

МОНГОЛ УЛСЫН ИХ СУРГУУЛЬ
ХЭРЭГЛЭЭНИЙ ШИНЖЛЭХ УХААН, ИНЖЕНЕРЧЛЭЛИЙН СУРГУУЛЬ
МЭДЭЭЛЭЛ, КОМПЬЮТЕРИЙН УХААНЫ ТЭНХИМ

Liu Xinyao

**Хиймэл оюун ухаанд сууриссан орон сууцны
өгөгдлийн дүн шинжилгээ**
(AI-based data analysis on Mongolian Real Estates)

Мэдээллийн технологи (D061303)
Бакалаврын судалгааны ажил

Улаанбаатар

2022 оны 04 сар

МОНГОЛ УЛСЫН ИХ СУРГУУЛЬ
ХЭРЭГЛЭЭНИЙ ШИНЖЛЭХ УХААН, ИНЖЕНЕРЧЛЭЛИЙН СУРГУУЛЬ
МЭДЭЭЛЭЛ, КОМПЬЮТЕРИЙН УХААНЫ ТЭНХИМ

**Хиймэл оюун ухаанд суурилсан орон сууцны өгөгдлийн дүн
шинжилгээ**

(AI-based data analysis on Mongolian Real Estates)

Мэдээллийн технологи (D061303)
Бакалаврын судалгааны ажил

Удирдагч: _____ Др. Б.Хуягбаатар

Хамтран удирдагч: _____

Гүйцэтгэсэн: _____ Liu Xinyao (18B1NUM0118)

Улаанбаатар

2022 оны 04 сар

Зохиогчийн баталгаа

Миний бие Liu Xinyao ”Хиймэл оюун ухаанд суурилсан орон сууцны өгөгдлийн дүн шинжилгээ” сэдэвтэй судалгааны ажлыг гүйцэтгэсэн болохыг зарлаж дараах зүйлсийг баталж байна:

- Ажил нь бүхэлдээ эсвэл ихэнхдээ Монгол Улсын Их Сургуулийн зэрэг горилохоор дэвшүүлсэн болно.
- Энэ ажлын аль нэг хэсгийг эсвэл бүхлээр нь ямар нэг их, дээд сургуулийн зэрэг горилохоор оруулж байгаагүй.
- Бусдын хийсэн ажлаас хуулбарлаагүй, ашигласан бол ишлэл, зүүлт хийсэн.
- Ажлыг би өөрөө (хамтарч) хийсэн ба миний хийсэн ажил, үзүүлсэн дэмжлэгийг дипломын ажилд тодорхой тусгасан.
- Ажилд тусалсан бүх эх сурвалжид талархаж байна.

Гарын үсэг: _____

Огноо: _____

ГАРЧИГ

УДИРТГАЛ	1
1. СЭДВИЙН ТАНИЛЦУУЛГА	2
1.1 Оршил	2
1.2 Зорилго	2
1.3 Зорилт	2
1.4 Системийн танилцуулга	3
2. ИЖИЛ СИТЕМИЙН СУДАЛГАА	4
2.1 Ижил төстэй системүүд	4
2.2 Бүлгийн дүгнэлт	6
3. ХОЛБОГДОХ ОНОЛЫН СУДАЛГАА	7
3.1 Data scraping судалгаа	7
3.2 Python судалгаа	7
3.3 Google colab судалгаа	8
3.4 BeautifulSoup судалгаа	9
3.5 Seaborn data visualization судалгаа	10
3.6 Sentence Transformers судалгаа	11
3.7 Linear Regression судалгаа	12
3.8 SentenceBert судалгаа	13
3.9 Бүлгийн дүгнэлт	14
4. СИТЕМИЙН ШИНЖИЛГЭЭ ЗОХИОМЖ	15
4.1 Өгөгдлийн сангийн диаграм	15
4.2 Өгөгдлийн сангийн хүснэгтүүдийн тайлбар	16
4.3 Өгөгдлийн сангийн холбоосын тайлбар	19
5. ХЭРЭГЖҮҮЛЭЛТ, ҮР ДҮН	20
5.1 Хэрэгжүүлсэн байдал	20

5.2 Бүлгийн дүгнэлт	37
ДҮГНЭЛТ	38
НОМ ЗҮЙ	38
ХАВСРАЛТ	39
A. БАКАЛАВРЫН СУДАЛГААНЫ АЖЛЫН ҮЕЧИЛСЭН ТӨЛӨВЛӨГӨӨ	40
B. КОДЫН ХЭРЭГЖҮҮЛЭЛТ	41

ЗУРГИЙН ЖАГСААЛТ

2.1 Америкчууд хэрхэн хооллодог өгөгдлийн дүрслэлийн жишээ	4
2.2 Өндөг хэрхэн хэлбэрээ олж авдаг өгөгдлийн дүрслэлийн жишээ	5
2.3 WHO bot-ийн жишээ	6
3.1 Data scraping үйл явц	7
3.2 Python лого	8
3.3 Google colab лого	9
3.4 BeautifulSoup	10
3.5 Seaborn лого	11
3.6 SBERT.net лого	12
3.7 Linear Regression лого	13
3.8 SentenceBERT-ийн хос сүлжээний архитектур	14
4.1 Өгөгдлийн сангийн диаграм	15
5.1 Судалгааны ажлын үйл явцын диаграм-1	20
5.2 Судалгааны ажлын үйл явцын диаграм-2	21
5.3 Хуудасны линкуудыг татаж буй үйл явц ба үр дүн	22
5.4 Орон сууцны өгөгдлүүдийг татаж буй үйл явц ба үр дүн	23
5.5 Орон сууцны өгөгдөл	24
5.6 Өгөгдлийн өөрчлөлт	25
5.7 Data cleaning хийсэн дараах өгөгдөл	26
5.8 Өгөгдлийн статистик-1	28
5.9 Өгөгдлийн статистик-2	29
5.10 Өгөгдлийн статистик-3	30
5.11 Өгөгдлийн статистик-4	31
5.12 Өгөгдлийн статистик-5-1	32
5.13 Өгөгдлийн статистик-5-2	32

5.14	Өгөгдлийн статистик-6	33
5.15	Өгөгдлийн статистик-7	34
5.16	Өгөгдлийн статистик-8	35
5.17	Өгөгдлийн статистик-9	36
5.18	Өгөгдлийн статистик-10	37
A.1	Бакалаврын судалгааны ажлын үучилсэн төлөвлөгөө.....	40

ХҮСНЭГТИЙН ЖАГСААЛТ

4.1	Зарын өгөгдлийн мэдээлэл	16
4.2	Хэрэглэгчийн өгөгдлийн мэдээлэл	18
4.3	Зургийн өгөгдлийн мэдээлэл	18
4.4	Байршилын өгөгдлийн мэдээлэл	19
5.1	Орон сууцны өрөөний мэдээллүүд	27

Кодын жагсаалт

B.1	Data Scraping эх код	41
B.2	Data Cleaning эх код	43

УДИРТГАЛ

Компьютер болон ухаалаг гар утасны хэрэглээ өсөн нэмэгдэж, техник технологи эрчимтэй хөгжиж буй өнөөгийн нийгэмд мэдээлэл харилцаа холбооны технологи хөгжихийн хэрээр бидний өдөр тутмын үйл ажиллагаа ч энэ бүгдээс хамааралтай болж байгаа билээ. Албан газар, байгууллага бүр үйл ажиллагаа эхэлсэн цагаас эхлэн өөрсдинй гэсэн дата-г бий болгодог. Харин эдгээр дата-гаа хэрхэн ашиглах вэ? гэсэн асуулт байгууллага бүрд байдаг. Тухайн дата-г ашиглан тодорхой нэг зүйл дээр ямар нэгэн харьцуулалт хийж хүссэн мэдээллийг олж авах боломжтой.

Unegui.mn зарын сайтаас орон сууцны мэдээллийн өгөгдлүүдийг Data Scraping хийж үүний дараагаар цуглуулсан дата дээр үндэслээд орон сууцны өгөгдлийн харьцуулалт хийж болох ба эдгээр харьцуулалт дээр ямар нэгэн дүн шинжилгээ хийж, хэрэглэгчид аль болох бага хугацааны дотор өөрийн хэрэгтэй мэдээллийг сонирхолтой байдлаар олж авахыг зорьсон болно.

1. СЭДВИЙН ТАНИЛЦУУЛГА

1.1 Оршил

Энэхүү бакалаврын судалгааны ажлын хүрээнд сонгож авсан сэдэв маань ”Хиймэл оюун ухаанд суурилсан орон сууцны өгөгдлийн дүн шинжилгээ” бөгөөд энэ нь орон сууцны мэдээллийг тавьсан зарын дагуу өгөгдүүдлийг цуглуулж, дүн шинжилгээ хийж өгөгдлийн дурслэлээр мэдээллүүдийг харуулах.

1.2 Зорилго

Unegui.mn зар дээрх тавигдсан орон сууцны мэдээллийн өгөгдлүүдийг ашиглан хиймэл оюун ухаан дээр суурьлаж дүн шинжилгээ хийж өгөгдлүүдийг нэгтгэн харьцуулж график дурслэлээр мэдээллүүдийг харуулах мөн Python хэл дээр суурилан BotFramework-ын үүрэг бүтэц, хэрэглээг судалж Google Colab дээр ажиллахуйц хэрэглэгчийн туслах Чатбот системийн хөгжүүлэлтийг хийж орон сууцтай холбоотой асуултанд хариулна.

1.3 Зорилт

Дээрх зорилгод хүрэхийн тулд дараах зорилтуудыг тавьсан. Үүнд:

- Орон сууцны мэдээллийн өгөгдлүүдийг цуглуулах
- Ижил төстэй системийн судалгаа хийх
- Холбогдох онол болон ашиглагдах технологийн судалгаа хийх
- Системийн шинжилгээ ба зохиомж хийх
- Хөгжүүлэлтийг хийх

- Үр дүнгийг харуулах

1.4 Системийн танилцуулга

Судалгааны ажлын хүрээнд зарын сайтаас орон сууцны мэдээллийн өгөгдлүүдийг татаж графикаар өгөгдлийн статистик байгуулж гарсан үр дүн дээр анализ дүгнэлт бичнэ. Энэ нь Seaborn data visualization-ийг ашиглан өгөгдлүүдийн харьцуулалтуудыг их сонирхолтой байдлаар харуулах боломжийг олгосон. Дараа нь орон сууцны мэдээллийн талаарх хайж буй хэрэглэгчийн туслах Чатбот системийг хөгжүүлнэ. Нээлттэй өгөгдлийн санд буй орон сууцны нэр, үнэ, байршил, ашиглалтанд орсон он, давхар, лизингээр авах боломж зэрэг хүснэгтүүдийг нэгтгэж үүнээсээ хэрэглэгчидэд тоон хариулт өгөх үндсэн зарчимтай болно.

Хэрэглэгчид ”Саруул хотхоны 2 өрөөтэй байр ямар үнэтэй байгаа вэ?” гэсэн асуулт тулгарлаа гэж бодоход Bot системээс тухайн асуултыг асууснаар Саруул хотхонд байгаа 2 өрөөтэй байрны үнийг харах боломжтой. Энэ Bot системийг Sentence Transformers-ийг ашиглан хэрэглэгчийн оруулсан асуулт өгүүлбэрийг таамаглан ойлгож түүнд тохирсон тоон векторуудыг хариулдаг. Харин эдгээр тоон хариултуудыг илүү оновтой болгохын тулд Linear Regression-ийг ашиглаад Sentence Transformers дээр гарсан тоон утгуудыг харьцуулж оруулсан өгүүлбэртэй хамгийн төстэй буюу хамгийн ойр байгаа тоог олж өгөх боломжтой юм.

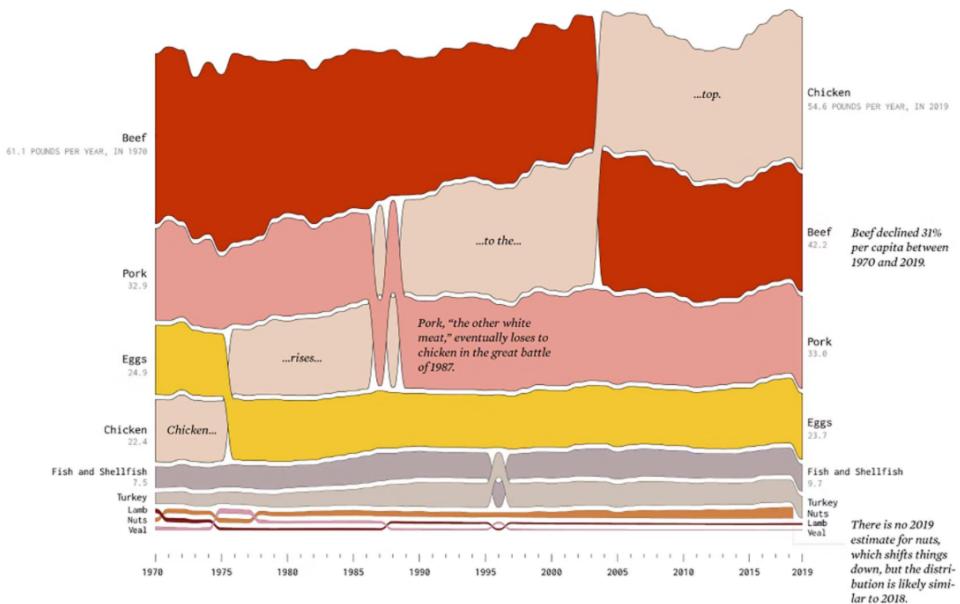
2. ИЖИЛ СИТЕМИЙН СУДАЛГАА

2.1 Ижил төстэй системүүд

2.1.1 Өгөгдлийн дүрслэлийн жишээ

Өгөгдлийн дүрслэл нь Excel хүснэгтүүдийн боломжоос гадна өгөгдлийг хөгжилтэй, бүтээлч байдлаар дүрслэх урлаг юм. Мэдээллийн дүрслэлийг янз бүрийн салбарт тайлагнах гэх мэт мэргэжлийн хүрээнд ихэвчлэн ашигладаг боловч зарим дүрслэл нь поп соёл, өдөр тутмын сэдэвтэй холбоотой өгөгдлийг харуулдаг. Жишээ нь:

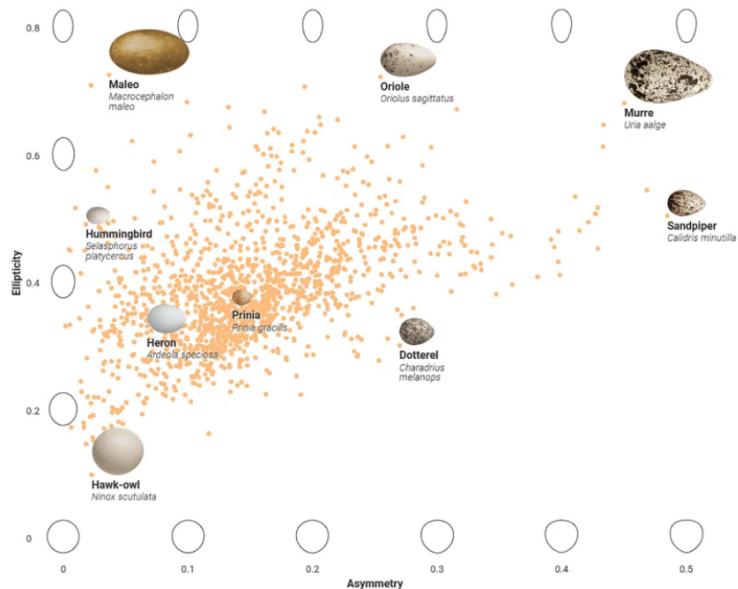
1. Энэхүү өгөгдлийн дүрслэлийн жишээ нь американчуудын үндэсний хэмжээнд хэрхэн



Зураг 2.1: Америкчууд хэрхэн хооллодог өгөгдлийн дүрслэлийн жишээ

иддэгийг тооцоолох, дүрслэх зорилгоор USDA-аас гаргасан хүнсний олдоцын талаарх мэдээллийг ашигладаг. 1970-2019 он хүртэл уургийн гол эх үүсвэрийн нэг хүнд ногдох жилийн фунт хэрхэн өөрчлөгдсөнийг дээрх графикаас харж болно.

