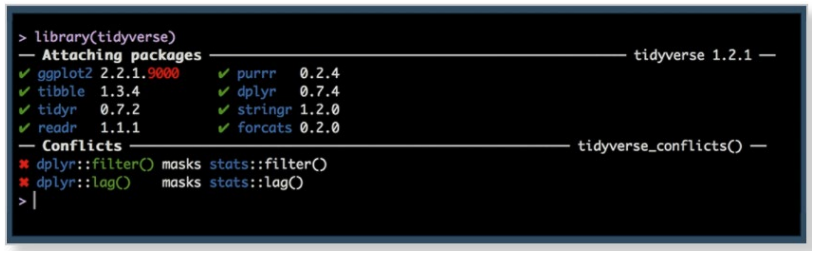
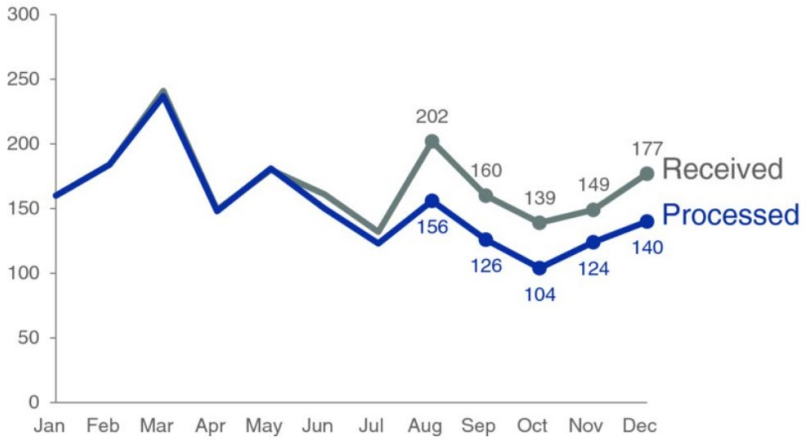
Midterm Exam Review

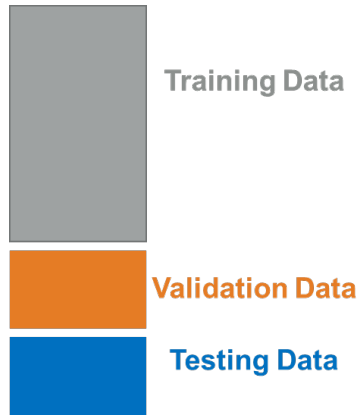
John Rios

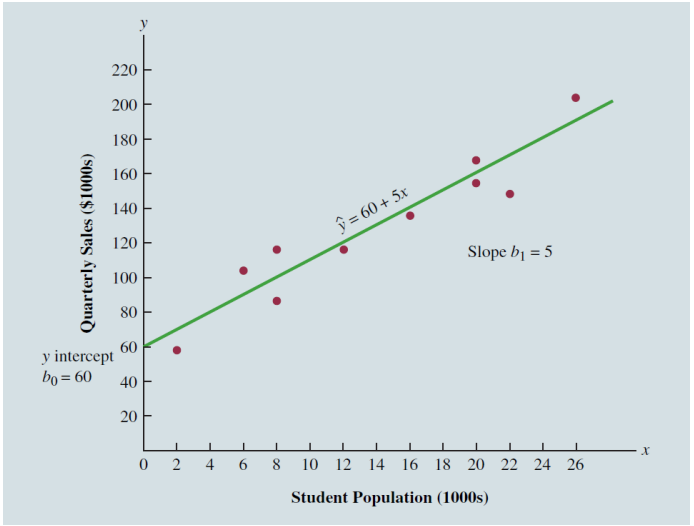
Business Intelligence and Analytics

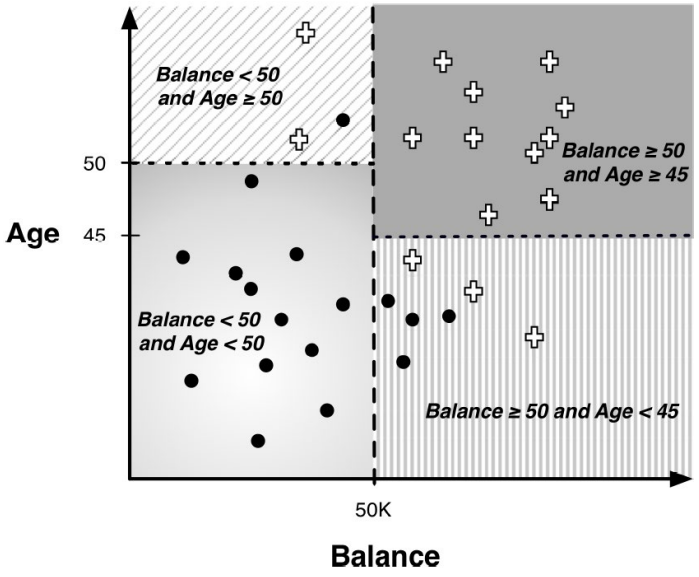
Terry College of Business
UNIVERSITY OF GEORGIA











Data Value Creation Model

- Connect Tableau to various Datasets: Excel and CSV files

Import

- Create Area Charts
- Create Barcharts
- Create Scatterplots
- Create Maps
- Create Treemaps
- Create Storylines
- Create Interactive Dashboards

ggplot2

Explore

```
> # how did we do with test set? confusion matrix
> confusionMatrix(data = churn_test$pred_churn,
+                 reference = churn_test$Churn,
+                 mode = "prec_recall",
+                 positive = "Yes")
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  906 186
       Yes 126 187

               Accuracy : 0.7779
                 95% CI : (0.7553, 0.7994)
    No Information Rate : 0.7345
    P-Value [Acc > NIR] : 9.989e-05

                  Kappa : 0.3998

 Mcnemar's Test P-Value : 0.0008371

              Precision : 0.5974
                 Recall : 0.5013
                     F1 : 0.5452
             Prevalence : 0.2655
         Detection Rate : 0.1331
   Detection Prevalence : 0.2228
      Balanced Accuracy : 0.6896

       'Positive' Class : Yes
```

Root Mean Squared Error (RMSE)

Evaluate

Question

Problem Formulation & Research Design

Wrangle

- Understand Types of Joins and how they work

CLASSICAL MACHINE LEARNING

Model

```
case_when(condition-1 ~ output-1  # IF
         ,condition-2 ~ output-2  # ELIF
         ,condition-3 ~ output-3  # ELIF
         ,TRUE ~ output-else      # ELSE
         )
```

actionable

Test

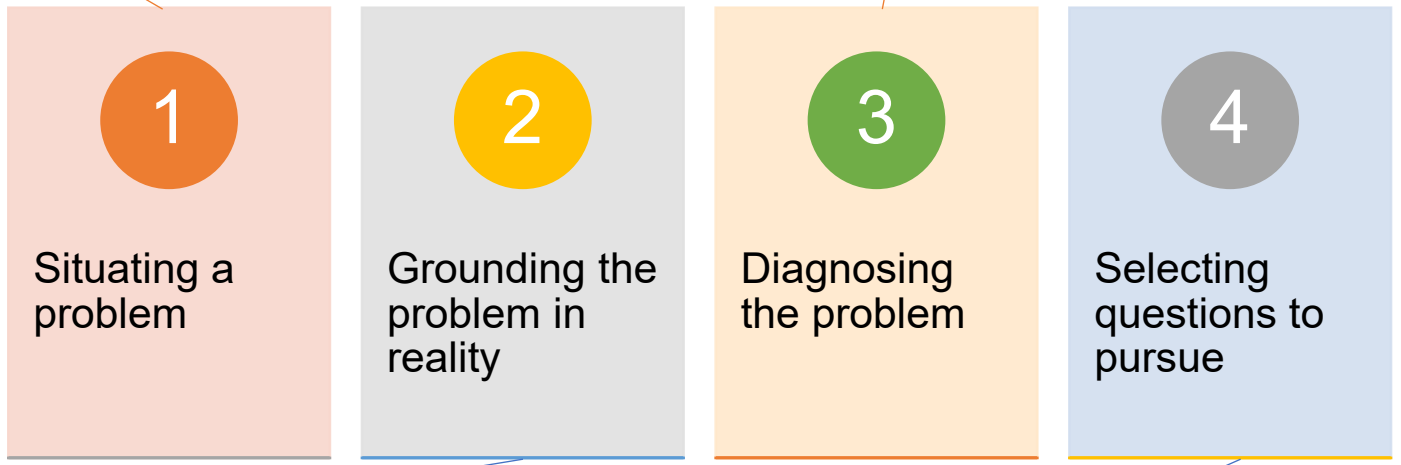
Deployment

Database → Statistics → Machine Learning

Problem Formulation

Identify whose point of view
Who are the intended users?
Foreground vs background
What is focus, level and scope?

Classify elements of the problem into categories
Aggregate categories to infer problem
Heuristic match and refine solutions
Does a solution exist? What anomalies surfaced?



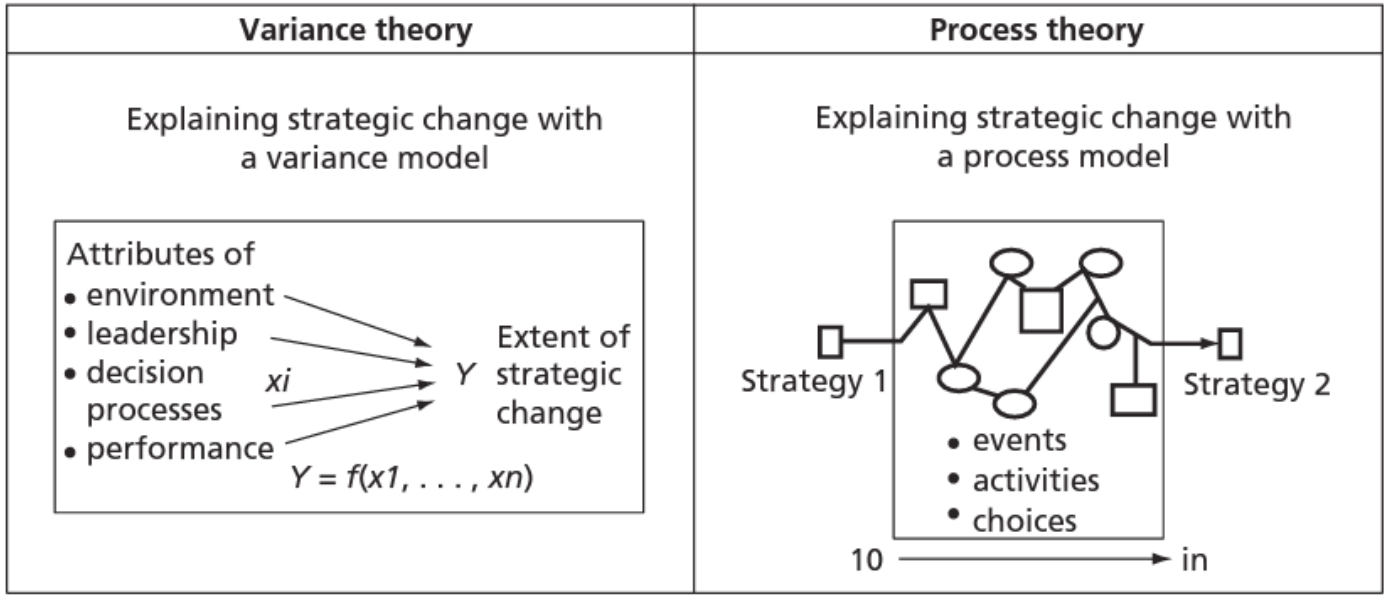
Address who, what, when, why, where
Describe the problem up-close and in general
Talk to people who experience the problem
Read articles, studies about the problem

What part of the problem merits attention?
Ensure you have a question, not a statement
Connect the question with the problem
Consider consequences

