МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное образовательное учреждение

высшего образования

**«Северный (Арктический) федеральный университет имени М.В. Ломоносова»**

Высшая школа информационных технологий и автоматизированных систем

(наименование высшей школы / филиала / института / колледжа)

**КУРСОВАЯ РАБОТА**

|  |  |  |
| --- | --- | --- |
| По дисциплине/междисциплинарному курсу/модулю | | Сбор и подготовка данных |
|  | | |
|  | | |
| На тему | Анализ набора данных | |
|  | | |

|  |  |
| --- | --- |
|  | Выполнил обучающийся:  Архаров Никита Михайлович |
|  | (Ф.И.О.) |
|  | Направление подготовки / специальность:  09.03.02 Информационные системы и технологии |
|  | (код и наименование) |
|  | Курс: 3 |
|  | Группа: 351018 |
|  | Руководитель:  Березовский Владимир Валерьевич, профессор |
|  | (Ф.И.О. руководителя, должность / уч. степень / звание) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Отметка о зачете |  |  |  |  |
|  |  | (отметка прописью) |  | (дата) |
| Руководитель |  |  |  | В. В. Березовский |
|  |  | (подпись руководителя) |  | (инициалы, фамилия) |

Архангельск 2023

Министерство науки и высшего образования Российской Федерации

федеральное государственное автономное образовательное учреждение

высшего образования

**«Северный (Арктический) федеральный университет имени М.В. Ломоносова»**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | | | | | | | | | | | | | | | | |  |
|  | Кафедра информационных систем и технологий | | | | | | | | | | | | | | | | | | | | |  |
|  |  | | | | | | | | | | | | | | | | | | | | |  |
|  |  | | | | | | | | | | | | | | | | | | | | |  |
|  | **Задание на курсовую работу** | | | | | | | | | | | | | | | | | | | | |  |
|  |  | | | | | | | | | | | | | | | | | | | | |  |
|  | по дисциплине | | | Сбор и подготовка данных | | | | | | | | | | | | | | | | | |  |
|  |  | | | |  | | | | | | | | | | | | | | | | |  |
|  | студенту | ИТАС | | | высш. школы | | | | | | 3 | | | курса | 351018 | | | | | группы | |  |
|  | Архаров Никита Михайлович | | | | | | | | | | | | | | | | | | | | |  |
|  |  | | | | | | | | | | | | | | | | | | | | |  |
|  | |  | | --- | | 09.03.02 «Информационные системы и технологии» | |  | | | | | | | | | | | | | | | | | | | | | |  |
|  | ТЕМА: | | Анализ набора данных | | | | | | | | | | | | | | | | | | |  |
|  | ИСХОДНЫЕ ДАННЫЕ: | | | | | | | | | | | | | | | | | | | | |  |
|  | Цель: получить представление о наборе данных с помощью анализа и визуализации  Задачи:  1. Очистить и подготовить набор данных к анализу  2. Использовать различные графические методы для визуализации данных  3. Делать выводы  4. Использовать знания, полученные в ходе лабораторных работ по данной дисциплине, для успешного завершения анализа | | | | | | | | | | | | | | | | | | | | |  |
|  |  | | | | | | | | | | | | | | | | | | | | |  |
|  | Срок проектирования с | | | | | | « | 1 | » | сентября | | | 2022 г. по | | « | 10 | | » | января | | 2023 г. |  |
|  | Руководитель работы | | | | | профессор | | | | | |  |  | | | |  | В. В. Березовский | | | |  |
|  |  | | | |  |  | | | | | |  | (подпись) | | | |  |  | | | |  |
|  |  | | | |  |  | | | | | |  |  | | | |  |  | | | |  |
|  |  | | | |  |  | | | | | |  |  | | | |  |  | | | |  |

Архангельск 2023

Лист для замечаний

Оглавление

[Введение 5](#_Toc124120582)

[Основная часть 6](#_Toc124120583)

[1 Описание набора данных 6](#_Toc124120584)

[2 Первичный анализ данных 7](#_Toc124120585)

[3 Очистка и подготовка данных 8](#_Toc124120586)

[4 Повторный анализ 10](#_Toc124120587)

[5 Визуализация 12](#_Toc124120588)

[Заключение 19](#_Toc124120589)

[Список используемых источников 20](#_Toc124120590)

Введение

В этой работе будет проанализирован набор данных, чтобы лучше понять тенденции и закономерности в нём. Целью анализа является получение информации, которая может помочь в принятии решений или выделить области для дальнейшего изучения. Для этого необходимо выполнить следующие задачи:

- очистить данные;

- использовать различные графические методы для визуализации данных. Эти методы включают в себя гистограммы, круговые диаграммы, тепловые карты и точечные графики, которые позволяют нам эффективно донести наши выводы.

