DANO

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Т БАНК

Задача 1

На маркетплейсе «Е-шопинг» продаются различные товары. Одна из задач аналитика — прогнозировать, сколько товаров будет продаваться при определенной цене. В ходе исследований и экспериментов был выявлен вид зависимости:

$$Q(P) = Q_0 \times e^{E \times \frac{P_0 - P}{P_0}}$$

где Q — это количество проданных единиц товара при цене P,

Q_0 — количество проданных единиц товара при цене P_0,

E — коэффициент чувствительности количества проданных единиц товара к изменению цены.

1. Кофемашину «Кофе каждый день» купили 200 000 раз (Q_0) при цене 20 000 ₽ (P_0). Позже продавец поднял цену на 5 000 ₽, при этом продажи сократились на 24 000 штук. Какой коэффициент чувствительности E имеет этот товар? Ответ округлите до двух знаков после запятой.
2. Потом продавец решил поставить новую цену на эту же модель: 22 000 ₽. Сколько продаж согласно нашей зависимости будет у этого товара? Используйте результаты предыдущего пункта. Ответ округлите до целых.
3. Другой продавец предлагает на нашем маркетплейсе кухонные ножи и сковородки. Благодаря исследованиям были получены следующие формулы зависимостей количества проданных товаров:

$$Q_{\text{ножи}}(P) = 5000 \times e^{3.2 \times \frac{2000-P}{2000}}$$

$$Q_{\text{сковородки}}(P) = 3000 \times e^{2.05 \times \frac{4000-P}{4000}}$$

Найдите, сколько заработает продавец при цене по 3 000 ₽ за нож и сковороду при условии, что себестоимость ножа — 1 000 ₽, а сковородки — 2 000 ₽.Ответ округлите до целых.

Задача 2

Небольшой интернет-магазин собрал данные о действиях пользователей на своем сайте за последние несколько месяцев.

ecommerce_logs.csv — журнал действий пользователей:

- user_id — идентификатор пользователя.
- action — тип действия пользователя:
  - visit — посещение сайта;
  - click — клик на карточку товара;
  - cart — добавление товара в корзину;
  - delete — удаление товара из корзины;
  - purchase — покупка товаров.
- date_time — время совершения действия.
- product_id — идентификатор товара.
- quantity — количество добавленного в корзину товара.
- delivery_price — стоимость доставки.
- sex — пол пользователя.
- region — регион пользователя.
- price — цена товара.

Ваша задача — проанализировать поведение пользователей, выявить возможные проблемы при покупке и предложить решения. Ваш анализ поможет понять, на каком этапе воронки магазин теряет покупателей и какие изменения можно внести, чтобы улучшить процесс покупок в интернет-магазине.

Как правило, количество пользователей на каждом последующем шаге уменьшается, и такая ситуация называется “воронкой”. Конверсия — это отношение количества пользователей на каком-то одном шаге к количеству пользователей на одном из предыдущих шагов. Например, конверсия из визита сайта в добавление товара в корзину рассчитывается так: количество пользователей, добавивших товар в корзину, делится на количество пользователей, посетивших сайт.

Вам нужно изучить воронку конверсии, которая показывает, как пользователи переходят от одного шага к другому на сайте. В нашем случае воронка состоит из следующих шагов:

1. Посещение сайта.
2. Просмотр карточки товара.
3. Добавление товара в корзину.
4. Покупка.

1.) **Посчитайте конверсию** (округлите ответ до 3 знаков после запятой):

- Из визита на сайт в клик на карточку товара.
- Из клика в добавление в корзину.
- Из добавления в корзину в покупку.
- Из визита на сайт в добавление в корзину.
- Из визита на сайт в покупку.

2. **Постройте воронку конверсии** с помощью столбчатой диаграммы:
   - По оси X — шаги воронки.
   - По оси Y — количество уникальных пользователей на каждом шаге.
3. **Определите**, на каком этапе конверсия из предыдущего шага ниже всего.

Сформулируйте одну гипотезу, связанную с поведением пользователей, которая может объяснить падение конверсии именно на этом этапе. Обоснуйте механизм работы приведенной гипотезы.