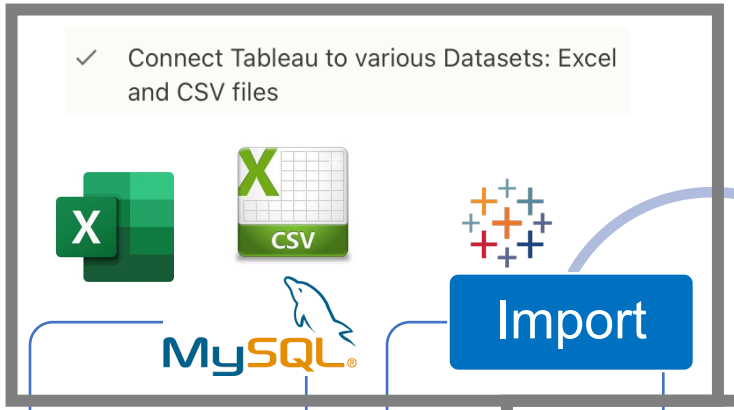
Variance and Process Models

Variance and process models are used to empirically examine two different types of research questions that are often asked about an issue being studied:

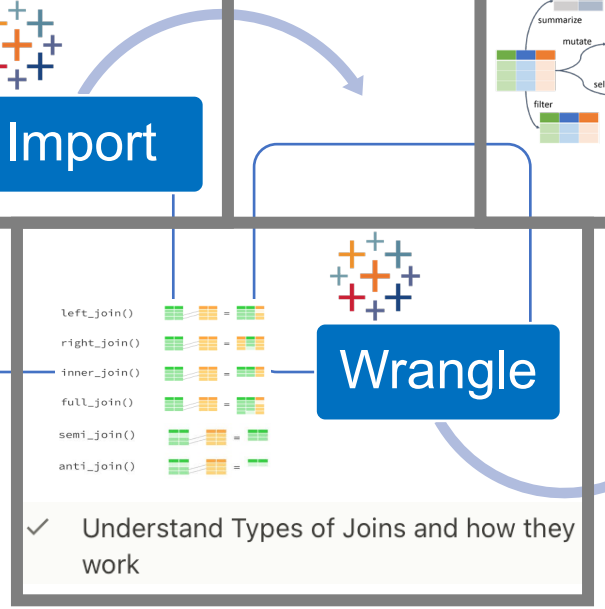
- What are the antecedents or consequences of the issue?
- How does the issue emerge, develop, grow, or terminate over time?



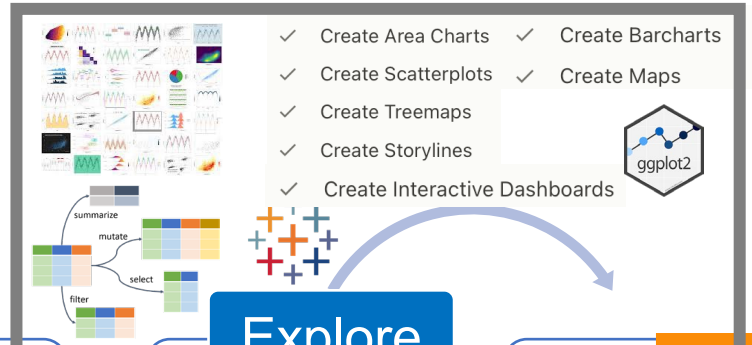
Data Value Creation Model



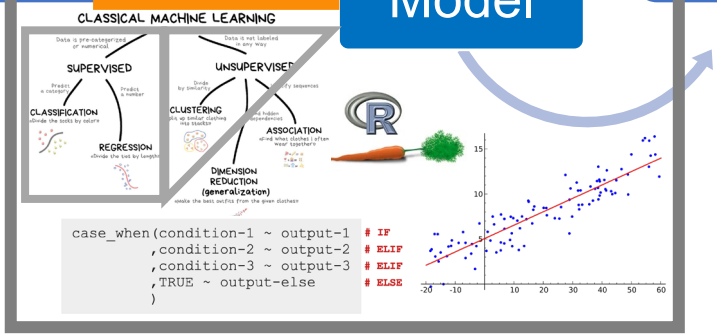
Import



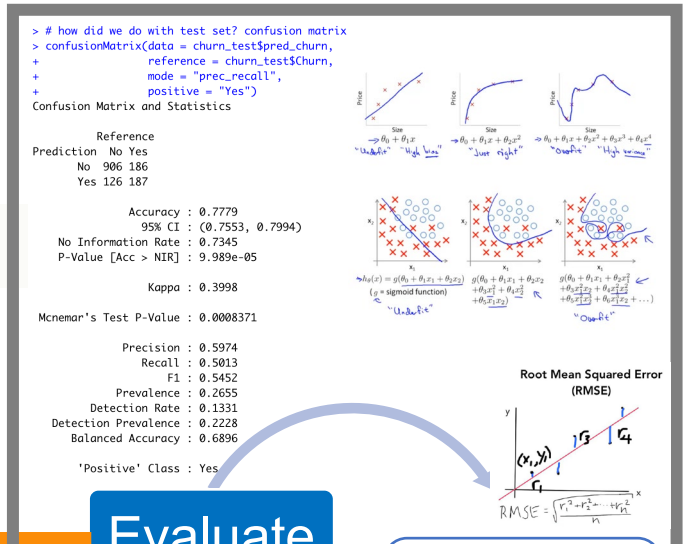
Wrangle



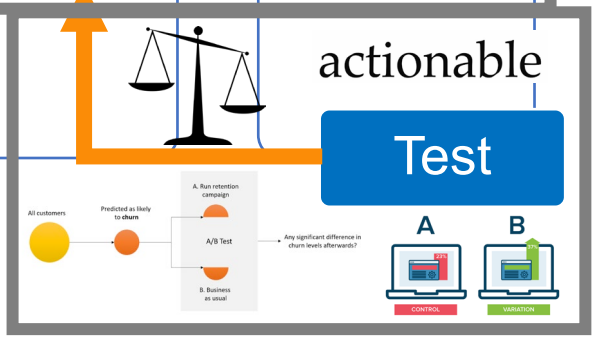
Explore



Model



Evaluate



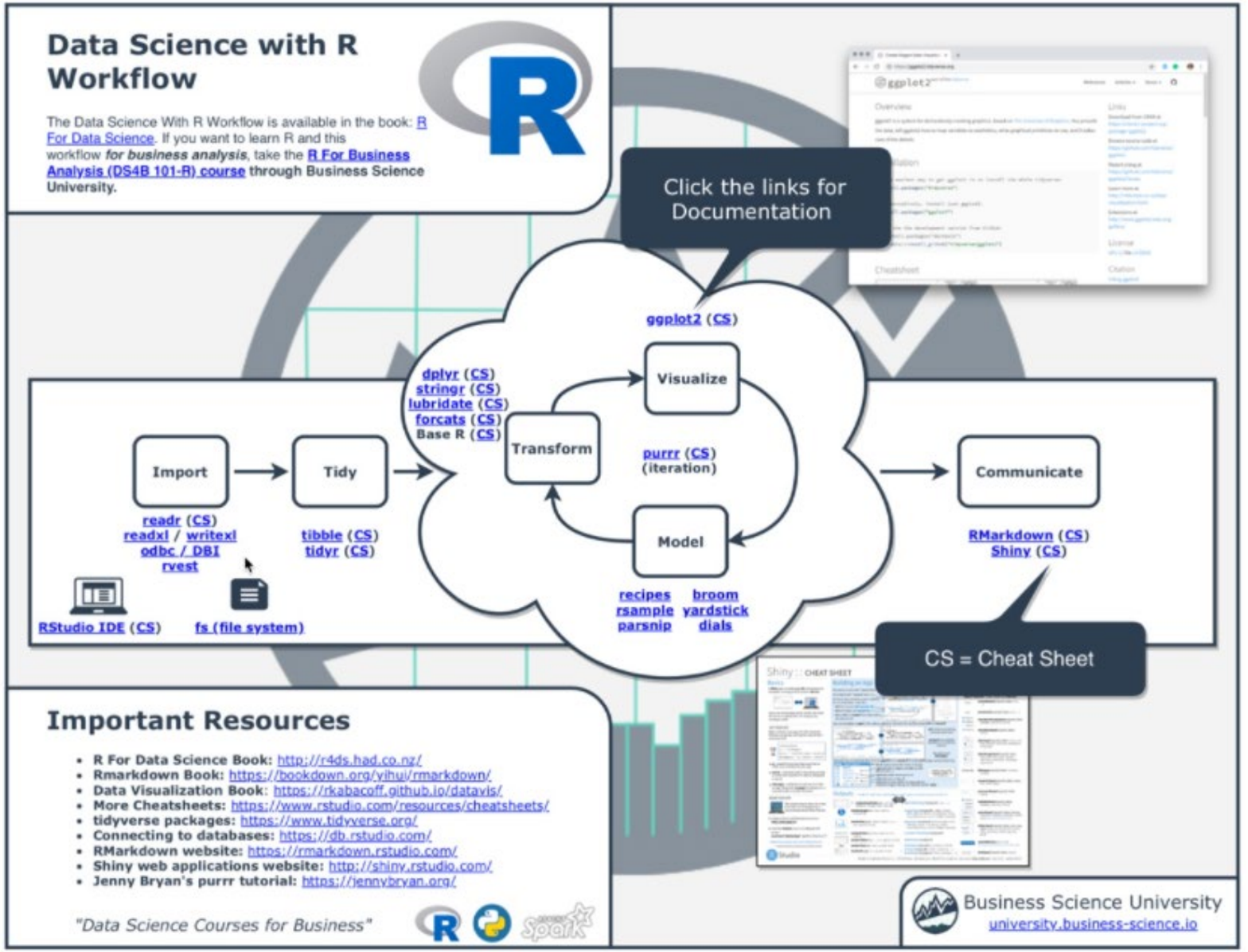
Test

Deployment

Question

Problem Formulation & Research Design

Database → Statistics → Machine Learning


Data Science with R Workflow
The Data Science With R Workflow is available in the book: R For Data Science. If you want to learn R and this workflow for business analysis, take the R For Business Analysis (DS4B 101-R) course through Business Science University.
Click the links for Documentation
ggplot2 (CS)
dplyr (CS)
stringr (CS)
lubridate (CS)
forcats (CS)
Base R (CS)
Visualize
Transform
purrr (CS)
(iteration)
Model
Import
Tidy
Communicate
readr (CS)
readxl / writexl
odbc / DBI
rvest
tibble (CS)
tidyr (CS)
RMarkdown (CS)
Shiny (CS)
RStudio IDE (CS)
fs (file system)
recipes
rsample
parsnip
broom
yardstick
dials
CS = Cheat Sheet
Important Resources
R For Data Science Book: http://r4ds.had.co.nz/
Rmarkdown Book: https://bookdown.org/yihui/rmarkdown/
Data Visualization Book: https://rkabacoff.github.io/datavis/
More Cheatsheets: https://www.rstudio.com/resources/cheatsheets/
tidyverse packages: https://www.tidyverse.org/
Connecting to databases: https://db.rstudio.com/
RMarkdown website: https://rmarkdown.rstudio.com/
Shiny web applications website: http://shiny.rstudio.com/
Jenny Bryan's purrr tutorial: https://jennybryan.org/
"Data Science Courses for Business"
Business Science University
university.business-science.io

Why Tidyverse?

