This Week

Monday

- Data and Summary Measures

Wednesday

- Lab Exercise: Acme Order Analysis (Part II)

Topics

- **Types of Data**
- **Summary Measures (or “Descriptive Statistics”)**
- **Relaying Quantitative Information**

A Problem Solving Framework

1. *Define the Problem*
2. *Collect and Organize Data*
3. *Characterize Uncertainty and Data Relationships*
4. *Build an Evaluation Model*
5. *Formulate a Solution Approach*
6. *Evaluate Potential Solutions*
7. *Recommend a Course of Action*

Types of Data

- **Numerical** vs. **Categorical**
- **Discrete** vs. **Continuous**
- **Cross-sectional** vs. **Time series**

Types of Data

- **Numerical** vs. **Categorical**:
  - Can you ***meaningfully perform arithmetic*** with the data?
  - Examples:
    - Duration of cell phone calls (numeric)

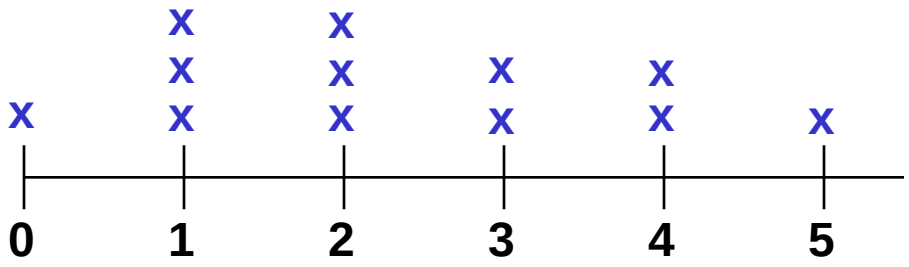
      0.75 minutes, 1.46 minutes, 2.39 minutes, …
    - Cellular phone service provider (categorical)

      AT&T, Verizon, T-Mobile, …
    - Performance Rating 1=Best to 5=Worst (??)

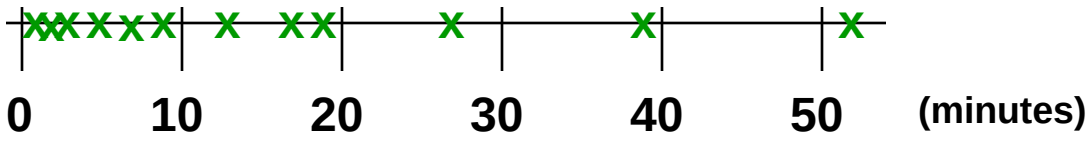
      1, 1, 3, 4, 2, 3, 1, 4, 2, 2, 2, 3, 4, 1, …

Types of Data

- **Discrete** vs. **Continuous**:
  - Do the data take on a ***countable number*** of values, or values in a ***continuous range***?
  - Examples:
    - Number of people in family with cell phones (discrete)

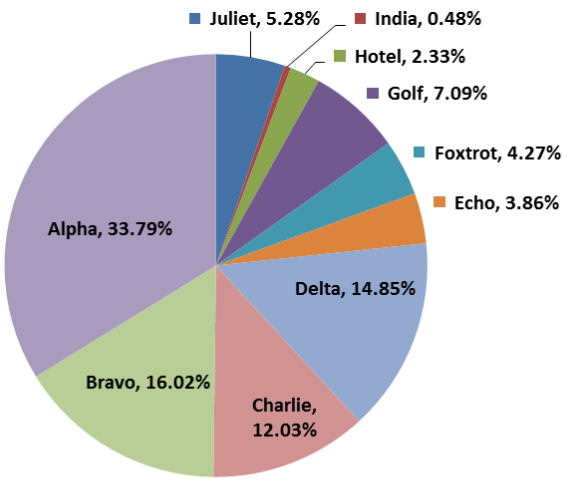


    - Duration of cell phone calls (continuous)

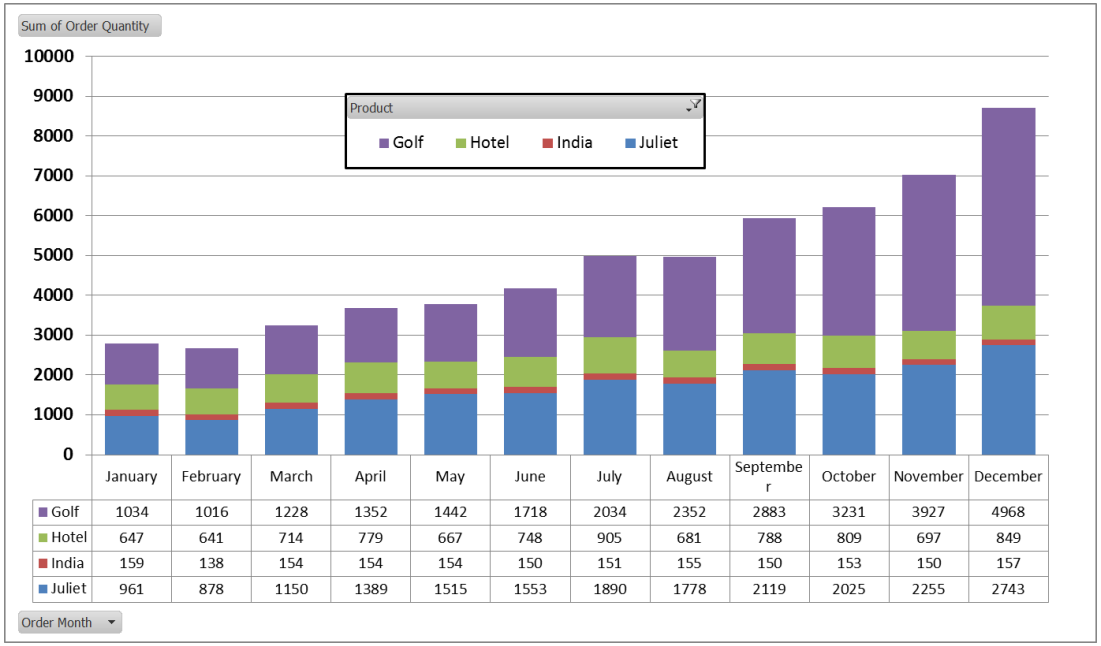


Types of Data

- **Cross-sectional** vs. **Time series**
  - Are the data shown for a ***single time period*** or are ***many periods of time distinguished***?



