CIS 602-01: Scalable Data Analysis

Data Analysis Tools

Dr. David Koop
Analysis is a cycle

Start with basic computations, analyze results, and ask new questions
Types of EDA

- Univariate (one attribute) vs. multivariate (2+ attributes)
- Non-graphical vs. graphical
  - Non-graphical ~ statistics
  - Graphical ~ visualizations
- All are important!
**Univariate Non-Graphical: Word Counts**

- "This is the house that Jack built. This is the malt that lay in the house that Jack built. This is the rat that ate the malt That lay in the house that Jack built."

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>6</td>
<td>0.171</td>
</tr>
<tr>
<td>that</td>
<td>5</td>
<td>0.143</td>
</tr>
<tr>
<td>This</td>
<td>3</td>
<td>0.086</td>
</tr>
<tr>
<td>is</td>
<td>3</td>
<td>0.086</td>
</tr>
<tr>
<td>house</td>
<td>3</td>
<td>0.086</td>
</tr>
<tr>
<td>Jack</td>
<td>3</td>
<td>0.086</td>
</tr>
<tr>
<td>built.</td>
<td>3</td>
<td>0.086</td>
</tr>
<tr>
<td>malt</td>
<td>2</td>
<td>0.057</td>
</tr>
<tr>
<td>lay</td>
<td>2</td>
<td>0.057</td>
</tr>
<tr>
<td>in</td>
<td>2</td>
<td>0.057</td>
</tr>
</tbody>
</table>
Univariate Graphical: Histograms

Observed ranks of posts by subreddit

["The reddit Front Page is Not a Meritocracy", T. W. Schneider]
Multivariate Non-Graphical: Crosstabs

- Count groups and subgroups
- At least two different attributes
- Can subdivide vertically and horizontally for more subgroups
- Sometimes totals are useful

```
<table>
<thead>
<tr>
<th>region</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>56</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>206</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>36</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>33</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>51</td>
</tr>
</tbody>
</table>
```
Multivariate Graphical: Parallel Coordinates
Progressive Visualization
Blocking vs. Progressive Visualization


time = dataset-delay

---

blocking

progressive

---

Our system is implemented in C# / Direct2D. We tested our system and ran all sessions on a quad-core 3.60 GHz, 16 GB RAM, Microsoft Windows 10 desktop machine with 1,080 pixel display.

For example, participant P14 was assigned dataset-ordering (blocking, instantaneous, progressive). In total, this adds up to 24 possible combinations of visualization, dataset-ordering, and dataset-delay conditions, all users saw all visualization conditions. To control against ordering and learning effects, we included visualization condition as a within-subject factor and dataset-delay and dataset-ordering as between-subject factors. Note that the dataset-delay has no direct influence on visualizations. We approximate the instantaneous condition, but it is used as a between-subject factor.

We recruited 24 participants from a research university in the United States. The participants were currently enrolled in and halfway through an introductory data science course. Our experiment has been shown to be problematic to comprehend in certain conditions.

Our system is implemented in C# / Direct2D. We tested our system and ran all sessions on a quad-core 3.60 GHz, 16 GB RAM, Microsoft Windows 10 desktop machine with 1,080 pixel display.

Upon startup, the system loads the selected dataset into memory. The system randomly shuffles the data points in their size can change significantly based on the dataset and analysis. The two possible dataset-ordering values were 123 (DS1 followed by DS2) and 123 (DS2 followed by DS1) within the range of different visualization conditions to the user and counterbalanced across dataset-ordering. We seed our system's random number generator differently for each session. Instead of fully randomizing the sequence in which we presented the different visualization conditions, all users saw all visualization conditions. To control against ordering and learning effects, we included visualization condition as a within-subject factor and dataset-delay and dataset-ordering as between-subject factors. Note that the dataset-delay has no direct influence on visualizations. We approximate the instantaneous condition, but it is used as a between-subject factor.
Experiment Results

