"Describe and visualize: enhancing your research using data preprocessing."

Dr. Regier for WVCTSI research bootcamp

10/23/2015

Before you analyze

Take a step back

With the proliferation of data analysis software, especially those with an excellent user interface, is putting the power of data analysis in the hands of the researcher, but,

- Resist the urge to jump to the analysis stage
  - Interim analysis has hidden costs
  - Overly zealous data analysis has hidden costs
    - Incorrect actual alpha level
    - Biased data analysis
    - Increase in the likelihood of reporting spurious results
    - Erosion of reproducibility

Useful approximate solutions

Good models rest on a good understanding of the data

- "The formulation of the problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill." - Albert Einstein
- There is a saying that all models are wrong, but some are more useful.
  - Statisticians (applied) are in the business of providing the most useful approximate solution to a problem.
  - Data description and visualization provides a foundation of understanding to build approximate solutions.

Components of statistical analysis

To formulate the problem and solution, we must:

1. Understand the physical, biological, or systems background
2. Understand the objective of the study.
   - Credible research has a clearly defined objective and is not a fishing expedition for any relationship
   - A relationship can be found in almost every dataset.
3. Understand the client's objectives.
   - Perform the analysis that is the best approximate solution for the question asked - simplicity is better than complexity.
4. Put the problem into statistical terms.
   - Translating the problem into statistical language will identify the appropriate set of approaches (e.g. methods, models) to consider.

Further things to consider

1. Understand how the data were collected.
2. Is the data observational or experimental?
3. Is there non-response, missing values or measurement error?
4. How was the data coded? Do I have a data dictionary or code book for the project?
5. What are the units of measurement?
6. Check for data entry errors. How was the data entered and what are the data sources.

Initial data analysis

Once the data is loaded

After loading the data in your chosen program (e.g. R, Stata), take a look at the data.

- This is a critical step and involves univariate and bivariate descriptions as well as some basic plots.
- From the univariate descriptions and graphs,
  - Look for unusually large or small summary measures (e.g. min, max),
  - Large magnitude of skew (e.g. skew > 0.4),
  - Mean, median, percentiles, standard deviation, and possibly kurtosis,
  - Use histograms (e.g. modality, outliers, skew), density plots, and boxplots.

Once the data is loaded

- From bivariate descriptions, look at
  - Correlation structure (e.g. correlation matrix, correlograms),
  - Contingency structures (e.g. tables for categorical variables, mosaic plots),
  - Scatterplots

A sad reality

Getting familiar with a dataset is the most time consuming part of a project.

- Expect to spend about 80% of your time understanding your data (e.g. preprocessing, preliminary analyses)
- Expect to spend about 20% of your time on the deliverable (final) analyses.

The time spent understanding the data and the research problem will prevent garbage in, garbage out research.

Example: Pima dataset

Pima dataset

Description:

A study on 768 adult female Pima Indians living near Phoenix.

Variables

- pregnant: Number of times pregnant
- glucose: Plasma glucose concentration at 2 hours in an oral glucose tolerance test
- diastolic: Diastolic blood pressure (mm Hg)
- triceps: Triceps skin fold thickness (mm)
- insulin: 2-Hour serum insulin (mu U/ml)
- bmi: Body mass index (weight in kg/(height in meters squared))
- diabetes: Diabetes pedigree function
- age: Age (years)
- test: test whether the patient shows signs of diabetes (coded 0 if negative, 1 if positive)

Numeric summaries: univariate

First look at the data

We begin by looking at the first few observations

pregnant	glucose	diastolic	triceps	insulin	bmi	diabetes	age
6	148	72	35	0	33.6	0.627	50
1	85	66	29	0	26.6	0.351	31
8	183	64	0	0	23.3	0.672	32
1	89	66	23	94	28.1	0.167	21
0	137	40	35	168	43.1	2.288	33

First look at the data

Then look at the last few observations

	pregnant	glucose	diastolic	triceps	insulin	bmi	diabetes
764	10	101	76	48	180	32.9	0.171
765	2	122	70	27	0	36.8	0.340
766	5	121	72	23	112	26.2	0.245
767	1	126	60	0	0	30.1	0.349
768	1	93	70	31	0	30.4	0.315

