Module 2: A Comparison of R, SAS, and Python

Brett Amione

Colorado State University – Global Campus

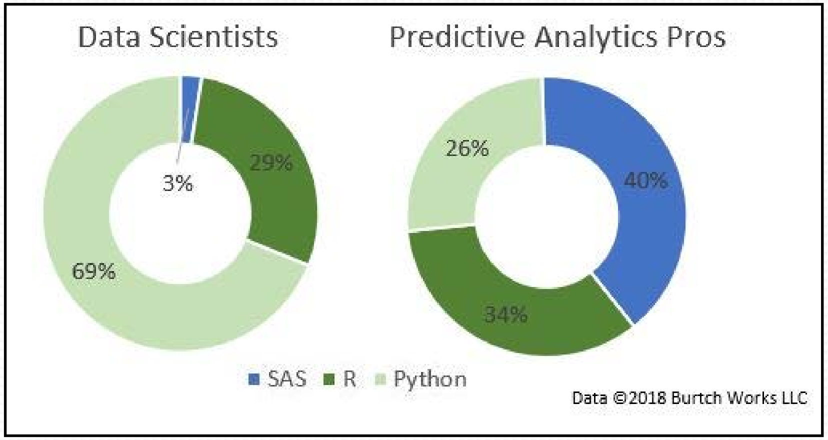


Figure 1: Comparing SAS, R, or Python Preferences: Data Scientists vs. Traditional Predictive Analytics Professionals (Burtch Works, 2018)

Author Note:

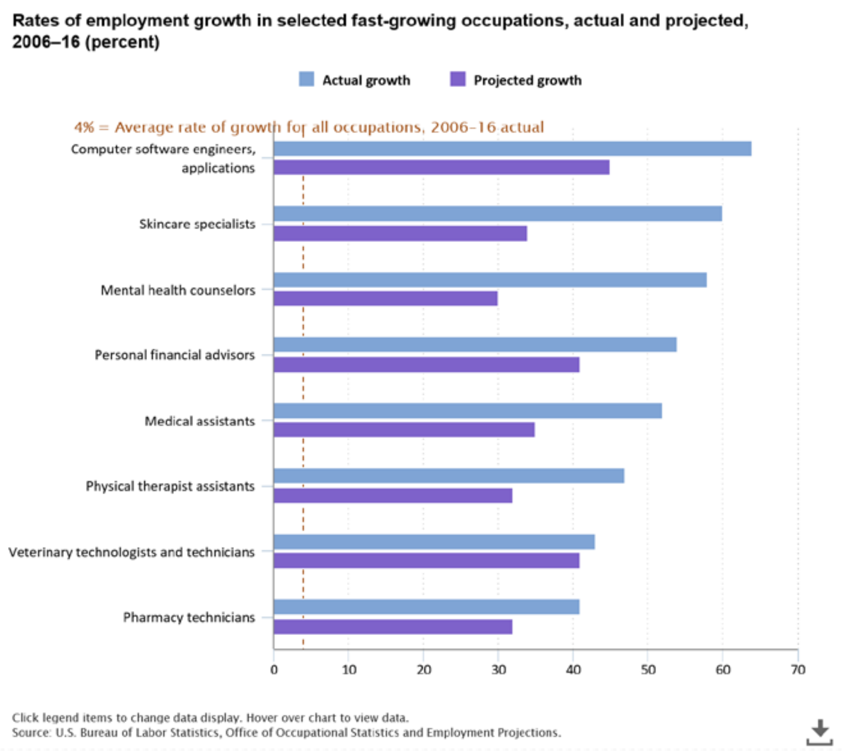
Presentation prepared for MIS500-1

Comparing R, SAS, and Python

- Discussion Points:
  - A note on the profession of data science and analytics
  - Software History
  - Job Perspective/Outlook
  - Availability/Cost
  - Ease of usage/learning
  - Software Capabilities
    - Data handling, visualization, statistical packages, deep learning support, and Customer service support and Community
  - Data Task Scenarios
  - Future Outlook of Tool

The Data Scientist/Analyst

- When comparing the job perspective of R, SAS, and Python users, it is important to recognize that the professionals utilizing these software packages are in a rapidly growing field
- $91,530: Mean annual wage of those in the Computer and Mathematical Occupations field (BLS, 2018)



Rates of employment growth in selected fast-growing occupations, actual and projected, 2006–16 (percent)

