Data Science - Data cleaning & visualization

MOHAMMAD SHABIB
19290116

Data Science:

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.

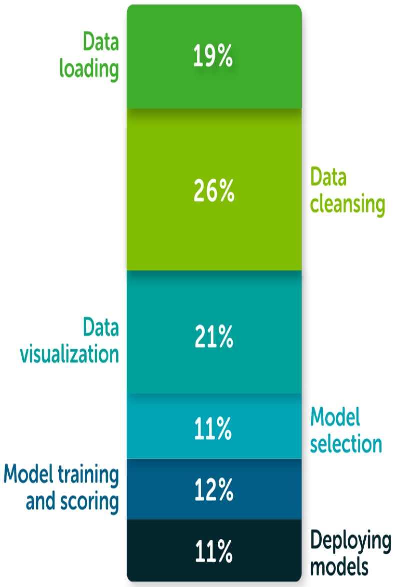
One of the main objects of Data science is The interpretation of big data. According to a research study by Forrester, Data-driven businesses are worth $1.2 trillion in 2020.

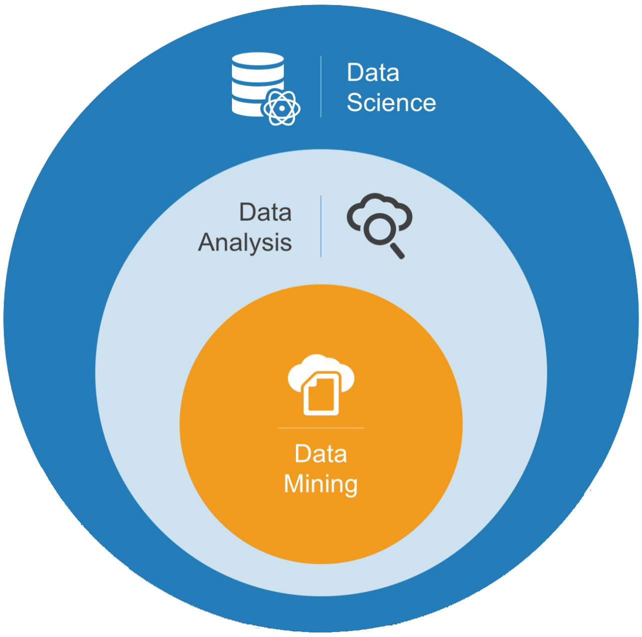
Data Science Vs Data Analysis Vs Data Mining

Data Mining: finding useful information in a dataset and utilizing that information to uncover hidden patterns.

Data Analysis: explain the history of the data and finding Conclusion.

Data Science: explain the history of the data and identifying the occurrence of a particular event in the future.





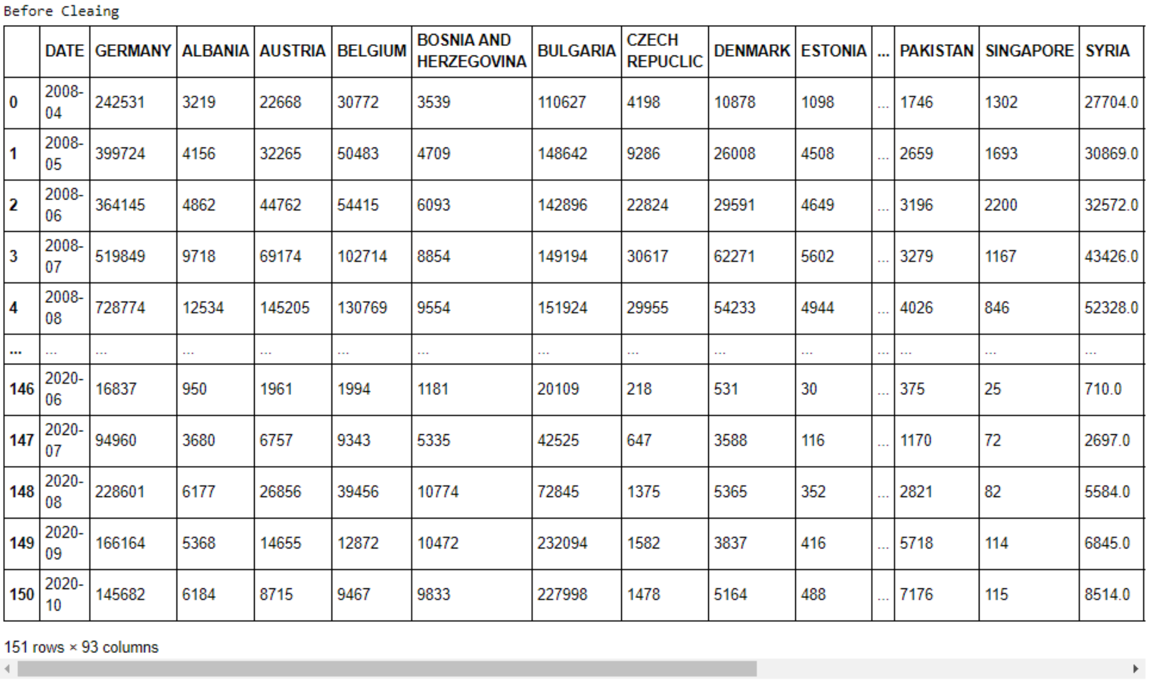
Python Code Implementation For Data Cleaning & Visualization

