CRIME AND TOURISM

CMPS004

Data Science in Crime

Haifa Sidani	202400882
Yosr Najjar	202402846

Why Crime and Tourism?

Crime threatens the safety of many people including tourists, and thus it's important to study the relation between crimes and tourism.

Crime can take many faces such as: Homicide

Sexual assault - Assault

Robbery

Burglary

Motor Vehicle Theft

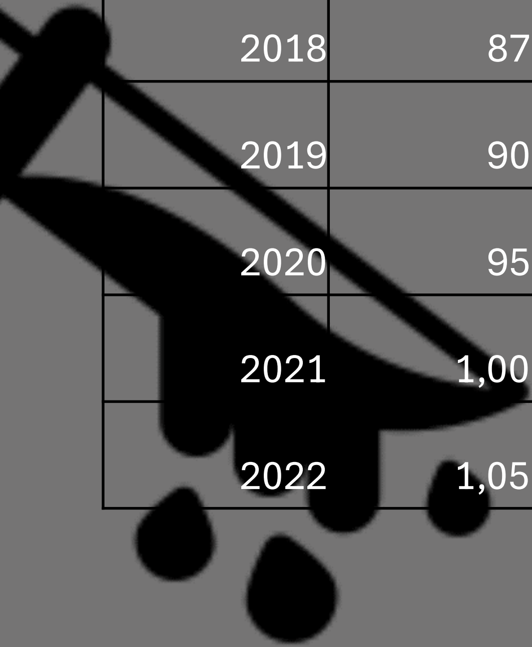
Data Collection

United States

Years	Homicide	Sexual assault	Assault	Robbery	Burglary	Motor Vehicle Theft	Tourism in millions
2015	15,696	1,134,000	1,183,500	366,000	1,579,527	707,758	77.5
2016	17,250	1,235,000	1,244,000	362,000	1,515,096	765,484	75.9
2017	19,547	1,258,000	1,319,000		1,401,840	773,139	76.9
2018	16,425	1,236,000	1,308,000	321,000	1,235,200	748,841	79.7
2019	16,425	1,320,000	1,344,000	298,000	1,117,696	721,885	79.3
2020	21,570	1,330,000	1,380,000	295,000	903,627	810,400	19.4
2021	22,900	1,340,000	1,400,000	290,000	816,355	899,340	Twenty two
2022	23,780	1,350,000	1,420,000	280,000	750,000	932,329	50.9

United Kingdom

Years	Homicide	Sexual assault	Assault	Robbery	Burglary	Motor Vehicle Theft	Tourism in millions
2015	750	1,457,000	2,378,000	236,000	701,000	81,000	36.1
2016	800	1,730,000	2,741,000	284,000	686,000	92,000	37.6
2017	850	1,896,000	3,032,000	302,000	682,000	103,000	39.2
2018	870	2,000,000	3,200,000	310,000	631	106,000	36.3
2019	900	2,100,000	3,300,000	320,000	402,000	107,198	40.9
2020	950	2,150,000	3,400,000	330,000	356,017	107,575	11.1
2021	1,000	2,200,000	3,500,000	340,000	267,931	108,542	6.2
2022	1,050	2,250,000	3,600,000	350,000	288,250	110,000	25.3



Data Preprocessing

United States

Years	Homicide	Sexual assault	Assault	Robbery	Burglary	Motor Vehicle Theft	Tourism in millions
2015	15,696	1,134,000	1,183,500	366,000	1,579,527	707,758	77.5
2016	17,250	1,235,000	1,244,000	362,000	1,515,096	765,484	75.9
2017	19,547	1,258,000	1,319,000		1,401,840	773,139	76.9
2018	16,425	1,236,000	1,308,000	321,000	1,235,200	748,841	79.7
2019	16,425	1,320,000	1,344,000	298,000	1,117,696	721,885	79.3
2020	21,570	1,330,000	1,380,000	295,000	903,627	810,400	19.4
2021	22,900	1,340,000	1,400,000	290,000	816,355	899,340	**Twenty two**
2022	23,780	1,350,000	1,420,000	280,000	750,000	932,329	50.9

United Kingdom

Years	Homicide	Sexual assault	Assault	Robbery	Burglary	Motor Vehicle Theft	Tourism in millions
2015	750	1,457,000	2,378,000	236,000	701,000	81,000	36.1
2016	800	1,730,000	2,741,000	284,000	686,000	92,000	37.6
2017	850	1,896,000	3,032,000	302,000	682,000	103,000	39.2
2018	870	2,000,000	3,200,000	310,000	631	106,000	36.3
2019	900	2,100,000	3,300,000	320,000	402,000	107,198	40.9
2020	950	2,150,000	3,400,000	330,000	356,017	107,575	11.1
2021	1,000	2,200,000	3,500,000	340,000	267,931	108,542	6.2
2022	1,050	2,250,000	3,600,000	350,000	288,250	110,000	25.3

Data Cleansing

United States

Years	Homicide	Sexual assault	Assault	Robbery	Burglary	Motor Vehicle Theft	Tourism in millions
2015	15,696	1,134,000	1,183,500	366,000	1,579,527	707,758	77.5
2016	17,250	1,235,000	1,244,000	362,000	1,515,096	765,484	75.9
2017	19,547	1,258,000	1,319,000	316,000	1,401,840	773,139	76.9
2018	16,425	1,236,000	1,308,000	321,000	1,235,200	748,841	79.7
2019	16,425	1,320,000	1,344,000	298,000	1,117,696	721,885	79.3
2020	21,570	1,330,000	1,380,000	295,000	903,627	810,400	19.4
2021	22,900	1,340,000	1,400,000	290,000	816,355	899,340	**22**
2022	23,780	1,350,000	1,420,000	280,000	750,000	932,329	50.9

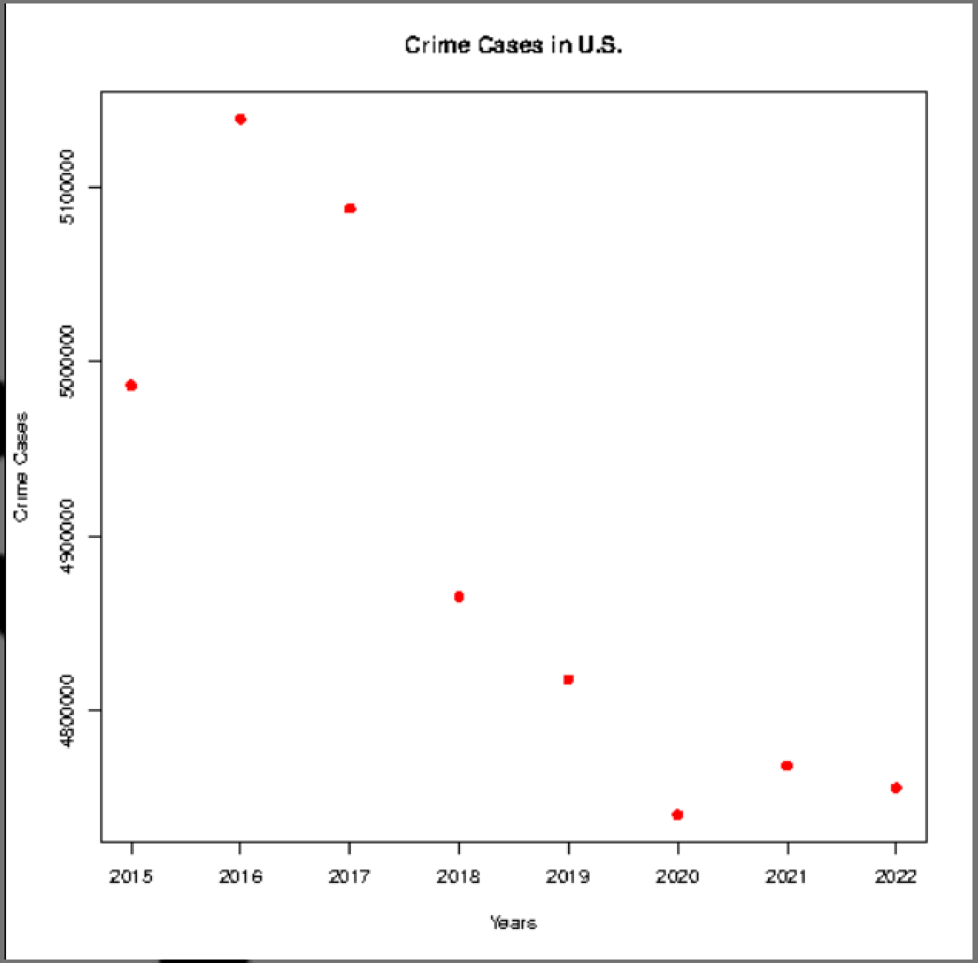
United Kingdom

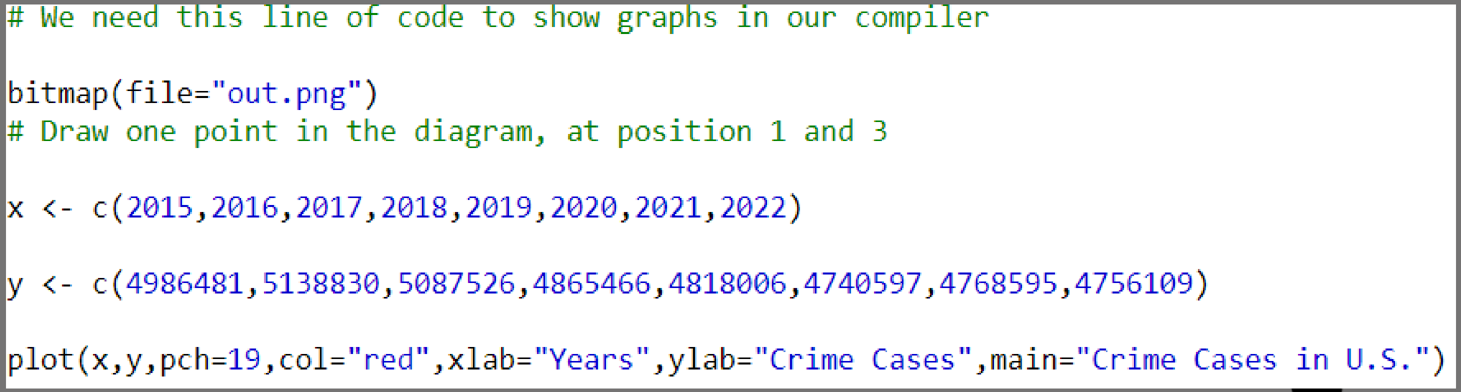
Years	Homicide	Sexual assault	Assault	Robbery	Burglary	Motor Vehicle Theft	Tourism in millions
2015	750	1,457,000	2,378,000	236,000	701,000	81,000	36.1
2016	800	1,730,000	2,741,000	284,000	686,000	92,000	37.6
2017	850	1,896,000	3,032,000	302,000	682,000	103,000	39.2
2018	870	2,000,000	3,200,000	310,000	631,000	106,000	36.3
2019	900	2,100,000	3,300,000	320,000	402,000	107,198	40.9
2020	950	2,150,000	3,400,000	330,000	356,017	107,575	11.1
2021	1,000	2,200,000	3,500,000	340,000	267,931	108,542	6.2
2022	1,050	2,250,000	3,600,000	350,000	288,250	110,000	25.3

