В современном мире образование и навыки играют ключевую роль в достижении карьерного успеха. Однако до сих пор остается не до конца изученным, какие именно факторы (академическая успеваемость, навыки, стажировки и т.д.) наиболее значимо влияют на карьерные результаты, такие как уровень зарплаты, удовлетворенность карьерой и скорость продвижения по службе.

Исследование взаимосвязи между образованием, навыками и карьерными результатами актуально для студентов, образовательных учреждений и работодателей. Понимание этих взаимосвязей поможет студентам лучше планировать свою карьеру, университетам — улучшать учебные программы, а компаниям — эффективнее подбирать кандидатов.

**Исследовательский вопрос:**

Какие факторы (академическая успеваемость, навыки, стажировки и т.д.) наиболее значимо влияют на уровень стартовой зарплаты выпускников?

**Цель:**

Выявить ключевые факторы, влияющие на уровень стартовой зарплаты выпускников, и оценить их значимость.

Для анализа был взять датасет «Education & Career Success Analysis» с сайта Kaggle (<https://www.kaggle.com/datasets/adilshamim8/education-and-career-success/data>)

Основные гипотезы исследования:

* Гипотеза 1: Чем выше средний балл в университете (University\_GPA), тем выше стартовая зарплата выпускников.
* Гипотеза 2: Количество завершенных стажировок (Internships\_Completed) положительно влияет на уровень стартовой зарплаты.
* Гипотеза 3: Стартовая зарплата среди выпускников мужского и женского пола примерна равна.
* Гипотеза 4: Выпускники технических специальностей (например, Computer Science) имеют более высокую стартовую зарплату по сравнению с выпускниками гуманитарных специальностей.

1. **Описание данных**

Описание столбцов есть в источнике, но продублирую на русском.

1. **Информация о студенте**
   * **Student\_ID** – Уникальный идентификатор каждого студента.
   * **Age** – Возраст студента (от 18 до 30 лет).
   * **Gender** – Пол студента (Мужской, Женский или Другой).
2. **Академическая успеваемость**
   * **High\_School\_GPA** – Средний балл в школе (по шкале от 2.0 до 4.0).
   * **SAT\_Score** – Балл за стандартизированный тест (от 900 до 1600).
   * **University\_Ranking** – Рейтинг университета, в котором учится студент (от 1 до 1000).
   * **University\_GPA** – Средний балл в университете (по шкале от 2.0 до 4.0).
   * **Field\_of\_Study** – Специальность или направление обучения (например, Компьютерные науки, Медицина, Бизнес).
3. **Навыки и внеучебная деятельность**
   * **Internships\_Completed** – Количество завершенных стажировок (от 0 до 4).
   * **Projects\_Completed** – Количество выполненных личных/академических проектов (от 0 до 9).
   * **Certifications** – Количество полученных дополнительных сертификатов (от 0 до 5).
   * **Soft\_Skills\_Score** – Оценка мягких навыков (от 1 до 10).
   * **Networking\_Score** – Оценка профессиональных связей и нетворкинга (от 1 до 10).
4. **Карьерные результаты**
   * **Job\_Offers** – Количество предложений о работе после выпуска (от 0 до 5).
   * **Starting\_Salary** – Зарплата на первой работе в долларах США (от 25,000до25,000до150,000).
   * **Career\_Satisfaction** – Уровень удовлетворенности карьерой (от 1 до 10).
   * **Years\_to\_Promotion** – Время, необходимое для получения первого повышения (от 1 до 5 лет).
   * **Current\_Job\_Level** – Текущий уровень карьеры (Начальный, Средний, Старший, Руководящий).
   * **Work\_Life\_Balance** – Оценка баланса между работой и личной жизнью (от 1 до 10).
   * **Entrepreneurship** – Наличие предпринимательской деятельности (Да/Нет).

Вот основные характеристики этих данных:

| **index** | **Starting\_Salary** | **Age** | **SAT\_Score** | **University\_GPA** |
| --- | --- | --- | --- | --- |
| **count** | 5000.0 | 5000.0 | 5000.0 | 5000.0 |
| **mean** | 50563.54 | 23.4422 | 1253.832 | 3.0200280000000004 |
| **std** | 14494.958207289195 | 3.473711630878275 | 203.22895408271802 | 0.5760472943370758 |
| **min** | 25000.0 | 18.0 | 900.0 | 2.0 |
| **25%** | 40200.0 | 20.0 | 1076.0 | 2.52 |
| **50%** | 50300.0 | 23.0 | 1257.0 | 3.03 |
| **75%** | 60500.0 | 26.0 | 1432.0 | 3.51 |
| **max** | 101000.0 | 29.0 | 1600.0 | 4.0 |

Так как в таблице 20 ячеек, то я отобрал наиболее интересные для исследования и вывел основные описательные статистики, такие как кол-во значений, среднее, квантили, стандартное отклонение. В файле с кодом можно ознакомится с оставшимися значениями ячеек.

