Универзитет „Св. Кирил и Методиј“ во Скопје
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

ДИПЛОМСКА РАБОТА

Анализа на перформанси на NoSQL бази на податоци со користење на податоци за успех на ученици

Ментор:

Проф. Д-р. Слободан Калајџиски

Студент:

Филип Ставров, 183054

Скопје, 2022

Содржина

Апстракт

Во последните неколку години, сведоци сме на брзиот раст и развој на Големите податоци (Big Data), кои значително го променија начинот на складирање и толкување на податоците. Релационите бази на податоци не успеваат да се соочат со непредвидливоста и разновидноста на овие „големи податоци“, па поради тоа на сцена настапуваат Неструктурираните бази на податоци, кои се фокусираат на врските помеѓу податоците, а не на нивната структурираност и форма. Од друга страна, големиот квантитет на податоци, уште повеќе го зголемува влијанието и го става акцентот на нивниот квалитет. Моќните механизми и визуелизации на процесот на претпроцесирање на податоци ни овозможуваат да навлеземе длабоко во природата на атрибутите, како и нивните меѓусебни зависности, а со тоа да извлечеме знаење за истите и да го подобриме нивниот квалитет.

Во дипломската работа ќе се направи перформансна анализа на NoSQL базите на податоци со користење на податочно множество за успех на ученици од две средни училишта. Акцент ќе се стави на процесот на претпроцесирање на податоците, преку користење на најразлични техники и алгоритми за нормализација, скалирање и визуелизација на податоците, како и определување на меѓузависности во податочните атрибути. По претпроцесирањето, податочното множество ќе се прилагоди за импортирање во две NoSQL бази на податоци. Ќе се дефинираат повеќе прашалници преку кои ќе се мерат перформансите, па со нивна имплементација, ќе се добијат соодветните времиња на извршување. На крај ќе следи дискусија и заклучок на дипломската работа.

Клучни зборови: податоци, неструктутирани, претпроцесирање, перформанси, ученици, бази, успех

Вовед

Неструктурираните бази на податоци земаат се поголем замав и се стекнуваат со се поголема популарност. Корисничкото искуство полека, но сигурно станува една од најважните компетитивни предности во денешната индустрија. Во коренот на корисничкото искуство лежат податоците достапни на интернет, кои се појавуваат во сè понеструктурирани или целосно неструктурирани формати, како резултат на брзорастечката популарност на Big Data и Internet of Things. Големите податоци секако носат и свои предизвици, како што се огромното количество на податоци, големата комплексност и непредвидливост. Како главни карактеристики на Big Data ги имаме поголемата развновидност, зголемениот волумен на податоци и брзина на генерирање, кои се клучни во решавањето и адресирањето на нови бизнис проблеми и идеи, кои до пред некое време беа незамисливи. Овие полуструктурирани и целосно нестуктурирани податоци претставуваат предизвик за традиционалните системи за менаџирање на релациони бази на податоци, па решението се наоѓа токму во неструктурираните бази на податоци, кон кои ќе го насочиме нашето внимание.

Проблемот со релационите бази на податоци лежи во нивната имплементација на ACID принципот, кој носи свои ограничувања, кои не можат да се справат со огромните количини и новите формати на податоци. Поради тоа, фокусот е ставен на неструктурираните бази на податоци, познати под името NoSQL, што значи “Not only SQL”. Наместо табели со редици и колони, неструктурираните бази на податоци ги складираат информациите во Json документи, а притоа разликуваме неколку типови бази на податоци: document-based, key-value, column-family и граф бази на податоци.

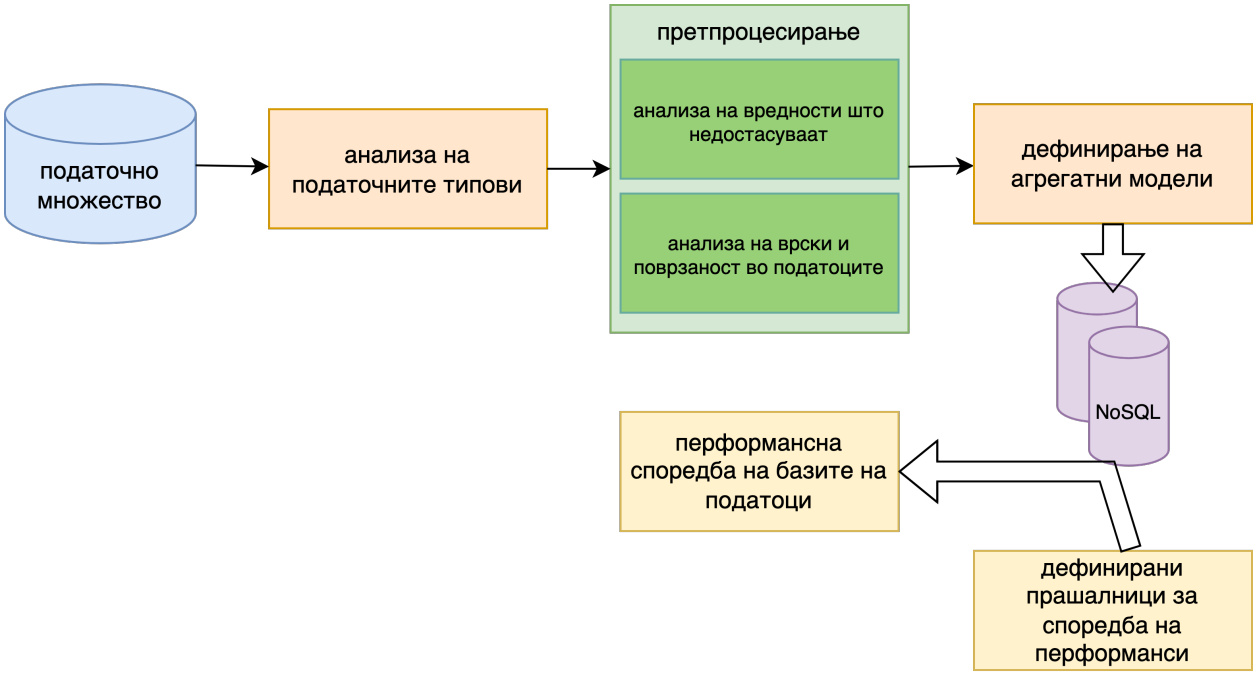
Во фокусот на истражувањето во оваа дипломска работа се документ базираните базите на податоци. Користејќи податочно множество кое се однесува на факторите кои играат улога врз перформансите на учениците во училиште, ќе се направи компаративна анализа на два претставници од документ базираните бази на податоци, MongoDB и CouchDB.

Избраното податочно множеството се однесува на перформансите на учениците во средни училишта, како и на факторите кои влијаат врз истите. Секој од нас, се соочувал со различни препреки и предизвици во текот на своето образование, па сметав дека ќе е интересно да се направи подлабока анализа и да се истражат врските кои, иако навидум незначајни, во комбинација и заедничка корелација се клучни за успехот и иднината на младите.

Со цел успешна анализа на податочното множесто, како и успешна компарација на двете бази на податоци, потребно е најпрво да добиеме добар увид во врските и поврзаностите кои постојат во самото податочно множество. Овде, на сцена доаѓа процесот на претпроцесирање, како фаза која ни овозможува да навлеземе подлабоко во корелациите помеѓу атрибутите на множеството и ни дава детален преглед на сите значајни фактори врз перформансите на учениците. Големиот број на хистограми, дијаграми на корелации и графикони ни овозможуваат да добиеме одлична претстава за структурата на податоците, зголемувајќи го нивниот квалитет и нудејќи ни подлабоко и помоќно толкување на сите релации.

Методологија

За почеток, да го погледнеме дијаграмот на слика број еден, кој ќе ни даде детален приказ на чекорите кои ќе ги преземеме низ текот на оваа анализа во факторите кои влијаат врз перформансите на учениците.



Слика 1. Методологија за анализа на податоците и перформансна анализа на базите на податоци

Најпрво ќе започнеме со опис и анализа на податочното множество, со цел да се запознаеме со темата од интерес и да добиеме одредено почетно знаење за истата.

Понатаму, со помош на Python и неговите библиотеки ќе направиме вчитување на множеството и анализа на податочните типови на секој од атрибутите. Фокусот ќе го ставиме на нивните вредности, нивните рангови, ограничувања, како и нивното значење во однос на перформансите на учениците.

Следно, имајќи во предвид дека вредностите кои недостасуваат, како и невалидните вредности се честопати виновници за неточни предвидувања и резултати, ќе пробаме да ги отстраниме истите, како и да ја пронајдеме причината за нивната невалидност.

Пред да преминеме кон имплементација на базите и податоците, ќе навлеземе длабоко во врските, меѓузависностите и релациите кои постојат помеѓу податоците. Вниманието ќе го насочиме кон зависностите помеѓу атрибутите и нивната корелација, со цел да добиеме

добра претстава за самото податочно множество, како и за факторите кои најмногу влијаат врз успехот на учениците.

Понатаму, ќе преминеме кон креирање и поврзување на соодветните бази на податоци и импортирање на податоците. Базите кои ќе бидат од наш интерес во тековната анализа се документ базираните бази на податоци MongoDB и CouchDB. Ќе ги погледнеме моделите за импортирање на податоците во соодветната база, како и предностите и недостатоците на истите. И двете бази, MongoDB и CouchDB, овозможуват креирање и надградување на програми без потреба да следат некоја главна шема. Менаџирање со содржини и справување со податоци во мобилни апликации се две од полињата каде што употреба на ваков тип на бази на податоци е соодветна.

Понатаму, акцентот ќе го ставиме на моделите на агрегација кои ги нуди секоја од базите на податоци. Ќе погледнеме прашалници, кои ги истражуваат врските кои веќе сме ги забележале во процесот на претпроцесирање и ќе се насочиме кон издвојување на факторите кои се од клучно значење за успехот на учениците.

На крај, следува компаративна анализа на перформансите и резултатите на двете бази на податоци, посветувајќи го вниманието на нивната брзина, едноставност за користење и интиуитивност.

Опис на податочното множество

Податочното множество кое ни е од интерес, како и врз кое ќе правиме споредба на перформансите на двете неструктурирани бази на податоци се однесува на успехот на учениците во средно училиште.

Средното училиште е период од животот на човекот, кој сам по себе е доста стресен и предизвикувачки. Притоа, самата личност е доста сензитивна во овој период, а воедно и лесно подлежи на надворешни влијанија, фактори и искушенија.

Ова истражување ќе ни овозможи да навлеземе во суштината на самите перформанси на учениците и да добиеме претстава за клучните фактори и проблеми со кои тие се соочуваат, а воедно и со кои секој од нас се соочил.

Истражувањето е направено во две средни училишта во Португалија, а притоа тестирањата се правени врз оценките од два предмети, математика и португалски јазик. Земени се во предвид полот на учениците, нивните воншколски активности, желбата за виско образование, средината на живеење, времето на патување, како и времето на учење.

Имајќи го во обзир и влијанието кое родителите го имаат врз своите деца, акцентот е ставен и на образованието на родителите, нивните професии, како и односите во семејството, кои во големо влијаат врз самото дете, а со тоа и врз неговата мотивација за учење и неговите резултати.

Вредностите, ограничувањата и разновидностите, како и толкувањата на секој од атрибутите кои претставуваат составен дел од ова податочно множество ќе ги погледнеме во продолжение.

Анализа на податочните типови и вредности на секој од атрибутите

Прв чекор од ваквата анализа е вчитување на множеството, користејќи го pandas модулот во Python. Ова ни овозможува да направиме увид во структурата на множеството и да добиеме претстава за карактеристиките со кои работиме. Бидејќи самото множество беше поделено во два документи, едниот за предметот математика, а другиот документ за предметот португалски јазик, најпрво се вчитаа двете множества. Следно, додадена беше нова колона кон множествата, именувана “subject”, со цел при спојување на податочните множества во едно, да не ја изгубиме информацијата за кој предмет станува збор. Дел од множеството е прикажан на сликата 2.

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3	subject
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	3	4	1	1	3	6	5	6	6	math
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	3	3	1	1	3	4	5	5	6	math
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	3	2	2	3	3	10	7	8	10	math
3	GP	F	15	U	GT3	T	4	2	health	services	...	2	2	1	1	5	2	15	14	15	math
4	GP	F	16	U	GT3	T	3	3	other	other	...	3	2	1	2	5	4	6	10	10	math
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1039	MS	F	19	R	GT3	T	2	3	services	other	...	4	2	1	2	5	4	10	11	10	portuguese
1040	MS	F	18	U	LE3	T	3	1	teacher	services	...	3	4	1	1	1	4	15	15	16	portuguese
1041	MS	F	18	U	GT3	T	1	1	other	other	...	1	1	1	1	5	6	11	12	9	portuguese
1042	MS	M	17	U	LE3	T	3	1	services	services	...	4	5	3	4	2	6	10	10	10	portuguese
1043	MS	M	18	R	LE3	T	3	2	services	other	...	4	1	3	4	5	4	10	11	11	portuguese

1044 rows × 34 columns

Слика 2. “Дел од содржината на податочното множество”

Следно, направен е осврт кон податочните типови на атрибутите, со цел да се увиди дали дел од нив треба да подлежат на одредени конверзии. На слика 3, може да ги видиме атрибутите и нивните податочни типови.

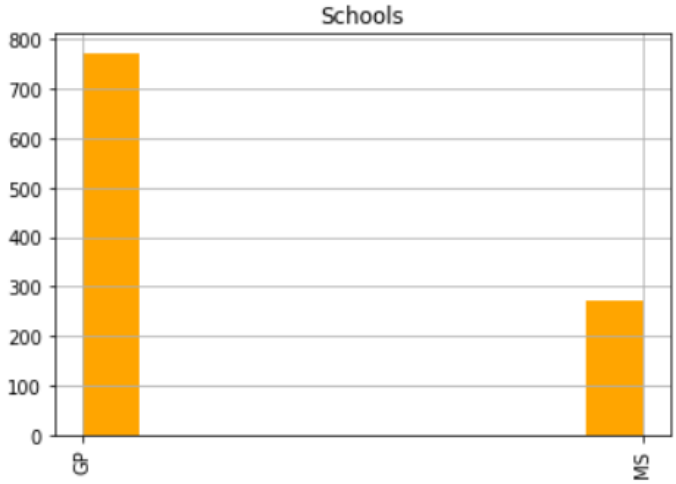
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1044 entries, 0 to 1043
Data columns (total 34 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      1044 non-null   object
 1   sex         1044 non-null   object
 2   age         1044 non-null   int64
 3   address     1044 non-null   object
 4   famsize     1044 non-null   object
 5   Pstatus     1044 non-null   object
 6   Medu        1044 non-null   int64
 7   Fedu        1044 non-null   int64
 8   Mjob        1044 non-null   object
 9   Fjob        1044 non-null   object
 10  reason      1044 non-null   object
 11  guardian    1044 non-null   object
 12  traveltime  1044 non-null   int64
 13  studytime   1044 non-null   int64
 14  failures    1044 non-null   int64
 15  schoolsup   1044 non-null   object
 16  famsup      1044 non-null   object
 17  paid        1044 non-null   object
 18  activities  1044 non-null   object
 19  nursery     1044 non-null   object
 20  higher      1044 non-null   object
 21  internet    1044 non-null   object
 22  romantic    1044 non-null   object
 23  famrel      1044 non-null   int64
 24  freetime    1044 non-null   int64
 25  goout       1044 non-null   int64
 26  Dalc        1044 non-null   int64
 27  Walc        1044 non-null   int64
 28  health      1044 non-null   int64
 29  absences    1044 non-null   int64
 30  G1          1044 non-null   int64
 31  G2          1044 non-null   int64
 32  G3          1044 non-null   int64
 33  subject     1044 non-null   object
dtypes: int64(16), object(18)
memory usage: 277.4+ KB
```

Слика 3. "Приказ на податочните типови на секој од атрибутите"

Сликата 3 дава до знаење дека секој од атрибутите кој има нумеричка вредност, е класифициран како integer, додека пак атрибутите кои имаат текстуална вредност, се претставени како објекти. Ваквата класификација и доделување на податочните типови е соодветна, па поради тоа самото множество нема потреба да подлежи на никакви конверзии.

Следен чекор е детална анализа на секој од атрибутите. Овде ќе се направи детално истражување на вредностите на секој од атрибутите, а притоа да се земе во предвид и нивната распределба.

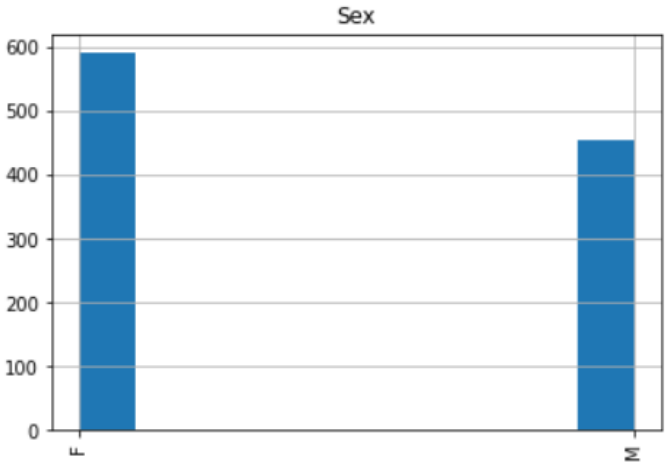
Да започнеме со ред. Најпрво да ја погледнеме колоната “school”.



Слика 4. “Распределба вредностите на колоната school”

Како што спомнавме и претходно, а воедно и како што може да видеме на сликата 4, истражувањето за перформансите на учениците е направено во две различни училишта. Едното од нив е Gabriel Pereira, означено со GP, а второто е Mousinho da Silveira, означено како MS. Самата визуелизација ни овозможува да заклчучиме дека поголем дел од учениците се од училиштето Gabriel Pereira, додека пак од училиштето Mousinho de Silveira имаме трипати помалку ученици.

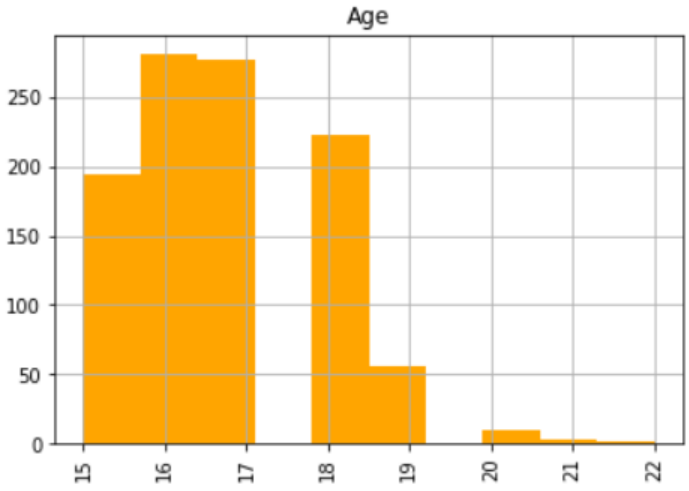
Во продолжение ќе направиме анализа и на колоната која се однесува на полот на испитаните ученици.



Слика 5. “Распределба на вредностите на колоната која се однесува на полот на учениците”

На хистограмот прикажан на слика 5, може да забележиме дека за малку преовладуваат ученици од женски пол.

Понатаму, се истражува колоната за возраст, а хистограмот за истата е видлив на слика 6.



Слика 6. "Распределба на вредностите на колоната за возраст"

Како што е видливо од самиот хистограм, годините на учениците се движат во рангот од 15, до 22 години. Притоа, преовладуваат ученици кои имаат 16 и 17 години, а има многу мал број на ученици кои имаат 21 или 22 години. Ова се потврди и со испитување на средната вредност која има вредност од 16.72, како и вредноста за стандардната девијација која изнесува 1.23, чие толкување е дека во голема мера вредностите за колоната возраст се движат околу самата средна вредност.

Средната вредност, како и стандардната девијација се видливи на сликата во продолжение.

