

Зертханалық жұмыс 6

Деректерді жинау және дайындау негіздері

Жұмыстың мақсаты: Жасанды интеллект үлгілерін оқыту, тексеру немесе сынау үшін әртүрлі дереккөздерден жинау және дайындау бойынша негізгі дағдыларды меңгеру.

ЖИ және машиналық оқытуда деректерді жинау нақты сценарийлерді қамтитын жан-жақты және репрезентативті деректер жиынтығын жинаудан тұрады. Бұл деректер жиынтығы әзірлеушілер мен зерттеушілер алгоритмдерді үлгіні тануға, болжам жасауға немесе басқа когнитивті тапсырмаларды орындауға үйрете алатын негізгі блоктар ретінде қызмет етеді.

Деректерді талдау үшін қолданылатын құрал:

- Google Collab: <https://colab.research.google.com/>

Мысал ретінде зертханалық жұмыста қолдануға арналған мәліметтер жиынтығы:

- Деректер жиынтығы-Титаник, <https://www.kaggle.com/c/titanic/data> бұл жиынтық талдау және өңдеу үшін Google Collab-қа жүктеледі.

Жұмысқа арналған барлық деректер жиынтығы курсқа жүктеледі!

Мысал ретінде жұмыс істеуге арналған негізгі файл: **train.csv**

Деректермен жұмыс жасамас бұрын, деректерді тазарту және дайындау қажет:

- Өткізіп алған мәндерді өңдеу
- Деректер форматтарын стандарттау
- Сандық мәндерді қалыпқа келтіру
- Категориялық айнымалыларды кодтау

Все наборы данных для работы будут загружены на курс!

Ол үшін сізге қажет:

1. Кітапхананы импорттау- **import** pandas **as** pd. Pandas-деректерді манипуляциялау және талдау үшін Python бағдарламалау тілі үшін жазылған бағдарламалық кітапхана. Сандық кестелер мен уақыт қатарларын басқару үшін деректер құрылымдары мен операцияларды ұсынады.
2. Csv файлын оқу - `df = pd.read_csv('train.csv')`

Импортирование библиотеки

название библиотеки сокращенное название библиотеки

```
• import pandas as pd
```

ключевое слово для импортирования библиотек в Питоне alias (переименование для обращения внутри кода)

Считывание csv файла

Обращение к библиотеке pandas через alias Функция для чтения файла формата csv (является частью pandas)

