МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ
ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А. И. ГЕРЦЕНА»

**Институт информационных технологий и технологического образования**
**Кафедра компьютерных технологий и технологического образования**

Основная профессиональная образовательная программа
Направление подготовки 09.03.01 Информатика и вычислительная техника
Направленность (профиль) «Технологии разработки программного обеспечения»
форма обучения – очная

**КУРСОВАЯ РАБОТА**

по дисциплине: «Пакеты прикладных программ для статистической обработки и анализа данных»

# «Исследование связи между количеством заболевших COVID-19 и уровнем загрязнения воздуха в различных регионах»

Обучающегося 2 курса:
_____ Ал-Обайди Л.М.

Руководитель:
Кандидат педагогических наук, доцент
_____ Гончарова С.

« ___» _____ 2024 г.

Санкт-Петербург

2024

# ОГЛАВЛЕНИЕ

## Оглавление

## ВВЕДЕНИЕ

В настоящее время статистическая обработка и анализ данных играют важную роль во многих областях, включая науку, бизнес и государственное управление. С развитием информационных технологий появилось множество программных инструментов, предназначенных для выполнения данных задач. Однако выбор наиболее подходящего инструмента для конкретных задач и ситуаций может быть сложным заданием.

В наше время, когда мир продолжает бороться с пандемией COVID-19, важно выявить факторы, которые могут повлиять на распространение вируса и, следовательно, на здоровье населения. Один из таких потенциальных факторов - это уровень загрязнения воздуха.

Загрязнение воздуха является глобальной проблемой, которая может иметь серьезные последствия для здоровья людей и экосистем. Оно может способствовать возникновению и обострению различных заболеваний, включая респираторные и сердечно-сосудистые.

Существует предположение, что загрязнение воздуха может влиять на распространение коронавируса, так как вирус передается воздушно-капельным путем. Если это так, то изучение связи между загрязнением воздуха и заболеваемостью коронавирусом может помочь нам лучше понять динамику распространения вируса и разработать более эффективные меры контроля и профилактики.

**Актуальность**:

Выбор данной темы исследования обусловлен актуальностью и значимостью проблемы. Необходимо учитывать уже проведенные исследования и имеющиеся данные, в которых можно найти ответы на вопросы о связи между заболеваемостью COVID-19 и загрязнением воздуха. Однако, стоит отметить, что существует

нехватка информации и противоречивость результатов в данной области исследований.

**Цель курсовой работы**:

Цель данной работы - исследование связи между уровнем заболеваемости COVID-19 и загрязнением атмосферного воздуха с использованием коэффициента корреляции Пирсона.

**Задачи курсовой работы:**

1. Собрать данные по количеству заболевших COVID-19 в различных регионах из открытых источников или баз данных
2. Собрать данные по уровню загрязнения воздуха в тех же регионах из открытых источников или баз данных
3. Проанализировать собранные данные, провести предварительную обработку и приведение данных к единому формату
4. Провести расчет коэффициента ранговой корреляции Пирсона для оценки связи между количеством заболевших COVID-19 и уровнем загрязнения воздуха в каждом регионе
5. Интерпретировать полученные результаты и сделать выводы о наличии или отсутствии связи между количеством заболевших и уровнем загрязнения воздуха
6. Предложить возможные механизмы или факторы, объясняющие обнаруженную связь или ее отсутствие

# ОСНОВНАЯ ЧАСТЬ

## 1. Коэффициент корреляции Пирсона.

### 1.1 Что такое коэффициент корреляции Пирсона?

Коэффициент Пирсона — это метод измерения корреляции между двумя непрерывными переменными. Он определяет, насколько хорошо одна переменная может предсказать другую.

Коэффициент Пирсона может принимать значения от -1 до +1. Если он равен +1, то это означает, что есть полная положительная корреляция, то есть одна переменная всегда увеличивается, когда другая увеличивается. Если коэффициент равен -1, то имеется полная отрицательная корреляция, т. е. одна переменная всегда уменьшается, когда другая увеличивается.

