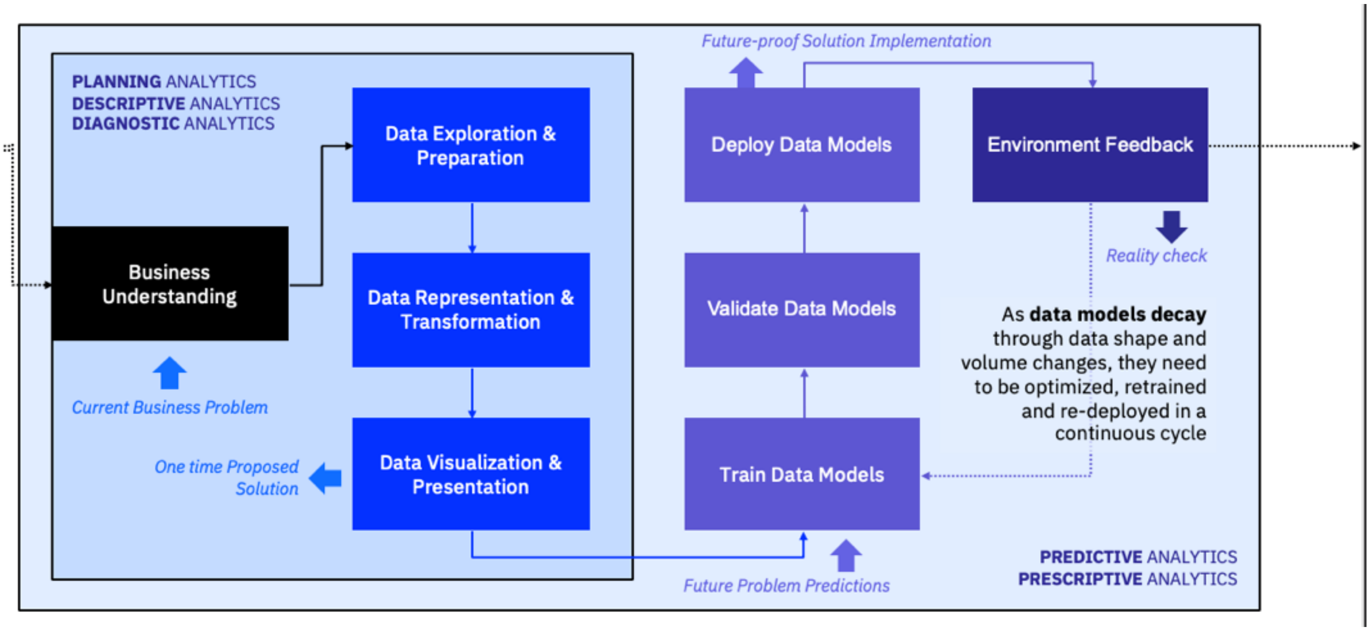
Data Wrangling Using Automotive Data Set

Section 1. Preface

Targeting New Opportunities

The Purpose of This Lab

The exercises in this lab, elaborate on data **exploration**, data **representation** and **visualization** approaches. Note that the visualization techniques used here illicit a preliminary outlook on the raw data and if further preparation or transformation is needed.



PLANNING ANALYTICS
DESCRIPTIVE ANALYTICS
DIAGNOSTIC ANALYTICS
Business Understanding
Current Business Problem
One time Proposed Solution
Data Exploration & Preparation
Data Representation & Transformation
Data Visualization & Presentation
Future-proof Solution Implementation
Deploy Data Models
Environment Feedback
Validate Data Models
Train Data Models
Reality check
As data models decay through data shape and volume changes, they need to be optimized, retrained and re-deployed in a continuous cycle
Future Problem Predictions
PREDICTIVE ANALYTICS
PRESCRIPTIVE ANALYTICS

Objective

This document comprises three labs. Each lab corresponds to a particular stage stemming from Business Understanding: data exploration and preparation, data representation and transformation, data visualization and presentation. In this lab you perform tasks by the Data Journalist and the Data Engineer as depicted in this list of objectives:

- Access IBM Cloud and provision the Watson Studio Service
- Import automobile data from open source communities
- Cleanse, analyze and reshape automotive data
- Visualize preliminary data wrangling results
- Run summary statistics on the results
- Validate automotive data

Car Manufacture Company

Headquartered in Detroit, Michigan, American automotive manufacturer, CAR MANUFACTURE COMPANY produces 1.6 million units annually and employs over 50,00 people across the continental United States.

During its fifty-year history, Car Manufacture Company has been known for its powerful vehicles, from sporty coupes to luxury sedans and diesel trucks.

Business challenge story

Already one of the world's largest makers of high horsepower vehicles, **Car Manufacture Company** realized in the mid-2010s that it would need to move beyond its mainline business to continue to grow.

As it watched car sales in Western Europe climb by almost 30% to all-time annual rate highs over the past five years, the company decided to focus its expansion in that area.

Additional analysis confirmed the demand, and in late 2017, Car Manufacture Company secured its first major order request, which would boost company profits nearly 20%.

However, recent legislation has imposed new fuel efficiency laws that apply to all new cars sold in the European Union. Manufacturers must meet these ever-tightening fuel economy rules or face bottom-line busting fines.

To capitalize on the new market opportunity and fulfill the order request, Car Manufacture Company needed to figure out how to transform their vehicle manufacturing operations, reinventing them for a future of fuel efficiency without sacrificing vehicle performance.

The first step in identifying a new manufacturing strategy is for Car Manufacture Company to benchmark company fuel efficiencies to the market. Car Manufacture Company purchased a data set from a research company.

The company then hired a data science team to analyze the data and to report back to leadership their findings on KPIs to build a car that meets EU fuel efficiency laws.

Overview

A. Get and prepare data in a project

After you create a project, or join one, the next step is to add data to the project and prepare the data for analysis. You can add data assets from your local system, from Watson Knowledge Catalog, from the IBM Watson Community, or from connections to data sources.

You can add these types of data assets to a project:

- Data assets from files from your local system, including structured data, unstructured data, and images. The files are stored in the project's IBM Cloud Object Storage bucket.
- Connection assets that contain information for connecting to data sources. You can add connections to IBM or third-party data sources.
- Connected data assets that specify a table, view, or file that is accessed through a connection to a data source.
- Folder assets that specify a path in IBM Cloud Object Storage.

You can see a preview of the contents of the data asset and a profile of the textual content of the data.

If you plan to refine data by cleansing and shaping it, first add the data to the project, then choose **Tools > Data Refinery**. The Data Engineering tool must be enabled in the **Tools** section of the project **Settings** page.

B. Refine Data

Refining data consists of cleansing and shaping it. When you cleanse data, you fix or remove data that is incorrect, incomplete, improperly formatted, or duplicated. And when you shape data, you customize it by filtering, sorting, combining or removing columns, and performing operations.

As you manipulate your data, you build a customized data flow that you can modify in real time and save for future re-use. When you save the refined data set, you typically load it to a different location than where you read it from. In this way, your source data can remain untouched by the refinement process.

- Prerequisites
- Refine your data
- Data set previews
- Data flows and steps

Prerequisites

In order to complete this lab, it is recommended you have a familiarity with statistics and a firm grasp on IT Basics.

Lab 2. Exploring and Preparing Automotive Data

You will use data about cars to graph the relationships between various properties, for example, how horsepower affects gas mileage. The cars data set was used for the 1983 American Statistical Association Data Exposition. This data set was collected by Ernesto Ramos and David Donoho and obtained from StatLib.

By now you have already registered with IBM Cloud and applied your promo code. Let’s begin our journey:

1. Login into IBM Cloud: https://console.bluemix.net/catalog/

   Note: Ensure that you have the promocode applied.
2. Click the **Catalog** tab and remove the label:lite filter.

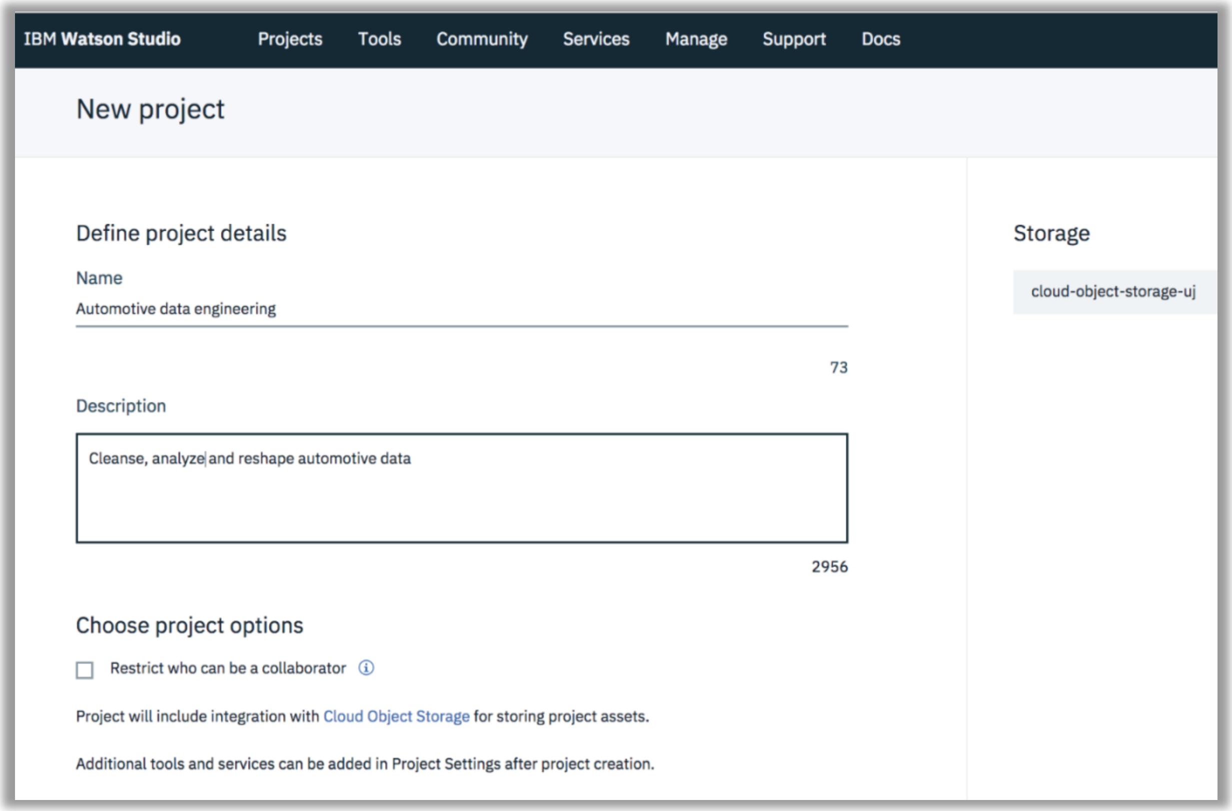
3. Search for the **Watson Studio** service and click that tile.
4. Click **Create**.
5. Click the **Get Started** button and when **Done**, click **Get Started** again.