2. Эрдэмтэд янз бүрийн шувуудын өндөгний хэлбэр яагаад өөр байдгийг саяхан олж



Зураг 2.2: Өндөг хэрхэн хэлбэрээ олж авдаг өгөгдлийн дурслэлийн жишээ

мэдсэн. Энэхүү судалгаа нь сүүлийн 100 жилийн хугацаанд цуглувансан бараг 50,000 шувууны өндөгний мэдээллийг цуглувсан. Өндөгний хэмжээсийг 1400 зүйлээр дүрсэлсэн бөгөөд энэ график нь тэгш бус байдал (цэгц) ба эллипс (төгс бөмбөрцөгөөс хазайх) хоорондын хамаарлыг харуулж байна [11].

2.1.2 Bot системийн жишиээ

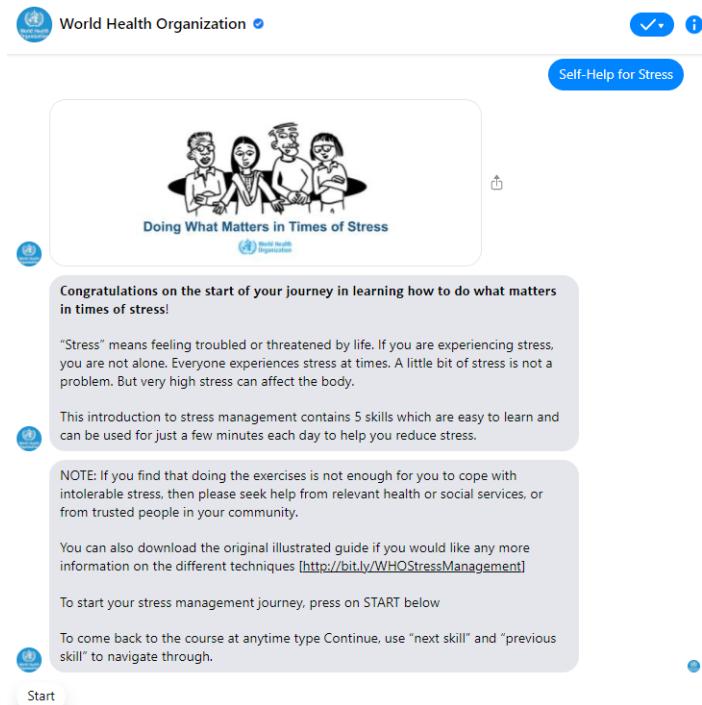
World Health Organization's WHO Health Alert

Коронавирусын цар тахал газар авч байгаатай холбогдуулан дэлхийн өнцөг булан бүрт байгаа хүмүүс эрүүл мэндийн талаархи албан ёсны, найдвартай мэдээлэл, зөвлөгөө өгөх зорилготой бот юм. Энэхүү бот нь Коронавирусын талаар олон нийтийн сонирхсон асуултанд хариулж дэлхийн өнцөг булан бүрт 24 цагийн турш шуурхай, найдвартай, хамгийн сүүлийн үеийн албан ёсны мэдээллийг өгдөг.

Түүнчлэн ДЭМБ-ын Эрүүл мэндийн сэрэмжлүүлэг буюу өөрийгөө халдвараас хэрхэн

хамгаалах, аяллын зөвлөгөө өгөх, Коронавирусын домгийг устгах зэрэг сэдвээр албан ёсны мэдээлэл өгөх болно [12].

Стрессээс өөрийгөө хамгаалах гэж сонгосноор ”COVID-19” цар тархалтын үеэр стрессээс гарах арга замуудын мэдээллийг ”Start” гэж дарснаар харж болно.



Зураг 2.3: WHO bot-ийн жишээ

2.2 Бүлгийн дүгнэлт

Энэ бүлгийн хүрээнд судалгааны ажлын сэдэвтэй төстэй жишээ болон системүүдийн талаарх судалгаа хийсэн болно.

3. ХОЛБОГДОХ ОНОЛЫН СУДАЛГАА

3.1 Data scraping судалгаа

Data scraping буюу web scraping нь вэб сайтаас мэдээллийг таны компьютер дээр хадгалагдсан хүснэгт эсвэл дотоод файл руу импортлох үйл явц юм. Энэ нь вэбээс мэдээлэл авах, үнийн өөрчлөлтийг онлайнаар хянах, үнийн харьцуулалт хийх, зарим тохиолдолд өөр вэбсайт руу дамжуулах, өрсөлдөгчид өөрсдийн вэбсайтаас мэдээлэл авах замаар хэр сайн ажиллаж байгааг харахад ашиглагддаг хамгийн үр дүнтэй аргуудын нэг юм [1].



Зураг 3.1: Data scraping үйл явц

Дипломын ажлын хүрээнд inegui.mn сайтаас орон сууцны мэдээллийн өгөгдлүүдийг data scraping хийж судалгааны ажлын орон сууцны өгөгдлийн дүн шинжилгээ хийх хамгийн эхний үе шат болно.

3.2 Python судалгаа

Python бол олон парадигмтай, ерөнхий зориулалттай, өндөр түвшний програмчлалын хэл юм. Python нь программистуудад өөр өөр програмчлалын хэв маягийг ашиглан энгийн эсвэл нарийн төвөгтэй програмуудыг үүсгэж, илүү хурдан үр дунд хүрч, бараг хүний хэлээр ярьж байгаа мэт код бичих боломжийг олгодог юм. Python нь модуль болон багцуудыг дэмждэг

хэл бөгөөд энэ нь программын модульчлагдсан байдал, кодыг дахин ашиглахыг дэмждэг болно. Python interpreter болон өргөн хүрээний стандарт сан нь бүх томоохон платформд төлбөргүй эх эсвэл хоёртын хэлбэрээр байдаг бөгөөд чөлөөтэй ашиглах боломжийг олгодог [2].



Зураг 3.2: Python лого

Судалгааны ажлын хүрээнд python хэлээр unegui.mn сайтаас орон сууцны мэдээллийн өгөгдлүүдийг data scraping хийх код бичиж ашигласан болно.

3.3 Google colab судалгаа

Colabatory буюу товчоор ”Colab” нь Google Research-ийн бүтээгдэхүүн юм. Colab нь хөтчөөр дамжуулан дурын python код бичиж, ажиллуулах боломжийг хэн бүхэнд олгодог бөгөөд ялангуяа машин сургалт, өгөгдөл дүн шинжилгээ хийх, боловсрол олгоход маш тохиромжтой байдаг. Colab дэвтэр нь зураг, HTML, LaTeX болон бусад зүйлсийн хамт нэг баримт бичигт гүйцэтгэх код, баялаг текстийг нэгтгэх боломжийг олгодог. Та өөрийн Colab дэвтэр үүсгэх үед таны Google Drive бүртгэлд хадгалагдана. Та Colab дэвтэрээ хамтран ажиллагсад эсвэл найзуудтайгаа хялбархан хуваалцаж, тэмдэглэлийн дэвтэр дээрээ сэтгэгдэл бичих эсвэл бүр засварлах боломжтой [3].



Зураг 3.3: Google colab лого

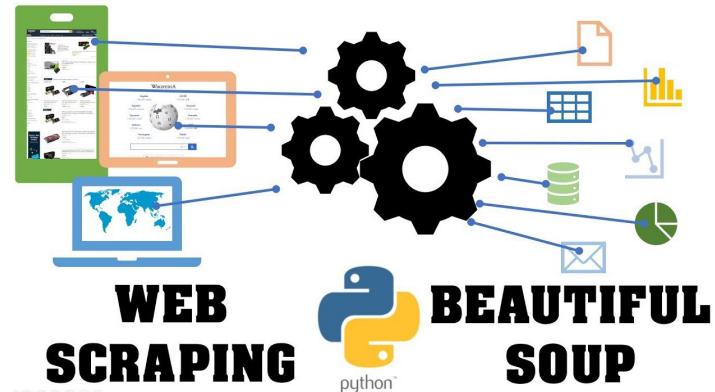
Судалгааны ажлын хүрээнд Google colab дээр python хэлээр data scraping код бичиж орон сууцны мэдээллийн тодорхой өгөгдлүүдийг татах ажлыг гүйцэтгэсэн болно. Google colab-ийн нэг давуу тал бол бичсэн кодыг ажиллах хооронд дараагийн хэсгийн кодыг бичих боломжийг олгодог.

3.4 BeautifulSoup судалгаа

BeautifulSoup нь HTML болон XML баримт бичгүүдийг задлан шинжлэхэд зориулагдсан Python сан юм. Энэ нь HTML-ээс өгөгдлийг задлахад ашиглаж болох ба задлан шинжлэх модыг үүсгэдэг бөгөөд энэ нь web scraping хийхэд их тустай байдаг [4].

Beautiful Soup-ийг өвөрмөц болгодог зарим гол шинж чанарууд:

- Beautiful Soup нь задлан шинжлэх модыг чиглүүлэх, хайх, өөрчлөхөд зориулсан цөөн хэдэн энгийн аргууд болон Pythonic хэлц үгсээр хангадаг.
- Beautiful Soup нь автоматаар ирж буй баримтуудыг Юникод руу, гарч буй баримтуудыг UTF-8 болгон хувиргадаг.



Зураг 3.4: BeautifulSoup

- BeautifulSoup нь lxml болон html5lib зэрэг алдартай Python задлагчдын дээр байрладаг бөгөөд энэ нь бидэнд уян хатан байдлын үүднээс өөр өөр задлан шинжлэх стратеги эсвэл ажиллах хурдыг туршиж үзэх боломжийг олгодог.

Судалгааны ажлын хүрээнд python хэлний багцаас BeautifulSoup-ийг ашиглан data scraping хийж орон сууцны мэдээллийн өгөгдлүүдийг цуглувулсан.

3.5 Seaborn data visualization судалгаа

Seaborn бол matplotlib дээр суурилсан Python-ын өгөгдөл дүрслэх сан юм. Энэ нь сэтгэл татам, мэдээлэл сайтай статистик график зурах өндөр түвшний интерфейсээр хангадаг.

Seaborn нь өгөгдлийг судалж, ойлгоход их хялбар байдаг. Түүний график функциуд нь бүхэл өгөгдлийн багцыг агуулсан data фрейм болон массив дээр ажиллаж, мэдээллийн график үүсгэхийн тулд шаардлагатай семантик зураглал, статистикийн нэгтгэлийг дотооддоо гүйцэтгэдэг. Seaborn-ийн өгөгдлийн багцад чиглэсэн, тунхаглалын API нь хэрхэн зурах тухай нарийн ширийн зүйлээс илүүтэйгээр талбайн дээрх элементүүд нь ямар утгатай болохыг анхаарч үзэх боломжийг олгодог [5].

Түүний онцлог:

- Seaborn бол статистикийн графикийн номын сан юм.
- Энэ нь үзэсгэлэнтэй анхдагч хэв маягтай.
- Энэ нь Pandas dataframe объектуудтай маш сайн ажиллахад зориулагдсан.



Зураг 3.5: Seaborn лого

Судалгааны ажлын хүрээнд Seaborn-ийг ашиглаад орон сууцны өгөгдлийн анализийг график дүрслэлээр харуулсан болно.

3.6 Sentence Transformers судалгаа

Sentence Transformers нь хамгийн сүүлийн үеийн өгүүлбэр, текст болон дүрс оруулахад зориулагдсан Python framework юм. Энэхүү framework-ийг ашиглан 100 гаруй хэлний өгүүлбэр болон текст оруулгыг тооцоолох боломжтой. Дараа нь эдгээр оруулгыг харьцуулж болно, жишээлбэл, косинус-төст байдлаар ижил утгатай өгүүлбэрүүдийг олох боломжтой. Энэ нь семантик текстийн ижил төстэй байдал, семантик хайлт эсвэл задлан олборлолтод хэрэгтэй байдаг.

Энэхүү framework нь PyTorch болон Transformers дээр суурилсан бөгөөд янз бүрийн даалгаварт тохирсон, урьдчилан бэлтгэгдсэн загваруудын томоохон цуглуулгыг санал болгодог. Цаашилбал,

өөрийн загвараа нарийн тааруулахад хялбар байдаг [7].