Exploratory Data Analysis (EDA)

U.S. (Using R)

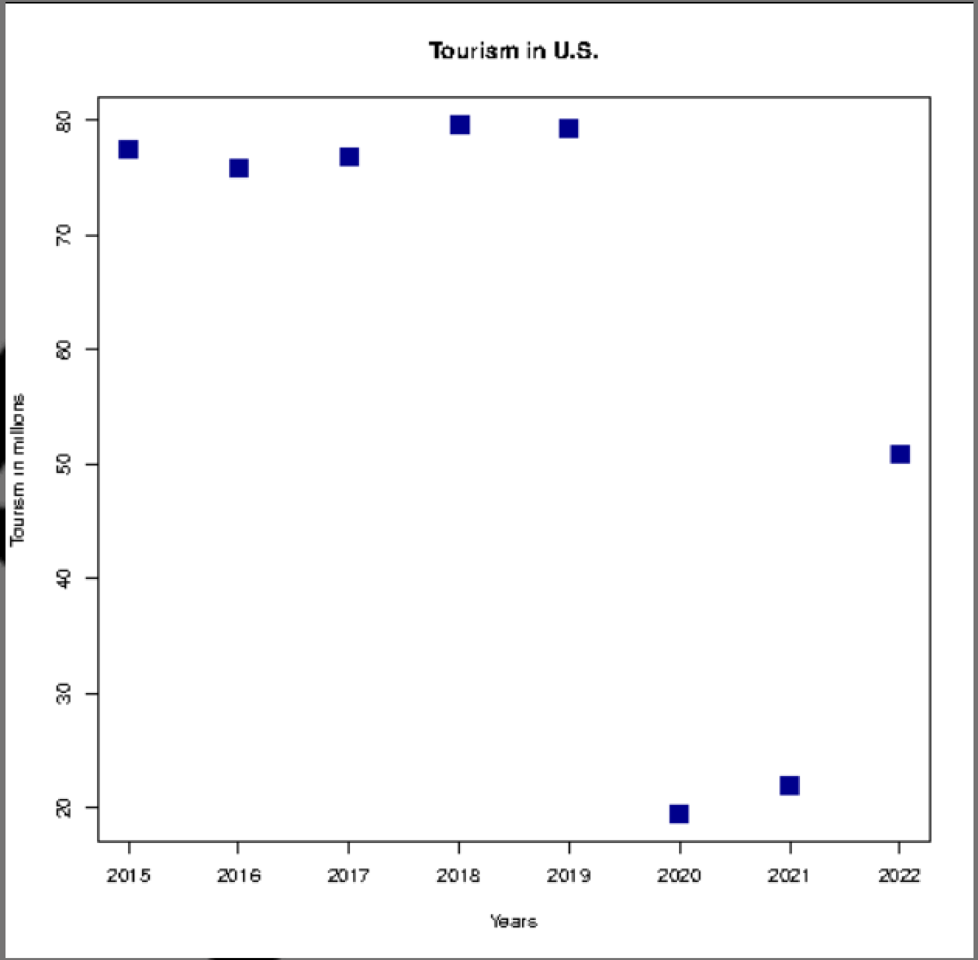
For Crime Cases:

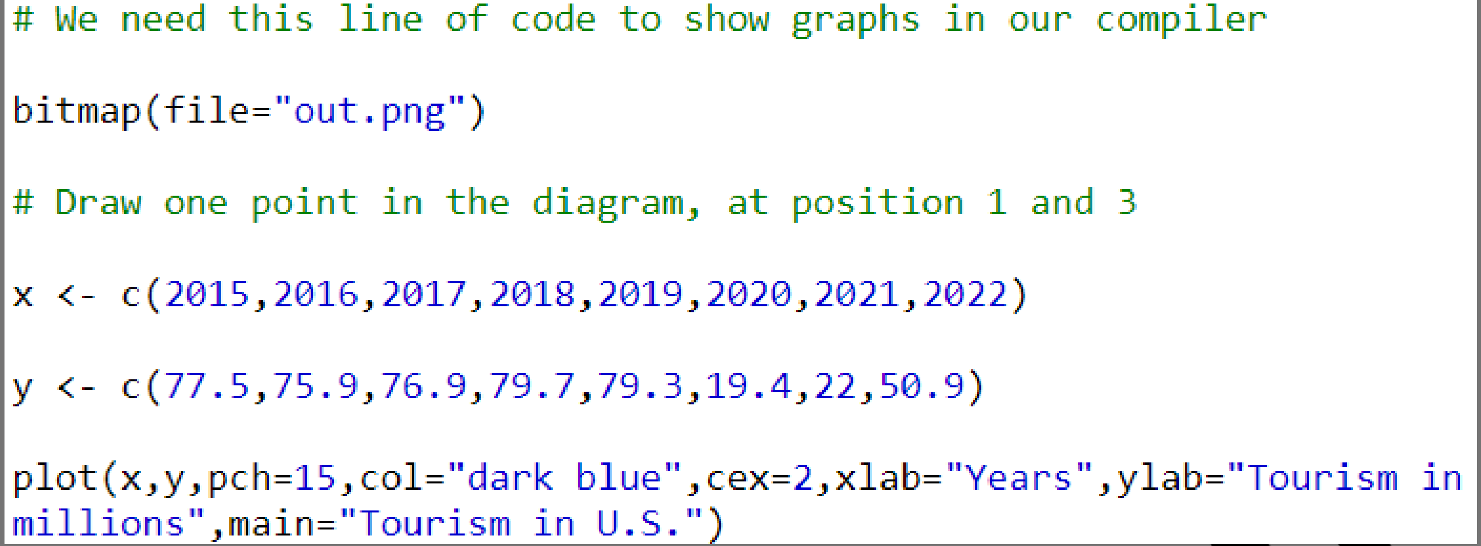




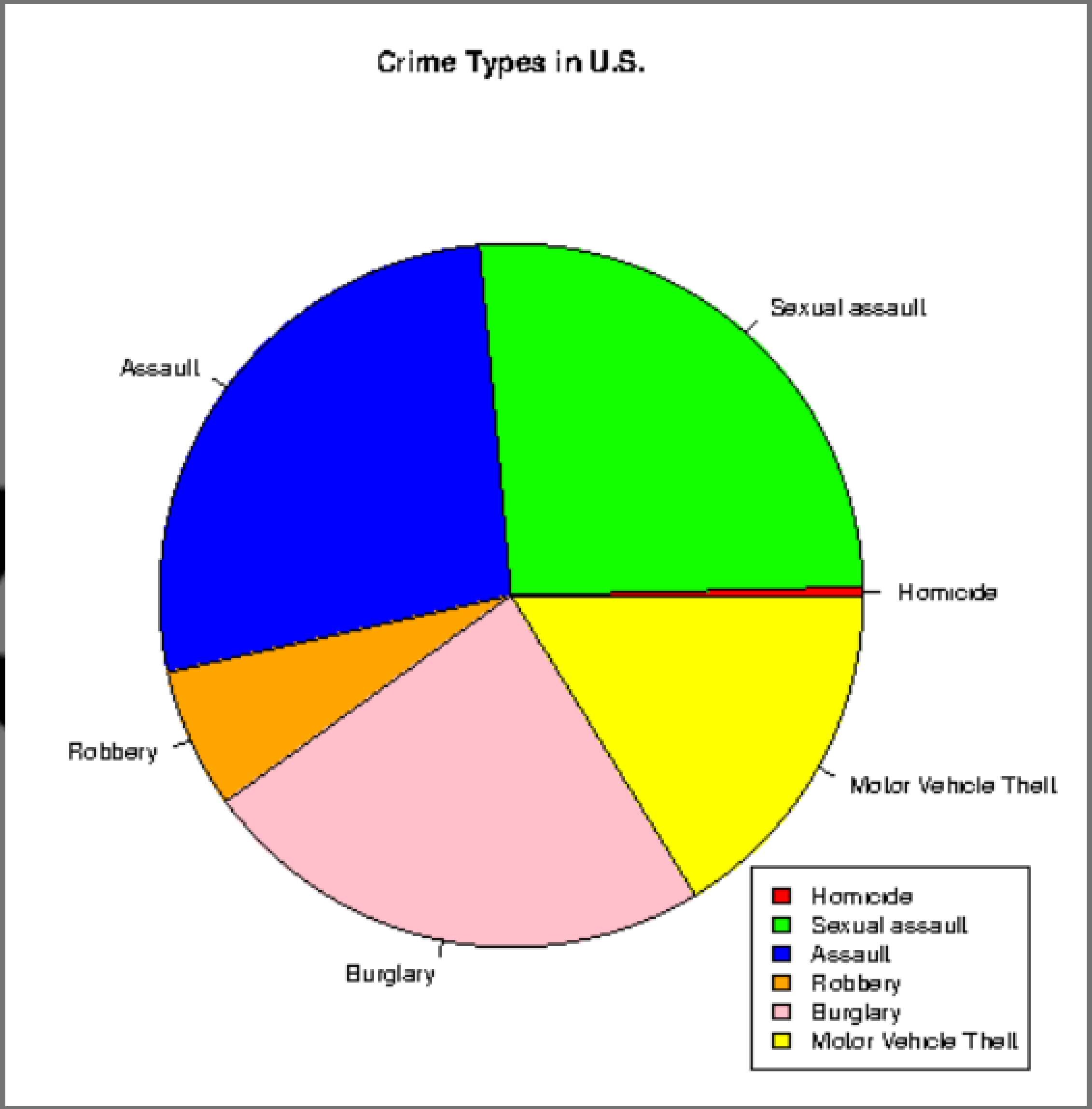
```
# We need this line of code to show graphs in our compiler

bitmap(file="out.png")
# Draw one point in the diagram, at position 1 and 3

x <- c(2015,2016,2017,2018,2019,2020,2021,2022)

y <- c(4986481,5138830,5087526,4865466,4818006,4740597,4768595,4756109)

plot(x,y,pch=19,col="red",xlab="Years",ylab="Crime Cases",main="Crime Cases in U.S.")
```