Complete the following steps:

After your collection has been created, you can immediately start uploading content using the upload area at the right of the screen. However, before you add your own content to the Discovery service, best practice is to configure the service to process the content the way that you want.

1. Click **New project**.
2. Select the **Standard** tile.
3. Click **OK**.
4. Specify a name. In this example, it is Automotive data engineering.
5. Specify a description; for example, **Cleanse, analyze and reshape automotive data**.
6. Click **Create**.

7. Now you have a cloud object storage available to you, click **Create**



New project

Define project details

Name
Automotive data engineering

73

Description

Cleanse, analyze and reshape automotive data

2956

Choose project options

□ Restrict who can be a collaborator

Project will include integration with Cloud Object Storage for storing project assets.

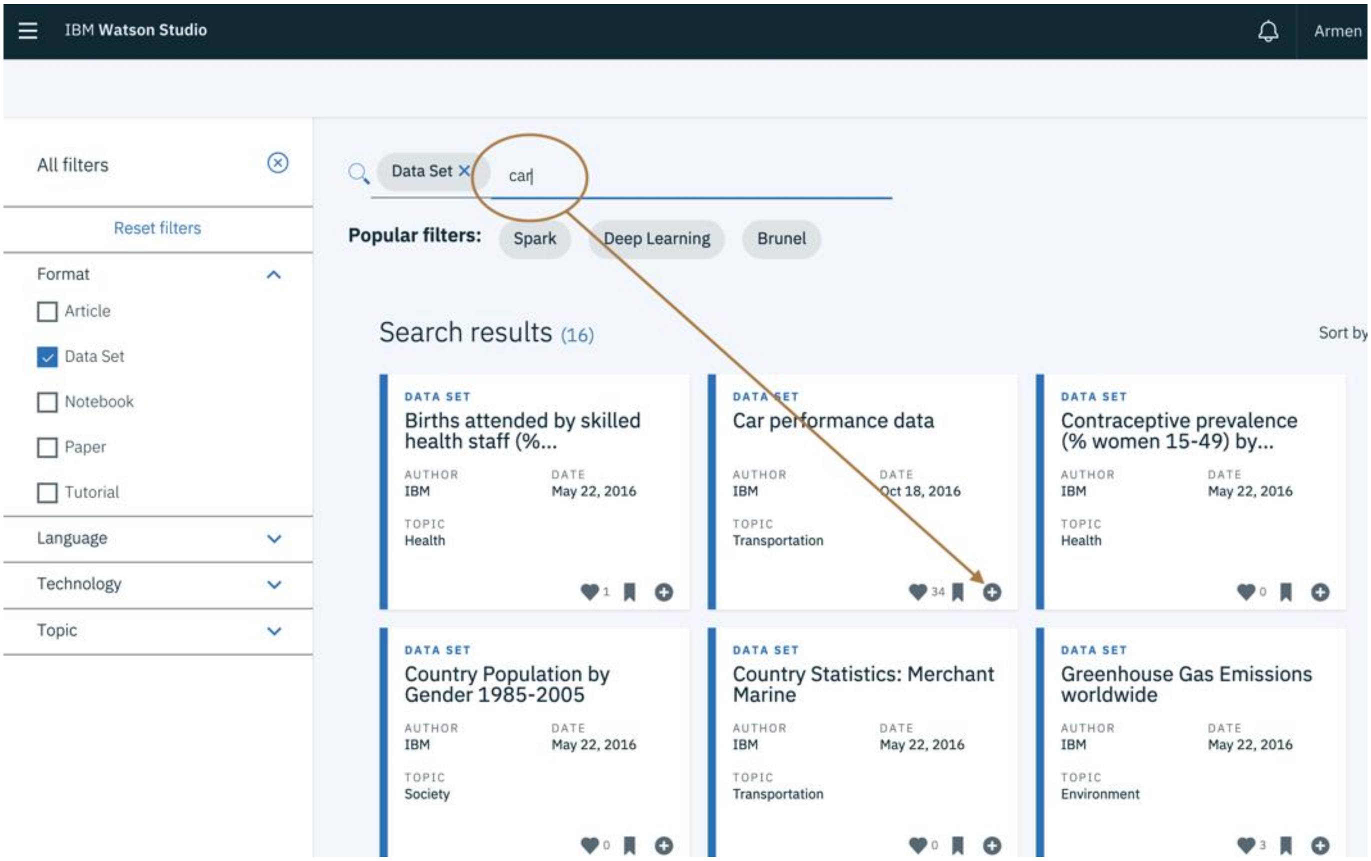
Additional tools and services can be added in Project Settings after project creation.

Storage

cloud-object-storage-uj

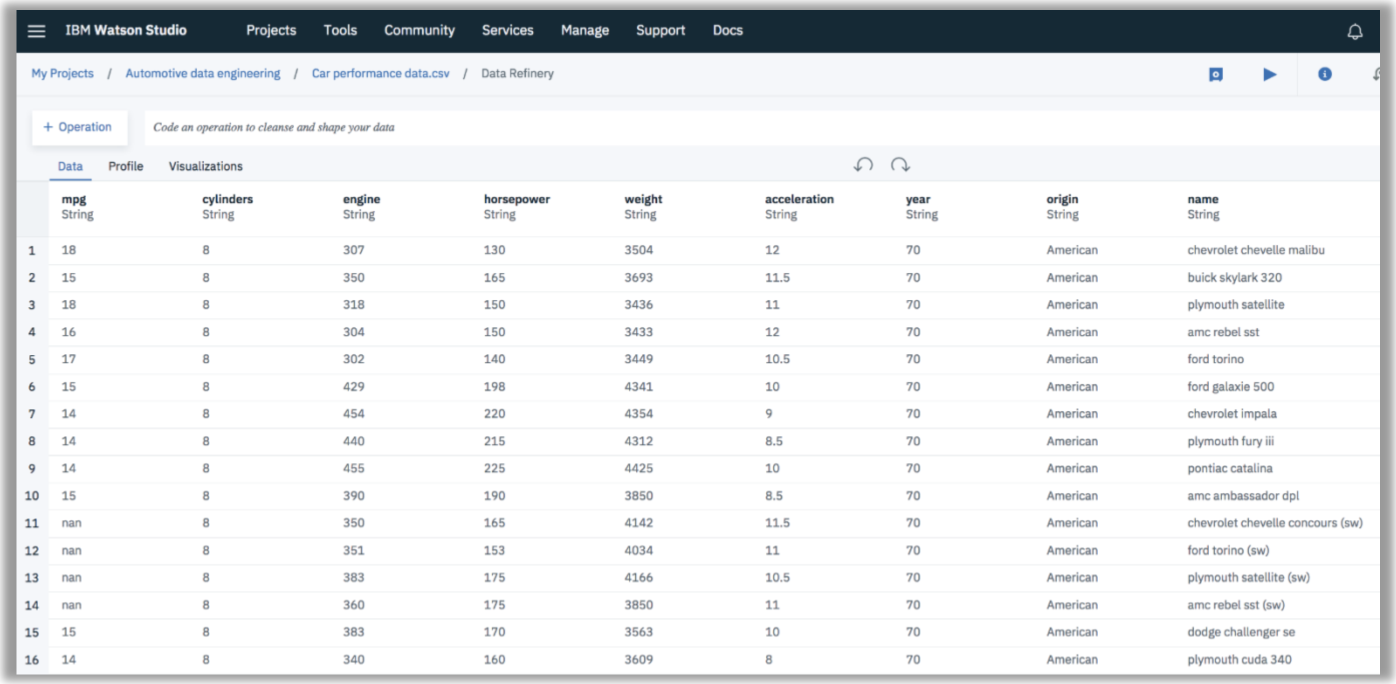
You are now ready to add data to your project. You can upload from a local drive, from a database or from the **Communities** link on the menu bar from the top. In this example you will upload data from the Communities.

1. Click **Bookmarks** from the top menu bar.
2. Click **Explore Community**.

3. 

4. Open **All Filters** tile in the left panel click **Data Set**
5. Type: *car* in the search field.
6. Once you find the tile, click the plus sign inside the Car performance data tile.
7. Select your project name to add.
8. Click **Add**.
9. Click **View Project**.
10. Click the **Assets** tab.
11. Click the data set (the CSV file) to preview. Notice that the columns are in string format and some clearly need to be numeric.
12. Close the View data assets panel.

