Exploratory Data Analysis

Exploratory Data Analysis

- **Exploratory Data Analysis (EDA)** is that part of statistical practice concerned with reviewing, communicating and using data where there is a low level of knowledge about its cause system.
- Many **EDA techniques** have been adopted into data mining and are being taught to young students **as a way to introduce them to statistical thinking**.

- **www.wikipedia.org**

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical first step in analyzing data from an experiment.

Main reasons to use EDA:

- ❑ Detection of mistakes
- ❑ Checking of assumptions
- ❑ Preliminary selection of appropriate models
- ❑ Determining relationships among the explanatory variables, and
- ❑ Assessing direction and rough size of relationships between explanatory and outcome variables.

➢ Loosely speaking, any method of looking at data that does not include formal statistical modeling and inference falls under term exploratory data analysis.

Typical data format and the types of EDA

- ❑ The data from an experiment are generally collected into a rectangular array.
- ❑ One row per experimental subject and one column for each subject identifier, outcome variable, and explanatory (independent) variable.
- ❑ Each column contains numeric values or the levels.
- ❑ People are not very good at looking at a column of numbers or a whole spread- sheet and then determining important characteristics of the data.
- ❑ Looking at numbers to be tedious, boring, and/or overwhelming.
- ❑ Exploratory data analysis techniques have been devised as an aid in this situation.
- ❑ Most of these techniques work in part by hiding certain aspects of the data while making other aspects more clear.

Typical data format and the types of EDA

- ❑ Exploratory data analysis is generally cross-classified in two ways.
  - ✓ non-graphical or
  - ✓ graphical.
- ❑ And second, each method is either
  - ✓ univariate or
  - ✓ multivariate (usually just bivariate).
- ❑ Non-graphical methods generally involve calculation of summary statistics.
- ❑ Graphical summarizes data in a diagrammatic or pictorial way.
- ❑ Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships.
- ❑ Usually our multivariate EDA will be bivariate, but occasionally it will involve three or more variables.
- ❑ It is almost always a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.

Typical data format and the types of EDA

- ❑ Beyond the four categories created by the above cross-classification, each of the categories of EDA have further divisions based on
  - ❑ the role (outcome or explanatory) and
  - ❑ type (categorical or quantitative) of the variable(s) being examined.
- ❑ Although there are guidelines about which EDA techniques are useful in what circumstances, there is an important degree of looseness and art to EDA.
- ❑ EDA need not be restricted to techniques discussed; sometimes you need to invent a new way of looking at your data.
- ❑ **The four types of EDA are:**
  1. univariate non-graphical,
  2. multivariate non-graphical,
  3. univariate graphical, and
  4. multivariate graphical.

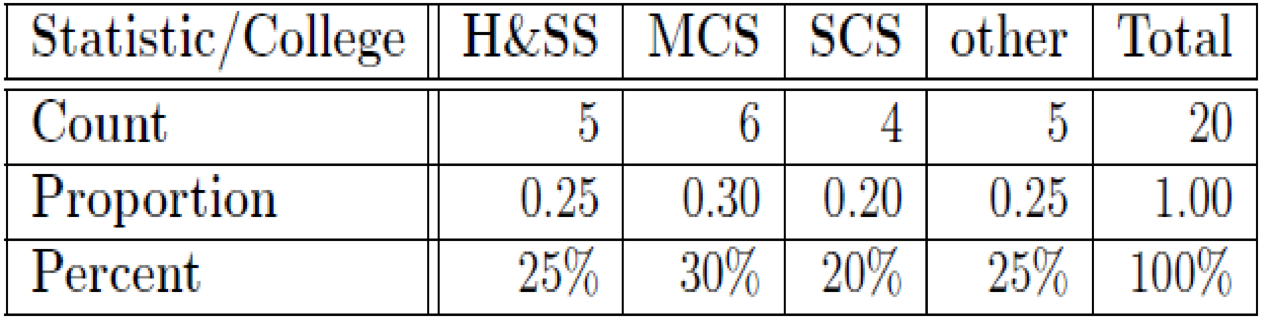
Univariate non-graphical EDA

- ❑ The usual goal of univariate non-graphical EDA is to better appreciate “sample distribution“ and also to make some tentative conclusions about what population distribution(s) is/are compatible with the sample distribution.
- ❑ Outlier detection is also a part of this analysis.

Univariate non-graphical EDA - Categorical data

Categorical data:

- ❑ The characteristics of interest for a categorical variable are simply the range of values and the frequency (or relative frequency) of occurrence for each value.
- ❑ Useful for univariate non-graphical techniques for categorical variables.
- ❑ For example if we categorize subjects by College at Carnegie Mellon University as H&SS, MCS, SCS and \other", there is a true population of all students enrolled



Statistic/College	H&SS	MCS	SCS	other	Total
Count	5	6	4	5	20
Proportion	0.25	0.30	0.20	0.25	1.00
Percent	25%	30%	20%	25%	100%

A simple tabulation of the frequency of each category is the best univariate non-graphical EDA for categorical data.

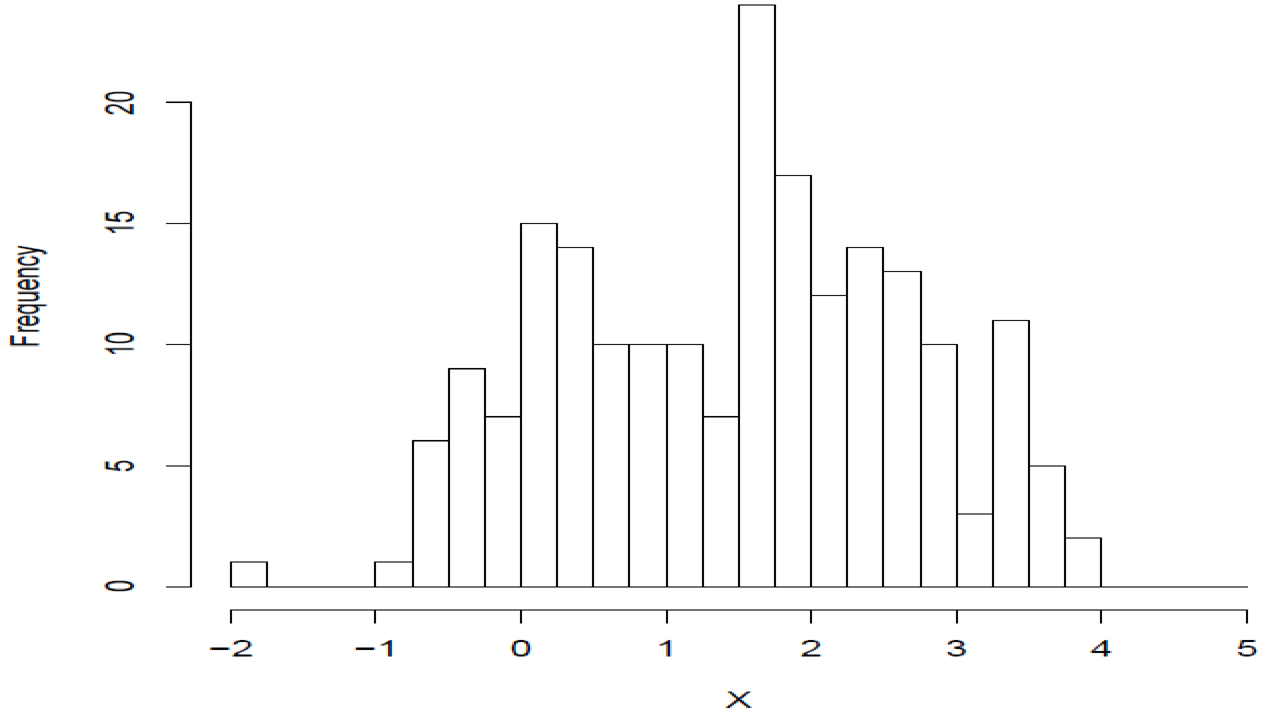
Univariate non-graphical EDA

Characteristics of quantitative data

- ❑ Univariate EDA for a quantitative variable is a way to make preliminary assessments about the population distribution of the variable using the data of the observed sample.
- ❑ Can calculate sample statistics" from data, such as sample mean, sample variance, sample standard deviation, sample skewness and sample kurtosis.
- ❑ Characteristics of population distribution of a quantitative variable are its center, spread, modality, shape, and outliers.
- ❑ Sample's distributional characteristics are seen qualitatively in the univariate graphical EDA technique of a histogram

Univariate non-graphical EDA

Histogram of a sample of size 200



- ❑ The bi-modality is visible, as is an outlier at X=-2.
- ❑ **Outlier:** Sample data values which correspond to areas of the population pdf (or pmf) with low density (or probability).

Univariate non-graphical EDA

- ❑ For quantitative variables (and possibly for ordinal variables) it is worthwhile looking at the central tendency, spread, skewness, and kurtosis of the data.
- ❑ For categorical variables, none of these make any sense.

Univariate non-graphical EDA

❑ Central tendency

- ❑ The central tendency or location" of a distribution has to do with typical or middle values.
- ❑ Useful measures of central tendency are the statistics called (arithmetic) mean, median, and sometimes mode.
- ❑ The arithmetic mean is simply the sum of all of the data values divided by the number of values.
- ❑ The probability distribution of the sample mean is referred to as its sampling distribution.
- ❑ The median is another measure of central tendency.
- ❑ For symmetric distributions, the mean and the median coincide.
- ❑ The median has a very special property called robustness.
- ❑ Rarely used measure of central tendency is the mode

- ➢ Most common measure of central tendency is mean.
- ➢ For skewed distribution or (concern about outliers), median is preferred.

Univariate non-graphical EDA

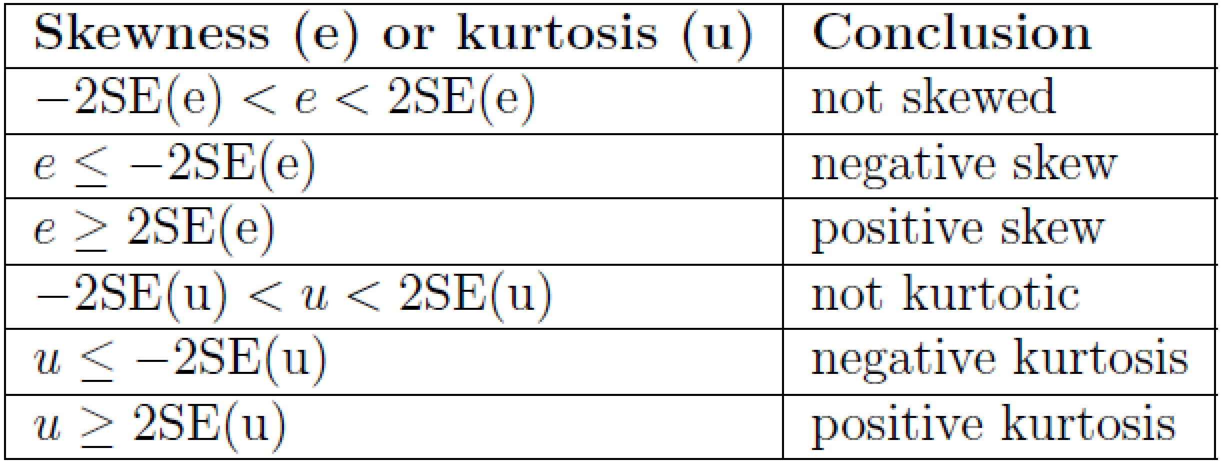
Spread

- ❑ Commonly used as a measure of spread of a distribution, including variance, standard deviation, and interquartile range.
- ❑ Spread is an indicator of how far away from the center
- ❑ The variance is a standard measure of spread.
- ❑ Variance of a population is defined as the mean squared deviation
- ❑ Measure of spread, because the bigger the deviations from the mean, the bigger the variance gets.
- ❑ The standard deviation is simply the square root of variance.
- ❑ A third measure of spread is the interquartile range.
- ❑ The IQR is a more robust measure of spread than variance or standard deviation.
- ❑ In contrast to IQR, range of the data is not very robust.

Univariate non-graphical EDA

Skewness and kurtosis

- ❑ Useful univariate descriptors are the skewness and kurtosis of a distribution.
- ❑ Skewness is a measure of asymmetry.
- ❑ Kurtosis is a measure of peaked-ness" relative to a Gaussian shape.
- ❑ If the sample skewness and kurtosis are calculated along with their standard errors, we can roughly make conclusions according to the following table



Skewness (e) or kurtosis (u)	Conclusion
$-2\mathrm{SE(e)} < e < 2\mathrm{SE(e)}$	not skewed
$e \leq -2\mathrm{SE(e)}$	negative skew
$e \geq 2\mathrm{SE(e)}$	positive skew
$-2\mathrm{SE(u)} < u < 2\mathrm{SE(u)}$	not kurtotic
$u \leq -2\mathrm{SE(u)}$	negative kurtosis
$u \geq 2\mathrm{SE(u)}$	positive kurtosis

Where

e is an estimate of skewness and

u is an estimate of kurtosis, and

SE(e) and SE(u) are the corresponding standard errors

Univariate non-graphical EDA

- ❑ Skewness is a measure of asymmetry.
- ❑ Kurtosis is a more subtle measure of peaked-ness compared to a Gaussian distribution.

$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

$$\text{kurtosis: } a_4 = m_4 \,/\, {m_2}^2$$

where

$$m_4 = \sum(x-\bar{x})^4 \,/\, n \quad \text{and} \quad m_2 = \sum(x-\bar{x})^2 \,/\, n$$

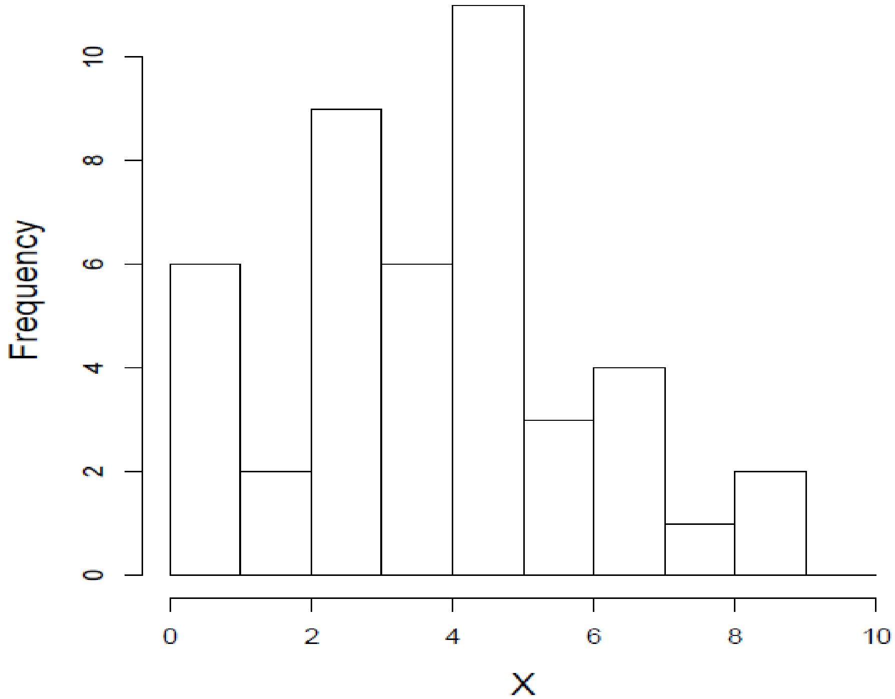
Univariate graphical EDA

- ❑ Non Graphical EDA focuses on sample statistics
- ❑Need to look graphically at distribution of the sample.
- ❑Non-graphical and graphical methods complement each other.
- ❑Non-graphical methods are quantitative and objective- do not give a full picture of the data;
- ❑Graphical methods, more qualitative and involve a degree of subjective analysis

Univariate graphical EDA

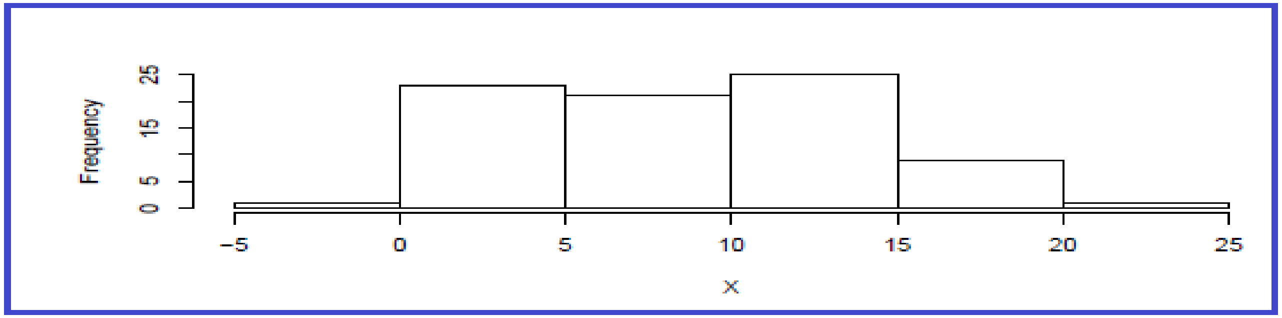
Histograms

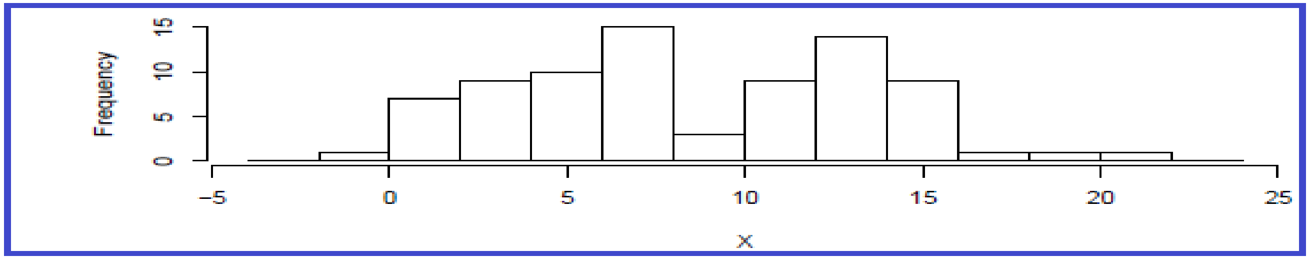
- ❑ Histograms makes sense for categorical data
- ❑ Histogram, which is a bar plot in which each bar represents frequency (count) or proportion
- ❑ Need to decide number of bins in histogram
- ❑ Generally choose between about 5 and 30 bins

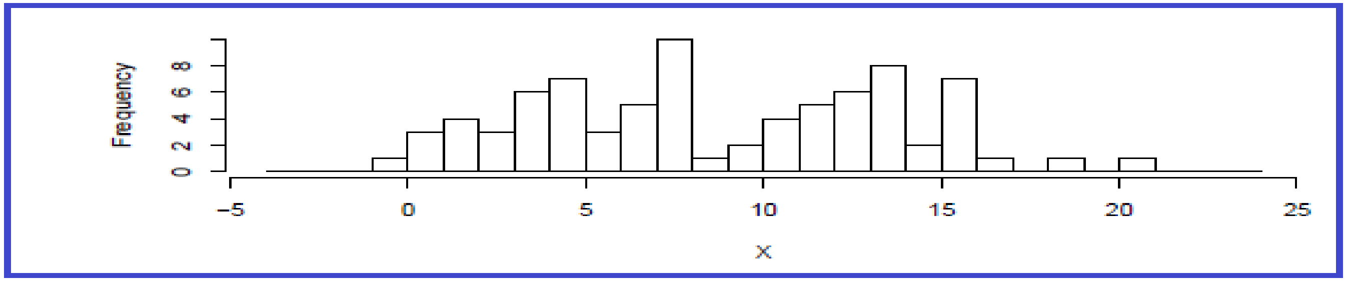


Univariate graphical EDA

Histograms of same sample from a bimodal population using three different bin widths (5, 2 and 1).