Зураг 3.6: SBERT.net лого

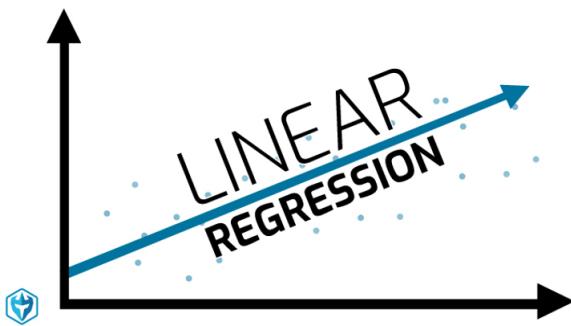
Судалгааны ажлын хүрээнд Sentence Transformers-ийг ашиглаад оруулсан өгүүлбэрүүдийг вектор тоон утга руу хувиргаж семантик хайлт хийх боломжийг олгосон болно.

3.7 Linear Regression судалгаа

Linear Regression буюу шугаман регрессийн загвар нь шулуун шугамыг ашигладаг бол логистик болон шугаман бус регрессийн загвар нь муруй шугамыг ашигладаг. Regression нь бие даасан хувьсагч (хувьсагч) өөрчлөгдөхөд хамааралтай хувьсагч хэрхэн өөрчлөгдөхийг тооцоолох боломжийг олгодог [6].

Хоёр тоон хувьсагчийн хоорондын хамаарлыг тооцоолоход энгийн шугаман регрессийг ашигладаг. Дараах мэдээллүүдийг мэдэхийг хүсвэл энгийн шугаман регрессийг ашиглаж болно.

- Хоёр хувьсагчийн хоорондын хамаарал хэр хүчтэй вэ (жишээлбэл, хур тунадас, хөрсний элэгдэл хоорондын хамаарал).
- Бие даасан хувьсагчийн тодорхой утга дахь хамааралтай хувьсагчийн утга (жишээлбэл, хур тунадасны тодорхой түвшинд хөрсний элэгдлийн хэмжээ).



Зураг 3.7: Linear Regression лого

Энгийн шугаман регресс нь параметрийн тест бөгөөд энэ нь өгөгдлийн талаар тодорхой таамаглал дэвшүүлдэг гэсэн үг юм. Эдгээр таамаглалууд нь:

- Вариацын нэгэн төрлийн байдал
- Ажиглалтын бие даасан байдал
- Нормал байдал

Судалгааны ажлын хүрээнд Linear Regression- ийг ашиглаад Sentence Transformers дээр гарсан тоон утгуудыг харьцуулж оруулсан өгүүлбэртэй хамгийн төстэй буюу хамгийн ойр байгаа тоог олж өгөх боломжийг олгосон болно.

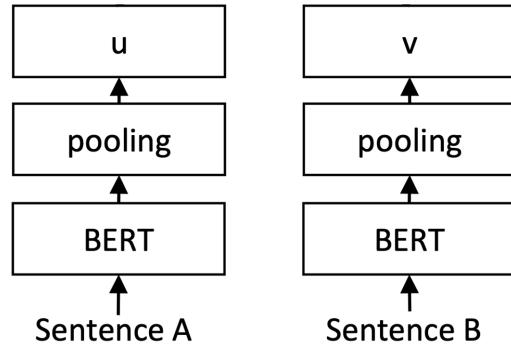
3.8 SentenceBert судалгаа

BERT нь 'Bidirectional Encoder Representations from Transformers' буюу Transformers-aac хоёр чиглэлт кодлогчийн төлөөлөл гэсэн үг бөгөөд 3.3 сая англи үгээр бэлтгэгдсэн хэлний дүрслэлийн загвар юм. BERT болон хэлний загваруудын өмнөх хувилбаруудын хоорондох асар том ялгаа бол BERT нь үг хэрэглэж буй нөхцөл байдлыг ”ойлгодог” явдал юм.

Sentence-BERT (SBERT)

SBERT нь хоёр өгүүлбэрийг нэгэн зэрэг боловсруулах боломжийг олгодог хос сүлжээ гэж

нэрлэгддэг сүлжээ юм. Эдгээр хоёр ихэр нь параметр бүрээрээ ижилхэн байдаг (жин нь холбоотой) бөгөөд энэ нь архитектурыг олон удаа ашигласан нэг загвар гэж үзэх боломжийг бидэнд олгодог [8].



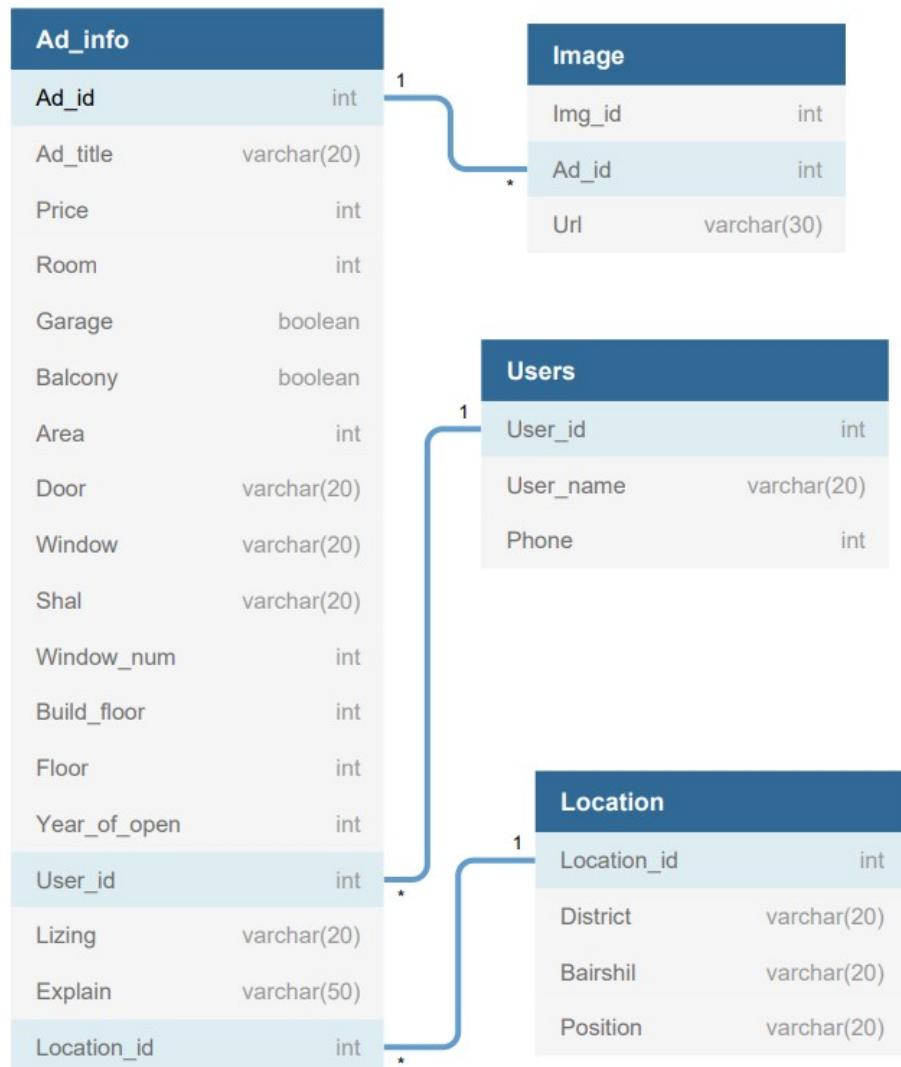
Зураг 3.8: SentenceBERT-ийн хос сүлжээний архитектур

3.9 Бүлгийн дүгнэлт

Энэ бүлгийн хүрээнд судалгааны ажлыг бүрэн хийж хэрэгжүүлэхийн тулд ашиглахад тохиромжтой технологиудыг сонгон онол болон хэрэгжүүлэлтийн талаар судалсан ба уг технологийг хэрхэн яаж ашигласан тухай тайлбарлан бичсэн болно.

4. СИСТЕМИЙН ШИНЖИЛГЭЭ ЗОХИОМЖ

4.1 Өгөгдлийн сангийн диаграм



Зураг 4.1: Өгөгдлийн сангийн диаграм

4.2 Өгөгдлийн сангийн хүснэгтүүдийн тайлбар

4.2.1 Зарын мэдээлэл

Энэ хүснэгтэнд зар дээр тавигдсан орон сууцны мэдээллийг агуулна. Мэдээллээс гадна зар бүрт харгалзах дугаартай байна.

Table 4.1: Зарын өгөгдлийн мэдээлэл

Баганын нэр	PK	FK	Төрөл ба утга	Тайлбар нэр	Хоосон утга	Тайлбар
Ad_id	+		int	Зарын дугаар	no	Тухайн орон сууц зар дээр байгаа онцгой дугаар
Ad_title			string	Зарын гарчиг	no	Тухайн орон сууц зар дээр байгаа гарчиг
Price			int	Үнэ	no	Тухайн орон сууц зар дээр байгаа үнэ
Room			int	Өрөө	no	Өрөөний тоо
Garage			boolean	Граж	no	Граж байгаа эсэх
Balcony			boolean	Тагт	no	Тагт байгаа эсэх
Area			int	Талбай	no	Нийт талбайны хэмжээ
Door			string	Хаалга	no	Хаалганы төрөл
Window			string	Цонх	no	Цонхны төрөл
Shal			string	Шал	no	Шалны төрөл
Win-dow_num			int	Цонхны тоо	no	Нийт цонхны тоо

Баганын нэр	PK	FK	Төрөл ба утга	Тайлбар нэр	Хоосон утга	Тайлбар
Build_floor			int	Барилгын давхар	no	Тухайн барилгын нийт давхар
Floor			int	Хэдэн давхарт	no	Тухайн орон сууцны байрлах давхар
Year_of_open			int	Ашиглалтанд орсон он	no	Тухайн барилга ашиглалтанд орсон он
User_id		+	string	Хэрэглэгчийн ID	no	Тухайн орон сууцны эзэмшигчийн ID дугаар
Lizing			string	Лизингээр авах боломж	no	Лизингээр орон сууцыг авах боломжтой эсэх
Explanation			string	Тайлбар	no	Тухайн орон сууцны талаарх товч тайлбар
Location_id		+	string	Байршлын ID	no	Тухайн орон сууцны байрлах хаягийн ID дугаар

4.2.2 Хэрэглэгчийн мэдээлэл

Энэ хүснэгтэнд зар дээр тавигдсан орон сууцны эзэмшигчийн мэдээллийг агуулна. Мэдээллээс гадна хэрэглэгч бүрт харгалзах дугаартай байна.

Table 4.2: Хэрэглэгчийн өгөгдлийн мэдээлэл

Баганын нэр	PK	FK	Төрөл ба утга	Тайлбар нэр	Хоосон утга	Тайлбар
User_name			string	Хэрэглэгчийн нэр	no	Тухайн орон сууцны эзэмшигчийн нэр
Phone			int	Утас	no	Тухайн орон сууцны эзэмшигчийн утасны дугаар
User_id	+		int	Хэрэглэгчийн ID	no	Тухайн орон сууцны эзэмшигчийн ID дугаар

4.2.3 Зургийн мэдээлэл

Энэ хүснэгтэнд зар дээр тавигдсан орон сууцны зургийн мэдээллийг агуулна. Мэдээллээс гадна зураг бүрт харгалзах дугаартай байна.

Table 4.3: Зургийн өгөгдлийн мэдээлэл

Баганын нэр	PK	FK	Төрөл ба утга	Тайлбар нэр	Хоосон утга	Тайлбар
Img_id			int	Зургийн ID	no	Тухайн орон сууц зар дээр тавигдсан зурагны ID
Ad_id			int	Зарын дугаар	no	Тухайн орон сууц зар дээр байгаа онцгой дугаар
Url		+	string	URL	no	Тухайн орон сууц зар дээр байгаа зурагны URL

4.2.4 Байршилын мэдээлэл

Энэ хүснэгтэнд зар дээр тавигдсан орон сууцны байршилын мэдээллийг агуулна. Мэдээллээс гадна байршил бүрт харгалзах дугаартай байна.

Table 4.4: Байршилын өгөгдлийн мэдээлэл

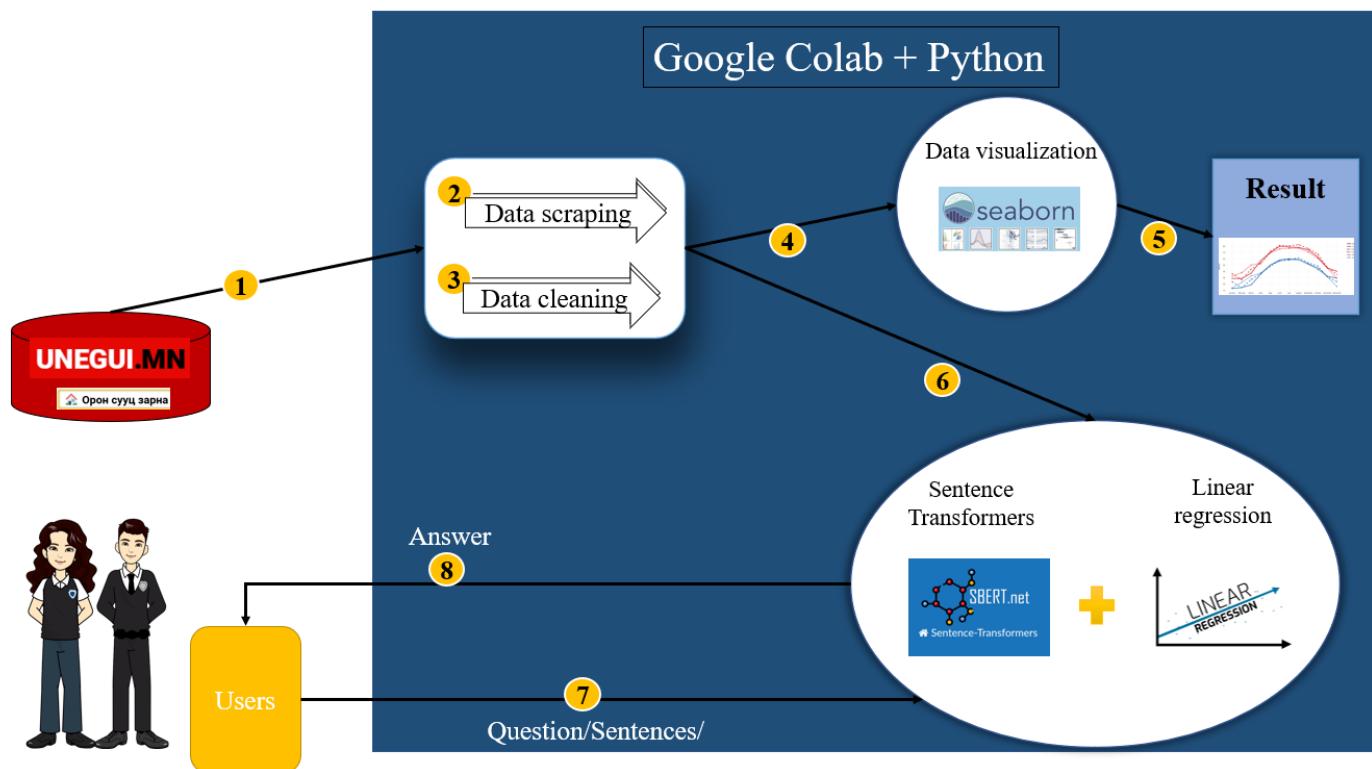
Баганын нэр	PK	FK	Төрөл ба утга	Тайлбар нэр	Хоосон утга	Тайлбар
District			string	Дүүрэг	no	Тухайн орон сууцны байрлах дүүрэг
Bairshil			string	Байршил	no	Тухайн орон сууцны байрлах тодорхой хаяг
Position			string	Байрлал	no	Тухайн орон сууцны байрлах хот аймаг
Location_id	+		string	Байршлын ID	no	Тухайн орон сууцны байрлах хаягийн ID дугаар

