Eksploracyjna Analiza Danych – Ryzyko zachorowania na raka płuc w 25 krajach

Karolina Kawulska, Daminika Dzeranhouskaya

Wyznaczenie celu biznesowego

Zmienna objaśniana - "Lung_Cancer_Diagnosis", chcemy na podstawie zmiennych objaśniających wykryć raka płuc.

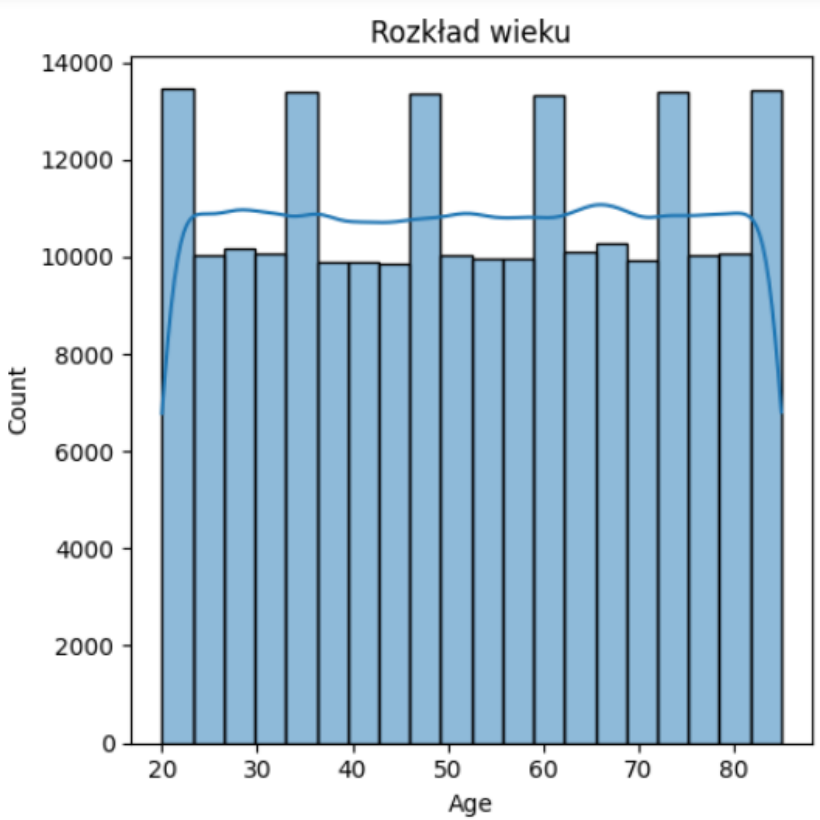
Zmienne objaśniane - wszystkie kolumny, oprócz tych dotyczących stadium choroby (np. wiek, płeć, palenie)

