```
var df = spark.read.parquet("/user/yx3494_nyu_edu/scr_data/funding_safe")SPARK JOB FINISHED
val cols: List[String] = List("N/RC_Index_Description", "N/RC_Index")
df = df.filter(df("N/RC_Index_Description").isNotNull)
df = df.withColumn("N/RC_Index",
    when(df("N/RC_Index") === 2, 1)
    .when(df("N/RC_Index") === 3, 1)
    .when(df("N/RC_Index") === 4, 1)
    .when(df("N/RC_Index") === 5, 2)
    .when(df("N/RC_Index") === 6, 2)
   .otherwise(df("N/RC_Index")))

df = df.withColumn("N/RC_Index_Description",
    when(df("N/RC_Index") === 1, "High (1-4)")
    .when(df("N/RC_Index") === 2, "Low (5-6)")
   .otherwise(df("N/RC_Index")))

df = df.withColumn("County_Name",
    when(df("County_Name") === "BRONX", "NYC")
    .when(df("County_Name") === "BROOKLYN", "NYC")
    .when(df("County_Name") === "NEW YORK", "NYC")
    .when(df("County_Name") === "NYC CENTRAL OFFICE", "NYC")
    .when(df("County_Name") === "RICHMOND", "NYC")
    .when(df("County_Name") === "QUEENS", "NYC")
    .when(df("County_Name") === "ALBANY", "ALBANY")
   .otherwise("Non-NYC"))

df = df.select(
    "School_BEDS_Code",
    "Year",
    "School_Name",
    "County_Name",
    "N/RC_Index",
    "N/RC_Index_Description",
    "Total_Enrollment",
    "Total_Funding",
    "Total_Funding_per_Pupil",
    "Federal_Funding",
    "Federal_Funding_per_Pupil",
    "State_&_Local_Funding",
    "State_&_Local_Funding_per_Pupil",
    "Total_Teachers",
    "Teacher_per_Pupil",
    "Total_Staff",
    "Staff_per_Pupil"
    )

df.printSchema
z.show(df.groupBy("County_Name").count)
```

```
root
 |-- School_BEDS_Code: string (nullable = true)
 |-- Year: integer (nullable = true)
 |-- School_Name: string (nullable = true)
 |-- County_Name: string (nullable = false)
 |-- N/RC_Index: integer (nullable = true)
 |-- N/RC_Index_Description: string (nullable = true)
 |-- Total_Enrollment: decimal(21,2) (nullable = true)
 |-- Total_Funding: decimal(20,2) (nullable = true)
 |-- Total_Funding_per_Pupil: decimal(20,2) (nullable = true)
 |-- Federal_Funding: decimal(20,2) (nullable = true)
 |-- Federal_Funding_per_Pupil: decimal(20,2) (nullable = true)
 |-- State_&_Local_Funding: decimal(20,2) (nullable = true)
 |-- State_&_Local_Funding_per_Pupil: decimal(20,2) (nullable = true)
 |-- Total_Teachers: decimal(21,2) (nullable = true)
 |-- Teacher_per_Pupil: decimal(20,2) (nullable = true)
 |-- Total_Staff: decimal(20,2) (nullable = true)
 |-- Staff_per_Pupil: decimal(20,2) (nullable = true)
```

County_Name	Cou
ALBANY	241
Non-NYC	127
NYC	725

```
df: org.apache.spark.sql.DataFrame = [School_BEDS_Code: string, Year: int ... 15 more field
s]
cols: List[String] = List(N/RC_Index_Description, N/RC_Index)
df: org.apache.spark.sql.DataFrame = [School_BEDS_Code: string, Year: int ... 15 more field
s]
df: org.apache.spark.sql.DataFrame = [School_BEDS_Code: string, Year: int ... 15 more field
s]
df: org.apache.spark.sql.DataFrame = [School_BEDS_Code: string, Year: int ... 15 more field
s]
df: org.apache.spark.sql.DataFrame = [School_BEDS_Code: string, Year: int ... 15 more field
s]
df: org.apache.spark.sql.DataFrame = [School_BEDS_Code: string, Year: int ... 15 more field
s]
```

Took 43 sec. Last updated by anonymous at December 10 2024, 4:44:57 PM.

```
var groupedDf = df.groupBy("N/RC_Index_Description").agg(sum("Total_Enrollment").alias("To
z.show(groupedDf)
```

settings

Available Fields

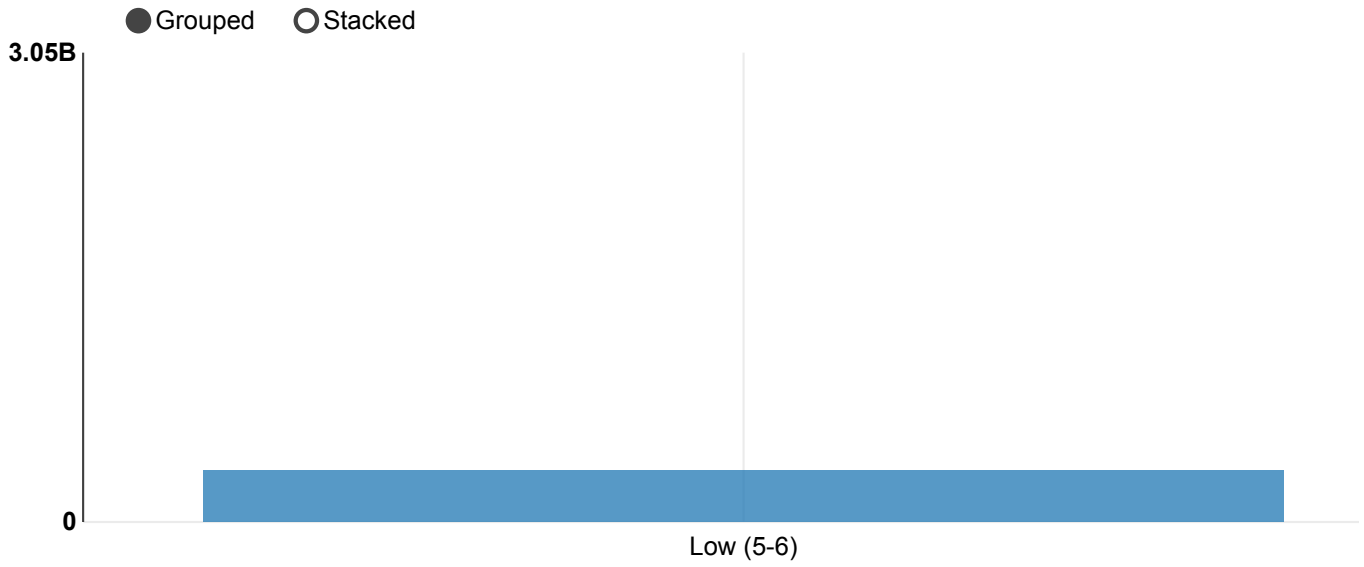
N/RC_Index_Description | Total_Enrollment

keys: N/RC_Index_Description ✖

groups:

values: Total_Enrollment **AVG** ✖

xAxis : Default | Rotate | Hide




```
groupedDf: org.apache.spark.sql.DataFrame = [N/RC_Index_Description: string, Total_Enrollme
nt: decimal(31,2)]
```