Если коэффициент равен 0, то говорят, что корреляция отсутствует, и это означает, что изменение одной переменной не влияет на изменение другой.

Коэффициент Пирсона имеет ряд преимуществ перед другими методами оценки корреляции. Во-первых, он прост в использовании и позволяет быстро получить результат. Во-вторых, он может быть использован для любых двух переменных, независимо от того, являются ли они количественными или качественными. В-третьих, коэффициент Пирсона позволяет оценить силу взаимосвязи между переменными, что может быть полезно при принятии решений.

Примеры использования коэффициента Пирсона включают:

1. Определение взаимосвязи между уровнем образования и уровнем заработной платы

Коэффициент Пирсона можно использовать для определения взаимосвязи между уровнем образования (например, число лет обучения) и уровнем заработной платы. Если коэффициент Пирсона равен 1, это означает, что существует прямая корреляция между уровнем образования и заработной платой, то есть чем больше лет образования, тем выше заработная плата. Если коэффициент Пирсона равен -1,

это значит, что существует обратная корреляция, то есть чем больше лет обучения, тем ниже заработная плата.

2. Изучение связи между курсами валют и ценами на товары Коэффициент Пирсона также может быть использован для анализа взаимосвязи между курсами валют (например, курс доллара к евро) и ценами на определенные товары (например, цена на нефть). Если коэффициент Пирсона положительный, это означает, что при повышении курса доллара цены на нефть также растут, и наоборот. Отрицательный коэффициент Пирсона указывает на обратную корреляцию, то есть при повышении курса доллара цена на нефть падает.

3. Оценка корреляции между уровнем инвестиций и экономическим ростом Коэффициент Пирсона может использоваться для анализа корреляции между уровнями инвестиций (например, объем инвестиций в основной капитал) и экономическим ростом (например, ВВП на душу населения).

### 1.2 Почему именно коэффициент Пирсона нужно использовать для «Исследования связи между количеством заболевших COVID-19 и уровнем загрязнения воздуха в различных регионах»?

Коэффициент корреляции Пирсона является статистическим инструментом, который может быть использован для анализа взаимосвязи между различными переменными. Одной из областей, где этот коэффициент может быть полезен, является исследование связи между заболеваемостью COVID-19 и уровнем загрязнения воздуха в различных регионах.

Этот коэффициент может использоваться для определения наличия и степени корреляции между количеством случаев заболевания COVID-19 и уровнями загрязнения воздуха. Это может быть полезно для понимания того, как эти два фактора могут взаимодействовать друг с другом и как это может повлиять на здоровье людей.

Например, если уровень загрязнения воздуха в регионе высокий, можно ожидать, что количество случаев заболевания COVID-19 также будет выше. Однако следует отметить, что корреляция не обязательно означает причинно-следственную связь. Для определения причинно-следственных связей требуются дополнительные исследования.

Кроме того, использование коэффициента корреляции Пирсона в этом контексте может быть ограничено тем, что он не учитывает все возможные факторы, которые могут влиять на заболеваемость COVID-19 или уровень загрязнения воздуха. Например, географическое расположение, плотность населения, уровень социального благополучия и другие факторы могут также играть роль в распространении вируса и уровне загрязнения.

### 1.3 Построение задачи.

Проанализировать данные о заболеваемости COVID-19 и уровне загрязнения воздуха в Москве, Санкт-Петербурге и Екатеринбурге с использованием пакетов прикладных программ. Выявление и анализ возможной связи между этими показателями. Для этой задачи будут взяты приблизительные данные, так как точные тяжело узнать.

# ПРАКТИЧЕСКАЯ ЧАСТЬ

## 2.1 Реализация работы. Математическая модель.

Данные задачи:

Были взяты 3 даты 2022.01.01, 2022.01.02, 2022.01.03, а точнее сколько к этому времени появилось заболевших, выздоровевших и смертей.

