Settings

correlation_analysis.ipynb

Interpreter binding

Clicking the restart icon before the interpreter can restart interpreter associated with this note when your interpreter is isolated per note.
Drag and drop to reorder interpreters. The first interpreter on the list becomes default.
To create/remove interpreters, go to Interpreter menu.

spark %spark (default), %sql, %pyspark, %ipyspark, %r, %ir, %shiny

Save Cancel

FINISHED

SPARK JOB FINISHED

```
val filePath = "/user/yx3494_nyu_edu/scr_data/funding_safety.parquet"
val df = spark.read.parquet(filePath)

z.show(df)

df.createOrReplaceTempView("all")
```

settings ▾

School_BEDS_Code	Year	County_Name	District_Name	School_Name	Homocide	Sexual_Offense
10100010016	2022	ALBANY	ALBANY CITY SCHOOL DISTRICT	PINE HILLS ELEMENTARY SCHOOL	0	0
10100010018	2022	ALBANY	ALBANY CITY SCHOOL DISTRICT	DELAWARE COMMUNITY SCHOOL	0	0
10100010019	2022	ALBANY	ALBANY CITY SCHOOL DISTRICT	NEW SCOTLAND ELEMENTARY SCHOOL	0	0
10100010023	2022	ALBANY	ALBANY CITY SCHOOL DISTRICT	ALBANY SCHOOL OF HUMANITIES	0	0

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

Took 1 sec. Last updated by yz6956_nyu_edu at December 13 2024, 1:15:22 PM.

SPARK JOB FINISHED

```
val temp1 = spark.sql("""
    select count(distinct(School_BEDS_Code))
    from all
""")

z.show(temp1)
```

settings ▾

count(DISTINCT School_BEDS_Code)
4376

```
temp1: org.apache.spark.sql.DataFrame = [count(DISTINCT School_BEDS_Code): bigint]
```

Took 1 sec. Last updated by yz6956_nyu_edu at December 13 2024, 1:15:23 PM.

correlation_analysis.ipynb

FINISHED

```
val safetyIssueColumns = df.columns.slice(5,15)
```

```
safetyIssueColumns: Array[String] = Array(Homocide, Sexual_Offense, Assault, Weapons_Possession, Dignity_Act_Excluding_Cyberbullying, Dignity_Act
_Cyberbullying, Bomb_Threat, False_Alarm, Drugs, Alcohol)
```

Took 0 sec. Last updated by yz6956_nyu_edu at December 13 2024, 1:15:23 PM.

SPARK JOB (http://nyu-dataproc-w-1.c.hpc-dataproc-19b8.internal:37899/jobs/job?id=289) FINISHED

```
val fundingSafetyDF = df.withColumn(
  "Sum_Safety_Issues",
  safetyIssueColumns.map(colName => col(colName)).reduce(_ + _)
)

z.show(fundingSafetyDF)

fundingSafetyDF.createOrReplaceTempView("fundingSafety")
```

settings ▾

School_BEDS_Code	Year	County_Name	District_Name	School_Name	Homocide	Sexual_Offense
140203060005	2022	ERIE	WILLIAMSVILLE CENTRAL SCHOOL DISTRICT	HEIM ELEMENTARY SCHOOL	0	0
132101060008	2022	DUTCHESS	WAPPINGERS CENTRAL SCHOOL DISTRICT	SHEAFE ROAD ELEMENTARY SCHOOL	0	0
140203060002	2022	ERIE	WILLIAMSVILLE CENTRAL SCHOOL DISTRICT	DODGE ELEMENTARY SCHOOL	0	0
132101060003	2022	DUTCHESS	WAPPINGERS CENTRAL SCHOOL	FISHKILL PLAINS ELEMENTARY	0	0

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

×

Took 0 sec. Last updated by yz6956_nyu_edu at December 13 2024, 1:15:23 PM.

SPARK JOB FINISHED

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

// Step 1: Calculate min and max for each column
val stats = fundingSafetyDF.agg(
  min(col("Total_Funding").cast("double")).alias("Total_Funding_min"),
  max(col("Total_Funding").cast("double")).alias("Total_Funding_max"),
  min(col("Sum_Safety_Issues").cast("double")).alias("Sum_Safety_Issues_min"),
  max(col("Sum_Safety_Issues").cast("double")).alias("Sum_Safety_Issues_max")
).collect()(0)

// Extract min and max values as scalars
val totalFundingMin = stats.getAs[Double]("Total_Funding_min")
val totalFundingMax = stats.getAs[Double]("Total_Funding_max")
val sumSafetyIssuesMin = stats.getAs[Double]("Sum_Safety_Issues_min")
val sumSafetyIssuesMax = stats.getAs[Double]("Sum_Safety_Issues_max")

// Step 2: Normalize the columns using Min-Max Normalization
val normalizedData = fundingSafetyDF
  .withColumn("Total_Funding_normalized",
    (col("Total_Funding").cast("double") - lit(totalFundingMin)) / lit(totalFundingMax - totalFundingMin))
  .withColumn("Sum_Safety_Issues_normalized",
    (col("Sum_Safety_Issues").cast("double") - lit(sumSafetyIssuesMin)) / lit(sumSafetyIssuesMax - sumSafetyIssuesMin))