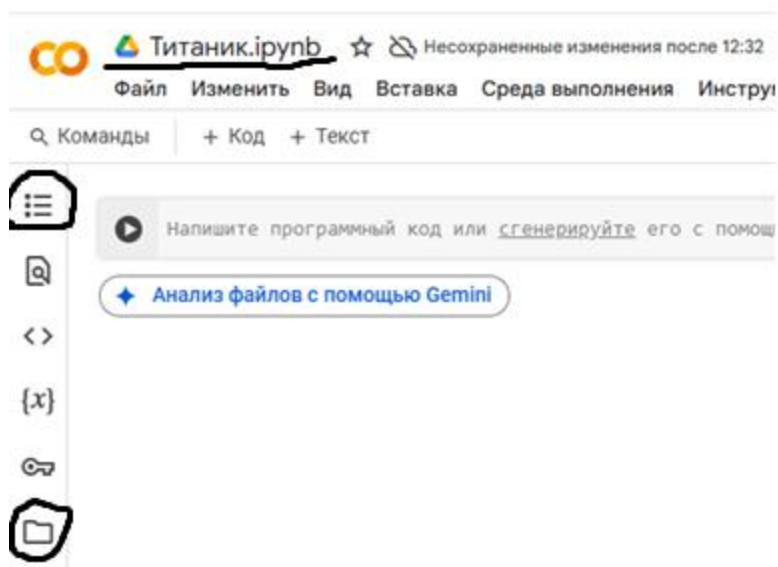
```
• df = pd.read_csv('train.csv')
```

Название переменной для сохранения содержимого Название файла

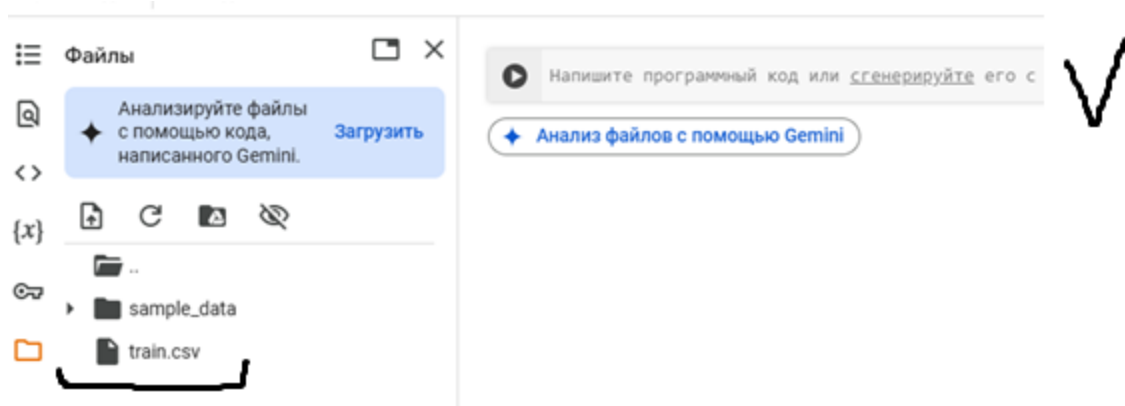
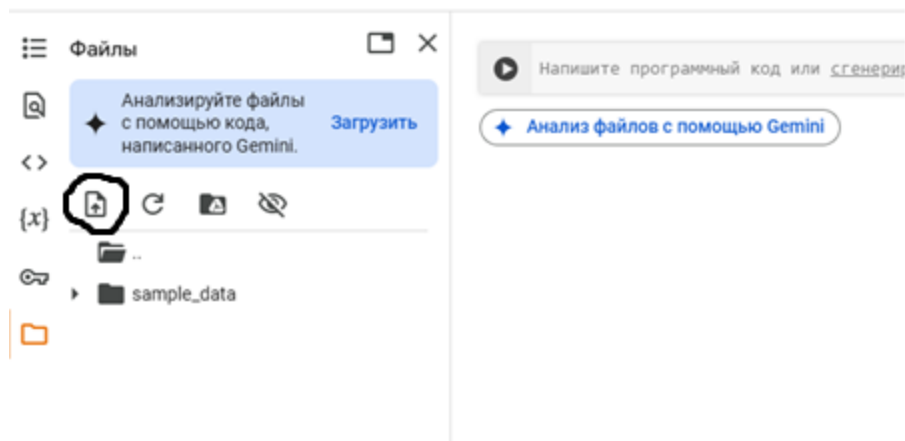
Активация Windows
Чтобы активировать Windows "Параметры".

Google Collab-қа тіркеліңіз Google аккаунтыңызды пайдаланып, құжатты құрып, кітапхананы жүктеп және файлды оқыңыз.

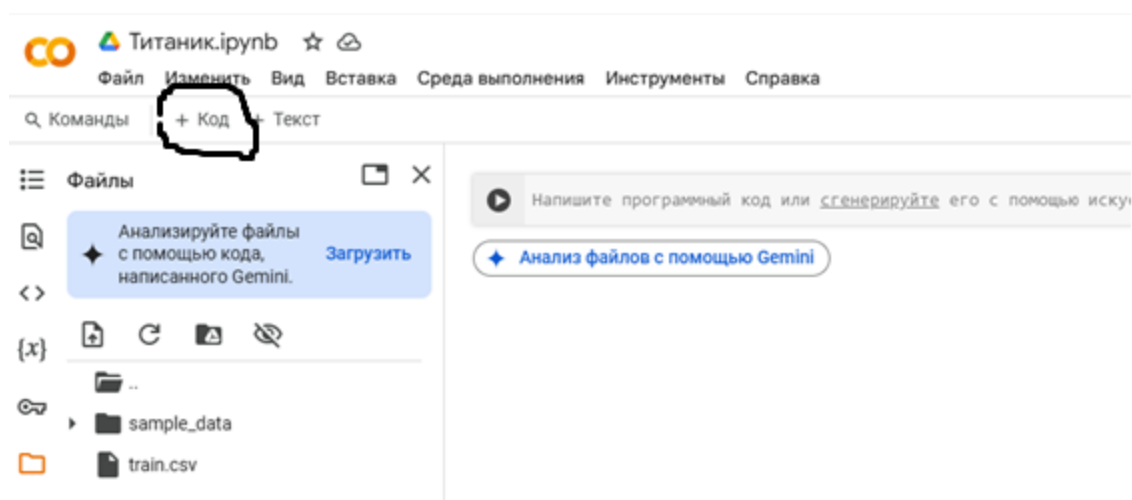
Google Collab – қа кіріңіз-оны Титаник деп атаңыз, train.csv файлын жүктеу үшін , алдымен мазмұнды, содан кейін буманы басыңыз.