Gather, Store, Access

- Read (readr, readxl)
- Scrape (rvest)
- Database (DBI)
- Web APIs (httr)
- XML (xml2)
- JSON (jsonlite)

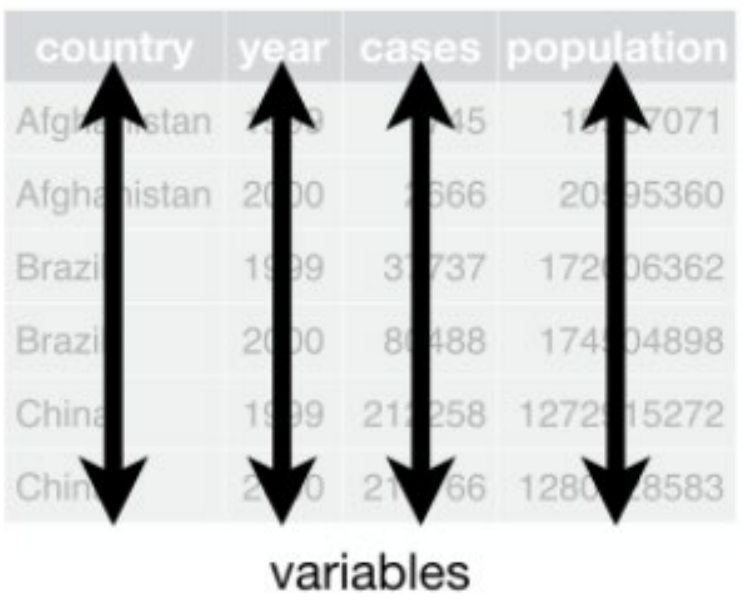
Analyze

- Visualize (ggplot2)
- Transform (dplyr)
- Wrangle (tidyr)
- Model (modelr)

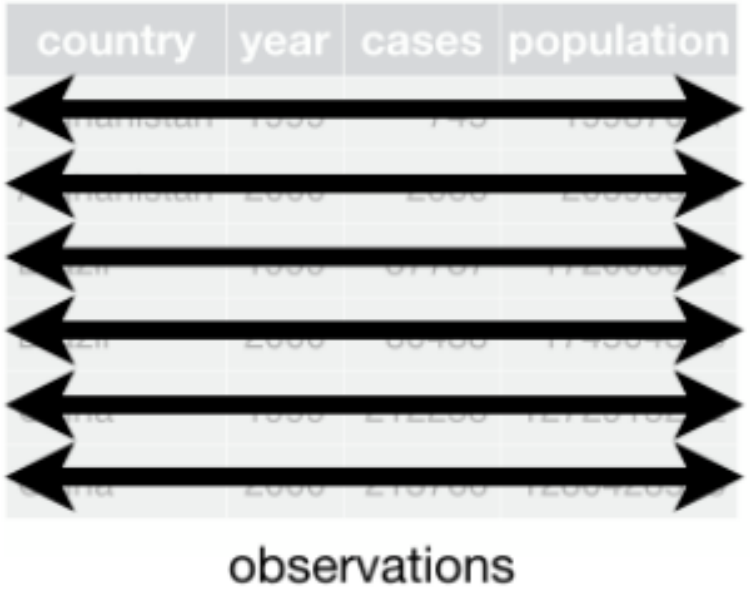
```
> library(tidyverse)
── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
✔ ggplot2 2.2.1.9000     ✔ purrr   0.2.4
✔ tibble  1.3.4          ✔ dplyr   0.7.4
✔ tidyr   0.7.2          ✔ stringr 1.2.0
✔ readr   1.1.1          ✔ forcats 0.2.0
── Conflicts ────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
>
```

Tidy Data

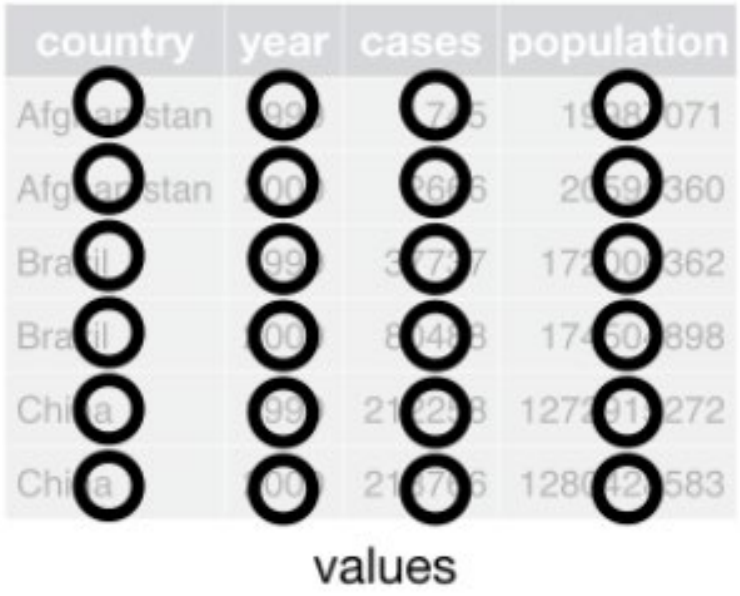
Each **variable** must have **its own column**.



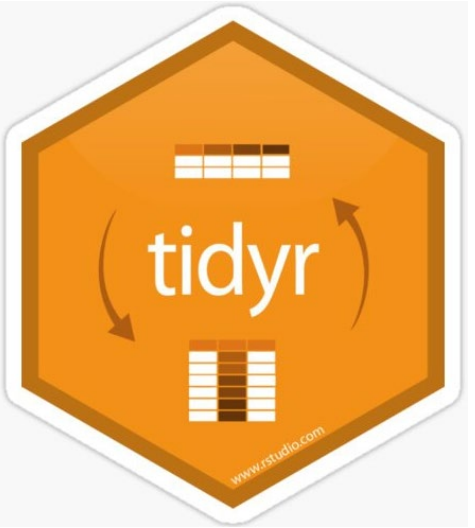
Each **observation** must have **its own row**



Each **value** must have **its own cell**



Pivoting



Most untidy datasets:

1. One **observation** scattered across **multiple rows**. **Pivot Wider**

2. One **variable** spread across **multiple columns**. **Pivot Longer**

Additional transformations

Other types of “not so common” problems

1. **One column** contains data about **two variables**. **Separate**.

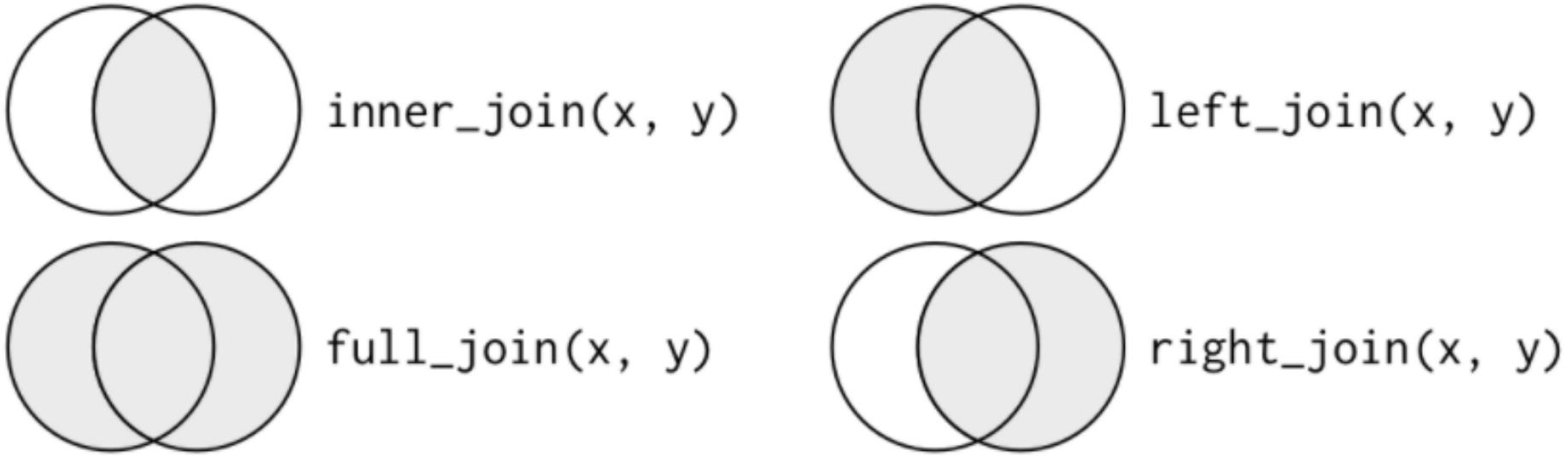
2. **Two columns** contain data about **one variable**. **Unite**.

dplyr – additional functions

dplyr also enables the transformation of data

- Look at a subset of the rows—**filter()**
- Reorder rows—**arrange()**
- Rename variables—**rename()**
- Create new variables—**mutate()**
- Collapse values down to a summary—**summarise()**

Mutating Joins



A Few Lessons from Knaflic (2015)

1. Understand the context
2. Choose an appropriate visual display
3. Eliminate clutter
4. Focus attention where you want it
5. Think like a designer
6. Tell a story

Knaflic, C. G. (2015). Storytelling with Data.

A Few Lessons from Knaflic (2015)

1. Understand the context
2. Choose an appropriate visual display
3. Eliminate clutter
4. Focus attention where you want it
5. Think like a designer
6. Tell a story

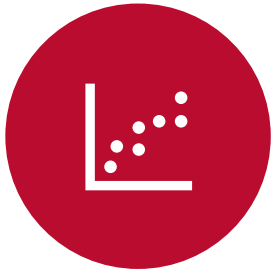
Knaflic, C. G. (2015). Storytelling with Data.

A Few Lessons from Knaflic (2015)

1. Understand the context
2. Choose an appropriate visual display
3. Eliminate clutter
4. Focus attention where you want it
5. Think like a designer
6. Tell a story

Knaflic, C. G. (2015). Storytelling with Data.

