Dataset Analysis

Table of Contents

Analysis of Aquaculture_Exports.csv

Statistic	Value
SOURCE_ID	count 369910.0 mean 63.0 std 0.0 min 63.0 25% 63.0 50% 63.0 75% 63.0 max 63.0 Name: SOURCE_ID, dtype: float64
HS_CODE	count 2.251200e+05 mean 6.280089e+08 std 5.618281e+08 min 3.011000e+08 25% 3.031900e+08 50% 3.061600e+08 75% 3.079900e+08 max 1.605906e+09 Name: HS_CODE, dtype: float64
GEOGRAPHY_CODE	count 369910.000000 mean 3344.115236 std 1999.328320 min 1.000000 25% 2010.000000 50% 3510.000000 75% 5081.000000 max 7990.000000 Name: GEOGRAPHY_CODE, dtype: float64

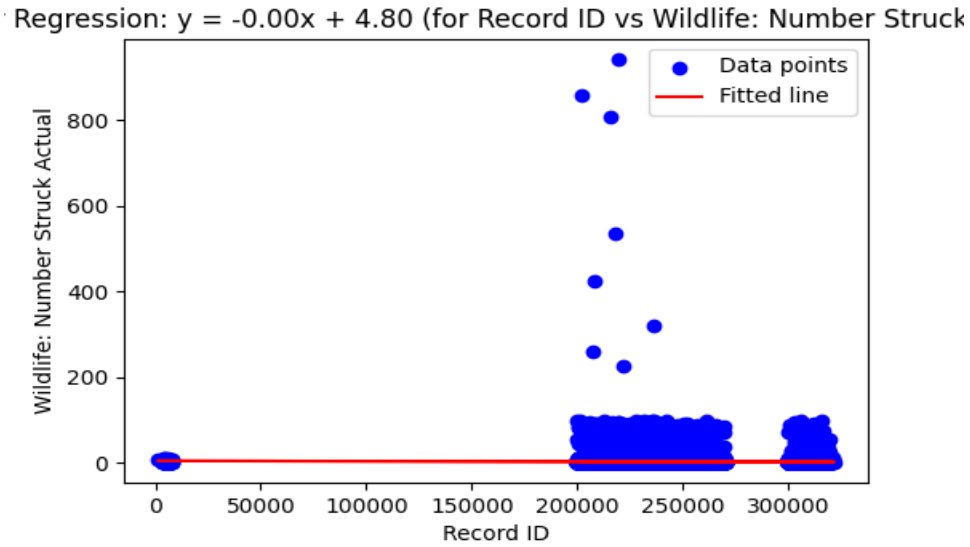
YEAR_ID	count 369910.000000 mean 2004.027904 std 7.777524 min 1989.000000 25% 1998.000000 50% 2005.000000 75% 2011.000000 max 2016.000000 Name: YEAR_ID, dtype: float64
TIMEPERIOD_ID	count 369910.000000 mean 6.506639 std 3.473561 min 1.000000 25% 3.000000 50% 7.000000 75% 10.000000 max 12.000000 Name: TIMEPERIOD_ID, dtype: float64
AMOUNT	count 3.699100e+05 mean 3.584079e+05 std 3.420339e+06 min 1.500000e+01 25% 4.082000e+03 50% 1.953500e+04 75% 9.813925e+04 max 3.214732e+08 Name: AMOUNT, dtype: float64

One or both columns have no variability, cannot perform linear regression.

Analysis of bsf.csv

Statistic	Value
Record ID	count 25558.000000 mean 253916.085609 std 38510.453382 min 1195.000000 25% 225783.750000 50% 248749.000000 75% 269168.750000 max 321909.000000 Name: Record ID, dtype: float64
Wildlife: Number Struck Actual	count 25558.000000 mean 2.691525 std 12.793975 min 1.000000 25% 1.000000 50% 1.000000 75% 1.000000 max 942.000000 Name: Wildlife: Number Struck Actual, dtype: float64
Number of people injured	count 25558.000000 mean 0.001056 std 0.050420 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 6.000000 Name: Number of people injured, dtype: float64

Linear Regression Equation: **y = -0.00x + 4.80 (for Record ID vs Wildlife: Number Struck Actual)**



Inferred Equations:

• **y = -0.00x + 4.80 (for Record ID vs Wildlife: Number Struck Actual)**

1. Collected data from 'Record ID' and 'Wildlife: Number Struck Actual'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as -0.00),

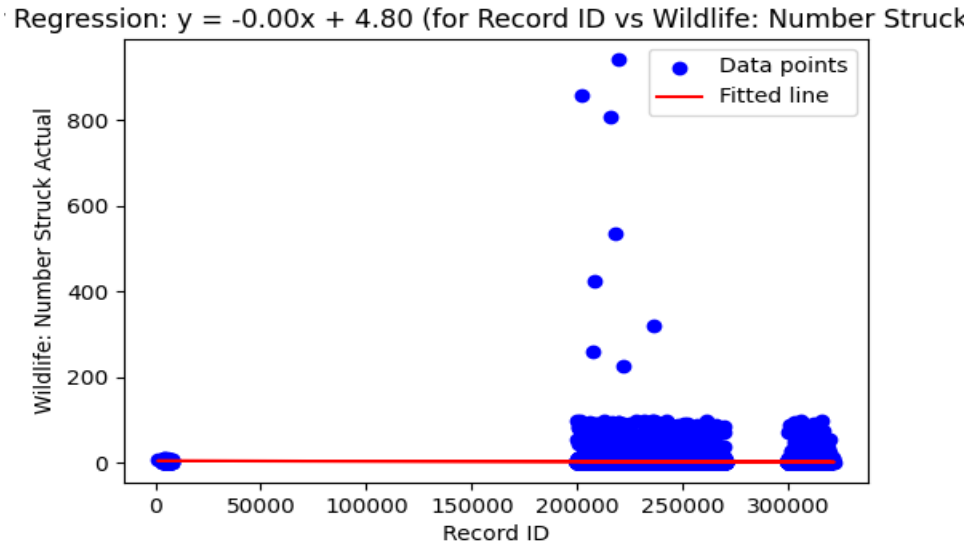
b = y-intercept (computed as 4.80).

Analysis of bsf.xlsx

Statistic	Value
Record ID	count 25558.000000 mean 253916.085609 min 1195.000000 25% 225783.750000 50% 248749.000000 75% 269168.750000 max 321909.000000 std 38510.453382 Name: Record ID, dtype: float64
Wildlife: Number Struck Actual	count 25558.000000 mean 2.691525 min 1.000000 25% 1.000000 50% 1.000000 75% 1.000000 max 942.000000 std 12.793975 Name: Wildlife: Number Struck Actual, dtype: float64
FlightDate	count 25429 mean 2007-01-22 13:20:40.017303040 min 2000-01-02 00:00:00 25% 2004-06-17 00:00:00 50% 2007-07-29 00:00:00 75% 2009-11-01 00:00:00 max 2011-12-31 00:00:00 std NaN Name: FlightDate, dtype: object

Cost: Total $	count 2.555800e+04 mean 5.567354e+03 min 0.000000e+00 25% 0.000000e+00 50% 0.000000e+00 75% 0.000000e+00 max 1.239775e+07 std 1.219713e+05 Name: Cost: Total $, dtype: float64
Feet above ground	count 25429.000000 mean 799.028432 min 0.000000 25% 0.000000 50% 50.000000 75% 700.000000 max 18000.000000 std 1740.079843 Name: Feet above ground, dtype: float64
Number of people injured	count 25558.000000 mean 0.001056 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 6.000000 std 0.050420 Name: Number of people injured, dtype: float64

Linear Regression Equation: **y = -0.00x + 4.80 (for Record ID vs Wildlife: Number Struck Actual)**



Inferred Equations:

• **y = -0.00x + 4.80 (for Record ID vs Wildlife: Number Struck Actual)**

1. Collected data from 'Record ID' and 'Wildlife: Number Struck Actual'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as -0.00),

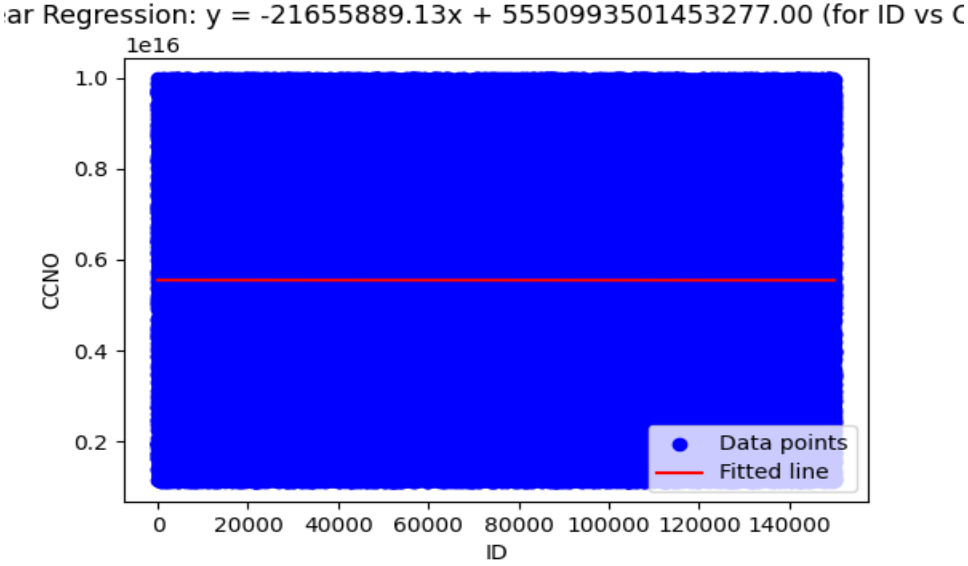
b = y-intercept (computed as 4.80).

