## 

## Цели и задачи аналитической работы

В современном мире проблема суицида является одной из основных причин смерти людей в мире. Причинами суицида может быть множество факторов, таких как:

1. Культурные особенности страны.
2. Проблемы со здоровьем.
3. Социально-экономические факторы.

Целью работы является анализ связи между количеством самоубийств в мире и уровнем счастья населения.

Для достижения данной цели были поставлены и выполнены следующие задачи:

1. Подбор и предобработка данных для последующего анализа.
2. Анализ полученных датасетов и формирование необходимых таблиц для статистического анализа.
3. Проведение статистического анализа данных.
4. Визуализация полученных результатов.

В ходе работы будут проанализированы данные по количеству случаев самоубийств, Happy Planet Index (HPI) и World Happiness Report (WHR). В результате анализа ожидается получить подтверждения или отклонение гипотез о том, что чем счастливее население, тем ниже уровень самоубийств в стране.

## Описание данных

В работе используются данные из трех различных источников.

1. Уровень самоубийств в мире по странам мира с 2015 по 2019 гг. от Института показателей оценки и здоровья (Institute for Health Metrics and Evaluation).

Данные содержат информацию о всех случаях самоубийства по 203 странам, в разрезе полов, 12 возрастных категории и количество смертей на 10 тыс человек.

Таблица содержит 15 столбцов.

| measure\_id | Код метрики  1 |
| --- | --- |
| measure\_name | Метрика  Deaths - смерть |
| location\_id | Код города |
| location\_name | Город |
| sex\_id | Код пола |
| sex\_name | Пол  Both  Female  Male |
| age\_id | Код возрастного диапазона |
| age\_name | Возрастной диапазон. Делится на 12 категорий  10-14 years  15-19 years  20-24 years  25-29 years  30-34 years  35-39 years  40-44 years  45-49 years  50-74 years  75-84 years  85+ years  All ages |
| cause\_id | Код причины  718 |
| cause\_name | Причина  Self-harm - Самоубийство |
| metric\_id | Код метрики  3 |
| metric\_name | Имя метрики  Rate - относительный показатель на 100 тыс чел. |
| year | Год |
| val | Значение  количество умерших человек на 100 тыс чел. |
| upper  lower | 95% интервал неопределенности -  Верхняя граница  Оценка процентиля 97,5%.  нижний 95% интервал неопределенности -  Нижняя граница  2,5% оценка процентиля. |

В таблице содержится 39720 значений, пропуски отсутствуют. (Приложение 1)

1. Данные индекса Happy planet Index с 2015 по 2019 гг.

Таблица данных содержит 13 столбцов

| HPI rank | Номер рейтинга |
| --- | --- |
| Country | Страна |
| ISO | Код страны ISO |
| Continent | Номер континента |
| Unnamed: 4 | Объединенный столбец кода страны и года |
| Population (thousands) | Численность |
| Year | Год |
| LifeExp (years) | Продолжительность жизни |
| Ladder of life (0-10) | Качество жизни, оценка от 1-10 |
| Footprint (g ha) | Экологический след на глобальный гектар. Страна с большим экологическим следом на душу населения использует ресурсы, превышающие ее справедливую долю, как за счет ресурсов других стран, так и за счет нанесения непоправимого ущерба планете, который затронет будущие поколения. |
| HPI | Индекс, показывает итоговую оценку по всем показателям влияющим на индекс |
| Biocapacity for year (g ha) | Биоемкость (биоразнообразие на га) |
| GDP per capita ($) | Уровень ВВП на душу населения |

В HPI имеются пропуски данных, максимальное количество данных 760, минимальное 725 по оценке индекса, т.о. у нас оказалось всего 725 оценок за 5 летний период. (приложение 2)

Из приложения 2 видно, что не все страны получили участвовали в индексе.

Об этом говорят пропуски в HPI,Ladder of life, Footprint и GDP per capita ($).

Авторы индекса при его составлении не используют в расчете уровень ВВП, это видно и по визуализации пропусков. Т.к. есть страны без значения в столбце GDP per capita ($), но при этом они получили оценку в РЗШ. Это говорит о том, что эта информация носит информационный характер.

При этом качество жизни и экологических след являются основным элементов расчета индекса счастья.

14 стран у нас имеют от одного пропуска в анализируемый период.