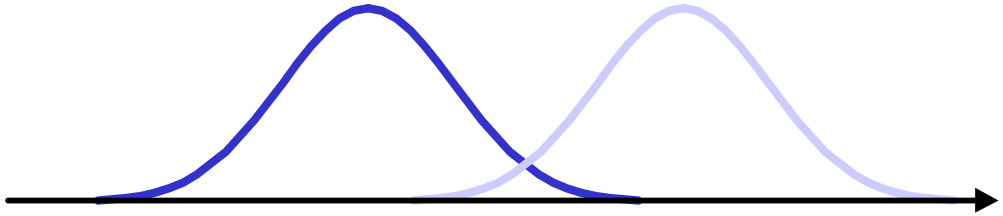
2011 *CallSign* Orders



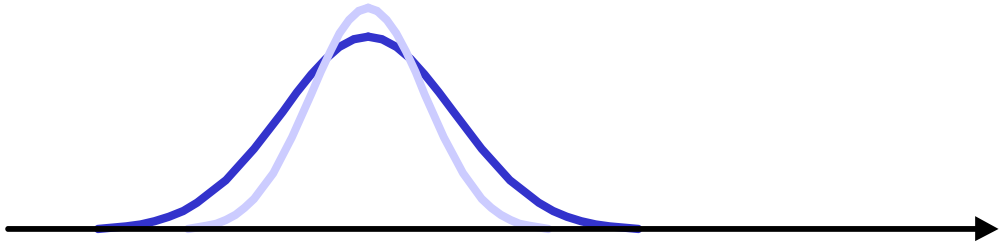
	January	February	March	April	May	June	July	August	September	October	November	December
Golf	1034	1016	1228	1352	1442	1718	2034	2352	2883	3231	3927	4968
Hotel	647	641	714	779	667	748	905	681	788	809	697	849
India	159	138	154	154	154	150	151	155	150	153	150	157
Juliet	961	878	1150	1389	1515	1553	1890	1778	2119	2025	2255	2743

Summary Measures

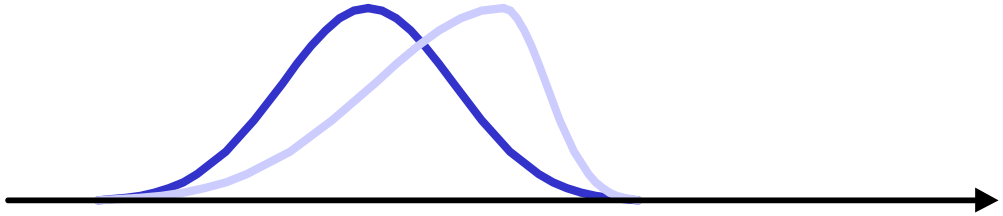
Central Tendency



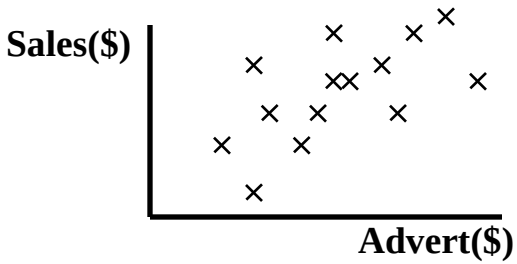
Dispersion



Shape



Relationships



Measures of Central Tendency

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	$\sum$
Demand	10	10	10	10	11	11	11	12	13	13	14	15	17	18	20	195

- **Mean**: The ***average*** value = $\overline{X} = \sum_{i=1}^{n} X_i / n$ = 195/15 = **13**.

  Notation: $\overline{X}$ denotes a ***sample mean*** .

  For an entire population, μ is used to denote the ***population mean***.

  Excel: =AVERAGE(data_range)

- **Median**: The ***middle*** value when the data are arranged in increasing order. Here, the median is the 8th value (of 15) = **12**. If there were 16 values, the median would be the average of the 8th and 9th (ordered) values. Excel: =MEDIAN(data_range)

- **Mode**: The ***most frequently occurring*** value in the data.

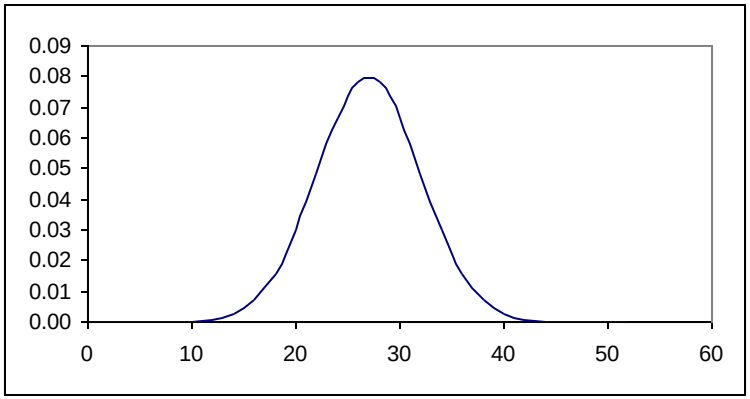
  Here, the mode = **10**. Excel: =MODE.SNGL(data_range)

Central Tendency and Shape

- For ***symmetrical*** data:

  mean = median = mode

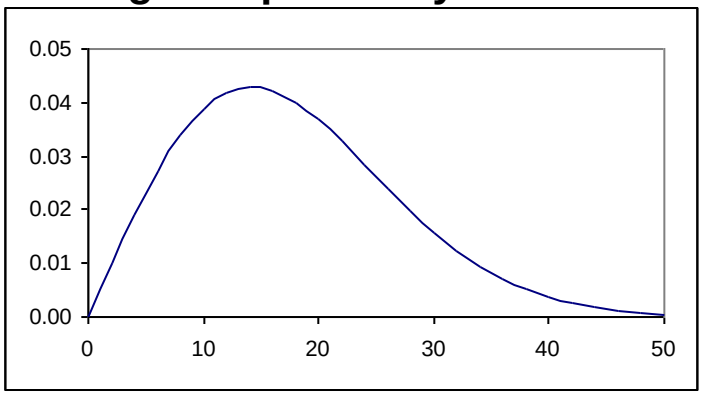
  Always *Usually*



- For ***non-symmetrical*** data:

