NYC Citywide Administrative Services

Introduction to Statistical Analysis

Instructor: Richard Dunks

10 August 2016

Supplemental Material: http://bit.ly/stats-supplement

Datapolitan

Data Solutions for the Modern Metropolis

Introductions

- Name
- Agency
- Why you signed up for this course
- The number of siblings in your family (not including you)
- The number of years working for the City of New York
- Your height (in inches)

Goals for the Course

- Learn common statistical measures, including mean, median, mode, standard deviation, and variance
- Calculate correlation coefficients for bivariate data and apply the technique of simple regression analysis
- Demonstrate techniques used for forecasting
- Communicate data meaningfully to a broad audience using charts and graphs in Microsoft Excel

Key Takeaways for the Course

- You will be familiar with common statistical measures
- You will be able to calculate correlation coefficients for bivariate data and perform simple linear regression analysis
- You will be familiar with the basic techniques of forecasting
- You will be better able to communicate analysis using charts and graphs in Microsoft Excel

Key Assumptions

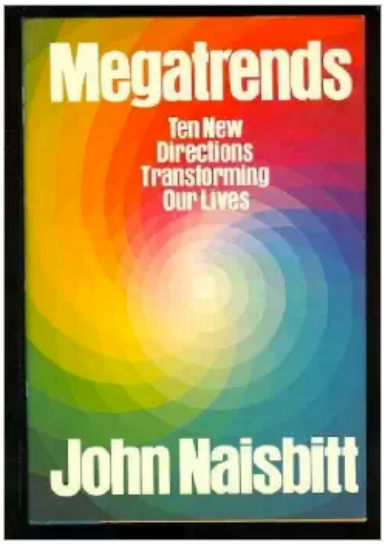
- You've had some previous experience with statistics and probability
- You're familiar with using Excel to manipulate data and calculate values
- You're familiar with using formulas in Excel

Disclaimer

- I'm not a statistician
- You won't be a statistician by the end of this course
- I often apply statistical tools and understanding in the work I do
- I'm assuming you all do the same, which is why you're here

Goals for this Morning

- Review basic statistical measures
- Practice using statistics in real-world applications
- Familiarize you with how to use Excel for statistical analysis



We are drowning in information but starved for knowledge.
- John Naisbitt

Why statistics?

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

LET'S RUN THE STATISTICS ON OUR CLASS TODAY

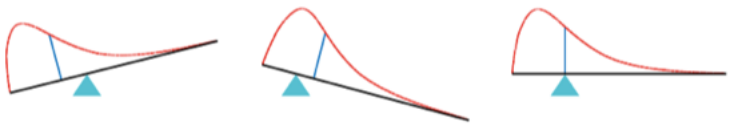
Open http://bit.ly/stats-data

Mean

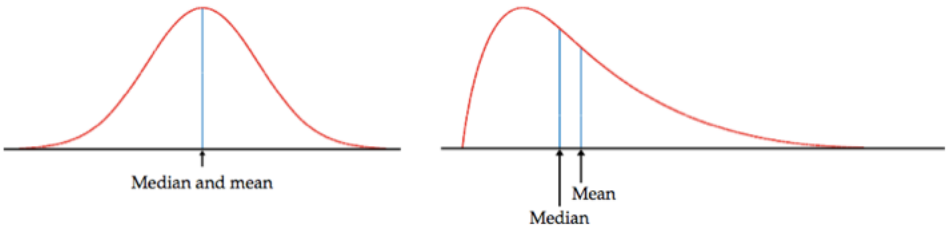
- A representative value for the data
- Usually what people mean by “average”
- Calculate by adding all the values together and dividing by the number instances
- Sensitive to extremes
- What’s the mean for the data we collected today? (Use the AVERAGE function)

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

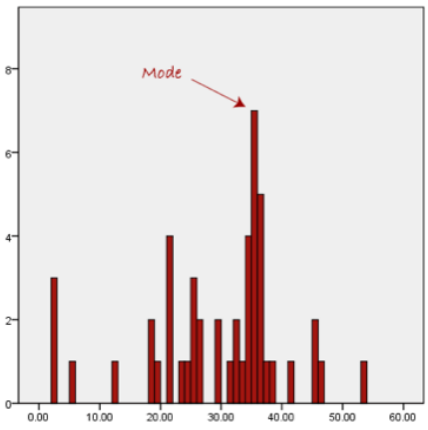


Mean vs Median



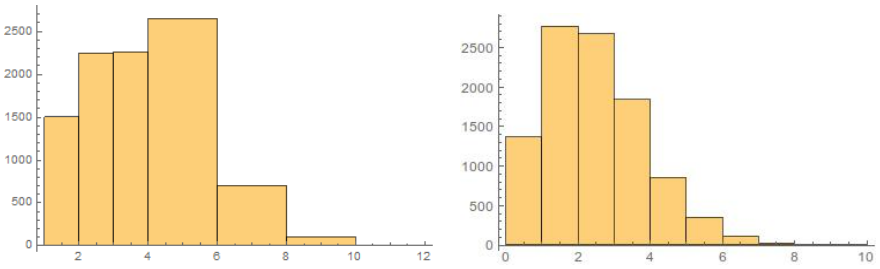
Mode

- The most frequent value in a dataset
- Often used for categorical data
- Use the MODE function in Excel to calculate



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

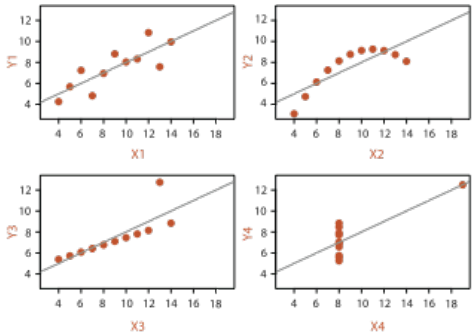


http://mathematica.stackexchange.com/questions/59520/histogram-with-variable-bin-size

Anscombe's Quartet

Anscombe's Quartet: Raw Data

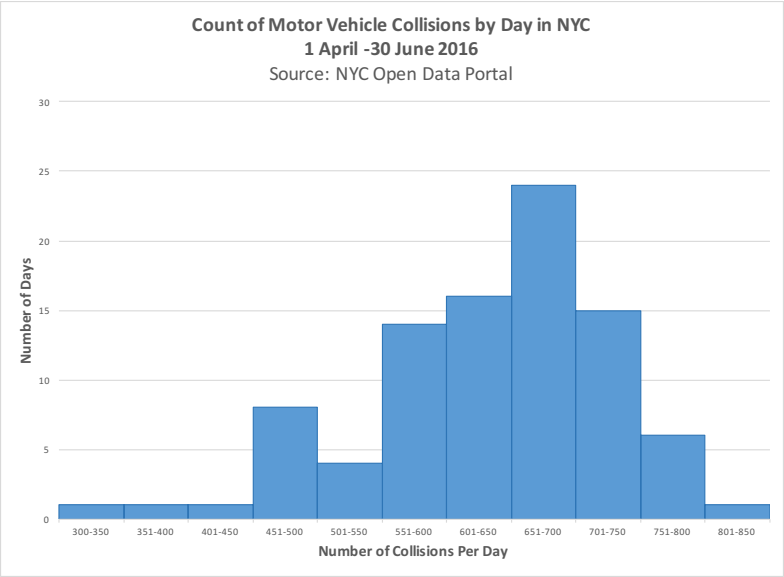
	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	



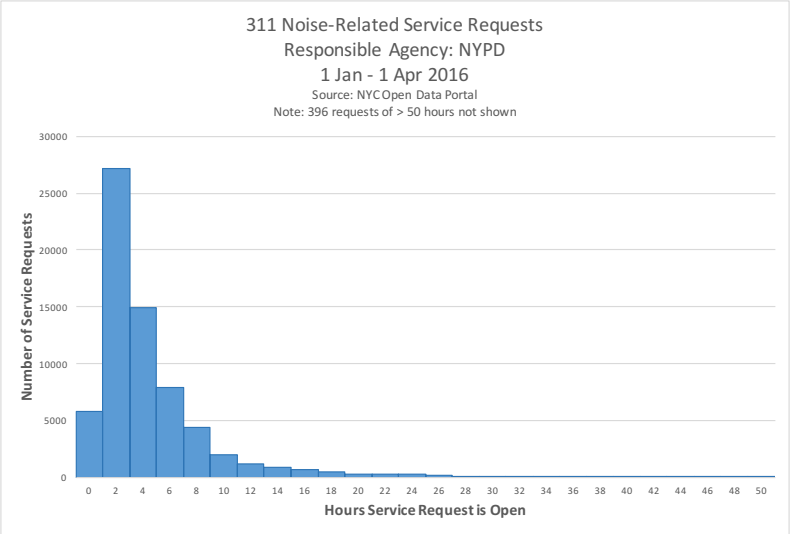
http://data.heapanalytics.com/anscombes-quartet-and-why-summary-statistics-dont-tell-the-whole-story/

DATA DISTRIBUTIONS

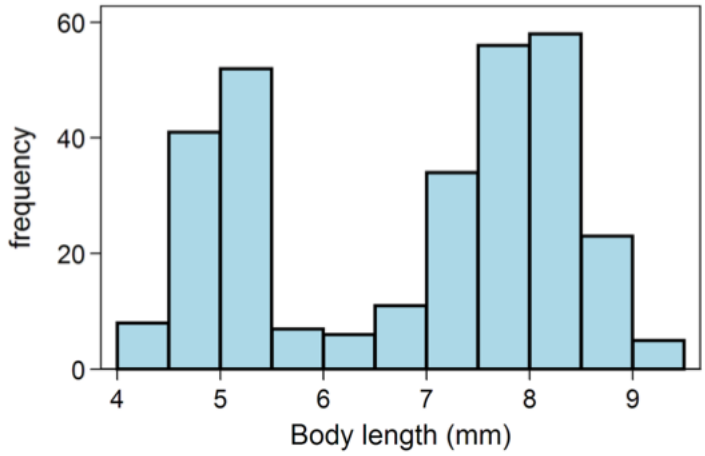
Normal Distribution



Long-tail Distribution



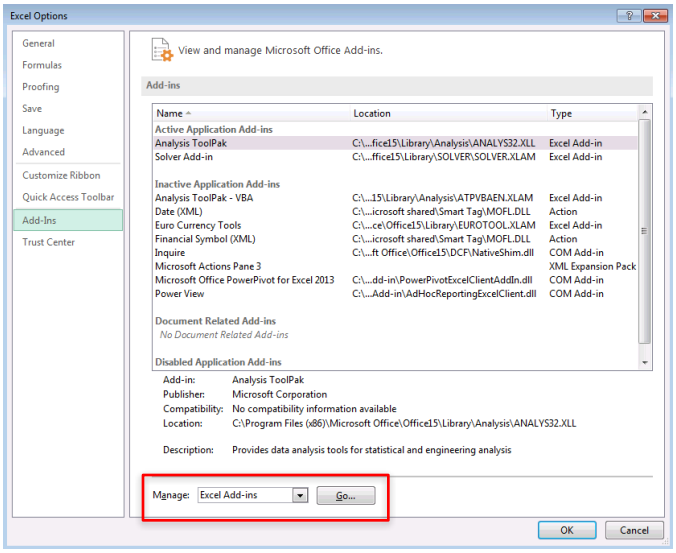
Bi-modal



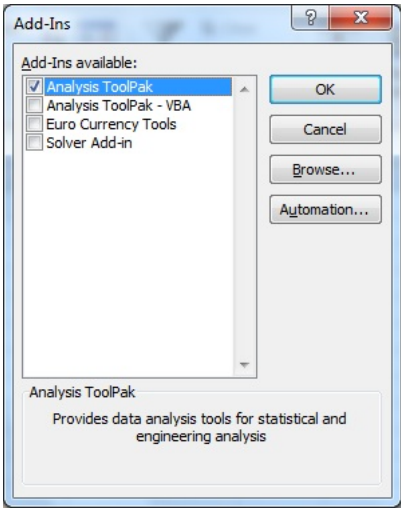
LET'S CREATE A HISTOGRAM OF OUR DATA

Installing Data Analysis ToolPak

- File
- Options
- Add-ins
- Manage
- "Go…"



Installing Data Analysis ToolPak



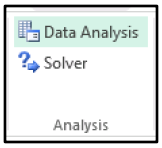
Setup Your Bins

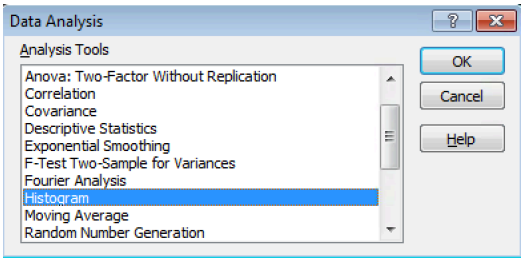
- Use an empty column and label it “Bins”
- Start with the minimum
- Create an entry for each bin you want
- Use a formula to save time

