Practical Data Science for Biologists

Paul Sochacki, MS
Bio Lab Analytics, LLC
03/01/2019

Overview

Introduction

Database Basics (SQL)

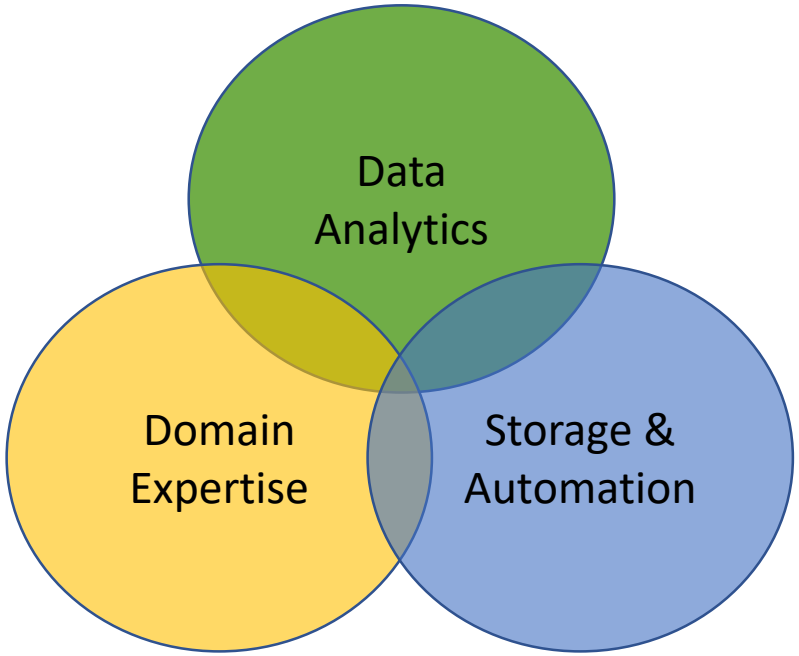
Data Munging

Descriptive Statistics & EDA

Inferential Statistics

Machine Learning Intro

Introduction – What is Data Science?



Introduction – Aspects of Data Science

Data Analytics

- Statistics/Biostatistics
- Data Visualization
- Dashboards & Reporting
- Bioinformatics

Storage & Automation

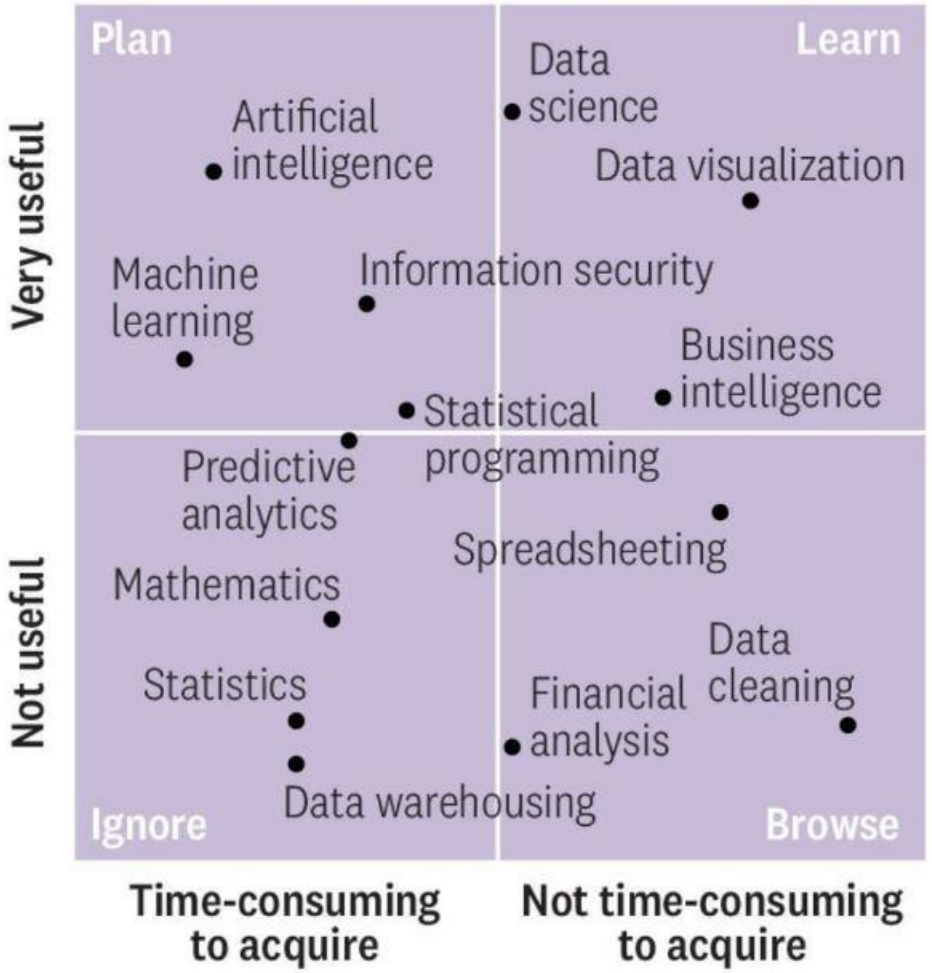
- “Big Data”
- Databases, Pipelines, and Interfaces
- Machine Learning (ML)
- Artificial Intelligence (AI)