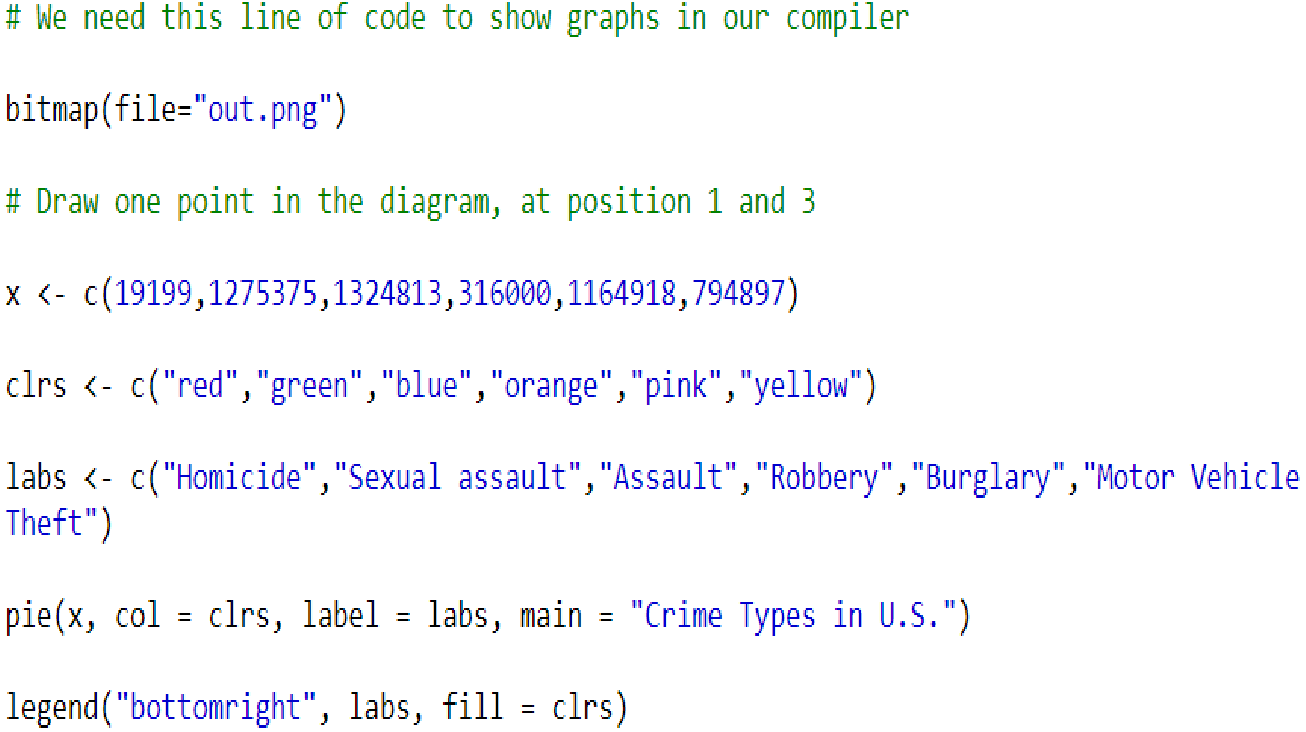
For Tourism:





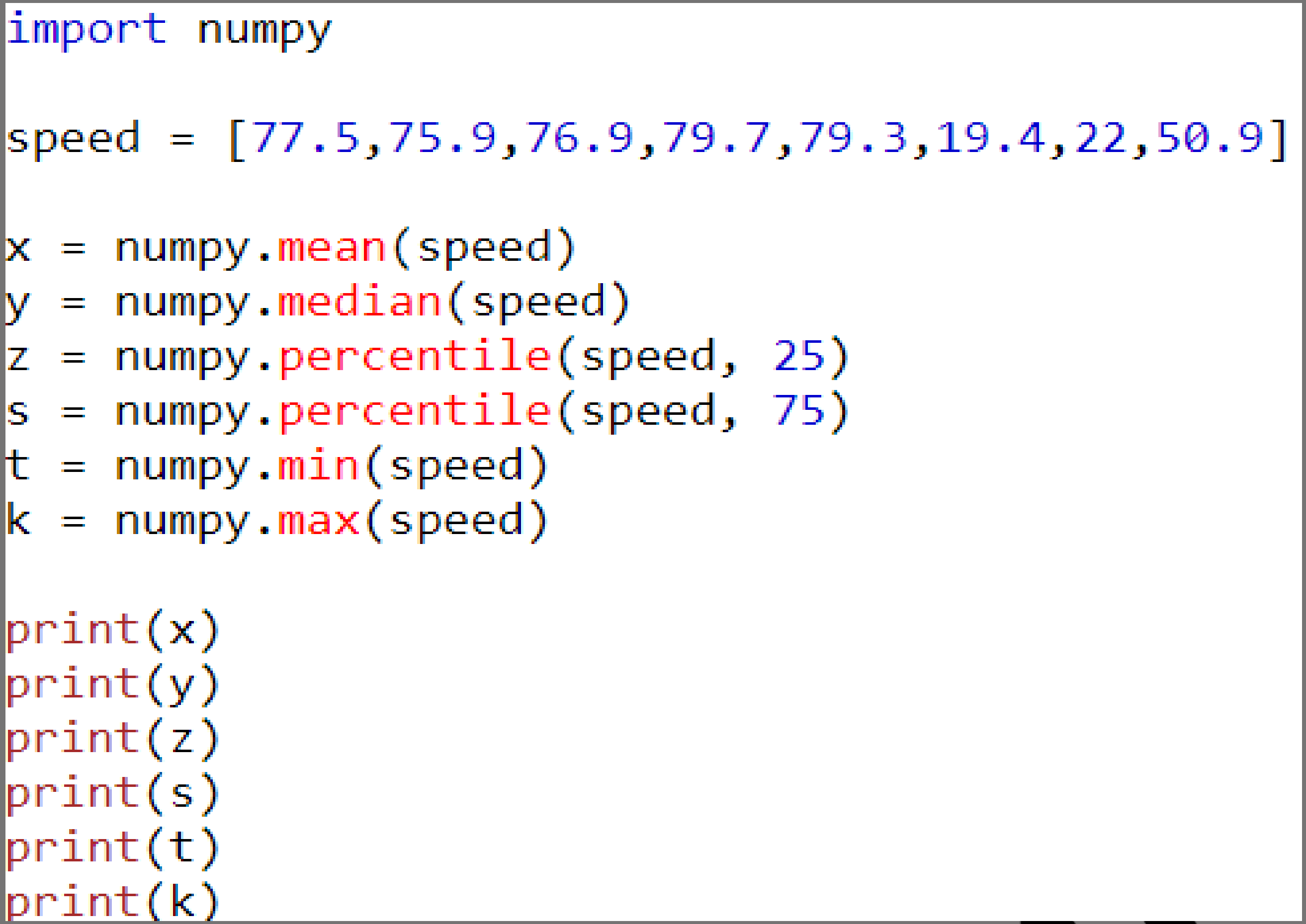
```
# We need this line of code to show graphs in our compiler

bitmap(file="out.png")

# Draw one point in the diagram, at position 1 and 3

x <- c(2015,2016,2017,2018,2019,2020,2021,2022)

y <- c(77.5,75.9,76.9,79.7,79.3,19.4,22,50.9)

plot(x,y,pch=15,col="dark blue",cex=2,xlab="Years",ylab="Tourism in
millions",main="Tourism in U.S.")
```





```
# We need this line of code to show graphs in our compiler

bitmap(file="out.png")

# Draw one point in the diagram, at position 1 and 3

x <- c(19199,1275375,1324813,316000,1164918,794897)

clrs <- c("red","green","blue","orange","pink","yellow")

labs <- c("Homicide","Sexual assault","Assault","Robbery","Burglary","Motor Vehicle Theft")

pie(x, col = clrs, label = labs, main = "Crime Types in U.S.")

legend("bottomright", labs, fill = clrs)
```

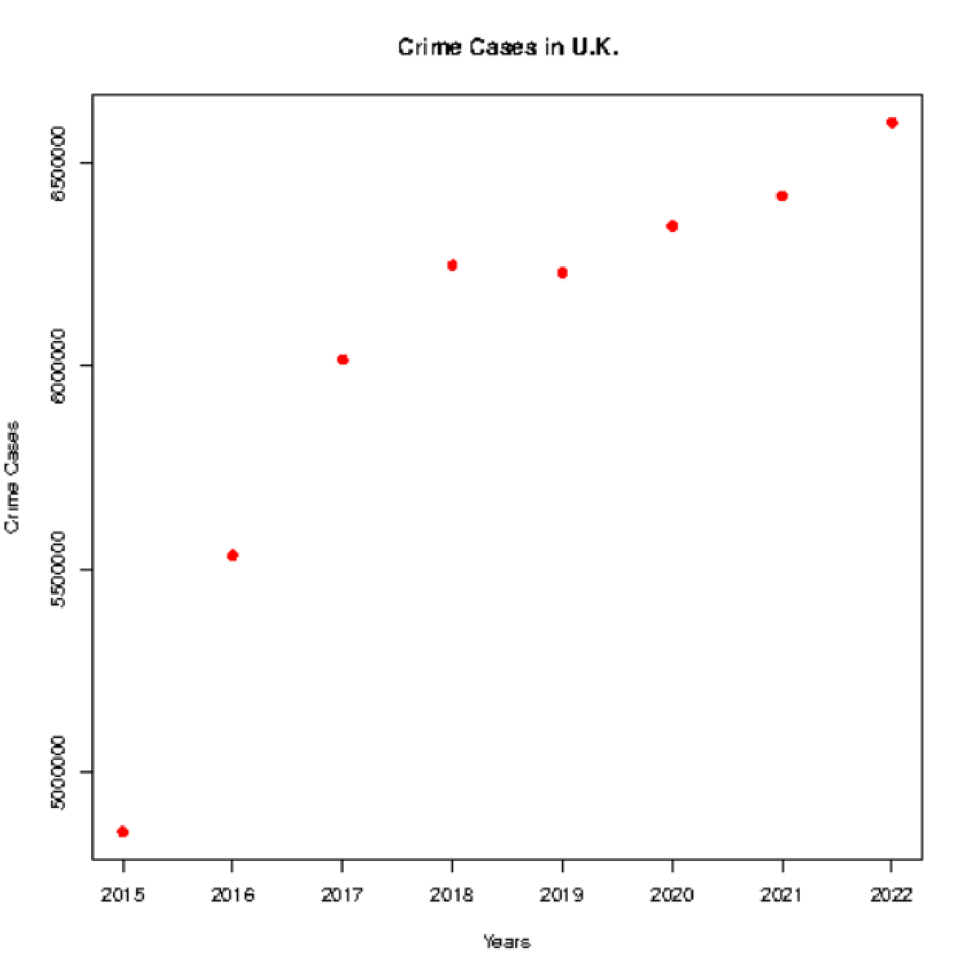
Tourism in U.S. (using python)	
Mean	60.2
Median	76.4
First Quartile	43.675
Third Quartile	77.95
Minimum	19.4
Maximum	79.7