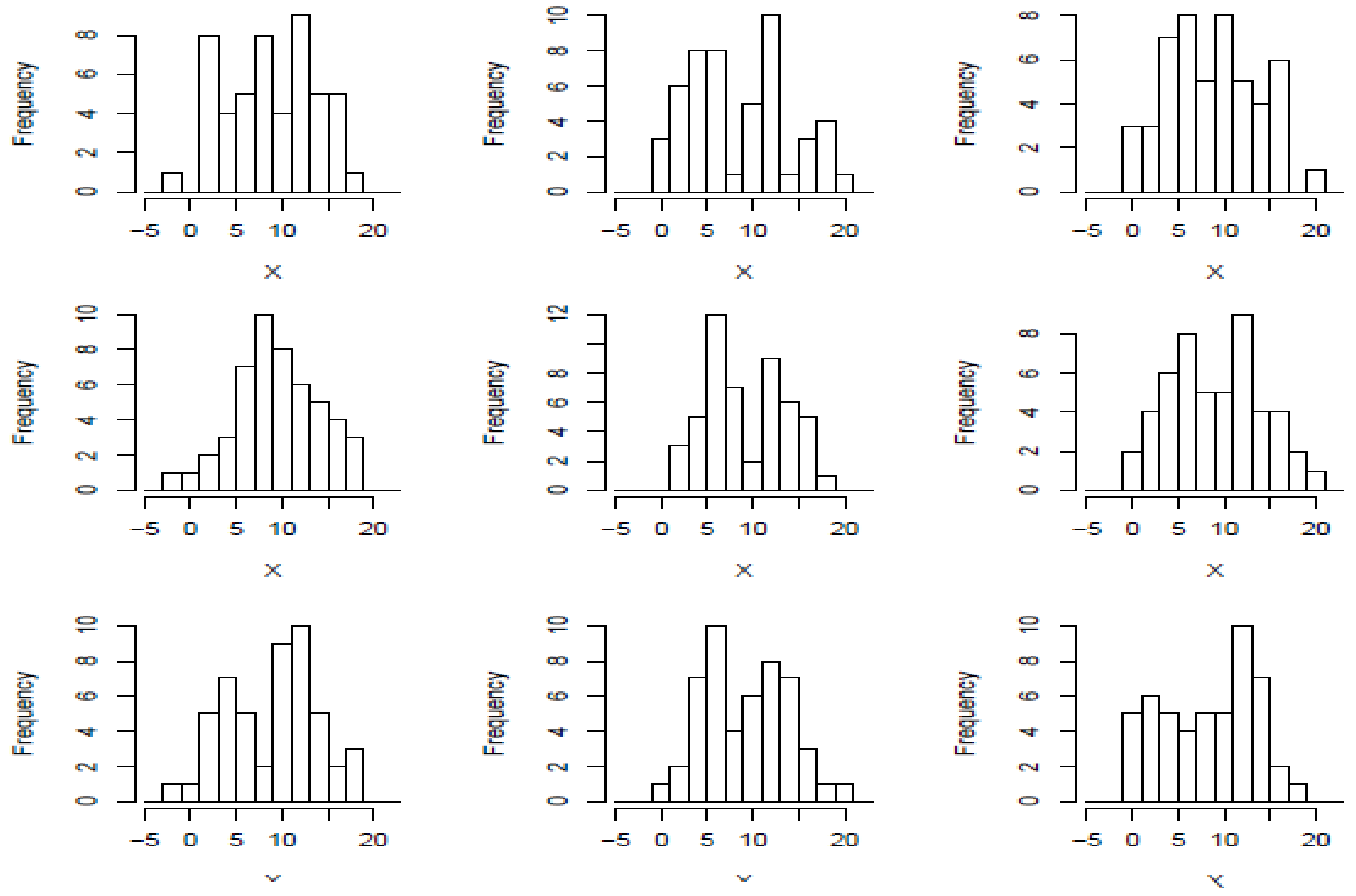






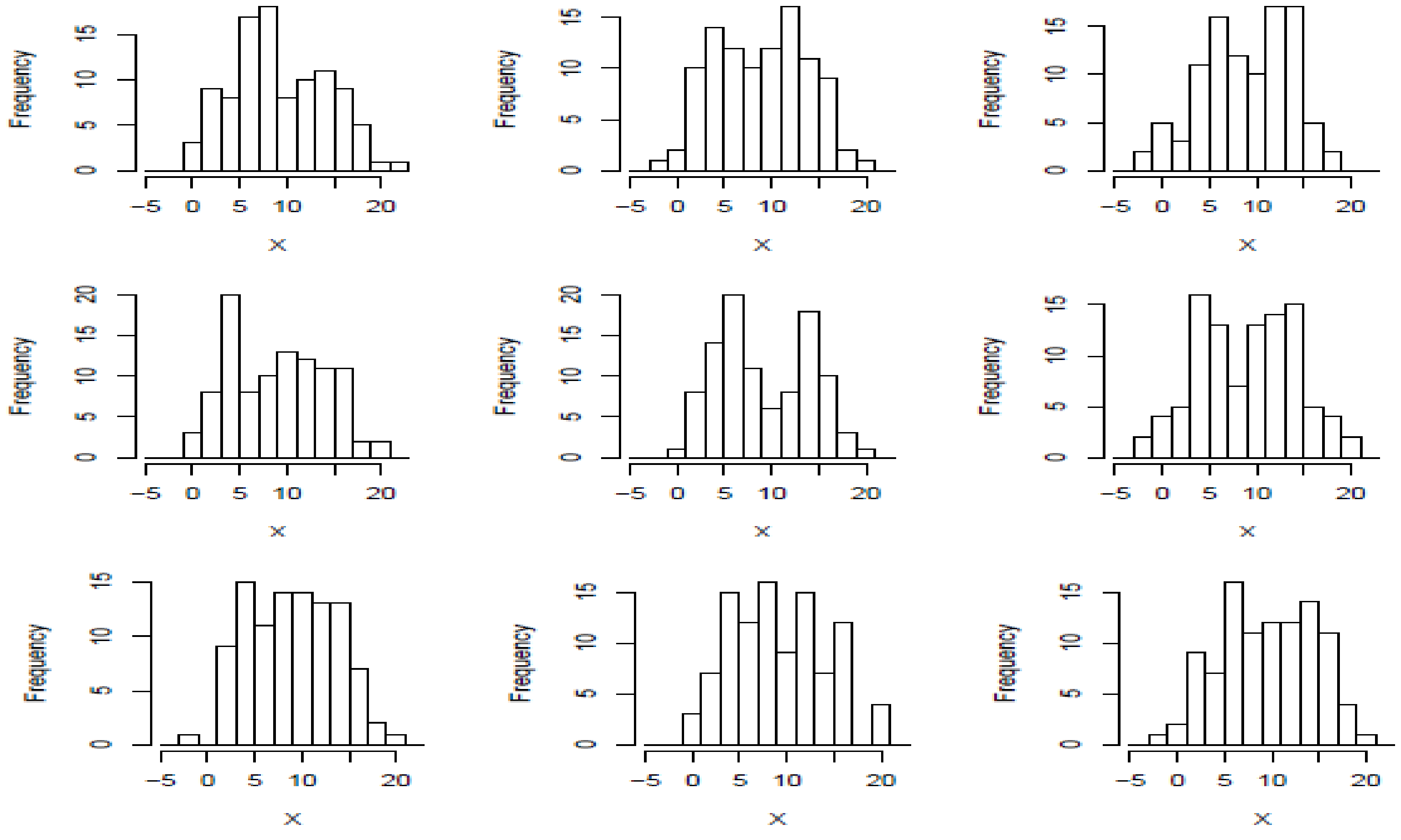
Univariate graphical EDA

Histograms from multiple samples of size 50 from same population



Histograms of multiple samples of size 50.

Univariate graphical EDA



Histograms of multiple samples of size 100.

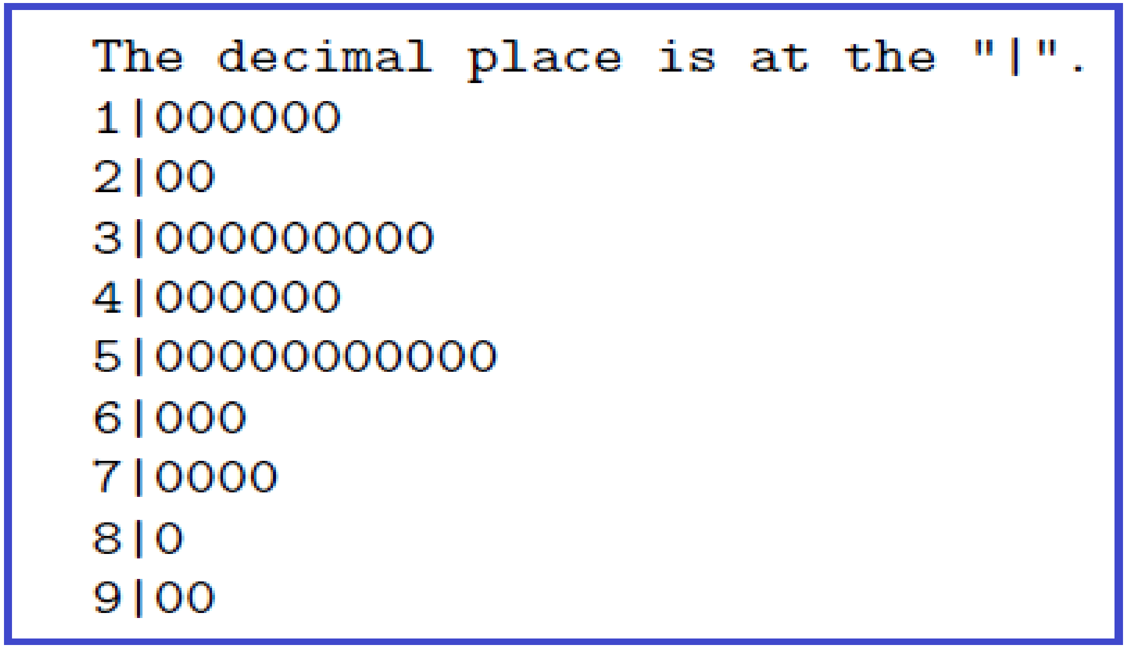
- High variability is quite high, for smaller sample size,
- incorrect impression (particularly of unimodality) is quite possible

Univariate graphical EDA

- ❑ With practice, histograms are one of the best ways to quickly learn a lot about your data, including
  - ✓ central tendency,
  - ✓ spread,
  - ✓ modality,
  - ✓ shape and
  - ✓ outliers.

Univariate graphical EDA

Stem-and-leaf plots



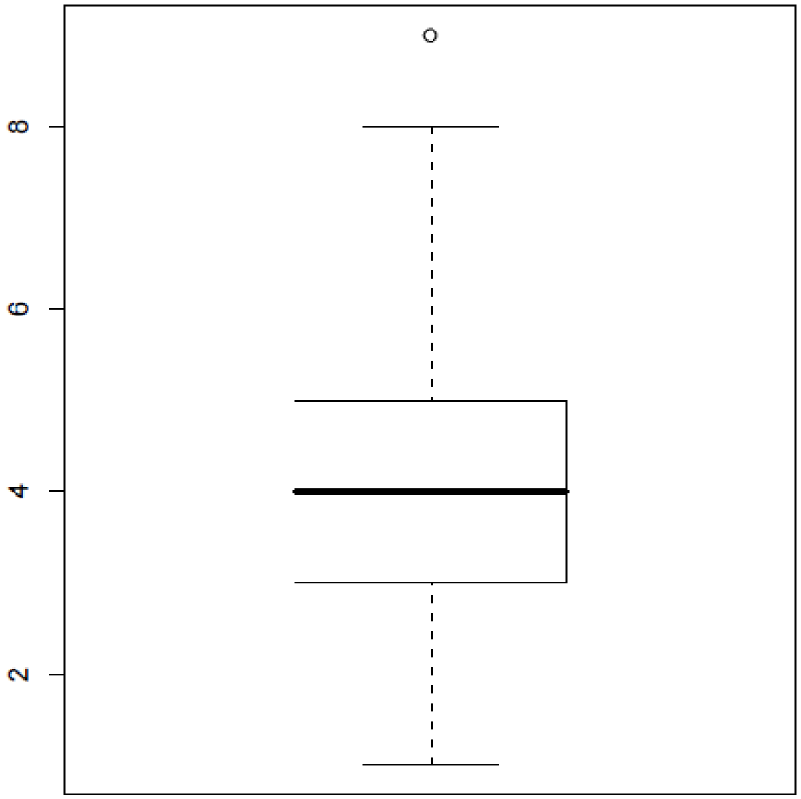
```
  The decimal place is at the "|".
  1|000000
  2|00
  3|000000000
  4|000000
  5|00000000000
  6|000
  7|0000
  8|0
  9|00
```

A stem and leaf plot shows all data values and shape of distribution.

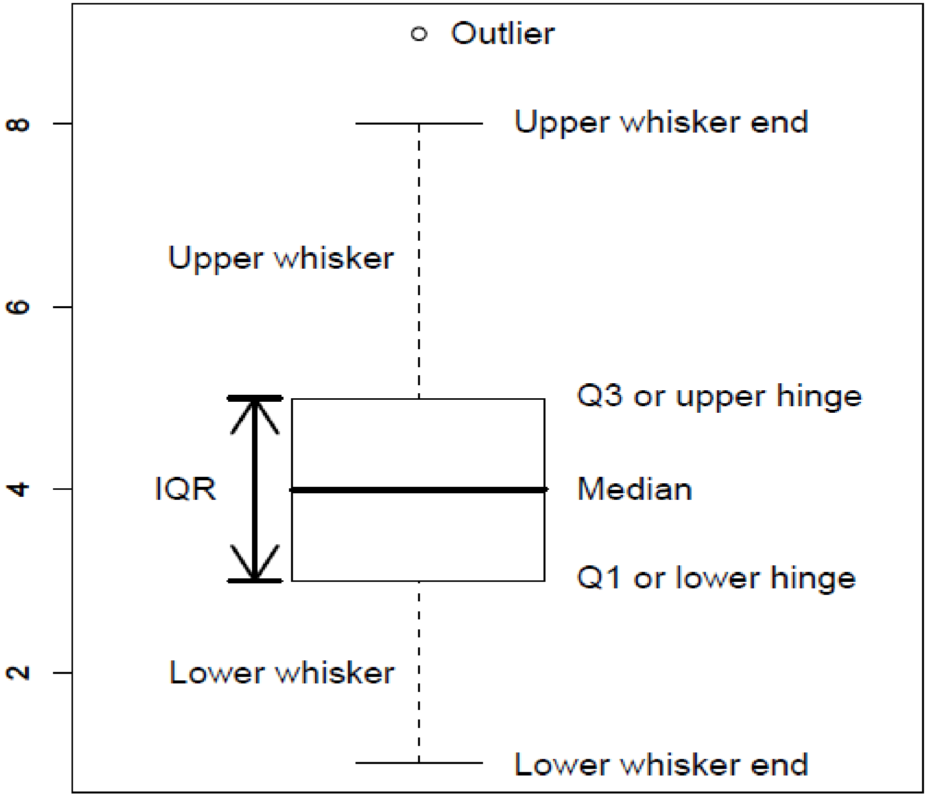
Univariate graphical EDA

Boxplots

- ❑ Very good at presenting information - central tendency, symmetry and skew, & outliers,
- ❑ Misleading about aspects such as multimodality



A boxplot of the data



Annotated boxplot.

Univariate graphical EDA

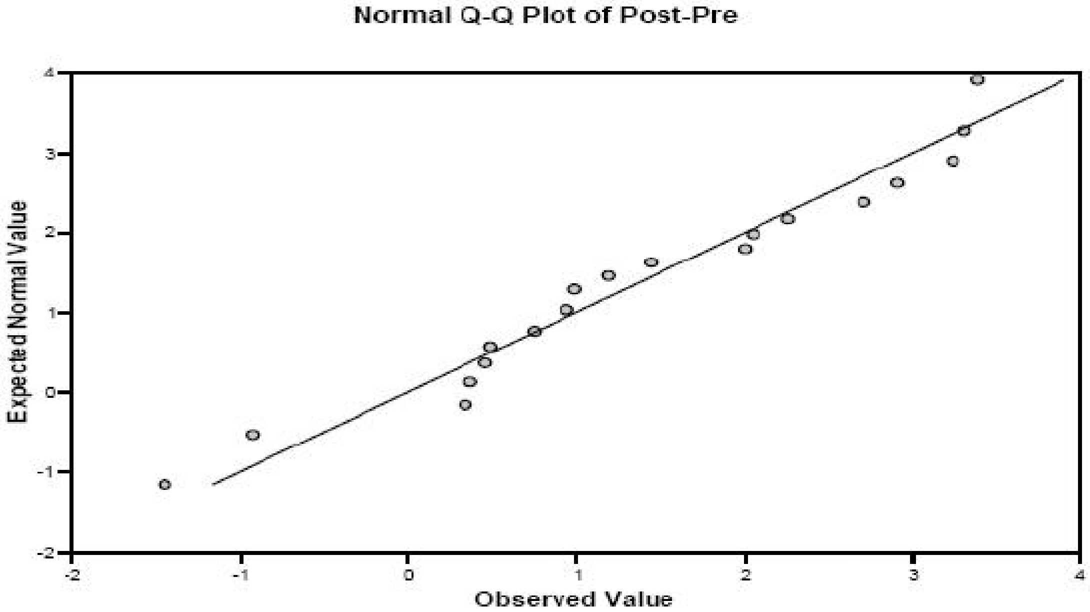
Boxplots

- ❑ Boxplots are excellent EDA plots because - rely on robust statistics like median and IQR rather than more sensitive ones such as mean and standard deviation.
- ❑ With boxplots it is easy to compare distributions (usually for one variable at different levels of another) with a high degree of reliability.
- ❑ Boxplots show robust measures of location and spread as well as providing information about symmetry and outliers.

Univariate graphical EDA

Quantile-normal plots

- Called quantile-normal or QN plot or quantile-quantile or QQ plot
- Used to know how well sample theoretical distribution
- Do not confuse a quantile-normal plot with a simple scatter plot of two variables.
- QN plot is not for EDA, but for examining something called “residuals”

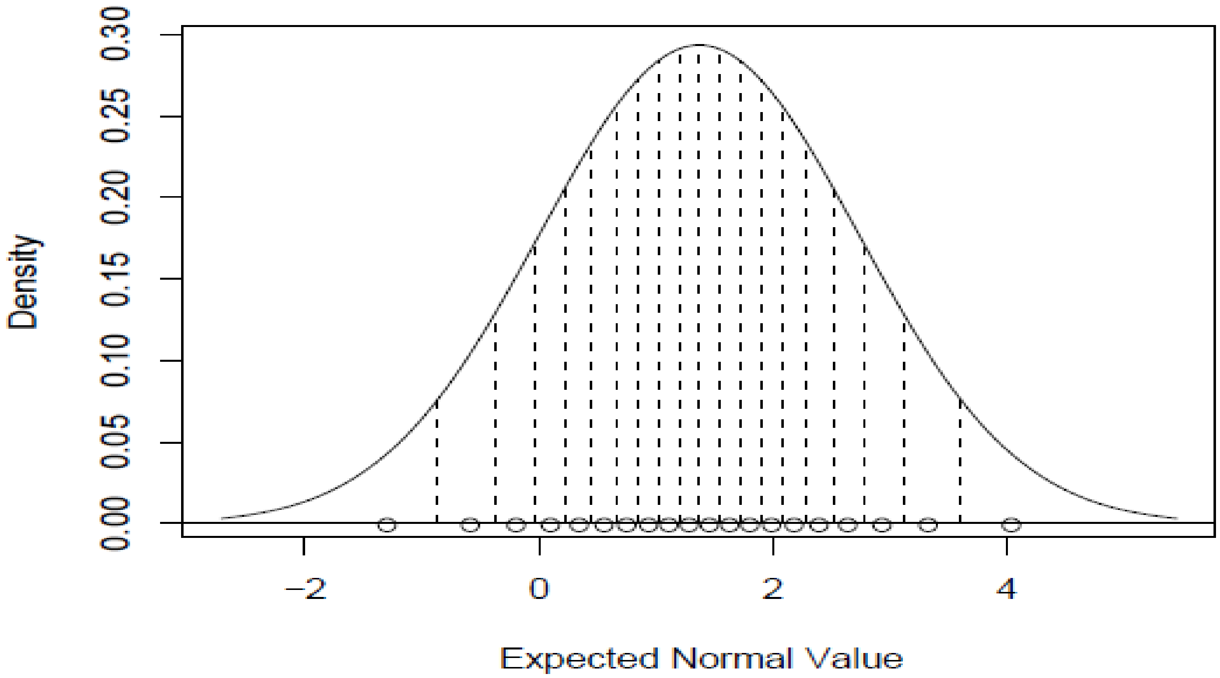


A quantile-normal plot.

Univariate graphical EDA

Quantile-normal plots

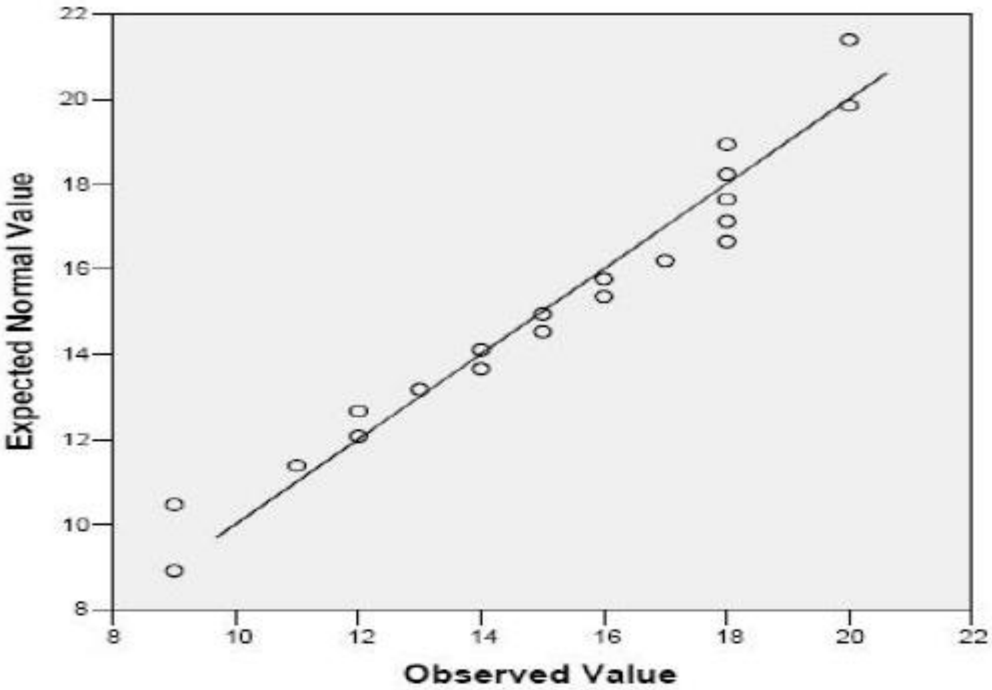
- ❑ Dotted lines divide the bell curve up into 20 equally probable zones,
- ❑ 20 points are at the probability mid-points of each zone.
- ❑ These 20 points, which are more tightly packed near the middle than in the ends, are used as Expected Normal Values" in the QN plot of our actual data.



A way to think about QN plots.

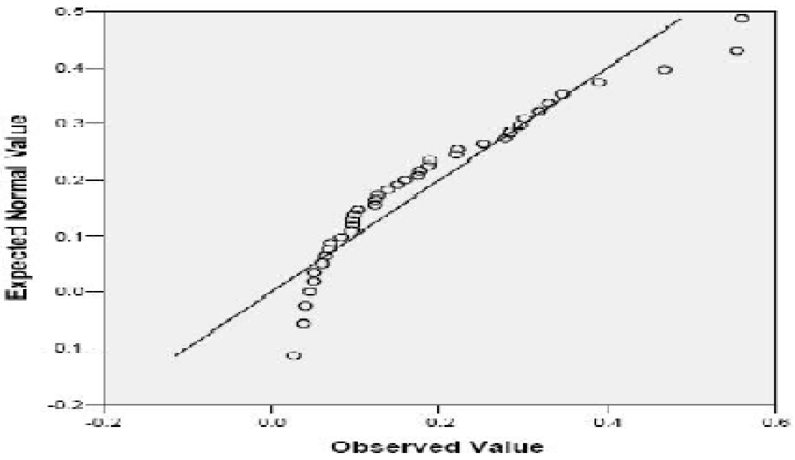
Univariate graphical EDA

- ❑x-axis is observed values, y-axis expected normal value
- ❑Bands indicate ties, i.e., multiple points with the same values.