Domain Expertise

- Terminology & conventions
- Software requirements/preferences
- Subject matter knowledge
- Organizational requirements

Introduction – Aspects of Data Science

Harvard Business Review, 10/23/2018
(unedited "plot" of data skills)



Introduction – Database Basics

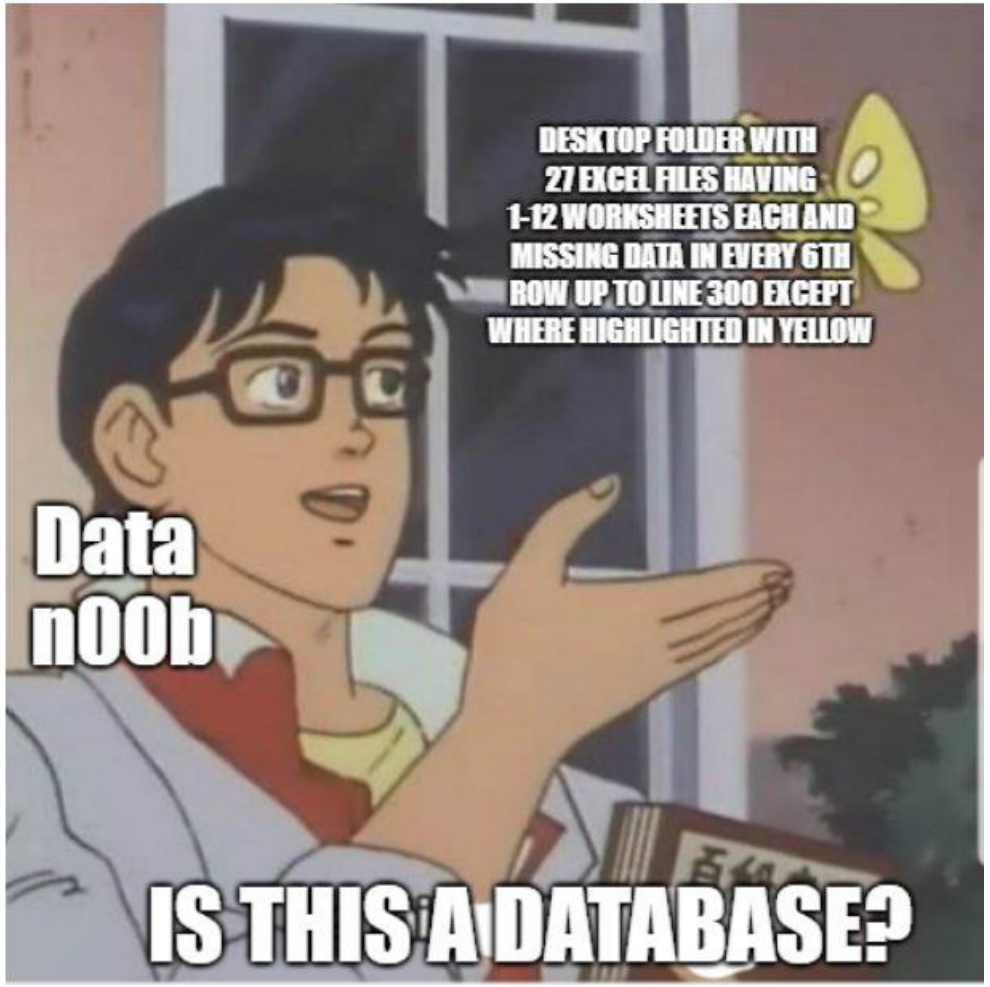
Data Storage Hierarchy

Flat Files (i.e., CSV, Text, Excel)

Databases (Relational, Non-Relational)

Database Management System (DBMS)

Servers (Local, Dedicated, Cloud, Virtual, etc.)