Основная часть

1 Описание набора данных

Выбранный набор данных [1] предоставляет уникальную возможность изучить структуру и тенденции уровня заболеваемости раком в Соединенных Штатах на уровне отдельных округов. Используя данные с сайта cancer.gov и Американского общественного опроса переписи населения США, эта база данных позволяет получить представление о том, как изменяются показатели смертности с поправкой на возраст, среднее число смертей в год и последние тенденции в разных округах - наряду с другими ключевыми показателями, такими как среднегодовое число заболевших, последние тенденции в показателях смертности и т.д., отраженные в глубокой многомерной базе данных.

Описание столбцов набора данных представлено в таблице 1.

Таблица 1 – Столбцы набора данных

|  |  |
| --- | --- |
| Название колонки | Описание |
| index | Индекс (int) |
| County | Название округа (str) |
| FIPS | Код округа по Федеральному стандарту обработки информации (int) |
| Met Objective of 45.5? (1) | Двоичное значение, указывающее, достиг ли округ цели 45,5 смертей с поправкой на возраст на 100 000 человек (bool) |
| Age-Adjusted Death Rate | Коэффициент смертности с поправкой на возраст на 100 000 человек (float) |
| Lower 95% Confidence Interval for Death Rate | Нижний 95% доверительный интервал для коэффициента смертности (float) |
| Upper 95% Confidence Interval for Death Rate | Верхний 95% доверительный интервал для коэффициента смертности (float) |
| Average Deaths per Year | Среднее количество смертей в год в округе (int) |

Продолжение таблицы 1

|  |  |
| --- | --- |
| Recent Trend (2) | Последние тенденции в уровне смертности с поправкой на возраст на 100 000 человек. (enum) |
| Recent 5-Year Trend (2) in Death Rates | Последние 5 лет тенденции в показателях смертности с поправкой на возраст на 100000 человек (float) |
| Lower 95% Confidence Interval for Trend | Нижний 95% доверительный интервал для тенденции (float) |
| Upper 95% Confidence Interval for Trend | Верхний 95% доверительный интервал для тенденции (float) |

2 Первичный анализ данных

Краткая информация о наборе данных, включая имена столбцов, типы данных каждого столбца и количество ненулевых значений представлено на рисунке 1.

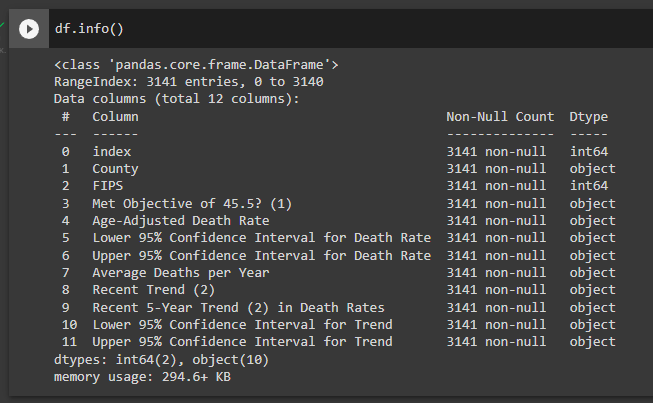


Рисунок 1 – Краткая информация о необработанном наборе данных

Как мы видим, большинство числовых столбцов имеет тип данных object. В дальнейшем нужно будет преобразовать в числовой формат некоторые столбцы.

Сводная статистика для числовых столбцов в DataFrame, включая среднее значение, стандартное отклонение, минимальное и максимальное значения, а также квартили представлена на рисунке 2.

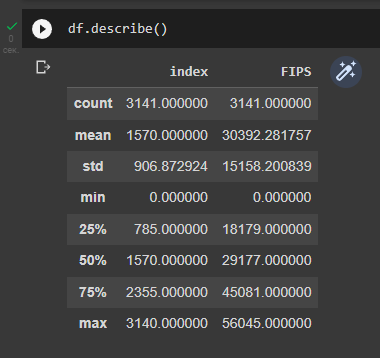


Рисунок 2 – Сводная статистика неподготовленных данных

Поскольку данные неподготовлены, результат выше даёт неполную информацию.

3 Очистка и подготовка данных

Для подготовки данных, можно выделить следующие задачи:

- убрать записи, в которых содержится «\*»;

- убрать записи, в которых содержится «\*\*»;

- переформатировать числа (вместо «157,376» нужно «157376»);

- преобразить некоторые столбцы в числовой тип данных.

Удаление строк с определенными символами позволяет нам избавиться от пропусков в данных. Результат такого очищение показан на рисунке 3.

