



Award #: 1640864

CIF21 DIBBs: El: Vizier, Streamlined Data Curation

PI: Oliver Kennedy¹, Co-Pis: Juliana Freire², Boris Glavic³

Institutions: ¹ University at Buffalo, ² New York University, ³ Illinois Institute of Technology



- Scalable open-source platform for data debugging and exploration

<https://github.com/VizierDB>

- Combines notebooks and spreadsheets for ease of use
- Captures provenance both for data and computations
 - Reproducibility, re-use, and transparency
 - Automated versioning
 - Tracking of potential data errors
- Vizier(Dirty Data) = Better Data + Provenance
 - Publish data with provenance and known caveats

goog

	High	Low	Open	Close	Volume	Adj. Close
-	1052.3199462890625	1015.7100219726562	1016.5700073242188	1045.8499755859375	1532600	1045.8499755859375
↑		1014.0700073242188	1041	1016.0599975585938	1841100	1016.0599975585938
↓		1027.41796875	1032.5899658203125	1070.7099609375	2093900	1070.7099609375
☒		1054.760009765625	1071.5	1068.3900146484375	1981900	1068.3900146484375
✚		1060.530029296875	1076.1099853515625	1076.280029296875	1764900	1076.280029296875
		1066.4000244140625	1081.6500244140625	1074.6600341796875	1199300	1074.6600341796875
		1057.7099609375	1067.6800341796875	1070.3299560546875	1456400	1070.3299560546875
-	1063.7750244140625	1048.47998046875	1063.1800537109375	1057.18994140625	1520800	1057.18994140625

```
[1] LOAD DATASET causes FROM FILE New_York_City_Leading_Causes_of_Death_12_11_2018.csv
```

```
causes (
  YEAR(int),
  LEADING_CAUSE(varchar),
  SEX(varchar),
  RACE_ETHNICITY(varchar),
  DEATHS(int),
  DEATH_RATE(real),
  AGE_ADJUSTED_DEATH_RATE(real)
)
1894 row(s)
```

```
[2] 1 SELECT sex, SUM(deaths) FROM causes GROUP BY sex
```

Output Dataset: by_gender

by_gender		SEX	SUM
0	M		207927
1	F		217071

```
[3] CREATE PLOT 'By Gender' FOR by_gender
```

Charts: Bar Chart

<https://vizierdb.info>