**АНАЛИЗ СМЕРТЕЙ В США В ПЕРИОД ЗА**

**2005 - 2015 ГОД**

**Курсовой проект**

по учебной дисциплине

«Наука о данных и аналитика больших объемов информации»

Направление подготовки магистров 09.04.04 «Программная инженерия»

|  |  |  |
| --- | --- | --- |
|  |  | Группа 3540904/10201 |
| Студенты: |  | **Е.С. Русаков** |
|  | **Изображение выглядит как текст  Автоматически созданное описание**Изображение выглядит как вычерчивание линий  Автоматически созданное описание | **Б.В. Репин** |
|  |  | **Ло Пиньяо** |
| Руководитель: |  |  |
| доцент, к. т. н. |  | **Никифоров И. В.** |

**Введение**

В данном отчёте излагается проделанная работа над датасетом из ресурса Kaggle [6], в котором представлена статистика по смертям в США за период с 2005 по 2015 года. В датасете были представлены обезличенные данные о погибших и умерших в промежуток с 2005 по 2015 год, включающие в себя подробные сведения о причинах наступления смерти, поле, возрасте, расе, этническом происхождении, образовании и т. п. умершего.

Над этим датасетом были проведены работы, обеспечивающие хранение в горизонтально-масштабируемом виде (NoSQL Data Base). Также проведён анализ данных с целью выявления некоторых зависимостей причин и времени наступления смерти с определенными обстоятельствами. Полученные данные представлены в графическом виде для более наглядного отображения выявленных зависимостей. После этого был собран дистрибутив, позволяющий повторить обработку данных и воссоздать графики и диаграммы, предстывленные в данном отчёте.

**Описание предметной области исследования**

Теоретическое значение выбранной темы:

Данная работа представляет большой интерес для различных государственных, частных и некоммерческих структур, нацеленных на улучшение качества и продолжительности жизни людей. Хоть данная работа основана только на данных о жителях США, это не умаляет её актуальности и для других стран первого и второго мира, поскольку общество и демография в них подчиняется схожим закономерностям. Примером тому является общемировая ранняя мужская смертность, большее, чем женских, число мужских суицидов, преобладание женщин в структуре населения, распространение деменции ввиду старения населения и сердечно-сосудистых заболеваний. В том числе, результаты данного исследования возможно экстраполировать на Россию и другие станы СНГ.

Практическое значение выбранной темы:

Эта работа может представлять большой интерес для органов здравоохранения многих стран. В ней не только показана возрастно-половая структура усопших, но и выявлены основные заболевания, приводящие к смерти граждан. Таким образом, органы здравоохранения могут оценить результаты своей деятельности и, если требуется, выявить текущие проблемы и сместить на них своё внимание.

Кроме того, представленный анализ может представлять интерес для различных социологических структур, некоммерческих и благотворительных организаций, занимающихся помощью уязвимым стратам населения и заинтересованных в социальном благополучие граждан.

Разработанность проблемы

Каждое государство, имеющее министерство здравоохранения, ЗАГС и статистическую службу располагает огромным массивом данных о здоровье и продолжительности граждан. Этот массив данных нещадно исследуется медиками, демографами и антропологами. Существует инфографика и исследования от Мирового Банка [1]. Проводятся многочисленные исследования в масштабах отдельных стран, регионов и городов. Например, исследование связи смертности с жарой и уровнем урбанизации в Огайо [2], сельско-городского разрыва уровня смертности в США за 1969-2009 года [3] или растущего с 1980-х годов разнообразия с типе смертности по всему миру [4].

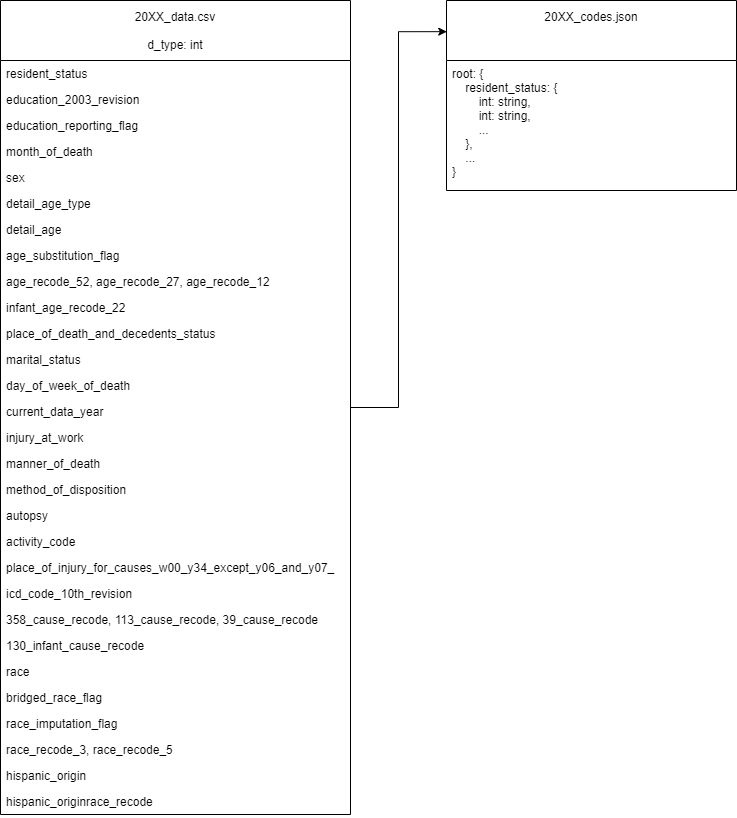
**Описание конкретных задач исследования**

В рамках данной работы планируется рассчитать и визуализировать следующие показатели и зависимости.

