**Национальный исследовательский университет «Высшая школа экономики» Московский институт электроники и математики им. А.Н. Тихонова, Москва**

Направление 01.03.04. Прикладная математика Бакалаврская программа "Прикладная математика"

Отчет о самостоятельной работе по дисциплине «Статистический анализ в Python»

Бригада №2

Зябкин Даниэль, 3 курс

Борисов Эрдни, 3 курс

Москва 2024

**Оглавление**

[1 Общая постановка задачи 3](#)

[2 Предварительный анализ собранных данных 4](#_jp77rkqdylxi)

[2.1 Анализ особенностей данных: потенциальные ошибки и пропущенные значения, 4](#_ou7xxe194tcw)

[группы и выбросы 4](#_i5c2so2mjcg7)

[2.1.1 Анализ количественных переменных 6](#_ne8etbv6xcfv)

[2.1.2. Анализ качественных переменных. 10](#_lj2vi3thi0qr)

[3. Анализ статистической связи. 12](#_585vdph83dzh)

[3.1. Графический анализ связи качественных переменных 15](#_2y99tjbo6hf3)

[3.2. Графический анализ пары “числовая зависимая переменная - числовая независимая переменная” 16](#_p1bv2ahlo8rn)

[3.3. Графический анализ связи двух независимых переменных 18](#_f91pbia7oqo2)

## 

## 1 Общая постановка задачи

В таблице 1 представлена информация о рейтинге университетов в зависимости от различных характеристик

| № | Характеристика объекта/явления | Название переменной | Шкала измерения | Роль: зависимая/независимая |
| --- | --- | --- | --- | --- |
| 1 | Место в рейтинге | University rank | Номинальная |  |
| 2 | Название университета | Name of University | Номинальная |  |
| 3 | Локация | Location | Номинальная |  |
| 4 | Количество студентов | No of student | Относительная |  |
| 5 | Количество студентов на преподавателя | No of student per staff | Относительная |  |
| 6 | Количество международных студентов | International Student | Относительная |  |
| 7 | Отношение женщины : мужчины | Female:Male Ratio | Относительная |  |
| 8 | Общая оценка | OverAll score | Относительная |  |
| 9 | Оценка качества преподавания | Teaching Score | Относительная |  |
| 10 | Оценка исследовательской деятельности | Research Score | Относительная |  |
| 11 | Цитируемость | Citation Score | Относительная |  |
| 12 | Доходы от индустрии | Industry Income Score | Относительная |  |
| 13 | Международная интеграция | International Outlook Score | Относительная |  |

*Таблица 1. Описание факторов, учтенных в анализе.*

Исследуемые данные были получены из набора данных, взятых с сайта:

<https://www.kaggle.com/datasets/alitaqi000/world-university-rankings-2023>

## 2 Предварительный анализ собранных данных

## 2.1 Анализ особенностей данных: потенциальные ошибки и пропущенные значения,

## группы и выбросы

Исследуемыми данными является рейтинг университетов мира за 2023 год. Всего имеем 3 номинальных и 10 относительных переменных. Стоит отметить, что некоторые переменные были изменены для того, чтобы получить больше категориальных переменных, некоторые переменные были удалены. Ниже будет изложено, какие манипуляции были проведены с исходным датасетом:

* Переменная Industry Income Score была заменена на переменную Industry Income, значением которой может быть либо ‘низкий’ либо ‘средний’ либо ‘высокий’. Эта переменная характеризует доходы университета, заработанные вследствие выполнения каких-либо работ в той или иной индустрии.
* Значения переменной International Student были также заменены на ‘низкий’, ‘средний’ либо ‘высокий’, которые характеризуют количество иностранных студентов в учебном заведении.
* Значения переменной Female:Male Ratio были заменены на ‘Более мужской’, ‘Более женский’ или ‘Сбалансированный’. Она показывает соотношение между студентами мужского и женского пола.

Были удалены переменные:

* University rank (Место в рейтинге), потому что эта переменная является производной OverAll Score (Место в рейтинге - нумерация университетов в порядке возрастания на основе OverAll Score: чем больше общая оценка - тем выше в рейтинге располагается вуз)
* Teaching Score
* Research Score
* Citation Score
* International Outlook Score

Последние 4 перечисленные поля удалены в связи с тем, что они - составляющие наиболее важной переменной OverAll Score

В ходе предварительного анализа были удалены строки с пропущенными записями, также в исходных данных значением той или иной переменной было ‘NaN’, они тоже были удалены. Кроме того, были удалены элементы выборки, в значениях переменных которых наблюдались выбросы (удалены они были по правилу трех сигм). В итоге количество переменных стало равно 8, количество элементов выборки сократилось до 1433.