// Step 3: Compute correlation between normalized columns
val correlation = normalizedData.stat.corr("Total_Funding_normalized", "Sum_Safety_Issues_normalized")

// Print the correlation
println(s"Correlation between Total_Funding and Sum_Safety_Issues: $correlation")

// Show normalized data if needed
z.show(normalizedData)
```

```
Correlation between Total_Funding and Sum_Safety_Issues: 0.4013578506602246
```

settings ▾

School_BEDS_Code	Year	County_Name	District_Name	School_Name	Homocide	Sexual_Offense

10100010016	2022	ALBANY	ALBANY CITY SCHOOL DISTRICT	PINE HILLS ELEMENTARY SCHOOL	0	0
10100010018	2022	ALBANY	ALBANY CITY SCHOOL DISTRICT	DELAWARE COMMUNITY SCHOOL	0	0
10100010019	2022	ALBANY	ALBANY CITY SCHOOL DISTRICT	NEW SCOTLAND ELEMENTARY SCHOOL	0	0

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT** ×

Took 1 sec. Last updated by yz6956_nyu_edu at December 13 2024, 1:15:24 PM.

```
val filePath2 = "/user/yx3494_nyu_edu/scr_data/funding_safety_nrc_inexp_gradRate.parquet"
val df2 = spark.read.parquet(filePath2)

z.show(df2)

df2.createOrReplaceTempView("all2")
```

SPARK JOB FINISHED

settings ▾

School_BEDS_Code	Year	Graduation_Rate	County_Name	District_Name	Homocide	Sexual_Offense
10100010034	2022	78.99999999999999	ALBANY	ALBANY CITY SCHOOL DISTRICT	5	0
10201040001	2022	93.73333333333335	ALBANY	BERNE-KNOX-WESTERLO CENTRAL SCHOOL DISTRICT	0	0
10306060008	2022	95.83333333333333	ALBANY	BETHLEHEM CENTRAL SCHOOL DISTRICT	0	0
10402060001	2022	90.96666666666665	ALBANY	RAVENA-COEYMANS-	2	0

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT** ×

Took 1 sec. Last updated by yz6956_nyu_edu at December 13 2024, 1:15:25 PM.

```
val temp2 = spark.sql("""
    select count(distinct(School_BEDS_Code))
    from all2
""")

z.show(temp2)
```

SPARK JOB FINISHED

settings ▾

count(DISTINCT School_BEDS_Code)
1172

```
temp2: org.apache.spark.sql.DataFrame = [count(DISTINCT School_BEDS_Code): bigint]
```

Took 1 sec. Last updated by yz6956_nyu_edu at December 13 2024, 1:15:26 PM.

SPARK JOB (http://nyu-dataproc-w-1.c.hpc-dataproc-19b8.internal:37899/jobs/job?id=300) FINISHED

```
val fundingSafetyGradDF = df2.withColumn(
  "Sum_Safety_Issues",
  safetyIssueColumns.map(colName => col(colName)).reduce(_ + _))

z.show(fundingSafetyGradDF)

fundingSafetyGradDF.createOrReplaceTempView("fundingSafetyGrad")
```

settings ▾

School_BEDS_Code	Year	Graduation_Rate	County_Name	District_Name	Homocide	Sexual_Offense
10100010034	2022	78.99999999999999	ALBANY	ALBANY CITY SCHOOL DISTRICT	5	0
10201040001	2022	93.73333333333335	ALBANY	BERNE-KNOX-WESTERLO CENTRAL SCHOOL DISTRICT	0	0
10306060008	2022	95.83333333333333	ALBANY	BETHLEHEM CENTRAL SCHOOL DISTRICT	0	0
10402060001	2022	90.96666666666665	ALBANY	RAVENA-COEYMANS-	2	0

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

Took 0 sec. Last updated by yz6956_nyu_edu at December 13 2024, 1:15:26 PM.

SPARK JOB FINISHED

```
val temp3 = spark.sql("""
    select *
    from fundingSafetyGrad
    where Graduation_Rate < 40
""")

z.show(temp3)
```

settings ▾

School_BEDS_Code	Year	Graduation_Rate	County_Name	District_Name	Homocide	Sexual_Offense
140600010133	2022	21.7	ERIE	BUFFALO CITY SCHOOL DISTRICT	0	0
140600010316	2022	29.100000000000005	ERIE	BUFFALO CITY SCHOOL DISTRICT	0	0
310200011570	2022	21.833333333333332	NEW YORK	NEW YORK CITY GEOGRAPHIC DISTRICT # 2	0	0
310200011575	2022	33.0	NEW YORK	NEW YORK CITY GEOGRAPHIC DISTRICT # 2	0	0
310200011404	2022	39.233333333333334	NEW YORK	NEW YORK CITY	0	0