Analysis of ccspend.csv

Statistic	Value
ID	count 150000.000000 mean 75000.500000 std 43301.414527 min 1.000000 25% 37500.750000 50% 75000.500000 75% 112500.250000 max 150000.000000 Name: ID, dtype: float64
CCNO	count 1.500000e+05 mean 5.549369e+15 std 2.563872e+15 min 1.111119e+15 25% 3.338369e+15 50% 5.552483e+15 75% 7.764596e+15 max 9.999993e+15 Name: CCNO, dtype: float64
TOTALSPEND%	count 150000.000000 mean 29.973973 std 11.842167 min 10.000000 25% 20.000000 50% 30.000000 75% 40.000000 max 50.000000 Name: TOTALSPEND%, dtype: float64

Max Limit	count 1.500000e+05 mean 2.503841e+06 std 1.442873e+06 min 1.000300e+04 25% 1.253584e+06 50% 2.505054e+06 75% 3.753341e+06 max 4.999994e+06 Name: Max Limit, dtype: float64
Amount Due	count 1.500000e+05 mean 7.505875e+05 std 5.517357e+05 min 1.010800e+03 25% 3.071280e+05 50% 6.266797e+05 75% 1.098379e+06 max 2.499746e+06 Name: Amount Due, dtype: float64

Linear Regression Equation: **y = -21655889.13x + 5550993501453277.00 (for ID vs CCNO)**



Inferred Equations:

• **y = -21655889.13x + 5550993501453277.00 (for ID vs CCNO)**

1. Collected data from 'ID' and 'CCNO'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as -21655889.13),

b = y-intercept (computed as 5550993501453277.00).

Analysis of county_population_by_race.csv

Statistic	Value
county	count 3222 unique 3222 top Autauga County, Alabama freq 1 Name: county, dtype: object
total_population_of_one_race	count 3222 unique 3143 top 6,602 freq 3 Name: total_population_of_one_race, dtype: object
total_population_of_one_race_white_alone	count 3222 unique 3144 top 8,080 freq 3 Name: total_population_of_one_race_white_alone, dtype: object
total_population_of_one_race_black_or_african_american_alone	count 3222 unique 2062 top 0 freq 30 Name: total_population_of_one_race_black_or_african_american_alone, dtype: object
otal_population_of_one_race_american_indian_and_alaska_native_alone	count 3222 unique 1207 top 37 freq 27 Name: total_population_of_one_race_american_indian_and_alaska_native_alone, dtype: objec
total_population_of_one_race_asian_alone	count 3222 unique 1276 top 0 freq 37 Name: total_population_of_one_race_asian_alone, dtype: object

_population_of_one_race_native_hawaiian_and_other_pacific_islander_alone	count 3222 unique 462 top 0 freq 446 Name: total_population_of_one_race_native_hawaiian_and_other_pacific_islander_alone, dtype: o
total_population_of_one_race_some_other_race_alone	count 3222 unique 1909 top 38 freq 14 Name: total_population_of_one_race_some_other_race_alone, dtype: object
total_population_of_two_or_more_races	count 3222 unique 2444 top 355 freq 7 Name: total_population_of_two_or_more_races, dtype: object

Not enough numerical data for analysis.

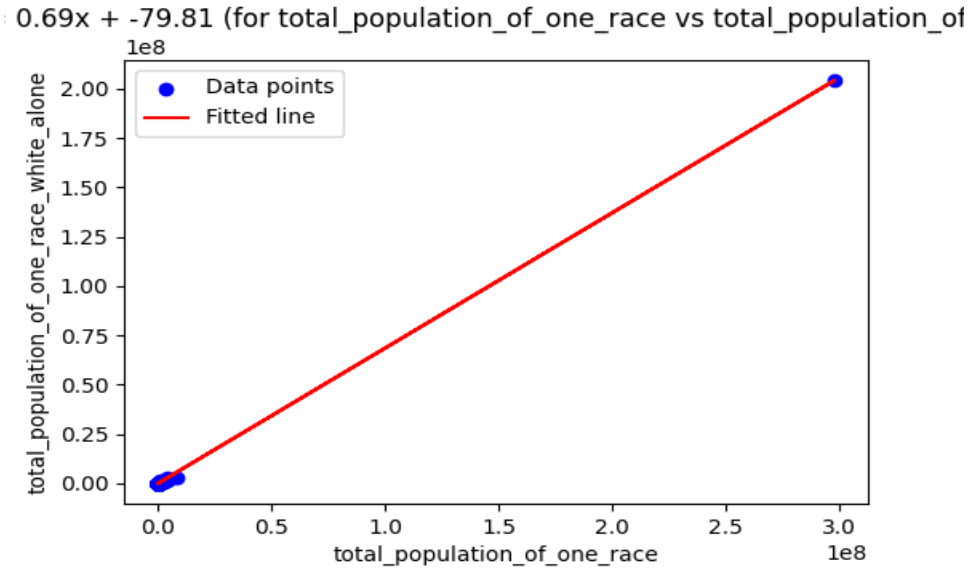
Analysis of county_population_by_race.xlsx

Statistic	Value
total_population_of_one_race	count 3.222000e+03 mean 1.852423e+05 std 5.249091e+06 min 5.800000e+01 25% 1.011475e+04 50% 2.372700e+04 75% 6.274000e+04 max 2.976003e+08 Name: total_population_of_one_race, dtype: float64
total_population_of_one_race_white_alone	count 3.222000e+03 mean 1.269755e+05 std 3.600920e+06 min 2.700000e+01 25% 7.771250e+03 50% 1.938200e+04 75% 5.252450e+04 max 2.042773e+08 Name: total_population_of_one_race_white_alone, dtype: float64
total_population_of_one_race_black_or_african_american_alone	count 3.222000e+03 mean 2.558570e+04 std 7.259532e+05 min 0.000000e+00 25% 9.000000e+01 50% 8.425000e+02 75% 5.305500e+03 max 4.110420e+07 Name: total_population_of_one_race_black_or_african_american_alone, dtype: float64

otal_population_of_one_race_american_indian_and_alaska_native_alone	count 3.222000e+03 mean 2.319100e+03 std 6.583857e+04 min 0.000000e+00 25% 4.800000e+01 50% 1.470000e+02 75% 5.432500e+02 max 3.727135e+06 Name: total_population_of_one_race_american_indian_and_alaska_native_alone, dtype: float6
total_population_of_one_race_asian_alone	count 3.222000e+03 mean 1.234516e+04 std 3.529591e+05 min 0.000000e+00 25% 3.400000e+01 50% 1.295000e+02 75% 7.912500e+02 max 1.988605e+07 Name: total_population_of_one_race_asian_alone, dtype: float64
_population_of_one_race_native_hawaiian_and_other_pacific_islander_alone	count 3222.000000 mean 428.468343 std 12351.096604 min 0.000000 25% 2.000000 50% 9.000000 75% 37.000000 max 689966.000000 Name: total_population_of_one_race_native_hawaiian_and_other_pacific_islander_alone, dtype: flo
total_population_of_one_race_some_other_race_alone	count 3.222000e+03 mean 1.758838e+04 std 4.957550e+05 min 0.000000e+00 25% 1.350000e+02 50% 5.400000e+02 75% 2.727250e+03 max 2.791572e+07 Name: total_population_of_one_race_some_other_race_alone, dtype: float64

total_population_of_two_or_more_races	count 3.222000e+03 mean 2.151883e+04 std 5.982174e+05 min 6.000000e+00 25% 5.232500e+02 50% 1.472000e+03 75% 5.524000e+03 max 3.384894e+07 Name: total_population_of_two_or_more_races, dtype: float64

Linear Regression Equation: **y = 0.69x + -79.81 (for total_population_of_one_race vs total_population_of_one_race_white_alone)**



Inferred Equations:

• **y = 0.69x + -79.81 (for total_population_of_one_race vs total_population_of_one_race_white_alone)**

