Хакатон “DANO”

Исследовательский проект

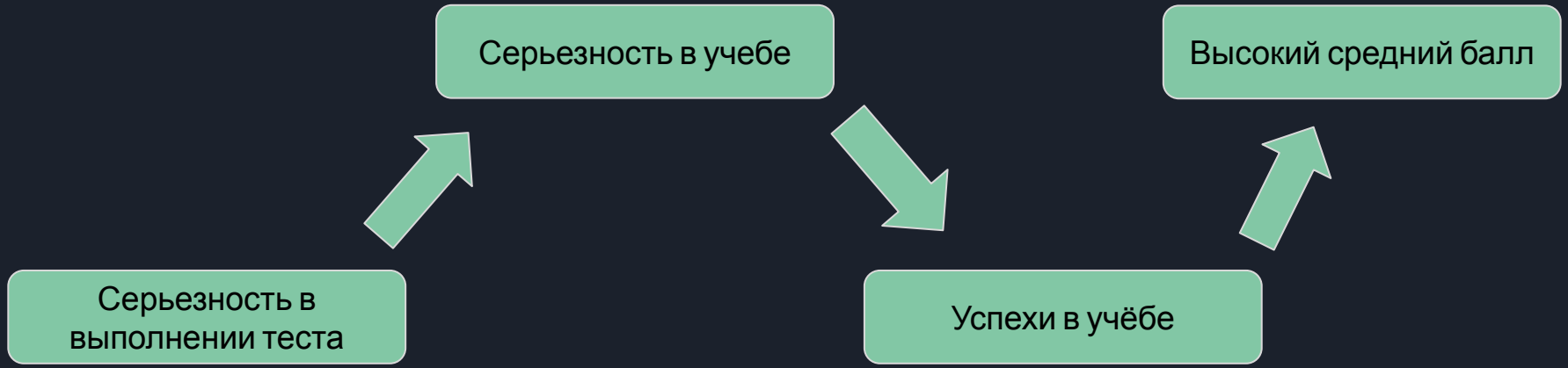
Роман Шваров
Павел Воробьёв
Арсений Чапайкин
Александр Смирнов
Марина Ильина
Мария Яхричева

Введение

Исследовательский вопрос: как стать успешным в учёбе?

Гипотеза: В среднем, студенты, серьезно подошедшие к заполнению теста, имеют более высокий средний балл.

Механизм гипотезы



Предварительный анализ

- выброс степеней выше магистра по причине отсутствия у них gpa
- построение таблицы корреляции всех данных - промежуточные выводы по ней (например, количество времени на отдых и сон, а также количество времени на учебу и работу коррелируют с GPA)

week_1	week_2	week_3	week_4	week_5	week_6
-0.144227	-0.000771	-0.148562	-0.111375	0.114834	0.314263

Предварительный анализ

Аномалии базы данных:

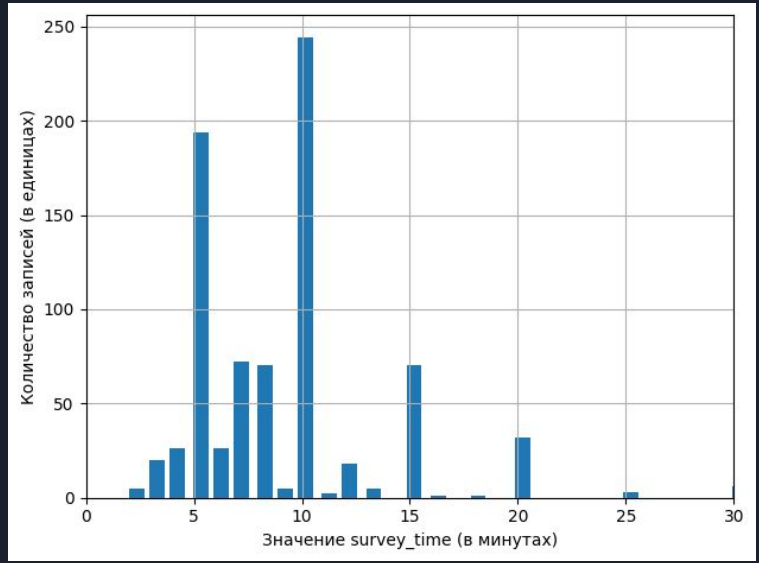
- максимальное кол-во семестров - 28 семестров
- максимальный возраст - 71 год
- макс. время прохождения теста - 240 минут
- максимальное количество сна “вчера” - 54
- “вчера” - 30 часов на подготовку к занятиям + 20 часов на обучение (в сутках 24 часа)