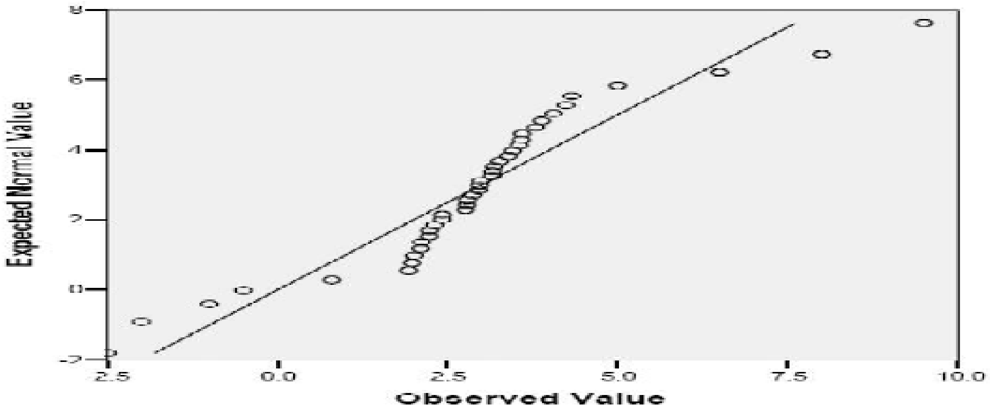


Quantile-normal plot with ties.

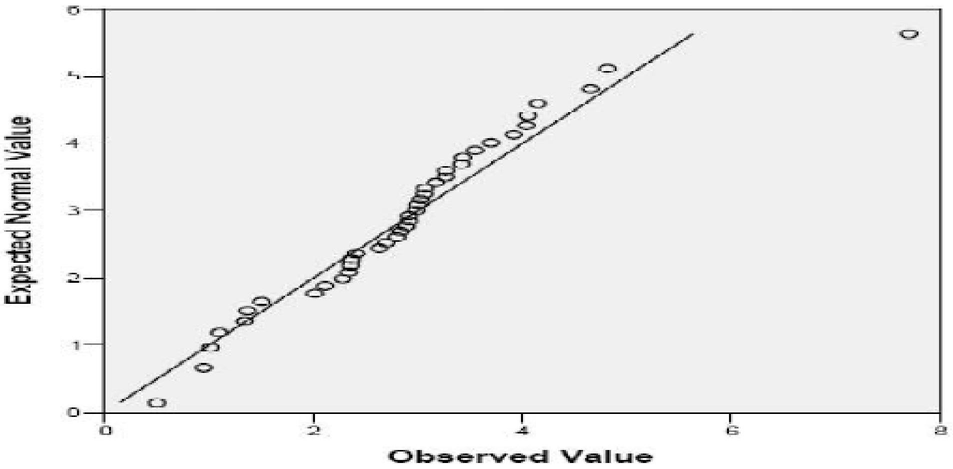
Univariate graphical EDA



Quantile-normal plot showing right skew.

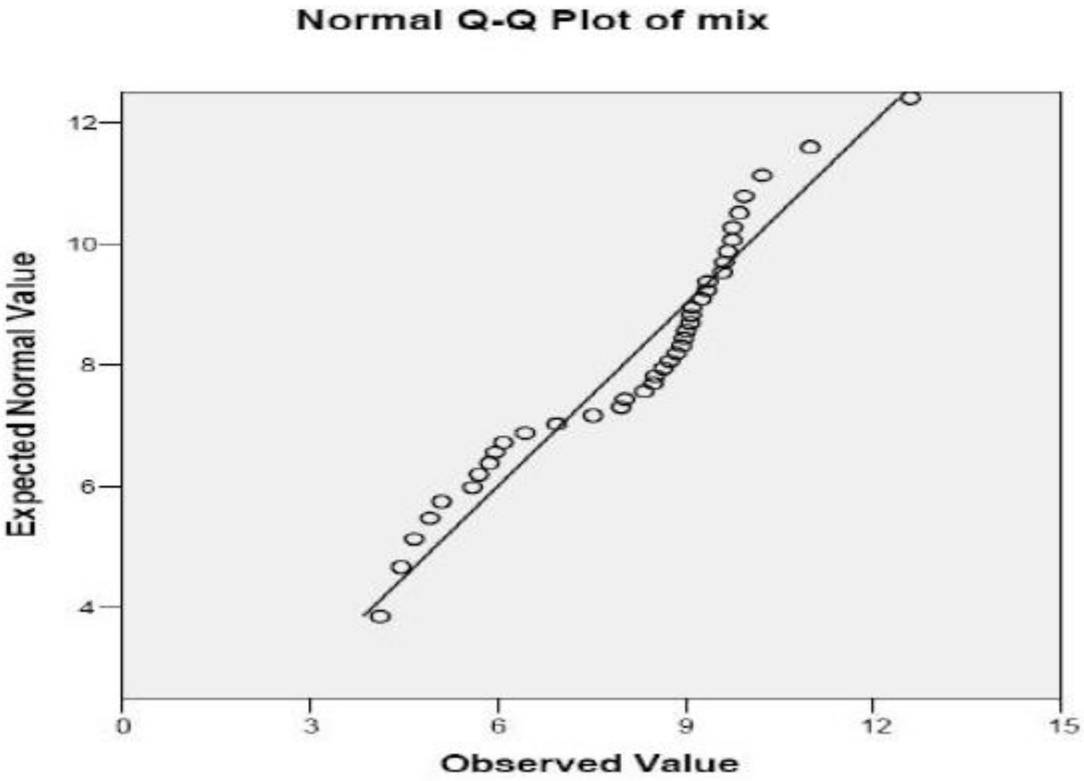


Quantile-normal plot showing fat tails.



Quantile-normal plot showing a high outlier.

Univariate graphical EDA



Quantile-normal plot showing bimodality.

Multivariate non-graphical EDA

❑ Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of

1. Cross-tabulation or
2. Statistics

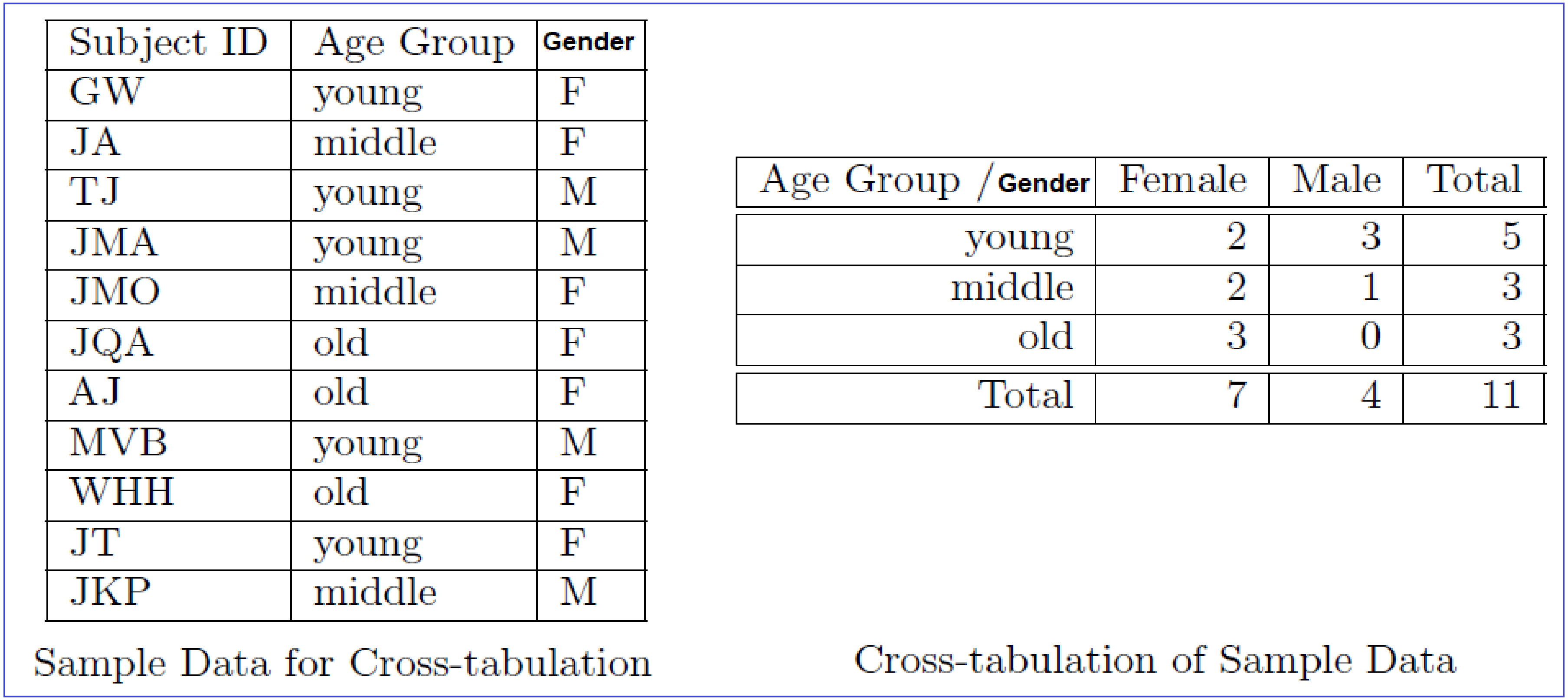
Multivariate non-graphical EDA

Cross-tabulation

- ❑ For categorical data extension of tabulation called cross-tabulation is very useful.
- ❑ For two variables, cross-tabulation is performed by
  - ✓ Making a two-way table with column headings that match the levels of one variable and
  - ✓ Row headings that match the levels of the other variable,
  - ✓ Then filling in the counts of all subjects that share a pair of levels.
- ❑ The two variables might be both explanatory, both outcome, or one of each.
- ❑ Depending on the goals, row percentages, column percentages, and/or cell percentages

Multivariate non-graphical EDA

Cross-tabulation



Subject ID	Age Group	Gender
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M

Sample Data for Cross-tabulation

Age Group / Gender	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

Cross-tabulation of Sample Data

Cross-tabulation is the basic bivariate non-graphical EDA technique.

Multivariate non-graphical EDA

Correlation for categorical data

- ❑ Another statistic that can be calculated for two categorical variables is their correlation.
- ❑ There are many forms of correlation for categorical variables
- ❑ Several coefficients have been defined and makes use the chi-square statistic.

  1. Goodman Kruskal's lambda
  2. Phi co-efficient (uses chi-squared statistic)
  3. Cramer's V (uses chi-squared statistic)
  4. Tschuprow's T (uses chi-squared statistic)
  5. Contingency coefficient C (uses chi-squared statistic)

Multivariate non-graphical EDA

Univariate statistics by category

- ❑For one categorical variable (usually explanatory) and one quantitative variable (usually outcome), it is common to produce some of the standard univariate non-graphical statistics for the quantitative variables separately for each level of the categorical variable
- ❑Then compare statistics across levels of categorical variable.
- ❑Comparing the means is an informal version of ANOVA.
- ❑Comparing medians is a robust informal version of one-way ANOVA.
- ❑Comparing measures of spread is a good informal test of the assumption of equal variances needed for valid analysis of variance.

Multivariate non-graphical EDA

Univariate statistics by category

Especially for a categorical explanatory variable and a quantitative outcome variable, it is useful to produce a variety of univariate statistics for the quantitative variable at each level of the categorical variable.

Multivariate non-graphical EDA

Correlation and covariance

- ❑ For two quantitative variables, the basic statistics of interest are the sample covariance and/or sample correlation
- ❑ The sample covariance is a measure of how much two variables co-vary", i.e., how much (and in what direction) should we expect one variable to change when the other changes.

Multivariate non-graphical EDA

Correlation and covariance

- Sample covariance is calculated by computing (signed) deviations of each measurement from the average of all measurements for that variable.

- Then deviations for the two measurements are multiplied together separately for each subject.

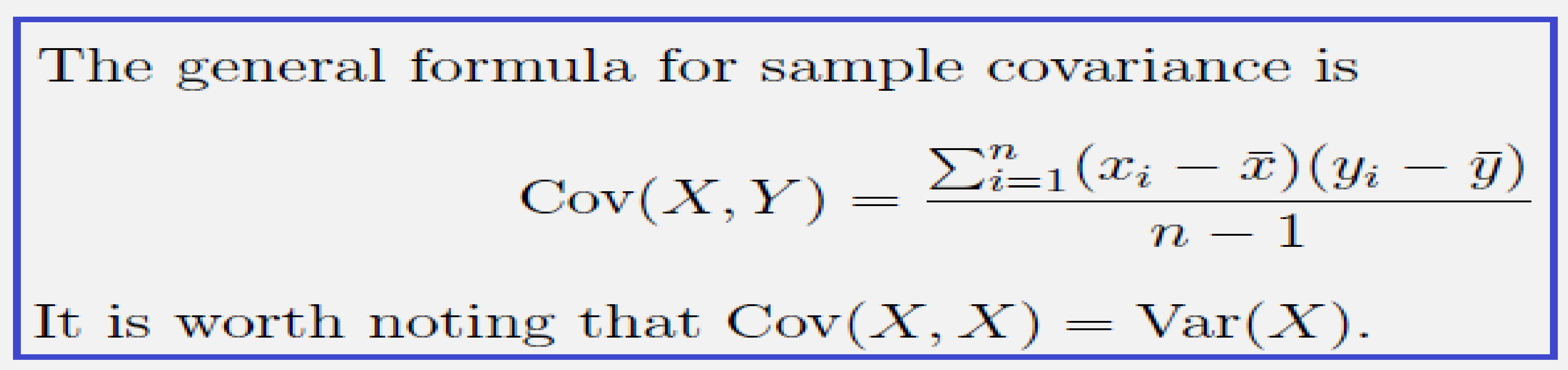
- Finally these values are averaged (actually summed and divided by n-1, to keep the statistic unbiased).

- Note that the units on sample covariance are the products of the units of the two variables.

Multivariate non-graphical EDA

Correlation and covariance

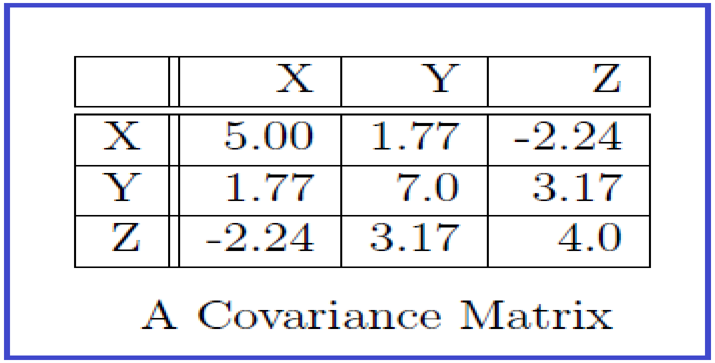
❑ When we have many quantitative variables the most common non-graphical EDA technique is to calculate all of the pairwise covariances and/or correlations and assemble them into a matrix.



The general formula for sample covariance is

$$\text{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

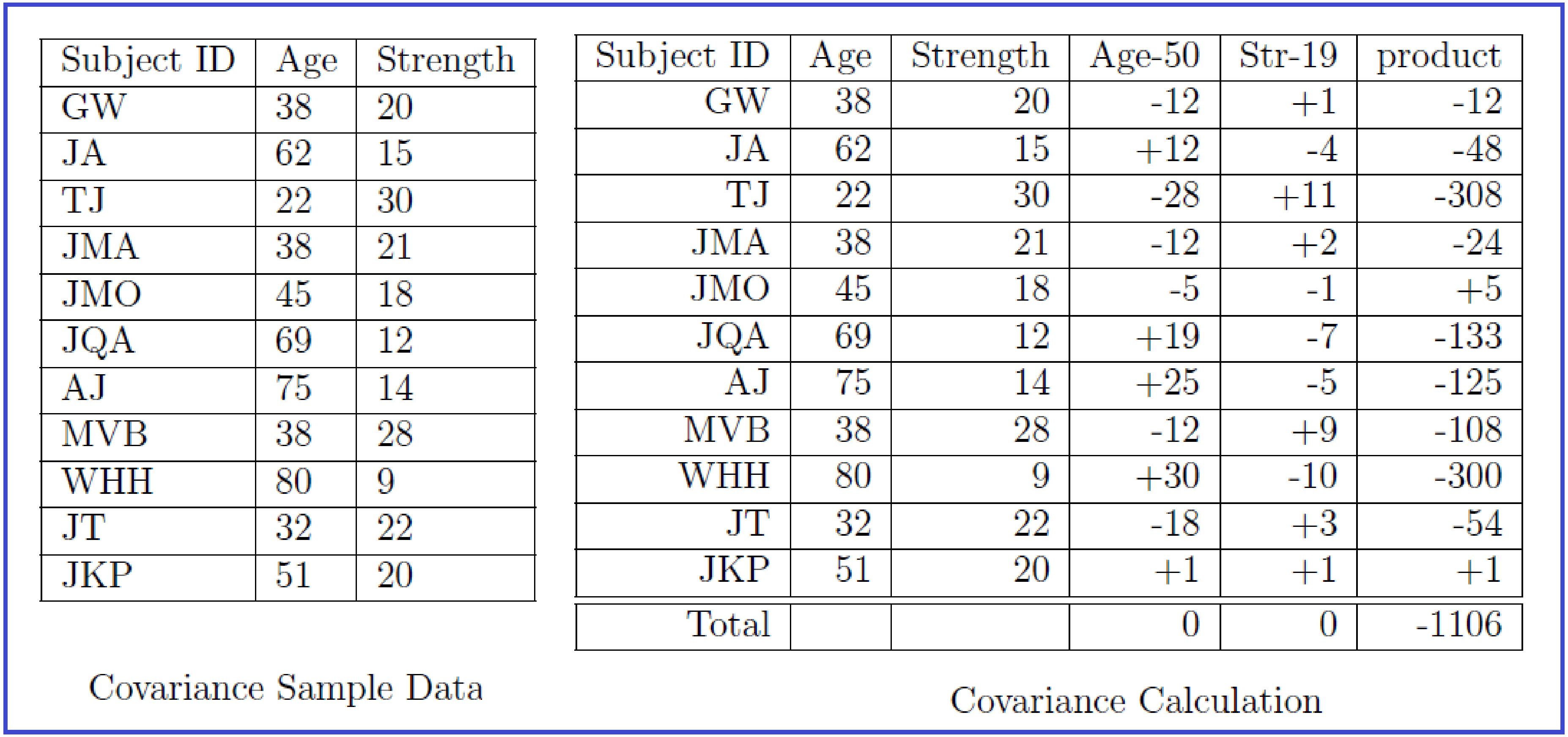
It is worth noting that $\text{Cov}(X,X) = \text{Var}(X)$.



	X	Y	Z
X	5.00	1.77	-2.24
Y	1.77	7.0	3.17
Z	-2.24	3.17	4.0

A Covariance Matrix

Multivariate non-graphical EDA



Subject ID	Age	Strength
GW	38	20
JA	62	15
TJ	22	30
JMA	38	21
JMO	45	18
JQA	69	12
AJ	75	14
MVB	38	28
WHH	80	9
JT	32	22
JKP	51	20

Covariance Sample Data

Subject ID	Age	Strength	Age-50	Str-19	product
GW	38	20	-12	+1	-12
JA	62	15	+12	-4	-48
TJ	22	30	-28	+11	-308
JMA	38	21	-12	+2	-24
JMO	45	18	-5	-1	+5
JQA	69	12	+19	-7	-133
AJ	75	14	+25	-5	-125
MVB	38	28	-12	+9	-108
WHH	80	9	+30	-10	-300
JT	32	22	-18	+3	-54
JKP	51	20	+1	+1	+1
Total			0	0	-1106

Covariance Calculation

- ❑ Since n=11, the covariance of x and y is -1106/10= **-110.6.**
- ❑ The fact that the covariance is negative indicates that as age goes up strength tends to go down (and vice versa).

Multivariate non-graphical EDA

Correlations:

❑ Covariances tend to be hard to interpret

❑ Often use correlation instead.

❑ Correlation has property that it is always between -1 and +1

❑ With -1 being a perfect negative linear correlation,

❑ With +1 being a perfect positive linear correlation

❑ With 0 indicating that X and Y are uncorrelated.

❑ The symbol **r or $r_{x;y}$** is often used for sample correlations.

Multivariate non-graphical EDA

Correlations:

□ The formula for the sample correlation is

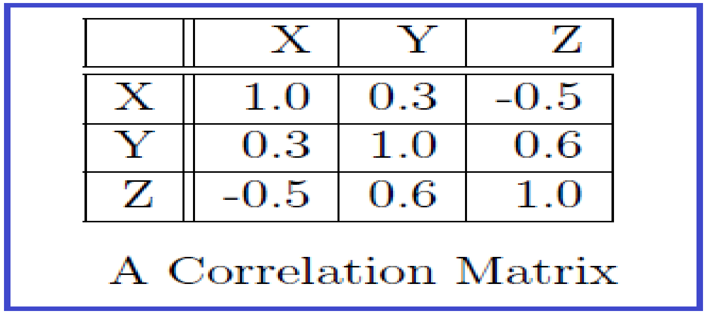
$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{s_x s_y}$$

where **s_x** is standard deviation of X and **s_y** is the standard deviation of Y

□ In above example, s_x = 18:96, s_y = 6:39,

$$r = \frac{-110.6}{18{:}96 \times 6{:}39} = -\,0{:}913$$

□ This is a strong **negative** correlation.