1. Collected data from 'total_population_of_one_race' and 'total_population_of_one_race_white_alone'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as 0.69),

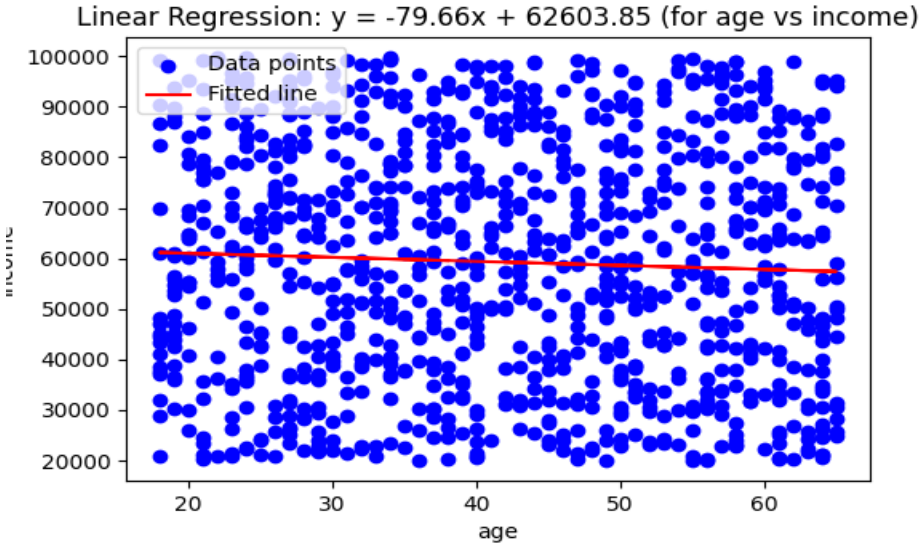
b = y-intercept (computed as -79.81).

Analysis of customer_data.csv

Statistic	Value
age	count 1000.000000 mean 41.754000 std 13.778582 min 18.000000 25% 30.000000 50% 42.000000 75% 54.000000 max 65.000000 Name: age, dtype: float64
income	count 1000.000000 mean 59277.852000 std 23258.377128 min 20031.000000 25% 38825.500000 50% 58972.000000 75% 79114.000000 max 99780.000000 Name: income, dtype: float64
purchase_frequency	count 1000.000000 mean 0.554600 std 0.284675 min 0.100000 25% 0.300000 50% 0.600000 75% 0.800000 max 1.000000 Name: purchase_frequency, dtype: float64

spending	count 1000.000000 mean 9613.296835 std 5484.707210 min 611.985000 25% 5020.425000 50% 9430.395000 75% 13645.507500 max 25546.500000 Name: spending, dtype: float64

Linear Regression Equation: **y = -79.66x + 62603.85 (for age vs income)**



Inferred Equations:

• **y = -79.66x + 62603.85 (for age vs income)**

1. Collected data from 'age' and 'income'.
2. Applied linear regression:

y = mx + b, where:

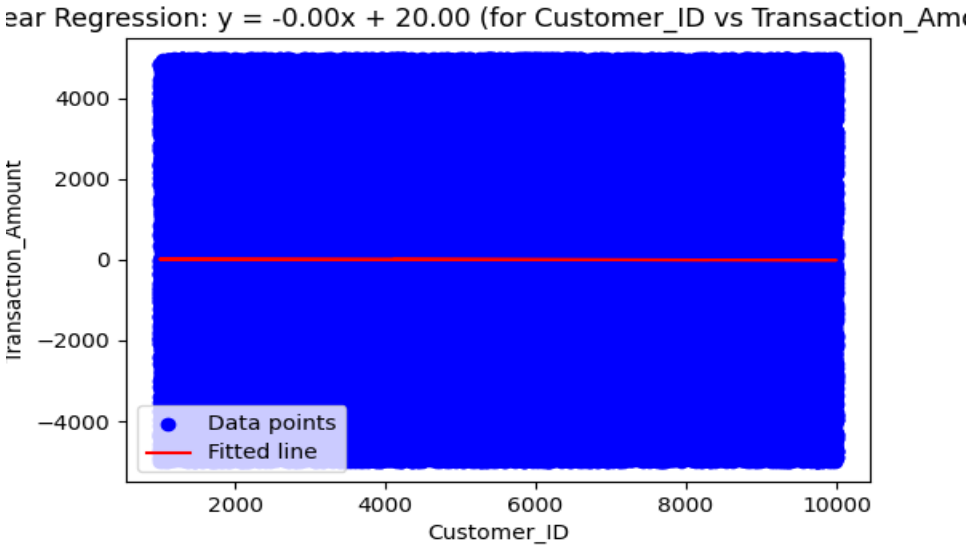
m = slope (computed as -79.66),

b = y-intercept (computed as 62603.85).

Analysis of customer_transactions.csv

Statistic	Value
Customer_ID	count 100000.000000 mean 5503.637790 std 2600.225771 min 1000.000000 25% 3251.000000 50% 5511.000000 75% 7750.000000 max 9998.000000 Name: Customer_ID, dtype: float64
Transaction_Amount	count 100000.000000 mean 5.542672 std 2886.631086 min -4999.980000 25% -2498.557500 50% 7.440000 75% 2501.370000 max 4999.840000 Name: Transaction_Amount, dtype: float64

Linear Regression Equation: **y = -0.00x + 20.00 (for Customer_ID vs Transaction_Amount)**



Inferred Equations:

• **y = -0.00x + 20.00 (for Customer_ID vs Transaction_Amount)**

1. Collected data from 'Customer_ID' and 'Transaction_Amount'.
2. Applied linear regression:

y = mx + b, where:

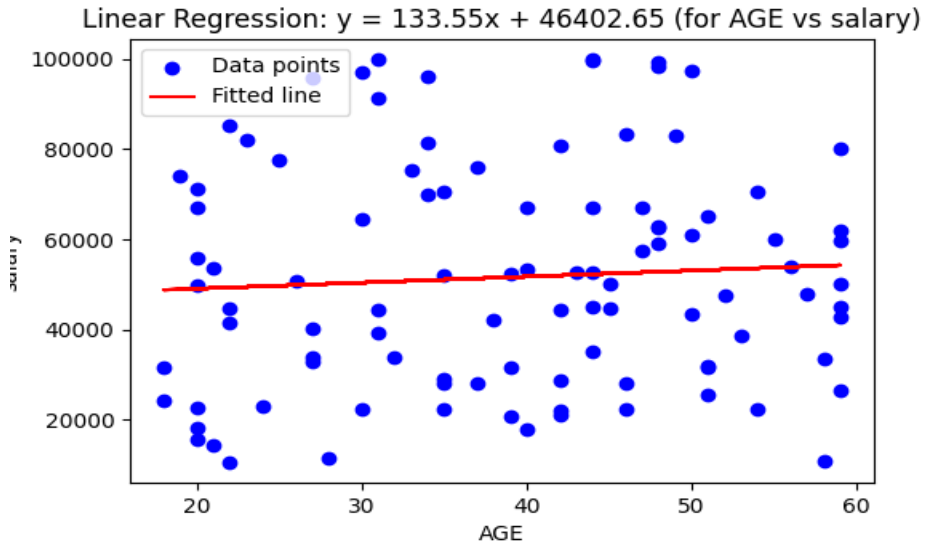
m = slope (computed as -0.00),

b = y-intercept (computed as 20.00).

Analysis of data.csv

Statistic	Value
AGE	count 99.000000 mean 38.949495 std 12.419022 min 18.000000 25% 29.000000 50% 40.000000 75% 48.000000 max 59.000000 Name: AGE, dtype: float64
salary	count 99.000000 mean 51604.515152 std 24764.866137 min 10581.000000 25% 31475.500000 50% 50165.000000 75% 68405.000000 max 99871.000000 Name: salary, dtype: float64

Linear Regression Equation: **y = 133.55x + 46402.65 (for AGE vs salary)**



Inferred Equations:

• **y = 133.55x + 46402.65 (for AGE vs salary)**

1. Collected data from 'AGE' and 'salary'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as 133.55),

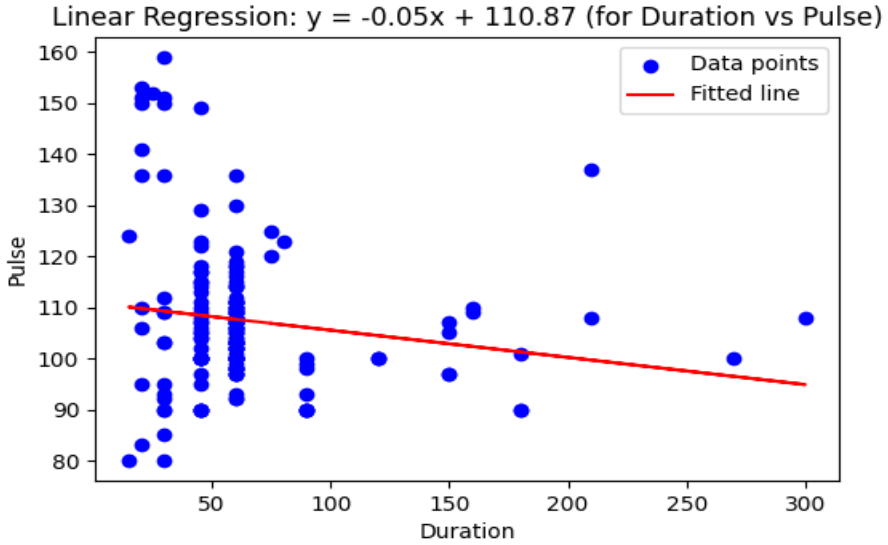
b = y-intercept (computed as 46402.65).