```
data['age'].mean()
```

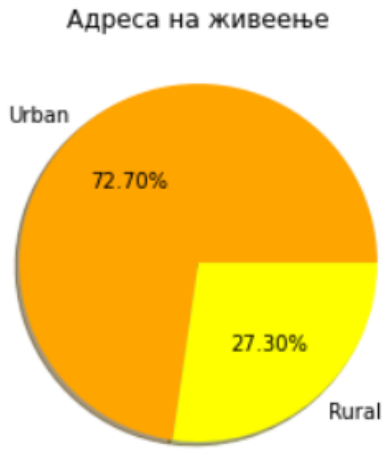
```
16.726053639846743
```

```
data['age'].std()
```

```
1.239974693164953
```

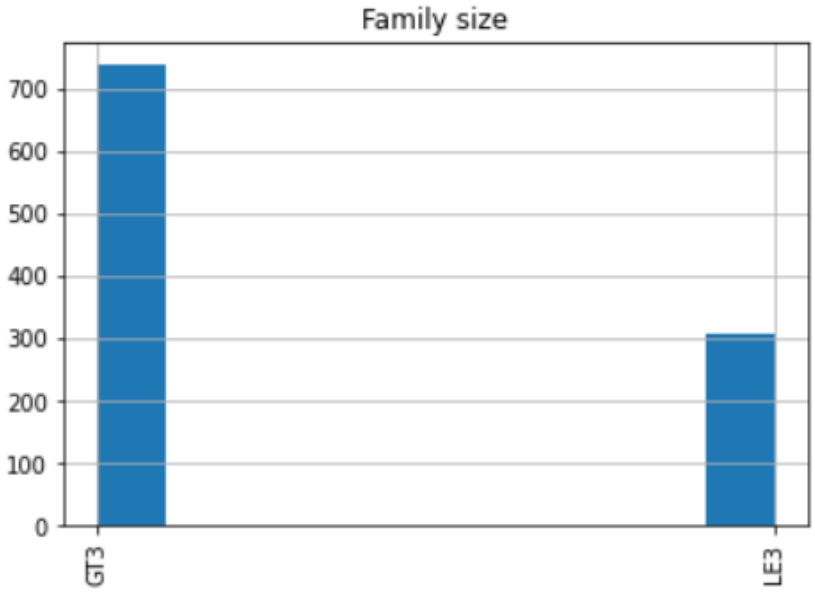
Слика 7. "Анализа на колоната за возраст"

Следно, се продолжи со истражување на вредностите на колоната која се однесува на адресата на учениците. Оваа колона разликува две вредности, односно U за урбана средина и R за рурални средини. Графиконот на сликата 8 ни укажува дека 72.70% од учениците доаѓаат од урабни средини, додека пак останатите 27.30% од рурални.



Слика 8. "Распределба на колоната која се однесува на адресата на живеење"

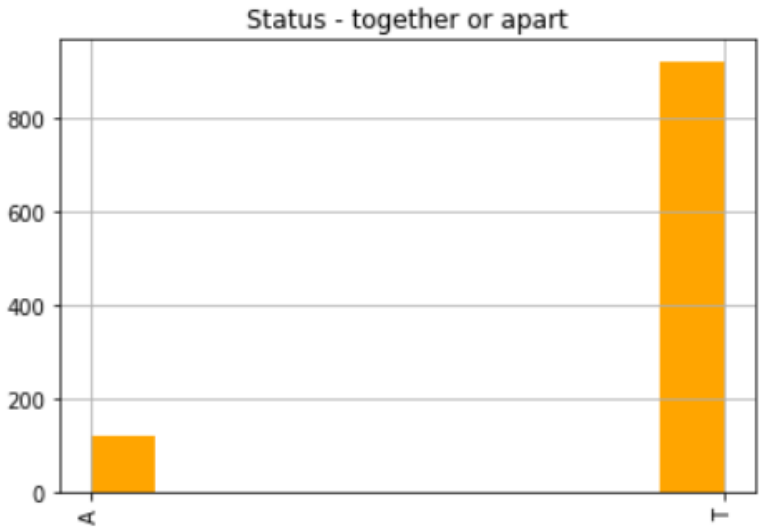
Понатаму, се премина на истражување на вредностите на атрибутот "famsize", кој фокусот го става на големината на семејството. Подетални информации можеме да добиеме од хистограмот прикажан на слика 9.



Слика 9. "Анализа на колоната за големина на семејството"

Од самиот хистограм можеме да забележиме дека разликуваме две вредности за оваа колона. Вредноста GT3 ни укажува на семејство поголемо од три членови, додека пак вредноста LE3 ни означува семејство помало или еднакво на три членови. Двојно помалку се ученици чијашто големина на семејство е помала или еднаква на три членови.

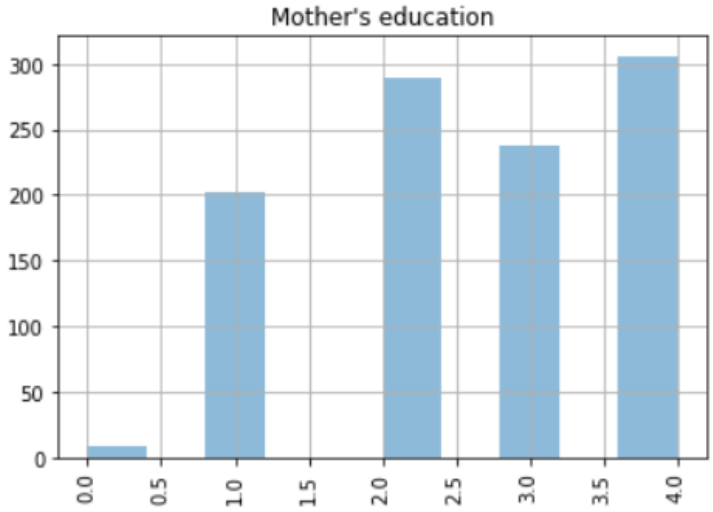
Следна колона од интерес е колоната означена со “Pstatus”. Оваа колона го истражува статусот на коегзистирање на родителите, односно дали тие живеат заедно, “T”, или пак посебно, т.е. “A”.



Слика 10. “Распределба на вредностите на колоната Pstatus”

Хистограмот на слика 10 ни покажува дека 90% од учениците имаат родители кои живеат заедно.

На хистограмите и графиконите во продолжение ќе ги погледнеме вредностите и распределбата на атрибутите кои се однесуваат на степенот на образование на мајката и степенот на образование на таткото.

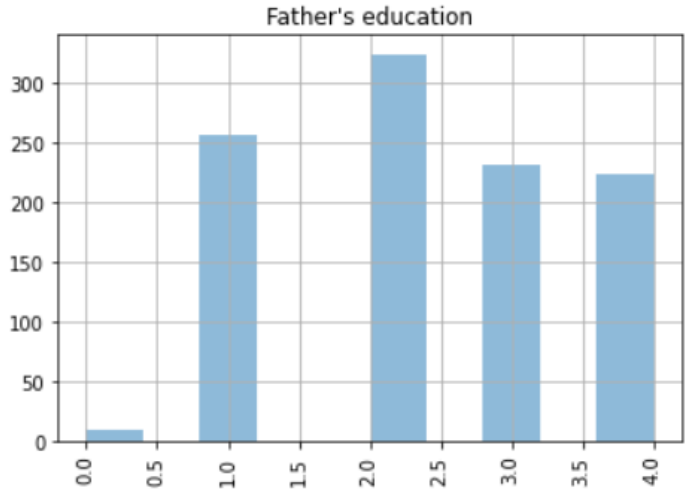


Слика 11. "Распределба на вредностите за колоната која се однесува на образование на мајката"

Слика 12. "Пита графикон – образование на мајка"

Вредностите кои се достапни за степенот на образование, независно дали станува за мајка или татко се во границите од 0, до 4. Притоа со 0 е ознаено дека нема образование, со 1, дека има основно образование до 4то одделение, со 2 е означено завршено основно образование, со 3 завршено средно образование, а со 4 завршено високо образование.

Конкретно, кај најголем дел од учениците, мајката има завршено високо образование. Мал е процентот на ученици, каде мајката е без образование, или со 4то одделение. Да погледнеме како стои ситуацијата со татковците на учениците.



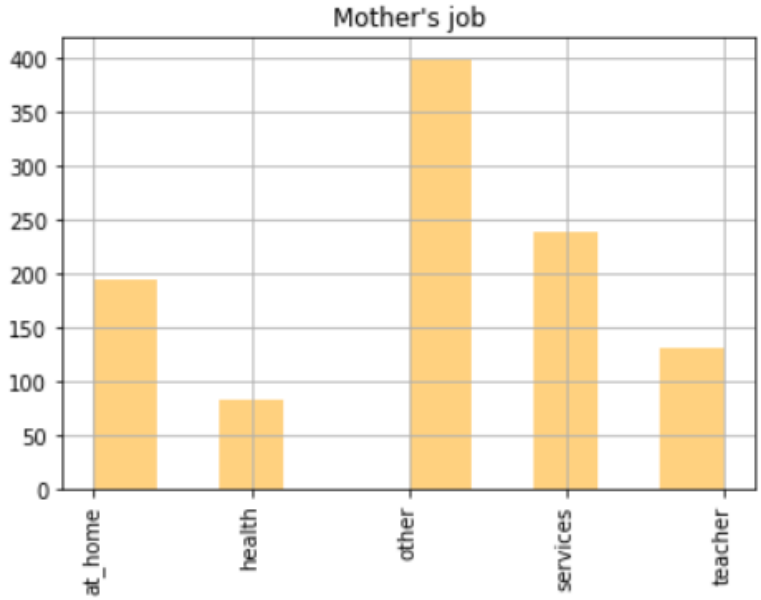
Слика 13. "Распределба на вредностите на колоната која се однесува на образованието на таткото"

Слика 14. "Пита графикон - образование на татко"

Од пита графиконот на слика 14, како и од хистограмот на слика 13, можеме да забележиме дека ситуацијата со степенот на образование на таткото на учениците е доста поразлична од онаа кај мајката. Во главно, најголем дел од учениците имаат татковци со завршено основно училиште, или со 4то одделение. Минимален е бројот на ученици со татко кој е без образование, а воедно не е многу голем и процентот на ученици каде таткото има завршено високо образование.

Доколку направиме споредба на овие два атрибути, можеме да забележиме дека најчесто мајката има повисок степен на образование од таткото, додека пак и во двата случаи минимални се поединци кои се без образование.

Следно, ќе ја погледнеме колоната која се однесува на работниот статус, односно професијата на мајката на ученикот. Повеќе информации можеме да добиеме од хистограмот прикажан на сликата 15.

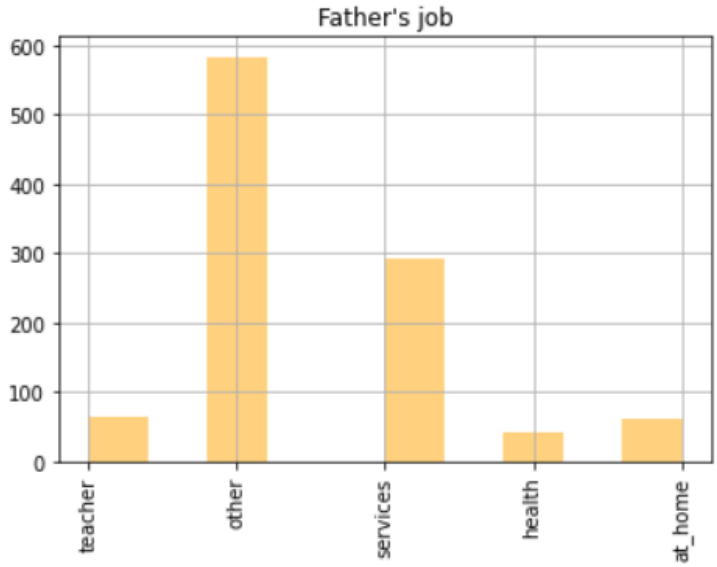


Слика 15. "Распределба на вредности – професија на мајка"

Како што можеме да видиме на хистограмот на слика 15, постојат пет категории кога станува збор за работата на мајката, а тоа се: не работи, односно stay at home mom, потоа работи во здравство, работи како професор, работи нешто друго, или пак работа во јавниот државен сектор.

Притоа, најголем дел од мајките на учениците работат нешто друго, односно категорија на работа која не е издвоена овде, но исто така голем дел од нив работат и во јавниот државен сектор.

Понатаму, можеме да продолжиме со разгледување на работниот статус на таткото. Овде ги имаме истите категории на професии, а нивната распределба можеме да ја погледнеме на хистограмот прикажан на слика 16.



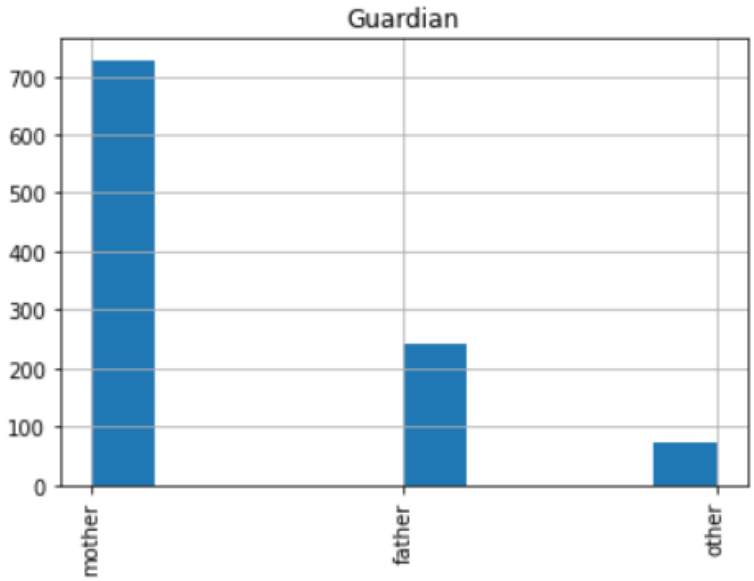
Слика 16. "Распределба на вредностите – професија на татко"

Следна колона која ќе ја истражиме е колоната која се однесува на причината која е виновник за изборот на училиштето. Графиконот за оваа колона е достапен на сликата во продолжение.

Слика 17. "Пита графикон – причина за избор на училиште"

Судејќи по самиот пита графикон, можеме да заклучиме дека најголем дел од учениците го избрале соодветното училиште поради програмите и предметите кое истото ги нуди. Сепак, значаен фактор се и близината на училиштето, како и неговата репутација.

Понатаму, да ја разгледаме колоната која го определува старателот на ученикот. Како достапни вредности ги имаме мајка, татко и останато, односно трето лице. Распределбата на овој атрибут е прикажана на слика 18.



Слика 18. "Распределба на вредности – старател на ученикот"

Од самиот хистограм можеме да забележиме дека кај најголем дел од учениците, над 70%, како старател ја имаме мајката.

Следна карактерисика која ќе ја разгледаме е времето на патување. Оваа карактеристика е тесно поврзана со адресата и средината на живеење на ученикот (урбана или рурална), како и со причината за избор на училиштето, имајќи во предвид дека речиси 30% од учениците се одлучиле за конкретно училиште само поради тоа што е во близина до нивното место на живеење. На графиконот на слика 19, можеме да ја погледнеме распределбата на овој атрибут.

Слика 19. "Пира графикон – време на патување до училиште"

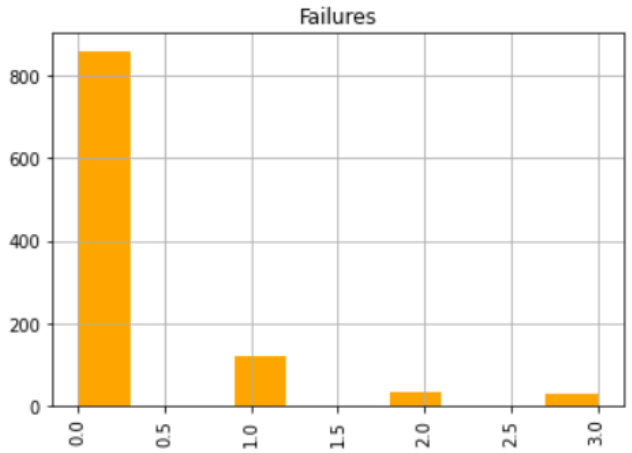
Од пита графиконот можеме да забележиме дека речиси 60% од учениците патуваат помалку од 15 минути до училиштето. Ова укажува на фактот дека близината на училиштето е значаен фактор при самиот избор. Помалку од 3% патуваат над еден час за да стигнат на училиште, а исто така мал е и процентот на ученици кои патуваат помеѓу половина и еден час – околу 8%.

Продолжуваме со разгледување на колоната посветена на времето на учење на неделно ниво. Врз овој атрибут имаат влијание повеќе фактори, како што се причината поради која е избрано училиштето, близината на училиштето, како и самата посветеност на ученикот. Подетални информации можеме да добиеме од пита графиконот кој се однесува на оваа карактеристика.

