Is R Outdated?

What is R?

R is the name for a programming language and environment for statistical analysis and graphics.

Use of R is declining

Is there a value to learning R?

Classes at Baruch are transitioning from R to Python

Is there a subset of users that still needs R?

Is R being replaced by another language?

RESEARCH QUESTIONS

1. What factors predict the use of R?
2. What languages are R users more likely to use than other people?

H1: Older respondents will be more likely to use R
H2: The use of R will increase the chance of using Python

METHODOLOGY

Question 1. What factors predict the use of R?

Answered with

- Logistical Regression
- Decision Tree

Question 2. What languages are R users more likely to use than other people?

Answered with

- Logistical Regression

kaggle

Data: 2021 Kaggle Machine Learning & Data Science Survey

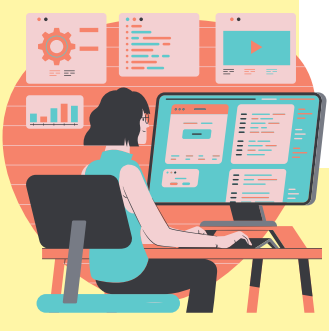
- 25,973 Responses
- 42 Questions

% of people using R is declining

- 2019: 23.3%
- 2020: 21.3%
- 2021: 20.5%

Common Job Titles for R Users

- Student 22.1%
- Data Scientist 21.4%
- Data/Business Analyst 17.4%



Only 5.6% of respondents recommend R as first language to learn

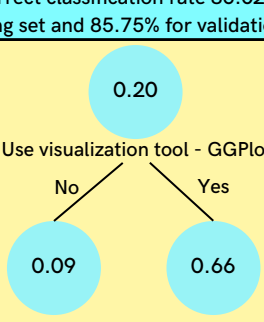
Most commonly used language is:

Python

77.8% of respondents recommend Python as first language to learn

Chance of Using R

(Correct classification rate 86.02% for training set and 85.75% for validation set)



92 variables were used in the decision tree and logistic regression

None of the variables were highly correlated with each other

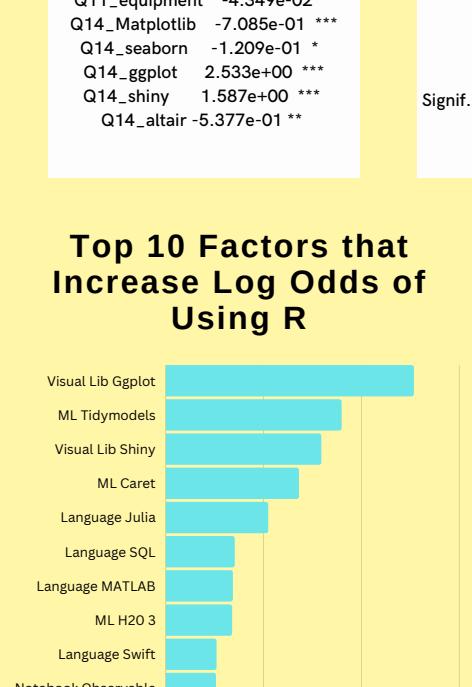
Chance of Using R in Log Odds

Logistic Regression Significant Results
(Full Results Link)