Analysis of data1.csv

Statistic	Value
Duration	count 169.000000 mean 63.846154 std 42.299949 min 15.000000 25% 45.000000 50% 60.000000 75% 60.000000 max 300.000000 Name: Duration, dtype: float64
Pulse	count 169.000000 mean 107.461538 std 14.510259 min 80.000000 25% 100.000000 50% 105.000000 75% 111.000000 max 159.000000 Name: Pulse, dtype: float64
Maxpulse	count 169.000000 mean 134.047337 std 16.450434 min 100.000000 25% 124.000000 50% 131.000000 75% 141.000000 max 184.000000 Name: Maxpulse, dtype: float64

	count 164.000000 mean 375.790244 std 266.379919 min 50.300000 25% 250.925000 50% 318.600000 75% 387.600000 max 1860.400000
Calories	Name: Calories, dtype: float64

Linear Regression Equation: **y = -0.05x + 110.87 (for Duration vs Pulse)**



Inferred Equations:

• **y = -0.05x + 110.87 (for Duration vs Pulse)**

1. Collected data from 'Duration' and 'Pulse'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as -0.05),

b = y-intercept (computed as 110.87).

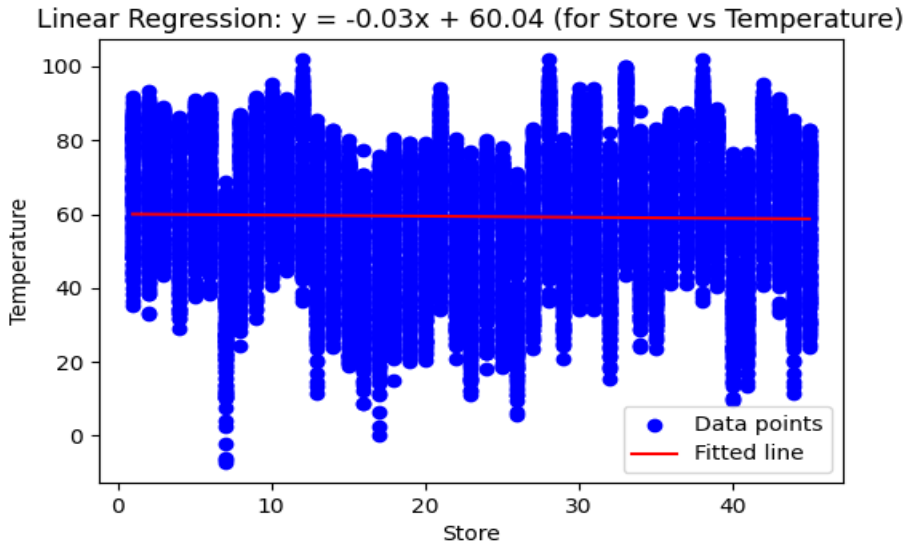
Analysis of Features data set.csv

Statistic	Value
Store	count 8190.000000 mean 23.000000 std 12.987966 min 1.000000 25% 12.000000 50% 23.000000 75% 34.000000 max 45.000000 Name: Store, dtype: float64
Temperature	count 8190.000000 mean 59.356198 std 18.678607 min -7.290000 25% 45.902500 50% 60.710000 75% 73.880000 max 101.950000 Name: Temperature, dtype: float64
Fuel_Price	count 8190.000000 mean 3.405992 std 0.431337 min 2.472000 25% 3.041000 50% 3.513000 75% 3.743000 max 4.468000 Name: Fuel_Price, dtype: float64

MarkDown1	count 4032.000000 mean 7032.371786 std 9262.747448 min -2781.450000 25% 1577.532500 50% 4743.580000 75% 8923.310000 max 103184.980000 Name: MarkDown1, dtype: float64
MarkDown2	count 2921.000000 mean 3384.176594 std 8793.583016 min -265.760000 25% 68.880000 50% 364.570000 75% 2153.350000 max 104519.540000 Name: MarkDown2, dtype: float64
MarkDown3	count 3613.000000 mean 1760.100180 std 11276.462208 min -179.260000 25% 6.600000 50% 36.260000 75% 163.150000 max 149483.310000 Name: MarkDown3, dtype: float64
MarkDown4	count 3464.000000 mean 3292.935886 std 6792.329861 min 0.220000 25% 304.687500 50% 1176.425000 75% 3310.007500 max 67474.850000 Name: MarkDown4, dtype: float64

MarkDown5	count 4050.000000 mean 4132.216422 std 13086.690278 min -185.170000 25% 1440.827500 50% 2727.135000 75% 4832.555000 max 771448.100000 Name: MarkDown5, dtype: float64
CPI	count 7605.000000 mean 172.460809 std 39.738346 min 126.064000 25% 132.364839 50% 182.764003 75% 213.932412 max 228.976456 Name: CPI, dtype: float64
Unemployment	count 7605.000000 mean 7.826821 std 1.877259 min 3.684000 25% 6.634000 50% 7.806000 75% 8.567000 max 14.313000 Name: Unemployment, dtype: float64

Linear Regression Equation: **y = -0.03x + 60.04 (for Store vs Temperature)**



Inferred Equations:

• **y = -0.03x + 60.04 (for Store vs Temperature)**

1. Collected data from 'Store' and 'Temperature'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as -0.03),

b = y-intercept (computed as 60.04).

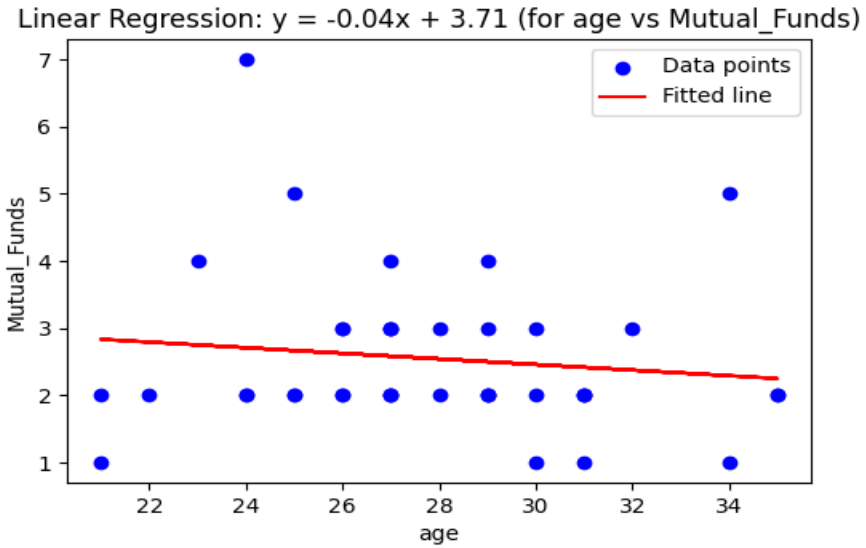
Analysis of Finance_data.csv

Statistic	Value
age	count 40.000000 mean 27.800000 std 3.560467 min 21.000000 25% 25.750000 50% 27.000000 75% 30.000000 max 35.000000 Name: age, dtype: float64
Mutual_Funds	count 40.000000 mean 2.550000 std 1.197219 min 1.000000 25% 2.000000 50% 2.000000 75% 3.000000 max 7.000000 Name: Mutual_Funds, dtype: float64
Equity_Market	count 40.000000 mean 3.475000 std 1.131994 min 1.000000 25% 3.000000 50% 4.000000 75% 4.000000 max 6.000000 Name: Equity_Market, dtype: float64

Debentures	count 40.000000 mean 5.750000 std 1.675617 min 1.000000 25% 5.000000 50% 6.500000 75% 7.000000 max 7.000000 Name: Debentures, dtype: float64
Government_Bonds	count 40.000000 mean 4.650000 std 1.369072 min 1.000000 25% 4.000000 50% 5.000000 75% 5.000000 max 7.000000 Name: Government_Bonds, dtype: float64
Fixed_Deposits	count 40.000000 mean 3.575000 std 1.795828 min 1.000000 25% 2.750000 50% 3.500000 75% 5.000000 max 7.000000 Name: Fixed_Deposits, dtype: float64
PPF	count 40.000000 mean 2.025000 std 1.609069 min 1.000000 25% 1.000000 50% 1.000000 75% 2.250000 max 6.000000 Name: PPF, dtype: float64

Gold	count 40.000000 mean 5.975000 std 1.143263 min 2.000000 25% 6.000000 50% 6.000000 75% 7.000000 max 7.000000 Name: Gold, dtype: float64

Linear Regression Equation: **y = -0.04x + 3.71 (for age vs Mutual_Funds)**



Inferred Equations:

• **y = -0.04x + 3.71 (for age vs Mutual_Funds)**

1. Collected data from 'age' and 'Mutual_Funds'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as -0.04),

b = y-intercept (computed as 3.71).