Fig. 5. Insights per minute: Boxplot (showing median and whiskers at 1.5 interquartile range) and overlaid swarmplot for insights per minute (left overall, (middle) by dataset-delay, and (right) by dataset-order. Higher values indicate better.
Assignment 1

- Boston Property Assessments
  - Initial exploratory analysis
  - Use a Python Notebook
  - May use pandas
  - Label subproblems and answers
  - Show work (even if it's not your final answer)
Test 1

• Thursday, October 5, 5-6:15pm, Dion 109 (here)
• Format:
  - Multiple Choice
  - Free Response
• Covers papers and topics discussed through this Thursday
• Next week:
  - No class on Tuesday, Oct. 3
  - Exam on Thursday, Oct. 5
  - No office hours
  - Email me with questions
Data Analysis Tools
Lots and lots of tools

- Tableau
- Qlikview
- Spotfire
- Trifacta
- KNIME
- Open Refine
- Octoparse
- Watson
- SAS
- Mahout
- Spark
- scikit-learn
Tables of Tools

• Data Architecture/Storage
• Data Cleaning
• Data Visualization
• Programming Languages
• Spreadsheets
• Machine Learning
O'Reilly's Data Science Salary Survey (2016)
Data Science Tasks

- **Conducting Data Analysis to Answer Research Questions**: 61%
- **BASIC EXPLORATORY DATA ANALYSIS**: 69%
- **Communicating Findings to Business Decision-Makers**: 58%
- **Developing Visualizations**: 49%
- **Identifying Business Problems to Be Solved with Analytics**: 47%
- **Developing Prototype Models**: 43%
- **Feature Extraction**: 43%
- **Collaborating on Code Projects (Reading/Editing Others’ Code, Using Git)**: 32%
- **Teaching/Training Others**: 31%
- **Planning Large Software Projects or Data Systems**: 30%
- **Developing Dashboards**: 30%
- **COMMUNICATING WITH PEOPLE OUTSIDE YOUR COMPANY**: 28%
- **ETL**: 29%
- **Organizing and Guiding Team Projects**: 39%
- **Implementing Models/Algorithms into Production**: 36%
- **Developing Data Analytics Software**: 20%
- **Implementing Models/Algorithms into Production**: 36%
- **Setting Up / Maintaining Data Platforms**: 24%
- **Developing Products That Depend on Real-Time Data Analytics**: 19%
- **Developing Hardware (or Working on Software Projects That Require Expert Knowledge of Hardware)**: 5%
- **Using Dashboards and Spreadsheets (Made by Others) to Make Decisions**: 19%
- **Developing Products That Depend on Real-Time Data Analytics**: 19%
- **ETL**: 29%
- **Developing Products That Depend on Real-Time Data Analytics**: 19%
- **Using Dashboards and Spreadsheets (Made by Others) to Make Decisions**: 19%
- **Developing Hardware (or Working on Software Projects That Require Expert Knowledge of Hardware)**: 5%
Data Science Tasks

Basic exploratory data analysis
Conducting data analysis to answer research questions
Communicating findings to business decision-makers
Data cleaning
Creating visualizations
Identifying business problems to be solved with analytics
Feature extraction
Developing prototype models
Organizing and guiding team projects
Implementing models / algorithms into production
Collaborating on code projects (reading / editing others’ code, using git)
Teaching / training others
Planning large software projects or data systems
Developing dashboards
ETL
Communicating with people outside your company
Setting up / maintaining data platforms
Developing data analytics software
Developing products that depend on real-time data analytics
Using dashboards and spreadsheets (made by others) to make decisions
Developing hardware (or working on software projects that require expert knowledge of hardware)
Programming Languages

Languages: SQL, R, Python, Bash, Java, JavaScript, Visual Basic/VBA, C++, Matlab, Scala, C, C#, Perl, SAS, Ruby, Octave, Go

