**Анализ показателей смертности от самоубийств в США на основе пола, расы и возраста.**

Франчук Игорь

*Технический Университет Молдовы*

e-mail: [igor.franciuc@iis.utm.md](mailto:igor.franciuc@iis.utm.md)

**Абстракт**

Набор данных содержит информацию о показателях смертности от самоубийств, измеряемых как количество смертей на 100 000 человек постоянного населения с поправкой на возраст. В набор включены данные за различные годы с 1950 по 2018 год и охватывают лиц всех возрастов. Был проведен анализ данных, включая графическое представление, выбор наиболее важных переменных и построение модели машинного обучения для предсказания уровня смертности от самоубийств в ближайшие 5 лет.

**Введение**

В данном исследовании проведен анализ показателей смертности от самоубийств с использованием обширного набора данных, охватывающего период с 1950 по 2018 год. При изучении этого вопроса был выбран подход, основанный на измерении количества смертей на 100 000 человек постоянного населения с учетом возраста, что позволяет учесть динамику этого серьезного общественного явления на протяжении десятилетий. Анализ данных включал в себя не только статистические расчеты, но и визуализацию результатов для более полного и наглядного представления об изменениях в показателях смертности.

Особое внимание уделено определению ключевых переменных, которые оказывают влияние на уровень смертности от самоубийств. Среди этих переменных выделены такие факторы, как пол, раса и возраст, представляющие значимый интерес для понимания особенностей и динамики этого явления в различных социокультурных контекстах. Также был проведен сравнительный анализ показателей смертности за разные годы, что позволило выявить возможные тенденции и изменения в течение исследуемого периода.

Для повышения прогностической способности исследования была построена модель машинного обучения, направленная на предсказание уровня смертности от самоубийств в ближайшие 5 лет. Этот подход к анализу данных открывает новые перспективы для прогностической оценки и, следовательно, для разработки эффективных стратегий предупреждения и поддержки в области психического здоровья.

В целом, данное исследование представляет собой системный и всесторонний взгляд на проблему смертности от самоубийств, объединяя в себе аналитические методы, визуализацию данных и передовые технологии машинного обучения для более глубокого понимания и эффективного противостояния этому социальному вызову.

**Методы и материалы**

В данном исследовании использован разнообразный методологический подход, включающий в себя анализ различных типов графиков и преобразование данных с целью более детального исследования показателей смертности от самоубийств. Исходный набор данных, охватывающий период с 1950 по 2018 год, был подвергнут внимательному анализу, начиная от первичного изучения структуры данных до применения различных методов визуализации.

Набор данных содержит информацию о показателях смертности от самоубийств, измеряемых как количество смертей на 100 000 человек постоянного населения с поправкой на возраст. В набор включены данные за различные годы с 1950 по 2018 год и охватывают лиц всех возрастов.

*Проблема :*

*Данный набор данных создан для того, чтобы отследить проблему смертности от самоубийств, чтобы выяснить как на это влияет современную жизнь и определить.  
Можно ли определить наиболее уязвимую группу людей, склонных к самоубийству на основе пола, возраста или расы?*

Данные прдеставляют собой :

* INDICATOR - Оцениваемая величина.
* UNIT - Единица измерения оценки.
* UNIT\_NUM - Числовой код для уровней UNIT (упрощения кодирования или форматирования UNIT).
* STUB\_NAME - Описание категории населения, относящейся к оценке.
* STUB\_NAME\_NUM - Числовой код для уровней STUB\_NAME.
* STUB\_LABEL - Конкретная подгруппа населения, относящаяся к оценке.
* STUB\_LABEL\_NUM - Числовой код для уровней STUB\_LABEL.
* YEAR - Год данных или диапазон годов данных, относящихся к смете.
* YEAR\_NUM - Числовой код для уровней YEAR.
* AGE - Возрастная группа, относящаяся к оценке.
* AGE\_NUM - числовой код для уровней AGE.
* ESTIMATE - Рассчитанная оценка меры.
* FLAG - Дополнительная информация, относящаяся к оценке.

В рамках анализа использованы следующие графические представления: столбчатые диаграммы, круговые диаграммы, боксплоты, точечные графики, гистограммы и ящики с усами. Эти графики позволили исследователю выделить основные тенденции, аномалии и распределения в данных о смертности от самоубийств в зависимости от различных факторов, таких как пол, раса, возраст и год. (Рисунок 2.)