1. Соотношение мужской и женской смертности
2. Распределение смертности по возрастно-половому составу населения
3. Влияние уровня образования на смертность
4. Зависимость смертности от дня недели
5. Зависимость смертности от месяца
6. Распределение причин смертности в общей смертности
7. Посмертная судьба тела усопшего
8. Распределение способа смерти

**Описание модели данных**

Данные в датасете представлены в виде табличных CSV-файлов по одному на каждый год с 2005 по 2015. Объём распакованных данных составляет 4 ГБ. Колонки таблиц см. на схеме данных. Данные практически во всех колонках имеют численный тип. Это идентификаторы, описание которых для каждой колонки хранится в JSON-файле.

Рис. 1 Схема данных анализируемого датасета

**Используемые алгоритмы, техники и подходы анализа**

Была предпринята попытка переместить данные из формата CSV в NoSQL-БД. Процесс конвертации форматов оказался затратным по ОЗУ и потребовал много процессорного времени. Кроме того, объём полученных данных возрос до 50 ГБ. Такое громадное увеличение (в 12 раз) объясняется конвертацией каждой строки таблицы в отдельную JSON-запись, что приводило к многократному дублированию названий полей.

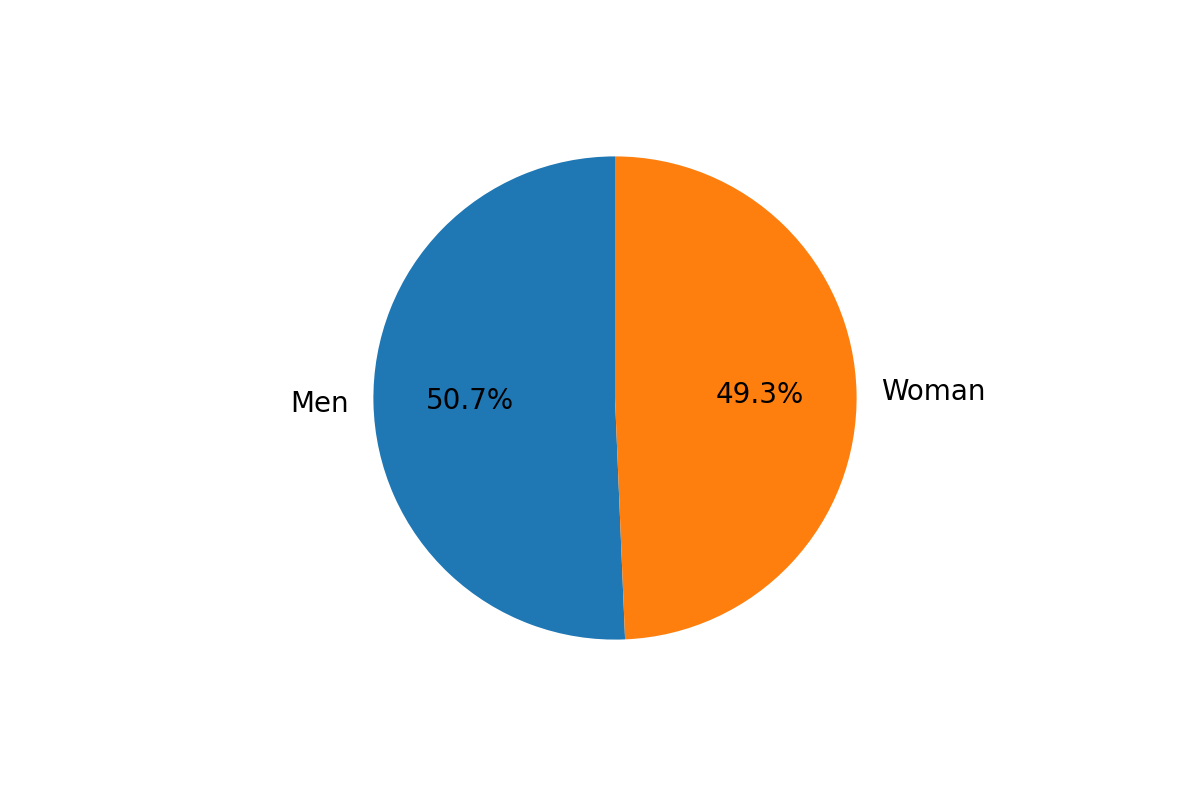
Ввиду объективных временны́х проблем, команда прибегла к линейным способам выборки и анализа данных датасета. В частности, использовался python-фреймворк pandas и библиотеки визуализации matplotlib и seaborn. Полученные графики, диаграммы и пояснения данных представлены в следующем разделе.

