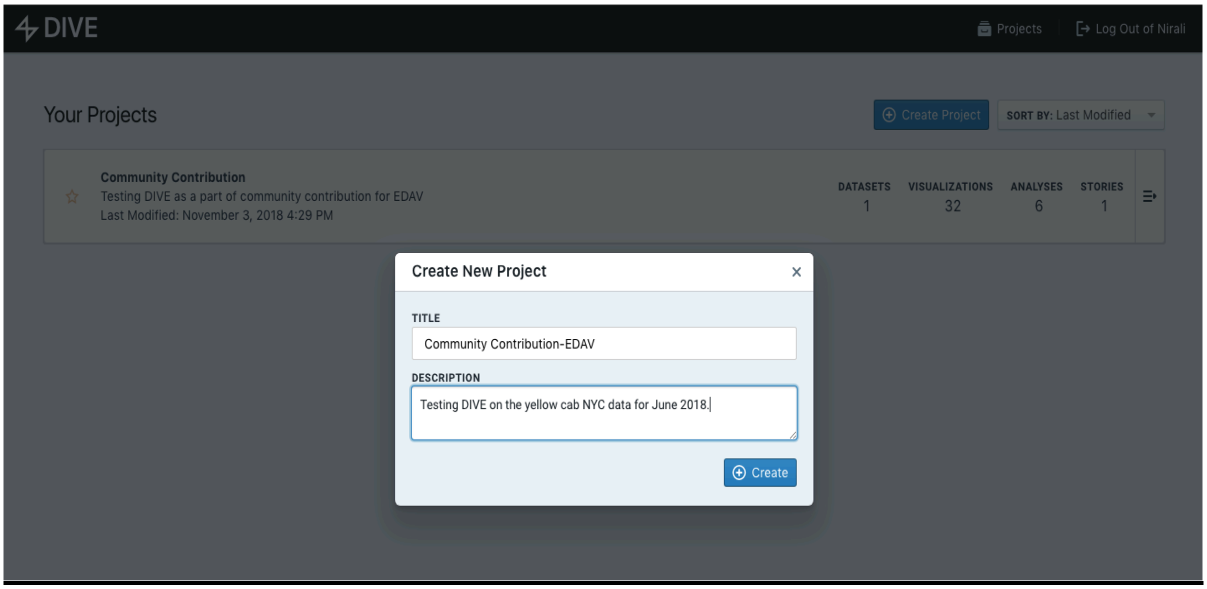
COMMUNITY CONTRIBUTION
TITLE- TUTORIAL ON DIVE (OPEN-SOURCE TOOL)
AUTHOR: NIRALI SHAH, UNI: NSS2173

DIVE

- DIVE is an open-source tool for recommendation-driven data exploration and visualization developed by Kevin Hu and César Hidalgo at the MIT Media Lab.
- It is a web-based data exploration system that allows users to visualize and explore their data without going through the hassle of writing any code.
- DIVE provides a single platform to perform:
  1. Semantic Data Ingestion
  2. Recommendation-based data visualization and analysis
  3. Dynamic story sharing
- Dive is an easy-to-learn data exploration tool.
- Dive is a mixed-initiative system combining recommender systems with point-and-click manual specification to support state-of-the-art data model inference, visualization, statistical analysis, and storytelling capabilities. [1]
- Tasks performed in order[2]:
  1. Intelligent Data Ingestion
  2. Semi-automated visualization recommendation
  3. Point-and-Click statistical analysis
  4. WYSIWYG (What You See Is What You Get) visual narratives

Steps to create a new project or work on an exisiting one:
1. To use the tool, visit the site https://dive.media.mit.edu/ .The website also contains a demo video that helps us understand the tool.
2. Create an account. Although the tool can be used without creating an account, it's better to create an account if one wants to switch between different projects while using the tool.
3. Log in to see the projects you've worked on. You can also create a new project.

I will be working on the NYC yellow cab data of June 2018.



Your Projects

Community Contribution
Testing DIVE as a part of community contribution for EDAV
Last Modified: November 3, 2018 4:29 PM

DATASETS	VISUALIZATIONS	ANALYSES	STORIES
1	32	6	1

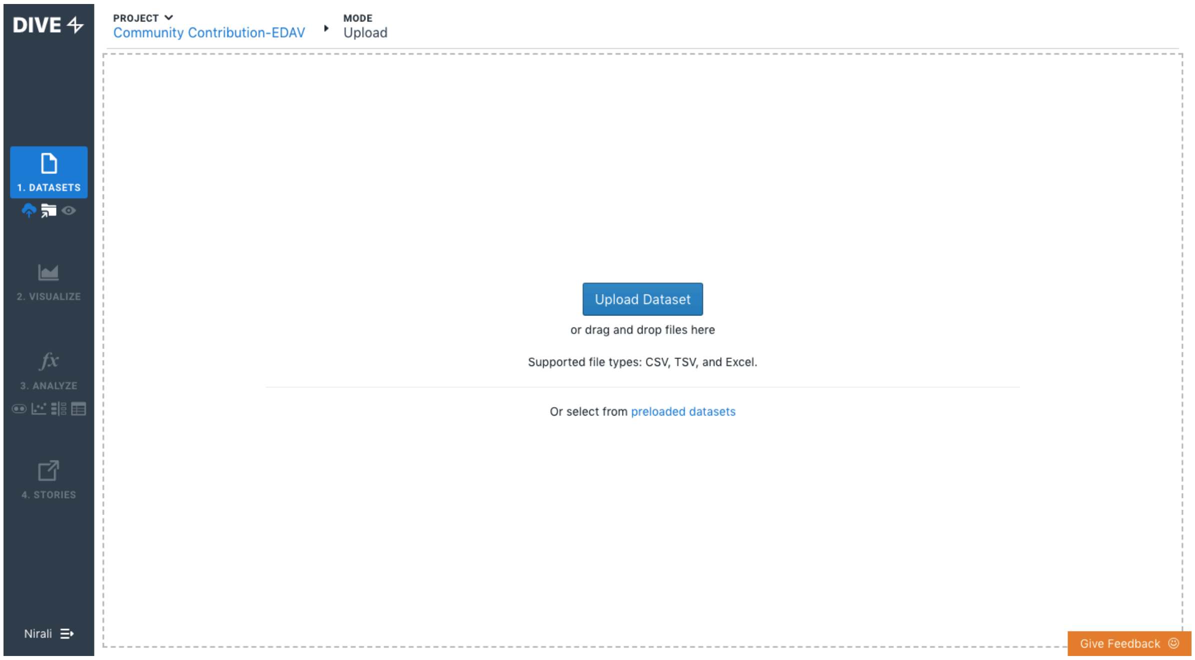
Create New Project

TITLE: Community Contribution-EDAV

DESCRIPTION: Testing DIVE on the yellow cab NYC data for June 2018.

1. Upload your dataset and inspect it using Inspect Mode:

It supports 3 file types: CSV, TSV and Excel



PROJECT: Community Contribution-EDAV

MODE: Upload

or drag and drop files here

Supported file types: CSV, TSV, and Excel.

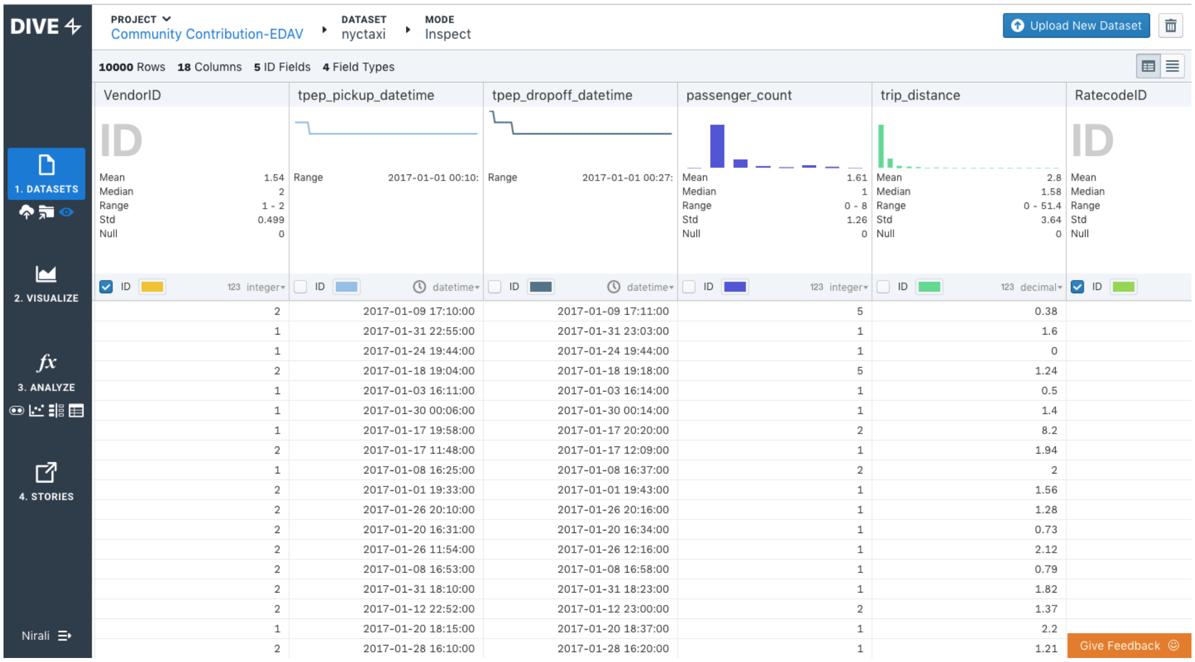
Or select from preloaded datasets

Once the data is uploaded, it computes dataset field properties and enters the **Inspect** mode.
The Inspect mode shows all the variables in the dataset and their field type.