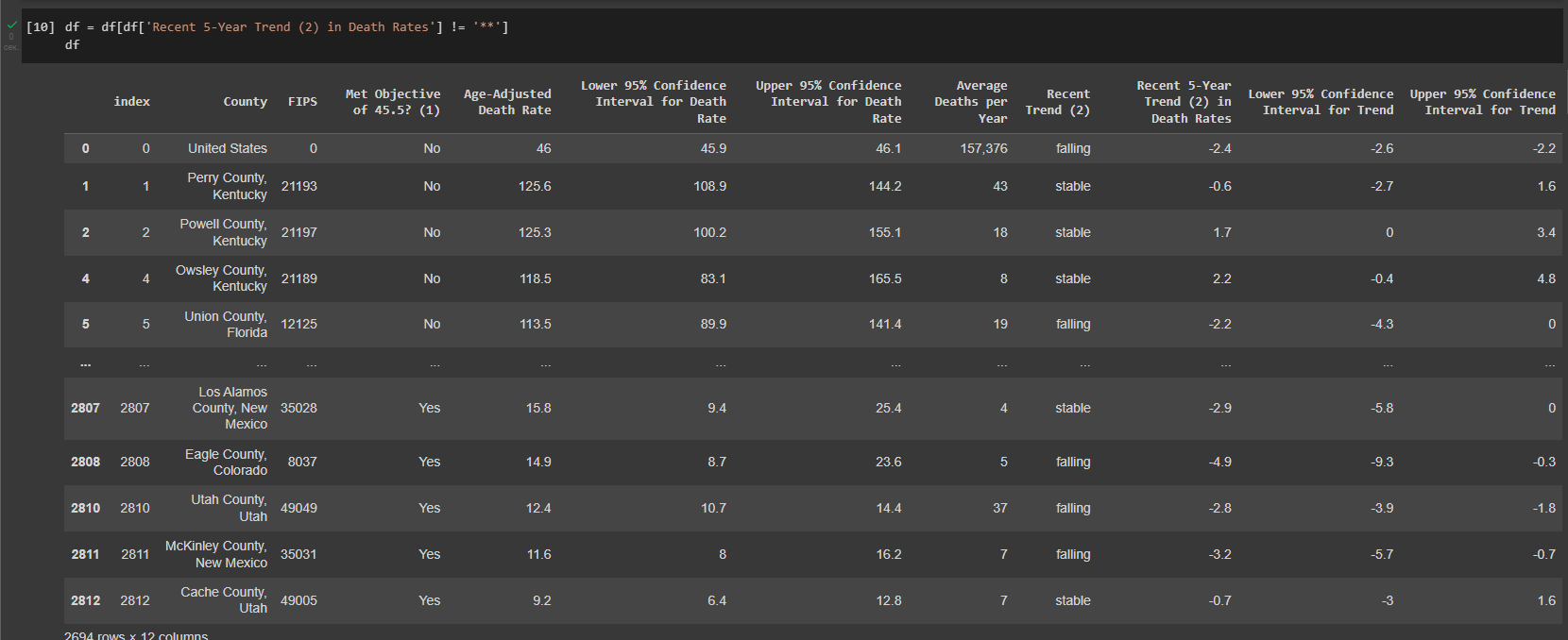


Рисунок 3 – Результат удаления строк с определенными символами

В столбце «Average Deaths per Year» присутствует форматирование чисел, что мешает преобразовать их из строкового формата в числовой. Например, вместо «157376» в наборе данных хранится «157,376». Нам нужно убрать эту запятую. Результат данной очистки показан на рисунке 4.

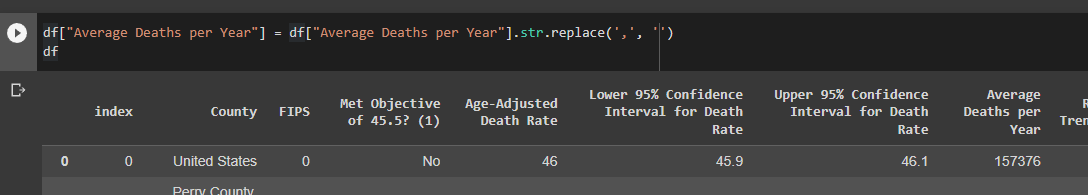


Рисунок 4 – Форматирование чисел

Действия выше позволяют преобразовать числа, которые хранятся в наборе данных строками, в числовой тип данных. Результат такого преобразования показан на рисунке 5.