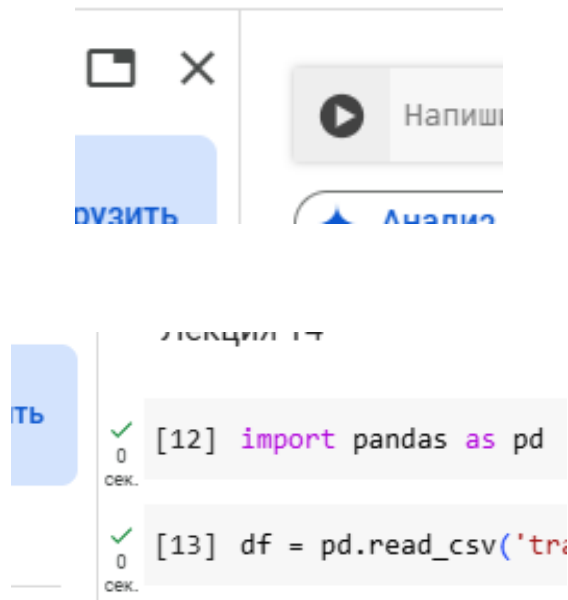
Буманы басқаннан кейін келесі белгіше пайда болады, файлды жүктеу үшін оны басыңыз.



Файл жүктелді, енді біз кітапхананы жүктеу және деректерді оқу үшін кодты жазамыз. Жұмыс істеу үшін барлық кодты зертханалық жұмыстан көшіруге болады. Кодты жазғыңыз келген сайын мәзірдегі КОД сөзін басыңыз.



Код жұмыс істеуі үшін кодты жазғаннан кейін оны жасыл белгі пайда болатындай етіп іске қосу керек.



Деректерді көру үшін жаза аласыз:

Тақырып бойынша деректерді қарау(басы):

- df
- df.head()
- df.head(5)
- df.head(10)

Соңы бойынша деректерді қарау :

- df.tail()
- df.tail(5)
- df.tail(10)

Деректерді сипаттау :

- df.describe()
- df.info()

Команды

Код

Текст

Файлы

Анализируйте файлы с помощью кода, написанного Gemini. Загрузить

sample_data

cleaned_titanic.csv

train.csv

import pandas as pd

[13] df = pd.read_csv('train.csv')

[14] df

	Id пассажира	Выжил	класс	Имя	Гендер	Возраст	# братьев и сестер / супругов на борту	# родителей/ детей на борту	\tНомер билета	Стоимость проезда пассажира	Номер каюты	Порт отправления
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cummings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	0	Allen, Mr. William	male	29.0	0	0	370150	8.0000	NaN	C

Файл Изменить Вид Вставка Среда выполнения Инструменты Справка

Команды + Код + Текст

Файлы

Анализируйте файлы с помощью кода, написанного Gemini. Загрузить

sample_data

cleaned_titanic.csv

train.csv

Далее: Создать код с переменной df Посмотреть рекомендованные графики New interactive sheet

[15] df.head()

	Id пассажира	Выжил	класс	Имя	Гендер	Возраст	# братьев и сестер / супругов на борту	# родителей/детей на борту	\tНомер билета	Стоимость проезда пассажира	Номер каюты	Порт отправления
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily)	female	35.0	1	0	113803	53.1000	C123	S

Файл Изменить Вид Вставка Среда выполнения Инструменты Справка

Команды + Код + Текст

Файлы

Анализируйте файлы с помощью кода, написанного Gemini. Загрузить

sample_data

cleaned_titanic.csv

train.csv

Далее: Создать код с переменной df Посмотреть рекомендованные графики New interactive sheet

[16] df.tail()

	Id пассажира	Выжил	класс	Имя	Гендер	Возраст	# братьев и сестер / супругов на борту	# родителей/детей на борту	\tНомер билета	Стоимость проезда пассажира	Номер каюты	Порт отправления
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W/C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

Файл Изменить Вид Вставка Среда выполнения Инструменты Справка

Команды + Код + Текст

Файлы

Анализируйте файлы с помощью кода, написанного Gemini. Загрузить

sample_data

cleaned_titanic.csv

train.csv

Далее: Создать код с переменной df Посмотреть рекомендованные графики New interactive sheet

[17] df.describe()

	Id пассажира	Выжил	класс	Возраст	# братьев и сестер / супругов на борту	# родителей/детей на борту	Стоимость проезда пассажира
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Файл Изменить Вид Вставка Среда выполнения Инструменты Справка

Команды + Код + Текст

Файлы

Анализируйте файлы с помощью кода, написанного Gemini. Загрузить

sample_data

cleaned_titanic.csv

train.csv

Далее: Создать код с переменной df Посмотреть рекомендованные графики New interactive sheet