Took 1 sec. Last updated by anonymous at December 10 2024, 11:26:52 AM.

```
var groupedDf = df.groupBy("Year", "N/RC_Index_Description").agg(avg("Total_Funding_per_Pu
    ("Stddev_Funding_per_Pupil"))
z.show(groupedDf)
```

settings

Available Fields

Year | N/RC_Index_Description | Total_Funding_per_Pupil | Stddev_Funding_per_Pupil

keys: Year ✖

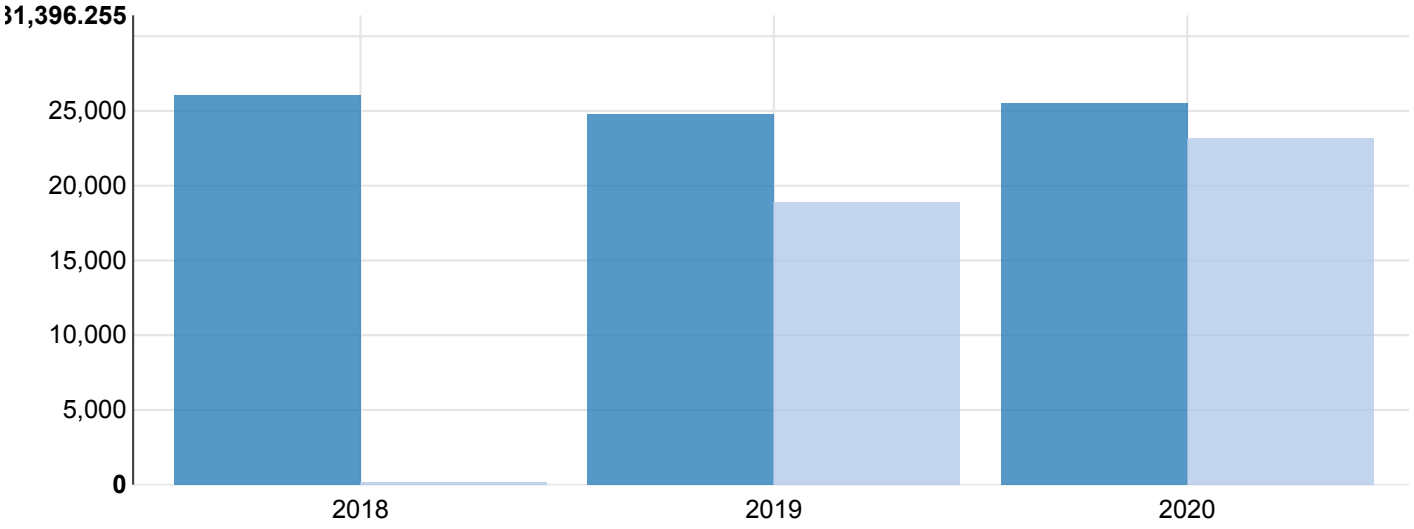
groups: N/RC_Index_Description ✖

values: Total_Funding_per_Pupil **AVG**

xAxis :

Default | Rotate | Hide

Grouped Stacked




```
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, N/RC_Index_Description: string ...
2 more fields]
```

Took 1 sec. Last updated by anonymous at December 10 2024, 9:45:44 AM. (outdated)

```
var groupedDf = df.groupBy("Year", "N/RC_Index_Description").agg(avg("Teacher_per_Pupil").
    ("Stddev_Teacher_per_Pupil"))
z.show(groupedDf)
```

settings

Available Fields

Year | N/RC_Index_Description | Teacher_per_Pupil | Stddev_Teacher_per_Pupil

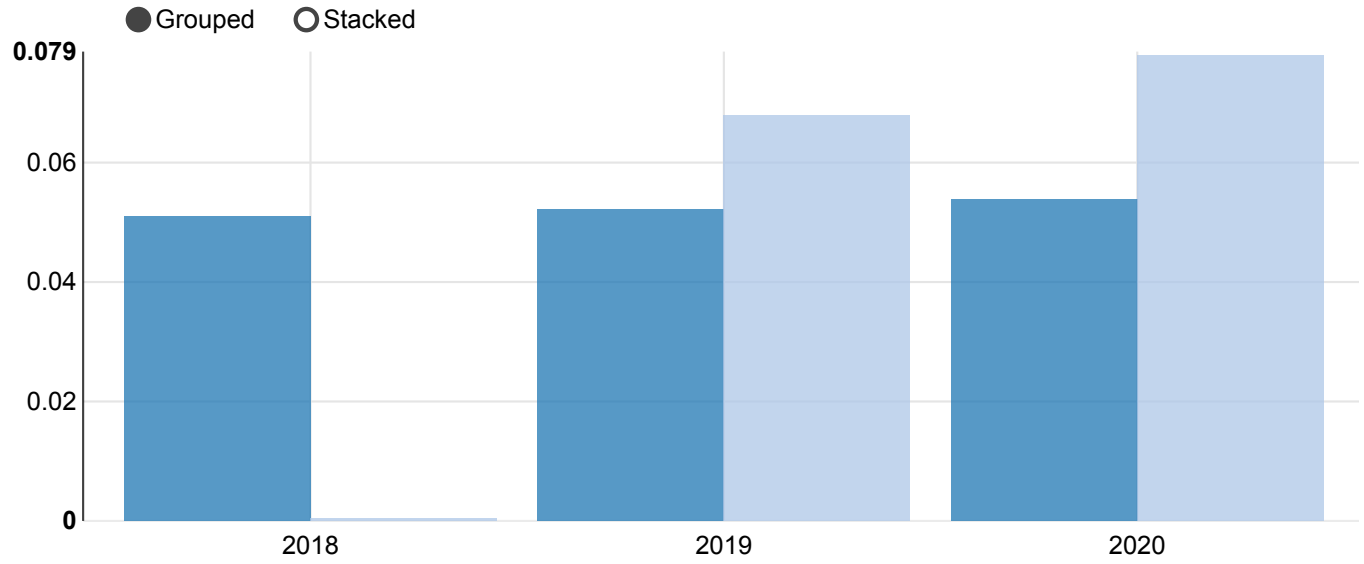
keys: Year ✖

groups: N/RC_Index_Description ✖

values: Teacher_per_Pupil **SUM** ✖

xAxis :

Default | Rotate | Hide




```
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, N/RC_Index_Description: string ...
2 more fields]
```

