Analiza podataka iz datoteke toy_dataset.csv

Kolone (varijable) su: 'Number', 'City', 'Gender', 'Age', 'Income', 'Illness'

Analiza podataka provedena je s ciljem utjecaja varijabli , 'City', 'Gender', 'Age', 'Income'
na razbolijevanje osobe.

Numeričke varijable su: 'Age' i 'Income'
Kategorijske varijable su: 'Illness', 'City' i 'Gender'

Opis kategorijskih varijabli:

Illness
count 150000
unique 2
top No
freq 137861
Name: Illness, dtype: object

City
count 150000
unique 8
top New York City
freq 50307
Name: City, dtype: object

Gender
count 150000
unique 2
top Male
freq 83800
Name: Gender, dtype: object

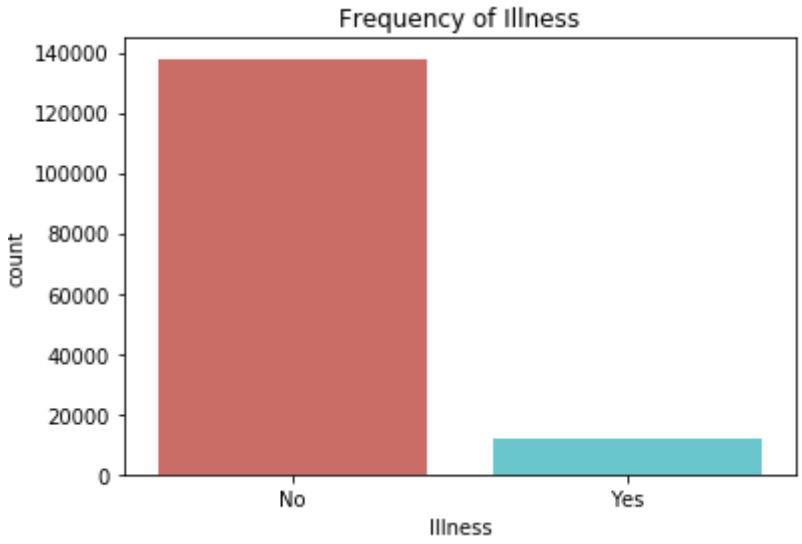
Postotak zdravih osoba: 91.9%
Postotak bolesnih osoba: 8.093%
Dakle, klase podataka su neuravnotežene, a omjer zdravih i bolesnih ispitanika je 92:8.

<u>Obsrevacije</u>

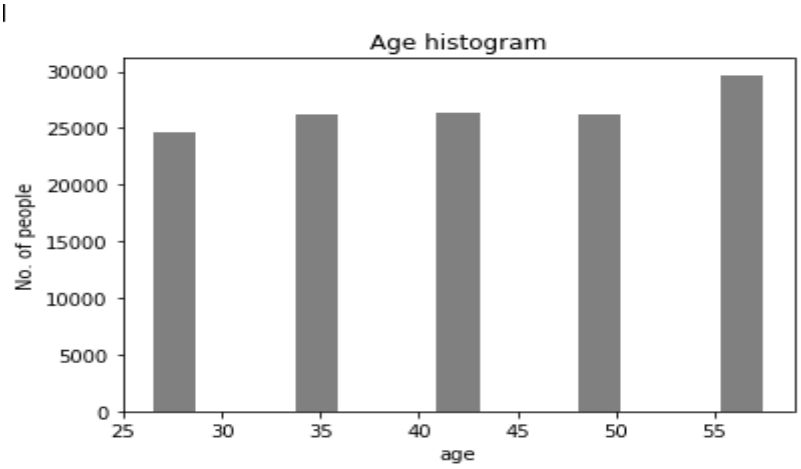
	Age	Income
Illness		
No	44.943980	91250.590174
Yes	45.020842	91277.875360

Prosječna dob osobe koja se razboljela je nešto veća od prosječne dobi osobe koja se nije razbolila.

Prosječna primanja osobe koja se razboljela je nešto veća od prosječnih primanja osobe koja se nije razbolila.

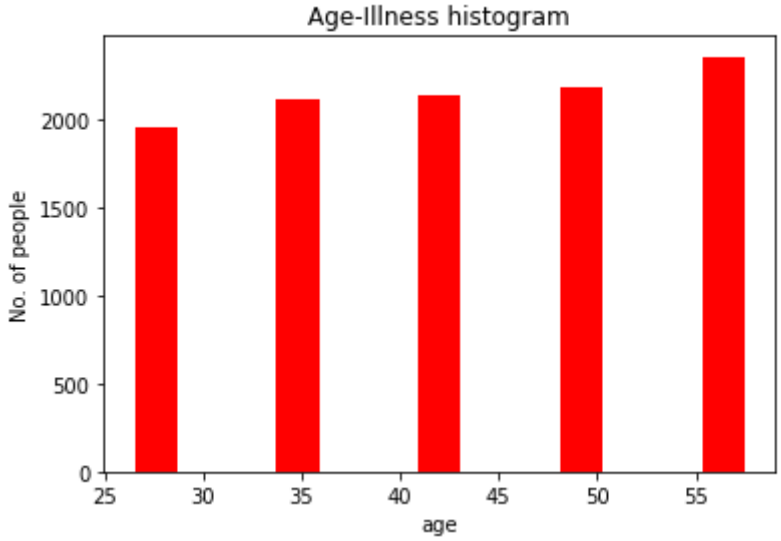


Prosječna dob ispitanika je 44.9502 i ona je veća od prosječne dobi bolesnog ispitanika te manja od prosječne dobi zdravog ispitanika.
Raspon godina (starosti) ispitanika je [25,65], a najveći broj ispitanika je u dobi između 55 i 65 godina, za koje se može očekivati da stopa oboljenja bude veća, a najmanji broj ispitanika je u dobi između 25 i 30 godina.