- встречаются у бакалавров и магистров

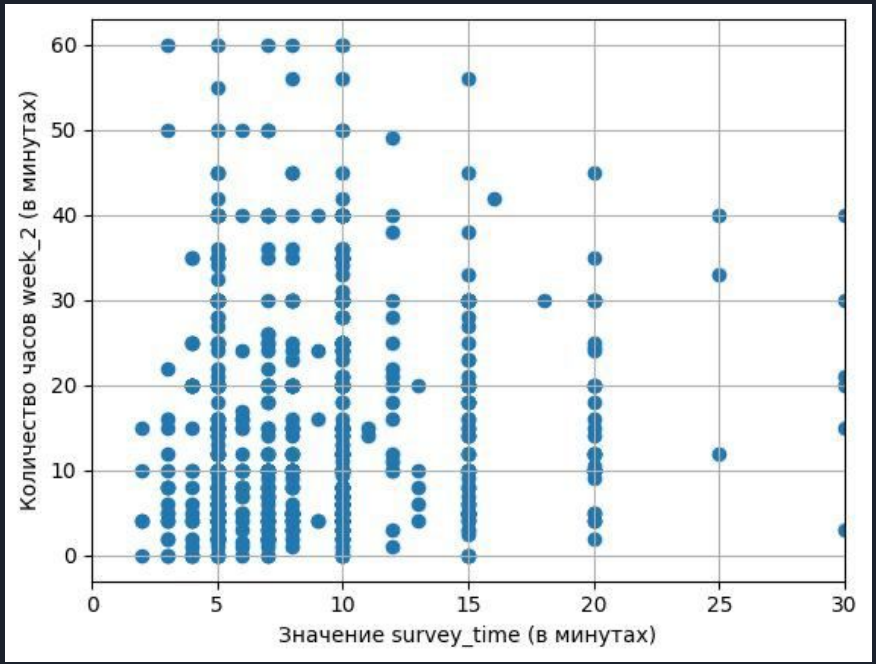
Коэффициент серьёзности

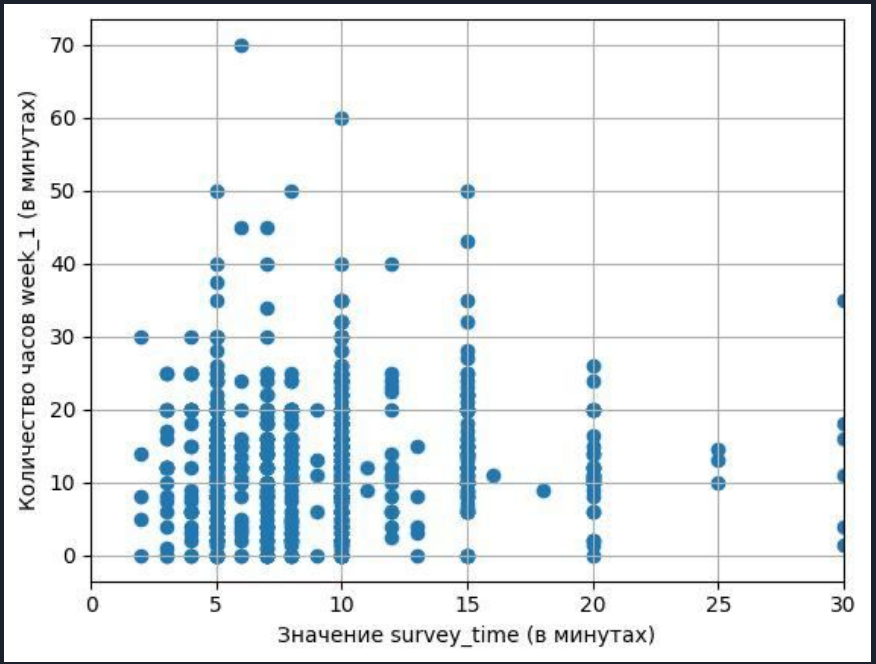
- Кол-во часов в неделю (week) < 168
- Кол-во часов в день (yesterday) < 24
- Часы сна (enough_6) < 20
- Кол-во пустых полей ≤ 2
- Время выполнения теста

  (5 ≤ survey_time ≤ 25)



Графики зависимости часов учебы от survey_time





Используемые величины

Эндогенная переменная	Обозначение в базе данных	Особенности измерения
средний балл	gpa	1 - лучший 4 - худший

Экзогенные переменные	Обозначение	Особенности измерения
коэффициент серьёзности	k	>0