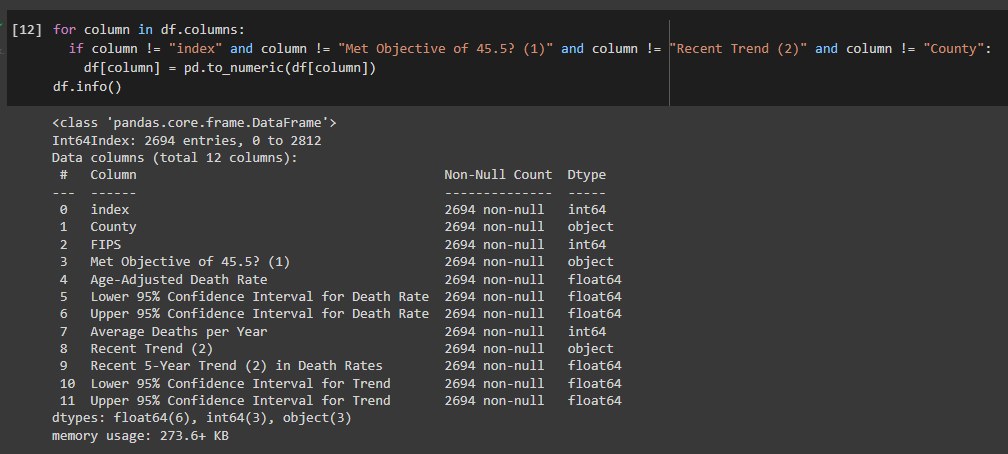


Рисунок 5 – Преобразование чисел в числовой тип данных

4 Повторный анализ

Повторная сводная статистика представлена на рисунке 6.

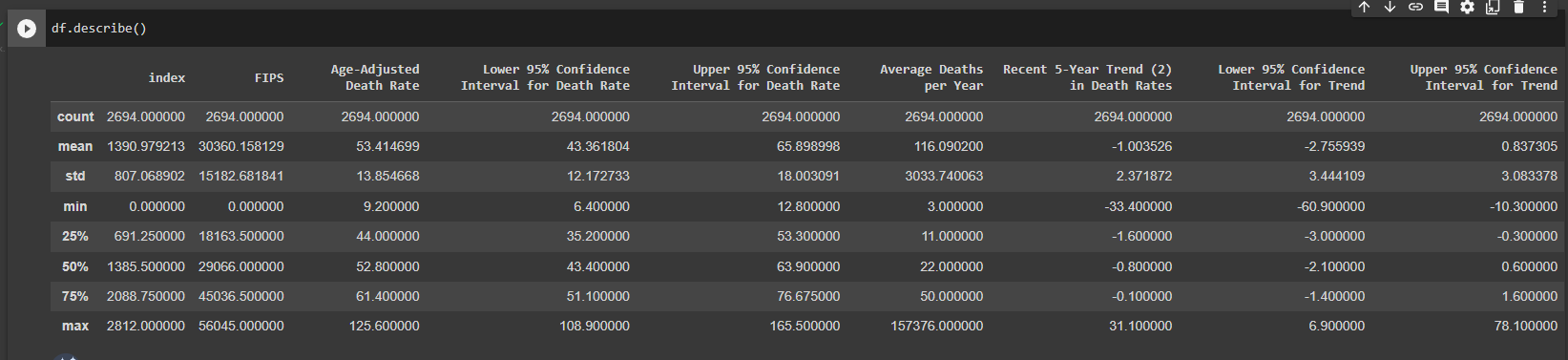


Рисунок 6 – Сводная статистика подготовленных данных

Из этого можно сделать несколько выводов:

- Данные состоят из 2694 наблюдений (как указано в строке «count»).

- Коэффициент смертности с поправкой на возраст в среднем составляет около 30360,158129 (как указано в строке «mean» для этой переменной).

- Уровень смертности имеет относительно высокое стандартное отклонение 15182,681841 (как указано в строке «std» для этой переменной), что говорит о значительном разбросе в уровне смертности среди различных наблюдений.

- Минимальное значение коэффициента смертности равно 0, а максимальное - 56045 (как указано в строках «min» и «max» для этой переменной).

- Первый квартиль (25-й процентиль) коэффициента смертности составляет 18163,5, а третий квартиль (75-й процентиль) - 45036,5 (как указано в строках «25%» и «75%» для этой переменной). Это говорит о том, что уровень смертности относительно перекошен, с длинным хвостом высоких значений.

- Недавняя 5-летняя тенденция в уровне смертности в среднем составляет около 116,09 (как указано в строке «mean» для этой переменной).

- Тенденция в уровне смертности имеет очень высокое стандартное отклонение 3033,74 (как указано в строке «std» для этой переменной), что говорит о том, что существует значительный разброс в тенденции среди различных наблюдений.