Bins
60
62
64
66
68
70
72
74

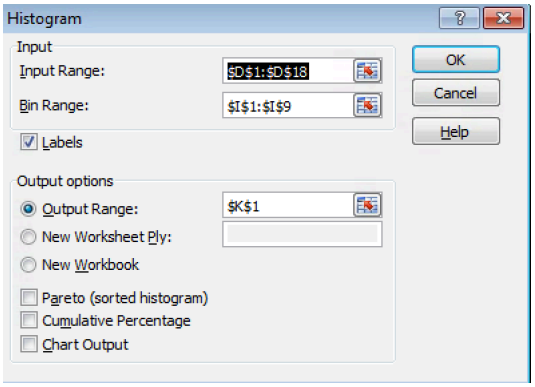
Creating a Histogram

Under the Data Ribbon



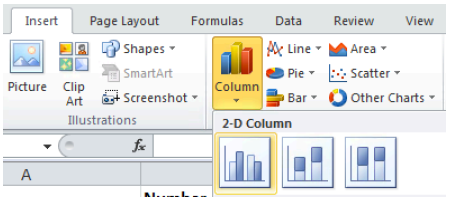


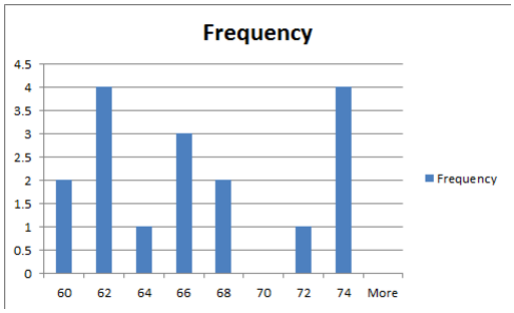
Creating a Histogram



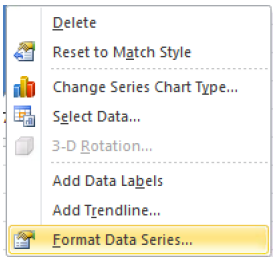
K	L
Bins	*Frequency*
60	2
62	4
64	1
66	3
68	2
70	0
72	1
74	4
More	0

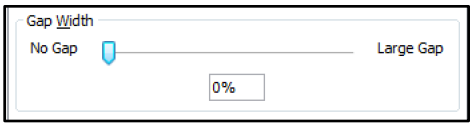
Creating a Histogram

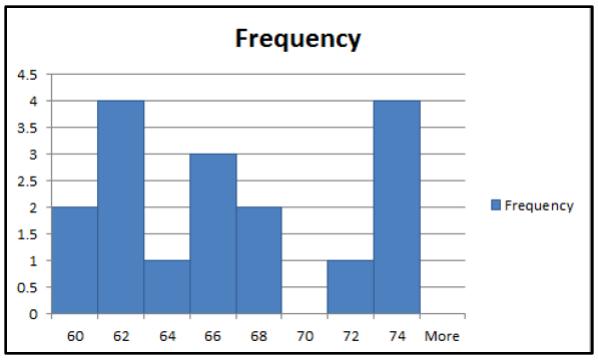




Creating a Histogram







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)
- Measurements
  - 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)

HOW DO WE MEASURE VARIABILITY IN A DATA SET?

Range

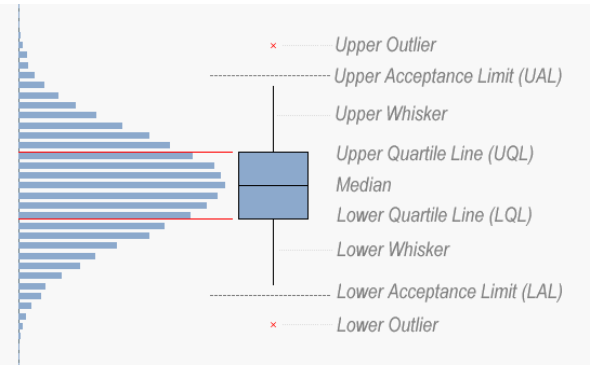
- The gap between the minimum value and the maximum value
- Calculated by subtracting the minimum from the maximum
- What's the range in the data we collected today? (Use the MAX and MIN functions)

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
- Use the QUARTILE function

Inter-Quartile Range

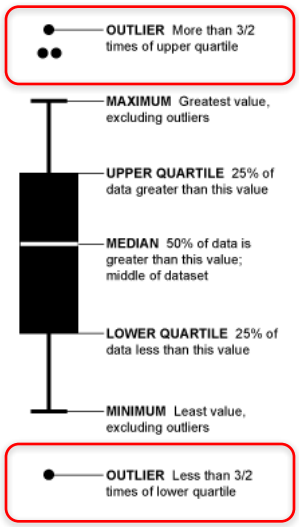
- "Middle" 50% of data (between 1st Quartile and 3rd Quartile)



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

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

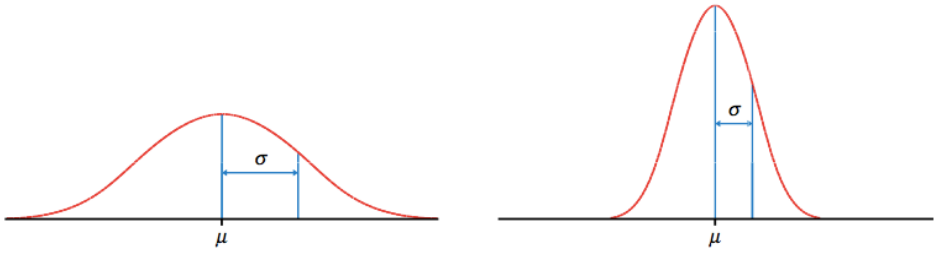


http://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/

WHAT ARE THE OUTLIERS IN OUR CLASS DATA?

Standard Deviation

- The average distance of each data point from the mean
- Larger the standard deviation, the greater the spread



Formula for Standard Deviation

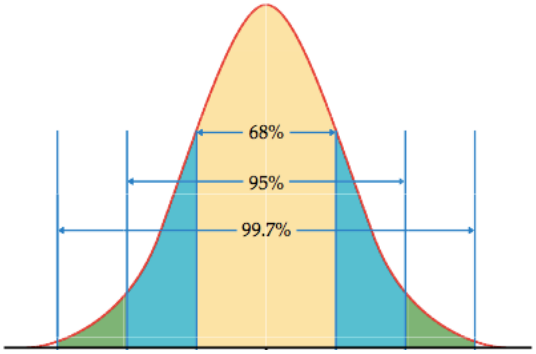
$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

Calculating the Standard Deviation

1. Subtract the mean from each data point
2. Square the result
3. Sum them together
4. Divide by the number of instances
5. Take the square root

- Or use the STDEV function in Excel

Standard Deviation



Measures of Variability

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

Descriptive Statistics

- Quantitatively describe the main features of a dataset
- Help distinguish distributions and make them comparable
- 5 number summary
  - Minimum
  - 1st Quartile
  - Median
  - 3rd Quartile
  - Maximum

Exploratory Data Analysis

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

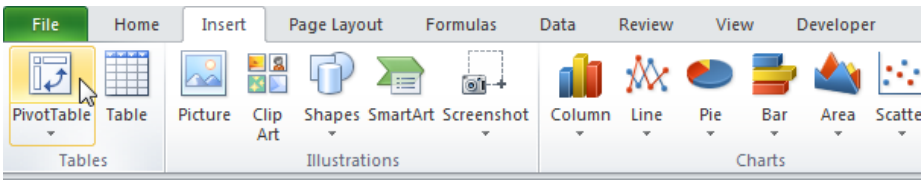
Tukey, J.W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

NYC MOTOR VEHICLE COLLISIONS

Find the Vehicle Collisions Data Supplemental Material: http://bit.ly/stats-supplement

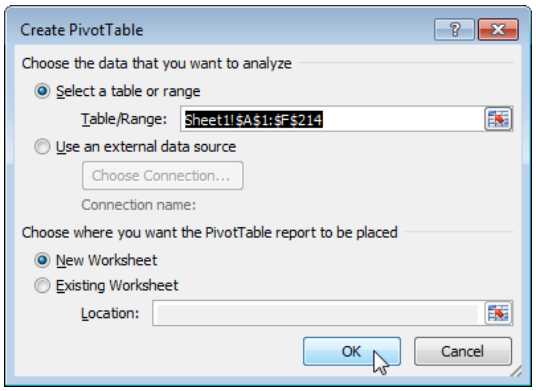
PivotTables

- What is a PivotTable?
  - A data summarization tool for quickly understanding and displaying the data you're analyzing
- How do I find it?

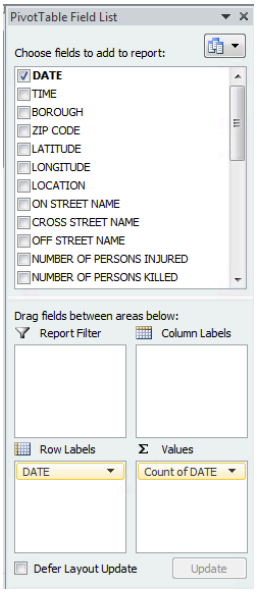


PivotTables

- Selecting range and destination

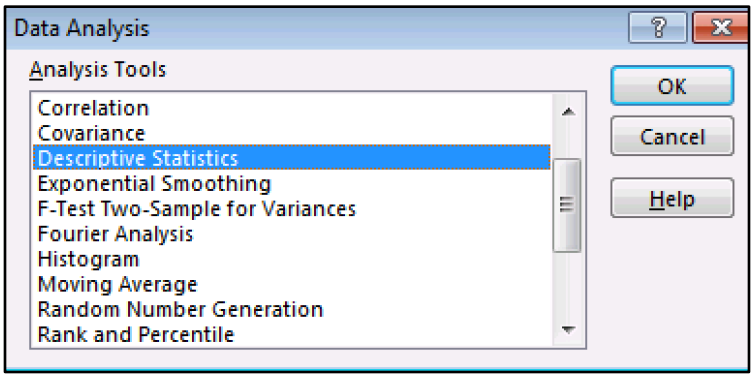


Create a PivotTable of Dates

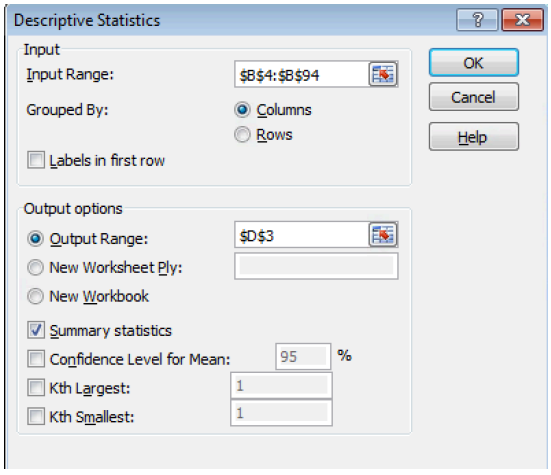


Row Labels	Count of DATE
4/1/2016	696
4/2/2016	587
4/3/2016	574
4/4/2016	675
4/5/2016	665
4/6/2016	612
4/7/2016	624
4/8/2016	673
4/9/2016	553
4/10/2016	490
4/11/2016	616
4/12/2016	630
4/13/2016	683
4/14/2016	668
4/15/2016	722

Calculating Descriptive Statistics



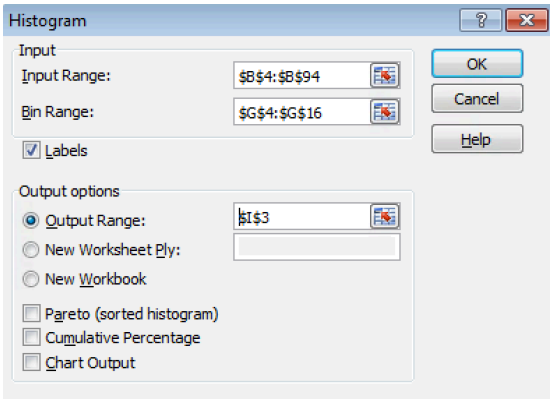
Calculating Descriptive Statistics



Column1	
Mean	632.6484
Standard Error	9.956442
Median	663
Mode	587
Standard Deviation	94.9784
Sample Variance	9020.897
Kurtosis	0.124912
Skewness	-0.73808
Range	462
Minimum	341
Maximum	803
Sum	57571
Count	91

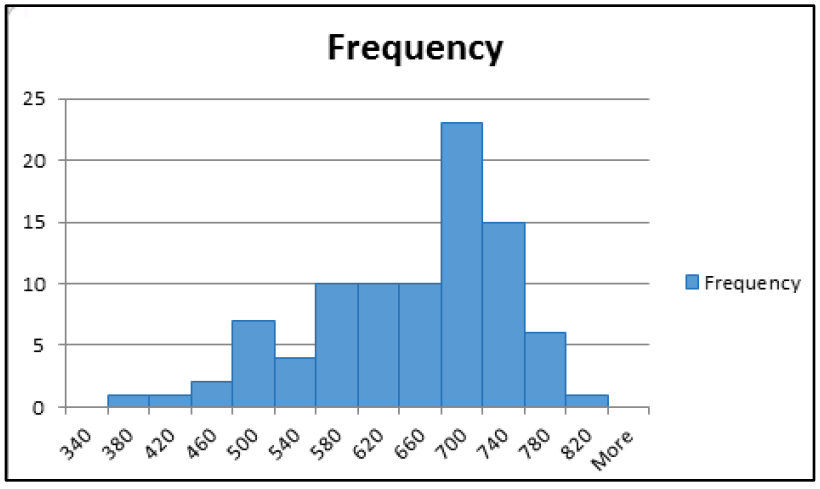
Create a Histogram

Bins
340
380
420
460
500
540
580
620
660
700
740
780
820



Bins	*Frequency*
340	0
380	1
420	1
460	2
500	7
540	4
580	10
620	10
660	10
700	23
740	15
780	6
820	1
More	0

Create a Histogram



Questions of this Data

- What is the mean number of accidents per day?
- Is mean or median the best way to describe this data?
- Are there any outliers in this data?