Excel: =SKEW(data_range)

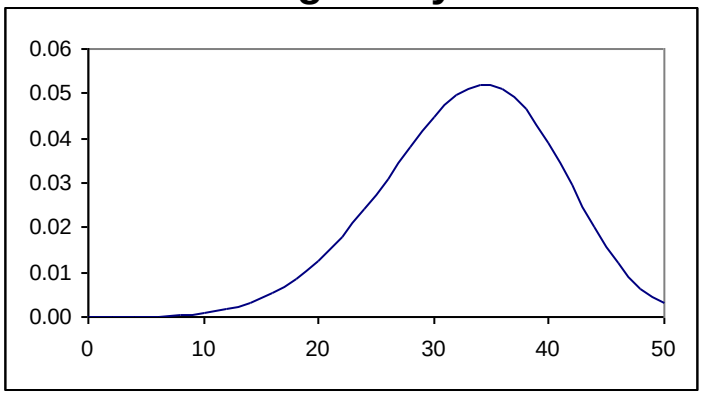
Right or positively-skewed



"mode < median < mean"

(Usually)

Left or negatively-skewed



"mean < median < mode"

(Usually)

Measures of Dispersion

Statistics based on ***deviations from the mean***:

- **Variance**: **Average squared deviation from the mean.**

Population

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}$$

Sample

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Excel: =VAR.P(data_range) or =VAR.S(data_range)

- **Standard deviation**: σ (population) or *s* (sample)

Excel: =STDEV.P(data_range) or =STDEV.S(data_range)

- **Mean absolute deviation** (MAD):

$$\frac{\sum_{i=1}^{n}\left|X_i - \bar{X}\right|}{n}$$

Excel: =AVEDEV(data_range)

Measures of Dispersion

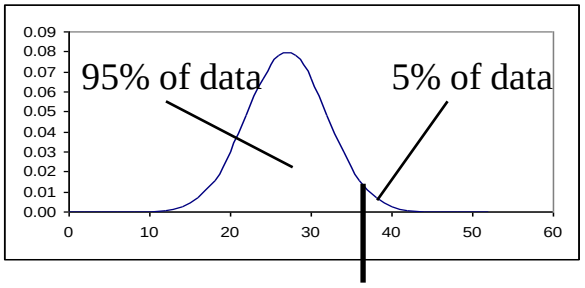
Other statistical measures of dispersion:

- **Range** = $X_{largest} - X_{smallest}$ Excel: =MAX(data_range) - MIN(data_range)

  Excel: =LARGE(data_range,*n*) returns the *n*th largest number

- **Percentiles and Quartiles**:

  The ***k*th *percentile*** is a value such that *k*% of the data fall at or below the value and (100 - k)% of the measurements fall at or above the value.



37 = 95th percentile

Excel: =PERCENTILE.EXC(data_range, *k%*)

Excel: =QUARTILE.EXC(data_range, *q*)

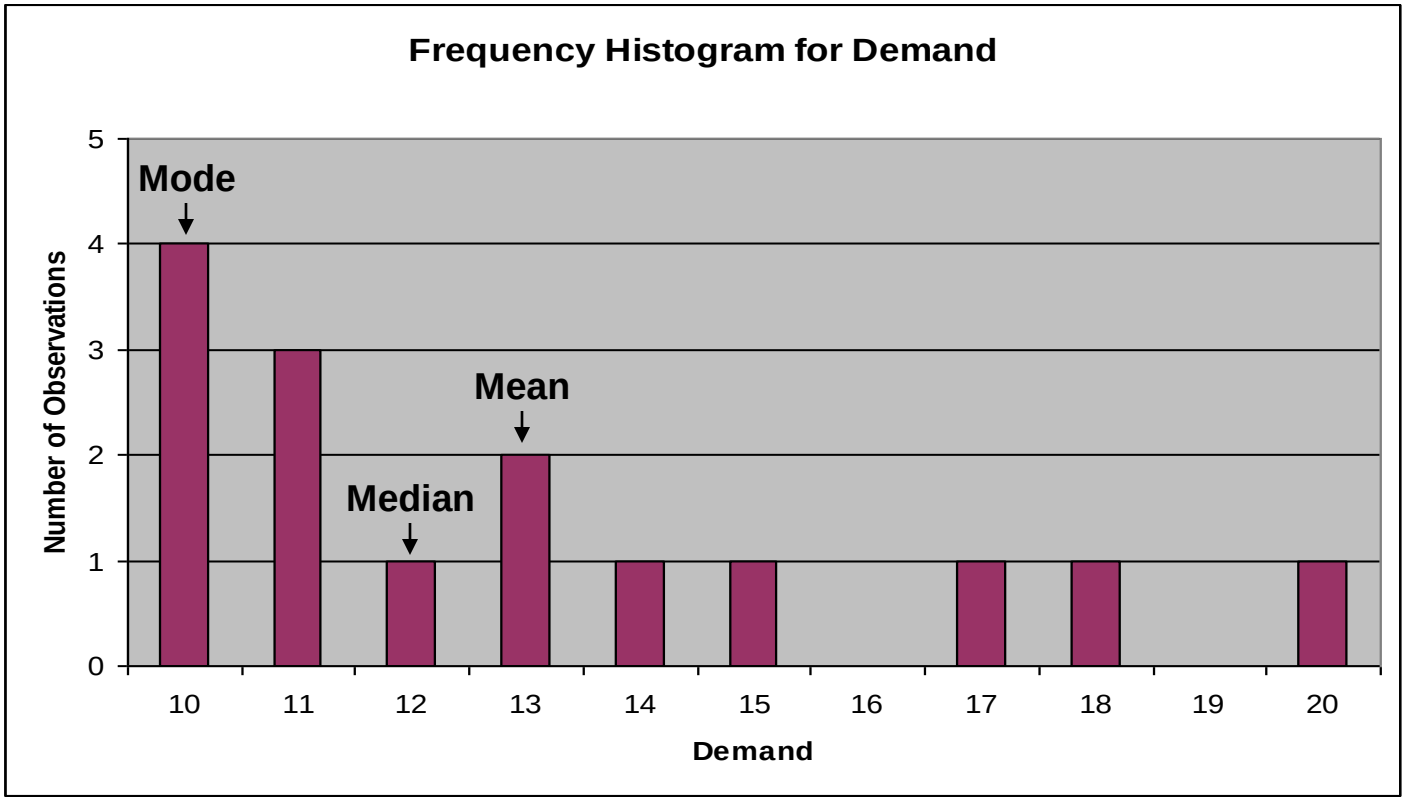
  - The ***first quartile (Q1)*** is the 25th percentile.
  - The ***second quartile*** (or median) is the 50th percentile.
  - The ***third quartile (Q3)*** is the 75th percentile.
  - The ***interquartile range (IQR)*** is Q3 - Q1.