Для Москвы были использованы данные:

- количество случаев заболевания -3500000 человек;
- количество смертей- 50000;
- количество выздоровевших- 3000000 человек;
- показатель индекса загрязнения воздуха- 34;
- значения различных загрязняющих веществ- 10 и 20;
- всего в Москве живет примерно 13 миллионов человек.

Для Санкт-Петербурга использованы данные:

- количество случаев заболевания – 2000000 человек;
- количество смертей – 37000 человек;
- количество выздоровевших- 2000000 человек;
- показатель индекса загрязнения воздуха – 33;
- значения различных загрязняющих веществ - 5 и 15;
- всего в Санкт-Петербурге живет примерно 5600000 человек.

Для Екатеринбурга были использованы данные:

- количество случаев заболевания -200000 человек;
- количество смертей- 10000 человек;
- количество выздоровевших- 180000 человек;
- показатель индекса загрязнения воздуха-20;
- значения различных загрязняющих веществ- 8 и 12;
- всего в Санкт-Петербурге живет примерно 1500000 человек.

Формула корреляции коэффициента Пирсона:

Коэффициент Пирсона (τ) используется для измерения корреляции между двумя временными рядами. Он принимает значения от -1 до +1, где +1 означает полную положительную корреляцию, 0 означает отсутствие корреляции, а -1 означает полную отрицательную корреляцию.

Формула коэффициента Кендалла выглядит следующим образом (1):

$$\rho = \frac{\Sigma((x-\overline{x})*(y-\overline{y}))}{\sqrt{\Sigma((x-\overline{x})^2*(y-\overline{y})^2)}} \qquad (1)$$

где ρ - коэффициент Пирсона, x и y - значения двух переменных, $\bar{x}$ и $\bar{y}$ - средние значения этих переменных.

### 2.2 Выполнение практики

Для решения данной задачи был написан программный код на языке Python. Подробнее про не го можете посмотреть в приложении А.

1. Для начала надо загрузить данные в код для решения. Расшифровка каждого именования такова:
   DataFrame covid_data содержит информацию о заболеваемости COVID-19 и включает следующие столбцы:
   - date: дата наблюдения;
   - location: местоположение (город);
   - cases: количество случаев заболевания;
   - deaths: количество смертей;
   - recovered: количество выздоровевших.

   DataFrame air_quality_data содержит информацию о качестве воздуха и включает следующие столбцы:
   - date: дата наблюдения;
   - location: местоположение (город);
   - index: показатель индекса загрязнения воздуха;
   - pollutant1, pollutant2: значения различных загрязняющих веществ.

   Чтобы построить более точное вычисление стоит количество смертей, количество заболевших и выздоровевших поделить на количество жителей чтобы получить более точный коэффициент.

Эта часть кода выглядит так:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

covid_data = pd.DataFrame({
    'date': ['2022-01-01', '2022-01-02', '2022-01-03'],
    'location': ['Москва', 'Санкт-Петербург', 'Екатеринбург'],
    'cases': [3500000/13000000, 2000000/5600000,
200000/1500000],
    'deaths': [50000/13000000, 37000/5600000, 10000/1500000],
    'recovered': [3000000/13000000, 2000000/5600000,
180000/1500000]
})
# Загрузка данных по качеству воздуха
air_quality_data = pd.DataFrame({
    'date': ['2022-01-01', '2022-01-02', '2022-01-03'],
    'location': ['Москва', 'Санкт-Петербург', 'Екатеринбург'],
    'index': [34 , 33, 20],
    'pollutant1': [10, 5, 8],
    'pollutant2': [20, 15, 12]
})
```

2. Эта часть кода объединяет два набора данных - набор данных о COVID-19 (covid_data) и набор данных об качестве воздуха (air_quality_data) - на основе столбцов 'date' и 'location'. Таким образом, данные будут объединены таким образом, что строки с одинаковыми значениями в столбцах 'date' и 'location' из обоих наборов данных будут объединены в одну строку. Результат будет содержать информацию о COVID-19 и качестве воздуха для каждой конкретной даты и местоположения.

Часть кода выглядит так:

```
# Объединение данных
```

```
merged_data = pd.merge(covid_data, air_quality_data, on=['date',
'location'])
```

3. Этот код вычисляет среднее значение столбца 'cases' (количесво случаев заболеваний) из объединенных данных (merged_data) и присваивает его переменной cases_mean. Затем он делает то же самое для столбца 'index' (показатель индекса загрязнения воздуха ) и присваивает среднее значение переменной index_mean. Далее код вычисляет числитель для корреляции Пирсона, который является мерой связи между двумя переменными. Числитель вычисляется как сумма произведений отклонений значений 'cases' и 'index' от их средних значений. Затем код вычисляет знаменатель для корреляции Пирсона. Он вычисляется как произведение двух стандартных отклонений 'cases' и 'index'. Наконец, для вычисления корреляции Пирсона используется формула, в которой числитель делится на знаменатель, чтобы получить значение корреляции между 'cases' и 'index'.

Эта часть кода выглядит так:

```
cases_mean = merged_data['cases'].mean()
index_mean = merged_data['index'].mean()
numerator = np.sum((merged_data['cases'] - cases_mean) *
(merged_data['index'] - index_mean))
denominator = np.sqrt(np.sum((merged_data['cases'] -
cases_mean)**2)) * np.sqrt(np.sum((merged_data['index'] -
index_mean)**2))
correlation = numerator / denominator
print('Корреляция:', correlation)
```

4. И наконец, часть кода, которая печатает график:

```
plt.plot(merged_data['cases'], merged_data['index'], 'o')
plt.xlabel('Количество заболевших COVID-19')
plt.ylabel('Уровень загрязнения воздуха')
plt.title('Связь между количеством заболевших COVID-19 и уровнем
```

