis and Design

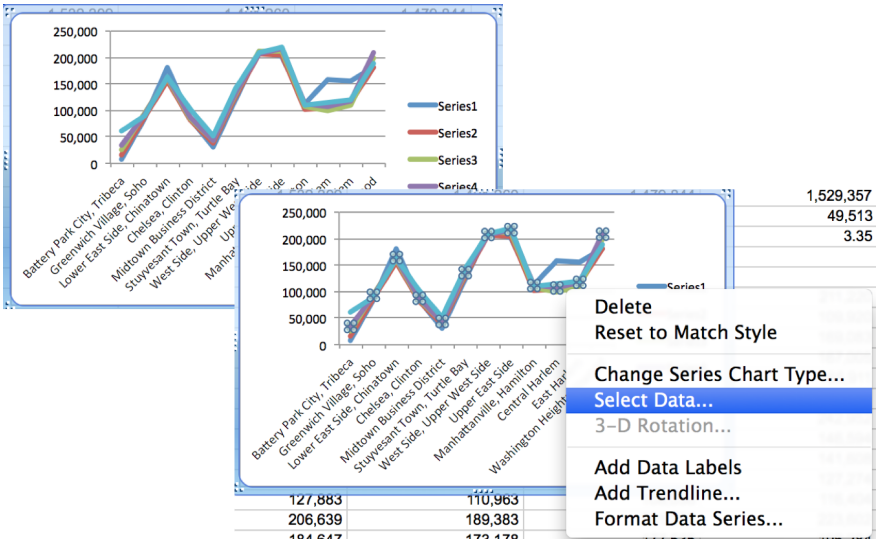
Create graph of CD population

COMMUNITY DISTRICT NAME	Total Population 1970	Total Population 1980	Total Population 1990	Total Population 2000	Total Population 2010
Battery Park City, Tribeca	7,706	15,918	25,366	34,420	60,978
Greenwich Village, Soho	84,337	87,069	94,105	93,119	90,016
Lower East Side, Chinatown	181,845	154,848	161,617	164,407	163,277
Chelsea, Clinton	83,601	82,164	84,431	87,479	103,245
Midtown Business District	31,076	39,544	43,507	44,028	51,673
Stuyvesant Town, Turtle Bay	122,465	127,554	133,748	136,152	142,745
West Side, Upper West Side	212,422	206,669	210,993	207,699	209,084
Upper East Side	200,851	204,305	210,880	217,063	219,920
Manhattanville, Hamilton Heights	113,606	103,038	106,978	111,724	110,193
Central Harlem	159,267	105,641	99,519	107,109	115,723
East Harlem	154,662	114,569	110,508	117,743	120,511
Washington Heights, Inwood	180,561	179,941	198,192	208,414	190,020
Total Borough Population	1,532,399	1,421,260	1,479,844	1,529,357	1,577,385
Population Change		-111,139	58,584	49,513	48,028
Percentage Change		-7.25	4.12	3.35	3.14