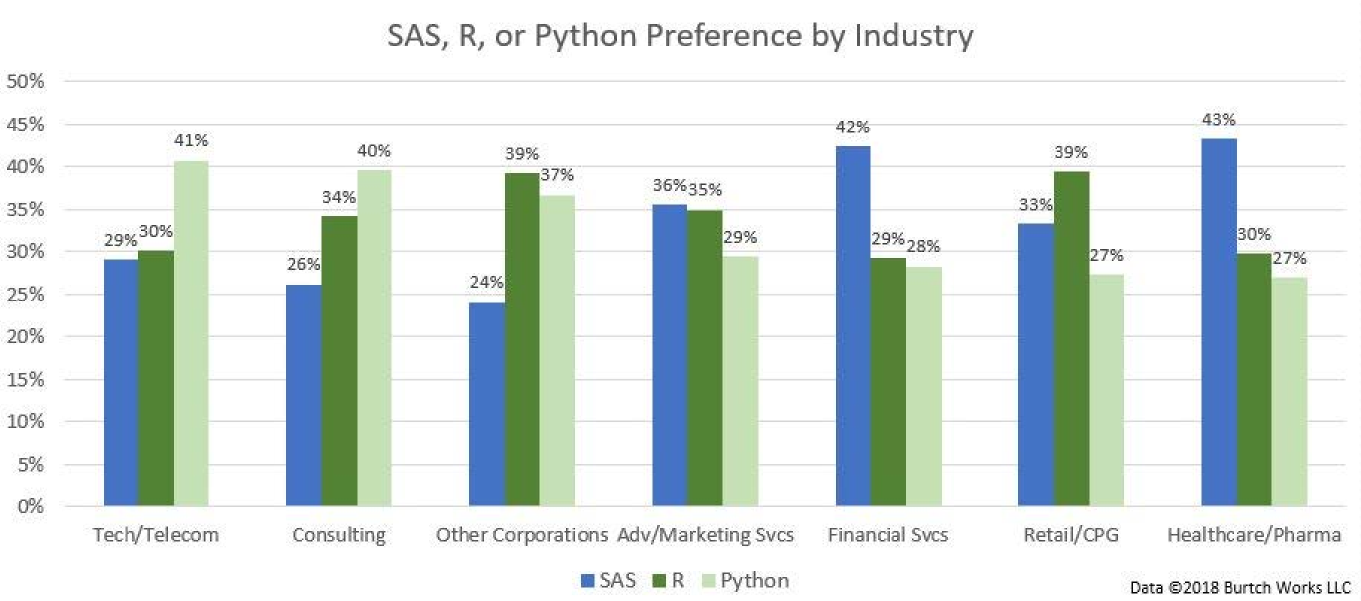
Click legend items to change data display. Hover over chart to view data.
Source: U.S. Bureau of Labor Statistics, Office of Occupational Statistics and Employment Projections.

Figure 2: Projected v. Actual Employment Growth by Occupation (Bureau of Labor Statistics, 2019)

Software Histories

- R:
  - Created in 1995
  - Influenced by the S and Scheme languages (Hornik, 2018)
  - Designed to deliver user-friendly data analysis, statistics, and graphical models (Data-Driven Science, 2018)
  - Originally used primarily in academic institutions; is now distributed under a GNU “copyleft” allowing for wide scale usage
  - Has an active online community which has developed thousands of packages
- SAS (SAS, 2019):
  - Created under a NIH Grant in 1966
  - Re-written in the 1980s to allow for use on PCs and to incorporate user-friendly GUIs
  - Expanded to include software packages across many fields of data analytics
  - Widely used in corporate institutions
- Python (Python, 2019)
  - Created in the 1980s and intended to be a successor the ABC language
  - Currently running on Python 3
  - Has an active online community which has developed thousands of packages

Job Outlook



Data ©2018 Burtch Works LLC

Figure 3: SAS, R, or Python Preference by Industry (Burtch, 2018)

- As stated previously, professions which make use of statistical software packages are in high demand
- While Python and R are very popular packages, SAS is widely used in corporate environments which may offer less flexibility in the particular software an employee uses

Availability/Cost

- R:
  - Distributed under a GNU copyleft effectively making it free to use and distribute the software
  - Free to download
- SAS:
  - Distributed via corporate licenses which cost hundreds of thousands of dollars
  - Free trials and academic versions
- Python:
  - Distributed as open source software
  - Free to download
- Note:
  - While R & Python are open source, there are still costs incurred for training users and there are likely costs involved in having staff to assist in providing systems support to users

Ease of Learning/Usage

- R:
  - Considered to be one more difficult languages to use due to its limited GUI and documentation
  - Users benefit from having a basic knowledge of programming, mathematics, statistics, and visualization (EDUCBA, 2019)
- SAS:
  - Considered to be a user-friendly software that allows users to easily import various data sources
  - Their Augmented Analytics software assists users in understanding which visualization technique is best to use and provides suggested insights
- Python:
  - Considered to be one of the easier programming languages to learn
  - Like R, users will benefit from already being skilled in related fields as the software does not guide the user in the same way that SAS can

