Teaching the Tidyverse in the Second Semester, Undergraduate Statistics Course

Kelly McConville, Swarthmore College, U.S.A.

Background

Goal: Modernize my second semester, undergraduate statistics course. Want course to satisfy two popular but conflicting ideas:

- Teach the entire data analysis workflow, of which modeling is only one step.
- Teach a more diverse set of models, especially statistical learning techniques.

Problem: How do I find time to teach more of the data analysis workflow and to cover new modeling techniques?

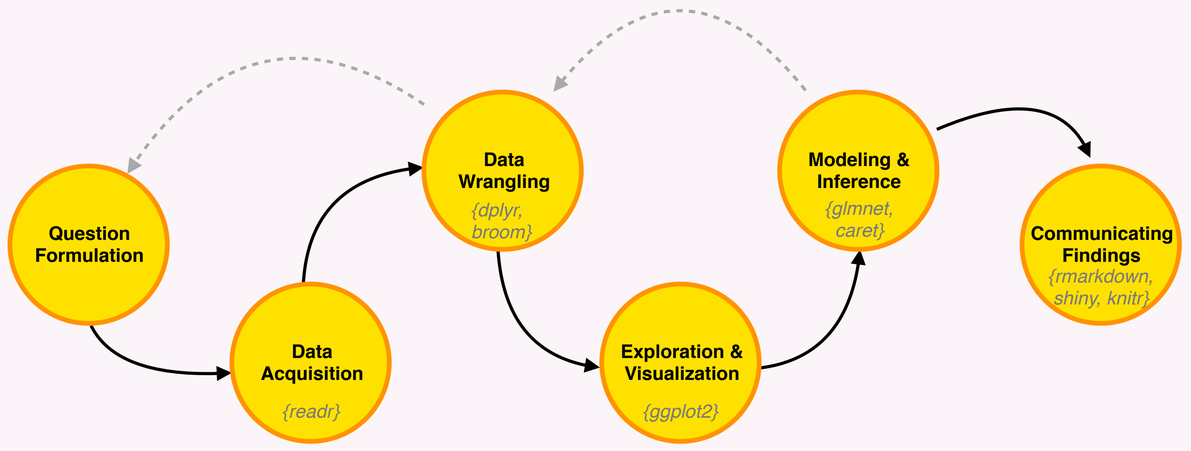
Proposed Solution:

- Streamline the process of teaching the data analysis workflow using the Tidyverse.
- Shorten the discussion of specific regression models.
- Use freed up class time to cover predictive modeling techniques.

Examples: In this poster, I present example activities which:

- Use Tidyverse packages.
- Emphasize the importance of the Data Wrangling and the Exploration and Visualization steps.
- Reflect an iterative approach to the data analysis workflow.
- Include statistical learning methods.
- Follow a reproducible workflow.

Data Analysis Workflow



Case Study 1: Are volcanic eruptions increasing?

Question Formulation:

- After learning simple linear regression, the students can frame this problem as:
  - *Is there a positive, linear relationship between time and number of eruptions?*

Data Acquisition:

- Data file from the Smithsonian Institution's Global Volcanism Program website.

```
read_csv()
```

```
Eruptions
```

```
## # A tibble: 11,078 × 24
##    Volcano_Number           Volcano_Name Eruption_Number
##             <int>                  <chr>           <int>
## 1          282080                   Aira           22203
## 2          300010              Kambalny           22198
## 3          262000               Krakatau           22188
```

Data Wrangling:

- Filter by date and confirmed eruptions.
- Group by start year.
- Record year, number of eruptions, and average size of eruptions.

```
dataset %>%
  filter() %>%
  group_by() %>%
  summarize()
```

```
Eruptions_yearly
```

```
## # A tibble: 118 × 3
##    Start_Year count  avg_VEI
##         <int> <int>    <dbl>
## 1        1900    20 1.500000
## 2        1901    16 2.066667
## 3        1902    34 1.941176
```

