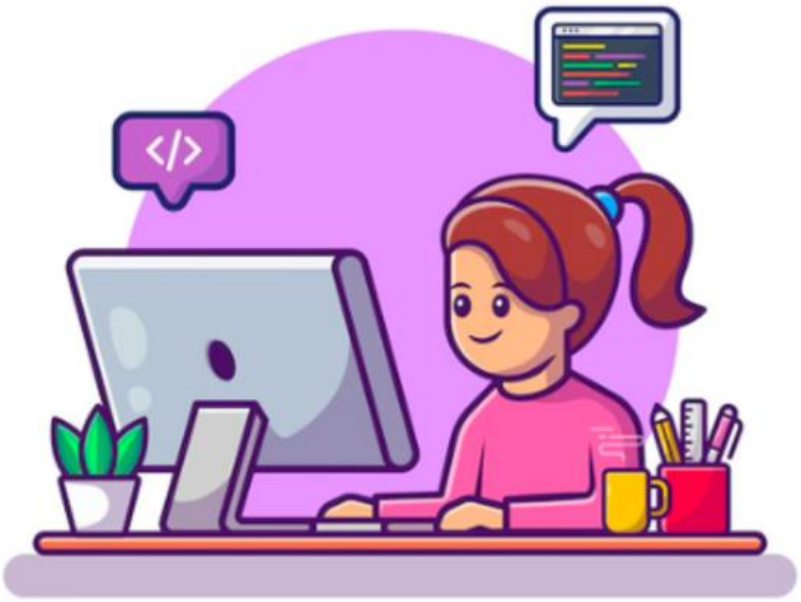
Универзитет “Св. Кирил и Методиј”

Факултет за информатички науки и компјутерско инженерство

Проектна задача по предметот: **Веб базирани системи**

Тема: **Анализа на податочното множество “Data Science Job Salaries”**



Изработила:
Тамара Малинова 185061

Професор:
Д-р Милош Јовановиќ

11.09.2022

1. Вовед

Како IDE за овој проект го избрав Spyder, кој е бесплатен, напишан во Python, за Python. Spyder е дизајниран за научници, инженери и аналитичари на податоци, кој меѓу другото се одликува и со визуелизација.

Со помош на софтверската библиотека Pandas, напишана за програмскиот јазик Python, која се користи за манипулација и анализа на податоци, во овој проект извршив анализа на податоците од множеството "Data Science Job Salaries".

2. Краток опис на податочното множество

"Data Science Job Salaries" е податочно множество кое содржи информации за платите на работните места од доменот на Data Science. За секој од податоците можеме да ги дознаеме work_year, experience_level, employment_type, job_title, salary, salary_currency, salary_in_usd, employee_residence, remote_ratio, company_location и company_size. Со work_year е прикажана годината за која е платена платата. Experience_level – нивото на искуство на работното место во текот на годината со зададени можни вредности: EN - Entry-level, MI – Mid-level, SE – Senior-level, EX – Executive-level. Employment_type – вид на вработување (PT Part-time, FT Full-time, CT Contract, FL Freelance. Job_title – работно место во текот на годината. Salary – вкупниот исплатен износ на бруто плата. Salary_currency – валута во која е исплатена платата. Salary_in_usd – исплатената плата пресметана во USD. Employee_residence – земјата на живеење во текот на работната година. Remote_ratio – работа извршена надвор од работното место. Company_location – земјата на главната канцеларија на работодавачот. Company_size – просечен број на вработени во текот на годината (S - помалку од 50 вработени, M - помеѓу 50 и 250 вреботени, L – повеќе од 250 вработени).

