ECE5984 – Applications of Machine Learning
Lecture 5 – More on Data Exploration

Creed Jones, PhD

BRADLEY DEPARTMENT
OF ELECTRICAL & COMPUTER ENGINEERING

VIRGINIA TECH.

Course Updates

- Quiz 2 on February 10
  - Covers lectures 4-7
- At the end of the semester, I will replace your lowest quiz grade with your next lowest grade
- HW1 is posted
  - Due on Feb 8
  - Submit via Canvas

A personal note

- Schedule
  - Past
  - Future

- Team selections will be extended through Saturday, February 5
  - Email me with your team selections!

Question 8 on Quiz 1 was poorly worded

- “True or False: the expected value of a random variable is the value that's most likely to occur; it's the value that we "expect" to get.”
- I meant this to check whether you understood that the expected value of a r.v. is a value that won’t necessarily occur, but rather a probability-weighted centroid of the result space that is used for decision making.
- But, one student pointed out that it was confusing, and upon re-reading it, I agree.
- I’ve regraded this question to allow either answer.

Today's Objectives

Chapter 3 – Data Exploration

- 3.1 The Data Quality Report
- 3.2 Getting to Know the Data
  - Tableau exploration -
- 3.3 Identifying Data Quality Issues
- 3.4 Handling Data Quality Issues
- 3.5 Advanced Data Exploration

CHAPTER 3 – DATA EXPLORATION

The *Data Quality Report* is a standard report generated from the analytic data set, to inform us on the use of the data for model development

- Consider a data set for the insurance claims example used in the book
- “ID” is the ID field (obviously) and “Fraud Flag” is the target variable
- The rest are descriptive features – but are they suited for modeling?

ID	TYPE	INC.	MARITAL STATUS	NUM. CLMNTS.	INJURY TYPE	HOSPITAL STAY	CLAIM AMT.	TOTAL CLAIMED	NUM CLAIMS	NUM. SOFT TISS.	% SOFT TISS.	CLAIM AMT. RCVD.	FRAUD FLAG
1	ci	0		2	soft tissue	no	1,625	3,250	2	2	1.0	0	1
2	ci	0		2	back	yes	15,028	60,112	1		0	15,028	0
3	ci	54,613	married	1	broken limb	no	-99,999	0	0	0	0	572	0
4	ci	0		4	broken limb	yes	5,097	11,661	1	1	1.0	7,864	0
5	ci	0		4	soft tissue	no	8,869	0	0	0	0	0	1
		⋮				⋮					⋮		
300	ci	0		2	broken limb	no	2,244	0	0	0	0	2,244	0
301	ci	0		1	broken limb	no	1,627	92,283	3	0	0	1,627	0
302	ci	0		3	serious	yes	270,200	0	0	0	0	270,200	0
303	ci	0		1	soft tissue	no	7,668	92,806	3	0	0	7,668	0
304	ci	46,365	married	1	back	no	3,217	0	0		0	1,653	0

