Fáza 1

Michaela Gubovská a Jakub Hajdu

V tejto fáze sa jednotlivo pozrieme na dáta z dvoch súborov - profiles.csv a labor.csv. Pre potreby našej analýzy sme dáta z týchto súborov nespájali do jednej veľkej tabuľky, avšak v ďalších fázach tak s najväčšou pravdepodobnosťou spravíme a spojíme tabuľky na základe stĺpca "ssn", ktorý sa v oboch nachádza.

In [1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm
import scipy.stats as stats
import statsmodels.stats.api as sms
```

Základný opis dát a ich charakteristika - dáta zo súboru profiles.csv

Najskôr si načítame dáta z "profiles.csv" do data frame.

In [2]:
```
filename_p = "data/profiles.csv"
dfp = pd.read_csv(filename_p, sep='\t')
## separátorom dát je znak tabulátora
```

Počet záznamov:

In [3]:
```
len(dfp)
```

Out[3]:
```
3082
```

Počet atribútov:

In [4]:
```
len(dfp.columns.values)
```

Out[4]:
```
10
```

Jednotlivé atribúty a ich typy:

In [5]:
```
dfp.dtypes
```

Out[5]:
```
Unnamed: 0      int64
job            object
address        object
blood_group    object
ssn            object
birthdate      object
residence      object
race           object
name           object
sex            object
dtype: object
```

Základné štatistiky stĺpcov "blood_group", "race" a "sex":

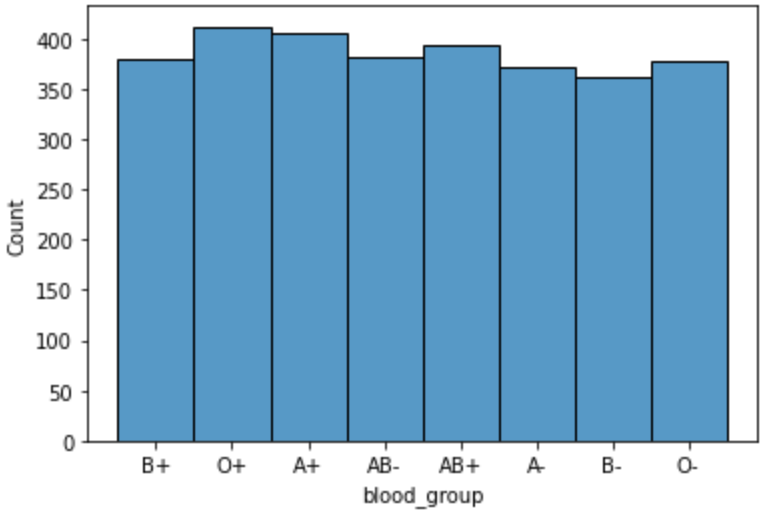
In [6]:
```
dfp[['blood_group', 'race', 'sex']].describe()
```

Out[6]:

	blood_group	race	sex
count	3082	3082	3082
unique	8	8	2
top	O+	White	F
freq	412	1569	1563

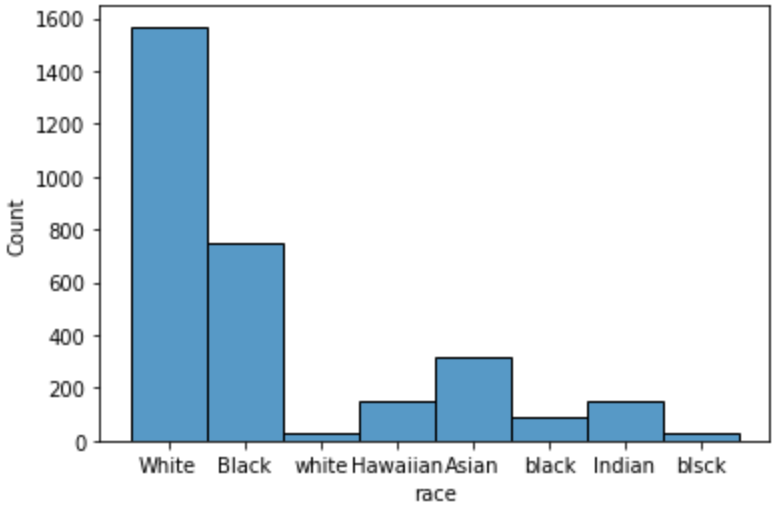
In [7]:
```
sns.histplot(data=dfp['blood_group'])
```

Out[7]:
```
<AxesSubplot:xlabel='blood_group', ylabel='Count'>
```



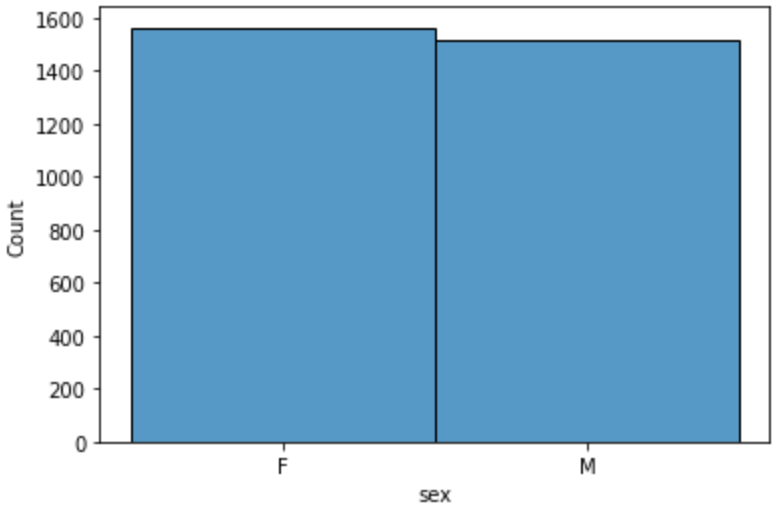
In [8]:
```
sns.histplot(data=dfp['race'])
```

Out[8]:
```
<AxesSubplot:xlabel='race', ylabel='Count'>
```



In [9]:
```
sns.histplot(data=dfp['sex'])
```

Out[9]:
```
<AxesSubplot:xlabel='sex', ylabel='Count'>
```



Overíme, či sa v dátach nevyskytujú chýbajúce hodnoty:

In [10]:
```
dfp.isnull().sum()
```

Out[10]:
```
Unnamed: 0      0
job             0
address         0
blood_group     0
ssn             0
birthdate       0
residence       0
race            0
name            0
sex             0
dtype: int64
```

Vypíšeme si začiatok dát. Už z výpisu atribútov bolo vidieť zvláštny stĺpec "Unnamed: 0".

In [11]:
```
dfp.head()
```

Out[11]:

	Unnamed: 0	job	address	blood_group	ssn	birthdate	residence	race	
0	0	Quarry manager	6269 Kelsey Cove\r\nJessicahaven, KY 76299	B+	818-49-6074	1925-12-31	30070 Anderson Branch Suite 371\r\nHoffmanberg...	White	K
1	1	Financial risk analyst	72241 Luna Divide Suite 877\r\nMelindachester,...	B+	617-52-9894	03/21/1912, 00:00:00	61551 Williams Shoal Apt. 506\r\nLunaborough, ...	Black	G
2	2	Waste management officer	6554 Nicole Lodge\r\nNorth Tanner, PA 59584	B+	690-88-6942	1960-11-19	77305 Palmer Valleys Suite 424\r\nJennifertown...	Black	
3	3	Publishing rights manager	847 Taylor Court Apt. 547\r\nGutierrezfort, SD...	O+	767-13-0171	1978/11/14	62021 Schwartz Roads\r\nJaimeville, SD 64413	Black	A
4	4	Medical sales representative	83815 Richard Causeway Suite 275\r\nLemouth, W...	B+	412-57-1910	03/14/1995, 00:00:00	Unit 4872 Box 1158\r\nDPO AA 51410	white	

Identifikácia problémov dát zo súboru profiles.csv s

navrhnutým riešením

Ako prvý problém vidíme nepotrebný index stĺpec s názvom "Unnamed: 0", ktorý môžme odstrániť.

In [12]:
```
dfp.drop('Unnamed: 0', axis=1, inplace=True)
dfp.head()
```

Out[12]:

	job	address	blood_group	ssn	birthdate	residence	race	name	sex
0	Quarry manager	6269 Kelsey Cove\r\nJessicahaven, KY 76299	B+	818-49-6074	1925-12-31	30070 Anderson Branch Suite 371\r\nHoffmanberg...	White	Kimberly Frazier	F
1	Financial risk analyst	72241 Luna Divide Suite 877\r\nMelindachester,...	B+	617-52-9894	03/21/1912, 00:00:00	61551 Williams Shoal Apt. 506\r\nLunaborough, ...	Black	Timothy Gutierrez MD	M
2	Waste management officer	6554 Nicole Lodge\r\nNorth Tanner, PA 59584	B+	690-88-6942	1960-11-19	77305 Palmer Valleys Suite 424\r\nJennifertown...	Black	Elaine Elliott	F
3	Publishing rights manager	847 Taylor Court Apt. 547\r\nGutierrezfort, SD...	O+	767-13-0171	1978/11/14	62021 Schwartz Roads\r\nJaimeville, SD 64413	Black	Richard Anderson	M
4	Medical sales representative	83815 Richard Causeway Suite 275\r\nLemouth, W...	B+	412-57-1910	03/14/1995, 00:00:00	Unit 4872 Box 1158\r\nDPO AA 51410	white	Jennifer Bass	F

Ďalší problém dát je nejednotný formát hodnôt v stĺpci "race":

In [13]:
```
dfp.race.unique()
```

Out[13]:
```
array(['White', 'Black', 'white', 'Hawaiian', 'Asian', 'black', 'Indian',
       'blsck'], dtype=object)
```

Tieto hodnoty zjednotíme, zachováme však rasu "Hawaiian" nakoľko sme sa dočítali o skutočnosti, že sa Havajčania považujú za samostatnú rasu. Po úprave sú hodnoty v konzistentnom formáte.

In [14]:
```
dfp['race'] = dfp['race'].str.replace('white','White')
dfp['race'] = dfp['race'].str.replace('black','Black')
dfp['race'] = dfp['race'].str.replace('blsck','Black')
dfp.race.unique()
```

Out[14]:
```
array(['White', 'Black', 'Hawaiian', 'Asian', 'Indian'], dtype=object)
```

Ako ďalší problém si môžeme všimnúť rôzne formáty dátumov v stĺpci "birthdate". Tieto dátumy preto upravíme na jednotný formát yy-mm-dd.

In [15]:
```
dfp['birthdate'] = pd.to_datetime(dfp.birthdate)

dfp.head()
```

Out[15]:

	job	address	blood_group	ssn	birthdate	residence	race	name	sex
0	Quarry manager	6269 Kelsey Cove\r\nJessicahaven, KY 76299	B+	818-49-6074	1925-12-31	30070 Anderson Branch Suite 371\r\nHoffmanberg...	White	Kimberly Frazier	F

	job	address	blood_group	ssn	birthdate	residence	race	name	sex
1	Financial risk analyst	72241 Luna Divide Suite 877\r\nMelindachester,...	B+	617-52-9894	1912-03-21	61551 Williams Shoal Apt. 506\r\nLunaborough, ...	Black	Timothy Gutierrez MD	M
2	Waste management officer	6554 Nicole Lodge\r\nNorth Tanner, PA 59584	B+	690-88-6942	1960-11-19	77305 Palmer Valleys Suite 424\r\nJennifertown...	Black	Elaine Elliott	F
3	Publishing rights manager	847 Taylor Court Apt. 547\r\nGutierrezfort, SD...	O+	767-13-0171	1978-11-14	62021 Schwartz Roads\r\nJaimeville, SD 64413	Black	Richard Anderson	M
4	Medical sales representative	83815 Richard Causeway Suite 275\r\nLemouth, W...	B+	412-57-1910	1995-03-14	Unit 4872 Box 1158\r\nDPO AA 51410	White	Jennifer Bass	F

Posledný nami identifikovaný problém sú nezmyselné dáta o osobách mladších ako 15 rokov pričom majú vyplnené zamestnanie.

In [16]:
```
sum_under_14 = dfp[dfp['birthdate'].dt.year > 2007]
sum_under_14
```

Out[16]:

	job	address	blood_group	ssn	birthdate	residence	race	name
32	Therapist, sports	05324 Jordan Grove Apt. 238\r\nNew Alexandria,...	O-	220-42-8328	2018-03-19	402 Alec Estates\r\nMarshchester, MO 55551	Hawaiian	Vanessa Serrano
37	Astronomer	8082 Davis Prairie\r\nWest Scottfort, IL 36181	A-	194-87-4332	2010-06-07	9928 Russell Mission\r\nPort Chadville, AK 10999	Black	Beth Hale
64	Social research officer, government	4399 Pamela Spurs\r\nWest Johnchester, SD 78725	AB-	605-22-5826	2013-10-29	3429 Fox Forges\r\nSuebury, TX 65082	White	Norma Evans
65	Occupational psychologist	95106 Janice Fall\r\nDavidsonburgh, HI 64306	AB+	262-43-2542	2011-09-23	89482 Jessica Tunnel Apt. 702\r\nNew Marcus, M...	White	Amy Bradshaw
88	Planning and development surveyor	67379 William Ways\r\nLake Victoriaville, MD 2...	O+	650-34-4074	2019-07-17	4517 Leon Groves\r\nAndrewfurt, RI 68424	Indian	Ryan Ross
...	...	...	...	...	...	...	...	...