После вышеперечисленных действий с данными таблица 1 приняла следующий вид:

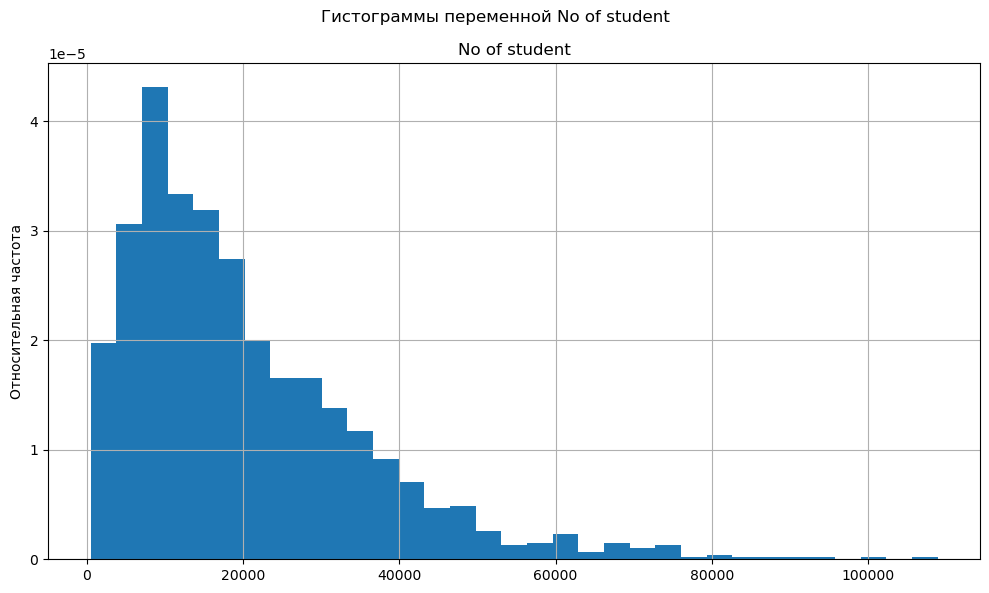
| № | Характеристика объекта/явления | Название переменной | Шкала измерения | Роль: зависимая/независимая |
| --- | --- | --- | --- | --- |
| 1 | Название университета | Name of University | Номинальная | Независимая |
| 2 | Локация | Location | Номинальная | Независимая |
| 3 | Количество студентов | No of student | Относительная | Независимая |
| 4 | Количество студентов на преподавателя | No of student per staff | Относительная | Независимая |
| 5 | Количество международных студентов | International Student | Относительная | Независимая |
| 6 | Отношение женщины : мужчины | Female:Male Ratio | Относительная | Независимая |
| 7 | Общая оценка | OverAll score | Относительная | Зависимая |
| 8 | Доходы от индустрии | Industry Income | Относительная | Независимая |

*Таблица 2. Описание модифицированных факторов, учтенных в анализе.*

## 2.1.1 Анализ количественных переменных

1. *No of student*

No of student - переменная, показывающая количество студентов в данном университете в 2023 году. На рисунке 1 можно наблюдать распределение этого показателя по Фридману-Диаконису.



*Рис. 1, гистограмма количества студентов по Фридману-Диаконису*

Таблица 3 отражает описательную статистику количества студентов в университетах

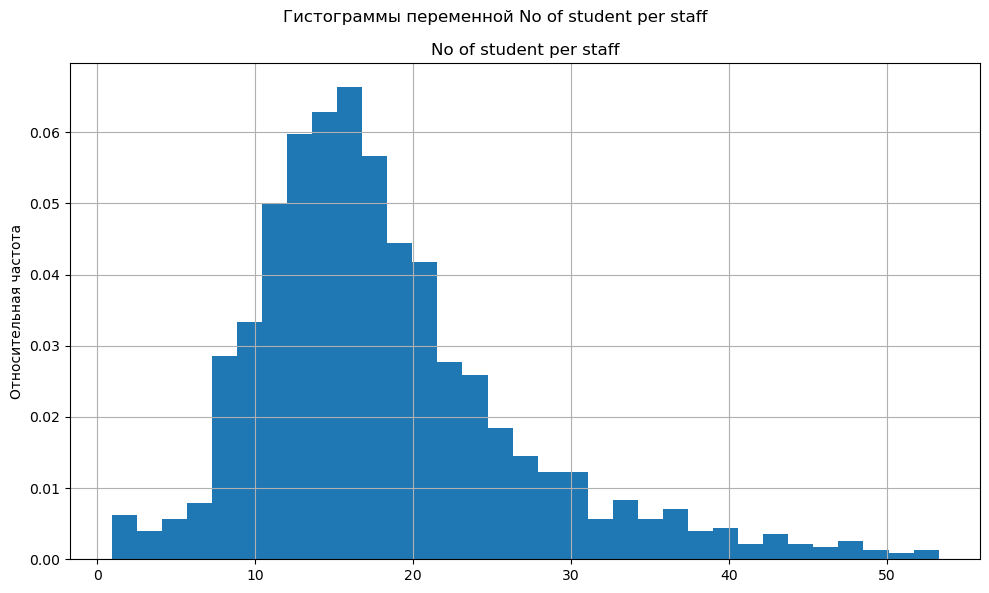
| count | 1433 |
| --- | --- |
| mean | 22802,5188172043 |
| std | 28763,0483402715 |
| min | 489 |
| 25% | 9212,25 |
| 50% | 16659 |
| 75% | 28637,75 |
| max | 460632 |
| IQR | 19425,5 |
| median | 16074 |
| asymmetry | 1,542619 |