4. **Постройте график динамики** (по оси X — дни) для каждой из конверсий:
   - Конверсия из визита в клик.
   - Конверсия из визита в добавление в корзину.
   - Конверсия из визита в покупку.
5. **На графике найдите просадку конверсии:** укажите, какая конверсия просела и в какой примерно период это произошло (допустимая погрешность — 1–3 дня).
6. Чем вызвано снижение конверсии в этот период? Какие изменения в бизнесе или поведении пользователей могли бы объяснить это? Ответьте на оба вопроса, опираясь на данные.

Задача 3

В аналитическом центре маркетплейса изучают удовлетворенность покупателей работой платформы. Владельцы бизнеса хотят, чтобы клиенты чаще покупали именно на их платформе. Для этого нужно повышать удовлетворенность клиентов.

Исследователи-аналитики провели работу и научились численно измерять удовлетворенность каждого клиента.

Стажеру аналитического центра поручили провести исследование и выявить, какие факторы сильнее влияют на удовлетворенность клиента. Другие аналитики хотят изучить его отчет и дать рекомендации бизнес-менеджерам, чтобы эффективнее развивать платформу с учетом клиентского опыта.

Для решения такой задачи стажеру предложили оценивать взаимосвязи методом построения множественной линейной регрессии. В общем виде такая задача описывается следующим уравнением:

$y = constant + a\_1 \times x\_1 + a\_2 \times x\_2 + ... + a\_n \times x\_n + eps,$

где y — зависимая переменная, x_1, x_2, ..., x_n — регрессоры (независимые переменные), constant — константа, a_1, ..., a_n — коэффициенты регрессии, eps — шум (случайная ошибка).

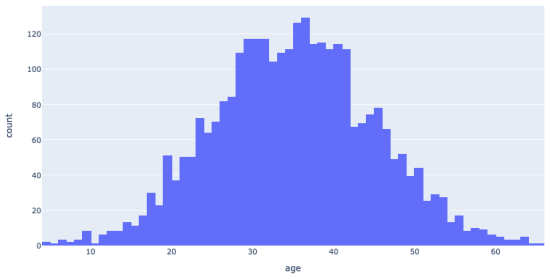
Ниже отчет стажера. При прочтении обратите внимание на неточности, противоречия и ошибки.

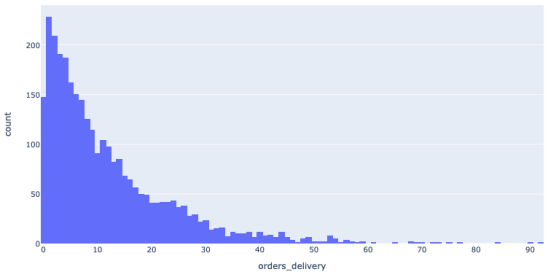
Я решил изучать факторы, влияющие на удовлетворенность, только для новых клиентов (которые первый раз совершили заказ не более 2 месяцев назад). Я собрал датасет, состоящий из следующих переменных:

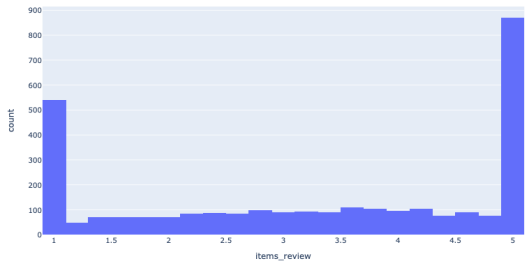
- satisfaction — удовлетворённость клиента.
- age — возраст клиента.
- orders_delivery — среднее время доставки за последний год в днях (для выполненных заказов это фактическое время доставки, для текущих заказов берется ожидаемое время доставки).
- items_review — средняя оценка товаров, которые покупал клиент (по шкале от 1 до 5, чем больше, тем лучше).

