|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное автономное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ Информатика и системы управления

КАФЕДРА Системы обработки информации и управления

**ОТЧЕТ**

***ПО ЛАБОРАТОРНОЙ РАБОТЕ №1***

***ПО ДИСЦИПЛИНЕ***

***«МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ»***

***НА ТЕМУ:***

**Создание "истории о данных"**

Студент \_\_ИУ5И-23М\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_Фу Чэньтянь\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Преподаватель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_Ю.Е.Гапанюк\_\_**

(Подпись, дата) (И.О.Фамилия)

2025 г.

**ВВЕДЕНИЕ**

Data Storytelling - это метод коммуникации, который объединяет данные и повествование, чтобы эффективно передать информацию или инсайты, полученные из данных, аудитории. Он использует элементы рассказывания историй для создания контекста и динамики в данных, делая их более понятными и запоминающимися для людей.

Синонимом термина "история о данных" является термин "история на основе данных". Цель лабораторной работы: изучение различных методов визуализация данных и создание истории на основе данных. Краткое описание: Построение графиков, помогающих понять структуру данных, и их интерпретация. Основой лабораторной работы является методология визуализации данных data-to-viz.

**ЗАДАНИЕ**

Выбрать набор данных (датасет).

Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
4. Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
5. История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

Сформировать отчет и разместить его в своем репозитории на github.

**ХОД ВЫПОЛНЕНИЯ РАБОТЫ**

# Часть 1. Текстовое описание набора данных

Этот набор данных собирает различную информацию о здоровье сна и образе жизни и используется для анализа влияния различных факторов на сон.

Он включает в себя такие данные, как пол, возраст, род занятий, продолжительность сна, качество сна, уровень физической активности, уровень стресса, категорию ИМТ, артериальное давление, частоту сердечных сокращений, количество шагов в день, а также наличие или отсутствие нарушений сна.

图形用户界面, 文本

AI 生成的内容可能不正确。

Рисунок 1: Информация о наборе данных

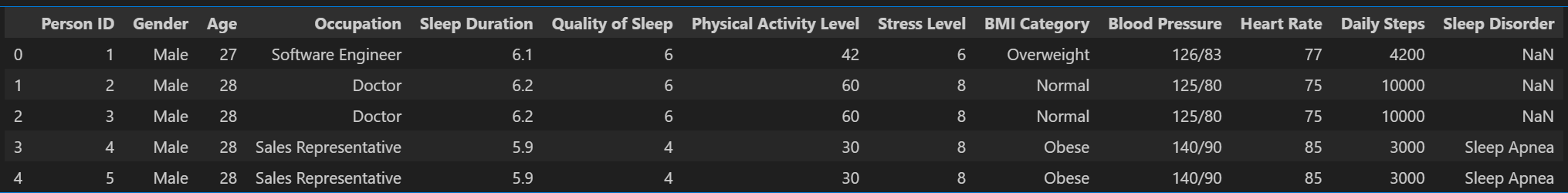


Рисунок 2: Первые 5 строк набора данных

# Часть 2. Визуальное исследование датасета

Для визуального исследования могут быть использованы различные виды диаграмм, мы построим только некоторые варианты диаграмм, которые используются достаточно часто.

1. Диаграмма рассеяния

Позволяет построить распределение двух колонок данных и визуально обнаружить наличие зависимости. Не предполагается, что значения упорядочены (например, по времени).

Проводим исследования о Daily Steps и Stress Level：

plt.figure(figsize=(10,6))

sns.scatterplot(

    data= df,

    x='Daily Steps',

    y='Stress Level',

    hue='BMI Category',

    palette='magma',

    size='Age',

    sizes=(20,200),

    alpha=0.7

)

plt.title('Steps vs stress level(group by bmi and age)')

plt.xlabel('steps')

plt.ylabel('stress level(1-10)')

plt.axvline(x=8000, color='green', linestyle='--', alpha=0.3)  # 推荐步数参考线

plt.legend(bbox\_to\_anchor=(1.05, 1), loc='upper left')

plt.tight\_layout()

plt.show()

图表, 散点图

AI 生成的内容可能不正确。

Рисунок 3: Диаграмма рассеяниях - Price vs Mileage

1. Гистограмма

Гистограмма -позволяет оценить плотность вероятности распределения данных. Строим гистограммы для числовых данных: 'age', ' Sleep Duration', Daily Steps', ' Stress Level'.