Database Basics – Database Types

- Relational (SQL) databases
  - T-SQL : PostgreSQL, MySQL (MariaDB), MSSS, SQLite
  - PL-SQL : Oracle
  - 90-95% similar syntax

- Non-relational databases
  - MongoDB, NoSQL, Cassandra
  - REDCap at Universities
  - Highly customizable
  - Not as common as relational Db

Database Basics – SQL Language Types

SQL

- SQL = Structured Query Language
- Build Relational Databases
- Analyze & Manipulate Data

DML

- Data Modification Language
- ALTER, CREATE – database level
- DELETE, INSERT, UPDATE – table level

DDL

- Data Definition Language
- SELECT, FROM, ORDER BY, GROUP BY
- Reporting (SUM, AVERAGE, COUNT, etc.)

Database Basics – Relational Db Terminology

Table (Entity)

- Collection of records defining an person, place, thing, or event

Column (Attribute)

- Variable of a fixed data type (text, numeric, image file, etc.) that contains information about an entity

Primary Key (PK)

- Uniquely identifies a single record within a table
- Each table should have one

Foreign Key (FK)

- Lives in one table, references a PK from another table
- Can have multiple FK within a table, depends on relationships

Referential Integrity

- Establishes FK/PK relationships among tables
- Defined through specification of constraints/references in SQL syntax.

Database Basics – PostgreSQL

- PostgreSQL is a popular, open-source database
- pgAdmin4 is a management tool for PostgreSQL databases
- Interface drivers available for both R & Python

Database Basics – Walkthrough

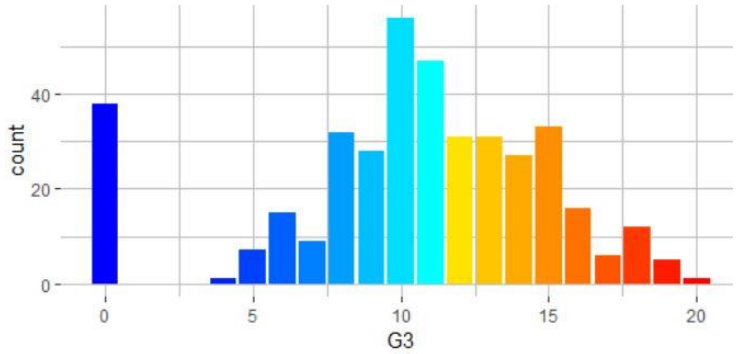
Software Requirements

- PostgresSQL 11.x
- pgAdmin 4.x

PostgresSQL Walkthrough

Data Munging

- **Knowing Your Data**
- **Cleaning Your Data**
- **Transforming Your Data**



Data Munging– Knowing Your Data

Data Processing & Validation

- Missing Data
  - Dealing with it
- Numeric Data Types
  - Quantitative
  - Ranked (Ordinal)
  - Binomial
  - Counts
- Date/Time Data Manipulation
  - ‘lubridate’ package

Data Munging– Knowing Your Data

Missing Data

There are a few different ways to handle this…

- Complete Cases
- Randomly Generate
- Impute Average
- Multiple Imputation

Data Munging– Knowing Your Data

Numeric Data Types

- Quantitative Data
  - Drugs, Hormones, Dates, Times
- Binomial Data
  - Critical Value, Condition (yes/no)
- Ranked (Ordinal)
  - Semi-quantitative ELISA, Age Groups
- Counts
  - Frequency of an event or trait
- Calculated
  - Rates, Ratios, Aggregate Functions

Data Munging– Knowing Your Data

View() generates a spreadsheet view of data frame. Allows sorting but not manipulation of data.