Frequency Histogram

Demand Value	10	11	12	13	14	15	16	17	18	19	20
Frequency	4	3	1	2	1	1	0	1	1	0	1

Excel: =FREQUENCY(data_range, bin_range)

(an *array* function)



Relationships Between Variables

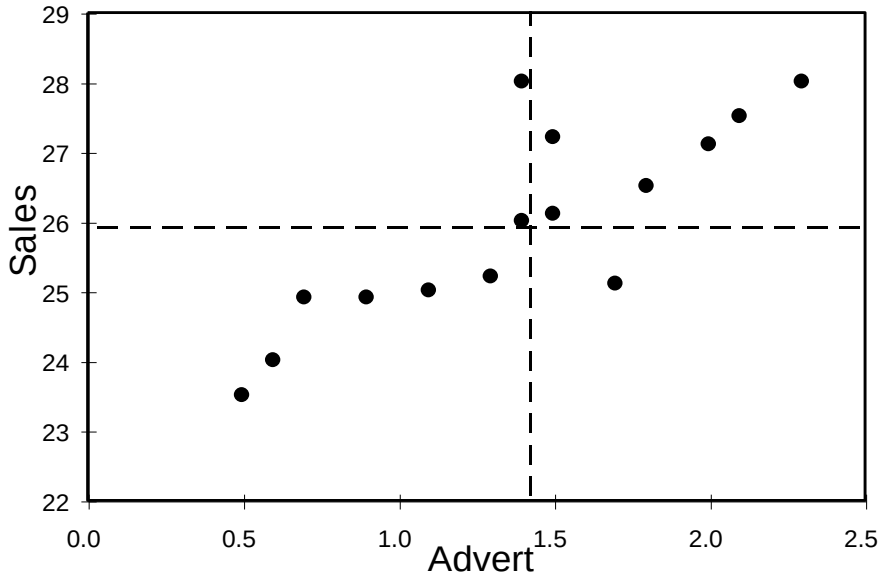
- ***Scatterplots*** are good visual tools for identifying data elements that may be related to one another:

X = (X_1, X_2, …, X_{15}) are Advertising dollars spent at 15 stores last quarter.

Y = (Y_1, Y_2, …, Y_{15}) are Sales dollars received at the same 15 stores last quarter.

Average advertising: $\overline{X}$ **= \$1.4M**

Average sales: $\overline{Y}$ **= \$25.9M**



- ***Covariance*** and ***correlation*** measure the ***strength of the linear relationship*** between two variables.

Covariance

The ***covariance*** statistic is typically defined as follows:

$$\text{cov}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

Excel: =COVARIANCE.S(X_data, Y_data) ***

*** **Note:** COVARIANCE.P has "*n*" in the denominator, NOT (*n* – 1).

- Covariance can be difficult to interpret because it ***depends on the scale of the data*.**
- Correlation is more commonly used in practice than covariance for descriptive purposes.

Correlation

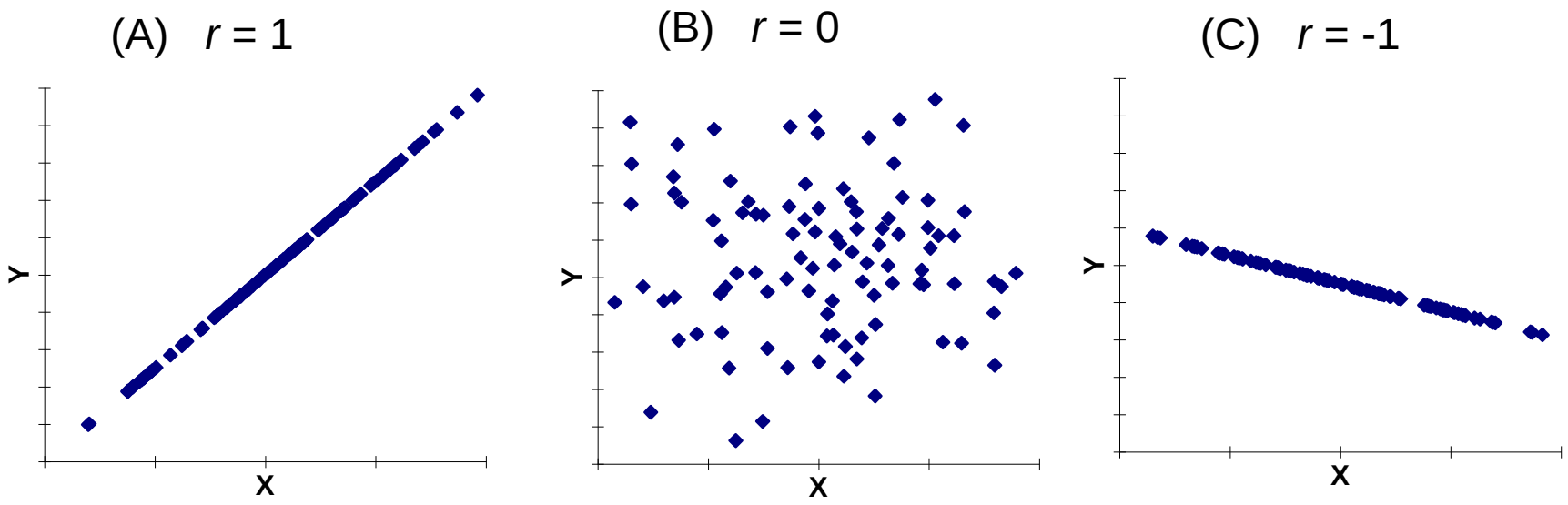
The ***correlation*** statistic (usually denoted ***r*** or ρ) is a ***standardized measure*** that is ***unitless*** :

$$\mathbf{corr}(\mathbf{X},\mathbf{Y}) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{(n-1)\,s_{\mathbf{X}}\,s_{\mathbf{Y}}} = \frac{\mathrm{cov}(\mathbf{X},\mathbf{Y})}{s_{\mathbf{X}}\,s_{\mathbf{Y}}}$$

Excel: =CORREL(X_data, Y_data)

- The ***maximum possible correlation is +1*** (perfect *positive* correlation).
- The ***minimum possible correlation is -1*** (perfect *negative* correlation).
- A correlation of ***zero*** implies that there is ***no discernable linear relationship*** between the data elements.