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2020	MI	FT	Data Scientist	70000	EUR	79833	DE	0	DE	L
1	2020	SE	FT	Machine Learning Scientist	260000	USD	260000	JP	0	JP	S
2	2020	SE	FT	Big Data Engineer	85000	GBP	109024	GB	50	GB	M
3	2020	MI	FT	Product Data Analyst	20000	USD	20000	HN	0	HN	S
4	2020	SE	FT	Machine Learning Engineer	150000	USD	150000	US	50	US	L
5	2020	EN	FT	Data Analyst	72000	USD	72000	US	100	US	L
6	2020	SE	FT	Lead Data Scientist	190000	USD	190000	US	100	US	S
7	2020	MI	FT	Data Scientist	11000000	HUF	35735	HU	50	HU	L
8	2020	MI	FT	Business Data Analyst	135000	USD	135000	US	100	US	L
9	2020	SE	FT	Lead Data Engineer	125000	USD	125000	NZ	50	NZ	S
10	2020	EN	FT	Data Scientist	45000	EUR	51321	FR	0	FR	S
11	2020	MI	FT	Data Scientist	3000000	INR	40481	IN	0	IN	L
12	2020	EN	FT	Data Scientist	35000	EUR	39916	FR	0	FR	M
13	2020	MI	FT	Lead Data Analyst	87000	USD	87000	US	100	US	L
14	2020	MI	FT	Data Analyst	85000	USD	85000	US	100	US	L
15	2020	MI	FT	Data Analyst	8000	USD	8000	PK	50	PK	L
16	2020	EN	FT	Data Engineer	4450000	JPY	41689	JP	100	JP	S
17	2020	SE	FT	Big Data Engineer	100000	EUR	114047	PL	100	GB	S
18	2020	EN	FT	Data Science Consultant	423000	INR	5707	IN	50	IN	M
19	2020	MI	FT	Lead Data Engineer	56000	USD	56000	PT	100	US	M
20	2020	MI	FT	Machine Learning Engineer	299000	CNY	43331	CN	0	CN	M
21	2020	MI	FT	Product Data Analyst	450000	INR	6072	IN	100	IN	L
22	2020	SE	FT	Data Engineer	42000	EUR	47899	GR	50	GR	L
23	2020	MI	FT	BI Data Analyst	98000	USD	98000	US	0	US	M
24	2020	MI	FT	Lead Data Scientist	115000	USD	115000	AE	0	AE	L
25	2020	EX	FT	Director of Data Science	325000	USD	325000	US	100	US	L
26	2020	EN	FT	Research Scientist	42000	USD	42000	NL	50	NL	L

3. Код

Најпрво треба да ја повикаме библиотеката Pandas со која како што кажав ќе ја спроведам анализата. Тоа го правиме со следната линија код:

```
import pandas as pd
```

За вчитување на csv фајлови се користи соодветната наредба во Pandas, read_csv. Со df = pd.read_csv (*r'G:\OneDrive\Desktop\WBS\ds_salaries.csv'*) го вчитувам ds_salaries.csv фајлот во кој ми е сместено податочното множество. Потоа со наредбата print(df) го печатам множеството и може да забележиме дека во ова множество има 607 записи.

```
     Unnamed: 0  work_year  ... company_location company_size
0             0        2020  ...               DE            L
1             1        2020  ...               JP            S
2             2        2020  ...               GB            M
3             3        2020  ...               HN            S
4             4        2020  ...               US            L
..          ...         ...  ...              ...          ...
602         602        2022  ...               US            M
603         603        2022  ...               US            M
604         604        2022  ...               US            M
605         605        2022  ...               US            M
606         606        2022  ...               US            L

[607 rows x 12 columns]
```