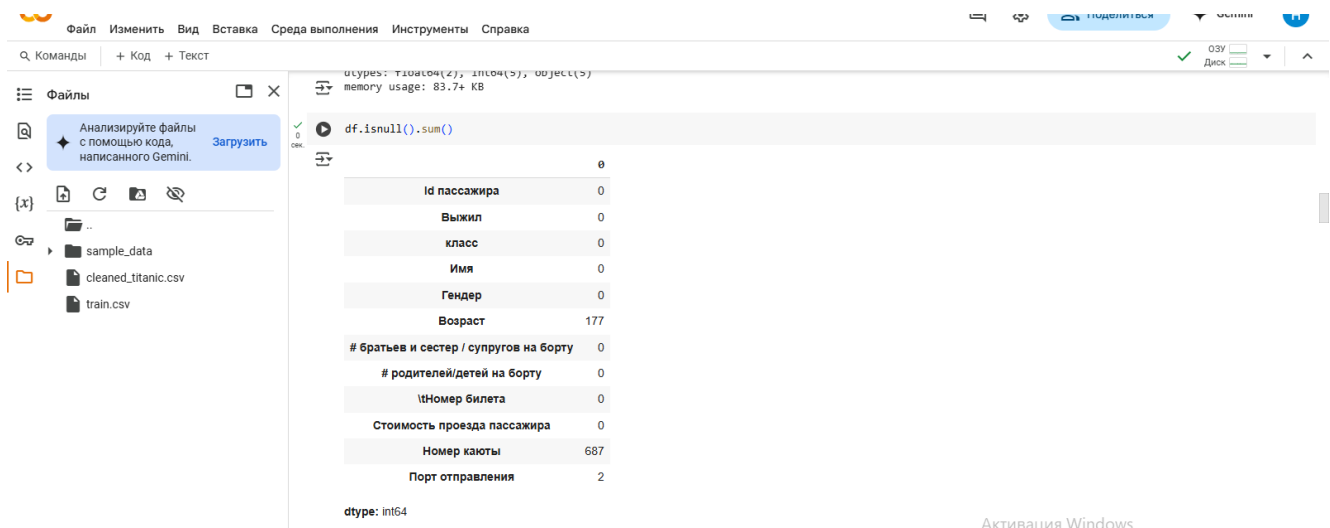
df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Id пассажира                        891 non-null    int64
1   Выжил                              891 non-null    int64
2   класс                              891 non-null    int64
3   Имя                                891 non-null    object
4   Гендер                             891 non-null    object
5   Возраст                             714 non-null    float64
6   # братьев и сестер / супругов на борту  891 non-null    int64
7   # родителей/детей на борту          891 non-null    int64
8   Номер билета                        891 non-null    object
9   Стоимость проезда пассажира        891 non-null    float64
10  Номер каюты                         284 non-null    object
11  Порт отправления                    889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Толтырылмаған деректер бар, жетіспейтін деректерді өңдеу. Оларды қалдыруға болмайды, өйткені біз талдау жасай алмаймыз. Оларды табу үшін бос мәндер бар жерде төмендегі кодты қолданамыз:

Жетіспейтін мәндерді анықтау:

`df.isnull().sum()`



Файл Изменить Вид Вставка Среда выполнения Инструменты Справка

Q Команды + Код + Текст

Анализируйте файлы с помощью кода, написанного Gemini. Загрузить

sample_data

cleaned_titanic.csv

train.csv

df.isnull().sum()

memory usage: 83.7+ KB

Id пассажира	0
Выжил	0
класс	0
Имя	0
Гендер	0
Возраст	177
# братьев и сестер / супругов на борту	0
# родителей/детей на борту	0
\tНомер билета	0
Стоимость проезда пассажира	0
Номер каюты	687
Порт отправления	2

dtype: int64

Активация Windows

Көріп отырғаныңыздай, бағандар Жасы, Кабинаның(каютаның) нөмірі, Шығу порты бос мәндерге ие. Бұл жағдайды түзетейік.

Жетіспейтін деректерді өңдеу

- # Жетіспейтін " Жасы " мәнін медианамен толтырыңыз.

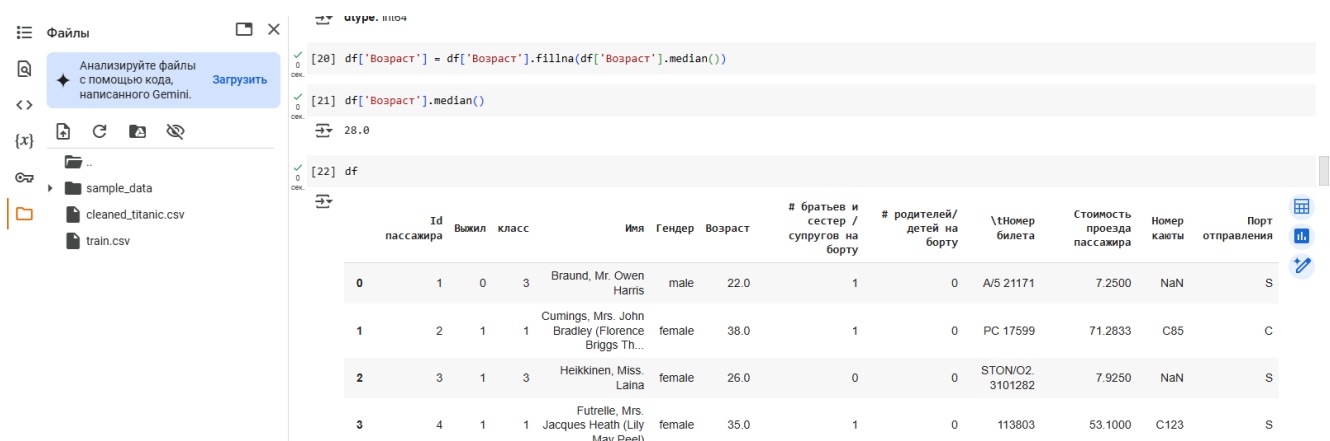
```
df['Возраст'] = df['Возраст'].fillna(df['Возраст'].median())
```