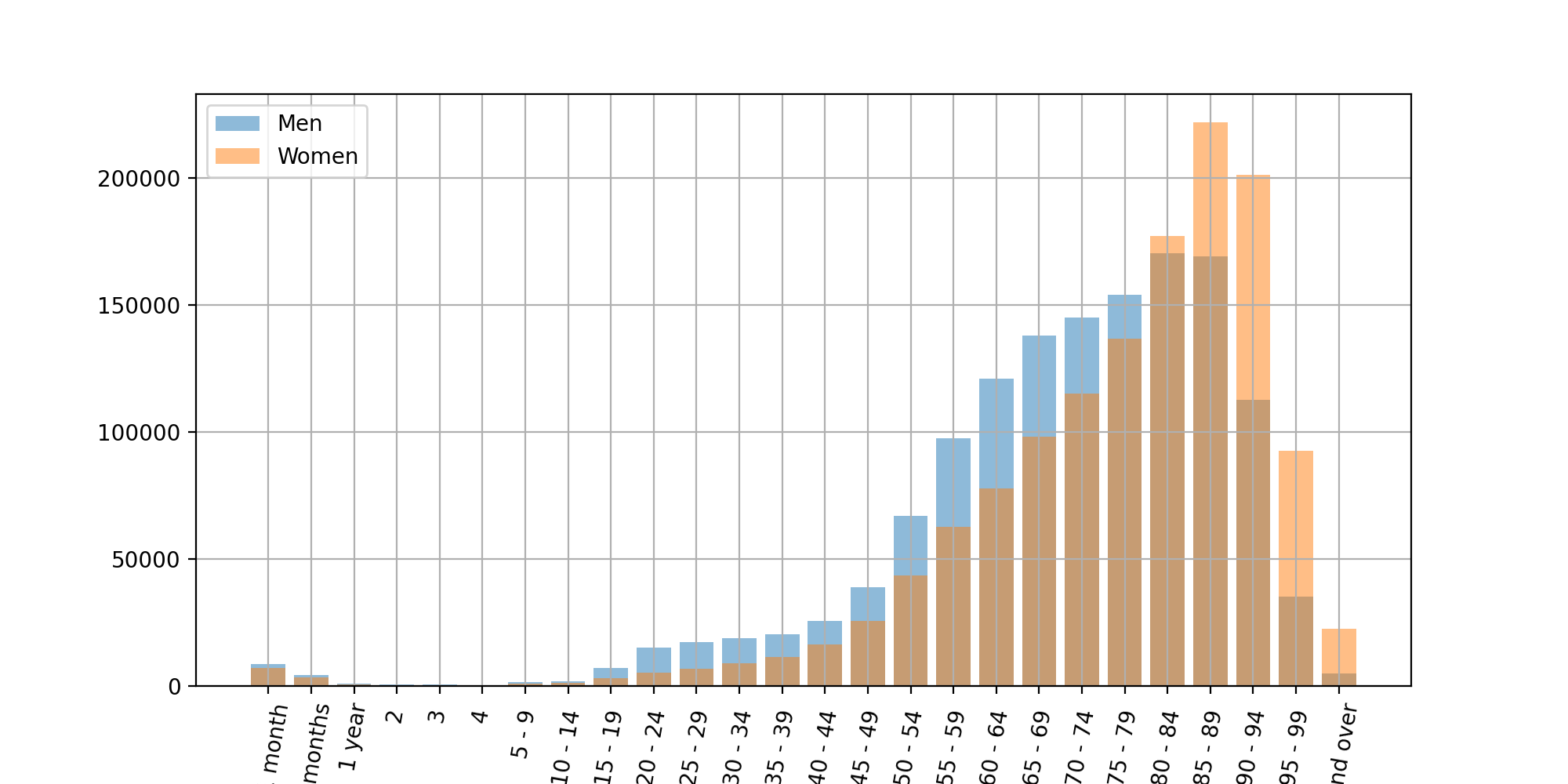
Предавшись фантазиям о бесконечных времени и мотивации, можно наделить этот проект серьёзной архитектурой. Из ЗАГСов поток JSONов с данными об усопших будет поступать в брокер сообщений (например Kafka), откуда может сохраняться в NoSQL-БД и (или) выгружаться в стек ELK, где можно будет почти что в реальном времени следить за процессом, строить визуализации и наблюдать за показателями. Но имеем, что есть.

**Графики и диаграммы**

1. Соотношение смертей мужчин и женщин.

Известно, что по природе мальчиков рождается слегка больше, чем девочек (примерно 105 на 100 девочек) [5]. Если принять, что оба пола умирают равномерно, то это объясняет небольшой перевес в сторону мужчин.