4.3 Өгөгдлийн сангийн холбоосын тайлбар

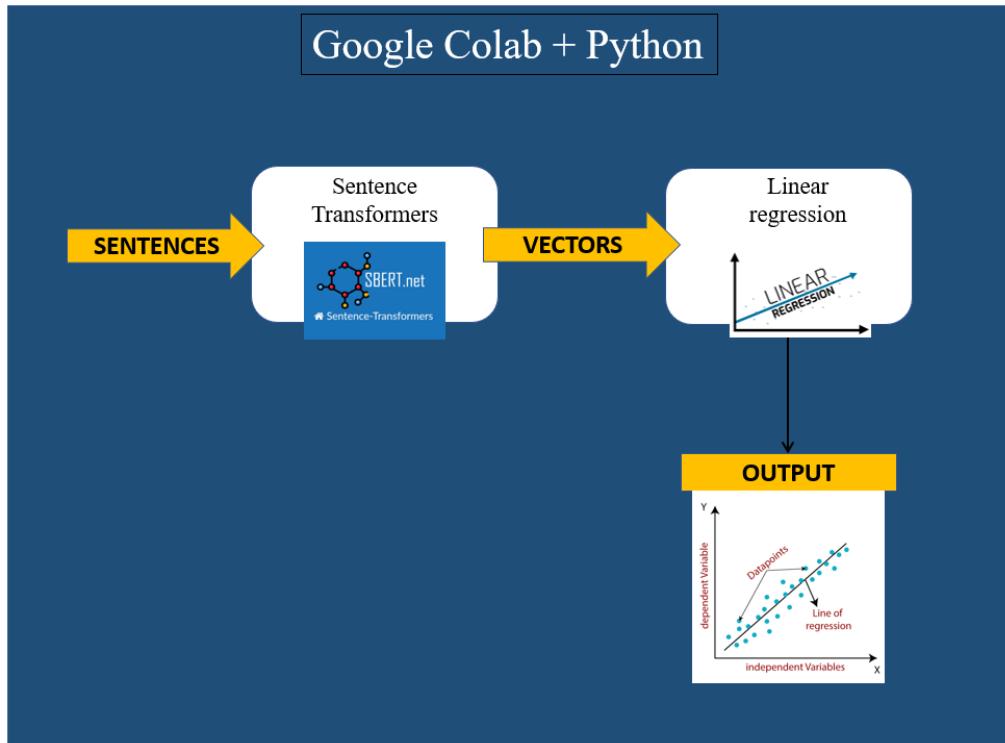
- Нэг хэрэглэгч олон зартай байж болох ба нэг зар нэг л хэрэглэгч дээр байна.
- Нэг байршил/хаяг/ дээр олон зартай байж болох ба нэг зар нэг л байршил/хаяг/ дээр байна.
- Нэг зар дээр олон зурагтай байж болох ба нэг зураг нэг л зар дээр байна.

5. ХЭРЭГЖҮҮЛЭЛТ, ҮР ДҮН

5.1 Хэрэгжүүлсэн байдал



Зураг 5.1: Судалгааны ажлын үйл явцын диаграмм-1

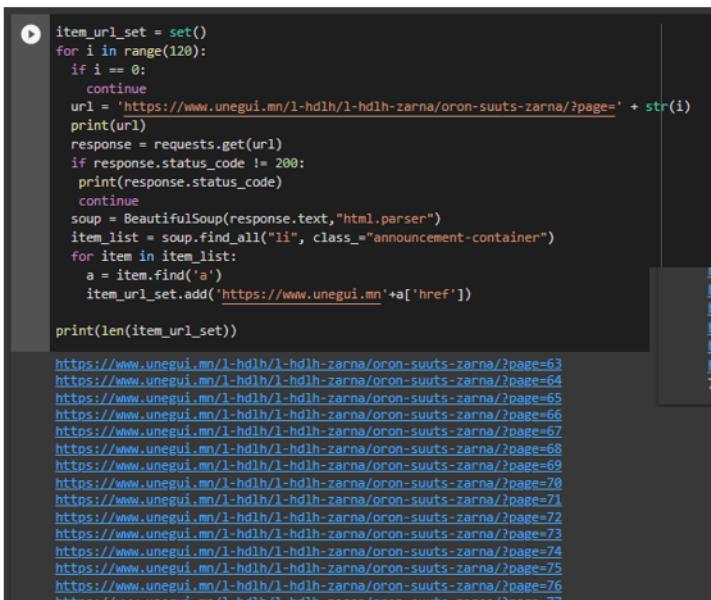


Зураг 5.2: Судалгааны ажлын үйл явцын диаграм-2

Судалгааны ажил нь Зураг 5.1, Зураг 5.2 дээрх үйлийн дагуу ажиллана. Зураг 5.1 бол өрөнхий үйл явцын диаграм, Зураг 5.2 бол өгүүлбэрүүдийн төст байдлыг харьцуулж хамгийн зөв хариултыг илгээх үйл явцыг харуулсан диаграм болно. Python хэлийг ашиглан Google colab дээр data scraping, data cleaning, data анализ хийх болон Sentence Transformer-ийг ашиглаж оруулсан өгүүлбэрийг таньж тохирсон хариулт илгээдэг гэсэн далгаваруудыг хийж гүйцэтгэсэн.

5.1.1 Орон сууцны өгөгдлүүдийг олж авсан үйл явц

Орон сууцны мэдээллийн өгөгдлүүдийг data scraping хийж тэдгээрийг татаж авахын тулд хамгийн эхэнд орон сууцны мэдээллүүдийг агуулсан хуудасны линкуудийг татах хэрэгтэй. Энэхүү үйл явц болон код ажилаж дууссныхаа дараа гарсан үр дүнгийг Зураг 5.3-т харуулсан болно.



The screenshot shows a Jupyter Notebook cell containing Python code for web scraping. The code uses the requests library to get the content of a page and BeautifulSoup to parse it. It extracts URLs from the 'a' tags within specific list items and adds them to a set. The final output is a list of URLs, which are then displayed in a separate text box.

```
item_url_set = set()
for i in range(120):
    if i == 0:
        continue
    url = 'https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=' + str(i)
    print(url)
    response = requests.get(url)
    if response.status_code != 200:
        print(response.status_code)
        continue
    soup = BeautifulSoup(response.text,"html.parser")
    item_list = soup.find_all("li", class_="announcement-container")
    for item in item_list:
        a = item.find('a')
        item_url_set.add('https://www.unegui.mn'+a['href'])

print(len(item_url_set))

https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=63
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=64
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=65
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=66
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=67
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=68
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=69
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=70
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=71
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=72
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=73
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=74
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=75
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=76
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=77
```

YP ДҮН :
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=114
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=115
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=116
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=117
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=118
https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=119
7104

Зураг 5.3: Хуудасны линкуудыг татаж буй үйл явц ба үр дүн

Дараа нь дээрх татсан хуудасны линкуудээс орон сууцны мэдээллийн тодорхой хэд хэдэн өгөгдлүүдийг татсан үйл явц болон үр дүнгийг Зураг 5.4-т харуулсан болно.

```

it = -1
ap_list = []
for item_url in item_url_set:
    url = item_url
    print(url)
    it = it + 1
    response = requests.get(url)
    if response.status_code != 200:
        print(response.status_code)
        continue
    soup = BeautifulSoup(response.text,"html.parser")
    ap = Apartment(url)

    el = soup.find("h1", class_="title-announcement")
    ap.title = el.text.strip()

    li_list = soup.find_all('li')

    ap.space = find_spec_list(li_list,'Талбай:')
    ap.location1 = find_spec_list(li_list,'Дүүрэг:')
    ap.location2 = find_spec_list(li_list,'Байршил')
    ap.date_in = find_spec_list(li_list,'Анги/лалтанд орсон он:')
    ap.lizing = find_spec_list(li_list,'Лизингээр авах боломж:')
    ap.floor = find_spec_list(li_list,'Хэдэн давхарт:')
    ap.build = find_spec_list(li_list,'Барилгын давхар:')

    el2 = soup.find("div", class_="announcement-price_cost")
    ap.price = el2.text.replace('\n','')

    el3 = soup.find("a",class_="announcement__location")
    ap.place = el3.text.strip()

    b = soup.find("span", class_="date-meta")
    ap.date = b.text.strip()

    c = soup.find("span", class_="number-announcement")
    ap.num = c.text.strip()

    print(ap.title)
    ap_list.append(ap)
    if it ==7104:
        break

Streaming output truncated to the last 5000 lines.
https://www.unegui.mn/adv/3973886\_tomor-zavyn-arktal-shar-bairand-3-oroog-zarna/
Гэмтлийн эмчилгэний хийн 2 эрэе байр
https://www.unegui.mn/adv/5890430\_sansurn-kfs-n-urd-upzarmaild-gal-tooro-tusdaa-2-oroog-bair-khudaldana/
Сансрын kfs н урд угсарчлаа гал тогтоо тусдаа 2 эрэе байр 49м2
https://www.unegui.mn/adv/5890830\_naturt-2-oroog-khudaldana-ulchilgee-iavuuulakh-bolezmzhtoi/

```

ҮР ДҮН :

+ Code
+ Text

```

https://www.unegui.mn/adv/5885886\_skhd-21r-khoroo-ol-khamag-mongol-khotkhond-29mkv-1oroog-balryg-zarna/
Схд 21р хороолол хамгийн хотхонд 29мкв 1өрөө байр
https://www.unegui.mn/adv/5534764\_tenuun-apartment-2-oroog/
Танунун апартмент 2 өрөө 51.5м2
https://www.unegui.mn/adv/5820036\_nisekhed-eel-tooro-tusdaa-1-oroog-tavileatai-bair-khudaldana/
Нисэхэд гал тогтоо тусдаа 1 өрөө тавилгатай байр
https://www.unegui.mn/adv/5859292\_nukht-paas-khotkhond-2-oroog-oron-suuts-khudaldana/
Нүхт палас хотхонц 2 өрөө орон сууц 72.81м2
https://www.unegui.mn/adv/5794684\_10-r-khoroo-ol/
10-р хороолол
https://www.unegui.mn/adv/5688687\_u1-khodlokh-khorongo-dondogduiam-rezidens-2oroog/
Дондогдуам резиданс-2өрөө
https://www.unegui.mn/adv/5791452\_teeri-apartment-d-harteryn-khiamb-bair-zarna/
Төгрүг apartment-д байр 52м2
https://www.unegui.mn/adv/5713590\_barguu-4-ram-hill-side-3oroog/
Баруун 4 зам hill side Зорилгоо
https://www.unegui.mn/adv/5878898\_dulaakhan-bair-ili-khotkhond-2oroog/
Дулаахан байр 3 хотхонц 2 өрөө
https://www.unegui.mn/adv/5889937\_sandin-neregzhit-surgutuullin-urd-2-oroog-bair/
Ганцуйн Нэрэмжийн сургуулийн үзүүлэлт 2 өрөө байр
https://www.unegui.mn/adv/5819715\_bzd-crystal-luxury-town-3-oroog-bair/
Бэд crystal luxury town 3 өрөө байр

```

Зураг 5.4: Орон сууцны өгөгдлүүдийг татаж буй үйл явц ба үр дүн

5.1.2 Орон сууцны өгөгдөл

Unegui.mn зарын сайтаас 1 сар гаруй орон сууцны мэдээллийг data scraping хийж өгөгдлүүдийг нь .tsv форматаар татаж авсан. Татсан өгөгдлүүд болох орон сууцны зарын нэр, үнэ, талбай, дүүрэг, байршил, ашиглалтанд орсон он, лизингээр авах боломж, хэдэн давхарт, барилгын давхар, байрлал, нийтэлсэн өдөр, зарын дугаар гэсэн мэдээллүүд байна. Үүнийг Зураг 5.5-т харуулсан болно.

The diagram illustrates the data processing workflow. At the top, a table shows 16 raw data files (0308_ap_data_all.tsv to 0316_ap_data_all.tsv) with columns: Name, Date modified, Type, and Size. A large blue downward arrow points from this table to a second table below.