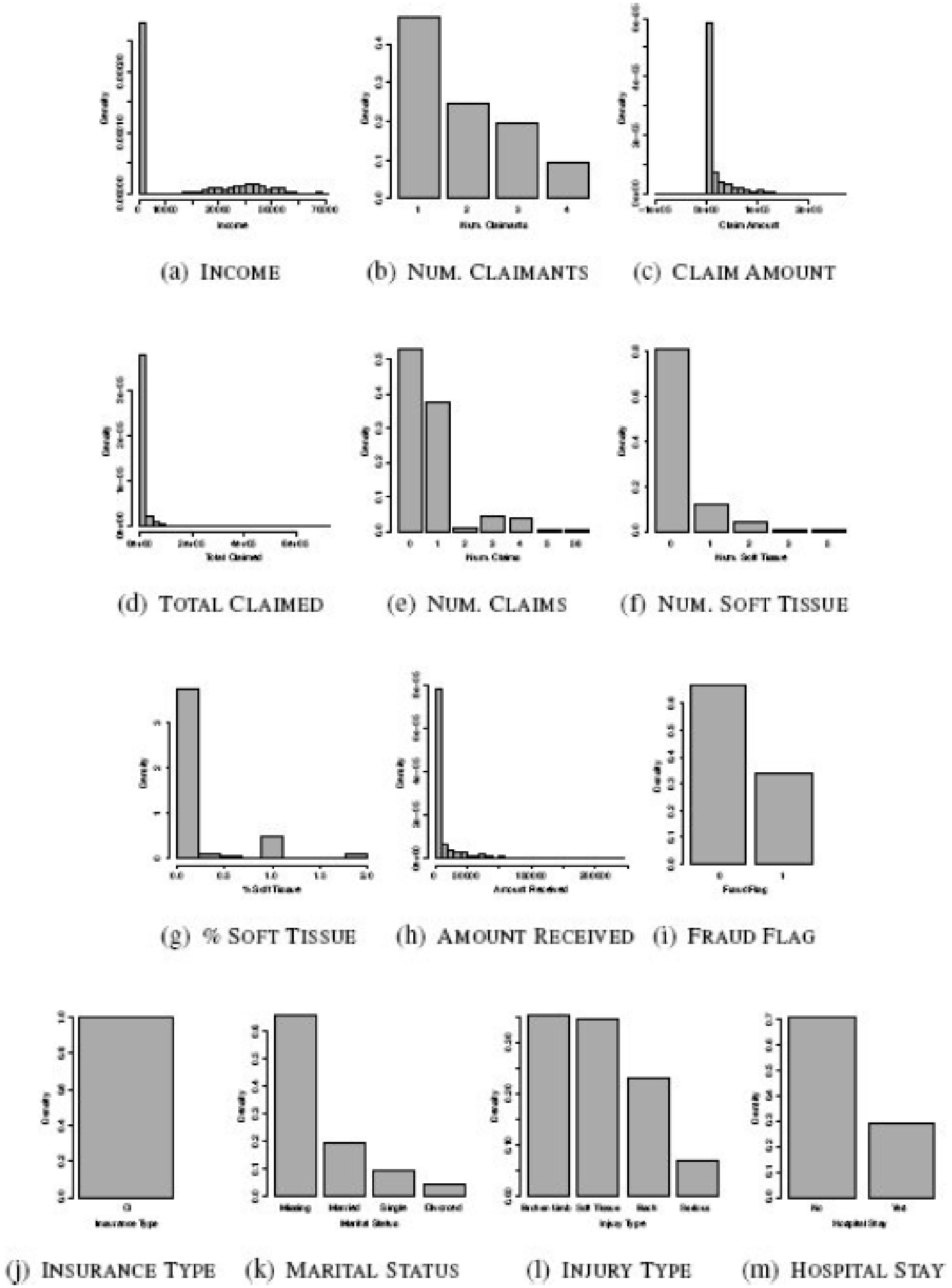
(a) Continuous Features

Feature	Count	% Miss.	Card.	Min	1^{st} Qrt.	Mean	Median	3^{rd} Qrt.	Max	Std. Dev.
INCOME	500	0.0	171	0.0	0.0	13,740.0	0.0	33,918.5	71,284.0	20,081.5
NUM. CLAIMANTS	500	0.0	4	1.0	1.0	1.9	2	3.0	4.0	1.0
CLAIM AMOUNT	500	0.0	493	-99,999	3,322.3	16,373.2	5,663.0	12,245.5	270,200.0	29,426.3
TOTAL CLAIMED	500	0.0	235	0.0	0.0	9,597.2	0.0	11,282.8	729,792.0	35,655.7
NUM. CLAIMS	500	0.0	7	0.0	0.0	0.8	0.0	1.0	56.0	2.7
NUM. SOFT TISSUE	500	2.0	6	0.0	0.0	0.2	0.0	0.0	5.0	0.6
% SOFT TISSUE	500	0.0	9	0.0	0.0	0.2	0.0	0.0	2.0	0.4
AMOUNT RECEIVED	500	0.0	329	0.0	0.0	13,051.9	3,253.5	8,191.8	295,303.0	30,547.2
FRAUD FLAG	500	0.0	2	0.0	0.0	0.3	0.0	1.0	1.0	0.5

(b) Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2^{nd} Mode	2^{nd} Mode Freq.	2^{nd} Mode %
INSURANCE TYPE	500	0.0	1	ci	500	1.0	–	–	–
MARITAL STATUS	500	61.2	4	married	99	51.0	single	48	24.7
INJURY TYPE	500	0.0	4	broken limb	177	35.4	soft tissue	172	34.4
HOSPITAL STAY	500	0.0	2	no	354	70.8	yes	146	29.2

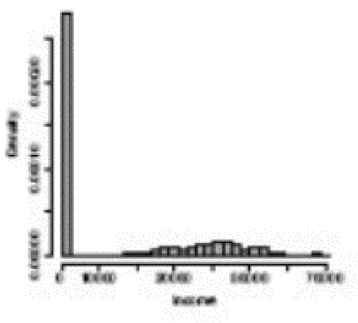
We also generate histograms of the distributions of continuous and categorical features – possibly interval features as well



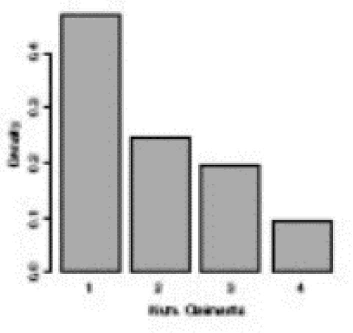
Very important – note that these are all univariate descriptions of the features – these say nothing about the joint distributions!