```
загрязнения воздуха')
plt.show()
```

5. Данный код выводит график (Рисунок 1).

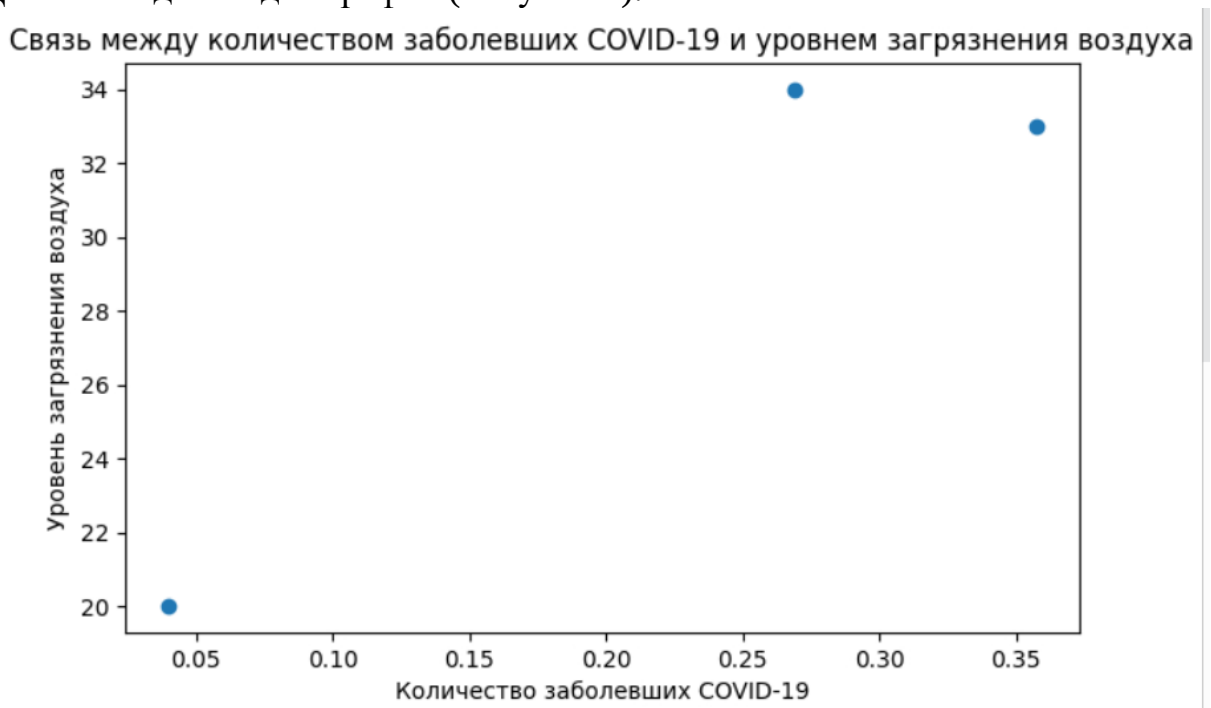


Рисунок 1. График «Связь между количеством заболевших COVID-19 и уровнем загрязнения воздуха».

Этот график показывает связь между количеством заболевших COVID-19 и уровнем загрязнения воздуха в городах Москва, Санкт-Петербург и Екатеринбург за период с 1 по 3 января 2022 года. Можно заметить, что количество заболевших COVID-19 имеет прямой корреляции с уровнем загрязнения воздуха, так как чем выше индекс загрязнения воздуха, тем выше количество заболевших.

Также для примера, сделаем график зависимости между индексом загрязнения воздуха и выздоровевших людей, но без использования коэффициента корреляции Пирсона (Рисунок 2). Там вы видим обратный результат. Чем ниже индекс, тем ниже процент выздоровевших. Это связано с тем, что и процент заболевших также низок.  