Сравнение бакалавров и магистров

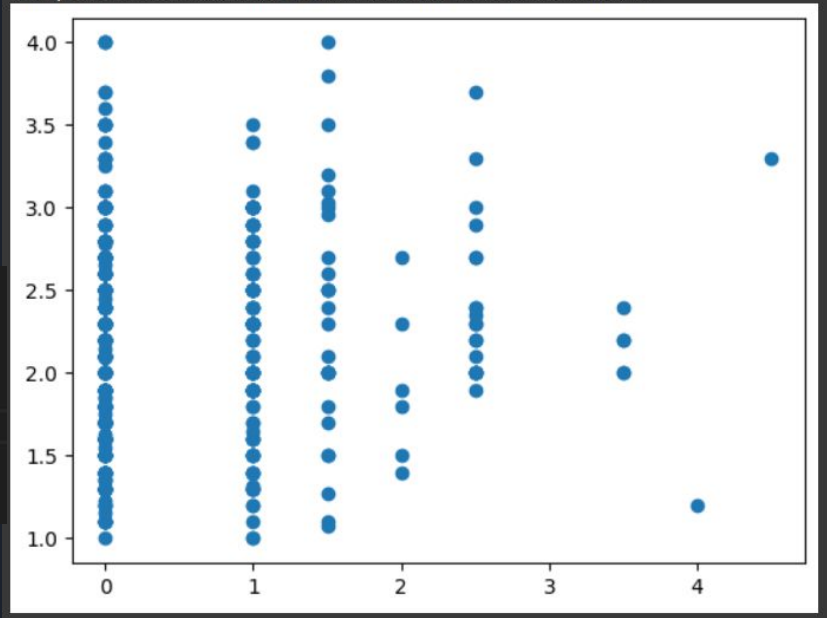
	Серьезные		Несерьезные	
	Количество	Ср. балл	Количество	Ср. балл
Бакалавры	435	2.17269	125	2.306596
Магистры	204	1.687892	55	1.792593

Прямая зависимость между серьезностью и gpa не была подтверждена

Использовались кросс-валидация и линейная регрессия

```
from sklearn.model_selection import KFold, cross_val_score
model = LinearRegression()
X = X.fillna(0)
kfold = KFold(n_splits=10,random_state=7,shuffle=True)
cv_results = cross_val_score(model, X,Y,cv=kfold,scoring='r2')
print('r2_score:',cv_results.mean())
plt.scatter(X,Y)
```

```
r2_score: -0.009378148987534318
<matplotlib.collections.PathCollection at 0x7ce0773ded40>
```



Проверка статистической значимости результата

Итоги U-теста Манна-Уитни

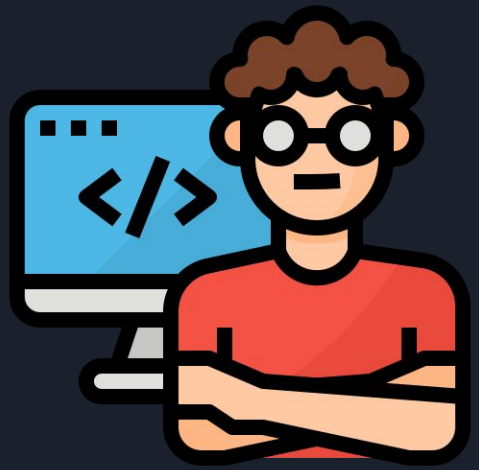
```
[ ]  import scipy.stats as stats
     dfs = pd.read_excel("/content/bach_ser.xlsx")
     dfuns = pd.read_excel("/content/mag_ser.xlsx")
     sum=0
     for i in range(3):
       group1,group2=dfs['gpa'].loc[i*30:(i+1)*30],dfuns['gpa'].loc[i*30:(i+1)*30]
       sum+=stats.mannwhitneyu(group1, group2).pvalue
     print(sum/3)
```

```
0.0014433643897318771
```

0.0014 < 0.05, значит результат не является статистически значимым

Результаты

Наша гипотеза подтвердилась, однако разница не является статистически значимой. Нет общей зависимости gpa от коэффициента серьезности для всех студентов. При этом для студентов бакалавриата эта зависимость проявляется сильнее, чем для студентов магистратуры



Перспективы и применимость исследования

- Результат не является статистически значимым, следовательно, у нас нет оснований, чтобы считать это общей особенностью
- В дальнейшем может применяться метод подсчета коэффициента серьезности
- Альтернативный подход для анализа: мы анализировали не данные, а способ их получения