1. **Предварительная обработка данных**

Пропущенных значений в данных нет, поэтому для начала достаточно проверить на наличие выбросов в количественных переменных. Результат на рис. 1. Как видно данные без выбросов, кроме стартовой зарплаты, от того еще интересней узнать причины такой анамалии. Сразу удалим выбросы, т.к. они плохо сказываются на линейной регрессии. Удалилось 8 значений, не критично для выборки.

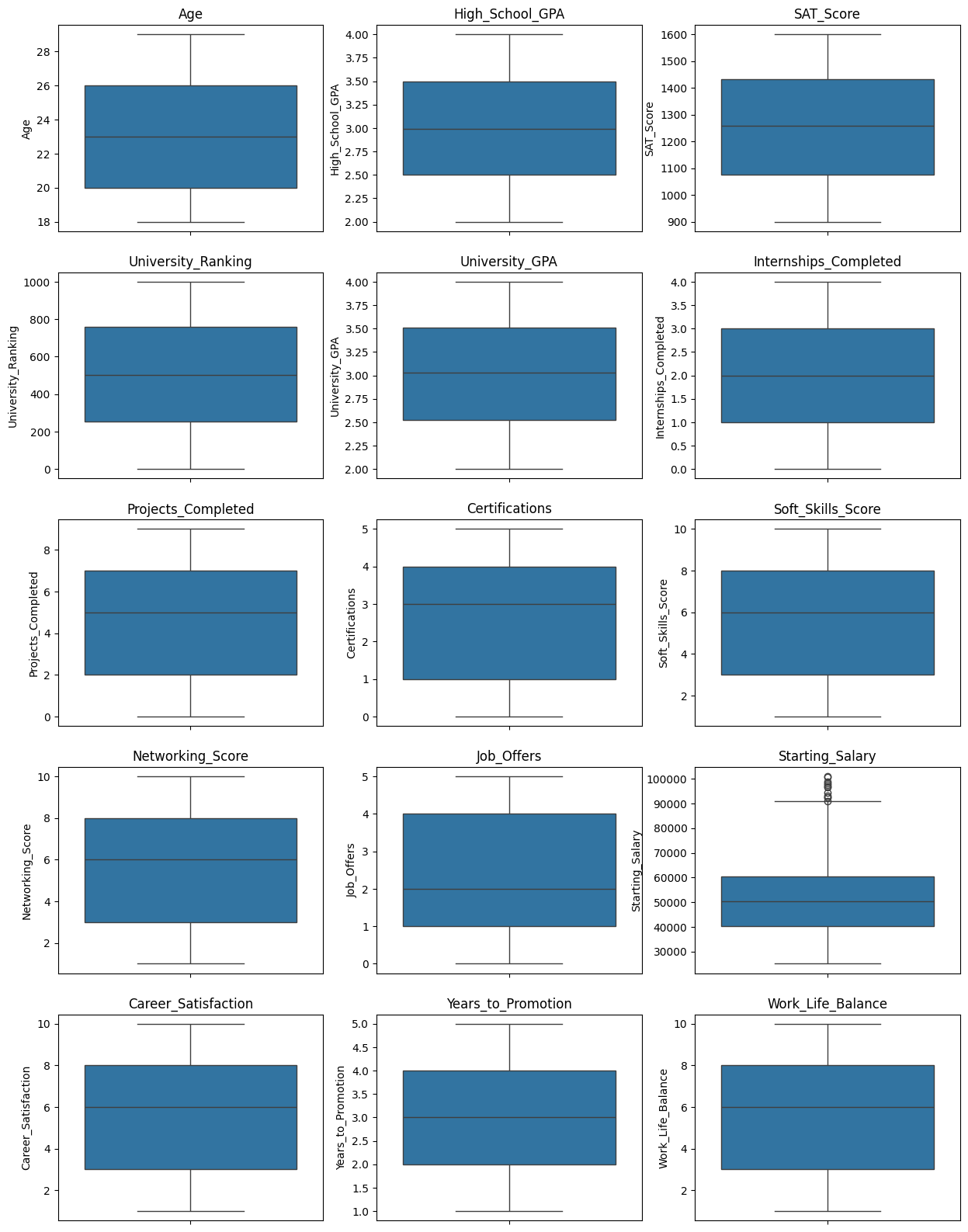


Рис.1 Выбросы

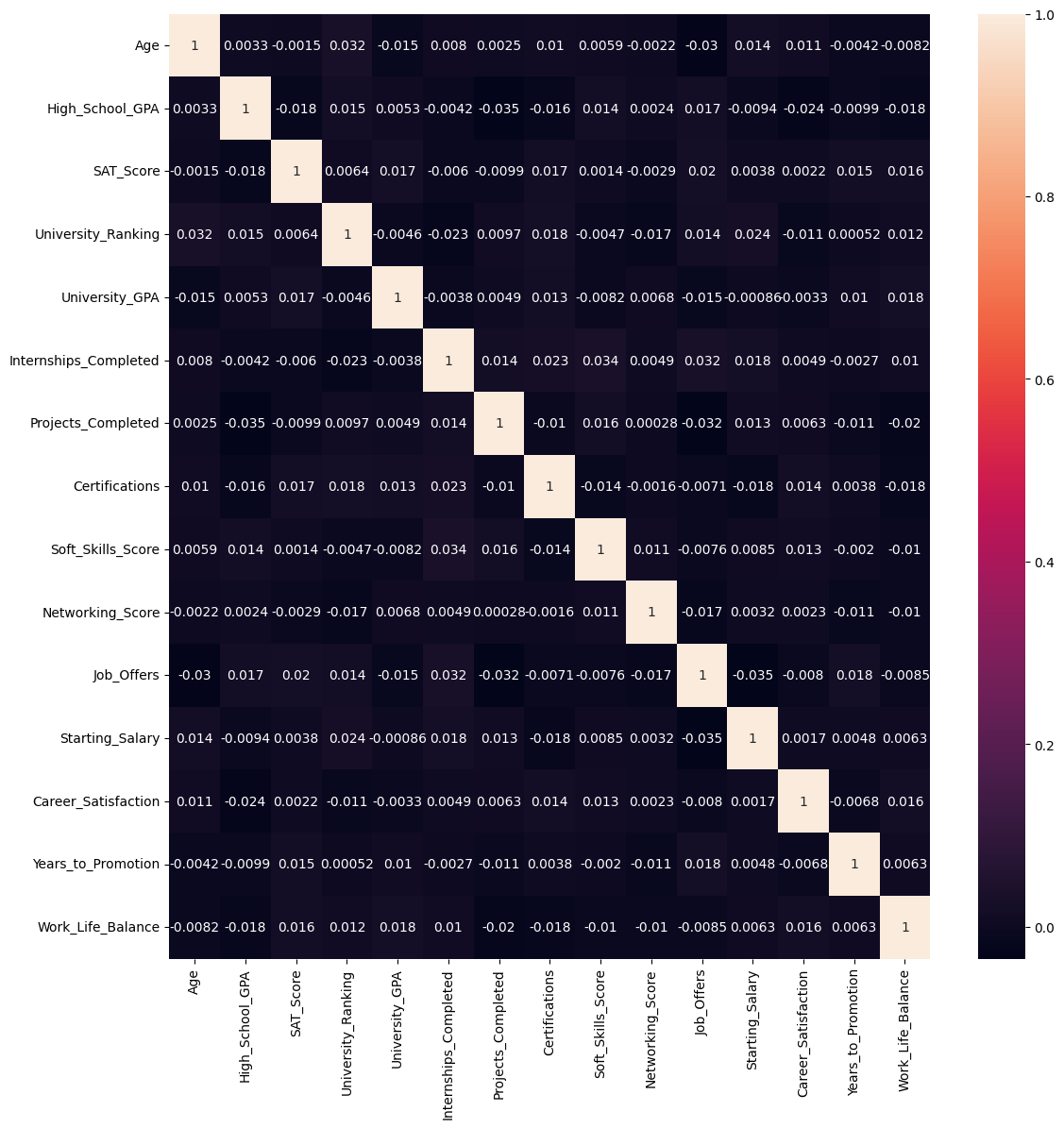
Теперь рассмотрим данные на наличие мультиколлиниарности. Построим корреляционную матрицу на рис.2. Видно, что корреляции почти нет, а значит дополнительной обработки не требуется.