- # Жетіспейтін "Шығу порты" сөзін режиммен толтырыңыз

```
df['Порт отправления'] = df['Порт отправления'].fillna(df['Порт отправления'].mode()[0])
```

- # Тым көп өткізіп алған мәндерге байланысты "Кабина нөмірін" жойыңыз.

```
df = df.drop(columns=['Номер каюты'])
```



Файл Изменить Вид Вставка Среда выполнения Инструменты Справка

Анализируйте файлы с помощью кода, написанного Gemini. Загрузить

sample_data

cleaned_titanic.csv

train.csv

[20] df['Возраст'] = df['Возраст'].fillna(df['Возраст'].median())

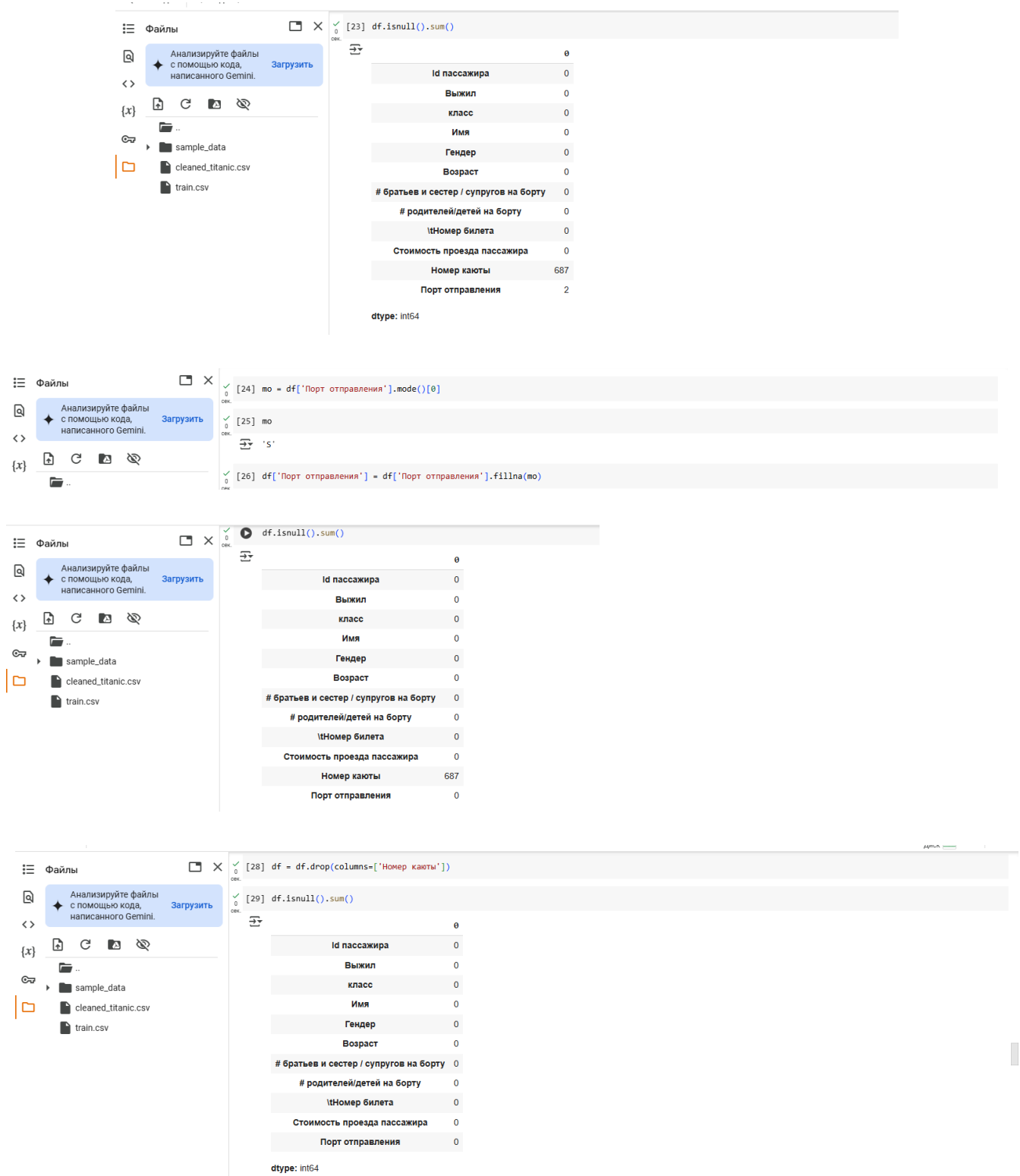
[21] df['Возраст'].median()