Analysis of hardcustomer_data.csv

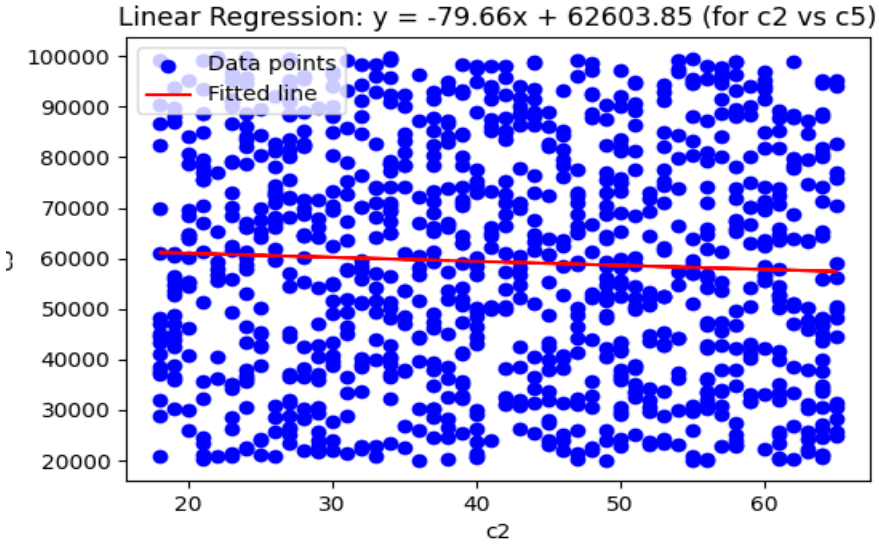
Statistic	Value
c2	count 1000.000000 mean 41.754000 std 13.778582 min 18.000000 25% 30.000000 50% 42.000000 75% 54.000000 max 65.000000 Name: c2, dtype: float64
c5	count 1000.000000 mean 59277.852000 std 23258.377128 min 20031.000000 25% 38825.500000 50% 58972.000000 75% 79114.000000 max 99780.000000 Name: c5, dtype: float64
c7	count 1000.000000 mean 0.554600 std 0.284675 min 0.100000 25% 0.300000 50% 0.600000 75% 0.800000 max 1.000000 Name: c7, dtype: float64

c8	count 1000.000000 mean 9613.296835 std 5484.707210 min 611.985000 25% 5020.425000 50% 9430.395000 75% 13645.507500 max 25546.500000 Name: c8, dtype: float64
Unnamed: 8	count 0.0 mean NaN std NaN min NaN 25% NaN 50% NaN 75% NaN max NaN Name: Unnamed: 8, dtype: float64
Unnamed: 9	count 0.0 mean NaN std NaN min NaN 25% NaN 50% NaN 75% NaN max NaN Name: Unnamed: 9, dtype: float64
Unnamed: 10	count 0.0 mean NaN std NaN min NaN 25% NaN 50% NaN 75% NaN max NaN Name: Unnamed: 10, dtype: float64

Unnamed: 11	count 0.0 mean NaN std NaN min NaN 25% NaN 50% NaN 75% NaN max NaN Name: Unnamed: 11, dtype: float64
Unnamed: 12	count 0.0 mean NaN std NaN min NaN 25% NaN 50% NaN 75% NaN max NaN Name: Unnamed: 12, dtype: float64
Unnamed: 13	count 0.0 mean NaN std NaN min NaN 25% NaN 50% NaN 75% NaN max NaN Name: Unnamed: 13, dtype: float64
Unnamed: 14	count 0.0 mean NaN std NaN min NaN 25% NaN 50% NaN 75% NaN max NaN Name: Unnamed: 14, dtype: float64

Unnamed: 15	count 0.0 mean NaN std NaN min NaN 25% NaN 50% NaN 75% NaN max NaN Name: Unnamed: 15, dtype: float64

Linear Regression Equation: **y = -79.66x + 62603.85 (for c2 vs c5)**



Inferred Equations:

• **y = -79.66x + 62603.85 (for c2 vs c5)**

1. Collected data from 'c2' and 'c5'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as -79.66),

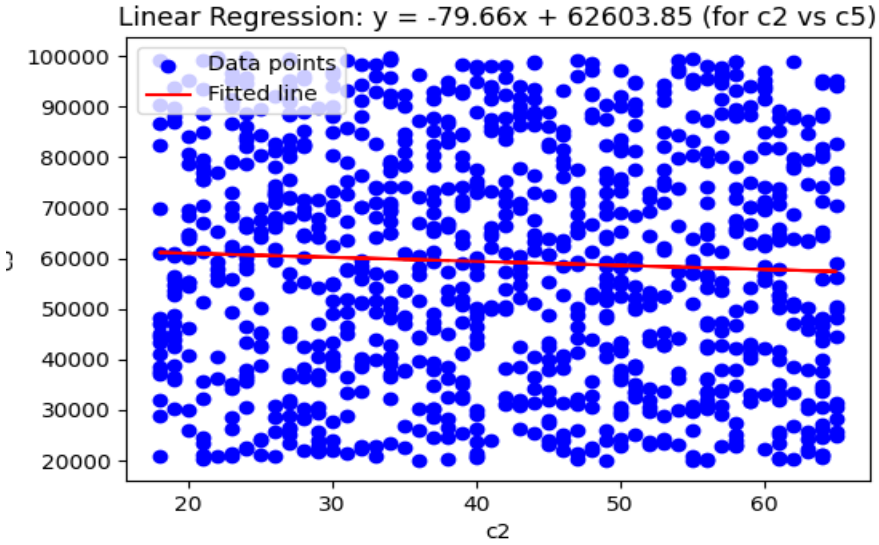
b = y-intercept (computed as 62603.85).

Analysis of hcdata2.csv

Statistic	Value
c2	count 1000.000000 mean 41.754000 std 13.778582 min 18.000000 25% 30.000000 50% 42.000000 75% 54.000000 max 65.000000 Name: c2, dtype: float64
c5	count 1000.000000 mean 59277.852000 std 23258.377128 min 20031.000000 25% 38825.500000 50% 58972.000000 75% 79114.000000 max 99780.000000 Name: c5, dtype: float64
c7	count 1000.000000 mean 0.554600 std 0.284675 min 0.100000 25% 0.300000 50% 0.600000 75% 0.800000 max 1.000000 Name: c7, dtype: float64

	count 1000.000000
	mean 9613.296835
	std 5484.707210
	min 611.985000
	25% 5020.425000
	50% 9430.395000
	75% 13645.507500
	max 25546.500000
c8	Name: c8, dtype: float64

Linear Regression Equation: **y = -79.66x + 62603.85 (for c2 vs c5)**



Inferred Equations:

• **y = -79.66x + 62603.85 (for c2 vs c5)**

1. Collected data from 'c2' and 'c5'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as -79.66),

b = y-intercept (computed as 62603.85).