- The data quality report tells us how each individual feature is distributed and how "well-behaved" it is
- Used for feature selection and for defining preprocessing
  - Missing values, for example

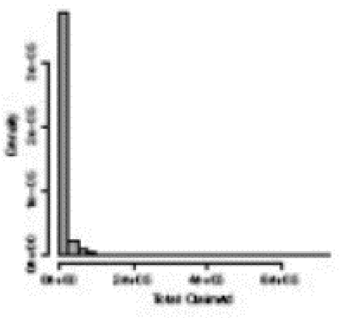
- To understand joint distributions (how one variable's distribution depends on the value of other variables), we must do correlation analysis



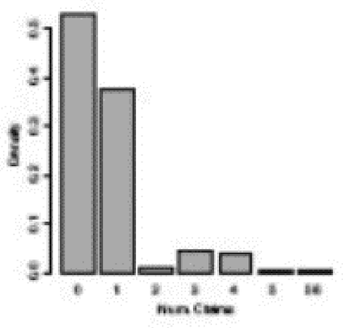
(a) INCOME



(b) NUM. CLAIMANTS



(d) TOTAL CLAIMED



(e) NUM. CLAIMS

The success or failure of any predictive modeling project is far more dependent on the data than on the modeling

- We need to understand the data
  - Its distributions (recall the data quality report)
  - Its time behavior
  - Its quality (missing values)
  - Its reliability (noise)

- Most predictive modelers are good at EDA – exploratory data analysis
  - It’s common to have a set of programs or scripts that assess distributions, interdependencies and trends over time
  - I like to use interactive tools for working with the data
  - Tableau is my favorite

There are many problems that can occur in data sets, that we can observe in the data quality report or the data itself

- First, determine whether the errors are due to *data invalidity*
  - Data entry problems
  - File corruption
  - Format problems

ID	OCCUPATION	AGE	STATE	LOAN-SALARY RATIO	OUTCOME
1	industrial	34	GA	2.96	repay
2	professional	41	VA	*NaN*	default
3	professional	36	VIRGINIA	3.22	?????
4	professional	411	IR	3.11	default
5	industrial	48	******	3.8	default
6	self	employed	55	TX	2.45
7	other	61		2.52	repay
8	professional	37	T	1.5	repay

Even in fully valid data, we can have data problems to address; the big three are – missing values, outliers and irregular cardinality

- Missing values may represent unreported or irrelevant fields
  - A missing value in the income field cannot be assumed to mean INCOME=0

Consider the possible results when we inquire about a person’s income:

- A given dollar amount
- Zero (the person truly has no income)
- Irrelevant for this example (the person is an infant, perhaps)
- Unknown, didn’t ask (will show up as missing)
- Unknown, asked but they didn’t answer (will show up as missing)
- Impossible values: $42 trillion or -$30K, for example (sign of an error)

What do we do?

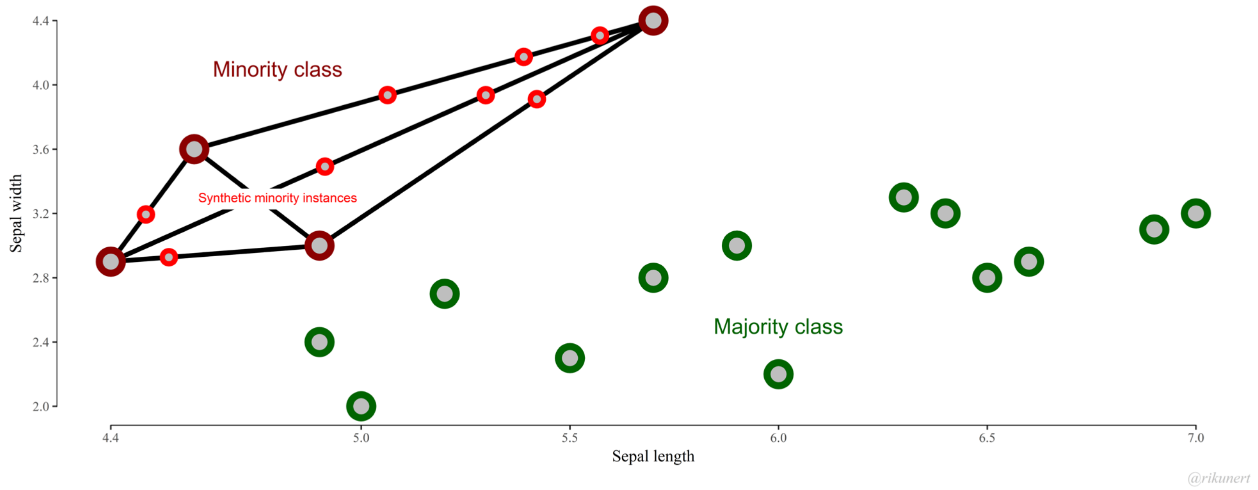
Possible ways to deal with missing values

1. Remove that feature from the dataset
   - Only if well over half of the values are missing
2. *Impute* the mean value of the data present
   - OK, but overly simplistic
3. Impute the mean value for an appropriate subset, using a categorical feature
   - Imagine a database of employees; if a salary is missing, replace it with the mean for other employees with the same job title
   - This is called stratified imputation
4. Replace it with zero
   - Sometimes, this is the right answer
5. More sophisticated techniques
   - SMOTE

SMOTE: Synthetic Minority Over-sampling Technique replaces missing values with a value randomly distributed between the values of like samples