Additional properties of each numerical or decimal variable:
1. Whether the variable is an ID.
2. Mean value
3.Median Value
4. Standard Deviation
5. Number of null values

For a string variable, it displays:
1. Number of unique values
2. Most frequent value
3. Number of occurences of the most frequent value
4. Number of null values

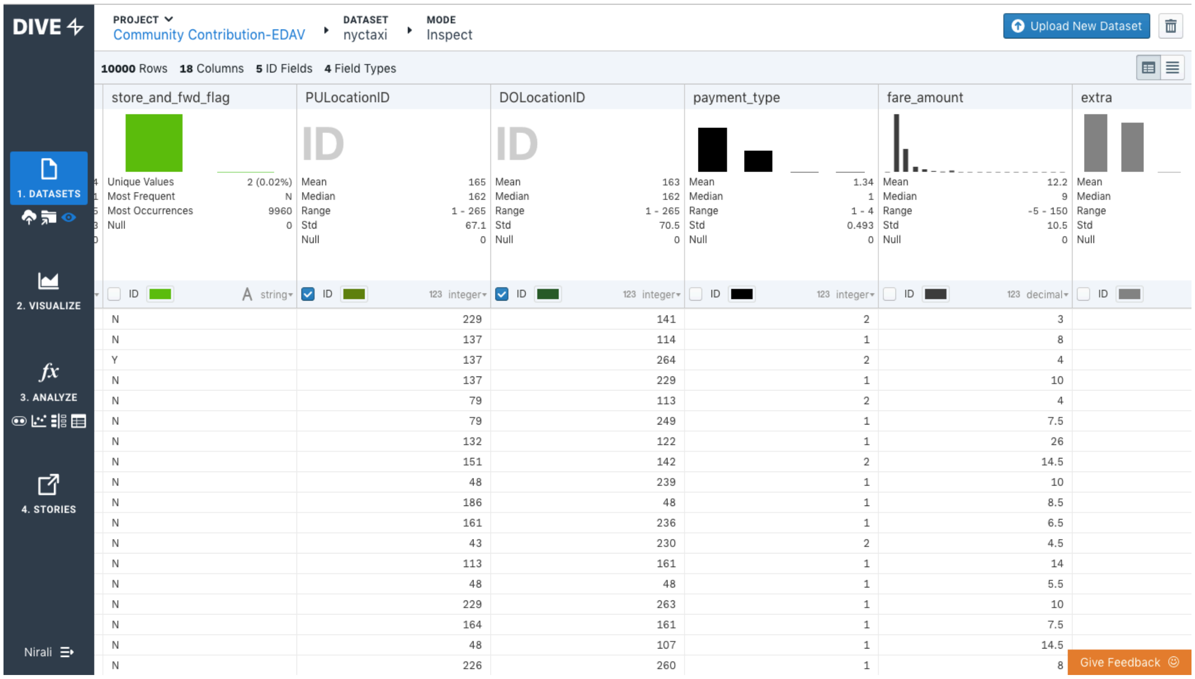
For a Date Time variable, it displays the range of values.
This helps us understand the dataset.
The type of the field can be changed and a field can be marked as a unique identifier if needed.



PROJECT
Community Contribution-EDAV ▸ DATASET nyctaxi ▸ MODE Inspect

10000 Rows **18** Columns **5** ID Fields **4** Field Types

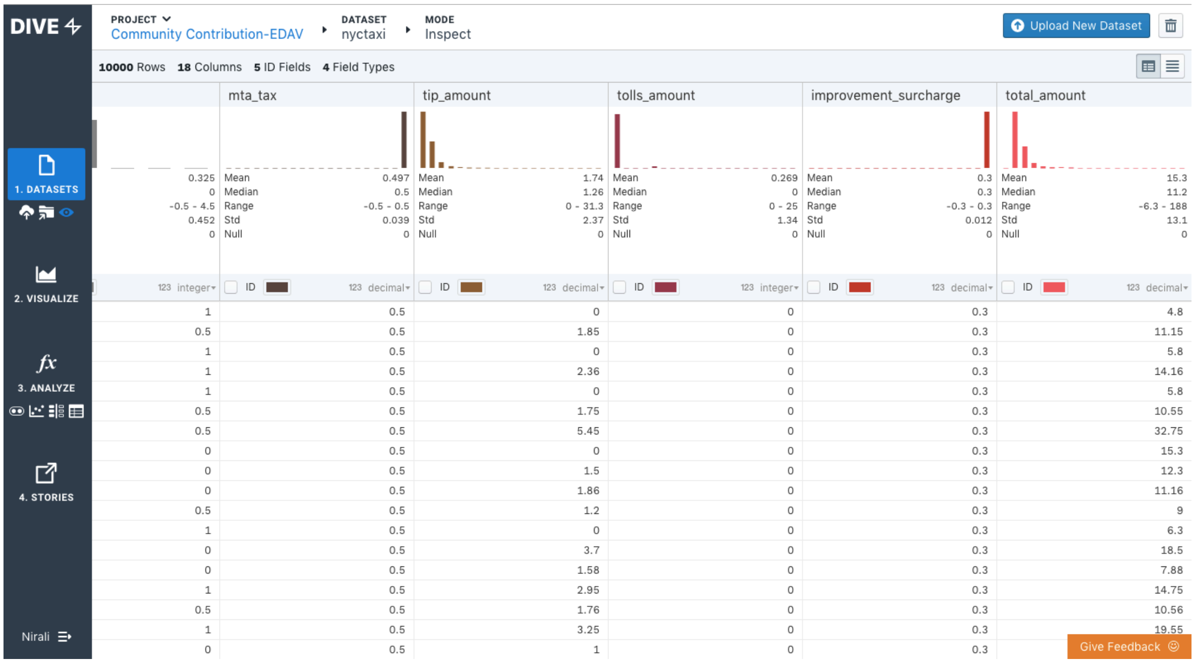
VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID
ID Mean 1.54 Median 2 Range 1 - 2 Std 0.499 Null 0	Range 2017-01-01 00:10:	Range 2017-01-01 00:27:	Mean 1.61 Median 1 Range 0 - 8 Std 1.26 Null 0	Mean 2.8 Median 1.58 Range 0 - 51.4 Std 3.64 Null 0	ID Mean Median Range Std Null
☑ ID 123 integer	☐ ID datetime	☐ ID datetime	☐ ID 123 integer	☐ ID 123 decimal	☑ ID
2	2017-01-09 17:10:00	2017-01-09 17:11:00	5	0.38	
1	2017-01-31 22:55:00	2017-01-31 23:03:00	1	1.6	
1	2017-01-24 19:44:00	2017-01-24 19:44:00	1	0	
2	2017-01-18 19:04:00	2017-01-18 19:18:00	5	1.24	
1	2017-01-03 16:11:00	2017-01-03 16:14:00	1	0.5	
1	2017-01-30 00:06:00	2017-01-30 00:14:00	1	1.4	
1	2017-01-17 19:58:00	2017-01-17 20:20:00	2	8.2	
2	2017-01-17 11:48:00	2017-01-17 12:09:00	1	1.94	
1	2017-01-08 16:25:00	2017-01-08 16:37:00	2	2	
2	2017-01-01 19:33:00	2017-01-01 19:43:00	1	1.56	
2	2017-01-26 20:10:00	2017-01-26 20:16:00	1	1.28	
2	2017-01-20 16:31:00	2017-01-20 16:34:00	1	0.73	
2	2017-01-26 11:54:00	2017-01-26 12:16:00	1	2.12	
2	2017-01-08 16:53:00	2017-01-08 16:58:00	1	0.79	
2	2017-01-31 18:10:00	2017-01-31 18:23:00	1	1.82	
2	2017-01-12 22:52:00	2017-01-12 23:00:00	2	1.37	
1	2017-01-20 18:15:00	2017-01-20 18:37:00	1	2.2	
2	2017-01-28 16:10:00	2017-01-28 16:20:00	1	1.21	



PROJECT Community Contribution-EDAV ▸ DATASET nyctaxi ▸ MODE Inspect

10000 Rows **18** Columns **5** ID Fields **4** Field Types

store_and_fwd_flag	PULocationID	DOLocationID	payment_type	fare_amount	extra
Unique Values 2 (0.02%) Most Frequent N Most Occurrences 9960 Null 0	Mean 165 Median 162 Range 1 - 265 Std 67.1 Null 0	Mean 163 Median 162 Range 1 - 265 Std 70.5 Null 0	Mean 1.34 Median 1 Range 1 - 4 Std 0.493 Null 0	Mean 12.2 Median 9 Range -5 - 150 Std 10.5 Null 0	Mean Median Range Std Null
□ ID 🟩 A string▾	☑ ID 🟩 123 integer▾	☑ ID 🟩 123 integer▾	□ ID ■ 123 integer▾	□ ID ■ 123 decimal▾	□ ID ■
N	229	141	2	3	
N	137	114	1	8	
Y	137	264	2	4	
N	137	229	1	10	
N	79	113	2	4	
N	79	249	1	7.5	
N	132	122	1	26	
N	151	142	2	14.5	
N	48	239	1	10	
N	186	48	1	8.5	
N	161	236	1	6.5	
N	43	230	2	4.5	
N	113	161	1	14	
N	48	48	1	5.5	
N	229	263	1	10	
N	164	161	1	7.5	
N	48	107	1	14.5	
N	226	260	1	8	



