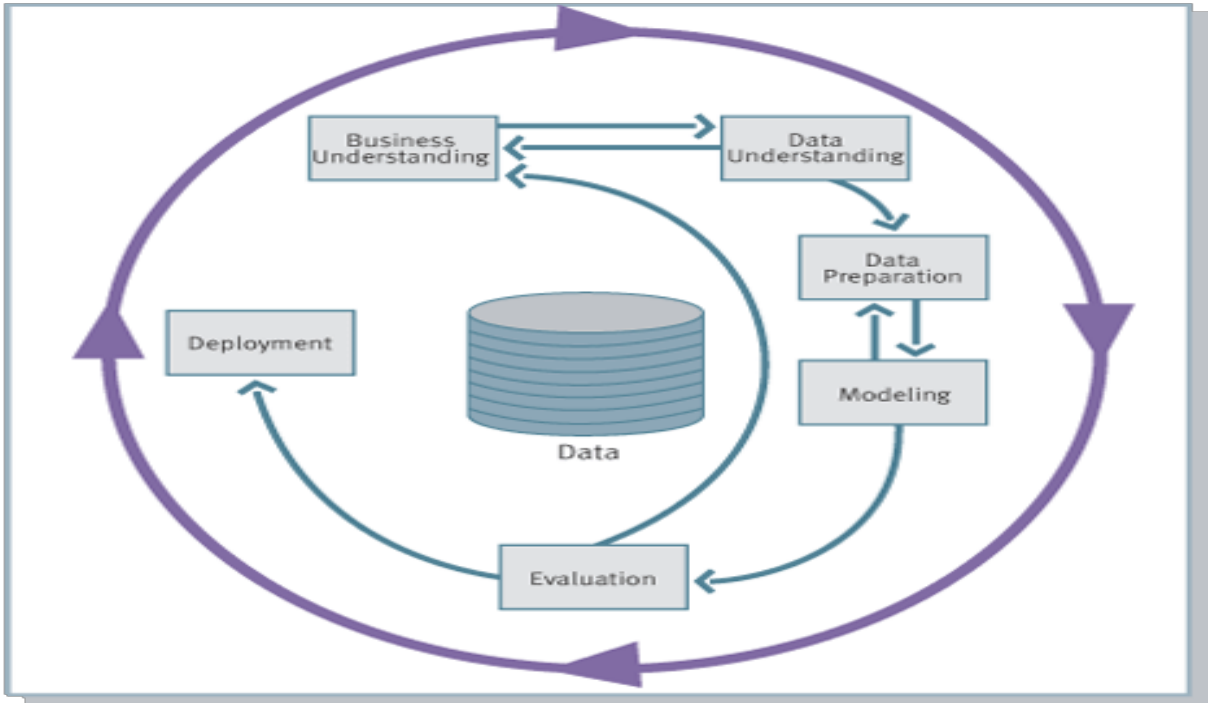
Exploratory Data Analysis

Crisp DM framework



Data Mining Tasks:
Description
Estimation
Prediction
Classification
Clustering
Association

