МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ
ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А. И. ГЕРЦЕНА»

Институт информационных технологий и технологического образования
кафедра информационных технологий и электронного обучения

Основная профессиональная образовательная программа
Направление подготовки 09.03.01 Информатика и вычислительная техника
Направленность (профиль) «Технологии разработки программного обеспечения»
форма обучения – очная

Курсовая работа

по дисциплине «Пакетам прикладных программ для статистической обработки и анализа данных»

Анализ опросных данных:
Воздействие контента на психическое здоровье

Обучающегося 3 курса
Зухир Амиры Саидовны

_____

Руководитель:
д.п.н, профессор
_____ Власова Е. З.

«_____» _____ 2024 г.

Санкт-Петербург
2024

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ....3
ТЕОРЕТИЧЕСКАЯ ЧАСТЬ....4
1.1 Корреляционный анализ:....4
1.1.1 Коэффициент корреляции Спирмена:....5
1.1.2 Коэффициент V Крамера:....6
1.2 Тест Хи-квадрат:....7
ПРАКТИЧЕСКАЯ ЧАСТЬ....9
ЗАКЛЮЧЕНИЕ....18
ЛИТЕРАТУРА....20
ПРИЛОЖЕНИЕ....21

ВВЕДЕНИЕ

Современное общество переживает период значительных изменений, вызванных активным развитием информационных технологий и интернета. С появлением новых социальных платформ и возросшим объемом потребляемого контента становится всё более актуальным вопрос о воздействии этого контента на психологическое здоровье человека. Несмотря на множество положительных аспектов цифровой эры, существует опасность, что избыточное воздействие разнообразных медийных форматов может оказывать негативное воздействие на эмоциональное, психическое и физическое состояние человека.

Анализ опросных данных представляет собой важный метод исследования, который позволяет получить количественные и качественные данные о восприятии контента, а также выявить возможные связи между типами контента, различными социальными сетями и психическим здоровьем человека.

Данная курсовая работа посвящена анализу опросных данных, направленных на выявление воздействия контента на психическое здоровье. Будет проведен анализ взаимосвязей между детоксом от социальных сетей, временем, проведенным в этих сетях, и психологическими аспектами, такими как потребность в детоксе и уровень стресса. С целью раскрытия этих взаимосвязей будут использованы различные статистические методы, включая коэффициенты V Крамера и Спирмена, а также тест хи-квадрат. Полученные результаты предоставляют информацию о степени связи между рассматриваемыми переменными, а также об их влиянии на психологическое состояние и потребности в детоксе. Ниже представлены основные выводы, которые могут оказать влияние на понимание влияния социальных сетей на психологическое благополучие и стрессовые уровни современного общества.

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Корреляционный анализ:

Корреляционный анализ — это метод по изучению статистической зависимости между случайными величинами с необязательным наличием строгого функционального характера, при которой динамика одной случайной величины приводит к динамике математического ожидания другой.

Ложная корреляция — это явление, при котором может наблюдаться статистическая связь между двумя переменными, не имеющими объективной причинной связи друг с другом. При проведении корреляционного анализа необходимо осознавать, что статистическая взаимосвязь может возникнуть случайно или из-за наличия третьего фактора, не имеющего непосредственного отношения к рассматриваемым переменным. Такие корреляции могут ввести исследователя в заблуждение, если не учитывать контекст и особенности данных. Однако, важно помнить, что статистическая корреляция не всегда указывает на наличие причинно-следственной связи между переменными.

Основные понятия в корреляционном анализе:

1) Корреляция: это степень взаимосвязи между двумя переменными. Корреляция может быть положительной, отрицательной или равной нулю. Положительная корреляция указывает на то, что увеличение одной переменной сопровождается увеличением другой. Отрицательная корреляция означает, что увеличение одной переменной связано с уменьшением другой. Корреляция равная нулю говорит о том, что между переменными нет статистической связи.
2) Коэффициент корреляции: это числовая мера, которая выражает степень корреляции между переменными. Коэффициент корреляции принимает значения от -1 до 1. Значение -1 указывает

на полную отрицательную корреляцию, 1 - на положительную, а 0 - на отсутствие корреляции.

3) Корреляционная матрица: это таблица, в которой показаны коэффициенты корреляции между всеми парами переменных в наборе данных.