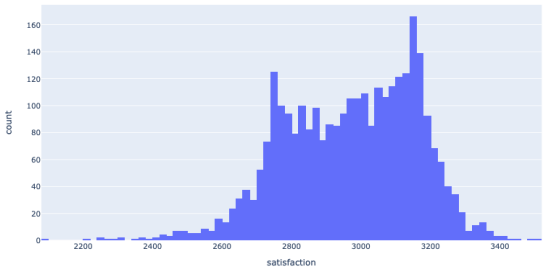
Дополнительно я добавил переменную orders_delivery_new — среднее время доставки за последний месяц (в днях), так как посчитал для новых клиентов эту метрику более релевантной.

На основе данных я построил визуализации с гистограммами, которые представлены ниже:

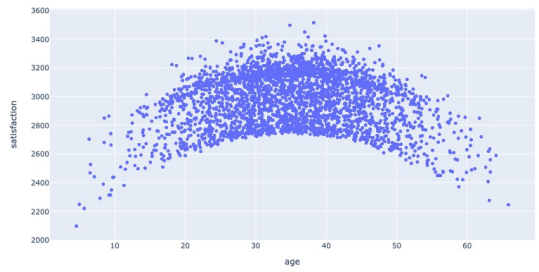


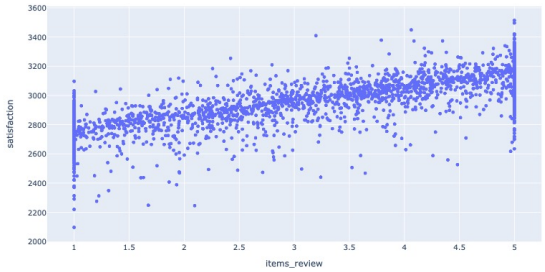


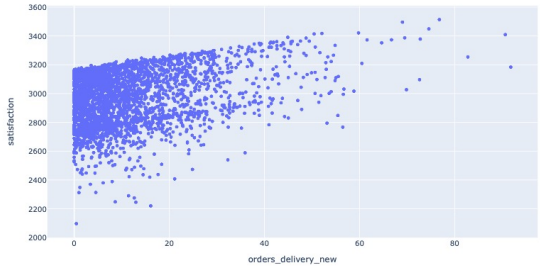




А также визуализировал зависимости между некоторыми переменными и удовлетворенностью клиента:







Для всех переменных я посчитал матрицу корреляций:

	satisfaction	age	orders_delivery	items_review	orders_delivery_new
satisfaction	1.000000	0.026678	0.329718	0.798202	0.330671
age	0.026678	1.000000	-0.015151	0.018060	-0.017924
orders_delivery	0.329718	-0.015151	1.000000	0.028127	0.996450
items_review	0.798202	0.018060	0.028127	1.000000	0.030646
orders_delivery_new	0.330671	-0.017924	0.996450	0.030646	1.000000

Я начал основную часть исследования с построения линейной регрессии переменной удовлетворенности от среднего времени доставки за последний месяц и получил следующие результаты, оценив регрессионную модель 1:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           satisfaction   R-squared:                       0.109
Model:                          OLS   Adj. R-squared:                  0.109
Method:               Least Squares   F-statistic:                     368.1
Date:              Sat, 16 Nov 2024   Prob (F-statistic):           1.81e-77
Time:                      21:52:38   Log-Likelihood:                -19842.
No. Observations:              3000   AIC:                         3.969e+04
Df Residuals:                  2998   BIC:                         3.970e+04
Df Model:                         1
Covariance Type:          nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
orders_delivery_new     5.5102      0.287     19.185      0.000       4.947       6.073
constant             2907.8210      4.665    623.354      0.000    2898.674    2916.968
```

P-value (на картинке обозначено как P > |t|) стремится к 0, значение коэффициента перед единственной независимой переменной равно 5,5102, значит, среднее время доставки заказа значимо положительно влияет на удовлетворенность клиента на маркетплейсе.

Но для корректной оценки модели в нее важно добавлять также факторы, которые имеют уже доказанную связь с исследуемой переменной. На графиках из предварительного анализа я увидел явную взаимосвязь возраста с удовлетворенностью — и решил добавить эту переменную в качестве контрольной переменной в модель. Получил следующие результаты оценки регрессионной модели 2:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           satisfaction   R-squared:                       0.110
Model:                            OLS   Adj. R-squared:                  0.110
Method:                 Least Squares   F-statistic:                     186.0
Date:                Sat, 16 Nov 2024   Prob (F-statistic):           7.30e-77
Time:                        21:52:43   Log-Likelihood:                -19840.
No. Observations:                3000   AIC:                         3.969e+04
Df Residuals:                    2997   BIC:                         3.970e+04
Df Model:                           2
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
age                         0.6321      0.334      1.893      0.058      -0.023       1.287
orders_delivery_new         5.5200      0.287     19.224      0.000       4.957       6.083
constant                 2885.8354     12.516    230.563      0.000    2861.294    2910.377
-------------------------------------------------------------------------------------------
```