*Таблица 3, описательная статистика переменной No of student*

На приведенной выше диаграмме можно наблюдать асимметрию вправо, что подтверждается описательной статистикой (среднее больше медианы, а также коэффициент асимметрии). На самом деле, это вполне закономерно, так как университеты с наибольшим количеством студентов из некоторых стран Африки и Южной Америки, а в рассматриваемом рейтинге большая часть вузов из США и стран Европы. Стоит отметить, что все-таки в университетах преобладают студенты из страны, в которой расположен вуз.

1. *No of student per staff*

No of student per staff - количество студентов на одного сотрудника университета. Распределение считалось по формуле Фридмана-Диакониса, его можно увидеть на рисунке 2.



*Рис. 2, гистограмма количества студентов на одного преподавателя.*

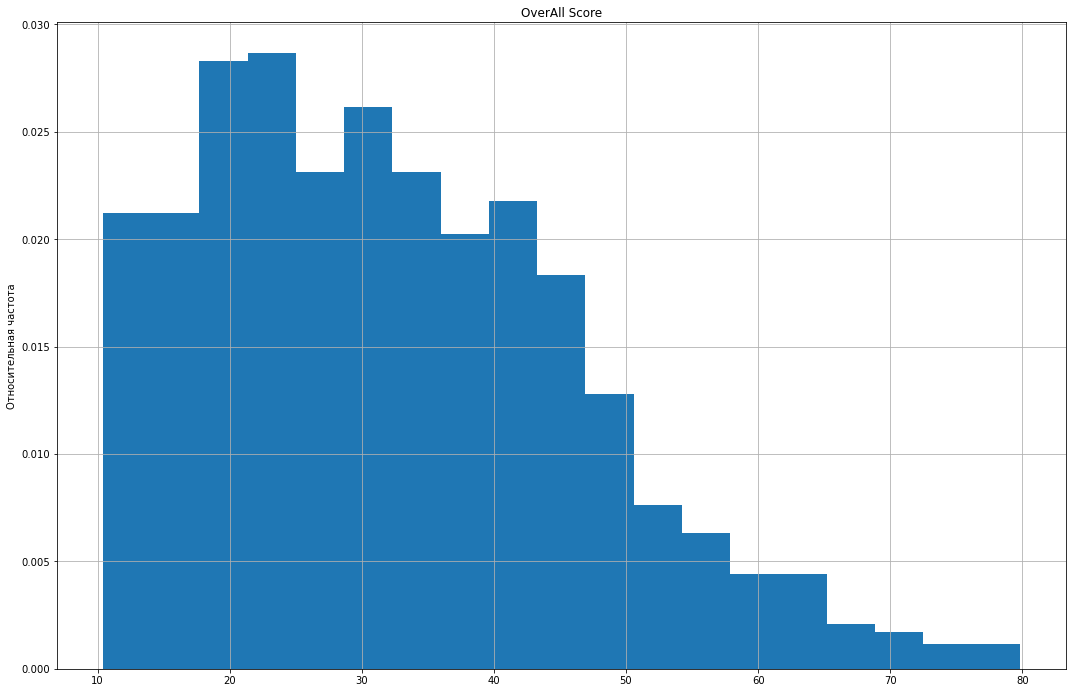
| count | 1433 |
| --- | --- |
| mean | 18,8331989247312 |
| std | 11,6644027641681 |
| min | 0,9 |
| 25% | 12,7 |
| 50% | 16,5 |
| 75% | 22,1 |
| max | 232,2 |
| IQR | 9,4 |
| median | 16,5 |
| asymmetry | 1,199584 |

*Таблица 4, описательная статистика переменной No of student per staff*

По приведенному графику можно заметить асимметрию вправо. Это подтверждает таблица основных статистик. Среднее больше медианы и коэффициент асимметрии равен 1.199584. Асимметрия вправо характеризует то, что в лучшее качество преподавания в университете достигается более индивидуальной работе со студентами. Отсюда и вытекает, что в большинстве рассматриваемых вузов количество студентов на преподавателя менее 23-х (ориентируемся на 3-й квартиль)

1. *OverAll score*

OverAll score - общая оценка университета, переменная, прямым и непосредственным образом влияющая на положение вуза в рейтинге. Гистограмма ниже показывает распределение этой переменной по Фридману-Диаконису. На ней можно увидеть пустые зоны в окрестностях 0, 20, 30 и 40. Такое происходит от того, что в исходных данных оценки университетов, располагающихся во второй половине рейтинга были даны диапазоном. Мы же при предварительных манипуляциях с данными присвоили каждому из таких университетов оценку, равную среднему значению указанного диапазона.