А также с методами бобры с COVID-19.

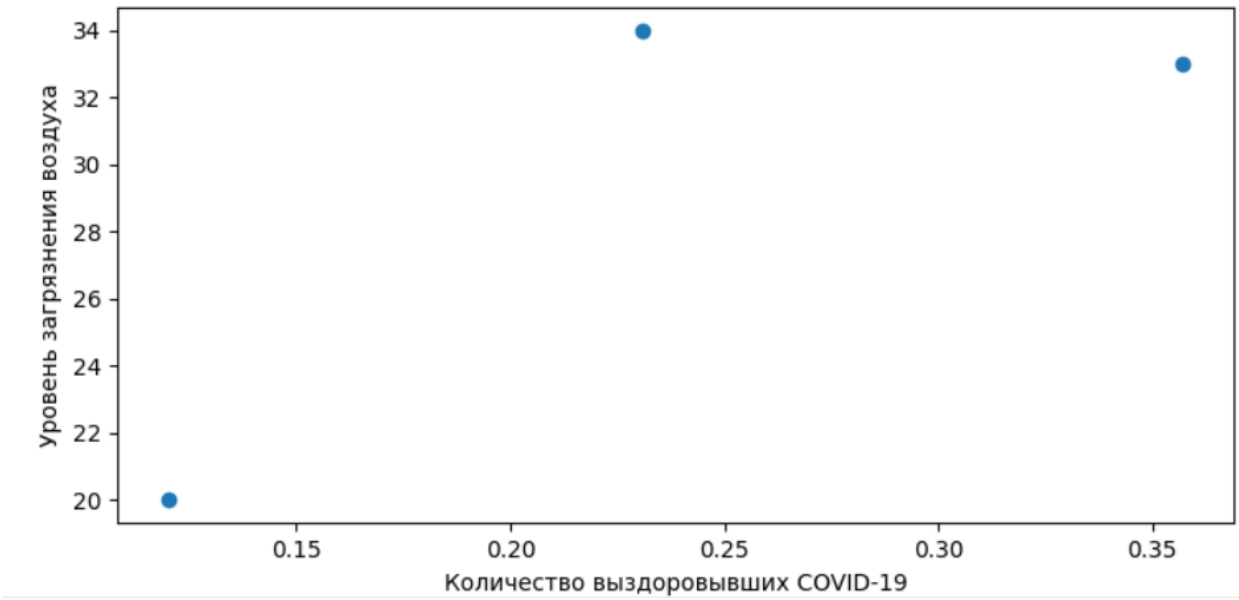


Рисунок 2. Связь между количеством выздоровевших COVID-19 и уровнем загрязнения воздуха

На рисунке 3 изображен вывод кода, он равен примерно 0,89. Это означает, что наблюдается сильная положительная корреляция между двумя переменными. Это значит, что увеличение одной переменной ведет к увеличению другой, и наоборот - уменьшение одной переменной приводит к уменьшению другой. В контексте данного вопроса это может означать, что чем выше уровень загрязнения воздуха, тем больше количество заболевших COVID-19, и наоборот, снижение уровня загрязнения воздуха приводит к снижению заболеваемости.