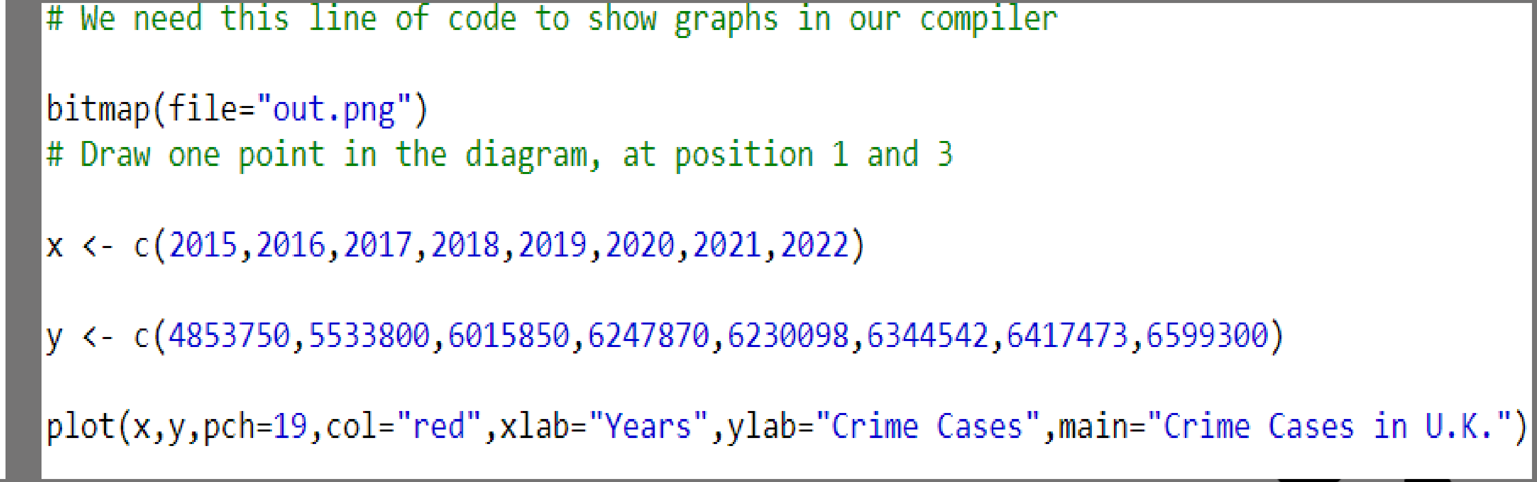


```
import numpy

speed = [77.5,75.9,76.9,79.7,79.3,19.4,22,50.9]

x = numpy.mean(speed)
y = numpy.median(speed)
z = numpy.percentile(speed, 25)
s = numpy.percentile(speed, 75)
t = numpy.min(speed)
k = numpy.max(speed)

print(x)
print(y)
print(z)
print(s)
print(t)
print(k)
```

U.K (Using R)

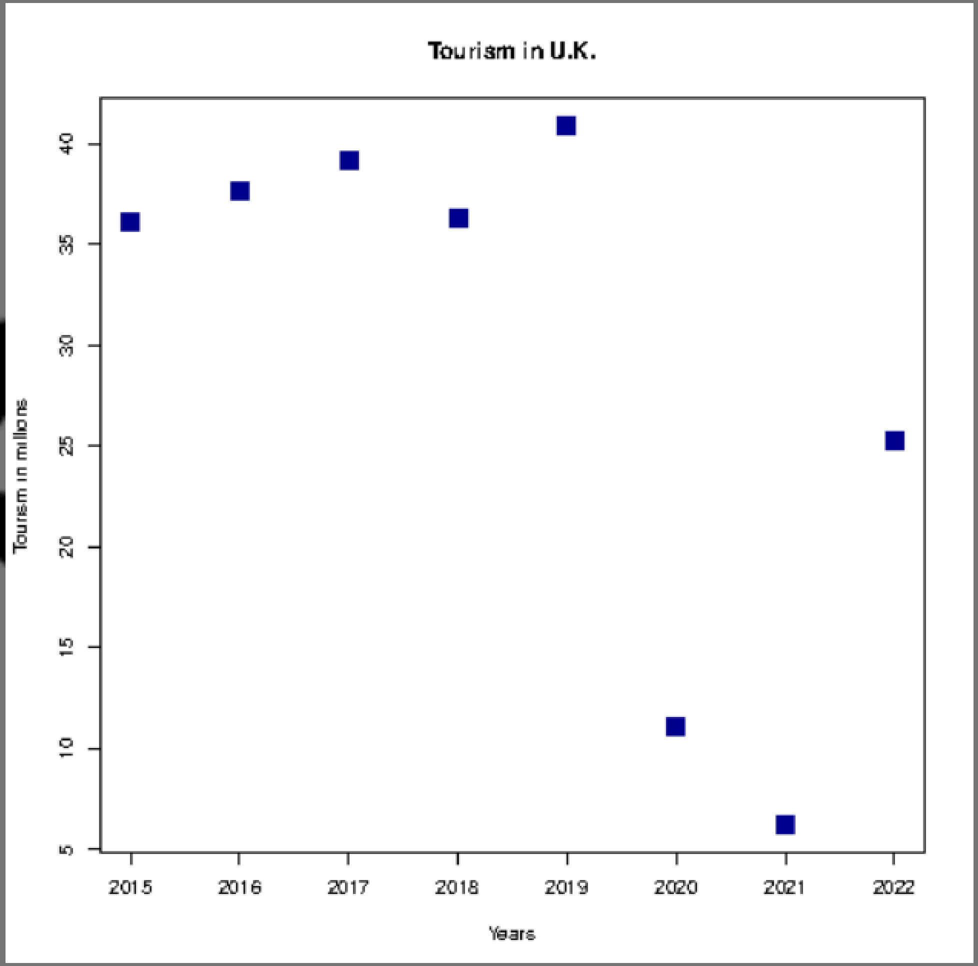
For Crime Cases:

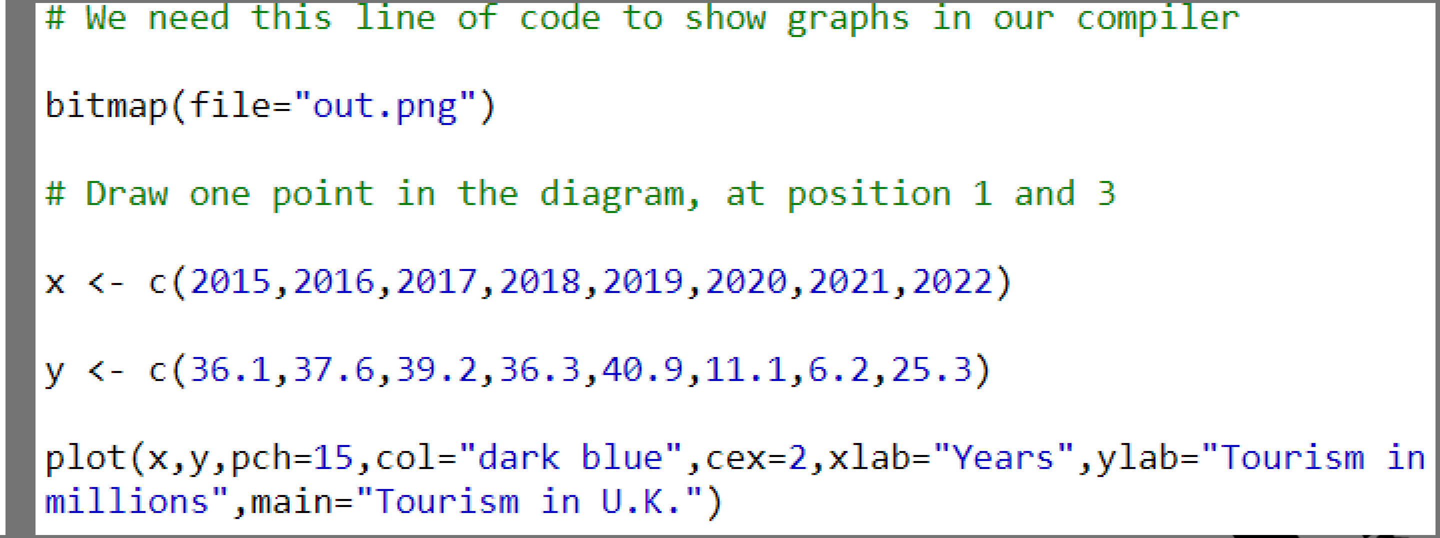




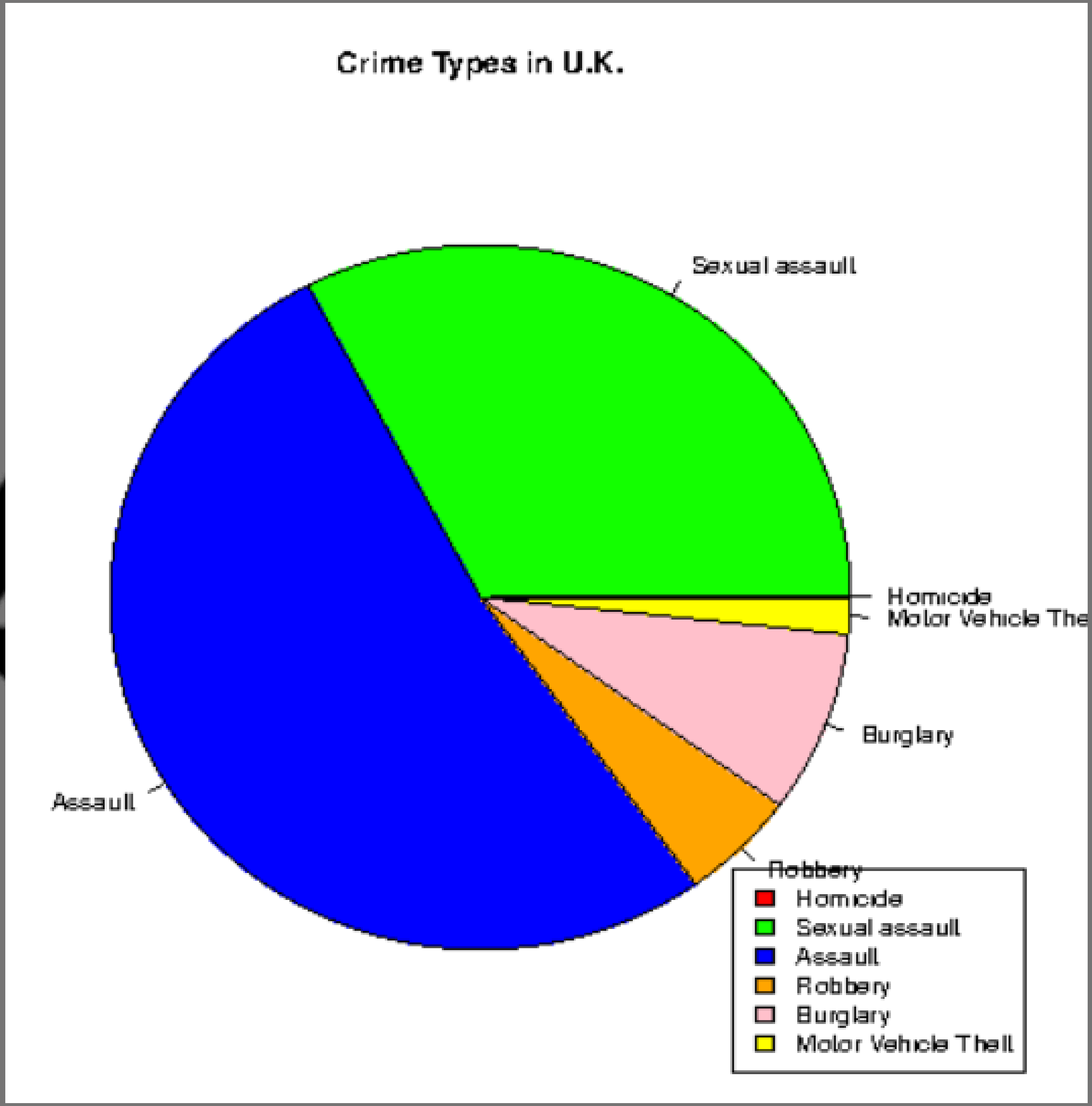
```
# We need this line of code to show graphs in our compiler

bitmap(file="out.png")
# Draw one point in the diagram, at position 1 and 3

x <- c(2015,2016,2017,2018,2019,2020,2021,2022)

y <- c(4853750,5533800,6015850,6247870,6230098,6344542,6417473,6599300)

plot(x,y,pch=19,col="red",xlab="Years",ylab="Crime Cases",main="Crime Cases in U.K.")
```