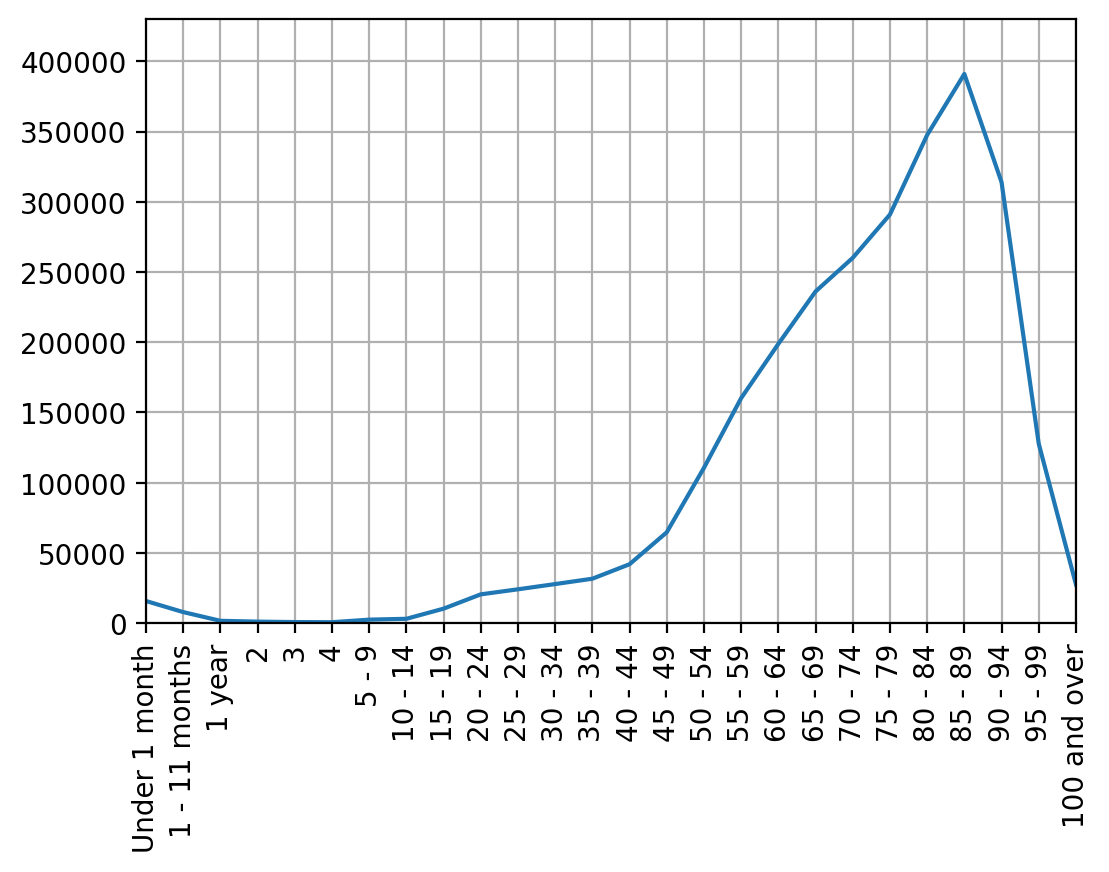


1. Смертность по половозрастному составу

В начале диаграммы видна младенческая и детская смертность, равная у обоих полов. Затем видно, что дети в возрасте от года до 15 лет погибают одинаково крайне редко. Однако в следующей страте (15 – 19 лет) смертность начинает расти, особенно быстро – у мужчин. Максимума разрыв мужской и женской смертности достигает в стратах 60 – 64 и 65 – 69 лет.

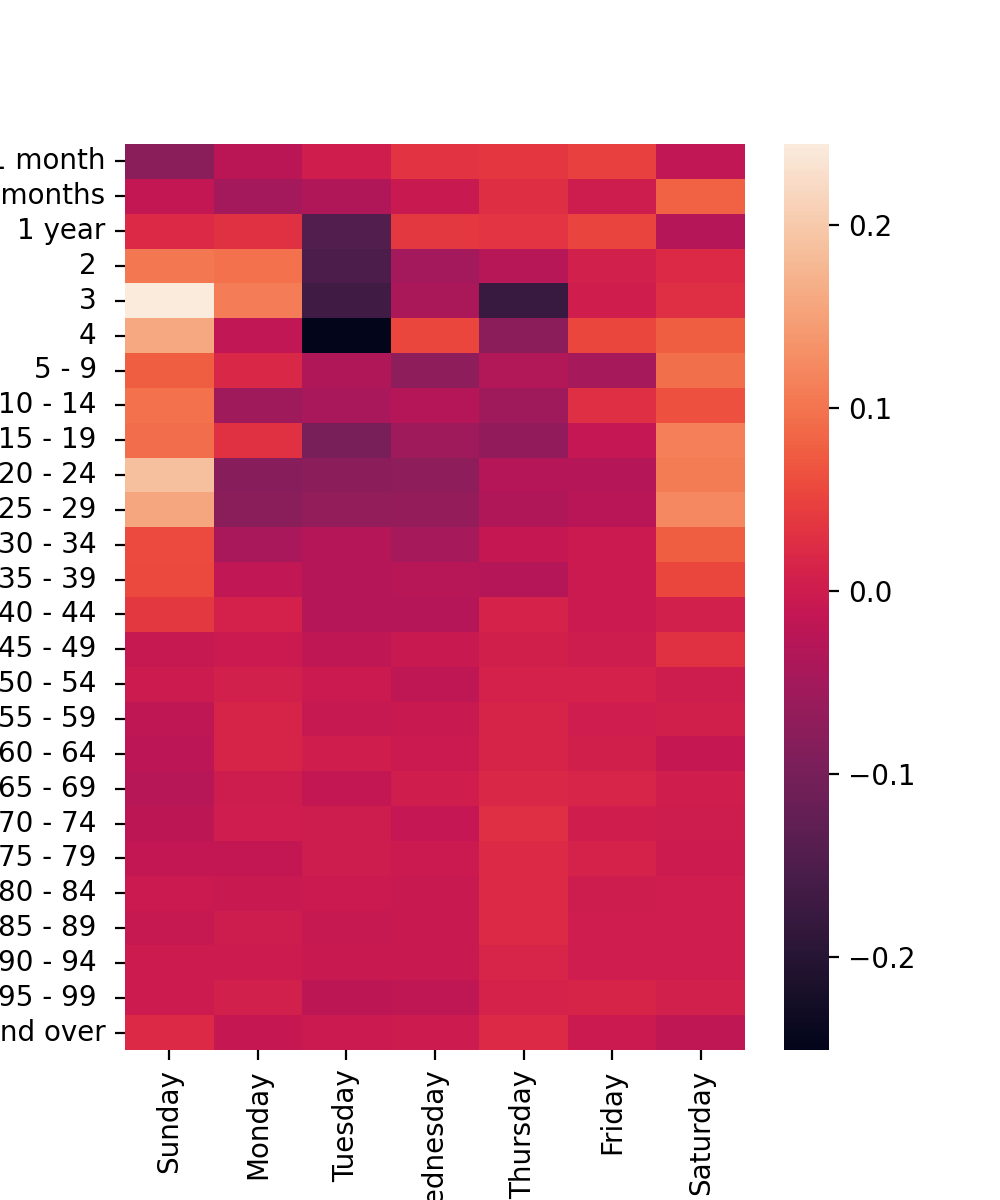
1. Смертность обоих полов по возрастным стратам