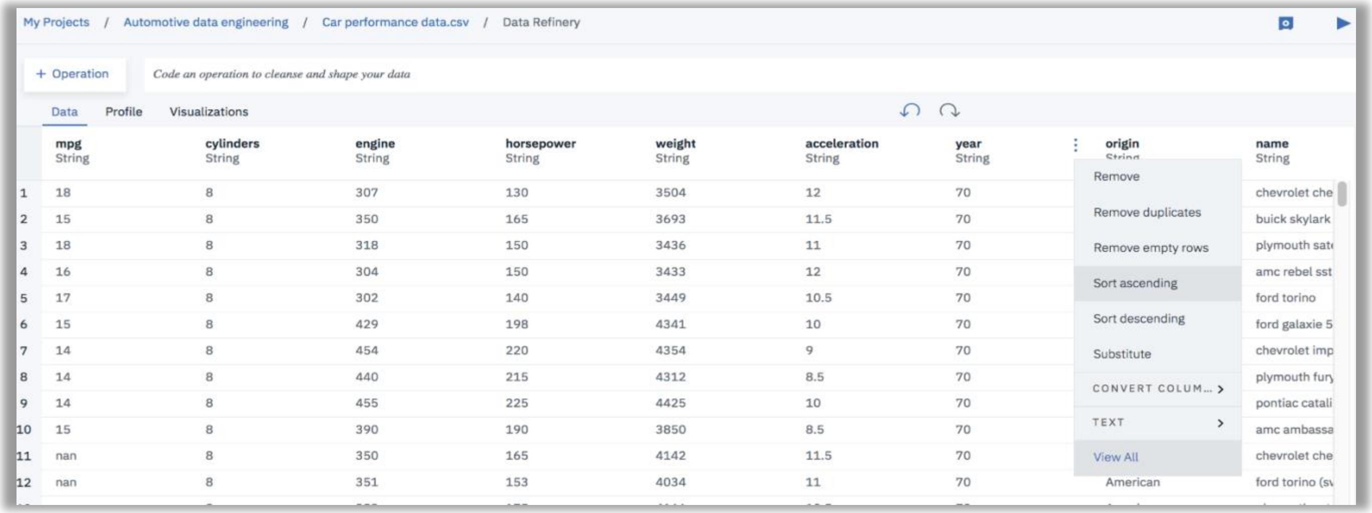
13. Click **Refine** ▶.


IBM Watson Studio
Projects
Tools
Community
Services
Manage
Support
Docs
My Projects / Automotive data engineering / Car performance data.csv / Data Refinery
+ Operation
Code an operation to cleanse and shape your data
Data
Profile
Visualizations

	mpg String	cylinders String	engine String	horsepower String	weight String	acceleration String	year String	origin String	name String
1	18	8	307	130	3504	12	70	American	chevrolet chevelle malibu
2	15	8	350	165	3693	11.5	70	American	buick skylark 320
3	18	8	318	150	3436	11	70	American	plymouth satellite
4	16	8	304	150	3433	12	70	American	amc rebel sst
5	17	8	302	140	3449	10.5	70	American	ford torino
6	15	8	429	198	4341	10	70	American	ford galaxie 500
7	14	8	454	220	4354	9	70	American	chevrolet impala
8	14	8	440	215	4312	8.5	70	American	plymouth fury iii
9	14	8	455	225	4425	10	70	American	pontiac catalina
10	15	8	390	190	3850	8.5	70	American	amc ambassador dpl
11	nan	8	350	165	4142	11.5	70	American	chevrolet chevelle concours (sw)
12	nan	8	351	153	4034	11	70	American	ford torino (sw)
13	nan	8	383	175	4166	10.5	70	American	plymouth satellite (sw)
14	nan	8	360	175	3850	11	70	American	amc rebel sst (sw)
15	15	8	383	170	3563	10	70	American	dodge challenger se
16	14	8	340	160	3609	8	70	American	plymouth cuda 340

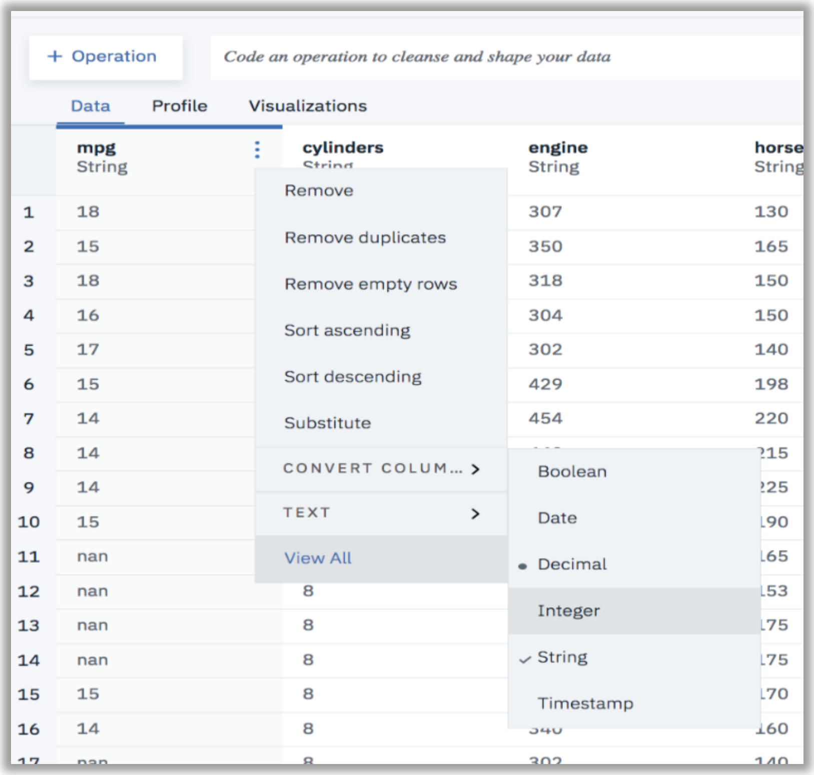
For this exercise, you may want to sort the data set by the **year** column.

1. Click the three dots in the **year** column, as you hover with your cursor over the column, to edit the content, in this example select **Sort ascending**.



	mpg String	cylinders String	engine String	horsepower String	weight String	acceleration String	year String	origin String	name String
1	18	8	307	130	3504	12	70		chevrolet che
2	15	8	350	165	3693	11.5	70		buick skylark
3	18	8	318	150	3436	11	70		plymouth sat
4	16	8	304	150	3433	12	70		amc rebel sst
5	17	8	302	140	3449	10.5	70		ford torino
6	15	8	429	198	4341	10	70		ford galaxie 5
7	14	8	454	220	4354	9	70		chevrolet imp
8	14	8	440	215	4312	8.5	70		plymouth fury
9	14	8	455	225	4425	10	70		pontiac catali
10	15	8	390	190	3850	8.5	70		amc ambassa
11	nan	8	350	165	4142	11.5	70		chevrolet che
12	nan	8	351	153	4034	11	70	American	ford torino (sv

2. Select each of the columns that have values stated as string and convert them to integer or decimal as suggested by the dot next to value that it should be. Do this for all columns. This will take a while.

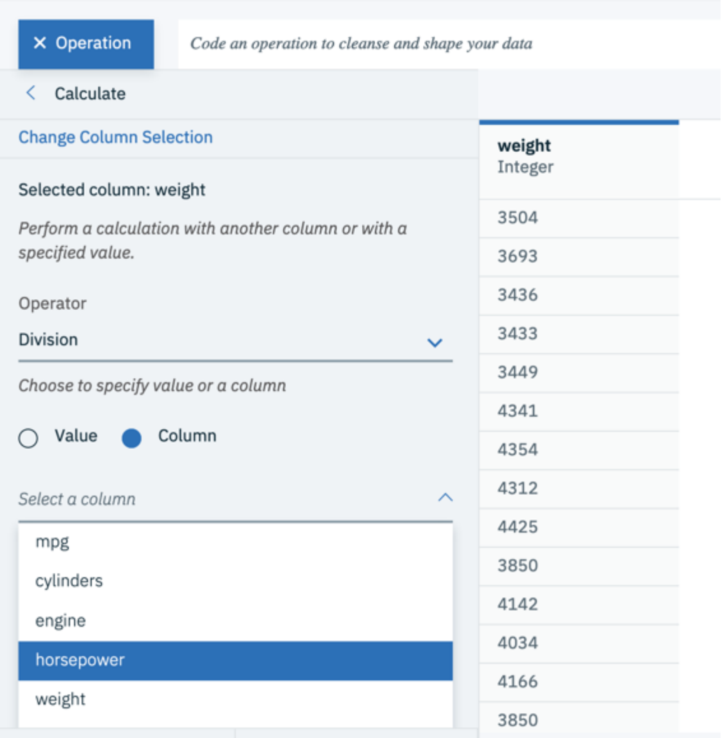


	mpg String	cylinders String	engine String	horse String
1	18		307	130
2	15		350	165
3	18		318	150
4	16		304	150
5	17		302	140
6	15		429	198
7	14		454	220
8	14			215
9	14			225
10	15			190
11	nan			165
12	nan	8		153
13	nan	8		175
14	nan	8		175
15	15	8		170
16	14	8	340	160
17	nan	8	302	140

3. Save the data flow.

Let’s say you want to calculate the weight per horse power ratio.

1. Select the **Weight** column.
2. Click **Operation**.
3. Click **Calculate**.
4. For the Operator, select **Division**.
5. For the second column, the denominator, select the **horsepower** column.

6. 

Code an operation to cleanse and shape your data

Calculate

Selected column: weight

Perform a calculation with another column or with a specified value.