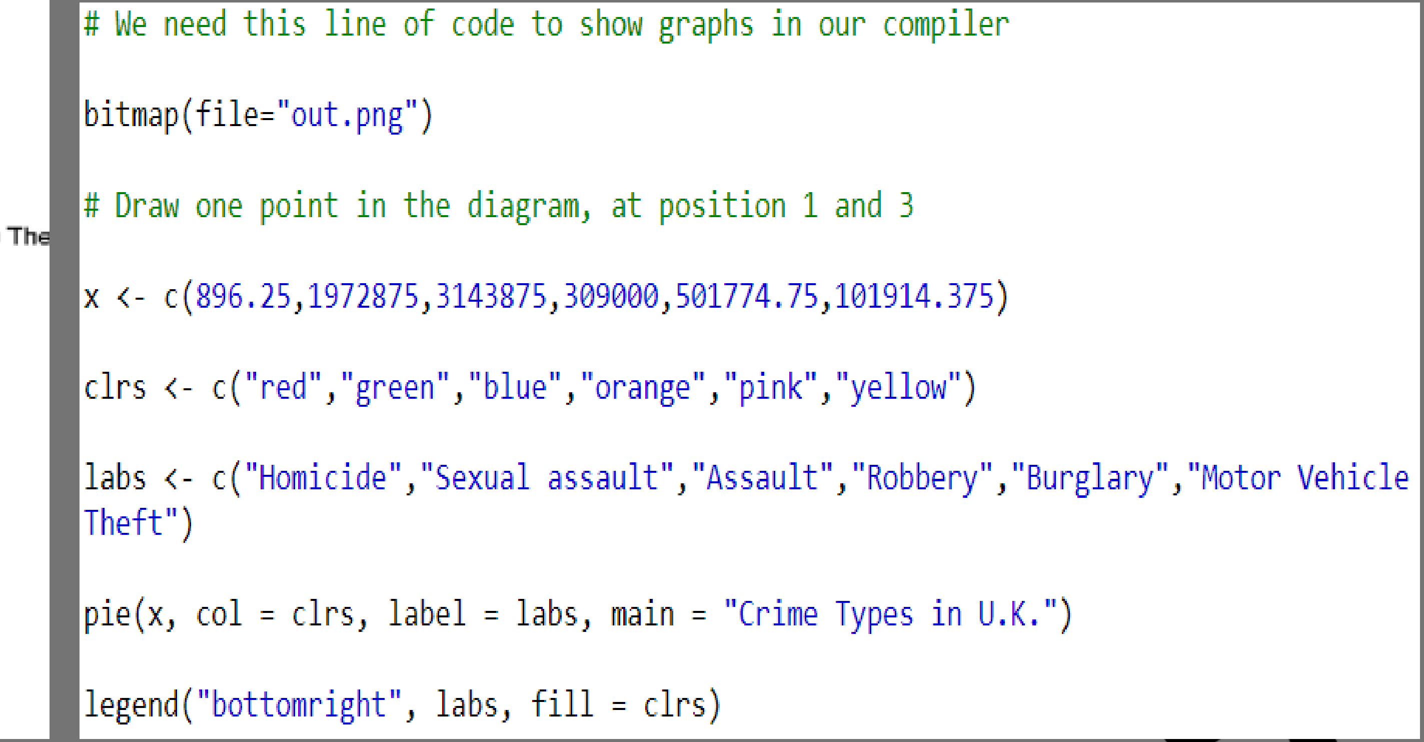
For Tourism:





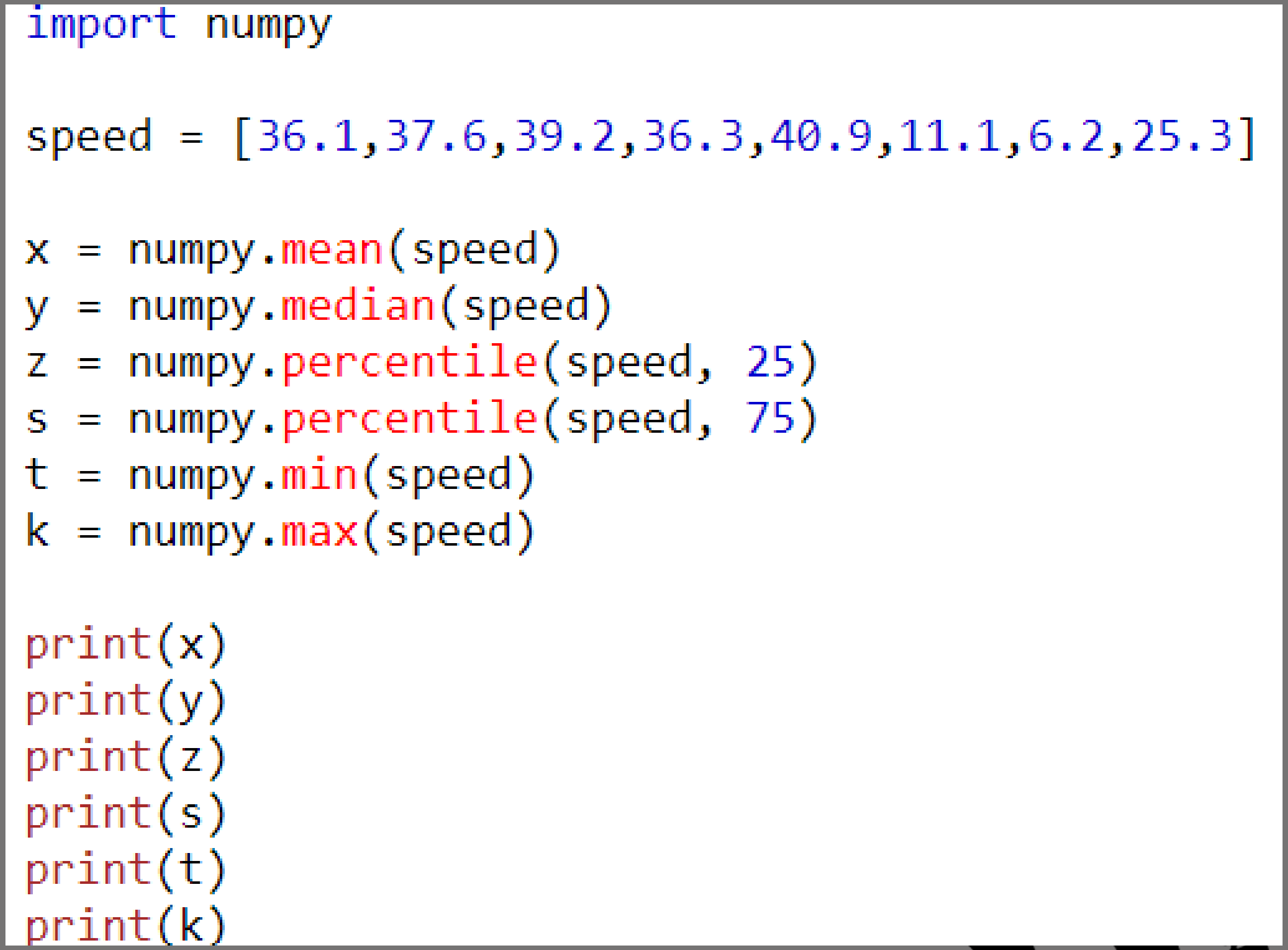
```
# We need this line of code to show graphs in our compiler

bitmap(file="out.png")

# Draw one point in the diagram, at position 1 and 3

x <- c(2015,2016,2017,2018,2019,2020,2021,2022)

y <- c(36.1,37.6,39.2,36.3,40.9,11.1,6.2,25.3)

plot(x,y,pch=15,col="dark blue",cex=2,xlab="Years",ylab="Tourism in millions",main="Tourism in U.K.")
```





```
# We need this line of code to show graphs in our compiler

bitmap(file="out.png")

# Draw one point in the diagram, at position 1 and 3

x <- c(896.25,1972875,3143875,309000,501774.75,101914.375)

clrs <- c("red","green","blue","orange","pink","yellow")

labs <- c("Homicide","Sexual assault","Assault","Robbery","Burglary","Motor Vehicle Theft")

pie(x, col = clrs, label = labs, main = "Crime Types in U.K.")

legend("bottomright", labs, fill = clrs)
```

Tourism in U.K. (using python)	
Mean	29.0875
Median	36.2
First Quartile	21.75
Third Quartile	38.0
Minimum	6.2
Maximum	40.9



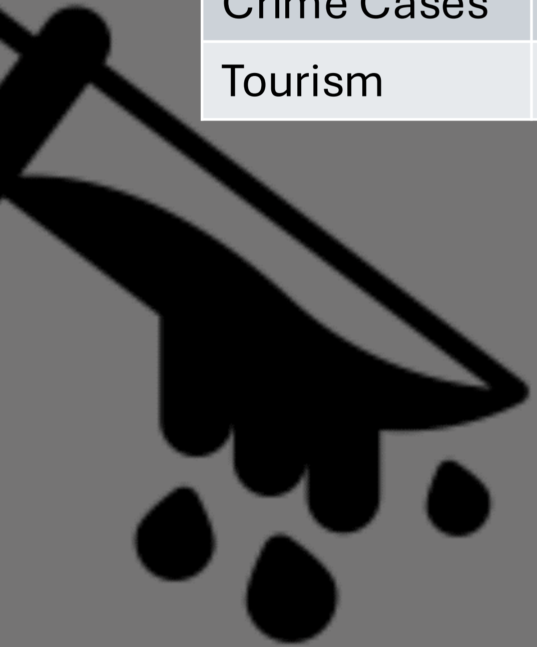
```
import numpy

speed = [36.1,37.6,39.2,36.3,40.9,11.1,6.2,25.3]

x = numpy.mean(speed)
y = numpy.median(speed)
z = numpy.percentile(speed, 25)
s = numpy.percentile(speed, 75)
t = numpy.min(speed)
k = numpy.max(speed)

print(x)
print(y)
print(z)
print(s)
print(t)
print(k)
```

Regression Analysis

Crime/Tourism Model – U.S.