Top Table (Raw Data Files):

Name	Date modified	Type	Size
0308_ap_data_all.tsv	3/8/2022 11:30 AM	TSV File	3,430 KB
0309_ap_data_all.tsv	3/9/2022 6:42 PM	TSV File	3,588 KB
0310_ap_data_all.tsv	3/10/2022 11:57 AM	TSV File	3,155 KB
0311_ap_data_all.tsv	3/11/2022 12:19 PM	TSV File	3,572 KB
0312_ap_data_all.tsv	3/12/2022 5:39 PM	TSV File	3,573 KB
0313_ap_data_all.tsv	3/13/2022 11:49 PM	TSV File	3,551 KB
0314_ap_data_all.tsv	3/14/2022 9:46 PM	TSV File	3,617 KB
0315_ap_data_all.tsv	3/15/2022 8:55 PM	TSV File	3,600 KB
0316_ap_data_all.tsv	3/16/2022 4:21 PM	TSV File	3,602 KB

Bottom Table (Structured Database):

id	title	price	space	district	location	built_year	leasing	floor	building_max_floor	city	published	
1	Худ 11 хороо river plaza-д 4 өрөө орон сууц 139м2	960 сая ₮	унз тохирно	139	Хан-Уул	River Garden	2022	Лизингтүй	13	24	Улаанбаатар	2022-03-26 11:17
2	3 өрөө байр 68м2	2,9 сая ₮	68	Баянгол	Темер зам	2012	Лизингтүй	6	9	Улаанбаатар	2022-03-02 17:02	
3	Жаст хотжон 2 өрөө байр 56.7м2	158 сая ₮	унз тохирно	56.7	Баянзүр х	14-р хороолол	2015	Банкны лизингтэй	8	15	Улаанбаатар	2022-03-16 17:26
4	Цирк, ub central residence д 3 өрөө байр	466 сая ₮	унз тохирно	108	Сүхбаатар	220 мянгат	2019	Лизингтүй	3	16	Улаанбаатар	2022-02-28 10:38
5	Хотын төв 4 өрөө спортын төв ордны зүн талд	366,3 сая ₮	унз тохирно	111	Сүхбаатар	Бага тойрог	2009	Банкны лизингтэй	9	10	Улаанбаатар	2022-03-04 17:22
6	Багасгийн яг баруун талын угсралмад 1 өрөө байр	95 сая ₮	унз тохирно	40	Баянгол	3, 4 хороолол	1999	Лизингтүй	3	9	Улаанбаатар	2022-03-01 11:44
7	Kh apartment хажууд 79м2 3 өрөө байр	4,3 сая ₮	унз тохирно	79	Хан-Уул	120 мянгат	2022	Лизингтүй	15	16	Улаанбаатар	2022-03-25 9:04
8	3-р эмнэлгийн замын урд 31м2 1өрөө	78 сая ₮	унз тохирно	31	Баянгол	10-р хороолол	2012	Лизингтүй	4	5	Улаанбаатар	2022-03-14 12:41
9	100 айлд цэвэрхэн 2 өрөө орон сууц	110 сая ₮	44.34	Сүхбаатар	Бусад	2016	Лизингтүй	7	12	Улаанбаатар	2022-03-24 14:50	
10	Гарден сити-2 95.80 м2 4 өрөө	431,1 сая ₮	95.8	Хан-Уул	Яармаг	2022	Лизингтүй	9	16	Улаанбаатар	2022-03-17 10:35	
11	Яармагт арцат2 Зөрөө байр 79.6м2	2,5 сая ₮	79.6	Хан-Уул	Яармаг	2022	Хувь лизингтэй	9	17	Улаанбаатар	2022-03-25 12:49	
12	Бэд, зүүн хүрээ хотжон 90м2 3 өрөө	252 сая ₮	90	Баянзүрх	Зүүн 4 зам	2014	Лизингтүй	11	15	Улаанбаатар	2022-03-24 11:02	
13	Нутгатд 1 өрөө байр 56м2	119 сая ₮	56	Хан-Уул	Яармаг	2022	Лизингтүй	1	3	Улаанбаатар	2022-03-19 15:15	
14	Хотын төвд 103 түргэн туслаамжийн хажууд байрлах lux center-т 5 өрөө байр	896 сая ₮	183	Сүхбаатар	Бусад	2014	Лизингтүй	14	15	Улаанбаатар	2022-03-16 7:43	
15	Баянмонголд 2 өрөө 44м2 Яармагт арцат-2 апартмент 79.6м2 3 өрөө	135 сая ₮	44	Баянзүрх	Баянмонгол хороолол	2010	Лизингтүй	6	12	Улаанбаатар	2022-03-22 21:52	
16	Encanto-2 хотжонд 4 өрөө байр 141м2	2,35 сая ₮	79.6	Хан-Уул	Яармаг	2022	Лизингтүй	11	16	Улаанбаатар	2022-03-14 13:49	
17	Moriton residence	3,6 сая ₮	45.1	Хан-Уул	Зайсан	2022	Банкны лизингтэй	10	13	Улаанбаатар	2022-03-13 14:10	
18	Бага тэнгэрийн аманд тансаг эзэртэгэлийн 3 өрөө байр 127м2	687,3 сая ₮	127	Хан-Уул	Бусад	2021	Лизингтүй	6	10	Улаанбаатар	2022-03-17 17:24	

Зураг 5.5: Орон сууцны өгөгдөл

5.1.3 Олж авсан өгөгдлүүдийг шүүх үйл явц

Хамгийн эхэнд татаж авсан өгөгдлүүдийг агуулсан файл доторх бүх мэдээллүүдийг дахин шүүж хараад орон сууцны талаарх яг чухал буюу тодорхой, үүн дээр дүн шинжилгээ хийх боломжтой мэдээллийг заасан өгөгдлүүдийг үлдээж үлдснийг хассан байх болно.

Unegui.mn сайтаас татсан орон сууцны мэдээллийн өгөгдөл дээрх ”Үнэ” гэсэн хэсэг бол хамгийн их асуудалтай багана байсан. Учир нь зарим хүмүүс орон сууцны м²-ийн үнийг тавьсан, харин зарим хүмүүс болхоор орон сууцны нийт үнийг тавьсан. Тиймээс хамгийн түрүүнд үнэ гэсэн хэсгийг өгөгдлийн шүүлт хийсэн болно. Энэ нь байрны м²-ийн үнэ болон нийт үнэ гэж 2 багана хуваасан юм. Дараа нь ”лизингээр авах боломж” гэсэн хэсгийг ”Лизингтүй” гэсэн утгатай байвал FALSE, ”Банкны лизингтэй” эсвэл ”Хувь лизингтэй” гэсэн утгатай байвал TRUE гэж сольж тэмдэглсэн билээ. Эдгээр өгөгдлийн шүүлт хийснийхээ дараа заасан форматанд нийцэхгүй байгаа өгөгдлүүд хасагдсан болно.

```

df = pd.read_csv('/content/drive/MyDrive/data/0328.tsv',sep='\t')

len(df)
7002

```

```

app_df = normalizeDataSet(filtered_df)

len(app_df)
6843

```

Зураг 5.6: Өгөгдлийн өөрчлөлт

Олж авсан орон сууцны мэдээллийн өгөгдлүүдийг шүүлт хийснийхээ дараа Зураг 5.7-т харуулсан байдалтай болсон.

	total	m2_price	space	district	title	year	floor	leasing
0	960	6.90647482	139	Хан-Уул	Худ 11 хороо river plaza- д 4 өрөө орон сууц 139м2	2022	13	FALSE
1	197.2	2.9	68	Баянгол	3 өрөө байр 68м2	2012	6	FALSE
2	158	2.78659612	56.7	Баянзүрх	Жаст хотхон 2 өрөө байр 56.7м2	2015	8	TRUE
3	466	4.314814815	108	Сүхбаатар	Цирк, ub central residence д 3 өрөө байр	2019	3	FALSE
4	366.3	3.3	111	Сүхбаатар	Хотын төв 4 өрөө спортын төв ордны зүүн талд	2009	9	TRUE
5	95	2.375	40	Баянгол	Өргөөгийн яг баруун талын угсармалд 1 өрөө байр	1999	3	FALSE
6	339.7	4.3	79	Хан-Уул	Kh apartment хажууд 79м2 3 өрөө байр	2022	15	FALSE
7	78	2.516129032	31	Баянгол	3-р эмнэлгийн замын урд 31м2 1өрөө	2012	4	FALSE
8	110	2.48082995	44.34	Сүхбаатар	100 айлд цэвэрхэн 2 өрөө орон сууц	2016	7	FALSE
9	431.1	4.5	95.8	Хан-Уул	Гарден сити-2 95.80 м2 4 өрөө	2022	9	FALSE
10	199	2.5	79.6	Хан-Уул	Яармагт арцат2 Зөрөө байр 79.6м2	2022	9	TRUE
11	252	2.8	90	Баянзүрх	Бэд, зүүн хүрээ хотхон 90мкв 3 өрөө	2014	11	FALSE
12	119	2.125	56	Хан-Уул	Нүхтэд 1 өрөө байр 56м2	2022	1	FALSE
13	896	4.896174863	183	Сүхбаатар	Хотын төвд 103 түргэн тусламжийн хажууд байрлах Iux center-т 5 өрөө байр	2014	14	FALSE
14	135	3.068181818	44	Баянзүрх	Баянмонголд 2 өрөө 44м2	2010	6	FALSE
15	187.06	2.35	79.6	Хан-Уул	Яармагт арцат-2 апартмент 79.6мкв 3 өрөө	2022	11	FALSE
16	724	5.134751773	141	Баянзүрх	Encanto-2 хотхонд 4 өрөө байр 141м2	2017	10	TRUE
17	162.36	3.6	45.1	Хан-Уул	Moriton residence	2022	11	TRUE
18	687.3	5.411811024	127	Хан-Уул	Бага тэнгэрийн аманд тансаг зэрэглэлийн 3 өрөө байр 127м2	2021	6	FALSE
19	145	2.071428571	70	Баянгол	Бичилд угсармалын 3 өрөө	1991	9	FALSE
20	412.75	3.25	127	Баянгол	Нарны хороолол 4 өрөө narnii khoroooli 4 игүү 127м2	2015	8	TRUE
21	323.7	3.9	83	Хан-Уул	Гарден сити хотхонд 83мкв 4 өрөө байр	2020	3	FALSE

Зураг 5.7: Data cleaning хийсэн дараах өгөгдөл

5.1.4 Орон сууцны өрөөний мэдээллүүд

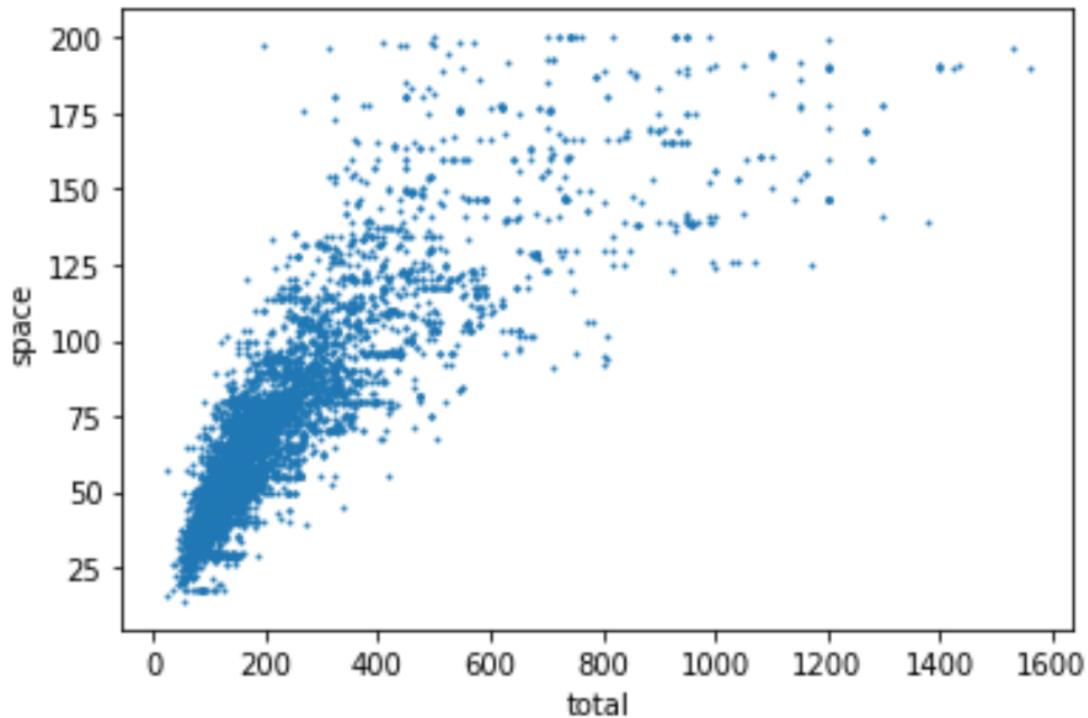
Дээрх шүүлт хийсэн орон сууцны мэдээллийн өгөгдлүүдийг дүүрэг бүрт 1, 2, 3 гэх мэт өрөөний орон сууц нийт хэд байгааг тоолж хүснэгтээр харуулсан болно.

Table 5.1: Орон сууцны өрөөний мэдээллүүд

Өрөөний тоо/Дүүрэг	СБД	БЗД	ХУД	БГД	ЧД	СХД	Налайх дүүрэг	Багануур дүүрэг	Орон нутаг	Нийт
1 өрөө	28	176	137	113	7	81	1	0	0	543
2 өрөө	251	957	916	570	66	214	5	0	3	2982
3 өрөө	121	588	1247	296	41	102	3	3	0	2401
4 өрөө	61	129	506	46	7	18	3	0	0	770
5+ өрөө	10	25	105	3	0	5	0	0	0	148
Нийт	471	1875	2911	1048	121	420	12	3	3	6844

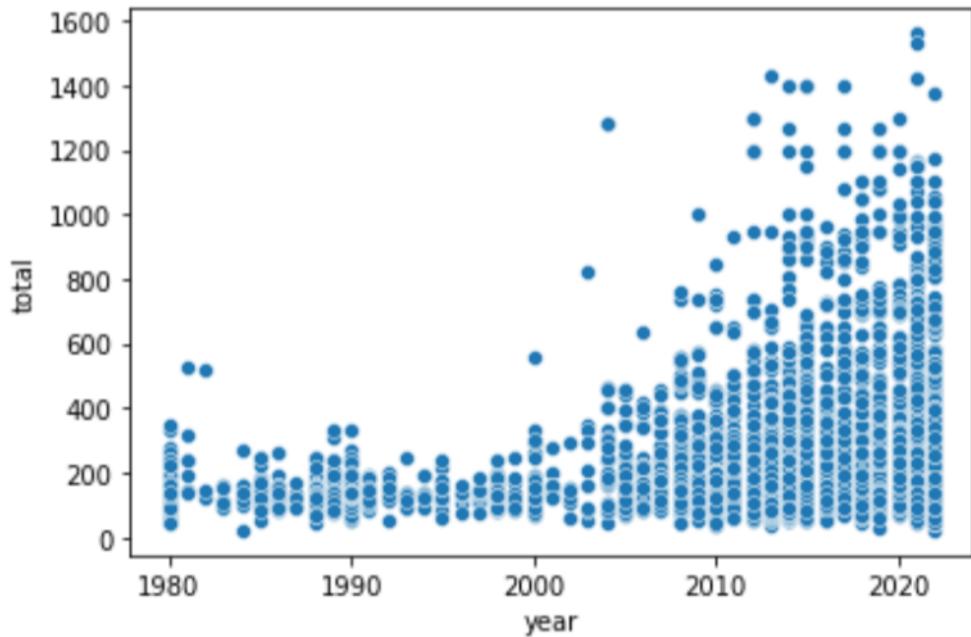
5.1.5 Өгөгдлийн статистик

1. Өгөгдлийн статистик-1 дээр нийт үнэ ба орон сууцны талбай /м²/ гэсэн 2 мэдээллийн өгөгдлүүд дээр график байгуулсан болно. Энэхүү график дүрслэл дээр харуулсан үр дүнгийг дүгнэж хэлбэл 100 м²-аас бага талбайтай орон сууцны дундаж үнэ нь ойролцоогоор 50-300 сая төгрөгтэй байгааг харж болно. Харин 100 м²-аас их 200 м²-аас бага талбайтай орон сууцны хамгийн дээд үнэ нь 1.6 тэрбум төгрөг хүрсэн байна.