PROJECT Community Contribution-EDAV ▸ DATASET nyctaxi ▸ MODE Inspect

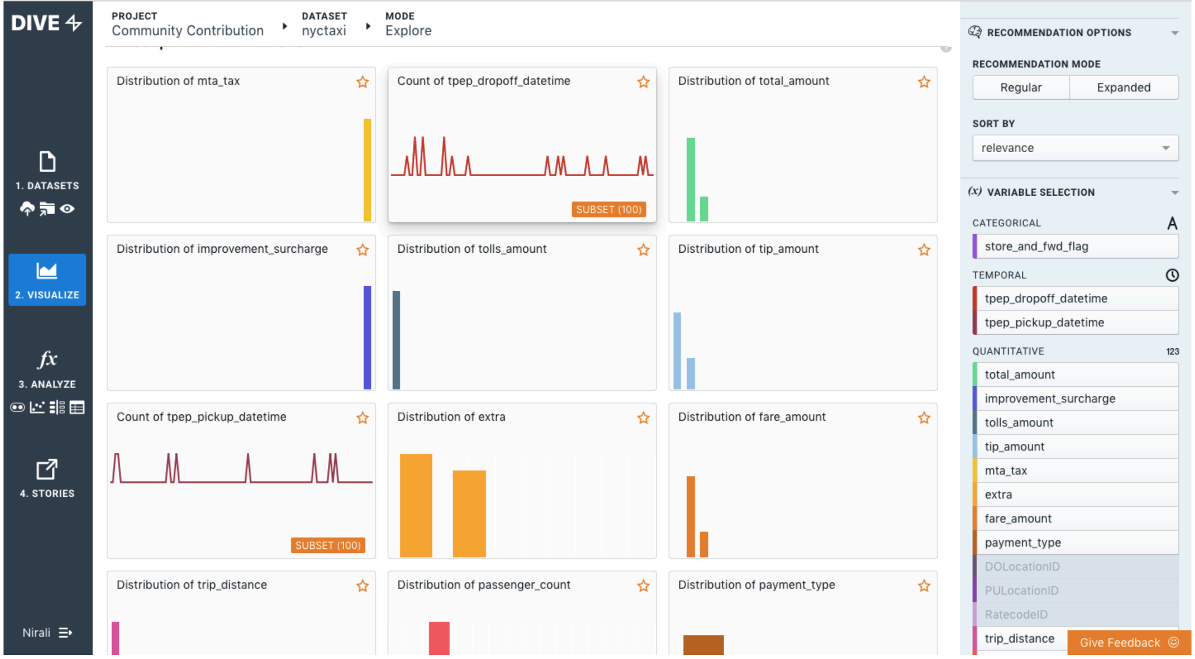
10000 Rows **18** Columns **5** ID Fields **4** Field Types

	mta_tax	tip_amount	tolls_amount	improvement_surcharge	total_amount
0.325 0 -0.5 - 4.5 0.452 0	Mean 0.497 Median 0.5 Range -0.5 - 0.5 Std 0.039 Null 0	Mean 1.74 Median 1.26 Range 0 - 31.3 Std 2.37 Null 0	Mean 0.269 Median 0 Range 0 - 25 Std 1.34 Null 0	Mean 0.3 Median 0.3 Range -0.3 - 0.3 Std 0.012 Null 0	Mean 15.3 Median 11.2 Range -6.3 - 188 Std 13.1 Null 0
123 integer▾	□ ID ■ 123 decimal▾	□ ID ■ 123 decimal▾	□ ID ■ 123 integer▾	□ ID 🟥 123 decimal▾	□ ID 🟥 123 decimal▾
1	0.5	0	0	0.3	4.8
0.5	0.5	1.85	0	0.3	11.15
1	0.5	0	0	0.3	5.8
1	0.5	2.36	0	0.3	14.16
1	0.5	0	0	0.3	5.8
0.5	0.5	1.75	0	0.3	10.55
0.5	0.5	5.45	0	0.3	32.75
0	0.5	0	0	0.3	15.3
0	0.5	1.5	0	0.3	12.3
0	0.5	1.86	0	0.3	11.16
0.5	0.5	1.2	0	0.3	9
1	0.5	0	0	0.3	6.3
0	0.5	3.7	0	0.3	18.5
0	0.5	1.58	0	0.3	7.88
1	0.5	2.95	0	0.3	14.75
0.5	0.5	1.76	0	0.3	10.56
1	0.5	3.25	0	0.3	19.55
0	0.5	1	0	0.3	

2. Descriptive Visualizations:

This section showcases univariate summary visualizations. To add a visualization to a story, the graph must be saved.

Visualizations of every variable



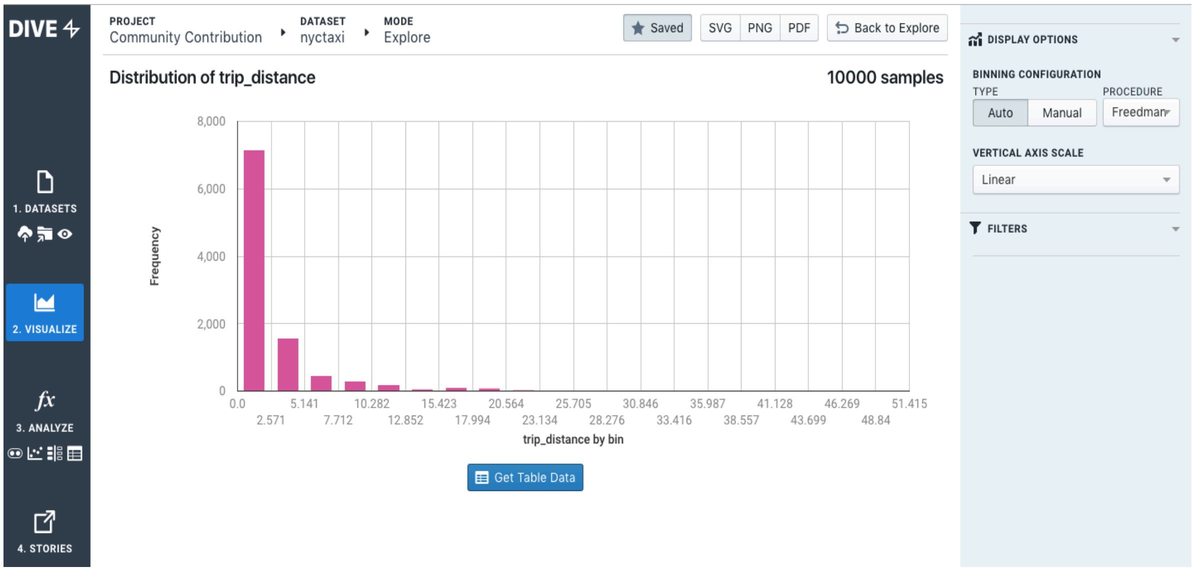
PROJECT Community Contribution ▸ DATASET nyctaxi ▸ MODE Explore

We are interested in the following graphs:
1. trip_distance
2. total_amount
3. tip_amount

The following display options are available:
1. Binning configuration:
 Type- Use automated number of bins or manually set the number of bins
Procedure- Select a procedure from a given list of procedures to create a visualization.
2. Vertical Axis Scale- it can be set to either Linear or Logarithmic.
3. Filter- Filters the data using multiple conditions. These conditions are based on values of different fields. Multiple conditions can be combined using OR or AND.

trip_distance

Default distribution of trip_distance


PROJECT Community Contribution
DATASET nyctaxi
MODE Explore

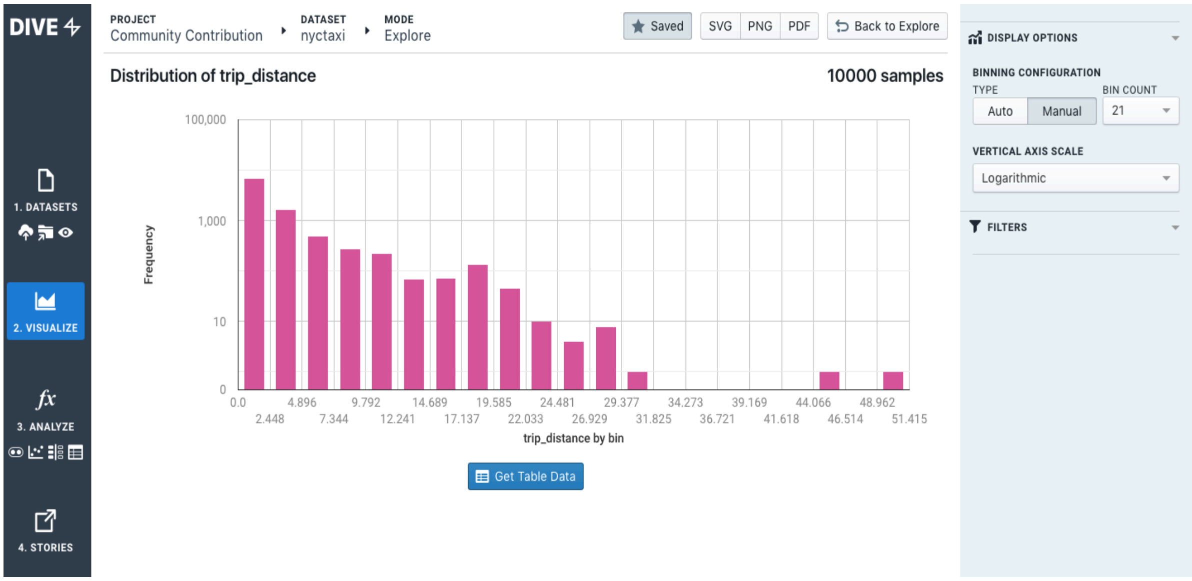
Distribution of trip_distance

10000 samples

BINNING CONFIGURATION
TYPE: Auto
PROCEDURE: Freedman

VERTICAL AXIS SCALE: Linear

Distribution of trip_distance with number of bins=21 and a logarithmic vertical scale