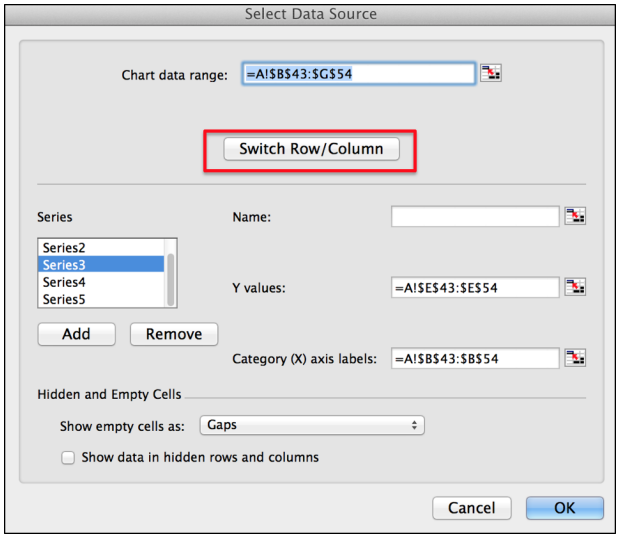
Create graph of CD population



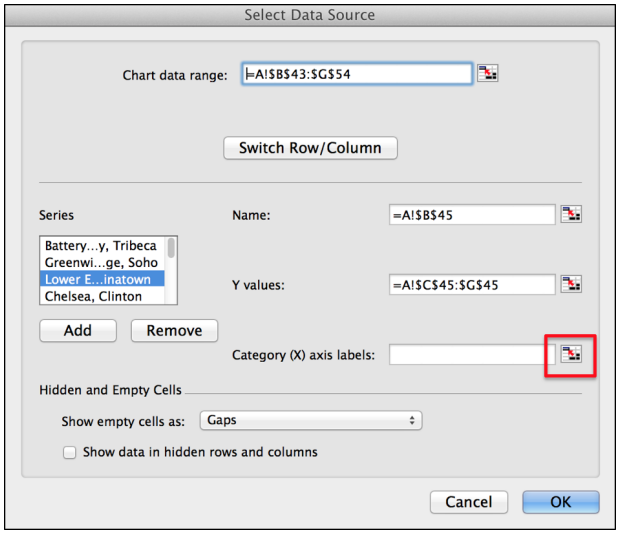
Create graph of CD population



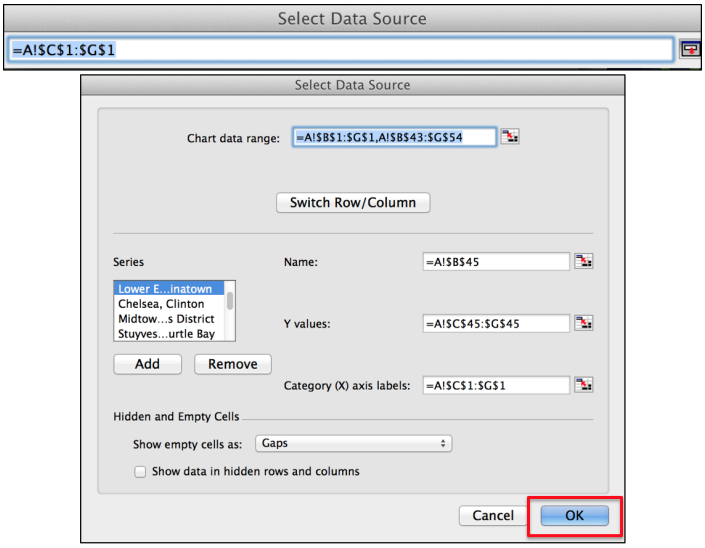
Create graph of CD population



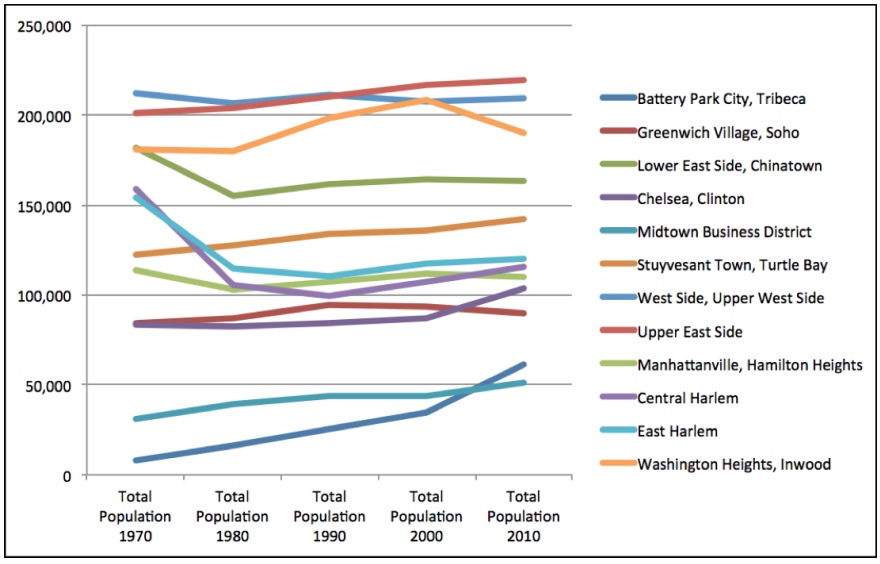
Create graph of CD population



Create graph of CD population



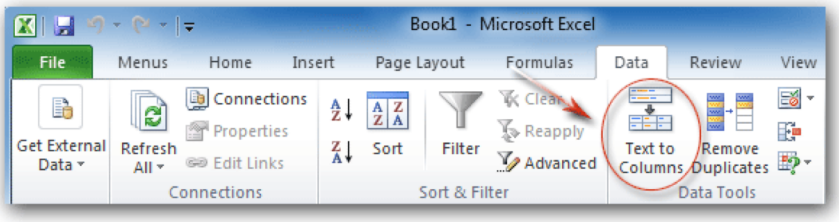
Create graph of CD population



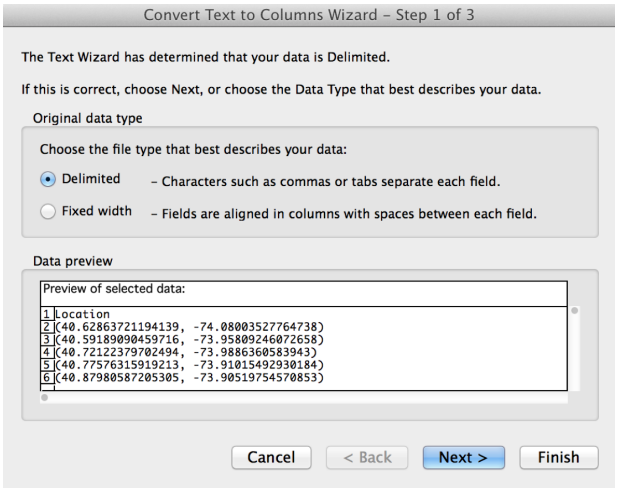
Splitting Data

	AX	AY	AZ
1	Latitude	Longitude	Location
2	40.62863721	-74.08003528	(40.62863721194139, -74.08003527764738)
3	40.5918909	-73.95809246	(40.59189090459716, -73.95809246072658)
4	40.7212238	-73.98863606	(40.72122379702494, -73.9886360583943)
5	40.77576316	-73.91015493	(40.77576315919213, -73.91015492930184)
6	40.87980587	-73.90519755	(40.87980587205305, -73.90519754570853)
7	40.57480362	-73.97352786	(40.57480361606634, -73.9735278635962)
8	40.68321658	-73.95387066	(40.68321657803109, -73.9538706637591)
9	40.6818709	-73.7660623	(40.68187089534303, -73.7660623046722)
10	40.70605964	-73.83143704	(40.70605964424251, -73.83143703998772)
11	40.74786736	-73.81847221	(40.74786736221416, -73.81847220968915)
12	40.64795817	-74.00018739	(40.64795817286927, -74.00018738922621)
13	40.6875289	-73.9717271	(40.68752889748776, -73.97172709811349)
14	40.86939583	-73.91661612	(40.86939583069894, -73.91661611865678)
15	40.71346027	-73.95874828	(40.7134602748696, -73.95874828450539)
16	40.86954934	-73.91634837	(40.86954933988228, -73.9163483681302)

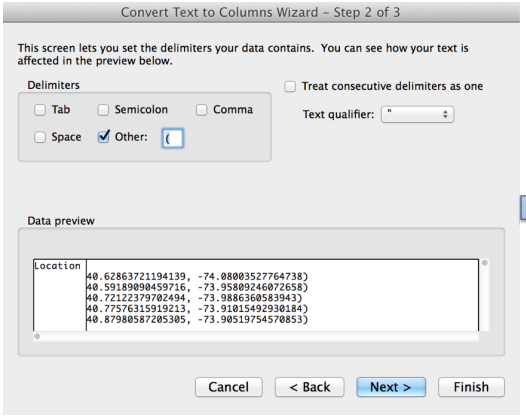
Splitting Data



Splitting Data

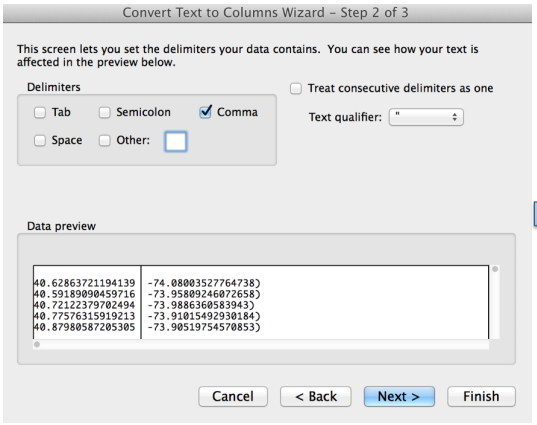


Splitting Data



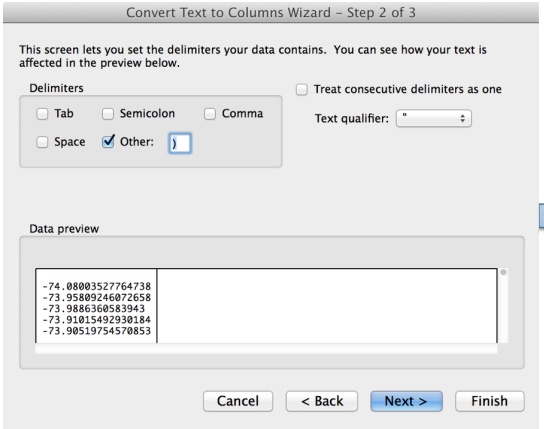
AZ	BA
Location	
	40.62863721194139, -74.08003527764738)
	40.59189090459716, -73.95809246072658)
	40.72122379702494, -73.9886360583943)
	40.77576315919213, -73.91015492930184)
	40.87980587205305, -73.90519754570853)
	40.57480361606634, -73.9735278635962)
	40.68321657803109, -73.9538706637591)
	40.68187089534303, -73.7660623046722)
	40.70605964424251, -73.83143703998772)
	40.74786736221416, -73.81847220968915)
	40.64795817286927, -74.00018738922621)
	40.68752889748776, -73.97172709811349)
	40.86939583069894, -73.91661611865678)
	40.7134602748696, -73.95874828450539)
	40.86954933988228, -73.9163483681302)

Splitting Data



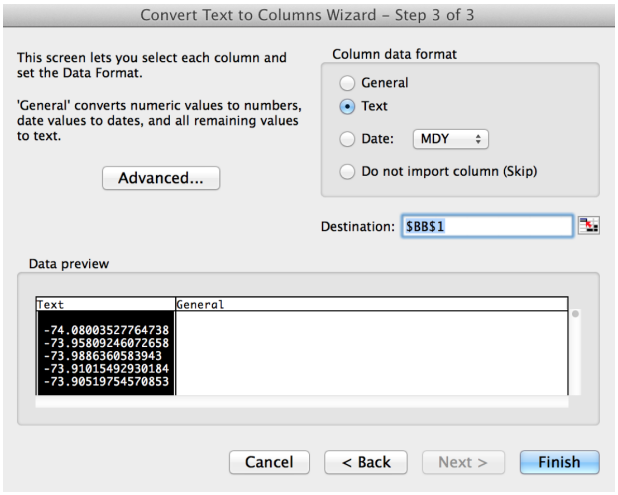
AZ	BA	BB
Location		
	40.62863721	-74.08003527764738)
	40.5918909	-73.95809246072658)
	40.7212238	-73.9886360583943)
	40.77576316	-73.91015492930184)
	40.87980587	-73.90519754570853)
	40.57480362	-73.9735278635962)
	40.68321658	-73.9538706637591)
	40.6818709	-73.7660623046722)
	40.70605964	-73.83143703998772)
	40.74786736	-73.81847220968915)
	40.64795817	-74.00018738922621)
	40.6875289	-73.97172709811349)
	40.86939583	-73.91661611865678)
	40.71346027	-73.95874828450539)
	40.86954934	-73.9163483681302)

Splitting Data



AZ	BA	BB
Location		
	40.62863721	-74.08003527764738
	40.5918909	-73.95809246072658
	40.7212238	-73.9886360583943
	40.77576316	-73.91015492930184
	40.87980587	-73.90519754570853
	40.57480362	-73.9735278635962
	40.68321658	-73.9538706637591
	40.6818709	-73.7660623046722
	40.70605964	-73.83143703998772
	40.74786736	-73.81847220968915
	40.64795817	-74.00018738922621
	40.6875289	-73.97172709811349
	40.86939583	-73.91661611865678
	40.71346027	-73.95874828450539
	40.86954934	-73.9163483681302

Splitting Data



Splitting Data with MID

MID, MIDB functions

This article describes the formula syntax and usage of the **MID** and **MIDB**function in Microsoft Excel.

Description

MID returns a specific number of characters from a text string, starting at the position you specify, based on the number of characters you specify.

MIDB returns a specific number of characters from a text string, starting at the position you specify, based on the number of bytes you specify.

IMPORTANT MID is intended for use with languages that use the single-byte character set (SBCS), whereas MIDB is intended for use with languages that use the double-byte character set (DBCS). The default language setting on your computer affects the return value in the following way:

- MID always counts each character, whether single-byte or double-byte, as 1, no matter what the default language setting is.
- MIDB counts each double-byte character as 2 when you have enabled the editing of a language that supports DBCS and then set it as the default language. Otherwise, MIDB counts each character as 1.

The languages that support DBCS include Japanese, Chinese (Simplified), Chinese (Traditional), and Korean.

Splitting Data with MID

Z	AA
Community Board	CB_Number
01 STATEN ISLAND	=MID(Z2,1,2)
15 BROOKLYN	
03 MANHATTAN	
01 QUEENS	
08 BRONX	
13 BROOKLYN	

Splitting Data with MID

Z	AA
Community Board	CB_Number
01 STATEN ISLAND	01
15 BROOKLYN	
03 MANHATTAN	
01 QUEENS	
08 BRONX	
13 BROOKLYN	

Other useful functions LEFT and RIGHT

LEFT, LEFTB functions

This article describes the formula syntax and usage of the **LEFT** and **LEFTB** function in Microsoft Excel.

Description

LEFT returns the first character or characters in a text string, based on the number of characters you specify.

RIGHT, RIGHTB functions

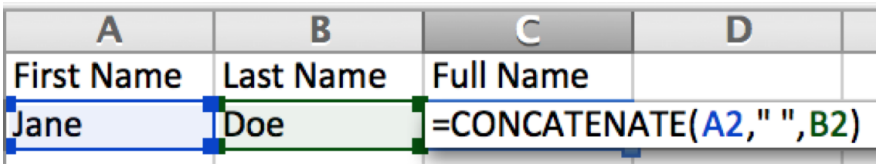
This article describes the formula syntax and usage of the **RIGHT** and **RIGHTB** functions in Microsoft Excel.