Корреляционный анализ широко применяется в различных областях, таких как экономика, психология, медицина и другие, для выявления связей между переменными и более глубокого понимания структуры данных.

1.1.1 Коэффициент корреляции Спирмена:

Коэффициент корреляции Спирмена - это статистическая мера, используемая для оценки силы и направления монотонной (не обязательно линейной) связи между двумя ранжированными переменными. Этот метод основан на рангах значений переменных вместо их фактических числовых значений.

Процесс вычисления коэффициента корреляции Спирмена включает в себя следующие шаги:

1. Ранжирование значений каждой переменной по возрастанию. Присвоение каждому значению его ранга. Ранг представляет собой порядковый номер. Далее, вычисление разностей рангов для соответствующих пар значений переменных.
2. Квадратирование этих разностей и суммирование квадратов разностей рангов.
3. Применение формулы (1) для вычисления коэффициента корреляции Спирмена.

$$r_s = 1 - 6\frac{\sum d^2}{n^3 - n}, \quad (1)$$

где: d – разность рангов, n – количество признаков, участвовавших в ранжировании.

Коэффициент корреляции Спирмена принимает значения от -1 до 1. Значение -1 указывает на полную обратную монотонную связь, 1 - на положительную монотонную связь, а 0 - на отсутствие монотонной связи. Данный коэффициент часто используется в случаях, когда данные не соответствуют предположению о нормальном распределении или когда связь между переменными не является линейной.

1.1.2 Коэффициент V Крамера:

Коэффициент V Крамера является мерой силы и направления ассоциации между двумя категориальными переменными, измеренной в кросс-таблице. Кросс-таблица представляет собой двумерную таблицу, где каждая ячейка показывает количество наблюдений, попавших в конкретную комбинацию категорий обеих переменных.

Процесс вычисления коэффициента V Крамера начинается с применения критерия хи-квадрат для оценки статистической значимости связи между двумя переменными. Критерий хи-квадрат проверяет гипотезу о том, что две категориальные переменные независимы.

Формула для вычисления коэффициента V Крамера (2).

$$V = \sqrt{\frac{x^2}{n \cdot \min(k-1, r-1)}}, \quad (2)$$

где: x^2– значение статистики хи-квадрат, полученное при анализе кросс-таблицы,

n – общее количество наблюдений,

k – количество категорий (уровней) в одной переменной,

r – количество категорий (уровней) в другой переменной.

Коэффициент V Крамера лежит в диапазоне от 0 до 1. Значение 0 означает отсутствие ассоциации между переменными, а значение 1 указывает на полную ассоциацию. Чем ближе значение V Крамера к 1, тем сильнее связь между переменными.

Этот коэффициент часто используется в исследованиях социальных наук, маркетинге, медицинских исследованиях и других областях, где необходимо оценить степень ассоциации между двумя категориальными переменными.

1.2 Тест Хи-квадрат:

Тест хи-квадрат — статистический метод, используемый для оценки статистической значимости ассоциаций между категориальными переменными в виде таблиц сопряженности (кросс-таблиц). Он позволяет проверить гипотезу о том, что наблюдаемое распределение частот в кросс-таблице не отличается от ожидаемого распределения при предположении независимости переменных.

Процесс теста хи-квадрат включает в себя следующие шаги:

1. Постановка гипотез:

Нулевая гипотеза (H0): нет статистически значимой связи между переменными (независимость).

Альтернативная гипотеза (H1): существует статистически значимая связь между переменными (зависимость).

2. Построение кросс-таблицы: данные разбиваются на категории, создавая кросс-таблицу, где значения ячеек представляют собой частоты наблюдений для каждой комбинации категорий.

3. Расчет ожидаемых частот: ожидаемые частоты вычисляются на основе предположения независимости переменных. Это делается путем умножения сумм частот по строкам и столбцам на соответствующие вероятности.

4. Вычисление статистики хи-квадрат: вычисляется статистика хи-квадрат, которая представляет собой сумму квадратов отклонений между наблюдаемыми и ожидаемыми частотами, нормированных на ожидаемые частоты.

5. Определение степеней свободы: определяются степени свободы для теста, которые зависят от размеров кросс-таблицы.