По оценке второй регрессионной модели видно, что коэффициент перед возрастом незначимый (p-value > 0,01). Но я решил оставить эту переменную в дальнейшем исследовании, поскольку взаимосвязь между возрастом и удовлетворенностью точно присутствует, просто вторая модель не смогла ее распознать.

Затем я решил добавить вторую важную переменную для изучения: среднее значение отзывов о купленных товарах. Предполагаю, что, если клиент выбирает товары с высокими средними отзывами, то он будет больше удовлетворен работой маркетплейса. Результаты оцененной регрессионной модели 3 приведены ниже:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           satisfaction   R-squared:                       0.731
Model:                            OLS   Adj. R-squared:                  0.731
Method:                 Least Squares   F-statistic:                     2718.
Date:                Sat, 16 Nov 2024   Prob (F-statistic):               0.00
Time:                        21:52:47   Log-Likelihood:                -18045.
No. Observations:                3000   AIC:                         3.610e+04
Df Residuals:                    2996   BIC:                         3.612e+04
Df Model:                           3
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
age                         0.3476      0.184      1.893      0.058      -0.012       0.708
orders_delivery_new         5.1129      0.158     32.378      0.000       4.803       5.423
items_review               99.1536      1.192     83.204      0.000      96.817     101.490
constant                 2571.2987      7.850    327.545      0.000    2555.906    2586.691
-------------------------------------------------------------------------------------------
```

Все переменные, кроме возраста, значимые. Увидев это, я захотел понять, что больше влияет на удовлетворенность: среднее время доставки заказов за последний год или за последний месяц. Для этого я добавил обе эти переменные в регрессионную модель 4. Результаты оценки модели 4 приведены ниже:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           satisfaction   R-squared:                       0.732
Model:                          OLS   Adj. R-squared:                  0.732
Method:               Least Squares   F-statistic:                     2044.
Date:              Sat, 16 Nov 2024   Prob (F-statistic):               0.00
Time:                      21:53:36   Log-Likelihood:                -18041.
No. Observations:              3000   AIC:                         3.609e+04
Df Residuals:                  2995   BIC:                         3.612e+04
Df Model:                         4
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
age                0.3320      0.184      1.809      0.071      -0.028       0.692
orders_delivery_new     0.2477      1.875      0.132      0.895      -3.428       3.923
orders_delivery    4.8930      1.879      2.604      0.009       1.209       8.577
items_review      99.2444      1.191     83.325      0.000      96.909     101.580
constant        2571.3582      7.843    327.867      0.000    2555.981    2586.736
```

Завершив построение моделей, я пришел к следующим выводам:

1. Возраст не взаимосвязан с удовлетворенностью клиента на маркетплейсе.
2. Средняя оценка купленных пользователем товаров в 20 раз сильнее влияет на удовлетворенность пользователя, чем среднее время доставки заказов за последний год.
3. Среднее время доставки товаров за последний месяц не взаимосвязано с удовлетворенностью клиента на маркетплейсе.

Какие ошибки в отчете допустил стажер? Найдите 3 ключевые ошибки, объясните, в чем они заключаются и как их можно исправить. Попробуйте сформулировать корректные выводы по построенным моделям, а если это невозможно — объясните почему. В ответе записывайте каждую новую ошибку отдельным пунктом.

Задача 4

Аналитик магазинчика «Ерунда, но», прогнозирует продажи. Все товары магазина делятся на три категории (каждый товар входит только в одну категорию): аксессуары для телефонов (A), веселые стикеры (B) и сборники сканвордов (C). Дальше эти категории будем называть A, B и C.

