Постановка первоначальных аналитических гипотез с использованием визуализации данных на примере изучения влияния социально-экономических факторов на состояние психического здоровья населения

Рабинович А.Е., к.э.н., доцент
Московский Политехнический университет, Москва, Россия
Седова Е.А., магистр направления «Системная аналитика больших данных»
Московский Политехнический университет, Москва, Россия

Аннотация. Данная статья посвящена проблеме выявления и постановки основных первоначальных гипотез на ранних стадиях проведения анализа больших данных по проблеме влияния социально-экономических факторов на состояние психического здоровья населения. Постановка гипотез производилась при помощи визуализации данных с использованием языка программирования Python. В результате выполнения работы, в процессе обработки и визуализации имеющихся данных, были сформулированы основные гипотезы, на которые можно будет опираться при построении дальнейшего анализа.

Ключевые слова: визуализация данных, Python, психическое здоровье, Big Data, SkikitLearn

Statement of initial analytical hypotheses using data visualization on the example of studying the influence of socio-economic factors on the state of mental health of the population

Rabinovich A.E., Ph. D., Associate Professor
Moscow Polytechnic University, Moscow, Russia
Sedova E.A., Master's Degree in System Analytics of Big Data,
Moscow Polytechnic University, Moscow, Russia

Annotation. This article is devoted to the problem of identifying and formulating the main initial hypotheses at the early stages of the analysis of Big Data on the topic of the influence of socio-economic factors on the state of mental health of the population. Hypotheses were formulated using data visualization using the Python programming language. As a result of the work performed in the process of processing and visualizing the available data, the main hypotheses were identified, which can be relied on when constructing further analysis.

Keywords: data visualization, Python, mental health, Big Data, SkikitLearn

В рамках изучения проблемы влияния различных факторов на здоровье населения было выявлено, что существует зависимость между различными социально-экономическими факторами и состоянием психического здоровья граждан [1]. Одним из показателей, по которому можно судить о состоянии психического здоровья населения страны, является статистика самоубийств. По статистике, ежедневно в мире добровольно уходят из жизни около 3-х тыс. человек, ежегодно – около 1 млн. человек [2]. Это является серьезной причиной для изучения влияния различных факторов, в том числе и социально-экономических, на состояние психического здоровья различных групп людей в разные промежутки времени.

Разрозненность и неоднородность больших данных не позволяет выявить главные зависимости и сформулировать первоначальные аналитические гипотезы, не прибегая к инструментам анализа данных. Первоначальным этапом анализа может стать визуализация данных при помощи современных средств анализа, например языка программирования Python. Язык Python обладает большим количеством готовых математических библиотек, позволяющих аналитикам в сфере Big Data различными способами визуализировать и анализировать большие данные.

Основной целью статьи является первоначальная визуализация данных при помощи средств Python для выявления общих закономерностей и постановки первоначальных гипотез, а также сопоставление полученных результатов с

имеющимися исследованиями по данной тематике. Полученные в результате выполнения работы гипотезы будут использованы в дальнейшем при построении прогнозной модели, позволяющей прогнозировать статистику самоубийств при различных параметрах. При выполнении работы использовалась программная среда Python под названием Jupyter Notebook. Основной источник данных – статистика самоубийств в разных странах с 1960 по 2017 годы [2].

Перед началом анализа необходимо понять, в каком виде представлены данные, с которыми предстоит работать. Для этого необходимо импортировать набор данных в программу и вывести основную информацию. Для того, чтобы произвести первоначальный импорт файла, необходимо подключить библиотеку Pandas, которая обладает широким функционалом, предназначенным для первичной обработки данных для последующего анализа. Основной результат представлен на рис. 1.

```
In [6]: import pandas as pd
        dataset = pd.read_csv('C:/Users/vzmgi/Desktop/suicides.csv')

In [7]: dataset.head()
Out[7]:
```

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	796	Generation X
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	796	Silent
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	796	Generation X
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	796	G.I. Generation
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	796	Boomers

Рис. 1 – Импорт и вывод файла с набором данных

В наборе данных представлена статистика по самоубийствам со следующими атрибутами: страна, год, пол, возрастная группа, количество самоубийств, популяция, количество самоубийств на 100 тысяч населения, годовой ВВП, ВВП на душу населения и принадлежность к поколению. Исходя из этих данных можно построить первичные графики для того, чтобы выявить основные зависимости.