Слика 20. "Пита графикон – време посветено на учење во тек на една недела"

Од самиот графикон можеме да увидиме дека речиси 50% од учениците посветуваат од 2 до 5 часа на учење во неделата. Понатаму, околу 30% од учениците посветуваат помалку од 2 часа. 15% од учениците учат од 5 до 10 часа во текот на една недела, а само 6% од нив посветуваат над 10 часа во учење во текот на една недела. Времето на учење секако се одразува на крајниот успех на учениците.

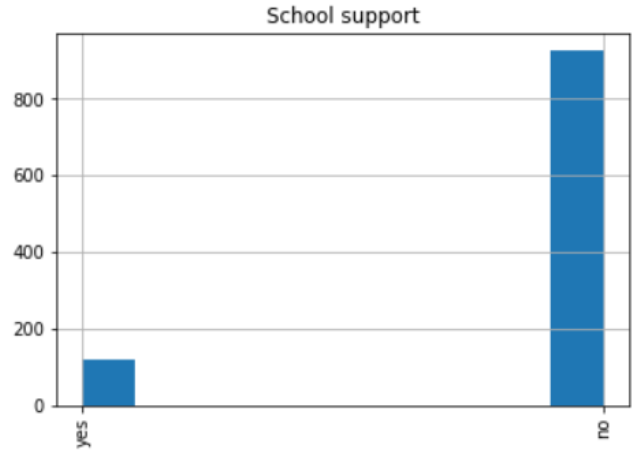
Следен атрибут кој ќе го истражиме е бројот на паѓања на одреден предмет. Хистограмот за овој атрибут е видлив на слика 21.



Слика 21. "Распределба на вредности – број на паѓања на предмет"

Хистограмот на слика 21 ни дава до знаење дека најголем дел од учениците, односно речиси 4/5 од нив немаат паднато предмет. Воедно, многу е мал бројот на ученици кои паднале даден предмет повеќе од еднаш, или двапати.

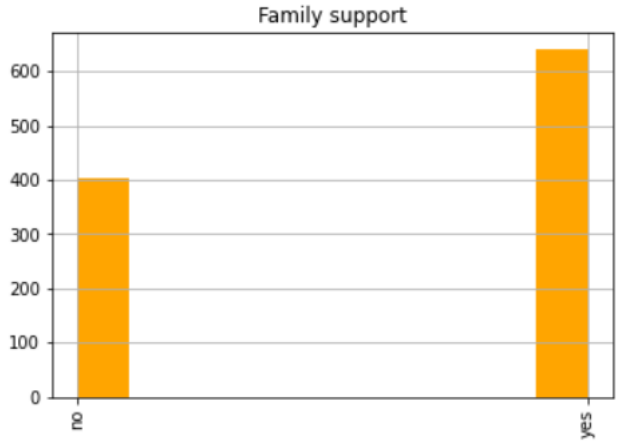
Понатаму, можеме да ја разгледаме колоната која се однесува на поддршка во образованието од страна на училиштето во вид на стипендија. Хистограмот во продолжение ќе ни покаже колкав процент од учениците земаат, а колкав процент од нив не земаат стипендија. Пресуден фактор за тоа дали ученикот ќе земе стипендија е неговиот успех, а најверојатно во предвид се зема и финансиската состојба на семејството.



Слика 22. "Распределба на вредности – земање на стипендија"

Хистограмот на слика 22 ни покажува дека само 10% од учениците земаат стипендија, односно околу 100тина ученици.

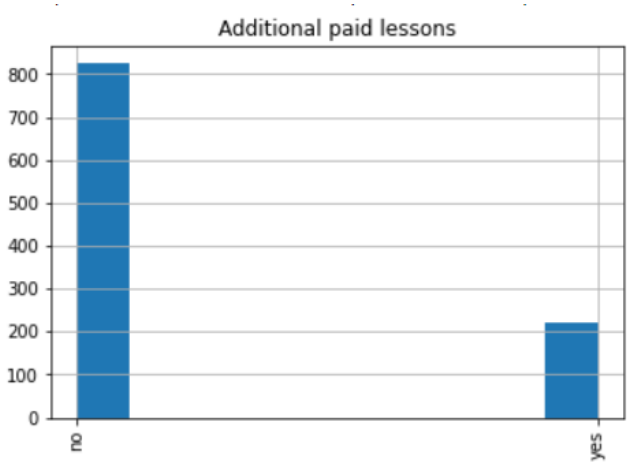
Во продолжение ќе ја погледнеме и карактеристиката која се однесува на семејната поддршка во образованието. Распределбата на истата е прикажана во продолжение.



Слика 23. "Распределба на вредности – семејна поддршка"

Хистограмот на слика 23 ни покажува дека 3/5 од учениците имаат семејна поддршка, додека 2/5 од нив немаат.

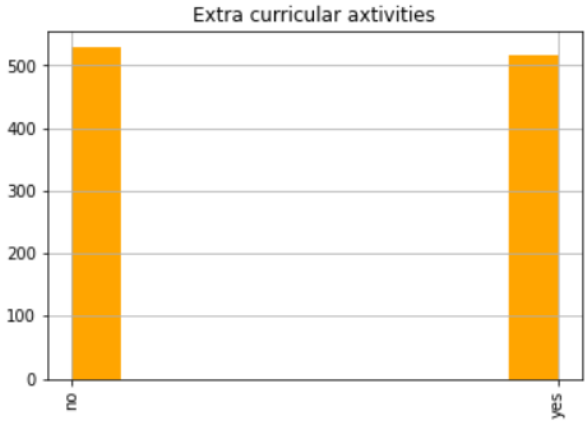
Следен атрибут кој ќе го разгледаме е дали ученикот платил дополнителни часови за еден од двата предмети, математика или португалски јазик. Вредноста за овој атрибут може да биде да, или не, а распределбата на вредностите е дадена на хистограмот во продолжение.



Слика 24. "Распределба на вредности – дополнителни платени часови"

Како што можеме да видиме од самиот хистограм, преовладуваат ученици кои не посетувале дополнителни часови. Процентуално, околу 80% од учениците немаат земано дополнителни платени часови, а останатите 20% имаат.

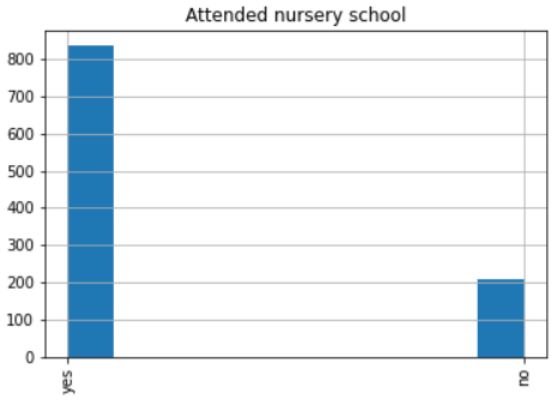
Следна колона која ќе ја истражуваме е колоната која се однесува на тоа дали учениците преземаат воншколски активности. И оваа колона има само две достапни вредности, а тоа се да и не, а нивната распределба можеме да ја погледнеме на хистограмот достапен на сликата 25.



Слика 25. "Распределба на вредности – воншколски активности"

Хистограмот прикажан погоре ни укажува на фактот дека вредностите за овој атрибут се еднакво распределени, односно, половина од учениците преземаат воншколски активности, а останатата половина не преземаат.

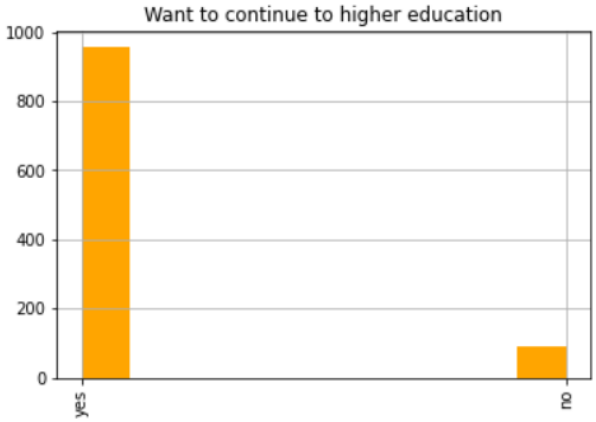
Продолжуваме со разгледување на следната колона, која се однесува на тоа дали учениците посетувале медицинско училиште. Распределбата на вредностите на оваа колона е дадена во продолжение.



Слика 26. "Распределба на вредности – посетувале медицинско училиште"

На хистограмот на слика 26 можеме да забележиме дека речиси 4/5 од учениците посетувале медицинско училиште.

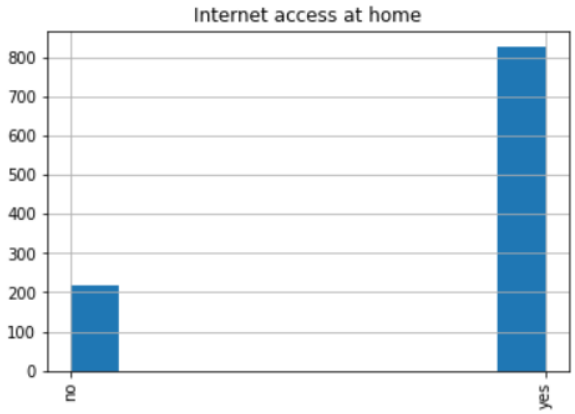
Следниот атрибут кој ќе го разгледаме се однесува на прашањето дали учениците планираат да продолжат со високо образование. Овој атрибут е од големо значење, бидејќи успехот од средно училиште е пресуден за тоа дали ученикот ќе биде примен на факултетот кој е од негов интерес. Распределбата на вредностите на овој атрибут е дадена во продолжение.



Слика 27. "Распределба на вредности – планира да продолжи со високо образование"

Хистограмот на сликата 27 покажува дека најголем дел од учениците, односно, речиси сите ученици планираат да го продолжат своето образование. Помалку од 100тина ученици одговориле со не, односно немаат желба да продолжат со високо образование.

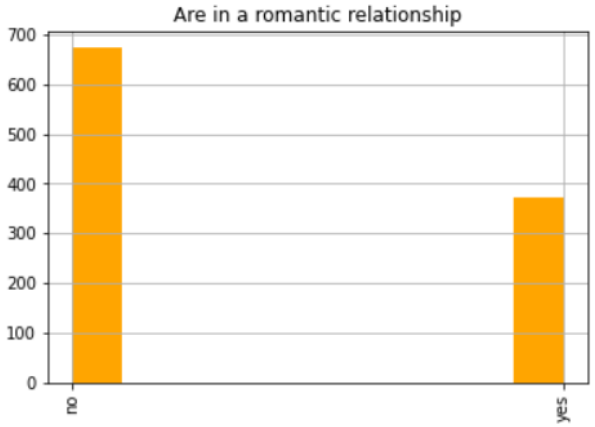
Имајќи го во предвид динамичниот живот, различните финансиски состојби, како и средини на живеење, интересна карактеристика за истражување е и достапноста на интеренет пристап од дома, односно, колку од учениците имаат, а колку од нив немаат интернет во своите домови.



Слика 28. "Распределба на вредности – пристап на интернет во домот"

Хистограмот на сликата 28 укажува на фактот дека 1/5 од учениците немаат интернет во своите домови. Во време кога најголем дел од материјалите се достапни онлајн и се електронски овој процент е зачудувачки голем, а секако недостапноста на интернет во рамки на домот придонесува и за помали резултати и перформанси во учењето. Меѓутоа, од друга страна за дел од учениците, немањето на интернет ослободува повеќе време за учење, па секако има и случаи каде овој фактор придонесува за повисоки резултати.

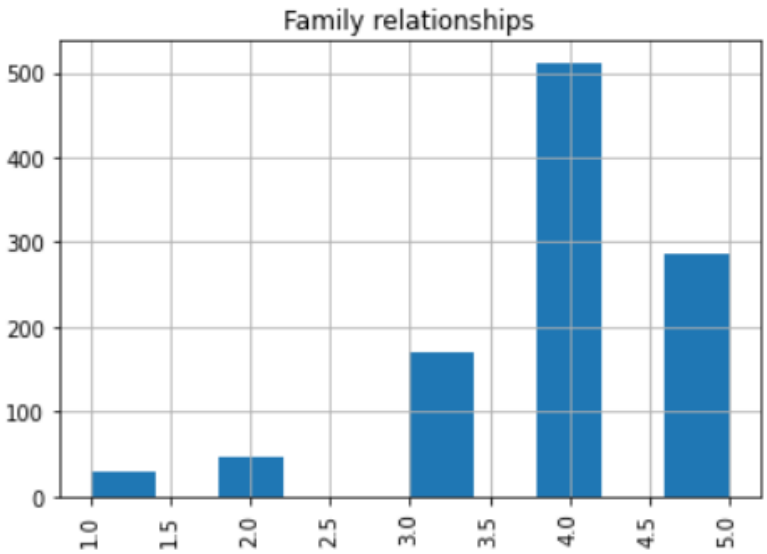
Знаејќи дека средношколските години се еден од најмалку предвидливите периоди во животот на луѓето, интересен атрибут за истражување е и тоа дали ученикот е во љубовна врска, или не, и дали и како истата влијае врз неговите/нејзините перформанси на училиште.



Слика 29. "Распределба на вредности – дали е во љубовна врска"

Од хистограмот на слика 29 можеме да заклучиме дека преовладуваат ученици кои не се во љубовна врска. Бројот на ученици кои се во љубовна врска е двапати помал од бројот на ученици кои не се во љубовна врска.

Знаеме дека перформансите на учениците зависат од многу фактори, а како еден од нив можеме да го издвоиме и односите во семејството. На хистограмот на слика 30 можеме да ги погледнеме достапните вредности за овој атрибут, како и нивната распределба.



Слика 30. "Распределба на вредности – односи во семејството"

Хистограмот на сликата 30 ни покажува дека односите во семејството се движат во интервалот од 1, до 5, каде вредноста 1 означува многу лоши односи во семејството, а вредноста 5 означува одлични односи во семејството. Најчеста вредност за односите во семејството е вредноста 4, која означува многу добри односи во семејството. Притоа, најчестата вредност е близу и до средната вредност од 3.93, а вредноста од 0.93 на стандардната девијација укажува на фактот дека најголем дел од вредностите за овој атрибут се движат околу самата средна вредност, што секако го потврдува и самиот хистограм.

Средната вредност, како и стандардната девијација за овој атрибут се достапни во продолжение.