Доколку сакаме да ги дознаеме работните места во ова множество тоа може да го направиме со слетната наредба:

```
In [2]: df.job_title.unique()
Out[2]:
array(['Data Scientist', 'Machine Learning Scientist',
       'Big Data Engineer', 'Product Data Analyst',
       'Machine Learning Engineer', 'Data Analyst', 'Lead Data Scientist',
       'Business Data Analyst', 'Lead Data Engineer', 'Lead Data Analyst',
       'Data Engineer', 'Data Science Consultant', 'BI Data Analyst',
       'Director of Data Science', 'Research Scientist',
       'Machine Learning Manager', 'Data Engineering Manager',
       'Machine Learning Infrastructure Engineer', 'ML Engineer',
       'AI Scientist', 'Computer Vision Engineer',
       'Principal Data Scientist', 'Data Science Manager', 'Head of Data',
       '3D Computer Vision Researcher', 'Data Analytics Engineer',
       'Applied Data Scientist', 'Marketing Data Analyst',
       'Cloud Data Engineer', 'Financial Data Analyst',
       'Computer Vision Software Engineer',
       'Director of Data Engineering', 'Data Science Engineer',
       'Principal Data Engineer', 'Machine Learning Developer',
       'Applied Machine Learning Scientist', 'Data Analytics Manager',
       'Head of Data Science', 'Data Specialist', 'Data Architect',
       'Finance Data Analyst', 'Principal Data Analyst',
       'Big Data Architect', 'Staff Data Scientist', 'Analytics Engineer',
       'ETL Developer', 'Head of Machine Learning', 'NLP Engineer',
       'Lead Machine Learning Engineer', 'Data Analytics Lead'],
      dtype=object)
```

Оваа наредба ни враќа низа од сите работни места, притоа не прикажувајќи ги

дупликатите.

Следно, сакам да дознаам за секое работно место колку записи има во ова податочно множество.

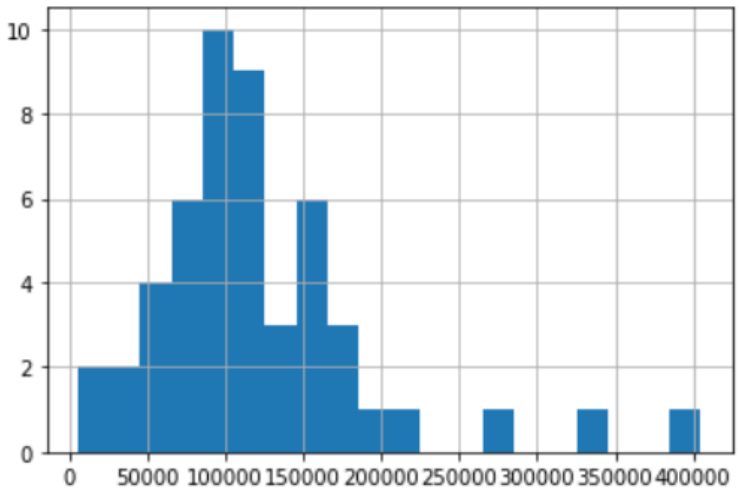
```
In [4]: df.groupby('job_title')['job_title'].count()
Out[4]:
job_title
3D Computer Vision Researcher                1
AI Scientist                                 7
Analytics Engineer                           4
Applied Data Scientist                       5
Applied Machine Learning Scientist           4
BI Data Analyst                              6
Big Data Architect                           1
Big Data Engineer                            8
Business Data Analyst                        5
Cloud Data Engineer                          2
Computer Vision Engineer                     6
Computer Vision Software Engineer            3
Data Analyst                                97
Data Analytics Engineer                      4
Data Analytics Lead                          1
Data Analytics Manager                       7
Data Architect                              11
Data Engineer                              132
Data Engineering Manager                     5
Data Science Consultant                      7
Data Science Engineer                        3
Data Science Manager                        12
Data Scientist                             143
Data Specialist                              1
Director of Data Engineering                 2
Director of Data Science                     7
ETL Developer                                2
Finance Data Analyst                         1
Financial Data Analyst                       2
Head of Data                                 5
Head of Data Science                         4
Head of Machine Learning                     1
Lead Data Analyst                            3
Lead Data Engineer                           6
Lead Data Scientist                          3
Lead Machine Learning Engineer               1
ML Engineer                                  6
Machine Learning Developer                   3
Machine Learning Engineer                   41
Machine Learning Infrastructure Engineer     3
Machine Learning Manager                     1
Machine Learning Scientist                   8
Marketing Data Analyst                       1
NLP Engineer                                 1
Principal Data Analyst                       2
Principal Data Engineer                      3
Principal Data Scientist                     7
Product Data Analyst                         2
Research Scientist                          16
Staff Data Scientist                         1
Name: job_title, dtype: int64
```

Со помош на groupby(‘job_title’) ги групирам според job_title и соодветно спрема job_title бројам колку се со иста работна позиција.