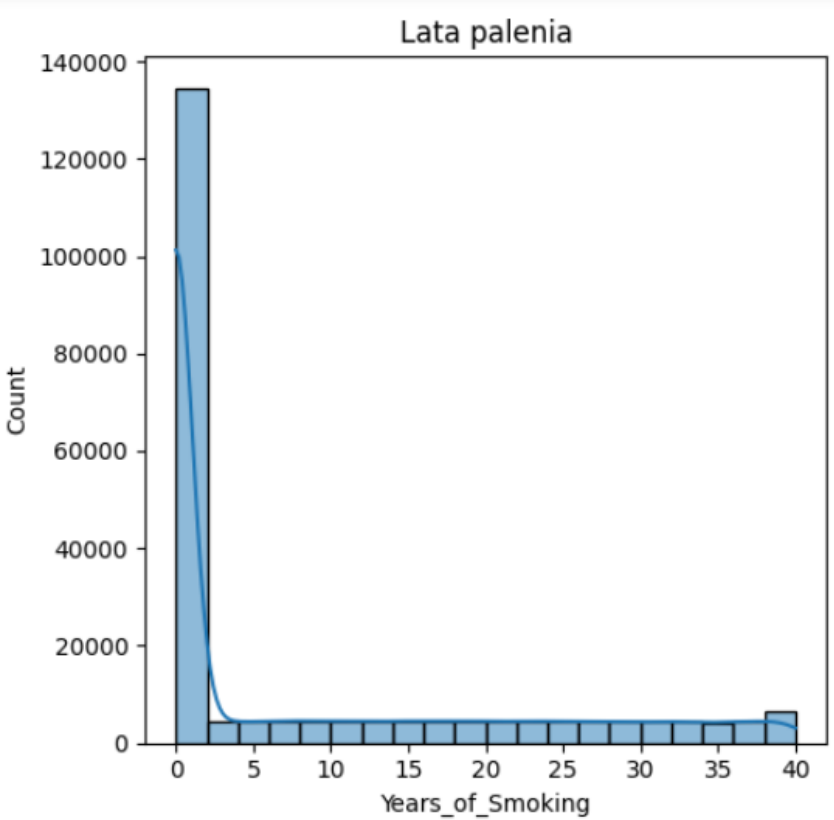
Analiza statystyczna i wizualizacja

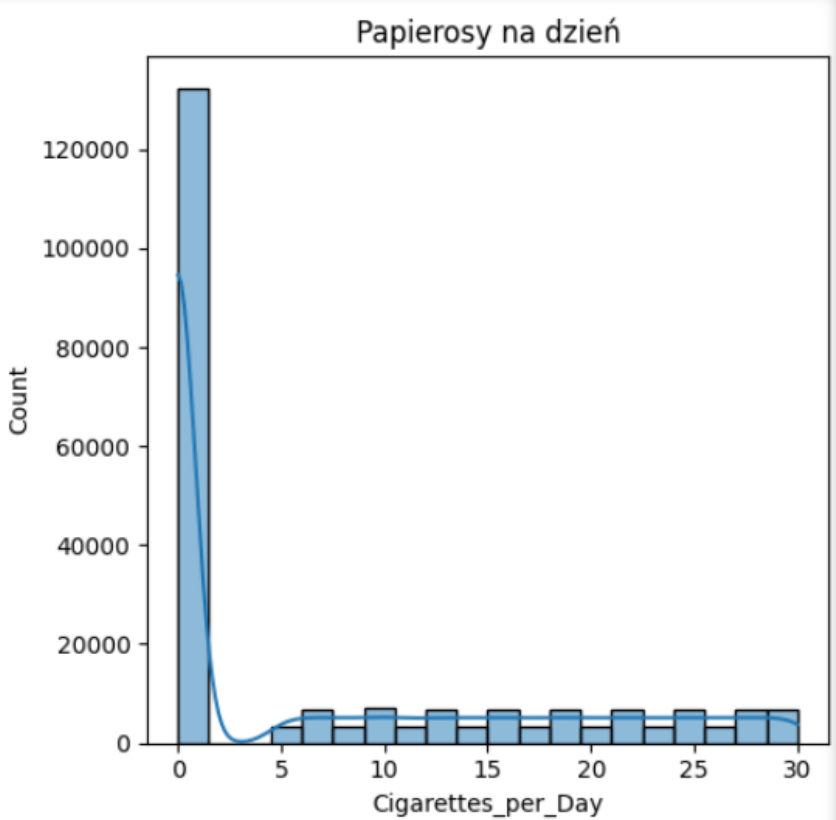
	Lung_Cancer_Prevalence_Rate	Mortality_Rate
count	220632.000000	220632.000000
mean	1.502085	3.049802
std	0.578043	14.924169
min	0.500000	0.000000
25%	1.000000	0.000000
50%	1.500000	0.000000