3055	Engineer, building services	543 Debra Loop\r\nWest Kristenborough, MT 84034	B+	473-75-3051	2013-11-10	287 Jennifer Corner Apt. 632\r\nDarrellbury, C...	Hawaiian	Kimberly Knight
3058	Designer, jewellery	247 Rivera Extension\r\nPort Andreafurt, TN 79770	A+	123-83-9055	2016-09-27	9887 Alexander Union\r\nEast Calvin, TX 84450	White	Francisco Johnson
3066	Radio broadcast assistant	603 Miller Coves\r\nTabithatown, OH 03642	AB+	553-68-4985	2012-07-07	54071 Jasmine Springs Suite 425\r\nHugheston, ...	Black	Alexandria Kennedy

	job	address	blood_group	ssn	birthdate	residence	race	name
3067	Designer, graphic	071 Smith Extensions\r\nWest Emily, NY 01082	O-	095-31-1095	2014-02-16	08293 Vega Unions Suite 021\r\nMirandaside, NC...	Black	Crystal Moran
3072	Public relations account executive	076 Jennifer Vista\r\nNew Samantha, LA 09776	O-	280-01-6370	2020-10-10	536 James Meadow\r\nWest Debra, AK 59069	Black	Robert Giles

357 rows × 9 columns

Počet týchto záznamov je:

In [17]:
```
len(sum_under_14)
```

Out[17]:
```
357
```

In [18]:
```
len(sum_under_14)/len(dfp)*100
```

Out[18]:
```
11.583387410772227
```

Zo všetkých povolaní sú unikátne tieto:

In [19]:
```
dfp.job.unique()
```

Out[19]:
```
array(['Quarry manager', 'Financial risk analyst',
       'Waste management officer', 'Publishing rights manager',
       'Medical sales representative', 'Engineer, water', 'Lobbyist',
       'Theatre stage manager', 'Building control surveyor',
       'Special educational needs teacher',
       'Environmental health practitioner',
       'Development worker, international aid', 'Sports administrator',
       'Operations geologist', 'Arts administrator', 'Ecologist',
       'Human resources officer', 'Environmental manager',
       'Doctor, hospital', 'Water engineer',
       'Exhibitions officer, museum/gallery', 'Haematologist',
       'Psychologist, prison and probation services', 'Music tutor',
       'Horticultural therapist', 'Health and safety inspector',
       'Conference centre manager', 'Psychiatrist', 'Ophthalmologist',
       'Administrator, local government', 'Pharmacologist',
       'Therapist, sports', 'Games developer',
       'Teacher, secondary school', 'Civil Service fast streamer',
       'Scientist, physiological', 'Astronomer', 'Immigration officer',
       'Nurse, mental health', 'Scientist, research (physical sciences)',
       'Building services engineer', 'Engineer, manufacturing systems',
       'Hospital pharmacist',
       'Historic buildings inspector/conservation officer',
       'Fashion designer', 'Cabin crew',
       'Control and instrumentation engineer', 'Dancer',
       'Health physicist', 'Financial manager', 'Product designer',
       'Psychologist, educational', 'Soil scientist',
       'Investment banker, operational', 'Merchandiser, retail',
       'Agricultural consultant', 'Company secretary', 'Barista',
       'Social research officer, government', 'Occupational psychologist',
       'Mining engineer', 'Broadcast engineer',
       'Armed forces technical officer',
       'Designer, blown glass/stained glass', 'Translator',
       'Financial planner', 'Industrial buyer', 'Optician, dispensing',
```

'Midwife', 'Photographer', 'Warehouse manager',
'Buyer, industrial', 'Engineer, electrical', 'Fish farm manager',
'Pilot, airline', 'Set designer', 'Gaffer',
'Therapist, speech and language', 'Geophysical data processor',
'TEFL teacher', 'Risk analyst',
'Planning and development surveyor', 'Sports therapist',
'Therapist, occupational', 'Dealer',
'Teaching laboratory technician', 'Manufacturing engineer',
'Surveyor, minerals', 'Charity fundraiser',
'Commercial art gallery manager', 'Administrator, sports',
'Marine scientist', 'Television camera operator',
'Production designer, theatre/television/film', 'Learning mentor',
'Lecturer, further education', 'Psychotherapist, dance movement',
'Exhibition designer', 'Firefighter', 'Retail merchandiser',
'Podiatrist', 'Physicist, medical', 'Programmer, systems',
'Best boy', 'Public affairs consultant',
'Television floor manager', 'Legal executive',
'Medical illustrator', 'Engineer, structural',
'Surveyor, quantity', 'Applications developer', 'Technical author',
'Chief Technology Officer', 'Biomedical engineer', 'Geoscientist',
'Brewing technologist', 'Designer, exhibition/display',
'Teacher, special educational needs', 'Orthoptist',
'Television production assistant', 'Secretary/administrator',
'Writer', 'Water quality scientist', 'Production manager',
'Science writer', 'Fitness centre manager', 'Sub',
'Systems developer', 'Warden/ranger', 'Trade mark attorney',
"Nurse, children's", 'Surveyor, land/geomatics',
'Rural practice surveyor', 'Computer games developer',
'Engineer, drilling', 'Designer, jewellery',
'Designer, interior/spatial', 'Therapist, music',
'Clinical biochemist', 'Futures trader', 'Cytogeneticist',
'Education officer, museum',
'Armed forces logistics/support/administrative officer',
'Armed forces training and education officer',
'Structural engineer', 'Medical secretary',
'Accounting technician', 'Producer, television/film/video',
'Horticulturist, amenity', 'Accountant, chartered certified',
'Conservation officer, historic buildings',
'Industrial/product designer', 'Herbalist', 'Theme park manager',
'Passenger transport manager', 'Intelligence analyst',
'Programmer, applications', 'Marketing executive',
'Microbiologist', 'Pharmacist, community', 'Tax inspector',
'Volunteer coordinator',
'Administrator, charities/voluntary organisations',
'Geologist, engineering', 'Equality and diversity officer',
'IT technical support officer', 'Trading standards officer',
'Energy manager', 'Designer, industrial/product',
'Chief Marketing Officer', 'Designer, textile',
'Advertising account executive', 'Quantity surveyor',
'Furniture conservator/restorer', 'Advertising copywriter',
'Ergonomist', 'Public librarian', 'Architectural technologist',
'Air cabin crew', 'Sports development officer',
'Designer, television/film set', 'Education administrator',
'Therapist, drama', 'Personal assistant',
'Designer, fashion/clothing', 'Surveyor, building control',
'Housing manager/officer', 'Medical technical officer',
'Armed forces operational officer', 'Psychologist, occupational',
'Editor, film/video', 'English as a second language teacher',
'Medical physicist', 'Lecturer, higher education',
'Electronics engineer', 'Sports coach',
'Research scientist (maths)', 'IT consultant', 'Surgeon',
'Engineer, mining', 'Conservator, museum/gallery',
'Fast food restaurant manager', 'Oncologist',
'Engineer, automotive', 'Engineer, civil (consulting)',
'Garment/textile technologist', 'Forest/woodland manager',
'Race relations officer', 'Hydrogeologist',

'Estate manager/land agent', 'Administrator, Civil Service',
'Clothing/textile technologist', 'Psychotherapist',
'Community arts worker', 'Surveyor, rural practice',
'Maintenance engineer', 'Tax adviser', 'Psychiatric nurse',
'Advertising art director', 'Museum/gallery exhibitions officer',
'Mudlogger', 'Chartered management accountant',
'Engineering geologist', 'Commercial/residential surveyor',
'Actuary', 'Web designer', 'Hotel manager',
'Engineer, maintenance (IT)', 'Scientist, forensic',
'Academic librarian', 'Nutritional therapist',
'Pension scheme manager', 'Designer, graphic', 'Film/video editor',
'Senior tax professional/tax inspector',
'Research scientist (medical)', 'Police officer',
'Financial trader', 'Scientist, audiological',
'Scientist, research (maths)', 'Animal technologist',
'Museum education officer', 'Social worker',
'Education officer, community', 'Health promotion specialist',
'Therapist, horticultural', 'Scientist, research (medical)',
'Location manager', 'Broadcast presenter',
'Further education lecturer', 'Clinical embryologist',
'Programme researcher, broadcasting/film/video', 'Estate agent',
'Social researcher', 'Teacher, music', 'Archivist',
'Records manager', 'Call centre manager', 'Theatre director',
'Teacher, English as a foreign language', 'Management consultant',
'Therapist, art', 'Arts development officer', 'Musician',
'Chiropodist', 'Librarian, public', 'Public relations officer',
'Energy engineer', 'Scientist, product/process development',
'Licensed conveyancer', 'English as a foreign language teacher',
'Product/process development scientist', 'Immunologist',
'Engineer, manufacturing', 'Chief Strategy Officer',
'Sales executive', 'Product manager', 'Surveyor, insurance',
'Geologist, wellsite', 'Wellsite geologist',
'Geophysicist/field seismologist', 'Psychologist, forensic',
'Scientist, water quality', 'Loss adjuster, chartered',
'Medical laboratory scientific officer', 'Special effects artist',
'Lexicographer', 'Journalist, magazine', 'Toxicologist',
'Health service manager', 'Network engineer', 'Media planner',
'Financial controller', 'Banker', 'Counselling psychologist',
'Tour manager', 'Geographical information systems officer',
'Radiographer, therapeutic', 'Youth worker', 'Landscape architect',
'Logistics and distribution manager', 'Chiropractor',
'Leisure centre manager', 'Petroleum engineer',
'Glass blower/designer', 'Press photographer',
'Chief Executive Officer', 'Clinical cytogeneticist',
'Clinical scientist, histocompatibility and immunogenetics',
'Insurance broker', 'Dispensing optician',
'Horticulturist, commercial', 'Merchant navy officer',
'Risk manager', 'Patent attorney', 'Information systems manager',
'Engineer, agricultural', 'Civil engineer, contracting',
'Scientist, marine', 'Training and development officer',
'Architect', 'Health and safety adviser',
'Operational investment banker',
'Scientific laboratory technician', 'Therapeutic radiographer',
'Museum/gallery conservator', 'Technical brewer',
'Multimedia specialist', 'Designer, furniture', 'Chief of Staff',
'Engineer, electronics', 'International aid/development worker',
'Retail banker', 'Neurosurgeon', 'Journalist, newspaper',
'Research officer, government', 'Restaurant manager, fast food',
'Plant breeder/geneticist', 'Customer service manager',
'Private music teacher', 'Pathologist', 'Exercise physiologist',
'Forensic scientist', 'Illustrator', 'Biomedical scientist',
'Animal nutritionist', 'Metallurgist', 'Engineer, production',
'Biochemist, clinical', 'Secondary school teacher',
'Chief Operating Officer', 'Contracting civil engineer',
'Acupuncturist', 'Artist', 'Solicitor', 'Press sub',
'Radiation protection practitioner',

'Chartered legal executive (England and Wales)',
'Multimedia programmer', 'Furniture designer',
'Lighting technician, broadcasting/film/video',
'Minerals surveyor', 'Electrical engineer',
'Audiological scientist', 'Commercial horticulturist',
'Charity officer', 'Diagnostic radiographer',
"Politician's assistant", 'Secretary, company',
'Technical sales engineer', 'Librarian, academic',
'Clinical research associate', 'Textile designer', 'Ship broker',
'Administrator, education',
'Diplomatic Services operational officer',
'Research scientist (life sciences)',
'Trade union research officer',
'Government social research officer', 'Clinical psychologist',
'Manufacturing systems engineer', 'Materials engineer',
'Clinical molecular geneticist', 'Advice worker',
'Psychologist, sport and exercise', 'Psychologist, counselling',
'Phytotherapist', 'Transport planner',
'Dance movement psychotherapist',
'Production assistant, television', 'Oceanographer',
'Interior and spatial designer', 'Recycling officer',
'Occupational hygienist', 'Radiographer, diagnostic', 'Contractor',
'Psychotherapist, child', 'Aid worker', 'Copywriter, advertising',
'Surveyor, hydrographic', 'Administrator',
'Educational psychologist', 'Personnel officer',
'Editor, commissioning', 'Research scientist (physical sciences)',
'Statistician', 'Magazine journalist', 'Chemist, analytical',
'Claims inspector/assessor', 'Meteorologist', 'Bonds trader',
'Lawyer', 'Engineer, technical sales', 'Consulting civil engineer',
'Recruitment consultant', 'Art gallery manager', 'Homeopath',
'Mechanical engineer', 'Civil engineer, consulting', 'Printmaker',
'Pensions consultant', 'Environmental consultant',
'Surveyor, mining', 'Financial adviser', 'Fisheries officer',
'Newspaper journalist', 'Dentist', 'Publishing copy',
'Adult nurse', 'Engineer, control and instrumentation',
'Interpreter', 'Automotive engineer', 'Land/geomatics surveyor',
'Stage manager', 'Amenity horticulturist', 'Facilities manager',
'Communications engineer', 'Camera operator', 'Farm manager',
'Professor Emeritus', 'Engineer, building services', 'Paramedic',
'Environmental education officer', 'Animator',
'IT sales professional', 'Teacher, primary school',
'Conservation officer, nature',
'Scientist, research (life sciences)', 'Arboriculturist',
'Accountant, chartered', 'Emergency planning/management officer',
'Ambulance person', 'Food technologist',
'Runner, broadcasting/film/video', 'Economist',
'Seismic interpreter', 'Holiday representative', 'Media buyer',
'Presenter, broadcasting', 'Journalist, broadcasting', 'Actor',
'Dietitian', 'Field trials officer', 'Town planner',
'Physiological scientist', 'Psychologist, clinical',
'Probation officer', 'Land', 'Art therapist',
'Telecommunications researcher', 'Office manager', 'Proofreader',
'Make', 'Editorial assistant', 'Sales promotion account executive',
'Radio broadcast assistant', 'Video editor', 'Careers adviser',
'Surveyor, building', 'Insurance underwriter',
'Therapist, nutritional', 'Commissioning editor', 'IT trainer',
'Agricultural engineer', 'Accountant, chartered public finance',
'Chartered public finance accountant', 'Retail buyer',
'Editor, magazine features', 'Embryologist, clinical',
'Engineer, aeronautical', 'Associate Professor',
'Print production planner', 'Museum/gallery curator',
'Education officer, environmental', 'Careers information officer',
'Engineer, chemical', 'Tourism officer', 'Adult guidance worker',
'Pharmacist, hospital', 'Tourist information centre manager',
'Investment analyst', 'Data scientist', 'Curator',
'Theatre manager', 'Site engineer', 'Dramatherapist',

```
'Corporate investment banker', 'Chief Financial Officer',
'Patent examiner', 'Investment banker, corporate',
'Scientist, biomedical', 'Database administrator', 'Archaeologist',
'Aeronautical engineer', 'Chartered loss adjuster',
'Legal secretary', 'Physiotherapist',
'Community development worker', 'Child psychotherapist',
'Fine artist', 'Programmer, multimedia', 'Comptroller',
'Music therapist', 'Solicitor, Scotland', 'Hospital doctor',
'Osteopath', 'Optometrist', 'Jewellery designer',
'Mental health nurse', 'Counsellor', 'Data processing manager',
'Quality manager', 'Equities trader', 'Visual merchandiser',
'Engineer, civil (contracting)', 'Insurance account manager',
'Drilling engineer', 'Early years teacher',
'Speech and language therapist',
'Accountant, chartered management', 'Market researcher',
'Naval architect', 'Local government officer', 'Hydrologist',
'Public house manager', 'Building surveyor',
'Research officer, political party', 'Production engineer',
'Doctor, general practice', 'Occupational therapist',
'Higher education careers adviser', 'Systems analyst',
'Community education officer',
'Outdoor activities/education manager', 'Horticultural consultant',
'Higher education lecturer', 'Copy', 'Heritage manager',
'Scientist, clinical (histocompatibility and immunogenetics)',
'Public relations account executive', 'Engineer, site',
'Purchasing manager', 'Production assistant, radio',
'Prison officer', 'Nature conservation officer', 'Tree surgeon',
'Airline pilot', 'Accommodation manager', 'Health visitor',
'Development worker, community', 'Analytical chemist',
'Advertising account planner', 'Conservator, furniture',
'Teacher, early years/pre', 'Corporate treasurer',
'Broadcast journalist', 'Forensic psychologist',
'Engineer, broadcasting (operations)', 'Teacher, adult education',
'Civil Service administrator', 'Freight forwarder', 'Geochemist',
'Sound technician, broadcasting/film/video',
'Magazine features editor', 'Engineer, communications',
'Nurse, adult', 'Cartographer', 'Surveyor, commercial/residential',
'Community pharmacist', 'Regulatory affairs officer',
'Restaurant manager', 'Engineer, land', 'Bookseller',
'Colour technologist', 'Retail manager', 'Graphic designer',
'Surveyor, planning and development', 'Geneticist, molecular',
'Engineer, maintenance', 'Chartered certified accountant',
'Herpetologist', 'Insurance claims handler', 'Producer, radio',
'Designer, ceramics/pottery', 'Engineer, materials',
'Sales professional, IT', 'Nurse, learning disability',
'Veterinary surgeon', 'Ranger/warden', 'Primary school teacher',
'Software engineer', 'Learning disability nurse',
'Air traffic controller', 'Operational researcher', 'Barrister',
'Paediatric nurse', 'General practice doctor',
'Administrator, arts', 'Chartered accountant', "Barrister's clerk",
'Information officer', 'Designer, multimedia', 'Chemical engineer',
'Sport and exercise psychologist', 'Insurance risk surveyor',
'Hydrographic surveyor', 'Engineer, petroleum',
'Research officer, trade union', 'Engineer, biomedical',
'Radio producer', 'Travel agency manager', 'Field seismologist',
'Engineer, energy', 'Television/film/video producer',
'Event organiser'], dtype=object)
```

Skontrolovali sme teda, či sa v stĺpci "job" nachádzajú iba legitímne pracovné pozície. Zistili sme, že áno a preto môžeme prehlásiť, že dáta o ľuďoch mladších ako 14 rokov sú nezmyselné, keďže v USA človek v tomto veku nemôže byť zamestnaný. Týchto nezmyselných dát sa v tabuľke nachádza 357, čo tvorí zaokrúhlene až 11.6% všetkých záznamov. Ako riešenie tohto problému sme navrhli a implementovali náhradu pracovných pozícií ľudí mladších ako 14 rokov statusom "unemployed", keďže by nebolo správne tieto dáta odstrániť nakoľko je ich počet vyšší ako 10%.