Salary Median and IQR (US Dollars)

Range/Median

Languages

SQL: 150K - 200K
R: 100K - 150K
Python: 150K - 200K
Bash: 0 - 50K
Java: 0 - 50K
JavaScript: 0 - 50K
Visual Basic/VBA: 0 - 50K
C++: 0 - 50K
Matlab: 0 - 50K
Scala: 0 - 50K
C: 0 - 50K
C#: 0 - 50K
Perl: 0 - 50K
SAS: 0 - 50K
Ruby: 0 - 50K
Octave: 0 - 50K
Go: 0 - 50K

O'Reilly
Database Tools

RELATIONAL DATABASES

- MySQL: 37%
- Microsoft SQL Server: 33%
- Oracle: 23%
- PostgreSQL: 22%
- SQLite: 11%
- IBM DB2: 10%
- Teradata: 10%
- EMC/Greenplum: 2%
- Aster Data (Teradata): 2%
- EMC/Greenplum: 2%
- Redshift: 1%
- SAP HANA: 1%
- Oracle Exascale: 1%
- Netezza (IBM): 1%
- Netezza (IBM): 4%
- IBM DB2: 5%
- Teradata: 11%
- PostgreSQL: 22%
- Oracle: 23%
- Microsoft SQL Server: 33%
- SQLite: 11%
- IBM DB2: 10%
- Teradata: 10%
- EMC/Greenplum: 2%
- Aster Data (Teradata): 2%
- EMC/Greenplum: 2%
- Redshift: 1%
- SAP HANA: 1%
- Oracle Exascale: 1%
- Netezza (IBM): 1%
- Netezza (IBM): 4%
- IBM DB2: 5%
- Teradata: 11%
- PostgreSQL: 22%
- Oracle: 23%
- Microsoft SQL Server: 33%
- SQLite: 11%
- IBM DB2: 10%
- Teradata: 10%
- EMC/Greenplum: 2%
- Aster Data (Teradata): 2%
- EMC/Greenplum: 2%
- Redshift: 1%
- SAP HANA: 1%
- Oracle Exascale: 1%
- Netezza (IBM): 1%
- Netezza (IBM): 4%
- IBM DB2: 5%
- Teradata: 11%
- PostgreSQL: 22%
- Oracle: 23%
- Microsoft SQL Server: 33%
- SQLite: 11%
- IBM DB2: 10%
- Teradata: 10%
- EMC/Greenplum: 2%
- Aster Data (Teradata): 2%
- EMC/Greenplum: 2%
- Redshift: 1%
- SAP HANA: 1%
- Oracle Exascale: 1%
- Netezza (IBM): 1%
- Netezza (IBM): 4%
- IBM DB2: 5%
- Teradata: 11%
- PostgreSQL: 22%
- Oracle: 23%
- Microsoft SQL Server: 33%
- SQLite: 11%
- IBM DB2: 10%
- Teradata: 10%
- EMC/Greenplum: 2%
- Aster Data (Teradata): 2%
- EMC/Greenplum: 2%
- Redshift: 1%
- SAP HANA: 1%
- Oracle Exascale: 1%
- Netezza (IBM): 1%
- Netezza (IBM): 4%
- IBM DB2: 5%
- Teradata: 11%
- PostgreSQL: 22%
- Oracle: 23%
- Microsoft SQL Server: 33%
- SQLite: 11%
- IBM DB2: 10%
- Teradata: 10%
- EMC/Greenplum: 2%
- Aster Data (Teradata): 2%
- EMC/Greenplum: 2%
- Redshift: 1%
- SAP HANA: 1%
- Oracle Exascale: 1%
- Netezza (IBM): 1%
- Netezza (IBM): 4%
- IBM DB2: 5%
- Teradata: 11%
- PostgreSQL: 22%
- Oracle: 23%
- Microsoft SQL Server: 33%
- SQLite: 11%
- IBM DB2: 10%
- Teradata: 10%
- EMC/Greenplum: 2%
- Aster Data (Teradata): 2%
- EMC/Greenplum: 2%
- Redshift: 1%
- SAP HANA: 1%
- Oracle Exascale: 1%
- Netezza (IBM): 1%
- Netezza (IBM): 4%
- IBM DB2: 5%
- Teradata: 11%
- PostgreSQL: 22%
- Oracle: 23%
- Microsoft SQL Server: 33%
- SQLite: 11%
- IBM DB2: 10%
- Teradata: 10%
- EMC/Greenplum: 2%
- Aster Data (Teradata): 2%
- EMC/Greenplum: 2%
- Redshift: 1%
- SAP HANA: 1%
- Oracle Exascale: 1%
- Netezza (IBM): 1%
- Netezza (IBM): 4%
- IBM DB2: 5%
- Teradata: 11%
- PostgreSQL: 22%
- Oracle: 23%
- Microsoft SQL Server: 33%
- SQLite: 11%
- IBM DB2: 10%
- Teradata: 10%
- EMC/Greenplum: 2%
- Aster Data (Teradata): 2%
- EMC/Greenplum: 2%
- Redshift: 1%
- SAP HANA: 1%
- Oracle Exascale: 1%
- Netezza (IBM): 1%
- Netezza (IBM): 4%
- IBM DB2: 5%
- Teradata: 11%
- PostgreSQL: 22%
- Oracle: 23%
- Microsoft SQL Server: 33%
- SQLite: 11%
- IBM DB2: 10%
- Teradata: 10%
- EMC/Greenplum: 2%
- Aster Data (Teradata): 2%
- EMC/Greenplum: 2%
- Redshift: 1%
- SAP HANA: 1%
- Oracle Exascale: 1%
- Netezza (IBM): 1%
- Netezza (IBM): 4%
- IBM DB2: 5%
- Teradata: 11%
- PostgreSQL: 22%
- Oracle: 23%
- Microsoft SQL Server: 33%
- SQLite: 11%
- IBM DB2: 10%
- Teradata: 10%
- EMC/Greenplum: 2%
- Aster Data (Terada
Hadoop and Search Tools