PROJECT Community Contribution
DATASET nyctaxi
MODE Explore

Distribution of trip_distance

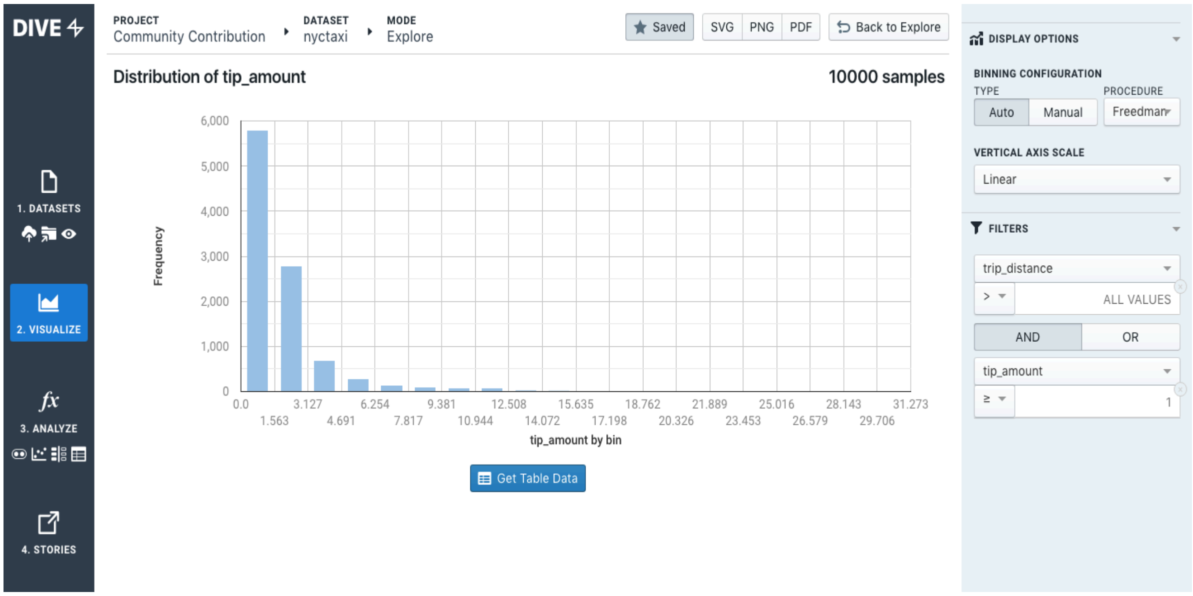
10000 samples

BINNING CONFIGURATION
TYPE: Manual
BIN COUNT: 21

VERTICAL AXIS SCALE: Logarithmic

tip_amount

Default distribution of tip_amount



PROJECT Community Contribution ▸ DATASET nyctaxi ▸ MODE Explore

Distribution of tip_amount

10000 samples

BINNING CONFIGURATION

TYPE: Auto

PROCEDURE: Freedman

VERTICAL AXIS SCALE: Linear

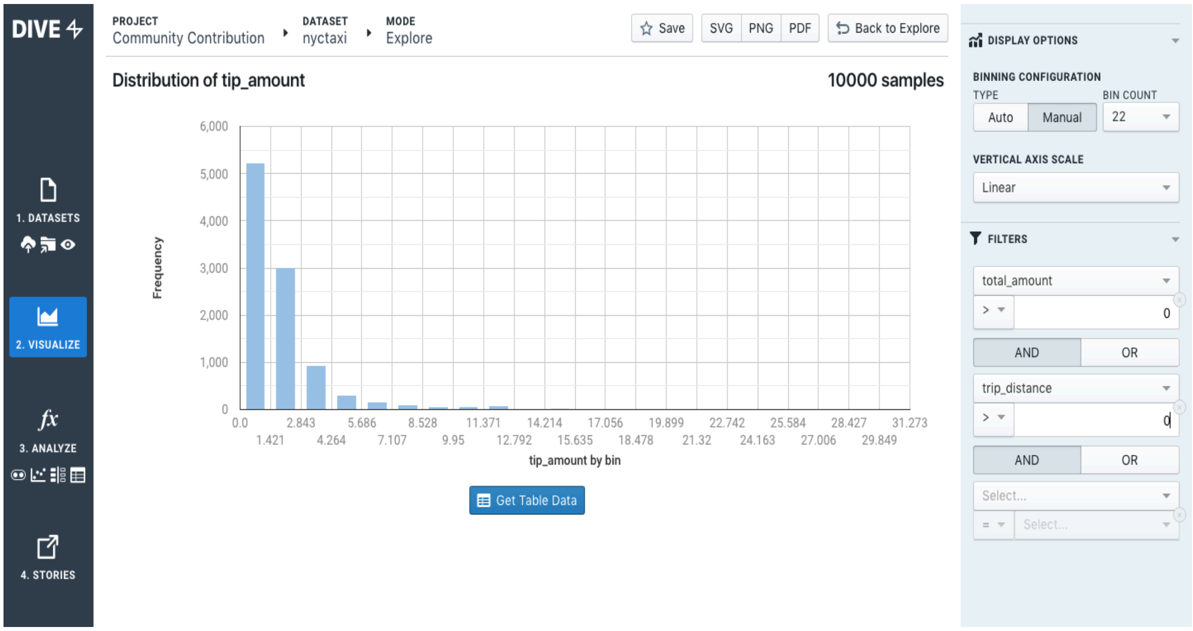
FILTERS

trip_distance > ALL VALUES

AND

tip_amount ≥ 1

Distribution of tip_amount with bins=22 and trip_distance>0 and total_amount>0



PROJECT Community Contribution ▸ DATASET nyctaxi ▸ MODE Explore

Distribution of tip_amount

10000 samples

BINNING CONFIGURATION

TYPE: Manual

BIN COUNT: 22

VERTICAL AXIS SCALE: Linear

FILTERS

total_amount > 0

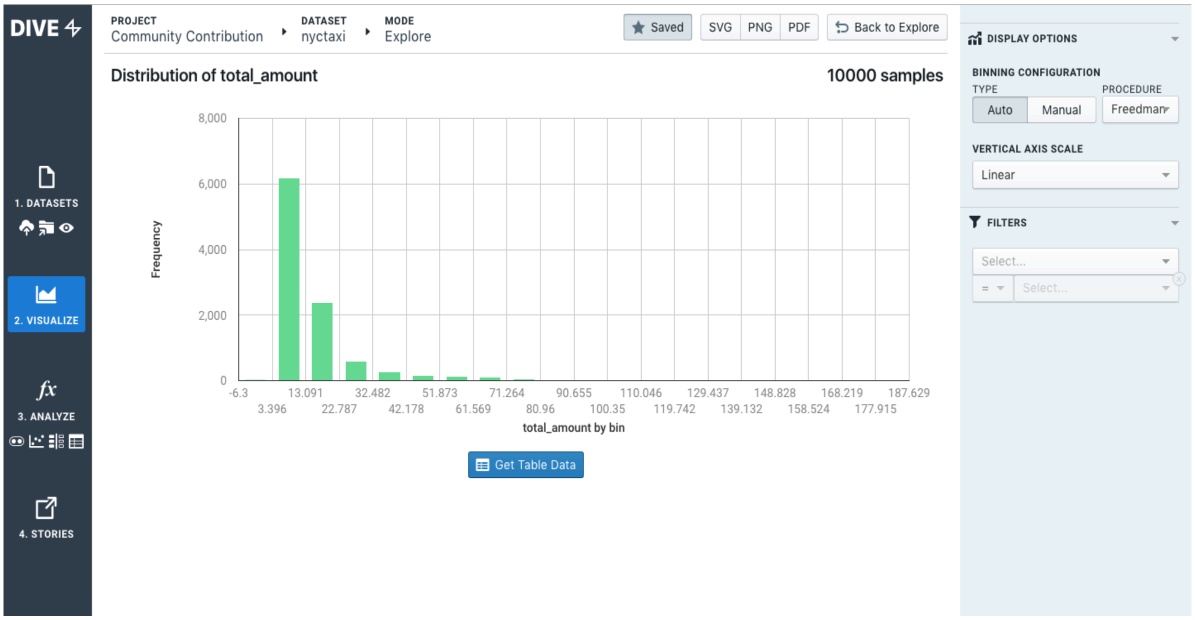
AND

trip_distance > 0

AND

total_amount

Default distribution of total_amount



PROJECT Community Contribution ▸ DATASET nyctaxi ▸ MODE Explore

Distribution of total_amount

10000 samples

BINNING CONFIGURATION

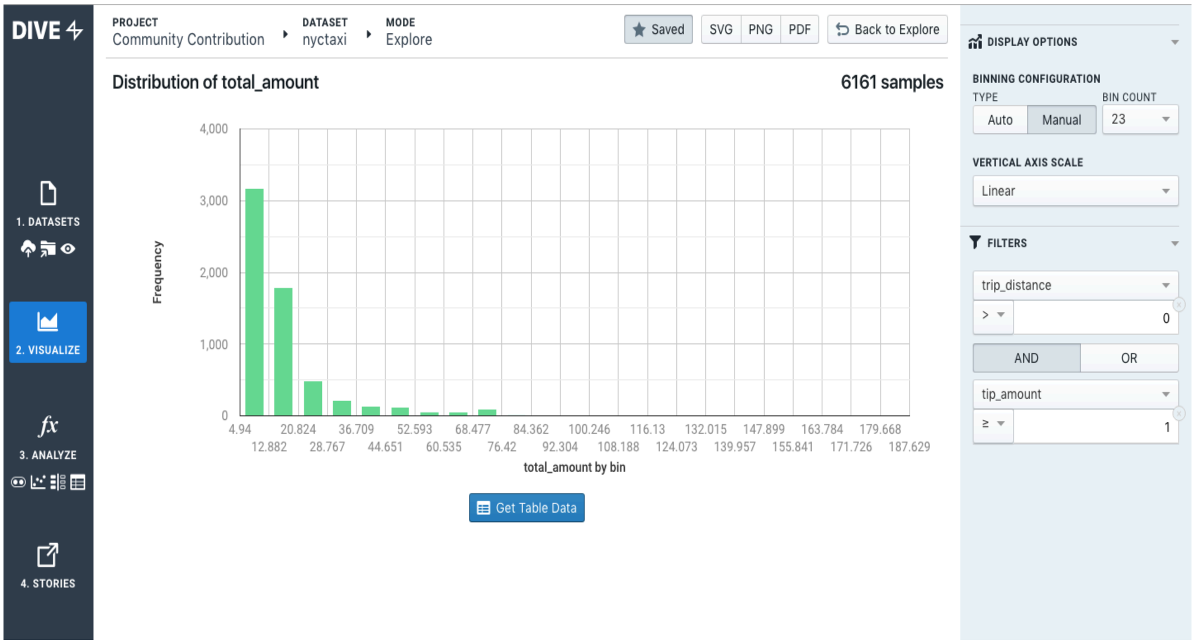
TYPE: Auto

PROCEDURE: Freedman

VERTICAL AXIS SCALE: Linear

FILTERS: Select... = Select...

Distribution of total_amount with bins=23 and trip_distance>0 and tip_amount ≥ 1



PROJECT Community Contribution ▸ DATASET nyctaxi ▸ MODE Explore

Distribution of total_amount

6161 samples

BINNING CONFIGURATION

TYPE: Manual

BIN COUNT: 23

VERTICAL AXIS SCALE: Linear

FILTERS

trip_distance > 0

AND

tip_amount ≥ 1

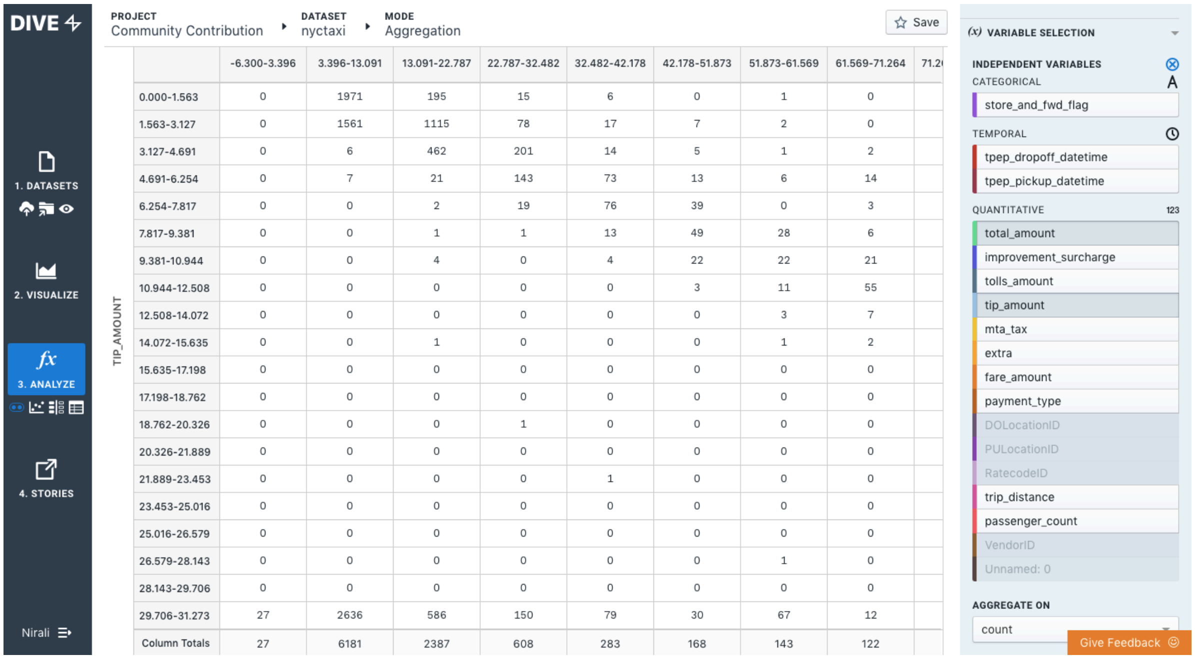
3. Analyze mode:

This mode gives 4 options:
a) Aggregation
b) Correlation
c) Comparison
d) Regression

a) Aggregation

This mode helps aggregate the data based on the variables selected.
On selecting two variables, the values of the variables are divided into different bins and a count for combinations of bins of the two variables is displayed.
This aggregation can be performed on any other variable of the dataset.
Example: aggregating total_amount and tip_amount on trip_distance.
One can choose to display either the sum of trip_distance or the mean of trip_distance for every combination of bins of total_amount and tip_amount.
The binning of the variables chosen can also be changed and filters can be applied just as explained previously.

Aggregating total_amount and tip_amount on Count

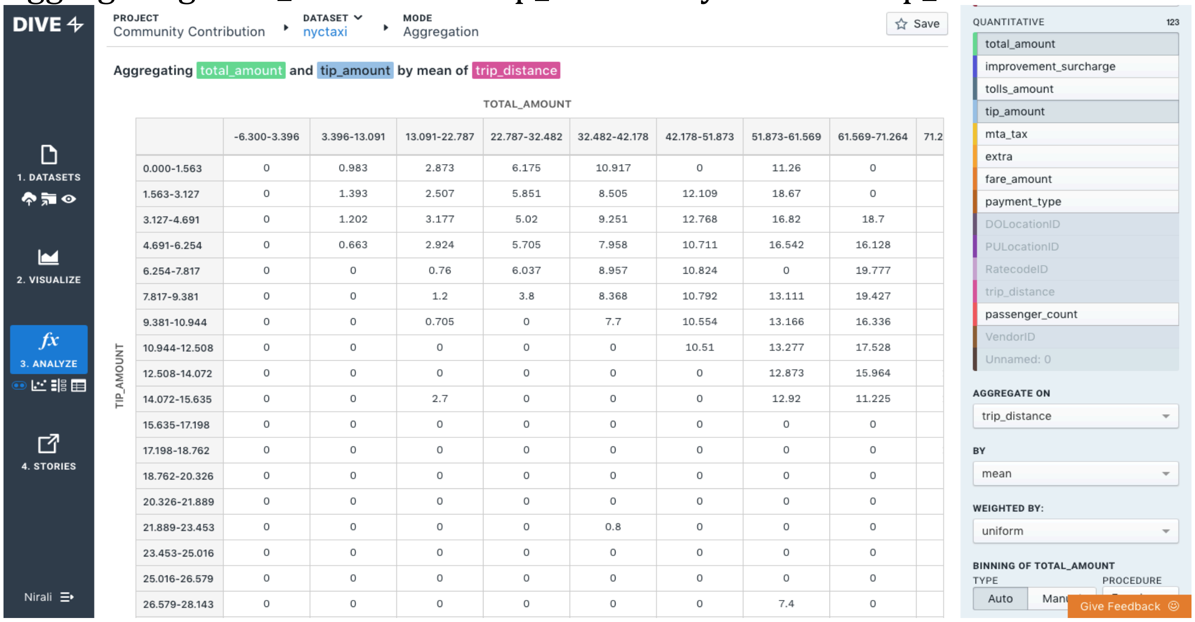


PROJECT Community Contribution ▸ DATASET nyctaxi ▸ MODE Aggregation

	-6.300-3.396	3.396-13.091	13.091-22.787	22.787-32.482	32.482-42.178	42.178-51.873	51.873-61.569	61.569-71.264	71.2
0.000-1.563	0	1971	195	15	6	0	1	0	
1.563-3.127	0	1561	1115	78	17	7	2	0	
3.127-4.691	0	6	462	201	14	5	1	2	
4.691-6.254	0	7	21	143	73	13	6	14	
6.254-7.817	0	0	2	19	76	39	0	3	
7.817-9.381	0	0	1	1	13	49	28	6	
9.381-10.944	0	0	4	0	4	22	22	21	
10.944-12.508	0	0	0	0	0	3	11	55	
12.508-14.072	0	0	0	0	0	0	3	7	
14.072-15.635	0	0	1	0	0	0	1	2	
15.635-17.198	0	0	0	0	0	0	0	0	
17.198-18.762	0	0	0	0	0	0	0	0	
18.762-20.326	0	0	0	1	0	0	0	0	
20.326-21.889	0	0	0	0	0	0	0	0	
21.889-23.453	0	0	0	0	1	0	0	0	
23.453-25.016	0	0	0	0	0	0	0	0	
25.016-26.579	0	0	0	0	0	0	0	0	
26.579-28.143	0	0	0	0	0	0	1	0	
28.143-29.706	0	0	0	0	0	0	0	0	
29.706-31.273	27	2636	586	150	79	30	67	12	
Column Totals	27	6181	2387	608	283	168	143	122	