Choosing an effective visual – graphs



POINTS

LINES

BARS

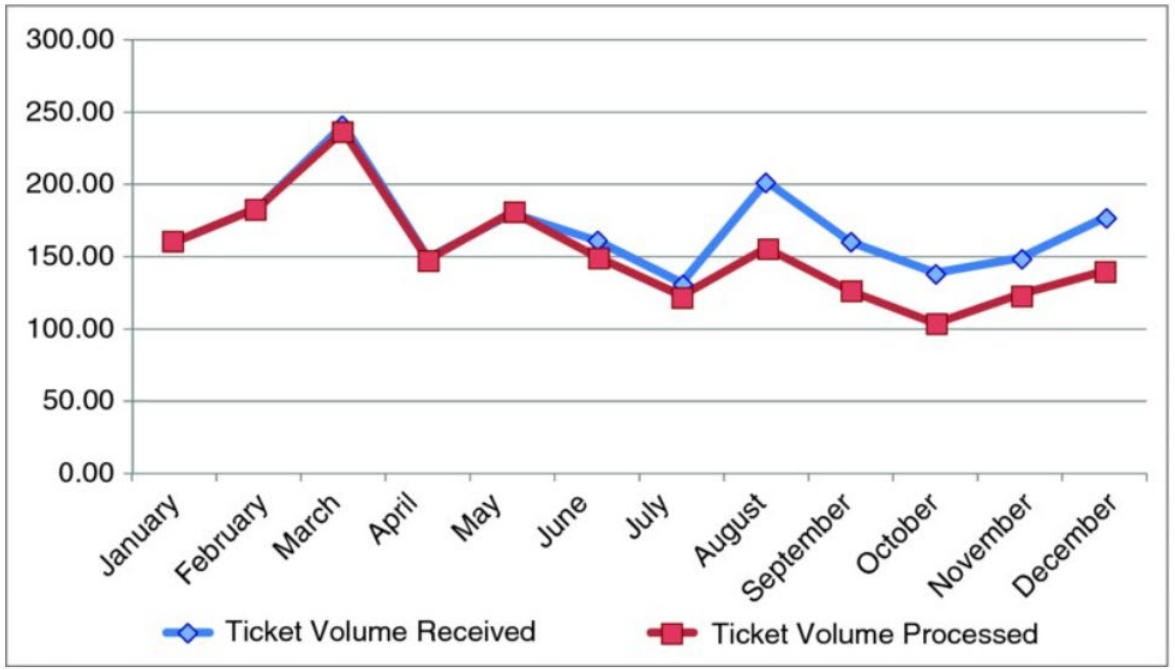
AREA

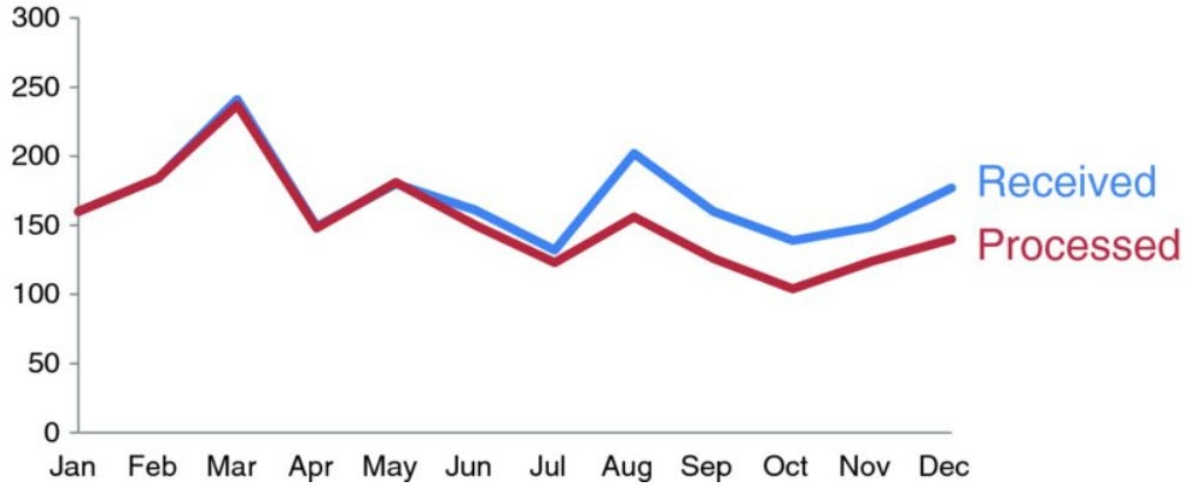
A Few Lessons from Knaflic (2015)

1. Understand the context
2. Choose an appropriate visual display
3. Eliminate clutter
4. Focus attention where you want it
5. Think like a designer
6. Tell a story

Knaflic, C. G. (2015). Storytelling with Data.

Before and after





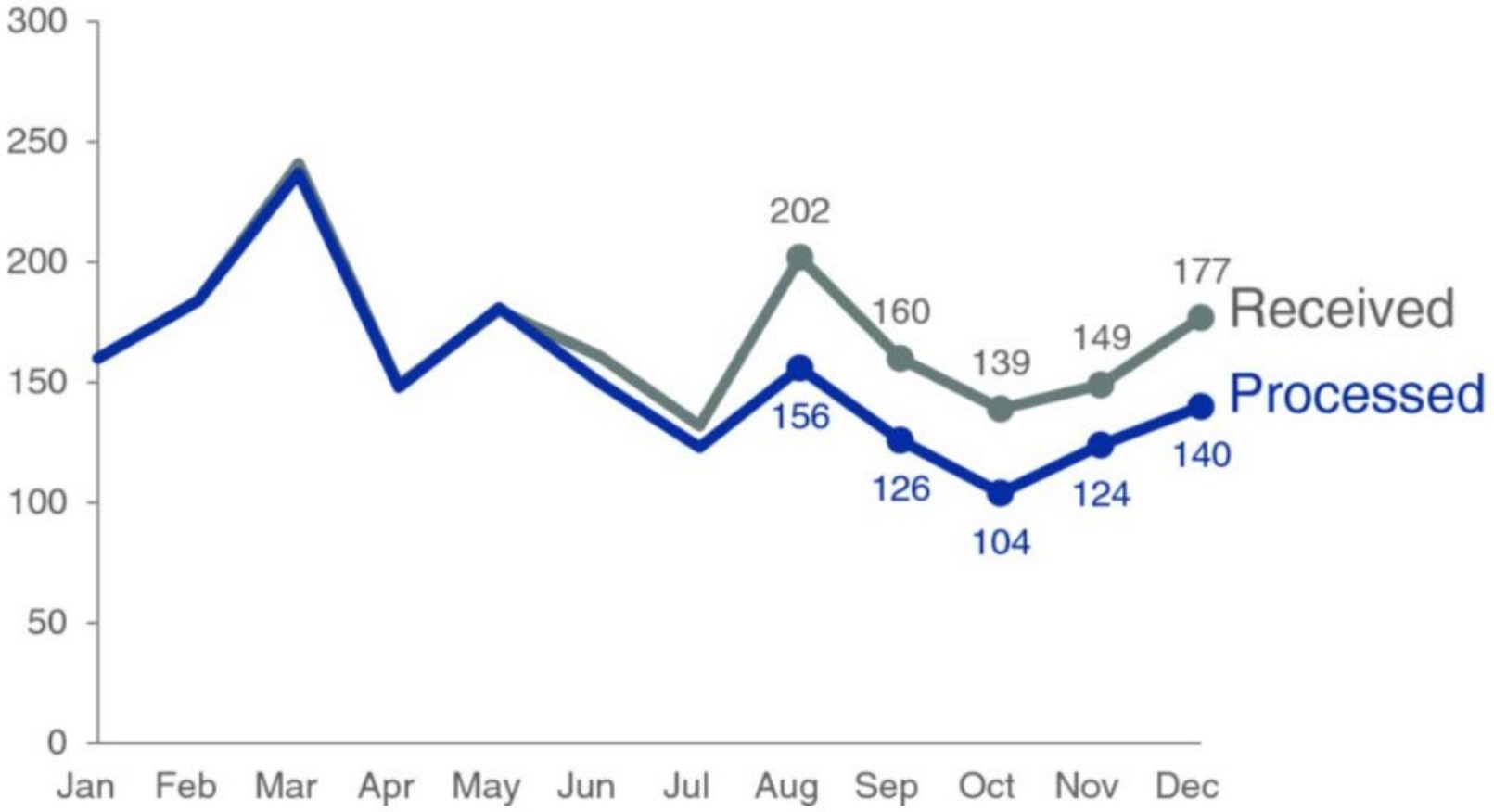
Knaflic, C. G. (2015). Storytelling with Data.

A Few Lessons from Knaflic (2015)

1. Understand the context
2. Choose an appropriate visual display
3. Eliminate clutter
4. Focus attention where you want it
5. Think like a designer
6. Tell a story

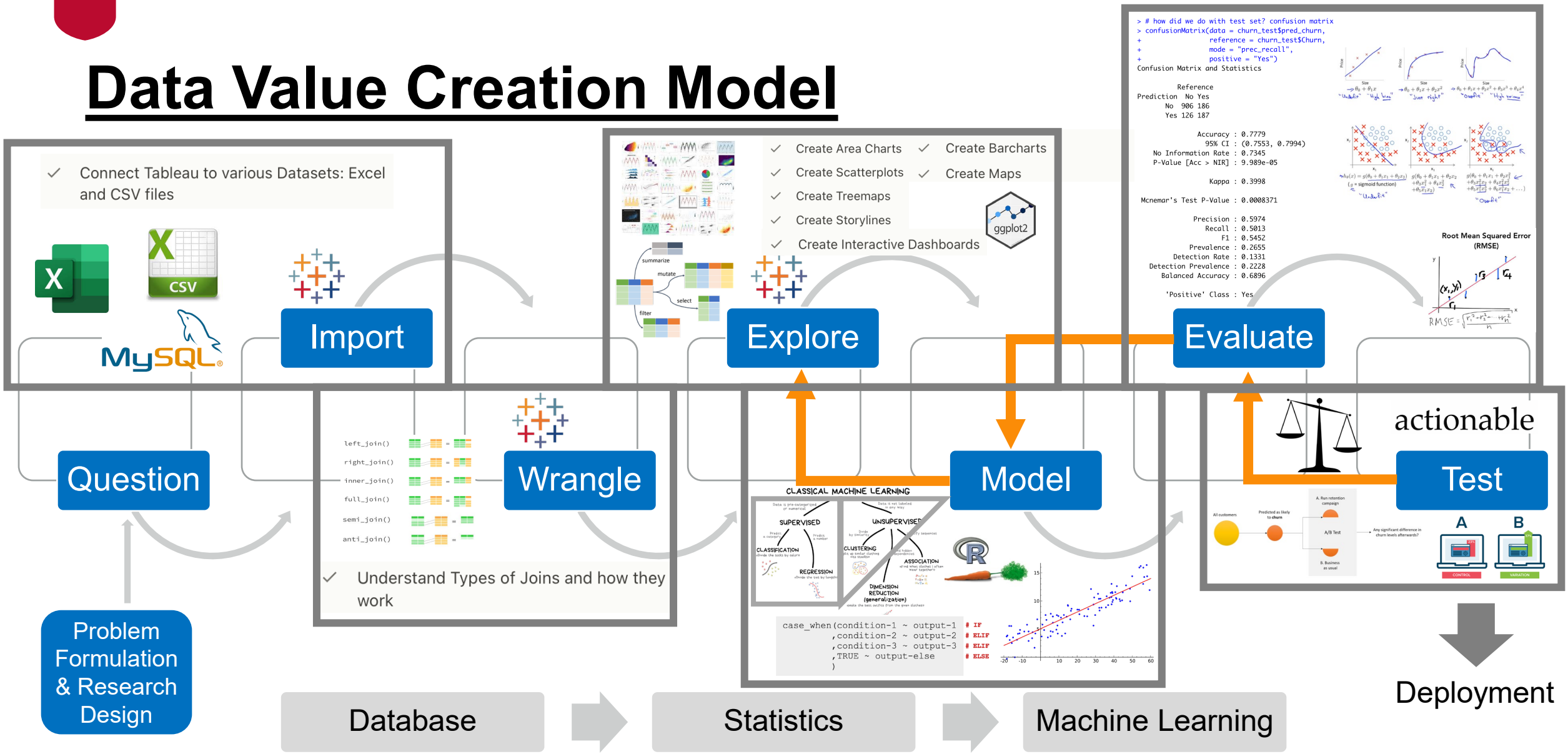
Knaflic, C. G. (2015). Storytelling with Data.