Wrap-up

- Reviewed basic descriptive statistics
- Calculated basic descriptive statistics in Excel
- Discussed histograms
- Created histograms in Excel
- Analyzed NYC motor vehicle collision data

Goals for the Afternoon

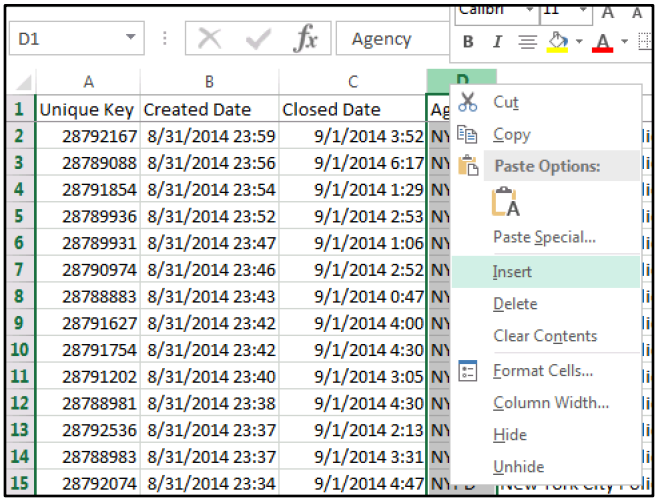
- Perform descriptive analysis on 311 noise complaints
- Introduce prediction and linear regression
- Create a linear regression model from NYC data
- Discuss decision models and perform a simple decision model example

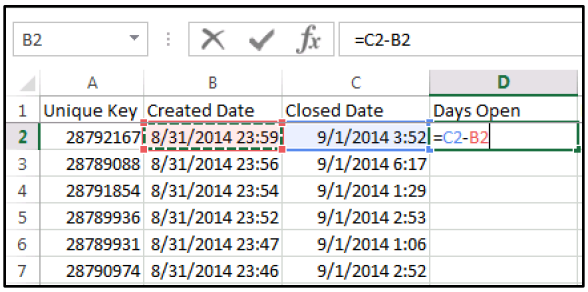
311 NOISE COMPLAINTS

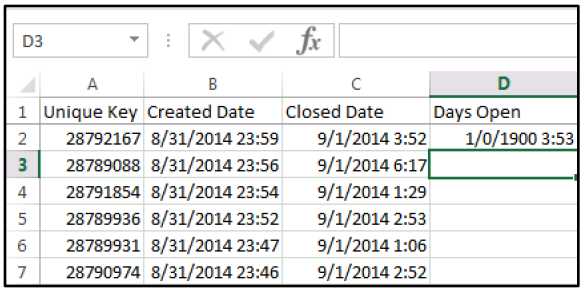
Open 20140601_20140901_311_noise.csv

PREPARING THE DATA

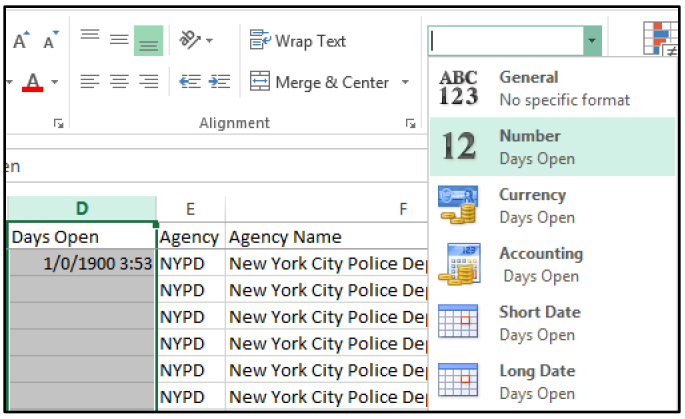
Insert a New Column



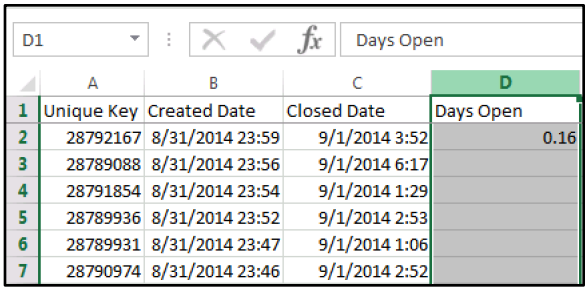




Reformat the Column Data Type

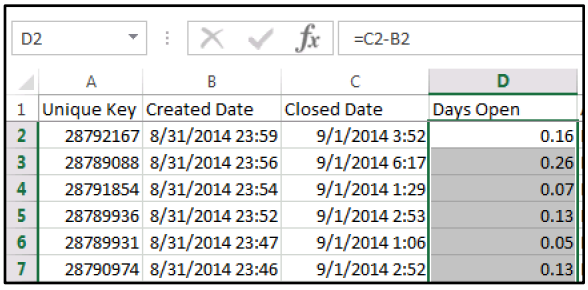


Paste Formulas Into Cells



D1 | Days Open

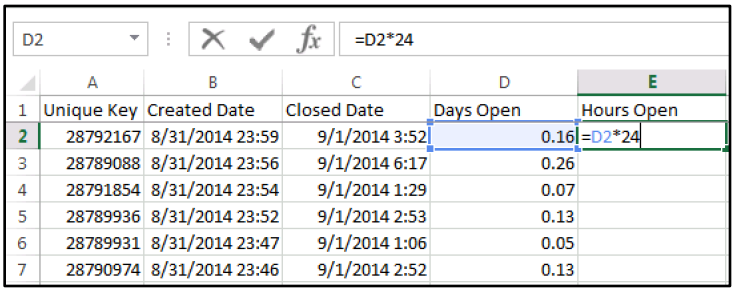
	A	B	C	D
1	Unique Key	Created Date	Closed Date	Days Open
2	28792167	8/31/2014 23:59	9/1/2014 3:52	0.16
3	28789088	8/31/2014 23:56	9/1/2014 6:17	
4	28791854	8/31/2014 23:54	9/1/2014 1:29	
5	28789936	8/31/2014 23:52	9/1/2014 2:53	
6	28789931	8/31/2014 23:47	9/1/2014 1:06	
7	28790974	8/31/2014 23:46	9/1/2014 2:52	



D2 | =C2-B2

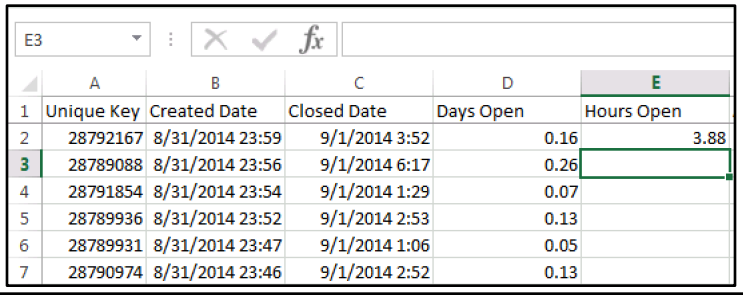
	A	B	C	D
1	Unique Key	Created Date	Closed Date	Days Open
2	28792167	8/31/2014 23:59	9/1/2014 3:52	0.16
3	28789088	8/31/2014 23:56	9/1/2014 6:17	0.26
4	28791854	8/31/2014 23:54	9/1/2014 1:29	0.07
5	28789936	8/31/2014 23:52	9/1/2014 2:53	0.13
6	28789931	8/31/2014 23:47	9/1/2014 1:06	0.05
7	28790974	8/31/2014 23:46	9/1/2014 2:52	0.13

Convert Time Open to Hours



D2 | =D2*24

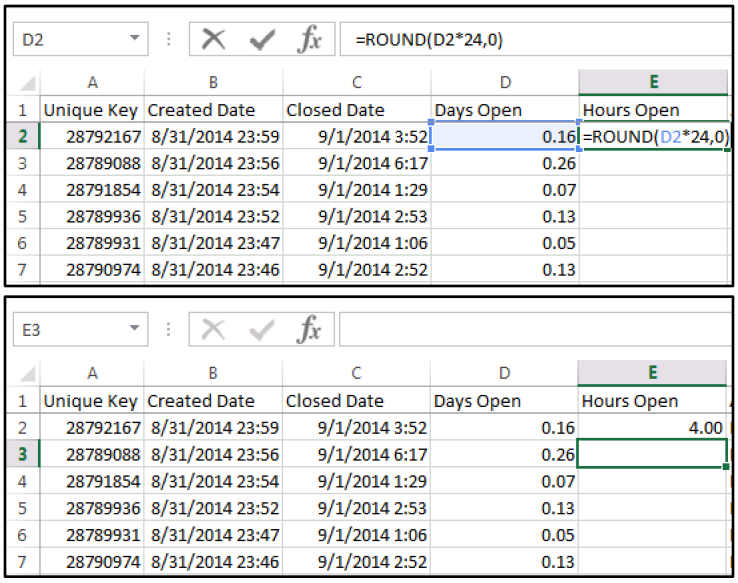
	A	B	C	D	E
1	Unique Key	Created Date	Closed Date	Days Open	Hours Open
2	28792167	8/31/2014 23:59	9/1/2014 3:52	0.16	=D2*24
3	28789088	8/31/2014 23:56	9/1/2014 6:17	0.26	
4	28791854	8/31/2014 23:54	9/1/2014 1:29	0.07	
5	28789936	8/31/2014 23:52	9/1/2014 2:53	0.13	
6	28789931	8/31/2014 23:47	9/1/2014 1:06	0.05	
7	28790974	8/31/2014 23:46	9/1/2014 2:52	0.13	



E3

	A	B	C	D	E
1	Unique Key	Created Date	Closed Date	Days Open	Hours Open
2	28792167	8/31/2014 23:59	9/1/2014 3:52	0.16	3.88
3	28789088	8/31/2014 23:56	9/1/2014 6:17	0.26	
4	28791854	8/31/2014 23:54	9/1/2014 1:29	0.07	
5	28789936	8/31/2014 23:52	9/1/2014 2:53	0.13	
6	28789931	8/31/2014 23:47	9/1/2014 1:06	0.05	
7	28790974	8/31/2014 23:46	9/1/2014 2:52	0.13	

Let's Round the Hours Instead



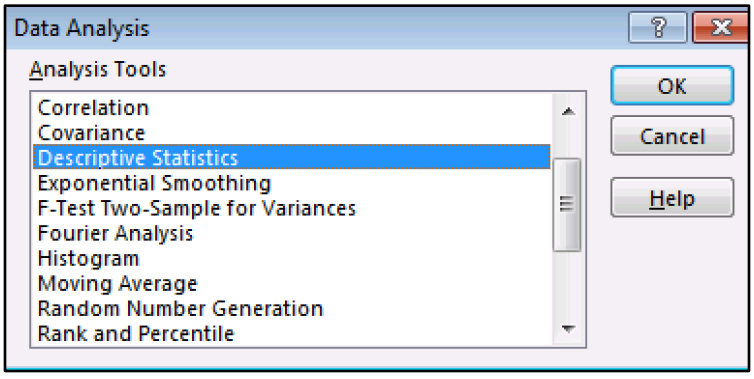
D2 =ROUND(D2*24,0)

	A	B	C	D	E
1	Unique Key	Created Date	Closed Date	Days Open	Hours Open
2	28792167	8/31/2014 23:59	9/1/2014 3:52	0.16	=ROUND(D2*24,0)
3	28789088	8/31/2014 23:56	9/1/2014 6:17	0.26	
4	28791854	8/31/2014 23:54	9/1/2014 1:29	0.07	
5	28789936	8/31/2014 23:52	9/1/2014 2:53	0.13	
6	28789931	8/31/2014 23:47	9/1/2014 1:06	0.05	
7	28790974	8/31/2014 23:46	9/1/2014 2:52	0.13	