- Consider the following data:

Years	2015	2016	2017	2018	2019	2022
Crime Cases	4,986,481	5,138,830	5,087,526	4,865,466	4,818,006	4,756,109
Tourism	77.5	75.9	76.9	79.7	79.3	50.9



C → Crime

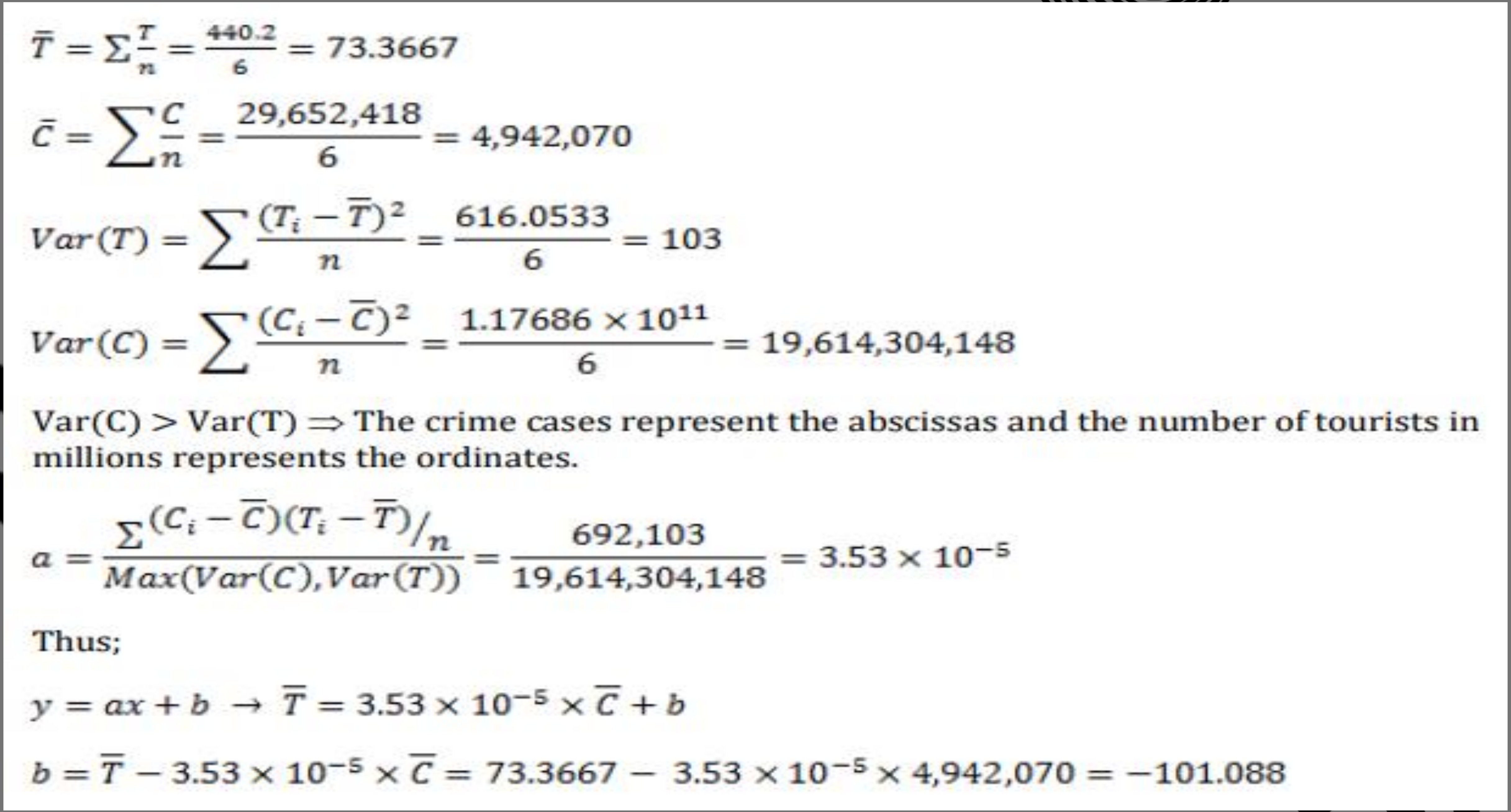
T→ Tourism

n → number of years

Var → variance

Statistical Study and Regression Line

	n	C_i	T_i	$C_i – C'_i$	$T_i – T'_i$	$(C_i – C'_i)(T_i – T'_i)$	$(C_i – C'_i)^2$	$(T_i – T'_i)^2$
	2015	4,986,481	77.5	44411.3	4	183566.8444	1972336921	17.0844444
	2016	5,138,830	75.9	196760.3	3	498459.5111	38714497600	6.41777778
	2017	5,087,526	76.9	145456.3	4	513945.7111	21157447936	12.4844444
	2018	4,865,466	79.7	-76603.7	6	-485156.5556	5868019609	40.1111111
	2019	4,818,006	79.3	-124063.7	6	-736111.0889	15391627969	35.2044444
	2022	4,756,109	50.9	-185960.7	-22	4177916.311	34581121600	504.751111
Sum		29652418	440.2	0	0	4152620.733	1.17685E+11	616.053333
Average		4942069.7	73.3667	0	0	692103	19614175273	103



$$\bar{T} = \sum \frac{T}{n} = \frac{440.2}{6} = 73.3667$$

$$\bar{C} = \sum \frac{C}{n} = \frac{29,652,418}{6} = 4,942,070$$

$$Var(T) = \sum \frac{(T_i - \overline{T})^2}{n} = \frac{616.0533}{6} = 103$$

$$Var(C) = \sum \frac{(C_i - \overline{C})^2}{n} = \frac{1.17686 \times 10^{11}}{6} = 19,614,304,148$$

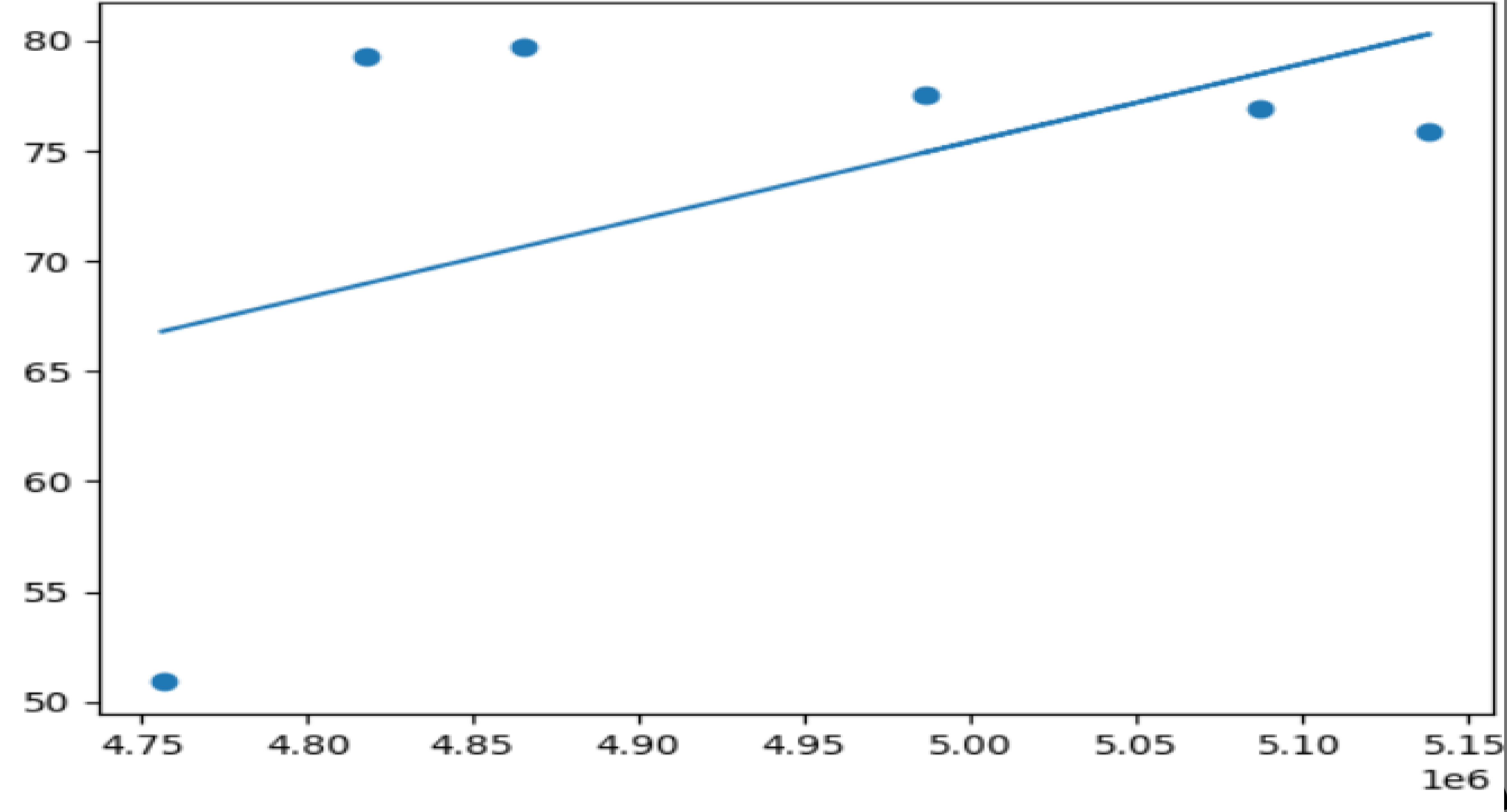
Var(C) > Var(T) ⇒ The crime cases represent the abscissas and the number of tourists in millions represents the ordinates.