На этом графике так же явно видна смертность детей до года и огромная выживаемость 1 – 15 летних детей. Больше всего взрослых умирает в возрасте 85 – 89 лет.



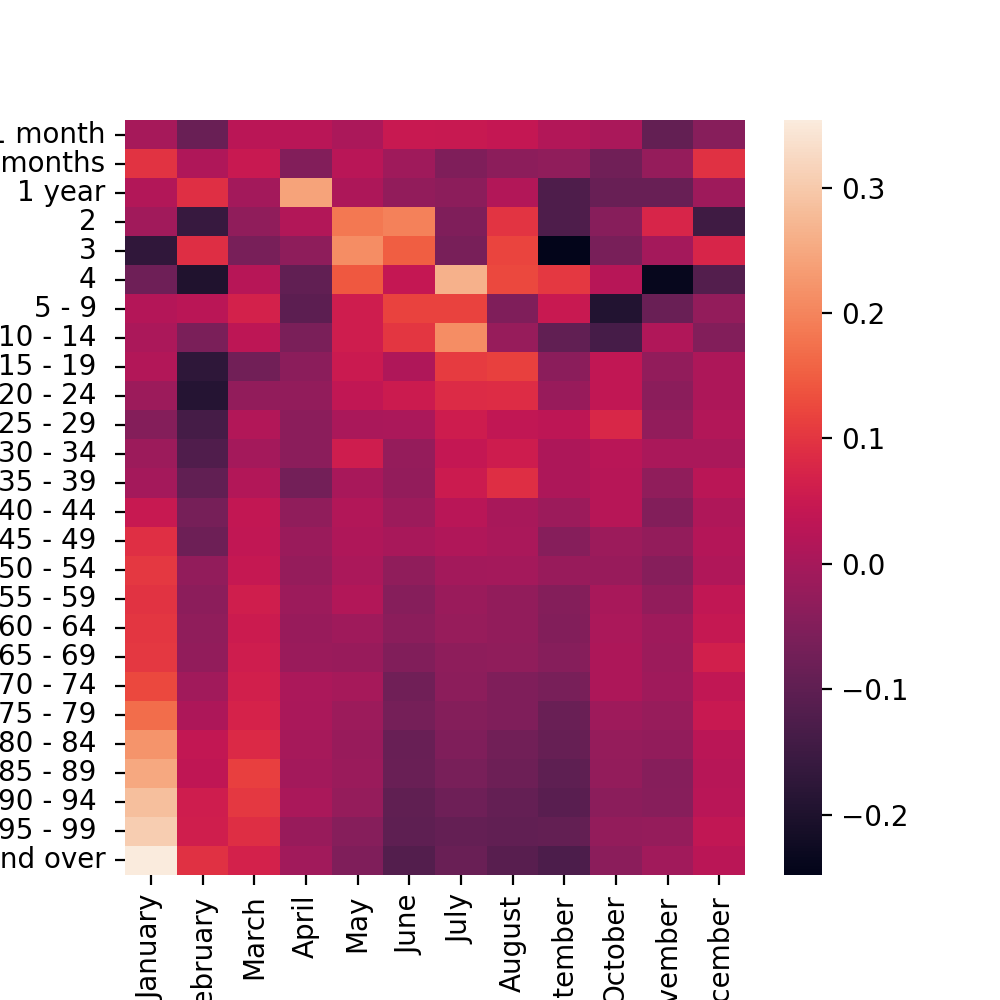
1. Зависимость смертности от дня недели

На этой диаграмме представлена карта отклонений фактической смертности в долях от статистически прогнозируемой (1/7 ≈ 0,14). Среди детей трёх и четырёх лет выделяются повышенной смертностью воскресенья, а пониженной – вторники. Объяснением этому может быть малое количество данных для этих страт (дети после года умирают очень редко). Также выделяются повышенной смертностью людей 20 – 40 лет субботы и воскресенья. Объяснение этого феномена – отдельная темя для исследования.