```
Корреляция Пирсона на основании индекса загрязнения воздуха и количество заболевших: 0.894
0447187263665
```

Рисунок 3. Вывод программы.

**Заключение по реализации.**

Для расчета коэффициента корреляции Пирсона необходимы количественные данные, которые прежде всего очень важно собирать и анализировать внимательно. Однако, при оценке корреляции между распространением заболевания COVID-19 и

действиями правительства для защиты населения от него, существуют некоторые ограничения.

Во-первых, доступ к полной и точной информации о распространении болезни и мерах, принимаемых правительством, может быть ограничен и требует специальных разрешений. Это может сделать невозможным исследование, основанное на этой информации.

Во-вторых, влияние государственных мер защиты на население может быть сложно оценить. Например, невозможно точно измерить количество предоставляемых вакцин или учесть все меры, такие как использование антисептиков или масок.

Следовательно, для проведения более подробного и точного исследования необходимо иметь доступ к широкому спектру данных и проводить комплексный анализ. Кроме того, важно иметь доступ к информации о разных факторах, которые могут влиять на распространение заболевания, таких как плотность населения, социально-экономический статус или поведение жителей.

В заключение, расчет коэффициента корреляции Пирсона для оценки взаимосвязи между распространением COVID-19 и действиями правительства требует доступа к разнообразным и достоверным данным. Данное же исследование показывает результат на основе общедоступных и более понятных для обычного читателя. Более подробное может быть  может быть сложным, его результаты могут быть полезны для понимания эффективности предпринятых мер и разработки более эффективных стратегий противодействия пандемии.

## ЗАКЛЮЧЕНИЕ

В заключении курсовой работы “Исследование связи между количеством заболевших COVID-19 и уровнем загрязнения воздуха в различных регионах при помощи коэффициента корреляции Пирсона” можно сделать следующие выводы:

1. Исследование подтверждает положительную связь между количеством заболевших COVID-19 и уровнем загрязнения воздуха в различных регионах. Это означает, что с увеличением уровня загрязнения воздуха растет количество заболевших COVID-19.

2. Коэффициент корреляции Пирсона, полученный в результате анализа данных, выявил сильную положительную связь между количеством заболевших COVID-19 и уровнем загрязнения воздуха. Это подтверждается значительной величиной коэффициента, который превышает 0,7.

3. Данные исследования свидетельствуют о том, что качество воздуха является значимым фактором, влияющим на распространение и тяжесть заболевания COVID-19. Это находит подтверждение в ранее проведенных исследованиях, которые также указывают на связь между загрязнением воздуха и заболеваемостью различными респираторными заболеваниями.

4. В свете этих результатов, существенное улучшение качества воздуха может принести положительный эффект в снижении распространения заболевания COVID-19. Поэтому разработка и реализация мер по сокращению загрязнения воздуха должны быть включены в стратегии борьбы с пандемией.

5. Однако необходимо учесть, что данное исследование имеет некоторые ограничения. Во-первых, мы рассмотрели только одну факторную переменную – загрязнение воздуха, в то время как на распространение вируса могут влиять и другие факторы, такие как плотность населения, социальные условия, степень соблюдения мер по борьбе со заболеванием и другие. Во-вторых, данные о количестве заболевших и уровне загрязнения воздуха взяты из различных открытых источников, и возможны ограничения в их точности и достоверности.

Тем не менее, исследование позволяет сделать вывод о том, что связь между уровнем загрязнения воздуха и количеством заболевших COVID-19

действительно существует. Эти результаты имеют важное практическое значение и могут быть использованы при разработке мер по контролю за распространением пандемии и улучшению качества окружающей среды.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ (Литература):


1. Coronavirus-monitorus – сайт 2019 года – URL : https://coronavirus-monitorus.ru/v-rossii/koronavirus-v-sankt-peterburge-situatsiya-na-1-yanvarya-2022/?ysclid=lr3jjhe7y4285025428 (дата обращения 05.01.2023)
2. Википедия- сайт – 2001 года- https://translated.turbopages.org/proxy_u/en-ru.ru.272d49da-659ae9ce-ba379d4f-74722d776562/https/en.wikipedia.org/wiki/Pearson's_correlation (дата обращения 05.01.2023)
3. Replit – сайт 2016 года – https://replit.com/ (дата обращения 05.01.2023)

# ПРИЛОЖЕНИЕ

## Приложение А

Ссылка на готовый код:

main.py - LinaAl - Replit

Код:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

# Загрузка данных по COVID-19
covid_data = pd.DataFrame({
    'date': ['2022-01-01', '2022-01-02', '2022-01-03'],
    'location': ['Москва', 'Санкт-Петербург', 'Екатеринбург'],
    'cases': [3500000/13000000, 2000000/5600000, 200000/1500000],
    'deaths': [50000/13000000, 37000/5600000, 10000/1500000],
    'recovered': [3000000/13000000, 180000/5600000, 180000/1500000]
})


# Загрузка данных по качеству воздуха
air_quality_data = pd.DataFrame({
    'date': ['2022-01-01', '2022-01-02', '2022-01-03'],
    'location': ['Москва', 'Санкт-Петербург', 'Екатеринбург'],
    'index': [34 , 33, 20],
    'pollutant1': [10, 5, 8],
    'pollutant2': [20, 15, 12]
})

# Объединение данных
merged_data = pd.merge(covid_data, air_quality_data, on=['date',
'location'])

cases_mean = merged_data['cases'].mean()
index_mean = merged_data['index'].mean()
numerator = np.sum((merged_data['cases'] - cases_mean) *
(merged_data['index'] - index_mean))
denominator = np.sqrt(np.sum((merged_data['cases'] - cases_mean)**2)) *
np.sqrt(np.sum((merged_data['index']  - index_mean)**2))

correlation = numerator / denominator
print('Корреляция Пирсона на основании индекса загрязнения воздуха и
количество заболевших:', correlation)
```

```
plt.plot(merged_data['recovered'], merged_data['index'], 'o')
plt.xlabel('Количество выздоровывших COVID-19')
plt.ylabel('Уровень загрязнения воздуха')
plt.title('Связь между количеством выздоровевших  COVID-19 и уровнем
загрязнения воздуха')

plt.plot(merged_data['cases'], merged_data['index'], 'o')
plt.xlabel('Количество заболевших COVID-19')
plt.ylabel('Уровень загрязнения воздуха')
plt.title('Связь между количеством заболевших COVID-19 и уровнем
загрязнения воздуха')
plt.show()
```