DT::DataTable() allows some convenient and efficient table-level manipulations; however, for some operations must convert data back to Data Frame object for analysis

R Data Objects (*very* limited list)

- Lists, Arrays, Matrices
- Data Frames, Data Tables, Tibbles
- Use str() to return object structure
- 'Tidyverse' objects share a common API

Data Munging– Cleaning Your Data

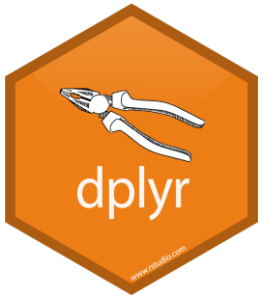
Date/Time Data

A workshop topic in itself

- Date vs DateTime
- Date & Time Formats
- Clock (AM/PM vs 24h)
- Time Zones

Data Munging– Transforming Your Data

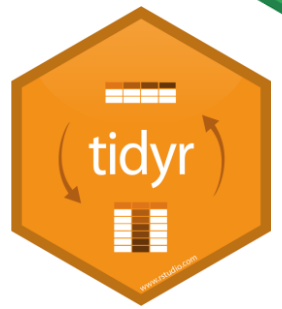
R/tidyverse (‘dplyr’) – official documentation
R/dplyr Tutorial – diagrams, more examples



R/tidyverse (‘stringr’) – for working with strings/text

R/tidyverse (‘lubridate’) – for working with dates/times

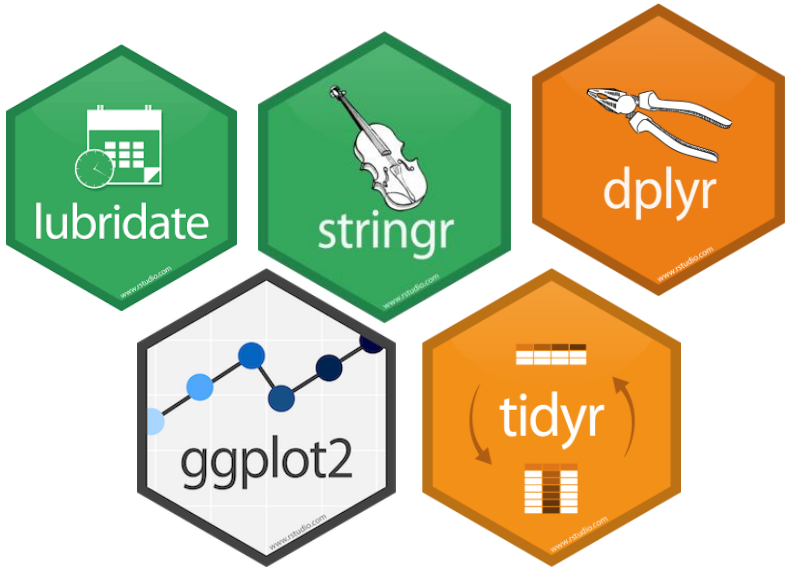
R/tidyverse (‘tidyr’) – for manipulating table structure