*Рис. 3, гистограмма общей оценки университетов.*

| count | 1433 |
| --- | --- |
| mean | 33,2549731182796 |
| std | 15,6384533355956 |
| min | 14,35 |
| 25% | 21,35 |
| 50% | 31,85 |
| 75% | 43,5 |
| max | 96,4 |
| IQR | 22,15 |
| median | 31,85 |
| asymmetry | 0.679145 |

*Таблица 4, описательная статистика переменной OverAll score*

На гистограмме можем наблюдать то, что оценок больше 50 менее 25 процентов от всей выборки. Это говорит о том, что в выборке сильно выделяются университеты, расположенные на верхних строчках составленного рейтинга, то есть об элитарности вузов из топа рейтинга.

## 2.1.2. Анализ качественных переменных.

1. *Name of University*

Name of University - переменная, имеющая исключительно формальный характер, наименование университета. Для каждого вуза она уникальная.

1. *Location*

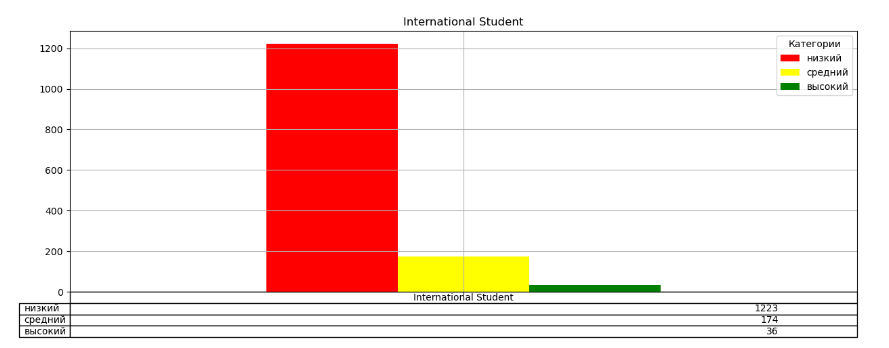
Location - страна, в которой расположен университет.

3 страны с наибольшим количеством рассматриваемых вузов:

1. США - 148 университетов;
2. Япония - 112 университетов;
3. Великобритания - 85 университетов.
4. *International Student*

International Student характеризует сколько иностранных студентов учится в вузе.

Может принимать следующие значения: ‘низкий’, ‘средний’ либо ‘высокий’.



*Рис. 4, гистограмма количества студентов в университете.*

По приведенному графику, можно заметить, что в большей части университетов учится малое число иностранных студентов. Этому есть ряд причин:

1. Финансовый аспект:

Обучение за рубежом может быть дорогостоящим для студентов

1. Языковой барьер:

Неспособность справляться с учебными программами на языке, отличном от родного,может быть значительным препятствием для иностранных студентов.

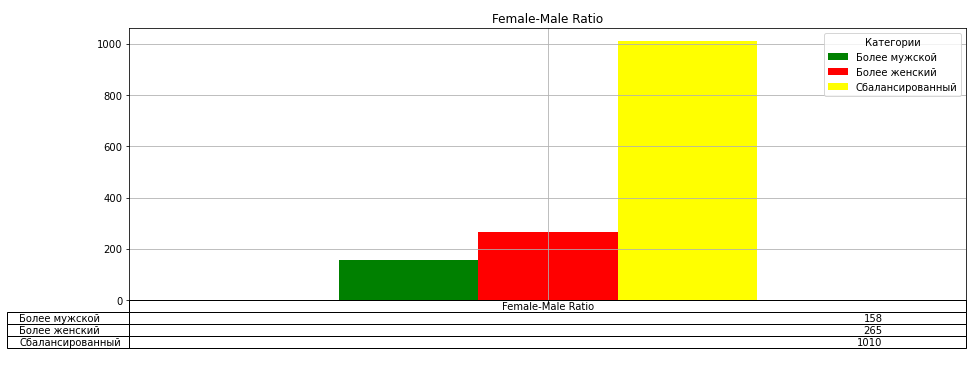
1. Иммиграционная политика:

Ограничения на визы для студентов могут сделать невозможным обучение иностранных студентов в университетах определенных стран.

1. *Female:Male Ratio*

Female:Male Ratio - пропорции студентов женского и мужского пола.

Принимает значения: ‘Более мужской’, ‘Более женский’ или ‘Сбалансированный’.