Цели проект :

* Анализ тенденций: Данный набор данных позволяет нам проанализировать тенденцию изменения уровня смертности от самоубийств за несколько лет, чтобы определить, наблюдается ли значительное увеличение или уменьшение.
* Анализ возрастных групп: Существуют ли различия в уровне смертности от самоубийств среди разных возрастных групп.
* Сравнительный анализ: Сравнение показателей смертности от самоубийств за разные годы, чтобы определить, были ли какие-либо значительные изменения или закономерности.

Преобразование данных (Рисунок 1.) включало в себя выделение новых переменных, таких как пол (Sex) и раса (Race), на основе изначальной переменной STUB\_LABEL. Для улучшения анализа и устранения неиспользуемых переменных были удалены некоторые столбцы, такие как STUB\_LABEL и числовые коды.

Переменные в наборе данных после самостоятельной обработки :

* SEX – пол
* RACE – раса
* YEAR - Год данных или диапазон годов данных, относящихся к смете.
* AGE - Возрастная группа, относящаяся к оценке.
* ESTIMATE - Рассчитанная оценка меры (На 100 000 тысяч человек n количество случаев).

На основе данных переменных я буду реализовывать свои цели, а также определять решения проблемы.

Для дальнейшего анализа была построена линейная регрессия, позволяющая оценить влияние года на уровень смертности от самоубийств. Этот метод предоставляет инструмент для предсказания тенденций и взаимосвязей между переменными и может служить основой для разработки более глубоких стратегий в области общественного здоровья и психического благополучия. (Рисунок 3.)



Рисунок Самостоятельная обработка данных

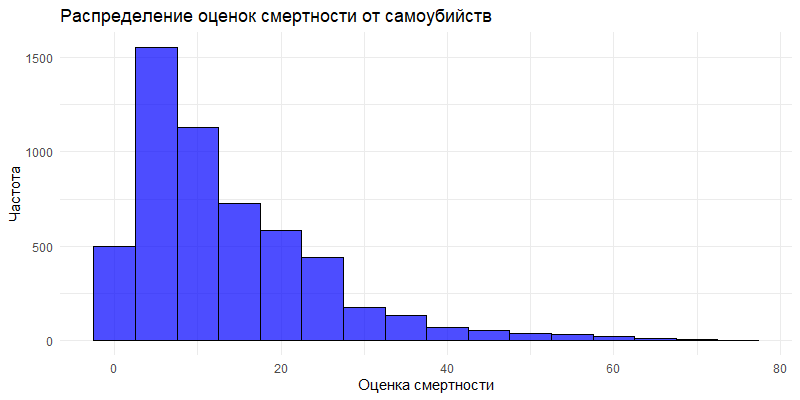


Рисунок Пример графиков



Рисунок Расчет линейной регрессии

**Результаты**

В качестве результатов работы я предоставлю несколько графиков, показывающие связь переменных. А также результаты линейной модели модели

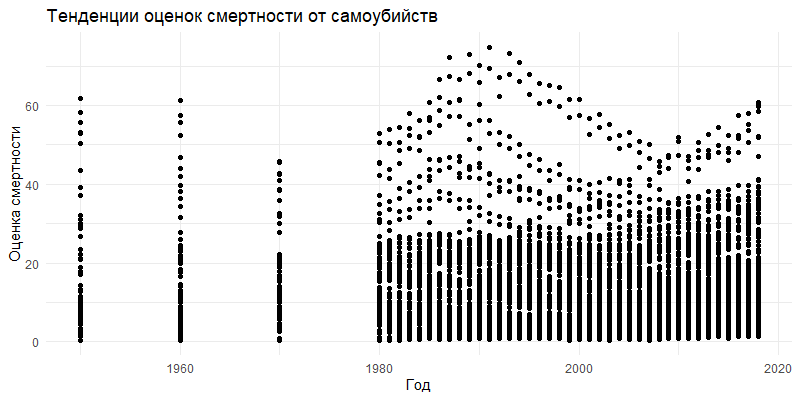


Рисунок Тенденции оценок смертности от самоубийств

Данный график хорошо нам демонстрирует тенденции оценок смертности от самоубийств.