Operator: Division

Choose to specify value or a column

○ Value ● Column

Select a column: horsepower

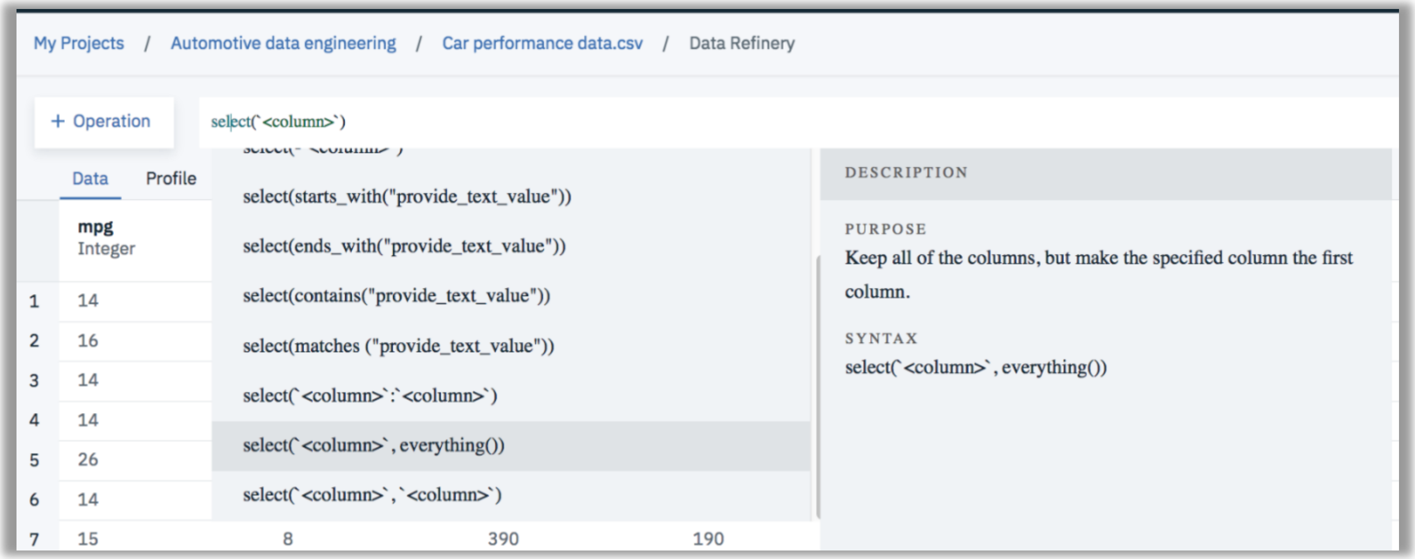
weight Integer
3504
3693
3436
3433
3449
4341
4354
4312
4425
3850
4142
4034
4166
3850

7. Check the **Create a new column for results**, check box.
8. Specify a new column name; for example, lbs/hp.
9. Sort the new column with ascending values.
10. Who knew! The Buick Estate Wagon is the best performance car; it has the highest horse power per pound of weight.

Now, let’s move that column to the front as your starting column.

1. Place your cursor in the **Code** field and the operation is to cleanse and reshape your data.

2. For the select function, click and choose the **select(‘column’,everything())** option.


My Projects / Automotive data engineering / Car performance data.csv / Data Refinery
+ Operation
select(`<column>`)
Data
Profile
select(`<column>`)
select(starts_with("provide_text_value"))
select(ends_with("provide_text_value"))
select(contains("provide_text_value"))
select(matches ("provide_text_value"))
select(`<column>`:`<column>`)
select(`<column>`, everything())
select(`<column>`, `<column>`)

	mpg Integer			
1	14			
2	16			
3	14			
4	14			
5	26			
6	14			
7	15	8	390	190

DESCRIPTION

PURPOSE

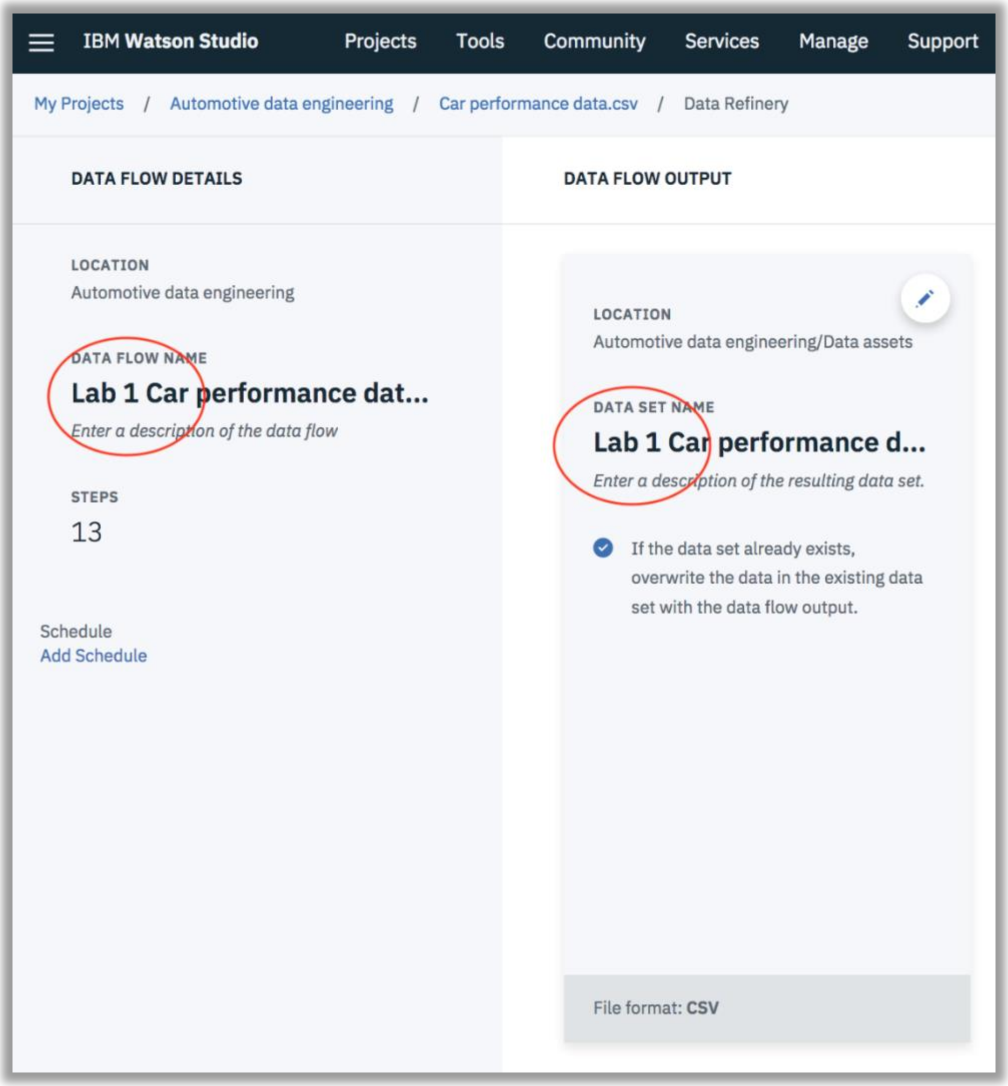
Keep all of the columns, but make the specified column the first column.

SYNTAX

```
select(`<column>`, everything())
```

3. Specify the **lbs/hp** as the desired column and click **Apply** (to the far right).
4. Your turn now; round the values of the lbs-hp feature to 2 decimal points. Hint: it has something to do with the **Math** Operation.
5. You can always delete or redo your operation from the steps panel in the right pane.

You are now ready to change the **DATA FLOW DETAILS** name and the **DATA FLOW OUTPUT** name.


IBM Watson Studio
Projects
Tools
Community
Services
Manage
Support
My Projects / Automotive data engineering / Car performance data.csv / Data Refinery

DATA FLOW DETAILS

LOCATION
Automotive data engineering

DATA FLOW NAME
Lab 1 Car performance dat...
Enter a description of the data flow

STEPS
13

Schedule
Add Schedule

DATA FLOW OUTPUT

LOCATION
Automotive data engineering/Data assets

DATA SET NAME
Lab 1 Car performance d...
Enter a description of the resulting data set.

✓ If the data set already exists, overwrite the data in the existing data set with the data flow output.

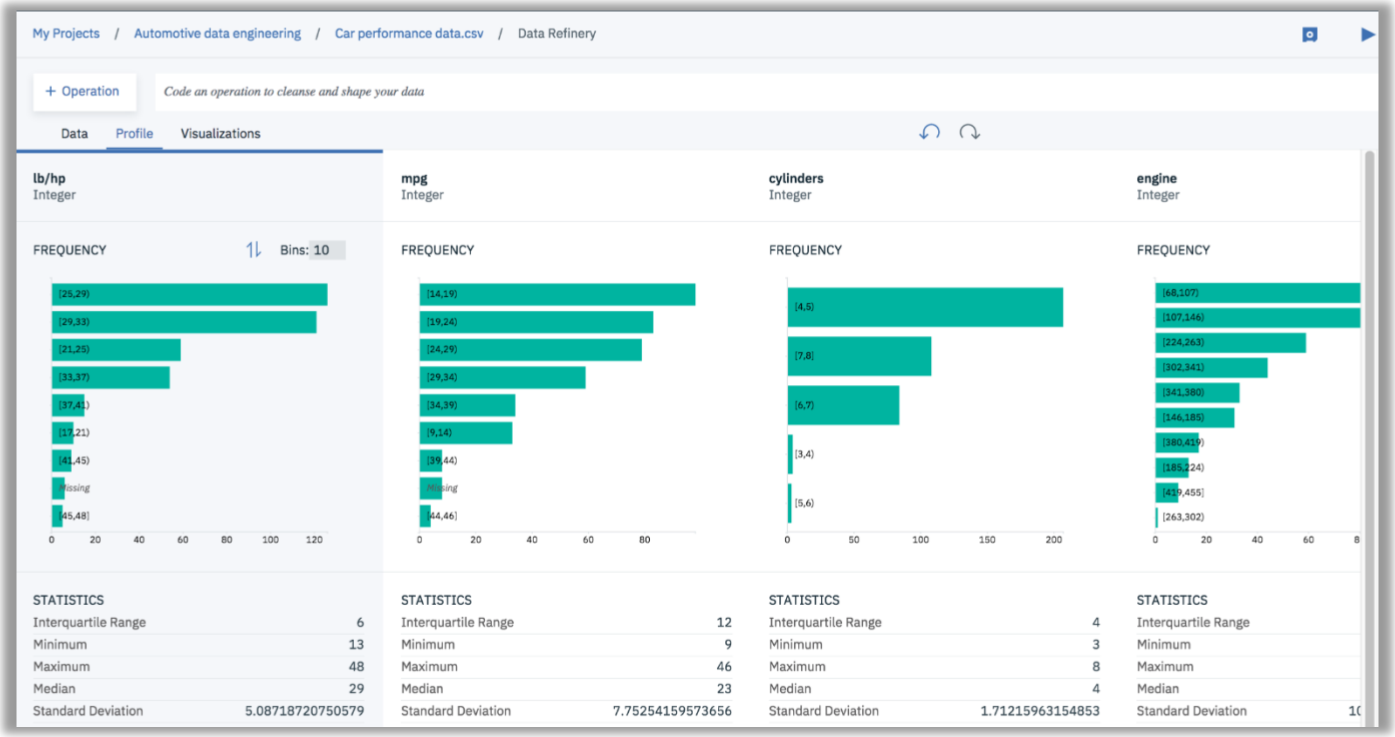
File format: **CSV**

1. Click **Apply** after you change the name in the DATA FLOW DETAILS box and enable the check mark in the OUTPUT box.
2. Click **Save and Run** the data flow (bottom right of the form).