Numerical summaries

	mean	sd	median	min	max	skew
pregnant	3.85	3.37	3.00	0.00	17.00	0.90
glucose	120.89	31.97	117.00	0.00	199.00	0.17
diastolic	69.11	19.36	72.00	0.00	122.00	-1.84
triceps	20.54	15.95	23.00	0.00	99.00	0.11
insulin	79.80	115.24	30.50	0.00	846.00	2.26
bmi	31.99	7.88	32.00	0.00	67.10	-0.43
diabetes	0.47	0.33	0.37	0.08	2.42	1.91
age	33.24	11.76	29.00	21.00	81.00	1.13
test	0.35	0.48	0.00	0.00	1.00	0.63

Observation 1: number of pregnancies

We observe some peculiarities from the summaries.

1. The max number of pregnancies is 17. This is large, but may be accurate.
    - check out the percentiles,

```
quantile(raw$pregnant, probs=c(.1,.25,.5,.75,.9,.95,
                               0.99, 1), na.rm=TRUE)
```

```
##  10%  25%  50%  75%  90%  95%  99% 100%
##    0    1    3    6    9   10   13   17
```

Observation 2: glucose, diastolic, triceps, insulin, bmi

We observed that the minimum for glucose, diastolic, triceps, insulin, and bmi was 0.

1. The absence of these values is biologically problematic.
2. Consider the sorted values

```
sort(raw$bmi)[1:30]
```

```
##  [1]  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
## [15] 18.4 19.1 19.3 19.4 19.5 19.5 19.6 19.6 19.6 19.9 2
## [29] 20.8 20.8
```

Observation 3: glucose, diastolic, triceps, insulin, bmi

We observed that the minimum for glucose, diastolic, triceps, insulin, and bmi was 0.

```
sort(raw$diastolic)[1:40]
```

```
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [24]  0  0  0  0  0  0  0  0  0  0  0  0 24 30 30 38 40
```

Postential coding of missing values

3. There are 35 subjects with a glucose reading of 0.
4. Given the large number of individuals with 0 and the lack of information about 0s in the data, we may conclude that 0 is a missing value code.
5. Recode these variables to properly account for the missing information.

```
recode <- c("glucose", "diastolic", "triceps",
            "insulin", "bmi" )
for(i in recode){
raw[raw[,i] == 0,i] <- NA
}
```

Numeric summaries: bivariate

Correlation

	glucose	diastolic	triceps	insulin	bmi	diabetes	age
glucose	1.00	0.21	0.20	0.58	0.21	0.14	0.34
diastolic	0.21	1.00	0.23	0.10	0.30	-0.02	0.30
triceps	0.20	0.23	1.00	0.18	0.66	0.16	0.17
insulin	0.58	0.10	0.18	1.00	0.23	0.14	0.22
bmi	0.21	0.30	0.66	0.23	1.00	0.16	0.07
diabetes	0.14	-0.02	0.16	0.14	0.16	1.00	0.09
age	0.34	0.30	0.17	0.22	0.07	0.09	1.00

Correlation and regression

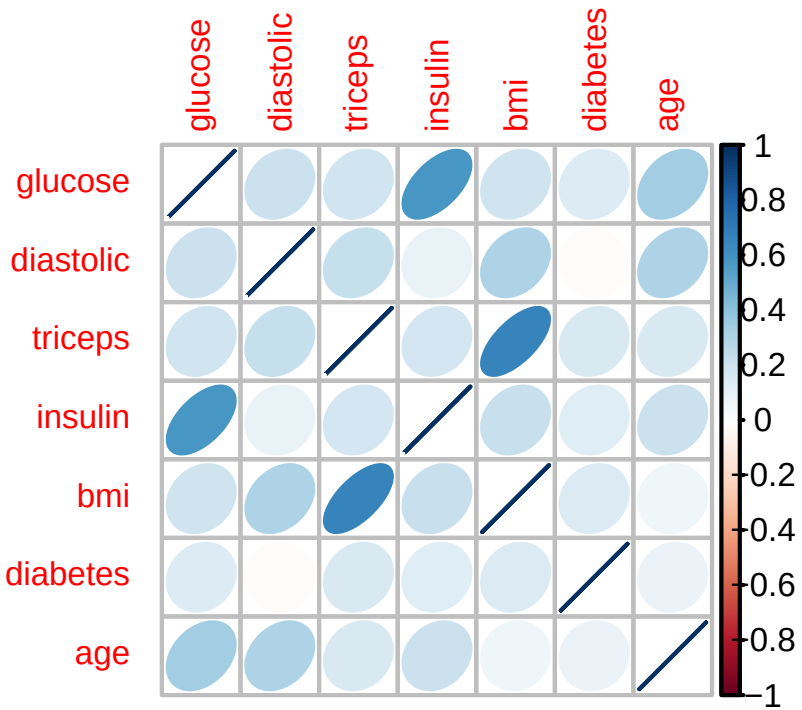
- Regression-type problems date back to 1805 when Legendre developed the method of least squares.
- In 1875, Francis Galton coined the term regression to mediocrity in reference to the relationship