**HADOOP**

**SHARE OF RESPONDENTS**

- **2%** IBM
- **4%** Oracle
- **7%** Amazon Elastic MapReduce (EMR)
- **8%** Hortonworks
- **12%** Cloudera
- **17%** Apache Hadoop
- **1%** EMC/Greenplum

**SALARY MEDIAN AND IQR (US DOLLARS)**

- Apache Hadoop
- Cloudera
- Hortonworks
- Amazon Elastic MapReduce (EMR)
- MapR
- Oracle
- EMC/Greenplum
- IBM

**SEARCH**

**SHARE OF RESPONDENTS**

- **10%** ElasticSearch
- **5%** Solr
- **4%** Lucene

**SALARY MEDIAN AND IQR (US DOLLARS)**

- ElasticSearch
- Solr
- Lucene

[O'Reilly]
Data Management Tools

DATA MANAGEMENT, BIG DATA PLATFORMS

SHARE OF RESPONDENTS

SALARY MEDIAN AND IQR (US DOLLARS)

Big Data Platforms

Spark
Hive
MongoDB
Amazon RedShift
Hbase
Kafka
Pig
Impala
Toad
Cassandra
Zookeeper
Redis
Neo4J
Google BigQuery/Fusion Tables
Splunk
Amazon DynamoDB
Storm
Couchbase

Range/Median
0 50K 100K 150K 200K
Visualization Tools

- **SHARE OF RESPONDENTS**
  - GGPlot: 35%
  - Tableau: 33%
  - Matplotlib: 26%
  - D3: 16%
  - Shiny: 16%
  - Bokeh: 8%
  - Python (via Plotly): 6%
  - Processing: 1%
  - JavaScript InfoVis Toolkit: 1%
  - Google Charts: 1%