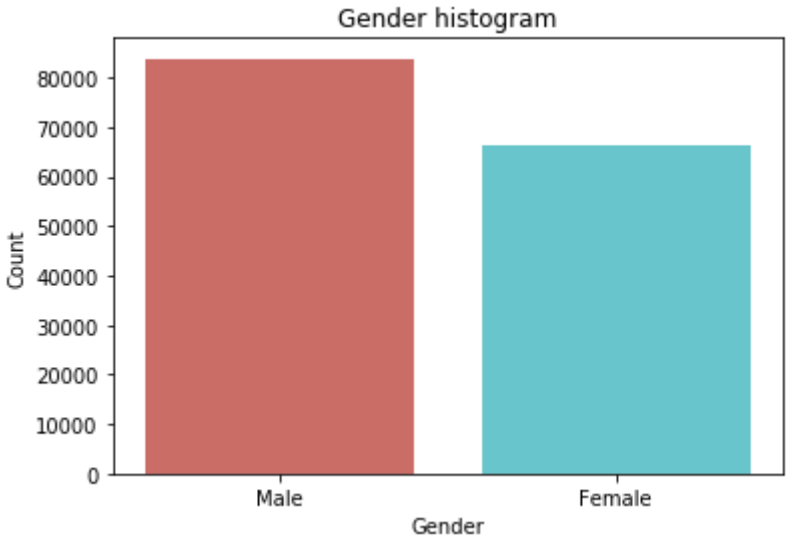
Hist 1

Među ispitanim oboljelim ispitanicima najveći broj je onih između 55 i 65 godina što i pokazuje (Hist2).

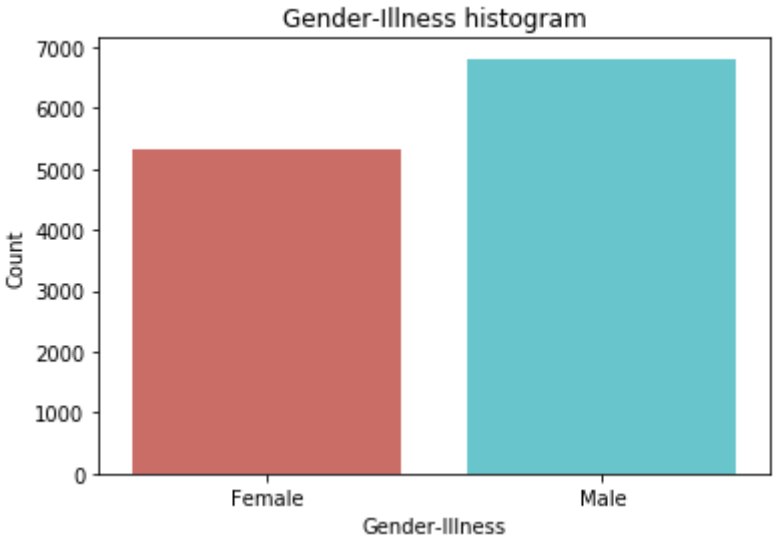


Hist 2

Broj ispitanika muškog spola (6808) je veći od broja ispitanika ženskog spola (5331) što prikazuje Hist3.