AGGREGATE ON
count

Aggregating total_amount and tip_amount by mean of trip_distance



PROJECT Community Contribution ▸ DATASET nyctaxi ▸ MODE Aggregation

Aggregating total_amount and tip_amount by mean of trip_distance

TOTAL_AMOUNT

TIP_AMOUNT	-6.300-3.396	3.396-13.091	13.091-22.787	22.787-32.482	32.482-42.178	42.178-51.873	51.873-61.569	61.569-71.264	71.2
0.000-1.563	0	0.983	2.873	6.175	10.917	0	11.26	0	
1.563-3.127	0	1.393	2.507	5.851	8.505	12.109	18.67	0	
3.127-4.691	0	1.202	3.177	5.02	9.251	12.768	16.82	18.7	
4.691-6.254	0	0.663	2.924	5.705	7.958	10.711	16.542	16.128	
6.254-7.817	0	0	0.76	6.037	8.957	10.824	0	19.777	
7.817-9.381	0	0	1.2	3.8	8.368	10.792	13.111	19.427	
9.381-10.944	0	0	0.705	0	7.7	10.554	13.166	16.336	
10.944-12.508	0	0	0	0	0	10.51	13.277	17.528	
12.508-14.072	0	0	0	0	0	0	12.873	15.964	
14.072-15.635	0	0	2.7	0	0	0	12.92	11.225	
15.635-17.198	0	0	0	0	0	0	0	0	
17.198-18.762	0	0	0	0	0	0	0	0	
18.762-20.326	0	0	0	0	0	0	0	0	
20.326-21.889	0	0	0	0	0	0	0	0	
21.889-23.453	0	0	0	0	0.8	0	0	0	
23.453-25.016	0	0	0	0	0	0	0	0	
25.016-26.579	0	0	0	0	0	0	0	0	
26.579-28.143	0	0	0	0	0	0	7.4	0	

QUANTITATIVE 123

total_amount
improvement_surcharge
tolls_amount
tip_amount
mta_tax
extra
fare_amount
payment_type
DOLocationID
PULocationID
RatecodeID
trip_distance
passenger_count
VendorID
Unnamed: 0

AGGREGATE ON: trip_distance

BY: mean

WEIGHTED BY: uniform

BINNING OF TOTAL_AMOUNT

TYPE: Auto | Manu

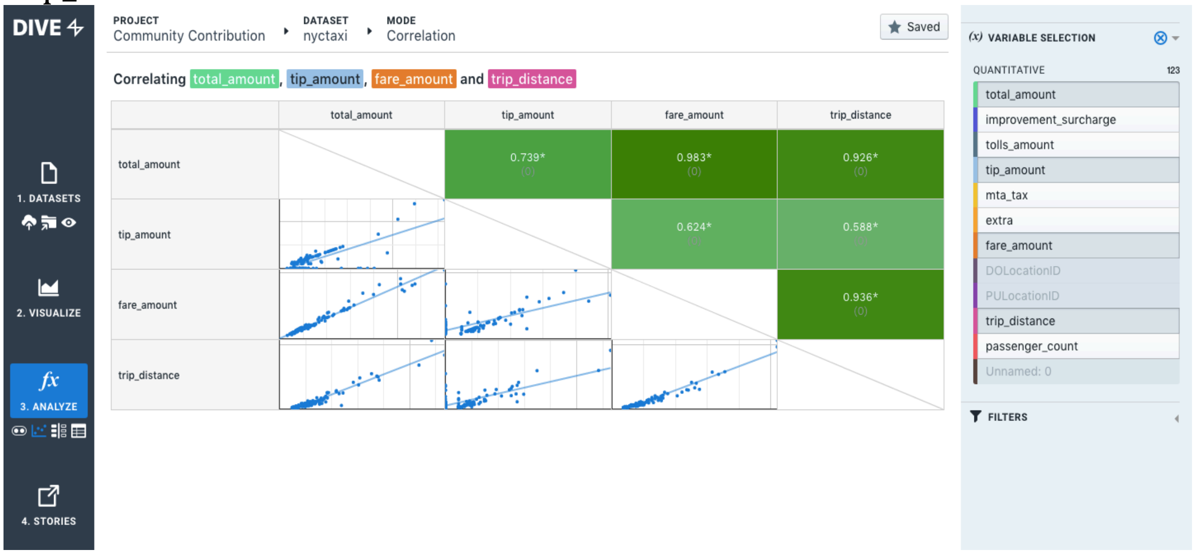
PROCEDURE

b) Correlation

This mode helps find the relationship between the variables in a dataset. It displays a correlation matrix that displays the correlation between the variables to indicate if the variables are positively correlated, negatively correlated or not correlated.

We just need to select the variables in the right pane and a correlation matrix will be generated. Filters can also be applied.

Correlation between total_amount, tip_amount, fare_amount and trip_distance



PROJECT Community Contribution ▸ DATASET nyctaxi ▸ MODE Correlation

Correlating total_amount, tip_amount, fare_amount and trip_distance

	total_amount	tip_amount	fare_amount	trip_distance
total_amount		0.739* (0)	0.983* (0)	0.926* (0)
tip_amount			0.624* (0)	0.588* (0)
fare_amount				0.936* (0)
trip_distance				

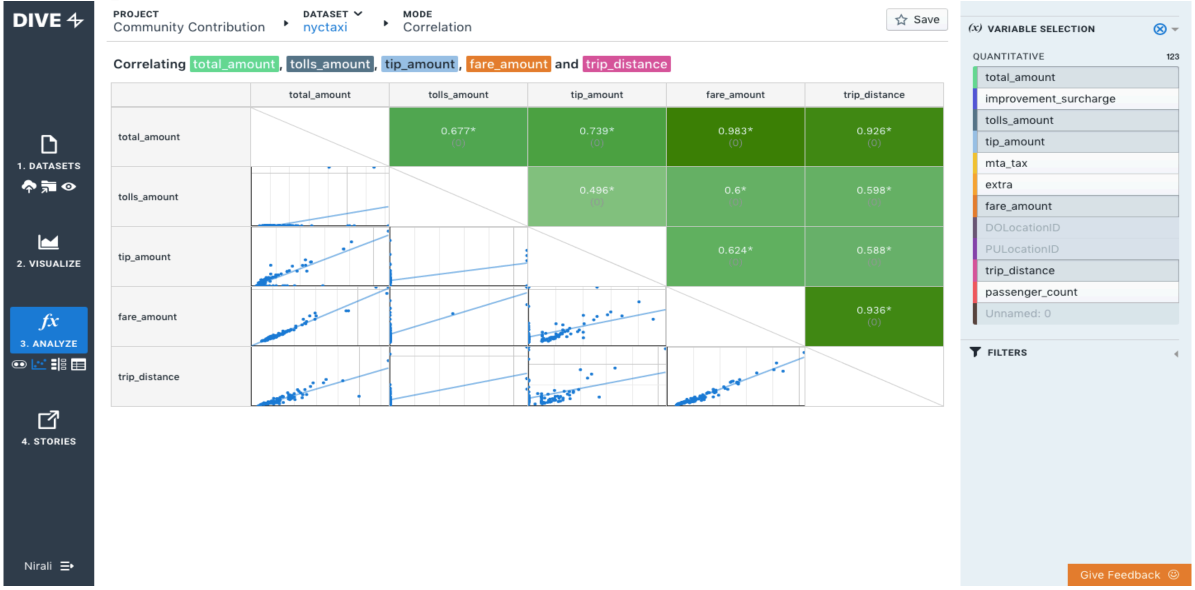
(x) VARIABLE SELECTION

QUANTITATIVE 123

total_amount
improvement_surcharge
tolls_amount
tip_amount
mta_tax
extra
fare_amount
DOLocationID
PULocationID
trip_distance
passenger_count
Unnamed: 0

FILTERS

Correlation between total_amount, tip_amount, tolls_amount, fare_amount and trip_distance



Correlating total_amount, tolls_amount, tip_amount, fare_amount and trip_distance