1. Данные отчета ООН - World Happiness Report с 2015 по 2019 гг.

| Country | Страна |
| --- | --- |
| Happiness score | Общая оценка уровня счастья |
| GDP per capita | Уровень ВВП на душу населения |
| Social support | Социальная поддержка |
| Healthy life expectancy | Продолжительность жизни |
| Freedom to make life choices | Свобода выбора |
| Generosity | Щедрость |
| Perceptions of corruption | Уровень коррупции |
| Dystopia Residual | Антиутопия — это воображаемая страна, в которой проживают наименее счастливые люди в мире. Цель создания антиутопии состоит в том, чтобы иметь контрольный показатель, с которым можно было бы выгодно сравнивать все страны (ни одна страна не работает хуже, чем антиутопия) Таким образом, самые низкие оценки по шести ключевым переменным характеризуют антиутопию. Поскольку жизнь была бы очень неприятной в стране с самыми низкими в мире доходами, самой низкой продолжительностью жизни, самой низкой щедростью, самой большой коррупцией, наименьшей свободой и наименьшей социальной поддержкой, ее называют «антиутопией», в отличие от утопии. |
| Year | Год |

WHR не имеет пропуски данных, все 782 строки имеют значения по всем столбца. (приложение 3)

## 

## Анализ данных

В работе использовались следующие библиотеки:

1. pandas потребуется для работы с данными в табличном представлении.
2. numpy потребуется для округления данных в корреляционной матрице.
3. matplotlib, seaborn, plotly.express и pycountry потребуется для визуализации данных.
4. sklearn потребуется для тестирования гипотез.

Для каждого датасета, участвующего в работе, был составлен список стран, написание которых отличалось от написания в библиотеке pycountry. Это было сделано для того, чтобы в последующем создать тепловую карту мира, а также упростить объединение таблиц для статистического анализа.

Данные по количеству самоубийств были проанализированы в динамике за 5 лет, с 2015-2019 гг.. Также было проведен анализ в разрезе в возрастных групп и пола.

Построена тепловая географическая карта по средним значениям за анализируемый период и в соответствии с полом людей, так же горизонтальные диаграммы по странам с 10 наибольшим и наименьшим количеством самоубийств.

Индексы счастья рассчитываются на основе из показателей, для каждого индекса была построена диаграмма корреляции, ящик с усами и тепловая карта. Это позволяет визуализировать корреляцию показателей в индексе.

Данные по HPI имеют пропуски, в работе они могут дать искажения, поэтому было принято решений их удалить. Также для снижения искажений при анализа составлен список стран, где был хоть один пропуск в оценке индекса за анализируемый период. Эти страны также были исключены из последующего анализа. Аналогично с данными по самоубийствами епловая географическая карта по индексу счастья, на ней видны страны которые не попали в анализ. И были выявлены 10 самых счастливых и несчастных стран.

WHR состоял из 5 разных файлов. В процессе работы они были объединены в одну таблицу, столбцы были приведены к одному названию, добавлен столбец с годом. Проверяем на наличие стран , которые не всегда участвовали в оценке в анализируемом периоде. Аналогично HPI страны с пропусками были убраны из дальнейшего анализа.

Для статистического анализа были выбраны разные периоды. Уровень самоубийств был взят за 2017 год, а индексы счастья за 2016 г. Это обусловлено тем, что уровень счастья за предыдущий год может повлиять на решение о самоубийстве в текущем, т.к. негативные события имеют отложенное влияние на человека.

2016 год был выбран потому, что именно в это год наблюдается резкий спад уровня счастья по world happiness report.

## Статистический анализ и проверка гипотез

Для проведения статистического анализа были выбраны два метода:

1. Корреляция Пирсона
2. Парные линейные регрессии

На этапе подготовки данных для статистического анализа было выявлено, что с каждым годом количество самоубийств снижается, а индексы счастья в мире увеличиваются. На основании этого можно предположить, что эти два показателя взаимосвязаны. Для того, чтобы принять или отклонить эту гипотезу была использована корреляция Пирсона.

Перед проверкой гипотезы о наличии корреляции между уровнем счастья и количеством самоубийств, проверим корреляцию индексов счастья между собой.

Нулевая гипотеза(H0) предполагает отсутствие корреляции.

В результате расчетов коэффициент корреляции Пирсона равен 0.520922 говорит о высокой положительной корреляции

На основании полученных значений p-value: 2.4737072615282194 есть основания отклонить H0. Зависимость есть.

Т.к. H0 была отклонена, то проверим наличие или отсутствие корреляции между количеством самоубийств и средним индексов счастья.

Для начала был построен график корреляции:



По визуализации корреляции между количеством суицидов средним индексов счастья, нельзя сказать, что чем счастливее население, тем ниже смертность. У нас есть несколько выбросов по смертности, при этом оценки счастья находятся на расстоянии друг от друга.

При оценке счастья до 3,5 баллов наблюдается низкое количество самоубийств.

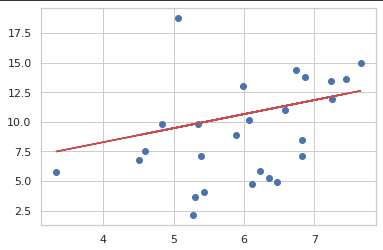
На основании визуализации была принята нулевая гипотеза: корреляция полностью отсутствует.