Рис. 2. Корреляционная матрица

1. **EDA**

Теперь, когда данные готовы можно провести разведывательный анализ. Построим графики между теми переменными, которые наиболее интересны с точки зрении гипотез на рис.3. Глядя на данные графики предварительно можно сделать следующие выводы:

* студенты из более престижных университетов (меньшее значение University\_Ranking) имеют тенденцию к более высокой стартовой зарплате, но зависимость не линейная, так как наблюдается большая дисперсия.
* Диаграмма рассеяния показывает слабую взаимосвязь между GPA и стартовой зарплатой. Это говорит о том, что высокий средний балл не гарантирует высокую зарплату.
* чем больше стажировок у студента, тем выше стартовая зарплата. Это подтверждает, что опыт работы во время учебы положительно влияет на карьерные перспективы.
* Наблюдается равномерное распределение, что говорит о том, что высокий рейтинг университета не всегда гарантирует высокий средний балл у студентов.
* большинство выпускников получают стартовую зарплату в диапазоне **от 40 000 до 75 000**.

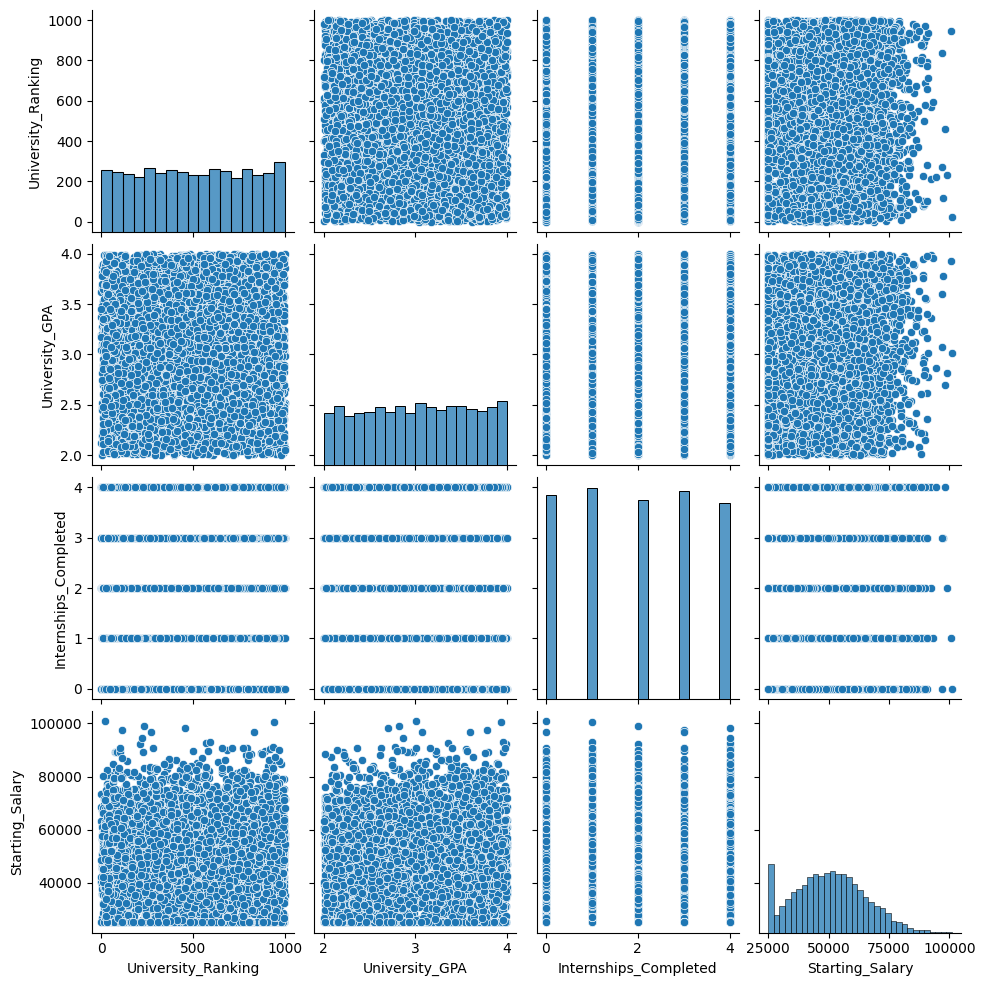
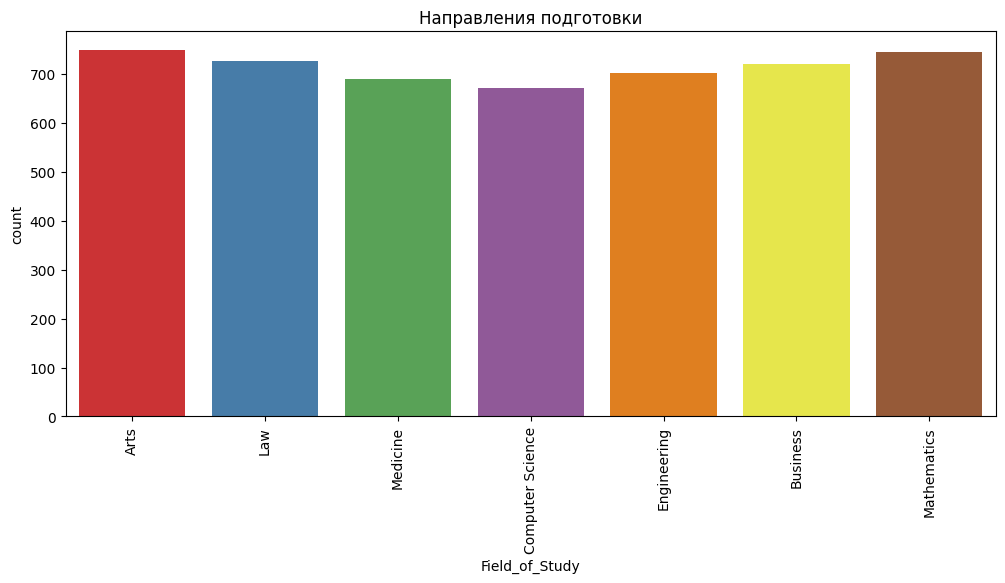


