.Intermediate Report

1. Describe project goals and why it is interesting

FINISHED

We plan to use this data to predict which languages are being used, what they're being used for (commercial, fun, etc). This data is interesting is because github is one the most used git repositories and allows the public to see the trend of languages throughout the years and see what type of projects people create.

Took 5 sec. Last updated by anonymous at March 23 2018, 9:10:50 AM.

2. Describe data collection/source of data, data format, data preprocessing

FINISHED

This dataset gives us info on what languages are used, its files (including their contents) and their size along with info on commits. We will use a combination of the data from bigquery along with using Github's official api using javascript. Bigquery's data is listed as tables, while the api will give us info in the form of a json file. With bigquery, we have to eliminate rows that have a date before the creation of github (which is about 0.04% (10 million) of the total rows. With the api, we eliminate some of the redundant information such as owner of repo (since unless you're searching, you get the repo based off the username).

Took 0 sec. Last updated by anonymous at March 23 2018, 9:11:15 AM.

3. Describe contents of data in detail (write code to analyse and visualize it)

FINISHED

Took 0 sec. Last updated by anonymous at March 23 2018, 9:11:30 AM.

FINISHED

4. Describe possible applications of this data, Including your ideas for the next phase

Took 0 sec. Last updated by anonymous at March 23 2018, 9:11:56 AM.

FINISHED

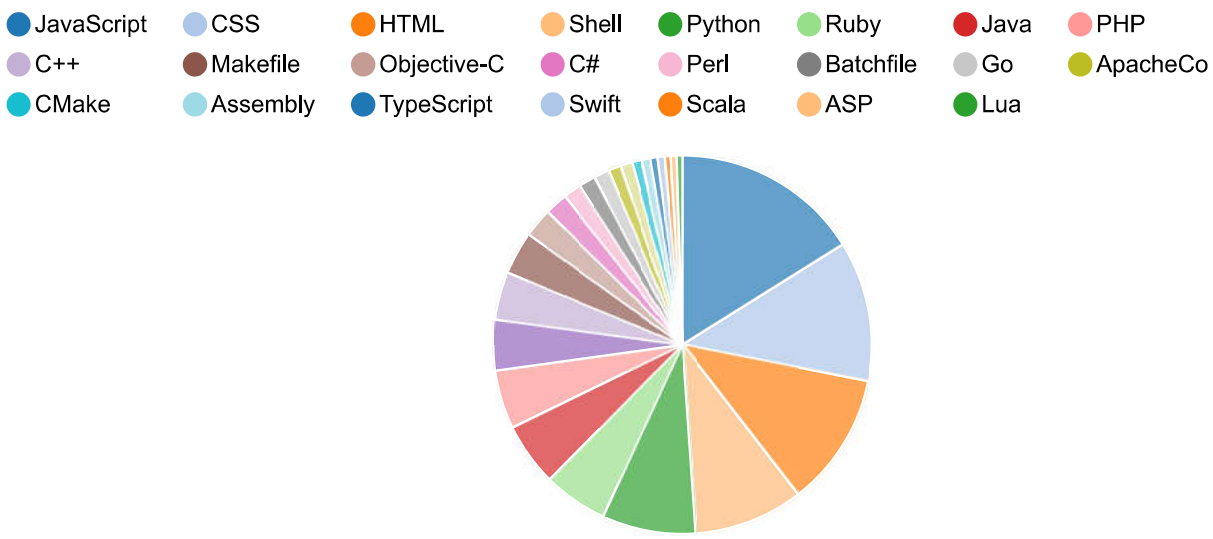
```
%r
test <- read.csv("c:/users/danny/documents/github/githubdatajs/repos.csv", header = TRUE, nrows = 1000
test = createDataFrame(test)
registerTempTable(test, "test")
```

Took 1 sec. Last updated by anonymous at March 14 2018, 10:09:59 PM.