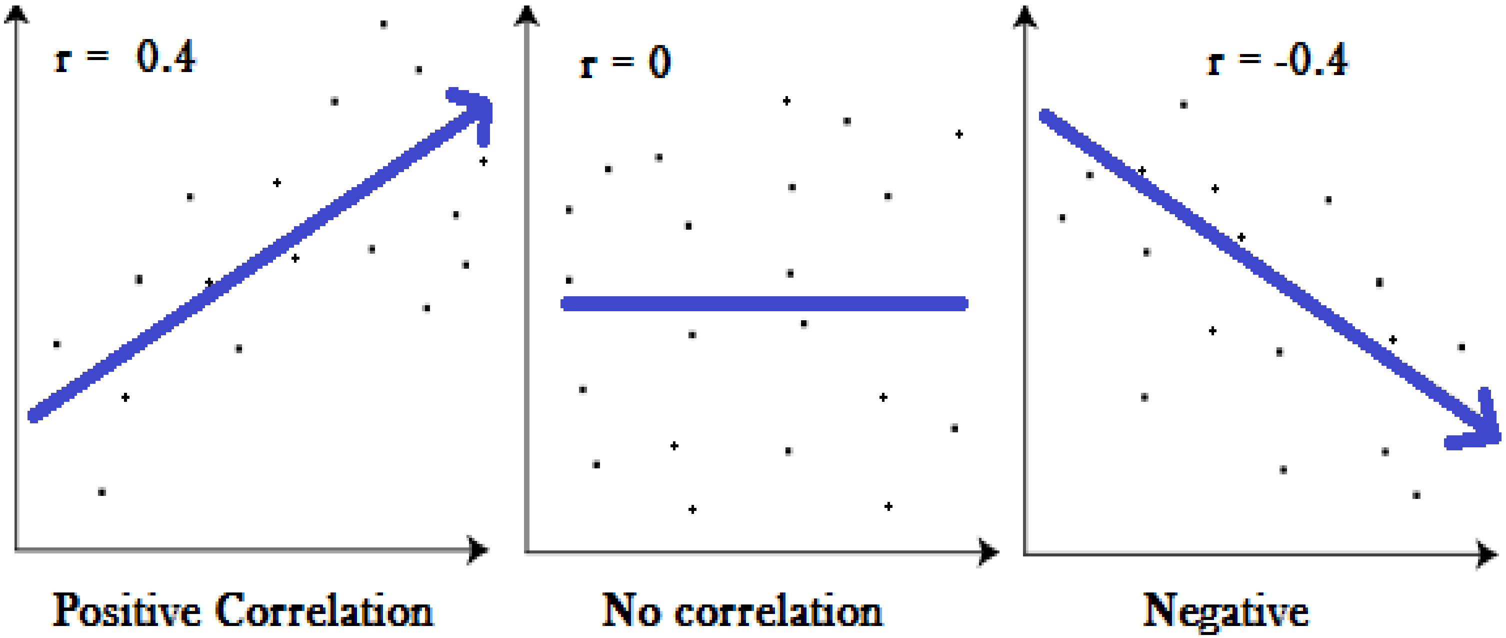


	X	Y	Z
X	1.0	0.3	-0.5
Y	0.3	1.0	0.6
Z	-0.5	0.6	1.0

A Correlation Matrix

Multivariate non-graphical EDA

Correlations:



The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

Multivariate graphical EDA

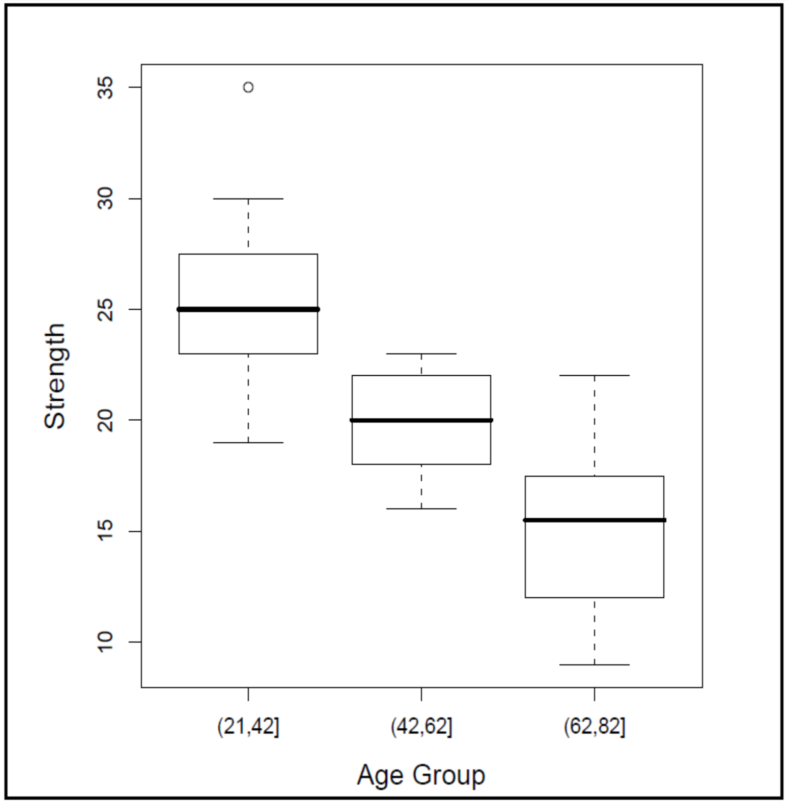
- ❑ There are few useful techniques for graphical EDA of two categorical random variables.
- ❑ One used commonly is a grouped bar plot with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

Multivariate graphical EDA

Univariate graphs by category

- ❑ When we have one categorical (usually explanatory) and one quantitative (usually outcome) variable, graphical EDA usually takes the form of “conditioning" on the categorical random variable.
- ❑ This simply indicates that we focus on all of subjects with a particular level of the categorical random variable, then make plots of quantitative variable for those subjects.
- ❑ Repeat this for each level of the categorical variable, then compare the plots.
- ❑ The most commonly used of these are side-by-side boxplots

Multivariate graphical EDA



- ❑You can see the downward trend in the median as the ages increase.
- ❑The spreads (IQRs) are similar for the three groups.
- ❑All three groups are roughly symmetrical with one high strength outlier in the youngest age group.

Side-by-side boxplots are best graphical EDA technique for examining relationship between a categorical variable and a quantitative variable, as well as distribution of quantitative variable at each level of categorical variable.

Multivariate graphical EDA

Scatterplots

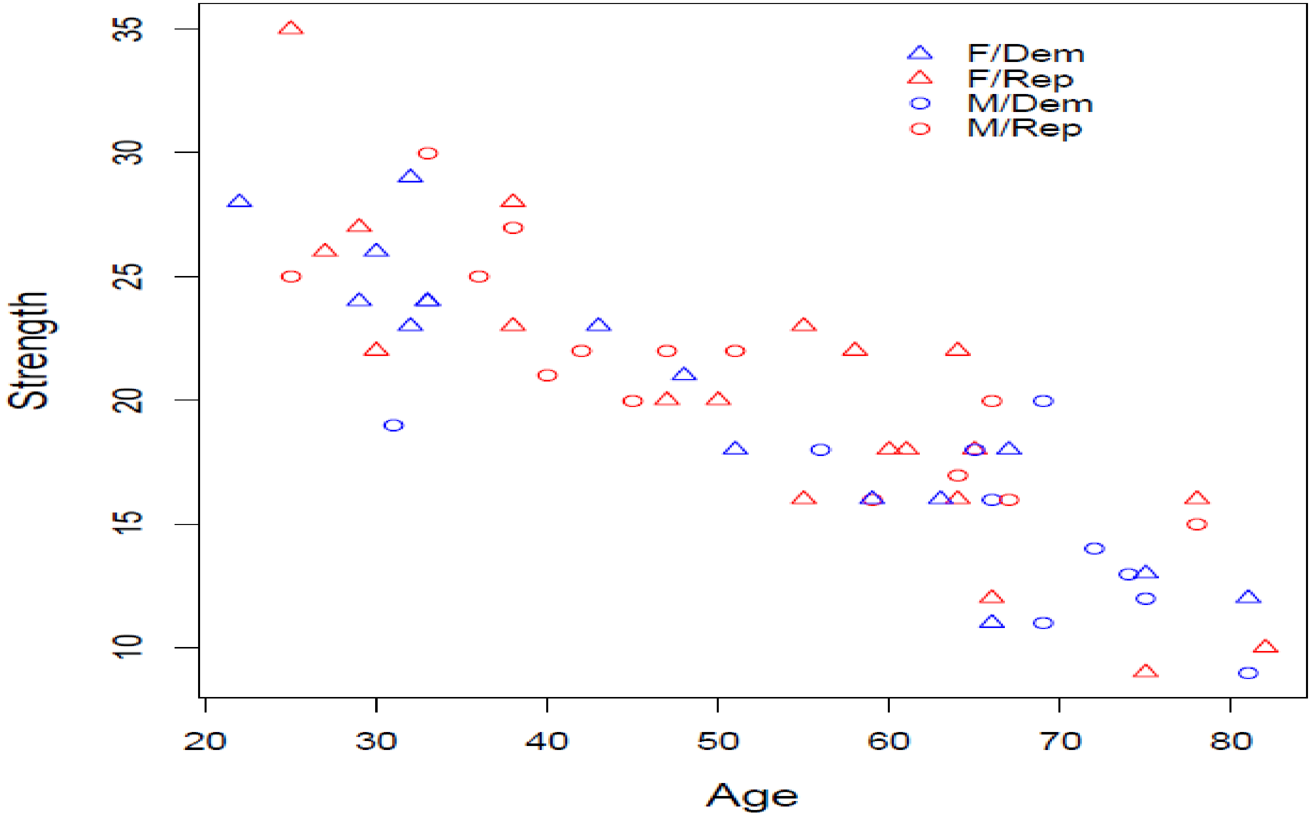
- For two quantitative variables, the basic graphical EDA technique is the scatterplot which has one variable on the x-axis, one on the y-axis and a point for each case in your dataset.

- **If one variable is explanatory and the other is outcome, it is a very, very strong convention to put the outcome on the y (vertical) axis.**

- One or two additional categorical variables can be accommodated on scatterplot by encoding additional information in the symbol type and/or color.

Multivariate graphical EDA

Scatterplots



scatterplot with two additional variables.

Age vs. strength is shown, and different colors and symbols are used to code political party and gender.

EDA

In a nutshell:

- ❑ You should always perform appropriate EDA before further analysis of your data.
- ❑ Perform whatever steps are necessary to become more familiar with your data, check for obvious mistakes, learn about variable distributions, and learn about relationships between variables.
- ❑ EDA is not an exact science - it is a very important art!

EDA - Example

- Graphs, plots, and tables often uncover important relationships.

- Relationships that could indicate important areas for further investigation.

- We will use exploratory methods to delve into the **churn data set** from the **UCI Repository of Machine Learning Databases** at the University of California

- **Churn**, also called attrition, is a term used to indicate a customer leaving the service of company.

Churn data set

The data set contains **20 predictors**.

- *State*: Categorical, for the 50 states and the District of Columbia.
- *Account length:* Integer-valued, how long account has been active.
- *Area code:* Categorical
- *Phone number*: Essentially a surrogate for customer ID.
- *International plan*: categorical, yes or no.
- *Voice mail plan:* categorical, yes or no.
- *Number of voice mail messages:* Integer-valued.
- *Total day minutes:* Continuous, minutes customer used service during the day.
- *Total day calls:* Integer-valued.
- *Total day charge:* Continuous, perhaps based on above two variables.

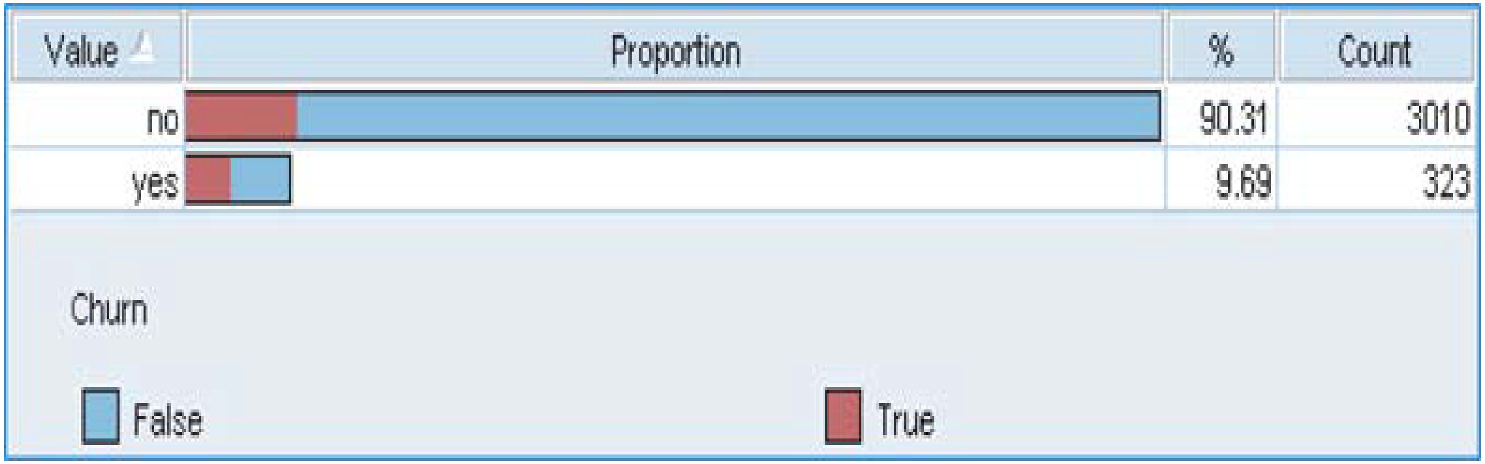
Churn data set

- *Total eve minutes*: Continuous, minutes customer used service during the evening.
- *Total eve calls*: Integer-valued.
- *Total eve charge*: Continuous, based on above two variables.
- *Total night minutes*: Continuous, minutes customer used service during the night.
- *Total night calls*: Integer-valued.
- *Total night charge*: Continuous, perhaps based on above two variables.
- *Total international minutes*: Continuous, minutes customer used service to make international calls.
- *Total international calls*: Integer-valued.
- *Total international charge*: Continuous, based on above two variables.
- *Number of calls to customer service*: Integer-valued.
- ***Churn***: **Target**. Indicator of whether customer has left company **(true or false).**

EXPLORING CATEGORICAL VARIABLES

Investigation of categorical variable *International Plan*

Comparison bar chart of churn proportions, by international plan participation



Value	Proportion	%	Count
no		90.31	3010
yes		9.69	323

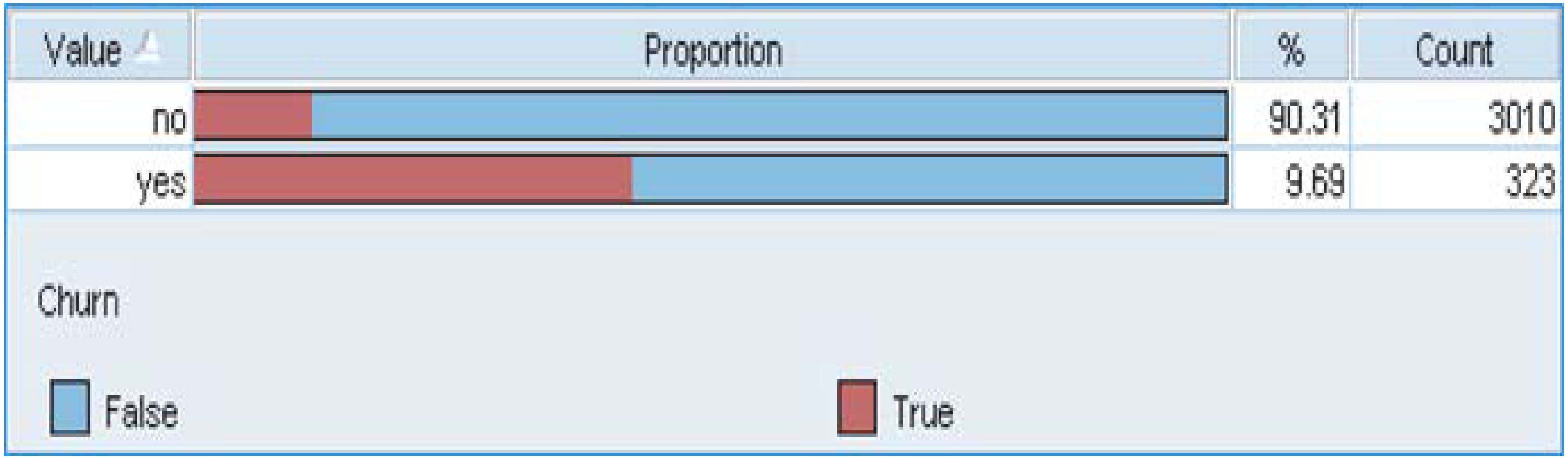
Churn

🟦 False 🟥 True

Greater proportion of International Plan holders are churning, but it is difficult to be sure.

EXPLORING CATEGORICAL VARIABLES

Comparison bar chart of churn proportions, by international plan participation, with equal bar length.



Value	Proportion	%	Count
no		90.31	3010
yes		9.69	323

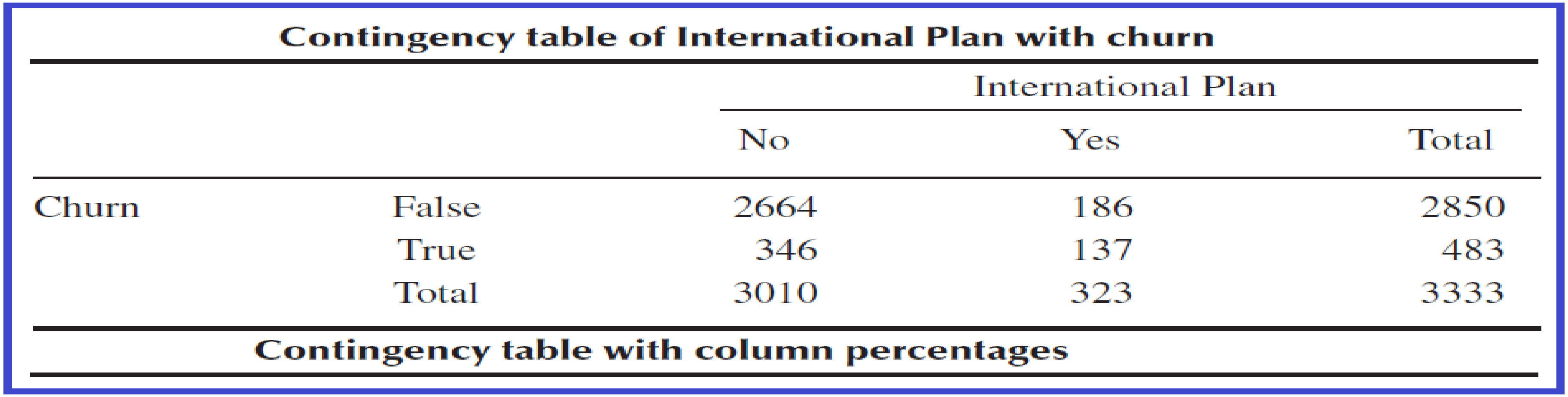
Churn

🟦 False 🟥 True

- ➢Clearly, those who have selected the International Plan have a greater chance of leaving the company’s service

EXPLORING CATEGORICAL VARIABLES

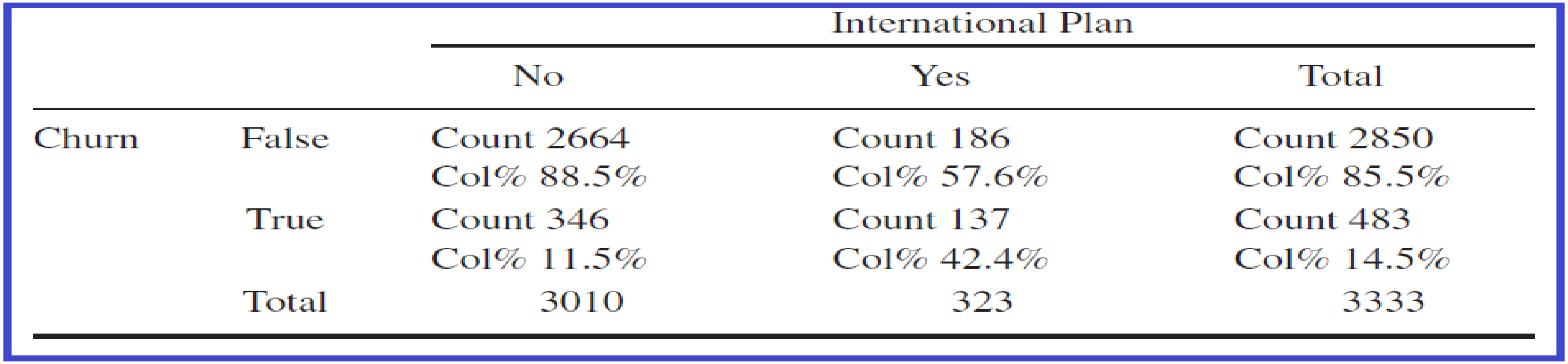
- Graphics above tell us that International Plan holders tend to churn more frequently, but they do not **quantify the relationship**
- Use a contingency table as both variables are categorical