- SMOTE determines the nearest neighbors in modeling space to an instance with a missing value (in the non-missing dimensions)
- It then replaces the missing value with a randomly weighted sum of the existing values of the neighbors for that attribute
- SMOTE is also used to create synthetic samples to correct class imbalance
- https://www.jair.org/index.php/jair/article/download/10302/24590

Addressing class imbalance problems of ML via SMOTE: synthesising new dots between existing dots



Journal of Artificial Intelligence Research 16 (2002) 321–357 Submitted 09/01; published 06/02

SMOTE: Synthetic Minority Over-sampling Technique

Nitesh V. Chawla CHAWLA@CSEE.USF.EDU
Department of Computer Science and Engineering, ENB 118
University of South Florida
4202 E. Fowler Ave.
Tampa, FL 33620-5399, USA

Kevin W. Bowyer KWB@CSE.ND.EDU
Department of Computer Science and Engineering
384 Fitzpatrick Hall
University of Notre Dame
Notre Dame, IN 46556, USA

Lawrence O. Hall HALL@CSEE.USF.EDU
Department of Computer Science and Engineering, ENB 118
University of South Florida
4202 E. Fowler Ave.
Tampa, FL 33620-5399, USA

W. Philip Kegelmeyer WPK@CALIFORNIA.SANDIA.GOV
Sandia National Laboratories
Biosystems Research Department, P.O. Box 969, MS 9951
Livermore, CA, 94551-0969, USA

Abstract

An approach to the construction of classifiers from imbalanced datasets is described. A dataset is imbalanced if the classification categories are not approximately equally represented. Often real-world data sets are predominately composed of "normal" examples with only a small percentage of "abnormal" or "interesting" examples. It is also the case that the cost of misclassifying an abnormal (interesting) example as a normal example is often much higher than the cost of the reverse error. Under-sampling of the majority (normal) class has been proposed as a good means of increasing the sensitivity of a classifier to the minority class. This paper shows that a combination of our method of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance (in ROC space) than only under-sampling the majority class.

This paper is on Canvas

Take a look at it

Possible ways to deal with outliers

1. Invalid outliers are generally data errors (negative age, for example)
    1. If we can't fix the error, delete the value and consider as missing

2. Valid outliers are real data
    1. If I analyze the income of the class roster, and one of you is secretly a billionaire, then averages and other stats will be skewed
    2. Perhaps I leave it and use it as is?
    3. Perhaps cap certain values (all incomes over $1M become $1M)?
        1. The book calls this a *clamp transformation*
    4. Perhaps change to a categorical (low, medium, high and very high incomes)?

The proper approach depends on the modeling technique used
Keep in mind that predictions based on rare outlier values will not generalize well

Possible ways to deal with irregular cardinality

- *Cardinality* refers to the number of distinct values present for a feature
- If a feature has cardinality of 1, then all values are the same and the feature offers no real value
  - May indicate a data error
- If the cardinality of a categorical value is large, then we may actually have a continuous numeric feature
  - Or, we need to do some grouping of the feature
- If the cardinality does not match the meaning of the feature, then check for data errors
  - A common one is a US_STATE field with cardinality well above 50
  - Often we need to group "VA", "Va", "Virginia", etc…