75%	2.000000	0.000000
max	2.500000	90.000000

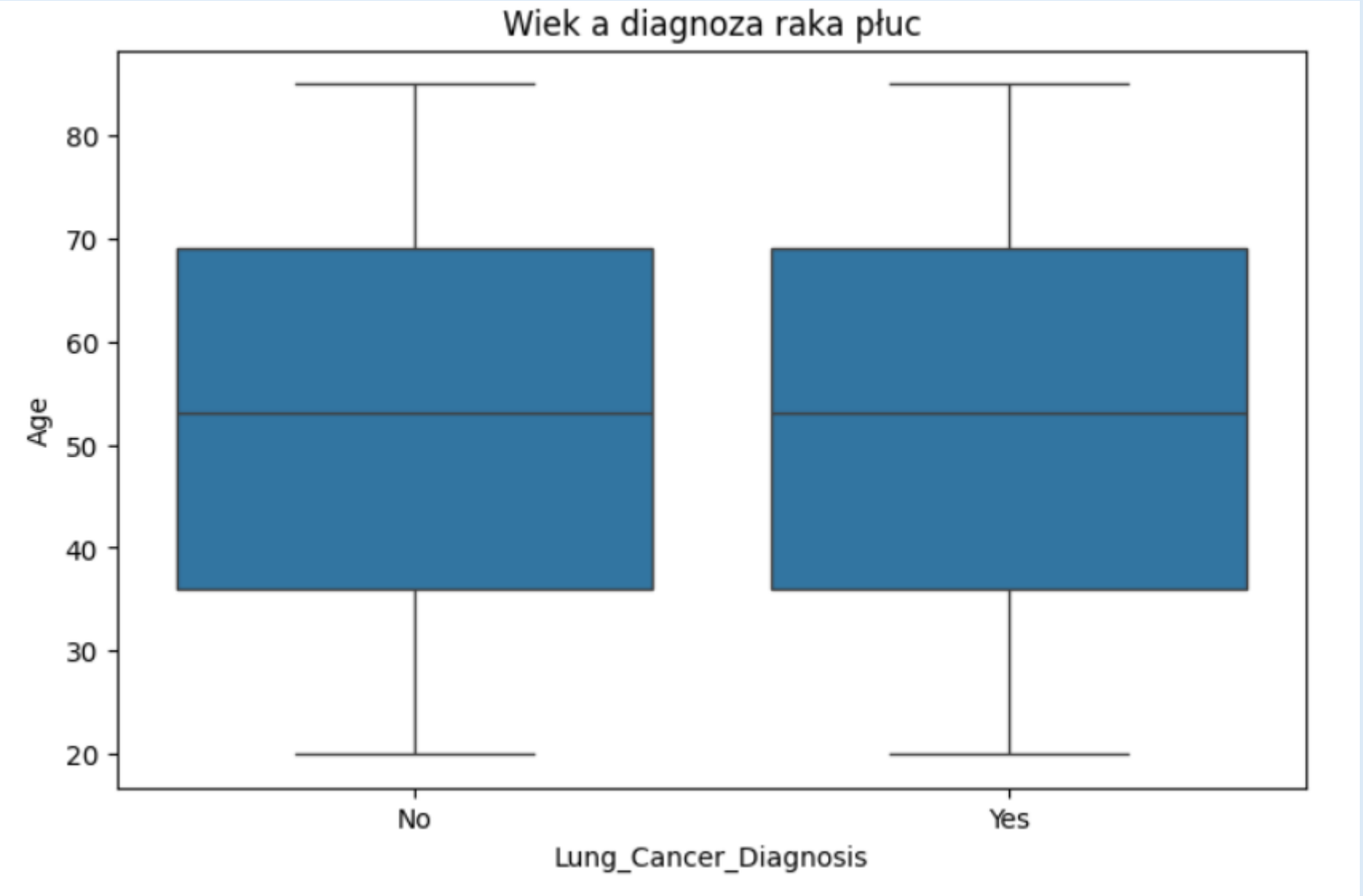
	Age	Years_of_Smoking
count	220632.000000	220632.000000
mean	52.518352	8.175274
std	19.078215	12.377248
min	20.000000	0.000000
25%	36.000000	0.000000
50%	53.000000	0.000000
75%	69.000000	15.000000
max	85.000000	40.000000