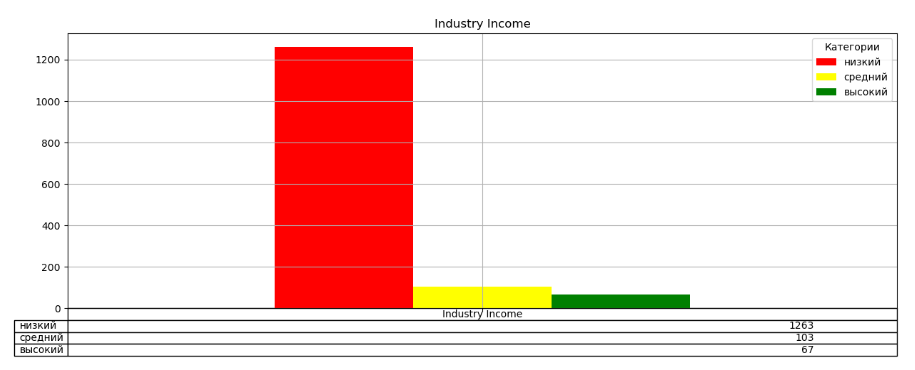
*Рис. 5, гистограмма пропорции женщин и мужчин*

По приведенному графику, можно заметить, что большинство имеют сбалансированное количество мужчин и женщин. Но некоторые университеты имеют больше мужчин или женщин. Факторы, влияющие на это, включают культурные нормы, политику приема, доступность образования для разных групп населения, а также социально-экономические факторы. Например, в некоторых областях обучения, таких как инженерия, информационные технологии, может быть больше мужчин, в то время как в других, таких как медицина или образование, может быть больше женщин.

1. *Industry Income*

Industry Income - характеризует доходы университета, получаемые от разных индустрий, за 2022 - 2023 гг.

Принимает значения: ‘низкий’, ‘средний’ либо ‘высокий’.



*Рис. 6, гистограмма прибыли университетов от индустрии.*

По графику, можно заметить, что большинство университетов имеют низкие доходы от индустрии. Большинство университетов имеют низкие доходы от различных промышленных секторов по нескольким причинам:

1. Основная цель: Основной целью университетов является образование, наука и исследования, а не максимизация прибыли от промышленных секторов. Университеты часто фокусируются на обучении студентов, проведении академических исследований и развитии новых знаний, а не на коммерческих интересах.

2. Финансовые ограничения: Многие университеты имеют ограниченные финансовые ресурсы, которые могут затруднять развитие коммерческих проектов или инвестиции в промышленные сектора.

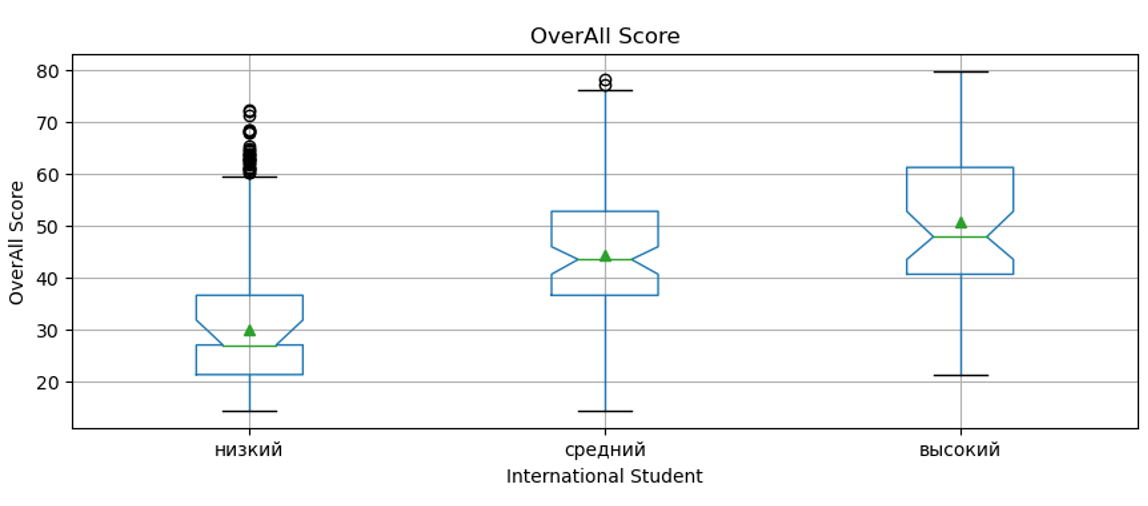
3. Академические приоритеты:Приоритеты университетов могут быть связаны с академическими исследованиями и образованием, а не с коммерческими или промышленными инициативами.

## 2.1.3. Гипотезы

1. С ростом дохода университета от индустрии растет его общая оценка.
2. Между кол-вом студентов на одного преподавателя и оценкой вуза существует сильная зависимость.

## 3. Анализ статистической связи.

2.2.1. Графический анализ пары “числовая зависимая переменная - качественная независимая”



*Рис. 7, гистограмма зависимости общей оценки от кол-ва иностранных студентов*

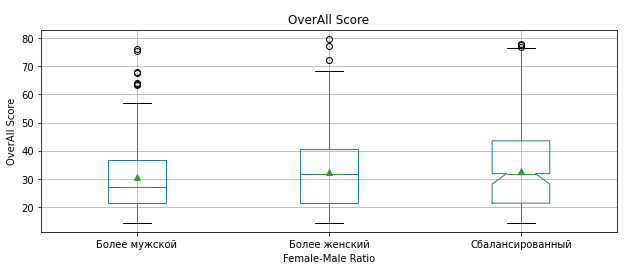
Из диаграммы Бокса-Уискера можно увидеть зависимость оценки университета от количества иностранных студентов. На это есть ряд причин:

1) привлечение международных студентов может свидетельствовать о мировом признании и престиже университета.