На сликата поделе може да ја забележиме средната исплатена плата во USD за секое работно место.

```
In [7]: df.groupby('job_title')['salary_in_usd'].mean()
Out[7]:
job_title
3D Computer Vision Researcher                  5409.000000
AI Scientist                                  66135.571429
Analytics Engineer                           175000.000000
Applied Data Scientist                       175655.000000
Applied Machine Learning Scientist           142068.750000
BI Data Analyst                               74755.166667
Big Data Architect                            99703.000000
Big Data Engineer                             51974.000000
Business Data Analyst                         76691.200000
Cloud Data Engineer                          124647.000000
Computer Vision Engineer                      44419.333333
Computer Vision Software Engineer            105248.666667
Data Analyst                                  92893.061856
Data Analytics Engineer                       64799.250000
Data Analytics Lead                          405000.000000
Data Analytics Manager                       127134.285714
Data Architect                               177873.909091
Data Engineer                                112725.000000
Data Engineering Manager                     123227.200000
Data Science Consultant                       69420.714286
Data Science Engineer                         75803.333333
Data Science Manager                         158328.500000
Data Scientist                               108187.832168
Data Specialist                              165000.000000
Director of Data Engineering                 156738.000000
Director of Data Science                     195074.000000
ETL Developer                                 54957.000000
Finance Data Analyst                          61896.000000
Financial Data Analyst                       275000.000000
Head of Data                                 160162.600000
Head of Data Science                         146718.750000
Head of Machine Learning                      79039.000000
Lead Data Analyst                             92203.000000
Lead Data Engineer                           139724.500000
Lead Data Scientist                          115190.000000
Lead Machine Learning Engineer                87932.000000
ML Engineer                                  117504.000000
Machine Learning Developer                    85860.666667
Machine Learning Engineer                    104880.146341
Machine Learning Infrastructure Engineer     101145.000000
Machine Learning Manager                     117104.000000
Machine Learning Scientist                   158412.500000
Marketing Data Analyst                        88654.000000
NLP Engineer                                  37236.000000
Principal Data Analyst                       122500.000000
Principal Data Engineer                      328333.333333
Principal Data Scientist                     215242.428571
Product Data Analyst                          13036.000000
Research Scientist                           109019.500000
Staff Data Scientist                         105000.000000
Name: salary_in_usd, dtype: float64
```

Доколку ова го претставам на хистограм, ќе забележиме дека најголем број од средните исплатени плати се помеѓу 90 000 и 120 000. Овој хистограм го добиваме со наредбата df.groupby('*job_title*')['*salary_in_usd*'].mean().hist(bins=20)



Разликата помеѓу најголемата и најмалата исплатена плата, во одредена работна позиција може да ја прочитаме од следната листа. Може да приметиме дека во некои оддели имаме огромна разлика.

```
In [14]: df.groupby('job_title')['salary_in_usd'].max()-
df.groupby('job_title')['salary_in_usd'].min()
Out[14]:
job_title
3D Computer Vision Researcher                  0
AI Scientist                              188000
Analytics Engineer                         70300
Applied Data Scientist                    325762
Applied Machine Learning Scientist        391125
BI Data Analyst                           140728
Big Data Architect                             0
Big Data Engineer                         108165
Business Data Analyst                     116558
Cloud Data Engineer                        70706
Computer Vision Engineer                  115000
Computer Vision Software Engineer          80000
Data Analyst                              193928
Data Analytics Engineer                    90000
Data Analytics Lead                            0
Data Analytics Manager                     44860
Data Architect                            175700
Data Engineer                             320000
Data Engineering Manager                  114697
Data Science Consultant                    97293
Data Science Engineer                      87032
Data Science Manager                      186906
Data Scientist                            409141
Data Specialist                                0
Director of Data Engineering               86524
Director of Data Science                  194974
ETL Developer                                  0
Finance Data Analyst                           0
Financial Data Analyst                    350000
Head of Data                              202026
Head of Data Science                      139000
Head of Machine Learning                       0
Lead Data Analyst                         150391
Lead Data Engineer                        220000
Lead Data Scientist                       149430
Lead Machine Learning Engineer                 0
ML Engineer                               254034
Machine Learning Developer                 21209
Machine Learning Engineer                 230000
Machine Learning Infrastructure Engineer  144820
Machine Learning Manager                       0
Machine Learning Scientist                248000
Marketing Data Analyst                         0
NLP Engineer                                   0
Principal Data Analyst                     95000
Principal Data Engineer                   415000
Principal Data Scientist                  267739
Product Data Analyst                       13928
Research Scientist                        408000
Staff Data Scientist                           0
Name: salary_in_usd, dtype: int64
```