Missing data
Outliers
Cardinality issues

(a) Continuous Features

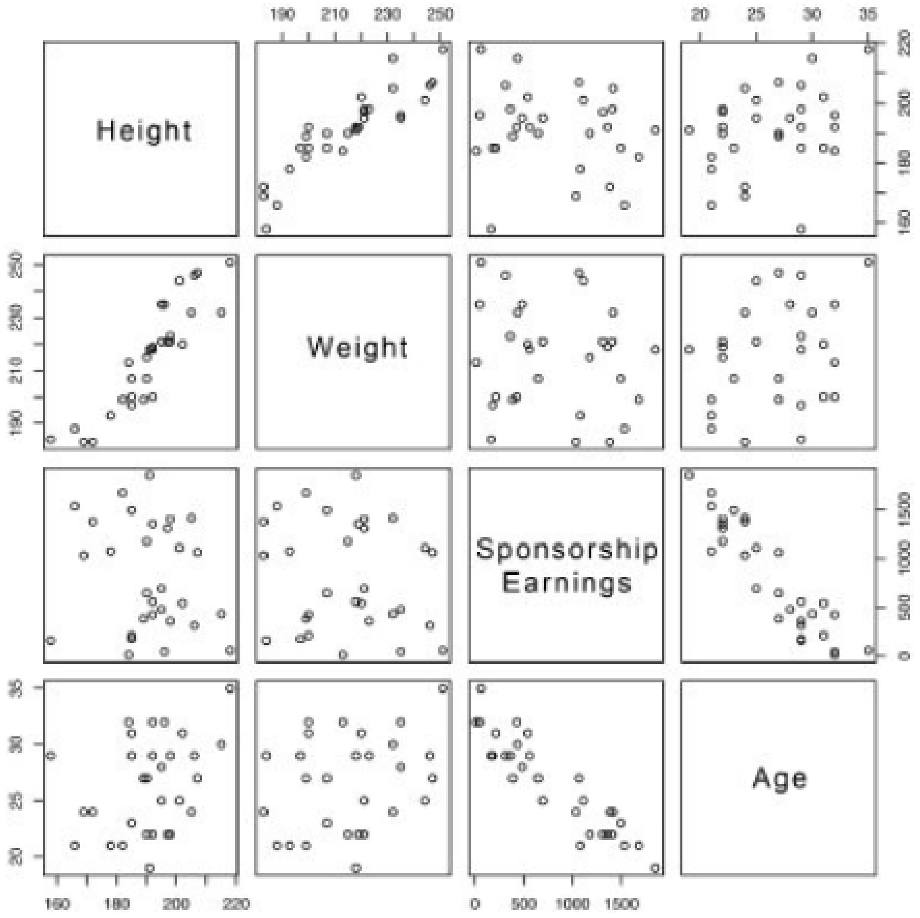
Feature	Count	% Miss.	Card.	Min	1st Qrt.	Mean	Median	3rd Qrt.	Max	Std. Dev.
INCOME	500	0.0	171	0.0	0.0	13,740.0	0.0	33,918.5	71,284.0	20,081.5
NUM. CLAIMANTS	500	0.0	4	1.0	1.0	1.9	2	3.0	4.0	1.0
CLAIM AMOUNT	500	0.0	493	-99,999	3,322.3	16,373.2	5,663.0	12,245.5	270,200.0	29,426.3
TOTAL CLAIMED	500	0.0	235	0.0	0.0	9,597.2	0.0	11,282.8	729,792.0	35,655.7
NUM. CLAIMS	500	0.0	7	0.0	0.0	0.8	0.0	1.0	56.0	2.7
NUM. SOFT TISSUE	500	2.0	6	0.0	0.0	0.2	0.0	0.0	5.0	0.6
% SOFT TISSUE	500	0.0	9	0.0	0.0	0.2	0.0	0.0	2.0	0.4
AMOUNT RECEIVED	500	0.0	329	0.0	0.0	13,051.9	3,253.5	8,191.8	295,303.0	30,547.2
FRAUD FLAG	500	0.0	2	0.0	0.0	0.3	0.0	1.0	1.0	0.5

(b) Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
INSURANCE TYPE	500	0.0	1	ci	500	1.0	–	–	–
MARITAL STATUS	500	61.2	4	married	99	51.0	single	48	24.7
INJURY TYPE	500	0.0	4	broken limb	177	35.4	soft tissue	172	34.4
HOSPITAL STAY	500	0.0	2	no	354	70.8	yes	146	29.2

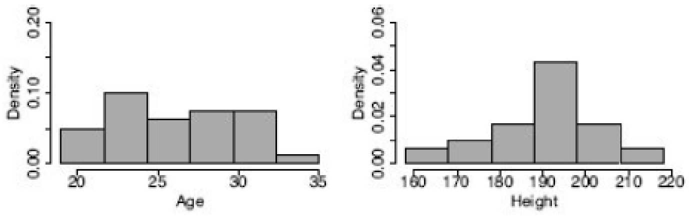
The data exploration phase is important, involves a little bit of art, and often moves imperceptibly into the modeling phase

- Often we explore data by understanding relationships between pairs of variables
- A set of scatter plots is a useful way of looking for variables that are related
- If the plot is a "cloud of points", then there's no simple relationship
- Linear, power or other relationships will be evident
- Often we need to stratify to see patterns (only scatter plot examples from Texas, for example)

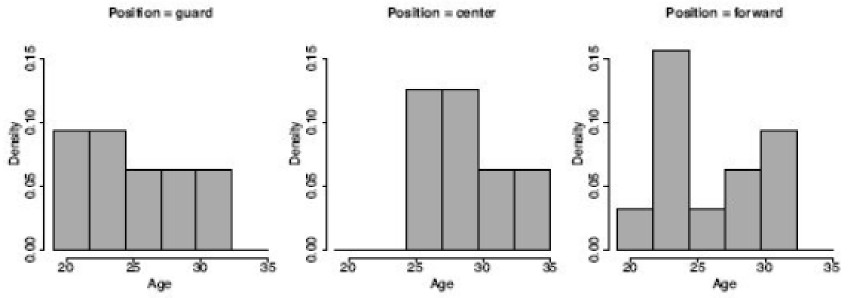


Many people like to use stacked bar plots to see relationships – do variables behave differently for certain values of other variables?