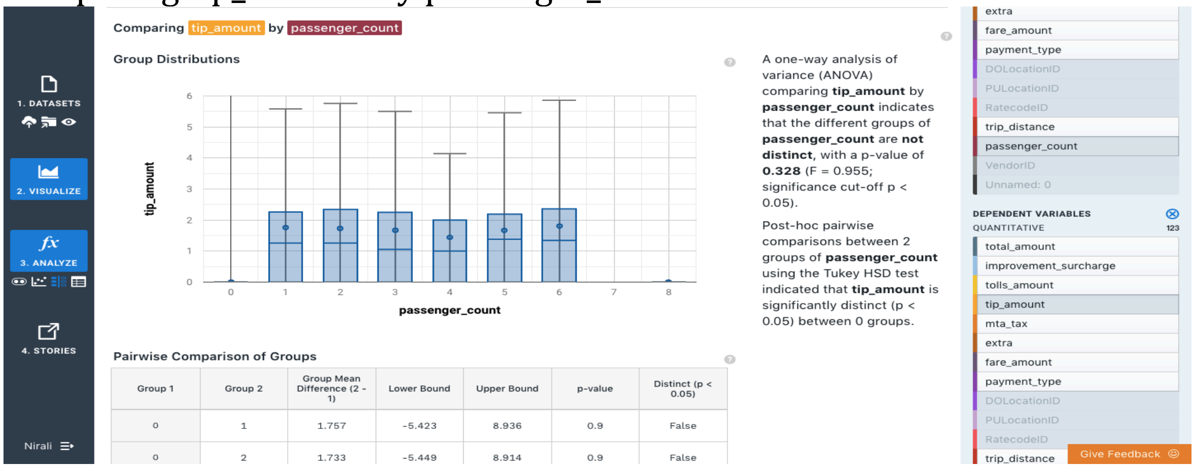
	total_amount	tolls_amount	tip_amount	fare_amount	trip_distance
total_amount		0.677* (0)	0.739* (0)	0.983* (0)	0.926* (0)
tolls_amount			0.496* (0)	0.6* (0)	0.598* (0)
tip_amount				0.624* (0)	0.588* (0)
fare_amount					0.936* (0)
trip_distance					

We see a positive correlation of 0.739 between total_amount and tip_amount.

c) Comparison

This mode performs a one-way analysis of variance (ANOVA) comparing the values of the dependent variable for different groups of the independent variable. It tells us if the groups of the independent varaible are distinct with a p-value. It also performs Post-hoc pairwise comparison between 2 groups of the dependent variable using the Tukey HSD test if those groups are distinct.

Comparing tip_amount by passenger_count



Comparing tip_amount by passenger_count

A one-way analysis of variance (ANOVA) comparing **tip_amount** by **passenger_count** indicates that the different groups of **passenger_count** are **not distinct**, with a p-value of **0.328** (F = 0.955; significance cut-off p < 0.05).

Post-hoc pairwise comparisons between 2 groups of **passenger_count** using the Tukey HSD test indicated that **tip_amount** is significantly distinct (p < 0.05) between 0 groups.

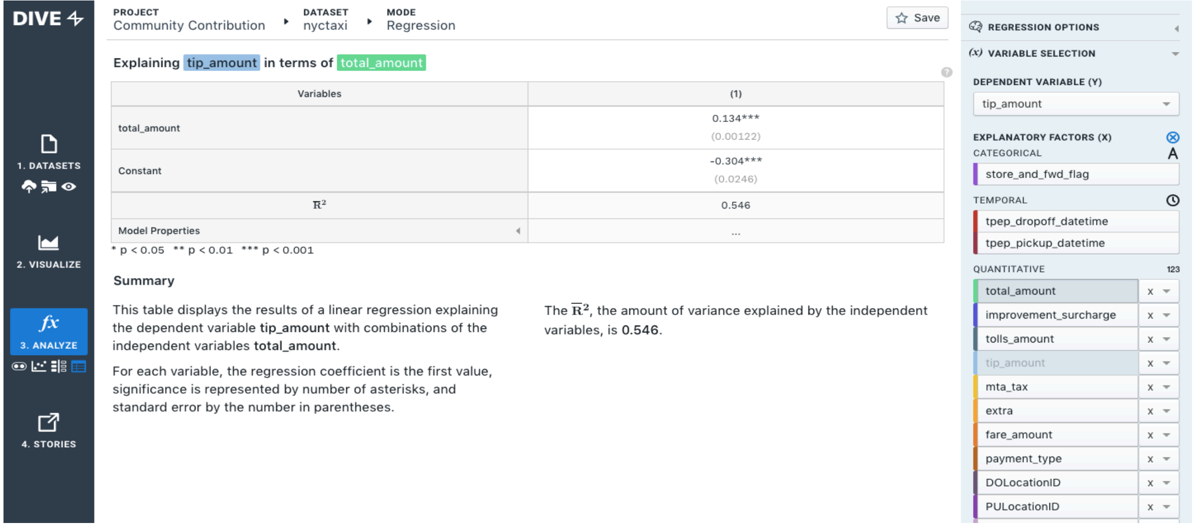
Pairwise Comparison of Groups

Group 1	Group 2	Group Mean Difference (2 - 1)	Lower Bound	Upper Bound	p-value	Distinct (p < 0.05)
0	1	1.757	-5.423	8.936	0.9	False
0	2	1.733	-5.449	8.914	0.9	False

d) Regression

Here, we need to select the dependent variable. By clicking on 'Recommend Model', it returns a set of models with regression coefficients and constants along with the contribution of the independent variables to the value of the dependent variable. We can also manually select a list of independent variables of our choice along with the degree (x or x^2) to test their contribution to the dependent variable. It also displays $\bar{R}^2$, which is the amount of variance explained by independent variables.

tip_amount in terms of total_amount



PROJECT Community Contribution ▸ DATASET nyctaxi ▸ MODE Regression

Explaining tip_amount in terms of total_amount

Variables	(1)
total_amount	0.134*** (0.00122)
Constant	-0.304*** (0.0246)
$\bar{R}^2$	0.546
Model Properties	...

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Summary

This table displays the results of a linear regression explaining the dependent variable **tip_amount** with combinations of the independent variables **total_amount**.

For each variable, the regression coefficient is the first value, significance is represented by number of asterisks, and standard error by the number in parentheses.

The $\bar{R}^2$, the amount of variance explained by the independent variables, is **0.546**.

DEPENDENT VARIABLE (Y): tip_amount

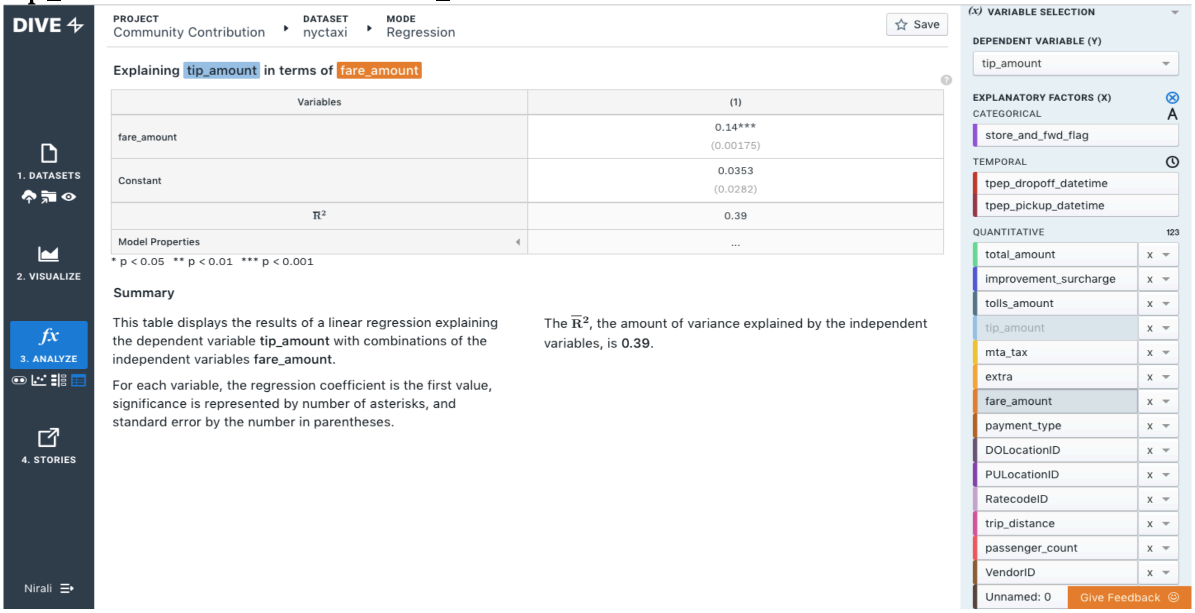
EXPLANATORY FACTORS (X)

CATEGORICAL: store_and_fwd_flag

TEMPORAL: tpep_dropoff_datetime, tpep_pickup_datetime

QUANTITATIVE: total_amount, improvement_surcharge, tolls_amount, tip_amount, mta_tax, extra, fare_amount, payment_type, DOLocationID, PULocationID

tip_amount in terms of fare_amount



PROJECT Community Contribution ▸ DATASET nyctaxi ▸ MODE Regression

Explaining tip_amount in terms of fare_amount

Variables	(1)
fare_amount	0.14*** (0.00175)
Constant	0.0353 (0.0282)
$\bar{R}^2$	0.39
Model Properties	...

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Summary

This table displays the results of a linear regression explaining the dependent variable **tip_amount** with combinations of the independent variables **fare_amount**.

For each variable, the regression coefficient is the first value, significance is represented by number of asterisks, and standard error by the number in parentheses.

The $\bar{R}^2$, the amount of variance explained by the independent variables, is **0.39**.