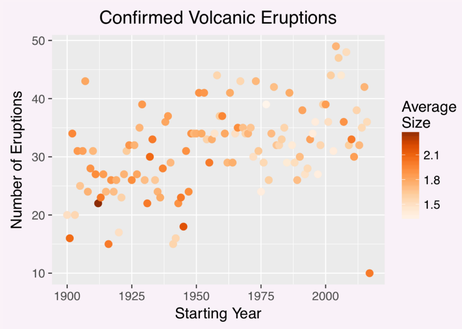
Exploration and Visualization:

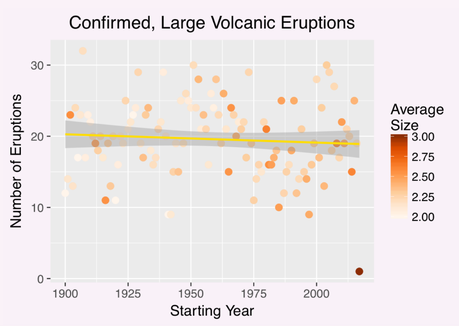
- Create scatterplots.

```
ggplot() +
  geom_point() +
  stat_smooth()
```

- Sampling bias issues:
  - World events impacting reporting.
  - Detection dependent on size of the eruption over time.
- Add one more wrangling argument to try to minimize bias.

```
filter()
```





Modeling and Inference:

- Construct model and summary table.

```
lm() %>%
  tidy() %>%
  kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	42.42	28.37	1.5	0.14
Start_Year	-0.01	0.01	-0.8	0.42

- Not a significant relationship.

Communicating Findings:

- Students write up their work using RMarkdown.
- Students also use this data to construct interactive maps of the world's volcanoes using *shiny* and *leaflet*.

Case Study 2: Build a model for household income.

Question Formulation:

- When covering model selection techniques, the students complete the following task:
  - *Build a model for income. Conduct model selection to determine an appropriate set of predictors.*

Data Acquisition:

- Data from the US Bureau of Labor Statistics Consumer Expenditure Survey.
- Two files from the fourth quarter of 2015:
  - Household data
  - Data on each individual

Data Wrangling:

- Merge the principal earner's information into the household dataset.

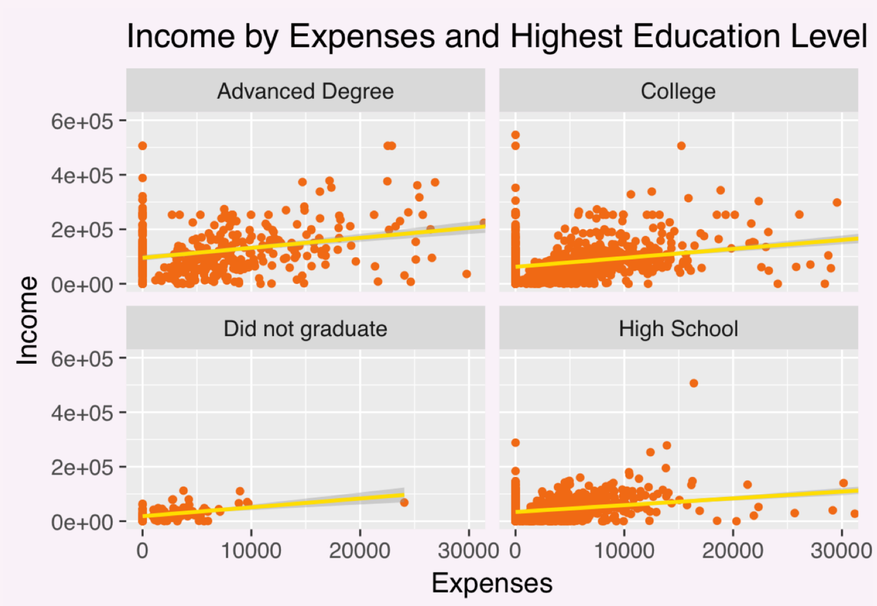
```
left_join()
```

- Resulting in 2,469 households.

Exploration and Visualization

- Students construct graphics to explore multivariate relationships.

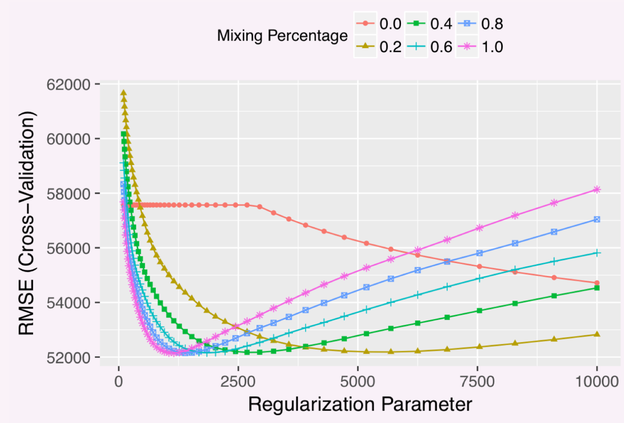
```
ggplot() +
  geom_point() +
  stat_smooth() +
  facet_wrap()
```



Modeling and Inference:

- Consider full two-way interaction model with 1,030 potential variables.
- Fit an elastic net model.
- Use cross-validation to select hyperparameters.