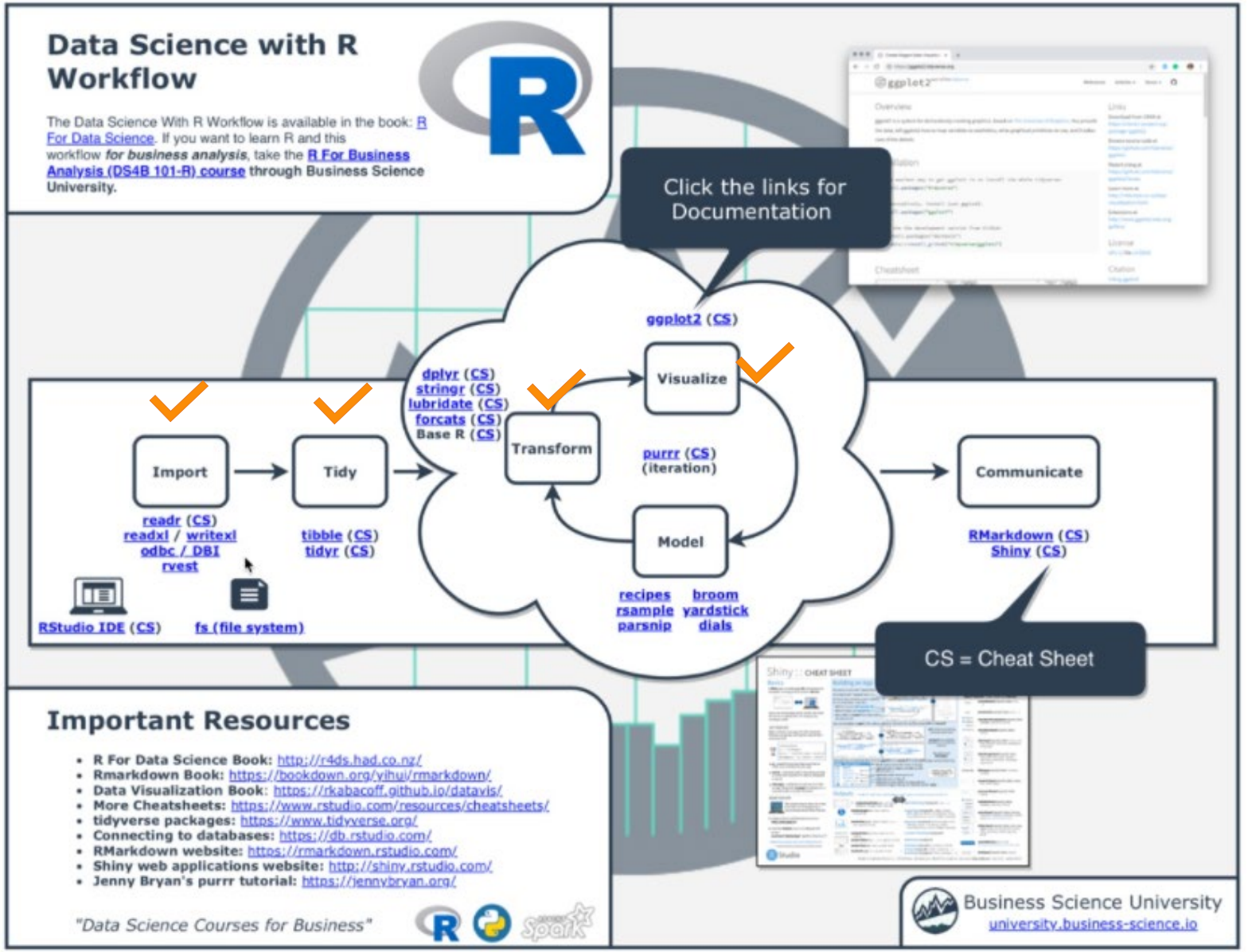
Be strategic about markers



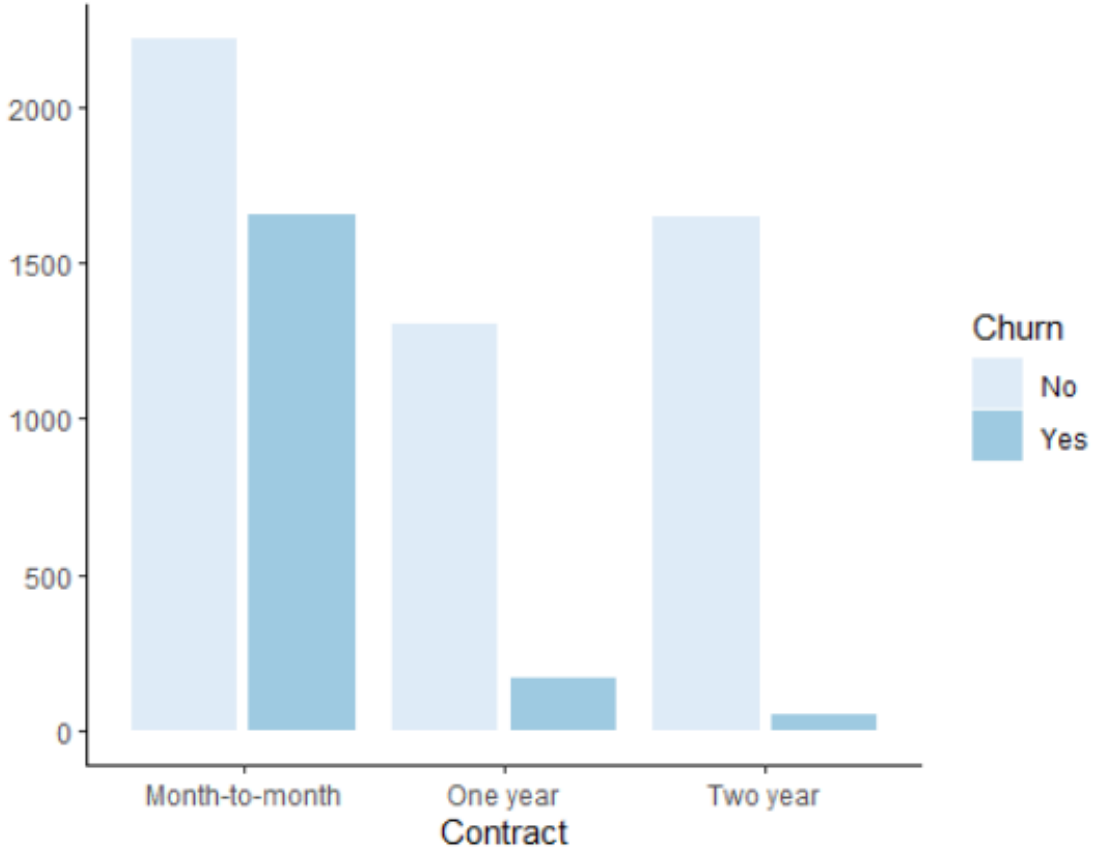
Knaflic, C. G. (2015). Storytelling with Data.

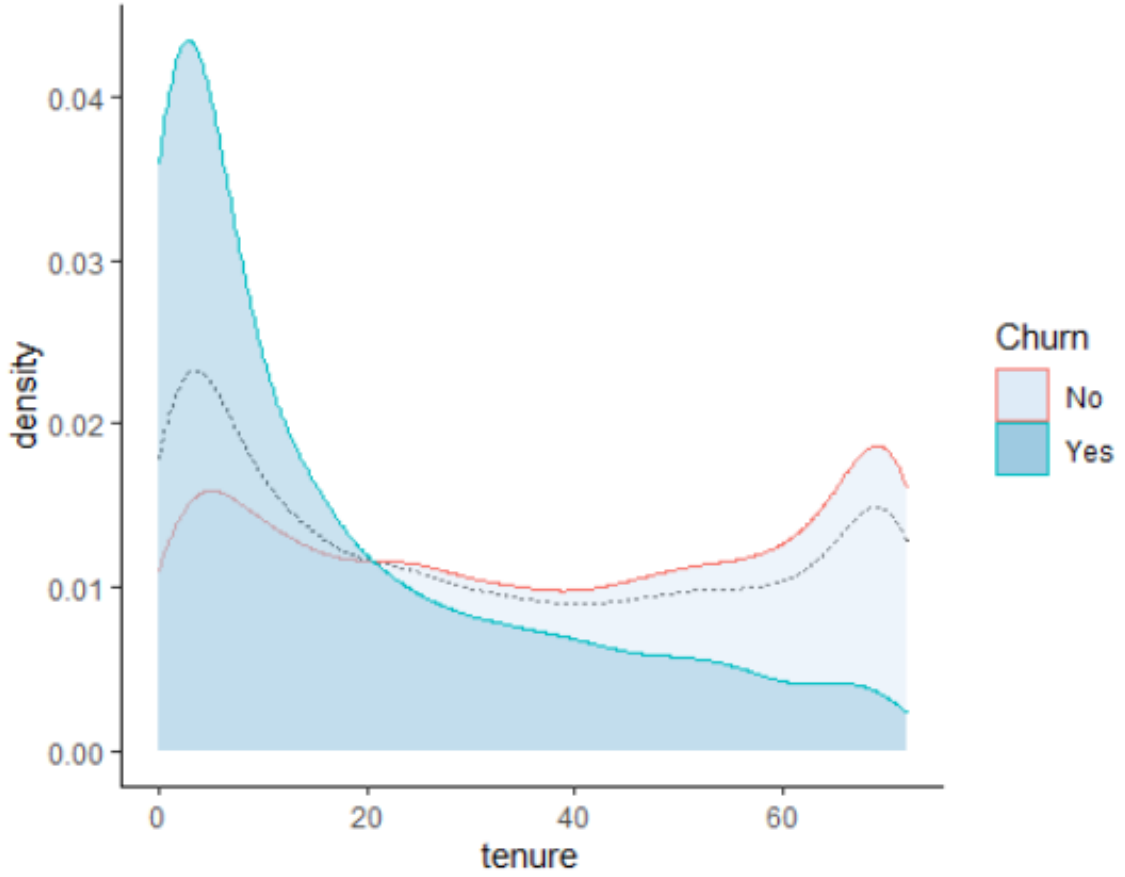
Data Value Creation Model




Data Science with R Workflow
The Data Science With R Workflow is available in the book: R For Data Science. If you want to learn R and this workflow for business analysis, take the R For Business Analysis (DS4B 101-R) course through Business Science University.
Click the links for Documentation
ggplot2 (CS)
dplyr (CS)
stringr (CS)
lubridate (CS)
forcats (CS)
Base R (CS)
Visualize
Transform
purrr (CS)
(iteration)
Model
Import
Tidy
Communicate
readr (CS)
readxl / writexl
odbc / DBI
rvest
tibble (CS)
tidyr (CS)
RMarkdown (CS)
Shiny (CS)
RStudio IDE (CS)
fs (file system)
recipes
rsample
parsnip
broom
yardstick
dials
CS = Cheat Sheet
Important Resources
R For Data Science Book: http://r4ds.had.co.nz/
Rmarkdown Book: https://bookdown.org/yihui/rmarkdown/
Data Visualization Book: https://rkabacoff.github.io/datavis/
More Cheatsheets: https://www.rstudio.com/resources/cheatsheets/
tidyverse packages: https://www.tidyverse.org/
Connecting to databases: https://db.rstudio.com/
RMarkdown website: https://rmarkdown.rstudio.com/
Shiny web applications website: http://shiny.rstudio.com/
Jenny Bryan's purrr tutorial: https://jennybryan.org/
"Data Science Courses for Business"
Business Science University
university.business-science.io

Exploratory analysis





Supervised vs Unsupervised Learning

Supervised

a. Based on well defined target, *and*
b. Training set *includes* labels for target

Unsupervised

- May apply when one can't satisfy (a) or (b) or both

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

CLASSIFICATION

«Divide the socks by color»

Predict a number

REGRESSION

«Divide the ties by length»

Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

Find hidden dependencies

Identify sequences

ASSOCIATION

«Find what clothes I often wear together»

Supervised Learning Model

A formula for estimating the unknown value of interest:

- **The target!**
  - The formula could be mathematical or a logical statement, such as a rule. Often, it is a hybrid

Things to Consider

- Is there a specific, quantifiable target that you are interested in predicting?
  - If yes, is it a **class or a number**?
    - Think about the decision

- Do you have data on the target?
  - Do you have **enough** data?
    - If the target is a class, a min of ~500 for each class type is needed

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

CLASSIFICATION

«Divide the socks by color»

Predict a number

REGRESSION

«Divide the ties by length»

Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

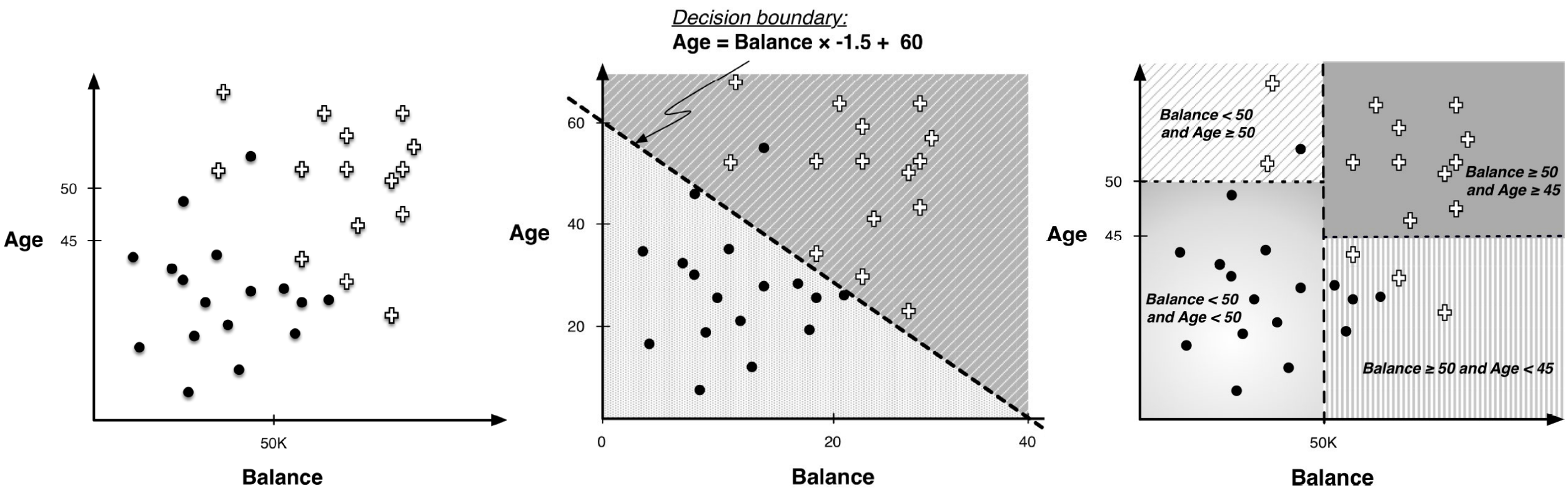
Find hidden dependencies

Identify sequences

ASSOCIATION

«Find what clothes I often wear together»

Classification Model



Regression Model

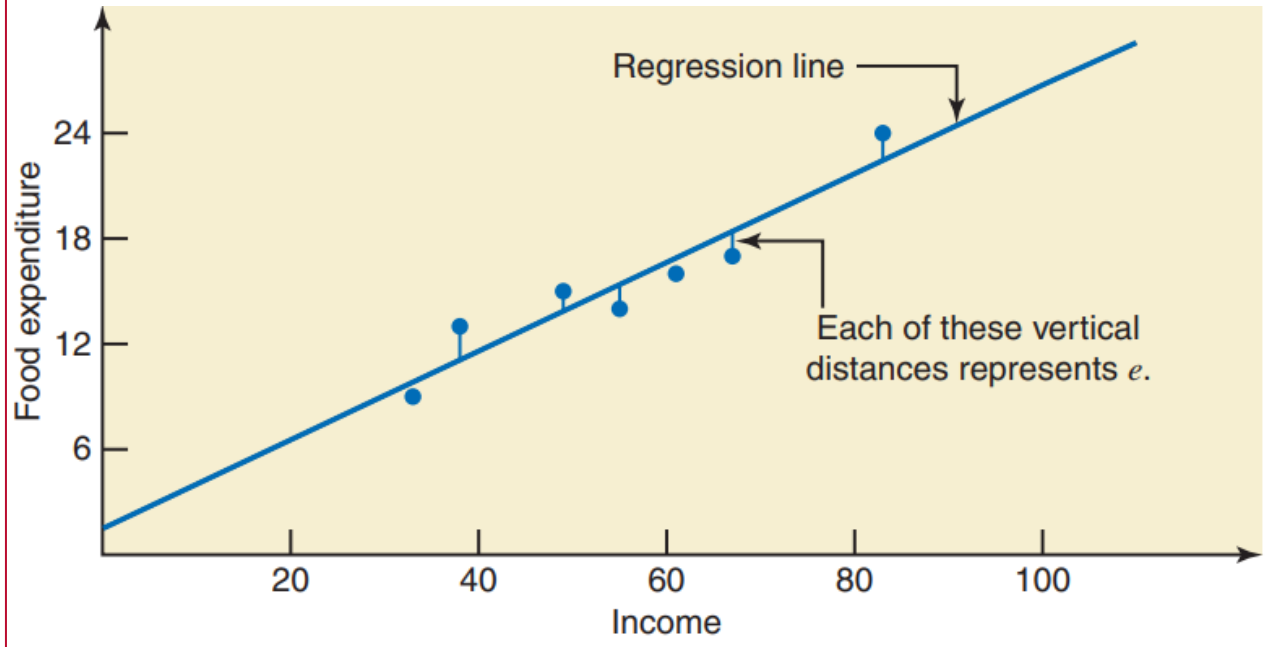
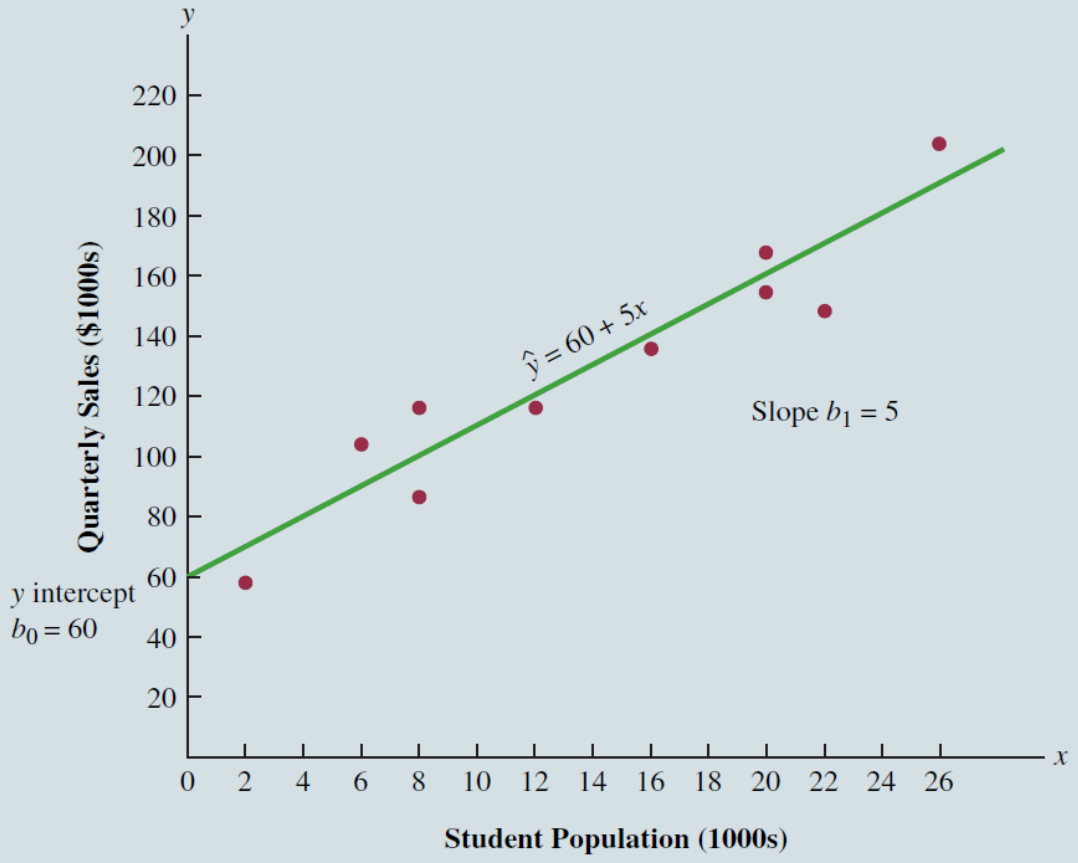


FIGURE 14.4 Graph of the Estimated Regression Equation for Armand's Pizza Parlors: $\hat{y} = 60 + 5x$



A Note on Inference/Prediction

The inference (explanation) is concerned with understanding the drivers of a business outcome

- Models of inference are interpretable but less accurate

The prediction itself is the main goal

- Not easily interpretable (“black-box”) but more accurate

Inference

Which of these affect the fraud probability the most?

	Transaction data A	Transaction data B	Transaction data C	Transaction data D
Transaction 1				
Transaction 2				
Transaction 3				
Transaction ...				
Transaction N				

Fraud probability

Prediction

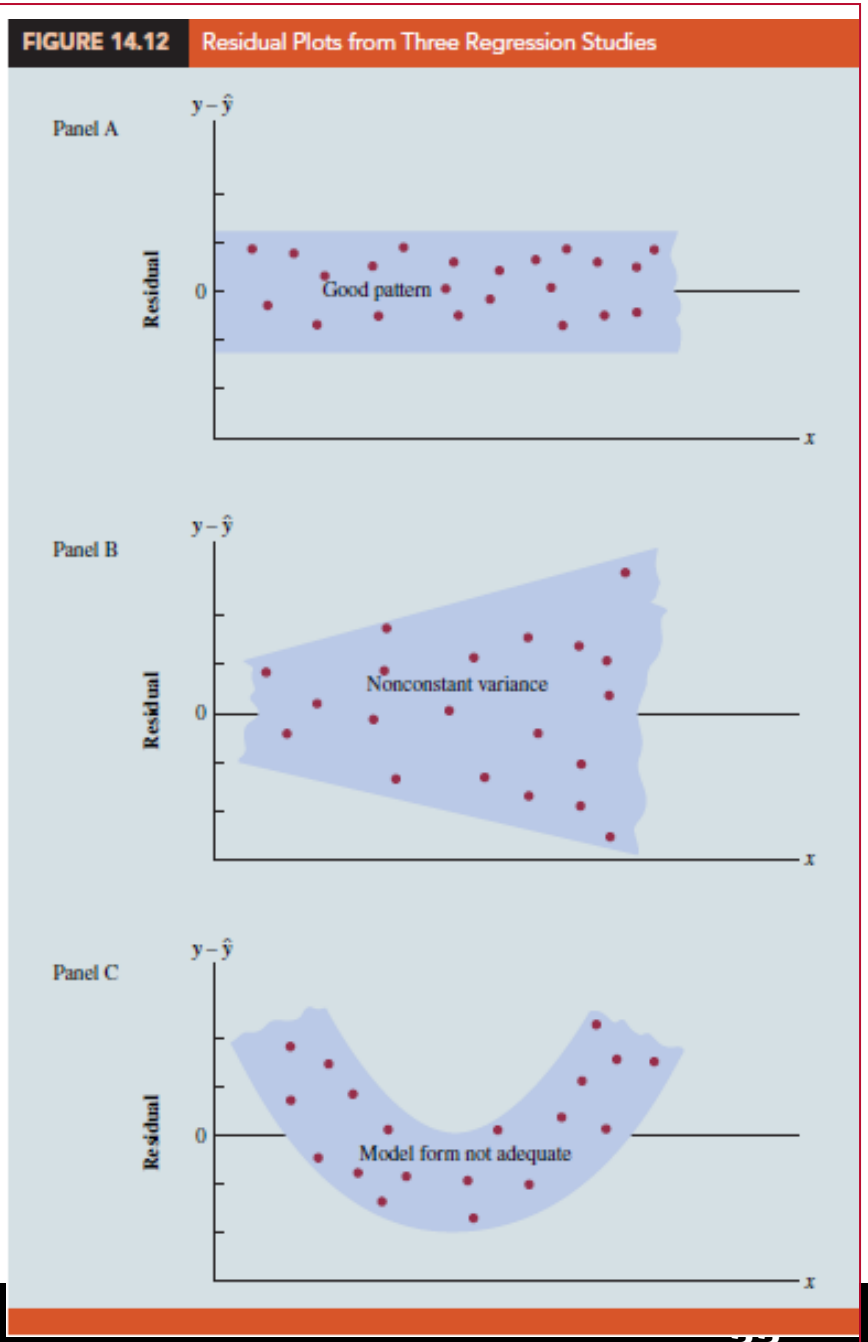
	Transaction data A	Transaction data B	Transaction data C	Transaction data D
Transaction 1				
Transaction 2				
Transaction 3				
Transaction ...				
Transaction N				