28.0

[22] df

	Id пассажира	Выжил	класс	Имя	Гендер	Возраст	# братьев и сестер / супругов на борту	# родителей/детей на борту	\tНомер билета	Стоимость проезда пассажира	Номер каюты	Порт отправления
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cummings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

Жасы бағанасында, енді нөлдік бос мәндер бар екенін тексерейік. Төменде біз қалған деректермен келесі әрекеттерді орындаймыз, Шығу портын толтырамыз, Кабиначарды алып тастаймыз.



Барлық деректер толтырылғаннан кейін деректерді машинаға түсінікті мәндерге түрлендіру қажет.

Деректер түрлерін түрлендіру

- Қажет болса, категориялық бағандарды сандық форматтарға түрлендіріңіз.
- Мысалы: Машиналық оқыту мақсатында "Гендерлік" және "Шығу портын" кодтаңыз.

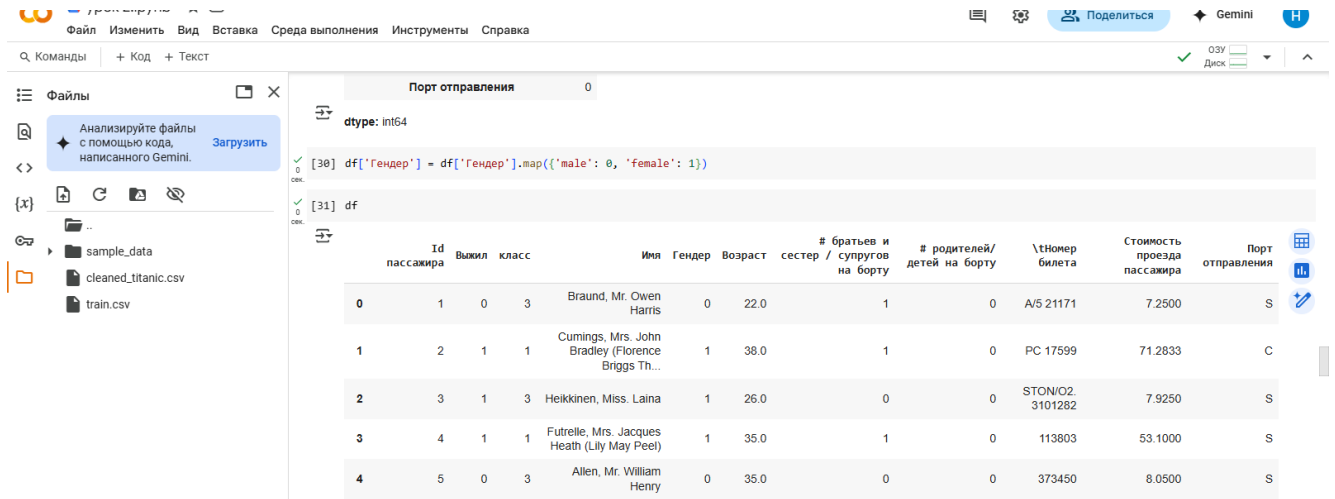
Деректер түрлерін түрлендіру:

- # "Жынысты" сандық мәнге түрлендіру (ерлер үшін 0 (male), әйелдер үшін 1 (female))

```
df['Гендер'] = df['Гендер'].map({'male': 0, 'female': 1})
```