Description

RIGHT returns the last character or characters in a text string, based on the number of characters you specify.

Combining data with CONCATENATE

- Useful for combining text fields
- Basic Syntax "=CONCATENATE(text1, [text2], ...)"



A	B	C	D
First Name	Last Name	Full Name	
Jane	Doe	Jane Doe	

Calculate the number of noise-related service requests per 1000 people in Community Board

1. Complaints per Community Board using PivotTable

A	B	C
Count of Unique Key		
Row Labels	**Total**	CB Population
0 Unspecified	60	
01 BRONX	56	
01 BROOKLYN	1023	
01 MANHATTAN	567	
01 QUEENS	666	
01 STATEN ISLAND	242	

2. Create matching identifier in population sheet

	A	B
1	**COMMUNITY DISTRICT NUMBER**	
2	BRONX COMMUNITY DISTRICTS	
3	1	=CONCATENATE("0",A3," BRONX")
4	2	
5	3	
6	4	
7	5	

	A	B
1	**COMMUNITY DISTRICT NUMBER**	
2	BRONX COMMUNITY DISTRICTS	
3	1	01 BRONX
4	2	
5	3	
6	4	
7	5	

2. Create matching identifier in population sheet

1	**COMMUNITY DISTRICT NUMBER**	
2	BRONX COMMUNITY DISTRICTS	
3	1	01 BRONX
4	2	02 BRONX
5	3	03 BRONX
6	4	04 BRONX
7	5	05 BRONX
8	6	06 BRONX
9	7	07 BRONX
10	8	08 BRONX
11	9	09 BRONX
12	10	010 BRONX
13	11	011 BRONX
14	12	012 BRONX

```
=IF(
    LEN( A20 ) < 2,
        CONCATENATE( "0", A20, " BROOKLYN" ),
            CONCATENATE( A20, " BROOKLYN" )
    )
```

3. Insert VLOOKUP formula

```
=VLOOKUP(A6,
    '[NYC_Population_1970-2010.xlsx]A'!$C$1:$I$81,
        7,
        FALSE)
```


3. Insert VLOOKUP formula

```
=VLOOKUP(A6,
    '[NYC_Population_1970-2010.xlsx]A'!$C$1:$I$81,
        7,
        FALSE)
```

Count of Unique Key Row Labels	Total	CB Population
0 Unspecified	60	
01 BRONX	6	91497
01 BROOKLYN	[illegible]	
01 MANHATTAN	567	
01 QUEENS	666	
01 STATEN ISLAND	242	

ata Analytics for Cities by Richard Dunks is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

4. Calculate complaints per 1000 residents

Count of Unique Key Row Labels	Total	CB Population	Complaints_per_1000
0 Unspecified	60		
01 BRONX	56	91497	=B6/(C6/1000)
01 BROOKLYN	1023	173083	
01 MANHATTAN	567	60978	
01 QUEENS	666	191105	
01 STATEN ISLAND	242	175756	

Count of Unique Key Row Labels	Total	CB Population	Complaints_per_1000
0 Unspecified	60		
01 BRONX	56	91497	0.612041925
01 BROOKLYN	1023	173083	
01 MANHATTAN	567	60978	
01 QUEENS	666	191105	
01 STATEN ISLAND	242	175756	

5. Revel in the data

Count of Unique Key Row Labels	Total	CB Population	Complaints_per_1000
0 Unspecified	60		
01 BRONX	56	91497	0.612041925
01 BROOKLYN	1023	173083	5.910459144
01 MANHATTAN	567	60978	9.298435501
01 QUEENS	666	191105	3.48499516
01 STATEN ISLAND	242	175756	1.376908896
02 BRONX	39	52246	0.746468629
02 BROOKLYN	513	99617	5.149723441
02 MANHATTAN	1139	90016	12.65330608
02 QUEENS	344	113200	3.038869258
02 STATEN ISLAND	68	132003	0.515139807
03 BRONX	36	79762	0.451342745
03 BROOKLYN	541	152985	3.536294408
03 MANHATTAN	1143	163277	7.000373598
03 QUEENS	240	171576	1.398797035
03 STATEN ISLAND	102	160209	0.636668352
04 BRONX	104	146441	0.710183623
04 BROOKLYN	325	112634	2.885451995
04 MANHATTAN	1096	103245	10.61552618
04 QUEENS	85	172598	0.492473841

What we've covered so far

- Value of data in government
- Exploratory data analysis
- Reviewed formulas in Excel
- Discussed visualization design
- Created charts in Excel
- Split and combined fields using Excel functions
- Joined data using VLOOKUP

Introduction to Statistics

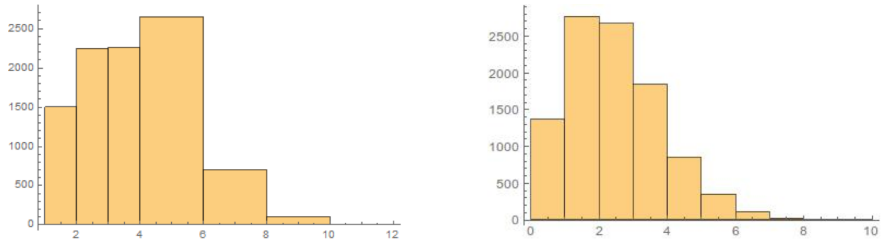
> We are drowning in information and starving for knowledge.
>
> Rutherford D. Roger

Why Statistics?

- Tools for extracting meaning from data
- Commonly understood ways of communicating meaning to others

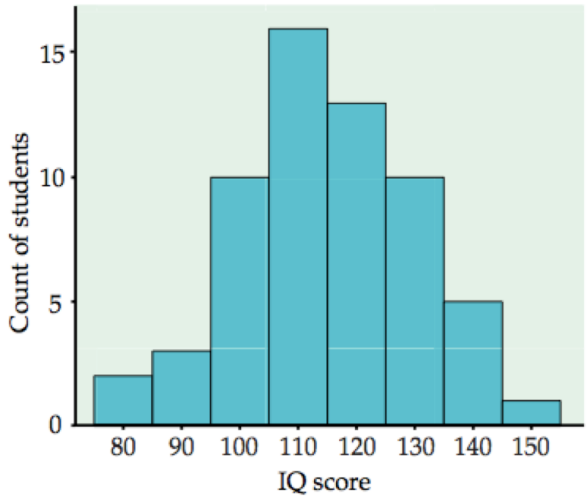
Histogram

- Charts the frequency of instances in the data
- Shows the frequency distribution
- Values are grouped into class intervals
- Best to have a consistent size to class intervals

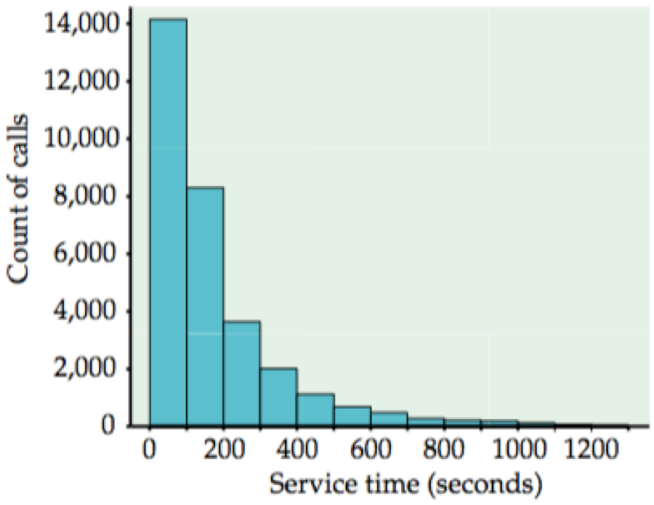


Source: http://mathematica.stackexchange.com/questions/59520/histogram-with-variable-bin-size