Data Preprocessing

Checking Range

Checking Membership Value

Whitespace Inconsistency

Case Inconsistency

Uniformity

Cross- Validation

Missing Data Analysis

Outliers Analysis

Exploratory Data Analysis

Exploring Variables

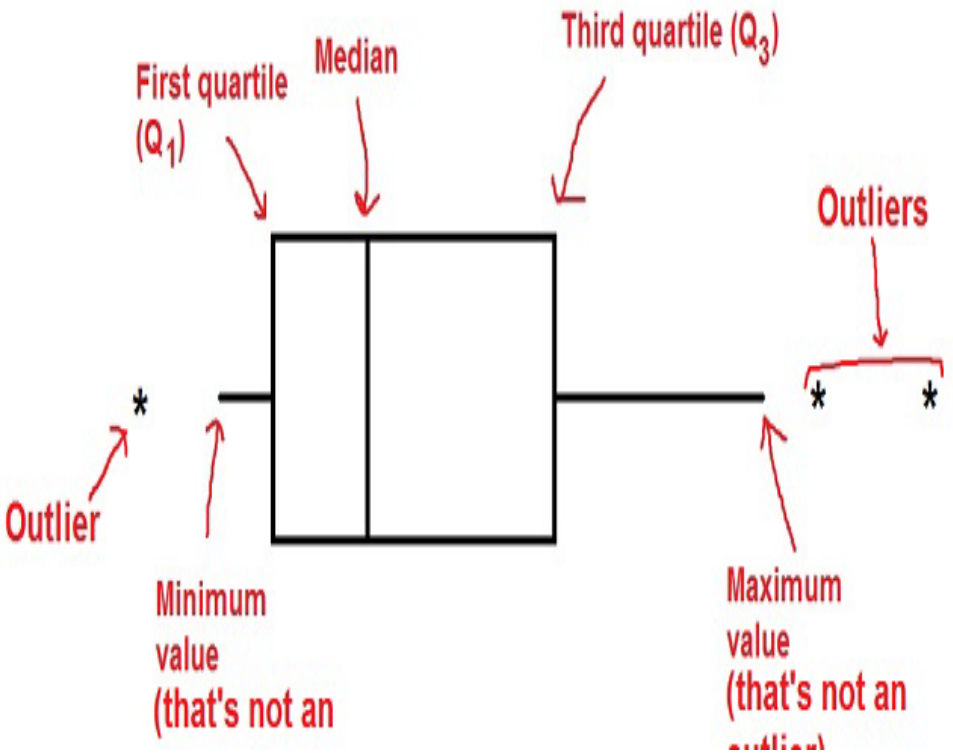
Goals: Exploratory Data Analysis

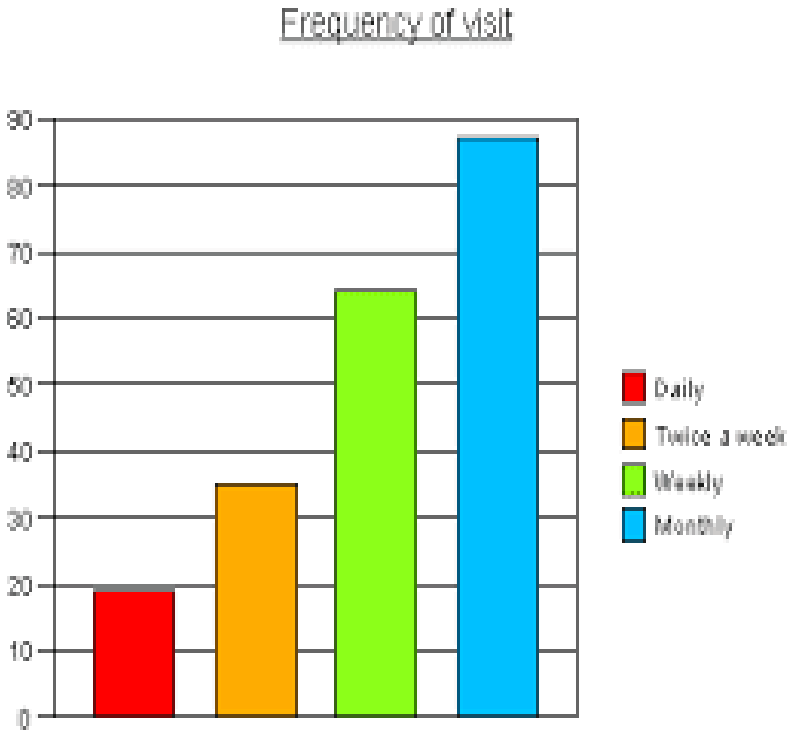
- Become familiar with data
- Explore relationships among variable sets
- While performing EDA, remain focused on objective
- That is, creating data mining model of customer likely to "churn"

- Investigate variables
  - Numeric → Analyze Histograms, Scatter Plots, Statistics, Correlations
  - Categorical → Bar-plot, Boxplots, Cross-tabulations

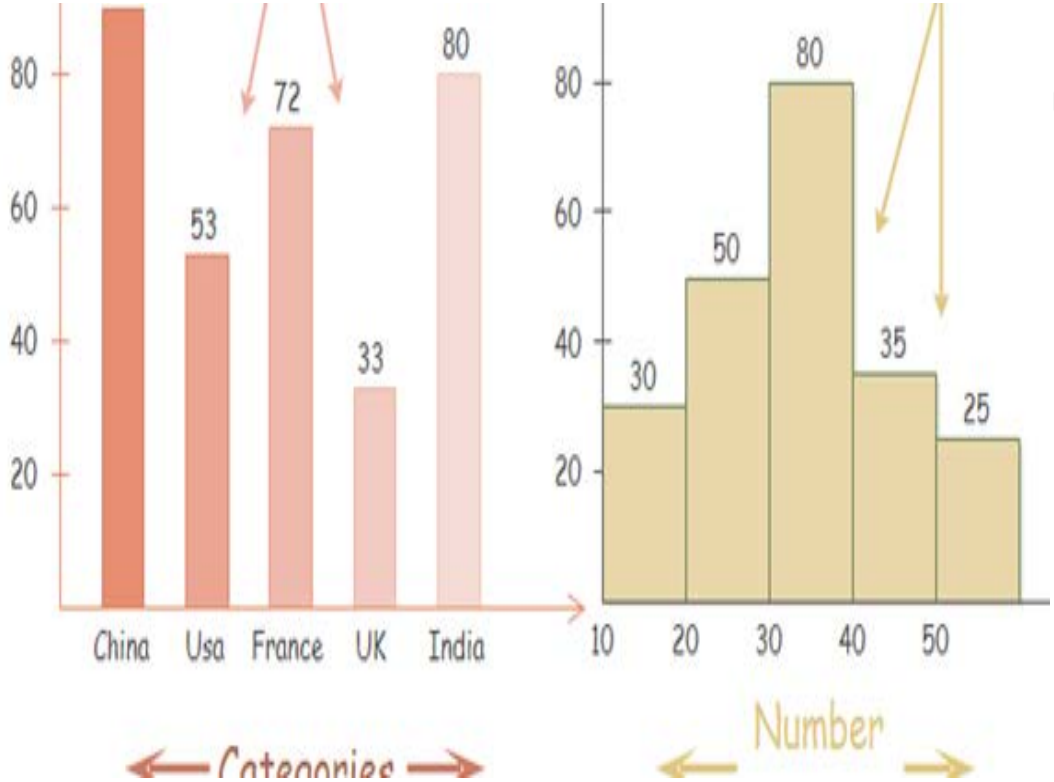
"The greatest value of a picture is when it forces us to notice what we never expected to see."