Data Munging– Transforming Your Data

R/tidyverse

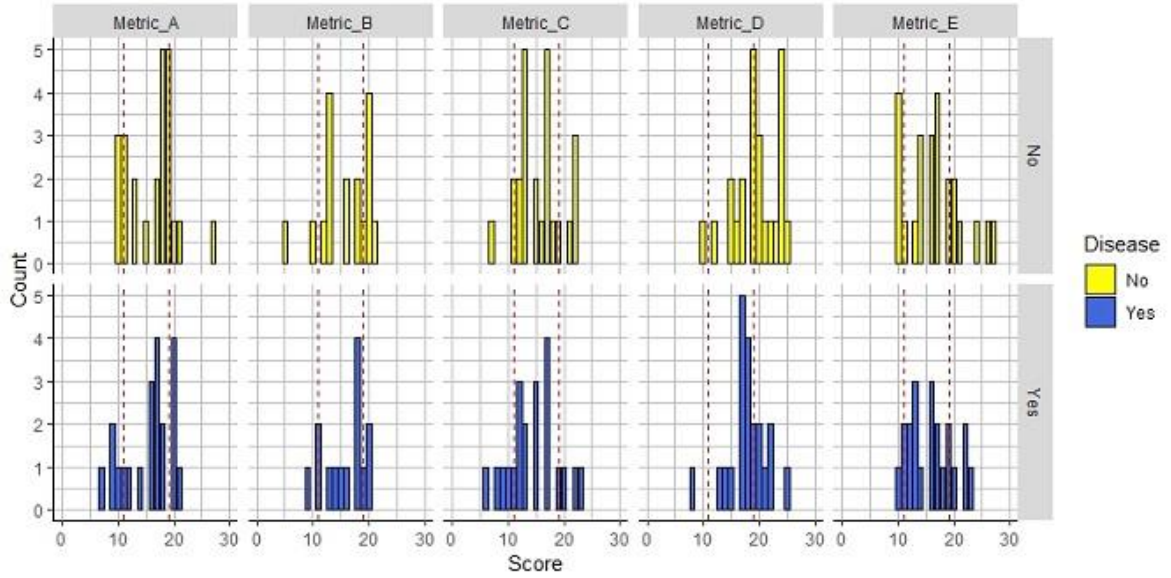
Data Munging Walkthrough



Introduction – Descriptive Statistics

- **Common Descriptive Statistics**

  - Mean (Average), Median, Mode
  - Counts or Frequencies
  - Minimum, Maximum
  - IQR – 25% to 75% percentiles
  - Standard Deviation from Mean (“noise in data”)
  - Standard Error of Mean (“noise around mean”)



Descriptive Statistics – Summary

Base R functions

sum() – column total
mean() – column average
sd() – column standard deviation
length() – number of records
levels() – number of levels within a variable

'psych' package

describe() – basic descriptive stats
describeBy() – grouped descriptive statistics

'dplyr' package

group_by() – group data
summarize() – describe grouped data

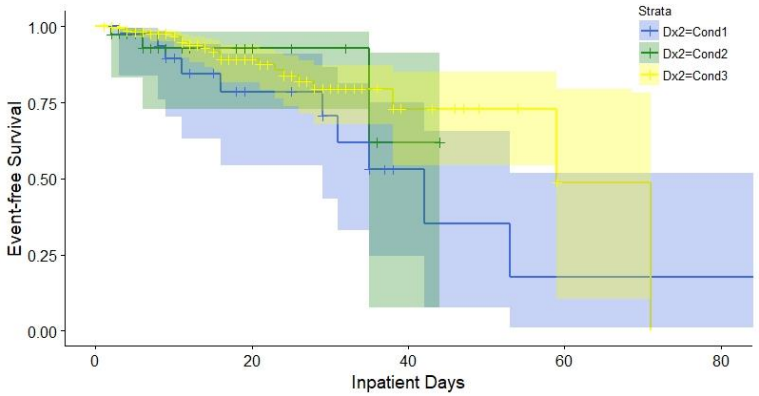
Descriptive Statistics – Visualization

Base R plotting functions

plot(), fit() – scatterplot with fitted line
hist() – histogram frequency plot
boxplot() – box and whiskers plot
barchart() – bar chart plot

Specialty & add-on packages

'PCA3D', 'survival' , 'randomForest' , 'gganimate', etc.
R packages for specialized graphics, animation



Descriptive Statistics – Visualization

R/Tidyverse ('ggplot2')

Standardized plotting functions for R

Pros: produces high-quality, customizable graphics
Cons: time required to learn specialized syntax

'ggplot2' walkthrough

Descriptive Statistics – Visualization

Add-on library ('ggmap')

Extension of 'ggplot2' library functions for overlaying data on geographic maps

Required Components

* *Geographic Info:* Lattitude, Longitude
  - Manually input from file
  - GoogleMaps API (subscription)

* *Density Data:* Population, Infections, Frequency

* *Plot Customization*: Shared API with 'ggplot2'