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

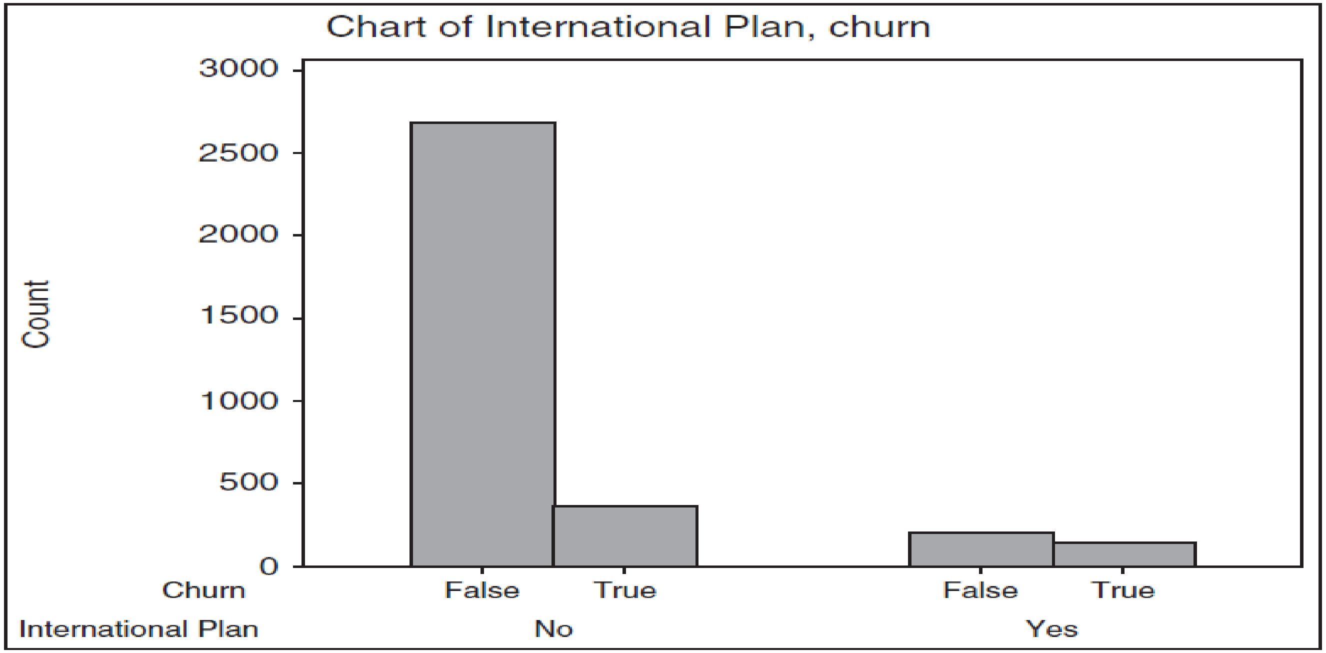
Contingency table with column percentages



		International Plan		
		No	Yes	Total
Churn	False	Count 2664 Col% 88.5%	Count 186 Col% 57.6%	Count 2850 Col% 85.5%
	True	Count 346 Col% 11.5%	Count 137 Col% 42.4%	Count 483 Col% 14.5%
	Total	3010	323	3333

EXPLORING CATEGORICAL VARIABLES

The graphical counterpart of the contingency table is the ***clustered bar chart.***

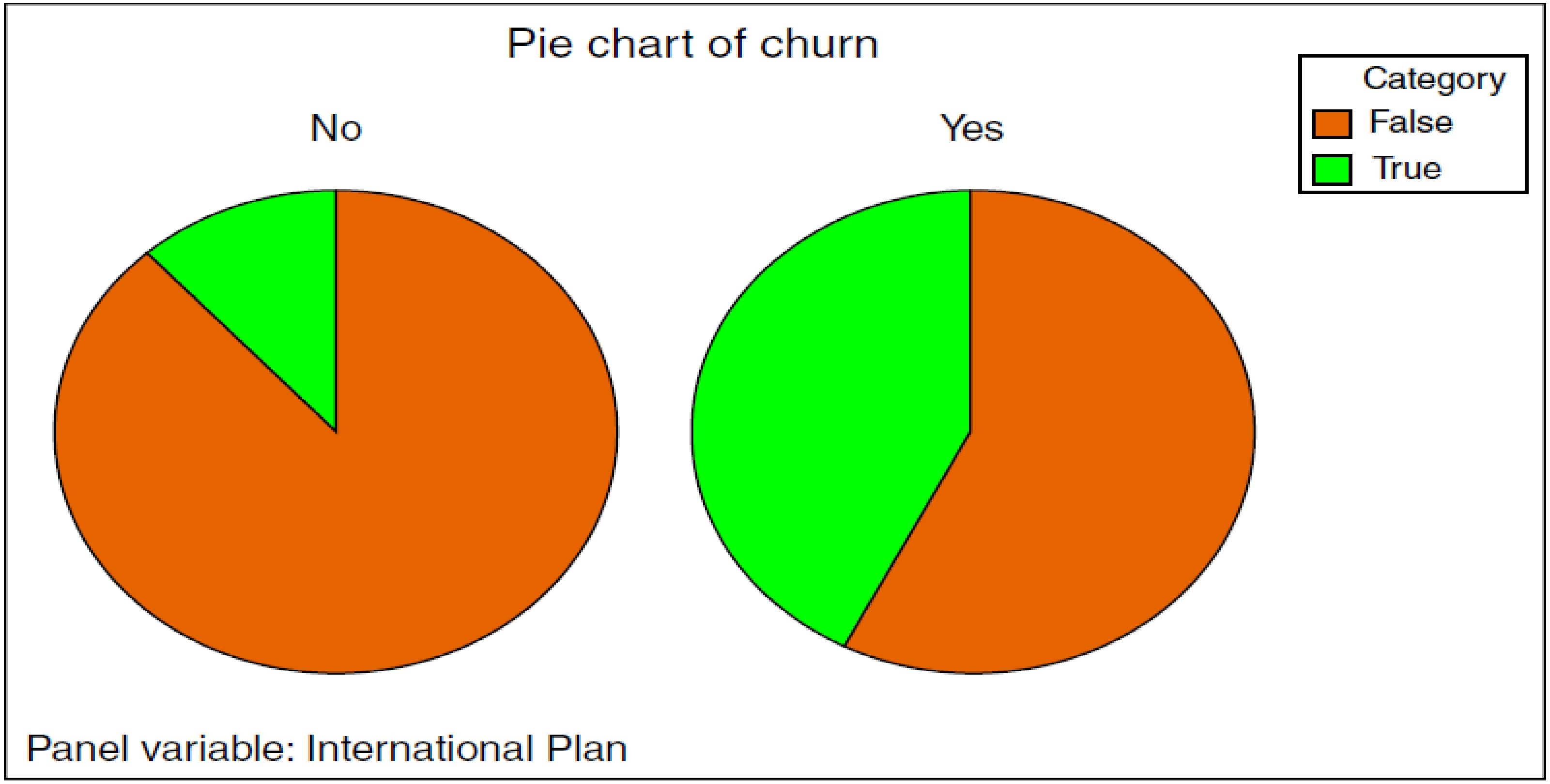


The clustered bar chart is the graphical counterpart of the contingency table.

Clearly, the proportion of churners is greater among those belonging to the **International plan**.

EXPLORING CATEGORICAL VARIABLES

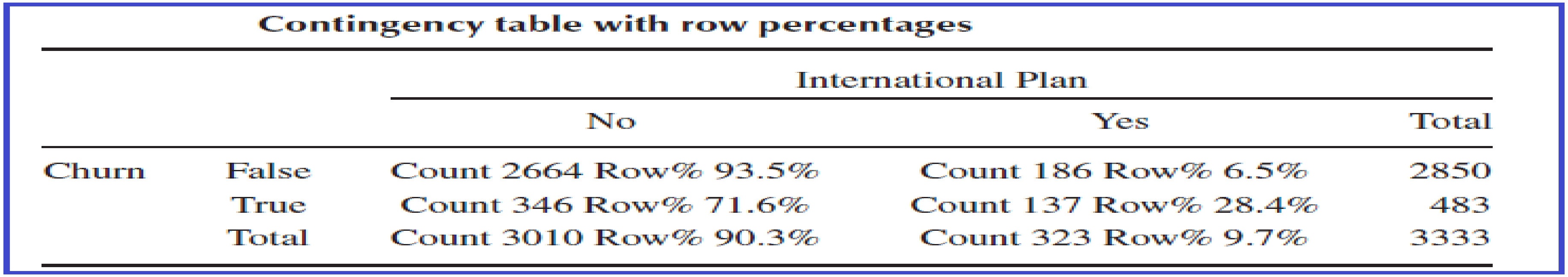
Another useful graphic for comparing two categorical variables is the *comparative pie chart*.



Comparative pie chart associated with Table 3.2.

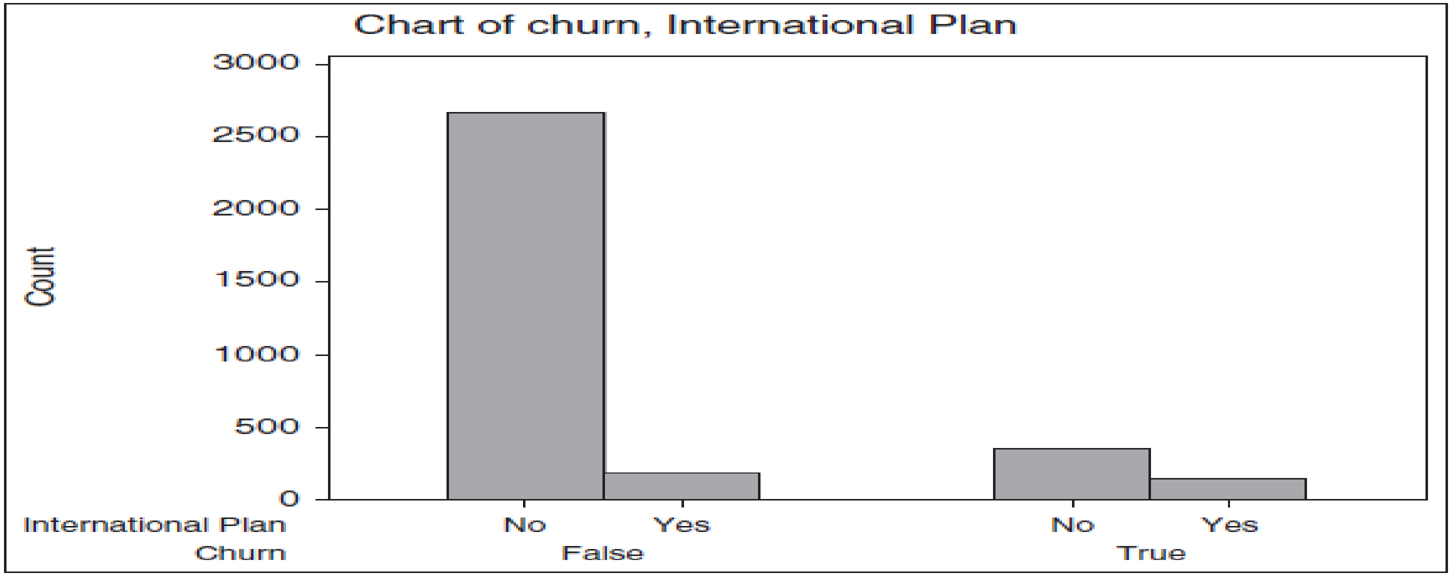
EXPLORING CATEGORICAL VARIABLES

Contrast with prev. Table, the contingency table with *row percentages*



Contingency table with row percentages

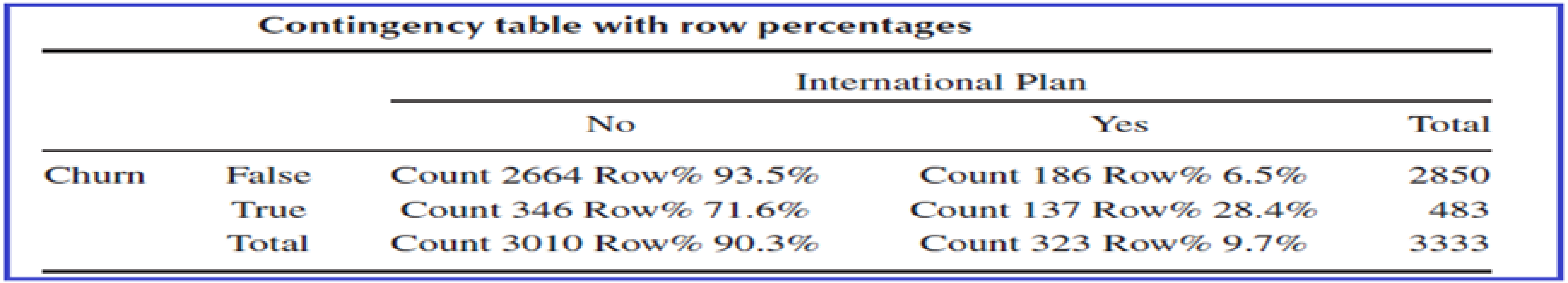
		International Plan		
		No	Yes	Total
Churn	False	Count 2664 Row% 93.5%	Count 186 Row% 6.5%	2850
	True	Count 346 Row% 71.6%	Count 137 Row% 28.4%	483
	Total	Count 3010 Row% 90.3%	Count 323 Row% 9.7%	3333



Clustered bar chart

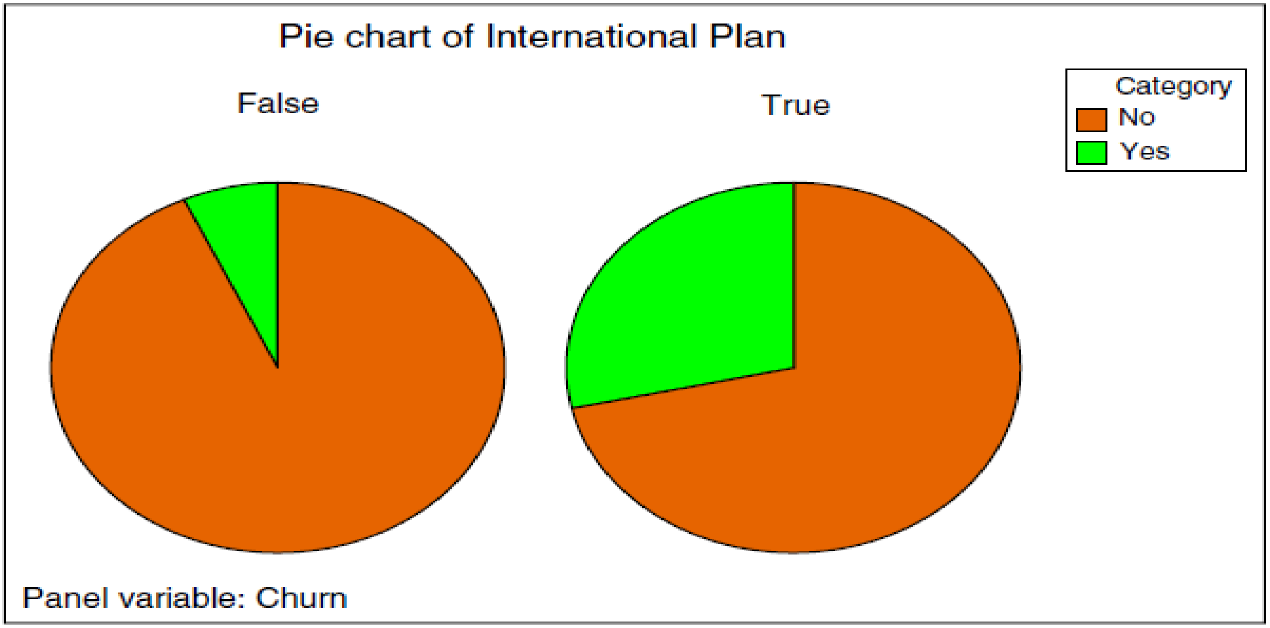
Proportion of International Plan holders is greater among churners

EXPLORING CATEGORICAL VARIABLES



Contingency table with row percentages

		International Plan		
		No	Yes	Total
Churn	False	Count 2664 Row% 93.5%	Count 186 Row% 6.5%	2850
	True	Count 346 Row% 71.6%	Count 137 Row% 28.4%	483
	Total	Count 3010 Row% 90.3%	Count 323 Row% 9.7%	3333



Comparative pie chart

Comparative pie chart associated with above Table

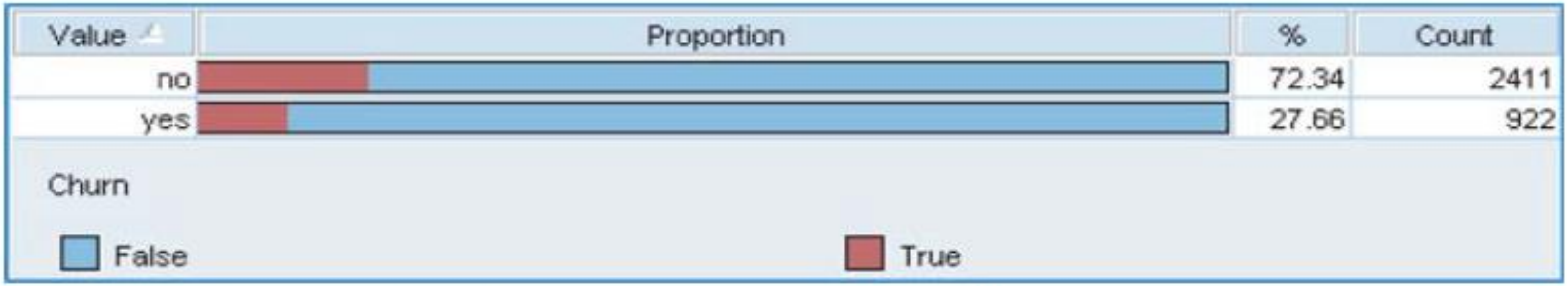
EXPLORING CATEGORICAL VARIABLES

To summarize, this EDA on the International Plan has indicated that

1. perhaps we should investigate what is it about our international plan that is inducing our customers to leave;
2. we should expect that, whatever data mining/machine learning algorithms we use to predict churn, the model will **probably include** whether or not the customer selected the International Plan.

EXPLORING CATEGORICAL VARIABLES

Let us now turn to the **Voice Mail Plan**

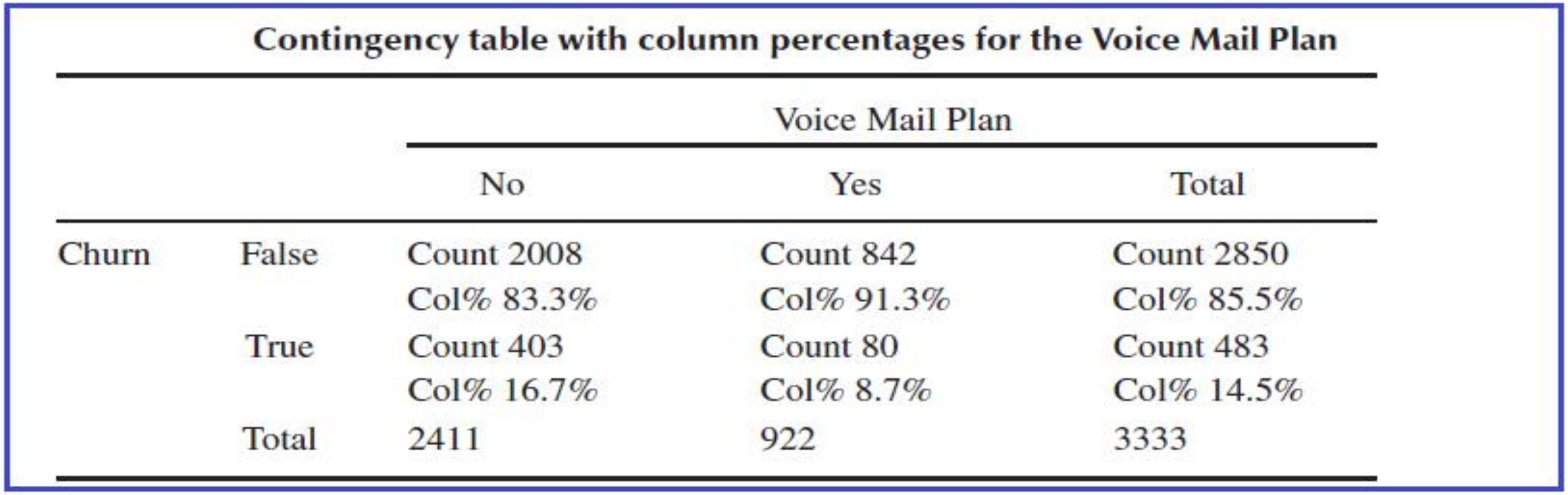


Value	Proportion	%	Count
no		72.34	2411
yes		27.66	922

Churn

🟦 False 🟥 True

Those without the voice mail plan are more likely to churn.



Contingency table with column percentages for the Voice Mail Plan

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

Without the Voice Mail Plan are churners, as compared to customers who do have the Voice Mail Plan.