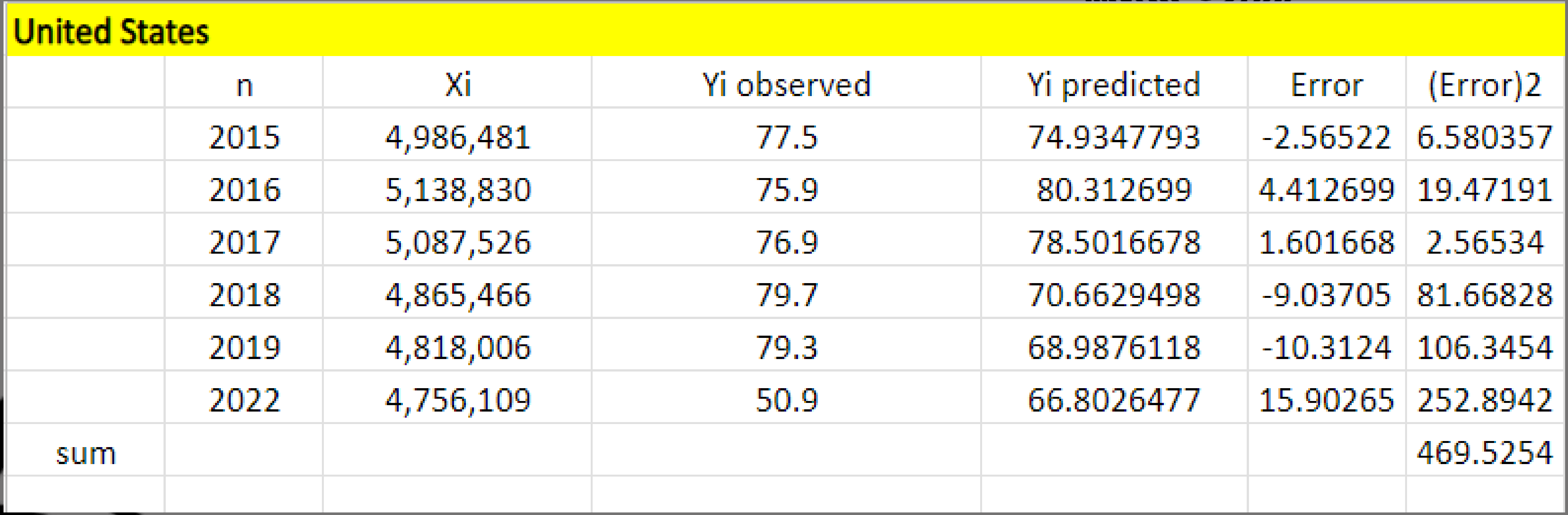
$$a = \frac{\sum^{(C_i - \overline{C})(T_i - \overline{T})}/_{n}}{Max(Var(C), Var(T))} = \frac{692,103}{19,614,304,148} = 3.53 \times 10^{-5}$$

Thus;

$$y = ax + b \ \rightarrow \ \overline{T} = 3.53 \times 10^{-5} \times \overline{C} + b$$

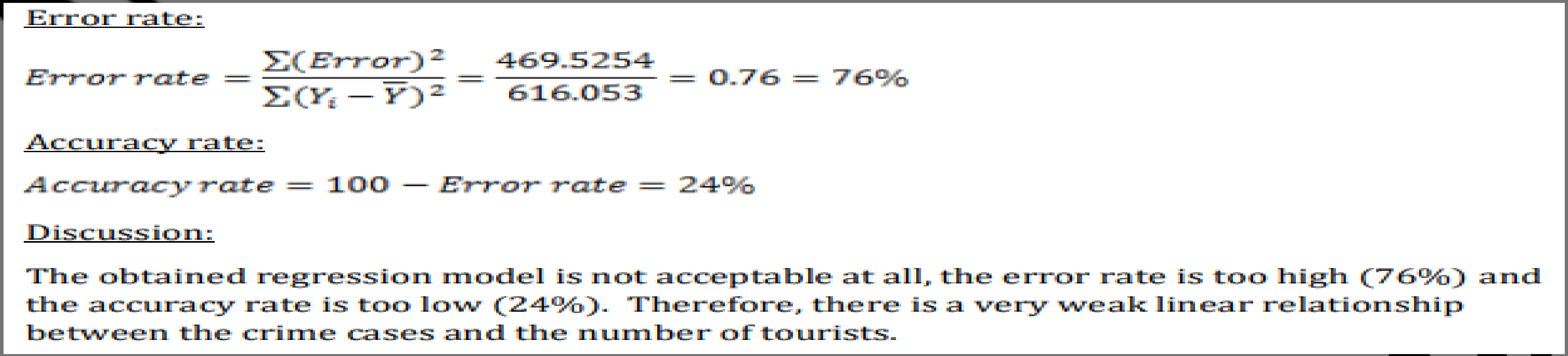
$$b = \overline{T} - 3.53 \times 10^{-5} \times \overline{C} = 73.3667 - \ 3.53 \times 10^{-5} \times 4,942,070 = -101.088$$


80
75
70
65
60
55
50
4.75
4.80
4.85
4.90
4.95
5.00
5.05
5.10
5.15
1e6



United States

	n	Xi	Yi observed	Yi predicted	Error	(Error)2
	2015	4,986,481	77.5	74.9347793	-2.56522	6.580357
	2016	5,138,830	75.9	80.312699	4.412699	19.47191
	2017	5,087,526	76.9	78.5016678	1.601668	2.56534
	2018	4,865,466	79.7	70.6629498	-9.03705	81.66828
	2019	4,818,006	79.3	68.9876118	-10.3124	106.3454
	2022	4,756,109	50.9	66.8026477	15.90265	252.8942
sum						469.5254



Error rate:

$$Error\,rate = \frac{\sum(Error)^2}{\sum(Y_i - \overline{Y})^2} = \frac{469.5254}{616.053} = 0.76 = 76\%$$

Accuracy rate:

$$Accuracy\,rate = 100 - Error\,rate = 24\%$$

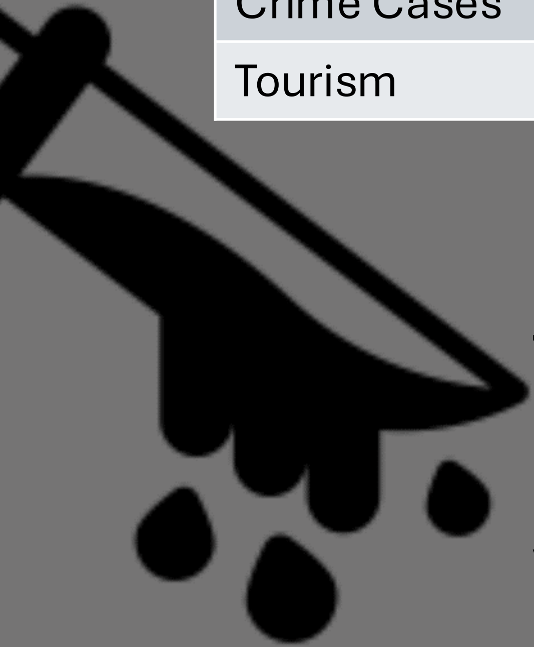
Discussion:

The obtained regression model is not acceptable at all, the error rate is too high (76%) and the accuracy rate is too low (24%). Therefore, there is a very weak linear relationship between the crime cases and the number of tourists.

Crime/Tourism Model – U.K.

- Consider the following data:

Years	2015	2016	2017	2018	2019	2022
Crime Cases	4,853,750	5,533,800	6,015,850	6,247,870	6,230,098	6,599,300
Tourism	36.1	37.6	39.2	36.3	40.9	25.3



C → Crime

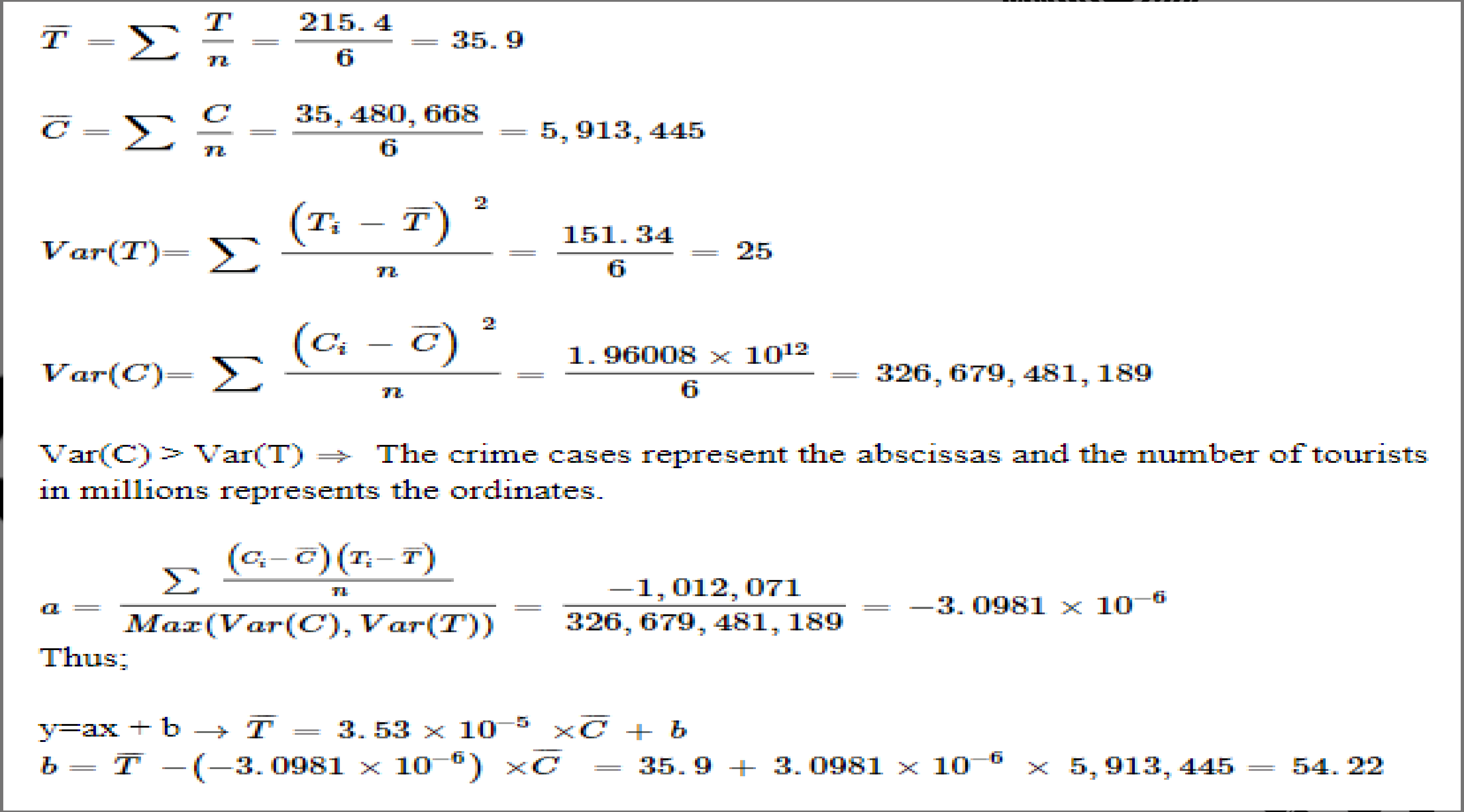
T→ Tourism

n → number of years

Var → variance

Statistical Study and Regression Line

	n	C_i	T_i	$C_i – C'_i$	$T_i – T'_i$	$(C_i – C'_i)(T_i – T'_i)$	$(C_i – C'_i)^2$	$(T_i – T'_i)^2$
	2015	4,853,750	36.1	-1,059,695	0	-211938.933333336	1122952786561.78	0.04
	2016	5,533,800	37.6	-379,645	2	-645395.933333335	144130072928.445	2.89
	2017	6,015,850	39.2	102,405	3	337937.599999999	10486852295.111	10.89
	2018	6,247,870	36.3	334,425	0	133770.133333333	111840303575.111	0.16
	2019	6,230,098	40.9	316,653	5	1583266.66666667	100269333511.111	25
	2022	6,599,300	25.3	685,855	-11	-7270066.53333333	470397538261.777	112.36
Sum		35,480,668	215.4	0	0	-6072427	1960076887133.33	151.34
Average		5,913,445	35.9	0	0	-1,012,071	326,679,481,189	25



$$\overline{T} = \sum \frac{T}{n} = \frac{215.4}{6} = 35.9$$

$$\overline{C} = \sum \frac{C}{n} = \frac{35,480,668}{6} = 5,913,445$$

$$Var(T) = \sum \frac{\left(T_i - \overline{T}\right)^2}{n} = \frac{151.34}{6} = 25$$

$$Var(C) = \sum \frac{\left(C_i - \overline{C}\right)^2}{n} = \frac{1.96008 \times 10^{12}}{6} = 326,679,481,189$$