图表, 直方图

AI 生成的内容可能不正确。

图表, 条形图, 直方图

AI 生成的内容可能不正确。

图表, 条形图

AI 生成的内容可能不正确。

图表, 条形图

AI 生成的内容可能不正确。

Рисунок 3: Гистограмма числовых данных

Комбинация гистограмм и диаграмм рассеивания：

sns.jointplot(x='Daily Steps', y='Stress Level', data=df, kind='scatter', marginal\_kws=dict(bins=20, color='lightblue'))

plt.show()

sns.jointplot(x='Daily Steps', y='Stress Level', data=df, kind='hex', marginal\_kws=dict(bins=20, color='lightblue'))

plt.show()

图表

AI 生成的内容可能不正确。

图表, 气泡图

AI 生成的内容可能不正确。

Рисунок 8: Комбинация гистограмм и диаграмм рассеивания 'Daily Steps vs Stress Level'

1. Ящик с усами

Отображает одномерное распределение вероятности：

plt.figure(figsize=(10, 6))

sns.boxplot(x='Occupation', y='Quality of Sleep', data=df)

plt.title('Quality of Sleep by Occupation')

plt.xticks(rotation=45)

plt.show()

图表, 箱线图

AI 生成的内容可能不正确。

Рисунок 9: Ящик с усами ' Quality of Sleep by Occupation'

1. Violin plot

Похоже на предыдущую диаграмму, но по краям отображаются распределения плотности - <https://en.wikipedia.org/wiki/Kernel_density_estimation>

plt.figure(figsize=(10, 6))

sns.violinplot(x='Occupation', y='Quality of Sleep', data=df)

plt.title('Quality of Sleep by Occupation')

plt.xticks(rotation=45)

plt.show()

图表

AI 生成的内容可能不正确。

Рисунок 11: Violin plot ' Quality of Sleep by Occupation'

Из приведенных графиков видно, что violinplot действительно показывает распределение плотности.

# Часть 3. Информация о корреляции признаков

Проверка корреляции признаков позволяет решить две задачи:

* Понять какие признаки (колонки датасета) наиболее сильно коррелируют с целевым признаком. Именно эти признаки будут наиболее информативными для моделей машинного обучения. Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. Нужно отметить, что некоторые алгоритмы машинного обучения автоматически определяют ценность того или иного признака для построения модели.
* Понять какие нецелевые признаки линейно зависимы между собой. Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если несколько признаков линейно зависимы, то для построения модели из них выбирают какой-то один признак.

Получена корреляционная матрица числовых данных в датасете:

numerical\_columns = df.select\_dtypes(include=['int64']).columns

correlation\_matrix = df[numerical\_columns].corr()

print(correlation\_matrix)

图片包含 图形用户界面

AI 生成的内容可能不正确。

Рисунок 13: Корреляционная матрица

После проведения корреляционного анализа набора данных мы обнаружили следующие важные корреляционные особенности:

**Ключевые связи с качеством сна**

Качество сна и уровень стресса: -0,90

Сильная отрицательная корреляция: чем сильнее стресс, тем значительно хуже качество сна.

Практическая значимость: Управление стрессом может быть наиболее эффективным методом улучшения сна.

Качество сна и частота сердечных сокращений: -0,66

Умеренная отрицательная корреляция: высокая частота сердечных сокращений (которая может быть признаком тревожности или проблем со здоровьем) была связана с плохим качеством сна.

Качество сна и возраст: 0,47

Слабая положительная корреляция: пожилые люди сообщают о более высоком качестве сна (может быть связано с предвзятостью выборки или изменениями в уровне стресса в жизни).

**Переменные, связанные с упражнениями**

Ежедневный объем шагов и упражнений: 0,77

Сильная корреляция с ожиданиями: это подтверждает рациональность данных (люди, которые делают больше шагов, занимаются спортом дольше).

Упражнения и стресс: -0,03

Отсутствие корреляции: продолжительность упражнений не имеет корреляции с уровнем стресса (вопреки здравому смыслу).

Ежедневные шаги и стресс: 0,19

Слабая положительная корреляция: противоречит здравому смыслу (обычно считается, что физические упражнения снижают стресс).

Возможное объяснение: люди с высоким уровнем стресса могут снимать стресс, увеличивая количество шагов.