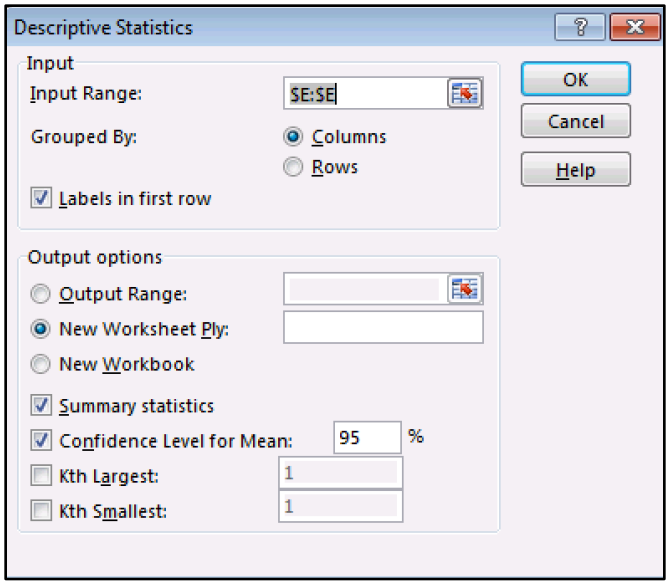
E3

	A	B	C	D	E
1	Unique Key	Created Date	Closed Date	Days Open	Hours Open
2	28792167	8/31/2014 23:59	9/1/2014 3:52	0.16	4.00
3	28789088	8/31/2014 23:56	9/1/2014 6:17	0.26	
4	28791854	8/31/2014 23:54	9/1/2014 1:29	0.07	
5	28789936	8/31/2014 23:52	9/1/2014 2:53	0.13	
6	28789931	8/31/2014 23:47	9/1/2014 1:06	0.05	
7	28790974	8/31/2014 23:46	9/1/2014 2:52	0.13	

Calculating Descriptive Statistics



Calculating Descriptive Statistics

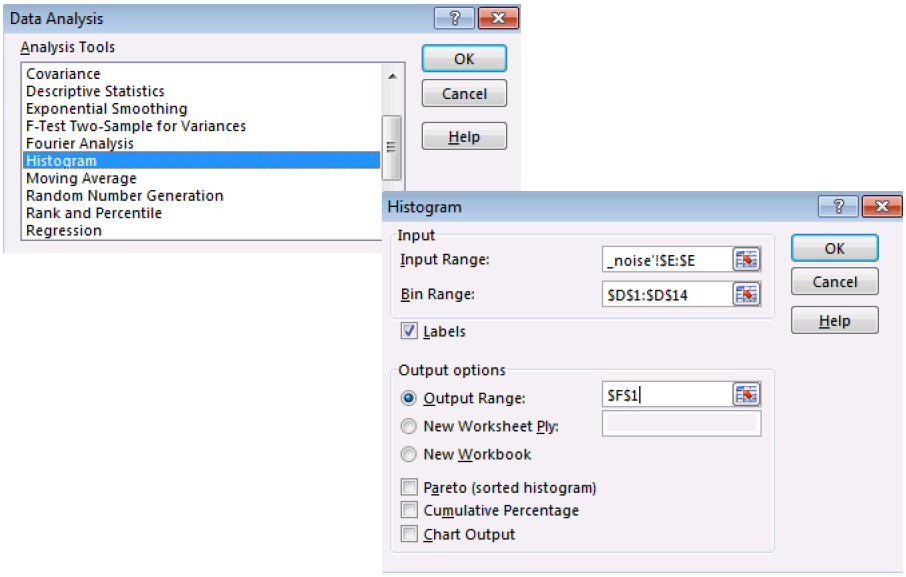


Specifying Bins for Histogram

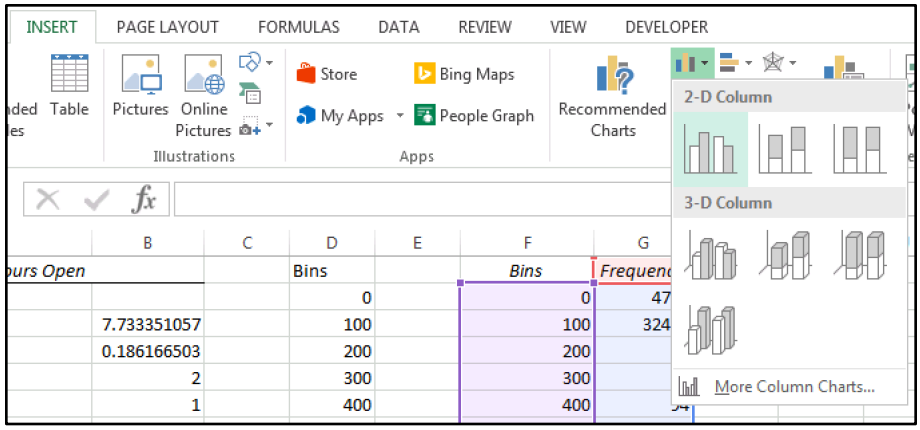
D23

	A	B	C	D
1	*Hours Open*			Bins
2				0
3	Mean	7.733351057		100
4	Standard Error	0.186166503		200
5	Median	2		300
6	Mode	1		400
7	Standard Deviation	36.10622415		500
8	Sample Variance	1303.659422		600
9	Kurtosis	229.5590789		700
10	Skewness	12.97119479		800
11	Range	1158		900
12	Minimum	0		1000
13	Maximum	1158		1100
14	Sum	290890		1200
15	Count	37615		
16	Confidence Level(95.0%)	0.364891383		

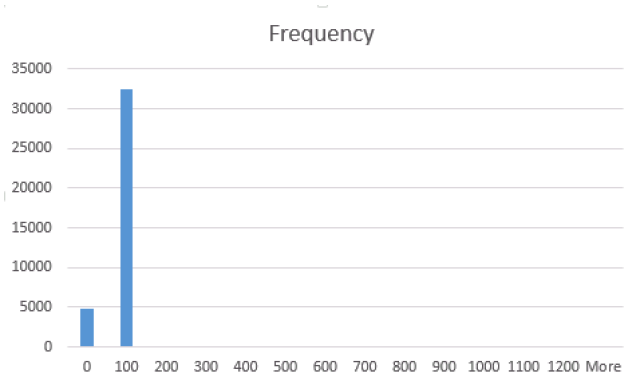
Creating a Histogram in Excel



Creating a Histogram in Excel

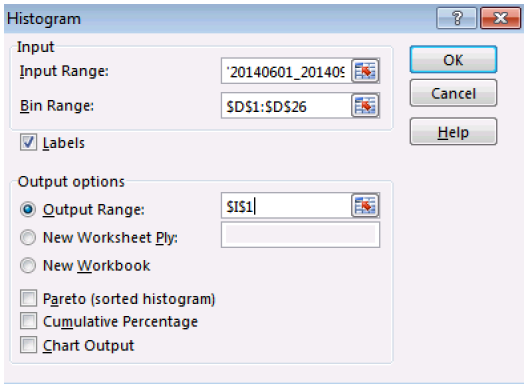


Histogram with Bin Size = 100

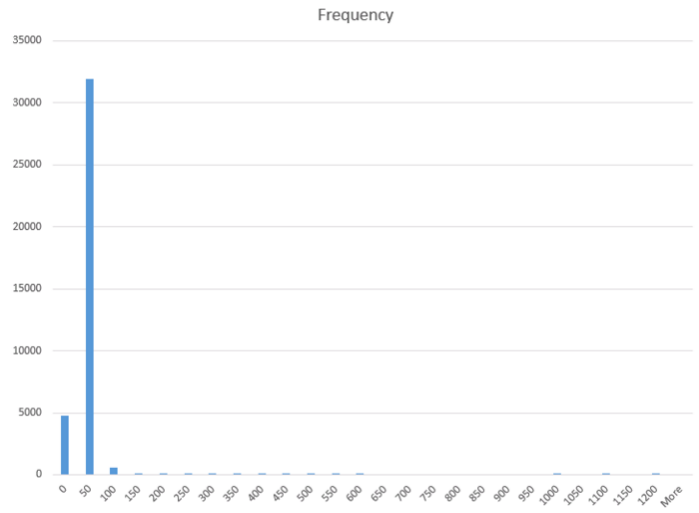


Change Bin Size

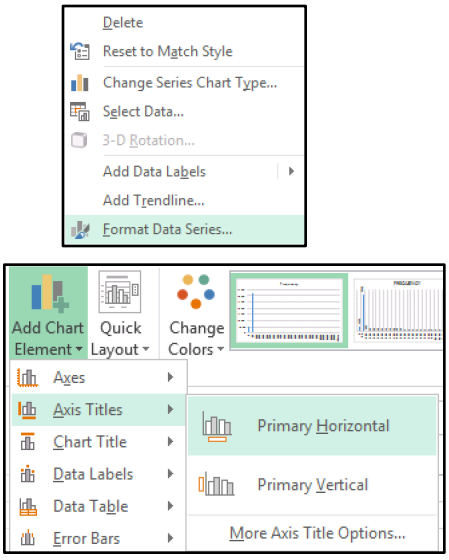
Bins
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1000
1050
1100
1150
1200

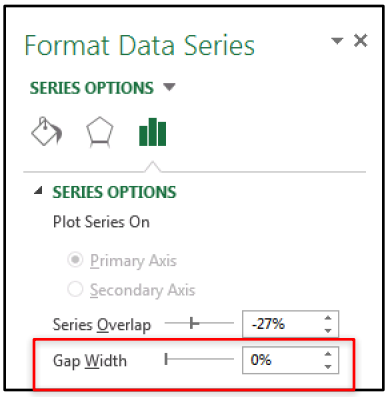


Histogram with Bin Size = 50

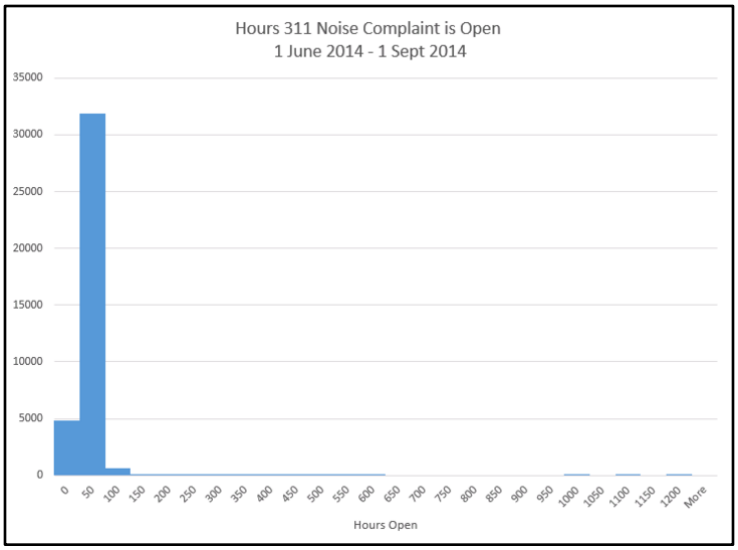


Format Chart





Formatting Chart



What do we know?

- The median time a noise complaint is open is 2 hours
- 50% of the noise complaints are closed between 1-4 hours (median is 2 hours, IQR is 3 hours)
- There is a long tail of complaints that take longer to close (range of 1158 hours, standard deviation of 36 hours)

Wrap-Up

- Descriptive Statistics
  - Measures of central tendency
  - Measures of variability
- Analyzed the results
- Created a histogram of the data

Correlation

- Values tend to have a relationship
- That relationship can be of several types
  - Proportional (increase in one increases the other)
  - Inversely proportional (increase in one decreases the other)
- Example
  - Height and weight

LET'S CHECK THAT OUT

Open height_weight.xlsx

Correlations

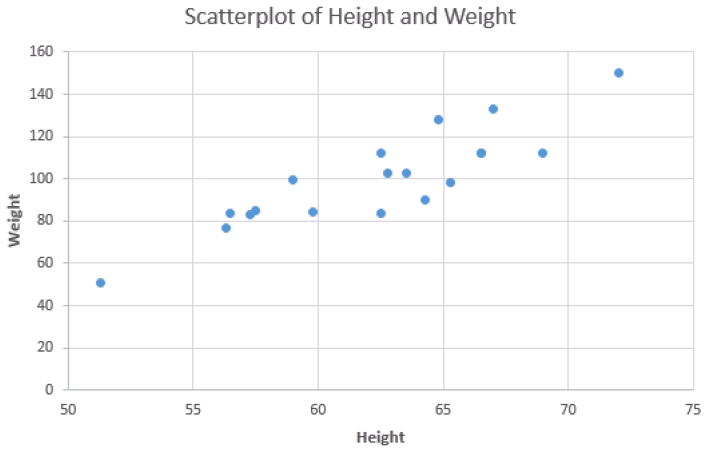
name	height	weight	age
Alfred	69	112.5	14
Alice	56.5	84	13
Barbara	65.3	98	13
Carol	62.8	102.5	14
Henry	63.5	102.5	14
James	57.3	83	12
Jane	59.8	84.5	12
Janet	62.5	112.5	15
Jeffrey	62.5	84	13
John	59	99.5	12
Joyce	51.3	50.5	11
Judy	64.3	90	14
Louise	56.3	77	12
Mary	66.5	112	15
Philip	72	150	16
Robert	64.8	128	12
Ronald	67	133	15
Thomas	57.5	85	11
William	66.5	112	15

Sorted by height and weight

name	height	weight	age
Joyce	51.3	50.5	11
Louise	56.3	77	12
Alice	56.5	84	13
James	57.3	83	12
Thomas	57.5	85	11
John	59	99.5	12
Jane	59.8	84.5	12
Jeffrey	62.5	84	13
Janet	62.5	112.5	15
Carol	62.8	102.5	14
Henry	63.5	102.5	14
Judy	64.3	90	14
Robert	64.8	128	12
Barbara	65.3	98	13
Mary	66.5	112	15
William	66.5	112	15
Ronald	67	133	15
Alfred	69	112.5	14
Philip	72	150	16

http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_introreg_sect003.htm

Correlations

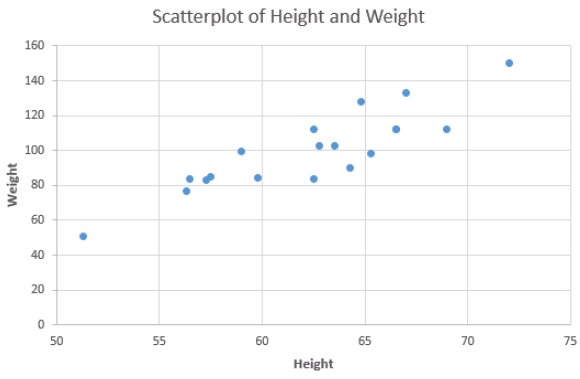


How do we measure this relationship?