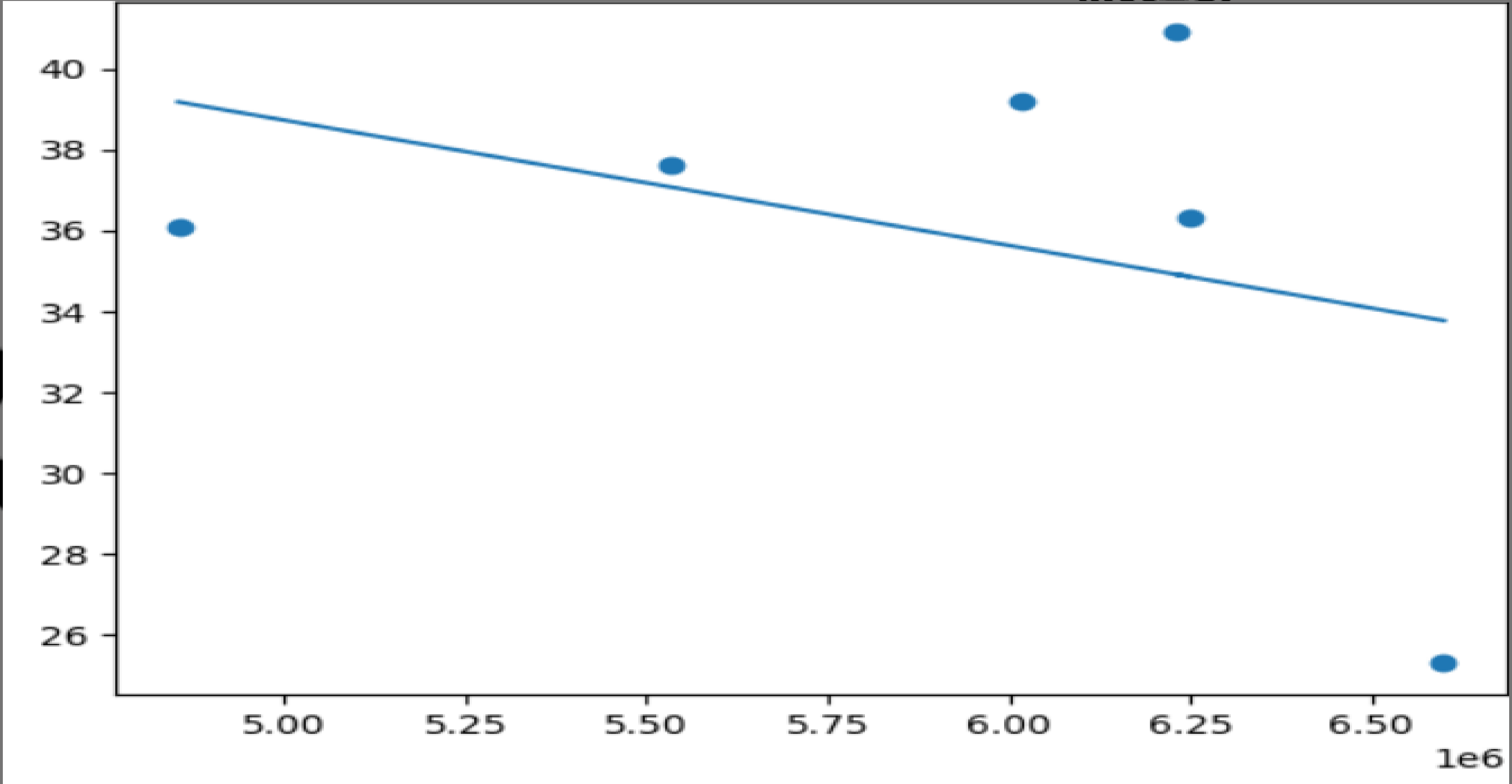
Var(C) > Var(T) ⇒ The crime cases represent the abscissas and the number of tourists in millions represents the ordinates.

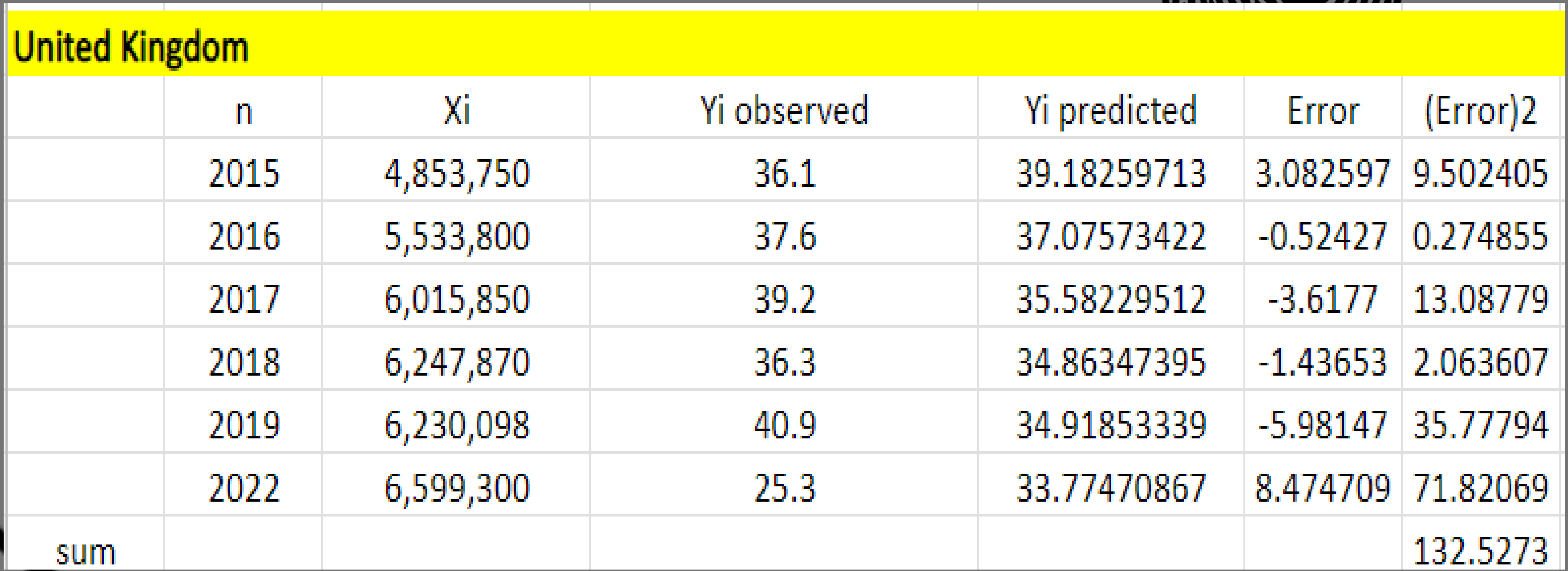
$$a = \frac{\sum \frac{\left(C_i - \overline{C}\right)\left(T_i - \overline{T}\right)}{n}}{Max(Var(C), Var(T))} = \frac{-1,012,071}{326,679,481,189} = -3.0981 \times 10^{-6}$$

Thus;

y=ax + b $\rightarrow \overline{T} = 3.53 \times 10^{-5} \times \overline{C} + b$

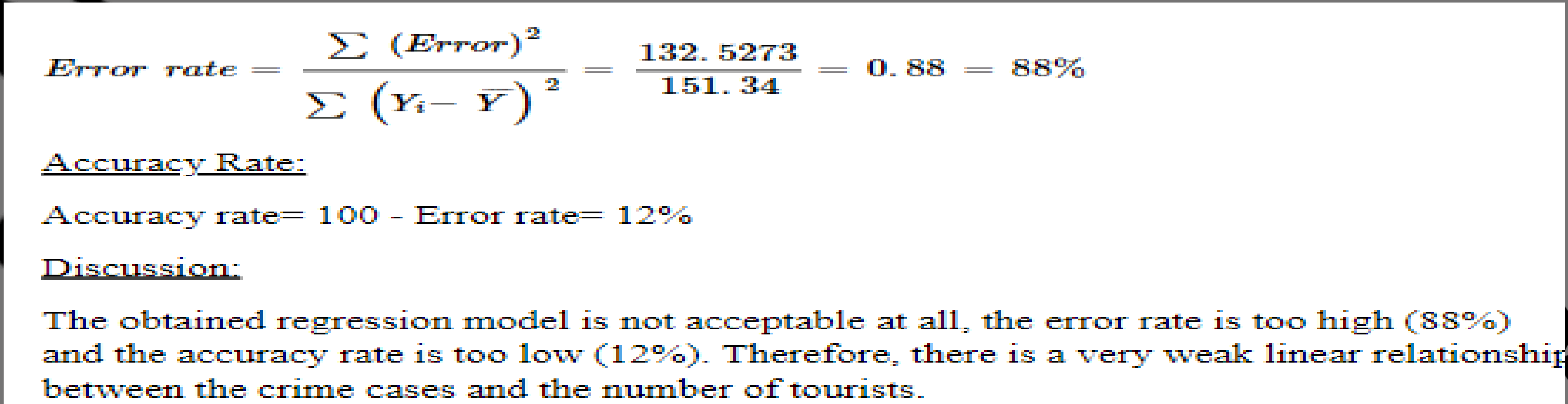
$b = \overline{T} - \left(-3.0981 \times 10^{-6}\right) \times \overline{C} = 35.9 + 3.0981 \times 10^{-6} \times 5,913,445 = 54.22$


40
38
36
34
32
30
28
26
5.00
5.25
5.50
5.75
6.00
6.25
6.50
1e6



United Kingdom

	n	Xi	Yi observed	Yi predicted	Error	(Error)2
	2015	4,853,750	36.1	39.18259713	3.082597	9.502405
	2016	5,533,800	37.6	37.07573422	-0.52427	0.274855
	2017	6,015,850	39.2	35.58229512	-3.6177	13.08779
	2018	6,247,870	36.3	34.86347395	-1.43653	2.063607
	2019	6,230,098	40.9	34.91853339	-5.98147	35.77794
	2022	6,599,300	25.3	33.77470867	8.474709	71.82069
sum						132.5273



$$Error\ rate = \frac{\sum\ (Error)^2}{\sum\ \left(Y_i - \bar{Y}\right)^2} = \frac{132.5273}{151.34} = 0.88 = 88\%$$

Accuracy Rate:

Accuracy rate= 100 - Error rate= 12%

Discussion:

The obtained regression model is not acceptable at all, the error rate is too high (88%) and the accuracy rate is too low (12%). Therefore, there is a very weak linear relationshi between the crime cases and the number of tourists.

Conclusion

This is an example of how data science integrates with business and help to identify the risk factors that might affect business and the economic level in order to improve and avoid these risks.

The studied models above prove that crime is not a risk of tourism.

Thank You!