Зураг 5.8: Θгөгдлийн статистик-1

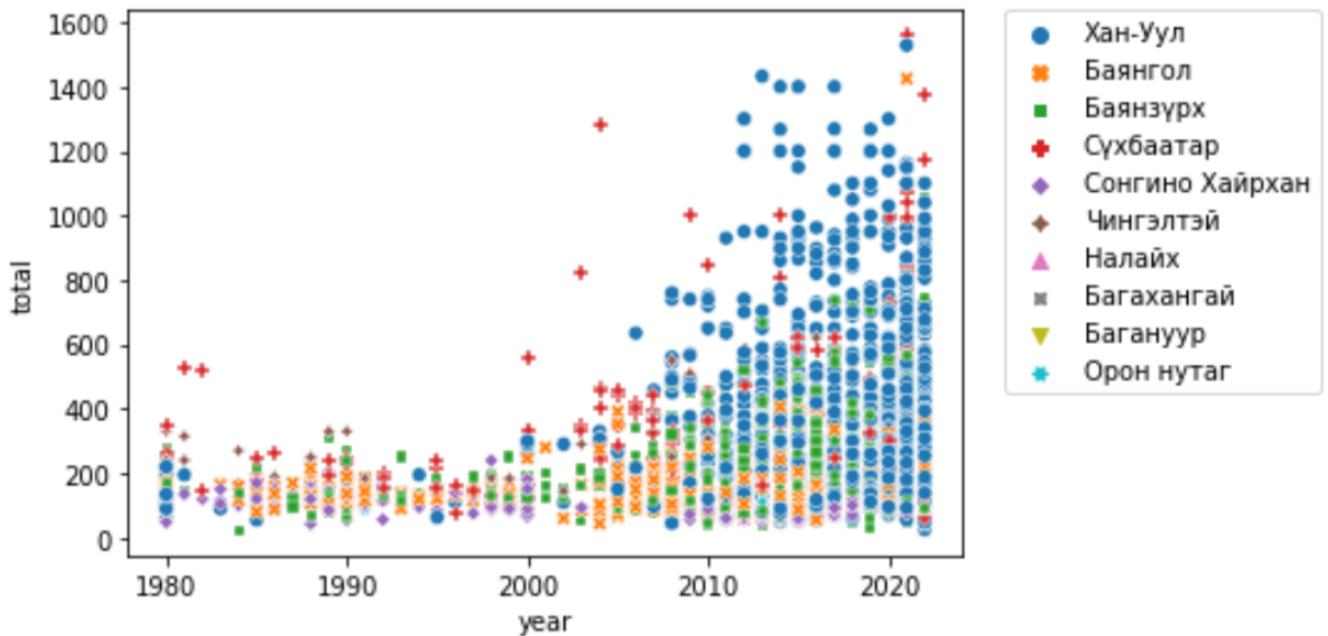
2. Θгөгдлийн статистик-2 дээр нийт үнэ ба ашиглалтанд орсон он гэсэн 2 мэдээллийн өгөгдлүүд дээр график байгуулсан болно. Энэхүү график дүрслэл дээр харуулсан үр дүнгийг дүгнэж хэлбэл 1980-2022 онд ашиглалтанд орсон орон сууцны хамгийн дээд үний дунджийг авч харьцуулбал 1980-1990 оны орон сууцууд бол дундаж 318 сая төгрөгийн үнэтэй, 1990-2000 оных бол 221 сая төгрөгийн үнэтэй, 2000-2010 оны орон сууцууд 660 сая төгрөгийн үнэтэй, 2010-2022 оных бол 1,4 тэрбум хүртэл өссөн байгааг харж болно.



Зураг 5.9: Өгөгдлийн статистик-2

Өгөгдлийн статистик-2 дээр дүүргийн өгөгдлийг нэмж харуулбал доорх зураг харагдах болно. Энэ зурагнаас нь бид 1980-1990 оны хооронд хамгийн үнэтэй орон сууцууд нь ихэнх хувь Сүхбаатар болон Чингэлтэй дүүрэгт, 1990-2000 оны хооронд хамгийн үнэтэй орон сууцууд нь ихэнх хувь Сүхбаатар болон Баянгол дүүрэгт, 2000-2010 оны хооронд хамгийн үнэтэй орон сууцууд нь ихэнх хувь Сүхбаатар болон Хан-Уул дүүрэгт, 2010-2022 оны хооронд хамгийн үнэтэй орон сууцууд нь ихэнх хувь Хан-Уул дүүрэгт байрладаг гэдгийг харж болно.

Мөн 2000 оноос өмнө бол эрэлт их нийлүүлт бага, харин 2000 оноос хойш бол эрэлт бага нийлүүлэлт их байгааг харуулсан байна.

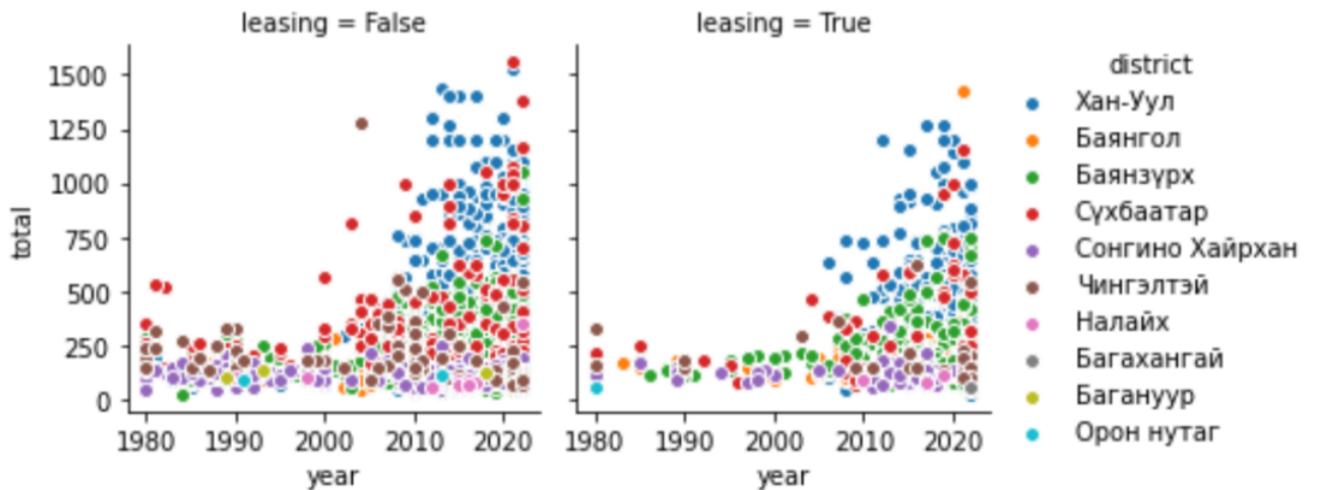


Зураг 5.10: Өгөгдлийн статистик-3

3. Өгөгдлийн статистик-4 дээр нийт үнэ, ашиглалтанд орсон он, дүүрэг болон лизингээр авах боломж гэсэн өгөгдлүүд дээр график байгуулсан болно. Энэхүү график дүрслэл дээр харуулсан үр дүнгийг дүгнэж хэлбэл 1980-2000 оны хоорондох 600 саяаас бага үнэтэй орон сууцуудыг лизинггүйгээр худалдаж авсан ихэнх хувь нь Сонгино Хайрхан болон Чингэлтэй дүүрэгт байна. 2000-2022 оны хооронд бол орон сууцны үнэ нь өсөж, 1.6 тэрбумээс бага үнэтэй лизинггүйгээр худалдаж авсан ихэнх орон сууц нь Хан-Уул, Сүхбаатар, Баянзүрх болон Чингэлтэй дүүрэгт болж өссөн байна.

Харин 1980-2000 оны хооронд 375 саяаас бага үнэтэй орон сууцыг хувь хүний лизинг болон банкны лизингээр хуудалдаж авсан нь Сонгино Хайрхан дүүрэгт байна. 2000-2022 оны хооронд бол орон сууцны үнэ нь өсөж, 1.4 тэрбумээс бага үнэтэй лизингээр орон сууцыг худалдаж авсан нь Хан-Уул, Баянзүрх болон Сүхбаатар дүүрэгт байна.

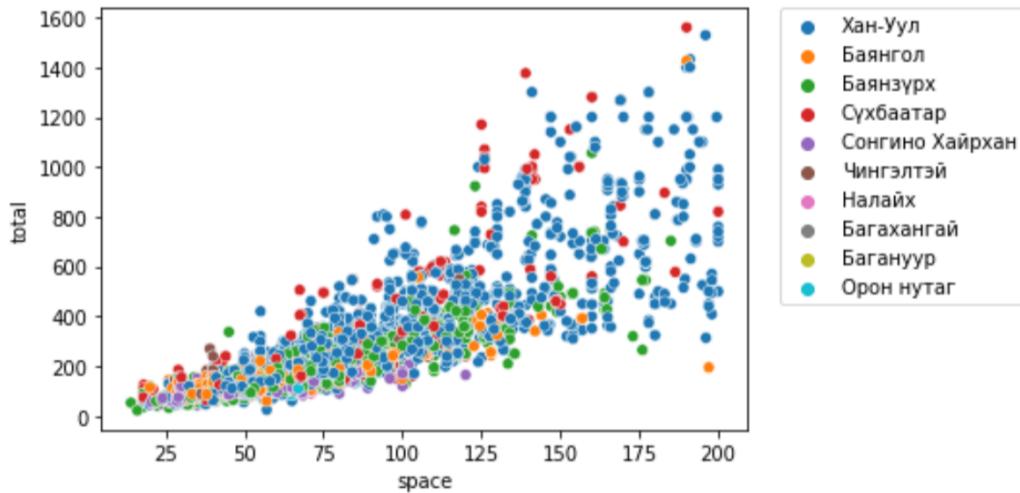
Энэ 2 графикийг нэгтгэн харьцуулж үзвэл хүмүүс лизинггүйгээр орон сууцыг худалдаж авах нь лизингтэй орон сууцыг худалдаж авахаас нь харьцангуй их байна.



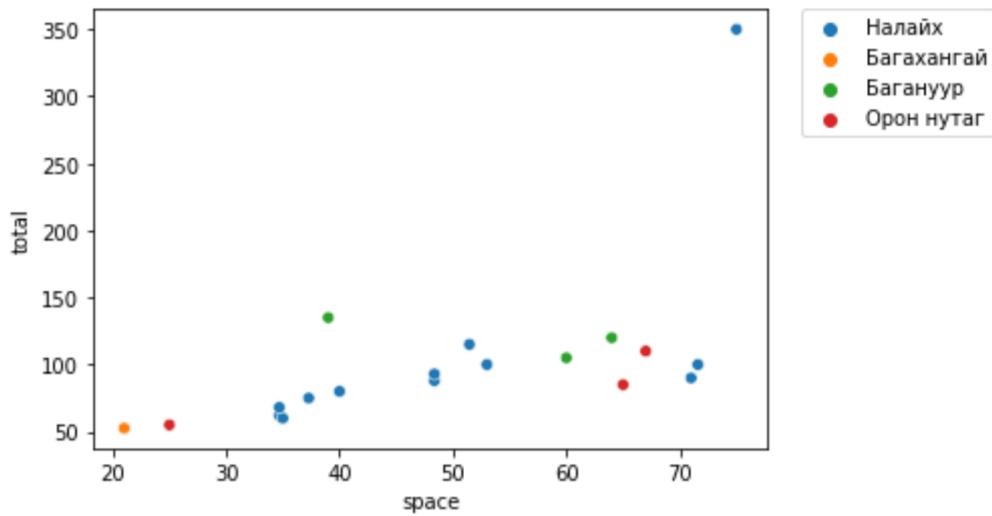
Зураг 5.11: Өгөгдлийн статистик-4

4. Өгөгдлийн статистик-5 дээр нийт үнэ, талбай болон дүүргийн өгөгдлүүд дээр график байгуулсан болно. Энэхүү график дүрслэл дээр харуулсан үр дүнгийг дүгнэж хэлбэл 125 m^2 -аас бага талбайтай орон сууцуудийн хамгийн дээд үнэ нь Хан-Уул болон Сүхбаатар дүүрэг дээр 850 сая, Баянзүрх дүүрэг дээр 780 сая, Чингэлтэй дүүрэг дээр бол 610 сая, Баянгол дүүрэг дээр бол 580 сая, Сонгино Хайрхан 320 сая, Налайх 355 сая, Багануур 135 сая, Орон нутагт 120 сая, Багахангай 55 сая төгрөгийн үнэтэй байна. Харин 125 m^2 -аас их тайлбайтай орон сууцуудийн хамгийн дээд үнэ нь Сүхбаатар дүүрэг дээр 1.58 тэрбум, Хан-Уул дүүрэг дээр бол 1.51 тэрбум, Баянгол дүүрэг дээр бол 1.42 тэрбум, Чингэлтэй дүүрэг дээр бол 1.28 тэрбум, Баянзүрх 1.03 тэрбум төгрөгийн үнэтэй байна.