В первый день своей работы аналитик решил просто посмотреть на графики и понять, какие метрики для оценки качества прогнозов использовать. Временные ряды продаж товаров каждой из трех категорий за последний месяц выглядят следующим образом (сборники сканвордов почему-то почти никто не покупает):

Аналитик выбирает между тремя стандартными метриками для оценки среднего размера ошибки прогнозов: MAE (Mean Absolute Error, средняя абсолютная ошибка), RMSE (Root Mean Squared Error, корень из среднеквадратической ошибки) и MAPE (Mean Absolute Percentage Error, средняя абсолютная процентная ошибка).

$$MAE = \frac{\sum_{t=1}^{T} |Y_t - \hat{Y}_t|}{T}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2}{T}}$$

$$MAPE = \frac{1}{T} \sum_{t=1}^{T} \frac{|Y_t - \hat{Y}_t|}{Y_t},$$

где Y_t — фактические значения ряда, $\hat{Y}_t$ — прогнозные, T — общее число наблюдений.

1. Какие рекомендации о возможности использования этих (MAE, RMSE и MAPE) метрик для каждого из трех рядов вы можете дать на основе этого графика? Учтите, что эти три метрики обладают разной чувствительностью к особенностям ряда и подбор конкретной метрики должен опираться на свойства самого ряда.

   На следующий день аналитик решил, что задача самостоятельного построения прогнозов для него слишком сложная и стоит нанять консалтинговую компанию для помощи. Консультанты из компании «Иннокентий и друзья» внимательно изучили данные, обучили ИИ-модель для предсказания продаж товаров группы A, протестировали ее на последней неделе данных и получили следующие результаты:

   В качестве метрики консультанты предлагают использовать R^2 (отношение дисперсии прогноза к дисперсии факта), и у предлагаемой модели оно составляет 0,46. Аналитик помнит, что R^2 всегда лежит от 0 до 1, и полученный консультантами результат вроде бы весьма неплох, но что-то его смущает.

2. Какую ошибку допустили консультанты в выборе метрики? Предложите любую простейшую модель, которая даст более точный по MAE/RMSE/MAPE прогноз, чем полученный консультантами.

   Аналитик понял: если хочешь сделать что-то хорошо — сделай это сам! После долгой и кропотливой работы он подобрал четыре хорошие модели: для прогнозирования продаж каждой из трех категорий товаров и отдельно для прогнозирования суммарных продаж. Прогнозы всех четырех моделей на пять дней вперед приведены в таблице (все продажи — в штуках, в том числе и суммарные). Суммировать аксессуары, стикеры и сканворды от аналитика потребовало начальство, ему этот показатель нужен!):

Продажи A	**Продажи B**	**Продажи C**	**Суммарные продажи**
21	11	0	30
23	10	5	41
22	34	0	49
21	10	0	28
19	13	0	32

Внимательно изучив таблицу, аналитик обнаружил еще один недочет в своей методологии, связанный со структурой полученного набора прогнозов.

3. Что это за недочет? Предложите простой способ его исправить, не требующий дополнительной информации. Предложите еще один способ исправить этот недочет — в ситуации, когда нам известны точности для всех четырех моделей и они значительно различаются.

Задача 5

На нашем маркетплейсе мы решили торговать не только вещами, но и продуктами питания. Часть из них скоропортящаяся, например молочные продукты, фрукты и овощи. Мы решили закупать такие товары у поставщиков по более низкой цене и продавать их самостоятельно. Чтобы получать максимальную прибыль, нам необходимо как можно точнее предсказывать спрос покупателей. Если мы закупим больше товара, чем сможем продать, он может испортиться, и мы потеряем стоимость закупки.