EXPLORING CATEGORICAL VARIABLES

To summarize, this EDA on the Voice Mail Plan has indicated that

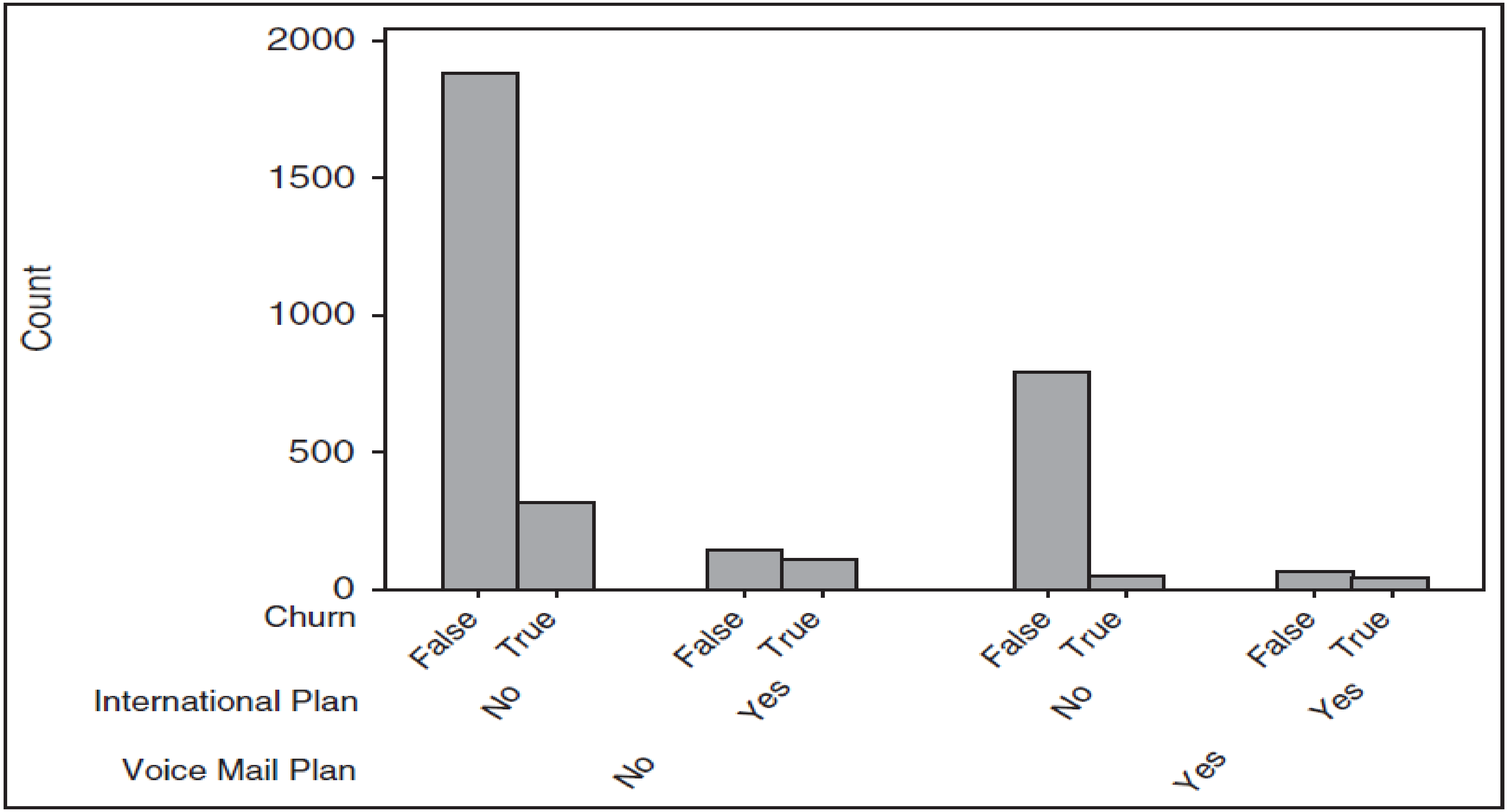
1. perhaps we should enhance our Voice Mail Plan still further, or make it easier for customers to join it, as an instrument for increasing customer loyalty;

2. whatever data mining algorithms/machine learning we use to predict churn, the model will **probably include** whether or not the customer selected the Voice Mail Plan

- **confidence in this expectation is perhaps not quite as high as for the International Plan**

EXPLORING CATEGORICAL VARIABLES

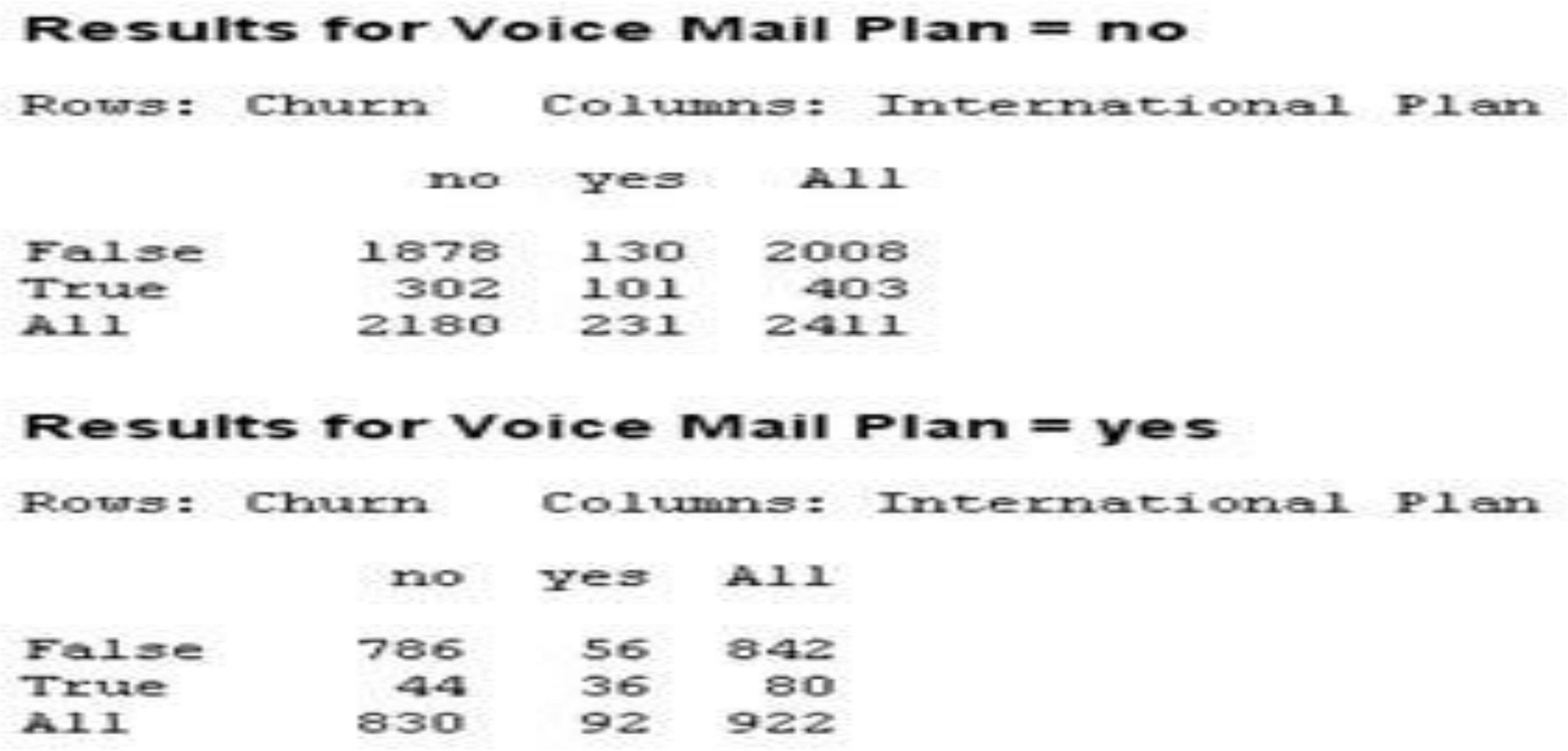
- May also explore the ***two-way interactions*** among categorical variables with respect to ***churn***.



Multilayer clustered bar chart.

EXPLORING CATEGORICAL VARIABLES

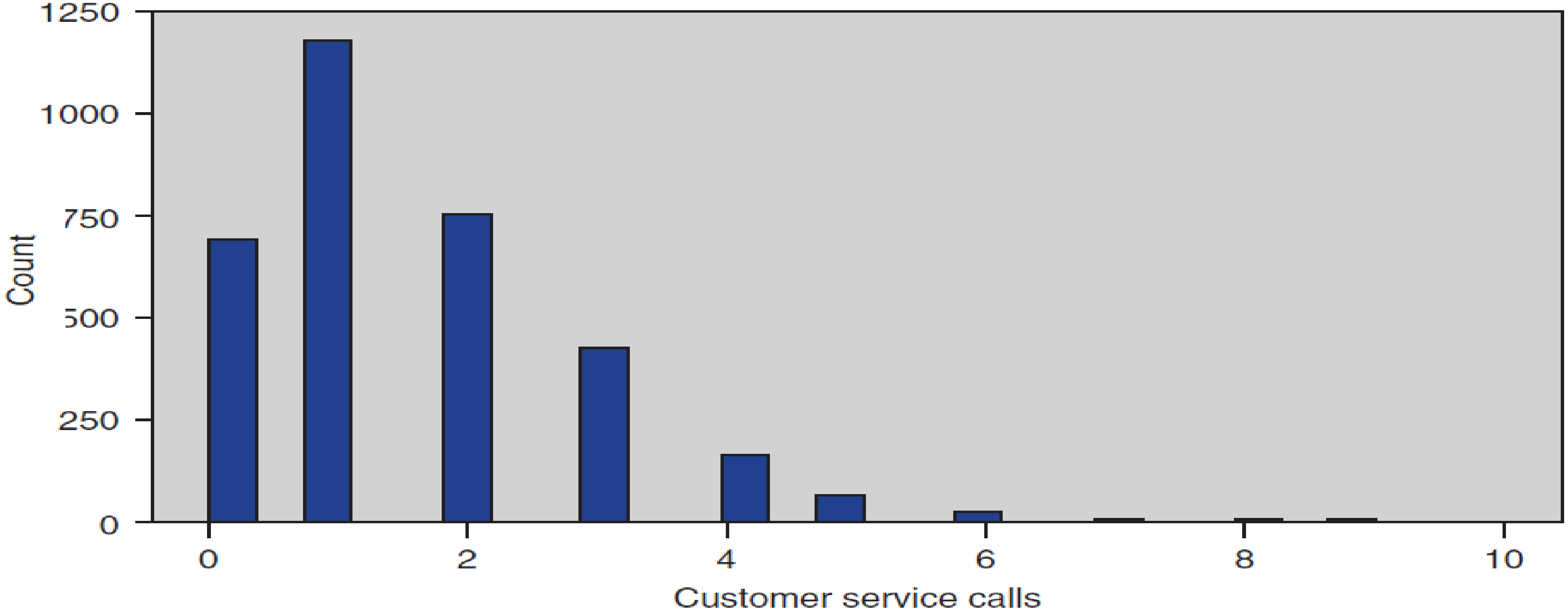
Statistics for multilayer clustered bar chart



```
Results for Voice Mail Plan = no

Rows: Churn    Columns: International Plan

           no   yes   All

False    1878   130  2008
True      302   101   403
All      2180   231  2411

Results for Voice Mail Plan = yes

Rows: Churn    Columns: International Plan

          no  yes  All

False    786   56  842
True      44   36   80
All      830   92  922
```

EXPLORING NUMERIC VARIABLES

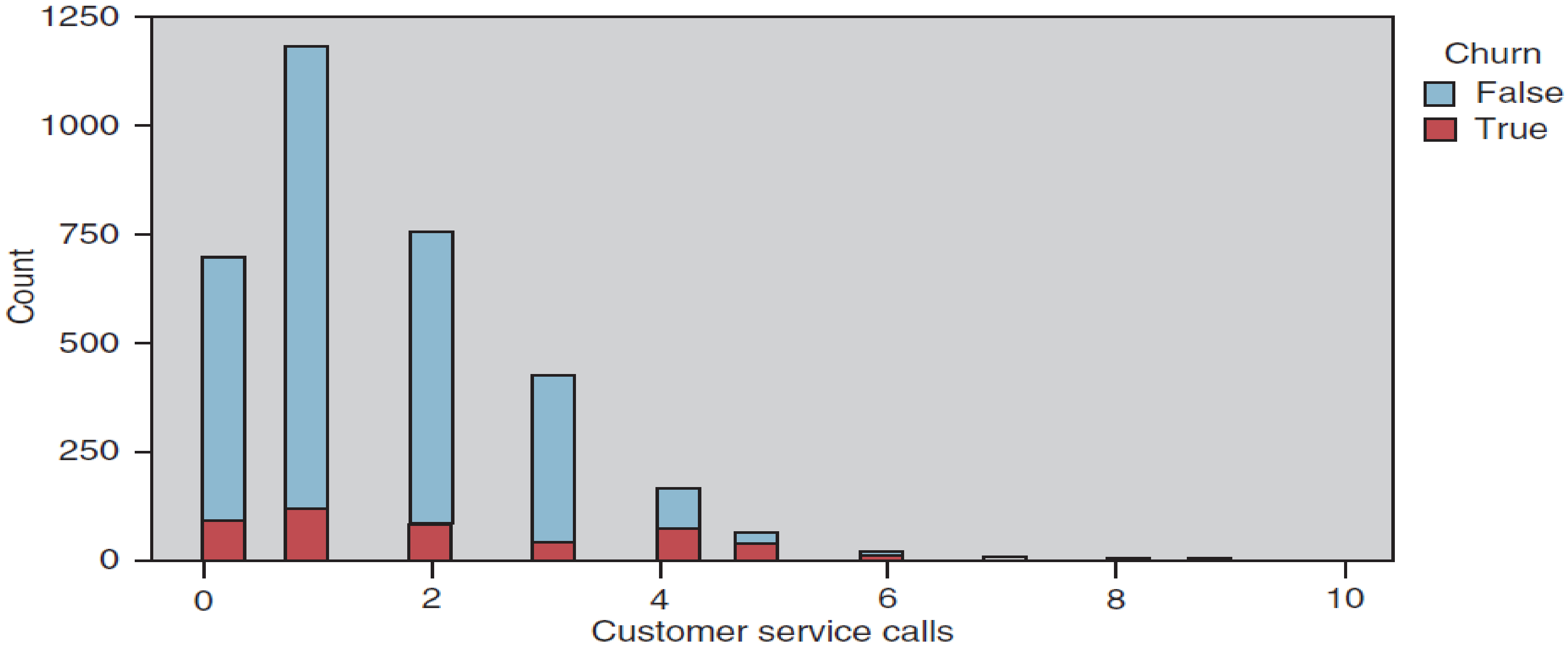
- Next, we turn to an exploration of the numeric predictive variables.
- Unfortunately, the usual type of histogram does not help us determine whether the predictor variables are associated with the target variable.



Histogram of customer service calls with no overlay.

EXPLORING NUMERIC VARIABLES

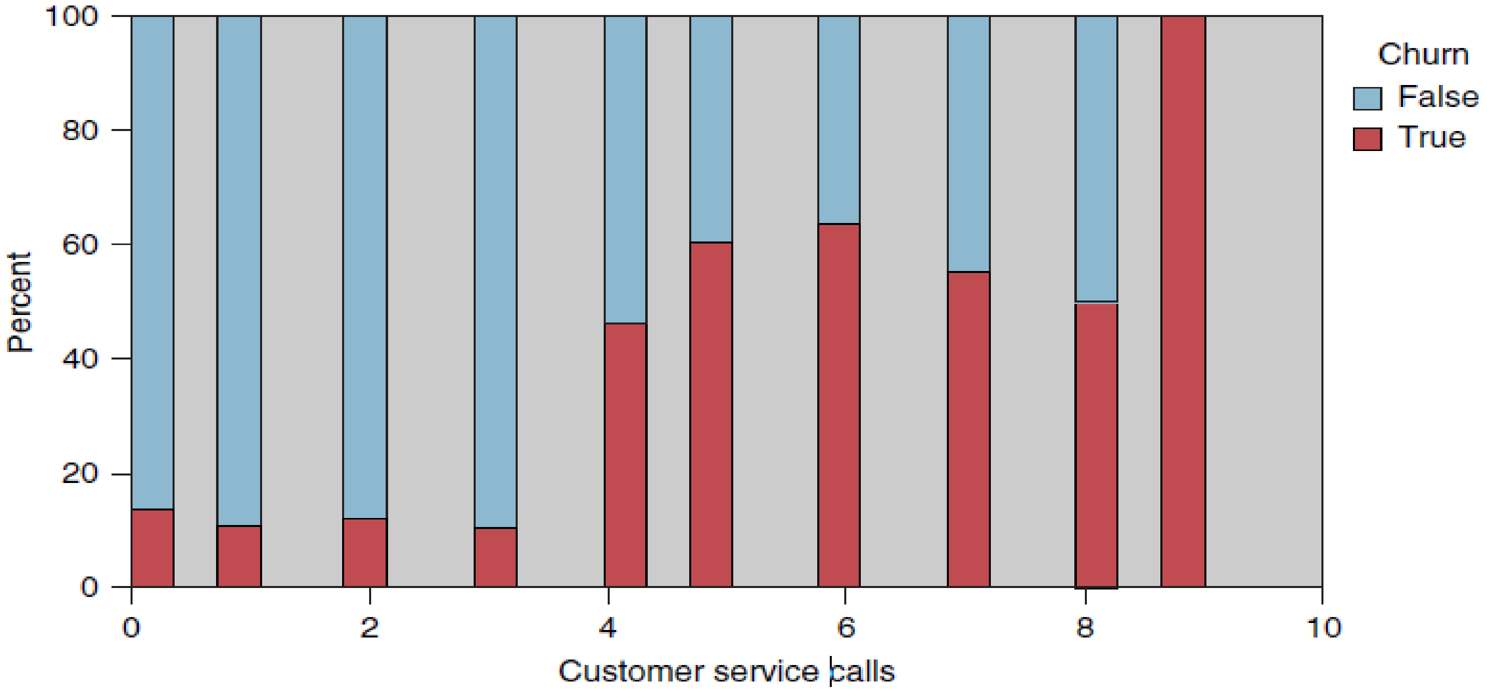
- Next, we turn to an exploration of the numeric predictive variables
- To explore whether a predictor is useful for predicting the target variable, use an overlay histogram,
- Which is a histogram where the rectangles are colored according to the values of the target variable.



Histogram of customer service calls with churn overlay.

EXPLORING NUMERIC VARIABLES

“stretching out” the rectangles that have low counts enables better definition and contrast.



“Normalized” histogram of customer service calls with churn overlay.

Customer called three times or less - lower churn rate
Customers called four or more times – higher churn rate .

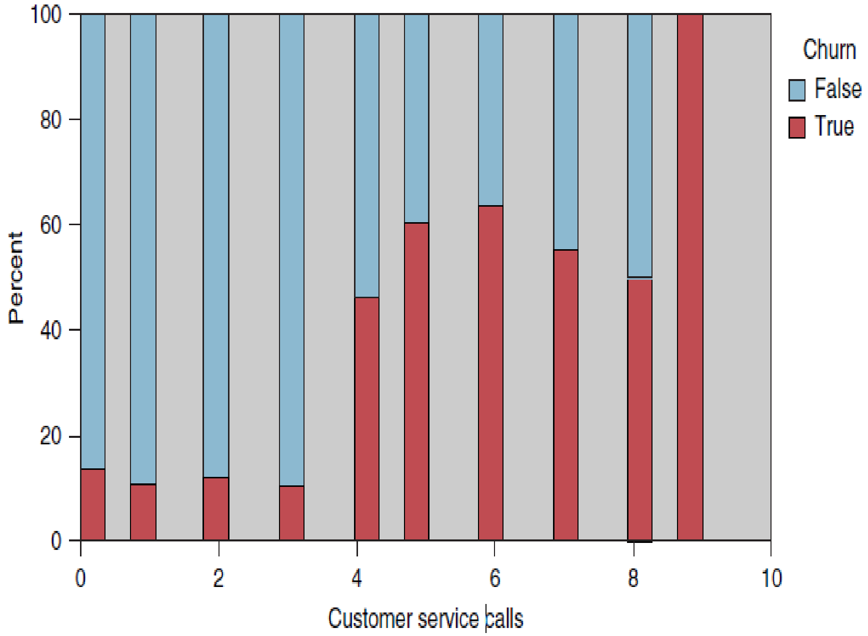
EXPLORING NUMERIC VARIABLES

This EDA on the customer service calls has indicated that

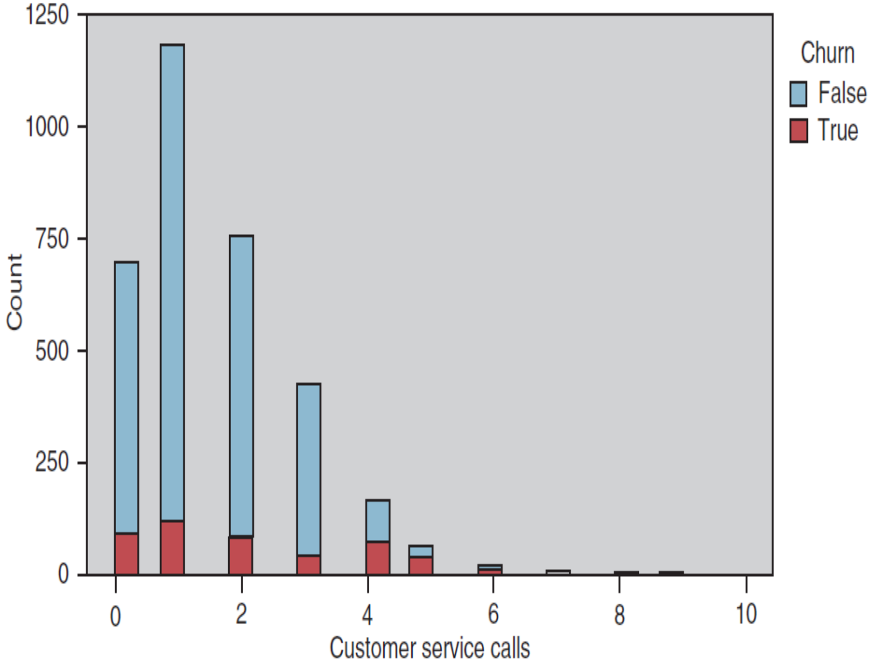
1. Carefully track the number of customer service calls made by each customer. By the third call, specialized incentives should be offered to retain customer loyalty, because, by the fourth call, the probability of churn increases greatly;
2. Whatever algorithms we use to predict churn, the model will **probably include** the number of customer service calls made by the customer.

EXPLORING NUMERIC VARIABLES

Important note: Data analysts always provide a **non-normalized histogram** along with the normalized histogram, because the normalized histogram does not provide any information on the frequency distribution of the variable.



"Normalized" histogram of customer service calls with churn overlay.