	Cigarettes_per_Day	Survival_Years	Annual_Lung_Cancer_Deaths
count	220632.000000	220632.000000	220632.000000
mean	7.007515	0.223526	63931.086928
std	9.802187	1.231025	130690.126777
min	0.000000	0.000000	10005.000000
25%	0.000000	0.000000	23000.000000
50%	0.000000	0.000000	30000.000000
75%	14.000000	0.000000	45000.000000
max	30.000000	10.000000	690000.000000


Rozkład wieku
Count
Age
0
2000
4000
6000
8000
10000
12000
14000
20
30
40
50
60
70
80


Lata palenia
Count
Years_of_Smoking
0
20000
40000
60000
80000
100000
120000
140000
0
5
10
15
20
25
30
35
40


Papierosy na dzień
Count
Cigarettes_per_Day
0
20000
40000
60000
80000
100000
120000
0
5
10
15
20
25
30

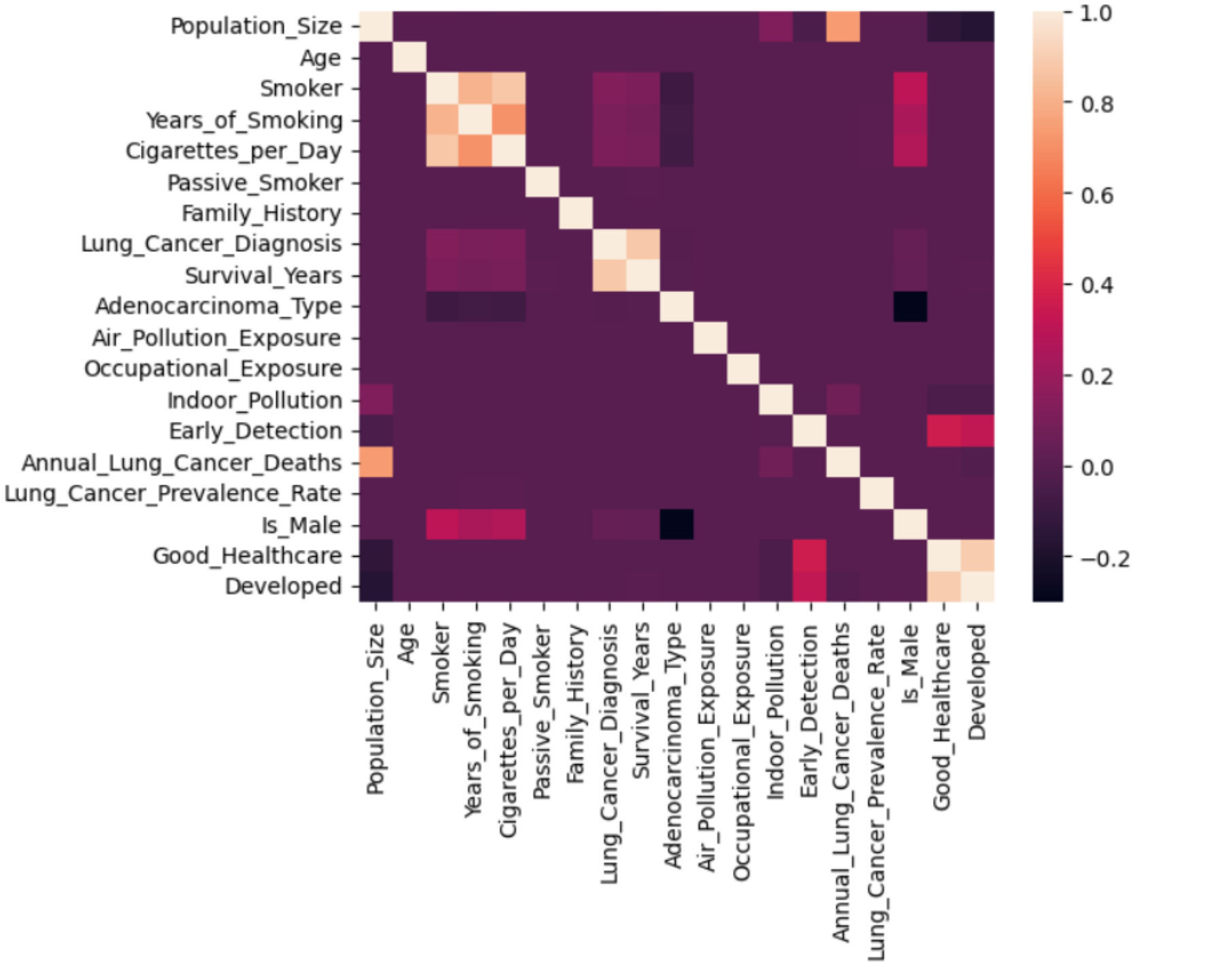

Wiek a diagnoza raka płuc
80
70
60
50
40
30
20
Age
No
Yes
Lung_Cancer_Diagnosis

Identyfikacja braków danych i anomalii

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 220632 entries, 0 to 220631
Data columns (total 24 columns):
 #   Column                      Non-Null Count   Dtype  
---  ------                      --------------   -----  
 0   ID                          220632 non-null  int64  
 1   Country                     220632 non-null  object 
 2   Population_Size             220632 non-null  int64  
 3   Age                         220632 non-null  int64  
 4   Gender                      220632 non-null  object 
 5   Smoker                      220632 non-null  object 
 6   Years_of_Smoking            220632 non-null  int64  
 7   Cigarettes_per_Day          220632 non-null  int64  
 8   Passive_Smoker              220632 non-null  object 
 9   Family_History              220632 non-null  object 
 10  Lung_Cancer_Diagnosis       220632 non-null  object 
 11  Cancer_Stage                8961 non-null    object 
 12  Survival_Years              220632 non-null  int64  
 13  Adenocarcinoma_Type         220632 non-null  object 