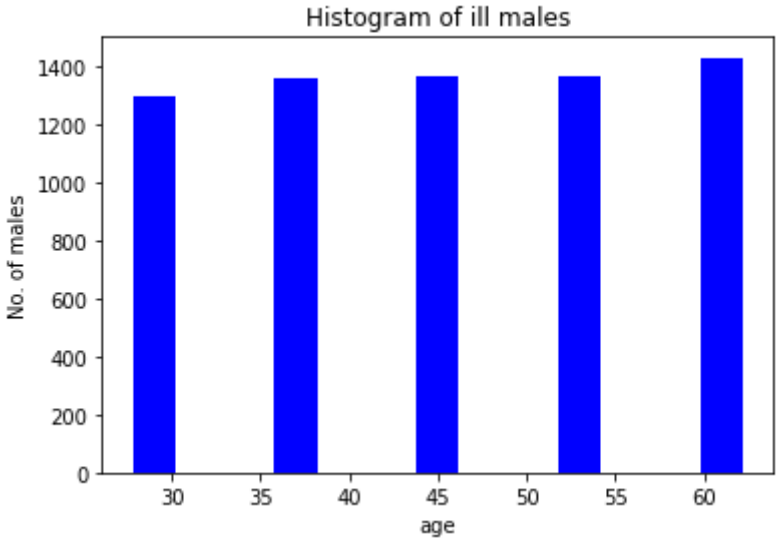
Hist 3



Hist 4

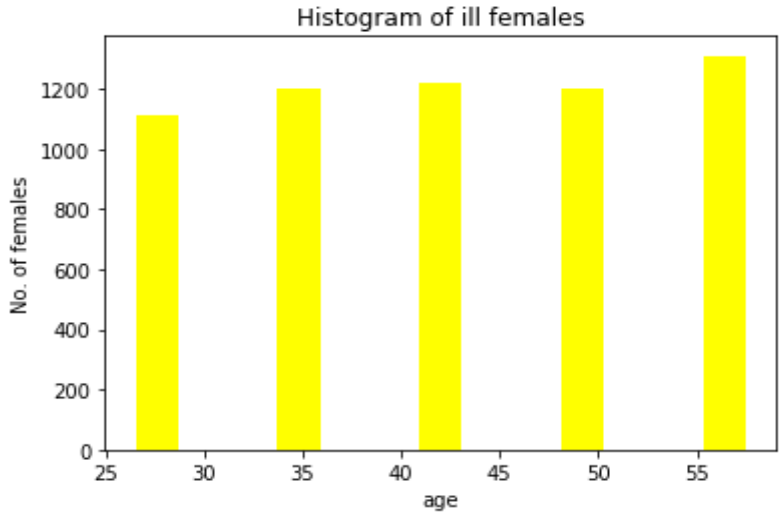
Broj bolesnih ispitanika muškog spola (6808) je veći od broja bolesnih ispitanika ženskog spola (5331) što je vidljivo iz Hist4.
Stopa obolijevanja muških ispitanika (8.124%) je veća od stope obolijevanja ženskih ispitanika (8.053%).

Najveći broj bolesnih ispitanika muškog i ženskog spola je iz dobnog intervala [55, 65] – Hist5 i Hist6



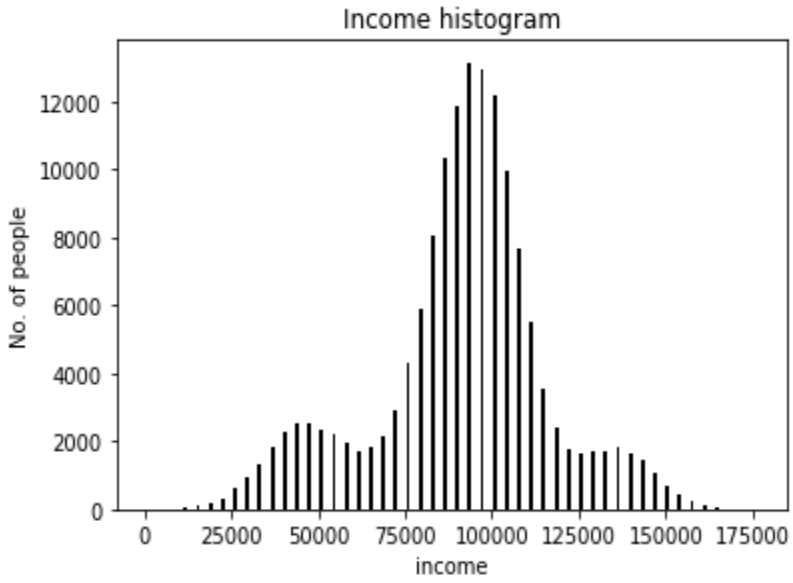
Hist 5

Odnosno obolijevanje kod jednog i kod drugog spola poveća se s godinama što je i očekivano.



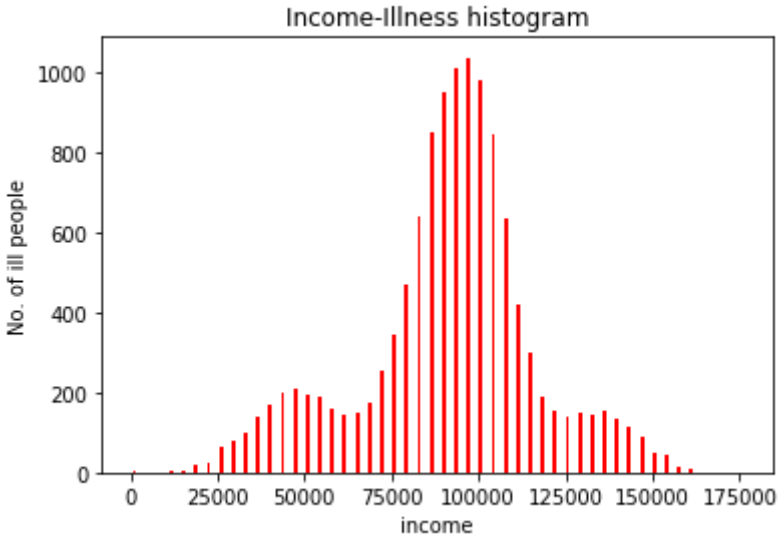
Hist 6

Srednja vrijednost primanja ispitanika je 91252.798273 odnosno najviše ispitanika ima primanja koja su približno jednaka toj vrijednosti. Standardna devijacija je: 24989.41764987958



Hist 7

Srednja vrijednost primanja bolesnih ispitanika je 91277.875360 koja je približno jednaka prosječnim primanjima svih ispitanika odnosno distribucija je ista – Hist 8



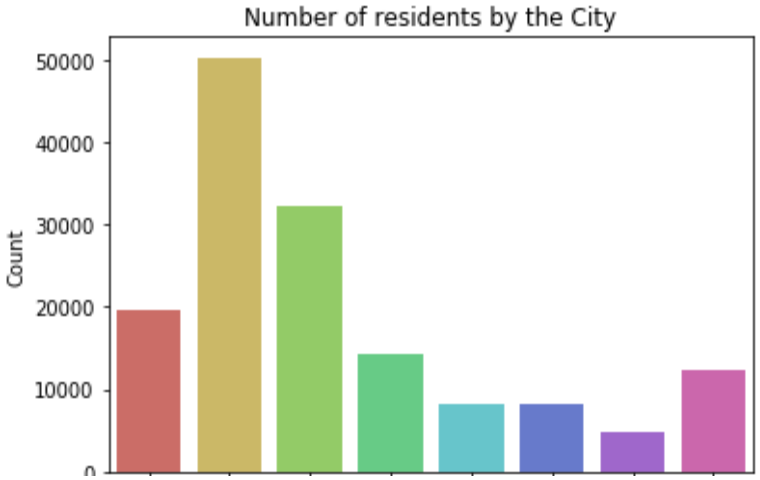
Hist 8

Prosječne vrijednosti starosti i primanja bolesnih ispitanika po gradovima su iz čega se ne može zaključiti linearnost:

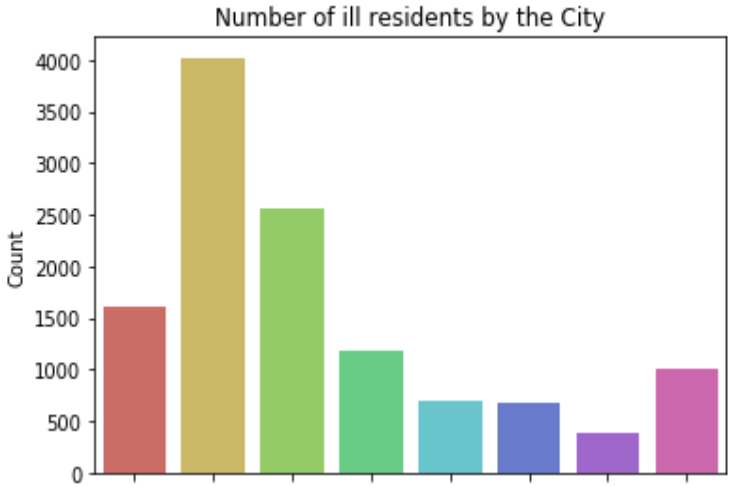
City	Age	Income
Austin	44.629080	90210.543027
Boston	45.408163	92013.594752
Dallas	44.992560	45275.147551
Los Angeles	45.297508	94976.434190
Mountain View	44.846350	134770.105263
New York City	44.996021	97061.376026
San Diego	45.210660	101940.631980
Washington D.C.	44.565868	71200.691617

Najveći broj ispitanika je iz New York City-u a najmanji iz San Diega- Hist 9. Također, najveći broj bolesnih ispitanika je je iz New York City-u a najmanji iz San Diega – Hist 10.

New York City	50307	4021
Los Angeles	32173	2568
Dallas	19707	1613
Mountain View	14219	1178
Austin	12292	1011
Boston	8301	686
Washington D.C.	8120	668
San Diego	4881	394



Hist 9



Hist 10

Također, analizom po gradovima dobivaju se sljedeći izračuni za stope obolijevanja iz čega je vidljivo da je stopa oboljenja najveća u Bostonu (8.264%) a najmanja u Los Angelesu (7.982%)

```
Rate of illness in NYC: 7.992923450016896
Rate of illness in LA: 7.981848133528113
Rate of illness in Dallas:  8.184908915613741
Rate of illness in Mountain View:  8.284689499964836
Rate of illness in Austin:   8.2248616986658
Rate of illness in Boston:  8.264064570533671
Rate of illness in WDC:   8.226600985221674
Rate of illness in San Diego:  8.072116369596394
```

Analiza oboljenja po gradovima za ispitanike muškog spola pokazuje da je stopa oboljenja ispitanika muškog spola najveća u San Diegu (8.66%) a najmanja u Los Angelesu (7.933%).

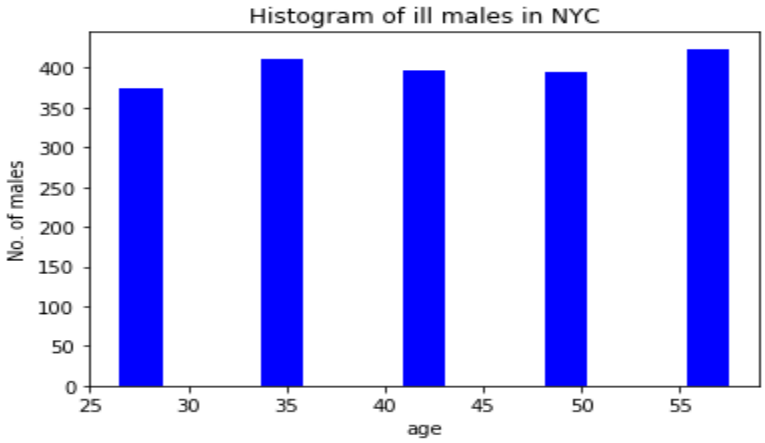
```
Rate of Illness for male in NY is:  7.982939399324684
Rate of Illness for male in LA is:  7.9326923076923075
Rate of Illness for male in Dallas is:  8.225777131430648
Rate of Illness for male in Mountain View is: 8.271433967014982
Rate of Illness for male in Austin is:  8.382800057912263
Rate of Illness for male in Boston is:  8.351504579153946
Rate of Illness for male in WDC is: 8.296751536435469
```

```
Rate of Illness for male in San Diego is: 8.659719726913401
```

Analiza oboloijavanja po gradovima za ispitanike ženskog spola pokazuje da je stopa obolijevanja ispitanika ženskog spola najveća u Mountain View-u u (8.3%) a najmanja u San Diegu (7.292%).