   This may take a few minutes. In the meantime, you can view the status by clicking **View Flow** in the ensuing dialog box. The status soon indicates **Completed**.
3. When the run is complete click **Refine** to further shape your data (this may take a few minutes).
4. Once back to your data flow, sort the **lbs-hp** column in descending order.
5. Save the data flow

6. Click the **Profile** tab.

Take a moment and view the results. The Profile tab reveals summary statistics results where you can decide, with a top down view, which of the columns, or feature sets needs further add/edit/update/delete activities.

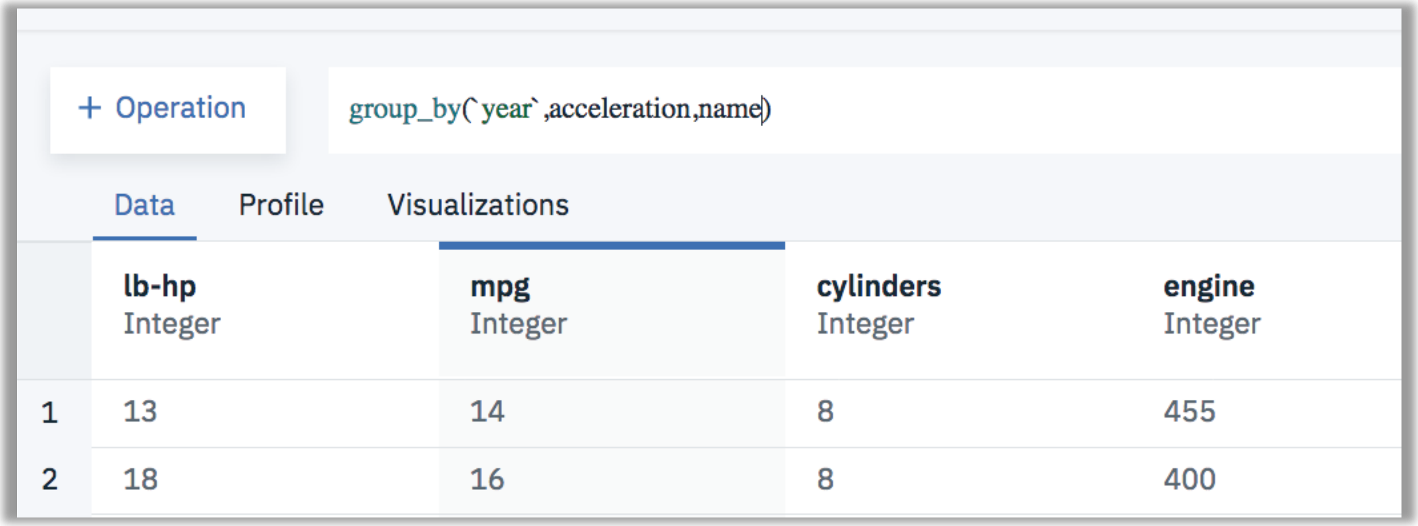


STATISTICS	lb/hp	mpg	cylinders	engine
Interquartile Range	6	12	4	
Minimum	13	9	3	
Maximum	48	46	8	
Median	29	23	4	
Standard Deviation	5.08718720750579	7.75254159573656	1.71215963154853	1

Let's do more data wrangling

1. In the coding box, select **group_by**, specify year as the column name and type, separated by comma: **acceleration** and **name**.

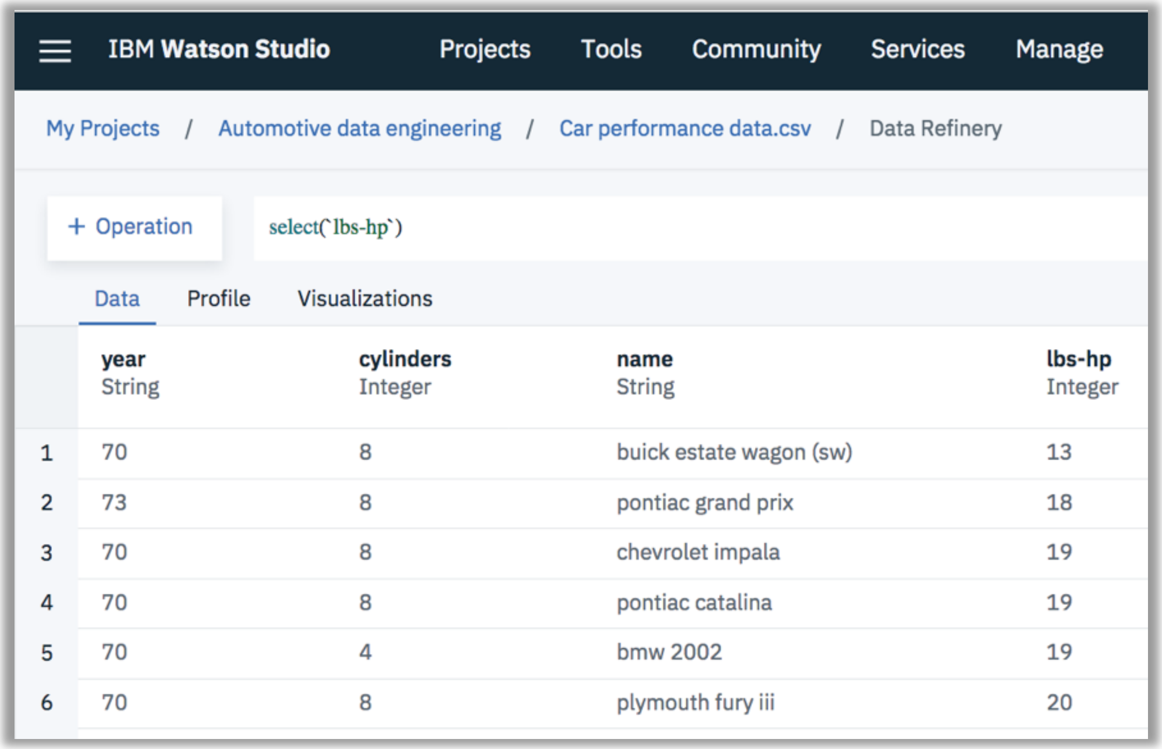
```
group_by(`year`,acceleration,name)
```



```
group_by(`year`,acceleration,name)
```

	lb-hp Integer	mpg Integer	cylinders Integer	engine Integer
1	13	14	8	455
2	18	16	8	400

2. Click **Apply**.
3. From the coding box, select the **lb-hp** column
4. Click **Apply**.
5. Using the select function, group by the **lbs-hp** column.



```
select(`lbs-hp`)
```

	year String	cylinders Integer	name String	lbs-hp Integer
1	70	8	buick estate wagon (sw)	13
2	73	8	pontiac grand prix	18
3	70	8	chevrolet impala	19
4	70	8	pontiac catalina	19
5	70	4	bmw 2002	19
6	70	8	plymouth fury iii	20

6. Save the data flow.

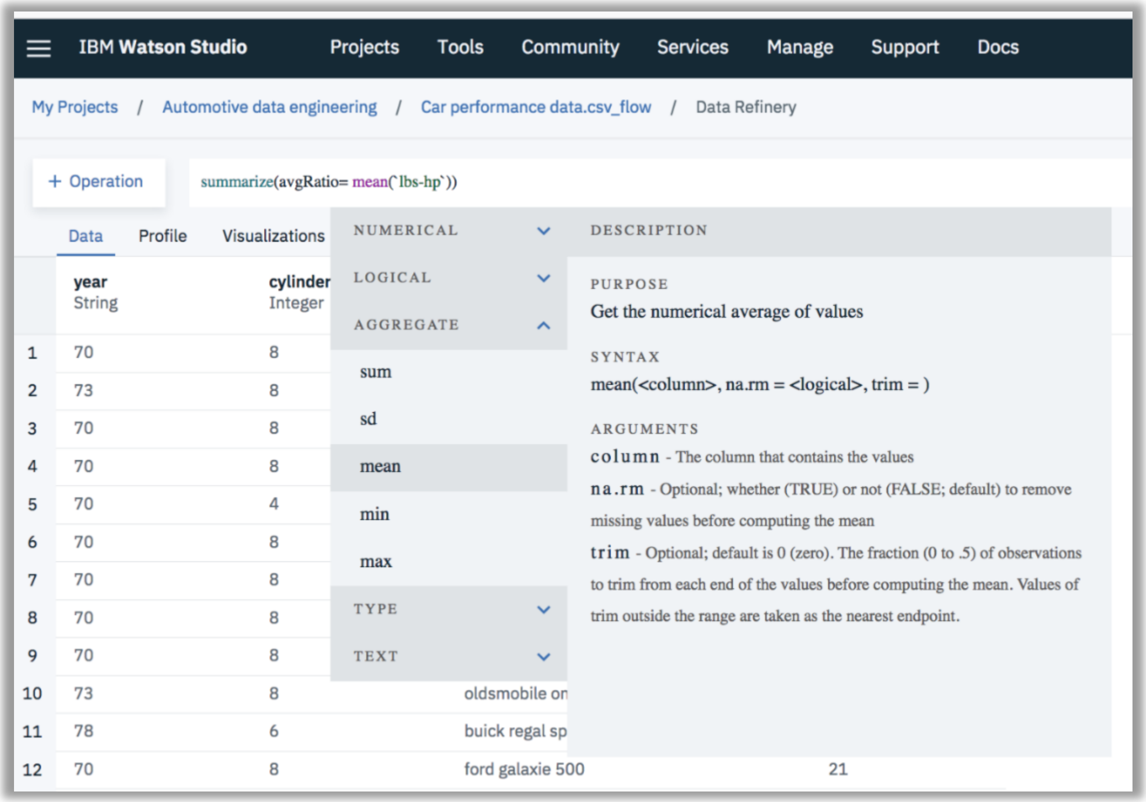
Let’s do some statistics

1. From the Code box, use the **summarize** operation to find the mean value of the weight to horse power ratio.

```
summarize(provide_new_column = <func>(‘<column>’))
```