Get the most accurate probability this is fraud

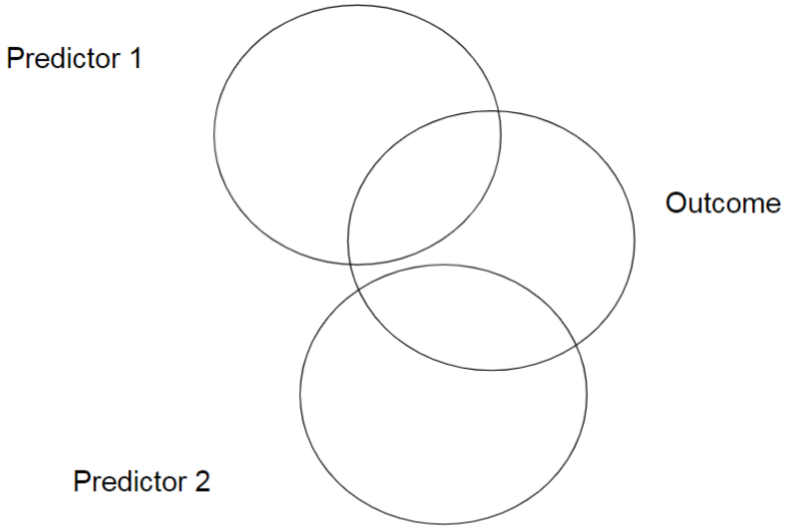
Fraud probability

Explanation - Reg. Assumptions

- Linear relationships
- No “lurking” or omitted variables
- Normally distributed errors (no pattern)
  - Independent
  - Similar variance across range of X
  - Eyeball test of plots

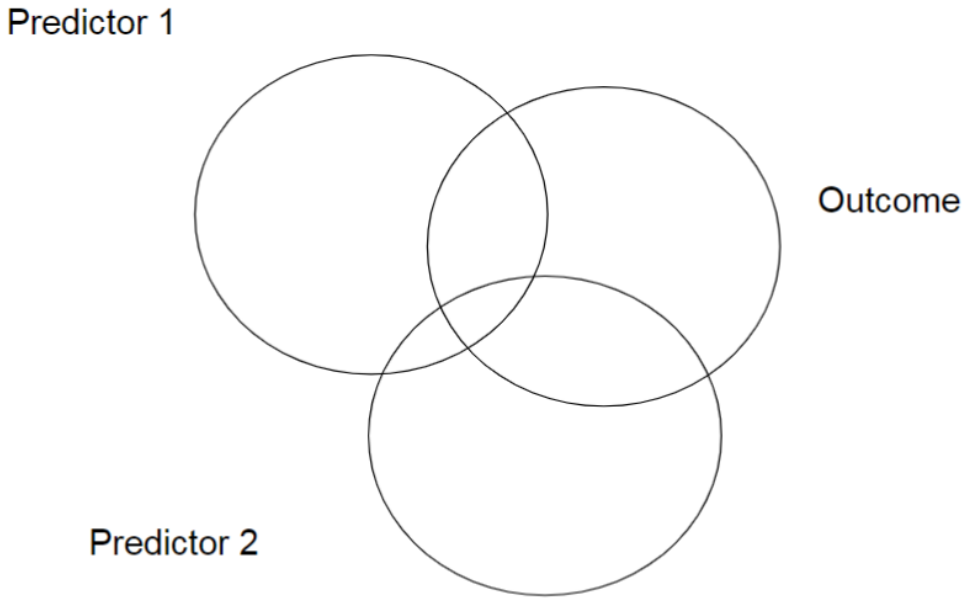


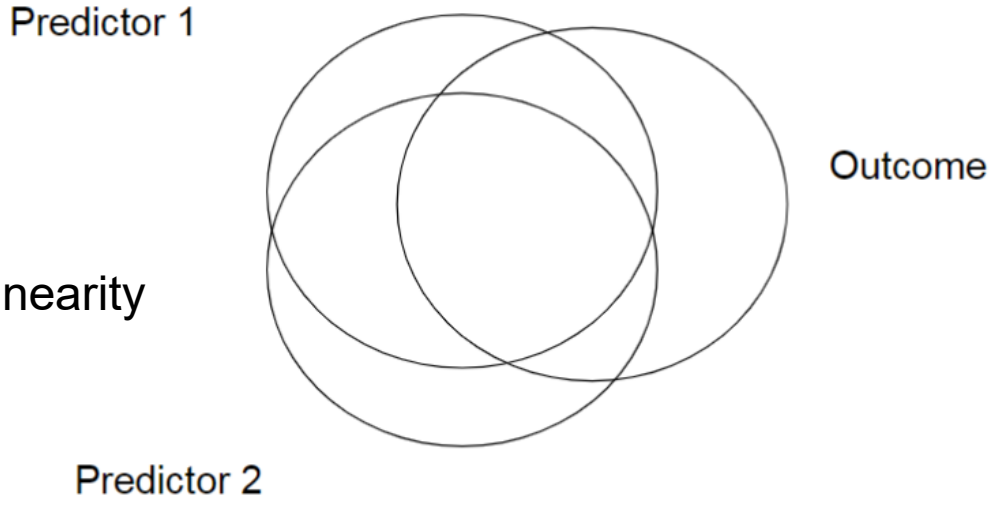
Explanation - Multicollinearity



Zero Multicollinearity

Minimal Multicollinearity





High Multicollinearity

Multicollinearity and Predictors Selection

```
# compute correlation between predictors and the target
cor(df1[,1:10])
```

```
> cor(df1[,1:10])
                    charges          age          sexN          bmi     children      smokerN region_northeast
charges         1.000000000  0.2990081933  0.057292062  0.198340969  0.06799823  0.787251430      0.006348771
age             0.299008193  1.0000000000 -0.020855872  0.109271882  0.04246900 -0.025018752      0.002474955
sexN            0.057292062 -0.0208558722  1.000000000  0.046371151  0.01716298  0.076184817     -0.002425432
bmi             0.198340969  0.1092718815  0.046371151  1.000000000  0.01275890  0.003750426     -0.138156224
children        0.067998227  0.0424689986  0.017162978  0.012758901  1.00000000  0.007673120     -0.022807598
smokerN         0.787251430 -0.0250187515  0.076184817  0.003750426  0.00767312  1.000000000      0.002811135
region_northeast  0.006348771  0.0024749545 -0.002425432 -0.138156224 -0.02280760  0.002811135      1.000000000
region_northwest -0.039904864 -0.0004074234 -0.011155728 -0.135995524  0.02480613 -0.036945474     -0.320177261
region_southeast  0.073981552 -0.0116419406  0.017116875  0.270024649 -0.02306575  0.068498410     -0.345561015
region_southwest -0.043210029  0.0100162342 -0.004184049 -0.006205183  0.02191358 -0.036945474     -0.320177261
                  region_northwest region_southeast region_southwest
charges             -0.0399048640       0.07398155     -0.043210029
age                 -0.0004074234      -0.01164194      0.010016234
sexN                -0.0111557280       0.01711688     -0.004184049
bmi                 -0.1359955237       0.27002465     -0.006205183
children             0.0248061293      -0.02306575      0.021913576
smokerN             -0.0369454740       0.06849841     -0.036945474
region_northeast    -0.3201772613      -0.34556102     -0.320177261
region_northwest     1.0000000000      -0.34626466     -0.320829220
region_southeast    -0.3462646614       1.00000000     -0.346264661
region_southwest    -0.3208292201      -0.34626466      1.000000000
```

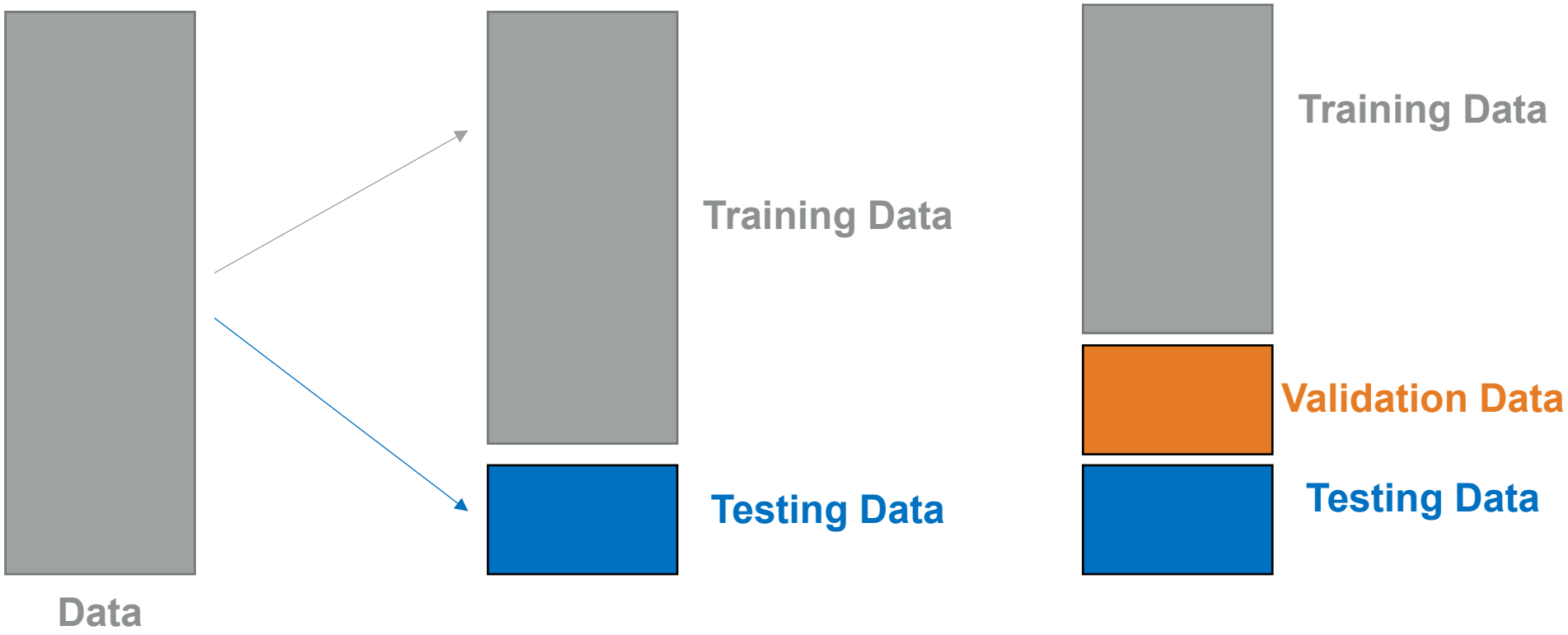
Model Performance

```
# check the results
summary(model)
```