Analysis of hotel.csv

Statistic	Value
is_canceled	count 119390.000000 mean 0.370416 std 0.482918 min 0.000000 25% 0.000000 50% 0.000000 75% 1.000000 max 1.000000 Name: is_canceled, dtype: float64
lead_time	count 119390.000000 mean 104.011416 std 106.863097 min 0.000000 25% 18.000000 50% 69.000000 75% 160.000000 max 737.000000 Name: lead_time, dtype: float64
arrival_date_year	count 119390.000000 mean 2016.156554 std 0.707476 min 2015.000000 25% 2016.000000 50% 2016.000000 75% 2017.000000 max 2017.000000 Name: arrival_date_year, dtype: float64

arrival_date_week_number	count 119390.000000 mean 27.165173 std 13.605138 min 1.000000 25% 16.000000 50% 28.000000 75% 38.000000 max 53.000000 Name: arrival_date_week_number, dtype: float64
arrival_date_day_of_month	count 119390.000000 mean 15.798241 std 8.780829 min 1.000000 25% 8.000000 50% 16.000000 75% 23.000000 max 31.000000 Name: arrival_date_day_of_month, dtype: float64
stays_in_weekend_nights	count 119390.000000 mean 0.927599 std 0.998613 min 0.000000 25% 0.000000 50% 1.000000 75% 2.000000 max 19.000000 Name: stays_in_weekend_nights, dtype: float64
stays_in_week_nights	count 119390.000000 mean 2.500302 std 1.908286 min 0.000000 25% 1.000000 50% 2.000000 75% 3.000000 max 50.000000 Name: stays_in_week_nights, dtype: float64

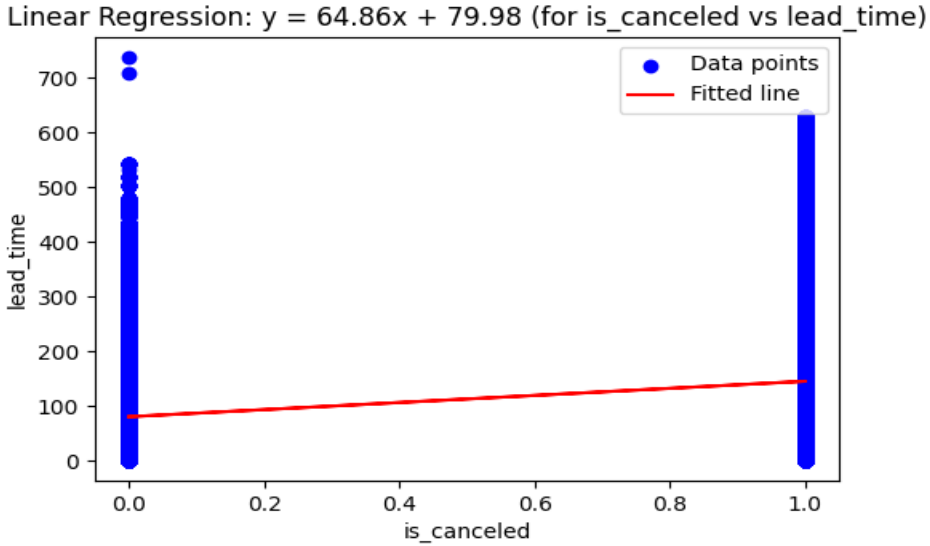
adults	count 119390.000000 mean 1.856403 std 0.579261 min 0.000000 25% 2.000000 50% 2.000000 75% 2.000000 max 55.000000 Name: adults, dtype: float64
children	count 119386.000000 mean 0.103890 std 0.398561 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 10.000000 Name: children, dtype: float64
babies	count 119390.000000 mean 0.007949 std 0.097436 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 10.000000 Name: babies, dtype: float64
is_repeated_guest	count 119390.000000 mean 0.031912 std 0.175767 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 1.000000 Name: is_repeated_guest, dtype: float64

previous_cancellations	count 119390.000000 mean 0.087118 std 0.844336 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 26.000000 Name: previous_cancellations, dtype: float64
previous_bookings_not_canceled	count 119390.000000 mean 0.137097 std 1.497437 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 72.000000 Name: previous_bookings_not_canceled, dtype: float64
booking_changes	count 119390.000000 mean 0.221124 std 0.652306 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 21.000000 Name: booking_changes, dtype: float64
agent	count 103050.000000 mean 86.693382 std 110.774548 min 1.000000 25% 9.000000 50% 14.000000 75% 229.000000 max 535.000000 Name: agent, dtype: float64

company	count 6797.000000 mean 189.266735 std 131.655015 min 6.000000 25% 62.000000 50% 179.000000 75% 270.000000 max 543.000000 Name: company, dtype: float64
days_in_waiting_list	count 119390.000000 mean 2.321149 std 17.594721 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 391.000000 Name: days_in_waiting_list, dtype: float64
adr	count 119390.000000 mean 101.831122 std 50.535790 min -6.380000 25% 69.290000 50% 94.575000 75% 126.000000 max 5400.000000 Name: adr, dtype: float64
required_car_parking_spaces	count 119390.000000 mean 0.062518 std 0.245291 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 8.000000 Name: required_car_parking_spaces, dtype: float64

total_of_special_requests	count 119390.000000 mean 0.571363 std 0.792798 min 0.000000 25% 0.000000 50% 0.000000 75% 1.000000 max 5.000000 Name: total_of_special_requests, dtype: float64

Linear Regression Equation: **y = 64.86x + 79.98 (for is_canceled vs lead_time)**



Inferred Equations:

• **y = 64.86x + 79.98 (for is_canceled vs lead_time)**

1. Collected data from 'is_canceled' and 'lead_time'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as 64.86),

b = y-intercept (computed as 79.98).

Analysis of hotel_booking.xlsx

Statistic	Value
is_canceled	count 119390.000000 mean 0.370416 std 0.482918 min 0.000000 25% 0.000000 50% 0.000000 75% 1.000000 max 1.000000 Name: is_canceled, dtype: float64
lead_time	count 119390.000000 mean 104.011416 std 106.863097 min 0.000000 25% 18.000000 50% 69.000000 75% 160.000000 max 737.000000 Name: lead_time, dtype: float64
arrival_date_year	count 119390.000000 mean 2016.156554 std 0.707476 min 2015.000000 25% 2016.000000 50% 2016.000000 75% 2017.000000 max 2017.000000 Name: arrival_date_year, dtype: float64

arrival_date_week_number	count 119390.000000 mean 27.165173 std 13.605138 min 1.000000 25% 16.000000 50% 28.000000 75% 38.000000 max 53.000000 Name: arrival_date_week_number, dtype: float64
arrival_date_day_of_month	count 119390.000000 mean 15.798241 std 8.780829 min 1.000000 25% 8.000000 50% 16.000000 75% 23.000000 max 31.000000 Name: arrival_date_day_of_month, dtype: float64
stays_in_weekend_nights	count 119390.000000 mean 0.927599 std 0.998613 min 0.000000 25% 0.000000 50% 1.000000 75% 2.000000 max 19.000000 Name: stays_in_weekend_nights, dtype: float64
stays_in_week_nights	count 119390.000000 mean 2.500302 std 1.908286 min 0.000000 25% 1.000000 50% 2.000000 75% 3.000000 max 50.000000 Name: stays_in_week_nights, dtype: float64

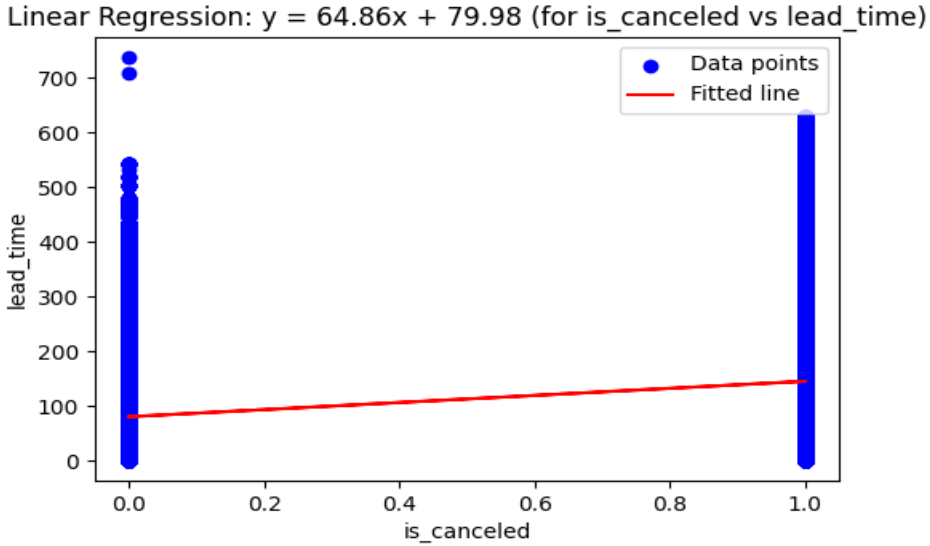
adults	count 119390.000000 mean 1.856403 std 0.579261 min 0.000000 25% 2.000000 50% 2.000000 75% 2.000000 max 55.000000 Name: adults, dtype: float64
children	count 119386.000000 mean 0.103890 std 0.398561 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 10.000000 Name: children, dtype: float64
babies	count 119390.000000 mean 0.007949 std 0.097436 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 10.000000 Name: babies, dtype: float64
is_repeated_guest	count 119390.000000 mean 0.031912 std 0.175767 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 1.000000 Name: is_repeated_guest, dtype: float64

previous_cancellations	count 119390.000000 mean 0.087118 std 0.844336 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 26.000000 Name: previous_cancellations, dtype: float64
previous_bookings_not_canceled	count 119390.000000 mean 0.137097 std 1.497437 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 72.000000 Name: previous_bookings_not_canceled, dtype: float64
booking_changes	count 119390.000000 mean 0.221124 std 0.652306 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 21.000000 Name: booking_changes, dtype: float64
agent	count 103050.000000 mean 86.693382 std 110.774548 min 1.000000 25% 9.000000 50% 14.000000 75% 229.000000 max 535.000000 Name: agent, dtype: float64

company	count 6797.000000 mean 189.266735 std 131.655015 min 6.000000 25% 62.000000 50% 179.000000 75% 270.000000 max 543.000000 Name: company, dtype: float64
days_in_waiting_list	count 119390.000000 mean 2.321149 std 17.594721 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 391.000000 Name: days_in_waiting_list, dtype: float64
adr	count 119390.000000 mean 101.831122 std 50.535790 min -6.380000 25% 69.290000 50% 94.575000 75% 126.000000 max 5400.000000 Name: adr, dtype: float64
required_car_parking_spaces	count 119390.000000 mean 0.062518 std 0.245291 min 0.000000 25% 0.000000 50% 0.000000 75% 0.000000 max 8.000000 Name: required_car_parking_spaces, dtype: float64

total_of_special_requests	count 119390.000000 mean 0.571363 std 0.792798 min 0.000000 25% 0.000000 50% 0.000000 75% 1.000000 max 5.000000 Name: total_of_special_requests, dtype: float64