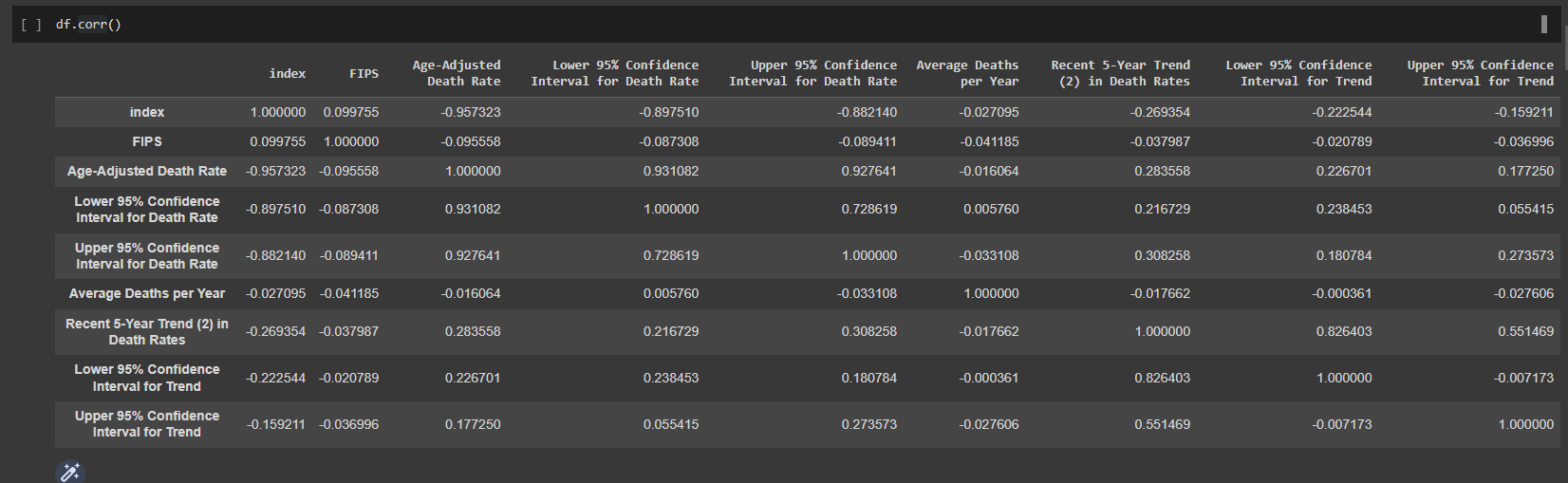


Рисунок 7 – Корреляция столбцов

На рисунке 7 показана корреляция между различными переменными. Значения в ячейках таблицы показывают силу и направление связи между двумя переменными в соответствующей строке и столбце.

Положительное значение указывает на положительную связь, то есть при увеличении одной переменной другая переменная также имеет тенденцию к увеличению. Отрицательное значение указывает на отрицательную связь, то есть при увеличении одной переменной другая переменная имеет тенденцию к уменьшению.

Например, значение в ячейке на пересечении строки «Смертность с поправкой на возраст» и столбца «Последние 5 лет тенденция (2) в показателях смертности» равно 0,283558. Это указывает на наличие положительной связи между коэффициентом смертности с поправкой на возраст и недавней пятилетней тенденцией в коэффициентах смертности.

Стоит также отметить, что значение в ячейке на пересечении строки «index» и столбца «index» равно 1. Это объясняется тем, что корреляция между переменной и самой собой всегда равна 1.

В целом, похоже, что существует сильная отрицательная связь между коэффициентом смертности с поправкой на возраст и индексом, а также сильная положительная связь между недавней 5-летней тенденцией в коэффициентах смертности и индексом. Между другими переменными есть и более слабые связи, но они не такие сильные, как эти две.

5 Визуализация

На рисунке 8 изображена тепловая карта корреляций между переменными в данных. Тепловая карта – это графическое представление данных, где значения представлены в виде цветов. В данном случае тепловая карта показывает корреляции между переменными в данных.

На основе этой тепловой карты можно увидеть силу и направление взаимосвязей между различными переменными в данных. Например, если цвет в ячейке белый, это указывает на положительную связь между переменными, а если чёрный - на отрицательную. Чем темнее или ярче цвет, тем сильнее взаимосвязь.

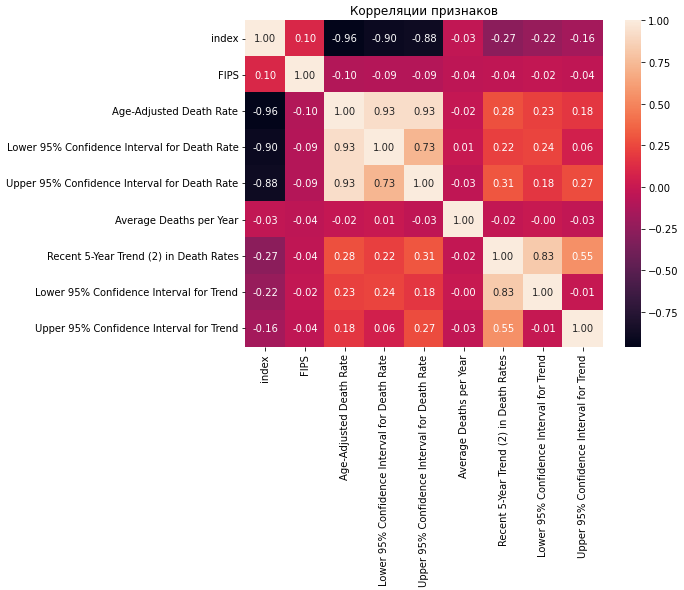


Рисунок 8 – Тепловая карта

В столбце «Достиг ли округ цели 45,5 скорректированных по возрасту показателей смертности на 100 000 человек» хранятся только значения «Да» или «Нет». По круговой диаграмме, представленной на рисунке 9, можно сделать вывод, что большинство округов не достигло цели.