Source Code:
https://github.com/Hii1/Data-Scince/blob/main/Tourism_Turkey.ipynb
Bonus Code -Perceptron Implementation-:
https://github.com/Hii1/Data-Scince/blob/main/Perceptron.ipynb

Code Objective:

The main idea of the code is to visualize the tourism numbers in Turkey, The initial idea was to create create a graph between countries and the size of the node depends on the tourism numbers and the directed edge weights depend on the tourism number from that country, but this idea failed due to lack of data.

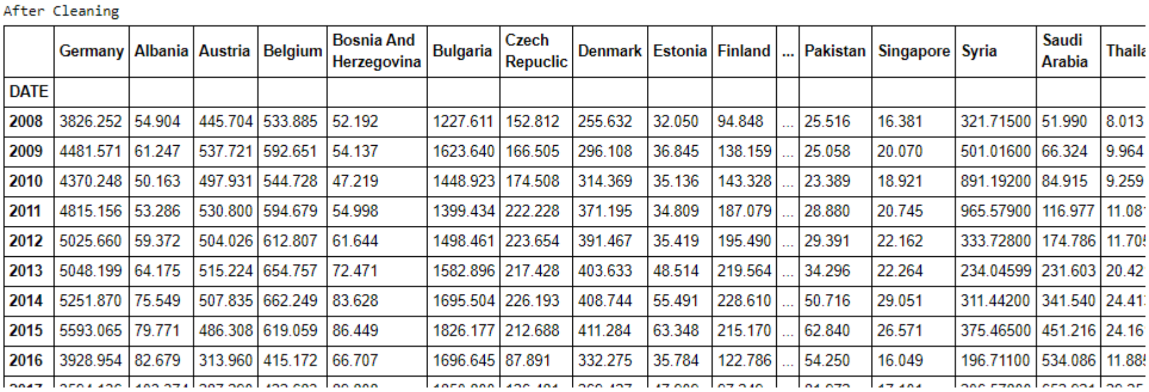
Data was cleaned and transformed, for example, the df_Turkey data set was transformed from numbers of tourists in each month to numbers each year and there was more cleaning too.



Before Cleaing

	DATE	GERMANY	ALBANIA	AUSTRIA	BELGIUM	BOSNIA AND HERZEGOVINA	BULGARIA	CZECH REPUCLIC	DENMARK	ESTONIA	...	PAKISTAN	SINGAPORE	SYRIA
0	2008-04	242531	3219	22668	30772	3539	110627	4198	10878	1098	...	1746	1302	27704.0
1	2008-05	399724	4156	32265	50483	4709	148642	9286	26008	4508	...	2659	1693	30869.0
2	2008-06	364145	4862	44762	54415	6093	142896	22824	29591	4649	...	3196	2200	32572.0
3	2008-07	519849	9718	69174	102714	8854	149194	30617	62271	5602	...	3279	1167	43426.0
4	2008-08	728774	12534	145205	130769	9554	151924	29955	54233	4944	...	4026	846	52328.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
146	2020-06	16837	950	1961	1994	1181	20109	218	531	30	...	375	25	710.0
147	2020-07	94960	3680	6757	9343	5335	42525	647	3588	116	...	1170	72	2697.0
148	2020-08	228601	6177	26856	39456	10774	72845	1375	5365	352	...	2821	82	5584.0
149	2020-09	166164	5368	14655	12872	10472	232094	1582	3837	416	...	5718	114	6845.0
150	2020-10	145682	6184	8715	9467	9833	227998	1478	5164	488	...	7176	115	8514.0