In [20]:
```
dfp.loc[dfp['birthdate'].dt.year > 2007, 'job'] = 'unemployed'
dfp[dfp['birthdate'].dt.year > 2007]
```

Out[20]:

	job	address	blood_group	ssn	birthdate	residence	race	name
32	unemployed	05324 Jordan Grove Apt. 238\r\nNew Alexandria,...	O-	220-42-8328	2018-03-19	402 Alec Estates\r\nMarshchester, MO 55551	Hawaiian	Vanessa Serrano
37	unemployed	8082 Davis Prairie\r\nWest Scottfort, IL 36181	A-	194-87-4332	2010-06-07	9928 Russell Mission\r\nPort Chadville, AK 10999	Black	Beth Hale
64	unemployed	4399 Pamela Spurs\r\nWest Johnchester, SD 78725	AB-	605-22-5826	2013-10-29	3429 Fox Forges\r\nSuebury, TX 65082	White	Norma Evans
65	unemployed	95106 Janice Fall\r\nDavidsonburgh, HI 64306	AB+	262-43-2542	2011-09-23	89482 Jessica Tunnel Apt. 702\r\nNew Marcus, M...	White	Amy Bradshaw
88	unemployed	67379 William Ways\r\nLake Victoriaville, MD 2...	O+	650-34-4074	2019-07-17	4517 Leon Groves\r\nAndrewfurt, RI 68424	Indian	Ryan Ross
...	...	...	...	...	...	...	...	...
3055	unemployed	543 Debra Loop\r\nWest Kristenborough, MT 84034	B+	473-75-3051	2013-11-10	287 Jennifer Corner Apt. 632\r\nDarrellbury, C...	Hawaiian	Kimberly Knight
3058	unemployed	247 Rivera Extension\r\nPort Andreafurt, TN 79770	A+	123-83-9055	2016-09-27	9887 Alexander Union\r\nEast Calvin, TX 84450	White	Francisco Johnson
3066	unemployed	603 Miller Coves\r\nTabithatown, OH 03642	AB+	553-68-4985	2012-07-07	54071 Jasmine Springs Suite 425\r\nHugheston, ...	Black	Alexandria Kennedy
3067	unemployed	071 Smith Extensions\r\nWest Emily, NY 01082	O-	095-31-1095	2014-02-16	08293 Vega Unions Suite 021\r\nMirandaside, NC...	Black	Crystal Moran
3072	unemployed	076 Jennifer Vista\r\nNew Samantha, LA 09776	O-	280-01-6370	2020-10-10	536 James Meadow\r\nWest Debra, AK 59069	Black	Robert Giles

357 rows × 9 columns

Základný opis dát a ich charakteristika - dáta zo súboru labor.csv

In [21]:
```
filename_l = "data/labor.csv"
dfl = pd.read_csv(filename_l, sep='\t')
dfl.head()
```

Out[21]:

	Unnamed: 0	leukocyty	ssn	name	smoker	hemoglobin	trombocyty	indicator	alt	relationship	wei

	Unnamed: 0	leukocyty	ssn	name	smoker	hemoglobin	trombocyty	indicator	alt	relationship	wei
0	0	5.90289	513-95-7625	Andrew Jacobs	no	7.54279	5.83096	1.0	17.60670	widowed	8.09
1	1	5.56403	025-71-2115	Ian Harrison	Y	7.87747	NaN	1.0	17.78037	single	73.58
2	2	6.24057	824-63-0108	Matthew Williams	no	4.72650	7.83234	1.0	25.25152	single	129.45
3	3	5.48374	157-32-2908	Charles Chavez	no	5.43079	5.36911	1.0	18.32802	divoced	18.61
4	4	6.04784	545-96-1267	Allen Chung MD	yes	8.85943	6.76682	1.0	11.03841	divoced	78.83

Počet záznamov:

In [22]:
```
len(dfl)
```

Out[22]:
```
10017
```

Počet atribútov:

In [23]:
```
len(dfl.columns.values)
```

Out[23]:
```
18
```

Jednotlivé atribúty a ich typy:

In [24]:
```
dfl.dtypes
```

Out[24]:
```
Unnamed: 0         int64
leukocyty        float64
ssn               object
name              object
smoker            object
hemoglobin       float64
trombocyty       float64
indicator        float64
alt              float64
relationship      object
weight           float64
ast              float64
alp              float64
hematokrit       float64
hbver            float64
etytr            float64
er-cv            float64
erytrocyty       float64
dtype: object
```

Základná charakteristika číselných dát v tabuľke:

In [25]:
```
dfl[['weight', 'leukocyty', 'hemoglobin', 'trombocyty', 'alt', 'ast', 'alp', 'hematokrit',
```

```
'er-cv', 'erytrocyty']].describe()
```

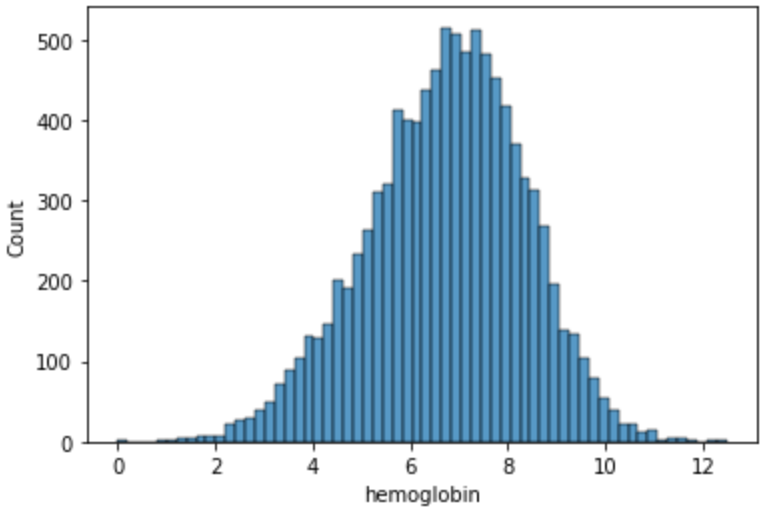
Out[25]:

	weight	leukocyty	hemoglobin	trombocyty	alt	ast	alp	hematokrit	
count	10017.000000	9986.000000	9985.000000	9987.000000	9987.000000	9987.000000	9987.000000	9987.000000	99
mean	70.699780	5.963060	6.735597	5.994497	15.191684	47.562467	57.371055	6.379837	
std	35.166726	0.997483	1.620583	0.993750	5.190542	13.140353	25.139268	1.593761	
min	-50.180020	2.039120	0.000000	2.114710	0.000000	0.000000	0.000000	1.806100	
25%	46.826400	5.306335	5.695290	5.319635	12.107925	38.774135	35.528960	5.162530	
50%	70.883910	5.963415	6.836120	5.999860	14.830500	47.549220	63.298140	6.447480	
75%	94.517900	6.633057	7.864160	6.657495	17.722370	56.475380	80.094585	7.563310	
max	204.381010	9.544190	12.483130	9.560170	100.000000	100.000000	100.000000	11.428950	

Rozdelenie hemoglobínu:

In [26]:
```
sns.histplot(data=dfl.hemoglobin)
```

Out[26]:
```
<AxesSubplot:xlabel='hemoglobin', ylabel='Count'>
```



Zistíme si základné vlastnosti atribútu hemoglobín, ako min hodnota, max hodnota, či priemerná hodnota:

In [27]:
```
dfl.hemoglobin.describe()
```

Out[27]:
```
count    9985.000000
mean        6.735597
std         1.620583
min         0.000000
25%         5.695290
50%         6.836120
75%         7.864160
max        12.483130
Name: hemoglobin, dtype: float64
```

Pozrieme sa na koeficient asymetrie rozdelenia hodnôt hemoglobínu:

In [28]:
```
dfl.hemoglobin.skew()
```

Out[28]:
```
-0.25039670757432236
```

Šikmosť je menšia ako 0 ale zároveň väčšia ako -0.5, čo znamená že rozdelenie hemoglobínu je celkom symetrické avšak trochu ťažšie na pravej strane (negatívna asymetria).

Pozrieme sa na koeficient špičatosti rozdelenia hodnôt hemoglobínu:

In [29]:
```
dfl.hemoglobin.kurtosis()
```

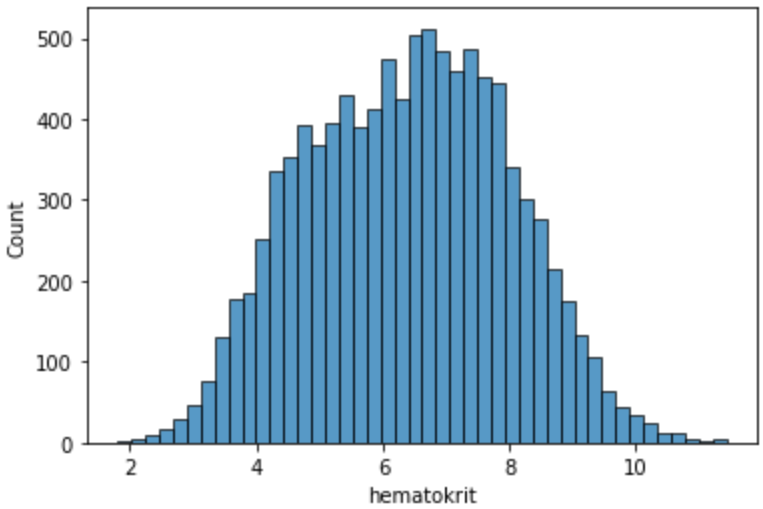
Out[29]:
```
-0.026888343109358726
```

Z hodnoty špičatosti vieme, že sa pre atribút hemoglobín nachádza menej hodnôt na okrajoch a rozdelenie je ploché.

Rozdelenie hematokritu:

In [30]:
```
sns.histplot(data=dfl.hematokrit)
```

Out[30]:
```
<AxesSubplot:xlabel='hematokrit', ylabel='Count'>
```



Zistíme si základné vlastnosti atribútu hematokrit, ako min hodnota, max hodnota, či priemerná hodnota:

In [31]:
```
dfl.hematokrit.describe()
```

Out[31]:
```
count    9987.000000
mean        6.379837
std         1.593761
min         1.806100
25%         5.162530
50%         6.447480
75%         7.563310
max        11.428950
Name: hematokrit, dtype: float64
```

Pozrieme sa na koeficient asymetrie rozdelenia hodnôt hematokritu:

In [32]:
```
dfl.hemoglobin.skew()
```

Out[32]:
```
-0.25039670757432236
```

Šikmosť je menšia ako 0 ale zároveň väčšia ako -0.5, čo znamená že rozdelenie hematokritu je celkom symetrické avšak trochu ťažšie na pravej strane (negatívna asymetria).

Pozrieme sa na koeficient špičatosti rozdelenia hodnôt hemoglobínu:

In [33]:
```
dfl.hemoglobin.kurtosis()
```

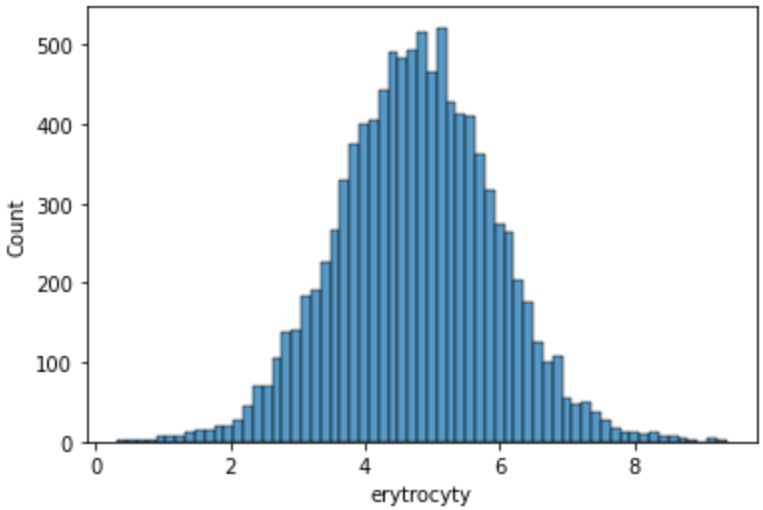
Out[33]:
```
-0.026888343109358726
```

Z hodnoty špičatosti vieme, že sa pre atribút hematokrit nachádza menej hodnôt na okrajoch a rozdelenie je ploché.

Rozdelenie erytrocytov:

In [34]:
```
sns.histplot(data=dfl.erytrocyty)
```

Out[34]:
```
<AxesSubplot:xlabel='erytrocyty', ylabel='Count'>
```



Zistíme si základné vlastnosti atribútu erytrocyty, ako min hodnota, max hodnota, či priemerná hodnota:

In [35]:
```
dfl.erytrocyty.describe()
```

Out[35]:
```
count    9987.000000
mean        4.769255
std         1.167590
min         0.328770
25%         3.995755
50%         4.773670
75%         5.545965
max         9.363550
Name: erytrocyty, dtype: float64
```

Pozrieme sa na koeficient asymetrie rozdelenia hodnôt erytrocytov:

In [36]:
```
dfl.erytrocyty.skew()
```

Out[36]:
```
0.015462758377710053
```

Šikmosť je väčšia ako 0 ale zároveň menšia ako 0.5, čo znamená že rozdelenie hematokritu je celkom symetrické avšak trochu ťažšie na ľavej strane (negatívna asymetria).

Pozrieme sa na koeficient špičatosti rozdelenia hodnôt erytrocytov:

In [37]:
```
dfl.erytrocyty.kurtosis()
```

Out[37]:
```
0.2853282235461396
```

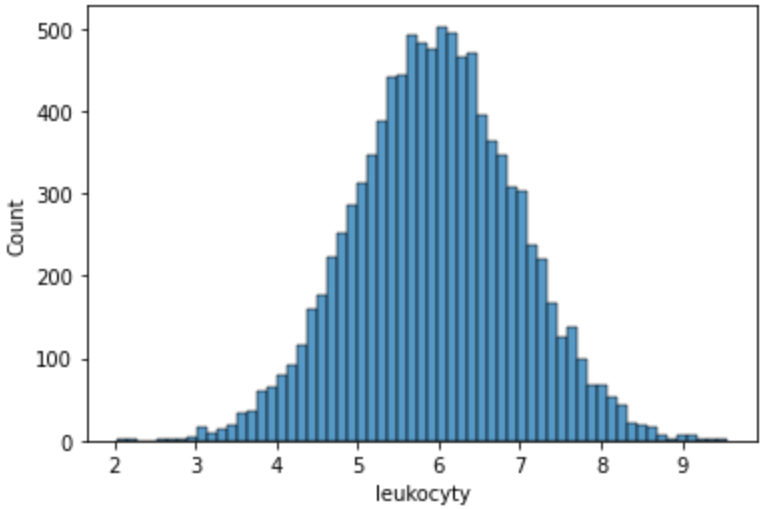
Z hodnoty špičatosti vieme, že sa pre atribút erytrocyty nachádza viac hodnôt na okrajoch a rozdelenie je špičaté

(nakoľko je hodnota kurtosis pozitívna).

Rozdelenie leukocytov:

In [38]:
```
sns.histplot(data=dfl.leukocyty)
```

Out[38]:
```
<AxesSubplot:xlabel='leukocyty', ylabel='Count'>
```



Zistíme si základné vlastnosti atribútu leukocyty, ako min hodnota, max hodnota, či priemerná hodnota:

In [39]:
```
dfl.leukocyty.describe()
```

Out[39]:
```
count    9986.000000
mean        5.963060
std         0.997483
min         2.039120
25%         5.306335
50%         5.963415
75%         6.633057
max         9.544190
Name: leukocyty, dtype: float64
```

Pozrieme sa na koeficient asymetrie rozdelenia hodnôt leukocytov:

In [40]:
```
dfl.leukocyty.skew()
```

Out[40]:
```
-0.010269243523144444
```

Šikmosť je menšia ako 0 ale zároveň väčšia ako -0.5, čo znamená že rozdelenie leukocytov je celkom symetrické avšak trochu ťažšie na pravej strane (negatívna asymetria).