6. Определение критического значения и принятие решения: сравнивается значение статистики хи-квадрат с критическим значением из таблицы распределения хи-квадрат. Если статистика хи-квадрат превышает критическое значение, то нулевая гипотеза отклоняется в пользу альтернативной.

Тест хи-квадрат часто используется в медицинских исследованиях, социологии, маркетинге и других областях, где требуется оценка влияния одной переменной на другую в виде категорий.

ПРАКТИЧЕСКАЯ ЧАСТЬ

Для проведения исследования, был создан специальный опрос и распространён среди 50 человек. Были заданы следующие вопросы:

1- Какое среднее количество часов в день вы обычно проводите в социальных сетях?
2- В каких социальных сетях Вы сидите?
3- Делаете ли Вы "детокс" от социальных сетей? Хотя бы на несколько часов
4- Какой контент Вы чаще всего потребляете?
5- Каково ваше самочувствие после проведенного времени в социальных сетях?
6- Хотели бы Вы проводить меньше времени в соц.сетях ?
7- Бывали ли случаи, когда вы испытывали стресс из-за событий в социальных сетях?

С помощью языка программирования Python и IDE VScode были построены гистограммы и круговые диаграммы результатов ответов представленные на рисунках 1-7.

Рисунок 1.Круговая диаграмма среднего кол-ва часов проведенных в соц.сетях

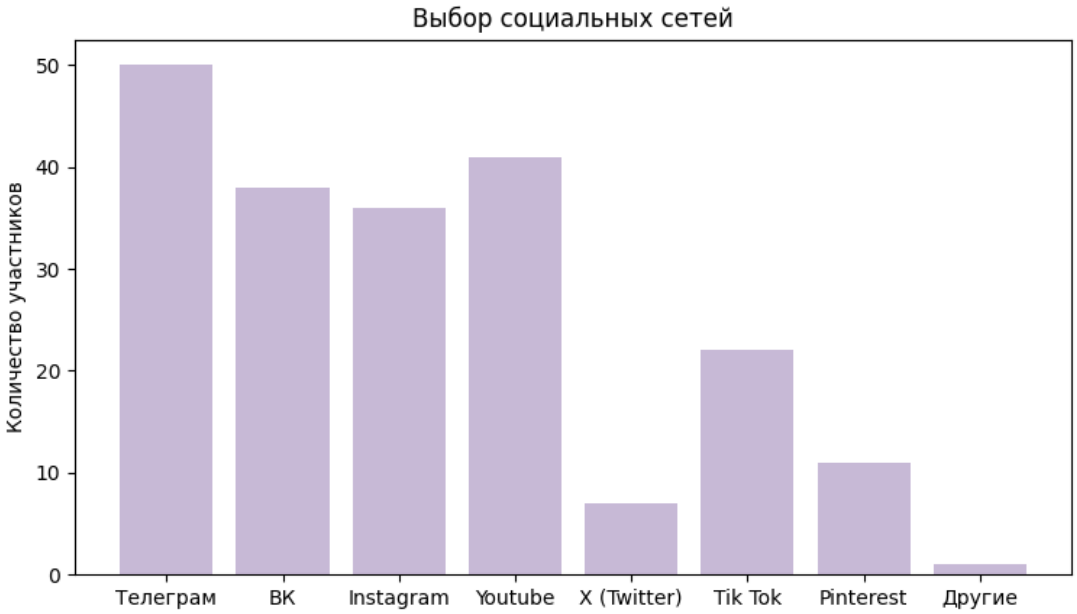

Рисунок 2. Выбор соц.сетей участниками опроса.