2) разнообразие в студенческом теле может способствовать обмену культурными и академическими идеями, что в конечном итоге повышает качество образования.

3) наличие иностранных студентов может улучшить международные партнерства и возможности для исследований и обмена опытом.

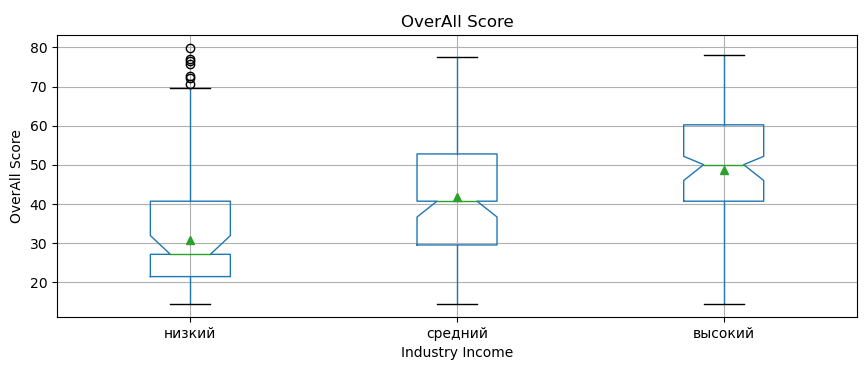
Это доказывает критерий Крускала-Уоллиса. P-значение 1.85e-30 практически равно нулю. Это говорит о том, что гипотеза об отсутствие связи отвергнута, что и говорит о том, что количество иностранных студентов влияет на оценку университета.



*Рис. 8, гистограмма зависимости общей оценки от соотношения женщин и мужчин.*

По диаграмме Бокса-Уискера можно увидеть, что отсутствует зависимость между оценкой университета и соотношения мужчин и женщин в университете. Так как в большинстве случаев, соотношение между мужчинами и женщинами не является непосредственным показателем качества учебного заведения. Рейтинг университета скорее зависит от таких факторов, как академическая репутация, качество преподавания, исследовательская деятельность, студенческие достижения. Таким образом, хотя соотношение мужчин и женщин может быть важным аспектом студенческой жизни, оно не всегда отражает общий рейтинг университета.

Это подтверждает критерий Крускала-Уоллиса, где P-значение составляет 0.221. Поскольку P-значение превышает общепринятый уровень значимости (обычно 0.05), то нет статистически значимых различий между оценкой университета и соотношения мужчин и женщин.



*Рис. 9, гистограмма зависимости общей оценки от доходов от индустрии.*

По диаграмме Бокса-Уискера видно, что существует зависимость между оценкой университета и уровня дохода университета от индустрий. На это есть ряд причин:

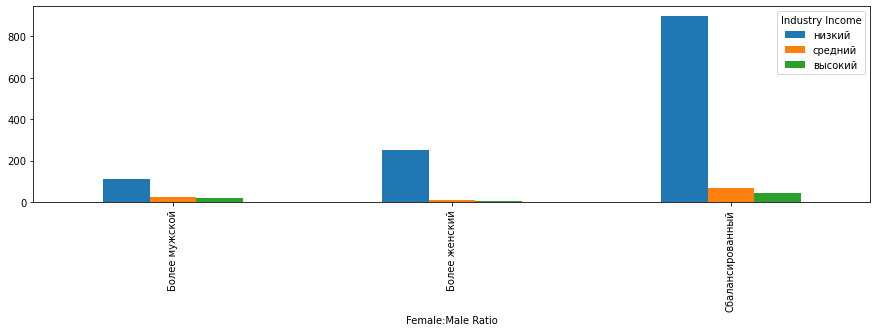
1) Большие доходы могут обеспечить университету больше ресурсов для инвестиций в качество обучения, исследовательскую деятельность, инфраструктуру и другие аспекты студенческой жизни. Эти ресурсы могут привлечь лучших преподавателей и исследователей, а также обеспечить студентам доступ к современным технологиям и учебным материалам.

2) Высокие доходы от индустрии могут способствовать университетским исследованиям и инновациям, что может привести к улучшению репутации университета в мировом сообществе науки и образования.

3) Университеты с высокими доходами от индустрии могут также иметь лучшие программы сотрудничества с бизнес-сектором, что способствует созданию возможностей для студентов для практического опыта и трудоустройства после окончания обучения.

Это подтверждает критерий Крускала-Уоллиса. P-значение близко к нулю (2.95e-25). Это также указывает на существенные статистически значимые различия. Следовательно, мы можем отвергнуть нулевую гипотезу об отсутствие связи. Другими словами, доходы университета от индустрий влияет на оценку университета.