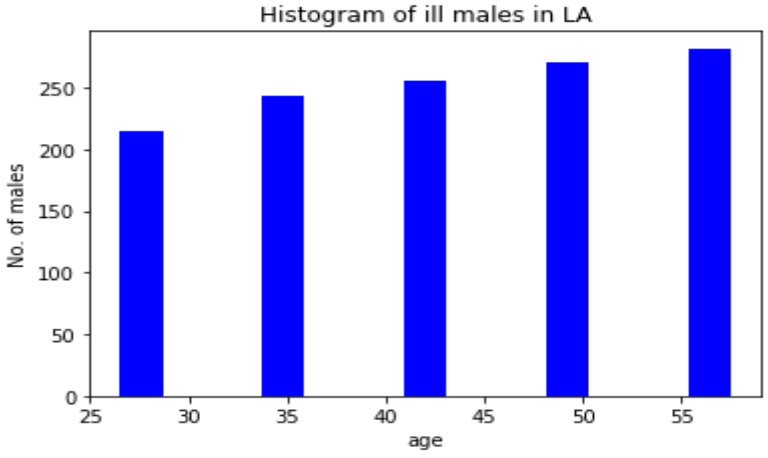
```
Rate of Illness for female in NY is: 8.005592639364965
Rate of Illness for female in LA is: 8.043402170108505
Rate of Illness for female in Dallas is: 8.13325674899483
Rate of Illness for female in Mountain View is: 8.30146590184831
Rate of Illness for female in Austin is: 8.022284122562674
Rate of Illness for female in Boston is: 8.156123822341858
Rate of Illness for female in WDC is: 8.136924803591471
Rate of Illness for female in San Diego is: 7.292659675881792
```

U nastavku je analiza oboljenja obzirom na spol, godine i mjesto življenja.
Broj bolesnih ispitanika muškog spola u NYC-u je najveći za dobni interval [55, 65] – Hist 11
Također, povećano oboljenje u NYC-u je za muškarce starosti približno 35 godina.



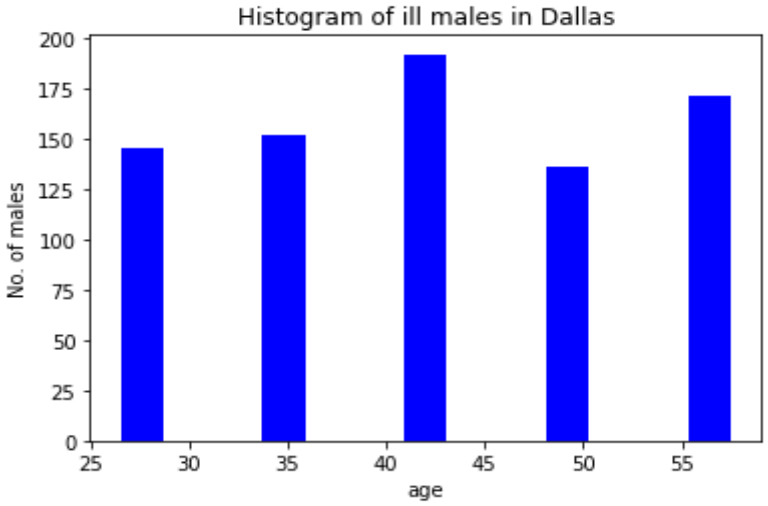
Hist 11

Broj bolesnih ispitanika muškog spola u Los angelesu povećava se s godinama starosti i on je najveći za dobni interval [55, 65] – Hist 12.



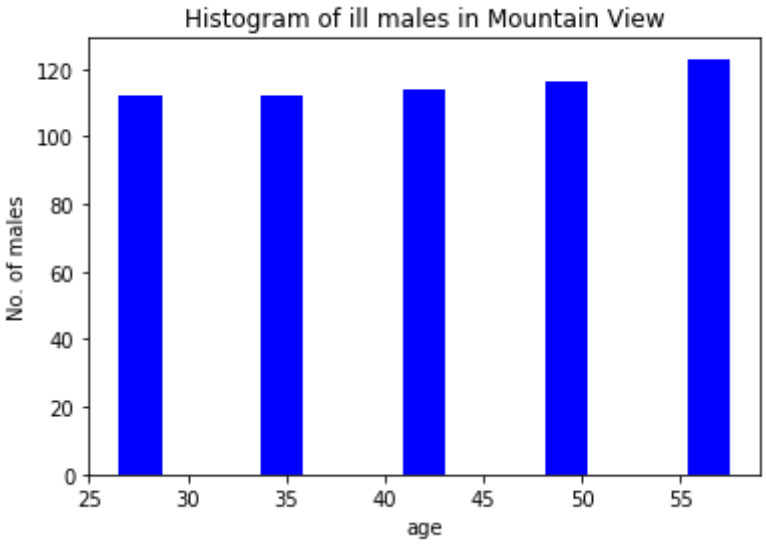
Hist 12

Broj bolesnih ispitanika muškog spola u Dallasu nije proporcionalan s godinama starosti i on je najveći za dobni interval [40, 45] – Hist 13.



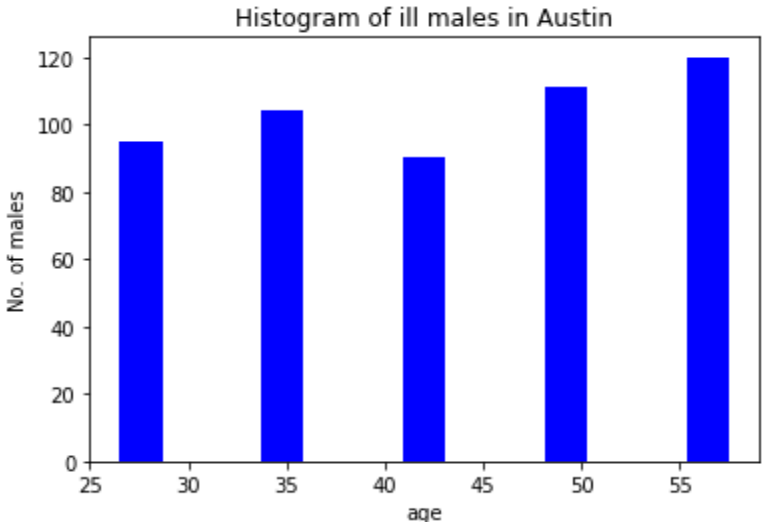
Hist 13

Broj bolesnih ispitanika muškog spola u Mountain View-u se postepeno povećava s godinama starosti i on je najveći za dobni interval [55, 65] – Hist 14.