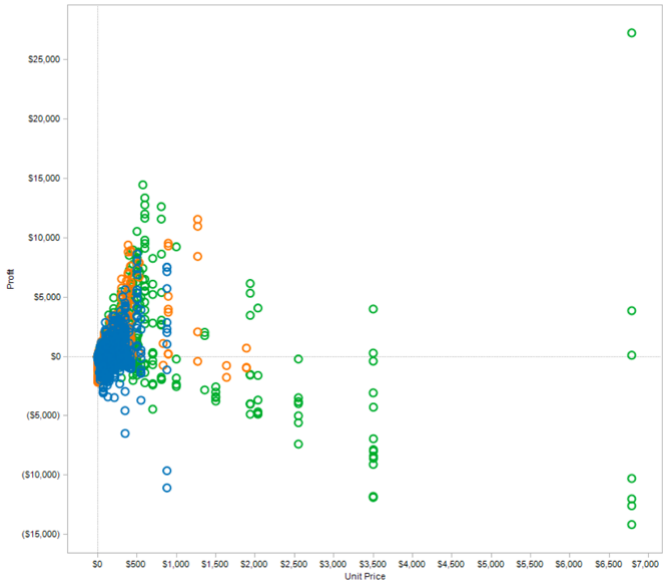
Scatter Plot

- Plots relationship between two continuous variables

name	height(X)	weight(Y)
Joyce	51.3	50.5
Louise	56.3	77
Alice	56.5	84
James	57.3	83
Thomas	57.5	85
John	59	99.5
Jane	59.8	84.5
Jeffrey	62.5	84
Janet	62.5	112.5
Carol	62.8	102.5
Henry	63.5	102.5
Judy	64.3	90
Robert	64.8	128
Barbara	65.3	98
Mary	66.5	112
William	66.5	112
Ronald	67	133
Alfred	69	112.5
Philip	72	150



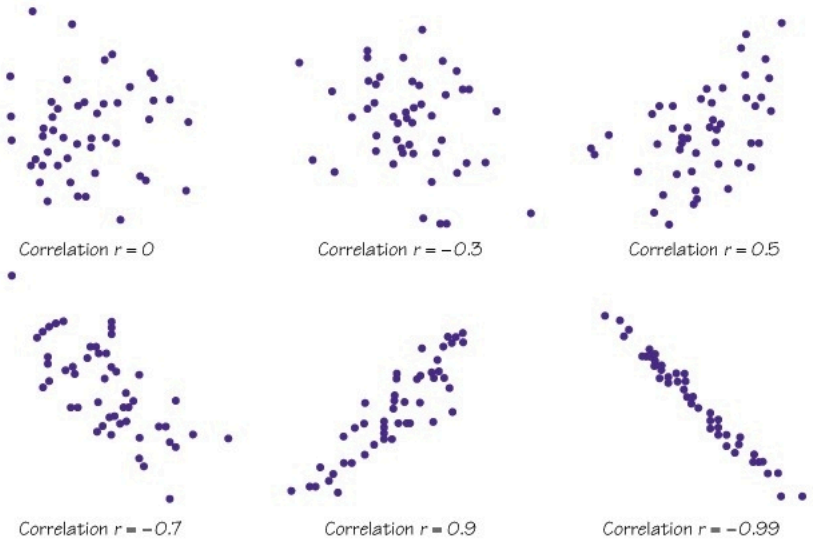
Scatter Plot



Correlation Coefficient

- 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

Correlation Coefficient



http://pixshark.com/correlation-examples.htm

Calculating the Correlation Coefficient

n = the number of instances (rows)

Σ means the sum (in this case the sum of X * Y)

The sum of X

The sum of Y

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

The sum of X squared (square each X then sum them together)

The square of the sum of X (sum each X then square the result)

The sum of Y squared (square each Y then sum them together)

The square of the sum of Y (sum each Y then square the result)

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

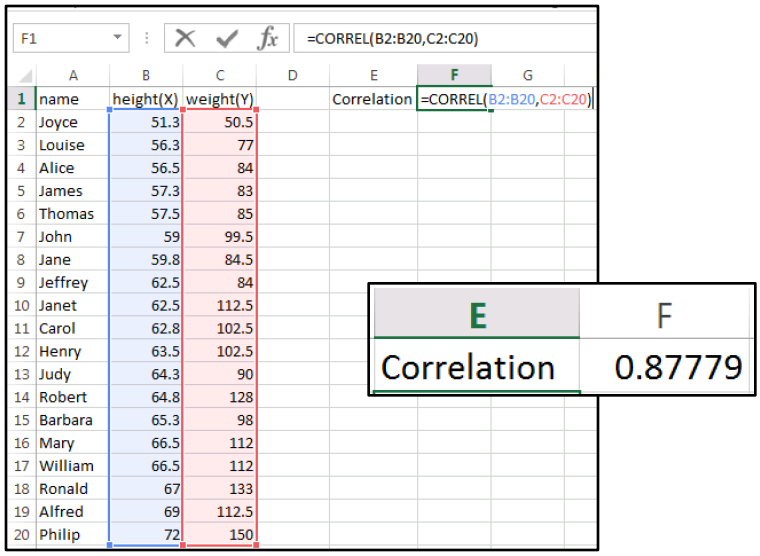
name	height(X)	weight(Y)	X^2	Y^2	ΣXY
Joyce	51.3	50.5	2631.69	2550.25	101.8
Louise	56.3	77	3169.69	5929	133.3
Alice	56.5	84	3192.25	7056	140.5
James	57.3	83	3283.29	6889	140.3
Thomas	57.5	85	3306.25	7225	142.5
John	59	99.5	3481	9900.25	158.5
Jane	59.8	84.5	3576.04	7140.25	144.3
Jeffrey	62.5	84	3906.25	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					

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

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

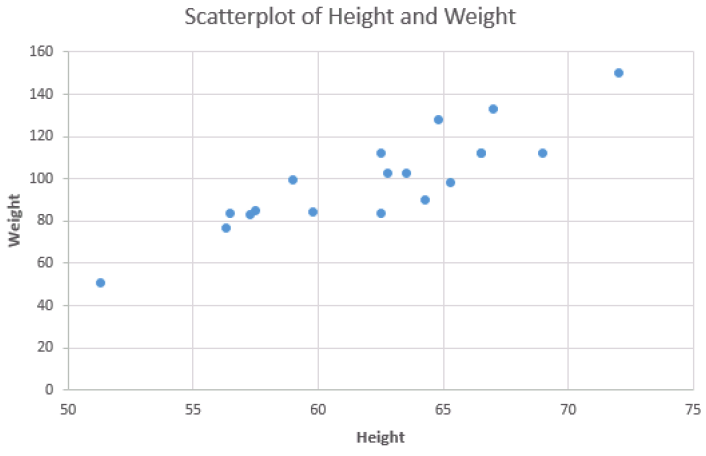
OR WE COULD USE EXCEL

Correlation Coefficient

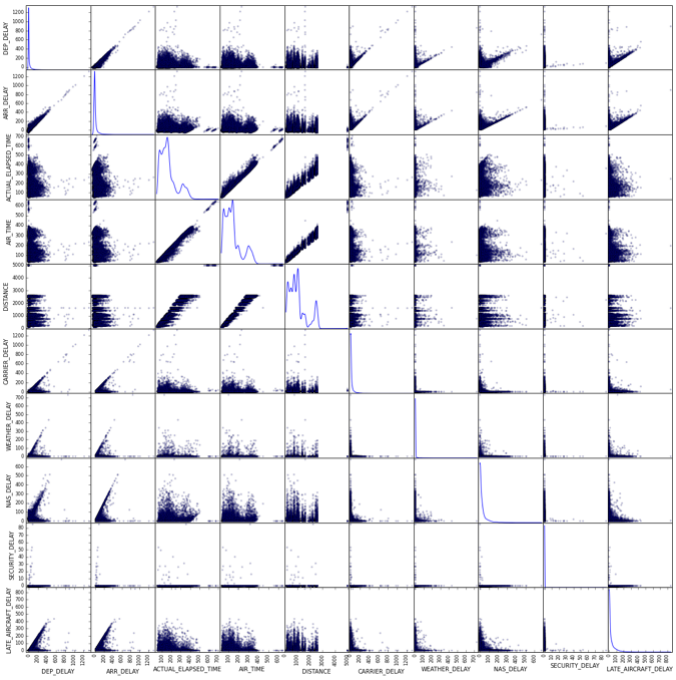


	A	B	C	D	E	F	G
1	name	height(X)	weight(Y)		Correlation	=CORREL(B2:B20,C2:C20)	
2	Joyce	51.3	50.5				
3	Louise	56.3	77				
4	Alice	56.5	84				
5	James	57.3	83				
6	Thomas	57.5	85				
7	John	59	99.5				
8	Jane	59.8	84.5				
9	Jeffrey	62.5	84				
10	Janet	62.5	112.5				
11	Carol	62.8	102.5				
12	Henry	63.5	102.5				
13	Judy	64.3	90				
14	Robert	64.8	128				
15	Barbara	65.3	98				
16	Mary	66.5	112				
17	William	66.5	112				
18	Ronald	67	133				
19	Alfred	69	112.5				
20	Philip	72	150				

Correlations



These are highly positively correlated (r=0.88)



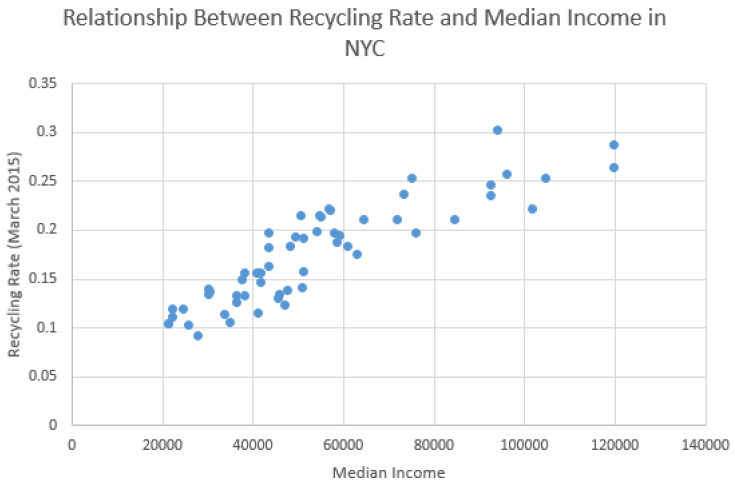
http://pandas.pydata.org/pandas-docs/stable/visualization.html#visualization-scatter-matrix

LET'S TRY THAT AGAIN

Is the recycling rate in NYC Community Districts correlated to income level in NYC?