2. Type a new column name, such as **avgRatio**.
3. Select the **lbs-hp** column for the mean parameter:

```
summarize(avgRatio = mean(‘ibs-hp’))
```



```
summarize(avgRatio= mean(`lbs-hp`))
```

	year String	cylinder Integer		
1	70	8		
2	73	8		
3	70	8		
4	70	8		
5	70	4		
6	70	8		
7	70	8		
8	70	8		
9	70	8		
10	73	8	oldsmobile on	
11	78	6	buick regal sp	
12	70	8	ford galaxie 500	21

DESCRIPTION

PURPOSE

Get the numerical average of values

SYNTAX

mean(<column>, na.rm = <logical>, trim =)

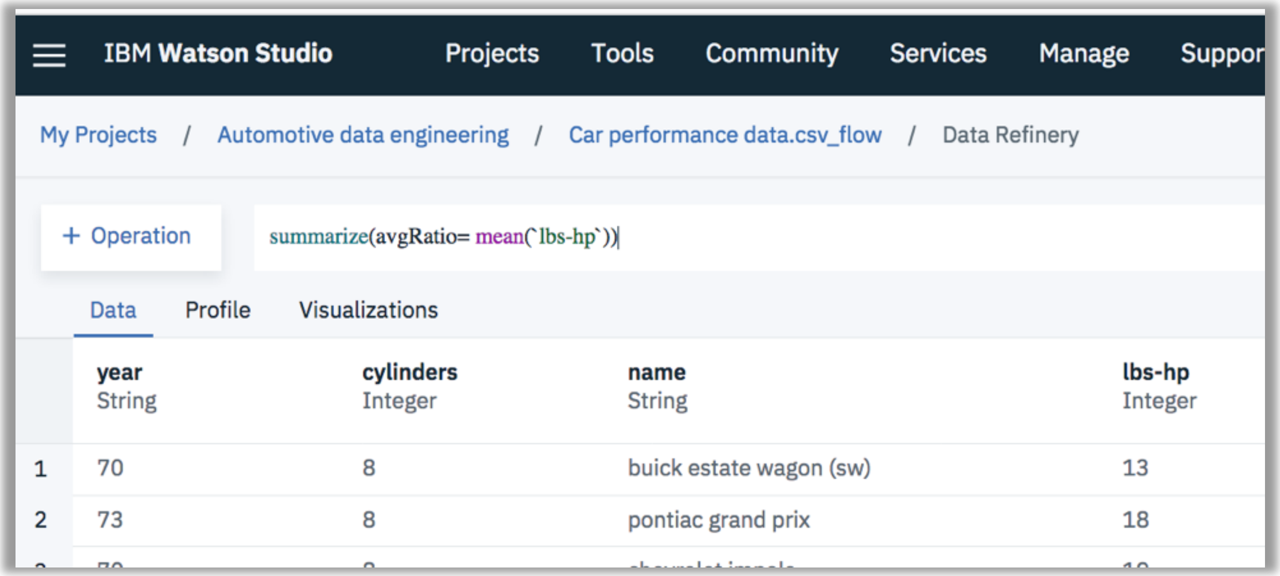
ARGUMENTS

`column` - The column that contains the values

`na.rm` - Optional; whether (TRUE) or not (FALSE; default) to remove missing values before computing the mean

`trim` - Optional; default is 0 (zero). The fraction (0 to .5) of observations to trim from each end of the values before computing the mean. Values of trim outside the range are taken as the nearest endpoint.

4. Click **Apply**. This is the average for the total weight to horse power ratio



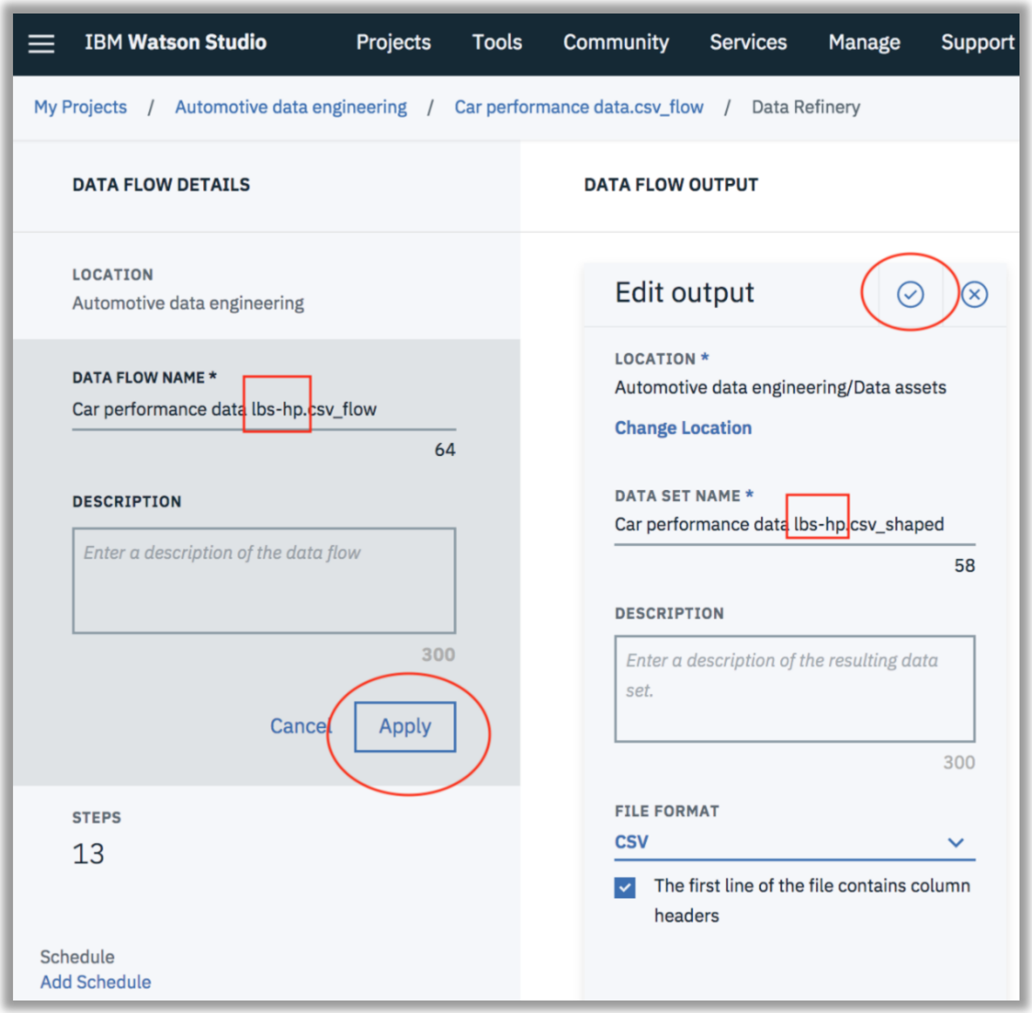
```
summarize(avgRatio= mean(`lbs-hp`))
```

	year String	cylinders Integer	name String	lbs-hp Integer
1	70	8	buick estate wagon (sw)	13
2	73	8	pontiac grand prix	18

5. You are now ready to run the data flow ▶

You can now change the DATA FLOW DETAILS name and the DATA FLOW OUTPUT name

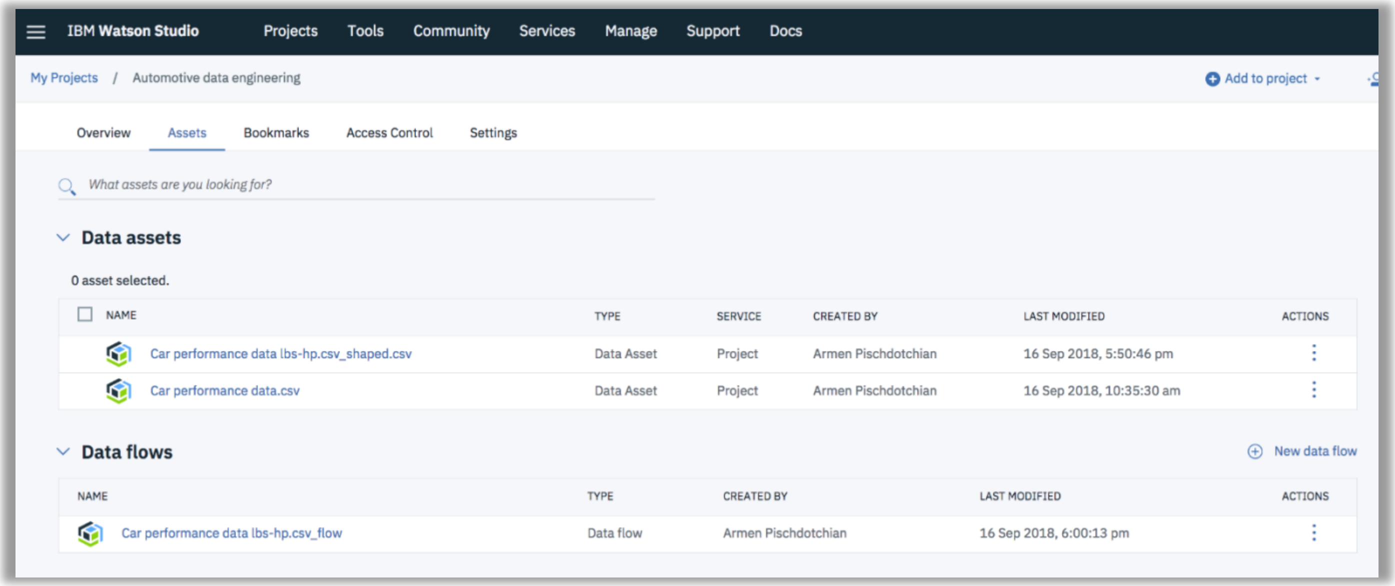
1. Click **Apply** after you change the name (again) in the DATA FLOW DETAILS box and the check mark in the OUTPUT box.