**VISUALIZATION TOOLS**

**SALARY MEDIAN AND IQR (US DOLLARS)**

- ggplot
- Tableau
- Matplotlib
- Shiny
- D3
- Google Charts
- Bokeh
- Processing
- JavaScript InfoVis Toolkit

Median and IQR values for salary distributions across different visualization tools.
Machine Learning Tools

- **Scikit-learn**: 31%
- **Spark MLLib**: 13%
- **WEKA**: 9%
- **H2O**: 5%
- **RAPIDMINER**: 4%
- **LIBSVM**: 4%
- **MATHEMATICA**: 3%
- **Stata**: 3%
- **Mahout**: 3%
- **LIBSVM**: 4%
- **RapidMiner**: 4%
- **Google Prediction**: 1%
- **IBM Big Insights**: 1%
- **Google Prediction**: 1%

**Salary Median and IQR (US Dollars)**

- Machine learning, statistics
- Share of respondents

**Salary Range:**
- 30K
- 60K
- 90K
- 120K
- 150K

---

D. Koop, CIS 602-01, Fall 2017
Trends in Data Analysis Tool Use

KDNuggets Analytics, Data Science, Machine Learning Software Poll, top tools share, 2015-2017

- Python
- R language
- SQL language
- RapidMiner
- Excel
- Spark
- Anaconda
- Tensorflow
- scikit-learn
- Tableau
- KNIME

[2017 %share]

[2016 %share]

[2015 %share]
## Trends in Data Analysis Tool Usage

### Table 1: Top Analytics/Data Science Tools in 2017 KDnuggets Poll

<table>
<thead>
<tr>
<th>Tool</th>
<th>2017 % Usage</th>
<th>% change 2017 vs 2016</th>
<th>% alone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python</td>
<td>52.6%</td>
<td>15%</td>
<td>0.2%</td>
</tr>
<tr>
<td>R language</td>
<td>52.1%</td>
<td>6.4%</td>
<td>3.3%</td>
</tr>
<tr>
<td>SQL language</td>
<td>34.9%</td>
<td>-1.8%</td>
<td>0%</td>
</tr>
<tr>
<td>RapidMiner</td>
<td>32.8%</td>
<td>0.7%</td>
<td>13.6%</td>
</tr>
<tr>
<td>Excel</td>
<td>28.1%</td>
<td>-16%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Spark</td>
<td>22.7%</td>
<td>5.3%</td>
<td>0.2%</td>
</tr>
<tr>
<td>Anaconda</td>
<td>21.8%</td>
<td>37%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Tensorflow</td>
<td>20.2%</td>
<td>195%</td>
<td>0%</td>
</tr>
<tr>
<td>scikit-learn</td>
<td>19.5%</td>
<td>13%</td>
<td>0%</td>
</tr>
<tr>
<td>Tableau</td>
<td>19.4%</td>
<td>5.0%</td>
<td>0.4%</td>
</tr>
<tr>
<td>KNIME</td>
<td>19.1%</td>
<td>6.3%</td>
<td>2.4%</td>
</tr>
</tbody>
</table>
# Top Increases