По умолчанию при построении матрицы используется коэффициент корреляции Пирсона. Возможно также построить корреляционную матрицу на основе коэффициентов корреляции Кендалла и Спирмена. На практике три метода редко дают значимые различия.

文本

AI 生成的内容可能不正确。

Рисунок 14: Корреляционная матрица по методу Пирсона

文本

AI 生成的内容可能不正确。

Рисунок 15: Корреляционная матрица по методу Кендалла

文本

AI 生成的内容可能不正确。

Рисунок 16: Корреляционная матрица по методу Спирмена

В случае большого количества признаков анализ числовой корреляционной матрицы становится неудобен.

Для визуализации корреляционной матрицы будем использовать "тепловую карту" heatmap которая показывает степень корреляции различными цветами.

plt.figure(figsize=(12, 8))

sns.heatmap(correlation\_matrix, annot=True, cmap=#'coolwarm'

            'YlGnBu'

            , fmt='.2f')

plt.title('Correlation Matrix Heatmap')

plt.show()

图表

AI 生成的内容可能不正确。

Рисунок 17: Тепловая карта ячейки с числовыми значениями

mask = np.triu(np.ones\_like(correlation\_matrix, dtype=bool))

plt.figure(figsize=(12, 8))

sns.heatmap(correlation\_matrix, mask=mask, annot=True, cmap='coolwarm', fmt='.2f')

plt.title('Lower Triangular Correlation Matrix Heatmap')

plt.show()

图表, 条形图, 瀑布图

AI 生成的内容可能不正确。

Рисунок 19: Тепловая карта треугольной матрицы

numerical\_columns = bmw\_data.select\_dtypes(include=['int64', 'float64']).columns

correlation\_matrix = bmw\_data[numerical\_columns].corr(

    #method='pearson'

    #method='kendall'

    method='spearman'

)

plt.figure(figsize=(12, 8))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt='.2f')

plt.title('Correlation Matrix Heatmap - spearman')

plt.show()

图形用户界面, 图表, 应用程序

AI 生成的内容可能不正确。

Рисунок 20: Тепловая карта по методу Пирсона

图形用户界面, 应用程序

AI 生成的内容可能不正确。

Рисунок 20: Тепловая карта по методу Кендалла

图形用户界面, 应用程序

AI 生成的内容可能不正确。

Рисунок 20: Тепловая карта по методу Спирмена

Необходимо отметить, что тепловая карта не очень хорошо подходит для определения корреляции нецелевых признаков между собой. Тепловая карта показывает корреляцию более интуитивно, и вывод тот же, что и в матрице корреляции, поскольку их данные одинаковы.

**ЗАКЛЮЧЕНИЕ**

Благодаря систематическому визуальному анализу наборов данных о здоровье сна и образе жизни мы выявляем ключевые факторы, влияющие на качество сна, и их взаимодействие.

Исследование показало, что уровень стресса имеет значительную отрицательную корреляцию с качеством сна (r=-0,90), что указывает на то, что управление психологическим стрессом является основным направлением улучшения сна; Продолжительность и качество сна имеют перевернутую U-образную зависимость, при этом оптимальный диапазон составляет 7–8 часов. Слишком длинный или слишком короткий может привести к снижению качества.

С точки зрения физиологических показателей положительная корреляция между возрастом и артериальным давлением, а также аномально высокая тенденция артериального давления у людей с ожирением указывают на то, что следует обратить внимание на влияние метаболического здоровья на сердечно-сосудистые заболевания.

Анализ поведенческих характеристик показал, что количество ежедневных шагов показало бимодальное распределение между «сидячими» и «активными» группами, а количество упражнений тесно коррелировало с количеством шагов (r=0,77), но его влияние на снятие стресса не достигло значительного уровня, что ставит под сомнение традиционные предположения о состоянии здоровья.

Исследование также выявило проблемы с качеством данных (например, аномально высокую корреляцию между персональным идентификатором и возрастом), которые необходимо устранить при последующем анализе.

Этот эксперимент преобразует сложные взаимосвязи в сфере здравоохранения в интуитивные идеи с помощью многомерных методов визуализации (таких как тепловые карты и комбинированные диаграммы рассеяния). Он не только обеспечивает научную основу для персонализированных медицинских вмешательств (например, составления планов сна для людей, испытывающих высокий уровень стресса), но и закладывает основу для проектирования признаков для построения моделей прогнозирования сна.