- # "Шығу порты" жылдам кодтау

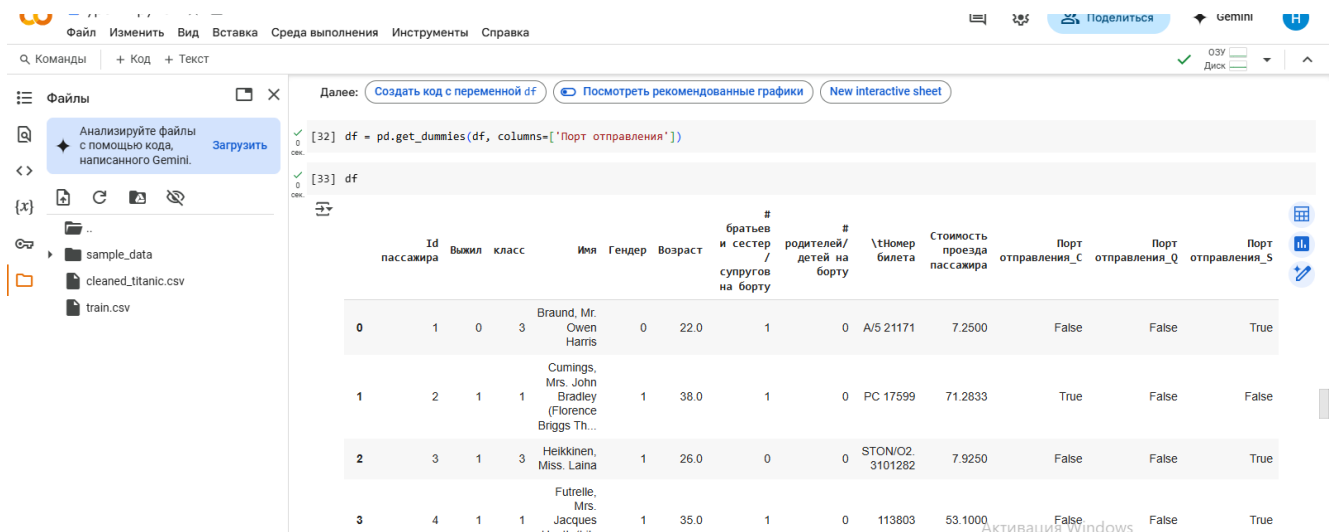
```
df = pd.get_dummies(df, columns=['Порт отправления'], drop_first=True)
```



```
[30] df['Гендер'] = df['Гендер'].map({'male': 0, 'female': 1})
```

```
[31] df
```

	Id пассажира	Выжил	класс	Имя	Гендер	Возраст	# братьев и сестер / супругов на борту	# родителей/детей на борту	\tНомер билета	Стоимость проезда пассажира	Порт отправления
0	1	0	3	Braund, Mr. Owen Harris	0	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	1	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	1	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	0	35.0	0	0	373450	8.0500	S



```
[32] df = pd.get_dummies(df, columns=['Порт отправления'])
```

```
[33] df
```

	Id пассажира	Выжил	класс	Имя	Гендер	Возраст	# братьев и сестер / супругов на борту	# родителей/детей на борту	\tНомер билета	Стоимость проезда пассажира	Порт отправления_C	Порт отправления_Q	Порт отправления_S
0	1	0	3	Braund, Mr. Owen Harris	0	22.0	1	0	A/5 21171	7.2500	False	False	True
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	1	38.0	1	0	PC 17599	71.2833	True	False	False
2	3	1	3	Heikkinen, Miss. Laina	1	26.0	0	0	STON/O2. 3101282	7.9250	False	False	True
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.0	1	0	113803	53.1000	False	False	True

Жаңа белгілерді жасау. Feature Engineering

- Жаңа мағыналы функцияларды жасаңыз(мысалы, 'Отбасы мөлшері', 'Жалғыз жолаушы').

'Отбасы мөлшері' белгісін жасау

```
df['Величина семьи'] = df['# братьев и сестер / супругов на борту'] + df['# родителей/детей на борту'] + 1
```

'Жалғыз жолаушы' белгісін жасау

```
df['Одинокий пассажир'] = (df['Величина семьи'] == 1).astype(int)
```

The screenshot shows a Jupyter Notebook environment. On the left, a file explorer displays 'sample_data' containing 'cleaned_titanic.csv' and 'train.csv'. The main area shows three lines of Python code using pandas:


```
[38] df['Величина семьи'] = df['# братьев и сестер / супругов на борту'] + df['# родителей/детей на борту'] + 1
[39] df['Одинокий пассажир'] = (df['Величина семьи'] == 1).astype(int)
[40] df.head()
```

 Below the code, a preview of the 'train.csv' data is shown as a table with 14 columns: Id пассажира, Выжил, класс, Имя, Гендер, Возраст, # братьев и сестер / супругов на борту, # родителей/детей на борту, \tНомер билета, Стоимость проезда пассажира, Порт отправления_C, Порт отправления_Q, Порт отправления_S, and Величина семьи. The first three rows of data are visible.

Соңында, барлық деректер толтырылып, қалыпқа келтірілгенде, файлды сақтау керек. Ол үшін **Мәзірден** таңдаңыз **Файл-жүктеу-жүктеу IPYNB**(Файл-Скачать-Скачать IPYNB). Файлға атау беріңіз және сақтаңыз. Дәл осы файл тексеруге жіберілуі керек.