Took 0 sec. Last updated by anonymous at December 10 2024, 9:47:34 AM. (outdated)

```
var groupedDf = df.groupBy("Year", "N/RC_Index_Description").agg(avg("Staff_per_Pupil").
z.show(groupedDf)
```

settings

Available Fields

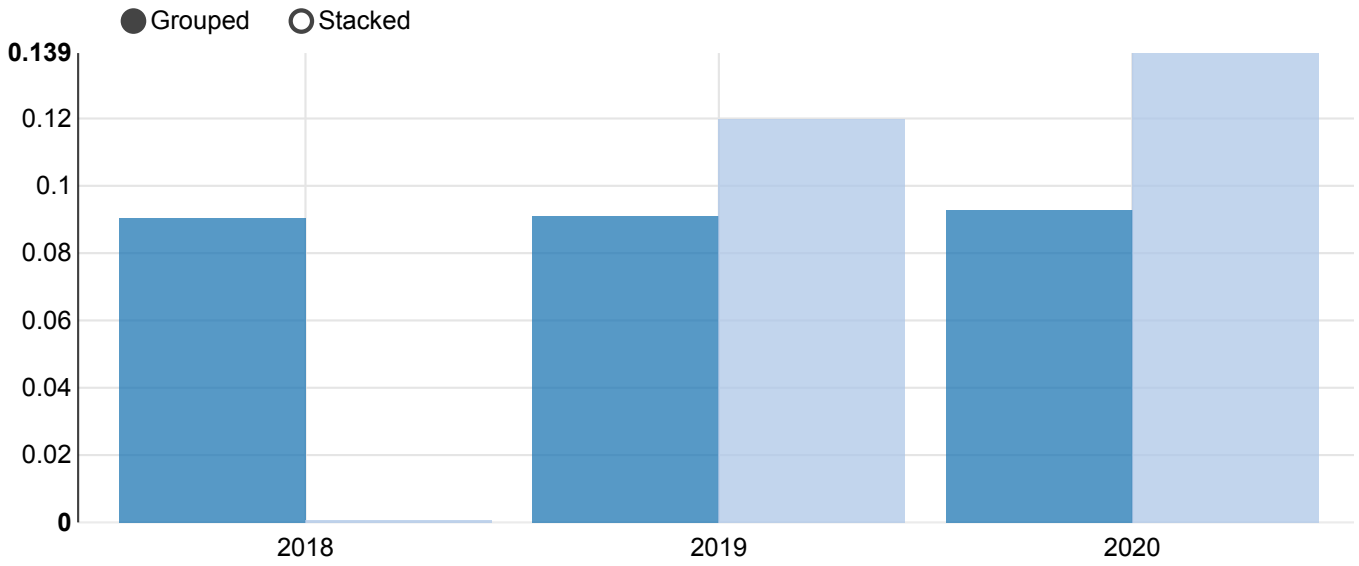
Year | N/RC_Index_Description | Staff_per_Pupil

keys | groups | values

Year ✖ | N/RC_Index_Description ✖ | Staff_per_Pupil **SUM** ✖

xAxis :

Default | Rotate | Hide




```
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, N/RC_Index_Description: string ...
1 more field]
```

Took 1 sec. Last updated by anonymous at December 10 2024, 9:43:02 AM. (outdated)

```
var groupedDf = df.groupBy("Year", "County_Name").agg(avg("Total_Funding").as("Total_Fu
z.show(groupedDf)
```

settings

Available Fields

Year | County_Name | Total_Funding

keys

Year ✖

groups

County_Name ✖

values

Total_Funding **SUM** ✖

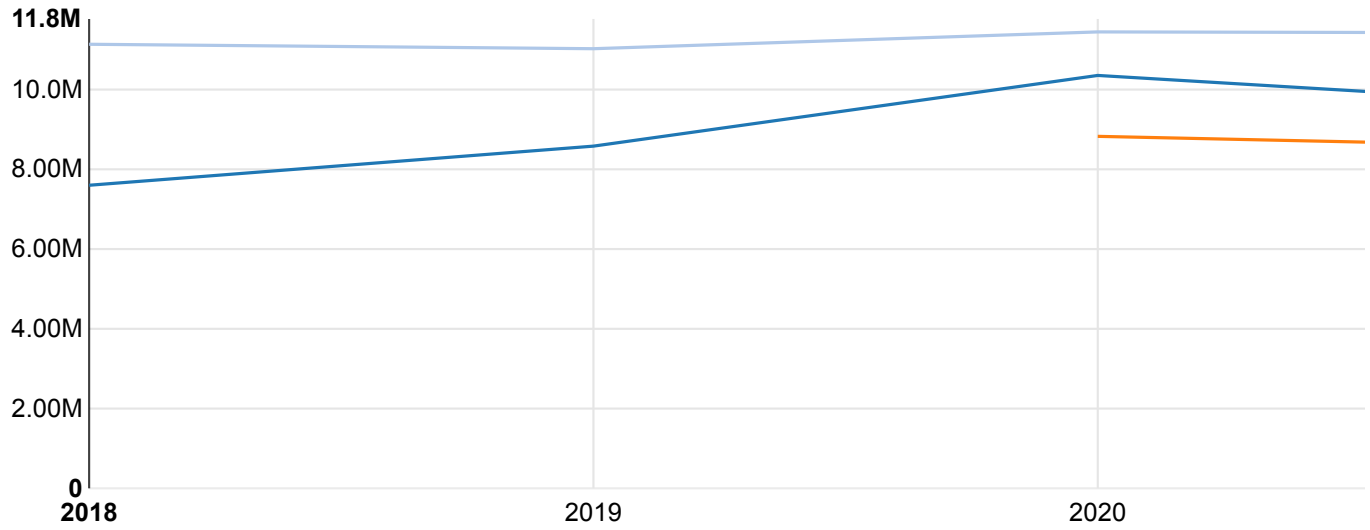
force Y to 0

zoom

Date format

xAxis :

Default | Rotate | Hide




```
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, County_Name: string ... 1 more fiel
d]
```

Took 1 sec. Last updated by anonymous at December 10 2024, 4:45:10 PM.

```
var groupedDf = df.groupBy("Year", "County_Name").agg(avg("Total_Enrollment").as("Total
z.show(groupedDf)
```