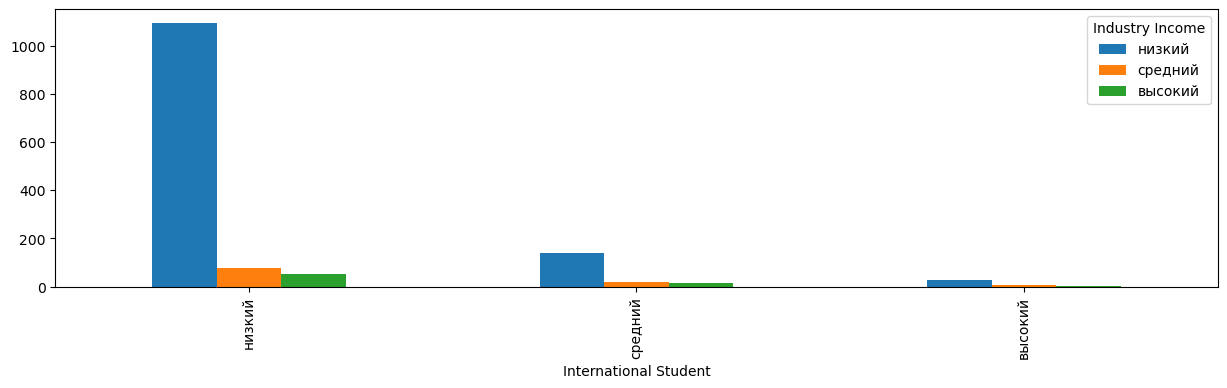
## 3.1. Графический анализ связи качественных переменных



*Рис. 10, гистограмма зависимости доходов от соотношения женщин и мужчин.*

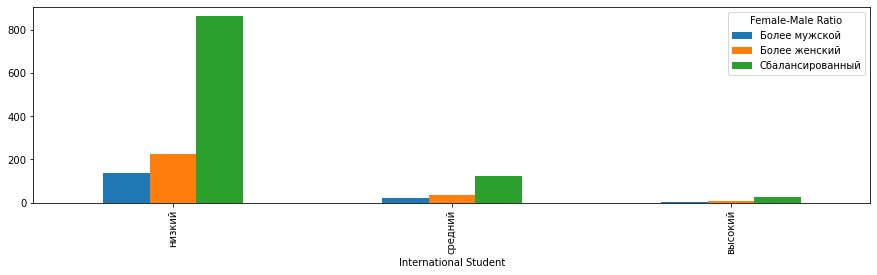
По данным, которые, мы получили в ходе вычислений, такие как: Коэффициент V Крамера и критерий Хи-квадрат. Коэффициент V Крамера составляет 0.1128, что указывает на слабую взаимосвязь между соотношением мужчин и женщин и доходами от индустрии. Это означает, что соотношение мужчин и женщин не сильно влияет на доходы от индустрии в университете.

По критерию Хи-квадрат значение p-значения близко к нулю (1.39e-08), что указывает на статистически значимую связь между соотношением мужчин и женщин и доходами от индустрии.



*Рис. 11, гистограмма зависимости доходов от кол-ва иностранных студентов.*

По данным, которые мы получили в ходе вычислений такие как: Коэффициент V Крамера и критерий Хи-квадрат. Коэффициент V Крамера равен 0.0695, что указывает на очень слабую взаимосвязь между количеством иностранных студентов и доходами от индустрии. Это означает, что количество иностранных студентов не сильно влияет на доходы от индустрии в университете. По критерию Хи-квадрат: значение p-значения равно 0.0138, что указывает на статистически значимую связь между статусом студента и доходами от индустрии.

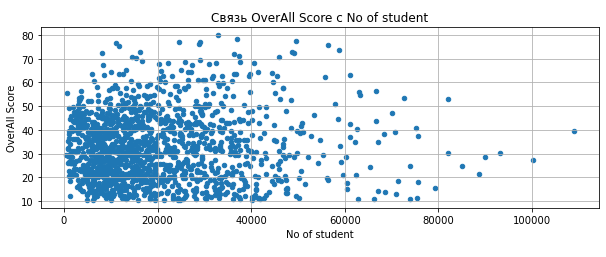


*Рис. 12, гистограмма зависимости соотношения женщин и мужчин от кол-ва иностранных студентов.*

По данным которые мы получили в ходе вычислений такие как: Коэффициент V Крамера и критерий Хи-квадрат. Значение коэффициента V Крамера равно 0.0127, что указывает на очень слабую взаимосвязь между количеством иностранных студентов и соотношением мужчин и женщин в университете. Это означает, что количество иностранных студентов не сильно влияет на распределение между мужчинами и женщинами в университете. По критерию Хи-квадрат значение p-значения составляет 0.9999, что указывает на отсутствие статистически значимой связи между количеством иностранных студентов и соотношением полов.