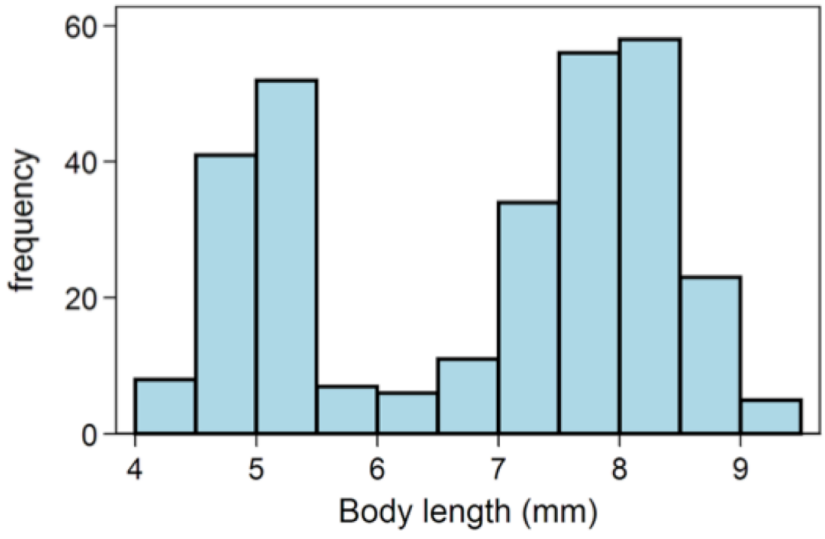
Normal Distribution



Long-tail Distribution



Bi-Modal Distribution

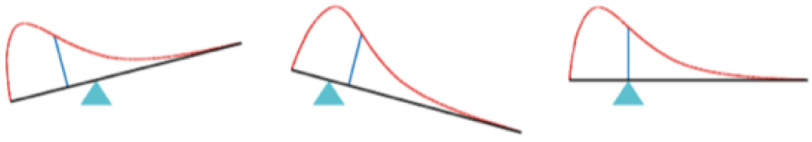


Mean

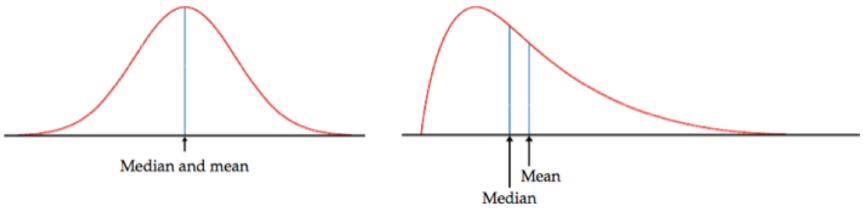
- A representative value for the data
- Usually what people mean by "average"
- Calculate by adding all the values together and dividing by the number instances
- Sensitive to extremes

Median

- The "middle" value of a data set
- Center value of a data set with an odd number of values
- Sum of two middle values divided by 2 if the number of items in a data set is even
- Resistant to extreme values

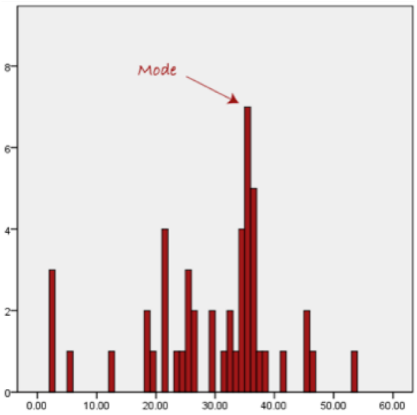


Median vs Mean



Mode

- The most frequent value in a dataset
- Often used for categorical data



Measures of Central Tendency

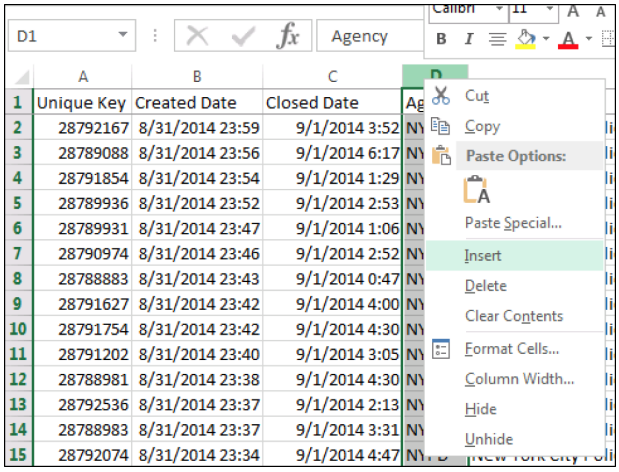
- Quantitative data tends to cluster around some central value
- Contrasts with the spread of data around that center (i.e. the variability in the data)
- Mean is a more precise measure and more often used
- Median is better when there are extreme outliers
- Mode is used when the data is categorical (as opposed to numeric)

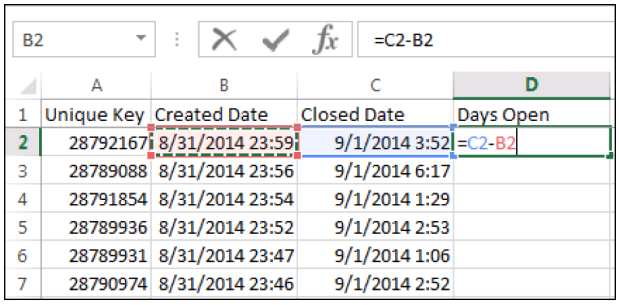
Let's do this in Excel

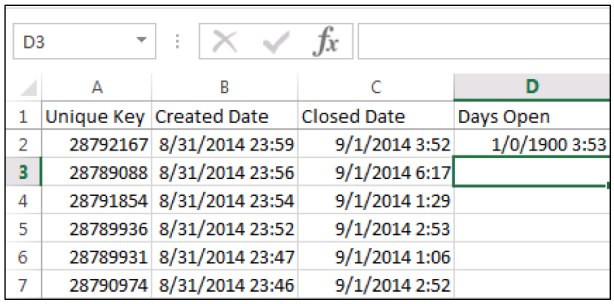
What is the mean time a 311 noise-related service request remains open? The median? The mode?

How does this look by agency?

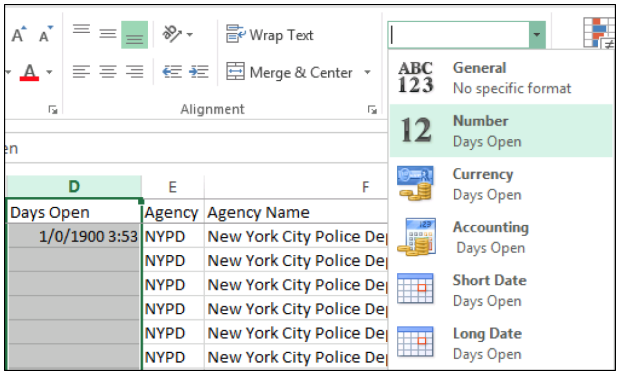
Preparing the Data



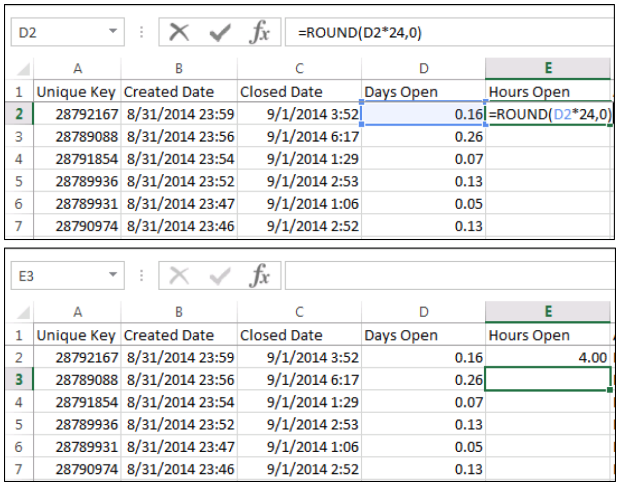




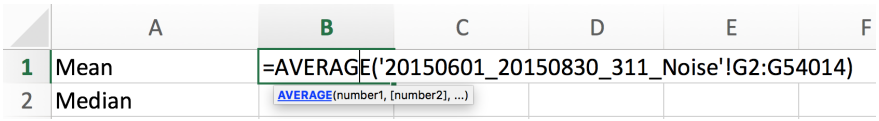
Preparing the Data



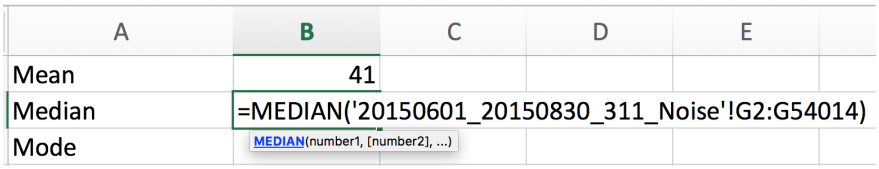
Preparing the Data



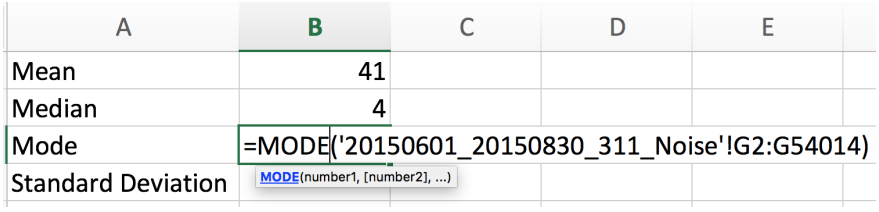
Calculating Mean



Calculating Median



Calculating Mode



Now do the same with 2014 noise-related complaints for the same period

How have the times changed? Better or worse? How much?

Range

- The gap between the minimum value and the maximum value
- Calculated by subtracting the minimum from the maximum

2	Median	4
3	Mode	1
4	Min	=MIN('20150601_20150830_311_Noise'!G2:G54014)
5	Max	

MIN(**number1**, [number2], ...)

3	Mode	1
4	Min	0
5	Max	=MAX('20150601_20150830_311_Noise'!G2:G54014)
6	Range	

MAX(**number1**, [number2], ...)

4	Min	0
5	Max	2500
6	Range	=B5-B4

Quartiles

- Median splits the data set into two equal groups
- Quartiles split the data into four equal groups
- First quartile is 0-25% of the data
- Second quartile is 25-50% of the data
- Third quartile is 50-75% of the data
- Fourth quartile is 75-100% of the data

QUARTILE function

Returns the quartile of a data set. Quartiles often are used in sales and survey data to divide populations into groups. For example, you can use QUARTILE to find the top 25 percent of incomes in a population.