Statistical Programming Resources

Data Carpentry

Data Analysis and Visualization in R for Ecologists

Biological Statistics Handbook and R Companion

The Handbook

Its Companion

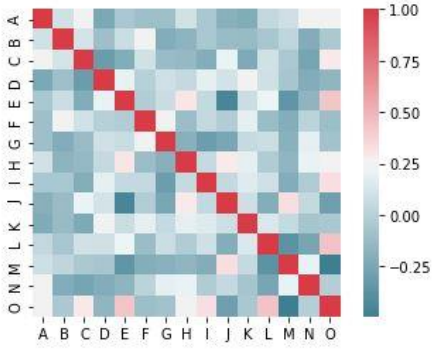
MOOC Learning

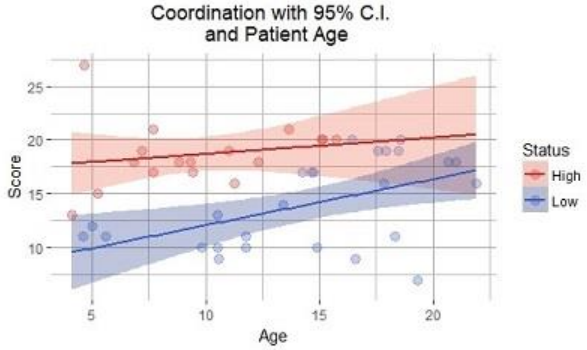
Udemy

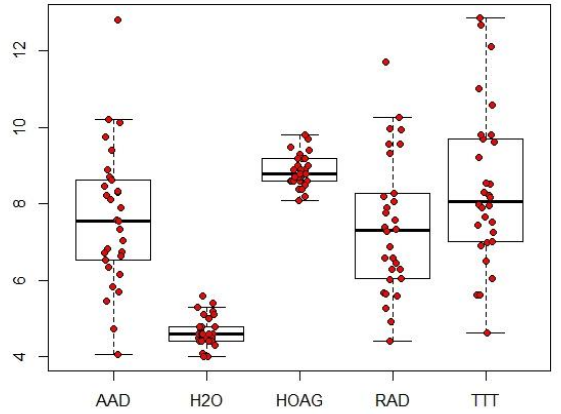
Coursera

Introduction – Inferential Statistics

- **Model Building and Hypothesis Testing**
  - Models often rely on assumptions about underlying the data
  - Lots of very specific terminology… Wikipedia can be helpful, but also has a (deserved?) nickname of "Wickedpedia"…
  - **KNOW YOUR USE CASE**







Inferential Statistics – t-tests

Testing for differences in means between two samples

- Welch's t-tests (replaces ye olde Student's t-tests)
  - Compare means between two groups
  - Both groups assumed to have a normal distribution in response
  - By default, test assumes unequal variance

Testing for before/after mean differences in one group

- Paired T-tests
  - Compare means between two timepoints or treatments for group of individuals
  - Subjects must have measures for both times/treats
  - Same assumptions as for regular t-test
  - More power than two-sample t-test

Inferential Statistics – t-test Assumptions

T-test Assumptions - Diagnostics

Normal (Gaussian) distribution among residuals of response variable… however, often very similar to distribution of response variable, in practice.

shapiro.test()

- conventional cut-off = 0.90
- marginally acceptable = 0.80
- square-root and log transformations can help

qqnorm()

- generates Q-Q plot of residuals
- “eyeballing” the data is a common practice

Inferential Statistics – Statistical Power

Power and Sample Size Calculations

Power Analysis

- Given means, standard deviation, number of experimental units (N), & alpha, calculate power

Sample Size Analysis

- Given means, standard deviation, power, and alpha, calculate sample size required

Base R has functions for power calculations, including for t-tests and ANOVA models

Generally speaking, a paired t-test (or repeated measures ANOVA) will have greater power because it accounts for intra-individual variability among measurements

Inferential Statistics – Statistical Power

Statistical Power and Significance

Table 1. Types of Statistical Errors

	H_0 is actually:	
	True	False
Reject H_0	Type I error	Correct
Accept H_0	Correct	Type II error

A **Type I error** is often represented by the Greek letter alpha (α) and a Type II error by the Greek letter beta (β). In choosing a level of probability

Alpha (α) - % chance of getting a false positive

Beta (β) - % chance of getting a false negative

Inferential Statistics – ANOVA

Testing for differences in means among 2+ groups

- Analysis of Variance (ANOVA)
  - Compare means among multiple groups
  - Response assumed to have a normal distribution
  - For two groups, essentially same as a t-test but has an F-value instead of a t-value

Testing for differences in means, adjusting for multiple measurements on the same subject

- Repeated Measures ANOVA
  - Compare means across timepoints or treatments for group of individuals
  - Each individual has data for each interval
  - Same assumptions as for ANOVA
  - Analogous to paired t-test but for >2 groups

Inferential Statistics – ANOVA Assumptions

ANOVA Modeling Assumptions

Multicollinearity

- Correlation among predictors, can skew estimates
- 'ols_corr' package
- Verify VIF < 4.0 for all predictors
- If VIF > 4.0 then should drop from the model, but this can still be problematic sometimes…