Linear Regression Equation: **y = 64.86x + 79.98 (for is_canceled vs lead_time)**



Inferred Equations:

• **y = 64.86x + 79.98 (for is_canceled vs lead_time)**

1. Collected data from 'is_canceled' and 'lead_time'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as 64.86),

b = y-intercept (computed as 79.98).

Analysis of milkquotabycountry.csv

Statistic	Value
Unnamed: 0	count 22 unique 4 top ENGLAND freq 12 Name: Unnamed: 0, dtype: object
County	count 74 unique 74 top Avon freq 1 Name: County, dtype: object
Unnamed: 2	count 1023 unique 15 top 1994/95 freq 73 Name: Unnamed: 2, dtype: object
Unnamed: 3	count 1023 unique 989 top 16,568,188 freq 12 Name: Unnamed: 3, dtype: object
% Change	count 1023 unique 620 top - freq 73 Name: % Change, dtype: object
Unnamed: 5	count 1023 unique 800 top - freq 73 Name: Unnamed: 5, dtype: object

Net Quota	count 1022 unique 1000 top 2,763,984 freq 11 Name: Net Quota, dtype: object
% Change.1	count 1023 unique 684 top - freq 73 Name: % Change.1, dtype: object
Unnamed: 8	count 1023 unique 833 top - freq 73 Name: Unnamed: 8, dtype: object

Not enough numerical data for analysis.

Analysis of name_gender.csv

Statistic	Value
	count 95025.000000 mean 0.984792 std 0.066169 min 0.500000 25% 1.000000 50% 1.000000 75% 1.000000 max 1.000000
probability	Name: probability, dtype: float64

Not enough numerical data for analysis.

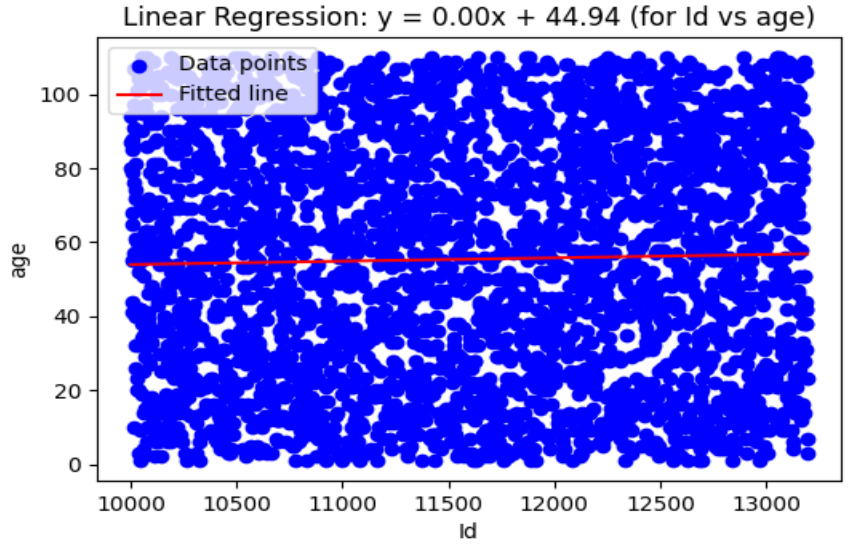
Analysis of senseessdata.csv

Statistic	Value
Id	count 3193.000000 mean 11597.000000 std 921.884031 min 10001.000000 25% 10799.000000 50% 11597.000000 75% 12395.000000 max 13193.000000 Name: Id, dtype: float64
age	count 3193.000000 mean 55.419042 std 31.927016 min 1.000000 25% 27.000000 50% 56.000000 75% 83.000000 max 110.000000 Name: age, dtype: float64
Income	count 3.193000e+03 mean 4.936105e+08 std 2.874596e+08 min 2.245560e+05 25% 2.460148e+08 50% 4.899553e+08 75% 7.411076e+08 max 9.997406e+08 Name: Income, dtype: float64

Spend	count 3193.000000 mean 47.393987 std 18.929946 min 15.000000 25% 31.000000 50% 47.000000 75% 64.000000 max 80.000000 Name: Spend, dtype: float64
Savings	count 3193.000000 mean 45.146884 std 8.959078 min 30.000000 25% 37.000000 50% 45.000000 75% 53.000000 max 60.000000 Name: Savings, dtype: float64
Debt	count 3193.000000 mean 5.602255 std 32.128798 min -50.000000 25% -22.000000 50% 6.000000 75% 34.000000 max 60.000000 Name: Debt, dtype: float64
Credit Rating	count 3193.000000 mean 4.961478 std 3.199421 min 0.000000 25% 2.000000 50% 5.000000 75% 8.000000 max 10.000000 Name: Credit Rating, dtype: float64

Unemployed For	count 3193.000000 mean 30.546195 std 16.967881 min 1.000000 25% 17.000000 50% 30.000000 75% 45.000000 max 60.000000 Name: Unemployed For, dtype: float64

Linear Regression Equation: **y = 0.00x + 44.94 (for Id vs age)**



Inferred Equations:

• **y = 0.00x + 44.94 (for Id vs age)**

1. Collected data from 'Id' and 'age'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as 0.00),

b = y-intercept (computed as 44.94).

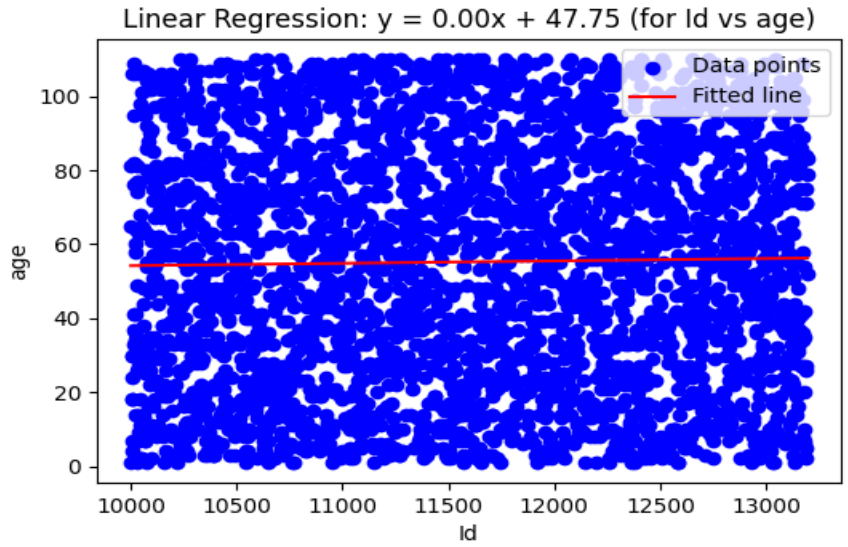
Analysis of ssns2.csv

Statistic	Value
Id	count 3193.000000 mean 11597.000000 std 921.884031 min 10001.000000 25% 10799.000000 50% 11597.000000 75% 12395.000000 max 13193.000000 Name: Id, dtype: float64
age	count 3193.00000 mean 55.23207 std 31.83426 min 1.00000 25% 28.00000 50% 56.00000 75% 82.00000 max 110.00000 Name: age, dtype: float64
Income	count 3.193000e+03 mean 5.005401e+08 std 2.880665e+08 min 8.168500e+04 25% 2.459586e+08 50% 5.081371e+08 75% 7.524832e+08 max 9.996266e+08 Name: Income, dtype: float64