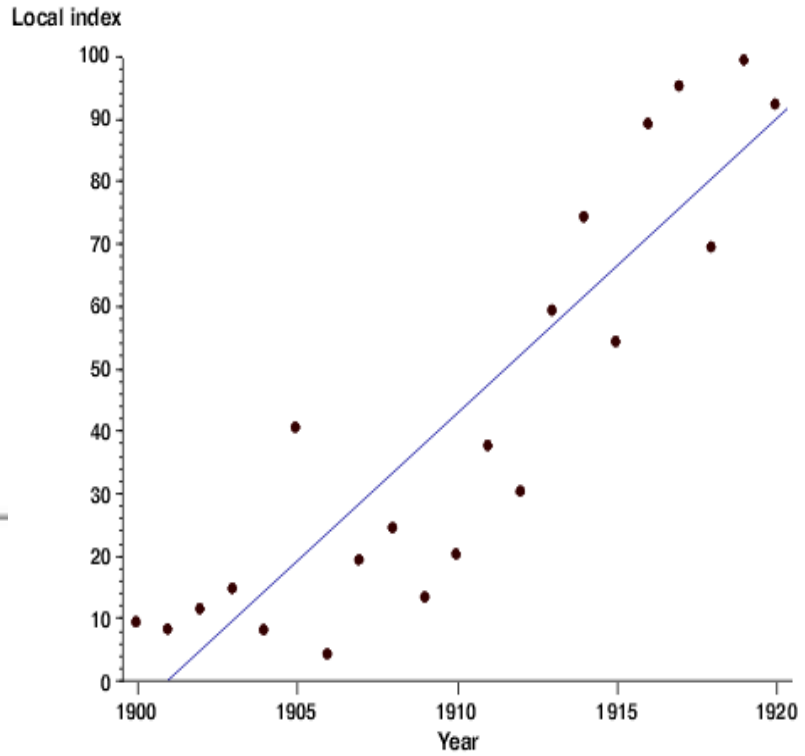
—John W. Tukey, Exploratory Data Analysis (1977)