Закупки производятся на каждый день. Вам дана таблица покупок клиентов 10 различных товаров на протяжении двух месяцев, а также выручка с одного товара и его закупочная стоимость. Учитываете, что цена закупки и продажи остается неизменной и сезонность закупки товара тоже не меняется. Предскажите на семь дней вперед, сколько товаров надо закупать, чтобы получить максимальную прибыль. Прибыль считается по формуле:

$$\sum_{7day=1} \sum_{10i=1} \min(Y_i, Ypred_i) \times revenue_i - Ypred_i \times cost_i$$

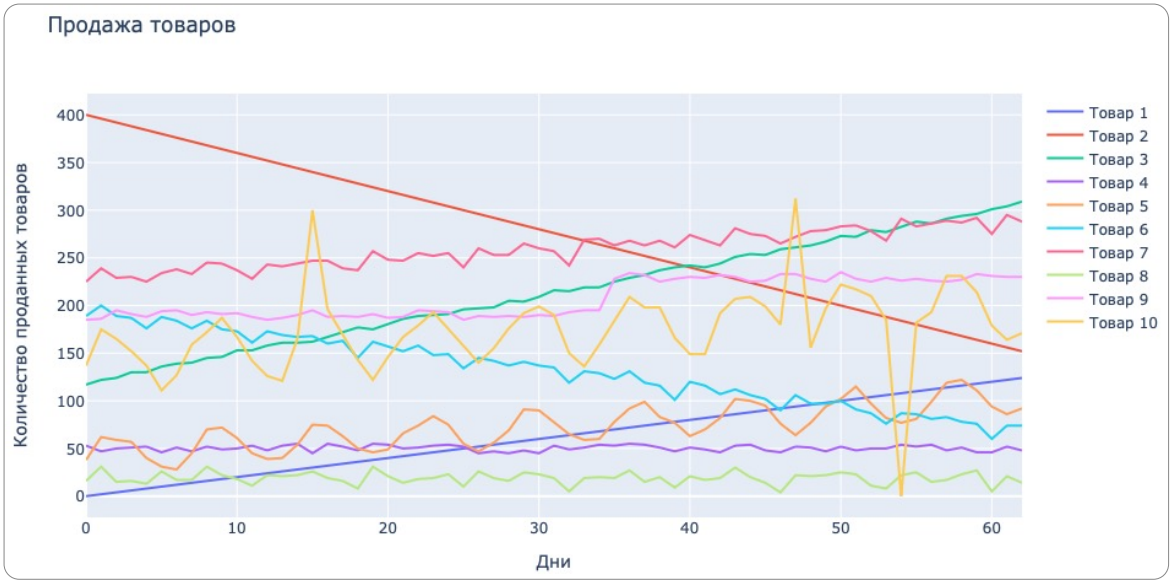
где Y_i — фактические спрос i-го товара, $Ypred_i$ — количество закупленного товара i, $revenue_i$ — доход с продажи одного товара i, $cost_i$ — цена закупки одной единицы товара.

Цены закупки и продажи товаров

Товар	Товар 1	Товар 2	Товар 3	Товар 4	Товар 5	Товар 6	Товар 7	Товар 8	Товар 9	Товар 10
Цена закупки 1-го товара	20	80	10	150	60	90	40	35	60	130
Доход с продажи 1-го товара	25	100	15	170	75	130	55	40	70	150

Дополнение: для прогноза временного ряда можно использовать различные подходы и методы. Вы можете пользоваться любым удобным для вас. Если раньше вы этим не занимались, можно оттолкнуться от линейной регрессии на прошлых значениях ряда. Чтобы получить наиболее точный прогноз, вам нужно понять, какое количество предыдущих компонентов ряда брать для прогноза и нужны ли они все, и подобрать к ним коэффициенты регрессии. Основные параметры, которые стоит учитывать, — это тренд и переодичность. Также есть случайный шум, который нельзя описать моделью.

Даны файл task_4_before и график (на графике нумерация товаров идет по порядку и начинается с нуля) ежедневных продаж товаров



Ответ дайте в виде таблицы в формате xlsx или в таблице в документе вида:

День	**Товар1**	**Товар2**	**...**	**Товар 10**
63				
64				
65				
66				
67				
68				
69				

Дополнение: Баллы по данному пункту будут даваться пропорционально приближению к максимальной прибыли с продаж товаров за предсказанные вами семь дней. в эти семь дней товары продаются по тем же законам, что и на предыдущих 63-х днях.