```
temp3: org.apache.spark.sql.DataFrame = [School_BEDS_Code: bigint, Year: int ... 32 more fields]
```

Took 1 sec. Last updated by yz6956_nyu_edu at December 13 2024, 1:15:27 PM.

SPARK JOB FINISHED

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

// Step 1: Calculate min and max for each column
val stats2 = fundingSafetyGradDF.agg(
  min(col("Total_Funding").cast("double")).alias("Total_Funding_min"),
  max(col("Total_Funding").cast("double")).alias("Total_Funding_max"),
  min(col("Sum_Safety_Issues").cast("double")).alias("Sum_Safety_Issues_min"),
  max(col("Sum_Safety_Issues").cast("double")).alias("Sum_Safety_Issues_max")
).collect()(0)

// Extract min and max values as scalars
val totalFundingMin2 = stats2.getAs[Double]("Total_Funding_min")
val totalFundingMax2 = stats2.getAs[Double]("Total_Funding_max")
val sumSafetyIssuesMin2 = stats2.getAs[Double]("Sum_Safety_Issues_min")
val sumSafetyIssuesMax2 = stats2.getAs[Double]("Sum_Safety_Issues_max")

// Step 2: Normalize the columns using Min-Max Normalization
val normalizedData2 = fundingSafetyGradDF
  .withColumn("Total_Funding_normalized",
    (col("Total_Funding").cast("double") - lit(totalFundingMin2)) / lit(totalFundingMax2 - totalFundingMin2))
  .withColumn("Sum_Safety_Issues_normalized",
    (col("Sum_Safety_Issues").cast("double") - lit(sumSafetyIssuesMin2)) / lit(sumSafetyIssuesMax2 - sumSafetyIssuesMin2))

// Step 3: Compute correlation between normalized columns
val correlation2 = normalizedData2.stat.corr("Total_Funding_normalized", "Sum_Safety_Issues_normalized")

// Print the correlation
println(s"Correlation between Total_Funding and Sum_Safety_Issues: ${correlation2}")

// Show normalized data if needed
z.show(normalizedData2)
```

```
Correlation between Total_Funding and Sum_Safety_Issues: 0.45725311057082835
```

settings

School_BEDS_Code	Year	Graduation_Rate	County_Name	District_Name	Homocide	Sexual_Offense
				DISTRICT		
60601040003	2022	76.8	CHAUTAUQUA	PINE VALLEY CENTRAL SCHOOL DISTRICT (SOUTH DAYTON)	0	0
60701040003	2022	96.15	CHAUTAUQUA	CLYMER CENTRAL SCHOOL DISTRICT	0	0
60800010009	2022	78.83333333333333	CHAUTAUQUA	DUNKIRK CITY SCHOOL DISTRICT	1	0
61001040005	2022	98.39999999999999	CHAUTAUQUA	BEMUS POINT CENTRAL SCHOOL	0	0

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

Took 1 sec. Last updated by yz6956_nyu_edu at December 13 2024, 1:15:28 PM.

SPARK JOB FINISHED

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

val stats3 = fundingSafetyGradDF.agg(
  min(col("Graduation_Rate").cast("double")).alias("Graduation_Rate_min"),
  max(col("Graduation_Rate").cast("double")).alias("Graduation_Rate_max")
).collect()(0)

val graduationRateMin = stats3.getAs[Double]("Graduation_Rate_min")
val graduationRateMax = stats3.getAs[Double]("Graduation_Rate_max")

val normalizedData3 = fundingSafetyGradDF
  .withColumn("Graduation_Rate_normalized",
    (col("Graduation_Rate").cast("double") - lit(graduationRateMin)) / lit(graduationRateMax - graduationRateMin))
  .withColumn("Sum_Safety_Issues_normalized",
    (col("Sum_Safety_Issues").cast("double") - lit(sumSafetyIssuesMin2)) / lit(sumSafetyIssuesMax2 - sumSafetyIssuesMin2))

val correlation3 = normalizedData3.stat.corr("Graduation_Rate_normalized", "Sum_Safety_Issues_normalized")

println(s"Correlation between Graduation_Rate and Sum_Safety_Issues: ${correlation3}")

z.show(normalizedData3)
```

```
Correlation between Graduation_Rate and Sum_Safety_Issues: 0.055747624233918054
```

settings

School_BEDS_Code	Year	Graduation_Rate	County_Name	District_Name	Homocide	Sexual_Offense
10100010034	2022	78.99999999999999	ALBANY	ALBANY CITY SCHOOL DISTRICT	5	0
10201040001	2022	93.73333333333335	ALBANY	BERNE-KNOX-WESTERLO CENTRAL SCHOOL DISTRICT	0	0
10306060008	2022	95.83333333333333	ALBANY	BETHLEHEM CENTRAL SCHOOL DISTRICT	0	0
10402060001	2022	90.96666666666665	ALBANY	RAVENA-COEYMANS-	2	0

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

Took 1 sec. Last updated by yz6956_nyu_edu at December 13 2024, 1:15:29 PM.

FINISHED

correlation_analysis.ipynb