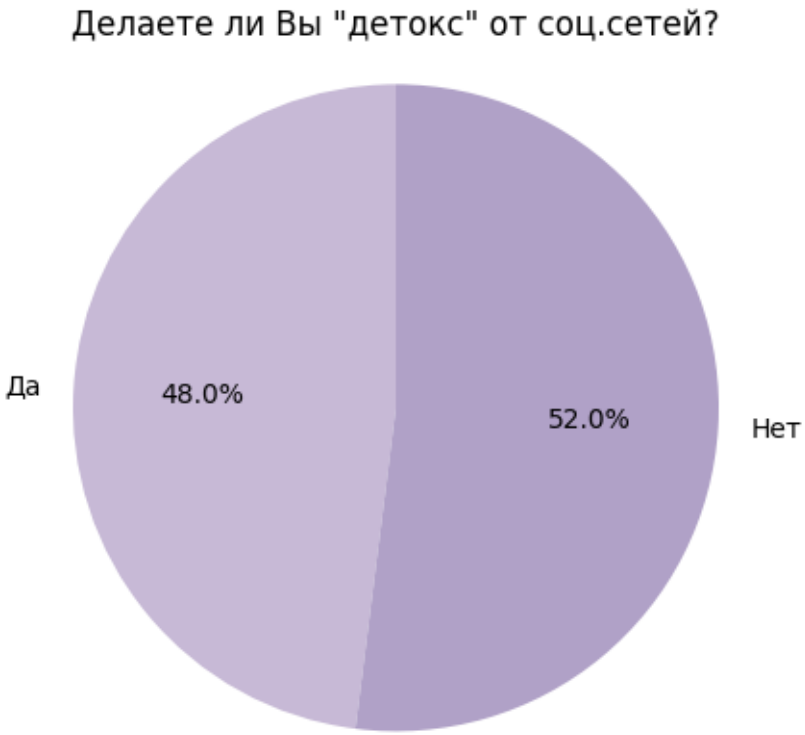

Рисунок 3. Процент опрошенных, выполняющих диджитал детокс (да/нет)

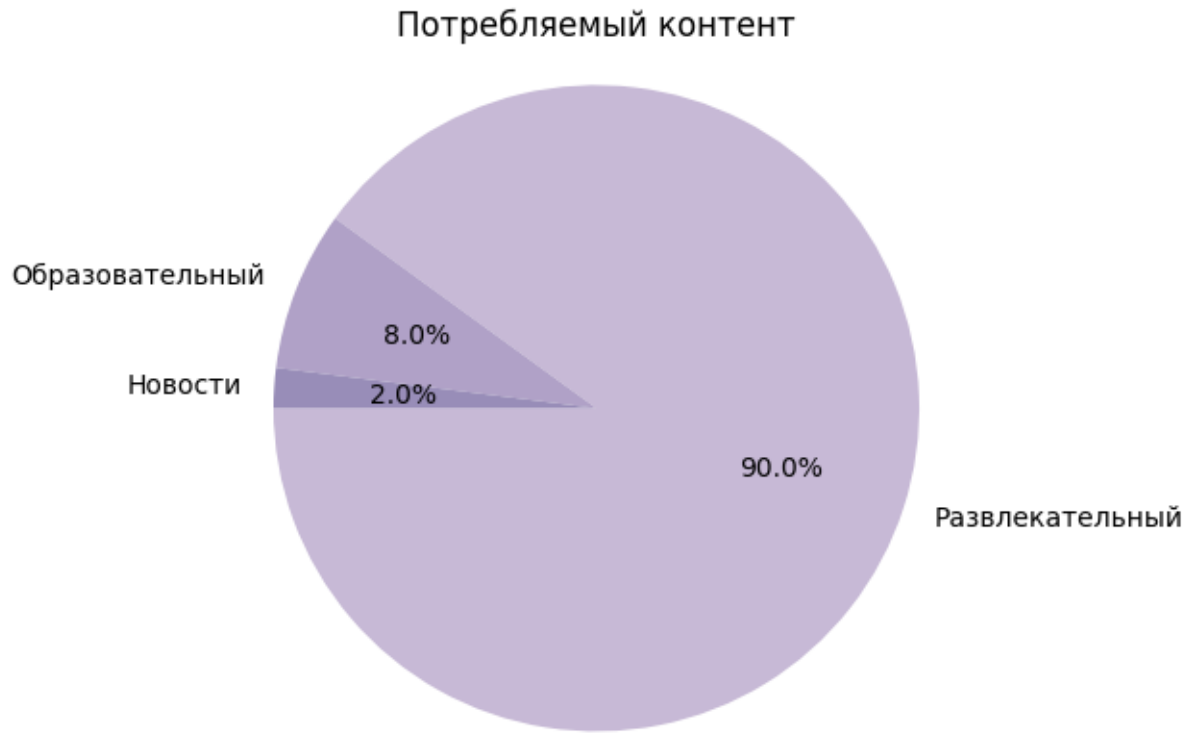

Рисунок 4. Выбор потребляемого контента.