$$\frac{y - \bar{y}}{SD_y} = r\frac{x - \bar{x}}{SD_x}$$

  where r is the correlation between x and y.
- This equation was used to explain the regression effect; sons of tall fathers were not as tall as their fathers and sons of short fathers were not as short as their fathers.

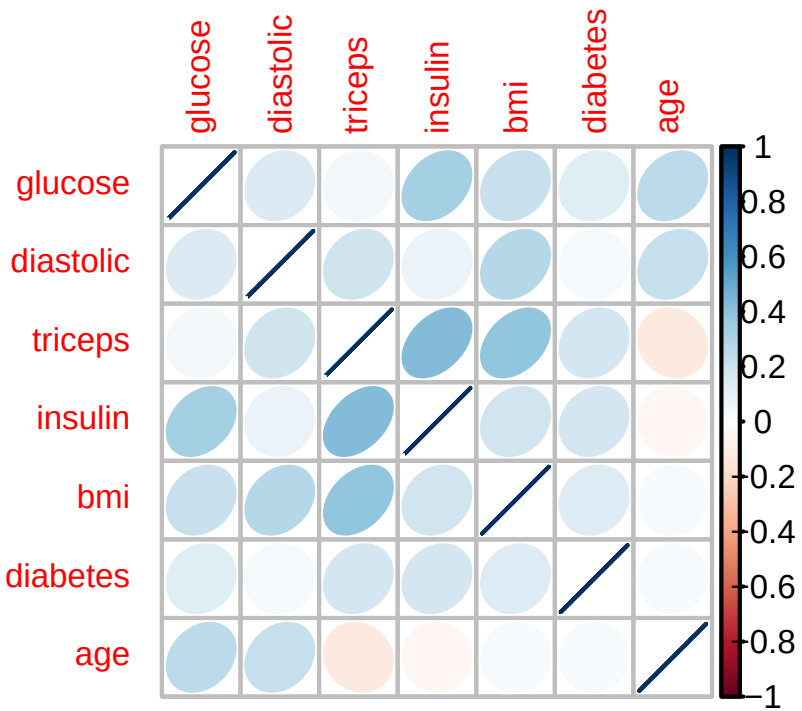
Correlation and correlogram

- The correlogram visualizes the correlation matrix.
- It gives a quick impression of the strength of correlations
- There is a moderate correlation between triceps and BMI, and between glucose and insulin.



Effect of missing data details

- If we used the raw data without investigating it, we would have incorrect bivariate relationships.
- Without recoding the missing data, we observe negative correlations when positive correlations should be observed, and weaker correlations than we should have observed.