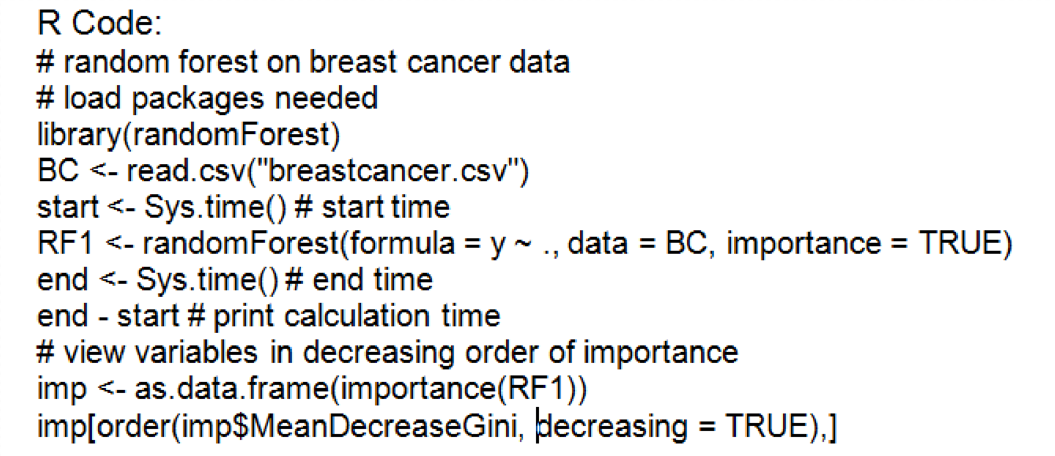
Software Capabilities

	R	SAS	Python
Data Handling	On a 64bit system, R can run a nearly infinite dataset Not the fastest at running large data sets (Pandey, 2019).	Can handle 9.2 quintillion observations in 1 data set Maximum size of 1 dataset is 2086GB (Stack Overflow, 2013).	Python's data handling is improved using packages such as NumPy. NumPy has been able to up to 1.4 billion rows of data (Stagg, 2018).
Data Visualization	4 basic presentation types (Comparison, Composition, Distribution, Relationship), 7 common charts (Scatter, Histogram, Bar & Stack, Box, Area, Heat Map, and Correlogram). Many packages exist, with ggplot2 being the most popular. (R-Studio, 2019).	Typical charts such as bar, pie, and scatter. Analytical visualizations such as decision trees and forecasting. Augmented Analytics which allow for autocharting, automated explanations, and suggested insights (SAS, 2019).	Many popular thirdparty packages for visualizaiton which includes: Matplotlib, Seaborn, ggplot (ported from R), and Bokeh (Tanner, 2019).
Statistical Capabilities	25 packages supplied with R and many more offered through CRAN (R-Studio, 2019)	4 featured statistical analysis software packages and many other related packages (SAS, 2019).	Python comes with a basic statistical package but external packages such as NumPy, SciPy, and Pandas are more popular for performing statistical analyses (Bobriakov, 2018)

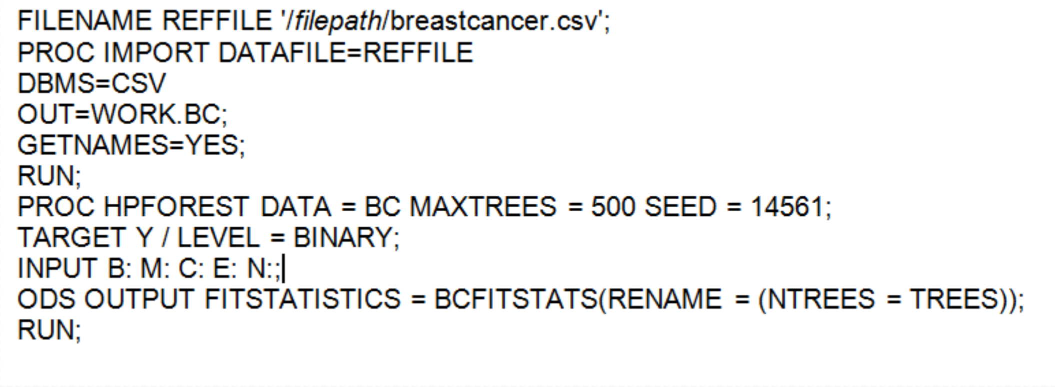
Software Capabilities (Cont’d)