DEPENDENT VARIABLE (Y): tip_amount

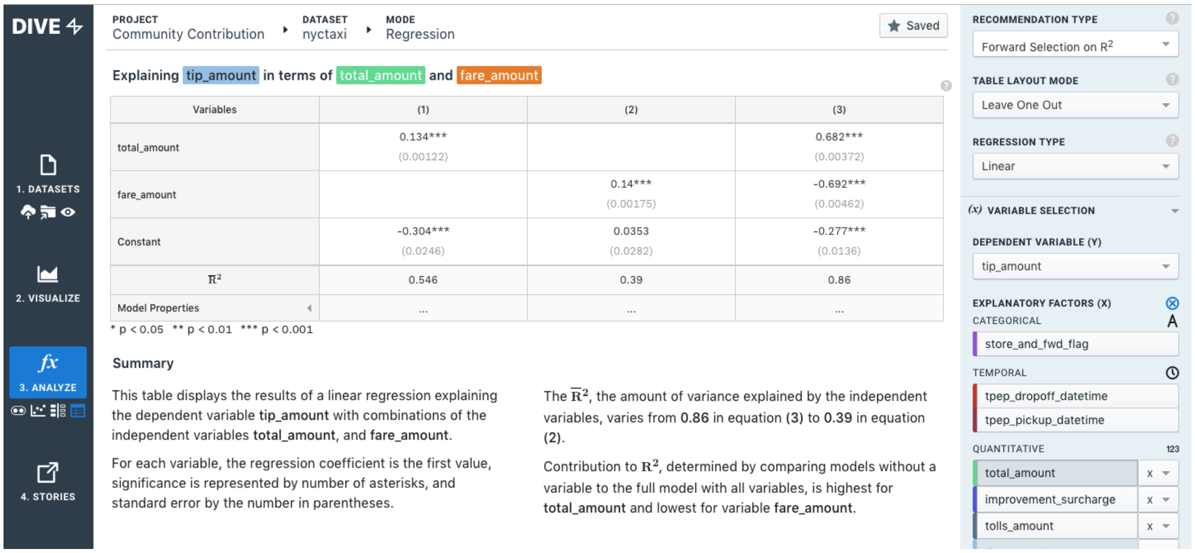
EXPLANATORY FACTORS (X)

CATEGORICAL: store_and_fwd_flag

TEMPORAL: tpep_dropoff_datetime, tpep_pickup_datetime

QUANTITATIVE: total_amount, improvement_surcharge, tolls_amount, tip_amount, mta_tax, extra, fare_amount, payment_type, DOLocationID, PULocationID, RatecodeID, trip_distance, passenger_count, VendorID, Unnamed: 0

tip_amount in terms of total_amount and fare_amount
This is the recommended linear regression model with regression type as 'Forward Selection on R^2 '



PROJECT Community Contribution ▸ DATASET nyctaxi ▸ MODE Regression

Explaining tip_amount in terms of total_amount and fare_amount

Variables	(1)	(2)	(3)
total_amount	0.134*** (0.00122)		0.682*** (0.00372)
fare_amount		0.14*** (0.00175)	-0.692*** (0.00462)
Constant	-0.304*** (0.0246)	0.0353 (0.0282)	-0.277*** (0.0136)
$\overline{R}^2$	0.546	0.39	0.86
Model Properties	...	...	...

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Summary

This table displays the results of a linear regression explaining the dependent variable **tip_amount** with combinations of the independent variables **total_amount**, and **fare_amount**.

For each variable, the regression coefficient is the first value, significance is represented by number of asterisks, and standard error by the number in parentheses.

The $\overline{R}^2$, the amount of variance explained by the independent variables, varies from **0.86** in equation **(3)** to **0.39** in equation **(2)**.

Contribution to R^2, determined by comparing models without a variable to the full model with all variables, is highest for **total_amount** and lowest for variable **fare_amount**.

RECOMMENDATION TYPE: Forward Selection on R^2

TABLE LAYOUT MODE: Leave One Out

REGRESSION TYPE: Linear

(x) VARIABLE SELECTION

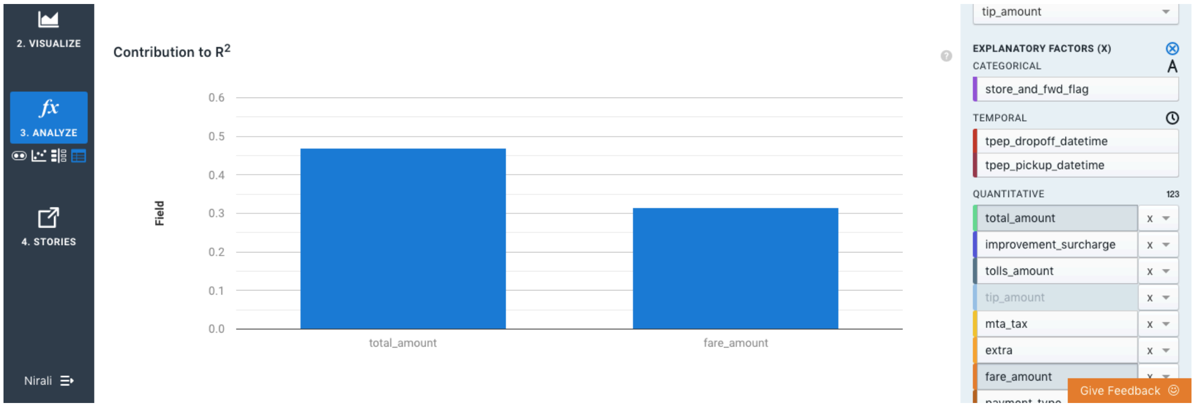
DEPENDENT VARIABLE (Y): tip_amount

EXPLANATORY FACTORS (X)

CATEGORICAL: store_and_fwd_flag

TEMPORAL: tpep_dropoff_datetime, tpep_pickup_datetime

QUANTITATIVE: total_amount, improvement_surcharge, tolls_amount



DEPENDENT VARIABLE (Y): tip_amount

EXPLANATORY FACTORS (X)

CATEGORICAL: store_and_fwd_flag

TEMPORAL: tpep_dropoff_datetime, tpep_pickup_datetime

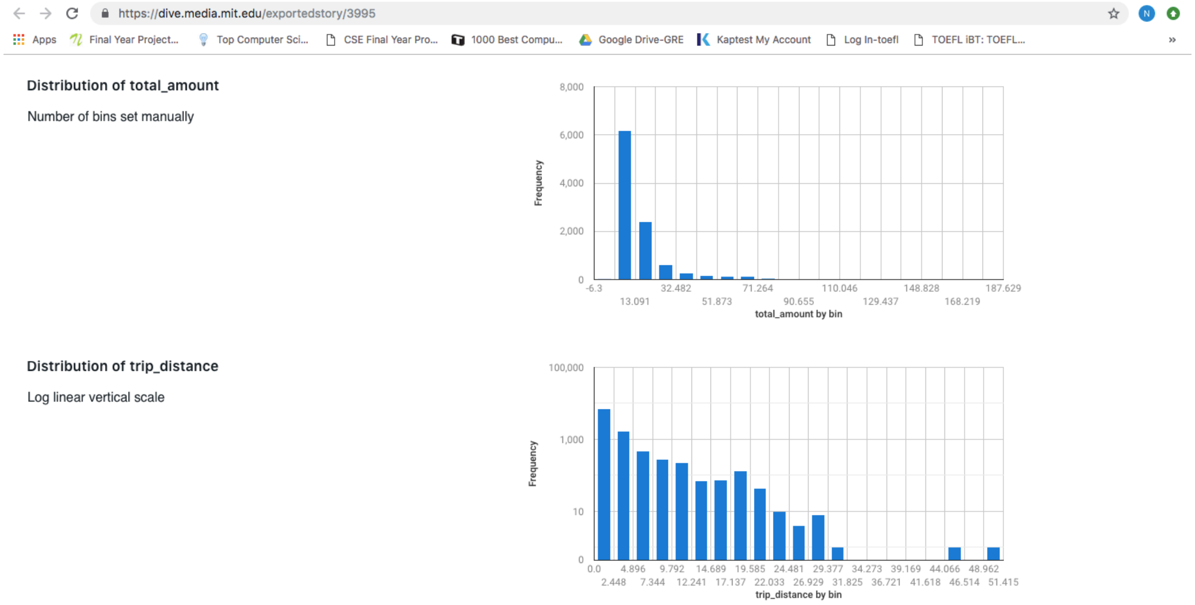
QUANTITATIVE: total_amount, improvement_surcharge, tolls_amount, tip_amount, mta_tax, extra, fare_amount, payment_type

We see that total_amount contributes more as compared to fare_amount.

4. Stories

In this mode, all the saved visualizations appear in the right pane.
To include a particular visualization in the story, we just need to click on it. The visualization appears in the story along with some space alloted for text, which can be used to explain the visualization.
In this way, an entire document can be created with all the necessary visualizations and this document can be shared.
Once you click the 'Share' option, it creates a web page for your document. This allows dynamic sharing of your work.

Document created.



https://dive.media.mit.edu/exportedstory/3995

Distribution of total_amount

Number of bins set manually

Distribution of trip_distance

Log linear vertical scale

References:

[1] Kevin Hu, Diana Orghian, and César Hidalgo. 2018. DIVE: A Mixed-Initiative System Supporting Integrated Data Exploration Workflows. In ACM SIGMOD Workshop on Human-In-the-Loop Data Analytics (HILDA), June 10, 2018, Houston, TX, USA. ACM, New York, NY, USA, Article 4, 7 pages.
https://doi.org/10.1145/3209900.3209910

[2] DIVE Tool videos
https://dive.media.mit.edu/

[3] DIVE demo video
https://www.youtube.com/watch?v=J3FceN2lYdA&t=17s