 14  Air_Pollution_Exposure      220632 non-null  object 
 15  Occupational_Exposure       220632 non-null  object 
 16  Indoor_Pollution            220632 non-null  object 
 17  Healthcare_Access           220632 non-null  object 
 18  Early_Detection             220632 non-null  object 
 19  Treatment_Type              6664 non-null    object 
 20  Developed_or_Developing     220632 non-null  object 
 21  Annual_Lung_Cancer_Deaths   220632 non-null  int64  
 22  Lung_Cancer_Prevalence_Rate 220632 non-null  float64
 23  Mortality_Rate              220632 non-null  float64
dtypes: float64(2), int64(7), object(15)
memory usage: 40.4+ MB
None
```

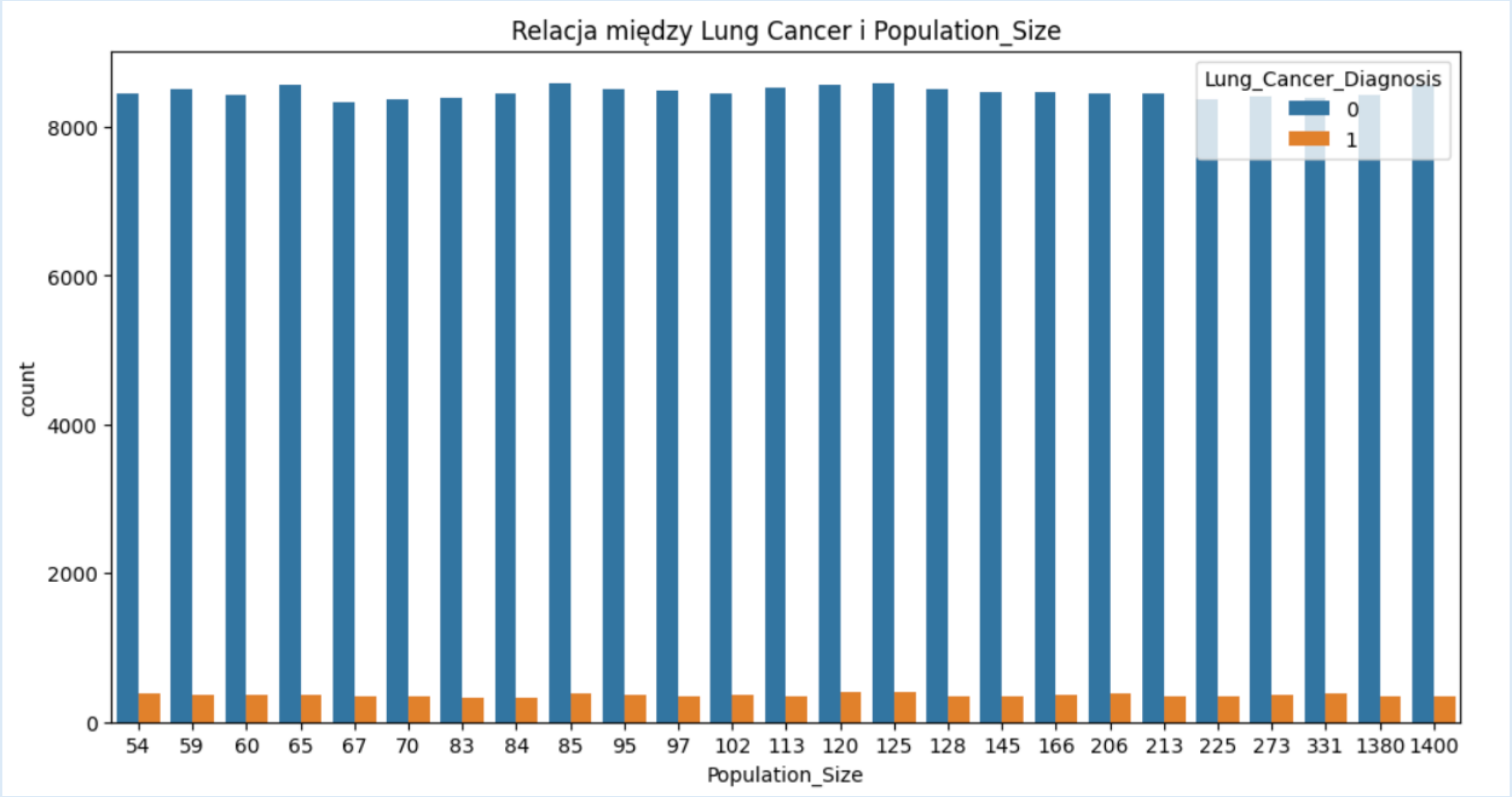
```
 Procent wartości null w każdej kolumnie:
 ID                                0.000000
Country                            0.000000
Population_Size                    0.000000
Age                                0.000000
Gender                             0.000000
Smoker                             0.000000
Years_of_Smoking                   0.000000
Cigarettes_per_Day                 0.000000
Passive_Smoker                     0.000000
Family_History                     0.000000
Lung_Cancer_Diagnosis              0.000000
Cancer_Stage                      95.938486
Survival_Years                     0.000000
Adenocarcinoma_Type                0.000000
Air_Pollution_Exposure             0.000000
Occupational_Exposure              0.000000
Indoor_Pollution                   0.000000
Healthcare_Access                  0.000000
Early_Detection                    0.000000
Treatment_Type                    96.979586
Developed_or_Developing            0.000000
Annual_Lung_Cancer_Deaths          0.000000
Lung_Cancer_Prevalence_Rate        0.000000
Mortality_Rate                     0.000000
dtype: float64