	R	SAS	Python
Deep Learning	R has recently became more competitive. TensorFlow and R are now compatible and there are numerous other packages such as nnet and h2o (Walia, 2017 & Willems, 2019).	Four featured Deep Learning packages including SAS Visual Data and Machine Learning, SAS Optimization, SAS Visual Forecasting, and SAS Visual Text Analytics (SAS, 2019).	Many Deep Learning packages such as Scitkit-Learn, XGBoost, and TensorFlow (Bobriakov, 2018).
Support & Customer Service	Built-in help features and package specific documentation. Many third-party support networks such as Stack Overflow and CRAN Task Views (R-Studio, 2019)	As a private software, SAS is able to dedicate a considerable amount of attention towards support and customer service. SAS provides technical documentation, support for specific software packages, and systems support. SAS also provides training and certification courses (SAS, 2019).	Python FAQs, Python Tutor list, and Python Newsgroup are three official Python support groups. Python also supports a bug tracker and security reporting system that allows for more collaborative support. Stack Overflow, Quora, and online learning are other popular places to go for support (Python, 2019).

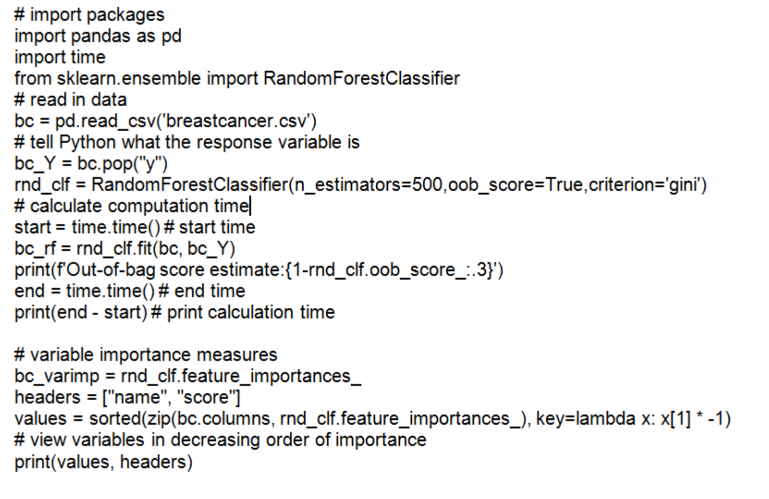
Data Task Scenarios



```
R Code:
# random forest on breast cancer data
# load packages needed
library(randomForest)
BC <- read.csv("breastcancer.csv")
start <- Sys.time() # start time
RF1 <- randomForest(formula = y ~ ., data = BC, importance = TRUE)
end <- Sys.time() # end time
end - start # print calculation time
# view variables in decreasing order of importance
imp <- as.data.frame(importance(RF1))
imp[order(imp$MeanDecreaseGini, decreasing = TRUE),]
```



```
FILENAME REFFILE '/filepath/breastcancer.csv';
PROC IMPORT DATAFILE=REFFILE
DBMS=CSV
OUT=WORK.BC;
GETNAMES=YES;
RUN;
PROC HPFOREST DATA = BC MAXTREES = 500 SEED = 14561;
TARGET Y / LEVEL = BINARY;
INPUT B: M: C: E: N:;
ODS OUTPUT FITSTATISTICS = BCFITSTATS(RENAME = (NTREES = TREES));
RUN;
```



```
# import packages
import pandas as pd
import time
from sklearn.ensemble import RandomForestClassifier
# read in data
bc = pd.read_csv('breastcancer.csv')
# tell Python what the response variable is
bc_Y = bc.pop("y")
rnd_clf = RandomForestClassifier(n_estimators=500,oob_score=True,criterion='gini')
# calculate computation time
start = time.time() # start time
bc_rf = rnd_clf.fit(bc, bc_Y)
print(f'Out-of-bag score estimate:{1-rnd_clf.oob_score_:.3}')
end = time.time() # end time
print(end - start) # print calculation time

# variable importance measures
bc_varimp = rnd_clf.feature_importances_
headers = ["name", "score"]
values = sorted(zip(bc.columns, rnd_clf.feature_importances_), key=lambda x: x[1] * -1)
# view variables in decreasing order of importance
print(values, headers)
```

Figure 4: Comparison on R (top-left), SAS (bottom-left), and Python (top-right) to implement a Random Forest Method which utilizes Machine Learning principles (Soifua, 2018).