```
trainControl(method = "cv")
train(..., method = "glmnet")
```



- Resulting model contains 163 variables.

Communicating Findings:

- In an RMarkdown report, students compare the performance of the selected models between stepwise selection and elastic net and draw conclusions about how the predictors relate with income.

Conclusions

- Students get a lot of satisfaction out of making impressive plots with *ggplot2* and polished reports with *RMarkdown*.
  - This provides motivation to improve their skills and to overcome errors.
- Students struggle with data wrangling. My suggestions are:
  - Make LOTS of pictures.
  - Use the pipes to breakdown each step.
  - Stress the importance of the wrangling step to the entire workflow.
- Must drop some topics.
- With freely available or "found" data, it is so important to emphasize the potential pitfalls of generalizing results.

Acknowledgments

I would like to thank the Smithsonian Institution and the US Bureau of Labor Statistics for providing public use datasets. My classes have also greatly benefitted from the RStudio Server.

References

- Allaire, J. J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., Hyndman, R., and R. Arslan (2017). rmarkdown: Dynamic Documents for R. R package version 1.5. http://CRAN.R-project.org/package=rmarkdown
- American Statistical Association Undergraduate Guidelines Workgroup. 2014. 2014 curriculum guidelines for undergraduate programs in statistical science. Alexandria, VA: American Statistical Association. http://www.amstat.org/education/curriculumguidelines.cfm
- Friedman, J. Hastie, T. and R. Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.
- Global Volcanism Program, 2013. Volcanoes of the World, v. 4.4.3. Venzke, E (ed.). Smithsonian Institution. Downloaded 06 May 2016. http://dx.doi.org/10.5479/si.GVP.VOTW4-2013
- Kuhn, M. (2017). caret: Classification and Regression Training. R package version 6.0-76. https://CRAN.R-project.org/package=caret
- Robinson, D. (2017). broom: Convert Statistical Analysis Objects into Tidy Data Frames. R package version 0.4.2. https://CRAN.R-project.org/package=broom
- United States Bureau of Labor Statistics, 2015. Consumer Expenditure Survey. Downloaded 01 January 2017. https://www.bls.gov/cex/pumd.htm
- Wickham, H. 2016. Tidyverse. http://tidyverse.org/.
- Wickham, H. ggplot2: elegant graphics for data analysis. Springer New York, 2009.
- Wickham, H. and R. Francois (2015). dplyr: A Grammar of Data Manipulation. R package version 0.4.3. http://CRAN.R-project.org/package=dplyr
- Wickham, H., and G. Grolemund. 2016. R for Data Science. http://r4ds.had.co.nz/; O'Reilly Media.
- Wickham, H., Hester, J. and R. Francois (2017). readr: Read Rectangular Text Data. R package version 1.1.0. https://CRAN.R-project.org/package=readr
- Xie, Y. (2016). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.15.1.

kmcconv1@swarthmore.edu