 Całkowita liczba zduplikowanych rekordów:  0
```

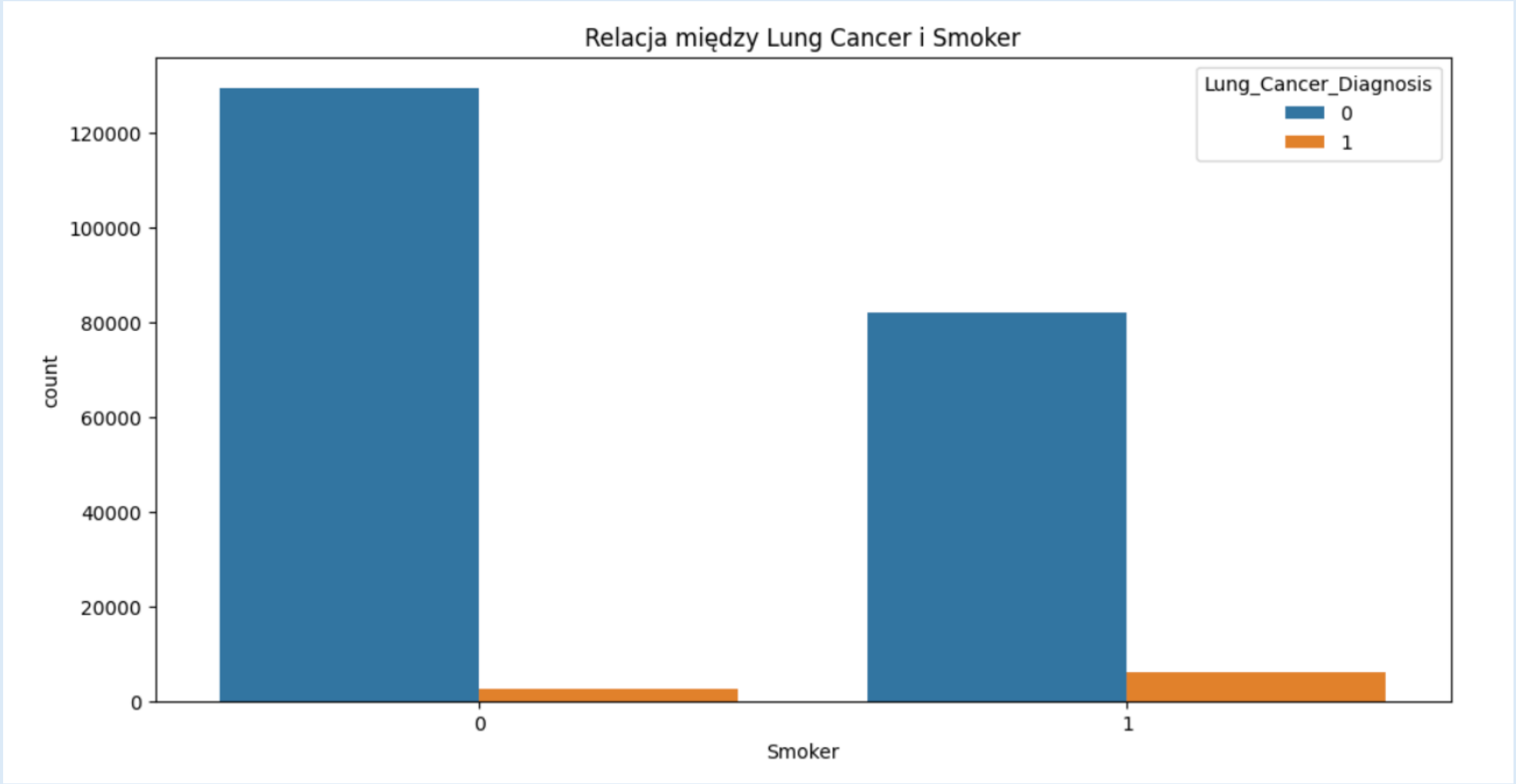
Badanie zależności między zmiennymi

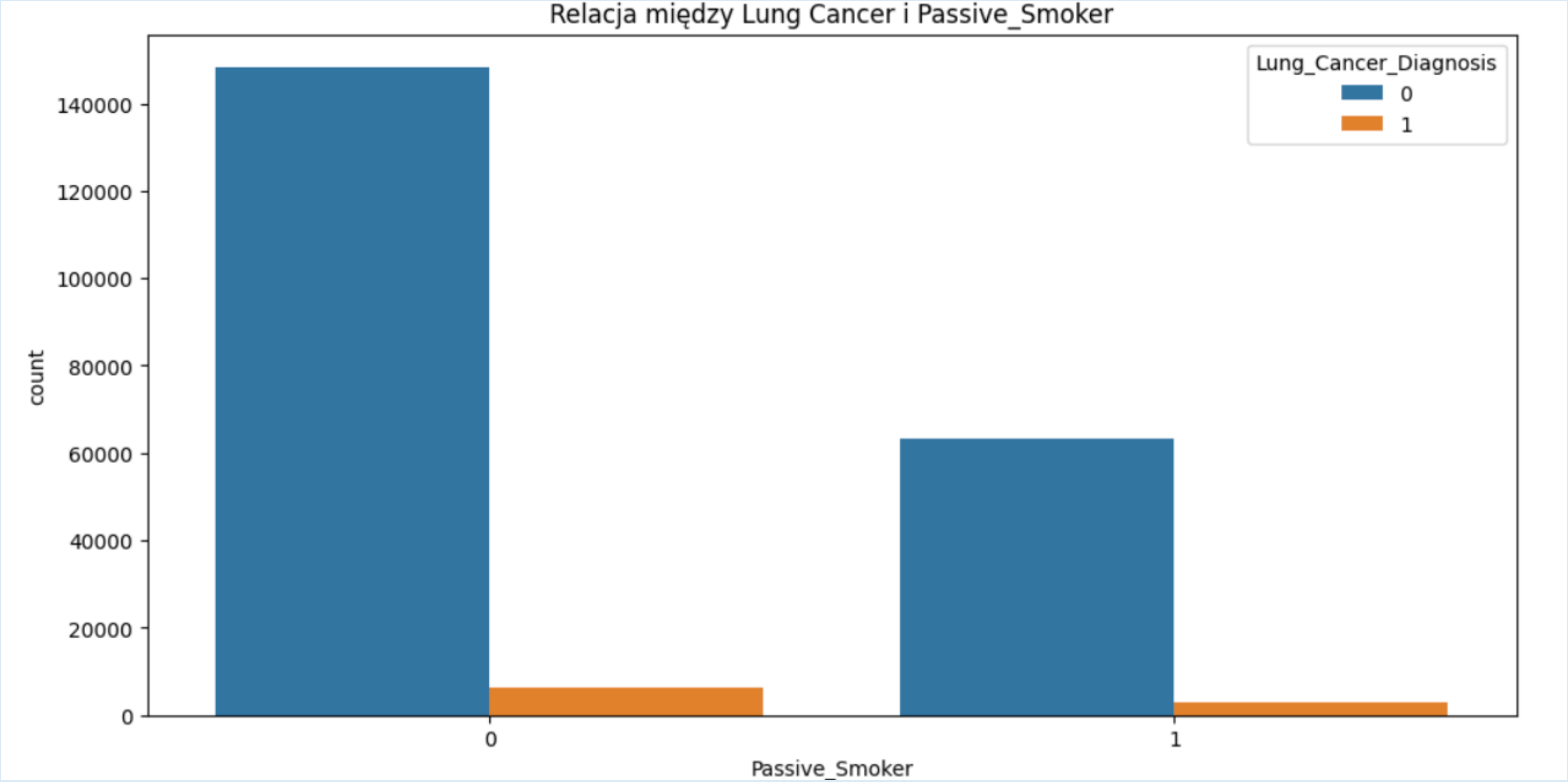

Population_Size
Age
Smoker
Years_of_Smoking
Cigarettes_per_Day
Passive_Smoker
Family_History
Lung_Cancer_Diagnosis
Survival_Years
Adenocarcinoma_Type
Air_Pollution_Exposure
Occupational_Exposure
Indoor_Pollution
Early_Detection
Annual_Lung_Cancer_Deaths
Lung_Cancer_Prevalence_Rate
Is_Male
Good_Healthcare
Developed