Для начала необходимо построить общий график количества самоубийств по годам для того, чтобы посмотреть в какое время наблюдаются максимальные и минимальные значения рассматриваемого признака. Для этого необходимо

сгруппировать значения количества самоубийств по годам и вывести полученный результат. Результат представлен на рис. 2.

```
In [5]: dataset.groupby(['year']).count()['suicides_no'].plot()
Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x24ac75949c8>
```

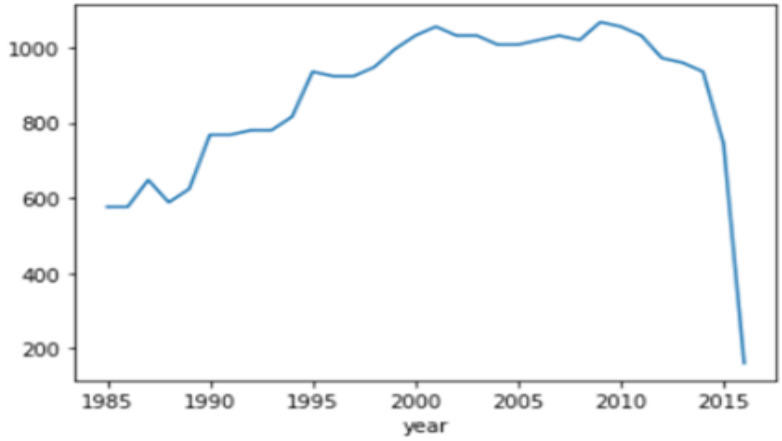

Рис. 2 – График общего количества самоубийств по годам

Для того, чтобы более наглядно отобразить статистику по экстремальным значениям выборки, можно воспользоваться функцией вывода n-количества максимальных и минимальных значений. В рассматриваемом случае было решено вывести по 5 наибольших и наименьших значений количества самоубийств, сгруппированных по годам (рис. 3).

```
In [6]: dataset.groupby(by=['year'])['suicides_no'].sum().nlargest(5)
Out[6]: year
        1999    256119
        2002    256095
        2003    256079
        2000    255832
        2001    250652
        Name: suicides_no, dtype: int64

In [7]: dataset.groupby(by=['year'])['suicides_no'].sum().nsmallest(5)
Out[7]: year
        2016     15603
        1985    116063
        1986    120670
        1988    121026
        1987    126842
        Name: suicides_no, dtype: int64
```

Рис. 3 – Вывод экстремальных значений количества самоубийств по годам

Исходя из полученной статистики видно, что наибольшее количество самоубийств в мире происходило в период с 1999 по 2003 годы. На это могло

повлиять большое количество факторов, главным из которых можно считать период глобальных финансовых кризисов [3].

Наименьшее количество самоубийств происходило в начале исследуемого периода – с 1985 по 1988 годы. Также из графика видно, что в 2014 году наметилась тенденция к снижению числа самоубийств, достигающая своего минимума в 2016 году. Это можно объяснить тем, что к настоящему времени количество самоубийств резко уменьшилось. Имеющиеся исследования по данной тематике утверждают, что увеличение общего уровня жизни населения по всему миру способствует уменьшению количества самоубийств в процентном соотношении к общему числу населения [4]. Конечно, для получения подробной статистики необходимо рассматривать факторы и их значения в разрезе каждой конкретной страны, однако, из полученного результата можно сформировать первоначальную гипотезу о том, что ухудшение мирового экономического состояния влияет на увеличение количества самоубийств по всему миру.

Так как была выдвинута гипотеза о том, что существует зависимость между экономическим состоянием в мире и статистикой самоубийств, было решено проследить зависимость количества самоубийств от значения мирового ВВП. Для построения графика была использована библиотека SkikitLearn. Функция Preprocessing в ней отвечает за трансформацию и нормализацию данных по определенному алгоритму. Нормализация данных была применена для того, чтобы привести несовместимые единицы измерений к соизмеримому виду [5]. Благодаря этому был исключен риск получения некорректных данных на графике. График зависимости количества самоубийств и ВВП изображен на рис. 4.