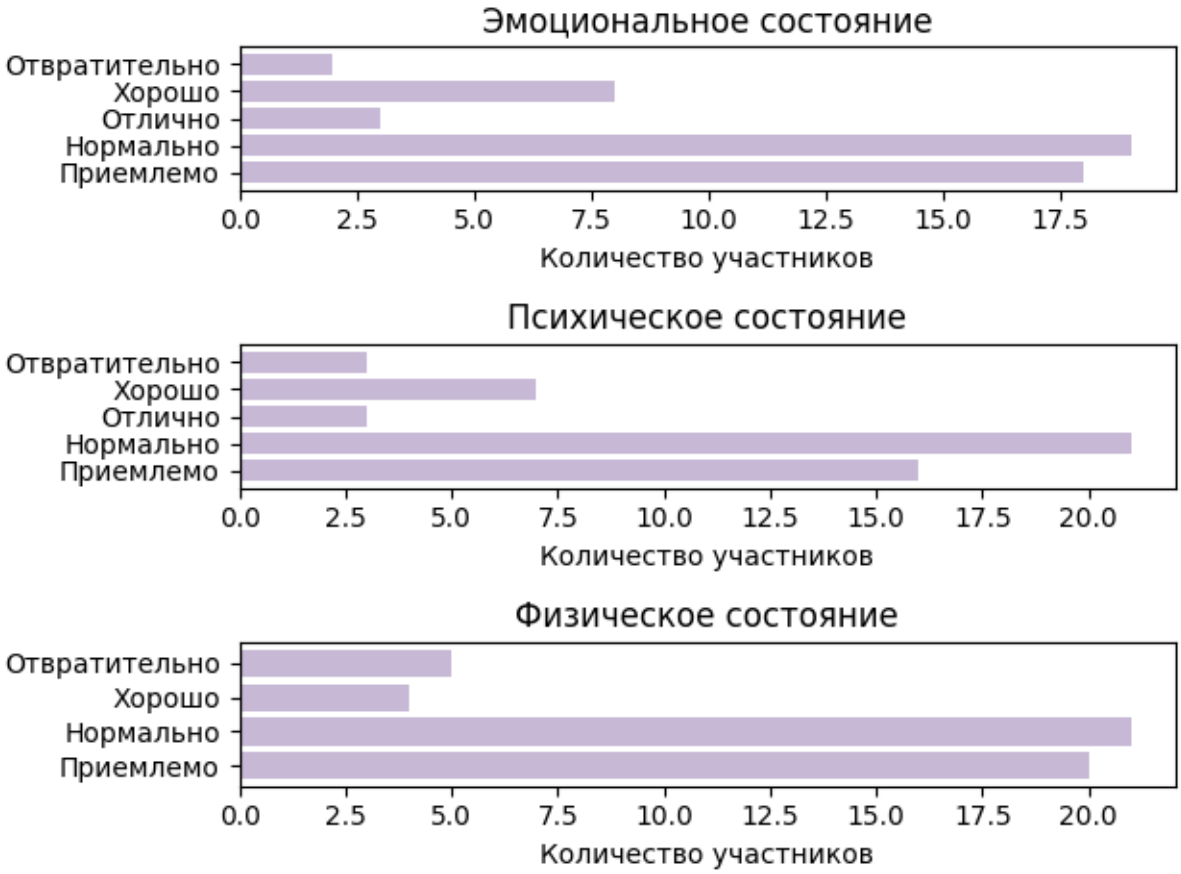

Рисунок 5. Эмоциональное, психическое и физическое состояния после времени, проведенного в социальных сетях.

Рисунок 6. Желание сократить время в социальных сетях.

Рисунок 7. Стресс из-за событий в социальных сетях.

Исследуем, есть ли статистически значимая связь между желанием уменьшить время в социальных сетях и опытом стресса из-за событий в социальных сетях.