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

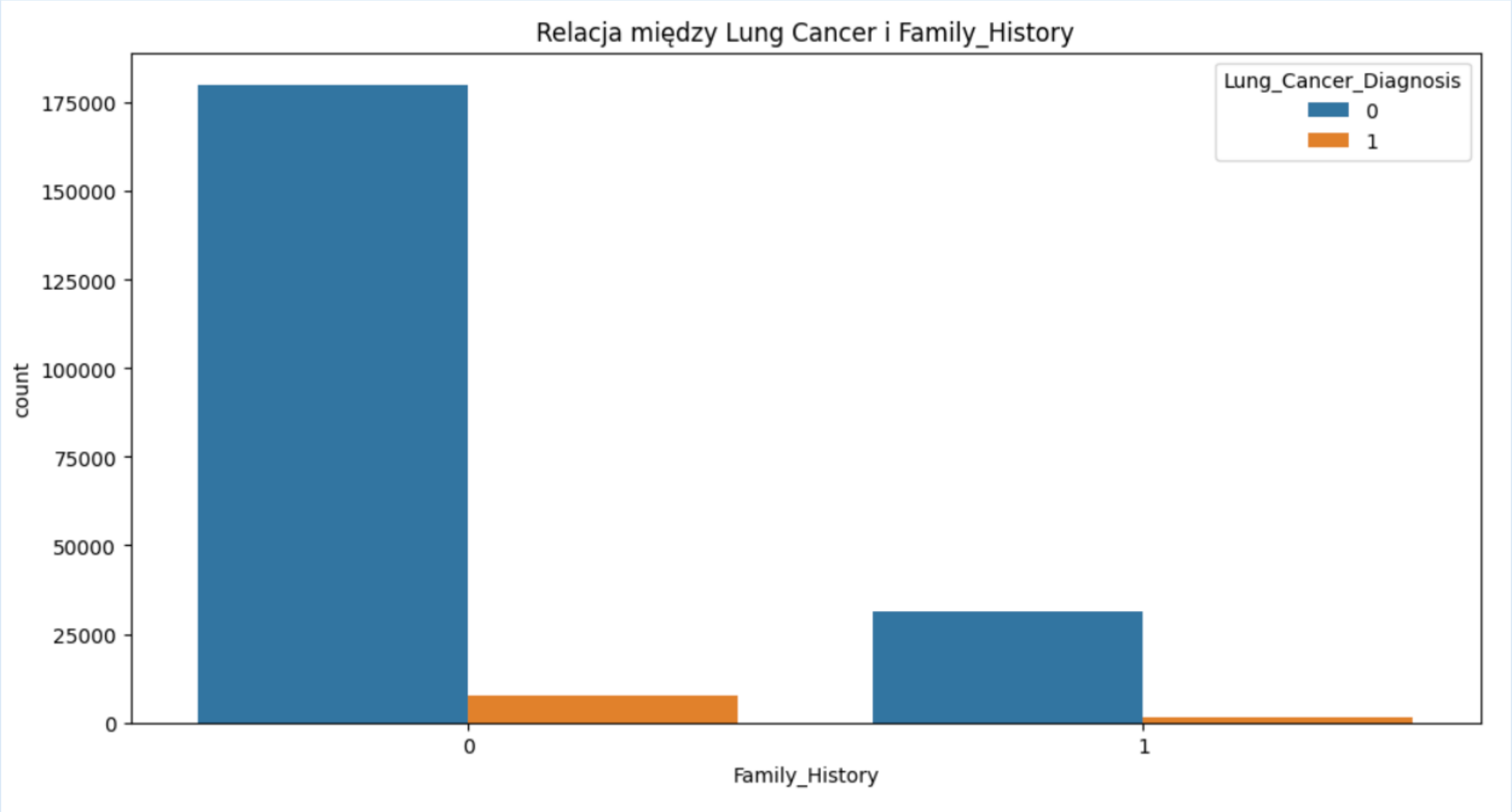
Największe korelacje z "Lung_Cancer_Diagnosis"

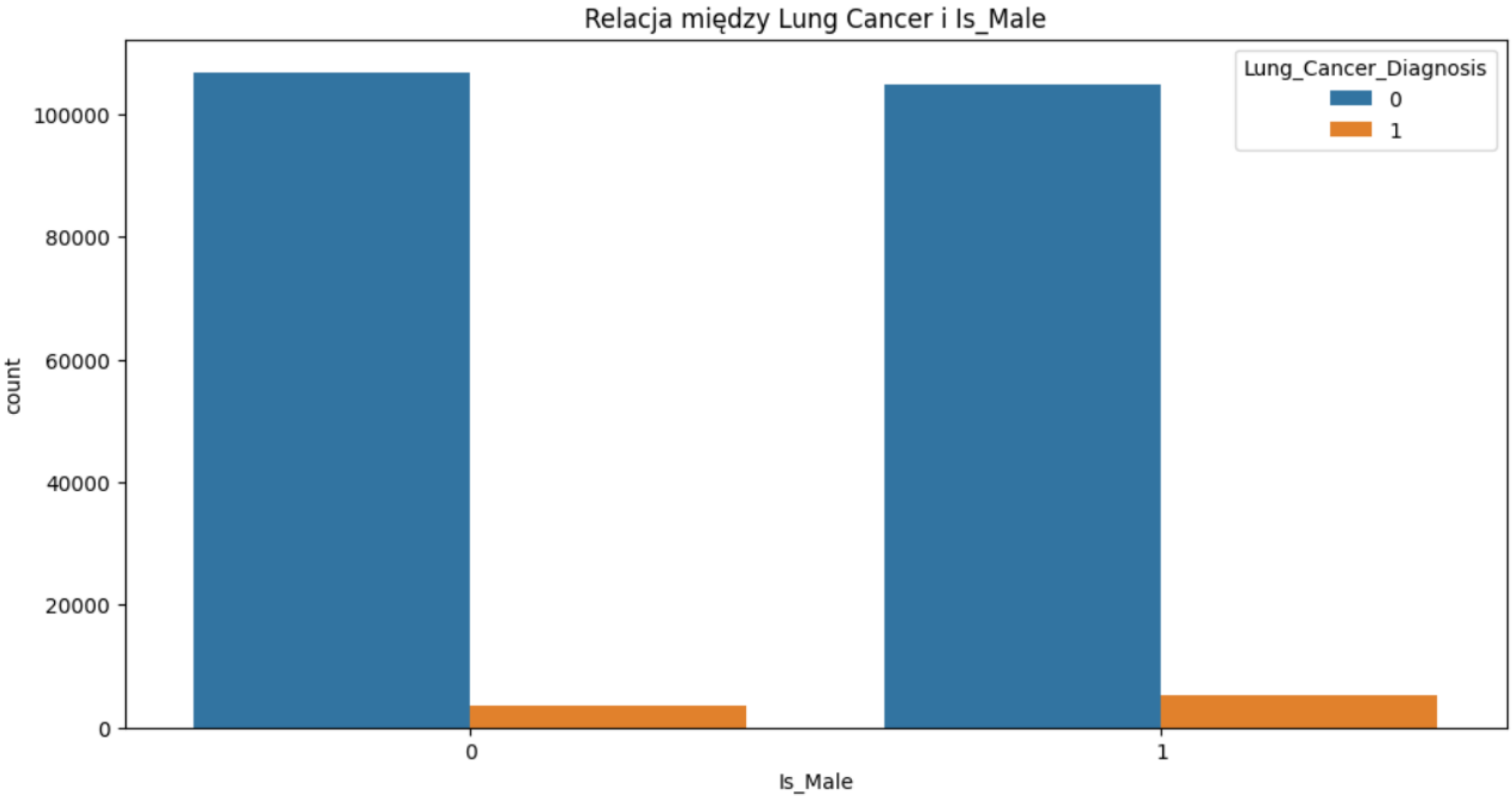
- Ciggarettes_per_day
- Years_of_Smoking
- Smoker


Relacja między Lung Cancer i Population_Size
Lung_Cancer_Diagnosis
0
1
count
8000
6000
4000
2000
0
54
59
60
65
67
70
83
84
85
95
97
102
113
120
125
128
145
166
206
213
225
273
331
1380
1400
Population_Size


Relacja między Lung Cancer i Smoker
Lung_Cancer_Diagnosis
0
1
120000
100000
80000
60000
40000
20000
0
count
0
1
Smoker


Relacja między Lung Cancer i Passive_Smoker
Lung_Cancer_Diagnosis
0
1
140000
120000
100000
80000
60000
40000
20000
0
count
0
1
Passive_Smoker


Relacja między Lung Cancer i Family_History
Lung_Cancer_Diagnosis
0
1
175000
150000
125000
100000
75000
50000
25000
0
count
0
1
Family_History


Relacja między Lung Cancer i Is_Male
Lung_Cancer_Diagnosis
0
1
100000
80000
60000
40000
20000
0
count
0
1
Is_Male