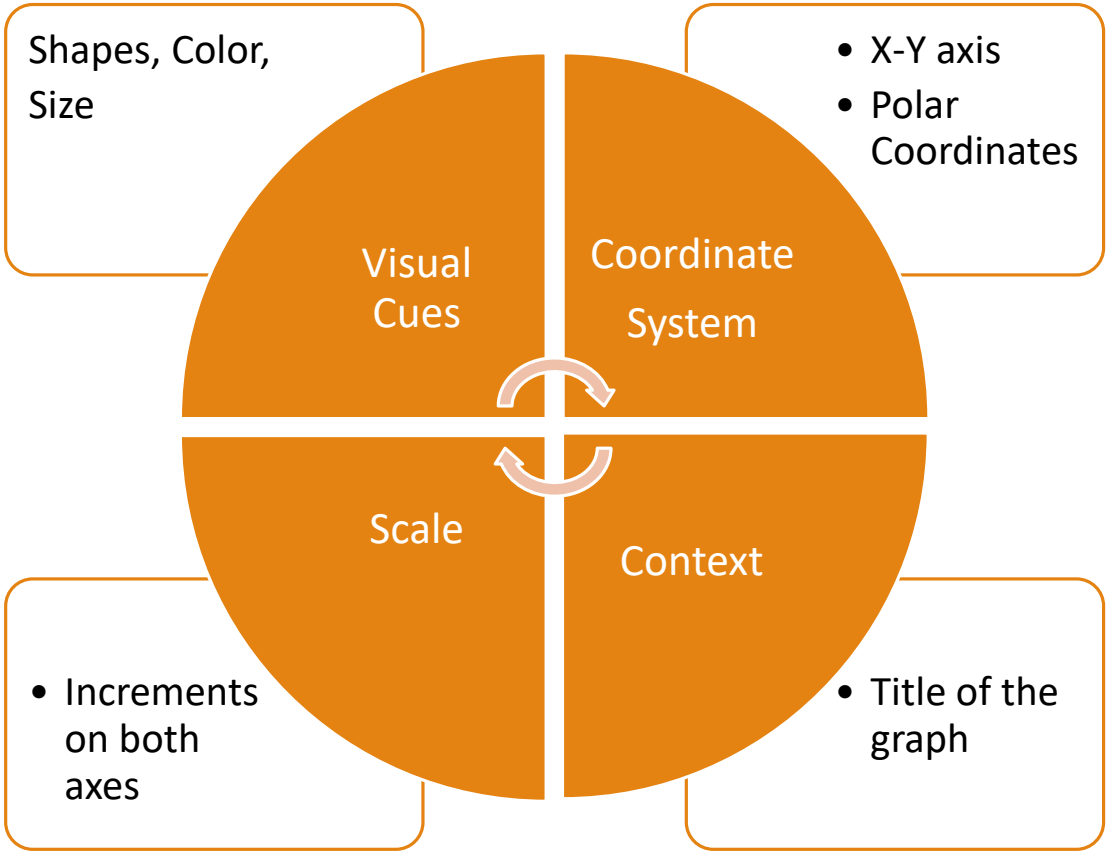
Bar Chart, Boxplot





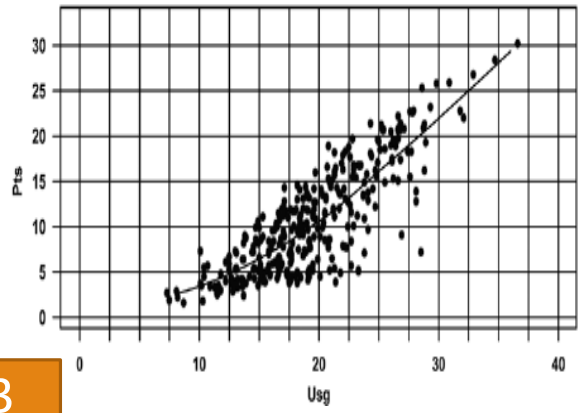
Histogram, Line Graph, Scatterplot

Working Parts of a Graph

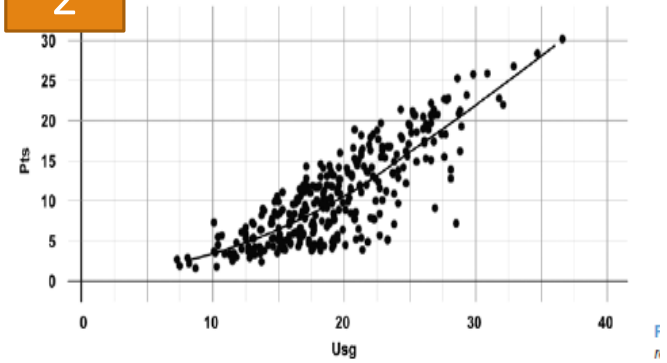


NBA player's usage percentage versus points per game

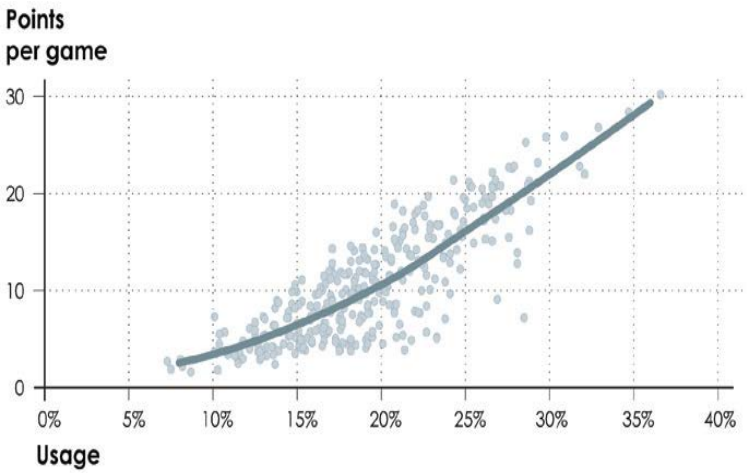
1



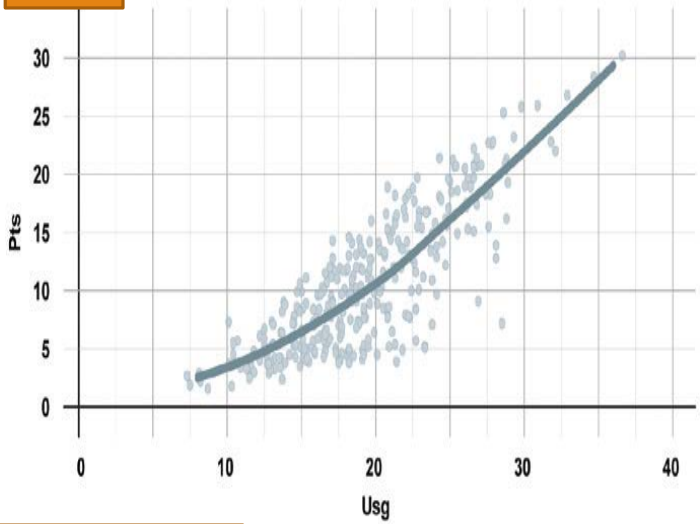
2



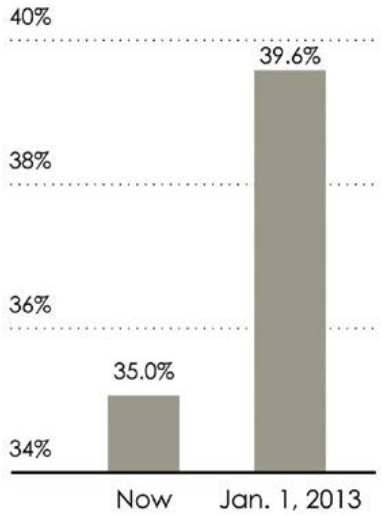
3



4

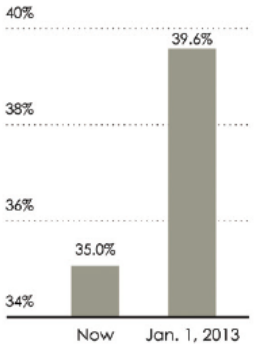