Open 2013_NYC_CD_MedianIncome_Recycle.xlsx

Correlation



Correlation Coefficient

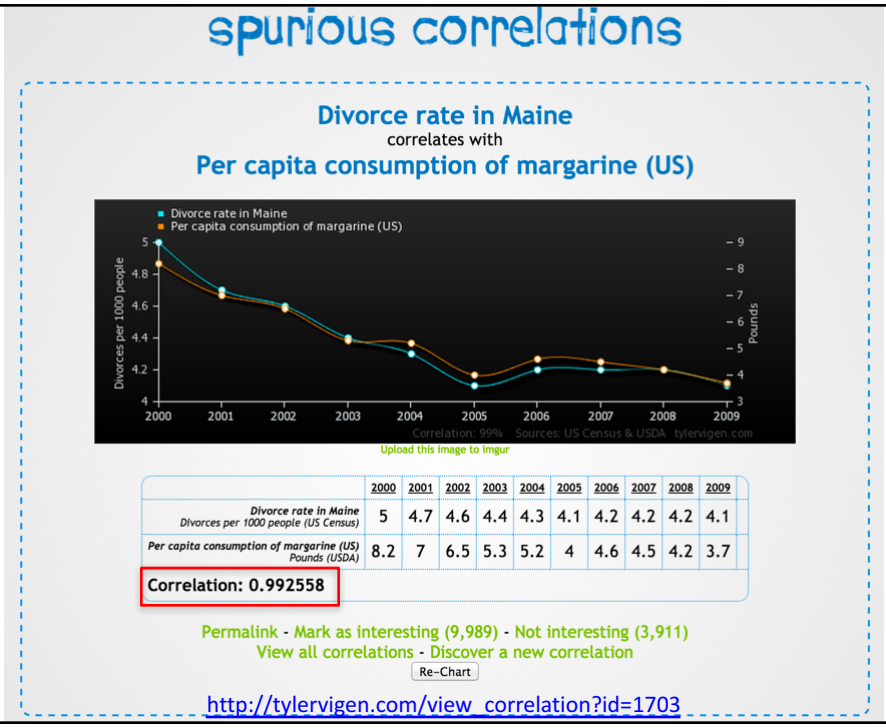
VLOOKUP =CORREL(C2:C60,D2:D60)

	A	B	C	D	E	F	G	H
1	Boro_CD	CD_Name	MdHHIncE	RecycleRate		Correlation Coefficient	=CORREL(C2:C60,D2:D60)	
26	301	Greenpoint & Williamsburg	50778	0.141620918				
27	302	Brooklyn Heights & Fort Greene	73290	0.237204648				
28	303	Bedford-Stuyvesant	36528	0.125817529				
29	304	Bushwick	38274	0.132462808				
30	305	East New York & Starrett City	33700	0.114029688				
31	306	Park Slope, Carroll Gardens & Red Hook	93969	0.302797999				
32	307	Sunset Park & Windsor Terrace	43351	0.197696644				
33	308	Crown Heights North & Prospect Heights	41075	0.156240603				
34	309	Crown Heights South, Prospect Lefferts & Wingate	41095	0.115119055				
35	310	Bay Ridge & Dyker Heights	57006	0.220854749				
36	311	Bensonhurst & Bath Beach	48252	0.18339298				
37	312	Borough Park, Kensington & Ocean Parkway	38215	0.156079659				

Correlation Coefficient	0.884783

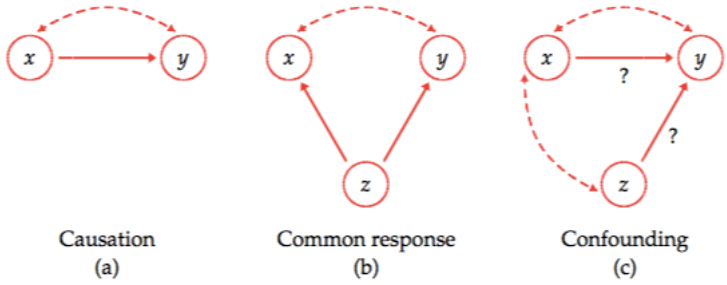
Correlation of Determination

- The percentage of variance in one variable shared with the other
- More shared variability implies a stronger relationship
- Calculate by squaring the correlation coefficient
  - Ex. The correlation of determination for median income vs recycling rates is 78%



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine Divorces per 1000 people (US Census)	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) Pounds (USDA)	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Causation



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

$$y = mx + b$$

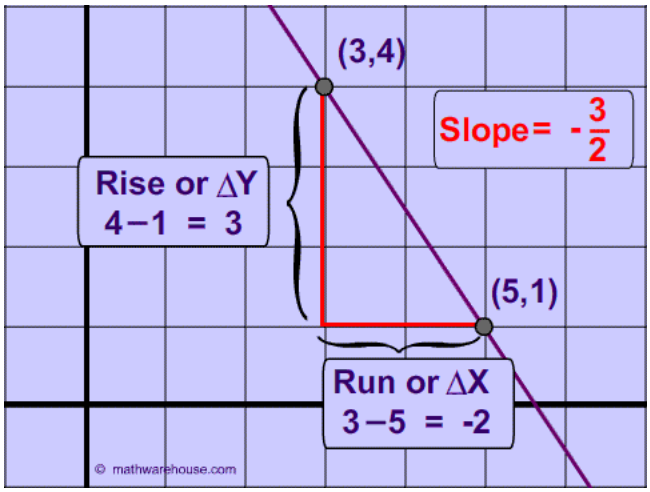
slope y-intercept

$$y = 3x - 5$$

slope y-intercept

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

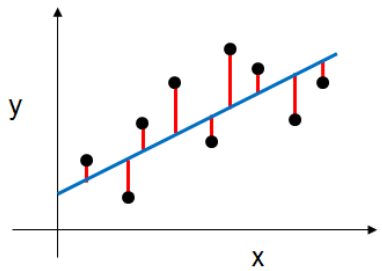
Slope



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

Linear Regression

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



Model Prediction

$$SS_{residuals} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

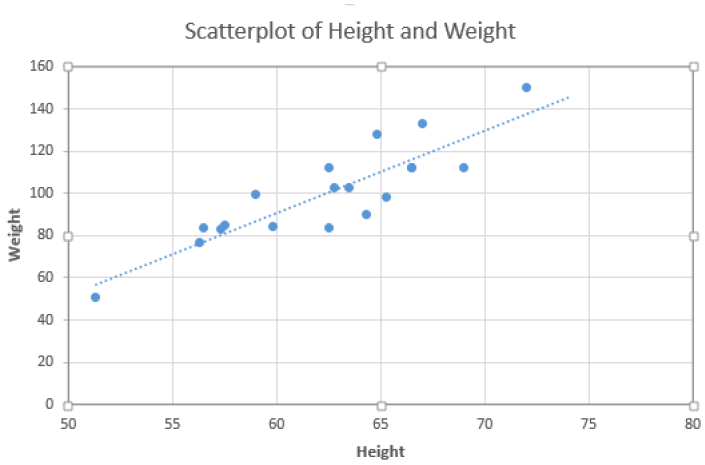
Observed Result

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

Linear Regression

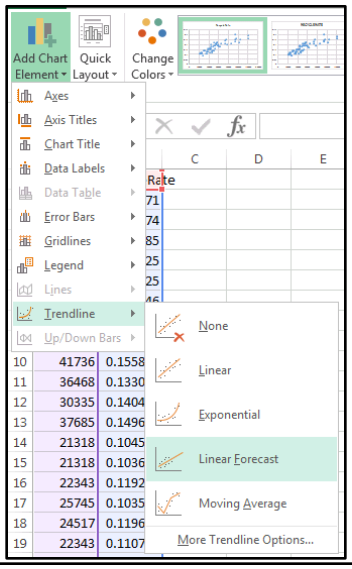
- Characteristics of the line defines the relationship
- Slope -> relationship between independent and dependent variable (how Y increases per unit of X)
- Intercept -> expected mean value of Y at X=0
- Values along the line are the predicted values for any given value X

Linear Regression

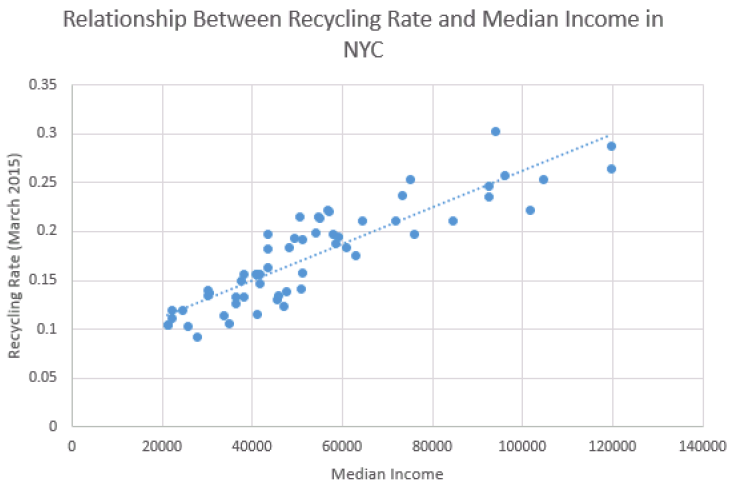


Adding a Linear Trendline in Excel

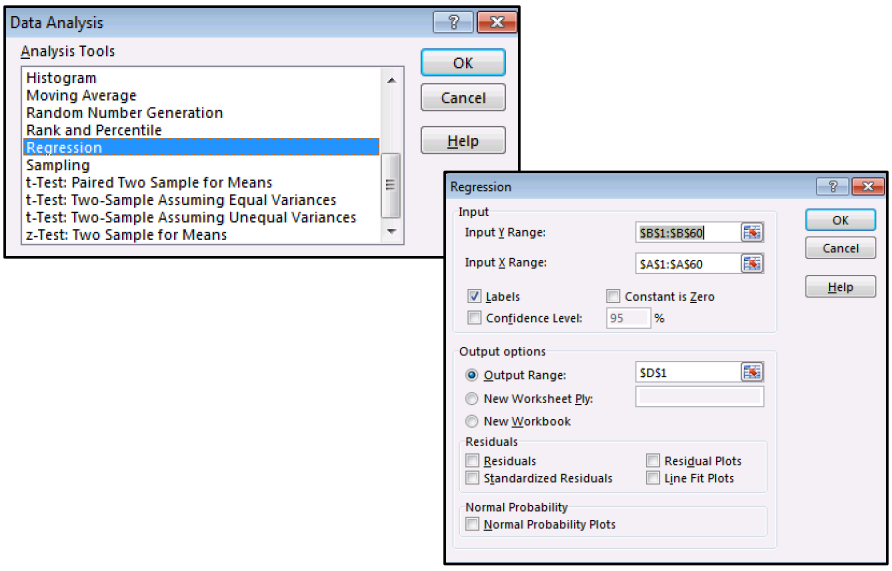
Chart Tools -> Design



Linear Regression



Linear Regression in Excel



Linear Regression 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 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 in Excel

	Coefficients
Intercept	0.07480414
MdHHIncE	1.8696E-06

Recycling Rate = 0.0000001869 * MedianIncome + 0.07480414

What would the predicted recycling rate be for a community district with a median household income of $70,000?

0.0000001869 * 70000 + 0.07480414 = 0.0879

WE'VE NOW CREATED A MODEL TO MAKE PREDICTIONS!

They just may not be very good predictions…

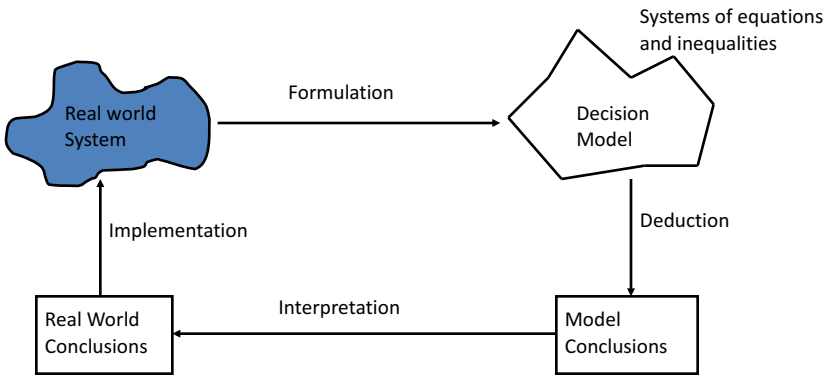
Making Decisions in a Resource Constrained World

- Types of constraints
  - Money
  - Time
  - Resources
  - Political Concerns
- Need ways to optimize around what's available

Decision Models

- *Decision modeling* refers to the use of mathematical or scientific methods to determine an allocation of scarce resources that improves or optimizes the performance of a system.
- The terms *operations research* and *management science* are also used to refer to decision modeling.