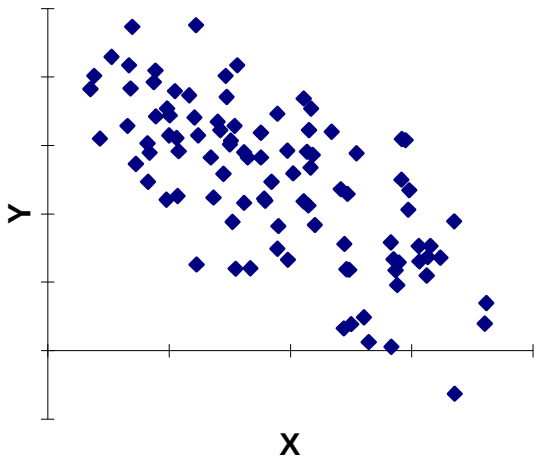
Some Examples of Correlation



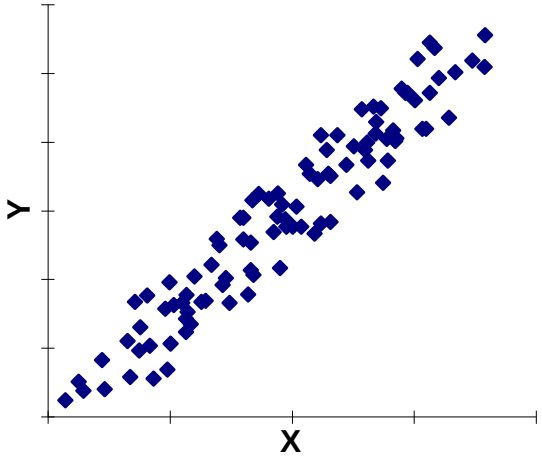
- If the plotted points resemble a cloud with no discernable linear pattern or slope, then the correlation will be near zero.
- The magnitude of the “slope” is a scale factor and does not affect the magnitude of the correlation (compare A and C above).

More Examples of Correlation

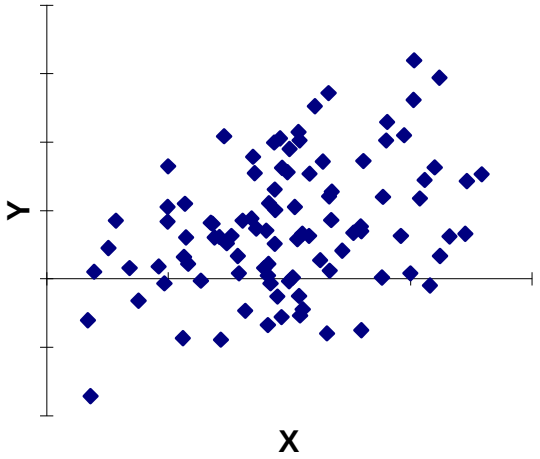
r = -0.7



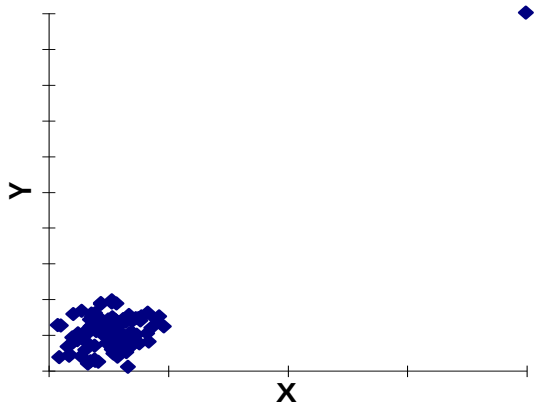
r = +0.9



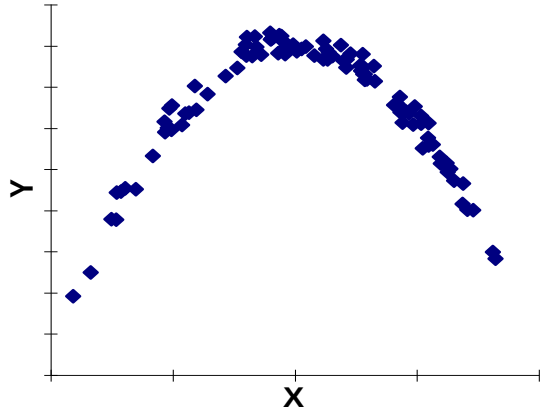
r = +0.4



r = 0.8 ?