FINISHED

```
%bigquery.sql
SELECT
  name,
  COUNT(*) AS occurences
FROM
  `bigquery-public-data.github_repos.languages` AS l,
  UNNEST(l.LANGUAGE)
GROUP BY
  name
ORDER BY
  occurences DESC
  limit 25
```

settings



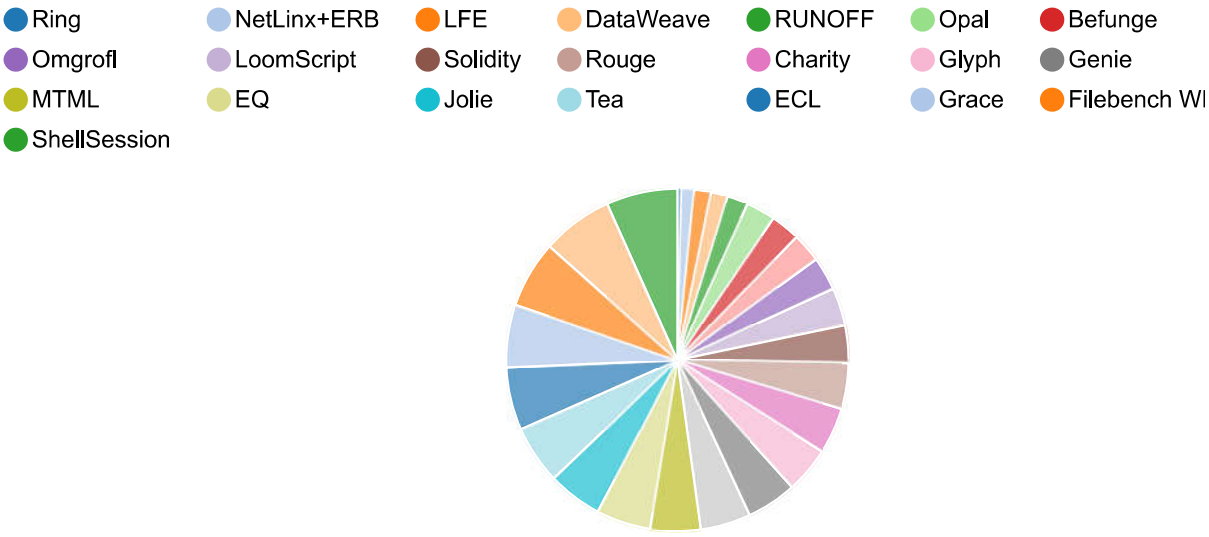
Took 3 sec. Last updated by anonymous at March 14 2018, 10:47:57 PM. (outdated)

FINISHED

```
%bigquery.sql
SELECT
  name,
  COUNT(*) AS occurences
FROM
  `bigquery-public-data.github_repos.languages` AS l,
```

```
  UNNEST(l.LANGUAGE)
GROUP BY
  name
ORDER BY
  occurences
```