```
data['famrel'].mean()
```

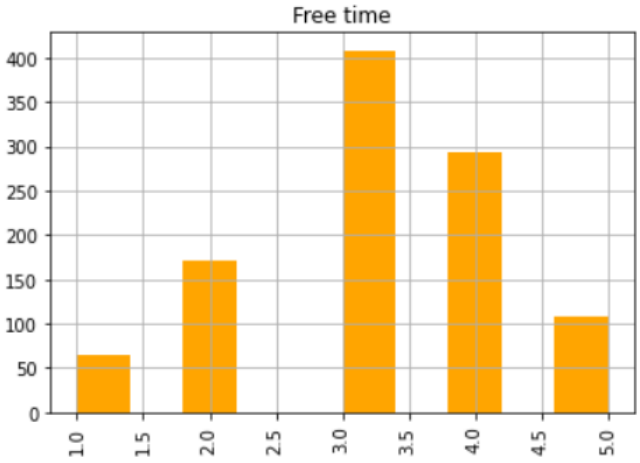
```
3.935823754789272
```

```
data['famrel'].std()
```

```
0.9334007716663543
```

Слика 31. "Средна вредност и стандардна девијација за атрибутот односи во семејството"

Следена карактеристика која ќе ја разгледаме е слободното време на учениците. Ќе го ставиме фокусот на рангот на вредностите на оваа карактеристика, како и нејзината распределба. Хистограмот за истата е прикажан на слика 32.



Слика 32. "Распределба на вредности – слободно време"

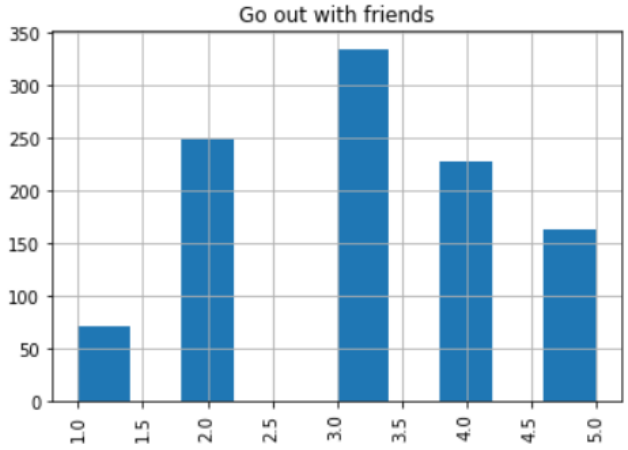
Со цел подобар приказ, како и толкување на вредностите за слободно време во рангот од 1 до 5, можеме да го погледнеме графиконот во продолжение.

Слика 33. "Пита графикон – слободно време после училиште"

Пита графиконот ни покажува дека имаме пет различни вредности за слободното време на учениците. Како најзастапена вредност имаме дека учениците имаат задоволително ниво на слободно време, а многу е мал бројот на ученици кои немаат, или пак имаат многу слободно време. Врз слободното време влијаат времето за патување до училиште, како и

времето посветено на учење, а пак овие карактеристики заедно во комбинација имаат значаен удел во крајните перформанси на учениците.

Следната карактеристика која ќе ја разгледаме е во голема мера поврзана и блиска до слободното време, а тоа е излегувањето со пријатели, кое варира од ниско ниво на излегување, до многу излегување со пријатели.



Слика 34. "Распределба на вредности – излегување со пријатели"

Како што спомнавме претходно, а и како што можеме да забележиме од самиот хистограм, излегувањето со пријатели е прикажано со нумерички вредности од 1, до 5, каде вредноста 1 значи малку излегување со пријатели, а вредноста 5 значи многу излегување со пријатели. Од хистограмот можеме да забележиме дека најголем дел од учениците имаат некое средно ниво на излегување со пријателите, а ова го потврдуваат и вредностите за средна вредност и стандардна девијација кои се дадени во продолжение.

```
data['goout'].mean()
```

```
3.1561302681992336
```

```
data['goout'].std()
```

```
1.152574660205394
```

Слика 35. "Средна вредност и стандардна девијација на атрубутот излегување со пријатели"

Од сликата 35 може да го потврдиме заклучокот дека најголем дел од учениците имаат некое средно ниво на излегување. Вредноста на стандардната девијација од 1.15 укажува дека најголем дел од вредностите за овој атрибут се фокусирани и се движат околу самата средна вредност.

Следно, ќе продолжиме со разгледување на два атрибути, кои се меѓусебно доста блиски и поврзани. Најпрво ќе го погледнеме атрибутот именуван како “Dalc”, кој се однесува на количеството на консумирање на алкохол во текот на работните денови, а потоа ќе го погледнеме и атрибутот со име “Walc”, кој се однесува на количеството на консумирање на алкохол во текот на викендот. Како што може да се забележи од хистограмите во продолжение, вредностите на овие атрибути се движат во рангот од 1 до 5, каде 1 означува многу мали количини на алкохол, а 5 означува високи количини на алкохол.

Слика 36. “Распределба на вредности – консумирање на алкохол во тек на работни денови”

Од хистограмот за атрибутот кој се однесува на истражување на консумирањето на алкохол во текот на работните денови, можеме да забележиме дека над 90% од учениците не практикуваат да консумираат алкохол во текот на работната недела. Многу мал број од учениците консумираат алкохол во поголеми количини во текот на работните денови.

Слика 37. “Распределба на вредности – консумирање на алкохол за викенди”

Од хистограмот на слика 38 можеме да забележиме драстично поразлична распределба, во споредба со распределбата која се однесува на консумирање на алкохол во текот на работната недела. Имено, повторно преовладуваат ученици кои не консумираат многу алкохол, но поголем е бројот на ученици и кои консумираат алкохол во умерни, но и во не баш толку умерени количини.

Со цел подобра споредба на овие два блиски атрибути, можеме да ги погледнеме нивните средни вредности, со цел да видиме колку истите се разликуваат. Притоа, нивните стандардни девијации ќе допринесат за донесување на заклучок за тоа дали најголем дел од вредностите се натрупани околу самата средна вредност.

```
data['Dalc'].mean()
```

```
1.4942528735632183
```

```
data['Dalc'].std()
```

```
0.9117142815596279
```

Слика 38. "Dalc mean & std"

```
data['Walc'].mean()
```

```
2.2844827586206895
```

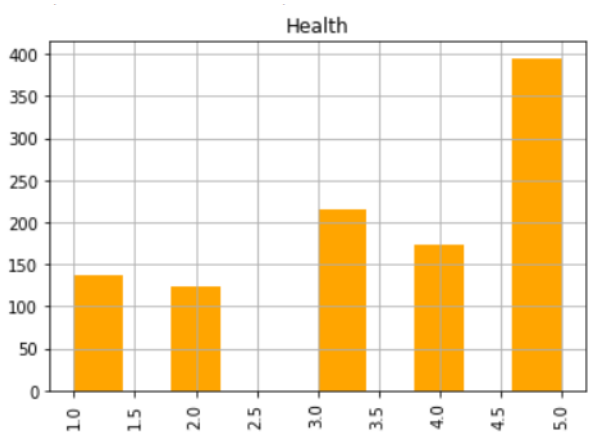
```
data['Walc'].std()
```

```
1.285104806261859
```

Слика 39. "Walc mean & std"

Како што и претпоставивме, средните вредности на овие две карактеристики се разликуваат во голема мера. Од една страна, за атрибутот кој се однесува на консумирањето на алкохол во текот на работната недела имаме средна вредност од 1.49, додека пак, за атрибутот кој се однесува на консумирањето на алкохол во текот на викендот, имаме средна вредност од 2.28, што е речиси двојно повеќе. Воедно, ова е и логички поддржано, имајќи во предвид дека за време на викенд учениците излегуваат, па најголем дел од нив и консумираат алкохол, во споредба со работните денови, кога најчесто се излегува на кафе, после училиште.

Понатаму, продолжуваме со разгледување на следниот атрибуот, кој се однесува на тековната здравствена состојба на ученикот. За подетални информации во врска со опсегот на вредностите за овој атрибут, како и неговата распределба, можеме да го погледнеме хистограмот на слика 40.



Слика 40. "Распределба на вредности – здравствена состојба"

Доколку го погледнеме хистограмот на слика 40, можеме да забележиме дека вредностите и на овој атрибут се движат во опсегот од 1, до 5. Притоа, како и претходно, вредноста 1 означува лоша здравствена состојба, додека пак вредноста 5 означува одлична здравствена состојба. Иако преовладуваат ученици со одлична здравствена состојба, сепак и останатите вредности за самата здравствена состојба на учениците се доста застапени. Средната вредност за овој атрибут изнесува 3.54, како што можеме да видиме на сликата во продолжение, што укажува на фактот дека преовладуваат ученци кои имаат добра здравствена состојба. Ниската вредност на стандардната девијација од 1.42, се придружува кон ова тврдење, укажувајќи на фактот дека најголем дел од вредностите за овој атрибут се наоѓаат околу самата средна вредност.

```
data['health'].mean()
```

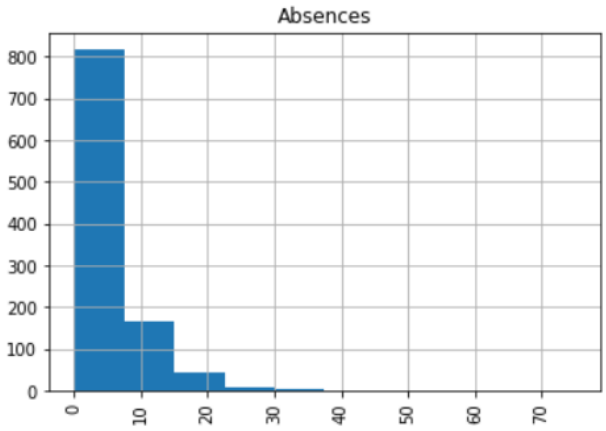
```
3.543103448275862
```

```
data['health'].std()
```

```
1.424703412249021
```

Слика 41. "Здравствена состојба – mean & std"

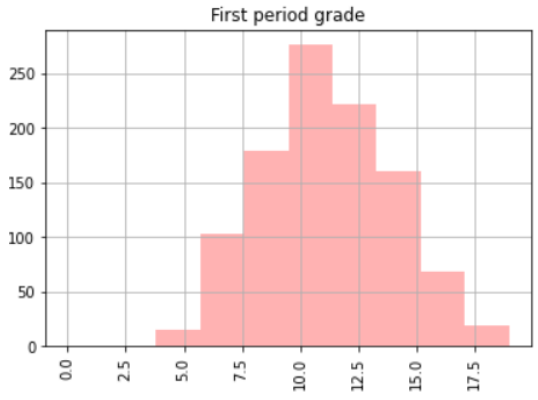
Следно, ќе ја разгледаме колоната која се однесува на отстуствата на учениците и ќе ја истражиме распределбата на вредностите за овој атрибут.



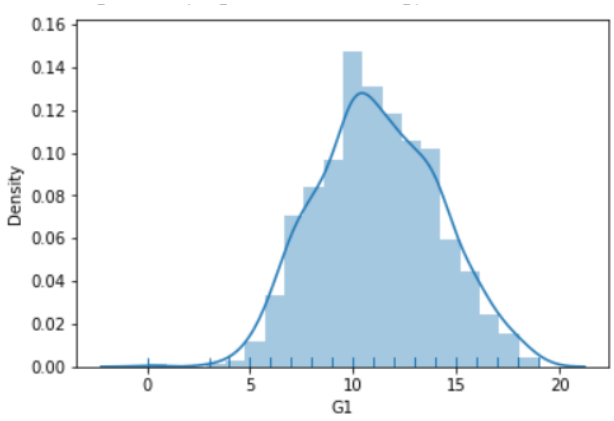
Слика 42. "Распределба на вредности - отсуства"

Од хистограмот на слика 42 можеме да забележиме дека најголем дел од учениците немаат воопшто отсуства, или имаат до 3 отсуства. Потоа мал број од нив, имаат отсуства во опсегот од 5 до 30, а речиси никој нема повеќе од 50 отсуства.

Понатаму, ќе продолжиме со разгледување на три доста слични и корелирани атрибути. Имињата на овие атрибути се G1, G2 и G3, а истите се однесуваат на оценките од прво полугодие, оценките од второ полугодие, како и конечните оценки соодветно. Нивниот опсег е ист, а претпоставуваме дека и нивната распределба ќе е во голема мера блиска меѓусебно. Најпрво ќе ги разгледаме оценките за првото полугодие, а потоа ќе преминеме на останатите.



Слика 43. "Распределба на вредности – оценки за прво полугодие"

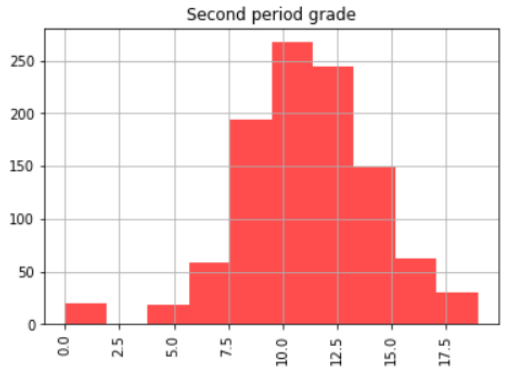


Слика 44. "Распределба на вредности 2 - оценки за прво полугодие"

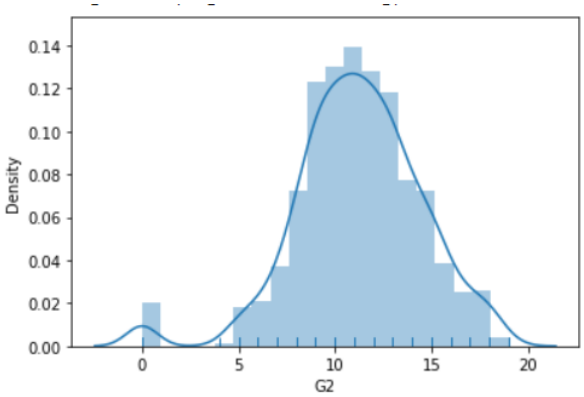
Како што можеме да забележиме од двата хистограми, опсегот на вредности за оценките на учениците се движи во границите од 0, до 20. Најголем дел од учениците имаат оценки

кои се движат во рангот од 8 до 15, а многу е мал бројот на ученици кои имаат оценки поголеми од 17.

Да ја разгледаме распределбата за оценките кои се однесуваат на второто полугодие.



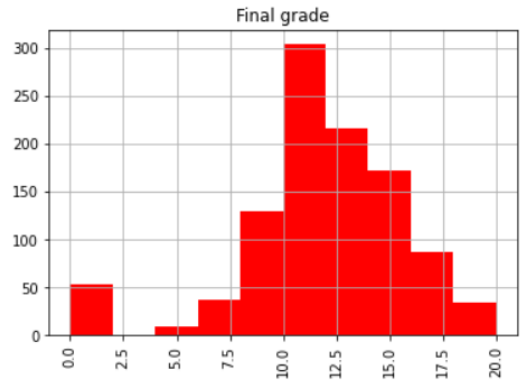
Слика 45. "Распределба на вредности – оценки за второ полугодие"



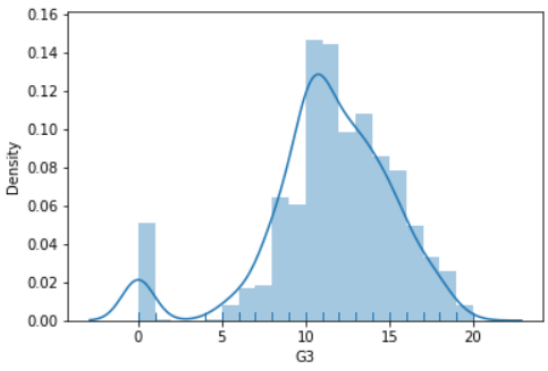
Слика 46. "Распределба на вредности 2 – оценки за второ полугодие"

Од хистограмите на сликите 45 и 46, може да забележиме дека имаме доста слична распределба, односно, најголем дел од учениците имале оценки во рангот од 8 до 14 и на второто полугодие. Сепак, можеме да забележиме минимален раст на овие вредности, кои водат и до минимален раст на самата средна вредност и просек на учениците. Притоа, можеме да забележиме дека кај одреден дел од учениците се појавуваат и многу ниски оценки, во рангот од 0 до 3, што не беше случај кај оценките од прво полугодие.

Да ја погледнеме и ситуацијата со крајните оценки на учениците.



Слика 47. "Распределба на вредности – крајни оценки"



Слика 48. "Распределба на вреднсоти 2 - крајни оценки"

Воочливо е дека распределбата на вредностите за крајните оценки е во голема мера слична со оценките од првото и второто полугодие. Она што можеме да забележиме е дека бројот на ниски оценки е зголемен, во споредба со второто полугодие, но имаме и одредено

зголемување кога станува збор за оценките во рангот од 16 до 20. Како и кај претходните два случаи, така и овде, преовладуваат оценки во границите од 10 до 15.

Интересно би било да ги погледнеме и средните вредности на секој од овие атрибути, со цел да заклучиме дали целокупниот успех во училиштето расте, или опаѓа во текот на годината.