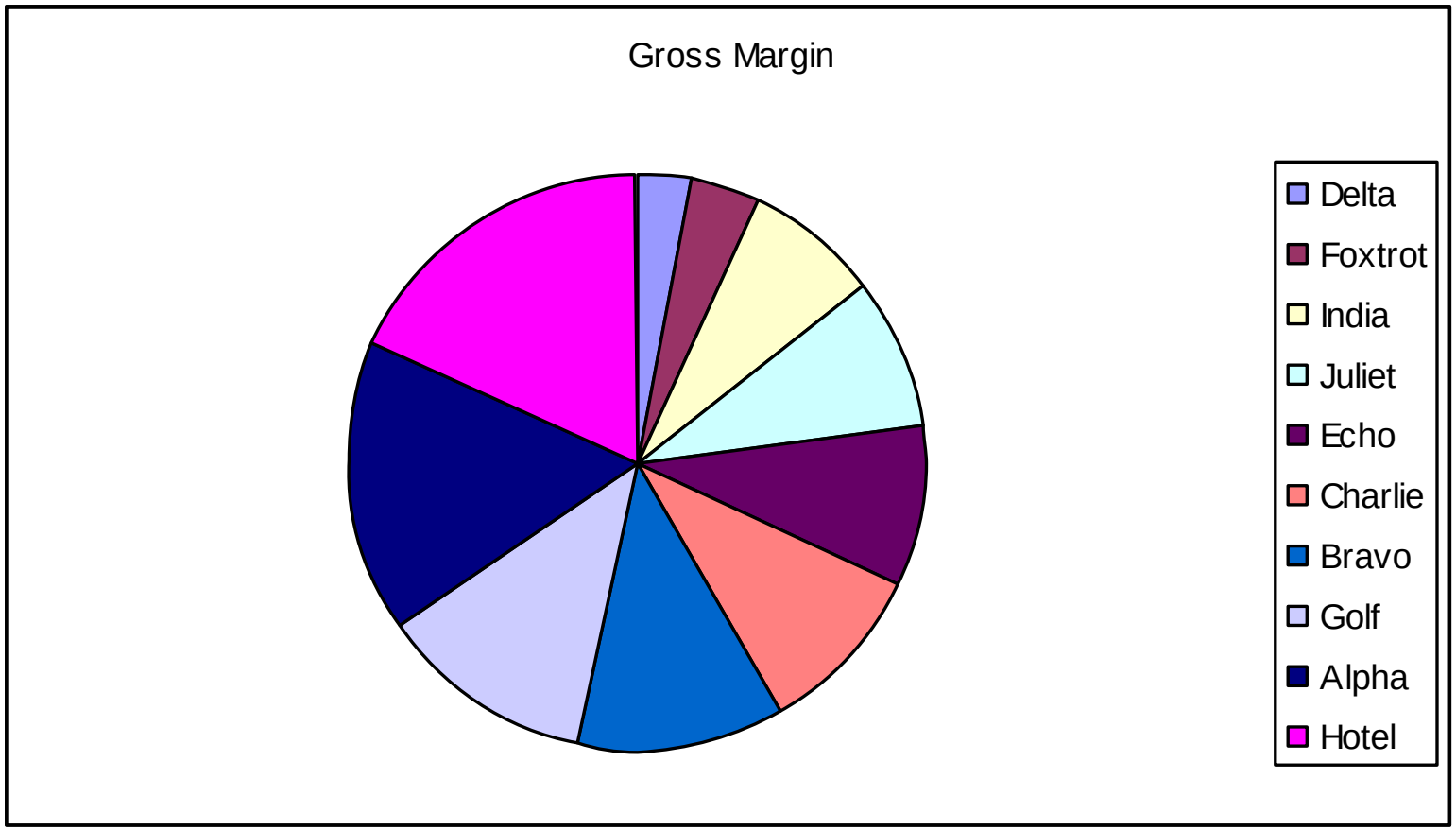
r = 0 ?