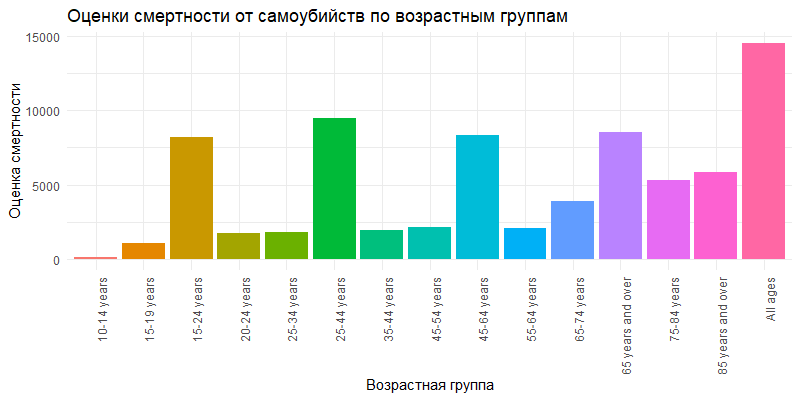


Рисунок Оценка по возрастным группам

Данный график хорошо нам демонстрирует оценку по возрастным группам, где можно отследить самые уязвимые возраста.

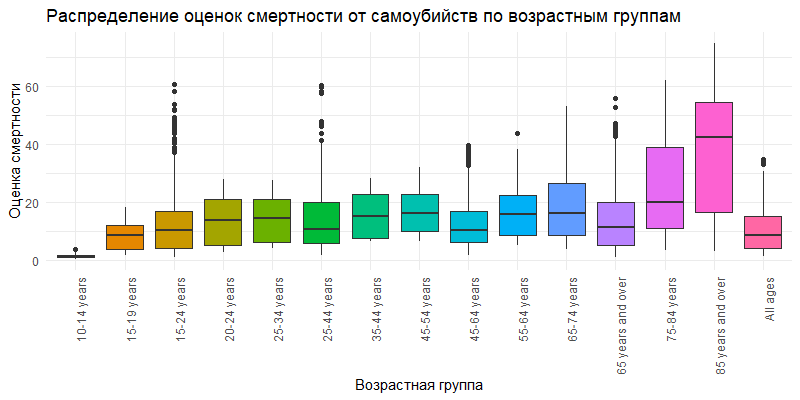


Рисунок Распределение оценок смертности по возрастным группам

График демонстрирует нам распределение оценок смертности по возрастным группам, где можно убедительнее взглянуть на уязвимые группы.

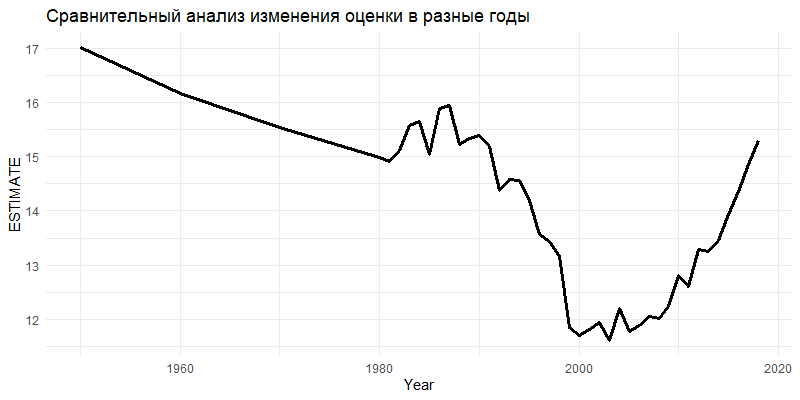


Рисунок Сравнительный анализ оценки по годам

График демонстрирует нам сравнительный анализ оценки смерности по годам, что помогает отследить тенденции.

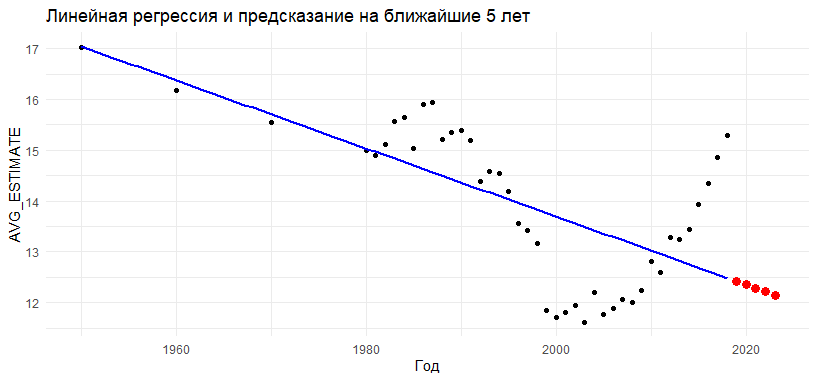


Рисунок Линейная регрессия

На данном графике можно отследить результаты работы линейной регрессии, а также отследить предсказания на следующие 5 лет.