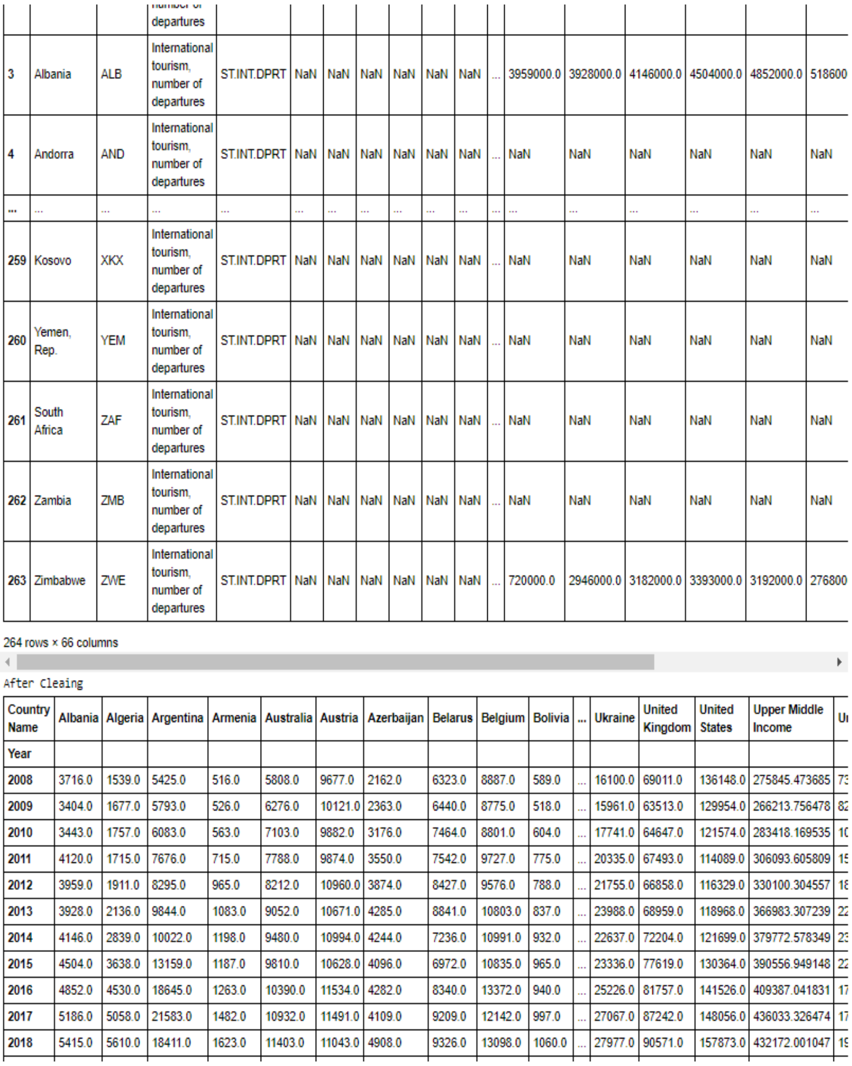
151 rows × 93 columns



After Cleaning

	Germany	Albania	Austria	Belgium	Bosnia And Herzegovina	Bulgaria	Czech Repuclic	Denmark	Estonia	Finland	...	Pakistan	Singapore	Syria	Saudi Arabia	Thaila
DATE																
2008	3826.252	54.904	445.704	533.885	52.192	1227.611	152.812	255.632	32.050	94.848	...	25.516	16.381	321.71500	51.990	8.013
2009	4481.571	61.247	537.721	592.651	54.137	1623.640	166.505	296.108	36.845	138.159	...	25.058	20.070	501.01600	66.324	9.964
2010	4370.248	50.163	497.931	544.728	47.219	1448.923	174.508	314.369	35.136	143.328	...	23.389	18.921	891.19200	84.915	9.259
2011	4815.156	53.286	530.800	594.679	54.998	1399.434	222.228	371.195	34.809	187.079	...	28.880	20.745	965.57900	116.977	11.08
2012	5025.660	59.372	504.026	612.807	61.644	1498.461	223.654	391.467	35.419	195.490	...	29.391	22.162	333.72800	174.786	11.70
2013	5048.199	64.175	515.224	654.757	72.471	1582.896	217.428	403.633	48.514	219.564	...	34.296	22.264	234.04599	231.603	20.42
2014	5251.870	75.549	507.835	662.249	83.628	1695.504	226.193	408.744	55.491	228.610	...	50.716	29.051	311.44200	341.540	24.41
2015	5593.065	79.771	486.308	619.059	86.449	1826.177	212.688	411.284	63.348	215.170	...	62.840	26.571	375.46500	451.216	24.16
2016	3928.954	82.679	313.960	415.172	66.707	1696.645	87.891	332.275	35.784	122.786	...	54.250	16.049	196.71100	534.086	11.88