Relaying Quantitative Information

- **What *NOT* to do: Examples**
- **Helpful Questions**

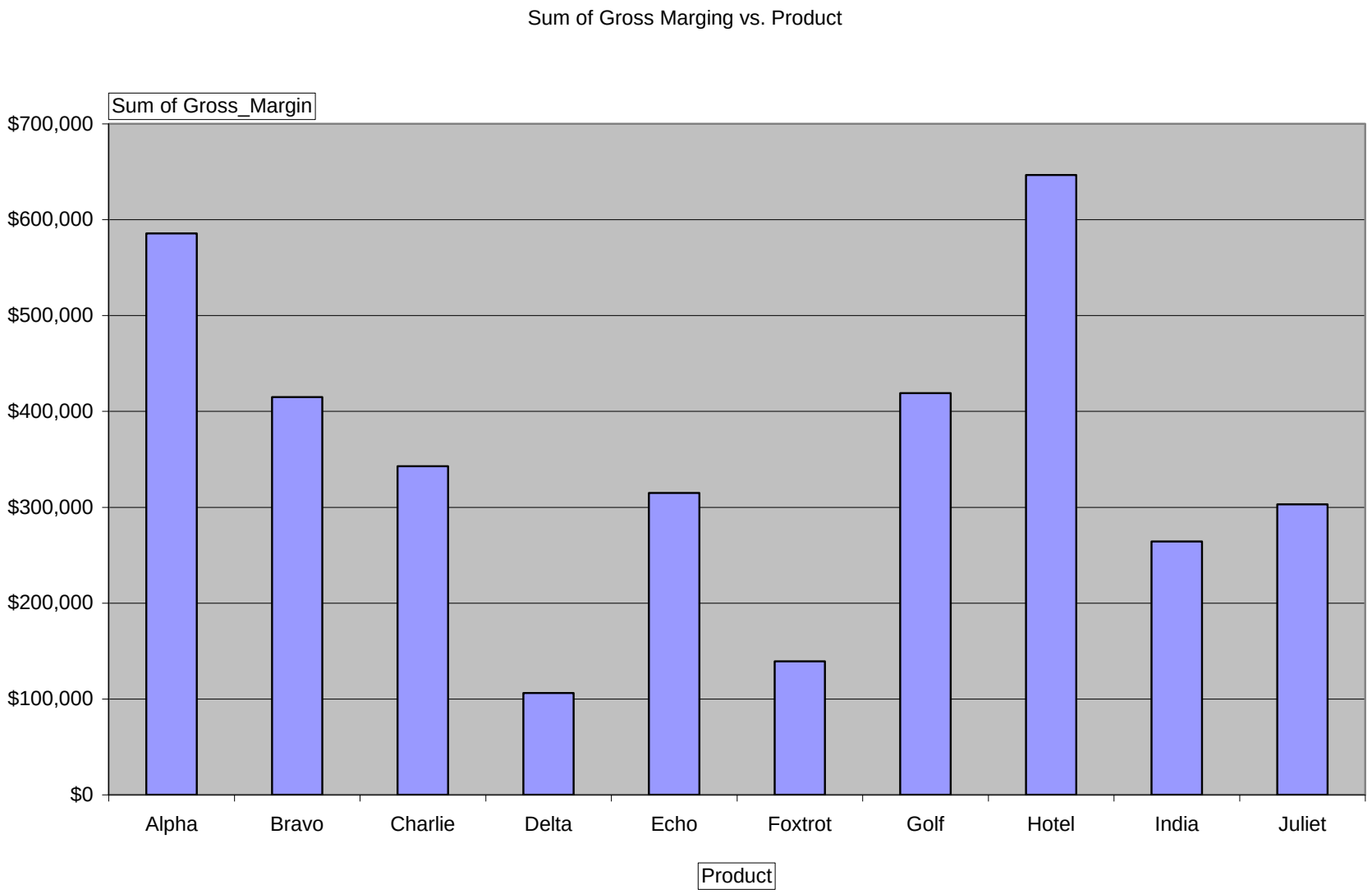
Bad Example #1



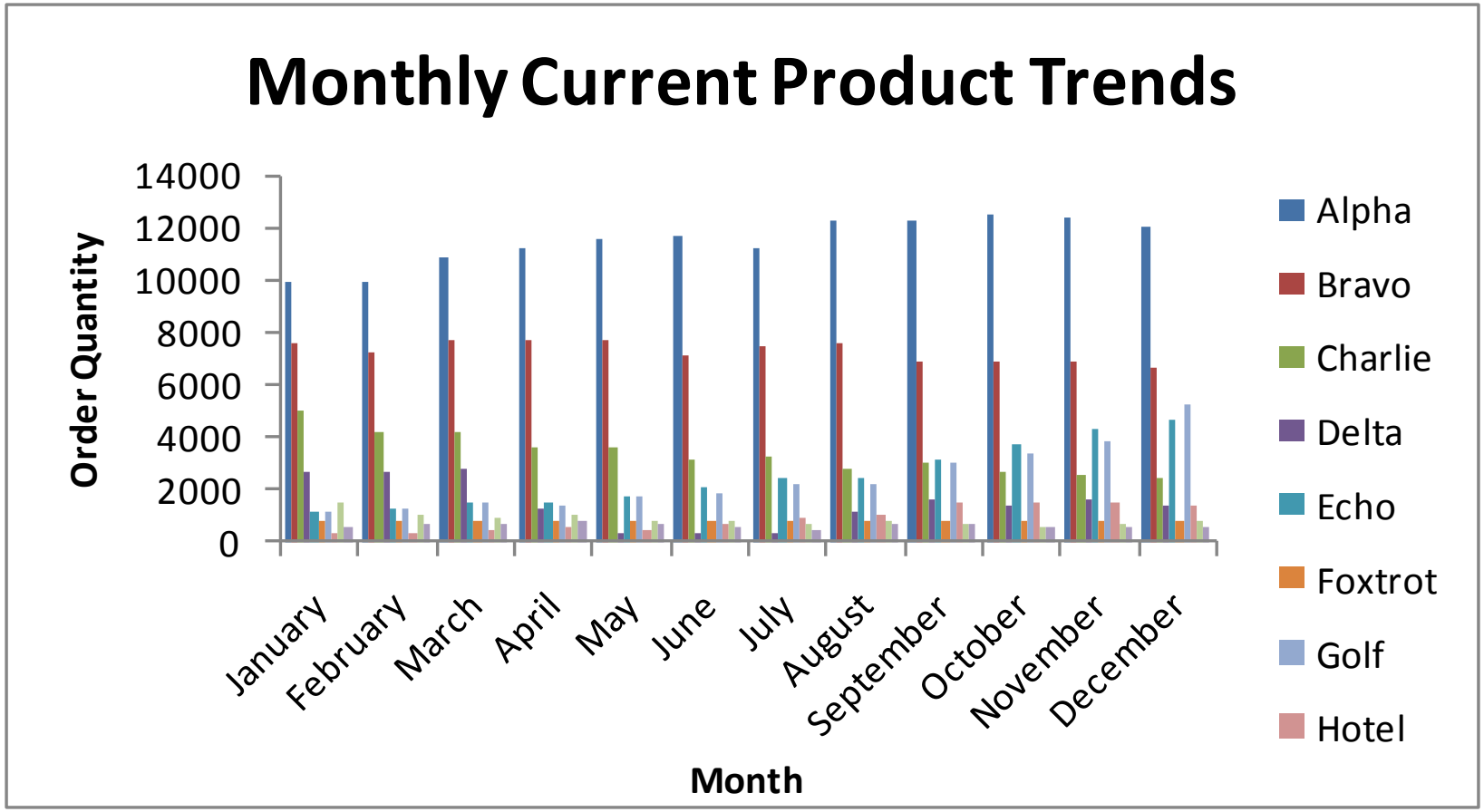
Bad Example #2

Product	Total Sales Revenue Attributed to Product
Alpha	$1,102,168.00.
Bravo	$829,910.50
Charlie	$564,732.00
Delta	$370,010.00
Echo	$753,511.50
Foxtrot	$318,112.00
Golf	$1,227,910.00
Hotel	$1,293,000.00
India	$528,700.00
Juliet	$568,050.00