SPARK JOB FINISHED

settings

Available Fields

Year County_Name Total_Enrollment

keys: Year ✖

groups: County_Name ✖

values: Total_Enrollment **SUM** ✖

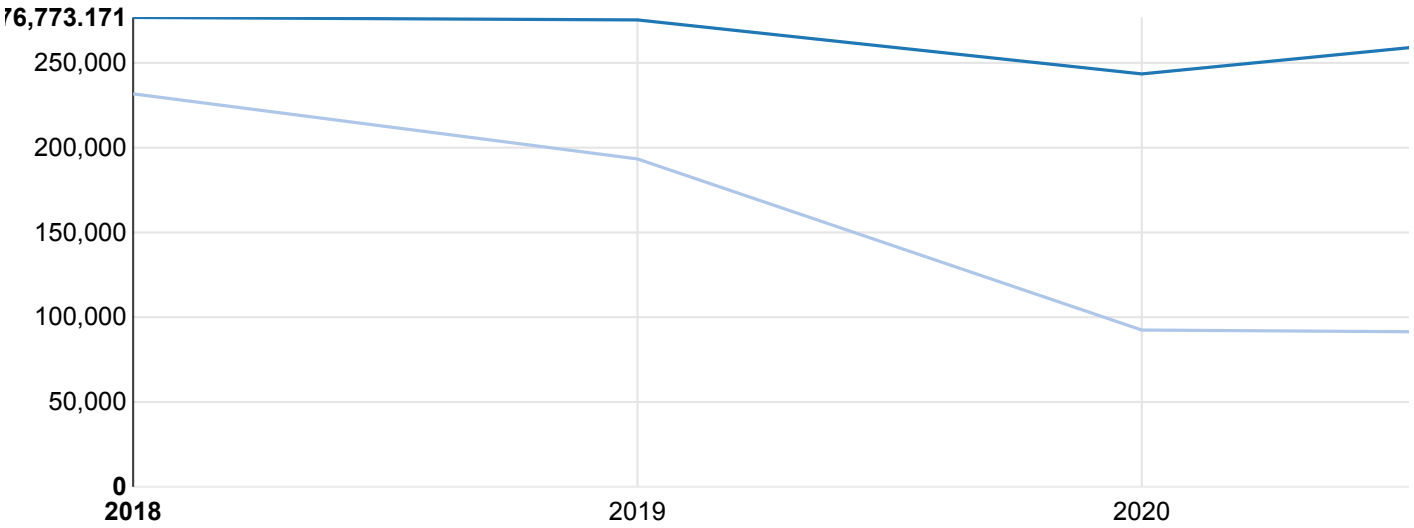
☑ **force Y to 0**

☐ **zoom**

☐ **Date format**

xAxis :

Default | Rotate | Hide




```
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, County_Name: string ... 1 more fiel
d]
```

Took 0 sec. Last updated by anonymous at December 10 2024, 10:50:24 AM.

```
var groupedDf = df.groupBy("Year", "County_Name").agg(avg("Total_Funding_per_Pupil").alias(
z.show(groupedDf)
```

SPARK JOB FINISHED

settings

Available Fields

Year | County_Name | Total_Funding_per_Pupil

keys: Year ✖

groups: County_Name ✖

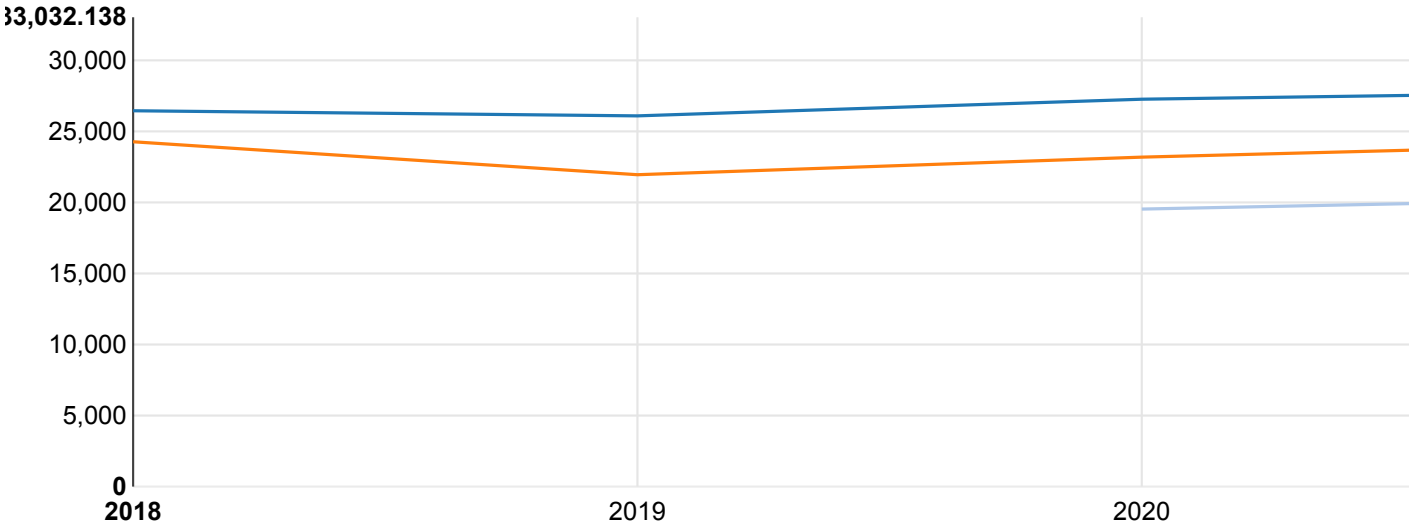
values: Total_Funding_per_Pupil SUM

☑ force Y to 0

☐ zoom

☐ Date format

xAxis : Default | Rotate | Hide



```
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, County_Name: string ... 1 more fiel
d]
```

Took 0 sec. Last updated by anonymous at December 10 2024, 4:45:22 PM.

```
var groupedDf = df.groupBy("Year").agg(avg("Total_Funding").alias("Total_Funding"))
z.show(groupedDf)
```