Decision Modeling Process



Requirements of DM Process

- You have to understand the real world process
- You have to be able to quantify the real world process
- You need to test your assumptions
- The decisions made based on the model will have an impact that need to be accounted for in the future

OPTIMIZING PARKING TICKET REVENUE

Constraints

- Density of illegally parked vehicles varies by location
- Number of illegally parked vehicles varies by location
- Amount of fine varies by location
- Number of agents is limited
- Only so many tickets an agent can write in a day
- Only so many tickets are actually paid
- Some neighborhoods are more concerned about illegal parking than others
- Every borough must have at least one ticket agent

Constraints

Location	Density	No. of Illegally Parked Vehicles	Fine/Ticket	% Collect	Min # of Agents	Max # of Agents
Manhattan	40%	24,000	$90	75%	200	600
Bronx	20%	4,000	$40	80%	50	200
Brooklyn	15%	9,000	$40	80%	50	300
Queens	10%	6,000	$40	90%	50	300
Staten Island	10%	2,000	$40	95%	25	100

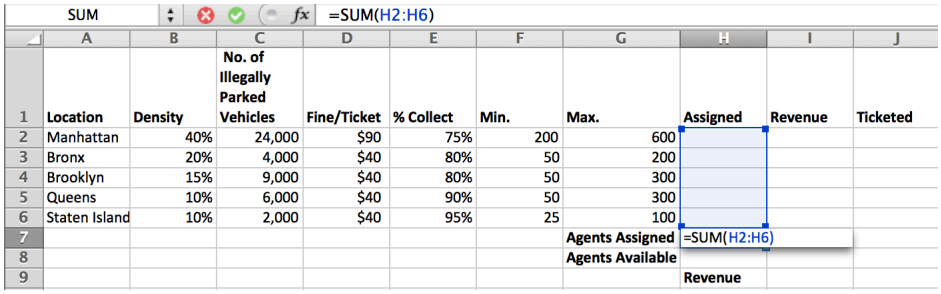
NOW LET'S OPTIMIZE!

Open decision_model.xlsx

Set up Spreadsheet

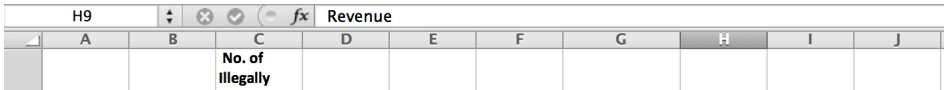
	A	B	C	D	E	F	G	H	I	J
1	**Location**	**Density**	**No. of Illegally Parked Vehicles**	**Fine/Ticket**	**% Collect**	**Min.**	**Max.**	**Assigned**	**Revenue**	**Ticketed**
2	Manhattan	40%	24,000	$90	75%	200	600			
3	Bronx	20%	4,000	$40	80%	50	200			
4	Brooklyn	15%	9,000	$40	80%	50	300			
5	Queens	10%	6,000	$40	90%	50	300			
6	Staten Island	10%	2,000	$40	95%	25	100			
7							**Agents Assigned**			
8							**Agents Available**			
9								**Revenue**		

Sum Assigned Agents



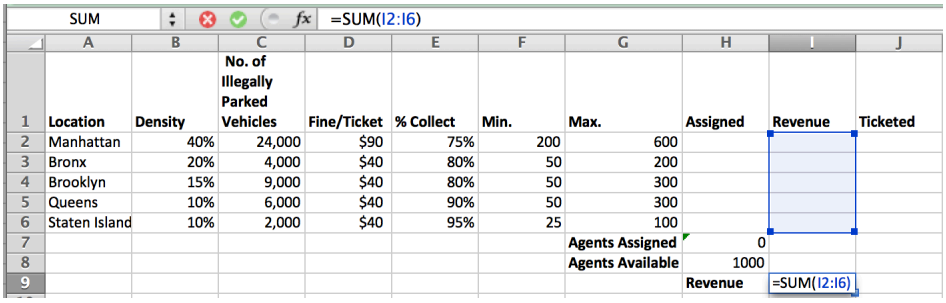
	A	B	C	D	E	F	G	H	I	J
1	Location	Density	No. of Illegally Parked Vehicles	Fine/Ticket	% Collect	Min.	Max.	Assigned	Revenue	Ticketed
2	Manhattan	40%	24,000	$90	75%	200	600			
3	Bronx	20%	4,000	$40	80%	50	200			
4	Brooklyn	15%	9,000	$40	80%	50	300			
5	Queens	10%	6,000	$40	90%	50	300			
6	Staten Island	10%	2,000	$40	95%	25	100			
7							Agents Assigned	=SUM(H2:H6)		
8							Agents Available			
9								Revenue		

Add Total Number of Agents (1000)



	A	B	C	D	E	F	G	H	I	J
1	Location	Density	No. of Illegally Parked Vehicles	Fine/Ticket	% Collect	Min.	Max.	Assigned	Revenue	Ticketed
2	Manhattan	40%	24,000	$90	75%	200	600			
3	Bronx	20%	4,000	$40	80%	50	200			
4	Brooklyn	15%	9,000	$40	80%	50	300			
5	Queens	10%	6,000	$40	90%	50	300			
6	Staten Island	10%	2,000	$40	95%	25	100			
7							Agents Assigned	0		
8							Agents Available	1000		
9								Revenue		

Sum Revenue



SUM =SUM(I2:I6)

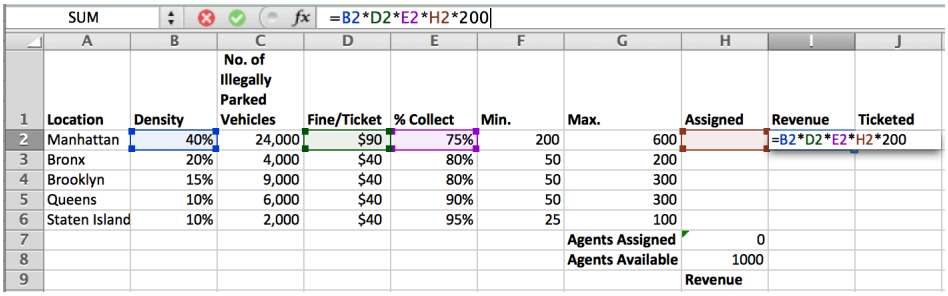
	A	B	C	D	E	F	G	H	I	J
1	Location	Density	No. of Illegally Parked Vehicles	Fine/Ticket	% Collect	Min.	Max.	Assigned	Revenue	Ticketed
2	Manhattan	40%	24,000	$90	75%	200	600			
3	Bronx	20%	4,000	$40	80%	50	200			
4	Brooklyn	15%	9,000	$40	80%	50	300			
5	Queens	10%	6,000	$40	90%	50	300			
6	Staten Island	10%	2,000	$40	95%	25	100			
7							Agents Assigned	0		
8							Agents Available	1000		
9								Revenue	=SUM(I2:I6)	

Objective Revenue Function

Revenue = # of Agents * 200 * Fine * % Collect * Density of Illegally Parked Cars

- Example:
  - 100 Agents in Manhattan
  - Fine is $90
  - They collect 75% of fines
  - The density of illegally parked cars is 40%
    - Max number of tickets an agent can write is 200 * 40% or 80 tickets
  - Revenue = 100 *200 * 0.4 * 90 * 0.75 = $540,000

Add Revenue Function



	A	B	C	D	E	F	G	H	I	J
1	Location	Density	No. of Illegally Parked Vehicles	Fine/Ticket	% Collect	Min.	Max.	Assigned	Revenue	Ticketed
2	Manhattan	40%	24,000	$90	75%	200	600		=B2*D2*E2*H2*200	
3	Bronx	20%	4,000	$40	80%	50	200			
4	Brooklyn	15%	9,000	$40	80%	50	300			
5	Queens	10%	6,000	$40	90%	50	300			
6	Staten Island	10%	2,000	$40	95%	25	100			
7							Agents Assigned	0		
8							Agents Available	1000		
9								Revenue		

Objective Ticketing Function

Ticketed = # of Agents * 200 * Density of Illegally Parked Cars

- Example
  - 100 Agents in Manhattan
  - Maximum number of tickets is 200
  - Density of illegally parked cars is 40%
  - # of expected tickets = 100 * 200 * 0.4 = 8,000

Adding Ticketing Function

SUM =H2*200*B2

	A	B	C	D	E	F	G	H	I	J
1	Location	Density	No. of Illegally Parked Vehicles	Fine/Ticket	% Collect	Min.	Max.	Assigned	Revenue	Ticketed
2	Manhattan	40%	24,000	$90	75%	200	600		$0	=H2*200*B2
3	Bronx	20%	4,000	$40	80%	50	200		$0	
4	Brooklyn	15%	9,000	$40	80%	50	300		$0	
5	Queens	10%	6,000	$40	90%	50	300		$0	
6	Staten Island	10%	2,000	$40	95%	25	100		$0	
7							Agents Assigned	0		
8							Agents Available	1000		
9								Revenue	$0	

Spreadsheet Setup

	A	B	C	D	E	F	G	H	I	J
1	Location	Density	No. of Illegally Parked Vehicles	Fine/Ticket	% Collect	Min.	Max.	Assigned	Revenue	Ticketed
2	Manhattan	40%	24,000	$90	75%	200	600		$0	0
3	Bronx	20%	4,000	$40	80%	50	200		$0	0
4	Brooklyn	15%	9,000	$40	80%	50	300		$0	0
5	Queens	10%	6,000	$40	90%	50	300		$0	0
6	Staten Island	10%	2,000	$40	95%	25	100		$0	0
7							Agents Assigned	0		
8							Agents Available	1000		
9								Revenue	$0	

NOW WE'RE READY TO OPTIMIZE THE NUMBER OF ASSIGNED AGENTS

Spreadsheet Setup

	A	B	C	D	E	F	G	H	I	J
1	Location	Density	No. of Illegally Parked Vehicles	Fine/Ticket	% Collect	Min.	Max.	Assigned	Revenue	Ticketed
2	Manhattan	40%	24,000	$90	75%	200	600		$0	0
3	Bronx	20%	4,000	$40	80%	50	200		$0	0
4	Brooklyn	15%	9,000	$40	80%	50	300		$0	0
5	Queens	10%	6,000	$40	90%	50	300		$0	0
6	Staten Island	10%	2,000	$40	95%	25	100		$0	0
7							Agents Assigned	0		
8							Agents Available	1000		
9								Revenue	$0	

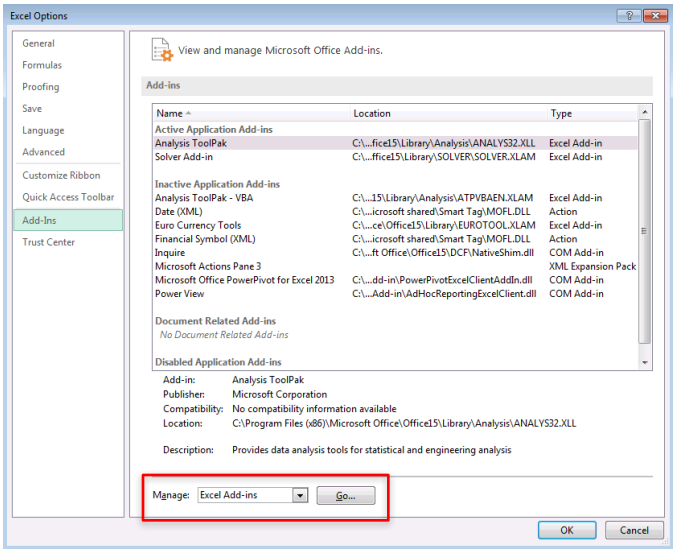
This is what Excel is going to optimize for us

Constraints

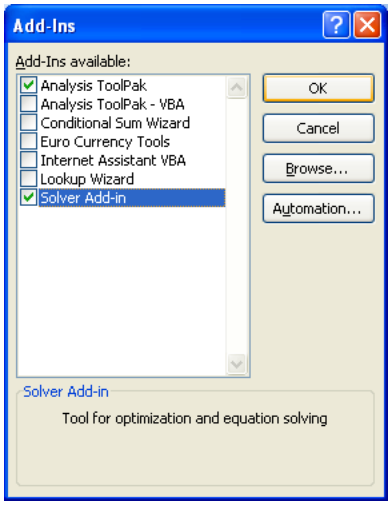
- Number of assigned agents must be greater than or equal to minimum but less than or equal to the maximum
- The number of tickets can't exceed the estimated number of illegally parked cars
- The total number of assigned agents must be less than or equal to 1000

Installing Solver

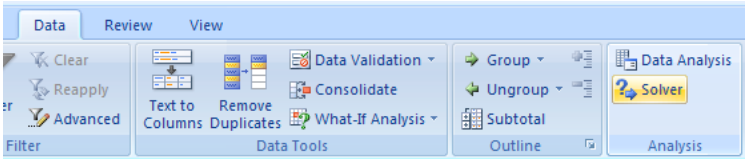
- File
- Options
- Add-ins
- Manage
- "Go…"