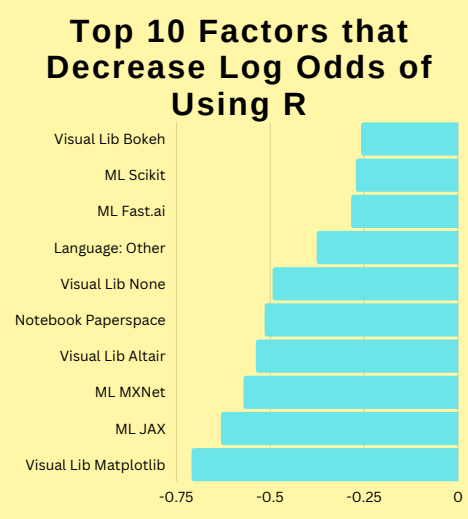
(Intercept)	-2.676e+00***
age	1.001e-02 ***
woman	1.222e-01 **
Country	1.576e-02
job	-2.992e-02 ***
yrs_coding	-3.492e-02 .
Q7_python	5.083e-01 ***
Q7_sql	7.023e-01 ***
Q7_c..	-2.432e-01 ***
Q7_java	-1.449e-01 *
Q7_javascript	-1.738e-01 **
Q7_julia	1.044e+00 ***
Q7_swift	5.162e-01 **
Q7_bash	1.347e-01 .
Q7_matlab	6.833e-01 ***
Q7_other2	-3.758e-01 ***
Q10_Kaggle	1.557e-01 **
Q10_Colab	9.378e-02 .
Q10_paper	-5.142e-01 *
Q10_sage	-2.537e-01 *
Q10_google_datalab	2.725e-01 ***
Q10_observable	5.096e-01 2 *
Q10_none	3.885e-01 ***
Q11_equipment	-4.349e-02 *
Q14_Matplotlib	-7.085e-01 ***
Q14_seaborn	-1.209e-01 *
Q14_ggplot	2.533e+00 ***
Q14_shiny	1.587e+00 ***
Q14_altair	-5.377e-01 **
Q14_bokeh	-2.579e-01 *
Q14_geoplotlib	-2.253e-01 *
Q14_none	-4.931e-01 ***
Q15_yrs_ml	9.519e-02 ***
Q16_scikit	-2.716e-01 ***
Q16_fast.ai	-2.842e-01 *
Q16_mxnet	-5.703e-01 **
Q16_prophet	2.493e-01 *
Q16_h2o	6.745e-01 ***
Q16_caret	1.358e+00 ***
Q16_tidymodels	1.794e+00 ***
Q16_jax	-6.304e-01 *
Q16_pytorch_lightning	2.019e-01 .
Q16_none	2.355e-01 **
Q17_linear	2.062e-01 ***
Q17_decision	1.245e-01 *
Q17_bayesian	3.225e-01 ***
Q17_convolutional	-1.363e-01 *
Q17_recurrent	1.321e-01 *
Q17_transformer	-2.514e-01 **
Q17_other	3.485e-01 *
Q21_num_emp	-3.312e-02 .
Q22_num_data_empl	-3.591e-02 **
Q24_ch2	-1.607e-01 **
Q24_ch3	-1.234e-01 .
Q24_ch6	2.249e-01 **
Q24_ch7	-2.315e-01 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Top 10 Factors that Increase Log Odds of Using R



Top 10 Factors that Decrease Log Odds of Using R



Top 10 Factors with the Highest Significance Levels

1. **MATLAB** 6.833e-01 p< 2e-16 ***
2. **SQL** 7.023e-01 p< 2e-16
3. **ML Caret** 1.358 p< 2e-16
4. **ML Tidymodels** 1.794 p< 2e-16
5. **Visual Lib GGPlot** 2.533 p< 2e-16
6. **Visual Lib Matplotlib** -7.085 p< 2e-16
7. **Visual Lib Shiny** 1.587 p< 2e-16
8. **Python** 5.083e-01 p=1.51e-12
9. **Years using ML** 9.519e-02 p=1.03e-11
10. **Julia** 1.044 p=1.06e-10

R Use Predicting Chance of Using Another Language in Log Odds

Logistic Regression Results

- **Julia** 1.27442 p<2e-16
- **SQL** 0.83330 p<2e-16
- **Swift** 0.616700 p=8.37E-06
- **MATLAB** 0.438690 p<2E-16
- **Bash** 0.384760 p=2.98E-14
- **Python** 0.16046 p=0.000228
- **Other** -0.130360 p=1.40E-02
- **Javascript** -0.16620 p=9.74e-05
- **Java** -0.19518 p=2.23e-06
- **C** -0.19842 p=1.74e-06
- **C++** -0.389630 p<2E-16
- **None** -16.411900 p=0.946

DISCUSSION

Chance of Using R

H1: Older respondents will be more likely to use R

Age and experience in machine learning increase likelihood of using R.

Greatest predictors were visualization and machine learning tools that pair with a coding language

Income wasn't significantly related to the use of R but job title was.

Chance of Using Other Languages in addition to R

H2: The use of R will increase the chance of using Python

Logistical regression was chosen to isolate the impact of using R from how popular each language is in general.

INCREASED CHANCE OF USING:
Julia, SQL, Swift, MATLAB, Bash, Python

DECREASED CHANCE OF USING:
C, C++, Java, Javascript, none, other

Limitations:
1. Questions weren't made for our study.
2. All results were correlations with no causality

FUTURE STUDIES

- Determine why less people are using R
- Explore what languages are being used in the place of R
- Investigate the reasons some users are staying with R
- Determine which languages predict a higher income