FINISHED

settings

Available Fields

Year Total_Funding

keys: Year ✖

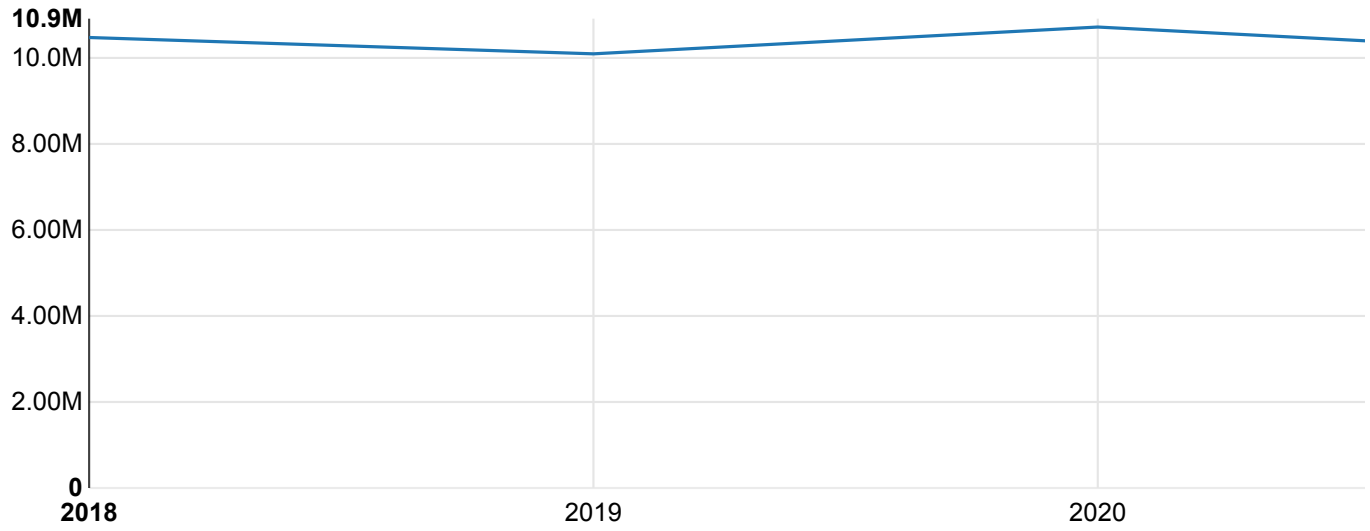
groups:

values: Total_Funding **SUM** ✖

- [x] **force Y to 0**
- [ ] **zoom**
- [ ] **Date format**

xAxis :

Default | Rotate | Hide




```
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, Total_Funding: decimal(24,6)]
```

Took 0 sec. Last updated by anonymous at December 10 2024, 10:53:22 AM. (outdated)

```
var groupedDf = df.groupBy("Year").agg(avg("Total_Funding_per_Pupil").as("Total_Funding"))
z.show(groupedDf)
```

SPARK JOB FINISHED

settings

Available Fields

Year | Total_Funding_per_Pupil

keys: Year

groups:

values: Total_Funding_per_Pupil SUM

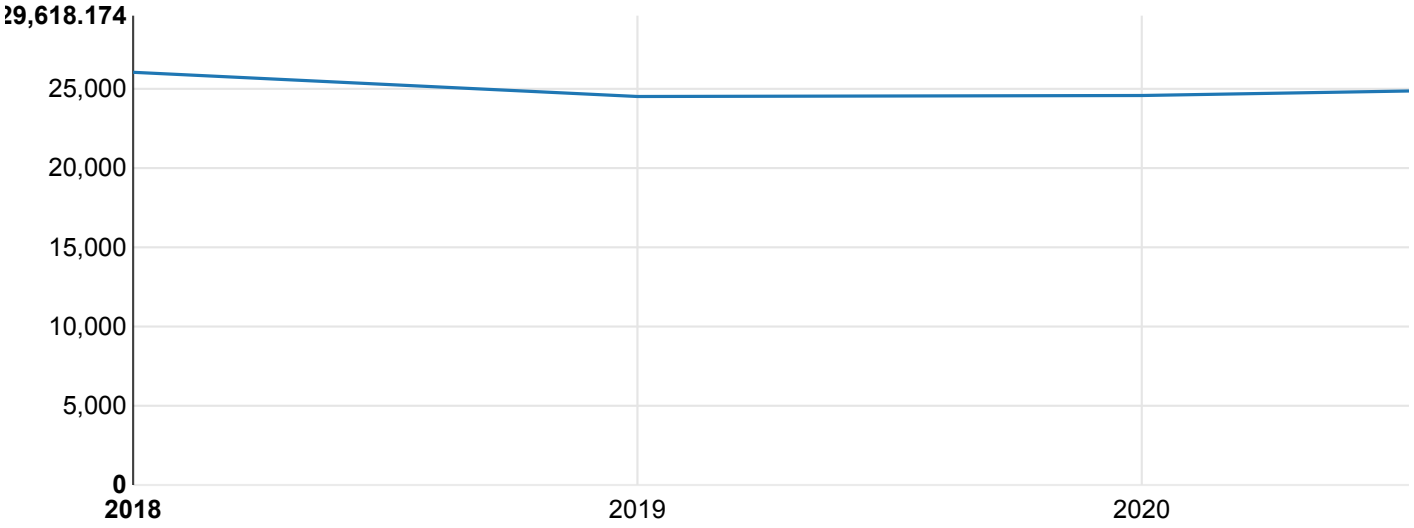
force Y to 0

zoom

Date format

xAxis :

Default | Rotate | Hide




```
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, Total_Funding_per_Pupil: decimal(24
,6)]
```

Took 1 sec. Last updated by anonymous at December 10 2024, 10:58:59 AM. (outdated)

```
var groupedDf = df.groupBy("Year").agg(avg("State_&_Local_Funding").a
z.show(groupedDf)
```

settings

Available Fields

Year | State_&_Local_Funding | Federal_Funding

keys | groups | values

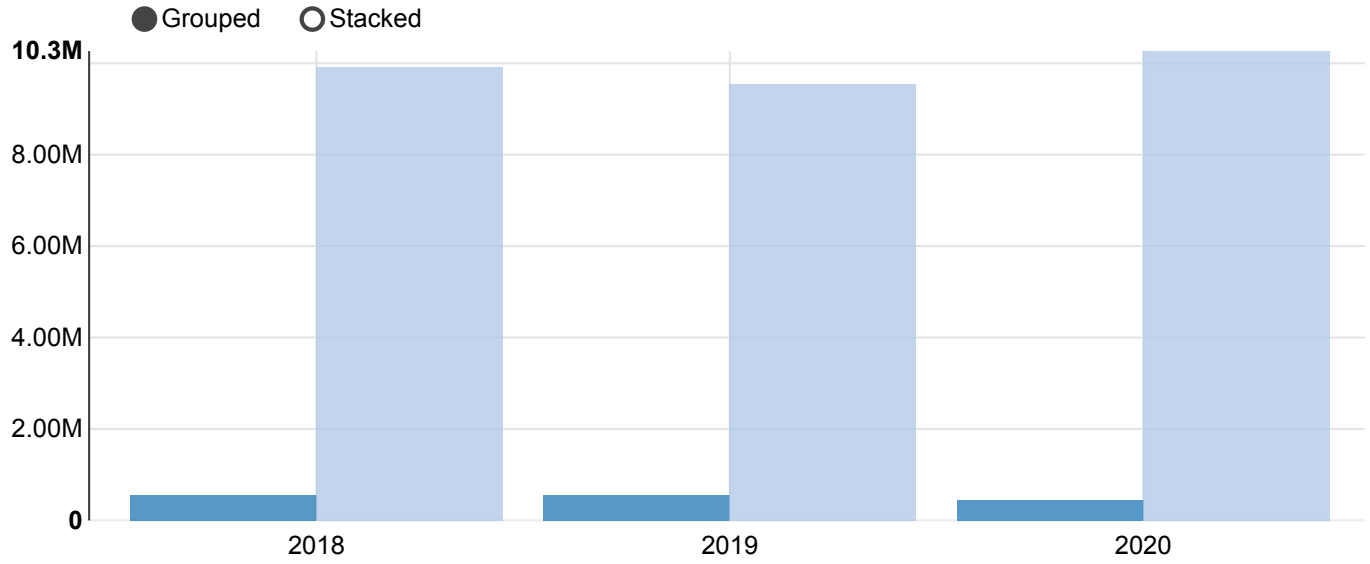
Year ✖

Federal_Funding SUM ✖

State_&_Local_Funding SUM

xAxis :

