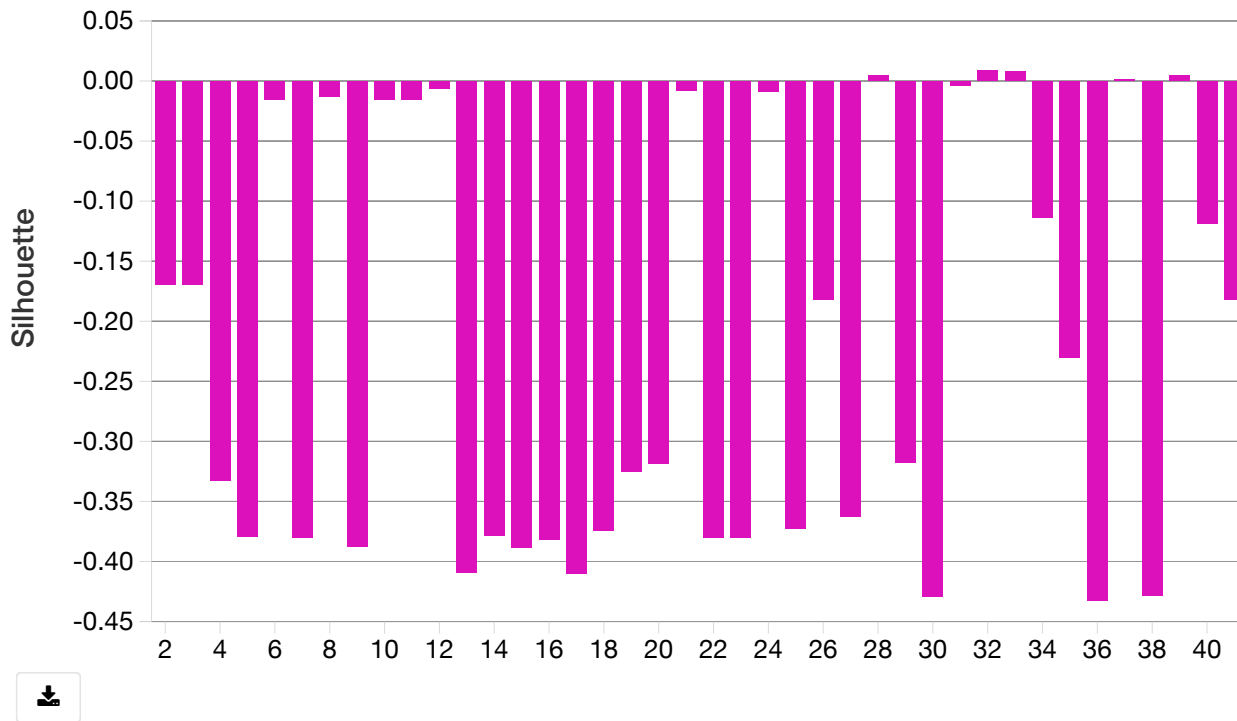**databricks** project

```
var df =
spark.read.format("csv").option("header","false").load("/FileStore/tables/part_
00000_0d2853db_22cc_41b1_ae97_82b2e116316d_c000-243e8.csv").toDF("K Value",
"Silhouette")
```

```
df: org.apache.spark.sql.DataFrame = [K Value: string, Silhouette: string]
```

Silhouette Width Analysis

```
display(df)
```



```
var df1 =
spark.read.format("csv").option("header","false").load("/FileStore/tables/part_
00000_e2ce2d25_42c4_4118_931f_0d6ddbbb2f34_c000-e5602.csv")
```

```
df1: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 3 more fiel
ds]
```

```
display(df1)
```

_c0
1848192576_15444933_2015-05-02
145162445X_20595117_2012-06-29
055341805X_52925878_2015-05-02
0425263908_40062567_2012-06-29
1416556141_47221489_2015-05-02
1455523003_20709973_2012-06-29
0990388697_18052694_2015-05-02
0345803485_48279565_2012-06-29
1469245167_14207627_2015_05_02

⬇

```
var df2 =
spark.read.format("csv").option("header","false").load("/FileStore/tables/part_
00000_e2ce2d25_42c4_4118_931f_0d6ddbbb2f34_c000-
e5602.csv").toDF("a","b","c","d","Real Review")

df2: org.apache.spark.sql.DataFrame = [a: string, b: string ... 3 more fields]

display(df2)
```

a		b
1848192576_15444933_2015-05-02		18
145162445X_20595117_2012-06-29		14
055341805X_52925878_2015-05-02		05
0425263908_40062567_2012-06-29		04
1416556141_47221489_2015-05-02		14
1455523003_20709973_2012-06-29		14
0990388697_18052694_2015-05-02		09
0345803485_48279565_2012-06-29		03
1469245167_14207627_2015_05_02		14

⬇

```
import org.apache.spark.sql._

import org.apache.spark.sql._
```

```
var df3 = df2.groupBy(df2("Real Review"))
  .count()

df3: org.apache.spark.sql.DataFrame = [Real Review: string, count: bigint]
```

Real Review Pie Chart

Real Rev

■ false
■ true

4%

96%