Рис 3.

Теперь посмотрим на распределение студентов по специальностям на рис.4. Видно, что количество студентов на разных направлениях примерно равно.

рис. 4.

1. **Тестирование гипотез**

Начнем с гипотезы 3: Стартовая зарплата среди выпускников мужского и женского пола примерна равна. Так как у нас две группы используем парный t-тест (если данные окажутся нормальными), где нулевая гипотеза (H₀): Средняя стартовая зарплата мужчин и женщин равна, а альтернативная гипотеза (H₁): Средняя стартовая зарплата мужчин и женщин не равна.

Отметим, что в данных имеется и третий гендер «Others». Чтобы не подвергать сомнению российские скрепы и законы, было принято решение забыть о существовании других гендеров помимо Male и Female.

Визуализировав данные по мужским и женским зарплатам получаем рис.5.

Проведем тест Левина на однородность дисперсий. P-value > 0.05, а значит дисперсии однородны, что позволяет проводить t-тест, однако можно увидеть, что данные, скорее всего не распределены нормально. Для проверки проведем Тест Шапиро-Уилка.

Тест Шапиро-Уилка для мужчин: pvalue=8.654144266116744e-15)

Тест Шапиро-Уилка для женщин: pvalue=5.7996027574578e-14)

Отсюда делаем вывод, что данные распределены ненормально, а значит будем использовать аналог t-теста - U-тест Манна-Уитни.

По результатам теста, p-value равно 0.8203215280258793, а значит при уровне значимости 0.05 мы не можем утверждать, что зарплаты отличаются, следовательно H1 отвергается – зарплаты равны.