The data set df_INT was harder to clean and transfer to the requested format, there were a lot of Nan values -Null values- so I used a threshold for deleting the country of 3 or more Nan values for each country and I used methods to transform columns to values and values to columns



			departures														
3	Albania	ALB	International tourism, number of departures	ST.INT.DPRT	NaN	NaN	NaN	NaN	NaN	NaN	...	3959000.0	3928000.0	4146000.0	4504000.0	4852000.0	518600
4	Andorra	AND	International tourism, number of departures	ST.INT.DPRT	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
259	Kosovo	XKX	International tourism, number of departures	ST.INT.DPRT	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
260	Yemen, Rep.	YEM	International tourism, number of departures	ST.INT.DPRT	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
261	South Africa	ZAF	International tourism, number of departures	ST.INT.DPRT	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
262	Zambia	ZMB	International tourism, number of departures	ST.INT.DPRT	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
263	Zimbabwe	ZWE	International tourism, number of departures	ST.INT.DPRT	NaN	NaN	NaN	NaN	NaN	NaN	...	720000.0	2946000.0	3182000.0	3393000.0	3192000.0	276800

264 rows × 66 columns

```
After Cleaing
```

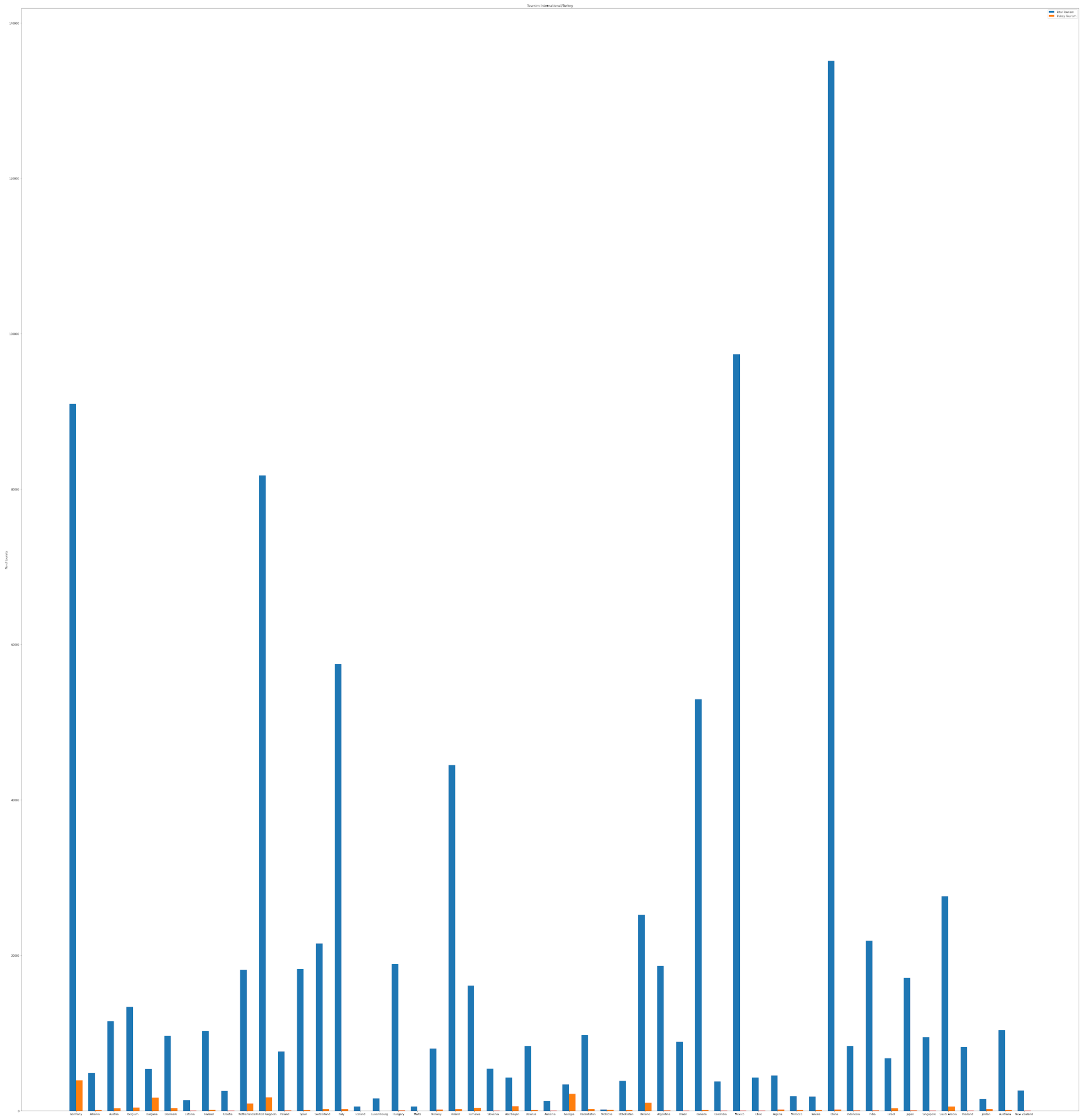
Country Name	Albania	Algeria	Argentina	Armenia	Australia	Austria	Azerbaijan	Belarus	Belgium	Bolivia	...	Ukraine	United Kingdom	United States	Upper Middle Income	Ui