Future Outlook

- As stated earlier in the presentation, computer and mathematic fields are all growing.
- Burtch Works' research (2018) which suggests that many large corporations utilize SAS and therefore it would seem that those who are interested in working within a larger organization may want to learn SAS.
- R and Python are two popularly used languages. Many organizations use them and they are extremely popular amongst start ups too.
- None of these products are experiencing a significant decline in use and all will remain relevant.

Summary

- R, SAS, and Python are all important programming languages that allow data scientists to contribute to the fields of Data Science and Data Analytics
- R and Python are open source packages which make use of third party packages to perform rich statistical analyses, machine learning techniques, and visualization.
- SAS is an enterprise software package suite with official modules that allow license administrators to tailor each user’s SAS module to their needs.

References

- Bierly, M. (2016, June 8). *10 Useful Python Data Visualization Libraries for Any Discipline*. Retrieved from: https://mode.com/blog/python-data-visualization-libraries
- Bobriakov, I. (2018, June 11). *Top 20 Python libraries for data science in 2018*. Retrieved from: https://medium.com/activewizards-machine-learning-company/top-20-python-libraries-for-data-science-in-2018-2ae7d1db8049
- Burtch Works (2018, July 16). *2018 SAS, R, or Python Survey Results: Which do Data Analysts & Analyst Pros Prefer?* Retrieved from: (https://www.burtchworks.com/2018/07/16/2018-sas-r-or-python-survey-results-which-do-data-scientists-analytics-pros-prefer/)
- Bureau of Labor Statistics (2018, May 1). *Occupational Employment and Wages, May 2018.* Retrieved from: https://www.bls.gov/oes/current/oes150000.htm
- Data-Driven Science (2018, January 30). *Python vs R for Data Science: And the winner is…* Retrieved from: *https://medium.com/@data_driven/python-vs-r-for-data-science-and-the-winner-is-3ebb1a968197*
- Educba (2019). *Careers in R Programming*. Retrieved from: https://www.educba.com/careers-in-r-programming/
- Hornik, K. (2018, October 18). *Frequently Asked Questions on R*. Retrieved from https://cran.r-project.org/doc/FAQ/R-FAQ.html#What-is-R_003f
- Jariwala, D. (2016, December 29). 7 Visualizations You Should Learn in R [Blog Post]. Retrieved from: https://www.r-bloggers.com/7-visualizations-you-should-learn-in-r/
- Pandey, P. (2019, March 7). *From 'R vs Python' to 'R and Python'*. Retrieved from: https://towardsdatascience.com/from-r-vs-python-to-r-and-
- Python (2019). *History and License*. Retrieved from: https://docs.python.org/3/license.html
- python-aa25db33ce17.
- Python Software Foundation (2019, August 25). *The Python Standard Library*. Retrieved from: https://docs.python.org/3/library/statistics.html
- R-Studio (2019, July 5). *An Introduction to R*. Retrieved from: https://cran.r-project.org/doc/manuals/r-release/R-intro.html#R-and-statistics
- SAS. (2019). *Company Information*. Retrieved from: https://www.sas.com/en_us/company-information/profile.html
- SAS. (2019). *SAS Visual Analytics (SAS Viya) Features List*. Retrieved from: https://www.sas.com/en_us/software/visual-analytics/viya-features.html
- SAS. (2019). *SAS Products, Technology, & Solutions.* Retrieved from https://www.sas.com/en_us/software/all-products.html#e148b0e5-288b-417d-a1f5-423c4bfdf6fe

References

- Soifua, Breckell, “A Comparison of R, SAS, and Python Implementations of Random Forests” (2018). *All Graduate Plan B and other Reports.* 1268. https://digitalcommons.usu.edu/gradreports/1268
- Stack Overflow (2011, April 4). How much data can R handle? [Forum post]. Retrieved from: https://stackoverflow.com/questions/5527850/how-much-data-can-r-handle
- Stagg, S. (2018, March 27). *Analysing 1.4 billion rows with Python.* Retrieved from: https://hackernoon.com/analysing-1-4-billion-rows-with-python-6cec86ca9d73
- Torpey, E. (2019, February 1). *The 2006-2016 projections: How did fast-growing occupations fare?* Retrieved from: https://www.bls.gov/careeroutlook/2019/data-on-display/projections-evaluation.htm
- Walia, A. (2017, June 19). *How to Implement Deep Learning in R using Keras and TensorFlow*. Retrieved from: https://towardsdatascience.com/how-to-implement-deep-learning-in-r-using-keras-and-tensorflow-82d135ae4889
- Willems, K. (2019, February 12). *keras: Deep Learning in R.* Retrieved from: https://www.datacamp.com/community/tutorials/keras-r-deep-learning