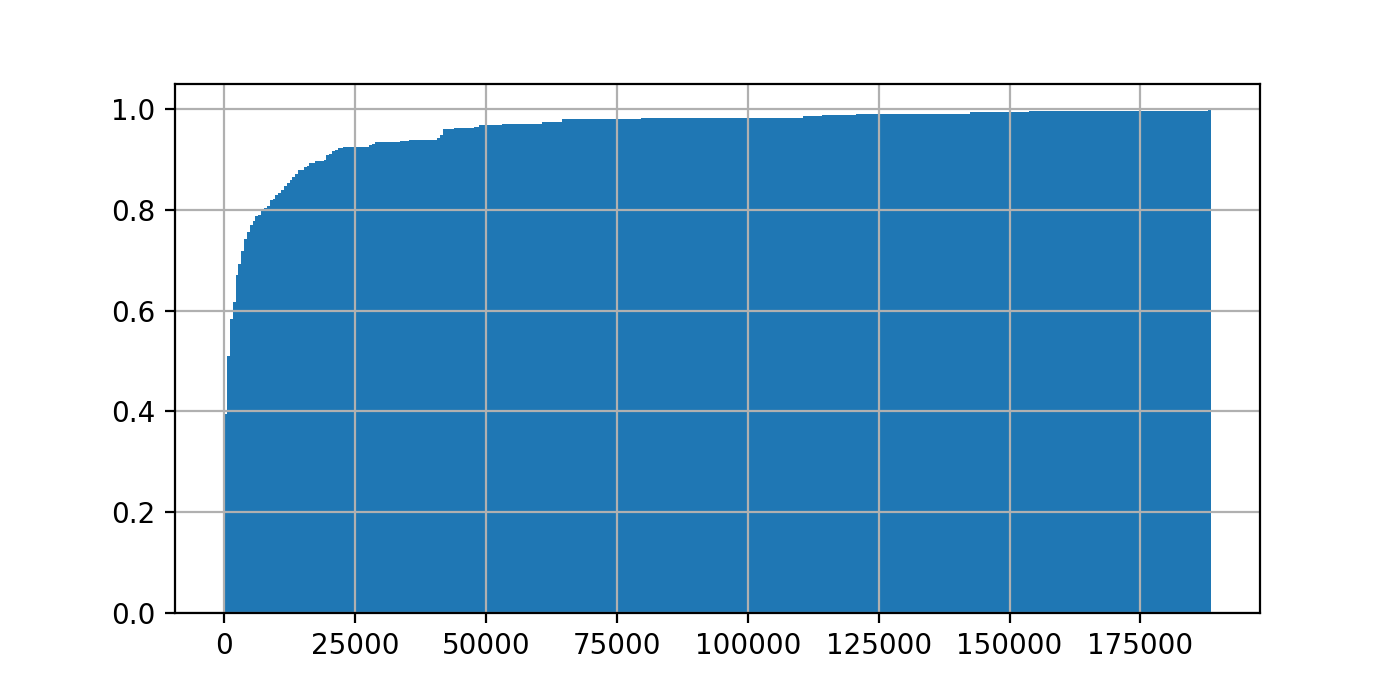
1. Зависимость смертности от месяца года

Здесь изображена аналогичная карта, но уже по месяцам. Статистически ожидаемая смертность – 1/12 ≈ 0,08. Сильные колебания среди детей объясняются аналогично – недостаточностью данных. Заметим также, что январь – очень вероятный месяц смерти пожилых людей, а в тёплое время года наоборот – смертность пожилых граждан снижается. Ещё примечательно выглядит снижение смертности молодых людей (15 – 30 лет) в феврале и наоборот – повышение в летние месяцы, особенно в августе.



1. Структура смертности

В этом пункте все данные сгруппированы по причине смерти и отсортированы по числу умерших. После этого построена накопительная гистограмма. Каждый новый столбик – добавление новой причины смерти в общую сумму. По вертикали отложена доля от общего числа.



По гистограмме хорошо видно, что большинство смертей приходится на малое число причин. Так, около 40% смертей вызваны всего десятью причин:

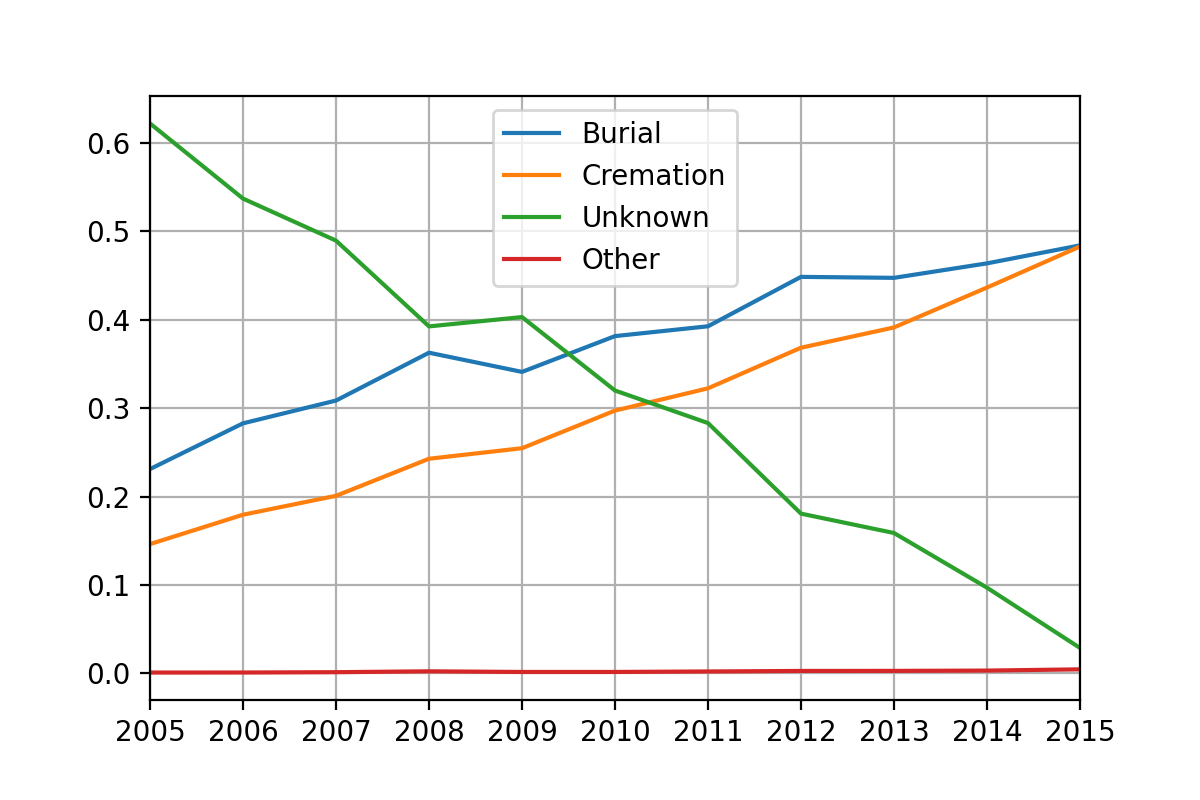
1. All other forms of chronic ischemic heart disease (иные формы ишемической болезни сердца)
2. Of trachea, bronchus and lung (заболевания трахеи, бронхов и лёгкий)
3. Other chronic obstructive pulmonary disease (иные хронические обструктивные болезни лёгких)
4. Organic dementia (органическая деменция)
5. Acute myocardial infarction (острый инфаркт миокарда)
6. Alzheimer's disease (болезнь Альцгеймера)
7. Diabetes mellitus (сахарный диабет)
8. Congestive heart failure (застойная сердечная недостаточность)
9. Stroke, not specified as hemorrhage or infarction (инсульт, не уточненный как кровоизлияние или инфаркт)