Pozrieme sa na koeficient špičatosti rozdelenia hodnôt leukocytov:

In [41]:
```
dfl.leukocyty.kurtosis()
```

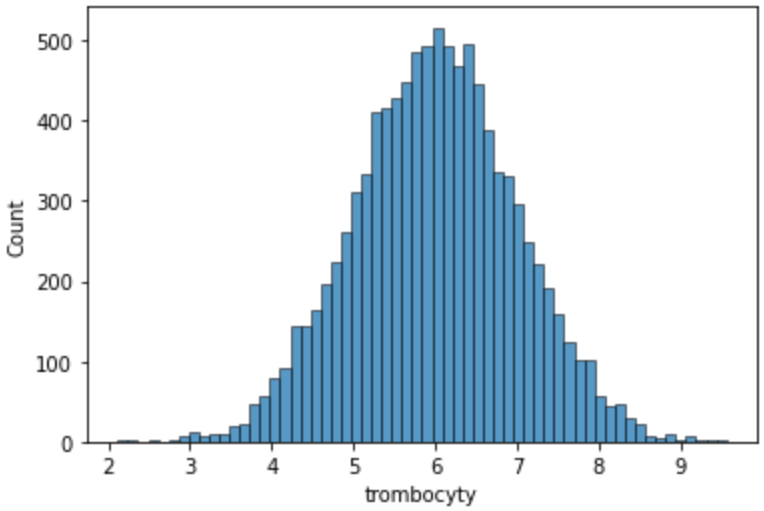
Out[41]:
```
-0.01805868137407929
```

Z hodnoty špičatosti vieme, že sa pre atribút leukocyty nachádza menej hodnôt na okrajocha rozdelenie je ploché.

Rozdelenie trombocytov:

In [42]:
```
sns.histplot(data=dfl.trombocyty)
```

Out[42]:
```
<AxesSubplot:xlabel='trombocyty', ylabel='Count'>
```



Zistíme si základné vlastnosti atribútu trombocyty, ako min hodnota, max hodnota, či priemerná hodnota:

In [43]:
```
dfl.trombocyty.describe()
```

Out[43]:
```
count    9987.000000
mean        5.994497
std         0.993750
min         2.114710
25%         5.319635
50%         5.999860
75%         6.657495
max         9.560170
Name: trombocyty, dtype: float64
```

Pozrieme sa na koeficient asymetrie rozdelenia hodnôt trombocytov:

In [44]:
```
dfl.trombocyty.skew()
```

Out[44]:
```
-0.0086023430585214
```

Šikmosť je menšia ako 0 ale zároveň väčšia ako -0.5, čo znamená že rozdelenie trombocytov je celkom symetrické avšak trochu ťažšie na pravej strane (negatívna asymetria).

Pozrieme sa na koeficient špičatosti rozdelenia hodnôt trombocytov:

In [45]:
```
dfl.trombocyty.kurtosis()
```

Out[45]:
```
-0.041982113756123596
```

Z hodnoty špičatosti vieme, že sa pre atribút trombocyty nachádza menej hodnôt na okrajoch a rozdelenie je ploché.

Identifikácia problémov dát zo súboru labor.csv s navrhnutým riešením

Ako prvý problém znova vidíme nepotrebný index stĺpec s názvom "Unnamed: 0", ktorý môžeme odstrániť.

In [46]:
```
dfl.drop('Unnamed: 0', axis=1, inplace=True)
dfl.head()
```

Out[46]:

	leukocyty	ssn	name	smoker	hemoglobin	trombocyty	indicator	alt	relationship	weight	a
0	5.90289	513-95-7625	Andrew Jacobs	no	7.54279	5.83096	1.0	17.60670	widowed	8.09544	42.9287
1	5.56403	025-71-2115	Ian Harrison	Y	7.87747	NaN	1.0	17.78037	single	73.58725	50.2550
2	6.24057	824-63-0108	Matthew Williams	no	4.72650	7.83234	1.0	25.25152	single	129.45079	25.1594
3	5.48374	157-32-2908	Charles Chavez	no	5.43079	5.36911	1.0	18.32802	divoced	18.61698	45.0209
4	6.04784	545-96-1267	Allen Chung MD	yes	8.85943	6.76682	1.0	11.03841	divoced	78.83355	59.0639

V dátach sa nachádzajú duplicitné záznamy.

In [47]:
```
len(dfl) - len(dfl.drop_duplicates()) ## 99 duplicitnych zaznamov
```

Out[47]:
```
99
```

In [48]:
```
dfl = dfl.drop_duplicates() ## odstranene duplicitne zaznamy
len(dfl)
```

Out[48]:
```
9918
```

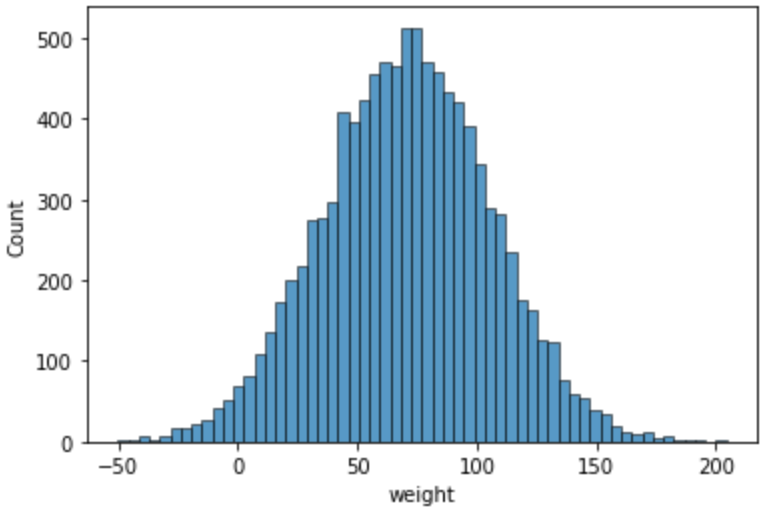
Identifikovali sme problém s dátami vo forme zlých hodnôt (prípadne rozdelenia hodnôt) hmotnosti. Možným riešením by bolo odstránenie záznamov s hmotnosťou <= 0, keďže takýchto záznamov je z celkového počtu približne 2,3%.

In [49]:
```
len(dfl[dfl.weight <= 0]) / len(dfl) * 100
```

Out[49]:
```
2.3089332526719097
```

In [50]:
```
sns.histplot(data=dfl['weight'])
```

Out[50]:
```
<AxesSubplot:xlabel='weight', ylabel='Count'>
```



Takisto sa v dátach vyskytujú chýbajúce hodnoty. Vzhľadom na ich relatívne nízku početnosť v porovnaní s celkovým počtom záznamov by možným riešením bolo nahradenie chýbajúcich hodnôt priemernými hodnotami daného atribútu (berúc do úvahy zistenie, že chýbajúce hodnoty sa vyskytujú len a práve pri číselných atribútoch, ktoré sú výsledkom samotného vyšetrenia/merania).

In [51]:
```
dfl.isnull().sum()
```

Out[51]:
```
leukocyty       30
ssn              0
name             0
smoker           0
hemoglobin      30
trombocyty      30
indicator        0
alt             30
relationship     0
weight           0
ast             30
alp             30
hematokrit      30
hbver           30
etytr           30
er-cv           30
erytrocyty      30
dtype: int64
```

In [52]:
```
len(dfl[dfl.isnull().any(axis=1)])
```

Out[52]:
```
324
```

Ďalším problémom je nekonzistentné vyplnenie stĺpca smoker. Tento problém už teraz vieme jednoducho opraviť.

In [53]:
```
dfl.smoker.unique()
```

Out[53]:
```
array(['no', 'Y', 'yes', 'N'], dtype=object)
```

In [54]:
```
dfl['smoker'] = dfl['smoker'].str.replace('N','no')
dfl['smoker'] = dfl['smoker'].str.replace('Y','yes')
dfl.smoker.unique()
```

Out[54]:
```
array(['no', 'yes'], dtype=object)
```

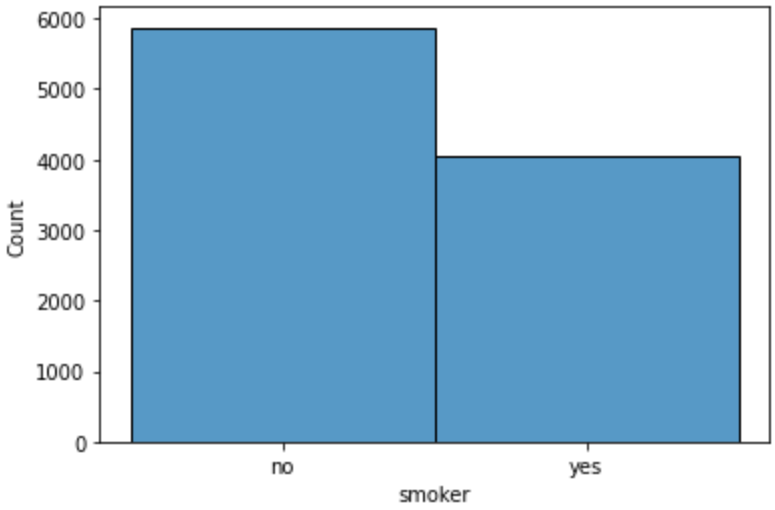
Párová analýza dát

Fajčiari vs. nefajčiari

Ako prvé sa pozrieme na celkový podiel fajčiarov a nefajčiarov.

In [55]:
```
sns.histplot(data=dfl['smoker'])
```

Out[55]:
```
<AxesSubplot:xlabel='smoker', ylabel='Count'>
```



In [56]:
```
len(dfl[dfl['smoker'] == 'no']) / len(dfl) * 100
```

Out[56]:
```
59.20548497680984
```

Z grafu je nám jasné, že v datasete sa nachádza väčší počet nefajčiarov ako fajčiarov. Nefajčiari tvoria zaokrúhlene 59.2% z celkového počtu záznamov.

Krvné skupiny a pohlavia

Ako ďalšie sa pozrieme na rozdelenie krvných skupín a ich zastúpenie medzi mužmi a ženami. Tieto dáta sa nachádzajú v prvej tabuľke profiles.

In [57]:
```
dfp.groupby(['blood_group', 'sex']).size()
```

Out[57]:
```
blood_group  sex
A+           F      200
             M      206
A-           F      186
             M      186
AB+          F      209
             M      184
AB-          F      175
             M      206
B+           F      203
             M      176
B-           F      188
             M      173
O+           F      204
             M      208
O-           F      198
             M      180
dtype: int64
```

Vo viacerých krvných skupinách môžeme pozorovať približne rovnaké zastúpenie mužov a žien (napr. A+, A-, B-,

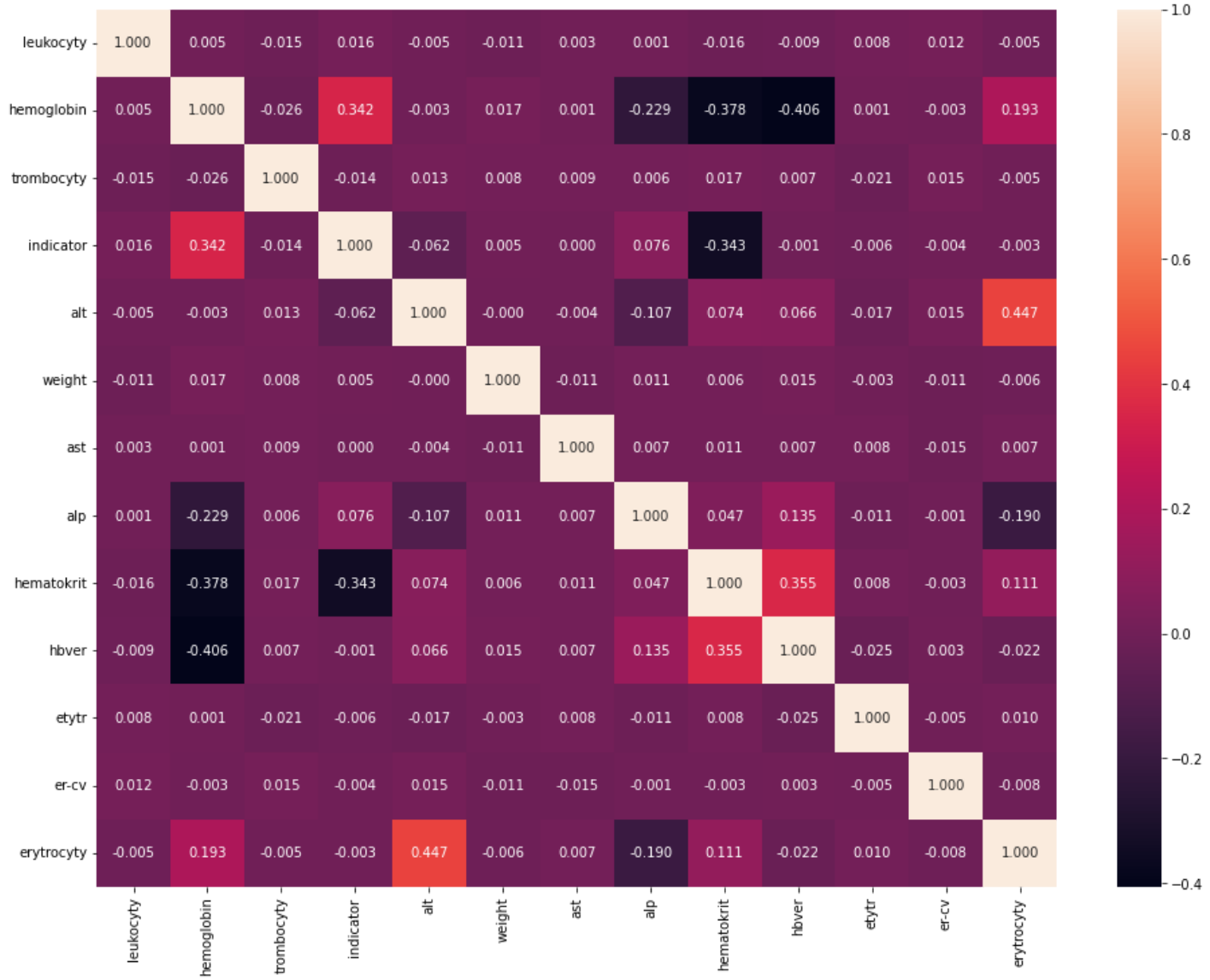
0+).

Závislosti medzi číselnými atribútmi tabuľky labor

Pre zistenie závislostí medzi dvojicami číselných atribútov si najskôr vykreslíme heatmapu a pairplot všetkých z nich.