По результатам расчета корреляции коэффициент корреляции Пирсона составляет 0.07783801662081977 - это говорит о низкой положительной корреляции. А на основании полученного значений p-value ( 0.38059736132471456.) нет оснований отклонить H0. Не достаточно оснований утверждать, что по уровню счастья можно предсказать количество случаев самоубийств.

Поскольку у нас высокий уровень корреляции между индексами счастья, пробуем посмотреть зависимость HPI на основе WHR через модель линейной регрессии. По результатам построения модели линейной регрессии наблюдаем прямую зависимость оценки индекса Наppy Planet Index и оценки по индексу World Happiness Report, c высоким уровнем зависимости, т.к. коэффициент детерминации равен 0.1414.

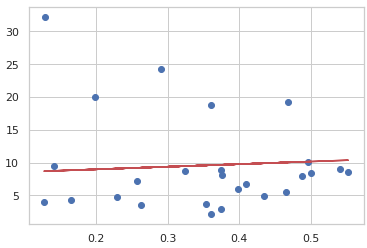
Коэффициент Пирсона показал низкую зависимость между счастьем и количеством самоубийств. Возможно, отдельные показатели каждого индекса могут оказать влияние на уровень самоубийств. Для этого возьмем из каждого индекса наиболее субъективный показатель, тот которые определяется на основе опросов населения: для World Happiness Report - это “Freedom to make life choices”, либо на основе слов авторов индекса : для Happy Planet Index - Ladder of life (0-10)”.

По результатам построения модели линейной регрессии между Ladder of life (0-10) (в качестве независимой переменной) и количеством самоубийств наблюдаем прямую зависимость при этом эта зависимость высока, т.к. коэффициент детерминации равен 0.016.



Показатель Freedom to make life choices показывает насколько свободен человек свободен в своем выборе в жизни, можно предположить о том, что этот показатель говорит о политическом режиме, либо религиозной культуре.

При построении модели между Freedom to make life choices и количеством самоубийств, мы также видим прямую зависимость, но с низким уровнем значимости, т.к. коэффициент детерминации отрицательный и равен -0.032.



## Шаги по улучшению работы

В данной работе осталось несколько неохваченных моментов, которые можно рассматривать дополнительно.

В индексах счастья есть страны которые были исключены из анализа, т.к. не всегда участвовали в оценках уровня счастья за период с 2015 по 2019 гг. Их можно рассмотреть отдельно и изучить их корреляцию с самоубийствами.

Также можно проверить как влияет количество самоубийств на экономические показатели по странам. Добавить анализ по географическим регионам. Дополнительно проанализировать географию самоубийств по возрастам.

Для регрессионного анализа возможно менять периоды. Например: сдвигать уровень счастья на год, т.е. проверить корреляцию с самоубийствами год в год, либо на год +1. В работе применялся только парный регрессионный анализ, его можно дополнить множественной регрессией. Также остались различные показатели индексов, которые можно использовать для поиска связи.

## Выводы

Проведенный анализ показал, что с каждым годом относительный уровень самоубийств снижается. Это может говорить как об общей тенденции к снижению, так и о том, что увеличивается количество населения в мире и это влияет на расчет.

Самой большой страной по количеству самоубийств является Гренландия, наименьшее число самоубийств зафиксировано в Сан-Томе и Принсипи. СТраны Южной Америки и Африки имеют низкий уровень суицидов среди населения. Мужчины чаще женщин заканчивают жизнь самоубийством, люди старше 75 лет также больше заканчивают жизнь самоубийством - это может быть связано с состоянием здоровья.

География смертей между мужчинами и женщинами различна, если мужская, в основном, соответствует тем же странам, что и без разделения по половым признакам, то женская отличается. Неизменным остается только Гренландия.

Happy Planet Index имеет тенденцию к росту с каждым годом. Показатели индекса хорошо коррелируют между собой, за исключением Биоразнообразия. По данному индексу, наиболее счастливым странами являются страны Южной Америки.

World Happiness Report так же имеет тенденцию к росту, за исключением падения уровня счастья в 2016 году. WHR, в отличии от HPI, считает наиболее счастливыми скандинавские страны и страны в Северной Америке, самыми несчастными являются страны Африки.

Визуально видна корреляция между самоубийствами и счастьем, за исключением стран Африки, низкий индекс счастья не равен высокому уровню самоубийств.

При проверке гипотезы о корреляции счастья и самоубийств, выяснилось, что корреляция низка и мы не можем утверждать, что по уровню счастья можно предсказать количество самоубийств.