Visual Hierarchy

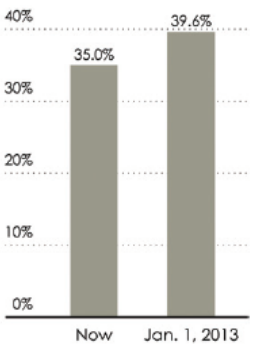


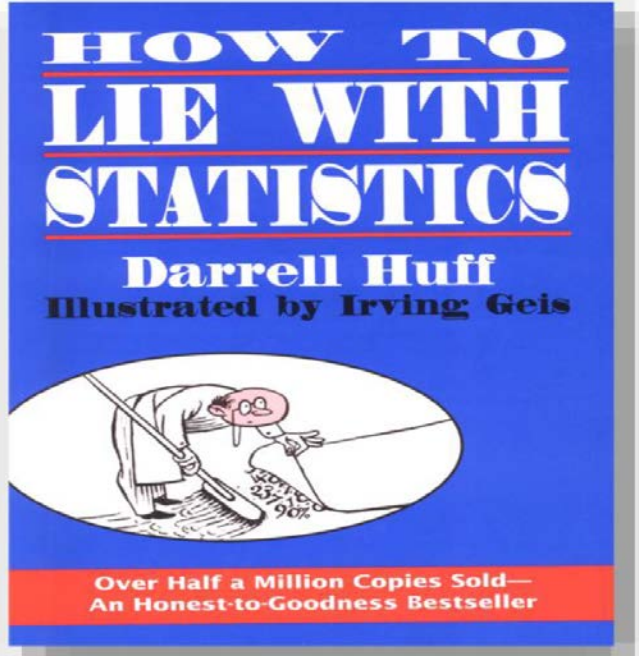
Tax Rate in 2013 and 2014

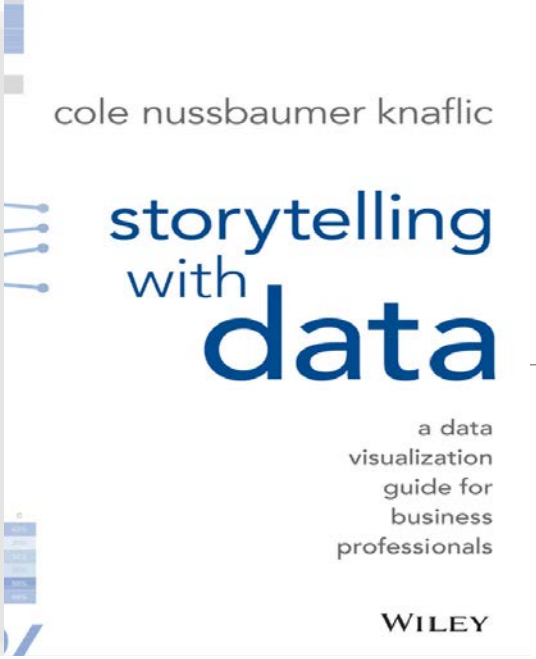
Axis starting at 34 percent

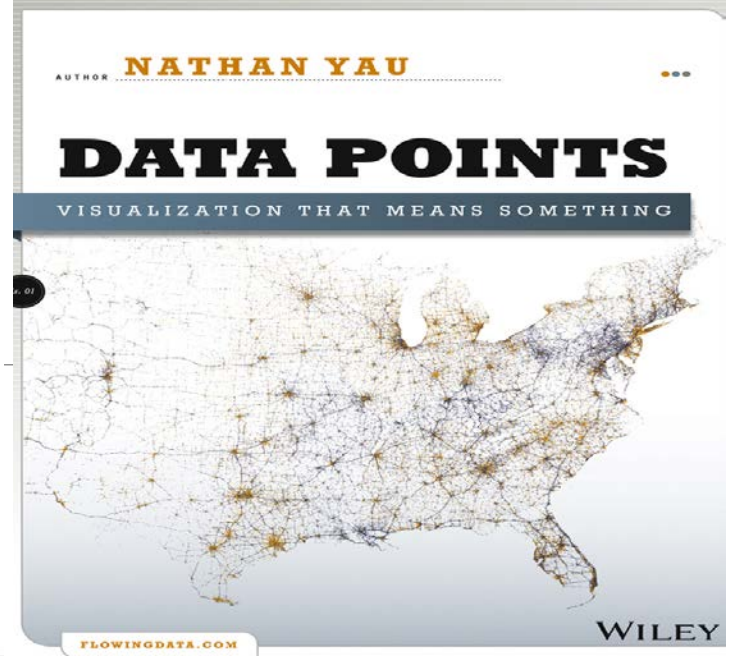


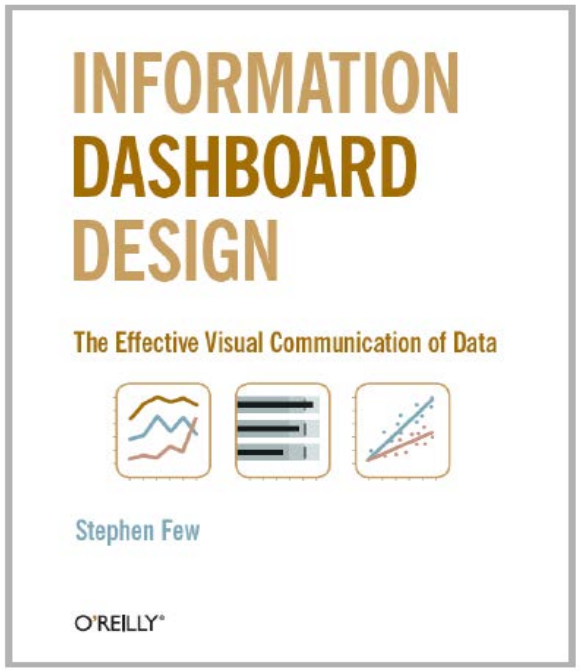
Axis starting at 0 percent

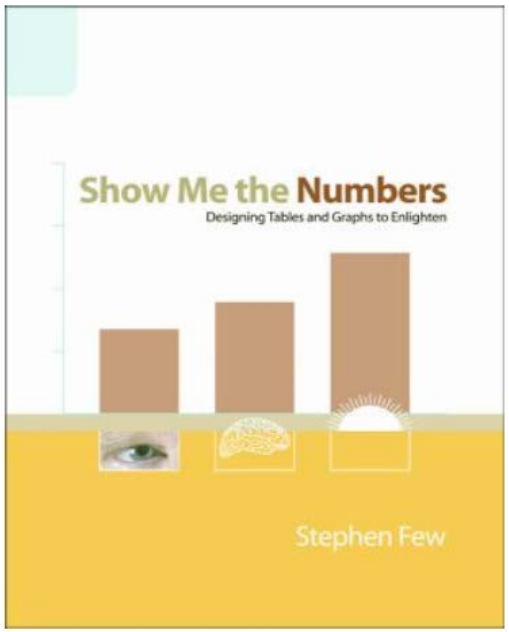












https://www.perceptualedge.com/blog/
https://postgraphics.tumblr.com/tagged/Behind-the-scenes/