Default | Rotate | Hide




```
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, State_&_Local_Funding: decimal(24,6
) ... 1 more field]
```

Took 0 sec. Last updated by anonymous at December 10 2024, 11:00:39 AM. (outdated)

```
var df = spark.read.parquet("/user/yx3494_nyu_edu/scr_data/funding_safety_SPARK JOB FINISHED
val cols: List[String] = List("N/RC_Index_Description", "N/RC_Index")
df = df.filter(df("N/RC_Index_Description").isNotNull)
df = df.withColumn("N/RC_Index",
    when(df("N/RC_Index") === 2, 1)
    .when(df("N/RC_Index") === 3, 1)
    .when(df("N/RC_Index") === 4, 1)
    .when(df("N/RC_Index") === 5, 2)
    .when(df("N/RC_Index") === 6, 2)
   .otherwise(df("N/RC_Index")))

df = df.withColumn("N/RC_Index_Description",
    when(df("N/RC_Index") === 1, "High (1-4)")
```

```
    .when(df("N/RC_Index") === 2, "Low (5-6)")
   .otherwise(df("N/RC_Index")))

df = df.withColumn("County_Name",
    when(df("County_Name") === "BRONX", "NYC")
    .when(df("County_Name") === "BROOKLYN", "NYC")
    .when(df("County_Name") === "NEW YORK", "NYC")
    .when(df("County_Name") === "NYC CENTRAL OFFICE", "NYC")
    .when(df("County_Name") === "RICHMOND", "NYC")
    .when(df("County_Name") === "QUEENS", "NYC")
    .when(df("County_Name") === "ALBANY", "ALBANY")
   .otherwise(df("County_Name")))

df = df.select(
    "School_BEDS_Code",
    "Year",
    "School_Name",
    "County_Name",
    "N/RC_Index",
    "N/RC_Index_Description",
    "Total_Enrollment",
    "Total_Funding",
    "Total_Funding_per_Pupil",
    "Federal_Funding",
    "Federal_Funding_per_Pupil",
    "State_&_Local_Funding",
    "State_&_Local_Funding_per_Pupil",
    "Total_Teachers",
    "Teacher_per_Pupil",
    "Total_Staff",
    "Staff_per_Pupil"
    )

var groupedDf = df.groupBy("Year", "County_Name").agg(avg("Total_Funding_per_Pupil").alias(
z.show(groupedDf)
```

Available Fields

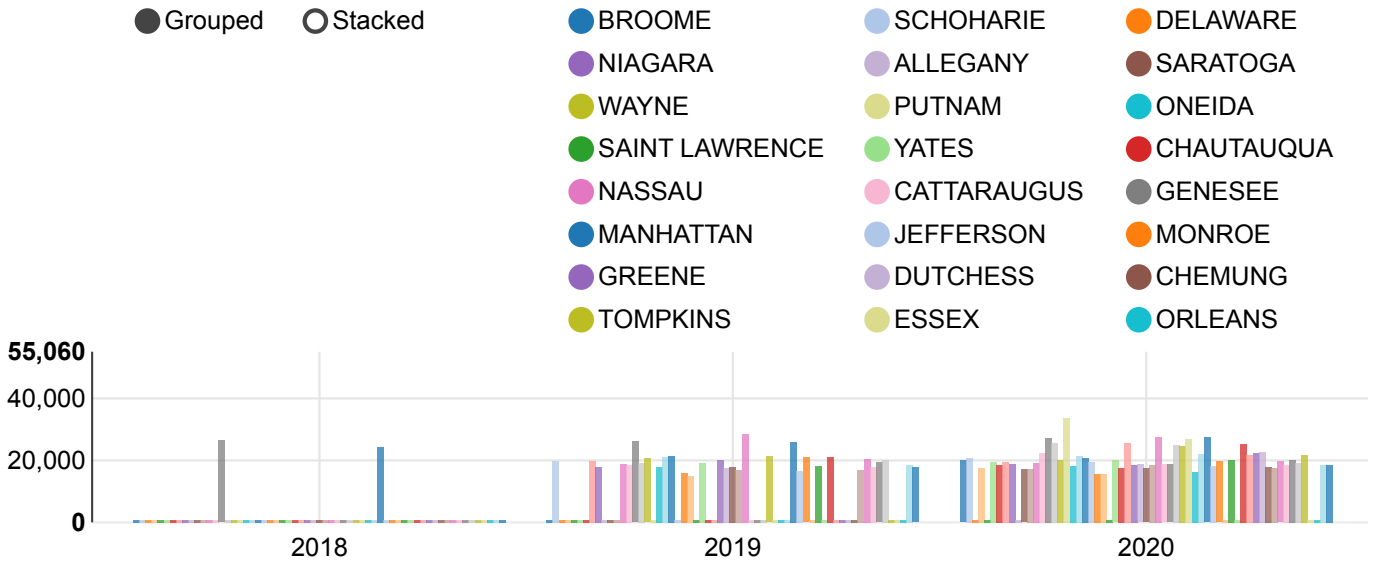
Year | County_Name | Total_Funding_per_Pupil

keys | groups | values

Year ✖ County_Name ✖ Total_Funding_per_Pupil SUM

xAxis :

Default | Rotate | Hide




```
df: org.apache.spark.sql.DataFrame = [School_BEDS_Code: string, Year: int ... 15 more field
s]
cols: List[String] = List(N/RC_Index_Description, N/RC_Index)
df: org.apache.spark.sql.DataFrame = [School_BEDS_Code: string, Year: int ... 15 more field
s]
df: org.apache.spark.sql.DataFrame = [School_BEDS_Code: string, Year: int ... 15 more field
s]
df: org.apache.spark.sql.DataFrame = [School_BEDS_Code: string, Year: int ... 15 more field
s]
df: org.apache.spark.sql.DataFrame = [School_BEDS_Code: string, Year: int ... 15 more field
s]
df: org.apache.spark.sql.DataFrame = [School_BEDS_Code: string, Year: int ... 15 more field
s]
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, County_Name: string ... 1 more fiel
d]
```

Took 1 sec. Last updated by anonymous at December 10 2024, 4:49:51 PM. (outdated)

```
var fundingDf = spark.read.parquet("schools-funding/2018-2023-schools-funding.parquet")
z.show(fundingDf.limit(5))
```