The screenshot shows the 'File' menu of the Jupyter Notebook interface. The menu options include:

- Показать расположение на Диске
- Открыть в режиме Playground
- Создать блокнот на Диске
- Открыть блокнот (Ctrl+O)
- Загрузить блокнот
- Переименовать
- Переместить
- Отправить в корзину
- Сохранить копию на Диске
- Сохранить копию как файл Gist для GitHub
- Создать копию в GitHub
- Сохранить (Ctrl+S)
- Сохранить и закрепить версию (Ctrl+M+S)
- История версий
- Скачать (highlighted)
- Печать (Ctrl+P)

 A sub-menu for 'Скачать' is open, showing two options: 'Скачать IPYNB' and 'Скачать PY'.

P.S.: Назар аударыңыз, мысалы, сіз Google Collab-та жұмыс істеп, содан кейін жұмысты тоқтатып, келесі күні кодты ашу арқылы жалғастыруды шешкен кезде (барлық код Google Drive-та Google Collab бумасында автоматты түрде сақталады), алдымен train.csv, файлын қайтадан жүктеп алуыңыз керек немесе сіз жұмыс істейтін файл атауы өзгеруі мүмкін. Жүктелгеннен кейін Мәзірден таңдау арқылы- Барлық кодты қайтадан іске қосу(Меню-Среда выполнения-Выполнить все) керек-жұмыс уақыты-бәрін

орындаңыз немесе оны басу арқылы барлық кодты қолмен іске қосып, белгілерді басу арқылы. Өзіңізге қаншалықты ыңғайлы солай таңдаңыз.

Жеке практикалық тапсырма .

Деректер жиынтығының сипаттамасы :

-Деректер жиынтығы **population.csv** файлында орналасқан, ол сондай-ақ осы тапсырма бойынша ДОТ платформасына жүктелген.

- Бұл файлда 2024 жылғы 1 шілдедегі жағдай бойынша Қазақстан Республикасының толықтырылған демографиялық деректері бар. Деректер жынысы мен рельеф түрі (қалалық және ауылдық) бойынша сегменттелген, бұл халықтың аймақтарда орналасуын талдауға мүмкіндік береді.

ТАПСЫРМА

Жоғарыдағы мысалда үйренген әдістерді қолдана отырып, төмендегі тапсырмаларды орындаңыз.

1. Бастапқы деректерді талдау

1.1. Деректер жинағын population.csv файлынан жүктеп алыңыз pandas кітапханасын пайдаланып. Файлды келесідей оқу керек:

```
df = pd.read_csv('population.csv', thousands='r',')
```

Файлды оқудың сәл өзгеше синтаксисі бар екенін ескеріңіз, абай болыңыз.

1.2. Жолдар мен бағандар санын көрсетіңіз

1.3. Әр бағандағы бос ұяшықтардың санын анықтаңыз.

2. Өткізілген деректермен жұмыс

2.1. Жетіспейтін деректерді қандай әдістермен толтыруға болатындығын сипаттаңыз.

2.2. Таңдалған әдістерді қолданыңыз және бос ұяшықтарды толтырыңыз.

3. Деректерді дайындау және статистикалық талдау

3.1. Деректерді дайындау үшін тағы қандай әдістерді қолдануға болады? Оларды сипаттаңыз.

3.2. Осы әдістерді осы деректер жиынтығына қолданыңыз.

3.3. Бағандар үшін медиана мен арифметикалық орташа мәнді табыңыз:

- "Жалпы халық"
- "Қала халқы"
- "Ауыл халқы"

Қосымша сұрақтар:

- Есептелген орташа және медианалық мәндер негізінде қандай қорытынды жасауға болады?
- Қала мен ауыл тұрғындарын бөлуде қандай тенденцияларды анықтауға болады?

Жұмысты қандай түрде тапсыру керек:

1. Аяқталған тапсырманы Google Colab-та орындау ұсынылады. Жұмысты аяқтағаннан кейін блокнотты (блокнотты) жүктеу керек (.Ipynb) Google Colab көмегімен және оны тапсырыңыз. 2. Сондай-ақ, әр тармақ бойынша тапсырма бойынша түсініктемелер мен жауаптары бар скриншоттармен есеп беру қажет. Скриндар курстағы талаптарға сай орындалуы керек, жүйелік жұмыс уақыты мен күнімен толық болуы керек. Егер скриндар талаптарға сәйкес келмесе, бағалау 50% - ға төмендейді.

Соңында сіз екі файлды, бірінші файл *.docx форматында тіркеуіңіз керек, бір файл *.ipynb форматта болу керек, файлдарды архивке(мұрағатқа) тіркемеңіз, өйткені архив(мұрағат) әрдайым ашылмайды.