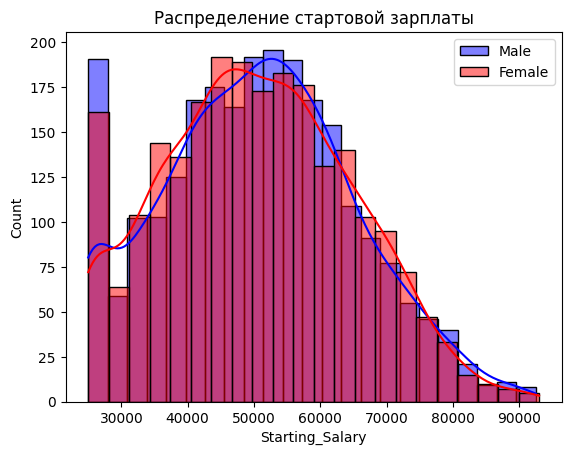
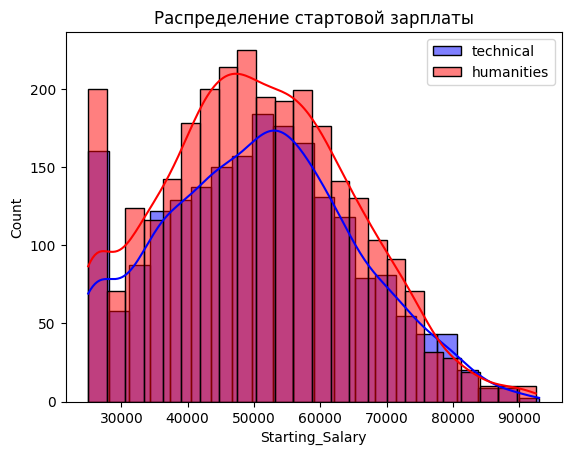


Рис.5 Распределение стартовой зарплаты

Теперь проверим гипотезу номер 4: Выпускники технических специальностей (например, Computer Science) имеют более высокую стартовую зарплату по сравнению с выпускниками гуманитарных специальностей.

Пусть гуманитарии это: medicine, law, arts, business. А технические: Computer science, mathematics, engineering

Для начала проанализируем данные (рис.6) так же как в прошлой гипотезе на нормальность и однородность дисперсий. Тест Левина показал что данные однородные (pvalue=0.55), а тест шапиро-Уилка, что данные распределены ненормально. Значит используем U-тест Манна-Уитни.

рис. 6

Несмотря на то, что визуально кажется, что различие между зарплатами имеется, мы получаем результат: p-value = 0.58. Следовательно у нас нет оснований, чтобы отвергнуть H0, а значит делаем вывод, что средние начальные зарплаты среди студентов гуманитариев и студентов технических специальностей равны.

1. **Линейная регрессия**

Чтобы проверить гипотезу 1 и 2 можно построить линейную регрессию. Для начала преобразуем категориальные данные в числовые с помощью One Hot Encoding.

Также чтобы в дальнейшем проверить нашу модель разделим ее на обучающую и тестовую выборки.

За целевую переменную возьмем Starting\_Salary и построив модель посмотрим как каждый предиктор в виде всех остальных столбцов (кроме ID) влияет на зарплату. Также зарание нормализируем данные, проверим на гетероскедастичность с помощью теста Бреуша-Пагана и автокорреляцию с помощью теста Дарбина-Уотсона.

Проведя тесты получаем, что p-value для теста Бреуша-Пагана составляет 0.77, а значения теста Дарбина-Уотсона 1.99, что указывает на отсутствие гетероскедастичности и автокорреляции. Следовательно можем продолжить работать с данными.

Значение R2 на тестовой выборке составляет **-0.0114**, что указывает на крайне низкую объясняющую способность модели. отрицательное значение R2говорит о том, что модель на тестовой выборке работает хуже, чем простое предсказание среднего значения стартовой зарплаты для всех наблюдений. Это свидетельствует о том, что построенная линейная регрессия не способна адекватно предсказывать зарплату на основе выбранных признаков, что может быть следствием переобучения, недостатка значимых факторов или нелинейных зависимостей в данных.

Intercept (50674.75) представляет собой базовый уровень стартовой зарплаты, который модель предсказывает в случае, если все независимые переменные равны нулю.

Однако коэффициенты модели остаются актуальными в том смысле, что они отражают направления связей между предикторами и целевой переменной в рамках данной линейной модели, поэтому имеет смысл, хоть и с осторожностью, интерпретировать результаты и проверить гипотезы.

Наибольшее положительное влияние оказывает возраст (Age, β = 364.49), что свидетельствует о тенденции роста профессионального успеха с увеличением стажа и накопленного опыта. Существенный вклад вносят также количество завершённых проектов (Projects\_Completed, β = 312.08) и уровень удовлетворённости карьерой (Career\_Satisfaction, β = 288.32), что подчёркивает важность профессиональной вовлечённости и личного восприятия работы. Развитие мягких навыков (Soft\_Skills\_Score, β = 263.25) также положительно коррелирует с целевой переменной, подтверждая гипотезу о том, что коммуникативные и управленческие компетенции способствуют профессиональному росту.