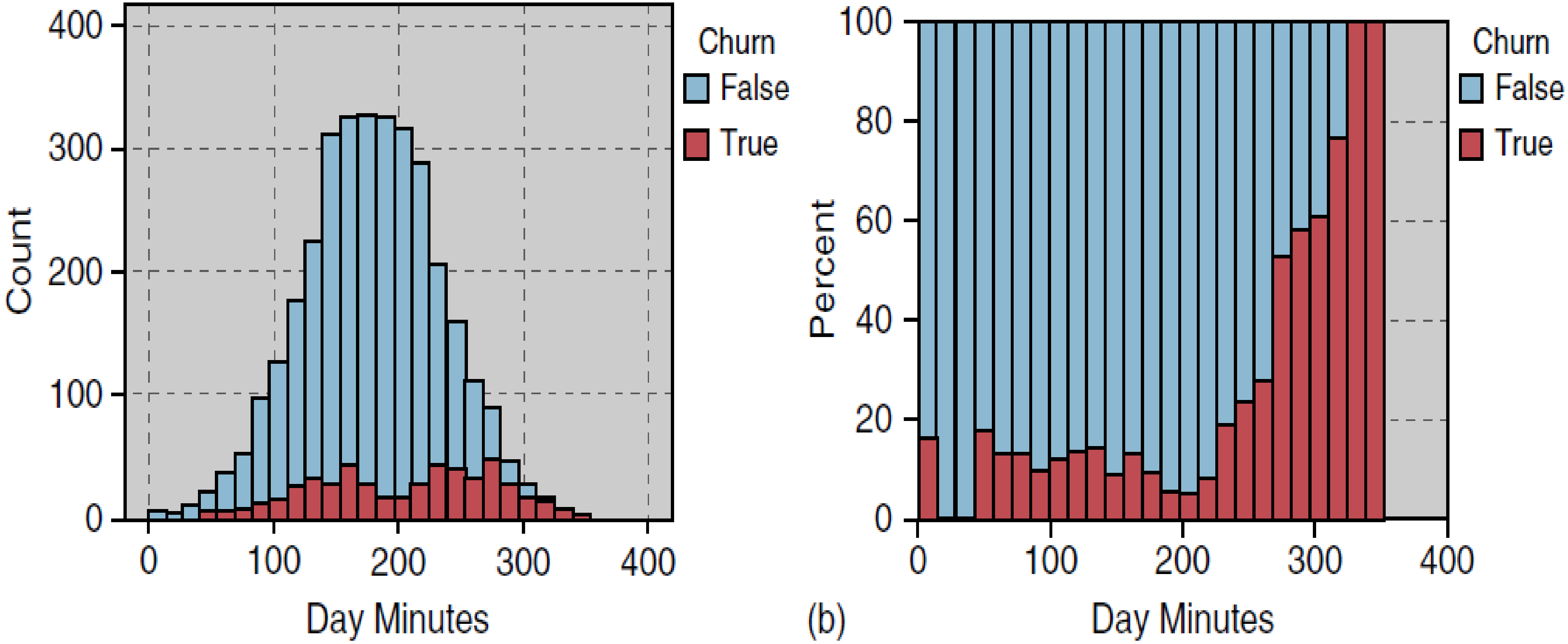
Histogram of customer service calls with churn overlay.

Indicates that the churn rate for customers logging nine service calls is 100%;

Shows that there are only two customers with this number of calls

EXPLORING NUMERIC VARIABLES

Let us now turn to the Day Minutes



(a) Non-normalized histogram of day minutes. (b) Normalized histogram of day

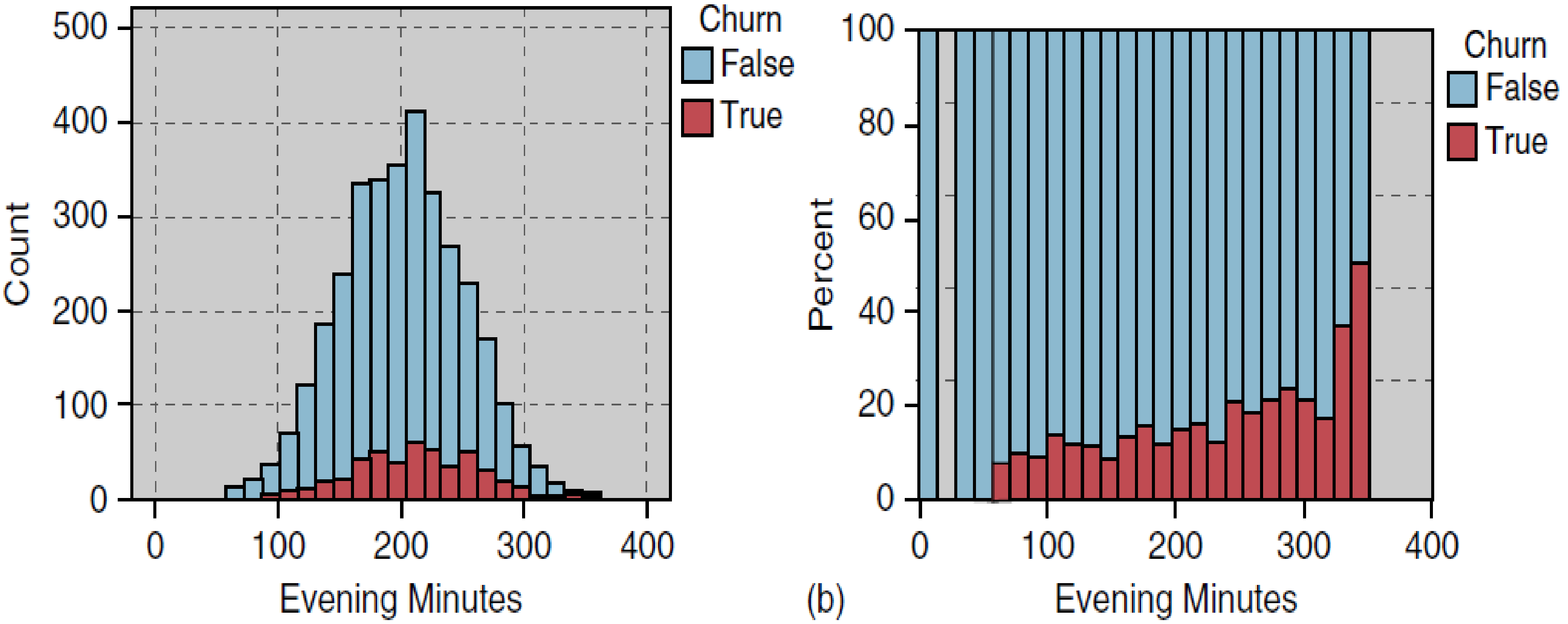
EXPLORING NUMERIC VARIABLES

The normalized histogram of ***Day Minutes*** shows that high day-users tend to churn at a higher rate. Therefore,

1. we should carefully track the number of day minutes used by each customer. As the number of day minutes passes 200, we should consider special incentives;
2. we should investigate why heavy day-users are tempted to leave;
3. we should expect that our eventual model will **include** *day minutes* as a **predictor of churn**.

EXPLORING NUMERIC VARIABLES

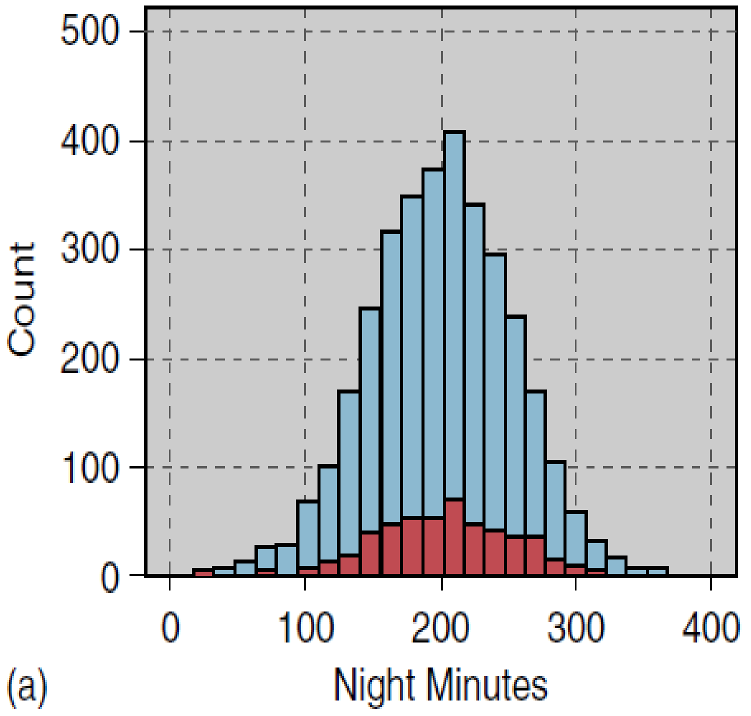
- **slight** tendency for customers with higher *evening minutes* to churn



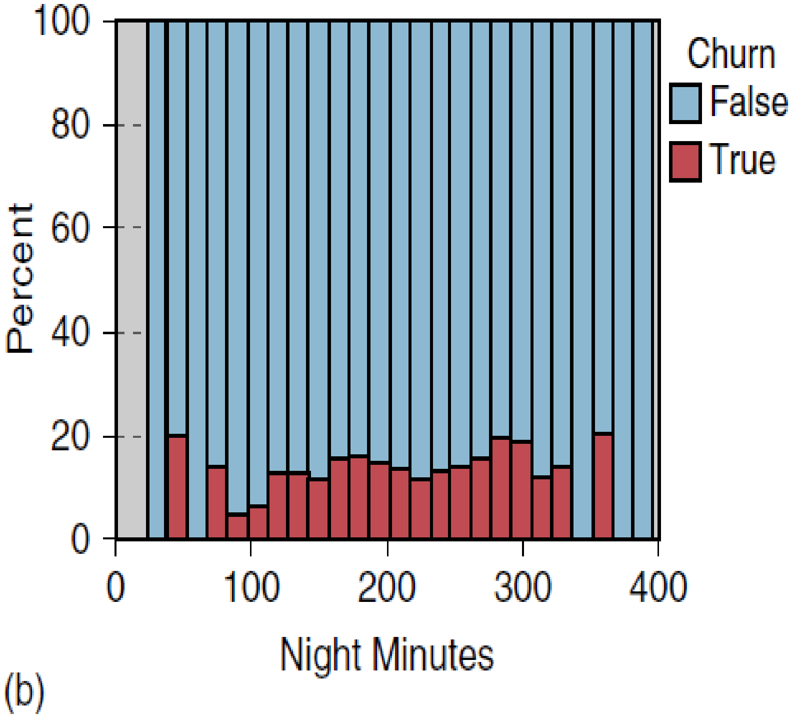
(a) Non-normalized histogram of evening minutes. (b) Normalized histogram of evening minutes.

EXPLORING NUMERIC VARIABLES

Graph indicates that there is no obvious association between churn and night minutes



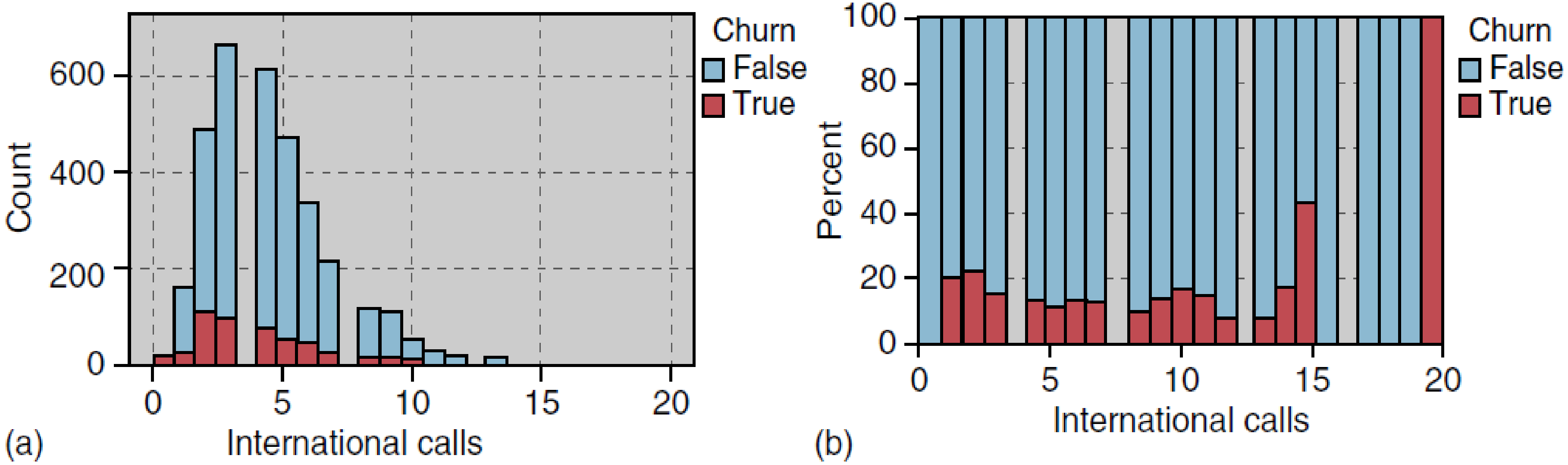
(a) Non-normalized histogram of night minutes.



(b) Normalized histogram of night minutes.

EXPLORING NUMERIC VARIABLES

The lack of obvious association at the EDA stage between a predictor and a target variable is not sufficient reason to omit that predictor from the model.

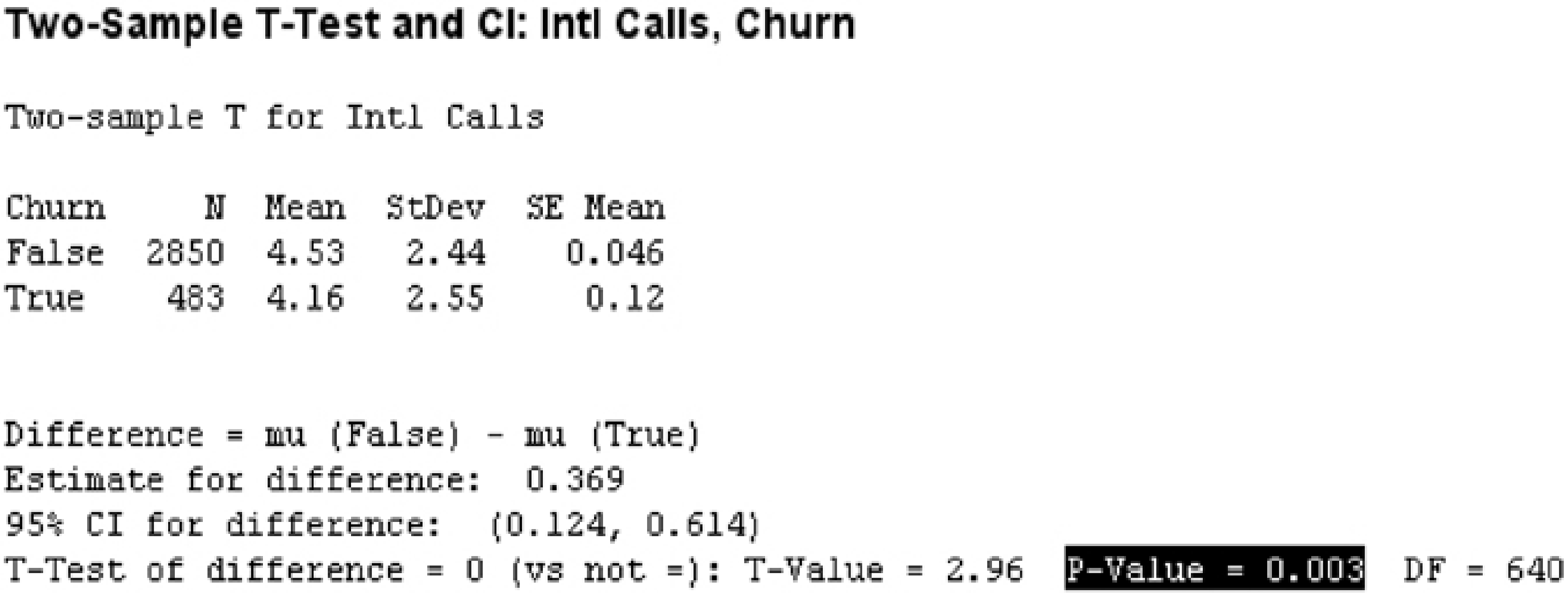


(a) Non-normalized histogram of *international calls*. (b) Normalized histogram of *international calls*.

predictor International Calls with churn overlay, do not indicate strong graphical evidence of predictive importance of International Calls.

EXPLORING NUMERIC VARIABLES

- However, a *t*-test for the difference in mean number of international calls for churners and non-churners is statistically significant
- This variable is indeed useful for predicting churn:
- Churners tend to place a lower mean number of international calls



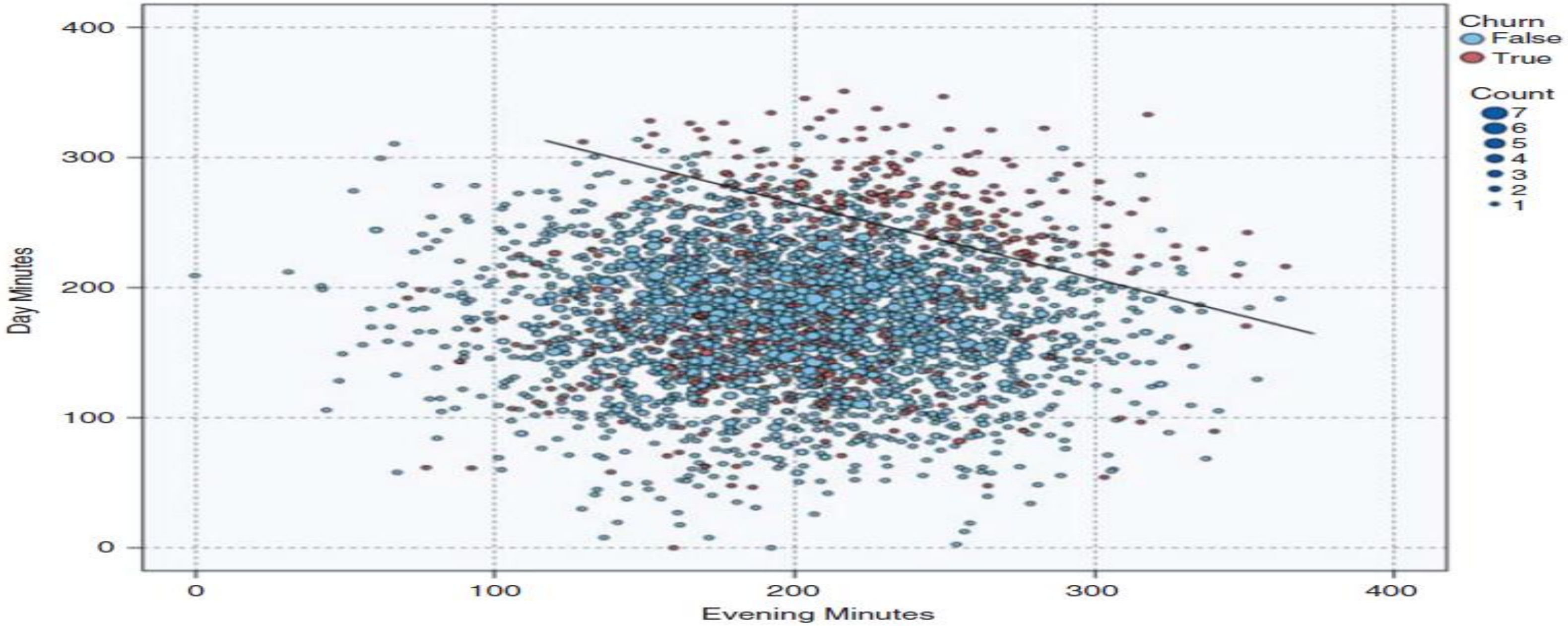
Two-Sample T-Test and CI: Intl Calls, Churn

```
Two-sample T for Intl Calls

Churn      N  Mean  StDev  SE Mean
False   2850  4.53   2.44    0.046
True     483  4.16   2.55     0.12


Difference = mu (False) - mu (True)
Estimate for difference:  0.369
95% CI for difference:  (0.124, 0.614)
T-Test of difference = 0 (vs not =): T-Value = 2.96  P-Value = 0.003  DF = 640
```

- **Omitting international calls – would have committed a mistake**
- **A hypothesis test, such as this t-test lies beyond the scope of EDA**

EXPLORING MULTIVARIATE RELATIONSHIPS

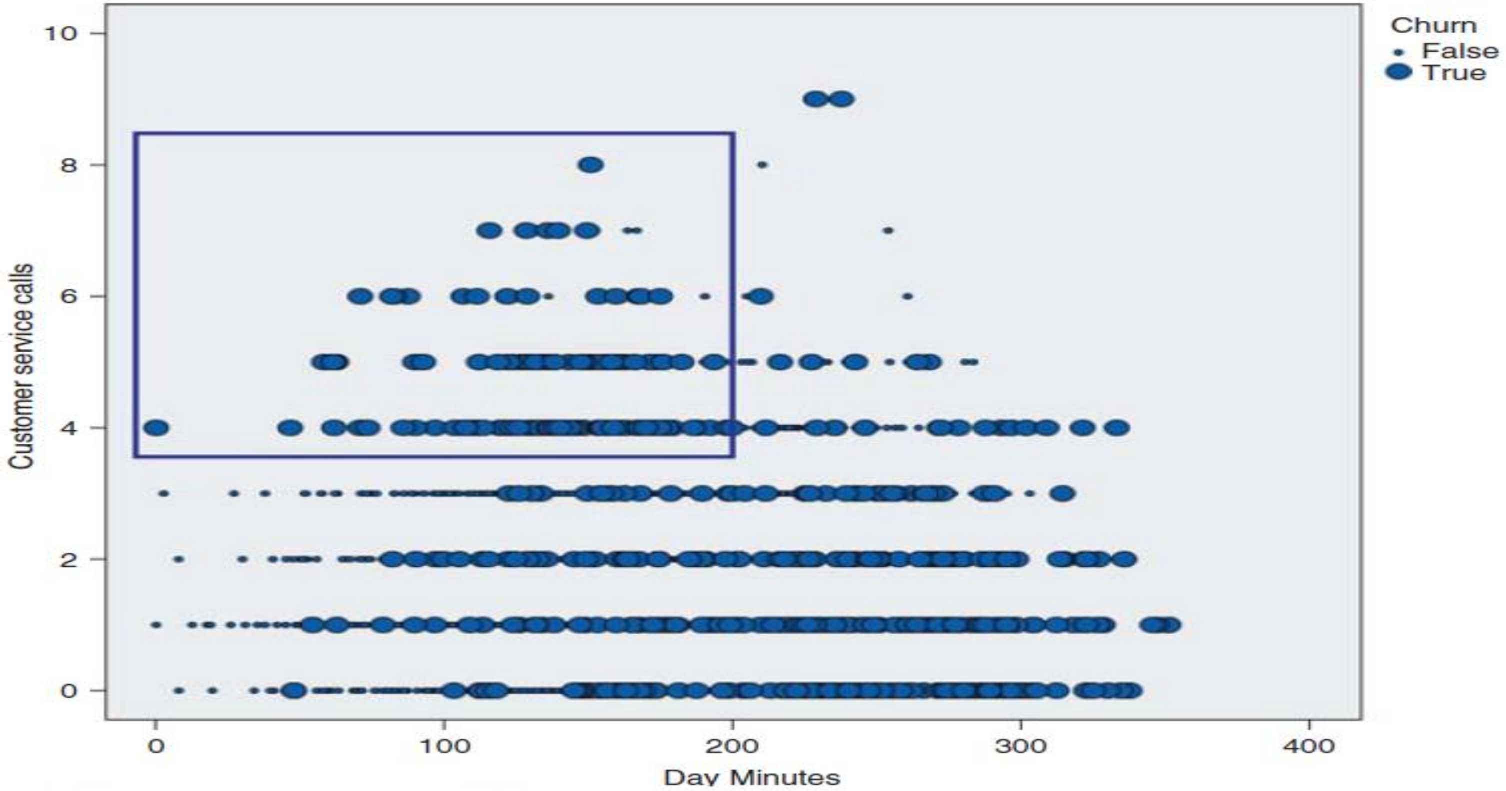
- Scatter plots can be used for examination of the possible multivariate associations



Customers with both high day minutes and high evening minutes are at greater risk of churning.