IMPORTANT This function has been replaced with one or more new functions that may provide improved accuracy and whose names better reflect their usage. This function is still available for compatibility with earlier versions of Excel. However, if backward compatibility is not required, you should consider using the new functions from now on, because they more accurately describe their functionality.

For more information about the new functions, see QUARTILE.EXC function and QUARTILE.INC function.

Syntax

QUARTILE(array,quart)

The QUARTILE function syntax has the following arguments:

- **Array** Required. The array or cell range of numeric values for which you want the quartile value.
- **Quart** Required. Indicates which value to return.

If quart equals	QUARTILE returns
0	Minimum value
1	First quartile (25th percentile)
2	Median value (50th percentile)
3	Third quartile (75th percentile)
4	Maximum value

ata Analytics for Cities by Richard Dunks is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Calculating Quartiles

3	Mode	1
4	Min	0
5	Max	2500
6	Range	2500
7	First Quartile	=QUARTILE('20150601_20150830_311_Noise'!G2:G54014,1)
8	Third Quartile	

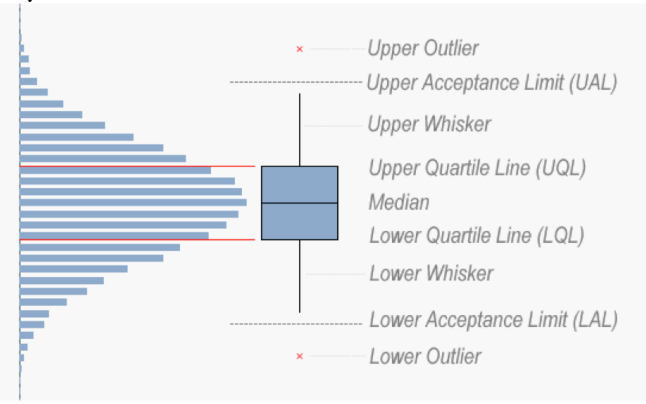
QUARTILE(array, quart)

5	Max	2500
6	Range	2500
7	First Quartile	1
8	Third Quartile	=QUARTILE('20150601_20150830_311_Noise'!G2:G54014,3)
9	IQR	

QUARTILE(array, quart)

Interquartile Range

- "Middle" 50% of data (between 1st Quartile and 3rd Quartile)



https://community.qlik.com/blogs/qlikviewdesignblog/2014/08/18/recipe-for-a-box-plot

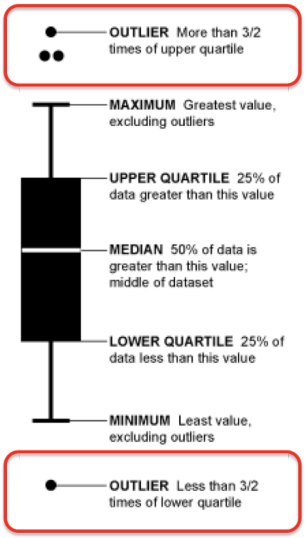
Calculate IQR

5	Max	2500
6	Range	2500
7	First Quartile	1
8	Third Quartile	18
9	IQR	=B8-B7

Outliers

- Any data points less than 1.5x the IQR or greater than 1.5x the IQR are considered outliers
- Helps identify data points that may skew the analysis
- Focus on the "meat" of the data

Outliers



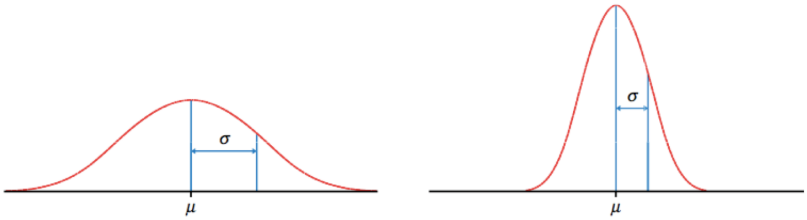
http://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/

Standard Deviation

- The average distance of each data point from the mean

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

- Larger the standard deviation, the greater the spread



Calculating Standard Deviation in Excel

7	First Quartile	1
8	Third Quartile	18
9	IQR	17
10	Standard Deviation	=STDEV('20150601_20150830_311_Noise'!$G$2:$G$54014)
11		

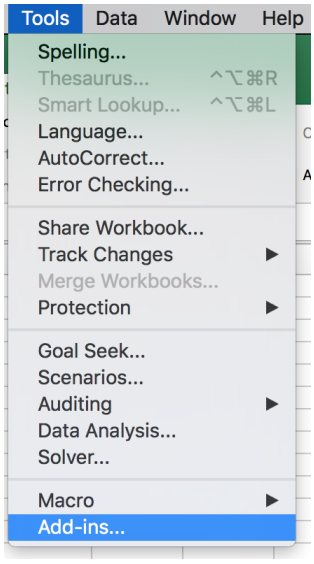
STDEV(number1, [number2], ...)

Measures of Variability

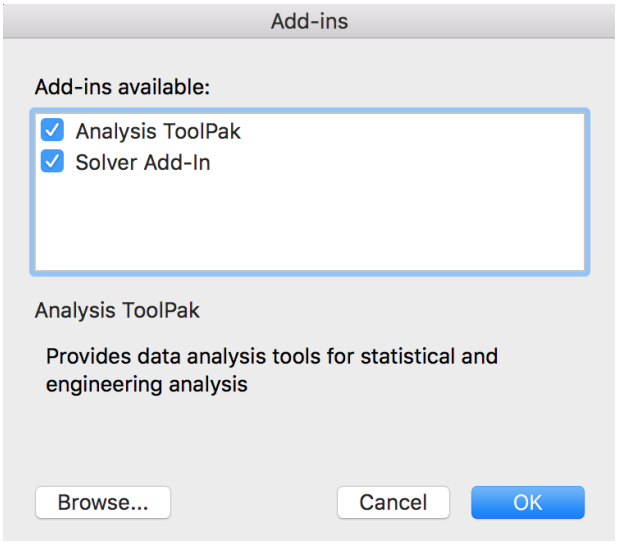
- Describe the distribution of our data
- Range (Maximum – Minimum)
- Inter-quartile Range
- Standard Deviation
- Identification of outliers (1.5 x IQR)

Enabling the Data Analysis Toolpak in Office 2016 for Mac

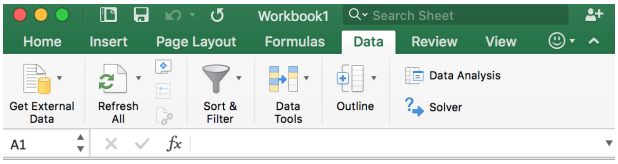
Enable Add-Ins



Enable Add-Ins



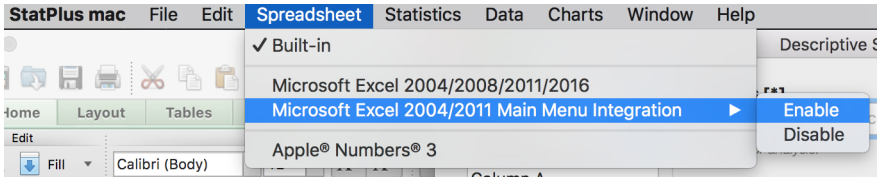
Restart Excel

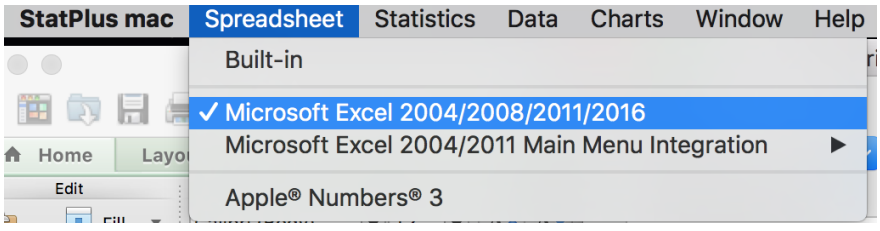


Enabling the Data Analysis Toolpak in Office 2016 for Mac

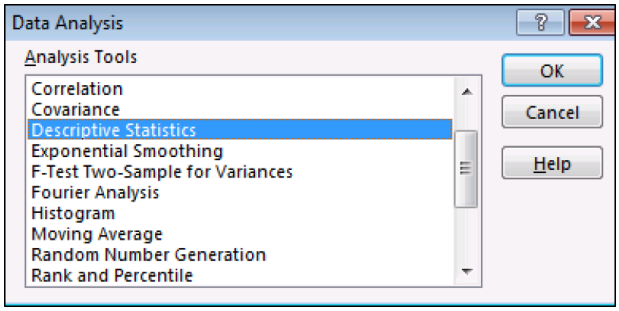
...there isn't one

Download StatPlus:mac

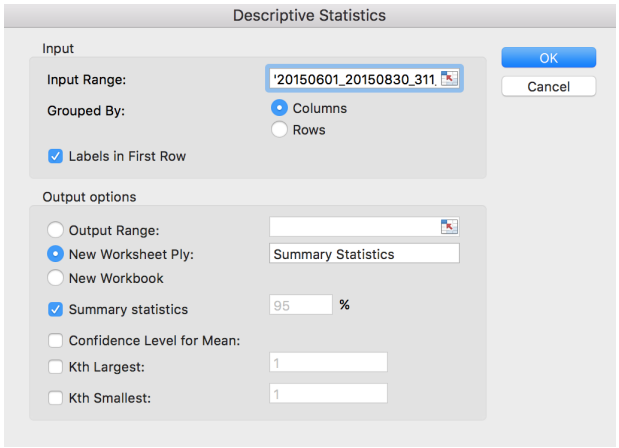




Descriptive Statistics in Excel 2013/2016



Descriptive Statistics in Excel 2013/2016

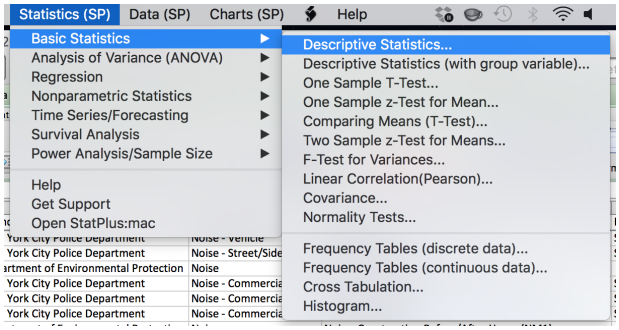


ata Analytics for Cities by Richard Dunks is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