SPARK JOB FINISHED

settings

Year	School_BEDS_Code	District_BEDS_Code	School_Type	To
2022	10100010014	10100	Elementary School	78
2022	10100010016	10100	Elementary School	34
2022	10100010018	10100	Elementary School	21
2022	10100010019	10100	Elementary School	45
2022	10100010023	10100	Elementary School	37

```
fundingDf: org.apache.spark.sql.DataFrame = [Year: int, School_BEDS_Code: string ... 13 more fields]
```

Took 1 sec. Last updated by yl12043_nyu_edu at December 13 2024, 5:31:53 PM.

```
var groupedDf = fundingDf.groupBy("Year", "School_Type").agg(avg("Total_Funding_per_Pupil")
z.show(groupedDf)
```

SPARK JOB FINISHED

settings

Available Fields

Year School_Type Total_Funding_per_Pupil

keys groups values

Year ✖ School_Type ✖ Total_Funding_per_Pupil SUM

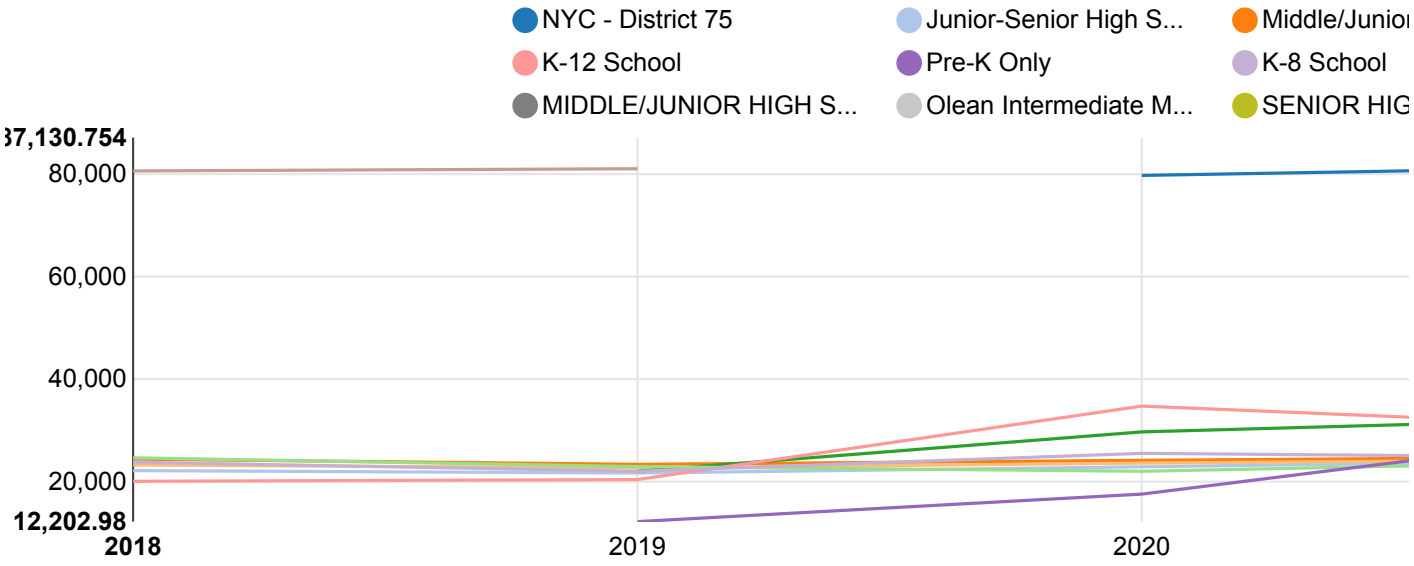
force Y to 0

zoom

Date format

xAxis :

Default | Rotate | Hide




```
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, School_Type: string ... 1 more fiel
d]
```

Took 1 sec. Last updated by yl12043_nyu_edu at December 13 2024, 5:32:44 PM. (outdated)

```
var groupedDf = fundingDf.groupBy("Year", "School_Type").agg(avg("Tea
z.show(groupedDf)
```

SPARK JOB FINISHED

settings

Available Fields

Year School_Type Teacher_per_Pupil

keys: Year ✖

groups: School_Type ✖

values: Teacher_per_Pupil **SUM** ✖

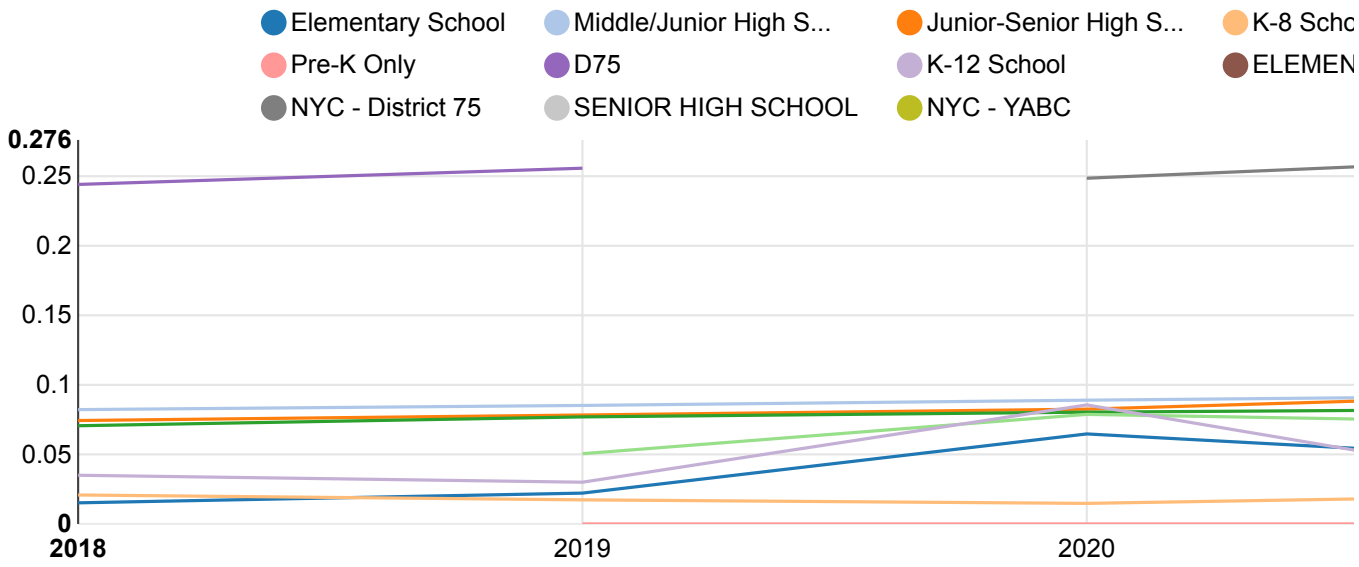
force Y to 0

zoom

Date format

xAxis :