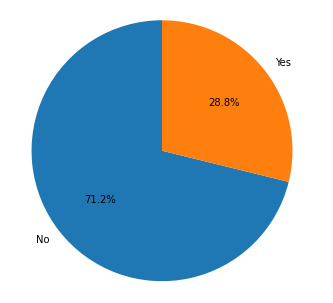


Рисунок 9 – Круговая диаграмма «Достиг ли округ цели 45,5 показателя смертности на 100 000 человек»

В столбце «Последние тенденции» хранятся только значения «Увеличивается», «Падает» или «Стабильная». По круговой диаграмме, представленной на рисунке 10, можно сделать вывод, что у большинства округов наблюдается стабильная тенденция смертности на 100 000 человек, меньше – у округов с падающей тенденцией, минимальный процент у округов с увеличивающейся тенденцией смертности.

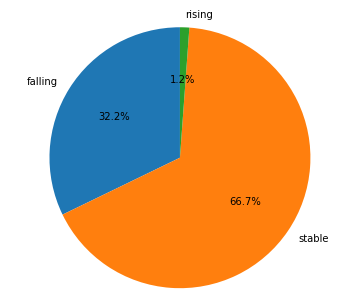


Рисунок 10 – Круговая диаграмма «Последняя тенденция смертности на 100 000 человек»

Эти графики можно представить в виде гистограмм, чтобы посмотреть на данные не в процентах, а в значениях. Такие гистограммы представлены на рисунках 11 и 12.

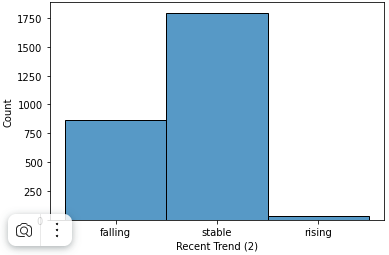


Рисунок 11 – Гистограмма для признака «Recent Trend»

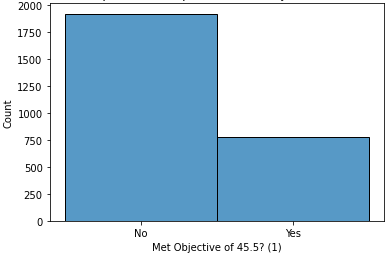


Рисунок 12 - Гистограмма для признака «Met Objective of 45.5?»

На рисунке 13 можно увидеть гистограмму коэффициента смертности с поправкой на возраст. По нему можно сделать вывод, что большинство смертей происходит в возрасте от 40 до 70 лет.

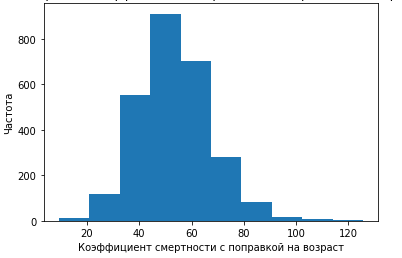


Рисунок 13 – Гистограмма коэффициента смертности с поправкой на возраст

На рисунке 14 можно наблюдать график смертностей с поправкой на возраст. Ось x представляет возрастной диапазон [4].

"Усы", отходящие от рамки при x=~20 и x=~90, представляют минимальные и максимальные значения в возрастных диапазонах ~20-45 и ~90-130 лет соответственно. Точки за пределами этих диапазонов могут представлять выбросы или точки данных, которые значительно выше или ниже остальных значений в наборе данных.

Прямоугольник между x=~45 и x=~63 представляет диапазон от первого квартиля (25-й процентиль) до третьего квартиля (75-й процентиль) в возрастном диапазоне 45-90 лет. Линия внутри прямоугольника при x=~55 представляет среднее значение в этом возрастном диапазоне. Это говорит о том, что большинство значений в этом возрастном диапазоне попадают в диапазон первого и третьего квартилей, при этом медианное значение находится в центре.

График предоставляет информацию о разбросе и асимметрии данных, а также о любых потенциальных выбросах.

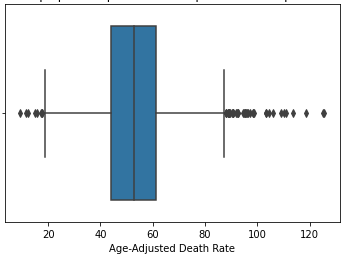


Рисунок 14 – График смертностей с поправкой на возраст

Между столбцами «Коэффициент смертности с поправкой на возраст» и «Средняя смертность в год» были найдены коэффициент Пирсона [2] и число p [3]. Результат можно увидеть на рисунке 15.