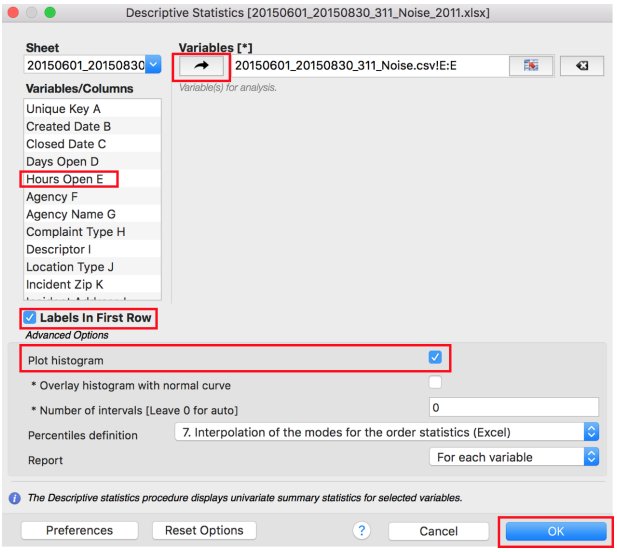
Descriptive Statistics in Excel 2013/2016

Hours Open	
Mean	41.44115306
Standard Error	0.47355833
Median	4
Mode	1
Standard Deviation	110.0582568
Sample Variance	12112.8199
Kurtosis	70.82804164
Skewness	6.720440262
Range	2500
Minimum	0
Maximum	2500
Sum	2238361
Count	54013

Descriptive Statistics in Excel 2011 for Mac

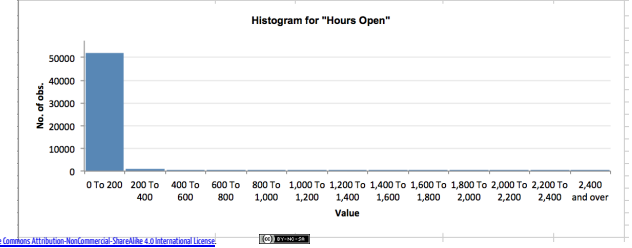


Descriptive Statistics in Excel 2011 for Mac



Descriptive Statistics in Excel 2011 for Mac

	A	B	C	D	E
	Alpha (for confidence interval)	5.%			
	Variable #1 (Hours Open)				
	Count	54013	*Mean Deviation*	58.33471	
	Mean	41.44115	*Second Moment*	12,112.59564	
	Mean LCL	40.51297	*Third Moment*	8,958,627.78165	
	Mean UCL	42.36933	*Fourth Moment*	1.08307E+10	
	Variance	12,112.8199			
	Standard Deviation	110.05826	*Sum*	2,238,361.	
	Mean Standard Error	0.47356	*Sum Standard Error*	25,578.30606	
	Coefficient of Variation	2.65577	*Total Sum Squares*	746,997,889.	
			Adjusted Sum Squares	654,237,628.20499	
	Minimum	0.			
	Maximum	2,500.	*Geometric Mean*	6.25357	
	Range	2,500.	*Harmonic Mean*	3.02523	
			Mode	1.	
	Median	4.			
	Median Error	0.00255	*Skewness*	6.72025	
	Percentile 25% (Q1)	1.	*Skewness Standard Error*	0.01054	
	Percentile 75% (Q3)	18.	*Kurtosis*	73.82137	
	IQR	17.	*Kurtosis Standard Error*	0.02108	
	MAD (Median Absolute Deviation)	117.	*Skewness (Fisher's)*	6.72044	
	Coefficient of Dispersion (COD)	9.9824	*Kurtosis (Fisher's)*	70.82804	



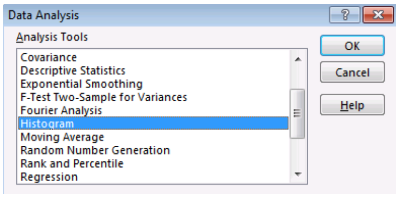
ata Analytics for Cities by Richard Dunks is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

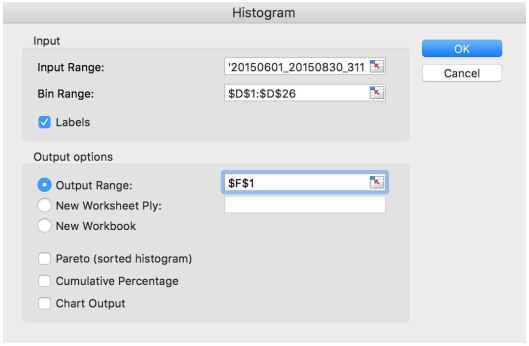
Histograms in Excel 2013/2016

Create Bins

A	B	C	D
Hours Open			Bins
			100
Mean	41.44115306		200
Standard Error	0.47355833		300
Median	4		400
Mode	1		500
Standard Deviation	110.0582568		600
Sample Variance	12112.8199		700
Kurtosis	70.82804164		800
Skewness	6.720440262		900
Range	2500		1000
Minimum	0		1100
Maximum	2500		1200
Sum	2238361		1300
Count	54013		1400
			1500
			1600
			1700
			1800
			1900
			2000
			2100
			2200
			2300
			2400
			2500

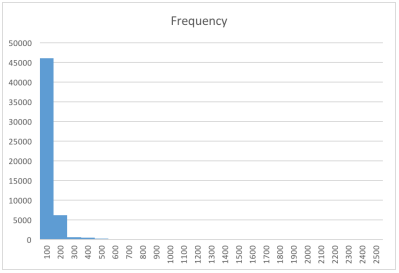
Open Data Analysis Toolpak

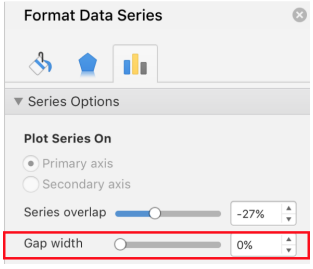




Results

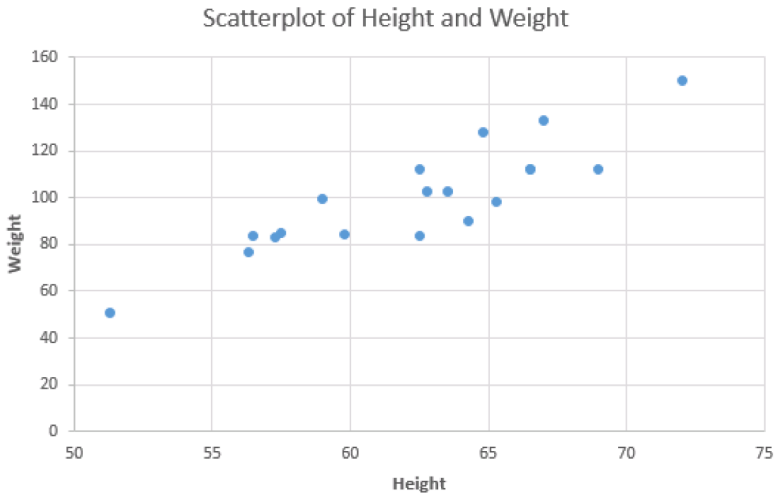
Bins		*Bins*	*Frequency*
100		100	46005
200		200	6206
300		300	582
400		400	484
500		500	168
600		600	145
700		700	71
800		800	91
900		900	106
1000		1000	37
1100		1100	19
1200		1200	30
1300		1300	36
1400		1400	10
1500		1500	2
1600		1600	6
1700		1700	0
1800		1800	3
1900		1900	1
2000		2000	2
2100		2100	5
2200		2200	1
2300		2300	2
2400		2400	0
2500		2500	1
		More	0





Bi-variate Analysis

Correlations

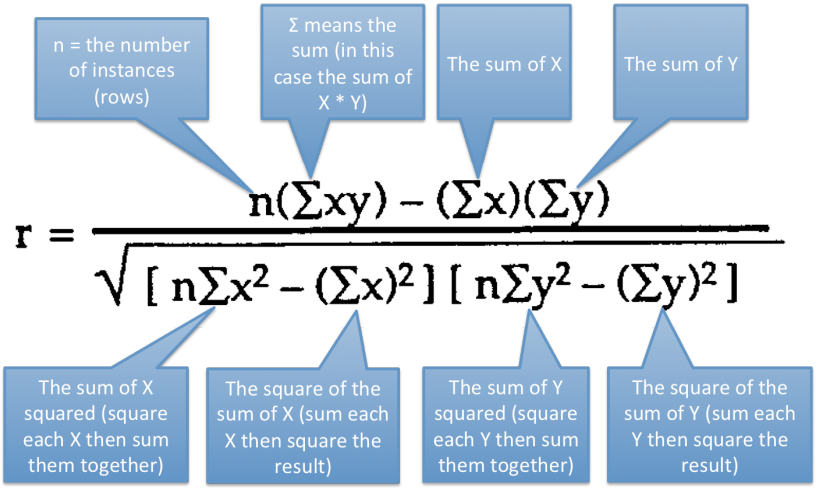


How do we measure this relationship?