При изучении связи самоубийств и уровня жизни населения выявлена прямая зависимость, т.о можно сказать, что чем выше уровень жизни, тем выше уровень самоубийств. Изучение свободы выбора человека и самоубийств наоборот показал, что корреляция низка и нельзя сделать вывод о том, что политический режим, религиозная культура могут повлиять на решение человека покончить жизнь самоубийством.

Таким образом цель работы достигнута, связь уровня самоубийств с различными изучена, а поставленные задачи были выполнены.

Приложение 1. Описание данные для отчета по количеству самоубийств

| measure\_id | location\_id | sex\_id | age\_id | cause\_id | metric\_id | year | val | upper | lower |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| count | 36720 | 36720 | 36720 | 36720 | 36720 | 36720 | 36720 | 36720 | 36720 |
| mean | 1 | 136.299 | 2 | 45.83333 | 718 | 3 | 2017 | 15.618172 | 21.541222 |
| std | 0 | 97.92294 | 0.81651 | 72.18689 | 0 | 0 | 1.41423 | 20.980677 | 29.837299 |
| min | 1 | 6 | 1 | 7 | 718 | 3 | 2015 | 0.02427 | 0.04658 |
| 25% | 1 | 62.75 | 1 | 9.75 | 718 | 3 | 2016 | 3.990835 | 5.696688 |
| 50% | 1 | 122.5 | 2 | 12.5 | 718 | 3 | 2017 | 8.840984 | 12.276364 |
| 75% | 1 | 182.25 | 3 | 26.75 | 718 | 3 | 2018 | 18.286434 | 24.424073 |
| max | 1 | 522 | 3 | 243 | 718 | 3 | 2019 | 246.2682 | 440.7315 |

Приложение 2. Описание данных HPI

|  | Population (thousands) | Year | LifeExp (years) | Ladder of life (0-10) | Footprint (g ha) | HPI | Biocapacity for year (g ha) | GDP per capita ($) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| count | 7.600000e+02 | 760.000000 | 760.000000 | 735.000000 | 750.000000 | 725.000000 | 760.000000 | 744.000000 |
| mean | 4.867277e+04 | 2017.000000 | 72.586919 | 5.471724 | 3.266578 | 44.231272 | 1.588000 | 21351.736380 |
| std | 1.616391e+05 | 1.415145 | 7.838368 | 1.122965 | 2.430140 | 7.832556 | 0.019403 | 21413.919487 |
| min | 2.711280e+02 | 2015.000000 | 50.900000 | 2.375092 | 0.515749 | 22.348251 | 1.560000 | 751.663957 |
| 25% | 4.818905e+03 | 2016.000000 | 66.775000 | 4.630803 | 1.481374 | 38.579845 | 1.570000 | 4792.287556 |
| 50% | 1.107750e+04 | 2017.000000 | 74.450000 | 5.481311 | 2.458404 | 44.453013 | 1.600000 | 12914.175155 |
| 75% | 3.640515e+04 | 2018.000000 | 78.300000 | 6.264960 | 4.469769 | 49.734753 | 1.600000 | 31537.834882 |
| max | 1.433784e+06 | 2019.000000 | 84.900000 | 7.858107 | 15.037589 | 64.166492 | 1.610000 | 114304.044200 |

Приложение 3. Описание данные WHR

|  | **Happiness score** | **GDP per capita** | **Social support** | **Healthy life expectancy** | **Freedom to make life choices** | **Generosity** | **Perceptions of corruption** | **Dystopia Residual** | **Year** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 782.000000 | 782.000000 | 782.000000 | 782.000000 | 782.000000 | 782.000000 | 782.000000 | 782.000000 | 782.000000 |
| **mean** | 5.379010 | 0.915243 | 1.079052 | 0.612551 | 0.411579 | 0.218688 | 0.125508 | 2.016385 | 2016.993606 |
| **std** | 1.127457 | 0.405628 | 0.329457 | 0.248362 | 0.153093 | 0.122347 | 0.105765 | 0.555952 | 1.417364 |
| **min** | 2.693000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.183000 | 2015.000000 |
| **25%** | 4.509750 | 0.606500 | 0.870406 | 0.440181 | 0.310499 | 0.130000 | 0.054249 | 1.680337 | 2016.000000 |
| **50%** | 5.322000 | 0.982000 | 1.125000 | 0.647281 | 0.431330 | 0.202000 | 0.091000 | 2.022000 | 2017.000000 |
| **75%** | 6.189500 | 1.236250 | 1.327250 | 0.808000 | 0.531000 | 0.279705 | 0.156023 | 2.369250 | 2018.000000 |
| **max** | 7.769000 | 1.871000 | 1.644000 | 1.141000 | 0.724000 | 0.838000 | 0.551910 | 3.837715 | 2019.000000 |