IBM Watson Studio
Projects
Tools
Community
Services
Manage
Support
My Projects / Automotive data engineering / Car performance data.csv_flow / Data Refinery
DATA FLOW DETAILS
DATA FLOW OUTPUT
LOCATION
Automotive data engineering
DATA FLOW NAME *
Car performance data lbs-hp.csv_flow
64
DESCRIPTION
Enter a description of the data flow
300
Cancel
Apply
STEPS
13
Schedule
Add Schedule
Edit output
LOCATION *
Automotive data engineering/Data assets
Change Location
DATA SET NAME *
Car performance data lbs-hp.csv_shaped
58
DESCRIPTION
Enter a description of the resulting data set.
300
FILE FORMAT
CSV
The first line of the file contains column headers

2. Click **Save and Run** the data flow (bottom right of the form).

3. This may take a few minutes. In the meantime, you can view the status by clicking **View Flow** in the ensuing dialog box.



Data assets

0 asset selected.

□ NAME	TYPE	SERVICE	CREATED BY	LAST MODIFIED	ACTIONS
Car performance data lbs-hp.csv_shaped.csv	Data Asset	Project	Armen Pischdotchian	16 Sep 2018, 5:50:46 pm	⋮
Car performance data.csv	Data Asset	Project	Armen Pischdotchian	16 Sep 2018, 10:35:30 am	⋮

Data flows

NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
Car performance data lbs-hp.csv_flow	Data flow	Armen Pischdotchian	16 Sep 2018, 6:00:13 pm	⋮

4. The status should indicate **Completed**.
5. When the run is complete click **Refine** to further shape your data (this may take a few minutes).
6. Once back to your data flow, sort the **lbs-hp** column in descending order.
7. Save the data flow
8. Click the **Automotive Data Engineering** project name and view the two data assets and the data flow. It is good practice to save often, with a unique name, at each interval of your data shaping journey.

Lab 3. Validating Automotive Data

At any time after you've added data to Data Refinery, you can validate your data. Typically, you'll want to do this at multiple points in the refinement process.
To validate your data:

1. From Data Refinery, click the Profile tab.
2. Review the metrics for each column.
3. Take appropriate actions, as described in the following sections, depending on what you learn.

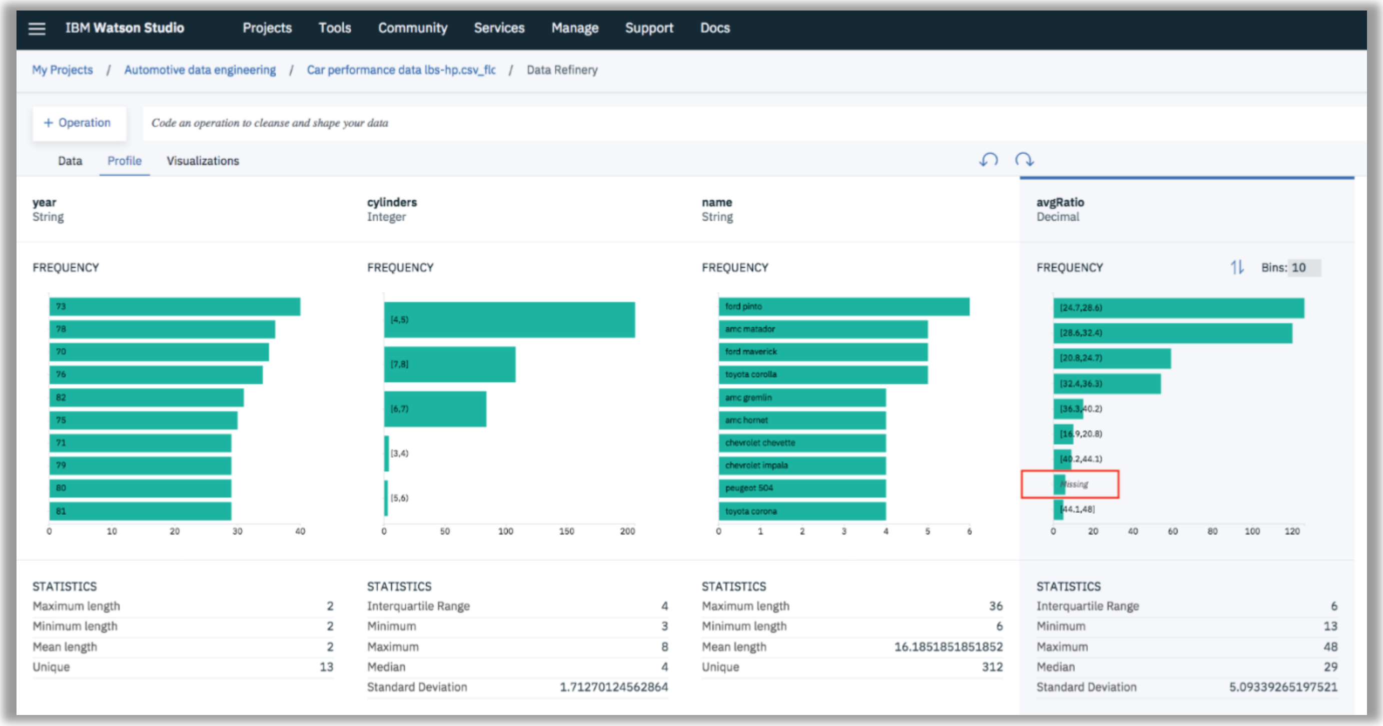
Frequency

Frequency is the number of times that a value, or a value in a specified range, occurs. Each frequency distribution (bar) shows the count of unique values in a column.
Review the frequency distribution to find anomalies in your data. If you want to cleanse your data of those anomalies, simply remove the values.
For Integer and Date/Time columns, you can customize the number of bins (groupings) that you want to see. In the default multi-column view, the maximum is 20. If you expand the frequency chart row, the maximum is 50.

Statistics

Statistics are a collection of quantitative data. The statistics for each column show the minimum, maximum, mean, and number of unique values in that column.
Depending on a column's data type, the statistics for each column will vary slightly. For example, statistics for a column of data type integer have minimum, maximum, and mean values while statistics for a column of data type string have minimum length, maximum length, and mean length values.

Study the metrics. Notice that the avgRatio profile contains 6 missing values (hover your mouse over it). Let’s remove those values.



STATISTICS

Maximum length	2
Minimum length	2
Mean length	2
Unique	13

STATISTICS

Interquartile Range	4
Minimum	3
Maximum	8
Median	4
Standard Deviation	1.71270124562864

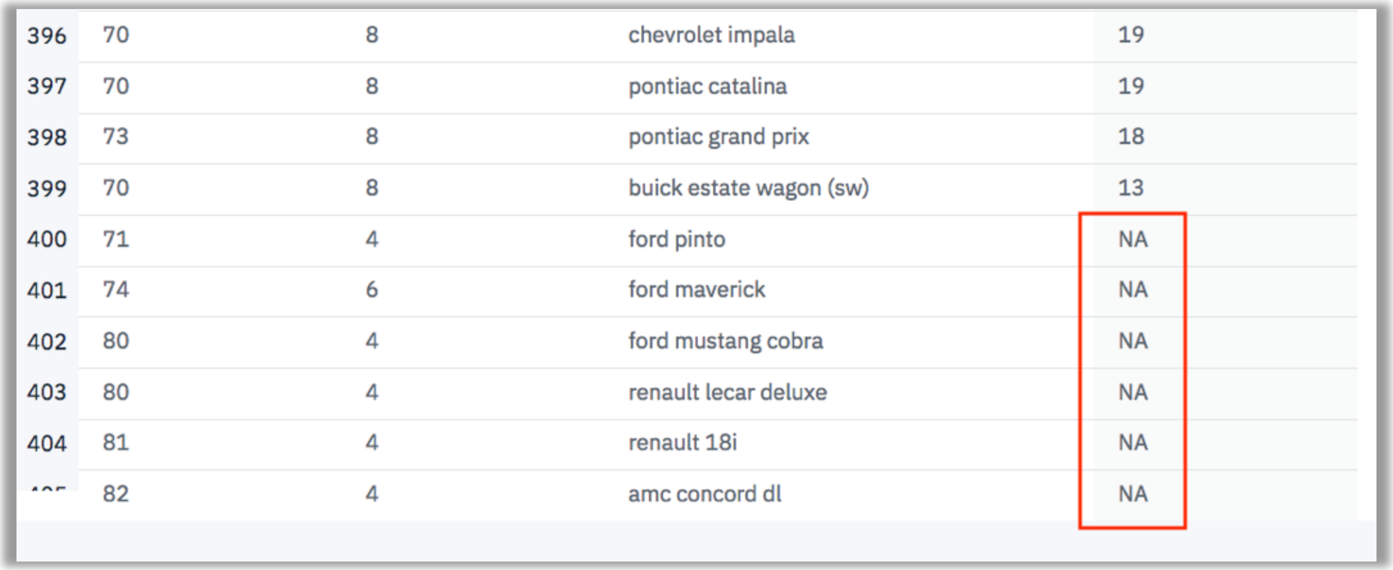
STATISTICS

Maximum length	36
Minimum length	6
Mean length	16.1851851851852
Unique	312

STATISTICS

Interquartile Range	6
Minimum	13
Maximum	48
Median	29
Standard Deviation	5.09339265197521