Для анализа связи между желанием уменьшить время в социальных сетях и опытом стресса из-за событий в социальных сетях, воспользуемся тестом Хи-квадрат.

```
def chi_analyses(data):
    # Подготовка данных для теста Хи-квадрат
    reduce_time_stress_data = {'Да_Да': 0, 'Да_Нет': 0, 'Нет_Да': 0, 'Нет_Нет': 0}

    for participant_data in data:
        reduce_time = next(answer for question, answer in participant_data
                           if question == 'Хотели бы Вы проводить меньше времени в соц.сетях ?')
        stress = next(answer for question, answer in participant_data
                      if question == 'Бывали ли случаи, когда вы испытывали стресс из-за событий в социальных сетях?')

        key = f"{reduce_time}_{stress}"
        reduce_time_stress_data[key] += 1

    # Проведение теста Хи-квадрат
    observed_values = [
        [reduce_time_stress_data['Да_Да'], reduce_time_stress_data['Да_Нет']],
        [reduce_time_stress_data['Нет_Да'], reduce_time_stress_data['Нет_Нет']]
    ]

    chi2, p, _, _ = chi2_contingency(observed_values)

    # Вывод результатов теста
    print(f"Хи-квадрат: {chi2}")
    print(f"P-value: {p}")

    # Интерпретация результатов
    if p < 0.05:
        print("Есть статистически значимая связь.")
    else:
        print("Статистической значимости связи не выявлено.")
```

Рисунок 8. Код выполнения теста Хи-квадрат.

На рисунке 8 представлен код, в нем используем библиотеку scipy, которая предназначена для выполнения научных и инженерных вычислений.

В этом коде chi2_contingency используется для выполнения теста Хи-квадрат на категориальных данных. Результаты теста включают значение Хи-квадрат и p-значение. Если p-значение меньше выбранного уровня значимости (обычно 0.05), то можно отклонить нулевую гипотезу и считать, что существует статистически значимая связь.

```
Хи-квадрат: 0.0
P-value: 1.0
Статистической значимости связи не выявлено.
```

Рисунок 9. результат выполнения функции

На рисунке 9 виден результат выполнения кода для вычисления теста хи-квадрата.

Значение Хи-квадрата равно 0.0, что указывает на то, что нет статистических различий между наблюдаемыми и ожидаемыми частотами. Это может произойти, когда все наблюдаемые значения точно соответствуют ожидаемым.

Значение P-value равно 1.0, говорит о том, что нет статистической значимости и нельзя отклонить нулевую гипотезу.

Исходя из этого, можно сказать, что данные не представляют статистически значимых доказательств для отклонения нулевой гипотезы в гипотезы в контексте теста хи-квадрат.

Далее вычислим коэффициент Спирмена для времени проведенного в социальных сетях и детокса от них. Для этого была использована функция spearmanr из библиотеки scipy.

```python
def spreamnan(data):
    # Извлекаем ответы для каждой группы для вопросов о социальных сетях и времени
    group1_social_media_data = [
        item[1] for item in data[0] if "социальных сетях" in item[0] or "Делаете ли Вы" in item[0]
    ]
    group2_social_media_data = [
        item[1] for item in data[1] if "социальных сетях" in item[0] or "Делаете ли Вы" in item[0]
    ]

    # Преобразуем категории в числовые значения (предполагая, что "больше 3 часов" = 1, "меньше 3 часов" = 0)
    category_mapping = {"больше 3 часов": 1, "меньше 3 часов": 0}
    group1_numeric_data = [category_mapping.get(answer, answer) for answer in group1_social_media_data]
    group2_numeric_data = [category_mapping.get(answer, answer) for answer in group2_social_media_data]

    # Вычисляем коэффициент корреляции Спирмена
    correlation_coefficient, _ = spearmanr(group1_numeric_data, group2_numeric_data)

    print(f"Коэффициент корреляции Спирмена для вопросов о социальных сетях и времени: {correlation_coefficient}")
```

Рисунок 10. Код вычисления коэффициента Спирмена.

После выполнения кода на рисунке 10 получим следующий результат: 0.963

```
Коэффициент корреляции Спирмена для вопросов о социальных сетях и времени: 0.9629302235284835
```

Рисунок 11. результат выполнения программы

Коэффициент Спирмена лежит в диапазоне от -1 до 1, где:
1 указывает на полную положительную корреляцию (то есть, если одна переменная увеличивается, другая тоже увеличивается с постоянной скоростью),
-1 указывает на полную отрицательную корреляцию (то есть, если одна переменная увеличивается, другая уменьшается с постоянной скоростью),
0 указывает на отсутствие корреляции.