## Table 2: Major Analytics/Data Science Tools with the largest increase in usage

<table>
<thead>
<tr>
<th>Tool</th>
<th>% change</th>
<th>2017 % usage</th>
<th>2016 % usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft, CNTK</td>
<td>294%</td>
<td>3.4%</td>
<td>0.9%</td>
</tr>
<tr>
<td>Tensorflow</td>
<td>195%</td>
<td>20.2%</td>
<td>6.8%</td>
</tr>
<tr>
<td>Microsoft Power BI</td>
<td>84%</td>
<td>10.2%</td>
<td>5.6%</td>
</tr>
<tr>
<td>Alteryx</td>
<td>76%</td>
<td>5.3%</td>
<td>3.0%</td>
</tr>
<tr>
<td>SQL on Hadoop tools</td>
<td>42%</td>
<td>10.3%</td>
<td>7.3%</td>
</tr>
<tr>
<td>Microsoft other ML/Data Science tools</td>
<td>40%</td>
<td>2.2%</td>
<td>1.6%</td>
</tr>
<tr>
<td>Anaconda</td>
<td>37%</td>
<td>21.8%</td>
<td>16.0%</td>
</tr>
<tr>
<td>Caffe</td>
<td>32%</td>
<td>3.1%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Orange</td>
<td>30%</td>
<td>4.0%</td>
<td>3.1%</td>
</tr>
<tr>
<td>DL4J</td>
<td>30%</td>
<td>2.2%</td>
<td>1.7%</td>
</tr>
<tr>
<td>Other Deep Learning Tools</td>
<td>30%</td>
<td>4.8%</td>
<td>3.7%</td>
</tr>
<tr>
<td>Microsoft Azure Machine Learning</td>
<td>26%</td>
<td>6.4%</td>
<td>5.1%</td>
</tr>
</tbody>
</table>
# Top Decreases

Table 3: Major Analytics/Data Science Tools with the largest decline in usage

<table>
<thead>
<tr>
<th>Tool</th>
<th>% change</th>
<th>2017 % usage</th>
<th>2016 % usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turi (former Dato/GraphLab)</td>
<td>-93%</td>
<td>0.2%</td>
<td>2.4%</td>
</tr>
<tr>
<td>RapidInsight/Veera</td>
<td>-92%</td>
<td>0.2%</td>
<td>3.0%</td>
</tr>
<tr>
<td>Salford SPM/CART/RF/MARS/TreeNet</td>
<td>-89%</td>
<td>0.4%</td>
<td>3.5%</td>
</tr>
<tr>
<td>MLlib</td>
<td>-61%</td>
<td>4.5%</td>
<td>11.6%</td>
</tr>
<tr>
<td>C4.5/C5.0/See5</td>
<td>-38%</td>
<td>1.2%</td>
<td>2.0%</td>
</tr>
<tr>
<td>Hadoop: Open Source Tools</td>
<td>-32%</td>
<td>15.0%</td>
<td>22.1%</td>
</tr>
<tr>
<td>Other free analytics/data mining tools</td>
<td>-29%</td>
<td>4.8%</td>
<td>6.8%</td>
</tr>
<tr>
<td>Rattle</td>
<td>-28%</td>
<td>2.6%</td>
<td>3.6%</td>
</tr>
<tr>
<td>Perl</td>
<td>-27%</td>
<td>1.7%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Pentaho</td>
<td>-23%</td>
<td>1.8%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Gnu Octave</td>
<td>-22%</td>
<td>2.4%</td>
<td>3.1%</td>
</tr>
<tr>
<td>QlikView</td>
<td>-21%</td>
<td>4.2%</td>
<td>5.3%</td>
</tr>
</tbody>
</table>
Data Science & Machine Learning Associations