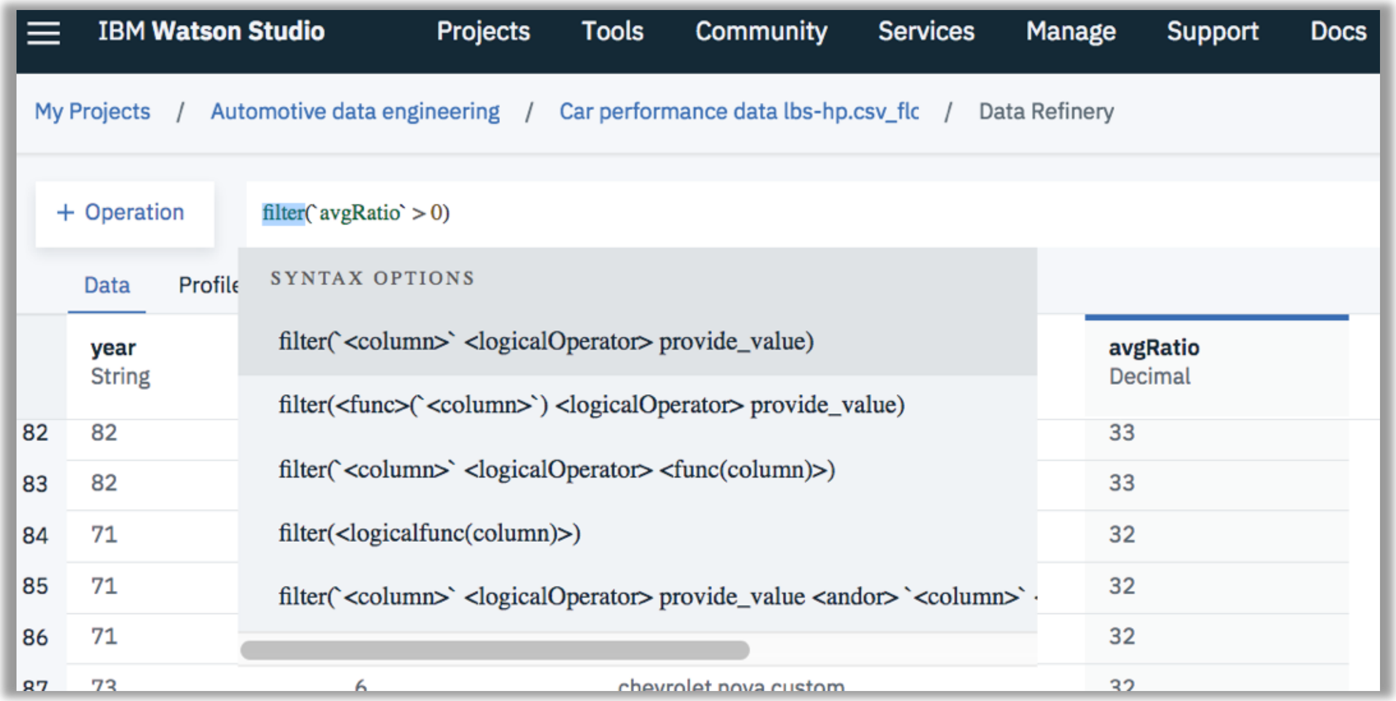
1. Click the **Data** tab to go back to your data flow.
2. If you have not sorted the column by descending order yet (or ascending) do so now.
3. Scroll down, if you sorted descending, and notice 6 of the records have a value of NA.



396	70	8	chevrolet impala	19
397	70	8	pontiac catalina	19
398	73	8	pontiac grand prix	18
399	70	8	buick estate wagon (sw)	13
400	71	4	ford pinto	NA
401	74	6	ford maverick	NA
402	80	4	ford mustang cobra	NA
403	80	4	renault lecar deluxe	NA
404	81	4	renault 18i	NA
405	82	4	amc concord dl	NA

4. From the **Code** box, use the filter operation to find the row in the avgRatio column that have a value of NA.

```
filter(`avgRatio` > 0)
```



```
filter(`avgRatio` > 0)
```

```
filter(`<column>` <logicalOperator> provide_value)
filter(<func>(`<column>`) <logicalOperator> provide_value)
filter(`<column>` <logicalOperator> <func(column)>)
filter(<logicalfunc(column)>)
filter(`<column>` <logicalOperator> provide_value <andor> `<column>`
```

	year String	avgRatio Decimal
82	82	33
83	82	33
84	71	32
85	71	32
86	71	32
87	73	32

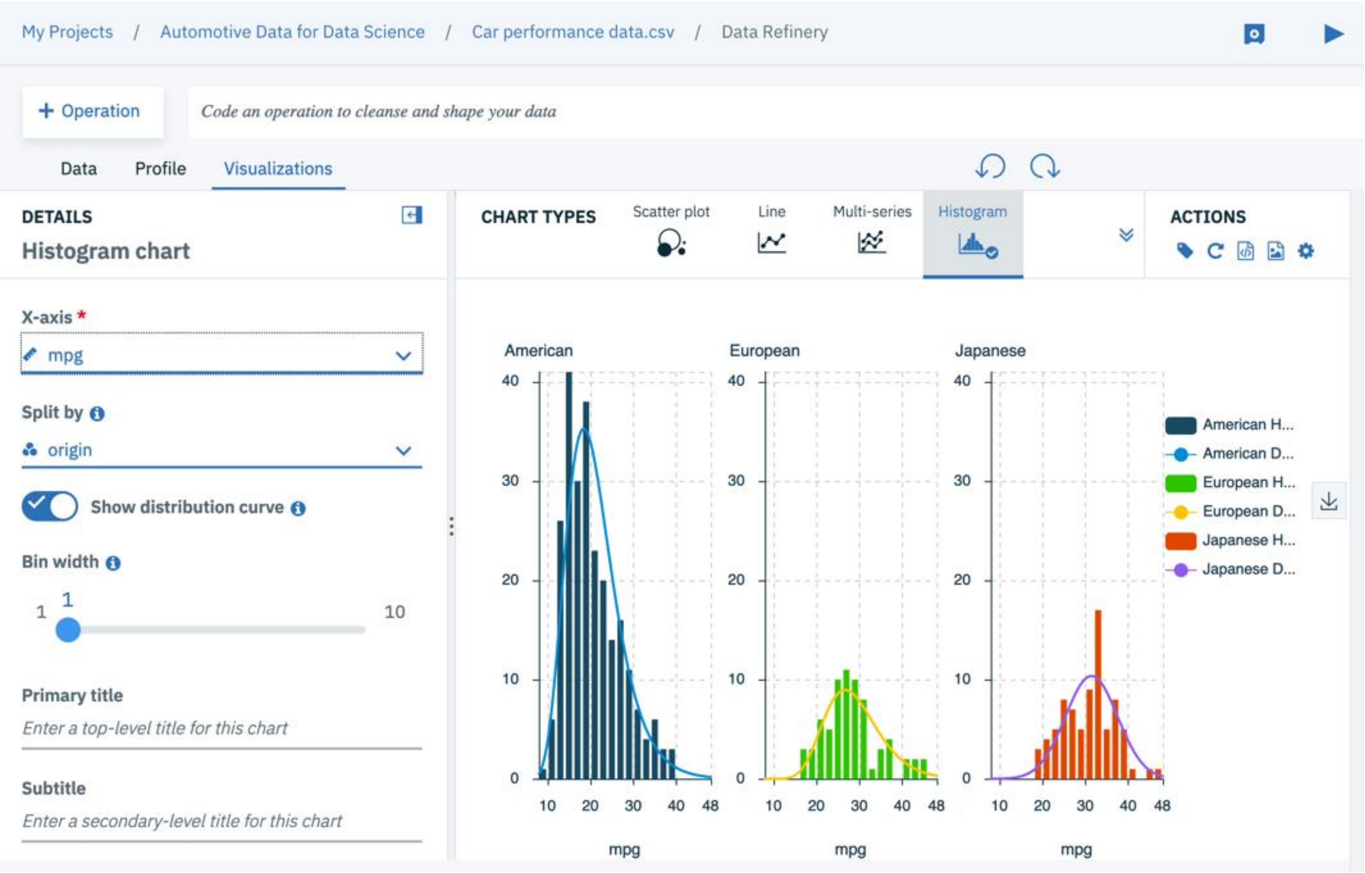
5. Select the **avgRatio** column and set it's value greater than 0.
6. Click **Apply**. This is the average for the total weight to horse power ratio.
7. View the **Profile**. Notice the 6 records have been removed.
8. You are now ready to run the data flow ▶
9. Change the output name and the Data Flow name with a prefix of Lab 2 to include what you just did; for example: **Lab 2 Automotive data in both boxes**.
10. Save the data flow 💾

Lab 4. Visualizing Automotive Data

Visualizing information in graphical ways can give you insights into your data. By enabling you to look at and explore data from different perspectives, visualizations can help you identify patterns, connections, and relationships within that data as well as understand large amounts of information very quickly. At any time after you've added data to Data Refinery, you can visualize your data.
To visualize your data:

1. Retrieve your latest Data Asset.
2. Click **Refine**.
3. From Data Refinery, click the **Visualizations** tab.
4. Select the columns that you want to work with, then click **Visualize Data**.
5. Select your X (first column you specify) and the Y (second column you specify) as variables.
6. Try the following variations:

   a) Columns = avgRatio and cylinders; Chart Types = Stacked Count
   b) Columns = avgRatio and name; Chart Types = Pie chart
   c) Columns = name and avgRatio; Chart Types = scatter plot



My Projects / Automotive Data for Data Science / Car performance data.csv / Data Refinery

DETAILS
Histogram chart

X-axis * mpg

Split by origin

Show distribution curve (on)

Bin width 1

Primary title: Enter a top-level title for this chart

Subtitle: Enter a secondary-level title for this chart

7. Experiment with various renderings.

8. If the type of visualization that you want to see isn't currently displayed, select it from the Chart types list.

   Tip: The chart types are ordered from most relevant to least relevant, based on the selected columns.

9. Optional: If you're familiar with Brunel Visualization Language, you can modify the visualization by editing the syntax and then clicking Update Visualization.

Brunel Visualization Language

Brunel Visualization Language is a high-level language developed by IBM and open-sourced in 2015. Brunel describes visualizations in terms of composable actions and drives a visualization engine (D3) that performs the actual rendering and interactivity.

Brunel visualizations

Many types of Brunel visualizations are interactive: you can zoom and pan across a graph, for example. You can also view more complex visualizations that display multiple dimensions.