Исходя из полученных данных видно, что существует обратно пропорциональная зависимость между значением мирового ВВП и количеством самоубийств: при увеличении ВВП, количество самоубийств резко падает и, наоборот, с понижением уровня ВВП - растет. Предположительно, искажения с обеих сторон графиков связаны с неполнотой тестовых данных на ранние и поздние периоды.

```
In [6]: from sklearn import preprocessing
        suicides_gdp= dataset.pivot_table(['suicides/100k pop','gdp_per_capita ($)'],
                        ['year'], aggfunc='mean')
        x = suicides_gdp.values
        min_max_scaler = preprocessing.MinMaxScaler()
        x_scaled = min_max_scaler.fit_transform(x)
        suicides_gdp_scaled = pd.DataFrame(x_scaled)
        suicides_gdp_scaled.columns = suicides_gdp.columns
        suicides_gdp_scaled.index = suicides_gdp.index

In [7]: suicides_gdp_scaled.plot()

Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x222ad2d3048>
```

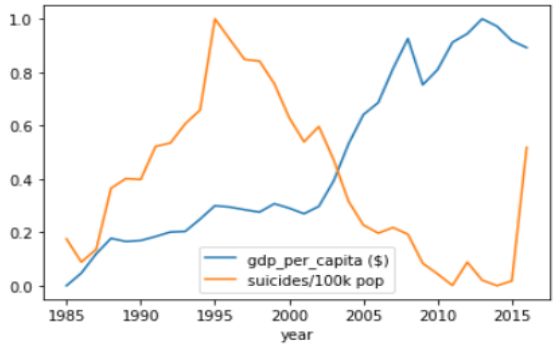

Рис. 4 – График зависимости ВВП и количества самоубийств

Далее было решено проследить распределение количества самоубийств по возрастным группам на протяжении всего наблюдаемого периода. Для построения графика было использовано несколько методов. Для начала данные были сгруппированы по годам и возрастным группам: количество самоубийств было просуммировано по представленным признакам. Далее были выделены основные категории, по которым строился график – возрастные группы. Для построения графика была использована библиотека Seaborn, которая основана на библиотеке Matplotlib, но обладает расширенным функционалом для визуализации данных [6]. Динамика числа самоубийств по годам для разных возрастных групп представлена на рис. 5.

Исходя из полученных данных видно, что на протяжении всего наблюдаемого периода, наибольшие количества самоубийств совершаются людьми в возрастной группе от 35 до 54 лет. В то время как наименьший показатель у людей в самой старшей возрастной группе от 75 лет.

Исходя из результатов проводимых исследований, в глобальном масштабе действительно наибольшее количество самоубийств совершается людьми среднего возраста, причем, в данной категории более склонны к суициду

мужчины. Среди людей 30–49 лет самоубийство — пятая по частоте причина смерти: на её долю приходится 4,1% общей смертности [7].

```
In [11]: import seaborn as sns
         from matplotlib.colors import LogNorm
         df = dataset.groupby(['year','age']).suicides_no.sum().reset_index()
         df['age'] = df.age.astype(pd.api.types.CategoricalDtype(categories = ['5-14 years','15-24 years','25-34 years',
                                                                                '35-54 years','55-74 years','75+ years']))
         sns.set(rc={'figure.figsize':(15,10)})
         sns.lineplot('year','suicides_no',hue='age',style='age',data=df,hue_norm=LogNorm(),palette="ch:2.5,.25",sort=False)

Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x1c93aeb5508>
```