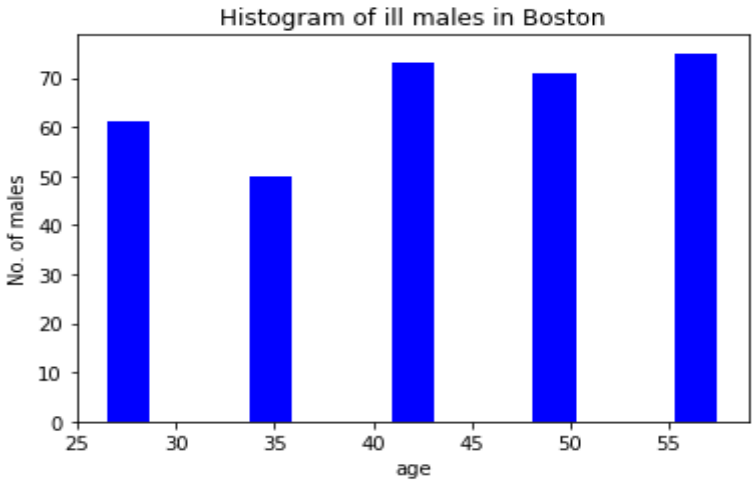


Hist – 14

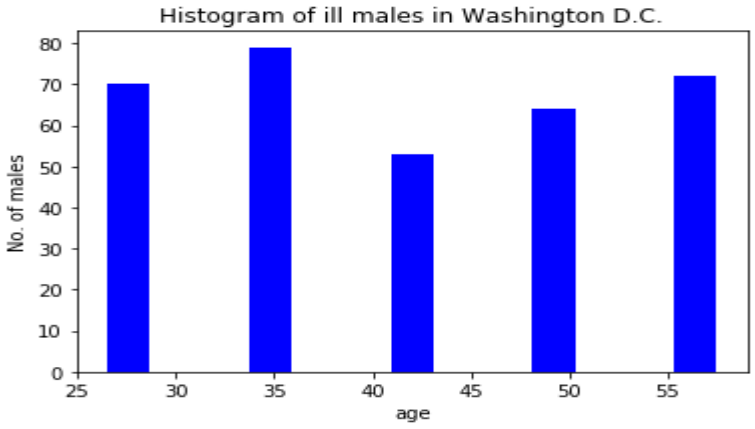
Slične obzervacije dobivaju se za ispitanike muškog spola i za ostale gradove temeljem kojih se ne može općenito zaključiti da se vjerojatnost oboljenja ne povećava nužno s godinama starosti budući da je u nekim gradovima stopa oboljenja veća kod mlađih ispitanika.