Spend	count 3193.000000 mean 47.853743 std 19.204518 min 15.000000 25% 31.000000 50% 48.000000 75% 65.000000 max 80.000000 Name: Spend, dtype: float64
Savings	count 3193.000000 mean 45.009082 std 8.913198 min 30.000000 25% 38.000000 50% 45.000000 75% 53.000000 max 60.000000 Name: Savings, dtype: float64
Debt	count 3193.000000 mean 5.902286 std 31.563525 min -50.000000 25% -22.000000 50% 7.000000 75% 33.000000 max 60.000000 Name: Debt, dtype: float64
Credit Rating	count 3193.000000 mean 4.954588 std 3.185593 min 0.000000 25% 2.000000 50% 5.000000 75% 8.000000 max 10.000000 Name: Credit Rating, dtype: float64

Is Employable	count 3193.000000 mean 0.500157 std 0.500078 min 0.000000 25% 0.000000 50% 1.000000 75% 1.000000 max 1.000000 Name: Is Employable , dtype: float64
HasCriminalBackground	count 3193.000000 mean 0.500157 std 0.500078 min 0.000000 25% 0.000000 50% 1.000000 75% 1.000000 max 1.000000 Name: HasCriminalBackground, dtype: float64
Unemployed For	count 3193.000000 mean 30.352960 std 17.178384 min 1.000000 25% 16.000000 50% 30.000000 75% 45.000000 max 60.000000 Name: Unemployed For, dtype: float64

Linear Regression Equation: **y = 0.00x + 47.75 (for Id vs age)**



Inferred Equations:

• **y = 0.00x + 47.75 (for Id vs age)**

1. Collected data from 'Id' and 'age'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as 0.00),

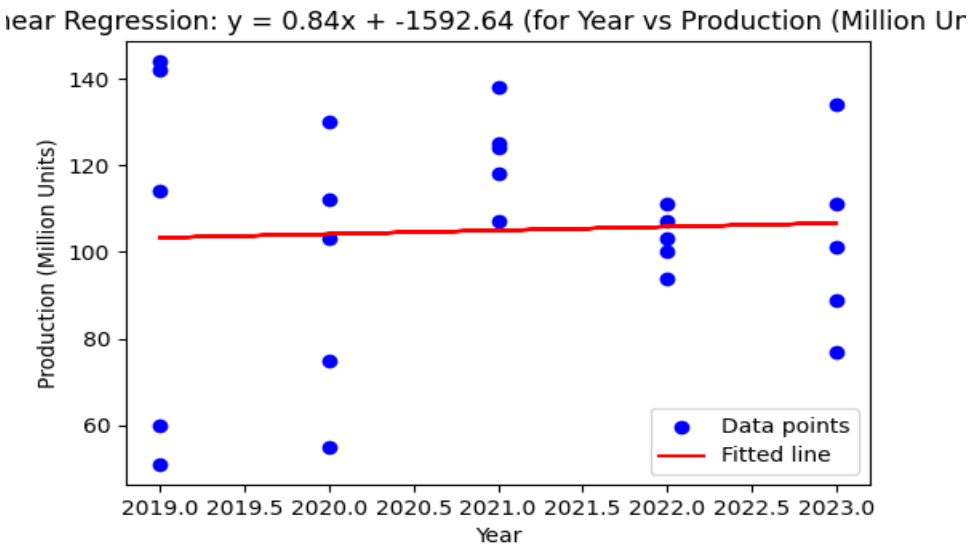
b = y-intercept (computed as 47.75).

Analysis of stapler_pin_market_data.csv

Statistic	Value
Year	count 25.000000 mean 2021.000000 std 1.443376 min 2019.000000 25% 2020.000000 50% 2021.000000 75% 2022.000000 max 2023.000000 Name: Year, dtype: float64
Production (Million Units)	count 25.000000 mean 105.000000 std 25.929391 min 51.000000 25% 94.000000 50% 107.000000 75% 124.000000 max 144.000000 Name: Production (Million Units), dtype: float64
Sales (Million Units)	count 25.000000 mean 107.281123 std 28.765482 min 49.740969 25% 93.451202 50% 107.656880 75% 129.461973 max 155.273202 Name: Sales (Million Units), dtype: float64

Unnamed: 4	count 0.0 mean NaN std NaN min NaN 25% NaN 50% NaN 75% NaN max NaN Name: Unnamed: 4, dtype: float64
Demand	count 25.000000 mean 137.320000 std 91.901088 min 18.000000 25% 56.000000 50% 128.000000 75% 231.000000 max 277.000000 Name: Demand, dtype: float64
Supply	count 25.00000 mean 76.52000 std 37.81988 min 10.00000 25% 39.00000 50% 82.00000 75% 109.00000 max 128.00000 Name: Supply, dtype: float64

Linear Regression Equation: **y = 0.84x + -1592.64 (for Year vs Production (Million Units))**



Inferred Equations:

• **y = 0.84x + -1592.64 (for Year vs Production (Million Units))**

1. Collected data from 'Year' and 'Production (Million Units)'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as 0.84),

b = y-intercept (computed as -1592.64).

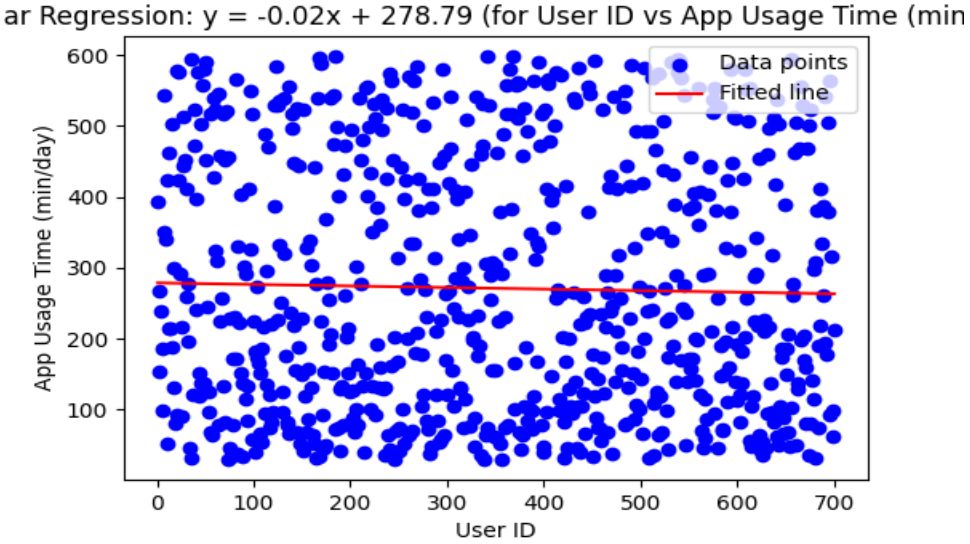
Analysis of user_behavior_dataset.csv

Statistic	Value
User ID	count 700.00000 mean 350.50000 std 202.21688 min 1.00000 25% 175.75000 50% 350.50000 75% 525.25000 max 700.00000 Name: User ID, dtype: float64
App Usage Time (min/day)	count 700.000000 mean 271.128571 std 177.199484 min 30.000000 25% 113.250000 50% 227.500000 75% 434.250000 max 598.000000 Name: App Usage Time (min/day), dtype: float64
Screen On Time (hours/day)	count 700.000000 mean 5.272714 std 3.068584 min 1.000000 25% 2.500000 50% 4.900000 75% 7.400000 max 12.000000 Name: Screen On Time (hours/day), dtype: float64

Battery Drain (mAh/day)	count 700.000000 mean 1525.158571 std 819.136414 min 302.000000 25% 722.250000 50% 1502.500000 75% 2229.500000 max 2993.000000 Name: Battery Drain (mAh/day), dtype: float64
Number of Apps Installed	count 700.000000 mean 50.681429 std 26.943324 min 10.000000 25% 26.000000 50% 49.000000 75% 74.000000 max 99.000000 Name: Number of Apps Installed, dtype: float64
Data Usage (MB/day)	count 700.000000 mean 929.742857 std 640.451729 min 102.000000 25% 373.000000 50% 823.500000 75% 1341.000000 max 2497.000000 Name: Data Usage (MB/day), dtype: float64
Age	count 700.000000 mean 38.482857 std 12.012916 min 18.000000 25% 28.000000 50% 38.000000 75% 49.000000 max 59.000000 Name: Age, dtype: float64

User Behavior Class	count 700.000000 mean 2.990000 std 1.401476 min 1.000000 25% 2.000000 50% 3.000000 75% 4.000000 max 5.000000 Name: User Behavior Class, dtype: float64

Linear Regression Equation: **y = -0.02x + 278.79 (for User ID vs App Usage Time (min/day))**



Inferred Equations:

• **y = -0.02x + 278.79 (for User ID vs App Usage Time (min/day))**

1. Collected data from 'User ID' and 'App Usage Time (min/day)'.
2. Applied linear regression:

y = mx + b, where:

m = slope (computed as -0.02),

b = y-intercept (computed as 278.79).