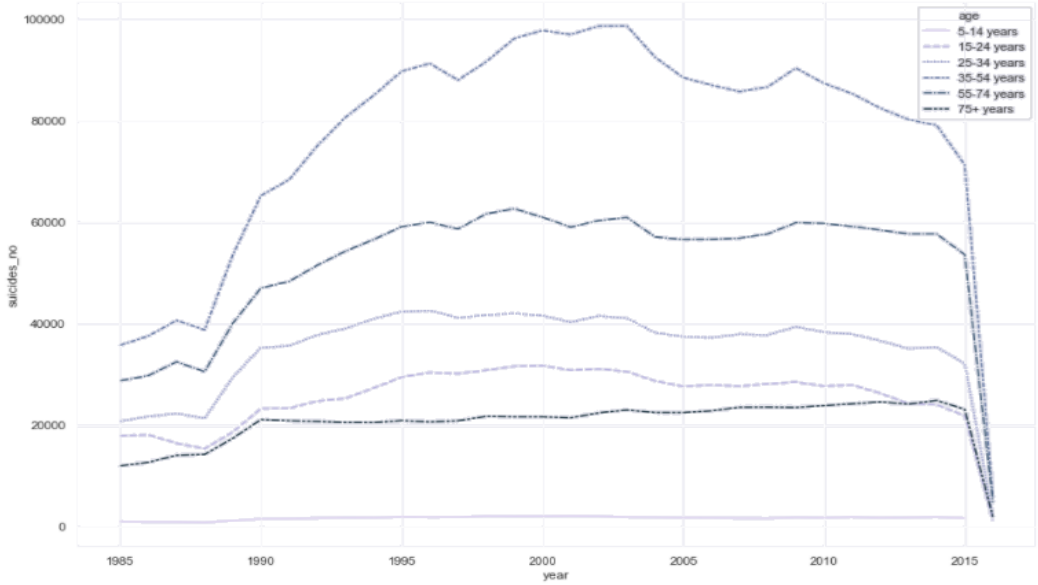

Рис. 5 – Динамика числа самоубийств по годам для разных возрастных групп

В результате выполнения работы было построено несколько графиков, позволяющих проследить некоторые основные закономерности. Были представлены основные гипотезы, на которые можно будет опираться при построении прогнозной модели:

1. После построения графика общего количества самоубийств по годам видно, что в настоящее время наблюдается самое минимальное количество самоубийств за весь рассматриваемый период.

2. Во времена тяжелых экономических кризисов происходят резкие скачки количества самоубийств по всему миру.

3. Исходя из графика статистики самоубийств по возрастным группам видно, что наибольшее количество самоубийств совершают люди среднего возраста (35-54 года), наименьшее – люди старшего возраста (75+).

4. Существует обратно-пропорциональная зависимость между общим количеством самоубийств в стране и размером ее ВВП: в странах с более

стабильной экономической обстановкой и высоким уровнем жизни количество самоубийств меньше, чем в странах с низким уровнем жизни.

Благодаря визуализации данных можно наглядно представить зависимость результата от других факторов, не прибегая к высокоуровневым инструментам анализа данных, что позволяет на первых этапах проведения анализа отыскать важные зависимости, на которые впоследствии можно будет опираться, при выполнении других операций. Таким образом, в результате выполнения работы была описана проблема выявления первоначальных аналитических гипотез в разрезе рассматриваемой темы, и предложены пути ее решения. Предлагаемы в статье методики могут применяться на данных любой структуры и размерности. Поставленные в данной работе гипотезы будут проверяться при выполнении дальнейшей работы по построению прогнозной модели.

Список используемых источников

1. Рабинович А.Е., Седова Е.А. Применение технологий больших данных для исследования влияния социально-экономических факторов на состояние психического здоровья населения // Российский Экономический Интернет Журнал. – 2020. – №1.

2. Preventing suicide: a global imperative // Geneva: World health organization. – 2015.

3. Лялюцкая М.Ю., Галакова Г.А., Юрковец Н.В., Потапаева О.П. Суицидальное поведение людей в период кризиса // Актуальные проблемы авиации и космонавтики. 2015. №11. URL: https://cyberleninka.ru/article/n/suitsidalnoe-povedenie-lyudey-v-period-krizisa (дата обращения: 27.09.2020).

4. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet. 2018 // http://dx.doi.org/10.1016/S0140-6736(18)32203-7.

5. Preprocessing Data // SkikitLearn URL: https://scikit-learn.org/stable/modules/preprocessing.html (дата обращения: 09.09.2020).

6. Seaborn: statistical data visualization // Seaborn URL: https://seaborn.pydata.org/ (дата обращения: 13.09.2020).

7. Суицид: статистика // Batenka URL: https://batenka.ru/resource/suicide/math/ (дата обращения: 16.09.2020).

References

1. Rabinovich A.E., Sedova E.A. The use of big data technologies to study the influence of socio-economic factors on the state of mental health of the population // Russian Economic Internet Journal. – 2020. – № 1.

2. Preventing suicide: a global imperative // Geneva: World health organization. – 2015.

3. Lyalutskaya M.Yu., Galakova G.A., Yurkovets N.V., Potapaeva O.P. Suicidal behavior of people during the crisis // Actual problems of aviation and cosmonautics. – 2015. – № 11. URL: https://cyberleninka.ru/article/n/suitsidalnoe-povedenie-lyudey-v-period-krizisa (date accessed: 09.27.2020).

4. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet. 2018 // http://dx.doi.org/10.1016/S0140-6736(18)32203-7.

5. Preprocessing Data // SkikitLearn URL: https://scikit-learn.org/stable/modules/preprocessing.html (date accessed: 09.09.2020).

6. Seaborn: statistical data visualization // Seaborn URL: https://seaborn.pydata.org/ (date accessed: 09.13.2020).

7. Suicide: statistics // Batenka URL: https://batenka.ru/resource/suicide/math/ (date accessed: 16.09.2020).