```
data['G1'].mean()
```

```
11.21360153256705
```

Слика 49. "G1 mean"

```
data['G2'].mean()
```

```
11.246168582375478
```

Слика 50. "G2 mean"

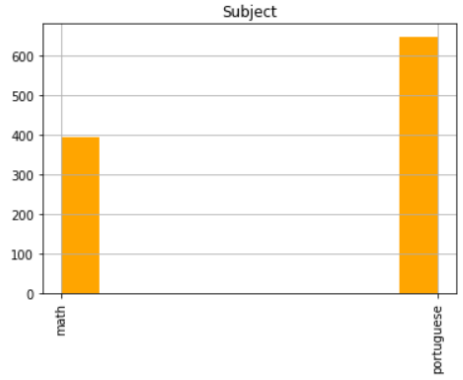
```
data['G3'].mean()
```

```
11.341954022988507
```

Слика 51. "G3 mean"

Како што и заклучивме од самите хистограми, средната вредност на оценките на учениците е околу 11. Притоа, можеме да забележиме раст од 0.03 кај оценките од прво до второ полугодие. Кај оценките од прво, до крајните оценки имаме раст од дури 0.13. Ова укажува на фактот дека крајните оценки кај најголем дел од учениците се поголеми од почетните, односно тие имаат тенденција да растат, укажувајќи на фактот дека самите ученици посветуваат поголемо внимание и повеќе се грижат за крајните оценки кои се потоа пресудни за запишување на факултет, во споредба со оценките во текот на самата година.

За крај, да ја разгледаме и распределбата на вредности на колоната која сами ја додадовме, а истата се однесува на предметот, односно, дали станува збор за математика, или португалски јазик



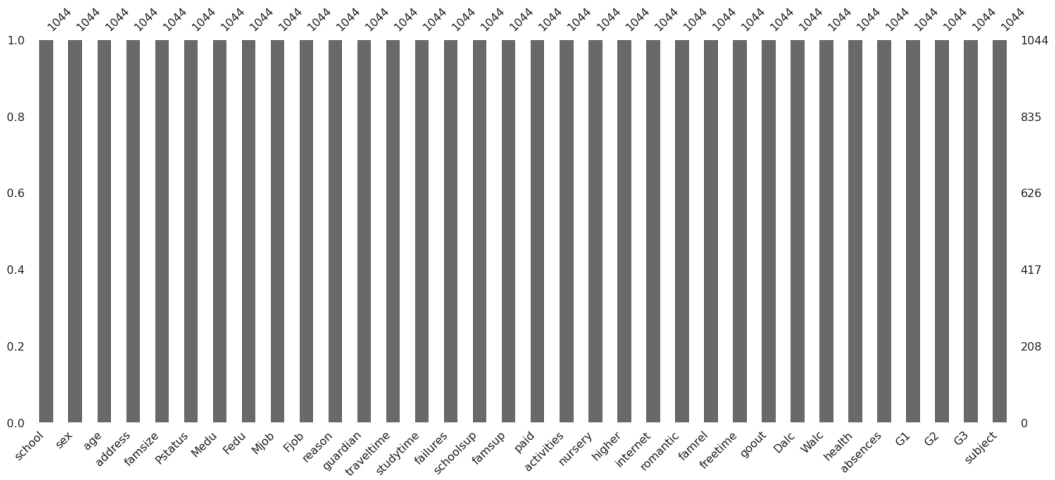
Слика 52. "Распределба на вредности - предмет"

Од самиот хистограм можеме да забележиме дека за 1/5 повеќе имаме ученици со оценки по португалски јазик, во споредба со ученици со оценки по математика.

Анализа на вредностите кои недостасуваат во дадените атрибути на множеството

Несоодвените вредности, како и вредностите кои недостасуваат се чест случај и проблем при истражување на врските и релациите кај голем број на податочни множества. Тие можат да бидат причинители за несоодветно толкување на корелациите помеѓу атрибутите, а соодветно на тоа и изведување на неточни заклучоци.

Поради таа причина, пред да преминеме кон истражување на врските и релациите кои постојат помеѓу атрибутите, потребно е да се справиме со ваквите невалидни вредности. Меѓутоа, множеството кое е од наш интерес, нема вредности кои недостасуваат, а ова можеме да го забележиме на дијаграмот во продолжение.

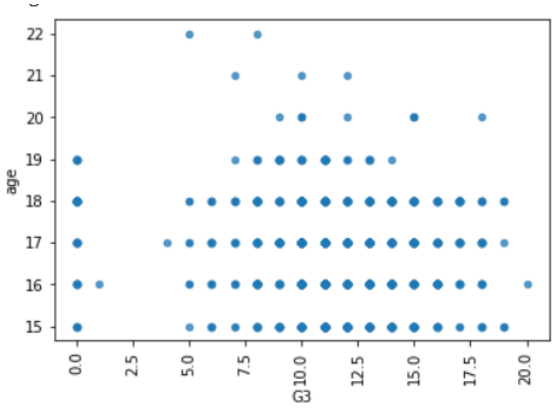


Слика 53. "Дијаграм на вредности кои недостасуваат"

Откако се уверивме дека невалидните и вредностите кои недостасуваат не претставуваат пречка во нашите истражувања, може да се премине кон следниот чекор, односно да навлеземе и да започнеме со проучување на врските кои постојат помеѓу различните атрибути на податочното множество.

Анализа на врските и поврзаноста помеѓу атрубутите на податочното множество

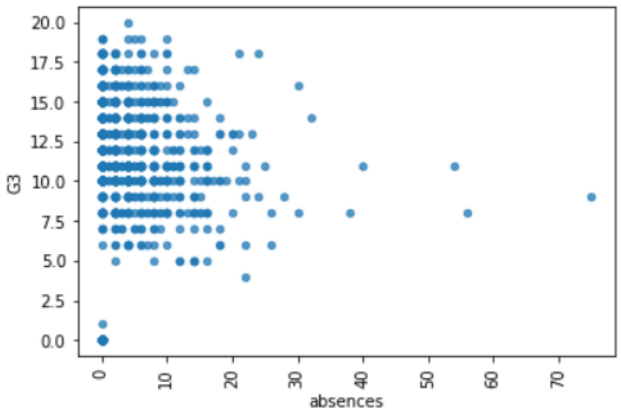
Со цел пишување на добри прашалници, од клучно значење е добро познавање на врските и корелациите кои постојат помеѓу атрибутите. Поради таа причина, во продолжение следуваат неколку дијаграми на поврзаност, кои ги истражуваат поврзаностите помеѓу атрибутите.



Слика 54. "Дијаграм на поврзаност – G3 & age"

На слика 54 имаме приказ на поврзаноста која постои помеѓу годините и успехот на учениците. Забележливо е дека не постои некоја силна поврзаност, односно, годините не се значаен фактор за перформансите на учениците.

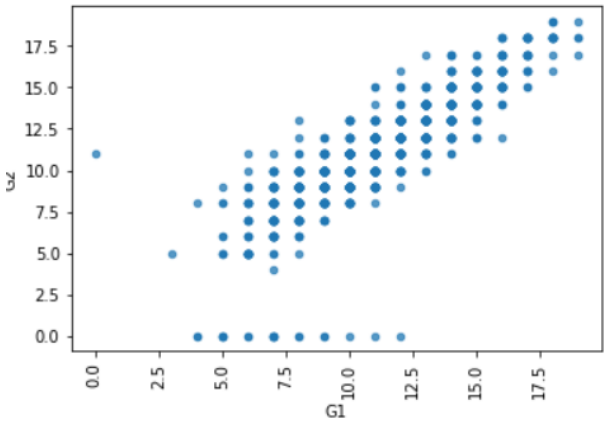
Влијанието на отсуствата врз крајниот успех на учениците е прикажано на слика 55.



Слика 55. "Дијаграм на поврзаност – absences & G3"

Самиот дијаграм ни покажува одредени корелации помеѓу овие атрибути. Односно, можеме да забележиме позитивна врска, меѓутоа истата не е многу изразена. Ова може да се толкува како минимално намалување на бројот на отсуства со зголемување на оценките.

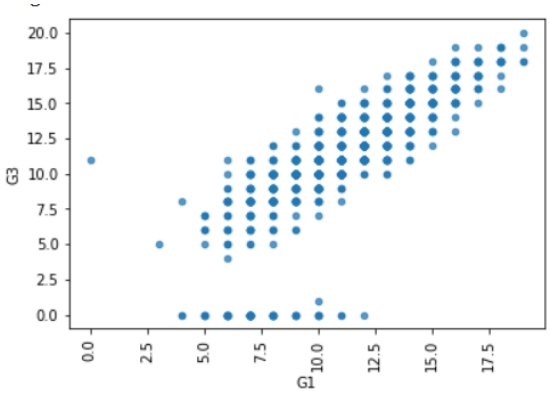
Претпоставуваме дека оценките од прво и второ полугодие би биле силно поврзани, па ова можеме да го истражиме на дијаграмот на поврзаност од сликата 56.



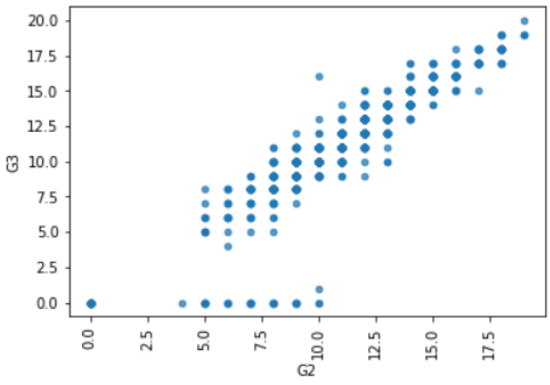
Слика 56. "Дијаграм на поврзаност – G1 & G2"

Како што и претпоставивме, помеѓу атрибутите G1 и G2, односно, помеѓу оценките од прво и оценките од второ полугодие постои силна линеарна поврзаност. Односно, колку е повисока оценката од прво полугодие, иста, или повисока ќе е и оценката од второ полугодие.

Во продолжение да ги погледнеме влијанијата кои оценките од прво и второ полугодие ги имаат врз крајните оценки на учениците. Ова е прикажано на сликите 57 и 58.



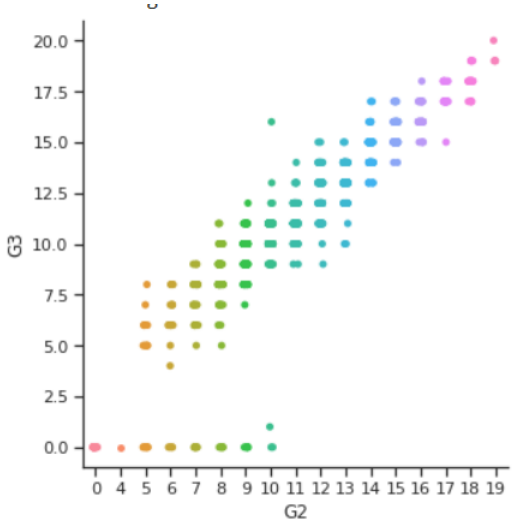
Слика 57. "Дијаграм на поврзаност – G1 & G3"



Слика 58. "Дијаграм на поврзаност – G2 & G3"

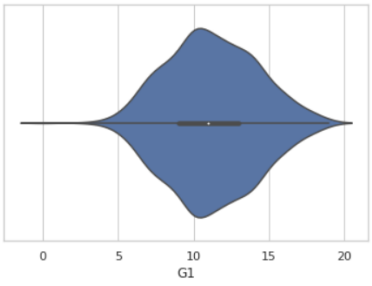
Од овие дијаграми на поврзаност можеме да заклучиме дека крајната оценка на учениците е во силна линеарна поврзаност и зависи од оценката на ученикот од прво, а воедно и од второ полугодие. Секако, постојат и исклучоци, меѓутоа, како што можеме да видиме на самите дијаграми, истите се минимални.

Особено е силна поврзаноста помеѓу оценките од второ полугодие и крајната оценка, па поради тоа во продолжение на слика 59 можеме да го погледнеме истиот дијаграм на поврзаност, меѓутоа со дополнителни кластерирачки информации.

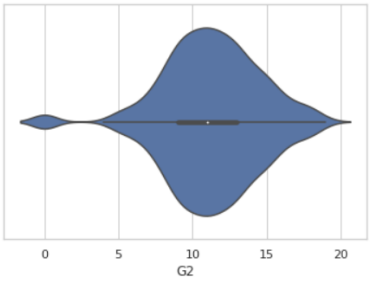


Слика 59. "Дијаграм на поврзаност со кластери – G2 & G3"

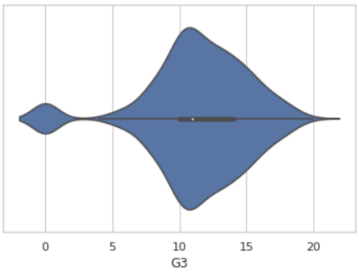
Во продолжение повторно ќе ја погледнеме распределбата на вредностите на оценките на учениците, притоа воочувајќи ја и уште еднаш нагласувајќи ја нивната силна зависност и меѓуповрзаност.



Слика 60. "G1 хистограм"



Слика 61. "G2 хистограм"

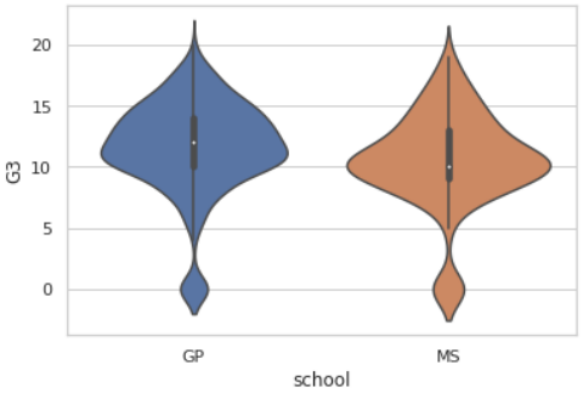


Слика 62. "G3 хистограм"

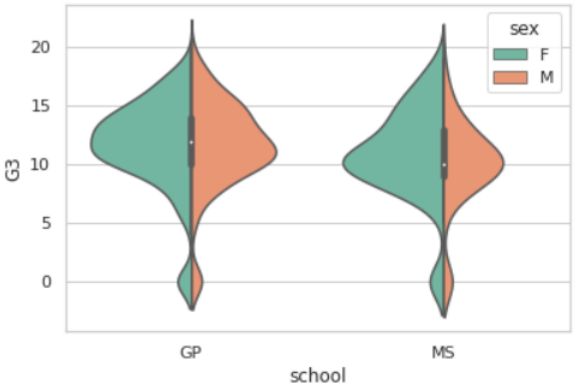
Можеме да воочиме колку се трите хистограми слични меѓусебно, што укажува на нивната зависност и силна корелација. Притоа, забележливи се поостри рабови околу оценките

10/11 кај атрибутите кои се однесуваат на оценки од второ полугодие, како и финалните оценки.

Следно, можеме да ја разгледаме распределбата на оценките и врската на истите во зависност од училиштата. Притоа, на слика 64 можеме да ја погледнеме оваа иста распределба, но притоа да направиме разлика и по полот на учениците, со цел да навлеземе подлабоко во истражувањето на овие атрибути.



Слика 63. "Поврзаност - school & G3"

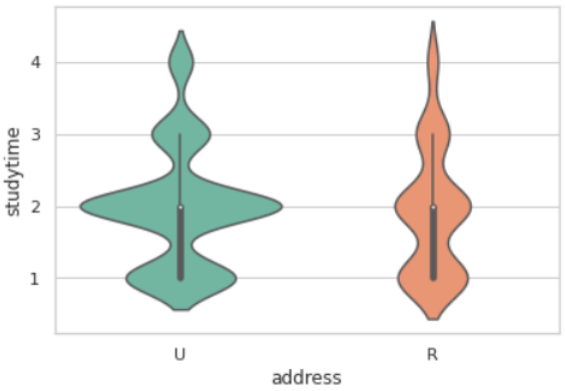


Слика 64. "Поврзаност – school, G3 & sex"

Од хистограмот на слика 63, можеме да забележиме дека распределбата на оценките е многу слична и во двете училишта. Сепак, забележливо е дека просечната вредност е под 10 кај училиштето Mousinho da Silveira, додека пак во училиштето Gabriel Pereira просечната вредност на крајни оценки е речиси 12. Дополнително, двете училишта се карактеризираат со многу малку ученици кои имаат оценки поголеми од 18.

Доколку во игра го додадеме и полот и вниманието го насочиме на хистограмот на слика 64, можеме да забележиме дека во училиштето Mousinsho de Silveira, освен што преовладуваат ученици од женски пол, истите во голема мера имаат и предност кога станува збор за повисоки оценки. Односно, во ова училиште, учениците од машки пол имаат забележливо помалку високи оценки од учениците од женски пол, што не е случај во училиштето Gabriel Pereira.

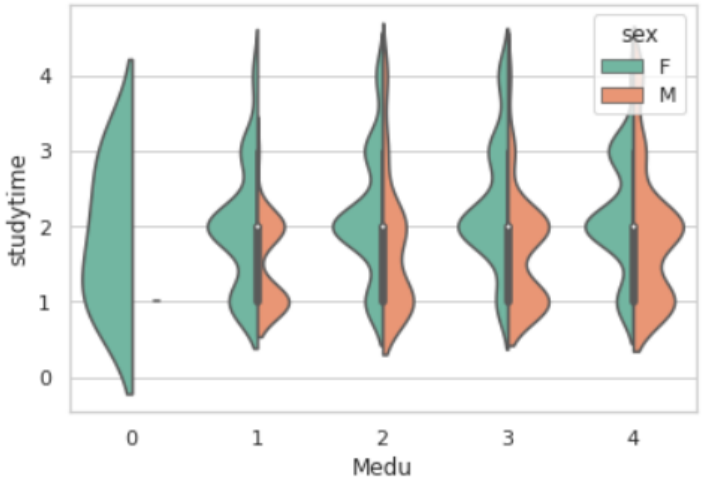
Понатаму можеме да направиме анализа на тоа колку адресата и средината на живеење влијаат на времето посветено на учење. Подетални информации во врска со ова истражување можеме да добиеме од хистограмот во продолжение.



Слика 65. "Хистограм на поврзаност – address & studytime"

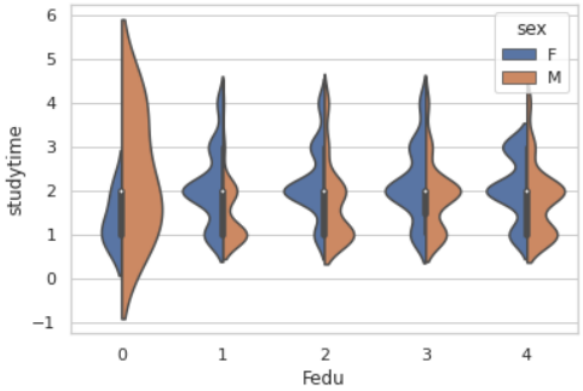
Доколку го разгледаме овој хистограм ќе увидиме дека учениците од рурална средина посветуваат помалку време на учење, што е секако разбирливо. Овие ученици губат повеќе време на патување, па разбирливо е да постои минимална предност кај учениците од урбани средини за времето посветено на учење.

Следно, фокусот го ставив на испитување на врската која постои помеѓу времето кое учениците го посветуваат на учење, во зависност од степенот на образование на нивните родители. Притоа, оваа анализа ја сегментирав и по самиот пол на учениците, со цел да истражам дали мајката/таткото имаат повисоко влијание кон децата од истиот пол како нив. На слика 66 е прикажан хистограмот кој се однесува на поврзаноста на учењето со степенот на образование на мајките, а на хистограмот на слика 67 е прикажана корелацијата со степенот на образование на татковците.



Слика 66. "Хистограм на поврзаност на образованието на мајката со времето посветено на учење, во зависност од полот на учениците"

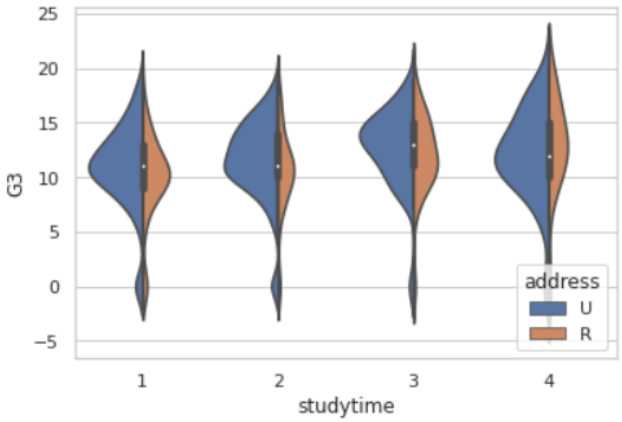
Забележливо е дека степенот на образование на мајката има одреден удел во времето на учење кое го посветуваат учениците. Притоа, забележливо е поголемо влијание кај учениците од женски пол, каде, колку е повисок степенот на образвание на мајката, толку повеќе време овие ученици вложуваат на учење. Слична е и тенденцијата кај учениците од машки пол, но истата е многу помалку изразена.