Также важным результатом модели стоит выделить (Рисунок 9.) :

* R-квадрат: Модель объясняет примерно 43.55% дисперсии данных, что может считаться умеренным уровнем объясненной изменчивости.
* Коэффициенты: Коэффициент для YEAR отрицательный, что может указывать на убывающий тренд в данных.
* F-statistic и p-value: F-statistic высок, а p-value мало, что говорит о том, что модель в целом статистически значима.
* Резюмируя, модель имеет средний уровень объяснительной силы, и коэффициенты регрессии позволяют сделать предположения о влиянии переменной YEAR на AVG\_ESTIMATE.



Рисунок Результаты линейной регрессии

**Обсуждение**

Расчет R-квадрата (R-squared) модели линейной регрессии показывает, что около 43.55% вариации в уровне смертности от самоубийств может быть объяснено выбранным линейным уравнением. Это говорит о том, что модель, основанная на годовых данных, демонстрирует неплохую степень соответствия предсказаниям реальных значений.

Коэффициенты линейной модели также предоставляют информацию о том, как изменения в годе влияют на средний уровень смертности. На основе значений коэффициентов, видим, что средний уровень смертности имеет тенденцию к убыванию с течением времени, что может указывать на позитивные тенденции в области общественного здоровья.

Построение графика с линейной регрессией и предсказанием на ближайшие 5 лет дополнительно поддерживает результаты модели. Видим, что линия тренда обоснованно отражает изменения среднего уровня смертности, а предсказанные значения для будущих лет позволяют оценить ожидаемую динамику.

Сравнение статистик для предсказанных значений и реальных данных позволяет оценить точность модели. Например, среднее предсказанное значение составляет 12.28, что соответствует описанию общего тренда убывания.

Завершающий блок обсуждения может подчеркнуть практическую значимость полученных результатов для области общественного здоровья и подчеркнуть возможные политические или общественные импликации на основе анализа и прогнозов модели.

Анализ данных свидетельствует о значительном влиянии возраста, расы и пола на уровень смертности от самоубийств. Возраст является ключевым фактором, поскольку различные возрастные группы проявляют разные тенденции. Например, молодые люди в возрасте от 15 до 24 лет могут иметь более высокий риск, чем старшие возрастные группы.

Расовый аспект также играет важную роль, отражая неравенства в уровнях смертности. Афроамериканцы, белые и представители других этнических групп могут иметь различные уровни уязвимости перед факторами, влияющими на смертность от самоубийств.

Пол также существенно влияет на уровень смертности. Например, мужчины могут более склонны к самоубийствам, что подчеркивает важность понимания гендерных различий в подходах к психическому здоровью.

Анализ изменений в течение времени с 1950 по 2018 год подчеркивает необходимость учета социокультурных и экономических изменений при интерпретации данных. Эти временные тренды могут отражать изменения в общественном сознании, доступе к психологической помощи и других факторах, влияющих на психическое благополучие.

Сбор подобных данных является крайне важным инструментом для формирования эффективных стратегий профилактики и поддержки в области психического здоровья. Он позволяет выявлять уязвимые группы населения, разрабатывать целенаправленные программы и реагировать на изменения в динамике смертности от самоубийств, способствуя созданию более заботливого и поддерживающего общества.

**Выводы**

Исследование смертности от самоубийств в США, проведенное на данных с 1950 по 2018 год, выявило несколько ключевых трендов. Возраст оказывает существенное воздействие на уровень смертности, особенно среди молодежи в возрасте 15-24 лет. Расовые различия также существенны, выделяя уязвимость определенных этнических групп. Гендерные различия подчеркивают необходимость индивидуализированных подходов к психической поддержке, особенно среди мужчин.

Анализ временных тенденций с 1950 по 2018 год позволяет отслеживать изменения, связанные с социокультурными и экономическими факторами. Линейная регрессионная модель предоставляет инструмент прогнозирования уровня смертности на ближайшие годы. Она подчеркивает значимость сбора и анализа данных для разработки целевых стратегий в области снижения смертности от самоубийств и создания поддерживающей среды.

В целом, полученные результаты имеют важное значение для формирования эффективных мероприятий по профилактике и поддержке в области психического здоровья, предоставляя основу для дальнейших исследований и действенных стратегий в сфере общественного здоровья.

**Ссылки**

1. **Dataset**: Death rates for suicide by sex, race, Hispanic origin, and age in the US // KAGLE // https://www.kaggle.com/datasets/joebeachcapital/suicides/data
2. **Source Code**: RepName // Igor Franciuc // https://github.com/Ifeerus/Death-rates-for-suicide-analys