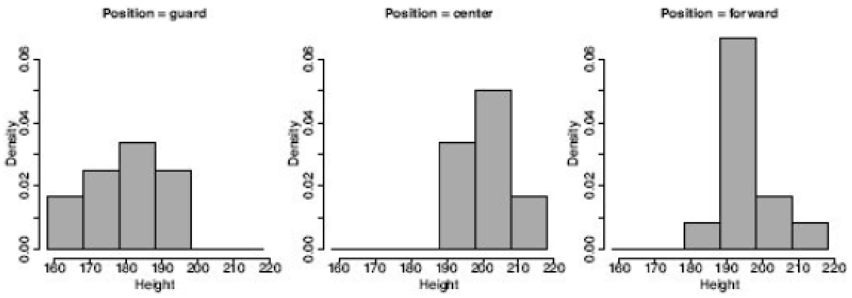
VIRGINIA TECH
ece
BRADLEY DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING



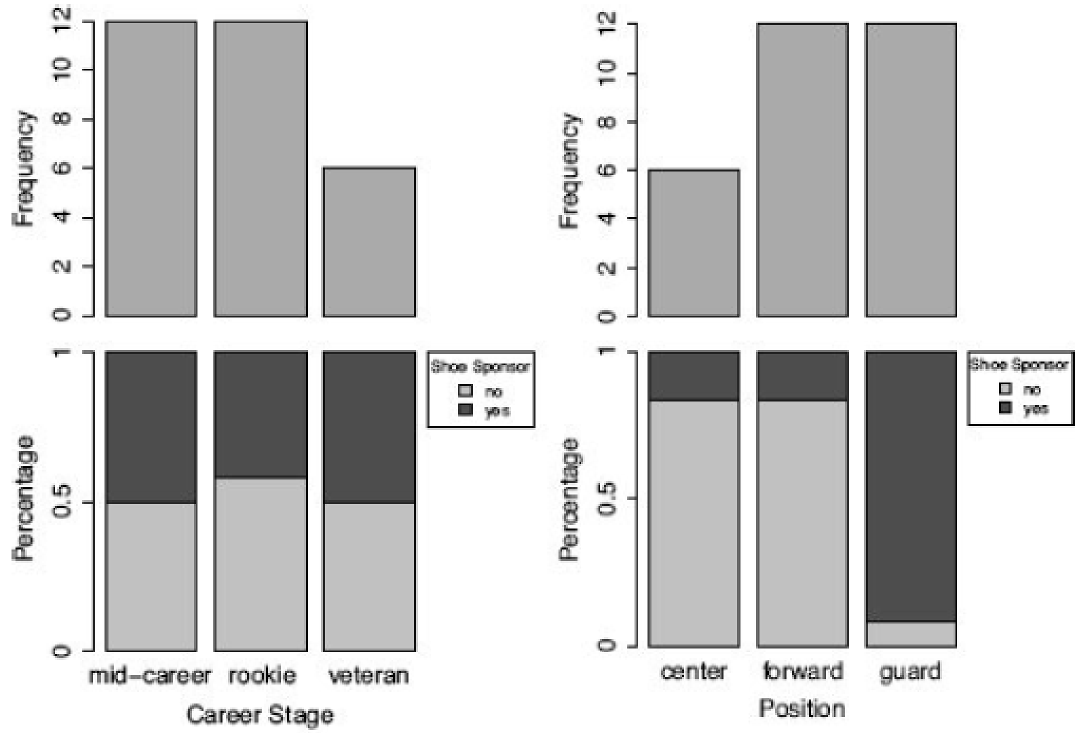
(a) Age (b) Height



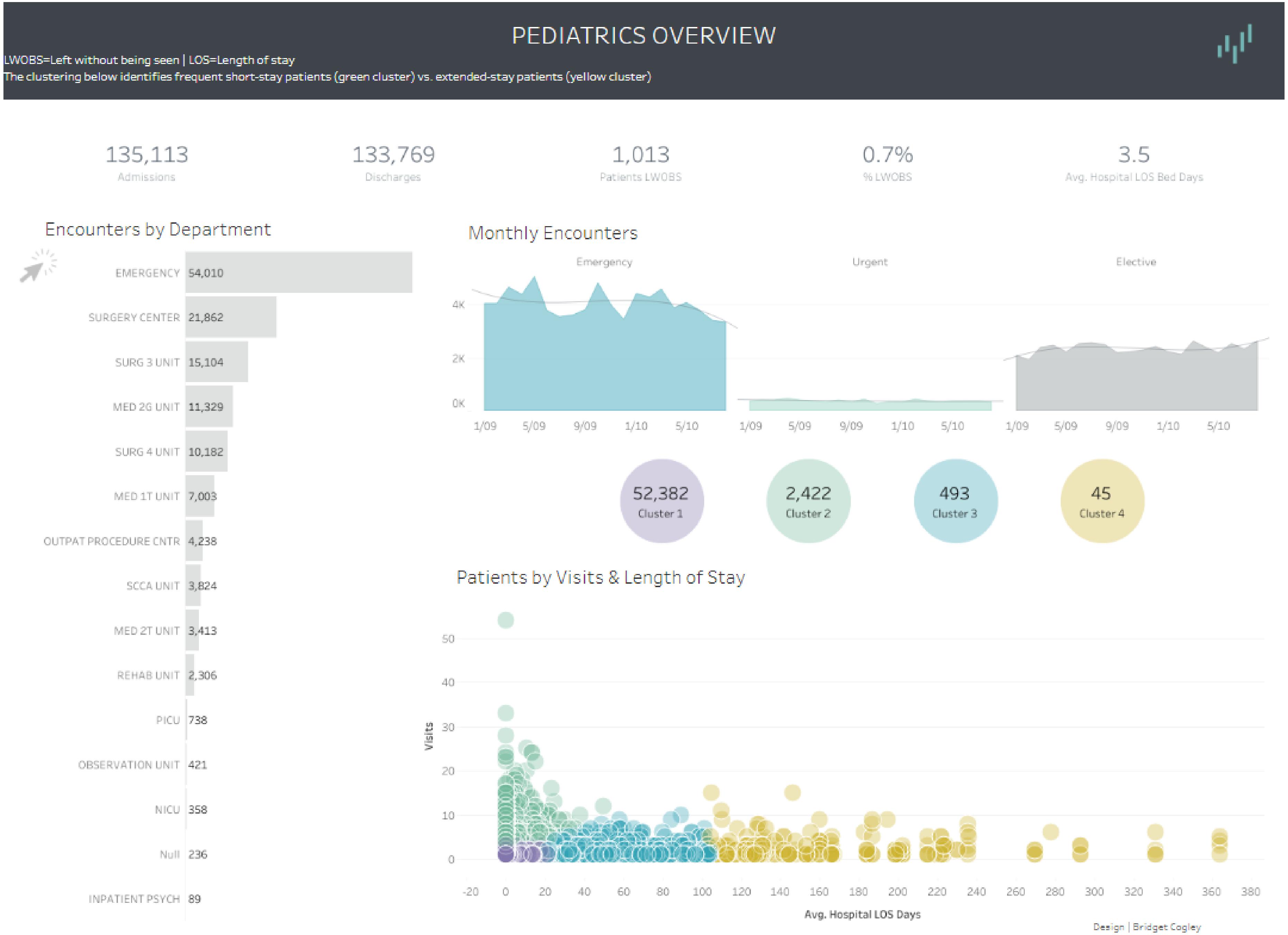
(c) Age and Position



(d) Height and Position



(a) Career Stage and Shoe Sponsor (b) Position and Shoe Sponsor


PEDIATRICS OVERVIEW
LWOBS=Left without being seen | LOS=Length of stay
The clustering below identifies frequent short-stay patients (green cluster) vs. extended-stay patients (yellow cluster)
135,113
Admissions
133,769
Discharges
1,013
Patients LWOBS
0.7%
% LWOBS
3.5
Avg. Hospital LOS Bed Days
Encounters by Department
EMERGENCY 54,010
SURGERY CENTER 21,862
SURG 3 UNIT 15,104
MED 2G UNIT 11,329
SURG 4 UNIT 10,182
MED 1T UNIT 7,003
OUTPAT PROCEDURE CNTR 4,238
SCCA UNIT 3,824
MED 2T UNIT 3,413
REHAB UNIT 2,306
PICU 738
OBSERVATION UNIT 421
NICU 358
Null 236
INPATIENT PSYCH 89
Monthly Encounters
Emergency
Urgent
Elective
4K
2K
0K
1/09 5/09 9/09 1/10 5/10
52,382
Cluster 1
2,422
Cluster 2
493
Cluster 3
45
Cluster 4
Patients by Visits & Length of Stay
Visits
50 40 30 20 10 0