- **Records above this diagonal line** (customers high day minutes and evening minutes), - higher proportion of churners than records below line.

EXPLORING MULTIVARIATE RELATIONSHIPS



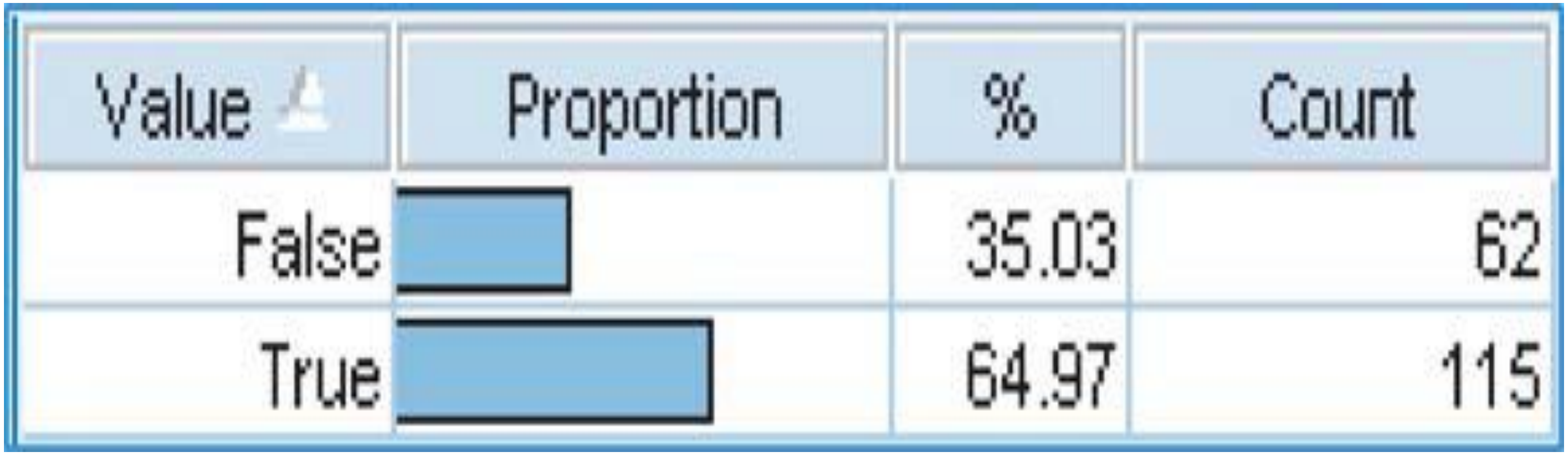
There is an interaction effect between *customer service calls* and *day minutes* with respect to churn.

- Consider the records inside the rectangle partition - indicates a high-churn area
- These records represent combination of a high number of customer service calls and a low number of day minutes used.
- This group of customers could not have been identified with **univariate exploration**

EXPLORING MULTIVARIATE RELATIONSHIPS

- Graphical EDA can uncover subsets of records that call for further investigation
- About 65% (115 of 177) of the selected records are churners

- Those with high customer service calls and low day minutes have a **65%** probability of churning

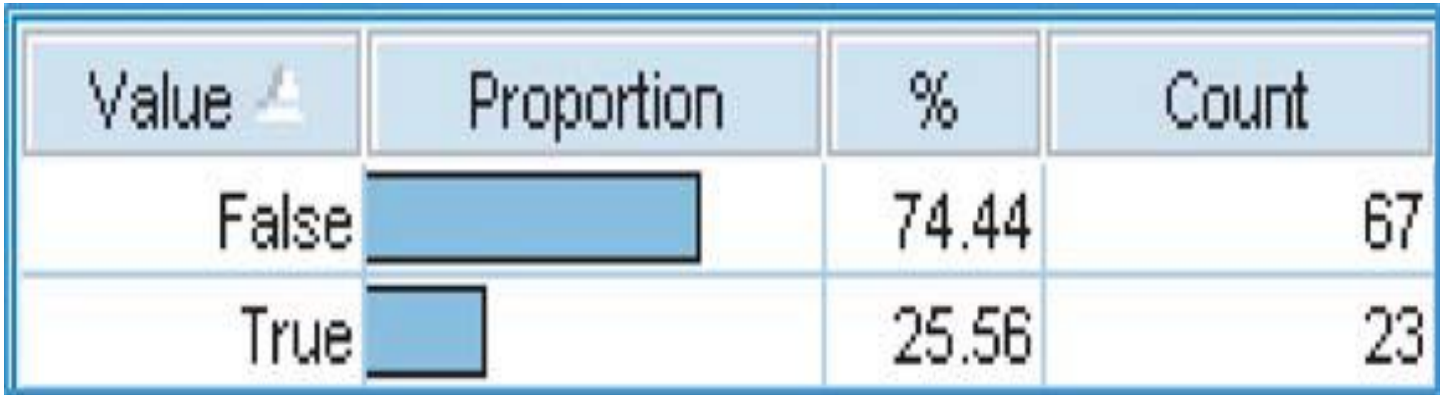


Value	Proportion	%	Count
False	▇▇	35.03	62
True	▇▇▇▇	64.97	115

SELECTING INTERESTING SUBSETS OF THE DATA FOR FURTHER INVESTIGATION

- Compare this to the records with high customer service calls and high day minutes

- About 26% of customers with high customer service calls and high day minutes are churners



Value	Proportion	%	Count
False		74.44	67
True		25.56	23

SELECTING INTERESTING SUBSETS OF THE DATA FOR FURTHER INVESTIGATION

To summarize, the strategy we implemented here is as follows:

1. Generate multivariate graphical EDA, such as scatter plots with a flag overlay.
2. Use these plots to uncover subsets of interesting records.
3. Quantify the differences by analyzing the subsets of records.

SUMMARY OF OUR EDA

- The four *charge* fields are linear functions of the *minute* fields, and should be omitted.
- The *area code* field and/or the *state* field are anomalous, and should be omitted until further clarification is obtained.

Insights with respect to *churn* are as follows:

- Customers with the *International Plan* tend to churn more frequently.
- Customers with the *Voice Mail Plan* tend to churn less frequently.
- Customers with four or more *Customer Service Calls* churn more than four times as often as the other customers.

SUMMARY OF OUR EDA

- Customers with both high DayMinutes and high Evening Minutes tend to churn at a higher rate than the other customers.
- Customers with both high Day Minutes and high Evening Minutes churn at a rate about six times greater than the other customers.
- Customers with low Day Minutes and high Customer Service Calls churn at a higher rate than the other customers.
- Customers with lower numbers of International Calls churn at a higher rate than do customers with more international calls.
- For the remaining predictors, EDA uncovers no obvious association of churn.

Exploratory Data Analysis of Text Data

- ❑Exploratory Data Analysis (EDA) is the process by which the data analyst becomes acquainted with their data to drive intuition and begin to formulate testable hypotheses.
- ❑This process typically makes use of descriptive statistics and visualizations

Exploratory Data Analysis of Text Data

Sentiment Analysis

- ❑**Sentiment analysis** is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social **sentiment** of their brand, product or service while monitoring online conversations.

- ❑Sentiment analysis is the process of determining the writer's attitude or opinion ranging from -1 (negative attitude) to 1 (positive attitude).

Exploratory Data Analysis of Text Data

Exploratory Data Analysis for Text Mining Steps

- ❑ Loading of Data Set
- ❑ Basic Summary and Statistics
- ❑ Sampling and Cleaning of Data
- ❑ Exploratory Data Analysis

Exploratory Data Analysis of Text Data

Basic Summary and Statistics

- ➢ Generate summary of text documents which includes
  - ❑ Line Count
  - ❑ Total Word Count
  - ❑ Max Word Count
  - ❑ Average Word Count

Exploratory Data Analysis of Text Data

Sampling and Cleaning of Data

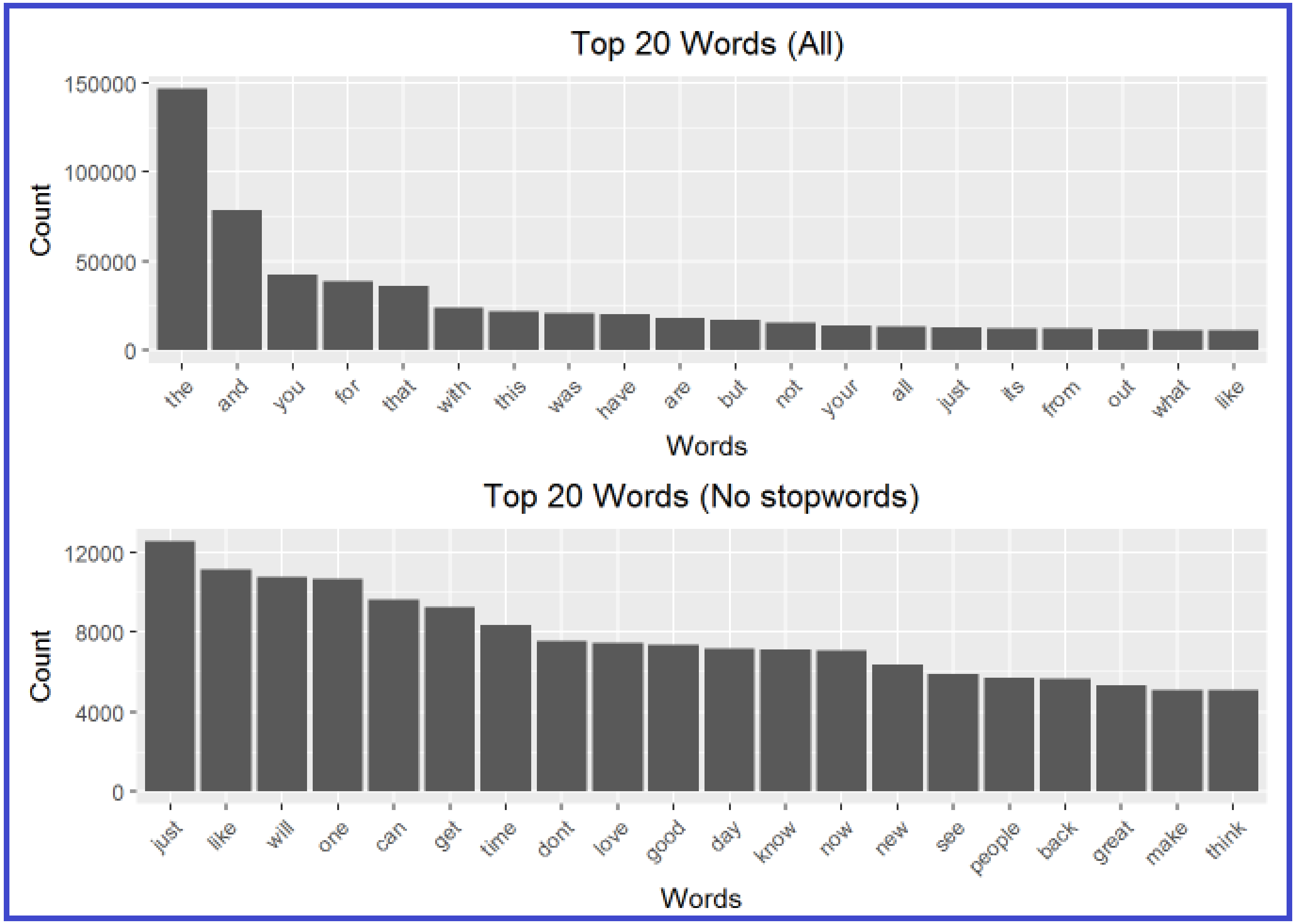
- ❑ Each individual data files is pretty huge.
- ❑ As a start, a sample of each text file will be taken to perform exploratory data analysis.
- ❑ Before that, it is important to clean up the data by removing
  - ➢ punctuations,
  - ➢ numbers,
  - ➢ profanities (offensive or obscene word or phrase)
  - ➢ some unwanted symbols.
- ❑ As white space will be introduced during the process, stripping of white space will be done at the end.

Exploratory Data Analysis of Text Data

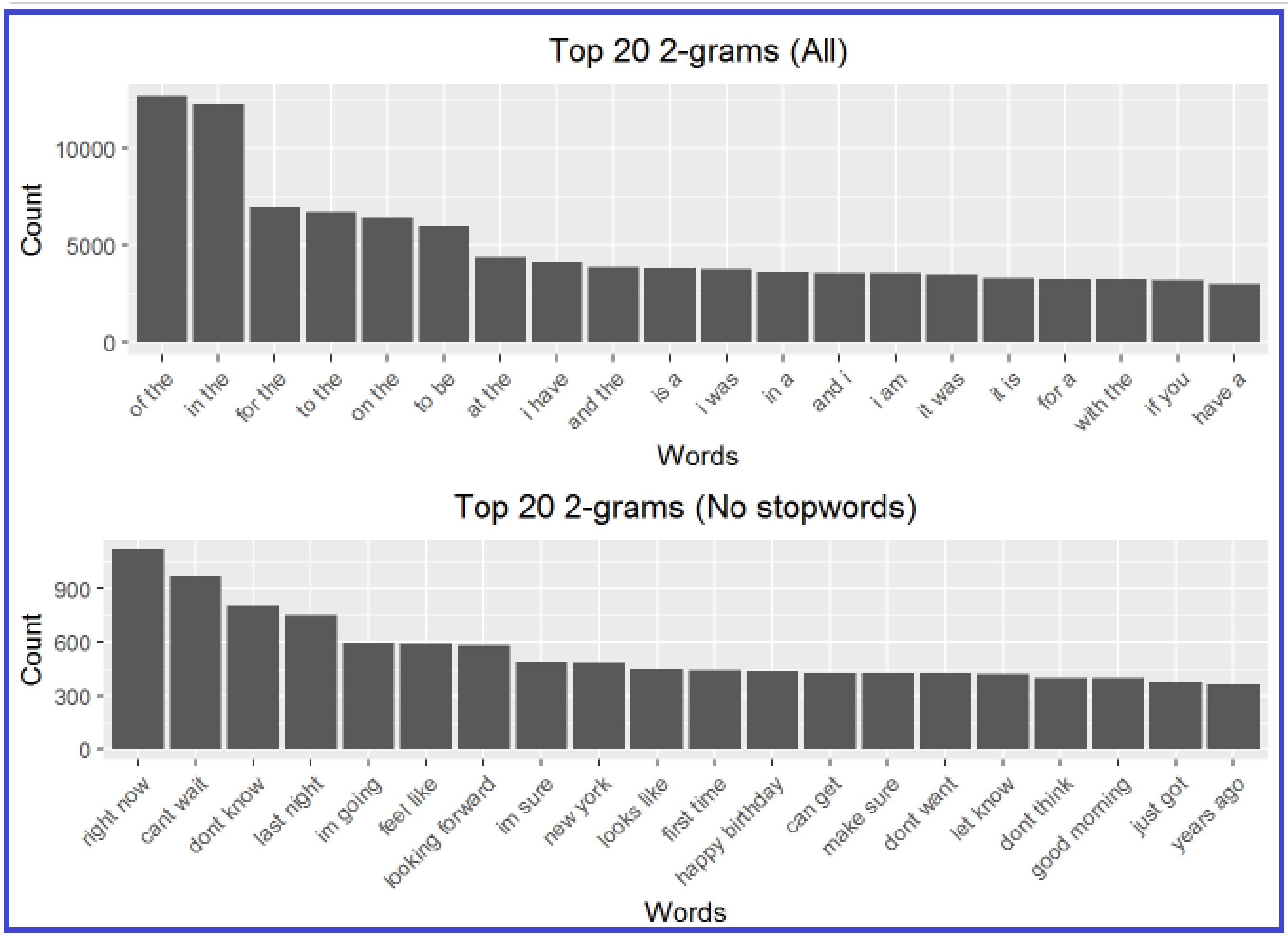
Exploratory Data Analysis

- ❑ Exploratory data analysis will be performed on the cleaned text corpuses.
- ❑ Look at frequencies of words/phrases used in sample data.
- ❑ N-grams are used to describe the number of words used as observation points, e.g., 1-gram means singly-worded, 2-gram means 2-worded phrase, and etc.
- ❑ Term document matrices are used to represent the frequencies of the word/phrase used in each document.
- ❑ The more the words and documents considered, the larger the dimension of the matrix.
- ❑ Since limitation of the computing power, limit the sample size of our data for exploratory data analysis.
- ❑ After generating the term document matrices, to find the most frequent terms, we need to sum the total number of entries in each row.

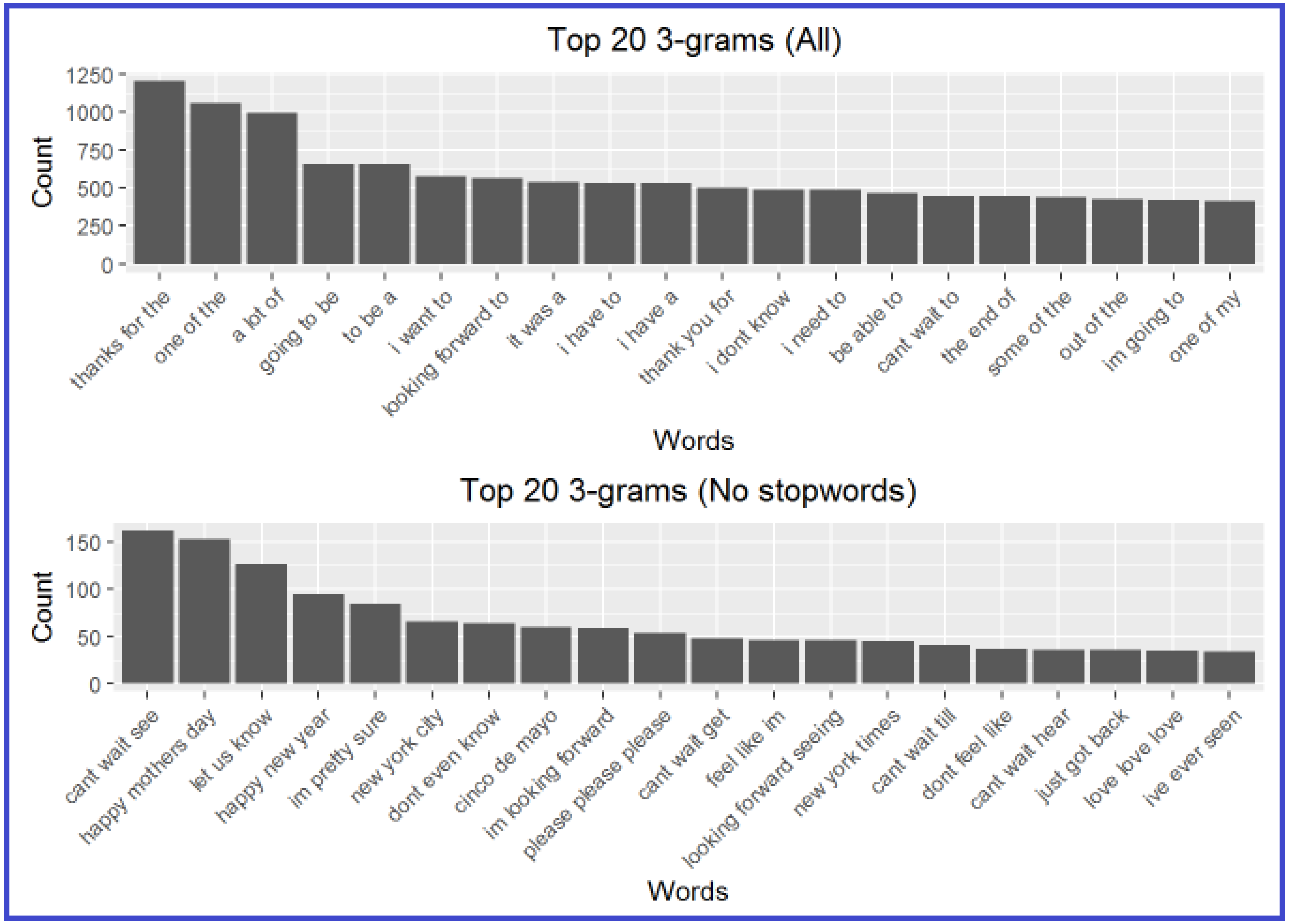
Exploratory Data Analysis of Text Data



Exploratory Data Analysis of Text Data



Exploratory Data Analysis of Text Data



Summary

have considered

- univariate non-graphical,
- multivariate non-graphical,
- univariate graphical, and
- multivariate graphical.
- EDA Example,
- EDA for Text Data

Thank You !!!

Exploratory Data Analysis of Text Data

Exploratory Data Analysis of Text Data

https://rstudio-pubs-static.s3.amazonaws.com/231095_0e6f05290f3b4f82bba74f97edb31744.html