Дээрх харьцуулалтаас нь дүгнэвэл том талбайтай орон сууцууд ихэнх нь Хан-Уул, Баянзүрх болон Сүхбаатар дүүрэгт байрладаг гэдгийг харж болно.



Зураг 5.12: Өгөгдлийн статистик-5-1



Зураг 5.13: Өгөгдлийн статистик-5-2

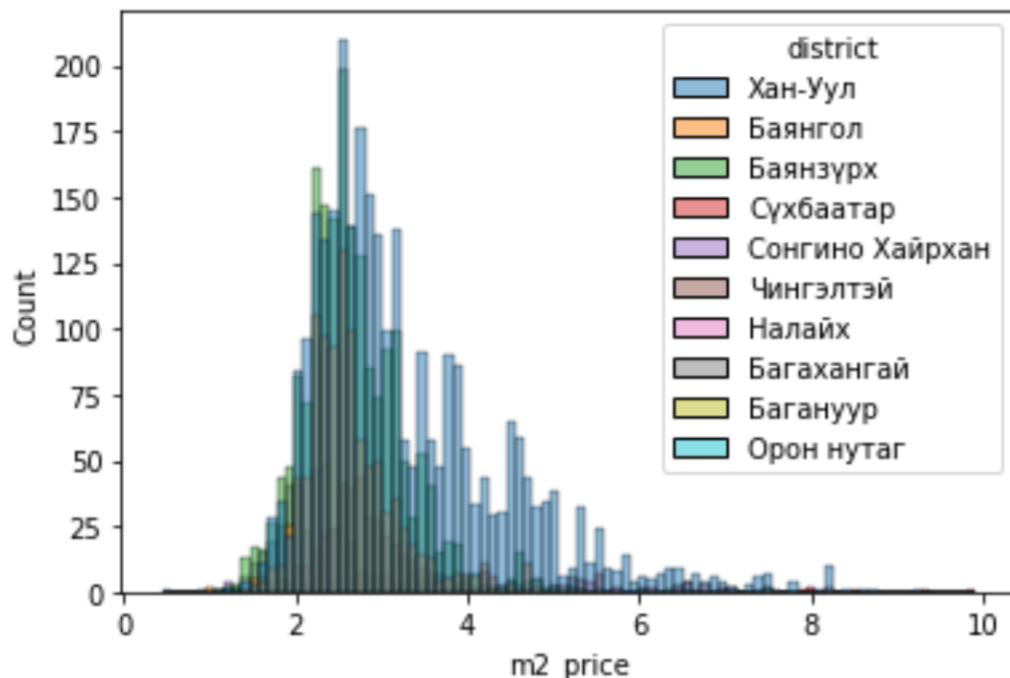
5. Өгөгдлийн статистик-6 дээр орон сууцны m^2 -ын үнэ ба дүүргийн өгөгдлүүд дээр график байгуулсан болно. Энэхүү график дүрслэл дээр харуулсан үр дүнгийг дүгнэж хэлбэл сайт дээр байгаа нийт зарын :

- **44 хувийн орон сууц нь** Хан-Уул дүүрэгт байрладаг. Хан-Уул дүүргийн орон сууц нь 1-9 сая төгрөгийн m^2 -ын үнэтэй байна.
- **22 хувийн орон сууц нь** Баянзүрх дүүрэгт байрладаг. Баянзүрх дүүргийн орон сууц нь

1.2-7.8 сая төгрөгийн м²-ын үнэтэй байна.

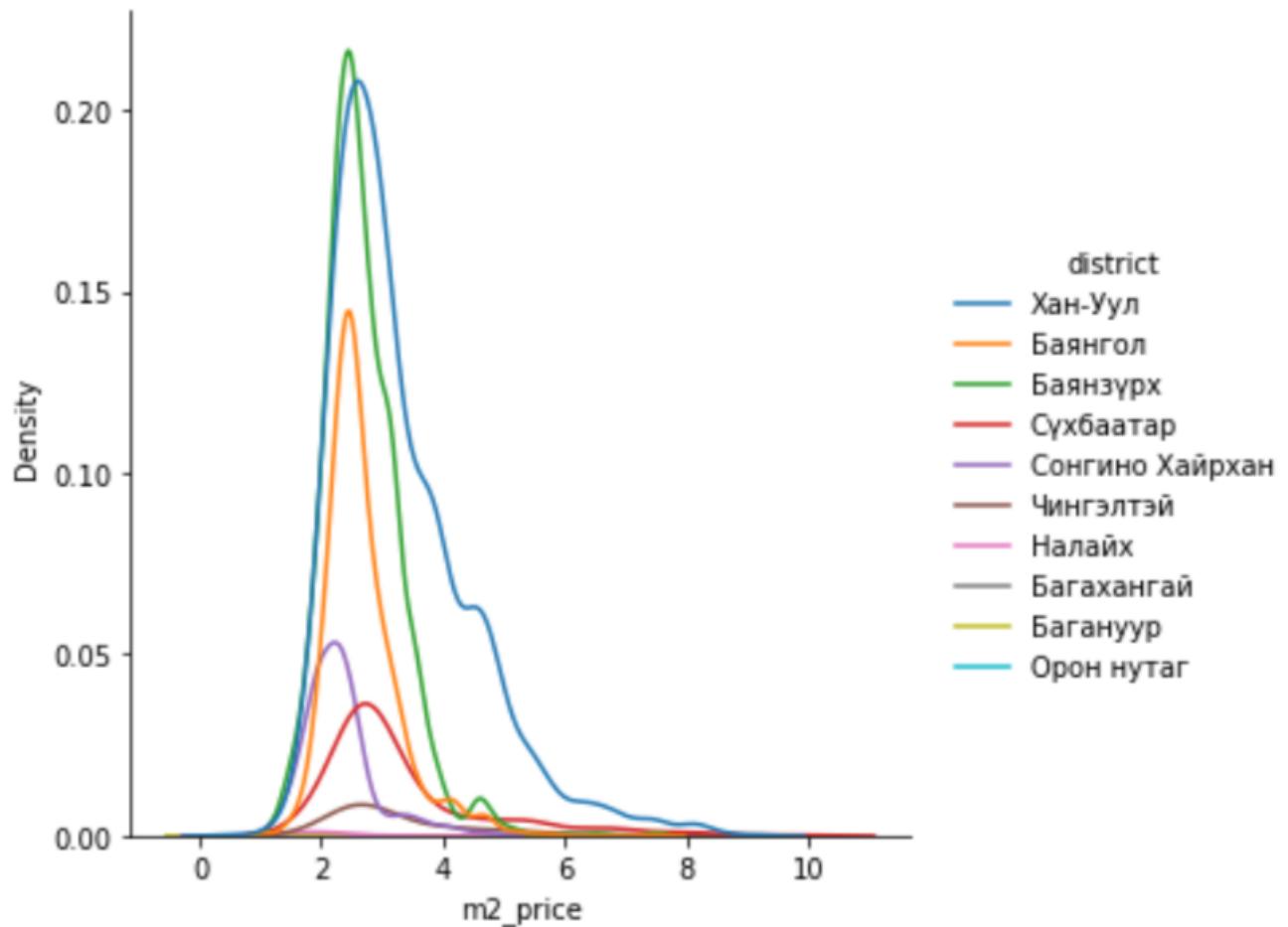
- **15 хувийн орон сууц нь** Баянгол дүүрэгт байрладаг. Баянгол дүүргийн орон сууц нь 1-5.8 сая төгрөгийн м²-ын үнэтэй байна.
- **7 хувийн орон сууц нь** Сүхбаатар дүүрэгт байрладаг. Сүхбаатар дүүргийн орон сууц нь 1-9.8 сая төгрөгийн м²-ын үнэтэй байна.
- **6 хувийн орон сууц нь** Сонгино-Хайрхан дүүрэгт байрладаг. Сонгино-Хайрхан дүүргийн орон сууц нь 1.2-5.2 сая төгрөгийн м²-ын үнэтэй байна.
- Бусад

Дээрх харьцуулалтуудыг авч үзвэл м²-ын дундаж үнэ нь 2.3 сая төгрөгийн үнэтэй байгаа ба Хан-Уул, Баянзүрх дүүргийн орон сууцны эрэлтийн хувьд их хувийг эзэлсэн байна.



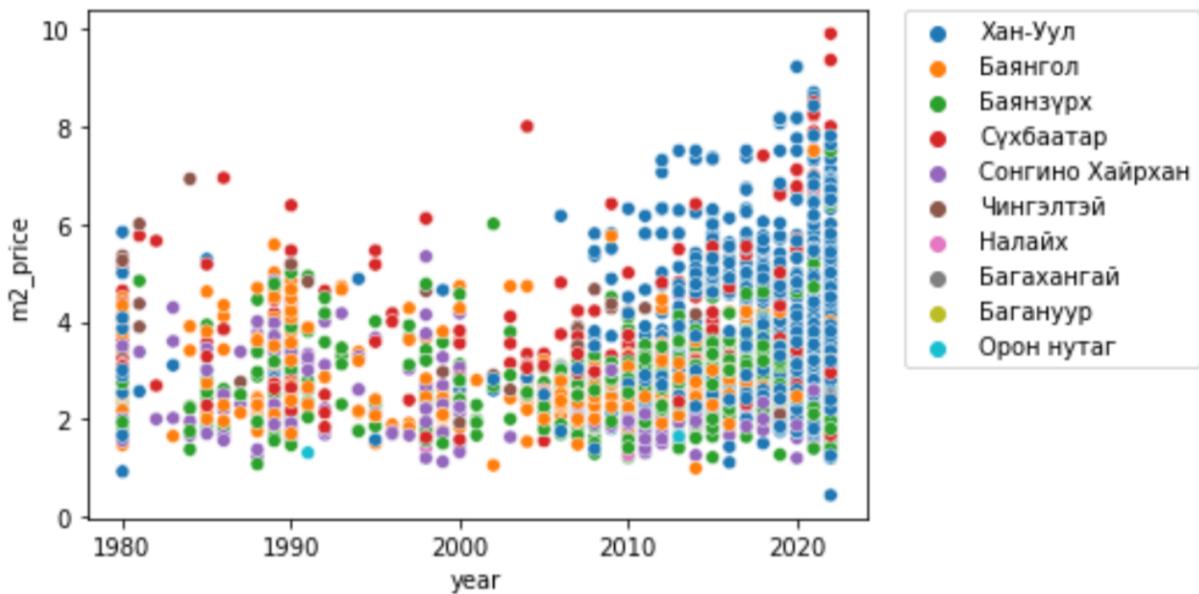
Зураг 5.14: Өгөгдлийн статистик-6

6. Өгөгдлийн статистик-7 дээр орон сууцны m^2 -ын үнэ ба дүүргийн өгөгдлүүд дээр график байгуулсан болно. Энэхүү график дүрслэл дээр харуулсан үр дүнгийг дүгнэж хэлбэл Хан-Уул ба Баянзүрх дүүрэгт иргэдийн төвлөрөл их байгаа бөгөөд орон сууцны m^2 -ын дундаж үнэ ч гэсэн бусад дургүүдтэйгээ харьцуулахад их үнэтэй байгааг харж болно.



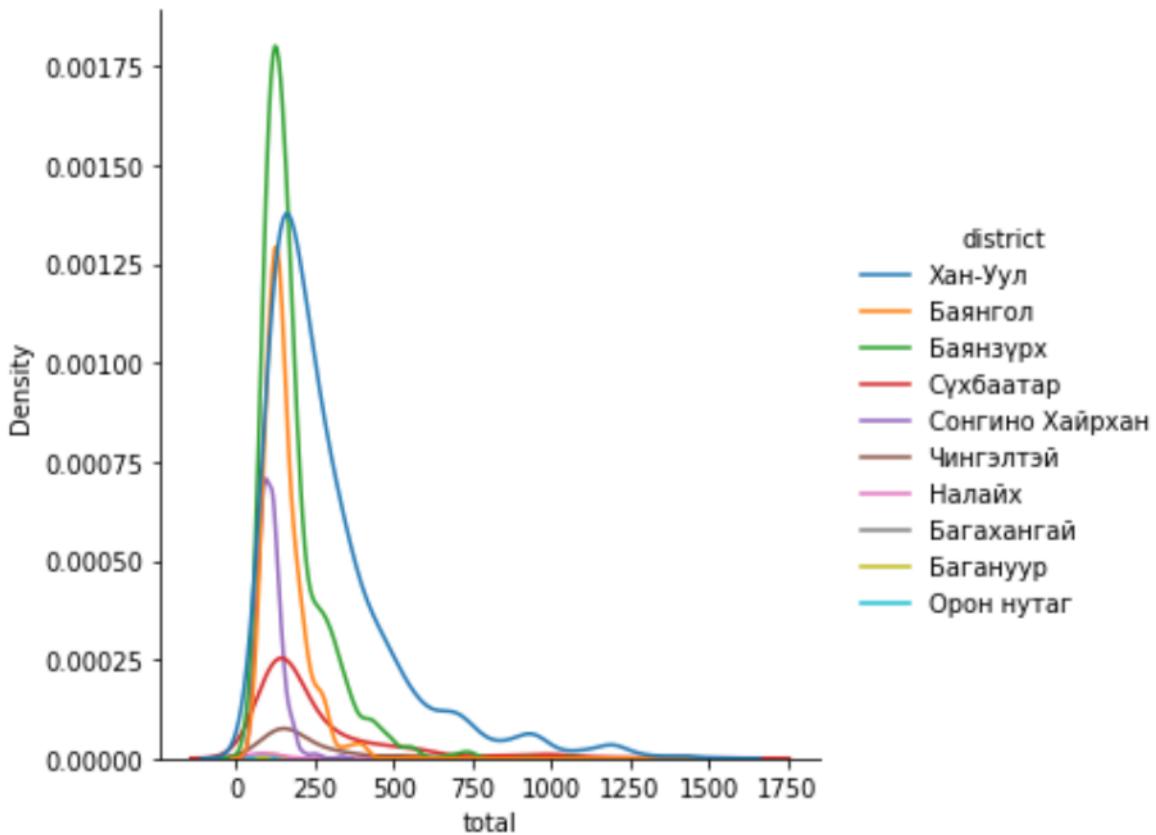
Зураг 5.15: Өгөгдлийн статистик-7

7. Өгөгдлийн статистик-8 дээр орон сууцны m^2 -ын үнэ, ашиглалтанд орсон он ба дүүргүүдийн өгөгдлүүд дээр график байгуулсан болно. Энэхүү график дүрслэл дээр харуулсан үр дүнгийг дүгнэж хэлбэл 2005 оноос хойш ашиглалтанд орсон байрнууд ихэнх нь Хан-Уул болон Баянзүрх дүүрэгт харьялагдаж байгаа бөгөөд Хан-Уул ба Сүхбаатар дүүрэгт байгаа орон сууцны үнэ нь харьцангуй өндөр байна.