Слика 67. "Хистограм на поврзаност на образованието на таткото со времето посветено на учење, во зависност од полот на учениците"

Од друга страна, степенот на образование на таткото нема голема зависност со времето кое учениците од машки пол го посветуваат на учење. Генерално, без разлика кое е образованието кое нивниот татко го има завршено, учениците од машки пол одвојуваат исто време за учење. Ова не е случајот со учениците од женски пол, каде се забележува најпрво растење во времето на учење, со зголемување на степенот на образование на таткото, но сепак, на крај и мало опаѓање на истото.

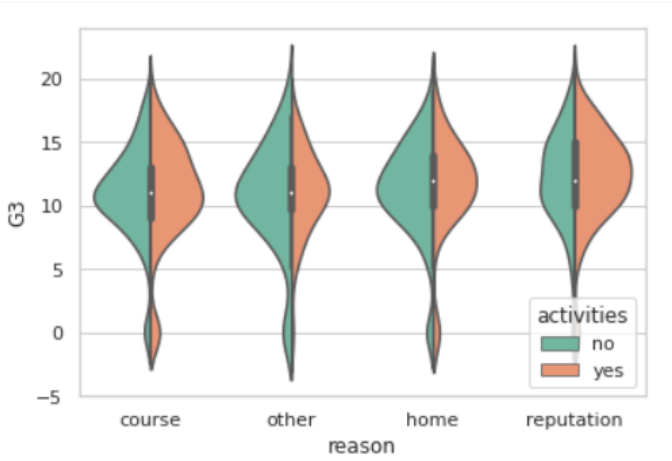
На хистограмот од сликата 68 е дадена корелацијата помеѓу оценките на учениците, времето на учење, како и адресата на живеење.



Слика 68. "Хистограм на поврзаност на времето на учење со крајните оценки, во зависност од адресата на живеење на учениците"

Адресата на живеење во голема мера влијае на времето кое учениците го посветуваат на учење, а времето на учење пак директно влијае врз перформансите на учениците. На хистограмот, јасно е забележливо дека учениците од урабни средини посветуваат повеќе време за учење, а притоа и во секое од скалилата за време на учење, тие добиваат повисоки оценки во споредба со учениците од рурални средини.

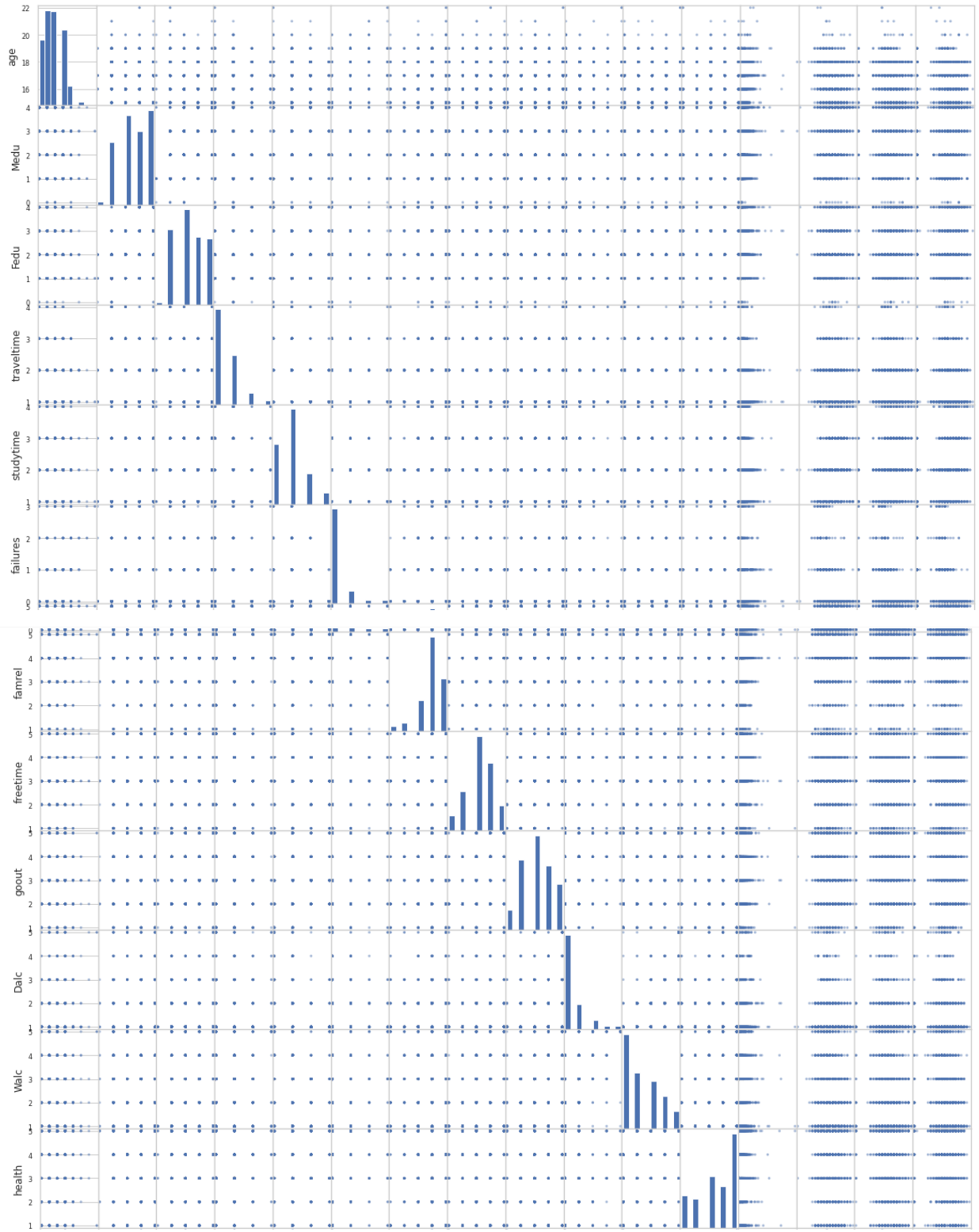
Ќе погледнеме и како причината поради која учениците го избрале училиштето влијае врз нивните перформанси. Со цел да навлеземе подлабоко во тематиката, ќе направиме и поделба според тоа дали учениците преземаат воншколски активности, или не.

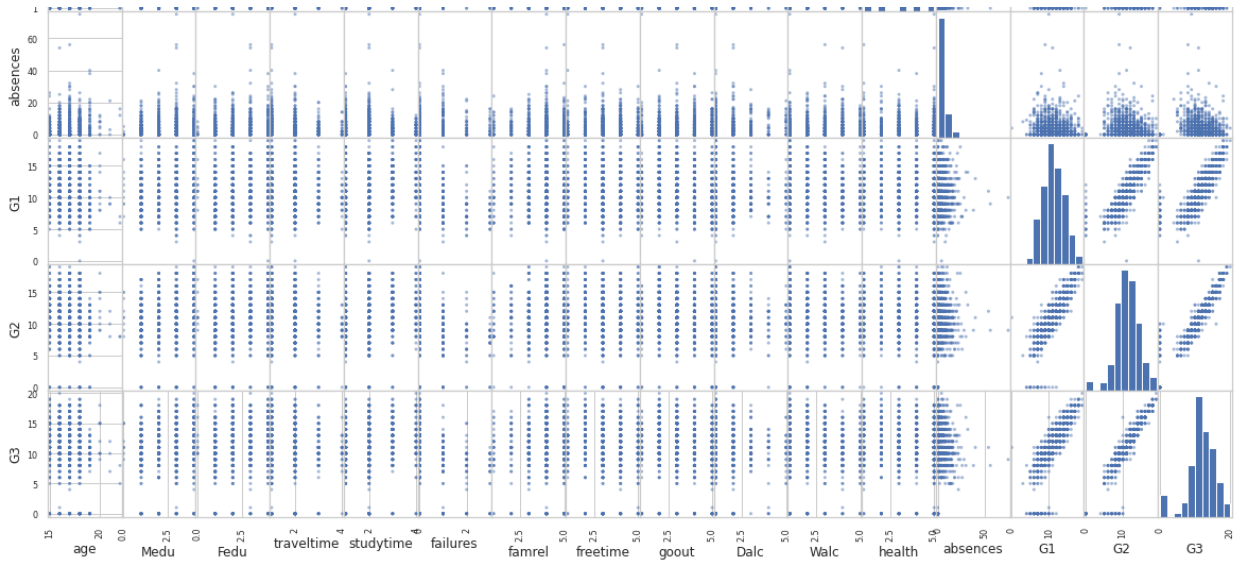


Слика 69. "Хистограм на поврзаност на причината поради која го избрале училиштето, со крајниот успех, во зависност од тоа дали преземаат воншколски активности, или не."

Од хистограмот, јасно е забележливо, дека учениците кои го избрале училиштето поради предметите кое истото ги нуди, како и поради репутацијата, вообичаено покажуваат повисоки перформанси. Супериорни во високи оценки се учениците кои покрај овие две причини за избор на училиштето, преземаат и воншколски активности.

Во продолжение, на сликата 70, ќе биде дадена матрицата на корелација, која ни дава добар визуелен приказ во корелациите што постојат помеѓу нумеричките атрибути од податочното множество.


age
Medu
Fedu
traveltime
studytime
failures
famrel
freetime
goout
Dalc
Walc
health



Слика 70. "Матрица на корелација"

Од матрицата на корелација можеме да ја забележиме силната линеарна поврзаност помеѓу оценките од прво и второ полугодие, како и нивната силна меѓузависност со крајните оценки. Воедно, се забележува и позитивна поврзаност со отсуствата.

Со цел поголема прегледност и нумеричка достапност, ќе погледнеме табела во која нумерички се покажани степените на поврзаност на секој од атрибутите.

Correlation matrix

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
age	1.00	-0.13	-0.14	0.05	-0.01	0.28	0.01	0.00	0.12	0.13	0.10	-0.03	0.15	-0.12	-0.12	-0.13
Medu	-0.13	1.00	0.64	-0.24	0.09	-0.19	0.02	0.00	0.03	0.00	-0.03	-0.01	0.06	0.23	0.22	0.20
Fedu	-0.14	0.64	1.00	-0.20	0.03	-0.19	0.01	0.00	0.03	-0.00	0.02	0.03	0.04	0.20	0.18	0.16
traveltime	0.05	-0.24	-0.20	1.00	-0.08	0.09	-0.01	-0.01	0.05	0.11	0.08	-0.03	-0.02	-0.12	-0.14	-0.10
studytime	-0.01	0.09	0.03	-0.08	1.00	-0.15	0.01	-0.09	-0.07	-0.16	-0.23	-0.06	-0.08	0.21	0.18	0.16
failures	0.28	-0.19	-0.19	0.09	-0.15	1.00	-0.05	0.10	0.07	0.12	0.11	0.05	0.10	-0.37	-0.38	-0.38
famrel	0.01	0.02	0.01	-0.01	0.01	-0.05	1.00	0.14	0.08	-0.08	-0.10	0.10	-0.06	0.04	0.04	0.05
freetime	0.00	0.00	0.00	-0.01	-0.09	0.10	0.14	1.00	0.32	0.14	0.13	0.08	-0.03	-0.05	-0.07	-0.06
goout	0.12	0.03	0.03	0.05	-0.07	0.07	0.08	0.32	1.00	0.25	0.40	-0.01	0.06	-0.10	-0.11	-0.10
Dalc	0.13	0.00	-0.00	0.11	-0.16	0.12	-0.08	0.14	0.25	1.00	0.63	0.07	0.13	-0.15	-0.13	-0.13
Walc	0.10	-0.03	0.02	0.08	-0.23	0.11	-0.10	0.13	0.40	0.63	1.00	0.11	0.14	-0.14	-0.13	-0.12
health	-0.03	-0.01	0.03	-0.03	-0.06	0.05	0.10	0.08	-0.01	0.07	0.11	1.00	-0.03	-0.06	-0.09	-0.08
absences	0.15	0.06	0.04	-0.02	-0.08	0.10	-0.06	-0.03	0.06	0.13	0.14	-0.03	1.00	-0.09	-0.09	-0.05
G1	-0.12	0.23	0.20	-0.12	0.21	-0.37	0.04	-0.05	-0.10	-0.15	-0.14	-0.06	-0.09	1.00	0.86	0.81
G2	-0.12	0.22	0.18	-0.14	0.18	-0.38	0.04	-0.07	-0.11	-0.13	-0.13	-0.09	-0.09	0.86	1.00	0.91
G3	-0.13	0.20	0.16	-0.10	0.16	-0.38	0.05	-0.06	-0.10	-0.13	-0.12	-0.08	-0.05	0.81	0.91	1.00

Слика 71. "Табеларен приказ на матрица на корелација"

Од овој табеларен приказ можеме да забележиме дека најсилна е поврзаноста на оценките од второ полугодие со крајните оценки. Потоа, силна е поврзаноста и помеѓу оценките од прво полугодие со оценките од второ полугодие, а малку послаба, за 0.05 е и корелацијата помеѓу оценките од прво полугодие со крајните оценките.

Следно што можеме да забележиме, дека блага предност кон подобри оценки има степенот на образование на мајката, над оној на таткото со 0.04.

Можеме да забележиме дека времето на учење е многу позависно со оценките од првото полугодие, во споредба со оние на крај. Ова укажува на фактот на кој секој од нас бил сведок, односно дека добриот прв впечаток е од огромно значење. Ова ни остава простор за толкување дека учениците кои на почетокот се посветиле на учење и оставиле добар впечаток и постигнале одличен успех, понатаму успеале да го одржат истиот со помалку учење.

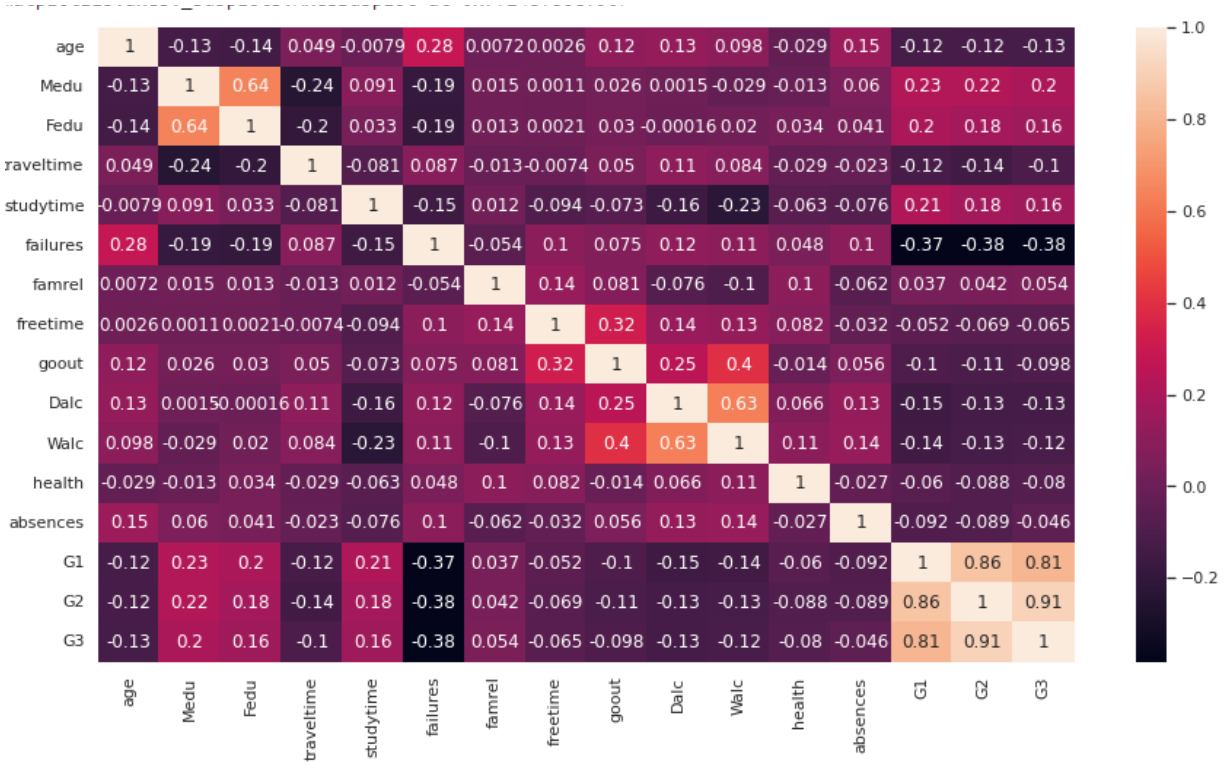
Бројот на презапишани предмети е во обратна, односно негативна корелација со успехот. Логичко, колку е поголем бројот на паднати предмети, толку оценките се пониски.

Забележлива е и минимална негативна поврзаност на изостаноците со крајните оценки, односно, повеќе изостаноци, донесуваат до пониски крајни оценки.

Како што споменавме и претходно, времето на учење е во негативна корелација со времето потребно за патување, како и со самите оценки. Односно, колку подолго време учениците патуваат, нивните оценки имаат тенденција минимално да опаѓаат.

Позитивна зависност гледаме помеѓу слободното време и времето на излегување со пријатели, што исто така е многу логично. Додека пак, негативна поврзаност имаме помеѓу излегувањето со пријатели и времето посветено на учење.

Во продолжение можеме да ја погледнеме и т.н. heat-мапа, која ни овозможува уште еднаш да направиме осврт кон врските и поврзаноста на атрибутите, притоа нагласувајќи ја нивната корелација со градиент бои.



Слика 72. "Heatmap"

Следно, можеме да направиме и анализа на тоа како секој од атрибутите влијае врз оценките на учениците, како на прво, така и на второ полугодие. Графиконите од оваа анализа можеме да ги погледнеме на сликите 73 и 74.

На графиконите можеме да ги забележиме различните влијанија и врски кои постојат и се значаен фактор врз оценките на учениците. Можеме да увидиме дека просечната оценка е повисока во училиштето Mousinho da Silveira. Забележливи се и најниски оценки кај учениците кај кои мајката не работи, односно е at home mom. Исто така, многу пониски оценки имаат учениците кои за старател имаат трето лице.

Понатаму, можеме да воочиме дека учениците кои немаат намера да продолжат со високо образование имаат значително пониска средна вредност на оценки, во споредба со оние кои планираат да продолжат. Интересно е да се забележи дека еднакви се оценките и на учениците кои преземаат, но и на оние кои не преземаат воншколски активности.

Исто така, највисоки оценки имаат учениците кои патуваат помалку од 15 минути, додека најниски имаат оние кои патуваат повеќе од еден час. Воедно, значително пониски оценки имаат и учениците кои консумираат алкохол во поголеми количини, како за викенд, така и во текот на работната недела.

Како што може да се претпостави, здравствената состојба е исто така фактор во перформансите на учениците, па оние кои имаат одредени здравствени проблеми, најчесто имаат пониски оценки.

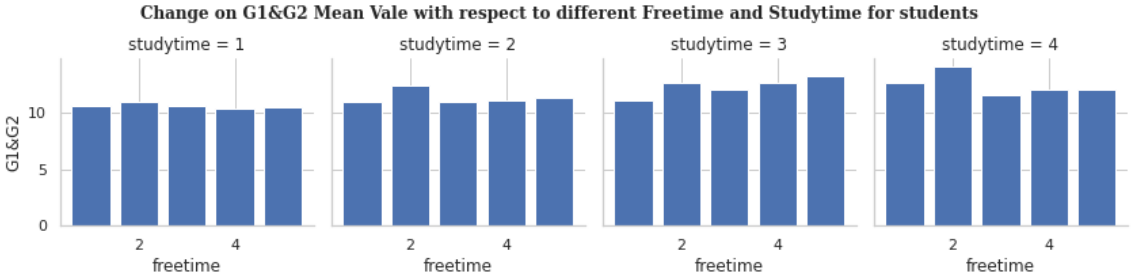
Образованието на родителите е исто така значаен фактор, кој е позитивно корелиран со оценките на учениците.

Слика 73. "Графикон на влијание на атрибутите врз оценките од прво и второ полугодие"

Слика 74. "Графикон на влијание на атрибутите врз оценките од прво и второ полугодие - продолжение"

Следно, имајќи во предвид дека времето кое учениците го посветуваат на учење влијае врз нивните перформанси, а врз времето кои учениците го посветуваат на учење влијае

нивното слободно време, ќе ја разгледаме поврзаноста помеѓу овие атрибути. Ова е детално претставено на хистограмот во продолжение.



Слика 75. "Поврзаност на атрибутите freetime, studytime and grades"

На сликата 75 можеме да забележиме приказ од корелациите помеѓу гореспоментатите атрибути: слободно време, време за учење и оценки.

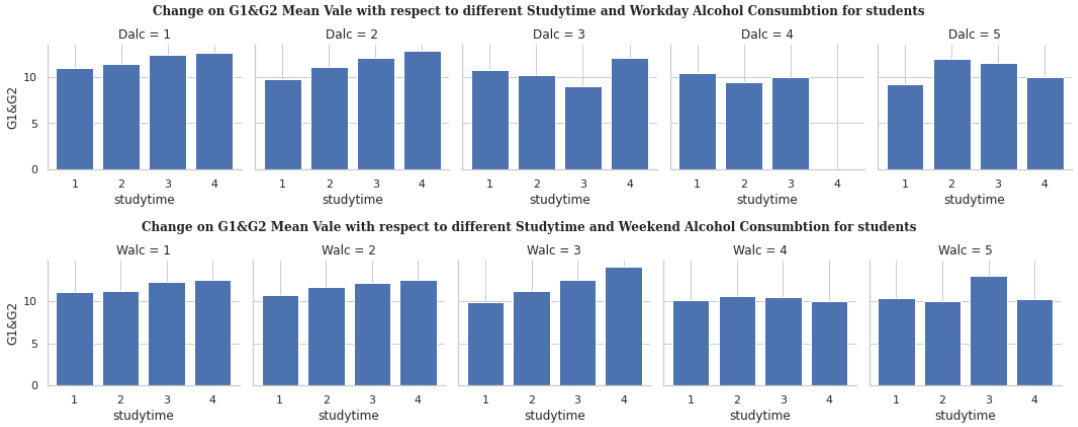