In [58]:
```
fig, ax = plt.subplots(figsize=(16,12))
sns.heatmap(dfl.corr(), ax=ax, annot=True, fmt=".3f")
```

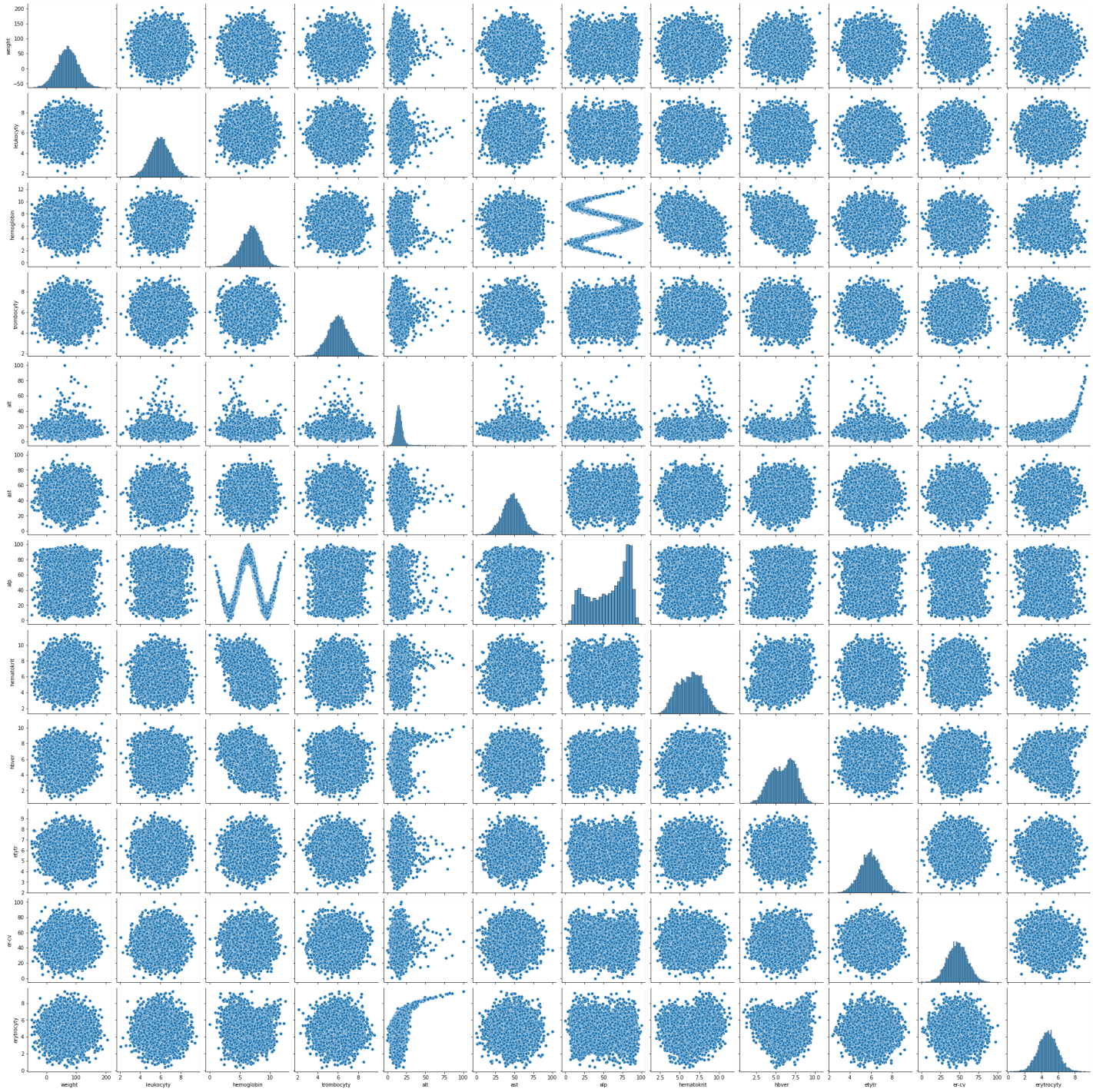
Out[58]:
```
<AxesSubplot:>
```



Z výsledkov heatmapy vidíme, že väčšina atribútov má medzi sebou koreláciu s hodnotou blížiacou s k 0, čo znamená, že medzi ich hodnotami nie je skoro žiadny vzťah a teda nie sú od seba skoro vôbec závislé. Pozitívna korelácia, kedy sa hodnota korelácie približuje alebo rovná 1 a teda hodnoty oboch atribútov sa pohybujú rovnakým smerom je výraznejšia napríklad medzi atribútmi **indikátor + hemoglobín**, **alt + erytrocyty** alebo **hematokrit + hbver**. Negatívna korelácia, kedy sa hodnota korelácie približuje alebo rovná -1 a teda hodnoty oboch atribútov sa pohybujú opačným smerom je výraznejšia napríklad medzi atribútmi **hemoglobín + alp**, **hemoglobín + hematokrit**, **hemoglobín + hbver** alebo **indikátor + hematokrit**.

In [59]:
```
sns.pairplot(dfl[['weight', 'leukocyty', 'hemoglobin', 'trombocyty', 'alt', 'ast', 'alp',
     'er-cv', 'erytrocyty']])
```

Out[59]:
```
<seaborn.axisgrid.PairGrid at 0x2dfc35a0130>
```

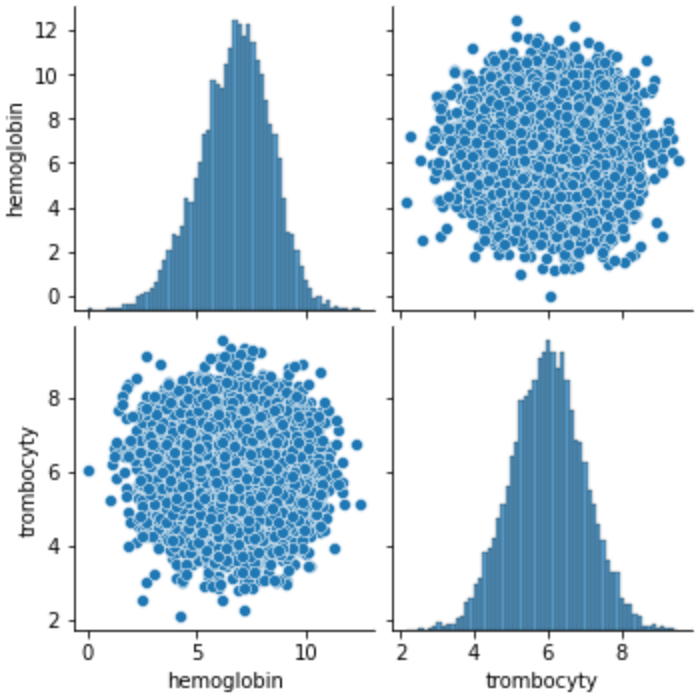


Väčšina z dvojíc atribútov medzi sebou má veľmi nízku, takmer žiadnu závislosť a ich grafy vyzerajú podobne ako tento príklad:

Hemoglobín a trombocyty

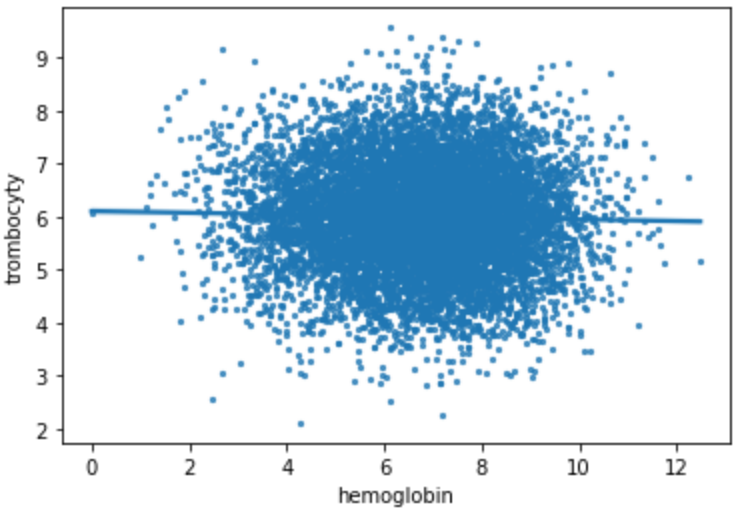
In [60]:
```
sns.pairplot(dfl[['hemoglobin', 'trombocyty']])
```

Out[60]:
```
<seaborn.axisgrid.PairGrid at 0x2dfcd0cff40>
```



In [61]:
```
sns.regplot(data=dfl, x="hemoglobin", y="trombocyty", scatter_kws={'s':5})
```

Out[61]:
```
<AxesSubplot:xlabel='hemoglobin', ylabel='trombocyty'>
```

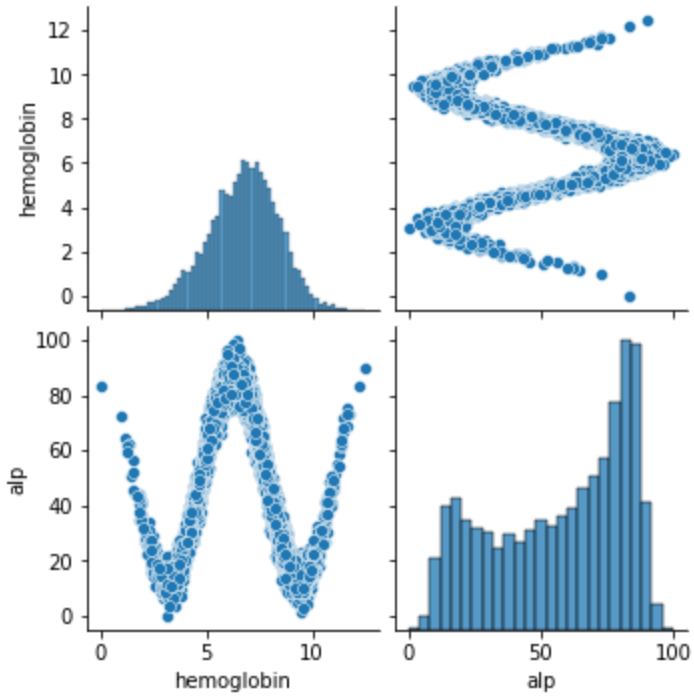


Atribúty hemoglobín a trombocyty medzi sebou nemajú žiadnu koreláciu.

Hemoglobín a alp

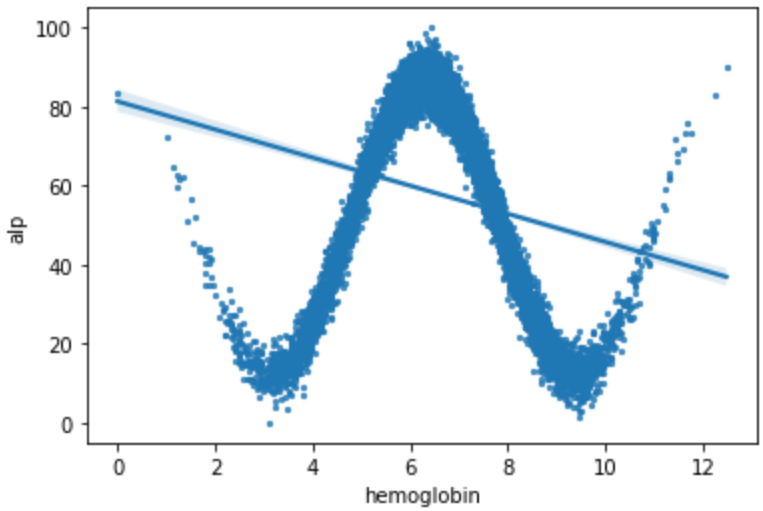
In [62]:
```
sns.pairplot(dfl[['hemoglobin', 'alp']])
```

Out[62]:
```
<seaborn.axisgrid.PairGrid at 0x2dfcf48b790>
```



In [63]:
```
sns.regplot(data=dfl, x="hemoglobin", y="alp", scatter_kws={'s':5})
```

Out[63]:
```
<AxesSubplot:xlabel='hemoglobin', ylabel='alp'>
```



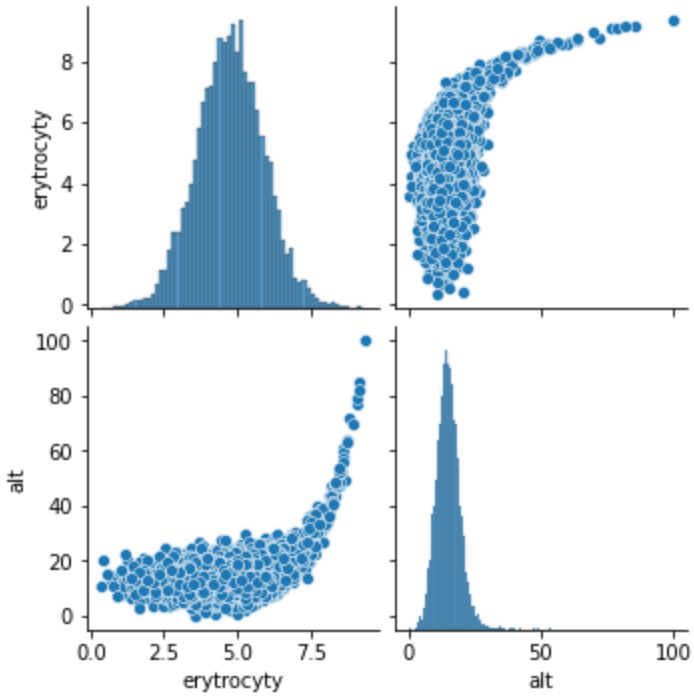
Atribúty hemoglobín a alp majú medzi sebou malú negatívnu koreláciu, čo znamená že so zvyšujúcou sa hodnotou jedného atribútu sa znižuje hodnota druhého atribútu a opačne.

Erytrocyty a alt

Dvojica, ktorá sa podľa grafu vizuálne najviac približuje nejakej forme závislosti, je *alt* (alanín transmináza) a *erytrocyty*. S vyššou hodnotou erytrocytov sa zvyšuje aj hodnota alt. Krivka však nemá príliš blízko k tvaru priamky y=x, pripomína skôr exponenciálnu funkciu (resp. logaritmickú, podľa zvolenia osí).

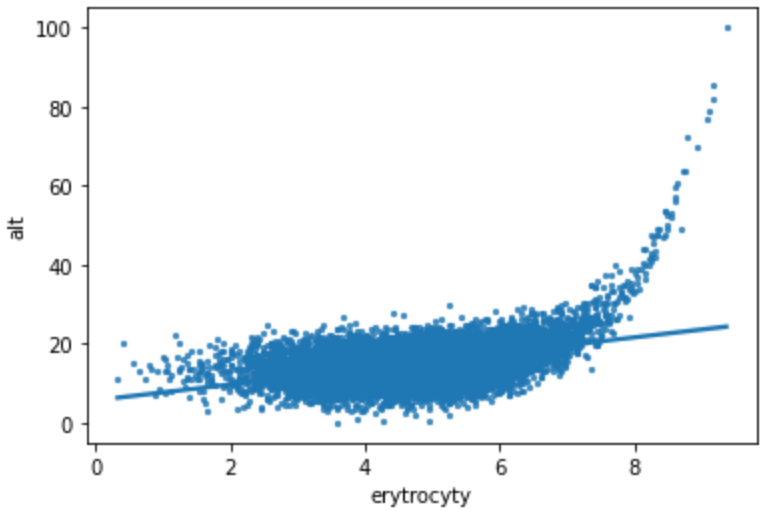
In [64]:
```
sns.pairplot(dfl[['erytrocyty', 'alt']])
```

Out[64]:
```
<seaborn.axisgrid.PairGrid at 0x2dfce469430>
```



In [65]:
```
sns.regplot(data=dfl, x="erytrocyty", y="alt", scatter_kws={'s':5})
```

Out[65]:
```
<AxesSubplot:xlabel='erytrocyty', ylabel='alt'>
```



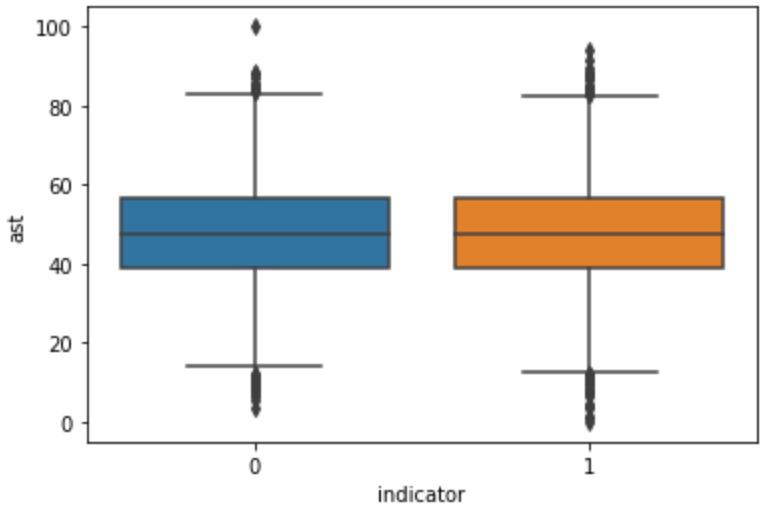
Atribúty erytrocyty a alt majú medzi sebou tzv. "curvilinear relationship", čo znamená, že medzi hodnotami ktoré nadobúdajú je určitá miera korelácie (v tomto prípade pozitívnej), avšak ich korelácia nie je lineárna.