Balanced design

- Ideally similar sample size for all treatments
- Generic rule of thumb is 10 replicates/treatment
- Unequal cell-sizes will not necessarily "ruin" an analysis; however, estimates of effect size (i.e., coefficients) may be skewed

Inferential Statistics – Correlation Analysis

Simple Correlation

- Describes general relationship between two continuous variables
- Can be positive or negative ($-1.0 \leq r \leq 1.0$)
- Has a p-value, interpreted as consistency
- Correlation ≠ Causation

Correlation Matrix

- Describes general relationships between many continuous (sometimes ranked) variables
- Can be positive or negative ($-1.0 \leq r \leq 1.0$)
- Has a p-value, interpreted as consistency
- Often displayed graphically with heatmaps

Inferential Statistics – Regression Analysis

Simple Regression

- Defines quantitative relationship between two continuous variables
- Positive or negative slope ($-\infty \leq \beta \leq +\infty$)
- P-value : interpreted as consistency
- R^2 : interpreted as strength of relationship
- Requires SME input to validate model

Multiple Regression

- Defines quantitative relationships among 2+ continuous predictors and an outcome
- All the same statistics as simple regression
- Model selection can benefit from comparison of Akaike's Information Criterion (AIC) scores
- AIC applies penalty as more predictors are added

Inferential Statistics – Resources

Biological Statistics Handbook and R Companion (also SAS)
The Handbook
Its Companion

Statistical Consulting, Resources, and Workshops
The Analysis Factor

Statistical Theories, Approaches, and Discussions
Stack Exchange

Machine Learning – Introduction

“Statistics” for Automation & Big Data

- Machines are not that smart, really
- For smaller data sets, ML methods often provide same answer as statistics
- Often few (if any) assumptions about underlying data
- Instead, assumptions are made at system level

Machine Learning– General Concepts

How does it work?

- Some data is collected about a system of interest
- Assumption made that the system is very well characterized

- System may include elaborate data pipeline with numerous inputs (sources)
- Clouds, servers, devices, sensors, etc., can all be sources of data

- A model type fitting the use case is selected
  - ML Linear Regression
  - ML Logistic Regression
  - SVM (basically Chi-square)
  - Cluster Analysis
  - Naïve Bayes

Machine Learning– General Concepts

Training & Validation

- Collected data is split into “Training” and “Validation” data sets
- Specialized ML libraries such as R ‘caret’ and Python ‘scikit-learn’ are used to split the data into training & validation sets
  - 80% vs 20% split (Training vs Validation)
- Validation set is tested against model built using the Training set data
- Software reports metrics for precision, recall, sensitivity, and/or specificity
- Check metrics, possibly repeat process

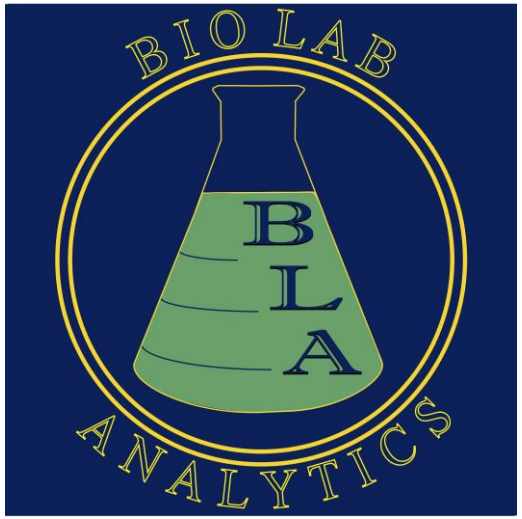
Machine Learning– General Concepts

Production

- If metrics are acceptable, move the ML model into Production and use it to predict/identify patterns based on the Training set. Ideally, refresh the training set every once in a while.
- ML models in Production vary widely in performance – context matters
- ***Feature selection*** (also known as variable or attribute selection) selection is active area of computer science & math research
- Commonly found in retail & web analytics, some recent applications in life science (monitoring & diagnostic devices)

Thank you ORTWS!

Dr. Meghan Martin-Wintle for (ORTWS) President !



Follow us on Twitter!
#rstats #datascience #dataviz

SQL script, data sets, & R script from this workshop available on GitHub
https://github.com/Cyberskout99/DSBworkshop