10)Atherosclerotic cardiovascular disease, so described (атеросклеротические сердечно-сосудистые заболевания, так описанные)

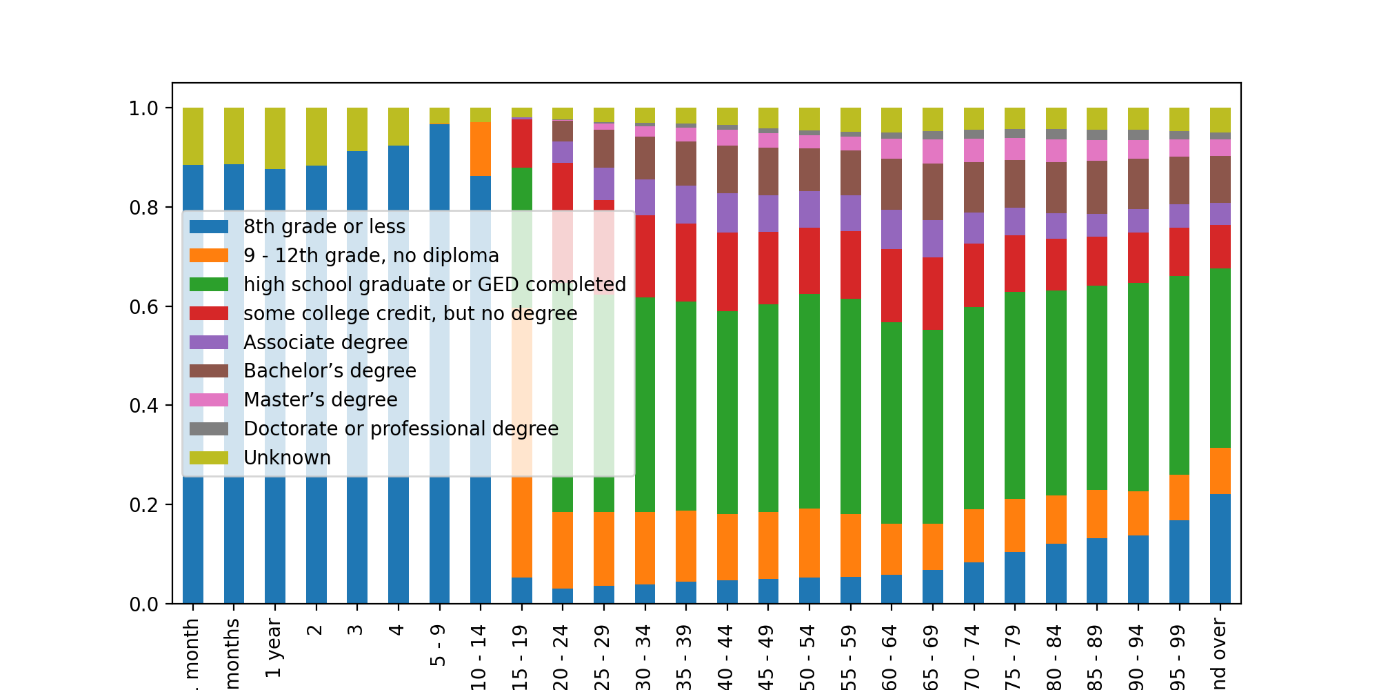
11)дайте знать если вы прочли это сообщение

1. Способ погребения

На этом графике представлена дальнейшая судьба тела усопшего. Удивляет высокая доля тел с неизвестной судьбой (зелёная линия). Также заметна тенденция к росту популярности кремации. К 2015 году кремировали тел больше, чем погребали.