Коэффициент Спирмена равный 0.963 довольно близок к единице, что может указывать на сильную положительную корреляции. Между временем, проведенным в социальных сетях, и необходимостью проведения детокса от них. То есть, люди, которые проводят больше времени в социальных сетях, чаще ощущают потребность в детоксе от них.

Вычислим еще один коэффициент Спирмена, но для времени, проведенного в социальных сетях и уровня стресса. Реализация кода представлена на рисунке 12.

```python
def spearman_2(data):
    stress_data = ["Да", "Нет", "Да", "Да", "Нет"]  # Пример данных по стрессу
    time_spent_data = ["Да", "Да", "Нет", "Нет", "Да"]  # Пример данных по времени в соц.сетях

    # Преобразование к числовым значениям
    numeric_stress_data = [1 if answer == "Да" else 0 for answer in stress_data]
    numeric_time_spent_data = [1 if answer == "Да" else 0 for answer in time_spent_data]

    # Рассчитываем коэффициент корреляции Спирмена
    spearman_corr, _ = spearmanr(numeric_stress_data, numeric_time_spent_data)

    print(f"Коэффициент корреляции Спирмена: {spearman_corr}")
```

Рисунок 12. Код вычисления коэффициента Спирмена.

```
Коэффициент корреляции Спирмена: -0.6666666666666667
```

Рисунок 13. результат выполнения программы

На рисунке 13 видим результат выполнения программы, где коэффициент Спирмена равен -0.667. Это говорит о том, что существует отрицательная

монотонная связь между временем, проведенным в социальных сетях, и уровнем стресса, и наоборот.

Однако важно отметить, что корреляция не означает причинно-следственную связь. В данном случае, нельзя с уверенностью сказать, что использование социальных сетей напрямую влияет на уровень стресса, так как другие факторы также могут оказывать влияние на оба параметра.

Вычислим коэффициент V Камера между детоксом и желанием проводить меньше времени в социальных сетях. На рис.14 представлен код.

```python
def v_kramer(data):
    # Создаем DataFrame
    df = pd.DataFrame([dict(item) for item in data])

    # Пересчитываем категориальные переменные в бинарные значения
    df['Детокс'] = (df['Делаете ли Вы "детокс" от социальных сетей ? Хотя бы на несколько часов'] == 'Да').astype(int)
    df['Меньше времени'] = (df['Хотели бы Вы проводить меньше времени в соц.сетях ?'] == 'Да').astype(int)

    # Убираем лишние столбцы
    df = df[['Детокс', 'Меньше времени']]

    # Создаем таблицу сопряженности
    contingency_table = pd.crosstab(df['Детокс'], df['Меньше времени'])

    # Рассчитываем коэффициент V Крамера
    chi2, _, _, _ = chi2_contingency(contingency_table)
    n = contingency_table.sum().sum()
    min_dim = min(contingency_table.shape) - 1
    cramers_v = np.sqrt(chi2 / (n * min_dim)) if (n * min_dim) != 0 else 0

    print(f"Коэффициент V Крамера между детоксом и желанием проводить меньше времени в соц.сетях: {cramers_v}")
```

Рисунок 14. Код вычисления коэффициента V Крамера.

```
Коэффициент V Крамера между детоксом и желанием проводить меньше времени в соц.сетях: 0.0854433718937274
```

Рисунок 15. Результат выполнения программы

В результате получили коэффициент V Крамера равный 0.085. Напомним, что коэффициент V Крамера является мерой ассоциации между двумя категориальными переменными в таблице сопряженности. Он принимает значения от -1 до 1, где 0 означает отсутствие ассоциации, а 1 (или -1) указывает на полную ассоциацию.

Значение 0.085 для коэффициента V Крамера, скорее всего, указывает на низкую степень ассоциации между переменными "детокс" и "желание

проводить меньше времени в соц.сетях". То есть, существует некоторая связь между этими двумя переменными, но она не сильная.

В контексте проведенного опроса это может означать, что желание проводить меньше времени в социальных сетях не обязательно связано с процессом детоксикации в значительной степени.