Coefficient of Correlation

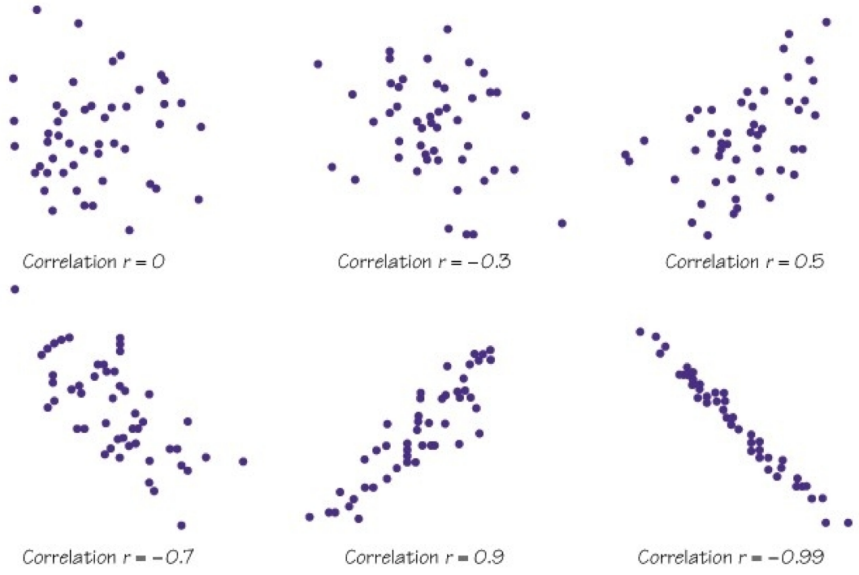
- Quantifies the amount of shared variability between variables
- Ranges between -1 and +1
- Negative numbers are inversely proportional
- Positive numbers are directly proportional
- The closer to either -1 or +1, the greater the correlation

Coefficient of Correlation

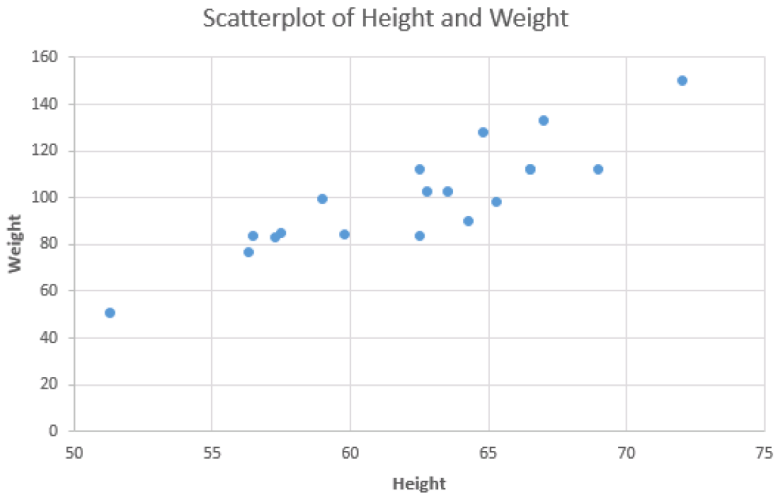


http://www.statisticshowto.com/what-is-the-correlation-coefficient-formula/

Coefficient of Correlation



Correlations - Height and Weight



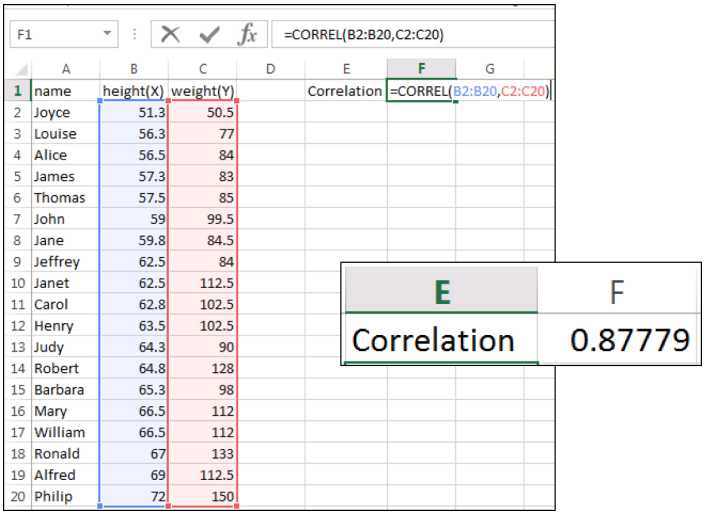
Download the data

name	height(X)	weight(Y)	X^2	Y^2	ΣXY
Joyce	51.3	50.5	2631.69	2550.25	101.8
Louise	56.3	77	3169.69	5929	133.3
Alice	56.5	84	3192.25	7056	140.5
James	57.3	83	3283.29	6889	140.3
Thomas	57.5	85	3306.25	7225	142.5
John	59	99.5	3481	9900.25	158.5
Jane	59.8	84.5	3576.04	7140.25	144.3
Jeffrey	62.5	84	3906.25	7056	146.5
Janet	62.5	112.5	3906.25	12656.25	175
Carol	62.8	102.5	3943.84	10506.25	165.3
Henry	63.5	102.5	4032.25	10506.25	166
Judy	64.3	90	4134.49	8100	154.3
Robert	64.8	128	4199.04	16384	192.8
Barbara	65.3	98	4264.09	9604	163.3
Mary	66.5	112	4422.25	12544	178.5
William	66.5	112	4422.25	12544	178.5
Ronald	67	133	4489	17689	200
Alfred	69	112.5	4761	12656.25	181.5
Philip	72	150	5184	22500	222
SUM	1184.4	1900.5	74304.92	199435.8	3084.9
n=19					

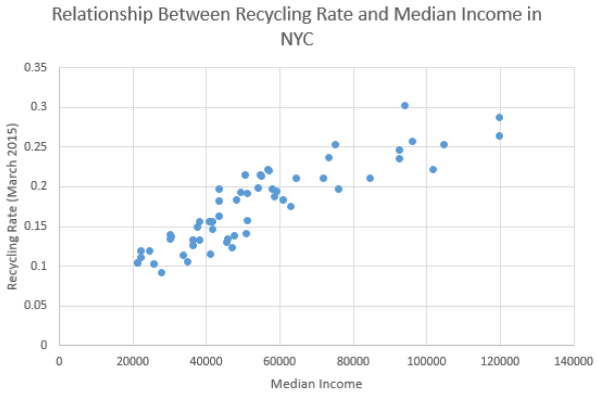
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

$$\frac{19(3084.9) - (1184.4)(1900.5)}{\sqrt{[(19 * 74304.92) - (1184.4)^2][(19 * 199435.75) - (1900.5)^2]}}$$

Or we could use Excel



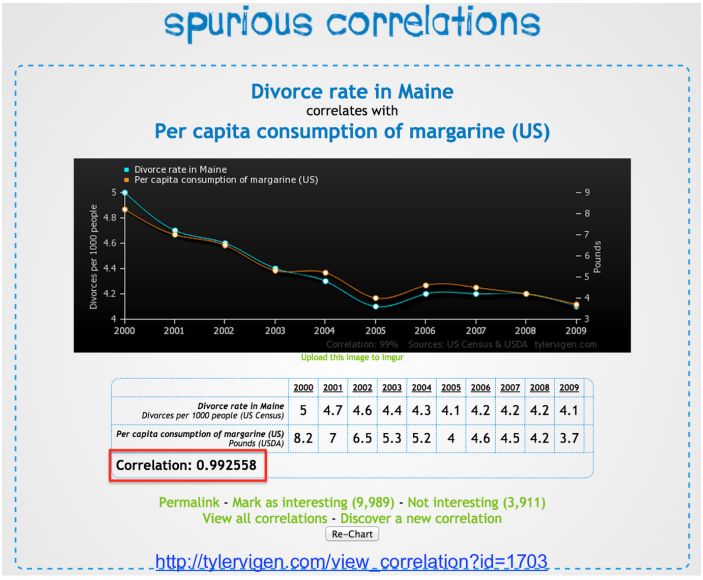
Correlations - Recycling and Median Income



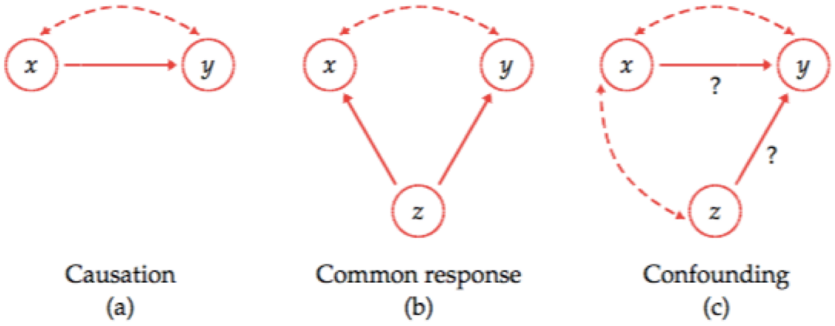
These are even slightly more correlated (r=0.88478) [Check it yourself]

http://iquantny.tumblr.com/post/79846201258/the-huge-correlation-between-median-income-and

Correlations



Correlations



Causation (a) | Common response (b) | Confounding (c)

Correlation does not imply causation

Prediction

- Knowing the relationship between variables (i.e. the correlation), we can predict values based on the relationship
- Can estimate the magnitude as well as the general trend
- More data points, the better the prediction
- Example -> Knowing the relationship between median income and recycling rates, what can we predict about recycling rates as median incomes grow in communities?