****

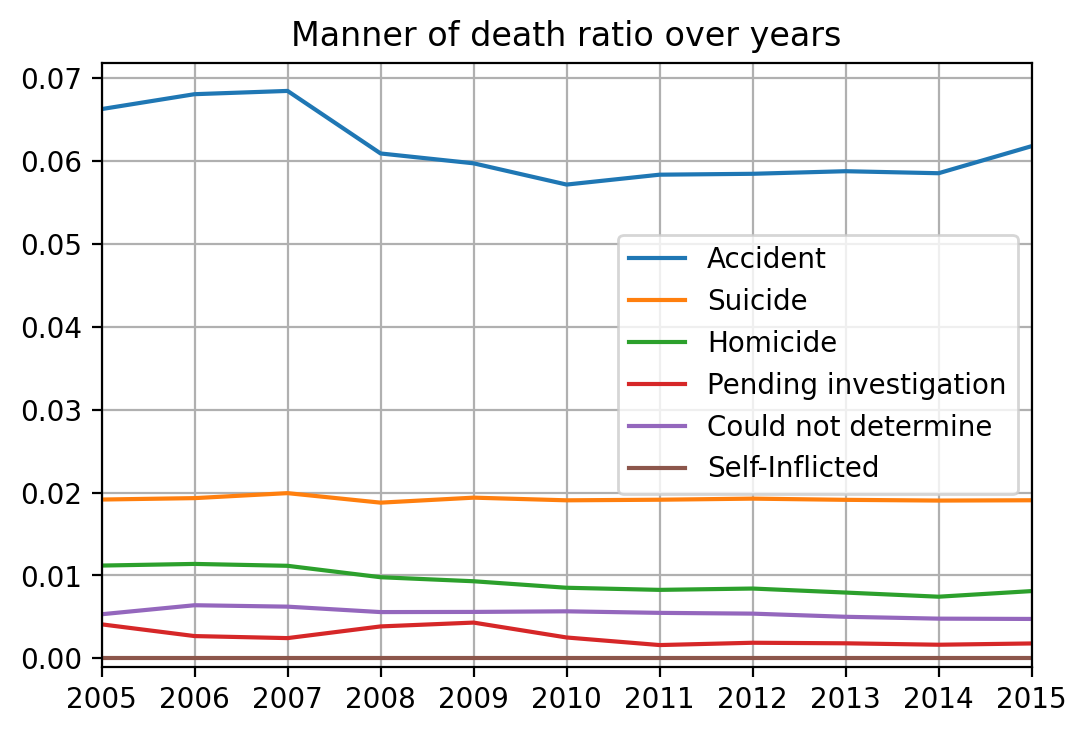
1. Зависимость возраста наступления смерти от образования

****

На этой диаграмме показаны доли усопших всех возрастных групп и их образование на момент смерти. Хорошо заметно, когда люди идут в старшую школу и поступаются в университет. Основной возраст получения образования всех видов – 15 – 25 лет. Видно, что высшее образование не влияет на продолжительность жизни. Даже наоборот: большой хвост пожилых умерших с восьмью и менее классами школы за плечами определяет их как очень живучую страту. В то же время, причины такого преобладания этой группы заслуживает отдельного исследования.

1. Распределение способа смерти

В данном пункте не рассматривается естественная смерть (это неинтересно). Рассматриваются все остальные причины: несчастные случаи, самоубийства, убийства, неопределимые или где расследование ещё не закончилось на момент формирования данных. Единственное изменение на графиках – падение доли несчастных случаев в 2008 году. Среди остальных причин изменений не наблюдается.



**Выводы**

Исходя из полученных аналитических данных можно прийти к следующим выводам:

1. Мужчины и женщины умирают почти одинаково часто, но мужчины немного чаще. Это можно объяснить тем, что мужчин статистически рождается чуть больше, чем женщин.
2. Женщины умирают реже мужчин и доживают до более старшего возраста.
3. Дети в возрасте от года до пятнадцати лет умирают крайне редко.
4. Пик смертности жителей США приходится в промежуток с 85 до 89 лет, что по общемировым меркам свидетельствуют о весьма развитой системе здравоохранения и благополучии граждан
5. Вероятность смерти почти никак не зависит от дня недели для большинства возрастных групп, но имеются небольшие аномалии у 20 – 40 летних граждан.
6. Вероятность смерти значительно зависит от месяца года.
7. Меньшинство причин вызывает большинство смертей.
8. Со временем кремация становится всё популярнее и уже обгоняет обычное захоронение по распространённости.
9. Образование почти никак не влияет на продолжительность жизни
10. После 2008 года кол-во смертей от несчастных случаев резко упало.

**Используемая литература**

1. The World Bank // URL: https://www.worldbank.org/en/home
2. Sheridan S.C, Dolney T.J. Heat, mortality, and level of urbanization: Measuring vulnerability across Ohio, USA. DOI: 10.3354/cr024255
3. Singh G.K., Siahpush M. Widening rural-urban disparities in all-cause mortality and mortality from major causes of death in the USA, 1969-2009. DOI: 10.1007/s11524-013-9847-2
4. Moser K., Shkolnikov V., Leon D.A. World mortality 1950-2000: Divergence replaces convergence from the late 1980s. ISSN: 00429686
5. Резник Н., Мальчики и девочки: соотношение полов и умозрительная генетика; «Химия и жизнь» №2, 2021, URL: https://elementy.ru/nauchno-populyarnaya\_biblioteka/435827/Malchiki\_i\_devochki\_sootnoshenie\_polov\_i\_umozritelnaya\_genetika
6. Death in the United States | Kaggle // URL: https://www.kaggle.com/cdc/mortality