Интересен случај е најдесниот хистограм, каде имаме вредност за време на учење еднаква на четири, што значи повеќе од 10 часа. Притоа, учениците кои имаат помалку слободно време, со вредност 2, имаат подобри перформанси од учениците кои имаат повеќе слободно време.

Ова може да го разгледаме од две перспективи, односно, учениците кои имаат помалку слободно време, повеќе се посветуваат на учењето, или пак, истите имаат помалку слободно време поради големиот број на дополнителни активности, кои секако им ги подобруваат перформансите.

За крај, ќе го погледнеме влијанието кое времето на учење, во комбинација со консумирањето на алкохол во текот на неделата, но и во текот на викендот, го имаат врз оценките на учениците. Детален приказ на ова имаме на слика 76, која е дадена во продолжение.

Како што споменавме и претходно, оценките на учениците се во инверзна релација со консумирањето на алкохол. Ова истражување можеме да го продлабочиме со погоре прикажаните хистограми, каде ги споредуваме средните вредности од оценките на учениците во зависност од нивното време на учење со консумирањето на алкохол во работни денови и преку викенд.

Видливо е дека највисоки средни оценки се постигнати при консумирање на мали количини на алкохол во текот на неделата – хистограмите лево 1 и 2.



Слика 76. "Поврзаност на атрибутите: studytime, dalc, walc and grades"

За разлика од консумирањето на алкохол во текот на неделата, кое е посилно изразено, консумирањето алкохол за викенд нема силна тенденција, меѓутоа можеме да забележиме дека при исто консумирање на алкохол, се зголемуваат оценките, доколку се зголеми времето на учење.

Од спроведената анализа на атрибутите од податочното множество, стекнавме добри информации и добивме добра претстава за генералната структура и тенденциите на поврзаност во податочното множество. Длабоките знаења за корелациите и поврзаностите на атрибутите ќе бидат добра основа за интересни прашалници, кои уште еднаш ќе ги потврдат и истражат главните фактори за успехот на учениците.

Импортирање и поврзување на двете бази соодветно, MongoDB и CouchDB

За креирање и поврзување на MongoDB користев MongoDB Atlas, како и MongoDB Compass. Atlas е сервис базиран во облак за работа со бази на податоци, направен од истиот тим кој работел на MongoDB. Истиот го упростува менаџирањето со бази на податоци нудејќи ја потребната разновидност за градење на глобални апликации со добри перформанси базирани во облак. Од друга страна, Compass претставува графички кориснички интерфејс кој овозможува лесен начин на внесување на податоци. Освен можноста за внесување и креирање на посебни JSON документи, постои и опција за директно додавање на цела табела, при што имаме можност да го избереме податочниот тип за секој од атрибутите. Ова е од голема помош и многу корисно особено во вакви случаи каде самото податочно можество има огромен број на атрибути. На овој начин внесувањето на податочното множество во база беше пребрзо и многу интуитивно. По самото импортирање на податоците во база, се поврзав локално со базата користејќи MongoDB Shell и преминав кон извршување на веќе креираните прашалници во терминал.

За креирање и поврзување на CouchDB користев Python и Project Fauxton. Project Fauxton претставува веб базиран интерфејс, кој го користев во процесот на дефинирање на логиката и операциите на самите прашалници. Истиот е базиран на map-reduce концепт, кој нуди неколку можности за избор на reduce функции, како што се sum, count, stats, а покрај тоа нуди и можност за custom дефинирање на сопствен метод. Базиран е на JavaScript, кој е доста едноставен и нуди едноставна манипулација и “играње” со документите.

Најпрво, пред процесот на импортирање, се премина кон креирање на JSON документ, за секој ред од податочното множество. Самиот процес на импортирање на податоците беше овозможен од couchdb пакетот во Python, со чијашто помош на брз и едноставен начин успеав да ги внесам JSON документите во базата. По импортирање на податоците во базата на податоци, со истата поврзувањето беше локално и се започна со извршување на веќе креираните прашалници во терминал.

Модели на агрегација и имплементација

Со цел да се истражат врските и поврзаностите кои веќе беа пронајдени во податочното множество, треба да се дефинираат соодветни прашалници, дадени во продолжение, притоа користејќи различни модели за агрегација. Прашалниците се во насока да направат анализа на веќе пронајдените меѓузависности и да помогнат во увидот на најзначајните фактори кои влијаат врз перформансите на учениците.

Притоа, во нив се сретнуваат следните агрегациски модели:

- Average (AVG): Просек на податочните вредности.
- Sum: Вкупна вредност од податочните вредности.
- Count: Број на записи.
- Maximum (Max): Максимална вредност на дадени податочни вредности.
- Minimum (Min): Минимална вредност на дадени податочни вредности.
- Group: Групирање на дадени податочни вредности според одреден критериум.
- Map-Reduce: За групирање на сите податоци базирани на клуч-вредност и reduce функција која се користи за изведување на операции на селектираните податоци.
- Stats: ни враќа информации за сума, минимум, максимум, бројач како и сума од квадратите на вредноста која сме одлучиле да и правиме reduce.

Во продолжение може да се видат самите прашалници и нивната имплементација во двете бази на податоци MongoDB и CouchDB.

Прашалник	MongoDB	CouchDB

Најди ги учениците кои учат помалку од 2 часа а имаат поголема крајна оценка од 18.	db.students.find({$and: [{studytime: {$eq: 1}}, {G3: {gt: 15}}]})	//Map
		function (doc) {
		if(doc.studytime == 1 && doc.G3 > 18)
		emit(doc);
		}
		//Reduce

Најди ги учениците кои учат од 5 до 10 часа, а воедно консумираат минимални количини на алкохол, како за викенди, така и во текот на работната недела.	db.students.find({$and: [{studytime: {$eq: 3}}, {Dalc: {$eq: 1}}, {Walc: {$eq: 1}}]})	//Map
		function (doc) {
		if(doc.studytime == 3 && doc.Dalc == 1 && doc.Walc == 1)
		emit(doc);
		}
		//Reduce

Најди ја просечната финална оценка.	db.students.aggregate([{$group: { _id: null, "Average math grade": {$avg: "$G3"}}}])	//Map
		function (doc) {
		  if(doc.G3)
		  emit('grade', doc.G3);
		}
		//Reduce
		function (keys, values, rereduce) {
		  return sum(values)/values.length
		}

Најди ја просечната оценка од прво полугодие по португалски јазик, кај учениците од училиштето Gabriel Pereira.	db.students.aggregate([{$match: {$and: [{school: "GP"}, {subject: "portuguese"}]}},{$group: { _id: null, "Average portuguese first period grade": {$avg: "$G1"}}}])	//Map
		function (doc) {
		  if(doc.school == 'GP' && doc.subject == "portuguese")
		  emit('grade', doc.G1);
		}
		//Reduce
		function (keys, values, rereduce) {
		  return sum(values)/values.length
		}

Најди го бројот на ученици кои како старател имаат друга личност освен мајка и татко, а имаат крајна оценка по математика поголема од 17.	db.students.countDocuments({$and: [{guardian: "other"}, {subject: "math"}, {G3: {$gt: 17}}]})	//Map
		function (doc) {
		  if(doc.guardian == 'other' && doc.subject == 'math' && doc.G3 > 17)
		  emit(doc);
		}
		//Reduce
		_count

Кои се учениците кои имаат паднато еден предмет, а потоа по истиот успеале да добијат оценка на прво полугодие поголема или едаква на 13, а крајна оценка поголема или еднаква на 15?	db.students.find({$and: [{failures: {$eq: 1}}, {G1: {$gte: 13}}, {G3: {$gte: 15}}]})	//Map
		function (doc) {
		  if(doc.failures == 1 && doc.G1 >= 13 && doc.G3 >= 15)
		  emit(doc);
		}
		//Reduce

Која е највисоката оценка по математика во училиштето Mousinho de Silveira, кај учениците кои имаат повеќе од 10 изостаноци?	db.students.aggregate([{$match: {$and: [{school: "MS"}, {subject: "math"}, {absences: {$gt: 10}}]}},{$group: { _id: null, "Highest math grade": {$max: "$G3"}}}])
	[{ _id: null, 'Highest math grade': 10 }]
		//Map
		function (doc) {
		  if(doc.school == 'MS' && doc.subject == 'math' && doc.absences>10)
		  emit('grade', doc.G3);
		}
		//Reduce
		_stats

Која е највисоката оценка по португалски јазик во училиштето Gabriel Pereira, кај учениците кои имаат повеќе од 20 изостаноци?	db.students.aggregate([{$match: {$and: [{school: "GP"}, {subject: "portuguese"}, {absences: {$gt: 20}}]}},{$group: { _id: null, "Highest portuguese grade": {$max: "$G3"}}}])
	[{ _id: null, 'Highest portuguese grade': 16 }]
		//Map
		function (doc) {
		  if(doc.school == 'GP' && doc.subject == 'portuguese' && doc.absences > 20)
		  emit('grade', doc.G3);
		}
		//Reduce
		_stats

Дали има ученици кои доаѓаат од рурални средини, нивните родители имаат завршено средно	db.students.find({$and: [{address: "R"},{Medu: {$eq: 3}}, {Fedu: {$eq: 3}}, {traveltime: {$eq: 3}}, {paid: "no"}, {G3: {$gt: 11}}]})	//Map