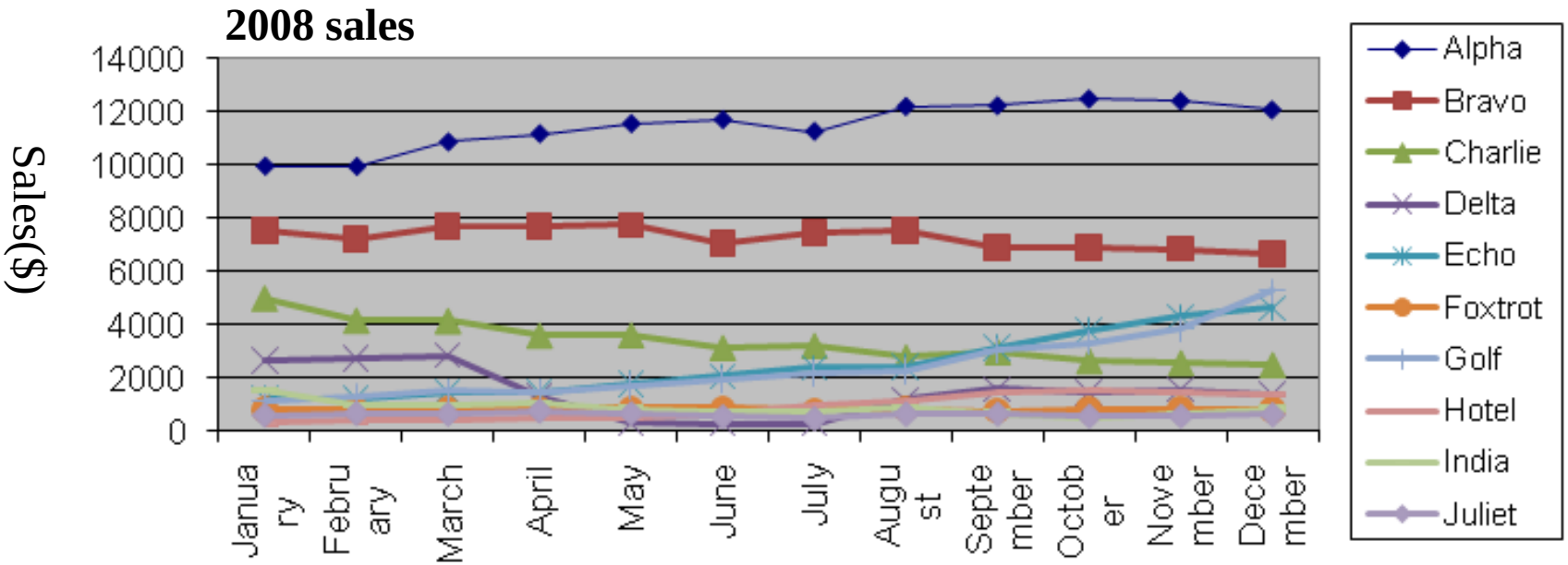
Bad Example #3



Bad Example #4



Bad Example #5

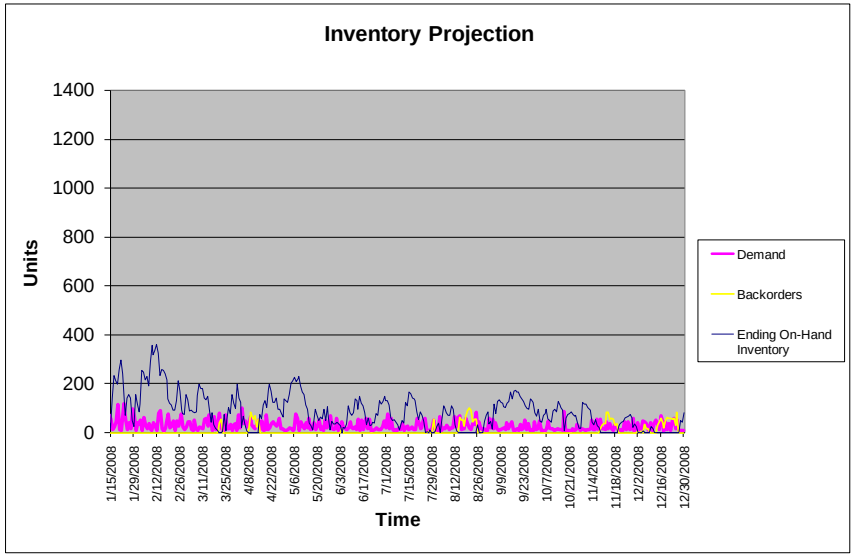


Bad Example #6

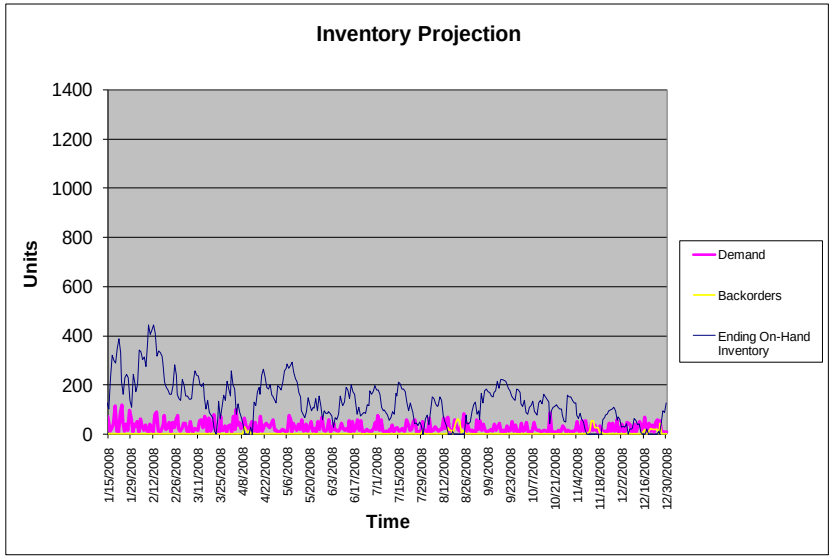
Gross Margin	Customer					
Order Month	**1**	**2**	**3**	**4**	**5**	**Total**
January	94942.25	56905.5	27500	31500.75	48127.5	258976
February	88914	52924.25	28261.25	24216	53082	247397.5
March	94813.5	56113.25	30745	21915	58044.5	261631.25
April	88286.75	51576.75	31970.75	20549.5	67925.5	260309.25
May	88216.25	49700	35518	19922.75	59259.5	252616.5
June	97503.75	42527.5	38598.75	22323.25	54095	255048.25
July	107279	45966.5	43966.5	21439.75	56522	275173.75
August	107758.5	49879.5	45518.5	22267.5	77413.5	302837.5
September	105078.25	53203.5	57223	21641	101850.5	338996.25
October	106989.25	51458.5	67128.5	25266	95590.5	346432.75
November	103484.75	56836.5	72609.5	26640.5	98160.5	357731.75
December	104633.25	54316.25	90556.5	29250.75	100333	379089.75
Total	1187899.5	621408	569596.25	286932.75	870404	**$3536240.5**

Bad Example #7

India Before



India After



Question *EVERY* piece of *EVERY* slide

- *What **point(s)** are you trying to make?*
- *How is the information **best displayed** (e.g., in a table, graph, bulleted text, or some combination)?*
- *Are the titles, axes, and data series **clearly labeled** and **properly formatted** (including units)?*
- *Is every element **necessary**?*
- *Is the **ordering** of the elements appropriate?*
- *Is the value of the information **worth the space** you are using?*
- *Is there a way to **efficiently relay more information** while maintaining clarity?*