Getting to Know the Data Set

- Graphs, plots, and tables often uncover important relationships in data
- The 3,333 records and 20 variables in *churn* data set are explored
- Simple approach looks at field values of records

	State	Account Length	Area Code	Phone	Intl Plan	VMail Plan	VMail Message	Day Mins	Day Calls	Day Charge	Eve Mins
1	KS	128	415	382-4657	no	yes	25	265.100	110	45.070	197.400
2	OH	107	415	371-7191	no	yes	26	161.600	123	27.470	195.500
3	NJ	137	415	358-1921	no	no	0	243.400	114	41.380	121.200
4	OH	84	408	375-9999	yes	no	0	299.400	71	50.900	61.900
5	OK	75	415	330-6626	yes	no	0	166.700	113	28.340	148.300
6	AL	118	510	391-8027	yes	no	0	223.400	98	37.980	220.600
7	MA	121	510	355-9993	no	yes	24	218.200	88	37.090	348.500
8	MO	147	415	329-9001	yes	no	0	157.000	79	26.690	103.100
9	LA	117	408	335-4719	no	no	0	184.500	97	31.370	351.600
10	WV	141	415	330-8173	yes	yes	37	258.600	84	43.960	222.000

	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge	CustServ Calls	Churn
1	99	16.780	244.700	91	11.010	10.000	3	2.700	1	False
2	103	16.620	254.400	103	11.450	13.700	3	3.700	1	False
3	110	10.300	162.600	104	7.320	12.200	5	3.290	0	False
4	88	5.260	196.900	89	8.860	6.600	7	1.780	2	False
5	122	12.610	186.900	121	8.410	10.100	3	2.730	3	False
6	101	18.750	203.900	118	9.180	6.300	6	1.700	0	False
7	108	29.620	212.600	118	9.570	7.500	7	2.030	3	False
8	94	8.760	211.800	96	9.530	7.100	6	1.920	0	False
9	80	29.890	215.800	90	9.710	8.700	4	2.350	1	False
10	111	18.870	326.400	97	14.690	11.200	5	3.020	0	False

Figure 3.1 Field Values of the First Ten Records in the *churn* Data Set

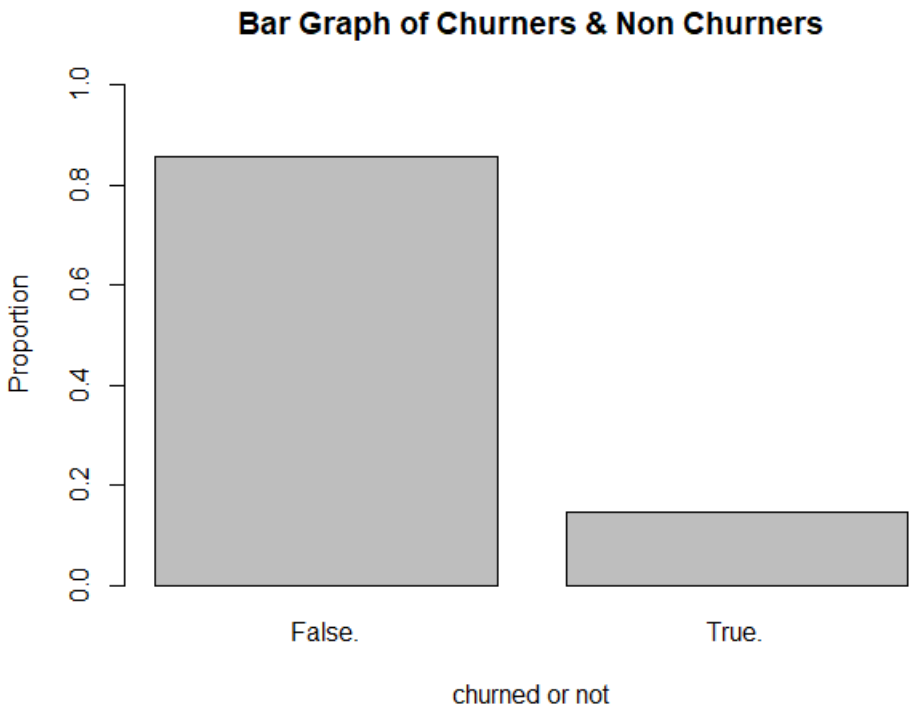
Getting to Know the Data Set *(cont'd)*

- Eight attributes shown in Figure 3.1:
  - State: categorical
  - Account Length: numeric
  - Area Code: categorical
  - Phone: categorical
  - Intl Plan: Boolean
  - VMail Plan: Boolean
  - Vmail Messages: numeric
  - Day Mins: numeric

- "churn" attribute (not shown) indicates customers leaving one company in favor of another company's products or services

Explore Target Variable

False.	True.
2850 (85%)	483 (15%)



Explore Categorical Variables

Are the customers who subscribe to the international plan more likely to churn the company than those without the plan? Yes/No.