Závislosti jednotlivých atribútov od predikovanej premennej

Predikovaná premenná (indikátor) nám dáta rozdeľuje na dve skupiny (hodota indikátora 0 a 1). Pomocou grafov si môžeme vykresľovať distribúcie jednotlibých premenných rozdelené podľa indikátora a následne tieto dve skupiny porovnávať. Pri väčšine atribútov sme nepozorovali takmer žiaden rozdiel v závislosti od hodnoty indikátora, ako vidíme na nasledujúcom príklade atribútov **Indikátor a ast**:

In [66]:
```
sns.boxplot(data=[dfl[dfl['indicator'] == 0]['ast'], dfl[dfl['indicator'] == 1]['ast']]).s
```

Out[66]:
```
[Text(0.5, 0, 'indicator'), Text(0, 0.5, 'ast')]
```

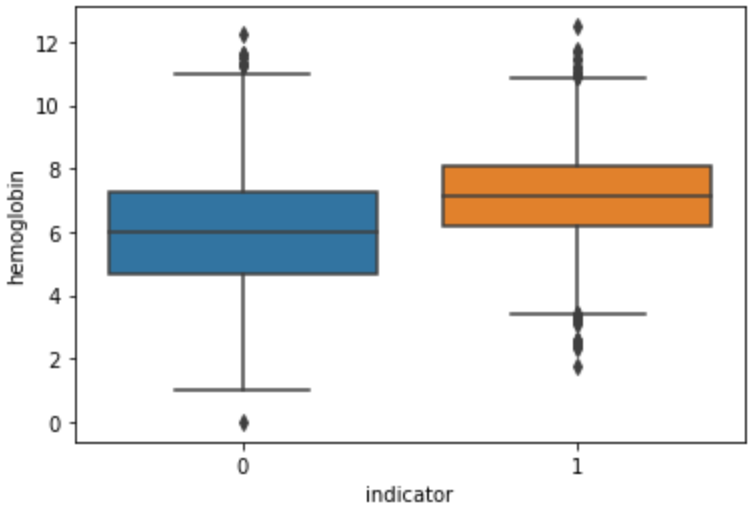


Tri atribúty v heat mape vizuálne preukázali nejakú formu korelácie vzhľadom na indikátor: *hemoglobín*, *alp (alkalická fosfatáza)* a *hematokrit*. Aby bola závislosť lepšie viditeľáý, vykreslili sme ju okrem box-plotu aj cez joint-plot a kde-plot.

Indikátor a hemoglobín

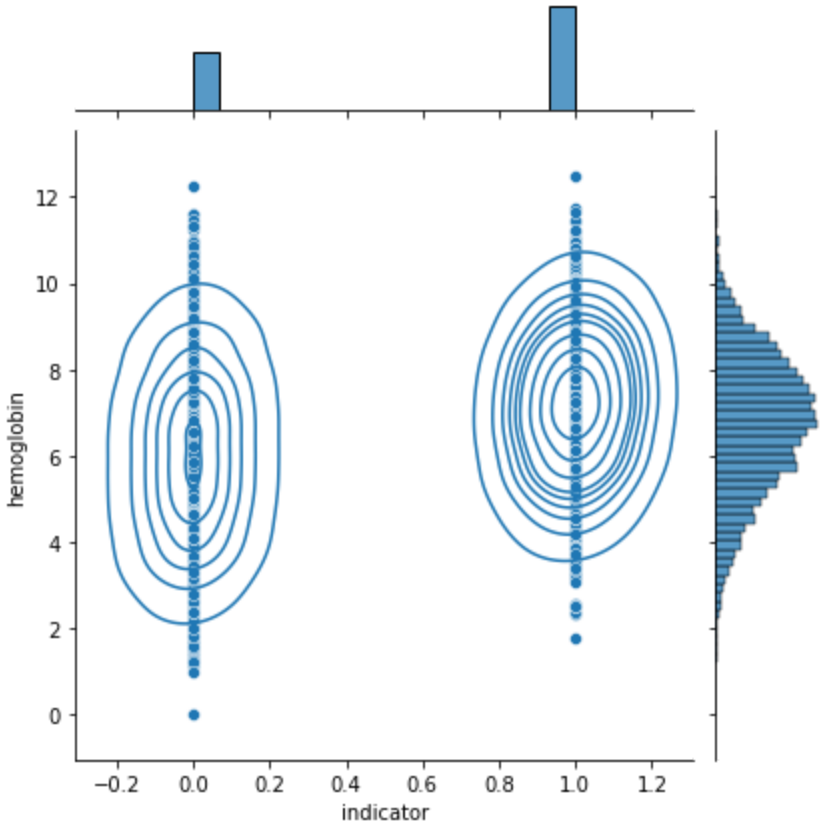
In [67]:
```
sns.boxplot(data=[dfl[dfl['indicator'] == 0]['hemoglobin'], dfl[dfl['indicator'] == 1]['he
```

Out[67]:
```
[Text(0.5, 0, 'indicator'), Text(0, 0.5, 'hemoglobin')]
```



In [68]:
```
sns.jointplot(x='indicator', y='hemoglobin', data=dfl).plot_joint(sns.kdeplot, n_levels=1(
```

Out[68]:
```
<seaborn.axisgrid.JointGrid at 0x2dfcfc2b760>
```

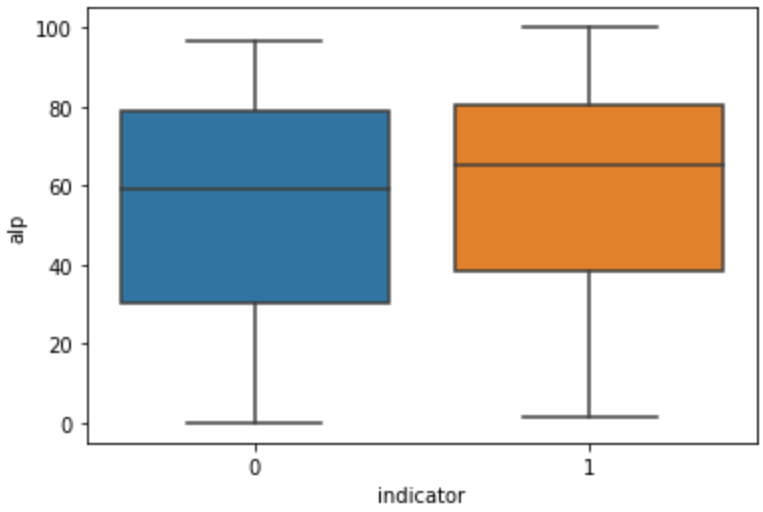


Z grafu môžeme pozorovať hustejší výskyt vyšších hodnôt hemoglobínu pri indikátore s hodnotou 1. Hodnoty hemoglobínu sú výrazne nižšie v prípade indikátora s hodnotou 0.

Indikátor a alp

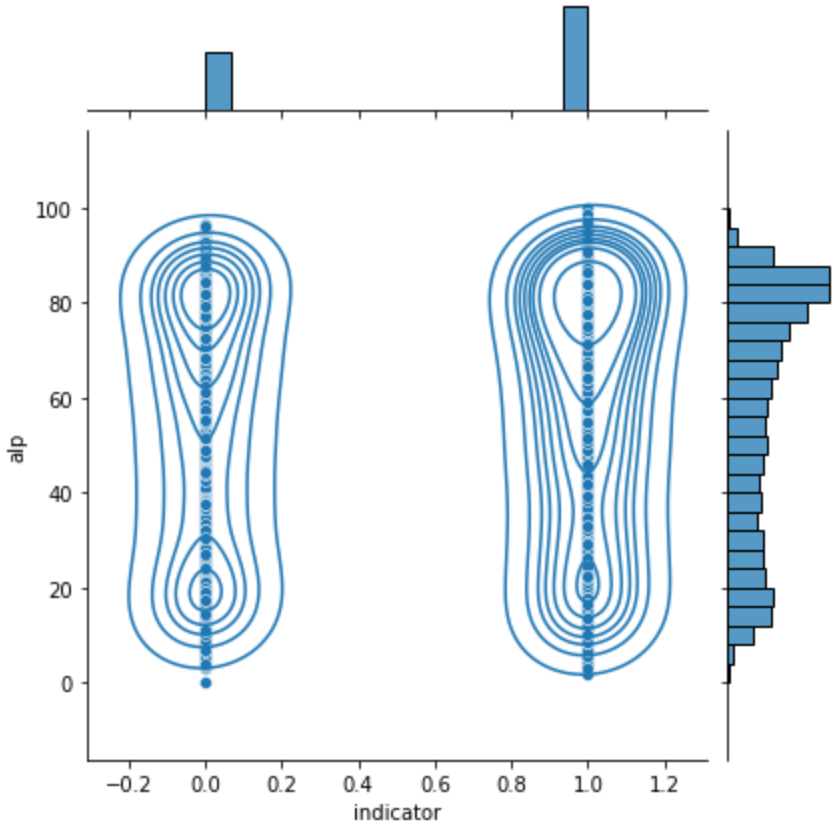
In [69]:
```
sns.boxplot(data=[dfl[dfl['indicator'] == 0]['alp'], dfl[dfl['indicator'] == 1]['alp']]).s
```

Out[69]:
```
[Text(0.5, 0, 'indicator'), Text(0, 0.5, 'alp')]
```



In [70]:
```
sns.jointplot(x='indicator', y='alp', data=dfl).plot_joint(sns.kdeplot, n_levels=10)
```

Out[70]:
```
<seaborn.axisgrid.JointGrid at 0x2dfcfdfca90>
```

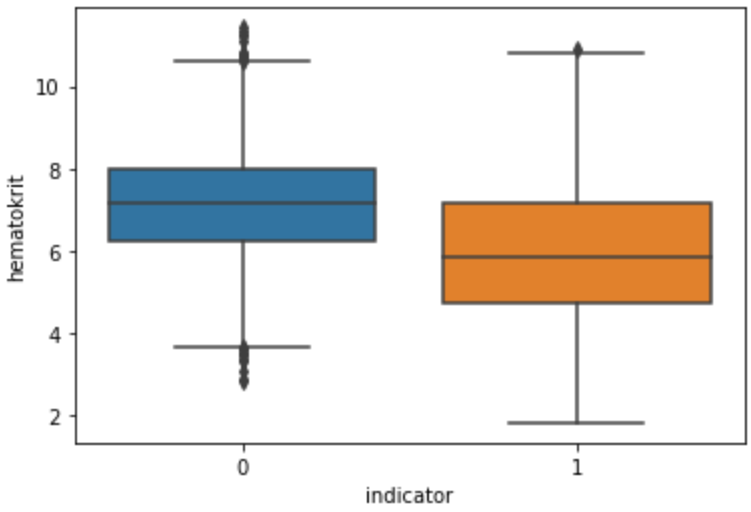


Z grafov môžeme pozorovať výrazne hustejší výskyt vysokých hodnôt alp pri indikátore s hodnotou 1. Pri indikátore 0 je v prvom grafe vidno výskyt nižších hodnôt alp.

Indikátor a hematokrit

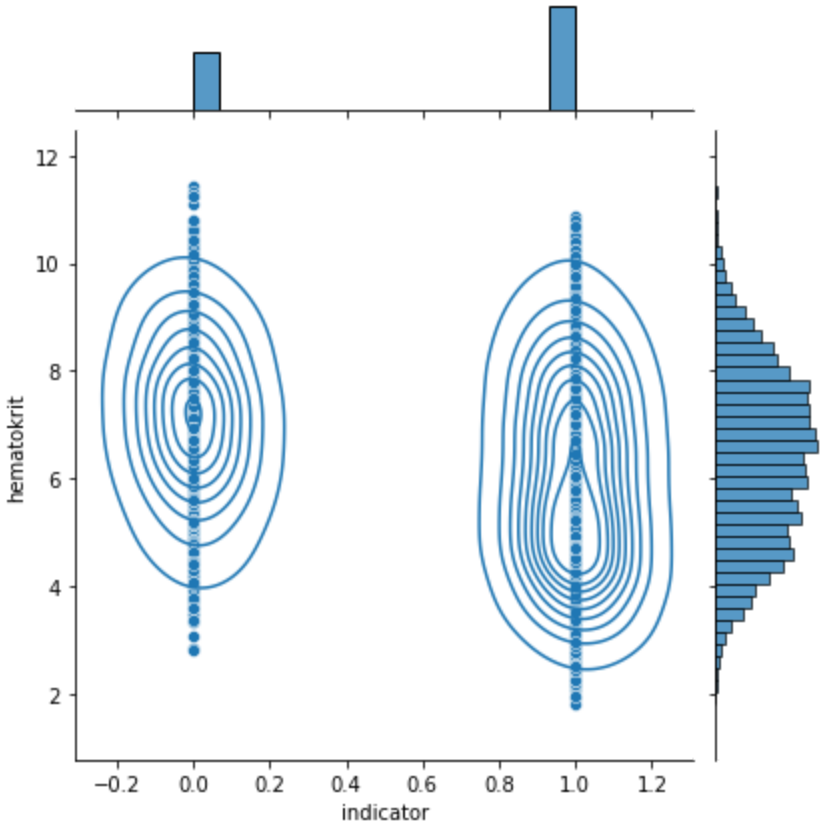
In [71]:
```
sns.boxplot(data=[dfl[dfl['indicator'] == 0]['hematokrit'], dfl[dfl['indicator'] == 1]['he
```

Out[71]:
```
[Text(0.5, 0, 'indicator'), Text(0, 0.5, 'hematokrit')]
```



In [72]:
```
sns.jointplot(x='indicator', y='hematokrit', data=dfl).plot_joint(sns.kdeplot, n_levels=1(
```

Out[72]:
```
<seaborn.axisgrid.JointGrid at 0x2dfd313b340>
```



Z grafu môžeme pozorovať výrazne väčší počet nízkych hodnôt hematokritu ak je hodnota indikátora rovná 1. Pri indikátore s hodnotou 0 je väčší výskyt vyšších hodnôt hematokritu.

Formulácia a štatistické overenie hypotéz o dátach

Prvá hypotéza

Chceme overiť, či má hladina hemoglobínu vplyv na indikátor.

Určíme si naše hypotézy nasledovne:

H_0 **(nulová hypotéza)**: Hladina hemoglobínu pacientov s indikátorom 0 **je** v priemere **rovnaká** ako hladina hemoglobínu pacientov s indikátorom 1.

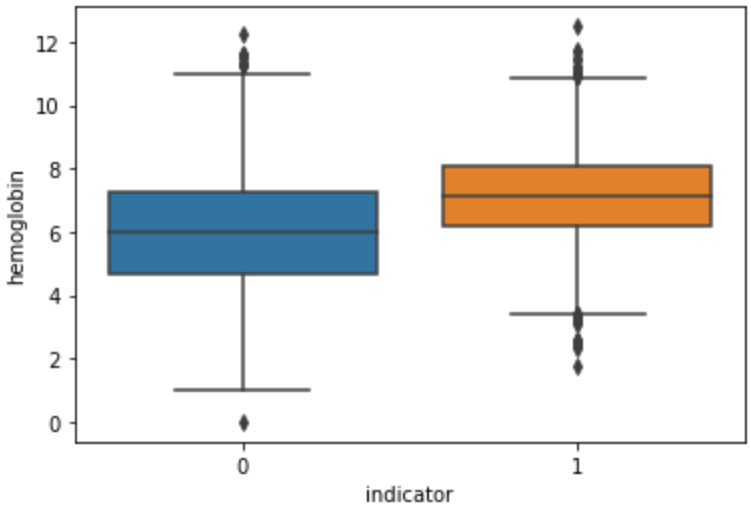
$H_1 = H_A$ **(alternatívna hypotéza)**: Hladina hemoglobínu pacientov s indikátorom 0 **je** v priemere **iná/väčšia/menšia** ako hladina hemoglobínu pacientov s indikátorom 1.

Rozdelíme si hladiny hemoglobínu na dve skupiny podľa indikátora (odstránime NaN, ktorých bolo dokopy len 30).