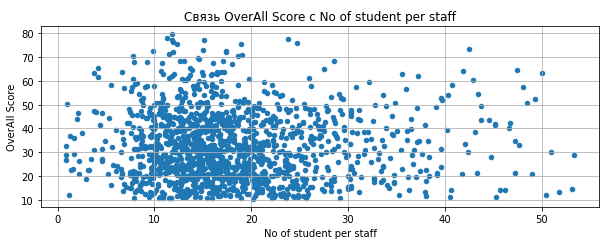
Исходя из этого, можно заключить, что хотя связь между переменными не является сильной, наблюдается статистически значимая связь между количеством иностранных студентов и доходами от индустрии, а также между соотношением мужчин и женщин и доходами от индустрии, в то время как, статистически значимая связь между количеством иностранных студентов и соотношением мужчин и женщин отсутствует.

## 3.2. Графический анализ пары “числовая зависимая переменная - числовая независимая переменная”



*Рис. 13, диаграмма зависимости общей оценки от кол-ва студентов.*

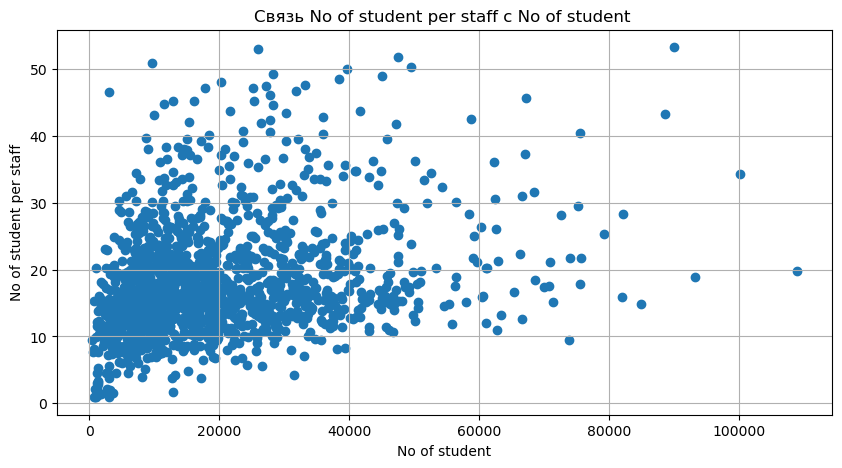
Исходя из графика, можно сказать, что оценка университета не сильно зависит от количества студентов в университете. Это подтверждает коэффициент корреляции Спирмена, который равен 0.031, что указывает на наличие очень слабой положительной корреляции между этими двумя переменными. Значимость оценки коэффициента корреляции Пирсона составляет приблизительно 0.2399, что также указывает на статистическую не значимость корреляции между этими переменными. То есть гипотезу об отсутствие связи принимаем.



*Рис. 13, диаграмма зависимости общей оценки от кол-ва студентов на одного преподавателя.*

Аналогично с прошлым графиком, можно заметить, что оценка университета практически не зависит от количества студентов на одного преподавателя. Это подтверждает коэффициент корреляции Спирмена, который равен -0.039, что указывает на очень слабую отрицательную корреляцию между этими двумя переменными. Также это подтверждает оценка значимости, которая равна 0.1344, что означает, что эта корреляция статистически незначима. То есть гипотезу об отсутствие связи принимаем.

## 3.3. Графический анализ связи двух независимых переменных



*Рис. 13, зависимость кол-ва студентов на одного преподавателя от общего кол-ва студентов.*

Исходя из графика, можно заметить, существует зависимость между количеством студентов и количеством студентов на преподавателя. Это подтверждает коэффициент корреляции Спирмена и оценка значимости коэффициента корреляции. Коэффициент корреляции Спирмена в размере 0.31 говорит о том, что существует положительная связь между количеством студентов на сотрудника и общим числом студентов. То есть, когда количество студентов на одного сотрудника растет, скорее всего, растет и общее количество студентов, и наоборот.

Оценка значимости коэффициента, близкая к нулю (4.66e-34), подтверждает, что эта корреляция является статистически значимой. Это означает, что вероятность такого результата случайно при отсутствии реальной связи между переменными крайне мала.

Таким образом, можно сделать вывод, что между оценкой университета и количеством студентов отсутствует зависимость, как и между оценкой университета и количеством студентов на одного преподавателя. Однако между количеством студентов и количеством студентов на одного преподавателя имеется зависимость.

## Предварительная проверка гипотез

1. С ростом дохода университета от индустрии растет его общая оценка.

P-значение близко к нулю (2.95e-25), следовательно доходы от индустрии существенно влияют на рейтинг университета, следовательно мы не можем отвергнуть поставленную гипотезу

1. Между кол-вом студентов на одного преподавателя и оценкой вуза существует сильная зависимость.

Коэффициент корреляции Спирмена, который равен -0.039, говорит об очень слабой отрицательной корреляцией между этими двумя переменными. Также это подтверждает оценка значимости, которая равна 0.1344, что означает, что эта корреляция статистически незначима.