Figure 3.4 Comparison Bar Chart of Churn Proportions, by International Plan Participation

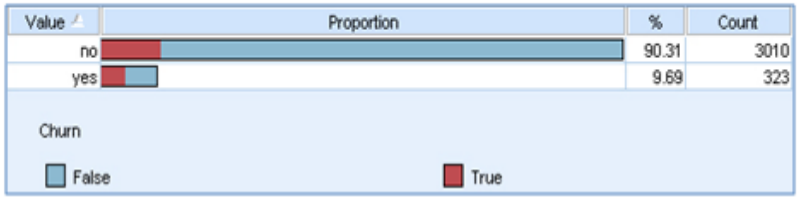
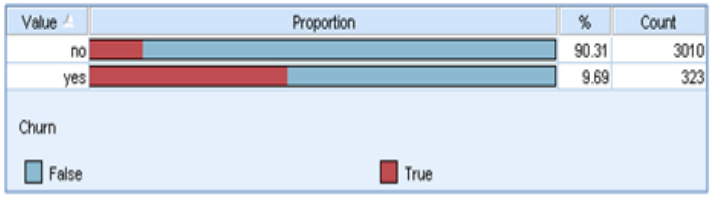


Figure 3.5 Comparison Bar Chart of Churn Proportions, by International Plan Participation, with Equal Bar Length



- Those selecting *International Plan* more likely to churn
- However, relationship not quantified

- Cross-tabulation quantifies relationship between *Churn* and *International Plan*
- *International plan* and *Churn* variables both categorical

Table 3.1 Contingency table of International Plan with Churn

		International Plan		
		No	**Yes**	**Total**
Churn	**False**	2664	186	**2850**
	True	346	137	**483**
	Total	**3010**	**323**	**3333**

Quantifying the relationship:

- (137 / (137 + 186)), 42.4% of customers in *International Plan* churned
- (346 / (346 + 2664)), 11.5% of customers not in *International Plan* churned

Insights:

- Customers selecting *International Plan* more than 3 times likely to leave company, as compared to those not in plan

Diagnostic & Predictive Analytics

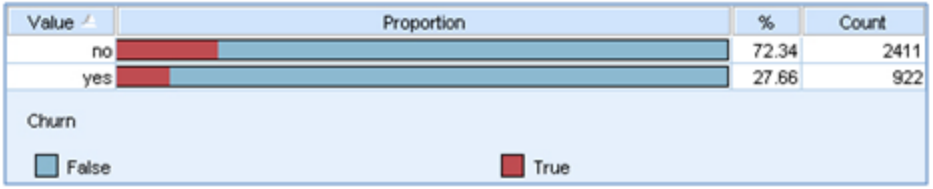
- Why does *International Plan* apparently cause customers to leave?
- Data models predicting *churn* will likely include *International Plan* as predictor

Are the customers who subscribe to the voice mail plan more likely to churn the company than those without the plan? Yes/No.

Table 3.4 Contingency table with column percentages for the Voice Mail Plan

		Voice Mail Plan		
		No	**Yes**	**Total**
Churn	**False**	Count 2008	Count 842	**Count 2850**
		Col % 83.3%	Col % 91.3%	**Col % 85.5%**
	True	Count 403	Count 80	**Count 483**
		Col % 16.7%	Col % 8.7%	**Col % 14.5%**
	Total	**2411**	**922**	**3333**

Figure 3.10 Those Without the VoiceMail Plan are More Likely to Churn



- Only 8.7% = 80/922 of those in plan are churners
- Of those not in plan, 16.7% = 403/2,411 are churners
- Therefore, those not participating in plan ~2X likely to churn, as compared to those in plan
- Perhaps customer loyalty can be increased by simplifying enrollment into *Voice Mail Plan.*
- Data models predicting *churn* likely to include *Voice Mail Plan* as predictor

Exploring Numeric Variables

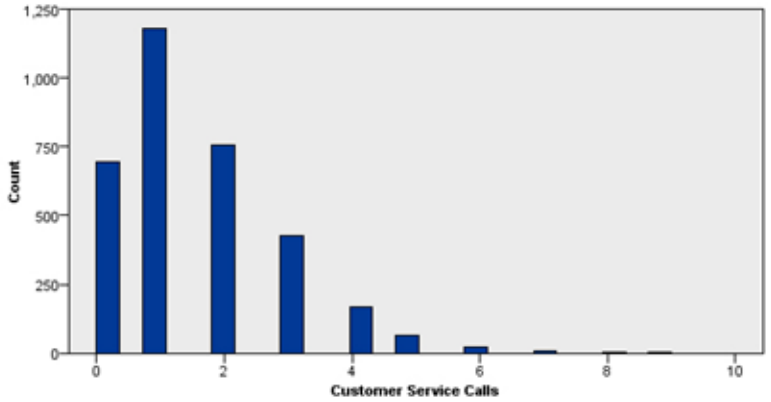


Figure 3.13 Histogram of Customer Service Calls with No Overlay

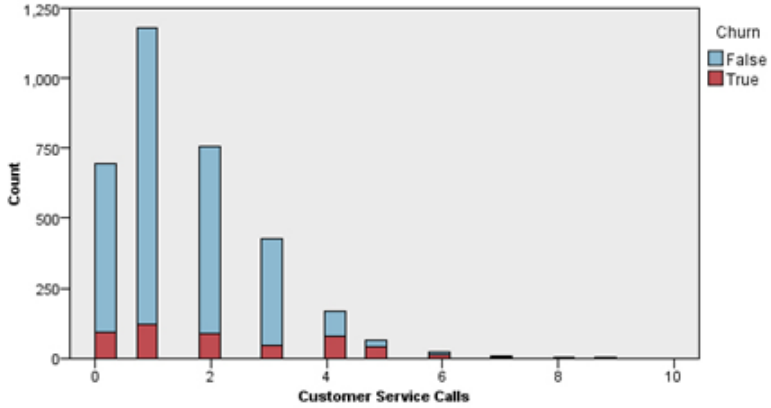


Figure 3.14 Histogram of Customer Service Calls, with Churn Overlay

Figure 3.15 "Normalized" Histogram of Customer Service Calls, with Churn Overlay

Is there any relation between the customer churn and number of customer service calls?