Значительный положительный эффект оказывает стартовая зарплата (Starting\_Salary, β = 217.16), что указывает на её влияние на последующее карьерное развитие. Важную роль играют также количество пройденных стажировок (Internships\_Completed, β = 186.46) и уровень текущей должности (Current\_Job\_Level\_Senior, β = 161.70), что подтверждает значимость начального карьерного пути и накопленного опыта для долгосрочного успеха. Положительное влияние оказывает и уровень нетворкинга (Networking\_Score, β = 147.21), что указывает на важность профессиональных связей.

Ряд факторов, связанных с образованием и карьерным балансом, оказывают умеренное положительное влияние: направление Field\_of\_Study\_Medicine (β = 130.80), количество лет до повышения (Years\_to\_Promotion, β = 89.45), баланс между работой и личной жизнью (Work\_Life\_Balance, β = 87.95), а также уровень текущей должности (Current\_Job\_Level\_Mid, β = 73.59). Роль академических достижений представлена университетским рейтингом (University\_Ranking, β = 71.53) и успеваемостью в университете (University\_GPA, β = 60.43), однако их влияние сравнительно слабее, что свидетельствует о более важной роли профессионального опыта.

В то же время наблюдается ряд факторов, оказывающих отрицательное влияние. Наиболее выраженный негативный эффект наблюдается у направлений обучения: Field\_of\_Study\_Engineering (β = -653.92), Field\_of\_Study\_Law (β = -438.77), Field\_of\_Study\_Computer Science (β = -434.05) и Field\_of\_Study\_Mathematics (β = -286.66). Это может свидетельствовать о высокой конкуренции и насыщенности рынка в этих областях, а также о специфике карьерных траекторий.

Другие значимые отрицательные предикторы включают пол (Gender\_Male, β = -485.26), что может указывать на гендерные различия в карьерном развитии, и количество предложений о работе (Job\_Offers, β = -356.93), что, возможно, связано с компромиссами между количеством и качеством предложений. Негативное влияние оказывают также уровень текущей должности на высшем уровне (Current\_Job\_Level\_Executive, β = -177.44), наличие сертификатов (Certifications, β = -205.92) и результаты SAT (SAT\_Score, β = -134.06), что может отражать снижение значимости формальных образовательных достижений по сравнению с практическим опытом.

Таким образом, ключевые положительные факторы связаны с возрастом, практическим опытом, удовлетворённостью карьерой и развитием мягких навыков, тогда как отрицательное влияние чаще всего связано с формальным образованием и рыночными особенностями конкретных профессиональных областей.

Вернемся к гипотезе 1: Результаты регрессии показывают положительный, но слабый коэффициент для переменной University\_GPA (β = 60.43). Это означает, что средний балл в университете действительно имеет некоторое влияние на стартовую зарплату, однако его эффект невелик по сравнению с другими факторами, такими как количество завершённых проектов (Projects\_Completed) или количество предложений о работе (Job\_Offers). Следовательно, хотя гипотеза в целом подтверждается, её практическая значимость относительно невысока.

Говоря о гипотезе номер 2, можно сказать что переменная Internships\_Completed имеет положительный коэффициент (β = 186.46), что указывает на значимое влияние количества стажировок на уровень стартовой зарплаты. Это подтверждает гипотезу о том, что практический опыт, полученный во время обучения, играет важную роль при выходе на рынок труда. Данный фактор оказывается более значимым, чем академические показатели, что подчёркивает важность раннего приобретения профессионального опыта для успешного начала карьеры.

Таким образом, обе гипотезы подтверждаются, но с разной степенью влияния: средний балл в университете оказывает слабый эффект, тогда как количество завершённых стажировок имеет более выраженное положительное влияние на стартовую зарплату выпускников.

В математическом виде нормированная модель выглядит следующим образом:

Starting\_Salary = 364.49 \* Age + 312.08 \* Projects\_Completed + 288.32 \* Career\_Satisfaction + 263.25 \* Soft\_Skills\_Score + 217.16 \* Starting\_Salary + 186.46 \* Internships\_Completed + 161.70 \* Current\_Job\_Level\_Senior + 147.21 \* Networking\_Score + 130.80 \* Field\_of\_Study\_Medicine + 89.45 \* Years\_to\_Promotion + 87.95 \* Work\_Life\_Balance + 73.59 \* Current\_Job\_Level\_Mid + 71.53 \* University\_Ranking + 60.43 \* University\_GPA - 36.58 \* High\_School\_GPA - 111.73 \* Field\_of\_Study\_Business - 134.06 \* SAT\_Score - 177.44 \* Current\_Job\_Level\_Executive - 205.92 \* Certifications - 286.66 \* Field\_of\_Study\_Mathematics - 356.93 \* Job\_Offers - 434.05 \* Field\_of\_Study\_Computer\_Science - 438.77 \* Field\_of\_Study\_Law - 485.26 \* Gender\_Male - 653.92 \* Field\_of\_Study\_Engineering