settings

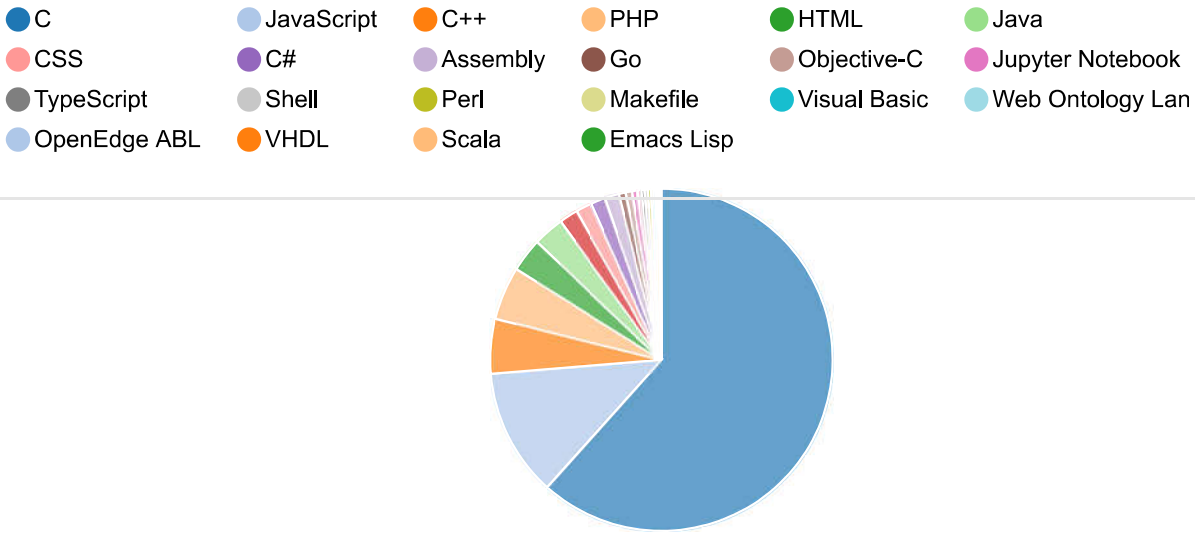


Took 4 sec. Last updated by anonymous at March 14 2018, 10:49:46 PM. (outdated)

FINISHED

```
%bigquery.sql
SELECT
  name,
  SUM(bytes) AS totalBytes
FROM
  `bigquery-public-data.github_repos.languages` AS l,
  UNNEST(l.LANGUAGE)
GROUP BY
  name
ORDER BY
  totalBytes DESC
limit 25
```

settings

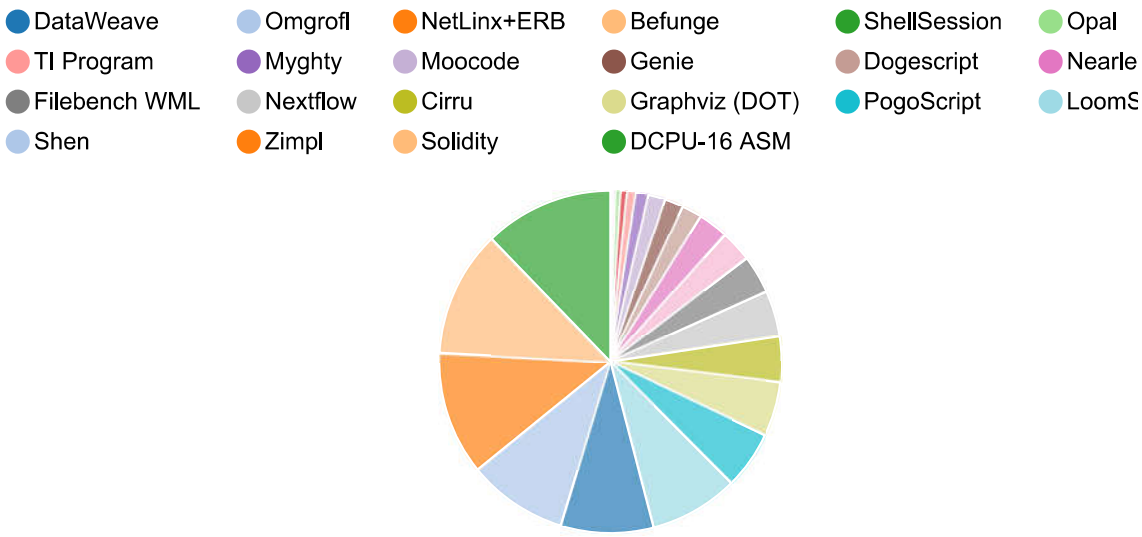


Took 5 sec. Last updated by anonymous at March 14 2018, 10:50:23 PM.

FINISHED

```
%bigquery.sql
SELECT
  name,
  SUM(bytes) AS totalBytes
FROM
  `bigquery-public-data.github_repos.languages` AS l,
  UNNEST(l.LANGUAGE)
GROUP BY
  name
ORDER BY
  totalBytes
limit 25
```

settings



Took 5 sec. Last updated by anonymous at March 14 2018, 10:51:35 PM. (outdated)

```
%bigquery.sql
```

READY