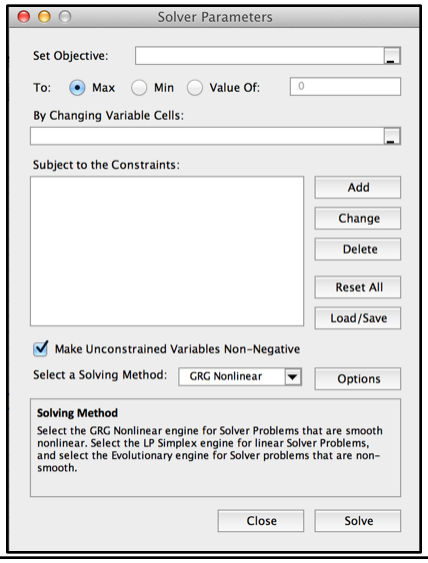
Installing Solver



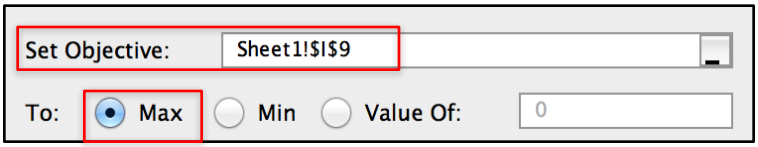
Using Solver



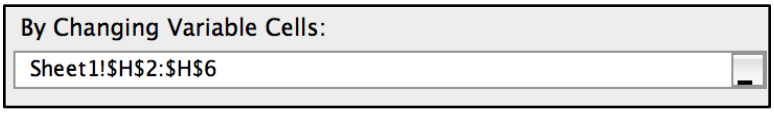
Configuring Solver



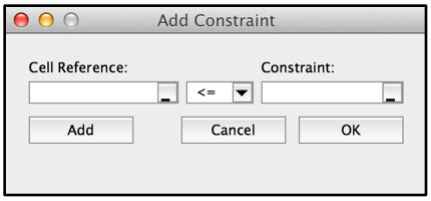
Set Objective (Maximize Revenue)

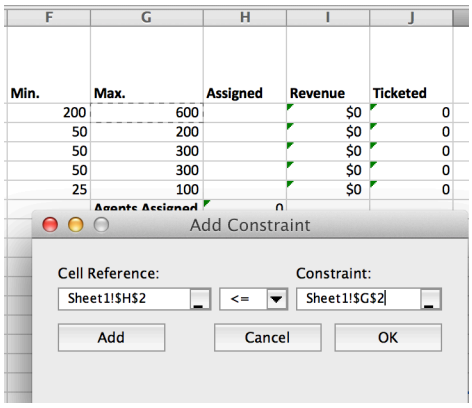


Set Variable Cells (To Be Optimized)



Add Constraints





Constraints

- Number of assigned agents must be greater than or equal to minimum but less than or equal to the maximum

Subject to the Constraints:

```
$H$2 <= $G$2
$H$2 >= $F$2
$H$3 <= $G$3
$H$3 >= $F$3
$H$4 <= $G$4
$H$4 >= $G$4
$H$5 <= $G$5
$H$5 >= $F$5
$H$6 <= $G$6
$H$6 >= $F$6
```

Constraints

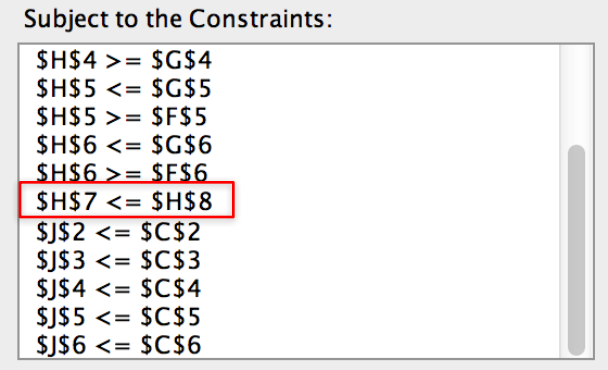
- The number of tickets can't exceed the estimated number of illegally parked cars

Subject to the Constraints:

```
$H$4 <= $G$4
$H$4 >= $G$4
$H$5 <= $G$5
$H$5 >= $F$5
$H$6 <= $G$6
$H$6 >= $F$6
$J$2 <= $C$2
$J$3 <= $C$3
$J$4 <= $C$4
$J$5 <= $C$5
$J$6 <= $C$6
```

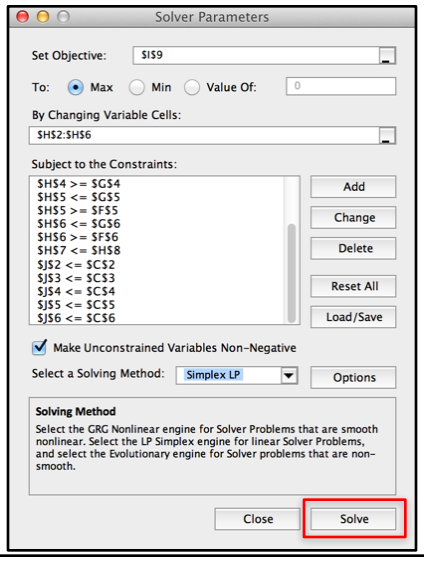
Constraints

- The total number of assigned agents must be less than or equal to 1000

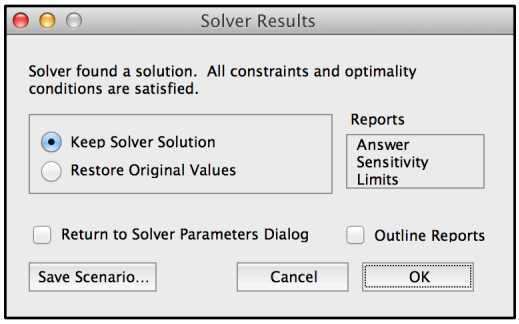


Set Solving Method

Check and then Click “Solve”



Results



Results

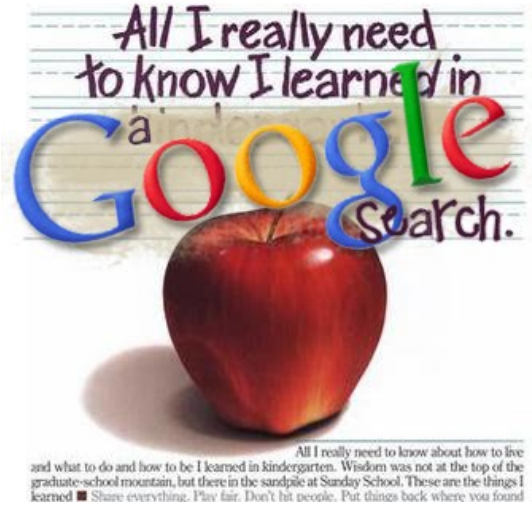
	A	B	C	D	E	F	G	H	I	J
1	**Location**	**Density**	**No. of Illegally Parked Vehicles**	**Fine/Ticket**	**% Collect**	**Min.**	**Max.**	**Assigned**	**Revenue**	**Ticketed**
2	Manhattan	40%	24,000	$90	75%	200	600	300	$1,620,000	24000
3	Bronx	20%	4,000	$40	80%	50	200	100	$128,000	4000
4	Brooklyn	15%	9,000	$40	80%	50	300	300	$288,000	9000
5	Queens	10%	6,000	$40	90%	50	300	200	$144,000	4000
6	Staten Island	10%	2,000	$40	95%	25	100	100	$76,000	2000
7							**Agents Assigned**	1000		
8							**Agents Available**	1000		
9								**Revenue**	$2,256,000	

Goals for the Course

- Learn common statistical measures, including mean, median, mode, standard deviation, and variance
- Calculate correlation coefficients for bivariate data and apply the technique of simple regression analysis
- Demonstrate techniques used for forecasting
- Communicate data meaningfully to a broad audience using charts and graphs in Microsoft Excel

Key Takeaways for the Course

- You will be familiar with common statistical measures
- You will be able to calculate correlation coefficients for bivariate data and perform simple linear regression analysis
- You will be familiar with the basic techniques of forecasting
- You will be better able to communicate analysis using charts and graphs in Microsoft Excel



Resources

Online Resources

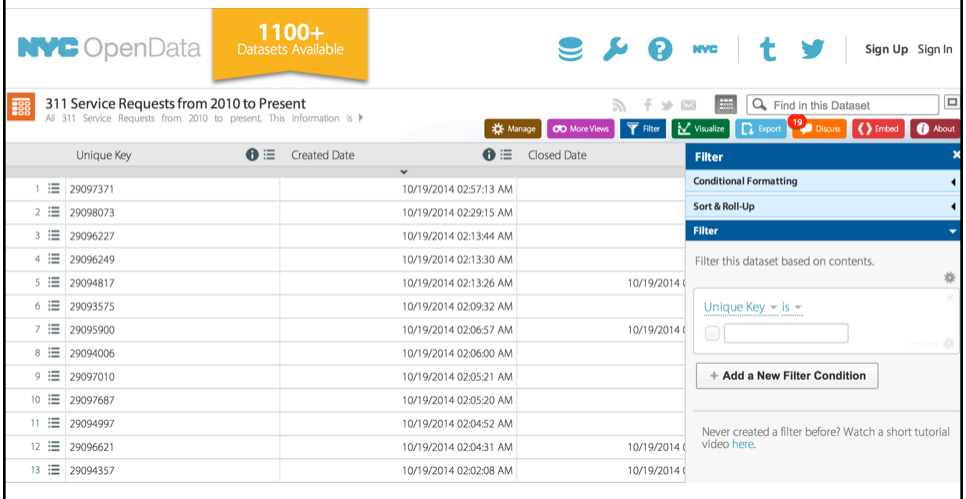
- http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3096219/
- https://www.udacity.com/course/st101
- https://www.khanacademy.org/math/probability

Technical Resources

- Stack Overflow
  - http://stackoverflow.com/
  - One of the best Q&A sites for technical questions of all kinds
- Microsoft Office Support
  - http://office.microsoft.com/en-us/support/
  - Documentation on various MS Office products
- Excel Tips
  - http://excel.tips.net/
  - Various tips and tricks for using Excel

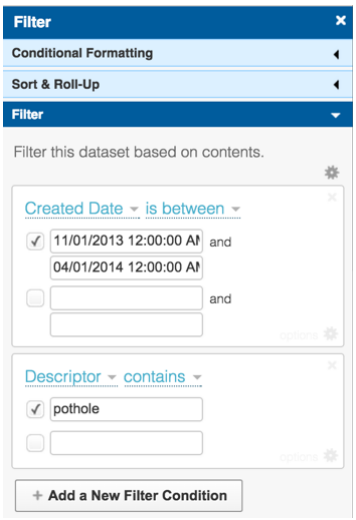
SOURCES

NYC Open Data Portal – 311 Data

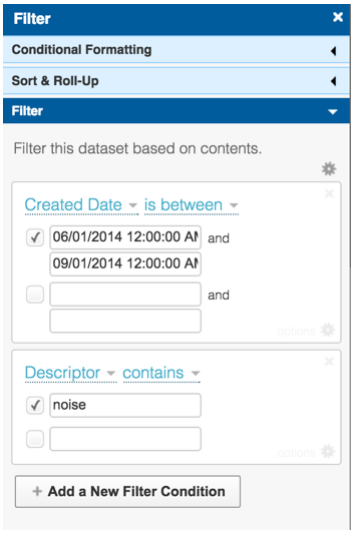


https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

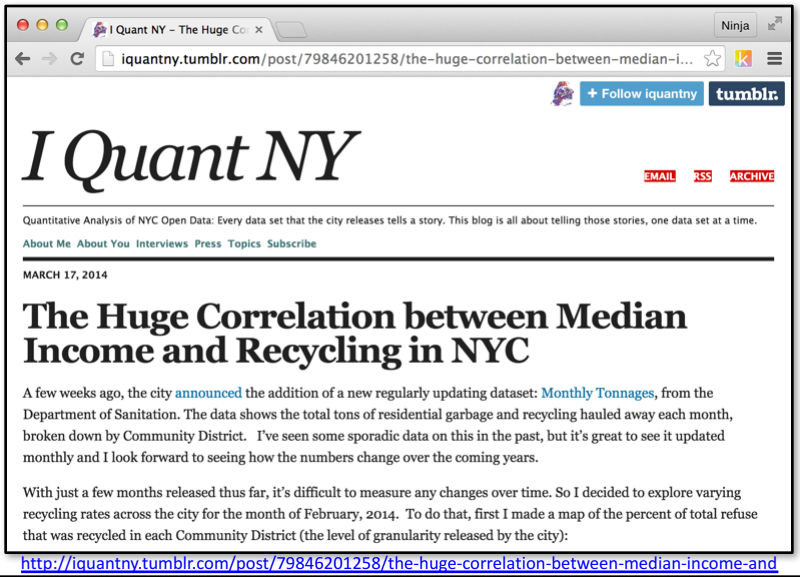
Potholes



Noise Complaints



Median Income and Recycling



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

Contact Information for MODA

Mayor's Office of Data Analytics

- Lindsay Mollineaux
  - Email: lmollineaux@cityhall.nyc.gov
- MODA website - http://www.nyc.gov/html/analytics/html/home/home.shtml

Contact Information

Instructor

- Name: Richard Dunks
- Email: richard@datapolitan.com
- Website: http://www.datapolitan.com
- Blog: http://blog.datapolitan.com
- Twitter: @rdunks1/@datapolitan

Functions List

- =AVERAGE: Calculates the arithmetic mean for a range of numbers
- =MEDIAN: Calculates the median for a range of numbers
- =MODE: Calculates the mode for a range of numbers
- =MAX and =MIN: Calculates the maximum and minimum number for a range of numbers
- =QUARTILE: Calculate quartiles for a range of numbers
- =CORREL: Calculates the coefficient of correlation between two ranges of numbers