Linear Regression

- Using the known relationship between continuous variables, we can predict unseen values
- Assumes relationship is linear

Formula for a line

$$y = mx + b$$

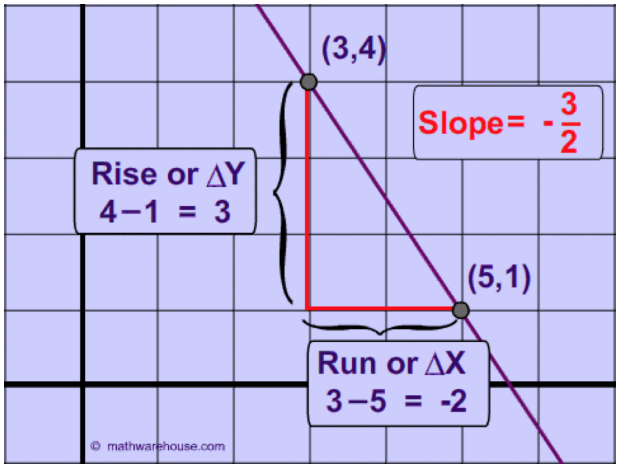
slope y-intercept

$$y = 3x - 5$$

slope y-intercept

http://www.algebra-class.com/slope-formula.html

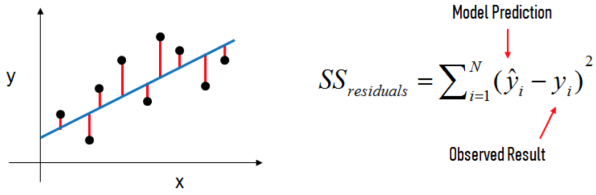
Formula for a line



http://www.mathwarehouse.com/algebra/linear_equation/slope-of-a-line.php

Formula for a line

- Draw a line that minimizes the distance between each point
- "Line of best fit" -> minimizes the sum of squared residuals

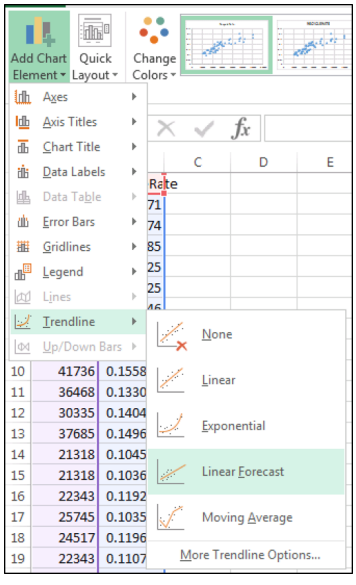


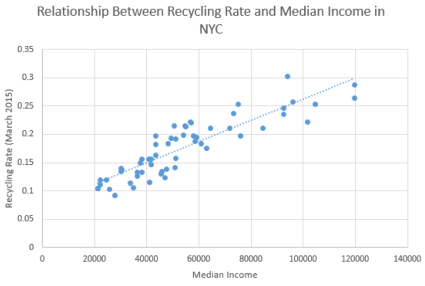
http://nbviewer.ipython.org/github/justmarkham/DAT4/blob/master/notebooks/08_linear_regression.ipynb

Linear Regression

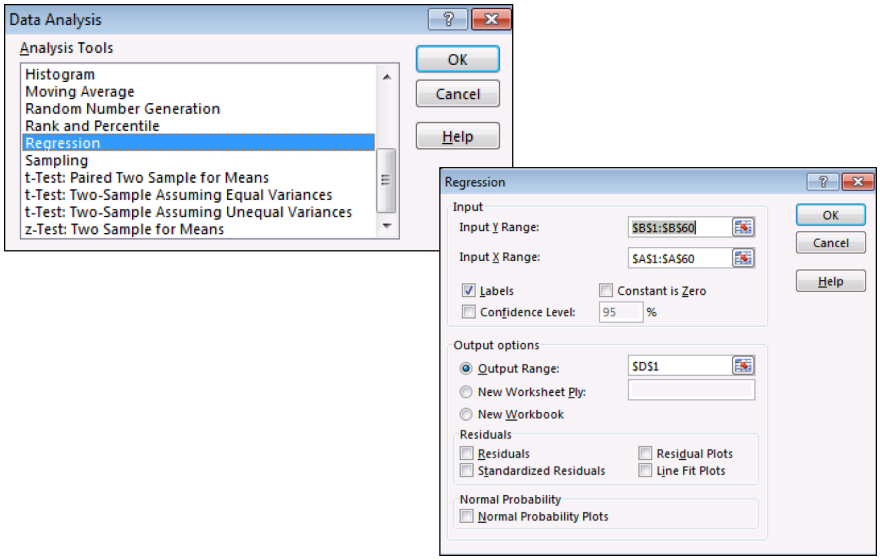
- Characteristics of the line defines the relationship
- Slope -> relationship between independent and dependent variable (how Y increases per unit of X)
- Intercept -> expected mean value of Y at X=0
- Values along the line are the predicted values for any given value X

Displaying a Trendline in Excel





Calculating coefficients in Excel



Calculating coefficients in Excel

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.884783183							
R Square	0.78284128							
Adjusted R Square	0.779031478							
Standard Error	0.024208426							
Observations	59							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	0.120421606	0.120421606	205.4808251	1.48386E-20			
Residual	57	0.03340473	0.000586048					
Total	58	0.153826336						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.074804136	0.007703657	9.710211521	1.09248E-13	0.05937783	0.090230443	0.05937783	0.090230443
MdHHIncE	1.86961E-06	1.30427E-07	14.33460237	1.48386E-20	1.60844E-06	2.13079E-06	1.60844E-06	2.13079E-06

Calculating coefficients in Excel

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.884783183							
R Square	0.78284128							
Adjusted R Square	0.779031478							
Standard Error	0.024208426							
Observations	59							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	0.120421606	0.120421606	205.4808251	1.48386E-20			
Residual	57	0.03340473	0.000586048					
Total	58	0.153826336						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.074804136	0.007703657	9.710211521	1.09248E-13	0.05937783	0.090230443	0.05937783	0.090230443
MdHHIncE	1.86961E-06	1.30427E-07	14.33460237	1.48386E-20	1.60844E-06	2.13079E-06	1.60844E-06	2.13079E-06

Linear Regression Line

	Coefficients
Intercept	0.07480414
MdHHIncE	1.8696E-06

Recycling Rate = 0.0000001869 * MedianIncome + 0.07480414

LibreOffice and OpenOffice

- Free and open-source office suites of software
- Fully featured with similar functionality to Microsoft Office
- And they're free
- http://www.openoffice.org/
- https://www.libreoffice.org/

OpenRefine (Google Refine)

- Free and open-source data cleaning tool
- Works in the browser
- Works with lots of data types
- Great for converting between Excel files, CSV, and JSON
- Facets are powerful (facet on anything)
- Meant specifically to clean data
- Available at: http://openrefine.org/

Python

- Computer language useful for working with data
- Easy to learn syntax for simple operations
- Lots of mathematical and scientific packages for advanced analysis
- Visualization packages for creating charts and graphs
- Can (generally) install on city computers
- Distribution available at: https://www.continuum.io/why-anaconda

R

- Open-source programming language for statistical analysis and visualization
- Implementation of the S programming language
- Relatively easy to use
- Packages for virtually any statistical technique or visualization
- Limited utility outside statistical analysis and visualization
- RStudio is an integrated development environment (IDE) that makes working with R easier

What we covered this afternoon

- Basic statistical measures, including measures of central tendency and measures of variability
- Using the Data Analysis Toolpak (and StatPlus:mac)
- Creating histograms in Excel
- Calculating coefficients of correlation
- Making predictions using linear regression
- Life beyond Excel

Possible topics for next training

- Data modeling with Excel (decision modeling, predictive analysis, clustering, etc)
- Mapping data with open-source tools
- Working with databases
- Statistical analysis in Python
- The practice of teaching urban analytics

A few warnings

- Don't be a solution in search of a problem
- Data and technology solutions are wasted if there wasn't a problem to begin with
- Context is king

Thank You

- richard@datapolitan.com
- http://blog.datapolitan.com
- @rdunks1/@datapolitan

ata Analytics for Cities by Richard Dunks is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.