KDnuggets 2017 Data Science Software Poll: Top Tools Associations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Python</td>
<td>17%</td>
<td>22%</td>
<td>-17%</td>
<td>50%</td>
<td>82%</td>
<td>68%</td>
<td>86%</td>
<td>18%</td>
<td>18%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2: R language</td>
<td>17%</td>
<td>22%</td>
<td>18%</td>
<td>22%</td>
<td>19%</td>
<td>36%</td>
<td>74%</td>
<td>36%</td>
<td>28%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3: SQL language</td>
<td>22%</td>
<td>22%</td>
<td>48%</td>
<td>27%</td>
<td>41%</td>
<td>66%</td>
<td>79%</td>
<td>36%</td>
<td>66%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4: RapidMiner</td>
<td>-17%</td>
<td>-32%</td>
<td>-29%</td>
<td>-21%</td>
<td>40%</td>
<td>21%</td>
<td>25%</td>
<td>40%</td>
<td>21%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5: Excel</td>
<td>18%</td>
<td>48%</td>
<td>-22%</td>
<td>-22%</td>
<td>53%</td>
<td>106%</td>
<td>179%</td>
<td>50%</td>
<td>74%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6: Spark</td>
<td>50%</td>
<td>27%</td>
<td>-32%</td>
<td>-22%</td>
<td>106%</td>
<td>79%</td>
<td>50%</td>
<td>40%</td>
<td>-16%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7: Anaconda</td>
<td>82%</td>
<td>41%</td>
<td>-29%</td>
<td>43%</td>
<td>53%</td>
<td>112%</td>
<td>159%</td>
<td>40%</td>
<td>-16%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8: Tensorflow</td>
<td>68%</td>
<td>-21%</td>
<td>106%</td>
<td>112%</td>
<td>159%</td>
<td>40%</td>
<td>40%</td>
<td>-16%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9: scikit-learn</td>
<td>86%</td>
<td>36%</td>
<td>24%</td>
<td>79%</td>
<td>178%</td>
<td>159%</td>
<td>40%</td>
<td>16%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10: Tableau</td>
<td>18%</td>
<td>21%</td>
<td>74%</td>
<td>50%</td>
<td>40%</td>
<td>16%</td>
<td>40%</td>
<td>-16%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11: KNIME</td>
<td>18%</td>
<td>28%</td>
<td>66%</td>
<td>-25%</td>
<td>36%</td>
<td>16%</td>
<td>16%</td>
<td>-16%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

lifting
Python vs. R

KDnuggets 2017 Data Science Poll: Python vs R associations

- scikit-learn
- PyCharm
- Keras
- Anaconda
- Dataiku
- TensorFlow
- C/C++
- Scala
- Spark
- Unix shell/awk/gawk
- Java
- MATLAB
- Open Source Hadoop Tools
- Other programming and data lang...
- Commercial Hadoop Tools
- SQL on Hadoop tools
- SQL language
- IBM SPSS Modeler
- Excel
- KNIME
- RapidMiner
- Tableau
- Weka
- Microsoft SQL Server
- Microsoft Azure Machine Learning
- IBM SPSS Statistics
- Microsoft Power BI
- SAS Base

log(Py/R)

-0.768 0.768
Hadoop/Spark Association

[Image: KDnuggets 2017 Data Science Software Poll: Deep Learning vs Hadoop/Spark affinity]
Structured Query Language (SQL)

• Used with relational database management systems (RDBMS)
• Based on relational algebra
• Supports:
  - Data schema definitions
  - Add/delete/update data
  - Data query
Excel

- One of the most universal data analysis tools
- Supports:
  - Spreadsheets
  - Calculated cells
  - Visualizations
- Many similar products
Tableau

• Grew out of research at Stanford University on how to explore multidimensional datasets & relational databases
• Tableau Desktop: standalone (free trial, student license)
• Tableau Public: cloud-based system (free)
• Tableau Vizable: mobile app
• High-level GUI that connects to data, helps organize it, and provides intuitive routines for visualizing it plus customization
• Lots of possibilities for exploring data
Trifacta

• Grew out of research at Stanford University and UC-Berkeley
• Data Wrangling: Data Cleaning Tool
• Prepare data for analysis
• Trifacta Wrangler: free desktop app, integrate with Tableau
• Trifacta Enterprise Wrangler: integration with Hadoop
• Google's Cloud Dataprep integrates Trifacta Enterprise
Python Tools

- **matplotlib:**
  - Python's "base" visualization library
  - Originally mimicked the functions in matlab
  - Integrated with pandas .plot calls
  - Jupyter Notebook: `%matplotlib inline` or `notebook`
  - seaborn adds extras on top of matplotlib

- **scikit-learn**
  - Machine learning library
  - Fit and predict using models
  - Classification and regression