In [73]:
```
hemo_0 = dfl[dfl['indicator'] == 0]['hemoglobin'].dropna()
hemo_1 = dfl[dfl['indicator'] == 1]['hemoglobin'].dropna()
```

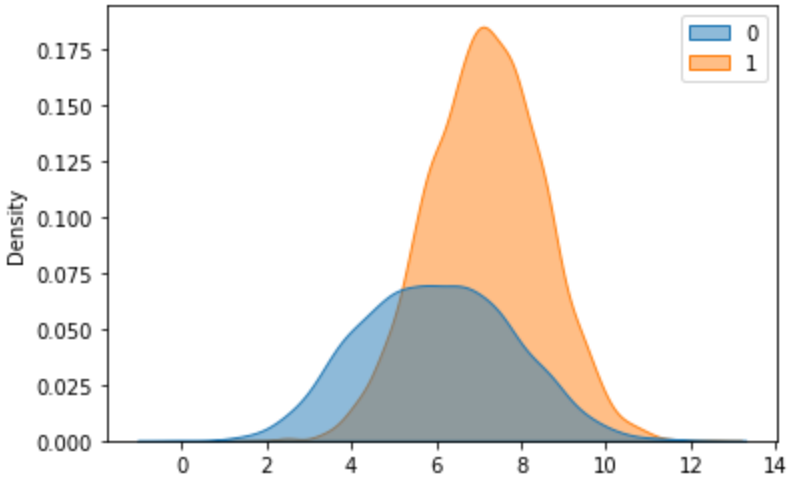
In [74]:
```
sns.boxplot(data=[hemo_0, hemo_1]).set(xlabel='indicator', ylabel='hemoglobin')
```

Out[74]:
```
[Text(0.5, 0, 'indicator'), Text(0, 0.5, 'hemoglobin')]
```



In [75]:
```
d = {'0': hemo_0, '1': hemo_1}
df = pd.DataFrame(data=d)
sns.kdeplot(data=df, fill = True, alpha = 0.5)
```

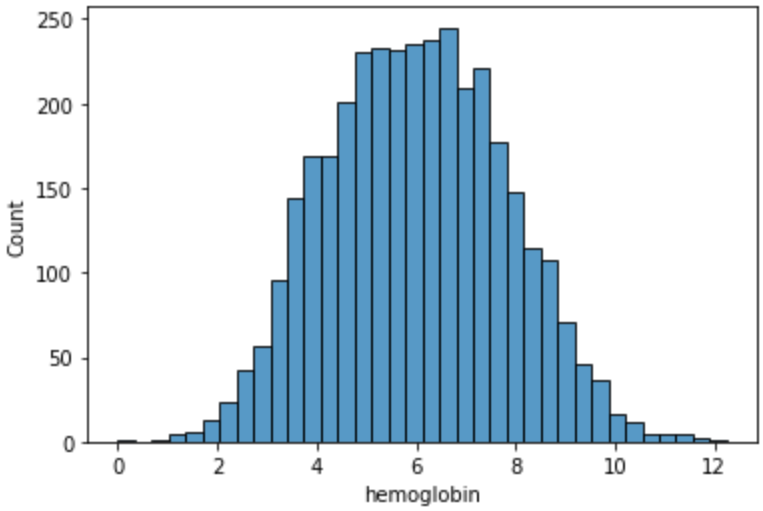
Out[75]:
```
<AxesSubplot:ylabel='Density'>
```



Z prvotných grafov sa vizuálne zdá, že rozdiel existuje. Musíme však zistiť, či je štatisticky signifikantný. Na to sa často používa **Studentov t-test**, pre ktorý ale musia skupiny dát spĺňať určité podmienky. Prvou je, že obe skupiny musia pochádzať z normálneho rozdelenia.

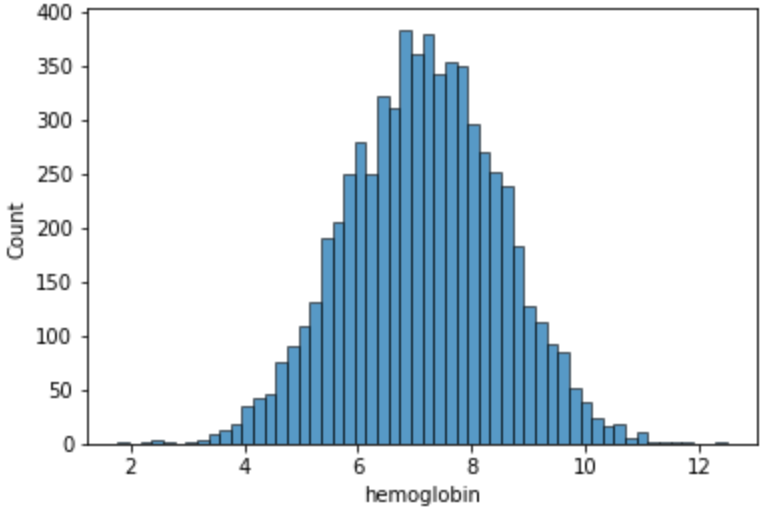
In [76]:
```
sns.histplot(data=hemo_0)
```

Out[76]:
```
<AxesSubplot:xlabel='hemoglobin', ylabel='Count'>
```



In [77]:
```
sns.histplot(data=hemo_1)
```

Out[77]:
```
<AxesSubplot:xlabel='hemoglobin', ylabel='Count'>
```



Prvotne sa môže zdať, že obe supiny majú normálne rozdelenie. Tento predpoklad si však musíme lepšie overiť.

In [78]:
```
##funkcia na detekciu outlierov
def identify_outliers(a):
    lower = a.quantile(0.25) - 1.5 * stats.iqr(a)
    upper = a.quantile(0.75) + 1.5 * stats.iqr(a)

    return a[(a > upper) | (a < lower)]
```

In [79]:
```
hemo_0_out = identify_outliers(hemo_0)
hemo_1_out = identify_outliers(hemo_1)
print(len(hemo_0_out))
print(len(hemo_1_out))
```

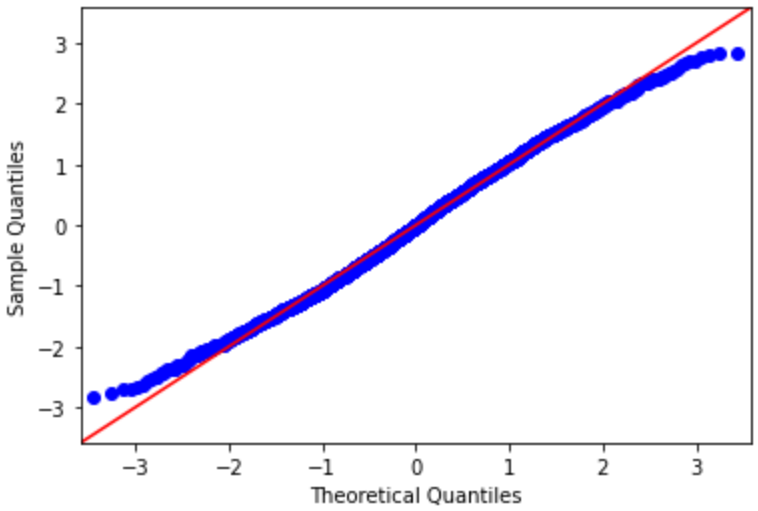
```
9
35
```

Obe skupiny obsahujú outlierov, preto ich odstránime:

In [80]:
```
hemo_0 = hemo_0.drop(hemo_0_out.index)
hemo_1 = hemo_1.drop(hemo_1_out.index)
```

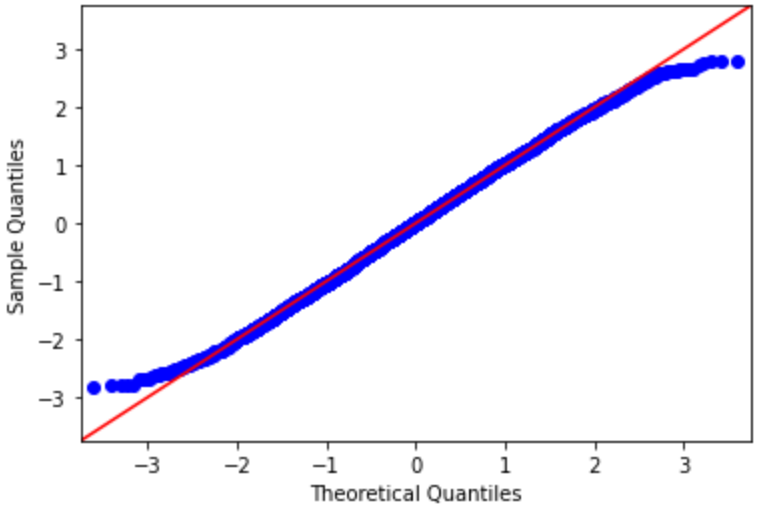
In [81]:
```
_ = sm.ProbPlot(hemo_0, fit=True).qqplot(line='45')
```

```
C:\Users\misul\AppData\Local\Programs\Python\Python39\lib\site-packages\statsmodels\graphics\gofplots.py:993: UserWarning: marker is redundantly defined by the 'marker' keyword argument and the fmt string "bo" (-> marker='o'). The keyword argument will take precedence.
  ax.plot(x, y, fmt, **plot_style)
```



In [82]:
```
_ = sm.ProbPlot(hemo_1, fit=True).qqplot(line='45')
```

```
C:\Users\misul\AppData\Local\Programs\Python\Python39\lib\site-packages\statsmodels\graphics\gofplots.py:993: UserWarning: marker is redundantly defined by the 'marker' keyword argument and the fmt string "bo" (-> marker='o'). The keyword argument will take precedence.
  ax.plot(x, y, fmt, **plot_style)
```



Z QQ-plotu sa javí, že skupiny nepochádzajú z normálneho rozdelenia, ale ich rozdelenia sa na seba podobajú.

In [83]:
```
stats.kurtosis(hemo_0)
```

Out[83]:
```
-0.4852596592842815
```

In [84]:
```
stats.kurtosis(hemo_1)
```

Out[84]:
```
-0.3138142815914269
```

In [85]:
```
stats.skew(hemo_0)
```

Out[85]:
```
0.049008205858826454
```

In [86]:
```
stats.skew(hemo_1)
```

Out[86]:
```
-0.030752183231020084
```

Rozdelenia našich skupín majú podobné vlastnosti špičatosti (kurtosis), avšak mierne sa líšia v asymetrii, nakoľko skupina s indikátorom 0 je viac ťažká na ľavej strane a skupina s indikátorom 1 na pravej strane.

Či naozaj nepochádzajú z normálneho rozdelenia môžeme overiť **Shapiro-Wilkovým testom**.

In [87]:
```
stats.shapiro(hemo_0)
```

Out[87]:
```
ShapiroResult(statistic=0.9959529042243958, pvalue=3.599223319383782e-08)
```

In [88]:
```
stats.shapiro(hemo_1)
```

Out[88]:
```
C:\Users\misul\AppData\Local\Programs\Python\Python39\lib\site-packages\scipy\stats\morest
ats.py:1760: UserWarning: p-value may not be accurate for N > 5000.
  warnings.warn("p-value may not be accurate for N > 5000.")
ShapiroResult(statistic=0.998154878616333, pvalue=7.563602366644773e-07)
```

Testovali sme nulovú hypotézu H_0, že dáta pochádzajú z normálneho rozdelenia. Ak je $p < 0,05$, nulovú hypotézu H_0 zamietame a dáta pravdepodobne pochádzajú z iného ako normálneho rozdelenia. Ak je $p > 0,05$, nulovú hypotézu H_0 nezamietame, teda na základe dát nemôžeme prehlásiť, že by dáta pochádzali z iného, ako normálneho rozdelenia.

Pre rozdelenia oboch skupín nám vyšla hodnota $p < 0,05$, nulovú hypotézu teda H_0 zamietame a dáta pravdepodobne pochádzajú z iného ako normálneho rozdelenia. Mali by sme teda použiť neparametrickú verziu t-testu, t. j. **Mann-Whitneyho U-test**

Studentov t-test vyžaduje tiež, aby rozdelenia oboch skupín mali podobnú varianciu. Môžme si ukázať tiež **Levenov test**, ktorý na overenie tejto podmienky slúži, aj keď už z predošlých testov je jasné, že Studentov t-test nemôžeme použiť. Levenov test testuje nulovú hypotézu H_0, že všetky vstupné vzorky pochádzajú z rozdelení s rovnakými varianciami. Ak H_0 nezamietame ($p > 0,05$), znamená to, že na základe dát nemôžeme prehlásiť, že by vzorky pochádzali z distribúcií s rôznymi varianciami.

In [89]:
```
stats.levene(hemo_0, hemo_1)
```

Out[89]:
```
LeveneResult(statistic=438.1176485022613, pvalue=3.237783967301522e-95)
```

p-value je blížiaci sa k nule, to znamená, že rozdelenia našich skupín nemajú ani podobné variancie.

Nebol teda splnený ani jeden z predpokladov na použitie t-testu, preto spustíme **Mann-Whitneyho U-test**.

In [90]:
```
stats.mannwhitneyu(hemo_0, hemo_1)
```

Out[90]:
```
MannwhitneyuResult(statistic=6737046.0, pvalue=6.312924688938223e-230)
```

Keďže $p < 0,001$, pravdepodobnosť chyby 1. rádu (že H_0 je pravdivá a my ju zamietame) je menej ako 1 promile. Našu nulovú hypotézu H_0 teda zamietame v prospech alternatívnej hypotézy H_A. Rozdiel v priemernej hladine hemoglobínu medzi pacientami s indikátorom 0 a indikátorom 1 je štatisticky signifikantný.

Môžeme si vizualizovať rozdiel medzi dvoma priemermi spolu s intervalmi spoľahlivosti, ktoré nám hovoria, že s N% pravdepodobnosťou (najčastejšie sa používa 95) sa skutočná hodnota priemeru bude nachádzať niekde v danom intervale.

In [91]:
```
sms.DescrStatsW(hemo_0).tconfint_mean()
```

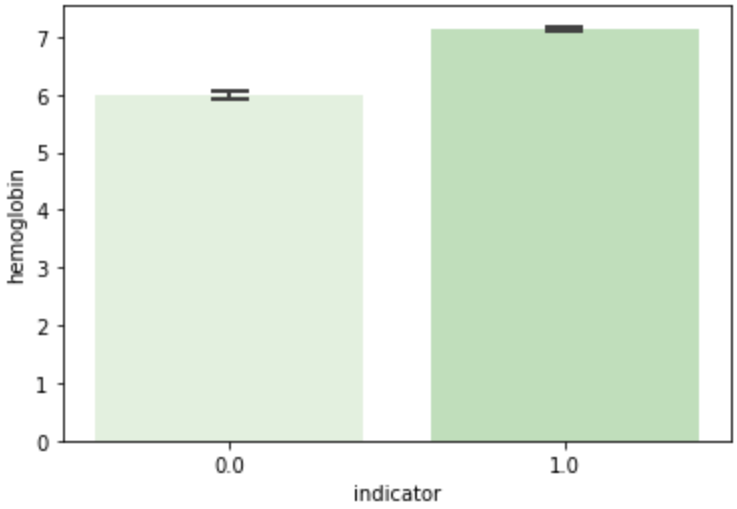
Out[91]:
```
(5.921212637445691, 6.038326141093122)
```

In [92]:
```
sms.DescrStatsW(hemo_1).tconfint_mean()
```

Out[92]:
```
(7.11533480518697, 7.1805706900811686)
```

In [93]:
```
sns.barplot(x='indicator', y='hemoglobin', data=dfl[(dfl.indicator == 0) | (dfl.indicator
            capsize=0.1, errwidth=2, palette=sns.color_palette("Greens"))
```

Out[93]:
```
<AxesSubplot:xlabel='indicator', ylabel='hemoglobin'>
```



Záver: hypotézu H_0, že priemerná hladina hemoglobínu u pacientov s indikátorom 0 a pacientov s indikátorom 1 je rovnaká, sme zamietli v prospech alternatívnej hypotézy H_A. Priemerná hladina hemoglobínu u pacientov s indikátorom 0 je nižšia ako u pacientov s indikátorom 1 a tento rozdiel je štatisticky signifikantný.

Druhá hypotéza

Chceme overiť, či má hodnota hematokritu vplyv na indikátor.