Посчитаем еще один коэффициент V Крамера, но между стрессом и желанием проводить меньше времени в социальных сетях. Реализация кода представлена на рисунке 16.

```python
def v_kramer_2(data):
    df = pd.DataFrame([dict(item) for item in data])

    # Пересчитываем категориальные переменные в бинарные значения
    df['Стресс'] = (df['Бывали ли случаи, когда вы испытывали стресс из-за событий в социальных сетях?'] == 'Да').astype(int)
    df['Меньше времени'] = (df['Хотели бы Вы проводить меньше времени в соц.сетях ?'] == 'Да').astype(int)

    # Убираем лишние столбцы
    df = df[['Стресс', 'Меньше времени']]

    # Создаем таблицу сопряженности
    contingency_table = pd.crosstab(df['Стресс'], df['Меньше времени'])

    # Рассчитываем коэффициент V Крамера
    chi2, _, _, _ = chi2_contingency(contingency_table)
    n = contingency_table.sum().sum()
    min_dim = min(contingency_table.shape) - 1
    cramers_v = np.sqrt(chi2 / (n * min_dim)) if (n * min_dim) != 0 else 0

    print(f"Коэффициент V Крамера между стрессом и желанием проводить меньше времени в соц.сетях: {cramers_v}")
```

Рисунок 16. Код вычисления коэффициента V Крамера.

```
Коэффициент V Крамера между стрессом и желанием проводить меньше времени в соц.сетях: 0.0
```

Рисунок 17. Результат выполнения программы.

В результате получили 0, это означает, что в опросных данных не обнаружено статистической связи между фактом проведения детокса и желанием проводить меньше времени в социальных сетях. Другими словами, эти два фактора не коррелируют между собой на основе предоставленных данных.

ЗАКЛЮЧЕНИЕ

В ходе проведенного исследования была оценена связь между детоксом от социальных сетей, временем, проведенным в них, и некоторыми психологическими аспектами, такими как потребность в детоксе и уровень стресса. Результаты позволяют сделать следующие основные выводы:

1. Отсутствие связи между детоксом и желанием проводить меньше времени в социальных сетях: коэффициент V Крамера и его значение (0.0 и 0.085) указывают на отсутствие статистически значимой связи между фактом проведения детокса и желанием уменьшить время в социальных сетях.
2. Связь между временем в социальных сетях и психологическими аспектами: коэффициент Спирмена 0.963 указывает на сильную положительную корреляцию между временем, проведенным в социальных сетях, и потребностью в детоксе. Тем не менее, важно отметить, что корреляция не подразумевает причинно-следственной связи.
3. Связь между временем в социальных сетях и уровнем стресса: коэффициент Спирмена -0.667 указывает на отрицательную монотонную связь между временем в социальных сетях и уровнем стресса, что может свидетельствовать о том, что более высокое время в соцсетях ассоциируется с более низким уровнем стресса.

Важно отметить, что корреляция не обозначает причинно-следственную связь, и существуют другие факторы, влияющие на эти взаимосвязи. Данные основаны на опросах, что подразумевает возможную субъективность.

Полученные результаты предостерегают от однозначных интерпретаций и подчеркивают сложность взаимосвязей между использованием социальных сетей, детоксом и психологическим благополучием. Дополнительные

исследования с учетом дополнительных факторов могут более точно выявить тенденции и динамику в данной области.

ЛИТЕРАТУРА

1. Гмурман В. Е. Теория вероятностей и математическая статистика: учебное пособие для вузов / В. Е. Гмурман. – М.: Высш. шк., 2003. – 479 с.
2. Новиков Д. А. Статистические методы в педагогических исследованиях (типовые случаи) / Д. А. Новиков. – М.: МЗ-Пресс, 2004. – 67 с.
3. Шилова З. В. Теория вероятностей и математическая статистика: учебное пособие / З. В. Шилова, О. И. Шилов. – Киров: Изд-во ВГГУ, 2015. – 158 с
4. Шихалёв А.М. Корреляционный анализ. Непараметрические методы / А.М. Шихалёв. – Казань: Казан. ун-т, 2015. – 58 с.
5. Сидоренко Е. Методы математической обработки в психологии. СПб., 2002;

ПРИЛОЖЕНИЕ

Ссылка на листинг кода: ссылка

Ссылка на опрос: ссылка