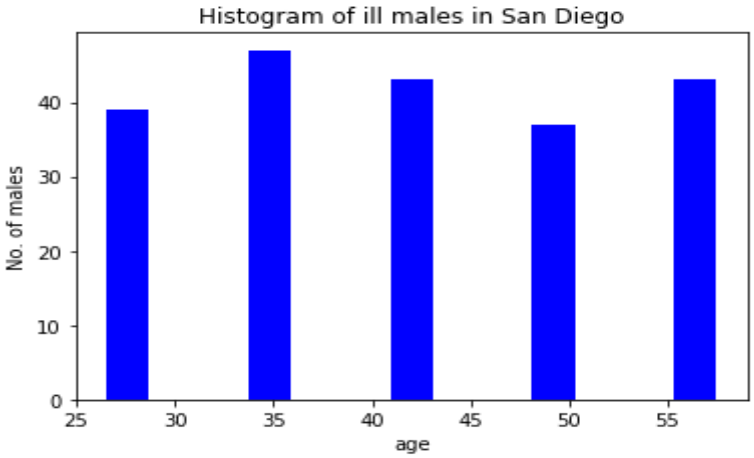
Hist - 15



Hist - 16

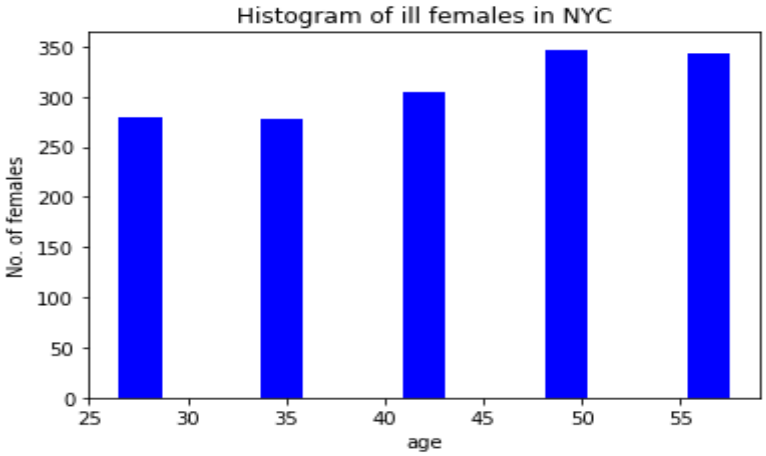


Hist - 17

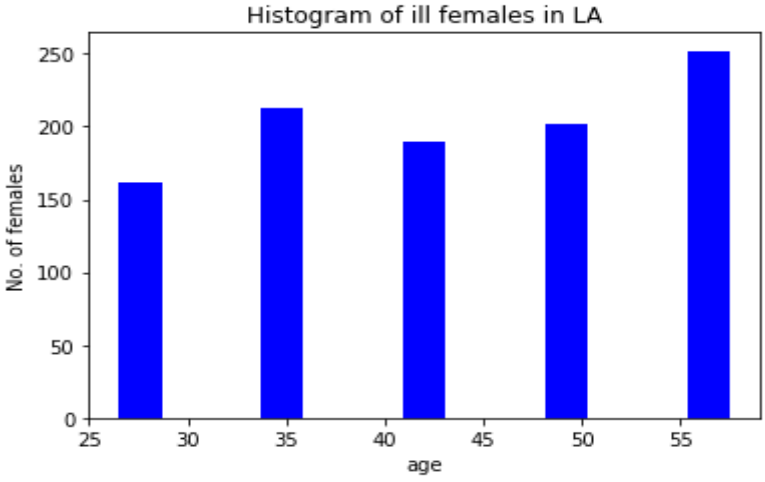


Hist – 18

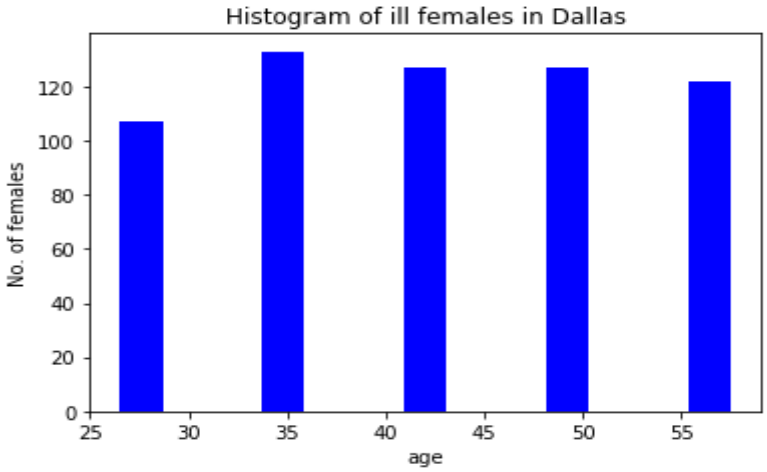
U nastavku je analiza oboljenja ispitanika ženskog spola (Hist – 19 do Hist – 26) kod koje se također pokazuje da se vjerojatnost oboljenja ne povećava nužno s godinama starosti budući da je u nekim gradovima stopa oboljenja veća kod mlađih ispitanica .