-20 0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380
Avg. Hospital LOS Days
Design | Bridget Cogley

Proper normalization of the data usually leads to better model training and performance

- In real-world data sources, continuous features often have very different numeric ranges
  - A feature representing customer ages might cover the range [16, 96], whereas a feature representing customer salaries might cover the range [10,000, 100,000].
- Range normalization (or min-max normalization) equalizes the range of all variables
- $a'_i = \frac{a_i - min(a)}{max(a) - min(a)}(high - low) + low$
  - a is the old variable
  - $high$ and low are the new extrema

	Height			Sponsorship Earnings		
	Values	Range	Standard	Values	Range	Standard
	192	0.500	-0.073	561	0.315	-0.649
	197	0.679	0.533	1,312	0.776	0.762
	192	0.500	-0.073	1,359	0.804	0.850
	182	0.143	-1.283	1,678	1.000	1.449
	206	1.000	1.622	314	0.164	-1.114
	192	0.500	-0.073	427	0.233	-0.901
	190	0.429	-0.315	1,179	0.694	0.512
	178	0.000	-1.767	1,078	0.632	0.322
	196	0.643	0.412	47	0.000	-1.615
	201	0.821	1.017	1111	0.652	0.384
Max	206			1,678		
Min	178			47		
Mean	193			907		
Std. Dev.	8.26			532.18		