Určíme si naše hypotézy nasledovne:

H_0 **(nulová hypotéza)**: Hodnota hematokritu pacientov s indikátorom 0 **je** v priemere **rovnaká** ako hodnota hematokritu pacientov s indikátorom 1.

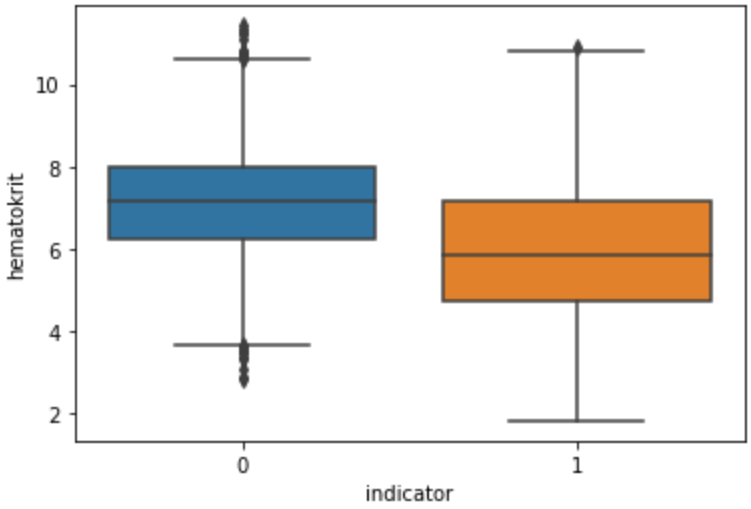
$H_1 = H_A$ **(alternatívna hypotéza)**: Hodnota hematokritu pacientov s indikátorom 0 **je** v priemere **iná/väčšia/menšia** ako hodnota hematokritu pacientov s indikátorom 1.

Rozdelíme si hodnoty hematokritu na dve skupiny podľa indikátora (odstránime NaN, ktorých bolo dokopy len 30).

In [94]:
```
hema_0 = dfl[dfl['indicator'] == 0]['hematokrit'].dropna()
hema_1 = dfl[dfl['indicator'] == 1]['hematokrit'].dropna()
```

In [95]:
```
sns.boxplot(data=[hema_0, hema_1]).set(xlabel='indicator', ylabel='hematokrit')
```

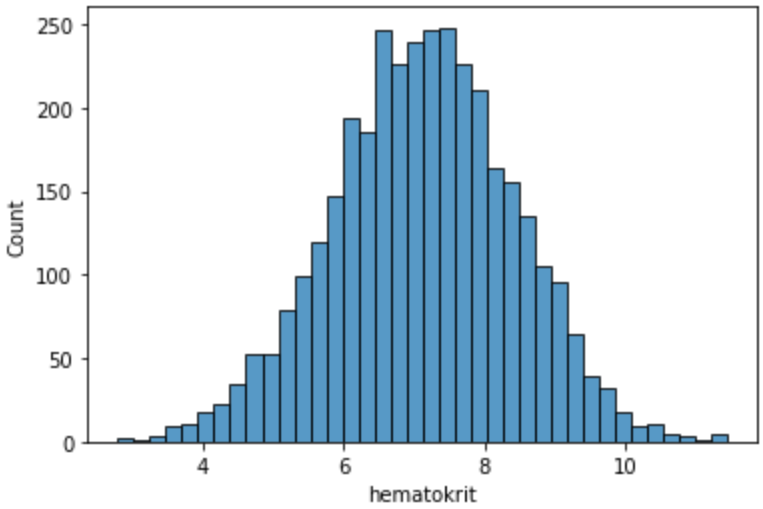
Out[95]:
```
[Text(0.5, 0, 'indicator'), Text(0, 0.5, 'hematokrit')]
```



Z prvotného grafu sa vizuálne zdá, že rozdiel existuje. Musíme však zistiť, či je štatisticky signifikantný. Na to sa často používa **Studentov t-test**, pre ktorý ale musia skupiny dát spĺňať určité podmienky. Prvou je, že obe skupiny musia pochádzať z normálneho rozdelenia.

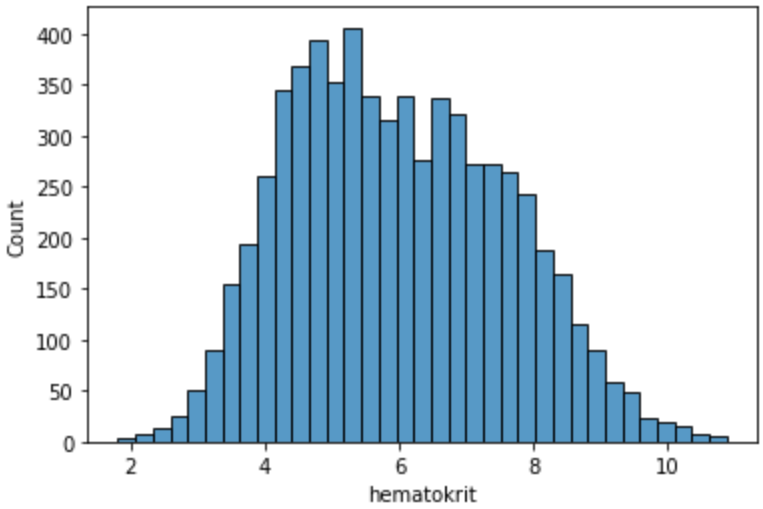
In [96]:
```
sns.histplot(data=hema_0)
```

Out[96]:
```
<AxesSubplot:xlabel='hematokrit', ylabel='Count'>
```



In [97]:
```
sns.histplot(data=hema_1)
```

Out[97]:
```
<AxesSubplot:xlabel='hematokrit', ylabel='Count'>
```



Prvotne sa môže zdať, že skupina s indikátorom 0 má normálne rozdelenie a skupina s indikátorom 1 nie je z normálneho rozdelenia. Tento predpoklad si však musíme lepšie overiť.

In [98]:
```
hema_0_out = identify_outliers(hema_0)
hema_1_out = identify_outliers(hema_1)
print(len(hema_0_out))
print(len(hema_1_out))
```

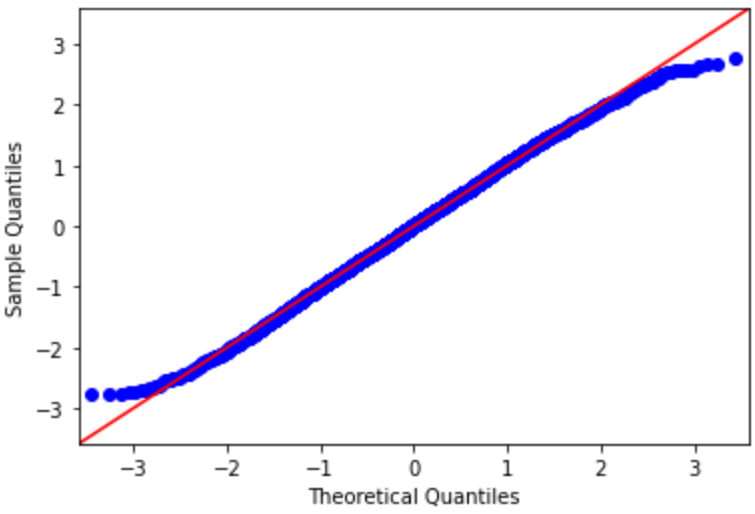
```
22
1
```

Obe skupiny obsahujú outlierov, preto ich odstránime:

In [99]:
```
hema_0 = hema_0.drop(hema_0_out.index)
hema_1 = hema_1.drop(hema_1_out.index)
```

In [100…
```
_ = sm.ProbPlot(hema_0, fit=True).qqplot(line='45')
```

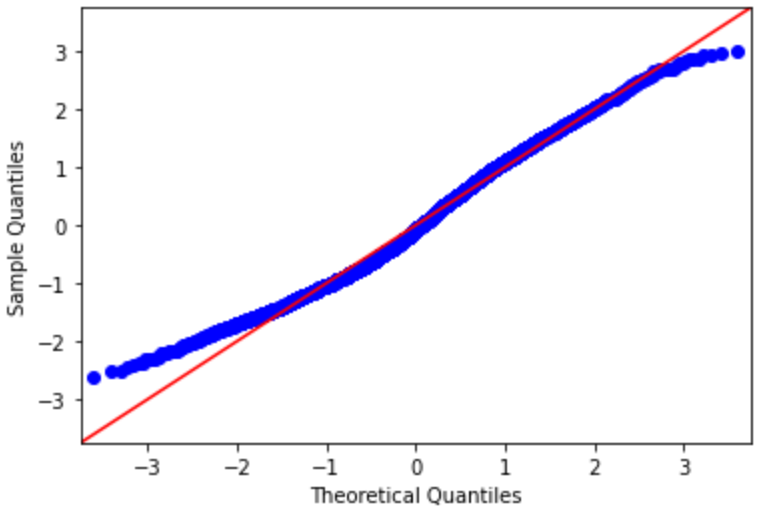
```
C:\Users\misul\AppData\Local\Programs\Python\Python39\lib\site-packages\statsmodels\graphics\gofplots.py:993: UserWarning: marker is redundantly defined by the 'marker' keyword argument and the fmt string "bo" (-> marker='o'). The keyword argument will take precedence.
  ax.plot(x, y, fmt, **plot_style)
```



In [101…
```
_ = sm.ProbPlot(hema_1, fit=True).qqplot(line='45')
```

```
C:\Users\misul\AppData\Local\Programs\Python\Python39\lib\site-packages\statsmodels\graphics\gofplots.py:993: UserWarning: marker is redundantly defined by the 'marker' keyword arg
```

```
ument and the fmt string "bo" (-> marker='o'). The keyword argument will take precedence.
  ax.plot(x, y, fmt, **plot_style)
```



Z QQ-plotu sa javí, že skupiny nepochádzajú z normálneho rozdelenia, a aj ich rozdelenia sa od seba mierne líšia.

In [102…
```
stats.kurtosis(hema_0)
```

Out[102…
```
-0.28235530662169195
```

In [103…
```
stats.kurtosis(hema_1)
```

Out[103…
```
-0.6021828788323154
```

In [104…
```
stats.skew(hema_0)
```

Out[104…
```
-0.06495368345547202
```

In [105…
```
stats.skew(hema_1)
```

Out[105…
```
0.24961028345168101
```

Rozdelenia našich skupín majú podobné vlastnosti špičatosti (kurtosis), ale odlišné vlastnosti asymetrie (skew). Skupina s indikátorom 1 má rozdelenie s výraznejším zošikmením smerom doľava.

Či naozaj nepochádzajú z normálneho rozdelenia môžeme overiť **Shapiro-Wilkovým testom**.

In [106…
```
stats.shapiro(hema_0)
```

Out[106…
```
ShapiroResult(statistic=0.9979398846626282, pvalue=0.00014385193935595453)
```

In [107…
```
stats.shapiro(hema_1)
```

Out[107…
```
C:\Users\misul\AppData\Local\Programs\Python\Python39\lib\site-packages\scipy\stats\morestats.py:1760: UserWarning: p-value may not be accurate for N > 5000.
  warnings.warn("p-value may not be accurate for N > 5000.")
ShapiroResult(statistic=0.9866166114807129, pvalue=4.154766699161236e-24)
```

Testovali sme nulovú hypotézu H_0, že dáta pochádzajú z normálneho rozdelenia. Ak je $p < 0,05$, nulovú

hypotézu H_0 zamietame a dáta pravdepodobne pochádzajú z iného ako normálneho rozdelenia. Ak je $p > 0,05$, nulovú hypotézu H_0 nezamietame, teda na základe dát nemôžeme prehlásiť, že by dáta pochádzali z iného, ako normálneho rozdelenia.

Pre rozdelenia oboch skupín nám vyšla hodnota $p < 0,05$, nulovú hypotézu teda H_0 zamietame a dáta pravdepodobne pochádzajú z iného ako normálneho rozdelenia. Mali by sme teda použiť neparametrickú verziu t-testu, t. j. **Mann-Whitneyho U-test**

Studentov t-test vyžaduje tiež, aby rozdelenia oboch skupín mali podobnú varianciu. Môžme si ukázať tiež **Levenov test**, ktorý na overenie tejto podmienky slúži, aj keď už z predošlých testov je jasné, že Studentov t-test nemôžeme použiť. Levenov test testuje nulovú hypotézu H_0, že všetky vstupné vzorky pochádzajú z rozdelení s rovnakými varianciami. Ak H_0 nezamietame ($p > 0,05$), znamená to, že na základe dát nemôžeme prehlásiť, že by vzorky pochádzali z distribúcií s rôznymi varianciami.

In [108…
```
stats.levene(hema_0, hema_1)
```

Out[108…
```
LeveneResult(statistic=329.5792052245033, pvalue=1.7836576653218764e-72)
```

p-value je blížiaci sa k nule, to znamená, že rozdelenia našich skupín nemajú ani podobné variancie.

Nebol teda splnený ani jeden z predpokladov na použitie t-testu, preto spustíme **Mann-Whitneyho U-test**.

In [109…
```
stats.mannwhitneyu(hema_0, hema_1)
```

Out[109…
```
MannwhitneyuResult(statistic=15799037.5, pvalue=2.1815414409207245e-261)
```

Keďže $p < 0,001$, pravdepodobnosť chyby 1. rádu (že H_0 je pravdivá a my ju zamietame) je menej ako 1 promile. Našu nulovú hypotézu H_0 teda zamietame v prospech alternatívnej hypotézy H_A. Rozdiel v priemernej hodnote hematokritu medzi pacientami s indikátorom 0 a indikátorom 1 je štatisticky signifikantný.

Môžeme si vizualizovať rozdiel medzi dvoma priemermi spolu s intervalmi spoľahlivosti, ktoré nám hovoria, že s N% pravdepodobnosťou (najčastejšie sa používa 95) sa skutočná hodnota priemeru bude nachádzať niekde v danom intervale.

In [110…
```
sms.DescrStatsW(hema_0).tconfint_mean()
```

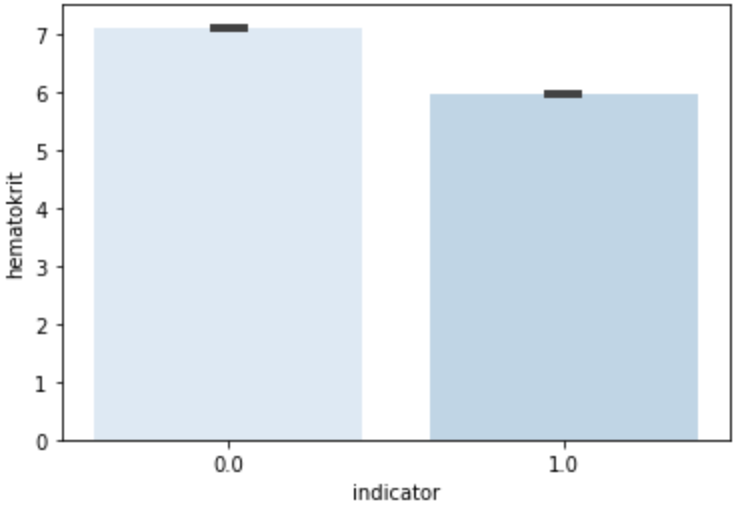
Out[110…
```
(7.072315102928971, 7.155624879888896)
```

In [111…
```
sms.DescrStatsW(hema_1).tconfint_mean()
```

Out[111…
```
(5.934521163963179, 6.01310805304608)
```

In [112…
```
sns.barplot(x='indicator', y='hematokrit', data=dfl[(dfl.indicator == 0) | (dfl.indicator
            capsize=0.1, errwidth=2, palette=sns.color_palette("Blues"))
```

Out[112…
```
<AxesSubplot:xlabel='indicator', ylabel='hematokrit'>
```



Záver: hypotézu H_0, že priemerná hodnota hematokritu u pacientov s indikátorom 0 a pacientov s indikátorom 1 je rovnaká, sme zamietli v prospech alternatívnej hypotézy H_A. Priemerná hodnota hematokritu u pacientov s indikátorom 0 je vyššia ako u pacientov s indikátorom 1 a tento rozdiel je štatisticky signifikantný.