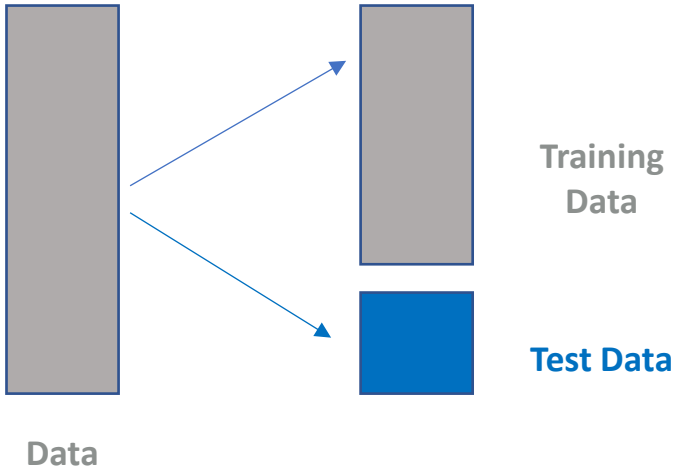
```
Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexN              -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokerN          23848.5      413.1  57.723  < 2e-16 ***
region_northwest  -353.0      476.3  -0.741 0.458769
region_southeast -1035.0      478.7  -2.162 0.030782 *
region_southwest  -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

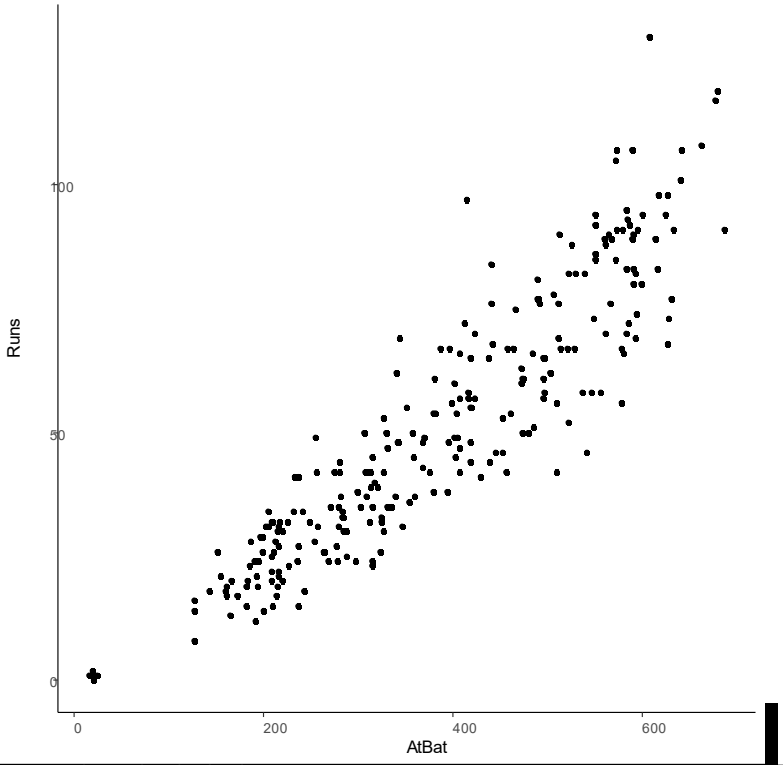
Predictive Model: Data Splitting

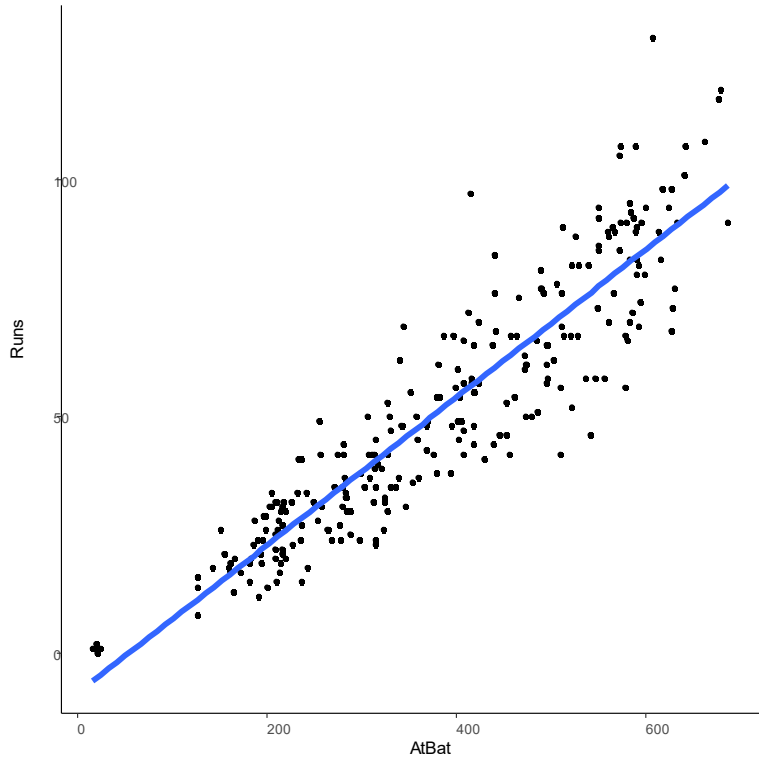
Use training set to build model, then predict using the test set

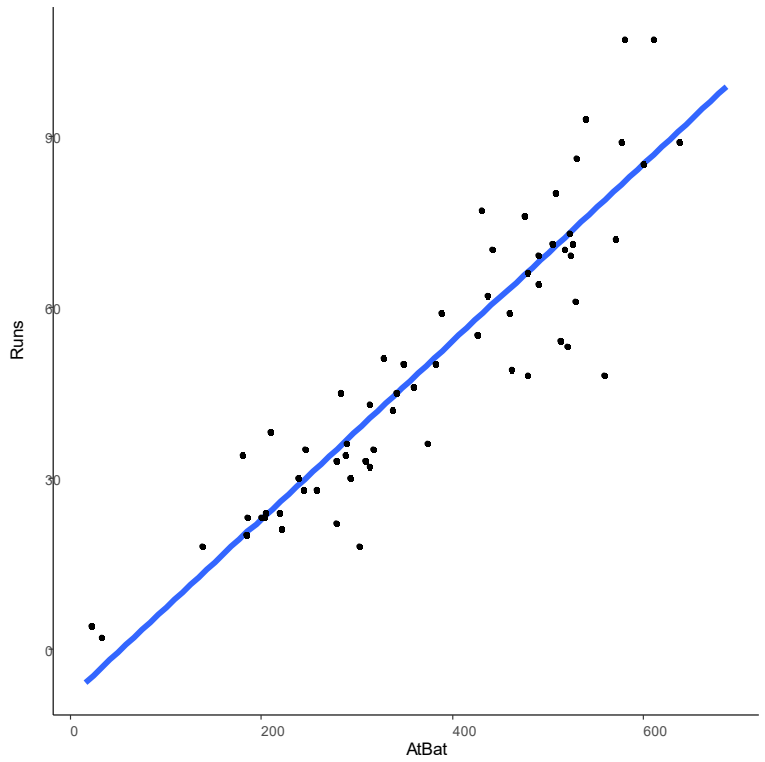


Train and Test the Model









Model Evaluation – Regression Analysis

Error for data record = predicted (p) minus actual (a)

RMSE: Root Mean Squared Error: $\sqrt{\frac{1}{n}\sum_1^n\left(Y_i-\hat{Y}_i\right)^2}$

MAE: Mean Absolute Error: $\frac{1}{n}\sum_1^n\left|\left(Y_i-\hat{Y}_i\right)\right|$

MAPE: Mean Absolute Percentage Error: $\frac{100}{n}\sum_1^n\left|\frac{Y_i-\hat{Y}_i}{Y_i}\right|$

Total SSE: Total Sum of Squared Errors: $\sum_1^n\left(Y_i-\hat{Y}_i\right)^2$

Model Performance

Use training set to build model, then predict insurance cost using the test set

```
# how did we do? calculate performance across resamples
# RMSE and R-squared
postResample(pred = p, obs = charges_test$charges)
# on average, our prediction is off by $5,790.49
```

```
> postResample(pred = p, obs = charges_test$charges)
         RMSE      Rsquared           MAE
5790.4940335    0.7999239 4169.6005174
```

Model Evaluation - Classification Performance

The confusion matrix

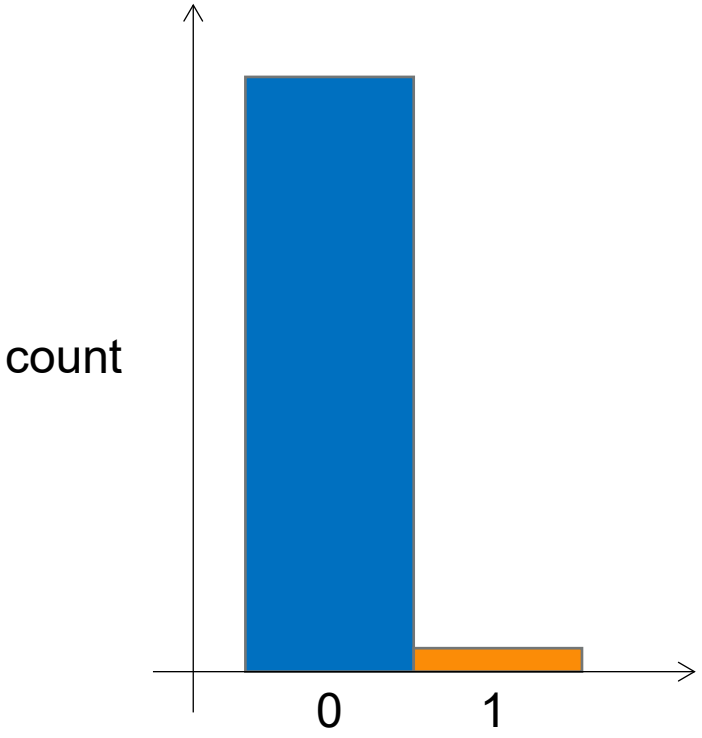
- Separates out the decisions made by the model, making explicit how one class is being confused for another

		Actual	
		Positive	**Negative**
Predicted	**Positive**	True Positives	False Positives
	Negative	False Negatives	True Negatives

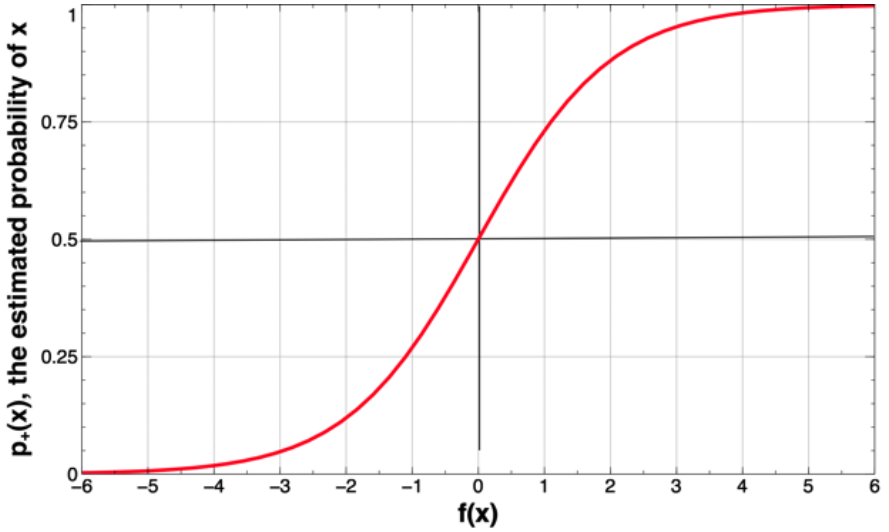
Accuracy

Inappropriate for unbalanced (or skewed) classes

0 = no fraud
1 = yes fraud



Resampling
Threshold analysis



Predictive Model: Classification Performance

Additional metrics

		Actual	
		Positive	**Negative**
Predicted	**Positive**	**True Positives**	**False Positives**
	Negative	**False Negatives**	**True Negatives**

- Precision = true positives / (true positives + false positives)
- Recall = true positives / (true positives + false negatives)
- F1-measure = (2 * precision * recall) / (precision + recall)

Model Performance

Use training set to build model, then predict churn using the test set

```
# how did we do? confusion matrix
confusionMatrix(data = churn_test$pred_churn,
                reference = churn_test$Churn,
                mode = "prec_recall",
                positive = "Yes")
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  926 171
       Yes 106 202

               Accuracy : 0.8028
                 95% CI : (0.7811, 0.8234)
    No Information Rate : 0.7345
    P-Value [Acc > NIR] : 1.364e-09

                  Kappa : 0.4647

 Mcnemar's Test P-Value : 0.0001204

              Precision : 0.6558
                 Recall : 0.5416
                     F1 : 0.5932
             Prevalence : 0.2655
         Detection Rate : 0.1438
   Detection Prevalence : 0.2192
      Balanced Accuracy : 0.7194

       'Positive' Class : Yes
```

- Of all customers where we predicted churn, ~66% actually churned
- Of all customers that actually churned, we only correctly predicted about half (~54%)

Leakage

- Do you have relevant data **prior** to the decision?
  - Think about the timing of decision and action leading up to it

Provost, F. & Fawcett, T. (2013). Data Science for Business.

Avoiding Leakage

- Do you have relevant data **prior** to the decision?
  - Think about the timing of decision and action leading up to it