Според податоците направени со следната наредба, воочуваме дека најголем број на компании, 326, се средни, односно со вработени помеѓу 50 и 250. Потоа имаме 198 големи компании, кои имаат вработени над 250, а најмал број на компании се мали со вработени под 50. Такви се 83 компании.

```
In [15]: df.groupby('company_size')['company_size'].count()
Out[15]:
company_size
L    198
M    326
S     83
Name: company_size, dtype: int64
```

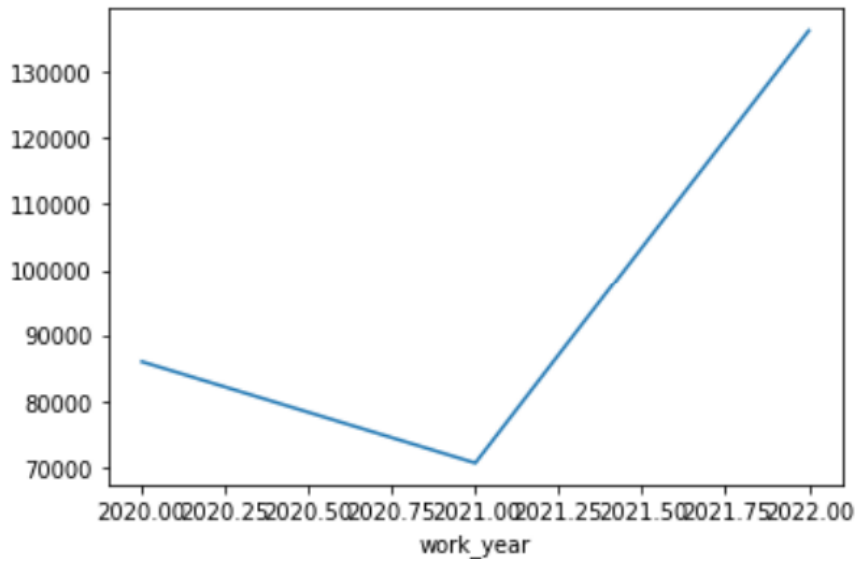
Доколку направиме мала статистика на исплатената плата со помош на describe(), забележуваме дека од 607 записи, средната исплатена плата е 112 297USD. Најниско исплатената плата е 2 859USD, додека највисоко исплатената плата е 600 000USD.

```
In [17]: df['salary_in_usd'].describe()
Out[17]:
count       607.000000
mean     112297.869852
std       70957.259411
min        2859.000000
25%       62726.000000
50%      101570.000000
75%      150000.000000
max      600000.000000
Name: salary_in_usd, dtype: float64
```

Исто така разгледувајќи го ова множество воочив дека најголем број од вработените се SE – Senior-level, потоа се MI – Mid-level, EN - Entry-level, EX – Executive-level.

```
In [18]: df.groupby('experience_level')['experience_level'].count()
Out[18]:
experience_level
EN     88
EX     26
MI    213
SE    280
Name: experience_level, dtype: int64
```

Конкретно за работната позиција Data Scientist од спроведената анализа за средна вредност во секоја од годините 2020, 2021 и 2022, може да се забележи дека во 2021 е најниска средната исплатена плата, додека во 2022 таа е драстично поголема од 2020 и 2021.



4. Користена литература

- Data Science Job Salaries | Kaggle
- Getting started tutorials — pandas 1.4.4 documentation (pydata.org)