Year																
2008	3716.0	1539.0	5425.0	516.0	5808.0	9677.0	2162.0	6323.0	8887.0	589.0	...	16100.0	69011.0	136148.0	275845.473685	73
2009	3404.0	1677.0	5793.0	526.0	6276.0	10121.0	2363.0	6440.0	8775.0	518.0	...	15961.0	63513.0	129954.0	266213.756478	82
2010	3443.0	1757.0	6083.0	563.0	7103.0	9882.0	3176.0	7464.0	8801.0	604.0	...	17741.0	64647.0	121574.0	283418.169535	10
2011	4120.0	1715.0	7676.0	715.0	7788.0	9874.0	3550.0	7542.0	9727.0	775.0	...	20335.0	67493.0	114089.0	306093.605809	15
2012	3959.0	1911.0	8295.0	965.0	8212.0	10960.0	3874.0	8427.0	9576.0	788.0	...	21755.0	66858.0	116329.0	330100.304557	18
2013	3928.0	2136.0	9844.0	1083.0	9052.0	10671.0	4285.0	8841.0	10803.0	837.0	...	23988.0	68959.0	118968.0	366983.307239	22
2014	4146.0	2839.0	10022.0	1198.0	9480.0	10994.0	4244.0	7236.0	10991.0	932.0	...	22637.0	72204.0	121699.0	379772.578349	23
2015	4504.0	3638.0	13159.0	1187.0	9810.0	10628.0	4096.0	6972.0	10835.0	965.0	...	23336.0	77619.0	130364.0	390556.949148	22
2016	4852.0	4530.0	18645.0	1263.0	10390.0	11534.0	4282.0	8340.0	13372.0	940.0	...	25226.0	81757.0	141526.0	409387.041831	17
2017	5186.0	5058.0	21583.0	1482.0	10932.0	11491.0	4109.0	9209.0	12142.0	997.0	...	27067.0	87242.0	148056.0	436033.326474	17
2018	5415.0	5610.0	18411.0	1623.0	11403.0	11043.0	4908.0	9326.0	13098.0	1060.0	...	27977.0	90571.0	157873.0	432172.001047	19

Data Visualization:

I decided to make 3 plots

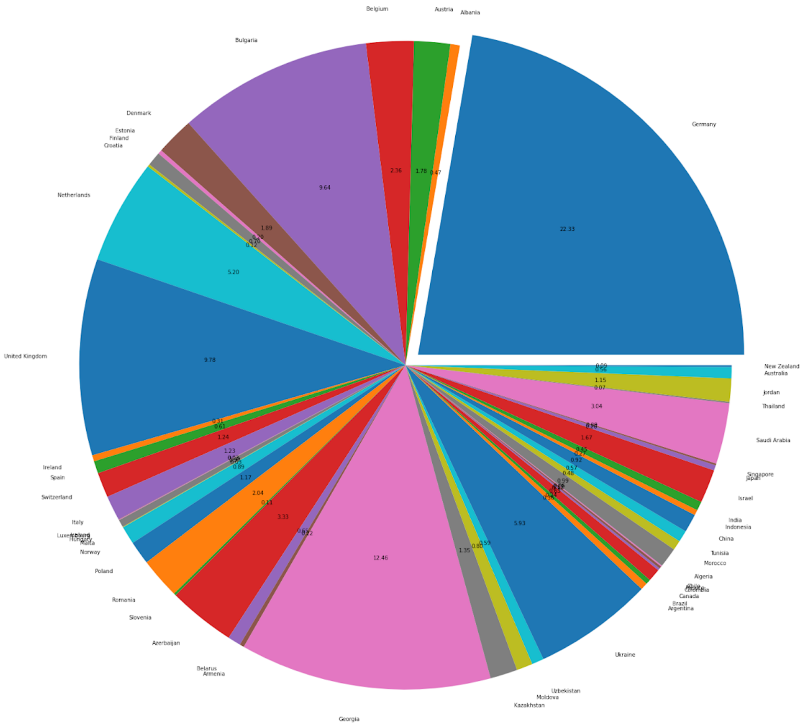
1’st plot - Bar Digram:

The first bar represents International tourism from each country and the second bar represents Torusim from each country to Turkey



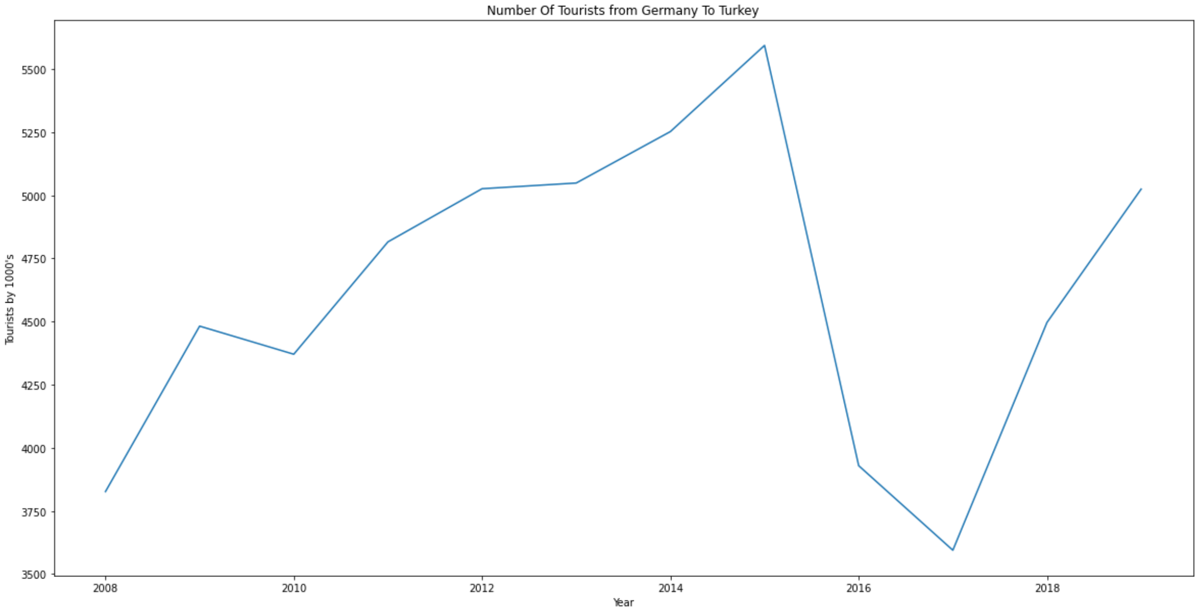
2’nd plot - Pie Diagram:

Each slice represents the percentage of country tourism to turkey, and the biggest percentage will be exploded out like Germany in the image below.



3’rd plot - Simple Plot:

The graph represents the tourism numbers on a specific country from 2008-2019


Number Of Tourists from Germany To Turkey
Tourists by 1000's
Year