- Observations
- Customer Service Call Distribution is right-skewed and has mode = 1
- Customers calling customer service 3 or fewer times, far less likely to churn
- Diagnostics & Predictive
- Carefully track number of customer service calls made by customers; Offer incentives to retain those making higher number of calls
- Data mining model will probably include *Customer Service Calls* as predictor

Exploring Numeric Variables *(cont'd)*

- Normalized histogram of *Day Minutes* shown with *Churn* overlay (Top)
- Indicates high usage customers churn at significantly greater rate
- Results: Carefully track customer *Day Minutes* as total exceeds 200
- Investigate why those with high usage tend to leave
- Normalized histogram of *Evening Minutes* shown with *Churn* overlay (Bottom)
- Higher usage customers churn slightly more
- Results: Based on graphical evidence, no specific conclusions drawn

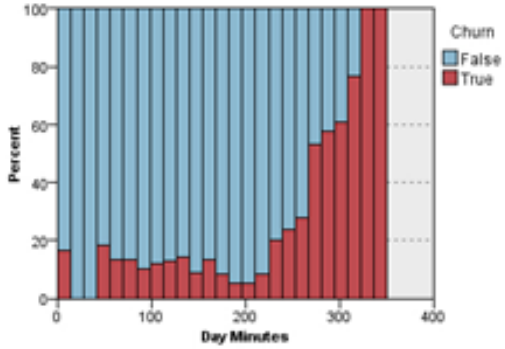


Figure 3.16b Normalized histogram of day minutes

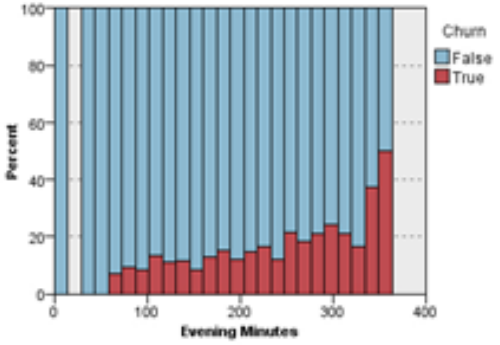
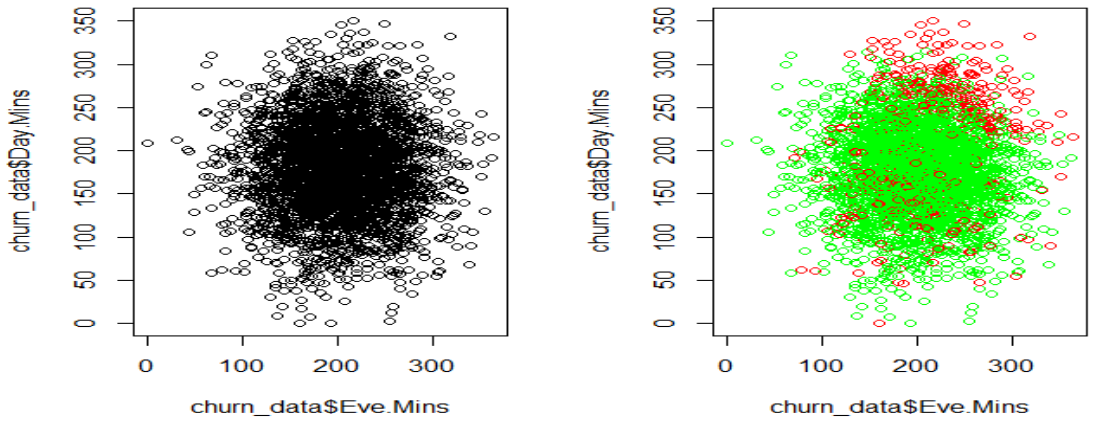


Figure 3.17b Normalized histogram of evening minutes

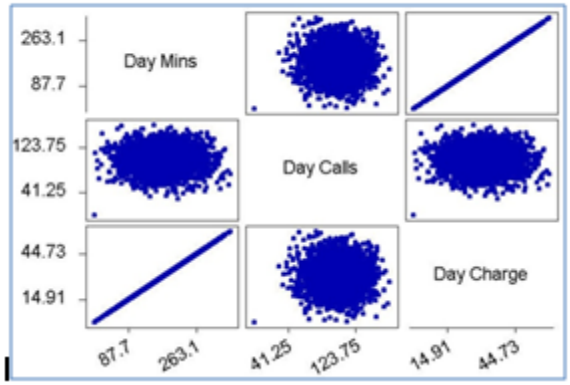
Exploring Multivariate Relationships

- Multivariate graphics can uncover new interaction effects which our univariate exploration missed
- Figure below shows a scatter plot of *day minutes* vs. *evenings minutes*, with churners indicated by red circles



CORRELATION AMONG PREDICTOR VARIABLES

Figure 3.27 Matrix Plot of *Day Minutes*, *Day Calls*, and *Day Charge*



	Day.Mins	Day.Calls	Day.Charge
Day. Mins	1	0.00675	0.999
Day.Calls	0.00675	1	0.00675
Day.Charge	0.999	0.00675	1

Day. Mins and Day. Charge are perfectly correlated.

Conclusion

- The four charge fields (Check for Day Charge) are linear functions of the minutes field, and should be omitted.
- Customers with International plan tend to churn more frequently.
- Customers with Voice Mail plan tend to churn less frequently.
- Customers with four or more service calls are more likely to churn
- Customers with high day Minutes tend to churn at a higher rate than do customers with more low Day Minutes
- Customers with both high Day Mins. And High Evening Minutes tend to churn at a higher rate than other customers.
- Customers with both low Day Mins. And High Customer Service Calls tend to churn at a higher rate than other customers.
- For the remaining predictors, EDA uncovers no obvious association of churn.
- However these variables are still retained for input to downstream models and techniques.