$$a'_i = \frac{a_i - min(a)}{max(a) - min(a)} \times (high - low) + low$$

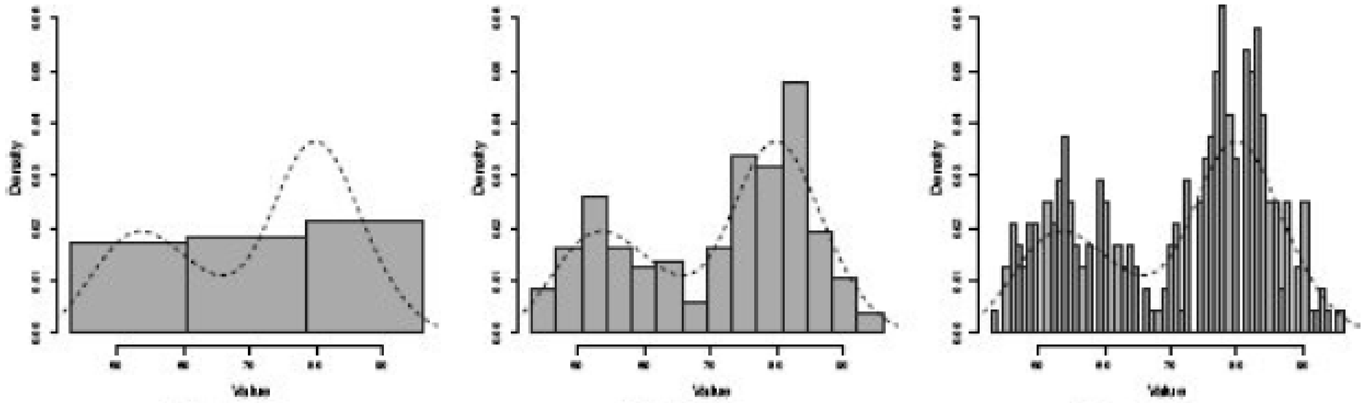
Standard score normalization (or mean-sigma normalization) transforms features to the same mean and standard deviation (often 0 and 1)

- Based on the presumption that the data is normally distributed, or close anyway

	HEIGHT			SPONSORSHIP EARNINGS		
	Values	Range	Standard	Values	Range	Standard
	192	0.500	-0.073	561	0.315	-0.649
	197	0.679	0.533	1,312	0.776	0.762
	192	0.500	-0.073	1,359	0.804	0.850
	182	0.143	-1.283	1,678	1.000	1.449
	206	1.000	1.622	314	0.164	-1.114
	192	0.500	-0.073	427	0.233	-0.901
	190	0.429	-0.315	1,179	0.694	0.512
	178	0.000	-1.767	1,078	0.632	0.322
	196	0.643	0.412	47	0.000	-1.615
	201	0.821	1.017	1111	0.652	0.384
Max	206			1,678		
Min	178			47		
Mean	193			907		
Std Dev	8.26			532.18		

Binning is the process of assigning a continuous variable to a categorical value – to mitigate noise and to allow use in stratifying

- Equal-width binning (0-10, 11-20, 21-30, etc.)
- Equal-frequency binning (lowest 10%, next 10%, etc)
- Often we will keep both the original continuous variable *and* the binned result as possible modeling features
- Need to determine the proper number of bins



(a) 3 bins (b) 14 bins (c) 60 bins

Sometimes the dataset we have is so large that we do not use all the data available to us in an ABT and instead *sample* a smaller percentage from the larger dataset

- We need to be careful when sampling, however, to ensure that the resulting datasets are still representative of the original data and that no unintended bias is introduced during this process.

- Common forms of sampling include:
  - top sampling
  - random sampling
  - stratified sampling
  - under-sampling
  - over-sampling

When we only deal with part of the dataset, think about how to choose the instances

- **Top sampling** simply selects the top s% of instances from a dataset to create a sample
  - It can introduce bias dependent on the order of the data – don't do it
- **Random sampling** randomly selects a proportion of s% of the instances from a large dataset to create a smaller set.
  - The most common practice
- **Stratified sampling** ensures that the relative frequencies of the levels of a specific stratification feature are maintained in the sampled dataset.
  - The instances in a dataset are divided into groups containing only instances that have a particular level for the stratification feature
  - s% of the instances in each stratum are randomly selected
  - these selections are combined to give an overall sample of s% of the original dataset.

Sometimes we want to modify the proportion of the data set having a particular value or values; this calls for under-sampling or over-sampling

Under-sampling begins by dividing a dataset into groups, containing only instances that have a particular level for the feature to be under-sampled.

- The number of instances in the smallest group is the under-sampling target size.
- Each group containing more instances than the smallest one is then randomly sampled by the appropriate percentage to create a subset that is the under-sampling target size.
- These under-sampled groups are then combined to create the overall under-sampled dataset.

Over-sampling addresses the same issue as under-sampling but in the opposite way.

- After dividing the dataset into groups, the number of instances in the largest group becomes the over-sampling target size.
- From each smaller group, we then create a sample containing that number of instances using *random sampling with replacement (or SMOTE).*
- These larger samples are combined to form the overall over-sampled dataset.

Today’s Objectives

Chapter 3 – Data Exploration

- 3.1 The Data Quality Report
- 3.2 Getting to Know the Data
  - Tableau exploration -
- 3.3 Identifying Data Quality Issues
- 3.4 Handling Data Quality Issues
- 3.5 Advanced Data Exploration