		function (doc) {
		  if(doc.address == 'R' && doc.Medu == 3 && doc.Fedu == 3 && doc.traveltime == 3 && doc.paid == 'no' && doc.G3 > 11)

образование, патуваат повеќе од половина час, не земаат дополнителни часови, а имаат крајна оценка поголема од 11 по еден од двата предмети?		emit(doc);
		}
		//Reduce

Која е највисоката оценка независно од предметот и училиштето на учениците кои немаат интернет, имаат лоши фамилијарни односи, имаат повеќе од 15 отсуства и консумираат поголеми количини на алкохол за време на викенд?	db.students.aggregate([{$match: {$and: [{internet: "no"}, {famrel: {$gte: 3}}, {absences: {$gt: 15}}, {Walc: {$gte: 3}}]}},{$group: { _id: null, "Highest grade": {$max: "$G3"}}}])	//Map
		function (doc) {
		if(doc.internet == 'no' && doc.famrel >= 3 && doc.absences > 15 && doc.Walc >= 3)
		emit('grade', doc.G3);
		}
		//Reduce
		_stats

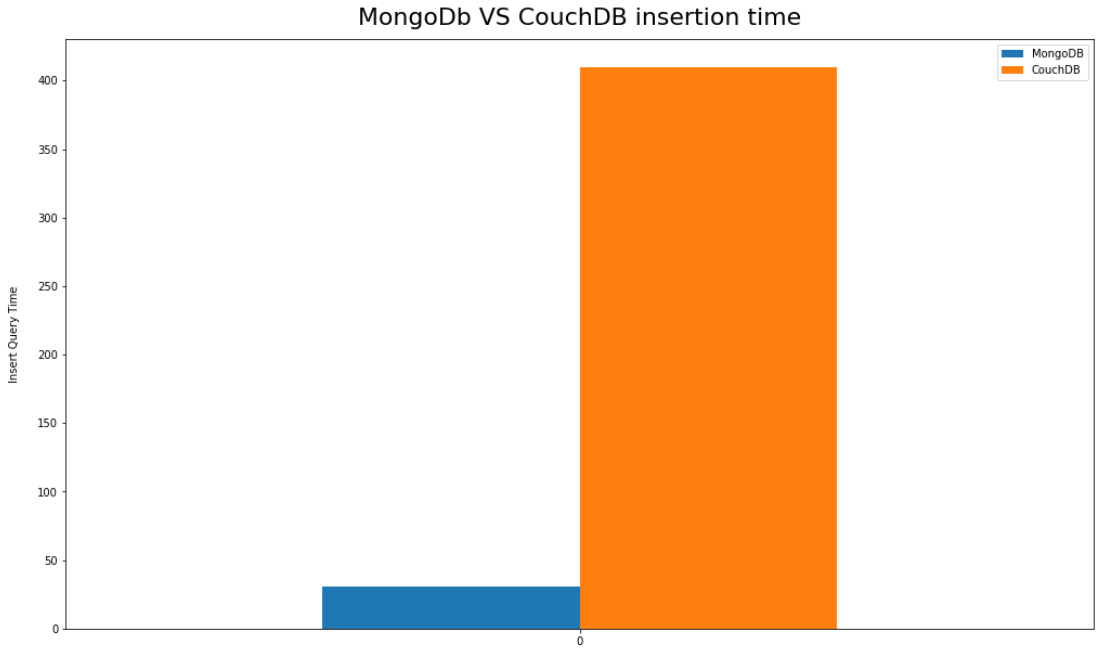
- На 9то прашање, има само еден студент кој ги исполнува критериумите, а на 10то прашање, највисоката оценка е 7.

Процесот на импортирање на податоците е доста едноставен и кај двете бази на податоци, меѓутоа, пишувањето на прашалниците е во голема мера поедноставно и поинтуитивно кај MongoDB. Ова можеме да го заклучиме и од самата табела, прикажана погоре. Иако CouchDB нуди многу моќни начини на комбинирање и филтрирање на податоците, како и полесна скалабилност, сепак при пишување на прашалниците се соочуваме со препреки и предизвици, кои не се случај при користењето на MongoDB. MongoDB е синтаксички, како и логички многу поблиска до SQL базите на податоци, со кои поголем дел од луѓето се запознати, меѓутоа при одреден период на користење, воопшто не е тешко да се увидат предностите во агрегација кои ги нуди CouchDB.

Резултати

MongoDB и CouchDB, како два претставници на document-based неструктурираните бази на податоци се во голема мера слични и ги овозможуваат истите манипулации и анализи со множествата. Сепак, тие имаат и големи разлики, како во начинот на пишување и логиката при самите прашалници, така и во начините на импортирање, а најзначано и во времињата на извршување.

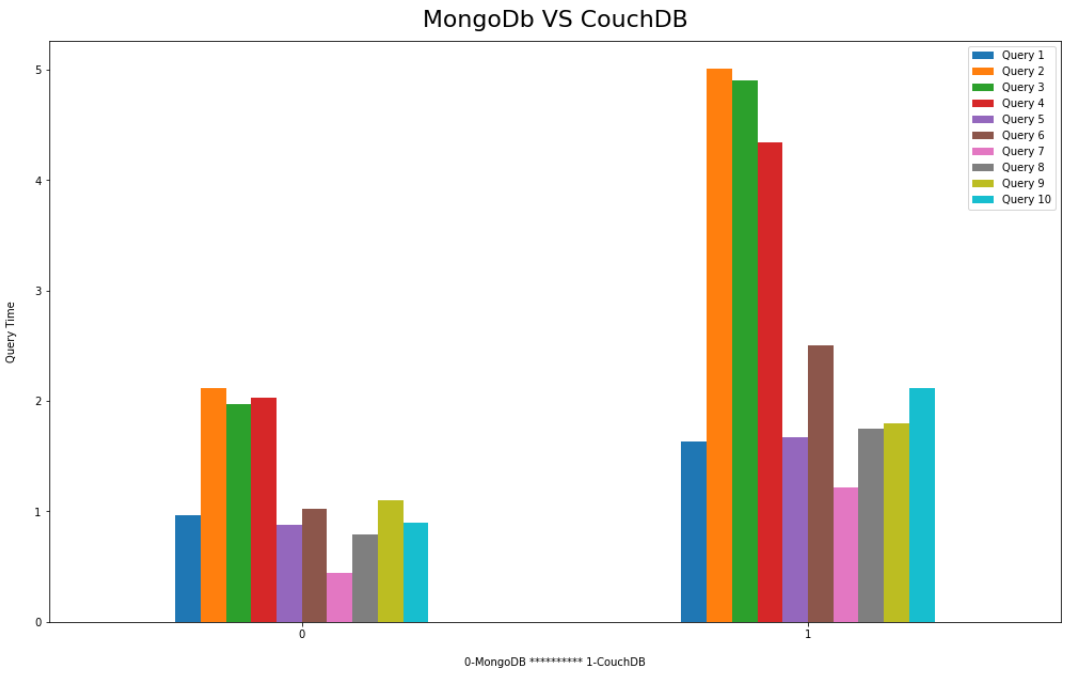
На хистограмот во продолжение можеме да ги погледнеме времињата на импортирање на податоците во двете бази на податоци, па да направиме споредба.



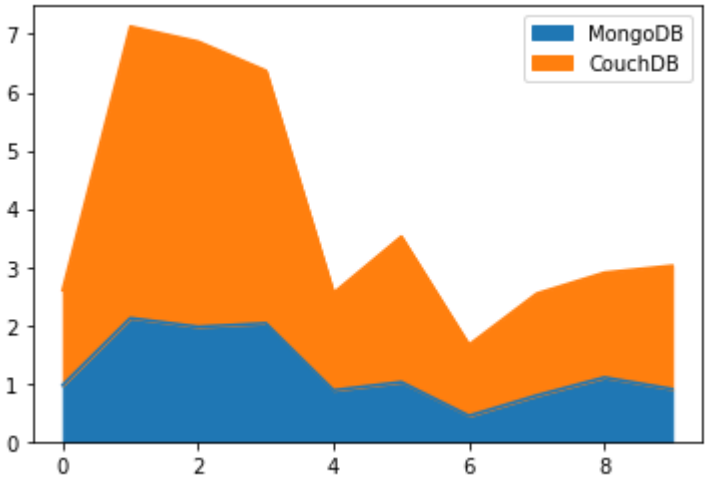
Слика 77. "Времиња на импортирање на податоците во двете бази на податоци"

Од самиот хистограм очигледно е дека времето на импортирање на податоците во MongoDB е речиси десетина пати помало од времето потребно за импортирање на податоците во CouchDB.

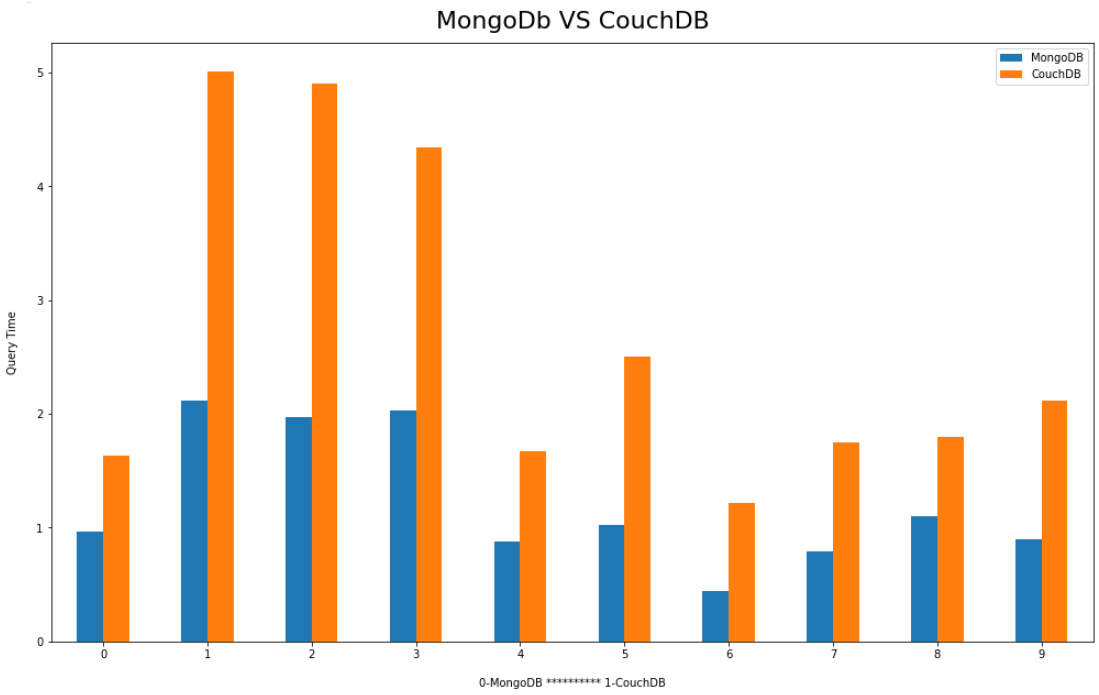
Следно, да преминеме кон анализа на времињата на извршување на прашалниците и споредба на перформансите на двете бази на податоци. За оваа цел, можеме да ги погледнеме графиконите кои се дадени во продолжение.



Слика 78. "Споредба на времиња на извршување на прашалниците"



Слика 79. "Споредба на времињата на извршување на прашалниците"



Слика 80. "Споредба на времињата на извршување на прашалниците"

По направените визуелизации, а воедно и при самиот процес на извршување на прашалниците можевме да забележиме дека има доста голема разлика во потребното време на извршување на прашалниците на двете неструктурирани бази на податоци. Имено, MongoDB е евидентно подобар од гледна точка на брзина на извршување на различните операции. Дополнително, со самото зголемување на нивото на агрегации кои ги користев во прашалниците, временската разлика помеѓу потребните времиња на извршување кај двете неструктурирани бази на податоци стануваше се поголема. Односно, со секоја покомплицирана агрегација, времето на извршување на CouchDB растеше многу побрзо, во споредба со времето на извршување на MongoDB, кое многу побавно растеше.

Заклучок

По завршената анализа на податочното множество, како и по резултатите добиени од самите прашалници, со сигурност можеме да кажеме дека околина на живеење, односите на луѓето, како и нивната мотивација имаат значајна улога во нивните перформанси, како во средно училиште, така и во секоја друга сфера од животот.

Секојдневно сме сведоци на личности со огромен потенцијал, кој не може да дојде до израз од непознати причини, како и на личности кои се соочуваат со проблеми, поради несоодветен избор на предмети и училишта, па не можат да блеснат во својот вистински сјај, поради грешниот избор.

Односите во семејството, образованието на родителите, како и нивниот работен статус, исто така се фактор и влијаат врз успехот на секое од децата. Дел од нив, постигнатиот успех на нивните родители го користат како мотивација, па така секојдневно се трудат и целат кон што поголеми успеси. Други пак, живеат во зоната на комфорот и го вложуваат само неопходното. Учениците, чиишто родители немаат високо образование, исто така дејствуваат на два начини. Кај некои од нив, се гледа желба и максимално вложување за успех, додека пак кај други преовладуа песимизмот.

Видовме дека и количеството на алкохол влијае врз перформансите на учениците, како и врз нивните отстуства и презапишувања на предмети.

Меѓутоа, видовме и дека независно од нивната средина, најголем дел од учениците имаат намера да продолжат со високо образовние, што укажува на фактот дека младите се во голема мера свесни за функционирањето на светот и секојдневно се трудат и инвестираат во зголемување на нивниот човечки капитал и потенцијал.

Од друга страна, ова истражување ни помогна да сфатиме колку е важен квалитетот, а не само квантитетот на податоците со кои работиме. Самиот процес на претпроцесирање на податоците и истражување на нивните корелации ни овозможи да извлечеме значајни информации за множеството, кои носат своја одредена бизнис вредност во реалниот свет. Слободната форма на податоците носи свои предизвици и препреки, меѓутоа деталната анализа и алатките кои ни ги нудат неструктурираните бази на податоци успешно се справуваат со истите.

Конкретно, двете неструктурирани бази на податоци CouchDB и MongoDB, се едни од најкористените и најпопуларните document-based бази на податоци. Доколку направиме осврт на имплементацијата, со сигурност можеме да кажеме дека MongoDB е доста поблиска до релационите бази на податоци, особено по синтакса, па со тоа е и полесна за користење од страна на поширокатa публика. Сепак, иако е покомлексна, CouchDB, нуди многу покомплексни и поелеганти решенија за голем дел од проблемите.

Од страна на перформанси, како што можевме да погледнеме и на визуелизациите, MongoDB со право го има статусот на најкористена неструктурирана база на податоци. Таа во голема мера е побрза, особено со зголемувањето на комплексноста на прашалниците. Односно, колку е поголема комплексноста, времето на раст на извршување на прашалницте многу побавно расте кај MongoDB, во споредба со CouchDB. Од друга страна пак, CouchDB нуди поддршка за iOS и Android уреди, што не е случај со MongoDB. Ова ни укажува на фактот дека двете бази имаат свои предности и недостатоци, како и на тоа дека се претставува компромис, па секоја од нив би доминирала во различни сценарија.

Направената анализа прави невозможно да не се помисли на Големите податоци (Big Data) коишто стануваат сè повеќе актуелни во секоја сфера, а за кои релационите бази на податоци не се соодветни, па неструктурираните бази на податоци доминираат во големи компании како што се Netflix, Amazon, Google и Facebook, кои секојдневно се стекнуваат со големо количество на разновидни и далеку од структурирани податоци. Меѓутоа, употребата на релациони бази на податоци е сеуште од големо значење, особено во банкарски системи, каде прецизноста и конзистентноста се од клучно значење.

Одовде доаѓаме до фактот дека не е се онака, како што изгледа на површина, односно, со самиот процес на претпроцесирање успеавме длабоко да навлеземе во сите фактори кои влијаат врз перформансите на учениците, кои обично ги занемаруваме и судиме без никакви информации. Истите се од огромно значење и треба да се искористат за мотивација на сите млади индивидуи, охрабрувајќи ги истите да ја изградат најдобрата верзија од самите себе.

Користена литература:

- MongoDB official documentation
- Simplelearn: MongoDB Full Course
- MongoDB tutorial
- The Battle of the NoSQL Databases - Comparing MongoDB and CouchDB
- CouchDB vs. MongoDB: What You Need to Know
- CouchDB official documentation
- Mark Grimes: CouchDB Tutorial
- Apache CouchDB
- CouchDB Tutorial
- CouchDB Fauxton
- Python CouchDB tutorial
- Difference between Couchdb and Mongodb
- Performance Evaluation of Query Response Time in The Document Stored NoSQL Database
- System Properties Comparison CouchDB vs. MongoDB
- Difference Between CouchDB vs MongoDB