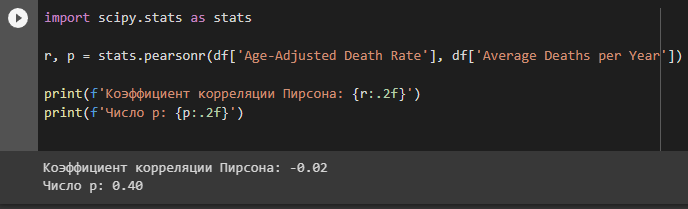


Рисунок 15 – Коэффициент Пирсона и число p

Основываясь на этих данных, можно сделать вывод, что существует слабая отрицательная корреляция между двумя столбцами: "Коэффициент смертности с поправкой на возраст" и "Средняя смертность в год". Коэффициент корреляции Пирсона, представленный переменной "r", равен -0,02, что близко к 0 и, следовательно, указывает на слабую корреляцию. Кроме того, значение p, представленное переменной "p", равно 0,40, что больше 0,05. Это говорит о том, что корреляция между двумя переменными не является статистически значимой.

Заключение

В заключение следует отметить, что целью данной работы был анализ набора данных для получения информации, которая может быть использована для принятия решений или определения областей для дальнейшего изучения. Для достижения этой цели было выполнено несколько задач, включая очистку данных и использование различных графических методов для визуализации данных.

В качестве графических методов использовались тепловые карты, круговые диаграммы, гистограммы и квадратные диаграммы. Эти усилия увенчались успехом в эффективной передаче выводов и достижении цели анализа.

Однако первоначальный анализ был затруднен тем, что данные не были в пригодной для использования форме: числа хранились в виде строк, а в колонках с числовыми данными присутствовали нечисловые символы. Это потребовало дополнительной очистки и подготовки данных. Процесс очистки данных включал удаление записей с символами «\*» и «\*\*», переформатирование чисел и преобразование некоторых столбцов в числовые типы данных.

В ходе работы удалось использовать знания, полученные в ходе лабораторных работ по данной дисциплине, для подготовки, анализа и визуализации набора данных.

Список используемых источников

1. Cancer County-Level // Kaggle : сайт. – URL: https://www.kaggle.com/datasets/thedevastator/exploring-county-level-correlations-in-cancer-ra?select=death+.csv (дата обращения: 08.01.2023)
2. Коэффициент корреляции Пирсона // Studfiles : сайт. – URL: https://studfile.net/preview/2966946/page:34/ (дата обращения: 08.01.2023)
3. Объясняем p-значения для начинающих Data Scientist’ов // Хабр : сайт. – URL: https://habr.com/ru/company/skillfactory/blog/510688/ (дата обращения: 08.01.2023)
4. Ящик с усами // Wikipedia : сайт. – URL: https://ru.wikipedia.org/wiki/Ящик\_с\_усами (дата обращения: 08.01.2023)