В ненормированном:

Starting\_Salary = 639.54 \* Current\_Job\_Level\_Executive + 401.48 \* Entrepreneurship\_Yes + 203.48 \* Current\_Job\_Level\_Senior + 203.04 \* Years\_to\_Promotion + 175.94 \* Gender\_Male + 132.35 \* Internships\_Completed + 108.56 \* Projects\_Completed + 105.24 \* University\_GPA + 104.76 \* Age + 92.28 \* Soft\_Skills\_Score + 75.75 \* Career\_Satisfaction + 51.62 \* Networking\_Score + 31.02 \* Work\_Life\_Balance + 0.25 \* University\_Ranking - 0.66 \* SAT\_Score - 63.57 \* High\_School\_GPA - 120.85 \* Certifications - 208.72 \* Job\_Offers - 329.38 \* Field\_of\_Study\_Computer\_Science - 384.39 \* Current\_Job\_Level\_Mid - 830.38 \* Field\_of\_Study\_Medicine - 1234.57 \* Field\_of\_Study\_Mathematics - 1250.24 \* Field\_of\_Study\_Engineering - 1379.94 \* Field\_of\_Study\_Business - 1857.36 \* Field\_of\_Study\_Law

1. **Выводы**

В ходе исследования была проведена проверка ключевых гипотез о влиянии академической успеваемости, практического опыта и других факторов на уровень стартовой зарплаты выпускников. Основной инструмент анализа — линейная регрессия — показал низкую предсказательную способность модели (R² < 0), что свидетельствует о слабой объясняющей силе включённых в модель факторов. Проверка данных на соответствие требованиям регрессионного анализа выявила отсутствие гетероскедастичности (тест Бреуша-Пагана) и автокорреляции (тест Дарбина-Уотсона), однако негативное значение R² указывает на то, что построенная модель на тестовой выборке работает хуже, чем простое усреднение.

Анализ влияния отдельных переменных выявил, что академические достижения, в частности средний балл в университете, действительно оказывают влияние на стартовую зарплату, но оно незначительно. Гораздо более выраженным предиктором оказался практический опыт: количество завершённых стажировок положительно коррелировало с уровнем стартовой зарплаты, что подтверждает гипотезу о важности раннего профессионального опыта. Пол выпускников не показал статистически значимого различия в стартовой зарплате, что подтвердило гипотезу о её равенстве между мужчинами и женщинами. Направление обучения также не дало значимого различия между гуманитарными и техническими специальностями, что противоречит распространённому мнению о преимуществах технического образования с точки зрения доходов.

Кроме того, анализ показал, что ряд факторов, традиционно считающихся важными, имеют слабое или даже отрицательное влияние на уровень зарплаты. Например, количество полученных сертификатов и результаты стандартизированных тестов (SAT) не продемонстрировали высокой значимости, что может свидетельствовать о снижении роли формальных образовательных достижений на рынке труда. Напротив, развитие мягких навыков и профессиональные связи (нетворкинг) показали положительное влияние, что подчёркивает важность личных и профессиональных компетенций.

Основной недостаток исследования заключается в использовании линейной регрессии, которая плохо справляется с моделированием сложных зависимостей. Улучшение качества анализа возможно за счёт применения более сложных моделей, таких как градиентный бустинг или нейронные сети, которые могут лучше выявлять нелинейные взаимосвязи между факторами. Также следует рассмотреть возможность включения дополнительных предикторов, таких как социально-экономическое происхождение студентов, их активность в профессиональных сообществах, а также тип и репутация работодателя.

В целом, исследование подтвердило, что академические достижения сами по себе не являются определяющим фактором успеха, а практический опыт, профессиональные связи и развитие личных качеств оказывают более значимое влияние на карьерные результаты. Это подчёркивает необходимость адаптации образовательных программ в сторону более тесного взаимодействия с рынком труда, а также повышения внимания студентов к профессиональному развитию ещё в процессе обучения.