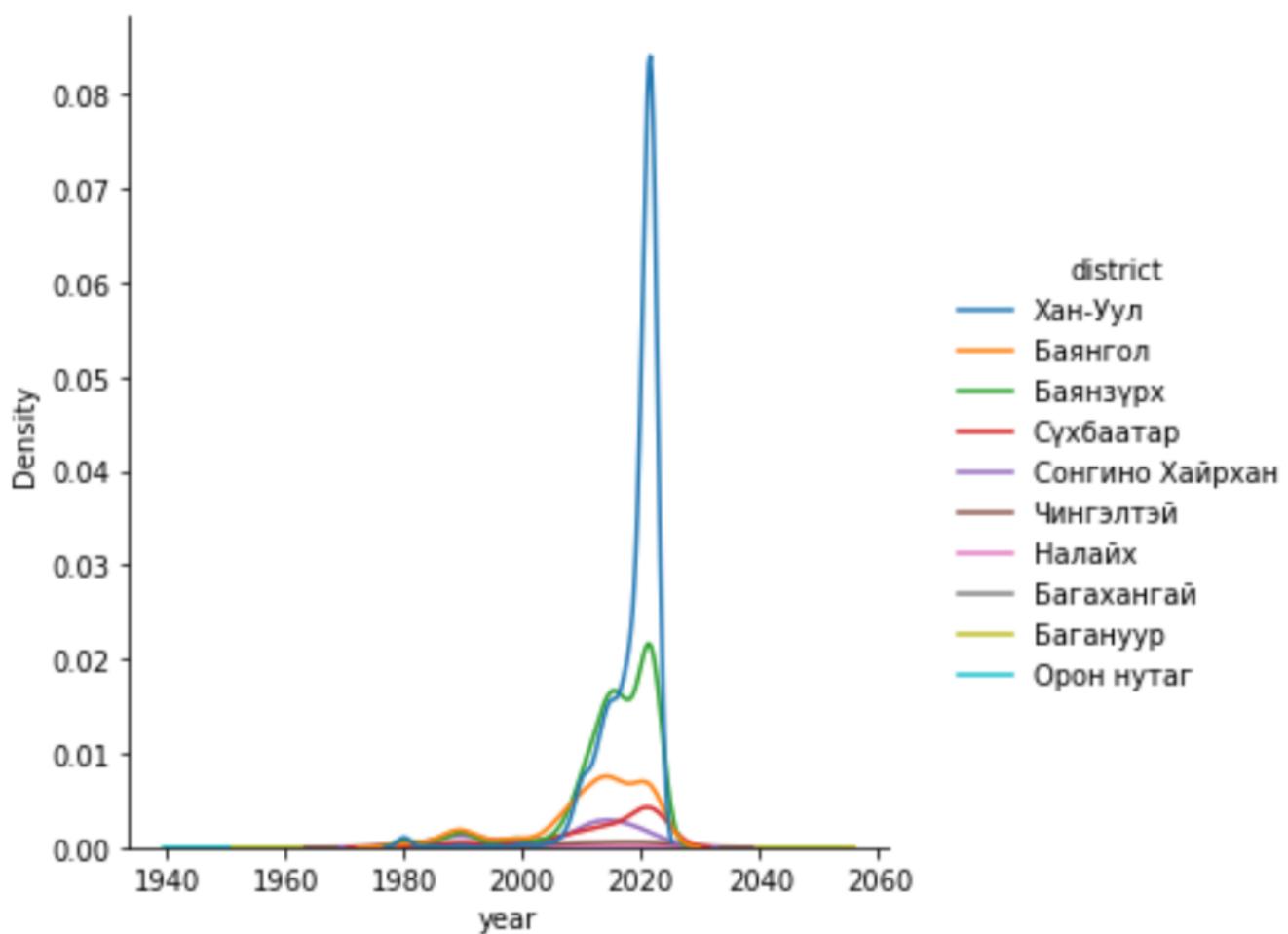
Зураг 5.16: Өгөгдлийн статистик-8

8. Өгөгдлийн статистик-9 дээр нийт үнэ болон дүүргүүдийн өгөгдлүүд дээр график байгуулсан болно. Энэхүү график дүрслэл дээр харуулсан үр дүнгийг дүгнэж хэлбэл орон сууцны үнэ хямд байх тусам нягтаршил нь их байна, харин үнэ өсөх тусам нягтаршил маань багасаж байна. Энэ нь үнэ болон нягтаршил 2 урвуу хамааралтай байгаа гэдгийг харж болно.



Зураг 5.17: Өгөгдлийн статистик-9

9. Өгөгдлийн статистик-10 дээрашиглалтанд орсон он болон дүүргүүдийн өгөгдлүүд дээр график байгуулсан болно. Энэхүү график дүрслэл дээр харуулсан үр дүнгийг дүгнэж хэлбэл 2000-2020 оны орон сууцууд бусад оны орон сууцаас илүү нягтаршилтай байгаа ба энэ дундаас Хан-Уул дүүрэг нь бусад дүргээсээ харьцангуй илүү байгаа гэдгийг харж болно.



Зураг 5.18: Өгөгдлийн статистик-10

5.2 Бүлгийн дүгнэлт

ДҮГНЭЛТ

Bibliography

- [1] Data Scraping, холбогдох онолын судалгаа,
<https://www.targetinternet.com/what-is-data-scraping-and-how-can-you-use-it/>
- [2] Python, холбогдох онолын судалгаа,
<https://www.python.org/doc/essays/blurb/>
- [3] Google Colab, холбогдох онолын судалгаа,
https://colab.research.google.com/?utm_source=scs-index#scrollTo=1SrWNr3MuFUS
- [4] Beautiful Soup, холбогдох онолын судалгаа,
<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
<https://www.educative.io/edpresso/what-is-beautiful-soup>
- [5] Seaborn data visualization, холбогдох онолын судалгаа,
<https://seaborn.pydata.org/>
- [6] Linear Regression, холбогдох онолын судалгаа,
<https://www.scribbr.com/statistics/simple-linear-regression/#:~:text=Linear>
- [7] Sentence Transformers, холбогдох онолын судалгаа,
<https://www.sbert.net/>
<https://huggingface.co/blog/sentence-transformers-in-the-hub>
- [8] SentenceBert, холбогдох онолын судалгаа,
<https://www.searchcandy.uk/nlp/sentence-bert/#Sentence-BERT-SBERT>
<https://medium.com/dair-ai/tl-dr-sentencebert-8dec326daf4e#:~:text=SBERT>
- [9] Unegui.mn, орон сууцны мэдээлэл,
<https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/>
- [10] Өгөгдлийн сангийн диаграм,
<https://dbdiagram.io/d/622c520561d06e6eadebd45f>
- [11] Өгөгдлийн дурслэлийн жишээ, ИЖИЛ СИТЕМИЙН СУДАЛГАА
<https://wwwqlik.com/us/data-visualization/data-visualization-examples>
- [12] World Health Organization, ИЖИЛ СИТЕМИЙН СУДАЛГАА
<https://www.who.int/>

А.БАКАЛАВРЫН СУДАЛГААНЫ АЖЛЫН ҮЕЧИЛСЭН ТӨЛӨВЛӨГӨӨ

Батлав.

2022 оны 02 сарын 15

МОНОГОН НЭРДЛЭЭЗИЙН ОУЮН УХААН СУРИГСАН ОРОН СҮҮДЛЭН ЕРГӨДЛИЙН ДУН ШААНЖИЙН
Английн нэрдэл А-based data analysis on Mongolian Real Estates
Сэргээж баялагчарын судалгааны ажлын
7 ХОНХИЙН ЧИМЧИЙН ТӨВЛӨҮҮС

Хүгэцаа: 2022.02.07-оос 2022.05.06 хүртэл 13 долоо хоног

Зөвхөрсөн: Удирдагч багш / Б.Хүягаар
Болсовтуулсан: Оюутан / Мэдээллийн Технологи, LIU ХТНУА /
Оруулна ID: 1881NUM0118
Холбогдуулж: 95870008

Зураг А.1: Бакалаврын судалгааны ажлын үечилсэн төлөвлөгөө

В. КОДЫН ХЭРЭГЖҮҮЛЭЛТ

```
1 import requests
2 import pandas as pd
3 import urllib.request
4 from bs4 import BeautifulSoup
5
6
7 item_url_set = set()
8 for i in range(123):
9     if i == 0:
10         continue
11     url = 'https://www.unegui.mn/l-hdlh/l-hdlh-zarna/oron-suuts-zarna/?page=' + str(i)
12     print(url)
13     response = requests.get(url)
14     if response.status_code != 200:
15         print(response.status_code)
16         continue
17     soup = BeautifulSoup(response.text, "html.parser")
18     item_list = soup.find_all("li", class_="announcement-container")
19     for item in item_list:
20         a = item.find('a')
21         item_url_set.add('https://www.unegui.mn'+a['href'])
22
23 print(len(item_url_set))
24
25
26 def find_spec_list(li_list, search_key):
27     ret = 'None'
28     for li in li_list:
29         text = li.text.strip()
30         if text.startswith(search_key):
31             ret = text.replace('\n', ' ')
32             break
33     return ret
34
35
36 class Appartment:
37     def __init__(self, _url):
38         self.url = _url
39         self.space = 0
40         self.price = ''
41         self.location1 = ''
42         self.location2 = ''
43         self.date_in = 0
44         self.lizing = ''
45         self.floor = 0
46         self.build = 0
47         self.place = ''
```

```

48     self.date = ''
49     self.num = ''
50
51
52     it = -1
53     ap_list = []
54     for item_url in item_url_set:
55         url = item_url
56         print(url)
57         it = it + 1
58         response = requests.get(url)
59         if response.status_code != 200:
60             print(response.status_code)
61             continue
62         soup = BeautifulSoup(response.text, "html.parser")
63         ap = Apartment(url)
64
65         el = soup.find("h1", class_="title-announcement")
66         ap.title = el.text.strip()
67
68         li_list = soup.find_all('li')
69
70         ap.space = find_spec_list(li_list, '    :')
71         ap.location1 = find_spec_list(li_list, '    :')
72         ap.location2 = find_spec_list(li_list, '    :')
73         ap.date_in = find_spec_list(li_list, "        :")
74         ap.lizing = find_spec_list(li_list, "        :")
75         ap.floor = find_spec_list(li_list, "        :")
76         ap.build = find_spec_list(li_list, "        :")
77
78
79         el2 = soup.find("div", class_="announcement-price__cost")
80         ap.price=el2.text.replace('\n', '')
81
82         el3 = soup.find("a",class_="announcement__location")
83         ap.place = el3.text.strip()
84
85         b = soup.find("span", class_="date-meta")
86         ap.date = b.text.strip()
87
88         c = soup.find("span", class_ = "number-announcement")
89         ap.num = c.text.strip()
90
91         print(ap.title)
92         ap_list.append(ap)
93         if it ==7193:
94             break
95
96
97
98     f = open('0327_ap_data_all.tsv', 'w', encoding='utf-8')
99

```

```

100 for ap in ap_list:
101     f.write(ap.title+"\t"+ap.price+"\t"+ap.space+"\t"+ap.location1+"\t"+
102             ap.location2+"\t"+ap.date_in+"\t"+ap.lizing+"\t"+ap.floor+"\t"+ap.
103             build+"\t"+ap.place+"\t"+ap.date+"\t"+ap.num+"\n")
104 f.close()

```

Код B.1: Data Scraping эх код

```

1 import requests
2 import pandas as pd
3 import numpy as np
4 import csv
5 import urllib.request
6 from bs4 import BeautifulSoup
7
8 from google.colab import drive
9 drive.mount('/content/drive')
10
11 import random
12 with open('/content/drive/MyDrive/data/0328.tsv', 'r', encoding='utf-8') as f:
13     app_lines = f.read().split('\n')
14
15 df = pd.read_csv('/content/drive/MyDrive/data/0328.tsv', sep='\t')
16
17 len(df)
18
19 filtered_df = df[df.city == '']
20
21 print(len(df), len(filtered_df))
22
23 for index, row in filtered_df.iterrows():
24     if ('' in row['price']) is False:
25         print(row)
26
27 def extractTprice(price):
28     ans = .0
29     if '' in price:
30         ans = float(price.split('')[0].strip().replace(',', '.'))
31     elif '' in price:
32         ans = float(price.split('')[0].strip().replace(',', '.')) * 1000
33     else:
34         return None
35     return ans
36
37
38 def normalizeDataSet(app_set):
39     ret = pd.DataFrame(columns=["total", "m2_price", 'space', 'district',
40                           'title', 'year', 'floor', 'leasing'])
41     for index, row in app_set.iterrows():
42         t_price = extractTprice(row['price'])
43         if t_price is None:

```

```

43     continue
44     if float(row['space']) < 10 or float(row['space']) > 200:
45         continue
46     #print(float(row['space']))
47     if row['leasing'].strip()=='':
48         leas = False
49     else:
50         leas = True
51     if float(t_price) > 21:
52         total = float(t_price)
53         m2 = t_price / float(row['space'])
54     else:
55         total = float(t_price * float(row['space']))
56         m2 = float(t_price)
57     if m2 > 10:
58         continue
59     ret = ret.append({'total': total, 'm2_price':m2,
60                       'space': float(row['space']), 'floor': row['
61                           floor'], 'leasing' : leas,
62                           'district': row['district'], 'title': row['title'], ' '
63                           'year': row['built_year']}, ignore_index=True)
64
65 return ret
66
67 app_df = normalizeDataSet(filtered_df)
68
69 len(app_df)
70
71 app_df.to_csv('data0328.csv')

```

Код B.2: Data Cleaning өх код