Graphical summaries

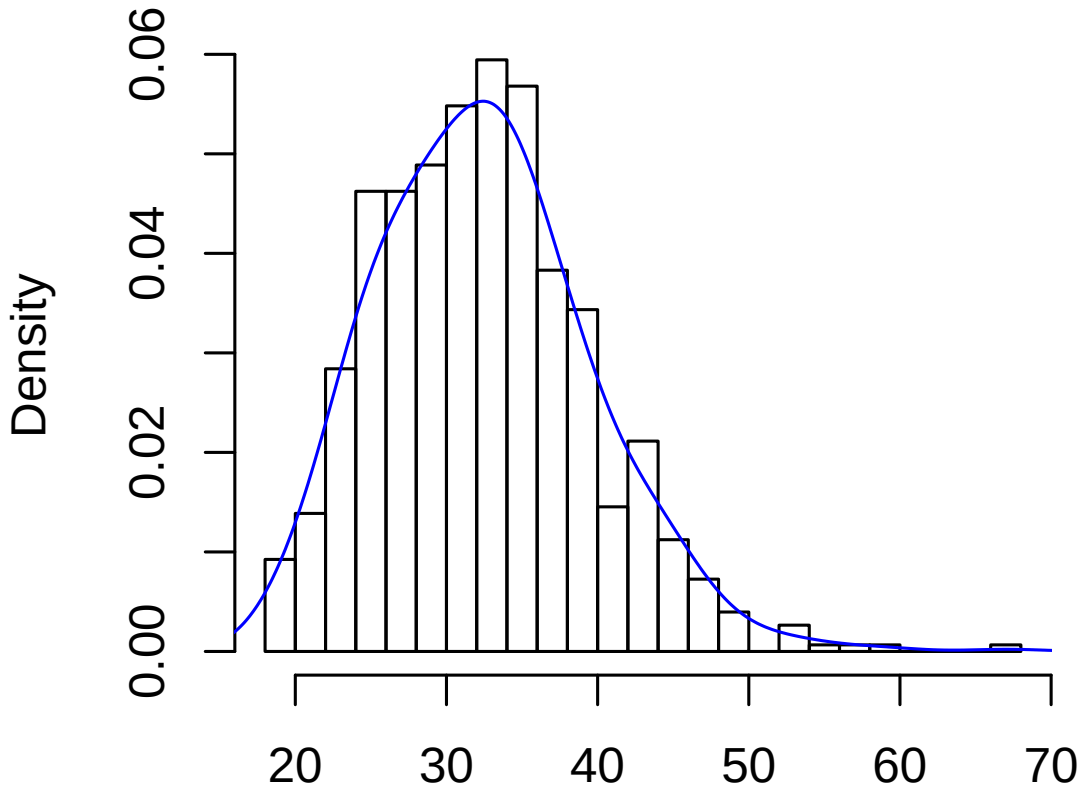
Graphic summaries

Graphical summaries of the data can

1. Describe the distribution
2. Indicate modalities
3. Identify potential outliers
4. Identify potential subgroups

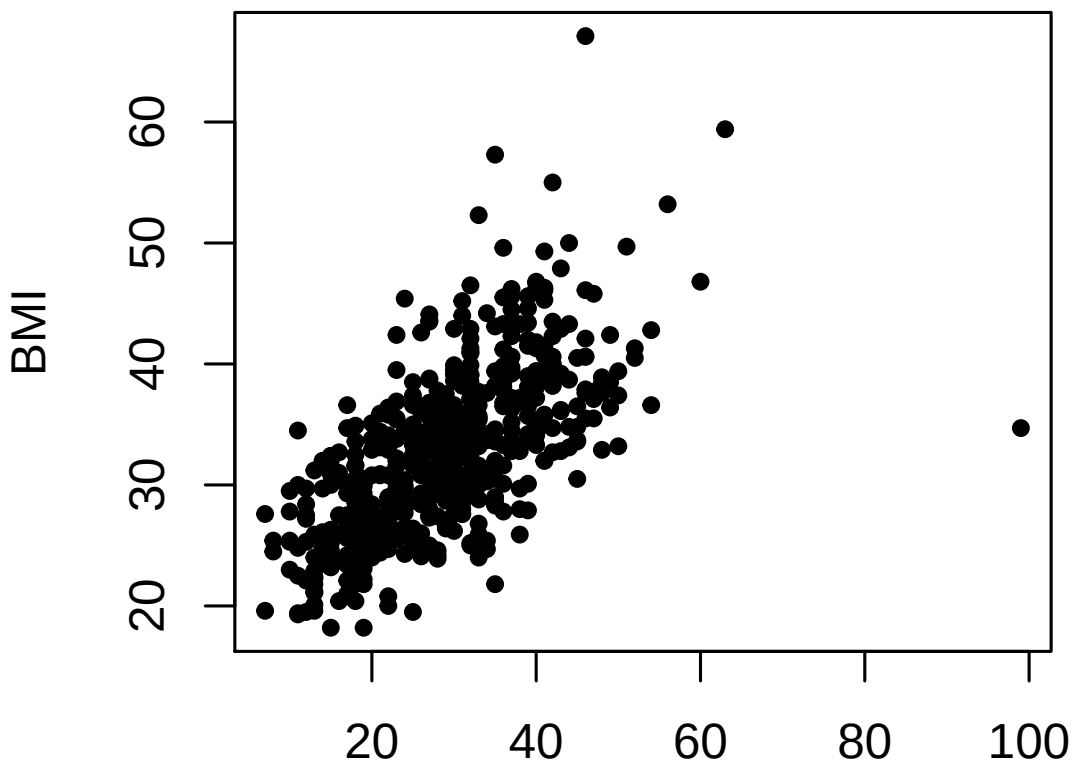
Density histogram, smoothed density

- A histogram is an estimate of the empirical distribution of the variable
- We can smooth the histogram to obtain a general form of the distribution (blue line)



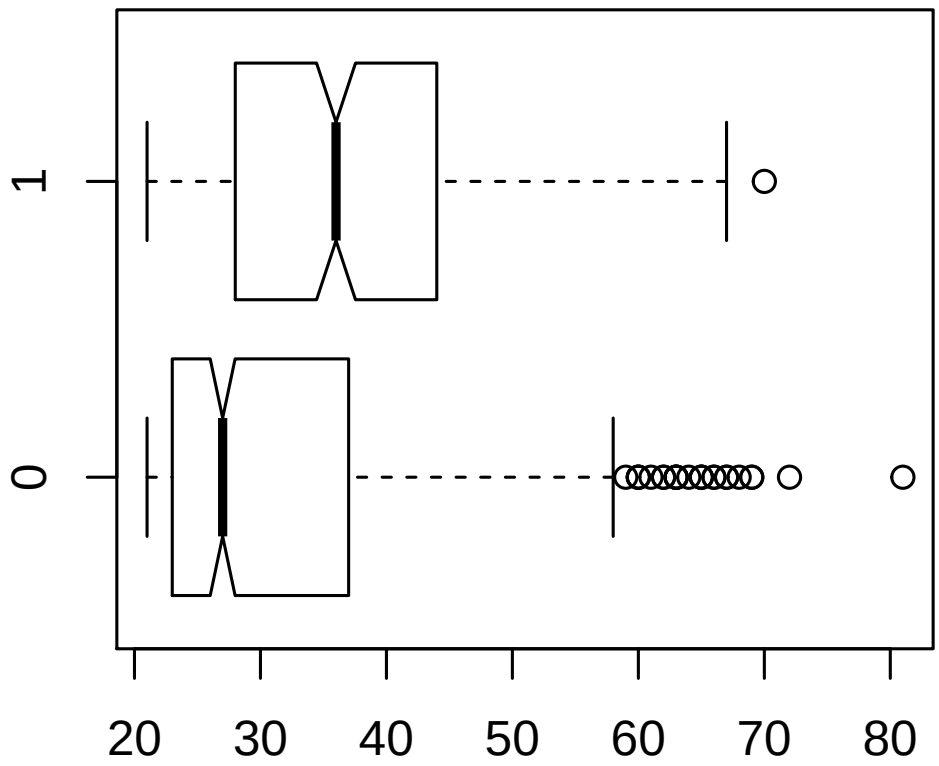
Scatterplots

- Scatterplots are a bivariate visualization
- Scatterplots can identify outliers in multiple dimensions as well as possible subgroups (cluster of outliers)



Comparative boxplots

- Boxplots can be used to compare the distribution of two groups, focusing on the median and interquartile range
- These are violin plots which have information about statistical significance without explicit, formal hypothesis testing



Conclusions

Final thoughts

- Simple approaches can yield deep insights.
- Simple approaches are more easily understood and communicated.
- Data visualization preceding data-preprocessing and statistical modeling is an essential step in the analysis process.
  - Suggest data transformations (e.g. Box-Cox, Box-Tidwell, spatial sign, exponential).
  - Suggest potential model families (e.g. PCA, ridge regression, time-series, non-linear regression).
  - Identify analytic pitfalls.