Default | Rotate | Hide




```
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, School_Type: string ... 1 more fiel
d]
```

Took 0 sec. Last updated by yl12043_nyu_edu at December 13 2024, 5:34:32 PM. (outdated)

```
fundingDf = fundingDf.withColumn("Teacher_percentage_Total_Staff", (fu[illegible]
var groupedDf = fundingDf.groupBy("Year", "School_Type").agg(avg("Teacher_percentage_Total_
z.show(groupedDf)
```

settings

Available Fields

Year | School_Type | Teacher_percentage_Total_Staff

keys	groups	values
Year ✖	School_Type ✖	Teacher_percentage_Total_Sta

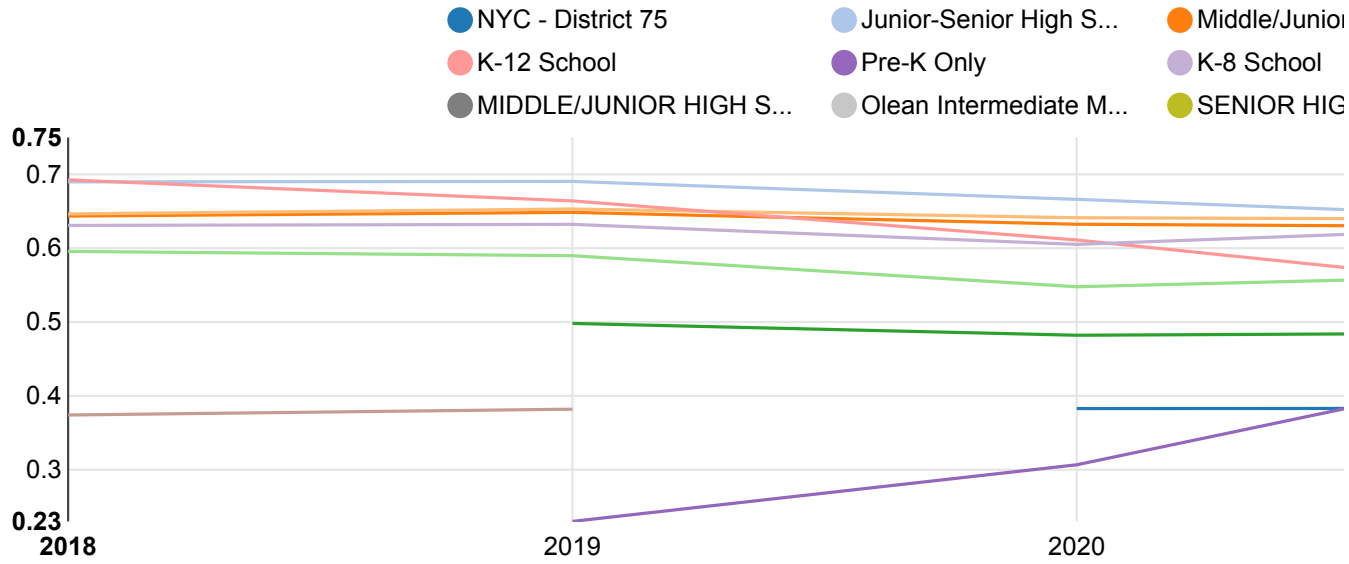
force Y to 0

zoom

Date format

xAxis :

Default | Rotate | Hide



```
fundingDf: org.apache.spark.sql.DataFrame = [Year: int, School_BEDS_Code: string ... 14 mor
e fields]
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, School_Type: string ... 1 more fiel
d]
```

Took 1 sec. Last updated by yl12043_nyu_edu at December 13 2024, 5:36:58 PM. (outdated)

```
var groupedDf = fundingDf.groupBy("Year").agg(avg("Teacher_percentage_Total_Staff"))
z.show(groupedDf)
```

Available Fields

Year Teacher_percentage_Total_Staff

keys: Year ✖

groups:

values: Teacher_percentage_Total_Sta

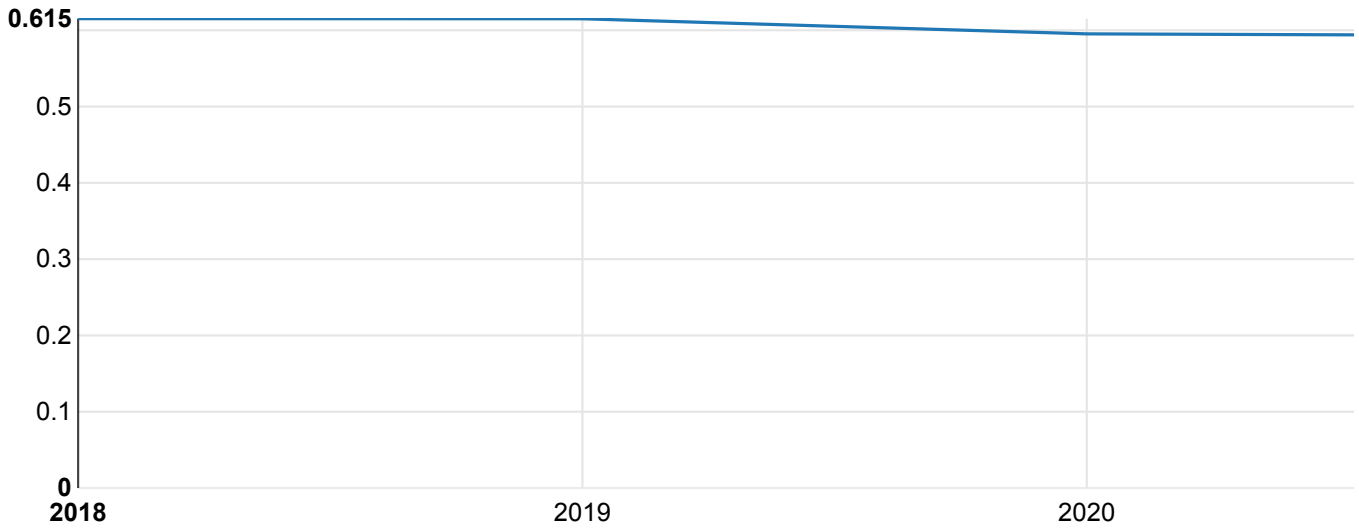
☑ **force Y to 0**

☐ **zoom**

☐ **Date format**

xAxis :

Default | Rotate | Hide




```
groupedDf: org.apache.spark.sql.DataFrame = [Year: int, Teacher_percentage_Total_Staff: dec
imal(24,6)]
```

Took 0 sec. Last updated by yl12043_nyu_edu at December 13 2024, 5:39:19 PM. (outdated)

READY