Приложение 1. Листинг кода курсовой работы

|  |  |
| --- | --- |
| # -\*- coding: utf-8 -\*-  """Arharov\_data2\_RGR.ipynb  Automatically generated by Colaboratory.  Original file is located at  https://colab.research.google.com/drive/1ty8CqfWWqzxy4w4SW2lUo1iN5OLjnMtl  ## Подключение библиотек, загрузка датасета  """  from google.colab import drive  drive.mount('/content/gdrive')  import pandas as pd  import numpy as np  import matplotlib.pyplot as plt  import matplotlib.patches as mpatches  import seaborn as sns  from collections import Counter  import warnings  warnings.filterwarnings('ignore')  df = pd.read\_csv('/content/gdrive/My Drive/data/death.csv')  """Выведем первые несколько строк DataFrame, чтобы убедиться в успешной загрузке данных."""  df.head()  """## Первичный анализ  Получим краткую информацию о DataFrame, включая имена столбцов, типы данных каждого столбца и количество ненулевых значений.  plt.show()  df.plot.scatter(x='index', y='Upper 95% Confidence Interval for Death Rate',  title="Верхний 95% доверительный интервал для\nкоэффициента смертности по округам")  df.plot.scatter(x='index', y='Age-Adjusted Death Rate',  title="Коэффициент смертности с поправкой \nна возраст по округам")  df.plot.scatter(x='index', y='Lower 95% Confidence Interval for Death Rate',  title="Нижний 95% доверительный интервал \nдля коэффициента смертности по округам")  df.plot.scatter(x='Upper 95% Confidence Interval for Death Rate', y='Age-Adjusted Death Rate',  title="Верхний 95% доверительный интервал для коэффициента \nсмертности в зависимотси от возраста")  df.plot.scatter(x='index', y='Recent 5-Year Trend (2) in Death Rates',  title="Последние 5 лет тенденции в показателях смертности с\nпоправкой на возраст на 100 000 человек по округам")  #Данные  data = Counter(df['Met Objective of 45.5? (1)'])  labels = data.keys()  sizes = data.values()  #Круговая диаграмма  plt.figure(figsize=(6,6))  plt.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90)  plt.title('Достиг ли округ цели 45,5 показателей смертности на 100 000 человек.')  plt.show()  #Данные  data = Counter(df['Recent Trend (2)'])  labels = data.keys()  sizes = data.values()  #Круговая диаграмма  plt.figure(figsize=(6,6))  plt.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90)  plt.title('Последняя тенденция изменения скорректированного по \nвозрасту коэффициента смертности на 100 000 человек.')  plt.show()  #Гистограмма  sns.histplot(df,x="Recent Trend (2)")  plt.title("Гистограмма для признака Recent Trend")  plt.show()  #Гистограмма  sns.histplot(df,x="Met Objective of 45.5? (1)")  plt.title("Гистограмма для признака Met Objective of 45.5?")  plt.show()  plt.hist(df['Age-Adjusted Death Rate'])  plt.title('Гистограмма коэффициента смертности с поправкой на возраст')  plt.xlabel('Коэффициент смертности с поправкой на возраст')  plt.ylabel('Частота')  plt.show()  sns.scatterplot(x='Age-Adjusted Death Rate', y='Average Deaths per Year', data=df)  в год, последней пятилетней тенденции (2) в коэффициентах \nсмертности и последней тенденции (2)')  plt.show()  sns.boxplot(x='Met Objective of 45.5? (1)', y='Age-Adjusted Death Rate', data=df)  plt.title('Ящик с усами скорректированной по возрасту \nсмертности по достижению цели 45,5? (1)')  plt.show()  grouped\_data = df.groupby('Met Objective of 45.5? (1)')  for name, group in grouped\_data:  plt.plot(group.index, group['Age-Adjusted Death Rate'], label=name)  plt.legend()  plt.title('Линейный график скорректированной по возрасту смертности \nпо достижению цели 45,5? (1)')  plt.show() | """  df.info()  """Как мы видим, большинство числовых столбцов имеет тип данных object. В дальнейшем нужно будет это преобразовать в числовой формат.  Получим сводную статистику для числовых столбцов в DataFrame, включая среднее значение, стандартное отклонение, минимальное и максимальное значения, а также квартили.  """  df.describe()  """Поскольку данные неподготовлены, результат выше даёт неполную информацию.  ## Очистка данных  Для начала уберем записи, в которых содержится строковый символ "\*":  """  df = df[df['Met Objective of 45.5? (1)'] != '\*']  df  """Повторим это для "\*\*":"""  df = df[df['Recent 5-Year Trend (2) in Death Rates'] != '\*\*']  df  """В столбце "Average Deaths per Year" присутствует форматирование чисел, что мешает преобразовать их из строкового формата в числовой. Например, вместо "157376" в наборе данных хранится "157,376". Нам нужно убрать эту запятую:"""  df["Average Deaths per Year"] = df["Average Deaths per Year"].str.replace(',', '')  df  """Наконец, преобразуем типы object в числа:"""  for column in df.columns:  if column != "index" and column != "Met Objective of 45.5? (1)" and column != "Recent Trend (2)" and column != "County":  df[column] = pd.to\_numeric(df[column])  df.info()  """## Повторный анализ"""  df.describe()  df.corr()  """## Визуализация"""  df.columns  #Данные  data = df.corr()  #Тепловая карта  plt.figure(figsize=(8,6))  sns.heatmap(data,annot=True, fmt=".2f")  plt.title('Корреляции признаков')  plt.title('Диаграмма разброса показателей смертности с поправкой \nна возраст в сравнении со средней смертностью за год')  plt.show()  grouped\_data = df.groupby('Met Objective of 45.5? (1)')  mean\_death\_rate = grouped\_data['Age-Adjusted Death Rate'].mean()  print(mean\_death\_rate)  sns.boxplot(x=df['Age-Adjusted Death Rate'])  plt.title('График смертности с поправкой на возраст')  plt.show()  import scipy.stats as stats  r, p = stats.pearsonr(df['Age-Adjusted Death Rate'], df['Average Deaths per Year'])  print(f'Коэффициент корреляции Пирсона: {r:.2f}')  print(f'Число p: {p:.2f}')  sns.pairplot(data=df, x\_vars=['Age-Adjusted Death Rate', 'Average Deaths per Year'], y\_vars=['Recent 5-Year Trend (2) in Death Rates', 'Recent Trend (2)'])  plt.title('Парный график коэффициента смертности с поправкой на возраст, среднего \nчисла смертей |