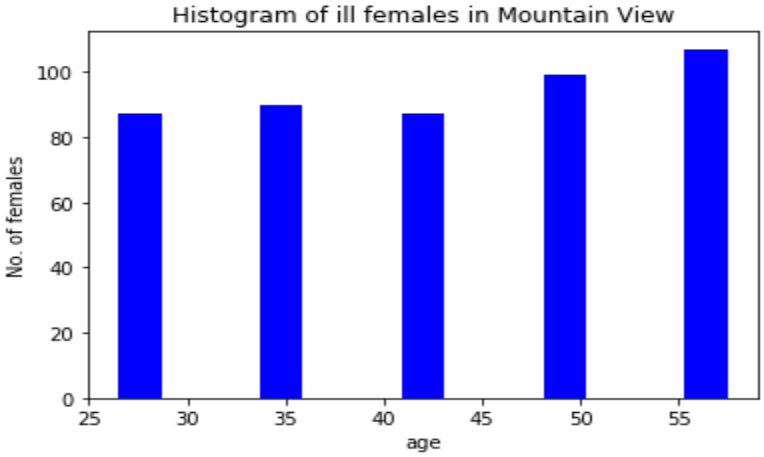
Hist - 19



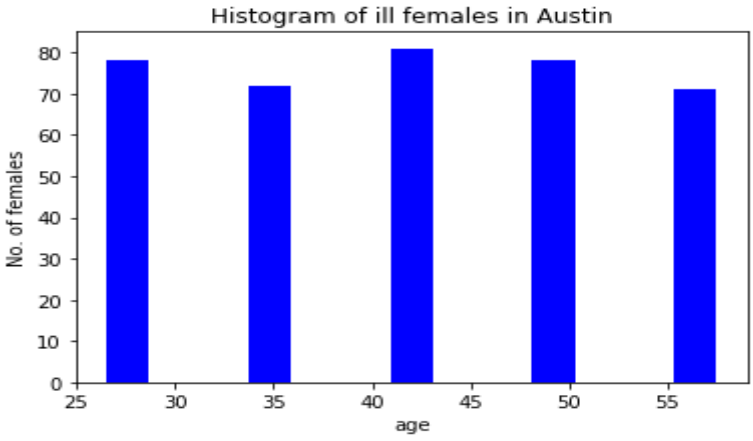
Hist - 20



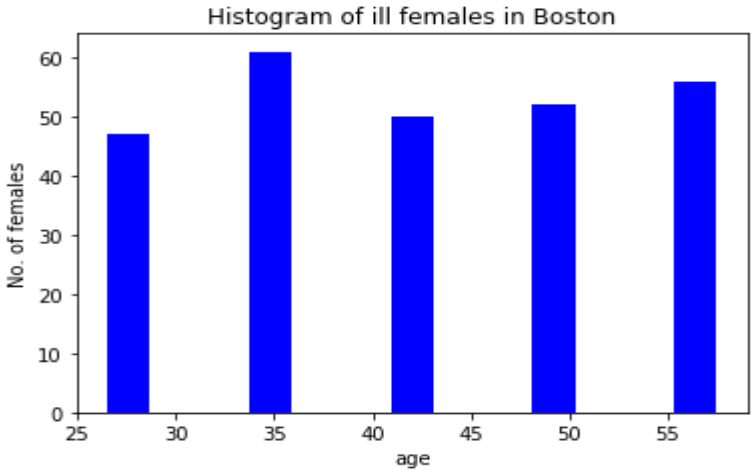
Hist - 21



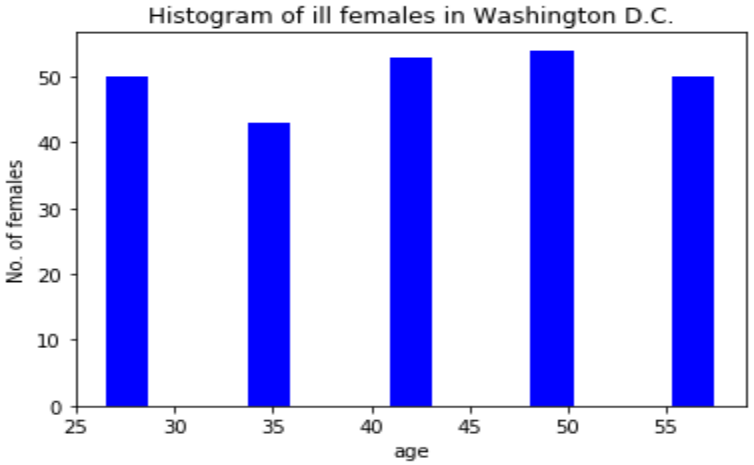
Hist - 22



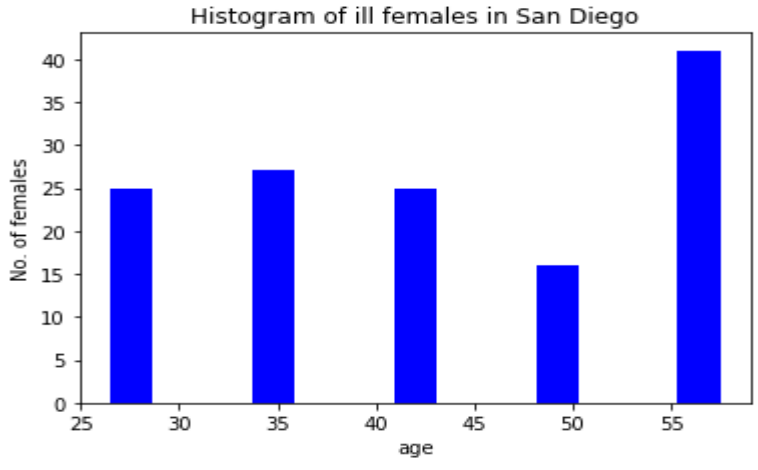
Hist - 23



Hist - 24



Hist - 25

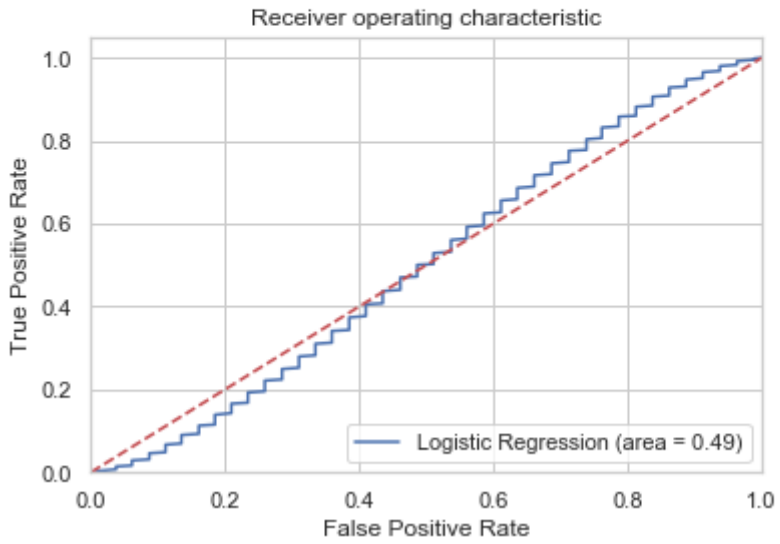


Hist - 26

Dakle, temeljem obzervacija dobivenih iz provedene analize može se zaključiti da populacijsko obolijevanje ne ovisni nužno o godinama starosti, spolu, mjestu življenja i prihodima.

Navedeno pokazuje i model u kojem je za identifikaciju prediktivnih varijabli korišten logistički regresijski model. U prilogu je ROC krivulja iz koje je vidljivo navedeno. Isprekidana linija predstavlja ROC krivulju slučajnog klasifikatora koji je u našem slučaju obolijevanje (1/0) ; Klasifikator je tim bolji što je dalje od te linije (prema gornjem lijevom kutu) što se ovdje ne može reći.

	precision	recall	f1-score	support
0	0.49	0.61	0.55	28789
1	0.49	0.37	0.43	29098
avg / total	0.49	0.49	0.49	57887



utu).